Ελληνικό Στατιστικό Ινστιτούτο

Greek Statistical Institute

28ο Πανελλήνιο Συνέδριο Στατιστικής

Αθήνα

15-18 Απριλίου 2015

Θέμα συνεδρίου

«Χωρική Στατιστική και εφαρμογές της Στατιστικής στη Γενετική και την Πληροφορική»

Χαροκόπειο Πανεπιστήμιο

Πληροφορίες 210 3303909 & www.esi-stat.gr

Περιεχόμενα

ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ
(Ε.Σ.Ι)
GREEK STATISTICAL INSTITUTE
(G.S.I)

Π Ρ Α Κ Τ Ι Κ Α

28ου Πανελληνίου
Συνεδρίου Στατιστικής

PROCEEDINGS

of the 28th Panhellenic
Statistics Conference

Χωρική Στατιστική και Εφαρμογές της Στατιστικής στη Γενετική και την Πληροφορική

Spatial Statistics and Applications of Statistics in Genetics and Informatics

Αθήνα, 15-18 Απριλίου 2015

ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ
(Ε.Σ.Ι)
GREEK STATISTICAL INSTITUTE
(G.S.I.)

Π Ρ Α Κ Τ Ι Κ Α

28ου Πανελληνίου
Συνεδρίου Στατιστικής

Χωρική Στατιστική και Εφαρμογές της Στατιστικής στη Γενετική και την Πληροφορική

Οργάνωση

ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ

ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Αθήνα, 15-18 Απριλίου 2015

ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ
Σολωμού 5 (Πλατεία Εξαρχείων)
Τηλ.& Fax 210 33.03.909
Email: esi-stat@hol.gr
http://www.esi-stat.gr

ISBN: 978-618-80672-6-4
ISSN: 1792-2461

ΠΕΡΙΕΧΟΜΕΝΑ

Εργασίες στα Αγγλικά

ΠΡΟΛΟΓΟΣ

Το 28ο Πανελλήνιο Συνέδριο Στατιστικής διοργανώθηκε από το Ελληνικό Στατιστικό Ινστιτούτο (Ε.Σ.Ι.) στην Αθήνα την περίοδο 15-18 Απριλίου 2015 σε συνεργασία με το Χαροκόπειο Πανεπιστήμιο. Το θέμα του Συνεδρίου ήταν *Χωρική Στατιστική και Εφαρμογές της Στατιστικής στη Γενετική και την Πληροφορική*.

Έλαβαν μέρος 182 Σύνεδροι, από τους οποίους 132 ήταν φοιτητές και 13 συνοδεύοντα μέλη από την Ελλάδα, την Κύπρο και τις ΗΠΑ. Η συμμετοχή των προπτυχιακών και των μεταπτυχιακών φοιτητών υπήρξε ικανοποιητική, συγκρίσιμη με το παρελθόν.

Στο Συνέδριο προσκεκλημένοι ομιλητές ήταν ο κ. Ιωάννης Τσιτσικλής, Πρόεδρος του Συμβουλίου του Χαροκοπείου Πανεπιστημίου και Καθηγητής στο Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Ηλεκτρονικών Υπολογιστών του Τεχνολογικού Ινστιτούτου της Μασαχουσέτης (MIT), ο κ. Ευθύμιος Λέκκας, Καθηγητής Δυναμικής Τεκτονικής Εφαρμοσμένης Γεωλογίας του Τμήματος Γεωλογίας και Γεωπεριβάλλοντος του ΕΚΠΑ, Αντιπρόεδρος Επιτροπής Ερευνών Στατιστικής, Πρόεδρος του Οργανισμού Αντισεισμικού Σχεδιασμού και Προστασίας και Πρόεδρος της Ελληνικής Γεωλογικής Εταιρείας, η Dr. *I. Tachmazidou*, The Wellcome Trust Sanger Institute, Human Genetics και ο Καθηγητής *M. Viana* του University of Illinois at Chicago με τίτλο «Dihedral reduction of cyclic DNA sequences».

Το Επιστημονικό Πρόγραμμα ήταν ιδιαίτερα πλούσιο και περιελάμβανε ογδόντα μία (81) ανακοινώσεις και δώδεκα (12) αναρτημένες ανακοινώσεις (posters) σε 22 παράλληλες Συνεδρίες σε θέματα Πειραματικών σχεδιασμών, Βιοστατιστικής, Πρόβλεψης-Χρονοσειρών, Στατιστικής στην Πληροφορική και την Οικονομία, Εφαρμοσμένης Στατιστικής, Εφαρμοσμένων Πιθανοτήτων, Στοχαστικών Διαδικασιών, Κοινωνικής Στατιστικής, Πιθανοτήτων-Αναλογισμού κτλ.

Η εναρκτήρια τελετή του Συνεδρίου πραγματοποιήθηκε στο αμφιθέατρο Γ. Καραμπατζός του Χαροκοπείου Πανεπιστημίου. Την έναρξη του Συνεδρίου κήρυξε ο Πρύτανης του Χαροκοπείου Πανεπιστημίου Καθηγητής Δημοσθένης Αναγνωστόπουλος. Χαιρετισμούς απηύθυναν ο Πρόεδρος του ΕΣΙ Καθηγητής Χαράλαμπος Δαμιανού, ο Καθηγητής Δημοσθένης Παναγιωτάκος, μέλος του ΔΣ του ΕΣΙ και της Οργανωτικής Επιτροπής του Συνεδρίου. Επίσης χαιρετισμούς απηύθυνε η Αντιπρύτανης Ακαδημαϊκών Υποθέσεων και Προσωπικού Καθηγήτρια Ευαγγελία Γεωργιτσογιάννη. Η τελετή έναρξης συνεχίστηκε με τις διακεκριμένες διαλέξεις των προσκεκλημένων ομιλητών *Ιωάννη Τσιτσικλή* με τίτλο «On

Influence Learning in Networks» και *Ευθύμιου Λέκκα,* με τίτλο «Χωροχρονική κατανομή των σεισμικών φαινομένων»του προσκεκλημένου ομιλητή, Καθηγητή Γεώργιου Μιχαηλίδη με τίτλο «*Big Data: Some theoretical and practical challenges*». Μετά την εναρκτήριο τελετή ακολούθησε Ειδική Συνεδρία με θέμα «*Σύζευξη της αγοράς με τη Στατιστική, Πληροφορική και Διοίκηση Επιχειρήσεων*». Την Παρασκευή 17 Απριλίου στα πλαίσια της συνεδρίας της Ολομέλειας στην Αίθουσα Β πραγματοποιήθηκαν δύο ακόμα διακεκριμένες διαλέξεις των Dr. *I. Tachmazidou* με τίτλο «Genome-wide association studies: In search of common and low frequency variants in complex traits» και Καθηγητή *M. Viana* με τίτλο «Dihedral reduction of cyclic DNA sequences».

Στη συνέχεια ο Πρόεδρος του ΕΣΙ Αν. Καθηγητής Χαράλαμπος Δαμιανού απένειμε το **Ελένειο Βραβείο Διδακτορικής Διατριβής** στη Στατιστική στον κ. *Δημήτριο Λυμπερόπουλο*, ο οποίος παρουσίασε τη διατριβή του με τίτλο «Martingale- ισοδύναμες κατανομές πιθανότητας με εφαρμογές στις αρχές υπολογισμού ασφαλίστρων». Ο κ. Δημήτριος Λυμπερόπουλος εκπόνησε τη διδακτορική διατριβή του στο Πανεπιστήμιο Πειραιώς με τον Αν. Καθηγητή κ. Νικόλαο Μαχαιρά. Μέλη της Επιτροπής του Ελένειου Βραβείου ήταν ο Καθηγητής Σταύρος Κουρούκλης, ο Αν. Καθηγητής Ιωάννης Ντζούφρας και ο Καθηγητής Γεώργιος Ηλιόπουλος.

Για το βραβείο υποβλήθηκαν και οι ακόλουθες πέντε (5) διατριβές:

1. **Βερυκούκη Ελένη**, *Stochastic Modelling and Bayesian Inference for the Effect of Antimicrobial Treatments on Transmission and Carriage of Nosocomial Pathogens*, University of Nottingham, Supervisors: Philip O' Neil & Theodore Kypraios

2. **Λυμπερόπουλος Δημήτρης**, *Martingale- Ισοδύναμες κατανομές πιθανότητας με εφαρμογές στις αρχές υπολογισμού ασφαλίστρου*, Πανεπιστήμιο Πειραιώς, Επιβλέπων: Νικόλαος Μαχαιράς

3. **Ξιφαρά Τατιανή**, *Bayesian inference on a coupled hidden Markov model for disease interactions and a new position dependent Metropolis adjusted*, Lancaster University, Supervisor: Chris Sherlock

4. **Τσαγρής Μιχαήλ**, *Novel methods for the statistical analysis of compositional data*, University of Nottingham, Supervisors: Andy Wood & Simon Preston

Για ένατη χρονιά απονεμήθηκε το **Βραβείο Καλύτερης Εργασίας Νέου Στατιστικού.** Για το σκοπό αυτό υποβλήθηκαν τέσσερις (4) εργασίες, οι οποίες και παρουσιάστηκαν σε ξεχωριστή συνεδρία (βλ. Πρόγραμμα Συνεδρίου). Την τελευταία μέρα του Συνεδρίου, η τριμελής Επιτροπή, αποτελούμενη από τους Καθηγήτρια Α. Καλαματιανού, Καθηγητή Α. Μπουρνέτα και η Αν. Καθηγήτρια Μ. Βαμβακάρη, κατέθεσε τη βαθμολογία της και το Βραβείο Καλύτερης Εργασίας Νέου Στατιστικού για το 2015 απονεμήθηκε στον ***κ. Κωνσταντίνο Τασιά***, για την από κοινού εργασία του με τον Αν. Καθηγητή κ. Γεώργιο Νένε με τίτλο: «A fully adaptive control scheme for joint monitoring of location and scale of processes subject to a multiplicity of assignable causes».

Οι κοινωνικές εκδηλώσεις ήταν πλούσιες: Cocktail προσφέρθηκε την πρώτη μέρα του Συνεδρίου στο χώρο του αμφιθεάτρου Γ. Καραμπατζός του Χαροκοπείου Πανεπιστημίου. Οι Σύνεδροι είχαν τη δυνατότητα να παρακολουθήσουν την Πέμπτη 16 Απριλίου την ταινία μεγάλης επιφάνειας «Κρυμμένο Σύμπαν» στο Ίδρυμα Ευγενίδου. Το βράδυ της ίδιας ημέρας διοργανώθηκε το επίσημο δείπνο του Συνεδρίου στο restaurant «Πισίνα» στη Μαρίνα Ζέας.

Στον τόμο αυτό περιλαμβάνονται εργασίες που παρουσιάστηκαν στο Συνέδριο και υποβλήθηκαν για δημοσίευση. Όλες οι εργασίες κρίθηκαν από κριτές με την φροντίδα των υπευθύνων έκδοσης πρακτικών.

Οι παρατηρήσεις και τα σχόλια των κριτών, σύμφωνα με την πάγια πολιτική που ακολουθεί το Ε.Σ.Ι., αφορούν κυρίως στον τρόπο παρουσίασης της εργασίας και στην παρουσία ή όχι τυπογραφικών και σοβαρών επιστημονικών λαθών. Οι εργασίες πρέπει να έχουν ικανό στατιστικό περιεχόμενο, να αναδεικνύουν το πρόβλημα που μελετούν, να μην περιορίζονται μόνο σε Περιγραφική Στατιστική κτλ. Για το σκοπό αυτό υπάρχουν κριτήρια δημοσίευσης εργασιών στα Πρακτικά του Ε.Σ.Ι. τα οποία είναι αναρτημένα στην ιστοσελίδα του Ε.Σ.Ι.: www.esi-stat.gr. Όλες οι εργασίες, για τις οποίες ζητήθηκε αναθεώρηση, κρίθηκαν εκ νέου από τους κριτές ή από τους υπεύθυνους έκδοσης των πρακτικών.

Συνολικά υποβλήθηκαν τριάντα επτά (37) εργασίες, από τις οποίες δύο (2) εργασίες ανακλήθηκαν από τους συγγραφείς μετά την πρώτη αξιολόγηση, μία (1) θα δημοσιευθεί στα επόμενα Πρακτικά και μία (1) απορρίφθηκε, σύμφωνα με απόφαση του Δ.Σ. και ύστερα από σχετική πρόταση των κριτών και των υπευθύνων της έκδοσης των Πρακτικών. *Ως κριτές των εργασιών συνεργάστηκαν οι*: Δ. Αντζουλάκος, Χ. Δαμιανού, Σ. Δασκαλάκη, Γ. Δονάτος, Δ. Ιωαννίδης, Θ. Κάκουλλος, Α. Καραγρηγορίου, Δ. Καρλής, Χ. Καρώνη, Σ. Κουρούκλης, Χ. Κουκουβίνος, Σ. Κουνιάς, Μ. Κούτρας, Ε. Μακρή, Λ.

Μελιγκοτσίδου, Μ. Μπούτσικας, Χ. Μωυσιάδης, Π. Οικονόμου, Ε. Παπαγεωργίου, Δ. Παναγιωτάκος, Γ. Παπαδόπουλος, Τ. Παπαϊωάννου, Δ. Παπαναστασίου, Χ. Παυλόπουλος, Γ. Πετράκος, Α. Ρήγας, Α. Σαχλάς, Π. Σύψας, Γ. Τσακλίδης, Κ. Φερεντίνος, Μ. Χαλικιάς, Χ. Χαραλαμπίδης, Θ. Χατζηπαντελής. Η Επιτροπή Έκδοσης Πρακτικών του ΕΣΙ εκφράζει τις ευχαριστίες της προς όλους τους κριτές για την επιμελημένη και προσεκτική αξιολόγηση των εργασιών.

Η σειρά παρουσίασης των εργασιών στον παρόντα τόμο είναι αλφαβητική με βάση το επώνυμο του πρώτου συγγραφέα.

Το Διοικητικό Συμβούλιο του Ε.Σ.Ι αισθάνεται την ανάγκη να ευχαριστήσει το Χαροκόπειο Πανεπιστήμιο και την Οργανωτική Επιτροπή για την πολύ καλή οργάνωση και προσφορά τους.

Τέλος, το ΔΣ του Ε.Σ.Ι. εκφράζει τις ευχαριστίες του στη Γραμματέα του ΕΣΙ Μαρία Χιώλου για την τεχνική επιμέλεια της έκδοσης των Πρακτικών και των CD.

ΕΚ ΜΕΡΟΥΣ ΤΟΥ Δ.Σ. ΤΟΥ Ε.Σ.Ι

Οι υπεύθυνοι Έκδοσης Πρακτικών 28ου Συνεδρίου

Χαράλαμπος Δαμιανού Χρήστος Κίτσος Ιωάννης Κουτρουβέλης

Τάκης Παπαϊωάννου Χαράλαμπος Χαραλαμπίδης

ΧΟΡΗΓΟΙ ΣΥΝΕΔΡΙΟΥ

Coca-Cola

Lipton ICE TEA

Εκδόσεις ΠΡΟΠΟΜΠΟΣ
ΠΑΝΕΠΙΣΤΗΜΙΑΚΑ ΣΥΓΓΡΑΜΜΑΤΑ

ΛΟΥΜΙΔΗΣ παπαγάλος

NESCAFÉ

ΑΝΤΩΝΗΣ ΓΙΑΝΝΟΥΛΗΣ

ΕΜΠΟΡΙΑ ΔΙΑΦ/ΚΩΝ & ΑΛΛΩΝ ΕΙΔΩΝ

ΠΥΡΓΟΥ 1-3, 163 45 ΗΛΙΟΥΠΟΛΗ – ΑΘΗΝΑ

ΤΗΛ : 210 9948 118 / 6947 428826

Nestlé Fitness

ΠΡΟΓΡΑΜΜΑ ΣΥΝΕΔΡΙΟΥ

ΤΕΤΑΡΤΗ 15/4			
15:00	**Εγγραφή συνέδρων**		
16:00	**Έναρξη του Συνεδρίου - Χαιρετισμοί**		
17:00	**Διακεκριμένη Διάλεξη:** Tsitsiklis J. On Influence Learning in Networks. **(Προεδρεύουσα: Νικολαΐδη Μ.)** *(Αμφιθέατρο Γ. Καραμπατζός)*		
17:30	**Διακεκριμένη Διάλεξη:** Λέκκας Ε. Χωροχρονική κατανομή των σεισμικών φαινομένων**.** **(Προεδρεύων: Δαμιανού Χ.)** *(Αμφιθέατρο Γ. Καραμπατζός)*		
18:00	**Ελένειο Βραβείο:** Λυμπερόπουλος Δ. Martingale-ισοδύναμες κατανομές πιθανότητας με εφαρμογές στις αρχές υπολογισμού ασφαλίστρων. **(Προεδρεύων: Ηλιόπουλος Γ.)** *(Αμφιθέατρο Γ. Καραμπατζός)*		
	Ελαφρύ γεύμα *Κατά τη διάρκεια του δείπνου, η κ. Παλόγλου Σοφία-Μαρία, μεταπτυχιακή φοιτήτρια του Χαροκοπείου Πανεπιστημίου,* *θα ερμηνεύσει στο πιάνο έργα Ελλήνων συνθετών.*		
ΠΕΜΠΤΗ 16/4			
	ΑΙΘΟΥΣΑ Α	***ΑΙΘΟΥΣΑ Β***	***Εργαστήριο ΗΥ***
	ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΕΦΑΡΜΟΓΕΣ	**ΣΤΑΤΙΣΤΙΚΗ**	**Σεμινάριο ΣΕΠ (09:00 - 14:00)**
	(Προεδρεύων: Μωυσιάδης Χ.)	**(Προεδρεύων: Κουτρουβέλης Ι.)**	
9:00	Charalambides Ch.A. An euler stochastic process.	Κατηνάς Α. Σύγκριση φθινουσών κατανομών και παραμέτρων μέσω δειγματοληψίας και μεταβλητότητας.	Βασικές αρχές ελέγχου ποιότητας, Μπερσίμης Σ.
9:20	Kyriakousis A., Vamvakari M. Fitting interarrival times of pageviews on Harokopio University's web site to a *q*-exponential distribution.	Jiménez-Gamero M.D., Batsidis A., Alba-Fernández M.V. Έλεγχοι για την επιλογή μοντέλου που στηρίζονται στην εμπειρική χαρακτηριστική συνάρτηση.	
9:40	Κυρίτσης Ζ., Παπαδοπούλου Α.	Ταφιάδη Μ., Ηλιόπουλος Γ.	

	Αξιολόγηση της ποιότητας ζωής με χρήση ημιμαρκοβιανού μοντέλου ανταμοιβών.	Ακριβείς έλεγχοι για την ισότητα των μέσων τιμών δύο κατανομών Laplace.	Εισαγωγή στον Σχεδιασμό και στην Ανάλυση Πειραμάτων, Μπερσίμης Σ
10:00	Μπενιουδάκης Μ., Μπουρνέτας Α. Στρατηγικές ισορροπίας σε ουρές αναμονής με αποζημίωση και πελάτες με αποστροφή κινδύνου.	Αυλογιάρης Γ., Μιχέας Σ., Ζωγράφος Κ. Από τη φ-απόκλιση στην τοπική φ-απόκλιση.	
10:20	Burnetas A., Katehakis M. Adaptive sampling policies under incomplete information and nonstationary distributions.	Tsagris M. A novel, divergence based, regression for compositional data.	
	ΕΦΑΡΜΟΓΕΣ ΤΗΣ ΣΤΑΤΙΣΤΙΚΗΣ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ & ΣΤΗΝ ΟΙΚΟΝΟΜΙΑ	**ΣΤΑΤΙΣΤΙΚΗ & ΠΙΘΑΝΟΤΗΤΕΣ**	
	(Προεδρεύουσα: Βαμβακάρη Μ.)	**(Προεδρεύων: Χαραλαμπίδης Χ.)**	
11:30	Sardianos C., Varlamis I. Finding the optimal graph partitioning scheme for boosting collaborative filtering algorithms performance in large-scale social networks.	Παπαϊωάννου Τ. Η γραμμή του χρόνου της Στατιστικής στην Ελλάδα.	Δείκτες Ικανότητας Διεργασίας, Μπερσίμης Σ.
11:50	Μητροπούλου Π., Φιλιοπούλου Ε., Μιχαλακέλης Χ, Νικολαΐδου Μ. Κατασκευή δείκτη τιμών για υπηρεσίες υπολογιστικού νέφους.	Μωυσιάδης Π., Αντωνίου Ι. Συμπτώσεις και τυχαιότητα.	Εισαγωγή στα Διαγράμματα Ελέγχου για Μεταβλητές, Μπερσίμης Σ
12:10	Κετζάκη Ε. Μέθοδος υπολογισμού του δείκτη Gini για ομαδοποιημένα δεδομένα που βασίζεται στην αναπαράστασή του ως γινόμενο πινάκων.	Δημητρίου Ξ.Δ. Μερικά αποτελέσματα πάνω στις Πιθανότητες.	
12:30	Θεοδοσιάδου Ο., Τσακλίδης Γ., Πολυμένης Β. Προσδιορισμός φίλτρου Kalman για τη μελέτη των αποδόσεων μετοχών με βάση τα θετικά και αρνητικά άλματα των αποδόσεων	Δονάτος Γ. Αξιολόγηση εκτιμητριών περιορισμένης πληροφόρησης για κανονικά και μη κανονικά τυχαία σφάλματα στην περίπτωση μικρών δειγμάτων.	Εισαγωγή στα Διαγράμματα Ελέγχου για Ιδιότητες, Μπερσίμης Σ.
12:50	Δονάτου Α., Λεβεντίδης Ι. Οι επιπτώσεις ακραίων συνθηκών της αγοράς στο δείκτη κεφαλαιακής επάρκειας των ελληνικών τραπεζών.	Κάκουλλος Θ. Οι κατανομές t και F ισχύουν υπό σφαιρική συμμετρία.	
	ΜΟΝΤΕΛΟΠΟΙΗΣΗ, ΑΝΑΛΥΣΗ ΚΑΙ ΣΧΕΔΙΑΣΜΟΣ ΜΗΧΑΝΟΛΟΓΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ	**ΒΡΑΒΕΙΟ ΚΑΛΥΤΕΡΟΥ ΝΕΟΥ ΣΤΑΤΙΣΤΙΚΟΥ**	
	(Προεδρεύουσα: Μαλεφάκη Σ.)	**(Προεδρεύων: Μπουρνέτας Α.)**	

15:30	Δημητράκος Θ., Κυριακίδης Ε. Βέλτιστη διανομή πολλών προϊόντων με συνεχείς κατανομές ζητήσεων.	Chasiotis V., Farmakis N., Kounias S. 2^k fractional factorials and construction of the saturated optimal design with 22 observations.	
15:50	Barbu V.S., Μακρίδης Α., Καραγρηγορίου Α. Ημιμαρκοβιανά μοντέλα σε τεχνικά συστήματα πολλαπλών επιπέδων.	Μποζίκας Α., Πιτσέλης Γ. Πρόβλεψη θνησιμότητας για τον ελληνικό πληθυσμό και οι επιπτώσεις μακροζωίας στα ασφαλιστικά ταμεία.	
16:10	Μαλεφάκη Σ., Κούτρας Β., Πλατής Α. Βελτιστοποίηση πολιτικών συντήρησης τεχνολογικών συστημάτων.	Konstantinides D.G., Zaxos G.C. Accuracy of betas by using a comparative methodology.	
16:30	Ελευθέρογλου Ν., Λούτας Θ., Μαλεφάκη Σ. Μοντελοποίηση της εξέλιξης της βλάβης σε σύνθετα υλικά με τη χρήση κρυμμένων ημιμαρκοβιανών μοντέλων και δεδομένων μη καταστροφικών ελέγχων.	Tasias K.A., Nenes G. A fully adaptive control scheme for joint monitoring of location and scale of processes subject to a multiplicity of assignable causes	
18:30	**Κοινωνική Εκδήλωση[1]** **Ίδρυμα Ευγενίδου, "*Κρυμμένο Σύμπαν*"**		
20:30	***ΕΠΙΣΗΜΟ ΔΕΙΠΝΟ ΣΥΝΕΔΡΙΟΥ*** ***Πισίνα, Μαρίνα Ζέας***		

ΠΑΡΑΣΚΕΥΗ 17/4

	ΑΙΘΟΥΣΑ Α	***ΑΙΘΟΥΣΑ Β***	***Εργαστήριο ΗΥ***
	ΚΟΙΝΩΝΙΚΗ ΣΤΑΤΙΣΤΙΚΗ	**ΒΙΟΣΤΑΤΙΣΤΙΚΗ**	**Σεμινάριο ΣΕΠ (09:00 - 14:00)**
	(Προεδρεύουσα: Καλαματιανού Α.)	**(Προεδρεύων: Καρλής Δ.)**	
9:00	Δημητριάδης Ε., Βαχλιώτη Ε. Διαδικτυακός εθισμός και μαθητές λυκείων: Η περίπτωση των μαθητών της περιφερειακής ενότητας Καβάλας.	Stogiannis D., Meligkotsidou L., Siannis F. Nonparametric meta-analysis of time to event data.	Διαγράμματα Ελέγχου CUSUM & EWMA, Ρακιτζής Α.
9:20	Καμπέρης Ν., Βουδούρη Ε., Μπερσίμης Φ.Γ. Διερεύνηση χαρακτηριστικών που διέπουν τη λειτουργία της δομής "Βοήθεια στο σπίτι".	Καράκος Π., Λιαλιάρης Θ., Καράκος Α. Επιδημιολογική μελέτη της επίδρασης του βιοσυντονισμού στον ανθρώπινο οργανισμό.	

[1] 17:30 Υπάρχει η δυνατότητα για μετάβαση των συνέδρων με λεωφορεία από τον χώρο του Χαροκοπείου Πανεπιστημίου στο Ίδρυμα Ευγενίδου.

9:40	Φρονιμάκη Ε., Μαύρη Μ. Εκτίμηση της συμπεριφοράς των καταναλωτών αναφορικά με τη χρήση των έξυπνων συσκευών.	Μπερσίμης Φ.Γ., Βαμβακάρη Μ., Παναγιωτάκος Δ.Β. Η απόδοση ειδικών σταθμίσεων στις συνιστώσες ενός σύνθετου δείκτη υγείας αυξάνει τη διαγνωστική του ικανότητα.	
10:00	Ρούσσης Ι.Γ., Γεωργίου Β.Λ. Συγκριτική μελέτη της χωρικής κατανομής της εγκληματικότητας στην Ελλάδα.	Σκιαδάς Χ.Χ., Ζαφείρης Κ.Ν. Μια μέθοδος για την αποτίμηση των χρόνων υγιούς ζωής στις χώρες της Ευρώπης.	Δειγματοληψία Αποδοχής, Ρακιτζής Α.
10:20	Ματαλλιωτάκης Γ. Ανάλυση παραμέτρων της ανεργίας με στόχο τη μοντελοποίηση.	Τσανούσα Α., Αγγελής Ε., Ντούφα Σ., Παπακωνσταντίνου Ν., Σταματόπουλος Κ. Εξερευνώντας το γονιδιακό μονοπάτι σηματοδότησης στην χρόνια λεμφοκυτταρική λευχαιμία με τη χρήση μοντέλων δομικών εξισώσεων.	
11:00	**Διακεκριμένη Διάλεξη:** Tachmazidou I. Genome-wide association studies: In search of common and low frequency variants in complex traits. **(Προεδρεύων: Δεδούσης Γ.)** *(Αίθουσα Β)*		
11:30	**Διακεκριμένη Διάλεξη:** Viana M. Dihedral reductions of cyclic DNA sequences. **(Προεδρεύων: Παπαϊωάννου Τ.)** *(Αίθουσα Β)*		
12:00 - 14:00	**Παρουσιάσεις αναρτημένων ανακοινώσεων** *(Κτήριο Δ. Χαροκόπου, Ε. Χαροκόπου - Πετρούτση)* Georgousopoulou E., Panagiotakos D., Pitsavos C., Chrysohoou C., Metaxa V., Ntertimani M., Pitaraki E., Skoumas J., Tousoulis D., Stefanadis C. Inclusion of dietary evaluation in cardiovascular disease risk prediction models increases accuracy and reduces bias of the estimations. Hatjispyros S.J., Kaloudis K., Merkatas C. Bayesian reconstruction of nonlinear dynamic systems. Hatjispyros S.J., Merkatas C. Joint estimation of future values of a nonlinear noisy time series. Ioannidis K., Karagrigoriou A., Lekkas D.F. Analysis and modelling of rainfall events. Kountzakis C.E., Koutsouraki M.P. On quantum risk measures. Toulias T.L., Kitsos C.P. Aspects of the generalized lognormal distribution. Yang Y., Konstantinides D.G., Wang K. Tail behaviour of randomly weighted infinite sums. Βέρδης Α., Καλογερόπουλος Κ., Παπαδόπουλος Α., Χαλκιάς Χ. Η γεωγραφική διάσταση της επίδοσης των μαθητών στις πανελλήνιες εξετάσεις. Μια γεωστατιστική προσέγγιση.		

	Γυλού Σ., Κολυβά-Μαχαίρα Φ. Αποθορυβοποίηση σημάτων με τη μέθοδο ανάλυσης ανεξάρτητων συνιστωσών. Καλογήρου Σ., Τσίμπος Κ. Χωρικά πρότυπα της γονιμότητας του πληθυσμού της Ελλάδος 2010-2012: Ανάλυση σε επίπεδο Δήμου Καλλικράτη. Ματαλλιωτάκης Γ. Δημοσκόπηση για την ανάπτυξη. Μοσχονά Θ. Μοντέλα διαγενεακής επαγγελματικής και εκπαιδευτικής κινητικότητας, Ελλάδα 2011. Τσαμτσακίρη Π., Κολυβά-Μαχαίρα Φ. A και D_A-βέλτιστοι πειραματικοί σχεδιασμοί για τον έλεγχο διαφορών διαδοχικών μέσων τιμών. Τσαναξίδου Ζ., Μάτης Κ., Σταματέλλος Γ. Αξιολόγηση του εκτιμητή μεγέθους δείγματος με τη μέθοδο bootstrap για την κατάρτιση μαζοπίνακα.		
	ΠΡΟΒΛΕΨΗ - ΧΡΟΝΟΣΕΙΡΕΣ	**ΣΤΑΤΙΣΤΙΚΗ**	**Σεμινάριο SPSS (15:00 - 19:00)**
	(Προεδρεύων: Κίτσος Χ.)	**(Προεδρεύων: Κούτρας Μ.)**	
15:30	Κανούλας Ε., Κουγιουμτζής Δ. Εκτίμηση τάξης αυτοπαλίνδρομων μοντέλων με δεσμευμένη αμοιβαία πληροφορία.	Παπατσούμα Ι., Φαρμάκης Ν. Πολυωνυμική έκφραση συμμετρικών κατανομών: Η περίπτωση σ.π.π. τριγωνομετρικής μορφής	Εισαγωγή στη διαχείριση δεδομένων, Παναγιωτάκος Δ.
15:50	Siggiridou E., Koutlis C., Tsimpiris A., Kugiumtzis D. Assessment of Granger causality in multivariate time series.	Ηλιόπουλος Γ. Ακριβή διαστήματα εμπιστοσύνης για την παράμετρο σχήματος της κατανομής γάμμα.	Δημιουργία στατιστικής βάσης δεδομένων, Παναγιωτάκος Δ.
16:10	Siggiridou E., Kugiumtzis D. Granger causality in multivariate systems using canonical correlation analysis.	Γεράρδη Δ., Σταματέλλος Γ. Μη παραμετρική εκτίμηση τρισδιάστατων συζεύξεων για τη διάμετρο και το ύψος δένδρων.	Περιγραφική Στατιστική, Γεωργουσοπούλου Ε.
16:30	Tsimpiris A., Koutlis C. Identification of connectivity changes in multivariate time series using feature selection.	Τζαβελάς Γ., Δουλή Μ., Οικονόμου Π. Τα αποτελέσματα της επιλογής λάθους μοντέλου στην εκτίμηση παραμέτρων από μεροληπτικά δείγματα.	
16:50	Koutlis C., Kimiskidis V.K. Covert states of excitability in the human epileptic brain revealed by multivariate time series analysis.	Οικονόμου Π., Ψαρράκος Γ. Μεροληπτική δειγματοληψία ως προς τη μέση υπολειπόμενη διάρκεια ζωής.	
	ΣΤΑΤΙΣΤΙΚΕΣ ΜΕΘΟΔΟΙ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΓΕΩΓΡΑΦΙΚΗ ΑΝΑΛΥΣΗ	**ΠΕΙΡΑΜΑΤΙΚΟΙ ΣΧΕΔΙΑΣΜΟΙ**	**Σεμινάριο SPSS (15:00 - 19:00)**
	(Προεδρεύων: Χαλκιάς Χ.)	**(Προεδρεύων: Φαρμάκης Ν.)**	
17:30	Τσαρπαλής Κ., Βάρλας Γ., Κατσαφάδος Π. Στατιστικές προσεγγίσεις και αξιολόγηση της	Ευαγγελάρας Χ. Κατασκευή κλασματικών παραγοντικών σχεδιασμών δύο επιπέδων με	

	εποχιακής πρόγνωσης καιρού.	ελάχιστη γενικευμένη απόκλιση.	Περιγραφική Στατιστική, Γεωργουσοπούλου Ε.
17:50	Σκριμιζέας Π. Εφαρμογή και αξιολόγηση μεθόδων χωρικής ανάλυσης βροχομετρικών δεδομένων.	Chasiotis V., Kounias S., Farmakis N. Optimal designs for estimating the contrasts of factor level effects in 2^k fractional factorials.	
18:10	Μαλαματάρης Δ., Τζελατίδης Ι. Υπολογισμός ισοδύναμου βροχομετρικού ύψους νομού Θεσσαλονίκης με τη χρήση γεωγραφικών συστημάτων πληροφοριών.	Παρπούλα Χ. Ανάλυση υπερκορεσμένων σχεδιασμών μέσω των ποινικοποιημένων μηχανών διανυσματικής υποστήριξης.	
	ΣΤΑΤΙΣΤΙΚΕΣ ΜΕΘΟΔΟΙ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΓΕΩΓΡΑΦΙΚΗ ΑΝΑΛΥΣΗ	**ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**	**Σεμινάριο SPSS (15:00 - 19:00)**
	(Προεδρεύων: Τσίμπος Κ.)	**(Προεδρεύων: Καραγρηγορίου Α.)**	
18:50	Χατζηχρήστος Θ., Παπακωνσταντίνου Δ. Ανάπτυξη γεωδημογραφικού συστήματος για το νομό Αττικής.	Καλαμάρας Δ.Α., Καλαματιανού Α.Γ. Κίνητρα, προσδοκίες, αρχικές δεσμεύσεις και διάρκεια πανεπιστημιακών σπουδών: Μια προσέγγιση στο πλαίσιο ενός μοντέλου αναλογικού κινδύνου με λανθάνουσες συμμεταβλητές.	Γραφήματα, Γεωργουσοπούλου Ε.
19:10	Θεοδωρίδου Π., Καρατζάς Γ., Βαρουχάκης Ε., Παπαδοπούλου Μ. Διερεύνηση μέτρων απόστασης στη χωρική ανάλυση περιβαλλοντικών δεδομένων.	Καραγάνης Α., Στάμου Μ. Υποδείγματα μαζικής επανεκτίμησης στεγαστικών ακινήτων.	
19:30	Αρτελάρης Π., Χαλκιάς Χ. Παχυσαρκία και οικονομική ανάπτυξη: Μία σχέση ανάστροφου U;	Πέτσα Α., Παπαδόπουλος Γ., Καλύβας Δ. Χωρική μεταβλητότητα της σχέσης πραγματικής και αντικειμενικής αξίας ακινήτων σε οκτώ δήμους του λεκανοπεδίου Αττικής.	

ΣΑΒΒΑΤΟ 18/4			
	ΑΙΘΟΥΣΑ Α	***ΑΙΘΟΥΣΑ Β***	***Εργαστήριο ΗΥ***
	ΣΤΑΤΙΣΤΙΚΗ ΣΤΗ ΣΕΙΣΜΟΛΟΓΙΑ	**ΕΦΑΡΜΟΣΜΕΝΕΣ ΠΙΘΑΝΟΤΗΤΕΣ**	**Σεμινάριο SPSS (9:00 - 13:00)**
	(Προεδρεύων: Τσακλίδης Γ.)	**(Προεδρεύων: Κωνσταντινίδης Δ.)**	

9:00	Ζήμερας Σ. Χωρική σεισμολογία.	Μπούτσικας Μ., Ρακιτζής Α., Αντζουλάκος Δ. Το πλήθος των αποζημιώσεων έως την χρεοκοπία σε ένα ανανεωτικό μοντέλο κινδύνου με άνω επίπεδο ασφαλείας.	Σύγκριση μέσων τιμών, Μπερσίμης Φ.
9:20	Βουγιούκα Γ., Καραγρηγορίου Α., Μακρίδης Α., Τσάπανος Θ. Τεχνικά συστήματα πολλαπλών επιπέδων με εφαρμογές στη σεισμολογία.	Κούτρας Β. Μ., Κούτρας Μ.Β. Σύνθετες συναρτήσεις σάρωσης και εφαρμογές στα χρηματοοικονομικά.	
9:40	Χορόζογλου, Δ. και Παπαδημητρίου, Ε. Μελέτη της δομής και έλεγχος τυχαιότητας σεισμικών δικτύων συσχέτισης από πολυμεταβλητές χρονοσειρές.	Κούτρας Μ., Λυμπερόπουλος Δ. Ασυμπτωτικά αποτελέσματα για την πολλαπλή συνάρτηση σάρωσης.	
10:00	Μαγγίρα Ο., Τσακλίδης Γ., Παπαδημητρίου Ε., Βότση Ε. Εκτίμηση της σεισμικής επικινδυνότητας του Κορινθιακού κόλπου με τη χρήση στοχαστικού μοντέλου απελευθέρωσης τάσης.	Αράπης Α.Ν., Μακρή Φ.Σ., Ψυλλάκης Ζ.Μ. Μήκος και θέση της μικρότερης ακολουθίας που περιέχει ροές επιτυχιών.	
	ΣΤΑΤΙΣΤΙΚΗ ΣΤΗ ΣΕΙΣΜΟΛΟΓΙΑ	**ΣΤΑΤΙΣΤΙΚΟΣ ΕΛΕΓΧΟΣ ΔΙΕΡΓΑΣΙΩΝ**	
	(Προεδρεύων: Παπαδημητρίου Ε.)	**(Προεδρεύων: Ρακιτζής Α.)**	
11:00	Μεσημέρη, Μ., Καρακώστας, Β. Χωροχρονικές ιδιότητες της σεισμικότητας στο Δυτικό Κορινθιακό κόλπο.	Sgora A, Psarakis S, Bersimis S. Methods for interpreting the out-of-control signal of multivariate control chart: A comparison study.	Συσχέτιση, Μπερσίμης Φ.
11:20	Gkarlaouni C., Lasocki S., Papadimitriou E. Investigation of earthquake magnitude and interevent time distribution in Corinth Gulf and Mygdonia basin with the use of stochastic tools.	Κούτρας Μ.Β., Σοφικίτου Ε.Μ. Ένα διδιάστατο ημιπαραμετρικό διάγραμμα ελέγχου για το ζευγάρι μιας διατεταγμένης παρατήρησης και της συμμεταβλητής της..	Παλινδρόμηση, Μπερσίμης Φ. .
11:40	Karlis D., Pedeli X. Modelling multivariate count time series of earthquake data: Selection of covariance structure.	Sachlas A, Bersimis S. Simultaneous monitoring of bivariate random variables defined on contingency tables.	
	ΠΙΘΑΝΟΤΗΤΕΣ - ΑΝΑΛΟΓΙΣΜΟΣ	**ΣΤΑΤΙΣΤΙΚΟΣ ΕΛΕΓΧΟΣ ΔΙΕΡΓΑΣΙΩΝ**	
	(Προεδρεύων: Κάκουλλος Θ.)	**(Προεδρεύων: Μπερσίμης Σ.)**	
12:20	Kountzakis C.E. On the strong sensitivity of risk measures on L^1-spaces.	Rakitzis A, Castagliola P, Maravelakis P. A new memory-type control chart for count	

		data.	Παλινδρόμηση, Μπερσίμης Φ
12:40	Λυμπερόπουλος Δ.Π., Μαχαιράς Ν.Δ., Τζανίνης Σ.Μ. Ισοδυναμία ορισμών μικτών στοχαστικών διαδικασιών Poisson.	Bourazas K., Kiagias D., Tsiamirtzis P. Bayesian statistical process control: Predictive control charts for continuous distributions in the regular exponential family.	
13:00	Hadjikyriakou M. A comparison theorem for functions of vectors of associated and negatively associated random variables.	Demertzi E., Psarakis S. Control charts for the logarithmic distribution.	
13:20	Konstantinides D.G., Kountzakis C.E. Regular variation in Orlicz spaces.		
14:00	**ΟΛΟΜΕΛΕΙΑ:** ΛΗΞΗ ΣΥΝΕΔΡΙΟΥ *(Αμφιθέατρο Γ. Καραμπατζός)*		

Κοινωνικές Εκδηλώσεις

Τετάρτη 15 Απριλίου 2015

Welcome Cocktail: Αμφιθέατρο Γ. Καραμπατζός
Ώρα: 20:00

Πέμπτη 16 Απριλίου 2015

<u>Ίδρυμα Ευγενίδου «Κρυμμένο Σύμπαν»</u>
Ώρα : 18:30
<u>Επίσημο Δείπνο Συνεδρίου</u>
Μαρίνα Ζέας, *Πισίνα*
Ώρα : 20:30

Επιστημονική Επιτροπή

Αναγνωστόπουλος Δ.
Αντζουλάκος Δ.
Δαμιανού Χ.
Ζωγράφος Κ.
Ηλιόπουλος Γ.
Κάκουλλος Θ.
Καλαματιανού Α.
Καραγρηγορίου Α.
Καρλής Δ.
Καρώνη Χ.
Κίτσος Χ.
Κουνιάς Σ.
Κούτρας Μ.
Κουτρουβέλης Ι.
Μαχαιράς Ν.
Μπάγκος Π.
Μωυσιάδης Π.
Νάκας Χ.
Παπαδόπουλος Γ.
Παπαϊωάννου Τ.
Παπαναστασίου Δ.
Σκιαδάς Χ.
Σύψας Π.
Τσακλίδης Γ.
Τσίμπος Κ.
Χαραλαμπίδης Χ.

Οργανωτική Επιτροπή

Ε Βαμβακάρη Μ.
Δεδούσης Γ.
Ηλιόπουλος Γ.
Καλογήρου Σ.
Κυριακούσης Α.
Μπερσίμης Σ.
Μπουρνέτας Α.
Παναγιωτάκος Δ.
Σιάννης Φ.
Χαλκιάς Χ.

Διοικητικό Συμβούλιο Ε.Σ.Ι.

Πρόεδρος: Δαμιανού Χαράλαμπος
Αντιπρόεδρος: Κουτρουβέλης Ιωάννης
Γενικός Γραμματέας: Κίτσος Χρήστος
Ειδικός Γραμματέας: Παναγιωτάκος Δημοσθένης
Ταμίας: Μπερσίμης Σωτήρης
Μέλη: Παπαϊωάννου Τάκης, *έφορος βιβλιοθήκης*
Χαραλαμπίδης Χαράλαμπος

εργασίες

στα ελληνικά

Greek Statistical Institute
Proceedings of the 28th Panhellenic Statistics Conference (2015), pp. 22−31

ΜΗΚΟΣ ΚΑΙ ΘΕΣΗ ΤΗΣ ΜΙΚΡΟΤΕΡΗΣ ΑΚΟΛΟΥΘΙΑΣ ΠΟΥ ΠΕΡΙΕΧΕΙ ΡΟΕΣ ΕΠΙΤΥΧΙΩΝ

Α.Ν. Αράπης[1], **Φ.Σ. Μακρή**[1], **Ζ.Μ. Ψυλλάκης**[2]
[1] Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών
tasosarp@hotmail.com, makri@math.upatras.gr
[2] Τμήμα Φυσικής, Πανεπιστήμιο Πατρών
psillaki@physics.upatras.gr

ΠΕΡΙΛΗΨΗ

Θεωρούμε μια ακολουθία πεπερασμένου μήκους δυαδικών (αποτυχία - επιτυχία) πειραμάτων διατεταγμένων σε μια γραμμή. Ορίζονται οι τυχαίες μεταβλητές που παριστάνουν το μήκος και τη θέση της μικρότερης υποακολουθίας η οποία περιέχει όλες τις ροές επιτυχιών μήκους μεγαλύτερου ή ίσου από ένα σταθερό αριθμό. Υπό τη συνθήκη ότι υπάρχουν δύο τουλάχιστον τέτοιες ροές στην αρχική ακολουθία των πειραμάτων, μελετάμε τις κατανομές πιθανότητας των τυχαίων μεταβλητών που προαναφέραμε. Η μελέτη παρουσιάζεται για ακολουθίες ανεξάρτητων και ισόνομων πειραμάτων. Αριθμητικά παραδείγματα διευκρινίζουν περαιτέρω τα θεωρητικά αποτελέσματα.

Λέξεις Κλειδιά: Πειράματα Bernoulli, Ροές, DNA.

1. ΕΙΣΑΓΩΓΗ ΚΑΙ ΠΡΟΚΑΤΑΡΚΤΙΚΕΣ ΕΝΝΟΙΕΣ

Διαδοχικά παρατηρούμενα δυαδικά αποτελέσματα δημιουργούν στοχαστικά πρότυπα τα οποία εμφανίζονται πολύ συχνά στην πράξη. Ροές και γενικότερα σύνθετοι σχηματισμοί οι οποίοι απαριθμούνται από κατάλληλες στατιστικές συναρτήσεις, χρησιμοποιούνται για την ανάλυση αποτελεσμάτων τα οποία εμφανίζονται σε πολλά επιστημονικά πεδία όπως είναι ο έλεγχος ποιότητας, η αξιοπιστία μηχανικών συστημάτων, η συμπίεση και μετάδοση πληροφορίας και η μοριακή βιολογία.

Η χρήση των ροών - σχηματισμών για τη μελέτη ακολουθιών δυαδικών τυχαίων μεταβλητών (τ.μ.) απαιτεί τον προσδιορισμό των κατανομών των εγγενών με τους σχηματισμούς στατιστικών συναρτήσεων υπό διαφορετικές κάθε φορά υποθέσεις για τη δομή των ακολουθιών. Τέτοιες υποθέσεις, μεταξύ άλλων, είναι η υπόθεση της ανεξαρτησίας, η υπόθεση της ανταλλαξιμότητας καθώς και η υπόθεση της Μαρκοβιανής εξάρτησης μεταξύ των όρων της ακολουθίας. Όλες οι περιπτώσεις αυτές

δίνουν ως ειδική περίπτωση την ακολουθία ανεξαρτήτων και ισόνομων τ.μ. την οποία θα μελετήσουμε στην παρούσα εργασία.

Θεωρούμε μια ακολουθία $\{X_i, i \geq 1\}$ από δυαδικές [Επιτυχία (S ή 1) - Αποτυχία (F ή 0)] τ.μ. διατεταγμένες σε μια γραμμή. Ως ροή επιτυχιών ή ροή από 1 ορίζεται μία (μερική) ακολουθία (ή υποακολουθία) της $\{X_i, i \geq 1\}$ η οποία αποτελείται από συνεχόμενες μονάδες (1) των οποίων προηγούνται ή έπονται μηδέν (0) ή τίποτα. Ο αριθμός των μονάδων σε μια ροή αναφέρεται ως μήκος ή μέγεθος της ροής.

Θεωρώντας τα πρώτα n, $n \geq 1$, δυαδικά πειράματα της ακολουθίας $\{X_i, i \geq 1\}$ και έναν θετικό ακέραιο k, $1 \leq k \leq n$, μπορούμε να ορίσουμε στατιστικές συναρτήσεις (τ.μ.) οι οποίες παρέχουν χρήσιμες πληροφορίες για τον αριθμό των ροών από 1, για τα μήκη των ροών από 1 καθώς και για το χρόνο αναμονής μέχρι τον σχηματισμό (την εμφάνιση) ροών από 1. Οι τ.μ. αυτές συνδέονται μεταξύ τους λόγω του ότι αναφέρονται σε ροές από 1 με μήκος μεγαλύτερο ή ίσο από το k και οι οποίες θα ονομάζονται στο εξής k-ροές από 1 ή απλώς k-ροές. Ορίζουμε τις τ.μ.:
(α) $G_{n,k}$, που παριστάνει τον αριθμό των k-ροών.
(β) $W_{r,k}$, που παριστάνει τον χρόνο αναμονής μέχρι να εμφανισθούν r, $r \geq 1$, k-ροές. Ισχύει ότι, $W_{r,k} = \min\{n \geq r(k+1) - 1 : G_{n,k} = r\}$.
(γ) L_n, που παριστάνει το μήκος της μεγαλύτερης ροής από 1. Ισχύει ότι, $L_n = \max\{k \leq n : G_{n,k} > 0\}$ εάν $\{k \leq n : G_{n,k} > 0\} \neq \emptyset$, και 0, διαφορετικά.
(δ) $D_{n,k}$, που παριστάνει την απόσταση (τον αριθμό πειραμάτων) μεταξύ της πρώτης και της τελευταίας k-ροής. Οι μονάδες (1) της πρώτης και της τελευταίας k-ροής συμπεριλαμβάνονται (απαριθμούνται) στην $D_{n,k}$. Ισχύει ότι, $D_{n,k} \leq n$ εάν $G_{n,k} \geq 1$ (δηλ. $W_{1,k} \leq n$). Επίσης, $D_{n,k} = L_n$ εάν $G_{n,k} = 1$ και $D_{n,k} > L_n$ εάν $G_{n,k} > 1$.
(ε) $\mathbf{U}_{n,k} = (U_{n,k}^{(1)}, U_{n,k}^{(2)})$ με τις (συνιστώσες) τ.μ. $U_{n,k}^{(1)}$ και $U_{n,k}^{(2)}$ να ορίζουν τις θέσεις της αρχής και του τέλους, αντίστοιχα, της πρώτης και της τελευταίας k-ροής, έτσι ώστε $D_{n,k} = U_{n,k}^{(2)} - U_{n,k}^{(1)} + 1$.

Όταν $G_{n,k} > 0$, οι τ.μ. $D_{n,k}$ και $\mathbf{U}_{n,k}$ καθορίζουν το μήκος και τη θέση αντίστοιχα της μικρότερης υποακολουθίας (τμήματος) της $\{X_i\}_{i=1}^n$ η οποία περιέχει όλες τις, $G_{n,k}$ το πλήθος, k-ροές από επιτυχίες. Ειδικά οι $D_{n,1}$ και $\mathbf{U}_{n,1}$ δίνουν το μήκος και τη θέση, αντίστοιχα, της ελαχίστου μήκους υποακολουθίας στην οποία περιέχονται όλες οι επιτυχίες της ακολουθίας των n πειραμάτων.

Το παράδειγμα που ακολουθεί διευκρινίζει περαιτέρω τους ορισμούς (α)-(ε).

Παράδειγμα. Θεωρούμε μια ακολουθία $\{Y_i\}_{i=1}^{2n}$, $n \geq 1$, με τιμές από ένα πεπερασμένο αλφάβητο $\mathcal{A}$. Ως εφαρμογή θεωρούμε μια ακολουθία DNA με $\mathcal{A} = \{A, C, G, T\}$ και τα γειτονικά της τμήματα $\{Y_1, Y_2, \ldots Y_n\}$, $\{Y_{n+1}, Y_{n+2}, \ldots Y_{2n}\}$. Ορίζουμε τη δυαδική ακολουθία $\{X_i\}_{i=1}^n$ ως εξής: $X_i = 1$, αν $Y_{i+n} = Y_i$ και 0, διαφορετικά. Έστω ότι έχουμε τα γειτονικά πρότυπα μήκους $n = 12$ το καθένα, Y_i : $CAAGTGTGGGTC$ και Y_{i+12} : $GAAGTGAGGAGC$, $i = 1, 2, \ldots, 12$. Τότε, X_i : 011111011001 και $L_{12} = 5$, $G_{12,1} = 3$, $U_{12,1}^{(1)} = 2$, $U_{12,1}^{(2)} = 12$, $D_{12,1} = 11$, $G_{12,2} = 2$, $U_{12,2}^{(1)} = 2$, $U_{12,2}^{(2)} = 9$, $D_{12,2} = 8$, και για $k = 3, 4, 5$, $G_{12,k} = 1$, $U_{12,k}^{(1)} = 2$, $U_{12,k}^{(2)} = 6$, $D_{12,k} = L_{12}$. □

Οι τ.μ. $G_{n,k}$, $W_{r,k}$ και L_n έχουν μελετηθεί από πολλούς ερευνητές. Ενδεικτικά αναφέρουμε τα βιβλία (Balakrishnan and Koutras (2002), Fu and Lou (2003)) και τις εργασίες (Fu and Koutras (1994), Koutras (2003), Lou (2003), Sinha and Sinha (2009), Demir and Eryilmaz (2010), Makri and Psillakis (2011)). Η τ.μ. μήκους $D_{n,k}$ έχει εισαχθεί από τον Benson (1999) και χρησιμοποιήθηκε για την ανίχνευση επαναλαμβανόμενων αλληλουχιών νουκλετιδίων που εμφανίζονται σε μια ακολουθία DNA. Η τ.μ. θέσης $\mathbf{U}_{n,k}$ έχει εισαχθεί από τους Makri et al. (2015).

Σκοπός της εργασίας είναι η μελέτη των τ.μ. $D_{n,k}$ και $\mathbf{U}_{n,k}$ δοθείσης της πραγματοποίησης του ενδεχομένου

$$\mathcal{M}_{n,k} = \{G_{n,k} \geq 2\}, \tag{1}$$

δηλ. υπό τη συνθήκη ότι υπάρχουν δύο τουλάχιστον k-ροές στη δυαδική ακολουθία $\{X_i\}_{i=1}^n$. Η μελέτη παρουσιάζεται για ακολουθίες ανεξάρτητων και ισόνομων τ.μ. $X_1, X_2, \ldots, X_n$, $n \geq 3$, με σταθερή πιθανότητα επιτυχίας

$$P(X_i = 1) = p = 1 - P(X_i = 0) = 1 - q, \quad i = 1, 2, \ldots, n. \tag{2}$$

Συγκεκριμένα στην ενότητα 2 προσδιορίζουμε τη δεσμευμένη συνάρτηση πιθανότητας των τ.μ. $D_{n,k}$ και $\mathbf{U}_{n,k}$ δοθέντος του $\mathcal{M}_{n,k}$. Στην ενότητα 3 παρουσιάζονται αριθμητικά παραδείγματα τα οποία αφ΄ ενός μεν διευκρινίζουν περαιτέρω τα θεωρητικά αποτελέσματα της εργασίας και αφ΄ ετέρου προτείνουν κάποια κριτήρια εφαρμογής - χρήσης των τ.μ. που μελετήθηκαν.

2. ΚΥΡΙΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Αρχικά παρουσιάζουμε δύο αριθμούς οι οποίοι δίνουν το πλήθος των διαφορετικών τρόπων τοποθέτησης όμοιων σφαιρών σε διακεκριμένες κάλπες περιορισμένης χωρητικότητας. Οι αριθμοί αυτοί θα χρησιμοποιηθούν στην εύρεση της συνάρτησης πιθανότητας των τ.μ. $G_{n,k}$, $1 \leq k \leq n$ και $D_{n,k}$, $n \geq 2k+1$, $k \geq 1$.

Λήμμα 1. Ο αριθμός των τρόπων τοποθέτησης α όμοιων σφαιρών σε r διακεκριμένες κάλπες, εκ των οποίων m, $0 \leq m \leq r$, συγκεκριμένες έχουν χωρητικότητα k, δίνεται από τον τύπο

$$H_m(\alpha, r, k) = \sum_{j=0}^{\lfloor \frac{a}{k+1} \rfloor} (-1)^j \binom{m}{j} \binom{a-(k+1)j+r-1}{a-(k+1)j}. \tag{3}$$

Ο αριθμός $H_m(\alpha, r, k)$ είναι ο αριθμός των ακεραίων μη αρνητικών λύσεων της εξίσωσης $x_1 + x_2 + \ldots + x_r = \alpha$ με τον περιορισμό $x_i \leq k$, $i = 1, 2, \ldots, m$.

Πόρισμα 1. Για $m = r$, ο $H_r(\alpha, r, k)$ είναι ο αριθμός τοποθέτησης α όμοιων σφαιρών σε r διακεκριμένες κάλπες, κάθε μια με χωρητικότητα k και δίνεται από τον τύπο

$$C(\alpha, r, k) \equiv H_r(\alpha, r, k) = \sum_{j=0}^{\lfloor \frac{a}{k+1} \rfloor} (-1)^j \binom{r}{j} \binom{a-(k+1)j+r-1}{a-(k+1)j} \tag{4}$$

(βλέπε, Riordan (1964), Charalambides (2002) ή Makri et al. (2007)). □

Στη συνέχεια δίνουμε την πιθανότητα πραγματοποίησης του ενδεχομένου $\mathcal{M}_{n,k}$

$$\alpha_{n,k} = P(G_{n,k} \geq 2) = P(W_{2,k} \leq n). \tag{5}$$

Λήμμα 2. Η πιθανότητα $\alpha_{n,k} = \alpha_{n,k}(p)$ δίνεται από τις σχέσεις
(I) Για $k = 1$,

$$\alpha_{n,k} = 1 - q^n - \frac{p[p^{n+1} - (n+1)pq^n + nq^{n+1}]}{(p-q)^2},\ p \neq q; 1 - \frac{2 + n(n+1)}{2^{n+1}},\ p = q.$$

(II) Για $k \geq 1$, $2k + 1 \leq n \leq 3k + 1$,

$$\alpha_{n,k} = \frac{1}{2}(n - 2k)qp^{2k}[2 + q(n - 2k - 1)].$$

(III) Γενικά για $n \geq k \geq 1$,

$$\alpha_{n,k} = 1 - \sum_{i=0}^{1} \sum_{y=0}^{n-ik} p^{n-y} q^y \binom{y+1}{i} H_{y+1-i}(n - y - ik, y + 1, k - 1). \tag{6}$$

Απόδειξη Για $k = 1$, παρατηρούμε ότι

$$\begin{aligned}
\alpha_{n,k} &= 1 - P(G_{n,k} = 0) - P(G_{n,k} = 1) \\
&= 1 - P(\prod_{m=1}^{n} (1 - X_m) = 1) - P\Big[\cup_{i=1}^{n} \cup_{j=1}^{n-i+1} \Big[(\prod_{m=1}^{j-1} (1 - X_m) = 1) \\
&\quad \cap (\prod_{m=j}^{j+i-1} X_m = 1) \cap (\prod_{m=j+i}^{n} (1 - X_m) = 1) \Big] \Big]
\end{aligned}$$

και για $k \geq 1$, $2k + 1 \leq n \leq 3k + 1$, ότι

$$\begin{aligned}
\alpha_{n,k} &= P(G_{n,k} = 2) \\
&= P\Big[\cup_{j=0}^{n-2k-1} \cup_{i=j+k+1}^{n-k} [(1 - X_j)(\prod_{m=j+1}^{j+k} X_m)(1 - X_i)(\prod_{m=i+1}^{i+k} X_m) = 1] \Big].
\end{aligned}$$

Τα αποτελέσματα προκύπτουν μετά από αλγεβρικές πράξεις και με την παραδοχή ότι $P(X_0 = 0) = 1$. Τέλος, το αποτέλεσμα (III) προκύπτει χρησιμοποιώντας την έκφραση για τη συνάρτηση πιθανοτητας της $G_{n,k}$, που δίνεται στο Θεώρημα 3.3 της εργασίας Makri et al. (2007). □

Για $n \geq 2k + 1$, $k \geq 1$ έστω

$$g_{n,k}(d) = P(D_{n,k} = d, \mathcal{M}_{n,k}), \quad d = 2k + 1, 2k + 2, \ldots, n. \tag{7}$$

Τότε η $g_{n,k}(d)$ μπορεί να προσδιορισθεί μέσω της Πρότασης που ακολουθεί.
Πρόταση 1. Η πιθανότητα $g_{n,k}(d) = g_{n,k}(d;p)$ δίνεται από τις σχέσεις
(I) Για $k = 1$, $d = 3, 4, \ldots, n$,

$$g_{n,k}(d) = (n-d+1)q^{n-d}(p^2 - p^d).$$

(II) Για $k \geq 1$, $2k+1 \leq n \leq 3k+1$,

$$g_{n,k}(d) = q(p^{2k} - p^d)[(n-d-1)q+2],\ d = 2k+1, 2k+2, \ldots, n-1; p^{2k} - p^d,\ d = n.$$

(III) Για $k \geq 1$, $n \geq 2k+1$, $d = 2k+1, 2k+2, \ldots, n$,

$$g_{n,k}(d) = \sum_{y_1=1}^{d-2k} \binom{d-2k}{y_1} \sum_{y=y_1}^{y_1+n-d} p^{n-y} q^y (y-y_1+1) C(n-d-y+y_1, y-y_1, k-1). \quad (8)$$

Απόδειξη Για $k = 1$, $d = 3, 4, \ldots, n$, ισχύει ότι

$$\begin{aligned} g_{n,k}(d) &= P\Big[\cup_{j=1}^{n-d+1} \Big[[(X_j = 1) \cap (X_{j+d-1} = 1) \cap (\prod_{m=j}^{j+d-1} X_m = 1)^c] \\ &\quad \cap [(\prod_{m=1}^{j-1} (1 - X_m) = 1) \cap (\prod_{m=j+d}^{n} (1 - X_m) = 1)] \Big] \Big]. \end{aligned}$$

Για $k \geq 1$, $2k+1 \leq n \leq 3k+1$ και για $d = 2k+1, \ldots, n-1$, έχουμε ότι

$$\begin{aligned} g_{n,k}(d) &= P\Big[\cup_{j=0}^{n-d} \Big[(X_j = 0) \cap (X_{j+d+1} = 0) \cap [(\prod_{m=j+1}^{j+k} X_m = 1) \\ &\quad \cap (\prod_{m=j+d-k+1}^{j+d} X_m = 1) \cap (\prod_{m=j+1}^{j+d} X_m = 1)^c] \Big] \Big], \end{aligned}$$

ενώ για $d = n$, ισχύει ότι

$$g_{n,k}(d) = P\big[(\prod_{m=1}^{k} X_m = 1) \cap (\prod_{m=n-k+1}^{n} X_m = 1) \cap (\prod_{m=1}^{n} X_m = 1)^c \big].$$

Θέτοντας $P(X_0 = 0) = P(X_{n+1} = 0) = 1$, προκύπτουν τα αντίστοιχα αποτελέσματα. Για την απόδειξη της γενικής περίπτωσης (III) έστω Y_n ο συνολικός αριθμός αποτυχιών στην ακολουθία των n πειραμάτων, Y_d εκ των οποίων βρίσκονται στο τμήμα της ακολουθίας των d πειραμάτων στην οποία περιέχονται όλες οι k-ροές. Για την εύρεση της $P(D_n = d, Y_n = y, Y_d = y_1, \mathcal{M}_{n,k})$ θεωρούμε ότι οι y αποτυχίες δημιουργούν κάλπες στις οποίες τοποθετούνται οι $n-y$ επιτυχίες υπό συγκεκριμένους

περιορισμούς. Οι $d-y_1$ επιτυχίες της d-υποακολουθίας τοποθετούνται στις y_1+1 κάλπες με $\binom{(d-y_1-2k)+(y_1+1)-1}{(y_1+1)-1}$ διαφορετικούς τρόπους. Οι υπόλοιπες $y-y_1$ αποτυχίες δημιουργούν $y-y_1$ κάλπες χωρητικότητας $k-1$ στις οποίες οι $n-(y-y_1)-d$ υπόλοιπες επιτυχίες μπορούν να τοποθετηθούν με $C(n-d-y+y_1, y-y_1, k-1)$ τρόπους. Παρατηρώντας ότι οι $y-y_1$ αποτυχίες μπορούν να εμφανισθούν στην ακολουθία με $y-y_1+1$ διαφορετικούς τρόπους, έχουμε ότι $P(D_{n,k}=d, Y_n=y, Y_d=y_1, \mathcal{M}_{n,k}) = (y-y_1+1)p^{n-y}q^y\binom{d-2k}{y_1}C(n-d-y+y_1, y-y_1, k-1)$. Αθροίζοντας ως προς y_1 και y προκύπτει το αποτέλεσμα. □

Ένας εναλλακτικός τρόπος υπολογισμού της $g_{n,k}(d)$ δίνεται από την ακόλουθη Πρόταση.

Πρόταση 2. Για $d=2k+1, 2k+2, \ldots, n-3$ ισχύει ότι

$$g_{n,k}(d)=g_{n-1,k}(d)-qp^k g_{n-k-1,k}(d)+qp^{2k}(1-p^{d-2k})[h_{n-d-1}(k-1)-ph_{n-d-2}(k-1)], \tag{9}$$

όπου $h_r(k-1)=P(L_r\le k-1)$, $g_{n,k}(n-2)=3q^2p^2(1-p^{n-4})$ για $k=1$, $g_{n,k}(n-2)=qp^{2k}(2+q)(1-p^{n-2-2k})$ για $k\ge 2$, $g_{n,k}(n-1)=2qp^{2k}(1-p^{n-2k-1})$, $g_{n,k}(n)=p^{2k}(1-p^{n-2k})$ και $g_{n,k}(d)=0$, για $d<2k+1$ ή $d>n$.

Απόδειξη Οι εκφράσεις της $g_{n,k}(d)$ για $d=n-2, n-1, n$ προκύπτουν άμεσα. Για $d=2k+1, 2k+2, \ldots, n-3$ παρατηρούμε ότι

$$g_{n,k}(d)=g_{n-1,k}(d)-P(D_{n,k}\ne d, D_{n-1,k}=d, \mathcal{M}_{n,k})+P(D_{n,k}=d, D_{n-1,k}\ne d, \mathcal{M}_{n,k}).$$

Το αποτέλεσμα προκύπτει αν στη συνέχεια παρατηρήσουμε ότι $D_n\ge D_{n-1}$,

$$P(D_{n,k}=d, D_{n-1,k}=d-1, \mathcal{M}_{n,k}) = q(p^{2k+1}-p^d)h_{n-d-1}(k-1),$$

$$P(D_{n,k}=d, D_{n-1,k}<d-1, \mathcal{M}_{n,k}) = q^2p^{2k}h_{n-d-1}(k-1),$$

και

$$P(D_{n,k}>d+1, D_{n-1,k}=d, \mathcal{M}_{n,k}) = qp^k g_{n-k-1}(d). \quad \square$$

Παρατήρηση 1. Για $1\le k<r$ ένας τρόπος υπολογισμού της $h_r(k)$ είναι μέσω της αναδρομικής σχέσης (βλέπε, Fu and Lou (2002), Corollary 3.1)

$$h_r(k)=P(L_r\le k)=qh_{r-1}(k)+q\sum_{i=1}^{k}p^i h_{r-i-1}(k), \tag{10}$$

με $h_r(0)=q^r$ και $h_r(k)=1$, $0\le r\le k$. □

Για $n\ge 2k+1$, $k\ge 1$, έστω

$$f_{n,k}(d)=P(D_{n,k}=d\mid \mathcal{M}_{n,k}), \quad d=2k+1, 2k+2, \ldots, n. \tag{11}$$

Τότε ισχύει το ακόλουθο Θεώρημα.

Θεώρημα 1. Η συνάρτηση πιθανότητας $f_{n,k}(d)=f_{n,k}(d;p)$ δίνεται από τον τύπο

$$f_{n,k}(d)=g_{n,k}(d)/\alpha_{n,k}, \tag{12}$$

όπου η πιθανότητα $\alpha_{n,k}$ προσδιορίζεται από το Λήμμα 2 και η $g_{n,k}(d)$ από την Πρόταση 1 ή την Πρόταση 2. $\square$

Για $n \geq 2k+1$, $k \geq 1$ ορίζουμε

$$u_{n,k}(i,j) = P(\mathbf{U}_{n,k} = (i,j) \mid \mathcal{M}_{n,k}),\ i = 1,2,\ldots,n-2k,\ j = i+2k,\ldots,n. \quad (13)$$

Τότε ακολουθώντας τη μέθοδο κατάλληλης διανομής επιτυχιών σε κάλπες που δημιουργούνται στα διαστήματα μεταξύ των αποτυχιών έχουμε το ακόλουθο Θεώρημα για τον προσδιορισμό της.

Θεώρημα 2. Η συνάρτηση πιθανότητας $u_{n,k}(i,j) = u_{n,k}(i,j;p)$ δίνεται από τις σχέσεις

(I) Για $k = 1$,

$$u_{n,k}(i,j) = \alpha_{n,k}^{-1} p^2 q^{n+i-j-1}(1-p^{j-i-1}).$$

(II) Για $k > 1$,

$$\begin{aligned} u_{n,k}(i,j) &= {\alpha_{n,k}}^{-1} p^n (p^{2k-j+i-1}-1) \sum_{y_2=\lfloor \frac{i+k-2}{k} \rfloor}^{i-1} C(i-1-y_2, y_2, k-1)(q/p)^{y_2} \\ &\quad \times \sum_{y_3=\lfloor \frac{n-j+k-1}{k} \rfloor}^{n-j} C(n-j-y_3, y_3, k-1)(q/p)^{y_3}. \end{aligned} \quad (14)$$

Το ακόλουθο πόρισμα παρέχει έναν εναλλακτικό τρόπο προσδιορισμού της $f_{n,k}(d)$ μέσω της $u_{n,k}(i,j)$.

Πόρισμα 2. Η συνάρτηση πιθανότητας $f_{n,k}(d)$, $d = 2k+1, 2k+2, \ldots, n$, $n \geq 2k+1$, $k \geq 1$ δίνεται από τη σχέση

$$f_{n,k}(d) = \sum_{i=1}^{n-d+1} u_{n,k}(i, i+d-1). \quad (15)$$

3. ΑΡΙΘΜΗΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Στην ενότητα αυτή παρουσιάζουμε αριθμητικά παραδείγματα τα οποία μέσω του σχολιασμού τους βοηθούν περαιτέρω στην κατανόηση των θεωρητικών αποτελεσμάτων της ενότητας 2. Για την εξαγωγή των αριθμητικών αποτελεσμάτων που εμφανίζονται στα παραδείγματα αυτά έχουν υλοποιηθεί οι τύποι της ενότητας 2.

Παράδειγμα 1. Στον Πίνακα 1 δίνονται οι πιθανότητες $u_{10,k}(i,j)$ για $p = 0.5$ και $k = 1,3$. Οι πιθανότητες αυτές υπολογίσθηκαν μέσω των σχέσεων (6-I,II) και (14-I,II). Στον Πίνακα 2 δίνονται οι πιθανότητες $f_{10,k}(d)$ για τα ίδιες τιμές των p και k που χρησιμοποιήθηκαν στον Πίνακα 1. Οι πιθανότητες αυτές έχουν υπολογισθεί μέσω της σχέσης (15) χρησιμοποιώντας τα στοιχεία του Πίνακα 1. Φυσικά οι $f_{10,k}(d)$, $k = 1,3$ ταυτίσθηκαν με αυτές που προκύπτουν αν χρησιμοποιήσουμε τις (6-I,II), ((8-I,II) ή (9)-(10)) και τη σχέση (12).

Πίνακας 1: $u_{10,k}(i,j)$ για $k=1,3$ και $p=0.5$.

i/j	3	4	5	6	7	8	9	10
$u_{10,1}(i,j)$, $i=1,2,\ldots,8$ και $j=i+2,i+3,\ldots,10$								
1	0.00103306	0.00309917	0.00723140	0.01549587	0.03202479	0.06508264	0.13119835	0.26342980
2		0.00103306	0.00309917	0.00723140	0.01549587	0.03202479	0.06508264	0.13119835
3			0.00103306	0.00309917	0.00723140	0.01549587	0.03202479	0.06508264
4				0.00103306	0.00309917	0.00723140	0.01549587	0.03202479
5					0.00103306	0.00309917	0.00723140	0.01549587
6						0.00103306	0.00309917	0.00723140
7							0.00103306	0.00309917
8								0.00103306
$u_{10,3}(i,j)$, $i=1,2,3,4$ και $j=i+2,i+3,\ldots,10$								
1					0.07142857	0.10714286	0.12500000	0.26785714
2						0.03571429	0.05357142	0.12500000
3							0.03571429	0.10714286
4								0.07142857

Πίνακας 2: $f_{10,k}(d)$ για $k=1,3$ και $p=0.5$.

d	3	4	5	6	7	8	9	10
$k=1$	0.00826448	0.02169419	0.04338840	0.07747935	0.12809916	0.19524792	0.26239670	0.26342980
$k=3$					0.21428572	0.26785714	0.25000000	0.26785714

Παράδειγμα 2. Αριθμητικός προσδιορισμός κριτηρίων ελαχίστου μήκους και λειτουργικής απόστασης.

I. Κριτήριο επιλογής ελαχίστου μήκους.

Η πιθανότητα $\alpha_{n,k}(p)$, $0<p<1$ μπορεί να χρησιμοποιηθεί για την επιλογή του ελαχίστου μήκους που πρέπει να έχει μια δυαδική ακολουθία ώστε να υπάρχουν τουλάχιστον 2 k-ροές με πιθανότητα τουλάχιστον ίση με β, $0<\beta<1$. Το μήκος αυτό (αριθμός δοκιμών) ορίζεται ως εξής

$$n^*(k,p;\beta)=\min\{n\geq 2k+1 : \alpha_{n,k}(p)\geq\beta\},\quad 0<\beta<1. \tag{16}$$

Ως αριθμητικό παράδειγμα, έστω $p=0.75$, $k=5$ (συνήθεις επιλογές σε ακολουθία DNA) και $\beta=95\%,99\%$. Τότε από τις σχέσεις (6-III) και (16) βρίσκουμε ότι $n^*(5,0.75;0.95)=56$ και $n^*(5,0.75;0.99)=74$. Δηλαδή, χρειάζεται να περιμένουμε την πραγματοποίηση τουλάχιστον 56 (74) πειραμάτων έτσι ώστε να έχουμε πιθανότητα τουλάχιστον 95% (99%) ότι θα παρατηρήσουμε τουλάχιστον 2 5-ροές από 1 σε μια ακολουθία από ανεξάρτητα και ισόνομα 0-1 πειράματα κάθε ένα από τα οποία έχει πιθανότητα $p=0.75$ να είναι 1. Για ακολουθίες με $p=0.75$, μεγέθους μικρότερου από 56 (74) χρειάζεται να χρησιμοποιήσουμε ένα μικρότερο k, δηλ. $k<5$, έτσι ώστε να εμφανισθούν τουλάχιστον 2 k-ροές από 1 με πιθανότητα τουλάχιστον 95% (99%).

II. Κριτήριο προσδιορισμού λειτουργικής απόστασης.

Για ένα δοθέν μήκος n ακολουθίας, με $n\geq n^*(k,p;\beta)$, δηλ. για ακολουθίες που το μήκος τους ικανοποιεί το κριτήριο ελαχίστου μήκους $\alpha_{n,k}(p)\geq\beta$, είναι πολλές φορές χρήσιμο να καθορίσουμε τη μέγιστη τιμή της απόστασης, έστω d_c, για την οποία η πιθανότητα εμφάνισης απόστασης d τουλάχιστον ίσης με d_c, είναι τουλάχιστον ίση

με γ, $0 < \gamma < 1$. Συγκεκριμένα η απόσταση αυτή $d_c = d_c(n, k, p; \beta, \gamma)$ για κάθε $n \geq n^*(k, p; \beta)$ ορίζεται από τις σχέσεις

$$P(D_{n,k} \geq d_c \mid \mathcal{M}_{n,k}) \geq \gamma, \quad P(D_{n,k} > d_c \mid \mathcal{M}_{n,k}) < \gamma, \quad 0 < \gamma < 1. \tag{17}$$

Το γ δεν απαιτείται να είναι ίσο με το β. Η απόσταση d_c μπορεί να θεωρηθεί ως κριτήριο της λειτουργικής (μέγιστης) απόστασης παρόμοιο στη λογική με εκείνο που ορίζεται στην εργασία του Benson (1999) και το οποίο είναι χρήσιμο για την επεξεργασία ακολουθιών DNA.

Ως αριθμητικό παράδειγμα θεωρούμε (όπως και στο κριτήριο I) ότι $p = 0.75$, $k = 5$ και $\gamma = \beta = 0.95$. Τότε για $n = 56 = n^*(5, 0.75; 0.95)$ και για $n = 100 > n^*$ βρίσκουμε, με χρήση των σχέσεων (6-III), ((8-III) ή (9)-(10)), της (12) και της (17), αντίστοιχα, ότι $d_c(56, 5, 0.75; 0.95, 0.95) = 22$ και $d_c(100, 5, 0.75; 0.95, 0.95) = 59$. Οπότε, π.χ. σε ακολουθία μήκους $n = 100$ η λειτουργική (ή φαινόμενη) απόσταση είναι 59. Το αποτέλεσμα αυτό στην περίπτωση ακολουθίας η οποία προέρχεται από DNA, θα μπορούσε να σημαίνει ότι αν οι 5-ροές περιορισθούν σε απόσταση $d < d_c = 59$ τότε είτε η επανάληψη αυτή δεν είναι δίδυμη (tandem repeat) ή ότι δεν έχουμε δει ακόμη αρκετό τμήμα της ακολουθίας ώστε να πεισθούμε για τέτοιες εμφανίσεις.

ABSTRACT

Let a sequence of binary (zero-one or failure-success) trials ordered on a line. We consider runs of successes of length at least equal to a fixed positive integer number. The statistics denoting the size (length) as well as the starting and ending positions of the minimum subsequence containing all such success runs are defined and studied. The present study deals with conditional probability distributions of these statistics given that the number of such success runs in the sequence is at least equal to two. The study is developed on sequences of independent and identically distributed binary random variables. Numerical examples illustrate further the theoretical results.

Ευχαριστίες: Ευχαριστούμε τον ανώνυμο κριτή για τις χρήσιμες και εύστοχες παρατηρήσεις του στην αρχική έκδοση της εργασίας.

ΑΝΑΦΟΡΕΣ

Balakrishnan, N. and Koutras, M. V. (2002). *Runs and Scans with Applications,* New York: John Wiley.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-580.

Charalambides, Ch. A. (2002). *Enumerative Combinatorics,* Boca Raton: Chapman & Hall/CRC.

Demir, S. and Eryilmaz, S. (2010). Run statistics in a sequence of arbitrarily dependent binary trials. *Stat. Papers* **51**, 959-973.

Fu, J.C and Koutras, M.V. (1994). Distribution theory of runs: a Markov chain approach. *J. Amer. Statist. Assoc.* **89**, 1050-1058.

Fu, J.C. and Lou, W.Y.W. (2003). *Distribution Theory of Runs and Patterns and its Applications: A Finite Markov Chain Approach*, River Edge: World Scientific.

Koutras, M.V. (2003). Applications of Markov chains to the distribution theory of runs and patterns, in: Shanbhag DN, Rao CR (Eds), *Handbook of Statistics,* **21**, North Holland: Elsevier, pp. 431-472.

Lou, W.Y.W. (2003). The exact distribution of the k-tuple statistic for sequence homology. *Statist. Probab. Lett.* **61**, 51-59.

Makri, F.S., Philippou, A.N. and Psillakis, Z.M. (2007). Success run statistics defined on an urn model. *Adv. Appl. Probab.* **39**, 991-1019.

Makri, F.S. and Psillakis, Z.M. (2011). On runs of length exceeding a threshold: normal approximation. *Stat. Papers* **52**, 531-551.

Makri, F.S., Psillakis, Z.M. and Arapis, A.N. (2015). Length of the minimum sequence containing repeats of success rus. *Statist. Probab. Lett.* **96**, 28-37.

Riordan, J. (1964). *An Introduction to Combinatorial Analysis, 2nd ed.,* New York: John Wiley.

Sinha, K. and Sinha, B.P. (2009). On the distribution of runs of ones in binary strings. *Comput. Math. Appl.* **58**, 1816-1829.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής (2015), σελ.32-46

ΣΤΑΤΙΣΤΙΚΕΣ ΠΡΟΣΕΓΓΙΣΕΙΣ ΚΑΙ ΑΞΙΟΛΟΓΗΣΗ ΤΗΣ ΕΠΟΧΙΑΚΗΣ ΠΡΟΓΝΩΣΗΣ ΚΑΙΡΟΥ

Γ. Βάρλας, Π. Κατσαφάδος
Τμήμα Γεωγραφίας, Χαροκόπειο Πανεπιστήμιο Αθηνών
{gvarlas, pkatsaf}@hua.gr

ΠΕΡΙΛΗΨΗ

Σε σημαντικά ερευνητικά και επιχειρησιακά κέντρα λειτουργούν τα τελευταία χρόνια μεσοπρόθεσμες και εποχιακές πιθανολογικές προγνώσεις (ensemble prediction) παράλληλα με τις αντίστοιχες ντετερμινιστικές. Η Ομάδα Δυναμικής της Ατμόσφαιρας και του Κλίματος (ΟΔΑΚ) του Τμήματος Γεωγραφίας του Χαροκόπειου Πανεπιστημίου παρέχει ελεύθερα από το Νοέμβριο του 2014 εποχιακές προγνώσεις για πρώτη φορά στην Ελλάδα. Το συγκεκριμένο προγνωστικό προϊόν αναφέρεται στην ευρύτερη περιοχή της Ευρώπης και προσφέρεται με τη μορφή θεματικών χαρτών. Η πρόσβαση του κοινού στην πληροφορία γίνεται ελεύθερα μέσω της ιστοσελίδας http://meteoclima.gr. Οι εποχιακές προγνώσεις προέρχονται από πιθανολογικές προσομοιώσεις και βασίζονται στην μεθοδολογία LAF (Lagged Average Forecast). Η πιθανολογική πρόγνωση της ΟΔΑΚ περιλαμβάνει 15 μέλη με παγκόσμια κάλυψη σε ανάλυση 1.4^ox1.4^o και προγνωστικό εύρος ενός έτους. Στο πλαίσιο της συγκεκριμένης εργασίας παρουσιάζεται η μεθοδολογία στοχαστικής ανάλυσης των αποτελεσμάτων καθώς και η αξιολόγησή τους με εφαρμογή στατιστικών μεθόδων. Για την εκτίμηση των εποχιακών τάσεων της θερμοκρασίας υπολογίζονται οι θερμοκρασιακές ανωμαλίες σε σχέση με τις μέσες μηνιαίες κλιματολογικές τιμές της περιόδου 1971-2000. Ο προσδιορισμός της αβεβαιότητας στο πεδίο της θερμοκρασίας προκύπτει από διαγράμματα τύπου spaghetti, ενώ για την ανάλυση των τάσεων του υετού εκτιμάται η χωρική απεικόνιση της πιθανότητας εμφάνισης μηνιαίου υετού που υπερβαίνει προκαθορισμένα όρια. Επιπλέον, τα αποτελέσματα αξιολογούνται με βάση δεδομένα ανάλυσης, που αποτελούν τη βέλτιστη αποτύπωση της συνοπτικής κατάστασης της ατμόσφαιρας σε πλεγματική μορφή. Η αξιολόγηση των προγνώσεων βασίζεται στον υπολογισμό καθιερωμένων στατιστικών δεικτών όπως Bias, RMSE, Mean, STD, Pearson, R^2, Scatter Index, σε σχέση με πλεγματικά δεδομένα ανάλυσης για την υπό εξέταση περίοδο. Η αξιολόγηση πραγματοποιήθηκε για τις μέσες μηνιαίες θερμοκρασίες στο ισοβαρικό επίπεδο των 850 hPa για το χειμώνα 2014-5. Τα αποτελέσματα της αξιολόγησης δείχνουν ότι οι εποχιακές προγνώσεις περιέχουν αξιοσημείωτη προγνωστική ικανότητα.

Λέξεις Κλειδιά: Εποχιακή πρόγνωση καιρού, ατμοσφαιρικό μοντέλο, αριθμητική προσομοίωση.

1. ΕΙΣΑΓΩΓΗ

Η ατμόσφαιρα αποτελεί ένα ιδιαίτερα περίπλοκο δυναμικό σύστημα με πολλούς βαθμούς ελευθερίας. Η κατάσταση της ατμόσφαιρας περιγράφεται με τις χωρικές κατανομές του ανέμου, της θερμοκρασίας και επιπλέον μετεωρολογικών μεταβλητών, όπως της ειδικής υγρασίας και της επιφανειακής πίεσης. Οι διαφορικές εξισώσεις που περιγράφουν τη χρονική και χωρική εξέλιξη του συστήματος περιλαμβάνουν τους θεμελιώδους νόμους κίνησης του Νεύτωνος καθώς και θερμοδυναμικούς νόμους. Το σύνολο των διαφορικών εξισώσεων περιγράφει προσεγγιστικά την εξέλιξη της ατμοσφαιρικής κατάστασης και, λόγω πολυπλοκότητας, δεν επιδέχεται αναλυτικές λύσεις.

Η πρόγνωση καιρού βασίζεται στη χρήση αριθμητικών μοντέλων και απαιτεί ακριβή προσομοίωση των επικρατούντων ατμοσφαιρικών φαινομένων. Σημείο εκκίνησης κάθε αριθμητικής ολοκλήρωσης αποτελούν οι αρχικές συνθήκες, οι οποίες υπολογίζονται με βάση θεωρίες αντικειμενικής ανάλυσης και εκτιμούν την κατάσταση της ατμόσφαιρας από τις διαθέσιμες παρατηρήσεις μετρητικών σταθμών και δορυφόρων. Το περιορισμένο πλήθος διαθέσιμων παρατηρήσεων (σε σχέση με τους βαθμούς ελευθερίας του συστήματος) εισάγουν αβεβαιότητες στις αρχικές συνθήκες. Η παρουσία αβεβαιοτήτων στις αρχικές συνθήκες αποτελούν την πρωτογενή πηγή προγνωστικού σφάλματος. Οι περιορισμένες δυνατότητες λεπτομερούς και αξιόπιστης περιγραφής των φυσικών διεργασιών σε συνδυασμό με τις προκαθορισμένες χωρικά και χρονικά διεργασίες, που προσομοιώνουν τα αριθμητικά μοντέλα, αποτελούν τις δευτερεύουσες πηγές προγνωστικού σφάλματος. Οι δύο πηγές προγνωστικού σφάλματος οδηγούν σε εκφυλισμό της προγνωστικής ικανότητας με την πάροδο του χρόνου.

Η πρόγνωση καιρού, ως ατμοσφαιρική διαδικασία, μπορεί να καθοριστεί με βάση τη χρονική εξέλιξη κατάλληλης συνάρτησης πυκνότητας πιθανότητας. Το συγκεκριμένο θέμα μπορεί να προσεγγιστεί με χρήση εξίσωσης συνέχειας πιθανότητας (Liouville), μέθοδος που δεν έχει επικρατήσει ευρέως, ή με εφαρμογή πεπερασμένου πλήθους ντετερμινιστικών ολοκληρώσεων για εκτίμηση της συνάρτησης πυκνότητας πιθανότητας ανεξάρτητα από το εύρος ανάπτυξης γραμμικού σφάλματος. Οι πιθανολογικές προγνώσεις (ensemble predictions) αποφέρουν εκτιμήσεις της προγνωστικής ικανότητας από ντετερμινιστικές προσομοιώσεις, ουσιαστικά προσφέροντας προγνώσεις της προγνωστικότητας.

Σε σημαντικά ερευνητικά και επιχειρησιακά κέντρα λειτουργούν τα τελευταία χρόνια μεσοπρόθεσμες και εποχιακές πιθανολογικές προγνώσεις (ensemble prediction) παράλληλα με τις ντετερμινιστικές. Η Ομάδα Δυναμικής της Ατμόσφαιρας και του Κλίματος (ΟΔΑΚ) του Τμήματος Γεωγραφίας του Χαροκόπειου Πανεπιστημίου παρέχει ελεύθερα από το Νοέμβριο του 2014 εποχιακές προγνώσεις για πρώτη φορά στην Ελλάδα. Το συγκεκριμένο προγνωστικό προϊόν αναφέρεται στην ευρύτερη περιοχή της Ευρώπης και προσφέρεται με τη μορφή θεματικών χαρτών. Η πρόσβαση του κοινού στην πληροφορία γίνεται ελεύθερα μέσω της ιστοσελίδας http://meteoclima.gr. Οι εποχιακές προγνώσεις προέρχονται από πιθανολογικές προσομοιώσεις και βασίζονται στην μεθοδολογία LAF (Lagged Average Forecast), οποία θα αναλυθεί σε επόμενη παράγραφο.

Η πιθανολογική πρόγνωση της ΟΔΑΚ περιλαμβάνει 15 μέλη με παγκόσμια κάλυψη σε ανάλυση 1.4^{o}x1.4^{o} και προγνωστικό εύρος ενός έτους. Στο πλαίσιο της συγκεκριμένης εργασίας παρουσιάζεται η μεθοδολογία στοχαστικής ανάλυσης των αποτελεσμάτων καθώς και η αξιολόγησή τους με εφαρμογή στατιστικών μεθόδων. Για την εκτίμηση των εποχιακών τάσεων της θερμοκρασίας υπολογίζονται οι θερμοκρασιακές ανωμαλίες σε σχέση με τις μέσες μηνιαίες κλιματολογικές τιμές της περιόδου 1971-2000. Ο προσδιορισμός της αβεβαιότητας στο πεδίο της θερμοκρασίας προκύπτει από διαγράμματα τύπου spaghetti, ενώ για την ανάλυση των τάσεων του υετού εκτιμάται η χωρική απεικόνιση της πιθανότητας εμφάνισης μηνιαίου υετού που υπερβαίνει προκαθορισμένα όρια. Η αξιολόγηση των προγνώσεων βασίζεται στον υπολογισμό καθιερωμένων στατιστικών δεικτών (Bias, RMSE, Mean, STD, Pearson, R^2, Scatter Index) σε σχέση με πλεγματικά δεδομένα ανάλυσης για την υπό εξέταση περίοδο. Η μεθοδολογία της στατιστικής αξιολόγησης των αποτελεσμάτων θα αναλυθεί σε επόμενη παράγραφο.

2. ΜΕΘΟΔΟΛΟΓΙΑ ΕΠΟΧΙΑΚΗΣ ΠΡΟΓΝΩΣΗΣ

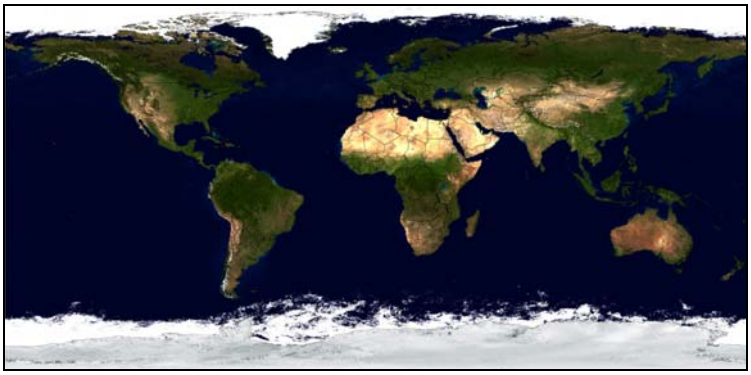
Η εποχιακή πρόγνωση βασίζεται σε επαναλαμβανόμενες μεσοπρόθεσμες προσομοιώσεις με διαφορετικές αρχικοποιήσεις κάθε φορά. Το αριθμητικό μοντέλο WRF-ARW Global ή GWRF (Skamarock et al., 2008) προσαρμόστηκε πρόσφατα για να προσομοιώνει την κατάσταση της ατμόσφαιρας σε παγκόσμιο επίπεδο, προσφέροντας τη δυνατότητα στους χρήστες για μεσοπρόθεσμες προγνώσεις (Zhang et al., 2012). Οι εποχιακές προγνώσεις της ΟΔΑΚ προέρχονται από προσομοιώσεις με το μοντέλο GWRF σε 257x129 σημεία επίλυσης σε παγκόσμιο επίπεδο (Εικόνα 1). Η ορίζοντια ανάλυση του πλέγματος επίλυσης είναι 1.4°x1.4° με 32 κατακόρυφα επίπεδα διακριτοποίησης έως ύψος 50 hPa στην ατμόσφαιρα και 10 λεπτά χρονικό βήμα επίλυσης. Για τις εποχιακές προγνώσεις πραγματοποιήθηκαν 15 πιθανολογικές προσομοιώσεις με αρχικοποιήσεις την περίοδο 17-31 Αυγούστου 2014. Οι αρχικές συνθήκες εξασφαλίστηκαν από τη βάση δεδομένων GFS (NCEP/NOAA) σε οριζόντια ανάλυση 0.5°x0.5° και ώρα 00:00UTC. Οι 15 πιθανολογικές προσομοιώσεις βασίστηκαν στην μεθοδολογία LAF.

Η LAF (Lagged Averaged Forecasting) προτάθηκε από τους Hoffman και Kalnay το 1983. Κατά τη συγκεκριμένη διαδικασία τα προγνωστικά μέλη προέρχονται από ολοκληρώσεις διαδοχικών κύκλων. Κατά τη μέθοδο LAF, η πιθανολογική πρόγνωση συντίθεται από προγνώσεις που εκκινούν από διαδοχικές αναλύσεις (Εικόνα 2). Η διαφοροποίηση ανάμεσα στην ανάλυση και στην ιδιαίτερα περιορισμένης χρονικής έκτασης πρόγνωση (very-short-range forecast) σε αντίστοιχες περιόδους μπορεί συνεπώς να θεωρηθεί ως αναπτυσσόμενη διατάραξη στο πεδίο των αρχικών συνθηκών. Η συγκεκριμένη μέθοδος εφαρμόστηκε και στην περίπτωση προσομοίωσης μίας μεγάλης κλίμακας κύματος καύσωνα στη Ρωσία το 2010 (15 Ιουλίου-15 Αυγούστου) με ικανοποιητικά αποτελέσματα (Katsafados et al., 2014). Για την τρέχουσα εποχιακή πρόγνωση οι 15 διαφορετικής αρχικοποίησης προσομοιώσεις είχαν προγνωστικό εύρος 1 έτος με επιπλέον κάποιες ημέρες, ώστε το κοινό προγνωστικό διάστημα να διαμορφωθεί σε 1 ολόκληρο έτος από την $1^{η}$ Σεπτεμβρίου 2014 00:00UTC έως την $1^{η}$ Σεπτεμβρίου 2015 00:00UTC. Τα

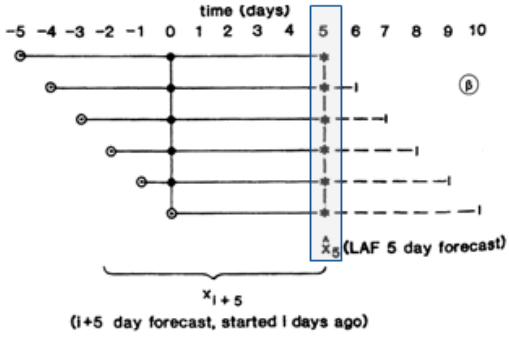
αποτελέσματα των προσομοιώσεων είναι διαθέσιμα κάθε 12 ώρες, δηλαδή για κάθε μέρα 00 και 12 UTC.

Επίσης, το μοντέλο τροποποίηθηκε κατάλληλα ώστε να χρησιμοποιηθεί ανανέωση της επιφανειακής θερμοκρασίας της θάλασσας (ΕΘΘ) και της κατανομής του θαλάσσιου πάγου με βάση κλιματολογικές τιμές. Πιο αναλυτικά, η ΕΘΘ παρέμενε σταθερή για τις πρώτες 15 μέρες της κάθε προσομοίωσης και ίση με την ανάλυση της ΕΘΘ την ημέρα της αρχικοποίησης για ώρα 00:00UTC. Έπειτα από τις πρώτες 15 μέρες προσομοίωσης, το πεδίο της ΕΘΘ δεχόταν ανανέωση από κλιματολογικές τιμές ΕΘΘ με βάση τη μηνιαία κλιματολογία NCEP/NOAA (1981-2010) (Saha et al., 2010). Παρόμοια ήταν η μεθοδολογία της ανανέωσης και για την κατανομή του θαλάσσιου πάγου, μόνο που η ανανέωση γινόταν κάθε 1 μήνα. Τα βασικά χαρακτηριστικά των προσομοιώσεων συνοψίζονται στον Πίνακα 1.

Εικόνα 1. *Περιοχή κάλυψης των αριθμητικών προσομοιώσεων.*



Εικόνα 2. *Παράδειγμα εφαρμογής της μεθόδου LAF σε 6 ιδεατές προσομοιώσεις με διαφορετική αρχικοποίηση και ίδιο προγνωστικό εύρος. Το διάστημα ημερών από 0-5 είναι κοινό και για τις 6 προσομοιώσεις (*Hoffman and Kalnay, 1983).



Πίνακας 1. *Βασικά χαρακτηριστικά των αριθμητικών προσομοιώσεων.*

Μοντέλο	GWRF
Οριζόντια ανάλυση	1.4°x1.4°

Κατακόρυφη ανάλυση	32 sigma-pressure επίπεδα μέχρι τα 50hPa
Σημεία επίλυσης	257x129
Χρονικό βήμα	600΄΄
Προγνωστικός ορίζοντας	1 έτος
Σύνολο προσομοιώσεων	15
Χρονικές στιγμές αποτελεσμάτων	00 και 12 UTC
Αρχικές συνθήκες	GFS NCEP/NOAA (0.5°x0.5°) ανάλυση 00:00UTC
Μοντέλο επιφάνειας-εδάφους	4-layer NOAH (Chen and Dudhia 2001)
Επιφανειακή θερμοκρασία θάλασσας (ΕΘΘ)	Ημερήσια ανανέωση της ΕΘΘ μετά τις πρώτες 15 ημέρες της προσομοίωσης από τη μηνιαία κλιματολογία NCEP/NOAA (1981-2010) (Saha et al., 2010)
Κατανομή θαλάσσιου πάγου	Μηνιαία ανανέωση της κατανομής του θαλάσσιου πάγου μετά τις πρώτες 15 ημέρες της προσομοίωσης με βάση παραμετροποιήσεις της μηνιαίας κλιματολογίας της ΕΘΘ των NCEP/NOAA (1981-2010)

3. ΜΕΘΟΔΟΛΟΓΙΕΣ ΑΝΑΛΥΣΗΣ ΚΑΙ ΣΤΑΤΙΣΤΙΚΗΣ ΑΞΙΟΛΟΓΗΣΗΣ

Η ανάλυση και η στατιστική αξιολόγηση των αποτελεσμάτων της εποχιακής πρόγνωσης βασίστηκε σε 2 επί μέρους μεθοδολογίες, με περίοδο αξιολόγησης το χειμώνα 2014-5 (Δεκέμβριος, Ιανουάριος και Φεβρουάριος). Η ανάλυση και η επεξεργασία των δεδομένων, καθώς και η διαδικασία στατιστικής αξιολόγησης έγινε με τη γλώσσα προγραμματισμού NCAR Command Language (NCL, 2014). Η πρώτη μεθοδολογία, που αποτελεί ουσιαστικά ένα είδος στοχαστικής ανάλυσης, σχετίζεται με την εκτίμηση της αβεβαιότητας κάθε πρόγνωσης από το σύνολο των 15 διαφορετικών προγνώσεων καθώς και σύγκριση της μέσης τιμής τους με κλιματολικά δεδομένα. Η δεύτερη μεθοδολογία σχετίζεται με τη στατιστική αξιολόγηση των προγνώσεων με δεδομένα αναλύσεων, που θεωρούνται η βέλτιση αναπαράσταση της συνοπτικής κατάστασης σε πλεγματική μορφή.

Κατά τη διάρκεια της πρώτης φάσης της αξιολόγησης των προγνώσεων, δημιουργήθηκε νέφος προγνώσεων (ensemble) της θερμοκρασίας στα 850 hPa για

την εκτίμηση της αβεβαιότητας κάθε πρόγνωσης (μέλους). Έπειτα, έγινε υπολογισμός των θερμοκρασιακών ανωμαλιών σε σχέση με την πλεγματική κλιματολογία των NCEP/NCAR της περιόδου 1968-1996 σε ανάλυση 2.5°x2.5° (Kalnay et al., 1996) για την ευρύτερη περιοχή της Ευρώπης, σε γεωγραφικό πλάτος από 25 έως 72 μοίρες και γεωγραφικό μήκος από -25 έως 50 μοίρες. Έτσι, δημιουργήθηκαν τα διαγράμματα τύπου spaghetti για κάθε μήνα, τα οποία αποτελούνται από τις χωρικές κατανομές κάποιας προκαθορισμένης τιμής θερμοκρασίας και των 15 μελών, της μέσης τιμής τους, της διαμέσου και της κλιματολογικής τιμής. Επίσης, έγινε χωρική απεικόνιση της διαφοράς των μέσων μηνιαίων θερμοκρασιών για όλα τα μέλη, με τις αντίστοιχες μηνιαίες κλιματολογικές τιμές. Για μια πιο λεπτομερή αξιολόγηση των προγνώσεων, εκτός από τη θερμοκρασία έγινε αξιολόγηση και για το συνολικό υετό (βροχή, χιόνι, χαλάζι). Έτσι λοιπόν, έγινε εκτίμηση της πιθανότητας κάθε προγνωστικό μέλος να υπερβεί ένα προκαθορισμένο ύψος μηνιαίου υετού (π.χ. 200 χιλιοστά υετού ή 200 τόνους ανά στρέμμα).

Στη δεύτερη φάση της αξιολόγησης των προγνώσεων, τα πεδία μέσων τιμών θερμοκρασίας στα 850 hPa διαμορφώθηκαν κατάλληλα, ώστε να συγκριθούν με τις μέσες τιμές των αντίστοιχων πεδίων ανάλυσης (GFS-ANL) του παγκόσμιου προγνωστικού μοντέλου GFS (NCEP/NOAA) σε οριζόντια ανάλυση 0.5°x0.5°. Η σύγκριση των δύο πεδίων έγινε με τη μέθοδο point to point, δηλαδή για κάθε τιμή του πεδίου ανάλυσης δημιουργήθηκε με βάση τα 4 γειτονικά σημεία και τη μέθοδο διγραμμικής παρεμβολής (bilinear interpolation) η αντίστοιχη τιμή του πεδίου προγνώσεων. Η στατιστική αξιολόγηση έγινε για 2 περιοχές. Αρχικά τα 2 πεδία αξιολογήθηκαν σε παγκόσμιο επίπεδο έως τις 70 μοίρες νότιο και γεωγραφικό πλάτος. Έπειτα, έγινε σύγκριση πάνω από την ευρύτερη περιοχή της Ευρώπης, στην περιοχή που προαναφέρθηκε. Στο πλαίσιο της στατιστικής αξιολόγησης δημιουργήθηκαν διαγράμματα διασποράς (scatter plots) και υπολογίστηκαν βασικοί στατιστικοί δείκτες (Bias, RMSE, Mean, STD, Pearson, R^2, Scatter Index), των οποίων οι εξισώσεις (Σχέσεις 1-9) ακολουθούν. Έστω ότι M_i και O_i είναι οι πλεγματικές τιμές της πρόγνωσης και της ανάλυσης αντίστοιχα, τότε:

$$MeanObs = \frac{\sum_{i=1}^{N} O_i}{N} \tag{1}$$

$$MeanMod = \frac{\sum_{i=1}^{N} M_i}{N} \tag{2}$$

$$STDObs = \sqrt{\frac{\sum_{i=1}^{N} (O_i - \overline{O})^2}{N}} \tag{3}$$

$$STDMod = \sqrt{\frac{\sum_{i=1}^{N} (M_i - \overline{M})^2}{N}} \tag{4}$$

$$Bias = \frac{\sum_{i=1}^{N}(M_i - O_i)}{N} \tag{5}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(M_i - O_i)^2}{N}} \tag{6}$$

$$Pearson = \frac{\sum_{i=1}^{N}(O_i - \overline{O})(M_i - \overline{M})}{\sqrt{\sum_{i=1}^{N}(O_i - \overline{O})^2}\sqrt{\sum_{i=1}^{N}(M_i - \overline{M})^2}} \tag{7}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(M_i - O_i)^2}{\sqrt{\sum_{i=1}^{N}(O_i - \overline{O})^2}} \tag{8}$$

$$SI = \frac{\sqrt{\frac{\sum_{i=1}^{N}(M_i - O_i)^2}{N}}}{\frac{\sum_{i=1}^{N}O_i}{N}} = \frac{RMSE}{MeanObs} \tag{9}$$

Στην επόμενη παράγραφο θα παρουσιαστούν τα αποτελέσματα της αξιολόγησης με βάση τις 2 επιμέρους μεθοδολογίες.

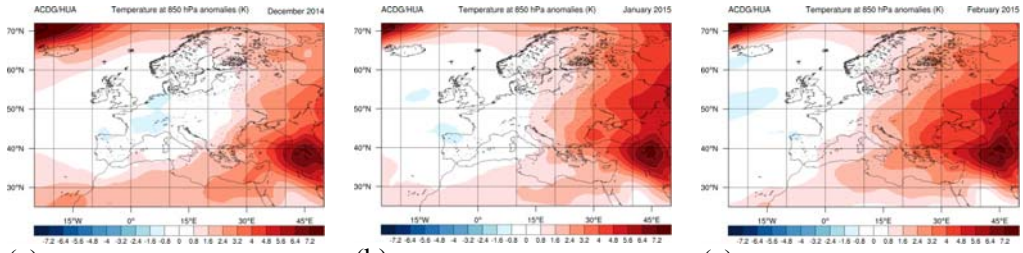
4. ΑΠΟΤΕΛΕΣΜΑΤΑ ΣΤΑΤΙΣΤΙΚΗΣ ΑΞΙΟΛΟΓΗΣΗΣ

Η στατιστική αξιολόγηση των αποτελεσμάτων της εποχιακής πρόγνωσης βασίστηκε σε στοχαστική ανάλυση των αποτελεσμάτων και σύγκρισή τους με δεδομένα κλιματολογίας και ανάλυσης, όπως αναφέρθηκε στην προηγούμενη παράγραφο. Αρχικά θα παρουσιαστούν τα αποτελέσματα της αξιολόγησης με βάση την κλιματολογία των NCEP/NCAR.

Στην Εικόνα 3 αποτυπώνεται η χωρική κατανομή της διαφοράς της μέσης μηνιαίας θερμοκρασίας στα 850 hPa για όλα τα μέλη με την αντίστοιχη μηνιαία θερμοκρασία στα 850 hPa από την κλιματολογία των δεδομένων NCEP/NCAR. Παρατηρείται παρόμοια συμπεριφορά και για τους 3 χειμερινούς μήνες, το Δεκέμβριο (Εικόνα 3a), τον Ιανουάριο (Εικόνα 3b) και το Φεβρουάριο (Εικόνα 3c). Γενικά αποτυπώνεται εκτίμηση για θερμότερες του κανονικού (κλιματολογίας) θερμοκρασίες στα 850 hPaστις περισσότερες περιοχές αν εξαιρεθεί η κεντρική και η δυτική Ευρώπη, όπου παρουσιάζονται μηδενικές και ελαφρώς αρνητικές διαφορές. Στην Εικόνα 4 παρουσιάζονται διαγράμματα τύπου spaghetti για θερμοκρασία 271 K στα 850 hPa για τον Δεκέμβριο (Εικόνα 4a) για τον Ιανουάριο (Εικόνα 4b) και για το Φεβρουάριο (Εικόνα 4c). Αρχικά, αποτυπώνονται οι υψηλότερες του κανονικού θερμοκρασίες σχεδόν από όλα τα μέλη στην Ανατολική Ευρώπη. Ωστόσο,

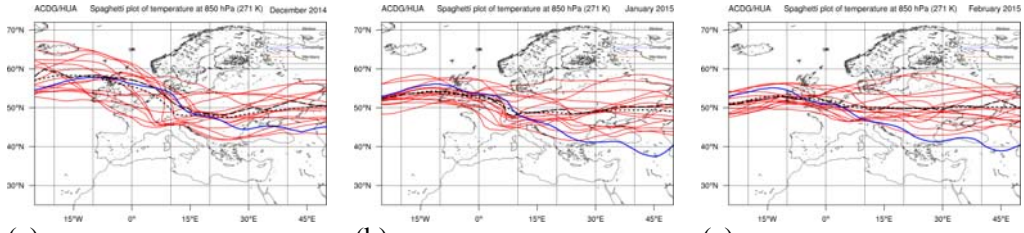
παρουσιάζουν μεγάλη διασπορά οι θέσεις των ισόθερμων για κάθε μέλος, οπότε αυξάνεται η αβεβαιότητα ενώ, αντίθετα, στη δυτική Ευρώπη παρατηρείται μεγαλύτερη σύγκλιση των ισόθερμων, και περιορισμό της αβεβαιότητας. Στην Εικόνα 5 αποτυπώνεται η χωρική κατανομή της πιθανότητας, ο συνολικός μηνιαίος υετός να υπερβεί τα 200 mm για τον Δεκέμβριο (Εικόνα 5a), για τον Ιανουάριο (Εικόνα 5b) και για τον Φεβρουάριο (Εικόνα 5c). Ένα κοινό χαρακτηριστικό που παρατηρείται και για τους 3 μήνες, είναι ότι η πιθανότητα εμφάνισης υετού πάνω από 200 mm είναι αυξημένη κυρίως στις δυτικές προσήνεμες πλευρές οροσειρών, ορεινών όγκων κτλ (π.χ. Οροσειρά της Πίνδου). Επιπλέον, φαίνεται ο πιο βροχερός μήνας συνολικά για την Ευρώπη να είναι ο Ιανουάριος, ενώ για την Ελλάδα ο Δεκέμβριος, με την πιθανότητα ο υετός να ξεπεράσει τα 200 mm, να εντοπίζεται κυρίως στα δυτικά και να ξεπερνάει το 50-60%.

Εικόνα 3. *Χωρική κατανομή της διαφοράς της μέσης μηνιαίας θερμοκρασίας στα 850 hPa για όλα τα μέλη με την αντίστοιχη μηνιαία θερμοκρασία στα 850 hPa από την κλιματολογία των NCEP/NCAR για τον a) Δεκέμβριο, b) Ιανουάριο, c) Φεβρουάριο.*



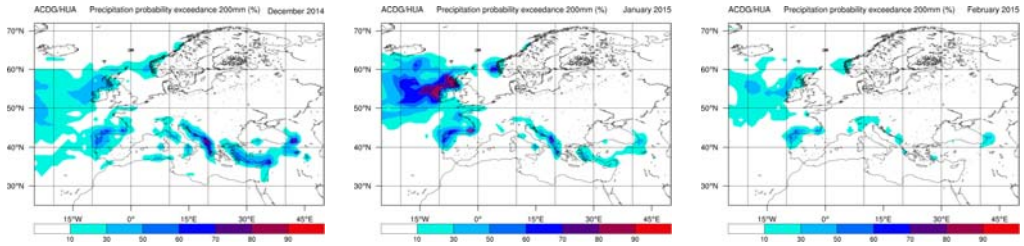
(a) (b) (c)

Εικόνα 4. *Διαγράμματα spaghetti για θερμοκρασία 271 K στα 850 hPa για τον a) Δεκέμβριο, b) Ιανουάριο, c) Φεβρουάριο.Με κόκκινο συμβολίζεται η χωρική κατανομή για κάθε μέλος, με μπλε για την κλιματολογία, με μαύρη μεγάλη διακεκομμένη καμπύλη η μέση τιμή των μελών και με μαύρη μικρή διακεκομμένη καμπύλη ο μέσος των μελών.*



(a) (b) (c)

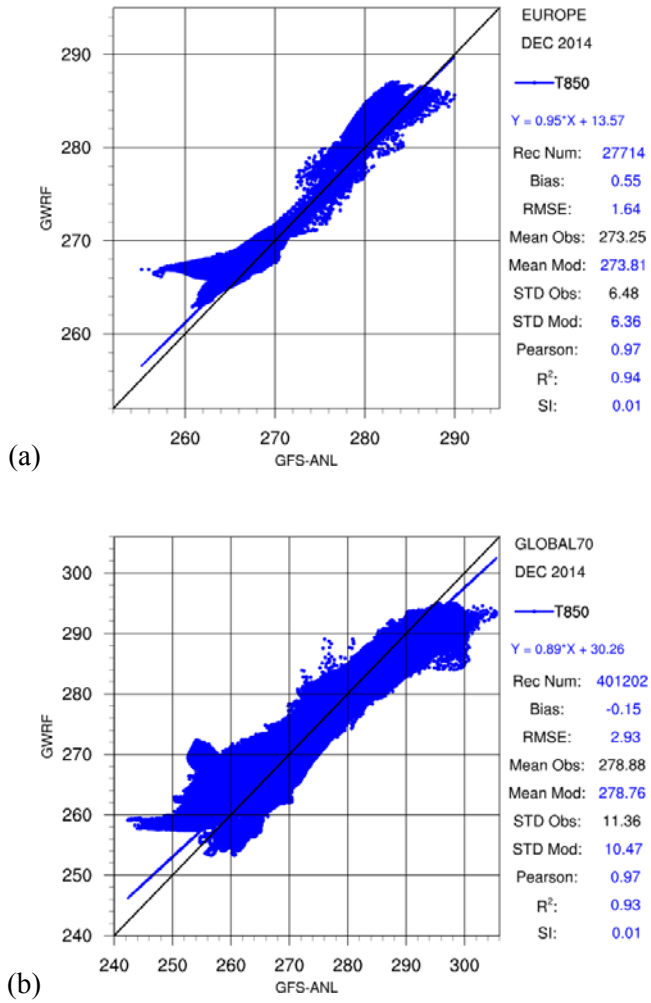
Εικόνα 5. *Χωρική κατανομή της πιθανότητας, ο συνολικός μηνιαίος υετός να ξεπεράσει τα 200 mm για τον a) Δεκέμβριο, b) Ιανουάριο, c) Φεβρουάριο.*



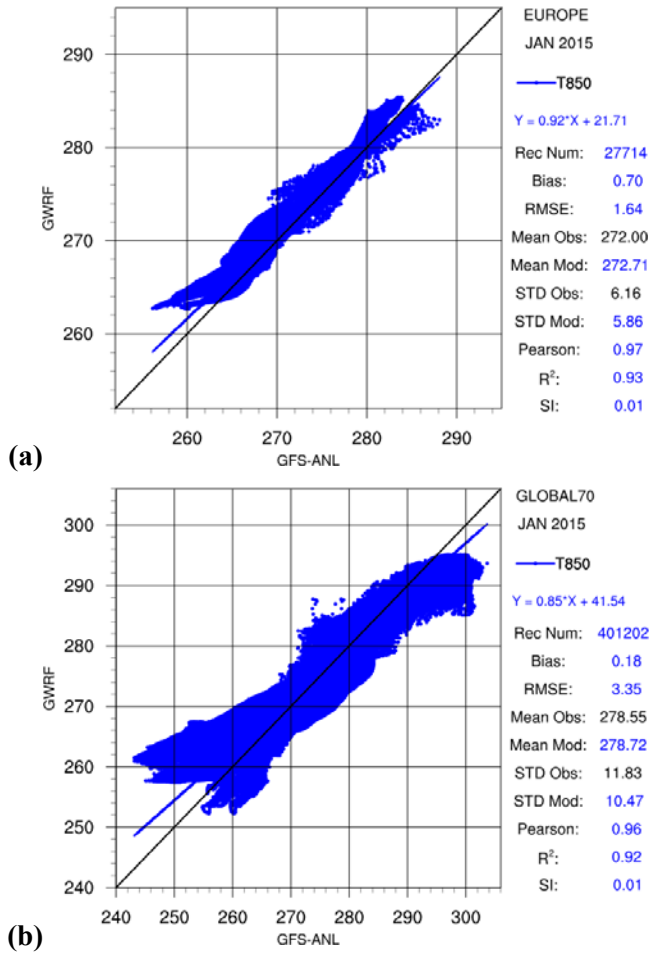
(a) (b) (c)

Η αξιολόγηση των εποχιακών προγνώσεων σε σχέση με τα πλεγματικά δεδομένα ανάλυσης GFS-ANL, εμφανίζει ικανοποιητική προγνωστική ικανότητα ακόμα και 6 μήνες μπροστά από την αρχικοποίηση. Αναλυτικότερα, στην Εικόνα 6 αποτυπώνεται ένα διάγραμμα διασποράς μεταξύ της μέσης θερμοκρασίας στα 850 hPa για όλα τα μέλη με την αντίστοιχη μηνιαία θερμοκρασία στα 850 hPa με βάση τα δεδομένα ανάλυσης GFS-ANL, για το μήνα Δεκέμβριο στην ευρύτερη περιοχή της Ευρώπης (Εικόνα 6a) και παγκοσμίως έως τις 70 μοίρες νότιο και βόρειο γεωγραφικό πλάτος (Εικόνα 6b). Τα δεδομένα εμφανίζουν ικανοποιητική συμφωνία μεταξύ τους και στις δύο περιοχές αξιολόγησης με συντελεστή συσχέτισης Pearson 0.97 και συντελεστή R^2 0.94 και 0.93 αντίστοιχα. Το σύνολο των μελών υπερεκτιμά τη θερμοκρασία στα 850 hPa με bias 0.55 και υποεκτιμά στην παγκόσμια περιοχή με bias -0.15. Το RMSE είναι σχετικά περιορισμένο, στο 1.64 για την Ευρώπη, ενώ στην παγκόσμια περιοχή αυξάνει στο στο 2.93. Επιπλέον, τα δεδομένα δεν παρουσιάζουν αξιόλογες διαφορές στη μέση τιμή (Mean) και στην τυπική απόκλιση (STD). Ο Scatter Index (SI) έχει τιμή 0.01 λόγω των περιορισμένων RMSEs, αλλά και επίσης και λόγω των μεγάλων τιμών που έχει η θερμοκρασία όταν η μονάδα μέτρησης είναι οι βαθμοί Kelvin. Γενικά, η εικόνα είναι παρόμοια και στους 2 άλλους χειμερινούς μήνες, με τη διαφορά ότι το μοντέλο υπερεκτιμά και στην Ευρώπη και παγκοσμίως με τιμές Bias 0.71 και 0.18 για τον Ιανουάριο και 0.7 και 0.18 αντίστοιχα (Εικόνες 7a, 7b, 8a, 8b). Ενδιαφέρον παρουσιάζει η συνολική αξιολόγηση του χειμώνα 2014-5, που περιλαμβάνει και τους 3 μήνες. Όπως φαίνεται στις Εικόνες 9a και 9b για την Ευρώπη και παγκοσμίως, η αξιολόγηση δείχνει καλή συσχέτιση με Pearson 0.96 και R^2 0.92. Η μέση τιμή και η τυπική απόκλιση εμφανίζουν παρόμοιες τιμές, ενώ το RMSE έχει τιμή 1.82 και 3.38 για την Ευρώπη και παγκοσμίως αντίστοιχα. Επίσης, από τις τιμές του Bias 0.66 και 0.07 αντίστοιχα για Ευρώπη και σε παγκόσμια βάση, προκύπτει μία συστηματική τάση του μοντέλου για υπερεκτίμηση της θερμοκρασίας στα 850 hPa, η οποία να περιορίζεται σημαντικά με την αύξηση του δείγματος (πχ. 1200000 ζεύγη τιμών στην παγκόσμια περιοχή).

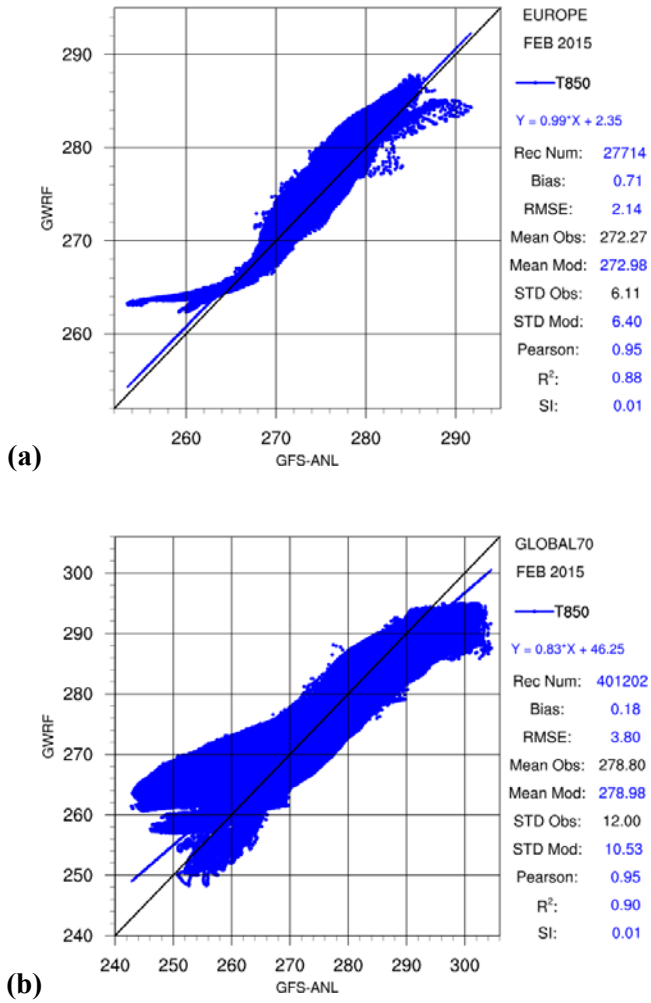
Εικόνα 6. *Διάγραμμα διασποράς μεταξύ της μέσης θερμοκρασίας στα 850 hPa για όλα τα μέλη με την αντίστοιχη μηνιαία θερμοκρασία στα 850 hPa με βάση τα δεδομένα ανάλυσης GFS-ANL, για το μήνα Δεκέμβριο a) στην ευρύτερη περιοχή της Ευρώπης b) παγκοσμίως.*



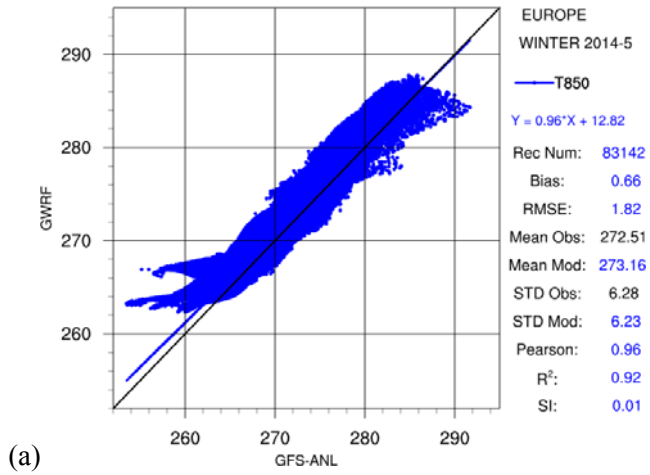
Εικόνα 7. *Διάγραμμα διασποράς μεταξύ της μέσης θερμοκρασίας στα 850 hPa για όλα τα μέλη με την αντίστοιχη μηνιαία θερμοκρασία στα 850 hPa με βάση τα δεδομένα ανάλυσης GFS-ANL, για το μήνα Ιανουάριο a) στην ευρύτερη περιοχή της Ευρώπης b) παγκοσμίως.*



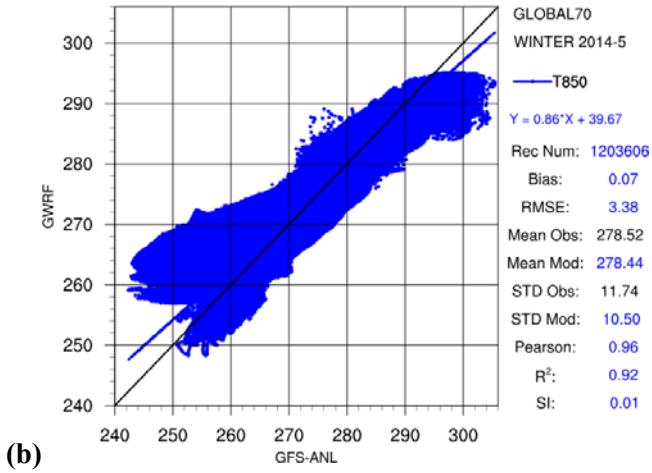
Εικόνα 8. *Διάγραμμα διασποράς μεταξύ της μέσης θερμοκρασίας στα 850 hPa για όλα τα μέλη με την αντίστοιχη μηνιαία θερμοκρασία στα 850 hPa με βάση τα δεδομένα ανάλυσης GFS-ANL, για το μήνα Φεβρουάριο a) στην ευρύτερη περιοχή της Ευρώπης b) παγκοσμίως.*



Εικόνα 9. *Διάγραμμα διασποράς μεταξύ της μέσης θερμοκρασίας στα 850 hPa για όλα τα μέλη με την αντίστοιχη μηνιαία θερμοκρασία στα 850 hPa με βάση τα δεδομένα ανάλυσης GFS-ANL, για το χειμώνα 2014-5 (Δεκέμβριος-Ιανουάριος-Φεβρουάριος) a) στην ευρύτερη περιοχή της Ευρώπης b) παγκοσμίως.*



(a)



(b)

4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Οι εποχιακές προγνώσεις της ΟΔΑΚ προέρχονται από προσομοιώσεις με το μοντέλο GWRF σε παγκόσμιο επίπεδο και ορίζοντια ανάλυση πλέγματος 1.4°x1.4°. Για τις εποχιακές προγνώσεις πραγματοποιήθηκαν 15 πιθανολογικές προσομοιώσεις που αρχικοποιήθηκαν με δεδομένα αρχικών συνθηκών από τις 15 τελευταίες μέρες του Αυγούστου 2014 (17-31 Αυγούστου 2014), τα οποία εξασφαλίστηκαν από τη βάση δεδομένων GFS (NCEP/NOAA). Οι 15 πιθανολογικές προσομοιώσεις βασίστηκαν στην μεθοδολογία LAF, κατά την οποία η πιθανολογική πρόγνωση συντίθεται από προγνώσεις που εκκινούν από διαδοχικές αναλύσεις. Τα αποτελέσματα αναλύθηκαν στοχαστικά για την εκτίμηση της αβεβαιότητας κάθε πρόγνωσης από το σύνολο των 15 διαφορετικών προγνώσεων καθώς και σύγκριση της μέσης τιμής τους με κλιματολογικά δεδομένα.

Τα αποτελέσματα της στοχαστικής ανάλυσης για τους 3 χειμερινούς μήνες Δεκέμβριο 2014 και Ιανουάριο-Φεβρουάριο 2015, έδειξαν υπερεκτίμηση στη θερμοκρασία στα 850 hPa σχεδόν σε ολόκληρη την έκταση της Ευρώπης, πλην περιοχών στην κεντρική δυτική Ευρώπη, κυρίως τους μήνες Ιανουάριο και Φεβρουάριο. Ωστόσο, η αβεβαιότητα ειδικά στην ανατολική Ευρώπη, όπου παρατηρήθηκαν τα μέγιστα της υπερεκτίμησης ήταν αυξημένη σε σχέση με την κεντρική και τη δυτική Ευρώπη. Επιπλέον, αυξημένη πιθανότητα (>50%) για μηνιαίο αθροιστικό υετό που υπερβαίνει τα 200 mm εκτιμήθηκε για τις δυτικές προσήνεμες περιοχές της Μ. Βρετανίας, της Νορβηγίας και της Βαλκανικής χερσονήσου. Για την Ανατολική Μεσόγειο ο πιο βροχερός μήνας, φαίνεται να είναι ο Δεκέμβριος, με τις μεγαλύτερες πιθανότητες για αυξημένα ποσά υετού στα δυτικά και στα πολύ ανατολικά τμήματα της χώρας.

Η στατιστική αξιολόγηση των εποχιακών προγνώσεων σε σχέση με πλεγματικά δεδομένα ανάλυσης, έδειξε ικανοποιητική συμφωνία για την περιοχή της Ευρώπης καθώς και σε παγκόσμιο επίπεδο. Το μοντέλο φαίνεται να υπερεκτιμά συστηματικά τη θερμοκρασία στα 850 hPa ενώ το RMSE αυξάνει σημαντικά σε παγκόσμιο επίπεδο.

ABSTRACT

Nowadays, medium-range and seasonal weather predictions (ensemble predictions) coexist with deterministic forecasts in many research and operational centers in the world. November 2014, the Atmosphere and Climate Dynamics Group (ACDG) of Geography department of Harokopio University provided free seasonal predictions for the first time in Greece. That forecast product is mainly referred to Europe and it is offered by thematic maps. Free public access is available at the website http://meteoclima.gr. The seasonal predictions are emanated from ensemble simulations and they are based on the LAF methodology (Lagged Average Forecast). The ensemble predictions of ACDG include 15 members and they have global coverage in a horizontal resolution of 1.4^ox1.4^o and forecast window about 1 year. In this study, a methodology of stochastic analysis and the statistical evaluation of the results are illustrated. For the estimation of the seasonal temperature trends, the temperature anomalies in contrast to the monthly climatic values of the period 1971-2000, are estimated. The temperature uncertainty is emerged by the spaghetti plots. Moreover, the spatial distribution of precipitation probability exceedance over preassigned precipitation thresholds is estimated. The seasonal predictions are statistically evaluated by gridded analyses, which optimally represent the atmospheric synoptic conditions. The statistical evaluation is based on the estimation of standard statistical indexes such as the Bias, RMSE, Mean, STD, Pearson, R^2, Scatter Index. To this end, the mean monthly temperature at 850 hPa has been evaluated for the winter 2014-15. The statistical scores indicate that the seasonal predictions include remarkable forecast skill.

ΑΝΑΦΟΡΕΣ

Hoffman R. N. and Kalnay E. (1983). Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus A*, **35(2)**, 100-118.

Kalnay E., Kanamitsu M., Kistler R., Collins W., Deaven D., Gandin L., Iredell M., Saha S., White G., Woollen J., Zhu Y., Leetmaa A., Reynolds R., Chelliah M., Ebisuzaki W., Higgins W., Janowiak J., Mo K. C., Ropelewski C., Wang J., Jenne R. and Joseph D. (1996). The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American meteorological Society*, ***77*****(3)**, 437-471.

Katsafados P., Papadopoulos A., Varlas G., Papadopoulou E., and Mavromatidis E. (2014). Seasonal predictability of the 2010 Russian heat wave, *Nat. Hazards Earth Syst. Sci*., **14**, 1531-1542.

NCL (The NCAR Command Language) (Version 6.2.1) [Software]. (2014). Boulder, Colorado: UCAR/NCAR/CISL/VETS. http://dx.doi.org/10.5065/D6WD3XH5.

Saha S., Moorthi S., Pan H., Wu X., Wang J., Nadiga S., Tripp P., Kistler R., Woollen J., Behringer D., Liu H., Stokes D., Grumbine R., Gayno G., Wang J., Hou Y., Chuang H., Juang H., Sela J., Iredell M., Treadon R., Kleist D., Delst P., Keyser D., Derber J., Ek M., Meng J., Wei H., Yang R., Lord S., Dool, V., Kumar A., Wang W., Long C., Chelliah M., Xue Y., Huang B., Schemm J., Ebisuzaki W., Lin R., Xie P., Chen M., Zhou S., Higgins W., Zou C., Liu Q., Chen Y., Han Y., Cucurull L., Reynolds R., Rutledge G. and Goldberg M. (2010). The NCEP Climate Forecast System Reanalysis. Bull. Amer. Meteor. Soc. **91**, 1015–1057.

Shamarock W., Klemp J. B., Dudhia J., Gill D. O., Barker D. M., Duda M., Huang X.-Y., Wang W. and Powers, J. G. (2008). A description of the advanced research WRF version 3. *NCAR technical note NCAR/TN/u2013475*.

Zhang Y., Hemperly J., Meskhidze N. and Skamarock W. C. (2012). The Global Weather Research and Forecasting (GWRF) Model: Model Evaluation, Sensitivity Study, and Future Year Simulation. *Atmospheric and Climate Sciences*, **2(3)**, 231-253.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής (2015), σελ.47-59

Η ΓΕΩΓΡΑΦΙΚΗ ΔΙΑΣΤΑΣΗ ΤΗΣ ΕΠΙΔΟΣΗΣ ΤΩΝ ΜΑΘΗΤΩΝ ΣΤΙΣ ΠΑΝΕΛΛΗΝΙΕΣ ΕΞΕΤΑΣΕΙΣ ΜΙΑ ΓΕΩΣΤΑΤΙΣΤΙΚΗ ΠΡΟΣΕΓΓΙΣΗ

Α. Βέρδης[1], Κ. Καλογερόπουλος[2]., Α.Γ.Παπαδόπουλος[2],Χ. Χαλκιάς[2]
[1]Τμήμα Φιλοσοφίας, Παιδαγωγικής και Ψυχολογίας, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών[2]
averdis@ppp.uoa.gr
Τμήμα Γεωγραφίας, Χαροκόπειο Πανεπιστήμιο
{kalogeropoulos, apospapa, xalkias}@hua.gr

ΠΕΡΙΛΗΨΗ

Η σχολική αποτυχία αποτελεί κεντρικό ζήτημα της Κοινωνιολογίας της Εκπαίδευσης. Το ζήτημα των ανισοτήτων στην επίδοση των μαθητών όπως και αυτό της αναπαραγωγής των κοινωνικών τάξεων μέσα από την απόκτηση τριτοβάθμιας εκπαίδευσης έχουν αναδειχθεί ως κυρίαρχα στις κοινωνιολογικές αναλύσεις που αφορούν την εκπαίδευση και τους εκπαιδευτικούς μηχανισμούς στην Ελλάδα. Η διαδικασία πρόσβασης στην τριτοβάθμια εκπαίδευση στην Ελλάδα είναι φυσικά διαφορετική από την αντίστοιχη σε άλλες χώρες. Η παρούσα εργασία αναλύει τα γεωγραφικά χαρακτηριστικά της επίδοσης των μαθητών δευτεροβάθμιας εκπαίδευσης κατά την προσπάθεια κατάκτησης μιας θέσης στο σύστημα της τριτοβάθμιας εκπαίδευσης στη χώρα.

Για τις ανάγκες της εργασίας δημιουργήθηκε χωρική βάση δεδομένων από τις επιδόσεις 81010 μαθητών της Γ' Λυκείου του εκπαιδευτικού έτους 2012-2013. Πραγματοποιήθηκε πρώτα αναγωγή σε επίπεδο σχολείου και κατόπιν σε επίπεδο Καλλικρατικού Δήμου. Στη συνέχεια πραγματοποιήθηκε αποτίμηση των χωρικών προτύπων της επίδοσης των μαθητών της Γ' Λυκείου, προκειμένου να μελετηθεί η χωρική μεταβλητότητα σε τοπικό και υπερτοπικό επίπεδο. Πραγματοποιείται, δηλαδή, μια χαρτογραφική απόδοση της επίδοσης των μαθητών σε επίπεδο Δήμου, καθώς επίσης εφαρμόζονται συγκεκριμένοι δείκτες χωρικής αυτοσυσχέτισης. Τέλος, πραγματοποιήθηκε μια προκαταρτική διερεύνηση της σχέσης της σχολικής επίδοσης με τον δείκτη εκπαίδευσης που αποτελεί μια έμμεση ένδειξη της επίδρασης των κοινωνικο-οικονομικών παραγόντων, οι οποίοι εκτιμάται ότι την επηρεάζουν.

Λέξεις κλειδιά: Μαθητική επίδοση, χωρικά πρότυπα, χωρική αυτοσυσχέτιση

1. ΕΙΣΑΓΩΓΗ

Η σχολική αποτυχία αποτελεί ένα από τα κεντρικά ζητήματα που αφορούν την κοινωνιολογία της εκπαίδευσης στη χώρα μας. Η μελέτη των εκπαιδευτικών θεσμών και των αποτελεσμάτων τους είναι κεφαλαιώδους σημασίας για τη μελέτη της κοινωνικής αναπαραγωγής σε μια κοινωνία, καθώς είναι σημαντικό να ελεγχθεί σε ποιό βαθμό και με ποιό τρόπο οι εκπαιδευτικοί θεσμοί αναπαράγουν τις κοινωνικοοικονομικές ανισότητες που υφίστανται στις σύγχρονες κοινωνίες. Η επίδοση των μαθητών και συνακόλουθα η πρόσβαση στην τριτοβάθμια εκπαίδευση αποτέλεσαν ένα μέσο βελτίωσης της κοινωνικής θέσης των ατόμων, ενώ παράλληλα θεωρήθηκε ότι θα μπορούσαν έτσι να αντιμετωπιστούν τα ζητήματα που συνδέονται με την αναπαραγωγή των κοινωνικών τάξεων και των κοινωνικών ανισοτήτων στην Ελλάδα.

Η μελέτη της πρόσβασης στην τριτοβάθμια εκπαίδευση στην Ελλάδα συνιστά το αντικείμενο της παρούσας ερευνητικής εργασίας. Στο τέλος της τελευταίας τάξης του λυκείου, δηλαδή του ανώτερου κύκλου της δευτεροβάθμιας εκπαίδευσης, διενεργούνται στη χώρα μας γενικές εξετάσεις οι οποίες, με διαφορετικούς κάθε φορά διαχωρισμούς σε κλάδους, δέσμες, κατευθύνεις, προσανατολισμούς κλπ. και διαφορετικούς κάθε τόσο συντελεστές βαρύτητας, οδηγούν σε κάποιο Πανεπιστημιακό Τμήμα.

Αποτελεί κοινό τόπο σε όλους όσοι ασχολούνται με τα εκπαιδευτικά ζητήματα στη χώρα μας ότι η σχολική αποτυχία αποτελεί τον πυρήνα της Κοινωνιολογίας της Εκπαίδευσης, της επιστήμης δηλαδή που σύμφωνα με την American Sociological Association (Ellwood, 1927) εξετάζει τους εκπαιδευτικούς θεσμούς μέσα στις κοινωνίες. Η παρούσα εργασία δεν έρχεται φυσικά να ασχοληθεί με την Κοινωνιολογία της Εκπαίδευσης σε θεωρητικό επίπεδο. Μας απασχολεί, όμως, σε γενικές γραμμές η εξέλιξη των κυρίαρχων τάσεων στην κοινωνιολογία της εκπαίδευσης ως επιστημονικού παραδείγματος στο βαθμό που αυτή έχει βιβλιογραφικά ασχοληθεί με το θέμα της εργασίας μας. Οι κυρίαρχες τάσεις στην Κοινωνιολογία της Εκπαίδευσης μπορούν να φανούν, σύμφωνα με τον Ramirez (2006), από τα περιεχόμενα των συλλογικών τόμων του Βρετανού κοινωνιολόγου και στοχαστή Albert Henry Halsey.

Στο πρώτο του βιβλίο τη δεκαετία του 1960 ο Halsey (1961) επιλέγει κείμενα με βάση την δομολειτουργική σχολή και την θεωρία του ανθρώπινου κεφαλαίου. Στο δεύτερο βιβλίο του με τίτλο *Power, Ideology and Education* ο ίδιος συγγραφέας παραθέτει κείμενα από μια μαρξιστική οπτική και στο πλαίσιο της κοινωνιολογίας των συγκρούσεων (Karabel & Halsey, 1977). Στο τρίτο βιβλίο, το οποίο εκδόθηκε στο τέλος της δεκαετία του 1990, ο Halsey και οι συνεργάτες του (Halsey, Lauder, Brown & Wells, 1997) συλλέγουν κείμενα στα οποία είναι φανερή η μεταμοντέρνα στροφή στην Κοινωνιολογία, καθώς και οι ειδικές επιστημολογίες σε σχέση με το φύλο και τις τοπικές κουλτούρες.

Στη χώρα μας το ζήτημα των ανισοτήτων στην επίδοση των μαθητών και το ζήτημα αναπαραγωγής των κοινωνικών τάξεων μέσα από την τριτοβάθμια εκπαίδευση

υπήρξαν κυρίαρχα στη διαδικασία της ανάπτυξης και διδασκαλίας της Κοινωνιολογίας της Εκπαίδευσης ως αυτόνομου διδακτικού αντικειμένου. Η μελέτη της πρόσβασης στην τριτοβάθμια εκπαίδευση στην Ελλάδα διαφοροποιείται σημαντικά από την αντίστοιχη σε άλλες χώρες, παρά το γεγονός ότι χρησιμοποιεί παρόμοια εργαλεία. Υπογραμμίζεται ότι, καθώς δεν υφίστανται στη χώρα ιδιωτικά πανεπιστήμια, η εισαγωγή στην πανεπιστημιακή και τεχνολογική ανώτατη εκπαίδευση γίνεται αμιγώς με ακαδημαϊκά κριτήρια. Στο τέλος της τελευταίας τάξης του λυκείου, δηλαδή του ανώτερου κύκλου της δευτεροβάθμιας εκπαίδευσης, διενεργούνται στη χώρα μας γενικές εξετάσεις οι οποίες, με διαφορετικούς κάθε φορά διαχωρισμούς σε κλάδους, δέσμες, κατευθύνεις, προσανατολισμούς κλπ. και διαφορετικούς κάθε τόσο συντελεστές βαρύτητας, οδηγούν σε κάποιο πανεπιστημιακό Τμήμα. Είναι άλλωστε γνωστή η ανάλυση του Λαμπριανίδη (1993) για τη δημιουργία περιφερειακών πανεπιστημίων στην Ελλάδα ως μια προσπάθεια περιφερειακής ανάπτυξης η οποία συνακόλουθα οδήγησε στην μεγάλη διασπορά πανεπιστημιακών τμημάτων σε σημαντικό τμήμα της επικράτειας της χώρας όπως και στον πολλαπλασιασμό των ιδρυμάτων τριτοβάθμιας εκπαίδευσης.

Στο πλαίσιο αυτό η εξήγηση των εκπαιδευτικών ανισοτήτων έχει επικεντρωθεί κυρίως στην δευτεροβάθμια εκπαίδευση. Η μελέτη της κοινωνικής αναπαραγωγής, αντίθετα, έχει επικεντρωθεί στην τριτοβάθμια εκπαίδευση. Στο ενδιάμεσο διάστημα μπορεί να βρει κανείς εργασίες σε σχέση με την μετάβαση από τη μία βαθμίδα στην άλλη. Τέτοιες εργασίες είναι αυτές της Γίτσας Κοντογιαννοπούλου-Πολυδωρίδη (1985, 1987, 1996), η εργασία της Σιάνου-Κύργιου (2006), καθώς και άλλες εργασίες της ίδιας συγγραφέως (Sianou-Kyrgiou, 2008), καθώς και η εργασία των Chrysakis, Balourdos και Capella (2009). Στο ίδιο πεδίο κατηγοριοποιούμε και εργασίες που αφορούν τη σχολική διαρροή, όπως είναι αυτή των Ρουσέα και Βρετάκου (2006) και σχετικά προσφάτως αυτή του Κυρίδη και των συνεργατών του (Kyridis, Tsakiridou, Zagkos, Koutouzis, & Tziamtzi, 2011). Τέλος, πέρα από την κοινωνιολογική οπτική, διάφοροι μελετητές όπως οι Zimdars και Sabbagh (2013), οι Meyer, St. John, Chankseliani και Uribe (2013), οι Κασσωτάκης και Παπαγγελή-Βουλιουρή (1996), καθώς και ο Ματθαίου (2009) γράφουν για τα ζητήματα της πρόσβασης στην τριτοβάθμια εκπαίδευση από μια περισσότερο παιδαγωγική και συγκριτική σκοπιά.

Σε σχέση με την περιοχή της μελέτης της κοινωνικής αναπαραγωγής, οι Μεϊμάρης και Νικολακόπουλος (1978) αναλύουν την κοινωνικοοικονομική προέλευση των φοιτητών στην τριτοβάθμια εκπαίδευση με τη βοήθεια της στατιστικής μεθόδου Path Analysis, μιας μεθόδου που ασχολείται με τη χαρτογράφηση αιτιωδών σχέσεων ανάμεσα σε εμφανείς και λανθάνουσες μεταβλητές (Wright, 1934). Οι Μεϊμάρης και Νικολακόπουλος (1978) μελέτησαν προφανώς τα σχετικά στατιστικά μοντέλα της κοινωνιολογικής «Σχολής του Wisconsin» (Alexander, Eckland, & Griffin, 1975), τα οποία αναπτύσσονταν στις Η.Π.Α. κατά τη δεκαετία του 1970 στο πλαίσιο μιας μαθηματικοποιημένης κοινωνιολογίας. Πράγματι, τη δεκαετία αυτή κυριαρχούσε στις σχετικές μελέτες το μοντέλο των Blau και Duncan (1967). Μάλιστα, ένας από τους πρωτεργάτες στη χρήση και στον εμπλουτισμό των εν λόγω μοντέλων ήταν και ο κοινωνιολόγος William Sewell (Sewell, Halle, & Ohlendorf, 1979, Sewell & Shan,

1967, Sewell, 1971). Για το θέμα της κοινωνικής διαστρωμάτωσης σε σχέση με την τριτοβάθμια εκπαίδευση στη χώρα μας έχουν γράψει επίσης οι Χρυσάκης και Μπαλούρδος (2007) και παλαιότερα ο Μυλωνάς (1982), ένας από τους πρώτους κοινωνιολόγους της εκπαίδευσης στην ελληνική βιβλιογραφία.

Άλλες εργασίες και βιβλία στη χώρα μας, τα οποία μπορούν να χρησιμεύσουν ως βασική κοινωνιολογική βιβλιογραφία για την κατανόηση των εκπαιδευτικών ανισοτήτων, είναι, ανάμεσα σε άλλα, τα βιβλία των Τομπαΐδη (1982), Τζάνη (1983) - που αφορά τη σχέση της σχολικής επιτυχίας με την ταξική προέλευση - και Φραγκουδάκη (1985), καθώς επίσης το βιβλίο της Χαρίτου (2011) που εμπεριέχει μια περισσότερο μαρξιστική οπτική στο θέμα που εξετάζουμε.

Στην παρούσα εργασία πραγματοποιήθηκε μια εκτεταμένη ανάλυση των αποτελεσμάτων των πανελληνίων εξετάσεων για το σχολικό έτος 2012-2013. Την βάση για την ανάλυση αυτή αποτέλεσαν τα αποτελέσματα 81000 μαθητών από όλη την ελληνική επικράτεια. Η ανάλυση αυτή δεν ήταν μονοσήμαντη καθώς, πραγματοποιήθηκε, αφενός, ποιοτική ανάλυση των αποτελεσμάτων και αφετέρου πραγματοποιήθηκε χωρική ανάλυση αυτών. Για την χωρική ανάλυση των αποτελεσμάτων των 81000 μαθητών έγινε αναγωγή σε επίπεδο Καλλικρατικού Δήμου. Το ζητούμενο στην εργασία ήταν κατά πόσο δημιουργούνται συστάδες (clusters) Δήμων με όμοια ή αντίθετα χαρακτηριστικά ως προς την απόδοση των μαθητών. Με σύχρονες μεθόδους και λογισμικά χωρικής ανάλυσης ανιχνεύθηκαν διάφορα πρότυπα όσον αφορά στην επίδοση των μαθητών στις πανελλήνιες εξετάσεις για το σχολικό έτος 2012-2013.

2. ΔΕΔΟΜΕΝΑ - ΜΕΘΟΔΟΛΟΓΙΑ

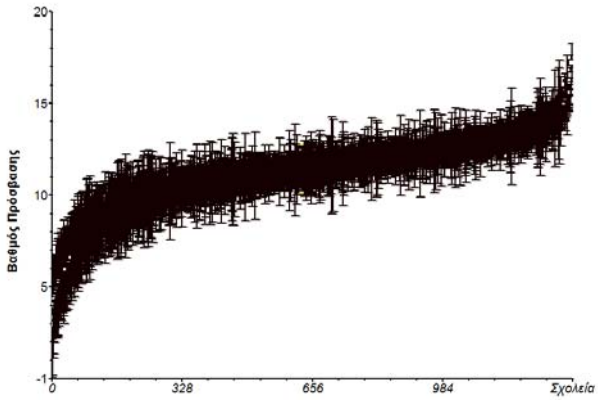
Για τις ανάγκες της εργασίας δημιουργήθηκε μια χωρική βάση δεδομένων από τις επιδόσεις 81000 περίπου μαθητών (1300 Λύκεια) της Γ' Λυκείου του εκπαιδευτικού έτους 2012-2013 (Πίνακας 1). Πραγματοποιήθηκε πρώτα αναγωγή της επίδοσης στις πανελλήνιες εξετάσεις σε επίπεδο σχολείου και κατόπιν σε επίπεδο Καλλικρατικού Δήμου. Στη συνέχεια πραγματοποιήθηκε αποτίμηση των χωρικών προτύπων της επίδοσης των μαθητών της Γ' Λυκείου, προκειμένου να μελετηθεί η χωρική μεταβλητότητα σε τοπικό και υπερτοπικό επίπεδο. Πραγματοποιείται δηλαδή μια χαρτογραφική απόδοση της επίδοσης (χαμηλή, υψηλή και εισαγωγή στην Τριτοβάθμια Εκπαίδευση) των μαθητών σε επίπεδο Δήμου, καθώς και εφαρμογή δεικτών αυτοσυσχέτισης. Τέλος, πραγματοποιήθηκε μια αρχική διερεύνηση της σχέσης της σχολικής επίδοσης με τον δείκτη εκπαίδευσης που αποτελεί μια έμμεση ένδειξη της επίδρασης των κοινωνικο-οικονομικών παραγόντων, οι οποίοι εκτιμάται ότι την επηρεάζουν.

Ο επόμενος πίνακας παρουσιάζει τα περιγραφικά στατιστικά στοιχεία της βάσης των δεδομένων των μαθητών.

Πίνακας 1. Εισαγωγικές εξετάσεις 2012-2013

Μέσος όρος	11,78
Διάμεσος	12,1
Δεσπόζουσα τιμή	15,4
Τυπική απόκλιση	4,31
Διακύμανση	18,6
Εύρος	19,7
Ελάχιστο	0,00
Μέγιστο	19,7
Λοξότητα	-0,22
Κύρτωση	-0,96

Η κατανομή των βαθμών για όλους τους μαθητές που έλαβαν μέρος στις εξετάσεις του 2013 με τα αντίστοιχα διαστήματα μ± 2σ (Εικόνα 1).



Εικόνα 1. Πληθυσμός. Κατανομή των βαθμών πρόσβασης των εξετάσεων του 2013 με τα σχολεία σε αύξουσα σειρά ανάλογα με τον μέσο όρο βαθμολογίας των μαθητών τους (N1= 81.010 μαθητές, N2=1.310 σχολεία)

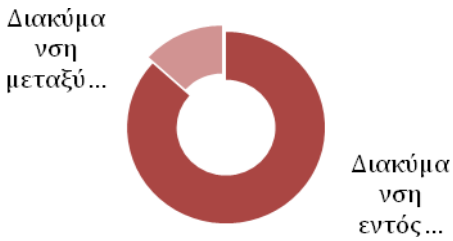
Η σκούρα σιγμοειδής λωρίδα δημιουργείται από χίλιες τριακόσιες δέκα κάθετες γραμμές, μία για κάθε λύκειο. Στο μέσον κάθε γραμμής, αν και δεν διακρίνεται λόγου της πυκνότητας των γραμμών, βρίσκεται ο μέσος όρος του κάθε σχολείου. Οι γραμμές αναπαριστούν τα διαστήματα μ± 2σ από τους αντίστοιχους μέσους. Το σχολείο με τον υψηλότερο μέσο όρο βαθμών πρόσβασης στις εξετάσεις του 2013 (πάνω δεξιά στις σκούρες κάθετες γραμμές του Σχήματος 3) ήταν το 2ο Λύκειο Αρσάκειου Ψυχικού, οι μαθητές του οποίου είχαν μέσο όρο πρόσβασης 17,38.

Πολλά άλλα μεγάλα ιδιωτικά λύκεια σαν τα Αρσάκεια Ψυχικού, καθώς και άλλα λιγότερο γνωστά σε περιοχές όπως το Κορωπί και το Χαϊδάρι είχαν υψηλούς μέσους όρους στις κατανομές των βαθμών πρόσβασης. Στον αντίποδα, όλα τα εσπερινά

δημόσια σχολεία είχαν τους χαμηλότατους βαθμούς πρόσβασης ανεξάρτητα από περιοχή, ενώ στα ημερήσια λύκεια χαμηλές βαθμολογίες σημειώθηκαν σε περιοχές όπως το Καστελόριζο, η Πάτμος, η Νίσυρος και η Ξάνθη.

Οι μέσοι όροι, όμως, πολλές φορές αποκρύπτουν την πραγματική εικόνα της κατανομής των βαθμών πρόσβασης. Για παράδειγμα, από τους λίγους υποψήφιους στις λυκειακές τάξεις μιας νησιωτικής περιοχής ο μοναδικός απόφοιτος του 2013 πέτυχε βαθμό πρόσβασης 7,97. Οι υπόλοιποι ελάχιστοι υποψήφιοι ήταν απόφοιτοι παλαιότερων ετών και κανένας από αυτούς δεν είχε βαθμό πρόσβασης μεγαλύτερο από τη μονάδα. Ο μέσος όρος για τους υποψήφιους στο εν λόγω σχολείο εμφανίζεται να είναι εξαιρετικά μικρός (ίσος με 1,89). Στο σημείο αυτό η διακύμανση της κατανομής, ίση με 11,69, δίνει πιο πιστή εικόνα της πραγματικότητας.

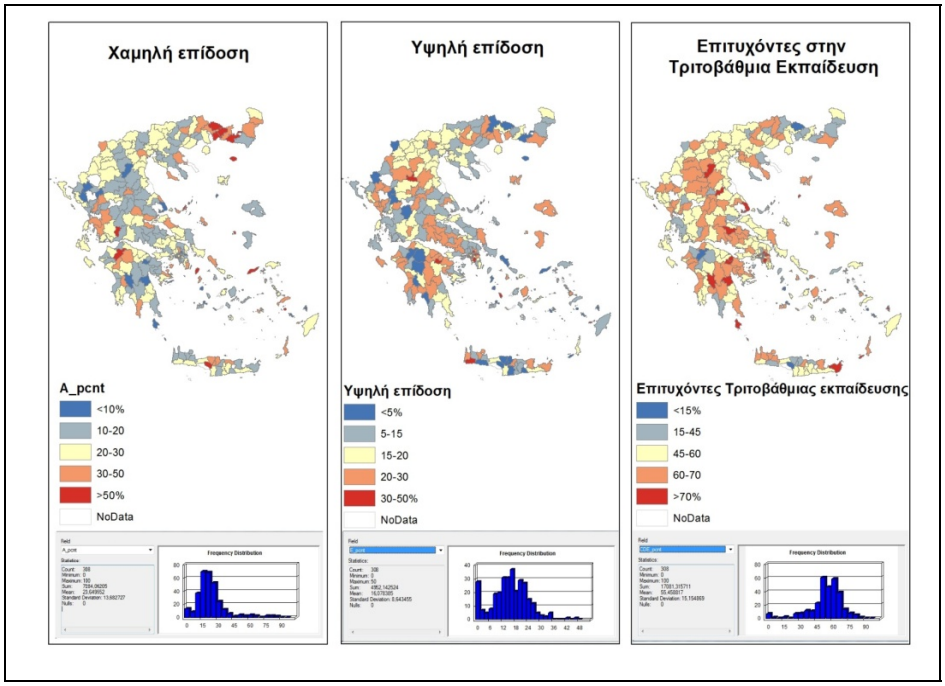
Η συνολική διακύμανση στη βαθμολογία των μαθητών είναι 18,29. Από αυτήν ένα ποσοστό της τάξης του 93 τοις εκατό (17,03) αποδίδεται στις διαφορές μεταξύ μαθητών εντός των σχολείων. Μόνο το 7 τοις εκατό της συνολικής διακύμανσης (1,26) ανιχνεύεται μεταξύ σχολείων. Ο συντελεστής συνάφειας μεταξύ των βαθμών των μαθητών σε κάθε σχολείο είναι 1,26/(1,26+17,03), δηλαδή περίπου ίσος με 0,07. Αυτός είναι ο «ενδοσχολικός συντελεστής συνάφειας». Δηλαδή το ποσοστό της μαθητικής επίδοσης που εξηγείται στατιστικά από τον παράγοντα «Σχολείο». Με άλλα λόγια η επίδραση του σχολείου. Πράγματι, στη διεθνή βιβλιογραφία το ποσοστό αυτό είναι το λεγόμενο “school effect”(Creemers, Kyriakides, & Sammons, 2010).



Εικόνα 2. Η διακύμανση σε μαθητές και σχολεία στους βαθμούς πρόσβασης για το έτος 2013

Το πρώτο βήμα, όπως προειπώθηκε, αφορά στη δημιουργία της χωρικής βάσης δεδομένων, καθώς και τη σύνδεση των χωρικών οντοτήτων (Καλλικρατικοί Δήμοι) με τα δεδομένα της επίδοσης. Τα αποτελέσματα της επίδοσης στην αρχική τους μορφή είχαν σαν χωρική αναφορά το σχολείο του κάθε μαθητή. Για το λόγο αυτό πραγματοποιήθηκε γεωαναφορά του κάθε σχολείου με τη μέθοδο της γεωκωδικοποίησης. Αυτή η διαδικασία (γεωκωδικοποίηση – geocoding) αναφέρεται στην εύρεση των γεωγραφικών συντεταγμένων με βάση τη γεωγραφική αναφορά στα πινακοποιημένα δεδομένα (Χαλκιάς, 2006). Μετά τη διαδικασία της γεωκωδικοποιήσης των σχολείων πραγματοποιήθηκε αναγωγή των σχολείων σε επίπεδο Καλλικρατικού Δήμου.

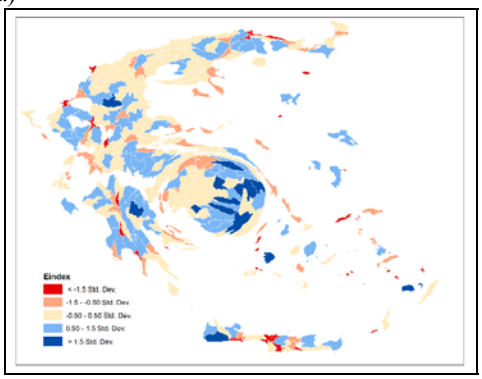
Η επόμενη εικόνα παρουσιάζει την γεωγραφία της επίδοσης των μαθητών στις πανελλήνιες εξετάσεις για το σχολικό έτος 2012-2013 σε επίπεδο Καλλικρατικού Δήμου.



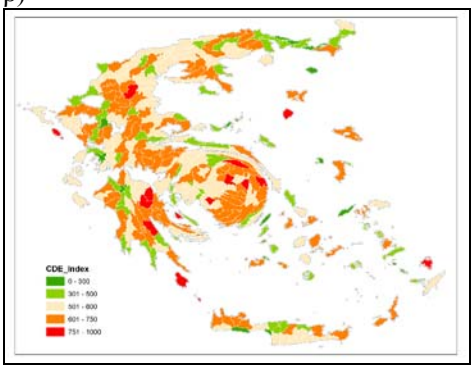
Εικόνα 3. Η γεωγραφία της επίδοσης των μαθητών στις πανελλήνιες εξετάσεις (2012-2013)

Η επόμενη εικόνα παρουσιάζει τα χαρτογράμματα της πολύ υψηλής επίδοσης, καθώς και την επιτυχία στην τριτοβάθμια εκπαίδευση. Τα χαρτογράμματα αυτά παρουσιάζουν την επίδοση των μαθητών, ενώ παράλληλα παρουσιάζεται και το πλήθος των μαθητών (ανάλογα με την παραμόρφωση του χάρτη) σε κάθε καλλικρατικό δήμο της χώρας.

α)



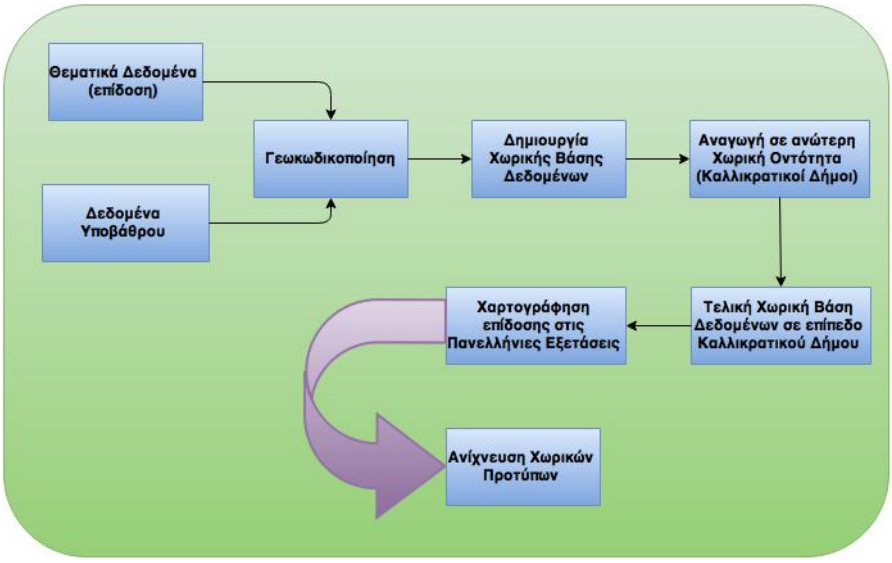
β)



Εικόνα 4. Χαρτογράμματα. α) Πολύ υψηλές επιδόσεις (επιτυχία σε τμήματα με υψηλές βάσεις εισαγωγής), β) Υψηλές επιδόσεις (επιτυχία σε ΑΕΙ ή ΤΕΙ)

Οι παραπάνω χάρτες αποτελούν μια πρώτη ένδειξη της σημαντικής γεωγραφικής ετερογένειας των δήμων που απαρτίζουν τον ελλαδικό χώρο. Για την ανάλυση αυτής της ετερογένειας πραγματοποιήθηκε αρχικά η ανάλυση αυτοσυσχέτισης για όλους τους δήμους, χρησιμοποιώντας τον ολικό χωρικό δείκτη Moran's I (Global Moran's I). Για την ανάλυση αυτή χρησιμοποιήθηκαν οι βαθμοί για τις εξής κατηγορίες: α) Χαμηλή επίδοση, β) Υψηλή επίδοση, γ) Επιτυχία στην Τριτοβάθμια εκπαίδευση. Από την εφαρμογή του Global Moran's I προέκυψε πως η χαμηλή επίδοση των μαθητών είναι ελαφρώς προτυποποιημένη, ενώ, αντίθετα, η υψηλή επίδοση καθώς και η επιτυχία στην Τριτοβάθμια εκπαίδευση παρουσιάζουν υψηλές συστάδες.

Το εννοιολογικό μοντέλο της μεθοδολογίας παρουσιάζεται στην επόμενη εικόνα.



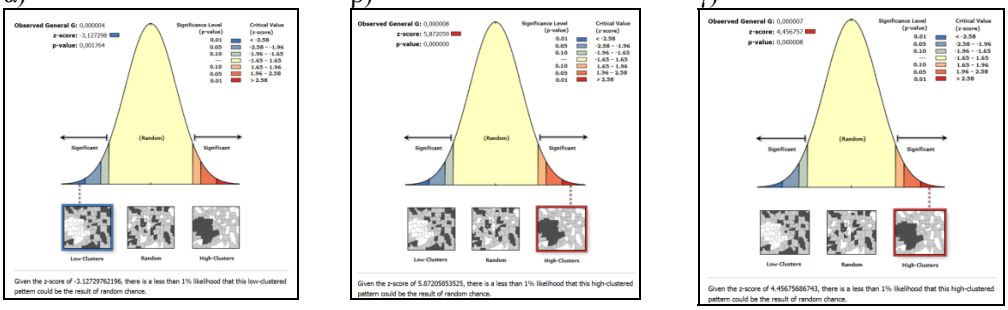
Εικόνα 5. Εννοιολογικό διάγραμμα μεθοδολογίας

3. ΑΠΟΤΕΛΕΣΜΑΤΑ

3.1 Ολική χωρική Αυτοσυσχέτιση (Global Moran's I)

Τα αποτελέσματα της εφαρμογής του δείκτη Ολικής Χωρικής Αυτοσυσχέτισης διακρίνονται στην επόμενη εικόνα (Εικόνα 6). Διακρίνονται τα πρότυπα τα οποία δημιουργούνται με βάση την επίδοση των μαθητών. Η προτυποποίηση αυτή πραγματοποιήθηκε με το ελεύθερο λογισμικό GEODA 1.6, καθώς και το εμπορικό λογισμικό ArcGIS 10.2.

α) β) γ)

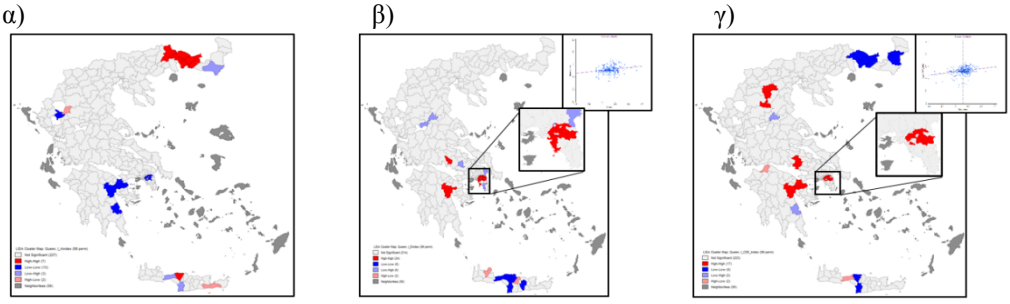


Εικόνα 6. Ανάλυση αυτοσυσχέτισης (Global Moran's I). α) Χαμηλή επίδοση, β) Υψηλή επίδοση, γ) Επιτυχία στην Τριτοβάθμια εκπαίδευση

Στη συνέχεια έλαβε χώρα ο υπολογισμός των τοπικών δεικτών χωρικής αυτοσυσχέτισης (Local Indicators of Spatial Autocorrelation). Συγκεκριμένα, υπολογίστηκε ο δείκτης ανίχνευσης χωρικών προτύπων LISA (Anselin, 1995). Με τον δείκτη αυτό εξετάστηκε ο βαθμός κατά τον οποίο υψηλές τιμές των τριών κατηγοριών που επιλέχθηκαν (α) Χαμηλή επίδοση, β) Υψηλή επίδοση, γ) Επιτυχία στην Τριτοβάθμια εκπαίδευση) σε ένα δήμο συσχετίζονται με υψηλές τιμές σε γειτονικούς δήμους διαμορφώνοντας με αυτόν τον τρόπο ένα χωρικό πρότυπο υψηλές-υψηλές τιμές (high-high pattern) κοκ.

Τα αποτελέσματα για τις εξής κατηγορίες παρουσιάζονται στην εικόνα που ακολουθεί (Εικόνα 7):

α) β) γ)



Εικόνα 7. Χωρικά πρότυπα (LISA Clustering). α) Χαμηλή επίδοση, β) Υψηλή επίδοση, γ) Επιτυχία στην Τριτοβάθμια εκπαίδευση

Όπως διακρίνεται, για την χαμηλή επίδοση παρουσιάζεται ένα συσσωμάτωμα όπου δήμοι με υψηλά ποσοστά χαμηλής επίδοσης περιβάλλονται από Δήμους με όμοια χαρακτηριστικά (Δ. Παρανεστίου, Ξάνθης, Μύκης, Αβδήρων. Ιάσμου & Κομοτηνής), για την υψηλή επίδοση παρουσιάζεται ένα συσσωμάτωμα όπου δήμοι του Μητροπολιτικού κέντρου της Αθήνας με υψηλά ποσοστά υψηλής επίδοσης περιβάλλονται από Δήμους με όμοια χαρακτηριστικά και για την επιτυχία στην Τριτοβάθμια εκπαίδευση παρουσιάζεται ένα συσσωμάτωμα γύρω από τον Δήμο Αθηναίων, όπου δήμοι (βόρεια του Δήμου Αθηναίων) με υψηλά ποσοστά επιτυχίας στην Τριτοβάθμια εκπαίδευση περιβάλλονται από Δήμους με όμοια χαρακτηριστικά.

3.2 Γραμμική παλινδρόμηση

Στη συνέχεια, επιχειρήθηκε μια πρώτη προσέγγιση συσχέτισης της επίδοσης με κοινωνικά δεδομένα. Για κάθε Καλλικρατικό δήμο δημιουργήθηκε ένας δείκτης εκπαίδευσης (αναλφάβητοι/1000 κατοίκους) και διερευνήθηκε η σχέση μεταξύ της επίδοσης των μαθητών με αυτή τη μεταβλητή. Εκτιμήθηκε μια αρνητική συσχέτιση μεταξύ της επιτυχίας στην Τριτοβάθμια εκπαίδευση με τον δείκτη αναλφαβητισμού που αποτελεί μια ένδειξη του χαμηλού κοινωνικοοικονομικού στάτους των περιοχών αυτών. Τα αποτελέσματα από αυτή την παλινδρόμηση παρουσιάζονται στον επόμενο πίνακα (Πίνακας 2). Όπως διακρίνεται το ποσοστό πρόβλεψης του συγκεκριμένου μοντέλου είναι 55% περίπου (R^2=0.5497).

Πίνακας 2. Αποτελέσματα παλινδρόμησης

R^2	0,55			
Adjusted R^2	0,55			
Variable	**Coefficient**	**Std. Error**	**t-Statistic**	**Probability**
Constant	6.077.045	1.990.959	3.052.321	0.000
CDE_Index	-0,67	0,0346	-1.932.847	0.000

4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Τα υψηλά ποσοστά χαμηλής επίδοσης είναι πιο συχνά σε ορεινές, απομακρυσμένες και συνοριακές περιοχές, γεγονός που καταδεικνύει τη σημασία της έλλειψης υποδομών και υπηρεσιών για την προετοιμασία των μαθητών της τρίτης Λυκείου. Επιπρόσθετα, τα χαμηλά ποσοστά υψηλής επίδοσης παρουσιάζουν μια παρόμοια εικόνα, καθώς η μειωμένη επιτυχία συνδέεται με τις χωρικές ανισότητες στην πρόσβαση στις υπηρεσίες υποστήριξης των μαθητών για την επιτυχία τους στην τριτοβάθμια εκπαίδευση. Όμως υπάρχουν και περιοχές που συνδυάζουν μεγάλα ποσοστά χαμηλών επιδόσεων και μικρά ποσοστά υψηλών επιδόσεων, γεγονός που τις κατατάσσει στις πλήρως περιθωριοποιημένες περιοχές, στις οποίες η επιτυχία πρόσβασης στην τριτοβάθμια εκπαίδευση αποτελεί μια αμυδρή πιθανότητα. Σε αυτές ακριβώς τις περιοχές σωρεύονται πολλαπλές μειονεξίες οι περισσότερες από τις οποίες συνδέονται με έντονες κοινωνικοοικονομικές διακρίσεις και εισοδηματικές

ανισότητες. Το υψηλό ποσοστό αναλφάβητων αποτελεί μια ένδειξη του χαμηλού κοινωνικοοικονομικού και εισοδηματικού επιπέδου συγκεκριμένων περιοχών.

Αντίθετα, οι περιοχές που συνδυάζουν μεγάλα ποσοστά υψηλών επιδόσεων με μικρά ποσοστά χαμηλών επιδόσεων είναι βασικά αστικές περιοχές όπου συγκεντρώνονται πληθυσμοί που κατέχουν υψηλό εκπαιδευτικό επίπεδο, διαθέτουν υψηλότερο εισόδημα και έχουν τη δυνατότητα πρόσβασης σε εκπαιδευτικές υποδομές και υπηρεσίες προκειμένου να επιτύχουν για τα παιδιά τους την ευρεία πρόσβαση στην τριτοβάθμια εκπαίδευση της χώρας.

Η ερευνητική ομάδα στοχεύει στο μέλλον να ασχοληθεί με την διερεύνηση της σχέσης των χαμηλών επιδόσεων με κοινωνικοοικονομικούς δείκτες που απεικονίζουν το ειδικό βάρος διαφορετικών οικονομικών κλάδων ανά περιοχή, των επιπτώσεων από την ύπαρξη εκπαιδευτικών υποδομών για την διασφάλιση της επιτυχούς πρόσβασης στην τριτοβάθμια εκπαίδευση κοντά στον τόπο κατοικίας των μαθητών, των κριτηρίων διαβάθμισης της πρόσβασης των μαθητών σε περιφερειακά ή κεντρικά πανεπιστήμια, καθώς επίσης και τη διαχρονική διακύμανση των επιδόσεων των μαθητών.

ABSTRACT

This paper analyzes the geographical characteristics of the performance of secondary school students of Greece, in the attempt to enter in the higher education system of the country. For the purposes of the paper, a spatial database (referring to 81010 students of the last class of the secondary education) was created for the academic year 2012-2013. The first step of our work was the aggregation of the data in two levels: at school level and then at municipality level. Next, we assessed the spatial variability of the performance of the students in the final exams at local and regional level with the usage of spatial autocorrelation indices. Finally, an initial investigation of the relationship between school performance and the education index, which is a proxy for the impact of socio-economic factors, was carried out.

ΑΝΑΦΟΡΕΣ

Alexander, K., Eckland, B., & Griffin, L. (1975). The Wisconsin model of socioeconomic achievement: A replication. *American Journal of Sociology*, 81(2), 324–342. Retrieved from http://www.jstor.org/stable/2777380.

Anselin L., (1995). Local Indicators of Spatial Association– LISA*, Geographical Analysis,* 27, 93-115.

Blau, P., & Duncan, O. (1967). *The American occupational structure*. New York, NY: Wiley and Sons.

Chrysakis, M., Balourdos, D., & Capella, A. (2009). Inequalities in access to tertiary education in Greece: an approach based on the official statistics (1984 - 2004). Athens: *National Center for Social Research of Social Policy*.

Creemers, B., Kyriakides, L., & Sammons, P. (2010). *Methodological Advances in Educational Effectiveness Research*. Taylor & Francis.

Ellwood, C. (1927). What is Educational Sociology? *Journal of Educational Sociology*, 1(1), 25–30.

Halsey, A. H. (1961*). Education, Economy, and Society: A Reader in the Sociology of Education*. Free Press of Glencoe.

Halsey, A. H., Lauder, H., Brown, P., Wells A. S. (1997). *Education: Culture, Economy, and Society*. New York: Oxford University Press, Inc.

Karabel, J.. Halsey A. H. (1977). *Power and Ideology in Education*. New York: Oxford University Press, Inc.

Kyridis, A., Tsakiridou, H., Zagkos, C., Koutouzis, M., & Tziamtzi, C. (2011). Educational Inequalities and School Dropout in Greece. *International Journal of Education*, 3(2), 1–15. doi:10.5296/ije.v3i2.855.

Meyer, H.-D., St. John, E., Chankseliani, M., & Uribe, L. (2013). *Fairness in Access to Higher Education in a Global Perspective*. (H.-D. Meyer, E. P. St. John, M. Chankseliani, & L. Uribe, Eds.). Rotterdam: Sense Publishers. doi:10.1007/978-94-6209-230-3.

Ramirez, F. (2006). Beyond Achievement and Attainment Studies: Revitalizing A Comparative Sociology of Education. *Comparative Education* 42, 1-19.

Sewell, W. (1971). Inequality of opportunity for Higher Education. *American Sociological Review*, 36(5), 793–809. Retrieved from http://www.ssc.wisc.edu/wlsresearch/publications/files/public/Sewell_Inequality.Opportunity.H.E.pdf

Sewell, W., & Shan, V. (1967). Socioeconomic status, Intelligence, and the attainment of Higher Education. *Sociology of Education*, 04(1), 1–23. Retrieved from http://www.jstor.org/stable/2112184.

Sewell, W., Halle, A., & Ohlendorf, G. (1979). The educational and early occupational status attainment process: replication and revision. *American Sociological Review*, 35(6), 1014–1027. Retrieved from http://www.ssc.wisc.edu/wlsresearch/publications/files/public/Sewell-Haller-Ohlendorf_The.Educational.and.Early.Occupational.Status.Attainment.Process.pdf

Sianou-Kyrgiou, E. (2008). Social class and access to higher education in Greece: supportive preparation lessons and success in national exams. *International Studies in Sociology of Education*, 18, 173–183.

Wright, S. (1934). *The method of path coefficients. Annals of Mathematical Statistics*, 5(3), 161–215. doi:10.1214/aoms/1177732676.

Zimdars, A., & Sabbagh, D. (2013). Call for Papers Special Issue in Volume 57 Access to Higher Education : "Fairness" in Comparative Perspective. *Comparative Education Review*, 57(3), 492–493. Retrieved from http://www.jstor.org/stable/10.1086/660852.

Κασσωτάκης, Μ., & Παπαγγελή-Βουλιουρή, Δ. (1996). *Η πρόσβαση στην ελληνική τριτοβάθμια εκπαίδευση*. Αθήνα: Γρηγόρη.

Κοντογιαννοπούλου-Πολυδωρίδη, Γ. (1985). *Η αξιολόγηση των εξετάσεων για την επιλογή στην τριτοβάθμια εκπαίδευση: εκπαιδευτικοί και κοινωνικοί παράγοντες που επηρεάζουν την επίδοση των υποψηφίων και των επιτυχόντων*. Ερευνητική Έκθεση αρ. 81056. Αθήνα.

Κοντογιαννοπούλου-Πολυδωρίδη, Γ. (1987). Εκπαιδευτικοί και κοινωνικοί παράγοντες που επηρεάζουν την επίδοση των υποψηφίων και των επιτυχόντων.Στο Ι. Λαμπίρη-Δημάκη (Επιμ.), *Η κοινωνιολογία στην Ελλάδα σήμερα*. Αθήνα: Παπαζήση.

Κοντογιαννοπούλου-Πολυδωρίδη, Γ. (1996). *Κοινωνιολογική ανάλυση της αξιολόγησης και της επίδοσης: Οι εισαγωγικές εξετάσεις*. Αθήνα: Gutenberg.

Λαμπριανίδης, Λ. (1993). *Περιφερειακά πανεπιστήμια στην Ελλάδα, Από το αίτημα για στρατόπεδα νεοσυλλέκτων στο αίτημα για περιφερειακά πανεπιστήμια*. Αθήνα: Παρατηρητής.

Ματθαίου, Δ. (2009). *Πολιτικές πρόσβασης στην Τριτοβάθμια Εκπαίδευση. Μαθήματα και παθήματα από τη διεθνή εμπειρία. Στο ΠΟΣΔΕΠ – ΑΕΙ, Πρακτικά επιστημονικής συνάντησης "Η αναβάθμιση του Λυκείου και τα συστήματα πρόσβασης στην Ανώτατη Εκπαίδευση: οι αναγκαίες αλλαγές"* (σελ. 19–26). Αθήνα: Πλέθρον.

Μεϊμάρης, Μ. Νικολακόπουλος, Η. (1978). Παραγοντική ανάλυση δεδομένων: σχέσεις κοινωνικοοικονομικής προέλευσης και σχολικής φοίτησης για τους σπουδαστές των ΑΕΙ. *Επιθεώρηση Κοινωνικών Ερευνών*, (33-34), 225–240.

Μυλωνάς, Θ. (1982). *Η αναπαραγωγή των κοινωνικών τάξεων μέσα από τους σχολικούς μηχανισμούς*. Αθήνα: Γρηγόρη.

Ρουσέας, Π., Βρετάκου, Β. (2006). *Η μαθητική διαρροή στη δευτεροβάθμια εκπαίδευση*. Υπουργείο Παιδείας, Παιδαγωγικό Ινστιτούτο.

Σιάνου-Κύργιου, Ε. (2006). *Εκπαίδευση και κοινωνικές ανισότητες. Η μετάβαση από τη Δευτεροβάθμια στην Ανώτατη Εκπαίδευση (1997-2004)*. Αθήνα: Μεταίχμιο.

Τζάνη, Μ. (1983). *Σχολική επιτυχία: Ζήτημα ταξικής προέλευσης και κουλτούρας*. Αθήνα: Γρηγόρη.

Τομπαΐδης, Δ. (1982). *Η ισότητα ευκαιριών στην εκπαίδευση: συμβολή στη μελέτη εκδημοκρατισμού της εκπαίδευσης*. Αθήνα: Γρηγόρης.

Φραγκουδάκη, Ά. (1985). *Κοινωνιολογία της εκπαίδευσης: Θεωρίες για την κοινωνική ανισότητα στο σχολείο*. Αθήνα: Παπαζήση.

Χαλκιάς, Χ. (2006). Όροι και έννοιες επιστήμης γεωγραφικών πληροφοριών (GIS). Αθήνα: Ίων.

Χαρίτου, Α. (2011). *Γονείς, μαθητές: ταξικοί και επαγγελματικοί προσδιορισμοί*. Αθήνα: Λευκή σελίδα.

Χρυσάκης, Μ., & Μπαλούρδος, Δ. (2007). *Ανισότητες πρόσβασης στην Ανώτατη Εκπαίδευση: Διαφοροποιήσεις σε προπτυχιακό και μεταπτυχιακό επίπεδο. Στο Κοινωνικό Πορτραίτο της Ελλάδας*, 2006. Αθήνα: Ινστιτούτο Κοινωνικής Πολιτικής.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής (2015), σελ. 60-70

ΧΡΗΣΗ ΤΡΙΔΙΑΣΤΑΤΩΝ ΣΥΖΕΥΞΕΩΝ ΓΙΑ ΤΙΣ ΔΙΑΜΕΤΡΟΥΣ ΚΑΙ ΤΟ ΎΨΟΣ ΔΕΝΔΡΩΝ

Δ. Γεράρδη , Γ. Σταματέλλος
Τμήμα Δασολογίας και Φυσικού Περιβάλλοντος, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
despogerardi@yahoo.com, stamatel@for.auth.gr

ΠΕΡΙΛΗΨΗ

Τα δασικά οικοσυστήματα είναι σύνθετα πολυδιάστατα συστήματα των οποίων η δομή παίζει σημαντικό ρόλο στη λειτουργία και ποικιλομορφία των οικοσυστημάτων. Στην παρούσα εργασία, κατασκευάζονται τριδιάστατες συναρτήσεις της από κοινού κατανομής της στηθιαίας διαμέτρου του κορμού, της διαμέτρου του κορμού στο μέσο και του ύψους του κορμού για ένα σύνολο δένδρων από το Πανεπιστημιακό δάσος Περτουλίου, χρησιμοποιώντας τις συζεύξεις (copulas), συναρτήσεις που συνδέουν μια πολυδιάστατη συνάρτηση κατανομής με τις περιθώριες μονοδιάστατες συναρτήσεις. Τα αποτελέσματα της ανάλυσης υποδεικνύουν ότι για τα τριδιάστατα δεδομένα μας, η χρήση τριδίαστατης σύζευξης περιγράφει επαρκώς τη σχέση ανάμεσα στις διαμέτρους και το ύψος και οι εκτιμήσεις του όγκου είναι ικανοποιητικές. Για τη σύζευξη αυτή έγινε έλεγχος καλής προσαρμογής στα δεδομένα. Προτείνουμε ακόμη τη χρήση μιας μη παραμετρικής εκτιμήτριας που να ορίζει σύζευξη καλά προσαρμοσμένη στα δεδομένα. Κατά τη μη παραμετρική εκτίμηση των εξεταζόμενων συζεύξεων, επιλέγεται να γίνει μη παραμετρικά και η εκτίμηση των τριών περιθώριων συναρτήσεων, προσδίδοντας έτσι μεγαλύτερη γενικότητα στη συνολική εκτίμηση.

Λέξεις κλειδιά: τριδιάστατες συζεύξεις, μη παραμετρική εκτίμηση, εκτιμητής πυρήνα

1. ΕΙΣΑΓΩΓΗ

Η πρόβλεψη είναι ζωτικής σημασίας για τις τεχνικές λήψης αποφάσεων στο σύγχρονο δασικό σχεδιασμό. Μοντέλα κατανομών της διαμέτρου και του ύψους των δέντρων παίζουν σημαντικό ρόλο στη δασική απογραφή καθώς εκτιμάται ο όγκος του δέντρου, και βοηθάει τη λήψη αποφάσεων όσον αφορά στη διαχείριση της ξυλείας. Η ποιότητα της καλής διαχείρισης των δασών βασίζεται στην ορθότητα και την ακρίβεια της αξιολόγησης των δασικών πόρων. Η εκτίμηση της κατανομής της διαμέτρου και του ύψους των δέντρων βελτιώνει την πρόβλεψη του συνολικού όγκου

των δέντρων και της δασικής συστάδας και εκτιμά την αξία της ξυλείας (Schreuder and Hafley 1977). Η κατάρτιση τέτοιων μοντέλων αν και ιδιαίτερα σημαντική, δεν είναι αρκετά διαδεδομένη. Λίγα μοντέλα και μέθοδοι είναι διαθέσιμα για να περιγράψουν την από κοινού συνάρτηση κατανομής της διαμέτρου και του ύψους των δέντρων. Οι Hafley and Schreuder (1976) εισήγαγαν την κατανομή S_{BB} στη δασολογία, για τη μοντελοποίηση της από κοινού συνάρτησης κατανομής της διαμέτρου και του ύψους των δέντρων, και από τότε αυτή χρησιμοποιείται ευρεύως (Zucchini et al. 2001, Li et al. 2002, Wang et al. 2007). Επέκταση αυτής της κατανομής στις τρεις διαστάσεις (στηθιαία διάμετρος D, ύψος H και όγκος V) είναι η S_{BBB} (Schreuder et al. 1982 a,b), η οποία περιγράφει ικανοποιητικά τη δομή της συστάδας. Πρόσφατα οι συζεύξεις (copulas) έχουν γίνει δημοφιλείς στη στατιστική και στην οικονομετρία. Στη δασολογία, Li et al. (2002) και Kershaw et al. (2010) χρησιμοποίησαν τις συζεύξεις για την προσομοίωση χωρικά συσχετιζόμενων συστάδων. Οι Wang et al. (2008) και (2010) εισήγαγαν τη διδιάστατη και τριδιάστατη κανονική σύζευξη για να περιγράψουν την από κοινού κατανομή της διαμέτρου και του ύψους, και της διαμέτρου, του ύψους και του όγκου αντίστοιχα. Οι συζεύξεις είναι ένα ισχυρό εργαλείο που παρέχουν ένα γενικό τρόπο κατασκευής των από κοινού κατανομών.

Στην παρούσα εργασία εισάγουμε μία τριδιάστατη σύζευξη που περιγράφει την από κοινού συνάρτηση κατανομής της στηθιαίας διαμέτρου, της διαμέτρου στο μέσο του κορμού και του ύψους του κορμού του δέντρου και προτείνουμε έναν τριδιάστατο εκτιμητή πυρήνα (kernel estimator) της σύζευξης.

Έστω $X_1, X_2, ..., X_d$ τυχαίες μεταβλητές με από κοινού συνάρτηση κατανομής $H(x_1, x_2, ..., x_d)$ και περιθώριες συναρτήσεις κατανομών $F_1, F_2, ..., F_d$ αντίστοιχα. Τότε, σύμφωνα με το θεώρημα του Sklar (1959) υπάρχει μια σύζευξη $C : [0,1]^d \to [0,1]$ τέτοια ώστε για όλα τα $u_i \in \mathbb{R}^d$ να ισχύει

$$C(u_1, u_2, ..., u_d) = H(F_1^{-1}(u_1), F_2^{-1}(u_2), ..., F_d^{-1}(u_d)).$$

Αν οι $F_1, F_2, ..., F_d$ είναι συνεχείς, τότε η C είναι μοναδική και μπορούμε εύκολα να κατασκευάσουμε τη μοναδική αυτή σύζευξη. Υπάρχει ένα πλούσιο σύνολο από διδιάστατες συζεύξεις που έχουν δημιουργηθεί με τη μέθοδο της αντιστροφής και από άλλες μεθόδους, μεταξύ αυτών η κανονική ή Gaussian σύζευξη, η Farlie-Gumbel-Morgenstern (FGM) καθώς και οι συζεύξεις από την Αρχιμήδεια οικογένεια (π.χ. Clayton, Gumbel, Frank). Από αυτές η κανονική και η FGM μπορούν να επεκταθούν σε περισότερες από δύο διαστάσεις, με έναν απλό τρόπο επιτρέποντας διαφορετική μορφή εξάρτησης ανάμεσα στα ζεύγη των μεταβλητών. Οι Αρχιμήδειες συζεύξεις μπορούν να επεκταθούν σε περισσότερες διαστάσεις αλλά ως επί το πλείστον, μια τέτοια επέκταση δεν επιτρέπει διαφορετική μορφή εξάρτησης μεταξύ των μεταβλητών. Συγκεκριμένα περιορίζει την εξάρτηση μεταξύ όλων των ζευγών των μεταβλητών να είναι η ίδια, ένας περιορισμός που είναι δεσμευτικός και όχι πάντα κατάλληλος. Από τις Αρχιμήδειες συζεύξεις, η Gumbel γενικεύεται με απλό τρόπο, χρησιμοποιώντας τη μέθοδο της αντιστροφής και επιτρέπει διαφορετική μορφή εξάρτησης για τις μεταβλητές της.

Οι Wang et al. (2008, 2010) διαπίστωσαν ότι, ανάμεσα στις συζεύξεις που χρησιμοποιούνται στη δασολογία, η κανονική σύζευξη ταιριάζει καλύτερα στα διδιάστατα δασικά δεδομένα. Επομένως περιμένουμε η κανονική σύζευξη να ταιριάζει καλά και στην τριδιάστατη περίπτωση.
Η εκτίμηση των συζεύξεων μπορεί να επιτευχθεί είτε πλήρως παραμετρικά με την παραδοχή παραμετρικών μοντέλων τόσο για τη σύζευξη όσο και για τις περιθώριες συναρτήσεις κατανομών, είτε ημιπαραμετρικά, είτε μη παραμετρικά. Κατά τη μη παραμετρική εκτίμηση των συζεύξεων, επιλέγεται να γίνει και η εκτίμηση των περιθώριων συναρτήσεων μη παραμετρικά, προσδίδοντας έτσι μεγαλύτερη γενικότητα στην εκτίμηση. Η μη παραμετρική εκτίμηση για πολυδιάστατα δεδομένα είναι μια σημαντική τεχνική που έχει ένα ευρύ φάσμα εφαρμογών. Ωστόσο, έχει τύχει λιγότερης προσοχής από την αντίστοιχη μονοδιάστατη περίπτωση. Αυτό μπορεί να οφείλεται στο ότι καθώς οι διαστάσεις των δεδομένων αυξάνονται, αυξάνεται συγχρόνως και η δυσκολία που υπάρχει για την εξαγωγή ενός βέλτιστου εύρους ζώνης (bandwidth).

2. ΕΚΤΙΜΗΤΗΣ ΠΥΡΗΝΑ

Συνήθως ένας μη παραμετρικός εκτιμητής σύζευξης σχηματίζεται σε δύο στάδια: πρώτα εκτιμώνται οι περιθώριες κατανομές, $\hat{F}_{1n}, \hat{F}_{2n}, \hat{F}_{3n}$ (για την τριδιάστατη περίπτωση), και ακολουθεί η εκτίμηση της σύζευξης με βάση τις εκτιμώμενες περιθώριες. Οι εκτιμήσεις των περιθώριων κατανομών μπορούν να γίνουν με βάση τις σχέσεις (Omelka et. al. (2009))

$$\hat{U}_{1i} = \frac{n}{n+1}\hat{F}_{1n}(x),\ \hat{U}_{2i} = \frac{n}{n+1}\hat{F}_{2n}(x),\ \hat{U}_{3i} = \frac{n}{n+1}\hat{F}_{3n}(x) \tag{1}$$

όπου $\hat{F}_{1n}$, $\hat{F}_{2n}$ και $\hat{F}_{3n}$ είναι οι εμπειρικές συναρτήσεις κατανομών των περιθώριων.

Έτσι με τη βοήθεια της (1), προκύπτει ο τοπικός γραμμικός εκτιμητής πυρήνα της σύζευξης, που δίνεται από τη σχέση

$$\hat{C}(u_1,u_2,u_3) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_1h_2h_3}K\left(\frac{u_1-\hat{U}_{1i}}{h_1}\right)K\left(\frac{u_2-\hat{U}_{2i}}{h_2}\right)K\left(\frac{u_3-\hat{U}_{3i}}{h_3}\right) \tag{2}$$

όπου $K(x) = \int_{-\infty}^{x} k\,du$ είναι το ολοκλήρωμα μιας συμμετρικής και φραγμένης στο διάστημα [-1,1] συνάρτησης πυρήνα *k*. Το εύρος ζώνης (bandwidth) h_i μπορεί να υπολογιστεί, για κάθε μεταβλητή ξεχωριστά από τη σχέση (Scott 1992; Bowman et al. 2004)

$$h_i = \left(\frac{4}{(d+2)n}\right)^{1/(d+4)}\sigma_i,\ \text{για}\ i=1,2,...,d,$$

όπου σ_i η τυπική απόκλιση της i-μεταβλητής και μπορεί να αντικατασταθεί από τη δειγματική εκτιμήτρια σε εφαρμογές. Αυτή η μέθοδος ονομάζεται “normal reference”

και χρησιμοποιείται συχνά στις εφαρμογές εξαιτίας της απουσίας οποιασδήποτε άλλης πρακτικής επιλογής εύρους ζώνης. Υπάρχει και το κριτήριο cross validation το οποίο μπορεί να εφαρμοσθεί για τον υπολογισμό του βέλτιστου εύρους ζώνης για τα πολυδιάστατα δεδομένα, που όμως, συνήθως, περιλαμβάνει αριθμητική βελτιστοποίηση η οποία καθίσταται δυσκολότερη όσο αυξάνονται οι διαστάσεις των δεδομένων (Sain et. al. (1994)).

3. ΕΦΑΡΜΟΓΗ

Στη συνέχεια παρουσιάζουμε εφαρμογή σε ένα δείγμα δένδρων από 4 επιφάνειες συνολικής έκτασης 0.2 εκταρίων του Πανεπιστημιακού Δάσους Περτουλίου. Το δάσος θεωρείται αμιγές, αποτελούμενο κυρίως από υβριδογενή ελάτη, και τα δέντρα είναι ανομήλικα. Δεδομένα του Δάσους του Περτουλίου χρησιμοποιήθηκαν από τον Σταματέλλο (1991, 1995) για την αξιολόγηση δισταδιακών δειγματοληπτικών σχεδίων, από τον Γκιούλο (2007) για την εκτίμηση της κοινής κατανομής της στηθιαίας διαμέτρου και του ύψους των δένδρων και από τους Γεράρδη κ.α. (2012) για τη μη παραμετρική εκτίμηση διδιάστατων συζεύξεων.

Έστω X_1 η στηθιαία διάμετρος του κορμού των δένδρων, X_2 η διάμετρος στο μέσο του κορμού των δένδρων και X_3 το ύψος του. Πρώτα μετασχηματίζουμε τα δεδομένα υπολογίζοντας τις ομοιόμορφες ψευδο-παρατηρήσεις (Omelka et. al. (2009)) με τις συναρτήσεις

$$\hat{U}_{1i} = \frac{n}{n+1}\hat{F}_{1n}(X_{1i}),\quad \hat{U}_{2i} = \frac{n}{n+1}\hat{F}_{2n}(X_{2i}), \hat{U}_{3i} = \frac{n}{n+1}\hat{F}_{3n}(X_{3i}),$$

όπου $\hat{F}_{1n}$, $\hat{F}_{2n}$ και $\hat{F}_{3n}$ είναι οι εμπειρικές συναρτήσεις κατανομών των περιθώριων.

Στη συνέχεια υποθέτουμε ότι οι νέες παρατηρήσεις $\hat{U}_{1i}, \hat{U}_{2i}, \hat{U}_{3i}$ ακολουθούν μια συγκεκριμένη σύζευξη, και ελέγχουμε την υπόθεση με τη δοκιμασία X^2.

Οι Αρχιμήδειες συζεύξεις χρησιμοποιούνται συχνά, λόγω της ευελιξίας τους. Οι συζεύξεις αυτής της οικογένειας έχουν την μορφή

$$C(u_1,u_2,...,u_n) = \phi^{-1}(\phi(u_1)+\phi(u_2)+...+\phi(u_n)),$$

όπου φ, είναι η γεννήτρια της σύζευξης.

Ωστόσο, για n>2 η συμμετρική σύζευξη έχει ως συνέπεια ότι οι συσχετίσεις μεταξύ όλων των ζευγών των μεταβλητών είναι πανομοιότυπες. Αυτό δεν είναι ρεαλιστικό για τις περισσότερες εφαρμογές. Μια μορφή ασύμμετρης σύζευξης σχηματίζεται "φωλιάζοντας" συμμετρικές συζεύξεις (nesting symmetric copulas). Έτσι στην τριδιάστατη περίπτωση έχουμε

$$\begin{aligned} C(u_1,u_2,u_3) &= C_1(C_2(u_1,u_2),u_3) \\ &= \phi_1^{-1}(\phi_1\{\phi_2^{-1}[\phi_2(u_1)+\phi_2(u_2)]+\phi_1(u_3)\}). \end{aligned}$$

Η σχέση των u_1, u_2 περιγράφεται από την σύζευξη C_2 ενώ το αποτέλεσμα αυτής της σχέσης, σχετίζεται με την u_3 μέσω της C_1. Τόσο η C_1 όσο και η C_2 είναι

διδιάστατες συζεύξεις με ϕ_1 και ϕ_2 να είναι αντίστοιχα, οι γεννήτριες τους.

Κατ' αρχήν θα χρησιμοποιήσουμε τη σύζευξη Gumbel and Hougaard (Wong et al. 2008), όπου

$$\begin{aligned} C(u_1,u_2,u_3) &= C_1(C_2(u_1,u_2),u_3) \\ &= \phi_1^{-1}(\phi_1\{\phi_2^{-1}[\phi_2(u_1)+\phi_2(u_2)]+\phi_1(u_3)\}) \\ &= \exp[-\{[(-\ln u_1)^{\theta_2}+(-\ln u_2)^{\theta_2}]^{\theta_1/\theta_2}+(-\ln u_3)^{\theta_1}\}^{1/\theta_1}], \end{aligned}$$

και φ είναι η γεννήτρια της σύζευξης $\varphi(t)=\left(-\ln t\right)^{\theta}$.

Για την εκτίμηση των παραμέτρων των συζεύξεων θα χρησιμοποιήσουμε τη μέθοδο των Genest and Rivest (1993), λύνοντας από τη σχέση που δίνει το συντελεστή συσχέτισης τ του Kendall των μεταβλητών-ζευγών (u_1,u_3) και (u_2,u_3) ως προς τη παράμετρο θ_1 (παίρνοντας ως θ_1 το μέσο των αντίστοιχων τιμών που προέκυψαν από τα ζεύγη (u_1,u_3) και (u_2,u_3)), και λύνοντας από τη σχέση που δίνει το συντελεστή συσχέτισης τ του Kendall του (u_1,u_2)(δηλαδή των u_1 και u_2) για την παράμετρο θ_2 (Wong et al. 2008). Για να ορίζεται σύζευξη θα πρέπει $\theta_i \geq 1$ και $\theta_1 \leq \theta_2$. Οι παράμετροι εδώ, εκτιμήθηκαν $\theta_1 = 3.3$ και $\theta_2 = 8.96$. Με τη βοήθεια της μεθόδου bootstrap percentile, κατασκευάστηκε από το παρατηρηθέν σύνολο δεδομένων ένας αριθμός B=3000 νέων δειγμάτων, ίσου μεγέθους. Στη συνέχεια υπολογίστηκε για κάθε δείγμα ο συντελεστής συσχέτισης τ του Kendall, και έπειτα το 95% διάστημα εμπιστοσύνης των παραμέτρων,

$$(2.301,\ 4.168) \text{ και } (6.9873,\ 17.763),$$

για τις παραμέτρους θ_1 και θ_2 αντίστοιχα.

Επομένως η σύζευξη παίρνει τη μορφή

$$C(u_1,u_2,u_3)=\exp[-[-\ln[\exp[-((-\ln u_1)^{8.96}+(-\ln u_2)^{8.96})^{0.11}]^{3.3}]+(-\ln u_3)^{3.3}]^{0.293}] \quad (3)$$

Για την ανωτέρω σύζευξη έγινε έλεγχος καλής προσαρμογής στα δεδομένα με χρήση του στατιστικού ελέγχου X^2. Η p-τιμή του στατιστικού ελέγχου X^2 για τις μετασχηματισμένες μεταβλητές προκύπτει ίση με 0,743.

Χρησιμοποιήσαμε στη συνέχεια και την τριδιάστατη κανονική σύζευξη, όπως αυτή προκύπτει από το θεώρημα του Sklar

$$C(u_1,u_2,u_3;\rho)=\Phi(\Phi^{-1}(u_1),\Phi^{-1}(u_2),\Phi^{-1}(u_3);\rho) \quad (4)$$

με Φ την τυπική μονοδιάστατη κανονική κατανομή, $u_1=\Phi(z_1)$, $u_2=\Phi(z_2)$ και $u_3=\Phi(z_3)$, και με συνάρτηση πυκνότητας πιθανότητας (Wang et al. 2010)

$$c(u_1,u_2,u_3;\rho)=\frac{\partial^3 C(u_1,u_2,u_3;\rho)}{\partial u_1\partial u_2\partial u_3}$$

$$=\frac{\exp\left(\begin{array}{l} z_1^2\left(\rho_{23}^2-1\right)+z_2^2\left(\rho_{13}^2-1\right)+z_3^2\left(\rho_{12}^2-1\right) \\ -\dfrac{+2\left[z_1z_2\left(\rho_{12}-\rho_{13}\rho_{23}\right)+z_1z_3\left(\rho_{13}-\rho_{12}\rho_{23}\right)+z_2z_3\left(\rho_{23}-\rho_{12}\rho_{13}\right)\right]}{2\left(\rho_{12}^2+\rho_{13}^2+\rho_{23}^2-2\rho_{12}\rho_{13}\rho_{23}\right)} \end{array}\right)}{(2\pi)^{3/2}\sqrt{1-\left(\rho_{12}^2+\rho_{13}^2+\rho_{23}^2+2\rho_{12}\rho_{13}\rho_{23}\right)}}$$

όπου ρ είναι ο συντελεστής συσχέτισης με $\rho\in\{\rho_{12},\rho_{13},\rho_{23}\}$. Για τα δεδομένα μας βρέθηκε ότι

$$\rho_{12}=0.9691,\ \rho_{13}=0.9059,\ \rho_{23}=0.8980.$$

Τα 95% διαστήματα εμπιστοσύνης για τις παραμέτρους ρ_{12}, ρ_{13} και ρ_{23} είναι

(0.9525, 0.9886), (0.8223, 0.9185) και (0.8794, 0.94407)

αντίστοιχα. Η p-τιμή του στατιστικού ελέγχου X^2 για τον έλεγχο καλής προσαρμογής στα δεδομένα υπολογίστηκε ίση με 0.762

Στη συνέχεια σχηματίσαμε τη μη παραμετρική εκτιμήτρια $\hat{C}_n$ (βλ. (2)),

$$\begin{aligned}\hat{C}_n(u_1,u_2,u_3) &=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_1h_2h_3}K\left(\frac{u_1-\hat{U}_{1i}}{h_1},\frac{u_2-\hat{U}_{2i}}{h_2},\frac{u_3-\hat{U}_{3i}}{h_3}\right)\\ &=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_1h_2h_3}K\left(\frac{u_1-\hat{U}_{1i}}{h_1}\right)\cdot K\left(\frac{u_2-\hat{U}_{2i}}{h_2}\right)\cdot K\left(\frac{u_3-\hat{U}_{3i}}{h_3}\right)\end{aligned} \qquad (5)$$

όπου $K(x)=\int_{-\infty}^{x}k\,du$ είναι το ολοκλήρωμα μιας συμμετρικής και φραγμένης συνάρτησης πυρήνα *k*. Εδώ χρησιμοποιήσαμε τον πυρήνα Epanechnikov k.

Το εύρος ζώνης h_i υπολογίστηκε από τη σχέση (Scott 1992; Bowman et al. 2004)

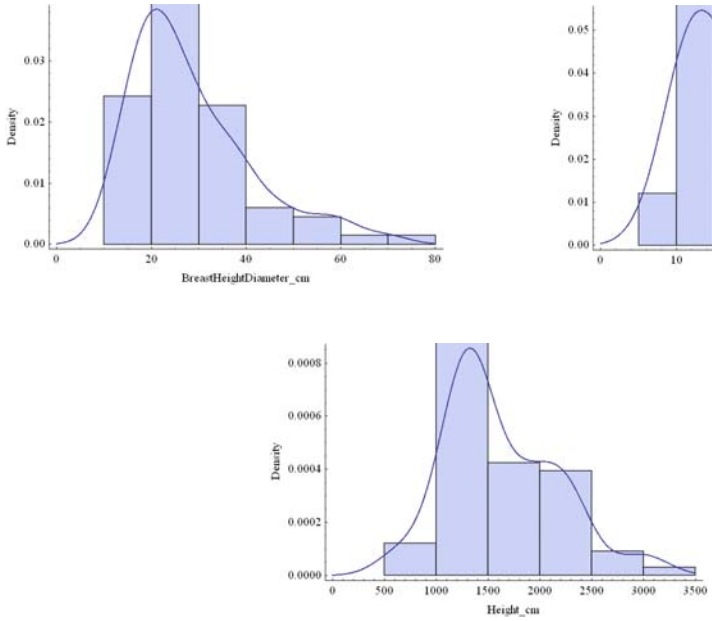
$$h_i=\left(\frac{4}{5}\right)^{1/7}\sigma_i n^{-1/7},\ \text{για}\ i=1,2,3,$$

όπου σ_i η τυπική απόκλιση της i-μεταβλητής. Το εύρος ζώνης υπολογίστηκε για τη στηθιαία διάμετρο $h_1=0.151$, για τη διάμετρο στο μέσο του κορμού $h_2=0.149$, και για το ύψος $h_3=0.151$. Για τον έλεγχο καλής προσαρμογής με βάση την εκτιμήτρια $\hat{C}_n$ εφαρμόσαμε τη δοκιμασία X^2 και βρέθηκε η p-τιμή ίση με 0.304.

Στο σχήμα 1, παρουσιάζουμε την πυκνότητα του εκτιμητή πυρήνα καθώς και τα

ιστογράμματα.

Σχήμα 1. *Ιστόγραμμα και μη παραμετρική εκτίμηση με χρήση συνάρτησης πυρήνα της στηθιαίας διαμέτρου, της διαμέτρου στο μέσο του κορμού και του ύψους.*



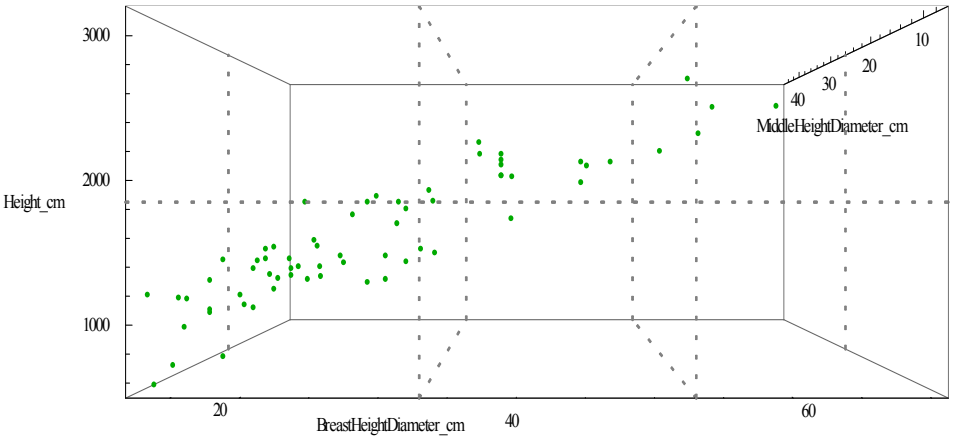
Υπάρχει μια ισχυρή συμφωνία ανάμεσα στις πιθανότητες από την εμπειρική κατανομή και στις εκτιμώμενες με χρήση τόσο της κανονικής σύζευξης όσο και της σύζευξης των Gumbel and Hougaard, γεγονός που φαίνεται και από τα αποτελέσματα του X^2 τεστ, υποδεικνύοντας ότι και οι δύο συζεύξεις περιγράφουν αρκετά καλά την εξάρτηση-σχέση των τριών χαρακτηριστικών του δέντρου της στηθιαίας διαμέτρου του κορμού, της διαμέτρου στο μέσο του κορμού και του ύψους. Για επιβεβαίωση της καλής προσαρμογής, παραθέτουμε στον Πίνακα 1, που ακολουθεί, τους συντελεστές συσχέτισης του δείγματος και των συζεύξεων Gumbel και κανονικής.

Πίνακας 1. *Συντελεστές συσχέτισης του δείγματος και των συζεύξεων*

Ζεύγος Μεταβλητών	Συντελεστής συσχέτισης		
	Δείγμα	Gumbel	Gaussian
(u_1, u_2)	0.9691	0.9827	0.9661
(u_1, u_3)	0.9059	0.8742	0.8978
(u_2, u_3)	0.8980	0.8747	0.8893

Στο σχήμα 2 φαίνεται το σύνολο των δένδρων που χρησιμοποιήθηκαν ως δεδομένα:

Σχήμα 2. *Παράσταση της στηθιαίας διαμέτρου του κορμού, της διαμέτρου στο μέσο του κορμού και του ύψους.*



Στη συνέχεια, δίνουμε στον Πίνακα 2 για διαφορετικές τυχαίες τιμές των X_1, X_2 και X_3 τις θεωρητικές πιθανότητες που προκύπτουν με βάση τις εκτιμήτριες των συζεύξεων που ορίζονται από τις (3), (4) και (5), καθώς και τις πιθανότητες που εκτιμώνται με βάση την εμπειρική κατανομή.

Πίνακας 2. *Πιθανότητες από εμπειρική κατανομή και θεωρητικές πιθανότητες των συζεύξεων, για τη στηθιαία διάμετρο, τη διάμετρο στο μέσο του κορμού και το ύψος.*

	Εμπειρική Κατανομή	Κανονική Σύζευξη	Εκτιμητής Πυρήνα $\hat{C}_n$	Σύζευξη Gumbell
$P[X_1 \le 20,\ X_2 \le 26,\ X_3 \le 1850]$	0.242	0.3275	0.2940	0.3257
$P[X_1 \le 33,\ X_2 \le 26,\ X_3 \le 1850]$	0.606	0.6381	0.5885	0.6458
$P[X_1 \le 33,\ X_2 \le 26,\ X_3 \le 3150]$	0.712	0.7164	0.6920	0.7164
$P[X_1 \le 53,\ X_2 \le 26,\ X_3 \le 1850]$	0.612	0.6701	0.6324	0.6757
$P[X_1 \le 53,\ X_2 \le 26,\ X_3 \le 3150]$	0.742	0.8027	0.7491	0.8059
$P[X_1 \le 53,\ X_2 \le 45,\ X_3 \le 1850]$	0.636	0.6855	0.6603	0.6858
$P[X_1 \le 53,\ X_2 \le 45,\ X_3 \le 3150]$	0.924	0.9100	0.8301	0.9103
$P[X_1 \le 73,\ X_2 \le 45,\ X_3 \le 3150]$	1	0.9761	0,8623	0.9809

Σύμφωνα με τα παραπάνω προκύπτουν τα εξής συμπεράσματα:

4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Με βάση την ανάλυση των δεδομένων μας, βρήκαμε ότι η τριδιάστατη κανονική σύζευξη μπορεί να χρησιμοποιηθεί αποτελεσματικά για την κατασκευή τριδιάστατων κατανομών κατάλληλων να περιγράψουν τη σχέση της στηθιαίας διαμέτρου, της διαμέτρου στο μέσο του κορμού και του ύψους των δέντρων. Θεωρούμε ότι η κανονική τριδιάστατη σύζευξη μπορεί να χρησιμοποιηθεί και σε άλλο σύνολο δασικών δεδομένων ή σε άλλα είδη δέντρων δίνοντας ικανοποιητική εκτίμηση.
Από τους παραπάνω πίνακες προκύπτει ότι

- και η σύζευξη Gumbel and Hougaard δίνει καλή εκτίμηση της σχέσης της στηθιαίας διαμέτρου, της διαμέτρου στο μέσο του κορμού και του ύψους των δέντρων (ιδιαίτερα εκτός των κελιών με τις (ακραίες) πολύ μικρές τιμές των μεταβλητών). Η ίδια σύζευξη περιγράφει με επιτυχία τη σχέση της στηθιαίας διαμέτρου του κορμού και του ύψους, στη διδιάστατη περίπτωση στην

εργασία Γεράρδη κ.α. (2012),

- η εκτίμηση που έγινε με τη συνάρτηση πυρήνα δίνει ικανοποιητικά αποτελέσματα, όπως φαίνεται στον Πίνακα 2, κυρίως στο εσωτερικό του κύβου, ενώ στο άνω άκρο παρουσιάζεται μεγαλύτερο σφάλμα (γεγονός που απαντάται στα άκρα, κατά την εκτίμηση με συζεύξεις).

Ευχαριστήρια. Ευχαριστούμε τον κο Γ. Τσακλίδη, καθηγητή του Τμήματος Μαθηματικών ΑΠΘ, για τις παρατηρήσεις και τα σχόλιά του για τη συγγραφή της παρούσας εργασίας.

ABSTRACT

Quantifying the dependence among tree diameters and heights has been an enduring task for forest researchers. Copulas are functions that join or "couple" multivariate distribution functions to their one dimensional marginal distribution functions. Recently copulas are of central importance in statistical applications as they allow to model and estimate the distribution of random vectors by estimating marginals and copula separately. The purpose of this study is to introduce and evaluate a three-dimension copula that could successfully describe the joint distribution function of tree breast height diameters, middle height diameters and heights. We also consider a nonparametric kernel estimator of the copula function and then we use a goodness-of-fit test for parametric copula models. The theoretical findings are confirmed by a simulation study using data set of trees which are located at the Pertouli's forest.

ΑΝΑΦΟΡΕΣ

Bowman, A.W., and Azzalini, A. (1997). Applied Smoothing Techniques for Data Analysis, London: Oxford University Press.

Deheuvels P. (1979). La fonction de dépendance empirique et ses propriétés Acad.Roy. Belg. Bull. Cl. Sci. 65: 274-292.

Genest C., Rivest. L. P. (1993). Statistical inference procedures for bivariate Archimedean Copulas. J. Amer. Statist. Assoc. 55: 698-707.

Hafley W.L., Schreuder H.T. (1976). Some non-normal bivariate distributions and their potential for forest application.P. 104–114 in the XVI world congress proc., Div. VI. Intern. Union of Forestry Research Organis., Oslo, Norway.

Kershaw J.A. Jr., Richards E.W., McCarter J.B., Oborn S. (2010). Original paper: spatially correlated forest stand structures: a simulation approach using copulas. Comput Electron Agric 74:120–128.

Li F., Zhang L., Davis C.J. (2002). Modeling the joint distribution of tree diameters and heights by bivariate generalized beta distribution. For. Sci. 48(1): 47-58. Nelsen. R. B. (2005). An Introduction to Copulas. Second Edition.

Omelka M., Gijbels I. and Veraverbeke N. (2009). Improved Kernel Estimation of Copulas: Weak Convergence and Goodness of Fit Testing. Annals of Statistics. Vol 37, No 5B, 3023-3058.

Petrauskas E., Rupšys P., Bartkevičius E. (2011). Volume Modeling of Individual Trees Using Stochastic Differential Equations and Trivariate Copula. Intern.

Conference on Circuits, System and Simulation IPCSIT vol.7: 258-263.
Sain, S.R., Baggerly, K.A., and Scott, D.W. (1994). Cross-Validation of MultivariateDensities. *Journal of the American Statistical Association*, 89: 807-817.
Schreuder H. T., Hafley W. L. (1977). A Useful Bivariate Distribution for Describing Stand Structure of Tree Heights and Diameters. Biometrics Vol. 33, (3): 471-478
Schreuder H.T., Bhattacharya H.T., Mcclure J.P. (1982a). Towards a unified distribution theory for stand variables using S_{BBB} distribution. Biometrics 38:137–142.
Schreuder H.T., Bhattacharya H.T., Mcclure J.P. (1982b). The S_{BBB} distribution: A potentially useful trivariate distribution. Can. J. For. Res. 12: 641– 645.
Scott D.W. (1992). Multivariate Density Estimation: Theory, Practice, and Visualization, New York: JohnWiley.
Silverman B. W., (1986). Density Estimation for Statistics and Data Analysis.Chapman and Hall. London.
Sklar A. (1959). Fonctions de repartition a n dimensions et leurs marges. Publ. Inst. Statist. Univ. Paris 8: 229 –231.
Stamatellos G. (1995). Comparison of Point and Point-3P sampling with cost analysis, Forest Ecology Management 74**:** 75-79.
Wang M., Rennolls K., Tang S. (2008). Bivariate distribution modeling of treediameters and heights: Dependency modeling using copulas. For. Sci. 54(3):284 –293.
Wang M., Upadhyay A., Zhang L. (2010). Trivariate Distribution Modeling of Tree Diameter, Height, and Volume. Forest Science 56(3): 290-300.
Wong G., Lambert M. F., Metcalfe A. V. (2008). Trivariate copulas for characterization of droughts. ANZIAM J. 49: C306-323.
Zucchini W., Schmidt M., Gadow K. V. (2001). A model for the diameter-height distribution in an uneven-aged beech forest and a method to assess the fit of such models. Silva Fenn. 35(2):169-183.
Γεράρδη Δ., Τσακλίδης Γ., Σταματέλλος Γ. (2012). Μη Παραμετρική Εκτίμηση Διδιάστατων Συζεύξεων. Πρακτικά 25ου Πανελλήνιου Συνεδρίου Στατιστικής. Βόλος 2012.
Γκιούλος Λ.Ν. (2007). Συζεύξεις (Copulas). Εκτίμηση της Κοινής Κατανομής Στηθιαίας Διαμέτρου και Ύψους Υβριδογενούς Ελάτης του Δάσους Περτουλίου. Διπλωματική Εργασία, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης.
Σταματέλλος Γ. (1991). Έρευνα των Δυνατοτήτων Ογκομέτρησης των Δασών με Δισταδιακά Δειγματοληπτικά Σχέδια. Διδακτορική Διατριβή. Α. Π. Θ..

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής (2015), σελ.71-80

ΑΠΟΘΟΡΥΒΟΠΟΙΗΣΗ ΣΗΜΑΤΩΝ ΜΕ ΤΗΝ ΜΕΘΟΔΟ ΑΝΑΛΥΣΗΣ ΑΝΕΞΑΡΤΗΤΩΝ ΣΥΝΙΣΤΩΣΩΝ: ΕΠΙΠΛΕΟΝ ΑΞΙΟΠΟΙΗΣΗ ΣΤΗΝ ΕΥΡΕΣΗ ΠΗΓΩΝ

Σ.Γυλού [1] ***, Φ.Κολυβά-Μαχαίρα*** [1] ***, Χ.Φραντζίδης*** [2] ***, Π.Μπαμίδης*** [2]
[1]ΤμήμαΜαθηματικών, Α.Π.Θ.
sotiriagilou@gmail.com , fkolyva@math.auth.gr
[2]Εργαστήριο Ιατρικής Φυσικής, Ιατρική Σχολή, Α.Π.Θ.
{christos.frantzidis,pdbamidis}@gmail.com

ΠΕΡΙΛΗΨΗ

Η Ανάλυση Ανεξάρτητων Συνιστωσών (IndependentComponentAnalysis), (I.C.A.), είναι μια σχετικά πρόσφατη μέθοδος ανάλυσης δεδομένων που εφαρμόζεται σε δεδομένα τα οποία δεν ακολουθούν κανονική κατανομή. Η μέθοδος ICA αναπτύχθηκε αρχικά για να εξεταστούν τα προβλήματα που σχετίζονται αρκετά με το πρόβλημα του «κοκτέιλ πάρτι». Πρόσφατα, το ενδιαφέρον έχει αυξηθεί και για άλλες εφαρμογές της μεθόδου, όπως στην Οικονομετρία όπου υπάρχουν συχνά παράλληλες χρονικές σειρές και η ICA μπορεί να τις αποσυνθέσει σε ανεξάρτητες συνιστώσες οι οποίες δίνουν μια πρόσβαση εσωτερικά στη δομή του συνόλου δεδομένων, στην εξαγωγή χαρακτηριστικών από εικόνες όπου σκοπός είναι να βρεθούν χαρακτηριστικά τα οποία είναι όσο το δυνατόν ανεξάρτητα, στην Ιατρική και συγκεκριμένα στην απεικόνιση του εγκεφάλου όπου υπάρχουν συχνά διαφορετικές πηγές στονεγκέφαλο που εκπέμπουν σήματα τα οποία αναμειγνύονται σε αισθητήρες έξωαπό το κεφάλι, όπως στο βασικό μοντέλο διαχωρισμού πηγών κλπ.

Στην εργασία αυτήθα ανιχνευτούν μέσω της Ανάλυσης Ανεξαρτήτων Συνιστωσών και τη χρήση ισοδύναμων διπόλων, περιοχές του εγκεφάλου που ενεργοποιούνται κατά την παθητική θέαση οπτικοποιημένων λεκτικών ερεθισμάτων σε 21 άτομα νεαρής ηλικίας. Τα ερεθίσματα (λέξεις) κατατάσσονται ανάλογα με το βαθμό ευχαρίστησης σε ευχάριστα, δυσάρεστα και ουδέτερα. Τόσο τα ευχάριστα, όσο και τα δυσάρεστα χαρακτηρίζονται από υψηλή διέγερση (ένταση συναισθήματος) σε αντίθεση με τα ουδέτερα που χαρακτηρίζονται από χαμηλή.Σε πρώτη φάση η Ανάλυση Ανεξαρτήτων Συνιστωσών εφαρμόζεται για την αποθορυβοποίηση των σημάτων από μυϊκά, οφθαλμολογικά ή άλλου τύπου παράσιτα (artifacts). Στη συνέχεια με τη βοήθεια του γραφικού περιβάλλοντος EEGLABτης Matlab, βρίσκονται οι 57 ανεξάρτητες συνιστώσες που αντιστοιχούν στα 57 ηλεκτρόδια που εφαρμόστηκαν σε κάθε συμμετέχοντα. Σε επόμενο βήμα, με τη βοήθεια του εργαλείου dipfit2 του EEGLAB της Matlab, βρίσκονται τα δίπολα τα οποία είναι σημειακά φορτία που δημιουργούν την εγκεφαλική δραστηριότητα για κάθε μια από τις 57 ανεξάρτητες συνιστώσες που έχουν βρεθεί προηγουμένως. Σε επόμενο στάδιο με χρήση μιας μετρικής που ονομάζεται «Σχετική Κυματιδιακή Εντροπία», υπολογίζεται ο συγχρονισμός μεταξύ ζευγών εγκεφαλικών περιοχών που επιλέχθηκαν από το προτεινόμενο μοντέλο. Τέλος, εφαρμόστηκε το μη-παραμετρικό κριτήριοWilcoxon για ζευγαρωτές παρατηρήσεις για κάθε ένα από τα αντίστοιχα ζευγάρια περιοχών του πίνακα βαθμού συγχρονισμού για 18 συμμετέχοντες και ελέγχθηκε εάν διαφέρει ο βαθμός συγχρονισμού σε 2 συναισθηματικές κατηγορίες, δηλαδή στις ευχάριστες-δυσάρεστες, δυσάρεστες-ουδέτερες και ευχάριστες-ουδέτερες λέξεις.

*Λέξεις Κλειδιά:*ICA., ανεξάρτητες συνιστώσες, αποθορυβοποίηση, λεκτικά ερεθίσματα, δίπολα, dipfit2, Σχετική Κυματιδιακή Εντροπία, βαθμός συγχρονισμού, εγκεφαλικές περιοχές

1. ΕΙΣΑΓΩΓΗ

Η Ανάλυση Ανεξαρτήτων Συνιστωσών(Independent Component Analysis, ICA) είναι μια μέθοδος εύρεσης στοχαστικά ανεξάρτητων συνιστωσών από στατιστικά δεδομένα πολλών μεταβλητών. Αυτό που ξεχωρίζει την ICA από άλλες μεθόδους είναι ότι οι συνιστώσες της δεν ακολουθούν την κανονική κατανομή. Στη βιβλιογραφία η μέθοδος αυτή παρομοιάζεται

με το πρόβλημα του «κοκτέιλ πάρτι», όπου υπάρχουν καταγραφέςkσυνομιλητών που μιλούν ταυτόχρονα σε k μικρόφωνα και τις οποίες συμβολίζουμε με $X_1(t), X_2(t), \ldots, X_k(t)$ και προσπαθούμε να βρούμε τα λεκτικά σήματα που εκπέμπονται από τον καθένα ξεχωριστά και τα οποία συμβολίζονται με $S_1(t), S_2(t), \ldots, S_k(t)$. Στο εξής, αν και τόσο οι καταγραφές όσο και τα σήματα είναι συναρτήσεις του χρόνου, χάρην απλότητας θα συμβολίζουμε $X_i(t) = X_i$ και $S_j(t) = S_j$. Θεωρούμε ότι κάθε μια από τις καταγραφές είναι γραμμικός συνδυασμός των λεκτικών σημάτων. Το πρόβλημα του «κοκτέιλ πάρτι» επεκτείνεται και στα εγκεφαλογραφήματα που είναι το αντικείμενο αυτής της εργασίας.

Εξετάζουμε για παράδειγμα, τις ηλεκτρικές καταγραφές της εγκεφαλικής δραστηριότητας όπως δίνεται από ένα ηλεκτροεγκεφαλογράφημα (EEG). Τα δεδομένα του ηλεκτροεγκεφαλογραφήματος (EEG) αποτελούνται από τις καταγραφές των ηλεκτρικών δυναμικών σε διαφορετικές θέσεις στο κρανίο. Αυτά τα ηλεκτρικά δυναμικά, πιθανώς παράγονται σε διάφορες περιοχές του εγκεφάλου που λειτουργούν ως ανεξάρτητες συνιστώσες. Θα επιθυμούσαμε να μπορούσαμε να βρούμε τις αρχικές συνιστώσες της δραστηριότητας του εγκεφάλου, αλλά μπορούμε μόνο να παρατηρήσουμε το συνδυασμό των συνιστωσών. Η μέθοδος ICA μπορεί να αποκαλύψει ενδιαφέρουσες πληροφορίες για τη δραστηριότητα του εγκεφάλου, παρέχοντας πρόσβαση στις ανεξάρτητες συνιστώσες οι οποίες είναι και οι ενεργοποιημένες εγκεφαλικές περιοχές που συμβάλλουν στα σήματα που εξάγουμε.

Η δομή της εργασίας έχει ως εξής: Στο $2^{ο}$ κεφάλαιο γίνεται παρουσίαση της μεθόδου Ανάλυσης Ανεξαρτήτων Συνιστωσών και των βασικών αρχών που τη διέπουν όπως η μη-κανονικότητα. Στην συνέχεια στο $3^{ο}$ κεφάλαιο αναφέρεται η μεθοδολογία που αναπτύχθηκε στην εργασία αυτή όπως είναι η δομή του πειράματος, πληροφορίες για τους συμμετέχοντες, η προεπεξεργασία δεδομένων και το πειραματικό πρωτόκολλο που ακολουθήθηκε. Στην συνέχεια στο $4^{ο}$ κεφάλαιο αναλύεται το μοντέλο που επιλέχθηκε και στη συνέχεια ο στατιστικός έλεγχος σε ζευγάρια περιοχών που αναφέραμε παραπάνω και για τα οποία βρήκαμε στατιστικά σημαντική διαφορά στους βαθμούς συγχρονισμού σε 2 βαθμούς ευχαρίστησης. Τέλος στο $5^{ο}$ κεφάλαιο γίνεται ένας σχολιασμός της εργασίας στο σύνολο της καθώς και τα μειονεκτήματα και οι περιορισμοί του μοντέλου που αναπτύχθηκε. Τέλος διατυπώνονται στόχοι για μια μελλοντική εργασία, ως συνέχεια αυτής της εργασίας.

2.Η μέθοδος Ανάλυσης Ανεξάρτητων Συνιστωσών

Υποθέτουμε ότι παρατηρούνται n τυχαίες μεταβλητές $X_1, X_2, ..., X_n$ που μοντελοποιούνται ως γραμμικοί συνδυασμοί n τυχαίων μεταβλητών $S_1, S_2, ..., S_n$:

$$X_i = a_{i1}S_1 + a_{i2}S_2 + ... + a_{in}S_n \quad i = 1, 2, ..., n \tag{1}$$

Όπου τα $a_{ij}, i = 1, 2, ..., n, j = 1, 2, ..., n$ είναι πραγματικοί συντελεστές. Εξορισμού, οι τυχαίες μεταβλητές S_j ,είναι στοχαστικά ανεξάρτητες.

Αυτό είναι το βασικό μοντέλο της ICA. Το μοντέλο αυτό περιγράφει πως τα παρατηρούμενα μεγέθη παράγονται με μια διαδικασία μίξης των συνιστωσών S_j. Οι ανεξάρτητες συνιστώσες S_j(ή με συντομογραφία IC_s) ονομάζονται λανθάνουσες μεταβλητές (latentvariables) , που σημαίνει ότι δεν μπορούν άμεσα να παρατηρηθούν. Επίσης οι συντελεστές μίξης a_{ij} είναι άγνωστοι. Αυτό που μπορούμε να παρατηρήσουμε είναι οι τυχαίες μεταβλητές X_i, και πρέπει να εκτιμήσουμε μαζί τους συντελεστές μίξης a_{ij} και τις ανεξάρτητες συνιστώσες S_j χρησιμοποιώντας τις X_i. Χωρίς περιορισμό της γενικότητας, υποθέτουμε ότι οι μεταβλητές μιγμάτων και οι ανεξάρτητες συνιστώσες έχουν μέση τιμή μηδέν.

Η μέθοδος ICA συσχετίζεται με τη μέθοδο BlindSourceSeparation (τυφλός διαχωρισμός πηγής) ή BlindSignalSeparation (τυφλός διαχωρισμός σημάτων). Ως "πηγή" εννοείται ένα αρχικό σήμα, δηλαδή ανεξάρτητη συνιστώσα, όπως ο ομιλητής σε ένα πρόβλημα του «κοκτέιλ-πάρτι». "Blind" σημαίνει ότι ξέρουμε πολύ λίγα για τον πίνακα μίξης και κάνουμε υποθέσεις για την πηγή των σημάτων. Η ICA είναι μια μέθοδος, ίσως ευρύτατα χρησιμοποιημένη, για την εκτέλεση του τυφλού διαχωρισμού πηγής.

Το πρότυπο μίξης γράφεται με τη βοήθεια πινάκων ως εξής:

$$\mathbf{X}= A\,\mathbf{S} \qquad (2)$$

Όπου **X**είναι το τυχαίο διάνυσμα με στοιχεία τις μίξεις $X_1, X_2, ..., X_n$, **S**το τυχαίο διάνυσμα που έχει ως στοιχεία τα $S_1, S_2, ..., S_n$ και ο πίνακας A έχει ως στοιχεία τα a_{ij}.

Ένα από τα βασικά μειονεκτήματα της μεθόδου είναι ότι δεν μπορεί να καθοριστεί η τάξη των ανεξάρτητων συνιστωσών. Λόγω του ότι τόσο το τυχαίο διάνυσμα **S**όσο και ο πίνακας A είναι άγνωστα, μπορούμε ελεύθερα να αλλάξουμε την τάξη των όρων στο άθροισμα της σχέσης και να αποκαλέσουμε οποιαδήποτε από τις ανεξάρτητες συνιστώσες πρώτη. Εάν P ένας αντιμεταθετικός πίνακας τότε ισχύει:$\mathbf{X}= AP^{-1}P\,\mathbf{S}$. Τα στοιχεία του P **S**είναι οι αρχικές ανεξάρτητες μεταβλητές S_j, αλλά σε μια άλλη τάξη. Έτσι ο πίνακας AP^{-1} είναι ο νέος άγνωστος πίνακας μίξης και μπορεί να επιλυθεί από τους ICA αλγορίθμους.

2.1 Περιορισμοί της Ανάλυσης Ανεξαρτήτων Συνιστωσών

Για να βεβαιωθούμε το ICA μοντέλο που θα μας δοθεί μπορεί να εκτιμηθεί, θα πρέπει να κάνουμε μερικές υποθέσεις και περιορισμούς.

- Οι ανεξάρτητες συνιστώσες είναι στοχαστικά ανεξάρτητες.
- Οι ανεξάρτητες συνιστώσες πρέπει να μην ακολουθούν κανονική κατανομή.
- Για λόγους ευκολίας, υποθέτουμε ότι ο άγνωστος πίνακας μίξης είναι τετραγωνικός. Μετά την εκτίμηση του πίνακα Α αν είναι αντιστρέψιμος, μπορούμε να υπολογίσουμε τον αντίστροφο του, έστω Β και έτσι να έχουμε τις ανεξάρτητες συνιστώσες από την σχέση : $\mathbf{X}= A\,\mathbf{S} \Leftrightarrow A^{-1}\mathbf{X}= A^{-1}\,A\,\mathbf{S} \Leftrightarrow \mathbf{S}=\mathrm{B}\mathbf{X}$
- Σε πολλές εφαρμογές, θα ήταν πιο ρεαλιστικό εάν υποθέταμε ότι υπάρχει κάποιος θόρυβος στις μετρήσεις μας, κάτι το οποίο θα είχε ως αποτέλεσμα να προστεθεί ο όρος του θορύβου στο μοντέλο μας. Για απλούστευση του προβλήματος όμως, παραλείπουμε οποιονδήποτε θόρυβο, αφού η μέθοδος είναι αρκετά δύσκολη και απαιτητική και χωρίς την υπόθεση θορύβου στα δεδομένα μας.
- Στην παρούσα εργασία θεωρούμε ότι ο αριθμός των παρατηρούμενων μεγεθών (μίξεων) είναι ίσος με τον αριθμό των ανεξαρτήτων συνιστωσών. Έτσι, ο πίνακας μίξης είναι τετραγωνικός και μπορούμε να λύσουμε το πρόβλημα. Όμως, τις περισσότερες περιπτώσεις αυτό δε συμβαίνει και συγκεκριμένα ο αριθμός των ανεξαρτήτων συνιστωσών είναι μεγαλύτερος από αυτόν των παρατηρήσεων. Στην περίπτωση αυτή ο πίνακας μίξης είναι μη-τετραγωνικός και το πρόβλημα γίνεται αρκετά πιο πολύπλοκο.

3.ΜΕΘΟΔΟΛΟΓΙΑ

3.1 Δομή πειράματος & Συμμετέχοντες

Στο πείραμα που αναπτύσσεται παρακάτω συμμετείχαν 21 υγιή νεαρά άτομα με ηλικία 28,31±5,982 χωρίς καμία νευροψυχολογική πάθηση, τα οποία αξιολογήθηκαν μέσω μιας νευροφυσιολογικής εκτίμησης η οποία ήταν μέρος μιας διαγνωστικής διαδικασίας στα πλαίσια του προγράμματος «μνήμες διαρκείας» (LongLastingMemories/ LLM). Το LLM ήταν ένα πολυκεντρικό χρηματοδοτούμενο πρόγραμμα από την Ευρωπαϊκή Κοινότητα το οποίο πρότεινε μια παρέμβαση που περιελάμβανε το συνδυασμό, της γνωστικής και φυσικής

άσκησης μέσω υπολογιστή, προκειμένου να προαχθεί η ανεξάρτητη διαβίωση των ηλικιωμένων (www.longlastingmemories.eu) (Bamidisetal, 2011; Gonzalez-Palauetal, 2014). Στην παρούσα εργασία πραγματοποιήθηκε η ανάλυση των νευροφυσιολογικών δεδομένων που ελήφθησαν από ένα δείγμα ελέγχου νεαρών, υγιών εθελοντών. Στο πείραμα, οι συμμετέχοντες κάθονται σε μία αναπαυτική καρέκλα, η οποία απέχει 1,5 μέτρα από μια οθόνη υπολογιστή στην οποία εμφανίζονται οι λέξεις (ερεθίσματα) με συγκεκριμένους χρόνους. Τα πειράματα διεξήχθησαν στην Εταιρεία Νόσου Alzheimer και Συγγενών Διαταραχών που βρίσκεται στην Θεσσαλονίκη.

3.2 Πειραματικό πρωτόκολλο

Το δείγμα αποτελείται από 21 νεαρούς εθελοντές που ήταν γνωστό ότι ήταν απαλλαγμένοι από νευρολογικές ή ψυχοσυναισθηματικές διαταραχές. Σε πρώτο στάδιο οι συμμετέχοντες συναινούν για τη συμμετοχή τους στο πείραμα καθώς επίσης γίνεται καταγραφή των δημογραφικών στοιχείων τους όπως ημερομηνία γέννησης κλπ.

Στο επόμενο στάδιο γίνεται η προετοιμασία των συμμετεχόντωνόπου γίνεται καθαρισμός της κεφαλής τους με οινόπνευμα ώστε να γίνει η καταγραφή του ηλεκτροεγκεφαλογραφήματος και να τοποθετηθούν τα ηλεκτρόδια.

Στο τρίτο στάδιο προβάλλονται σε κάθε συμμετέχοντα 120 οπτικοποιημένα λεκτικά ερεθίσματα (λέξεις). Τα ερεθίσματα αυτά είναι κατανεμημένα σε 3 συναισθηματικές κατηγορίες ευχάριστα, δυσάρεστα, ουδέτερα (40 λέξεις ανά κατηγορία). Θα πρέπει να σημειωθεί ότι κάθε λεκτικό ερέθισμα προκαλεί μια συναισθηματική απόκριση η οποία περιγράφεται μέσω της μεταβλητής του **σθένους (βαθμός ευχαρίστησης)** και **της διέγερσης (arousal)**. Το σθένος δείχνει το βαθμό ευχαρίστησης που προκαλείται στον κάθε συμμετέχοντα μετά την εμφάνιση του οπτικοποιημένου λεκτικού ερεθίσματος ενώ η διέγερση είναι μια πιο γενική ιδιότητα και αναφέρεται στο επίπεδο ενεργοποίησης.

3.3 Προεπεξεργασία δεδομένων

Για την προεπεξεργασία των δεδομένων χρησιμοποιήθηκε το εργαλείο EEGLAB τηςMatlab το οποίο χρησιμοποιείται για την επεξεργασία σημάτων, καθώς και εγκεφαλογραφημάτων. Τα βήματα που ακολουθήσαμε ήταν τα παρακάτω:

- **Εισαγωγή και αφαίρεση calibration**
- **Εφαρμογή RemoveBaseline**
- **Re-reference**
- **Εφαρμογή αλγορίθμου ICA**

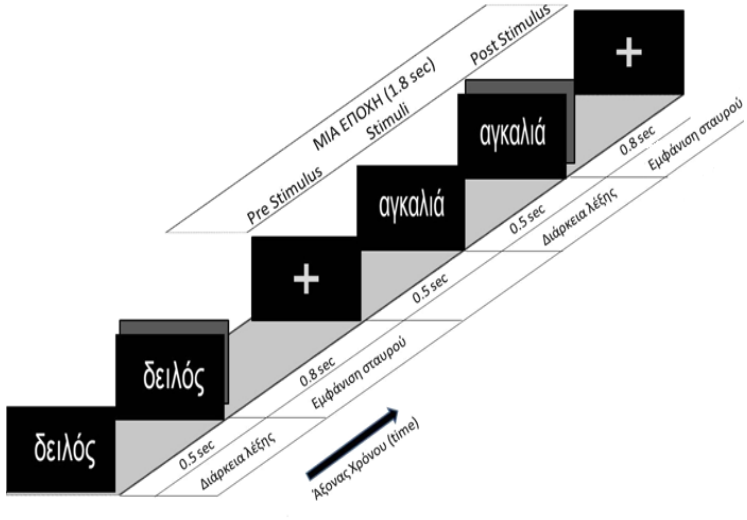
Ο αλγόριθμος ICA (Independent Component Analysis) υπολογίζει έναν πίνακα μίξης, με μέση τιμή 26 βήματα για το κάθε κανάλι, ενώ παράλληλα προκύπτουν από την εφαρμογή του, 57 ανεξάρτητες συνιστώσες (πηγές), όσες και τα ηλεκτρόδια από τα οποία εκπέμπονται 57 σήματα (μίξεις). Ο αλγόριθμος στο πείραμά μας σε πρώτη φάση χρησιμοποιείται για αποθορυβοποίηση. Στην ουσία, διώχνουμε τις πηγές θορύβου που μπορεί να προκύψουν από μυϊκά, οφθαλμολογικά παράσιτα ή και από κάποια κακή τοποθέτηση /λήψη σήματος ηλεκτροδίου.

3.3.1 Ορισμός εποχής

Το πείραμα έχει ως εξής : εμφανίζονται σήματα (120 λέξεις-events) με διάρκεια 1,3 sec. Μεταξύ των σημάτων υπάρχει παύση διάρκειας 0,5 sec. Ως εποχή ορίζεται η διάρκεια του σήματος μαζί με την παύση.

Η κάθε εποχήαντιστοιχεί σε χρόνο 1,8 δευτερολέπτων. Η συχνότητα δειγματοληψίας είναι 500 Hz, δηλαδή σε κάθε ηλεκτρόδιο πραγματοποιούνται 500 καταγραφέςανά δευτερόλεπτο, οπότε ανά εποχή υπάρχουν 900 καταγραφές.

Εικόνα 1. *Απεικόνιση εποχής κάθε ερεθίσματος*

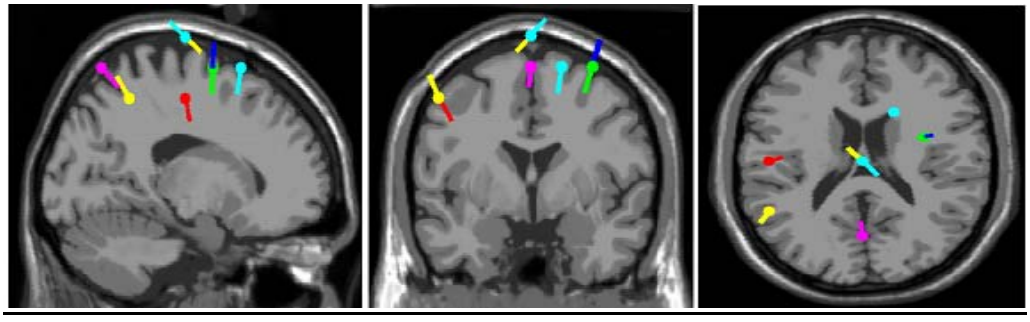


3.3.2 Εύρεση διπόλων

Το EEGLABτης Matlabδιαθέτει το εργαλείο dipfit 2 με το οποίο μπορούμε να βρούμε τα δίπολα για κάθε συναισθηματική κατηγορία κάθε συμμετέχοντα. ***Ως δίπολο ορίζουμε το σημειακό φορτίο που δημιουργεί την εγκεφαλική δραστηριότητα για τη συγκεκριμένη συνιστώσα***.

Στη συνέχεια μπορούμε να επιλέξουμε ποια δίπολα αντιστοιχούν σε συγκεκριμένες ανεξάρτητες συνιστώσες προκειμένου να απεικονιστούν σε σύστημα συντεταγμένων Talairach ώστε να μπορέσουμε να προσδιορίσουμε τις περιοχές του εγκεφάλου που ενεργοποιούνται κατά την εμφάνιση του ερεθίσματος.

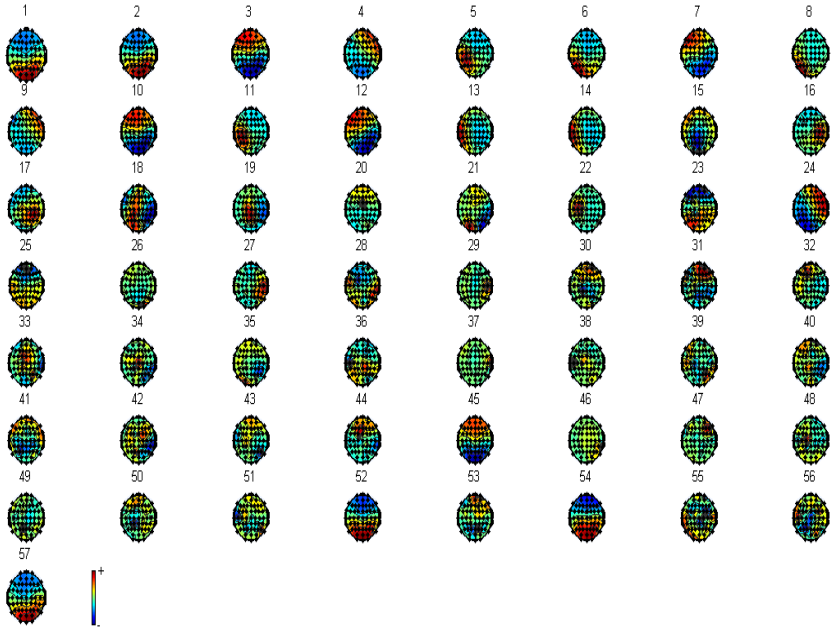
Εικόνα 2.*Απεικόνιση διπόλων για ένα συμμετέχοντα*



3.3.3Η εντολή Run-ICA& η εύρεση ανεξαρτήτων συνιστωσών

Κάθε epoched σήμα το οποίο αποτελείται από 40 λέξεις-events για κάθε συναισθηματική κατηγορία καθενός από τους συμμετέχοντες, το εισαγάγουμε στο EEGLAB και στη συνέχεια από την επιλογή ***tools->RunICA*** εφαρμόζουμε για δεύτερη φορά τον αλγόριθμο ανεξαρτήτων συνιστωσών για τα 57 κανάλια. Το σύνολο των 57 νέων ανεξαρτήτων συνιστωσών (ενεργοποιημένων περιοχών του εγκεφάλου) μπορούμε να το δούμε καθώς εκτελούμε την εντολή του EEGLAB***Plot->Componentmaps->In2-D.***

Εικόνα 3. *Το σύνολο των 57 ανεξαρτήτων συνιστωσών που αντιστοιχούν στα 57 ηλεκτρόδια. Κάθε απεικόνιση δείχνει την εγκεφαλική δραστηριότητα για κάθε συνιστώσα αντίστοιχα*



3.4 Υπολογισμός του πίνακα συγχρονισμού για κάθε συναισθηματική κατηγορία

Εφαρμόσαμε τον αλγόριθμο συγχρονισμού στις εγκεφαλικές περιοχές που περιέχονται στο προτεινόμενο μοντέλο. Οι περιοχές αυτές ορίστηκαν με βάση τις κοινές εγκεφαλικές περιοχές που βρέθηκαν στο dipfit2, συγκρινόμενες με αυτές της βιβλιογραφίας.Το προτεινόμενο μοντέλο για κάθε συναισθηματική κατηγορία καθενός από τους συμμετέχοντες, εφαρμόστηκε σε κάθε συναισθηματική κατηγορία. Στη συνέχεια, βρέθηκε ο πίνακας που περιέχει το συγχρονισμό ανά ζεύγος των επιλεγμένων εγκεφαλικών περιοχών.Ο πίνακας αυτός υπολογίζεται με τη χρήση μιας μετρικής που ονομάζεται «Σχετική Κυματιδιακή Εντροπία» (relativewaveletentropy) (Rossoetal, 2001; Frantzidis et al., 2010)

3.5 Στατιστική Ανάλυση

Στη συνέχεια για να διαπιστώσουμε αν διαφέρουν οι τιμές συγχρονισμού 2 ζευγών περιοχών του εγκεφάλου στον υψηλό με χαμηλό βαθμό ευχαρίστησης (ή ανάλογα στο χαμηλό-ουδέτερο, υψηλό-ουδέτερο) θα κάνουμε στατιστικό έλεγχο για ζευγαρωτές παρατηρήσεις. Επειδή έχουμε 2 δείγματα εξαρτημένα, και επειδή ο αριθμός των συμμετεχόντων είναι μικρότερος του 25, το κριτήριο που θα δοκιμάσουμε στο στατιστικό πακέτο SPSS θα είναι μη-παραμετρικό και συγκεκριμένα θα επιλέξουμε το κριτήριοWilcoxon. Στο επόμενο κεφάλαιο παρατίθενται αναλυτικά τα αποτελέσματα των ελέγχων αυτών.

4.ΑΠΟΤΕΛΕΣΜΑΤΑ

4.1 Ορισμός του μοντέλου

Για τη διαμόρφωση του μοντέλου επιλέχθηκαν τυχαία 3 άτομα στα οποία βρέθηκαν κοινές εγκεφαλικές περιοχές ενεργοποίησης εφαρμόζοντας το εργαλείο dipfit2 του EEGLAB, σύμφωνα με προηγούμενες μελέτες (Bullmore et al., 1996; Bedo, 2012; Klepousniotou et al., 2014). Οι περιοχές που καταλήξαμε ήταν :

1) **Ο Κατώτερος βρεγματικός λοβός**(inferiorparietallobule) με συντεταγμένες διπόλων (-50,-47,49),

2) **Το Προσφηνοειδές λοβίο** (precuneus) με συντεταγμένες διπόλων (0,-63,66),
3) **Η Μέση μετωπιαία έλικα** (middlefrontalgyrus) με συντεταγμένες (34,3,62) και την
4) **Η Άνω μετωπιαία έλικα** (superiorfrontalgyrus) με συντεταγμένες (17,20,62).

Στη συνέχεια εφαρμόσαμε το μοντέλο των τεσσάρων προκαθορισμένων περιοχών για τα υπόλοιπα (18) άτομα. **Στην περίπτωση πολλαπλών διπόλων στην ίδια περιοχή επιλέχθηκε εκείνο το οποίο αντιστοιχούσε στο μικρότερο ποσοστό RV(residualvariance).**

Τέλος, επιλέχθηκαν οι χρονοσειρές των συνιστωσών (ICAcomponents) που αντιστοιχούσαν στα επιλεγμένα δίπολα. Στις χρονοσειρές αυτές πραγματοποιήθηκε ανάλυση συγχρονικότητας με τη μέθοδο της «Σχετικής Κυματιδιακής Εντροπίας», όπως αναφέρθηκε παραπάνω.

Για το κάθε ένα από τα 18 άτομα του πειράματος, βρέθηκε ένας πίνακαςγια κάθε συναισθηματική κατηγορία, με στοιχεία τους βαθμούς συγχρονισμού μεταξύ των περιοχών. Οι τιμές του συγχρονισμού κυμαίνονται από το 0 έως το 1. Όσο πλησιάζουμε προς το 0 τόσο εντονότερο συγχρονισμό έχουμε σε αντίθεση με τις τιμές που τείνουν ανοδικά προς το 1, που σημαίνει ότι έχουμε λιγότερο συγχρονισμό. Παρακάτω δίνεται ενδεικτικά ο πίνακας του συγχρονισμού των 4 περιοχών του 1^{ου} συμμετέχοντα για την κατηγορία των ευχάριστων λέξεων.

Πίνακας 1. *Απεικόνιση των τιμών συγχρονισμού των περιοχών του μοντέλου ανά ζεύγη. Ο συγχρονισμός υπολογίστηκε για τον υψηλό βαθμό ευχαρίστησης (ευχάριστες λέξεις) του 1ου συμμετέχοντα*

	Κατώτερος βρεγματικός λοβός	**Προσφηνοειδές λοβίο**	**Μέση μετωπιαία έλικα**	**Άνω μετωπιαίαέλικα**
Κατώτερος βρεγματικός λοβός	0	0.238	0.344	0.370
Προσφηνοειδές λοβίο	0.275	0	0.347	0.377
Μέση μετωπιαία έλικα	0.357	0.360	0	0.381
Άνω μετωπιαίαέλικα	0.544	0.508	0.457	0

4.2 Στατιστικός έλεγχος

Στη συνέχεια εφαρμόστηκετο μη παραμετρικό κριτήριοWilcoxon για ζευγαρωτές παρατηρήσεις (εξαρτημένα δείγματα) για κάθε ένα από τα αντίστοιχα ζευγάρια περιοχών του πίνακα βαθμού συγχρονισμού για τα 18 άτομα και ελέγξαμε εάν διαφέρει ο βαθμός συγχρονισμού ανά 2 συναισθηματικές κατηγορίες, δηλαδή ευχάριστες-δυσάρεστες, δυσάρεστες-ουδέτερες και ευχάριστες-ουδέτερες. Βρέθηκαν στατιστικά σημαντικές διαφορές στο βαθμό συγχρονισμού των ζευγών των συναισθηματικών κατηγοριών όπως φαίνεται στον παρακάτω πίνακα:

Πίνακας 2. *Λίστα των στατιστικά σημαντικών ζευγών των περιοχών ανά ζεύγος βαθμού ευχαρίστησης*

Ζεύγος κατηγοριών βαθμού ευχαρίστησης	*Ζεύγος περιοχών του προτεινόμενου μοντέλου με βάση τις συντεταγμένες Talairach*
Υψηλός-Χαμηλός (HV-LV)	**Κάτω βρεγματικός λοβός-Προσφηνοειδές λοβίο** **P-value=0,022<0,05** **Άνω μετωπιαία έλικα- Μέση μετωπιαία έλικα** **P-value=0,011<0,05** **Προσφηνοειδές λοβίο- Κάτω βρεγματικός λοβός** **P-value=0,020<0,05**
Χαμηλός-Ουδέτερος (LV-NV)	**Κάτω βρεγματικός λοβός-Άνω μετωπιαία έλικα** **P-value=0,043<0,05** **Μέση μετωπιαία έλικα- Κάτω βρεγματικός λοβός** **P-value=0,025<0,05** **Άνω μετωπιαία έλικα- Κάτω βρεγματικός λοβός** **P-value=0,039<0,05** **Άνω μετωπιαία έλικα- Μέση μετωπιαία έλικα** **P-value=0,039<0,05**

5.ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα εργασία τα αποτελέσματα της μεθόδου που ακολουθήσαμε για την εύρεση των εγκεφαλικών περιοχών που ενεργοποιούνται κατά την παθητική θέαση οπτικοποιημένων λεκτικών ερεθισμάτων συμφωνούν με τα αποτελέσματα της προγενέστερης μεθοδολογίας προηγούμενων μελετών (Bedo, 2012; Klepousniotou et al., 2014). Το εργαλείο dipfit2 που χρησιμοποιήθηκε αποτελεί έναν αξιόπιστο τρόπο εντοπισμού των εγκεφαλικών περιοχών. Αυτό μπορούμε να το διαπιστώσουμε γιατί οι εγκεφαλικές περιοχές που βρέθηκαν σχετίζονται με την επεξεργασία των οπτικοποιημένων λέξεων σε σύγκριση με άλλες νευροαπεικονιστικές μελέτες (Bullmoreetal, 1996; Klepousniotou et al., 2014; Bedo 2012). Todipfit2, αποτελεί μια καινούρια μέθοδο η οποία είναι πολλά υποσχόμενη στην επιστημονική κοινότητα εφόσον μπορεί να ανιχνεύσει με αξιοπιστία ταηλεκτρικά δίπολα που δημιουργούνται από τα σημειακά φορτία. Κάθε ανεξάρτητη συνιστώσα αντιστοιχίζεται σε ένα δίπολο που προσδιορίζει με σχετικά μεγάλη ευκολία και εγκυρότητα την εγκεφαλική περιοχή που ενεργοποιείται κατά τη διάρκεια εμφάνισης λέξεων.

Η εύρεση των εγκεφαλικών περιοχών μπορεί να συμβάλλει στη διάγνωση πιθανής μελλοντικής εμφάνισης ασθενειών του εγκεφάλου στους συμμετέχοντες όπως της επιληψίας, διαταραχών του ύπνου, εγκεφαλικών αλλά και της νόσου Alzheimer και έτσι υπάρχει η

δυνατότητα πρόληψης ούτως ώστε να μειωθούν οι πιθανότητες εμφάνισής τους. Όσον αφορά την ιατρική έρευνα, η μελέτη των ενεργοποιημένων εγκεφαλικών περιοχών συμβάλλει στην «αποκρυπτογράφηση» των διεργασιών του εγκεφάλου παρέχοντας ένα μεγάλο όγκο πληροφοριών ο οποίος μπορεί να αξιοποιηθεί κατάλληλα.

Επίσης, στην εργασία αυτή, έχει συνδυαστεί η ICAμε μια τεχνική εκτίμησης συγχρονισμού που είναι ένας καινοτόμος συνδυασμός γιατί ανιχνεύει πέρα από τις εγκεφαλικές περιοχές το πώς λειτουργεί η κάθε μια από αυτές κατά τη διάρκεια της παθητικής θέασης των λεκτικών ερεθισμάτων του πειράματος.

Η σημαντικότερη αδυναμία της ICAοφείλεται στον περιορισμό της μεθόδου που ορίζει ότι όσα είναι τα σήματα (ηλεκτρόδια) τόσες θα είναι οι ανεξάρτητες συνιστώσες/πηγές (ενεργοποιημένες εγκεφαλικές περιοχές). Έτσιη τοποθέτηση 57 ηλεκτροδίων στην κεφαλή κάθε συμμετέχοντα είχε ως αποτέλεσμα την ανίχνευση μεγάλου αριθμού (57) ανεξάρτητωνσυνιστωσών.

Ένα από τα μειονεκτήματα της μελέτης αυτής, που αφορά το dipfit2 και όχι τόσο την εφαρμογή της μεθόδου ICA οφείλεται στην αδυναμία χωρικής ανάλυσης εστιασμού του μοντέλου και σχετίζεται με τηντοποθεσία του διπόλου που δεν είναι στο κέντρο μιας συγκεκριμένης εγκεφαλικής περιοχής. Για παράδειγμα, αν η εγκεφαλική περιοχή είναι σε μια έλικα, τότε η ισοδύναμη θέση του διπόλου θα είναι βαθύτερα της περιοχής αυτής (http://sccn.ucsd.edu/wiki/A08:_DIPFIT) .

Επομένως, το dipfit2 του EEGLAB ναι μεν είναι ένα αξιόλογο εργαλείο το οποίο αντιστοιχεί σε αρκετά ενδιαφέροντα συμπεράσματα για τις ενεργοποιημένες εγκεφαλικές περιοχές κατά τη διάρκεια εμφάνισης ερεθίσματος, αλλά δεν παύει δε να μη βρίσκει με ακρίβεια τη θέση των διπόλων που ευθύνονται για τις δραστηριότητες των εγκεφαλικών πηγών. Μια μελλοντική εργασία στοχεύει στην εύρεση πιο ακριβών εργαλείων, με εντοπισμό μεγαλύτερου αριθμού διπόλων που θα οδηγούν σε ακόμη πιο ισχυρά συμπεράσματα όχι μόνο από στατιστικής αλλά και από νευροφυσιολογικής άποψης.

ABSTRACT

The Independent Component Analysis / ICA is a method of finding components derived from multiple variables. ICA differs from other methodologies since its components do not follow normal (Gaussian distribution). The aim of the present study is to find dipole locations corresponding to active brain regions and then to estimate their co-operative activity. The adopted experimental procedure employs passive viewing of affective word stimuli. The study employed 21 healthy volunteer. The stimuli are classified into positive, negative and neutral. Electroencephalographic (EEG) activity was performed and employed event-related potentials, time-locked to the stimulus onset. ICA, was initially used to reject artifactual componentns (e.g. muscles, blinks, linear trends). Then, the EEGLAB graphic user interface was used to find dipole locations, which correspond to brain regions activated during the experimental procedure. A proposed model based on previous state of the art findings and validated by our data was proposed. The brain regions proposed by this model were identified in each participant and then their synchronization degree was estimated through the notion of the Relative Wavelet Entropy (RWE). Finally, the synchronization values were statistically analyzed through the non-parametric Wilcoxon test. The analysis demonstrated statistically significant synchronization differences due to differences in their valence degree.

*Keywords:*ICA.,independent components, affective word stimuli, dipoles, dipfit2, relative wavelet entropy, synchronization degree, brain regions

ΑΝΑΦΟΡΕΣ

Bamidis, P. D., Konstantinidis, E., Billis, A., Frantzidis, C., Tsolaki, M., Hlauschek, W., &Pattichis, C. S. (2011, August). A Web services-based exergaming platform for senior

citizens: the long lasting memories project approach to e-health care. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE* (pp. 2505-2509).

Bedo, N. (2012). Connectivity in cortical networks during word reading.

Bullmore, E. T., Rabe-Hesketh, S., Morris, R. G., Williams, S. C. R., Gregory, L., Gray, J. A., &Brammer, M. J. (1996). Functional magnetic resonance image analysis of a large-scale neurocognitive network.*NeuroImage*, *4*(1), 16-33.

Frantzidis, C. A., Bratsas, C., Papadelis, C. L., Konstantinidis, E., Pappas, C., &Bamidis, P. D. (2010).Toward emotion aware computing: an integrated approach using multichannel neurophysiological recordings and affective visual stimuli. IEEE Trans InfTechnol Biomed 14(3), 589-597.

González-Palau, F., Franco, M., Bamidis, P., Losada, R., Parra, E., Papageorgiou, S. G., &Vivas, A. B. (2014).The effects of a computer-based cognitive and physical training program in a healthy and mildly cognitive impaired aging sample.*Aging & mental health*, *18*(7), 838-846.

Klepousniotou, E., Gracco, V. L., & Pike, G. B. (2014). Pathways to lexical ambiguity: fMRI evidence for bilateral fronto-parietal involvement in language processing. *Brain and language*, *131*, 56-64.

Rosso, O. A., Blanco, S., Yordanova, J., Kolev, V., Figliola, A., Schürmann, M., &Başar, E. (2001). Wavelet entropy: a new tool for analysis of short duration brain electrical signals. *Journal of neuroscience methods*, *105*(1), 65-75.

http://sccn.ucsd.edu/wiki/A08:_DIPFIT

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής (2015) σελ.81-92

ΒΕΛΤΙΣΤΗ ΔΙΑΝΟΜΗ ΠΟΛΛΩΝ ΠΡΟΪΟΝΤΩΝ ΜΕ ΣΥΝΕΧΕΙΣ ΚΑΤΑΝΟΜΕΣ ΖΗΤΗΣΕΩΝ

***Θ. Δ. Δημητράκος*[1], *Ε. Γ. Κυριακίδης*[2]**
[1]*Τμήμα Μαθηματικών, Πανεπιστήμιο Αιγαίου*
dimitheo@aegean.gr
[2]*Τμήμα Στατιστικής, Οικονομικό Πανεπιστήμιο Αθηνών*
ekyriak@aueb.gr

ΠΕΡΙΛΗΨΗ

Θεωρούμε το πρόβλημα της βέλτιστης δρομολόγησης ενός οχήματος που διανέμει K προϊόντα σε N πελάτες. Υποθέτουμε ότι οι ζητήσεις των πελατών για κάθε προϊόν είναι συνεχείς τυχαίες μεταβλητές που ακολουθούν γνωστές κατανομές. Οι πελάτες εξυπηρετούνται από το όχημα σύμφωνα με μία προκαθορισμένη σειρά. Κάθε προϊόν τοποθετείται στο δικό του τμήμα στο όχημα. Το όχημα ξεκινάει τη διαδρομή του από μία αποθήκη με μία συγκεκριμένη ποσότητα από το κάθε προϊόν. Υπάρχει ένα κόστος για την μετάβαση του οχήματος από τον έναν πελάτη προς τον επόμενο πελάτη. Μετά από την εξυπηρέτηση του κάθε πελάτη, το όχημα έχει δύο επιλογές: (α) να ταξιδέψει προς τον επόμενο πελάτη ή (β) να επιστρέψει στην αποθήκη για να ανανεώσει το απόθεμά του σε κάθε προϊόν και να συνεχίσει τη διαδρομή του. Μετά από την εξυπηρέτηση όλων των πελατών θεωρούμε ότι το όχημα επιστρέφει στην αποθήκη. Υποθέτουμε ότι υπάρχει ένα κόστος για την επιστροφή του οχήματος στην αποθήκη. Αποδεικνύεται ότι, για κάθε πελάτη, η βέλτιστη πολιτική έχει μία συγκεκριμένη δομή τύπου κατωφλίου (threshold-type structure). Για την ειδική περίπτωση της διανομής δύο προϊόντων, μετά από κατάλληλη διακριτοποίηση του χώρου καταστάσεων, κατασκευάζουμε τον κλασσικό αλγόριθμο του δυναμικού προγραμματισμού και υπολογίζουμε εκείνη την πολιτική που ικανοποιεί τις ζητήσεις των πελατών για κάθε προϊόν ελαχιστοποιώντας το συνολικό αναμενόμενο κόστος. Σχεδιάζουμε επίσης έναν αλγόριθμο ειδικού σκοπού (special-purpose algorithm) για τον υπολογισμό της βέλτιστης πολιτικής ο οποίος είναι σημαντικά ταχύτερος του κλασσικού αλγορίθμου του δυναμικού προγραμματισμού. Παρουσιάζουμε ένα αριθμητικό παράδειγμα στο οποίο οι ζητήσεις των πελατών για κάθε προϊόν είναι ανεξάρτητες συνεχείς τυχαίες μεταβλητές που ακολουθούν γνωστές κατανομές.

Λέξεις-Κλειδιά: υπηρεσίες εφοδιασμού, δυναμικός προγραμματισμός, δρομολόγηση οχήματος με χωριζόμενα φορτία προϊόντων, διανομή πολλών προϊόντων, συνεχείς κατανομές ζητήσεων, αλγόριθμος ειδικού σκοπού

1. ΕΙΣΑΓΩΓΗ

Σε ένα πρόβλημα βέλτιστης δρομολόγησης ενός οχήματος με στοχαστικές ζητήσεις, ο στόχος είναι να βρεθεί εκείνη η διαδρομή που ελαχιστοποιεί το συνολικό αναμενόμενο κόστος για ένα όχημα που έχει πεπερασμένη χωρητικότητα και ξεκινά από μία αποθήκη διανέμοντας προϊόντα σε N πελάτες. Οι ζητήσεις για τα προϊόντα είναι είτε διακριτές είτε συνεχείς τυχαίες μεταβλητές οι οποίες ακολουθούν γνωστές κατανομές. Πολλοί ακριβείς και ευρετικοί αλγόριθμοι, έχουν σχεδιαστεί για διάφορες μορφές αυτού του προβλήματος. Για παράδειγμα, ενδεικτικά αναφέρουμε τις εργασίες των Mendoza et al. (2010), Goodson et al. (2012) και Marinakis et al. (2013). Η μέθοδος του δυναμικού προγραμματισμού έχει προταθεί και έχει εφαρμοσθεί σε διάφορα προβλήματα βέλτιστης δρομολόγησης ενός οχήματος με στοχαστικές διακριτές ζητήσεις στα οποία οι πελάτες εξυπηρετούνται από το όχημα σύμφωνα με μία προκαθορισμένη σειρά. Ενδεικτικά αναφέρουμε τις εργασίες των Yang et al. (2000), Minis and Tatarakis (2011) και Pandelis et al. (2012, 2013a, 2013b).

Οι Pandelis et al. (2012) μελέτησαν ένα στοχαστικό πρόβλημα βέλτιστης δρομολόγησης ενός οχήματος με χωριζόμενα φορτία προϊόντων υποθέτοντας ότι οι ζητήσεις των πελατών είναι διακριτές τυχαίες μεταβλητές. Στην παρούσα εργασία τροποποιούμε το πρόβλημα που μελέτησαν οι Pandelis et al. υποθέτοντας ότι οι ζητήσεις των πελατών είναι συνεχείς τυχαίες μεταβλητές. Στη συνέχεια, περιγράφουμε το τροποποιημένο πρόβλημα για την περίπτωση κατά την οποία μόνο δύο προϊόντα διανέμονται στους πελάτες. Τα αποτελέσματα μπορούν να επεκταθούν και για τις περιπτώσεις κατά τις οποίες το όχημα διανέμει $K > 2$ προϊόντα, αν και οι απαιτούμενοι υπολογισμοί είναι αρκετά πιο απαιτητικοί.

Θεωρούμε ένα σύνολο κορυφών $V = \{0, 1, \ldots, N\}$ με την κορυφή 0 να αναπαριστά την αποθήκη και τις κορυφές $1, \ldots, N$ να αντιστοιχούν στους πελάτες. Οι πελάτες εξυπηρετούνται σύμφωνα με την προκαθορισμένη σειρά $1, 2, \ldots, N$ από ένα όχημα που αποτελείται από $K = 2$ τμήματα με χωρητικότητες ίσες με Q_1 και Q_2, αντίστοιχα. Σε κάθε τμήμα του οχήματος τοποθετείται μόνο ένα είδος προϊόντος. Υπάρχουν δύο διαφορετικά είδη προϊόντων που διανέμονται στους πελάτες. Το όχημα ξεκινά τη διαδρομή του από την αποθήκη φορτωμένο με ποσότητες προϊόντων ίσες με $Q_i, i \in \{1, 2\}$ και μετά την εξυπηρέτηση όλων των πελατών επιστρέφει στην αποθήκη. Η ζήτηση του πελάτη $j, j \in \{1, \ldots, N\}$ για το προϊόν i είναι μία συνεχής τυχαία μεταβλητή ξ_i^j. Θεωρούμε ότι η ζήτηση του κάθε πελάτη γίνεται γνωστή με την άφιξη του οχήματος σε κάθε πελάτη. Υποθέτουμε ότι η ζήτηση κάθε πελάτη για ένα προϊόν δεν μπορεί να ξεπερνά την χωρητικότητα κάθε τμήματος του οχήματος, δηλαδή ισχύει ότι $\max_{j=1,2,\ldots,N} \xi_i^j \le Q_i$, για κάθε $i \in \{1, 2\}$. Όταν το όχημα επισκέπτεται τον πελάτη j για πρώτη φορά ικανοποιεί όση ζήτηση είναι δυνατόν να ικανοποιηθεί. Αν μέρος της ζήτησης για το προϊόν $i \in \{1, 2\}$ δεν ικανοποιείται, το όχημα πηγαίνει

προς την αποθήκη, συμπληρώνει τα τμήματά του με ποσότητες προϊόντων ίσες με Q_1 και Q_2 και επιστρέφει στον πελάτη για να ικανοποιήσει τη ζήτηση. Μετά την εξυπηρέτηση του τελευταίου πελάτη, το όχημα επιστρέφει στην αποθήκη. Έστω $c_{j,j+1},\ j=1,2,\ldots,N-1$, το κόστος για την μετάβαση του οχήματος από τον πελάτη j προς τον πελάτη $j+1$ και έστω $c_{j0},\ j=1,2,\ldots,N,$ το κόστος για την μετάβαση του οχήματος από τον πελάτη j προς την αποθήκη. Είναι λογικό να υποθέσουμε ότι αυτά τα κόστη είναι συμμετρικά και ικανοποιούν την τριγωνική ιδιότητα, δηλαδή ισχύει ότι:

$$c_{i0}=c_{0i}, i=1,\ldots,N \text{ και } c_{i,i+1}\le c_{i0}+c_{0,i+1},\ i=1,\ldots,N-1.$$

Έστω $z_i\in[0,Q_i], i\in\{1,2\},$ η ποσότητα του προϊόντος i που έχει απομείνει στο όχημα μετά την εξυπηρέτηση της ζήτησης ενός πελάτη. Τότε, το όχημα, είτε (i) πηγαίνει κατευθείαν προς τον επόμενο πελάτη ή (ii) πηγαίνει προς την αποθήκη, ανανεώνει το απόθεμά του με ποσότητες προϊόντων 1 και 2 ίσες με Q_1 και Q_2, αντίστοιχα και μετά επισκέπτεται τον επόμενο πελάτη. Ο στόχος είναι να προσδιοριστεί μία στρατηγική δρομολόγησης του οχήματος η οποία να ελαχιστοποιεί το αναμενόμενο κόστος κατά τη διάρκεια μίας ολοκληρωμένης διαδρομής του οχήματος. Όπως αναφέρεται στην εργασία των Pandelis et al. (2012), ένα ρεαλιστικό παράδειγμα εφαρμογής αυτού του μοντέλου μπορεί να είναι η περίπτωση κατά την οποία το όχημα διανέμει δύο διαφορετικά είδη πετρελαίου σε μία σειρά από σταθμούς πετρελαίου. Το όχημα έχει δύο τμήματα και σε κάθε τμήμα τοποθετείται μόνο ένα συγκεκριμένο είδος πετρελαίου. Οι Kyriakidis and Dimitrakos (2008) μελέτησαν αυτό το πρόβλημα στην περίπτωση κατά την οποία μόνο ένα προϊόν διανέμεται στους πελάτες. Οι Dimitrakos and Kyriakidis (2015) μελέτησαν το πρόβλημα βέλτιστης διανομής πολλών προϊόντων και συλλογής ληγμένων προϊόντων θεωρώντας συνεχείς κατανομές ζητήσεων. Για την ειδική περίπτωση των δύο προϊόντων, σχεδίασαν έναν αλγόριθμο ειδικού σκοπού για τον υπολογισμό της βέλτιστης πολιτικής.

Ο υπολογισμός της βέλτιστης στρατηγικής δρομολόγησης του οχήματος επιτυγχάνεται με τη χρήση ενός κατάλληλου αλγορίθμου του δυναμικού προγραμματισμού όπως στην περίπτωση των διακριτών ζητήσεων που μελετήθηκε στην εργασία των Pandelis et al. (2012). Στο παρόν πρόβλημα όμως, ο χώρος καταστάσεων μετά την πρώτη επίσκεψη του οχήματος σε κάθε πελάτη είναι ένα συνεχές σύνολο. Για το λόγο αυτό, προτείνουμε μία κατάλληλη διακριτοποίηση του χώρου καταστάσεων ώστε να μπορεί να εφαρμοστεί η μέθοδος του δυναμικού προγραμματισμού.

Στο επόμενο εδάφιο παρουσιάζουμε τη μορφή της βέλτιστης στρατηγικής δρομολόγησης του οχήματος και δείχνουμε πως μπορεί να υπολογισθεί σχεδιάζοντας έναν αλγόριθμο δυναμικού προγραμματισμού ειδικού σκοπού. Στο Εδάφιο 3 παρουσιάζουμε ένα αριθμητικό παράδειγμα στο οποίο οι ζητήσεις των πελατών για

κάθε προϊόν είναι ανεξάρτητες συνεχείς τυχαίες μεταβλητές που ακολουθούν γνωστές κατανομές.

2. Η ΒΕΛΤΙΣΤΗ ΠΟΛΙΤΙΚΗ

Ακολουθώντας την ίδια προσέγγιση με αυτή των Pandelis et al. (2012) για δύο προϊόντα, ορίζουμε τα διανύσματα $\bar{z} = (z_1, z_2)$, $\overline{Q} = (Q_1, Q_2)$ και θεωρούμε την από κοινού συνάρτηση πυκνότητας πιθανότητας $g^j(\bar{\xi})$, $\bar{\xi} = (\xi_1, \xi_2) \in S$, όπου $S = [0, Q_1] \times [0, Q_2]$, για τη ζήτηση του κάθε πελάτη $j \in \{1, \ldots, N\}$. Υποθέτουμε ότι η από κοινού συνάρτηση πυκνότητας πιθανότητας $g^j(\bar{\xi})$ είναι γνωστή. Έστω $f_j(\bar{z})$ το ελάχιστο αναμενόμενο μελλοντικό κόστος όταν η ποσότητα του προϊόντος i που έχει απομείνει στο όχημα μετά την ικανοποίηση της ζήτησης του πελάτη j είναι ίση με z_i. Τότε, η βέλτιστη στρατηγική δρομολόγησης του οχήματος, σε αυτό το τροποποιημένο μοντέλο μπορεί να προσδιοριστεί από τις ακόλουθες εξισώσεις δυναμικού προγραμματισμού. Για $j = 1, 2, \ldots, N-1$ έχουμε:

$$f_j(\bar{z}) = \min\{H_j(\bar{z}), A_j\}, \bar{z} = (z_1, z_2) \in [0, Q_1] \times [0, Q_2], \tag{1}$$

όπου,

$$\begin{aligned} H_j(\bar{z}) = {} & c_{j,j+1} + \int_0^{z_1}\int_0^{z_2} f_{j+1}(z_1 - \xi_1, z_2 - \xi_2) g^{j+1}(\xi_1, \xi_2) d\xi_2 d\xi_1 \\ & + \int_{z_1}^{Q_1}\int_{z_2}^{Q_2} [2c_{j+1,0} + f_{j+1}(Q_1 + (z_1 - \xi_1)^-, Q_2 + (z_2 - \xi_2)^-)] g^{j+1}(\xi_1, \xi_2) d\xi_2 d\xi_1 \\ & + \int_0^{z_1}\int_{z_2}^{Q_2} [2c_{j+1,0} + f_{j+1}(Q_1 + (z_1 - \xi_1)^-, Q_2 + (z_2 - \xi_2)^-)] g^{j+1}(\xi_1, \xi_2) d\xi_2 d\xi_1 \\ & + \int_{z_1}^{Q_1}\int_0^{z_2} [2c_{j+1,0} + f_{j+1}(Q_1 + (z_1 - \xi_1)^-, Q_2 + (z_2 - \xi_2)^-)] g^{j+1}(\xi_1, \xi_2) d\xi_2 d\xi_1 \end{aligned} \tag{2}$$

και

$$A_j = c_{j0} + c_{j+1,0} + \int_0^{Q_1}\int_0^{Q_2} f_{j+1}(Q_1 - \xi_1, Q_2 - \xi_2) g^{j+1}(\xi_1, \xi_2) d\xi_2 d\xi_1. \tag{3}$$

Η οριακή συνθήκη είναι:

$$f_N(\bar{z}) = c_{N0}, \bar{z} = (z_1, z_2) \in S. \quad (4)$$

Ο αριστερός όρος ανάμεσα στα άγκιστρα στην Εξίσωση (1) αντιστοιχεί στην ενέργεια του απευθείας ταξιδιού του οχήματος προς τον επόμενο πελάτη και ο δεξιός όρος αντιστοιχεί στην ενέργεια της επιστροφής του οχήματος στην αποθήκη για ανεφοδιασμό. Στην Εξίσωση (2) οι ποσότητες $(z_1 - \xi_1)^-$ και $(z_2 - \xi_2)^-$ ορίζονται ως $\min(z_1 - \xi_1, 0)$ και $\min(z_2 - \xi_2, 0)$, αντίστοιχα.

Το ελάχιστο συνολικό αναμενόμενο κόστος κατά τη διάρκεια μιας ολοκληρωμένης διαδρομής του οχήματος είναι:

$$f_0 = c_{10} + \int_0^{Q_1} \int_0^{Q_2} f_1(Q_1 - \xi_1, Q_2 - \xi_2) g^1(\xi_1, \xi_2) d\xi_2 d\xi_1.$$

Μπορεί να αποδειχθεί με παρόμοιο τρόπο όπως στο Θεώρημα 1 των Pandelis et al. (2012) ότι η βέλτιστη πολιτική για αυτό το πρόβλημα πεπερασμένου χρονικού ορίζοντα έχει την ίδια δομή τύπου κατωφλίου (threshold-type structure) με το αντίστοιχο πρόβλημα πεπερασμένου χρονικού ορίζοντα για την περίπτωση των διακριτών ζητήσεων. Το ακόλουθο θεώρημα χαρακτηρίζει τη μορφή της βέλτιστης στρατηγικής δρομολόγησης του οχήματος μετά την ικανοποίηση της ζήτησης του πελάτη j.

Θεώρημα 1. Για κάθε ποσότητα z_1 του προϊόντος 1, υπάρχει μία κρίσιμη ποσότητα $s_j(z_1) \in [0, Q_2]$ του προϊόντος 2 τέτοια ώστε η βέλτιστη απόφαση για το όχημα είναι να ταξιδέψει κατευθείαν προς τον επόμενο πελάτη $j+1$ αν η ποσότητα z_2 του προϊόντος 2 ικανοποιεί την ανισότητα $z_2 \geq s_j(z_1)$. Επιπλέον, η κρίσιμη ποσότητα $s_j(z_1)$ είναι μη-αύξουσα ως προς z_1.

Ο χώρος καταστάσεων μετά την πρώτη επίσκεψη του οχήματος στον πελάτη $j \in \{1, \ldots, N\}$ είναι το σύνολο:

$$S = \{(z_1, z_2) : z_1 \in [0, Q_1], z_2 \in [0, Q_2]\}.$$

Παρατηρούμε ότι οι μεταβλητές z_1 και z_2 είναι συνεχείς μεταβλητές που παίρνουν τιμές στα διαστήματα $[0, Q_1]$ και $[0, Q_2]$, αντίστοιχα. Μία διακριτοποίηση του χώρου καταστάσεων είναι απαραίτητη για την εφαρμογή του αλγορίθμου του δυναμικού προγραμματισμού. Έστω ρ ένας σχετικά μικρός αριθμός (π.χ. $\rho = 0.05$ ή $\rho = 0.01$). Διακριτοποιούμε το σύνολο S περιοριζόμενοι μόνο στα σημεία του $(k\rho, l\rho), k, l = 0, \ldots Q/\rho$.

Το ελάχιστο αναμενόμενο κόστος $f_N(k\rho,l\rho),0\le k,l\le Q/\rho,$ υπολογίζεται από την Εξίσωση (4) με $z_1=k\rho$ και $z_2=l\rho.$ Το ελάχιστο αναμενόμενο κόστος $f_j(k\rho,l\rho),0\le k,l\le Q/\rho$ και η αντίστοιχη βέλτιστη απόφαση υπολογίζονται αναδρομικά, για $j=N-1,\ldots,1$ χρησιμοποιώντας την εξίσωση δυναμικού προγραμματισμού (1) με $z_1=k\rho$ και $z_2=l\rho.$ Τα διπλά ολοκληρώματα στις Εξισώσεις (2) και (3) υπολογίζονται προσεγγιστικά. Για παράδειγμα, η ποσότητα $H_j(k\rho,l\rho)$ υπολογίζεται ως εξής:

$$H_j(k\rho,l\rho)=c_{j,j+1}+\sum_{x=0}^{k\rho}\sum_{y=0}^{l\rho}f_{j+1}(k\rho-x\rho,l\rho-y\rho)g^{j+1}(x\rho,y\rho)\rho^2$$
$$+\sum_{x=k\rho+\rho}^{Q_1/\rho-1}\sum_{y=l\rho+\rho}^{Q_2/\rho-1}[2c_{j+1,0}+f_{j+1}(Q_1+(k\rho-x\rho)^-,Q_2+(l\rho-y\rho)^-)]g^{j+1}(x\rho,y\rho)\rho^2$$
$$+\sum_{x=0}^{k\rho}\sum_{y=l\rho+\rho}^{Q_2/\rho-1}[2c_{j+1,0}+f_{j+1}(Q_1+(k\rho-x\rho)^-,Q_2+(l\rho-y\rho)^-]g^{j+1}(x\rho,y\rho)\rho^2$$
$$+\sum_{x=k\rho+\rho}^{Q_1/\rho-1}\sum_{y=0}^{l\rho}[2c_{j+1,0}+f_{j+1}(Q_1+(k\rho-x\rho)^-,Q_2+(l\rho-y\rho)^-]g^{j+1}(x\rho,y\rho)\rho^2.$$

Σύμφωνα με το Θεώρημα 1, η βέλτιστη πολιτική, δηλαδή οι κρίσιμες ποσότητες $s_j(k\rho),0\le k\le Q_1/\rho,$ για κάθε πελάτη $j\in\{1,\ldots,N-1\},$ μπορούν να υπολογιστούν από τον ακόλουθο αλγόριθμο δυναμικού προγραμματισμού ειδικού σκοπού (special-purpose dynamic programming algorithm).

Αλγόριθμος ειδικού σκοπού για τον προσδιορισμό των κρίσιμων ποσοτήτων $s_j(k\rho),0\le k\le Q_1/\rho$ για κάθε πελάτη $j\in\{1,\ldots,N-1\}$

Βήμα 0. Θέτουμε $f_N(k\rho,l\rho)=c_{N0},$ για $k=0,\ldots,Q_1/\rho,\ \ l=0,\ldots,Q_2/\rho.$

Θέτουμε $j=N-1.$

Βήμα 1. Υπολογίζουμε την ποσότητα A_j από την Εξίσωση (3). Θέτουμε $k\rho=0.$

Βήμα 2. Αν $A_j\le H_j(k\rho,Q_2)$ τότε θέτουμε $s_j(k\rho)=Q_2/\rho+1.$

και $f_j(k\rho,l\rho)=A_j,\ 0\le l\le Q_2/\rho,$

διαφορετικά, πηγαίνουμε στο Βήμα 3.

Βήμα 3. Αν $H_j(k\rho,0)\le A_j$ τότε θέτουμε $s_j(k\rho)=0$

και $f_j(k\rho,l\rho)=H_j(k\rho,l\rho),\ 0\le l\le Q_2/\rho,$

διαφορετικά, υπολογίζουμε την ποσότητα $H_j(k\rho,l\rho)$

για $l=Q_2/\rho-1,Q_2/\rho-2,\ldots$ μέχρι $A_j\le H_j(k\rho,l\rho).$

Θέτουμε $s_j(k\rho)=l\rho+\rho$

και $f_j(k\rho,l\rho)=H_j(k\rho,l\rho),\ s_j(k\rho)\le l\rho\le Q_2$

και $f_j(k\rho,l\rho)=A_j,\ 0\le l\rho\le s_j(k\rho)-\rho.$

Βήμα 4. Θέτουμε $k\rho=k\rho+\rho.$ Αν $k\rho\le Q_1,$ πηγαίνουμε στο Βήμα 2.

Βήμα 5. Θέτουμε $j=j-1.$ Αν $j\ge 1,$ πηγαίνουμε στο Βήμα 1. Διαφορετικά, σταματάμε.

Ο παραπάνω αλγόριθμος είναι σημαντικά ταχύτερος από τον κλασσικό αλγόριθμο του δυναμικού προγραμματισμού που βασίζεται στις Εξισώσεις (1)-(4) διότι, για $j=1,\ldots,N-1$ και $0\le k\le Q_1/\rho,$ δεν απαιτείται να υπολογιστούν οι ποσότητες $H_j(k\rho,l\rho),\quad 0\le l\rho<s_j(k\rho)-\rho.$ Στο επόμενο εδάφιο παρουσιάζουμε ένα αριθμητικό παράδειγμα.

3. ΑΡΙΘΜΗΤΙΚΟ ΠΑΡΑΔΕΙΓΜΑ

Υποθέτουμε ότι $N=10$ και $Q_1=Q_2=8.$ Τα κόστη μεταξύ των διαδοχικών πελατών j και $j+1,\ j=1,\ldots,9,$ είναι: $c_{12}=32,\ c_{23}=34,\ c_{34}=38,\ c_{45}=28,$ $c_{56}=26,\ c_{67}=24,\ c_{78}=18,\ c_{89}=28$ και $c_{9,10}=36.$ Τα κόστη μεταξύ των πελατών $j,\ j=1,\ldots,10,$ και της αποθήκης είναι: $c_{10}=18,\ c_{20}=14,\ c_{30}=22,$ $c_{40}=18,\ c_{50}=12,\ c_{60}=16,\ c_{70}=20,\ c_{80}=24,\ c_{90}=26$ και $c_{10,0}=22.$ Παρατηρούμε ότι αυτά τα κόστη ικανοποιούν την τριγωνική ιδιότητα. Υποθέτουμε ότι, για κάθε πελάτη $j\in\{1,\ldots,10\}$ οι ζητήσεις ξ_1^j και ξ_2^j για τα προϊόντα 1 και 2 είναι ανεξάρτητες συνεχείς τυχαίες μεταβλητές που ακολουθούν τη δεξιά-αποκοπτώμενη (right-truncated) κατανομή Gamma στα διαστήματα $[0,Q_1]$ και $[0,Q_2],$ αντίστοιχα. Οι συναρτήσεις πυκνότητας πιθανότητας είναι:

$$h^j(x)=[F(Q_1)]^{-1}\frac{\lambda_1 x^{\alpha_1-1}}{\Gamma(\alpha_1)}e^{-\lambda_1 x},\ x\in[0,Q_1]\ \text{και}$$

$$w^j(y)=[F(Q_2)]^{-1}\frac{\lambda_2 y^{\alpha_2-1}}{\Gamma(\alpha_2)}e^{-\lambda_2 y},\ y\in[0,Q_2],$$

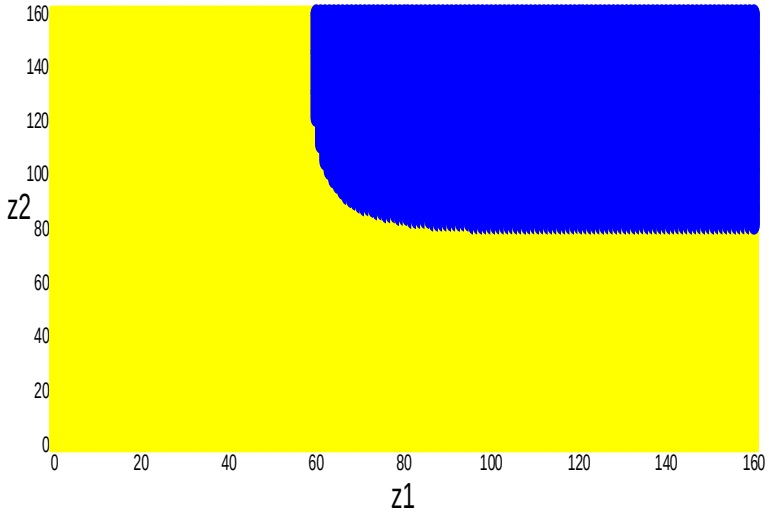
αντίστοιχα, όπου, $\alpha_i, \lambda_i > 0, i \in \{1,2\}$, $\Gamma(\alpha_i) = \int_0^\infty e^{-u} u^{\alpha_i - 1} du, a_i > 0, i \in \{1,2\}$ και $F_i(x) = [\Gamma(\alpha_i)]^{-1} \int_0^{\lambda_i x} e^{-u} u^{\alpha_i - 1} du$, $x \geq 0$. Η κατανομή Gamma φαίνεται να είναι μία λογική επιλογή για τις ζητήσεις για τα προϊόντα 1 και 2 διότι, όπως αναφέρεται στη σελ. 442 στο βιβλίο του Tijms (2003), σε εφαρμογές διαχείρισης αποθεμάτων (inventory applications) η κατανομή Gamma χρησιμοποιείται συχνά για τη μοντελοποίηση κατανομών ζήτησης. Όπως αναφέρεται στις σελ. 85-88 του Κεφ. 5 στο βιβλίο του Axsater (2006), εναλλακτικές λογικές επιλογές συνεχών κατανομών ζητήσεων είναι η κανονική κατανομή, η κατανομή Weibull, η κατανομή Log-Normal καθώς και μίξεις της κατανομής Erlang.

Λόγω της ανεξαρτησίας των τυχαίων μεταβλητών ξ_1^j και ξ_2^j ισχύει ότι $g^j(x,y) = h^j(x)w^j(y)$. Θέτουμε $\alpha_1 = 5, \lambda_1 = 4$ και $\alpha_2 = 3, \lambda_2 = 2$. Επιλέγουμε $\rho = 0.05$ έτσι ώστε ο διακριτοποιημένος χώρος καταστάσεων μετά την πρώτη επίσκεψη του οχήματος σε κάθε πελάτη είναι το σύνολο:

$$\{(k * 0.05, l * 0.05) : k, l = 0, \ldots, 160\}.$$

Εφαρμόζουμε τον κλασσικό αλγόριθμο του δυναμικού προγραμματισμού που βασίζεται στις Εξισώσεις (1)-(4) και τον αλγόριθμο δυναμικού προγραμματισμού ειδικού σκοπού εκτελώντας τα σχετικά προγράμματα σε Matlab σε προσωπικό υπολογιστή τύπου Intel Core i5-3230M, 2.6 GHz και 4 GB RAM. Στο Σχήμα 1, παρουσιάζουμε τις βέλτιστες αποφάσεις μετά την πρώτη επίσκεψη στον Πελάτη 3. Η βέλτιστη πολιτική, όπως αναμενόταν, είναι τύπου κατωφλίου όπως περιγράφεται στο Θεώρημα 1. Η ενέργεια του απευθείας ταξιδιού του οχήματος στον επόμενο πελάτη χρωματίζεται με βαθύ μπλε και η ενέργεια της επιστροφής του οχήματος στην αποθήκη για ανεφοδιασμό με ποσότητες προϊόντων 1 και 2, ίσες με Q_1 και Q_2, αντίστοιχα, χρωματίζεται με κίτρινο.

Σχήμα 1. *Οι βέλτιστες αποφάσεις μετά την πρώτη επίσκεψη στον πελάτη 3*



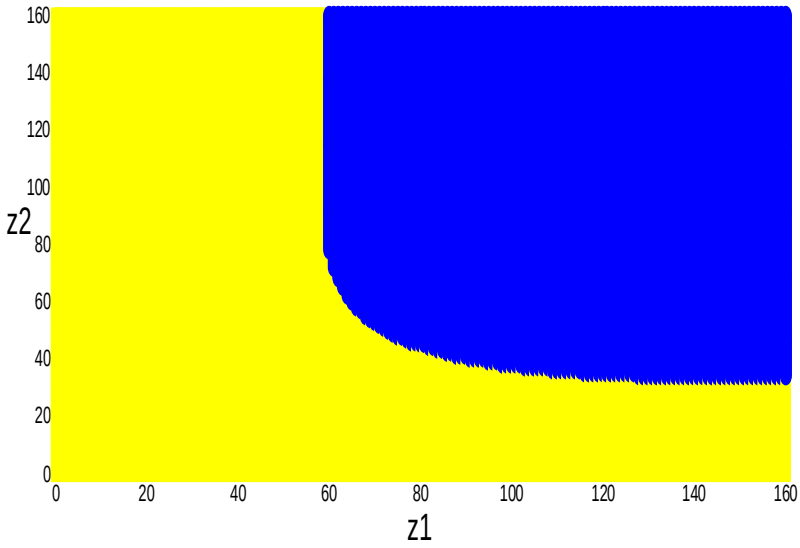
Η τιμή του ελάχιστου συνολικού αναμενόμενου κόστους f_0 βρέθηκε κατά προσέγγιση ίση με 318.11. Ο υπολογιστικός χρόνος για την εκτέλεση του αλγορίθμου ειδικού σκοπού είναι 116.75 δευτερόλεπτα. Είναι σημαντικά μικρότερος από τον αντίστοιχο υπολογιστικό χρόνο για την εκτέλεση του κλασσικού αλγορίθμου του δυναμικού προγραμματισμού που είναι 710.56 δευτερόλεπτα. Θεωρούμε τον ίδιο αριθμό πελατών και τα ίδια κόστη μετάβασης. Υποθέτουμε ότι, για κάθε πελάτη j, οι ζητήσεις ξ_1^j και ξ_2^j για τα προϊόντα 1 και 2 είναι ανεξάρτητες συνεχείς τυχαίες μεταβλητές που ακολουθούν τη διπλά-αποκοπτώμενη (double-truncated) κανονική κατανομή στα διαστήματα $[0,Q_1]$ και $[0,Q_2]$, αντίστοιχα.
Οι συναρτήσεις πυκνότητας πιθανότητας είναι:

$$h^j(x)=[F(Q_1)-F(0)]^{-1}\frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}},\ x\in[0,Q_1]$$

και $w^j(y)=[F(Q_2)-F(0)]^{-1}\frac{1}{\sigma_2\sqrt{2\pi}}e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}},\ y\in[0,Q_2],$

αντίστοιχα, όπου, $F(x)=\frac{1}{\sigma_i\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{(t-\mu_i)^2}{2\sigma_i^2}}dt,\quad x\in\Re,$ είναι η συνάρτηση κατανομής της κανονικής κατανομής $N(\mu_i,\sigma_i^2),\ \mu_i\in\Re,\ \sigma_i>0, i\in\{1,2\}$. Θέτουμε $\mu_1=3,\sigma_1=2$ και $\mu_2=2,\sigma_2=1.$ Επιλέγουμε πάλι $\rho=0.05.$ Στο Σχήμα 2, παρουσιάζονται οι βέλτιστες αποφάσεις μετά την πρώτη επίσκεψη στον Πελάτη 7.

Σχήμα 2. *Οι βέλτιστες αποφάσεις μετά την πρώτη επίσκεψη στον πελάτη 7*



Ο χρόνος εκτέλεσης του αλγορίθμου ειδικού σκοπού είναι 72.26 δευτερόλεπτα, σημαντικά μικρότερος του αντίστοιχου χρόνου εκτέλεσης του κλασσικού αλγορίθμου που είναι 779.44 δευτερόλεπτα. Η τιμή του ελάχιστου συνολικού αναμενόμενου κόστους f_0 βρέθηκε κατά προσέγγιση ίση με 363.04. Από ένα μεγάλο αριθμό αριθμητικών παραδειγμάτων, υπάρχει ισχυρή ένδειξη ότι οι υπολογιστικοί χρόνοι των δύο αλγορίθμων αυξάνονται μη-γραμμικά καθώς οι παράμετροι Q_1 και Q_2, αυξάνονται. Η διαφορά στους υπολογιστικούς χρόνους μεταξύ των δύο αλγορίθμων γίνεται σημαντική καθώς οι ποσότητες Q_1 και Q_2 αυξάνονται. Ο απαιτούμενος υπολογιστικός χρόνος για την εκτέλεση του αλγορίθμου ειδικού σκοπού είναι σημαντικά μικρότερος από τον αντίστοιχο απαιτούμενο υπολογιστικό χρόνο για την εκτέλεση του κλασσικού αλγορίθμου, ειδικά για υψηλές τιμές των παραμέτρων Q_1 και Q_2. Για παράδειγμα, αν θεωρήσουμε τον ίδιο αριθμό πελατών, τα ίδια κόστη μετάβασης, τις ίδιες δεξιά αποκοπτώμενες κατανομές ζητήσεων Gamma για τα δύο προϊόντα και επιλέξουμε $Q_1 = Q_2 = 16,$ τότε ο απαιτούμενος υπολογιστικός χρόνος για την εκτέλεση του αλγορίθμου ειδικού σκοπού είναι 3254.4 δευτερόλεπτα ενώ ο αντίστοιχος απαιτούμενος υπολογιστικός χρόνος για την εκτέλεση του κλασσικού αλγορίθμου που βασίζεται στις Εξισώσεις (1)-(4) είναι 25632.6 δευτερόλεπτα.

ABSTRACT

This paper extends the results of a particular capacitated single vehicle routing problem with compartmentalized load (see Pandelis et al. (2012)) to the case in which the demands for K products that are delivered to N customers are continuous random variables instead of discrete ones. The customers are served according to a particular order. The optimal policy that serves all customers has a specific threshold-type structure. This policy minimizes the total expected cost from the beginning of the route until its end. For the special case of two products, after an appropriate discretization of the state space, the optimal policy is computed by a suitable efficient special-purpose dynamic programming algorithm that operates over all policies having this structure. The computation time of the special-purpose algorithm is considerably smaller than the corresponding computation time of the standard dynamic programming algorithm. The structural result for the form of the optimal policy is illustrated by a numerical example in which the demands of the customers for the products are independent continuous random variables which follow known distributions.

Keywords: logistics, dynamic programming, vehicle routing with compartmentalized load, multiple-product delivery, continuous random demands, special-purpose algorithm

ΑΝΑΦΟΡΕΣ

Axsater S. (2006). Inventory Control, 2nd Edition, Springer, New York.

Dimitrakos T.D. and Kyriakidis E.G. (2015). A single vehicle routing problem with pickups and deliveries, continuous random demands and predefined customer order, *European Journal of Operational Research* **244**, 990-993.

Goodson J.C., Ohlmann J.W. and Thomas B.W. (2012). Cyclic-order neighborhoods with application to the vehicle routing problem with stochastic demand, *European Journal of Operational Research* **217**, 312-323.

Kyriakidis E.G. and Dimitrakos T.D. (2008). Single vehicle routing problem with a predefined customer sequence and stochastic continuous demands, *Mathematical Scientist* **33**, 148-152.

Marinakis Y., Iordanidou G.-R. and Marinaki M. (2013). Particle swarm optimization for the vehicle routing problem with stochastic demands, *Applied Soft Computing* **13**, 1693-1704.

Mendoza J.E., Castanier B., Gueret C., Medaglia A.L., Velasco N. (2010). A memetic algorithm for the multi-compartment vehicle routing problem with stochastic demands, *Computers and Operations Research* **37** (11), 1886-1898.

Minis I. and Tatarakis A. (2011). Stochastic single vehicle routing problem with delivery and pickup and a predefined customer sequence, *European Journal of Operational Research* **213**, 37-51.

Pandelis D.G., Kyriakidis E.G. and Dimitrakos T.D. (2012). Single vehicle routing problems with a predefined customer sequence, compartmentalized load and stochastic demands, *European Journal of Operational Research* **217**, 324-332.

Pandelis D.G., Karamatsoukis C.C. and Kyriakidis E.G. (2013a). Single vehicle routing problems with a predefined customer order, unified load and stochastic discrete demands, *Probability in the Engineering and Informational Sciences* **27** (1), 1-23.

Pandelis D.G., Karamatsoukis C.C. and Kyriakidis E.G. (2013b). Finite and infinite-horizon single vehicle routing problems with a predefined customer sequence and pickup and delivery, *European Journal of Operational Research* **231**, 577-586.

Tijms H. C. (2003). A First Course in Stochastic Models, Wiley, Chichester.

Yang W.-H., Mathur K. and Ballou R.H. (2000). Stochastic vehicle routing problem with restocking, *Transportation Science* **34**, 99-112.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής (2015), σελ.93-107

ΔΙΑΔΙΚΤΥΑΚΟΣ ΕΘΙΣΜΟΣ ΚΑΙ ΜΑΘΗΤΕΣ ΛΥΚΕΙΩΝ: Η ΠΕΡΙΠΤΩΣΗ ΤΩΝ ΜΑΘΗΤΩΝ ΤΗΣ ΠΕΡΙΦΕΡΕΙΑΚΗΣ ΕΝΟΤΗΤΑΣ ΚΑΒΑΛΑΣ

Ε. Δημητριάδης[1], Ε. Βαχλιώτη[2]
[1]Τμήμα Διοίκησης Επιχειρήσεων, Τ.Ε.Ι Ανατολικής Μακεδονίας και Θράκης
edimit@teiemt.gr, leonidaslena@yahoo.gr

ΠΕΡΙΛΗΨΗ

Το διαδίκτυο προσφέρει άπειρες δυνατότητες στην καθημερινότητα του ανθρώπου, στην επικοινωνία, στην εκπαίδευση και στην επιστήμη. Ταυτόχρονα πρόσφατες έρευνες επισημαίνουν και πολλές αρνητικές συνέπειες από την χρήση του. Η παρούσα εργασία έχει ως αντικείμενο το φαινόμενο του εθισμού των εφήβων στο διαδίκτυο, ενός νεοεμφανιζόμενου τύπου εθισμού, συνεπεία της χρήσης των νέων τεχνολογιών. Σκοπό της εργασίας αποτέλεσε η διερεύνηση του επιπέδου εθισμού των μαθητών των λυκείων. Έρευνα η οποία πραγματοποιήθηκε σε δείγμα 726 μαθητών Λυκείων της περιφερειακής ενότητας της Καβάλας έδειξε ότι οι μαθητές είναι αρκετά εξοικειωμένοι με τη χρήση του διαδικτύου. Ο κύριος λόγος που χρησιμοποιούν το διαδίκτυο είναι η επικοινωνία, η αναζήτηση πληροφοριών και λιγότερο η αγορά των προϊόντων. Ενθαρρυντικό είναι το γεγονός ότι έξι στους δέκα μαθητές έχουν τον πλήρη έλεγχο της χρήσης του διαδικτύου και μόνο ένας στους εκατό παρουσιάζει σοβαρά προβλήματα εθισμού στο διαδίκτυο. Το σημαντικότερο πρόβλημα το οποίο παρουσιάζουν οι μαθητές, αφορά την έλλειψη αυτοελέγχου και την παραμέληση των εργασιών τους (διάβασμα).

Λέξεις Κλειδιά: Διαδίκτυο, Εθισμός, Έφηβοι, Καβάλα

1. ΕΙΣΑΓΩΓΗ

Το 1995 οι χρήστες του διαδικτύου ήταν λιγότεροι από το 1% του πληθυσμού της γης. Σήμερα, περίπου 198 χώρες παγκοσμίως κάνουν χρήση του διαδικτύου και σχεδόν το 40% του παγκόσμιου πληθυσμού έχει μια σύνδεση στο διαδίκτυο. (Internet World Stats, 2014). Στην Ελλάδα το 59,9% των πολιτών κάνει χρήση του διαδικτύου, ενώ περίπου ίδιο είναι το ποσοστό των χρηστών στην περιφέρεια Ανατολικής Μακεδονίας και Θράκης στην οποία ανήκει η Περιφερειακή Ενότητα Καβάλας (Ελληνική Στατιστική Αρχή, 2014)

Το διαδίκτυο κρίνεται ως το πιο αναγκαίο μέσο για την ψυχαγωγία, την ενημέρωση και την επικοινωνία του ατόμου. Οι Έλληνες, σε ποσοστό 81% κάνουν καθημερινή χρήση του διαδικτύου για πληροφορίες σχετικές με προϊόντα και υπηρεσίες (76,7%), ηλεκτρονικό ταχυδρομείο (74,6%), ανάγνωση ηλεκτρονικών περιοδικών/εφημερίδων (70,5%), παιχνίδια, φωτογραφίες, ταινίες, μουσική (63,9%), συνομιλίες (49,7%), εκπαίδευση/κατάρτιση (29,4%), τραπεζικές συναλλαγές (21,7%) κ.λ.π. (Κοινωνία της Πληροφορίας, 2014).

Λόγω της ευκολίας πρόσβασης με χαμηλό κόστος (Paul και Bryant, 2005) και της δυνατότητας χρήσης πολυποίκιλων δραστηριοτήτων (Dannon και Iancu, 2007), το διαδίκτυο έχει καταστεί ιδιαίτερα δημοφιλές μεταξύ των εφήβων οι οποίοι αφιερώνουν πολύ μεγάλο μέρος του χρόνου τους συνδεδεμένοι σε αυτό, έχοντας ως δεδομένο την οικειότητα που έχουν αναπτύξει με τις τεχνολογίες (Suss, 2007). Η χρήση του διαδικτύου από τους εφήβους επιδρά θετικά στην μείωση του άγχους τους (Valaitis, 2005), στην κοινωνικοποίησή τους και στον σχηματισμό της προσωπικής τους ταυτότητας (Bradley, 2005). Ωστόσο, η υπερβολική χρήση του διαδικτύου από τους εφήβους αποτελεί σήμερα ένα νέο κοινωνικό φαινόμενο το οποίο αποκτά παγκόσμιες διαστάσεις και αναγνωρίζεται ως ψυχική διαταραχή.

2. ΕΘΙΣΜΟΣ ΚΑΙ ΔΙΑΔΙΚΤΥΟ

Σύμφωνα με το Marks (1990), ο εθισμός αποτελεί μια έλξη η οποία έχει κατεύθυνση προς την ικανοποίηση ενός συναισθήματος. Κάποιοι ερευνητές κάνουν λόγο για δύο τύπους εθισμών: (1) Τον εθισμό που οφείλεται σε ουσίες όπως το κάπνισμα, το αλκοόλ, τα ναρκωτικά και (2) Τον συμπεριφορικό εθισμό, όπως τα τυχερά παιχνίδια, οι δαπάνες, τα ψώνια, το φαγητό, η σεξουαλική δραστηριότητα και το διαδίκτυο (Encyclopedia of Children's Health, 2014).

Ο εθισμός στο Διαδίκτυο αποτελεί ένα κοινωνικό θέμα που αποκτά παγκόσμιες διαστάσεις και απαντάται σχεδόν σε όλα τα ηλικιακά και κοινωνικά στρώματα. Κατά τους Widyanto και Griffiths (2006), ορίζεται ως συμπεριφοριστικός ή μη χημικός εθισμός που περικλείει την αλληλεπίδραση μηχανής-ανθρώπου. Οι εξαρτήσεις αυτές μπορεί να είναι ενεργητικές, όπως να συνομιλεί κάποιος σε chat room ή παθητικές, όπως να παρακολουθεί κάποιος μια ταινία. Ένας πιο εξειδικευμένος ορισμός αποδίδει τον Διαδικτυακό εθισμό στην έλλειψη ικανότητας του ανθρώπου να ελέγξει την χρήση του Διαδικτύου, η οποία δημιουργεί ψυχολογικές, σχολικές, κοινωνικές και πιθανόν επαγγελματικές δυσκολίες στη ζωή του (Davis, 2001; Shapira et al., 2000; Young, 1998). Σύμφωνα με τους Young και Abreu (2011), η τυπική περιγραφή του εθισμού στο διαδίκτυο δίδεται ως το σημείο όπου το άτομο χάνει την ικανότητα να ελέγχει την χρήση του διαδικτύου και δεν μειώνει την αυξημένη χρήση του μέχρι να εμφανισθούν προβληματικές συμπεριφορές οι οποίες έχουν αρνητικό αντίκτυπο στην ζωή του. Το 1995, ο ψυχίατρος Ivan Goldberg χρησιμοποιώντας εν μέρει χιούμορ, είναι ο πρώτος που ανέφερε τον όρο «εθισμός στο διαδίκτυο» (internet addiction). Ένα χρόνο μετά η ψυχολόγος Kimberly Young καθιερώνει την έκφραση «εθισμός στο διαδίκτυο» (Παναγοπούλου, 2011; Καράπετσας *et al*., 2012). Παρ' όλα αυτά, υποστηρίζεται από πολλούς ερευνητές ότι μόνο σε περιπτώσεις χρήσης ουσιών μπορεί να γίνει χρήση αυτού του όρου (Rachlin, 1990; Walker, 1989). Η άποψη τους

δεν επικρατεί και έτσι ο όρος εθισμένος μπορεί να χαρακτηρίσει άτομα που παρουσιάζουν παθολογική σχέση με τα ηλεκτρονικά παιχνίδια, την παρακολούθηση τηλεοπτικών εκπομπών, τα τυχερά παιχνίδια, τη σωματική άσκηση, την υπερφαγία και τις ερωτικές σχέσεις (Καράπετσας *et al*., 2012).

Σύμφωνα με τους Dannon και Iancu (2007) υπάρχουν πέντε ειδικοί τύποι εθισμού στο διαδίκτυο: (1) Εθισμός στο διαδικτυακό σεξ και την πορνογραφία, (2) Εθισμός στις διαδικτυακές σχέσεις, (3) Καθαροί καταναγκασμοί, (4) Υπέρμετρη αναζήτηση πληροφοριών και (5) Εθισμός στους ηλεκτρονικούς υπολογιστές.

Τα αίτια που οδηγούν στον διαδικτυακό εθισμό αποτελούνται από ένα κράμα αιτιών τόσο του ψυχοκοινωνικού μοντέλου όσο και της εθιστικής φύσης του ίδιου του διαδικτύου. Πιο συγκεκριμένα, ένα πλήθος διαταραχών εκτιμάται ότι συνδέεται με την υπέρμετρη χρήση του διαδικτύου. Αυξημένη συχνότητα παρουσιάζει η συννοσηρότητα με διαταραχές της διάθεσης, ιδιαίτερα διπολικού τύπου και η κοινωνική φοβία (Caplan *et al.,* 2009; Morahan- Martin, 2005; Shapira *et al*., 2000). Εδώ θα πρέπει να σημειωθεί ότι ένας μεγάλος αριθμός ατόμων που παρουσιάζουν διαδικτυακό εθισμό, ικανοποιούν τα κριτήρια εξαρτησιογόνων διαταραχών από ψυχοτρόπες ουσίες (Anderson, 2001; Bai *et al*., 2001). Διάφοροι παράγοντες σχετικοί με τον εθισμό στο διαδίκτυο έχουν προσδιορισθεί όπως η ελλειμματική προσοχή και υπερκινητικότητα (Yen *et al.,* 2007), η κατάθλιψη (Morrison και Gore, 2010), η μοναξιά (Moody, 2001; Yao και Zhong, 2014), η χαμηλή αυτοεκτίμηση (Niemz *et al.,*2005), η μειωμένη φυσική υγεία (Kelley και Gruber, 2013) και η αλεξιθυμία (Dalbudak *et al*., 2013; De Berardis *et al.,* 2009). Ο κυριότερος ψυχολογικός παράγοντας για τον διαδικτυακό εθισμό είναι η ανάγκη που νιώθει το άτομο να μειώσει τα έντονα αρνητικά συναισθήματα που νιώθει (Morahan- Martin, 2005).

Το γεγονός ότι μπορεί ένα άτομο να εκφράσει τον πραγματικό του εαυτό μέσω του διαδικτύου καθίσταται ιδιαίτερο ελκυστικό για έναν έφηβο καθώς αυτή η περίοδος της ζωής του έχει ως χαρακτηριστικό της γνώρισμα την ανάπτυξη της ταυτότητάς του όπως επίσης ποικίλα ερωτήματα αυτογνωσίας (Tosun και Lajunen, 2009). Ελκυστική για τους εφήβους καθίσταται η «ανωνυμία» που χαρακτηρίζει το διαδίκτυο καθώς έχουν την δυνατότητα να συμμετέχουν σε δραστηριότητες που θα ήταν αδύνατον ή μη προσβάσιμες για αυτούς στον πραγματικό κόσμο (Cao και Su, 2007). Αυτό έχει σαν αποτέλεσμα οι έφηβοι να παρουσιάζουν μεγαλύτερη δεκτικότητα να διαπληκτιστούν ή να παρενοχλήσουν άλλα άτομα στο διαδίκτυο, να μελετήσουν ποικίλες σεξουαλικές συμπεριφορές και να είναι προσβάσιμοι σε πορνογραφικό υλικό (Dowell *et al.* 2009).

Παρ' όλα αυτά, η χρήση του σε καθημερινή βάση εγκυμονεί τον κίνδυνο παρουσίασης αρνητικών συνεπειών (Bremer, 2005; Suss, 2007; Yan, 2006) οι οποίες πιθανόν να οφείλονται σε παλαιότερη απουσία ψυχοκοινωνικής λειτουργικότητας (Caplan, 2007). Ποικίλα ερεθίσματα που ασκούν πίεση στους εφήβους είναι ικανά να τους ωθήσουν στην υπέρμετρη ενασχόληση τους με το διαδίκτυο (Lam et al. 2009). Οι έφηβοι, σε αντίθεση με τους ενήλικες, παρουσιάζουν αυξημένη πιθανότητα υιοθέτησης συμπεριφορών υπέρμετρης χρήσης του διαδικτύου, γεγονός που οφείλεται στην ευπάθεια που παρουσιάζει η διαδικασία της ενηλικίωσης τους (Pallanti *et al,.* 2006; Leung, 2007; Shapira *et al.,* 2000). Η μη φυσιολογική χρήση

συσχετίζεται με την διάρκεια και την συχνότητα, ιδιαιτέρως όταν επικεντρώνεται σε διαδικτυακά παιχνίδια ή επαφή μέσω ηλεκτρονικών μηνυμάτων και δωματίων συνομιλίας (Chak και Leung, 2004). Σοβαρές συνέπειες είναι οι μη λειτουργικές διαπροσωπικές σχέσεις (Suhail και Bargees, 2006; Borzekowski, 2006) και η μειωμένη σχολική επίδοση (Ng και Wiemer-Hastigs, 2005). Θετικός συσχετισμός παρουσιάζεται μεταξύ διαδικτυακής εθιστικής συμπεριφοράς των εφήβων και Διαταραχή Ελλειμματικής Προσοχής-Υπεραντιδραστικότητας, καθώς επίσης και με το υψηλό ποσοστό πιθανής εμφάνισης αυτοκτονικού ιδεασμού ή κατάθλιψης (Ha *et al.* 2006; Kim *et al*., 2008).

Υπερβολική χρήση του Η/Υ και του διαδικτύου παρατηρείται περισσότερο στους εφήβους που έχουν περιορισμένη συναισθηματική στήριξη από τους γονείς τους και λιγότερο σε εκείνους που έχουν ικανοποιητική συναισθηματική στήριξη. Η υπερβολική ενασχόληση με το διαδίκτυο είναι εντονότερη στα αγόρια σε σχέση με τα κορίτσια ενώ οι έφηβοι που κάνουν υπερβολική χρήση Η/Υ και διαδικτύου έχουν σχεδόν σε διπλάσιο ποσοστό ψυχοκοινωνικές δυσκολίες σε σύγκριση με τους εφήβους που δεν ασχολούνται το ίδιο συχνά με τον Η/Υ (ΕΠΙΨΥ, 2012). Είναι αξιοσημείωτο ότι το μεγαλύτερο ποσοστό εθισμένων εφήβων στο διαδίκτυο το παρουσιάζουν οι μαθητές της Τεχνολογικής κατεύθυνσης, ακολουθούμενοι από τους μαθητές της θετικής και τέλος της θεωρητικής κατεύθυνσης. Επίσης το μεγαλύτερο ποσοστό αυτών ανήκει στην κατηγορία των «καλών» μαθητών ενώ ακολουθούν οι «άριστοι» μαθητές και τέλος οι «μέτριοι» μαθητές. Επίσης η συντριπτική πλειοψηφία των μαθητών οι οποίο χαρακτηρίζονται ως εθισμένοι στο διαδίκτυο διαθέτει οικιακή σύνδεση (Σοφός *et al.,* 2011). Οι έφηβοι που ασχολούνται με τυχερά παιχνίδια έχουν τριπλάσια πιθανότητα να αναπτύξουν δυσλειτουργική διαδικτυακή συμπεριφορά σε σύγκριση με αυτούς που δεν το ασχολούνται ενώ οι έφηβοι που παίζουν ηλεκτρονικά/διαδικτυακά παιχνίδια έχουν διπλάσια πιθανότητα ανάπτυξης δυσλειτουργικής διαδικτυακής συμπεριφοράς (EU NET ADB, 2013).

3. ΕΡΕΥΝΗΤΙΚΗ ΜΕΘΟΔΟΛΟΓΙΑ

Βασικό στόχο της εργασίας αποτέλεσε η μέτρηση του επιπέδου εθισμού των μαθητών της περιφερειακής ενότητας Καβάλας στο διαδίκτυο. Επιπρόσθετο στόχο αποτέλεσε και η αποτύπωση της συμπεριφοράς των μαθητών με βάση διάφορα χαρακτηριστικά όπως το φύλο και ο τύπος του Λυκείου στο οποίο φοιτούν καθώς επίσης και η εξοικείωσή τους με τη χρήση του διαδικτύου. Για την υλοποίηση των στόχων της εργασίας πραγματοποιήθηκε έρευνα, στους μαθητές των Λυκείων, με τη χρήση δομημένου ερωτηματολογίου.

3.1 Πληθυσμός και Δείγμα

Η έρευνα πραγματοποιήθηκε τους μήνες Απρίλιο και Μάιο του 2014 και πληθυσμό της έρευνας αποτέλεσαν οι μαθητές και οι μαθήτριες όλων των Λυκείων (ΓΕΛ και ΕΠΑΛ) της περιφερειακής ενότητας Καβάλας. Το σύνολο των λυκείων ανέρχεται σε 19 και κατανέμεται σε 14 ΓΕΛ και 5 ΕΠΑΛ. Οι μαθητές/τριες οι οποίοι φοιτούν σε ΓΕΛ είναι 3021 και σε ΕΠΑΛ 1286. Από αυτούς οι 2251 είναι αγόρια και

2056 κορίτσια. Για την συλλογή των δεδομένων χρησιμοποιήθηκε ερωτηματολόγιο το οποίο αφού εγκρίθηκε από την Δευτεροβάθμια Εκπαίδευση νομού Καβάλας και κατόπιν άδειας διανομής και συμπλήρωσης του από τον Προϊστάμενο Εκπαίδευσης, μοιράστηκε στους μαθητές και συμπληρώθηκε από τους ιδίους κατά την διάρκεια μιας διδακτικής ώρας. Τα δεδομένα τα οποία συλλέχθηκαν καταχωρήθηκαν στο στατιστικό πακέτο S.P.S.S. 20. Το τελικό δείγμα αποτελείται από 726 μαθητές Λυκείων το 51,3% των οποίων προέρχεται από ΕΠΑΛ και το 48,7% από ΓΕΛ. Ειδικότερα, το 33,5% των μαθητών των ΕΠΑΛ προέρχεται από Λύκεια της Καβάλας και το 66,5% από Λύκεια της επαρχίας. Το 33% των μαθητών των ΓΕΛ προέρχεται από Λύκεια της Καβάλας και το υπόλοιπο 67% από Λύκεια της επαρχίας. Στο σύνολο των μαθητών το 59,4% είναι αγόρια και το 40,6% κορίτσια, ενώ στα ΕΠΑΛ τα αγόρια εκπροσωπούν το 70,% των μαθητών και στα ΓΕΛ το 47,6%. Η συντριπτική πλειοψηφία των μαθητών (71,6%) είναι ηλικίας μεταξύ 16 και 17 ετών, ενώ το 27% είναι ηλικίας μεταξύ 18 και 19 ετών. Ο τόπος μόνιμης κατοικίας του 62% των μαθητών είναι αγροτική περιοχή, του 13,9% ημιαστική περιοχή και του 23,9% αστική περιοχή. Στο 12,8% των μαθητών ο πατέρας είναι στοιχειώδους εκπαίδευσης, στο 61,5% μέσης εκπαίδευσης και το σ25,7% ανώτατης εκπαίδευσης. Όσον αφορά τις μητέρες των μαθητών το 10,6% είναι στοιχειώδους εκπαίδευσης, το 61,4% μέσης εκπαίδευσης και το 28% ανώτατης εκπαίδευσης. Ποσοστό 14,2% των μαθητών έχει πατέρα άνεργο, το 39% εργάζεται 8 ώρες ενώ σημαντικό ποσοστό (34%) εργάζεται πλέον των 8 ωρών. Το 29,6% των μητέρων των μαθητών είναι άνεργες, το 33% εργάζεται 8 ώρες ενώ αξιοσημείωτο ποσοστό μητέρων (15%) εργάζεται περισσότερες από 8 ώρες. Οι μαθητές, σε ποσοστό 49,3% ασχολούνται με διάφορα αθλήματα, το 33% δεν ασχολείται με τίποτα πέραν του Λυκείου, το 22% μαθαίνει ξένες γλώσσες, το 10,6% ασχολείται με το χορό και το 7,6% πηγαίνει σε ωδείο.

3.2 Ερευνητικό Εργαλείο

Για την συλλογή των απαραίτητων πληροφοριών χρησιμοποιήθηκε ως ερευνητικό εργαλείο δομημένο ερωτηματολόγιο αποτελούμενο από τρία μέρη.
Το πρώτο μέρος έχουμε δημογραφικά χαρακτηριστικά όπως το φύλο, η ηλικία, ο τόπος μόνιμης κατοικίας καθώς και κάποιες σχετικές πληροφορίες όπως το μορφωτικό επίπεδο των γονέων κ.α.

Στο δεύτερο μέρος, με τρεις ερωτήσεις οι οποίες υιοθετήθηκαν από την εργασία των (Maenpaa et al., 2008; Castaneda et al, 2007), γίνεται προσπάθεια προσδιορισμού του επιπέδου εξοικείωσης των μαθητών με το διαδίκτυο. Το ελάχιστο σκορ είναι 3 και το μέγιστο 11. Όσο μεγαλύτερο είναι το σκορ τόσο πιο εξοικειωμένοι με το διαδίκτυο είναι οι χρήστες. Επιπλέον οι συνήθεις λόγοι χρήσεις του διαδικτύου, από τους μαθητές, αποτυπώνονται με μία ερώτηση πολλαπλών επιλογών.

Το τρίτο και σημαντικότερο μέρος του ερωτηματολογίου αποτελείται από 20 ερωτήσεις στις οποίες οι μαθητές κλήθηκαν να αξιολογήσουν, σε μία πενταβάθμια κλίμακα Likert (1= Σπάνια, 2= Περιστασιακά, 3= Τακτικά, 4= Συχνά και 5= Πάντα), τον βαθμό κατά τον οποίο η χρήση του διαδικτύου επηρεάζει την καθημερινή τους ζωή. Το ελάχιστο σκορ είναι 20 μονάδες και το μέγιστο 100. Όσο υψηλότερο είναι το

σκορ τόσο περισσότερα είναι τα προβλήματα τα οποία δημιουργεί το διαδίκτυο στον χρήστη. Η Young (1996), αναφέρει ότι ένα σκορ από 20 έως 39 μονάδες δηλώνει ότι ο χρήστης έχει τον πλήρη έλεγχο της χρήσης, σκορ από 40 έως 69 μονάδες υποδηλώνει συχνά προβλήματα εξαιτίας της χρήσης του διαδικτύου, ενώ σκορ από 70 έως 100 μονάδες σημαίνει ότι η χρήση του διαδικτύου δημιουργεί σοβαρά προβλήματα στον χρήστη. Οι είκοσι ερωτήσεις που αποτελούν το εργαλείο ελέγχου του εθισμού στο διαδίκτυο (Internet Addiction Test) σχεδιάστηκαν από την Young (1996).

3.3 Επικύρωση και Αξιοπιστία Ερευνητικού Εργαλείου

Το πρώτο μέρος του ερωτηματολογίου αναφέρεται στα δημογραφικά χαρακτηριστικά των μαθητών και δεν απαιτεί ιδιαίτερο έλεγχο καθώς χρησιμοποιήθηκαν οι συνήθεις ερωτήσεις που προσδιορίζουν τα χαρακτηριστικά του δείγματος. Το δεύτερο μέρος αναφέρεται στην συχνότητα χρήσης του διαδικτύου και οι τρεις ερωτήσεις που υιοθετήθηκαν αποτελούν τον πλέον διαδεδομένο τρόπο μέτρησης της εξοικείωσης με οποιαδήποτε δραστηριότητα/ λειτουργία. Για την πιστοποίηση της εγκυρότητας του τρίτου μέρους του ερωτηματολογίου το οποίο μετρά το επίπεδο εθισμού στο διαδίκτυο πραγματοποιήθηκε: (1) έλεγχος εγκυρότητας περιεχομένου, (2) έλεγχος μονοδιάστατης δομής και (3) έλεγχος αξιοπιστίας.

Ο έλεγχος εγκυρότητας του περιεχομένου του παρόντος ερωτηματολογίου, το οποίο χρησιμοποιείται στις περισσότερες έρευνες παγκοσμίως καθώς τυγχάνει της εμπιστοσύνης των ερευνητών, πραγματοποιήθηκε με συζήτηση με ειδικούς και ακαδημαϊκούς του κλάδου. Στη συνέχεια υλοποιήθηκε πιλοτικό τεστ σε μικρό αριθμό μαθητών έτσι ώστε να διαπιστωθεί η λειτουργικότητά του καθώς και η κατανόηση των ερωτήσεων οι οποίες μεταφράστηκαν από την αγγλική γλώσσα και έγινε προσπάθεια να αποδοθούν σωστά όλοι οι όροι.

Για τον έλεγχο της μονοδιάστατης δομής των 20 ερωτήσεων πραγματοποιήθηκε Διερευνητική Παραγοντική Ανάλυση. Ο έλεγχος της καταλληλότητας των δεδομένων για παραγοντική ανάλυση βασίστηκε στον έλεγχο σφαιρικότητας του Barteltt η τιμή του οποίου (3110,532) είναι στατιστικά σημαντική (Sig.=0,000) και στον δείκτη Κ.Μ.Ο του οποίου η τιμή 0,876 είναι σημαντικά ανώτερη του ορίου 0,7 που προτείνουν οι Hair et al. (1995). Δημιουργήθηκαν έξι παράγοντες με ιδιοτιμή μεγαλύτερη της μονάδας οι οποίοι ερμηνεύουν το 54,5% της συνολικής διακύμανσης. Οι παράγοντες οι οποίοι προέκυψαν από την ανάλυση των δεδομένων του δείγματος συμπίπτουν με τους αντίστοιχους παράγοντες του αυθεντικού ερωτηματολογίου (Addiction Test) και είναι: (1) *Περίοπτη Θέση*, (2) Υ*περβολική Χρήση*, (3) *Παραμέληση Εργασίας*, (4) *Ανυπομονησία*, (5) *Έλλειψη Αυτό-Ελέγχου* και (6) *Παραμέληση Κοινωνικής Ζωής*. Οι φορτίσεις όλων των μεταβλητών στους αντίστοιχους παράγοντες κυμαίνονται από 0,436 έως 0,941 και είναι ικανοποιητικές καθώς υπερβαίνουν το ελάχιστο όριο του 0,35 για δείγματα μεγαλύτερα των 350 ατόμων (Hair et al., 1995). Τα αποτελέσματα της παραγοντικής ανάλυσης παρουσιάζονται στους πίνακες 1 και 2.

Πίνακας 1. *Παραγοντική Ανάλυση*

	Ερωτήσεις	Φορτίσεις	Παράγοντες
1	Επιλέγεις να ξοδέψεις χρόνο στο διαδίκτυο από το να βγεις έξω με άλλους;	0,667	**Περίοπτη Θέση 16,534%**
2	Εκνευρίζεσαι, φωνάζεις, ή συμπεριφέρεσαι ενοχλημένα αν κάποιος σε απασχολήσει την ώρα που είσαι στο διαδίκτυο;	0,587	
3	Φοβάσαι, ότι η ζωή χωρίς το διαδίκτυο θα ήταν βαρετή, κενή και δυσάρεστη;	0,697	
4	Σε απασχολεί το διαδίκτυο όταν δεν είσαι συνδεδεμένος, ή φαντασιώνεσαι ότι είσαι σε σύνδεση;	0,447	
5	Ξεχνάς ενοχλητικές σκέψεις για τη ζωή σου με καθησυχαστικές σκέψεις για το διαδίκτυο;	0,474	
6	Αμελείς δουλειές του δωματίου σου για να μείνεις περισσότερη ώρα στο διαδίκτυο;	0,471	**Υπερβολική Χρήση 9,744%**
7	Χάνεις ύπνο εξαιτίας συνδέσεων στο διαδίκτυο αργά το βράδυ;	0,645	
8	Νοιώθεις θλίψη, δυσαρέσκεια ή εκνευρισμό όταν είσαι εκτός διαδικτύου και παύεις να νιώθεις έτσι όταν επιστρέψεις στο διαδίκτυο;	0,484	
9	Αντιλαμβάνεσαι ότι βρίσκεσαι συνδεδεμένος στο διαδίκτυο περισσότερο από όσο είχες σκοπό;	0,469	
10	Προσπαθείς να κρύψεις πόση ώρα ήσουν στο διαδίκτυο;	0,667	
11	Επιβαρύνονται οι επιδόσεις (βαθμοί) στο σχολείο ή αλλού, εξαιτίας της ποσότητας του χρόνου που δαπανάς στο διαδίκτυο;	0,843	**Παραμέληση Εργασίας 9,203%**
12	Η απόδοσή σου στο σχολείο ή αλλού, επιβαρύνεται λόγω του διαδικτύου;	0,858	
13	Γίνεσαι αμυντικός ή μυστικοπαθής, όταν κάποιος σε ρωτήσει τι κάνεις στο διαδίκτυο ;	0,695	
14	Αντιλαμβάνεσαι τον εαυτό σου να ανυπομονεί, πότε θα ξανασυνδεθεί στο διαδίκτυο;	0,615	**Ανυπομονησία 8,128%**
15	Ελέγχεις το ηλεκτρονικό σου ταχυδρομείο, πριν κάνεις κάτι άλλο που χρειάζεται;	0,941	
16	Προσπαθείς να μειώσεις το χρόνο που είσαι στο διαδίκτυο;	0,744	**Έλλειψη Ελέγχου 5,924%**
17	Οι γύρω σου, σου παραπονιούνται για την ποσότητα του χρόνου που δαπανάς στο διαδίκτυο;	0,680	

18	Βρίσκεις τον εαυτό σου να λέει «λίγα λεπτά ακόμα» όταν είσαι σε σύνδεση;	0,547	
19	Συνάπτεις νέες σχέσεις με χρήστες του διαδικτύου σαν και εσένα;	0,436	**Παραμέληση Κοινωνικής Ζωής 5,307%**
20	Προτιμάς τον ενθουσιασμό του διαδικτύου από το να έρθεις κοντά με το αγόρι/κορίτσι σου;	0,732	
	Συνολική Ερμηνευμένη Διακύμανση		**54,84%**

Πίνακας 2. *KMO and Bartlett's Test*

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0,876
Bartlett's Test of Sphericity	Approx. Chi-Square	3110,532
	df	190
	Sig.	0,000

Ο έλεγχος αξιοπιστίας, ο οποίος μετρά την εσωτερική συνοχή των παραγόντων, πραγματοποιήθηκε με τη χρήση του δείκτη *a* Cronbach. Τιμές του δείκτη μεγαλύτερες του 0,7 χαρακτηρίζουν τον παράγοντα αξιόπιστο (Nunally, 1978). Τα αποτελέσματα της ανάλυσης των έξι παραγόντων έδωσαν για 4 παράγοντες οριακές τιμές ενώ για δύο παράγοντες οι τιμές είναι αισθητά ανώτερες του ορίου (Πίνακας 3).

Πίνακας 3. *Ανάλυση Αξιοπιστίας*

Παράγοντες	**Cronbach's alpha**
Περίοπτη Θέση	0,847
Υπερβολική Χρήση	0,800
Παραμέληση Εργασίας	0,713
Ανυπομονησία	0,701
Έλλειψη Αυτό-ελέγχου	0,711
Παραμέληση Κοινωνικής Ζωής	0,700

4. ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ- ΑΠΟΤΕΛΕΣΜΑΤΑ

4.1 Εξοικείωση με το Διαδίκτυο

Η δεύτερη ενότητα του ερωτηματολογίου αναφέρεται στην συχνότητα χρήσης (εξοικείωση) του διαδικτύου καθώς επίσης και στους λόγους για τους οποίους το χρησιμοποιούν συνήθως οι μαθητές. Από τα αποτελέσματα προκύπτει ότι το 18,3% των μαθητών το χρησιμοποιεί, όταν συνδέεται, λιγότερο από μια ώρα, οι μισοί περίπου (46%) από μία έως τρεις ώρες, ενώ σημαντικό είναι το ποσοστό (35,7%) αυτών που το χρησιμοποιεί περισσότερο από 3 ώρες. Το 66,9% των μαθητών χρησιμοποιεί το διαδίκτυο περισσότερο από τρία χρόνια, το 22,3% από ένα έως τρία χρόνια και ένα μικρό ποσοστό λιγότερο από ένα χρόνο. Η χρήση του διαδικτύου

φαίνεται ότι είναι για τους περισσότερους (74%) καθημερινή απασχόληση, το 11,7% αυτών το χρησιμοποιεί από 4 έως 6 φορές την εβδομάδα και μόνο το 2% μια φορά το μήνα.

Συνολικά, η εξοικείωση των μαθητών, όπως προέκυψε από την Ανάλυση Διακύμανσης (ANOVA), δεν παρουσιάζει στατιστικά σημαντική διαφορά η οποία θα μπορούσε να οφείλεται στον τύπο του Λυκείου ή στο φύλο του μαθητή καθώς σε κάθε περίπτωση έχουμε Sig. F>0,05.

Πίνακας 4. *Εξοικείωση στο Internet και: Τύπος Λυκείου και Φύλο*

	Τύπος Λυκείου		Φύλο	
	F	**Sig.**	**F**	**Sig.**
Εξοικείωση στο Internet	2,489	0,115	0,480	0,489

Κύρια χρήση του διαδικτύου αποτελεί η επικοινωνία (78,5%) και ακολουθούν η αναζήτηση πληροφοριών (61,3%), η ενημέρωση (59,1%), η διασκέδαση (50,8%) και η αγορά προϊόντων/ υπηρεσιών (37,5%). Γίνεται αντιληπτό από τον πίνακα 5.7 ότι ανεξαρτήτως τύπου λυκείου οι μαθητές έχουν ως κύριο λόγο χρήσης του διαδικτύου την επικοινωνία. Ωστόσο ο δεύτερος σπουδαιότερος λόγος για τους μαθητής των ΓΕΛ είναι η αναζήτηση πληροφοριών (71,7%) ενώ για τους μαθητές των ΕΠΑΛ η ενημέρωση (52,8%). Η αγορά προϊόντων και υπηρεσιών είναι ο σπανιότερος λόγος χρήσης και για τις δύο κατηγορίες λυκείων.

Σχετικά με το φύλο των μαθητών, διαπιστώνουμε ότι αγόρια και κορίτσια έχουν ως κύρια προτεραιότητα την επικοινωνία (76,6% και 81,4%) αντίστοιχα. Οι μαθήτριες ως δεύτερη προτεραιότητα έχουν την αναζήτηση πληροφοριών (65,1%), ενώ οι μαθητές την ενημέρωση (60,8%). Αξιοσημείωτο ότι το 44,1% των μαθητών χρησιμοποιεί το διαδίκτυο για αγορά προϊόντων/ υπηρεσιών, σε αντίθεση με τις μαθήτριες οι οποίες σε ποσοστό μόνο 27,8% το χρησιμοποιούν για την ίδια αιτία.

4.2 Επίπεδο Εθισμού στο Διαδίκτυο

Σύμφωνα με τα αποτελέσματα της ανάλυσης των δεδομένων, στο σύνολο των μαθητών, το 61,8% έχει τον πλήρη έλεγχο της χρήσης του διαδικτύου, το 36,8% παρουσιάζει συχνά προβλήματα εξαιτίας της χρήσης του διαδικτύου και μόνο το 1,4% παρουσιάζει σοβαρά προβλήματα (Πίνακας 5). Έρευνα που διεξήχθη το 2012 με συμμετοχή 7 κρατών της Ε.Ε. έδειξε ότι το ποσοστό των εφήβων που χαρακτηρίζονται εθισμένοι στην Ελλάδα ανέρχεται σε 1,7% και είναι το υψηλότερο μαζί με αυτό της Ρουμανίας (EU NET ADB, 2013).

Πίνακας 5. *Διαβάθμιση επιπέδου Εθισμού στο Διαδίκτυο*

Διαβάθμιση επιπέδου Εθισμού στο Διαδίκτυο	**Ποσοστό**	**Σκορ**
Πλήρη Έλεγχο	61,8%	20-39
Συχνά Προβλήματα	36,8%	40-69
Σοβαρά Προβλήματα	1,4%	70-100

Κατά τον έλεγχο της εγκυρότητας του ερωτηματολογίου προέκυψαν 6 παράγοντες οι οποίοι εκφράζουν τους τομείς στους οποίους επιδρά ενδεχομένως αρνητικά η χρήση του διαδικτύου. Από την ανάλυση των δεδομένων διαπιστώνουμε ότι οι παράγοντες *Περίοπτη Θέση* και *Παραμέληση Εργασίας* οριακά υπερβαίνουν το όριο των 39 μονάδων που δηλώνει ότι η χρήση του διαδικτύου δημιουργεί συχνά προβλήματα. Οι παράγοντες *Υπερβολική Χρήση*, *Ανυπομονησία* και *Παραμέληση Κοινωνικής Ζωής* βρίσκονται στο διάστημα 20-39 που δηλώνει ότι χρήση του διαδικτύου βρίσκεται υπό τον πλήρη έλεγχο των χρηστών. Τέλος, ο παράγοντας *Έλλειψη Αυτό-ελέγχου* είναι ο μοναδικός στον οποίο οι χρήστες παρουσιάζουν εμφανώς συχνά προβλήματα (Πίνακας 6).

Πίνακας 6. *Βασικά στατιστικά μέτρα*

Παράγοντες	**Μέση τιμή**	**Τυπική απόκλιση**
Περίοπτη Θέση	39,28	12,85
Υπερβολική Χρήση	37,99	15,06
Παραμέληση Εργασίας	39,96	19,00
Ανυπομονησία	35,48	17,41
Έλλειψη Αυτό-ελέγχου	42,82	16,59
Παραμέληση Κοινωνικής Ζωής	26,29	13,06

Όπως φαίνεται στον πίνακα 7, ο τύπος του Λυκείου και το φύλο του μαθητή είναι αιτίες διαφοροποίησης του βαθμού εθισμού στο διαδίκτυο μόνο για τους παράγοντες *Έλλειψη Αυτό-ελέγχου* και *Παραμέληση Κοινωνικής Ζωής*. Ειδικότερα, οι μαθητές των ΓΕΛ παρουσιάζουν μεγαλύτερη έλλειψη αυτό- ελέγχου αλλά παραμελούν λιγότερο την κοινωνική τους ζωή. Επίσης, τα κορίτσια παρουσιάζουν, σε σχέση με τα αγόρια, μεγαλύτερη έλλειψη αυτοελέγχου αλλά παραμελούν λιγότερο την κοινωνική τους ζωή.

Πίνακας 7. *Παράγοντες και: Τύπος Λυκείου και Φύλο*

	Τύπος Λυκείου		Φύλο	
Παράγοντες	**F**	**Sig.**	**F**	**Sig.**
Περίοπτη Θέση	0,073	0,787	0,517	0,472
Υπερβολική Χρήση	2,191	0,139	0,025	0,873
Παραμέληση Εργασίας	0,073	0,787	0,361	0,548
Ανυπομονησία	0,919	0,328	0,332	0,565
Έλλειψη Αυτό-ελέγχου	6,362	0,012*	7,260	0,007*
Παραμέληση Κοινωνικής Ζωής	4,993	0,026*	2,958	0,086*

Είναι προφανές, από τα αποτελέσματα τα οποία παρουσιάζονται στον επόμενο πίνακα, ότι η σχέση μεταξύ των παραγόντων και του επιπέδου εξοικείωσης είναι θετική αλλά δεν μπορεί να χαρακτηριστεί ως έντονη. Όσο περισσότερο χρόνο αφιερώνουν οι μαθητές στη χρήση του διαδικτύου, τόσο υψηλότερος είναι ο βαθμός εθισμού στο διαδίκτυο. Ωστόσο, στατιστικά σημαντική είναι η σχέση μεταξύ της *εξοικείωσης* και των παραγόντων *Περίοπτη Θέση, Υπερβολική Χρήση, Παραμέληση*

Εργασίας και *Ανυπομονησία*. Διαφαίνεται επίσης μια θετική σχέση μεταξύ της *εξοικείωσης* και των παραγόντων *Έλλειψη Αυτό-ελέγχου και Παραμέληση Κοινωνικής Ζωής* η οποία όμως δεν είναι στατιστικά σημαντική (Πίνακας 8).

Πίνακας 8. *Συσχετίσεις μεταξύ των παραγόντων και της εξοικείωσης*

Παράγοντας	**Εξοικείωση**
Π1. Περίοπτη Θέση	0,227*
Π.2 Υπερβολική Χρήση	0,213*
Π3. Παραμέληση Εργασίας	0,164*
Π4. Ανυπομονησία	0,171*
Π5. Έλλειψη Αυτό-ελέγχου	0,050
Π6. Παραμέληση Κοινωνικής Ζωής	0,053

*Συσχετίσεις σημαντικές σε επίπεδο 0,01.

ΣΥΜΠΕΡΑΣΜΑΤΑ

Από την ανάλυση των αποτελεσμάτων προκύπτει πως οι μαθητές είναι αρκετά εξοικειωμένοι με τη χρήση του διαδικτύου καθώς επτά στους δέκα το χρησιμοποιούν καθημερινά για περισσότερο από τρία χρόνια και από 1-3 ώρες την ημέρα. Πολύ σημαντικό είναι να αναφερθεί ότι ένας στους τρείς μαθητές χρησιμοποιεί το διαδίκτυο σε κάθε σύνδεσή του, περισσότερο από 3 ώρες. Ωστόσο το επίπεδο εξοικείωσης αγοριών και κοριτσιών, μαθητών ΕΠΑΛ και ΓΕΛ δεν παρουσιάζει σημαντική διαφορά. Ο κύριος λόγος που χρησιμοποιούν το διαδίκτυο είναι η επικοινωνία, η αναζήτηση πληροφοριών και λιγότερο η αγορά προϊόντων. Ενθαρρυντικό είναι το γεγονός ότι έξι στους δέκα μαθητές έχουν τον πλήρη έλεγχο της χρήσης του διαδικτύου και μόνο ένας στους εκατό παρουσιάζει σοβαρά προβλήματα εθισμού στο διαδίκτυο. Το σημαντικότερο πρόβλημα το οποίο παρουσιάζουν οι μαθητές αφορά την έλλειψη αυτοελέγχου και την παραμέληση των εργασιών τους (διάβασμα). Φαίνεται ότι οι μαθητές των γενικών Λυκείων παρουσιάζουν μεγαλύτερη έλλειψη αυτοελέγχου σε σχέση με τους μαθητές των ΕΠΑΛ, κάτι που χαρακτηρίζει και τα κορίτσια σε σχέση με τα αγόρια. Ωστόσο, τα κορίτσια παρότι εμφανίζουν έλλειψη αυτοελέγχου, παραμελούν λιγότερο τη κοινωνική τους ζωή σε σχέση με τα αγόρια. Τέλος, γίνεται φανερό από την ανάλυση των δεδομένων ότι όσο περισσότερο χρόνο αφιερώνουν οι μαθητές στη χρήση του διαδικτύου τόσο υψηλότερος είναι ο βαθμός εθισμού τους στο διαδίκτυο που ωστόσο, δεν είναι ανησυχητικός.

Τα αποτελέσματα της συγκεκριμένης έρευνας μπορούν να αξιολογηθούν από ερευνητές και ειδικούς που ασχολούνται με το φαινόμενο του εθισμού στο Διαδίκτυο για περαιτέρω διερεύνηση ή ανάδειξη του φαινομένου. Επιπλέον, στα πλαίσια της ενημέρωσης γονέων και εκπαιδευτικών η γνωστοποίηση των ευρημάτων της έρευνας μπορεί να λειτουργήσει ως κινητήριος μοχλός πρόληψης αυτού του φαινομένου. Τέλος, τα ευρήματα μπορούν να συμβάλουν στην προώθηση προγραμμάτων Αγωγής

Υγείας, υπό την διεύθυνση της Δευτεροβάθμιας Εκπαίδευσης, με σκοπό την ψυχική ενδυνάμωση των εφήβων μαθητών.

ABSTRACT

Internet offers infinite possibilities in everyday human life, in communication, in education and in sciences. At the same time recent surveys point out a lot of negative consequences of its use. The present work has as an object the phenomenon of addiction of adolescents on the Internet, an emerging type of addiction that results of the use of new technologies. The aim of this study was the investigation of the level of addiction of students of high schools. Research was carried out on a sample of 726 students of High schools of regional unit of Kavala, showed that students are quite familiar with the use of internet. The main reason of using the Internet is communication, consultation and less for buying products. It is encouraging that six out of ten students have full control of internet use and only one out of hundred have serious addiction problems. The major problem which students present is the lack of self-control and the neglect of their work (reading).

Keywords: Internet, Addiction, Adolescents, Kavala

ΑΝΑΦΟΡΕΣ

Anderson, K. J. (2001). Internet use among college students: an exploratory study. *Journal of American College Health*, 50 (1), 21-26.

Bai, Y. M.; Lin, C. C. και Chen, J. Y. (2001). Internet addiction disorder among clients of a virtual clinic. *Psychiatric Services*, 52 (10), 1397.

Borzekowski, D. L. (2006). Adolescents' use of the internet: a controversial, coming of age resource. *Adolescence Medical Clinic*, 17 (1), 205-216.

Bradley, K. (2005). Internet lives: social context and moral domain in adolescent development. *New Dir Youth Dev*, 108, 57-76.

Bremer, J. (2005). The internet and children: advantages and disadvantages. *Child Adolescence Psychiatric Clinic*, 14 (3), 405-428.

Cao, F. και Su, L. (2007). Internet addiction among Chinese adolescents: prevalence and psychological feauters. *Child Care Dev.*, 33 (3), 275-281.

Caplan, S. E. (2007). Relations among loneliness, social anxiety and problematic internet use. *Cyberpsychology Behaviour*, 10 (2), 234-242.

Caplan, S.E., Williams, D. και Yee, N. (2009). Problematic internet use and phychosocial wellbeing among MMO players. *Computers in human Behaviour*, 18 (5), 553-575.

Castaneda, J. A., Munoz-Leiva, F. και Luque, T. (2007). Web Acceptance Model (WAM): Moderating effects of user experience., *Information and management*, 44, 384-396.

Chak, K. και Leung, L. (2004). Shyness and locus of control as predictors of internet addiction and internet use. *Cybepsycholy Behaviour*, 7 (5), 559-570.

Dalbudak, E., Evren, C., Aldemir, S., Coskun, K. S., Ugurlu, H., και Yildirim, F. G. (2013). Relationship of Internet Addiction Severity with Depression, Anxiety and

Alexithymia, Temperament and Character in Univercity Studients. *Cyberpsychology, Behaviour, and Social Networking*, 16 (4), 272-278.
Dannon, P.N. και Iancu, I. (2007). Internet addiction. *Harefuah,* 146 (7), 549-553.
Davis, R.A. (2001). A cognitive-behavioural model of pathological Internet use. *Computers in Human Behaviour,* 17, 185-195.
De Berardis, D., D'Albenzio, A., Gambi, F., Sepede, G., Valchera, A., Conti, C. M. και Salerno, R. M. (2009). Alexithymia and its relationships with dissociative experiences and Internet addiction in a nonclinical sample. *CyberPsychology & Behaviour*, 12 (1), 67-69.
Dowell, E.B., Burgess, A.W. και Cavanaugh, D.J. (2009). Clustering of Internet risk behaviors in a middle school student population. *J. Sch Health*, 79 (11), 547-553.
Ha, J. H., Yoo, H. J., Cho, I. H., Chin, B., Shin, D. και Kim, J. H. (2006). Psychiatric comorbidity assessed in Korean children and adolescents who screen positive for Internet addiction. *Journal of Clinical Psychiatry*, 67 (5), 821-826.
Hair, F., Anderson, R., Tatham, R. και Black, W. (1995). *Multivariate Data Analysis with Readings 4th edit*., Prentice-Hall International, London.
Καράπετσας, Α., Φώτης, Α. και Ζυγούρης, Ν. (2012). Νέοι και εθισμός στο διαδίκτυο: Ερευνητική προσέγγιση συχνότητας του φαινομένου. *Εγκέφαλος*, 49, 67-42.
Kelley, K.J. και Gruber, E.M. (2013). Problematic Internet use and physical health. *Journal of Behavioural Addictions*, 2 (2), 108-112.
Kim, E.J., Namkoong, K., Ku, T. και Kim, S.J. (2008). The relationship between on line game addiction and aggression, self-control and narcissistic personality traits. *European Psychiatry*, 23 (3), 212-218.
Lam, L. T., Peng, Z. W., Mai, J. και Jing, J. (2009). Factors associated with Internet addiction among adolescents. *Cyberpsychology Behaviour*, 12 (5), 551-555.
Leung, L. (2007). Stressful life events, motives for Internet use and social suort among digital kids. *Cyberpsychology behavior*, 10 (2), 204-214.
Maenpaa, K., Kale, S.H., Kuusela, H. και Mesiranta, N. (2008). Consumer perceptions of Internet banking in Finland: The moderating role of familiarity. *Journal of retailing and Consumer Services*, 15, 266-276.
Marks, I. (1990). Behavioural (not-chemical) addictions. *British journal of Addictions*, 85(11), 1429-1431.
Moody, E.J. (2001). Internet use and its relationship to loneliness. *CyberPsychology & Behaviour*, 4 (3), 393-401.
Morahan-Martin, J. (2005). Internet abuse: addiction? Disorder? Symptom? Alternative explanations?. *Social Science Computer Review,* 23 (1), 39-48.
Morrison, C.M. και Gore, H. (2010). The relationship between excessive Internet use and depression: a questionnaire-based study of 1,319 young people and adults. *Psychopathology*, 43(2), 121-126.
Ng, B.D. και Wiemer-Hastigs, P. (2005). Addiction to the internet and online gaming. *Cyberpsychology behaviour*, 8 (2), 110-113.
Niemz, K., Griffiths, M.D και Banyard, P. (2005). Prevalence of pathological Internet use among university students and correlations with self-esteem, the General

Health Questionnaire (GHQ), and disinhibition. *CyberPsychology & Behavior*, 8 (6), 562-570.

Nunnally, J.C. (1978). *Psychometric Theory, 2nd edit*. McGraw-Hill, New York.

Pallanti, S., Bernadi, S., Quercioli, L. (2006). The Shorter PROMIS Questionnaire and the Internet Addiction Scale in the assessment of multiple addictions in a high-school population: prevalence and related disability. *CNS Spectrums*, 11 (12), 966-974.

Παναγοπούλου, Α. (2011). *Συμπεριφορές Εξάρτησης στο Διαδίκτυο*. Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών. Ειδικός Λογαριασμός κονδυλίων έρευνας, Μονάδα Εφηβικής Υγείας της Β΄ Παιδιατρικής Κλινικής, Κέντρο Επαγγελματικής Κατάρτισης του Ε.Κ.Π.Α , Αθήνα.

Paul, B. και Bryant, J. A. (2005). Adolescents and the internet. *Adolescence Medical Clinic*, 16 (2), 413-26.

Rachlin, H. (1990). Why do people gamble and keep gambling despite heavy losses? *Psychological Science*, 1, 294-297.

Shapira, N. A., Goldsmith, T. D., Keck, P. E. Jr., Khosla, U. M. and McElroy, S. L. (2000). Psychiatric features of individuals with problematic Internet use. *Journal of Affective Disorders*, 57 (1-3), 267-272.

Suhail, K. και Bargees, Z. (2006). Effects of excessive internet use on undergraduate students in Pakistan. *Cyberpsychology behaviour*, 9 (3), 297-307.

Suss, D. (2007). Impacts of computer- and media usage on the personality development of children and young people. *Ther Umsch*, 64 (2), 103-108.

Tosun, L.P. και Lajunen, T. (2009). Why do young adults develop a passion for Internet activities? The associations among personality, revealing "true self" on the Internet, and passion for the Internet. *Cyberpsychology Behaviour*, 12 (4), 401-406.

Valaitis, RK. (2005). Computers and the internet: tools for youth empowerment. *J Med Internet Res*, 7 (5), 51-59.

Walker, M. B. (1989). Some problems with the concept of gambling addiction: should theories of addiction be generalized to include excessive gambling?. *Journal of Gambling Behavior*, 5, 179-200.

Widyanto, L. και Griffiths, M. (2006). Internet addiction: A critical Review. Int J Ment Health Addict, 4, 31-51.

Yan, Z. (2006). What influences children's and adolescents' understanding of the complexity of the internet? Dev Phychol, 42 (3), 418-428.

Yao, M.Z. και Zhong, Z.-J. (2014). Loneliness, social contacts and Internet addiction: A cross-lagged panel study. *Computers in Human Behaviour*, 30, 164-170.

Yen, J.-Y., Ko, C.-H., Yen, C.-F., Wu, H.-Y. και Yang, M.-J. (2007). The co morbid psychiatric symptoms of Internet addiction: attention deficit and hyperactivity disorder (ADHD) depression, social phobia, and hostility. *Journal of Adolescent Health*, 41 (1), 93-98.

Young, K. S. (1996). Internet Addiction: the emergence of a new clinical disorder. *Cyber Psychology and behaviour*, 1, 237-244.

Young, K. και Abreu, C. (2011*). Internet addiction. A handbook and guide to evaluation and treatment*, John Wiley & Sons, Hoboken, NJ.

Ιστοσελίδες

Ελληνική Στατιστική Αρχή, (2014), *«Στατιστικά στοιχεία».* Ανακτήθηκε στις 28-07-2014, από http://www.report24.gr/statistika-stixia-to-599-twn-ellinwn-kani-xrisi-ke-enimerwnete-apo-to-internet

Encyclopedia of Children's Health, (2014), "*Addiction*". Ανακτήθηκε στις 31-07-2014, από http://www.healthofchildren.com/A/Addiction.

ΕΠΙΨΥ, (2012), *«Χρήση Η/Υ και ίντερνετ από τους εφήβους»*. Ανακτήθηκε στις 24-07-2014, από /http://www.epipsi.gr/pdf/2011/08_HBSC_2010_EPIPSI_2012.pdf

EU NET ADB, (2013), *«Διαδίκτυο και Συμπεριφορές Εξάρτησης: Μελέτη σε Ευρωπαίους Εφήβους».*Ανακτήθηκε στις 08-08-2014, από http://www.eunetadb.eu/files/docs/FinalResearchInternet-GR.pdf

Internet World Stats. Ανακτήθηκε στις 25-07-2014, από http://www.internetlivestats.com/internet-users/

Κοινωνία της πληροφορίας, (2014), *«Οι νέες τεχνολογίες στην καθημερινή ζωή των πολιτών».* Ανακτήθηκε στις 29-07-2014, από http://icteval.ktpae.gr/stats/delivery2/

Σοφός, Α., Αθανασιάδης, Η., Διάκος, Κ. και Δούκα, Α. (2011). *Εθισμός στο Διαδίκτυο-Έρευνα στην Ελλάδα.* 2ο Πανελλήνιο Συνέδριο - Πάτρα 28-30/4/2011. Ανακτήθηκε στις 12-09-2014, από www.cetl.elemedu.upatras.gr/proc2/proceedings.html

Greek Statistical Institute
Proceedings of the 28th Panhellenic Statistics Conference (2015), pp. 108−119

ΑΞΙΟΛΟΓΗΣΗ ΕΚΤΙΜΗΤΩΝ ΠΕΡΙΟΡΙΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΗΣΗΣ ΓΙΑ ΚΑΝΟΝΙΚΟΥΣ ΚΑΙ ΜΗ ΚΑΝΟΝΙΚΟΥΣ ΔΙΑΤΑΡΑΚΤΙΚΟΥΣ ΟΡΟΥΣ ΣΤΗΝ ΠΕΡΙΠΤΩΣΗ ΜΙΚΡΩΝ ΔΕΙΓΜΑΤΩΝ

Γεώργιος Σ. Δονάτος
Πανεπιστήμιο Αθηνών,
Τμήμα Οικονομικών Επιστημών
gdonat@econ.uoa.gr

ΠΕΡΙΛΗΨΗ

Στο άρθρο αυτό θεωρούμε κάποιες εναλλασσόμενες μορφές κανονικών και μη κανονικών διαταρακτικών όρων και παρουσιάζουμε μια μελέτη Monte Carlo της συμπεριφοράς τεσσάρων k τάξης οικονομετρικών εκτιμητών (προβλεπτών). Με βάση 1000 επαναλήψεις δείγματος μεγέθους 21, περιγράφονται τα Monte Carlo πειράματα για μία υπερταυτοποιημένη εξίσωση ενός οικονομικού υποδείγματος, γνωστού ως υπόδειγμα Klein I, και για δεδομένα από την ελληνική οικονομία. Περαιτέρω, βρίσκουμε τις διαβαθμίσεις για μικρά δείγματα των οικονομετρικών εκτιμητών (προβλεπτών), σύμφωνα με τα μέτρα της μεροληψίας και της διασποράς και παρουσιάζουμε τα συμπεράσματα.

Λέξεις Κλειδιά: k τάξης οικονομετρικοί εκτιμητές, κανονικοί και μη κανονικοί διαταρακτικοί όροι, Monte Carlo πειράματα, μικρά δείγματα.

1. ΕΙΣΑΓΩΓΗ

Τα τελευταία χρόνια μελετάται η συμπεριφορά ορισμένων οικονομετρικών εκτιμητών με τη χρήση πειραμάτων Monte Carlo για κανονικούς και μη κανονικούς διαταρακτικούς όρους (Raj 1980, Knight 1985, Prucha and Kielejian 1984, Donatos 1989, Donatos and Michailidis 1996). Η εξειδικευμένη αυτή έρευνα οφείλεται στο ότι η κανονικότητα των διαταρακτικών όρων δεν ισχύει πολλές φορές για μικρά δείγματα (Mandelbrot 1963, Fama 1963, Fieltz and Roselle 1983, Tucker and Pond 1988 κ.ά.). Εξάλλου, οι k τάξης εκτιμητές δεν εξαρτώνται από την κανονικότητα των διαταρακτικών όρων (Raj 1980). Στο πείραμα χρησιμοποιείται το

Υπόδειγμα I του Klein, για γνωστές τιμές των μεταβλητών του. Ακόμα, είναι δυνατό να προσδιοριστούν οι τιμές των διαταρακτικών όρων για εναλλακτικές κατανομές τους. Έτσι, εφαρμόζοντας έναν ικανοποιητικό αριθμό επαναλήψεων στο πλαίσιο του αναπτυσσόμενου πειράματος Monte Carlo, επιτυγχάνεται η εκτίμηση των διαρθρωτικών συντελεστών για τους εκτιμητές τάξης k, που επελέγησαν, για τα διάφορα είδη κατανομής των διαταρακτικών όρων. Στόχο της μελέτης αποτελεί η αξιολόγηση των τεσσάρων οικονομετρικών εκτιμητών k τάξης των διαρθρωτικών συντελεστών και των αντίστοιχων προβλεπτών των μέσων μελλοντικών τιμών της ενδογενούς μεταβλητής, με βάση κριτήρια μεροληψίας και διασποράς.

2. ΥΠΟΔΕΙΓΜΑ ΚΑΙ ΔΕΔΟΜΕΝΑ

Το οικονομικό υπόδειγμα που χρησιμοποιήθηκε στα πειράματα Monte Carlo έχει τρεις εξισώσεις και τρεις ταυτότητες. Έχει κατασκευαστεί από τον Klein (1950) και ονομάζεται Υπόδειγμα I του Klein.

Η πρώτη εξίσωση του υποδείγματος αντιστοιχεί στη συνάρτηση κατανάλωσης:

$$C_\alpha = \beta_0 + \beta_1 P_\alpha + \beta_2 P_{\alpha-1} + \beta_3 (W_\alpha + W'_\alpha) + \varepsilon_\alpha \tag{1}$$

όπου: $\alpha = 1, 2, \ldots, 21$ που αντιστοιχούν στη χρονική περίοδο 1959-1979, C_α είναι η συνολική κατανάλωση για τη χρονική περίοδο, P_α είναι τα συνολικά κέρδη για τη χρονική περίοδο και $P_{\alpha-1}$ είναι τ' αντίστοιχα κέρδη μ' ένα χρόνο υστέρηση, $W_\alpha + W'_\alpha$ είναι το σύνολο των μισθών και ημερομισθίων για τη χρονική περίοδο και ε_α είναι διαταρακτικός όρος. Με W_α και W'_α συμβολίζονται οι μισθοί και τα ημερομίσθια που πληρώνονται από το Δημόσιο (W'_α) και από τον ιδιωτικό τομέα (περιλαμβανομένων και των δημόσιων επιχειρήσεων) (W_α).

Η δεύτερη εξίσωση είναι:

$$I_\alpha = \beta'_0 + \beta'_1 P_\alpha + \beta'_2 P_{\alpha-1} + \beta'_3 K_{\alpha-1} + \varepsilon'_\alpha \tag{2}$$

όπου: I_α είναι οι καθαρές επενδύσεις, K_α είναι το κεφάλαιο της οικονομίας στο τέλος της χρονικής περιόδου, $K_{\alpha-1}$ είναι το κεφάλαιο της οικονομίας στην αρχή της χρονικής περιόδου και ε'_α είναι διαταρακτικός όρος.

Η τρίτη εξίσωση είναι:

$$W_\alpha = \beta''_0 + \beta''_1 X_\alpha + \beta''_2 X_{\alpha-1} + \beta''_3 (\alpha - 1970) + \varepsilon''_\alpha \tag{3}$$

όπου: X_α είναι η συνολική παραγωγή του ιδιωτικού τομέα (περιλαμβανομένων και των δημόσιων επιχειρήσεων) για τη χρονική περίοδο και $X_{\alpha-1}$ μ' ένα χρόνο υστέρηση, και ε''_α είναι διαταρακτικός όρος.

Ακολουθούν οι παρακάτω τρεις ταυτότητες:

$$X_\alpha = C_\alpha + I_\alpha + G_\alpha + R_\alpha \tag{4}$$

όπου: Το G_α αντιπροσωπεύει κυβερνητικές δαπάνες που δεν αφορούν μισθούς, ενώ το R_α αναφέρεται σε στατιστικές διαφορές και στη μεταβολή αποθεμάτων. Η αντίστοιχη ταυτότητα του Υποδείγματος I του Klein τροποποιήθηκε με πρόσθεση της μεταβλητής R, εξαιτίας του τρόπου που δίνονται τα δεδομένα, για την ελληνική οικονομία, τα οποία αφορούν την ακαθάριστη δαπάνη της οικονομίας και το ακαθάριστο εγχώριο προϊόν.

$$P_\alpha = X_\alpha - W_\alpha - T_\alpha \tag{5}$$

όπου: T_α είναι φόροι επιχειρήσεων.

$$K_\alpha = K_{\alpha-1} + I_\alpha. \tag{6}$$

Συνολικά, στο παραπάνω σύστημα των έξι αλληλεξαρτημένων εξισώσεων περιέχονται έξι ενδογενείς μεταβλητές οι: C_α, P_α, W_α, I_α, K_α, X_α και οκτώ προκαθορισμένες μεταβλητές (με χρονική υστέρηση: $X_{\alpha-1}$, $P_{\alpha-1}$, $K_{\alpha-1}$ και εξωγενείς: α, W'_α, C_α, T_α, R_α).

Οι εξωγενείς μεταβλητές και οι ενδογενείς μεταβλητές με χρονική υστέρηση θεωρούνται ανεξάρτητες, σχετικά με τη λειτουργία του συστήματος, αν οι διαταρακτικοί όροι e_α, e'_α, e''_α είναι στοχαστικά ανεξάρτητοι.

Σκοπεύοντας να εξετάσουμε τις ιδιότητες των εκτιμητών για μικρά δείγματα, θα χρησιμοποιήσουμε την υπερταυτοποιημένη εξίσωση κατανάλωσης του Υποδείγματος I του Klein και δεδομένα από την ελληνική οικονομία. Όλες οι μεταβλητές του υποδείγματος μετριούνται σε εκατομμύρια δραχμές και σταθερές τιμές για το 1970.

Η συνάρτηση κατανάλωσης (1) μπορεί να γραφεί:

$$C_1 = Z_1\delta_1 + E_1 \tag{7}$$

όπου:

$$C_1 = [c_1 c_2 \dots c_{21}]', \quad Z_1 = [P_\alpha \;\; W_\alpha + W'_\alpha \;\; I \;\; P_{\alpha-1}]$$
$$\delta_1 = [\beta_1 \beta_3 \mid \beta_0 \beta_2]', \quad E_1 = [\varepsilon_1 \varepsilon_2 \dots \varepsilon_{21}]'.$$

Οι διαταρακτικοί όροι ακολουθούν την κανονική, ομοιόμορφη και λογαριθμοκανονική κατανομή, με τυπική απόκλιση $\sigma = 1200$, αντί $\sigma = 2520,35$ (τυπικό σφάλμα παλινδρόμησης). Αυτή η μικρή τυπική απόκλιση επελέγη προκειμένου οι εκτιμήσεις των διαρθρωτικών συντελεστών της εξίσωσης κατανάλωσης του υποδείγματος I του Klein (1950), που εκτιμήθηκε για δεδομένα από την ελληνική οικονομία, να είναι θετικές σύμφωνα με την οικονομική θεωρία.

3. ΕΚΤΕΛΕΣΗ ΠΕΙΡΑΜΑΤΟΣ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ

Η δημιουργία των διαταρακτικών όρων επιτυγχάνεται ως εξής:

Αρχικά κατασκευάζονται χίλια δείγματα μεγέθους εικοσιένα, τα οποία αποτελούνται από ομοιόμορφα κατανεμημένους ανεξάρτητους τυχαίους αριθμούς στο διάστημα $(0,1)$. Στη συνέχεια, τα σύνολα αυτά των τυχαίων αριθμών μετασχηματίζο-

νται σε σύνολα ανεξάρτητα κατανεμημένων κανονικών και μη κανονικών αποκλίσεων. Ειδικότερα, κάθε σύνολο των ομοιόμορφα κατανεμημένων τυχαίων αριθμών στο διάστημα $(0,1)$ μετασχηματίζεται:

α) σ' ένα σύνολο αποκλίσεων, που ακολουθεί την τυποποιημένη κανονική κατανομή ή

β) σ' ένα σύνολο αποκλίσεων, που ακολουθεί την ομοιόμορφη κατανομή $U(0,1)$ ή

γ) σ' ένα σύνολο αποκλίσεων, που ακολουθεί την λογαριθμοκανονική κατανομή Log(- .3476, .693).

Ακολούθως, η ανάπτυξη του πειράματος μπορεί να περιγραφεί, ως εξής:

Από τον τύπο (7) είναι φανερό ότι ο Πίνακας C_1, που αντιστοιχεί στην κατανάλωση, εκφράζεται με μορφή πινάκων ως συνάρτηση των πινάκων Z_1, δ_1 και E_1. Οι διαταρακτικοί όροι, όπως αναφέρθηκε, ακολουθούν τρεις διαφορετικούς τύπους κατανομής. Το πρόγραμμα επιστημονικών εφαρμογών για τις κατανομές διαταρακτικών όρων εφαρμόστηκε για εικοσιμία παρατηρήσεις, χίλιες επαναλήψεις και για κάθε τύπο κατανομής. Εξάλλου, για τ' αρχικά δεδομένα και δια μέσου της μεθόδου OLS προσδιορίσθηκαν οι τιμές των στοιχείων του σταθερού διανύσματρος δ_1^0:

$$\delta_1^0 = [.226 \quad .962 \quad | \quad 20120.226 \quad .327]' \tag{8}$$

που θεωρείται ότι παρέχει την "αληθινή" τιμή των συντελεστών, με τυπικό σφάλμα παλινδρόμησης $\sigma = 2520$.

Στη συνέχεια, κατασκευάστηκαν για κάθε είδος κατανομής των διαταρακτικών όρων, χίλια διανύσματα C_1, που καθένα περιέχει εικοσιμία τιμές. Η δημιουργία αυτών των διανυσμάτων πραγματοποιήθηκε δια της προσθέσεως του αντίστοιχου κάθε φορά διαταρακτικού όρου στο σταθερό διάνυσμα C, που δίνεται από την έκφραση:

$$C = Z_1\, \delta_1^0. \tag{9}$$

Στη συνέχεια, εφαρμόστηκαν ορισμένοι εκτιμητές τάξης k, δηλαδή ο OLS ($k = 0$), ο 2SLS ($k = 1$), ο αμερόληπτος τάξης k εκτιμητής του Nagar κατά προσέγγιση τάξης T^{-1} (Unbiased k-class (UBK)) και ο εκτιμητής που αντιστοιχεί στην τιμή $k = 1.4$. Η τιμή $k = 1.4$ είναι ένα άνω όριο, επειδή για εκτιμητές τάξης k που η τιμή του k είναι μεγαλύτερη από 1.4 οι αντίστοιχες εκτιμήσεις των διαρθρωτικών συντελεστών είναι αρνητικές και δεν έχουν οικονομική έννοια. Έτσι, με τη βοήθεια των "τιμών" των προκαθοριζόμενων μεταβλητών και των παραγόμενων διανυσμάτων C_1 εκτιμήθηκαν, για κάθε μέθοδο και για κάθε είδος κατανομής των διαταρακτικών όρων, τα διανύσματα δ_1 που αναφέρονται στους διαρθρωτικούς συντελεστές. Μετά τις εκτιμήσεις των διανυσμάτων δ_1, υπολογίζονται οι αντίστοιχες προβλέψεις των διανυσμάτων C_1 από τον τύπο:

$$\widehat{C}_1 = Z_1\, \widehat{\delta}_1 \tag{10}$$

όπου: $\widehat{\delta}_1$ είναι η εκτίμηση του διανύσματος δ_1 και $\widehat{C}_1$ η εκτίμηση του αντίστοιχου διανύσματος C_1.

Η επιλογή των κριτηρίων καταλληλότητας ενός εκτιμητή εξαρτάται κυρίως από τον κύριο στόχο της εκτίμησης. Εξάλλου, συνήθως δεν υπάρχει ιδιαίτερο ενδιαφέρον για την εύρεση της κατανομής που ακολουθεί κάποιος εκτιμητής, αλλά για τη γνώση των αντίστοιχων ροπών και ειδικότερα της μέσης τιμής και της διακύμανσης. Με βάση αυτή την αρχή έχουν επιλεγεί ως κριτήρια καταλληλότητας των εκτιμητών που χρησιμοποιήθηκαν το σφάλμα μεροληψίας, η τυπική απόκλιση και η ρίζα του μέσου τετραγωνικού σφάλματος.

Το σφάλμα μεροληψίας που αντιστοιχεί σε κάθε εκτιμητή, για τους διαφόρους τύπους κατανομής των διαταρακτικών όρων, δίνεται από τον τύπο:

$$b(\widehat{\delta}_{21}) = \overline{\widehat{\delta}}_{21} - \delta_1^0 \tag{11}$$

όπου: $\overline{\widehat{\delta}}_{21}$ εκφράζει τη μέση τιμή των εκτιμήσεων $\widehat{\delta}_j$ του παραμετρικού διανύσματος που αντιστοιχούν στο δείγμα j.

Ο πίνακας διακύμανσης-συνδιακύμανσης γύρω από την "αληθινή" τιμή του παραμετρικού διανύσματος είναι:

$$\overline{\mathrm{Cov}}(\widehat{\delta}_{21}) = \frac{1}{1000}\sum_{j=1}^{1000}(\widehat{\delta}_j - \delta_1^0)(\widehat{\delta}_j - \delta_1^0)'. \tag{12}$$

Ο πίνακας διακύμανσης-συνδιακύμανσης για την κατανομή του μικρού δείγματος είναι:

$$\mathrm{Cov}(\widehat{\delta}_{21}) = \frac{1}{1000}\sum_{j=1}^{1000}(\widehat{\delta}_j - \overline{\widehat{\delta}}_{21})(\widehat{\delta}_j - \overline{\widehat{\delta}}_{21})'. \tag{13}$$

Τα διαγώνια στοιχεία των πινάκων $\overline{\mathrm{Cov}}(\widehat{\delta}_{21})$ και $\mathrm{Cov}(\widehat{\delta}_{21})$ δίνουν, αντίστοιχα, το μέσο τετραγωνικό σφάλμα (MSE) και τη διακύμανση $(Var(\widehat{\delta}_j))$. Απ' όπου λαμβάνεται η ρίζα του μέσου τετραγωνικού σφάλματος $(RMSE)$ και η τυπική απόκλιση $SD(\widehat{\delta}_j)$:

$$RMSE = (MSE))^{1/2} \tag{14}$$

$$SD(\widehat{\delta}_j) = (Var(\widehat{\delta}_j))^{1/2}. \tag{15}$$

Ένα άλλο ενδιαφέρον μέτρο που εφαρμόστηκε, για την εύρεση του αποτελεσματικού εκτιμητή δίνεται από τον τύπο:

$$e_1 = \frac{\left|\overline{\mathrm{Cov}}(\widehat{\delta}_{21})_{k_a}\right|}{\left|\overline{\mathrm{Cov}}(\widehat{\delta}_{21})_{k_{a'}}\right|} \tag{16}$$

όπου: $\overline{\mathrm{Cov}}(\widehat{\delta}_{21})_{k_a}$ και $\overline{\mathrm{Cov}}(\widehat{\delta}_{21})_{k_{a'}}$, είναι οι πίνακες διακύμανσης-συνδιακύμανσης γύρω από την "αληθινή" τιμή του παραμετρικού διανύσματος για τους εκτιμητές

τάξης k που χαρακτηρίζονται από τους δείκτες k_a και $k_{a'}$ αλλά για ίδια κατανομή διαταρακτικών όρων. Μάλιστα, το μέτρο (16) είναι ο λόγος των "γενικευμένων διακυμάνσεων" γύρω από την "αληθινή" τιμή του παραμετρικού διανύσματος, που ταυτόχρονα αναφέρεται στα χαρακτηριστικά μεροληψίας. Με $\|$ συμβολίζεται η ορίζουσα.

Επίσης, χρησιμοποιήθηκε το μέτρο (17) τροποποιημένο για την περίπτωση διακύμανσης-συνδιακύμανσης για την κατανομή του μικρού δείγματος:

$$e_2 = \frac{\left|\mathrm{Cov}(\widehat{\delta}_{21})_{k_a}\right|}{\left|\mathrm{Cov}(\widehat{\delta}_{21})_{k_{a'}}\right|} \tag{17}$$

Για την ενδογενή μεταβλητή υπολογίστηκαν τα μέτρα που αντιστοιχούν στα μέτρα (11) μέχρι (17) και αφορούν κάθε προβλεπτή του μέσου της μεταβλητής κατανάλωσης και όλα τα είδη κατανομής των διαταρακτικών όρων.

Από την εφαρμογή των προαναφερομένων κριτηρίων προέκυψαν τα σχετικά αποτελέσματα. Λόγω περιορισμένου χώρου, παραθέτουμε ενδεικτικά μόνο τα αποτελέσματα που περιέχονται στους Πίνακες 1, 2, 3, 4, 5, 6 και 7.

Πίνακας 1
Κανονική κατανομή

		$\bar{\widehat{\delta}}_{21}$		$b(\widehat{\delta}_{21})$		$SD(\widehat{\delta}_j)$		$RMSE(\widehat{\delta}_j)$	
OLS	β_0	.2011980	E 05	- .4257813	E 00	.8996610	D 00	.8767030	D 00
	β_1	.2269810	E 00	- .1895428	E-04	.1064397	D-03	.1075830	D-03
	β_2	.3269813	E 00	- .1865625	E-04	.1062432	D-03	.1070243	D-03
	β_3	.9629582	E 00	- .4178286	E-04	.2987638	D-03	.3008032	D-03
$2SLS$	β_0	.2012026	E 05	.3515625	E-01	.9271552	D 00	.9363463	D 00
	β_1	.2269871	E 00	- .1287460	E-04	.3387708	D-04	.3536071	D-04
	β_2	.3269810	E 00	- .1895428	E-04	.1063411	D-03	.1071448	D-03
	β_3	.9629896	E 00	- .1037121	E-04	.2234267	D-04	.2184463	D-04
UBK	β_0	.2086771	E 05	.7474844	E 03	.7913131	D 01	.7477558	D 03
	β_1	.1653141	E 00	- .6168586	E-01	.2331592	D-03	.6168557	D-01
	β_2	.3779719	E 00	.5097198	E-01	.1700670	D-03	.5097739	D-01
	β_3	.9739250	E 00	.1092499	E-01	.8281846	D-04	.1093189	D-01
$k = 1.4$	β_0	.2278766	E 05	.2667434	E 04	.1241224	D 02	.2067714	D 04
	β_1	.8715268	E-02	- .2182847	E 00	.1231925	D-02	.2182881	D 00
	β_2	.5046096	E 00	.1776097	E 00	.4560077	D-01	.1833758	D 00
	β_3	.1004164	E 01	.4116374	E-01	.1952988	D-03	.4116997	D-01

Πίνακας 2
Ομοιόμορφη κατανομή

		$\bar{\hat{\delta}}_{21}$		$b(\hat{\delta}_{21})$		$SD(\hat{\delta}_j)$		$RMSE(\hat{\delta}_j)$	
OLS	β_0	.2012034	E 05	.1171875	E 00	.3954531	D 00	.4886613	D 00
	β_1	.2269638	E 00	- .3623962	E-04	.1723629	D-03	.1757765	D-03
	β_2	.3269905	E 00	- .9477139	E-05	.9886804	D-05	.9850744	D-05
	β_3	.9629870	E 00	- .1299381	E-04	.8225151	D-05	.9284931	D-05
$2SLS$	β_0	.2012083	E 05	.6054688	E-00	.3432931	D 00	.8625486	D 00
	β_1	.2269903	E 00	- .9655952	E-05	.1036802	D-04	.1277890	D-04
	β_2	.3269902	E 00	- .9775162	E-05	.1178785	D-04	.1212362	D-04
	β_3	.9629880	E 00	- .1204014	E-04	.7961275	D-05	.8670642	D-05
UBK	β_0	.2087070	E 05	.7504766	E 03	.6680934	D 01	.7507415	D 03
	β_1	.1652088	E 00	- .6179172	E-01	.1942937	D-03	.6179143	D-01
	β_2	.3780382	E 00	.5103821	E-01	.1669765	D-03	.5104318	D-01
	β_3	.9642138	E 00	.1213849	E-02	.9690761	D-01	.9691545	D-01
$k = 1.4$	β_0	.2278460	E 05	.2664375	E 04	.1561233	D 02	.2664687	D 04
	β_1	.9192370	E-02	- .2178076	E 00	.1316465	D-02	.2178115	D 00
	β_2	.5088153	E 00	.1818153	E 00	.1290998	D-02	.1818255	D 00
	β_3	.1004073	E 01	.4107314	E-01	.1882646	D-03	.4108004	D-01

Πίνακας 3
Λογαριθμοκανονική κατανομή

		$\bar{\hat{\delta}}_{21}$		$b(\hat{\delta}_{21})$		$SD(\hat{\delta}_j)$		$RMSE(\hat{\delta}_j)$	
OLS	β_0	.2012081	E 05	.5859375	E 00	.6627622	D 00	.1051528	D 01
	β_1	.2269946	E 00	- .5424023	E-05	.2202416	D-04	.2202666	D-04
	β_2	.3269854	E 00	- .1454353	E-04	.2194192	D-04	.2286325	D-04
	β_3	.9629881	E 00	- .1192093	E-04	.1362605	D-04	.1349980	D-04
$2SLS$	β_0	.2012128	E 05	.1054688	E-01	.6594976	D 00	.1429972	D 01
	β_1	.2269915	E 00	- .8363860	E-05	.2370983	D-04	.2434206	D-04
	β_2	.3269902	E 00	- .9775162	E-05	.3976883	D-04	.3964653	D-04
	β_3	.9629886	E 00	- .1144409	E-04	.1380199	D-04	.1346172	D-04
UBK	β_0	.2087237	E 05	.7521445	E 03	.2669932	D 02	.7528913	D 03
	β_1	.1652800	E 00	- .6172001	E-01	.1931482	D-03	.6171963	D-01
	β_2	.3780088	E 00	.5100888	E-01	.1496099	D-03	.5101393	D-01
	β_3	.9739276	E 00	.1092756	E-01	.8484216	D-04	.1093450	D-01
$k = 1.4$	β_0	.2279330	E 05	.2673074	E 04	.3428532	D 02	.2673561	D 04
	β_1	.8658968	E-02	- .2183410	E 00	.1163065	D-02	.2183441	D 00
	β_2	.5091992	E 00	.1821992	E 00	.1259958	D-02	.1822085	D 00
	β_3	.1004180	E 01	.4117996	E-01	.1897294	D-03	.4118708	D-01

Πίνακας 4
Γενικευμένες διακυμάνσεις

(α) $\lvert\overline{\mathrm{Cov}}(\widehat{\delta}_{21})\rvert$	**Εκτιμητές**			
(β) $\lvert\mathrm{Cov}(\widehat{\delta}_{21})\rvert$	OLS	$2SLS$	UBK	$k = 1.4$
Κανονική Κατανομή (α)	-.55773 D-18	.18465 D-16	.37464 D+11	.92990 D+15
Κανονική Κατανομή (β)	.32362 D-16	.35898 D-15	.58135 D-07	.12198 D-03
Ομοιόμορφη Κατανομή (α)	.16641 D-16	.75536 D-16	.38384 D+11	.96722 D+15
Ομοιόμορφη Κατανομή (β)	.25764 D-17	-.14073 D-18	-.18223 D-07	.19694 D-03
Λογαριθμο-κανονική Κατανομή (α)	.30979 D-15	.39244 D-15	.38778 D+11	.98981 D+15
Λογαριθμο-κανονική Κατανομή (β)	.79406 D-17	.32532 D-16	.18746 D-06	.24949 D-03

Πίνακας 5
Κανονική κατανομή

e_1	**Εκτιμητές**			
Αριθμητής / Παρανομαστής	OLS	$2SLS$	UBK	$k = 1.4$
OLS	1.0	- .33107 D+02	- .67172 D+29	-1.66729 D+33
$2SLS$	-3.02047 D-02	1.0	2.02891 D+27	5.03601 D+31
UBK	-1.48870 D-29	.49287 D-27	1.0	2.48211 D+04
$k = 1.4$	- .59977 D-33	.19990 D-31	.40288 D-04	1.0
e_2	**Εκτιμητές**			
Αριθμητής / Παρανομαστής	OLS	$2SLS$	UBK	$k = 1.4$
OLS	1.0	1.10926 D+01	1.79639 D+09	.37692 D+13
$2SLS$	.90149 D-01	1.0	1.61944 D+08	.33979 D+12
UBK	.55666 D-09	.61749 D-08	1.0	.20982 D+04
$k = 1.4$	2.65305 D-13	2.94294 D-12	4.76594 D-04	1.0

Πίνακας 6

Ομοιόμορφη κατανομή

e_1	**Εκτιμητές**			
Αριθμητής / Παρανομαστής	OLS	$2SLS$	UBK	$k = 1.4$
OLS	1.0	4.53915	2.30659 D+27	5.81227 D+31
$2SLS$	.22030	1.0	.50815 D+27	1.28047 D+31
UBK	.43354 D-27	1.96790 D-27	1.0	2.51985 D+04
$k = 1.4$	.17204 D-31	.78095 D-31	.39684 D-04	1.0
e_2	**Εκτιμητές**			
Αριθμητής / Παρανομαστής	OLS	$2SLS$	UBK	$k = 1.4$
OLS	1.0	- .54622 D-01	- .70730 D+10	.76439 D+14
$2SLS$	-1.83073 D+01	1.0	1.29489 D+11	-1.39941 D+15
UBK	-1.41381 D-10	.77226 D-11	1.0	-1.08072 D+04
$k = 1.4$	1.30821 D-14	-.71458 D-15	.92530 D-04	1.0

Πίνακας 7

Λογαριθμοκανονική κατανομή

e_1	**Εκτιμητές**			
Αριθμητής / Παρανομαστής	OLS	$2SLS$	UBK	$k = 1.4$
OLS	1.0	1.26679	1.25175 D+26	3.19509 D+30
$2SLS$	.78939	1.0	.98812 D+26	2.52219 D+30
UBK	.79888 D-26	1.01201 D-26	1.0	2.55250 D+04
$k = 1.4$	.31297 D-30	.39648 D-30	.39177 D-04	1.0
e_2	**Εκτιμητές**			
Αριθμητής / Παρανομαστής	OLS	$2SLS$	UBK	$k = 1.4$
OLS	1.0	.40969 D+01	.23607 D+11	.31419 D+14
$2SLS$	2.44085 D-01	1.0	.57623 D+10	.76690 D+13
UBK	4.23589 D-11	1.73541 D-10	1.0	1.33089 D+03
$k = 1.4$	3.18273 D-14	1.30394 D-13	.75137 D-03	1.0

4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Γενικά, με βάση τις εκτιμήσεις που προέκυψαν, μπορεί να διατυπωθούν τα ακόλουθα συμπεράσματα:

1) Από τους Πίνακες 1, 2 και 3 εξάγονται, για τους εκτιμητές των διαρθρωτικών συντελεστών, τα εξής:

   α) Τα σφάλματα μεροληψίας των εκτιμήσεων των διαρθρωτικών συντελεστών για τους εκτιμητές OLS και $2SLS$ είναι ιδιαίτερα μικρά και οι αντίστοιχες διαφορές τους κυμαίνονται σε πολύ χαμηλά επίπεδα.

Αναλυτικότερα: Ο εκτιμητής $2SLS$ είναι λιγότερο μεροληπτικός από τους άλλους εκτιμητές, όταν οι διαταρακτικοί όροι κατανέμονται κανονικά, ενώ στην περίπτωση που χρησιμοποιείται ομοιόμορφη ή λογαριθμοκανονική κατανομή οι εκτιμητές OLS και $2SLS$ παρουσιάζουν το μικρότερο σφάλμα μεροληψίας. Εξάλλου, ο εκτιμητής $k = 1.4$ είναι ο πιο μεροληπτικός και από τους τέσσερις εκτιμητές, για όλες τις κατανομές των διαταρακτικών όρων που χρησιμοποιήθηκαν.

β) Οι διαφορές των $RMSE$ ή των SD μεταξύ των διαφόρων μεθόδων δεν μπορεί να θεωρηθούν σημαντικές. Ιδιαίτερα οι διαφορές αυτές για τους εκτιμητές OLS και $2SLS$ είναι οπωσδήποτε πολύ μικρές. Ακόμα, οι διασπορές των εκτιμητών OLS και $2SLS$ γύρω από τους μέσους είναι παρόμοιες με τις διασπορές τους γύρω από την "αληθινή" τιμή, όπως βέβαια αναμενόταν εξαιτίας των μικρών σφαλμάτων μεροληψίας.
Ειδικότερα παρατηρείται ότι:

i) Οι εκτιμητές OLS και $2SLS$ έχουν το μικρότερο $RMSE$, όταν οι διαταρακτικοί όροι κατανέμονται σύμφωνα με την κανονική ή την ομοιόμορφη κατανομή. Ο εκτιμητής OLS παρουσιάζει το μικρότερο $RMSE$ στην περίπτωση που οι διαταρακτικοί όροι ακολουθούν τη λογαριθμοκανονική κατανομή. Ο εκτιμητής $k = 1.4$ έχει το μεγαλύτερο $RMSE$ για κάθε χρησιμοποιούμενη στο πείραμα κατανομή διαταρακτικών όρων.

ii) Οι εκτιμητές OLS και $2SLS$ εμφανίζουν το μικρότερο SD, εφόσον οι διαταρακτικοί όροι ακολουθούν την κανονική κατανομή. Οι εκτιμητές OLS και $2SLS$ παρουσιάζουν το μικρότερο SD όταν κατανομή των διαταρακτικών όρων είναι, αντίστοιχα, η λογαριθμοκανονική και η ομοιόμορφη. Ο εκτιμητής $k = 1.4$ έχει το μεγαλύτερο SD και για τις τρεις κατανομές των διαταρακτικών όρων.

2) Με τη βοήθεια των Πινάκων 4, 5, 6 και 7, εξάγονται τα εξής:

α) Όταν λαμβάνεται ως μέτρο ο λόγος των "γενικευμένων διακυμάνσεων" γύρω από το μέσο παραμετρικό διάνυσμα, προκύπτει ότι:

i) Για κανονική ή λογαριθμοκανονική κατανομή διαταρακτικών όρων ο εκτιμητής OLS είναι αποτελεσματικότερος από τους εκτιμητές $2SLS$, UBK και $k = 1.4$, ο εκτιμητής $2SLS$ από τους εκτιμητές UBK και $k = 1.4$ και ο εκτιμητής UBK από τον εκτιμητή $k = 1.4$.

ii) Για ομοιόμορφη κατανομή διαταρακτικών όρων ο εκτιμητής OLS είναι αποτελεσματικότερος από τον εκτιμητή $k = 1.4$ και ο εκτιμητής $2SLS$ από τον εκτιμητή UBK.

β) Όταν λαμβάνεται ως μέτρο ο λόγος των "γενικευμένων διακυμάνσεων" γύρω από την "αληθινή" τιμή, προκύπτει ότι:

i) Για κανονική κατανομή διαταρακτικών όρων ο εκτιμητής $2SLS$ είναι αποτελεσματικότερος από τους εκτιμητές UBK και $k = 1.4$ και ο εκτιμητής UBK από τον εκτιμητή $k = 1.4$.

ii) Για ομοιόμορφη ή λογαριθμοκανονική κατανομή διαταρακτικών όρων ο εκτιμητής OLS είναι αποτελεσματικότερος από τους εκτιμητές $2SLS$, UBK και $k = 1.4$, ο εκτιμητής $2SLS$ από τους εκτιμητές UBK και $k = 1.4$ και ο εκτιμητής UBK από τον εκτιμητή $k = 1.4$.

3) Από τα αποτελέσματα που απορρέουν για την περίπτωση της ενδογενούς μεταβλητής, τα οποία λόγω περιορισμένου χώρου δεν περιέχονται στο κείμενο, συνάγονται τα εξής:

α) Τα σφάλματα μεροληψίας για τον προβλεπτή του μέσου της ενδογενούς μεταβλητής είναι μεγάλα. Ο προβλεπτής OLS έχει το μικρότερο σφάλμα μεροληψίας και ο προβλεπτής $k = 1.4$ εμφανίζεται περισσότερο μεροληπτικός από τους άλλους προβλεπτές και για τις τρεις κατανομές των διαταρακτικών όρων.

β) Οι διασπορές για τις προβλεπόμενες τιμές γύρω από την "αληθινή" τιμή είναι πολύ μεγάλες. Όμως, οι διαφορές μεταξύ των χρησιμοποιούμενων προβλεπτών είναι τόσο μικρές, ώστε κάθε διαβάθμισή τους πρέπει να θεωρηθεί παρακινδυνευμένη. Οπωσδήποτε καμιά μέθοδος δε φαίνεται, με βεβαιότητα, να είναι καλύτερη ή χειρότερη από τις άλλες, αλλά θα μπορούσε να λεχθεί ότι οι προβλεπτές OLS και $k = 1.4$ δείχνουν ότι έχουν, αντίστοιχα, το μικρότερο και μεγαλύτερο $RMSE$.

γ) Οι διασπορές για τις προβλεπόμενες τιμές γύρω από τους μέσους δεν είναι μεγάλες. Επίσης, αρκετές τιμές, για τις διάφορες εκτιμητικές μεθόδους είναι σχεδόν ίσες. Ιδιαίτερα, οι προβλεπτές OLS και $2SLS$ παρουσιάζουν πολύ μικρές διασπορές γύρω από τους μέσους. Ακόμα, ο προβλεπτής $2SLS$ εμφανίζει το μικρότερο SD από τους άλλους προβλεπτές για την κανονική και ομοιόμορφη κατανομή των διαταρακτικών όρων, ενώ ο προβλεπτής OLS για τη λογαριθμοκανονική κατανομή των διαταρακτικών όρων. Τέλος, ο προβλεπτής $k = 1.4$ έχει το μεγαλύτερο SD από τους άλλους προβλεπτές και για τις τρεις κατανομές των διαταρακτικών όρων.

ABSTRACT

In this paper we consider some alternative forms of normal and non normal disturbances and we report a Monte Carlo study of the behaviour of four k class estimators (predictors). On the basis of 1000 replications of sample size 21, we describe the Monte Carlo experiments on an overidentified equation of an economic moded, known as the Klein I model, and for data from the Greek economy. Further, we found the small sample rankings of econometric, estimators (predictors), according the measures of bias and dispersion and we present the conclusions.

ΑΝΑΦΟΡΕΣ

Donatos, G. S. (1989). A Monte Carlo study of k class estimators for small samples with normal and non normal disturbances, *The Statistician*, **38**, 11-20.

Donatos, G. S. and Michailidis, G. S. (1996). Small sample properties of some limited and full information simultaneous equation estimator for normal and non normal autocorrelated disturbances, special issue on econometric methology, *Journal of Statistical Planning and Inference*, **50**, 732-282.

Famma, E. F. (1963). Mandelbrot and the stable paretian hypothesis, *Journal of Business*, **36**, 420-429.

Fieltz, B. and Roselle, J. (1983). Stable distributions and the mixture of distributions hypotheses for common stock returns, *Journal of the American Statistical Association*, **78**, 28-36.

Klein, L. R. (1950). *Economic fluctuations in the United States, 1921-1941*, New York, John Wiley.

Knight, J. L. (1985). The moments of OLS and 2SLS when the disturbances are non normal, *Journal of Econometrics*, **27**, 39-60.

Mandelbrot, B. (1963). The variation of certain speculative prices, *Journal of Business*, **36**, 394-419.

Prucha, I. R. and Kelejian, H. H. (1984). The structure of simultaneous equation estimators: a generalisation towards non normal disturbances, *Econometrica*, **52**, 721-736.

Raj, B. (1980). A Monte Carlo study of small sample properties of simultaneous equation estimators with normal and nonnormal disturbances, *Journal of the American Statistical Association*, **75**, 221-229.

Tucker, A. and Pond, L. (1988). The probability distribution of foreign exchange price changes, *Review of Economics and Statistics*, **70**, 638-647.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής (2015), σελ.120-132

ΟΙ ΕΠΙΠΤΩΣΕΙΣ ΤΩΝ ΑΚΡΑΙΩΝ ΣΥΝΘΗΚΩΝ ΤΗΣ ΑΓΟΡΑΣ ΣΤΟ ΔΕΙΚΤΗ ΚΕΦΑΛΑΙΑΚΗΣ ΕΠΑΡΚΕΙΑΣ ΤΩΝ ΕΛΛΗΝΙΚΩΝ ΤΡΑΠΕΖΩΝ: ΜΙΑ ΕΦΑΡΜΟΓΗ ΤΗΣ ΜΕΘΟΔΟΥ ΕΣΩΤΕΡΙΚΩΝ ΔΙΑΒΑΘΜΙΣΕΩΝ

Α. Δονάτου[1], Ι. Λεβεντίδης[2]
[1] Τμήμα Οικονομικών Επιστημών, Πανεπιστήμιο Αθηνών
adonatou@yahoo.gr, ylevent@econ.uoa.gr

ΠΕΡΙΛΗΨΗ

Στην εργασία εξετάζεται η επίδραση ακραίων καταστάσεων στα δανειακά χαρτοφυλάκια του ελληνικού τραπεζικού συστήματος. Τα χαρτοφυλάκια αυτά έχουν ομαδοποιηθεί σε τρεις ομάδες, ανάλογα με το μέγεθος της Τράπεζας στην οποία ανήκουν, δηλαδή σε μεγάλου, μεσαίου και μικρού μεγέθους. Εφαρμόζεται μια σειρά σεναρίων ακραίων καταστάσεων στα τρία χαρτοφυλάκια και εκτιμάται με την προσέγγιση εσωτερικών διαβαθμίσεων του συμφώνου της Βασιλείας ΙΙ η αύξηση των κεφαλαιακών απαιτήσεων για κάθε σενάριο. Τα αποτελέσματα, που προέκυψαν, καταδεικνύουν την αύξηση του πιστωτικού κινδύνου στις περιόδους κρίσης, τη διαφοροποίηση του κινδύνου, ανάλογα με το μέγεθος του τραπεζικού οργανισμού, καθώς και τα επιπρόσθετα κεφάλαια που θα χρειασθούν για την αντιστάθμιση του κινδύνου.

Λέξεις κλειδιά:Διαχείριση κινδύνων, δείκτης κεφαλαιακής επάρκειας, Σύμφωνα Βασιλείας Ι και ΙΙ, Πιστωτικός κίνδυνος, μέθοδος εσωτερικών διαβαθμίσεων.

1. ΕΙΣΑΓΩΓΗ

Οι Τράπεζες οφείλουν να διακρατούν κατάλληλο ύψος ιδίων κεφαλαίων, προκειμένου να είναι σε θέση να αντιμετωπίσουν επαρκώς και αποτελεσματικά τους αναλαμβανόμενους κινδύνους. Το ύψος των κεφαλαιακών αυτών απαιτήσεων προσδιορίζεται, κυρίως, από το μέγεθος και τη δομή των δανειακών χαρτοφυλακίων, την ποιότητα των οποίων ποσοτικοποιούν ορισμένες παράμετροι (πιθανότητα αθέτησης, άνοιγμα σε αθέτηση κ.ά.), όπως καθορίζει η επιλεγόμενη μέθοδος υπολογισμού με βάση τα Σύμφωνα της Βασιλείας Ι και ΙΙ, (Basel Committee on Banking Supervision (1988), Basel Committee on Banking Supervision (1999), Basel Committee on Banking Supervision (2004)). Ιδιαίτερη σημασία, για μια

ολοκληρωμένη αντιμετώπιση των κεφαλαιακών απαιτήσεων, που κάθε Τράπεζα πρέπει να διακρατεί, έχει η ανάπτυξη σεναρίων προσομοίωσης ακραίων καταστάσεων, που ακολουθούν τις απαιτήσεις τις οποίες θέτει το εποπτικό πλαίσιο, αλλά και τις υποθέσεις της ορθής και πλήρους εκτέλεσης διαχείρισης κινδύνων.

Ο βαθμός κάλυψης των αναλαμβανόμενων κινδύνων, σε σχέση με τα ανοίγματα που έχουν πραγματοποιηθεί, αντιμετωπίζεται και στα δύο Σύμφωνα της Βασιλείας με τον δείκτη κεφαλαιακής επάρκειας. Πράγματι, ο δείκτης κεφαλαιακής επάρκειας, αποτελεί ένα αξιόπιστο μέτρο αποτίμησης των κεφαλαιακών απαιτήσεων που πρέπει να διακρατούνται για την κάλυψη της Τράπεζας, έναντι των κινδύνων. Ειδικότερα, ο δείκτης κεφαλαιακής επάρκειας ορίζεται ως ένας λόγος, του οποίου ο αριθμητής συμπεριλαμβάνει το σύνολο των εποπτικών κεφαλαίων (βασικών και συμπληρωματικών) και ο παρονομαστής όλα τα ανοίγματα του ενεργητικού, σταθμισμένα, ως προς τους κινδύνους. Στον υπολογισμό του παρονομαστή συμπεριλαμβάνονται και άλλες πληροφορίες (καθυστερήσεις, εξασφαλίσεις κλπ.), οι οποίες δεν περιέχονται στις οικονομικές καταστάσεις και στα ετήσια δελτία, επειδή θεωρούνται ότι αφορούν εμπιστευτικά στοιχεία, των οποίων δεν επιτρέπεται η δημοσίευση. Επίσης, για τον υπολογισμό του δείκτη κεφαλαιακής επάρκειας λαμβάνονται υπόψη επιπρόσθετες κεφαλαιακές απαιτήσεις, έναντι του κινδύνου αγοράς και του λειτουργικού κινδύνου. Σημειώνεται ότι στο πρώτο Σύμφωνο της Βασιλείας, οι κίνδυνοι που λαμβάνονταν υπόψη ήταν ο πιστωτικός κίνδυνος και ο κίνδυνος της αγοράς, ενώ στο δεύτερο Σύμφωνο συμπεριλαμβάνεται και ο λειτουργικός κίνδυνος. Μάλιστα, ενώ ο υπολογισμός των κινδύνων με το πρώτο Σύμφωνο στηρίζεται στη χρήση προκαθορισμένων συντελεστών, οι οποίοι αντιστοιχούν σε κάθε άνοιγμα, το δεύτερο Σύμφωνο αναπτύσσει μία πιο σύνθετη μεθοδολογία, όπου λαμβάνονται υπόψη παράγοντες οι οποίοι προσδιορίζουν το βαθμό (υψηλός ή χαμηλός κίνδυνος), και, ανάλογα, το ύψος του ανοίγματος. Ειδικότερα, η Βασιλεία ΙΙ (Black and Scholes (1973)), προτείνει τη χρήση δύο προσεγγίσεων (Τυποποιημένη Μέθοδος (RW), Μέθοδος Εσωτερικών Διαβαθμίσεων (RWIRB) (Basel Committee on Banking Supervision (2008), Ozdemir and Miu (2009)). Η εφαρμογή της πρώτης προσέγγισης (Τυποποιημένη Μέθοδος) παρουσιάζει ομοιότητες με την αντίστοιχη προσέγγιση της Βασιλείας Ι (υπολογισμός κινδύνων βάσει εποπτικών συντελεστών), αλλά συμπεριλαμβάνει και επιπρόσθετους παράγοντες για τον υπολογισμό του τελικού αποτελέσματος (καθυστερήσεις, εξασφαλίσεις, διαβάθμιση πιστοληπτικής ικανότητας κλπ.). Η δεύτερη προσέγγιση (Μέθοδος Εσωτερικών Διαβαθμίσεων) είναι πληρέστερη και επιτρέπει την ανάπτυξη και χρήση υποδειγμάτων μέτρησης των τριών κινδύνων (πιστωτικός, αγοράς και λειτουργικός), μετά από τη σύμφωνη έγκριση και αποδοχή τους από την Κεντρική Τράπεζα.

Μέχρι τώρα, οι επιπτώσεις που θα επιφέρουν, επί των κεφαλαιακών απαιτήσεων του ελληνικού τραπεζικού συστήματος, τα διακρατούμενα κεφάλαια για την αντιστάθμιση των αναλαμβανόμενων κινδύνων, δεν έχουν σαφώς αποτιμηθεί ως προς την ύπαρξη, την έκταση και το μέγεθος των μεταβολών (Basel Committee on Banking Supervision (2013)). Σημαντικά ζητήματα, όπως ποιές κατηγορίες

Τραπεζών θα επηρεασθούν περισσότερο, ποιά τμήματα των χαρτοφυλακίων θα παρουσιάσουν τις μεγαλύτερες, ως προς το μέγεθος και την έκταση, μεταβολές, ποιός θα είναι ο βαθμός επίδρασής τους κλπ., τα οποία δεν αφορούν μόνο θέματα ποιοτικού χαρακτήρα, δεν έχουν τύχει της αναγκαίας ποσοτικής τεκμηρίωσης.

Η Επιτροπή της Βασιλείας έχει φροντίσει να εκπονηθούν μελέτες προσομοίωσης ακραίων καταστάσεων στις οποίες συμμετέχει μεγάλος αριθμός Τραπεζών, από τις χώρες μέλη της Επιτροπής, προκειμένου να διαμορφωθούν αξιόπιστες στρατηγικές διαχείρισης κινδύνου και να δημιουργηθεί ένα πλαίσιο προληπτικής εποπτείας των πιστωτικών ιδρυμάτων (Basel Committee on Banking Supervision (1988), Alexander (2008), Basel Committee on Banking Supervision (2009), Foglia (2009)). Επίσης, δημοσιεύθηκαν εμπειρικές αναλύσεις οι οποίες εξετάζουν θέματα σχετικά με την αντιμετώπιση ή αποφυγή ακραίων δυσμενών καταστάσεων (Financial Services Authority (2009), Haldane (2009), Bank of Greece (1999)). Στην εργασία αυτή εφαρμόζεται η μέθοδος εσωτερικών διαβαθμίσεων και επιχειρείται να εξετασθεί, για συνθήκες ακραίων καταστάσεων, η επίδραση του Συμφώνου της Βασιλείας ΙΙ στις ελληνικές Τράπεζες που δραστηριοποιούνται στην Ελλάδα, είναι εισηγμένες στο Ελληνικό Χρηματιστήριο Αξιών και εποπτεύονται από την Κεντρική Τράπεζα της Ελλάδος.

2. ΔΕΔΟΜΕΝΑ ΚΑΙ ΑΝΑΠΤΥΞΗ ΥΠΟΔΕΙΓΜΑΤΟΣ

2.1. Γενικά

Στην εμπειρική μελέτη εφαρμόσθηκε κατάλληλο υπόδειγμα το οποίο βασίσθηκε στις απαιτήσεις που ορίζει το δεύτερο Σύμφωνο της Βασιλείας, στις αρχές που θέτουν τα Διεθνή Λογιστικά Πρότυπα καθώς και στη λειτουργική δομή των ελληνικών Τραπεζών. Κατά τη σύνθεση του υποδείγματος, ελήφθη υπόψη ότι κάθε Τράπεζα χαρακτηρίζεται από ιδιαιτερότητες, που αφορούν το είδος και το εύρος των εργασιών της, τους αναλαμβανόμενους τραπεζικούς κινδύνους και την αποτελεσματικότητα της πολιτικής καθορισμού των απαιτούμενων ιδίων κεφαλαίων, που χρήζουν εξειδικευμένης διερεύνησης και αξιολόγησης.

Το υπόδειγμα δέχεται στοιχεία που χαρακτηρίζουν το είδος των πιστωτικών ανοιγμάτων, δηλαδή το είδος του προϊόντος, το υπόλοιπο του ανοίγματος, την υφιστάμενη καθυστέρηση, το ύψος της κάλυψης/ καλύμματος του ανοίγματος, τις προβλέψεις έναντι των αναλαμβανόμενων κινδύνων, την πιθανότητα αδυναμίας εκπλήρωσης (αθέτησης) της υποχρέωσης του αντισυμβαλλόμενου (Probability of Default – PD), τη ζημία σε περίπτωση αδυναμίας εκπλήρωσης της υποχρέωσης του αντισυμβαλλόμενου (Loss Given Default – LGD) και την έκθεση έναντι αντισυμβαλλόμενου (Exposure At Default – EAD). Έτσι, το υπόδειγμα είναι σε θέση να εκτιμά τον κίνδυνο των τραπεζικών χαρτοφυλακίων, αφού προσμετρά τον κίνδυνο από δανεισμό, προκειμένου να προσδιοριστούν οι αντίστοιχες κεφαλαιακές απαιτήσεις, έναντι του πιστωτικού κινδύνου καθώς και η κεφαλαιακή επάρκεια των Τραπεζών.

Τα στοιχεία που τροφοδοτούν το υπόδειγμα προέρχονται από ισολογισμούς, ετήσιες αναφορές και εποπτικές αναφορές για το έτος 2007. Τα δεδομένα, που συλλέχθησαν, αφορούν: Οικονομικά στοιχεία και καθυστερήσεις, που εμφανίζουν τα χαρτοφυλάκια των ελληνικών Τραπεζών, προβλέψεις και τιμές των παραμέτρων, τις οποίες τα Σύμφωνα Βασιλείας και κατ'επέκταση η Κεντρική Τράπεζα της Ελλάδος επιβάλλουν, για την εκτίμηση του πιστωτικού και λειτουργικού κινδύνου και για τον υπολογισμό της κεφαλαιακής επάρκειας των Τραπεζικών Ιδρυμάτων. Σημειώνεται ότι το έτος 2007 είναι ένα κρίσιμο μεταβατικό χρονικό διάστημα για τις ελληνικές Τράπεζες, αφού από τον Αύγουστο του 2007 η χρηματοπιστωτική κρίση άρχισε να ασκεί ανοδικές πιέσεις στα τραπεζικά επιτόκια, στις αγορές χρήματος των αναπτυγμένων οικονομιών, ενώ τα νομισματικά και πιστωτικά μεγέθη εξακολουθούσαν να εμφανίζουν μεγάλη αύξηση.

Αρχικά, επιχειρήθηκε διαχωρισμός του δείγματος των Τραπεζών που περιέχει τις κυριότερες ελληνικές Τράπεζες σε 3 κατηγορίες, με κριτήριο το μέγεθος του ενεργητικού τους. Σημειώνεται ότι στο δείγμα ελήφθησαν τα οικονομικά στοιχεία 17 ελληνικών Τραπεζών που δραστηριοποιούνται στην Ελλάδα, είναι εισηγμένες στο Ελληνικό Χρηματιστήριο, εποπτεύονται από την Κεντρική Τράπεζα της Ελλάδος και το ενεργητικό τους καλύπτει το 80 – 90% του ενεργητικού ολόκληρου του τραπεζικού κλάδου της χώρας. Η εφαρμογή της μεθοδολογίας, που αναπτύχθηκε (Hartigan (1975), Bartholomew et al. (2002)), μας οδήγησε σε ομαδοποίηση των Τραπεζών στις εξής κατηγορίες: (α) Την κατηγορία των μεγάλων Τραπεζών που περιέχει τις τέσσερις μεγαλύτερες Τράπεζες (Εθνική Τράπεζα, Τράπεζα Πειραιώς, Alpha Bank και Eurobank). (β) Την κατηγορία των μεσαίου μεγέθους Τραπεζών που περιλαμβάνει τις επόμενες τέσσερις σε μέγεθος Τράπεζες (Εμπορική Τράπεζα, Αγροτική Τράπεζα Ελλάδος, Ταχυδρομικό Ταμιευτήριο και Marfin Bank). (γ) Την κατηγορία των μικρών Τραπεζών που περιλαμβάνει τις δέκα μικρότερες Τράπεζες (Millenium Bank, Γενική Τράπεζα, Τράπεζα Αττικής, Ασπίς Τράπεζα, Probank, Proton Bank, FBB, Aegean Baltic, Επενδυτική Τράπεζα και Πανελλήνια Τράπεζα).

Κατά την ανάπτυξη του υποδείγματος, υπολογίσθηκε το κατά περίπτωση σταθμισμένο ενεργητικό. Ακολούθως, εφαρμόσθηκαν σενάρια στα οποία μεταβάλλονται τα στοιχεία κινδύνου, όπως η μεταβολή του ύψους των ανοιγμάτων, η μεταβολή της πιθανότητας αδυναμίας εκπλήρωσης της υποχρέωσης του αντισυμβαλλόμενου, η μεταβολή της ζημίας σε περίπτωση αδυναμίας εκπλήρωσης της υποχρέωσης του αντισυμβαλλόμενου και, στη συνέχεια, υπολογίσθηκαν εκ νέου οι προκύπτουσες κεφαλαιακές απαιτήσεις, με βάση τις νέες συνθήκες (επιδεινωμένες ή μη). Δηλαδή, εφαρμοζόντας τη μέθοδο εσωτερικών διαβαθμίσεων, διενεργήσαμε σενάρια προσομοίωσης ακραίων καταστάσεων για να εκτιμήσουμε την επίπτωση των ακραίων συνθηκών της αγοράς στην οικονομική κατάσταση της Τράπεζας. Στην προκειμένη περίπτωση, βρήκαμε ποιά είναι η επίπτωση της αύξησης των δανείων σε αθέτηση στο δείκτη κεφαλαιακής επάρκειας. Η αύξηση των δανείων σε αθέτηση, θα επηρεάσει κατά πρώτο λόγο τις προβλέψεις και, συνεπώς, την κερδοφορία και τα κεφάλαια και κατά δεύτερο λόγο θα αυξήσει το σταθμισμένο ενεργητικό. Στο σύνολό

τους, οι επιπτώσεις αυτές θα αποτυπωθούν ως μείωση του δείκτη κεφαλαιακής επάρκειας.

2.2. Προσέγγιση Εσωτερικών Διαβαθμίσεων

Για τον υπολογισμό των κεφαλαιακών απαιτήσεων με την προσέγγιση εσωτερικών διαβαθμίσεων πρώτα διακρίθηκαν τα ανοίγματα σε κατηγορίες ανάλογα με τον εποπτικό χειρισμό τους. Στη συνέχεια, χρησιμοποιήθηκε η αναλογιστική μέθοδος, δηλαδή ο διαχωρισμός των ανοιγμάτων σε κατηγορίες ομοειδών δανείων, ως προς τον κίνδυνο που ενέχουν. Ακολούθως, με τη χρήση της προσέγγισης εσωτερικών διαβαθμίσεων επιχειρήθηκε να εξετασθεί αν οι μεταβολές που εισάγονται με το Σύμφωνο της Βασιλείας ΙΙ, οδηγούν σε μείωση ή αύξηση των κεφαλαιακών απαιτήσεων και να εκτιμηθεί αυτή τη μεταβολή. Για κάθε Τράπεζα εφαρμόσθηκαν τρία σενάρια τα οποία στηρίζονται στο βαθμό κάλυψης των αθετημένων δανείων από προβλέψεις. Το ποσοστό κάλυψης των καθυστερήσεων από προβλέψεις αποτελεί βασική παράμετρο για τον τελικό υπολογισμό των κεφαλαιακών απαιτήσεων που θα διακρατήσουν οι Τράπεζες. Πιο συγκεκριμένα, μεγαλύτερη κάλυψη των καθυστερήσεων από προβλέψεις θα οδηγήσει στη διακράτηση χαμηλότερων κεφαλαιακών απαιτήσεων. Το ποσοστό αυτό, εξαρτάται εκτός από το μέγεθος της Τράπεζας και από την πολιτική προβλέψεων της Τράπεζας καθώς και από την επάρκεια των προβλέψεών της (Bank of Greece (1999)).

Στους Πίνακες, που ακολουθούν, δίνονται τα αποτελέσματα που προέκυψαν από τα σχετικά σενάρια που εφαρμόστηκαν για κάθε Τράπεζα.

Στους Πίνακες 1, 3 και 5 η πρώτη στήλη αφορά τα είδη των χαρτοφυλακίων της Τράπεζας. Οι τρεις επόμενες στήλες αναφέρονται στα οικονομικά στοιχεία των χαρτοφυλακίων καθώς και στις σταθμίσεις των κινδύνων που ορίζουν για κάθε άνοιγμα οι προσεγγίσεις της Βασιλείας Ι (RWBασιλείαΙ) και Βασιλείας ΙΙ (RWIRB). Η αναμενομένη ζημιά (EL), είναι το ποσό που έχει αναγνωρίσει η Τράπεζα ως αναμενομένη αξία για το άνοιγμα. Στην προκειμένη περίπτωση, το ποσοστό που αναφέρεται στον Πίνακα 1 είναι η πιθανότητα αθέτησης επί τη ζημία σε περίπτωση αθέτησης. Αποτελεί μια ένδειξη του κινδύνου που ενέχει το κάθε άνοιγμα. Ομοίως, ορίζονται η έκτη και η έβδομη στήλη, που απεικονίζουν τα αποτελέσματα του σεναρίου ακραίων καταστάσεων για την Τράπεζα. Για τα ανοίγματα που είναι σε αθέτηση, η πιθανότητα αθέτησης έχει θεωρηθεί ίση με 100%, με αποτέλεσμα να εμφανίζονται πολύ υψηλά ποσοστά, τα οποία οφείλονται στη ζημία σε περίπτωση αθέτησης.

Σχετικά με τους Πίνακες 2, 4 και 6 σημειώνεται ότι : (α) Η πρώτη γραμμή (% Αύξηση του PD), αφορά όλα τα σενάρια αύξησης της πιθανότητας αθέτησης που εκτελέστηκαν, με ευνοϊκότερο την αύξηση της πιθανότητας σε 0%, και δυσμενέστερο την αύξηση της πιθανότητας κατά 160%. (β) Η δεύτερη γραμμή (RWAIRBΠιστωτικός) αποτελεί το συνολικό απόθεμα κεφαλαίων που θα πρέπει να διακρατά η Τράπεζα για όλη την κλίμακα των σεναρίων αύξησης της πιθανότητας που υϊοθετήθηκε, για την αντιμετώπιση του πιστωτικού κινδύνου. (γ) Η τρίτη γραμμή (RWAΛειτουργικός) αποτελεί τις κεφαλαιακές απαιτήσεις που η Τράπεζα 1 πρέπει

να διακρατά, έναντι του λειτουργικού κινδύνου. (δ) Η τέταρτη γραμμή (RWA Έλλειμμα Προβλέψεις), αντιστοιχεί στις επιπρόσθετες κεφαλαιακές απαιτήσεις, που οφείλονται στις ελλιπείς προβλέψεις, που πραγματοποιήθηκαν, αφού το ενδεχόμενο πραγματοποίησης αυτής της κατάστασης, αποτελεί μια ακραία έκφραση, η οποία ήταν μη αναμενόμενη από την Τράπεζα. Καθώς αυξάνεται το ποσοστό της πιθανότητας των αθετήσεων και το έλλειμμα των προβλέψεων θα αυξάνεται ανάλογα. (ε) Η πέμπτη γραμμή (Συνολικό RWA (εκτός κινδύνου αγοράς)) αποτελεί το τελικό σύνολο των κεφαλαιακών απαιτήσεων που αντιστοιχεί για κάθε σενάριο αύξησης της πιθανότητας των αθετήσεων. Στα σταθμισμένα αυτά στοιχεία ενεργητικού, δεν συμπεριλαμβάνονται οι κεφαλαιακές απαιτήσεις έναντι του κινδύνου αγοράς, επειδή ο κίνδυνος αυτός δεν αποτελεί σημείο εξέτασης στην παρούσα εργασία. (στ) Η έκτη γραμμή (% Αύξηση RWA), αποτελεί την ποσοστιαία έκφραση της μεταβολής των επιπρόσθετων κεφαλαιακών απαιτήσεων που διακρατούνται, ανάλογα με το σενάριο αύξησης της πιθανότητας αθέτησης. Αύξηση της πιθανότητας, όπως είναι αναμενόμενο, θα οδηγήσει σε αύξηση του ποσοστού αυτού.

Για την Τράπεζα 1 τα αποτελέσματα συνοψίζονται στους Πίνακες 1 και 2.

Πίνακας 1 *Ανάλυση Μεγεθών ανά Χαρτοφυλάκιο της Τράπεζας 1 για Κανονικές Συνθήκες και Συνθήκες Έκτακτων Καταστάσεων, σε χιλ. Ευρώ*

Τράπεζα 1			**Κανονικές Συνθήκες**		**Stress +100%**	
Είδος Ανοίγματος	**Ισολογισμός**	**RW Βασιλεία I**	**RW IRB**	**%EL**	**RW IRB**	**%EL**
Μικρών Επιχ. (μη αθετημένα)	20,317,708.53	100%	94.57%	1.52%	120.65%	3.03%
Μικρών Επιχ. (αθετημένα)	882,523.47	100%	0.00%	45.10%	0.00%	45.10%
Μεγάλων Επιχ. (μη αθετημένα)	63,693,041.47	100%	101.56%	0.74%	132.06%	1.48%
Μεγάλων Επιχ. (αθετημένα)	2,719,584.53	100%	0.00%	52.01%	0.00%	52.01%
Στεγαστικά LTV<75% (μη αθετημένα)	30,035,466.00	50%	24.44%	0.17%	38.88%	0.34%
Στεγάστηκα LTV<75% (αθετημένα)	1,093,453.00	50%	0.00%	30.00%	0.00%	30.00%
Στεγάστηκα LTV>75% (μη αθετημένα)	3,397,417.00	50%	40.73%	0.29%	64.80%	0.57%
Στεγάστηκα LTV>75% (αθετημένα)	152,839.00	50%	0.00%	50.00%	0.00%	50.00%
Καταναλωτικά (μη αθετημένα)	19,329,601.00	100%	55.10%	2.02%	86.83%	4.04%
Καταναλωτικά (αθετημένα)	901,719.00	100%	0.00%	65.54%	0.00%	65.54%

Ακολουθεί, ο Πίνακας 2 των ακραίων σεναρίων αύξησης των πιθανοτήτων αθέτησης, που πραγματοποιήθηκαν για την Τράπεζα 1.

Πίνακας 2 *Ακραία Σενάρια για Διάφορα Επίπεδα Μεταβολής της Πιθανότητας Αθέτησης για την Τράπεζα 1, σε χιλ. Ευρώ*

Τράπεζα 1					
%Αύξηση του PD	**RWA IRB Πιστωτικός**	**RWA Λειτουργικού**	**RWA Έλλειμμα Προβλέψεις**	**Συνολικό RWA (εκτός κινδύνου αγοράς)**	**%Αύξηση RWA**
0%	103,276,900	11,738,372	17,298,780	**132,314,053**	**-**
20%	111,967,114	11,738,372	20,376,918	**144,082,404**	**8.89%**
40%	119,705,155	11,738,372	23,455,055	**154,898,583**	**17.07%**
60%	126,738,816	11,738,372	26,533,192	**165,010,382**	**24.71%**
80%	133,230,889	11,738,372	29,611,329	**174,580,592**	**31.94%**
100%	139,292,584	11,738,372	32,689,466	**183,720,424**	**38.85%**
120%	145,002,340	11,738,372	35,767,603	**192,508,317**	**45.49%**
140%	150,416,993	11,738,372	38,845,741	**201,001,108**	**51.91%**
160%	155,578,701	11,738,372	41,923,878	**209,240,953**	**58.14%**

Η σύνθεση των σεναρίων που αναπτύχθηκαν οδηγεί στην παρακάτω εκτιμημένη συνάρτηση παλινδρόμησης του σταθμισμένου ενεργητικού (Y).

$$\hat{Y} = 0.3605PD - 0.3411 \qquad R^2 = 0.9567 \tag{1}$$

Η κλίση της συνάρτησης (1), είναι ίση με 0,36. Δηλαδή, μία αύξηση της πιθανότητας αθέτησης (PD) κατά x% προκαλεί αύξηση του σταθμισμένου ενεργητικού κατά 0.36 * x%.

Ομοίως, για την Τράπεζα 2 και Τράπεζα 3, πραγματοποιείται μια σειρά από σενάρια προσομοίωσης. Η μεθοδολογία που εφαρμόζεται είναι η ίδια όπως και στην περίπτωση της Τράπεζας 1.

Στους Πίνακες 3 και 4 που έπονται εμφανίζονται τα αποτελέσματα που αφορούν την Τράπεζα 2.

Πίνακας 3. *Ανάλυση Μεγεθών ανά Χαρτοφυλάκιο της Τράπεζας 2 για Κανονικές Συνθήκες και Συνθήκες Ακραίων Καταστάσεων, σε χιλ Ευρώ*

Τράπεζα 2			**Κανονικές Συνθήκες**		**stress +100%**	
Είδος Ανοίγματος	**Ισολογισμός**	**RW ΒασιλείαΙ**	**RW IRB**	**%EL**	**RW IRB**	**%EL**
Μικρών Επιχ. (μη αθετημένα)	6,784,562.74	100%	88.95%	1.32%	113.0%	2.6%
Μικρών Επιχ. (αθετημένα)	541,833.26	100%	0.00%	41.09%	0.0%	41.1%
Μεγάλων Επιχ. (μη αθετημένα)	12,787,654.26	100%	109.37%	0.95%	141.1%	1.9%
Μεγάλων Επιχ. (αθετημένα)	2,004,114.74	100%	0.00%	47.63%	0.0%	47.6%
Στεγαστικά LTV<75% (μη αθετημένα)	15,065,652.00	50%	15.38%	0.10%	24.6%	0.2%
Στεγαστικά LTV<75% (αθετημένα)	826,638.00	50%	0.00%	20.80%	0.0%	20.8%
Στεγαστικά LTV>75% (μη αθετημένα)	1,380,297.00	50%	33.28%	0.22%	53.3%	0.4%
Στεγαστικά LTV>75% (αθετημένα)	86,289.00	50%	0.00%	45.00%	0.0%	45.0%
Καταναλωτικά (μη αθετημένα)	5,891,266.00	100%	68.88%	2.78%	106.0%	5.6%
Καταναλωτικά (αθετημένα)	617,096.00	100%	0.00%	70.00%	0.0%	70.0%

Ακολουθεί, ο Πίνακας 4 των ακραίων σεναρίων αύξησης των πιθανοτήτων αθέτησης, που πραγματοποιήθηκαν για την Τράπεζα 2.

Πίνακας 4. *Διενέργεια Ακραίων Σεναρίων για Διάφορα Επίπεδα Μεταβολών Πιθανότητας Αθέτησης για την Τράπεζα 2, σε χιλ Ευρώ*

Τράπεζα 2					
%Αύξηση του PD	**RWA IRB Πιστωτικός**	**RWA Λειτουργικού**	**RWA Έλλειμμα Προβλέψεις**	**Συνολικό RWA (εκτός κινδύνου αγοράς)**	**%Αύξηση RWA**
0%	26,855,534	4,049,827	896,262	**31,801,624**	**-**
20%	29,126,924	4,049,827	1,879,899	**35,056,652**	**10.24%**
40%	31,166,858	4,049,827	2,863,537	**38,080,223**	**19.74%**
60%	33,035,450	4,049,827	3,847,174	**40,932,452**	**28.71%**
80%	34,771,310	4,049,827	4,830,811	**43,651,950**	**37.26%**
100%	36,400,240	4,049,827	5,814,449	**46,264,517**	**45.48%**
120%	37,940,052	4,049,827	6,798,086	**48,787,967**	**53.41%**
140%	39,403,402	4,049,827	7,781,723	**51,234,955**	**61.11%**
160%	40,799,518	4,049,827	8,765,361	**53,614,708**	**68.59%**

Η εκτιμημένη συνάρτηση παλινδρόμησης του σταθμισμένου ενεργητικού (Y), για τα σενάρια που εφαρμόστηκαν, είναι η παρακάτω :

$$\hat{Y} = 0.4250PD - 0.4060 \qquad R^2 = 0.9575 \tag{2}$$

Δηλαδή, μία αύξηση της πιθανότητας αθέτησης (PD) κατά x% προκαλεί αύξηση του σταθμισμένου ενεργητικού κατά: 0.425 * x%.

Για την Τράπεζα 3, εφαρμόζονται, επίσης, μια σειρά από σενάρια προσομοίωσης. Η μεθοδολογία είναι η ίδια με την περίπτωση της Τράπεζα 1 και της Τράπεζας 2, τα δε αποτελέσματα που προκύπτουν απεικονίζονται στους Πίνακες 5 και 6.

Πίνακας 5. *Ανάλυση Μεγεθών ανά Χαρτοφυλάκιο της Τράπεζας 3 για Κανονικές Συνθήκες και Συνθήκες Ακραίων Καταστάσεων, σε χιλ. Ευρώ*

Τράπεζα 3			**Κανονικές Συνθήκες**		**stress +100%**	
Είδος ανοίγματος	**Ισολογισμός**	**RW Βασιλεία I**	**RW IRB**	**% EL**	**RW IRB**	**%EL**
Μικρών Επιχ. (μη αθετημένα)	3,029,998.20	100%	92.72%	1.41%	117.92%	2.81%
Μικρών Επιχ. (αθετημένα)	133,887.90	100%	0.00%	43.64%	0.00%	43.64%
Μεγάλων Επιχ. (μη αθετημένα)	8,787,838.80	100%	102.32%	0.89%	132.03%	1.77%
Μεγάλων Επιχ. (αθετημένα)	562,380.10	100%	0.00%	44.60%	0.00%	44.60%
Στεγάστηκα LTV<75% (μη αθετημένα)	2,707,838.00	50%	14.79%	0.10%	23.67%	0.20%
Στεγαστικά LTV<75% (αθετημένα)	96,409.00	50%	0.00%	20.00%	0.00%	20.00%
Στεγαστικά LTV>75% (μη αθετημένα)	1,227,767.00	50%	33.28%	0.22%	53.25%	0.45%
Στεγαστικά LTV>75% (αθετημένα)	86,822.00	50%	0.00%	45.00%	0.00%	45.00%
Καταναλωτικά (μη αθετημένα)	2,007,165.00	100%	75.77%	3.06%	116.58%	6.12%
Καταναλωτικά (αθετημένα)	301,757.00	100%	0.00%	77.00%	0.00%	77.00%

Ακολουθεί, ο Πίνακας 6 των ακραίων σεναρίων αύξησης των πιθανοτήτων αθέτησης, που πραγματοποιήθηκαν για την Τράπεζα 3.

Η εκτιμημένη συνάρτηση παλινδρόμησης του σταθμισμένου ενεργητικού (Y), για τα σενάρια που εφαρμόστηκαν, είναι η εξής:

$$\hat{Y} = 0.3665PD - 0.3483 \qquad R^2 = 0.9563 \tag{3}$$

Δηλαδή, μία αύξηση της πιθανότητας αθέτησης (PD) κατά x% προκαλεί αύξηση του σταθμισμένου ενεργητικού κατά 0.3665 * x%.

Πίνακας 6. *Διενέργεια Ακραίων Σεναρίων για Διάφορα Επίπεδα Μεταβολών Πιθανότητας Αθέτησης για την Τράπεζα 3, σε χιλ. Ευρώ*

Τράπεζα 3					
%Αύξηση του PD	**RWA IRB Πιστωτικός**	**RWA Λειτουργικού**	**RWA Έλλειμμα Προβλέψεις**	**Συνολικό RWA (εκτός κινδύνου αγοράς)**	**%Αύξηση RWA**
0%	14,130,925	1,470,158	2,278,487	**17,879,571**	**-**
20%	15,256,094	1,470,158	2,746,897	**19,473,151**	**8.91%**
40%	16,258,985	1,470,158	3,215,307	**20,944,452**	**17.14%**
60%	17,172,713	1,470,158	3,683,717	**22,326,590**	**24.87%**
80%	18,018,469	1,470,158	4,152,127	**23,640,757**	**32.22%**
100%	18,810,337	1,470,158	4,620,537	**24,901,035**	**39.27%**
120%	19,557,971	1,470,158	5,088,948	**26,117,079**	**46.07%**
140%	20,268,164	1,470,158	5,557,358	**27,295,682**	**52.66%**
160%	20,945,811	1,470,158	6,025,768	**28,441,740**	**59.07%**

3. ΣΥΜΠΕΡΑΣΜΑΤΙΚΕΣ ΠΑΡΑΤΗΡΗΣΕΙΣ

Για να εξετασθούν οι επιπτώσεις στην κεφαλαιακή επάρκεια των ελληνικών τραπεζών από την εφαρμογή της προσέγγισης εσωτερικών διαβαθμίσεων, εκτελέστηκε μια σειρά σεναρίων ακραίων καταστάσεων με ομοιόμορφη αύξηση της πιθανότητας αθέτησης, για όλα τα χαρτοφυλάκια. Μια τέτοια αύξηση επιδρά στις κεφαλαιακές απαιτήσεις, αυξάνοντας τις προβλέψεις και αυξάνοντας τους δείκτες στάθμισης. Επιδρά, επίσης, και στον αριθμητή και στον παρονομαστή του δείκτη κεφαλαιακής επάρκειας. Για να υπάρχει ενιαία αντιμετώπιση, αποτυπώθηκαν οι δύο παραπάνω επιδράσεις ως αύξηση του σταθμισμένου ενεργητικού (παρονομαστής). Τα αποτελέσματα που ελήφθησαν είναι τα εξής: Το σταθμισμένο ενεργητικό αυξάνει σχεδόν γραμμικά, σε σχέση με την αύξηση της πιθανότητας αθέτησης. Πιο συγκεκριμένα, κάθε αύξηση 10% του PD προκαλεί αύξηση στο σταθμισμένο ενεργητικό κατά 3,6%, 4,3% και 3,7%, για τις Τράπεζες, δηλαδή για Τράπεζα 1, Τράπεζα 2 και Τράπεζα 3, αντίστοιχα. Αυτό σημαίνει ότι, η Τράπεζα 2 είναι πιο ευαίσθητη, από τις δύο άλλες, στους ελέγχους ακραίων καταστάσεων. Όμως, εδώ δεν παίζει ρόλο μόνο η κατανομή των ανοιγμάτων στα διάφορα χαρτοφυλάκια, αλλά και τα PD και LGD των ανοιγμάτων.

Σε όλες τις Τράπεζες, σύμφωνα με τα σενάρια που πραγματοποιήθηκαν, παρατηρείται πτώση του δείκτη κεφαλαιακής επάρκειας. Αυτό σημαίνει ότι σε όλες τις Τράπεζες η δυνατότητα ανάπτυξης περιορίζεται και υπάρχει ανάγκη αποτελεσματικότερης διαχείρισης του κινδύνου. Ειδικά, για την Τράπεζα 2, που ο δείκτης κεφαλαιακής επάρκειας ήταν ήδη οριακός, θα απαιτηθεί αύξηση κεφαλαίου. Ισχυρότερη Τράπεζα δείχνει ότι είναι η Τράπεζα 1, η οποία έχει και περισσότερα κεφάλαια και υψηλότερο σχετικό δείκτη. Η μέθοδος εσωτερικών διαβαθμίσεων μπορεί να πλήξει την κεφαλαιακή της επάρκεια, αν αναλάβει υψηλότερους κινδύνους από αυτούς που αντέχει. Φαίνεται ότι, όπως είναι το χαρτοφυλάκιο της Τράπεζας

αυτής, ευνοούνται τα δάνεια λιανικής τραπεζικής. Παρόλα αυτά, η Τράπεζα αυτή θα μπορούσε να αναλάβει και καλής ποιότητας μεγάλα επιχειρηματικά δάνεια. Η Τράπεζα 3 έρχεται δεύτερη, αλλά λόγω του μικρού μεγέθους της, καθώς και των κεφαλαίων που έχει, δε θα μπορούσε να αναλάβει κινδύνους, έναντι μεγάλων επιχειρηματικών δανείων. Με «προσεκτική» πολιτική θα μπορούσε να επεκταθεί στο χώρο της λιανικής τραπεζικής. Τέλος, η Τράπεζα 2 έχει ανάγκη αύξησης κεφαλαίων. Ο δείκτης της με τη μέθοδο εσωτερικών διαβαθμίσεων διαμορφώνεται κάτω από το 8%, πράγμα που σημαίνει ότι δεν έχει καθόλου δυνατότητες ανάπτυξης. Ενδεχομένως, την Τράπεζα αυτή συμφέρει να εφαρμόζει την τυποποιημένη προσέγγιση και όχι τη μέθοδο εσωτερικών διαβαθμίσεων. Σε περίοδο κρίσης, οι κεφαλαιακές απαιτήσεις θα πληγούν διπλά και από την αύξηση των αθετήσεων και από την αύξηση της πιθανότητας αθέτησης. Αυτό σημαίνει ότι οι Τράπεζες θα πρέπει να λάβουν έγκαιρα μέτρα, προκειμένου να αντισταθμίσουν αυτόν τον κίνδυνο. Στη μέθοδο εσωτερικών διαβαθμίσεων αυτό είναι δυνατόν, αφού υπάρχουν τα εργαλεία και τα υποδείγματα για τη μέτρηση και τη διαχείριση του πιστωτικού κινδύνου. Η πιστοδοτική πολιτική μπορεί να καθοριστεί κατάλληλα, μέσω των συστημάτων βαθμολόγησης των χορηγήσεων, των εξασφαλίσεων, της κατάλληλης τιμολόγησης και διαχείρισης κινδύνων, τόσο στο επίπεδο πελάτη, όσο και στο επίπεδο χαρτοφυλακίου.

Με την εφαρμογή της μεθόδου εσωτερικών διαβαθμίσεων παρατηρήθηκε σημαντική κεφαλαιακή ελάφρυνση, όταν το χαρτοφυλάκιο είναι καλής ποιότητας. Η Τράπεζα, σε αυτή την περίπτωση, θα έχει το όφελος να έχει χαμηλές κεφαλαιακές απαιτήσεις σε καλές περιόδους, ενώ θα πάρει το ρίσκο να έχει υψηλές κεφαλαιακές απαιτήσεις σε κακές περιόδους. Δηλαδή, η ευαισθησία, που παρατηρήσαμε να εμφανίζεται στους υπολογισμούς των παραμέτρων κινδύνων, μπορεί να σημαίνει πολύ μεγάλες μεταβολές στις κεφαλαιακές απαιτήσεις στις κακές περιόδους. Τότε η Τράπεζα θα πρέπει ή να μπορεί, έγκαιρα, να σφίξει την πιστοδοτική της πολιτική, ώστε να αποφύγει την επιδείνωση του χαρτοφυλακίου (αυτό σημαίνει να έχουν αναπτυχθεί κατάλληλοι πιστωτικοί και διαχείρισης κινδύνων μηχανισμοί – credit και risk management –) ή στη χειρότερη περίπτωση να υπάρχει η δυνατότητα να ενισχυθεί κεφαλαιακά. Κάτι τέτοιο φαίνεται ρεαλιστικό για τις μεγαλύτερες Τράπεζες που έχουν τη δυνατότητα να επενδύουν, σε ανθρώπινο εξειδικευμένο δυναμικό, αναγκαίες διαδικασίες και πληροφοριακά συστήματα για την αποτελεσματικότερη διαχείριση κινδύνου, αλλά που διαθέτουν και τα κεφάλαια να απορροφήσουν τον κίνδυνο, όταν αυτός ξεπεράσει κάποια όρια. Οι μικρότερες Τράπεζες, για να αποφύγουν υψηλές κεφαλαιακές απαιτήσεις, θα πρέπει να επενδύσουν σε προϊόντα χαμηλού πιστωτικού κινδύνου, όπως σε δάνεια προς υψηλής ποιότητας πιστούχους, μεγάλης διασποράς και καλών εξασφαλίσεων. Η απόδοση, όμως, τέτοιας μορφής χαρτοφυλακίων θα είναι πολύ χαμηλή. Αξίζει να σημειωθεί ότι και μια μεγαλύτερη Τράπεζα θα κινδύνευε από την υιοθέτηση της μεθόδου εσωτερικών διαβαθμίσεων, αν δεν δύναται να ελέγξει επαρκώς τον πιστωτικό κίνδυνο. Ένα κακό χαρτοφυλάκιο θα μπορούσε να οδηγήσει τον δείκτη κεφαλαιακής επάρκειας σε κατάρρευση. Η χρήση της μεθόδου εσωτερικών διαβαθμίσεων αναγκάζει την Τράπεζα, ανεξάρτητα του μεγέθους της, να αποκτήσει μια πιο υπεύθυνη στάση στον έλεγχο του πιστωτικού

κινδύνου και να εστιάσει την προσοχή της στη μέτρηση, στη διαχείριση και στην αντιστάθμισή του.

ABSTRACT

In this paper we study the impact of extreme events on the loan portfolios of the Greek banking system. These portfolios are grouped into three separate groups based on the size of the bank to which they belong and, in particular, large, medium, and small size. A series of extreme scenarios was performed and the increase in capital requirements was calculated for each scenario based on the internal ratings approach of the Basel II accord. The results obtained show an increase of credit risk during the crisis periods, and the differentiation of risk depending on the size of the banking organization as well as the added capital that will be needed in order to hedge that risk.

ΑΝΑΦΟΡΕΣ

Alexander, C. and E. Sheedy. Developing a stress testing framework based on market risk models, Journal of Banking and Finance, 2008, 32:10, 2220-2236.

Bank of Greece. Bank of Greece Governor's Act 2442/19 January 1999; Bank of Greece: Athens, Greece, 1999.

Bartholomew, D. Multivariate Analysis Made Clear; Chapman & Hall: New York, NY, USA, 2002.

Basel Committee on Banking Supervision. International Convergence of Capital Measurement and Capital Standards: A Revised Framework. Available online: http://www.bis.org/publ/bcbs107.htm, Bank for International Settlements Press & Communications: 2004.

Basel Committee on Banking Supervision. International Coverage of Capital Measurement and Capital Standards: Available online: http://www.bis.org/publ/bcbsc111.pdf (accessed on 20 March 2015), Bank for International Settlements Press & Communications: 1988.

Basel Committee on Banking Supervision. Proposed revisions to the Basel II market risk framework: Available online: http://www.bis.org/publ/bcbs140.pdf (accessed on 12 November 2015), Bank for International Settlements Press & Communications: 2008.

Basel Committee on Banking Supervision. A New Capital Adequacy Framework. Available online: http://www.bis.org/publ/bcbs50.pdf (accessed on 20 March 2015), Bank for International Settlements Press & Communications: 1999.

Basel Committee on Banking Supervision. Regulatory Consistency Assessment Program (RCAP). Analysis of Risk-Weighted Assets for Credit Risk in the Banking Book. Available online: http://www.bis.org/publ/bcbs256.pdf (accessed on 20 March 2015), Bank for International Settlements Press & Communications: 2013.

Basel Committee on Banking Supervision. Principles for sound stress testing practices and supervision, Available at: http://www.bis.org/publ/bcbs155.pdf (accessed on 20 March 2015), Bank for International Settlements Press & Communications: 2009.

Black, F.; Scholes, M. The pricing of options and corporate liabilities. J. Polit. Econ. 1973, 81, 637–654.

Financial Services Authority. FSA Statement on Its Use of Stress Tests; FSA: London, United Kingdom, 2009.

Foglia. A. Stress Testing Credit Risk: A Survey of Authorities' Approaches. ,2009, Available online: http://www.ijcb.org/journal/ijcb09q3a1.htm (accessed on 20 March 2015).

Haldane, A.G. Why Banks Failed the Stress Test. Speech Given at the 2009 Marcus-Evans Conference on Stress-Testing. Available online: http://www.bankofengland.co.uk/archive/ documents/historicpubs/speeches/2009/speech374.pdf (accessed on 20 March 2015).

Hartigan, J.A. Clustering Algorithms; John Wiley & Sons, Inc.: New York, NY, USA, 1975.

Ozdemir, B.; Miu, P. Basel II Implementation: A Guide to Developing and Validating a Compliant Internal Risk Rating System; McGraw-Hill: New York, NY, USA, 2009.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής (2015), σελ.133-144

ΠΡΟΣΔΙΟΡΙΣΜΟΣ ΦΙΛΤΡΟΥ KALMAN ΓΙΑ ΤΗΝ ΕΚΤΙΜΗΣΗ ΤΩΝ ΘΕΤΙΚΩΝ ΚΑΙ ΑΡΝΗΤΙΚΩΝ ΑΛΜΑΤΩΝ ΤΩΝ ΑΠΟΔΟΣΕΩΝ ΧΡΗΜΑΤΙΣΤΗΡΙΑΚΩΝ ΔΕΙΚΤΩΝ.

Ο. Θεοδοσιάδου, Γ. Τσακλίδης
Τμήμα Μαθηματικών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
rania.theo@hotmail.com, tsaklidi@math.auth.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία γίνεται μια εκτίμηση των θετικών και αρνητικών αλμάτων στην ημερήσια απόδοση του δείκτη Nasdaq, των οποίων η διαφορά ορίζει την ημερήσια απόδοση του δείκτη. Τα θετικά ή αρνητικά άλματα καθορίζονται από τις θετικές ή αρνητικές ειδήσεις και δεν είναι παρατηρήσιμα. Για την εκτίμηση των κρυφών αυτών ημερήσιων αλμάτων γίνεται χρήση κατάλληλου ομογενούς ως προς το χρόνο, γραμμικού φίλτρου Kalman, διακριτού χρόνου. Προκειμένου τα εκτιμώμενα άλματα με το φίλτρο Kalman να είναι μη αρνητικές ποσότητες, χρησιμοποιείται για την εκτίμηση η μέθοδος της αποκοπής της κατανομής των αλμάτων (pdf truncation) με βάση τους περιορισμούς.

Λέξεις κλειδιά: Θετικά-αρνητικά άλματα απόδοσης, Φίλτρο Kalman, αποκοπή κατανομής

1. ΕΙΣΑΓΩΓΗ

Οι Black-Scholes (1973) και Merton (1973) μοντελοποίησαν τις αποδόσεις περιουσιακών στοιχείων (π.χ. μετοχών) θεωρώντας ότι εξελίσσονται κατά συνεχή τρόπο. Οι Carr et al. (2002) διερεύνησαν ωστόσο την ύπαρξη αλμάτων στις αγορές, τα οποία μπορεί να συμβαίνουν είτε με μικρές είτε με μεγάλες συχνότητες. Το μοντέλο που προτείνεται στο Polimenis (2012), θεωρεί ότι οι αποδόσεις εκτίθενται τόσο σε μεταβλητότητα τύπου κίνησης Brown, όσο και σε άλματα, που είναι κρυφές τυχαίες μεταβλητές, δηλαδή δεν αποτελούν παρατηρήσιμες ποσότητες και τα οποία χωρίζονται σε δυο κατηγορίες: Τα θετικά άλματα (άλματα προς τα πάνω), τα οποία οφείλονται στην άφιξη θετικών ειδήσεων στην αγορά και τα αρνητικά άλματα (άλματα προς τα κάτω), τα οποία οφείλονται στην άφιξη αρνητικών ειδήσεων στην αγορά. Στην εργασία Theodosiadou et al. (2013) έγινε μια ανάλυση ευαισθησίας του μοντέλου Polimenis (2012), ενώ στο Theodosiadou et al. (2014) εξετάστηκε η προσαρμογή του σε πραγματικά δεδομένα.

Στην παρούσα εργασία χρησιμοποιούνται οι αποδόσεις του δείκτη Nasdaq για τη χρονική περίοδο 2006-2008 (755 ημερήσιες μετρήσεις), οι οποίες αντλήθηκαν από

την ιστοσελίδα finance.yahoo.com. Επιχειρείται η εκτίμηση των μη παρατηρήσιμων αλμάτων με τη χρήση κατάλληλου γραμμικού φίλτρου Kalman, (Kalman, 1960), το οποίο είναι ομογενές ως προς το χρόνο, δηλαδή ο πίνακας μετάβασης θεωρείται χρονικά αμετάβλητος. Οι εκτιμητές που παρέχει το φίλτρο Kalman είναι οι βέλτιστοι, εφόσον οι θόρυβοι ακολουθούν κανονικές κατανομές. Διαφορετικά, είναι οι βέλτιστοι γραμμικοί εκτιμητές. Οι ημερήσιες αποδόσεις του δείκτη Nasdaq θεωρούνται κατά τη συνήθη πρακτική ως η διαφορά των αντίστοιχων θετικών και αρνητικών αλμάτων (με την προσθήκη θορύβου). Έτσι, προκειμένου τα εκτιμώμενα άλματα να αποτελούν μη αρνητικές ποσότητες, επιλέγεται η αποκοπή των κατανομών των αλμάτων, ώστε αυτά να είναι μη αρνητικά και οι νέες εκτιμήσεις προκύπτουν ως οι αναμενόμενες τιμές των αποκομμένων κατανομών.

2. ΠΡΟΣΔΙΟΡΙΣΜΟΣ ΜΟΝΤΕΛΟΥ

Στην παρούσα ενότητα θα επιχειρήσουμε με τη χρήση του φίλτρου Kalman να εκτιμήσουμε τους κρυφούς παράγοντες X_t και Y_t, t=1,2,.., όπου

X_t: το θετικό άλμα του δείκτη Nasdaq στο χρόνο t,

Y_t: το αρνητικό άλμα του δείκτη Nasdaq στο χρόνο t.

Επιλέγουμε αρχικά τη βασική γενική μορφή του γραμμικού μοντέλου καταστάσεων (state space model) σε διακριτό χρόνο, που εκφράζεται με το σύστημα

$$\mathbf{x}_t = \mathbf{F}_{t-1}\mathbf{x}_{t-1} + \mathbf{w}_{t-1}\ , \tag{1}$$

$$\mathbf{z}_t = \mathbf{H}_t\mathbf{x}_t + \mathbf{e}_t\ , \tag{2}$$

όπου:

- $\mathbf{x_t}$ το m×1 *διάνυσμα κατάστασης*, με συνιστώσες τις m ($m\epsilon N^+$) πλήθους κρυφές καταστάσεις του συστήματος ,
- $\mathbf{z_t}$ το p×1 *διάνυσμα μετρήσεων*, με συνιστώσες τα p ($p\ \epsilon N^+$) παρατηρούμενα μεγέθη ,
- $\{\mathbf{w_t}\}$, $\{\mathbf{e_t}\}$ γκαουσιανοί λευκοί θόρυβοι, ανεξάρτητοι μεταξύ τους με πίνακες συνδιακύμανσης $\mathbf{Q}_t$ και $\mathbf{R}_t$ αντίστοιχα. Δηλαδή, έχουμε

$$\mathbf{w}_t \sim N(\mathbf{0}, \mathbf{Q}_t)\ ,\ \mathbf{e}_t \sim N(\mathbf{0}, \mathbf{R}_t)\ ,$$

$$E[\mathbf{w}_\kappa \mathbf{w}_j^T] = \mathbf{Q}_{k\ \ k-j},\ E[\mathbf{e}_\kappa \mathbf{e}_j^T] = \mathbf{R}_{\kappa\ \ \kappa-j},\ E[\mathbf{e}_\kappa \mathbf{w}_j^T] = \mathbf{0},$$

όπου δ_{k-j} είναι η συνάρτηση δέλτα του Kronecker. Η εξίσωση (1) ονομάζεται *εξίσωση μετάβασης* (*transition equation*) ή *εξίσωση καταστάσεων* (*state equation*) και η (2) ονομάζεται *εξίσωση των μετρήσεων* (*measurement equation*). Στη συνέχεια σε μια πρώτη προσέγγιση της δυναμικής εξέλιξης του δείκτη θα θεωρήσουμε ότι το μοντέλο είναι ομογενές ως προς το χρόνο, δηλαδή ότι οι πίνακες $\mathbf{F}_t$, $\mathbf{H}_t$, $\mathbf{Q}_t$, $\mathbf{R}_t$ στις εξισώσεις (1)–(2) δε μεταβάλλονται με το χρόνο, οπότε μπορούμε να απαλείψουμε τον δείκτη τους t.

Η εκτίμηση του διανύσματος κατάστασης $\mathbf{x}_t$ του μοντέλου (1)-(2) που περιέχει τις κρυφές καταστάσεις, θα γίνει με βάση τη μεθοδολογία του φιλτραρίσματος Kalman (Kalman filtering), η οποία αναπτύχθηκε για πρώτη φορά στην εργασία (Kalman, 1960) και έκτοτε αναπτύχθηκε-χρησιμοποιήθηκε πολύπλευρα τόσο θεωρητικά όσο και στις εφαρμογές. Έτσι, συμβολίζουμε με $\hat{\mathbf{x}}_t^-$ την *εκ των προτέρων (a priori) εκτίμηση* της κατάστασης $\mathbf{x}_t$, με δεδομένες όλες τις μετρήσεις πριν τη χρονική στιγμή t, και με $\hat{\mathrm{x}}_t^+$ την *εκ των υστέρων (a posteriori) εκτίμηση* της $\mathbf{x}_t$, με δεδομένες όλες τις μετρήσεις $\mathbf{z}_t$ μέχρι το χρόνο t. Επιπλέον, συμβολίζουμε με $\mathbf{P}_t^-$, $\mathbf{P}_t^+$ τους πίνακες συνδιακύμανσης των εκ των προτέρων και των εκ των υστέρων σφαλμάτων εκτίμησης της $\mathbf{x}_t$, αντίστοιχα, δηλαδή

$$\mathbf{P}_t^- = \mathrm{E}\left[\left(\mathbf{x}_t - \hat{\mathbf{x}}_t^-\right)\left(\mathbf{x}_t - \hat{\mathbf{x}}_t^-\right)^{\mathrm{T}}\right] \quad \text{και} \quad \mathbf{P}_t^+ = \mathrm{E}\left[\left(\mathbf{x}_t - \hat{\mathbf{x}}_t^+\right)\left(\mathbf{x}_t - \hat{\mathbf{x}}_t^+\right)^{\mathrm{T}}\right].$$

Για να ξεκινήσει η διαδικασία εκτίμησης με το φίλτρο Kalman, υποθέτουμε ότι

$$\hat{\mathbf{x}}_0^+ \sim \mathrm{N}\left(\mathbf{x}_0^+, \mathbf{P}_0^+\right), \text{ όπου } \mathbf{P}_0^+ = \mathrm{E}\left[\left(\mathbf{x}_0 - \hat{\mathbf{x}}_0^+\right)\left(\mathbf{x}_0 - \hat{\mathbf{x}}_0^+\right)^{\mathrm{T}}\right].$$

Στη συνέχεια, για τον υπολογισμό του διανύσματος κατάστασης $\mathbf{x}_t$ για t=1,2,..., χρησιμοποιούνται οι επαναληπτικές σχέσεις του φίλτρου Kalman που δίνονται στον πίνακα που ακολουθεί (Simon, 2006).

Πίνακας 1*.Εξισώσεις του φίλτρου Kalman για διακριτό χρόνο*

Εκ των προτέρων εκτίμηση των $\mathbf{x}_t$	$\hat{\mathbf{x}}_t^- = \mathbf{F}\hat{\mathbf{x}}_{t-1}^+, \quad \hat{\mathbf{x}}_t^- \sim \mathrm{N}\left(\mathbf{x}_t^-, \mathbf{P}_t^-\right)$
Πίνακας συνδιακύμανσης των εκ των προτέρων σφαλμάτων	$\mathbf{P}_t^- = \mathbf{F}\mathbf{P}_{t-1}^+\mathbf{F}^{\mathrm{T}} + \mathbf{Q}$
Πίνακας Kalman (*Kalman gain matrix*)	$\mathbf{K}_t = \mathbf{P}_t^-\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{P}_t^-\mathbf{H}^{\mathrm{T}} + \mathbf{R}_t\right)^{-1}$
Εκ των υστέρων εκτίμηση των $\mathbf{x}_t$	$\hat{\mathbf{x}}_t^+ = \hat{\mathbf{x}}_t^- + \mathbf{K}_t\left(\mathbf{z}_t - \mathbf{H}\hat{\mathbf{x}}_t^-\right), \ \hat{\mathbf{x}}_t^+ \sim \mathrm{N}\left(\mathbf{x}_t^+, \mathbf{P}_t^+\right)$
Πίνακας συνδιακύμανσης των εκ των υστέρων σφαλμάτων	$\mathbf{P}_t^+ = \left(\mathbf{I} - \mathbf{K}\mathbf{H}\right)\mathbf{P}_t^-\left(\mathbf{I} - \mathbf{K}\mathbf{H}\right)^{\mathrm{T}} + \mathbf{K}_t\mathbf{R}\mathbf{K}_t^{\mathrm{T}}$ $= \left(\mathbf{I} - \mathbf{K}_t\mathbf{H}\right)\mathbf{P}_t^-$

Το φίλτρο εφαρμόζεται σε δυο φάσεις: Αρχικά, γίνεται μια πρόβλεψη της επόμενης τιμής της κατάστασης και του αντίστοιχου πίνακα συνδιακύμανσής της, χρησιμοποιώντας όλη την πληροφορία του παρελθόντος, προτού η παρατήρηση στο χρόνο t γίνει γνωστή, και στο δεύτερο βήμα γίνεται ανανέωση (updating) της παραπάνω εκτίμησης, λαμβάνοντας πλέον υπόψη και την παρατήρηση στο χρόνο t. Το σφάλμα της πρόβλεψης για ένα βήμα προς τα εμπρός και η διασπορά του δίνονται από τις σχέσεις

$$\mathbf{u}_t = \mathbf{z}_t - \mathbf{H}\hat{\mathbf{x}}_t^- \ ,$$

$$\mathbf{\Omega}_t = \mathbf{H}\mathbf{P}_t^-\mathbf{H}^T + \mathbf{R} \ ,$$

ενώ, ο πίνακας Kalman είναι το στοιχείο εκείνο που καθορίζει τη συνεισφορά του όρου $\mathbf{u}_t$ στην εκτίμηση της κατάστασης $\mathbf{x}_t$.

Στο γραμμικό, ομογενές ως προς το χρόνο, φίλτρο Kalman που θα χρησιμοποιήσουμε στην παρούσα εργασία για την εκτίμηση των κρυφών καταστάσεων X_t, Y_t, η εξίσωση καταστάσεων δίνεται (με βάση τις (1)-(2)) από τη σχέση

$$\left.\begin{aligned} X_t &= f_{11}X_{t-1} + f_{12}Y_{t-1} + w_{1,t} \\ Y_t &= f_{21}X_{t-1} + f_{22}Y_{t-1} + w_{2,t} \end{aligned}\right\} \ , \ t=1.2\ldots \ ,$$

ή σε μορφή πινάκων

$$\mathbf{x}_t = \mathbf{F}\mathbf{x}_{t-1} + \mathbf{w}_{t-1} \ , \quad t=1.2\ldots \ , \tag{3}$$

όπου,

$$\mathbf{x}_t = \begin{pmatrix} X_t \\ Y_t \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{pmatrix} \ , \quad \mathbf{w}_t = \begin{pmatrix} w_{1,t} \\ w_{2,t} \end{pmatrix} \ .$$

Σημειώνεται ότι για τα θετικά και αρνητικά άλματα, X_t και Y_t αντίστοιχα, των αποδόσεων του δείκτη Nasdaq πρέπει να ισχύει $X_t, Y_t \geq 0$, για $t=1,2\ldots$. Η εξίσωση των μετρήσεων δίνεται από τη σχέση

$$R_t = X_t - Y_t + e_t \ , \ t=1,2\ldots \ ,$$

ή σε μορφή πινάκων

$$R_t = \mathbf{H}\mathbf{x}_t + e_t \ , \ t=1.2\ldots \ , \tag{4}$$

όπου $\mathbf{H}=(1,-1)$ και R_t είναι οι αποδόσεις του δείκτη Nasdaq στο χρόνο t.

3. ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ

Προκειμένου να εκτιμήσουμε τα στοιχεία f_{ij} του πίνακα μετάβασης **F**, καθώς και τις διασπορές των θορύβων, θα χρησιμοποιήσουμε τις αναδρομικές σχέσεις του Πίνακα 1 για την εκτίμηση των κρυφών (μη παρατηρούμενων) αλμάτων X_t, Y_t . Η εκτίμηση θα γίνει με τη μέθοδο της μέγιστης πιθανοφάνειας και για το σκοπό αυτό χρησιμοποιείται το υπολογιστικό περιβάλλον της R. Η λογαριθμική συνάρτηση πιθανοφάνειας έχει τη μορφή (Durbin and Koopman, 2012)

$$\mathrm{LogL}\left(\mathrm{R}_1,...,\mathrm{R}_n\right)=-\frac{n}{2}log\left(2\pi\right)-\frac{1}{2}\sum_{t=1}^{n}\left(log\left(\left|\boldsymbol{\Omega}_t\right|\right)+\mathbf{u}_t^{T}\boldsymbol{\Omega}_t^{-1}\mathbf{u}_t\right),\quad n=755,$$

από την οποία παίρνουμε τις εκτιμήσεις

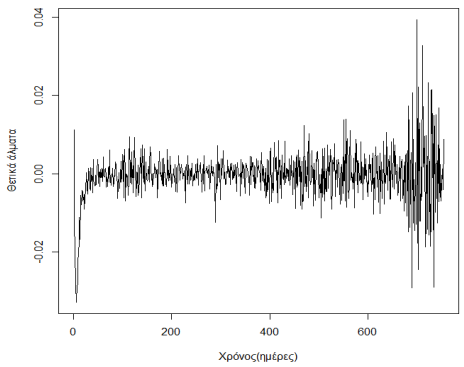
$$\mathbf{F}=\begin{pmatrix}0.3882 & 0.5254\\ 0.5106 & 0.4018\end{pmatrix},\ \mathbf{Q}=\begin{pmatrix}6.1085\times10^{-5} & 0\\ 0 & 6.1085\times10^{-5}\end{pmatrix},\ \mathrm{R}=6.1085\mathrm{x}\ 10^{-5}.$$

Προκειμένου να κάνουμε χρήση των επαναληπτικών σχέσεων του φίλτρου Kalman που δίνονται στον Πίνακα 1 χρησιμοποιώντας τις παραπάνω εκτιμήσεις για τις άγνωστες παραμέτρους του μοντέλου (3)-(4), θεωρούμε ως αρχικό σημείο το

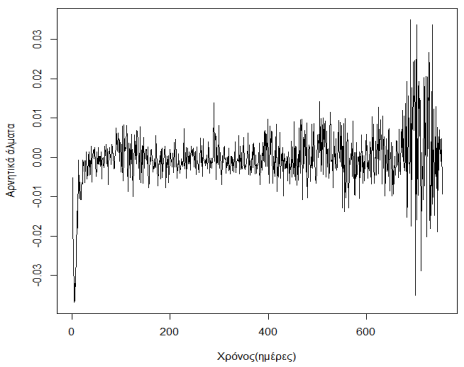
$$\begin{pmatrix}x_0^{+}\\ y_0^{+}\end{pmatrix}=\begin{pmatrix}0.0092\\ 0.0108\end{pmatrix}\quad \text{με}\quad \mathbf{P}_0^{+}=\begin{pmatrix}0.03 & 0\\ 0 & 0.03\end{pmatrix}.$$

Ο πίνακας συνδιασποράς επιλέγεται να περιέχει σχετικά μεγάλες τιμές στη διαγώνιο (η διασπορά των αποδόσεων του δείκτη Nasdaq είναι της τάξης του 10^{-4}). Τα αποτελέσματα στη συνέχεια δεν επηρεάζονται από την επιλογή των τιμών αυτών. Όμοια, δεν επηρεάζει η συγκεκριμένη επιλογή του αρχικού διανύσματος. Συνεπώς, οι εκ των υστέρων εκτιμήσεις των θετικών και των αρνητικών αλμάτων του δείκτη εμφανίζονται στα γραφήματα που ακολουθούν.

Σχήμα 1. *(α) Τα θετικά άλματα του δείκτη Nasdaq με βάση το μοντέλο (3)-(4), (β) Τα αρνητικά άλματα του δείκτη Nasdaq με βάση το μοντέλο (3)-(4).*



(α)



(β)

Παρατηρώντας το Σχήμα 1, διαπιστώνουμε ότι από τη χρησιμοποίηση των εξισώσεων του φίλτρου Kalman προκύπτουν και αρνητικές τιμές των αλμάτων, δηλαδή δεν ικανοποιείται απαραίτητα ο περιορισμός $X_t, Y_t \geq 0$. Σημειώνεται ότι οι τιμές που υπολογίζονται για τα άλματα είναι *αναμενόμενες τιμές* και ότι τα αντίστοιχα 95% διαστήματα εμπιστοσύνης κατά κανόνα περιλαμβάνουν και θετικές τιμές. Εντούτοις, κρίνεται σημαντικό οι προκύπτουσες αναμενόμενες τιμές να είναι μη αρνητικές. Έτσι, στην επόμενη ενότητα θα αναφερθούμε σε έναν τρόπο ενσωμάτωσης του περιορισμού στη μοντελοποίηση του φίλτρου Kalman.

4. ΦΙΛΤΡΑΡΙΣΜΑ ΥΠΟ ΠΕΡΙΟΡΙΣΜΟΥΣ

Προκειμένου να προκύψουν με τη χρήση του φίλτρου Kalman τιμές $X_t, Y_t \geq 0$, θα χρησιμοποιήσουμε τη μέθοδο της αποκοπής για τις συναρτήσεις πυκνότητας πιθανότητας των τ.μ. X_t, Y_t, με βάση αυτούς τους περιορισμούς. Η εκτίμηση του διανύσματος κατάστασης σε αυτήν την περίπτωση προκύπτει ως η αναμενόμενη τιμή της αποκομμένης κατανομής.

Στο μοντέλο (3)-(4), όταν δεν λαμβάνονται υπόψη οι περιορισμοί, η συνάρτηση πυκνότητας πιθανότητας της εκ των υστέρων εκτίμησης του διανύσματος κατάστασης $\mathbf{x}_t$, είναι μια διδιάστατη κανονική κατανομή,

$$\begin{pmatrix} \hat{X}_t^+ \\ \hat{Y}_t^+ \end{pmatrix} \sim N\left(\begin{pmatrix} X_t^+ \\ Y_t^+ \end{pmatrix}, \mathbf{P}_t^+ \right), \quad t=1,2,....$$

Από τη σ.π.π. της πολυδιάστατης κανονικής κατανομής με μέση τιμή $\boldsymbol{\mu} \in \mathbb{R}^d$ και πίνακα συνδιακύμανσης $\boldsymbol{\Sigma}$, που δίνεται από τη σχέση

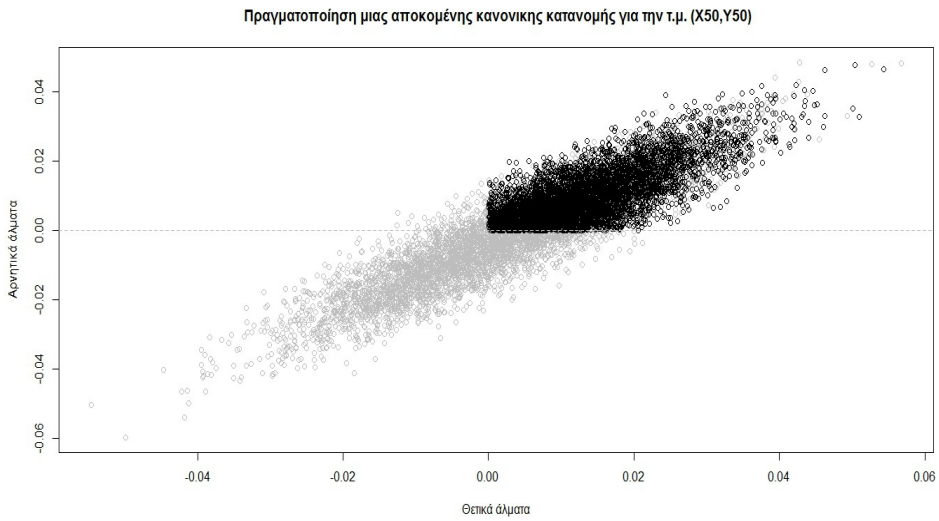
$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} exp\left\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \right\}, \quad \mathbf{x} \in \mathbb{R}^d,$$

προκύπτει η αποκομμένη κανονική κατανομή με άκρα $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^d$, από τη σχέση

$$f(\mathbf{x}\boldsymbol{\mu}\ \boldsymbol{\Sigma}\ \boldsymbol{\alpha}\ \boldsymbol{\beta}\ )=\begin{cases} \dfrac{f\left(\mathbf{x}\boldsymbol{\mu}\ \boldsymbol{\Sigma}\ \right)}{P\left(\boldsymbol{\alpha}\le \mathbf{X}\le \boldsymbol{\beta}\right)}, & \boldsymbol{\alpha}\le \mathbf{x}\le \boldsymbol{\beta} \\ 0\acute{\upsilon} \qquad \alpha\lambda\lambda o\ . \end{cases}$$

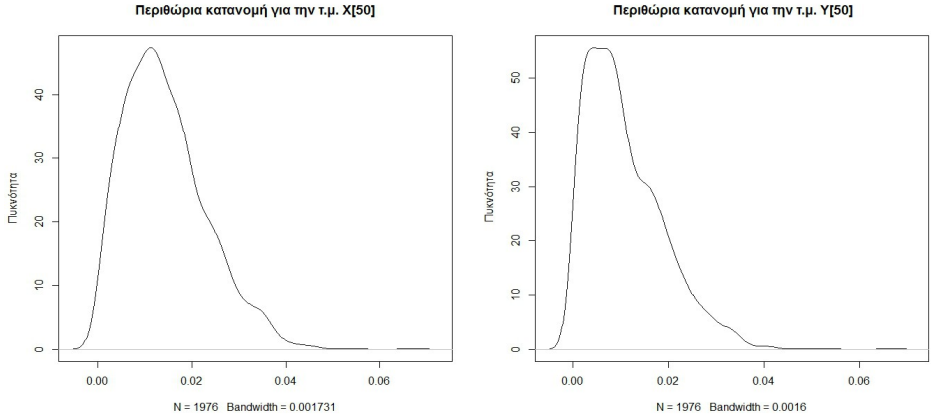
Τα άκρα στα οποία γίνεται η αποκοπή της κανονικής κατανομής στο μοντέλο για τα θετικά και αρνητικά άλματα του δείκτη Nasdaq, είναι $\boldsymbol{\alpha}=(0,0)'$ και $\boldsymbol{\beta}=(\infty, \infty)'$. Στη συνέχεια, παρουσιάζεται με βάση τα ανωτέρω, ενδεικτικά, μια πραγματοποίηση της διδιάστατης κανονικής κατανομής για t=50, δηλαδή της τ.μ. $\left(X_{50}^{+}, Y_{50}^{+}\right)'$, καθώς και της αντίστοιχης αποκομμένης κανονικής κατανομής, τέτοιας ώστε $X_{50}^{+}, Y_{50}^{+} \geq 0$.

Σχήμα 2. *Μια πραγματοποίηση της κανονικής κατανομής της τ.μ.* $\left(X_{50}^{+}, Y_{50}^{+}\right)'$ *και της αντίστοιχης αποκομμένης κανονικής κατανομής.*



Οι περιθώριες κατανομές της διδιάστατης τ.μ. $\left(X_{50}^{+}, Y_{50}^{+}\right)'$ που προκύπτουν από την αποκομμένη κατανομή φαίνονται στα γραφήματα που ακολουθούν.

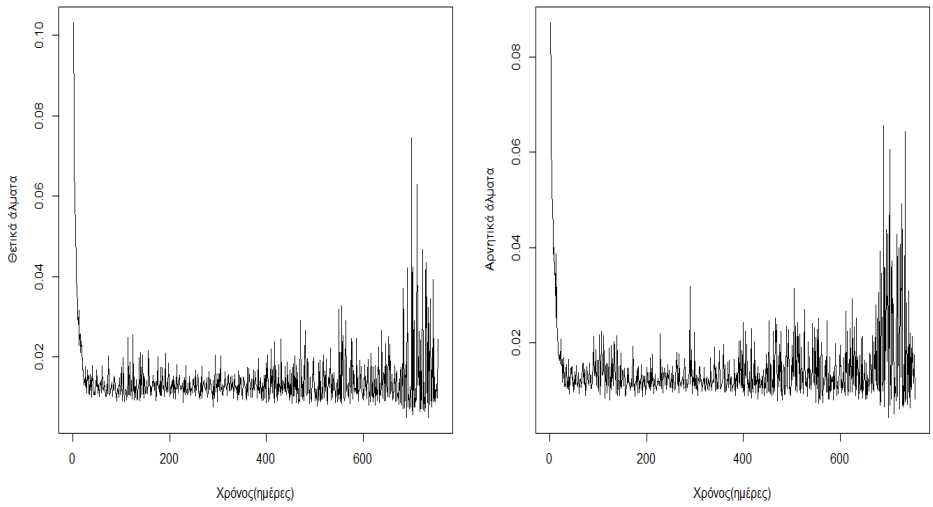
Σχήμα 3. *(α) Περιθώρια κατανομή της τ.μ.* X_{50}^{+} *με βάση την αποκομμένη κανονική κατανομή, (β) Περιθώρια κατανομή της τ.μ.* Y_{50}^{+} *με βάση την αποκομμένη κανονική κατανομή.*



(α) (β)

Ως εκτιμήσεις των (κρυφών) θετικών και αρνητικών αλμάτων των αποδόσεων του δείκτη, που ικανοποιούν τον περιορισμό $X_t, Y_t \geq 0$, μπορούν να ληφθούν πλέον οι αναμενόμενες τιμές της διδιάστατης αποκομμένης κανονικής κατανομής της τ.μ. $\left(X_t^{+}, Y_t^{+}\right)'$, για t=1,2,…,755. Οι εκτιμήσεις των αλμάτων με την περιγραφείσα μεθοδολογία πραγματοποιήθηκαν χρησιμοποιώντας το πακέτο ‘‘tmvtnorm’’ της R και παρουσιάζονται στα γραφήματα που ακολουθούν.

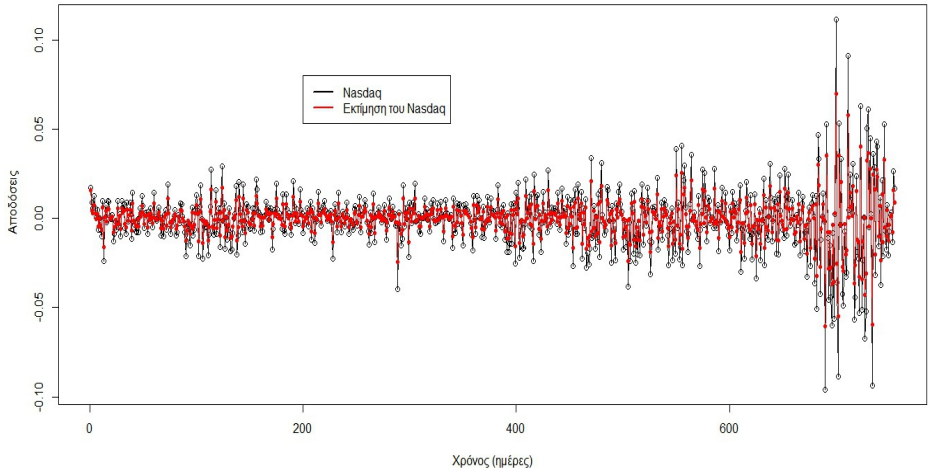
Σχήμα 4. *(α) Θετικά άλματα του δείκτη Nasdaq με βάση την αποκομμένη κανονική κατανομή, (β) Αρνητικά άλματα του δείκτη Nasdaq με βάση την αποκομμένη κανονική κατανομή.*



(α) (β)

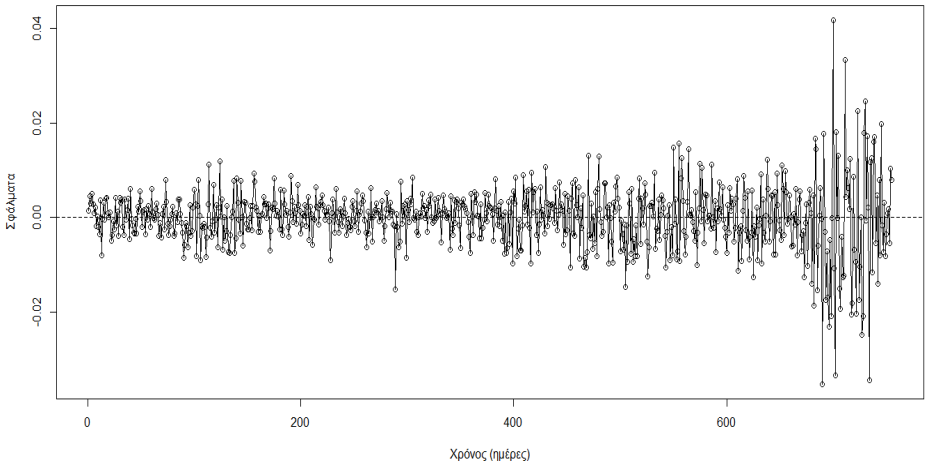
Στο Σχήμα 4 επιβεβαιώνεται ότι οι εκτιμήσεις των θετικών και των αρνητικών αλμάτων για κάθε t ικανοποιούν τον περιορισμό που τέθηκε. Η προσαρμογή στα εμπειρικά δεδομένα του δείκτη, με βάση τη σχέση (4) και τον περιορισμό $X_t, Y_t \geq 0$, φαίνεται στο γράφημα που ακολουθεί.

Σχήμα 5. *Προσαρμογή του μοντέλου στις αποδόσεις του δείκτη Nasdaq για την χρονική περίοδο 2006-2008.*



Σύμφωνα με το Σχήμα 5, οι εκτιμήσεις των αποδόσεων που προκύπτουν με βάση το φιλτράρισμα υπό περιορισμούς ακολουθούν το πρόσημο των εμπειρικών δεδομένων, ενώ η διακύμανση των αποδόσεων υποεκτιμάται. Τα σφάλματα εκτίμησης e_t (βλ. σχέση (4)) των αποδόσεων δίνονται στο παρακάτω γράφημα.

Σχήμα 6. *Σφάλματα εκτίμησης των αποδόσεων του δείκτη Nasdaq για τη χρονική περίοδο 2006-2008.*



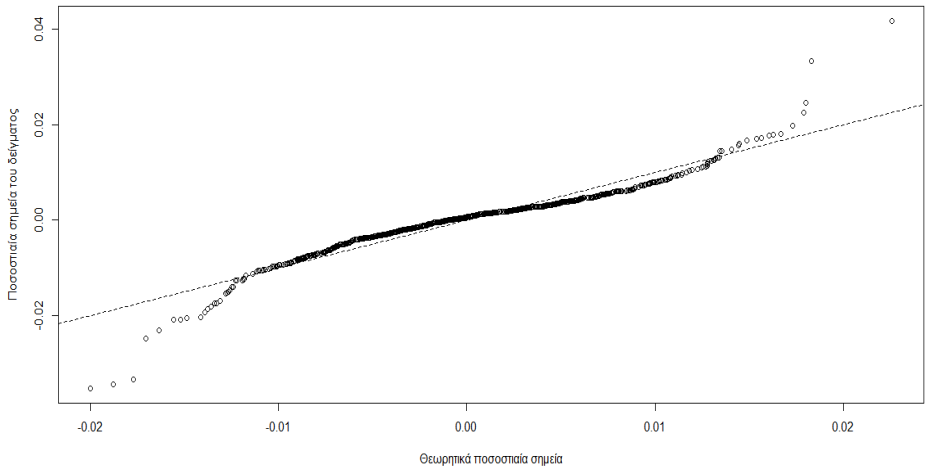
Χρησιμοποιώντας το τεστ Shapiro-Wilk, διαπιστώνουμε ότι τα σφάλματα e_t που προκύπτουν με βάση τις αποκομμένες συναρτήσεις κατανομής των X_t, Y_t, δεν μπορεί

πλέον να θεωρηθεί ότι ακολουθούν κανονική κατανομή (σε 5% επίπεδο σημαντικότητας), όπως φαίνεται ακολούθως:

S.W. στατιστικό = 0.9215, p-τιμή < 2.2e-16

Το γράφημα των ποσοστιαίων σημείων που ακολουθεί επιβεβαιώνει τον παραπάνω ισχυρισμό, όπου παρουσιάζεται αυξημένη πιθανότητα εμφάνισης σφαλμάτων στα άκρα (ουρές) της προκύπτουσας κατανομής των αποδόσεων σε σχέση με την (θεωρητική) κανονική κατανομή.

***Σχήμα** 7. Γράφημα ποσοστιαίων σημείων των σφαλμάτων εκτίμησης και της αντίστοιχης θεωρητικής κανονικής κατανομής*



6. ΣΥΜΠΕΡΑΣΜΑΤΑ-ΠΑΡΑΤΗΡΗΣΕΙΣ

Μια εκτίμηση των κρυφών (μη παρατηρούμενων) θετικών και αρνητικών αλμάτων στην απόδοση του δείκτη Nasdaq μπορεί να επιτευχθεί με χρήση του φίλτρου Kalman, θεωρώντας ότι ισχύουν οι προϋποθέσεις κανονικότητας των σφαλμάτων και γραμμικότητας του πίνακα μετάβασης, όπως εξηγείται στη θεμελίωση του μοντέλου με τις εξισώσεις (3)-(4). Η καταλληλότητα του μοντέλου μπορεί να ελεγχθεί με σύγκριση των αποδόσεων που προκύπτουν από το φίλτρο Kalman και των πραγματικών (παρατηρούμενων) αποδόσεων. Όπως φαίνεται στο Σχήμα 5, οι εκτιμήσεις των αποδόσεων με βάση το φιλτράρισμα υπό περιορισμούς ακολουθούν το πρόσημο των εμπειρικών δεδομένων, ενώ η διακύμανση των αποδόσεων (η απόλυτη τιμή μεγέθους της απόδοσης) υποεκτιμάται. Επόμενο βήμα για πιθανή άρση της υποεκτίμησης των αποδόσεων με όμοια μεθοδολογία φιλτραρίσματος, μπορεί να είναι η χρήση μη γραμμικού φίλτρου Kalman (extended filter Kalman) ή η χρήση φιλτραρίσματος με σωματίδια (particle filtering).

ABSTRACT

The positive and negative jumps of the Nasdaq daily log returns which constitute the daily return are estimated. These jumps are determined by the arrival of the positive and negative news in the market and are not observable. In order to estimate the jumps, the discrete time-homogeneous linear Kalman filter is applied. In order for the estimated jumps to be nonnegative, the method of theirs pdfs truncation, according to the nonnegativity constraints, is used.

ΑΝΑΦΟΡΕΣ

Θεοδοσιάδου Ο., Τσακλίδης Γ., και Πολυμένης Β. (2014). Διερεύνηση των αποδόσεων χρηματιστηριακών δεικτών ως προς τα θετικά και αρνητικά άλματα των αποδόσεων. Η περίπτωση της μετοχής Google, *Πρακτικά 27ου Πανελλήνιου Συνέδριου Στατιστικής.*

Θεοδοσιάδου Ο., Πολυμένης Β., και Τσακλίδης Γ. (2013). Διερεύνηση ενός μοντέλου για τον διαχωρισμό του θορύβου από την άφιξη πληροφορίας στην εξέλιξη της τιμής μιας μετοχής, *Πρακτικά 26ου Πανελλήνιου Συνέδριου Στατιστικής.*

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities*, J. Polit. Econ.* **81**, pp. 637–659

Carr, P., Geman, H., Madan, D.B. and M. Yor (2002). The fine structure of asset returns: An empirical investigation, *J. Bus.* 75, p.305–332.

Durbin, J., and Koopman, S.J. (2012). Time Series Analysis by State Space Methods (Second Edition) , Oxford University Press.

Kalman, R. (1960).A new approach to linear filtering and prediction problems.

Trans. ASME (J. Basic Engineering), 82 D, 35-45.

Polimenis, V. (2012). Information arrival as price jumps. *Optimization: A Journal of*

Mathematical Programming and Operations Research **61:10**, 1179-1190.

Wilhelm, S. and Manjunath, B.G. (2015). tmvtnorm: Truncated Multivariate Normal and Student t distribution, URL *http://CRAN.R-project.org/package= tmvtnorm.*

Merton, R. C. (1973). Theory of rational option pricing, *Bell Journal of Economics and Management Science,* **4 (1)**, 141-183.

Simon, D. (2006). Optimal State Estimation, John Wiley & Sons

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής (2015), σελ. 145−154

Ανάλυση και Μοντελοποίηση Επεισοδίων Βροχόπτωσης

Κ. Ιωαννίδης, Α. Καραγρηγορίου, Δ.Φ. Λέκκας
Analysis and Simulation of Environmental Systems (ASES), Πανεπιστήμιο Αιγαίου
alex.karagrigoriou@aegean.gr

ΠΕΡΙΛΗΨΗ

Ο σκοπός της παρούσας μελέτης είναι η στατιστική ανάλυση επεισοδίων βροχόπτωσης με στόχο τη διερεύνηση ύπαρξης μοτίβων (patterns) στον τρόπο εμφάνισής τους και τα χαρακτηριστικά τους. Πιο συγκεκριμένα, στην εργασία αυτή γίνεται χρήση των καμπυλών έντασης–διάρκειας–συχνότητας (intensity–duration–frequency (IDF) curves), οι οποίες χρησιμοποιούνται ευρέως για τη μοντελοποίηση βροχοπτώσεων. Αξιοποιώντας δεδομένα από το μετεωρολογικό σταθμό της Ερεσού στη Λέσβο, έγινε η εκτίμηση των παραμέτρων του μοντέλου για προκαθορισμένες περιόδους επαναφοράς (return periods). Οι εκτιμήσεις που προκύπτουν υπόκεινται σε ανάλυση ευαισθησίας για να διερευνηθεί αν έχουν τις βέλτιστες τιμές και αν είναι σημαντικές. Τέλος, ένα γενικότερο μοντέλο εφαρμόζεται που επιτρέπει την ταυτόχρονη μοντελοποίηση της διάρκειας, της έντασης και της συχνότητας (μέσω περιόδου επαναφοράς) των επεισοδίων βροχόπτωσης. Η γενίκευση των μοντέλων είναι δυνατόν να συμβάλλει στην ακριβέστερη πρόγνωση των βροχοπτώσεων σε περιπτώσεις ελλειπουσών ή περικομμένων δεδομένων.

Λέξεις Κλειδιά: IDF curves; ανάλυση ευαισθησίας; Monte Carlo προσομοίωση.

1. ΕΙΣΑΓΩΓΗ

Είναι κρίσιμο για την διαχείριση των υδάτων και τον υδρολογικό σχεδιασμό να αναπτυχθούν μοντέλα που μπορούν να περιγράψουν τα χαρακτηριστικά της διεργασίας της βροχόπτωσης, τα οποία συνθέτουν επεισόδια βροχόπτωσης και να προβλέψουν με ακρίβεια τη μελλοντική συμπεριφορά τους. Πιο συγκεκριμένα είναι απαραίτητο να εντοπιστούν οι σχέσεις που συνδέουν την ένταση και τη διάρκεια των επεισοδίων βροχόπτωσης λαμβάνοντας ταυτόχρονα υπ'οψιν και τη συχνότητα εμφάνισής τους. Παραδόξως, ένα μοντέλο που αναπτύχθηκε πριν από σχεδόν έναν αιώνα από τον Bernard (1932) είναι μέχρι και σήμερα ένα από τα πιο ευρέως αποδεκτά και χρησιμοποιούμενα για τη μοντελοποίηση βροχοπτώσεων. Μια πιο γενικευμένη έκφραση για τον ίδιο σκοπό προτάθηκε και διερευνήθηκε σε βάθος πιο πρόσφατα από τους Koutsoyiannis et al. (1998).

Για τη μοντελοποίηση τέτοιων χαρακτηριστικών χρησιμοποιούνται οι καμπύλες

ένταση–διάρκειας–συχνότητας (intensity–duration–frequency (IDF) curves). Μια IDF καμπύλη είναι μια γραφική αναπαράσταση της πιθανότητας να πραγματοποιηθεί ένα επεισόδιο βροχόπτωσης δοσμένης έντασης και διάρκειας. Οι μεταβλητές που αποτελούν τους τρεις άξονες του γραφήματος είναι:

- Διάρκεια επεισοδίου (σε ώρες)
- Ένταση επεισοδίου (σε χιλιοστά ανά ώρα)
- Συχνότητα επεισοδίου (μέσω της περιόδου επαναφοράς) (σε χρόνια)

Μελέτες που έχουν βασιστεί στο συγκεκριμένο είδος μοντελοποίησης έχουν γίνει σε διάφορες περιοχές (βλέπε Stern and Coe, 1984; Lee, 2005; Rao and Kao, 2006). Σημειώνεται επίσης ότι είναι δυνατόν να μελετηθεί πέραν του πιο πάνω και ένα γενικευμένο μοντέλο που συνδέει ταυτόχρονα τη διάρκεια, την ένταση και τη συχνότητα των επεισοδίων βροχόπτωσης.

Στην παρούσα εργασία επιχειρούμε μοντελοποίηση της διεργασίας της βροχόπτωσης αξιοποιώντας τα πιό πάνω μοντέλα και χρησιμοποιώντας τα δεδομένα από το μετεωρολογικό σταθμό της Ερεσού στη Λέσβο για μια περίοδο τριών περίπου ετών. Στην επόμενη ενότητα γίνεται παρουσίαση των βασικών εννοιών που αφορούν στα επεισόδια βροχόπτωσης και ανάλυση των προτεινομένων μοντέλων. Στη συνέχεια η μεθοδολογία εφαρμόζεται στα δεδομένα της Λέσβου οπότε και δίνονται εκτιμήσεις (σημειακοί και σε διάστημα) για τις παραμέτρους των μοντέλων που χρησιμοποιούνται. Τέλος, η αξιοπιστία των εκτιμητών διερευνάται μέσω της ανάλυσης ευαισθησίας για τη διαπίστωση και επιβεβαίωση των βέλτιστων τιμών των εκτιμητών.

2. ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ ΚΑΙ ΤΑ ΜΟΝΤΕΛΑ ΠΡΟΣΑΡΜΟΓΗΣ

2.1 ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ

Οι μετρήσεις της ημερήσιας βροχόπτωσης που συνήθως συλλέγονται από κάποιο μετεωρολογικό σταθμό τυγχάνουν επεξεργασίας με την παραδοχή ότι 2 επεισόδια βροχόπτωσης θεωρούνται ανεξάρτητα αν 2 διαδοχικές μη μηδενικές καταγραφές απέχουν μεταξύ τους πάνω από k ώρες, για κάποια τιμή του k (Κωτούλας, 2001α; 2001β). Τότε για κάθε επεισόδιο βροχόπτωσης που προκύπτει καταγράφονται οι παρακάτω μεταβλητές:

- (R) Συνολική ποσότητα βροχόπτωσης (σε χιλιοστά)
- (D) Διάρκεια επεισοδίου (σε ώρες)
- (I) Ένταση επεισοδίου (σε χιλιοστά ανά ώρα)

οι οποίες ορίζονται ως ακολούθως:

Ορισμός 1. *Η συνολική ποσότητα βροχόπτωσης (R) ορίζεται ως το άθροισμα όλων των καταγραφών βροχόπτωσης κατά τη διάρκεια ενός επεισοδίου και μετριέται σε χιλιοστά.*

Ορισμός 2. *Η διάρκεια επεισοδίου (D) ορίζεται ως η χρονική διαφορά μεταξύ της πρώτης και της τελευταίας μη μηδενικής καταγραφής εντός ενός επεισοδίου βροχόπτωσης και μετριέται σε ώρες.*

Ορισμός 3. *Η ένταση* (I) *ορίζεται ως το πηλίκο της συνολικής ποσότητας βροχόπτωσης προς τη διάρκεια του επεισοδίου* $(I = R/D)$ *και μετριέται σε χιλιοστά ανά ώρα.*

2.2 ΠΕΡΙΓΡΑΦΗ ΜΟΝΤΕΛΟΥ ΚΑΙ ΕΚΤΙΜΗΣΗ

Έστω $F(r)$ η συνάρτηση κατανομής της συνολικής ποσότητας βροχόπτωσης R.

Ορισμός 4. *Η περίοδος επαναφοράς* T_r *ορίζεται για κάθε τιμή* r *της συνολικής βροχόπτωσης ως*

$$T_r = \frac{1}{1 - F(r)} = \frac{1}{S(r)}, \tag{1}$$

όπου $S(r)$ *η συνάρτηση επιβίωσης και μετριέται σε χρόνια.*

Επειδή, εξ'ορισμού, η $S(r)$ που ορίζεται ως $S(r) = 1 - F(r) = \bar{F}(r)$ αντιπροσωπεύει την πιθανότητα η συνολική ποσότητα βροχόπτωσης να υπερβεί τα r χιλιοστά (στη μονάδα του χρόνου) είναι προφανές ότι το αντίστροφο της πιθανότητας αυτής αντιπροσωπεύει *τον αριθμό των μονάδων του χρόνου* T_r που απαιτείται να παρέλθουν μέχρις επαναλήψεως του φαινομένου (της βροχόπτωσης με περισσότερα από r χιλιοστά). Έτσι η περίοδος επαναφοράς T_r αποτελεί έναν εναλλακτικό τρόπο έκφρασης της συχνότητας. Για κάθε περίοδο επαναφοράς ζητείται να προσδιοριστεί μια IDF καμπύλη, η οποία να περιγράφει τη σχέση μεταξύ διάρκειας και έντασης. Όλες οι τυπικές καμπύλες IDF για συγκεκριμένη περίοδο επαναφοράς αποτελούν ειδικές περιπτώσεις της εξίσωσης:

$$I = \frac{c}{(D^{\nu} + \theta)^{\eta}}, \tag{2}$$

όπου οι συντελεστές είναι μη αρνητικοί. Η πιο πάνω εξίσωση δεν έχει προκύψει από κάποια θεωρητική προσέγγιση αλλά αποτελεί εμπειρική φόρμουλα, που προέκυψε μέσα από την εμπειρία της μελέτης πολλών καμπυλών IDF. Απλούστερες μορφές της εξίσωσης εμφανίζονται στην βιβλιογραφία για $\nu = 1$, $\eta = 1$ & $\theta = 0$.

Για τιμές $\nu \neq 1$ και $\eta \neq 1$ το κλάσμα $\frac{1}{(D^{\nu}+\theta)^{\eta}}$ προσεγγίζεται ικανοποιητικά από το $\frac{1}{(D+\theta^*)^{\eta^*}}$. Αριθμητικές μελέτες (Koutsoyiannis et al., 1998) έχουν δείξει το βαθμό αξιοπιστίας της προσέγγισης, έτσι ώστε θέτοντας $a = \theta^*$ και $b = \eta^*$ η πιο πάνω εξίσωση να παίρνει τελικά (προσεγγιστικά) τη μορφή

$$I = \frac{c}{(D + a)^{b}}, \tag{3}$$

όπου a, b και c είναι οι παράμετροι της εξίσωσης.

Είναι συχνό φαινόμενο οι τιμές των παραμέτρων a, b να είναι παρόμοιες για κάθε περίοδο επαναφοράς, όταν τα δεδομένα προέρχονται από τον ίδιο σταθμό ή γειτονικούς. Μάλιστα δεν είναι δύσκολο να δειχθεί (Koutsoyiannis et al., 1998) ότι από τις 3 παραμέτρους της εξίσωσης, ουσιαστικά μόνο η c εξαρτάται από την περίοδο επαναφοράς. Σε τέτοιες περιπτώσεις, είναι δυνατό να προσαρμοσθεί ένα γενικευμένο μοντέλο που συνδέει ταυτόχρονα τη διάρκεια, την ένταση και τη συχνότητα των επεισοδίων βροχόπτωσης:

$$I = \frac{c(T_r)}{(D+a)^b}, \tag{4}$$

Ως συνάρτηση $c(T_r)$ μπορούμε να ορίσουμε με τη βοήθεια της (1) εκείνο το σημείο r της κατανομής F τέτοιο ώστε

$$r = F^{-1}(1 - 1/T_r) \equiv c(T_r). \tag{5}$$

Λόγω του ότι ποσότητες βροχόπτωσης πάνω από μια τιμή r συνδέονται με ακραία φαινόμενα και την κατανομή του maximum, συνήθεις κατανομές που θα μπορούσαν να χρησιμοποιηθούν είναι από το 1^o θεώρημα της θεωρίας ακραίων τιμών (Fisher-Tippett-Gnedenko theorem, Fisher and Tippett, 1928; Gnedenko, 1943) οι κατανομές Gumbel, Frechet και Weibull. Η επιλογή της κατανομής καθορίζει, μέσω της (4), την τελική μορφή της καμπύλης IDF. Ειδικά για την κατανομή Gumbel σημειώνεται ότι στην υδρολογία χρησιμοποιείται για την ανάλυση μηνιαίων ή ετήσιων μέγιστων τιμών βροχόπτωσης αλλά και ξηρασίας. Η ευκολία της κατανομής έγκειται στο γεγονός ότι οι παράμετροι θέσης και κλίμακος της κατανομής α και β εκτιμώνται εύκολα με τη μέθοδο των ροπών. Σχετικώς σημειώνεται ότι με τη βοήθεια της δειγματικής μέσης τιμής $\bar{X}$ και της δειγματικής τυπικής απόκλισης S οι ροποεκτιμήτριες των παραμέτρων της κατανομής δίνονται από τις σχέσεις:

$$\hat{\alpha} = \bar{X} - \frac{\gamma\sqrt{6}s}{\pi} \;\&\; \hat{\beta} = \frac{\sqrt{6}s}{\pi} \tag{6}$$

όπου γ η σταθερά Euler.

Παρά την πιο πάνω πιθανοθεωρητική θεώρηση, στη βιβλιογραφία προτείνονται 2 εναλλακτικές σχέσεις που είναι ιδιαίτερα δημοφιλείς (Berbard, 1932; Raudkivi, 1979; Singh, 1992; Efstratiadis, 2011):

$$I = \frac{c + K\ln(T_r^d)}{(D+a)^b}, \tag{7}$$

και

$$I = \frac{KT_r^d}{(D+a)^b}, \tag{8}$$

Για την εκτίμηση των παραμέτρων μη–γραμμικών μοντέλων, όπως τα παραπάνω, μπορεί να χρησιμοποιηθεί η μη–γραμμική μέθοδος ελαχίστων τετραγώνων, η οποία είναι μια αριθμητική μέθοδος που προσεγγίζει το μη–γραμμικό μοντέλο με ένα γραμμικό αξιοποιώντας τον αλγόριθμό Gauss–Newton (Seber & Wild, 2003).

3. ΑΝΑΛΥΣΗ & ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΔΕΔΟΜΕΝΩΝ ΕΡΕΣΟΥ

Τα δεδομένα που χρησιμοποιήθηκαν στην εργασία αυτή προέρχονται από το μετεωρολογικό σταθμό της Ερεσού στη Λέσβο και καλύπτουν την περίοδο από 17.11.2009 έως 16.12.2012 με κάποια διαστήματα κενά από καταγραφές λόγω τεχνικών αδυναμιών. Τα συλλεχθέντα δεδομένα έτυχαν επεξεργασίας με την παραδοχή ότι $k = 2$ ισοδυναμεί με ανεξαρτησία. Με άλλα λόγια αν δύο μη μηδενικές καταγραφές απέχουν πάνω από $k = 2$ ώρες μεταξύ τους (δηλαδή δεν υπάρχει βροχομετρική καταγραφή για τουλάχιστον 2 ώρες), τότε τα επεισόδια βροχόπτωσης είναι ανεξάρτητα. Συνολικά βρέθηκαν 234 επεισόδια, για καθένα από τα οποία καταγράφηκαν οι 3 μεταβλητές R, D και I.

Κάποια στατιστικά χαρακτηριστικά των δεδομένων δίνονται συνοπτικά στον Πίνακα 1.

Πίνακας 1: *Περιγραφικά στατιστικά επεισοδίων βροχόπτωσης*

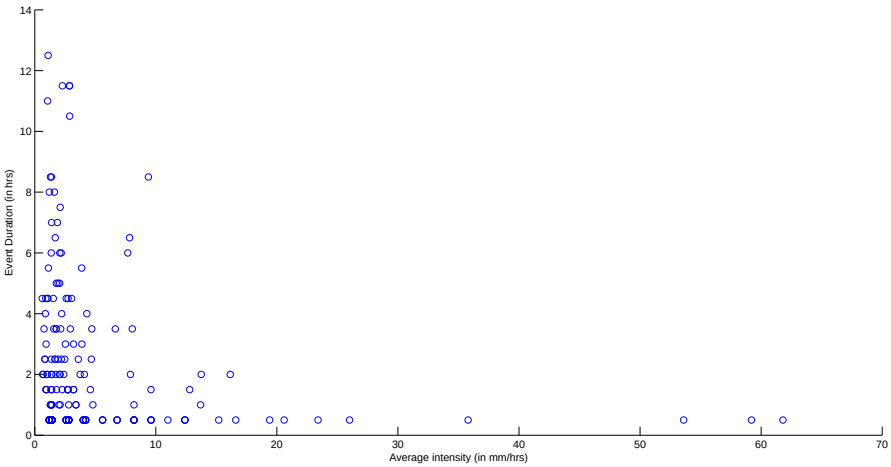
Στατιστικό	Ποσότητα (R)	Διάρκεια (D)	Ένταση (I)
Μέγιστη Τιμή	79.8	12.50	61.8
Ελάχιστη Τιμή	0.60	0.50	0.62
Μέση Τιμή	5.87	1.89	4.21
Τυπική Απόκλιση	9.31	2.41	7.61
1^o Τεταρτημόριο	0.70	0.50	1.40
2^o Τεταρτημόριο	2.10	0.50	1.69
3^o Τεταρτημόριο	6.90	2.38	3.55

ενώ το Σχήμα 1 δίνει μια οπτική απεικόνιση της σχέσης διάρκειας–έντασης μέσω του διαγράμματος διασποράς.

Εφαρμόσαμε το μοντέλο (3) για 4 διαφορετικές περιόδους επαναφοράς και πιο συγκεκριμένα για βροχοπτώσεις που εμφανίζονται μια φορά στα 2, 5, 10 και 20 χρόνια. Οι εκτιμήσεις των παραμέτρων μαζί με τα 95% διαστήματα εμπιστοσύνης δίνονται στον Πίνακα 2.

Πίνακας 2: *Εκτιμήσεις παραμέτρων και* 95% *Δ.Ε. για το μοντέλο (3)*

T_r	2 χρόνια	5 χρόνια	10 χρόνια	20 χρόνια
a	-0.47	-0.29	-0.36	-0.26
95% δ.ε.	(-0.70,-0.24)	(-0.91,0.33)	(-0.82,0.10)	(-0.67,0.16)
b	0.37	0.73	0.62	0.69
95% δ.ε.	(-0.31,1.06)	(-0.06,1.52)	(-0.04,1.29)	(0.27,1.11)
c	4.69	12.98	15.39	21.86
95% δ.ε.	(1.60,7.79)	(0.39,25.57)	(2.89,27.89)	(8.56,35.15)



Σχήμα 1: *Διάγραμμα διασποράς διάρκειας–έντασης*

Οι εκτιμηθείσες εξισώσεις παίρνουν τη μορφή:

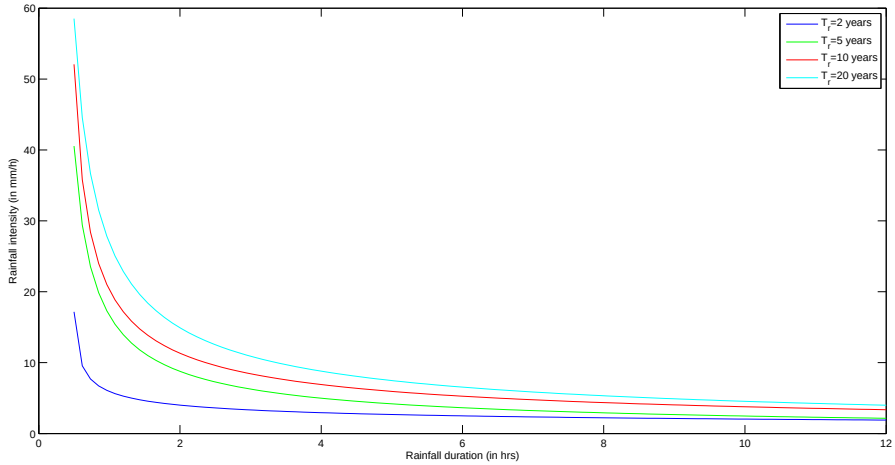
$$I = \frac{4.69}{(D-0.47)^{0.37}}, T_r = 2 \quad years$$
$$I = \frac{12.98}{(D-0.29)^{0.73}}, T_r = 5 \quad years$$
$$I = \frac{15.39}{(D-0.36)^{0.62}}, T_r = 10 \quad years$$
$$I = \frac{21.86}{(D-0.26)^{0.69}}, T_r = 20 \quad years$$

Όπως έχει προαναφερθεί, μια IDF καμπύλη είναι μια γραφική μέθοδος με 3 άξονες, τη διάρκεια, την ένταση και την περίοδο επαναφοράς. Οι παραπάνω εξισώσεις σε κοινό γράφημα απεικονίζονται στο Σχήμα 2.

Εφόσον όπως διαπιστώνουμε οι εκτιμήσεις των παραμέτρων a και b δεν απέχουν σημαντικά, επιλέγουμε, μεταξύ των γενικευμένων μοντέλων που αναφέρθηκαν ανωτέρω, να εφαρμόζουμε το μοντέλο (8). Οι νέες εκτιμήσεις δίνονται στον Πίνακα 3.

Πίνακας 3: *Εκτιμήσεις παραμέτρων και* 95% *διαστήματα εμπιστοσύνης για μοντέλο (8)*

Παράμετρος	Εκτίμηση	95% δ.ε.
a	0.9325	(0.3229,1.542)
b	1.921	(1.449,2.393)
K	5.893	(0.1919,11.59)
d	0.91	(0.8813,0.9387)

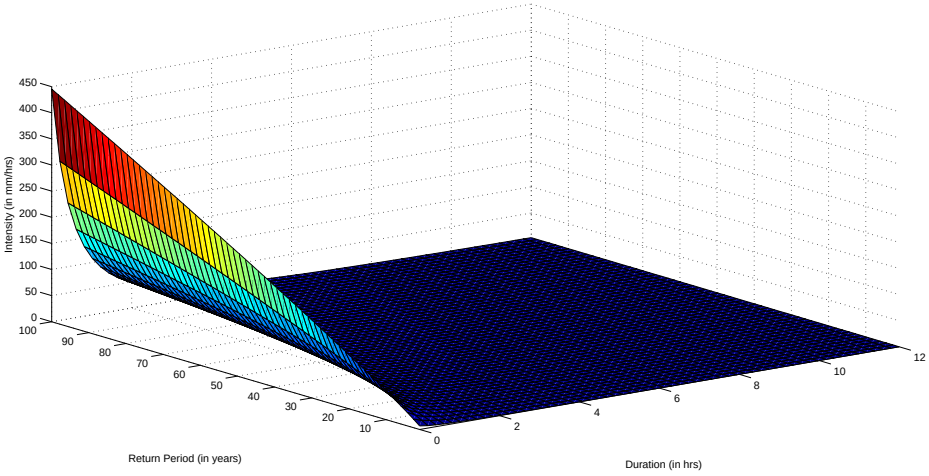


Σχήμα 2: *IDF καμπύλες για T_r 2,5,10 και 20 χρόνια*

Η εξίσωση της συνολικής IDF καμπύλης είναι:

$$I = \frac{5.893 T_r^{0.91}}{(D + 0.9325)^{1.921}}$$

και μπορεί να αποδοθεί γραφικά με το τρισδιάστατο Σχήμα 3.



Σχήμα 3: *IDF καμπύλη με διάρκεια, ένταση και περίοδο επαναφοράς*

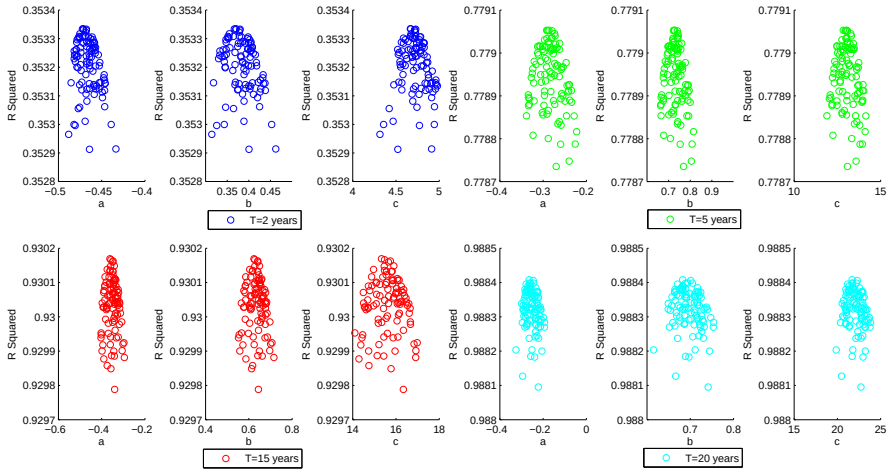
4. ΑΝΑΛΥΣΗ ΕΥΑΙΣΘΗΣΙΑΣ

Η ανάλυση ευαισθησίας είναι τεχνική που μελετά τις συνέπειες που υφίσταται η βέλτιστη (ή η εκτιμώμενη) λύση ενός μοντέλου, ως συνέπεια αλλαγών στις τιμές των δεδομένων ή των παραμέτρων του. Ουσιαστικά σε κάθε φαινόμενο θα πρέπει να λαμβάνεται υπόψη το δυναμικό περιβάλλον των συνεχών αλλαγών μέσα στο οποίο λειτουργεί και υπάρχει το φαινόμενο.

Σε μη γραμμικά μοντέλα σαν αυτό που χρησιμοποιήθηκε εδώ, ενδέχεται οι μέθοδοι εκτίμησης να οδηγήσουν σε μη βέλτιστες τιμές των παραμέτρων. Αυτό μπορεί να συμβεί όταν η αντικειμενική συνάρτηση η οποία πρέπει να μεγιστοποιηθεί, συγκλίνει σε τοπικό και όχι ολικό μέγιστο ή όταν το μοντέλο δεν είναι καλά ορισμένο για κάποιες αρχικές τιμές των παραμέτρων.

Για να διερευνηθεί κατά πόσον οι εκτιμήσεις των παραμέτρων στο μοντέλο είναι όντως βέλτιστες αποφασίστηκε η μεγιστοποίηση του συντελεστή προσδιορισμού R^2 χρησιμοποιώντας τη Monte Carlo μεθοδολογία. Για κάθε ξεχωριστή περίοδο επαναφοράς T_r, δημιουργήθηκαν 10000 τυχαία δείγματα (a_i, b_i, c_i), όπου τα a_i, b_i, c_i ακολουθούν Uniform κατανομή σε διάφορα διαστήματα και για κάθε τριάδα υπολογίσθηκε η ένταση I και επιλέχθηκε εκείνη με το μέγιστο R^2. Η διαδικασία επαναλήφθηκε 100 φορές και τελικά επιλέχθηκαν συνολικά 100 τριάδες παραμέτρων μαζί με την αντίστοιχη τιμή του συντελεστή R^2.

Από το Σχήμα 4 προκύπτει ότι οι κατανομές των παραμέτρων είναι μονοκόρυφες για κάθε περίοδο επαναφοράς με κορυφές τις εκτιμήσεις, όπως προέκυψαν στην προηγούμενη ενότητα, και συνεπώς αυτές καθίστανται βέλτιστες.



Σχήμα 4: *Διαγράμματα διασποράς παραμέτρων και* R^2

5. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΠΕΡΑΙΤΕΡΩ ΕΡΕΥΝΑ

Εφαρμόσαμε ένα μοντέλο που χρησιμοποιείται ευρέως στη βιβλιογραφία για τη μοντελοποίηση επεισοδίων βροχόπτωσης. Παρόμοιες μελέτες έχουν γίνει σε διάφορες περιοχές από τους Stern and Coe (1984), Lee (2005) και Rao and Kao (2006). Οι παράμετροι του μοντέλου εκτιμήθηκαν και γραφικές αναπαραστάσεις των IDF καμπυλών δημιουργήθηκαν. Οι εκτιμήσεις των παραμέτρων υποβλήθηκαν σε ανάλυση ευαισθησίας και επιβεβαιώθηκε ότι αυτές είναι βέλτιστες.

Η παρούσα εργασία αποτελεί το πρώτο μέρος μιας εν εξελίξει έρευνας, η οποία μπορεί να γενικευτεί σε δύο κατευθύνσεις. Γεωγραφικά, καθώς χρησιμοποιήθηκαν δεδομένα από έναν μόνο μετεωρολογικό σταθμό στην Ερεσό της Λέσβου και το μοντέλο μπορεί να επεκταθεί χρησιμοποιώντας δεδομένα από άλλους μετεωρολογικούς σταθμούς της Λέσβου ή και της ευρύτερης περιοχής του Βορειοανατολικού Αιγαίου. Χρονικά, συγκρίνοντας τις προβλέψεις μέσω του μοντέλου για τη συμπεριφορά των επεισοδίων βροχόπτωσης με τα πραγματικά μελλοντικά δεδομένα και επαληθεύοντας την ακρίβειά του.

ABSTRACT

The purpose of this study is the statistical analysis of rainfall events to explore patterns and dependencies that would allow the generalization in cases of missing or truncated data. More specifically, in this paper we estimate intensity–duration–frequency (IDF) curves, which are widely used to model rainfall. We use data from Eresos meteorological station and estimate the parameters of the model for fixed return periods. Sensitivity analysis is conducted to check whether the estimates are optimal. Finally, a more general model is applied that allows for simultaneous modeling of rainfall duration, intensity and frequency (via return periods).

ΑΝΑΦΟΡΕΣ

Bernard, M. (1932). Formulas for rainfall intensities of long duration, *Transactions of the American Society of Civil Engineers*, **96(1)**, 592–606.

Efstratiadis, A., *Lecture notes on flood hydrology and design of sewage networks*, 44 pages, https://www.itia.ntua.gr/en/docinfo/1154/.

Fisher, R.A., Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest and smallest member of a sample, *Proc. Cambridge Phil. Soc.*, **24(2)**, 180-190.

Gnedenko, B.V. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire, *Annals of Mathematics*, **44**, 423-453.

Koutsoyannis D., Kozonis D., Manetas A. (1998). A mathematical framework for studying rainfall intensity-duration-frequency relationships, *Journal of Hydrology*, **206(1)**, 118–135.

Κωτούλας, Δ. (2001α). *Υδρολογία και Υδραυλική Φυσικού Περιβάλλοντος*, Θεσσαλονίκη, Τμήμα Εκδόσεων ΑΠΘ.

Κωτούλας, Δ. (2001β). *Ορεινή Υδρονομική Τόμος Ι, Τα Ρέοντα Ύδατα*, Θεσσαλονίκη, Τμήμα Εκδόσεων ΑΠΘ.

Lee, C. (2005). Application of Rainfall Frequency Analysis on studying rainfall distribution characteristics of Chia-Nan plain area in southern Taiwan, *Crop, Environment and Bioinformatics*, **2**, 31–38.

Rao, A.R & Kao, S.C. (2006). Statistical Analysis of Indiana rainfall data, *Joint Transportation Research Program Report*, **C-36-62R**, Purdue University, West Lafayette, IN.

Seber, G. A. F.; Wild, C. J. (2003). *N*onlinear Regression. New York: John Wiley and Sons.

Stern, R.D. and Coe, R. (1984). A model fitting analysis of daily rainfall data, *Journal of the Royal Statistical Society. Series A (General)*, **147**, 1–34.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής (2015), σελ.155-164

ΜΕΘΟΔΟΣ ΥΠΟΛΟΓΙΣΜΟΥ ΤΟΥ ΔΕΙΚΤΗ GINI ΠΟΥ ΒΑΣΙΖΕΤΑΙ ΣΤΗΝ ΑΝΑΠΑΡΑΣΤΑΣΗ ΤΟΥ ΩΣ ΓΙΝΟΜΕΝΟ ΠΙΝΑΚΩΝ ΓΙΑ ΔΕΔΟΜΕΝΑ ΚΑΤΗΓΟΡΙΟΠΟΙΗΜΕΝΑ ΣΕ ΚΛΑΣΕΙΣ

Ε. Κετζάκη
Τμήμα Μαθηματικών Α.Π.Θ.
eketzaki@yahoo.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία θα αποδειχθεί μια μέθοδος υπολογισμού του δείκτη Gini, ως γινόμενο πινάκων, στην περίπτωση που τα δεδομένα είναι κατηγοριοποιημένα σε κλάσεις. Η προτεινόμενη μέθοδος μειώνει το πλήθος των απαιτούμενων πράξεων για τον υπολογισμό των συνιστωσών που συνθέτουν τον δείκτη Gini, στην περίπτωση που τα δεδομένα είναι κατηγοριοποιημένα. Με αυτό τον τρόπο, εξασφαλίζεται, ο υπολογισμός του δείκτη πραγματοποιώντας λιγότερες πράξεις, σε σύγκριση με τις υπάρχουσες μεθόδους, χωρίς το αποτέλεσμα να στερείται υπολογιστικής ακρίβειας. Στην πρώτη ενότητα της εργασίας γίνεται αναφορά στην έννοια του δείκτη Gini, στην δεύτερη ενότητα διατυπώνεται και αποδεικνύεται η προτεινόμενη μέθοδος και στην τρίτη ενότητα παρουσιάζεται ένα αριθμητικό παράδειγμα, ώστε να γίνουν περισσότερο κατανοητές οι υπάρχουσες αλλά και οι προτεινόμενες μέθοδοι.

Λέξεις κλειδιά: δείκτης Gini, δεδομένα κατηγοριοποιημένα σε κλάσεις.

1. ΕΙΣΑΓΩΓΗ

Ο δείκτης Gini είναι ένας από τους πιο διαδεδομένους δείκτες, που χρησιμοποιούνται για την μέτρηση της ανομοιότητας εισοδηματικών κυρίως δεδομένων, χωρίς όμως να περιορίζεται η επέκταση της εφαρμογής του και σε άλλου είδους οικονομικά ή κοινωνικά δεδομένα (Xu, 2003).

Η σχέση του δείκτη Gini με την καμπύλη Lorenz, αποτελεί χαρακτηριστικό πλεονέκτημα του συγκεκριμένου δείκτη, συγκριτικά με άλλους δείκτες που μετρούν την κοινωνική ή οικονομική ανομοιότητα (Lorenz, 1905, Gastwirth, 1971). Η καμπύλη Lorenz είναι η γραφική αναπαράσταση της κατανομής εισοδηματικών δεδομένων n ατόμων, τα σημεία της οποίας προκύπτουν από την σχέση (1),

$$L_i = \left[i/n,\ y_i \Big/ \left(n\bar{y} \right) \right] \tag{1}$$

όπου y_i εκφράζει το εισόδημα του i ατόμου και $\bar{y}$ την μέση τιμή των εισοδημάτων των n ατόμων. Στην περίπτωση που τα εισοδήματα είναι ομοιόμορφα κατανεμημένα στα n άτομα, δηλαδή σε ίσο πλήθος ατόμων να αντιστοιχεί ίδιο ποσοστό εισοδήματος, τότε η καμπύλη Lorenz θα ονομάζεται ευθεία τέλειας ισότητας. Η απόκλιση της καμπύλης Lorenz από την ευθεία τέλειας ισότητας εκφράζει το μέγεθος που διαφοροποιούνται τα εισοδήματα, συγκεκριμένου πλήθους ατόμων, από την «ιδανική» περίπτωση στην οποία θα έχουν όλοι το ίδιο εισόδημα. Με αυτόν τον τρόπο ελέγχεται εάν υπάρχει η όχι ανομοιότητα στην κατανομή των εισοδημάτων.

Η τιμή του δείκτη Gini προκύπτει άμεσα από την καμπύλη Lorenz, αφού ισούται με το πηλίκο του εμβαδού της περιοχής Α, που είναι η περιοχή η οποία βρίσκεται ανάμεσα στην ευθεία τέλειας ισότητας και στην καμπύλη Lorenz, προς το εμβαδόν Α+Β που είναι η περιοχή που βρίσκεται κάτω από την ευθεία τέλεια ισότητας. (Γράφημα 1) Επομένως ο δείκτης Gini θα ισούται με

$$I_G = A/(A+B) \tag{2}$$

Κατά την διάρκεια του τελευταίου αιώνα πολλοί επιστήμονες (Xu, 2003, Santos and Guerro, 2010), πρότειναν διαφορετικές αλγεβρικές μορφές του δείκτη Gini που είχαν σαν στόχο να μετασχηματίσουν και να βελτιώσουν την μορφή του, μερικοί από αυτούς είναι οι Sen (1973), Donaldson and Weymark (1980), Berrebi and Silber (1987), Silber (1989). Η αλγεβρική αναπαράσταση που προτάθηκε από τον Sen (1973) δίνεται από την σχέση (3)

$$I_G = \left[1/\left(2n^2\bar{y}\right)\right]\cdot\sum_{i=1}^{n}\sum_{j=1}^{n}\left|y_i - y_j\right| \tag{3}$$

Η αλγεβρική μορφή του δείκτη αποδείχθηκε (Silber, 1989), ότι μπορεί να μετασχηματιστεί σε γινόμενο πινάκων ώστε να καθίσταται ευκολότερος ο υπολογισμός του δείκτη, καθώς δεν απαιτούνται ιδιαίτερες προγραμματιστικές ικανότητες από τον ερευνητή. Ο δείκτης ανομοιότητας Gini, για δεδομένα πλήθους n, υπολογίζεται ως γινόμενο διανυσμάτων και πινάκων, από την σχέση (4)

$$I_G = e' \cdot G \cdot s \tag{4}$$

Όπου e' παριστάνει το διάνυσμα γραμμή *n* στοιχείων που έχουν όλα αριθμητική τιμή ίση με *1/n*. G είναι ο *n* x *n* τετραγωνικός πίνακας, τα στοιχεία του οποίου προκύπτουν ως εξής: όταν $i < j$ τότε η τιμή των στοιχείων $g_{ij} = -1$, όταν $i > j$ τότε η τιμή των στοιχείων $g_{ij} = 1$ και όταν $i = j$ η τιμή των στοιχείων $g_{ij} = 0$. Το διάνυσμα στήλη *s* αποτελείται από τα *n* στοιχεία s_i, τοποθετημένα σε φθίνουσα σειρά. Το s_i ισούται με τον λόγο του εισοδήματος που αντιστοιχεί στο i-άτομο, προς το συνολικό άθροισμα των εισοδημάτων των n ατόμων.

Στην παρούσα εργασία αρχικά θα περιγραφεί η υφιστάμενη μέθοδος υπολογισμού του δείκτη Gini για δεδομένα τα οποία είναι κατηγοριοποιημένα σε κλάσεις. Στην συνέχεια θα αποδειχθεί μια νέα μέθοδος υπολογισμού του δείκτη για κατηγοριοποιημένα δεδομένα, εμπνευσμένη από την μεθοδολογία που περιγράφηκε στην εργασία Κετζάκη και Φαρμάκης (2014). Η προτεινόμενη μέθοδος αποσκοπεί στον υπολογισμό των συνιστωσών που συνθέτουν τον δείκτη Gini για κατηγοριοποιημένα δεδομένα χρησιμοποιώντας λιγότερες πράξεις.

2. ΜΕΘΟΔΟΣ ΥΠΟΛΟΓΙΣΜΟΥ ΤΟΥ ΔΕΙΚΤΗ GINI ΓΙΑ ΔΕΔΟΜΕΝΑ ΚΑΤΗΓΟΡΙΟΠΟΙΗΜΕΝΑ ΣΕ ΚΛΑΣΕΙΣ

Αν υποθέσουμε ότι $y_1, y_2, \ldots y_n$ είναι τα εισοδηματικά δεδομένα των n ατόμων, κατηγοριοποιημένα σε m κλάσεις και n_h είναι το πλήθος των ατόμων που ανήκουν στην h-κλάση, ώστε να ισχύει $n = \sum_{h=1}^{m} n_h$.

Διαμερίζοντας κάθε έναν από τους παράγοντες του γινομένου της σχέσης (4) e΄, s και G σε m υποδιανύσματα και υποπίνακες αντίστοιχα, η τιμή του δείκτη Gini γράφεται (Silber, 1989)

$$I_G = \begin{bmatrix} e'(n_1) & e'(n_2) & \cdots & e'(n_m) \end{bmatrix} \begin{bmatrix} G(n_1,n_1) & G(n_1,n_2) & \cdots & G(n_1,n_m) \\ G(n_2,n_1) & G(n_2,n_2) & \cdots & G(n_2,n_m) \\ \vdots & \vdots & \ddots & \vdots \\ G(n_m,n_1) & G(n_m,n_2) & \cdots & G(n_m,n_m) \end{bmatrix} \begin{bmatrix} s(n_1) \\ s(n_2) \\ \vdots \\ s(n_m) \end{bmatrix} \quad (5)$$

Όπου $e'(n_h) = \begin{bmatrix} 1/n & 1/n & \cdots & 1/n \end{bmatrix}$ το υποδιάνυσμα γραμμή του e΄, μεγέθους n_h, κάθε στοιχείο του οποίου ισούται με *1/n*. $G(n_k, n_\lambda)$ είναι ο υποπίνακας διαστάσεων n_κ x n_λ με κ, λ = *1, 2, ..., m* του πίνακα G. Σε περίπτωση που το $\kappa = \lambda$ τότε ο πίνακας $G(n_k, n_\lambda)$ θα είναι ο τετραγωνικός πίνακας οι τιμές των στοιχείων του οποίου προκύπτουν ως εξής: τα στοιχεία που βρίσκονται πάνω από την κύρια διαγώνιο είναι ίσα με -1, τα στοιχεία τα οποία βρίσκονται κάτω από την κύρια διαγώνιο είναι ίσα με +1 με 0 τα στοιχεία της κύριας διαγώνιου. Εάν το $\kappa < \lambda$ τότε ο πίνακας $G(n_k, n_\lambda)$ θα είναι ο πίνακας διαστάσεων n_κ x n_λ με όλα του τα στοιχεία να έχουν τιμή ίση με -1 και εάν το $\kappa > \lambda$ τότε ο πίνακας $G(n_k, n_\lambda)$ θα είναι ό πίνακας διαστάσεων n_κ *x* n_λ του οποίου όλα τα στοιχεία έχουν τιμή ίση με 1. Το διάνυσμα στήλη $s(n_h)$ είναι το υποδιάνυσμα του διανύσματος s μεγέθους n_h. Το *s (i,h)* είναι το i στοιχείο του $s(n_h)$ και περιγράφει τον λόγο του εισοδήματος του i- ατόμου που ανήκει στην κλάση h προς το συνολικό άθροισμα των εισοδημάτων. Ο συμβολισμός $s_{\bullet h}$ εκφράζει το άθροισμα όλων των στοιχείων *s(i,h)* της h ομάδας, δηλαδή: $s_{\bullet h} = \sum_{i=1}^{n_h} s(i,h)$.

Αποδείχθηκε από τον Silber (1989), ότι από την σχέση (5) μπορεί να προκύψει σχέση που μετρά την ανομοιότητα εντός των m κλάσεων (within classes)

$$I_W = \sum_{p=1}^{m} e'(n_p) \cdot G(n_p, n_p) \cdot s(n_p) \quad (6)$$

αλλά και την ανομοιότητα ανάμεσα στις m κλάσεις (between classes)

$$I_B = \sum_{p=1}^{m} \left[\sum_{q \neq p}^{m} e'(n_p) \cdot G(n_p, n_q) \cdot s(n_q) \right] \quad (7)$$

Στην ενότητα αυτή θα αποδείξουμε ότι κάθε ένας από τους δύο δείκτες I_W και I_B μπορεί να γραφεί σαν ένα διαφορετικό γινόμενο πινάκων. Για την ακρίβεια σαν ένα γινόμενο δυο πινάκων οι οποίοι έχουν μειωμένη διάσταση κατά το ήμισυ συγκριτικά

με τους αρχικούς πίνακες, για να αποδειχθεί αυτό είναι χρήσιμη η πρόταση που ακολουθεί

Πρόταση

Αν υποθέσουμε ότι $y_1,y_2,...y_n$ είναι τα εισοδηματικά δεδομένα των n ατόμων, κατηγοριοποιημένα σε m κλάσεις και n_h είναι το πλήθος των ατόμων που ανήκουν στην h-κλάση, τότε ο δείκτης Gini για τα n_h στοιχεία της h-κλάσης, με $n_h \le n$, θα είναι ίσος με το γινόμενο των πινάκων

$$I_{G_h} = e_h' \cdot s_{d_h} \tag{8}$$

όπου e_h' είναι το διάνυσμα γραμμή μεγέθους $[n_h/2]$, κάθε στοιχείο του οποίου συμβολίζεται με $e_h(i)$ και προκύπτει από την σχέση

$$e_h(i) = (n_h - 2i + 1)/n, \quad \forall i = 1,2,\cdots,[n_h/2] \tag{9}$$

και s_{d_h} είναι το διάνυσμα στήλη μεγέθους $[n_h/2]$, κάθε στοιχείο του οποίου συμβολίζεται με $s_{d_h}(i)$ και προκύπτει από την σχέση

$$s_{d_h}(i) = s(i,h) - s(n_h + 1 - i, h) \tag{10}$$

Όπου το $[n_h/2]$ παριστάνει το ακέραιο μέρος της ποσότητας $n_h/2$.

Απόδειξη:

Σε περίπτωση που το πλήθος των στοιχείων της κλάσης n_h είναι άρτιο, οπότε το $n_h = 2\rho,\ \rho \in \mathbb{Z}_+^*$ το γινόμενο των πινάκων $e_h' \cdot s_{d_h}$, θα ισούται με

$$e_h' \cdot s_{d_h} = \left[(2\rho-1)/n \quad (2\rho-3)/n \quad \cdots \quad 1/n\right] \cdot \begin{bmatrix} s(1,h) - s(2\rho,h) \\ s(2,h) - s(2\rho-1,h) \\ \vdots \\ s(\rho,h) - s(\rho+1,h) \end{bmatrix} =$$

$$= \sum_{t=1}^{\rho} \left[\left[2\rho - (2t-1)\right]/n\right] \cdot \left[s(t,h) - s(2\rho+1-t,h)\right] =$$

$$= \sum_{t=1}^{\rho} \left[\left[2\rho - (2t-1)\right]/n\right] \cdot s(t,h) + \sum_{t=\rho+1}^{2\rho} \left[\left[2\rho - (2t-1)\right]/n\right] \cdot s(t,h) =$$

$$= \sum_{t=1}^{2\rho} \left[\left[2\rho - (2t-1)\right]/n\right] \cdot s(t,h) \tag{11}$$

Στην h-κλάση της σχέσης (6) θεωρώντας $n_h = 2\rho$ και $s_i = s(i,h)$, προκύπτει η ακόλουθη σχέση για κάθε $i = 1,2,...,n_h = 2\rho$,

$$I_{G_h} = e'(2\rho) \cdot G(2\rho, 2\rho) \cdot s(2\rho) = \sum_{t=1}^{2\rho} \left[\left[2\rho - (2t-1)\right]/n\right] \cdot s(t,h) \tag{12}$$

Διαπιστώθηκε ότι τα αποτελέσματα των σχέσεων (11) και (12) ταυτίζονται στην περίπτωση που το πλήθος των στοιχείων της κλάσης είναι άρτιος αριθμός. Με όμοιο

τρόπο αποδεικνύεται ότι ισχύει η πρόταση και σε περίπτωση που το πλήθος n_h είναι περιττός αριθμός δηλαδή $n_h = 2\rho + 1,\ \rho \in \mathbb{Z}_+^*$ θα ισχύει

$$e_h' \cdot s_{d_h} = \begin{bmatrix} 2\rho/n & (2\rho-2)/n & \cdots & 2/n \end{bmatrix} \cdot \begin{bmatrix} s(1,h) - s(2\rho,h) \\ s(2,h) - s(2\rho-1,h) \\ \vdots \\ s(\rho,h) - s(\rho+1,h) \end{bmatrix} =$$

$$= \sum_{t=1}^{\rho} \left[(2\rho - 2t + 2)/n \right] \cdot \left[s(t,h) - s(2\rho+1-t,h) \right] =$$

$$= \sum_{t=1}^{\rho} \left[(2\rho - 2t + 2)/n \right] \cdot s(t,h) + \sum_{t=\rho+1}^{2\rho+1} \left[(2\rho - 2t + 2)/n \right] \cdot s(t,h) =$$

$$= \sum_{t=1}^{2\rho+1} \left[(2\rho - 2t + 2)/n \right] \cdot s(t,h)$$

Εάν στην h κλάση της σχέσης (6) αντικατασταθεί το $n_h = 2\rho + 1$ και $s_i = s(i,h)$ για κάθε $i = 1, 2, ..., n_h = 2\rho + 1$ θα προκύψει

$$I_{G_h} = e'(2\rho+1) \cdot G(2\rho+1, 2\rho+1) \cdot s(2\rho+1) = \sum_{t=1}^{2\rho+1} \left[(2\rho - 2t + 2)/n \right] \cdot s(t,h) \qquad (14)$$

Επομένως αποδείχθηκε ότι: $I_{G_h} = e'(n_h) \cdot G(n_h, n_h) \cdot s(n_h) = e_h' \cdot s_{d_h}$

□

Χρησιμοποιώντας την προηγούμενη πρόταση θα αποδειχθούν τα ακόλουθα δυο θεωρήματα τα οποία υπολογίζουν την τιμή του δείκτη Gini που μετρά την ανομοιότητα εντός των κλάσεων αλλά και μεταξύ των κλάσεων.

Θεώρημα 1

Η ανομοιότητα εντός των m κλάσεων, στις οποίες είναι χωρισμένα n πλήθους δεδομένα, μπορεί να υπολογιστεί από την σχέση.

$$I_W = \sum_{h=1}^{m} e_h' \cdot s_{d_h} \qquad (15)$$

Όπου e_h' είναι το διάνυσμα γραμμή, κάθε ένα από τα στοιχεία του οποίου υπολογίζονται από την σχέση $e_h(i) = (n_h - 2i + 1)/n$ και s_{d_h} το διάνυσμα στήλη κάθε ένα από τα στοιχεία του οποίου υπολογίζονται από την σχέση $s_{d_h}(i) = s(i,h) - s(n_h + 1 - i, h),\ \forall i = 1, 2, \cdots, \lfloor n_h / 2 \rfloor$ με $\lfloor n_h / 2 \rfloor$ συμβολίζεται το ακέραιο μέρος της ποσότητας $n_h / 2$

Απόδειξη:

Η μέτρηση της ανομοιότητας εντός των κλάσεων σύμφωνα με την σχέση (6) είναι ίση με $I_w = \sum_{h=1}^{m} e'(n_h) \cdot G(n_h, n_h) \cdot s(n_h)$. Λαμβάνοντας υπόψη την πρόταση και αντικαθιστώντας την ποσότητας $e'(n_h) \cdot G(n_h, n_h) \cdot s(n_h)$ με $e_h' \cdot s_{d_h}$ η τιμή του δείκτη γράφεται

$$I_W = \sum_{h=1}^{m} e'(n_h) \cdot G(n_h, n_h) \cdot s(n_h) = \sum_{h=1}^{m} e_h' \cdot s_{d_h}$$

επομένως αποδεικνύεται το ζητούμενο

□

Θεώρημα 2

Η ανομοιότητα μεταξύ των κλάσεων, για n ομαδοποιημένα δεδομένα χωρισμένα σε m κλάσεις υπολογίζεται από την σχέση.

$$I_B = \sum_{p=1}^{m} \sum_{q>p}^{m} e'_{pq} \cdot s_{d_{pq}} \tag{16}$$

Όπου e'_{pq} και $s_{d_{pq}}$ είναι τα διάνυσμα γραμμή και στήλη αντίστοιχα τα στοιχεία των οποίων δίνονται από τις σχέση

$$e_{pq}(i) = \left(n_p + n_q - 2i + 1\right)/n \quad \text{και} \tag{17}$$

$$s_{d_{pq}}(i) = s(i, p+q) - s\left(n_p + n_q + 1 - i, p+q\right) \tag{18}$$

$\forall i = 1,2,\cdots,\left[\left(n_p + n_q\right)/2\right]$ όπου $\left[\left(n_p + n_q\right)/2\right]$συμβολίζεται το ακέραιο μέρος του $\left(n_p + n_q\right)/2$ και

$$s(i, p+q) = \begin{cases} \overline{s_p} & \varepsilon\alpha\nu \quad 1 \le i \le n_p \\ \overline{s_q} & \varepsilon\alpha\nu \quad n_p < i \le n_p + n_q \end{cases} \tag{19}$$

με $\overline{s_p}$ και $\overline{s_q}$ ο λόγος της μέσης τιμής των εισοδημάτων που αφορούν την κλάση p και q αντίστοιχα, προς το συνολικό εισόδημα.

Απόδειξη:
Η ανομοιότητα μεταξύ των κλάσεων που περιγράφεται στην σχέση (7), μπορεί να γραφεί ως συνάρτηση της ποσότητας I_{pq}. (Silber, 1989) . Με το I_{pq} να προκύπτει από

$$I_{pq} = \left[1/\left(n_p + n_q\right) \quad \cdots \quad 1/\left(n_p + n_q\right)\right] \cdot G\left(n_{p+q}, n_{p+q}\right) \cdot \begin{bmatrix} \overline{s_p}/\left(n_p \cdot \overline{s_p} + n_q \cdot \overline{s_q}\right) \\ \vdots \\ \overline{s_p}/\left(n_p \cdot \overline{s_p} + n_q \cdot \overline{s_q}\right) \\ \overline{s_q}/\left(n_p \cdot \overline{s_p} + n_q \cdot \overline{s_q}\right) \\ \vdots \\ \overline{s_q}/\left(n_p \cdot \overline{s_p} + n_q \cdot \overline{s_q}\right) \end{bmatrix} \begin{matrix} \} n_p \text{ όροι} \\ \\ \\ \} n_q \text{ όροι} \end{matrix} \tag{20}$$

Δηλαδή,

$$I_B = \sum_{p=1}^{m} \left[\sum_{q \ne p}^{m} e\left(n_p\right) \cdot G\left(n_p, n_q\right) \cdot s\left(n_q\right)\right] = \sum_{p=1}^{m} \sum_{q>p}^{m} \left[\left(\left(n_p + n_q\right)/n\right) \cdot I_{pq} \cdot \left(n_p \cdot \overline{s_p} + n_q \cdot \overline{s_q}\right)\right] \tag{21}$$

Εάν γίνει αντικατάσταση του γινομένου $\left(\left(n_p + n_q\right)/n\right) \cdot I_{pq} \cdot \left(n_p \cdot \overline{s_p} + n_q \cdot \overline{s_q}\right)$ με $I_{G_{pq}}$, όπου

$$I_{G_{pq}} = \underbrace{\left[1/n \quad \cdots \quad 1/n\right]}_{p+q \text{ όροι}} \cdot G\left(n_{p+q}, n_{p+q}\right) \cdot \begin{bmatrix} \overline{s_p} \\ \vdots \\ \overline{s_p} \\ \overline{s_q} \\ \vdots \\ \overline{s_q} \end{bmatrix} \begin{matrix} \} n_p \text{ όροι} \\ \\ \\ \} n_q \text{ όροι} \end{matrix} \tag{22}$$

Τότε $\sum_{p=1}^{m} \sum_{q>p}^{m} \left[\left(\left(n_p + n_q \right) / n \right) \cdot I_{pq} \cdot \left(n_p \cdot \overline{s_p} + n_q \cdot \overline{s_q} \right) \right] = \sum_{p=1}^{m} \sum_{q>p}^{m} I_{G_{pq}}$

όμως η σχέση (22) μπορεί να γραφεί λόγω της σχέσης (8) της πρότασης,

$$I_{G_{pq}} = e'_{pq} \cdot s_{d_{pq}} \tag{23}$$

εάν θεωρήσουμε ότι το πλήθος των στοιχείων της h κλάσης n_h θα είναι ίσο με το άθροισμα του πλήθους των στοιχείων που βρίσκονται στην κλάση p και q, δηλαδή $n_h = n_p + n_q$ και τα στοιχεία $s_{d_h}(i) = s_{d_{pq}}(i) = s(i, p+q) - s\left(n_p + n_q + 1 - i, p+q\right)$

με $s(i, p+q) = \begin{cases} \overline{s_p} & \varepsilon\alpha\nu \quad 1 \le i \le n_p \\ \overline{s_q} & \varepsilon\alpha\nu \quad n_p < i \le n_p + n_q \end{cases}$ τότε ο δείκτης,

$$I_B = \sum_{p=1}^{m} \left[\sum_{q>p}^{m} e'\left(n_p\right) \cdot G\left(n_p, n_q\right) \cdot s\left(n_q\right) \right] = \sum_{p=1}^{m} \sum_{q>p}^{m} e'_{pq} \cdot s_{d_{pq}}$$

οπότε αποδεικνύεται το ζητούμενο □

3. ΠΑΡΑΔΕΙΓΜΑ

Στην ενότητα αυτή θα πραγματοποιηθεί μια εφαρμογή των θεωρητικών μεθόδων, που παρουσιάστηκαν στις προηγούμενες ενότητες της παρούσας εργασίας. Μέσω ενός απλού αριθμητικού παραδείγματος θα περιγραφεί ο τρόπος υπολογισμού του δείκτη Gini, αρχικά για μεμονωμένα δεδομένα και έπειτα για δεδομένα τα οποία ανήκουν σε κλάσεις. Ο υπολογισμός του δείκτη Gini, για μεμονωμένα δεδομένα, θα γίνει χρησιμοποιώντας την γεωμετρική, την αλγεβρική έκφραση του δείκτη και την μορφή του δείκτη που βασίζεται στον υπολογισμό του ως γινόμενο πινάκων.

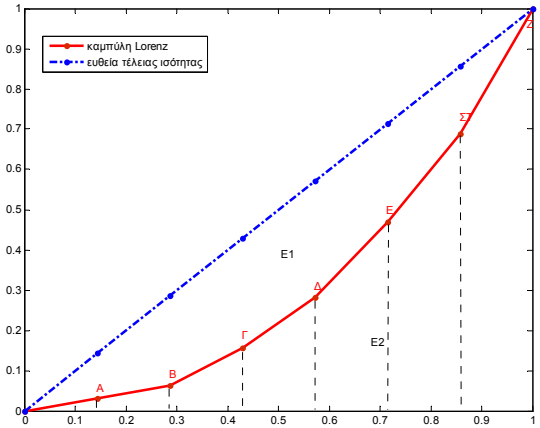
Τα αριθμητικά δεδομένα που χρησιμοποιούνται στην συγκεκριμένη εφαρμογή, είναι εμπνευσμένα από το αριθμητικό παράδειγμα το οποίο παρατίθεται στο παράρτημα της εργασίας του Silber (1989).

Θεωρούμε ότι οι τιμές 1, 1, 3, 4, 6, 7, 10 περιγράφουν τα κέρδη 7 ατόμων. Η καμπύλη Lorenz για τα δεδομένα που δίνονται είναι η τεθλασμένη γραμμή που προκύπτει από την ένωση των σημείων που την ορίζουν σύμφωνα με τον τύπο (1) και είναι τα ακόλουθα: Ο(0, 0), Α(1/7, 1/32), Β(2/7, 2/32), Γ(3/7, 5/32), Δ(4/7, 9/32), Ε(5/7, 15/32), ΣΤ(6/7, 22/32), Ζ (7/7, 32/32). Η αριθμητική τιμή του δείκτη Gini, για τα δεδομένα του παραδείγματος, υπολογίζεται από το Γράφημα 1. διαιρώντας το

εμβαδόν του χωρίου E_1, που περικλείεται από την ευθεία τέλειας ισότητας και την καμπύλη Lorenz, με το συνολικό εμβαδόν του χωρίου E_1+E_2, που βρίσκεται κάτω από την ευθεία τέλειας ισότητας. Δηλαδή

$$I_G = E_1 / (E_1 + E_2) = E_1 / (1/2) = 2E_1 = 1 - 2E_2 .$$

Γράφημα 1. Καμπύλη Lorenz και ευθεία τέλειας ισότητας



Αθροίζοντας τα εμβαδά αυτά προκύπτει ότι E_2= 140/448. Επομένως η τιμή του δείκτη Gini θα ισούται με I_G=1-2(140/448)=84/224=3/8.

Χρησιμοποιώντας στην συνέχεια την αλγεβρική έκφραση της σχέση (2), προκύπτει η ίδια αριθμητική τιμή για τον δείκτη. Πιο συγκεκριμένα αντικαθιστώντας στον αλγεβρικό τύπο την μέση τιμή του δείγματος, η οποία είναι ίση με μ=32/7 και το άθροισμα των απόλυτων διαφορών που προκύπτουν εάν αφαιρεθεί κάθε τιμή του δείγματος από τις υπόλοιπες συμπεριλαμβανομένου και της ίδιας, προκύπτει

$$I_G = \left[1/\left(2n^2\mu\right)\right] \cdot \sum_{i=1}^{n}\sum_{j=1}^{n}\left|y_i - y_j\right| = \left[1/\left(2 \cdot 7^2 \cdot (32/7)\right)\right] \cdot 168 = 84/224 = 3/8$$

Στην συνέχεια υπολογίζεται αρχικά ο δείκτης Gini όπως προτάθηκε από Κετζάκη και Φαρμάκης (2014) με την χρήση του τύπου που τον περιγράφει ως γινόμενο πινάκων και έπειτα όπως προτάθηκε από τον Silber (1989),

και $I_G = e' \cdot s_d = [6/7 \quad 4/7 \quad 2/7][9/32 \quad 8/32 \quad 3/32]' = 84/224 = 3/8$

$$I_G = [1/7 \quad 1/7 \quad 1/7 \quad 1/7 \quad 1/7 \quad 1/7 \quad 1/7] \begin{bmatrix} 0 & -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 0 & -1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 0 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 0 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 0 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 0 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 10/32 \\ 7/32 \\ 6/32 \\ 4/32 \\ 3/32 \\ 1/32 \\ 1/32 \end{bmatrix} = 3/8$$

Ας θεωρήσουμε ότι τα 7 άτομα τα οποία μελετήθηκαν στον προηγούμενο παράδειγμα ανήκουν σε 3 διαφορετικές ομάδες των 2,3 και 2 ατόμων αντίστοιχα. Τα

κέρδη που αντιστοιχούν σε κάθε ομάδα είναι: 1 και 3 για την πρώτη ομάδα, 1,4 και 7 για την δεύτερη ομάδα και 6 και 10 για την τρίτη ομάδα.
Αρχικά θα υπολογιστεί η τιμή του δείκτη Gini εντός των κλάσεων χρησιμοποιώντας τον τύπο της σχέσης (6),

$$I_W = \sum_{h=1}^{3} e^{'}(n_h)\cdot G(n_h,n_h)\cdot s(n_h) =$$
$$= e^{'}(n_1)\cdot G(n_1,n_1)\cdot s(n_1) + e^{'}(n_2)\cdot G(n_2,n_2)\cdot s(n_2) + e^{'}(n_3)\cdot G(n_3,n_3)\cdot s(n_3) =$$
$$= [1/7 \quad 1/7]\cdot G(2,2)\cdot [3/32 \quad 1/32] + [1/7 \quad 1/7 \quad 1/7]\cdot G(3,3)\cdot [7/32 \quad 4/32 \quad 1/32] +$$
$$+ [1/7 \quad 1/7]\cdot G(2,2)\cdot [10/32 \quad 6/32] = 18/224$$

Το ίδιο αποτέλεσμα θα προκύψει για την τιμή του δείκτη Gini εντός των κλάσεων εφαρμόζοντας την σχέση (15)

$$I_W = \sum_{h=1}^{m} e'_h \cdot s_{d_h} = e'_1 \cdot s_{d_1} + e'_2 \cdot s_{d_2} + e'_3 \cdot s_{d_3} = (1/7)\cdot(2/32) + (2/7)\cdot(6/32) + (1/7)(4/32) =$$
$$= 18/224$$

Στην συνέχεια θα υπολογιστεί η τιμή του δείκτη ανομοιότητα μεταξύ των κλάσεων αρχικά σύμφωνα με την σχέση (16) και στην συνέχεια με την σχέση (21)

$$\mathrm{I_B} = \sum_{p=1}^{m}\sum_{q>p}^{m} e'_{pq}\cdot s_{d_{pq}} = e'_{12}\cdot s_{d_{12}} + e'_{13}\cdot s_{d_{13}} =$$
$$= [4/7 \quad 2/7]\cdot\begin{bmatrix}2/32\\2/32\end{bmatrix} + [3/7 \quad 1/7]\cdot\begin{bmatrix}12/32\\12/32\end{bmatrix} + [4/7 \quad 2/7]\cdot\begin{bmatrix}4/32\\4/32\end{bmatrix} = 60/224$$

και στην συνέχεια με την σχέση (21)

$$I_B = \sum_{p=1}^{m}\sum_{q>p}^{m}\left[\left((n_p+n_q)/n\right)\cdot I_{pq}\cdot\left(n_p\cdot\overline{s_p}+n_q\cdot\overline{s_q}\right)\right] =$$
$$= \left[(n_1+n_2)/n\right]\cdot I_{12}\cdot\left(n_1\cdot\overline{s_1}+n_2\cdot\overline{s_2}\right) + \left[(n_1+n_3)/n\right]\cdot I_{13}\cdot\left(n_1\cdot\overline{s_1}+n_3\cdot\overline{s_3}\right) + \left[(n_2+n_3)/n\right]\cdot I_{23}\cdot\left(n_2\cdot\overline{s_2}+n_3\cdot\overline{s_3}\right) =$$
$$= (5/7)\cdot(3/20)\cdot(16/32) + (4/7)\cdot(3/10)\cdot(20/32) + (5/7)\cdot(6/35)\cdot(28/32) = 60/224$$

αφού και τα I_{12}, I_{13}, I_{23} προκύπτουν από τις ακόλουθες σχέσεις:

$$I_{12} = \left[\underbrace{1/5 \quad \cdots \quad 1/5}_{5\ \text{όροι}}\right] G(5,5)\begin{bmatrix}(4/32)/(16/32)\\(4/32)/(16/32)\\(4/32)/(16/32)\\(2/32)/(16/32)\\(2/32)/(16/32)\end{bmatrix} = 3/20, \quad I_{13} = \left[\underbrace{1/4 \quad \cdots \quad 1/4}_{4\ \text{όροι}}\right] G(4,4)\begin{bmatrix}(16/32)/(40/32)\\(16/32)/(40/32)\\(4/32)/(40/32)\\(4/32)/(40/32)\end{bmatrix} = 3/11$$

$$I_{23} = \left[\underbrace{1/5 \quad \cdots \quad 1/5}_{5\ \text{όροι}}\right] G(5,5)\begin{bmatrix}(8/32)/(28/32)\\(8/32)/(28/32)\\(4/32)/(28/32)\\(4/32)/(28/32)\\(4/32)/(28/32)\end{bmatrix} = 6/35$$

4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στη συγκεκριμένη εργασία παρουσιάστηκε μια έκφραση που υπολογίζει τον δείκτη Gini για κατηγοριοποιημένα δεδομένα. Πιο συγκεκριμένα αποδείχθηκε ότι οι συνιστώσες που περιγράφουν τον δείκτη Gini για δεδομένα που ανήκουν σε κατηγορίες και μετρούν την ανομοιότητα εντός των κλάσεων και μεταξύ των κλάσεων, μπορούν να γραφούν κάθε μια χωριστά σε απλούστερη μορφή. Η προτεινόμενη σχέση υπολογίζει την ανομοιότητα πραγματοποιώντας λιγότερες πράξεις και κατά συνέπεια με μεγαλύτερη υπολογιστική ταχύτητα διατηρώντας παράλληλα την υπολογιστική ακρίβεια της τιμής του δείκτη αλλά χρησιμοποιώντας λιγότερα τεχνικά χαρακτηριστικά του Η/Υ.

ABSTRACT

In this paper an expression is proposed, that calculates precisely Gini index for classified data. This expression has the form of the product of two matrices. The proposed expression improves current expressions as it calculates the Gini index within and between classes of classified data using matrices with reduced dimensions. An expression that computes the Gini index of inequality for classified data as a matrix multiplication allows an easy and quick computation, since many computer programs have subroutines for matrix multiplication. Moreover reducing matrices' dimension is very useful especially for big data because the speed of the calculation becomes higher and the required memory for the calculation of Gini index is reduced.

ΑΝΑΦΟΡΕΣ

Berrebi Z. M. and Silber J. (1987). Regional differences and the components of growth and inequality change. *Economic Letters,* **25**, 295-298.

Donaldson D. and Weymark J. A. (1980). A single-parameter generalization of the Gini indices of inequality. *Journal of Economic Theory,* **22**, 67-86.

Gastwirth J. L. (1971). A general definition of the Lorenz curve. *Econometrica,* **39**, 1037-1039.

Lorenz M. O. (1905). Methods of measuring the concentration of wealth*, Publications of the American Statistical Association,* **9**, 209-219.

Santos B. J. and Guerro J. B. J. (2010). Gini's Concentration Ratio. *Electronic Journal for History of Probability and Statistics,* **6**, 1-42.

Sen, A. (1973). *On economic inequality*, New York: Oxford University Press.

Silber J. (1989). Factor components, population subgroups and the computation of the Gini index of inequality. *The Review of Economics and Statistics,* **71**, 107-115.

Xu K. (2003). How has the literature on Gini's index evolved in the past 80 Years? *Department of Economics, Dalhousie University.*

Κετζάκη Ε. και Φαρμάκης Ν. (2014). Μέθοδος υπολογισμού του δείκτη Gini που βασίζεται στην αναπαράσταση του ως γινόμενο πινάκων. *Πρακτικά 27ου Πανελλήνιου Συνέδριού Στατιστικής,* 109 -116.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής (2015), σελ.165-177

ΣΥΝΘΕΤΕΣ ΣΥΝΑΡΤΗΣΕΙΣ ΣΑΡΩΣΗΣ ΚΑΙ ΕΦΑΡΜΟΓΕΣ ΣΤΑ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΑ

Β. Μ. Κούτρας, Μ. Β. Κούτρας
Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς
vkoutras@icloud.com, mkoutras@unipi.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία εισάγεται μια σύνθετη κατανομή η οποία αφορά το άθροισμα τυχαίων μεταβλητών των οποίων το πλήθος είναι τυχαίο και καθορίζεται με βάση την τιμή μιας συνάρτησης σάρωσης.

Αρχικά δείχνεται πως η προτεινόμενη κατανομή μπορεί να εφαρμοσθεί αποτελεσματικά στο χώρο των χρηματοοικονομικών. Πιο συγκεκριμένα διαμορφώνεται και μελετάται ένα μοντέλο που μπορεί να χρησιμοποιηθεί από μια εποπτεύουσα αρχή (regulator) για την ανίχνευση προβλήματος στην κεφαλαιακή επάρκεια ενός εποπτευόμενου χρηματοπιστωτικού οργανισμού, για την ανίχνευση κινδύνων (πιστωτικών ή λειτουργικών) με τυχαιοποιημένα σχήματα ελέγχου των υποκείμενων μονάδων κ.ά.

Για την κατανομή που εμπλέκεται στο προαναφερθέν μοντέλο, δίνονται τύποι υπολογισμού των παραμέτρων της (μέση τιμή και διακύμανση), της πιθανογεννήτριας και ροπογεννήτριας καθώς επίσης και τρόποι υπολογισμού της συνάρτησης πιθανότητας με χρήση αναδρομικών σχημάτων που χρησιμοποιούν πληροφορίες από κατάλληλους πίνακες μετάβασης αλυσίδων Markov ή τους συντελεστές της ρητής έκφρασης που βρίσκεται για την πιθανογεννήτρια της κατανομής.

Λέξεις Κλειδιά: Σύνθετες κατανομές, συναρτήσεις σάρωσης, επιτήρηση χρηματοπιστωτικών οργανισμών.

1. ΕΙΣΑΓΩΓΗ

Πολλά προβλήματα που απαντώνται στις περιοχές

- της χρηματοικονομικής διοίκησης κινδύνου,
- του αναλογισμού,
- του ελέγχου ποιότητας και την αξιοπιστία,
- στις κοινωνικές επιστήμες και την ψυχολογία

μπορούν να περιγραφούν με στοχαστικά μοντέλα που περιλαμβάνουν δίτιμες τυχαίες μεταβλητές οι οποίες παίρνουν τις τιμές 1 (επιτυχία - success, *S*) ή 0 (αποτυχία - failure, *F*).

Σε τέτοια μοντέλα το ενδιαφέρον εστιάζεται συνήθως σε τυχαίες μεταβλητές που σχετίζονται με το χρόνο αναμονής μέχρι να ικανοποιηθεί ένα συγκεκριμένο κριτήριο (stopping rule).

Στην παρούσα εργασία διαμορφώνεται και μελετάται ένα μοντέλο που μπορεί να χρησιμοποιηθεί από μια εποπτεύουσα αρχή (regulator) για την ανίχνευση προβλήματος στην κεφαλαιακή επάρκεια ενός εποπτευόμενου χρηματοπιστωτικού οργανισμού, για την ανίχνευση κινδύνων (πιστωτικών ή λειτουργικών) με τυχαιοποιημένα σχήματα ελέγχου των υποκείμενων μονάδων κ.ά.

Θα δώσουμε λοιπόν αρχικά την περιγραφή του προβλήματος το οποίο αποτέλεσε το έναυσμα για τη μελέτη του στοχαστικού μοντέλου που θα αναλυθεί στα πλαίσια της παρούσης εργασίας.

Ας υποθέσουμε ότι μία τράπεζα υπόκειται σε μια σειρά stress tests σε διάφορες χρονικές στιγμές. Με βάση κάποιους δείκτες οι οποίοι αντικατοπτρίζουν την χρηματοπιστωτική σταθερότητα της τράπεζας, αυτή μπορεί να θεωρηθεί είτε ότι λειτουργεί ικανοποιητικά (και επομένως έχει μικρή πιθανότητα χρεοκοπίας) είτε όχι (άρα υπάρχει αυξημένος κίνδυνος χρεοκοπίας). Για παράδειγμα, είναι γενικώς αποδεκτό ότι, η κεφαλαιοποίηση μιας τράπεζας αποτελεί εάν σημαντικό δείκτη οικονομικής υγείας αφού μέσω αυτής μπορούν να αντιμετωπισθούν πιθανές μη αναμενόμενες ζημιές. Επομένως μια τράπεζα με επαρκή κεφαλαιοποίηση έχει μικρή πιθανότητα χρεοκοπίας.

Ένας ικανοποιητικός δείκτης καλής κεφαλαιοποίησης μιας τράπεζας αποτελεί ο «λόγος κεφαλαιοποίησης» *CAR* (Capital Ratio), ο οποίος ορίζεται ως ο λόγος μεταξύ των κεφαλαίων τύπου *Tier*1+*Tier*2 προς τα λεγόμενα *Risk-Weighted Assets* (*RWA*), δηλαδή

$$CAR = \frac{Tier1 + Tier2}{RWA}.$$

Για εποπτικούς σκοπούς, το σύμφωνο της Βασιλείας (Basel Accord) έχει υιοθετήσει μια απλή διχοτομική ταξινόμηση των εποπτευόμενων ιδρυμάτων, σύμφωνα με την οποία μια τράπεζα χαρακτηρίζεται ως υποκεφαλαιοποιημένη ή μη (undercapitalized ή well capitalized) ανάλογα με το αν ο δείκτης *CAR* είναι μικρότερος ή μεγαλύτερος από 8%, βλέπε για παράδειγμα Berger *et al.* (2008), Estrella *et al.* (2000), Goldberg and Hudgins (2002), Lindquist (2004).

Προφανώς τα αποτελέσματα μιας σειράς stress tests θα δημιουργούν μια ακολουθία δίτιμων μεταβλητών (indicator variables) με τιμές 1 (success, *S*) ή 0 (failure, *F*) όπου το 1 υποδηλώνει ένα αρνητικό stress test (η τράπεζα βρέθηκε υποκεφαλαιοποιημένη κατά τη διάρκεια του *i*-οστού stress test, δηλαδή διαπιστώθηκε ότι $CAR<8\%$) και το 0 υποδηλώνει ένα θετικό stress test (κατά τη διάρκεια του *i*-οστού stress test, διαπιστώθηκε ότι $CAR\geq 8\%$). Σημειώνεται ότι με το συμβολισμό

που υιοθετήθηκε η τιμή 1 που αντιστοιχεί στην επιτυχία υποδηλώνει αρνητικό αποτέλεσμα ενώ η τιμή 0 που αντιστοιχεί σε αποτυχία, δηλώνει θετικό αποτέλεσμα.

Δύο «λογικά» κριτήρια για να θεωρηθεί ότι για την τράπεζα υπάρχει κίνδυνος για επικείμενη χρεοκοπία είναι τα εξής:

a. Η τράπεζα αποτυγχάνει σε πολλά διαδοχικά stress tests δηλαδή παρατηρείται ένα αποτέλεσμα της μορφής *SS*....*S* (εμφάνιση μιας μεγάλου μήκους ροής επιτυχιών)

b. Παρατηρούνται δύο αρνητικά stress test τα οποία βρίσκονται πολύ κοντά μεταξύ τους, δηλαδή έχουμε αποτελέσματα της μορφής *SS*, *SFS*, *SFFS* κτλ.

Προφανώς το δεύτερο κριτήριο είναι πιο ρεαλιστικό από το πρώτο αφού μέσω αυτού, ακόμη και αν η τράπεζα αντέξει σε κάποια από τα stress tests αμέσως μετά την διαπίστωση κακής κεφαλαιοποίησης, η σύντομη επαναφορά της στην κατάσταση υποκεφαλαιοποίησης δίνει σήμα κινδύνου χρεοκοπίας.

Στα πλαίσια της παρούσας εργασίας θα παρουσιάσουμε ένα μοντέλο στο οποίο οι χρονικές στιγμές κατά τις οποίες γίνεται το stress test είναι τυχαίες μεταβλητές και το κριτήριο για να χαρακτηρισθεί η τράπεζα επικίνδυνη για χρεοκοπία είναι η αδυναμία της να περάσει επιτυχώς stress tests που απέχουν μεταξύ τους απόσταση λιγότερη ή ίση από k. Ένα ενδιαφέρον πρόβλημα σε αυτή την περίπτωση είναι να μελετηθεί η κατανομή και τα χαρακτηριστικά (μέση τιμή, ροπές κτλ) του χρόνου κατά τον οποίο θα ανιχνευθεί (μέσω του κριτηρίου που χρησιμοποιούμε) ο επικείμενος κίνδυνος χρεοκοπίας του ελεγχόμενου οργανισμού.

2. ΕΝΑ ΜΟΝΤΕΛΟ ΑΝΙΧΝΕΥΣΗΣ ΚΙΝΔΥΝΟΥ ΜΕ ΤΥΧΑΙΟΥΣ ΧΡΟΝΟΥΣ ΕΛΕΓΧΟΥ

Ας υποθέσουμε ότι μία τράπεζα υπόκειται σε μια σειρά stress tests σε διάφορες χρονικές στιγμές οι οποίες επιλέγονται από την εποπτεύουσα αρχή με τυχαίο τρόπο, έτσι ώστε να μην μπορεί η ελεγχόμενη τράπεζα να προβλέψει το χρόνο ελέγχου και επομένως μην έχει τη δυνατότητα να παρουσιάσει αλλοιωμένα στοιχεία στην ελέγχουσα αρχή. Ας συμβολίσουμε με

- Y_1 τη χρονική στιγμή που διενεργείται το πρώτο stress test.
- Y_t τον ενδιάμεσο χρόνο μεταξύ του t-1 και του t stress test (t>1).

Τότε, ο συνολικός χρόνος μέχρι να ανιχνευθεί ο επικείμενος κίνδυνος χρεοκοπίας του ελεγχόμενου οργανισμού θα περιγράφεται από το τυχαίο άθροισμα

$$S_k = \sum_{t=1}^{T_k} Y_t$$

όπου $Y_1, Y_2, \ldots$ είναι μια ακολουθία θετικών (συνήθως ανεξάρτητων και ισόνομων) τυχαίων μεταβλητών και T_k είναι ο χρόνος αναμονής μέχρι να ικανοποιηθεί το

κριτήριο που χρησιμοποιούμε (stopping rule). Στη συνέχεια οι τυχαίες μεταβλητές Y_1, Y_2,… θα θεωρούνται ανεξάρτητες από την τυχαία μεταβλητή T_k κάτι το οποίο ισχύει στις περισσότερες εφαρμογές που παρουσιάζουν πρακτικό ενδιαφέρον.

Όπως αναφέρθηκε και στην Ενότητα 1, το κριτήριο που θα χρησιμοποιήσουμε για να χαρακτηρισθεί η τράπεζα επικίνδυνη για χρεοκοπία είναι η αδυναμία της να περάσει επιτυχώς stress tests που απέχουν μεταξύ τους μικρή απόσταση η οποία θα καθορίζεται στη συνέχεια μέσω μιας (ακέραιας) παραμέτρου k. Πιο συγκεκριμένα η τράπεζα θα χαρακτηρίζεται ως επικίνδυνη για χρεοκοπία αν η απόσταση ανάμεσα σε δύο αποτυχημένα stress tests είναι μικρότερη ή ίση από k.

Προκειμένου να μπορέσουμε να περιγράψουμε μαθηματικά το στοχαστικό μοντέλο που αποτυπώνει τον προαναφερθέντα μηχανισμό ελέγχου, θα συμβολίσουμε με $\xi_1, \xi_2, \ldots$ τα αποτελέσματα των διαδοχικών stress tests, δηλαδή

$$\xi_t = \begin{cases} 1(Success), & \text{αν η τράπεζα απέτυχε κατά το } t-\text{οστό stress test} \\ 0(Failure), & \text{αν η τράπεζα δεν απέτυχε κατά το } t-\text{οστό stress test.} \end{cases}$$

Στη συνέχεια θα συμβολίζουμε με T_k το χρόνο αναμονής μέχρι την πρώτη εμφάνιση δύο επιτυχιών που απέχουν μεταξύ τους απόσταση το πολύ k δηλαδή παρεμβάλλονται μεταξύ τους το πολύ $k-2$ αποτυχίες. Προφανώς η τυχαία μεταβλητή T_k απαριθμεί το πλήθος των δίτιμων δοκιμών που χρειάζονται μέχρι να παρατηρηθεί για πρώτη φορά ένας από τους επόμενους σχηματισμούς

$$SS, SFS, \ldots, S\overbrace{F \ldots F}^{k-2}S.$$

Σημειωτέον ότι η τυχαία μεταβλητή T_k μπορεί να θεωρηθεί ως ο χρόνος αναμονής μέχρι την πρώτη εμφάνιση μιας συνάρτησης σάρωσης ή ισοδύναμα μιας γενικευμένης ροής τύπου r/k με $r=2$ (βλέπε Balakrishnan and Koutras (2002), Chen and Glaz (1997), Glaz and Naus (1991)).

Αν συμβολίσουμε με $P_{T_k}(z), P_Y(z), P_{S_k}(z)$ τις πιθανογεννήτριες των τυχαίων μεταβλητών T_k, Y_t, S_k αντίστοιχα, δηλαδή

$$P_{T_k}(z) = E(z^{T_k}) = \sum_{t=1}^{\infty} P(T_k = t) z^t ,$$

$$P_Y(z) = E(z^{Y_t}) = \sum_{x=1}^{\infty} P(Y_t = x) z^x ,$$

$$P_{S_k}(z) = E(z^{S_k}) = \sum_{t=1}^{\infty} f_k(t) z^t = \sum_{t=1}^{\infty} P(S_k = t) z^t$$

τότε είναι γνωστό ότι η πιθανογεννήτρια του τυχαίου αθροίσματος $S_k = \sum_{t=1}^{T_k} Y_t$ θα δίνεται από τον τύπο (βλέπε Bowers *et al.* (1997))

$$P_{S_k}(z) = P_k(P_Y(z)). \tag{1}$$

Αντίστοιχα, αν συμβολίσουμε με

$$M_Y(z) = E(e^{zY_t}) = \sum_{x=1}^{\infty} P(Y_t = x)e^{zx},$$

$$M_{S_k}(z) = E(e^{zS_k}) = \sum_{t=1}^{\infty} f_k(t)e^{zt} = \sum_{t=1}^{\infty} P(S_k = t)e^{zt}$$

τις ροπογεννήτριες των τυχαίων μεταβλητών Y_t, S_k, η ροπογεννήτρια του τυχαίου αθροίσματος S_k θα δίνεται από τον τύπο

$$M_{S_k}(z) = P_{T_k}(M_Y(z)). \tag{2}$$

Αξίζει να αναφερθεί ότι για τη μέση τιμή και τη διακύμανση του τυχαίου αθροίσματος S_k ισχύουν οι παρακάτω απλοί στη μορφή τύποι

$$\begin{aligned} E(S_k) &= E\left(\sum_{t=1}^{T_k} Y_t\right) = E(T_k)E(Y_t) \\ Var(S_k) &= Var\left(\sum_{t=1}^{T_k} Y_t\right) = E(T_k)Var(Y_t) + \left(E(Y_t)\right)^2 Var\left(T_k\right). \end{aligned} \tag{3}$$

Κλείνοντας την παράγραφο αυτή σημειώνουμε ότι, η κατανομή του τυχαίου αθροίσματος

$$S_k = \sum_{t=1}^{T_k} Y_t$$

όταν $Y_1, Y_2, \ldots$ είναι μια ακολουθία θετικών (συνήθως ανεξάρτητων και ισόνομων) τυχαίων μεταβλητών και T_k είναι ο χρόνος αναμονής μέχρι να εμφανισθεί μια ροή επιτυχιών συγκεκριμένου μήκους (δηλαδή, μέχρι να βρεθεί η τράπεζα σε κατάσταση αποτυχίας για πολλά διαδοχικά stress tests) έχει μελετηθεί σχετικά πρόσφατα από τους Koutras and Eryilmaz (2015).

3. ΜΕΛΕΤΗ ΤΗΣ ΚΑΤΑΝΟΜΗΣ ΤΟΥ ΜΟΝΤΕΛΟΥ ΚΙΝΔΥΝΟΥ

Στη συνέχεια θα χρησιμοποιούμε το σύμβολα p, q για να δηλώνουμε την πιθανότητα αποτυχίας, επιτυχίας στα διαδοχικά stress tests δηλαδή για $t = 1,2,...$ θα έχουμε $p = P(\xi_t = 1), q = P(\xi_t = 0)$ αντίστοιχα. Η μέση τιμή και η διακύμανση του χρόνου αναμονής μέχρι να δοθεί σήμα ότι η τράπεζα θα πρέπει να χαρακτηρισθεί επικίνδυνη για χρεοκοπία, βρίσκεται εύκολα με εφαρμογή των τύπων (3) και (4),

κάνοντας χρήση των παρακάτω εκφράσεων που υπάρχουν για τη μέση τιμή και τη διακύμανση της τυχαίας μεταβλητής T_k (βλέπε Balakrishnan and Koutras (2002))

$$E(T_k) = \frac{2 - q^{k-1}}{p(1 - q^{k-1})}$$

$$Var(T_k) = \frac{q}{p^2} + (2k-1)\frac{q^{k-1}}{p(1-q^{k-1})^2} + \frac{q^k}{p^2(1-q^{k-1})^2}.$$

Με απλή αντικατάσταση των τελευταίων τύπων στις (3) βρίσκουμε τις επόμενες εκφράσεις

$$E(S_k) = \frac{2 - q^{k-1}}{p(1 - q^{k-1})} E(Y_t)$$

$$Var(S_k) = \frac{2 - q^{k-1}}{p(1 - q^{k-1})} Var(Y_t) + \tag{4}$$

$$+ (E(Y_t))^2 \left[\frac{q}{p^2} + (2k-1)\frac{q^{k-1}}{p(1-q^{k-1})^2} + \frac{q}{p^2(1-q^{k-1})^2} \right].$$

Όσον αφορά τις πιθανογεννήτριες και ροπογεννήτριες της κατανομής της S_k, υπολογίζονται εύκολα με χρήση των τύπων (1), (2) και της παρακάτω έκφρασης για την πιθανογεννήτρια της τυχαίας μεταβλητής T_k (βλέπε Balakrishnan and Koutras (2002))

$$P_{T_k}(z) = E(z^{T_k}) = \frac{(pz)^2 A(z)}{1 - qz - pq^{k-1}z^{k-1}} \text{ όπου } A(z) = \frac{1 - (qz)^{k-1}}{1 - qz}. \tag{5}$$

Για μικρές τιμές της παραμέτρου k και απλές διακριτές κατανομές των ενδιαμέσων χρόνων Y_t, θα μπορούσε κανείς να κατασκευάσει αποτελεσματικά αναδρομικά σχήματα που επιτρέπουν τον γρήγορο υπολογισμό της συνάρτησης πιθανότητας της τυχαίας μεταβλητής S_k. Για παράδειγμα, αν k=3 και οι τυχαίες μεταβλητές Y_t ακολουθούν τη συνήθη γεωμετρική κατανομή με συνάρτηση πιθανότητας

$$P(Y_t = y) = \theta(1-\theta)^{y-1}, y = 1,2,\ldots$$

θα έχουμε $P_{Y_t}(z) = E(z^{Y_t}) = \theta z/(1-(1-\theta)z)$ και αντικαθιστώντας στο τύπο (1) τις ποσότητες $P_{T_k}(z)$, $P_{Y_t}(z)$ καταλήγουμε εύκολα στην παρακάτω έκφραση για την πιθανογεννήτρια της τυχαίας μεταβλητής $S_k = S_3$

$$P_{S_3}(z) = E(z^{S_3}) = \frac{(p\theta z)^2[B^2(z) - (1-p)^2\theta^2 z^2]}{((p\theta - 1)z + 1)[B^3(z) - p(1-p)^2\theta^3 z^3 - (1-p)\theta z B^2(z)]}$$

όπου $B(z) = (\theta - 1)z + 1$. Αφού

$$P_{S_3}(z) = E(z^{S_3}) = \sum_{t=1}^{\infty} f_3(t) z^t$$

όπου $f_3(t) = P(S_3 = t)$, $t = 1,2,...$, η παραπάνω σχέση μπορεί να γραφεί ισοδύναμα στη μορφή

$$((p\theta - 1)z + 1)[B^3(z) - p(1-p)^2\theta^3 z^3 - (1-p)\theta z B^2(z)] \sum_{t=1}^{\infty} f_3(t) z^t =$$

$$= (p\theta z)^2[B^2(z) - (1-p)^2\theta^2 z^2]$$

και από την τελευταία προκύπτει εύκολα ο επόμενος αναδρομικός τύπος για τη συνάρτηση πιθανότητας της τυχαίας μεταβλητής $S_k = S_3$

$$f_3(t) = 2(1-p\theta) f_3(t-1) - [(1-p\theta)^2 + (1-p)^2\theta^2] f_3(t-2) -$$

$$-(1-p)^2(p-2)\theta^3 \sum_{i=1}^{t-1} ((p-2)\theta + 1)^{i-3} f_3(t-i), \quad t \geq 3.$$

Χρησιμοποιώντας την παραπάνω αναδρομική σχέση, σε συνδυασμό με τις αρχικές συνθήκες $f_3(2) = (p\theta)^2$, $f_3(1) = 0$ μπορεί κανείς εύκολα και γρήγορα να προβεί στον υπολογισμό της συνάρτησης πιθανότητας της τυχαίας μεταβλητής S_k.

Θα δώσουμε στη συνέχεια έναν εναλλακτικό τρόπο υπολογισμού της κατανομής της τυχαίας μεταβλητής S_k με χρήση της θεωρίας των κατανομών phase-type. Η προσέγγιση αυτή επιτρέπει τη δημιουργία αναδρομικών σχέσεων στην πλέον γενική περίπτωση, δηλαδή για οποιαδήποτε κατανομή των τυχαίων μεταβλητών Y_t. Η μόνη προϋπόθεση για την εφαρμογή της συγκεκριμένης τεχνικής είναι να υπάρχει δυνατότητα υπολογισμού της συνέλιξης j τυχαίων μεταβλητών από τις $Y_1, Y_2, ...$ για $j = 1,2,...$.

Θα παρουσιάσουμε αρχικά σε συντομία την οικογένεια των κατανομών phase-type περιοριζόμενοι στη διακριτή περίπτωση. Μια κατανομή phase-type τάξης d σχετίζεται με την κατανομή του χρόνου T εισόδου στην κατάσταση απορρόφησης, για μια πεπερασμένη αλυσίδα Markov διακριτού χρόνου με d μεταβατικές καταστάσεις (transient states) και μία κατάσταση απορρόφησης. Ας συμβολίσουμε με Λ_0 τον πίνακα μεταπήδησης (τάξης $(d+1)\times(d+1)$) της αλυσίδας Markov και με

$\boldsymbol{\pi}_0 = (\pi_1, \pi_2, ..., \pi_d, \pi_{d+1})'$ το αντίστοιχο διάνυσμα-στήλη των αρχικών πιθανοτήτων, θεωρώντας ότι η κατάσταση απορρόφησης έχει τοποθετηθεί τελευταία στη σειρά δηλαδή στη θέση $d+1$. Τότε, η συνάρτηση πιθανότητας του χρόνου T εισόδου στην κατάσταση απορρόφησης θα δίνεται από τον τύπο

$$P(T = t) = \boldsymbol{\pi}' \Lambda^{t-1} \mathbf{u}, \; t = 1,2,... \tag{6}$$

όπου Λ είναι ο πίνακας διάστασης $d \times d$ που περιλαμβάνει τις πιθανότητες μεταπήδησης μεταξύ των d μεταβατικών καταστάσεων της αλυσίδας, $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_d)'$ είναι το διάνυσμα-στήλη των αρχικών πιθανοτήτων που προκύπτει αν αφαιρέσουμε τη συντεταγμένη που αντιστοιχεί στην κατάσταση απορρόφησης και $\mathbf{u} = (I_d - \Lambda)\mathbf{1}$ είναι ένα διάνυσμα-στήλη που περιλαμβάνει τις πιθανότητες μεταπήδησης από τις μεταβατικές καταστάσεις στην κατάσταση απορρόφησης ($\mathbf{1} = (1,1,...,1)'$ ενώ το σύμβολο I_d συμβολίζει τον ταυτοτικό πίνακα διάστασης $d \times d$).

Για μια διακριτή τυχαία μεταβλητή με συνάρτηση πιθανότητας της μορφής (6) θα λέμε ότι ακολουθεί μια κατανομή phase-type τάξης d με παραμέτρους $\boldsymbol{\pi}, \Lambda$, και θα χρησιμοποιούμε το συμβολισμό $T \sim PH_d(\boldsymbol{\pi}, \Lambda)$. Για περισσότερες πληροφορίες για τις κατανομές phase-type ο ενδιαφερόμενος αναγνώστης παραπέμπεται στις μονογραφίες των Neuts (1981) και He (2014).

Ο Eisele (2006) παρουσίασε ένα αναδρομικό σχήμα για τον υπολογισμό της συνάρτησης πιθανότητας του τυχαίου αθροίσματος $S = \sum_{t=1}^{T} Y_t$ όταν οι τυχαίες μεταβλητές $Y_1, Y_2, \ldots$ είναι ανεξάρτητες και ισόνομες με κοινή συνάρτηση πιθανότητας $f_Y(t)$ και $T \sim PH_d(\boldsymbol{\pi}, \Lambda)$. Το αναδρομικό σχήμα του Eisele (2006) χρησιμοποιεί δύο οικογένειες συντελεστών οι οποίες υπολογίζονται μέσω του υποστοχαστικού πίνακα Λ. Η πρώτη οικογένεια συντελεστών $b_1, \ldots, b_d$ δεν είναι τίποτε άλλο παρά οι συντελεστές του χαρακτηριστικού πολυωνύμου του πίνακα Λ, πιο συγκεκριμένα έχουμε

$$|xI_d - \Lambda| = x^d + \sum_{i=1}^{d} b_i x^{d-i}. \tag{7}$$

Η δεύτερη οικογένεια $a_1, \ldots, a_d$ υπολογίζεται μέσω των $b_1, \ldots, b_d$ και των πιθανοτήτων $P(T = t)$, $t = 1,2,...,d$ μέσω των επόμενων αναδρομικών τύπων

$$a_1 = P(T = 1), \qquad a_t = P(T = 1) + \sum_{i=1}^{t-1} b_i P(T = t - i) \text{ για } t = 2, \ldots, d. \tag{8}$$

Παρουσιάζουμε στη συνέχεια το επόμενο βασικό αποτέλεσμα που αφορά τον υπολογισμό της συνάρτησης πιθανότητας $f_k(t) = P(S_k = t)$, $t = 1,2,...$ για το μοντέλο κινδύνου που μελετάμε.

Πρόταση 1. Ας υποθέσουμε ότι το στήριγμα των τυχαίων μεταβλητών $Y_1, Y_2, \ldots$ είναι το σύνολο $\{y_0, y_0 + 1, ...\}$ και ας συμβολίσουμε με $f_Y^{*j}(t)$ την $j-$οστή συνέλιξη των $Y_1, Y_2, ..., Y_j$, δηλαδή

$$f_Y^{*j}(t) = P\left(\sum_{i=1}^{j} Y_i = t\right), \qquad j = 1,2,... \ . \tag{9}$$

Τότε η συνάρτηση πιθανότητας $f_k(t) = P(S_k = t)$, $t = 1,2,...$ του τυχαίου αθροίσματος $S_k = \sum_{t=1}^{T_k} Y_t$ ικανοποιεί το αναδρομικό σχήμα

$$f_k(t) = \begin{cases} \displaystyle\sum_{j=2}^{k} p^2(1-p)^{j-2} f_Y^{*j}(t) - (p-1)\sum_{u=1}^{t-1} f_k(u) P(Y = t-u) - \\ \qquad - p(1-p)^{k-1} \displaystyle\sum_{u=1}^{t-1} f_k(u) f_Y^{*j}(t-u), \quad \text{αν } t > y_0 k \\ \displaystyle\sum_{j=2}^{k} p^2(1-p)^{j-2} f_Y^{*j}(t) - (p-1)\sum_{u=1}^{t-1} f_k(u) P(Y = t-u), \qquad \text{αν } 1 < t \le y_0 k \end{cases}$$

με αρχικές συνθήκες $f_k(0) = f_k(1) = 0$.

Απόδειξη. Θα υπολογίσουμε αρχικά τις οικογένειες συντελεστών $b_1, \ldots, b_d$ και $a_1, \ldots, a_d$ που εμπλέκονται στο αναδρομικό σχήμα του Eisele (2006). Οι Balakrishnan and Koutras (2002) έχουν δείξει ότι η κατανομή της τυχαίας μεταβλητής T_k μπορεί να περιγραφεί με χρήση κατάλληλης εμφύτευσης σε αλυσίδα Markov με πίνακα μεταπήδησης

$$\Lambda_0 = \left[\begin{array}{cccccccc|c} 1-p & p & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 1-p & 0 & \cdots & 0 & 0 & 0 & p \\ 0 & 0 & 0 & 1-p & \cdots & 0 & 0 & 0 & p \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & p \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1-p & 0 & p \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1-p & p \\ 1-p & p & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 1 \end{array}\right]_{(k+2)\times(k+2)}$$

Το αποτέλεσμα αυτό ουσιαστικά καταδεικνύει ότι, η κατανομή τυχαίας μεταβλητής T_k ανήκει στην οικογένεια των κατανομών Phase-type, πιο συγκεκριμένα ότι $T \sim PH_{k+1}(\boldsymbol{\pi}, \Lambda)$ με παραμέτρους $\boldsymbol{\pi} = (1,0,...,0)'$ και

$$\Lambda = \begin{bmatrix} 1-p & p & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 1-p & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 1-p & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1-p & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1-p \\ 1-p & p & 0 & 0 & \cdots & 0 & 0 & 0 \end{bmatrix}_{(k+1)\times(k+1)}$$

Προκειμένου να υπολογίσουμε την πρώτη οικογένεια συντελεστών $b_1,\ldots,b_d$ του αναδρομικού σχήματος του Eisele (2006) χρειαζόμαστε τους συντελεστές του χαρακτηριστικού πολυωνύμου του πίνακα Λ. Μετά από κάποιους αλγεβρικούς υπολογισμούς προκύπτει ότι

$$|xI_k - \Lambda| = x^{k+1} + (p-1)x^k + (-1)^k (p-1)^{k-1} px = x^{k+1} + \sum_{i=1}^{k+1} b_i x^{k+1-i}$$

οπότε οι συντελεστές $b_1,\ldots,b_d$ θα δίνονται από τους τύπους

$$b_1 = p-1, \quad b_2 = \ldots = b_{k-1} = b_{k+1} = 0, \quad b_k = -p(1-p)^{k-1}$$

Για την τυχαία μεταβλητή T_k είναι εύκολο να διαπιστώσει κανείς ότι

$P(T_k = 0) = P(T_k = 1) = 0$ και

$$P(T_k = x) = (x-1)p^2 q^{x-2}, \qquad 1 < x \le k.$$

Χρησιμοποιώντας επιπλέον την αναδρομική σχέση (βλέπε Koutras (1996))

$$P(T_k = x) = qP(T_k = x-1) + pq^{x-1}P(T_k = x-k), \quad x > k$$

βρίσκουμε

$$P(T_k = x) = (x-1)p^2 q^{x-2}.$$

Αντικαθιστώντας τις παραπάνω τιμές στις σχέσεις (8) καταλήγουμε στους επόμενους τύπους για την οικογένεια συντελεστών $a_1,\ldots,a_d$ που εμφανίζονται στο αναδρομικό σχήμα του Eisele (2006)

$$a_i = \begin{cases} 0 & \text{αν } i = 1, \\ p^2(1-p)^{i-2} & \text{αν } i = 2,3,...,k, \\ 0 & \text{αν } i = k+1. \end{cases}$$

Ο αναδρομικός τύπος του Eisele (2006) για τη συνάρτηση πιθανότητας του τυχαίου αθροίσματος $S = \sum_{t=1}^{T} Y_t$ όταν οι τυχαίες μεταβλητές $Y_1, Y_2, \ldots$ είναι ανεξάρτητες και ισόνομες με κοινή συνάρτηση πιθανότητας $f_Y(t)$ και $T \sim PH_d(\boldsymbol{\pi}, \Lambda)$ έχει την εξής μορφή

$$P(S=t) = \sum_{j=1}^{\min(d,t)} a_j f_Y^{*j}(t) - \sum_{j=1}^{\min(d,t-1)} b_j \left(\sum_{u=1}^{t-1} P(S=u) f_Y^{*j}(t-u) \right), \quad t \geq 1.$$

Η απόδειξη της πρότασης μπορεί πλέον να ολοκληρωθεί εύκολα αντικαθιστώντας στον τελευταίο τύπο τις εκφράσεις που βρέθηκαν για τις οικογένειες συντελεστών $b_1, \ldots, b_{k+1}$ και $a_1, \ldots, a_{k+1}$.

$\rightleftharpoons$

Κλείνοντας την παράγραφο αυτή κρίνουμε σκόπιμο να αναφέρουμε κάποια ερωτήματα τα οποία παρουσιάζουν ενδιαφέρον για μελλοντική μελέτη του μοντέλου που εισήχθη.

Θεωρώντας ότι για μια εποπτεύουσα αρχή (regulator) είναι κεφαλαιώδους σημασίας να έχει τη δυνατότητα να λαμβάνει, μέσω των μηχανισμών που υιοθετεί, έγκαιρα σήματα για εμφάνιση προβλήματος στην κεφαλαιακή επάρκεια ενός εποπτευόμενου χρηματοπιστωτικού οργανισμού, ένα πρόβλημα που θα μπορούσε να μελετηθεί είναι η παραμετροποίηση του μοντέλου ώστε να επιτυγχάνονται συγκριμένοι στόχοι. Για παράδειγμα, θα μπορούσε κάποιος να πειραματισθεί με διάφορες κατανομές για τους ενδιάμεσους χρόνους μεταξύ διαδοχικών stress tests και να μελετήσει τις τιμές των παραμέτρων (επομένως τη συχνότητα με την οποία θα γίνονται τα stress tests) οι οποίες οδηγούν σε προκαθορισμένους μέσους χρόνους αναμονής μέχρι να δοθεί σήμα επικινδυνότατος.

Επίσης, χρησιμοποιώντας τις εκφράσεις που δόθηκαν για τη συνάρτηση πιθανότητας του $S_k = \sum_{t=1}^{T_k} Y_t$, θα μπορούσε η εποπτεύουσα αρχή να καθορίζει την τιμή της παραμέτρου *k* ώστε να υπάρχει συγκεκριμένη (υψηλή) πιθανότητα εκπομπής σήματος κινδύνου, για δεδομένη τιμή της πιθανότητας *p* (η οποία σχετίζεται με την πιθανότητα μια τράπεζα να βρεθεί υποκεφαλαιοποιημένη σε κάποιο stress test).

Ένα άλλο σημείο το οποίο παρουσιάζει ενδιαφέρον είναι η ανάπτυξη παρόμοιων μοντέλων με αυτό που εξετάσθηκε στην παρούσα εργασία για την περίπτωση που οι δίτιμες μεταβλητές μέσω των οποίων καθορίζεται η τυχαία μεταβλητή T_k είναι εξαρτημένες. Ένα τέτοιο μοντέλο είναι πιο γενικό αλλά και πιο ρεαλιστικό από αυτό που εξετάσαμε στα πλαίσια της παρούσας εργασίας αφού για παράδειγμα, όταν οι

δίτιμες μεταβλητές καθορίζονται με βάση το «λόγο κεφαλαιοποίησης» *CAR*, είναι πολύ λογικό να θεωρεί κανείς ότι υπάρχει Μαρκοβιανή εξάρτηση μεταξύ των τυχαίων μεταβλητών *CAR* σε διάφορες χρονικές στιγμές, άρα και μεταξύ των αντίστοιχων δίτιμων μεταβλητών που καθορίζονται μέσω των τελευταίων. Μοντέλα τα οποία επιτρέπουν την ύπαρξη εξάρτησης με την έννοια που περιγράψαμε παραπάνω θα αποτελέσουν αντικείμενο μελλοντικής έρευνας στη συγκεκριμένη περιοχή.

4. ΣΥΝΟΨΗ

Στην παρούσα εργασία εισήχθη μια σύνθετη κατανομή που αφορά το τυχαίο άθροισμα τυχαίων μεταβλητών των οποίων το πλήθος καθορίζεται με βάση την τιμή μιας συνάρτησης σάρωσης.

Για τη νέα κατανομή δόθηκαν τύποι υπολογισμού των παραμέτρων της (μέση τιμή και διακύμανση) καθώς επίσης και τρόποι υπολογισμού της αντίστοιχης συνάρτησης πιθανότητας μέσω αναδρομικών σχημάτων που κάνουν χρήση

α. των πιθανογεννητριών της κατανομής
β. ποσοτήτων οι οποίες εξάγονται μέσω κατάλληλων πινάκων μετάβασης αλυσίδων Markov.

Τέλος δείχνηκε πως η προτεινόμενη κατανομή μπορεί να εφαρμοσθεί αποτελεσματικά στο χώρο των χρηματοοικονομικών. Πιο συγκεκριμένα διαμορφώθηκε και μελετήθηκε ένα μοντέλο που θα μπορούσε να χρησιμοποιηθεί από μια εποπτεύουσα αρχή (regulator) για την ανίχνευση προβλήματος στην κεφαλαιακή επάρκεια ενός εποπτευόμενου χρηματοπιστωτικού οργανισμού, για την ανίχνευση κινδύνων (πιστωτικών ή λειτουργικών) με τυχαιοποιημένα σχήματα ελέγχου των υποκείμενων μονάδων κ.ά.

ABSTRACT

In this article, we consider a random variable related to a binary scan statistic which is of special interest in the area of Financial risk management (bank regulation). More specifically, we introduce and study a compound distribution that makes use of the waiting time for the first appearance of a pair of successes separated by a pre-specified number of failures. Several formulae are provided for the probability mass function, probability generating function and moments of the distribution.

The underlying probability model can be briefly described as follows: consider an infinite sequence of binary outcomes (success, *S*-failure, *F*) and denote by T_k the waiting time for the first occurrence of two successes which lie at most *k* places apart from each other, i.e. they are separated by at most $k-2$ failures. Assume further that $Y_1, Y_2, \ldots$ is a sequence of independent and identically distributed discrete random

variables which are independent of T_k. In the present article we study the distribution of the random sum $S_k = \sum_{t=1}^{T_k} Y_t$.

Ευχαριστίες

Η παρούσα έρευνα έχει συγχρηματοδοτηθεί από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο - ΕΚΤ) και από εθνικούς πόρους μέσω του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» του Εθνικού Στρατηγικού Πλαισίου Αναφοράς (ΕΣΠΑ) – Ερευνητικό Χρηματοδοτούμενο Έργο: Αριστεία ΙΙ. Επένδυση στην κοινωνία της γνώσης μέσω του Ευρωπαϊκού Κοινωνικού Ταμείου.

ΑΝΑΦΟΡΕΣ

Balakrishnan, N. and Koutras, M. V. (2002). *Runs and Scans with Applications*, New York: John Wiley.

Berger, A. R. DeYoung, M. Flannery, D. Lee and O. Oztekin, (2008). How do large banking organizations manage their capital ratios? *Journal of Financial Services Research*, **34**, 123-149.

Bowers, N.L., Hickman, J.C., Gerber, H.U., Nesbitt, C.J. and D.A. Jones (1997). *Actuarial Mathematics* (2nd Edition), Society of Actuaries, Schaumburg, Illinois.

Chen, J. and Glaz, J. (1997). Approximations and inequalities for the distribution of a scan statistic for 0-1 Bernoulli trials. In *Advances in Combinatorial Methods and Applications to Probability and Statistics* (Ed. N. Balakrishnan), Birkhaeuser, Boston.

Eisele K-T (2006). Recursions for compound phase distributions, *Insurance: Mathematics & Economics*, **38**, 149-156.

Estrella, A., Park, S. and S. Peristiani (2000). Capital ratios as predictors of bank failure, *Federal Reserve Bank of New York Economic Policy Review ,* **6**, 33-52.

Glaz, J. and J. Naus (1991). Tight bounds and approximations for scan statistic probabilities for discrete data. *The Annals of Applied Probability*, 1, 306-318.

Goldberg, L. and S. Hudgins, (2002), Depositor discipline and changing strategies for regulating thrift institutions, *Journal of Financial Economics,* **63**, 263-274.

He Q-M. (2014). *Fundamentals of Matrix-Analytic Methods*, Springer.

Koutras, M. V. (1996). On a waiting time distribution in a sequence of Bernoulli trials. *Annals of the Institute of Statistical Mathematics*, **48**, 789-806.

Koutras, M. V. and S. Eryilmaz (2015). Compound geometric distribution of order *k*. *Submitted for publication*.

Lindquist, K. (2004). Banks buffer capital: How important is risk, *Journal of International Money and Finance,* **23**, 493-513.

Neuts M. F. (1981). *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. The Johns Hopkins University Press, Baltimore.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής (2015), σελ. 178−192

ΑΣΥΜΠΤΩΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΓΙΑ ΤΗΝ ΠΟΛΛΑΠΛΗ ΣΥΝΑΡΤΗΣΗ ΣΑΡΩΣΗΣ

Μ. Κούτρας[1], Δ. Λυμπερόπουλος[2]
[1]Πανεπιστήμιο Πειραιώς
mkoutras@unipi.gr
[2]Πανεπιστήμιο Πειραιώς
dilyber@webmail.unipi.gr

ΠΕΡΙΛΗΨΗ

Η έννοια του σχηματισμού εμφανίζεται σε πολλές εφαρμογές που περιλαμβάνουν τη μελέτη πειραματικών δοκιμών με δύο ή περισσότερα δυνατά αποτελέσματα ανά δοκιμή. Η δίτιμη σάρωση τύπου r/k είναι ένας ειδικός σχηματισμός που αναφέρεται στην εμφάνιση υπακολουθιών μήκους k, και αφορά συνήθως σε ακολουθίες δοκιμών τύπου «επιτυχία-αποτυχία», οι οποίες περιλαμβάνουν τουλάχιστον r-επιτυχίες (r, k θετικοί ακέραιοι με $r \leq k$). Η πολλαπλή συνάρτηση σάρωσης $W_{t,k,r}$ ορίζεται ως η τυχαία μεταβλητή που απαριθμεί τα επικαλυπτόμενα κυλιόμενα παράθυρα που εμφανίζονται μέχρι τον χρόνο t και περιλαμβάνουν μια σάρωση τύπου r/k. Επειδή τόσο η κατανομή πιθανότητας της πολλαπλής συνάρτησης σάρωσης όσο και η στοχαστική συμπεριφορά τυχαίων μεταβλητών που σχετίζονται με την $W_{t,k,r}$ δεν μπορούν συνήθως να καθοριστούν αναλυτικά, η αναζήτηση φραγμάτων για τις πιθανότητες ή τις αναμενόμενες τιμές των εν λόγω τυχαίων μεταβλητών αποδεικνύεται ιδιαίτερα χρήσιμη. Στην παρούσα εργασία θεωρούμε μια ακολουθία ανεξάρτητων δίτιμων δοκιμών με όχι κατ΄ ανάγκη ίσες πιθανότητες επιτυχίας, και εξάγουμε άνω φράγματα για την πιθανότητα του ενδεχομένου η πολλαπλή συνάρτηση σάρωσης να παρουσιάσει ένα τουλάχιστον άλμα από το ℓ στο $\ell + 1$ (όπου ℓ είναι δοσμένος θετικός ακέραιος) σε πεπερασμένο χρονικό ορίζοντα.

Λέξεις Κλειδιά: πολλαπλή συνάρτηση σάρωσης, άνω φράγμα, demisubmartingale, N - demisupermartingale, demimartingale.

1. ΕΙΣΑΓΩΓΗ

Στη στατιστική βιβλιογραφία ως *πολλαπλή συνάρτηση σάρωσης* αναφέρεται η απαριθμήτρια τυχαία μεταβλητή (τ.μ.) των επικαλυπτόμενων κυλιόμενων παραθύρων, που εμφανίζονται μέχρι τον χρόνο t και περιλαμβάνουν μια σάρωση τύπου r/k. Η εν λόγω συνάρτηση, μεταξύ άλλων, βρίσκει παρόμοιες εφαρμογές στη μελέτη αρκετών

προβλημάτων στα οποία έχει χρησιμοποιηθεί η αρκετά απλούστερη έννοια της ροής επιτυχιών, βλ. τους Balakrishnan & Koutras (2002, Chapter 12).

Υπενθυμίζουμε ότι αν η $\{X_n\}_{n\in\mathbb{N}}$ ($\mathbb{N} := \{1, 2, \ldots\}$) είναι μια ακολουθία δίτιμων δοκιμών σε έναν αυθαίρετο αλλά σταθερό χώρο πιθανότητας (Ω, Σ, P), τέτοιων ώστε η έκβαση καθεμίας από αυτές να είναι είτε *επιτυχία* (δηλ. $\{X_n = 1\}$) είτε *αποτυχία* (δηλ. $\{X_n = 0\}$) με όχι κατ' ανάγκη ίσες πιθανότητες επιτυχίας p_n ($0 < p_n < 1$), τότε για οποιοδήποτε σταθερό $k \in \mathbb{N}$ και για κάθε $m \in \mathbb{N}$ ώστε $m \leq k$, μια ακολουθία $X_n, X_{n+1}, \ldots, X_{n+m}$ τ.μ. επάνω στο Ω ονομάζεται ένα *κυλιόμενο παράθυρο (για την $\{X_n\}_{n\in\mathbb{N}}$) μήκους m*. Ιδιατέρως, αν το ενδεχόμενο $\{\sum_{j=n}^{n+m-1} X_j \geq r\}$ δεν είναι ένα σύνολο μηδενικής πιθανότητας (κάτω από το μέτρο P), η παραπάνω υπακολουθία της $\{X_n\}_{n\in\mathbb{N}}$ ονομάζεται μια *$(P-)$σάρωση ή $(P-)$γενικευμένη ροή τύπου r/k*. Δηλαδή, ο όρος «σάρωση τύπου r/k» αναφέρεται σε υπακολουθίες $X_n, X_{n+1}, \ldots, X_{n+m}$ μήκους $m \leq k$ που περιέχουν με θετική πιθανότητα τουλάχιστον r-επιτυχίες.

Η πολλαπλή συνάρτηση σάρωσης $W_{t,k,r}$ ορίζεται ως η τυχαία μεταβλητή που απαριθμεί τα επικαλυπτόμενα κυλιόμενα παράθυρα που εμφανίζονται μέχρι τον χρόνο t και περιλαμβάνουν μια σάρωση τύπου r/k (*ο πλήρης ορισμός δίνεται στην Ενότητα 3*). Στην παρούσα εργασία παρέχονται μερικά ασυμπτωτικά αποτελέσματα για πιθανότητες σχετιζόμενες με τη στοχαστική διαδικασία (σ.δ.) $\{W_{t,k,r}\}_{t\in\mathbb{N}}$ πολλαπλής συνάρτησης σάρωσης. Ειδικότερα, αρχικά αναπτύσσεται ένα άνω φράγμα για την πιθανότητα του ενδεχομένου η πολλαπλή συνάρτηση σάρωσης να παρουσιάσει ένα άλμα από το ℓ στο $\ell + 1$ μέχρι τον χρόνο t, όπου $t, \ell \in \mathbb{N}$ με $t \geq k$ (βλ. Ενότητα 3). Ως συνέπεια, δίνεται ένα άνω φράγμα για τον αναμενόμενο αριθμό μοναδιαίων αλμάτων της ίδιας σ.δ. και πάλι σε πεπερασμένο χρονικό ορίζοντα.

Τα παραπάνω αποτελέσματα βασίζονται στο ότι η σ.δ. $\{W_{t,k,r}\}_{t\in\mathbb{N}}$ αποδεικνύεται πως είναι demisubmartingale (βλ. Λήμμα 1). Ο τελευταίος ισχυρισμός χρησιμοποιείται σε συνδυασμό με μια σχετικά απλή γενίκευση μιας γνωστής ανισότητας για demisubmartingales, για την εξαγωγή των ζητουμένων φραγμάτων. Με αφορμή το γεγονός αυτό, καθώς και το ότι για οποιαδήποτε σταθερά $r, k \in \mathbb{N}$ με $r \leq k$ οι τ.μ. $W_{t,k,r}$ είναι εξαρτημένες, στην τρίτη ενότητα διερευνάται το αν η σ.δ. πολλαπλής συνάρτησης σάρωσης ανήκει στις κλάσεις των demimartingales και N-demisupermartingales.

Οι Ενότητες 4 και 5 περιλαμβάνουν μια αριθμητική μελέτη των εξαχθέντων φραγμάτων, καθώς και μια συζήτηση πιθανών πεδίων εφαρμογής τους, αντίστοιχα, ολοκληρώνοντας έτσι τους σκοπούς αυτής της εργασίας.

2. ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ ΚΑΙ ΕΝΑ ΧΡΗΣΙΜΟ ΑΠΟΤΕΛΕΣΜΑ

Με $\mathbb{R}$ συμβολίζεται το σύνολο των πραγματικών αριθμών, ενώ για $d \in \mathbb{N}$, με $\mathbb{R}^d$ συμβολίζεται ο Ευκλείδιος χώρος διάστασης d. Επί πλέον, θα χρησιμοποιήσουμε τους συμβολισμούς $x \wedge y := \min\{x, y\}$ και $x^+ := \max\{x, 0\}$ για $x, y \in \mathbb{R}$. Για κάθε $n \in \mathbb{N}$ και $i \in \{1, \ldots, n\}$ η i-κανονική προβολή από το $\mathbb{R}^n$ επάνω στο $\mathbb{R}$ θα σημειώνεται με π_i.

Στο εξής, θεωρούμε έναν χ.π. (Ω, Σ, P). Ένα σύνολο $N \in \Sigma$ με $P(N) = 0$ ονομάζεται *P-μηδενικό σύνολο* (ή απλώς μηδενικό σύνολο). Η οικογένεια όλων των

P-μηδενικών συνόλων θα συμβολίζεται με Σ_0. Για οποιεσδήποτε δύο μετρήσιμες απεικονίσεις Z_1, Z_2 επάνω στο Ω γράφουμε $Z_1 = Z_2\ P-$σχεδόν βέβαια (σ.β.), αν $\{Z_1 \neq Z_2\} \in \Sigma_0$. Με $\sigma(Z) := \{Z^{-1}(B) : B \in \mathfrak{B}\}$ συμβολίζεται η ελάχιστη σ-άλγεβρα που παράγεται από τη Σ-μετρήσιμη συνάρτηση Z, όπου με $\mathfrak{B} := \mathfrak{B}(\mathbb{R})$ δηλώνεται η Borel σ-άλγεβρα υποσυνόλων του $\mathbb{R}$.

Θέτοντας $T_Z := \{B \subseteq \mathbb{R} :\ Z^{-1}(B) \in \Sigma\}$, προφανώς έχουμε $\mathfrak{B} \subseteq T_Z$. Με P_Z συμβολίζουμε το μέτρο-εικόνα του P κάτω από την Z, και πάλι με P_Z τον περιορισμό του στη $\mathfrak{B}$, ενώ με R_Z θα σημειώνεται το σύνολο τιμών της τ.μ. Z. Ο συμβολισμός $\mathbf{PB}(n, p_1, \ldots, p_n)$, όπου $n \in \mathbb{N}$ και $p_j \in (0,1)$ για $j \in \{1, \ldots, n\}$, δηλώνει τη διωνυμική κατανομή του Poisson. Πιο συγκεκριμένα, η $\mathbf{PB}(n, p_1, \ldots, p_n)$ αποτελεί την κατανομή πιθανότητας του αθροίσματος n-ανεξάρτητων τ.μ. Bernoulli με πιθανότητες επιτυχίας $p_1, \ldots, p_n$ (βλ. π.χ. Wang (1993, Section 3) για περισσότερες λεπτομέρειες).

Η οικογένεια όλων των πραγματικών P-ολοκληρώσιμων συναρτήσεων του Ω συμβολίζεται με $\mathcal{L}^1(P)$. Συναρτήσεις που είναι P-σ.β. ίσες δεν ταυτίζονται. Η μέση τιμή της τ.μ. Z θα σημειώνεται με $\mathbb{E}_P[Z]$. Αν η $Z \in \mathcal{L}^1(P)$ και η $\mathcal{F}$ είναι μια σ-υποάλγεβρα της Σ, τότε κάθε συνάρτηση $W \in \mathcal{L}^1(P \mid \mathcal{F})$ που ικανοποιεί για κάθε $F \in \mathcal{F}$ την ισότητα $\int_F Z\, dP = \int_F W\, dP$ ονομάζεται *μία εκδοχή της δεσμευμένης μέσης τιμής* της Z δοθείσης της $\mathcal{F}$, και συμβολίζεται με $\mathbb{E}_P[Z \mid \mathcal{F}]$. Επί πλέον, για κάθε $E \in \Sigma$ θέτουμε $P(E \mid F) := \mathbb{E}_P[\chi_E \mid \mathcal{F}]$, όπου με χ_E σημειώνεται η δείκτρια (ή χαρακτηριστική) συνάρτηση του συνόλου E.

Υπενθυμίζουμε ακόμη ότι μια οικογένεια $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ σ-υποαλγεβρών της Σ, τέτοια ώστε $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$ για κάθε $n \in \mathbb{N}$, ονομάζεται *διύλιση* για τον μετρήσιμο χώρο (Ω, Σ). Επί πλέον, μια ακολουθία $\{Z_n\}_{n\in\mathbb{N}}$ τ.μ. επάνω στο Ω ονομάζεται *προσαρμοσμένη σε μια διύλιση* $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ αν κάθε Z_n είναι $\mathcal{F}_n$-μετρήσιμη. Αν $\mathcal{F}_n = \sigma(\bigcup_{j=1}^n \sigma(Z_j))$ για κάθε $n \in \mathbb{N}$, τότε η $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ καλείται η *κανονική διύλιση* για την $\{Z_n\}_{n\in\mathbb{N}}$, και θα σημειώνεται με $\{\mathcal{F}_n^{(Z)}\}_{n\in\mathbb{N}}$.

Ορισμοί 1. Έστω $\{Z_n\}_{n\in\mathbb{N}}$ μια ακολουθία στο $\mathcal{L}^1(P)$. Τότε η $\{Z_n\}_{n\in\mathbb{N}}$ ονομάζεται:

(a) *martingale* επάνω στον (Ω, Σ, P) προσαρμοσμένο σε μια διύλιση $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ ή απλώς $(P, \{\mathcal{F}_n\}_{n\in\mathbb{N}})$-*martingale*, αν η $\{Z_n\}_{n\in\mathbb{N}}$ είναι προσαρμοσμένη στην $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ και για κάθε $n \in \mathbb{N}$ ισχύει η συνθήκη

$$\mathbb{E}_P[Z_{n+1} \mid \mathcal{F}_n] = Z_n\ \ P \mid \mathcal{F}_n - \text{σ.β.}.$$

Ιδιαιτέρως, αν $\mathcal{F}_n = \mathcal{F}_n^{(Z)}$ για κάθε $n \in \mathbb{N}$, τότε η $\{Z_n\}_{n\in\mathbb{N}}$ θα ονομάζεται απλώς P-*martingale*.

(b) *demimartingale* κάτω από το P ή απλώς P-*demimartingale*, αν

$$\mathbb{E}_P[(Z_{n+1} - Z_n) f(Z_1, \ldots, Z_n)] \geq 0 \quad \text{για κάθε } n \in \mathbb{N} \tag{1}$$

και για κάθε f μη φθίνουσα κατά συντεταγμένες συνάρτηση επάνω στο $\mathbb{R}^n$ για την οποία ορίζονται οι παραπάνω μέσες τιμές.

(c) *demisubmartingale* αν η συνθήκη (1) ικανοποιείται για κάθε f όπως στο (b) αλλά με $f \geq 0$.
(d) *N-demimartingale κάτω από το P*, αν η συνθήκη (1) αλλά με την αντίστροφη ανισότητα ικανοποιείται για κάθε f όπως στο (b). Ιδιαιτέρως, αν η $f \geq 0$ τότε η $\{Z_n\}_{n\in\mathbb{N}}$ ονομάζεται *N-demisupermartingale κάτω από το P*.

Παρατηρήσεις 1. (a) Η κλάση όλων των P-martingales είναι υποσύνολο των demimartingales, τα οποία με τη σειρά τους αποτελούν μια υποκλάση αυτής όλων των demisubmartingales. Για περισσότερα πάνω στους Ορισμούς 1 και τον τρόπο που οι έννοιες των martingales, demimartingales και demisubmartingales σχετίζονται μεταξύ τους παραπέμπουμε στον Prakasa Rao (2012, Section 2.1). Είναι ακόμη ξεκάθαρο ότι κάθε N-demimartingale είναι επίσης N-demisupermartingale.
(b) Μπορεί εύκολα να αποδειχθεί ότι αν μια ακολουθία $\{Z_n\}_{n\in\mathbb{N}}$ τ.μ. επάνω στο Ω είναι P-demisubmartingale ή N-demisupermartingale κάτω από το P, τότε το ίδιο ισχύει και για την ακολουθία $\{Z_n^{(m_1,m_2)}\}_{n\in\mathbb{N}}$ που ορίζεται μέσω της σχέσης

$$Z_n^{(m_1,m_2)} := \begin{cases} Z_{m_1} & \text{αν } \ n \in \{1,\ldots,m_1-1\} \\ Z_n & \text{αν } \ n \in \{m_1,\ldots,m_2\} \\ Z_{m_2} & \text{αν } \ n \in \{m_2+1,\ldots\}, \end{cases}$$

όπου $m_1, m_2 \in \mathbb{N}$ με $m_1 < m_2$.

Έστω, τώρα, μια ακολουθία $\{Z_n\}_{n\in\{1,\ldots,m\}}$ τ.μ. επάνω στο Ω, και ας θεωρήσουμε για οποιαδήποτε σταθερά $a, b \in \mathbb{R}$ με $a < b$ την τ.μ. $U_{a,b}$, που ορίζεται επάνω στο Ω μέσω των σχέσεων

$$U_{a,b} := U_{a,b}(Z_1,\ldots,Z_m) := \max\{k \in \mathbb{N} : J_{2k} < m+1\},$$

όπου

$$J_{2k-1} := \begin{cases} m+1, & \text{αν } \ \{n \in \mathbb{N} : J_{2k-2} < n \leq m,\ Z_n \leq a\} = \emptyset \\ \min\{n \in \mathbb{N} : J_{2k-2} < n \leq m,\ Z_n \leq a\}, & \text{αλλιώς} \end{cases}$$

και

$$J_{2k} := \begin{cases} m+1, & \text{αν } \ \{n \in \mathbb{N} : J_{2k-1} < n \leq m,\ Z_n \geq b\} = \emptyset \\ \min\{n \in \mathbb{N} : J_{2k-1} < n \leq m,\ Z_n \geq b\}, & \text{αλλιώς} \end{cases}$$

για κάθε $k \in \mathbb{N}$, και $J_0 := 0$ (βλ. π.χ. Prakasa Rao (2012, σελ. 4)).

Το επόμενο αποτέλεσμα είναι ιδιαίτερα χρήσιμο για τους σκοπούς αυτής της εργασίας.

Πόρισμα 1. *Έστω $m_1, m_2 \in \mathbb{N}$ με $m_1 < m_2$. Αν η ακολουθία $\{Z_n\}_{n\in\mathbb{N}}$ είναι P-demisubmartingale, τότε*

$$\mathbb{E}_P[U_{a,b}(Z_{m_1},\ldots,Z_{m_2})] \leq \frac{\mathbb{E}_P[(Z_{m_2}-a)^+ - (Z_{m_1}-a)^+]}{b-a}$$

για όλα τα $a, b \in \mathbb{R}$ ώστε $a < b$.

Απόδειξη. Από την Παρατήρηση 1, (b) έπεται ότι η σ.δ. $\{Z_n^{(m_1,m_2)}\}_{n\in\mathbb{N}}$ είναι P-demisubmartingale. Ας θεωρήσουμε, τώρα, την ακολουθία $\{\widetilde{Z}_n\}_{n\in\mathbb{N}}$ τ.μ. επάνω στο Ω, που ορίζεται μέσω της σχέσης

$$\widetilde{Z}_n := \begin{cases} Z_{n+m_1-1}, & \text{αν } n \in \{1,\ldots,m_2-m_1+1\} \\ Z_{m_2}, & \text{αλλιώς.} \end{cases}$$

Ας θεωρήσουμε επίσης για οποιαδήποτε σταθερά $a, b \in \mathbb{R}$ με $a < b$ την τ.μ.

$$\widetilde{U}_{a,b} := U_{a,b}(\widetilde{Z}_1,\ldots,\widetilde{Z}_{m_2-m_1+1}) = U_{a,b}(Z_{m_1},\ldots,Z_{m_2}).$$

Επειδή η $\{\widetilde{Z}_n\}_{n\in\mathbb{N}}$ είναι P-demisubmartingale, αφού το ίδιο ισχύει και για την $\{Z_n^{(m_1,m_2)}\}_{n\in\mathbb{N}}$, μπορούμε να εφαρμόσουμε μια γνωστή ανισότητα για demisubmartingales (βλ. π.χ. Prakasa Rao (2012, Theorem 2.4.1)) για να συμπεράνουμε ότι

$$\mathbb{E}_P[\widetilde{U}_{a,b}] \leq \frac{\mathbb{E}_P[(\widetilde{Z}_{m_2-m_1+1}-a)^+ - (\widetilde{Z}_1-a)^+]}{b-a},$$

κάτι που αποδεικνύει το πόρισμα. □

3. Η ΔΙΑΔΙΚΑΣΙΑ ΠΟΛΛΑΠΛΗΣ ΣΥΝΑΡΤΗΣΗΣ ΣΑΡΩΣΗΣ ΩΣ ΕΝΑ DEMISUBMARTINGALE

Στην παρούσα ενότητα θεωρούμε ότι η $\{X_n\}_{n\in\mathbb{N}}$ είναι μια ακολουθία από ανεξάρτητες δίτιμες δοκιμές τέτοιες ώστε $p_n := P(X_n = 1) \in (0,1)$ και $q_n := 1 - p_n$ για κάθε $n \in \mathbb{N}$. Για κάθε $n \in \mathbb{N}$ και $k \in \mathbb{N}$ ας θεωρήσουμε την τ.μ. $Y_{n,k}$ που ορίζεται επάνω στο Ω μέσω της σχέσης

$$Y_{n,k} := \sum_{j=\max\{n-k+1,1\}}^{n} X_j.$$

Στο εξής, θέτουμε $X_0 := 0$ και υποθέτουμε ότι κάθε άθροισμα επάνω σε ένα κενό σύνολο δεικτών είναι ίσο με το μηδέν.

Η πολλαπλή συνάρτηση σάρωσης δηλώνει τον συνολικό αριθμό των επικαλυπτόμενων κυλιόμενων παραθύρων, μέχρι τον χρόνο t, που περιέχουν μία σάρωση τύπου r/k, όπου $r, k \in \mathbb{N}$ με $r \leq k$ και $t \in \mathbb{N}$. Επομένως, αν τα r, k, t είναι όπως παραπάνω, τότε η τ.μ. $W_{t,k,r}$ που ορίζεται επάνω στο Ω μέσω της σχέσης

$$W_{t,k,r} := \sum_{n=k}^{t} \chi_{[r,\infty)}(Y_{n,k})$$

θα ονομάζεται *η (t,r,k)-πολλαπλή συνάρτηση σάρωσης, η σχετιζόμενη με την ακολουθία* $\{Y_{n,k}\}_{n\in\mathbb{N}}$ ή απλώς *η πολλαπλή συνάρτηση σάρωσης* αν δεν προκαλείται σύγχυση. Επί πλέον, η ακολουθία $\{W_{t,k,r}\}_{t\in\mathbb{N}}$ θα ονομάζεται *η διαδικασία πολλαπλής συνάρτησης σάρωσης*. Σημειώνεται ότι $W_{t,k,r} = 0$ για κάθε $t < k$, και $R_{W_{t,k,r}} = \{0,\ldots,t-k+1\}$ για κάθε $t \geq k$.

Λήμμα 1. *Έστω $k \in \mathbb{N}$ και $r \in \mathbb{N}$ ώστε $r \leq k$. Τότε ισχύουν τα εξής:*

(*i*) *Η διαδικασία πολλαπλής συνάρτησης σάρωσης $\{W_{t,k,r}\}_{t\in\mathbb{N}}$ είναι P - demisubmartingale.*

Αν η ακολουθία $\{W_{t,k,r}\}_{t\in\mathbb{N}}$ είναι επιπροσθέτως P-demimartingale ή N - demisupermartingale κάτω από το P, τότε

(*ii*) *υπάρχει ένα P-μηδενικό σύνολο $O_W \in \Sigma$ τέτοιο ώστε για κάθε $\omega \notin O_W$ να ισχύει η συνθήκη $W_{t,k,r}(\omega) = 0$ για κάθε $t \in \mathbb{N}$.*

(*iii*) $P(\bigcap_{j=1}^{r}\{X_j = 1\}) = 0$.

Απόδειξη. Αρχικά ας σταθεροποιήσουμε $k, r \in \mathbb{N}$ τέτοια ώστε $r \leq k$.
(*i*) Τότε για κάθε $t \in \mathbb{N}$ έχουμε ότι $\mathbb{E}_P[W_{t,k,r}] = \sum_{n=r}^{t} P(Y_{n,k} \geq r) < \infty$, καθώς και ότι

$$W_{t+1,k,r} - W_{t,k,r} = \chi_{[r,\infty)}(Y_{t+1,k}).$$

Συνεπώς για οποιαδήποτε f μη φθίνουσα κατά συντεταγμένες συνάρτηση επάνω στο $\mathbb{R}^t$ για την οποία ορίζεται η μέση τιμή

$$\widetilde{H}_{t,k,r}(f) := \widetilde{H}_{t,k,r}(W_{1,k,r}, \ldots, W_{t,k,r}; f) := \mathbb{E}_P[(W_{t+1,k,r} - W_{t,k,r}) f(W_{1,k,r}, \ldots, W_{t,k,r})],$$

προκύπτει ότι

$$\widetilde{H}_{t,k,r}(f) = \mathbb{E}_P[\chi_{[r,\infty)}(Y_{t+1,k}) f(W_{1,k,r}, \ldots, W_{t,k,r})] \geq 0, \tag{2}$$

κάτι που αποδεικνύει το (*i*).
(*ii*) Ας υποθέσουμε ότι η $\{W_{t,k,r}\}_{t\in\mathbb{N}}$ είναι P-demimartingale. Τότε η συνθήκη (2) θα ικανοποιείται για κάθε $t \in \mathbb{N}$ και για οποιαδήποτε f μη φθίνουσα κατά συντεταγμένες συνάρτηση επάνω στο $\mathbb{R}^t$ για την οποία ορίζεται η μέση τιμή $\widetilde{H}_{t,k,r}(f)$. Συνεπώς, εφαρμόζοντας την τελευταία συνθήκη για $f = f_1 := -1$ προκύπτει η ισχύς της $P(Y_{t+1,k} \geq r) = 0$ για κάθε $t \in \mathbb{N}$, απ΄ όπου έπεται ότι $P(\bigcup_{t\in\mathbb{N}}\{Y_{t+1,k} \geq r\}) = 0$ ή ισοδύναμα ότι $P(\bigcap_{t\in\mathbb{N}}\{Y_{t+1,k} < r\}) = 1$. Επομένως, υπάρχει ένα καθολικό P-μηδενικό σύνολο $O_W := O_{Y;r,k} \in \Sigma$ τέτοιο ώστε για κάθε $\omega \notin O_W$ να ισχύει η συνθήκη $Y_{n+1,k}(\omega) < r$ για κάθε $n \in \mathbb{N}$, κάτι που εξασφαλίζει το συμπέρασμα του ισχυρισμού (*ii*).

Υποθέτοντας, τώρα, ότι η $\{W_{t,k,r}\}_{t\in\mathbb{N}}$ είναι N-demisupermartingale κάτω από το P, από το (*i*) έπεται ότι για κάθε $t \in \mathbb{N}$ και για οποιαδήποτε f μη αρνητική μη φθίνουσα κατά συντεταγμένες συνάρτηση επάνω στο $\mathbb{R}^t$, για την οποία ορίζεται η μέση τιμή $\widetilde{H}_{t,k,r}(f)$, ισχύει η συνθήκη (2) αλλά με «=» αντί του «≥». Επομένως εφαρμόζοντας την τελευταία συνθήκη για $f = f_2 := 1$ εξασφαλίζουμε και πάλι ότι $P(Y_{t+1,k} \geq r) = 0$ για κάθε $t \in \mathbb{N}$, κάτι που αποδεικνύει το (*ii*).
(*iii*) Αν η $\{W_{t,k,r}\}_{t\in\mathbb{N}}$ είναι επιπροσθέτως P-demimartingale ή N - demisupermartingale κάτω από το P, τότε από την απόδειξη του (*ii*) έχουμε ότι $P(Y_{r,k} = r) = 0$ ή ισοδύναμα $P(\bigcap_{j=1}^{r}\{X_j = 1\}) = 0$. □

Παρατηρήσεις 2. Βάσει του Λήμματος 1, παρατηρούμε τα ακόλουθα για τη διαδικασία πολλαπλής συνάρτησης σάρωσης $\{W_{t,k,r}\}_{t\in\mathbb{N}}$.

(a) Αν η ακολουθία $\{W_{t,k,r}\}_{t\in\mathbb{N}}$ είναι P-demimartingale ή N - demisupermartingale κάτω από το P, τότε από τον ισχυρισμό (ii) του παραπάνω λήμματος έχουμε ότι $P(T_r^{(k)} \le t) = 1 - P(W_{t,k,r} = 0) = 0$ για κάθε $t \in \mathbb{N}$.
(b) Αν η ακολουθία $\{X_n\}_{n\in\mathbb{N}}$ είναι P-ανεξάρτητη, τότε η $\{W_{t,k,r}\}_{t\in\mathbb{N}}$ δεν μπορεί να είναι ούτε P-demimartingale ούτε N-demisupermartingale κάτω από το P, αφού αν ήταν τότε θα ίσχυε ο ισχυρισμός (iii) του Λήμματος 1, κάτι που θα συνεπάγονταν ότι $p_j = 0$ για κάποιο $j \in \{1, \ldots, r+1\}$, που είναι άτοπο.
(c) Ομοίως με το (b), συμπεραίνουμε ότι αν η $\{X_n\}_{n\in\mathbb{N}}$ είναι μια P-ομογενής αλυσίδα Markov πρώτης τάξης, με πιθανότητες μετάβασης $\{p_{s,t}\}_{s,t\in\{0,1\}}$ (βλ. π.χ. Shiryaev (1984)), τέτοια ώστε

$$p_{1,1} = P(X_{n+1} = 1 \mid X_n = 1) > 0 \quad \text{για κάθε } n \in \mathbb{N},$$

τότε η $\{W_{t,k,r}\}_{t\in\mathbb{N}}$ δεν μπορεί να είναι ούτε P-demimartingale ούτε N - demisupermartingale κάτω από το P, αφού η μαρκοβιανή ιδιότητα συνεπάγεται την ισχύ της συνθήκης $P(\prod_{j=1}^{r} X_j = 1) = p_{1,1}^{r-1} p_1 > 0$ για $r \in \mathbb{N}$.

Ας σταθεροποιήσουμε, τώρα, αυθαίρετα $r, k \in \mathbb{N}$ με $r \le k$, καθώς επίσης και αυθαίρετα $0 < \varepsilon < 1/2$ και $\ell \in \mathbb{N}$. Έπειτα, ας θεωρήσουμε την ακολουθία $\{U_t(\ell,\varepsilon)\}_{t\in\mathbb{N}}$ τ.μ. επάνω στο Ω τέτοιων ώστε $U_t(\ell,\varepsilon) = 0$ για κάθε $t < k$ και

$$U_t(\ell,\varepsilon) := U_{\ell+\varepsilon,\ell+1-\varepsilon}(W_{k,k,r}, \ldots, W_{t,k,r}) \quad \text{για κάθε } t \in \{k, k+1, \ldots\}.$$

Η τ.μ. $U_t(\ell,\varepsilon)$ δηλώνει των αριθμό των ανοδικών διελεύσεων (*upcrossings*) της διαδικασίας πολλαπλής συνάρτησης σάρωσης από μια τιμή μικρότερη του $\ell+\varepsilon$ σε μια μεγαλύτερη του $\ell+1-\varepsilon$, δηλαδή τον αριθμό των αλμάτων από το ℓ στο $\ell+1$, μέχρι τον χρόνο t. Είναι ξεκάθαρο ότι η $U_t(\ell,\varepsilon)$ δεν εξαρτάται από την επιλογή του ε, οπότε για κάθε $t \in \mathbb{N}$ με $t \ge k$ η τ.μ.

$$U_t(\ell) := \lim_{\varepsilon\to 0^+} U_t(\ell,\varepsilon)$$

θα δηλώνει επίσης τον αριθμό των αλμάτων της πολλαπλής συνάρτησης σάρωσης από το ℓ στο $\ell+1$ μέχρι τον χρόνο t.

Πρόταση 1. *Έστω $\theta \in (0,\infty)$ και έστω $r, k \in \mathbb{N}$ αυθαίρετα αλλά σταθερά με $r \le k$. Αν $t \to \infty$, $p_t \to 0^+$ έτσι ώστε να ικανοποιείται η συνθήκη*

$$\lim_{t\to\infty} (t-k+1)\binom{k-1}{r-1} p_t^r q_t^{k-r+1} = \theta \tag{3}$$

τότε για κάθε $\ell \in \mathbb{N}$ ισχύει η ακόλουθη ανισότητα:

$$P\Big(\bigcup_{n=k}^{\infty} \{W_{n+1,k,r} = \ell+1,\ W_{n,k,r} = \ell\}\Big) \le \sum_{n=k+\ell}^{\infty} \big[P(Y_{n,k} \ge r) \wedge [1 - F_*(\ell;\theta,r,k)]\big], \tag{4}$$

όπου F_* είναι η συνάρτηση κατανομής μιας σύνθετης κατανομής Poisson με παραμέτρους το θ και τη συνάρτηση κατανομής

$$G(x) := G(x;r,k) := \begin{cases} 0, & \alpha\nu \quad x \le 0 \\ 1 - \frac{\binom{k-x-1}{r-1}}{\binom{k-1}{r-1}}, & \alpha\nu \quad x \in \{1,\dots,k-r\} \\ 1, & \alpha\nu \quad x \ge k-r+1. \end{cases}$$

Απόδειξη. Αρχικά σταθεροποιούμε αυθαίρετα $r,k,\ell \in \mathbb{N}$ με $r \le k$. Από τον ορισμό της τ.μ. $U_t(\ell)$ και το γεγονός ότι η πολλαπλή συνάρτηση σάρωσης μπορεί να πάρει μόνο θετικές ακέραιες τιμές, προκύπτει ότι για κάθε $t \in \mathbb{N}$ με $t \ge k$ το σύνολο τιμών της $U_t(\ell)$ ισούται με το $\{0,1\}$, κάτι που συνεπάγεται ότι

$$\mathbb{E}_P[U_t(\ell)] = P(U_t(\ell)=1) = P\Big(\bigcup_{n=k}^{t}\{W_{n+1,k,r}=\ell+1, W_{n,k,r}=\ell\}\Big),$$

οπότε

$$\lim_{t\to\infty}\mathbb{E}_P[U_t(\ell)] = P\Big(\bigcup_{n=k}^{\infty}\{Y_{n+1,k}>r, W_{n,k,r}=\ell\}\Big). \tag{5}$$

Επομένως, το όριο $\lim_{t\to\infty}\mathbb{E}_P[U_t(\ell)]$ όντως υπάρχει.

Επί πλέον το Πόρισμα 1 σε συνδυασμό με το Λήμμα 1, (i) συνεπάγεται ότι για $0<\varepsilon<1/2$ και για κάθε $t\in\mathbb{N}$ με $t\ge k$ έχουμε

$$\begin{aligned}
\mathbb{E}_P[U_t(\ell,\varepsilon)] &\le \frac{\mathbb{E}_P\big[\big(W_{t,k,r}-(\ell+\varepsilon)\big)^+ - \big(W_{k,k,r}-(\ell+\varepsilon)\big)^+\big]}{\big(\ell+1-\varepsilon\big)-\big(\ell+\varepsilon\big)} \\
&= \frac{1}{1-2\varepsilon}\mathbb{E}_P\Big[\Big(W_{t,k,r}-\ell-\varepsilon\Big)\chi_{\{W_{t,k,r}>\ell+\varepsilon\}}\Big] \\
&- \frac{1}{1-2\varepsilon}\mathbb{E}_P\Big[\Big(W_{k,k,r}-\ell-\varepsilon\Big)\chi_{\{W_{k,k,r}>\ell+\varepsilon\}}\Big] \\
&= \frac{1}{1-2\varepsilon}\mathbb{E}_P\Big[\Big(W_{t,k,r}-\ell-\varepsilon\Big)\chi_{\{W_{t,k,r}\ge\ell+1\}}\Big] \\
&- \frac{1}{1-2\varepsilon}\mathbb{E}_P\Big[\Big(W_{k,k,r}-\ell-\varepsilon\Big)\chi_{\{W_{k,k,r}\ge\ell+1\}}\Big] \\
&\le \frac{1}{1-2\varepsilon}\mathbb{E}_P\Big[\Big(W_{t,k,r}-\ell-\varepsilon\Big)\chi_{\{W_{t,k,r}\ge\ell+1\}}\Big] \\
&- \frac{1-\varepsilon}{1-2\varepsilon}P(W_{k,k,r}\ge\ell+1);
\end{aligned}$$

οπότε όταν το $\varepsilon\to 0^+$, από το Θεώρημα Μονότονης Σύγκλισης εξασφαλίζουμε ότι

$$\begin{aligned}
\mathbb{E}_P[U_t(\ell)] &\le \mathbb{E}_P\big[(W_{t,k,r}-\ell)\chi_{\{W_{t,k,r}\ge\ell+1\}}\big] \\
&\le \mathbb{E}_P\big[(W_{t,k,r}-W_{k+\ell-1,k,r})\chi_{\{W_{t,k,r}\ge\ell+1\}}\big]
\end{aligned} \tag{6}$$

$$= \mathbb{E}_P\Big[\sum_{n=k+\ell}^{t} \chi_{[r,\infty)}(Y_{n,k})\chi_{\{W_{t,k,r}\geq \ell+1\}}\Big]$$

$$\leq \sum_{n=k+\ell}^{t} [P(Y_{n,k}\geq r)\wedge P(W_{t,k,r}\geq \ell+1)]. \tag{7}$$

Επειδή, όμως, για κάθε $t\in\mathbb{N}$ με $t\geq k+\ell$ και για κάθε $n\in\{k+\ell,\ldots,t\}$ έχουμε

$$P(Y_{n,k}\geq r)\leq P(Y_{n,k}\geq 1)=1-P(Y_{n,k}=0)=1-\prod_{j=n-k+1}^{n} p_j<1$$

με την τελευταία ισότητα να ισχύει λόγω της P-ανεξαρτησίας της $\{X_n\}_{n\in\mathbb{N}}$, τότε προκύπτει ότι $\limsup_{n\to\infty} P(Y_{n,k}\geq r)<1$, που συνεπάγεται ότι

$$\sum_{n=k+\ell}^{\infty} [P(Y_{n,k}\geq r)\wedge P(W_{t,k,r}\geq \ell+1)]<\infty.$$

Επομένως, από την (7) έπεται ότι για κάθε $t\in\mathbb{N}$ με $t\geq k$ ισχύει η ανισότητα

$$\mathbb{E}_P[U_t(\ell)]\leq \sum_{n=k+\ell}^{\infty} [P(Y_{n,k}\geq r)\wedge P(W_{t,k,r}\geq \ell+1)],$$

και άρα αφήνοντας το $t\to\infty$ εξασφαλίζουμε βάσει του Θεωρήματος Fubini ότι

$$\lim_{t\to\infty}\mathbb{E}_P[U_t(\ell)] \leq \sum_{n=k+\ell}^{\infty}\lim_{t\to\infty}[P(Y_{n,k}\geq r)\wedge P(W_{t,k,r}\geq \ell+1)]$$

$$\leq \sum_{n=k+\ell}^{\infty}[P(Y_{n,k}\geq r)\wedge \lim_{t\to\infty} P(W_{t,k,r}\geq \ell+1)].$$

Το τελευταίο σε συνδυασμό με τη συνθήκη (5), το Corollary 3.3.3 των Boutsikas et al. (2009) και την (3) αποδεικνύει την ισχύ της (4). □

Παρατηρήσεις 3. Σχετικά με την Πρόταση 1 παρατηρούμε τα εξής:
(a) Για οποιαδήποτε σταθερά $r,k,\ell\in\mathbb{N}$ με $r\leq k$ και για κάθε $t\in\mathbb{N}$ με $t\geq k$ η πιθανότητα (που εμπλέκεται και στην απόδειξη της παραπάνω πρότασης) η πολλαπλή συνάρτηση σάρωσης να παρουσιάσει μέχρι τον χρόνο t ένα άλμα από το ℓ στο $\ell+1$, δίνεται από την

$$\varpi_{t,k,r}(\ell) := P\Big(\bigcup_{n=k}^{t}\{W_{n+1,k,r}=\ell+1, W_{n,k,r}=\ell\}\Big)$$

$$= P\Big(\bigcup_{n=k+\ell-1}^{t}\{W_{n+1,k,r}=\ell+1, W_{n,k,r}=\ell\}\Big)$$

$$= \sum_{n=k+\ell-1}^{t} P(W_{n+1,k,r}=\ell+1, W_{n,k,r}=\ell)$$

$$= \sum_{n=k+\ell-1}^{t} P(Y_{n,k} \geq r, W_{n,k,r} = \ell),$$

αφού τα ενδεχόμενα $\{\{W_{n+1,k,r} = \ell + 1, W_{n,k,r} = \ell\}\}_{n\in\mathbb{N}}$ είναι ασυμβίβαστα. Επομένως, έχουμε

$$\varpi_{t,k,r}(\ell) \leq \sum_{n=k+\ell}^{t+1} P(Y_{n,k} \geq r) =: u_0(r,k,\ell,t).$$

Προφανώς, $u_0(r,k,\ell,t)$ είναι ένα από τα λιγότερο εκλεπτυσμένα ή περισσότερο «αφελή» άνω φράγματα (αν όχι το πιο «αφελές») για την πιθανότητα $\varpi_{t,k,r}(\ell)$. Είναι επίσης φανερό ότι είναι λιγότερο αποτελεσματικό από το άνω φράγμα της συνθήκης (7), οπότε το $u_0(r,k,\ell) := \lim_{t\to\infty} u_0(r,k,\ell,t) = \sum_{n=k+\ell}^{\infty} P(Y_{n,k} \geq r) < \infty$ θα είναι λιγότερο αποτελεσματικό από το άνω φράγμα της παραπάνω πρότασης.

(b) Εξαιτίας του (a), μπορούμε να ισχυριστούμε ότι ο υπολογισμός της πιθανότητας του ενδεχομένου $\{Y_{n+1,k} \geq r, W_{n,k,r} = \ell\}$, όπου $n \in \{k+\ell, k+\ell+1, \ldots\}$, μας παρέχει μια ένδειξη για τις δυσκολίες του υπολογισμού της πιθανότητας $\varpi_{t,k,r}(\ell)$ (καθώς και της αντίστοιχης πιθανότητας $P(\bigcup_{n=k}^{\infty}\{W_{n+1,k} = \ell+1, W_{n,k,r} = \ell\})$ που αφορά σε άπειρο χρονικό ορίζοντα), καταδεικνύοντας με αυτόν τον τρόπο τη σημασία του άνω φράγματος της συνθήκης (4). Έτσι, σε αυτή τη λογική και μέσω ορισμένων εύκολων υπολογισμών διαπιστώνουμε ότι

$$\begin{aligned} P(Y_{t+1,k} \geq r, W_{t,k,r} = \ell) &= P(Y_{t,k} \geq r, W_{t,k,r} = \ell) \\ &+ p_{t+1} P(Y_{t,k} \geq r+1, W_{t,k,r} = \ell) \\ &- q_{t+1} P(X_{(t-k+1)^+} = 1, Y_{t,k} = r, W_{t,k,r} = \ell) \end{aligned}$$

για όλα τα r, k, t όπως παραπάνω.

(c) Για οποιαδήποτε σταθερά $r, k, \ell \in \mathbb{N}$ με $r \leq k$ και για κάθε $t \in \mathbb{N}$ με $t \geq k$, μπορούμε να φράξουμε τη μέση τιμή του δεξιού μέλους της ανισότητας (6) και ως εξής:

$$\begin{aligned} \mathbb{E}_P\big[(W_{t,k,r} - \ell)\chi_{\{W_{t,k,r} \geq \ell+1\}}\big] &= \mathbb{E}_P[W_{t,k,r}] - \ell - \mathbb{E}_P[(W_{t,k,r} - \ell)\chi_{\{W_{t,k,r} \leq \ell\}}] \\ &\leq \sum_{n=k+\ell}^{\infty} P(Y_{n,k} \geq r) - \ell + \ell P(W_{t,k,r} \leq \ell) \\ &= \sum_{n=k+\ell}^{\infty} P(Y_{n,k} \geq r) - \ell[1 - P(W_{t,k,r} \leq \ell)]. \end{aligned}$$

Επομένως, το

$$u_1(r,k,\ell,t) := \sum_{n=k+\ell}^{\infty} P(Y_{n,k} \geq r) - \ell[1 - P(W_{t,k,r} \leq \ell)]$$

είναι ένα άνω φράγμα για την πιθανότητα $\varpi_{t,k,r}(\ell)$, κάτι που σε συνδυασμό με το Θεώρημα Μονότονης Σύγκλισης και το Corollary 3.3.3 των Boutsikas et al. (2009) συνεπάγεται ότι αν η ικανοποιείται η συνθήκη (3) τότε το

$$u_1(r,k,\ell) := \lim_{t\to\infty} u_1(r,k,\ell,t) = \sum_{n=k+\ell}^{\infty} P(Y_{n,k} \geq r) - \ell[1 - F_*(\ell;\theta,r,k)]$$

(όπου η F_* είναι όπως στην Πρόταση 1) είναι ένα άνω φράγμα για την πιθανότητα $\varpi_{k,r}(\ell) := P(\bigcup_{n=k}^{\infty}\{W_{n+1,k,r} = \ell+1,\ W_{n,k,r} = \ell\})$. Προφανώς, το $u_1(r,k,\ell)$ είναι καλύτερο του $u_0(r,k,\ell)$. Σε αντίθεση με το άνω φράγμα της Πρότασης 1, όμως, δεν μπορούμε να ισχυριστούμε ότι είναι *αποδεκτό*, δηλαδή ότι παίρνει τιμές στο διάστημα $(0,1)$, για όλα τα r,k,ℓ όπως πιο πάνω. Επί πλέον, το αντίστοιχό του φράγμα που αναφέρεται σε πεπερασμένο χρονικό ορίζοντα αποδεικνύεται αποτελεσματικότερο του αντίστοιχου «αφελούς» άνω φράγματος, δηλαδή $u_1(r,k,\ell,t) < u_0(r,k,\ell,t)$, όταν και μόνο όταν

$$P(W_{t,k,r} \leq \ell) < 1 - \frac{1}{\ell}\sum_{n=k}^{k+\ell-1} P(Y_{n,k} \geq r) - \frac{1}{\ell}\sum_{n=t+2}^{\infty} P(Y_{n,k} \geq r).$$

(d) Η P-ανεξαρτησία των δίτιμων δοκιμών $\{X_n\}_{n\in\mathbb{N}}$ συνεπάγεται ότι για οποιαδήποτε σταθερά $r,k,\ell \in \mathbb{N}$ με $r \leq k$, για κάθε $t \in \mathbb{N}$ με $t \geq k$ και για κάθε $n \in \{k,\ldots,t\}$ ισχύει $P_{Y_{n,k}} = \mathbf{PB}(k; p_{n-k+1},\ldots,p_n)$, οπότε το άνω φράγμα της Πρότασης 1 καθώς και τα $u_0(r,k,\ell)$ και $u_1(r,k,\ell)$ μπορούν να υπολογιστούν ακριβώς. Μάλιστα, ο υπολογισμός της σειράς που εμπλέκεται στον υπολογισμό του φράγματος, απαιτεί αυτόν της συνάρτησης πιθανότητας της διωνυμικής κατανομής του Poisson μέσω της σχέσης

$$P(Y_{n,k} = y) = \begin{cases} \prod_{j=1}^{k} q_{n-k+1+j} & \text{αν } y = 0 \\ (1/y)\sum_{j=1}^{y}(-1)^{j-1}P(Y_{n,k} = y-j)T(j) & \text{αν } y > 0, \end{cases} \tag{8}$$

όπου $T(j) := \sum_{i=1}^{k}(p_{n-k+1+i}/q_{n-k+1+i})^j$ για κάθε j (βλ. π.χ. Shah (1994)), ενώ για τον υπολογισμό εκείνου του μέρους του φράγματος στο οποίο εμπλέκεται η σύνθετη κατανομή Poisson μπορεί να χρησιμοποιηθεί το αναδρομικό σχήμα των Bowers et al. (1997, equation (12.4.6)) ή οι αναδρομικές σχέσεις του Panjer (βλ. π.χ. Bowers et al. (1997, Theorem 12.4.3)).

(e) Από την (8) έπεται ότι αν $p_t \to 0^+$ όταν $t \to \infty$, το ίδιο θα ισχύει και για την πιθανότητα $P(Y_{t,k} = y)$ για οποιοδήποτε $y \in \{0,\ldots,k\}$, όπου τα r,k,ℓ είναι όπως στο (d). Επομένως θα υπάρχει ένας θετικός ακέραιος $t_0 \geq k$ τέτοιος ώστε $\sum_{n=k+\ell}^{\infty} P(Y_{n,k} \geq r) \simeq \sum_{n=k+\ell}^{t_0} P(Y_{n,k} \geq r)$. Το γεγονός αυτό υποδηλώνει ότι είναι πολύ πιθανό να υπάρχουν τιμές των r,k,ℓ,t για τις οποίες όλα ή κάποια από τα προαναφερθέντα άνω φράγμα είναι αποδεκτά.

Από τον ορισμό της τ.μ. $U_t(\ell)$ έχουμε ότι για οποιοδήποτε σταθερό $\ell \in \mathbb{N}$ η συνάρτηση $t \longmapsto U_t(\ell)$ είναι μη φθίνουσα, και εφαρμόζοντας το Θεώρημα Μονότονης Σύγκλισης προκύπτει ότι $\lim_{t\to\infty} \mathbb{E}_P[U_t(\ell)] = \mathbb{E}_P[\lim_{t\to\infty} U_t(\ell)] = \mathbb{E}_P[U_\infty(\ell)]$, όπου η τ.μ. $U_\infty(\ell) := \lim_{t\to\infty} U_t(\ell)$ δηλώνει τον αριθμό των αλμάτων από το ℓ στο $\ell+1$ για τη πολλαπλή συνάρτηση σάρωσης.

Ας θεωρήσουμε, τώρα, τις τ.μ. $U_\infty := \sum_{\ell\in\mathbb{N}} U_\infty(\ell)$ και $U_t := \sum_{\ell\in\mathbb{N}} U_t(\ell)$, οι οποίες δηλώνουν τον αριθμό των μοναδιαίων αλμάτων της πολλαπλής συνάρτησης σάρωσης σε άπειρο χρονικό ορίζοντα και μέχρι τον χρόνο t, αντίστοιχα. Τότε το ακόλουθο αποτέλεσμα προκύπτει ως μια συνέπεια του ορισμού της U_t, της συνθήκης (7) και του Θεωρήματος Fubini.

Πόρισμα 2. *Κάτω από τις υποθέσεις της Πρότασης 1 η συνθήκη*

$$P\Big(\bigcup_{\ell=1}^{t-k+1}\bigcup_{n=k}^{\infty}\{W_{n+1,k,r}=\ell+1,\ W_{n,k,r}=\ell\}\Big) \le \sum_{\ell=1}^{t-k+1}\sum_{n=k+\ell}^{t}[P(Y_{n,k}\ge r)\wedge P(W_{t,k,r}\ge \ell+1)]$$

ισχύει για κάθε $t \in \mathbb{N}$ με $t \ge k$.

Με αφορμή το Πόρισμα 2, εγείρεται το ερώτημα για το αν μπορούν να δοθούν άνω φράγματα και για τη μέση τιμή $\mathbb{E}_P[U_\infty]$ παρόμοια με αυτό της Πρότασης 1. Κάτι τέτοιο, όμως, είναι ένα έργο κάθε άλλο παρά τετριμμένο και για τον λόγο αυτό δεν θα μας απασχολήσει περαιτέρω στην παρούσα εργασία.

4. ΑΡΙΘΜΗΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Σε αυτή την ενότητα, εκμεταλλευόμαστε την Πρόταση 1 σε συνδυασμό με την Παρατήρηση 3, (d) για τον υπολογισμό άνω φραγμάτων της πιθανότητας $\varpi_{t,k,r}(\ell)$ για διάφορες τιμές των παραμέτρων r, k, ℓ, θ, καθώς και των πιθανοτήτων επιτυχίας p_t, που εμπλέκονται σε αυτό τον υπολογισμό και οι οποίες επί πλέον πρέπει να ικανοποιούν τις συνθήκες (3) και $\lim_{t\to 0^+} p_t = 0$.

Ακολούθως, περιγράφεται μια απλή προσέγγιση για την εξασφάλιση αυτών των φραγμάτων έτσι ώστε να ικανοποιούνται οι προαναφερθείσες δύο συνθήκες: Αρχικά σταθεροποιούμε ένα αυθαίρετο $\theta \in (0,\infty)$. Έπειτα θέτουμε τα $r, k, \ell \in \mathbb{N}$, όπου $r \le k$, ίσα με συγκεκριμένες επιθυμητές τιμές (λ.χ. $r=2,\ k=4,\ \ell=7$) και αποδίδουμε μια πολύ μικρή τιμή (λ.χ. $p_0 = 10^{-i}$, όπου $i \in \{1,2,3\}$) σε όλες τις πιθανότητες επιτυχίας p_t, ενώ θέτουμε $t_* := \lfloor \frac{\theta}{\binom{k-1}{r-1} p_0^r (1-p_0)^{k-r+1}} \rfloor + k - 1$, όπου με $\lfloor x \rfloor$ συμβολίζεται το ακέραιο μέρος του πραγματικού αριθμού x. Τότε, καθίσταται δυνατή η εφαρμογή της Πρότασης 1 σε συνδυασμό με την Παρατήρηση 3, (d), και επομένως ο υπολογισμός των ζητούμενων άνω φραγμάτων για όλα τα $t \ge t_*$.

Σύμφωνα με την παραπάνω προσέγγιση, πραγματοποιήθηκε μια αριθμητική μελέτη για το άνω φράγμα της Πρότασης 1, μέσω της οποίας διαπιστώσαμε ότι εξασφαλίζονται αποδεκτές τιμές του εν λόγω φράγματος για

- *μικρές τιμές του* θ, δηλαδή για $\theta \in (0,1)$ ή ακόμα καλύτερα για $\theta \simeq 0$,

- *μικρές ή μέτριες τιμές του* r,
- *μικρές πιθανότητες επιτυχίας* p_t, δηλαδή μεταξύ 0.001 και 0.1 (και προτιμότερα τιμές κοντά στο 0.1).

Από την άλλη μεριά, θέτοντας λ.χ. $p = 0.075,\ \theta = 1.2,\ (r,k) = (2,4),\ \ell = 7$ προκύπτει ότι τα εξαχθέντα φράγματα είναι μη αποδεκτά, αφού όλα δίνουν τιμές μεγαλύτερες της μονάδος. Στους παρακάτω πίνακες παρουσιάζονται ορισμένα απτά παραδείγματα αποδεκτών φραγμάτων. Αναφέρουμε ότι σε καθέναν από αυτούς τους πίνακες, η χρονική στιγμή t_* είναι ο πρώτος (μικρότερος) χρόνος στον οποίο έχει γίνει ο υπολογισμός του φράγματος.

I	t	55	56	57	58	59
	Άνω Φράγμα	.187234	.191489	.195744	.200000	.204255
II	t	227	228	229	230	231
	Άνω Φράγμα	.00124131	.00124693	.00125255	.00125816	.00126378
III	t	120	121	122	123	124
	Άνω Φράγμα	.00767801	.00774976	.00782152	.00789328	.00796504
IV	t	155	156	157	158	159
	Άνω Φράγμα	.00145779	.00146805	.00147832	.00148858	.00149885

Πίνακας 1. Φράγματα για την $\varpi_{t,k,r}(\ell)$ για 4 διαφορετικές επιλογές παραμέτρων
(I: $p = 0.1,\ \theta = 0.1,\ (r,k) = (4,8),\ \ell = 4$, II: $p = 0.025,\ \theta = 0.01,\ (r,k) = (3,4),\ \ell = 3$, III: $p = 0.1,\ \theta = 0.15,\ (r,k) = (4,7),\ \ell = 7$, IV: $p = .005,\ \theta = .0005,\ (r,k) = (3,9),\ \ell = 5$)

5. ΠΡΟΤΕΙΝΟΜΕΝΕΣ ΕΦΑΡΜΟΓΕΣ

Το πραγματικό ζητούμενο αυτής της ενότητας είναι η εύρεση λιγότερο ή περισσότερο πρακτικών προβλημάτων που εμφανίζονται σε διάφορες πτυχές της καθημερινότητας για τα οποία θα παρουσίαζε ενδιαφέρον η μελέτη του ενδεχομένου $\bigcup_{n=k}^{t}\{W_{n,k,r} = \ell, W_{n+1,k,r} = \ell+1\}$ ή του $\bigcup_{n=k}^{\infty}\{W_{n,k,r} = \ell, W_{n+1,k,r} = \ell+1\}$, και άρα και των πιθανοτήτων $\varpi_{t,k,r}(\ell)$ ή $\varpi_{k,r}(\ell)$ (όπου $r,k,\ell,t \in \mathbb{N}$ με $t \geq k \geq r$), αντίστοιχα. Για τον λόγο αυτό παραθέτουμε, αρχικά, το ακόλουθο παράδειγμα, που είχε ως αφορμή τη δουλειά των Boutsikas & Koutras (2002, Section 4).

Παράδειγμα 1. Αν κάθε δίτιμη τ.μ. X_n είναι η δείκτρια τ.μ. του ενδεχομένου το αποθεματικό μιας ασφαλιστικής εταιρείας να είναι κάτω από ένα επιθυμητό ή κρίσιμο επίπεδο (έστω $\alpha > 0$) στον χρόνο n, τότε η πολλαπλή συνάρτηση σάρωσης $W_{t,k,r}$ θα μετράει το πόσες φορές μέχρι τον χρόνο t η σ.δ. αποθεματικού (διακριτού χρόνου) θα παρουσιάσει r-τέτοιες πτώσεις σε χρονικά παράθυρα εύρους k. Υπό την παραδοχή ότι τιμές του $W_{t,k,r}$ που υπερβαίνουν το ℓ συνιστούν μια προειδοποίηση για την εξέλιξη του αποθεματικού της εταιρείας, η $\varpi_{k,r}(\ell)$ αποτελεί την πιθανότητα ενός «συναγερμού φερεγγυότητας» για την ασφαλιστική εταιρεία.

Στη συνέχεια προτείνεται ένα εναλλακτικό παράδειγμα.

Παράδειγμα 2. Ένας ανιχνευτής παικτών μιας ομάδας καλαθοσφαίρησης θέλει να καταρτίσει μια (σύντομη) λίστα εκτελεστών (δηλ. σουτέρ) τριών πόντων. Για το σκοπό αυτό, ρίχνει πρώτα μια ματιά στα στατιστικά των σουτέρ και αποφασίζει για το αν θα τους συμπεριλάβει στην αρχική του λίστα βάσει του ακόλουθου κριτηρίου: *εντοπισμός εκείνων των παικτών που σε περισσότερες από* $\ell = 5$ *περιπτώσεις πετύχαιναν* $r = 7$ *από τα τελευταία* $k = 10$ *τρίποντα που επιχειρούσαν κατά τη διάρκεια* $t = 100$ *προσπαθειών*. Υποθέτουμε, τώρα, ότι οι εκβάσεις των τρίποντων προσπαθειών κάθε παίκτη είναι ανεξάρτητες η μια της άλλης, ότι η πιθανότητα επιτυχίας για κάθε προσπάθεια φθίνει προϊόντος του χρόνου, και ότι ο αναμενόμενος αριθμός επιτυχημένων προσπαθειών τελικά θα καταλήξει να είναι ίσος με έναν θετικό αριθμό θ (δηλ. τις προϋποθέσεις της Πρότασης 1). Πράγματι, οι τελευταίες υποθέσεις μπορούν εύκολα να θεωρηθούν ως πολύ ρεαλιστικές, διότι θα μπορούσαν άνετα να ερμηνευτούν ως το «αντίκτυπο της κόπωσης» ($\lim_{t\to\infty} p_t = 0$) στην απόδοση ενός σουτέρ (αφού πρόκειται για σουτέρ είναι λογικό να θεωρήσουμε ότι $\theta > 0$, διότι αλλιώς δεν μπορεί να χαρακτηρίζεται ως τέτοιος!). Έτσι σε αυτό το πλαίσιο, η $\varpi_{100,10,7}(5)$ δηλώνει την πιθανότητα εντοπισμού ενός μέλους της αρχικής λίστας του ανιχνευτή.

Νύξεις και για άλλες πιθανές περιοχές εφαρμογών (π.χ. μοριακή βιολογία) των αποτελεσμάτων μας μπορούν να βρεθούν στους Boutsikas & Koutras (2002).

ABSTRACT

The concept of pattern arises in many applications comprising experimental trials with two or more possible outcomes in each trial. A binary scan of type r/k is a special pattern referring usually to "success-failure" strings of fixed length k that contain at least r-successes, where r, k are positive integers with $r \leq k$. The multiple scan statistic $W_{t,k,r}$ is defined as the enumerating random variable for the overlapping moving windows occurring until time t and including a scan of type r/k. Since its probability distribution as well as the stochastic behavior of random variables related to $W_{t,k,r}$ cannot be usually determined explicitly, it is really useful to seek some upper bounds concerning probabilities or expectations involving these variables. In the current work, considering a sequence of independent binary trials with not necessarily equal success probabilities, some upper bounds are obtained for the probability of the event that the multiple scan statistic will perform a jump from ℓ to $\ell + 1$ (where ℓ is a positive integer) in a finite time horizon.

Ευχαριστίες: Η παρούσα έρευνα έχει συγχρηματοδοτηθεί από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο - ΕΚΤ) και από εθνικούς πόρους μέσω του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» του Εθνικού Στρατηγικού Πλαισίου Αναφοράς (ΕΣΠΑ) - Ερευνητικό Χρηματοδοτούμενο Έργο: Αριστεία ΙΙ. Επένδυση στην κοινωνία της γνώσης μέσω του Ευρωπαϊκού Κοινωνικού Ταμείου.

ΑΝΑΦΟΡΕΣ

Balakrishnan, N. and Koutras, M.V. (2002). *Runs and scans with applications.* Wiley, New York.

Boutsikas, M.V. and Koutras, M.V. (2002). Modeling claim exceedances over thresholds. *Insurance Math. Econom.* **30**, 67–83.

Boutsikas, M.V., Koutras, M.V. and Milienos, F.S. (2009). Extreme Value Results for Scan Statistics in: *Scan Statistics: Methods and Applications*, chapter 3. Birkhäuser, Boston.

Bowers, N.L., Gerber, H.U., Hickman, J., Jones, D.A. and Nesbitt, C.J. (1997). *Actuarial Mathematics*, 2nd Edition. The society of Actuaries, Illinois.

Prakasa Rao, B.L.S. (2012). *Associated Sequences, Demimartingales and Nonparametric Inference.* Probability and its Applications, Springer Basel AG.

Shirayev, A.N. (1984). *Probability*, 2nd Edition. Graduate Texts in Mathematics 95, Springer, Berlin.

Shah, B.K. (1994). On the distribution of the sum of independent integer valued random variables. *Amer. Statist.* **27** (3), 123–124.

Wang, Y.H. (1993). On the number of successes in independent trials. *Stat. Sinica* **3**, 295–312.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής (2015), σελ.193-206

ΈΝΑ ΔΙΔΙΑΣΤΑΤΟ ΗΜΙΠΑΡΑΜΕΤΡΙΚΟ ΔΙΑΓΡΑΜΜΑ ΕΛΕΓΧΟΥ ΓΙΑ ΤΟ ΖΕΥΓΑΡΙ ΜΙΑΣ ΔΙΑΤΕΤΑΓΜΕΝΗΣ ΠΑΡΑΤΗΡΗΣΗΣ ΚΑΙ ΤΗΣ ΣΥΜΜΕΤΑΒΛΗΤΗΣ ΤΗΣ

Μ. Β. Κούτρας, Ε. Μ. Σοφικίτου
Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς
mkoutras@unipi.gr, esofikit@unipi.gr

ΠΕΡΙΛΗΨΗ

Η έννοια του Στατιστικού Ελέγχου Ποιότητας είναι συνδεδεμένη με τα διαγράμματα ελέγχου, καθώς αποτελούν ένα βασικό στατιστικό εργαλείο για την ανίχνευση, τον έλεγχο και τη βελτίωση της ποιότητας ενός προϊόντος ή ακόμη και ολόκληρης της παραγωγικής διαδικασίας. Τα παραμετρικά διαγράμματα ελέγχου μπορούν να χρησιμοποιηθούν μόνο όταν οι μετρήσεις που λαμβάνονται ακολουθούν κάποια γνωστή κατανομή, συνήθως την κανονική. Σε διαφορετική περίπτωση, η χρήση τους μπορεί να οδηγήσει σε λανθασμένα συμπεράσματα. Αν η κατανομή του χαρακτηριστικού ή των χαρακτηριστικών που μελετάμε δεν είναι γνωστή, τότε είναι σχεδόν αναπόφευκτη η χρήση μη παραμετρικών διαγραμμάτων.

Στην παρούσα εργασία, προτείνεται ένα νέο διδιάστατο διάγραμμα ελέγχου στο οποίο χρησιμοποείται το ζευγάρι των τυχαίων μεταβλητών $(X_{(r:n)}, Y_{[r:n]})$, όπου $X_{(r:n)}$ είναι η r-οστή διατεταγμένη παρατήρηση των παρατηρήσεων X_i, $i = 1,2,...,n$ και $Y_{[r:n]}$ η συμμεταβλητή (concomitant) που αντιστοιχεί στη $X_{(r:n)}$. Το διάγραμμα ελέγχου που προκύπτει είναι ημιπαραμετρικό, αφού τα όρια ελέγχου και η απόδοσή του δεν επηρεάζεται από τις περιθώριες κατανομές των X_i και Y_i, επηρεάζεται όμως από τη δομή εξάρτησης των υπό μελέτη μεταβλητών (X, Y).

Στα πλαίσια της εργασίας αυτής δίνεται ακριβής τύπος για τον υπολογισμό της χαρακτηριστικής συνάρτησης λειτουργίας (Operating Characteristic Function) του διαγράμματος και του ποσοστού λανθασμένων συναγερμών (False Alarm Rate). Επίσης, δίνεται έκφραση για το ποσοστό συναγερμών για συγκεκριμένες κατανομές. Τέλος, παρουσιάζονται ορισμένα αριθμητικά αποτελέσματα και συγκρίσεις.

Λέξεις Κλειδιά: Διατεταγμένες παρατηρήσεις (order statistics), συμμεταβλητές (concomitants), μη παραμετρικά διαγράμματα ελέγχου, πιθανότητα λανθασμένου συναγερμού, στατιστικός έλεγχος ποιότητας, copulas.

1. ΕΙΣΑΓΩΓΗ

Είναι κοινώς αποδεκτό ότι η βελτίωση της ποιότητας των παραγόμενων προϊόντων αποτελεί το βασικό σκοπό του Στατιστικού Ελέγχου Ποιότητας ή Διεργασιών, ο οποίος μπορεί να επιτευχθεί με τον εντοπισμό και τη μείωση των ειδικών ή συστηματικών αιτιών μεταβλητότητας, στο βαθμό βέβαια που είναι εφικτό. Τα τελευταία χρόνια, τα μη παραμετρικά διαγράμματα ελέγχου έχουν κεντρίσει το ενδιαφέρον αρκετών ερευνητών, καθώς μπορούν να αποτελέσουν χρήσιμες εναλλακτικές των αντίστοιχων παραμετρικών διαγραμμάτων σε περιπτώσεις που δεν υπάρχει πληροφορία για την κατανομή της υποκείμενης διεργασίας ή η κατανομή της δεν είναι κανονική.

Η διαδικασία κατασκευής των μη παραμετρικών διαγραμμάτων είναι η εξής. Αρχικά, λαμβάνεται ένα τυχαίο δείγμα αναφοράς (reference sample) από μία εντός ελέγχου άγνωστη συνεχή κατανομή και βάσει του δείγματος αυτού καθορίζονται τα όρια ελέγχου, συνήθως με χρήση διατεταγμένων παρατηρήσεων. Στη συνέχεια, λαμβάνονται διαδοχικά τυχαία δείγματα ελέγχου (test samples) με σκοπό να διαπιστωθεί εάν η διεργασία παραμένει εντός ελέγχου ή έχει μετατοπισθεί σε κατάσταση εκτός ελέγχου.

Οι Janacek and Meikle (1997) πρότειναν ένα μονοδιάστατο, μη παραμετρικό, δίπλευρο διάγραμμα ελέγχου (τύπου Shewhart) για τη διάμεσο. Το βασικό πλεονέκτημα του διαγράμματος αυτού βασίζεται στη χρήση της διαμέσου και ως εκ τούτου δεν είναι ευαίσθητο σε πιθανά λάθη μετρήσεων και μπορεί να χρησιμοποιηθεί σε περιπτώσεις που η μέτρηση του χαρακτηριστικού περιλαμβάνει υποκειμενική αξιολόγηση διατάξιμων δεδομένων. Γενικεύοντας την παραπάνω ιδέα, οι Chakraborti *et al.* (2004) πρότειναν ένα διάγραμμα, στο οποίο χρησιμοποιούνται τα ποσοστημόρια. Πρόσφατα, οι Balakrishnan *et al.* (2010) βελτίωσαν το παραπάνω διάγραμμα με την προσθήκη μίας επιπλέον συνθήκης, η οποία προϋποθέτει ότι τουλάχιστον r από τις n παρατηρήσεις βρίσκονται μεταξύ των ορίων ελέγχου. Επεκτείνοντας την ιδέα των Chakraborti *et al.* (2004), οι Κούτρας και Σοφικίτου (2014) πρότειναν ένα διδιάστατο διάγραμμα ελέγχου, το οποίο βασίζεται σε διατεταγμένες παρατηρήσεις και αποτελεί άμεση γενίκευση των κλασσικού μονοδιάστατου μη παραμετρικού διαγράμματος ελέγχου για τη διάμεσο. Στην παρούσα εργασία, θα παρουσιάσουμε ένα διδιάστατο ημιπαραμετρικό διάγραμμα ελέγχου για το ζευγάρι μιας διατεταγμένης παρατήρησης και της συμμεταβλητής της. Το διάγραμμα αυτό κατασκευάστηκε με σκοπό την ανίχνευση μετατοπίσεων του συντελεστή συσχέτισης.

2. ΤΟ ΝΕΟ ΔΙΔΙΑΣΤΑΤΟ ΗΜΙΠΑΡΑΜΕΤΡΙΚΟ ΔΙΑΓΡΑΜΜΑ

Έστω ότι ένα τυχαίο διδιάστατο δείγμα αναφοράς $(X_1^{(R)}, Y_1^{(R)})$, $(X_2^{(R)}, Y_2^{(R)})$, ..., $(X_m^{(R)}, Y_m^{(R)})$ συλλέγεται από μία εντός ελέγχου κατανομή, με συνεχή από κοινού αθροιστική συνάρτηση κατανομής (α.σ.κ.) $F_{X,Y}^{(R)}(x,y) = F(x,y)$ και αντίστοιχες περιθώριες συναρτήσεις κατανομών $F_X(x)$, $F_Y(y)$. Εάν οι μεταβλητές X τοποθετηθούν

σε αύξουσα σειρά, $X_{(1:m)}^{(R)} \le X_{(2:m)}^{(R)} \le ... \le X_{(m:m)}^{(R)}$, τότε οι μεταβλητές Y που συνδέονται με τις παραπάνω διατεταγμένες παρατηρήσεις συμβολίζονται με $Y_{[1:m]}^{(R)}, Y_{[2:m]}^{(R)}, ..., Y_{[m:m]}^{(R)}$ και ονομάζονται συμμεταβλητές (concomitants) των $X_{(1:m)}^{(R)}, X_{(2:m)}^{(R)}, ..., X_{(m:m)}^{(R)}$, αντίστοιχα.

Στη συνέχεια, συλλέγονται διαδοχικά τυχαία δείγματα ελέγχου από την υποκείμενη διεργασία, τα οποία είναι μεταξύ τους ανεξάρτητα και ανεξάρτητα από το δείγμα αναφοράς. Βάσει των δειγμάτων αυτών, μπορούμε να αποφασίσουμε εάν η διεργασία έχει μετατοπιστεί σε μία εκτός ελέγχου κατανομή $F_{X,Y}^{(T)}(x,y) = G(x,y)$ με περιθώριες κατανομές $G_X(x)$ και $G_Y(y)$. Για το σκοπό αυτό, επιλέγονται δύο συγκεκριμένα ζευγάρια από το δείγμα αναφοράς, τα οποία χρησιμοποιούνται ως όρια ελέγχου του προτεινόμενου διαγράμματος, για παράδειγμα $(X_{(a:m)}^{(R)}, Y_{[a:m]}^{(R)})$ και $(X_{(b:m)}^{(R)}, Y_{[b:m)]}^{(R)})$ με $1 \le a < b \le m$.

Αφού συλλεχθεί το δείγμα ελέγχου, υπολογίζεται η διδιάστατη συνάρτηση $(X_{(r:n)}^{(T)}, Y_{[r:n]}^{(T)})$, όπου η $X_{(r:n)}^{(T)}$ υποδηλώνει την *r*-οστή διατεταγμένη παρατήρηση της μεταβλητής X (του δείγματος ελέγχου) και η $Y_{[r:n]}^{(T)}$ συμβολίζει τη συμμεταβλητή της $X_{(r:n)}^{(T)}$ και συγκρίνεται με τα παραπάνω όρια ελέγχου. Συγκεκριμένα, η διεργασία θεωρείται ότι βρίσκεται εντός ελέγχου, εάν ισχύουν οι επόμενες δύο συνθήκες

$$X_{(a:m)}^{(R)} \le X_{(r:n)}^{(T)} \le X_{(b:m)}^{(R)} \text{ και } \min(Y_{[a:m]}^{(R)}, Y_{[b:m]}^{(R)}) \le Y_{[r:n]}^{(T)} \le \max(Y_{[a:m]}^{(R)}, X_{[b:m]}^{(R)}).$$

Αξίζει να σημειωθεί ότι όταν χρησιμοποιείται η διάμεσος του δείγματος ελέγχου, το *r* τίθεται ίσο με $(n+1)/2$, για περιττό *n* και ίσο με $n/2$ για άρτιο *n*. Σε αυτή την περίπτωση, είναι λογικό τα όρια ελέγχου να επιλεχθούν συμμετρικά, δηλαδή να θέσουμε $b = m - a + 1$.

Στη συνέχεια, θα συμβολίζουμε με $D(u,v)$ και $C(u,v)$ τα διδιάστατα copulas που συνδέονται με τις $F(x,y)$ και $G(x,y)$, αντίστοιχα. Σύμφωνα με το θεώρημα του Sklar (Sklar (1959)), οι α.σ.κ. $F(x,y)$ και $G(x,y)$ θα δίνονται από τους τύπους

$$F(x,y) = D(F_X(x), F_Y(y)) \text{ και } G(x,y) = C(G_X(x), G_Y(y)).$$

Σύμφωνα με τις παραπάνω εκφράσεις των από κοινού α.σ.κ., οι αντίστοιχες συναρτήσεις πυκνότητας (σ.π.) θα δίνονται από τους τύπους

$$f(x,y) = \frac{\partial^2 D(F_X(x), F_Y(y))}{\partial x \partial y} = d(F_X(x), F_Y(y)) \cdot f_X(x) \cdot f_Y(y)$$

$$g(x,y) = \frac{\partial^2 C(G_X(x), G_Y(y))}{\partial x \partial y} = c(G_X(x), G_Y(y)) \cdot g_X(x) \cdot g_Y(y),$$

όπου

$$d(u,v) = \frac{\partial^2 D(u,v)}{\partial u \partial v} \text{ και } c(u,v) = \frac{\partial^2 C(u,v)}{\partial u \partial v}. \quad (1)$$

Γενικά, ένα διδιάστατο copula συνδέει την α.σ.κ. με τις μονοδιάστατες περιθώριες και χρησιμοποιείται για να περιγράψει τη δομή εξάρτησης μεταξύ δύο τυχαίων μεταβλητών. Εάν οι περιθώριες συναρτήσεις είναι συνεχείς, τότε το copula είναι μοναδικό.

Στην περίπτωση που το δείγμα ελέγχου προέρχεται από μία συνεχή κατανομή $G(x,y)$, η πιθανότητα το προτεινόμενο διάγραμμα να μη δώσει σήμα εκτός ελέγχου, δίνεται από την έκφραση

$$p = p_{F,G}(m,n;a,b;r)$$

$$= P\,[X_{(a:m)}^{(R)} \le X_{(r:n)}^{(T)} \le X_{(b:m)}^{(R)} \text{ και } \min(Y_{[a:m]}^{(R)}, Y_{[b:m]}^{(R)}) \le Y_{[r:n]}^{(T)} \le \max(Y_{[a:m]}^{(R)}, Y_{[b:m]}^{(R)})], \quad (2)$$

η οποία είναι στην πραγματικότητα η χαρακτηριστική συνάρτηση κατανομής (operating characteristic function) του νέου διαγράμματος, ενώ το $1-p$ εκφράζει την πιθανότητα το διάγραμμα να δώσει σήμα ότι η διεργασία μετατοπίστηκε σε εκτός ελέγχου κατάσταση.

Στη συνέχεια, θα παρουσιάσουμε το ακόλουθο Λήμμα, το οποίο θα χρησιμοποιηθεί αργότερα για τον υπολογισμό της παραπάνω πιθανότητας p.

Λήμμα 1. Έστω (X_i, Y_i), $i = 1,2,..,n$ ανεξάρτητες τυχαίες μεταβλητές με α.σ.κ. $F(x,y)$, περιθώριες συναρτήσεις $F_X(x)$, $F_Y(y)$ και αντίστοιχο copula $C(x,y)$. Επιπλέον, η $f(x,y)$ παριστάνει την σ.π. της $F(x,y)$. Τότε η από κοινού σ.π. των τυχαίων μεταβλητών

$$U_{(r:n)} = F_X(X_{(r:n)}) \text{ και } V_{[r:n]} = F_Y(Y_{[r:n]})$$

υπολογίζεται από την έκφραση

$$f^{(C)}_{U_{(r:n)},V_{[r:n]}}(u,v) = \frac{1}{B(r,n-r+1)} u^{r-1}(1-u)^{n-r}\, c(u,v) = f_r^{(C)}(u,v)$$

για $u,v \in [0,1]$ και $1 \le r \le n$, όπου

$$B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$$

είναι η γνωστή συνάρτηση Βήτα (Beta function).

Απόδειξη. Εφόσον τα (X_i, Y_i), $i = 1,2,..,n$ είναι ανεξάρτητα και ισόνομα τυχαία ζεύγη, συνεπάγεται ότι

$$f_{Y_{[r:n]}|X_{(r:n)}}(y \mid x) = f_{Y|X}(y \mid x).$$

Επομένως,

$$f_{X_{(r:n)},Y_{[r:n]}}(x,y) = f_{Y|X}(y \mid x) \cdot f_{X_{(r:n)}}(x).$$

Παρατηρώντας ότι η σ.π. της r-οστής διατεταγμένης παρατήρησης δίνεται από τη σχέση

$$f_{X_{(r:n)}}(x) = \frac{1}{B(r,n-r+1)} [F_X(x)]^{r-1}[1-F_X(x)]^{n-r} f_X(x), x \in (-\infty,\infty),$$

προκύπτει το εξής (βλ. David (1973) και Yang (1977))

$$f_{X_{(r:n)},Y_{[r:n]}}(x,y)=\frac{1}{B(r,n-r+1)}[F_X(x)]^{r-1}[1-F_X(x)]^{n-r}f_{X,Y}(x,y),x,y\in(-\infty,\infty). \quad (3)$$

Τότε, η από κοινού σ.π. των τυχαίων μεταβλητών $U_{(r:n)}=F_X(X_{(r:n)})$ και $V_{[r:n]}=F_Y(Y_{[r:n]})$, υπολογίζεται από την έκφραση

$$f^{(C)}_{U_{(r:n)},V_{[r:n]}}(u,v)=\frac{1}{B(r,n-r+1)}u^{r-1}(1-u)^{n-r}c(u,v).$$

όπου $U_{(r:n)}$ είναι η *r*-οστή διατεταγμένη παρατήρηση και $V_{[r:n]}$ η συμμεταβλητή της και προέρχονται από τυπικές κανονικές ομοιόμορφες κατανομές, συμβ. $U,V\sim U(0,1)$. Το αποτέλεσμα είναι άμεσο παρατηρώντας ότι $F_U(u)=u,\ 0\le u\le 1$ και $f_{U,V}(u,v)=c(u,v),$ όπου το $c(u,v)$ ορίζεται στη Σχέση (1). ■

Λήμμα 2. Η από κοινού σ.π. των $(U_{(a:n)},V_{[a:n]}),(U_{(b:n)},V_{[b:n]})$ δίνεται από τον τύπο

$$f^{(D)}_{U_{(a:m)},U_{(b:m)};V_{[a:m]},V_{[b:m]}}(u_1,u_2;v_1,v_2)=$$

$$\frac{m!}{(a-1)!(m-b)!(b-a-1)!}u_1^{a-1}(1-u_2)^{m-b}(u_2-u_1)^{b-a-1}d(u_1,v_1)d(u_2,v_2)$$

για $0\le u_1\le u_2\le 1,\ 0\le v_1,v_2\le 1$ και $1\le a<b\le m$.

Απόδειξη. Η από κοινού σ.π. των τυχαίων μεταβλητών $(X_{(r_1:n)},Y_{[r_1:n]})$, $(X_{(r_2:n)},Y_{[r_2:n]})$, ..., $(X_{(r_k:n)},Y_{[r_k:n]})$ υπολογίζεται μέσω της έκφρασης

$$f_r(x_1,x_2,...,x_k;y_1,y_2,...,y_k)=\frac{n!}{(r_1-1)!(n-r_k)!}[F_X(x)]^{r_1-1}[1-F_X(x)]^{n-r_k}$$
$$\times\prod_{i=2}^{k}\frac{[F_X(x_i)-F_X(x_{i-1})]^{r_i-r_{i-1}-1}}{(r_i-r_{i-1}-1)!}\prod_{i=1}^{k}f_{X,Y}(x_i,y_i),$$

η οποία αποδείχθηκε από τους Abo-Eleneen and Nagaraja (2002) και αποτελεί γενίκευση της Σχέσης (3). Συνεπώς, η από κοινού σ.π. των $(U_{(a:n)},V_{[a:n]})$, $(U_{(b:n)},V_{[b:n]})$, δίνεται από την έκφραση

$$f^{(D)}_{U_{(a:m)},U_{(b:m)};V_{[a:m]},V_{[b:m]}}(u_1,u_2;v_1,v_2)=\frac{m!}{(a-1)!(m-b)!}u_1^{a-1}(1-u_2)^{m-b}\frac{(u_2-u_1)^{b-a-1}}{(b-a-1)!}d(u_1,v_1)d(u_2,v_2)$$

$$=\frac{m!}{(a-1)!(m-b)!(b-a-1)!}u_1^{a-1}(1-u_2)^{m-b}(u_2-u_1)^{b-a-1}d(u_1,v_1)d(u_2,v_2)$$

για $1\le a<b\le m$ και $0\le u_1\le u_2\le 1,\ 0\le v_1,v_2\le 1$. Το αποτέλεσμα του Λήμματος προκύπτει εύκολα κάνοντας χρήση των σχέσεων $F_U(u)=u,\ 0\le u\le 1$ και $f_{U,V}(u)=d(u,v)$, όπου το $d(u,v)$ ορίζεται στη Σχέση (1). ■

Στην Πρόταση 1 που ακολουθεί, δίνεται η ακριβής έκφραση της χαρακτηριστικής συνάρτησης λειτουργίας (2) του προτεινόμενου διαγράμματος.

Πρόταση 1. Η χαρακτηριστική συνάρτηση λειτουργίας του νέου διαγράμματος δίνεται από την έκφραση

$$p_{F,G}(m,n;a,b;r) =$$

$$\int_0^1\int_0^1 F_r^{(C)}(G_X(F_X^{-1}(u_1)),G_Y(F_Y^{-1}(v_1)))f^{(D)}_{U_{(a:m)};V_{[a:m]}}(u_1;v_1)du_1dv_1$$

$$-2\int_0^1\int_0^1 F_r^{(C)}(G_X(F_X^{-1}(u_1)),G_Y(F_Y^{-1}(v_1)))\left(\int_0^{v_1} f^{(D)}_{U_{(a:m)};V_{[a:m]},V_{[b:m]}}(u_1;v_1,v_2)dv_2\right)dv_1du_1$$

$$+\int_0^1\int_0^1 F_r^{(C)}(G_X(F_X^{-1}(u_2)),G_Y(F_Y^{-1}(v_2)))f^{(D)}_{U_{(b:m)};V_{[b:m]}}(u_2;v_2)dv_2du_2$$

$$-2\int_0^1\int_0^1 F_r^{(C)}(G_X(F_X^{-1}(u_2)),G_Y(F_Y^{-1}(v_2)))\left(\int_{v_2}^1 f^{(D)}_{U_{(b:m)};V_{[a:m]},V_{[b:m]}}(u_2;v_1,v_2)dv_1\right)dv_2du_2$$

$$+\int_0^1\int_0^1 F_r^{(C)}(G_X(F_X^{-1}(u_1)),G_Y(F_Y^{-1}(v_2)))f^{(D)}_{U_{(a:m)};V_{[b:m]}}(u_1;v_2)du_1dv_2$$

$$-2\int_0^1\int_0^1 F_r^{(C)}(G_X(F_X^{-1}(u_1)),G_Y(F_Y^{-1}(v_2)))\left(\int_0^{v_2} f^{(D)}_{U_{(a:m)};V_{[a:m]},V_{[b:m]}}(u_1;v_1,v_2)dv_1\right)du_1dv_2$$

$$+\int_0^1\int_0^1 F_r^{(C)}(G_X(F_X^{-1}(u_2)),G_Y(F_Y^{-1}(v_1)))f^{(D)}_{U_{(b:m)};V_{[a:m]}}(u_2,v_1)du_2dv_1$$

$$-2\int_0^1\int_0^1 F_r^{(C)}(G_X(F_X^{-1}(u_2)),G_Y(F_Y^{-1}(v_1)))\left(\int_{v_1}^1 f^{(D)}_{U_{(b:m)};V_{[a:m]},V_{[b:m]}}(u_2;v_1,v_2)dv_2\right)du_2dv_1$$

όπου

$$f^{(D)}_{U_{(k:m)};V_{[k:m]}}(u,v) = \frac{m!}{(k-1)!(m-k)!}u^{k-1}(1-v)^{m-k}\,d(u,v), \quad 0\le u,v\le 1, \quad k=a,b\,,$$

$F_r^{(C)}(u,v)$ είναι η α.σ.κ. της σ.π. που περιγράφεται στο Λήμμα 1 και $f^{(D)}_{U_{(a:m)};V_{[a:m]},V_{[b:m]}}(u_1;v_1,v_2)$, $f^{(D)}_{U_{(b:m)};V_{[a:m]},V_{[b:m]}}(u_2;v_1,v_2)$, $f^{(D)}_{U_{(a:m)};V_{[b:m]}}(u_1;v_2)$, $f^{(D)}_{U_{(b:m)};V_{[a:m]}}(u_2,v_1)$ είναι οι περιθώριες κατανομές της σ.π. του Λήμματος 2.

Απόδειξη. Γράφουμε αρχικά την πιθανότητα p της Σχέσης (2) στην ισοδύναμή της μορφή

$$P[F_X(X^{(R)}_{(a:m)}) \le F_X(X^{(T)}_{(r:n)}) \le F_X(X^{(R)}_{(b:m)}) \text{ και}$$

$$\min(F_Y(Y^{(R)}_{[a:m]}),F_Y(Y^{(R)}_{[b:m]})) \le F_Y(Y^{(T)}_{[r:n]}) \le \max(F_Y(Y^{(R)}_{[a:m]})),F_Y(Y^{(R)}_{[b:m]}))]$$

και στη συνέχεια εφαρμόζουμε τη συνεχή έκδοση του Θεωρήματος Ολικής Πιθανότητας στις τυχαίες μεταβλητές

$$U_{(a:m)} = F_X(X^{(R)}_{(a:m)})\,,\ U_{(b:m)} = F_X(X^{(R)}_{(b:m)})\,,\ V_{[a:m]} = F_Y(Y^{(R)}_{[a:m]})\,,\ V_{[b:m]} = F_Y(Y^{(R)}_{[b:m]})$$

οπότε να προκύψει η ακόλουθη έκφραση

$$p = \int_0^1\int_0^{u_2}\int_0^1\int_0^1 P[u_1 \le F_X(X^{(T)}_{(r:n)}) \le u_2 \ \text{ και } \ \min(v_1,v_2) \le F_Y(Y^{(T)}_{[r:n]}) \le \max(v_1,v_2)]$$
$$\times f^{(D)}_{U_{(a:m)},U_{(b:m)};V_{[a:m]},V_{[b:m]}}(u_1,u_2,v_1,v_2)dv_1dv_2du_1du_2\,,$$

ή ισοδύναμα

$$p=\int_0^1\int_0^{u_2}\int_0^1\int_0^{v_2} P[u_1\le F_X(X_{(r:n)}^{(T)})\le u_2 \ \text{ και } \ v_1\le F_Y(Y_{[r:n]}^{(T)})\le v_2]$$
$$\times f^{(D)}_{U_{(a:m)},U_{(b:m)};V_{[a:m]},V_{[b:m]}}(u_1,u_2,v_1,v_2)dv_1dv_2du_1du_2$$
$$+\int_0^1\int_0^{u_2}\int_0^1\int_{v_2}^1 P[u_1\le F_X(X_{(r:n)}^{(T)})\le u_2 \ \text{ και } \ v_2\le F_Y(Y_{[r:n]}^{(T)})\le v_1]$$
$$\times f^{(D)}_{U_{(a:m)},U_{(b:m)};V_{[a:m]},V_{[b:m]}}(u_1,u_2,v_1,v_2)dv_1dv_2du_1du_2$$

όπου $f^{(D)}_{U_{(a:m)},U_{(b:m)};V_{[a:m]},V_{[b:m]}}(u_1,u_2,v_1,v_2)$ είναι η από κοινού σ.π. των $U_{(a:m)}$, $U_{(b:m)}$, $V_{[a:m]}$, $V_{[b:m]}$.

Γράφοντας τις πιθανότητες που εμφανίζονται μέσα στο τετραπλό ολοκλήρωμα ως αλγεβρικό άθροισμα ποσοτήτων της μορφής

$$P[F_X(X_{(r:n)}^{(T)})\le u_i, F_Y(Y_{[r:n]}^{(T)})\le v_j]=P[X_{(r:n)}^{(T)}\le F_X^{-1}(u_i),\ Y_{[r:n]}^{(T)}\le F_Y^{-1}(v_j)]=$$
$$=P[G_X(X_{(r:n)}^{(T)})\le G_X(F_X^{-1}(u_i)),\ G_Y(Y_{[r:n]}^{(T)})\le G_X(F_Y^{-1}(v_j))]=$$
$$=P[U_{(r:n)}^{(T)}\le G_X(F_X^{-1}(u_i)),\ V_{[r:n]}^{(T)}\le G_Y(F_Y^{-1}(v_j))]=F_r^{(C)}(G_X(F_X^{-1}(u_i)),G_Y(F_Y^{-1}(v_j)))$$

για $i,j=1,2$, όπου $F_r^{(C)}(u,v)$ είναι η σ.κ. της σ.π. $f_r^{(C)}(u,v)$ που περιγράφεται στο Λήμμα 1. Στη συνέχεια, λαμβάνοντας υπόψη την παραπάνω έκφραση, προκύπτουν τα εξής ολοκληρώματα

$$I_1=+\int_0^1\int_0^{u_2}\int_0^1\int_0^{v_2} F_r^{(C)}(G_X(F_X^{-1}(u_1)),G_Y(F_Y^{-1}(v_1)))$$
$$\times f^{(D)}_{U_{(a:m)},U_{(b:m)};V_{[a:m]},V_{[b:m]}}(u_1,u_2;v_1,v_2)dv_1dv_2du_1du_2$$
$$I_2=+\int_0^1\int_0^{u_2}\int_0^1\int_0^{v_2} F_r^{(C)}(G_X(F_X^{-1}(u_2)),G_Y(F_Y^{-1}(v_2)))$$
$$\times f^{(D)}_{U_{(a:m)},U_{(b:m)};V_{[a:m]},V_{[b:m]}}(u_1,u_2;v_1,v_2)dv_1dv_2du_1du_2$$
$$I_3=-\int_0^1\int_0^{u_2}\int_0^1\int_0^{v_2} F_r^{(C)}(G_X(F_X^{-1}(u_1)),G_Y(F_Y^{-1}(v_2)))$$
$$\times f^{(D)}_{U_{(a:m)},U_{(b:m)};V_{[a:m]},V_{[b:m]}}(u_1,u_2;v_1,v_2)dv_1dv_2du_1du_2$$
$$I_4=-\int_0^1\int_0^{u_2}\int_0^1\int_0^{v_2} F_r^{(C)}(G_X(F_X^{-1}(u_2)),G_Y(F_Y^{-1}(v_1)))$$
$$\times f^{(D)}_{U_{(a:m)},U_{(b:m)};V_{[a:m]},V_{[b:m]}}(u_1,u_2;v_1,v_2)dv_1dv_2du_1du_2$$
$$I_5=+\int_0^1\int_0^{u_2}\int_0^1\int_{v_2}^1 F_r^{(C)}(G_X(F_X^{-1}(u_1)),G_Y(F_Y^{-1}(v_2)))$$
$$\times f^{(D)}_{U_{(a:m)},U_{(b:m)};V_{[a:m]},V_{[b:m]}}(u_1,u_2;v_1,v_2)dv_1dv_2du_1du_2$$
$$I_6=+\int_0^1\int_0^{u_2}\int_0^1\int_{v_2}^1 F_r^{(C)}(G_X(F_X^{-1}(u_2)),G_Y(F_Y^{-1}(v_1)))$$
$$\times f^{(D)}_{U_{(a:m)},U_{(b:m)};V_{[a:m]},V_{[b:m]}}(u_1,u_2;v_1,v_2)dv_1dv_2du_1du_2$$

$$I_7 = -\int_0^1\int_0^{u_2}\int_0^1\int_{v_2}^1 F_r^{(C)}(G_X(F_X^{-1}(u_1)),G_Y(F_Y^{-1}(v_1)))$$
$$\times f^{(D)}_{U_{(a:m)},U_{(b:m)};V_{[a:m]},V_{[b:m]}}(u_1,u_2;v_1,v_2)dv_1dv_2du_1du_2$$
$$I_8 = -\int_0^1\int_0^{u_2}\int_0^1\int_{v_2}^1 F_r^{(C)}(G_X(F_X^{-1}(u_2)),G_Y(F_Y^{-1}(v_2)))$$
$$\times f^{(D)}_{U_{(a:m)},U_{(b:m)};V_{[a:m]},V_{[b:m]}}(u_1,u_2;v_1,v_2)dv_1dv_2du_1du_2.$$

Τέλος, αλλάζοντας τη σειρά ολοκλήρωσης προκύπτει η ζητούμενη έκφραση. ■

3. ΑΡΙΘΜΗΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Στον Πίνακα 1 που ακολουθεί φαίνονται πως διαμορφώνονται οι τιμές της πιθανότητας συναγερμού, καθώς ο συντελεστής συσχέτισης μεταβάλλεται. Θα πρέπει να αναφερθεί πως, ως εντός ελέγχου κατανομή της διεργασίας έχει θεωρηθεί η διδιάστατη κανονική με περιθώριες μονοδιάστατες τυπικές κανονικές κατανομές. Στη στήλη, όπου ο συντελεστής συσχέτισης είναι ίσος με την εντός ελέγχου τιμή του $(\rho_{in} = 0.9)$, απεικονίζεται η πιθανότητα λανθασμένου συναγερμού (False Alarm Rate, *FAR*) του διαγράμματος. Παρατηρούμε ότι, καθώς ο συντελεστής συσχέτισης απομακρύνεται από την τιμή 0.9, η πιθανότητα συναγερμού αυξάνεται, γεγονός που επιβεβαιώνει ότι το προτεινόμενο διάγραμμα είναι ευαίσθητο σε μετατοπίσεις συσχέτισης. Αξίζει να σημειωθεί ότι, η συμμετρία των τιμών που παρατηρείται στα αποτελέσματα της προσομοίωσης, οφείλεται στις συναρτήσεις min και max, οι οποίες χρησιμοποιήθηκαν στον κανόνα της απόφασης.

Στη συνέχεια, θεωρούμε πως το copula που συνδέεται με την κατανομή της διεργασίας είναι στη μία περίπτωση το Gaussian και στην άλλη το Gumbel-Hougaard, τα οποία περιγράφονται αντίστοιχα από τις σχέσεις που ακολουθούν:

$$C_\rho(u,v) = \frac{1}{2\pi\sqrt{1-\rho^2}}\int_{-\infty}^{\Phi^{-1}(v)}\int_{-\infty}^{\Phi^{-1}(v)} \exp\left[\frac{-(s^2-2\rho st+t^2)}{2(1-\rho^2)}\right] dsdt \quad (\rho_1 = \rho_2 = 0.9)$$

$$C_\alpha(u,v) = \exp(-[(-\log u)^a + (-\log v)^a]^{1/a}) \quad (a_1 = a_2 = 4).$$

Πίνακας 1. *Πιθανότητες Συναγερμού για δεδομένο σχεδιασμό (Διδιάστατη Κανονική Κατανομή)*

Παράμετροι Σχεδιασμού			**Μετατόπιση Συντελεστή Συσχέτισης:** ρ_{out}									
m	***n***	**(*a*,*b*)**	**-0.1**	**-0.3**	**-0.5**	**-0.7**	**-0.9**	**0.9**	**0.7**	**0.5**	**0.3**	**0.1**
100	15	(6, 95)	0.20	0.17	0.15	0.11	0.04	**0.05**	0.11	0.15	0.18	0.20
		(8, 93)	0.24	0.22	0.20	0.14	0.07	**0.07**	0.15	0.20	0.23	0.25
		(10, 91)	0.28	0.27	0.24	0.18	0.10	**0.10**	0.18	0.24	0.27	0.29
	35	(8, 93)	0.24	0.23	0.19	0.13	0.05	**0.05**	0.14	0.20	0.23	0.24
		(10, 91)	0.28	0.26	0.24	0.18	0.08	**0.07**	0.18	0.23	0.27	0.29
		(12, 89)	0.33	0.32	0.28	0.21	0.10	**0.10**	0.21	0.27	0.30	0.32
	55	(8, 93)	0.24	0.23	0.20	0.13	0.05	**0.05**	0.14	0.19	0.23	0.24
		(10, 91)	0.28	0.26	0.23	0.17	0.07	**0.07**	0.17	0.24	0.27	0.29
		(12, 89)	0.33	0.31	0.27	0.21	0.10	**0.10**	0.21	0.27	0.30	0.33
200	35	(15, 186)	0.23	0.22	0.18	0.13	0.05	**0.05**	0.13	0.18	0.21	0.23
		(20, 181)	0.28	0.27	0.23	0.17	0.07	**0.07**	0.18	0.23	0.27	0.29
		(24, 177)	0.33	0.31	0.28	0.21	0.10	**0.10**	0.21	0.28	0.31	0.33
	55	(17, 184)	0.26	0.24	0.20	0.14	0.05	**0.05**	0.14	0.20	0.24	0.25
		(21, 180)	0.29	0.29	0.24	0.19	0.08	**0.07**	0.18	0.24	0.28	0.29
		(25, 176)	0.34	0.32	0.28	0.22	0.11	**0.10**	0.23	0.28	0.32	0.34
	75	(16, 185)	0.24	0.22	0.19	0.13	0.05	**0.05**	0.14	0.19	0.23	0.25
		(20, 181)	0.28	0.27	0.24	0.17	0.07	**0.07**	0.17	0.24	0.27	0.29
		(26, 175)	0.35	0.34	0.30	0.23	0.11	**0.10**	0.23	0.29	0.33	0.35

Πίνακας 2. *Πιθανότητες Συναγερμού για δεδομένο σχεδιασμό (Gaussian Copula)*

Παράμετροι Σχεδιασμού			**Μετατόπιση Μέσης Τιμής:** $\mu_{out} = \mu_X^{(T)} = \mu_Y^{(T)}$										
m	***n***	**(*a*,*b*)**	**-2.5**	**-2**	**-1.5**	**-1**	**-0.5**	**0**	**0.5**	**1**	**1.5**	**2**	**2.5**
100	25	(7, 94)	1	0.97	0.72	0.32	0.11	**0.05**	0.11	0.32	0.72	0.97	1
		(9, 92)	1	0.99	0.83	0.40	0.14	**0.07**	0.15	0.41	0.83	0.99	1
		(11, 90)	1	1	0.90	0.50	0.18	**0.10**	0.19	0.50	0.90	1	1
	35	(8, 93)	1	1	0.79	0.35	0.12	**0.05**	0.12	0.35	0.79	1	1
		(10, 91)	1	1	0.88	0.44	0.15	**0.07**	0.16	0.44	0.88	1	1
		(12, 89)	1	1	0.94	0.55	0.19	**0.10**	0.21	0.54	0.94	1	1
	55	(8, 93)	1	0.99	0.81	0.33	0.12	**0.05**	0.12	0.34	0.80	0.99	1
		(10, 91)	1	1	0.89	0.43	0.15	**0.07**	0.16	0.44	0.90	1	1
		(12, 89)	1	1	0.96	0.52	0.20	**0.10**	0.20	0.53	0.95	1	1
200	35	(16, 185)	1	1	0.81	0.35	0.12	**0.05**	0.12	0.36	0.82	1	1
		(19, 182)	1	1	0.88	0.42	0.15	**0.07**	0.16	0.42	0.89	1	1
		(24, 177)	1	1	0.96	0.54	0.20	**0.10**	0.21	0.54	0.96	1	1
	55	(17, 184)	1	1	0.86	0.36	0.13	**0.05**	0.12	0.36	0.86	1	1
		(20, 181)	1	1	0.92	0.43	0.15	**0.07**	0.15	0.43	0.93	1	1
		(24, 177)	1	1	0.97	0.53	0.20	**0.10**	0.19	0.52	0.97	1	1
	75	(17, 184)	1	1	0.87	0.36	0.13	**0.05**	0.13	0.36	0.87	1	1
		(20, 181)	1	1	0.94	0.43	0.15	**0.07**	0.16	0.43	0.94	1	1
		(25, 176)	1	1	0.98	0.54	0.21	**0.10**	0.20	0.54	0.99	1	1

Πίνακας 3. *Πιθανότητες Συναγερμού για δεδομένο σχεδιασμό (Gumbel-Hougaard Copula)*

Παράμετροι Σχεδιασμού			**Μετατόπιση Μέσης Τιμής:** $\mu_{out} = \mu_X^{(T)} = \mu_Y^{(T)}$										
m	***n***	**(*a*,*b*)**	**-2.5**	**-2**	**-1.5**	**-1**	**-0.5**	**0**	**0.5**	**1**	**1.5**	**2**	**2.5**
100	15	(7, 94)	1	0.96	0.72	0.36	0.13	**0.05**	0.06	0.23	0.66	0.95	1
		(10, 91)	1	0.99	0.85	0.49	0.19	**0.07**	0.11	0.39	0.81	0.99	1
		(12, 89)	1	0.99	0.90	0.56	0.23	**0.10**	0.16	0.49	0.88	0.99	1
	25	(8, 93)	1	0.98	0.80	0.37	0.14	**0.05**	0.06	0.26	0.74	0.98	1
		(11, 90)	1	1	0.90	0.51	0.19	**0.07**	0.12	0.42	0.89	1	1
		(14, 87)	1	1	0.96	0.64	0.24	**0.10**	0.18	0.58	0.96	1	1
	35	(9, 92)	1	1	0.85	0.41	0.15	**0.05**	0.07	0.30	0.81	1	1
		(11, 90)	1	1	0.92	0.50	0.18	**0.07**	0.10	0.40	0.90	1	1
		(14, 87)	1	1	0.97	0.64	0.24	**0.10**	0.16	0.58	0.96	1	1
200	35	(19, 182)	1	1	0.89	0.44	0.16	**0.05**	0.07	0.32	0.87	1	1
		(23, 178)	1	1	0.95	0.53	0.20	**0.07**	0.11	0.45	0.94	1	1
		(28, 173)	1	1	0.98	0.64	0.24	**0.10**	0.16	0.59	0.98	1	1
	55	(20, 181)	1	1	0.93	0.44	0.16	**0.05**	0.08	0.34	0.91	1	1
		(25, 176)	1	1	0.98	0.56	0.21	**0.07**	0.12	0.48	0.97	1	1
		(30, 171)	1	1	0.99	0.69	0.25	**0.10**	0.17	0.64	0.99	1	1
	75	(20, 181)	1	1	0.94	0.43	0.17	**0.05**	0.07	0.33	0.95	1	1
		(24, 177)	1	1	0.98	0.53	0.20	**0.07**	0.11	0.45	0.98	1	1
		(29, 172)	1	1	1	0.67	0.24	**0.10**	0.16	0.60	1	1	1

Και στις δύο περιπτώσεις, τα copulas έχουν μονοδιάστατες τυπικές κανονικές κατανομές. Στους Πίνακες 2 και 3, απεικονίζονται οι τιμές της πιθανότητας συναγερμού όταν η μέση τιμή της υποκείμενης διεργασίας απομακρύνεται από την εντός ελέγχου τιμή $\mu_{in} = \mu_X^{(R)} = \mu_Y^{(R)} = 0$ ως εξής $\mu_{out} = \mu_X^{(T)} = \mu_Y^{(T)} = -2.5\,(0.5)\,2.5$.

Είναι προφανές ότι το νέο διάγραμμα είναι αρκετά ευαίσθητο και στις μετατοπίσεις της μέσης τιμής της διεργασίας ανεξάρτητα από το μέγεθος του δείγματος που χρησιμοποιείται. Όπως αναφέραμε και προηγουμένως, στις στήλες με έντονη γραφή υπολογίζεται το *FAR*, το οποίο κυμαίνεται μεταξύ 5% και 10%. Συνεπώς, ανάλογα με την προκαθορισμένη τιμή του *FAR* που θέλουμε για δεδομένο σχεδιασμό του διαγράμματος, μπορούμε να κάνουμε κατάλληλη επιλογή των παραμέτρων *m*, *n*, *a* και *b*. Η βέλτιστη επιλογή της τιμής μιας παραμέτρου, μπορεί να επιτευχθεί σταθεροποιώντας τις υπόλοιπες από αυτές ή εναλλακτικά θεωρώντας έναν αποδεκτό συνδυασμό των παραπάνω παραμέτρων, ο οποίος ικανοποιεί συγκεκριμένες απαιτήσεις. Για να μειωθεί το πλήθος των παραμέτρων, έχουμε επιλέξει η *r*-οστή διατεταγμένη παρατήρηση να είναι η διάμεσος, για αυτό το λόγο έχουν χρησιμοποιηθεί παντού συμμετρικά όρια ελέγχου.

4. ΜΕΛΕΤΗ ΠΙΘΑΝΟΤΗΤΑΣ ΣΥΝΑΓΕΡΜΟΥ ΓΙΑ ΣΥΓΚΕΚΡΙΜΕΝΗ ΟΙΚΟΓΕΝΕΙΑ ΚΑΤΑΝΟΜΩΝ

Στην παράγραφο αυτή θα δοθεί ακριβής τύπος υπολογισμού της πιθανότητας συναγερμού (alarm rate), όταν $F_X = G_X$, $F_Y = G_Y$ και τόσο το δείγμα αναφοράς όσο και το δείγμα ελέγχου προέρχονται από τη διδιάστατη κατανομή των Farlie-Gumbel-Morgenstern (Morgenstern (1956), Gumbel (1958), Farlie (1960)) με από κοινού α.σ.κ.

$$F_\theta(x,y) = F_X(x)F_Y(y)[1+\theta(1-F_X(x))(1-F_X(x))],\ x,y \geq 0$$

με διαφορετικές παραμέτρους συσχέτισης. Τα copulas του δείγματος αναφοράς και του δείγματος ελέγχου, τα οποία αντιστοιχούν στην παραπάνω κατανομή, δίνονται αντίστοιχα από τις σχέσεις

$$D_{\theta_1}(u,v) = uv[1+\theta_1(1-u)(1-v)],\ -1 \leq \theta_1 \leq 1$$

$$C_{\theta_2}(u,v) = uv[1+\theta_2(1-u)(1-v)],\ -1 \leq \theta_2 \leq 1.$$

Η σ.π. των παραπάνω copulas δίνεται από τους τύπους

$$d_{\theta_1}(u,v) = 1+\theta_1(1-2u)(1-2v) = 1+\theta_1-2\theta_1 u-2\theta_1 v+4\theta_1 uv$$

$$c_{\theta_2}(u,v) = 1+\theta_2(1-2u)(1-2v) = 1+\theta_2-2\theta_2 u-2\theta_2 v+4\theta_2 uv.$$

Λαμβάνοντας υπόψη τις παραπάνω εκφράσεις, η α.σ.κ. της σ.π. του Λήμματος 1 υπολογίζεται ως εξής

$$F_r^{(C)}(u,v) = \frac{1}{B(r,n-r+1)}\int_0^u\int_0^v s^{r-1}(1-s)^{n-r}c_{\theta_2}(s,t)\,dtds$$

και χρησιμοποιώντας βασικές ιδιότητες της συνάρτησης Βήτα (beta function) και της μη-πλήρους συνάρτησης βήτα (incomplete beta function) προκύπτει η έκρφαση

$$F_r^{(C)}(u,v) = \sum_{j=r}^{n+1}[a_j v + b_j\theta_2 v(1-v)]\binom{n+1}{j}u^j(1-u)^{n+1-j},$$

όπου

$$a_j = \begin{cases}(n-r+1)/(n+1), & \text{εάν } j=r\\ 1, & \text{εάν } r<j\leq n\end{cases} \text{ και } b_j = \begin{cases}(n-r+1)/(n+1), & \text{εάν } j=r\\ (n-2r+1)/(n+1), & \text{εάν } r<j\leq n.\end{cases}$$

Σε αυτήν την περίπτωση, παρατηρώντας ότι

$$d_{\theta_1}(u,v) = 1+\theta_1\sum_{x=0}^{1}\sum_{y=0}^{1}(-1)^{x+y}u^x(1-u)^{1-x}v^y(1-v)^{1-y} = 1+\theta_1 D_1(u,v)$$

το γινόμενο $d_{\theta_1}(u_1,v_1)d_{\theta_1}(u_2,v_2)$ παίρνει τη μορφή

$$\begin{aligned}d_{\theta_1}(u_1,v_1)d_{\theta_1}(u_2,v_2) &= (1+\theta_1 D_1(u_1,v_1))(1+\theta_1 D_1(u_2,v_2))\\ &= 1+\theta_1 D_1(u_1,v_1)+\theta_1 D_1(u_2,v_2)+\theta_1^2 D_2(u_1,v_1;u_2,v_2),\end{aligned}$$

όπου

$$D_2(u_1,v_1;u_2,v_2) =$$

$$\sum_{x_1=0}^{1}\sum_{y_1=0}^{1}\sum_{x_2=0}^{1}\sum_{y_2=0}^{1}(-1)^{x_1+y_1+x_2+y_2}u_1^{x_1}(1-u_1)^{1-x_1}v_1^{y_1}(1-v_1)^{1-y_1}u_2^{x_2}(1-u_2)^{1-x_2}v_2^{y_2}(1-v_2)^{1-y_2}.$$

Συνεπώς, ο τύπος του Λήμματος 2 οδηγεί στην επόμενη έκφραση

$$f^{(D)}_{U_{(a:m)},U_{(b:m)};V_{[a:m]},V_{[b:m]}}(u_1,u_2;v_1,v_2)=\frac{m!}{(a-1)!(m-b)!(b-a-1)!}$$
$$\times[S_1(u_1,u_2)+\theta_1 S_2(u_1,u_2;v_1)+\theta_1 S_3(u_1,u_2;v_2)+\theta_1^2 S_4(u_1,u_2;v_1,v_2)].$$

Στον παραπάνω τύπο χρησιμοποιείται ο συμβολισμός

$$S_1(u_1,u_2)=u_1^{a-1}(1-u_2)^{m-b}(u_2-u_1)^{b-a-1},$$

ενώ το $S_2(u_1,u_2;v_1)$ και το $S_3(u_1,u_2;v_2)$ είναι άθροισμα τεσσάρων προσθετέων της μορφής

$$s_2(u_1,u_2;v_1)=u_1^{a-1+x_1}(1-u_1)^{1-x_1}(1-u_2)^{m-b}(u_2-u_1)^{b-a-1}v_1^{y_1}(1-v_1)^{1-y_1},$$
$$s_3(u_1,u_2;v_2)=u_1^{a-1}u_2^{x_2}(1-u_2)^{m-b+1-x_2}(u_2-u_1)^{b-a-1}v_2^{y_2}(1-v_2)^{1-y_2},$$

και το $S_4(u_1,u_2;v_2)$ γράφεται ως άθροισμα οχτώ προσθετέων της μορφής

$$s_4(u_1,u_2;v_1,v_2)=u_1^{a-1+x_1}(1-u_1)^{1-x_1}v_1^{y_1}(1-v_1)^{1-y_1}u_2^{x_2}(1-u_2)^{m-b+1-x_2}$$
$$\times(u_2-u_1)^{b-a-1}v_2^{y_2}(1-v_2)^{1-y_2}.$$

Συνεπώς, για να υπολογίσουμε τις περιθώριες συναρτήσεις της σ.π. που περιγράφεται στο Λήμμα 2, πρέπει να ολοκληρωθούν κατάλληλα οι επιμέρους προσθετέοι, κάτι το οποίο γίνεται εύκολα παρατηρώντας τα εξής

$$f^{(D)}_{U_{(a:m)};V_{[a:m]}}(u_1,v_1)=\frac{m!}{(a-1)!(m-a)!}u_1^{a-1}(1-v_1)^{m-a}(1+\theta_1 D_1(u_1,v_1))$$

$$f^{(D)}_{U_{(b:m)};V_{[b:m]}}(u_2,v_2)=\frac{m!}{(b-1)!(m-b)!}u_2^{b-1}(1-v_2)^{m-b}(1+\theta_1 D_1(u_2,v_2))$$

$$f^{(D)}_{U_{(a:m)};V_{[a:m]},V_{[b:m]}}(u_1;v_1,v_2)=\int_{u_1}^1 f^{(D)}_{U_{(a:m)},U_{(b:m)};V_{[a:m]},V_{[b:m]}}(u_1,u_2;v_1,v_2)du_2$$

όπου

$$\int_{u_1}^1 S_1(u_1,u_2)du_2=u_1^{a-1}(1-u_1)^{m-a}\sum_{k=0}^{b-a-1}(-1)^{b-a-1-k}\binom{b-a-1}{k}\frac{1}{m-a-k}$$

$$\int_{u_1}^1 s_2(u_1,u_2;v_1)du_2=u_1^{a-1+x_1}(1-u_1)^{m-a+1-x_1}v_1^{y_1}(1-v_1)^{1-y_1}\sum_{k=0}^{b-a-1}(-1)^{b-a-1-k}\binom{b-a-1}{k}\frac{1}{m-a-k}$$

$$\int_{u_1}^1 s_3(u_1,u_2;v_2)du_2=u_1^{a-1}(1-u_1)^k v_2^{y_2}(1-v_2)^{1-y_2}\sum_{k=0}^{b-a-1}(-1)^{b-a-1-k}\binom{b-a-1}{k}$$
$$\times\Big(B(x_2+1,m-x_2-a-k+1)-B_{u_1}(x_2+1,m-x_2-a-k+1)\Big)$$

$$\int_{u_1}^1 s_4(u_1,u_2;v_1,v_2)du_2=u_1^{a-1+x_1}(1-u_1)^{1-x_1+k}v_1^{y_1}(1-v_1)^{1-y_1}v_2^{y_2}(1-v_2)^{1-y_2}\sum_{k=0}^{b-a-1}(-1)^{b-a-1-k}\binom{b-a-1}{k}$$
$$\times\Big(B(x_2+1,m-x_2-a-k+1)-B_{u_1}(x_2+1,m-x_2-a-k+1)\Big).$$

Οι υπόλοιπες περιθώριες συναρτήσεις που εμφανίζονται στον τύπο του *AR*, μπορούν να υπολογισθούν όμοια. Στην πραγματικότητα, η παραπάνω διαδικασία μπορεί να εφαρμοσθεί για τον υπολογισμό του *AR* για όλα τα πολυωνυμικά (polynomial) copulas (Nelsen (2006)). Προφανώς, στην περίπτωση που ο

συντελεστής συσχέτισης είναι κοινός, η προαναφερθείσα μεθοδολογία υπολογίζει την πιθανότητα λανθασμένου συναγερμού της διεργασίας.

ABSTRACT

In the present article, we introduce a new bivariate semiparametric Shewhart-type control chart for the pair $(X_{(r:n)}, Y_{[r:n]})$, where $X_{(r:n)}$ is the *r*-th order statistic of the test sample and $Y_{[r:n]}$ denotes the concomitant of $X_{(r:n)}$. The proposed chart is quite simple, whose key advantage is its sensitivity to correlation shifts. The proposed chart is a semiparametric one, since its performance is typically affected by the dependence structure of the bivariate observations under study. However, the simulation study carried out reveals that the values of the false alarm rate do not seem to change dramatically when different copulas or distributions are used.

An expression for the operating characteristic function of the new control chart is obtained. In addition, an exact formula is provided for the Alarm Rate (*AR*) when the bivariate observations follow a specific bivariate distribution. In addition, a procedure is described for the calculation of the *AR*, when polynomial copulas are used. Finally, tables are presented for the implementation of the chart for some typical *AR* values for given designs.

ΑΝΑΦΟΡΕΣ

Abo-Eleneen, Z.A. and Nagaraja, H.N. (2002). Fisher Information in an Order Statistics and its Concomitant, *Annals of the Institute of Statistical Mathematics*, **54**, 667-680.

Balakrishnan, N., Triantafyllou, I. S. and Koutras, M. V. (2010). A Distribution-Free Control Chart Based on Order Statistics, *Communications in Statistics: Theory and Methods*, **39**, 3652-3677.

Chakraborti, S., van der Laan, P. and van de Wiel (2004). A Class of Distribution-Free Control Charts, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **53**, 443-462.

David, H.A. (1973). Concomitants of Order Statistics. *Bulletin of the International Statististical Institute,* **45**, 743-745.

Farlie, D.J.G. (1960). The Performance of Some Correlation Coefficients for a General Bivariate Distribution, *Biometrika*, **47**, 307-323.

Gumbel, E.J. (1958). Distributions à plusieurs variables dont les marges sont données (with remarks by M. Fréchet). *Comptes Rendus de l'Académie des Science*, Paris, **246**, 2717–2720.

Janacek, G.J. and Meikle, S. E. (1997). Control Charts Based on Medians, *Journal of the Royal Statistical Society: Series D*, **46**, 19–31.

Morgenstern, D. (1956). Einfache Beispiele Zweidimensionaler Verteilungen, *Mitteilungsblatt fór Mathematische Statistik*, **8**, 234-235.

Nelsen, R. B. (2006). *An Introduction to Copulas*, 2nd Ed., Springer, USA.

Sklar, A. (1959). Fonctions de Répartition à n Dimensions et Leurs Marges, *Publications de l'Institut de Statistique de l'Université de Paris*, **8**, 229–231.

Yang, S.S. (1977). General Distribution Theory of the Concomitants of Order Statistics, *The Annals of the Institute of Statistical Mathematics*, **5**, 996-1002.

Κούτρας, Μ.Β. και Σοφικίτου Ε.Μ. (2014). Ένα διδιάστατο μη παραμετρικό διαγραμμα ελέγχου που βασίζεται σε διατεταγμένες παρατηρήσεις, *Πρακτικά του 27ου Πανελληνίου Συνεδρίου Στατιστικής*, 130-141.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής (2015), σελ.207-221

ΔΙΕΡΕΥΝΗΣΗ ΧΡΗΣΗΣ ΤΟΥ ΣΥΖΕΥΓΜΕΝΟΥ ΜΟΝΤΕΛΟΥ ΑΠΕΛΕΥΘΕΡΩΣΗΣ ΤΑΣΗΣ ΣΤΟΝ ΚΟΡΙΝΘΙΑΚΟ ΚΟΛΠΟ. ΕΚΤΙΜΗΣΗ ΤΗΣ ΣΕΙΣΜΙΚΗΣ ΕΠΙΚΙΝΔΥΝΟΤΗΤΑΣ.

Ο. Μαγγίρα[1], Γ. Τσακλίδης[1], Ε. Παπαδημητρίου[2], Ε. Βότση[1]
[1]Τμήμα Μαθηματικών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
ouraniamg@gmail.com, tsaklidi@math.auth.gr , votsiire@gmail.com
[2]Τμήμα Γεωφυσικής, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
ritsa@geo.auth.gr

ΠΕΡΙΛΗΨΗ

Οι χωρο-χρονικές μεταβολές των τάσεων σε μία περιοχή αποτελούν τη σημαντικότερη συνιστώσα εκτίμησης της σεισμικής επικινδυνότητας, δεδομένου ότι συνδέονται άμεσα με τη γένεση ισχυρών σεισμών. Ιδιαίτερο ενδιαφέρον παρουσιάζουν οι αλληλεπιδράσεις γειτονικών ρηγμάτων ή ομάδων ρηγμάτων μέσω του πεδίου των τάσεων. Επειδή η αιτιοκρατική προσέγγιση της αλληλεπίδρασης δεν είναι πάντοτε εφικτή, ακολουθείται στοχαστική προσέγγιση για την εκτίμησή της. Συγκεκριμένα, για την εκτίμηση της σεισμικής επικινδυνότητας του Κορινθιακού Κόλπου, μία περιοχή με ιδιαίτερα έντονη σεισμικότητα, επιχειρείται η στοχαστική προσέγγιση στις δύο υποπεριοχές, το δυτικό και ανατολικό τμήμα, η διάκριση των οποίων βασίζεται σε σεισμοτεκτονικά κριτήρια. Μετά από ανασκόπηση του Απλού Μοντέλου Απελευθέρωσης Τάσης (AMAT-Stress Release Model), αναπτύσσεται στη συνέχεια το Συζευγμένο Μοντέλο Απελευθέρωσης Τάσης (ΣΜΑΤ-Linked Stress Release Model). Εφαρμόζεται η θεωρία των σημειακών διαδικασιών (point processes) μέσω της υπό συνθήκη συνάρτησης του θετικού ρυθμού των αφίξεων. Στην περίπτωση που εξετάζουμε, η υπό συνθήκη συνάρτηση του θετικού ρυθμού των αφίξεων, δηλαδή η συνάρτηση κινδύνου για τους ενδιάμεσους χρόνους μεταξύ των σεισμών, είναι αυτή της εκθετικής κατανομής. Από τα αποτελέσματα προκύπτει ότι το ΣΜΑΤ προσαρμόζεται στα δεδομένα και παρέχει πληροφορίες για το βαθμό αλληλεπίδρασης μεταξύ των δύο υποπεριοχών. Εναλλακτικά, θεμελιώνουμε το Συζευγμένο Μοντέλο Απελευθέρωσης Ροπής (ΣΜΑΡ), για περαιτέρω έλεγχο των αποτελεσμάτων μας.

Λέξεις-Κλειδιά: μεταφορά τάσης, συζευγμένο μοντέλο απελευθέρωσης τάσης

1. ΕΙΣΑΓΩΓΗ

Οι αιτιοκρατικές επιστημονικές προσπάθειες που χρησιμοποιούνται για την εκτίμηση της σεισμικής επικινδυνότητας βασίζονται κυρίως στη θεωρία Ελαστικής Ανάπαλσης (Elastic Rebound theory) του Reid (1910), όσο και σε άλλες σχετικές θεωρίες και έννοιες (σεισμικός κύκλος, χαρακτηριστικός σεισμός). Οι παρατηρήσεις του Reid μετά τον ισχυρό σεισμό του Σαν Φρανσίσκο το 1906 κατά μήκος του ρήγματος του Αγίου Ανδρέα οδήγησαν στο συμπέρασμα ότι οι σεισμοί είναι αποτέλεσμα ελαστικής ανάπαλσης, λόγω συσσωρευμένων ελαστικών τάσεων στα πετρώματα εκατέρωθεν της ρηξιγενούς επιφάνειας. Όταν οι συσσωρευμένες αυτές τάσεις ξεπεράσουν ένα ορισμένο επίπεδο, όταν δηλαδή υπερβούν την αντοχή των πετρωμάτων, τότε παρατηρείται ολίσθηση, απελευθερώνεται η συσσωρευμένη ενέργεια και εκδηλώνεται σεισμός. Παρόλο που το μοντέλο αυτό και τα άμεσα παραγόμενα από αυτό (τα μοντέλα πρόγνωσης χρόνου και ολίσθησης-time predictable και slip predictable model) έχουν χρησιμοποιηθεί ευρέως για μακροπρόθεσμη πρόγνωση, οι ακολουθίες των ισχυρών σεισμών είναι πιο πολύπλοκες. Κι αυτό γιατί η θεωρία Ελαστικής Ανάπαλσης υποστηρίζει ότι μετά από ένα μεγάλο σεισμό θα έπρεπε να ακολουθεί μία περίοδος αδράνειας-ηρεμίας, ενώ στην πραγματικότητα ένας μεγάλος σεισμός ενδέχεται να ακολουθείται από έντονη μετασεισμική δραστηριότητα, με αυξημένη πιθανότητα να ακολουθήσει σεισμός συγκρίσιμος σε μέγεθος. Σύμφωνα με τους Kagan & Jackson (1991) μακροπρόθεσμη ασθενής συσταδοποίηση (clustering) χαρακτηρίζει τις κύριες ακολουθίες σεισμών. Μετά από έναν ισχυρό σεισμό αυτή η συσταδοποίηση οδηγεί σε μία αύξηση του ρυθμού των μεγάλων σεισμών για αρκετές δεκαετίες.

Αναπτύσσοντας το στοχαστικό Μαρκοβιανό μοντέλο που προτάθηκε από τον Knopoff (1971) σχετικά με τις κύριες ακολουθίες των σεισμών, ο Vere-Jones (1978) πρότεινε το μοντέλο απελευθέρωσης τάσης (stress release model (SRM)), μία στοχαστική εκδοχή της θεωρίας Ελαστικής Ανάπαλσης που ενσωματώνει την αιτιοκρατική συσσώρευση της τάσης σε μία περιοχή και τη στοχαστική απελευθέρωσή της μέσω των σεισμών.

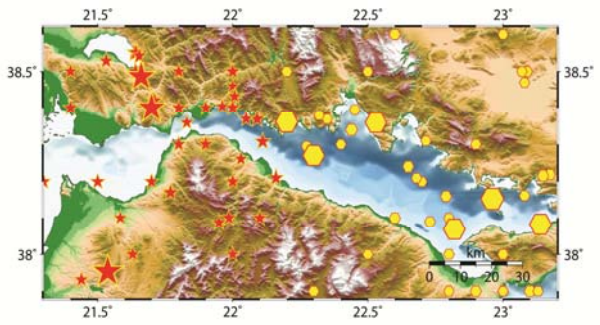
Στην παρούσα εργασία γίνεται εφαρμογή των μοντέλων ΣΜΑΤ και ΣΜΑΡ σε ισχυρούς σεισμούς, μεγέθους $M \geq 5.0$, που έγιναν στον Κορινθιακό Κόλπο από το 1911 έως το 2010, με σκοπό να εκτιμηθεί η σεισμική επικινδυνότητα της περιοχής.

2. ΔΕΔΟΜΕΝΑ ΠΑΡΑΤΗΡΗΣΗΣ

Ο Κορινθιακός Κόλπος αποτελεί μία από τις πλέον σεισμικά ενεργές δομές στην περιοχή του Αιγαίου, η οποία χωρίζει την ηπειρωτική Ελλάδα από την Πελοπόννησο (McKenzie et al, 1978), με χαρακτηριστικό έντονο εφελκυσμό με διεύθυνση Β-Ν. Ο κατάλογος των σεισμών που χρησιμοποιήσαμε προέρχεται από την τράπεζα δεδομένων του Τομέα Γεωφυσικής του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης (http://geophysics.geo.auth.gr/ss/). Η περιοχή οριοθετείται από τα γεωγραφικά μήκη 21.2^{o} Α και 23.2^{o} Α, και τα γεωγραφικά πλάτη 37.85^{o}Β και 38.65^{o} Β. Με σκοπό να χρησιμοποιήσουμε ένα πλήρες δείγμα δεδομένων, λήφθηκαν υπόψη οι σεισμοί με μέγεθος $M \geq 5.0$, οι οποίοι έγιναν στην περιοχή από το 1911 έως το 2010.

Η Εικόνα 1 δείχνει την κατανομή των σεισμών στην υπό μελέτη περιοχή, όπου με κόκκινους αστερίσκους παριστάνονται οι σεισμοί που έγιναν στο δυτικό τμήμα του Κορινθιακού Κόλπου και με κίτρινα εξάγωνα οι σεισμοί που έγιναν στο ανατολικό τμήμα του Κόλπου.

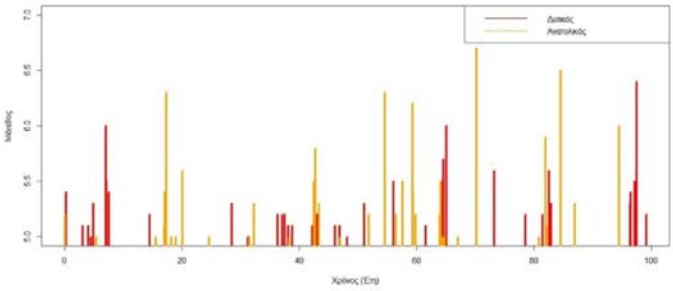
Εικόνα 1. Χάρτης σεισμικότητας του Κορινθιακού Κόλπου. Περιλαμβάνονται σεισμοί με μέγεθος $M \geq 5.0$, οι οποίοι έγιναν από το 1911 έως το 2010. Με κόκκινους αστερίσκους παριστάνονται οι σεισμοί που έγιναν στο δυτικό τμήμα του Κόλπου και με κίτρινα εξάγωνα οι σεισμοί που έγιναν στο δυτικό τμήμα του Κορινθιακού Κόλπου.



Το σύνολο δεδομένων περιλαμβάνει 100 σεισμούς, εκ των οποίων οι 53 έγιναν στο δυτικό και οι 47 στον ανατολικό Κορινθιακό Κόλπο. Ο ισχυρότερος σεισμός είχε μέγεθος $M = 6.7$ και έγινε στην περιοχή των Αλκυονίδων Νήσων το 1981.

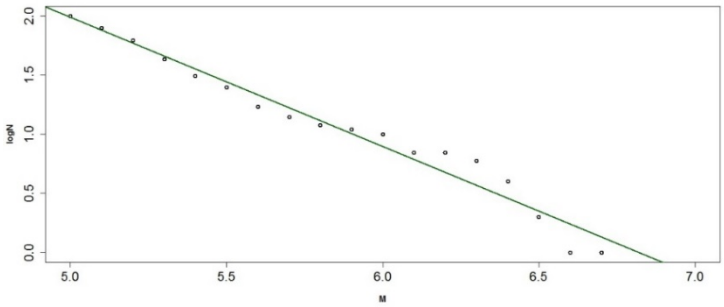
Στην Εικόνα 2 που ακολουθεί παρουσιάζεται η χρονική ακολουθία μεγεθών για το ίδιο σύνολο δεδομένων, με την ίδια χρωματική διάκριση ανάλογα με το τμήμα του Κορινθιακού Κόλπου στο οποίο έγινε ο σεισμός.

Εικόνα 2. *Χρονική ακολουθία μεγεθών για σεισμούς που έγιναν από το 1911.*



Στην Εικόνα 3 φαίνεται η κατά μέγεθος κατανομή. Στον οριζόντιο άξονα παρατίθενται τα μεγέθη και στον κατακόρυφο άξονα ο δεκαδικός λογάριθμος της αθροιστικής συχνότητας σεισμών με μέγεθος πάνω από x, x=5.0, 5.1,…,6.7. Η καλή προσαρμογή της ευθείας των ελαχίστων τετραγώνων στα δεδομένα επιβεβαιώνει την πληρότητα των δεδομένων για αυτό το χρονικό διάστημα.

Εικόνα 3. *Κατά μέγεθος κατανομή των σεισμών.*



3. ΠΕΡΙΓΡΑΦΗ ΤΩΝ ΜΟΝΤΕΛΩΝ

Τα τελευταία χρόνια η θεωρία σημειακών διαδικασιών έχει χρησιμοποιηθεί ευρέως στην εκτίμηση της σεισμικής επικινδυνότητας. Κι αυτό γιατί οι στοχαστικές σημειακές διαδικασίες θεωρούνται από τα καταλληλότερα μοντέλα για την περιγραφή φυσικών φαινομένων που εκδηλώνονται σε άτακτα και ακανόνιστα χρονικά σημεία (τυχαία ή ημιπεριοδικά), όπως είναι η γένεση των σεισμών.

Μία από τις πιο θεμελιώδεις συναρτήσεις στη θεωρία των σημειακών διαδικασιών είναι αυτή της υπό συνθήκης συνάρτησης του θετικού ρυθμού των αφίξεων (conditional intensity function). Η υπό συνθήκη συνάρτηση του θετικού ρυθμού των αφίξεων $\lambda^*(t)$ ορίζεται ως

$$\lambda^*(t) = \lambda(t \mid H_t) = \lim_{\Delta(t)\to 0} \frac{P\{N(t+\Delta(t)) - N(t) \geq 1 \mid H_t\}}{\Delta(t)},$$

όπου *N(t)* είναι το πλήθος των αφίξεων στο χρονικό διάστημα $(0,t]$ και H_t η ιστορία της διαδικασίας μέχρι τη χρονική στιγμή *t*, αλλά χωρίς τη χρονική στιγμή *t* (Daley & Vere-Jones, 2003). Έτσι, η υπό συνθήκη συνάρτηση του θετικού ρυθμού των αφίξεων παριστά την πιθανότητα να συμβεί ένα γεγονός στο απειροστό διάστημα $(t, t+\Delta(t))$, δεδομένης της ιστορίας της διαδικασίας ως τη χρονική στιγμή *t*.

Στην παρούσα μελέτη, οι παράμετροι των μοντέλων εκτιμώνται μεγιστοποιώντας τη λογαριθμική πιθανοφάνεια μέσω της υπό συνθήκη συνάρτησης του θετικού ρυθμού των αφίξεων. Αρχικά περιγράφεται το Απλό Μοντέλο Απελευθέρωσης Τάσης και στη συνέχεια το Συζευγμένο Μοντέλο Απελευθέρωσης Τάσης, όπως και το Συζευγμένο Μοντέλο Απελευθέρωσης Ροπής.

3.1. Απλό Μοντέλο Απελευθέρωσης Τάσης

Στο μονοδιάστατο μοντέλο απελευθέρωσης τάσης η μεταβλητή-κλειδί είναι το επίπεδο της τάσης (stress level) σε μία περιοχή, το οποίο καθορίζει την πιθανότητα να συμβεί ένας σεισμός (Vere-Jones and Deng, 1988). Το επίπεδο της τάσης $X(t)$ αυξάνεται συνεχώς μεταξύ δύο σεισμών, λόγω της συνεχούς τεκτονικής

φόρτισης, και απελευθερώνεται απότομα κατά τη γένεση του σεισμού. Ο τρόπος με τον οποίο απελευθερώνεται η ενέργεια (θεωρείται ότι) ορίζει μια Μαρκοβιανή διαδικασία. Η εξέλιξη της τάσης ως προς το χρόνο ακολουθεί την εξίσωση

$$X(t) = X(0) + \rho t - S(t),$$

όπου $X(0)$ είναι το αρχικό επίπεδο τάσης, ρ είναι ο σταθερός ρυθμός τεκτονικής φόρτισης (loading rate) και $S(t)$ είναι η απελευθέρωση τάσης κατά τη γένεση των σεισμών κατά την περίοδο $(0,t)$, δηλαδή, $S(t) = \sum_{t_i<t} S_i$, όπου t_i, S_i είναι ο χρόνος και η απελευθέρωση τάσης αντίστοιχα, που συνδέονται με τον *i*-στό σεισμό.

3.2. Συζευγμένο Μοντέλο Απελευθέρωσης Τάσης (ΣΜΑΤ)

Η μεταφορά τάσης και οι αλληλεπιδράσεις τάσεων μεταξύ γειτονικών περιοχών δεν μπορούν να ληφθούν υπόψη στο Απλό Μοντέλο Απελευθέρωσης Τάσης. Οι Zheng και Vere-Jones (1994) βρήκαν ότι μεγάλες γεωγραφικές περιοχές προσαρμόζονται καλύτερα στο μοντέλο απελευθέρωσης τάσης όταν διαιρεθούν σε υποπεριοχές. Παρουσίασαν περιπτώσεις που συνδέονται με κάποια μορφή δράσης από απόσταση, δηλαδή μεταφορά τάσης και αλληλεπίδραση. Οι παρατηρήσεις αυτές οδήγησαν στην τροποποίηση του απλού μοντέλου απελευθέρωσης τάσης και στη θεμελίωση του Συζευγμένου Μοντέλου Απελευθέρωσης Τάσης – ΣΜΑΤ (Linked Stress Release Model). Για να ληφθούν υπόψη αλληλεπιδράσεις μεταξύ διαφορετικών υποπεριοχών, η εξέλιξη της τάσης $X_i(t)$ στην *i*-στή υποπεριοχή μπορεί να θεωρηθεί στη μορφή

$$X_i(t) = X_i(0) + \rho_i t - \sum_j \theta_{ij} S(t,j), \qquad (1)$$

όπου $S(t,j)$ είναι η απελευθέρωση τάσης στην υποπεριοχή j στο χρονικό διάστημα $(0,t)$, ρ_i είναι ο σταθερός ρυθμός τεκτονικής φόρτισης στην υποπεριοχή *i*, και ο συντελεστής θ_{ij} αντιπροσωπεύει τη σταθερή αναλογία της πτώσης της τάσης, που από την υποπεριοχή j μεταφέρεται στην υποπεριοχή *i*.

Το ποσό της τάσης που απελευθερώνεται κατά τη διάρκεια ενός σεισμού μπορεί να υπολογιστεί από το μέγεθος του σεισμού. Σύμφωνα με τους Kanamori & Anderson (1975) το μέγεθος M θεωρείται ανάλογο του λογαρίθμου της σεισμικής ενέργειας που απελευθερώνεται κατά τη διάρκεια ενός σεισμού, σύμφωνα με τη σχέση $M = 2/3 \log E + const$. Οι Bufe & Varnes (1993) προτείνουν την αθροιστική τάση Benioff (Benioff strain) ως μέτρο της ολικής ενέργειας που απελευθερώνεται, δηλαδή

$$S(t) = \sum_{i=1}^{n(t)} E_i^{1/2},$$

όπου E_i είναι η σεισμική ενέργεια του *i*-οστού σεισμού και $n(t)$ είναι το πλήθος των σεισμών στο χρόνο *t*. Επομένως, προκύπτει ότι

$$S = 10^{0.75(M-M_{\min})}, \qquad (2)$$

όπου M είναι το μέγεθος του σεισμού και $M_{\min}$ το μέγεθος αναφοράς, δηλαδή το μικρότερο μέγεθος που περιλαμβάνεται στα δεδομένα μας.

Στο μοντέλο αυτό θέτουμε θ_{ii}=1 για κάθε *περιοχή i*, καθώς δεχόμαστε ότι όταν γίνεται σεισμός, απελευθερώνεται όλη η ενέργεια στη συγκεκριμένη περιοχή. Υποθέτουμε ότι κάθε περιοχή χαρακτηρίζεται από μία εκθετική συνάρτηση κινδύνου

$$\Psi(X_i(t)) = \exp[\mu_i + \nu_i X_i(t)], \ i=1,2,$$

με παραμέτρους που διαφέρουν ανά περιοχή, υποδεικνύοντας τις διαφορετικές τεκτονικές ιδιότητες. Ειδικότερα η παράμετρος ν_i παριστά την ευαισθησία στην αύξηση της τάσης της i υποπεριοχής. Και πάλι το σημείο-κλειδί στη στατιστική ανάλυση είναι ότι τα σεισμολογικά δεδομένα μπορούν να θεωρηθούν ως μία σημειακή διαδικασία στο πεδίο του χρόνου-τάσης με την υπό συνθήκη συνάρτηση του θετικού ρυθμού των αφίξεων (conditional intensity function) να γίνεται με τη βοήθεια της (1),

$$\lambda_i^*(t) = \Psi(X_i(t)) = \exp\left\{a_i + \nu_i\left[\rho_i t - \sum_j \theta_{ij} S(t,j)\right]\right\}, \ i=1,2, \qquad (3)$$

για κάθε περιοχή i όπου α_i (=μ_i+$\nu_i X_i(0)$), ν_i και ρ_i είναι οι παράμετροι που πρέπει να προσδιοριστούν για κάθε περιοχή i. Επιλέξαμε να παραμετροποιήσουμε στην (3) την υπό συνθήκη συνάρτηση του θετικού ρυθμού των αφίξεων επειδή γίνεται περισσότερο κατανοητή.

Μία απλούστερη παραμετροποίηση (Liu et al.,1998) επιτυγχάνεται θέτοντας $b_i=\nu_i\,\rho_i$ και $c_{ij}=\theta_{ij}/\rho_i$. Η μεταφορά τάσης μεταξύ των υποπεριοχών καθορίζεται από τον πίνακα C=(c_{ij}). Σε αυτή την περίπτωση η υπό συνθήκη συνάρτηση του θετικού ρυθμού των αφίξεων γίνεται

$$\lambda_i^*(t) = \exp\left\{a_i + b_i\left[t - \sum_j c_{ij} S(t,j)\right]\right\},$$

για κάθε υποπεριοχή i, όπου η $S(t,j)$ περιλαμβάνει όχι μόνο τα γεγονότα στην υποπεριοχή j αλλά και στις υπόλοιπες περιοχές. Επομένως τώρα πρέπει να προσδιοριστούν οι παράμετροι α_i , b_i και c_{ij} .

Για την εκτίμηση των παραμέτρων του μοντέλου, χρησιμοποιείται η μέθοδος μέγιστης πιθανοφάνειας. Στην περίπτωση του Συζευγμένου Μοντέλου Απελευθέρωσης Τάσης για δύο υποπεριοχές, η συνάρτηση λογαριθμικής πιθανοφάνειας έχει τη μορφή

$$\log L = \sum_{i=1}^{N_1(T)} \log \lambda_1^*(t_i) + \sum_{i=1}^{N_2(T)} \log \lambda_2^*(t_i) - \int_0^T (\lambda_1^*(u) + \lambda_2^*(u))du \quad ,$$

όπου $N_1(T)$, $N_2(T)$ το πλήθος σεισμών που έγιναν στις υποπεριοχές *1* και *2* αντίστοιχα στο χρονικό διάστημα $(0,T)$.

3.3. Εφαρμογή του Συζευγμένου Μοντέλου Απελευθέρωσης Τάσης στον Κορινθιακό Κόλπο

Επειδή η αναλυτική μορφή της συνάρτησης πιθανοφάνειας είναι πολύπλοκη, καταλήγουμε σε επαναληπτικές, αριθμητικές μεθόδους για τον υπολογισμό του μεγίστου. Για το λόγο αυτό κατασκευάσαμε κώδικα στη γλώσσα R. Η υπό μελέτη περιοχή (Εικόνα 1) χωρίστηκε σε δύο υποπεριοχές, το δυτικό Κορινθιακό (υποπεριοχή 1) και τον ανατολικό Κορινθιακό (υποπεριοχή 2).

Για τις εκτιμώμενες παραμέτρους υπάρχουν κάποιοι περιορισμοί, που πρέπει να ληφθούν υπόψη, έτσι ώστε να αποδίδεται το φυσικό τους νόημα. Οι παράμετροι $b_i=\nu_i\rho_i$, και $c_{ii}=1/\rho_i$ πρέπει να είναι θετικές, καθώς ο ρυθμός φόρτισης ρ_i και η ευαισθησία στην αύξηση της τάσης ν_i , είναι θετικές ποσότητες. Αντιθέτως, τα c_{12} και c_{21} μπορούν να πάρουν είτε θετικές είτε αρνητικές τιμές. Επίσης, ανάλυση γεωδαιτικών δεδομένων (Briole et al 2000, Chousianidis et al 2015) έδειξε μεταβολές στο ρυθμό εφελκυσμού μεταξύ του δυτικού (15 mm/έτος) και ανατολικού (θεωρούμενου ως κεντρικού στη βιβλιογραφία) Κορινθιακού Κόλπου (10 mm/έτος). Έτσι έχουμε ρ_1/ρ_2=3/2, οπότε c_{11}/c_{22} =2/3.

Για την εκτίμηση των παραμέτρων θα πρέπει να ελεγχθεί ένας μεγάλος αριθμός αρχικών τιμών και να παρατηρήσουμε την εξέλιξη των τιμών της συνάρτησης λογαριθμικής πιθανοφάνειας καθώς επίσης και της κλίσης της έτσι ώστε να βρεθεί το ολικό μέγιστό της. Η εκτίμηση πραγματοποιείται μέσω ενός αλγορίθμου τύπου Newton.

Στον τύπο (2) που αφορά στην πτώση τάσης περιέχεται η μεταβλητή $M_{\min}$, που παριστά το μέγεθος αναφοράς. Στην περίπτωση που θεωρήσουμε το μικρότερο επιτρεπτό μέγεθος στα δεδομένα μας να είναι το $M_{\min}$=5.0, η συνάρτηση λογαριθμικής πιθανοφάνειας λαμβάνει τη μέγιστη τιμή -167.266 για τις τιμές παραμέτρων που παρουσιάζονται στον Πίνακα 1.

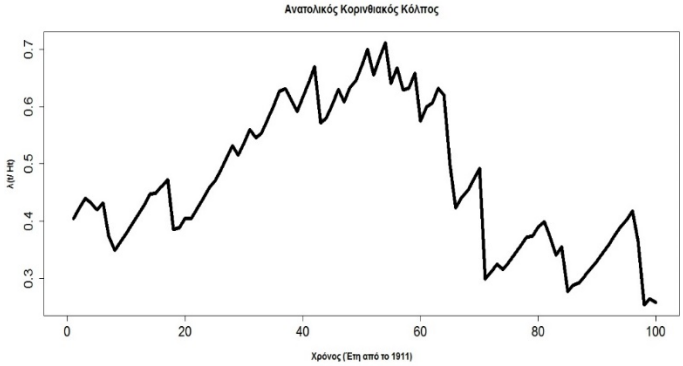
Πίνακας 1. Παράμετροι, τυπικά σφάλματα και 90% διαστήματα εμπιστοσύνης.

Παράμετρος	Εκτίμηση	Τυπικό σφάλμα	90% Διάστημα Εμπιστοσύνης
a_1	-0.8119	2.0058	(-4.1114, 2.4876)
a_2	-0.7967	2.4102	(-4.7615, 3.1680)
b_1	0.0149	0.0032	(0.0143, 0.0153)
b_2	0.0413	0.0009	(0.0397, 0.0428)
c_{11}	0.2513	0.0264	(0.2079, 0.2947)
c_{12}	0.4550	0.0130	(0.4336, 0.4764)
c_{21}	0.6112	0.0414	(0.5431, 0.6793)
c_{22}	0.3773	0.0264	(0.3339, 0.4208)

Τα τυπικά σφάλματα υπολογίστηκαν χρησιμοποιώντας τον πίνακα πληροφορίας του Fisher. Τα διαστήματα εμπιστοσύνης κατασκευάστηκαν υποθέτοντας (ασυμπτωτική) κανονικότητα. Οι παράμετροι b_1, b_2, c_{11}, c_{22}, οι οποίες πρέπει να είναι θετικές, αποδόθηκαν ως (μετασχηματίστηκαν σε) εκθετικές συναρτήσεις στη συνάρτηση λογαριθμικής πιθανοφάνειας, γεγονός που τελικά οδήγησε σε 90% διαστήματα εμπιστοσύνης με το κάτω άκρο κάθε διαστήματος να παίρνει θετική τιμή και συνεπώς όλο το διάστημα να βρίσκεται στο επιτρεπτό πεδίο τιμών. Σημειώνεται ότι χωρίς το μετασχηματισμό, όλα τα σχετικά διαστήματα εμπιστοσύνης θα περιείχαν οριακά και αρνητικές τιμές. Με χρήση των εκτιμώμενων παραμέτρων προκύπτουν οι υπό συνθήκη συναρτήσεις του θετικού ρυθμού των αφίξεων για το δυτικό και τον ανατολικό Κορινθιακό Κόλπο αντίστοιχα (Εικόνες 4,5).

Εικόνα 4. *Υπό συνθήκη συνάρτηση του θετικού ρυθμού των αφίξεων για το δυτικό Κορινθιακό Κόλπο με την εφαρμογή του ΣΜΑΤ.*

Εικόνα 5. *Υπό συνθήκη συνάρτηση του θετικού ρυθμού των αφίξεων για τον ανατολικό Κορινθιακό Κόλπο κατά την εφαρμογή του ΣΜΑΤ.*



Από τις συναρτήσεις που παρουσιάζονται στις ανωτέρω εικόνες μπορεί να γίνει εκτίμηση της σεισμικής επικινδυνότητας στις δυο υποπεριοχές. Έτσι, πχ για την υποπεριοχή 1 προκύπτει, από την Εικόνα 4, ότι η πιθανότητα γένεσης σεισμού στο χρονικό διάστημα 40-60 (1951-1971) είναι (ήταν) αυξημένη, κυμαινόμενη από 0.56

έως 0.63. Η σεισμική επικινδυνότητα με τις τιμές που παρουσιάζονται στις Εικόνες 4 και 5 μπορεί να θεωρηθεί αυξημένη και για τις δυο υποπεριοχές.

Εφόσον οι εκτιμώμενες παράμετροι ικανοποιούν τους περιορισμούς που θέσαμε, είναι δυνατός ο μετασχηματισμός τους. Στον Πίνακα 2 παρατίθενται οι μετασχηματισμένες παράμετροι για τις δύο περιοχές, από τις οποίες προκύπτουν συμπεράσματα για τους ρυθμούς τεκτονικής φόρτισης, αλλά και για τους συσχετισμούς μεταξύ των σεισμών που γίνονται σε κάθε περιοχή και το πώς αυτές αλληλοεπηρεάζονται.

Πίνακας 2. *Μετασχηματισμένες παράμετροι*

Υποπεριοχή i	a_i	ν_i	ρ_i	θ_{i1}	θ_{i2}
1 (Δυτικός Κορινθιακός)	-0.8119	0.0004	3.97	1	1.806
2(Ανατολικός Κορινθιακός)	-0.7967	0.0156	2.65	1.619	1

Ο μετασχηματισμός των παραμέτρων στον Πίνακα 2 σε συνδυασμό με τις Εικόνες 4 και 5 βοηθάει να ερμηνεύσουμε τα αποτελέσματα και τις πιθανές αλληλεπιδράσεις μεταξύ των δύο υποπεριοχών. Ελέγχοντας τις παραμέτρους θ_{12} και θ_{21} στον Πίνακα 2 διαπιστώνουμε πως οι σεισμοί που γίνονται στο δυτικό Κορινθιακό Κόλπο προκαλούν αποφόρτιση (αποδιέγερση) του ανατολικού Κορινθιακού, δηλαδή μείωση της σεισμικής δραστηριότητας, και αντίστροφα, καθώς οι παράμετροι θ_{12} και θ_{21} είναι θετικές.

Αν στον τύπο (4) που αφορά την πτώση τάσης σε μία περιοχή, μειώσουμε το μέγεθος αναφοράς θέτοντας $M_{\min}$=4.0, προκύπτουν ενδιαφέροντα αποτελέσματα από την εφαρμογή του συζευγμένου μοντέλου απελευθέρωσης τάσης. Μετά τον έλεγχο ενός μεγάλου αριθμού αρχικών τιμών, διαπιστώνουμε πως το μέγιστο της συνάρτησης λογαριθμικής πιθανοφάνειας ελάχιστα διαφοροποιείται απ' ότι για $M_{\min}$=5.0 (διαφοροποιείται μετά το 5ο δεκαδικό ψηφίο). Οι σχετικές εκτιμήσεις για τις παραμέτρους παρουσιάζονται στον Πίνακα 3.

Πίνακας 3. *Παράμετροι, τυπικά σφάλματα και 90% διαστήματα εμπιστοσύνης με δεδομένο μέγεθος αναφοράς* $M_{\min}$=4.0.

Παράμετρος	Εκτίμηση	Τυπικό σφάλμα	90% Διάστημα Εμπιστοσύνης
a_1	-0.8119	2.0058	(-4.1114, 2.4876)
a_2	-0.7967	2.4102	(-4.7615, 3.1680)
b_1	0.0149	0.0003	(0.0143, 0.0154)
b_2	0.0413	0.0009	(0.0397, 0.0428)
c_{11} (=2/3 c_{22})	0.0447	0.0176	(0.0157, 0.0737)
c_{12}	0.0809	0.0130	(0.0595, 0.1023)
c_{21}	0.1087	0.0414	(0.0406, 0.1768)
c_{22}	0.0671	0.0264	(0.0237, 0.1105)

Παρατηρούμε πως καθώς άλλαξε το μέγεθος αναφοράς, οι εκτιμήσεις για τα a_1, a_2, b_1, b_2 δεν μεταβλήθηκαν με ακρίβεια $4^{ου}$ δεκαδικού ψηφίου (μεταβάλλονται στο $7^{ο}$ δεκαδικό ψηφίο), ενώ μεταβάλλονται σημαντικά οι τιμές για τα c_{ij}. Ακόμη, μετασχηματίζοντας τις παραμέτρους (Πίνακας 4) διαπιστώνουμε πως οι σταθερές θ_{12}, θ_{21} που αποδίδουν τη μεταφορά τάσης μεταξύ των δύο υποπεριοχών και τις αλληλεπιδράσεις μεταξύ τους, παραμένουν σχεδόν ίδιες (ίδιες με ακρίβεια τριών δεκαδικών ψηφίων).

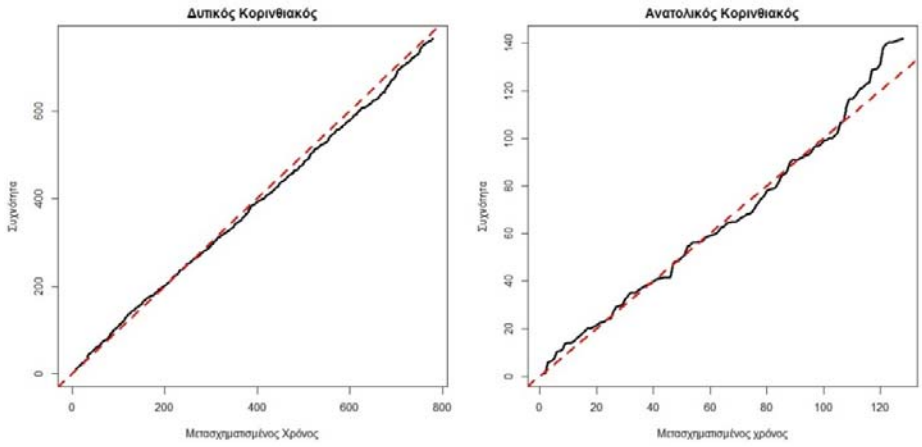
Πίνακας 4. *Μετασχηματισμένες παράμετροι στην περίπτωση που* $M_{min}=4.0$.

Υποπεριοχή *i*	a_i	ν_i	ρ_i	θ_{i1}	θ_{i2}
1 (Δυτικός Κορινθιακός)	-0.8119	$6 * 10^{-4}$	22.371	1	1.806
2 (Ανατολικός Κορινθιακός)	-0.7967	$27 * 10^{-4}$	14.903	1.619	1

Αυτό που αλλάζει σημαντικά σε αυτή την περίπτωση είναι οι ρυθμοί τεκτονικής φόρτισης για τις δύο περιοχές, ενώ σχεδόν ίδιες παραμένουν οι υπό συνθήκη συναρτήσεις του θετικού ρυθμού των αφίξεων για τις δύο περιοχές. Κατά συνέπεια, δεν αλλάζουν σημαντικά (τουλάχιστον με ακρίβεια $3^{ου}$ δεκαδικού ψηφίου) οι πιθανότητες να συμβεί ένας σεισμός δεδομένης της ιστορίας, αν αλλάξει με βάση τα παραπάνω το μέγεθος αναφοράς.

Προκειμένου, τέλος, να διαπιστωθεί αν το μοντέλο που εφαρμόσαμε προσαρμόζεται στα δεδομένα, ελέγχουμε τη σχέση της αθροιστικής συχνότητας των σεισμών, με την αθροιστική συχνότητα των σεισμών προσομοιωμένου μοντέλου. Αν η προσέγγιση είναι καλή, τότε η προκύπτουσα καμπύλη προσεγγίζει τη διχοτόμο y=x (Εικόνα 6). Από την Εικόνα 6 διαπιστώνουμε πως πρόκειται για μία αρκετά καλή προσέγγιση.

Εικόνα 6. *Προσαρμογή του ΣΜΑΤ στα δεδομένα.*



Η μοντελοποίηση του ΣΜΑΤ που εξετάσαμε, πέρα από τη στατιστική ανάλυση και τα συμπεράσματα για σεισμική αποφόρτιση που προσφέρει, μπορεί να χρησιμοποιηθεί και για εκτίμηση της σεισμικής επικινδυνότητας. Έτσι από την Εικόνα 4 που αφορά το δυτικό τμήμα του Κορινθιακού Κόλπου, διαφαίνεται από την ανοδική πορεία της καμπύλης στο τέλος του διαστήματος και το ύψος των τιμών, αυξημένη η πιθανότητα (περίπου 0.55) να συμβεί σεισμός μετά το 2010. Πράγματι την 7η και 20η Αυγούστου 2011 συνέβησαν δυο σεισμοί, μεγέθους 5.0 και 5.1 αντίστοιχα. Στο ανατολικό τμήμα του Κορινθιακού Κόλπου συνέβη σεισμός την 22α Σεπτεμβρίου 2012 μεγέθους 5.0, όμως για το χρονικό διάστημα 2010 -2012 δεν διαφαινόταν (Εικόνα 5) αυξημένη η σεισμική επικινδυνότητα. Έτσι οι χαμηλές τιμές της υπό συνθήκη συνάρτησης του θετικού ρυθμού των αφίξεων δεν διασφαλίζουν τη μη γένεση σεισμού. Προφανώς και άλλα στοιχεία πρέπει να προσμετρηθούν ώστε να επιτευχθεί αξιόπιστη εκτίμηση της μεσοπρόθεσμης σεισμικής επικινδυνότητας.

3.4. Συζευγμένο Μοντέλο Απελευθέρωσης Ροπής

Εναλλακτικά, για περαιτέρω έλεγχο των αποτελεσμάτων μας, η αλληλεπίδραση μεταξύ των δύο υποπεριοχών μπορεί να προσδιοριστεί με τη βοήθεια της σεισμικής ροπής. Στο μοντέλο αυτό, η εξέλιξη της ροπής $X_i(t)$ στην *i*-οστή περιοχή σε σχέση με το χρόνο μπορεί να γραφεί ως

$$X_i(t) = X_i(0) + m_i t - \sum_j \theta_{ij} M(t, j),$$

όπου $X_i(0)$ είναι το αρχικό επίπεδο ροπής, m_i είναι ο ρυθμός φόρτισης ροπής (moment rate), ο συντελεστής θ_{ij} μετράει τη σταθερή αναλογία πτώσης ροπής που μεταφέρεται από την υποπεριοχή j στην υποπεριοχή *i* και *M(t,j)* είναι η σωρευτική απελευθέρωση ροπής στην υποπεριοχή j στο χρονικό διάστημα $(0, t)$. Η σχέση που χρησιμοποιούμε για την απελευθέρωση ροπής είναι αυτή που προτάθηκε από τους Hanks & Kanamori (1979) σύμφωνα με την οποία η σεισμική ροπή M_0 συνδέεται με το μέγεθος ροπής του σεισμού M_W ως εξής:

$$\log M_0 = \frac{3}{2} M_W + 16 \cdot$$

Με τις ίδιες παραδοχές σε σχέση με το ΣΜΑΤ, θεωρούμε ως συνάρτηση κινδύνου την εκθετική, δηλαδή

$$\lambda_i^*(t) = \exp\left\{ a_i + \nu_i \left[m_i t - \sum_j \theta_{ij} M(t, j) \right] \right\} \cdot$$

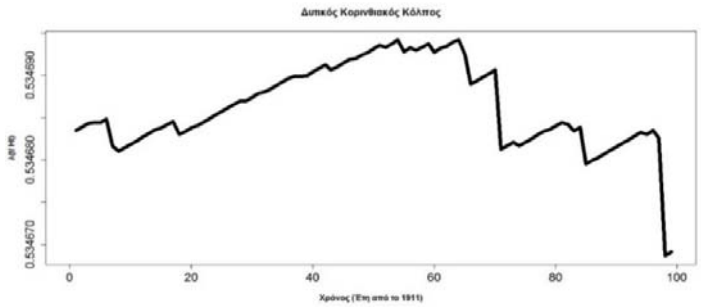
Για την εκτίμηση των παραμέτρων χρησιμοποιείται και πάλι η μέθοδος μέγιστης πιθανοφάνειας μέσω ενός αλγορίθμου τύπου Newton. Η συνάρτηση λογαριθμικής πιθανοφάνειας λαμβάνει μέγιστη τιμή log*L*=-166.57 για τις παραμέτρους που παρουσιάζονται στον Πίνακα 7.

Τα αποτελέσματα υποδεικνύουν και πάλι (εξαιτίας των θετικών τιμών των παραμέτρων θ_{12} και θ_{21}) πως οι σεισμοί που συμβαίνουν στο δυτικό τμήμα του Κόλπου μπορεί να θεωρηθεί ότι αποφορτίζουν τον ανατολικό Κορινθιακό Κόλπο και το αντίστροφο, δηλαδή προκαλείται μείωση της σεισμικής δραστηριότητας από τη μια περιοχή στην άλλη.

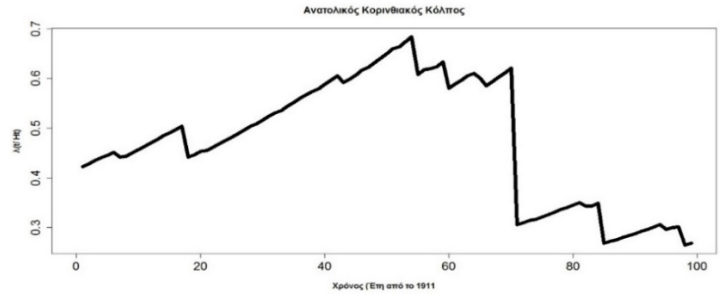
Πίνακας 5. *Εκτιμώμενες παράμετροι μετά την εφαρμογή του ΣΜΑΡ για τους σεισμούς με μεγέθη* $M \geq 5.0$ *που συνέβησαν στην περιοχή του Κορινθιακού Κόλπου.*

Υποπεριοχή i	a_i	ν_i	m_i	θ_{i1}	θ_{i2}
1(Δυτικός)	-0.6261	$7.9*10^{-34.1}$	$0.079*10^{25}$	1	0.1888
2 (Ανατολικός)	-0.8639	$1.48*10^{-27}$	10^{25}	0.7147	1

Εικόνα 7. *Υπό συνθήκη συνάρτηση του θετικού ρυθμού των αφίξεων για το δυτικό Κορινθιακό Κόλπο μετά την εφαρμογή του ΣΜΑΡ*



Εικόνα 8. *Υπό συνθήκη συνάρτηση του θετικού ρυθμού των αφίξεων για το δυτικό Κορινθιακό Κόλπο μετά την εφαρμογή του ΣΜΑΡ*



Η μορφή των συναρτήσεων στις ανωτέρω εικόνες είναι ανάλογη εκείνων του ΣΜΑΤ. Όμως οι προκύπτουσες πιθανότητες παρουσιάζουν πολύ μικρές διακυμάνσεις, κατά συνέπεια το μοντέλο ΣΜΑΡ στην παρούσα βασική του μορφή παρουσιάζεται «άκαμπτο» και με μικρή διακριτική ικανότητα στην εκτίμηση της σεισμικής επικινδυνότητας έναντι του ΣΜΑΤ.

4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Μέσω της εφαρμογής του Συζευγμένου Μοντέλου Απελευθέρωσης Τάσης (ΣΜΑΤ) θεμελιώσαμε μοντέλο για την αλληλεπίδραση μεταξύ των ρηγμάτων του δυτικού και του ανατολικού Κορινθιακού Κόλπου. Με βάση το συγκεκριμένο μοντέλο ΣΜΑΤ καταλήξαμε στο βασικό συμπέρασμα ότι η όποια αλληλεπίδραση των δύο περιοχών αφορά αποφόρτιση (αποδιέγερση) της μιας περιοχής εξαιτίας των σεισμών που συμβαίνουν στην άλλη. Τέλος παρουσιάζεται και το συζευγμένο μοντέλο απελευθέρωσης ροπής (ΣΜΑΡ), η εφαρμογή του οποίου οδηγεί σε ανάλογα συμπεράσματα όπως και το ΣΜΑΤ. Έτσι, μέσω του ΣΜΑΡ, τα αποτελέσματα υποδεικνύουν και πάλι αποφόρτιση της μιας περιοχής εξαιτίας των σεισμών που συμβαίνουν στην άλλη περιοχή.

Από την εφαρμογή του μοντέλου ΣΜΑΤ, παρατηρείται αυξημένη πιθανότητα γένεσης σεισμού μεγέθους $M \geq 5.0$ όταν η πορεία της καμπύλης της υπό συνθήκη συνάρτησης του θετικού ρυθμού των αφίξεων είναι ανοδική και συγχρόνως οι τιμές της είναι υψηλές (Εικόνες 4 και 5). Τα αποτελέσματα αυτά εμφανίζονται ελπιδοφόρα ώστε να αποτελέσουν ένα εργαλείο για την εκτίμηση της μεσοπρόθεσμης σεισμικής επικινδυνότητας και παρέχουν ένδειξη για την αποτελεσματικότητα της αξιοποίησής τους όταν συνεκτιμηθούν και με άλλα σχετικά αποτελέσματα.

ABSTRACT

The spatio-temporal stress changes in a region consist the most important component of seismic hazard assessment, since they are connected with the genesis of strong earthquakes. Particularly interesting are the interactions between adjacent faults or group of faults. The deterministic approach of the interaction is not always feasible, and yet a stochastic approach is followed. In order to study the long-term probabilistic seismic hazard of the Corinth Gulf, an area that accommodates high seismicity, the whole area is divided into two distinct subregions, namely western Gulf of Corinth and eastern Gulf of Corinth, based on their seismotectonic features. After reviewing the genesis of the simple stress release model (SSRM), we apply the linked stress release model (LSRM). Point process theory is applied by means of the conditional intensity function. In the model proposed, the conditional intensity function has the form of the exponential distribution. The results demonstrate that the LSRM fits adequately the dataset and evidence the existence of interaction between the two subregions. Alternatively, for further investigating the interactions between the two subregions, the Linked Moment Release Model is applied.

Ευχαριστίες: Η εργασία αυτή υποστηρίχθηκε από το Πρόγραμμα ΘΑΛΗΣ του Υπουργείου Παιδείας και Θρησκευμάτων και της Ευρωπαϊκής Ένωσης με τίτλο: Ενοποιημένη Προσέγγιση στην ερμηνεία της σεισμικότητας με τη συνδυασμένη χρήση Εργαστηριακών Πειραμάτων θραύσης και καινοτόμων μεθοδολογιών επεξεργασίας σεισμολογικών δεδομένων & Στατιστικής Φυσικής- Εφαρμογή στο γεωδυναμικό σύστημα του Ελληνικού Τόξου (SEISMO FEAR HELLARC) και κωδικό MIS 380208».

ΑΝΑΦΟΡΕΣ

Aki, K.(1989). Ideal probabilistic earthquake prediction, Tectonophysics, 169, 197-198.

Bebbington, M. and Harte, D. (2003).The Linked Stress Release Model for Spatio-Temporal Seismicity: Formulations, Procedures and Applications. Geophys. J. Int, **154**, 925-946.

Briole, P., Rigo, A., Lyon–Caen, H., Ruegg, JC, Papazissi, K., Mitsakaki, C., Balodimou, A., Veis, G., Hatzfeld, D., Deschamps, A. (2000). Active deformation of the Corinth rift, Greece: Results from repeated Global Positioning surveys between 1990 and 1995. J Geophys Res. doi: 105:25605–25625.

Bufe, C., Varnes, D. (1993). Predictive Modeling of the Seismic Cycle of the Greater San Francisco Bay Region, Journal of Geophys. Res., 98, 9871-9883.

Chousianitis, K., Ganas, A., Evangelidis, C.P. (2015). Strain and rotation rate patterns of mainland Greece from continuous GPS data and comparison between seismic and geodetic moment release. J Geophys Res 120. doi:10.1002/2014JB011762.

Console, R., Carluccio, R., Papadimitriou, E., Karakostas, V., (2015). Synthetic earthquake catalogs simulating seismic activity in the Corinth Gulf, Greece, fault system. J. Geophys. Res. Solid Earth, doi: 10.1002/2014JB011765.

Daley, D., Vere-Jones, D. (1988): An Introduction to the Theory of Point Processes, Springer, Berlin.

Hanks, T. C., Kanamori H. (1979). A Moment – Magnitude Scale, Journal of Geophys. Res., **84**.

Kagan, Y. Y., Jackson, D. D. (1991). Long-term earthquake clustering, Geophys. J. Int., 104, 117-133.

Kanamori, H., Anderson, D.L. (1975). Theoretical basis of some empirical relations in seismology, Bull. Seismol. Soc. Am. **65**, 5, 1073-1095.

Knopoff, L. (1971). A stochastic model for the occurrence of main-sequence eartquakes, Rev. Geophys. **9**, 1, 175-188, DOI: 10.1029/RG009i001p00175.

Lu, C., Harte, D. and Bebbington, M. (1999). A Linked Stress Release Model for Historical Japanese Earthquakes: Coupling among Major Seismic Regions, EarthPlanets Space, **51**, 907-916.

Lu, C., Vere-Jones, D. (2000). Application of Linked Stress Release Model to Historical Earthquake Data: Comparison between Two Kinds of Tectonic Seismicity, Pure appl. geophys. , **157**, 2351–2364.

Ogata, Y. (1981). On Lewis' simulation method for point processes. IEEE Transactions on Information Theory 27(1), 23-31.

Papazachos, B. C., Kiratzi, A.A., Karakostas, B., G. (1997). Towards a Homogeneous Moment- Magnitude Determination for Earthquakes in Greece and the Surrounding Area, Bull. Seismol. Soc. Am., **87,** No. 2, 474-483.

Papazachos, B.C., Karakaisis, G.F., Papadimitriou, E.E., Papaioannou, Ch.A. (1997). Time dependent seismicity in the Alpine-Himalayan Belt, Tectonophysics, **271**, 295-324, DOI: 10.1016/S0040-1951(96)00252-1.

Papazachos, B.C., Papazachou, C. (2003). The earthquakes of Greece, Ziti Publ. Co, Thessaloniki, Greece (in Greek).

Reid, H.F., (1910). The mechanism of the earthquake, in The California Earthquake of April 18, 1906, Report of the State Earthquake Investigation Commission, Vol 2, pp. 16-28, Carnegie Institute of Washongton, Washington, DC.

Vere-Jones,D. (1978). Earthquake prediction- a statistician's view, J. Phys. Earth, **26**, 129-146.

Vere-Jones, D., Deng, Y.L. (1988). A point process analysis of historical earthquakes from North China, Earthquake Res. China, **2**,165-181.

Vere-Jones, D. (1995). Forecasting earthquakes and earthquake risk, Int. J. Forecasting**, 11**, 503-538.

Votsi, I., Tsaklidis, G., Papadimitriou, E. (2011). Seismic Hazard Assessment in Central Ionian Islands Area Based on Stress Release Models, Acta Geophys. , **59,** 701-727, DOI: 10, 2478/s11600-011-0020-6.

Zheng, X., Vere-Jones, D. (1991). Applications of stress release models to earthquakes from North China, Pure appl. Geophys. , **135**, 559-576.

Zheng, X., Vere-Jones, D. (1994). Further applications of stress release models to historical earthquake data, Tectonophysics **229**, 101-121, DOI: 10.1016/0040-1951(94)90007-8.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής (2015), σελ.222-236

ΧΩΡΟΧΡΟΝΙΚΕΣ ΙΔΙΟΤΗΤΕΣ ΣΕΙΣΜΙΚΟΤΗΤΑΣ ΣΤΟ ΔΥΤΙΚΟ ΚΟΡΙΝΘΙΑΚΟ ΚΟΛΠΟ

Μ. Μεσημέρη[1], Β. Καρακώστας[1], Ε. Παπαδημητρίου[1], Γ. Τσακλίδης[2]
[1]Τμήμα Γεωλογίας, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
{mmesimer, vkarak, ritsa}@geo.auth.gr,
[2]Τμήμα Μαθηματικών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
tsaklidi@math.auth.gr

ΠΕΡΙΛΗΨΗ

Ο Κορινθιακός κόλπος αποτελεί μία από τις πιο ενεργές σεισμικά περιοχές του ελληνικού χώρου με την εκδήλωση πολλών σεισμών υπό τη μορφή σεισμικών εξάρσεων οι οποίες διακρίνονται σε μετασεισμικές ακολουθίες (MS-AS) και σε σμηνοσεισμούς (swarms). Με σκοπό την κατανόηση του τρόπου εξέλιξης της σεισμικής δραστηριότητας και της σχέσης της με τα σεισμοτεκτονικά χαρακτηριστικά της περιοχής, γίνεται διάκριση των εξάρσεων που εκδηλώθηκαν στο χρονικό διάστημα 2011 – 2014 με τη χρήση στατιστικών μεθόδων. Με βάση τον τρόπο έκλυσης της σεισμικής ροπής, εξετάζονται τα ιδιαίτερα χαρακτηριστικά της εξέλιξης των σεισμικών εξάρσεων και γίνεται προσπάθεια ταξινόμησής τους. Τα στοιχεία αυτά χρησιμοποιούνται για την κατανόηση της χωρικής και χρονικής κατανομής της σεισμικής δραστηριότητας στην περιοχή και του τρόπου αλληλεπίδρασης των σεισμών.

Λέξεις Κλειδιά: Σεισμικές ακολουθίες, μέτρα ασυμμετρίας, Κορινθιακός κόλπος.

1. ΕΙΣΑΓΩΓΗ

Οι σεισμοί συχνά γίνονται σε συγκεκριμένο χώρο και χρόνο υπό τη μορφή ομάδων. Η αύξηση του ρυθμού γένεσης των σεισμών σε μία περιοχή σε σχέση με τη σεισμικότητα υποβάθρου ορίζεται ως σεισμική έξαρση. Οι σεισμικές εξάρσεις διακρίνονται σε δύο κατηγορίες: σε μετασεισμικές ακολουθίες όπου ο κύριος σεισμός, που έχει μεγαλύτερο μέγεθος, γίνεται συνήθως στην αρχή της ακολουθίας, και σε σμηνοσεισμούς όπου δεν υπάρχει κάποιος σεισμός που να υπερέχει σε μέγεθος σε σχέση με τους υπόλοιπους της ίδιας έξαρσης (Mogi, 1963). Η διάκριση των σεισμικών εξάρσεων σε δύο τύπους είναι σημαντική για τη μελέτη της σεισμικής επικινδυνότητας μιας περιοχής διότι παρουσιάζουν διαφορετικά χωροχρονικά χαρακτηριστικά. Ο Δυτικός Κορινθιακός κόλπος έχει πληγεί από ισχυρούς σεισμούς στο παρελθόν τόσο στην ιστορική όσο και στην ενόργανη περίοδο. Στην ίδια περιοχή

παρατηρείται ιδιαίτερα έντονη σεισμικότητα με την εκδήλωση πολλών σεισμών μικρών σε μέγεθος υπό τη μορφή σεισμικών εξάρσεων.

Πρωταρχικό στόχο της εργασίας αυτής αποτελεί η αναγνώριση των σεισμικών εξάρσεων μέσα από ένα κατάλογο σεισμών με τη χρήση διαφόρων αλγορίθμων (Reasenberg, 1985; van Stiphout et al., 2012; Jacobs et al., 2013). Στη συνέχεια γίνεται ταξινόμηση των σεισμικών εξάρσεων με βάση τις τιμές του συντελεστή λοξότητας και κύρτωσης της σεισμικής ροπής σε σχέση με το χρόνο, το χρόνο εκδήλωσης του μεγαλύτερου σεισμού στη σεισμική έξαρση, t_{max} (σε ημέρες), σε σχέση με τη συνολική διάρκεια αυτής και τη διαφορά στο μέγεθος των δύο μεγαλύτερων σεισμών της σεισμικής έξαρσης (ΔΜ). Έχει δειχθεί από προηγούμενες εργασίες ότι ο συντελεστής λοξότητας για τις μετασεισμικές ακολουθίες λαμβάνει πολύ υψηλές θετικές τιμές (~30) ενώ αντίστοιχα για τους σμηνοσεισμούς οι τιμές είναι αρνητικές ή μικρές θετικές του διαστήματος (-2, 2) (Roland and McGuire, 2009). Όσον αφορά το συντελεστή κύρτωσης, οι μετασεισμικές ακολουθίες παρουσιάζουν υψηλές θετικές τιμές ενώ οι σμηνοσεισμοί τιμές κάτω από 10 (Mesimeri et al., 2013; 2014). Οι Chen and Shearer (2011) βρήκαν ότι οι μετασεισμικές ακολουθίες παρουσιάζουν $t_{max}<0.40$. Τις περισσότερες δε φορές είναι $t_{max}=0$, διότι ο μεγαλύτερος σε μέγεθος σεισμός γίνεται με την έναρξη της σεισμικής έξαρσης. Αντίθετα οι σμηνοσεισμοί έχουν $t_{max}\geq0.40$, διότι ο μεγαλύτερος σε μέγεθος σεισμός στους σμηνοσεισμούς γίνεται αργότερα. Ο Båth (1965) παρατήρησε πως η μέση διαφορά των δύο μεγαλύτερων σε μέγεθος σεισμών ήταν 1.2 για τις μετασεισμικές ακολουθίες.

Με βάση τα κριτήρια αυτά ταξινομούνται οι σεισμικές εξάρσεις και εξετάζονται ως προς τα επιμέρους χαρακτηριστικά τους με τη χρήση των μεθοδολογιών που αναφέρονται στη συνέχεια.

2. ΜΕΘΟΔΟΛΟΓΙΑ

2.1 Αλγόριθμος εύρεσης σεισμικών συγκεντρώσεων

Η αναγνώριση των σεισμικών συγκεντρώσεων από ένα κατάλογο σεισμών έγινε με την εφαρμογή σχετικού αλγόριθμου (Reasenberg, 1985), ο οποίος χρησιμοποιείται συχνά σε σεισμολογικές εργασίες. Ο αλγόριθμος εφαρμόζει τη μέθοδο των ροπών δεύτερης τάξης σε έναν κατάλογο σεισμών όπου ο j-στος σεισμός, e_j, ενός καταλόγου παριστάνεται από το διάνυσμα $\mathbf{x}_j = (x_j^k), k = 1,...,5$, όπου (x_j^1, x_j^2) οι επικεντρικές συντεταγμένες, x_j^3 το εστιακό βάθος, x_j^4 το μέγεθος και x_j^5 ο χρόνος γένεσης. Η ροπή πρώτης τάξης $m_1(\mathbf{x})$ της διαδικασίας είναι η αναμενόμενη τιμή του πλήθους των σεισμών στην κατάσταση **x**. Η ροπή δεύτερης τάξης $m_2(\mathbf{x}_1, \mathbf{x}_2)$ ορίζεται (Cox and Lewis, 1966; Kagan and Knopoff, 1976) ως

$$m_2(\mathbf{x}_1, \mathbf{x}_2) = m_1(\mathbf{x}_1) \cdot m(\mathbf{x}_2 \mid \mathbf{x}_1), \quad (1)$$

όπου $m(\mathbf{x}_2 \mid \mathbf{x}_1)$ είναι η δεσμευμένη ροπή της διαδικασίας, δηλαδή η αναμενόμενη τιμή του πλήθους των σεισμών στην κατάσταση $\mathbf{x}_2$ δεδομένου ότι συνέβη σεισμός της κατάστασης $\mathbf{x}_1$. Η δεύτερης τάξης ροπή εκφράζει συσχέτιση δύο σημείων και περιγράφει την κατανομή ζευγών σεισμών στον κατάλογο, δηλαδή $m_2(\mathbf{x}_i, \mathbf{x}_j)$ είναι η αναμενόμενη τιμή του πλήθους των σεισμών (e_i, e_j). Όταν e_i και e_j λαμβάνονται από τον ίδιο πληθυσμό τότε $m_2(\mathbf{x}_i, \mathbf{x}_j)$ είναι ο (συμμετρικός) πίνακας αυτοσυσχέτισης ενώ όταν λαμβάνονται από ξένα μεταξύ τους υποσύνολα τότε m_2 είναι μη συμμετρικός πίνακας συσχέτισης

Η εφαρμογή του αλγόριθμου έγινε σε περιβάλλον MATLAB με χρήση του λογισμικού ZMAP (Wiemer, 2001) όπως περιγράφεται από τους van Stiphout et al. (2012). Αρχικά είναι απαραίτητο να προσδιορισθούν ορισμένες παράμετροι που αφορούν τη διάρκεια της έξαρσης, το μέγεθος καθώς και το χώρο που καταλαμβάνει. Ο ελάχιστος και μέγιστος χρόνος σε ημέρες ώστε ο επόμενος σεισμός να ανήκει στη σεισμική έξαρση με συγκεκριμένη πιθανότητα (P_1) ορίζεται από τις παραμέτρους T_{min} και T_{max}. Τις τρεις αυτές παραμέτρους συνδέει η σχέση

$$\tau = \frac{-\ln(1-P_1)t}{10^{\frac{2(\Delta M-1)}{3}}} \quad , \tag{2}$$

όπου τ το διάστημα σε ημέρες για το οποίο ο επόμενος σεισμός θα ανήκει στη σεισμική έξαρση ή όχι, t ο χρόνος σε ημέρες από τη γένεση του κύριου σεισμού και $\Delta M = M_{max} - x_{meff.}$ Το x_{meff} αντιστοιχεί στο ελάχιστο μέγεθος του καταλόγου το οποίο λαμβάνεται υπόψη, ενώ με M_{max} συμβολίζεται το μέγεθος του μεγαλύτερου σεισμού στη σεισμική έξαρση. Ορίζεται επίσης ο συντελεστής x_k με τον οποίο πολλαπλασιάζεται το μέγιστο μέγεθος σε κάθε σεισμική έξαρση και η ποσότητα που προκύπτει προστίθεται στο x_{meff} για τον προσδιορισμό της τιμής x'_{meff} για κάθε σεισμική έξαρση ως εξής:

$$x'_{meff} = x_{meff} + x_k \cdot \mathrm{M}_{\max} . \tag{3}$$

Τέλος, ορίζεται η παράμετρος R_{fact} η οποία αφορά τη μέγιστη ακτίνα, σε χιλιόμετρα (km), στην οποία θεωρούμε ότι λαμβάνει χώρα μία σεισμική έξαρση (Kanamori and Anderson, 1975). Παρ' ότι βιβλιογραφικά έχουν δοθεί εμπειρικές τιμές αυτών των παραμέτρων (Schorlemmer and Gerstenberger, 2007), θεωρείται απαραίτητη η αναπροσαρμογή αυτών σύμφωνα με τις ανάγκες κάθε μελέτης και ανάλογα με τις σεισμοτεκτονικές ιδιότητες της υπό μελέτη περιοχής.

Εκτός από την παραπάνω μέθοδο εφαρμόστηκε και ο αλγόριθμος CURATE (CUmulative RATE) ο οποίος χρησιμοποιεί το ρυθμό σεισμικότητας για τη συσχέτιση των σεισμών (Jacobs et al., 2013). Κάνοντας χρήση του ρυθμού σεισμικότητας απορρίπτεται η υπόθεση ότι οι σεισμικές εξάρσεις γίνονται λόγω αλληλεπίδρασης μεταξύ των σεισμών μιας έξαρσης αλλά ενισχύεται η υπόθεση πως

όλοι οι σεισμοί που γίνονται σε μία έξαρση έχουν κοινά φυσικά αίτια. Τα αποτελέσματα από αυτήν την εφαρμογή είναι παρόμοια με αυτά του αλγορίθμου Reasenberg.

2.2 Ταξινόμηση σεισμικών συγκεντρώσεων

Προκειμένου να εφαρμοστεί η μεθοδολογία που έχει προταθεί από τους Mesimeri et al. (2014) υπολογίζονται οι συντελεστές λοξότητας (skewness)

$$\text{συντ. λοξότητας} = \frac{m_3}{s^3}. \quad (4)$$

και κύρτωσης (kurtosis)

$$\text{συντ. κύρτωσης} = \frac{m_4}{s^4}. \quad (5)$$

όπου m_3 και m_4 η σταθμισμένη ως προς τη σεισμική ροπή, (δειγματική) ροπή $3^{ης}$ και $4^{ης}$ τάξης, αντίστοιχα, και s η τυπική απόκλιση της κάθε σεισμικής συγκέντρωσης. Η σεισμική ροπή χρησιμοποιείται διότι εκφράζει την ισχύ κάθε σεισμού και σε αντίθεση με το μέγεθος μπορεί να αθροιστεί και να αποτελέσει συγκρίσιμη ποσότητα μεταξύ των σεισμών. Με τον τρόπο αυτό λαμβάνουμε υπόψη το χρόνο γένεσης του μεγαλύτερου σεισμού, δηλαδή το χρόνο κατά τον οποίο εκλύθηκε η περισσότερη ενέργεια. Συνεπώς η ταξινόμηση των σεισμικών συγκεντρώσεων γίνεται με βάση τις τιμές που προκύπτουν για τους συντελεστές λοξότητας και κύρτωσης της κάθε σεισμικής έξαρσης. Επιπλέον εξετάζεται η σχέση μεταξύ των δύο αυτών παραμέτρων.

2.3 Η παράμετρος b

Η παράμετρος b είναι η κλίση της ευθείας του διαγράμματος κατανομής της αθροιστικής συχνότητας σεισμών ως προς το μέγεθος. Για δείγματα που αφορούν μεγάλα χρονικά διαστήματα, με μεγάλο εύρος στην κατανομή των μεγεθών των σεισμών, η τιμή της παραμέτρου b προσεγγίζει τη μονάδα. Για μικρότερης κλίμακας δείγματα οι τιμές κυμαίνονται στο διάστημα [0.5, 2.5]. Μία ειδική περίπτωση αποτελούν οι σμηνοσεισμοί οι οποίοι λόγω της απουσίας ενός κύριου σεισμού (μεγαλύτερου σε μέγεθος από τους υπόλοιπους) λαμβάνουν τιμές της παραμέτρου b μεγαλύτερες από την μονάδα. Η παράμετρος b είναι σημαντική για τη Σεισμολογία διότι συνδέεται με τις σεισμοτεκτονικές ιδιότητες της περιοχής. Αυξομειώσεις της παραμέτρου b συνδέονται με τη μεταβολή του πεδίου των τάσεων σε μία περιοχή, την ετερογένεια του υλικού, τη ροή των ρευστών και την προετοιμασία για επικείμενο μεγάλο σεισμό (π.χ Scholz, 1968; 2002).

Η τιμή της παραμέτρου b εκτιμάται με τη μέθοδο της μέγιστης πιθανοφάνειας από τη σχέση (Kendall and Stuart, 1961; Aki,1965)

$$b = \frac{\log e}{\overline{M} - M_0}, \quad (6)$$

όπου $\overline{M}$ είναι το μέσο μέγεθος και M_0 είναι το ελάχιστο μέγεθος των διαθέσιμων δεδομένων. Το τυπικό σφάλμα της εκτίμησης προσεγγίζεται για μεγάλο αριθμό σεισμών, N, σε κάθε σεισμική συγκέντρωση από την ποσότητα

$$S_b = \frac{b}{\sqrt{N}} \ . \tag{7}$$

2.4 Μήκος σεισμικής ζώνης

Ο χώρος τον οποίο καταλαμβάνει μία σεισμική έξαρση υπολογίζεται από τη γυροσκοπική ακτίνα R σε km (radius of gyration) η οποία ορίζεται ως η τετραγωνική ρίζα της απόστασης κάθε σεισμού μέλους της σεισμικής έξαρσης από το βαρύκεντρό της και δίνεται από τη σχέση

$$R_g = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[(x_i - X_0)^2 + (y_i - Y_o)^2 + (z_i - Z_o)^2]}, \tag{8}$$

όπου x_i , y_i και z_i είναι το γεωγραφικό μήκος, το γεωγραφικό πλάτος και το εστιακό βάθος του i-στου σεισμού της κάθε σεισμικής έξαρσης. Με X_0, Y_o και Z_o συμβολίζονται οι συντεταγμένες του βαρύκεντρου κάθε σεισμικής έξαρσης οι οποίες υπολογίζονται από το μέσο όρο των επιμέρους στοιχείων της σεισμικής έξαρσης. Το πλήθος των σεισμών σε κάθε σεισμική έξαρση συμβολίζεται με N.

Η γυροσκοπική ακτίνα (R_g) δίνει έναν εκτιμητή της διακύμανσης στη Γκαουσιανή προσέγγιση του νέφους των σεισμών. Η διάμετρος $D_g=2R_g$ του νέφους ορίζει την περιοχή που περιέχει περίπου το 66% (για Γκαουσιανή κατανομή) των σεισμών και το διπλάσιο της διαμέτρου ορίζει περιοχή που περιέχει το 96%. Η τελευταία ποσότητα χρησιμοποιείται ως εκτιμητής του μήκους της ζώνης της σεισμικής έξαρσης $L=2D_g=4R_g$.

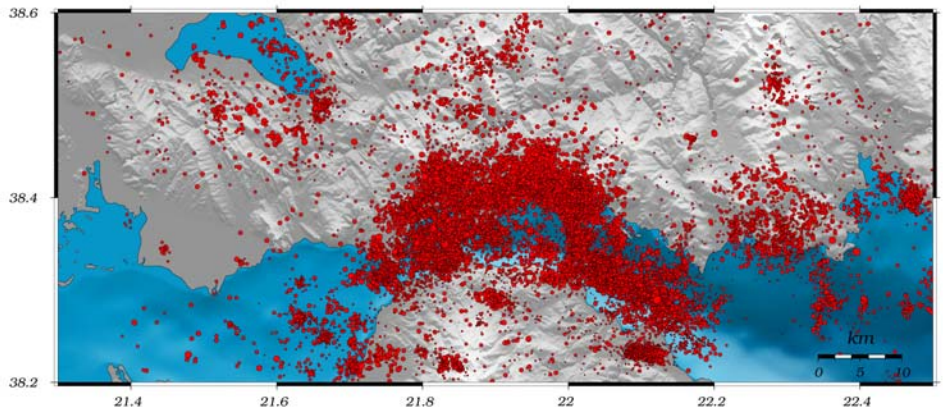
3. ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ

Η περιοχή του Δυτικού Κορινθιακού κόλπου παρουσιάζει έντονη σεισμικότητα (Σχ. 1) η οποία εκδηλώνεται πολύ συχνά υπό τη μορφή σεισμικών εξάρσεων. Κατά το παρελθόν, όπως προκύπτει τόσο από ιστορικά όσο και από ενόργανα δεδομένα, έχουν γίνει ισχυροί σεισμοί (M>6.0) οι οποίοι είχαν καταστροφικές συνέπειες για την περιοχή. Λόγω και της κατανομής του πληθυσμού, η οποία είναι μάλλον πυκνή στην περιοχή, γίνεται αντιληπτό πως η μελέτη της είναι ιδιαίτερης σημασίας.

Στην παρούσα εργασία χρησιμοποιείται ο κατάλογος σεισμών όπως προκύπτει από τις καταγραφές των μηνιαίων δελτίων σεισμών του Τομέα Γεωφυσικής του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης (http://geophysics.geo.auth.gr/ss) και του Γεωδυναμικού Ινστιτούτου του Εθνικού Αστεροσκοπείου Αθηνών (http://bbnet.gein.noa.gr/HL/) για την περιοχή του Δυτικού Κορινθιακού κόλπου και για τη χρονική περίοδο 2011-2014 με πλήθος σεισμών 15,435. Με σκοπό τον ακριβή

προσδιορισμό των εστιακών παραμέτρων και του χρόνου γένεσης των σεισμών έγινε νέος προσδιορισμός αυτών με τη χρήση των διαθέσιμων καταγραφών και εφαρμογή σύγχρονων μεθοδολογιών. Η βελτίωση της ακρίβειας στον προσδιορισμό των εστιακών συντεταγμένων των σεισμών συμβάλλει τόσο στην καλύτερη εφαρμογή του αλγορίθμου αναγνώρισης και διάκρισης των σεισμικών εξάρσεων όσο και στον ακριβέστερο προσδιορισμό κυρίως των χωρικών ιδιοτήτων των επιμέρους ομάδων.

Σχήμα 1. *Χάρτης σεισμικότητας Δυτικού Κορινθιακού για το διάστημα 2011-2014*



4. ΕΦΑΡΜΟΓΗ

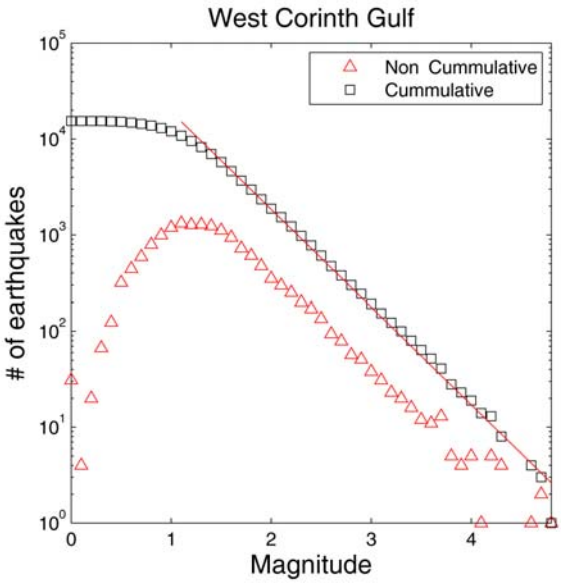
4.1 Πληρότητα δείγματος

Οι μελέτες σεισμικότητας απαιτούν τη χρήση καταλόγου σεισμών ο οποίος έχει ελεγχθεί για την πληρότητά του. Ως μέγεθος πληρότητας, M_c, ορίζεται το ελάχιστο μέγεθος πάνω από το οποίο θεωρούμε ότι έχουν καταγραφεί όλοι οι σεισμοί που έχουν γίνει στην περιοχή κατά το χρονικό διάστημα που επιλέγεται για την έρευνα (π.χ. Wiemer and Wyss, 2000). Η σωστή εκτίμηση του M_c αποτελεί κρίσιμο σημείο για την αξιοποίηση των δεδομένων του καταλόγου σεισμών. Υπερεκτίμηση του M_c οδηγεί σε μικρότερο αριθμό δεδομένων με συνέπεια την απόρριψη χρήσιμων στοιχείων, ενώ αντίθετα υποεκτίμηση οδηγεί σε λανθασμένο προσδιορισμό των σεισμικών παραμέτρων.

Για την εύρεση του μεγέθους πληρότητας (Σχ. 2) χρησιμοποιείται η μέθοδος της μέγιστης καμπυλότητας της συχνότητας των σεισμών σε σχέση με το μέγεθός τους (Wiemer and Wyss, 2000). Τα πλήρη δεδομένα αντιστοιχούν στο ευθύγραμμο τμήμα της καμπύλης, η κλίση της οποίας εκφράζεται με την παράμετρο b (Gutenberg and Richer, 1944). Για το σύνολο των δεδομένων υπολογίστηκε το μέγεθος πληρότητας M_c=1.1. Η ευθεία που προσεγγίζει τη συμπληρωματική αθροιστική (complementary cumulative) συχνότητα των σεισμών (N) για M≥1.1 είναι η $\log N = 5.29 - 1.01M$, όπου N το πλήθος των σεισμών του καταλόγου με μεγέθη μεγαλύτερα ή ίσα του M.

Το πλήθος των σεισμών N μειώθηκε με βάση τη συνθήκη M≥1.1 έναντι του αρχικού σε 10,859.

Σχήμα 2. *Κατανομή συχνότητας και συμπληρωματικής αθροιστικής συχνότητας των σεισμών σε σχέση με το μέγεθος.*



4.2 Εύρεση σεισμικών συγκεντρώσεων

Για την εφαρμογή του αλγορίθμου Reasenberg (1985) δοκιμάστηκαν διάφορες τιμές των αρχικών παραμέτρων που αφορούν στη διάρκεια και την ακτίνα αλληλεπίδρασης μιας σεισμικής έξαρσης. Παρατηρήθηκε ότι όταν επιλέγονταν μεγάλες τιμές για τη μέγιστη διάρκεια της έξαρσης ο αλγόριθμος ομαδοποιούσε στην ίδια σεισμική συγκέντρωση σεισμούς που γινόταν αργότερα στο χρόνο. Για παράδειγμα αν μία σεισμική συγκέντρωση ήταν αρκετά έντονη τις πρώτες ημέρες και μετά από μία εβδομάδα γινόταν άλλος ένας σεισμός, οι δυο συγκεντρώσεις θα ομαδοποιούνταν ως μια. Όμως αυτό οδηγεί σε λανθασμένο υπολογισμό των παραμέτρων που βασίζονται στη διάρκεια (π.χ. συντελεστές λοξότητας και κύρτωσης). Με σκοπό τη διόρθωση της χρονικής διάρκειας των σεισμικών συγκεντρώσεων έγινε χρήση νέου αλγορίθμου λαμβάνοντας υπόψη το ρυθμό σεισμικότητας σε κάθε σεισμική συγκέντρωση. Με την εφαρμογή του νέου αλγορίθμου παρατηρήθηκε απότομη διακοπή στη χρονική εξέλιξη κάθε σεισμικής συγκέντρωσης και γι' αυτό εγκαταλείφθηκε. Εναλλακτικά χρησιμοποιήθηκαν μικρές τιμές ως μέγιστο κατώφλι διάρκειας της σεισμικής έξαρσης. Όσον αφορά την ακτίνα αλληλεπίδρασης μεταξύ διαδοχικών σεισμών δοκιμάστηκε ένα εύρος τιμών στο διάστημα [1, 5] km. Βάσει των σεισμοτεκτονικών ιδιοτήτων της περιοχής και της ακρίβειας των εστιακών παραμέτρων του καταλόγου των σεισμών έγινε χρήση της ελάχιστης τιμής της ακτίνας αλληλεπίδρασης. Οι τελικές τιμές παραμέτρων που χρησιμοποιήθηκαν φαίνονται στον Πίνακα 1.

Πίνακας 1. *Τελικές τιμές παραμέτρων στον αλγόριθμο Reasenberg (1985)*

Παράμετροι	Τιμές
T_{min} (ημέρες)	1
T_{max} (ημέρες)	1
P_1	0.95
x_{meff}	1.1
x_k	0.5
R_{fact} (km)	1

Συνολικά προέκυψαν 810 σεισμικές συγκεντρώσεις με πλήθος σεισμών N≥2. Εφαρμόζοντας το κριτήριο του Mogi (1963) σύμφωνα με το οποίο ως σεισμική έξαρση ορίζεται αυτή που έχει αριθμό σεισμών N≥10 καθορίστηκαν τελικά 83 σεισμικές συγκεντρώσεις οι οποίες και μελετήθηκαν στη συνέχεια ως προς τα επιμέρους χαρακτηριστικά τους.

4.3 Ταξινόμηση σεισμικών συγκεντρώσεων

Για κάθε μία από τις 83 αναγνωρισμένες σεισμικές συγκεντρώσεις υπολογίστηκαν οι συντελεστές λοξότητας και κύρτωσης της σεισμικής ροπής ως προς το χρόνο. Από τη χαρτογράφηση των τιμών αυτών (Σχ. 3) προκύπτει πως τη μεταξύ τους σχέση την προσεγγίζει καλύτερα η εξίσωση της παραβολής. Έτσι αν K συμβολίζει την κύρτωση και S τη λοξότητα (όπως αυτά προσδιορίζονται από τους αντίστοιχους συντελεστές κύρτωσης και λοξότητας) τότε αναζητούμε σχέση μεταξύ αυτών της μορφής $K = aS^2 + bS + c + e$, όπου $a,b,c \in \mathbb{R}$ και e δηλώνει τυχαίο σφάλμα. Με χρήση της μεθόδου ελαχίστων τετραγώνων προσδιορίζονται οι εκτιμητές $\hat{a},\hat{b},\hat{c}$ και το μοντέλο $\hat{K} = \hat{a}S^2 + \hat{b}S + \hat{c}$ προκύπτει να είναι το

$$\hat{K} = 0.904S^2 + 9.406S - 3.948. \quad (9)$$

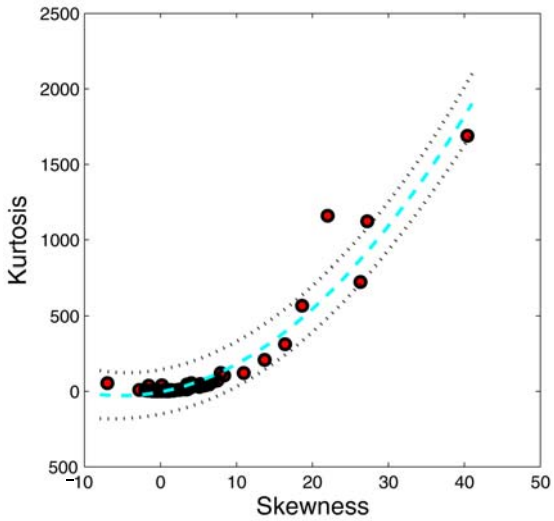
Τα 95% διαστήματα εμπιστοσύνης για τους συντελεστές a,b και c είναι

a: (0.865, 0.943), *b:* (8.742, 10.070) και *c*: (-153.120, 145.223).

Από τις 83 σεισμικές συγκεντρώσεις οι 31 χαρακτηρίζονται ως μετασεισμικές ακολουθίες και οι 52 ως σμηνοσεισμοί. Με βάση τη διαφορά στο μέγεθος των δύο μεγαλύτερων σεισμών στην κάθε σεισμική έξαρση προκύπτει για τους σμηνοσεισμούς μέση διαφορά $\Delta\overline{M} = 0.3$ ενώ για τις μετασεισμικές ακολουθίες $\Delta\overline{M} = 0.8$. Για τους σμηνοσεισμούς αυτούς έχουμε S<2 και K<10 ενώ για τις μετασεισμικές ακολουθίες S>25 και K>700. Υπάρχουν επίσης ορισμένες σεισμικές εξάρσεις στα διαστήματα 2<S<25 και 10<K<700 οι οποίες δεν διαχωρίζονται σαφώς

μεταξύ τους. Κάποιες από αυτές τις περιπτώσεις εμπίπτουν στα προαναφερθέντα κριτήρια και χαρακτηρίζονται ως μετασεισμικές ακολουθίες ενώ κάποιες άλλες παρουσιάζουν μικρή διαφορά στο μέγεθος των δύο μεγαλύτερων σεισμών με αποτέλεσμα να μην μπορούν να ταξινομηθούν. Για την ομάδα αυτή των σεισμικών εξάρσεων χρειάζεται περαιτέρω διερεύνηση για να μπορέσει να γίνει με αυστηρότητα η ταξινόμησή τους. Αναφορικά με το χρόνο γένεσης του μεγαλύτερου σεισμού στην κάθε σεισμική έξαρση βρέθηκε ότι για 79% των σεισμικών εξάρσεων που χαρακτηρίστηκαν ως σμηνοσεισμοί έχουν τιμή $t_{max} \geq 0.40$, ενώ αντίστοιχα το 65% των μετασεισμικών ακολουθιών έχουν τιμή $t_{max} < 0.40$. Υπάρχουν περιπτώσεις στους σμηνοσεισμούς όπου ο σεισμός με το μεγαλύτερο μέγεθος γίνεται νωρίτερα στη σεισμική έξαρση καθώς και μετασεισμικές ακολουθίες με το μέγιστο σε μέγεθος σεισμό να συμβαίνει αργότερα. Παρ' όλα αυτά για τις μετασεισμικές ακολουθίες προκύπτει $\overline{t}_{max} = 0.28$ ενώ για τους σμηνοσεισμούς προκύπτει $\overline{t}_{max} = 0.83$.

Σχήμα 3. *Παραβολική σχέση μεταξύ λοξότητας- κύρτωσης*

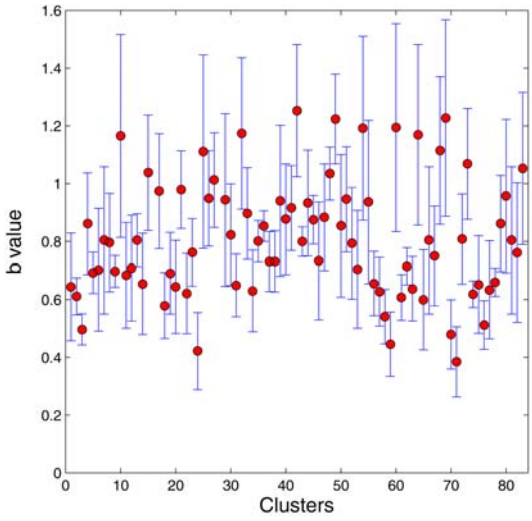


4.4 Χαρακτηριστικά σμηνοσεισμών – μετασεισμικών ακολουθιών

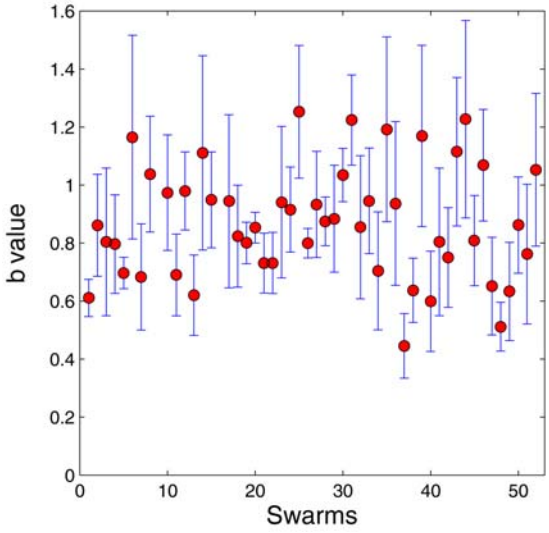
Για κάθε μια από τις αναγνωρισμένες σεισμικές συγκεντρώσεις υπολογίστηκε η τιμή της παραμέτρου b σύμφωνα με τη σχέση (6) καθώς και το τυπικό σφάλμα (σχέση (7)) και χαρτογραφήθηκαν όπως φαίνεται στο Σχήμα 4. Επιπλέον με βάση τη διάκριση που έγινε στην προηγούμενη παράγραφο χαρτογραφήθηκαν οι τιμές της παραμέτρου b για τις σεισμικές συγκεντρώσεις που ταξινομήθηκαν ως σμηνοσεισμοί (Σχ. 5) και για τις μετασεισμικές ακολουθίες (Σχ. 6). Παρατηρείται πως οι μετασεισμικές ακολουθίες παρουσιάζουν μικρότερες τιμές της παραμέτρου b με $\overline{b} = 0.71$. Αντίθετα οι τιμές για τους σμηνοσεισμούς είναι μεγαλύτερες από τις αντίστοιχες των μετασεισμικών ακολουθιών με $\overline{b} = 0.93$. Τα μεγαλύτερα σφάλματα, τα οποία παρατηρούνται κυρίως στις μετρήσεις της παραμέτρου b στους

σμηνοσεισμούς, οφείλονται στο μικρό πλήθος δεδομένων των σεισμικών συγκεντρώσεων.

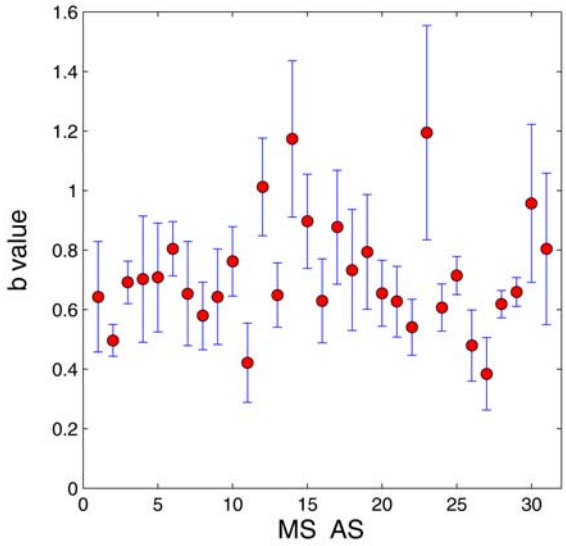
Σχήμα 4. *Τιμές της παραμέτρου b για τις αναγνωρισμένες σεισμικές εξάρσεις μαζί με τις ράβδους σφαλμάτων*



Σχήμα 5. *Τιμές της παραμέτρου b για τους σμηνοσεισμούς μαζί με τις ράβδους σφαλμάτων*



Σχήμα 6. *Τιμές της παραμέτρου b για τις μετασεισμικές ακολουθίες μαζί με τις ράβδους σφαλμάτων*



Στη συνέχεια χρησιμοποιώντας τη σχέση (8) για κάθε σεισμική συγκέντρωση βρέθηκε το μήκος της σεισμικής ζώνης το οποίο και χαρτογραφήθηκε σε σχέση με το μεγαλύτερο σε μέγεθος σεισμό (M_{max}). Τα αποτελέσματα απεικονίζονται στο Σχήμα 7 διατηρώντας την ταξινόμηση που έχει προηγηθεί.

Παρατηρείται ότι το μήκος της σεισμικής ζώνης για σεισμούς με μέγεθος M<3.0, που κατά κύριο λόγο είναι σμηνοσεισμοί, δεν παρουσιάζει κάποια εξάρτηση από το μέγιστο μέγεθος του σεισμού της σεισμικής έξαρσης. Αντίθετα, για σεισμούς με M>3.0, όπου οι σεισμικές εξάρσεις χαρακτηρίζονται κυρίως ως μετασεισμικές ακολουθίες, παρατηρείται γραμμική σχέση μεταξύ του μήκους της σεισμικής ζώνης και του μέγιστου μεγέθους. Έτσι αν L συμβολίζει το μήκος της σεισμικής ζώνης και M το μέγεθος του μεγαλύτερου σεισμού τότε αναζητούμε σχέση μεταξύ αυτών της μορφής $\log L = aM + b + e$, όπου $a, b \in \mathbb{R}$, και e δηλώνει τυχαίο σφάλμα.

Για την εύρεση της γραμμικής σχέσης εφαρμόστηκε η μέθοδος των ελαχίστων τετραγώνων και προκύπτει

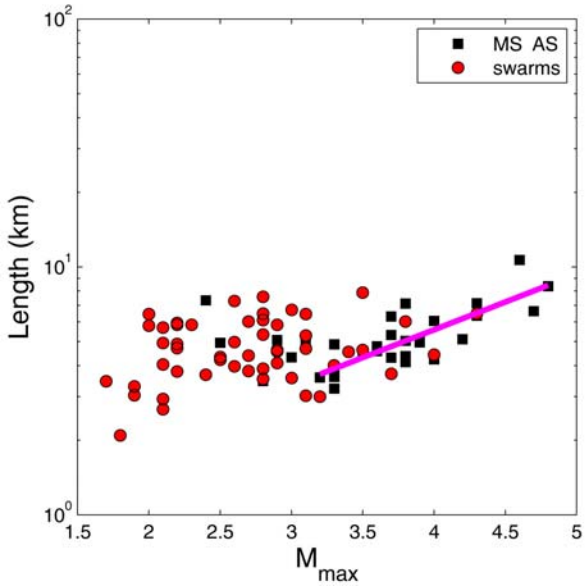
$$\log L = 0.224M - 0.151 . \qquad (10)$$

Τα 95% διαστήματα εμπιστοσύνης για τους συντελεστές *a* και *b* είναι

α: (0.219, 0.229) και *b*: (-0.287, -0.015).

Η σχέση (10) διατυπώνεται με επιφύλαξη επειδή το εύρος των μεγεθών για το οποίο φαίνεται να ισχύει είναι περιορισμένο.

Σχήμα 7. *Διάγραμμα μήκους σεισμικής ζώνης σε συνάρτηση με το μέγιστο μέγεθος σε κάθε σεισμική έξαρση*



Στον Πίνακα 2 φαίνεται το εύρος τιμών των εξεταζόμενων παραμέτρων των σεισμικών συγκεντρώσεων μετά την ταξινόμησή τους σε μετασεισμικές ακολουθίες και σε σμηνοσεισμούς.

Πίνακας 2. *Εύρος τιμών των εξεταζόμενων παραμέτρων*

Παράμετροι	Σμηνοσεισμοί	Μετασεισμικές ακολουθίες
Λοξότητα	[-6.98, 1.96]	[2.28, 40.39]
Κύρτωση	[1.52, 5.00]	[9.36, 1689.93]
Διάρκεια έξαρσης (ημέρες)	[0.16, 55.17]	[0.74, 8.79]
Μέγιστο Μέγεθος (M_{max})	[1.7, 4.3]	[2.4, 4.8]
Μήκος σεισμικής ζώνης (km)	[2.08, 7.89]	[3.22, 10.66]
Παράμετρος b	[0.44, 2.38]	[0.38, 1.19]
Σφάλμα παραμέτρου b	[0.05, 0.72]	[0.04, 0.36]

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η σχέση μεταξύ λοξότητας και κύρτωσης (όπως προσδιορίζονται μέσω των αντίστοιχων συντελεστών) αποτελεί κριτήριο για την ταξινόμηση των σεισμικών συγκεντρώσεων σε μετασεισμικές ακολουθίες και σμηνοσεισμούς. Οι τιμές που εκτιμήθηκαν για την περιοχή του Δυτικού Κορινθιακού κόλπου είναι σε συμφωνία με τα αποτελέσματα σε προηγούμενες εργασίες (Mesimeri et al., 2013; 2014). Το κατώφλι της διάκρισης για το συντελεστή λοξότητας προσεγγίζει το 2 ενώ για το συντελεστή κύρτωσης το 10. Η ταξινόμηση αυτή επαληθεύεται και από τις τιμές που προκύπτουν για την παράμετρο t_{max} η οποία παρουσιάζει μεγαλύτερες τιμές για την

πλειονότητα των σμηνοσεισμών σε αντίθεση με τις μετασεισμικές ακολουθίες. Ένα επιπλέον κριτήριο που χρησιμοποιήθηκε για την αξιολόγηση της ταξινόμησης ήταν η διαφορά των δύο μεγαλύτερων σεισμών στη σεισμική έξαρση. Για τους σμηνοσεισμούς η μέση διαφορά βρέθηκε ίση με 0.3 ενώ για τις μετασεισμικές ακολουθίες 0.8. Για την παράμετρο b, παρατηρήθηκαν μεγαλύτερες τιμές στους σμηνοσεισμούς και μικρότερες στις μετασεισμικές ακολουθίες. Τέλος, το μήκος της σεισμικής ζώνης για τις μετασεισμικές ακολουθίες εξαρτάται από το μέγεθος του μεγαλύτερου σεισμού στην ακολουθία, για M>3.0, σύμφωνα με τη σχέση (8). Σε αντίθεση, οι σμηνοσεισμοί δεν παρουσιάζουν τέτοια εξάρτηση αλλά η διασπορά των τιμών του μήκους της σεισμικής ζώνης σε συνάρτηση με το μέγιστο μέγεθος είναι μεγάλη.

Η κατάταξη των σεισμικών εξάρσεων με βάση τη λοξότητα και κύρτωση της σεισμικής ροπής ως προς το χρόνο μπορεί να συμβάλει στη μελέτη των σεισμοτεκτονικών χαρακτηριστικών μιας περιοχής και στη συσχέτιση της κανονικής σεισμικότητας με τους ισχυρούς σεισμούς. Η γνώση αυτή μπορεί να ενισχύσει σημαντικά την προσπάθεια εκτίμησης της σεισμικής επικινδυνότητας, ειδικότερα σε περιοχές με έντονη σεισμική δραστηριότητα που βρίσκονται κοντά σε αστικά κέντρα.

ABSTRACT

The western part of Corinth rift is one of the most seismically active areas in Greece with seismicity mainly manifested as earthquake sequences. Earthquake sequences are divided into two types, Mainshok - Aftershock sequences and earthquake swarms. Aiming to understand the physical process of seismicity and relate it to seismotectonic properties of the area, earthquake sequences were identified for the period 2011-2014 by using several statistical parameters. Based on the moment release history of each cluster their properties were examined and an effort was made for their classification. These properties were used for the interpretation of spatio-temporal distribution of seismic activity in the area.

Ευχαριστίες: Οι συγγραφείς ευχαριστούμε τον κριτή για τις παρατηρήσεις του οι οποίες βοήθησαν στη βελτίωση της εργασίας μας.

ΑΝΑΦΟΡΕΣ

Aki K. (1965). Maximum likelihood estimate of b in the formula logn=a-bm and its confidence limits, *Bulletin of the Earthquake Research Institute,* **43**, 237-239.

Båth M. (1965). Lateral inhomogeneities of the upper mantle. *Tectonophysics*, **2**(6), 483–514. doi:10.1016/0040-1951(65)90003-X

Chen X. and Shearer P. M. (2011). Comprehensive analysis of earthquake source spectra and swarms in the Salton Trough, California. *Journal of Geophysical Research B: Solid Earth*, **116**(B09). doi:10.1029/2011JB008263

Cox D. R. and Lewis P. W. A. (1966). *The Statistical Analysis of Series of Events*, London: Methuen.

Gutenberg B and Richter C.F. (1944). Frequency of earthquakes in California. *Bulletin of Seismological Society of America,* 34, 185-188.

Jacobs K. M., Smith E. G. C., Savage M. K. and Zhuang J. (2013). Cumulative rate analysis (CURATE): A clustering algorithm for swarm dominated catalogs. *Journal of Geophysical Research: Solid Earth*, **118**(2), 553–569. doi:10.1029/2012JB009222

Kagan Y. Y. and Knopoff L. (1976). Statisticial search for non-random features of strong earthquakes. *Physics of the Earth and Planetary Interiors*, **12**, 291–318.

Kanamori H. and Anderson D. L. (1975). Theoretical basis of some empirical relations in seismology. *Bulletin of the Seismological Society of America*, **65**(5), 1073–1095.

Kendall M. G. and Stuart A. (1961). *The Advanced Theory of Statistics, vol. 2: Inference and Relationship* (3rd ed.). New York: Hafner Publishing.

Mesimeri M., Karakostas V. and Papadimitriou, E. (2014). Σχέση κύρτωσης και λοξότητας για τον χαρακτηρισμό των σεισμικών εξάρσεων. Πρακτικά *27ου Πανελληνίου Συνεδρίου Στατιστικής* (pp. 142–154).

Mesimeri M., Papadimitriou E., Karakostas V. and Tsaklidis, G. (2013). Earthquake clusters in NW Peloponnese. In *Bulletin of the Geological Society of Greece, vol. XLVII .Proceedings of the 13th International Congress* (Vol. XLVII).

Mogi K. (1963). Some discussions on aftershocks,foreshocks and earthquake swarms - the fracture of a semi-infinite body caused by an inner stress origin and its relation to the earthquake phenomena. *Bulletin of the Earthquake Research Institute*, **40**, 831–853.

Reasenberg P. (1985). Second-order moment of central California seismicity, 1969–1982. *Journal of Geophysical Research*, **90**(B7), 5479 - 5495. doi: 10.1029/JB090iB07p05479

Roland E. and McGuire J. J. (2009). Earthquake swarms on transform faults, *Geophysical Journal International*, **178**(3), 1677–1690.

Scholz C. (2002). *The Mechanics of Earthquakes and Faulting* (Vol. 2nd), Cambridge.

Scholz C. H. (1968). The frequency - magnitude relation of microfracturing in rock and its relation to earthquakes. *Bulletin of Seismological Society of America*, **58**(1), 399–415.

Schorlemmer D. and Gerstenberger M. C. (2007). RELM testing center. *Seismological Research Letters*, **78**(1), 30–36.

Van Stiphout T., Zhuang J. and Marsan D. (2012). Seismicity declustering. *Community Online Resource for Statistical Seismicity Analysis*. doi:10.5078/corssa-52382934

Wiemer S. (2001). A Software Package to Analyze Seismicity: ZMAP. *Seismological Research Letters*, **72**(3), 373–382. doi:10.1785/gssrl.72.3.373

Wiemer S. and Wyss M. (2000). Minimum magnitude of completeness in earthquake catalogs: Examples from Alaska, the Western United States, and Japan. *Bulletin of the Seismological Society of America*, **90**(4), 859–869. doi:10.1785/0119990114

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής (2015), σελ.237-251

Η ΧΡΗΣΗ ΕΙΔΙΚΩΝ ΣΤΑΘΜΙΣΕΩΝ ΣΤΙΣ ΣΥΝΙΣΤΩΣΕΣ ΕΝΟΣ ΣΥΝΘΕΤΟΥ ΔΕΙΚΤΗ ΥΓΕΙΑΣ ΑΥΞΑΝΕΙ ΤΗ ΔΙΑΓΝΩΣΤΙΚΗ ΤΟΥ ΙΚΑΝΟΤΗΤΑ

Φ.Γ. Μπερσίμης[1], Μ. Βαμβακάρη[1], Δ.Β. Παναγιωτάκος[2]
[1] Τμήμα Πληροφορικής & Τηλεματικής, Χαροκόπειο Πανεπιστήμιο
[2] Τμήμα Επιστήμης Διαιτολογίας - Διατροφής, Χαροκόπειο Πανεπιστήμιο
{fbersim, mvamv, dbpanag}@hua.gr

ΠΕΡΙΛΗΨΗ

Ένας σύνθετος δείκτης υγείας Τ δημιουργείται συνήθως από το απλό άθροισμα διακριτών ή συνεχών συνιστωσών μεταβλητών X_i, i=1,2,…m και έχει ως στόχο την αποτίμηση ενός κλινικού ή βιοχημικού χαρακτηριστικού Υ ενός ατόμου με απώτερο σκοπό την έγκαιρη διάγνωση κάποιας πιθανής νόσου. Η παρούσα εργασία έχει ως στόχο να αξιολογήσει αν η χρήση ειδικών βαρών (π.χ. σχετικοί λόγοι πιθανοτήτων Λογιστικής Παλινδρόμησης, συντελεστές Διαχωριστικής Ανάλυσης) στις συνεχείς συνιστώσες μεταβλητές ενός σύνθετου δείκτη υγείας βελτιώνει την διαγνωστική ικανότητα αυτού, κάνοντας χρήση της ευαισθησίας, της ειδικότητας και της καμπύλης ROC, με χρήση προσομοιωμένων δεδομένων. Για το σκοπό αυτό, κατασκευάστηκαν οκτώ σύνθετοι δείκτες υγείας που παράγονται από το άθροισμα πέντε συνιστωσών τυχαίων μεταβλητών με διαφορετική μέθοδο στάθμισης έκαστος και εξετάστηκε η ευαισθησία και η ειδικότητα τους. Τα αποτελέσματα αυτής της εργασίας προτείνουν τη χρήση ειδικών σταθμίσεων για την κατασκευή σύνθετων δεικτών υγείας προκειμένου να αυξηθεί η ευαισθησία, η ειδικότητα και συνολικά το εμβαδό κάτω από την καμπύλη ROC. Τα ευρήματα αυτά παρέχουν μια μεθοδολογία για την ανάπτυξη συνεχών δεικτών που σχετίζονται με την υγεία για την πρόβλεψη της κλινικής κατάστασης ενός ατόμου.

Λέξεις Κλειδιά: Καμπύλη λειτουργικών χαρακτηριστικών, ευαισθησία, ειδικότητα.

1. ΕΙΣΑΓΩΓΗ

Οι δείκτες υγείας χρησιμοποιούνται εκτεταμένα σε επιστημονικούς τομείς, όπως είναι η Ιατρική, η Βιομετρία, η Βιοστατιστική, η Ψυχομετρία, η Διατροφολογία κ.α. (Kant, 1996; Mcdowell I, 2002; Jackson, 1970). Οι δείκτες υγείας, απλοί ή σύνθετοι, διακριτοί ή συνεχείς, αποσκοπούν στη μέτρηση διαφόρων κλινικών

χαρακτηριστικών, τα οποία είναι δύσκολο ή αδύνατο να μετρηθούν ποσοτικά (Streiner & Norman, 2008; Panagiotakos, 2009). Για παράδειγμα, στον τομέα της ψυχομετρίας γίνεται χρήση σύνθετων διακριτών δεικτών για τη μέτρηση του άγχους ή της κατάθλιψης (Zung, 1965; Jackson, 1970), στον τομέα της διατροφολογίας γίνεται χρήση σύνθετων διακριτών δεικτών (π.χ. MedDietScore που εκφράζει το βαθμό προσήλωσης στο πρότυπο της Μεσογειακής διατροφής και συνδέεται με την πρόληψη γενικότερα του καρδιαγγειακού κινδύνου) που ως στόχο έχουν τη μέτρηση της επάρκειας και της ποιότητας της διατροφής και πως συνδέεται με την εμφάνιση ασθενειών (Panagiotakos et al.; 2007, Kourlaba G, Panagiotakos D.B., 2009a). Επίσης, στο χώρο των επιστήμων υγείας γίνεται χρήση του συνεχούς δείκτη VAS (Carlsson M., 1983) που παρέχει μια συνεχή κλίμακα για την εκτίμηση του πόνου ενός ασθενούς.

Ένας σύνθετος δείκτης υγείας T, συνήθως, δημιουργείται από το μη σταθμισμένο άθροισμα m διακριτών ή συνεχών συνιστωσών μεταβλητών που επιλέγονται κατάλληλα, όπως παρατίθεται ακολούθως:

$$T = \sum_{j=1}^{m} X_j$$

Οι m συνιστώσες μεταβλητές εκφράζουν m διαφορετικές διαστάσεις της κλινικής κατάστασης ενός ατόμου και το άθροισμα τους παρέχει την τιμή του αντίστοιχου δείκτη. Για την κατασκευή ενός σύνθετου δείκτη υγείας απαιτούνται:

- Επιλογή των κατάλληλων μεταβλητών, σχετικών με το κλινικό χαρακτηριστικό που αποβλέπει να προβλέπει ο δείκτης.
- Επιλογή του κατάλληλου αριθμού των διαμερίσεων (αριθμός πιθανών απαντήσεων) για κάθε διακριτή συνιστώσα μεταβλητή ή επιλογή του κατάλληλου πεδίου ορισμού για κάθε συνεχή συνιστώσα μεταβλητή.
- Επιλογή ειδικών σταθμίσεων για την εκάστοτε συνιστώσα μεταβλητή.

Η επιλογή των συνιστωσών μεταβλητών πραγματοποιείται κατά περίπτωση από τον ερευνητή, αλλά μια βασική αρχή είναι ότι ο σύνθετος δείκτης θα πρέπει να έχει την ικανότητα να διακρίνει, ή αλλιώς να ταξινομεί, τα υπό εξέταση άτομα ως ασθενείς ή μη, ως προς το κλινικό χαρακτηριστικό (π.χ. κατάθλιψη) που ο δείκτης σκοπεύει να εκτιμήσει (Οι ασθενείς εμφανίζουν το αντίστοιχο κλινικό χαρακτηριστικό ενώ οι υγιείς όχι). Η χρήση ενός σύνθετου δείκτη υγείας προσφέρει τη δυνατότητα έγκαιρης διάγνωσης μιας νόσου με συνέπεια την έγκαιρη εφαρμογή θεραπείας στο άτομο που πιθανόν νοσεί (McCullough et al, 2000). Η ορθότητα της διάγνωσης έγκειται στην επάρκεια ανίχνευσης ασθενών και υγιών από το διαγνωστικό έλεγχο/δείκτη, ή διαφορετικά, στον έγκυρο διαχωρισμό αυτών. Ένας δείκτης υγείας χρησιμοποιείται ως εξής: αν η τιμή του δείκτη είναι μικρότερη ή μεγαλύτερη από μια δεδομένη τιμή (η τιμή αυτή αποτελεί το διαχωριστικό όριο μεταξύ υγιών και ασθενών), το εκάστοτε άτομο, κατατάσσεται ως υγιές ή μη. Για παράδειγμα, ο διακριτός δείκτης του Zung (Zung, 1965) είναι ένας δείκτης που αποτιμά το βαθμό κατάθλιψης ενός ατόμου και αποτελείται από 20 ερωτήσεις (διακριτές μεταβλητές) όπως π.χ. «Αισθάνομαι αποκαρδιωμένος ή λυπημένος», κλπ). Ο αριθμός των δυνατών απαντήσεων εκφράζει το πλήθος των διαμερίσεων της

αντίστοιχης μεταβλητής. Η απάντηση καθεμιάς βαθμονομείται με σκορ από 1 έως 4, ανάλογα με τη συχνότητα εμφάνισης του αντίστοιχου συμπτώματος (καθόλου, μερικές φορές, συχνά, πάντοτε). Η συνολική τιμή του δείκτη προκύπτει από το άθροισμα των επιμέρους τιμών που λαμβάνουν οι συνιστώσες. Μεγάλες τιμές του δείκτη για ένα άτομο, εκφράζουν ότι το άτομο αυτό πιθανόν να πάσχει κατάθλιψη. Ο αριθμός των διαμερίσεων, στην περίπτωση των διακριτών δεικτών, προτείνεται να είναι ο μεγαλύτερος που δύναται να χρησιμοποιηθεί (Bersimis et al, 2013).

Στόχος της παρούσας εργασίας είναι η αξιολόγηση της τυχόν επίδρασης προτεινόμενων σταθμίσεων που αποδίδονται στις συνιστώσες μεταβλητές ενός σταθμισμένου σύνθετου δείκτη υγείας, στη διαγνωστική ακρίβεια αυτού, με χρήση του εμβαδού (AUC) κάτω από την Καμπύλη Λειτουργικών Χαρακτηριστικών (ROC), της ευαισθησίας και της ειδικότητας (Kourlaba, G., Panagiotakos D.B., 2009b) σε συγκεκριμένα διαχωριστικά σημεία (cutoff points). Πιο συγκεκριμένα γίνεται χρήση του βέλτιστου διαγνωστικού κατωφλιού (optimum cutoff point - δηλαδή του σημείου της καμπύλης ROC όπου ισχύει (max(Se+Sp))) και ενός τυχαίου σημείου (3ο τεταρτημόριο) που σχετίζεται με το λόγο ασθενών-υγειών (1-3) του εκάστοτε δείκτη που παράγεται. Πιο συγκεκριμένα, για τον εκάστοτε δείκτη υπολογίστηκαν (α) το βέλτιστο διαγνωστικό κατώφλι του και (β) το τρίτο τεταρτημόριο του. Η ευαισθησία (S_e) ενός σύνθετου δείκτη υγείας εκφράζει την ικανότητα εντοπισμού των πραγματικά παθολογικών περιπτώσεων, η ειδικότητα (S_p) ενός διαγνωστικού ελέγχου ή ενός σύνθετου δείκτη υγείας εκφράζει την ικανότητα εντοπισμού των πραγματικά φυσιολογικών (υγιών) περιπτώσεων και η καμπύλη ROC συνεκτιμά τους δύο δείκτες (S_e & S_p) κάνοντας χρήση του εμβαδού (AUC) κάτω από την καμπύλη. Η καμπύλη ROC ορίζεται στο μοναδιαίου εμβαδού τετράγωνο [0,1]x[0,1] του καρτεσιανού επιπέδου και σχηματίζεται από τα σημεία με συντεταγμένες (Se, 1-Sp), δηλαδή από τα σημεία με τετμημένη την πιθανότητα αληθώς θετικών (ευαισθησία) και τεταγμένη την πιθανότητα ψευδώς θετικών (1-ειδικότητα), για όλα τα δυνατά διαγνωστικά κατώφλια (διαχωριστικά όρια - cutoff points). Το εμβαδόν (AUC) κάτω από την Καμπύλη (ROC) αποτελεί ένα μέτρο/κριτήριο για το διαχωρισμό ασθενών – υγειών και λαμβάνει τιμές στο σύνολο [0.5,1]. Το μέγιστο AUC είναι το επιθυμητό διότι όσο μεγαλύτερη είναι η τιμή του εμβαδού κάτω από την καμπύλη, τόσο μεγαλύτερη είναι η ακρίβεια του διαγνωστικού ελέγχου προς ανίχνευση ασθενών, ή μη, ατόμων. Στην περίπτωση που οι κατανομές ασθενών και υγειών συμπίπτουν, τότε η τιμή του εμβαδού AUC είναι 0.5. Στην περίπτωση που οι κατανομές ασθενών και υγειών δεν συμπίπτουν, τότε η τιμή του εμβαδού AUC είναι 1. Το εμβαδό AUC συνδέεται με τον έλεγχο του Wilcoxon test και επίσης με το στατιστικό U του Mann-Whitney (Hanley J. A., McNeil B J, 1982). Ο υπολογισμός του AUC (παρέχεται από κατάλληλο λογισμικό). έγινε με χρήση μη παραμετρικής προσέγγισης της πιθανότητας $P(T_{A\Theta}>T_{AA})$, όπου $T_{A\Theta}$ είναι οι τιμές του δείκτη για τις αληθώς θετικές περιπτώσεις, δηλαδή τους ασθενείς, και T_{AA} είναι οι τιμές του δείκτη για τις αληθώς αρνητικές περιπτώσεις, δηλαδή τους μη ασθενείς με χρήση του τύπου:

$$\frac{1}{n_{\text{ΑΘ}} \cdot n_{\text{ΑΑ}}} \cdot \sum S\left(T_{\text{ΑΘ}}, T_{\text{ΑΑ}}\right)$$, για όλους τους δυνατούς συνδυασμούς ($T_{ΑΘ}$,$T_{ΑΑ}$).

όπου $S\left(T_{\text{ΑΘ}}, T_{\text{ΑΑ}}\right) = \begin{cases} 1, & \text{αν } T_{\text{ΑΘ}} > T_{\text{ΑΑ}} \\ 1/2, & \text{αν } T_{\text{ΑΘ}} = T_{\text{ΑΑ}} \\ 0, & \text{αν } T_{\text{ΑΘ}} < T_{\text{ΑΑ}} \end{cases}$ και $n_{\text{ΑΘ}}, n_{\text{ΑΑ}}$ είναι τα πλήθη ασθενών και υγειών αντιστοίχως (Hanley J. A., McNeil B J, 1982). Η ευαισθησία και η ειδικότητα εκφράζονται μέσω δεσμευμένων πιθανοτήτων και υπολογίστηκαν ως λόγοι όπως ακολούθως:

$$S_e\left(T\right) = P\left(T > c \mid Y = 1\right) = \frac{\text{ΑΘ}}{\text{ΑΘ} + \Psi\text{Α}} \text{ και } S_p\left(T\right) = P\left(T < c \mid Y = 0\right) = \frac{\text{ΑΑ}}{\text{ΑΑ} + \Psi\Theta}$$

όπου ΑΘ είναι οι ασθενείς που ταξινομήθηκαν αληθώς ως ασθενείς, ΨΑ είναι οι ασθενείς που ταξινομήθηκαν ψευδώς ως υγιείς, ΑΑ είναι οι υγιείς που ταξινομήθηκαν αληθώς ως υγιείς, ΨΘ είναι οι υγιείς που ψευδώς ταξινομήθηκαν ως ασθενείς.

Η εργασία αυτή περιλαμβάνει, εκτός της εισαγωγής και της συζήτησης, τα ακόλουθα κυρίως μέρη: Το πρώτο μέρος παρουσιάζει τις μεθόδους στάθμισης που χρησιμοποιήθηκαν (με χρήση Λογιστικής Παλινδρόμησης & Διαχωριστικής Ανάλυσης), ομοιότητες και διαφορές μεταξύ των παραπάνω μεθόδων, τη μεθοδολογία κατασκευής των σύνθετων δεικτών και τα αποτελέσματα που παρήχθησαν με χρήση του λογισμικού IBM SPSS Statistics 21.

2. ΜΕΘΟΔΟΙ ΣΤΑΘΜΙΣΗΣ

Οι συντελεστές στάθμισης, w_{ij}, *j=1, 2,...,5, i=1,2,...,8*, που χρησιμοποιήθηκαν για την κατασκευή των προτεινόμενων σταθμισμένων δεικτών προέρχονται από δύο ευρέως χρησιμοποιούμενες στατιστικές τεχνικές που ως στόχο έχουν το διαχωρισμό παρατηρήσεων σε δύο ομάδες, τη Λογιστική Παλινδρόμηση και τη Διαχωριστική Ανάλυση. Με τη χρήση ειδικών σταθμίσεων παρήχθησαν δείκτες που δίνονται από τον παρακάτω τύπο:

$$\mathrm{T}_i = \sum_{j=1}^{5} w_{ij} X_j, \quad j = 1,2,...,5, \quad i = 1,2,...,8$$

όπου w_{ij},είναι, για παράδειγμα, οι σχετικοί λόγοι πιθανοτήτων από τη λογιστική παλινδρόμηση ή οι μη τυποποιημένοι συντελεστές της διαχωριστικής συνάρτησης από τη διαχωριστική ανάλυση. Η λογική που κρύβεται πίσω από τους προτεινόμενους σταθμισμένους δείκτες είναι ότι μεταβλητές που διαχωρίζουν με μεγαλύτερη σαφήνεια τους υγιείς από τους ασθενείς, θα συνεισφέρουν περισσότερο στο τελικό αποτέλεσμα (σκορ) του δείκτη. Οι σταθμισμένοι δείκτες παρουσιάζονται αναλυτικά σε επόμενη παράγραφο.

2.1 Λογιστική Παλινδρόμηση

Το λογιστικό μοντέλο παλινδρόμησης είναι ένα μη γραμμικό μοντέλο παλινδρόμησης που εφαρμόζεται στην περίπτωση που η εξαρτημένη μεταβλητή απόκρισης Y δεν είναι ποσοτική, αλλά κατηγορική με δύο ή περισσότερες κατηγορίες. Στην παρούσα εργασία, εφαρμόστηκε λογιστική παλινδρόμηση με δίτιμη εξαρτημένη μεταβλητή (Binary Logistic Regression) που εκφράζει παρουσία (Y=1) ή απουσία νόσου (Y=0). Το απλό λογιστικό μοντέλο είναι της μορφής (Σταυρινός, Παναγιωτάκος, 2007) που δίνεται από τη σχέση (1):

$$P\left(Y=1 \mid X_j\right)=\pi\left(X_j\right)=\frac{e^{a+\beta X_j}}{1+e^{a+\beta X_j}}=\left(1+e^{-\left(a+\beta X_j\right)}\right)^{-1}, \quad j=1,2,..,5 \tag{1},$$

όπου $P(Y=1/X_j)=\pi(X_j)$ εκφράζει το λογαριθμικό υπόδειγμα και είναι η πιθανότητα παρουσίας της νόσου δεδομένου του χαρακτηριστικού Χ. Επομένως η πιθανότητα απουσίας της νόσου δίνεται από την μαθηματική έκφραση της σχέσης (2):

$$P\left(Y=0 \mid X_j\right)=1-\pi\left(X_j\right)=\frac{1}{1+e^{a+\beta X_j}}=\left(1+e^{\left(a+\beta X_j\right)}\right)^{-1}, \quad j=1,2,...,5 \tag{2}$$

Το πηλίκο $\pi(X_j)/(1-\pi(X_j))$ εκφράζει το λόγο των συμπληρωματικών πιθανοτήτων (odds) και δίνεται από τον τύπο της σχέσης (3) (Σταυρινός, Παναγιωτάκος, 2007) λύνοντας τη σχέση (1) ως προς τον όρο $\exp(\alpha+\beta X_j)$:

$$e^{a+\beta X_j}=\frac{\pi\left(\mathrm{X}_j\right)}{1-\pi\left(\mathrm{X}_j\right)}, \quad j=1,2,...,5 \tag{3}$$

Λαμβάνοντας το λογάριθμο κατά μέλη στη σχέση (3), ο γραμμικός όρος $\alpha+\beta X$ δίνεται από τη σχέση (4):

$$a+\beta \mathrm{X}_j=\ln\left(\frac{\pi\left(\mathrm{X}_j\right)}{1-\pi\left(\mathrm{X}_j\right)}\right), \quad j=1,2,...,5 \tag{4}$$

Ο όρος e^{β} (σχέση 1 & 2) ισούται με το σχετικό λόγο συμπληρωματικών πιθανοτήτων να εμφανιστεί η νόσος σε όσους έχουν ένα δίτιμο χαρακτηριστικό Χ σε σύγκριση με αυτούς που δεν το έχουν, οπότε ισχύει:

$$OR=\frac{\pi(1)\cdot\left[1-\pi(0)\right]}{\pi(0)\cdot\left[1-\pi(1)\right]}=e^{\beta} \tag{5}$$

Στο ίδιο αποτέλεσμα καταλήγουμε αν η ανεξάρτητη μεταβλητή X_j είναι ποσοτική. Η διαφοροποίηση εντοπίζεται στον τρόπο ερμηνείας του αντίστοιχου σχετικού λόγου συμπληρωματικών πιθανοτήτων. Το πολλαπλό λογιστικό μοντέλο παλινδρόμησης που χρησιμοποιήθηκε στην παρούσα εργασία είναι της μορφής:

$$P\left(Y=1 \mid \tilde{X}\right)=\pi\left(\tilde{X}\right)=\frac{e^{g\left(\tilde{X}\right)}}{1+e^{g\left(\tilde{X}\right)}} \tag{6}$$

όπου $\tilde{\mathrm{X}}=\left(X_1, X_2, X_3, X_4, X_5\right)$ το διάνυσμα των ανεξάρτητων κανονικών μεταβλητών X_j, j=1,2,..,5 και $g\left(\tilde{\mathrm{X}}\right)=a+\beta_1 \mathrm{X}_1+\beta_2 \mathrm{X}_2+\beta_3 \mathrm{X}_3+\beta_4 \mathrm{X}_4+\beta_5 \mathrm{X}_5$.

Στα πλαίσια των προτεινόμενων σταθμίσεων για τους παραγόμενους δείκτες της εργασίας αυτής έγινε χρήση της απόκλισης (Deviance) μεταξύ παρατηρούμενων και προβλεπόμενων τιμών του λογιστικού υποδείγματος, δηλαδή της απόκλισης μεταξύ

πλήρους (θεωρητικού) υποδείγματος (Full) και του υποδείγματος που εκτιμήθηκε συναρτήσει των διαθέσιμων ανεξάρτητων μεταβλητών (Estimated). Το επιθυμητό αποτέλεσμα είναι το εκτιμώμενο υπόδειγμα να «πλησιάζει» το πλήρες μοντέλο και τότε η απόκλιση λαμβάνει μικρές τιμές. Η απόκλιση (Deviance) εκφράζεται με τη βοήθεια του λόγου πιθανοφανειών (-2LL) μεταξύ των δυο προαναφερθέντων υποδειγμάτων, δηλαδή είναι ένα μέτρο της καλής προσαρμοστικότητας του μοντέλου λογιστικής παλινδρόμησης και δίνεται από τη μαθηματική έκφραση της σχέσης (7):

$$D = -2\ln l\left(estimated\right) - \left[-2\ln l\left(full\right)\right] \tag{7}$$

Η στατιστική D ακολουθεί χ^2 κατανομή με n-p βαθμοί ελευθερίας, όπου n το μέγεθος του δείγματος και p ο αριθμός των ερμηνευτικών μεταβλητών που συμμετέχουν στο μοντέλο λογιστικής παλινδρόμησης.

2.2 Διαχωριστική Ανάλυση

Η διαχωριστική ανάλυση είναι μια στατιστική τεχνική η οποία ως στόχο έχει την κατάταξη παρατηρήσεων σε δύο ή περισσότερους ήδη γνωστούς πληθυσμούς (Καρλής, 2005). Βασική διαφορά της διαχωριστικής ανάλυσης από την ανάλυση κατά συστάδες είναι ότι στην τελευταία οι ομάδες δεν είναι γνωστές ενώ στην πρώτη είναι γνωστές. Για την υλοποίηση της μεθόδου της γραμμικής διαχωριστικής ανάλυσης (Linear Discriminant Analysis) απαιτείται η κατασκευή ενός κανόνα διαχωρισμού των ομάδων (π.χ. Μέγιστης Πιθανοφάνειας, Μέθοδος Bayes, Ελαχιστοποίηση του κόστους εσφαλμένης κατάταξης) με τη βοήθεια του οποίου, οι παρατηρήσεις κατατάσσονται στις διακριτές ομάδες μέσω μιας γραμμικής διαχωριστικής συνάρτησης (διαχωριστική συνάρτηση του Fisher). Υποθέτοντας κανονικότητα των πληθυσμών και ισότητα των πινάκων συνδιακύμανσης, τότε για τυχαία παρατήρηση $x \sim N\left(\mu_k, \Sigma\right)$, για την αντίστοιχη πυκνότητα της πολυδιάστατης κανονικής κατανομής ισχύει:

$$f_k\left(x \mid \mu_k, \Sigma\right) = \left(2\pi\right)^{-p/2} \cdot \left|\Sigma\right|^{-1/2} \cdot \exp\left(-\frac{1}{2}\left(x-\mu_k\right)'\Sigma^{-1}\left(x-\mu_k\right)\right), \quad k = 1,2 \tag{8}$$

όπου η ποσότητα $\left(x-\mu_k\right)'\Sigma^{-1}\left(x-\mu_k\right)$ είναι η απόσταση Mahalanobis της παρατήρησης x από τον μέσο της k ομάδας. Στην περίπτωση αυτή ο αντίστοιχος διαχωριστικός κανόνας δίνεται από τη σχέση (9):

$$\log\left(\frac{f_1\left(x \mid \mu_1, \Sigma\right)}{f_2\left(x \mid \mu_2, \Sigma\right)}\right) = \left(\mu_1-\mu_2\right)'\Sigma^{-1}x - \frac{1}{2}\left(\mu_1-\mu_2\right)'\Sigma^{-1}\left(\mu_1+\mu_2\right) \geq k_0 \tag{9}$$

όπου k_0 εξαρτάται από τον κανόνα διαχωρισμού των ομάδων (π.χ. στην περίπτωση του κανόνα μέγιστης πιθανοφάνειας ισχύει k_0=1)

Ο διαχωριστικός κανόνας του Fisher μετατρέπει τις αρχικές μεταβλητές σε μονοδιάστατες τιμές μέσω της παραγόμενης διαχωριστικής συνάρτησης (discriminant function) (Καρλής, 2005). Αρχικά, παράγονται δύο γραμμικές συναρτήσεις με τη μέθοδο του Fisher, W_1 και W_2, μία για τους ασθενείς και μια για τους υγιείς, που ονομάζονται διαχωριστικές συναρτήσεις του Fisher (Classification Function Coefficients). Θεωρώντας τη διαφορά μεταξύ των διαχωριστικών συναρτήσεων του Fisher $Z=W_1-W_2$, λαμβάνουμε μια συνάρτηση, της οποίας οι συντελεστές είναι

ανάλογοι των μη τυποποιημένων συντελεστών της κανονικοποιημένης διαχωριστικής συνάρτησης (Canonical Discriminant Function Coefficients). Έτσι, αν ισχύει ότι Z>0, τότε κατατάσσουμε την αντίστοιχη παρατήρηση, στην $1^{η}$ ομάδα, διαφορετικά στη $2^{η}$. Για την εύρεση των σταθμίσεων δεικτών $T_5,\ldots,T_8$, που προέρχονται από τη Διαχωριστική Ανάλυση, γίνεται χρήση των μη τυποποιημένων συντελεστών της κανονικοποιημένης διαχωριστικής συνάρτησης (CDFC), οι οποίοι είναι ανάλογοι με τους συντελεστές της διαχωριστικής συνάρτησης $Z=W_1-W_2$ (Καρλής, 2005). Οι μη τυποποιημένοι συντελεστές της κανονικοποιημένης διαχωριστικής συνάρτησης προκύπτουν από την εύρεση των ιδιοτιμών της σχέσης $(T-W)V=\lambda WV$, όπου V η μήτρα των συντελεστών της διαχωριστικής συνάρτησης, T είναι ο πίνακας τετραγωνικών συνολικών αθροισμάτων, W είναι ο πίνακας τετραγωνικών αθροισμάτων εντός των ομάδων και το λ είναι ένας διαγώνιος πίνακας των ιδιοτιμών. Οι σταθμίσεις για τους δείκτες $T_5,\ldots,T_8$ λαμβάνονται με χρήση της προαναφερθείσας διαχωριστικής συνάρτησης Z, με θετικό πρόσημο. Ένα μέτρο που εκφράζει τη διαχωριστική ικανότητα της προαναφερθείσας συνάρτησης είναι το Λ του Wilks, όπου εκφράζει το ποσοστό της ανερμήνευτης διακύμανσης από το αντίστοιχο μοντέλο της ανάλυσης διακύμανσης κατά ένα παράγοντα, ή διαφορετικά, το ποσοστό μεταβλητότητας που δεν εξηγείται από τη διαχωριστική συνάρτηση και λαμβάνει τιμές από το 0 έως το 1, συνεπώς μικρές τιμές του Λ είναι επιθυμητές (Καρλής, 2005).

2.3 Λογιστική Παλινδρόμηση & Διαχωριστική Ανάλυση

Αμφότερες οι στατιστικές μέθοδοι της λογιστικής παλινδρόμησης και της διαχωριστικής ανάλυσης χρησιμοποιούνται για να κατατάξουν παρατηρήσεις σε γνωστές ομάδες και συναντώνται συχνά στη βιβλιογραφία σε προβλήματα διαχωρισμού πληθυσμών (Worth, Cronin, 2003; Pohar et al, 2004). Στη λογιστική παλινδρόμηση δεν είναι απαραίτητες οι υποθέσεις της διαχωριστικής ανάλυσης που αναφέρονται στην κανονική κατανομή των ανεξάρτητων μεταβλητών καθώς επίσης και στην υπόθεση περί ισότητας διασπορών μεταξύ των πληθυσμών. Βασική διαφορά των δύο μεθόδων είναι ότι στη διαχωριστική ανάλυση γίνεται χρήση αποκλειστικά ποσοτικών μεταβλητών ως ανεξάρτητων, ενώ, στην περίπτωση της λογιστικής παλινδρόμησής δύναται να χρησιμοποιηθούν και ποιοτικές μεταβλητές. Συνεπώς, θα μπορούσε να υποτεθεί ότι η λογιστική παλινδρόμηση είναι πιο εύχρηστη μέθοδος από αυτή της διαχωριστικής ανάλυσης, αλλά στην περίπτωση που ικανοποιούνται οι υποθέσεις περί κανονικότητας και ισότητας των πινάκων διακύμανσης, τότε η διαχωριστική ανάλυση θα δίνει καλύτερα αποτελέσματα.

2.4 Δημιουργία Δεδομένων

Κατασκευάστηκε πλήθος δεδομένων για τις ανεξάρτητες μεταβλητές X_j, *j=1,2,..,5*. Ενδεικτικά παρουσιάζονται δύο διαφορετικά σενάρια στα οποία οι παραγόμενες μεταβλητές προέκυψαν από την κανονική κατανομή. Στο $1^{ο}$ σενάριο υποθέτουμε σταθερή μέση τιμή με μεταβαλλόμενη τυπική απόκλιση και στο $2^{ο}$ σενάριο υποθέτουμε το αντίστροφο. Το μέγεθος δείγματος για την κάθε μεταβλητή είναι ίσο με 10.000 και η αναλογία ασθενών – υγιών είναι 1 προς 3, επομένως

παρήχθησαν δεδομένα για 2.500 ασθενείς και 7.500 υγιείς. Ο γενικός κανόνας που υιοθετήθηκε αναφέρει ότι η μέση τιμή ασθενών είναι μεγαλύτερη της μέσης τιμής των υγιών, διότι για πολλές ασθένειες, οι αντίστοιχοι βιοχημικοί δείκτες των πασχόντων ατόμων εμφανίζουν μεγαλύτερες τιμές σε σχέση με εκείνες των υγειών ατόμων, που κρίνονται ως φυσιολογικές. Πιο συγκεκριμένα, στο $1^{ο}$ σενάριο δημιουργήθηκαν υποθετικές μεταβλητές που η μέση τιμή των ασθενών ήταν 15 και η αντίστοιχη μέση τιμή των υγιών ήταν 12, ενώ, στο $2^{ο}$ σενάριο δημιουργήθηκαν μεταβλητές που η μέση τιμή των ασθενών ήταν μεταβαλλόμενη (από 15 έως 13) και η αντίστοιχη μέση τιμή των υγιών ήταν επίσης μεταβαλλόμενη (από 10 έως 12), αλλά σταθερά χαμηλότερη από αυτή των ασθενών. Για τη σαφή διάκριση ασθενών και υγειών, έγινε χρήση δίτιμης μεταβλητής Y, όπου η τιμή Y=1 εκφράζει παρουσία νόσου και η τιμή Y=0 εκφράζει απουσία νόσου.

***Πίνακας 1**: Κατανομή και παράμετροι συνιστωσών μεταβλητών των δεικτών*

Σενάριο $1^{ο}$:Σταθερή Μέση Τιμή – Μεταβ. Τυπ. Απόκ.		Σενάριο $2^{ο}$:Μεταβ. Μέση Τιμή – Σταθερή. Τυπ. Απόκ.	
Υγιείς	**Ασθενείς**	**Υγιείς**	**Ασθενείς**
X_1~N(12,3)	**X_1~N(15,3)**	**X_1~N(10,2)**	**X_1~N(15,2)**
X_2~N(12,2.5)	**X_2~N(15,2.5)**	**X_2~N(10.5,2)**	**X_2~N(14.5,2)**
X_3~N(12,2)	**X_3~N(15,2)**	**X_3~N(11,2)**	**X_3~N(14,2)**
X_4~N(12,1.5)	**X_4~N(15,1.5)**	**X_4~N(11.5,2)**	**X_4~N(13.5,2)**
X_5~N(12,1)	**X_5~N(15,1)**	**X_5~N(12,2)**	**X_5~N(13,2)**

Στον πίνακα 1 παρατίθενται αναλυτικά τα χαρακτηριστικά των παραγόμενων μεταβλητών. Στο σημείο αυτό πρέπει να σημειωθεί ότι παρήχθη πλήθος όμοιων σεναρίων με τα 2 προαναφερθέντα (με μικρότερη ή μεγαλύτερη απόκλιση των μέσων τιμών μεταξύ των ασθενών και των υγιών, με μικρότερες ή μεγαλύτερες αποκλίσεις μεταξύ των μεταβαλλόμενων τυπικών αποκλίσεων) και τα αποτελέσματα ήταν σταθερά.

2.5 Μεθοδολογία Κατασκευής Σύνθετων Δεικτών

Στα πλαίσια αυτής της εργασίας κατασκευάστηκαν οκτώ δείκτες T_i, *i=1,2,,…,8* με διαφορετική μεθοδολογία ο καθένας. Για το δείκτη T_1 δεν χρησιμοποιήθηκε στάθμιση, οπότε παράγεται από το απλό αλγεβρικό άθροισμα των 5 μεταβλητών X_j, *j=1,2,..,5*. Στη συνέχεια εκτελέστηκε λογιστική παλινδρόμηση και διαχωριστική ανάλυση με εξαρτημένη μεταβλητή την Y που εκφράζει την κλινική κατάσταση των ατόμων και ανεξάρτητες τις μεταβλητές X_j, j=1, 2,..,5. Με χρήση των σχετικών λόγων πιθανοτήτων και των μη τυποποιημένων συντελεστών της κανονικοποιημένης διαχωριστικής συνάρτησης που παρήχθησαν από τις δύο αυτές μεθόδους, λαμβάνονται οι προτεινόμενες σταθμίσεις. Πιο συγκεκριμένα, οι δείκτες T_2 και T_3 παρήχθησαν με χρήση στάθμισης, τους σχετικούς λόγους πιθανοτήτων από την μονομεταβλητή (SLR OR) και πολυμεταβλητή (MLR OR) λογιστική παλινδρόμηση αντιστοίχως της εξαρτημένης μεταβλητής Y με ανεξάρτητες μεταβλητές τις X_1, X_2, X_3, X_4, X_5. Στον δείκτη T_4 χρησιμοποιήθηκαν, ως σταθμίσεις, οι λόγοι πιθανοτήτων από τη μονομεταβλητή λογιστική παλινδρόμηση, διορθωμένοι με τη χρήση της μέγιστης πιθανοφάνειας, λαμβάνοντας το πηλίκο τους (SLR OR/(-2LL)). Χρησιμοποιήθηκε η απόκλιση (Deviance) -2LL διότι εκφράζει το ποσοστό της

απόκλισης μεταξύ εκτιμώμενου και πραγματικού μοντέλου παλινδρόμησης, οπότε με στάθμιση το αντίστοιχο πηλίκο, ενισχύονται εκείνες οι μεταβλητές που αντιστοιχούν σε χαμηλότερες τιμές απόκλισης. Οι δείκτες T_5 και T_6 παρήχθησαν με χρήση σταθμίσεων, τους μη τυποποιημένους συντελεστές της κανονικοποιημένης διαχωριστικής συνάρτησης από τη μονομεταβλητή (SDA CDFC) και πολυμεταβλητή (MDA CDFC) διαχωριστική ανάλυση αντιστοίχως, με μεταβλητή ομαδοποίησης την Y και ανεξάρτητες μεταβλητές τις X_1, X_2, X_3, X_4, X_5.. Ως στάθμιση για τους Δείκτες T_7 και T_8 χρησιμοποιήθηκαν τα πηλίκα των μη τυποποιημένων συντελεστών της κανονικοποιημένης διαχωριστικής συνάρτησης από την μονομεταβλητή και πολυμεταβλητή Διαχωριστική Ανάλυση διορθωμένοι με τη χρήση του Λ του Wilks ((SDA CDFC)/LW & (MDA CDFC)/LW). Χρησιμοποιήθηκε το Λ του Wilk διότι εκφράζει το ποσοστό της μεταβλητότητας που δεν εξηγείται από την αντίστοιχη διαχωριστική συνάρτηση, οπότε με στάθμιση το αντίστοιχο πηλίκο, ενισχύονται εκείνες οι μεταβλητές που αντιστοιχούν σε χαμηλότερες τιμές του Λ. Η μαθηματική έκφραση των σταθμισμένων δεικτών δίνεται από τον τύπο:

$$\mathrm{T}_i = \sum_{j=1}^{5} w_{ij} X_j, \quad j = 1,2,...,5, \quad i = 1,2,...,8$$

με σταθμίσεις που πριγράφονται ακολούθως:

$$w_{1j} = 1, w_{2j} = \frac{(SLR\,OR)_j}{\sum_{j=1}^{5}(SLR\,OR)_j}, w_{3j} = \frac{(MLR\,OR)_j}{\sum_{j=1}^{5}(MLR\,OR)_j}, w_{4j} = \frac{(SLR\,OR)_j/(DS)_j}{\sum_{j=1}^{5}\left((SLR\,OR)_j/(DS)_j\right)}, w_{5j} = \frac{(SDA\,CDFC)_j}{\sum_{j=1}^{5}(SDA\,CDFC)_j}, \quad j = 1,2,...,5$$

$$w_{6j} = \frac{(MDA\,CDFC)_j}{\sum_{j=1}^{5}(MDA\,CDFC)_j}, w_{7j} = \frac{(SDA\,CDFC)_j/(LW)_j}{\sum_{j=1}^{5}\left((SDA\,CDFC)_j/(LW)_j\right)}, w_{8j} = \frac{(MDA\,CDFC)_j/(LW)_j}{\sum_{j=1}^{5}\left((MDA\,CDFC)_j/(LW)_j\right)}, \quad j = 1,2,...,5$$

Από τη μαθηματική έκφραση των παραγόμενων δεικτών και λόγω της υπόθεσης της κανονικής κατανομής των συνιστωσών μεταβλητών, οι τιμές των σύνθετων δεικτών κυμαίνονται αριθμητικά σε εύρος που παράγεται βάση των παραμέτρων των κατανομών, για υγιείς και ασθενείς (π.χ. ο δείκτης T_1 λαμβάνει τιμές στο σύνολο [30, 105]).

2.6 Αλγόριθμος Bootstrap

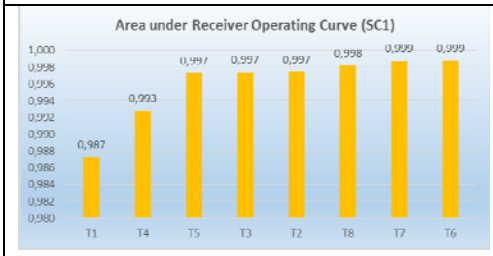
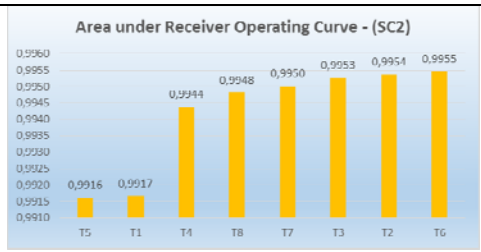
Ο αλγόριθμος Bootstrap είναι μια αναλυτική διαδικασία που περιλαμβάνει δειγματοληψία παρατηρήσεων με επανάθεση από το διαθέσιμο δείγμα, με σκοπό την εκτίμηση παραμέτρων της αντίστοιχης κατανομής (Ross, 2006). Έτσι παράγεται μια κατανομή (Bootstrap Distribution) που τείνει στην εμπειρική κατανομή της αντίστοιχης μεταβλητής και δίνει τη δυνατότητα παραγωγής Διαστημάτων Εμπιστοσύνης, με απώτερο σκοπό την ενίσχυση ενός αποτελέσματος με την αντίστοιχη στατιστική σημαντικότητα.

Για την αξιολόγηση της στατιστικής σημαντικότητας των διαφορών της ευαισθησίας και ειδικότητας μεταξύ των δεικτών έγινε χρήση του αλγόριθμου bootstrap (bootstrap method). Με αυτό τον τρόπο παρήχθησαν διαστήματα εμπιστοσύνης για την ευαισθησία και την ειδικότητα του κάθε δείκτη στο βέλτιστο διαγνωστικό κατώφλι (cutoff point) και στο $3^{ο}$ τεταρτημόριο, με παραγωγή 1000 bootstrap δειγμάτων, ιδίου μεγέθους με το αρχικό δείγμα. Πιο συγκεκριμένα, η

επαναδειγματοληψία (resampling) των 1000 Bootstrap δειγμάτων έγινε σε κάθε σενάριο χωριστά και για ασθενείς και υγιείς επίσης χωριστά. Αυτή η διαδικασία πραγματοποιήθηκε παράγοντας νέες μεταβλητές που εκφράζουν την πρόβλεψη του εκάστοτε δείκτη (ασθενής - υγιής), βασιζόμενες στους αντίστοιχους δείκτες και στη συνέχεια κάνοντας χρήση του αντίστοιχου αλγορίθμου bootstrap υπολογίστηκαν κατάλληλα διαστήματα εμπιστοσύνης. Η χρήση της τεχνικής Bootstrap για παραγωγή νέων δειγμάτων εφαρμόζεται κυρίως για την εκτίμηση παραμέτρων (σημειακή και διαστήματος) από πραγματικά δεδομένα.

2.7 Αποτελέσματα – Αξιολόγηση Δεικτών

Η αξιολόγηση των δεικτών πραγματοποιήθηκε με χρήση του εμβαδού (AUC) κάτω από την καμπύλη των λειτουργικών χαρακτηριστικών (ROC), της ευαισθησίας S_e και της ειδικότητας S_p. Πιο συγκεκριμένα, η ευαισθησία και η ειδικότητα υπολογίστηκαν στο βέλτιστο διαγνωστικό κατώφλι και στο 3^o τεταρτημόριο του εκάστοτε δείκτη. Το βέλτιστο διαγνωστικό κατώφλι αντιστοιχεί στην τιμή του εκάστοτε δείκτη που μεγιστοποιεί το άθροισμα της ευαισθησίας και της ειδικότητας του, δηλαδή βρίσκεται στο σημείο της καμπύλης ROC όπου μεγιστοποιούνται οι ορθές αποφάσεις (max(Se+Sp)) και είναι εκείνο με τη μέγιστη απόσταση από τη διαγώνιο του γραφήματος. Το 3^o τεταρτημόριο του εκάστοτε δείκτη επιλέχτηκε ως τυχαία τιμή για τον υπολογισμό της ευαισθησίας και της ειδικότητας λόγω της σχέσης του με το λόγο ασθενών – υγιών. Στα σχήματα 1 και 2 παρατίθεται γραφικά η αξιολόγηση των οκτώ δεικτών ως προς το κριτήριο του AUC, από τα οποία γίνεται εμφανές ότι ο δείκτης T_6 είναι εκείνος που αντιστοιχεί στη μέγιστη τιμή σε αμφότερα τα δύο σενάρια που εξετάστηκαν. Επιπλέον, η τιμή AUC του δείκτη T_6, που παράγεται με σταθμίσεις που προέρχονται από τους μη τυποποιημένους συντελεστές της κανονικοποιημένης διαχωριστικής συνάρτησης της μεθόδου της διαχωριστικής ανάλυσης, είναι στατιστικώς σημαντικά υψηλότερη σε επίπεδο 5% σε σχέση με τον αστάθμιστο δείκτη T_1, αλλά και σε σχέση με κάποιους σταθμισμένους, όπως είναι ο δείκτης T_5, στο 2^o σενάριο. Τα σχήματα 1 και 2 καταδεικνύουν μια σταθερότητα του δείκτη T_6, στη σχετική θέση που καταλαμβάνει ως προς τους άλλους δείκτες για το κριτήριο του AUC, αλλά δεν είναι σταθερή η σχετική διάταξη όλων των υπολοίπων δεικτών. Αυτή η σταθερότητα των αποτελεσμάτων, ως προς το ποιος δείκτης καταλαμβάνει την μέγιστη τιμή στο κριτήριο του AUC, θα μπορούσε να αποδοθεί στο γεγονός ότι οι συνιστώσες μεταβλητές ακολουθούν την κανονική κατανομή, που αποτελεί βασική προϋπόθεση για τη χρήση της διαχωριστικής ανάλυσης.

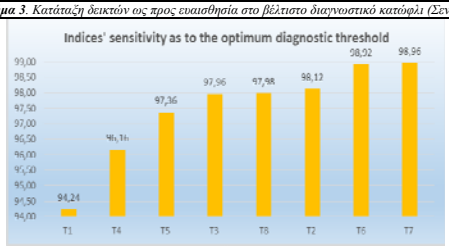
Σχήμα 1. *Κατάταξη δεικτών ως προς το εμβαδό κάτω από την καμπύλη λειτουργικών χαρακτηριστικών (Σενάριο 1)*	***Σχήμα 2***. *Κατάταξη δεικτών ως προς το εμβαδό κάτω από την καμπύλη λειτουργικών χαρακτηριστικών (Σενάριο 2)*
	

Στη συνέχεια, παρατίθεται γραφικά η αξιολόγηση των οκτώ δεικτών με κριτήριο την ευαισθησία, ως προς το βέλτιστο διαγνωστικό κατώφλι και ως προς το 3^{ο} τεταρτημόριο στα σχήματα 3, 4, 5 και 6, από τα οποία είναι εμφανές ότι ο δείκτης T_6 είναι εκείνος που αντιστοιχεί στη μέγιστη ευαισθησία στο 75% των περιπτώσεων.

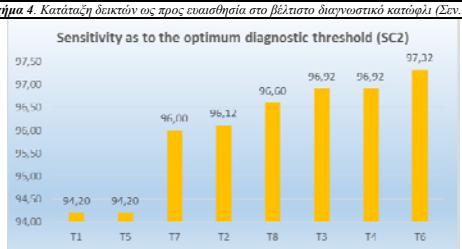
Στην περίπτωση που ο σταθμισμένος δείκτης T_6 δεν αντιστοιχεί στη μέγιστη ευαισθησία, αντιστοιχεί στη 2η μεγαλύτερη τιμή ευαισθησίας χωρίς να είναι στατιστικώς σημαντικά χαμηλότερη από αυτή του T_7, που είναι ο δείκτης που παράγεται με χρήση σταθμίσεων τους διορθωμένους συντελεστές, με την τιμή του Λ του Wilks, της διαχωριστικής συνάρτησης της διαχωριστικής ανάλυσης και αντιστοιχεί στη μέγιστη τιμή ευαισθησίας. Η αξιολόγηση της ειδικότητας των οκτώ δεικτών ως προς το βέλτιστο διαγνωστικό κατώφλι και ως προς το 3^{ο} τεταρτημόριο τους παρατίθεται γραφικά στα σχήματα 7, 8, 9 και 10, από τα οποία είναι εμφανές ότι ο δείκτης T_2 αντιστοιχεί στη μέγιστη τιμή ειδικότητας, στην περίπτωση του 2^{ου} σεναρίου (μεταβαλλόμενη μέση τιμή). Στην περίπτωση του 1^{ου} σεναρίου, δείκτες προερχόμενοι από τη διαχωριστική ανάλυση αντιστοιχούν σε μεγάλες τιμές ευαισθησίας. Από τις παραπάνω επισημάνσεις, με βάση την αναλυτική προσέγγιση της προσομοίωσης, μπορεί να υποτεθεί ότι η διαγνωστική ακρίβεια ενός σύνθετου δείκτη υγείας αυξάνεται (στατιστικώς σημαντικά σε πολλές περιπτώσεις), όταν αποδίδονται ειδικές σταθμίσεις στις 5 συνιστώσες του. Πιο συγκεκριμένα, παρατηρήθηκε ότι όσο πιο πολύ απέχουν οι μέσες τιμές των μεταβλητών X_j, j=1,2,..,5 μεταξύ ασθενών και υγιών, τόσο μεγαλύτερη διαγνωστική ικανότητα έχουν οι σταθμισμένοι δείκτες. Ο βέλτιστος σύνθετος δείκτης υγείας ως προς το κριτήριο του AUC και στα 2 σενάρια είναι ο δείκτης που σχηματίζεται με σταθμίσεις από την Διακρίνουσα Ανάλυση και δείχνει να είναι ο «ανθεκτικότερος» σε περιπτώσεις που διαφοροποιείται η μέση τιμή ή η τυπική απόκλιση των κατανομών των ασθενών και των υγιών & παραμένει σε υψηλές θέσεις ως προς τα κριτήρια αξιολόγησης (κυρίως AUC και Ευαισθησίας). Αναλόγως που εφαρμόζεται μια διαγνωστική διαδικασία, σε γενικό ή ειδικό πληθυσμό, τότε επιθυμητή είναι η αύξηση της ευαισθησίας ή της ειδικότητας αντιστοίχως, επομένως μπορεί να γίνεται χρήση του κατάλληλου δείκτη που αντιστοιχεί σε μεγάλες τιμές του αντίστοιχου κριτηρίου. Για παράδειγμα ο

δείκτης T_2 είναι σταθερά σε υψηλές θέσεις ως προς την Ειδικότητα, στην περίπτωση που είναι σταθερή η τυπική απόκλιση.

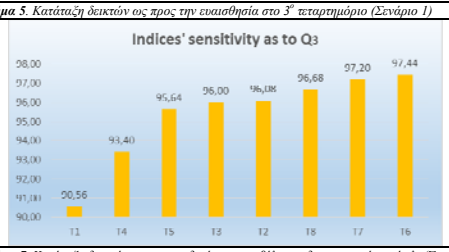
***Σχήμα 3**. Κατάταξη δεικτών ως προς ευαισθησία στο βέλτιστο διαγνωστικό κατώφλι (Σεν. 1)*



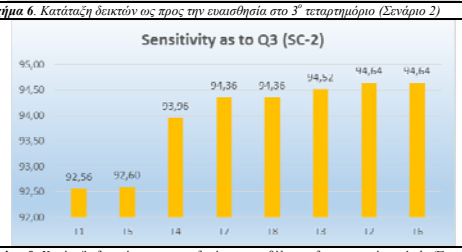
***Σχήμα 4**. Κατάταξη δεικτών ως προς ευαισθησία στο βέλτιστο διαγνωστικό κατώφλι (Σεν. 2)*



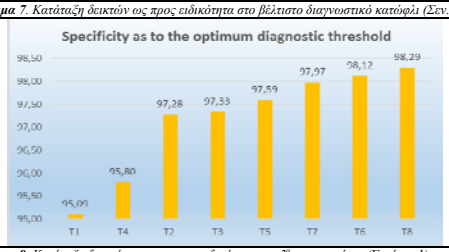
***Σχήμα 5**. Κατάταξη δεικτών ως προς την ευαισθησία στο 3ο τεταρτημόριο (Σενάριο 1)*



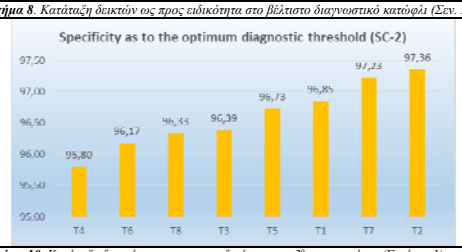
***Σχήμα 6**. Κατάταξη δεικτών ως προς την ευαισθησία στο 3ο τεταρτημόριο (Σενάριο 2)*



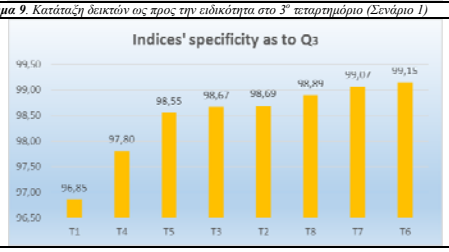
***Σχήμα 7**. Κατάταξη δεικτών ως προς ειδικότητα στο βέλτιστο διαγνωστικό κατώφλι (Σεν. 1)*



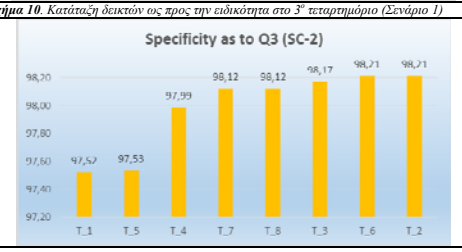
***Σχήμα 8**. Κατάταξη δεικτών ως προς ειδικότητα στο βέλτιστο διαγνωστικό κατώφλι (Σεν. 2)*



***Σχήμα 9**. Κατάταξη δεικτών ως προς την ειδικότητα στο 3ο τεταρτημόριο (Σενάριο 1)*



***Σχήμα 10**. Κατάταξη δεικτών ως προς την ειδικότητα στο 3ο τεταρτημόριο (Σενάριο 1)*



3. ΣΥΖΗΤΗΣΗ

Οι δείκτες αποτελούν πολύτιμο εργαλείο πολλών επιστημονικών κλάδων, όπως είναι η ψυχομετρία, η βιομετρία, κτλ. Για παράδειγμα, στο χώρο της υγείας χρησιμοποιούνται δείκτες υγείας για την ανίχνευση κλινικών χαρακτηριστικών ή στάσεων συμπεριφοράς (Kant, 1996). Συνεπώς, παρότι χρησιμοποιούνται εκτεταμένα στο χώρο της υγείας, στην πλειονότητα των περιπτώσεων οι *m* συνιστώσες μεταβλητές σχηματίζουν το σύνθετο δείκτη υγείας μέσω ενός απλού (αστάθμιστου) αθροίσματος.

Στηριζόμενοι σε προηγούμενες μελέτες (Kourlaba, G., Panagiotakos D.B., 2009b), σε αυτή την εργασία, επιχειρήθηκε να διερευνηθεί αν η διαγνωστική ικανότητα ενός

σύνθετου συνεχούς δείκτη υγείας βελτιώνεται όταν αποδοθούν στις συνιστώσες του, ειδικού τύπου σταθμίσεις. Από τα αποτελέσματα της εργασίας, στο σύνολο των περιπτώσεων που εξετάστηκαν, παρουσιάστηκε αύξηση στο εμβαδόν κάτω από την καμπύλη λειτουργικών χαρακτηριστικών των περισσοτέρων σταθμισμένων δεικτών (π.χ. T_6, T_7 κτλ) ως προς τον αστάθμιστο δείκτη (T_1). Το ίδιο συμπέρασμα παράγεται ως προς την ευαισθησία σε αμφότερα τα δύο σενάρια που εξετάστηκαν, είτε ως προς τυχαίο σημείο ($3^ο$ τεταρτημόριο), είτε ως προς το βέλτιστο διαγνωστικό κατώφλι. Ο αστάθμιστος δείκτης T_1 είχε χαμηλότερη τιμή σε σχέση με τους σταθμισμένους δείκτες που παράγονται με χρήση σταθμίσεων, είτε από τους συντελεστές της διαχωριστικής συνάρτησης, είτε με τη διόρθωση του Λ του Wilks. Στην περίπτωση της ειδικότητας τα παραγόμενα αποτελέσματα δείχνουν ότι οι σταθμισμένοι δείκτες εμφάνισαν υψηλότερη τιμή σε σχέση με τον αστάθμιστο δείκτη, όπως συμβαίνει στην περίπτωση του $2^{ου}$ σεναρίου (Μεταβαλλόμενη μέση τιμή & σταθερή τυπική απόκλιση) ως προς το βέλτιστο διαγνωστικό κατώφλι. Σε κάθε περίπτωση, οι σταθμισμένοι δείκτες αντιστοιχούσαν σε υψηλότερες τιμές ειδικότητας.

Γενικότερα, θα μπορούσε να υποτεθεί ότι η διαγνωστική ακρίβεια ενός σύνθετου δείκτη υγείας αυξάνεται όταν αποδίδονται ειδικές σταθμίσεις στις συνιστώσες του και μάλιστα, όσο πιο πολύ απέχουν οι μέσες τιμές μεταξύ των πληθυσμών ασθενών και υγιών, τόσο μεγαλύτερη διαγνωστική ικανότητα παρουσιάζουν οι σταθμισμένοι δείκτες. Ο βέλτιστος σύνθετος δείκτης υγείας ως προς το κριτήριο του AUC και στα 2 σενάρια που μελετήθηκαν είναι ο δείκτης T_6 που σχηματίζεται με σταθμίσεις από την Διακρίνουσα Ανάλυση και δείχνει να είναι ο «ανθεκτικότερος» σε περιπτώσεις που διαφοροποιείται η μέση τιμή ή η τυπική απόκλιση των κατανομών ασθενών και υγιών με αποτέλεσμα να παραμένει σε υψηλές θέσεις ως προς τα κριτήρια αξιολόγησης. Αναλόγως με το που εφαρμόζεται μια διαγνωστική διαδικασία, σε γενικό ή ειδικό πληθυσμό, τότε επιθυμητή είναι η αύξηση της ευαισθησίας ή της ειδικότητας αντιστοίχως, επομένως μπορεί να γίνεται χρήση του κατάλληλου δείκτη (π.χ. σχήμα 8, 10). Για παράδειγμα ο δείκτης T_2, που προκύπτει με χρήση σταθμίσεων των σχετικών λόγων πιθανοτήτων από τη μονοδιάστατη λογιστική παλινδρόμηση, είναι σταθερά σε υψηλές θέσεις ως προς την Ειδικότητα, στην περίπτωση ανομοιογενών πληθυσμών.

Η παρούσα εργασία παρέχει ένα εφόδιο για την επιλογή του κατάλληλου σύνθετου δείκτη υγείας με χρήση κατάλληλων σταθμίσεων σε σχέση με την έκβαση υγείας που στοχεύει να ανιχνεύσει, όπως ψυχολογικές διαταραχές ή επάρκεια διατροφικής κατάστασης, κτλ. (McCullought et al., 2000; Kranz et al., 2006). Συνήθης πρακτική είναι να επιλέγονται δείκτες χωρίς στάθμιση καθώς ίσως είναι απλούστεροι στη χρήση τους και στον υπολογισμό του σκορ τους. Στην παρούσα εργασία, παρατηρήθηκε ότι οι αστάθμιστοι δείκτες δεν επιτυγχάνουν τόσο υψηλή διαγνωστική ακρίβεια, όσο οι περισσότεροι σταθμισμένοι. Συνεπώς, είναι προτιμότερο να γίνεται χρήση σταθμισμένων δεικτών με χρήση κατάλληλης μεθόδου στάθμισης. Σε κάθε περίπτωση χρειάζεται περαιτέρω έρευνα σε αυτόν τον τομέα καθώς η θεωρητική προσέγγιση του θέματος δεν έχει επαρκώς διερευνηθεί.

ABSTRACT

A composite health related index T is usually constructed by the unweighted sum of discrete or continuous component variables X_j, j = 1,2, ... m and aims to evaluate a person's clinical or biochemical characteristic Y. The present study aims to assess whether the use of special weights (e.g. odds ratios from binary logistic regression, coefficients from Discriminant Analysis) in continuous variables components of a composite health index T improves its diagnostic capability, by using the sensitivity, specificity and curve ROC, with simulated data. For this purpose, eight composite health indices were constructed derived by the sum of five components (random variables), each with a different weighting method and sensitivity and specificity were examined. The results of this work suggest using specific weights to construct composite health indices to increase the sensitivity, specificity and overall the area under the curve ROC. These findings provide a methodology for development continuous indices related to health for predicting the clinical status of a person.

ΑΝΑΦΟΡΕΣ

Kant, A.K. (1996) Indexes of overall diet quality: a review. J. Am. Diet. Assoc. 96, pp. 785-91.

Mcdowell I. Health Measurement Scales. Encyclopedia of Public Health. The Gale Group Inc. 2002. Encyclopedia.com. (July 24, 2009). http://www.encyclopedia.com/doc/1G2-3404000406.html

Jackson, D.N. (1970) a sequential system for personality scale development. In Current topics in clinical and community psychology. New York, Academic Press.

Streiner, D.L., Norman, G.F. (2008) Introduction. In Health Measurement Scales, 4th Ed.; Oxford University Press, USA, pp. 1-4.

Panagiotakos D. (2009) Health measurement scales: methodological issues. Open Cardiovasc. Med. J.3, pp. 160-5.

Zung W.K. William, MD, (1965) A Self-Rating Depression Scale. Arch Gen Psychiatry. 12(1):63-70

Panagiotakos D.B., Pitsavos C., Arvaniti F., Stefanadis C. (2007) Adherence to the Mediterranean food pattern predicts the prevalence of hypertension, hypercholesterolemia, diabetes and obesity, among healthy adults; the accuracy of the MedDietScore. Prev. Med. 44, pp. 335-40.

Kourlaba, G., Panagiotakos D.B. (2009a) Dietary quality indices and human health: a review. Maturitas 62, pp. 1-8.

Carlsson AM. (1983), Assessment of chronic pain. I. Aspects of the reliability and validity of the visual analogue scale. Pain, 16 (1):87-101.

Bersimis, F., Panagiotakos, D., Vamvakari, M.: Sensitivity of health related indices is a non-decreasing function of their partitions. Journal of Statistics Applications & Probability 2(3), 183–194 (2013).

Kourlaba, G., Panagiotakos D.B. (2009b) The diagnostic accuracy of a composite index increases as the number of the partitions of the components increases and when specific weights are assigned to each component. J. Appl. Stat. 37, pp. 537-554.

Hanley J. A., McNeil B J (1982). "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve". diagnostic radiology 143 (1), pp. 29-36.

Stavrinos B., Panagiotakos D.B. Biostatistics, 2007, Gudenberg Publications

Karlis D., Multivariate Statistical Analysis, 2005, Stamoulis Publications

Andrew P. Worth, Mark T.D. Cronin, The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects, Journal of Molecular Structure (Theochem) 622 (2003) 97–111

Maja Pohar, Mateja Blas, and Sandra Turk, Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study. Metodološki zvezki, Vol. 1, No. 1, 2004, 143-161

Sheldon M. Ross, Simulation (4th edition), Simulation, 2006, Elsevier Academic Press

McCullough M.L., Feskanich D., Stampfer M.J., Rosner B.A., Hunter D.J., Variyam J.N., Colditz G.A., Willet W.C. (2000) Adherence to the Dietary Guidelines for Americans and risk of major chronic disease in women. Am. J. Clin. Nutr. 72, pp. 1214-22.

Kranz, S., Hartman, T. et al. (2006). A diet quality index for American preschoolers based on current dietary intake recommendations and an indicator of energy balance. J. Am. Diet. Assoc. 106, pp.1594-604.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής (2015), σελ.252-267

ΠΡΟΒΛΕΨΗ ΘΝΗΣΙΜΟΤΗΤΑΣ ΓΙΑ ΤΟΝ ΕΛΛΗΝΙΚΟ ΠΛΗΘΥΣΜΟ ΚΑΙ ΟΙ ΕΠΙΠΤΩΣΕΙΣ ΜΑΚΡΟΖΩΙΑΣ ΣΤΑ ΑΣΦΑΛΙΣΤΙΚΑ ΤΑΜΕΙΑ

Α. Μποζίκας, Γ. Πιτσέλης
Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς
{bozikas,pitselis}@unipi.gr

ΠΕΡΙΛΗΨΗ

Ο 20ος αιώνας έφερε σημαντικές αλλαγές στα ποσοστά θνησιμότητας για όλες τις ηλικίες, για άνδρες και γυναίκες, σε παγκόσμιο επίπεδο. Στην Ελλάδα, ο αριθμός των ηλικιωμένων αυξήθηκε, ως συνδυασμός της αύξησης στο προσδόκιμο ζωής και των χαμηλών ποσοστών γεννητικότητας. Αυτό είχε ως αποτέλεσμα την επιβάρυνση των ταμείων κοινωνικής ασφάλισης, συμπεριλαμβανομένων των δαπανών για συντάξεις και ιατροφαρμακευτική περίθαλψη. Η παρούσα εργασία παρουσιάζει την εξέλιξη των ποσοστών θνησιμότητας του παρελθόντος, κατά τα επόμενα 40 χρόνια. Η μεθοδολογία των Lee-Carter (1992) αναλύεται και εφαρμόζεται για πρώτη φορά σε ελληνικά δεδομένα. Στη συνέχεια, παρουσιάζονται οι επιπτώσεις της διαφαινόμενης μακροζωίας του ελληνικού πληθυσμού στην αναλογιστική υποχρέωση (και το κανονικό κόστος), για διαφορετικά αναλογιστικά προγράμματα καθορισμένης συνταξιοδοτικής παροχής. Τέλος, εξάγονται συμπεράσματα για τη μελλοντική θνησιμότητα στην Ελλάδα, καθώς και τον ρόλο που καλούνται να έχουν οι αρμόδιες κρατικές αρχές ώστε να μπορέσουν να διαχειριστούν, εκτός των άλλων χρηματοοικονομικών κινδύνων, τις επιπτώσεις μακροζωίας στα ασφαλιστικά ταμεία.

Λέξεις Κλειδιά: Προβολή Θνησιμότητας, Μοντέλο Lee-Carter, Πίνακες Επιβίωσης, Ρίσκο Μακροζωίας, Συνταξιοδοτικά προγράμματα.

1. ΕΙΣΑΓΩΓΗ

Κατά τη διάρκεια του 20ού αιώνα το προσδόκιμο ζωής έχει αυξηθεί σημαντικά σε παγκόσμιο επίπεδο. Αυτό οφείλεται κυρίως στη βελτίωση των συνθηκών διαβίωσης, μέσω της τεχνολογίας και της ανάπτυξης του ιατροφαρμακευτικού κλάδου. Η αύξηση του προσδόκιμου ζωής έχει σαν αποτέλεσμα την εμφάνιση του "Κινδύνου" Μακροζωίας (Longevity Risk) που προκύπτει, όταν οι μελλοντικοί δείκτες θνησιμότητας, εκ των υστέρων, δεν αντικατοπτρίζουν τους αναμενόμενους. Για την αντιμετώπιση του κινδύνου αυτού είναι απαραίτητη η χρήση εξειδικευμένων τεχνικών πρόβλεψης για την κατασκευή προβαλλόμενων πινάκων επιβίωσης, που θα λαμβάνουν υπόψη παρατηρούμενες τάσεις βελτίωσης στα ποσοστά θνησιμότητας.

Η προσπάθεια εύρεσης της καταλληλότερης καμπύλης θνησιμότητας κατέχει μια σημαντική θέση στην ιστορία της δημογραφίας και της αναλογιστικής επιστήμης και στο παρελθόν απασχόλησε σημαντικούς ερευνητές, όπως οι De Moivre, Gompertz, Makeham, Sang και Weibull. Ωστόσο, μελέτες που έγιναν τα τελευταία χρόνια (Stoto, 1983; Keilman, 1998) αποκάλυψαν πολλά σφάλματα στις προβλέψεις. Το 1992, οι Ronald Lee και Lawrence Carter παρουσίασαν ένα εξαιρετικά αποτελεσματικό στοχαστικό μοντέλο για την πρόβλεψη των ειδικών κατά ηλικία δεικτών θνησιμότητας του πληθυσμού των Η.Π.Α.

Η εργασία είναι δομημένη ως εξής: Στη 2η Ενότητα γίνεται μια σύντομη παρουσίαση του μοντέλου Lee-Carter, περιγράφονται οι μεθοδολογίες εύρεσης των παραμέτρων του και επιλογής των κατάλληλων μοντέλων χρονολογικών σειρών ARIMA, για τη προβολή των μελλοντικών δεικτών θνησιμότητας. Στη 3η Ενότητα, προσαρμόζονται στο μοντέλο τα δεδομένα θνησιμότητας ανδρών και γυναικών του ελληνικού πληθυσμού, για την περίοδο 1961 έως και 2011, με σκοπό την κατασκευή προβαλλόμενων πινάκων επιβίωσης. Στη 4η Ενότητα μελετώνται οι επιπτώσεις που έχουν οι τάσεις μακροζωίας του ελληνικού πληθυσμού στις ατομικές προσόδους κατά την ηλικία κανονικής συνταξιοδότησης, καθώς και στην αναλογιστική υποχρέωση (και το κανονικό κόστος) των συνταξιοδοτικών προγραμμάτων για διαφορετικές αναλογιστικές μεθόδους καθορισμένης συνταξιοδοτικής παροχής. Τα συμπεράσματα της εργασίας και προτάσεις για περαιτέρω έρευνα δίνονται στην 5η Ενότητα.

2. Η ΜΕΘΟΔΟΣ LEE-CARTER ΚΑΙ Η ΑΝΑΛΥΣΗ ΧΡΟΝΟΣΕΙΡΩΝ

Το μοντέλο Lee-Carter περιγράφει μέσω τριών παραμέτρων $\{a_x, b_x, k_t\}$ τη θνησιμότητα που καταγράφεται σε κάθε ηλικιακή ομάδα του πληθυσμού ανά έτος και δίνεται από τη σχέση:

$$\ln(m_{x,t}) = a_x + b_x k_t + \varepsilon_{x,t} \text{ , ή } m_{x,t} = e^{a_x + b_x k_t + \varepsilon_{x,t}}, \quad (1)$$

όπου, $m_{x,t}$: οι παρατηρούμενοι ειδικοί κατά ηλικία δείκτες θνησιμότητας της ηλικιακής ομάδας *x* το έτος *t*

α_x: παράμετρος ηλικίας - εκφράζει τη μέση θνησιμότητα στην ηλικιακή ομάδα *x*

k_t: παράμετρος χρόνου - εκφράζει το γενικό επίπεδο θνησιμότητας του έτους *t*

b_x: παράμετρος ηλικίας - εκφράζει την απόκλιση από τη μέση θνησιμότητα, καθώς το γενικό επίπεδο θνησιμότητας αλλάζει

$\varepsilon_{x,t}$: σφάλμα - εκφράζει ειδικά αποτελέσματα ηλικίας και χρόνου, τα οποία δεν περιγράφονται από το μοντέλο.

Για την εκτίμηση των παραμέτρων του μοντέλου, χρησιμοποιούμε μία προσέγγιση της μεθόδου Διάσπασης Ιδιαζουσών Τιμών (SVD). Υποθέτοντας ότι τα σφάλματα έχουν μηδενική μέση τιμή και είναι ομοσκεδαστικά, χρησιμοποιούνται οι ακόλουθοι περιορισμοί:

$$\sum_{t=t_1}^{t_n} k_t = 0, \qquad \sum_x b_x = 1. \quad (2)$$

Διαφορίζοντας (μερικώς) τη σχέση:

$$f(a_x, b_x, k_t) = \sum_{x,t} [\ln(m_{x,t}) - a_x - b_x k_t]^2, \tag{3}$$

εξάγουμε τις εκτιμήσεις των παραμέτρων του μοντέλου:

$$\hat{a}_x = \frac{1}{h}\sum_{t=t_1}^{t_n} \ln(m_{x,t}) = \ln[\prod_{t=t_1}^{t_n} m_{x,t}^{\frac{1}{h}}], \quad h = t_n - t_1 + 1,$$

$$\hat{k}_t = \sum_x [\ln(m_{x,t}) - a_x], \quad \hat{b}_x = (\sum_t [\ln(m_{x,t}) - a_x] k_t) \Big/ \sum_t k_t^2 . \tag{4}$$

Οι Lee και Carter παρατήρησαν ότι κατά την εκτίμηση των παραμέτρων a_x, b_x και k_t, ο συνολικός αριθμός των παρατηρούμενων θανάτων δεν είναι κατ' ανάγκη ίσος με τον αριθμό θανάτων που εκτιμάται από το μοντέλο. Έτσι υπολογίζεται μια νέα εκτίμηση της παραμέτρου k_t, η $k_t^{(2)}$ έτσι ώστε:

$$\sum_x D_{x,t} = \sum_x E_{x,t} e^{(a_x + b_x k_t^{(2)})}, \tag{5}$$

όπου, $D_{x,t}$: ο αριθμός των θανάτων της ηλικιακής ομάδας x μέσα στο έτος t, $E_{x,t}$: ο εκτιθέμενος σε κίνδυνο πληθυσμός της ηλικιακής ομάδας x στο μέσον του έτους t.

Καθώς δεν υπάρχει αναλυτική λύση για την παραπάνω εξίσωση, οι Lee και Carter κατέληξαν σε μια λύση, μέσα από μια επαναληπτική διαδικασία αναζήτησης σε ένα εύρος τιμών της $k_t^{(2)}$.

Το μοντέλο Lee-Carter είναι άρρηκτα συνδεδεμένο με την ανάλυση χρονολογικών σειρών, αφού η k_t προσεγγίζεται από μια διαδικασία Αυτοπαλινδρομούμενων Ολοκληρωμένων Υποδειγμάτων Κινούμενου Μέσου με παραμέτρους *p, d* και *q* (ARIMA((*p,d,q*)), η γενική μορφή των οποίων δίνεται από την παρακάτω εξίσωση (Hyndman and Athanasopoulos, 2013):

$$(1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p)(1 - B)^d k_t = c + (1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q)\varepsilon_t, \tag{6}$$

όπου, το αριστερό μέλος αποτελείται από το AR(*p*) μέρος του μοντέλου και τις *d* - τάξης διαφορές της παραμέτρου k_t, οι οποίες δίνονται από τη σχέση:

$$\Delta^d k_t = (1 - B)^d k_t, \tag{7}$$

όπου, με *B* συμβολίζεται ο τελεστής χρονικής υστέρησης. Το δεξιό μέλος περιλαμβάνει την παράμετρο μετατόπισης *c* (drift parameter), το MA(q) μέρος του μοντέλου και τα κατάλοιπα $\{\varepsilon_t : t = 1, 2, \ldots\}$, τα οποία υποθέτουμε ότι αποτελούν μια διαδικασία λευκού θορύβου (white noise process), δηλαδή $E(\varepsilon_t) = 0$ για κάθε *t*, $V(\varepsilon_t) = \sigma^2$ για κάθε *t*, και $Cov(\varepsilon_t, \varepsilon_{t-s}) = 0$ για κάθε *t* με $1 \le s \le t - 1$.

Με εφαρμογή της μεθοδολογίας των Box και Jenkins (1976) ελέγχεται η στασιμότητα της σειράς μέσω των συναρτήσεων αυτοσυσχέτισης (autocorrelation functions), ή/και τους ελέγχους ύπαρξης μοναδιαίας ρίζας (unit root tests). Έπειτα, εκτιμώνται οι παράμετροι (*p,d,q*), οι οποίες καθορίζουν ένα σύνολο υποψήφιων μοντέλων ARIMA (*p,d,q*) και επιλέγεται εκείνο που εμφανίζει την καλύτερη προσαρμογή στα δεδομένα σε συνδυασμό με μια ικανοποιητική προβλεπτική ικανότητα.

Στη συνέχεια, πραγματοποιείται έλεγχος καλής προσαρμογής (goodness of fit) των επιλεγμένων μοντέλων. Χρησιμοποιώντας το επιλεγμένο ARIMA μοντέλο, εκτιμούμε τις προβλέψεις για την παράμετρο k_t. Υιοθετώντας το συμβολισμό s.e.(k_t) για το σφάλμα πρόβλεψης, μπορούμε να υπολογίσουμε το 95% διάστημα εμπιστοσύνης για τις προβλέψεις από τη σχέση:

$$k_t \pm 1.96 \; s.e.(k_t). \tag{8}$$

Μετά την πρόβλεψη της παραμέτρου k_t μπορούμε να προβάλλουμε τους ειδικούς κατά ηλικία δείκτες θνησιμότητας έως το έτος $t+h$ σύμφωνα με την παρακάτω σχέση:

$$\ln(\hat{m}_{x,t+h}) = \hat{a}_x + \hat{b}_x \hat{k}_{t+h}, \; h = t_n - t_1 + 1. \tag{9}$$

Παρατήρηση 1: Μετά από τη μελέτη αρκετών μοντέλων, οι Lee και Carter (1992) κατέληξαν στο συμπέρασμα ότι ο τυχαίος περίπατος με μετατόπιση (random walk with a drift parameter) είναι το καταλληλότερο μοντέλο για τα δεδομένα των Η.Π.Α. Ωστόσο, αν και θα μπορούσαν να εφαρμοστούν πολυπλοκότερα μοντέλα, στην πράξη ο τυχαίος περίπατος έχει καταλήξει να χρησιμοποιείται (σχεδόν) αποκλειστικά λόγω της απλότητας σε συνδυασμό με την αποτελεσματικότητά του.

3. ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΘΝΗΣΙΜΟΤΗΤΑΣ ΠΛΗΘΥΣΜΟΥ

Οι ειδικοί κατά ηλικία δείκτες θνησιμότητας του ελληνικού πληθυσμό, για τα έτη 1961-2011 προσαρμόστηκαν στο μοντέλο Lee-Carter, το οποίο αναλύθηκε εν συντομία στην 2η Ενότητα. Για την εφαρμογή του μοντέλου Lee-Carter χρησιμοποιήσαμε το πακέτο "demography" της γλώσσας στατιστικών υπολογισμών R (https://www.r-project.org). Στη συνέχεια, παρουσιάζονται τα αποτελέσματα εφαρμογής του Lee-Carter και των ARIMA μοντέλων του στα Ελληνικά δεδομένα.

3.1 Τα Δεδομένα

Τα ελληνικά απογραφικά δεδομένα (στατιστικά πληθυσμού και αριθμού θανάτων), αντλήθηκαν από τη βάση δεδομένων της Ευρωπαϊκής Στατιστικής Επιτροπής (Eurostat, 2012) και καλύπτουν μια περίοδο παρατήρησης 51 ετών, από το 1961 έως το 2011 (όλα τα διαθέσιμα δεδομένα για τον ελληνικό πληθυσμό). Ακολούθησε η κατηγοριοποίηση των δεδομένων ανά φύλο, για τις 19 ακόλουθες ηλικιακές ομάδες [0,1), [1,5), [5,10), …, [80,85), 85^+.

3.2 Εκτίμηση των Παραμέτρων του Μοντέλου Lee-Carter

Ο ειδικός κατά ηλικία δείκτης θνησιμότητας $m_{x,t}$ ορίζεται ως ο λόγος του αριθμού των θανάτων $D_{x,t}$ της ηλικιακής ομάδας x μέσα στο έτος t προς τον εκτιθέμενο σε κίνδυνο πληθυσμό $E_{x,t}$ της ηλικιακής ομάδας x στο μέσον του έτους t:

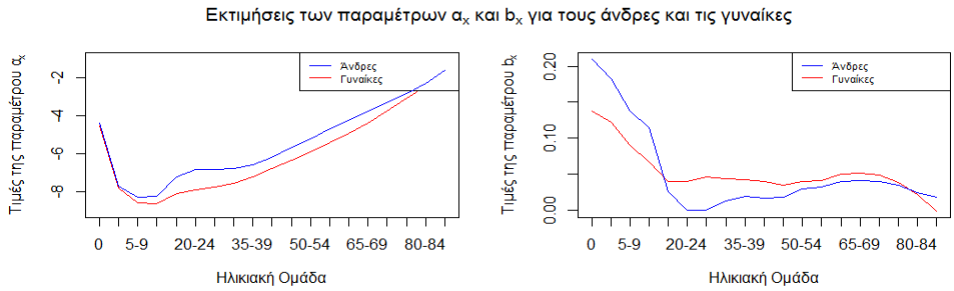
$$m_{x,t} = D_{x,t} / E_{x,t}, \tag{10}$$

όπου, x: η ηλικιακή ομάδα με x=[0,1), [1,5), [5,10),… ,[75,80), [80,85), 85^+, t: το ημερολογιακό έτος με t = 1961, 1962, …, 2010, 2011 και $D_{x,t}$, $E_{x,t}$ όπως ορίζονται αμέσως μετά τη σχέση (5).

Στο Διάγραμμα 1 παρουσιάζονται οι εκτιμήσεις των a_x και b_x. Παρατηρούμε ότι οι γραφικές παραστάσεις των a_x έχουν σε γενικές γραμμές την ίδια συμπεριφορά και για τα δύο φύλα, με τις γυναίκες να παρουσιάζουν χαμηλότερες τιμές από τους άνδρες. Ειδικότερα, στις ηλικίες [0,15) η μέση θνησιμότητα μειώνεται σημαντικά. Οι τιμές της παραμέτρου b_x δείχνουν κατά πόσο η θνησιμότητα σε κάποια ηλικιακή ομάδα

τείνει να αυξηθεί ή να μειωθεί, καθώς αλλάζει το γενικό επίπεδο θνησιμότητας k_t. Μπορούμε επίσης να δούμε, ότι στις ηλικίες [0,30) για τους άνδρες και [0,20) για τις γυναίκες, η παράμετρος b_x έχει μεγαλύτερες τιμές (με έντονα βέβαια πτωτική τάση), κάτι που δείχνει ότι τα ποσοστά θνησιμότητας σε αυτές τις ηλικίες μεταβάλλονται πιο γρήγορα, όταν αλλάζει το γενικό επίπεδο θνησιμότητας k_t.

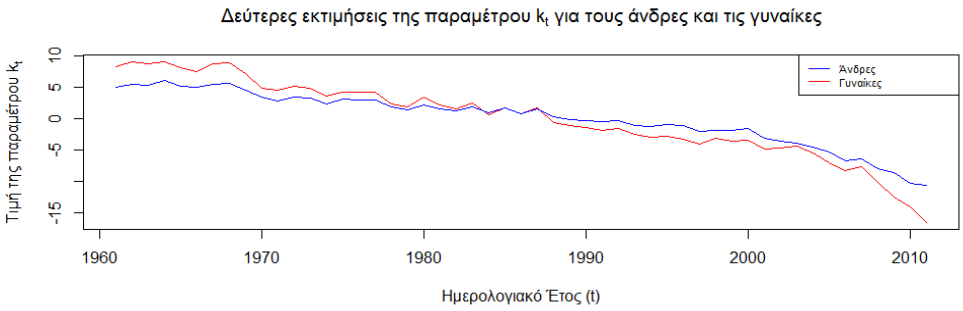
Διάγραμμα 1. Εκτιμήσεις παραμέτρων a_x και b_x για τους άνδρες και τις γυναίκες



3.2.1 Οι Χρονολογικές Σειρές των Εκτιμήσεων της Παραμέτρου

Ιδιαίτερη έμφαση δίνεται στην δεύτερη (διορθωτική) εκτίμηση (5) της παραμέτρου k_t, η οποία εμπεριέχει όλη την απαραίτητη πληροφορία για την ανάλυση της θνησιμότητας και οι τιμές της αποτελούν μία χρονολογική σειρά (Διάγραμμα 2).

Διάγραμμα 2. Δεύτερες εκτιμήσεις της παραμέτρου k_t για τους άνδρες και τις γυναίκες



Στο Διάγραμμα 2 μπορούμε εύκολα να δούμε ότι κατά την περίοδο μελέτης 1961 έως 2011, οι τιμές της παραμέτρου k_t των ανδρών μειώνονται σχεδόν γραμμικά, παρουσιάζοντας όμως μικρές αυξομειώσεις ανά διαστήματα. Απότομες αλλαγές στο γενικό επίπεδο θνησιμότητας παρατηρούνται στο διάστημα από το 1968 έως και το 1971, καθώς και μετά το 2000, με έντονα αρνητικές τιμές στα τέλη του 2010. Παρόμοια πορεία ακολουθεί και η παράμετρος k_t των γυναικών (Διάγραμμα 2).

3.3 Ανάλυση Χρονολογικών Σειρών

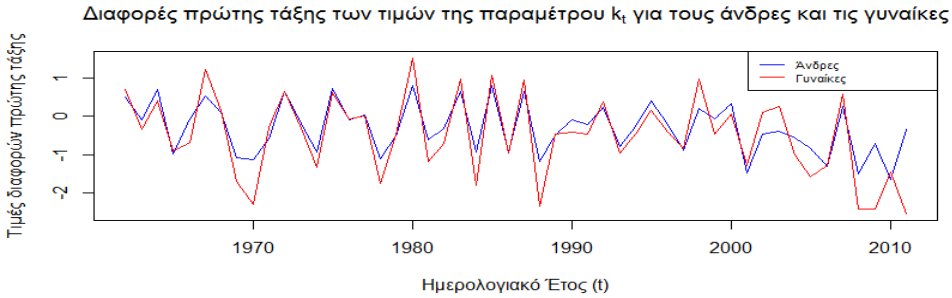
Η μεθοδολογία Box-Jenkins (1976) εφαρμόζεται στις χρονολογικές σειρές της παραμέτρου k_t και εν συνεχεία κατασκευάζουμε συνεπτυγμένους πίνακες επιβίωσης για το μελλοντικό προσδόκιμο ζωής του ελληνικού πληθυσμού.

3.3.1 Ταυτοποίηση Χρονολογικών Σειρών

Παρατηρώντας τις τιμές του δεύτερου σταδίου εκτίμησης της k_t για τους άνδρες και τις γυναίκες στο Διάγραμμα 2, μπορούμε εύκολα να διαπιστώσουμε ότι οι παρακάτω σειρές δεν είναι στάσιμες.

Οι διαφορές $1^{ης}$ τάξης (σχέση (7) για $d = 1$) εξαλείφουν σε ικανοποιητικό βαθμό τη μη-στασιμότητα των σειρών, όπως φαίνεται στο Διάγραμμα 3.

Διάγραμμα 3. Διαφορές πρώτης τάξης των σειρών της k_t για τους άνδρες και τις γυναίκες



Ο έλεγχος μοναδιαίας ρίζας των Philips-Perron καθώς και τα κριτήρια πληροφορίας AIC (Akaike), SIC (Schwarz,) οδήγησαν στην εκτίμηση των τάξεων p και q. Λόγω της έντονα πτωτικής τάσης της παραμέτρου k_t, θεωρήσαμε επίσης μια σταθερά (drift parameter), η οποία εκφράζει τη μέση ανά έτος μεταβολή των τιμών της παραμέτρου.

Στον Πίνακα 1 παρουσιάζονται τα υποψήφια μοντέλα ανδρών και γυναικών, με τις αντίστοιχες τιμές των κριτηρίων AIC και SIC.

Πίνακας 1. Υποψήφια μοντέλα ανδρών και γυναικών, με τις τιμές των AIC και SIC

Άνδρες	***AIC***	***SIC***	***Γυναίκες***	***AIC***	***SIC***
ARIMA(0,1,0) with drift	***103.48***	***107.05***	***ARIMA(0,1,0) with drift***	***146.90***	***150.47***
ARIMA(1,1,0) with drift	104.01	109.23	ARIMA(1,1,0) with drift	148.88	154.09
ARIMA(0,1,1) with drift	104.21	109.42	ARIMA(0,1,1) with drift	148.87	154.08
ARIMA(1,1,1) with drift	106.08	112.84	ARIMA(1,1,1) with drift	151.23	157.99
ARIMA(0,1,0)	113.22	115.05	ARIMA(0,1,0)	158.01	159.84

Κατά συνέπεια, το μοντέλο ARIMA(0,1,0) με μία σταθερά επιλέχθηκε για την προβολή των μελλοντικών δεικτών θνησιμότητας ανδρών και γυναικών στην Ελλάδα.

Παρατήρηση 2: Όπως είδαμε, για τα δεδομένα του ελληνικού πληθυσμού καταλήξαμε στο μοντέλο του τυχαίου περιπάτου με μετατόπιση (random walk with a drift parameter), το οποίο είναι αρκετά διαδεδομένο και χρησιμοποιείται ευρέως σε αντίστοιχες μελέτες διεθνώς.

3.3.2 Εκτίμηση των Παραμέτρων των Μοντέλων Χρονολογικών Σειρών

Σύμφωνα με τη γενική μορφή της εξίσωσης (6) για d=1, οι εκτιμηθείσες παράμετροι και οι τυπικές αποκλίσεις του ARIMA(0,1,0) με μια σταθερά δίνεται για τους άνδρες από τη σχέση:

$$k_t = -0,31 + k_{t-1} + \varepsilon_t, \tag{11}$$

όπου, $c = -0.31$ είναι η σταθερά μετατόπισης ενώ για τις γυναίκες από τη σχέση:

$$k_t = -0,49 + k_{t-1} + \varepsilon_t, \tag{12}$$

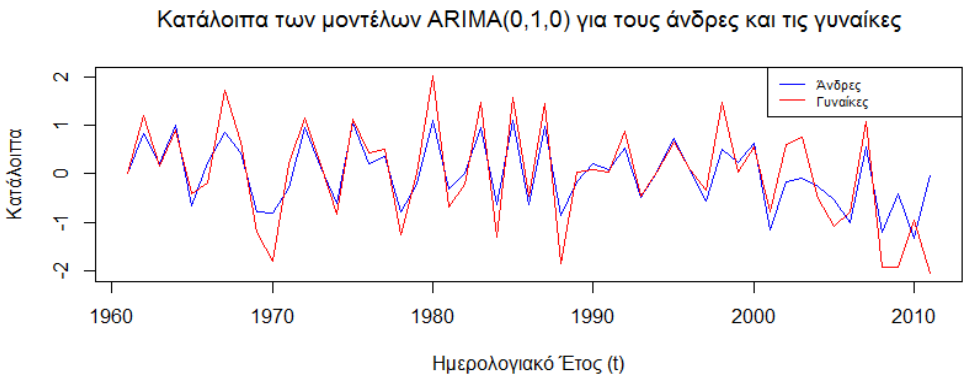
όπου, $c = -0.49$ είναι η αντίστοιχη σταθερά μετατόπισης.

Παρατήρηση 3: Ο υπολογισμός της σταθεράς c δίνεται από τη μέσο όρο των διαφορών μεταξύ των διαδοχικών τιμών της παραμέτρου k_t. Η αρνητική τιμή της σταθεράς μετατόπισης c υποδηλώνει την πτωτική τάση των τιμών της k_t που παρατηρείται κατά τα έτη προσαρμογής.

3.3.3 Διαγνωστικός Έλεγχος

Στο Διάγραμμα 4 βλέπουμε ότι κατά την προσαρμογή των μοντέλων στα δεδομένα ανδρών και γυναικών, η μέση τιμή των καταλοίπων παρατηρείται κοντά στο μηδέν. Μπορούμε επίσης να θεωρήσουμε ότι η διακύμανση των καταλοίπων παραμένει σταθερή σε όλο το εύρος των ιστορικών δεδομένων. Επιπλέον, από τα διαγράμματα αυτοσυσχετίσεων των καταλοίπων των μοντέλων ARIMA ανδρών και γυναικών, συμπεραίνουμε ότι δεν υπάρχουν ενδείξεις για μη-μηδενικές (στατιστικά σημαντικές) αυτοσυσχετίσεις στα κατάλοιπα για τις χρονικές υστερήσεις 1 έως 20.

Διάγραμμα 4. Τα κατάλοιπα των επιλεγμένων μοντέλων ARIMA για τους άνδρες και τις γυναίκες



3.3.4 Προβλέψεις

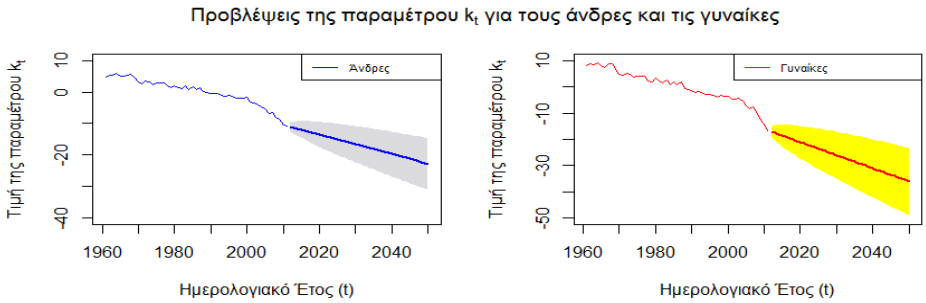
Στο σημείο αυτό, μπορούμε να εκτιμήσουμε τις μελλοντικές τιμές της k_t για τους άνδρες και τις γυναίκες έως το 2050, χρησιμοποιώντας τις σχέσεις (11) και (12), για τα δεδομένα προσαρμογής από το 1961 έως το 2011. Για τον υπολογισμό του διαστήματος εμπιστοσύνης των προβλέψεων επικεντρωθήκαμε μόνο στη μεταβλητότητα της παραμέτρου k_t, αγνοώντας άλλες πηγές σφαλμάτων. Το 95% διάστημα εμπιστοσύνης για τις προβλέψεις της παραμέτρου k_t δίνεται από τη σχέση:

$$\hat{k}_{2011+h} \pm 1.96 \; s.e.(\hat{k}_{2011+h}), \tag{13}$$

όπου, $h = 1, 2, \ldots, 39$ και $s.e.(k_{2011+h})$ υποδηλώνει το σφάλμα της πρόβλεψης.

Στο Διάγραμμα 5 παρατηρούμε ότι οι τιμές των προβλέψεων της k_t παρουσιάζουν πτωτική τάση, με ακόμα χαμηλότερες τιμές από την περίοδο προσαρμογής και για τα δύο φύλα.

Διάγραμμα 5. Προβλέψεις της k_t σε διάστημα εμπιστοσύνης 95%, για τους άνδρες και τις γυναίκες



3.3.5 Προβολή Θνησιμότητας

Μετά την πρόβλεψη της παραμέτρου θνησιμότητας k_t για τους άνδρες και τις γυναίκες μπορούμε να προβάλλουμε τους ειδικούς κατά ηλικία δείκτες θνησιμότητας έως το 2050, με τη σχέση:

$$\hat{m}_{x,2011+h} = m_{x,2011} e^{\hat{b}_x (\hat{k}_{2011+h} - k_{2011})}, \quad h = 1, 2, \ldots, 39. \tag{14}$$

Από τη σχέση (14) μπορούμε να υπολογίσουμε τους δείκτες θνησιμότητας για όλες τις ηλικιακές ομάδες x των ανδρών και των γυναικών χρησιμοποιώντας τον δείκτη θνησιμότητας για το έτος 2011 $m_{x,2011}$ και τις προβλέψεις της παραμέτρου $\{k_{2011+h} : h = 1, 2, \ldots, 39\}$. Συνδυάζοντας τις σχέσεις (1) και (13), το 95% διάστημα εμπιστοσύνης (ΔE) για τους προβαλλόμενους δείκτες θνησιμότητας δίνεται από την:

$$\Delta E(\hat{m}_{x,2011+h}) = \hat{m}_{x,2011+h} \; e^{\pm \hat{b}_x 1.96 \; s.e.(\hat{k}_{2011+h})}, \quad h = 1, 2, \ldots, 39. \tag{15}$$

3.3.6 Προβολή Πινάκων Επιβίωσης

Με τα αποτελέσματα της πρόβλεψης των μελλοντικών δεικτών θνησιμότητας (14), μπορούμε να κατασκευάσουμε συνεπτυγμένους πίνακες επιβίωσης και να υπολογίσουμε το μελλοντικό προσδόκιμο ζωής (Keyfitz, 1977), χρησιμοποιώντας τους παρακάτω συμβολισμούς:

- $[x, x + w_x)$: η ηλικιακή ομάδα, έτσι ώστε $x = x_0, x_1, \ldots, x_{k-1}$ με $x_k = 85^+$ για $k = 18$ και w_x : η διαφορά ανάμεσα στις διαδοχικές ακριβείς ηλικίες, έτσι ώστε $w_{x_i} = x_{i+1} - x_i$ με $i = 0,1,2,\ldots,k-1$
- ${}_nq_x$: η πιθανότητα θανάτου ανάμεσα στις ακριβείς ηλικίες x και $x+n$, $n = 1,4,5$
- f_x : διαχωριστικός παράγοντας που εκφράζει την ισοκατανομή των θανάτων μέσα σε κάθε ηλικιακή ομάδα. Υποθέτουμε ότι $f_x = 0.5$ για όλες τις ηλικιακές ομάδες, με εξαίρεση την [0,1) για την οποία ισχύει $f_x = 0.15$ και $f_x = 0.16$ για τους άνδρες και τις γυναίκες αντίστοιχα
- l_x : ο αριθμός των επιζώντων στην ηλικία x. Υποθέτουμε ότι $l_{x_0} = 100{,}000$
- ${}_nd_x$: ο αριθμός των ατόμων που απεβίωσαν στην ηλικιακή ομάδα $[x, x+n)$

- ${}_nL_x$: ο αριθμός ετών ζωής των ατόμων του πληθυσμού μέσα στην ηλικιακή ομάδα $[x, x+n)$
- T_x: ο αριθμός ετών ζωής που απομένει στα άτομα της ηλικιακής ομάδας $[x, x+n)$
- e_x: ο μέσος υπολειπόμενος χρόνος ζωής των ατόμων στην ηλικία x.

Πρώτα θα υπολογίσουμε την πιθανότητα θανάτου σύμφωνα με την ακόλουθη προσεγγιστική σχέση:

$$q_x \approx w_x m_x / (1 + f_x' w_x m_x), \text{ για } x = x_0,\ x_1,\ \ldots,\ x_{k-1},\ f_x' = 1 - f_x,\ m_x \equiv m_{x,t},\ q_{x_k} = 1. \quad (16)$$

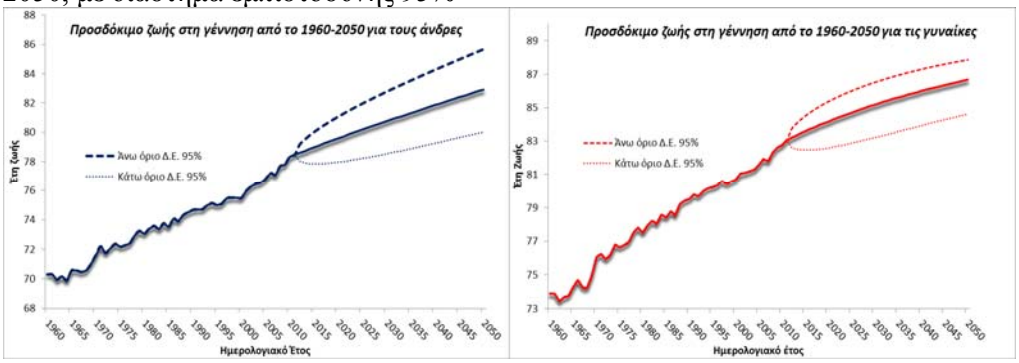
Έτσι για $x = x_0,\ x_1,\ \ldots,\ x_{k-1}$ και για κάθε έτος t, ο πίνακας επιβίωσης κατασκευάζεται χρησιμοποιώντας τις σχέσεις που ακολουθούν:

$$\begin{aligned} l_{x+w_x} &= l_x(1 - q_x),\ \text{με } l_{x_0} = 100000, \\ {}_{w_x}d_x &= l_x - l_{x+w_x} = l_x\ {}_{w_x}q_x,\ \text{με } d_{x_k} = l_{x_k}, \\ {}_{w_x}L_x &= w_x(l_x - f_x'\,{}_{w_x}d_x),\ \text{με } L_{x_k} = l_{x_k}/m_{x_k} \text{ και} \\ T_{x_i} &= \sum_{x=x_i}^{x_k} L_x. \end{aligned} \quad (17)$$

Τέλος, το προσδόκιμο ζωής στην ηλικία x_i δίνεται από την παρακάτω σχέση:

$$e_{x_i} = T_{x_i} / l_{x_i}. \quad (18)$$

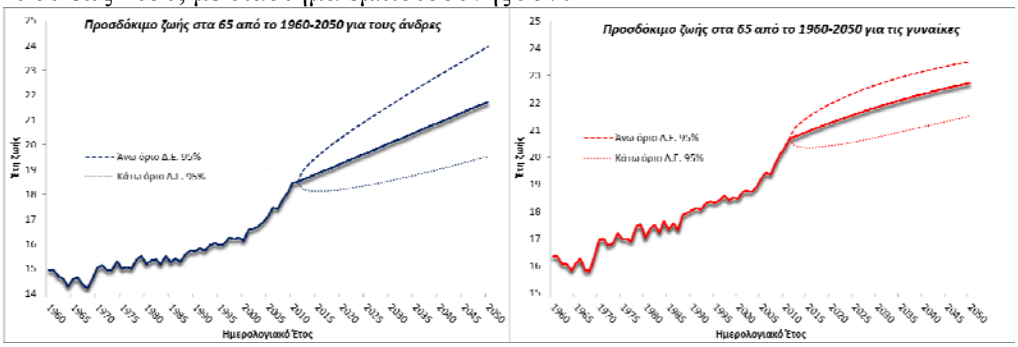
Διάγραμμα 6. Προσδόκιμο ζωής ανδρών κατά τη γέννηση για την περίοδο 1960 έως 2050, με διάστημα εμπιστοσύνης 95%



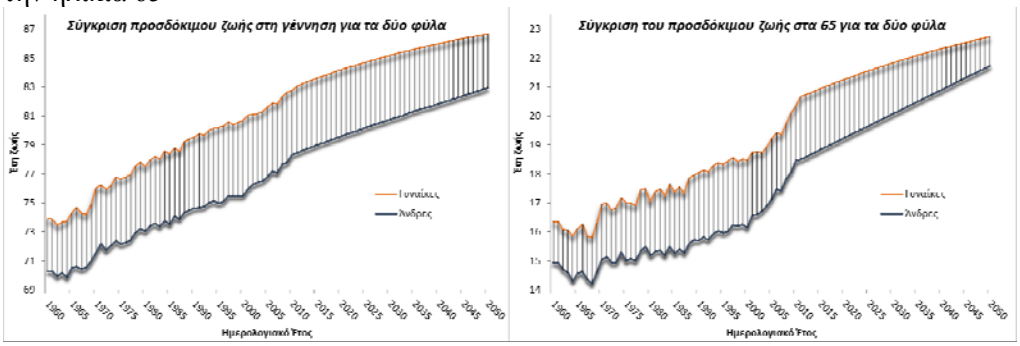
Στο Διάγραμμα 6 (αριστερά) παρατηρούμε ότι η αύξηση στο προσδόκιμο ζωής κατά τη γέννηση των ανδρών, το οποίο βρίσκεται σήμερα στα 78 περίπου έτη, αναμένεται να φτάσει και ίσως ξεπεράσει τα 83 έτη έως το 2050. Ανάλογη αύξηση παρουσιάζει και το προσδόκιμο ζωής των ανδρών κατά την ηλικία των 65 (Διάγραμμα 7, αριστερά) που σήμερα είναι 18.5 περίπου χρόνια και αναμένεται να φτάσει ή ίσως ξεπεράσει τα 22 έως το 2050. Εξίσου σημαντική φαίνεται να είναι η αύξηση στο προσδόκιμο ζωής κατά τη γέννηση των γυναικών (Διάγραμμα 6, δεξιά), καθώς κερδίζουν 4 έτη κατά την περίοδο προβολής και ενώ σήμερα βρίσκεται στα 83 περίπου έτη, αναμένεται να φτάσει και ίσως ξεπεράσει τα 87 έτη έως το 2050. Μία πιο μικρή αύξηση παρατηρείται στο προσδόκιμο ζωής των γυναικών κατά την ηλικία

των 65 (Διάγραμμα 7, δεξιά) που σήμερα είναι 21 περίπου χρόνια και αναμένεται να φτάσει στα 23 το 2050.

Διάγραμμα 7. Προσδόκιμο ζωής ανδρών κατά την ηλικία των 65 για την περίοδο 1960 έως 2050, με διάστημα εμπιστοσύνης 95%



Διάγραμμα 8. Σύγκριση προσδόκιμου ζωής ανδρών-γυναικών κατά τη γέννηση και την ηλικία 65



Συγκρίνοντας τα ιστορικά δεδομένα του προσδόκιμου ζωής στη γέννηση ανάμεσα στα δύο φύλα συμπεραίνουμε ότι οι γυναίκες παρουσιάζουν μια σημαντική υπεροχή στη μακροζωία. Η μεγαλύτερη διαφορά παρατηρήθηκε το 1996, με τις γυναίκες να ζουν κατά 5.2 χρόνια περισσότερο από τους άνδρες. Σήμερα, η διαφορά αυτή βρίσκεται στα 4.6 περίπου έτη και όπως μπορούμε να δούμε (Διάγραμμα 8-αριστερά) μειώνεται αισθητά κατά τα έτη προβολής και αναμένεται να φτάσει στα 3.7 έτη το 2050. Αντίστοιχα συμπεράσματα προκύπτουν και από τη σύγκριση του προσδόκιμου ζωής για τα δύο φύλα κατά την ηλικία των 65. Πιο συγκεκριμένα, η μεγαλύτερη διαφορά σημειώθηκε το 1996, με τις γυναίκες να ζουν 18.4 έτη επιπλέον των 65 έναντι 16 επιπλέον ετών που ζούσαν οι άνδρες αντίστοιχα, δηλαδή μια διαφορά 2.4 έτη. Στο διάγραμμα 8 (δεξιά) παρατηρούμε ότι η διαφορά αυτή μειώνεται σημαντικά, κατά την περίοδο προβολής. Ειδικότερα, μετά το 2038, η διαφορά πέφτει κάτω από τον 1.5 χρόνο και αναμένεται να είναι σχεδόν 1 χρόνος το 2050, αφού τότε οι άνδρες προβλέπεται να ζουν 21.7 έτη πέραν των 65 έναντι 22.7 των γυναικών αντίστοιχα.

4. ΟΙ ΕΠΙΠΤΩΣΕΙΣ ΜΑΚΡΟΖΩΙΑΣ ΣΤΑ ΑΣΦΑΛΙΣΤΙΚΑ ΤΑΜΕΙΑ

Η πρόβλεψη χαμηλότερων ποσοστών θνησιμότητας αποτελεί μια θετική είδηση για το ανθρώπινο είδος. Αντιθέτως, για τα συνταξιοδοτικά ταμεία καθορισμένων

παροχών ένα υψηλότερο προσδόκιμο ζωής αυξάνει το μελλοντικό κόστος, καθότι οι παροχές πρέπει να δίνονται για μεγαλύτερη περίοδο. Στις Η.Π.Α, έρευνες έχουν δείξει ότι, αύξηση του προσδόκιμου ζωής κατά 1 έτος αυξάνει τις συνταξιοδοτικές υποχρεώσεις κατά ένα ουσιαστικό οικονομικό μέγεθος της τάξης του 3 με 4%. Όταν οι άνθρωποι ζουν περισσότερο από το αναμενόμενο, έχουμε αύξηση του ρίσκου μακροζωίας, το οποίο επιδρά: α) στο κράτος, το οποίο οφείλει να παρέχει συντάξεις και ιατροφαρμακευτική περίθαλψη μέσω της φορολογίας, β) στους οργανισμούς που χορηγούν συνταξιοδοτικά και ιατροφαρμακευτικά προγράμματα, γ) σε ιδιώτες, οι οποίοι δεν βασίζονται εξ' ολοκλήρου στις παροχές του κράτους ή των επιχειρήσεων για τις οποίες εργάζονταν (Jones, 2013).

4.1 Η περίπτωση της Ελλάδας

Στην Ελλάδα, οι δικαιούχοι πλήρους σύνταξης οφείλουν να έχουν φτάσει την ηλικία των 65 ετών ή να έχουν συμπληρώσει 4500 ημέρες εργασίας έως την ηλικία αυτή. Για την περίπτωση της Ελλάδας το ρίσκο μακροζωίας επηρεάζει: α) το κράτος, το οποίο υποχρεούται να παρέχει συντάξεις και ιατροφαρμακευτική περίθαλψη στους εργαζόμενους του δημόσιου τομέα και άλλους δικαιούχους, β) το ίδρυμα κοινωνικών ασφαλίσεων (ΙΚΑ), το μεγαλύτερο οργανισμό ασφάλισης στην Ελλάδα, δεδομένου ότι καλύπτει 5,530,000 υπαλλήλους και άλλους εργαζομένους και παρέχει σήμερα, συντάξεις σε 830,000 περίπου δικαιούχους, γ) το ταμείο ελεύθερων επαγγελματιών (ΤΕΒΕ), με 545,100 ασφαλισμένους και δ) τον οργανισμό γεωργικών ασφαλίσεων (ΟΓΑ), με 1,149,000 περίπου ασφαλισμένους.

Παρακάτω, παρουσιάζουμε τις επιπτώσεις της αύξησης του προσδόκιμου ζωής στον ελληνικό πληθυσμό, αρχικά στις ατομικές προσόδους κατά την ηλικία κανονικής συνταξιοδότησης και δευτερευόντως, στην αναλογιστική υποχρέωση (και το κανονικό κόστος) των συνταξιοδοτικών προγραμμάτων, για διαφορετικές αναλογιστικές μεθόδους καθορισμένης συνταξιοδοτικής παροχής. Η εφαρμογή των μεθόδων βασίστηκε σε πλήρεις πίνακες επιβίωσης γενεών (cohort life tables) για αυτούς που γεννήθηκαν τα έτη 1960 έως 1985, χρησιμοποιώντας το μοντέλο Lee-Carter, όπως αυτό αναλύθηκε στις προηγούμενες ενότητες.

4.1.1 Πρόσοδοι κατά την Ηλικία Συνταξιοδότησης

Η αναλογιστική παρούσα αξία μιας προκαταβλητέας προσόδου, η οποία στην αρχή κάθε μήνα πληρώνει ποσό ίσο με $1/12$ $(m = 12)$ νομισματικές μονάδες σε άτομο ηλικίας r δίνεται από τη σχέση:

$$\ddot{a}_r^{(m)} = \sum_{t=0}^{\infty} {}_t p_r . v^t - \frac{m-1}{2m}, \qquad (19)$$

όπου, r είναι η ηλικία κανονικής συνταξιοδότησης (συνήθως στα 65), ${}_t p_r = l_{r+t}/l_r$ η πιθανότητα επιβίωσης ατόμου ηλικίας r μέχρι την ηλικία $r+t$ και $v = (1+i)^{-1}$, ο προεξοφλητικός παράγοντας

Στον Πίνακα 2 παρατηρούμε την αύξηση των ποσοστών στις ράντες ζωής των ανδρών, της τάξης του 0.76% κατά μέσο όρο ανά 5ετία και μια συνολική αύξηση σε ποσοστό 3.86% από τα έτη 2025 έως 2050. Τα ποσοστά αύξησης για τις γυναίκες είναι 0.57% και 2.86%, αντίστοιχα.

Πίνακας 2 Οι παρούσες αξίες των προκαταβλητέων προσόδων ζωής στην ηλικία r=65, υπολογισμένες από τους πίνακες γενεών αυτών που γεννήθηκαν τα έτη 1960 έως και 1985

Γέννηση ανδρών:	*Έτος 65^{ων} γενεθλίων ανδρών:*	$\ddot{a}_{65}^{(12)}$	*Αύξηση ποσοστών στα 65:*	
1960	**2025**	11.8134	**2025-2030**	0.70%
1965	**2030**	11.8959	**2030-2035**	0.83%
1970	**2035**	11.9951	**2035-2040**	0.79%
1975	**2040**	12.0903	**2040-2045**	0.75%
1980	**2045**	12.1815	**2045-2050**	0.72%
1985	**2050**	12.2688	***Συνολική αύξηση στα 65:***	
			2025-2050	3.86%
Γέννηση γυναικών:	*Έτος 65^{ων} γενεθλίων γυναικών:*	$\ddot{a}_{65}^{(12)}$	*Αύξηση ποσοστών στα 65:*	
1960	**2025**	12.9549	**2025-2030**	0.56%
1965	**2030**	13.0275	**2030-2035**	0.65%
1970	**2035**	13.1117	**2035-2040**	0.59%
1975	**2040**	13.1891	**2040-2045**	0.54%
1980	**2045**	13.2604	**2045-2050**	0.49%
1985	**2050**	13.3258	***Συνολική αύξηση στα 65:***	
			2025-2050	2.86%

4.1.2 Αναλογιστικές Μέθοδοι Συνταξιοδότησης

Στην υποενότητα αυτή παρουσιάζονται παραδείγματα εφαρμογής 4^{ων} βασικών μεθόδων συνταξιοδότησης καθορισμένης παροχής. Για όλες τις μεθόδους, υπολογίζονται η αναλογιστική υποχρέωση και το κανονικό κόστος, υποθέτοντας σταθερό επιτόκιο $i = 4\%$, 1 συμμετέχοντα ηλικίας 50 στο συνταξιοδοτικό πρόγραμμα, ηλικία εισαγωγής στο πρόγραμμα $e = 30$, ηλικία αποτίμησης $x = 50$ και ηλικία κανονικής συνταξιοδότησης $r = 65$. Για πιο λεπτομερείς πληροφορίες σχετικά με τις Αναλογιστικές Μεθόδους Συνταξιοδότησης, βλ. Aitken (1994).

α) Μέθοδος Πιστωτικής Μονάδας (Traditional Unit Credit-TUC): η αναλογιστική υποχρέωση στην ηλικία x (αξία των παροχών που έχει συσσωρευτεί από την ηλικία e μέχρι την ηλικία x) και το κανονικό κόστος ορίζονται ως:

$$AL_x = B_x \cdot (D_r / D_x) \cdot \ddot{a}_r^{(12)}, \; NC_x = b_x \cdot \frac{D_r}{D_x} \ddot{a}_r^{(12)}, \tag{20}$$

όπου, B_x είναι η ετήσια παροχή, πληρωτέα ανά μήνα, η οποία έχει συσσωρευτεί από την ηλικία e μέχρι την ηλικία x και $b_x = B_x/(x-e)$ είναι το μερίδιο της συνολικής παροχής σύνταξης που αντιστοιχεί σε κάθε έτος. Θεωρούμε επίσης, μια σταθερή μηνιαία παροχή 50€ για κάθε έτος υπηρεσίας. Με $D_x = l_x \cdot v^x$ συμβολίζεται η συνάρτηση μετατροπής στην ηλικία x και υπολογίζεται από τον πίνακα επιβίωσης γενεών.

β) Μέθοδος Προβαλλόμενης Πιστωτικής Μονάδας (Projected Unit Credit-PUC): η αναλογιστική υποχρέωση στην ηλικία x ορίζεται όπως στη σχέση (20) και το κανονικό κόστος όπως στην (21), με τη διαφορά ότι, η συνταξιοδοτική παροχή B_x υπολογίζεται ως το 3% του μέσου ετήσιου μισθού των τελευταίων 3 ετών, για κάθε έτος υπηρεσίας. Επίσης, υποθέτουμε ετήσιο μισθό 12,000€ κατά την εισαγωγική ηλικία και ετήσια αύξηση $s = 3\%$.

Πίνακας 3 *AL* και *NC* στην ηλικία αποτίμησης 50, για τις αναλογιστικές μεθόδους συνταξιοδότησης *TUC* και *PUC*, χρησιμοποιώντας πίνακες επιβίωσης γενεών, για άνδρες και γυναίκες που γεννήθηκαν τα έτη 1960 έως και 1985

Άνδρες	***Μέθοδος Πιστωτικής Μονάδας***				***Μέθοδος Προβ/μενης Πιστ/ής. Μονάδας***			
Γέννηση:	AL_{50}	***Αύξηση***	NC_{50}	***Αύξηση***	AL_{50}	***Αύξηση***	NC_{50}	***Αύξηση***
1960	70,082 €	-	3,504 €	-	111,561 €	-	5,578 €	-
1965	70,897 €	1.16%	3,545 €	1.17%	112,858 €	1.16%	5,643 €	1.17%
1970	71,859 €	1.36%	3,593 €	1.35%	114,390 €	1.36%	5,720 €	1.36%
1975	72,786 €	1.29%	3,639 €	1.28%	115,866 €	1.29%	5,793 €	1.28%
1980	73,679 €	1.23%	3,684 €	1.24%	117,287 €	1.23%	5,864 €	1.23%
1985	74,539 €	1.17%	3,727 €	1.17%	118,656 €	1.17%	5,933 €	1.18%
1960-1985	***Συνολική Αύξηση***		***Συνολική Αύξηση***		***Συνολική Αύξηση***		***Συνολική Αύξηση***	
	6.36%		6.36%		6.36%		6.36%	
Γυναίκες	***Μέθοδος Πιστωτικής Μονάδας***				***Μέθοδος Προβ/μενης Πιστ/ής. Μονάδας***			
Γέννηση:	AL_{50}	***Αύξηση***	NC_{50}	***Αύξηση***	AL_{50}	***Αύξηση***	NC_{50}	***Αύξηση***
1960	82,557 €	-	4,128 €	-	131,419 €	-	6,571 €	-
1965	83,273 €	0.87%	4,164 €	0.87%	132,560 €	0.87%	6,628 €	0.87%
1970	84,116 €	1.01%	4,206 €	1.01%	133,901 €	1.01%	6,695 €	1.01%
1975	84,892 €	0.92%	4,245 €	0.93%	135,137 €	0.92%	6,757 €	0.93%
1980	85,608 €	0.84%	4,280 €	0.82%	136,276 €	0.84%	6,814 €	0.84%
1985	86,266 €	0.77%	4,313 €	0.77%	137,324 €	0.77%	6,866 €	0.76%
1960-1985	***Συνολική Αύξηση***		***Συνολική Αύξηση***		***Συνολική Αύξηση***		***Συνολική Αύξηση***	
	4.49%		4.48%		4.49%		4.48%	

Στις τιμές της αναλογιστικής υποχρέωσης και του κανονικού κόστους των ανδρών και για τις 2 μεθόδους του Πίνακα 3 παρατηρείται μία μέση αύξηση ανά πενταετία της τάξης του 1.24% και μία συνολική αύξηση σε ποσοστό 6.36% από το έτος 1960 έως το 1985. Οι τιμές για τις γυναίκες είναι 0.88% και 4.49%, αντίστοιχα.

γ) Μέθοδος Κανονικής Εισαγωγικής Ηλικίας σε Επίπεδο Ευρώ (Entry age Normal Level Euro-EAN): Το κανονικό κόστος και η αναλογιστική υποχρέωση για τη μέθοδο EAN δίνονται από τις σχέσεις:

$$NC_e \cdot \ddot{a}_{e:\overline{r-e}|} = B_r \cdot \frac{D_r}{D_e} \ddot{a}_r^{(12)}, \tag{21}$$

$$AL_x = NC_e \cdot \ddot{s}_{e:\overline{x-e}|} = NC_e \cdot \frac{N_e - N_x}{D_x}, \tag{22}$$

όπου, B_r είναι η ετήσια παροχή ατόμου ηλικίας x που έχει συσσωρευτεί από την ηλικία e μέχρι την ηλικία r, $\ddot{a}_{e:\overline{r-e}|}$ είναι η αναλογιστική παρούσα αξία στην ηλικία e, μίας νομισματικής μονάδας, πληρωτέας στην αρχή των r-e ετών, $\ddot{s}_{e:\overline{x-e}|}$ είναι η συσσωρευμένη αξία από την ηλικία e στην ηλικία x και $N_x = \sum_{t=0}^{\infty} l_{x+t} \cdot v^{x+t}$ η αντίστοιχη συνάρτηση μετατροπής στην ηλικία x. Και εδώ, όπως στην μέθοδο **TUC**, θεωρούμε μία σταθερή μηνιαία παροχή των 50€ για κάθε έτος υπηρεσίας.

δ) Μέθοδος Κανονικής Εισαγωγικής Ηλικίας (Το Κανονικό Κόστος ως Ποσοστό του Μισθού): Εδώ, όπως και στην μέθοδο **PUC**, η παροχή B_r υπολογίζεται ως το 3% του μέσου ετήσιου μισθού των τελευταίων 3 ετών, για κάθε έτος υπηρεσίας. Επίσης, υποθέτουμε ετήσιο μισθό 12,000€ κατά την εισαγωγική ηλικία

και ετήσια αύξηση s=3%. Το κανονικό κόστος και η αναλογιστική υποχρέωση δίνονται από τις σχέσεις:

$$NC_e \cdot \ddot{a}^{s}_{\overline{e:r-e|}} = U \cdot S_e \cdot \ddot{a}^{s}_{\overline{e:r-e|}} = B_r \cdot \frac{D_r}{D_e} \ddot{a}^{(12)}_r , \tag{23}$$

$$AL_x = NC_x \cdot \ddot{s}^{s}_{\overline{e:x-e|}} = NC_x \cdot \frac{N^s_e - N^s_x}{D^s_x} , \tag{24}$$

όπου, $\ddot{a}^{s}_{\overline{e:r-e|}}$ είναι η βασισμένη στο μισθό ράντα από την ηλικία e έως την ηλικία r και $\ddot{s}^{s}_{\overline{e:x-e|}}$ είναι η βασισμένη στο μισθό συσσωρευμένη ράντα ζωής από την ηλικία e στην ηλικία x. Οι D^s_x και N^s_x είναι οι βασισμένες στο μισθό συναρτήσεις μετατροπής.

Πίνακας 4 *AL* και *NC* στην ηλικία 50 για *EAN* - Επίπεδο Ευρώ και *EAN* - % μισθού

Άνδρες:	***Μέθοδος EAN (Επίπεδο Ευρώ)***				***Μέθοδος EAN (Κόστος ως % μισθού)***			
Γέννηση	AL_{50}	***Αύξ.***	NC_{50}	***Αύξ.***	AL_{50}	***Αύξ.***	NC_{50}	***Αύξ.***
1960	90,874 €	-	2,854 €	-	122,690 €	-	5,392 €	-
1965	91,865 €	1.09%	2,889 €	1.23%	123,988 €	1.06%	5,455 €	1.17%
1970	93,042 €	1.28%	2,930 €	1.42%	125,531 €	1.24%	5,531 €	1.39%
1975	94,184 €	1.23%	2,969 €	1.33%	127,034 €	1.20%	5,602 €	1.28%
1980	95,290 €	1.17%	3,006 €	1.25%	128,492 €	1.15%	5,670 €	1.21%
1985	96,357 €	1.12%	3,041 €	1.16%	129,900 €	1.10%	5,734 €	1.13%
	Συν. Αύξ.		***Συν. Αύξ.***		***Συν. Αύξ.***		***Συν. Αύξ.***	
1960-1985	6.03%		6.55%		5.88%		6.34%	
Γυναίκες:	***Μέθοδος EAN (Επίπεδο Ευρώ)***				***Μέθοδος EAN (Κόστος ως % μισθού)***			
Γέννηση	AL_{50}	***Αύξ,***	NC_{50}	***Αύξ.***	AL_{50}	***Αύξ.***	NC_{50}	***Αύξ.***
1960	105,995 €	-	3,381 €	-	142,417 €	-	6,350 €	-
1965	106,869 €	0.82%	3,411 €	0.89%	143,563 €	0.80%	6,405 €	0.87%
1970	107,900 €	0.96%	3,445 €	1.00%	144,914 €	0.94%	6,468 €	0.98%
1975	108,844 €	0.87%	3,478 €	0.96%	146,147 €	0.85%	6,528 €	0.93%
1980	109,707 €	0.79%	3,509 €	0.89%	147,272 €	0.77%	6,583 €	0.84%
1985	110,503 €	0.73%	3,536 €	0.77%	148,310 €	0.70%	6,634 €	0.77%
	Συν. Αύξ.		***Συν. Αύξ.***		***Συν. Αύξ.***		***Συν. Αύξ.***	
1960-1985	4.25%		4.58%		4.14%		4.47%	

Στον Πίνακα 4 παρατηρούμε τα εξής: α) Για τη μέθοδο κανονικής εισαγωγικής ηλικίας σε επίπεδο ευρώ, έχουμε μια μέση αύξηση της αναλογιστικής υποχρέωσης κατά 1,18% ανά πενταετία και μια συνολική αύξηση της τάξης του 6,03% από το έτος 1960 έως το 1985, για τους άνδρες. Οι αυξήσεις για τις γυναίκες είναι 0.84% και 4.25%, αντίστοιχα. Για την ίδια μέθοδο, έχουμε μια μέση αύξηση του κανονικού κόστους κατά 1,28% ανά πενταετία και μια συνολική αύξηση της τάξης του 6,55% από το έτος 1960 έως το 1985, για τους άνδρες και οι αντίστοιχες αυξήσεις για τις γυναίκες είναι 0,9% και 4,58%. β) Για τη μέθοδο κανονικής εισαγωγικής ηλικίας, με το κόστος ως ποσοστό του μισθού, βλέπουμε μια μέση αύξηση της αναλογιστικής υποχρέωσης κατά 1,15% κάθε 5 χρόνια και μια συνολική αύξηση της τάξης του 5,88% από το έτος 1960 έως το 1985, για τους άνδρες ενώ οι αντίστοιχες τιμές για τις γυναίκες είναι 0,81% και 4,14%. Για την ίδια μέθοδο, έχουμε μια μέση αύξηση του κανονικού κόστους κατά 1,24% κάθε 5 χρόνια και μια συνολική αύξηση της τάξης του 6,34% από το έτος 1960 έως το 1985, για τους άνδρες με τις αντίστοιχες αυξήσεις για τις γυναίκες να είναι 0,88% και 4,47%.

Σκοπός των παραπάνω αριθμητικών εφαρμογών είναι να δείξουμε πώς η αύξηση του προσδόκιμου ζωής, σε σχέση με την αύξηση του μισθού, επιδρά στις αναλογιστικές υποχρεώσεις και τα αναλογιστικά κόστη των σπουδαιότερων αναλογιστικών μεθόδων αποτίμησης. Μία πλήρης πρόβλεψη του συνταξιοδοτικού κόστους για ένα συγκεκριμένο ταμείο απαιτεί περισσότερη ανάλυση και μία σειρά από ετήσιες αποτιμήσεις, λαμβάνοντας υπόψη και άλλους παράγοντες, όπως τα μελλοντικά ποσοστά αναπηρίας και απόσυρσης από την εργασία των ενεργών υπαλλήλων, καθώς και ο αριθμός των νέων εργαζομένων που θα προσληφθούν στο μέλλον, το ποσοστό του πληθωρισμού, οι αποδόσεις των επενδύσεων, κ.λ.π. (Winklevoss, 1976).

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η προβολή της θνησιμότητας έως το 2050 βασίστηκε στα διαθέσιμα δημογραφικά δεδομένα των προηγούμενων 50 ετών και τα αποτελέσματα μας δείχνουν ότι η θνησιμότητα θα παρουσιάσει αξιόλογη μείωση τα επόμενα χρόνια, με άμεση συνέπεια την αύξηση του προσδόκιμου ζωής και για τα δύο φύλα. Παρά το γεγονός ότι, το μοντέλο Lee-Carter έχει εφαρμοστεί για τις περισσότερες, ανεπτυγμένες χώρες, δεν είχε εφαρμοστεί μέχρι σήμερα για την πρόβλεψη της θνησιμότητας στην Ελλάδα. Ένας από τους βασικούς στόχους της έρευνας αυτής ήταν να καλυφθεί το βιβλιογραφικό κενό, της εφαρμογής του μοντέλου στα ελληνικά δεδομένα.

Στη εργασία μας, επίσης παρουσιάστηκαν οι επιπτώσεις της μακροζωίας του ελληνικού πληθυσμού στις ράντες ζωής των ατόμων κατά την ηλικία κανονικής συνταξιοδότησης και στην αναλογιστική υποχρέωση (και το κανονικό κόστος) των συνταξιοδοτικών προγραμμάτων διαφόρων αναλογιστικών μεθόδων καθορισμένης συνταξιοδοτικής παροχής. Η εφαρμογή των μοντέλων, με βάση τους προβαλλόμενους πίνακες επιβίωσης έδειξε ουσιαστική αύξηση των ατομικών προσόδων και κατά συνέπεια αύξηση των αναλογιστικών υποχρεώσεων των συνταξιοδοτικών ταμείων. Λόγω της δομής των συνταξιοδοτικών σχημάτων στην Ελλάδα, το κράτος κατέχει έναν σημαντικό ρόλο για τη διαχείριση του ρίσκου μακροζωίας. Πιθανές λύσεις για την αντιμετώπιση αυτού του προβλήματος είναι: $1^{ον}$) η ανάπτυξη ενός συστήματος αξιολόγησης της μακροζωίας από το κράτος που θα χρησιμοποιείται από όλους τους φορείς συνταξιοδοτικής ασφάλισης, $2^{ον}$) η άμεση αναθεώρηση των υποθέσεων για το προσδόκιμο ζωής από τα συνταξιοδοτικά ταμεία, $3^{ον}$) η σταδιακή αύξηση της έκδοσης ομολόγων μακράς διάρκειας και $4^{ον}$) η ανάπτυξη και εφαρμογή εσωτερικών μοντέλων (ή μερικώς εσωτερικών μοντέλων) ελέγχου του ρίσκου μακροζωίας, στα πλαίσια του συστήματος Solvency II (CEIOPS, 2007). Τέλος, η Εποπτική Αρχή της Ελλάδας θα μπορούσε να δημιουργήσει μια ερευνητική ομάδα για τη βελτίωση του υπάρχοντος μοντέλου προσομοίωσης του ρίσκου μακροζωίας, το οποίο θα μετράει α) το συστηματικό κίνδυνο που σχετίζεται με τις ετήσιες αναθεωρημένες προβλέψεις του προσδόκιμου ζωής και β) το μη συστηματικό κίνδυνο που προέρχεται από τις διακυμάνσεις μεταξύ ατόμων διαφορετικών κοινωνικό-οικονομικών στρωμάτων και ομάδων υγείας.

ABSTRACT

Worldwide, the 20th century brought significant changes in mortality rates at all ages, for both males and females. In Greece, the number of elderly was increased as a result of the increase of life expectancy and the declining trends in fertility rates, which imply an additional cost for the social security programs, including pensions and medical care. This paper presents how the past mortality rates may evolve, over the next 40 years. The Lee-Carter method and its ARIMA models incorporated for the first time with Greek data and analyzed. The impact of longevity improvements in Greek population to actuarial liabilities (and normal cost), for different actuarial pension plan methods, with defined retirement benefits, is also presented. Conclusions are drawn about future mortality rates in Greece and the important role the government must play in order to manage, in addition to other risks, the longevity risk.

Keywords: Mortality Projection, Lee-Carter Model, Life Tables, Longevity Risk, Pension Plans.

ΑΝΑΦΟΡΕΣ

Aitken, H. (1994). *A problem-solving approach to pension funding and valuation.* ACTEX Publication, Winsted, Connecticut.

Box, G. E. P. and Jenkins, G. M. (1976). *Time series analysis for Forecasting and Control.* Revised ed., San Francisco: Holden-Day.

CEIOPS (2007). QIS3. *Calibration of the underwriting risk, market risk and MCR*.

Eurostat (2012). *European Commission.* Available at: ec.europa.eu/eurostat/data.

Hyndman, R. J. and Athanasopoulos, G. (2013). *Forecasting: principles and practice.* OTexts. Available at: https://www.otexts.org/fpp.

Jones, G. (2013). *Longevity Risk and Reinsurance. Society of Actuaries*. Available at: https://www.soa.org. Accessed 10 December 2013.

Keilman, N. (1998). How Accurate Are The United Nations World Population Projections? *Population and Development Review*, **24**, 15-41.

Keyfitz, N. (1977). *Introduction to the Mathematics of Population with revisions.* Reading, Mass: Addison-Wesley Pub. Co.

Kisser, M., Kiff, J., Oppers, E. S. and Soto, M. (2012). *The Impact of Longevity Improvements on US Corporate Defined Benefit Pension Plans.* International Monetary Fund.

Lee, R. and Carter, L. (1992). Modeling and forecasting US sex differentials in mortality. *International Journal of Forecasting*, **8**, 393-411.

Stoto, M. A. (1983). The Accuracy of Population Projections. *Journal of the American Statistical Association*, **78**, 13-20.

Winklevoss, H.E. (1993). *Pension Mathematics with Numerical Illustrations.* 2nd ed. University of Pennsylvania Press, Philadelphia.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής (2015), σελ. 268-281

Η ΓΡΑΜΜΗ ΤΟΥ ΧΡΟΝΟΥ ΤΗΣ ΣΤΑΤΙΣΤΙΚΗΣ ΣΤΗΝ ΕΛΛΑΔΑ

Τ. Παπαϊωάννου
Πανεπιστήμια Ιωαννίνων και Πειραιώς
takpap@unipi.gr

ΠΕΡΙΛΗΨΗ

Στατιστική είναι συλλογή, επεξεργασία δεδομένων και παρουσίαση του τι μπορούν να μας πουν οι αριθμοί, τα δεδομένα. Περιλαμβάνει και τις πιθανότητες διότι στηρίζεται σ' αυτές. Η παρούσα εργασία αποτελεί ένα χρονολόγιο της Στατιστικής στην Ελλάδα από την αρχαιότητα μέχρι σήμερα.

1. ΕΙΣΑΓΩΓΗ

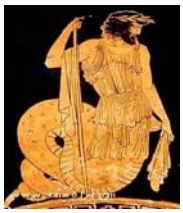
Η παρούσα εργασία αναφέρεται στην ιστορία της Στατιστικής στην Ελλάδα. Είναι συνέχεια της περσινής εργασίας μας που παρουσιάστηκε στο συνέδριο του Ελληνικού Στατιστικού Ινστιτούτου (Ε.Σ.Ι.) στη Θεσσαλονίκη και δημοσιεύτηκε στα πρακτικά του συνεδρίου (Πρακτικά 27ου Πανελλήνιου Συνέδριου Στατιστικής (2014), σελ. 200-215).

Όπως και η περσινή εργασία αποτελείται από χρονολογικά μηνύματα (bullets) που δείχνουν την ιστορική πορεία της Στατιστικής στην Ελλάδα. Είναι ένα χρονολόγιο της Στατιστικής όπως υποδηλώνεται και από τη 'γραμμή του χρόνου'. Πολλοί πιστεύουν ότι για να είσαι καλός στατιστικός πρέπει να γνωρίζεις την ιστορία της Στατιστικής. Ο Goethe είπε: «Η ιστορία μιας επιστήμης είναι η επιστήμη η ίδια» Αποτελείται από τρείς ενότητες: α) Από την αρχαιότητα μέχρι το 1453 μ.Χ., β) 1453 – 1900 και γ) 1900 – 2015 Σύγχρονη εποχή.

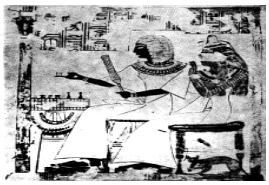
2. ΑΠΟ ΤΗΝ ΑΡΧΑΙΟΤΗΤΑ ΜΕΧΡΙ ΤΟ 1453 μ.Χ.

Μελετώντας τα αρχαία κείμενα, τους αρχαίους συγγραφείς βρίσκει κανείς αρκετές περιγραφές που αναφέρονται σε στατιστικές έννοιες και ποσότητες. Επιλεκτικά αναφέρουμε έξι περιπτώσεις. Είναι επίσης γνωστό ότι η σύγχρονη έννοια της πιθανότητας δεν είχε καλλιεργηθεί από τους αρχαίους Έλληνες αν και η έννοια του τυχαίου, του εικότος, ήταν γνωστή. Ομοίως και η Στατιστική, που είναι η

επιστήμη της εμπειρικής γνώσης στηριζόμενη στα Μαθηματικά, δεν είχε καλλιεργηθεί στην Αρχαία Ελλάδα.



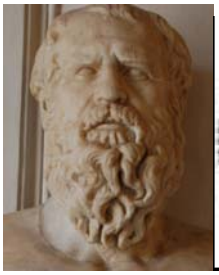
• **16ος αιώνας π.Χ.** Ο **Κέκροπας**, ο μυθικός πρώτος Βασιλιάς των Αθηνών, έκανε μια καταγραφή των υπηκόων του η οποία είχε όλα τα χαρακτηριστικά απογραφής πληθυσμού: Διέταξε κάθε άτομο να ρίξει μια και μόνο πέτρα και μετρώντας τις πέτρες βρήκε 20000 κατοίκους. Η πρώτη (?) απογραφή πληθυσμού στην Ιστορία.

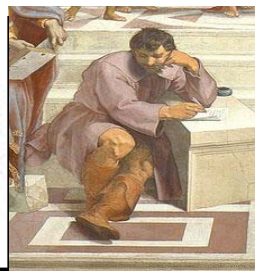


Oriental Institute, U. of Chicago

• **Αίγυπτος, αστράγαλος:** Αιγύπτιος ευγενής παίζει επιτραπέζιο παιχνίδι με αστράγαλο, πρόδρομο του ζαριού. Πηγή: David, F. N. (1962) *Games Gods and Gambling:* The origins and history of probability and statistical ideas from the earliest times to the Newtonian era, Ch. Griffin Ltd, London (επιτάφια ζωγραφική).

• **575-475 π. Χ., Ηράκλειτος:** 'πάντα ρει', μεταφυσική, επιστημολογία





Ηράκλειτος από τον Michaelangelo

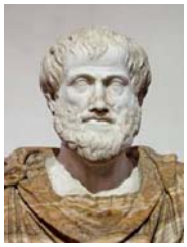
Ο Ηράκλειτος είναι ένας από τους μυστικοπαθείς φιλοσόφους γνωστός για το «πάντα ρεί» και τις ιδέες του στη μεταφυσική και την επιστημολογία. Ενέπνευσε τον Ιδρυτή του Ελληνικού Στατιστικού Ινστιτούτου (Ε.Σ.Ι.) να τον χρησιμοποιήσει ως λογότυπο-έμβλημά του.

Οπλιτοδρομία 550 π.Χ.

• **450 π. Χ.- Μέση τιμή:** Ο Ιππίας της αρχαίας Ηλείας χρησιμοποιεί τη **μέση διάρκεια (μέση τιμή** του χρόνου διάρκειας) μιας βασιλείας για να προσδιορίσει την ημερομηνία των πρώτων Ολυμπιακών Αγώνων, περίπου 300χρόνια πριν από την εποχή του. Γνώριζε πόσοι βασιλείς είχαν προηγηθεί.

• **431-404 π.Χ.-Θουκυδίδης:** Οι επιτιθέμενοι στις Πλαταιές κατά τον Πελοποννησιακό Πόλεμο (Θουκυδίδου Ιστορία) υπολογίζουν το ύψος των τειχών μετρώντας τον αριθμό (το πλήθος) των σειρών από τούβλα. Η μέτρηση

επαναλαμβάνεται αρκετές φορές από διαφορετικούς στρατιώτες. Η πλέον **συχνή τιμή (mode)** λαμβάνεται ως η πιο πιθανή. Πολλαπλασιάζοντας την τιμή αυτή με το ύψος του τούβλου επέτρεπε στους επιτιθέμενους να υπολογίζουν το μήκος των σκαλών που χρειάζονταν για να αναρριχηθούν στα τείχη. Στον Θουκυδίδη, επίσης, βρίσκεται και η αρχή του minmax στη Θεωρία Αποφάσεων.



• **384-322 π. Χ. - Αριστοτέλης**: Το βιβλίο Μετεωρολογικά του Αριστοτέλη περιέχει πλήθος ακριβών και εμπειρικών σεισμικών δεδομένων τα οποία χρησιμοποίησε, προφανώς με πρωτόλειο στατιστικό τρόπο, για να στηρίξει τη θεωρία του για τους σεισμούς. Ο Αριστοτέλης είναι επίσης γνωστός για τον πρώτο (?) ορισμό του *εύρους* των παρατηρήσεων και της εκτίμησης της *μέσης τιμής* από τις ακραίες παρατηρήσεις max και min.

Κατά την περίοδο του Μ. Αλεξάνδρου, την ελληνιστική περίοδο στην Αλεξάνδρεια, και τους αιώνες της ρωμαϊκής και βυζαντινής αυτοκρατορίας δεν καταχωρείται καμία αξιόλογη αναφορά στατιστικού περιεχομένου.

3. 1453 -1900

Την περίοδο αυτή υπάρχουν πολλές απογραφές κυρίως για τον πληθυσμό, την κτηνοτροφία κλπ. Μνημονεύουμε μόνο μία και ένα βιβλίο.

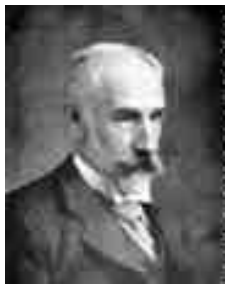
• **1895**: **Η Τουρκική Στατιστική της Ηπείρου στο Σαλμανέ του 1895** είναι μια απογραφή που απαριθμεί αναλυτικά και λεπτομερειακά το συνολικό πληθυσμό της Ηπείρου κατά πόλεις, χωριά, κατοικίες (χανέδες), φύλο κλπ πριν την κατάλυση της οθωμανική εξουσίας στον Α Βαλκανικό πόλεμο του 1912. Η Στατιστική έχει δημοσιευθεί σε τουρκική γλώσσα και οθωμανική γραφή στο 7^{ο} τεύχος του επίσημου Σαλμανέ με κόστος 12 γρόσια. Το Σαλμανέ είναι διοικητική επετηρίδα που εξέδιδαν σε τακτά διαστήματα οι αρχές του βιλαετίου. Πηγή: Κοκολάκης, Μ. (2003). *Το Ύστερο Γιαννιώτικο Πασαλίκι*: Χώρος, διοίκηση και πληθυσμός στην Τουρκοκρατούμενη Ήπειρο(1820-1913), Εθνικό Ίδρυμα Ερευνών, σελ. 1-552.

• **1895- Παλαιότερο ελληνικό βιβλίο Στατιστικής Block-Σακελλάριος:** Σακελλάριος, Πολύβιος Α. (1895) *Εγχειρίδιο Θεωρητικής και Πρακτικής Στατιστικής, (*Μετάφραση Γαλλικού βιβλίου του Maurice Block, Τύποις και Αναλώμασι Π.Δ. Σακελλαρίου, Εν Αθήναις, 1895. Πρόκειται για βιβλίο δημοσιονομικής στατιστικής και δημογραφίας. Περιέχει και ψήγματα θεωρίας (στατιστικοί νόμοι, ο Νόμος των Μεγάλων Αριθμών (ΝΜΑ) και οι μέσοι αριθμοί, πίνακες θνητότητας και μέση ζωή (Από τον Πρόλογο του μεταφραστή:
«Μετά την παλιγγενεσίαν της Ελλάδος οι μετά του αειμνήστου βασιλέως Όθωνος κατελθόντες Γερμανοί μετά πολλών άλλων αγαθών, άτινα μετέδοσαν εις την χώραν,

ίδρυσαν και εν τω Υπουργείω των Εσωτερικών **γραφείον στατιστικόν** αποτελούν μέρος του τμήματος της Δημόσιας Οικονομίας »

Αφιέρωση: ΤΩ ΕΜΩ ΠΑΤΡΙ **ΑΘΑΝΑΣΙΩ Α. ΣΑΚΕΛΛΑΡΙΩ** ΔΗΜΟΣΙΑ ΤΕ ΟΥΚ ΟΛΙΓΑ ΥΠΕΡ ΤΗΣ ΕΚΠΑΙΔΕΥΣΕΩΣ ΤΗΣ ΕΛΛΗΝΙΚΗΣ ΝΕΟΤΗΤΟΣ ΣΥΝΕΝΕΓΚΟΝΤΙ ΚΑΙ ΔΙ' ΑΔΡΩΝ ΣΥΓΓΡΑΜΜΑΤΩΝ ΤΗΝ ΕΛΛΗΝΙΚΗΝ ΕΠΙΣΤΗΜΗΝ ΘΕΡΑΠΕΥΣΑΝΤΙ ΚΑΙ ΙΔΙΑ ΥΠΕΡ ΤΗΣ ΜΟΡΦΩΣΕΩΣ ΤΩΝ ΕΑΥΤΟΥ ΤΕΚΝΩΝ ΠΟΛΛΑ ΠΟΝΗΣΑΝΤΑ ΤΗΝ ΜΕΤΑΦΡΑΣΙΝ ΤΗΝΔΕ ΕΥΓΝΟΜΩΝΟΣ ΑΝΑΤΙΘΗΜΙ

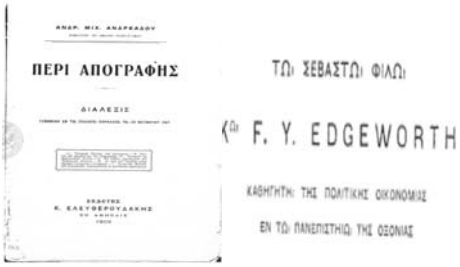
4. 1900-2015 ΣΥΓΧΡΟΝΗ ΕΠΟΧΗ



• **1906 -- Ανδρέας Ανδρεάδης (1876-1935)**: Το **1906** ο Α. Ανδρεάδης μνημονεύεται ως Καθηγητής Πολιτικής Οικονομίας και **Στατιστικής** ή Δημόσιας Οικονομικής στο Πανεπιστήμιο Αθηνών. Ήταν Ακαδημαϊκός, Ιστορικός της Οικονομίας και της Λογοτεχνίας και μελετητής της Δημοσιονομικής Ιστορίας του Ελληνισμού και πέτυχε διεθνή αναγνώριση για το πρωτοπόρο κολοσσιαίο επιστημονικό του έργο. Επίσης παρουσίασε αξιόλογο δημοσιογραφικό και συγγραφικό έργο περί την λογοτεχνία και το θέατρο.

Ο Ανδρεάδης είχε πλούσιο έργο, μεγάλο μέρος του οποίου αφορά στη δημογραφία και τη στατιστική π.χ. δημόσια οικονομία και στατιστική, πληθυσμός και πλούτος της Κωνσταντινούπολης κατά τους μέσους αιώνες, η κρίση του υπερπληθυσμού εν Αγγλία, le montant du budget de l'Etat Athenien au V et IV Siecles (*XX Session de l'Institut International de Statistique*, Madrid 1931) κλπ. Το 1916 υπήρχε Ελληνική Εταιρεία Πολιτικής Οικονομίας και Στατιστικής.

Το σκέλος της στατιστικής είναι απογραφικό, περιγραφικό και όχι συμπερασματολογικό με την έννοια της σημερινής στατιστικής συμπερασματολογίας


ΠΕΡΙ ΑΠΟΓΡΑΦΗΣ

ΤΩι ΣΕΒΑΣΤΩι ΦΙΛΩι

F. Y. EDGEWORTH

Ανδρεάδης, Περί απογραφής- με αφιέρωση στον Francis Ysidro Edgeworth

• **1914- Κωνσταντίνος Καραθεοδωρή (1873 – 1950):** Θεώρημα επέκτασης μέτρου

Über das lineare Mass von Punktmengen — eine Verallgemeinerung des Längenbegriffs.

Von

C. Carathéodory in Göttingen.

Vorgelegt von F. Klein in der Sitzung vom 24. Oktober 1914.

Einleitung.

Der Gedanke, die bahnbrechenden und äußerst fruchtbaren Theorien, die Herr Lebesgue für den Inhalt von Punktmengen entwickelt hat [1]), auf den Begriff der Länge zu übertragen, liegt sehr nahe: es genügt eine additive Mengenfunktion zu finden, deren Wert für jede rektifizierbare Kurve gleich der gewöhnlichen Länge dieser Kurve ist, und die im übrigen bei der Bildung von Vereinigungs- und Durchschnittsmengen die Eigenschaften des gewöhnlichen Lebesgueschen Maßes besitzt.

Es zeigt sich, daß man bei der Durchführung dieses Gedankens nicht nur auf keine nennenswerte Schwierigkeiten stößt, sondern daß man eine Theorie erhält, die trotz der großen Allgemeinheit ebenso einfach ist, wie die bisher üblichen. Der einzige Unterschied ist der, daß man die Beweise auf allgemeinere Eigenschaften der Punktmengen zu stützen hat, als die, die man früher gewöhnlich zu Hilfe zog.

Ich habe es deshalb für zweckmäßig gehalten, meine Darstellung mit einer rein formalen Theorie der Meßbarkeit zu beginnen. Dabei wird eine Definition der Meßbarkeit zu Grunde

1) Eine gute Darstellung dieser Theorien findet man in Ch. J. de la Vallée Poussin, Cours d'Analyse Infinitésimale (Louvain & Paris T. I 2e éd, 1914, T. II 2e éd. 1912), eine Zusammenstellung der Literatur in der neuesten unter den grundlegenden Arbeiten von H. Lebesgue, Sur l'intégration des fonctions discontinues (Ann. Éc. Norm. sup. (3) T. 27 (1910) p. 361—450).

Ο Καραθεοδωρή, ίσως ο διασημότερος Έλληνας Μαθηματικός της Νεωτέρας Ελλάδος, μαθητής του Αϊνστάιν είναι γνωστός και στη Θεωρία Πιθανοτήτων και κατ' επέκταση στη Στατιστική από το Θεώρημα Επέκτασης Μέτρου, το οποίο δημοσιεύθηκε στο περιοδικό *Nachrichten von der Koeniglichen Gessellschaft der Wissenschaften zu Goettingen* το **1914** (όχι όπως συνήθως διατυπώνεται, αλλά οι ιδέες είναι εδώ). Η συνήθης διατύπωση εμφανίζεται στο βιβλίο Caratheodory, C. (1918) *Vorlesungen uber reelle Funktionen*, Leipzig-Berlin και έκδοση 1927, κεφάλαιο V: Αν μ_0 είναι μια αριθμήσιμη προσθετική απεικόνιση από μία άλγεβρα $\mathcal{F}_0$ στο διάστημα $[0,\infty]$, τότε υπάρχει μέτρο μ στο $(\Omega, \mathcal{F})$ τέτοιο ώστε $\mu = \mu_0$ στη σ-άλγεβρα $\mathcal{F}$. Αν η μ_0 είναι πεπερασμένη, τότε η επέκταση αυτή είναι μοναδική. (χρήση: μέτρο Lebesgue και μέτρο Wiener). Ο Καραθεοδωρή έχει συνεισφέρει σημαντικά στη θεωρία συναρτήσεων, τη θεωρία μέτρου, τις μερικές διαφορικές εξισώσεις, το λογισμό μεταβολών, τη θεμελίωση θερμοδυναμικής κλπ

- **ΑΣΟΕΕ 1920**

Το **1920** ιδρύεται η Ανωτάτη Σχολή Εμπορικών Επιστημών υπαγόμενη στο Υπουργείο Εθνικής Οικονομίας, η οποία το **1926** μετονομάστηκε σε Ανωτάτη Σχολή Οικονομικών και Εμπορικών Επιστημών (ΑΣΟΕΕ). Το **1927-1928** προσφέρεται το πρώτο μάθημα Στατιστικής στην ΑΣΟΕΕ αποτελούμενο από τη Θεωρητική Στατιστική και τη Στατιστική Πληθυσμού. Διδάσκοντες: Δ. Καλλιτσουνάκης, Καθηγητής Δημοσιονομίας και Στατιστικής και το 1934-35 Καθηγητής Πολιτικής Οικονομίας και Α. Καλλιάβας, Υφηγητής Στατιστικής και το 1934-35 Υφηγητής Πολιτικής Οικονομίας.

1935-36: Στα Γενικά Μαθηματικά διδάσκονται στοιχειώδη προβλήματα πιθανοτήτων από τον Ν. Σακελλαρίου.

- **1936- Βασίλειος Βαλαώρας (1902-1996)**: Το **1936** ο Β. Βαλαώρας εκλέγεται Καθηγητής **Βιοστατιστικής** στην Υγειονομική Σχολή Αθηνών (1936-1946). Επίσης Υφηγητής της Υγιεινής, Επιδημιολογίας και Ιατρικής Στατιστικής στο Πανεπιστήμιο Αθηνών. Doctor of Public Health in Biostatistics, Johns Hopkins University, USA

Το 1941 εκδίδει τα *Στοιχεία Βιομετρίας και Στατιστικής: Δημογραφική μελέτη του πληθυσμού της Ελλάδος,* Εκδ. Βασιλείου, Αθήνα.

Το 1961 έγινε Καθηγητής Υγιεινής και Διευθυντής του Εργαστηρίου Υγιεινής και Επιδημιολογίας στο Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών (ΕΚΠΑ) (διαδέχθηκε τον Αλιβιζάτο) και το 1966 ίδρυσε το Κέντρο Βιομετρικών και Δημογραφικών Ερευνών

• **1939- Αγγελος Αγγελόπουλος (1904-1995). Πρώτος Έλληνας μέλος του ISI το 1939:** Ακαδημαϊκός, Πρύτανης Παντείου (1967-68), Καθηγητής στο ΕΚΠΑ (1937) Οικονομολόγος (Στο ISI ακολουθούν το 1955 ο Β. Βαλαώρας και το 1970 ο Θ. Κάκουλλος, κλπ). Τακτ. Καθηγητής Εφαρμοσμένης Οικονομικής στο Πάντειο Παν/μιο (1961) και Πρύτανης (1967-68). Διοικητής Εθνικής Τράπεζας της Ελλάδος (1974-79). Πρόεδρος της Επιτροπής Σοφών Ζολώτα.

• **1946- Αλιβιζάτος Γεράσιμος (1989-1976):** Ο Γ. Αλιβιζάτος το **1946** εκδίδει το Μνημόνιο Υγιεινής (βλ. και Β΄ έκδοση το 1953). Στο 2ο Κεφάλαιο που πραγματεύεται τη Δημογραφία έχει ένα εδάφιο **Στατιστικής Μεθοδολογίας**, σελ. 147-222. Το 1936 έγινε Καθηγητής Υγιεινής στο ΕΚΠΑ. Το 1938 έγινε διευθυντής του Εργαστηρίου Υγιεινής και Επιδημιολογίας του ΕΚΠΑ μέχρι το 1960. Τον διαδέχθηκε ο Βαλαώρας, μετά ήρθε ο Βασιλειάδης και το 1972 ο Τριχόπουλος

• **1947- Μαυρίκιος Μπρίκας (1896-1981)**: Το **1947** ο Μ. Μπρίκας (1896-1981), στο πλαίσιο των μονογραφιών του για τα Εφαρμοσμένα Μαθηματικά, εκδίδει το **Λογισμό των Πιθανοτήτων** (Γενική Εισαγωγή), Τεύχος 5ο Αθήνα, το πρώτο βιβλίο Πιθανοτήτων.

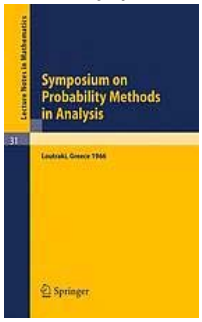
Μερικά περιεχόμενα του Λογισμού των Πιθανοτήτων : Μαθηματικός ορισμός της Πιθανότητας, Συμπλεκτική ανάλυσις, Προσδιορισμός της κατανομής τη βοήθεια των ροπών, Χαρακτηριστική συνάρτησις του Cauchy, Διωνυμική κατανομή, Κατανομή του de Moivre-Laplace-Gauss, Γεωμετρικαί πιθανότητες, ΝΜΑ, Θεώρημα του Bernoulli, Θεώρημα του Bayes

Το 1953-54 ο Μ. Μπρίκας, ως Καθηγητής του Αριστοτέλειου Πανεπιστημίου Θεσσαλονίκης (ΑΠΘ), διδάσκει Μαθήματα Στατιστικής στο 3ο έτος του Τμήματος Οικονομικών και Πολιτικών και Πολιτικών Επιστημών της Νομικής Σχολής του ΑΠΘ και εκδίδει τα *Μαθήματα Στατιστικής, Τεύχη Ι και ΙΙ*, Θεσσαλονίκη. Το 1956 διορίστηκε καθηγητής στο Μαθηματικό Αθηνών.

• **1952-Σπύρος Κανέλλος**: Το **1952** ο Σπύρος Κανέλλος, Υφηγητής στις Πιθανότητες στο Πανεπιστήμιο Αθηνών εκδίδει το βιβλίο του *Λογισμός των Πιθανοτήτων*, Αθήνα, 1952. Και στον Πρόλογο διαβάζουμε «Ο Λογισμός των Πιθανοτήτων δεν είναι μόνον μια μεγαλοπρεπής Μαθηματική θεωρία συνδεδεμένη αναποσπάστως με τα ονόματα διασήμων ανδρών της Μαθηματικής Επιστήμης, αλλ' αποτελεί σήμερον και βάθρον, επί του οποίου στηρίζονται διάφοροι άλλαι εφαρμοσμέναι επιστήμαι». Το βιβλίο του αποτελεί σημαντική συμβολή στην ελληνική ορολογία των πιθανοτήτων λαμβανομένου υπόψιν και του θεμελιώδους βιβλίου του H. Cramer *Mathematical Methods of Statistics,* Princeton University Press, Princeton NJ. (1946).

• **Δημήτριος Κάππος (1904-1985)**

Το **1952** ο Δημ. Κάππος, μαθητής του Καραθεοδωρή, εκλέγεται έκτακτος εντεταλμένος καθηγητής της Α΄ Τακτικής Έδρας των Μαθηματικών και τον Μάιο του 1956 γίνεται τακτικός Καθηγητής. Ο Κάππος στην Ελλάδα είναι γνωστός ως καθηγητής της Μαθηματικής Ανάλυσης. Στο εξωτερικό είναι ευρύτερα γνωστός από το έργο του (και βιβλίο) στις Άλγεβρες Πιθανοτήτων και Στοχαστικούς Χώρους. Συνέβαλε τα μέγιστα στην ανάπτυξη και διεθνή προβολή των σύγχρονων μαθηματικών στην Ελλάδα και παρακίνησε πολλούς μαθηματικούς και στατιστικούς να κάνουν μεταπτυχιακές σπουδές ή/και καριέρα στο εξωτερικό.



Κάππος Δημήτριος

• **1953-ΕΣΥΕ:** Σύσταση με Ν.Δ. της Εθνικής Στατιστικής Υπηρεσίας της Ελλάδος

ΓΕΝΙΚΗ ΓΡΑΜΜΑΤΕΙΑ ΕΘΝΙΚΗΣ ΣΤΑΤΙΣΤΙΚΗΣ ΥΠΗΡΕΣΙΑΣ ΕΛΛΑΔΟΣ
GENERAL SECRETARIAT OF NATIONAL STATISTICAL SERVICE OF GREECE

Πρόδρομος της ΕΣΥΕ ήταν η **Επιτροπή Καταγραφής του Όθωνος (1832)** που συγκροτήθηκε με σκοπό να καταγράψει τα πηγάδια της Ελλάδος, την αγροτική παραγωγή, τις μολυσματικές ασθένειες, και γιατί οι Έλληνες εγκατέλειπαν τους κάμπους και πήγαιναν στα βουνά (οι Έλληνες είχαν επιφυλάξεις να δίνουν στοιχεία στους «Βαυαρούς». Το 1920 λειτούργησε η **Γενική Διεύθυνση Απογραφών στο Υπουργείο Εμπορίου.**

- **ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ (ΠΑΠΕΙ)**

Κέντρο Στατιστικής Εκπαίδευσης

- **1953** Ίδρυση του **Κέντρου Στατιστικής Εκπαίδευσης** στην **Ανωτέρα Σχολή Βιομηχανικών Σπουδών** (1945) με Διευθυντή Σπουδών τον Ευσταθ. Μαργαρίτη. Το 1955 ιδρύεται από το Κέντρο η **Στατιστική Σχολή** και αναγνωρίζεται ως επίσημη Στατιστική Σχολή από το κράτος

Ο Μαργαρίτης είχε διδακτορικό (το **1940**) από τη Σχολή Θετικών Επιστημών του Πανεπιστημίου Αθηνών και η διατριβή του είχε τίτλο: Συμβολή εις την αναλυτικήν σπουδήν των νόμων πιθανοτήτων. Επιβλέπων Καθηγητής του φαίνεται να είναι ο Νείλος Σακελλαρίου. Στα μαθήματα του Κέντρου διακρίνουμε: Μαθηματική Στατιστική, Στατιστική Μέθοδος, Στατιστική Ανάλυση, Δειγματοληψία, Δημογραφία, Οικονομική Στατιστική, Στατιστική Βιομηχανικής Παραγωγής, Στατιστική Επιχειρήσεων, Στατιστική Κοινωνικών Φαινομένων, Οικονομετρική, Ειδικά Θέματα Στατιστικής, Στατιστική Νομοθεσία.

Το 1959 ο Κ. Δρακάτος έλαβε Δίπλωμα από το Κέντρο με άριστα.

Το 1977-78 λειτούργησε για πρώτη φορά το **Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης** στο πλαίσιο της Ανώτατης Βιομηχανικής Σχολής (1958). Το 1984-85 λειτούργησε ως ανεξάρτητο τμήμα και από το 1989 ως Τμήμα του **Πανεπιστημίου Πειραιώς** (1989).

Ελληνική Εταιρεία Στατιστικής

- Το **1955** ο Κάππος ιδρύει την Ελληνική Εταιρεία Στατιστικής με 1^{ο} Πρόεδρο τον Μαργαρίτη. Ουσιαστικά δεν λειτούργησε. Το 1965 ο Κ.Α. Αθανασιάδης τότε Τακτ.

Καθηγητής της ΑΣΟΕΕ αναφέρεται ως Πρόεδρος της Ελληνικής Εταιρείας Στατιστικής

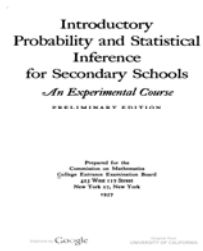
- **Επάνοδος της Στατιστικής στην ΑΣΟΕΕ το 1956-57**

Ο Καθηγητής **Κ. Αθανασιάδης** (1898-1971), Πλοίαρχος ε.α. Εμπορικού Ναυτικού, αυτοδίδακτος στη Στατιστική, το 1956-1957 διδάσκει μάθημα Στατιστικής με περιεχόμενο: Διωνυμική κατανομή, Κατανομές Poisson, -Laplace, Pareto, «έλεγχο ολιγοπληθών δειγμάτων δια της καμπύλης του Student, έλεγχο υποθέσεων, ανεξαρτησίας ιδιοτήτων κλπ δια του χ^2 Helmert-Pearson» και (το επόμενο έτος) διμεταβλητά μαθηματικά υποδείγματα (εξισώσεως παλινδρόμησεως, συντελεστής συσχετίσεως), ανάλυση της διακύμανσης, αριθμοδείκτες.

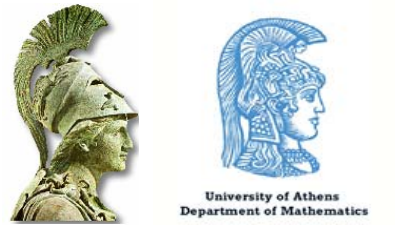
Το **1966** γίνεται Τακτικός Καθηγητής της Στατιστικής στην ΑΣΟΕΕ και εκδίδει Ασκήσεις Στατιστικής (Παπαζήσης) με ασκήσεις $2^{ου}$, $3^{ου}$ έτους και πτυχιακών. Το 1966 εκδίδει μέσω ΚΕΠΕ το ΑΓΓΛΟΕΛΛΗΝΙΚΟΝ ΓΛΩΣΣΑΡΙΟΝ ΤΩΝ ΣΤΑΤΙΣΤΙΚΩΝ ΟΡΩΝ, το πρώτο αγγλοελληνικό λεξικό στατιστικής ορολογίας.

Το **1962,** ο **Κ. Κεβόρκ** εκλέγεται Υφηγητής και το 1964 διορίζεται ως Εντεταλμένος Εισηγητής Στατιστικής. Το **1962-63** ο **Π. Στεριώτης** διδάσκει Λογισμό Πιθανοτήτων στο πλαίσιο των Οικονομικών και Ασφαλιστικών Μαθηματικών.

- **1964- Σιαδήμας:** *Εισαγωγή εις τας Πιθανότητας και τον Στατιστικόν Συμπερασμόν*, μετάφραση από τον Χ. Σιαδήμα του Introductory Probability and Statistical Inference for Secondary Schools Prepared for the Commission on Mathematics of the College Entrance Examination Board: E. C. Douglas (Taft School), F. Mosteller (Harvard), R. S. Pieters (Phillips Academy, Andover), R.A. Richmond, (Williams College), R. E. K. Rourke Kent School), G. B. Thomas, Jr (MIT), S.S. Wilks (Princeton)



- **1965 Έδρα Λογισμού Πιθανοτήτων και Στατιστικής στο Πανεπιστήμιο Αθηνών**



Πανεπιστήμιο Αθηνών

Το 1965 ιδρύεται η Έδρα, το 1966 προκηρύσσεται και το 1968 αναλαμβάνει ως Τακτικός Καθηγητής ο Θ. Κάκουλλος.

Στο Μαθηματικό Αθηνών:

- Το 1958-59 (β' Εξάμηνο) πρώτο μάθημα Λογισμού Πιθανοτήτων - Σαραντόπουλος στο β' έτος [Σαραντόπουλος (1961)]
- Το 1960-61 πρώτο μάθημα Στατιστικής – Σαραντόπουλος – Μπρίκας στο δ' έτος
- Το 1961 ο Σπ. Σαραντόπουλος εκδίδει τον *Λογισμό Πιθανοτήτων και Στατιστική*, Τόμος Α (ουσιαστικά εισαγωγή στις πιθανότητες)

- **1969 Οικονομικό της Νομικής Σχολής του Πανεπιστημίου Αθηνών**: Το 1969 ο Κ. Δρακάτος εκλέγεται Καθηγητής **Στατιστικής** στο νυν Τμήμα Οικονομικών Επιστημών του Πανεπιστημίου Αθηνών

- **1969 Έδρα Πιθανοτήτων και Στατιστικής στο Πανεπιστήμιο Πατρών**

Πρώτος Καθηγητής ο Γ. Ρούσσας το 1972. Ο Γ. Ρούσσας διετέλεσε και Πρύτανης του Πανεπιστημίου Πατρών το 1982-83 και παρέμεινε στο Πανεπιστήμιο μέχρι το 1984 οπότε και επέστρεψε στην Αμερική, Πανεπιστήμιο της Καλιφόρνιας (Davis). Το 1968-69 και το 1969-70 τα πρώτα μαθήματα Πιθανοτήτων και Στατιστικής διδάχτηκαν από τον Θ. Κάκουλλο.

- **1969**: **Ε΄ Έδρα Μαθηματικής Επιστήμης (Πιθανοτήτων και Στατιστικής) στο Πανεπιστήμιο Ιωαννίνων**.

Πρώτος Καθηγητής ο Δημ. Λαμπράκης το 1969-1972. Στη συνέχεια εκλέγεται Καθηγητής και αναλαμβάνει καθήκοντα ο Τ. Παπαϊωάννου από το 1976-1999. Το 1999 μετακλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς.

- **1970 Έδρα Πιθανοτήτων και Στατιστικής στο Πανεπιστήμιο Θεσσαλονίκης**.

Πρώτος Καθηγητής ο Στρατής Κουνιάς το 1976. Διετέλεσε και Αντιπρύτανης του Πανεπιστημίου Θεσσαλονίκης το 1982-1985. Το 1988 μετακλήθηκε στο Τμήμα Μαθηματικών του Πανεπιστημίου Αθηνών.

- **ΓΕΩΠΟΝΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ (ΓΠΑ)**

1972-73: Διδάσκεται το πρώτο αυτοτελές μάθημα Στατιστικής στο ΓΠΑ. Από το 1950 στοιχεία Στατιστικής διδάσκονται στο μάθημα «Βελτίωση Καλλιεργούμενων Φυτών και Γεωργικός Πειραματισμός», το επόμενο ακαδημαϊκό έτος 1951-1952 η Στατιστική εμφανίζεται ως Κεφάλαιο στο πλαίσιο των μαθημάτων Ανώτερα Μαθηματικά και Θεωρητική Μηχανική και Μαθηματικά- Στατιστική (Καθηγητής Δ. Παπαμιχαήλ). Ως αυτοτελές μάθημα διδάσκεται από το 1972-1973. Στις πρώτες 24 έδρες που ιδρύθηκαν με την ίδρυση της «Ανωτέρας Γεωπονικής Σχολής Αθηνών», το 1920 περιλαμβάνεται και η έδρα Μαθηματικών και Στοιχείων Μηχανικής η οποία αργότερα, το 1943, μετονομάστηκε σε έδρα Ανωτέρων Μαθηματικών και Θεωρητικής Μηχανικής.

- **ΠΑΝΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ**

1975 Αρχή διδασκαλίας στοιχείων στατιστικής από το Σπουδαστήριο Κοινωνιολογίας (Β. Φίλιας} στο πλαίσιο (εντός) του μαθήματος Εισαγωγή στη Μεθοδολογία και τις Τεχνικές Κοινωνικών Ερευνών με διδάσκοντες Νικολακόπουλο, Ψυχογιό, Κασιμάτη κλπ. Από το 1985 διδάσκονται δυο μαθήματα Στατιστική Ι και ΙΙ στα ανεξάρτητα Τμήματα Κοινωνιολογίας και Δημόσιας Διοίκησης (Καλαματιανού et al)

- **1981 Ίδρυση του Ελληνικού Στατιστικού Ινστιτούτου**

Πρώτο Δ.Σ. ΕΣΙ 1982-83

Θ. Κάκουλλος, Κ. Δρακάτος, Γ. Κοκολάκης, Δ. Ταμπουρατζής, Χ. Δαμιανού, Χ. Κελπερής, Τ. Παπαϊωάννου

- **1982: Ακαδημαϊκός –Γεωπόνος Ιωάννης Παπαδάκης (1903-1997):** Γίνεται μέλος της Ακαδημίας Αθηνών και το **1986** Επίτιμο μέλος του ΕΣΙ. Διαπρεπής επιστήμων στη γεωπονία και εδαφολογία, γεωργική οικολογία και κλιματολογία αλλά και στο σχεδιασμό και ανάλυση στατιστικών πειραμάτων. Ιδρυτής του Ινστιτούτου Σιτηρών στη Θεσσαλονίκη. Fisher στον κλάδο της Γεωπονίας – Μέθοδος Παπαδάκη: adjusting plot responses in randomized block designs due to correlations with neighboring fields(plots). Methode statistique pour des experiences sur cham*p, Δελτίο Ινστιτούτου Καλλιτερεύσεως Φυτών*, Θεσσαλονίκη, Αρ. 23 (1937). «Σιτάρκεια' το 1957.

- **1984 Πρώτα ακαδημαϊκά Τμήματα Στατιστικής στην Ελλάδα:**

Το 1984 ιδρύονται τα πρώτα τμήματα Στατιστικής στην Ελλάδα, το Τμήμα Στατιστικής και Πληροφορικής στην **ΑΣΟΕΕ**, και το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης στην **ΑΒΣΠ**. Το 2000 ιδρύθηκε στο **Πανεπιστήμιο του Αιγαίου** το Τμήμα Στατιστικής και Αναλογιστικών και Χρηματοοικονομικών Μαθηματικών. Βέβαια και σε άλλα Τμήματα ΑΕΙ [Μαθηματικών, Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών (ΣΕΜΦΕ), Οικονομικά κ.λπ.] ανθούν Τομείς Στατιστικής.

5. ΕΠΙΛΟΓΟΣ

Η προηγούμενη ιστορική παρουσίαση δεν περιλαμβάνει πληροφορίες για την έναρξη και πορεία της Στατιστικής στην Ελλάδα για το Εθνικό Μετσόβιο Πολυτεχνείο (ΕΜΠ), το Χαροκόπειο Πανεπιστήμιο και τις Επιστήμες Ψυχολογίας (στο Πανεπιστήμιο Ιωαννίνων και το Πανεπιστήμιο Αθηνών-Καθηγητής Ι. Παρασκευόπουλος). Η «Στατιστική» δραστηριότητα σ' αυτά τα Ιδρύματα αλλά και σε άλλα είναι νεότερη, κυρίως μετά το 1985 και αποτελεί αντικείμενο συμπληρωματικής έρευνας.

ΠΡΩΤΟΙ ΈΛΛΗΝΕΣ FELLOWS ΣΤΗ ΣΤΑΤΙΣΤΙΚΗ

- **Institute of Mathematical Statistics:** 1983: G. Roussas, 1984: Th. Cacoullos, S. Cambanis, F. Papangelou
- **American Statistical Association:** 1972 Ch. Tsokos, 1978 Suzan Horn-Dadakis
- **Royal Statistical Society**

ABSTRACT

Statistics is about collecting data, analyzing them and presenting what numbers can tell us. It includes Probability Theory since it depends on the calculus of Probability. The present paper is a "timeline", a chronological presentation (historical account) of Statistics in Greece from antiquity to present time.

Ευχαριστίες: Ευχαριστίες απευθύνονται στον κριτή για τις εύστοχες παρατηρήσεις του που συνέβαλαν στη βελτίωση της εργασίας.

ΠΗΓΕΣ

Google

Τα βιβλία που μνημονεύονται

Αγγελής, Λ. (2014) *Ιστορία Στατιστικής*. Μη δημοσιευμένο κείμενο, ΑΠΘ.

Μεϊντάνης, Σ. (2005) Σύντομη αναδρομή στην εξέλιξη της Στατιστικής στην ΑΣΟΕΕ (1927-1984), *Στατιστικό Περισκόπιο No.15.,* Ελληνικό Στατιστικό Ινστιτούτο

Missiakoulis, S. (2008), Aristotle and earthquake data: A historical note. International Statistical Review, 76,1,130-133

Missiakoulis, S. (2010), Cecrops, King of Athens: the first (?) recorded population census in history. International Statistical Review, 78,3, 413-418.

Μωυσιάδης, Χ. (2012) *Ιστορία της έννοιας της Πιθανότητας*, users.auth.gr/hara/courses/history.of.math/2012/Ιστορία της έννοιας της Πιθανότητας 23_05_12.pdf

Παπαϊωάννου, Τ. (2014). Μία ιστορική περιήγηση στην Επιστήμη της Στατιστικής. *Πρακτικά 27ου Πανελληνίου Συνεδρίου Στατιστικής*, Ελληνικό Στατιστικό Ινστιτούτο, 200-215

Σαραντόπουλος, Σ. Β. (1961). *Λογισμός των Πιθανοτήτων και Στατιστική*, τόμος α΄, Αθήνα.

Προσωπική επικοινωνία με Γ. Δονάτο, Κ. Φερεντίνο, Θ. Κάκουλλο, Α. Καλαματιανού, Σ. Κουρούκλη, Θ. Μπόλη, Φ. Παπαγγέλου, Γ. Παπαδόπουλο (με πολύτιμη βοήθεια από τον Δ. Παναγιωτόπουλο, υπεύθυνο αρχείου ΓΠΑ).

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής (2015), σελ.282-294

ΠΟΛΥΩΝΥΜΙΚΗ ΕΚΦΡΑΣΗ ΣΥΜΜΕΤΡΙΚΩΝ ΚΑΤΑΝΟΜΩΝ: Η ΠΕΡΙΠΤΩΣΗ Σ.Π.Π. ΤΡΙΓΩΝΟΜΕΤΡΙΚΗΣ ΜΟΡΦΗΣ

Ι. Παπατσούμα, Ν. Φαρμάκης
Τμήμα Μαθηματικών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
ioannapapatsouma@gmail.com, farmakis@math.auth.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία εξετάζονται κατανομές με συμμετρική συνάρτηση πυκνότητας πιθανότητας (σ.π.π.), οι οποίες εκφράζονται αναλυτικά με χρήση τριγωνομετρικών συναρτήσεων, όπως cosx, sinx, κλπ. Πιο συγκεκριμένα, επιχειρείται (για τις σ.π.π. που εκφράζονται από τριγωνομετρικές μορφές) μία προσεγγιστική έκφραση τους με χρήση πολυωνυμικής συνάρτησης προσέγγισης και με βάση δεδομένα που προέρχονται από συστηματική (κυρίως) δειγματοληψία. Στην όλη διαδικασία υπεισέρχεται ο συντελεστής μεταβλητότητας. Με κίνηση αντίθετης φοράς επιχειρείται να χρησιμοποιηθεί ο εκθέτης της πολυωνυμικής έκφρασης ως στοιχείο ταυτοποίησης της συμμετρικής σ.π.π. Αν προκύψει δηλαδή εκθέτης με συγκεκριμένη τιμή από δεδομένα να συμπεραίνεται ότι η αναλυτική μορφή της σ.π.π. με χρήση τριγωνομετρικών όρων ή παραγόντων να είναι συγκεκριμένη. Π.χ. να έχουμε σ.π.π. με τη μορφή $f(x)=\begin{cases}[\alpha\cdot sin(ax)]/2, & x\in[0,\pi/\alpha]\\ 0, & x\notin[0,\pi/\alpha]\end{cases}$ για συγκεκριμένη τιμή του εκθέτη που θα προκύψει από τα δειγματοληπτικά δεδομένα.

Λέξεις Κλειδιά: δειγματοληψία, συντελεστής μεταβλητότητας, τριγωνομετρική συνάρτηση, τυχαία μεταβλητή, κατανομή

AMS ταξινόμηση: 62D05, 62E17

1. ΕΙΣΑΓΩΓΗ

Η μελέτη κατανομών συνεχών τυχαίων μεταβλητών (τμ) είναι διαδεδομένη σε ένα ευρύ φάσμα της επιστημονικής βιβλιογραφίας. Οι ερευνητές συχνά έχουν να χειριστούν συνεχείς τμ, των οποίων η κατανομή μπορεί να περιγραφεί από μία συνάρτηση πυκνότητας πιθανότητας (σ.π.π.) συμμετρικής μορφής. Εύλογα προκύπτει

το ερώτημα πώς αυτή η σ.π.π. θα εκφραστεί με τέτοιο τρόπο, ώστε να καταστεί ένα εύχρηστο εργαλείο για την περαιτέρω επεξεργασία και αξιοποίησή της.

Η μορφή μιας τέτοιας σ.π.π. διαμορφώνεται από τη φύση του προβλήματος που εξετάζεται και μπορεί να είναι πολυωνυμική, αρνητική εκθετική, τριγωνομετρική κλπ. Στην παρούσα εργασία αντιμετωπίζονται ερωτήματα σχετικά με την προσεγγιστική μορφή των σ.π.π. που θεωρητικά είναι τριγωνομετρικής μορφής και αυτό μέσα από δειγματοληπτικά δεδομένα. Ειδικά, γίνεται μια προσπάθεια να εκφραστεί η σ.π.π. της ίδιας τμ και με πολυωνυμική μορφή με τη βοήθεια του συντελεστή μεταβλητότητας.

2. ΣΥΜΜΕΤΡΙΚΕΣ ΚΑΤΑΝΟΜΕΣ

Έστω συνεχής τμ X, η οποία παίρνει τιμές στο διάστημα [α, β], $a \geq 0$, με σ.π.π.:

$$f(x)=\begin{cases} h\cdot(x-a)^{\nu}, & x\in[\alpha,\frac{\alpha+\beta}{2}] \\ h\cdot(\beta-x)^{\nu}, & x\in[\frac{\alpha+\beta}{2},\beta] \\ 0, & x\notin[\alpha,\beta] \end{cases} \quad h=\frac{2^{\nu}\cdot(\nu+1)}{(\beta-\alpha)^{\nu+1}},\ \nu\neq -1 \qquad (2.1)$$

Για τον υπολογισμό του συντελεστή μεταβλητότητας και των παραμέτρων ν (εκθέτης) και h (συντελεστής) υπολογίζεται αρχικά η μέση τιμή $EX=\int_{-\infty}^{+\infty} x\cdot f(x)\cdot dx=\ldots=(a+\beta)/2$ και ομοίως, η ποσότητα EX^2. Στη συνέχεια, υπολογίζεται η διασπορά

$$VarX=EX^2-\left(EX\right)^2=\ldots=(\beta-a)^2/[2\left(\nu+2\right)\cdot(\nu+3)],\ \nu\in(-\infty,-3)\cup(-2,+\infty)-\{-1\},$$

ο συντελεστής μεταβλητότητας $Cv=\sqrt{2/[(\nu+2)\cdot(\nu+3)]}\cdot[(\beta-\alpha)/(\beta+\alpha)]$ και οι τιμές της παραμέτρου $\lambda=Cv^{-2}$, του εκθέτη $\nu=[-5+\sqrt{1+8\lambda}\,]/2$,και του συντελεστή $h=2^{\nu}\cdot(\nu+1)/(\beta-\alpha)^{\nu+1},\ \nu\neq -1$, Farmakis (2003).

Στην παρούσα εργασία εξετάζονται κατανομές με συμμετρική σ.π.π., οι οποίες εκφράζονται με χρήση τριγωνομετρικών συναρτήσεων, όπως cosx, sinx, κλπ.

Μια οικογένεια ημιτονοειδών σ.π.π. περιγράφεται από τη σχέση:

$$f(x)=\begin{cases} [\alpha\cdot sin(ax)]/2, & x\in[0,\pi/\alpha] \\ 0, & x\notin[0,\pi/\alpha] \end{cases} \qquad (2.2)$$

Η παραπάνω σ.π.π. θα προσεγγιστεί πολυωνυμικά ***με χρήση μόνο του συντελεστή μεταβλητότητας , Cv***.

Πίνακας 2.1. *Εκτίμηση παραμέτρων κατανομής*

μ	σ^2	σ
$\pi / 2\alpha$	$(\pi^2 - 8) / 4a^2$	$\sqrt{(\pi^2 - 8)} / 2\alpha$

Από τα δεδομένα του Πίνακα 2.1, εκτιμώνται ο συντελεστής μεταβλητότητας $CV = \sqrt{(\pi^2 - 8)} / \pi = 0.4352$, η παράμετρος $\lambda = Cv^{-2} = \pi^2 / (\pi^2 - 8) = 5.2774$, ο εκθέτης $\nu = [-5 + \sqrt{1 + 8\lambda}] / 2 = 0.7871$ και ο συντελεστής $h = [2^{\nu} \cdot (\nu + 1)] / (\pi / \alpha)^{\nu+1} = 0.3987 \cdot \alpha^{1.7871}$,

οπότε η αντίστοιχη σ.π.π. πολυωνυμικής μορφής είναι η:

$$f(x) = \begin{cases} 0.3987 \cdot \alpha^{1.7871} \cdot x^{0.7871}, & x \in [0,\ \pi / 2\alpha] \\ 0.3987 \cdot \alpha^{1.7871} \cdot (\dfrac{\pi}{\alpha} - x)^{0.7871}, & x \in [\pi / 2\alpha,\ \pi / \alpha] \\ 0, & x \notin [0,\ \pi / \alpha] \end{cases}$$

Σημείωση 1: Από τις παραπάνω εκτιμήσεις βγαίνει το (σημαντικό) συμπέρασμα ότι για κάθε συνάρτηση που ανήκει στην οικογένεια των ημιτονοειδών συναρτήσεων που περιγράφεται από τη σχέση (2.2) ***ο εκθέτης ν είναι σταθερός και ίσος με 0.7871***.

Τα παραδείγματα που ακολουθούν αποτυπώνουν τη χρησιμότητα των παραπάνω σ.π.π.

3. ΠΑΡΑΔΕΙΓΜΑΤΑ

Παράδειγμα 3.1: Να βρεθεί η προσέγγιση πολυωνυμικής μορφής της σ.π.π. που περιγράφεται από τη σχέση (2.2) για α=1 και να εξεταστεί η ισότητα των δύο κατανομών.

Για α=1 η τριγωνομετρική σ.π.π. γίνεται:

$$\varphi(x) = \begin{cases} (\sin x) / 2, & x \in [0, \pi] \\ 0, & x \notin [0, \pi] \end{cases}$$

Κατόπιν υπολογισμών, προκύπτει ο Πίνακας 3.1:

Πίνακας 3.1 *Εκτίμηση παραμέτρων κατανομής (α=1)*

μ	σ^2	σ	CV	λ	v	h
1.5708	0.4674	0.6837	0.4353	5.2774	0.7871	0.3987

Από τα δεδομένα του Πίνακα 1, προκύπτει η σ.π.π. πολυωνυμικής μορφής:

$$f(x)=\begin{cases}0.3987\cdot x^{0.7871}, & x\in[0,\frac{\pi}{2}]\\ 0.3987\cdot(\pi-x)^{0.7871}, & x\in[\frac{\pi}{2},\pi]\\ 0, & x\notin[0,\pi]\end{cases}$$

και η αντίστοιχη σ.κ. από την ολοκλήρωση της σ.π.π. στο εκάστοτε διάστημα:

$$F(x)=\begin{cases}0, & x<0\\ (0.3987/1.7871)\cdot x^{1.7871}, & x\in[0,\pi/2]\\ 1-(0.3987/1.7871)\cdot(\pi-x)^{1.7871}, & x\in[\pi/2,\pi]\\ 1, & x>\pi\end{cases}$$

Στη συνέχεια, υπολογίζεται η αντίστοιχη σ.κ. της σ.π.π. τριγωνομετρικής μορφής:

$$\Phi(x)=\begin{cases}0, & x<0\\ [1-\cos x]/2, & x\in[0,\pi/2]\\ 1, & x>\pi/2\end{cases}$$

Ο στατιστικός έλεγχος των υποθέσεων:

$$H_0: F(x)=\Phi(x)$$

$$H_1: F(x)\neq\Phi(x)$$

έγινε με το κριτήριο Kolmogorov-Smirnov, προκειμένου να εξεταστεί η ισότητα των δύο κατανομών.

Από τον Πίνακα 3.2, που ακολουθεί, προκύπτει $D_{6,6}=max|\Phi(x)-F(x)|=0.0015$. Από τους πίνακες της Kolmogorov-Smirnov κατανομής προκύπτει η κρίσιμη τιμή $c=D_{6,6;0.05}=0.6667>D_{6,6}$, συνεπώς δεν υπάρχουν επαρκή στοιχεία για την απόρριψη της μηδενικής υπόθεσης.

Πίνακας 3.2. *Αποτελέσματα Kolmogorov-Smirnov (α=1)*

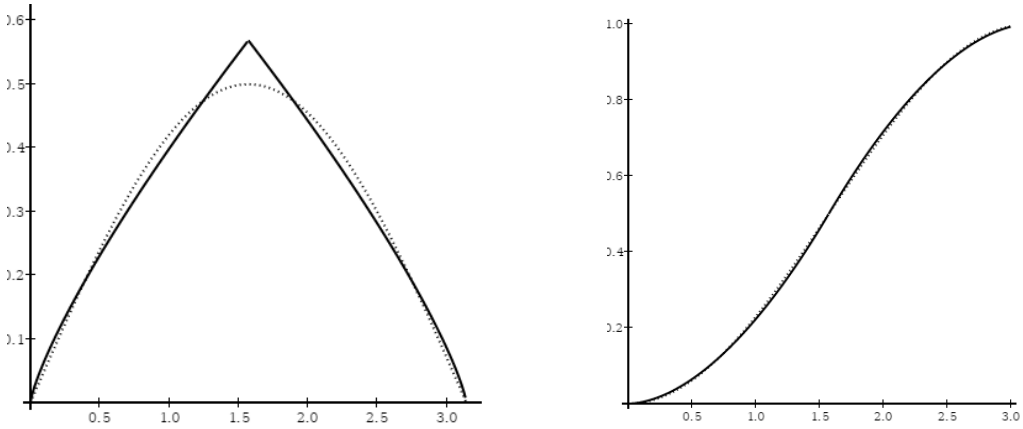
x	$\Phi(x)$	$F(x)$	$\lvert\Phi(x)-F(x)\rvert$

[0, π/6)	0.0670	0.0702	0.0032
[π/6, π/3)	0.2500	0.2423	***0.0077***
[π/3, π/2)	0.5000	0.5000	0.0000
[π/2, 2π/3)	0.7500	0.7577	***0.0077***
[2π/3, 5π/6)	0.9330	0.9298	0.0032
[5π/6, π)	1.0000	1.0000	0.0000

Σημείωση 2: Ο παραπάνω έλεγχος υποθέσεων έγινε, επίσης, για εύρος κλάσης *w=π/4* και *w=π/8* και δεν υπήρχαν επαρκή στοιχεία για την απόρριψη της μηδενικής υπόθεσης ($c' = D_{4,4;0.05} = 0.75 > max|\Phi(x)-F(x)| = 0.0015$ και $c'' = D_{8,8;0.05} = 0.625 > max|\Phi(x)-F(x)| = 0.0097$, αντίστοιχα).

Όλα τα παραπάνω αποτυπώνονται με εύληπτο τρόπο στις γραφικές παραστάσεις των αρχικών τριγωνομετρικών σ.π.π. και σ.κ. (διακεκομμένη γραμμή) και των πολυωνυμικών προσεγγίσεών τους (συνεχής γραμμή) (Σχήμα 1).

Σχήμα 1. *Γραφικές παραστάσεις των σ.π.π. φ και f (αριστερά) και των σ.κ. Φ και F (δεξιά)*



Παράδειγμα 3.2: Να βρεθεί η προσέγγιση πολυωνυμικής μορφής της σ.π.π. που περιγράφεται από τη σχέση (2.2) για α=2 και να εξεταστεί η ισότητα των δύο κατανομών.

Για α=2 η τριγωνομετρική σ.π.π. της (2.2) γίνεται:

$$\varphi(x)=\begin{cases}2\cdot \sin x\cdot \cos x, & x\in[0,\ \pi/2]\\ 0, & x\notin[0,\ \pi/2]\end{cases}$$

Κατόπιν υπολογισμών, προκύπτει ο Πίνακας 3.3:

Πίνακας 3.3. *Εκτίμηση παραμέτρων κατανομής (α=2)*

μ	σ^2	σ	CV	λ	ν	h
0.7854	0.1169	0.3419	0.4353	5.2774	0.7871	1.3760

Από τα δεδομένα του Πίνακα 3.3, προκύπτει η σ.π.π. πολυωνυμικής μορφής:

$$f(x)=\begin{cases}1.3760\cdot x^{0.7871}, & x\in[0,\ \pi/4]\\ 1.3760\cdot[((\pi/2)-x)^{0.7871}], & x\in[\pi/4,\ \pi/2]\\ 0, & x\notin[0,\ \pi/2]\end{cases}$$

και η αντίστοιχη σ.κ.:

$$F(x)=\begin{cases}0, & x<0\\ (1.3760/1.7871)\cdot x^{1.7871}, & x\in[0,\ \pi/4]\\ 1-(1.3760/1.7871)\cdot[((\pi/2)-x)^{1.7871}], & x\in[\pi/4,\ \pi/2]\\ 1, & x>\pi/2\end{cases}$$

Στη συνέχεια, υπολογίζεται η αντίστοιχη σ.κ. της σ.π.π. τριγωνομετρικής μορφής:

$$\Phi(x)=\begin{cases}0, & x<0\\ [1-\cos(2x)]/2, & x\in[0,\pi/2]\\ 1, & x>\pi/2\end{cases}$$

Ο στατιστικός έλεγχος των υποθέσεων:

$$H_0:\quad F(x)=\Phi(x)$$

$$H_1: \quad F(x) \neq \Phi(x)$$

έγινε με το κριτήριο Kolmogorov-Smirnov, προκειμένου να εξεταστεί η ισότητα των δύο κατανομών.

Από τον Πίνακα 3.4, που ακολουθεί, προκύπτει $D_{4,4} = max\ |\Phi(x)-F(x)| = 0.0015$. Από τους πίνακες της Kolmogorov-Smirnov κατανομής προκύπτει η κρίσιμη τιμή $c = D_{4,4;0.05} = 0.75 > D_{4,4}$, συνεπώς δεν υπάρχουν επαρκή στοιχεία για την απόρριψη της μηδενικής υπόθεσης.

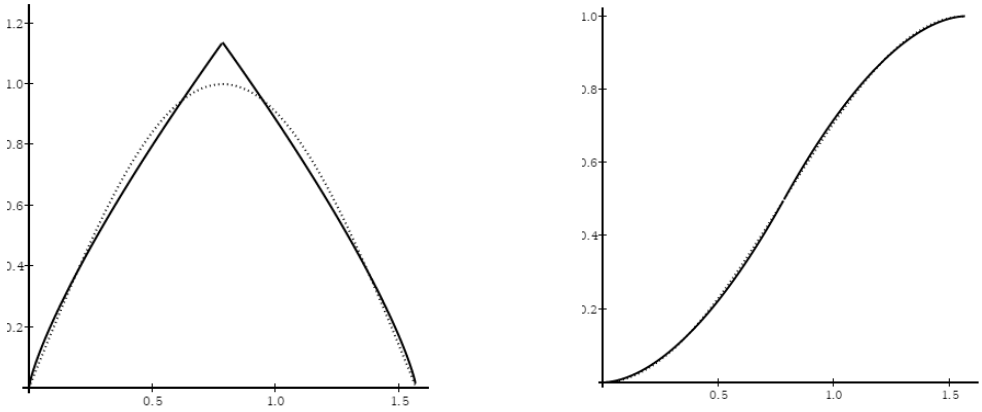
Πίνακας 3.4. *Αποτελέσματα Kolmogorov-Smirnov (α=2)*

x	***Φ(x)***	***F(x)***	***|Φ(x)-F(x)|***
[0, π/8)	0.1464	0.1449	***0.0015***
[π/8, π/4)	0.5000	0.5000	0.0000
[π/4, 3π/8)	0.8536	0.8551	***0.0015***
[3π/8, π/2)	1.0000	1.0000	0.0000

Σημείωση 3: Ο παραπάνω έλεγχος υποθέσεων έγινε, επίσης, για εύρος κλάσης $w=\pi/4$ και $w=\pi/6$. Στον πίνακα της Kolmogorov-Smirnov κατανομής για τις κρίσιμες τιμές δειγμάτων μεγέθους $n=m=2$ και $n=m=3$ αντίστοιχα, υπάρχει σημείωση ότι τα δείγματα προέρχονται από ίσες κατανομές (μηδενική υπόθεση).

Οι γραφικές παραστάσεις των σ.π.π. και των σ.κ. των αρχικών τριγωνομετρικών συναρτήσεων και των πολυωνυμικών προσεγγίσεών τους είναι παρόμοιες με αυτές του προηγούμενου παραδείγματος, με διαφορετικό πεδίο ορισμού, όπως φαίνεται παρακάτω στο Σχήμα 2.

Σχήμα 2. *Γραφικές παραστάσεις των σ.π.π. φ και f (αριστερά) και των σ.κ. Φ και F (δεξιά)*



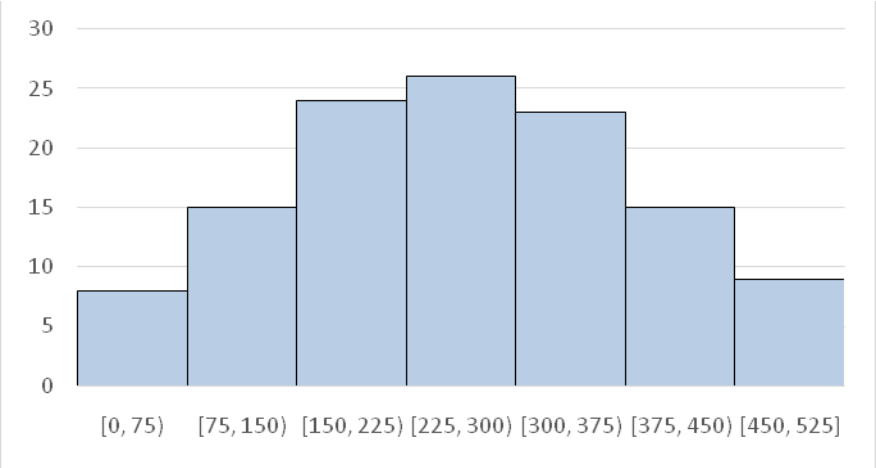
Παράδειγμα 3.3: Σε ένα πείραμα διαδικτυακής εκπαίδευσης 120 μαθητές προσπαθούν να βρουν, ο καθένας ξεχωριστά, την απάντηση στην ερώτηση «*Πότε κτίστηκε ο Παρθενώνας;*». Οι τιμές της τμ *X* αναπαριστούν τον *χρόνο,* σε sec, που χρειάστηκε ο κάθε μαθητής για να βρει τη σωστή απάντηση μέσω μίας μηχανής αναζήτησης.

Πίνακας 3.5. *Κατανομή μαθητών*

Κλάσεις	***[0, 75)***	***[75, 150)***	***[150, 225)***	***[225, 300)***	***[300, 375)***	***[375, 450)***	***[450, 525]***
Συχνότητες ni	*8*	*15*	*24*	*26*	*23*	*15*	*9*

Από τα δεδομένα του Πίνακα 3.5, κατασκευάζεται το ιστόγραμμα συχνοτήτων (Σχήμα 3), το οποίο αναδεικνύει με εύληπτο τρόπο ότι η δειγματική κατανομή προσιδιάζει με συμμετρική κατανομή.

Σχήμα 3. *Ιστόγραμμα συχνοτήτων*



Υποθέτουμε ότι οι παρατηρήσεις κάθε κλάσης είναι ομοιόμορφα κατανεμημένες και ότι οι τιμές της μεταβλητής σε κάθε κλάση εκπροσωπούνται από την αντίστοιχη κεντρική τιμή x_i, δηλαδή το ημιάθροισμα των άκρων της εκάστοτε κλάσης. Από τις κεντρικές τιμές υπολογίζονται οι τιμές $x_i' = (x_i - m)/w$, *i= 1, ..., 7,* όπου *w* το πλάτος της εκάστοτε κλάσης και *m* η προσωρινή μέση τιμή (κεντρική τιμή μεσαίας κλάσης συνήθως). Στη συνέχεια, υπολογίζονται οι ποσότητες $T_1 = \sum_{i=1}^{7} n_i x_i'$ και $T_2 = \sum_{i=1}^{7} n_i x_i'^2$ (Πίνακας 3.6), με τη βοήθεια των οποίων εκτιμώνται η δειγματική μέση τιμή $\bar{x} = m + (w \cdot T_1)/n$ και η δειγματική διασπορά $s^2 = [w^2 \cdot (T_2 - T_1^2/n)]/(n-1)$, Κολυβά-Μαχαίρα, Μπόρα-Σέντα (2012).

Πίνακας 3.6. *Υπολογισμός* T_1 & T_2

Κλάσεις	x_i	n_i	x_i'	$n_i x_i'$	$n_i x_i'^2$
[0, 75)	37.5	8	-3	-24	72
[75, 150)	112.5	15	-2	-30	60
[150,225)	187.5	24	-1	-24	24
[225, 300)	***262.5***	26	0	0	0
[300, 375)	337.5	23	1	23	23

[375, 450)	412.5	15	2	30	60
[450, 525]	487.5	9	3	27	81
Σύνολο		**120**		**T_1=2**	**T_2=320**

Κατόπιν υπολογισμών, προκύπτει ο Πίνακας 3.7:

Πίνακας 3.7. *Εκτίμηση παραμέτρων πειράματος*

$\bar{x}$	s^2	s	CV	λ	v	h
263.75 sec	15124.4748 sec	122.98 sec	0.4663	4.5991	0.5738	0.000123

Από τα δεδομένα του Πίνακα 7, προκύπτει η σ.π.π. πολυωνυμικής μορφής:

$$f(x)=\begin{cases}0.000123\cdot x^{0.5738}, & x\in[0,\ 262.5]\\ 0.000123\cdot(525-x)^{0.5738}, & x\in[262.5,\ 525]\\ 0, & x\notin[0,525]\end{cases}$$

και η αντίστοιχη σ.κ.:

$$F(x)=\begin{cases}0, & x<0\\ 0.000078\cdot x^{1.5738}, & x\in[0,\ 262.5]\\ 1-0.000078\cdot(525-x)^{1.5738}, & x\in[262.5,\ 525]\\ 1, & x>525\end{cases}$$

Ακολουθούν οι αντίστοιχες σ.π.π. και σ.κ. τριγωνομετρικής μορφής:

$$\varphi(x)=\begin{cases}(\pi/1050)\cdot sin[(\pi\cdot x)/525], & x\in[0,525]\\ 0, & x\notin[0,525]\end{cases}$$

$$\Phi(x)=\begin{cases}0, & x<0\\ (1/2)\cdot[1-cos((\pi\cdot x)/525)], & x\in[0,525]\\ 1, & x>525\end{cases}$$

Στη συνέχεια, έγινε ο στατιστικός έλεγχος των υποθέσεων:

$$H_0: F(x) = \Phi(x)$$

$$H_1: F(x) \neq \Phi(x)$$

με το κριτήριο Kolmogorov-Smirnov, προκειμένου να εξεταστεί η ισότητα των δύο κατανομών.

Από τον Πίνακα 3.8, που ακολουθεί, προκύπτει $D_{7,7} = max\ |\Phi(x)-F(x)| = 0.0202$. Από τους πίνακες της Kolmogorov-Smirnov κατανομής προκύπτει η κρίσιμη τιμή $c = D_{7,7;0.05} = 0.7143 > D_{7,7}$, συνεπώς δεν υπάρχουν επαρκή στοιχεία για την απόρριψη της μηδενικής υπόθεσης.

Πίνακας 3.8. *Αποτελέσματα Kolmogorov-Smirnov πειράματος*

x	***Φ(x)***	***F(x)***	***|Φ(x)-F(x)|***
[0, 75)	0.0495	0.0697	***0.0202***
[75, 150)	0.1882	0.2074	0.0192
[150, 225)	0.3887	0.3926	0.0039
[225, 300)	0.6113	0.6074	0.0039
[300, 375)	0.8118	0.7926	0.0192
[375, 450)	0.9505	0.9303	***0.0202***
[450, 525]	1.0000	1.0000	0.0000

Στη συνέχεια, υπολογίστηκαν οι τιμές των θεωρητικών συχνοτήτων $\theta_i = n \cdot \Delta\Phi_i$ και $\theta_i' = n \cdot \Delta F_i$, $i= 1, \ldots, 7,$ αντίστοιχα (Πίνακας 3.9).

Σημείωση 4: Όλες οι τιμές των θεωρητικών συχνοτήτων βρέθηκαν μεγαλύτερες του 5, συνεπώς δεν παραβιάζεται ο περιορισμός του Cochran και δεν απαιτείται σύμπτυξη μεταξύ των κλάσεων.

Ακολούθησε ο στατιστικός έλεγχος των υποθέσεων:

H_0: *Δεν υπάρχει διαφορά στον τρόπο διεξαγωγής του πειράματος*

H_1: *όχι η* H_0

με το κριτήριο καλής προσαρμογής X^2, προκειμένου να εξεταστεί αν υπάρχει διαφορά μεταξύ των δεδομένων που συλλέχθηκαν και παρουσιάστηκαν στον Πίνακα

3.5 (πραγματικές συχνότητες) και αυτών που θα περιμέναμε να εμφανιστούν αν ίσχυε η μηδενική υπόθεση (αναμενόμενες ή θεωρητικές συχνότητες). Με άλλα λόγια, εξετάστηκε αν στην πραγματικότητα δεν υπήρχε διαφορά στον τρόπο με τον οποίο οι μαθητές έψαχναν σε μία μηχανή αναζήτησης και απαντούσαν στην ερώτηση που τους δόθηκε, δηλαδή αν η προσαρμογή του μοντέλου (θεωρητική κατανομή) στη δειγματική κατανομή είναι καλή.

$$X^2 = \sum_{i=1}^{7} \frac{n_i^2}{\theta_i} - n = 120.0721 - 120 = 0.0721 < 12.6 = X^2_{6;0.05} \qquad (3.1)$$

$$X'^2 = \sum_{i=1}^{7} \frac{n_i^2}{\theta_i'} - n = 120.6042 - 120 = 0.6042 < 12.6 = X^2_{6;0.05} \qquad (3.2)$$

Βάσει των θεωρητικών συχνοτήτων των σ.π.π. τριγωνομετρικής και πολυωνυμικής μορφής, υπολογίστηκαν οι τιμές των στατιστικών X^2. Οι τιμές αυτές βρέθηκαν μικρότερες των αντίστοιχων κρίσιμων τιμών σε επίπεδο σημαντικότητας 0.05 (σχέσεις (3.1), (3.2)), συνεπώς δεν υπάρχουν επαρκή στοιχεία για την απόρριψη της μηδενικής υπόθεσης.

Πίνακας 3.9. *Πίνακας θεωρητικών συχνοτήτων*

x	n_i	θ_i	n_i^2/θ_i	θ_i'	n_i^2/θ_i'
[0, 75)	8	5.9	10.8474	8.4	7.6190
[75, 150)	15	16.6	13.5542	16.5	13.6364
[150, 225)	24	24.0	24.0000	22.2	25.9459
[225, 300)	26	26.7	25.3184	25.8	26.2016
[300, 375)	23	24.1	21.9502	22.2	25.9459
[375, 450)	15	16.6	13.5545	16.5	13.6364
[450, 525]	9	5.9	10.8474	8.4	7.6190

4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στις τριγωνομετρικές συμμετρικές συναρτήσεις της μορφής $\alpha / 2 \cdot sin(\alpha x)$, ο συντελεστής μεταβλητότητας, η παράμετρος λ και ο εκθέτης ν του πολυωνύμου είναι σταθερές ποσότητες, ανεξάρτητες από τα άκρα του εκάστοτε διαστήματος. Διαπιστώθηκε ότι οι μέθοδοι εκτίμησης των σ.π.π. πολυωνυμικής και τριγωνομετρικής μορφής είναι ισοδύναμες (εξίσου αποτελεσματικές) και ότι η

προσαρμογή του μοντέλου, δηλαδή της θεωρητικής κατανομής, στη δειγματική κατανομή είναι πολύ καλή. Τέλος, προκύπτει ότι τα δεδομένα προσαρμόζονται καλύτερα στο μοντέλο της σ.π.π. τριγωνομετρικής μορφής, συγκριτικά με το μοντέλο της σ.π.π. πολυωνυμικής μορφής.

Ευχαριστίες: Οι δύο συγγραφείς ευχαριστούν τον κριτή (ες) για τις εποικοδομητικές παρατηρήσεις που βελτίωσαν την ποιότητα της εργασίας αυτής.

ABSTRACT

This paper examines distributions with symmetric probability density function (pdf), expressed with a trigonometric function, such as *cosx*, *sinx*, etc. In particular, a polynomial approach of pdf is attempted, using systematic sampling and the coefficient of variation. The use of polynomial expression's exponent as identifier of the symmetric pdf is also attempted, in order to assess the analytical form using trigonometric conditions or factors, when the exponent is obtained from data. For example, for a specific exponent (ν=0.7871) we have the following pdf $f(x)=\begin{cases}[\alpha \cdot sin(ax)]/2, & x \in [0, \pi/\alpha] \\ 0, & x \notin [0, \pi/\alpha]\end{cases}$

ΑΝΑΦΟΡΕΣ

Φαρμάκης Ν. (2009$^{\alpha}$) "*Δημοσκοπήσεις και Δεοντολογία*", Εκδόσεις Α&Π Χριστοδουλίδη, Θεσσαλονίκη.

Φαρμάκης Ν. (2009$^{\beta}$) "*Εισαγωγή στη Δειγματοληψία*", Εκδόσεις Α&Π Χριστοδουλίδη, Θεσσαλονίκη.

Κολυβά-Μαχαίρα Φ., Μπόρα-Σέντα Ε. (2012) "*Στατιστική: Θεωρία και Εφαρμογές*", Εκδόσεις Ζήτη, Θεσσαλονίκη.

Farmakis, N., (2003). "Estimation of Coefficient of Variation: Scaling of Symmetric Continuous Distributions", *Statistics in Transition*, Vol. 6, No 1, pp 83-96.

Farmakis, N., Makris G., (2011). "Web-sampling: Probabilities of Specific Information Achievement", *Proceedings of Applied Stochastic Models Data Analysis (ASMDA 2011), Rome, Italy*, pp 447-454.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής (2015), σελ.295-309

ΕΦΑΡΜΟΓΗ ΚΑΙ ΑΞΟΛΟΓΗΣΗ ΜΕΘΟΔΩΝ ΧΩΡΙΚΗΣ ΑΝΑΛΥΣΗΣ ΒΡΟΧΟΜΕΤΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

Π.Σκριμιζέας
Εθνική Μετεωρολογική Υπηρεσία
pskrim@hnms.gr/pan.skrimizeas@gmail.com

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία επιχειρείται μία ανασκόπηση των χρησιμοποιούμενων μεθόδων χωρικής ανάλυσης κλιματικών δεδομένων και η εφαρμογή – αξιολόγηση τεσσάρων από αυτές σε τρία σετ δεδομένων βροχόπτωσης από 75 μετεωρολογικούς σταθμούς του δικτύου της Εθνικής Μετεωρολογικής Υπηρεσίας, για την περίοδο 1981-2000. Η χωρική ανάλυση υλοποιήθηκε σε περιβάλλον Γεωγραφικών Πληροφορικών Συστημάτων (ΓΠΣ) και παρήχθησαν οι αντίστοιχοι βροχομετρικοί χάρτες. Λαμβάνοντας υπόψη τη διασπορά των θέσεων παρατήρησης, τις σημαντικές διακυμάνσεις της πυκνότητας του δικτύου αλλά και την ειδική γεωμορφολογία του ελληνικού χώρου, οι παραγόμενοι χάρτες ετήσιας και εποχιακής βροχόπτωσης (υγρής και ξηρής περιόδου) κρίνονται ικανοποιητικοί καθόσον απεικονίζουν την βασική κατανομή της βροχόπτωσης στον ελληνικό χώρο αναγνωρίζοντας υφιστάμενα πρότυπα.

Λέξεις Κλειδιά: χωρική ανάλυση, βροχομετρικοί χάρτες, χωρική στατιστική, γεωγραφικά συστήματα πληροφοριών.

1. ΕΙΣΑΓΩΓΗ

Η Μετεωρολογία αποτελεί κλάδο των φυσικών επιστημών, με κύριο αντικείμενο την έρευνα της ατμόσφαιρας στο σύνολό της και τα φαινόμενα που συμβαίνουν σε αυτή, όπως αυτή περιγράφεται από ένα σύνολο μετεωρολογικών μεταβλητών, συνηθέστερες των οποίων είναι η θερμοκρασία του αέρα, η ατμοσφαιρική πίεση, η υγρασία, η διεύθυνση και ένταση του ανέμου, τα νέφη, τα καιρικά φαινόμενα, το ποσό βροχής ή χιονιού, η ορατότητα κ.α. Η καταγραφή και ανάλυση για μια μακρά περίοδο των δεδομένων του καιρού συνθέτει την εικόνα του μέσου καιρού μιας περιοχής ή το κλίμα. Η επιστήμη που ασχολείται με το κλίμα είναι η κλιματολογία (climatology) Για την διαμόρφωση της εικόνας του κλίματος μιας περιοχής, με βάση τον Παγκόσμιο Μετεωρολογικό Οργανισμό, μία αποδεκτή χρονοσειρά δεδομένων είναι τα τριάντα χρόνια. Ειδικά για την μετεωρολογική πληροφορία θα πρέπει να σημειωθεί ότι οι ερευνητές από πολύ νωρίς, είχαν διαπιστώσει την ανάγκη της ανάπτυξης τεχνικών χωρικής παρεμβολής (ή ανάλυσης της χωρικής κατανομής) των μελετώμενων παραμέτρων, μιας διαδικασίας δηλαδή μετάβασης από τις διακριτές τιμές (σημειακά δεδομένα - παρατηρήσεις) σε μία συνεχή επιφάνεια. Η κατασκευή κλιματικών χαρτών

υψηλής ευκρίνειας, είναι σήμερα, ιδιαίτερα απαιτητή αφού πέραν του διαγνωστικού ρόλου τους, στην κατανόηση και περιγραφή του κλίματος μιας περιοχής, αποτελούν πλέον βασική πηγή γνώσης για την προβολή του κλίματος στο μέλλον, καθόσον η στοχαστικά ασφαλής πρόβλεψη του μελλοντικού κλίματος ξεκινά από την γνώση του παρόντος και του παρελθόντος.

Έτσι αναπτύσσεται ένας νέος επιστημονικός κλάδος η γεωχωρική κλιματολογία (Geospatial Climatology - Daly, 2010. Αναζητώντας, όπως αναφέρει ο Gunst (1995), την βέλτιστη προσέγγιση στη διαδικασία μετάβασης από σημειακά δεδομένα σε επιφάνεια, οι Matheron (1962) και Gandin (1963), εισήγαγαν μεθόδους που έχουν ονομαστεί, αντίστοιχα, kriging και βέλτιστη παρεμβολή. Η ευρεία χρήση στατιστικών μεθόδων ουσιαστικά ξεκίνησε την τελευταία δεκαετία του προηγούμενου αιώνα.

2. ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ ΤΩΝ ΜΕΘΟΔΩΝ ΧΩΡΙΚΗΣ ΑΝΑΛΥΣΗΣ ΚΛΙΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

Περιγραφή των μεθόδων χωρικής παρεμβολής παρέχεται από τους Dobesch et al. (2007) και Tveito et al. (2006). Σύμφωνα με τον Sluiter (2009) οι μέθοδοι αυτές κατατάσσονται σε τρεις βασικές κατηγορίες: αιτιοκρατικές - προσδιοριστικές (deterministic), στοχαστικές-πιθανοθεωρητικές - στατιστικές (probabilistic) και συνδυαστικές – εμπειρικές. Οι αιτιοκρατικές παράγουν μια συνεχή επιφάνεια με τη χρήση μόνο των γεωμετρικών χαρακτηριστικών των σημειακών μετρήσεων. Οι στοχαστικές μέθοδοι χρησιμοποιούν την έννοια της τυχαιότητας, δεχόμενες ότι το παρεμβαλλόμενο πεδίο είναι ένα από πολλά δυνητικά υφιστάμενα. Οι συνδυαστικές – εμπειρικές μέθοδοι αποτελούν συνδυασμό ντετερμινιστικών, στοχαστικών αλλά και εμπειρικών μεθόδων. Σε όλες τις μεθόδους δίνεται ιδιαίτερη προσοχή στις τεχνικές ενσωμάτωσης των «βοηθητικών δεδομένων». «Βοηθητικά δεδομένα» (Sluiter, 2009) είναι, η μορφολογία με την έννοια της εναλλαγής γης και θάλασσας, η εγγύτητα στη θάλασσα, το υψόμετρο, η χρήση γης, το γεωγραφικό πλάτος, δεδομένα βροχόπτωσης από ραντάρ, δεδομένα από αριθμητικά μοντέλα καιρού, κυρίως για την παροχή αξιόπιστων πεδίων σαν υπόβαθρο κ.α. Ένας άλλος διαχωρισμός των μεθόδων παρεμβολής είναι σε «γενικευμένες» και «τοπικές» (Global – Local Interpolation Methods) (Sluiter, 2009). Οι γενικευμένες μέθοδοι χρησιμοποιούν όλα τα υπάρχοντα δεδομένα, ολόκληρης της περιοχής μελέτης, και παράγουν, μέσω μίας ενιαίας διαδικασίας, αποτελέσματα για το σύνολο της περιοχής, παράγοντας εν γένει περισσότερο «λείες» επιφάνειες, ενώ οι τοπικές μέθοδοι μέσω απλών ή περισσότερο σύνθετων διαδικασιών υλοποιούν την παρεμβολή διαδοχικά εντός υποπεριοχών με τη χρήση του αντίστοιχου υποσυνόλου των δεδομένων. Ένας τρίτος χωρισμός των μεθόδων είναι σε «ακριβείς» και «μη ακριβείς ή προσεγγιστικές» (exact – inexact or approximate). Στις «ακριβείς» μεθόδους η εκτιμώμενη τιμή παραμένει η ίδια στα σημεία που υπάρχουν μετρήσεις σε αντίθεση με τις «προσεγγιστικές» που δεν αναπαράγουν τις αρχικές τιμές στα σημεία μέτρησης υιοθετώντας την υπόθεση της αβεβαιότητας και σε αυτές.

Η περιγραφή των μεθόδων για την χωρική παρεμβολή των κλιματικών δεδομένων δεν θα ήταν πλήρης αν δεν γίνει αναφορά και σε τρεις άλλες μεθόδους οι οποίες έ-

χουν αναπτυχθεί σχεδόν αποκλειστικά για ανάλυση μετεωρολογικών δεδομένων και ενδεχομένως δεν μπορούν να ενταχθούν σε καμία από τις παραπάνω κατηγορίες λόγω διαφορετικής φιλοσοφίας προσέγγισης του θέματος ή χρήσης συνδυασμού τεχνικών, την μέθοδο MISH (Meteorological interpolation based on Surface Homogenized Data Basis), την μέθοδο PRISM (Parameter - elevation Relationships on Independent Slopes Model) και την AURELHY (Analysis Using the RELief for HYdrometeorology). Εκτεταμένη βιβλιογραφική αναφορά των μεθοδολογιών και τεχνικών της διεθνούς ερευνητικής κοινότητας στον τομέα της χωρικής παρεμβολής κλιματολογικών παραμέτρων παρέχεται από τον Σκριμιζέα (2014).

Η αξιολόγηση της εφαρμοζόμενης μεθόδου χωρικής παρεμβολής αποτελεί εκ των ων ουκ άνευ της όλης διαδικασίας, με στόχο, μέσω του βαθμού προσαρμογής, και την εκτίμηση της τάξης μεγέθους του σφάλματος της εφαρμοζόμενης μεθόδου. Τόσο στις αιτιοκρατικές όσο και τις στοχαστικές μεθόδους η εκτίμηση της καλής προσαρμογής ή η επιλογή της βέλτιστης μεθόδου βασίζεται στην απλή ιδέα της επαλήθευσης με τη χρήση ενός δείγματος μετρήσεων οι οποίες δεν συμμετείχαν στη διαδικασία δόμησης του μοντέλου. Οι μέθοδοι Kriging πλεονεκτούν στη δυνατότητα εκτίμησης του σφάλματος της παρεμβολής και κυρίως παρέμβασης σε αυτό, στην κατεύθυνση ελαχιστοποίησής του. Πάντως, σύμφωνα με τον Cressie (1991) η δυνατότητα αυτή δεν παρέχει από μόνη της συνολική αξιολόγηση της μεθόδου.

Μαθηματικά όλες οι μέθοδοι παρεμβολής εμπεριέχουν σφάλμα. Πόσο ακριβείς όμως είναι και οι διαθέσιμες μετρήσεις; Συνεπώς η δυνατότητα της μεθόδου να ποσοτικοποιεί το σφάλμα της εκτίμησης αποτελεί μονόδρομο στη κατεύθυνση της αξιολόγησης του βαθμού χρηστικότητας του αποτελέσματος. Η διαθέσιμη σήμερα υπολογιστική ισχύς επέτρεψε την υλοποίηση τεχνικών οι οποίες συνεισφέρουν σημαντικά στην αξιολόγηση τόσον της ποιότητας του μοντέλου της χωρικής παρεμβολής όσον και των δεδομένων (Kresic and Mikszewski, 2012).

Έχουν προταθεί διάφορες, παραμετρικές και μη, στατιστικές διαδικασίες εκ των οποίων οι πλέον χρησιμοποιούμενες είναι οι Cross Validation, Jackknife και bootstrap (Michaelsen, 1987, Wilks, 2011). Ειδικά η Cross – Validation είναι μια ποσοτική μέθοδος, και εκ των πλέον δημοφιλών, η οποία επιτρέπει εκτίμηση της σχετικής ποιότητας της ανάλυσης υπολογίζοντας το σφάλμα σε επίπεδο καταλοίπων (residuals). Υπάρχουν διάφορες παραλλαγές της μεθόδου (Refaeilzadeh et al. 2009). Σύμφωνα με μία εξ αυτών (leave-one-out), τα σφάλματα υπολογίζονται αφαιρώντας διαδοχικά μία παρατήρηση από το σύνολο των δεδομένων των τιμών N και χρησιμοποιώντας τα υπόλοιπα δεδομένα και τον αλγόριθμο παρεμβολής για την εκτίμηση της τιμής στην πρώτη θέση παρατήρησης. Από τα διαθέσιμα στο τέλος N σφάλματα παρεμβολής υπολογίζεται το μέσο σφάλμα (ME- mean error,) ως εκτίμηση της μεροληψίας (bias) και η τετραγωνική ρίζα του μέσου τετραγωνικού σφάλματος (RMSE – root mean square error) το οποίο παρέχει την ακρίβεια της εκτίμησης. Τα σφάλματα αυτά μπορούν να υπολογισθούν σε όλες τις μεθόδους παρεμβολής. Οι αιτιοκρατικές μέθοδοι εξαντλούν στο σημείο αυτό τη δυνατότητά τους σε αντίθεση με τις γεωστατιστικές μεθόδους που, όπως έχει ήδη αναφερθεί, μπορούν να εκτιμήσουν την αβεβαιότητα της παρεμβολής. Ένα καλό μοντέλο θα πρέπει να είναι αμερόληπτο και να

προσεγγίζει όσο το δυνατόν την πραγματικότητα, με την έννοια της επαλήθευσης στα σημεία όπου υπάρχουν, αλλά θα είναι καλλίτερο αν μπορεί να περιγράψει με την μέγιστη δυνατή ακρίβεια και την μεταβλητότητα των δεδομένων (Kresic and Mikszewski, 2012).

3. ΧΑΡΤΟΓΡΑΦΗΣΗ ΒΡΟΧΟΜΕΤΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

Στην παρούσα εργασία εφαρμόζονται τέσσερις διαφορετικές μέθοδοι χωρικής παρεμβολής, σε περιβάλλον ΓΠΣ και αξιολογούνται τα αποτελέσματά τους, στην χαρτογράφηση της ετήσιας και εποχιακής (υγρής και ξηράς περιόδου) βροχόπτωσης στον ελληνικό χώρο. Η υγρή περίοδος αναφέρεται στο εξάμηνο Οκτωβρίου – Μαρτίου κάθε έτους και η ξηρή στο διάστημα Απρίλιος – Σεπτέμβριος. Η επιλογή της χαρτογράφησης του ύψους βροχής δεν είναι τυχαία καθόσον πρόκειται για μία παράμετρο εν γένει μη συνεχή και ιδιαίτερα ευαίσθητη σε φυσιογραφικά (τοπογραφικά) χαρακτηριστικά. Είναι σαφές ότι η χωρική κατανομή των μετεωρολογικών και κλιματικών μεταβλητών εξαρτάται σε μεγάλο βαθμό από τις τοπικές συνθήκες, ιδίως σε περιοχές με έντονο ανάγλυφο (Tveito και Schöner, 2002). Ειδικά για την βροχόπτωση αυτή μπορεί να διαφοροποιείται από την ακτή προς την ενδοχώρα, μεταξύ μικρών και μεγάλων υψομέτρων αλλά και προσανατολισμών. Συνεπώς κάθε προσπάθεια για μια ακριβή, κατά το δυνατόν, χωρική κατανομή της σε κάθε περίπτωση θα πρέπει να λάβει υπόψη της και τις τοπογραφικές συνθήκες της περιοχής. Συμπερασματικά και σύμφωνα με όσα έχουν λεχθεί μέχρι τώρα θα πρέπει να διερευνηθεί η σχέση μεταξύ βροχόπτωσης και τοπογραφικών παραμέτρων όπως γεωγραφικό μήκος – πλάτος, υψόμετρο, κλίση, προσανατολισμός, ηπειρωτικότητα, απόσταση από τις ακτές –θάλασσα κλπ. Τα αποτελέσματα αφενός θα βοηθήσουν στην βελτίωση της κατανόησης του ρόλου των τοπογραφικών συνθηκών στην κατανομή της βροχόπτωσης και αφετέρου η γνώση αυτή θα συνεισφέρει στην εκτίμησή της σε περιοχές όπου δεν υφίστανται μετρητικά δεδομένα.

Η βροχόπτωση γενικά αυξάνεται με το υψόμετρο (Daly et al., 1994, Hutchinson, P., 1968, Hevesi et al., 1992a,1992b, Goodale et al., 1998, Goovaerts, 2000, Kyriakidis et al., 2001). Σχετικές μεθοδολογίες απαντώνται στη μέθοδο PRISM και στην AURELHY (Bénichou και Le Breton, 1987). Επεκτείνοντας την χρήση τοπογραφικών παραμέτρων στην εκτίμηση της χωρικής κατανομής της βροχόπτωσης, εφαρμογές, κυρίως με την μέθοδο της πολλαπλής παλινδρόμησης, συναντώνται στους Ninyerola et al, (2000), Perry et al. (2005) και Yamada (1990). Στα καθ' ημάς σύνθεση των παραπάνω παραμέτρων συναντάται στην εφαρμογή Geoklima (www.geoclima.eu). Στο σημείο αυτό θα πρέπει να αναφερθεί ότι το έντονο γεωγραφικό ανάγλυφο της Ελλάδας αποτελεί κρίσιμο στοιχείο στην αναζήτηση της μεθόδου που θα αποτυπώσει καλλίτερα την χωρική κατανομή των βροχοπτώσεων (ύψους βροχής). Οι Μαριολόπουλος και Καραπιπέρης (1955) είχαν σημειώσει ότι δεν υφίσταται σαφής νόμος της μεταβολής του υετού με το υψόμετρο και ως εκ τούτου είναι αδύνατη η αναγωγή των υψών βροχής στην επιφάνεια της θάλασσας. Οι Γκουβάς και Σακελλαρίου (2004) έδειξαν ότι η αύξηση των βροχοπτώσεων με το υψόμετρο δεν είναι απόλυτη και οφείλεται στην ύπαρξη θετικής συσχέτισης μεταξύ του υψομέτρου των σταθμών και του ανάγλυφου της γύρω από το σταθμό περιοχής.

Στην παρούσα εργασία για την διερεύνηση της χωρικής κατανομής των βροχοπτώσεων (ύψους βροχής) στον ελληνικό χώρο χρησιμοποιήθηκαν τα δεδομένα (μέσες μηνιαίες τιμές) 75 μετεωρολογικών σταθμών (74 της ΕΜΥ και ένας του Εθνικού Αστεροσκοπείου Αθηνών – Θησείο) για μια περίοδο είκοσι ετών (1981-2000).

Τα στατιστικά χαρακτηριστικά της βροχόπτωσης κάθε μελετώμενης περιόδου από του μετεωρολογικούς σταθμούς του ελληνικού χώρου παρέχονται στον Πίνακα 1. Η παράθεση αυτή κρίνεται αναγκαία για την κατά το δυνατόν περισσότερο «αντικειμενική» αξιολόγηση των εφαρμοζόμενων μεθόδων.

Η επιλογή της χρονικής περιόδου και του αριθμού των σταθμών σχετίζεται με την πληρότητα των δεδομένων βροχόπτωσης. Έτσι, η περίοδος χαρακτηρίζεται από σχεδόν απόλυτη πληρότητα δεδομένων ενώ παράλληλα τα δεδομένα έχουν «περάσει» από έναν ποιοτικό έλεγχο πριν την εισαγωγή τους στη βάση της ΕΜΥ. Οι μέσες μηνιαίες τιμές αποτελούν τον μέσο όρο των ανά μήνα και έτος αθροιστικών τιμών του ημερήσιου υετού (ύψους βροχής). Ο αθροιστικός ημερήσιος υετός αναφέρεται στο συνολικό ύψος βροχόπτωσης σε χιλιοστά μεταξύ των παρατηρήσεων 18UTC της προηγούμενης ημέρας μέχρι 18UTC της ημέρας αναφοράς. Για μία ποσοτική εκτίμηση της πυκνότητας του διαθέσιμου δικτύου εκτιμήθηκε ότι, η μέση απόσταση των διαθέσιμων σταθμών μέτρησης, δηλαδή η απόσταση ενός σταθμού από τον πλησιέστερό του, είναι περίπου 40Km (39233,69 m), με ελάχιστη τα 6Km (6140,85m) και μέγιστη τα 105Km (104905,74m). Επίσης για τις ανάγκες της εφαρμογής χρησιμοποιήθηκε ένα ψηφιακό μοντέλο εδάφους (DEM) ανάλυσης 100 μέτρων το οποίο προέρχεται από την ψηφιοποίηση ισοϋψών 1:50000 της Γεωγραφικής Υπηρεσίας Στρατού.

Πίνακας 1. *Παράμετροι περιγραφικής στατιστικής των βροχοπτώσεων στον ελληνικό χώρο βασισμένες στο δείγμα των διαθέσιμων 75 σταθμών παρατήρησης και για την περίοδο 1981-2000*

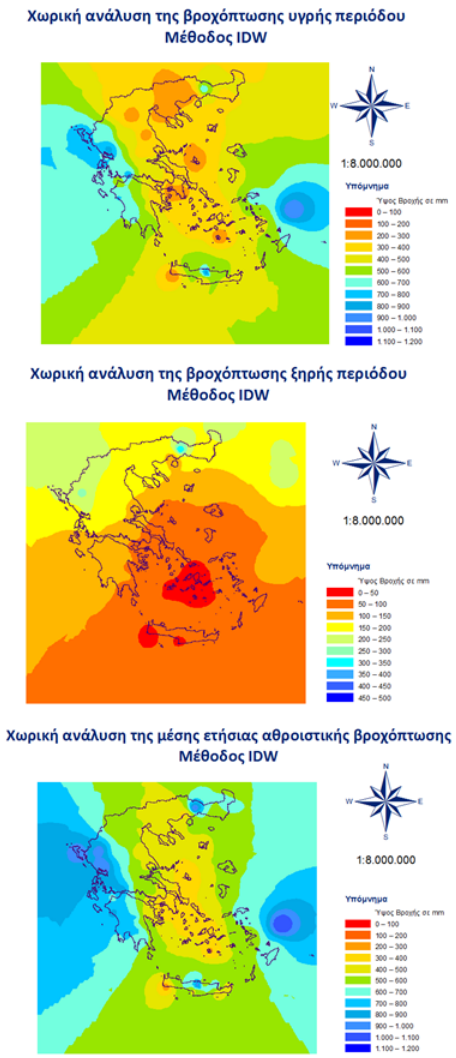
Βροχόπτωση	Έτος	Ξηρή περίοδος	Υγρή περίοδος
Mean	579,1353	116,9694	462,1658
Standard Error	22,51864	7,454647	19,39029
Median	554,0103	104,8128	430,235
Standard Deviation	195,0171	64,55914	167,9248
Range	818,4469	333,2722	700,2134
Minimum	244,2602	28,26049	195,8366
Maximum	1062,707	361,5327	896,0500

3.1 Η μέθοδος της «αντίστροφης απόστασης»

Η πρώτη μέθοδος που εφαρμόσθηκε ήταν της «αντίστροφης απόστασης», γνωστή με την συντομογραφία IDW στα περισσότερα λογισμικά που χρησιμοποιούνται. Η μέθοδος συνήθως εφαρμόζεται ως αντίστροφο τετράγωνο της απόστασης. Ας σημειωθεί ότι η ακανόνιστη κατανομή των θέσεων των μετρήσεων αποτελούσε και αποτε-

λεί το μείζον ζήτημα στην εφαρμογή των μεθόδων χωρικής ανάλυσης, καθόσον η συσσώρευση μετρήσεων προς μια κατεύθυνση οδηγεί σε ετεροβαρή αποτελέσματα.

Εικόνα 1: *Χαρτογράφηση της βροχόπτωσης με την μέθοδο IDW*



Ένα δεύτερο θέμα είναι η επιλογή των παρατηρήσεων που θα συνεισφέρουν στην εκτίμηση της τιμής σε μία θέση και εκφράζεται μέσω της «ακτίνας επίδρασης». Η ακτίνα επίδρασης προσεγγίζεται, είτε με βάση την απόσταση, είτε από τον ελάχιστο αριθμό μετρήσεων που απαιτούνται για την εκτίμηση μιας τιμής, είτε από συνδυασμό και των δύο.

Τέλος η χρήση της τιμής 2 (αντίστροφο τετράγωνο της απόστασης) πλέον τελεί υπό την αίρεση της βέλτιστης προσαρμογής καθόσον υφίστανται τα ανάλογα μέσα προσδιορισμού του σφάλματος της εκτίμησης (cross validation). Οι παραπάνω διαδικασίες έχουν ενσωματωθεί στη σχετική εργαλειοθήκη (geostatistical analyst tools) του ArcGIS και χρησιμοποιήθηκαν κατά την εφαρμογή της μεθόδου. Κατά πρώτον η ακτίνα επίδρασης ορίσθηκε μέσω του αριθμού των παρατηρήσεων που κατ' ελάχιστον θα πρέπει να συνεισφέρουν στην εκτίμηση κάθε άλλης θέσης και μέχρι ενός ανώτατου ορίου (π.χ. ελάχιστο 10, μέγιστο 15). Η ακανόνιστη κατανομή των παρατηρήσεων επιχειρείται να αμβλυνθεί με την ομαδοποίησή τους σε τομείς. Έτσι αυτομάτως οι λίγες παρατηρήσεις ενός τομέα αποκτούν, και είναι φυσικό, πολύ μεγαλύτερη αξία από τις πολλές ενός άλλου, χωρίς αυτό να σημαίνει πάντα και ακριβέστερο αποτέλεσμα.

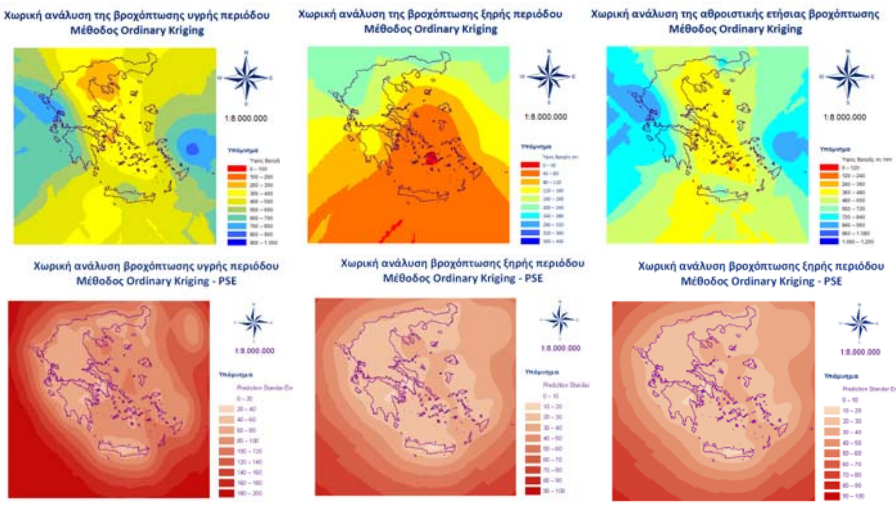
Η μέθοδος δοκιμάστηκε τόσο με τη χρήση τομέων όσο και χωρίς αυτούς. Επίσης με μεταβλητή ακτίνα επίδρασης που καθορίζονταν από τον ελάχιστο-μέγιστο αριθμό δεομένων που θα συνεισέφεραν στους υπολογισμούς. Τελικά, από την εφαρμογή της μεθόδου για την χωρική ανάλυση των βροχοπτώσεων της υγρής περιόδου, προέκυψε ως βέλτιστη η χρήση ακτίνας επίδρασης που θα περιλαμβάνει τουλάχιστον 10 και μέχρι 15 παρατηρήσεις και ως εκθέτη του συντελεστή βαρύτητας την τιμή 2,546864.

Η αξιολόγηση της μεθόδου έγινε με την Cross Validation (leave-one-out), γεγονός που σημαίνει ότι το σύνολο των διαθέσιμων παρατηρήσεων χρησιμοποιήθηκε στη διαδικασία της ανάλυσης. Με την ίδια διαδικασία εκτιμήθηκε η ανάλυση της ετήσιας βροχόπτωσης και αυτή της ξηρής περιόδου. Και στις περιπτώσεις αυτές η βροχόπτωση προσεγγίζεται καλλίτερα με την χρήση των ίδιων παραμέτρων όπως και στην περίπτωση της υγρής περιόδου. Η εφαρμογή της μεθόδου ολοκληρώνεται με την ανάλυση της ετήσιας βροχόπτωσης. Οι παραγόμενοι «βέλτιστοι» χάρτες φαίνονται στην εικόνα 1**.**

3.2 Μέθοδοι Kriging

Ως επιλογή εφαρμογής της μεθόδου Kriging παρουσιάζεται η Ordinary Kriging όπως αυτή υλοποιείται μέσω του λογισμικού ArcGIS. Το εν λόγω λογισμικό παρέχει ένα διαδραστικό περιβάλλον προσαρμογής των πλέον χρησιμοποιούμενων θεωρητικών μεθόδων προσαρμογής του ημιβαριογράμματος, βελτιστοποίησης της προσαρμογής, ελέγχου και άρσης της γεωμετρικής ανισοτροπίας και φυσικά αξιολόγησης της ακρίβειας της εκτίμησης μέσω της cross validation.

Εικόνα 2. *Χαρτογράφηση της βροχόπτωσης με την μέθοδο Ordinary Kriging και οι αντίστοιχοι χάρτες SPE*



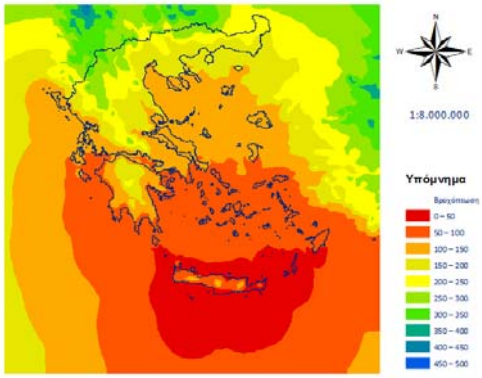
Η εφαρμογή παρέχει στα αποτελέσματά της τις βασικές παραμέτρους ελέγχου (μέσο σφάλμα και τετραγωνική ρίζα του τετραγώνου του μέσου σφάλματος-RMSE) αλλά και μέτρο εκτίμησης της ακρίβειας μέσω του τυπικού σφάλματος πρόβλεψης (Standard Error of Prediction – PSE), το οποίο περιγράφεται από την τυπική απόκλιση της εκτίμησης σε κάθε θέση παρατήρησης όπως αυτό προκύπτει μέσω της διαδικασίας cross validation (leave-one-out). Το αποτέλεσμα απεικονίζεται χαρτογραφικά παρέχοντας ένα επί πλέον εργαλείο αξιοπιστίας της μεθόδου, αφού η περιγραφόμενη χωρική κατανομή του τυπικού σφάλματος εκτίμησης, επιτρέπει να μην απορρίπτεται συλλήβδην το αποτέλεσμα, καθόσον ενδεχομένως η αστοχία σε μία περιοχή (για διάφορους λόγους όπως αραιά δεδομένα) δεν θα απορρίπτει την χρήση της για άλλες.

Τόσο η επιλογή του μοντέλου, όσο και αποδοχή ή όχι των προσαρμογών (ανισοτροπία) προέκυψαν μετά από διαδοχικές δοκιμές των προτεινόμενων εναλλακτικών και με γνώμονα τη βελτιστοποίηση του τελικού αποτελέσματος. Η εφαρμογή της μεθόδου έγινε χωρίς κανένα μετασχηματισμό των αρχικών δεδομένων και άρση τυχόν υφιστάμενων τάσεων (άλλωστε η χωρική ανάδειξή τους είναι το ζητούμενο). Για κάθε περίοδο παρουσιάζονται δύο χάρτες ο ένας αναφέρεται στην προτεινόμενη από το μοντέλο εκτίμηση και ο δεύτερος στην χωρική αξιοπιστία αυτής της εκτίμησης με την χωρική ανάλυση του SPE. Η ανάλυση της υγρής, ξηρής περιόδου και ετήσιας βροχόπτωσης και το αντίστοιχο τυπικό σφάλμα της εκτίμησης (PSE) περιγράφονται στην ομάδα χαρτών της Εικόνας 2.

3.3 Εφαρμογή γραμμικής παλινδρόμησης

Από τα πλέον βασικά θέματα, στην εφαρμογή της γραμμικής παλινδρόμησης, είναι η επιλογή των ανεξάρτητων μεταβλητών και εν προκειμένω των φυσιογραφικών παραμέτρων που σε κάποιο βαθμό αναμένεται να σχετίζονται με το ύψος βροχόπτωσης που δέχεται ένας τόπος. Οι παράμετροι που χρησιμοποιούνται, ως ανεξάρτητες μεταβλητές, στην παρούσα εργασία, είναι το υψόμετρο, το γεωγραφικό πλάτος, η απόσταση από τις ακτές και η ηπειρωτικότητα.

Εικόνα 3. *Χωρική ανάλυση της βροχόπτωσης ξηρής περιόδου με την χρήση ως ανεξάρτητων μεταβλητών στην OLS των παραγόντων του υψομέτρου, της απόστασης από τις ακτές και του γεωγραφικού πλάτους*



Για το υψόμετρο, με δεδομένο τα όσα έχουν αναφερθεί μέχρι τώρα, ορίστηκε ως η μέση τιμή του υψομέτρου σε μια περιοχή ακτίνας 10Km γύρω από τη θέση μέτρησης, ενώ η απόσταση από τη θάλασσα εκτιμάται ως η ελάχιστη ευκλείδεια απόσταση από την θάλασσα. Η ηπειρωτικότητα (continentality) εκφράζεται από τον Johansson Continentality Index, με βάση τη διαφορά μεταξύ της μέσης μηνιαίας θερμοκρασίας του θερμότερου μήνα από την αντίστοιχη του ψυχρότερου, συνήθως Ιουλίου – Ιανουαρίου) και το γεωγραφικό πλάτος του σταθμού. Στο σημείο αυτό θα πρέπει να ληφθεί υπόψη ότι, στην τελική φάση εκτίμησης, απαιτείται να είναι γνωστή η τιμή της σε κάθε θέση του χρησιμοποιούμενου ψηφιακού μοντέλου, συνεπώς θα πρέπει να έχει προηγηθεί ανάλογη παραγωγή του πεδίου των θερμοκρασιών. Κατά τον έλεγχο ανεξαρτησίας μεταξύ των ανεξάρτητων μεταβλητών, στον οποίο συμπεριλήφθηκε και η ηπειρωτικότητα, αποτυπώθηκε, η αναμενόμενη, υψηλή συσχέτιση της ηπειρωτικότητας με το γεωγραφικό πλάτος αλλά και την απόσταση από την ακτή με συνέπεια να παραληφθεί τελικώς από τις ανεξάρτητες μεταβλητές.

Διερευνώντας τις προϋποθέσεις εφαρμογής της (διερευνητική ανάλυση - OLS) , τόσον σε περιβάλλον ΓΠΣ όσο και με τη χρήση στατιστικού λογισμικού (SPSS), για την βροχόπτωση της υγρής περιόδου, προκύπτει πολύ χαμηλός συντελεστής προσδιορισμού είτε για το σύνολο των ανεξάρτητων μεταβλητών είτε για μέρος αυτών (R^2=0,17). Αντίστοιχα για την ετήσια αθροιστική βροχόπτωση προκύπτει R^2=0,14. Και στις δύο περιπτώσεις ο συντελεστής προσαρμογής καθιστούν επισφαλή τη χρήση της μεθόδου καθόσον είναι προφανές ότι οι ανεξάρτητες μεταβλητές δεν μπορούν να περιγράψουν την μεταβλητότητα της εξαρτημένης. Για την περίπτωση της βροχόπτωσης της ξηρής περιόδου προκύπτει R^2=0,66 που προδιαθέτει για ικανοποιητική προσαρμογή του μοντέλου.

Στη συνέχεια υλοποιήθηκε η OLS (σε περιβάλλον ArcGIS) μόνο για την βροχόπτωση ξηρής περιόδου και με την χρήση μόνο των τριών από τις τέσσερις ανεξάρτητες μεταβλητές (την απόσταση από την ακτή, το γεωγραφικό πλάτος και το μέσο υψόμετρο σε ζώνη 10Km). Ο βαθμός προσαρμογής του μοντέλου είναι R^2= 65,77% Παράλληλα, λόγω περιορισμού των διαθέσιμων υπολογιστικών πόρων, θα έπρεπε να γίνουν υπολογισμοί σε ένα πλήθος 1000Χ1020 σημείων, υποβιβάστηκε η ανάλυση του DEM στα 10km, όπου σε κάθε νέο κελί η τιμή του υψομέτρου ορίστηκε ως ο μέσος όρος των αρχικών τιμών, ώστε τελικά να μπορεί να χρησιμοποιηθεί άμεσα ως το μέσο υψόμετρο της περιοχής στη διαδικασία της παλινδρόμησης. Στη συνέχεια παρήχθη για κάθε σημείο του νέου DEM το εκτιμώμενο μέσω της ευθείας παλινδρόμησης ύψος βροχόπτωσης και το αποτέλεσμα απεικονίζεται χαρτογραφικά, μέσω της μεθόδου του τετραγώνου της αντίστροφης απόστασης. Κατά την παρεμβολή, για την αξιοπιστία της μεθόδου χρησιμοποιήθηκαν μόνο οι τέσσερις περιφερειακές συμμετρικές τιμές κάθε σημείου. Η αξιολόγηση της μεθόδου έγινε με παρεμβολή στις θέσεις των αρχικών δεδομένων και εκτίμηση των βασικών παραμέτρων (μέσο σφάλμα και RMSE). Η εκτίμηση του ύψους βροχόπτωσης προκύπτει από τη σχέση: ***Υψος Βροχόπτωσης(mm) = -768,836425+ 0,000409*απόσταση από τις ακτές + 0,000204*γεωγραφικό πλάτος + 0,092068*μέσο υψόμετρο σταθμού.***

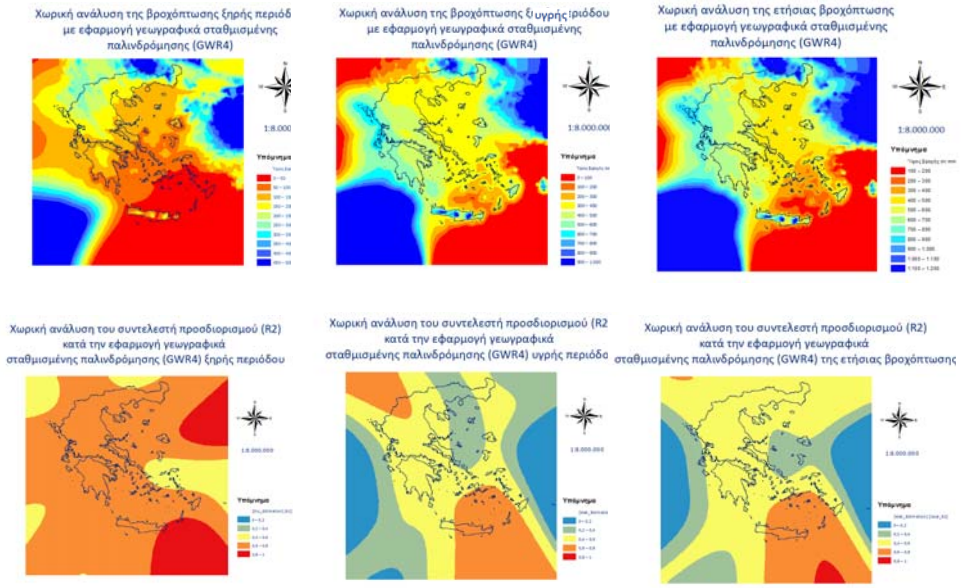
Η αξιολόγηση της μεθόδου (στις αρχικές θέσεις μέτρησης) έδωσε Μέσο σφάλμα: -1,941 και RMSE: 39,636. Η χαρτογραφική απεικόνιση των αποτελεσμάτων δίδεται στην εικόνα 3.

3.4 Γεωγραφικά σταθμισμένη παλινδρόμηση (GWR4)

Η γεωγραφικά σταθμισμένη παλινδρόμηση (Geographically Weighted Regre-ssion) επιτρέπει την εφαρμογή της κλασσικής γραμμικής παλινδρόμησης σε τοπικό επίπεδο λαμβάνοντας υπόψη ειδικά «τοπικά» χαρακτηριστικά. Φυσικά η έννοια «τοπικά» είναι πολύ σχετική και εξαρτάται εν πολλοίς από την χωρική κατανομή των υφιστάμενων μετρητικών δεδομένων. Σε μία επέκταση της μεθόδου μπορούν να ενσωματωθούν τόσο γενικοί όσο και τοπικοί παράγοντες. Η έκδοση που χρησιμοποιήθηκε είναι η GWR4 που αναπτύχθηκε από τον καθ. Tomoki Nakaya (http://gwr.nuim.ie). Η μέθοδος εφαρμόστηκε με τις προτεινόμενες ρυθμίσεις (μοντέλο Gauss) και με μόνο «τοπικές» μεταβλητές. Ο συντελεστής βάρους εξαρτάται εκθετικά από την ευκλείδεια απόσταση με αυτόματα ελεγχόμενο εύρος (bandwidth) με βάση την αξιοπιστία

του μοντέλου (corrected Akaike Information Criterion - http://gwr.nuim.ie/ downloads/ GWR_WhitePaper.pdf). Το βάρος της παρατήρησης i στη θέση j, εκτιμάται εκθετικά από τον λόγο των τετραγώνων της (ευκλείδειας) απόστασης μεταξύ i και j και το εύρος (ακτίνα) επίδρασης. Η επιλογή «Calculate Distance Band from Neighbor Count» έδωσε Μέγιστη απόσταση: 104906, Ελάχιστη: 6141 και Μέση: 39018 ενώ η εφαρμογή εντόπισε «Best bandwidth size 123866,837». Αν και στον οδηγό της προηγούμενης έκδοσης (3.0) η επιλογή Fixed Gaussian προέτρεπε να γίνεται σε καλώς κατανεμημένα στο χώρο δεδομένα (πχ gridded), προτιμήθηκε η επιλογή σταθερού εύρους ως συνηθέστερη στην ανάλυση μετεωρολογικών δεδομένων. Το ιδιαίτερο χαρακτηριστικό της GWR είναι ότι προσδιορίζονται τοπικοί συντελεστές παλινδρόμησης καθώς μπορεί να αναγνωρίσει χωρική συγγένεια (συνάφεια) των δεδομένων. Έτσι, η εκτίμηση βασίζεται σε γεωγραφικά σταθμισμένους συντελεστές και προφανώς η προσαρμογή (R^2) εμφανίζει χωρική μεταβλητότητα.

Εικόνα 4. *Χαρτογράφηση της βροχόπτωσης με εφαρμογή γεωγραφικά σταθμισμένης παλινδρόμησης (GWR4) οι αντίστοιχοι χάρτες χωρικής κατανομής του συντελεστή προσδιορισμού.*



Για την ανάλυση της βροχόπτωσης της ξηρής περιόδου εκτιμήθηκε ένα μέσο R^2= 84%, καλλίτερη συνεπώς προσαρμογή σε σχέση με την αντίστοιχη OLS. Η εφαρμογή εκτιμά τιμές σε κάθε άλλη θέση και έτσι κατ' αναλογία με την OLS εκτιμήθηκαν οι τιμές στα 1020 σημεία του υποβαθμισμένου (για λόγους υπολογιστικού κόστους) πεδίου, τα αποτελέσματα εισήχθησαν στο ArcGIS και παρήχθησαν οι σχετικοί χάρτες (και για τις τρεις μελετώμενες περιόδους). Οι εκτιμώμενες τιμές του μέσου σφάλματος και του RMSE, παράγονται από την επανεκτίμηση των τιμών στις αρχικές θέσεις παρατήρησης (ME= 0,116139 και RMSE=25,65646).

Αντίστοιχα για την βροχόπτωση υγρής περιόδου η μέθοδος έδωσε ικανοποιητικά χαρακτηριστικά προσαρμογής (σε αντίθεση με την OLS), μέσο R^2 = 60,39% . Επίσης Μέσο σφάλμα : 8,428151 και RMSE: 105,2044. Τέλος, από την εφαρμογή της μεθόδου για την ετήσια βροχόπτωση προέκυψαν μέσο R^2 = 59,30% , με μέσο σφάλμα : 8,716393 και RMSE: 123,6678. Στην εικόνα 4 απεικονίζονται οι παραγόμενοι, για κάθε περίπτωση χάρτες και η αντίστοιχη κατανομή του συντελεστή προσδιορισμού.

4. ΑΝΤΙΚΕΙΜΕΝΙΚΗ ΚΑΙ ΥΠΟΚΕΙΜΕΝΙΚΗ ΑΞΙΟΛΟΓΗΣΗ ΤΟΥ ΑΠΟΤΕΛΕΣΜΑΤΟΣ - ΣΥΜΠΕΡΑΣΜΑΤΑ

Οι αιτιοκρατικές μέθοδοι σαν κοινό παρονομαστή διαθέτουν το μέσο σφάλμα και την τετραγωνική ρίζα του τετραγώνου του μέσου σφάλματος παράμετροι που εκτιμώνται άμεσα από κοινό «εργαλείο» αξιολόγησης (Cross Validation – leave-one-out), το οποίο παρέχεται στο περιβάλλον του ArcGIS. Στις γεωστατιστικές μεθόδους (Ordinary Kriging) ο έλεγχος καλής προσαρμογής και η εκτίμηση του σφάλματος εμπεριέχονται στη διαδικασία. Στις μεθόδους παλινδρόμησης ο έλεγχος αξιοπιστίας προϋπάρχει του αποτελέσματος και είναι ο γνωστός έλεγχος καλής προσαρμογής (R^2) του γραμμικού μοντέλου και οι συνακόλουθοι έλεγχοι. Με δεδομένο ότι ο στόχος για βέλτιστη προσαρμογή σε κάθε περίπτωση φαντάζει ουτοπικός, η παραγωγή του μέσου σφάλματος και του RMSE στις θέσεις των αρχικών μετρήσεων αποτελεί εν δυνάμει κριτήριο αξιολόγησης. Με όλες τις αδυναμίες και ιδιαιτερότητες στη φύση και το αυστηρό θεωρητικό υπόβαθρο των μεθόδων η σύγκριση των αποτελεσμάτων τους σε επίπεδο ΜΕ και RMSE αποτελεί την πρώτη προσέγγιση αξιολόγησής τους (Πίνακας 2). Στην περίπτωση της OLS η τελική ύπαρξη clustered υπολοίπων αποτελεί τον βασικό λόγο αντιμετώπισης με σκεπτικισμό των αποτελεσμάτων καθόσον ενδεχομένως υφίστανται παράμετροι οι οποίοι προσδιορίζουν αυτή την συμπεριφορά και θα πρέπει να αναζητηθούν. Στη κατεύθυνση αυτή απαιτείται περισσότερη έρευνα και ειδικά στις περιπτώσεις της βροχόπτωσης της υγρής περιόδου και έτους, αφού οι επιλεγείσες ανεξάρτητες μεταβλητές ερμηνεύουν ελάχιστα την συμπεριφορά τους. Όσον αφορά την GWR θα πρέπει να σημειωθεί ότι η παρουσία clustered υπολοίπων στην OLS λειτουργεί απαγορευτικά στη χρησιμοποίησή της, αν και στον οδηγό που τη συνοδεύει προτείνεται η δοκιμή της. Συνεπώς και τα αποτελέσματά της αντιμετωπίζονται με επιφύλαξη. Η OLS δεν εφαρμόστηκε στην περίπτωση της υγρής περιόδου και του έτους καθόσον ο συντελεστής προσδιορισμού ήταν απαγορευτικός.

Με δεδομένες τις μέσες τιμές της παρατηρούμενης βροχόπτωσης κάθε περιόδου τα μέσα σφάλματα είναι απολύτως αποδεκτά σε όλες τις περιπτώσεις. Τα εκτιμώμενα σφάλματα προσεγγίζουν το μηδέν και τεκμηριώνουν την θεωρητική υποχρέωση για τον χαρακτηρισμό μιας μεθόδου ως βέλτιστης. Το RMSE με βάση τα υφιστάμενα εύρη των δεδομένων βροχόπτωσης είναι μάλλον υψηλό με εξαίρεση την περίπτωση της ξηρής περιόδου όπου γενικά η προσαρμογή όλων των μεθόδων κρίνεται ικανοποιητική.

Πίνακας 2. *Αντικειμενική αξιολόγηση των μεθόδων*

	Χωρική ανάλυση υγρής περιόδου		Χωρική ανάλυση ξηρής περιόδου		Χωρική ανάλυση ετήσιας βροχόπτωσης	
Μέθοδοι	ΜΕ	RMSE	ΜΕ	RMSE	ΜΕ	RMSE
IDW	-2,388	121,652	- 1,738	39,406	-4,402	151,225
Ordinary Kriging	-0,506	113,672	0,241	36,269	0,063	144,192
OLS	-	-	-1,941	39,636	-	-
GWR4	8,428	105,204	0,116	25,656	8,716	123,668

Οι βροχοπτώσεις στην Ελλάδα, όπως έχουν καταγραφεί ιστορικά, έχουν τα εξής χαρακτηριστικά. Στην αρχή της υγρής περιόδου η ατμοσφαιρική κυκλοφορία με την από δυτικά - νοτιοδυτικά κίνηση των βαρομετρικών συστημάτων αποδίδει υψηλά ποσά βροχοπτώσεων στη δυτική Ελλάδα. Η παρουσία του ορεινού όγκου της Πίνδου αποτελεί φραγμό στην επέκταση των βροχοπτώσεων στην ανατολική χέρσο χώρα, βροχοπτώσεις που στη συνέχεια εντοπίζονται στα νησιά του ανατολικού Αιγαίου και επιλεκτικά στη βόρεια Ελλάδα (κυρίως ανατολική Μακεδονία – Θράκη). Η σταδιακή στροφή της κυκλοφορίας σε βόρεια τους χειμωνιάτικους μήνες αποδίδει βροχές στα ανατολικά προσήνεμα της ηπειρωτικής χώρας και τα νησιά του Αιγαίου – Κρήτη. Με την έλευση της ξηρής περιόδου και τον σταδιακό περιορισμό της διέλευσης βαρομετρικών συστημάτων πάνω από τη χώρα, η ουσιαστική συνεισφορά στο υδατικό ισοζύγιο προέρχεται από τις απογευματινές κυρίως μπόρες, ως έκφραση θερμικής αστάθειας με ή χωρίς δυναμική υποβοήθηση. Από τον χαρακτήρα τους τα φαινόμενα αυτά αφορούν την ηπειρωτική Ελλάδα με έμφαση στις ορεινές περιοχές.

Με βάση τη γνώση αυτή κατά την χαρτογράφηση της βροχόπτωσης της υγρής περιόδου αναμένονται μεγαλύτερα ύψη βροχής στη δυτική Ελλάδα αλλά και τα νησιά του ανατολικού Αιγαίου και λιγότερα στην ανατολική ηπειρωτική χώρα και τα νησιά του Αιγαίου. Κατά την ξηρή περίοδο τα περισσότερα ποσά βροχής αναμένονται στο ηπειρωτικό κορμό, τα δυτικά παράλια θα πρέπει να έχουν λίγα, ενώ κάποια ποσά στα νησιά του ανατολικού Αιγαίου και τα Δωδεκάνησα είναι συνεισφορά της θερμικής αστάθειας της Μικράς Ασίας. Τα χωρικά αυτά πρότυπα απεικονίζονται και στους χάρτες κατανομής των υπολοίπων κατά την εφαρμογή της OLS. Οι παραγόμενοι χάρτες όλων των μεθόδων που εφαρμόστηκαν απεικονίζουν τα βασικά πρότυπα της κατανομής της βροχόπτωσης στον Ελληνικό χώρο. Με εξαίρεση την IDW, με τις γνωστές και αναμενόμενες συγκεντρώσεις γύρω από συγκεκριμένες θέσεις, οι υπόλοιπες απεικονίζουν ομαλά το πεδίο της βροχόπτωσης. Σημαντικό βοήθημα στην χρήση των αποτελεσμάτων αποτελούν οι χάρτες του τυπικού σφάλματος της εκτίμησης που παρέχουν την τοπική ακρίβεια της εκτίμησης.

Συμπερασματικά, με βάση τη διασπορά των θέσεων παρατήρησης, τις σημαντικές διακυμάνσεις της πυκνότητας του δικτύου αλλά και την ειδική γεωμορφολογία του

ελληνικού χώρου, τα αποτελέσματα κρίνονται ικανοποιητικά καθόσον απεικονίζουν την βασική κατανομή της βροχόπτωσης στον ελληνικό χώρο αναγνωρίζοντας υφιστάμενα πρότυπα. Παράλληλα κατέδειξαν την δυνατότητα που παρέχουν τα διαθέσιμα μέσα, στην ανάλυση σύνθετων θεμάτων, αν και η υψηλή ανάλυση απαιτεί και ανάλογη υπολογιστική ισχύ που έγινε κατανοητή κατά τη διάρκεια εκπόνησης της παρούσας εργασίας.

Στο σημείο αυτό θα πρέπει να σημειωθεί ότι τα αποτελέσματα της OLS, ακόμα και στην περίπτωση της ξηρής περιόδου όπου εμφανίζεται με έναν «αξιοπρεπή» δείκτη προσαρμογής, «δεν περνούν» τους σχετικούς ελέγχους. Όσον αφορά την GWR υλοποιείται στην παρούσα εργασία ως ένα επιπλέον εργαλείο χωρικής ανάλυσης, αφού όσο και ενθαρρυντικά αν παρουσιάζονται τα αποτελέσματά της θα πρέπει να αντιμετωπιστούν με επιφύλαξη λόγω της διασταλτικής αντιμετώπισης των προϋποθέσεων εφαρμογής της. Γενικά οι μέθοδοι παλινδρόμησης επιχειρούν να περιγράψουν την χωρική μεταβλητότητα ενός φαινομένου μέσω ενός συνόλου «ανεξάρτητων» μεταβλητών σε ένα πολύπλοκο περιβάλλον με ενδεχόμενες μη-γραμμικές σχέσεις παραμέτρων που δεν έχει προσδιοριστεί η συμμετοχή τους. Η παρουσία των τελευταίων, και η παράλειψή τους από ένα γενικό (OLS) μοντέλο πιθανόν να είναι η αιτία των διαφοροποιήσεων τοπικής κλίμακας. Συνεπώς, και ανεξάρτητα με την χρήση της GWR στην παρούσα εργασία όπως ήδη αναφέρθηκε, ένα τοπικής κλίμακας μοντέλο μπορεί να αναδείξει χαρακτηριστικά τα οποία δεν είχαν γίνει αντιληπτά στο μοντέλο γενικής κλίμακας επαναπροσδιορίζοντας το τελευταίο. Συνεπώς το μοντέλο γενικής κλίμακας δεν αντικαθίσταται από το τοπικό αλλά ενισχύεται από αυτό. Με αυτό το πνεύμα η GWR υλοποιεί γενική και τοπική παλινδρόμηση και αποτελεί ενδιαφέρον πεδίο έρευνας. Άλλωστε, όπως σαφώς αναφέρεται στο εγχειρίδιο χρήσης που τη συνοδεύει, απαιτεί αρκετή δουλειά ακόμα για την θεωρητική εκτίμηση και τεκμηρίωσή των αποτελεσμάτων της. Αλλά κυρίως απαιτούνται αξιόπιστα μετρητικά δεδομένα σε πυκνότερο δίκτυο. Στη κατεύθυνση αυτή, και με βάση τις δυνατότητες των μεθόδων περιοχικής - τοπικής ανάλυσης, προτεινόμενος στόχος για περαιτέρω έρευνα είναι η ανάπτυξη μικτών μεθόδων αξιολόγησης και αποτύπωσης της χωρικής ανάλυσης των κλιματικών παραμέτρων και ιδίως της βροχόπτωσης, εφαρμοζόμενων σε περιοχικό επίπεδο.

ABSTRACT

In this paper we attempt a brief description of the methods used for the spatial analysis of climate data. We also attempt to implement - evaluate four of these methods using three sets of rainfall data taken from a number of 75 meteorological stations of the Hellenic National Meteorological Service's network during the 1981-2000 period. Using a Geographic Information Systems-(GIS) environment for these data's spatial analysis, the corresponding pluviometric maps were produced. Given the dispersion of the positions of the observations, the significant fluctuations of the network's density, as also the specific geomorphology of Greece, these maps of annual and seasonal precipitation, for both wet and dry period, are deemed satisfactory as they reflect the

basic distribution of rainfall in Greece, not significantly deviating from established patterns.

ΑΝΑΦΟΡΕΣ

Bénichou P., Le Breton O. (1987). AURELHY : une méthode d'analyse utilisant le relief pour les besoins de l'hydrométéorologie. (http://horizon. documentation.ird.fr/exldoc/pleins_textes/pleins_textes_4/ colloques/25973.pdf - Τελευταία επίσκεψη 08/02/2014).

Cressie A.C.N. (1991). Statistics for spatial Data. J. Willey and Sons Inc.

Daly, C., R. P. Neilson, and D. L. Phillips (1994). A statistical- topographic model for mapping climatological precipitation over mountainous terrain, Journal of Applied Meteorology 33, 140- 158.

Daly, C. (2010). Overview of PRISM Spatial Climate Datasets. (https://www. ametsoc.org/meet/annual/annual90shortcourses/9.30am%20Daly.pdf. - Τελευταία επίσκεψη 20/03/2014).

Dobesch, H., P. Dumolard end I. Dryas, Eds. (2007). Spatial interpolation for climate data: the use of GIS in climatology and meteorology Geographical Information Systems Series. London: ISTE ltd.

Hevesi, J. A., J. D. Istok, and A. L. Flint (1992a). Precipitation estimation in mountainous terrain using multivariate geostatistics. Part I: Structural analysis, Journal of Applied Meteorology 31, 661 – 676.

Hevesi, J. A., J. D. Istok, and A. L. Flint (1992b). Precipitation estimation in mountainous terrain using multivariate geostatistics. Part II: Isohyetal Maps, Journal of Applied Meteorology 31, 677 – 688.

Hutchinson, P. (1968). An analysis of the effect of topography on rainfall in the Taieri Catchment area, Otago, Earth Science Journal 2, 51-68.

Gandin L.S. (1963). Objective analysis of meteorological fields. Leningrad. Hydromet. Press.

Goodale C.L., Aber J.D., S.V. Ollinger (1998). Mapping monthly precipitation, temperature, and solar radiation for Ireland with polynomial regression and a digital elevation model. Clim. Res. 10, 35-49.

Goovaerts, P. (2000). Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. Journal of Hydrology, 228, 113-129.

Gunst, F.R. (1995). Estimating Spatial Correlations from Spatial-Temporal Meteorological Data. AMS J. Climate 8, 2454-2470.

Kresic, N. and A. Mikszewski (2012). Hydrogeological Conceptual Site Models: Data Analysis and Visualization, CRC Press Taylor & Francis Group, 600pages.

Kyriakidis P.C., Jinwon Kim, N. L. Miller (2001). Geostatistical Mapping of Precipitation from Rain Gauge Data Using Atmospheric and Terrain Characteristics, J. Appl. Meteor., Vol. 40, 1855–1877.

Matheron, G. (1962). Traite de Geostatistique Appliquee, Tome I. Memoires du Bureau de Recherches Geologiques et Minieres. No. 14, Editions Techniq, 333 pp.

Michaelsen, J. (1987). Cross-Validation in Statistical Climate Forecast Models. J. Climate Appl. Meteor., 26, 1589–1600

Ninyerola M, Pons X, Roure J.M. (2000). A methodological approach of climatological modelling of air temperature and precipitation through GIS techniques. Int. J. Climatol. 20, 1823-1841.

Perry, M. and D. Hollis (2005). The generation of monthly gridded datasets for a range of climate variables over the UK, International Journal of Climatology 25, 1041 – 1054.

Refaeilzadeh, Payam, Lei Tang and Huan Liu (2009). Cross Validation. In Encyclopedia of Database Systems, Ed.: M. Tamer Özsu and Ling Liu, Springer, 2009. (http:// leitang.net/ papers/ - Τελευταία επίσκεψη 20/02/2014).

Sluiter, R. (2009). Interpolation methods for climate data. KNMI intern rapport IR 2009-04. http://www.knmi.nl/bibliotheek/knmipubIR/IR2009-04.pdf. (Τελευταία επίσκεψη 25/02/2014).

Tveito, O. E., and W. Schöner (2002). Application of spatial interpolation of climatological and meteorological elements by the use of geographical information systems (GIS), met.no REPORT NO. 28/02 KLIMA.

Tveito, O.E., M. Wegehenkel, F. Van der Wel & H. Dobesch (2006). The Use of geographic Information Systems in Climatology and Meteorology - Final Report COST Action 719. http://w3.cost.eu/fileadmin/domain_files/METEO/Action719/final_report/ final_report-719.pdf. (Τελευταία επίσκεψη 05/03/2014).

Wilks, D.S. (2011). Statistical methods in the atmospheric sciences, 3rd edition. Academic Press, 676pp.

Yamada, K. (1990). Estimation of monthly precipitation by geographical factors and meteorological variables, Hydrology in Mountainous Regions, I - Hydrological Measurements; the Water Cycle, 405 – 412. (Proceedings of two Lausanne Symposia, August 1990). IAHS Publ. no. 193.

Γκουβάς Μ. και Σακελλαρίου Ν. (2004). Σχέση του υψομέτρου των μετεωρολογικών σταθμών με το μέσο ετήσιο και μηνιαίο ύψος υετού. Πρακτικά 7ου Παν. Συν. Μετεωρολογίας, Κλιματολογίας και Φυσικής της Ατμόσφαιρας, Λευκωσία, 28-30 Σεπτεμβρίου 2004, Τόμος Β, σελ. 765-771.

Μαριολόπουλος Η. και Καραπιπέρης Λ. (1955). Αι βροχοπτώσεις εν Ελλάδι, Αθήναι.

Σκριμιζέας Π. (2014). Εφαρμογές χωρικής ανάλυσης κλιματικών δεδομένων. Χαρτογράφηση βροχομετρικών δεδομένων στον ελληνικό χώρο. Διπλωματική Εργασία, ΠΜΣ Γεωπληροφορική, Χαροκόπειο Πανεπιστήμιο, Αθήνα.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής (2015), σελ.310-317

ΑΞΙΟΛΟΓΗΣΗ ΤΟΥ ΕΚΤΙΜΗΤΗ ΜΕΓΕΘΟΥΣ ΔΕΙΓΜΑΤΟΣ ΜΕ ΤΗ ΜΕΘΟΔΟ BOOTSTRAP ΓΙΑ ΤΗΝ ΚΑΤΑΡΤΙΣΗ ΜΑΖΟΠΙΝΑΚΑ

Ζ. Τσαναξίδου, Κ. Μάτης, Γ. Σταματέλλος
Τμήμα Δασολογίας και Φυσικού Περιβάλλοντος, ΑΠΘ
{ztsanaxi, matis, stamatel}@for.auth.gr

ΠΕΡΙΛΗΨΗ

Στην κατάρτιση – εκτίμηση μοντέλων παλινδρόμησης, πολλές φορές το κόστος λήψης των στοιχείων είναι ένας αποφασιστικός παράγοντας. Αυτό έχει ιδιαίτερη σημασία στην κατάρτιση μαζοπινάκων για τη διαχείριση των δασικών οικοσυστημάτων. Στην έρευνα αυτή εκτιμάται ένα κατάλληλο μέγεθος δείγματος για την κατάρτιση ενός μαζοπίνακα, ο οποίος δίνει τις τιμές του όγκου των δέντρων, συναρτήσει εύκολα μετρούμενων χαρακτηριστικών τους. Αξιοποιείται ένα προκαταρκτικό δείγμα 50 παρατηρήσεων και χρησιμοποιείται μια μέθοδος των Demaerchalk και Kozak (1974). Η μέθοδος χρησιμοποιεί το διάστημα εμπιστοσύνης των προβλεπόμενων τιμών για διάφορες θεωρητικές κατανομές της ανεξάρτητης μεταβλητής. Από τις Bootstrap επαναλήψεις κατασκευάστηκαν BCa διαστήματα εμπιστοσύνης και εκτιμήθηκε η μεροληψία. Το μέγεθος δείγματος για κάθε κατανομή εκτιμήθηκε με ακρίβεια 6% και επιλέχτηκε εκείνο του πιο απαιτητικού παράγοντα. Για την κατάρτιση μαζοπινάκων διπλής εισόδου βρέθηκε ένα μέγεθος δείγματος 81 για την U κατανομή, 101 για την ομοιόμορφη κατανομή, 104 για την κανονική και 184 για τη Weibull κατανομή.

Λέξεις κλειδιά: μέγεθος δείγματος, Bootstrap, μαζοπίνακας, δασικά δεδομένα

1. ΕΙΣΑΓΩΓΗ

Ο προσδιορισμός του κατάλληλου μεγέθους δείγματος σε μια μελέτη-έρευνα είναι σημαντικός παράγοντας για οικονομικούς κυρίως λόγους. Στη διαχείριση γενικά των φυσικών πόρων, η επιλογή του μεγέθους δείγματος αποκτά εξαιρετική σπουδαιότητα. Ο υπολογισμός του παραγόμενου ξυλώδους κεφαλαίου που συνδέεται με την οικονομική εκμετάλλευση ενός δασικού οικοσυστήματος εκτιμάται με την εφαρμογή μοντέλων παλινδρόμησης. Οι Elsidig και Hetherington (1982), Cormier et al.(1992), Διαμαντοπούλου και Μάτης (1996), Διαμαντοπούλου (1996) και Kitikidou και Chatzilazarou (2008) χρησιμοποίησαν ένα σαφή τρόπο καθορισμού για το μέγεθος

δείγματος. Στις περισσότερες περιπτώσεις ωστόσο η κατάρτιση μαζοπινάκων γίνεται εκ των υστέρων από τις προβλεπόμενες υλοτομίες του διαχειριστικού σχεδίου για λόγους ευκολίας και κόστους χωρίς προηγούμενη εκτίμηση του μεγέθους δείγματος.

Σκοπός της εργασίας είναι η εκτίμηση του κατάλληλου μεγέθους δείγματος για κατάρτιση ενός μαζοπίνακα, που δίνει τις τιμές του όγκου των δέντρων συναρτήσει εύκολα μετρούμενων μεταβλητών τους και η αξιολόγηση του εκτιμητή με τη μέθοδο Bootstrap.

2. ΥΛΙΚΑ ΚΑΙ ΜΕΘΟΔΟΙ

Για την εκτίμηση του μεγέθους δείγματος χρησιμοποιείται ένα απλό τυχαίο δείγμα 50 παρατηρήσεων, το οποίο συλλέχθηκε από το δάσος «Αγίου Δημητρίου – Τετραλόφου – Ρυακίου» του Ν. Κοζάνης. Το δείγμα επιλέγεται από όλες κλάσεις των τιμών της ανεξάρτητης μεταβλητής των δέντρων οξιάς (*Fagus sylvatica L.*) του πληθυσμού. Η μέθοδος που εφαρμόστηκε, προτάθηκε από τους Demaerschalk και Kozak (1974) και χρησιμοποιεί το διάστημα εμπιστοσύνης του μέσου των προβλεπόμενων τιμών, για διάφορες θεωρητικές κατανομές της ανεξάρτητης μεταβλητής. Η απαιτούμενη επιθυμητή ακρίβεια, η οποία προαποφασίζεται, είναι ίση με ένα ποσοστό του μέσου της εξαρτημένης μεταβλητής. Για τη δειγματοληψία της ανεξάρτητης μεταβλητής έχουν προταθεί (Demaerschalk και Kozak 1974) και εφαρμοστεί (Μάτης και Διαμαντοπούλου 1995, Διαμαντοπούλου 1996, Kitikidou και Chatzilazarou 2008) οι παρακάτω κατανομές συχνοτήτων: η κατανομή σχήματος U (N_1), η ομοιόμορφη (N_2), η κανονική (N_3) και η Weibull (N_4).

Η επιλογή του μοντέλου παλινδρόμησης έγινε εξ αυτών που συνήθως χρησιμοποιούνται στη βιβλιογραφία για τους μαζοπίνακες διπλής εισόδου:

$$V = b_0 + b_1 d^2 h \quad (1)$$

όπου *V:* ο όγκος του δέντρου, d^2h*:* η στηθιαία διάμετρος (σε ύψος 1,3m από το έδαφος) στο τετράγωνο επί το συνολικό ύψος και b_0*,* b_1*:* οι συντελεστές παλινδρόμησης.

Για τον έλεγχο της κανονικότητάς των πρωτογενών δεδομένων χρησιμοποιήθηκε το τεστ Lilliefors (Lilliefors 1967, Conover 1980) που αποτελεί μία παραλλαγή του μη παραμετρικού ελέγχου Kolmogorov-Smirnov για ένα δείγμα για τις περιπτώσεις που ο μέσος όρος και η διασπορά του πληθυσμού δεν είναι γνωστές και εκτιμώνται από ένα δείγμα.

Με τη μέθοδο των Demaerschalk και Kozak (1974) για την εκτίμηση του μεγέθους δείγματος χρησιμοποιείται το διάστημα εμπιστοσύνης w_i, του μέσου των προβλεπόμενων τιμών της ανεξάρτητης μεταβλητής.

$$w_i = 2t_{n-2,\alpha/2}\,\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(X_i-\overline{X}_j)^2}{SSx}} \quad (2)$$

όπου $t_{n-2,a/2}$: η τιμή της t-κατανομής για n-2 βαθμούς ελευθερίας και επίπεδο σημαντικότητας α, $\hat{\sigma}$: το τυπικό σφάλμα εκτίμησης από την εφαρμογή της παλινδρόμησης στο δείγμα, X_i : η i τιμή της ανεξάρτητης μεταβλητής, $\overline{X}_j$: ο μέσος όρος της X για κάθε κατανομή δειγματοληψίας της και

$$SSx = \sum_{i=1}^{k+1} \sum_{l=1}^{n_i} (X_{il} + \overline{X}_j) = V_j n \tag{3}$$

όπου V_j : η διασπορά της σχετικής κατανομής για κάθε j κατανομή, $k+1$: ο αριθμός των τιμών που παίρνει η ανεξάρτητη μεταβλητή

Η απαιτούμενη επιθυμητή ακρίβεια, είναι ίση με ένα ποσοστό του μέσου της εξαρτημένης μεταβλητής και καθορίζεται από το μέγιστο πλάτος του διαστήματος εμπιστοσύνης W_i . Η σχέση που συνδέει το απαιτούμενο μέγεθος δείγματος με το μέγιστο πλάτος W_i του διαστήματος εμπιστοσύνης προκύπτει από την εξίσωση (2), με αντικατάσταση σε αυτή της (3) οπότε προκύπτει η παρακάτω εξίσωση :

$$W_i = 2t_{n-2,\alpha/2} \hat{\sigma} \sqrt{\frac{1 + (X_i - \overline{X}_j)^2 / V_j}{N}} \tag{4}$$

όπου N : το απαιτούμενο μέγεθος δείγματος

Μικρή αλγεβρική επεξεργασία της εξίσωσης (4) δίνει την εξίσωση από την οποία υπολογίζεται το μέγεθος δείγματος, για κάθε τιμή της ανεξάρτητης μεταβλητής X.

$$N = \frac{2t_{n-2,\alpha/2}\, \hat{\sigma}^2 (1 + (X_i - \overline{X}_j)^2 / V_j)}{W_i^2} \tag{5}$$

όπου $\hat{\sigma}^2$: η εκτίμηση της υπολειπόμενης διασποράς από την εφαρμογή της παλινδρόμησης στο δείγμα.

Η Bootstrap χρησιμοποιεί την εμπειρική κατανομή του δείγματος ως μια εκτίμηση της πραγματικής αλλά άγνωστης κατανομής του πληθυσμού και με βάση αυτή γίνεται δειγματοληψία με επανάθεση παίρνοντας πολλές επαναλήψεις. Από τις Bootstrap επαναλήψεις και τα αντίστοιχα δείγματα που δημιουργούνται υπολογίζεται μια τιμή $\hat{\theta}^*$ για κάθε επανάληψη. Οι τιμές αυτές θεωρούνται μια καλή προσέγγιση της κατανομής της θ και μπορεί να κατασκευαστούν διαστήματα εμπιστοσύνης για τη $\hat{q}$. Η εκτίμηση της μεροληψίας (bias) του εκτιμητή $\hat{\theta}$ δίνεται ως η διαφορά της αναμενόμενης τιμής της $\hat{q}$ από την τιμή της παραμέτρου θ. Η ακρίβεια των διαστημάτων εμπιστοσύνης εξαρτάται από την κατανομή της παραμέτρου αν π.χ. είναι συμμετρική και τη μεροληψία της. Σύμφωνα με τα προηγούμενα επιλέγονται να κατασκευαστούν τα BCa διαστήματα εμπιστοσύνης που υπολογίζονται με τη μέθοδο μεροληψίας και επιτάχυνσης (Bias corrected and accelerated) (Efron 1987).

Η εφαρμογή της μη παραμετρικής Bootstrap μεθόδου (Davison και Hinkley 1997, Canty και Ripley 2014) και η ανάλυση παλινδρόμησης έγινε μέσω του R-Studio με ένα πρόγραμμα που γράφτηκε σε γλώσσα R, χρησιμοποιώντας τις απαραίτητες βιβλιοθήκες (Ricci 2005, Fox και Weisberg 2011).

3. ΑΠΟΤΕΛΕΣΜΑΤΑ

Περιγραφικά στατιστικά για το δείγμα των 50 δέντρων παρουσιάζονται στον Πίνακα 1. Η μέση τιμή του συνολικού όγκου είναι 0,4802κ.μ. με τυπική απόκλιση 0,3092 και συντελεστή κύμανσης 64,39%. Η τιμή της λοξότητας είναι θετική και ίση με 1,096 που δείχνει μια θετική ασυμμετρία στην κατανομή. Από την τιμή της κύρτωσης 1,033 συμπεραίνουμε ότι η κατανομή της μεταβλητής όγκος είναι γενικά πλατύκυρτη.

Πίνακας 1*.Περιγραφικά στατιστικά του δείγματος*

Μετα-βλητή	**Μέσος όρος**	**Ελάχιστη τιμή**	**Μέγιστη τιμή**	**Τυπική απόκλιση**	**Λοξότητα**	**Κύρτωση**
V(κ.μ.)	0,4802	0,0807	1,3964	0,3092	1,096	1,033
d (μ.)	0,2908	0,1410	0,4360	0,0793	-0,088	-0,917
h (μ.)	15,2	8,0	23,5000	3,9615	0,168	-0,849
$\boldsymbol{d^2h\ (\mu.)^3}$	1,4979	0,2088	4,3722	1,0361	0,944	0,534

Για τις μεταβλητές: συνολικός όγκος, στηθιαία διάμετρος και συνολικό ύψος έγινε ο έλεγχος κανονικότητας με το τεστ Lilliefors για την εξέταση της κανονικότητας του πληθυσμού από όπου προέρχονται τα δεδομένα και δίνεται η εικόνα μιας καλής προσαρμογής των δεδομένων στην κανονική κατανομή.

Το μέγεθος δείγματος για κάθε κατανομή εκτιμήθηκε, αφού καθορίστηκε η επιθυμητή ακρίβεια σαν ποσοστό της μέσης τιμής της εξαρτημένης μεταβλητής ίση με 6%. Για κάθε κατανομή επιλέχθηκε το μέγεθος δείγματος του πιο απαιτητικού παράγοντα (ανεξάρτητη μεταβλητή) της κατανομής, για την οποία η τιμή του λόγου w_i/W_i γίνεται μέγιστη. Για την κατάρτιση μαζοπίνακα διπλής εισόδου προκύπτει μέγεθος δείγματος ίσο με N_1= 81 για την κατανομή σχήματος U, N_2=101 για την ομοιόμορφη κατανομή, N_3=104 για την κανονική κατανομή και N_4=184 για την Weibull κατανομή όπως φαίνεται στον Πίνακα 2. Το μικρότερο μέγεθος δείγματος δίνει η κατανομή σχήματος U και αυτό κατανέμεται στις διάφορες βαθμίδες διαμέτρου σύμφωνα με αυτή την κατανομή, ώστε να ληφθεί το δείγμα και να γίνει η εκτίμηση του μοντέλου παλινδρόμησης. Στον Πίνακα 3, φαίνεται το ελάχιστο μέγεθος δείγματος στις διάφορες βαθμίδες της ανεξάρτητης μεταβλητής για την ανάλυση παλινδρόμησης και διακρίνονται δυο περιπτώσεις. Στην πρώτη κατανομή γίνεται η στρογγυλοποίηση προς τα πάνω σε κάθε βαθμίδα της ανεξάρτητης μεταβλητής για να διατηρηθεί η ομοιομορφία στην κατανομή και το ελάχιστο μέγεθος γίνεται ίσο με 88 δέντρα, επτά δέντρα πάνω από τον εκτιμητή. Εάν επιλεγεί

να γίνει στρογγυλοποίηση στον πλησιέστερο ακέραιο, το δείγμα γίνεται ίσο με 80 δέντρα και το δέντρο που υπολείπεται κατανέμεται με τυχαίο τρόπο σε μια από τις βαθμίδες, με αποτέλεσμα μικρή αλλαγή στη μορφή της αρχικής κατανομής.

Πίνακας 2. *Αρχική εκτίμηση μεγέθους δείγματος για τις 4 κατανομές, εκτιμήσεις από τη Bootstrap και τυπικό σφάλμα της εκτίμησης, μεροληψία, Bca, Percentile και Normal Διαστήματα Εμπιστοσύνης*

	Σχήματος U	**Ομοιόμορφη**	**Κανονική**	**Weibull**
Αρχική Εκτίμηση	81	101	104	184
Εκτίμηση από Bootstrap	80	97	101	177
Τυπικό Σφάλμα	19,448	24,250	26,046	44,801
Μεροληψία	-1,843	-4,316	-3,415	-7,458
Bca Διάστημα Εμπιστοσύνης	(51 133)	(66 182)	(68 190)	(118 337)
Percentile Δ.Ε.	(46 121)	(55 148)	(55 155)	(98 270)
Normal Δ.Ε.	(45 121)	(58 153)	(57 159)	(104 280)

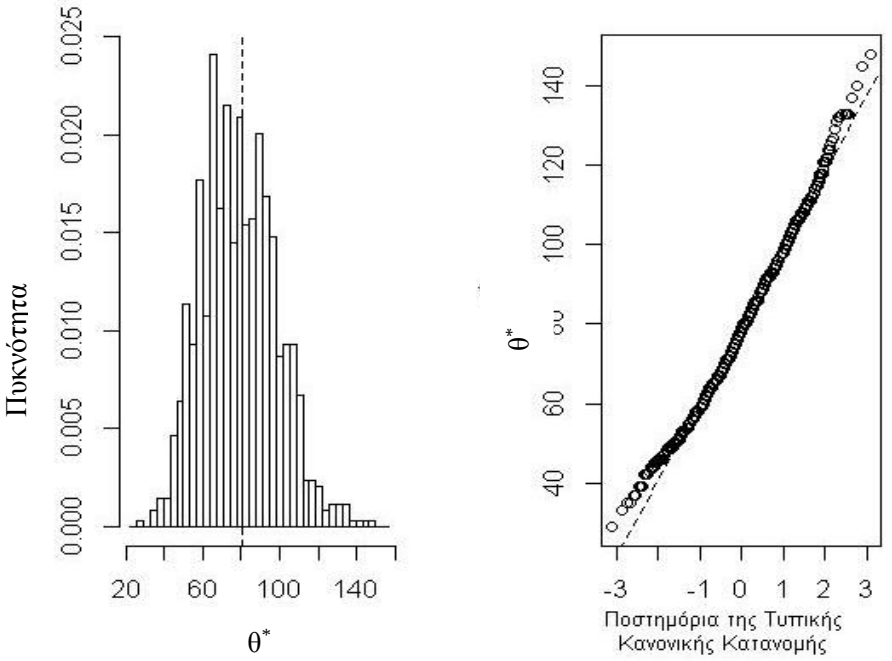
Οι μέσες τιμές για τον εκτιμητή από τη Bootstrap προσεγγίζουν σε αυτές που έδωσε το δείγμα όπως φαίνεται στον Πίνακα 2. Οι αποκλίσεις που παρατηρούνται δε θεωρούνται μεγάλες παρότι η εκτίμηση γίνεται από παραμέτρους που εκτιμώνται με τη μέθοδο της παλινδρόμησης όπου εισέρχεται η μεταβλητότητα όγκου των δέντρων η οποία είναι μεγάλη. Το τυπικό σφάλμα και η μεροληψία έχουν μικρότερη τιμή για τον εκτιμητή που τελικά επιλέγεται.

Πίνακας 3. *Κατανομή της αρχικής εκτίμησης του δείγματος ανά βαθμίδα διαμέτρου για τις δυο περιπτώσεις*

Xi	4,3722	3,7774	3,1827	2,5879	1,9931	1,3983	0,8036	0,2088
N ανά βαθμίδα (1)	17	13	9	5	5	9	13	17
N ανά βαθμίδα (2)	16	12	8	4	4	8	12	16

Στο Σχήμα 1 φαίνεται η κατανομή των εκτιμώμενων τιμών από τις 1000 Bootstrap επαναλήψεις και το Q-Q γραφικό κανονικής πιθανότητας, όπου φαίνεται να υπάρχει μια προσέγγιση της κανονικής κατανομής.

Σχήμα 1. *Κατανομή των εκτιμώμενων τιμών από μια Bootstrap επανάληψη και το Q-Q γραφικό κανονικής πιθανότητας για τη σχήματος U κατανομή της ανεξάρτητης μεταβλητής*



Όσον αφορά τις προϋποθέσεις για την ανάλυση παλινδρόμησης ελέγχθηκε η κανονικότητα και η ομοιογένεια της διασποράς των υπολοίπων. Η υπόθεση της κανονικότητας ικανοποιείται αφού η τιμή του στατιστικού Lilliefors D= 0,0823 και p-value=0,5409, ενώ δεν ικανοποιείται η υπόθεση της ομοιογένειας της διασποράς χ^2=4,945 και p=0,0262.

4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Το μέγεθος δείγματος για την κατάρτιση μαζοπίνακα εκτιμήθηκε ίσο με 81 δέντρα τα οποία θα επιλεγούν έτσι ώστε η ανεξάρτητη μεταβλητή να ακολουθεί την κατανομή σχήματος U. Η κατανομή αυτή επιλέχθηκε διότι έχει τη μεγαλύτερη σχετική αποτελεσματικότητα. Με αυτή την κατανομή δίνεται περισσότερο βάση στα άκρα της κατανομής του δείγματος τα οποία επηρεάζουν περισσότερο την εκτίμηση του μοντέλου παλινδρόμησης. Η εφαρμογή της Bootstrap έδωσε εκτίμηση του μεγέθους δείγματος πολύ κοντά στην αρχική εκτίμηση με μικρότερο τυπικό σφάλμα και μεροληψία, για την κατανομή που τελικά επιλέγεται. Λόγω της μικρής μεροληψίας τα BCa διαστήματα εμπιστοσύνης βρέθηκαν μεγαλύτερα σε σχέση με τα Normal και τα Percentile διαστήματα. Ενδέχεται το μεγάλο διάστημα εμπιστοσύνης

να οφείλεται στην ανομοιογένεια της διασποράς, ερευνητικό ερώτημα το οποίο θα εξεταστεί σε μελλοντική έρευνα.

ABSTRACT

In training – estimating regression models, many times the cost of taking the data is a decisive factor. This is particularly important in constructing volume tables for management of forest ecosystems. In this research, a suitable sample size is estimated for constructing volume tables, which give the values of the volume of trees, as a function of easily measured characteristics. A preliminary sample of 50 observations is exploited and the method of Demaerchalk and Kozak (1974) is used. That method uses the confidence interval of predicted values for various theoretical distributions of the independent variable. BCa confidence intervals were constructed from Bootstrap replications and their bias was estimated. The sample size for each distribution was assessed with accuracy of 6% and the one of the most demanding factor was chosen. For the construction of dual input volume tables a sample size of 81 for the U shape distribution, 101 for the uniform distribution, 104 for the normal and 184 for Weibull distribution was estimated.

ΑΝΑΦΟΡΕΣ

Canty, A. and Ripley, B. (2014). *boot: Bootstrap Functions (originally by Angelo Canty for S)*. R package version 1.3-11. http://CRAN.r-project.org/web/packages/boot/index. html

Conover, J.W. (1980). *Practical Nonparametric Statistics*. 2nd ed. USA: John Wiley and Sons, Inc.

Cormier, K.L. Reich, R.M. Czaplewski, R.L., Bechtold, W.A. (1992). Evaluation of weighted regression and sample size in developing a taper model for loblolly pine. *Forest Ecology and Management*. **53**,65-76.

Demaerschalk, J.P. and Kozak, A. (1974). Suggestions and Criteria for More Effective Sampling. *Canadian Journal of Forest Research*, **4**(3), 341-348.

Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap Methods and Their Applications*. Cambridge: Cambridge University Press.

Διαμαντοπούλου, Μ.Ι. (1996). *Δασοβιομετρικά μοντέλα για την ελάτη του Πανεπιστημιακού Δάσους Περτουλίου*. Διδακτορική διατριβή. Α.Π.Θ. Θεσσαλονίκη.

Διαμαντοπούλου, Μ.Ι. και Μάτης, Κ.Γ. (1996). Καθορισμός του μεγέθους του δείγματος για τη σχέση υψομορφαρίθμου – ύψους δέντρων οξιάς Σιδηροκάστρου και αριθμός τιμών της ανεξάρτητης μεταβλητής. Αξιοποίηση Δασικών Πόρων. *Πρακτικά 7ου Πανελλήνιου Δασολογικού Συνεδρίου. Καρδίτσα.* 364-371.

Efron, B. (1987). Better Bootstrap Confidence Intervals. *Journal of American Statistical Association.* **82**(397), 171-185.

Efron, B. and Tibshirani, R. (1994). *An Introduction to the Bootstrap*, New York: Chapman and Hall.

Elsiddig, E.A. and Hetherington, J.C. (1982). *The Stem and the Brunch Volume of Acacia Nicotica in the Fung Region in Sudan.* Univ. College of North Wales, Dep. Of Forestry and Wood Science.

Fox, J. and Weisberg, S. (2011). *An {R} Companion to Applied Regression, Second Edition.* Thousand Oaks CA: Sage.

Kitikidou, K. & Chatzilazarou, G. (2008). Estimating the sample size for fitting taper equations. *Journal of Forest Science*, **54**(4), 176–182.

Lilliefors, H. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association.* **62**, 399-402.

Μάτης, Κ.Γ. και Διαμαντοπούλου, Μ.Ι. (1995). Μέγεθος δείγματος για τη σχέση ύψους διαμέτρου δέντρων οξιάς Αριδαίας. *Πρακτικά 8ου Πανελλήνιου Συνεδρίου Στατιστικής. Δελφοί.* 168-177.

Ricci, V. (2005). *Fitting Distributions with R.* http://CRAN.r-project.org/doc/contrib/Ricci-distributions-en.pdf

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής (2015), σελ.318-325

ΒΕΛΤΙΣΤΟΙ ΣΧΕΔΙΑΣΜΟΙ ΓΙΑ ΤΗΝ ΕΚΤΙΜΗΣΗ ΤΩΝ ΑΝΤΙΘΕΣΕΩΝ ΣΕ 2^k ΚΛΑΣΜΑΤΙΚΟΥΣ ΣΧΕΔΙΑΣΜΟΥΣ

Β. Χασιώτης[1], Σ. Κουνιάς[2], Ν. Φαρμάκης[1]
[1] Τμήμα Μαθηματικών, Πανεπιστήμιο Θεσσαλονίκης
chasiotisv@math.auth.gr, farmakis@math.auth.gr
[2] Τμήμα Μαθηματικών, Πανεπιστήμιο Αθηνών
skounias@math.uoa.gr,

ΠΕΡΙΛΗΨΗ

Στους παραγοντικούς σχεδιασμούς με k παράγοντες, ο καθένας σε δύο στάθμες (επίπεδα), το ενδιαφέρον μας είναι στην εκτίμηση των αντιθέσεων των επιδράσεων του κάθε παράγοντα. Δίνεται το μοντέλο και ο πίνακας πληροφορίας Q των προς εκτίμηση παραμέτρων, οι παρατηρήσεις είναι ασυσχέτιστες. Υπάρχει αρκετή βιβλιογραφία για την εκτίμηση του γενικού μέσου και των αντιθέσεων των επιδράσεων του καθένα από τους k παράγοντες, αλλά όχι για την εκτίμηση μόνο των αντιθέσεων. Δίνονται οι βέλτιστοι σχεδιασμοί για την εκτίμηση των αντιθέσεων και η κατασκευή τους, όταν το πλήθος των παρατηρήσεων είναι $n \equiv 0,1,2,3 \bmod 4$. Παρουσιάζονται παραδείγματα για την κατανόηση της εφαρμογής σε συγκεκριμένες περιπτώσεις.

Λέξεις κλειδιά: Πίνακας πληροφορίας, φ βέλτιστοι σχεδιασμοί, πίνακας σχεδιασμού.

1. ΕΙΣΑΓΩΓΗ ΚΑΙ ΤΟ ΜΟΝΤΕΛΟ

Στους 2^k κλασματικούς παραγοντικούς σχεδιασμούς υπάρχουν k παράγοντες, ο καθένας σε δύο στάθμες, που θα τις συμβολίζουμε με 1 (χαμηλή στάθμη) και 2 (υψηλή στάθμη). Αν ο παράγοντας i μετέχει σε μια παρατήρηση στη χαμηλή στάθμη, η επίδρασή του είναι l_i και αν μετέχει στην υψηλή στάθμη η επίδρασή του είναι $q_i \quad i = 1,...,k$.

Σε σχεδιασμούς αναλυτικής τάξης (resolution) III, περιλαμβάνονται μόνο κύριες επιδράσεις και το μοντέλο είναι,

$$y_s = \mu + \sum_{i=1}^{k} \tau_{d(i,s)} + e_s \quad s = 1,...,k \qquad (1)$$

$\tau_{d(i,s)}$ είναι η επίδραση του παράγοντα i στην s παρατήρηση όταν εφαρμόζεται ο σχεδιασμός d, μ είναι ο γενικός μέσος, y_s είναι το αποτέλεσμα της παρατήρησης s και e_s είναι το τυχαίο σφάλμα με μέση τιμή 0 και διασπορά σ^2.

Κάθε παράγοντας μετέχει σε κάθε μία από τις s παρατηρήσεις, με μια από τις δύο στάθμες. Το μοντέλο (1) γράφεται και στη μορφή,

$$y_s = \sum_{i=1}^{k} \left(l_i x_{i,s} + q_i z_{i,s} \right) + e_s \quad s = 1,...,n \qquad (2)$$

$$x_{i.s} = \begin{cases} 1 & \text{αν} \quad d(i,s) = 1 \\ 0 & \text{αν} \quad d(i,s) = 2 \end{cases}, \quad z_{i.s} = \begin{cases} 0 & \text{αν} \quad d(i,s) = 1 \\ 1 & \text{αν} \quad d(i,s) = 2 \end{cases} \quad s = 1,...,n, \ i = 1,...,k$$

Επειδή $x_{is} + z_{is} = 1$, αν $k \geq 2$ οι παράμετροι $l_{i,}, q_i$, $i = 1,...,k$ δεν μπορεί να εκτιμηθούν, d(i,s) είναι το επίπεδο του i παράγοντα στην s παρατήρηση

Ο σχεδιασμός d είναι ένας σχηματισμός με n γραμμές και k στήλες με στοιχεία 1,2. Οι γραμμές αντιστοιχούν στις παρατηρήσεις και οι στήλες στους παράγοντες.

Παράδειγμα 1. Τέσσερις παρατηρήσεις τρεις παράγοντες,

$$d: \left[\begin{array}{ccc} 1 & 2 & 3 \\ \hline 1 & 2 & 2 \\ 2 & 1 & 2 \\ 1 & 1 & 2 \\ 2 & 1 & 2 \end{array}\right]$$ Στη δεύτερη παρατήρηση ο παράγοντας 3 μετέχει με τη στάθμη 2.

Στη μελέτη αυτή μας ενδιαφέρει να εκτιμήσουμε τις γραμμικές αντιθέσεις $q_i - l_i, \quad i = 1,...,k,$ που είναι η διαφορά των επιδράσεων του κάθε παράγοντα. Προς τούτο το μοντέλο (2) μετά από γραμμικό μετασχηματισμό γράφεται,

$$\mathbf{y} = \mu^* \mathbf{1}_n + \sum_{i=1}^{k} v_i \mathbf{w}_i + \mathbf{e} \qquad (3)$$

$v_i = (q_i - l_i)/2, \ \mathbf{w}_i = \mathbf{z}_i - \mathbf{x}_i = (w_{i1}, \cdots, w_{in})', \mathbf{y} = (y_1, \cdots, y_n)', \quad i = 1,...,k$ και

$$w_{is} = \begin{cases} -1 & \text{αν } \mathrm{d}(\mathrm{i},\mathrm{s}) = 1 \\ 1 & \text{αν } \mathrm{d}(\mathrm{i},\mathrm{s}) = 2 \end{cases} \quad i = 1,\cdots,k, \;\; s = 1,\cdots,n$$

(•) Αν $\sigma^2\mathbf{V}$ είναι ο πίνακας διασποράς των υπό εκτίμηση παραμέτρων, τότε ο $\mathbf{Q} = \mathbf{V}^{-1}$ ονομάζεται πίνακας πληροφορίας του σχεδιασμού.

Ορισμός 1. Αν $\mathbf{a} = (\lambda_1 \le \lambda_2 \le \cdots \le \lambda_m)$, $\mathbf{b} = (\mu_1 \le \mu_2 \le \cdots \le \mu_m)$ είναι δύο διανύσματα πραγματικών αριθμών, τότε το $\mathbf{a}$ κυριαρχείται από το $\mathbf{b}$ και γράφουμε $\mathbf{a} \prec \mathbf{b}$ αν $\lambda_1 \ge \mu_1,\ \lambda_1 + \lambda_2 \ge \mu_1 + \mu_2, \cdots, \lambda_1 + \cdots + \lambda_{m-1} \ge \mu_1 + \cdots + \mu_{m-1}$, $\lambda_1 + \cdots + \lambda_m = \mu_1 + \cdots + \mu_m$.•

Ορισμός 2. Ο σχεδιασμός d^* με πίνακα πληροφορίας $\mathbf{Q}^*$ και ιδιοτιμές $\lambda_1,\cdots,\lambda_k$ είναι φ-βέλτιστος στην κλάση R των σχεδιασμών, αν οι ιδιοτιμές του $\mathbf{Q}^*$ κυριαρχούνται από τις ιδιοτιμές του πίνακα πληροφορίας $\mathbf{Q}$ για κάθε $d \in R$.•

Αυτό είναι ισοδύναμο με την ελαχιστοποίηση του αθροίσματος $\varphi(\lambda_1) + \cdots + \varphi(\lambda_k)$ για όλες τις συνεχείς, φθίνουσες κυρτές συναρτήσεις φ(.) (Marshall and Olkin (1979), p. 11), Pukelsheim (1993).

2. ΕΚΤΙΜΗΣΗ ΟΛΩΝ ΤΩΝ ΠΑΡΑΜΕΤΡΩΝ

Υπάρχει μια εκτεταμένη βιβλιογραφία για την εκτίμηση του βέλτιστου σχεδιασμού για την εκτίμηση των k+1 παραμέτρων $\mu^*, v_1,\cdots,v_k$ του μοντέλου (3).

Αυτό επιμερίζεται στις περιπτώσεις,
$n \equiv 0 \bmod 4, n \equiv 1 \bmod 4, n \equiv 2 \bmod 4, n \equiv 3 \bmod 4$

(•) Αν $n \equiv 0 \bmod 4$ and $k+1 \le n$, ο φ-βέλτιστος σχεδιασμός είναι να πάρουμε k+1 στήλες από έναν $n \times n$ πίνακα Hadamard, ο πίνακας πληροφορίας είναι $\mathbf{Q}_0 = n\,\mathbf{I}_{k+1}$.

(•) Αν $n \equiv 1 \bmod 4$ και $k+1 \le n-1$, παίρνουμε k+1 στήλες από έναν $n \times n$ πίνακα Hadamard και επισυνάπτουμε μια γραμμή $\underbrace{1\cdots1}_{k+1}$, αυτός ο σχεδιασμός είναι E,A,D βέλτιστος και ο πίνακας πληροφορίας είναι $\mathbf{Q}_1 = (n-1)\,\mathbf{I}_{k+1} + \mathbf{J}_{k+1}$.

Αν $n = k+1$ έχουμε κορεσμένο σχεδιασμό. Για n=5,13,25,41 υπάρχουν E,A,D βέλτιστοι σχεδιασμοί αλλά για n =17, 21, 29,…,ο D βέλτιστος σχεδιασμός απαιτεί μια ιδιαίτερη κατασκευή. Στις περιπτώσεις n=17,21 ο D βέλτιστος σχεδιασμός έχει κατασκευαστεί, δες Galil and Kiefer (1982), Kounias and Moyssiadis (1984), Chadjipantelis, Moyssiadis and Kounias (1987), Ehlich (1964a).

(•) Αν $n \equiv 2 \bmod 4$ και $k+1 \le n-2,$ παίρνουμε k+1 στήλες από έναν $n \times n$ πίνακα Hadamard και επισυνάπτουμε τις δύο γραμμές,

$$\underbrace{1\cdots 1}_{k+1}, \quad \underbrace{-1\cdots -1}_{r}\underbrace{1\cdots 1}_{s}, \; r+s=k+1, \; |r-s| \le 1.$$

Αν $k+1=n$ ο σχεδιασμός είναι κορεσμένος και ο D βέλτιστος σχεδιασμός κατασκευάζεται χρησιμοποιώντας κυκλικούς πίνακες **A, B** τάξης n/2 , όπως

$$d = \begin{bmatrix} \mathbf{A} & -\mathbf{B}^{\mathrm{T}} \\ \mathbf{B} & \mathbf{A}^{\mathrm{T}} \end{bmatrix} \Rightarrow \mathbf{Q} = \begin{bmatrix} \mathbf{A}\mathbf{A}^{T} + \mathbf{B}^{\mathrm{T}}\mathbf{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}\mathbf{B}^{\mathrm{T}} + \mathbf{A}^{\mathrm{T}}\mathbf{A} \end{bmatrix} = \begin{bmatrix} (n-2)\mathbf{I}_{n/2} + 2\mathbf{J}_{n/2} & \mathbf{0} \\ \mathbf{0} & (n-2)\mathbf{I}_{n/2} + 2\mathbf{J}_{n/2} \end{bmatrix}$$

όπου $\mathbf{A}^{T}$ ή $\mathbf{A}'$ συμβολίζει τον ανάστροφο του **A.** Αυτή η κατασκευή είναι δυνατή μόνο αν το $n-1$ είναι άθροισμα δύο τετραγώνων Yang (1976), όπως όταν n=6,10, 14,18,26,…, Kounias, Koukouvinos, Nikolaou, Kakos (1994), αλλά όχι όταν n=22,34,58,…, στις περιπτώσεις αυτές δεν είναι γνωστοί οι D βέλτιστοι κορεσμένοι σχεδιασμοί.

Αν $n \equiv 3 \bmod 4,\ k \le n-1$ ο πίνακας πληροφορίας με τη μεγαλύτερη ορίζουσα έχει s διαγώνιους πίνακες τάξης $[\frac{k}{s}]+1$ και $[\frac{k}{s}]$ με στοιχεία n στη διαγώνιο και 3 εκτός διαγωνίου. Εκτός των διαγωνίων πινάκων κάθε στοιχείο είναι ίσο με -1, Ehlich (1964b). Για την κατασκευή αυτών των σχεδιασμών δες Galil and Kiefer (1982), Kounias and Chadjipantelis (1983), Kounias and Farmakis (1984).

3. ΕΚΤΙΜΗΣΗ ΤΩΝ k ΑΝΤΙΘΕΣΕΩΝ $v_i = q_i - l_i \quad i = 1,...,k$

Αν μας ενδιαφέρει να εκτιμήσουμε τις αντιθέσεις $v_1/2,...,v_k/2$ αλλά όχι το μέσο μ^*, τότε ο πίνακας πληροφορίας, με τη μέθοδο ελαχίστων τετραγώνων, είναι,

$$\mathbf{Q} = \mathbf{X}_1'\mathbf{X}_1 - \frac{1}{n}(\mathbf{X}_1'\mathbf{X}_2)(\mathbf{X}_1'\mathbf{X}_2)' \qquad (4)$$

όπου $\mathbf{X}_1 = (\mathbf{w}_1,\cdots,\mathbf{w}_k),\ \mathbf{X}_2 = \mathbf{1}_n,\ \mathbf{w}_i = (w_{i1},\cdots,w_{in})',\ \ \mathbf{X}_2'\mathbf{X}_2 = n,$

$$w_{is} = \begin{cases} -1 & \text{αν } d(i,s)=1 \\ 1 & \text{αν } d(i,s)=2 \end{cases}$$

$$\mathbf{X}_1'\mathbf{X}_1 = \begin{bmatrix} \mathbf{w}_1' \\ \vdots \\ \vdots \\ \mathbf{w}_k' \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 & \cdots & \cdots & \mathbf{w}_k \end{bmatrix} = \begin{bmatrix} n & w_1'w_2 & \cdots & w_1'w_k \\ w_2'w_1 & \ddots & & \\ & & \ddots & w_{k-1}'w_k \\ w_k'w_1 & & w_k'w_{k-1} & n \end{bmatrix}$$

$$\mathbf{X}_1'\mathbf{X}_2 = \mathbf{X}_1'\mathbf{1}_n = \begin{bmatrix} w_1' \\ \vdots \\ w_k' \end{bmatrix} \mathbf{1}_n = \begin{bmatrix} n_1^2 - n_1^1 \\ \vdots \\ n_k^2 - n_k^1 \end{bmatrix}, \quad w_i'w_j = (n_{ij}^{11} - n_{ij}^{12} - n_{ij}^{21} + n_{ij}^{22})$$

- $n_i^1 =$ Πλήθος εμφανίσεων του παράγοντα i στη χαμηλή στάθμη
- $n_i^2 =$ Πλήθος εμφανίσεων του παράγοντα i στην υψηλή στάθμη
- $n_{ij}^{11} =$ Πλήθος εμφανίσεων των παραγόντων i,j και οι δύο στη χαμηλή
- στάθμη, στη ίδια παρατήρηση.
- $n_{ij}^{22} =$ Πλήθος εμφανίσεων των παραγόντων i,j και οι δύο στην υψηλή
- στάθμη, στη ίδια παρατήρηση.
- $n_{ij}^{12} =$ Πλήθος εμφανίσεων των παραγόντων i,j ο πρώτος στη χαμηλή στάθμη και ο δεύτερος στην υψηλή, στην ίδια παρατήρηση.
-

3.1. **Βέλτιστοι σχεδιασμοί για τις k αντιθέσεις**

Λήμμα 1. Αν $\mathbf{X} : n \times k$ είναι ένας σχεδιασμός με πίνακα πληροφορίας **Q,** ο σχεδιασμός $\mathbf{X}^* = \mathbf{GPX}$ έχει τον ίδιο πίνακα πληροφορίας, όπου $\mathbf{G}$ είναι $k \times k$ διαγώνιος πίνακας με στοιχεία ± 1 και $\mathbf{P}$ είναι $k \times k$ μεταθετικός πίνακας.

Απόδειξη. $\mathbf{M}^* = (\mathbf{X}^*)'(\mathbf{X}^*) = \mathbf{X}'\mathbf{P}'\mathbf{G}'\mathbf{GPX} = \mathbf{X}'\mathbf{X} = \mathbf{M}, \quad \mathbf{M} : k \times k$.•

Λήμμα 2. Αν $n \equiv 0 \bmod 4$ και μια στήλη του $n \times n$ πίνακα Hadamard $\mathbf{H}$ είναι $\mathbf{1}_n$ και πάρουμε $k \le n-1$ άλλες στήλες $\mathbf{X}_1$ του $\mathbf{H}$, τότε ο σχεδιασμός $\mathbf{X}_1$ είναι φ βέλτιστος για την εκτίμηση των αντιθέσεων $v_1, \cdots, v_k$.

Απόδειξη. $\mathbf{Q} = \mathbf{X}_1'\mathbf{X}_1 - \frac{1}{n}\mathbf{X}_1'\mathbf{J}_n\mathbf{X}_1 \le \mathbf{X}_1'\mathbf{X}_1 = n\mathbf{I}_k$, με ιδιοτιμές $\lambda_1 = \cdots = \lambda_k = n$ οι οποίες κυριαρχούνται από τις k στήλες ενός άλλου $k \times k$ πίνακα πληροφορίας.•

Θεώρημα 1. Αν $n \equiv 1 \bmod 4$ και ο σχεδιασμός $\mathbf{X} = (\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_k)$ με $\mathbf{x}_0 = \mathbf{1}_n$ είναι D βέλτιστος για την εκτίμηση των $\mu^*, v_1, \cdots, v_k$, τότε ο σχεδιασμός $\mathbf{X}^* = (\mathbf{x}_1, \cdots, \mathbf{x}_k)$ είναι D βέλτιστος για την εκτίμηση των k αντιθέσεων $v_1, \cdots, v_k$.

Απόδειξη. Ο πίνακας πληροφορίας του $n \times (k+1)$ σχεδιασμού $\mathbf{X} = (\mathbf{1}_n, \mathbf{X}^*)$ είναι

$$\mathbf{M}_{k+1} = \begin{bmatrix} n & \mathbf{b}' \\ \mathbf{b} & \mathbf{M}_k \end{bmatrix}, \mathbf{b}' = \mathbf{1}_n' \mathbf{X}^*, \ \mathbf{M}_k = (\mathbf{X}^*)' \mathbf{X}^*, \text{ τότε}$$

$$\det(\mathbf{M}_{k+1)}) = n \det(\mathbf{M}_k - \frac{1}{n} \mathbf{b}\mathbf{b}') = n \det(\mathbf{M}_k - \frac{1}{n} (\mathbf{X}^*)' \mathbf{J}_n (\mathbf{X}^*).$$

Επομένως αν η $\det(\mathbf{M}_{k+1)})$ μεγιστοποιείται, το ίδιο συμβαίνει και με την $\det\left(\mathbf{M}_k - \frac{1}{n} (\mathbf{X}^*)' \mathbf{J}_n (\mathbf{X}^*) \right)$.•

Για n=5,9,13,17,21,25,41 ο D βέλτιστος κορεσμένος σχεδιασμός για την εκτίμηση των $\mu^*, v_1, \cdots, v_k$, είναι γνωστός, οπότε είναι γνωστός και ο σχεδιασμός $\mathbf{X}^* = (\mathbf{x}_1, \cdots, \mathbf{x}_k)$ για την εκτίμηση των $v_1, \cdots, v_k$.•

Το Θεώρημα 1 ισχύει και στις περιπτώσεις $n \equiv 2 \bmod 4$ και $3 \bmod 4$, στις περιπτώσεις αυτές αν $\mathbf{X}$ είναι ένας βέλτιστος σχεδιασμός για την εκτίμηση των $\mu^*, v_1, \cdots, v_k$, τότε ο σχεδιασμός $\mathbf{X}_1 = \mathbf{GPX}$ έχει μια στήλη ίση με $\mathbf{1}_n$. Παίρνουμε τις υπόλοιπες k στήλες του $\mathbf{X}_1$ και αυτές είναι ένας σχεδιασμός $\mathbf{X}^*$,

που είναι D βέλτιστος για την εκτίμηση των αντιθέσεων $v_1, \cdots, v_k$. $\mathbf{G}$ είναι διαγώνιος πίνακας με στοιχεία ± 1 και $\mathbf{P}$ ένας $n \times n$ μεταθετικός πίνακας.

Για ένα κατάλογο E,A,D βέλτιστων σχεδιασμών δες Kounias and Chadjipantelis (1985), Kounias, Koukouvinos, Nikolaou and Kakos (1994), Moyssiadis and Kounias (1983).

3.2 Εφαρμογή

Παράδειγμα 2: Πέντε παράγοντες, έξι παρατηρήσεις **.**

Παίρνουμε τους δύο κυκλικούς πίνακες $\mathbf{A} = [-1\ 1\ 1], \mathbf{B} = (1\ 1\ 1)$, τότε ο σχεδιασμός $\mathbf{X} = \begin{bmatrix} \mathbf{A} & -\mathbf{B}^T \\ \mathbf{B} & \mathbf{A}^T \end{bmatrix}$ είναι D βέλτιστος για την εκτίμηση των $\mu^*, v_1, \cdots, v_5$ Yang (1976).

$$\mathbf{X}=\begin{bmatrix} - & 1 & 1 & - & - & - \\ 1 & - & 1 & - & - & - \\ 1 & 1 & - & - & - & - \\ 1 & 1 & 1 & - & 1 & 1 \\ 1 & 1 & 1 & 1 & - & 1 \\ 1 & 1 & 1 & 1 & 1 & - \end{bmatrix} \Rightarrow \mathbf{X}_1=\begin{bmatrix} 1 & - & - & 1 & 1 & 1 \\ 1 & - & 1 & - & - & - \\ 1 & 1 & - & - & - & - \\ 1 & 1 & 1 & - & 1 & 1 \\ 1 & 1 & 1 & 1 & - & 1 \\ 1 & 1 & 1 & 1 & 1 & - \end{bmatrix} \Rightarrow \mathbf{X}'\mathbf{X}=\mathbf{Q}=\begin{bmatrix} 6 & 2 & 2 & 0 & 0 & 0 \\ 2 & 6 & 2 & 0 & 0 & 0 \\ 2 & 2 & 6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 6 & 2 & 2 \\ 0 & 0 & 0 & 2 & 6 & 2 \\ 0 & 0 & 0 & 2 & 2 & 6 \end{bmatrix}$$

$$\mathbf{X}^*=\begin{bmatrix} - & - & 1 & 1 & 1 \\ - & 1 & - & - & - \\ 1 & - & - & - & - \\ 1 & 1 & - & 1 & 1 \\ 1 & 1 & 1 & - & 1 \\ 1 & 1 & 1 & 1 & - \end{bmatrix},\ \mathbf{Q}^*=(\mathbf{X}^*)'(\mathbf{X}^*)-\frac{1}{6}(\mathbf{X}^*)'\mathbf{1}_6\mathbf{1}_6'\mathbf{X}^*=\begin{bmatrix} 6 & 2 & 0 & 0 & 0 \\ 2 & 6 & 0 & 0 & 0 \\ 0 & 0 & 6 & 2 & 2 \\ 0 & 0 & 2 & 6 & 2 \\ 0 & 0 & 2 & 2 & 6 \end{bmatrix}-\frac{1}{6}\begin{bmatrix} 2 \\ 2 \\ 0 \\ 0 \\ 0 \end{bmatrix}\begin{bmatrix} 2 & 2 & 0 & 0 & 0 \end{bmatrix}$$

όπου $-$ συμβολίζει το -1, ο πίνακας διασποράς είναι $\mathbf{V}=\sigma^2\mathbf{Q}^{-1}$ και D βέλτιστος σχεδιασμός έχει $\det(\mathbf{Q})=\dfrac{14\times 18\times 16^2\times 22}{3^5}=5840.59$.

Αν παίρναμε ως σχεδιασμό τις 5 τελευταίες στήλες από τον πίνακα **X,** τότε ο $\mathbf{X}_1$ και ο αντίστοιχος πίνακας πληροφορίας $\mathbf{Q}_1$ θα είναι,

$$\mathbf{X}_1=\begin{bmatrix} 1 & 1 & - & - & - \\ - & 1 & - & - & - \\ 1 & - & - & - & - \\ 1 & 1 & - & 1 & 1 \\ 1 & 1 & 1 & - & 1 \\ 1 & 1 & 1 & 1 & - \end{bmatrix},\mathbf{Q}_1=\mathbf{X}_1'\mathbf{X}_1-\frac{1}{6}(\mathbf{X}_1'1_6)(\mathbf{X}_1'1_6)'=\begin{bmatrix} 6 & 2 & 0 & 0 & 0 \\ 2 & 6 & 0 & 0 & 0 \\ 0 & 0 & 6 & 2 & 2 \\ 0 & 0 & 2 & 6 & 2 \\ 0 & 0 & 2 & 2 & 6 \end{bmatrix}-\frac{4}{6}\begin{bmatrix} 2 \\ 2 \\ - \\ - \\ - \end{bmatrix}\begin{bmatrix} 2 & 2 & - & - & - \end{bmatrix},$$

$$\det(\mathbf{Q}_1)=\frac{1}{3^5}\det\begin{bmatrix} 16 & 4 & 4 \\ 4 & 16 & 4 \\ 4 & 4 & 16 \end{bmatrix}\det\begin{bmatrix} 8 & -4 \\ -4 & 8 \end{bmatrix}=\frac{4^5\times 2}{3}=682.66<5840.66$$

ABSTRACT

In factorial designs with k factors, each at two levels, the interest is in estimating the contrasts of factor level effects. The model and the information matrix of the contrasts is given, the observations are uncorrelated. There is a considerable amount of literature for finding the optimal designs in estimating the mean and the k contrasts, but not for estimating alone the contrasts. Optimal designs for estimating the contrasts of factor level effects are given when

$n \equiv 0,1,2,3 \bmod 4$ and a specific example is presented as an application of the preceding theory.

Ευχαριστίες: Ευχαριστούμε τον κριτή για τις εύστοχες παρατηρήσεις του.

ΑΝΑΦΟΡΕΣ

Chadjipantelis T., Kounias S. and Moyssiadis C. (1987). The maximum Determinant of 21x21 (+1,-1)-Matrices and D-Optimal Designs. *Journal of Statistical Planning and Inference* **16**, 167-178.

Ehlich H., (1964a). Determinantenabschätzungen für binäre Matrizen. *Math. Zeitschr.* **83**, 123-132.

Ehlich H., (1964b). Determinantenabschätzungen für binäre Matrizen mit n≡3mod4. *Math. Zeitschr.* **84**, 438-447.

Galil Z. and Kiefer J. (1982). Construction methods for D-optimum weighing designs when n≡3mod4. *Ann. Statist.* **10**, 502-510.

Kounias S. and Chadjipantelis T. (1983). Some D-optimal weighing designs when n≡3mod4. *Journal of Statistical Planning and Inference* **8**, 117-127.

Kounias S. and Farmakis N. (1984). A construction for D-optimal weighing designs when n≡3mod4. *Journal of Statistical Planning and Inference* **10**, 177-187.

Kounias S. and Chadjipantelis T. (1985). Supplementary Difference Sets and D-optimal Designs for n≡2mod4, *Discrete Math.* **57**, 211-216.

Kounias S., Koukouvinos C., Nikolaou N. and Kakos A. (1994). The Non-equivalent Circulant D-optimal Designs for n≡3mod4, n≤54, n=66. *Journal of Combinatorial Theory, Series A,* **65**, No.1, 26-38.

Pukelsheim, F. (1993). *Optimal Designs of Experiments*. New York: John Wiley and Sons,Inc

Marshall, A. W. and Olkin, I. (1979). *Inequalities: Theory of Majorization and Its Applications*. New York: Academic Press.

Moyssiadis C. and Kounias S. (1982). The exact D-optimal first order saturated design with 17 observations. *Journal of Statistical Planning and Inference* **7**, 13-27.

Yang C. H. (1976). Maximal Binary Matrices and Sum of Two Squares. *Math. of Comp.*, **30**, No.133, 148-153.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής (2015), σελ.326-337

2^k ΠΑΡΑΓΟΝΤΙΚΟΙ ΣΧΕΔΙΑΣΜΟΙ - ΚΑΤΑΣΚΕΥΗ ΤΟΥ ΚΟΡΕΣΜΕΝΟΥ ΒΕΛΤΙΣΤΟΥ ΣΧΕΔΙΑΣΜΟΥ ΜΕ 22 ΠΑΡΑΤΗΡΗΣΕΙΣ

Β. Χασιώτης[1], Ν. Φαρμάκης[1], Σ. Κουνιάς[2]
[1]Τμήμα Μαθηματικών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
chasiotisv@math.auth.gr, farmakis@math.auth.gr
[2]Τμήμα Μαθηματικών, Πανεπιστήμιο Αθηνών,
skounias@math.uoa.gr

ΠΕΡΙΛΗΨΗ

Στους 2^k παραγοντικούς σχεδιασμούς με k παράγοντες, ο καθένας σε δύο στάθμες, μας ενδιαφέρει η εκτίμηση του μέσου και των αντιθέσεων των επιδράσεων των παραγόντων. Υπάρχει εκτεταμένη βιβλιογραφία για την κατασκευή βέλτιστων σχεδιασμών, όταν το πλήθος των παρατηρήσεων είναι 0,1,2,3 *mod*4. Αναπτύσσεται ένας αλγόριθμος για την κατασκευή βέλτιστων σχεδιασμών με 22 παρατηρήσεις και 22 ή 21 παράγοντες. Βρίσκουμε το βέλτιστο σχεδιασμό για κάθε μια από τις δύο περιπτώσεις και δίνουμε τους αντίστοιχους πίνακες πληροφορίας. Δείχνουμε ότι οι πίνακες πληροφορίας με ορίζουσα μεγαλύτερη αυτής που δίνουμε αντιστοιχούν σε σχεδιασμούς που δεν υπάρχουν. Υπάρχουν δύο μη ισοδύναμοι D-βέλτιστοι σχεδιασμοί, οι οποίοι είναι και A-βέλτιστοι, όμως δεν έχουν την ίδια τιμή για το E-κριτήριο.

Λέξεις κλειδιά: φ-βελτιστοποίηση, αλγόριθμοι, πίνακας πληροφορίας, συνάρτηση πληροφορίας, πίνακας σχεδιασμού.

1. ΕΙΣΑΓΩΓΗ

Έχουμε k παράγοντες $F_1,...,F_k$ ο καθένας σε δύο στάθμες. Συμβολίζουμε τις στάθμες αυτές με 1 και 2, τις οποίες ονομάζουμε χαμηλή και υψηλή. Αν ο παράγοντας F_i $i=1,...,k$ συμμετέχει στο πείραμα με τη χαμηλή ή την υψηλή στάθμη οι επίδραση είναι αντίστοιχα l_i ή q_i. Σε κάθε εκτέλεση του πειράματος ο κάθε παράγοντας συμμετέχει με μία από τις δύο στάθμες. Για διανύσματα και πίνακες χρησιμοποιούμε έντονα γράμματα.

Ορισμός 1.1 Ο 2^k παραγοντικός σχεδιασμός με n παρατηρήσεις είναι ένας $n \times k$ σχηματισμός με στοιχεία 1 και 2, όπου οι γραμμές αντιστοιχούν στις n παρατηρήσεις και οι στήλες στους k παράγοντες.

Παράδειγμα 1.1 Δίνεται ο σχεδιασμός *d* με 5 παρατηρήσεις και 4 παράγοντες,

$$d:\left[\begin{array}{c|cccc} & F_1 & F_2 & F_3 & F_4 \\ \hline 1 & 1 & 2 & 2 & 1 \\ 2 & 1 & 1 & 2 & 1 \\ 3 & 2 & 2 & 1 & 2 \\ 4 & 1 & 1 & 2 & 2 \\ 5 & 2 & 1 & 1 & 2 \end{array}\right],$$

Στην τρίτη παρατήρηση οι 4 παράγοντες συμμετέχουν στις στάθμες 2,2,1,2 και αν y_3 είναι το αποτέλεσμα της παρατήρησης αυτής, αγνοώντας αλληλεπιδράσεις δύο και περισσότερων παραγόντων, το μοντέλο λέγεται κύριων επιδράσεων και είναι $y_3 = \mu + q_1 + q_2 + l_3 + q_4 + e_3$, όπου μ είναι ο μέσος και e_3 είναι το σφάλμα.

2. ΤΟ ΜΟΝΤΕΛΟ

Στη μελέτη αυτή οι παράμετροι που μας ενδιαφέρουν είναι: $\mu, (q_1 - l_1), ..., (q_k - l_k)$ και το μοντέλο είναι,

$$y_s = \mu + \sum_{i=1}^{k} \tau_{d(i,s)} + e_s,\ s = 1, ..., n,$$

όπου $\tau_{d(i,s)}$ συμβολίζει την επίδραση του παράγοντα F_i στην *s* παρατήρηση, όπως καθορίζεται στο σχεδιασμό *d*. Το μοντέλο αυτό γράφεται επίσης ως

$$y_s = \mu + \sum_{i=1}^{k} (l_i x_{i,s} + q_i z_{i,s}) + e_s, \qquad (2.1)$$

όπου l_i είναι η επίδραση του παράγοντα F_i στη στάθμη 1 και q_i η επίδρασή του στη στάθμη 2. Οι συμβολισμοί είναι

$$x_{i,s} = \begin{cases} 1 & \text{αν} \quad d(i,s) = 1 \\ 0 & \text{αν} \quad d(i,s) = 2 \end{cases}, \quad z_{i,s} = \begin{cases} 0 & \text{αν} \quad d(i,s) = 1 \\ 1 & \text{αν} \quad d(i,s) = 2 \end{cases}.$$

Σε διανυσματική μορφή το μοντέλο γράφεται,

$$\mathbf{y} = \mu \mathbf{1}_n + \sum_{i=1}^{k} \left(l_i \mathbf{x}_i + q_i \mathbf{z}_i \right) + \mathbf{e}, \qquad (2.2)$$

όπου $\mathbf{y} = (y_1, ..., y_n)'$, $\mathbf{x}_i = (x_{i1}, ..., x_{in})'$, $\mathbf{z}_i = (z_{i1}, ..., z_{in})'$.

Εδώ $\mathbf{y}, \mathbf{x}_i, \mathbf{z}_i, \mathbf{e}$, $i = 1,...,k$ είναι $n \times 1$ διανύσματα, με $\mathbf{x}_i + \mathbf{z}_i = \mathbf{1}_n$, $i = 1,2,...,k$, επομένως οι παράμετροι l_i, q_i, $i = 1,...,k$ δεν εκτιμούνται.

Θεωρούμε ότι τα σφάλματα είναι ασυσχέτιστες τυχαίες μεταβλητές με διασπορά σ^2.

Ενδιαφερόμαστε να εκτιμήσουμε τις αντιθέσεις $v_i = \dfrac{(q_i - l_i)}{2}$, $i = 1,2,...,k$, προς τούτο το μοντέλο (2.2) γράφεται

$$\mathbf{y} = \mu^* \mathbf{1}_n + \sum_{i=1}^{k} v_i \mathbf{w}_i + \mathbf{e}, \tag{2.3}$$

όπου $\mu^* = \mu + (\sum_{i=1}^{k} (l_i + q_i)/2$, $\mathbf{w}_i = (\mathbf{z}_i - \mathbf{x}_i)$, $\mathbf{w}_i = (w_{i1},...,w_{in})'$, $i = 1,2,...,k$.

Οπότε: $w_{i,s} = \begin{cases} -1 & \text{αν} \quad d(i,s) = 1 \\ 1 & \text{αν} \quad d(i,s) = 2 \end{cases}$, $i = 1,2,\ldots,k$, $j = 1,2,\ldots,n$.

Το μοντέλο (2.3) γράφεται επίσης και ως $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{e}$, όπου $\mathbf{X} = (\mathbf{1}_n, \mathbf{w}_1, ..., \mathbf{w}_k)$, $\boldsymbol{\theta} = (\mu^*, v_1, ..., v_k)'$. Για ασυσχέτιστες παρατηρήσεις ο πίνακας διασποράς των εκτιμημένων παραμέτρων $\hat{\boldsymbol{\theta}}$ είναι $\mathbf{V} = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$ και ο πίνακας πληροφορίας είναι $\mathbf{Q} = \mathbf{X}'\mathbf{X}$. Αυτοί οι σχεδιασμοί λέγονται επίσης και spring balance.

Υπάρχει ένα τεράστιο πλήθος εργασιών για τον προσδιορισμό βέλτιστων σχεδιασμών, όταν ισχύει το μοντέλο (2.3), όπου εξετάζονται οι περιπτώσεις $n \equiv 0 \bmod 4,\ 1 \bmod 4,\ 2 \bmod 4,\ 3 \bmod 4$.

3. Η ΕΝΝΟΙΑ ΤΗΣ ΚΥΡΙΑΡΧΙΑΣ

Ορισμός 3.1 Η $g(\mathbf{Q})$ είναι συνάρτηση πληροφορίας, αν

(*i*) $\mathbf{C} \geq \mathbf{D} \Rightarrow g(\mathbf{C}) \geq g(\mathbf{D})$, για κάθε $\mathbf{C}, \mathbf{D} \in nnd(k)$

(*ii*) $g((1-a)\mathbf{C} + a\mathbf{D}) \geq (1-a)g(\mathbf{C}) + ag(\mathbf{D})$, για κάθε $0 < a < 1$, και $\mathbf{C}, \mathbf{D} \geq 0$.

(*iii*) $g(d\mathbf{C}) = dg(\mathbf{C})$, για κάθε $d > 0$, $\mathbf{C} \geq 0 \Rightarrow g(\mathbf{0}) = 0 \Rightarrow g(\mathbf{C}) \geq 0$, και $\mathbf{C} \geq \mathbf{0}$

(*iv*) $g(\mathbf{C}) = g(\mathbf{PCP}')$ *για κάθε* $\mathbf{C} \in nnd(k)$ •

Ο $\mathbf{Q}$ είναι ένας $k \times k$ μη αρνητικά ορισμένος πίνακας (nnd(k)), το $\mathbf{C} \geq \mathbf{0}$ συμβολίζει ότι ο C είναι nnd(k) και ο $\mathbf{P}$ είναι ένας $k \times k$ μεταθετικός πίνακας (Pukelsheim (1993, pp. 114-119 and pp. 352-353).

Η συνθήκη *(iv)* σημαίνει ότι με τη μετάθεση ορισμένων παραμέτρων, ο πίνακας διασποράς περιέχει την ίδια πληροφορία του αρχικού πίνακα διασποράς.

Στον ορισμό 3.1 η συνάρτηση g είναι κοίλη και αύξουσα, διότι χρησιμοποιούμε τον πίνακα πληροφορίας $\mathbf{Q} = \mathbf{X'X}$.

Σημειώνουμε ότι αν ο πίνακας $\mathbf{Q} \in nnd(k)$, τότε από τις *(ii)* και *(iv)* προκύπτει ότι ο πίνακας $\bar{\mathbf{Q}} = (1-a)\mathbf{Q} + a\mathbf{PQP'},\ 0 < a < 1$ είναι «καλύτερος» από τον $\mathbf{Q}$, διότι $g(\bar{\mathbf{Q}}) \geq g(\mathbf{Q})$. Επαναλαμβανόμενη εφαρμογή της κυρτής αυτής σχέσης καταλήγει σε πίνακα πληροφορίας που έχει ίσα όλα τα διαγώνια στοιχεία και ίσα όλα τα μη διαγώνια στοιχεία.

Ορισμός 3.2 Ένας σχεδιασμός d^* με πίνακα πληροφορίας $\mathbf{Q}^*$ είναι καθολικά βέλτιστος, αν $g(\mathbf{Q}^*) \geq g(\mathbf{Q})$ για όλες τις συναρτήσεις πληροφορίας g(.), όπου $\mathbf{Q}$ είναι ο πίνακας πληροφορίας κάθε άλλου ανταγωνιστικού σχεδιασμού.•

Ορισμός 3.3 Ένας σχεδιασμός d^* με $n \times n$ πίνακα πληροφορίας $\mathbf{Q}^*$ και ιδιοτιμές $\lambda_1, \lambda_2, \ldots, \lambda_n$ είναι φ-βέλτιστος στην κλάση R των σχεδιασμών, αν οι ιδιοτιμές του $\mathbf{Q}^*$ κυριαρχούνται από τις ιδιοτιμές του πίνακα πληροφορίας $\mathbf{Q}_d$, $\forall d \in R$.•

Αυτό ισοδυναμεί με την ελαχιστοποίηση του $\varphi(\lambda_1) + \cdots + \varphi(\lambda_n)$ για όλες τις συνεχείς φθίνουσες και κυρτές συναρτήσεις $\varphi(\lambda)$ (Marshall and Olkin (1979, pp. 11)).

Σημειώνουμε ότι η *MV*-βελτιστοποίηση, δηλαδή η ελαχιστοποίηση του μέγιστου διαγώνιου στοιχείου του πίνακα διασποράς $\mathbf{V} = (v_{ij})$, δεν καλύπτεται από τη φ-βελτιστοποίηση, διότι τα στοιχεία v_{ii} δεν μπορούν να εκφραστούν ως συναρτήσεις των ιδιοτιμών του πίνακα πληροφορίας (Fr. Pukelsheim (1993).

Η προηγούμενη ανάλυση οφείλεται στον Fr. Pukelsheim (1993), όπου μπορούν να αντληθούν περισσότερες λεπτομέρειες.

Αν δεν υπάρχει φ-βέλτιστος σχεδιασμός, καταφεύγουμε σε συγκεκριμένα κριτήρια βελτιστοποίησης, όπως *E, A, D, MV, G*- βελτιστοποίηση.

Ορισμός 3.4 Αν *F* είναι μια κλάση θετικά ορισμένων πινάκων τάξης m (pd(m)), τότε

E-βελτιστοποίηση: $\underbrace{\max}_{Q \in F} \underbrace{\min}_{i} (\lambda_i(\mathbf{Q}))$, *A*-βελτιστοποίηση: $\underbrace{\min}_{Q \in F} \left(\frac{1}{\lambda_1(\mathbf{Q})} + \cdots + \frac{1}{\lambda_m(\mathbf{Q})} \right)$,

D-βελτιστοποίηση: $\underbrace{\max}_{Q \in F} \left(\lambda_1(\mathbf{Q}) \times \cdots \times \lambda_m(\mathbf{Q}) \right)$, *MV*-βελτιστοποίηση: $\underbrace{\min}_{Q \in F} \underbrace{\max}_{i} (v_{ii})$,

όπου v_{ii} είναι το *i*-στο διαγώνιο στοιχείο του πίνακα διασποράς $\mathbf{V}$.•

Οι προηγούμενοι ορισμοί έχουν στατιστική ερμηνεία.
Ο σχηματισμός OA(n,r,2,2) έχει n γραμμές, r στήλες, δύο σύμβολα 0 και 1 και σε κάθε δύο στήλες το πλήθος των (00), (01), (10), (11) είναι ίσο με $n/4$. Λέγονται ορθογώνιοι σχηματισμοί (orthogonal arrays).

4. ΒΕΛΤΙΣΟΙ ΣΧΕΔΙΑΣΜΟΙ

Λήμμα 4.1 (i) Αν $n \equiv 0 \bmod 4$ και το πλήθος των παραγόντων είναι $k \le n-1$, ο OA(n,$k+1$,2,2) είναι φ-βέλτιστος για την εκτίμηση των $k+1$ παραμέτρων.

(ii) Αν $n \equiv 1 \bmod 4$, ο φ-βέλτιστος σχεδιασμός έχει πίνακα πληροφορίας με διαγώνια στοιχεία ίσα με n και όλα τα μη διαγώνια στοιχεία ίσα με 1, $\mathbf{Q} = (n-1)\mathbf{I}_{k+1} + \mathbf{J}_{k+1}$.

(*iii*) Αν $n \equiv 2 \bmod 4$, ο φ-βέλτιστος σχεδιασμός έχει πίνακα πληροφορίας της μορφής $\mathbf{Q} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} \end{bmatrix}$. Ο πίνακας $\mathbf{A} : (n/2) \times (n/2)$ έχει τα διαγώνια στοιχεία ίσα με n και τα μη διαγώνια στοιχεία ίσα με *2*.

(*iv*) Αν $n \equiv 3 \bmod 4$, ο D-βέλτιστος σχεδιασμός έχει στη διαγώνιο υποπίνακες διαστάσεων που διαφέρουν το πολύ κατά 1 με τα διαγώνια στοιχεία ίσα με n και τα μη διαγώνια στοιχεία ίσα με *3*. Τα στοιχεία εκτός των υποπινάκων είναι ίσα με *-1*.
Σχετικές εργασίες είναι: Kiefer (1961, 1975), Ehlich (1964 a, b) , Wojtas (1964). Για την κατασκευή των αντίστοιχων βέλτιστων σχεδιασμών δες: Galil, Kiefer (1982), Kounias, Chadjipantelis (1983), Kounias, Farmakis (1984, 1987). Για $n \equiv 2 \bmod 4$ δες Yang (1976), Kounias, Koukouvinos, Nikolaou and Kakos (1994). Για $n \equiv 1 \bmod 4$ δες Moyssiadis, Kounias (1982,1983), Chadjipantelis, Kounias, Moyssiadis (1987).

Λήμμα 4.2 (Marshall and Olkin p. 225) Αν $\mathbf{H}$ και $\bar{\mathbf{H}}$ είναι συμμετρικοί πίνακες της μορφής $\mathbf{H} = \begin{bmatrix} \mathbf{H_{11}} & \mathbf{H_{12}} \\ \mathbf{H_{21}} & \mathbf{H_{22}} \end{bmatrix}$, $\bar{\mathbf{H}} = \begin{bmatrix} \mathbf{H_{11}} & \mathbf{0} \\ \mathbf{0} & \mathbf{H_{22}} \end{bmatrix}$, όπου $\mathbf{H_{11}} : r \times r$, $\mathbf{H_{22}} : m \times m$ και $r + m = n$, τότε $\lambda(\bar{\mathbf{H}}) \prec \lambda(\mathbf{H})$. Αυτό συμβολίζει ότι οι ιδιοτιμές του $\bar{\mathbf{H}}$ κυριαρχούνται από τις ιδιοτιμές του $\mathbf{H}$.•

5. Ο D -ΒΕΛΤΙΣΤΟΣ ΚΟΡΕΣΜΕΝΟΣ ΣΧΕΔΙΣΜΌΣ ΓΙΑ n=22

(*i*) Αν $n \equiv 2 \bmod 4$ και το πλήθος των παραγόντων είναι $k \le n-3$, ο φ-βέλτιστος σχεδιασμός για τις $k+1$ παραμέτρους κατασκευάζεται επισυνάπτοντας σε έναν OA(n-*2*,k+*1*, 2,2) τις δύο γραμμές $\underbrace{1 \cdots 1}_{k+1}, \underbrace{-1 \cdots -1}_{r} \underbrace{1 \cdots 1}_{s}$, $r + s = k + 1$ και $|r - s| \le 1$.

(ii) Αν $n \equiv 2 \bmod 4$ και $k = n-1$, τότε χρησιμοποιούνται δύο κυκλικοί πίνακες **A**, **B** τάξης *n/2* για την κατασκευή του *φ*-βέλτιστου σχεδιασμού $d = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ -\mathbf{B}^{\mathrm{T}} & \mathbf{A}^{\mathrm{T}} \end{bmatrix}$.

Μια αναγκαία αλλά όχι ικανή συνθήκη για την ύπαρξη αυτών των κυκλικών πινάκων είναι ότι το *n-1* είναι άθροισμα τετραγώνων δύο περιττών αριθμών, δες Yang (1976), Chadjipantelis and Kounias (1985). Για *n*=6, 10, 14, 18, 26, 30, 38, 42, 46, 50 αυτοί οι πίνακες υπάρχουν, αλλά όχι για *n*=22,34,54,…, δες Kounias, Koukouvinos, Nikolaou and Kakos (1994).

Επομένως για τους κορεσμένους σχεδιασμούς *n*=22, *k*=21 ή 20 απαιτείται ειδική κατασκευή.

Λήμμα 5.1 Αν $\mathbf{w}_i, \mathbf{w}_j, \mathbf{w}_t$ είναι διανύσματα $n \times 1$ με στοιχεία ± 1, τότε

(*i*) $\mathbf{w}_i'\mathbf{w}_j \equiv n \bmod 4$ ή $(n+2) \bmod 4$.

(*ii*) $\mathbf{w}_1'\mathbf{w}_2 + \mathbf{w}_1'\mathbf{w}_3 - \mathbf{w}_2'\mathbf{w}_3 \equiv n \bmod 4$.

(*iii*) $\mathbf{w}_1'\mathbf{w}_2 \equiv c \bmod 4, \mathbf{w}_1'\mathbf{w}_3 \equiv c \bmod 4 \Rightarrow \mathbf{w}_2'\mathbf{w}_3 \equiv c \bmod 4,\ c = 0, 2$.

Η απόδειξη παραλείπεται.

Ορισμός 5.1 Αν $\mathbf{M}_1, \mathbf{M}_2 : n \times n$ και ισχύει $\mathbf{M}_1 = \mathbf{PDM}_2\mathbf{D'P'}$, όπου **P** είναι $n \times n$ μεταθετικός πίνακας και **D** είναι ένας $n \times n$ διαγώνιος πίνακας με στοιχεία ± 1, τότε οι πίνακες $\mathbf{M}_1, \mathbf{M}_2$ λέγονται ισοδύναμοι.•

Αυτός ο ορισμός μας λέει ότι αν σε ένα σχεδιασμό $X : n \times n$ με στοιχεία ± 1 αλλάξουμε γραμμές ή και πολλαπλασιάσουμε γραμμές με -1, δηλαδή $\mathbf{X}_1 = \mathbf{DPX}$, ο σχεδιασμός που προκύπτει είναι ισοδύναμος με τον αρχικό σχεδιασμό. Οι αντίστοιχοι πίνακες πληροφορίας είναι $\mathbf{X'X} = \mathbf{M}_2$, $\mathbf{XX'} = \mathbf{M}_1 = \mathbf{P'DM}_2\mathbf{DP}$. Οι δύο αυτοί σχεδιασμοί έχουν την ίδια ορίζουσα.

Θεώρημα 5.1 Αν $\mathbf{X} : n \times n$, $\mathbf{X'X} = \mathbf{M}_2$, $\mathbf{XX'} = \mathbf{M}_1$ και $\mathbf{M}_1 = \mathbf{P'DM}_2\mathbf{DP}$, τότε υπάρχει σχεδιασμός $\mathbf{R} : n \times n$ που ικανοποιεί τη σχέση $\mathbf{RR'} = \mathbf{R'R} = \mathbf{M}_2$.

Απόδειξη Παίρνουμε $\mathbf{R} = \mathbf{DPX} \Rightarrow \mathbf{R'R} = \mathbf{X'P'D'DPX} = \mathbf{X'X} = \mathbf{M}_2$ και

$\mathbf{RR'} = \mathbf{DPXX'P'D} = \mathbf{DPM}_1\mathbf{P'D} = \mathbf{DP(P'DM}_2\mathbf{DP)P'D} = \mathbf{M}_2$.•

Επομένως $\mathbf{RR'} = \mathbf{R'R} = \mathbf{M}_2$ και ο σχεδιασμός **R** επίσης ικανοποιεί την ίδια σχέση. Η σχέση αυτή χρησιμοποιείται για να δείξουμε ότι για ορισμένους πίνακες πληροφορίας $\mathbf{M} : n \times n$ δεν υπάρχουν σχεδιασμοί $\mathbf{X} : n \times n$ ώστε $\mathbf{X'X} = \mathbf{M}$.

Αν **X** είναι $m \times m$ πίνακας με στοιχεία ± 1 και $\mathbf{X'X} = \mathbf{M}$, τότε $\det(\mathbf{M}) = (\det(\mathbf{X}))^2$, που είναι το τετράγωνο ενός ακέραιου. Για *n=22* ο *φ*-βέλτιστος σχεδιασμός έχει πίνακα πληροφορίας της μορφής,

$$\mathbf{Q}=\begin{bmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{0} & \mathbf{B}\end{bmatrix},\ \mathbf{A}=\mathbf{B}=20\mathbf{I}_{11}+2\mathbf{J}_{11} \tag{5.1}$$

με $\det(\mathbf{Q})=(\det(\mathbf{A}))^2=(20^{10}\times 42)^2=184.9688\times 10^{27}$.

Θεώρημα 5.2 Δεν υπάρχει σχεδιασμός $\mathbf{X}: 22\times 22$ με πίνακα πληροφορίας $\mathbf{X'X}=\mathbf{Q}=\begin{bmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{0} & \mathbf{A}\end{bmatrix}$ και $\mathbf{A}=20\mathbf{I}_{11}+2\mathbf{J}_{11}$.

Απόδειξη Ο πίνακας πληροφορίας που δίνεται στην (6.1) έχει τη μεγαλύτερη ορίζουσα (Ehlich H (1964a, b). Αν υπάρχει σχεδιασμός $\mathbf{X}:\mathbf{X'X}=\mathbf{Q}$, τότε από το Θεώρημα 5.1 προκύπτει ότι υπάρχει σχεδιασμός $\mathbf{R}: 22\times 22$ που ικανοποιεί την $\mathbf{R}'\mathbf{R}=\mathbf{R}\mathbf{R}'=\mathbf{Q}=\begin{bmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{0} & \mathbf{B}\end{bmatrix},\ \mathbf{A}=\mathbf{B}=\mathbf{20I_{11}}+\mathbf{2J_{11}}$ με $\det(\mathbf{Q})=(\det(\mathbf{A}))^2$.

$\mathbf{R}=(\mathbf{R}_1,\mathbf{R}_2),\mathbf{R}_1,\mathbf{R}_2: 22\times 11,\ \mathbf{R}_1=(\mathbf{X}_1,...,\mathbf{X}_{11}),\mathbf{R}_2=(\mathbf{X}_{12},...,\mathbf{X}_{22}),$

$\mathbf{X}_i: 22\times 1,\mathbf{X}_i'\mathbf{X}_i=22,\mathbf{X}_i'\mathbf{X}_j=2,\ i\neq j,i,j\leq 11,\ \mathbf{X}_i'\mathbf{X}_j=0\ ,i=1,...,11,\ j=12,...,22$.

$\mathbf{R}\mathbf{R}'=\mathbf{R}_1\mathbf{R}_1'+\mathbf{R}_2\mathbf{R}_2',\ \mathbf{R}'\mathbf{R}=\begin{bmatrix}\mathbf{R}_1'\mathbf{R}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{R}_2'\mathbf{R}_2\end{bmatrix},\ \mathbf{R}_1'\mathbf{R}_1=\mathbf{R}_2'\mathbf{R}_2=\mathbf{A}=\mathbf{20I_{11}}+\mathbf{2J_{11}}$.

Παίρνουμε την πρώτη στήλη $\mathbf{X}_1$, τότε

$\mathbf{X}_1'\mathbf{R}\mathbf{R}'\mathbf{X}_1=\mathbf{X}_1'\mathbf{R}_1\mathbf{R}_1'\mathbf{X}_1+\mathbf{X}_1'\mathbf{R}_2\mathbf{R}_2'\mathbf{X}_1=(\mathbf{X}_1'\mathbf{X}_1)^2+\cdots(\mathbf{X}_1'\mathbf{X}_{11})^2=22^2+4\times 10=524$.

Αν $\mathbf{X}_1=\begin{bmatrix}\mathbf{X}_{1,1}\\ \mathbf{X}_{1,2}\end{bmatrix},\ \mathbf{X}_{1,1},\mathbf{X}_{1,2}: 11\times 1,$ τότε

$\mathbf{X}_1'\mathbf{R}'\mathbf{R}\mathbf{X}_1=\mathbf{X}_{1,1}'\mathbf{R}_1'\mathbf{R}_1\mathbf{X}_{1,1}+\mathbf{X}_{1,2}'\mathbf{R}_2'\mathbf{R}_2\mathbf{X}_{1,2}\Rightarrow$

$\mathbf{X}_1'\mathbf{R}'\mathbf{R}\mathbf{X}_1=\mathbf{X}_{1,1}'\mathbf{A}\mathbf{X}_{1,1}+\mathbf{X}_{1,2}'\mathbf{A}\mathbf{X}_{1,2}=20\times 22+2(X_{1,1}'1_{11})^2+2(X_{1,2}'1_{11})^2$.

Οπότε, $524=20\times 22+2(\mathbf{X}_{1,1}'1_{11})^2+2(\mathbf{X}_{1,2}'1_{11})^2\Rightarrow$

$$42=(\mathbf{X}_{1,1}'1_{11})^2+(\mathbf{X}_{1,2}'1_{11})^2. \tag{5.2}$$

Όμως το 42 δεν είναι άθροισμα τετραγώνων δύο περιττών αριθμών, επομένως ο σχεδιασμός $\mathbf{X}: 22\times 22$ με $(\det(\mathbf{X}))^2=(20^{10}\times 42)^2=184.9688\times 10^{27}$ δεν υπάρχει.•

Στο Παράρτημα, δίνεται ο σχεδιασμός $\mathbf{X}^*: 22\times 22$ με $\det(\mathbf{X}^*)=4.096\times 10^{14}$ με πίνακα πληροφορίας,

$$\mathbf{Q}=\begin{bmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{0} & \mathbf{A}\end{bmatrix},\ \mathbf{A}=\begin{bmatrix}22 & \mathbf{b'}\\ \mathbf{b} & \mathbf{C}\end{bmatrix},\quad \mathbf{b'}=\begin{bmatrix}-2\times\mathbf{1}_5' & 2\times\mathbf{1}_5'\end{bmatrix},\mathbf{C}=20\mathbf{I}_{10}+2\mathbf{J}_{10}. \quad (5.3)$$

Υπάρχουν πίνακες $\mathbf{Q}\in nnd(22)$ με $167.77216\times10^{27}<\det(\mathbf{Q})<184.968\times10^{27}$. Σε όλες αυτές τις περιπτώσεις δείχνουμε, όπως στο Θεώρημα 5.2, ότι δεν υπάρχουν σχεδιασμοί $\mathbf{X}:22\times22$ που να ικανοποιούν τη σχέση $\left(\det(\mathbf{X})\right)^2=\det(\mathbf{Q})$. Επομένως, οι δύο σχεδιασμοί $\mathbf{X}^*$, $\mathbf{X}^{**}:22\times22$ που δίνονται στο Παράρτημα, έχουν τη μέγιστη ορίζουσα. Ο περιορισμός των 15 σελίδων δεν μας επιτρέπει να παραθέσουμε αποδείξεις για όλες τις περιπτώσεις που $167.77216\times10^{27}<\det(\mathbf{Q})$.

6. ΑΛΓΟΡΙΘΜΟΣ ΓΙΑ ΤΗΝ ΚΑΤΑΣΚΕΥΗ ΤΟΥ ΚΟΡΕΣΜΕΝΟΥ ΣΧΕΔΙΑΣΜΟΥ ΜΕ n=22

Δίνεται ένας γνωστός αλγόριθμος στον οποίο ξεκινούμε από ένα αυθαίρετο 22×22 σχεδιασμό **X** με στοιχεία ±1 και βήμα-βήμα καταλήγουμε σε ένα σχεδιασμό, όπου η αλλαγή οποιουδήποτε στοιχείου δεν αυξάνει την ορίζουσά του. Αυτοί λέγονται τοπικά μέγιστοι σχεδιασμοί.

6.1 Αλγόριθμος

Βήμα 1: Ξεκινάμε από έναν αυθαίρετο 22×22 σχεδιασμό $\mathbf{X}$ με στοιχεία ±1.

Βήμα 2: Μεταβάλλουμε ένα στοιχείο s του $\mathbf{X}$ σε –s. Αν ο νέος σχεδιασμός $\mathbf{X}_s$ δίνει $\det(\mathbf{X}_s)>\det(\mathbf{X})$, τότε κρατάμε το σχεδιασμό $\mathbf{X}_s$. Αν $\det(\mathbf{X}_s)\le\det(\mathbf{X})$ κρατάμε το σχεδιασμό $\mathbf{X}$.

Βήμα 3: Επαναλαμβάνουμε το βήμα 2 μέχρις ότου η αλλαγή κάθε στοιχείου δεν αυξάνει την ορίζουσα του προηγούμενου σχεδιασμού. Τότε έχουμε ένα σχεδιασμό που δίνει τοπικό μέγιστο.

Για να αποφύγουμε πολλά τοπικά μέγιστα είναι προτιμότερο να ξεκινήσουμε με ένα «καλό» σχεδιασμό. Στην περίπτωσή μας αρχίσαμε με ένα OA(20,20,2,2), επισυνάπτοντας δύο γραμμές $\underbrace{+\,+\cdots+}_{20},\underbrace{+\,-\,+\cdots+}_{10}\underbrace{-\,-\cdots-}_{10}$ και δύο στήλες $\mathbf{y}_1,\mathbf{y}_2$, που έχουν με κάθε στήλη $\mathbf{x}$ του OA(20,20,2,2), τις επόμενες τιμές $\mathbf{x}'\mathbf{y}_i=0,\pm2,\pm4,\pm6,\ i=1,2$ και $\mathbf{y}_1'\mathbf{y}_2=0,\pm2$. Τις στήλες $\mathbf{y}_1,\mathbf{y}_2$ τις επιλέγουμε από τα $2^{20}=524288$ διανύσματα με στοιχεία ±1. Αυτή η διαδικασία μας έδωσε τους σχεδιασμούς $\mathbf{X}^*$, $\mathbf{X}^{**}$ που δίνονται στο παράρτημα.

6.2 The *D* -optimal design for *n*=22, *k+1*=21

Εάν διαγράψουμε την 1η ή την 12η στήλη από τον σχεδιασμό $\mathbf{X}^*$, προκύπτει ο *D*-βέλτιστος σχεδιασμός για την περίπτωση *n=22, k+1 =21*.

Εκτός από τον $\mathbf{X}^*$ υπάρχει και ο σχεδιασμός $\mathbf{X}^{**}$ με πίνακα πληροφορίας $\mathbf{M}^{**} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{bmatrix}$, όπου $\mathbf{A} = 20\mathbf{I}_{10} + 2\mathbf{J}_{10}$ και $\mathbf{B} = \begin{bmatrix} 22 & 2 & \mathbf{b}' \\ 2 & 22 & \mathbf{b}' \\ \mathbf{b} & \mathbf{b} & \mathbf{C} \end{bmatrix}$, με $\mathbf{b}' = \begin{bmatrix} -2 \times 1_5' & 2 \times 1_5' \end{bmatrix}, \mathbf{C} = 20\mathbf{I}_{10} + 2\mathbf{J}_{10}$.

Όμως $\min(\lambda(\mathbf{M}^{**})) = 12.83 < \min(\lambda(\mathbf{M}^{*})) = 14.59$, που σημαίνει ότι ο $\mathbf{X}^*$ είναι Ε- βέλτιστος συγκρινόμενος με τον $\mathbf{X}^{**}$.

ABSTRACT

In 2^k fractional factorials with *k* factors, each at two levels, the interest is in estimating the mean and the *k* contrasts of factor levels or in estimating some of the parameters of interest. For the construction of optimal designs a lot of papers have been published, when *n* is 0,1,2,3 mod4. An algorithm is developed for the construction of the optimal design with *n*=22 observations and 20 or 21 factors. We derive the optimal designs for the above mentioned two cases and present their information matrices. We also prove that information matrices with larger determinant correspond to designs that do not exist. There are two non equivalent *D* and *A*-optimal designs, which do not have the same value for the *E*-optimality criterion.

ΑΝΑΦΟΡΕΣ

Chadjipantelis T., Kounias S. and Moyssiadis C. (1987). The maximum Determinant of 21x21 (+1,-1)-Matrices and D-Optimal Designs. *Journal of Statistical Planning and Inference* **16**, 167-178.

Ehlich H., (1964a). Determinantenabschätzungen für binäre Matrizen. *Math. Zeitschr.* **83**, 123-132.

Ehlich H., (1964b). Determinantenabschätzungen für binäre Matrizen mit $n \equiv 3 \bmod 4$. *Math. Zeitschr.* **84**, 438-447.

Farmakis N. and Kounias S. (1987). Two new D-optimal designs (83,56,12), (83,55,12). *Journal of Statistical Planning and Inference* **15**, 247-257.

Farmakis N. and Kounias S. (1987b). The excess of Hadamard matrices and optimal designs. *Discrete Math.* **67**, 165-176.

Galil Z. and Kiefer J. (1982). Construction methods for D-optimum weighing designs when $n \equiv 3 \bmod 4$. *Ann. Statist.* **10**, 502-510.

Kiefer J. (1961). Optimum experimental designs V,with applications to systematic and rotatable designsrotatable. *In proceedings on the fourth Berkeley symposium in Mathematical Statistics and Probability*, **1**, 381-405.

Kiefer J. (1975). Construction and optimality of generalized Youden designs. In: Srivastava, J. N., ed. *A Survey of Statistical Design and Linear Models*, 333–353.

Kounias S. and Chadjipantelis T. (1983). Some D-optimal weighing designs when $n \equiv 3(\mathrm{mod}\, 4)$. *Journal of Statistical Planning and Inference* **8**, 117-127.

Kounias S. and Chalikias M. (2008). Estimability of parameters in a linear model and related characterizations. Statistics *and Probability letters 78, 2437-2439.*

Kounias S. and Farmakis N. (1984). A construction for D-optimal weighing designs when $n \equiv 3(\mathrm{mod}\, 4)$. *Journal of Statistical Planning and Inference* **10**, 177-187.

Kounias S. and Chadjipantelis T. (1985). Supplementary Difference Sets and D-optimal Designs for $n \equiv 2 \bmod 4$, *Discrete Math.* **57**, 211-216.

Kounias S., Koukouvinos C., Nikolaou N. and Kakos A. (1994). The Non-equivalent Circulant D-optimal Designs for $n \equiv 2 \bmod 4, n \le 54, n = 66$, *Journal of Combinatorial Theory, Series A,* **65**, No.1, 26-38.

Marshall, A. W. and Olkin, I. (1979). *Inequalities: Theory of Majorization and Its Applications*. New York: Academic Press.

Moyssiadis C. and Kounias S. (1982). The exact D-optimal first order saturated design with 17 observations. *Journal of Statistical Planning and Inference* **7**, 13-27.

Moyssiadis C. and Kounias S. (1983). Exact d-optimal n observations 2 designs of resolution III, when $n \equiv$ 1 or 2 mod 4, *Statistics* **14**, 367-379.

Pukelsheim, F. (1993). *Optimal Designs of Experiments*. New York: John Wiley and Sons,Inc.

Wojtas M. (1964). On Hadamard's inequality for determinants of order non-divisible by 4. *Colloq. Math.* **12**, 73-83.

Yang C. H. (1976). Maximal Binary Matrices and Sum of Two Squares. *Math. of Comp.*, **30**, No.133, 148-153.

ΠΑΡΑΡΤΗΜΑ

Ο *D, A , E* -βέλτιστος 22×22 σχεδιασμός $\mathbf{X}^*$

$$\left(\det(\mathbf{X}^*)\right)^2 = (4.096 \times 10^{14})^2 = 167.77216 \times 10^{27}$$

$$\mathbf{X}^* = \begin{bmatrix}
1 & 1 & 1 & 1 & -1 & 1 & 1 & 1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & -1 & 1 & -1 \\
1 & 1 & 1 & 1 & 1 & -1 & 1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 \\
1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 & -1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 \\
1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & -1 & 1 & 1 & 1 \\
1 & 1 & 1 & -1 & 1 & 1 & 1 & 1 & -1 & 1 & 1 & 1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 \\
1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 \\
1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 \\
1 & 1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & 1 & 1 & 1 & -1 & 1 & -1 & 1 \\
1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 \\
1 & 1 & -1 & -1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 \\
1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 & -1 \\
1 & -1 & 1 & 1 & 1 & 1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 \\
1 & -1 & 1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 & -1 & -1 & 1 & -1 & -1 & -1 & 1 & -1 & -1 \\
1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & 1 & 1 & 1 & 1 & -1 & 1 \\
1 & -1 & 1 & -1 & -1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 \\
1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 \\
1 & -1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & -1 & 1 & 1 & 1 & 1 & 1 & -1 & 1 & 1 \\
1 & -1 & -1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\
1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 \\
1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & -1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 \\
1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 \\
1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 & -1
\end{bmatrix}$$

Ο D και A αλλά όχι E-βέλτιστος 22×22 σχεδιασμός $\mathbf{X}^{**}$

$$\left(\det(\mathbf{X}^{**})\right)^2 = \left(\det(\mathbf{X}^{*})\right)^2$$

$$\mathbf{X}^{**} = \begin{bmatrix}
-1 & -1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & 1 & -1 & 1 \\
1 & 1 & 1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 \\
-1 & 1 & 1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 \\
-1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & -1 & 1 & 1 & 1 \\
1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 \\
-1 & 1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 \\
-1 & 1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 \\
1 & 1 & -1 & 1 & 1 & 1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & 1 & 1 & -1 & 1 & -1 & 1 \\
1 & 1 & -1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 \\
1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 \\
1 & -1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 & -1 \\
-1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 \\
1 & -1 & 1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 & -1 & -1 & -1 & 1 & -1 & -1 \\
1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & 1 & 1 & 1 & 1 & -1 & 1 \\
-1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 \\
-1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 \\
1 & -1 & -1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & 1 & 1 & 1 & -1 & 1 & 1 \\
-1 & -1 & -1 & -1 & -1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\
1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 \\
1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 \\
-1 & -1 & -1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 \\
-1 & -1 & -1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 & -1
\end{bmatrix}$$

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής (2015), σελ.338-352

ΜΕΛΕΤΗ ΤΗΣ ΔΟΜΗΣ ΚΑΙ ΕΛΕΓΧΟΣ ΤΥΧΑΙΟΤΗΤΑΣ ΣΕΙΣΜΙΚΩΝ ΔΙΚΤΥΩΝ ΣΥΣΧΕΤΙΣΗΣ ΑΠΟ ΠΟΛΥΜΕΤΑΒΛΗΤΕΣ ΧΡΟΝΟΣΕΙΡΕΣ

Δ. Χορόζογλου[1], Δ. Κουγιουμτζής[2], Ε. Παπαδημητρίου[1]
[1]Τομέας Γεωφυσικής, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
{chorozod, ritsa}@geo.auth.gr,
[2]Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης,
dkugiu@auth.gr,

ΠΕΡΙΛΗΨΗ

Στην εργασία αυτή εξετάζονται δίκτυα σεισμών ενδιαμέσου μεγέθους (Μ≥5.5) που έγιναν στον ελληνικό χώρο κατά το χρονικό διάστημα 1911 έως 2013. Οι κόμβοι του δικτύου παριστούν σεισμικές ζώνες και η μεταξύ τους σύνδεση δίνεται με δύο διαφορετικές προσεγγίσεις, από κάποιο δείκτη συσχέτισης όταν γίνεται χρήση πολυμεταβλητής χρονοσειράς ή από την διαδοχή δύο σεισμών σε διαφορετικές σεισμικές ζώνες. Μελετάται η δυναμική εξέλιξη της δομής του δικτύου με σκοπό τον προσδιορισμό εκείνων των χρονικών διαστημάτων που παρατηρείται σημαντική αλλαγή στις τιμές των μέτρων δικτύου, όπως ο συντελεστής συσταδοποίησης. Εξετάζεται αν το δίκτυο σεισμικότητας που παράγεται σε κάθε χρονικό διάστημα είναι τυχαίο, δηλαδή οι συνδέσεις του είναι τυχαία ορισμένες, και για το λόγο αυτό γίνεται σύγκριση με κατάλληλα σχηματισμένα τυχαία δίκτυα. Η σύγκριση γίνεται με τη μορφή στατιστικού ελέγχου, όπου ως μηδενική υπόθεση θεωρείται ότι το δίκτυο είναι τυχαίο, και το στατιστικό ελέγχου είναι κάποιο μέτρο δικτύου. Τα αποτελέσματα αναδεικνύουν την αλλαγή της δομής του δικτύου σεισμικότητας κατά το χρονικό διάστημα 1980-2000 και αυτό επιβεβαιώνεται από την απόρριψη της μηδενικής υπόθεσης για πολλά μέτρα δικτύου την περίοδο αυτή.

Λέξεις Κλειδιά: πολυμεταβλητή χρονοσειρά, συντελεστής συσχέτισης, δίκτυο συσχέτισης, μέτρο δικτύου, μέθοδοι τυχαιοποίησης, τυχαιοποιημένο δίκτυο.

1. ΕΙΣΑΓΩΓΗ ΣΤΗ ΣΕΙΣΜΙΚΟΤΗΤΑ ΩΣ ΔΙΚΤΥΟ

Γενικά με τον όρο δίκτυο εννοούμε το γράφημα που ορίζεται από τους κόμβους και τις συνδέσεις μεταξύ τους. Οι συνδέσεις του δικτύου μπορεί να είναι κατευθυνόμενες ή μη, καθώς και να δίνονται με βάρη (σταθμίσεις), έτσι ώστε το

δίκτυο *K* κόμβων να περιγράφεται πλήρως από έναν πίνακα γειτνίασης μεγέθους *K*x*K* με κενά (μηδενικά) διαγώνια στοιχεία και η τιμή σε κάθε θέση (*i*,*j*) του πίνακα να δηλώνει τη σύνδεση των κόμβων *i* και *j*. Ειδικότερα για πίνακα γειτνίασης σε δίκτυο με σταθμισμένες συνδέσεις η τιμή σε κάθε θέση (*i*,*j*) είναι θετικός αριθμός αν υπάρχει σύνδεση και μηδέν αν δεν υπάρχει (συνήθως όλες οι τιμές είναι μη-μηδενικές), ενώ αν οι συνδέσεις είναι απλές η τιμή σε κάθε θέση (*i*,*j*) είναι ένα (1) αν υπάρχει σύνδεση και μηδέν (0) αν δεν υπάρχει. Αν οι συνδέσεις είναι κατευθυνόμενες ο πίνακας γειτνίασης δεν είναι συμμετρικός, ενώ αν είναι μη-κατευθυνόμενες είναι συμμετρικός.

Η προσέγγιση του δικτύου είναι ένα ισχυρό εργαλείο για την ανάλυση δυναμικών δομών πολύπλοκων συστημάτων. Η κατασκευή σεισμικού δικτύου εισήχθη στον χώρο της Σεισμολογίας σχετικά πρόσφατα (Abe and Suzuki 2004), προκειμένου να εκφράσει την πολυπλοκότητα που παρουσιάζει η σεισμικότητα. Παγκόσμιες φυσικές ιδιότητες της σεισμικότητας μπορούν να διερευνηθούν εξετάζοντας γεωμετρικά (τοπολογικά) και δυναμικά χαρακτηριστικά. Το σεισμικό δίκτυο έχει μια σειρά από ενδιαφέρουσες ιδιότητες, μερικές από τις οποίες είναι κοινές με πολλά άλλα φυσικά όσο και τεχνητά πολύπλοκα συστήματα, όπως το μεταβολικό δίκτυο και το δίκτυο του παγκόσμιου ιστού (Albert and Barabasi 2002), οι οποίες παρέχουν τη δυνατότητα μελέτης της σεισμικότητας. Επομένως, η προσέγγιση του δικτύου προσφέρει ένα νέο τρόπο ανάλυσης των σεισμικών δεδομένων και ρίχνει νέο φως στην ερμηνεία της εμφάνισης και συμπεριφοράς της σεισμικότητας.

Στην ανάλυση πολυμεταβλητών χρονοσειρών, δηλαδή μεταβλητών που παρατηρούμε ταυτόχρονα, χρησιμοποιούνται δίκτυα που έχουν ως κόμβους τις παρατηρούμενες μεταβλητές και συνδέσεις που ορίζονται με κάποιο δείκτη συσχέτισης, για μη-κατευθυνόμενες συνδέσεις, ή αιτιότητας, για κατευθυνόμενες συνδέσεις (Newman 2010, Horvath 2011, Campanharo et al. 2011). Ειδικότερα για τα δίκτυα συσχέτισης χρησιμοποιείται ο δειγματικός συντελεστής συσχέτισης για να δώσει σταθμισμένες συνδέσεις ή η σημαντικότητά του για να δώσει μη-σταθμισμένες συνδέσεις (δηλαδή αν κρίνεται σημαντικός τότε υπάρχει σύνδεση αλλιώς όχι). Η σημαντικότητα του δειγματικού συντελεστή συσχέτισης μπορεί να οριστεί απλά από το αν η τιμή του υπερβαίνει κάποιο αυθαίρετο κατώφλι ή από την απόφαση ενός στατιστικού ελέγχου σημαντικότητας.

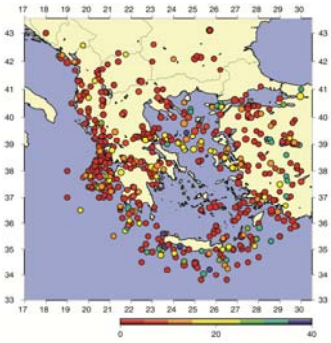
Σκοπός της παρούσας εργασίας είναι η κατασκευή μη-κατευθυνόμενων δικτύων σεισμών στα οποία οι κόμβοι παριστούν σεισμικές ζώνες, οι οποίες είναι ομογενείς με βάση σεισμοτεκτονικά κριτήρια (όπως τύπος διάρρηξης, επίπεδο σεισμικότητας) και οι οποίες περιέχουν υποσύνολα του συνολικού δείγματος δεδομένων. Η σύνδεση μεταξύ των ζωνών δίνεται είτε από κάποιο δείκτη συσχέτισης στην περίπτωση που το σεισμικό δίκτυο δημιουργείται με την βοήθεια πολυμεταβλητής χρονοσειράς, ή από την χρονική διαδοχή δύο σεισμών που έγιναν σε διαφορετικές σεισμικές ζώνες. Όταν, η σύνδεση στο σεισμικό δίκτυο δίνεται από την χρονική διαδοχή δύο σεισμών προτείνεται μία νέα προσέγγιση που αποδίδει σταθμισμένες συνδέσεις στο δίκτυο, δηλαδή συνδέσεις που περιέχουν βάρος ανάλογο με το πλήθος σεισμών που πραγματοποιήθηκε μεταξύ των κόμβων.

Μελετάται η εξέλιξη της δομής του σεισμικού δικτύου που παράγεται σε κάθε χρονικό παράθυρο διάρκειας 20 ετών με παρακολούθηση των τιμών κάποιων μέτρων δικτύου που το χαρακτηρίζουν. Τα αποτελέσματα αποκαλύπτουν αλλαγές στην δομή του δικτύου ανάλογα με την σεισμική δραστηριότητα που παρατηρείται στην περιοχή της μελέτης γεγονός που πιστοποιεί την δυνατή αξιοποίηση της θεωρίας δικτύων ως προς την μελέτη της σεισμικότητας. Η παραγωγή τυχαιοποιημένων δικτύων στην διερεύνηση ύπαρξης ή μη τυχαιότητας στις συνδέσεις του αρχικού σεισμικού δικτύου με την χρήση στατιστικού ελέγχου, αποτυπώνουν την μη τυχαιότητα των συνδέσεων κατά το χρονικό παράθυρο 1980-2000.

2. ΜΕΘΟΔΟΙ ΔΗΜΙΟΥΡΓΙΑΣ ΣΕΙΣΜΙΚΟΥ ΔΙΚΤΥΟΥ

Τα δεδομένα παρατήρησης λήφθηκαν από τον σεισμικό κατάλογο του ελληνικού χώρου (http://geophysics.geo.auth.gr/ss/). Για τον σκοπό της παρούσας μελέτης χρησιμοποιούνται επιφανειακοί σεισμοί (εστιακό βάθος μικρότερο από 40 Km) με μέγεθος M≥5.5 οι οποίοι έγιναν κατά το χρονικό διάστημα 1911-2013. Το δείγμα δεδομένων είναι πλήρες γι αυτό το χρονικό διάστημα (περιέχει όλους τους σεισμούς) και περιλαμβάνει 568 σεισμούς, η χωρική κατανομή των οποίων φαίνεται στην Εικόνα 1.

Εικόνα 1. *Τα επίκεντρα των 568 σεισμών ενδιαμέσου μεγέθους (M≥5.5) κατά την περίοδο 1911-2013 στον ελληνικό χώρο. Η διάμετρος των κύκλων είναι ανάλογη του μεγέθους των σεισμών και το χρώμα ανάλογο του εστιακού βάθους όπως δίνεται από την χρωματική κλίμακα.*



Με μοναδική πληροφορία τον σεισμικό κατάλογο γίνεται προσπάθεια δημιουργίας του σεισμικού δικτύου. Παρακάτω παρουσιάζονται δύο διαφορετικές προσεγγίσεις που αφορούν την δημιουργία του σεισμικού δικτύου με κύρια διαφορά την χρήση ή μη πολυμεταβλητής χρονοσειράς.

2.2 Δημιουργία σεισμικού δικτύου χωρίς χρήση χρονοσειράς

Για την κατασκευή του σεισμικού δικτύου η περιοχή μελέτης χωρίζεται συνήθως σε ισομεγέθεις κυψελίδες, και μία κυψελίδα θεωρείται ως κόμβος του δικτύου αν περιλαμβάνει σεισμό με οποιαδήποτε μέγεθος ή πάνω από ένα ορισμένο κατώφλι μεγέθους. Στην παρούσα εργασία ορίζουμε τους κόμβους διαφορετικά. Χωρίζουμε την γεωγραφική περιοχή που βρίσκεται υπό εξέταση σε σεισμικές ζώνες

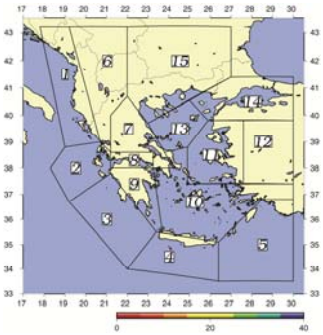
(βλέπε Εικόνα 2) με βάση τις σεισμοτεκτονικές τους ιδιότητες, όπως ο τύπος διάρρηξης και το επίπεδο σεισμικότητας (Papaioannou and Papazachos 2000). Οι *K* σεισμικές ζώνες που δημιουργούνται αποτελούν τους *K* κόμβους του σεισμικού δικτύου. Δύο διαδοχικοί σεισμοί ορίζουν την σύνδεση μεταξύ των κόμβων που πραγματοποιήθηκαν. Αν οι διαδοχικοί σεισμοί έγιναν στην ίδια γεωγραφική περιοχή τότε συνδέεται ένας βρόγχος στον συγκεκριμένο κόμβο. Οι συνδέσεις και οι βρόγχοι αντιπροσωπεύουν τις συσχετίσεις μεταξύ δύο διαδοχικών σεισμών. Με αυτόν τον τρόπο κατασκευάζεται ένα κατευθυνόμενο δίκτυο με κατεύθυνση από τον προηγούμενο στον επόμενο σεισμό. Έχει αποδειχθεί ότι ένας μελλοντικός σεισμός μπορεί να προκληθεί και να εκδηλωθεί από αμέσως προηγούμενο πολύ ισχυρό σεισμό που έγινε σε απόσταση μεγαλύτερη των 1.000 χιλιομέτρων (Steeples and Steeples 1996). Αυτό αποτελεί ένδειξη ότι δύο διαδοχικοί σεισμοί συσχετίζονται, ανεξάρτητα από την χωρική τους απόσταση.

Ένα συχνό φαινόμενο που παρατηρείται είναι η δημιουργία πολλαπλών συνδέσεων μεταξύ των κόμβων, δηλαδή η διαδοχή σεισμών σε ίδιες σεισμικές ζώνες περισσότερες από μία φορά στο χρονικό παράθυρο μελέτης. Οι παραπάνω προσεγγίσεις δημιουργίας του σεισμικού δικτύου δημιουργούν απλές συνδέσεις μεταξύ των κόμβων με αποτέλεσμα να χάνεται αρκετή πληροφορία από τα δεδομένα. Έτσι, προτείνουμε μία νέα προσέγγιση που δημιουργεί σταθμισμένες συνδέσεις στο σεισμικό δίκτυο. Η διαδικασία έχει ως εξής: Για κάθε χρονικό παράθυρο δημιουργούμε έναν τετραγωνικό πίνακα $S = \{s\}_{ij}$, όπου s_{ij} ακέραιος αριθμός που δηλώνει πόσες φορές εμφανίστηκε διαδοχή σεισμών μεταξύ των κόμβων *i* και *j*. Έχοντας δημιουργήσει *Z* τέτοιους τετραγωνικούς πίνακες, όσα είναι τα χρονικά παράθυρα μελέτης, ορίζουμε τον πίνακα βαρών W σε κάθε χρονικό παράθυρο ως $W = \frac{S}{\max\{\{s\}_{ij}\}}$, όπου $\max\{\{s\}_{ij}\}$ η μέγιστη τιμή όλων των *S* πινάκων. Με αυτόν τον τρόπο ορίζουμε σταθμισμένες συνδέσεις στο διάστημα [0,1].

2.3 Δημιουργία σεισμικού δικτύου συσχέτισης

Η κατασκευή του σεισμικού δικτύου συσχέτισης προϋποθέτει τον σχηματισμό της πολυμεταβλητής χρονοσειράς. Αρχικά, οι κόμβοι του δικτύου αντιπροσωπεύουν τις σεισμικές ζώνες που αναφέρθηκαν παραπάνω. Η σύνδεση, στην προσέγγιση αυτή, δεν δίνεται από την διαδοχή δύο σεισμών αλλά από κάποιον δείκτη συσχέτισης που υπολογίζεται στην πολυμεταβλητή χρονοσειρά. Επομένως, υπάρχει η ανάγκη της δημιουργίας πολυμεταβλητής χρονοσειράς η οποία προκύπτει από την αξιοποίηση των δεδομένων.

Εικόνα 2. *Οι 15 σεισμικές ζώνες που αποτελούν τους κόμβους του δικτύου.*



Η δημιουργία της πολυμεταβλητής χρονοσειράς βασίζεται στην χρήση ενός μέτρου σεισμικότητας, όπως πλήθους σεισμών (Jimenez et al. 2008) ή έκλυση σεισμικής ροπής (Tenenbaum et al. 2012), για την ποσοτικοποίηση της σεισμικής δραστηριότητας που παρατηρείται σε κάθε σεισμική ζώνη την ίδια χρονική στιγμή. Αυτό έχει ως αποτέλεσμα την δημιουργία χρονοσειράς για κάθε σεισμική ζώνη στο παράθυρο μελέτης, και το σύνολο των χρονοσειρών από όλες τις σεισμικές ζώνες σχηματίζει την πολυμεταβλητή χρονοσειρά.

Στην παρούσα εργασία, η πολυμεταβλητή χρονοσειρά δημιουργήθηκε από τον ετήσιο αριθμό σεισμών που έγιναν σε κάθε σεισμική ζώνη και την ετήσια σεισμική ροπή M_0 που εκλύεται σε κάθε σεισμική ζώνη. Η σεισμική ροπή υπολογίστηκε από την εμπειρική σχέση $logM_0 = 1.5M + 16.01$, όπου M το μέγεθος του σεισμού (Kanamori and Anderson 1975). Η πολυμεταβλητή χρονοσειρά, είτε του πλήθους σεισμών ή της σεισμικής ροπής, έχει 103 παρατηρήσεις για κάθε σεισμική ζώνη καθώς τα δεδομένα προέρχονται από την χρονική περίοδο 1911-2013. Η ανάλυση και με τις δύο προσεγγίσεις, διαδοχή σεισμών και πολυμεταβλητή χρονοσειρά, έγινε σε κυλιόμενα επικαλυπτόμενα χρονικά παράθυρα 20 παρατηρήσεων (20 ετών) με την επικάλυψη να εκτίνεται σε παράθυρο 10 παρατηρήσεων (10 ετών). Σε κάθε ένα από τα 9 χρονικά παράθυρα σχηματίστηκαν τα δίκτυα με τις δύο προσεγγίσεις, με και χωρίς χρήση πολυμεταβλητής χρονοσειράς.

2.4 Ορισμός σύνδεσης δικτύου συσχέτισης

Στα δίκτυα συσχέτισης οι κόμβοι αντιστοιχούν σε τυχαίες μεταβλητές και οι συνδέσεις δίνονται από κάποιο στατιστικό διασυσχέτισης, όπως ο δειγματικός συντελεστής συσχέτισης Pearson (Horvath 2011, Κεφ. 5). Άρα για ένα σύνολο K τυχαίων μεταβλητών, $X_1,...,X_K$, θεωρούμε ένα δείγμα n, $\{x_{1,t},\ldots,x_{K,t}\}$ για $t=1,\ldots,n$ ενδεχομένως εξαρτημένων παρατηρήσεων της ίδιας μεταβλητής, το οποίο δηλώνει την ύπαρξη σημαντικής αυτοσυσχέτισης, αλλά και μεταξύ των μεταβλητών, το οποίο δηλώνει την ύπαρξη σημαντικής διασυσχέτισης. Το ενδιαφέρον εδώ εστιάζεται σε μη-κατευθυνόμενες συνδέσεις δικτύου και γι' αυτό εξετάζουμε τη γραμμική διασυσχέτιση χωρίς χρονική υστέρηση, που είναι ο συντελεστής συσχέτισης Pearson. Για δύο μεταβλητές $X=X_i$ και $Y=X_j$, $i,j \in \{1,\ldots,K\}$, η εκτίμηση του συντελεστή συσχέτισης ορίζεται ως $r_{XY} = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}}$, όπου $s_{XY} = \frac{1}{n-1}\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})$ είναι

η δειγματική συνδιασπορά των X και Y και $s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ είναι η δειγματική διασπορά της X και $\bar{x}$ ο μέσος όρος.

Είναι γνωστό πως ο πίνακας συσχέτισης Σ, που έχει στη θέση (i,j) το δειγματικό συντελεστή συσχέτισης Pearson των X_i και X_j, είναι πάντα θετικά ημι-ορισμένος, δηλαδή οι ιδιοτιμές του είναι μη-αρνητικές. Αυτή η συνθήκη θα πρέπει να τηρείται ακόμα και αν οι τυχαίες μεταβλητές (κόμβοι) είναι ασυσχέτιστες, δηλαδή αν το αντίστοιχο δίκτυο είναι τυχαίο. Για μη-σταθμισμένες συνδέσεις, ο πίνακας γειτνίασης A προκύπτει από τον πίνακα συσχέτισης Σ με κάποια μέθοδο που θέτει σύνδεση (τιμή ένα) όταν η συσχέτιση κρίνεται σημαντική και δε θέτει σύνδεση (τιμή μηδέν) όταν η συσχέτιση κρίνεται ασήμαντη. Το κριτήριο μπορεί να είναι ένα αυθαίρετο κατώφλι, ή ένα κατώφλι που αντιστοιχεί σε καθορισμένη πυκνότητα συνδέσεων ή να προκύπτει από στατιστικό έλεγχο σημαντικότητας (Horvath 2011, Κεφ. 10). Για τον παραμετρικό έλεγχο σημαντικότητας με μηδενική υπόθεση H_0: $\rho = 0$, μετασχηματίζεται ο δειγματικός συντελεστής συσχέτισης του Pearson στο στατιστικό ελέγχου $t = r\sqrt{\frac{n-2}{1-r^2}}$ που ακολουθεί κατανομή Student με n-2 βαθμούς ελευθερίας, όπου n το πλήθος των παρατηρήσεων της χρονοσειράς. Αντίθετα, οι σταθμισμένες συνδέσεις δίνονται από τον πίνακα βαρών W ο οποίος προκύπτει με τον εξής τρόπο: Δημιουργούμε ένα νέο πίνακα M με στοιχεία $M_{i,j} = r_{i,j} - \bar{r}_{i,j} + 1$ (για να είναι τα βάρη πάντα θετικά), όπου $r_{i,j}$ οι τιμές του δειγματικού συντελεστή συσχέτισης του Pearson για τις μεταβλητές X_i και X_j και $\bar{r}_{i,j}$ η μέση τιμή των $r_{i,j}$, όπου $i, j = 1, \ldots, K$. Από τον πίνακα M δημιουργείται ο πίνακας βαρών $W_{i,j} = \frac{M_{i,j}}{\max(M)}$, όπου $\max(M)$ η μέγιστη τιμή των στοιχείων του πίνακα M. Γενικά, ο πίνακας βαρών ορίζει το σταθμισμένο δίκτυο με σταθμισμένες συνδέσεις στο διάστημα [0,1].

Η είσοδος της σύνδεσης μεταξύ των κόμβων αντιπροσωπεύει παρόμοια πρότυπα σεισμικής δραστηριότητας διαφορετικών σεισμικών ζωνών. Έτσι, ορίζουμε τις συνδέσεις μεταξύ των κόμβων με βάση την μακροπρόθεσμη ομοιότητα που παρουσιάζουν σε κάποιο μέτρο σεισμικότητας (μεγέθους σεισμών, πλήθους σεισμών, ρυθμού έκλυσης σεισμικής ροπής).

3. ΜΕΤΡΑ ΔΙΚΤΥΟΥ

Για τη μελέτη της εξέλιξης της δομής του σεισμικού δικτύου, υπολογίζουμε σε κάθε ένα από τα 9 χρονικά παράθυρα την τιμή που παρουσιάζουν κάποια αντιπροσωπευτικά μέτρα δικτύου. Η τιμή καθενός από τα μέτρα δικτύου αποδίδει διαφορετική σχετική πληροφορία (βλέπε Πίνακα 1). Παρακάτω παρατίθενται έννοιες που αφορούν την θεωρία δικτύων για την καλύτερη κατανόηση της ερμηνείας κάθε μέτρου δικτύου.

Σε ένα δίκτυο G, για κάθε δύο κόμβους i και j η γεωδαισιακή απόσταση τους $d(i,j)$ ορίζεται ως το μήκος της συντομότερης διαδρομής από τον i στο j, εφόσον οι κόμβοι αυτοί είναι συνδεδεμένοι ενώ $d(i,j) = \infty$ διαφορετικά.

Γείτονες: Έστω ο μη κατευθυνόμενος γράφος $G = (N, E)$ και $i, j \in N$ δύο κόμβοι του. Ο j λέγεται γείτονας του i όταν $(i,j) \in E$, όπου *N* είναι το σύνολο των κόμβων και *E* το σύνολο των συνδέσεων του δικτύου.

Πίνακας γειτνίασης *A*: Στην περίπτωση μη κατευθυνόμενου και μη σταθμισμένου δικτύου είναι ένας συμμετρικός πίνακας $[A = \{a\}_{ij}], i, j \in N$ τάξης $|N| * |N|$ τέτοιος ώστε $a_{ij} = 1$, όταν *i, j* είναι γείτονες και $a_{ij} = 0$, διαφορετικά.

Πίνακας βαρών *W*: Στην περίπτωση μη κατευθυνόμενου και σταθμισμένου δικτύου είναι ένας συμμετρικός πίνακας $[W = \{w\}_{ij}], i, j \in N$ τάξης $|N| * |N|$, όπου w_{ij} είναι το βάρος που χαρακτηρίζει την σταθμισμένη σύνδεση.

Η διακύμανση της τιμής πολλών από τα μέτρα δικτύου, όπως για παράδειγμα του συντελεστή συσταδοποίησης (Clustering coefficient), μπορεί ακόμα και να προειδοποιήσει αν αναμένεται ένας ισχυρός σεισμός στο άμεσο μέλλον καθώς έχει παρατηρηθεί σταδιακή αύξηση της τιμής τους πριν από έναν ισχυρό σεισμό και στη συνέχεια μείωση για να επανέλθει σε μία σταθερή τιμή (Abe and Suzuki 2009).

Το μέτρο του συντελεστή συσταδοποίησης (Clustering coefficient) εκφράζει την πιθανότητα που έχουν δυο κόμβοι να συνδέονται με έναν κοινό τους κόμβο αλλά και να είναι γείτονες μεταξύ τους. Υψηλή τιμή του συντελεστή συσταδοποίησης υποδεικνύει μεγάλη πιθανότητα ύπαρξης συστάδας στο δίκτυο, δηλαδή μιας ομάδας κόμβων στην οποία όλοι οι κόμβοι συνδέονται με όλους τους άλλους σχηματίζοντας ένα πλήρες δίκτυο.

Για τον υπολογισμό των τιμών των μέτρων δικτύου χρησιμοποιήθηκαν οι συναρτήσεις του Brain Connectivity Toolbox σε περιβάλλον Matlab, (https://sites.google.com/site/bctnet/measures).

4. ΜΕΛΕΤΗ ΤΗΣ ΔΟΜΗΣ ΤΟΥ ΣΕΙΣΜΙΚΟΥ ΔΙΚΤΥΟΥ

Η περιοχή μελέτης, δηλαδή ο ευρύτερος ελληνικός χώρος, χωρίστηκε σε 15 σεισμικές ζώνες οι οποίες αποτελούν και τους κόμβους του σεισμικού δικτύου όπως ήδη αναφέρθηκε. Η σύνδεση δίνεται από τον δειγματικό συντελεστή συσχέτισης του Pearson όταν γίνεται χρήση πολυμεταβλητής χρονοσειράς (αριθμού σεισμών ή σεισμικής ροπής) ή από την διαδοχή δύο σεισμών σε διαφορετικούς κόμβους. Έχοντας σχηματίσει το σεισμικό δίκτυο, και με τις δύο προσεγγίσεις, για κάθε ένα από τα 9 κυλιόμενα χρονικά παράθυρα διερευνώνται τυχόν αλλαγές στην δομή του με οπτική παρατήρηση. Τα αποτελέσματα της Εικόνας 3, όπου η σύνδεση δίνεται από την διαδοχή σεισμών και οι συνδέσεις είναι μη-σταθμισμένες, φανερώνουν σημαντικές αλλαγές στην εξέλιξη της δομής του δικτύου καθώς κατά την περίοδο 1920-1940 (βλέπε Εικόνα 3a) το δίκτυο είναι σχετικά αραιό ως προς τις συνδέσεις, στην συνέχεια κατά την περίοδο 1950-1970 γίνεται αρκετά πυκνό (βλέπε Εικόνα 3b) και τέλος την περίοδο 1980-2000 γίνεται πολύ αραιό (βλέπε Εικόνα 3c).

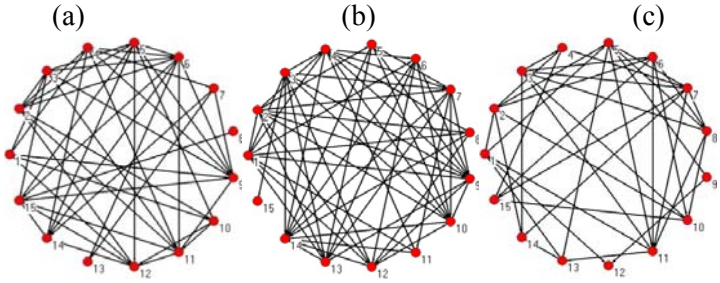
***Πίνακας 1**: Αποτύπωση της πληροφορίας που αποδίδει κάθε μέτρο δικτύου.*

Τιμή δικτύου / **Μέτρο δικτύου**	**Μικρή**	**Μεγάλη**
Degree/Strength	Λίγες συνδέσεις	Πολλές συνδέσεις
Density	Αραιό	Πυκνό
Clustering coefficient	Απουσία συνδεδεμένων κόμβων ανά τριάδες	Ύπαρξη συνδεδεμένων κόμβων ανά τριάδες
Transitivity	Ασθενής τάση ομαδοποίησης κόμβων ανά τριάδες	Ισχυρή τάση ομαδοποίησης κόμβων ανά τριάδες
Global efficiency	Μικρή αποδοτικότητα	Μεγάλη αποδοτικότητα
Assortativity	Μικρή τάση σύνδεσης κόμβων με όμοιο βαθμό (Degree)	Ισχυρή τάση σύνδεσης κόμβων με όμοιο βαθμό (Degree)
Betweenness centrality	Μικρός έλεγχος ροής	Μεγάλος έλεγχος ροής
Eigenvector centrality	Λίγοι ή λίγο σημαντικοί γείτονες	Πολλοί ή πολύ σημαντικοί γείτονες
Pagerank	Απουσία σημαντικών κόμβων στην ροή πληροφορίας	Ύπαρξη σημαντικών κόμβων στην ροή πληροφορίας
Characteristic path length	Μικρές αποστάσεις μεταξύ κόμβων	Μεγάλες αποστάσεις μεταξύ κόμβων
Eccentricity	Εύκολη επικοινωνία μεταξύ κόμβων	Δύσκολη επικοινωνία μεταξύ κόμβων
Diameter	Εύκολη μεταφορά πληροφορίας	Δύσκολη μεταφορά πληροφορίας

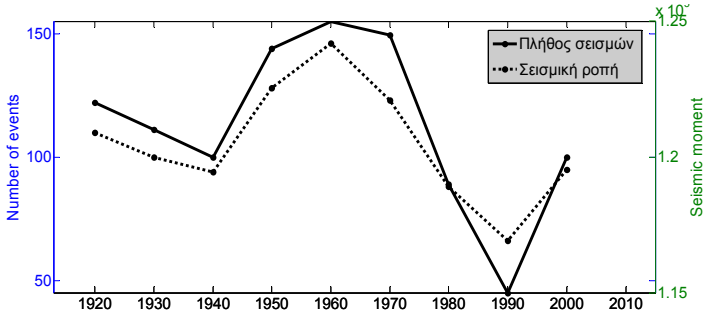
Τα αποτελέσματα της Εικόνας 3, σχετικά με το πλήθος των συνδέσεων, μπορούμε να δεχτούμε ότι ήταν αναμενόμενα καθώς αυτά σχετίζονται με το πλήθος των σεισμών που γίνονται σε κάθε χρονικό παράθυρο καθώς η σύνδεση δίνεται από την διαδοχή δύο σεισμών. Η παραπάνω παραδοχή πιστοποιείται από την Εικόνα 4 που δείχνει το συνολικό πλήθος σεισμών και σεισμικής ροπής από όλες τις ζώνες ανά χρονικό παράθυρο 20 ετών. Παρατηρούμε πως την περίοδο 1950-1970 (παρατήρηση 1960, Εικόνα 4) με τις περισσότερες συνδέσεις στο δίκτυο (βλέπε Εικόνα 3b) πραγματοποιήθηκαν οι περισσότεροι σεισμοί από οποιοδήποτε άλλο χρονικό παράθυρο οι οποίοι έφτασαν σε αριθμό τους 150.

Ιδιαίτερο ενδιαφέρον παρουσιάζει η Εικόνα 5 που παρουσιάζει τα δίκτυα για τα ίδια χρονικά παράθυρα με την Εικόνα 3, όπου η σύνδεση όμως δίνεται από τον υπολογισμό του δειγματικού συντελεστή συσχέτισης του Pearson στην πολυμεταβλητή χρονοσειρά η οποία ορίστηκε με μέτρο σεισμικότητας το πλήθος των σεισμών. Τα αποτελέσματα βρίσκονται σε συμφωνία με την Εικόνα 3 τονίζοντας την επιτυχία της προσέγγισης της πολυμεταβλητής χρονοσειράς να αποτυπώνει αξιόπιστα την εξέλιξη της σεισμικότητας.

Εικόνα 3. *Εξέλιξη της δομής του σεισμικού δικτύου που κατασκευάστηκε χωρίς την χρήση πολυμεταβλητής χρονοσειράς κατά τα χρονικά παράθυρα a) 1920-1940 b) 1950-1970 και c) 1980-2000.*

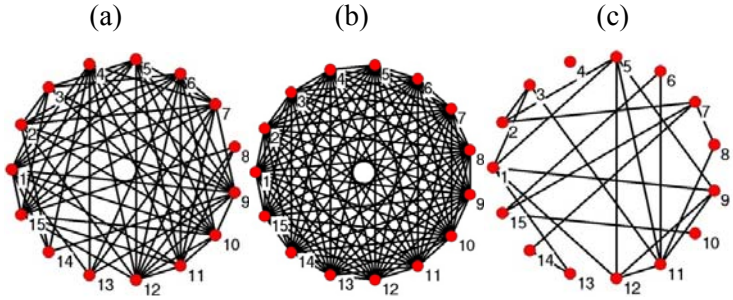


Εικόνα 4. *Αποτύπωση της σεισμικής δραστηριότητας σε κάθε ένα από τα 9 χρονικά παράθυρα με το συνολικό πλήθος σεισμών και την αθροιστική σεισμική ροπή* M_0.



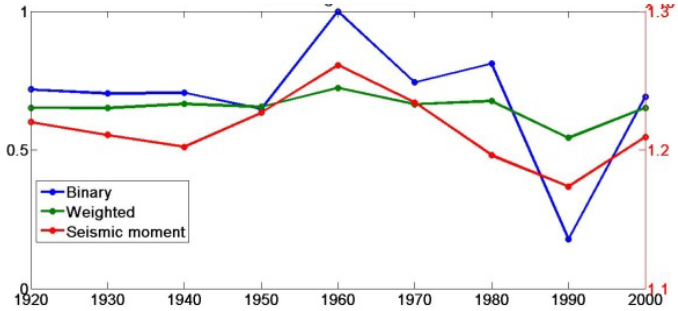
Αφού παρατηρήσαμε την εξέλιξη της δομής του δικτύου, προχωρούμε στην μελέτη της μεταβολής των τιμών των μέτρων δικτύου σε κάθε χρονικό παράθυρο. Ενδεικτικά, η Εικόνα 6 παρουσιάζει την εξέλιξη της τιμής του συντελεστή συσταδοποίησης (Clustering coefficient) στο σεισμικό δίκτυο, όπου οι συνδέσεις δίνονται από τον δειγματικό συντελεστή συσχέτισης του Pearson με μέτρο σεισμικότητας την έκλυση σεισμικής ροπής για να δημιουργηθεί η πολυμεταβλητή χρονοσειρά.

Εικόνα 5. *Εξέλιξη της δομής του σεισμικού δικτύου που κατασκευάστηκε με την χρήση πολυμεταβλητής χρονοσειράς κατά τα χρονικά παράθυρα a) 1920-1940 b) 1950-1970 και c) 1980-2000.*



Ο συντελεστής συσταδοποίησης υπολογίστηκε για σταθμισμένες συνδέσεις (Weighted) και για μη σταθμισμένες συνδέσεις από ένα αυθαίρετο κατώφλι (Binary). Παρατηρούμε την πολύ καλή προσαρμογή του μέτρου σε σχέση με την σεισμική δραστηριότητα που παρατηρείται, καθώς η τιμή του σε κάθε περίπτωση φτάνει σε μία μέγιστη τιμή εκεί που παρατηρείται έντονη σεισμικότητα και στην συνέχεια μειώνεται καθώς έχουμε ύφεση του φαινομένου (βλέπε Εικόνα 6).

Εικόνα 6. *Αποτύπωση της μεταβολής των τιμών του συντελεστή συσταδοποίησης* (Clustering coefficient) *σε σταθμισμένο (Weighted) και μη σταθμισμένο (Binary) δίκτυο σε σχέση με την αθροιστική σεισμική ροπή* M_0 *που εκλύεται σε κάθε παράθυρο.*



5. ΜΕΘΟΔΟΙ ΤΥΧΑΙΟΠΟΙΗΣΗΣ ΔΙΚΤΥΩΝ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ

Ένα πρώτο ερώτημα στην ανάλυση δικτύων είναι η ύπαρξη η μη τυχαιότητας. Για μία πολυμεταβλητή χρονοσειρά, το αντίστοιχο δίκτυο συσχέτισης είναι τυχαίο όταν οι κόμβοι αντιστοιχούν σε ανεξάρτητες (ή ασυσχέτιστες όταν περιοριζόμαστε σε γραμμική ανάλυση) τυχαίες μεταβλητές, δηλαδή οι υπάρχουσες συνδέσεις στο δίκτυο εμφανίζονται τυχαία. Το δίκτυο που δημιουργείται από την διαδοχή σεισμών θεωρείται τυχαίο πάλι αν οι συνδέσεις εμφανίζονται τυχαία.

5.1 Κλασικές μέθοδοι τυχαιοποίησης δικτύων

Το πρόβλημα που εξετάζουμε στην τυχαιοποίηση δικτύων είναι να δημιουργήσουμε ένα αντίστοιχο με το δεδομένο δίκτυο αλλά με τυχαίες συνδέσεις. Η πιο απλή κατασκευή τυχαιοποιημένου δικτύου με σταθμισμένες συνδέσεις ή μη-σταθμισμένες συνδέσεις γίνεται με τυχαίο ανασχηματισμό των αρχικών συνδέσεων ώστε να διατηρείται η συνολική ισχύς (total strength) ή αντίστοιχα ο συνολικός βαθμός (total degree) του αρχικού δικτύου (Newman 2010). Με αναφορά στην ισχύ ονομάζουμε αυτήν την μέθοδο τυχαιοποίησης RNavestr.

Μια πιο αυστηρή συνθήκη στην τυχαιοποίηση δικτύου είναι η διατήρηση του βαθμού (για μη-σταθμισμένες συνδέσεις) ή της ισχύος (για σταθμισμένες συνδέσεις) του κάθε κόμβου. Μεταξύ διαφόρων προσεγγίσεων του προβλήματος αυτού για μη-σταθμισμένες συνδέσεις χρησιμοποιείται ο αλγόριθμος των Maslov και Sneppen (2002). Η διαδικασία είναι η εξής: Επιλέγονται τυχαία δύο συνδέσεις του αρχικού δικτύου, π.χ. (i,j) και (k,l) και με εναλλαγή των δύο τελικών κόμβων j και l δημιουργούνται δύο νέες συνδέσεις (i,l) και (k,j). Η εναλλαγή γίνεται στους τελικούς κόμβους καθώς θέλουμε να διατηρήσουμε το βαθμό του κάθε κόμβου στο

τυχαιοποιημένο δίκτυο. Αν αυτές οι δύο νέες συνδέσεις δεν υπάρχουν ήδη τις αποδεχόμαστε και διαγράφουμε τις αρχικές. Στην περίπτωση που οι συνδέσεις που προκύπτουν υπάρχουν ήδη στο αρχικό δίκτυο τότε επιλέγεται τυχαία άλλο ζεύγος συνδέσεων. Μετά από πολλές επαναλήψεις, η διαδικασία αυτή δημιουργεί μία τυχαιοποιημένη παραλλαγή του αρχικού δικτύου που διατηρεί το πλήθος συνδέσεων κάθε κόμβου. Η μέθοδος τυχαιοποίησης των Maslov και Sneppen αναφέρεται στη συνέχεια ως μέθοδος RNnoddeg. Οι μέθοδοι RNavestr και RNnoddeg είναι κατάλληλοι για την τυχαιοποίηση δικτύων που ορίζονται με την διαδοχή σεισμών αλλά όχι για δίκτυα συσχέτισης (Χορόζογλου and Κουγιουμτζής 2014).

5.2. Μέθοδος τυχαιοποίησης δικτύων με τυχαιοποίηση στην πολυμεταβλητή χρονοσειρά

Για δίκτυα συσχέτισης από πολυμεταβλητές χρονοσειρές, προτείνεται μία άλλη μέθοδος τυχαιοποίησης που βασίζεται στη δημιουργία υποκατάστατης (surrogate) πολυμεταβλητής χρονοσειράς που μοιάζει με την αρχική πολυμεταβλητή χρονοσειρά αλλά οι μεταβλητές έχουν μηδενική διασυσχέτιση (Chorozoglou and Kugiumtzis 2014, Χορόζογλου and Κουγιουμτζής 2014). Άρα και το δίκτυο συσχέτισης που προκύπτει από αυτήν την υποκατάστατη πολυμεταβλητή χρονοσειρά είναι τυχαίο.

Για τη δημιουργία της υποκατάστατης πολυμεταβλητής χρονοσειράς εφαρμόζεται ξεχωριστά σε κάθε μια από τις K χρονοσειρές $\{X_{1,t},...,X_{K,t}\}$ για t=1,…,n, ένας αλγόριθμος δημιουργίας υποκατάστατων μονομεταβλητών χρονοσειρών που τυχαιοποιεί τις n παρατηρήσεις της χρονοσειράς διατηρώντας όμως την περιθώρια κατανομή και συνάρτηση αυτοσυσχέτισης (Kugiumtzis 2002) ή ισοδύναμα το φάσμα ισχύος (Schreiber and Schmitz 1996). Χρησιμοποιείται ο αλγόριθμος Iterative Amplitude Adjusted Fourier Transform (IAAFT) των Schreiber και Schmitz (1996) ως ο πλέον κατάλληλος για πολύ μικρές χρονοσειρές που χρησιμοποιούνται εδώ. Η διαδικασία σχηματισμού τυχαιοποιημένου δικτύου συσχέτισης με σταθμισμένες συνδέσεις δίνεται στα παρακάτω βήματα.

1) Για κάθε χρονοσειρά $\{X_{i,t}\}$, i=1,..,K, δημιουργείται υποκατάστατη χρονοσειρά με τον αλγόριθμο IAAFT, $\{X^*_{i,t}\}$, και έτσι προκύπτει η υποκατάστατη πολυμεταβλητή χρονοσειρά $\{X^*_{1,t}, ..., X^*_{K,t}\}$.

2) Υπολογίζεται ο πίνακας συσχέτισης Σ^* της $\{X^*_{1,t}, ..., X^*_{K,t}\}$.

3) Δημιουργείται το δίκτυο συσχέτισης από τον πίνακα βαρών W που προκύπτει από τον πίνακα συσχέτισης Σ^*.

4) Επαναλαμβάνονται τα βήματα 1) ως 3) B φορές για να δημιουργηθούν B τυχαιοποιημένα δίκτυα συσχέτισης.

Η παραπάνω μέθοδος τυχαιοποίησης δικτύου για σταθμισμένες συνδέσεις ονομάζεται RTSweight. Για τη δημιουργία τυχαιοποιημένων δικτύων με μη-σταθμισμένες συνδέσεις, η διαδικασία είναι παρόμοια με αλλαγή στο βήμα 3. Εδώ υπάρχουν δύο διαφορετικές προσεγγίσεις. Στην πρώτη ακολουθείται η ίδια διαδικασία μετατροπής του πίνακα συσχέτισης Σ σε πίνακα γειτνίασης A, όπως στο αρχικό δίκτυο, π.χ. με κάποιο κατώφλι ή με έλεγχο σημαντικότητας για το r_{xy}. Χρησιμοποιούμε παρακάτω το κατώφλι για τον σχηματισμό μη σταθμισμένων

συνδέσεων και ονομάζουμε RTSbinthr την μέθοδο αυτή. Σε αυτήν την περίπτωση δεν διατηρείται απαραίτητα ο συνολικός βαθμός. Για τον λόγο αυτό θεωρείται και μια δεύτερη προσέγγιση, όπου τίθεται ένα κατώφλι για τα στοιχεία Σ^* τέτοιο ώστε να προκύπτει ο ίδιος αριθμός συνδέσεων με το αρχικό δίκτυο. Για την εύρεση του κατωφλίου πρώτα διατάσσονται τα στοιχεία του άνω τριγωνικού μέρους του πίνακα συσχέτισης Σ^* σε φθίνουσα σειρά. Το κατώφλι είναι η τιμή του στοιχείου που έχει θέση διάταξης ίση με το συνολικό βαθμό του αρχικού δικτύου. Η μέθοδος αυτή ονομάζεται RTSbindeg.

5.3 Στατιστικός έλεγχος

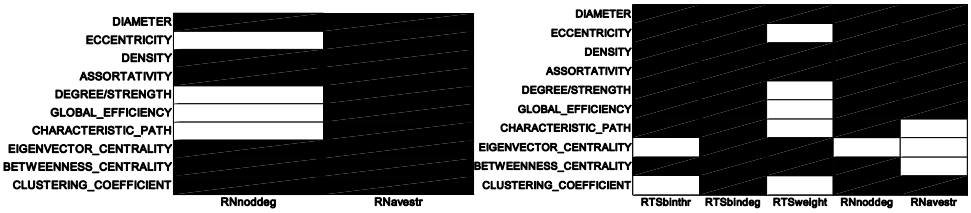
Μετά τη δημιουργία B τυχαιοποιημένων δικτύων ακολουθεί η πραγματοποίηση του ελέγχου για τη μηδενική υπόθεση H_0 ότι το αρχικό δίκτυο είναι τυχαίο, δηλαδή οι μεταβλητές οι οποίες αντιπροσωπεύουν τις σεισμικές ζώνες που παρατηρούνται με την πολυμεταβλητή χρονοσειρά είναι ασυσχέτιστες και οι όποιες συνδέσεις σχηματίστηκαν είναι τυχαίες. Ως στατιστικό ελέγχου θεωρείται ένα μέτρο δικτύου. Το κάθε μέτρο δικτύου q υπολογίζεται στο αρχικό δίκτυο και έστω q_0 η τιμή του, και στα B τυχαιοποιημένα δίκτυα οι τιμές του είναι $q_1,\ldots,q_B$. Εξετάζεται η θέση r_0 της τιμής q_0 στη διατεταγμένη σειρά των $q_0,q_1,\ldots,q_B$, και η p-τιμή του ελέγχου θεωρώντας επίπεδο σημαντικότητας $\alpha = 0.05$ δίνεται από:

(1)

Πραγματοποιώντας τον παραπάνω στατιστικό έλεγχο διαπιστώνουμε την απόρριψη της μηδενικής υπόθεσης H_0, ότι το αρχικό δίκτυο είναι τυχαίο, από πολλά μέτρα δικτύου που χρησιμοποιήθηκαν ως στατιστικό ελέγχου ειδικότερα στο χρονικό διάστημα 1980-2000 (βλέπε Εικόνα 7). Ειδικότερα, στην Εικόνα 7b και με την μέθοδο RTSweight η απόρριψη της μηδενικής υπόθεσης H_0 από αρκετά μέτρα δικτύου σηματοδοτεί την ανάγκη περαιτέρω ανάλυσης εκείνης της χρονικής περιόδου για ενδεχόμενη "ανασκαφή" πληροφορίας σχετικά με την φυσική των σεισμών, το οποίο αποτυπώνεται ανάλογα στην Εικόνα 7a με την μέθοδο RNnoddeg.

Εικόνα 7. *Συγκεντρωτικός πίνακας απορρίψεων της μηδενικής υπόθεσης H_0 κατά την χρονική περίοδο 1980-2000, όπου a) η σύνδεση του δικτύου δίνεται από την διαδοχή δύο σεισμών και b) από τον υπολογισμό του συντελεστή συσχέτισης στην πολυμεταβλητή χρονοσειρά με μέτρο σεισμικότητας την σεισμική ροπή M_0. Στον y-άξονα βρίσκεται το στατιστικό ελέγχου και στον x-άξονα η μέθοδος τυχαιοποίησης του δικτύου, με άσπρο χρώμα παριστάνεται η απόρριψη της μηδενικής υπόθεσης H_0 σε αντίθεση με το μαύρο που αντιπροσωπεύει την αποδοχή της.*

(a) (b)



5.4 Αποτύπωση τυχαιότητας δικτύων μέσω $Z-score$

Ένας τρόπος να μετρηθεί η απόκλιση του σεισμικού δικτύου από το τυχαίο δίκτυο, είναι με το $Z-score$. Η τιμή Z δίνεται από: $Z=\frac{q_0-\bar{q}}{S_q}$, όπου q_0 η τιμή του μέτρου δικτύου στο αρχικό δίκτυο, $\bar{q}$ και S_q ο μέσος όρος και η τυπική απόκλιση των τιμών του μέτρου δικτύου των B τυχαιοποιημένων δικτύων, q

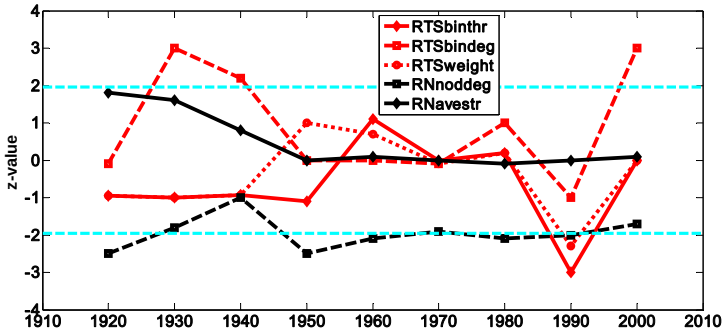
αντίστοιχα. Υποθέτοντας πως ο δείκτης $Z-score$ στα τυχαιοποιημένα δίκτυα ακολουθεί

κανονική κατανομή, το μπορεί να χρησιμοποιηθεί για την απόφαση του ελέγχου σε αντιστοιχία με την $p-τιμή$ που ορίστηκε στην (1) χωρίς την υπόθεση γνωστής μηδενικής κατανομής. Για επίπεδο σημαντικότητας a η περιοχή απόρριψης R του στατιστικού ελέγχου για την τιμή Z είναι $R=\{Z|Z<Z_1(a/2)v\,Z>Z_1(1-a/2)\}$. Για $a=0.05$ οι τιμές είναι $Z<-1.96$ ή $Z>1.96$. Το $Z-score$ δείχνει καλύτερα από την $p-τιμή$ την απόκλιση από την H_0 (τυχαίο δίκτυο) και το χρησιμοποιούμε για να σχηματίσουμε το προφίλ της απόκλισης από το τυχαίο δίκτυο, με βάση κάποιο μέτρο δικτύου, ως προς τα επικαλυπτόμενα χρονικά παράθυρα.

Στην Εικόνα 8 φαίνεται η χρονική μεταβολή του συντελεστή συσταδοποίησης με τις 5 μεθόδους τυχαιοποίησης δικτύου, όπου η πολυμεταβλητή χρονοσειρά κατασκευάστηκε με μέτρο σεισμικότητας την σεισμική ροπή M_0. Τα αποτελέσματα της Εικόνας 8 επιβεβαιώνουν την Εικόνα 7 καθώς παρατηρούνται συστηματικά απορρίψεις της μηδενικής υπόθεσης H_0 (βλέπε Εικόνα 8, παρατηρήσεις έξω από τα όρια των οριζοντίων γαλάζιων γραμμών) το χρονικό διάστημα 1980-2000 από πολλά μέτρα δικτύου για τις μεθόδους τυχαιοποίησης του αρχικού δικτύου.

Εικόνα 8. *Αποτύπωση της τυχαιότητας στις συνδέσεις του σεισμικού δικτύου μέσω $Z-score$ (y-άξονας) με μέτρο δικτύου τον συντελεστή συσταδοποίησης (Clustering coefficient).*

Παρατηρήσεις έξω από τα όρια των οριζόντιων γαλάζιων γραμμών εκφράζουν απόρριψη της μηδενικής υπόθεσης H_0.



6. ΣΥΜΠΕΡΑΣΜΑ

Η προσέγγιση του δικτύου ως εργαλείο μελέτης συμβάλλει σημαντικά στην διερεύνηση ιδιοτήτων πολύπλοκων φαινομένων, όπως είναι η σεισμική δραστηριότητα. Οι δύο διαφορετικές προσεγγίσεις για την εισαγωγή των συνδέσεων του σεισμικού δικτύου, δηλαδή η χρονική διαδοχή δύο σεισμών ή κάποιος δείκτης συσχέτισης, αποτυπώνουν το ίδιο αξιόπιστα τη σεισμική δραστηριότητα που παρατηρείται στην εξεταζόμενη περιοχή. Επιπρόσθετα, η ικανότητα του συντελεστή συσταδοποίησης να συλλαμβάνει τις μεταπτώσεις της σεισμικότητας αναδεικνύει το συγκεκριμένο μέτρο δικτύου σε αναπόσπαστο εργαλείο μελέτης σεισμικών δικτύων.

Η κατασκευή της πολυμεταβλητής χρονοσειράς με την βοήθεια νέων μέτρων σεισμικότητας, η χρήση μη γραμμικών μέτρων συσχέτισης για την εισαγωγή της σύνδεσης, η εισαγωγή νέων μέτρων δικτύου για την μελέτη της δομής του σεισμικού δικτύου καθώς και η δημιουργία κατευθυνόμενου δικτύου παρέχουν ένα νέο ισχυρό εργαλείο για την εκτίμηση μελλοντικής σεισμικής δραστηριότητας.

ABSTRACT

Earthquake networks are constructed from strong earthquakes (M≥5.5) that occurred in the Greek area during 1911–2013. The network nodes represent seismic zones and the connections are defined with two different approaches. In the first approach the connections between the seismic zones are given by a correlation index which is computed on multivariate time series of seismic moment. In the second approach, the connections are drawn from the succession of two earthquakes in different seismic zones. The investigation of the complex network at sliding windows throughout the historical record intends to identify the periods showing significant changes in the values of network measures, such as the clustering coefficient. Then, we examine whether the seismic network in each time window is random, i.e. the connections are randomly defined, and therefore compare it to properly formed random networks. The comparison is performed by statistical testing, where the null hypothesis reads that the network is random, and the test statistic is a network measure. The results point to the change of network structure during the period 1980-2000 as evidenced by the rejection of the null hypothesis in many network measures.

ΑΝΑΦΟΡΕΣ

Abe, S. and Suzuki, N. (2004). *Scale-free network of earthquakes*, Europhysics Letters, **65**, 581-586.

Abe, S. and Suzuki, N. (2009). *Main shocks and evolution of complex earthquake networks*, Brazilian Journal of Physics, 39(**2A)**, 428-430.

Albert, R. and Barabasi, A. L. (2002). *Statistical mechanics of complex networks*, Reviews of Modern Physics, vol. 74, 47-97.

Campanharo, A. S. L. O., Sirer, M. I., De Malmgren, R. D., Ramos, F. M., Amaral, L. A. N. (2011). *Duality between time series and networks*, Plos One, **6(8)**, art. no. e23378.

Chorozoglou, D. and Kugiumtzis, D. (2014). *Testing the randomness of causality networks from multivariate time series*, International Symposium on Nonlinear Theory and its Applications, pp. 229-232.

Horvath, S. (2011). *Weighted network analysis*. University of California.

Jimenez, A. , Tiampo, K. F. and Posadas, A. M. (2008). *Small world in a seismic network: the California case*, Nonlinear Processes in Geophysics, **15**, 389-395.

Kanamori, H. and Anderson, L. (1975). *Theoretical basis of some empirical relations in seismology*, Bulletin of the Seismological Society of America, 65(**5)**, 1073-1095.

Kugiumtzis, D. (2002). *Statistically transformed autoregressive process and surrogate data test for nonlinearity*, Physical Review E, **66**, 025201.

Maslov, S. and Sneppen K. (2002). *Specificity and stability in topology of protein networks*. Science, **296**, 910-913.

Newman, M. (2010). *Networks, an introduction*, Oxford University Press.

Papaioannou, Ch. A. and Papazachos, B. C. (2000). *Time-independent and time – dependent seismic hazard in Greece based on seismogenic sources*, Bulletin Seismology Society of America, **90**, 22-33.

Steeples, W. and Steeples, D. (1996). *Far-field aftershocks of the 1906 earthquake*, Bulletin of the Seismological Society of America, 86(**4)**, 921-924.

Schreiber, T. and Schmitz, A. (1996). *Improved surrogate data for nonlinearity tests*. Physical Review Letters, **77(4)**, 635-638.

Tenenbaum, J., Havlin, S. and Stanley, H. E. (2012). *Earthquake networks based on similar activity patterns*, Physical Review E, **86**, 046107.

Χορόζογλου, Δ. and Κουγιουμτζής, Δ. (2014). *Έλεγχος τυχαιότητας δικτύων συσχέτισης από πολυμεταβλητές χρονοσειρές*, Πρακτικά 27ου Πανελληνίου Συνεδρίου Στατιστικής, 301-314.

εργασίες

στα αγγλικά

Greek Statistical Institute
Proceedings of the 28th Panhellenic Statistics Conference (2015), pp. 354–363

The normal theory t and F distributions hold under spherical symmetry

Θεόφιλος Κάκουλλος
Τμήμα Μαθηματικών,
Πανεπιστήμιο Αθηνών

1. Introduction and Summary

The well-known t and F distributions are customarily derived as distributions of ratios involving independent identically distributed (iid) normal $N(0, \sigma^2)$ random variables (rv) $X_1, \ldots, X_n$. As regards the Cauchy distribution, it was shown, Arnold and Brockett (1992), Jones (1999), that spherical (SS) or elliptical symmetry (ES) of $\boldsymbol{X} = (X_1, \ldots, X_n)$ about the origin (see Definition 2.1 below) suffices for the ratio $Y = X_j/X_i$ of any two components X_i, X_j of $\boldsymbol{X}$ to have the standard or general Cauchy distribution. The case of $m \geq 2$ component ratios from $\boldsymbol{X}$ was considered by Cacoullos (2014), showing that their joint distribution is an m-variate Cauchy. Specifically, if $\boldsymbol{X} = (X_1, \ldots, X_n)'$ is SS or ES, then the vector ratio

$$\boldsymbol{Y} = (Y_1, \ldots, Y_{n-1}) = (\frac{X_2}{X_1}, \ldots, \frac{X_n}{X_1}) \sim C_{n-1}(\boldsymbol{\delta}, \Lambda), \tag{1.1}$$

where $C_m(\boldsymbol{\delta}, \Lambda)$ denotes an m-variate Cauchy distribution with location parameter $\boldsymbol{\delta}$ and scale matrix an $m \times m$ non-singular matrix Λ; in (1.1), if $\boldsymbol{X} \sim SS$, $\boldsymbol{\delta} = \boldsymbol{0}$ and $\Lambda = I_{n-1}$ (I_m denoting the identity matrix of order m). Geometrically, $\boldsymbol{Y}$ represents the *polar angle* θ_1 *tangent vector* (see (2.7) below and Cacoullos, 2014).

Note 1.1. For convenience, small letters, instead of capital ones, will be used to denote both angular random variables and corresponding values.

Here the main result (Theorem 3.1 and its extension Theorem 3.2) show that the normal theory t distributions hold under spherical symmetry. Specifically, if $\boldsymbol{X}' = (\boldsymbol{X}'_{(1)}, \boldsymbol{X}'_{(2)}) \sim SS$, then the vector ratio

$$\boldsymbol{t} = \sqrt{n}\boldsymbol{X}_{(1)}/\sqrt{\boldsymbol{X}'_{(2)}\boldsymbol{X}_{(2)}} \sim t_{n,m}, \tag{1.2}$$

$t_{n,m}$ denoting the (standard) m-variate t distribution with n degrees of freedom (df). The m-dimensional Cauchy distribution, $C(\boldsymbol{\delta}, \Lambda)$ is customarily defined (Kotz-Nadarajah, 2007) as a special case ($n = 1$) of the m-dimensional t distribution with n degrees of freedom, location parameter $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)'$ and scale matrix a positive definite $m \times m$ matrix Λ, with density $f_{n,m}(\boldsymbol{x}; \boldsymbol{\delta}, \Lambda)$:

$$f_{n,m}(\boldsymbol{x}; \delta, \Lambda) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{(n\pi)^{\frac{m}{2}}\Gamma\left(\frac{n}{2}\right)|\Lambda|^{1/2}}\left[1+\frac{1}{n}(\boldsymbol{x}-\boldsymbol{\delta})'\Lambda^{-1}(\boldsymbol{x}-\boldsymbol{\delta})\right]^{-\frac{m+n}{2}}, \ \boldsymbol{x} \in \mathbb{R}^m. \quad (1.3)$$

This is denoted by $t_{n,m}(\boldsymbol{\delta}, \Lambda)$ and simply by $t_{n,m}$ when $\boldsymbol{\delta} = \mathbf{0}$ and $\Lambda = I_m$. Thus, $t_{1,m}(\delta, \Lambda) = C_m(\boldsymbol{\delta}, \Lambda)$, and $t_{1,m} = C(\mathbf{0}, I_m) = C_m$, the standard m-variate symmetric Cauchy distribution.

An interesting result, presumably known, yet somewhat obscured and not clearly presented in the statistical literature (Dempster, 1969; Kelker, 1970 p. 428; Mardia et al, 1979) is that, as in the normal case, the F ratio

$$F = \frac{n}{m}\frac{\boldsymbol{X}'_{(1)}\boldsymbol{X}_{(1)}}{\boldsymbol{X}'_{(2)}\boldsymbol{X}_{(2)}} \sim F_{m,n}. \quad (1.4)$$

Here this follows immediately from (1.2), (2.9) and that $\boldsymbol{t}$ itself is SS.

An application of (3.18), (3.19) for testing the hypothesis $\mu = \delta = 0$ under spherical symmetry is briefly discussed in Section 4.

2. Some preliminaries

For our purposes we require the following.

Definition 2.1. *$\boldsymbol{X} = (X_1, \ldots, X_n)'$ (centered at zero) is said to have a SS distribution iff*

$$\boldsymbol{X} \stackrel{d}{=} O\boldsymbol{X} \quad (2.1)$$

for any $n \times n$ orthogonal matrix O, or when $\boldsymbol{X}$ has density $f(\boldsymbol{x})$:

$$f(\boldsymbol{x}) = c_n g(\boldsymbol{x}'\boldsymbol{x}), \quad g : \int_0^\infty t^{n-1} g(t^2) < \infty; \quad (2.2)$$

then we write $\boldsymbol{X} \sim SS(g)$.

(b) Similarly, $\boldsymbol{Y}$ is said to have an elliptically symmetric (ES) distribution with scale matrix an $n \times n$ positive definite matrix Λ, if there exists a positive definite matrix A such that $\boldsymbol{Y} = A\boldsymbol{X}$ with $\boldsymbol{X} \sim SS$, $AA' = \Lambda$; if $\boldsymbol{X} \sim SS(g)$, then $\boldsymbol{Y} \sim ES(g)$, with density

$$f_Y(\boldsymbol{y}) = \frac{c_n}{|\Lambda|^{1/2}} g(\boldsymbol{y}'\Lambda^{-1}\boldsymbol{y}). \quad (2.3)$$

Note 2.1. If $\boldsymbol{U}^{(n)}$ denotes the n-dimensional vector which is uniformly distributed on the n-sphere $\overline{S}_n$, obviously $\boldsymbol{U}^{(n)} \stackrel{d}{=} O\boldsymbol{U}^{(n)} \sim SS$, and though $\boldsymbol{U}^{(n)}$ does not have a density (in the usual sense), it may be regarded as the generator of all $\boldsymbol{X} \sim SS(g)$, in the sense that, Cambanis-Huang-Simmons (1981),

$$\boldsymbol{X} \sim SS \quad \text{iff} \quad \boldsymbol{X} \stackrel{d}{=} R\boldsymbol{U}^{(n)}, \tag{2.4}$$

where $R \geq 0$ is independent of $\boldsymbol{U}^{(n)}$.

A main tool in this paper is the elementary formula for the distribution of a vector ratio

$$\boldsymbol{Z} = \frac{\boldsymbol{X}}{Y}, \quad \boldsymbol{X} = (X_1, \ldots, X_m)' \sim SS, \quad Y \geq 0, \quad P[Y = 0] = 0. \tag{2.5}$$

Then the density of $\boldsymbol{Z}$ is given by

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \int_0^\infty y^m f_{\boldsymbol{X},Y}(y\boldsymbol{z}, y)dy, \tag{2.6}$$

where $f_{\boldsymbol{X},Y}(\boldsymbol{x}, y)$ denotes the joint density of $\boldsymbol{X}$ and Y.

As in Cacoullos (2014), the following basically mathematical result will be also used (cf. (2.4)).

Lemma 2.1. *(Goldman, 1976, Theorem 1; Muirhead, 1982, Theorem 1.5.5). Let $\boldsymbol{X} = (X_1, \ldots, X_m)' \sim SS$ with density function as in (2.2) and $\mathcal{U}^n = (\mathcal{U}_1, \ldots, \mathcal{U}_n)$ be uniformly distributed on the unit sphere $\overline{S}_m$. Then there exists a set of spherical rv's $R, \theta_1, \ldots, \theta_{m-1}$,*

$$\begin{aligned}
X_1 &= R\mathcal{U}_1 = R\cos\theta_1, \quad 0 \leq \theta_1 \leq \pi, \\
X_k &= R\mathcal{U}_k = R\Big(\prod_{i=1}^{k-1} \sin\theta_i\Big)\cos\theta_k, \quad 0 \leq \theta_{k-1} \leq 2\pi, \quad 2 \leq k \leq m-1, \\
X_m &= R\mathcal{U}_m = R\prod_{i=1}^{m-1} \sin\theta_i, \quad 0 \leq \theta_{m-1} < 2\pi,
\end{aligned} \tag{2.7}$$

such that $R, \theta_1, \ldots, \theta_{m-1}$ are independently distributed with density functions

$$f_R(u) = \frac{2c_m\pi^{\frac{m}{2}}}{\Gamma\Big(\frac{m}{2}\Big)}u^{m-1}g(u^2), \quad \frac{1}{c_m} = \frac{2\pi^{\frac{m}{2}}}{\Gamma\Big(\frac{m}{2}\Big)}\int_0^\infty u^{m-1}g(u^2)du, \tag{2.8}$$

$$f_{R^2}(u) = \frac{c_m\pi^{\frac{m}{2}}}{\Gamma\Big(\frac{m}{2}\Big)}u^{\frac{m}{2}-1}g(u), \tag{2.9}$$

$$f_k(\theta_k) = \frac{1}{B\left(\frac{1}{2}, \frac{n-k}{2}\right)} \sin^{m-k-1}\theta_k, \quad 0 \le \theta_k \le \pi, \quad 1 \le k \le m-2, \tag{2.10}$$

$$f_{m-1}(\theta_{m-1}) = \frac{1}{2\pi}, \quad 0 \le \theta_{m-1} < 2\pi. \tag{2.11}$$

Conversely, if $R, \theta_1, \ldots, \theta_{m-1}$ are independently distributed with densities (2.8)-(2.11), then $\boldsymbol{X} = (X_1, \ldots, X_m)'$, with $X_1, \ldots, X_m$ defined by (2.7), has density given by (1.2).

Remark 2.1. If sines and cosines in (2.7) are interchanged, i.e., $X_1 = R\cos\theta_1$ etc., then in (2.10) sin should be replaced by cos.

Definition 2.2. *(a) Let $\boldsymbol{X} = (X_1, \ldots, X_m)' \sim SS(g)$, with density (2.2). Then*

$$R^2 = \boldsymbol{X}'\boldsymbol{X} = X_1^2 + \cdots + X_m^2,$$

is said to have the generalized $g\chi_m^2$ distribution, with m df, and R the $g\chi_m$, to be denoted by $R^2 \sim g\chi_m^2$ and $R \sim g\chi_m$. The density of $g\chi_m$ is given by (2.8), of $g\chi_m^2$ by (2.9).

Note 2.2. The familiar χ_m^2 corresponds to $\boldsymbol{X} \sim N(\boldsymbol{0}, I_m)$, i.e.,

$$\chi_m^2 \equiv g\chi_m^2, \quad g(t) = g_0(t) = e^{-\frac{1}{2}t}, \quad f_{\boldsymbol{X}}(\boldsymbol{x}) = c_0 e^{-\frac{1}{2}\boldsymbol{x}'\boldsymbol{x}}, \quad c_0 = (2\pi)^{-\frac{m}{2}}, \tag{2.12}$$

Also, since the components of $\boldsymbol{X} \sim SS$ are uncorrelated but not independent unless they are normal, we need

Definition 2.3. *Let $\boldsymbol{X}' = (\boldsymbol{X}_{(1)}', \boldsymbol{X}_{(2)}') \sim SS(g)$, $R_1^2 = \boldsymbol{X}_1'\boldsymbol{X}_1 \sim g\chi_m^2$, $R_2^2 = \boldsymbol{X}_{(2)}'\boldsymbol{X}_{(2)} \sim g\chi_n^2$, where $\boldsymbol{X}_{(1)}$ and $\boldsymbol{X}_{(2)}$ are disjoint subvectors of $\boldsymbol{X}$:*

$$\boldsymbol{X}_{(1)} = (X_1, \ldots, X_m)' \sim SS(g), \quad \boldsymbol{X}_{(2)} = (X_{m+1}, \ldots, X_{m+n})' \sim SS(g). \tag{2.13}$$

*Such $g\chi_m^2$ and $g\chi_n^2$ are referred to as **disjoint** or **orthogonal** $g\chi^2$.*

Remark 2.2. It may be worth noting that the well-known χ^2 distribution does not follow only as distribution of a sum of squares of iid $N(0,1)$. Indeed, if $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_m)' \sim SS(g)$ where

$$g(t) = g_\alpha(t) = t^\alpha\, e^{-\lambda t}, \quad 2\alpha + m > 0$$

that is, $\boldsymbol{\xi}$ is a "gamma-type" SS distribution, then it readily follows from (2.9) that

$$f_{g_\alpha\chi_m^2}(u) = c u^{\frac{m}{2}-1} g_\alpha(u) = c \cdot u^{\frac{m}{2}+\alpha-1} e^{-\lambda u},$$

so that taking $\alpha = \frac{k}{2}$ $(k = 1, 2, \dots)$, $\lambda = \frac{1}{2}$, we have

$$\boldsymbol{\xi}'\boldsymbol{\xi} \sim g_\alpha \chi^2_m = \chi^2_{m+k}.$$

Thus, in general, if $\boldsymbol{\xi} \sim SS(g_\alpha)$ and $Q = \boldsymbol{\xi}' A \boldsymbol{\xi}$ with A idempotent $(A^2 = A)$ and $\text{rank}(A) = r$, Q may be χ^2_n distributed with $n \neq r$. For more details we refer to Cacoullos and Khatri (1991), including the following characterization of normality: Let $\boldsymbol{\xi} = (\xi_1, \dots, \xi_m)' \sim SS$, $Q = \boldsymbol{\xi}' A \boldsymbol{\xi}$ with $A = A'$. Then $Q \sim \chi^2_r$ with $r = \text{rank}(A)$ iff $\boldsymbol{\xi} \sim N(\mathbf{0}, I_m)$. Further discussion is beyond the scope of the present investigation.

3. Main results

The main result (Theorem 3.1) shows that the familiar t ratios, which are t distributed under the assumption of independent $N(0,1)$ (or $N(0, \sigma^2)$) $X_1, \dots, X_m$, $X_{m+1}, \dots, X_{m+n}$, follow the same t distributions under the weaker assumption of spherical symmetry.

Note 3.1. For convenience, in the following derivations normalizing and other constant multipliers will be absorbed into a single constant c, the proper normalizing constant required by the resulting distribution.

Theorem 3.1. *Let* $\boldsymbol{X}' = (\boldsymbol{X}'_{(1)}, \boldsymbol{X}'_{(2)}) \sim SS(g)$, $\boldsymbol{X}_{(1)} = (X_1, \dots, X_m)'$, $\boldsymbol{X}_{(2)} = (X_{m+1}, \dots, X_{m+n})'$ *and* $R_2^2 = X_{m+1}^2 + \cdots + X_{m+n}^2 \sim g\chi^2_n$ *(Definition 2.2). Then the vector* $\boldsymbol{t}$ *ratio*

$$\boldsymbol{t}^* = \frac{\sqrt{n}\boldsymbol{X}_{(1)}}{R_2} \sim t_{n,m}, \tag{3.1}$$

that is, $\boldsymbol{t}^*$ *follows the* m*-variate* t *distribution with* n *df, as under normality of* $X \sim N(\mathbf{0}, I_{m+n})$.

Proof. Letting

$$\boldsymbol{Z} = \frac{\boldsymbol{t}^*}{\sqrt{n}} = \frac{\boldsymbol{X}_{(1)}}{R_2}, \quad R_2 = Y, \tag{3.2}$$

we have by (2.6)

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \int_0^\infty y^m f_{\boldsymbol{X}_{(1)}, Y}(y\boldsymbol{z}, y) dy. \tag{3.3}$$

By (2.2) and (2.8)

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = c_{m+n} g(\boldsymbol{x}'\boldsymbol{x}) = c_{m+n} g(r_1^2 + r_2^2), \tag{3.4}$$

and

$$f_Y(y) = \frac{2c_n}{\Gamma\left(\frac{n}{2}\right)} y^{n-1} g(y^2); \tag{3.5}$$

hence (3.3) yields

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = c\int_0^\infty y^{m+n-1} g(y^2 \boldsymbol{z}'\boldsymbol{z} + y^2)dy = c\int_0^\infty y^{m+n-1} g((1+\boldsymbol{z}'\boldsymbol{z})y^2)dy, \tag{3.6}$$

which, by (2.8), giving the proper value of c, proves (3.1).

Alternative Proof. By (2.7) (see also (2.4)), $\boldsymbol{X} = (X_1, \ldots, X_m, X_{m+1}, \ldots, X_{m+n})'$ can be written as

$$\boldsymbol{X}' = (\boldsymbol{X}'_{(1)}, \boldsymbol{X}'_{(2)}) = R\ \ (\boldsymbol{\mathcal{U}}'_{(1)}, \boldsymbol{\mathcal{U}}'_{(2)}) = R(\mathcal{U}_1, \ldots, \mathcal{U}_m, \mathcal{U}_{m+1}, \ldots, \mathcal{U}_{m+n}), \tag{3.7}$$

where $R = \sqrt{\boldsymbol{X}'\boldsymbol{X}}$ is independent of $(\boldsymbol{\mathcal{U}}'_{(1)}, \boldsymbol{\mathcal{U}}'_{(2)})$, uniformly distributed on $\overline{S}_{m+n}$. Moreover, it is easily verified that

$$R_2 = R|\boldsymbol{\mathcal{U}}_{(2)}| = R\sin\theta_1 \ldots \sin\theta_m, \tag{3.8}$$

so that (3.2) takes the form

$$\boldsymbol{Z} = \frac{R\boldsymbol{\mathcal{U}}_{(1)}}{R|\boldsymbol{\mathcal{U}}_{(2)}|} = \frac{\boldsymbol{\mathcal{U}}_{(1)}}{|\boldsymbol{\mathcal{U}}_{(2)}|} = (Z_1, \ldots, Z_m) = \frac{\boldsymbol{\mathcal{U}}_{(1)}}{\prod_{i=1}^{m} \sin\theta_i}, \tag{3.9}$$

where

$$Z_k = \frac{X_k}{R_2} = \frac{\left(\prod_{i=1}^{k-1} \sin\theta_i\right)\cos\theta_k}{\prod_{i=1}^{m} \sin\theta_i} = \frac{\cot\theta_k}{\prod_{k+1}^{m} \sin\theta_i}, \quad 1 \le k \le m-1,\ Z_m = \cot\theta_m. \tag{3.10}$$

Hence, setting $\boldsymbol{\theta}_{(1)} = (\theta_1, \ldots, \theta_m)'$ and using the densities of $\theta_1, \ldots, \theta_m$ from (2.10), we have the density $f_{\boldsymbol{Z}}(\boldsymbol{z})$ of $\boldsymbol{Z}$,

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \frac{1}{|J(\boldsymbol{z}, \boldsymbol{\theta}_{(1)})|} f(\boldsymbol{\theta}_1(\boldsymbol{z})) = \frac{1}{|J(\boldsymbol{z}, \boldsymbol{\theta}_{(1)}|} \prod_{k=1}^{m} \sin^{m+n-k-1}\theta_k \tag{3.11}$$

where the Jacobian $J(\boldsymbol{z}, \boldsymbol{\theta}_{(1)})$ of (3.10) is found to be

$$[J(\boldsymbol{z}, \theta_{(1)})]^{-1} = \sin\theta_1^2 \sin^3\theta_2 \cdots \sin^{m+1}\theta_m.$$

Hence (3.11) gives

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = c(\sin^2\theta_1 \cdots \sin^2\theta_m)^{\frac{m+n}{2}}, \tag{3.12}$$

from which, noting that

$$1 + z_1^2 + \cdots + z_m^2 = (\sin^2\theta_1 \cdots \sin^2\theta_m)^{-1}, \tag{3.13}$$

we have

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = c(1 + z_1^2 + \cdots + z_m^2)^{-\frac{m+n}{2}}; \tag{3.14}$$

therefore (3.1) readily follows.

We now state certain results, corollaries of Theorem 3.1

Taking $\boldsymbol{X}_{(1)} \sim ES$ we obtain Theorem 3.2.

Theorem 3.2. *If $\boldsymbol{X}_{(1)} \sim ES(g)$ with density (cf. (2.3))*

$$f_{\boldsymbol{X}_{(1)}}(\boldsymbol{x}) = \frac{c_m}{|\Lambda|^{1/2}} g((\boldsymbol{x}-\boldsymbol{\delta})'\Lambda^{-1}(\boldsymbol{x}-\boldsymbol{\delta})) \tag{3.15}$$

and R_2 as in (3.1), then

$$\boldsymbol{t}^*(\boldsymbol{\delta}, \Lambda) = \frac{\sqrt{n}\boldsymbol{X}_{(1)}}{R_2} \sim t_{n,m}(\boldsymbol{\delta}, \Lambda). \tag{3.16}$$

$t_{n,m}(\boldsymbol{\delta}, \Lambda)$ as defined by (1.3).

Proof. As in Theorem 3.1 simply, replacing $\boldsymbol{z}'\boldsymbol{z}$ by $(\boldsymbol{z}-\boldsymbol{\delta})'\Lambda^{-1}(\boldsymbol{z}-\boldsymbol{\delta})$.

Another important result from Theorem 3.1 is as stated by (1.4); this is immediate in view of (2.9) and that $\boldsymbol{t}^*$ is SS. A more involved proof is found in Kelker (1970, p. 428), showing first that $\dfrac{R_1^2}{\boldsymbol{X}'\boldsymbol{X}} \sim \beta\Big(\dfrac{m}{2}, \dfrac{n}{2}\Big)$.

Taking $m = 1$ in (3.1), i.e., $\boldsymbol{X}_{(1)} = X_1 = Y_1$, and $\boldsymbol{X}_{(2)} = (Y_2, \ldots, Y_n)'$ we conclude

Corollary 3.1. *Let $\boldsymbol{Y} \equiv (Y_1, \ldots, Y_n)' \sim SS$. Then the familiar t ratio*

$$t = \frac{\sqrt{n-1}Y_1}{\sqrt{Y_2^2 + \cdots + Y_n^2}} \sim t_{n-1} \tag{3.17}$$

i.e., the same t_{n-1} distribution as under normality of $\boldsymbol{X} \sim N(\boldsymbol{0}, I_n)$.

This was also shown (Cacoullos, 2014, Theorem 3.5) directly from the distribution of $\cot\theta_k$, namely, that

$$t_{n-k}^* = \sqrt{n-k}\cot\theta_k \sim t_{n-k}, \quad 1 \le k \le n-1, \tag{3.18}$$

combined with the inverse transformation of (2.7):

$$\theta_k = \cot^{-1} \frac{Y_k}{\sqrt{Y_{k+1}^2 + \cdots + Y_n^2}}, \quad 1 \le k \le n-2. \tag{3.19}$$

Moreover, the t_{n-k}^* are also independent (Cacoullos 2014, Theorem 3.6), since the θ_k are independent, an important result not at all obvious even under the assumption of iid $N(0, \sigma^2)$ $Y_1, \ldots, Y_n$.

Conversely, using (2.7), it is easily verified that, as expected,

$$\frac{Y_k}{\sqrt{Y_{k+1}^2 + \cdots + Y_n^2}} = \cot \theta_k; \tag{3.20}$$

hence (3.17) – (3.18).

Finally applying Theorem 3.1 for $m \ge 2$, and $n = 1$, we obtain

Corollary 3.2. *Let* $\boldsymbol{X} = (X_1, \ldots, X_{m+1})' \sim SS$. *Then the vector ratio*

$$\boldsymbol{Y} = \frac{\boldsymbol{X}_{(1)}}{X_{m+1}} \sim C(\boldsymbol{0}, I_m). \tag{3.21}$$

(Cacoullos, 2014, Theorems 3.1-3.3).

Remark 3.1. Actually, all the preceding results readily follow by using (2.4), instead of (2.6), since the distributions of the ratios Z (3.9), t_{n-k}^* (see (3.18)-(3.20)), and F (see (2.4)) are independent of R, and hence are the same under spherical normality (cf. proofs of Theorems 3.1 and 3.2 in Cacoullos, 2014).

4. Testing means under spherical symmetry

Testing hypotheses about population means is associated with t tests, in general under the assumption of a random sample of iid normal $N(\mu, \sigma^2)$ rv's $X_1, \ldots, X_n$, so that for testing $H_0 : \mu = 0$ (or that the location parameter $\delta = 0$) one uses the statistic $t' = \sqrt{n}\overline{X}/s \sim t_{n-1}$.

It is noted, however, that by rotational symmetry (see (2.1), t of (3.17)) can be transformed to take the form of t', so that $t \stackrel{d}{=} t' \sim t_{n-1}$ (cf. Muirhead, 1982, pp. 38-39). That $t' \sim t_{n-1}$ was pointed out by Efron (1969), who was primarily concerned with the behavior of $S_n = \sum_{i=1}^{n} X_i / \sum_{i=1}^{n} X_i^2$ under "rotational symmetry" (axes permution), a weaker condition than spherical symmetry.

Furthermore, in view of (3.18), each of the $n-1$ t_{n-k}^* provides a statistic for testing $H_0 : \mu = \delta = 0$, and more importantly the t_{n-k}^* are independent if $\boldsymbol{X} = (X_1, \ldots, X_n) \sim SS$. Equivalently to each t_{n-k}^* there corresponds a typical t'_{n-k} statistic,

$$t'_{n-k} = \sqrt{n-k}\frac{\overline{X}_k}{s_k} \sim t_{n-k},$$

where

$$\overline{X}_k = \frac{1}{k}\sum_{i=1}^{k} X_i, \quad s_k^2 = \frac{1}{k-1}\sum_{i=1}^{k}(X_i - \overline{X}_k)^2, \quad (2 \le k \le n-1).$$

Remark 4.1. It may be of some interest to examine the behavior (e.g. p-values, t-values, etc) of t^*_{n-1} (or t') based on all n X_i, (i.e., on $\boldsymbol{X} = (X_1, \ldots, X_n)' \sim SS$) in relation to tests based on $k < n$ X_i $(k = 2, \ldots, n-1)$, such as the t^*_{n-k}.

Remark 4.2. It should be also noted that, by symmetry, any of the $\binom{n}{k}$ subvectors-combinations of $X_1, X_2, \ldots, X_n$ can be chosen for testing H_0.

This, however, is beyond the distributional scope of this note and will not be further examined here.

Addendum

In March 2015 I received some comments from two colleagues, indicating that, like (1.4), the main result (1.2) was presumambly known, though somewhat obscured, not having been clearly stated in the statistical literature. Indeed in Anderson and Fang (1990) and Anderson (2003) one can find several statistics in multivariate analysis which, under spherical symmetry, have the same distributions under normality. These results rest on stochastic representation (2.4), since such statistics (ratios), like t and F, turn out to be independent of R (c.f. Lemma 2.1, (3.9) in the Alternative Proof here, and Theorems 3.1 and 3.2 in Cacoullos 2014), so that the resulting distributions are the same as under normality. Equivalently, using (2.6) yields ratios which are functions only of θ's, spherical angular coordinates, which are independent of R.

Acknowledgement. Thanks are due to Apostolos Batsidis and Barry Arnold, who brought to my attention the two references included in the Addendum.

ΠΕΡΙΛΗΨΗ

Έστω ότι η $\boldsymbol{X}' = (\boldsymbol{X}'_{(1)}, \boldsymbol{X}'_{(2)})$ ακολουθεί σφαιρικά συμμετρική $(\Sigma\Sigma)$ κατανομή, με πυκνότητα $f_{\boldsymbol{X}}(\boldsymbol{x}) = cg(\boldsymbol{x}'\boldsymbol{x})$. Υπό κανονικότητα $(g(t) = g_0(t) = e^{-t/2})$ η $\boldsymbol{X}_{(1)} = (X_1, \ldots, X_m)'$ είναι ανεξάρτητη της $\boldsymbol{X}_{(2)} = (X_{m+1}, \ldots, X_{m+n})'$, άρα και της $\boldsymbol{X}'_{(2)}\boldsymbol{X}_{(2)} \sim g_0\chi^2_n$, και ως γνωστό το λογοδιάνυσμα $t = \sqrt{n}\boldsymbol{X}_{(1)}/g\chi_n$ ακολουθεί την m-διάστατη t κατανομή με n βαθμούς ελευθερίας. Υπό σφαιρική συμμετρία, αποδεικνύεται ότι η $t = \sqrt{n}\boldsymbol{X}_{(1)}/g\chi_n$ έχει την ίδια t κατανομή. Σημειωτέον ότι η γενικευμένη $g\chi_n$ δεν είναι ανεξάρτητη της $\boldsymbol{X}_{(1)}$ εκτός αν η $g(t) = g_0(t)$. Η πολυδιάστατη Cauchy προκύπτει, Cacoullos (2014), ως μερική περίπτωση της $t(n = 1)$.

Πολύ ενδιαφέρον είναι ότι από $(X_1, X_2, \ldots, X_\nu) \sim \Sigma\Sigma$ μπορούν να γίνουν $\nu - 1$ ανεξάρτητοι t έλεγχοι της υπόθεσης $\mu = 0$. Επιπλέον ο λόγος

$$F = \frac{n}{m} \frac{g\chi_m^2}{g\chi_n^2} \sim F_{m,n},$$

όπου $\boldsymbol{X}'_{(1)}\boldsymbol{X}_{(1)} \sim g\chi_m^2$. Συνεπώς ο γνωστός ANOVA F Table ισχύει και για σφαιρικές κατανομές, γεγονός πολύ σημαντικό στις στατιστικές εφαρμογές.

REFERENCES

Anderson, T. W., and Fang, K. T., "Theory and Applications of Elliptically Contoured and Related Distributions", Technical report No 24, Stanford University (1990):.

Anderson, T. W., (2003), An Introduction to Multivariate Statistical Analysis, 3rd edition, *Wiley, New York.*

Arnold, B. C., Brockett, P. L., (1992), "On distributions whose component ratios are Cauchy" *Amer. Statist.*, **46**, 25-26.

Cacoullos, T., (2014), "Polar angle tangent vectors follow Cauchy distributions under spherical symmetry" *J. of Multivariate Analysis*, **128**, 147-153.

Cacoullos, T. and Khatri, C. G., (1991), "Correcting remarks on Characterization of normality within the class of elliptical contoured distributions" *Statist. & Probability Letters*, **11**, 551-552.

Cambanis, S., Huang, S., Simons, G., (1981), "On the theory of elliptically contoured distributions" *J. Multivariate Anal.*, **11**, 368-385.

Dempster, A., (1969), *Elements of Continuous Multivariate Analysis*, Addison Wesley.

Efron, B., (1969), "Student's t-test under symmetry conditions" *J. Amer. Statist. Assoc.*, **64**, 1278-1302.

Goldman, J., (1976), "Detection in the presence of spherically symmetric random vectors" *IEEE Trans. Info. Theory*, **22**, 52-59.

Jones, N. C., (2008), " The distribution of the ratio X/Y for all centred elliptically symmetric distributions" *J. Multivariate Anal.*, **99**, 572-573.

Kelker, D., (1970), "Distribution theory of spherical distributions and a location-scale parameter generalization" *Sankyā*, A **32**, 419-430.

Kotz, S., Nadarajah, S., (2007), *Multivariate t distribution and its Applications*, Cambridge University Press.

Mardia, K. V., Kent, J. T. & Bibby, J. M., (1979), *Multivariate Analysis*, Academic Press.

Muirhead, R. J., (1982), *Aspects of Multivariate Statistical Analysis*, Wiley, New York.

Greek Statistical Institute
Proceedings of the 28th Panhellenic Statistics Conference (2015), pp. 364−370

AN EULER STOCHASTIC PROCESS

Ch. A. Charalambides
University of Athens
ccharal@math.uoa.gr

ABSTRACT

A stochastic model, which is developing in time and successes (event A) occur at continuous points, is considered. An Euler stochastic process is defined on this model and its distribution is derived. Also, the distribution of the waiting time until the occurrence of a fixed number of successes is deduced as a q-Erlang distribution.

Keywords: Discrete q-Distribution; q-Erlang distribution; q-Exponential distribution.

1. INTRODUCTION

The Heine and Euler distributions, which constitute q-analogues of the Poisson distribution, were derived by Benkherouf and Bather (1988) as feasible priors in a simple Bayesian model for oil exploration. Also, they noticed the expression of the Heine distribution as an infinite convolution of zero-one Bernoulli distributions and Kemp (1992a) discussed the expression of the Euler distribution as an infinite convolution of geometric distributions. Further, Kemp and Newton (1990) derived the Heine distribution as a limiting distribution of a q-Binomial distribution. Also, Kemp (1992b) derived the Heine and Euler distributions as stationary distributions of Markov chains.

In the present paper, we consider a stochastic model that is developing in time or space, in which a success or a failure (events A or A') may occur at continuous points. Then, an Euler process X_t, $t \geq 0$, with dependent and homogeneous increments, which constitutes a q-analogue of the Poisson process, is introduced by considering a q-partition of the time interval $(0, t]$. Further, using the condition on the transition probabilities in small time intervals, a system of q-differential equations is derived. Solving it, the probability function of the Euler process is obtained. Also, the distribution of the waiting time W_n until the occurrence of the nth success, which is connected to the distribution of the Euler process X_t, $t \geq 0$, is obtained as a q-analogue of the Erlang distribution.

2. q-FACTORIALS AND q-EXPONENTIAL FUNCTIONS

Let x and q be real numbers, with $q \neq 1$, and k be an integer. The number

$$[x]_q = \frac{1-q^x}{1-q},$$

is called *q-number* and in particular $[k]_q$ is called *q-integer*. The base (parameter) q, in the theory of discrete q-distributions, varies in the interval $0 < q < 1$ or in the interval $1 < q < \infty$.

The kth order factorial of the q-number $[x]_q$, which is defined by

$$[x]_{k,q} = [x]_q[x-1]_q \cdots [x-k+1]_q, \quad k = 1, 2, \ldots,$$

is called *q-factorial of x of order k*. In particular

$$[k]_q! = [1]_q[2]_q \cdots [k]_q, \quad k = 1, 2, \ldots,$$

is called *q-factorial*.

In general, the transition from a formula to its q-analogue is not unique. Thus, in addition to the *q-exponential function*

$$E_q(t) = \prod_{i=1}^{\infty}(1 + (1-q)q^{i-1}t) = \sum_{x=0}^{\infty} q^{\binom{x}{2}} \frac{t^x}{[x]_q!}, \quad -\infty < t < \infty,$$

which will be used in sequel, there is another *q-exponential function*

$$e_q(t) = \prod_{i=1}^{\infty}(1 - (1-q)q^{i-1}t)^{-1} = \sum_{x=0}^{\infty} \frac{t^x}{[x]_q!}, \quad |t| < 1/(1-q),$$

with $E_q(t) = e_q(-t) = 1$.

The *q-derivative operator*, denoted by $\mathcal{D}_q = d_q/d_q t$, is defined by

$$\mathcal{D}_q f(t) = \frac{d_q f(t)}{d_q t} = \frac{f(t) - f(qt)}{(1-q)t},$$

so that $\mathcal{D}_q 1 = 0$. The higher order q-derivatives are defined recursively by

$$\mathcal{D}_q^k f(t) = \mathcal{D}_q\big(\mathcal{D}_q^{k-1} f(t)\big), \quad k = 2, 3, \ldots .$$

The q-derivatives of the q-exponential functions can be readily deduced as

$$\mathcal{D}_q e_q(t) = e_q(t), \quad \mathcal{D}_q E_q(t) = E_q(qt).$$

The q-derivative of a product of two functions is given by the *q-Leibnitz formula* as

$$\mathcal{D}_q\big(f(t)g(t)\big) = g(t)\mathcal{D}_q f(t) + f(qt)\mathcal{D}_q g(t).$$

For more details on these and other related q-series and q-functions, the interested reader is advised to consult the book of Gasper and Rahman (2004) on the Basic hypergeometric series.

3. EULER PROCESS AND q-ERLANG DISTRIBUTION

In the stochastic model of a sequence of independent Bernoulli trials, the event of success, $A = \{s\}$, may occur at discrete points (trials) of its development. The possibility of an event to occur at continuous (time or space) points of the development of a stochastic model, is of great theoretical and practical interest. In this respect, let us consider a stochastic model that is developing in time or space, in which successes (or failures) may occur at continuous points. Further, let X_t be the number of successes (occurrences of event A) in the interval $(0,\, t]$. The family of random variables X_t, $t \geq 0$, is called *stochastic process*. A nonnegative integer valued stochastic process X_t, $t \geq 0$, with independent and homogeneous increments is called *Poisson process*, if in a small time interval either a success, $A = \{s\}$, occurs, with probability analogous to the length of the interval, or a failure, $A' = \{f\}$. A q-analogue of the Poisson process is introduced in the following definition, by considering the geometrically decreasing sequence of time differences

$$\delta t_i = (1-q)q^{i-1}t, \quad i = 1, 2, \dots, \quad 0 < q < 1,$$

with $\sum_{i=1}^{\infty} \delta t_i = t$, to partition the time interval $(0,\, t]$.

Definition 1. *Consider a stochastic model that is developing in time or space and let X_t, $t \geq 0$, be the number of successes (occurrences of event A) in the interval $(0,\, t]$. Assume that X_t, $t \geq 0$, is a stochastic process, with dependent and homogeneous increments, which starts at $t = 0$ from state (epoch) 0, $P(X_0 = 0) = 1$, and, in the small time intervals $(q^i t,\, q^{i-1}t]$, of lengths $\delta t_i = (1-q)q^{i-1}t$, $i = 1, 2, \dots,$ satisfies the condition*

$$p_{j,k}(\delta t_i) = P(X_{q^{i-1}t} = k | X_{q^i t} = j) = \begin{cases} 1 - \lambda(1-q)q^{i-j-1}t, & k = j, \\ \lambda(1-q)q^{i-j-1}t, & k = j+1, \\ 0, & k > j+1, \end{cases} \tag{1}$$

for $j = 0, 1, \dots, i-1$ and $i = 1, 2, \dots,$ with $0 < \lambda t < 1/(1-q)$ and $0 < q < 1$. Then, X_t, $t \geq 0$, is called Euler process, with parameters λ and q.

It is worth noticing that, in contrast to a Poisson process, an Euler process does not have independent increments. Also, the condition of the occurrence of at most one success in a small time interval is expressed in terms of a series of small time intervals of varying (q-decreasing) lengths.

Theorem 1. *The probability function of the Euler process X_t, $t \geq 0$, with parameters λ and q, is given by*

$$p_x(t) = P(X_t = x) = E_q(-\lambda t)\frac{(\lambda t)^x}{[x]_q!}, \quad x = 0, 1, \dots, \tag{2}$$

where $0 < \lambda t < 1/(1-q)$, $0 < q < 1$ and $E_q(u) = \prod_{i=1}^{\infty}(1 + u(1-q)q^{i-1})$ is a q-exponential function.

Proof. The probability function $p_x(q^{i-1}t)$ of the Euler process, by the total probability theorem,

$$p_x(q^{i-1}t) = p_x(q^i t + \delta t_i) = \sum_{k=0}^{x} p_{x-k}(q^i t) p_{x-k,x}(\delta t_i), \quad x = 0, 1, \ldots, i-1,$$

and condition (1), satisfies the system of equations

$$p_0(q^{i-1}t) = (1 - \lambda(1-q)q^{i-1}t)p_0(q^i t),$$

$$p_x(q^{i-1}t) = (1-\lambda(1-q)q^{i-x-1}t)p_x(q^i t)+\lambda(1-q)q^{i-x}tp_{x-1}(q^i t), \;\; x = 1, 2, \ldots, i-1.$$

Setting $u = q^{i-1}t$, this system of equations may be rewritten as

$$p_0(u) = (1 - \lambda(1-q)u)p_0(qu),$$

$$p_x(u) = (1 - \lambda(1-q)q^{-x}u)p_x(qu) + \lambda(1-q)q^{-(x-1)}up_{x-1}(qu), \quad x = 1, 2, \ldots,$$

or as

$$\frac{p_0(u) - p_0(qu)}{(1-q)u} = -\lambda p_0(qu),$$

$$\frac{p_x(u) - p_x(qu)}{(1-q)u} = -\lambda q^{-x} p_x(qu) + \lambda q^{-(x-1)} p_{x-1}(qu), \quad x = 1, 2, \ldots .$$

Introducing the q-derivative operator $\mathcal{D}_q$, with respect to u, we deduce the system of q-differential equations

$$\mathcal{D}_q p_0(u) = -\lambda p_0(qu),$$

$$\mathcal{D}_q p_x(u) = -\lambda q^{-x} p_x(qu) + \lambda q^{-(x-1)} p_{x-1}(qu), \quad x = 1, 2, \ldots .$$

Introducing the function $g(u)$ by

$$p_x(u) = g(u)\frac{(\lambda u)^x}{[x]_q!}, \quad x = 0, 1, \ldots, \tag{3}$$

and since

$$\mathcal{D}_q p_x(u) = \frac{(\lambda u)^x}{[x]_q!}\mathcal{D}_q g(u) + \lambda\frac{(\lambda u)^{x-1}}{[x-1]_q!}g(qu),$$

the system of q-differential equations reduces to the q-differential equation

$$\mathcal{D}_q g(u) = -\lambda g(qu),$$

with the initial condition $g(0) = p_0(0) = 1$. Its solution is readily obtained as $g(u) = E_q(-\lambda u)$, and so, by (3), expression (2) is established, with u instead of t. □

In an Euler process, the distribution of the waiting time until the occurrence of a fixed number of successes is connected to the distribution of the number of successes in a fixed time interval. In this respect, the following definition is introduced.

Definition 2. *Consider a stochastic model that is developing in time and successes occur according to an Euler process. Let W_n, be the waiting time until the occurrence of the nth success (event $A = \{s\}$) in the interval $(0, t]$. The distribution of W_n is called q-Erlang distribution of the second kind, with parameters n, λ and q. In particular, the distribution of the waiting time until the occurrence of the first success, $W \equiv W_1$, is called q-Exponential distribution of the second kind, with parameters λ and q.*

The distribution function, together with the q-density function and q-moments of the q-Erlang distribution of the second kind are derived in the following theorem.

Theorem 2. *The distribution function $F_n(w) = P(W_n \leq w)$, $-\infty < w < \infty$, of the q-Erlang distribution of the second kind, with parameters n, λ and q, is given by*

$$F_n(w) = 1 - \sum_{x=0}^{n-1} E_q(-\lambda w) \frac{(\lambda w)^x}{[x]_q!}, \quad 0 < w < \infty, \tag{4}$$

and $F_n(w) = 0$, $-\infty < w < 0$, where n is a positive integer, $0 < \lambda < \infty$ and $0 < q < 1$. Its q-density function $f_n(w) = d_q F_n(w)/d_q w$ is given by

$$f_n(w) = \frac{\lambda^n}{[n-1]_q!} w^{n-1} E_q(-\lambda q w), \quad 0 < w < \infty. \tag{5}$$

Also, its jth q-moment is given by

$$\mu'_{j,q} = E(W_n^j) = \frac{[n+j-1]_{j,q}}{\lambda^j}, \quad j = 1, 2, \ldots . \tag{6}$$

Proof. The event $\{W_n > w\}$, that the nth success occurs after time w, is equivalent to the event $\{X_w < n\}$, that the number of successes up to time w is less than n and so

$$P(W_n > w) = P(X_w < n) = \sum_{x=0}^{n-1} P(X_w = x).$$

Thus, the distribution function of the random variable W_n, on using the relation $F_n(w) = P(W_n \leq w) = 1 - P(W_n > w)$ and expression (2), is deduced as (4).

The q-density function of W_n, on taking the q-derivative of (4) and using the q-Leibnitz formula, with $f(w) = E_q(-\lambda w)$ and $g(w) = \sum_{x=0}^{n-1}(\lambda w)^x/[x]_q!$, is obtained in the form

$$f_n(w) = \lambda E_q(-\lambda q w)\sum_{x=0}^{n-1}\frac{(\lambda w)^x}{[x]_q!} - \lambda E_q(-\lambda q w)\sum_{x=1}^{n-1}\frac{(\lambda w)^{x-1}}{[x-1]_q!},$$

which reduces to (5). Note that, using the relation

$$\int_0^\infty u^{n-1}E_q(-qu)d_q u = [n-1]_q!, \tag{7}$$

which may be derived by integration by parts, it follows that

$$\int_0^\infty f_n(w)d_q w = 1,$$

which conforms with the definition of a q-density function.

The jth q-moment of W_n,

$$\mu'_{j,q} = E(W_n^j) = \frac{\lambda^n}{[n-1]_q!}\int_0^\infty w^{n+j-1}E_q(-\lambda q w)d_q w,$$

using the transformation $u = \lambda w$ and expression (7), is obtained as

$$\mu'_{j,q} = \frac{\lambda^n}{[n-1]_q!\lambda^{n+j}}\int_0^\infty u^{n+j-1}E_q(-qu)d_q u = \frac{[n+j-1]_q!}{[n-1]_q!\lambda^j}.$$

Since $[n+j-1]_q! = [n-1]_q![n+j-1]_{j,q}$, the last relation implies the required expression (6). □

Remark 1. The distribution function of the q-Erlang distribution of the second kind, in addition to expression (4), may be obtained as a q-integral of its q-density function as

$$F_n(w) = \int_0^w \frac{\lambda^n}{[n-1]_q!}u^{n-1}E_q(-\lambda q u)d_q u.$$

These two expressions of $F_n(w)$ imply the relation

$$\int_0^w \frac{\lambda^n}{[n-1]_q!}u^{n-1}E_q(-\lambda q u)d_q u = 1 - \sum_{x=0}^{n-1}E_q(-\lambda w)\frac{(\lambda w)^x}{[x]_q!}.$$

ΠΕΡΙΛΗΨΗ

Θεωρούμε ένα στοχαστικό πρότυπο, το οποίο εξελίσσεται χρονικά και επιτυχίες (ενδεχόμενο A) εμφανίζονται σε συνεχή σημεία. Ορίζεται μια στοχαστική ανέλιξη

Euler και συνάγεται η κατανομή της. Επίσης, συνάγεται η κατανομή του χρόνου αναμονής μέχρι την εμφάνιση συγκεκριμένου αριθμού επιτυχιών ως μια κατανομή q-Erlang.

REFERENCES

Benkherouf, L. and Bather, J. A. (1988). Oil exploration: sequential decisions in the face of uncertainty. *J. Appl. Probab.* **25**, 529-543.

Gasper, G. and Rahman, M. (2004) *Basic Hypergeometric Series.* Second Edition, Cambridge University Press, Cambridge.

Kemp, A. (1992a). Heine-Euler extensions of the Poisson distribution. *Comm. Statist. Theory Methods* **21**, 791-798.

Kemp, A. (1992b). Steady-state Markov chain models for the Heine and Euler distributions. *J. Appl. Probab.* **29**, 869-876.

Kemp, A. and Newton, J. (1990). Certain state-dependent processes for dichotomised parasite populations. *J. Appl. Probab.* **27**, 251-258.

Greek Statistical Institute
Proceedings of the 28th Panhellenic Statistics Conference (2015), pp.371-384

CONTROL CHARTS FOR THE LOGARITHMIC DISTRIBUTION

E. Demertzi, S. Psarakis
Department of Statistics, Athens University of Economics and Business
e_demertzi@yahoo.com, demertzi@aueb.gr, psarakis@aueb.gr

ABSTRACT

Logarithmic distribution is one of the discrete distributions with various applications in many fields such as in ecology, engineering, water resources, economics, pharmacology, biochemistry, molecular biology, genetics and biotechnology, environmental sciences, meteorology and atmospheric sciences, telecommunications, and others. As a result of its variety of applications, it appears to be important that control charts to detect shifts in mean and variability should be constructed based on the assumption that the distribution of the quality characteristic under study is the logarithmic distribution. Here we construct Shewhart-type and probability-type control charts for detecting parameter shifts for the logarithmic distribution and illustrate them using numerical examples with simulated data.

Keywords: Logarithmic distribution, Shewhart-type control charts, probability-type control charts, process mean monitoring, process variability monitoring.

1. INTRODUCTION

Control charts have been proposed for many types of distributions and have achieved great applications in many fields of our everyday lives, and the logarithmic distribution is a distribution with a lot of applications in many disciplines, as mentioned later in this section. As a result, it becomes obvious that control charts for the logarithmic distribution are required to be constructed.

Logarithmic distribution is an asymmetric discrete distribution with positive skewness. Its probability density function is given by $P(X = x) = -\frac{1}{\ln(1-\theta)}\frac{\theta^x}{x}$, $0 < \theta < 1$, $x = 1, 2, \ldots$. The moments of the logarithmic distribution are computed using the following formulas:

$$E(X) = -\frac{1}{\ln(1-\theta)}\frac{\theta}{1-\theta} \tag{1}$$

and

$$V(X) = -\frac{1}{\ln(1-\theta)}\frac{\theta}{(1-\theta)^2}\left(1+\frac{\theta}{\ln(1-\theta)}\right). \quad (2)$$

The coefficient of skewness of the logarithmic distribution is given by

$$sk = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{a\theta(1+\theta-3a\theta+2a^2\theta^2)}{\left[a\theta(1-a\theta)\right]^{3/2}},\ a = -\frac{1}{\ln(1-\theta)}. \quad (3)$$

The distribution was first introduced by [Fisher et al. (1943)] and has many applications in ecology ([Khang and Ong (2005)], [Williams (1944)], [Boswell and Patil (1970)], etc.), biology ([Williams (1947)], etc.), and economic applications ([Chatfield et al. (1966)], etc.). It can also be used in various fields, such as biochemistry, biochemical research methods, biology, molecular biology, genetics, and biotechnology ([Wolf et al. (2000)], [Wojtowicz and Tiuryn (2006a), (2006b), (2007)], etc.), meteorology and atmospheric sciences ([Williams (1952)], etc.), and many other scientific areas.

Since the logarithmic distribution has so many applications in our everyday lives, it would be useful to construct a control chart for this distribution. The following is an attempt to construct a control chart for the logarithmic distribution. More specifically, Section 2 deals with the construction of the logarithmic control charts for monitoring both the process mean and the process variability. All of the above are illustrated through an example in Section 3. Conclusions and further research topics are given in Section 4.

2. CONSTRUCTION OF LOGARITHMIC CONTROL CHARTS

The construction of logarithmic control charts is firstly going to be done based on the Shewhart-type control charts, by replacing the moments (mean and variance) of the distribution of concern (namely the logarithmic distribution) in the general form of the Shewhart-type control charts using the skewness correction as in [Chan and Cui (2003)]. Another way of finding the control limits of the chart is to derive them in terms of the probability of type I error or false alarm rate, α, using the logarithmic distribution (as [Chang and Gan (1999)] did for the case of the modified geometric distribution).

2.1 Logarithmic Control Charts for Detecting Shifts in the Process Mean

When we want to monitor a process, we usually monitor the process mean. Control of the process average or mean quality level is usually done with a control chart for the mean. Shewhart control charts for the process mean for asymmetric distributions are constructed as follows. The central line is placed at the mean of the distribution of concern, while its control limits are placed around the mean of the distribution at L times its standard deviation plus c_4^* times its standard deviation, where

$c_4^*\left(\bar{x}\right)=\dfrac{\frac{4}{3}\left[\mathrm{sk}\left(\bar{x}\right)\right]}{1+0.2\left[\mathrm{sk}\left(\bar{x}\right)\right]^2}$ is the skewness correction for the mean and sk(X) is the distribution's skewness coefficient. This rule is going to be applied here for the construction of the logarithmic control chart. The mean and variance of the logarithmic distribution are given by equations (1) and (2) respectively, while from Equation (3) the skewness coefficient for the mean is $\mathrm{sk}\left(\bar{\mathrm{x}}\right)=\dfrac{\mathrm{sk(X)}}{\sqrt{\mathrm{n}}}=\dfrac{\mathrm{a\theta}\left(1+\theta-3\mathrm{a\theta}+2\mathrm{a}^2\theta^2\right)}{\sqrt{\mathrm{n}}\left[\mathrm{a\theta}\left(1-\mathrm{a\theta}\right)\right]^{3/2}}$. This means that the skewness correction for the mean of the logarithmic distribution will be

$$\mathrm{c}_4^*\left(\bar{\mathrm{x}}\right)=\frac{4}{3}\frac{\left(1+\theta-3\mathrm{a\theta}+2\mathrm{a}^2\theta^2\right)\sqrt{\mathrm{n}}\left(\mathrm{a\theta}\right)^{1/2}\left(1-\mathrm{a\theta}\right)^{3/2}}{\mathrm{na\theta}\left(1-\mathrm{a\theta}\right)^3+0.2\left(1+\theta-3\mathrm{a\theta}+2\mathrm{a}^2\theta^2\right)^2}. \tag{4}$$

As a result, the central line (CL) and the upper and lower control limits (UCL and LCL, respectively) of the Shewhart-type logarithmic control chart for the process mean are as follows.

$$\mathrm{UCL}=-\frac{1}{\ln\left(1-\theta\right)}\theta\left(1-\theta\right)^{-1}+\left[\mathrm{L}+\mathrm{c}_4^*\left(\bar{\mathrm{x}}\right)\right]\sqrt{\frac{-\frac{1}{\ln\left(1-\theta\right)}\theta\left(1+\frac{1}{\ln\left(1-\theta\right)}\theta\right)\left(1-\theta\right)^{-2}}{\mathrm{n}}}$$

$$\mathrm{CL}=-\frac{1}{\ln\left(1-\theta\right)}\theta\left(1-\theta\right)^{-1} \tag{5}$$

$$\mathrm{LCL}=-\frac{1}{\ln\left(1-\theta\right)}\theta\left(1-\theta\right)^{-1}+\left[-\mathrm{L}+\mathrm{c}_4^*\left(\bar{\mathrm{x}}\right)\right]\sqrt{\frac{-\frac{1}{\ln\left(1-\theta\right)}\theta\left(1+\frac{1}{\ln\left(1-\theta\right)}\theta\right)\left(1-\theta\right)^{-2}}{\mathrm{n}}}$$

Each point outside the control limits is an indication that our logarithmic process is out of statistical control.

As previously mentioned, another way of finding the control limits of the chart is to derive them in terms of the probability of type I error or false alarm rate, α, using the logarithmic distribution. This is done by the use of the cumulative probability of the logarithmic distribution, which is given by the form

$P(X \le x) = 1 + \frac{1}{\ln(1-\theta)} \sum_{u=x+1}^{\infty} \frac{\theta^u}{u} = -\frac{1}{\ln(1-\theta)} \sum_{u=1}^{x} \frac{\theta^u}{u}$. Therefore, for a significance level α, we have $P(\bar{X} < LCL) = \frac{\alpha}{2}$ and $P(\bar{X} < LCL) = -\frac{1}{\ln(1-\theta)} \sum_{u=1}^{LCL} \frac{\theta^u}{u}$, from which we get that

$$\sum_{u=1}^{LCL} \frac{\theta^u}{u} = -\frac{\alpha}{2} \ln(1-\theta) \tag{6}$$

and solving this equation we obtain the expression for LCL (see below). Similarly, for the upper control limit, we have $P(\bar{X} > UCL) = \frac{\alpha}{2}$ and $P(\bar{X} > UCL) =$

$= 1 - P(\bar{X} \le UCL) = 1 + \frac{1}{\ln(1-\theta)} \sum_{u=1}^{UCL} \frac{\theta^u}{u}$, from which we get that

$$\sum_{u=1}^{UCL} \frac{\theta^u}{u} = \left(\frac{\alpha}{2} - 1\right) \ln(1-\theta). \tag{7}$$

and solving this equation we obtain the expression for UCL (see below).

For the computation of the sum required for finding the values of LCL and UCL, we will use the following equation ([Gradshteyn and Ryzhik (1980)], equation 2.735):

$$\int x^{2n+1} \ln|x^2 - a^2| \, dx = \frac{1}{2n+2} \left\{ \left(x^{2n+2} - a^{2n+2}\right) \ln|x^2 - a^2| - \sum_{k=1}^{n+1} \frac{1}{k} a^{2n-2k+2} x^{2k} \right\}.$$

For a = 1, and setting w = n + 1 and then $y = x^2$, the equation becomes

$$\sum_{k=1}^{w} \frac{y^k}{k} = \left(y^w - 1\right) \ln|y-1| - w \int y^{w-1} \ln|y-1| \, dy. \tag{8}$$

Combining equations (6) and (8) we obtain the following relationship: $\left(\theta^{LCL} - 1\right) \ln|\theta - 1| - LCL \int \theta^{LCL-1} \ln|\theta - 1| \, d\theta = -\frac{\alpha}{2} \ln(1-\theta)$, and differentiating with respect to θ and considering that c is a positive constant, we get $\left(\theta^{LCL} - 1\right) \frac{1}{|\theta - 1|} = \frac{\alpha}{2} \frac{1}{1-\theta} + c$.

Considering that $0 < \theta < 1$ and for an appropriate value of c so that both sides of the equation above are negative, we get $\theta^{LCL} = 1 - \frac{\alpha}{2} \frac{1}{1-\theta} |\theta - 1|$ and since

$$0 < \theta < 1 \Rightarrow \theta - 1 < 0 \Rightarrow |\theta - 1| = -(\theta - 1) = 1 - \theta, \tag{9}$$

we will finally have $\theta^{LCL} \overset{(9)}{=} 1-\frac{\alpha}{2}\frac{1}{1-\theta}(1-\theta) = 1-\frac{\alpha}{2} \Rightarrow LCL\ln(\theta) = \ln\left(1-\frac{\alpha}{2}\right) \Rightarrow$

$$LCL = \frac{\ln\left(1-\frac{\alpha}{2}\right)}{\ln(\theta)}. \tag{10}$$

Similarly, for UCL, when combining equations (7) and (8), and then differentiating with respect to θ and considering that c is a positive constant, we get $\left(\theta^{UCL}-1\right)\frac{1}{|\theta-1|} = \left(1-\frac{\alpha}{2}\right)\frac{1}{1-\theta} + c$. Considering that $0 < \theta < 1$ and for an appropriate value of c so that both sides of the equation above are negative, we have

$$\theta^{UCL} = 1-\left(1-\frac{\alpha}{2}\right)\frac{1}{1-\theta}|\theta-1| \overset{(9)}{=} 1-\left(1-\frac{\alpha}{2}\right) = \frac{\alpha}{2} \Rightarrow \ln\left(\theta^{UCL}\right) = \ln\left(\frac{\alpha}{2}\right) \Rightarrow UCL = \frac{\ln\left(\frac{\alpha}{2}\right)}{\ln(\theta)}.$$

Similarly for the central line we have $CL = \frac{\ln(1-0.5)}{\ln(\theta)} = \frac{\ln(0.5)}{\ln(\theta)}$. As a result from all the above, the control limits of the chart in terms of the probability of type I error, α, are as follows.

$$UCL_{\alpha} = \frac{\ln\left(\frac{\alpha}{2}\right)}{\ln(\theta)}$$

$$CL_{\alpha} = \frac{\ln(0.5)}{\ln(\theta)} \tag{11}$$

$$LCL_{\alpha} = \frac{\ln\left(1-\frac{\alpha}{2}\right)}{\ln(\theta)}$$

2.4 Logarithmic Control Charts for Detecting Shifts in the Process Variability

We previously showed how to construct logarithmic control charts for monitoring the process mean. It is, however, very important to monitor the process variation, too, besides the process mean, since a change in the process variation can affect the statistic plotted on the means chart and therefore, knowing that the process variation has changed can prevent an erroneous interpretation of the means chart. It should also be noted here that both increases and decreases in the process variability are

important to be monitored and detected since the former correspond to decreases in quality while the latter correspond to process improvement. As a result, the ability to quickly and effectively detect changes in the process variability is very important for process improvement.

Shewhart control charts for the process variability for asymmetric distributions are constructed as follows. The central line is placed at the standard deviation of the distribution of concern, while its control limits are placed around the standard deviation of the distribution at L times its standard deviation plus c_4^* times its standard deviation, where $c_4^*\left(s^2\right)=\frac{\frac{4}{3}\left[\mathrm{sk}\left(s^2\right)\right]}{1+0.2\left[\mathrm{sk}\left(s^2\right)\right]^2}$ is the skewness correction for the process variability and sk(X) is the distribution's skewness coefficient. This rule is going to be applied here for the construction of the logarithmic control chart. The variance of the logarithmic distribution is given by Equation (2), which means that the standard deviation is $\sqrt{-\frac{1}{\ln(1-\theta)}\theta\left(1+\frac{1}{\ln(1-\theta)}\theta\right)(1-\theta)^{-2}}$. For the computation of the variance of the standard deviation see Appendix.

Generally, we have $\mathrm{E}\left(s^2\right)=\sigma^2$ and $\mathrm{V}\left(s^2\right)=\frac{1}{n}\left(\mathrm{E}\left(X-\mu\right)^4-\frac{n-3}{n-1}\sigma^4\right)$ (see [Mood et al. (1974)]) and using equations (11)-(17) in [Choi and Sweetman (2010)] we have $\mathrm{E}\left[\left(X-\mu\right)^4\right]=\mathrm{E}\left(X^4\right)-4\mu\mathrm{E}\left(X^3\right)+6\mu^2\mathrm{E}\left(X^2\right)-3\mu^4$ or, equivalently, $\mathrm{E}\left[\left(X-\mu\right)^4\right]=$ $=\mathrm{E}\left(X^4\right)-4\mu\mathrm{E}\left(X^3\right)+6\mu^2\sigma^2+3\mu^4$, which will then become $\mathrm{E}\left[\left(X-\mu\right)^4\right]=$ $=\frac{\alpha\theta}{(1-\theta)^4}\left[1+4\theta+\theta^2-4\alpha\theta(1+\theta)+6\alpha^2\theta^2-3\alpha^3\theta^3\right]$ for the case of the logarithmic distribution, with $\alpha=-\frac{1}{\ln(1-\theta)}$. Using all the above, we find that the skewness coefficient for the variability is computed using the relationship $\mathrm{sk}\left(s^2\right)=\mathrm{E}\left(\frac{\left[s^2-\mathrm{E}\left(s^2\right)\right]^3}{\left[\mathrm{V}\left(s^2\right)\right]^{3/2}}\right)=\frac{\left(\sigma^2-n\sigma^4\right)^3}{(n-1)^3\left[\frac{1}{n}\mathrm{E}\left(X-\mu\right)^4-\frac{n-3}{n-1}\sigma^4\right]^{3/2}}$, which for the case of the logarithmic distribution becomes

$$sk\left(s^2\right)=\frac{\left(\frac{\alpha\theta\left(1-\alpha\theta\right)}{\left(1-\theta\right)^2}-n\left[\frac{\alpha\theta\left(1-\alpha\theta\right)}{\left(1-\theta\right)^2}\right]^2\right)^3}{\left(n-1\right)^3\left[\frac{\alpha\theta\left[1+4\theta+\theta^2-4\alpha\theta\left(1+\theta\right)+6\alpha^2\theta^2-3\alpha^3\theta^3\right]}{n\left(1-\theta\right)^4}-\frac{\left(n-3\right)\alpha^2\theta^2\left(1-\alpha\theta\right)^2}{\left(n-1\right)\left(1-\theta\right)^4}\right]^{3/2}} \quad (12)$$

Plugging this equation in

$$c_4^*\left(s^2\right)=\frac{\frac{4}{3}sk\left(s^2\right)}{1+0.2\left[sk\left(s^2\right)\right]^2}, \quad (13)$$

we get the equation for the skewness correction for the variability of the logarithmic distribution which we will use below for the construction of the logarithmic control charts for the process variability.

As a result, the central line (CL) and the upper and lower control limits (UCL and LCL, respectively) of the Shewhart-type logarithmic control chart are as follows.

$$UCL=-\frac{\sqrt{nm}-1}{\sqrt{nm-1}}\frac{1}{\ln\left(1-\theta\right)}\frac{\theta}{1-\theta}+\left[L+c_4^*\left(s^2\right)\right]\sqrt{\frac{-\frac{\theta}{\ln\left(1-\theta\right)}\left(1+\frac{\theta}{\ln\left(1-\theta\right)}\right)}{\left(1-\theta\right)^2}}$$

$$CL=-\frac{\sqrt{nm}-1}{\sqrt{nm-1}}\frac{1}{\ln\left(1-\theta\right)}\frac{\theta}{1-\theta} \quad (14)$$

$$LCL=-\frac{\sqrt{nm}-1}{\sqrt{nm-1}}\frac{1}{\ln\left(1-\theta\right)}\frac{\theta}{1-\theta}+\left[-L+c_4^*\left(s^2\right)\right]\sqrt{\frac{-\frac{\theta}{\ln\left(1-\theta\right)}\left(1+\frac{\theta}{\ln\left(1-\theta\right)}\right)}{\left(1-\theta\right)^2}}$$

On the other hand, as far as the probability-type control limits are concerned, we need to find the formula for the computation of $P\left(\sigma<LCL\right)=\frac{\alpha}{2}$ for the logarithmic distribution, for a significance level α. Considering that $\sigma=\sqrt{\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij}-\mu\right)^2}{nm-1}}$, where

n is the sample size of each sub-sample and m is the number of sub-samples, we obtain

$$P(X<A)=P\left((X-\mu)^2<(A-\mu)^2\right)=P\left(\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij}-\mu)^2<\sum_{i-1}^{m}\sum_{j=1}^{n}(A-\mu)^2\right)=$$

$$=P\left(\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij}-\mu)^2}{nm-1}<\frac{\sum_{i-1}^{m}\sum_{j=1}^{n}(A-\mu)^2}{nm-1}\right)=P\left(\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij}-\mu)^2}{nm-1}<\frac{nm(A-\mu)^2}{nm-1}\right)=$$

$$=P\left(\sqrt{\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij}-\mu)^2}{nm-1}}<\sqrt{\frac{nm(A-\mu)^2}{nm-1}}\right)=P\left(\sigma<\sqrt{\frac{nm}{nm-1}}(A-\mu)\right)$$

and using the previously mentioned equation $P(X<A)=-\frac{1}{\ln(1-\theta)}\sum_{u=1}^{A}\frac{\theta^u}{u}$ (resulting from the definition of the cumulative probability of the logarithmic distribution), we obtain the following equation: $P\left(\sigma<\sqrt{\frac{nm}{nm-1}}(A-\mu)\right)=-\frac{1}{\ln(1-\theta)}\sum_{u=1}^{\sqrt{\frac{nm}{nm-1}}(A-\mu)}\frac{\theta^u}{u}$, where $LCL=\sqrt{\frac{nm}{nm-1}}(A-\mu)$. Following the same procedure as for obtaining equations (6) and (10), we get $LCL=\frac{\ln\left(1-\frac{\alpha}{2}\right)}{\ln(\theta)}$. Similarly, $UCL=\frac{\ln\left(\frac{\alpha}{2}\right)}{\ln(\theta)}$ and $CL=\frac{\ln(0.5)}{\ln(\theta)}$. As a result from all the above, the control limits of the chart in terms of the probability of type I error, α, are as follows.

$$UCL_\alpha=\frac{\ln\left(\frac{\alpha}{2}\right)}{\ln(\theta)}$$

$$CL_\alpha=\frac{\ln(0.5)}{\ln(\theta)} \qquad (15)$$

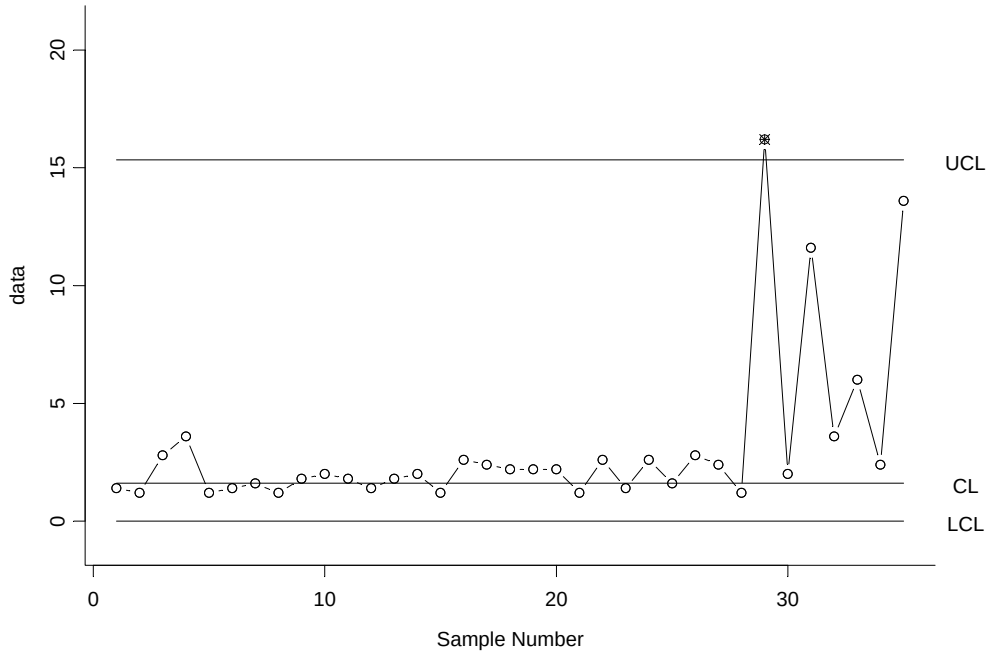
$$LCL_\alpha=\frac{\ln\left(1-\frac{\alpha}{2}\right)}{\ln(\theta)}$$

3. EXAMPLE ON THE LOGARITHMIC CONTROL CHARTS USING SIMULATED DATA

2.1 Probability-type Control Charts

Suppose we take 35 samples of n = 5 observations from a logarithmic process as follows. First, we take 25 samples of 5 observations from a logarithmic process with in-control theta value equal to 0.65. Now suppose that a shift of one standard deviation unit occurs in the process mean, and after that shift, we draw another set of 10 samples of 5 observations each, from the process. Then, for this data set, which can be seen in Table 1, in this subsection, we construct the probability-type logarithmic control charts for monitoring the process mean and variability at a significance level equal to the most commonly (when dealing with control charts) used value of 0.27%, which corresponds to 0.27% probability of falsely rejecting the null hypothesis that our process is in control. The Shewhart-type control charts for the same data set are going to be presented in the next subsection.

Figure 1. *Probability-type Control Chart for the Process Mean for the Data Set in Table 1*



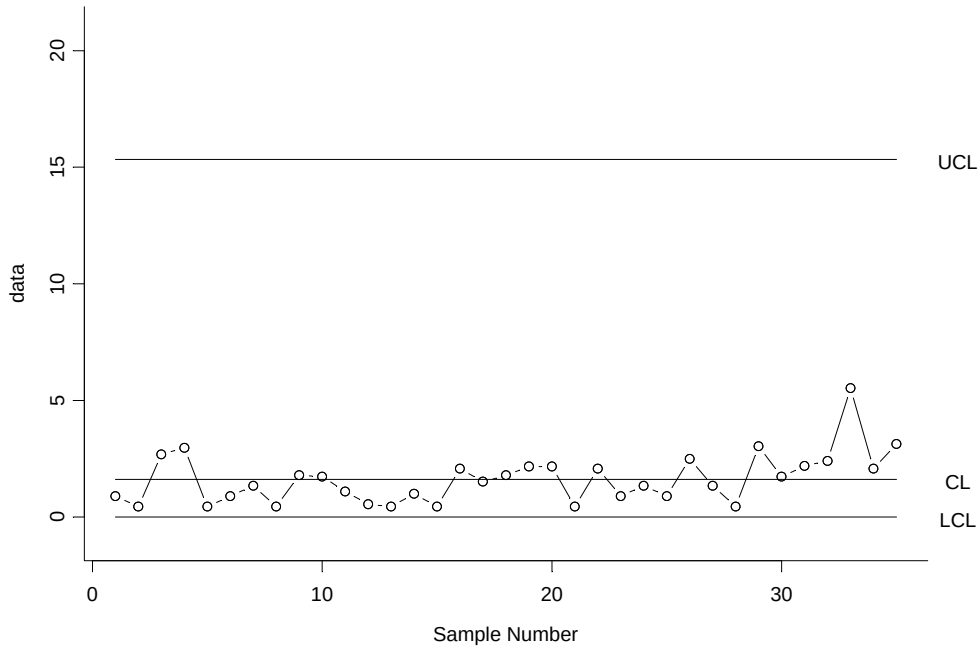
The probability-type logarithmic control charts for monitoring the process mean and variability are shown in Figure 2 and Figure 3, respectively. As we can see there, the probability-type logarithmic control chart for the process mean indicates that the process mean has shifted to an out-of-control level about after the first 28 samples.

The corresponding control chart for the process variability does not indicate any out-of-control samples, although there is evidence of a slightly increasing trend of the variability after the first 28 samples. This means that an assignable cause has occurred in the process affecting both its mean and variability and causing its mean to shift to an out-of-control level.

Table 1: *Data From A Logarithmic Process With In-Control $\theta = 0.65$ And A Shift Of One Standard Deviation Unit In The Process Mean After The First 125 Observations (25^{th} Sample Of 5 Observations)*

Data Set	1	1	3	1	1	1	1	2	1	1	7	1	4	1	1	1	1	3
	5	8	1	1	2	1	1	1	1	3	1	1	1	1	1	4	1	1
	1	2	1	1	1	5	1	1	1	5	1	2	1	1	3	1	1	3
	1	1	1	1	2	2	1	2	2	2	2	1	3	1	2	3	1	1
	2	1	1	3	1	1	2	6	1	5	2	2	2	5	1	3	1	1
	6	2	1	1	1	1	1	2	6	1	1	1	1	1	2	1	6	1
	2	3	1	1	1	3	1	2	4	1	2	4	1	1	3	1	2	1
	1	6	1	5	1	3	4	1	3	1	1	2	1	1	15	18	12	16
	20	1	1	2	5	1	14	12	8	12	12	5	1	1	6	5	12	1
	12	3	2	1	2	2	1	6	11	13	19	13	12					

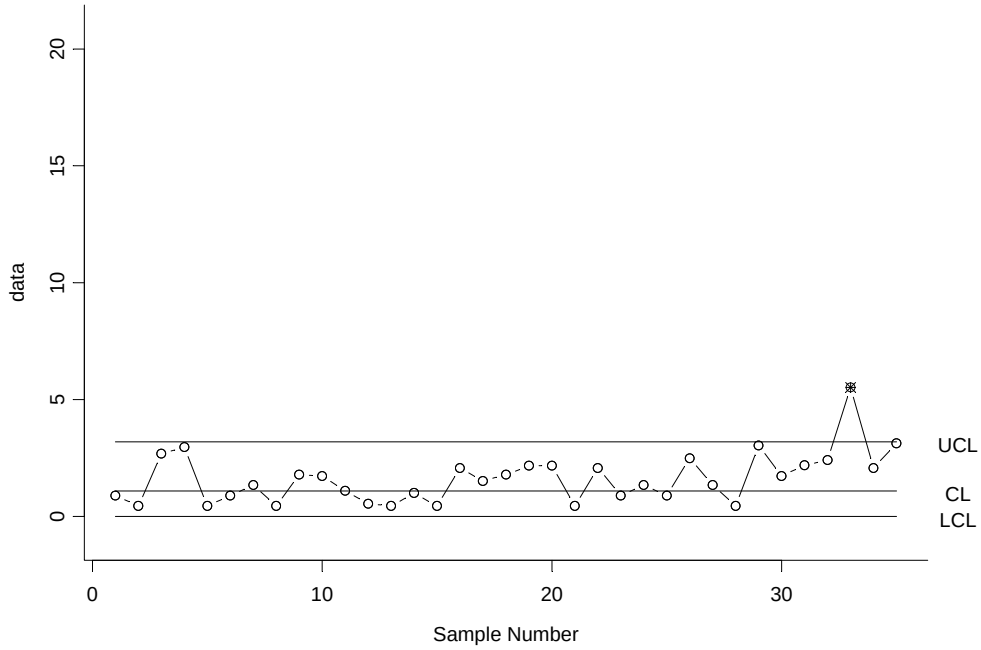
Figure 2. *Probability-type Control Chart for the Process Variability for the Data Set in Table 1*



3.2 Shewhart-type Control Charts

Suppose we have the same data set as in the previous subsection which can be seen in Table 1. We are now going to construct the Shewhart-type logarithmic control charts for those data for the cases of monitoring the mean and the variability of the process, using L = 3 standard deviations. These two control charts are shown in Figure 4 and Figure 5, respectively.

Figure 3. *Shewhart-type Control Chart for the Process Mean for the Data Set in Table 1*

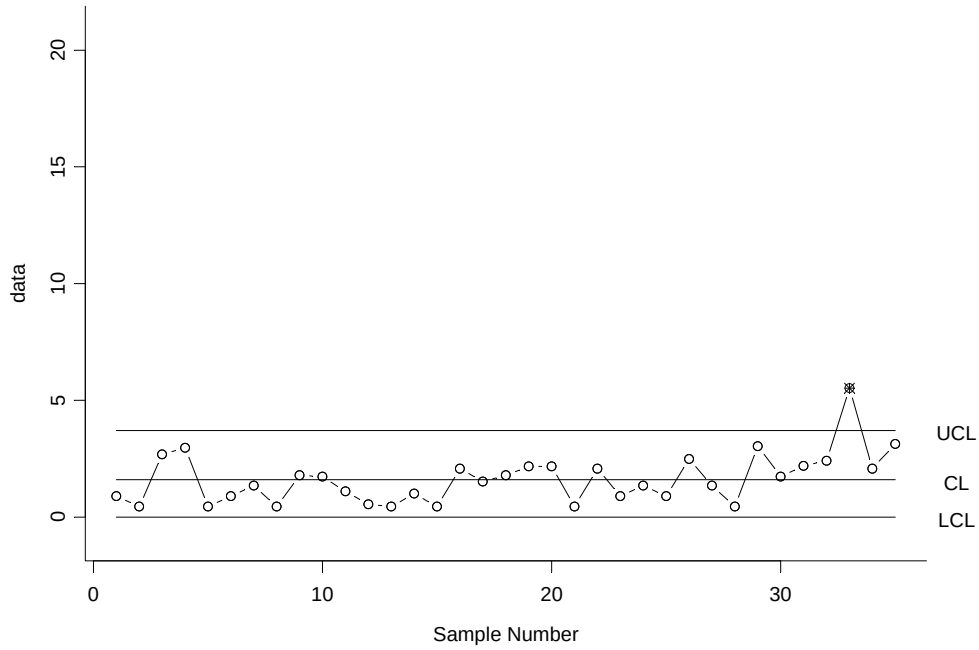


The two control charts seem to detect an increasing trend after the first 28 samples leading to an out-of-control point, indicating that an assignable cause has occurred in the process causing its mean and variability to shift to an out-of-control level.

4. CONCLUSIONS AND FURTHER RESEARCH

The logarithmic distribution has many applications in various fields of science and everyday life. This makes it clear that we need to construct control charts for logarithmically distributed data. As a result, here we attempted to construct control charts for the logarithmic distribution for the theoretical case of knowing the parameters of the distribution.

Figure 4. *Shewhart-type Control Chart for the Process Variability for the Data Set in Table 1*



However, there is still a lot to be done on the logarithmic control charts subject, such as the following. The case of the more realistic and usual in everyday life scenario of not knowing the actual value of the distribution's parameter and estimating it from the data needs be studied, and logarithmic CUSUM control charts need to be constructed, subjects which are already under study by the author as part of her PhD thesis. Moreover, the case of variable sample size should be studied and recommendations for the choice of the effective sample size are also required. Last but not least, control charts for generalizations of the logarithmic distribution and the multivariate case are yet to be dealt with.

APPENDIX

Calculation of the mean and variance of standard deviation of the logarithmic distribution

$$E(S)=E\left(\sqrt{\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij}-\bar{x}\right)^2}{nm-1}}\right)=\frac{1}{\sqrt{nm-1}}E\left(\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij}-\bar{x}\right)^2}\right)=$$

$$=\frac{1}{\sqrt{nm-1}}E\left(\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}x_{ij}^2-nm\bar{x}^2}\right)\le\frac{1}{\sqrt{nm-1}}E\left(\sum_{i=1}^{m}\sum_{j=1}^{n}\sqrt{x_{ij}^2}-\sqrt{nm\bar{x}^2}\right)=$$

$$=\frac{1}{\sqrt{nm-1}}E\left(\sum_{i=1}^{m}\sum_{j=1}^{n}x_{ij}-\sqrt{nm}\bar{x}\right)=\frac{1}{\sqrt{nm-1}}\left(\sum_{i=1}^{m}\sum_{j=1}^{n}E\left(x_{ij}\right)-\sqrt{nm}\bar{x}\right)=$$

$$=\frac{1}{\sqrt{nm-1}}nm\frac{-1}{\ln(1-\theta)}\frac{\theta}{1-\theta}+\frac{\sqrt{nm}}{\sqrt{nm-1}}\frac{1}{\ln(1-\theta)}\frac{\theta}{1-\theta}=-\frac{nm-\sqrt{nm}}{\sqrt{nm-1}}\frac{1}{\ln(1-\theta)}\frac{\theta}{1-\theta}=$$

$$=-\frac{\sqrt{nm}-1}{\sqrt{nm-1}}\frac{1}{\ln(1-\theta)}\frac{\theta}{1-\theta}$$

where n is the sample size of each sub-sample and m is the number of sub-samples.

Considering $V(\bar{x})=V\left(\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}x_i}{nm}\right)=\frac{1}{(nm)^2}\sum_{i=1}^{m}\sum_{j=1}^{n}V(x_i)=\frac{-\frac{\theta}{\ln(1-\theta)}\left(1+\frac{\theta}{\ln(1-\theta)}\right)}{nm(1-\theta)^2}$, then

$$V(S)=V\left(\sqrt{\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij}-\bar{x}\right)^2}{nm-1}}\right)=\frac{1}{nm-1}V\left(\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij}-\bar{x}\right)^2}\right)=$$

$$=\frac{1}{nm-1}V\left(\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}x_{ij}^2-nm\bar{x}^2}\right)\le\frac{1}{nm-1}V\left(\sum_{i=1}^{m}\sum_{j=1}^{n}\sqrt{x_{ij}^2}-\sqrt{nm\bar{x}^2}\right)=$$

$$=\frac{1}{nm-1}V\left(\sum_{i=1}^{m}\sum_{j=1}^{n}x_i-\sqrt{nm}\bar{x}\right)=\frac{1}{nm-1}\left(\sum_{i=1}^{m}\sum_{j=1}^{n}V(x_i)-nmV(\bar{x})\right)=$$

$$=\frac{-\frac{\theta}{\ln(1-\theta)}nm\left(1+\frac{\theta}{\ln(1-\theta)}\right)}{(nm-1)(1-\theta)^2}-\frac{-\frac{\theta}{\ln(1-\theta)}\left(1+\frac{\theta}{\ln(1-\theta)}\right)}{(nm-1)(1-\theta)^2}=\frac{-\frac{\theta}{\ln(1-\theta)}\left(1+\frac{\theta}{\ln(1-\theta)}\right)}{(1-\theta)^2}$$

REFERENCES

Boswell, M T., Patil, G P. (1970). Chance mechanisms generating the logarithmic series distribution used in analysis of number of species and individuals. *Statistical Ecology*, **1**, 99-130.

Chan, LK and Cui HJ. (2003). Skewness correction $\overline{X}$ and R charts for skewed distributions. *Naval Research Logistics*, **50**, 555-573.

Chang, TC and Gan, FF. (1999). Charting techniques for monitoring a random shock process. *Quality and Reliability Engineering International*, **15**, 295-301.

Chatfield, C, Ehrenberg, ASC and Goodhardt, GJ. (1966). Progress on a simplified model of stationary purchasing behaviour (with discussion). *Journal of the Royal Statistical Society (Series A)*, **129**, 317-367.

Choi, M. and Sweetman, B. (2010). Efficient calculation of statistical moments for structural health monitoring. *Journal of Structural Health Monitoring*, **9**, 13-24.

Fisher, RA, Corbett, RS, and Williams, CB. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology*, **12**, 42-58.

Gradshteyn, IS and Ryzhik, IM. (1980). *Table of Integrals, Series and Products,* Corrected and Enlarged Edition (prepared by Jeffrey, A.), Orlando, Florida: Academic Press, 206.

Khang, TF and Ong, SH. (2005). *A New Generalized Logarithmic Distribution with Applications in Ecology*, Petaling Jaya: International Statistics Conference, Statistics in the Technological Age, ISCM.

Mood, A.M., Graybill F.A., and Boes D.C. (1974). *Introduction to the Theory of Statistics*, 3rd ed., New York: McGraw-Hill, 229.

Williams, CB. (1944). Some applications of the logarithmic series and the index of diversity to ecological problems. *Journal of Ecology*, **32**, 1-44.

Williams, CB. (1947). The logarithmic series and its applications to biological problems. *Journal of Ecology*, **34**, 253-272.

Williams, CB. (1952). Sequences of wet and dry days considered in relation to the logarithmic series. *Quarterly Journal of the Royal Meteorological Society*, **78**, 91-96.

Wojtowicz, D and Tiuryn, J. (2006). On genome evolution with accumulated change and innovation, Comparative Genomics. *Proceedings Lecture Notes in Computer Science*, **4205**, 39-50.

Wojtowicz, D and Tiuryn, J. (2006). On genome evolution with innovation, Mathematical Foundations of Computer Science. *Proceedings Lecture Notes in Computer Science*, **4162**, 801-811.

Wojtowicz, D and Tiuryn, J. (2007). Evolution of gene families based on gene duplication, loss, accumulated change, and innovation. *Journal of Computational Biology*, **14**, 479-495.

Wolf, YI, Grishin, NV and Koonin, EV. (2000). Estimating the number of protein folds and families from complete genome data. *Journal of Molecular Biology*, **299**, 897-905.

Greek Statistical Institute
Proceedings of the 28th Panhellenic Statistics Conference (2015), pp.385-399

INVESTIGATION OF EARTHQUAKE MAGNITUDE AND INTEREVENT TIME DISTRIBUTION IN CORINTH GULF AND MYGDONIA BASIN WITH THE USE OF STOCHASTIC TOOLS

C. Gkarlaouni[1], S. Lasocki[2] E., Papadimitriou[1]
[1]Department of Geophysics, School of Geology, Aristotle University of Thessaloniki
{hagarl, ritsa}@geo.auth.gr
[2] Institute of Geophysics, Polish Academy of Sciences, Warszawa
lasocki@igf.edu.pl

ABSTRACT

Great effort has been exerted for the investigation of the seismicity parameters distribution, especially regarding magnitude and interevent time between consecutive earthquakes. An approach combining a parametric and a non-parametric methodology are tested for the instrumental seismicity, in two of the most active extensional fault zones in Greece (Corinth gulf and Mygdonia basin) that accommodate intense seismicity with devastating events. The non-parametric approach refers to the "smoothed bootstrap test for modality and bump-hunt test" for unveiling possible complex distribution characteristics. The input data regard complete and declustered seismicity catalogs ($M \geq M_C$) for the time interval: August 2008 - October 2014. The test is based upon setting the null hypotheses that the distributions of interevent time and magnitude are unimodal. A multimodal distribution is evidenced in most cases where clustering of earthquakes is dominant. When multimodality is identified, the next target is to define the exact locations of modes and bumps. The results are interpreted in terms of the seismotectonic properties of the fault populations in each area and earthquake clustering.

Keywords: magnitude, interevent time, distribution, multimodality

1. INTRODUCTION

One of the first objectives of seismicity studies is the construction of numerical seismicity models, in an attempt to a long-term forecasting of strong earthquakes and hazard assessment. Since the seismicity process can be decomposed into several parameters due to its multiparametrical nature, components of time, space, size and energy are separately approached. The investigation of earthquake magnitude frequency ended up at the Gutenberg-Richter (G-R) empirical relation (Gutenberg and Richter, 1944) for earthquakes with magnitude greater than the lowest threshold (magnitude of completeness, M_C). The parameter of this relation (*b*) provides information on the seismotectonic regime and stress state. The G-R law is given by the equation

$$\log_{10} N(M) = a + bM \ ,$$

where *N(M)* stands for the cumulative number of earthquakes with magnitude greater than or equal to M_C. The parameter *a* describes the level of seismic activity, and *b* is characteristic for the seismogenic layer of the corresponding region implying the relative abundance between strong and small events. However, thorough investigations revealed that there are several deviations from the G-R linearity (Utsu, 1999) depending on different seismicity data, although this fact does not hinder the vast functionality of this relation (e.g. Kagan, 1991; Main, 1996; Utsu, 1999; Kijko et al., 2001; Pisarenko and Sornette, 2003/2004). The main difference between all the proposed models is the fact that they are exhibiting either a unimodal or a multimodal density probability function (pdf) of the magnitudes (Lasocki and Papadimitriou, 2006)

On the other hand, the temporal characteristics of seismicity in terms of the time lag which mediates between two consecutive events (interevent time, IET) shows controversial results for interpreting the occurrence of a forthcoming earthquake. The existence of dependent (main shocks) and independent events (aftershocks) complicates the accurate identification of interevent times. It is suggested that the temporal process is connected with a memoryless and random behavior related to Poisson distributions (Gardner and Knoppof, 1974). In other cases there is an implication that strong earthquakes periodically occur along long rupture zones (Wesnousky, 1994). For this reason, a variety of exponential-like distributions have been checked for their applicability, like Weibull (Cornell and Winterstein, 1988), Gamma (Corral, 2003) or stretched exponential (Altmann and Kanz, 2005) for normalized interevent times. Lognormal distribution is associated with good results in certain cases (Nishenko and Buland, 1987) whereas, the Generalized Pareto distribution for the strong recurrence times has been tested by Pisharenko and Sornette (2003/ 2004) and Byrdina et al. (2006) with satisfying results.

We investigate the distribution of magnitude and interevent time for recent seismicity in two fault population comprised in Greece, the Mygdonian basin and the Corinth gulf. These two fault populations were selected among others because they share common seismotectonic properties as being developed in two extensional back–arc

basins, although with considerably different seismic moment rate. The methodology we follow, concerns a combination of both a parametric and a non–parametric statistical procedure, the so-called smoothed bootstrap test for multimodality, which is used to indicate the existence of the pdf complexity. This method is now applied to the interevent time of the consecutive events in addition to magnitudes (Lasocki and Papadimitriou, 2006) for both complete and declustered seismicity catalogs.

The results extracted from the application of the stochastic process enable us to provide an interpretation of seismicity behavior as well as they denote a further insight into the physical mechanisms for the current seismicity state in the two study regions.

3. DATA AND PROCESSES

The employed seismicity catalogs include earthquakes that occurred in the time span between August 2008 and October 2014 within the boundaries of the selected areas (λ=22.80°-23.90°, φ=40.10°-41.00° for Mygdonia basin and λ=21.00º-23.40º, φ=37.90º-38.70º for the Corinth gulf) and were instrumentally recorded by the Hellenic Unified Seismologic Network (HUSN). An important factor often affecting the reliability of the statistical results is the quality and the homogeneity of the observations. For this reason and because of the temporal variation of the completeness magnitude, different sub-catalogs were compiled. Complete earthquake catalogs over a certain magnitude threshold are important not only for the process itself but also for the interpretation of the results regarding complexity properties. The magnitude threshold (M_C) for each sub catalog was estimated according to the algorithm of Leptokaropoulos et al. (2013) based on the methodology proposed by Wiemer and Wyss (2002) (Table 1).

The second step is the catalog declustering, a process for removing the aftershocks from the initial seismicity catalog. For distinguishing independent (mainshocks) from dependent events (aftershocks) the Reasenberg's declustering algorithm was applied (Reasenberg, 1985). Two data sets were obtained for each region, the entire data sets with the lower magnitude cut off and the declustered catalogs, with a different magnitude threshold, since the second catalogs have different characteristics than the original ones. Information on the magnitude thresholds (M_C), maximum magnitude values (M_{Max}) and number of data (N) in each catalog are provided in Table 1.

Table 1. *Information on the four seismicity catalogs used, where N is the number of the observations, Mc the completeness magnitude and M_{Max} is the strongest earthquake for the study period.*

n	Seismicity Catalog	Code	N	M_C	M_{Max}
1	Corinth gulf, full catalog	A1	3191	2.3	6.2
2	Corinth gulf, declustered catalog	A2	2823	2.2	6.2
3	Mygdonia, full catalog	B1	559	1.7	4.8
4	Mygdonia, declustered catalog	B2	573	1.6	4.8

2. METHODOLOGY

2.1 Parametric Approach

In the majority of studies with similar objective, the probability density function (pdf) or the probability (mass) function (pf) of seismicity parameters is one of the most meaningful features to be investigated. For this reason, the parametric approach which was followed, relies on the application of goodness of fit tests on theoretical probabilistic distributions. The most popular theoretical distributions for earthquake magnitude and interevent times are the Exponential, the Weibull, the Lognormal, the Gamma and the Generalized Pareto distributions. In each case, the pdf along with the Maximum Likelihood values (MLE) were estimated. Now for x>0, the pdf for the Exponential distribution and the MLE are given, respectively, by

$$f(x|\mu) = (1/\mu)\exp\{-(x/\mu)\}$$

and

$$\ln L = -n\ln\mu - (1/\mu)\sum_{i=1}^{n} x_i \ .$$

For earthquake magnitudes investigation the truncated G-R distribution is widely used as a modification of the G-R law and imposes an upper and lower constraint to the magnitude values, related to the seismogenic crust properties and the magnitude of completeness (M_C=$M_{\min}$). For $M_{min} \le M \le M_{max}$ the respective pf is given by (Page, 1968)

$$f(M) = \frac{\beta e^{-\beta(M - M_{\min} + \Delta M/2)}}{1 - e^{-\beta(M_{\max} - M_{\min} + \Delta M/2)}} \ ,$$

where $\beta = b\log 10$. The b value is approximated by the estimator proposed by Aki (1965) and suggested by many researchers, which is based on MLE and is given by

$$b = \frac{1}{\ln(10)\left[\langle M \rangle - (M_C - \Delta M/2)\right]} \ ,$$

where $<M>$ is the average magnitude in the data sample and ΔM is the binning width of the catalog, taken here equal to 0.1. The pdf along with the MLE for the Weibull distribution are given by

$$f(x|a,b) = (b/a)(x/a)^{b-1} \exp\left\{-(x/a)^b\right\}$$

and

$$\ln L = n \ln b - (b-1)\sum_{i=1}^{n} \ln x_i - (1-a^b)\sum_{i=1}^{n} x_i^b \cdot$$

The pdf and the subsequent MLE for the Lognormal distribution are

$$f(x|\mu,\sigma) = \left\{1/x\sigma\sqrt{2\pi}\right\}\exp\left\{-(\ln x - \mu)^2/2\sigma^2\right\}$$

and

$$\ln L = -(n/2)\ln\left(2\pi\sigma^2\right) - \sum_{i=1}^{n} \ln x_i - \left(\sum_{i=1}^{n} \ln\left(x_i\right)^2/2\sigma^2\right) + \left(\sum_{i=1}^{n} \mu \ln\left(x_i\right)/\sigma^2 - n\mu^2/2\sigma^2\right) \cdot$$

The pdf of the Gamma distribution and its MLE are given by

$$f(x|a,b) = \left(1/b^a \Gamma(\alpha)\right)x^{a-1} \exp\left\{-x/b\right\}$$

and

$$\log\, p(D|a,b) = (a-1)\sum_{i=1}^{n} \log x_i - nl \log \Gamma(a) - na \log b - (1/b)\sum_{i=1}^{n} x_i$$

respectively, where $\Gamma(\cdot)$ stands for the Gamma function.

Finally, the pdf for the Generalized Pareto (GP) distribution is given by

$$f(x|k,\sigma,\theta) = (1/\sigma)\left(1 + k\frac{(x-\theta)}{\sigma}\right)^{-1-\frac{1}{k}} \cdot$$

In what follows the MLE of the (3-parameter) GP is computed by means of the MATLAB package.

2.2 Non - Parametric Approach

The characteristics of IET and magnitudes are further investigated through a non-parametric tool, the so called multimodality and bump-hunt test, first introduced by Silverman (1986) and Efron and Tibshirani (1993/1998). This tool has been applied in mining seismicity (Lasocki, 2001; Lasocki and Orlecka - Sikora, 2008) and natural seismicity (Lasocki and Papadimitriou, 2006) after modifications. The method is based on the null hypothesis of the unimodality of the distribution, a characteristic of the most probabilistic distributions and is fully described by Lasocki and Papadimitriou (2006). The pdfs for the studied parameters are tested upon the existence and the number of modes or bumps. The terms "mode" and "bump" refer to a pdf's local maximum and an interval in which the density is convex within a specific interval limited by two inflexion points. There are two null hypotheses: H_0^1 assumes that the pdf is unimodal and H_0^2 assumes that only one bump exists to the right of the mode. The complexity of the distributions indicating multimodality is

evidenced when the significance of either of the null hypothesis $\left(H_1^0, H_2^0\right)$ is low. A pdf estimator for a series of magnitudes M_i is given by

$$\hat{f}\left(M \mid \{M_i\}, h\right) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{M - M_i}{h}\right),$$

where M_i is the observations series of n given data, h is a positive smoothing factor and $K(\cdot)$ corresponds to a kernel function which follows a Gaussian distribution. The estimated $\hat{f}\left(M \mid \{M_i\}, h\right)$ strongly depends on h since the existence of the complexities is a decreasing function of the smoothing factor. The smallest values of h, such that there is one mode or one bump, are the critical smoothing factors, $h_{crit(1)}$ and $h_{crit(2)}$. In order to determine the significance p_1 and p_2 (the respective p-values) for H_1^0 and H_2^0 hypotheses it is required to determine the critical smoothing factor, for these two hypotheses. Alternatively, the two significances can be estimated as:

$$P_{(1)} = [\textit{the number of unimodal } \hat{f}\left(M \mid \left\{M_i^{(k)}(1)\right\}, h_{cr}(1)\right),\ k=1,..,R]/R,$$

$$P_{(2)} = [\textit{the number of one–bump } \hat{f}\left(M \mid \left\{M_i^{(k)}(2)\right\}, h_{cr}(2)\right),\ k=1,..,R]/R,$$

where R is the number of the random samples that are produced after estimating the zeros of the 1st and 2nd derivative of $\hat{f}\left(M \mid \{M_i\}, h_{cr}(l)\right)$ (Lasocki and Papadimitriou, 2006). Bootstrap techniques are the most suitable ones for the resampling process despite the small seismicity sample.

3. APPLICATION AND RESULTS

3.1 Parametric Approach

The assessment of the "Goodness-of-Fit" in the parametric approach, is evaluated with the use of numerical and graphical means. The χ^2 test, the Kolmogorov-Smirnov test (K-S test) and the Anderson-Darling test (A-D test) testify the rejection of the null hypothesis for each distribution at a 5% significance level. The corresponding significance (p) is calculated when a test fails to reject the null hypothesis. The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are additionally employed for the justification of the statistical results. Finally, comparative graphical means (Q-Q plots) for relating empirical with theoretical quantiles, are constructed. Pairs of quantiles are plotted and interpreted in case that the distribution is verified for a 95% confidence interval and is expected to be linear. Estimations on the approximations of IET and magnitude distributions, are numerically summarized in Table 2, where estimates for the distribution parameters in each case, along with MLE values are given. A descriptive view of the empirical pdfs (ECDF) compared to the theoretical ones is schematically given in the sub plots of Figure 1 for all IET data sets.

For full IET data of Corinth gulf, Weibull distribution cannot be rejected according to the K-S test with a very small significance, while, Gamma distribution cannot be

rejected according to results of the χ^2 test with p_G=0.65. The information criteria exhibit slightly smaller value in the case of Weibull distribution (AIC=2439.46 and BIC=2451.60) compared to the respective Gamma values (AIC=2466.04 and BIC=2478.17).

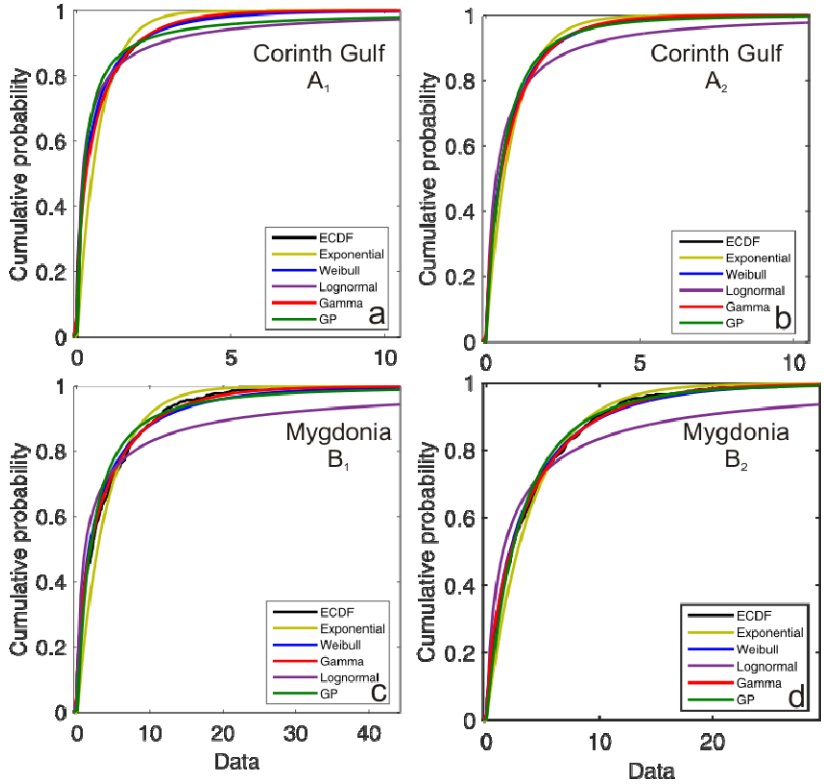


Figure 1. *Comparison for the interevent time ECDF, with the theoretical distributions for a) Corinth gulf full data b) Corinth gulf declustered data, c) Mygdonia full data d) Mygdonia declustered data.*

Table 2. *Estimations for the IET and magnitude distribution parameters for Corinth gulf and Mygdonia full and declustered catalogs*

		Distrib.	Parameters	95% confidence intervals	MLE
IET	A_1	Exp.	μ=0.72	[0.69, 0.74]	-2120.45
		Weibull	a=0.47, b= 0.59	[0.44, 0.50] [0.57, 0.60]	-1217.73
		Logn.	μ= -1.72, σ=2.11	[-1.79, -1.65] [2.06 , 2.16]	-1411.85
		Gamma	a=0.46, b= 1.54	[0.44, 0.48] [1.44, 1.65]	-1231.02
		GP	k= 0.99, σ=0.22, θ= 0	[0.90, 1.09] [0.20, 0.24]	-1550.97
	A_2	Exp.	μ=0.80	[0.77 0.83]	-2219.07
		Weibull	a=0.68, b= 0.76	[0.65, 0.72][0.74, 0.78]	-2025.24

		Logn.	μ=-1.13, σ= 1.71	[-1.20, -1.07] [1.67, 1.76]	-2314.69
		Gamma	a=0.65, b=1.22	[0.63, 0.68] [1.14, 1.30]	-2027.45
		GP	k=0.33, σ=0.55, θ= 0	[0.28, 0.39] [0.51, 0.59]	-2105.80
	B_1	Exp.	μ= 4.06	[3.75, 4.42]	-1343.24
		Weibull	a=2.95, b=0.61	[2.56, 3.40] [0.57, 0.65]	-1218.55
		Logn.	μ= 0.0705, σ=2.35	[-0.12, 0.26] [2.22, 2.49]	-1310.31
		Gamma	a=0.47, b=8.47	[0.43, 0.52] [7.26, 9.88]	-1204.66
		GP	k=0.57, σ=2.11, θ= 0	[0.39, 0.75] [1.74, 2.56]	-1301.61
	B_2	Exp.	μ= 3.96	[3.66, 4.31]	-1362.71
		Weibull	a=3.37, b=0.74	[3.01 3.78] [0.69 0.79]	-1318.51
		Logn.	μ= 0.39, σ=1.96	[0.23 0.55] [1.85 2.08]	-1425.16
		Gamma	a= 0.62, b= 6.36	[0.56, 0.68] [5.51, 7.33]	-1311.22
		GP	k=0.26, σ=2.96, θ= 0	[0.14, 0.39] [2.55, 3.43]	-1349.03
Magnitude	**A_1**	Exp.	μ=2.26	[2.24 2.28]	-
		Weibull	a= 2.78, b= 5.33	[2.76 2.80] [5.22 5.44]	-2774.51
		Logn.	μ=0.94, σ= 0.14	[0.93 0.94] [0.14 0.15]	-1791.71
		Gamma	a= 46.50, b= 0.05	[44.28 48.83] [0.05 0.06]	-1538.34
		GP	k=-0.13, σ=0.55, θ= 2.2	[-0.15, -0.10] [0.53, 0.57]	-890.90
	A_2	Exp.	μ=2.18	[2.14 2.17]	-
		Weibull	a= 2.80, b= 5.16	[2.78 2.83] [5.04 5.29]	-2103.22
		Logn.	μ=0.94, σ= 0.15	[0.94 0.95] [0.15 0.15]	-1408.61
		Gamma	a=39.49, b= 0.06	[37.49 41.59] [0.06 0.06]	-1505.55
		GP	k=-0.12, σ=0.57, θ= 2.1	[-0.15, -0.1] [0.54,0.60]	-917.830
	B_1	Exp.	μ=1.93	[1.90 1.97]	-
		Weibull	a= 2.36, b= 4.15	[2.31 2.41] [3.93 4.38]	-431.37
		Logn.	μ=0.75, σ= 0.19	[0.73 0.76] [0.18 0.20]	-300.63
		Gamma	a=24.60, b= 0.08	[21.90 27.63] [0.07 0.09]	-323.06
		GP	k=-0.13, σ=0.64, θ= 1.6	[-0.19, -0.07] [0.58, 0.71]	-234.94
	B_2	Exp.	μ=1.88	[1.84 1.92]	-
		Weibull	a= 2.27, b= 3.92	[2.22 2.32] [3.71 4.13]	-451.97
		Logn.	μ=0.70, σ= 0.20	[0.69 0.72] [0.19 0.22]	-322.428
		Gamma	a=21.46, b= 0.09	[19.13 24.07] [0.08 0.10]	-346.01
		GP	k=-0.13, σ=0.65, θ= 1.5	[-0.19, -0.07] [0.59, 0.72]	-255.58

For the declustered catalog it is assumed that even when dependent events are removed, Weibull and Gamma distributions still adapt with a good fit to the ECDF both implied by the results of the χ^2-test (p_W = 0.131 and p_G = 0.122), K-S test (p_W =0.645, p_G = 0.08) and the A-D test (p_W = 0.092, p_G = 0.317). Weibull values for AIC (4054.48) and BIC (4066.37) argue that there is an indication for a slightly better fit compared to Gamma distribution (AIC=4058.91 and BIC=4070.80). Q-Q plots are presented in sub plots of Figure 2 (2a, 2b, 2c and 2d) for the best fits, concerning Gamma distribution which shows a better fit than Weibull. For the two Mygdonia data sets, Gamma and Weibull distributions show a good fit in all cases (Figure 1a and 1b) for the IETs. For full data, Gamma distribution is fulfilling tests requirements at significance levels: p_G=0.052 (χ^2), p_G=0.417 (K-S) and p_G=0.239 (A-D). The

information criteria also demonstrate the lowest values, for the Gamma hypothesis (AIC=2413.33 and BIC=2421.98).

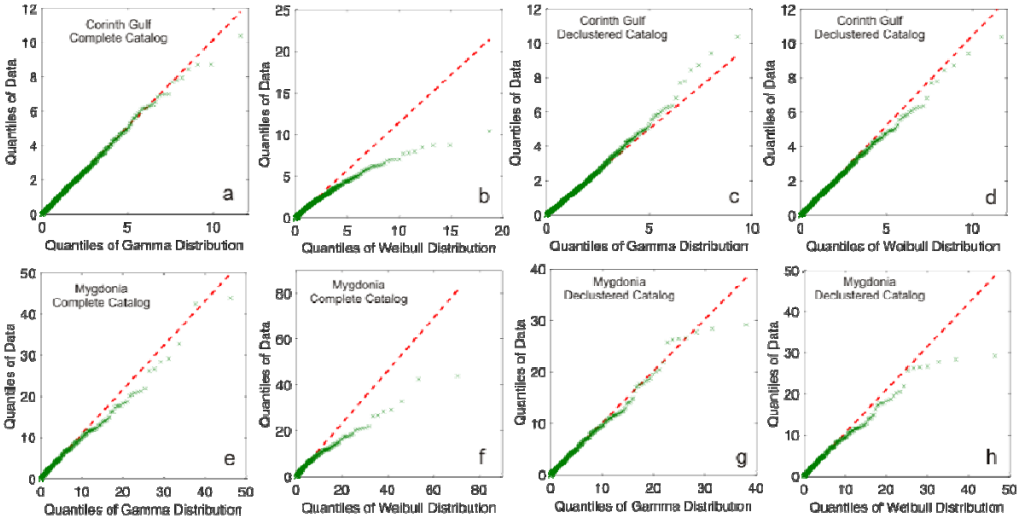


Figure 2. *Quantile –Quantile (Q-Q) diagrams for IET in the cases that a GF is implied by one of the tests, for a) Gamma quantiles for the full set of Corinth gulf, b) Weibull quantiles for the full set of Corinth gulf, c) Gamma quantiles for the declustered set of Corinth gulf, d) Weibull quantiles for the declustered set of Corinth gulf, e) Gamma quantiles for the complete set of Mygdonia, f) Weibull quantiles for the complete set of Mygdonia, g) Gamma quantiles for the declustered set of Mygdonia, h) Weibull quantiles for the declustered set of Mygdonia.*

In the case of the declustered catalogs, the distribution for IET of Mygdonia area shows that the observations vary from some minutes to 29.187 days. There is a good fit for Weibull, Gamma and the GP distributions. Gamma distribution shows a better approximation although all distributions fulfill χ^2, K-S and A-D statistical tests, the significance for which (in other words the respective p-values) are:

- for the Gamma distribution: $p_G = 0.368$, $p_G = 0.601$ and $p_G = 0.343$,
- for the Weibull distribution: $p_W = 0.135$, $p_W = 0.423$ and $p_W = 0.001$,
- for GP with θ=0: $p_{GP} = 0.076$, $p_{GP} = 0.005$ and $p_{GP} = 0.344$.

Q-Q plots are presented in Figure 2 for the two best approximations according to AIC and BIC criteria. Therefore the IET are described with the use of a Gamma theoretical distribution, with a= 0.62 and b= 6.36. Finally, comparative graphical means (Q-Q plots) for relating empirical quantiles with the theoretical distribution quantiles, are constructed (Figure 2e, 2f, 2g, 2h).

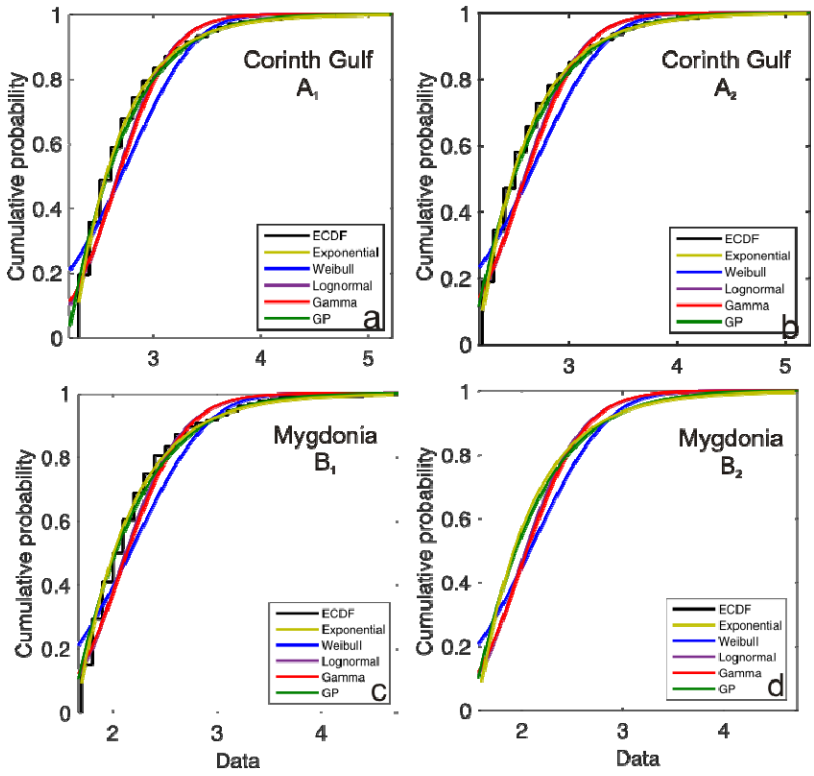


Figure 3. *Comparison for the Magnitude ECDF, with the theoretical distributions for a) Corinth gulf full data b) Corinth gulf declustered data, c) Mygdonia full data d) Mygdonia declustered data.*

The aforementioned distributions are testified for magnitudes (Figure 3) and results of parameter estimates are shown in Table 2. It should be mentioned that for the investigation of the exponential distribution for the magnitudes, Eq.4 has been used, where $M_{\min}$ and $M_{\max}$ are given in Table 1. For A_1 and A_2 data sets the lowest values correspond to the GP distribution (A_1: AIC=1787.81, BIC=1806.01, and A_2: AIC=1841.75, BIC=1859.52), however none of the used tests verify a good fit. In the case of Mygdonia full set (B_1), the lowest AIC and BIC are also met for the GP distribution, while in the declustered set (B_2), the good GP fit is also verified by a p_{GP}=0.297 significance (χ^2 -test).

3.2 Non-Parametric Approach

The non-parametric procedure was initially followed for the full data sets, however, in the case of rejecting the null hypothesis (when the estimated p-value was found significantly less than 0.15), 1000 bootstrap samples were compiled in each case for estimating in detail the significance of the null hypothesis. The rejection of the null hypothesis does not claim the validation of the alternative hypothesis. In the case that

the size distribution is complex and shows more than one bump or mode, the location of the minimum and maximum points found is significant since they provide information on the evolution of the seismicity processes and the prevalence of certain magnitudes or time lags. Two smoothing factors, h_{crit} and h_{critb} were calculated along with the corresponding significance level (Table 3).

Table 3. *Results for the verification of the null hypothesis for magnitude and IET distribution, where p_1 and p_2 stand for signifficance of the H_1^0, H_2^0 hypotheses. For the case of null hypothesis rejection the location of modes or bumps according to the test are given (NS: not significant result)*

Area	Cat	hcrit	p_1	$hcrit_b$	p_2	Locations		
						Modes	Bumps	
IET	A_1	0.690	0.534	1.230	0.40	N.S.	N.S.	
	A_2	1.00	0.240	1.463	**0.001**	N.S.	1.3	9.9
	B_1	0.510	0.302	1.731	0.25	N.S.	N.S.	
	B_2	0.414	0.261	10.269	**<0.001**	N.S.	6.08	10.30
Magnitude	A_1	0.378	**0.10**	0.434	**0.03**	5.9	2.9	6.2
	A_2	0.364	0.24	0.446	**0.06**	N.S.	2.6	4.3
	B_1	0.291	0.18	0.411	**0.04**	N.S.	2.4	4.4
	B_2	0.302	0.16	0.429	**0.03**	N.S.	2.3	4.3

In both data sets for Corinth gulf it is observed that both null hypotheses for unimodality are rejected and multiple modes and bumps are present in both cases (since p_1 and p_2 <0.15). Characteristic clustering of magnitudes is observed at small magnitudes in the case of the full catalog (locations of modes at M=2.6) as well as at the highest magnitude values located in the tail of the pdf distribution (M=6.2). The existence of bumps is indicated by the low significance of the test (p_2 =0.06). It is observed that in the case of the declustered catalog no modes are met but there are bumps at lower magnitudes compared to the full set. Bumps are met after the first modes and before the last mode of the largest magnitudes.

Results for magnitudes show that there is not an important indication for mode existence in Mygdonia, however, the significance for multi bumps is important (p_2 = 0.04) for all data sets. Further investigation for locations of zero first and second derivatives reveal that the locations of the bumps are located at magnitude values of 2.4 and 4.4. The bumps in the distribution correspond to lower magnitudes than expected according to G-R law. For the declustered catalog, the test signifies the existence of bumps, with a small probability of being wrong ($p_{(2)}$=0.03) stating that we cannot reject the hypothesis that the distribution is characterized by two bumps corresponding to magnitudes equal to 2.3 and 4.3.

4. DISCUSSION

The focus of our study is the exploitation of methodologies for revealing recent seismicity properties within two well-defined seismotectonic environments considered independent under the view of short-term and long-term interaction, from any adjacent zones. The study puts emphasis to the distribution of short IET and small to moderate magnitudes above a magnitude threshold for full and declustered seismicity catalogs. Accurate and complete data for the study period were constructed for revealing the parameters of microseismicity pattern for each area. The non-parametric tools chosen for this reason concern methodologies that they have been applied in regions with small magnitude and localized earthquakes such as mines that exhibit induced seismicity.

The parametric statistical analysis for the IET demonstrates that the full and declustered data sets do not show major contradictions and thus the shape of the distribution is not strongly influenced by aftershocks removal, in accordance with Davidsen and Kwiatek (2013). Weibull distribution is fitted better to the data of Corinth gulf, both complete and declustered, whereas Gamma distribution generally seems that is fitted better to the data set in the case of Mygdonia. However, in the case of Mygdonia, in addition with these two, GP distribution which is associated with extreme values shows a good fit. These exponential-like distributions reveal a general interrelation of IET times, also largely verified with the investigation of other statistical means (Gkarlaouni et al., 2014). Weibull distribution is also preferable for other data sources in Greece (Kourouklas et al., 2014). As far as magnitudes are concerned, it is indicated that none from the theoretical distribution fits well to the data, with the exception of the GP distribution.

The smoothed bootstrap test for multimodality, uncovers complex features of the aforementioned distribution with a good significance related with a multi–modes shape for the case of IET in declustered sets and for all cases in magnitudes. A bimodal magnitude distribution is ascertained in the case of Corinth gulf, showing a concentration of magnitudes around 5.9, before the maximum magnitude. In other cases multiple pumps are implied, exhibiting a magnitude deficit of magnitudes compared to the expected ones. The existence of bumps might be attributed to the short duration of the study.

ΠΕΡΙΛΗΨΗ

Οι κατανομές των μεγεθών και των ενδιαμέσων χρόνων μεταξύ διαδοχικών σεισμών αποτελούν αντικείμενο μελέτης, δεδομένου ότι αυτές εμπεριέχουν χαρακτηριστικά της σεισμικής διαδικασίας. Στην παρούσα εργασία, σε συνδυασμό με τη παραμετρική προσέγγιση της διερεύνησης της κατανομής επιχειρείται μια μη παραμετρική προσέγγιση, η λεγόμενη εξέταση της πολυτροπικότητας ή πολυκυρτότητας με στόχο τη διερεύνηση των κατανομών της σεισμικότητας σε δύο σεισμογενείς περιοχές της Ελλάδας. Η μέθοδος αυτή αρχικά χρησιμοποιήθηκε στην περίπτωση των μεγεθών, ενώ στην παρούσα μελέτη εφαρμόζεται και στην περίπτωση των ενδιαμέσων χρόνων

στις δύο περιοχές μελέτης, συγκεκριμένα τον Κορινθιακό Κόλπο και τη Μυγδονία λεκάνη, οι οποίες παρουσιάζουν έντονη σεισμική δράση με ισχυρούς και συχνούς σεισμούς. Τα δεδομένα παρατήρησης λήφθηκαν από τους σεισμολογικούς καταλόγους για το χρονικό διάστημα 2008-2014, και έγινε επεξεργασία ώστε να εμπεριέχουν σεισμούς μεγαλύτερους από το όριο πληρότητας. Επίσης υπέστησαν τη διαδικασία της από-ομαδοποίησης ή αποσυσταδοποίησης (de-clustering). Η δοκιμή βασίζεται στη μηδενική υπόθεση ότι η κατανομή παρουσιάζει μόνο μία επικρατούσα τιμή. Η παρουσία περισσότερων επικρατουσών τιμών ή περισσότερων σημείων όπου η συνάρτηση είναι κυρτή ενισχύεται όσο η p-τιμή προκύπτει μικρή. Στις περισσότερες των περιπτώσεων υπάρχει σημαντική ένδειξη για την πολυπλοκότητα της κατανομής με την παρουσία πολλαπλών μεγίστων και ομαδοποίησης των σεισμών. Οι θέσεις αυτές ανιχνεύτηκαν με τη βοήθεια της δοκιμής και αξιολογήθηκαν σύμφωνα με το εκάστοτε σεισμοτεκτονικό καθεστώς.

Acknowledgments:
1. We would like to thank Prof. G. Tsaklidis (Department of Mathematics, A.U.TH, Greece) for his helpful comments on the paper's content.
2. This research was co–financed by the European Union (European Social Fund–ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II Investing in knowledge society through the European Social Fund. Department of Geophysics, AUTh, Contribution number: 847/2015.

REFERENCES

Aki, K. (1965). Maximum likelihood estimate of b in the formula logN=a-bM and its confidence limits. *Bull. Earthquake Res. Inst.* Tokyo Univ, **43**, 237–239.

Altmann, E. and Kantz, H. (2005). Recurrence time analysis, long term correlations, and extreme events. *Phys. Rew. E.,* **71**, 056106.

Byrdina, S., Shebalin, P., Narteau, C. and Le Mouel, J. L. (2006). Temporal properties of seismicity and largest earthquakes in SE Carpathians. *Nonlin. Processes Geophys*., **13**, 629-639.

Corral, A. (2003). Local distributions and rate fluctuations in a unified scaling law for earthquakes. *Phys. Rev. E.*, **68**, 035 102, 2003.

Coral, A. (2006). Dependence of earthquake recurrence times and independence of magnitudes on seismicity history. *Tectonophysics*, **424**, 177–193.

Cornell, C. and Winterstein, S. (1988). Temporal and magnitude dependence in earthquake recurrence models. *Bull. Seismol. Soc. Am.*, **78**, 1522–1537.

Davidsen, J. and Kwiatek, G. (2013). Earthquake Interevent Time Distribution for Induced Micro-, Nano-, and Picoseismicity. *Phys.Rev.Lett.*, **110**, 068501.

Efron, B. and Tibshirani, R.J. (1993/1998). *An Introduction to the Bootstrap*. Chapman and Hall, London.

Gkarlaouni, C., Karakostas, V. and Tsaklidis, G. (2014). Investigation of earthquake interaction in terms of long and short term memory. *Proceedings of the 27th Hellenic Statistical Conference*, 53-67.

Gardner, J. K., and L. Knopoff (1974). Is the sequence of earthquakes in Southern California, with aftershocks removed, Poissonian? *Bull. Seis. Soc. Am.,* **64**(5), 1363–1367.

Gutenberg, B. and Richter, C. F. (1944). Frequency of earthquakes in California. *Bull. Seismol. Soc. Am.*, **34** 185–188.

Kagan, Y.Y. (1991). Likelihood analysis of earthquake catalogues. *Geophys. J. Int.,* **106**, 135-148.

Kijko, A., Lasocki, S. and Graham, G. (2001). Non-parametric Seismic Hazard in Mines. *Pure Appl. Geophys.,* **158**, 1655-1675.

Kourouklas, H., Papadimitriou, E. and Karakostas, V. (2014). Statistical distributions of earthquake recurrence times in central Ionian Islands (Greece). *Proceeding of the 27th Hellenic Statistical Conference*, 117-129.

Lasocki, S. (2001). Quantitative evidences of complexity of magnitude distribution in mining-induced seismicity: Implications for hazard evaluation, in 5th Int. Symp. Rockbursts and Seismicity in Mines "Dynamic rock mass response to mining" (G. van Aswegen, R. J. Durrheim, W. D. Ortlepp, eds.) SAIMM S27, Johannesburg, 543–550.

Lasocki, S. and Papadimitriou, E. (2006). Magnitude distribution complexity revealed in seismicity from Greece. *J. Geophys. Res.,* **111**, B11309, doi: 10.1029/2005JB003794.

Lasocki, S. and Orlecka–Sikora, B. (2008). Seismic hazard assessment under complex source size distribution of mining–induced seismicity. *Tectonophysics*, **456**, 28–37.

Leptokaropoulos, K., Karakostas, V., Papadimitriou, E. Adamaki, A. Tan, O. and Inan, S., (2013). A homogeneous earthquake catalogue compilation for western Turkey and magnitude completeness determination. *Bull. Seismol. Soc. Am.,* **103**, 2739–2751,

Main, I., (1996). Statistical physics, seismogenesis and seismic hazard. *Reviews of Geophysics,* **34**, 433 – 462.

Molchan, G. (2005). Interevent time distribution in seismicity: a theoretical approach. *Pure Appl. Geophys.*, **162**, 1135–1150.

Nishenko, S.P. and Buland, R. (1987). A generic recurrence interval distribution for earthquake forecasting. *Bull. Seismol. Soc. Am.,* **77** (4), 1382–1399.

Page, R. (1968). Aftershocks and microaftershocks of the Great Alaska Earthquake of 1964. *Bull. Seism. Soc. Am.,* **58**, 1131–1168.

Pisarenko, V.F. and Sornette, D. (2003). Characterization of Frequency of Extreme Earthquake

Events by the Generalized Pareto Distribution. *Pure Appl. Geophys.,* **160**, 2343-2364.

Pisarenko, V.F. and Sornette, D. (2004). Statistical Detection and Characterization of a Deviation from the Gutenberg-Richter Distribution above Magnitude 8. *Pure Appl. Geophys.,* **161**, 839-864.

Reasenberg, P. (1985). Second-order moment of central California seismicity, 1969–1982. *J. Geophys. Res.*, **90**, 5479– 5495.

Silverman B. W. (1986). *Density Estimation for Statistics and Data Analysis, Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Utsu, T. (1999). Representation and analysis of the earthquake size distribution: a historical review and some new approaches. *Pure Appl. Geophys*., **155**, 509–535.

Wesnousky, S.G. (1994). The Gutenberg-Richter or characteristic earthquake distribution, which is it? *Bull. Seismol. Soc. Am.*, **84**, 1940–1959.

Wiemer, S. and Wyss, M. (2000). Minimum magnitude of completeness in earthquake catalogs: examples from Alaska, the Western United States, and Japan. *Bull. Seismol. Soc. Am.*, **90**, 859–869.

Greek Statistical Institute
Proceedings of the 28th Panhellenic Statistics Conference (2015), pp. 400−413

ACCURACY OF BETAS BY USING A COMPARATIVE METHODOLOGY

D. G. Konstantinides[1] and G. C. Zachos[2]
[1]University of the Aegean
konstant@aegean.gr
[2]University of the Aegean
zachosg@aegean.gr

ABSTRACT

The Methodology that reduces implications occurring to price adjustment delays is studied. Namely the proposed technique examines if by confronting intervalling effect bias we have a positive effect to the accuracy of the risk estimations. For robust reasons, we filter our sample by excluding data that are suggested to provide noise and contribute to misleading results. In spite of the initial impression that intervalling effect free coefficients prove to be accurate, in fact by using models that take into account heteroskedasticity in residuals also maybe contribute to accuracy of risk estimators. We employ the proposed methodology in a sample from the Athens Stock Exchange and we put it to the test in a different market with similar features according to MSCI index provider. Through Vienna Stock Exchange we end up in some very useful conclusions in regard of some Greek market characteristics.

JEL classification: C22, G12, G14. *Keywords:* accuracy of betas, asymptotic risk estimators, adjusted risk coefficients, intervalling-effect bias, asymptotic beta.

1. INTRODUCTION

The beta concept has been used extensively by both practitioners and researchers. It quantifies the systematic risk of an investment yet its profound simplicity both in terms of theory and practice makes it a rather appealing methodology. A practitioner can measure the volatility of the investment, for instance a security, by regressing its returns to those of the market. The slope of the equation demonstrates how risky (volatile) is the investment. In addition, researchers are attempting to make models that will provide the most accurate risk coefficient possible. Researchers attempt to include other factors that capture risk or

confront biases. The most important models are the Capital Asset Pricing Model introduced by (Sharpe, 1964) and (Lintner, 1965), the Market Model, the three factors model, suggested by (Fama and French, 1993) and (Fama and French, 1995) and the four factors model introduced by (Carhart, 1997).By capturing as many risk components as possible a researcher is able to estimate up to which extend a market appears to be weak, semi-strong or strong form efficient.

We inspect the accuracy of the beta estimations that are suggested to be free of intevalling effect bias. Namely we investigate whether a methodology that reduces problems occurring to price adjustment delays due to microstructure of capital markets endows a positive effect to the accuracy of the risk estimations. For calculation of the betas that are free of the Intervalling effect we utilize the methodology of (Cohen *et al.*, 1983a) while in terms of accuracy inspection we are inspired from the technique suggested in (Blume, 1971), (Blume, 1975) and (Blume, 1979). Given the research interest of aforementioned methodologies, we mention the (Gordon and Chervany, 1980). The (Lusk and Koulayan, 2007) examine the Bloomberg Forecasting Heuristic, which is derived from Blume's work, as a functional model to forecast β. Unlike our work they used a sample from a mature market thus, 131 companies form S&P 500 and they proposed Bloomberg Forecasting Heuristic as a method of correcting risk estimators. Based on Blume's work we measure their accuracy. In parallel (Cwynar and Kasmierkiewicz, 2010) are using beta adjustment techniques in order to find the most suitable methodology for estimating risk factors with data from Warsaw Stock Exchange. Moreover the (Bian *et al.*, 2013) research effort on their modified maximum likelihood estimators is influenced by Blume's work as they are concerned on non-stationarity of betas. Finally Blume's technique was used in (Mantripragada, 1980) and recently in (Sarker, 2013). In regard with methodology of (Cohen *et al.*, 1983a), there are (Fung *et al.*, 1985), (Diacogiannis and Makri, 2008) and (Milonas and Rompotis, 2013). Moreover (Bian *et al.*, 2013) methodology should be compared to the one of (Cohen *et al.*, 1983).

Further in (Zachos, 2014) was suggested a comparative methodology approach as the betas that are free of intervalling effect were compared to the naïve and adjusted ones. Let remind that naïve and adjusted betas appeared as characterizations according to (Blume, 1971), (Blume, 1975). Adjusted are the betas that take into account the tendency to regress towards market mean. Naïve are the estimations that don't take that tendency into account. The suggested methodology differs in terms of how we use benchmark periods. While Blume employs benchmark and comparative periods of the same features, we suggest a different approach by choosing a stable benchmark period and alter the comparative ones. In Practical context we prefer such approach because Blume had only two sets of risk estimators to compare but our methodology involves more, for example OLS, OLS adjusted assessments, asymptotic estimation of betas, betas extracted according to models that take into account Heteroskedasticity in residuals and all

the above in daily and monthly intervals. In order to avoid to use adjusted assessments and asymptotic estimations of betas as benchmark periods, we examine which of those deviate less compared to their benchmark period.

The Initial evidence in (Zachos, 2014) is that the intervallig effect free estimators are more accurate with respect to naïve assessments in almost every occasion examined but less accurate compared to adjusted assessments. Furthermore, when we use as benchmark OLS and daily intervals in terms of observations and as comparison periods: a) OLS and Blume adjustments (daily int.) b) Asymptotic estimators and c) Naïve estimation (daily int.) intervalling effect betas are more accurate, compared to both naïve and adjusted assessments in two cases: when we don't take into account Corhay effect and when we do so. If the interval between data observations is more than a day we get a different result every time we choose a different starting day, as (Corhay, 1992) points out.

In this paper we attempt to make the initial findings more robust by filtering the sample in two ways: a) By excluding data that are not significant to the weight of the ASE. b) By taking into account R^2 of the regressions. In such way we expect to exclude data that provide noise or contribute to misleading results. Suggesting an intuitive interpretation we examine whether, certain defects existing in the market of ASE, bias our analysis.

Taking into consideration the above we: 1) Filter the sample and investigate the accuracy of the asymptotic estimators of betas by comparing them to OLS naïve assessments and to beta estimations adjusted according to Blume method. 2) Compare Blume's regression formulas with those in (Zachos, 2014). Those formulas serve in an intuitive sense as a proxy of the goodness of fit of the regressions in terms of how close are the prices of the slopes and intercepts we extract compared to the ones demonstrated in Blume's work. 3) Employ the comparison above for different intervals among data observations (daily and monthly intervals). 4) Re-examine the results after taking into account the Corhay effect. 5) Take into account Heteroskedasticity in residuals and proceed with the same accuracy inspection. 6) Employ it in another market besides the Athens Stock Exchange.

In the rest of the paper we have: Section 2 that includes Methodology Approach. Section 3 that discuses initial findings. Section 4 that analyzes reevaluation of the results depending on the filtering we perform each time. Section 5 that presents the results extracted from the Vienna Stock Exchange (VSE) and compares them with the ones from the Athens Stock Exchange (ASE). Section 6 that containes the conclusion.

2. METHODOLOGY APPROACH

We begin with the following formula of the Market Model:

$$E(r_i) = \alpha_i + \beta_i E(r_m)\,, \tag{2.1}$$

where $E(r_i)$ is the expected return of the capital asset, α_i is the residual return of asset i, $E(r_m)$ stands for the expected return of the market, β_i represents sensitivity of the asset returns compare to market returns or $\beta_i = \text{cov}(r_i, r_m)/\text{var}(r_m)$. The beta can be represented as the systematic risk of the portfolio. In Elton *et al.* (2011, p.152) was suggested that the Market Model lacks the assumption that all correlations among securities occur because of a common correlation with the market. Furthermore, we employ the methodology suggested by (Cohen *et al.*, 1983a), (Cohen *et al.*, 1983b) for the calculation of betas that are free of the intervalling effect bias. (Hawanini *et al.*, 1980) noted the importance of the friction in trading process and established its influence on the returns of an investor. In the first step we calculate the betas for a number of intervals, in particular from one day to one month.

$$r_{jLT} = a^1_{jL} + b^1_{jL}\, r_{MLT} + e^1_{jLT}\,, \tag{2.2}$$

In Second step we estimate the intervalling effect of risk coefficients. For this reason we obtain the regression the betas of first step (for all intervals) towards the monotonically decreased equation $f_j(l) = L^{-n}$, where it is assumed that:

$$\lim_{l\to\infty} f_j(l) = 0\,, \tag{2.3}$$

for any $j \in \mathbb{N}$. The monotonically decreased equation expresses the interval effect which reduces as intervals among observations increase. The formula of the second step is as follows:

$$b^1_{jL} = a^2_j + b^2_j\, L^{-n} + e^2_{jL}\,, \tag{2.4}$$

where a^2_j stands for the asymptotic estimator of beta. As L increases without bound the intervalling effect lessens and thus true beta will be converging observed ones. Let remind that true betas are these betas that should be obtained in case of a frictionless environment, while observed betas are the betas that can be calculated and actually observed by investors. For the exponent n we use the same methodology with (Cohen *et al.*, 1983a) and (Fung *et al.*, 1985). Apart from calculating betas free of intervalling effect we also examine their accuracy according to (Blume, 1971) and (Blume, 1975). Specifically, we calculate the future betas by taking into account the fact that betas are not constant over time but they have the tendency to fluctuate towards one. Practically, we divide dataset into two five year sub-periods and we calculate betas for each time period. For the first year period:

$$r_{i1} = a_{i1} + \beta_{i1}\, r_{m1} + e_{i1}\,, \tag{2.5}$$

For the second year period:

$$r_{i2} = a_{i2} + \beta_{i2}\, r_{m2} + e_{i2}\,, \tag{2.6}$$

Next we perform regressions to the betas of first period towards the betas of the second.

$$\beta_{i2} = A + \beta_{i1}\, B\,, \tag{2.7}$$

Afterwards we need the formula (2.7) and first period's betas and we calculate adjusted assessment of beta.

$$\beta_{iaa} = A + \beta_{i1}\, B\,, \tag{2.8}$$

Finally, we observe the accuracy of the assessments of the latter period that are based to historical data compared to the risk factors that take into account the tendency of the betas to fluctuate around the market mean. The methodology we select is Mean Square Error.

$$MSE = \frac{(\beta_{i2} - \beta_{i1})^2}{n}\,, \quad MSE = \frac{(\beta_{iaa} - \beta_{i1})^2}{n}\,. \tag{2.9}$$

3. INITIAL FINDINGS

The results from (Zachos, 2014) are presented in table 3 and 4 in appendix. The smallest Mean Square Error price denotes the more accurate estimation. Key finding appears when we use as benchmark OLS and daily intervals in terms of observations and as comparison periods: a) OLS and Blume adjustments (daily intervals) b) Asymptotic estimators and c) Naïve estimation (daily intervals). Namely free intervalling effect betas are more accurate compared to both naïve and adjusted assessments either we don't consider Corhay effect or when we consider Corhay effect. According to the Corhay effect we should adjust the regressions' outcomes by selecting every possible starting day within the interval of the sample. Finally the beta estimation is the average of these regression results.

In every other test we perform asymptotic estimators are more accurate compare to naïve assessments but less accurate compared to Blume's assessments.

4. RE-EVALUATING RESULTS

Initial results suggest that a methodology that reduces problems occurring to price adjustment delays, due to microstructure of capital markets, endows also a positive effect to the accuracy of the risk estimations. Besides these findings, there are also certain defects to be addressed. Specifically, we need to filter our sample firstly according to the Capitalization of ASE, secondly according to the R^2 of regression.

4.1 Considering ASE Capitalization

An important drawback when working with ASE concerns the index composition of 31/12/2011 trading day. In particular, the 99.996% of the index weight

corresponds to only 60 stocks, yet we are working with 224 stocks which is the whole market. Consequently, about 75% of the sample contributes nothing to the index weight and is uncorrelated to it. To that end we chose the stocks that actually contribute to the index weight and we evaluate the accuracy of risk estimators. Results are presented in table 5 and 6 in appendix and new Blume's regression formulas can be found in table 1. In detail the results suggest:

1. Blume's regression formulas include considerably larger figures in terms of both intercepts and slopes, which are more close to Blume's results. It suggests that first and second period's betas are more correlated to those of the initial study.
2. Compare to initial findings in (Zachos, 2014) our results are similar up to certain extent, yet they expose some differences: a) Intervalling effect free estimators are more accurate compare to naïve assessments and less accurate compared to adjusted assessments. b) There is no case where free intervalling effect betas are more accurate compare to both naïve and adjusted assessments. c) There is one case where naïve assessments are more accurate compare to free intervalling effect betas: when we work with daily intervals, as benchmark period we use GARCH and as comparison periods we also use GARCH methodology (with Corhay correction). Such result concludes that models that take into account heteroskedasticity in residuals contribute to the accuracy of the risk estimations.

4.2 Considering R^2

As mentioned in 4.1 the majority of the sample items appears to be uncorrelated to the market. Furthermore, uncorrelated time series are not expected to get good R^2 once regressed. As an example in (Zachos, 2014) R^2 mean is only 0.15 for the second period and daily intervals among observations. To that end we attempt to filter the sample according to R^2 of that occasion. For such analysis we filter our sample by excluding at first stocks that exhibit an R^2 mean less than 0.10 and at second less than 0.30.

The results concerning first case appear in table 7 and 8 in appendix while new Blume's regression formulas appear in table 2. Taking into account the points that we would like to address the results suggest:

1. Again Blume's regression formulas include considerably larger figures in terms of both intercepts and slopes.
2. Compare to initial findings, results are the same with a difference: a) Intervalling effect free estimators are more accurate compared to naïve assessments and less accurate compared to adjusted assessments. b) Again compare to initial findings there is no occasion where free intervalling effect betas are more accurate compare to both naïve and adjusted assessments.

Concerning the second case, results appear in table 9 and 10 in appendix and new Blume's regression formulas appear in table 2. The results suggest:

1. Once again Blume's regression formulas have intercepts and slopes more closer to the ones in Blume's work suggesting that such filtering contributes to robustness of results.
2. Intervalling effect free estimators are more accurate compared to naïve assessments and less accurate compared to adjusted assessments in every single comparison we perform.

5. VIENA STOCK EXCHANGE RESULTS

The results of Vienna Stock Exchange market appear in table 12 and 13 while Blume regression formulas appear in table 11. Concerning the Blume's formulas we notice that the figures of slopes and intercepts are much closer to the prices of Blume's work in comparison to the ones of ASE. Having in mind the fact that they serve as a proxy of the goodness of fit we realize that the sample from VSE appears to be more suitable compare to the one of the ASE.

In addition, there are differences compare to the results of ASE in many occasions: When we take into account the Corhay effect the results for the Intervalling effect free estimators are less accurate compared to naïve assessments and less accurate compare to adjusted assessments in every single comparison we perform. Without considering the Corhay effect the intervalling effect unbiased risk estimators also tend to perform equally poorly. Specifically Intervalling effect free estimators are less accurate compared to naïve assessments and less accurate compared to adjusted assessments in three out of six occasions. In the remaining three they are more accurate compared to naïve assessments and less accurate compared to adjusted assessments. In addition, there is not a single occasion where intervalling effect free betas outperform the adjusted assessments in terms of accuracy. Summing up there are two key findings: a) VSE as data provider appears to be more suitable compare to ASE. b) In regard of the accuracy of the intervalling effect unbiased risk estimators, they are less accurate than ASE findings. Similar to our work is the (Cwynar and Kasmierkiewicz, 2010) in terms they also examine the suitability of Warsaw Stock Exchange as a data provider. In addition, MSCI index provider categorizes both ASE and WSE as emerging and, they both exhibit similar characteristics like illiquidity. Additionally, in (Cwynar and Kasmierkiewicz, 2010) the results suggest an increase in the efficiency of WSE betas.

6. CONCLUDING REMARKS

In the paper at hand, we attempt to make the findings in (Zachos, 2014) robust by initially filtering our sample. Moreover, we compare Blume's regression formulas and we conclude that filtering our sample appears to be necessary, especially once working with sample of an emerging market with features like those of ASE. Initial findings in this paper are similar to the ones in (Zachos, 2014). In

addition: a) Evidence suggest that utilizing methodologies that take into account Heteroskdasticity in residuals, contribute to the accuracy of risk estimators. b) In a comparative pattern, the evidence in this paper suggests that asymptotic estimations of betas are accurate up to a certain extent, and they are not as accurate as initial findings demonstrated. Having in mind the necessity for data filtering, we employ the proposed methodology in another market with similar features. To this end, we chose a market with similar characteristics as Athens's Stock exchange market, which is the Vienna Stock Exchange. It is categorized as a mature market according to MSCI index provider, yet it exhibits similar features to ASE. For instance, it is not one of the strongest in capitalization terms. Moreover, we should mention the fact that ASE is downgraded to emerging market status in 2013, yet for the period 2001-2011 where the data were collected, ASE and VSE where both harmonized with mature markets standards. The Vienna Stock Exchange results exposed a different impression. Initially, they gave a different impression on the accuracy of the intervalling effect free betas. In addition, they highlighted the fact that ASE is most likely not an appropriate data provider.

Summing up, the findings of this paper highlight the necessity for future study. In particular: a) The methodology employed in (Zachos, 2014), should also be utilized to other markets with different features compared to the ones ASE and VSE markets expose. b) Apart from Blume's also Bayesian techniques could be included. for instance the technique inducted in (Vasicek, 1973) is used extensively by practitioners. Also in (Lam *et al.*, 2012) a pseudo-Bayesian model attempts to capture certain market behaviour biases. According to Elton *et al.* (2011, p.146), Bayesian techniques tend to perform slightly better in certain occasions.

ΠΕΡΙΛΗΨΗ

Στην εργασία αυτή εξετάζεται εάν η εφαρμογή μεθοδολογίας εκτίμησης συντελεστών κινδύνου που αντιμετωπίζει το φαινόμενο καθυστερημένης προσαρμογής των τιμών εξ αιτίας της μικροδομής των κεφαλαιακών αγορών, συμβάλει στην ακρίβεια των προαναφερθέντων συντελεστών. Καθώς χρειάζεται να εξετάσουμε εκ νέου τα αρχικά αποτελέσματα μας φιλτράρουμε το δείγμα που έχουμε από το ΧΑΑ (Χρηματιστήριο Αξιών Αθηνών) εξαιρώντας χρονοσειρές τιμών μετοχών οι οποίες παρέχουν θόρυβο και εν δυνάμει οδηγούν σε λάθος συμπεράσματα. Τα καινούργια αποτελέσματα ενισχύουν τα αρχικά έως ένα βαθμό. Επιπλέον προτείνουν ότι η χρήση μοντέλων που παίρνουν υπόψη την ετεροσκεδαστικότητα στα κατάλοιπα των παλινδρομήσεων συμβάλλουν στην ακρίβεια των εκτιμήσεων συντελεστών κινδύνου. Έχοντας υπόψη το γεγονός ότι χρησιμοποιούμε δείγμα από το ΧΑΑ, κάνουμε την ίδια μεθοδολογία και σε αγορά με παρόμοια χαρακτηριστικά. Επιλέγοντας το χρηματιστήριο Αξιών της Βιέννης και συγκρίνοντας τα αποτελέσματα που προκύπτουν με αυτά του ΧΑΑ καταλήγουμε σε κάποια συμπεράσματα αναφορικά με τα ιδιαίτερα χαρακτηριστικά του

Χρηματιστηρίου Αξιών Αθηνών.

REFERENCES

Bian, G., McAleer, M., Wong, W.(2013). Robust Estimation and Forecasting of the Capital Asset Pricing Model. *Annals of Financial Economics*, **8**, No. 2, Doi:10.1142/S201049513500073.

Blume, E. M.(1971.) On the Assessment of Risk. *The Journal of Finance*, **26**, No. 1, 1–10.

Blume, E. M.(1975). Betas and Their Regression Tendencies. *The Journal of Finance*, **30**, No. 3, 785–795.

Blume, E. M.(1979). Betas and Their Regression Tendencies: Some Further Evidence. *The Journal of Finance*, **34**, No. 1, 265–267.

Carhart, M. M.(1997). On Persistence in Mutual Fund Performance. *The Journal of Finance*, **52**, No. 1, 57–81.

Chan, C., Peretti de, C., Qiao, Z., Wong, W.(2012). Empirical Test of the Efficiency of the U.K. Covered Warrants Market: Stochastic Dominance and Likelihood Ratio Test Approach. *Journal of Empirical Finance*, **19**, 162–174.

Cohen, K., Hawanini, G., Maier, S., Schwartz, R., Whitecomb, D.(1983a). Estimating and Adjusting for the Intervalling Effect bias in Beta. *Management Science*, **29**, No. 1, 135–148.

Cohen, K., Hawanini, G., Maier, S., Schwartz, R., Whitecomb, D.(1983b). Friction in the Trading Process and the Estimation of Systematic Risk. *Journal of Financial Economics*, **12**, No. 2, 263–278.

Corhay, A.(1992). The intervalling effect bias in beta: A note. *Journal of Banking and Finance*, **16**, No. 1, 61–73.

Cwynar, W., Kazmierkiewicz, P.(2010). Testing the Maturity of the Polish Stock Market. *SSRN working papers series.*

Diacogiannis, G., Makri, P.(2008). Estimating Betas in Thinner Markets: The Case of The Athens Stoch Exchange. *International Research Journal of Finance and Economics*, **1**, No. 13, 108–123.

Elton, J. E., Gruber, J. M., Urich, J. T.(1978). Are Betas Best? *The Journal of Finance*, **33**, No. 5, 1375–1384.

Elton, J. E., Gruber, J. M., Brown, J. S., Goetzmann, N. W.(2011). *Modern Portfolio Theory and Investment Analysis*, Wiley, Eighth Edition.

Fama E., French K.(1993). Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics*, **33**, No. 1,3–56.

Fama E., French K.(1995). Size and Book-to-Market Factors in Earnings and Returns. *The Journal of Finance*, **50**, No. 1, 131–156.

Fung, W., Schwartz, R. and Whitecomb, D.(1985). Adjusting for the Intervalling Effect bias in Beta: A Test using Paris Bourse Data. *Journal of Banking and Finance*, **9**, No. 3, 443–460.

Gordon, A., Chervany, C.(1980). On the Estimation and Stability of Beta. *The Journal of Financial and Quantitative Analysis*, **15**, No. 1, 123–137.

Hawawini, G. A.(1980). Intertemporal Cross Dependence in Securities' Daily Returns and the Short-Run Intervalling Effect on Systematic Risk. *Financial Quantitave Anal.*, **15**, No. 1, 139–150.

Lam, K., Taisheng, L. Wing-Keung, W.(2012). A New Pseudo-Bayesian Model with Implications for Financial Anomalies and Investors' Behavior. *Journal of Behavioral Finance*, **13**, No. 2, 93—107.

Lintner, J.(1965). The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. *The Review of Economics and Statistics*, **47**, No. 1, 13–37.

Lusk, E., Koulayan, H.(2007). Forecasting β: An Evalution of the Bloumberg Heuristic. *Investment Management and Financial Innovations*, **4**, No. 1, 56–60.

Mantripragada, K.(1980). Beta Adjustment Methods. *Journal of Business Research*, **8**, No. 3, 329–339.

Milonas, T. N., Rompotis, G. G.(2013). Does Intervalling Effect affect ETF's? *Managerial Finance*, **39**, No. 9, 863–882.

Sarker, R. M.(2013). Forecast Ability of the Blume's and Vasicek's Technique: Evidence from Bangladesh. *Journal of Business and Management*, **9**, No. 6, 22–27.

Sharpe, F. W.(1964) Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk. *Journal of Finance*, **19**, No. 3, 425–442.

Vasicek, A. O.(1973). A note on Using Cross-Sectional information in Bayesian Estimation of Security Betas. *The Journal of Finance*, **28**, No. 5, 1233–1239.

Zachos, G.(2014). On the Accuracy of the Risk Estimators. *3rd SMTDA Conference Proceedings*, 851–863.

A appendix

Reg. Meth.	Init.Find.	Cap.
OLS D No Cor.	y=0.26+0.30x	y=0.38+0.44x
OLS M No Cor.	y=0.47+0.23x	y=0.74+0.16x
GARCH D. No Cor.	y=0.26+0.28x	y=0.31+0.47x
GARCH M. No Cor.	y=0.46+0.19x	y=0.64+0.17x
As.OLS No Cor.	y=0.43+0.24x	y=0.70+0.17x
As. GARCH No Cor.	y=0.38+0.27x	y=0.63+0.22x
As. EGARCH No Cor	y=0.37+0.25x	y=0.58+0.22x
OLS M Cor.	y=0.48+0.21x	y=0.72+0.15x
GARCH M Cor.	y=0.42+0.22x	y=0.63+0.18x
As. OLS Cor.	y=0.55+0.16x	y=0.77+0.11x
As.GARCH Cor.	y=0.50+0.18x	y= 0.75+0.12x
As EGARCH Cor.	y=0.46+0.18x	y=0.66+0.15x

Table 1. *Blume's Regression Formulas Init. Find. and Cap.*

Reg. Meth.	R^2 case 1	R^2 case 2
OLS D No Cor.	y=0.51+0.31x	y=0.07+0.81x
OLS M No Cor.	y=0.81+0.13x	y=0.40+0.39x
GARCH D. No Cor.	y=0.46+0.34x	y=0.05+0.81x
GARCH M. No Cor.	y=0.73+0.15x	y=0.35+0.38x
As.OLS No Cor.	y=0.76+0.16x	y=0.35+0.43x
As. GARCH No Cor.	y=0.73+0.17x	y=0.26+0.49x
As. EGARCH No Cor	y=0.69+0.17x	y=0.29+0.45x
OLS M Cor.	y=0.78+0.15x	y=0.38+0.40x
GARCH M Cor.	y=0.73+0.15x	y=0.34+0.38x
As. OLS Cor.	y=0.83+0.12x	y=0.47+0.32x
As.GARCH Cor.	y=0.80+0.12x	y=0.42+0.33x
As EGARCH Cor.	y=0.75+0.13x	y=0.38+0.33x

Table 2. *Blume's Regression Formulas filtering aording to* R^2

OLS Daily as Bench.	Adj. 07-11 0.286	Asymp. 07-11 0.271	Naive 07-11 0.388
OLS Monthly as Bench.	Adj. 07-11 0.257	Asymp. 07-11 0.368	Naive 07-11 0.379
GARCH Daily as Bench.	Adj. 07-11 0.264	Asymp. 07-11 0.274	Naive 07-11 0.373
GARCH Monthly as Bench.	Adj. 07-11 0.320	Asymp. 07-11 0.420	Naive 07-11 0.480
GARCH Daily as Benc.	Adj. 07-11 0.264	Asymp. 07-11 0.301	Naive 07-11 0.373
GARCH Monthly as Bench.	Adj. 07-11 0.320	Asymp. 07-11 0.453	Naive 07-11 0.480

Table 3. *MSE between Adj. Asymp. and Naive Betas (no Corhay)*

OLS Daily as Bench.	Adj. 07-11 0.286	Asymp. 07-11 0.254	Naive 07-11 0.388
OLS Monthly as Bench.	Adj. 07-11 0.375	Asymp. 07-11 0.467	Naive 07-11 0.487
GARCH Daily as Bench.	Adj. 07-11 0.260	Asymp. 07-11 0.271	Naive 07-11 0.373
GARCH Monthly as Bench.	Adj. 07-11 0.368	Asymp. 07-11 0.446	Naive 07-11 0.500
GARCH Daily as Bench.	Adj. 07-11 0.260	Asymp. 07-11 0.296	Naive 07-11 0.373
GARCH Monthly as Bench.	Adj. 07-11 0.368	Asymp. 07-11 0.494	Naive 07-11 0.500

Table 4. *MSE between Adj. Asymp. and Naive Betas (Corhay)*

OLS Daily as Bench.	Adj. 07-11 0.059	Asymp. 07-11 0.146	Naive 07-11 0.150
OLS Monthly as Bench.	Adj. 07-11 0.167	Asymp. 07-11 0.255	Naive 07-11 0.270
GARCH Daily as Bench.	Adj. 07-11 0.055	Asymp. 07-11 0.118	Naive 07-11 0.125
GARCH Monthly as Bench.	Adj. 07-11 0.223	Asymp. 07-11 0.284	Naive 07-11 0.336
GARCH Daily as Bench.	Adj. 07-11 0.055	Asymp. 07-11 0.115	Naive 07-11 0.125
GARCH Monthly as Bench.	Adj. 07-11 0.223	Asymp. 07-11 0.295	Naive 07-11 0.336

Table 5. *MSE between Adj. Asymp. and Naive Betas (No Corhay and Cap.)*

OLS Daily as Bench.	Adj. 07-11 0.059	Asymp. 07-11 0.146	Naive 07-11 0.150
OLS Monthly as Bench.	Adj. 07-11 0.225	Asymp. 07-11 0.317	Naive 07-11 0.319
GARCH Daily as Bench.	Adj. 07-11 0.042	Asymp. 07-11 0.119	Naive 07-11 0.117
GARCH Monthly as Bench.	Adj. 07-11 0.233	Asymp. 07-11 0.308	Naive 07-11 0.326
GARCH Daily as Bench.	Adj. 07-11 0.042	Asymp. 07-11 0.117	Naive 07-11 0.117
GARCH Monthly as Bench.	Adj. 07-11 0.223	Asymp. 07-11 0.320	Naive 07-11 0.326

Table 6. *MSE between Adj. Asymp. and Naive Betas (Corhay and Cap.)*

OLS Daily as Bench.	Adj. 07-11 0.099	Asymp. 07-11 0.127	Naive 07-11 0.159
OLS Monthly as Bench.	Adj. 07-11 0.157	Asymp. 07-11 0.215	Naive 07-11 0.237
GARCH Daily as Bench.	Adj. 07-11 0.082	Asymp. 07-11 0.120	Naive 07-11 0.131
GARCH Monthly as Bench.	Adj. 07-11 0.198	Asymp. 07-11 0.247	Naive 07-11 0.285
GARCH Daily as Bench.	Adj. 07-11 0.082	Asymp. 07-11 0.120	Naive 07-11 0.131
GARCH Monthly as Bench.	Adj. 07-11 0.198	Asymp. 07-11 0.258	Naive 07-11 0.285

Table 7. *MSE between Adj. Asymp. and Naive Betas (No Corhay and R^2 case 1)*

OLS Daily as Bench.	Adj. 07-11 0.099	Asymp. 07-11 0.128	Naive 07-11 0.159
OLS Monthly as Bench.	Adj. 07-11 0.216	Asymp. 07-11 0.270	Naive 07-11 0.278
GARCH Daily as Bench.	Adj. 07-11 0.079	Asymp. 07-11 0.131	Naive 07-11 0.132
GARCH Monthly as Bench.	Adj. 07-11 0.228	Asymp. 07-11 0.265	Naive 07-11 0.293
GARCH Daily as Bench.	Adj. 07-11 0.079	Asymp. 07-11 0.130	Naive 07-11 0.132
GARCH Monthly as Bench.	Adj. 07-11 0.228	Asymp. 07-11 0.265	Naive 07-11 0.293

Table 8. *MSE between Adj. Asymp. and Naive Betas (Corhay and R^2 case 1)*

OLS Daily as Bench.	Adj. 07-11 0.066	Asymp. 07-11 0.117	Naive 07-11 0.158
OLS Monthly as Bench.	Adj. 07-11 0.122	Asymp. 07-11 0.205	Naive 07-11 0.221
GARCH Daily as Bench.	Adj. 07-11 0.050	Asymp. 07-11 0.094	Naive 07-11 0.127
GARCH Monthly as Bench.	Adj. 07-11 0.150	Asymp. 07-11 0.197	Naive 07-11 0.277
GARCH Daily as Bench.	Adj. 07-11 0.050	Asymp. 07-11 0.010	Naive 07-11 0.127
GARCH Monthly as Bench.	Adj. 07-11 0.150	Asymp. 07-11 0.208	Naive 07-11 0.277

Table 9. *MSE between Adj. Asymp. and Naive Betas (No Corhay and R^2 case 2)*

OLS Daily as Bench.	Adj. 07-11 0.066	Asymp. 07-11 0.119	Naive 07-11 0.158
OLS Monthly as Bench.	Adj. 07-11 0.147	Asymp. 07-11 0.233	Naive 07-11 0.237
GARCH Daily as Bench.	Adj. 07-11 0.049	Asymp. 07-11 0.097	Naive 07-11 0.127
GARCH Monthly as Bench.	Adj. 07-11 0.165	Asymp. 07-11 0.213	Naive 07-11 0.252
GARCH Daily as Bench.	Adj. 07-11 0.049	Asymp. 07-11 0.104	Naive 07-11 0.127
GARCH Monthly as Bench.	Adj. 07-11 0.165	Asymp. 07-11 0.234	Naive 07-11 0.252

Table 10. *MSE between Adj. Asymp. and Naive Betas (Corhay and R^2 case 2)*

Reg. Meth.	ASE	VSE
OLS D No Cor.	y= 0.26 + 0.30 x	y= 0.27 + 0.66 x
OLS M No Cor.	y= 0.47 + 0.23 x	y= 0.66 + 0.18 x
GARCH D. No Cor.	y= 0.26 + 0.28 x	y= 0.20 + 0.74 x
GARCH M. No Cor.	y= 0.46 + 0.19 x	y= 0.56 + 0.31 x
As.OLS No Cor.	y= 0.43 + 0.24 x	y= 0.56 + 0.24 x
As. GARCH No Cor.	y= 0.38 + 0.27 x	y= 0.48 + 0.36 x
As. EGARCH No Cor	y= 0.37 + 0.25 x	y= 0.42 + 0.42 x
OLS M Cor.	y= 0.48 + 0.21 x	y= 0.59 + 0.24 x
GARCH M Cor.	y= 0.42 + 0.22 x	y= 0.49 + 0.38 x
As. OLS Cor.	y= 0.55 + 0.16 x	y= 0.74 + 0.16 x
As.GARCH Cor.	y= 0.50 + 0.18 x	y= 0.72 + 0.21 x
As EGARCH Cor.	y= 0.46 + 0.18 x	y= 0.62 + 0.27 x

Table 11. *Blume's Regression Formulas ASE and VSE*

OLS Daily as Bench.	Adj. 07-11 0.039	Asymp. 07-11 0.324	Naive 07-11 0.121
OLS Monthly as Bench.	Adj. 07-11 0.148	Asymp. 07-11 0.258	Naive 07-11 0.346
GARCH Daily as Bench.	Adj. 07-11 0.019	Asymp. 07-11 0.318	Naive 07-11 0.102
GARCH Monthly as Bench.	Adj. 07-11 0.120	Asymp. 07-11 0.231	Naive 07-11 0.310
GARCH Daily as Bench.	Adj. 07-11 0.019	Asymp. 07-11 0.261	Naive 07-11 0.102
GARCH Monthly as Bench.	Adj. 07-11 0.120	Asymp. 07-11 0.192	Naive 07-11 0.310

Table 12. *MSE between Adj. Asymp. and Naive Betas (No Corhay and VSE)*

OLS Daily as Bench.	Adj. 07-11 0.039	Asymp. 07-11 0.558	Naive 07-11 0.121
OLS Monthly as Bench.	Adj. 07-11 0.116	Asymp. 07-11 0.411	Naive 07-11 0.292
GARCH Daily as Bench.	Adj. 07-11 0.019	Asymp. 07-11 0.600	Naive 07-11 0.102
GARCH Monthly as Bench.	Adj. 07-11 0.077	Asymp. 07-11 0.409	Naive 07-11 0.208
GARCH Daily as Bench.	Adj. 07-11 0.019	Asymp. 07-11 0.442	Naive 07-11 0.102
GARCH Monthly as Bench.	Adj. 07-11 0.077	Asymp. 07-11 0.294	Naive 07-11 0.208

Table 13. *MSE between Adj. Asymp. and Naive Betas (Corhay and VSE)*

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής (2015), σελ.414-429

A FULLY ADAPTIVE CONTROL SCHEME FOR JOINT MONITORING OF LOCATION AND SCALE OF PROCESSES SUBJECT TO A MULTIPLICITY OF ASSIGNABLE CAUSES

Konstantinos A. Tasias and George Nenes
University of Western Macedonia, Department of Mechanical Engineering,
Bakola & Sialvera, 50100 Kozani, Greece
ktasias@uowm, gnenes@uowm.gr

ABSTRACT

In this paper a new fully adaptive control scheme for monitoring processes subject to a multiplicity of assignable causes is presented. The assignable causes may occur contemporarily and affect not only the mean but also the variance of a specific quality characteristic. A two-dimensional Discrete-Time Markov Chain (DTMC) is utilized to stochastically model the proposed scheme. In order to satisfy an acceptable statistical performance along with minimum total quality-related costs, an optimization problem is formulated. For a given economic and statistical data set of process characteristics, the optimal solution defines the optimum values of the design parameters of the scheme. The examination of a benchmark of examples leads to some utilitarian conclusions.

Keywords: Control Scheme, Adaptive Parameters, Markov Chain model, Multiple Assignable Causes, Economic-Statistical Optimization

1. INTRODUCTION

From the pioneering work of Shewhart (1931) until now, a wide variety of control charts have been utilized, as useful Statistical Process Control tools, for monitoring the quality of processes in an extensive variety of industries. Although the majority of control charts are designed to monitor only one process parameter, in many practice applications it is desirable to monitor simultaneously both the mean and the variance of a quality characteristic.

The benefit of reducing the quality-related costs combined with the demand for acceptable statistical performance from practitioners, urged many researchers to investigate the economic-statistical design of control charts. Serel and Moskowitz (2008) and Lu *et al*. (2013) presented the economic-statistical design of control charts for joint monitoring of process mean and variance.

The general conclusion in the literature that allowing the design parameters of a control chart to vary leads to an improvement of the performance of the chart has, also, been verified in $\overline{X}$ -s control charts. De Magalhães and Moura Neto (2005) and Tasias and Nenes (2012) presented fully adaptive control charts for monitoring location and scale of a process.

All cited references assume either one assignable cause whose occurrence may affect the mean and/or the standard deviation of the process or two independent assignable causes that affect the mean and/or the process variance. However, in many manufacturing scenarios, this simplified assumption is far from reality, because a process shift to the out-of-control condition may be the consequence of several assignable causes, which can occur at the same time or independently.

In the present paper, a fully adaptive control scheme for joint monitoring of the mean and variance is proposed, where multiple assignable causes affecting the mean and/or multiple assignable causes affecting the standard deviation of a process may occur independently, leading to a progressive deterioration of the process performance. The control scheme is economically and statistically optimized. A proper Markov chain approach is utilized for the stochastic model that describes the scheme's operation.

2. PROBLEM SETTING AND ASSUMPTIONS

A process is assumed to operate for an infinite horizon of time. A specific quality characteristic *X*, is assumed to be a continuous, normally distributed random variable with target mean μ_0 and target variance σ_0^2. A *Vp* $\overline{X} - s$ control scheme is utilized to monitor the process on-line. The process is assumed to have a perfect initial set up, starting its operation with mean and variance equal to μ_0 and σ_0^2, respectively. Multiple assignable causes may occur at random times affecting both the population mean and/or the standard deviation of the quality characteristic.

The occurrence of an assignable cause that affects the process mean does not prevent the occurrence of another assignable cause that may affect the standard deviation of the process and vice versa. Moreover, even if the process is out-of-control, either because the mean of the quality characteristic, its the standard deviation, or both have different values compared to their target ones, the occurrence of another assignable cause may deteriorate the process performance by shifting it to an inferior state. Namely, out-of-control operation consists of different levels of "bad-quality" performance, depending on the assignable causes that affect the process mean and/or the standard deviation of the process.

Let us consider *m* different assignable causes that may affect the process mean and *r* different assignable causes that may affect the standard deviation of the process. The occurrence of an assignable cause *i* that affects the process mean shifts the quality characteristic population mean away from μ_0 to $\mu_i = \mu_0 + \delta_i \cdot \sigma_0$. In a similar manner, the occurrence of an assignable cause *j* that affects the process standard deviation shifts the standard deviation from σ_0 to $\sigma_j = \gamma_j \cdot \sigma_0$. Obviously, by

considering $\delta_0 = 0$ and $\gamma_0 = 1$ as the in-control operation, there are m+1 possible values of δ with $\delta \in \{0, \delta_1, \delta_2, ..., \delta_{m-1}, \delta_m\}$ $0 < \delta_1 < \delta_2 < ... < \delta_{m-1} < \delta_m$ and r+1 possible values of γ with $\gamma \in \{1, \gamma_1, \gamma_2, ..., \gamma_{r-1}, \gamma_r\}$ $1 < \gamma_1 < \gamma_2 < ... < \gamma_{r-1} < \gamma_r$.

The time until the occurrence of an assignable cause is assumed to be an exponentially distributed random variable. This is a realistic assumption for many real applications such as mechanical or electric components with no wear, fatigue or corrosion during their expected life. For example, components of an aircraft radar, the hard disk of a computer etc.

In case the process operates under the effect of state i (i=0,1,...,m-1) and/or j (j=0,1,...,r-1) and because of the assumption that only transitions to inferior states may occur, the occurrence rate for a transition to any inferior state is also exponential and equal to $\lambda_{x_{ik}}$, $k = i+1,\ldots,m$ and $\lambda_{s_{jl}}$, $l = j+1,\ldots,r$.

The transition rates to any inferior state as regards: (a) the process mean, when the process operates under the effect of state i (i=0,1,...,m-1) (b) the standard deviation, when the process is under the effect of state j (j=0,1,...,r-1), are also exponential and equal to $v_{x_i} = \sum_{k=i+1}^{m} \lambda_{x_{ik}}$ and $v_{s_j} = \sum_{l=j+1}^{r} \lambda_{s_{jl}}$, respectively. Apparently, in the special case where $i=m$ ($j=r$) no further transition may occur and, so, $v_{x_m} = 0$ ($v_{s_r} = 0$).

Whenever an assignable cause occurs, the process remains under the effect of that assignable cause until the occurrence is detected and its effect is removed or until a new deterioration occurs and the process is shifted to an inferior state. It is obvious that the greater the shift size δ_i and/or γ_j of an assignable cause/es, the lower the quality output and/or the greater the operational costs. On the other hand, the greater the shift size δ_i and/or γ_j of an assignable cause/es, the lower the difficulty for the detection of the assignable cause/es by the control scheme.

The process is monitored by a control scheme determined by the values of the sample size n, the sampling interval h and the width of the warning w and control limits k, both for the chart that monitors the process mean and for the one that monitors the standard deviation of the process. These design parameters are allowed to vary at two different levels, a relaxed and tightened one.

Whenever the control scheme indicates a possible out-of control operation of the process, an inspection takes place, in order to reveal any possible assignable cause's effect. Then, if an assignable cause has indeed occurred, the process is certainly and perfectly restored to the in control state ($\mu = \mu_0$ and $\sigma = \sigma_0$). If no assignable cause has occurred, the process, obviously, remains in-control. Thus, after any signal that will reveal if a restoration is needed, the process will resume its operation from the in-control state.

The cost of lost production time for the process investigation and the cost of the time needed for the removal of an assignable cause are included in the costs of investigation and restoration, respectively. The time to search and remove an

assignable cause *i*, which affects the process mean, is denoted by T_{x_i} whereas the time for the search and removal of an assignable cause *j*, which affects the standard deviation of the process, is denoted by T_{s_j}. Finally, the time needed to reveal a false alarm is denoted by T_0. It is apparent that $T_0 \le T_{x_i}, T_{s_j}$ $i = 1,\ldots,m$ and $j = 1,\ldots,r$.

The cost elements that have an economic impact on the process and arise from the process monitoring can be divided into four different categories; (a) the sampling costs, which can be divided into a fixed cost per sample, denoted by *b* and a variable cost per sample unit denoted by *c*; (b) the false alarm cost, denoted by $L_{(0,0)}$, when the production is erroneously stopped for investigation of the process for possible assignable causes, (because no assignable cause has, indeed, occurred); (c) the cost per time unit for operation under the effect of assignable cause *i* and *j*, $i = 1,\ldots,m$ and $j = 1,\ldots,r$, denoted by $M_{(i,j)}$; (d) the cost for the restoration of the process to the in-control state by removing the effect of assignable causes *i* and *j* $(i, j > 0)$, denoted by $L_{(i,j)}$. It is apparent that the lower the effect of an assignable cause to the process mean and/or the standard deviation, the lower the out-of-control operation cost *M* and the cost of the removal of the assignable cause $L_{(i,j)}$.

3. MATHEMATICAL MODEL

Two parameter sets are employed for the sample size *n*, sampling interval *h*, upper warning limit coefficients w_x and w_s and control limit coefficients k_x and k_s, the *relaxed*, $\{n_1, h_1, w_{x_1}, w_{s_1}, k_{x_1}, k_{s_1}\}$ and the *tightened* one $\{n_2, h_2, w_{x_2}, w_{s_2}, k_{x_2}, k_{s_2}\}$, where $n_2 \ge n_1$, $h_2 \le h_1$, $w_{x_2} \le w_{x_1}$, $w_{s_2} \le w_{s_1}$, $k_{x_2} \le k_{x_1}$ and $k_{s_2} \le k_{s_1}$. At each sampling instance, the mean and the standard deviation of the collected sample are compared to the values of the respective warning and control limits in order for the following decisions to be made: a) whether to investigate the process or not, b) the value of the next sample's size - n_1 or n_2 -, c) the value of the next sampling interval - h_1 or h_2 -, d) the next warning and control limit coefficients for the chart that monitors the process mean - w_{x_1} or w_{x_2} and k_{x_1} or k_{x_2} - and e) the next warning and control limit coefficients for the chart that monitors the standard deviation of the process - w_{s_1} or w_{s_2} and k_{s_1} or k_{s_2}. Conclusively, there are twelve design parameters that fully define the proposed *Vp* $\bar{X}$ -s Shewhart control scheme: n_1 and n_2, h_1 and h_2, w_{x_1} and w_{x_2}, w_{s_1} and w_{s_2}, k_{x_1} and k_{x_2}, k_{s_1} and k_{s_2}.

A two-dimensional *DTMC* (Discrete-Time Markov Chain) (Y_t, a_t) is utilized, exploiting the assumption that the time for the occurrence of an assignable cause is an exponentially distributed random variable, to describe both the actual state of the

process and the decision made at each sampling instance. Specifically, the state of the process at any sampling instance t is denoted by Y_t and its possible values are: (a) $Y_t = (0,0)$, when no assignable cause has occurred; (b) $Y_t = (i,0)$, when an assignable cause i (i=1,…,m) that affects the process mean has occurred, leading to a shift size $\delta = \delta_i$, but the standard deviation is equal to its target value $(\sigma = \sigma_0)$; (c) $Y_t = (0,j)$, when an assignable cause j (j=1,..,r) that affects the standard deviation has occurred, leading to a shift size $\gamma = \gamma_j$, but the mean of the process is not affected $(\mu = \mu_0)$; (d) $Y_t = (i,j)$ i=1,...,m and j=1,...,r, when two assignable causes have contemporarily occurred and affect both the mean and the standard deviation of the process by shifting them to $\mu_i = \mu_0 + \delta_i \cdot \sigma_0$ and $\sigma_j = \gamma_j \cdot \sigma_0$, respectively.

The decision making procedure of the control scheme is based on the mean and the standard deviation of the collected sample at each sampling instance. There are three possible values for the variable a_t, that define the decision made. Specifically: a_t = 0, if both the standardized sample mean z_t and the sample's standard deviation s are below the respective threshold values (*central zone*), i.e, $z_t \le w_{x_{1or2}}$ and $s \le \left(c_{4,1or2} + w_{s_{1or2}}\sqrt{1-c_{4,1or2}^2}\right)\cdot\sigma_0$. In such a case, the process continues its operation and *relaxed* parameters $\left(n_1, h_1, w_{x_1}, w_{s_1}, k_{x_1}, k_{s_1}\right)$ are used for the next sampling. Moreover, a_t = 1, if either the standardized sample mean or the sample standard deviation or both of them, lie between the respective threshold and control limits (*warning zone*), but in none of the two charts is there an alarm signal. In particular: i) $w_{x_{1or2}} < z_t \le k_{x_{1or2}}$ and $s \le \left(c_{4,1or2} + k_{s_{1or2}}\sqrt{1-c_{4,1or2}^2}\right)\cdot\sigma_0$ or ii) $z_t \le w_{x_{1or2}}$ and $\left(c_{4,1or2} + w_{s_{1or2}}\sqrt{1-c_{4,1or2}^2}\right)\cdot\sigma_0 \le s \le \left(c_{4,1or2} + k_{s_{1or2}}\sqrt{1-c_{4,1or2}^2}\right)\cdot\sigma_0$. It should be noted that whenever a_t = 1 the decision for the process is to continue, but at the next sampling the *tightened* group of parameters $\left(n_2, h_2, w_{x_2}, w_{s_2}, k_{x_2}, k_{s_2}\right)$ should be used. Finally, a_t = 2, if the value of the chart statistic in at least one control chart outreaches the respective control limit (*action zone*), i.e., if either $z_t > k_{x_{1or2}}$ and/or $s > \left(c_{4,1or2} + k_{s_{1or2}}\sqrt{1-c_{4,1or2}^2}\right)\cdot\sigma_0$ .Then, the process is stopped for investigation and either a false alarm is discovered, or the process was actually out-of-control and is perfectly restored to the in-control condition. It should be noted that regardless of the control chart that issues the alarm, the investigation of the process reveals the effect of any possible assignable cause, that may affect the mean or the standard deviation,

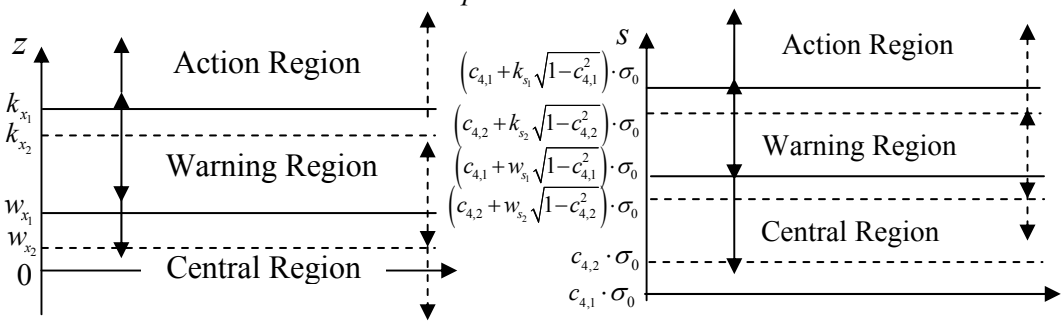
and the process is assumed to restart its operation from the in-control state. After this perfect set-up, relaxed parameters will be used, namely $\left(n_1, h_1, w_{x_1}, w_{s_1}, k_{x_1}, k_{s_1}\right)$.

It should be noted that:

$$c_{4,q} = \left(\sqrt{\frac{2}{n_q - 1}}\right) \cdot \frac{\Gamma\left(\frac{n_q}{2}\right)}{\Gamma\left(\frac{n_q - 1}{2}\right)} \quad \text{q} = 1,2.$$

The control policy for the two one-sided control charts is illustrated graphically:

Figure 1. *Regions of the two control charts for both relaxed and tightened parameters*



Based on the values of the two components that constitute the two-dimensional state, the Markov chain has $(m+1)\times(r+1)\times 3$ possible (Y_t,a_t) states for each possible combination of $Y_t = \left\{\left\{(0,0),(0,1)...,(0,r)\right\},\left\{(1,0),(1,1)...(1,r)\right\},...\left\{(m,0),(m,1)...(m,r)\right\}\right\}$ and a_t = 0,1,2, with the transition probabilities of the *DTCM* defined as follows:

$$P\left[a_t = v, Y_t = (k,l) \middle| a_{t-1} = u, Y_{t-1} = (i,j)\right] \; i,k \in [0,m], \; j,l \in [0,r], \; u,v{=}0,1,2$$

In order to compute the transition probabilities of the transition probability matrix P, every probability for the process moving from any state to any other state, must be first computed. Based on the approach described in Nenes *et al.* (2015) and because of the fact that the two sets of assignable causes affect independently the mean and the standard deviation of the process, the probability of a transition from state $i \geq 0$ to any other state $k \geq i+1$ and from state $j \geq 0$ to another state $l \geq j+1$, can be computed recursively from the following expression:

$$q_{\substack{xs_{ik} \\ jl}}(h) = q_{x_{ik}}(h) \cdot q_{s_{jl}}(h) =$$

$$= \left(\int_0^h \lambda_{x_{ik}} e^{-v_{x_i} t} \cdot e^{-v_{x_k}(h-t)} dt + \int_0^h \sum_{p=i+1}^{k-1} \lambda_{x_{ip}} e^{-v_{x_i} t} \cdot q_{x_{pk}}(h-t) dt\right) \cdot \tag{1}$$

$$\cdot \left(\int_0^h \lambda_{s_{jl}} e^{-v_{s_j} t} \cdot e^{-v_{s_l}(h-t)} dt + \int_0^h \sum_{q=j+1}^{l-1} \lambda_{s_{jq}} e^{-v_{s_j} t} \cdot q_{s_{ql}}(h-t) dt\right)$$

The probability of no transition, as regards the process mean, in *h* time units, is denoted by $q_{x_{ii}}(h)$, whereas, the respective probability for the standard deviation by $q_{s_{jj}}(h)$ ($q_{x_{ii}}\left(h\right)=e^{-v_{x_i}\cdot h}$, $q_{s_{jj}}\left(h\right)=e^{-v_{s_j}\cdot h}$). It is obvious that when $i = m$ and $j = r$, then $q_{x_{mm}}(h)=q_{s_{rr}}(h)=1$.

It should be noted that based on the assumption that X is a normally distributed random variable, the sample mean is also a normal variable $\overline{X}\sim N\left(\mu_i,\sigma_j^2/n\right)$ with the value of its mean and its variance depending on the actual state of the process. Moreover, in order to compute the probability of the sample's standard deviation to be in one of the central, warning or action zone, a simple transformation to variable $X^2=\left(n-1\right)\cdot s^2/\sigma_j^2$, which is a continuous random variable following a chi-square distribution with $n-1$ degrees of freedom, is necessary.

By denoting the upper warning and control limits of the control chart which monitors the standard deviation of the process by α_q $\left(\alpha_q=c_{4,q}+w_{s_q}\sqrt{1-c_{4,q}^2}\right)$ and β_q $\left(\beta_q=c_{4,q}+k_{s_q}\sqrt{1-c_{4,q}^2}\right)$, respectively, with $q=1$, in case relaxed parameters are used $(a_{t-1}=0,a_{t-1}=2)$, and $q=2$, if tightened parameters are utilized $(a_{t-1}=1)$, the probability of the decision made at each sampling instance, given that $Y_t=(k,l)$, is:

$$P(a_t=0)=\Phi\left(\frac{w_{x_q}-\delta_k\sqrt{n}q}{\gamma_l}\right)\cdot P\left(\chi^2_{n_q\text{-}1}<\frac{\alpha_q^2}{\gamma_l^2}\cdot\left(n_q\text{-}1\right)\right) \tag{2}$$

$$P(a_t=1)=\Phi\left(\frac{k_{x_q}-\delta_k\sqrt{n_q}}{\gamma_l}\right)\cdot P\left(\chi^2_{n_q\text{-}1}<\frac{\beta_q^2}{\gamma_l^2}\cdot\left(n_q\text{-}1\right)\right)- -\Phi\left(\frac{w_{x_q}-\delta_k\sqrt{n_q}}{\gamma_l}\right)\cdot P\left(\chi^2_{n_q-1}<\frac{\alpha_q^2}{\gamma_l^2}\cdot\left(n_q\text{-}1\right)\right) \tag{3}$$

$$P(a_t=2)=1-\Phi\left(\frac{k_{x_q}-\delta_k\sqrt{n_q}}{\gamma_l}\right)\cdot P\left(\chi^2_{n_q\text{-}1}<\frac{\beta_q^2}{\gamma_l^2}\cdot\left(n_q\text{-}1\right)\right) \tag{4}$$

The exact expressions for all the transition probabilities $P^{\,Y_{t-1}Y_t}_{a_{t-1}a_t}$ are equal to the product of the probability of the transition of the actual state of the process (Equation (1)) times the probability of the decision made at sampling instance t (Equations (2)-(4)).

The steady-state probabilities, which represent the long-term probability for the process being in state (Y_t, a_t) are denoted by $\pi_{Y_t a_t}$ and are computed by solving the following linear system:

$$\pi_{Y_t a_t} = \sum_{Y_{t-1}=(0,0)}^{(m,r)} \sum_{a_{t-1}=0}^{2} P_{\substack{Y_{t-1}Y_t \\ a_{t-1}a_t}} \cdot \pi_{Y_{t-1}a_{t-1}} \text{ and } \sum_{Y_t=(0,0)}^{(m,r)} \sum_{a_t=0}^{2} \pi_{Y_t a_t} = 1$$

As it is already mentioned, the assignable causes have an upward effect to the mean and/or the standard deviation of the quality characteristic. This is a realistic assumption for many real applications. For example, the Exhaust Gas Temperature (EGT) is an important measure of a jet engine's health and the occurrence of an/some assignable cause/es increases the mean and/or the standard deviation of this characteristic. Moreover, the occurrence of an assignable cause in the production of plastic parts by means of injection molding results in a shrinkage shift, which increases the mean and/or the standard deviation of mold temperature.

However, this assumption is not restrictive and can easily be negated by the model. It is worth noting that the proposed control scheme can be easily extended to cases where the effect of the assignable causes that affect the process mean, may be a downward shift of the mean $(\mu_i = \mu_0 + \delta_i \cdot \sigma_0, \delta_i < 0)$. In such cases, there are m+1 possible values of negative δ's with $\delta \in \{0, -\delta_1, -\delta_2, ..., -\delta_{m-1}, -\delta_m\}$ $0 > -\delta_1 > -\delta_2 > ... > -\delta_{m-1} > -\delta_m$ and the expressions of the transition probabilities are slightly different, due to the lower, instead of upper, warning $(-w_{x_1}, -w_{x_2})$ and control $(-k_{x_1}, -k_{x_2})$ limits utilized for the control chart that monitors the process mean. It is apparent that the value of each transition probability, either the assignable causes affect upwards or downwards the process mean, would be the same because of the symmetry of the normal distribution around μ_i. In a similar manner, an extension of the model to the rare cases where the model should detect a downward shift of the standard deviation can, also, easily, be made.

4. COST OF TRANSITION INTERVAL

The cost of a transition interval depends on the values of a_{t-1}, Y_{t-1}, Y_t and on the exact order of the assignable causes' occurrence within the interval, so as for the process to move from Y_{t-1} to Y_t.

In particular, in order to compute the expected out-of-control operation cost for a sampling interval, where $Y_{t-1} = (i, j)$ and $Y_t = (k, l)$, denoted by ECK, every possible combination/scenario of the chronological sequence of the assignable causes

that may have occurred within this interval in order for the process to move from state $Y_{t-1}=(i,j)$ to state $Y_t=(k,l)$, should be taken into account.

The out-of-control operation cost for each of the possible scenarios for a process transition is denoted by $CK(fn_1+fn_2)_{\substack{(i,j)\\(k,l)}}$. The variables fn_1 and fn_2 indicate the number of the assignable causes, in each scenario, that affect the mean and the standard deviation, respectively, in order for the process to move from state $Y_{t-1}=(i,j)$ to $Y_t=(k,l)$. The process transition from one state to another, during a sampling interval, may occur through more than one different ways as regards the number of the assignable causes. It is apparent that $fn_1 \le (k-i)$ and $fn_2 \le (l-j)$, ($fn_1, fn_2 \ge 0$).

For example, if we assume $Y_{t-1}=(0,0)$ and $Y_t=(2,2)$, the process may be shifted directly to state Y_t by the occurrence of assignable causes *i*=2 affecting the mean and *j*=2 affecting the standard deviation of the process $(fn_1=1, fn_2=1)$, with assignable cause *i* occurring earlier than *j* and vice versa. Another possible scenario is that assignable cause *i*=1 occurs first, then *i*=2 and finally *j*=2 $(fn_1=2, fn_2=1)$ or *j*=2, then *i*=1 and then *i*=2 and so on.

Consequently, for the computation of *CK*, the following parameters should be computed: a) the probability of a specific possible scenario when (fn_1+fn_2) assignable causes occur within the interval $(p_{(fn_1+fn_2)})$ b) the expected times of the occurrence of each assignable cause of a specific scenario in order to compute the time the process operates under each intermediate state ($\tau^{(1)}$ the precedent assignable cause, $\tau^{(2)}$ the second one,..., $\tau^{(fn_1+fn_2)}$ the last one to occur). The aforementioned parameters are computed from the following expressions (for the computation of their values, the occurrence rate λ should be substituted by λ_x, if the assignable cause affects the process mean and by λ_s, if it affects the standard deviation):

$$p_{(fn_1+fn_2)}=\int_0^h \lambda_1 e^{-\lambda_1 t_1}\int_{t_1}^h \lambda_2 e^{-\lambda_2 t_2}\ldots\int_{t_{[(fn_1+fn_2)-1]}}^h \lambda_{(fn_1+fn_2)} e^{-\lambda_{(fn_1+fn_2)}t_{(fn_1+fn_2)}}dt_{(fn_1+fn_2)}\ldots dt_2 dt_1$$

$$\tau^{(1)} = \frac{\int_0^h t_1 \lambda_1 e^{-\lambda_1 t_1} \int_{t_1}^h \lambda_2 e^{-\lambda_2 t_2} \ldots \int_{t_{[(fn_1+fn_2)-1]}}^h \lambda_{(fn_1+fn_2)} e^{-\lambda_{(fn_1+fn_2)} t_{(fn_1+fn_2)}} dt_{(fn_1+fn_2)} \ldots dt_2 dt_1}{\int_0^h \lambda_1 e^{-\lambda_1 t_1} \int_{t_1}^h \lambda_2 e^{-\lambda_2 t_2} \ldots \int_{t_{[(fn_1+fn_2)-1]}}^h \lambda_{(fn_1+fn_2)} e^{-\lambda_{(fn_1+fn_2)} t_{(fn_1+fn_2)}} dt_{(fn_1+fn_2)} \ldots dt_2 dt_1}$$

$$\tau^{(2)} = \frac{\int_0^h \lambda_1 e^{-\lambda_1 t_1} \int_{t_1}^h t_2 \lambda_2 e^{-\lambda_2 t_2} \ldots \int_{t_{[(fn_1+fn_2)-1]}}^h \lambda_{(fn_1+fn_2)} e^{-\lambda_{(fn_1+fn_2)} t_{(fn_1+fn_2)}} dt_{(fn_1+fn_2)} \ldots dt_2 dt_1}{\int_0^h \lambda_1 e^{-\lambda_1 t_1} \int_{t_1}^h \lambda_2 e^{-\lambda_2 t_2} \ldots \int_{t_{[(fn_1+fn_2)-1]}}^h \lambda_{(fn_1+fn_2)} e^{-\lambda_{(fn_1+fn_2)} t_{(fn_1+fn_2)}} dt_{(fn_1+fn_2)} \ldots dt_2 dt_1}$$

$$\tau^{(fn_1+fn_2)} = \frac{\int_0^h \lambda_1 e^{-\lambda_1 t_1} \int_{t_1}^h \lambda_2 e^{-\lambda_2 t_2} \ldots \int_{t_{[(fn_1+fn_2)-1]}}^h t_{(fn_1+fn_2)} \lambda_{(fn_1+fn_2)} e^{-\lambda_{(fn_1+fn_2)} t_{(fn_1+fn_2)}} dt_{(fn_1+fn_2)} \ldots dt_2 dt_1}{\int_0^h \lambda_1 e^{-\lambda_1 t_1} \int_{t_1}^h \lambda_2 e^{-\lambda_2 t_2} \ldots \int_{t_{[(fn_1+fn_2)-1]}}^h \lambda_{(fn_1+fn_2)} e^{-\lambda_{(fn_1+fn_2)} t_{(fn_1+fn_2)}} dt_{(fn_1+fn_2)} \ldots dt_2 dt_1}$$

The out-of-control operation cost per time unit after the effect of the first assignable cause that occurs is denoted by $M^{(1)}$, the second one by $M^{(2)}$ and so on.

Consequently, the out-of-control operation cost (CK) of a specific scenario can be, now, computed as:

$$CK(fn_1+fn_2) = p_{(fn_1+fn_2)} \left(M_{(i,j)} \tau^{(1)} + M^{(1)} \left(\tau^{(2)} - \tau^{(1)} \right) + \ldots + M^{(fn_1+fn_2)} \left(h - \tau^{(fn_1+fn_2)} \right) \right)$$

The expected out-of-control operation cost (*ECK*) for the process transition from a specific state to another specific state, is defined as the sum of the out-of-control operation costs (*CK*) for each possible scenario as regards each possible permutation of the assignable causes that occur within the interval and result in the transition.

The number of all possible permutations when a total number of $(fn_1 + fn_2)$ assignable causes occur within a sampling interval, with fn_1 assignable causes affecting the process mean and fn_2 affecting the standard deviation, is denoted by c, should be taken into account for the computation of *ECK*, and are equal to:

$$c = \binom{(fn_1+fn_2)}{fn_1} = \binom{(fn_1+fn_2)}{fn_2} = \frac{(fn_1+fn_2)!}{fn_1!\,fn_2!}$$

For example, by assuming that $fn_1 = 2$ and $fn_2 = 2$, there are six possible permutations as regards the chronological sequence of each assignable cause's occurrence $c = (2+2)!/(2!2!) = 6$. If the first assignable cause that occurs within the interval and affects the process mean is denoted by $x^{(1)}$, the second one by

$x^{(2)}$ and by $s^{(1)}$, $s^{(2)}$ the respective assignable causes that affect the standard deviation, then the six possible permutations are: $\left(x^{(1)}, x^{(2)}, s^{(1)}, s^{(2)}\right)$, $\left(x^{(1)}, s^{(1)}, x^{(2)}, s^{(2)}\right)$, $\left(x^{(1)}, s^{(1)}, s^{(2)}, x^{(2)}\right)$, $\left(s^{(1)}, s^{(2)}, x^{(1)}, x^{(2)}\right)$, $\left(s^{(1)}, x^{(1)}, x^{(2)}, s^{(2)}\right)$ and $\left(s^{(1)}, x^{(1)}, s^{(2)}, x^{(2)}\right)$.

It is apparent that, due to the assumption that only transitions to inferior states may occur, there is only one combination for the exact order of the assignable causes' occurrence when they either affect only the mean $\left(fn_1 > 0, fn_2 = 0\right)$ or only the standard deviation $\left(fn_1 = 0, fn_2 > 0\right)$ of the process (c=1). In such case, the expected out-of-control operation cost of a transition interval *ECK* equals the out-of-control operation cost (*ECK=CK*).

Based on the expected out-of-control operation cost of a transition interval *ECK*, described above, the expected cost of the process operation under the effect of an/some assignable cause/es can be computed. An analytic presentation of the exact computation of this cost is presented in Tagaras and Lee (1988).

The expected out-of-control operation cost when the process is under the effect of assignable causes (i, j) $i = 0,...,m$, $j = 0,...,r$, at the beginning of a sampling interval, is denoted by $K_{(i,j)}(h)$ and can be computed as the sum of the following three terms:

(i) the cost if no assignable cause occurs within the interval

(ii) the cost in case either only the mean or only the standard deviation of the process is affected

(iii) the cost when both the mean and the standard deviation of the process are shifted from their initial values

In general, the expected out-of-control operation cost for every possible initial state when up to *m* and *r* assignable causes may occur, can be derived from the following expression:

$$K_{(i,j)}(h) = M_{(i,j)} h e^{-\left(v_{x_i} + v_{s_j}\right)h} + \left(\begin{array}{l} e^{-v_{s_j} h} \cdot \int_0^h \sum_{k=i+1}^{m} \left[\lambda_{x_{ik}} e^{-v_{x_i} t} \left(tM_{(i,j)} + K_{(k,j)}(h-t) \right) \right] dt + \\ + e^{-v_{x_i} h} \cdot \int_0^h \sum_{l=j+1}^{r} \left[\lambda_{s_{jl}} e^{-v_{s_j} t} \left(tM_{(i,j)} + K_{(i,l)}(h-t) \right) \right] dt \end{array} \right) +$$

$$+ \sum_{k=i+1}^{m} \sum_{l=j+1}^{r} \left[\sum_{fn_1=1}^{k-i} \sum_{fn_2=1}^{l-j} ECK\left(fn_1 + fn_2\right)_{(k,l)}^{(i,j)} \right]$$

5. THE ECONOMIC MODEL AND STATISTICAL MEASURES

From the computation of the steady-state probabilities of the process and the computation of the mean cost of out-of-control operation, the average cost of a

transition step, denoted by *EC* and the average duration of a transition step, denoted by *ET*, can be evaluated. The long-run average cost per time unit, denoted by *ECT*, equals the ratio of *EC* over *ET*.

The expressions for the computation of the aforementioned measures are:

$$EC = \sum_{Y_t=(0,0)}^{(m,r)} \sum_{a_t=0}^{2} \pi_{Y_t a_t} \cdot C_{Y_t a_t} \quad ET = \sum_{Y_t=(0,0)}^{(m,r)} \sum_{a_t=0}^{2} \pi_{Y_t a_t} \cdot T_{Y_t a_t}$$

In more detail, the average cost of a transition step can be derived from the following expression:

$$EC = b + \sum_{k=0}^{m}\sum_{l=0}^{r} \pi_{(k,l)0} \cdot \left(cn_1 + K_{(k,l)}\left(h_1\right)\right) + \sum_{k=0}^{m}\sum_{l=0}^{r} \pi_{(k,l)1} \cdot \left(cn_2 + K_{(k,l)}\left(h_2\right)\right) +$$
$$+ \sum_{k=0}^{m}\sum_{l=0}^{r} \pi_{(k,l)2} \cdot \left(cn_1 + K_{(0,0)}\left(h_1\right) + L_{(k,l)}\right)$$

It should be noted that for $(k,l)=(0,0)$, $L_{(k,l)}$ represents the cost of a false alarm $\left(L_{(0,0)}\right)$. In a similar manner, the average duration of a transition step equals the sum of the relaxed (h_1) or tightened (h_2) duration of a sampling interval, plus the time to search and remove an/some assignable cause/es, if needed, multiplied by the respective long-run probabilities for each decision:

$$ET = h_1 \cdot \sum_{k=0}^{m}\sum_{l=0}^{r} \pi_{(k,l)0} + h_2 \cdot \sum_{k=0}^{m}\sum_{l=0}^{r} \pi_{(k,l)1} + \left(T_0 + h_1\right) \cdot \pi_{(0,0)2} +$$
$$+ \sum_{k=1}^{m} \pi_{(k,0)2} \cdot \left(T_{x_k} + h_1\right) + \sum_{l=1}^{r} \pi_{(0,l)2} \cdot \left(T_{s_l} + h_1\right) + \sum_{k=1}^{m}\sum_{l=1}^{r} \pi_{(k,l)2} \cdot \left(T_{xs_{kl}} + h_1\right)$$

It should be noted that the in-control average run length (ARL_0) and the weighted out-of-control average run length (WARL) are utilized as measures of the statistical performance of the proposed scheme. Their values are derived from the following expressions:

$$ARL_0 = \frac{\pi_{(0,0)0} + \pi_{(0,0)1} + \pi_{(0,0)2}}{\pi_{(0,0)2}}, \; WARL = \sum_{Y_t \neq (0,0)} \left(\frac{1 - \left(\pi_{(0,0)0} + \pi_{(0,0)1} + \pi_{(0,0)2}\right)}{\pi_{Y_t 2}} \right)$$

6. OPTIMIZATION PROBLEM

Let DP be the design parameters vector $DP = \left\{n_1, n_2, h_1, h_2, w_{x_1}, w_{x_2}, k_{x_1}, k_{x_2}, w_{s_1}, w_{s_2}, k_{s_1}, k_{s_2}\right\}$. The goal of the proposed scheme is to find the optimal DP in order to minimize the expected cost per time unit ECT (ECT=EC/ET) along with a constraint in WARL to be less or equal to five time units $\left(WARL \leq 5\right)$ and a lower bound for the in-control average run length ARL_0 to

be greater or equal to one hundred time units $\left(ARL_0 \geq 100\right)$. The aforementioned constraints are put in order to satisfy acceptable statistical performance of the proposed scheme.

The optimization problem is formulated as follows:

$$\min_{DP}(ECT)$$

$$\text{s.t} \quad WARL \leq 5,\ ARL_0 \geq 100,$$

$$DP > 0, n_2 \geq n_1 \geq 2,\ h_2 \leq h_1,\ w_{x_2} \leq w_{x_1}, k_{x_2},\ k_{x_1} \geq w_{x_1}, k_{x_2},\ w_{s_2} \leq w_{s_1}, k_{s_2},\ k_{s_1} \geq w_{s_1}, k_{s_2}$$

7. NUMERICAL ANALYSIS

In this section a numerical investigation is performed in order to explore the statistical performance and the potential cost savings of the proposed control scheme. The aforementioned problem for computing the optimum design parameters is applied to a benchmark of scenarios, which are defined by ten parameters that are allowed to vary at two levels, a *high* and a *low* one. Two assignable causes that affect the process mean are possible to occur and two assignable causes are possible to shift the standard deviation from its target value $\left(m = r = 2\right)$. Specifically, the ten parameters and their possible values are: $\lambda \in (0.01;0.1)$, $\delta \in (0.5;1.5)$, $\gamma \in (1.4;2.0)$, $\mathrm{T}_0 \in (0.0167;0.1)$, $\mathrm{T} \in (0.167;0.5)$, $b \in (0;5)$, $c \in (1;10)$, $M \in (100;1000)$, $L_{(0,0)} = L_0 \in (100;200)$ and $L_{(i,j)} = L_1 \in (200;400)$.

It should be mentioned that $\lambda_{q,q+1} = \lambda$ and $\lambda_{q,q+2} = \lambda / 2$ where q an assignable cause that affects either the mean or the standard deviation of the process. Moreover, $\delta_1 = \delta$, $\delta_2 = 1.5\delta$, $\gamma_1 = \gamma$ and $\gamma_2 = 2\gamma - 1$. As regards the time to search and remove an assignable cause $\mathrm{T}_{x_1} = T_{s_1} = T/2$ and $\mathrm{T}_{x_2} = T_{s_2} = T$. Finally, $M_{(0,1)} = M_{(1,0)} = M$, $M_{(0,2)} = M_{(2,0)} = 1.5M$, $M_{(i,j)} = 0.75(M_{(i,0)} + M_{(j,0)})$, $L_{(0,1)} = L_{(1,0)} = L_1$, $L_{(0,2)} = L_{(2,0)} = L_1 + 50$ and $L_{(i,j)} = 0.75(L_{(i,0)} + L_{(j,0)})$. Finally, the process is assumed not to operate during search and repair. It should be noted that the presumed assumption is not restrictive and can easily be negated by the model.

Table 1. *Economic and statistical data sets of examined cases*

Case	λ	δ	γ	T_0	T	b	c	M	L_0	L_1
1	0.01	1.5	1.4	0.0167	0.167	0	1	1000	100	200
2	0.01	0.5	2	0.0167	0.167	0	1	100	200	200
3	0.01	0.5	1.4	0.1	0.167	0	1	100	100	400
4	0.1	0.5	1.4	0.1	0.167	0	1	1000	200	200
5	0.01	1.5	2	0.1	0.167	0	1	1000	200	400
6	0.1	1.5	2	0.1	0.167	0	1	100	100	200
7	0.01	1.5	1.4	0.0167	0.5	0	1	1000	200	400
8	0.1	1.5	1.4	0.0167	0.5	0	1	100	100	200
9	0.01	0.5	2	0.0167	0.5	0	1	100	100	400

10	0.1	0.5	2	0.0167	0.5	0	1	1000	200	200
11	0.01	1.5	2	0.1	0.5	0	1	1000	100	200
12	0.1	1.5	2	0.1	0.5	0	1	100	200	400
13	0.01	1.5	2	0.0167	0.167	5	1	1000	100	400
14	0.1	1.5	2	0.0167	0.167	5	1	100	200	200
15	0.01	1.5	1.4	0.0167	0.167	5	1	1000	200	200
16	0.1	0.5	1.4	0.1	0.5	5	1	1000	200	400
17	0.01	1.5	2	0.1	0.5	5	1	1000	200	200
18	0.1	1.5	2	0.1	0.5	5	1	100	100	400
19	0.01	0.5	2	0.1	0.5	5	1	100	200	400
20	0.1	1.5	2	0.1	0.5	5	1	1000	200	400

Based on the conclusion of Park and Reynolds (1999) that the use of more than one warning limits leads to a small cost reduction, only one upper warning limit is utilized for each control chart, i.e. $w_{x_1} = w_{x_2} = w_x$ and $w_{s_1} = w_{s_2} = w_s$.

The solution to the optimization problem for every examined case, defines the optimum design parameters and the optimal values of ECT, WARL and ARL_0. The results are presented in Table 2.

Table 2. *Optimum design parameters and optimal solutions of examined cases*

Case	h_1	h_2	n_1	n_2	wx	kx_1	kx_2	ws	ks_1	ks_2	ECT	ARL_0	WARL
1	0.3	0.1	3	11	1.8	3.0	2.9	1.1	3.5	2.0	40.37	129.87	2.94
2	1.5	0.1	4	14	0.9	3.5	2.1	2.1	4.1	3.6	18.15	238.10	3.00
3	0.1	0.1	2	3	1.8	3.0	1.8	3.6	4.0	3.7	21.10	256.41	2.07
4	0.1	0.1	3	20	1.1	3.2	2.1	1.5	3.8	2.3	113.91	136.99	3.40
5	0.1	0.1	2	9	2.2	3.6	3.0	2.8	5.2	3.1	22.03	3333.33	3.29
6	0.1	0.1	2	4	2.1	3.3	2.4	3.1	4.6	3.4	45.54	909.09	3.65
7	0.3	0.1	3	16	1.9	3.3	3.2	1.3	4.1	2.2	47.38	312.50	3.10
8	0.2	0.1	2	5	1.7	3.0	2.5	1.6	4.0	2.1	78.00	166.67	4.48
9	1.3	0.1	3	10	0.8	3.2	1.8	1.9	4.0	3.3	23.17	103.09	3.19
10	0.1	0.1	3	16	1.0	3.2	1.9	2.4	3.9	3.3	157.77	166.67	2.79
11	0.1	0.1	2	7	2.1	3.3	2.7	2.7	4.7	3.0	18.73	1000.00	3.13
12	0.1	0.1	2	2	2.1	3.8	2.5	3.8	4.1	3.9	71.40	769.23	4.98
13	0.7	0.1	5	9	1.8	3.0	2.9	1.5	3.3	2.7	44.42	270.27	1.48
14	0.8	0.1	5	8	1.9	3.0	2.9	1.7	3.4	2.7	87.09	312.50	1.48
15	0.8	0.1	8	23	2.2	3.3	3.2	1.1	3.1	2.2	51.56	196.08	1.84
16	0.1	0.1	3	20	1.1	3.2	2.1	1.5	3.9	2.3	161.09	138.89	3.41
17	0.5	0.1	4	9	1.8	3.2	3.1	1.6	3.8	2.9	34.12	588.24	1.76
18	0.1	0.1	2	2	2.0	3.5	2.4	3.2	4.9	3.7	94.45	1428.57	4.85
19	2.5	0.1	7	17	0.9	3.0	2.0	2.1	3.7	3.6	25.44	163.93	2.24
20	0.1	0.1	2	2	1.5	2.7	1.9	2.3	3.7	2.9	217.43	114.94	2.75

8. CONCLUSIONS

It becomes evident that for all the cases, the value of the tightened sampling interval h_2 equals its minimum allowable value (0.1 time units) and, so, a second sample should be collected as soon as possible when a warning is issued by the scheme.

Moreover, for large values of M, λ and greater magnitude of the shifts of the assignable causes δ, γ, the value of ECT gets larger. Another conclusion is that the values of the warning and control limits are inversely related to the value of $1/ARL_0$ but lead to greater values of WARL. As regards WARL, it is concluded to be inversely related to the sample sizes (n_1, n_2), to λ, δ and γ, but gets smaller for "tighter" warning and control limits of the scheme.

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία προτείνεται ένα πλήρως δυναμικό σχήμα ελέγχου για την παρακολούθηση διαδικασιών, στις οποίες είναι πιθανή η έλευση πολλαπλών συστηματικών αιτιών. Οι συστηματικές αιτίες δύνανται να επέλθουν συγχρόνως και να επηρεάσουν τη μέση τιμή αλλά και την τυπική απόκλιση του υπό παρακολούθηση χαρακτηριστικού ποιότητας. Η στοχαστική μοντελοποίηση του προτεινόμενου σχήματος ελέγχου γίνεται μέσω μιας δισδιάστατης Μαρκοβιανής αλυσίδας διακριτού χρόνου. Η ανάγκη για διασφάλιση αποδεκτής στατιστικής συμπεριφοράς του σχήματος ελέγχου συνδυαζόμενης με την απαίτηση για ελαχιστοποίηση του συνολικού κόστους ποιότητας, οδηγεί στη διαμόρφωση ενός προβλήματος οικονομικο-στατιστικής βελτιστοποίησης. Η επίλυση του εν λόγω προβλήματος για δεδομένα οικονομικά και στατιστικά χαρακτηριστικά μιας διαδικασίας καθορίζει τις βέλτιστες τιμές των παραμέτρων σχεδίασης του προτεινόμενου σχήματος. Η εφαρμογή του

σχήματος ελέγχου σε μια σειρά διαφορετικών μεταξύ τους διαδικασιών οδηγεί σε ορισμένα χρήσιμα συμπεράσματα.

REFERENCES

De Magalhães M.S. and Moura Neto F.D. (2005). Joint economic model for totally adaptive $\overline{X}$ and R Charts. *European Journal of Operational Research*, **161**, 148-161.

Lu S.-L., Huang C.-J. and Chiu W.-C. (2013). Economic-statistical design of maximum exponentially weighted moving average control charts. *Quality and Reliability Engineering International*, **29**, 1005-1014.

Nenes G., Tasias A.K. and Celano G. (2015). A general model for the economic-statistical design of adaptive control charts for processes subject to multiple assignable causes. *International Journal of Production Research*, **53**, 2146-2164.

Park C. and Reynolds M.R. (1999). Economic design of a variable sampling rate $\overline{X}$ chart. *Journal of Quality Technology*, **31**, 427-443.

Serel D.A. and Moskowitz H. (2008). Joint economic design of EWMA control charts for mean and variance. *European Journal of Operational Research*, **184**, 157-168.

Shewhart W.A. (1931). *Economic Control of Manufactured Product*. Van Nostrand Reinhold Company, Inc., Princeton, N.J.

Tagaras G. and Lee H.L. (1988). Economic design of control charts with different control limits for different assignable causes. *Management Science*, **34**, 1347-1366.

Tasias A.K. and Nenes G. (2012). A Variable parameter Shewhart control scheme for joint monitoring of process mean and variance. *Computers and Industrial Engineering*, **63**, 1154-1170.

Greek Statistical Institute
Proceedings of the 28th Panhellenic Statistics Conference (2015), pp. 430−444

A novel, divergence based, regression for compositional data

M. Tsagris[1]
[1]Independent researcher, Athens, Greece
mtsagris@yahoo.gr

ABSTRACT

In compositional data, an observation is a vector with non-negative components which sum to a constant, typically 1. Data of this type arise in many areas, such as geology, archaeology, biology, economics and political science among others. The goal of this paper is to propose a new, divergence based, regression modelling technique for compositional data. To do so, a recently proved metric which is a special case of the Jensen-Shannon divergence is employed. A strong advantage of this new regression technique is that zeros are naturally handled. An example with real data and simulation studies are presented and are both compared with the log-ratio based regression suggested by Aitchison in 1986.

Keywords: compositional data, Jensen-Shannon divergence, regression, zero values, ϕ-divergence

1. INTRODUCTION

Compositional data is a special type of multivariate data in which the elements of each observation vector are non-negative and sum to a constant, usually taken to be unity.

$$\mathbb{S}^d = \left\{ (x_1, ..., x_D) | x_i \geq 0, \sum_{i=1}^{D} x_i = 1 \right\} \quad (d = D - 1)$$

A typical data set with $D = 3$ components has the following form:

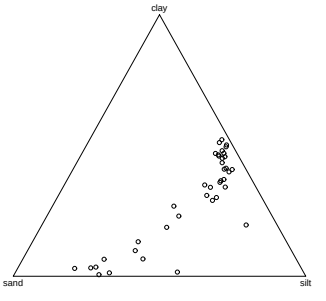
$$\begin{bmatrix} 0.775 & 0.195 & 0.030 \\ 0.719 & 0.249 & 0.320 \\ \vdots & \vdots & \vdots \end{bmatrix}$$

When $D = 3$, a good way to visualize compositional data is the so called ternary diagram. Figure 1 presents the Arctic lake data (we will see this data

set again later and describe it in more detail), which consist of three components, sand, silt and clay. These data are available in Aitchison (2003, p. 359) or the in R package *compositions* along with more compositional data and techniques to analyze such data. Sand is a naturally occurring granular material composed of finely divided rock and mineral particles. Silt is a granular material of a size somewhere between sand and clay whose mineral origin is quartz and feldspar. As for clay, it is a fine-grained natural rock or soil material that combines one or more clay minerals with traces of metal oxides and organic matter.

The closer a point is to a vertex, the higher its value in that corresponding component. If the point is on the vertex, it means that its value in the component corresponding to that vertex is 1 and the other two values are 0. On the opposite side, the further away a point is from a vertex, the lowest its value in the corresponding component. If a point lies on an edge, it means, that the value of the value of the component corresponding to that vertex is zero. If the point is on the middle of that edge, the values of the other two components are exactly 0.5. Points at the barycenter indicate that the values are equal (to 1/3). For example, the points lying at the base of the triangle in Figure 1 have very low values of clay and higher values sand than silt.

Figure 1: *Ternary diagram of the Arctic lake data.*



Analysis of such data may be carried out in various ways which can be summarized in two directions: either by transforming the data or not. In the first direction, the most popular transformations are the log of ratios formed by the components (Aitchison, 1982; 2003). Other approaches include taking the square root of the data, resulting in data which lie on the hypersphere (Stephens, 1982 and Scealy & Welsh, 2011). Another parametric model of this school of thought is the Dirichlet distribution (Gueorguieva et al., 2008). In all of these cases, regression models have been developed.

An important issue in compositional data is the presence of zeros, which cause problems for the logarithmic transformation. The issue of zero values in some components is not addressed in most papers, but see Neocleous et al.(2011) for

an example of discrimination in the presence of zeros. Only when treated as directional data, compositional data have no problem with zeros. Log-ratio and Dirichlet regression zero imputation techniques must be applied prior to fitting the models.

The paper proposes a novel regression method based on a recently suggested metric. It is a new metric for probability distributions (Endres & Schindelin, 2003 and Österreicher & Vajda, 2003), which is a special case of the Jensen-Shannon divergence. We will use it for compositional data, since they also sum to 1. The main advantage of this newly proposed regression is that it handles zeros naturally and in addition it can lead to better fits as will be seen later.

The paper is structured as follows. Section 2 reviews some regression approaches for compositional data analysis and introduces the new regression model. Section 3 contains an example using real data and simulation studies to illustrate the performance of the newly proposed regression. Some Miscellanea and the Conclusion appear in Section 4 and 5 respectively.

2. REGRESSION MODELS

The most popular approach is to use the logistic normal distribution (Aitchison, 1982; 2003)

$$\log\left(\frac{\mathbf{y}_{-D}}{y_D}\right) = \left(\log\frac{y_1}{y_D}, \ldots, \log\frac{y_d}{y_D}\right) \sim \mathrm{N}_d\left(\mathbf{x}^T\mathbf{B}, \boldsymbol{\Sigma}\right),$$

where $\mathbf{x}^T$ is a column vector of the design matrix $\mathbf{X}$, D is the number of components, $d = D - 1$ and

$$\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d) \quad \text{and} \quad \boldsymbol{\beta}_i = (\beta_{0i}, \beta_{1i}, ..., \beta_{pi})^T, \ i = 1, ..., d.$$

are the regression coefficients and p is the number of independent variables. We will denote this regression as Aitchison regression.

A Dirichlet distribution (Gueorguieva et al., 2008) with the parameters linked to some covariates is another approach

$$\mathbf{y} \sim \mathrm{Dir}\left(\frac{\phi}{1+\sum_{j=2}^{D} e^{\mathbf{x}^T\boldsymbol{\beta}_j}}, \frac{\phi e^{\mathbf{x}^T\boldsymbol{\beta}_2}}{1+\sum_{j=2}^{D} e^{\mathbf{x}^T\boldsymbol{\beta}_j}}, \ldots, \frac{\phi e^{\mathbf{x}^T\boldsymbol{\beta}_d}}{1+\sum_{j=2}^{D} e^{\mathbf{x}^T\boldsymbol{\beta}_j}}\right).$$

Note, that the precision parameter ϕ can also be linked to the independent variables via link function of course to ensure positivity of the estimated values. Maier (2014) has written an R package to fit Dirichlet regression models with and without covariates on ϕ.

Two approaches were examined in Murteira & Ramalho (2014) are the OLS regression

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left[\mathbf{y}_i - \mathbf{f}_i\left(\boldsymbol{\beta}; \mathbf{x}\right)\right]\left[\mathbf{y}_i - \mathbf{f}_i\left(\boldsymbol{\beta}; \mathbf{x}\right)\right]$$

and the multinomial logit regression

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} y_i \log \frac{y_i}{\mathbf{f}_i\left(\boldsymbol{\beta}; \mathbf{x}\right)} = \max \boldsymbol{\beta} \sum_{i=1}^{n} y_i \log \mathbf{f}_i\left(\boldsymbol{\beta}; \mathbf{x}\right),$$

where

$$\mathbf{f}_i\left(\boldsymbol{\beta}; \mathbf{x}\right) = \left(\frac{1}{\sum_{j=1}^{D} e^{\mathbf{x}_i^T \boldsymbol{\beta}_j}}, \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}_2}}{\sum_{j=1}^{D} e^{\mathbf{x}_i^T \boldsymbol{\beta}_j}}, \ldots, \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}_d}}{\sum_{j=1}^{D} e^{\mathbf{x}_i^T \boldsymbol{\beta}_j}} \right).$$

Finally, the Kent distribution Kent (1982) was employed by Scealy & Welsh (2011) after the square root was applied component-wise to the data, thus amping them to the unit (hyper)-sphere

$$\sqrt{\mathbf{y}} \sim \text{Kent}\left(\boldsymbol{\mu}\left(\boldsymbol{\beta}, \mathbf{x}\right), \boldsymbol{\Gamma}, \kappa, \boldsymbol{\delta}\right).$$

2.1 The ES-OV regression

We advocate that as a measure of the distance between two compositions we can use a special case of the Jensen-Shannon divergence

$$\text{ES-OV}\left(\mathbf{x}, \mathbf{y}\right) = \sum_{j=1}^{D} \left(x_j \log \frac{2x_j}{x_j + y_j} + y_j \log \frac{2y_j}{x_j + y_j} \right), \tag{1}$$

where $\mathbf{x}$ and $\mathbf{y} \in \mathbb{S}^d$. Endres & Schindelin (2003) and Österreicher & Vajda (2003) proved, independently, that (1) satisfies the triangular identity and thus it is a metric. The names ES-OV comes from the researchers' initials. In fact, (1) is the square of the metric, still a metric, and we will use this version.

The idea is simple and straightforward, minimization of the ES-OV metric between the observed and the fitted compositions with respect to the beta coefficients

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{D} \left(\mathbf{y}_i \log \frac{2\mathbf{y}_i}{\mathbf{y}_i + \mathbf{f}_i\left(\boldsymbol{\beta}; \mathbf{x}\right)} + \mathbf{f}_i\left(\boldsymbol{\beta}; \mathbf{x}\right) \log \frac{2\mathbf{f}_i\left(\boldsymbol{\beta}; \mathbf{x}\right)}{\mathbf{y}_i + \mathbf{f}_i\left(\boldsymbol{\beta}; \mathbf{x}\right)} \right). \tag{2}$$

Below we summarise a few properties of the ES-OV metric and its associated regression model.

- The ES-OV metric belongs to the class of ϕ-divergences. Let $f\left(t\right) = t \log \frac{2t}{1+t} + \log \frac{2}{1+t}$, then $\text{ES-OV}\left(\mathbf{y}, \mathbf{z}\right) = \sum_{j=1}^{D} z_j f\left(\frac{y_j}{z_j}\right)$. Therefore, we can say that this regression falls within the minimum ϕ-divergence regression algorithms.

- A weighted version of the ES-OV regression (2), such as $\sum_{i=1}^{D}\left(\lambda \mathbf{y}_i \log \frac{2\mathbf{y}_i}{\mathbf{y}_i+\hat{\mathbf{y}}_i} + (1-\lambda)\,\hat{\mathbf{y}}_i \log \frac{2\hat{\mathbf{y}}_i}{\mathbf{y}_i+\hat{\mathbf{y}}_i}\right)$ (Jensen-Shannon divergence) produces the best fits when $\lambda = 0$.
- A similar version of (2) is $\sum_{i=1}^{D}\left(y_i \log \frac{y_i}{\hat{y}_i} + \hat{y}_i \log \frac{\hat{y}_i}{y_i}\right)$, which gives nice results, but not as good (2).
- The ES-OV regression (2) can deal with zero values naturally as already mentioned in the Introduction.
- The ES-OV regression (2) can lead to better fits than the logistic normal.

The consistency and the asymptotic distribution of the regression parameters has not been studied. This is a task we have to do and we have confidence that the answers will be positive. As for the last property and in general the fit of regression models for compositional data we suggest the use of the Kulback-Leibler divergence (Kullback, 1997), which was also used as a measure of fit by Theil (1967)

$$\sum_{i=1}^{n} \mathrm{KL}\left(\mathbf{y}_i, \hat{\mathbf{y}}_i\right) = \sum_{i=1}^{n} \mathbf{y}_i \log \frac{\mathbf{y}_i}{\hat{\mathbf{y}}_i}. \tag{3}$$

3. REAL DATA ANALYSIS AND SIMULATION STUDIES

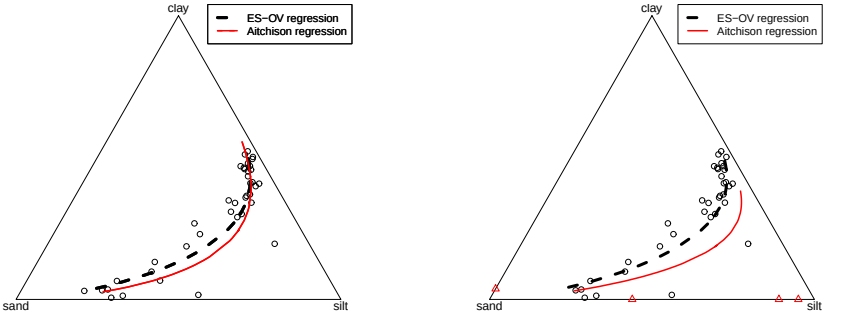
In both the example and the simulation studies presented below, we will compare the ES-OV regression with the Aitchison regression.

3.1 Real data analysis

In the first instance we will present an example with the Arctic lake (real) data we saw in Figure 1. Three ingredients, *sand*, *silt* and *clay* were measured at various depths (39 measurements) of an Arctic lake (Aitchison, 1986, p.359). The goal is to see how the composition in these three elements changes as the water depth increases and how good can our predictions be. We can see in Figure 2(a) that both methods fit the data well. As we move from left to right, we go from the surface to deeper levels of the lake. The composition of the samples has initially high percentage of sand, but as we move to the bottom of the lake, this percentage reduces, while the percentages of silt and clay increase. This makes absolute sense, since the percentages sum to 1, the derivatives, or the rates of change of the components must sum to zero. Figure 3 serves as an ancillary to Figure 2 for the readers who are not familiar with the ternary diagrams.

We then make the values of four components equal to zero. In specific, the percentage of clay becomes zero for three observations and the percentage of silt becomes zero for one observation. In order to apply the logistic normal regression, imputation or zero replacement techniques (Martín-Fernández, 2012) are required.

Figure 2: *Regression lines for the Arctic lake data. The ternary diagram on the left shows the data as they are, whereas the right diagram shows the data with four observations having one component with zero value (red triangles).*



(a) (b)

Templ et al. (2011) has created an R package (robCompositions) which performs zero value replacement for compositional data. Figure (2(b)) shows the Arctic lake data with the zero values (red triangles in the edges of the triangle). The Aitchison regression, applied to the new, zero value replaced data has been heavily affected by the zero values and the fitted values fall outside the bulk of the data. The ES-OV regression on the other hand seems unaffected.

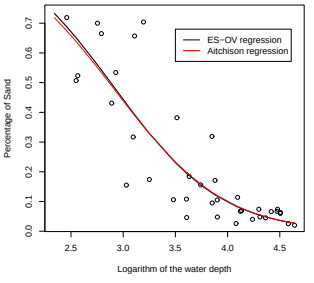
3.2 Simulation studies

Simulation studies were conducted in the spirit of prediction performance. As we mentioned before, the asymptotic properties of the ES-OV regression, thus we cannot use it for inference about the parameters. For this reason, we can only use it for prediction purposes. This is the current drawback of this regression.

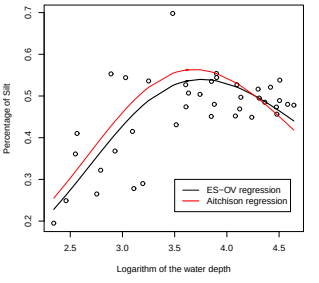
We generated data from logistic normal regression models with various sample sizes ($n = 25, 50, 75, 100$) and components ($d = 6, 11, 16$). We repeated this scenario and made about half of the observations (simulating from a uniform variable) each time contain zero values. There were 2, 4 and 6 components with zero values when the simulated data consisted of 6, 11 and 16 components. For each case, we applied one-fold cross validation. That is, we remove one observation and estimate the regression parameters using the remaining data. Then we predict the value of the observation and calculate the Kullback-Leibler divergence of the true from the fitted value. This was repeated for all observations and in the end

Figure 3: *All graphs contain the logarithm of the water depth against the observed and the estimated percentages of the three components. The graphs in first row correspond to the observed Arctic lake data, while the graphs in the second row correspond to the data with the four zero values.*

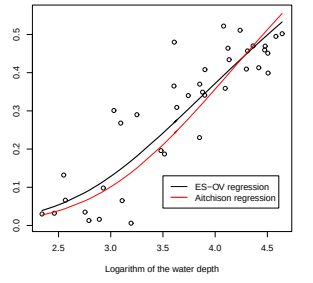
Arctic lake data



Sand

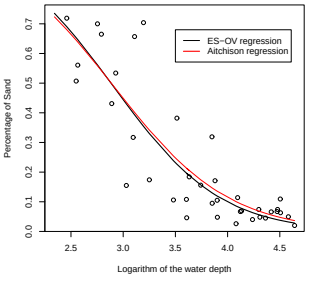


Silt

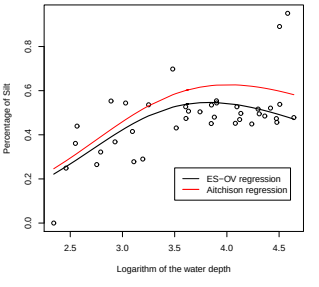


Clay

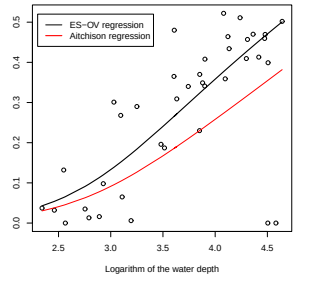
Arctic lake data with zeros



Sand



Silt



Clay

we summed all the divergences

$$\mathrm{KL}\left(\mathbf{y}, \hat{\mathbf{y}}\right) = \sum_{i=1}^{n} \mathbf{y}_i \log \frac{\mathbf{y}_i}{\hat{\mathbf{y}}_{-i}},$$

where $\hat{\mathbf{y}}_{-i}$ is the i-th predicted compositional observation with the i-th observation having been excluded from the estimation of the regression parameters.

The one-fold cross validation procedure was repeated 200 iterations, due to limited computational sources, for each combination of sample size and number of components. In the end we averaged the Kulback-Leibler divergences

$$D = \frac{1}{200} \sum_{j=1}^{200} \mathrm{KL}_j \left(\mathbf{y}, \hat{\mathbf{y}}\right). \tag{4}$$

The number 200 might seem small, but I think is large enough to make valid conclusions. The sample sizes considered are small, again due to limited computational resources. Even in this case, we believe we can extract some valid conclusions.

Scenario 1. Compositional data with no zeros

The pseudo-code for the first set of the simulation studies is given below

Step 1. Generate n data from a multivariate normal regression model $\mathbf{Z} \sim N_p(\mathbf{BX}, \boldsymbol{\Sigma})$, where $\mathbf{X}$ is a design matrix with 2 independent variables. $\mathbf{B}$ were chosen randomly from a standard normal distribution and $\boldsymbol{\Sigma}$ was a diagonal matrix with variances generated from an Exp(1) distribution.

Step 2. Make the $\mathbf{Z}$ compositional using

$$y_1 = \frac{1}{1 + \sum_{j=1}^{p} e^{z_j}}, \quad y_i = \frac{e^{z_i}}{1 + \sum_{j=1}^{p} e^{z_j}}, \quad \text{for} \quad i = 2, \ldots, p.$$

Step 3. Perform the ES-OV and Aitchison regressions and calculate (4) each time.

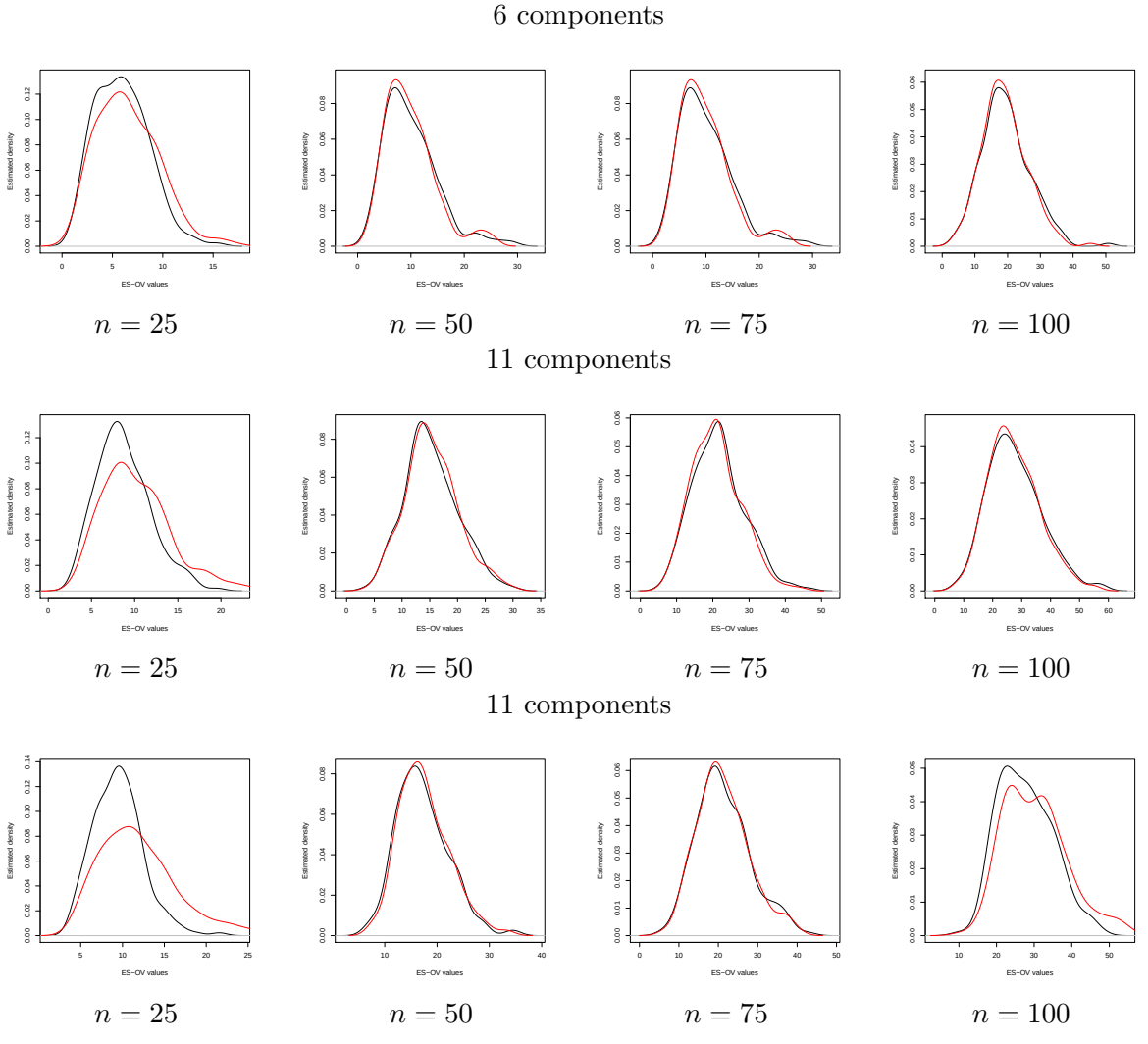
Step 4. Repeat Steps 1-3 for various small sample sizes $n = (25, 50, 75, 100)$ and number of variables $p = (5, 10, 15)$. Thus, we now have $D = (6, 11, 16)$ number of components and essentially $15, 30$ and 45 beta parameters to estimate.

The results of this simulation study are presented in Table 1. We can see that the proportion of times, the Kulback-Lleibler divergence for the fitted values of the ES-OV regression is smaller than the fitted values of the Aitchison regression, grows large as the sample size increases, but when the sample size is equal to 100 decays. An explanation that can be given is due to random error. In addition, if we see Figure 4 we will see that the Kulback-Leibler divergences of the two regression models are close in general.

	Number of components		
Sample sizes	6	11	16
$n = 25$	0.195	0.105	0.055
$n = 50$	0.615	0.430	0.345
$n = 75$	0.820	0.705	0.565
$n = 100$	0.670	0.580	0.294

Table 1: *Proportion of times the Kullback-Leibler divergence, of the true values from the fitted values, is smaller for the ES-OV than for the Aitchison regression. The data have no zero values.*

Figure 4: *All graphs present the kernel estimated densities of the* 200 *Kullback-Leibler divergences of the ES-OV (red line) and the Aitchison regression (black line).*



Scenario 2. Compositional data with zeros

The calculations are the same as before but the scenario of the simulations has one modification: some values were set to zero

Step 1. Generate data from a multivariate normal regression model $\mathbf{Z} \sim N_p(\mathbf{BX}, \mathbf{\Sigma})$, where $\mathbf{X}$ is a design matrix with 2 independent variables. $\mathbf{B}$ were chosen randomly from a standard normal distribution and $\mathbf{\Sigma}$ was a diagonal matrix with variances generated from an Exp(1) distribution.

Step 2. Make the $\mathbf{Z}$ compositional using

$$y_1 = \frac{1}{1+\sum_{j=1}^{p} e^{z_j}}, \quad y_i = \frac{e^{z_i}}{1+\sum_{j=1}^{p} e^{z_j}}, \quad \text{for} \quad i = 2, \ldots, p.$$

Step 3. Set about 50% (use a uniform distribution) of the elements of these components equal to zero and rescale these vectors so that their sum is again equal to 1.

Step 4. Perform the ES-OV regression and calculate (4).

Step 5. Use the R package *robCompositions* to replace the zero values and then apply the Aitchison regression and calculate (4).

Step 6. Repeat Steps 1-5 for various small sample sizes $n = (25, 50, 75, 100)$ and $p = (5, 10, 15)$. Thus, we now have $D = (6, 11, 16)$ number of components and essentially $15, 30$ and 45 beta parameters to estimate.

Table 2 presents the the proportion of times the Kulback-Lleibler divergence for the fitted values of the ES-OV regression is smaller than for the fitted values of the Aitchison regression. The results are clearer now, obviously the ES-Ov has managed to give better predictions, in terms of smaller Kullback-Leibler divergences.

	Number of components		
Sample sizes	6	11	16
$n = 25$	0.885	0.885	0.790
$n = 50$	0.975	0.995	0.995
$n = 75$	1.000	1.000	1.000
$n = 100$	1.000	1.000	0.990

Table 2: *Proportion of times the Kullback-Leibler divergence, of the true values from the fitted values, is smaller for the ES-OV than for the Aitchison regression. The data have zero values.*

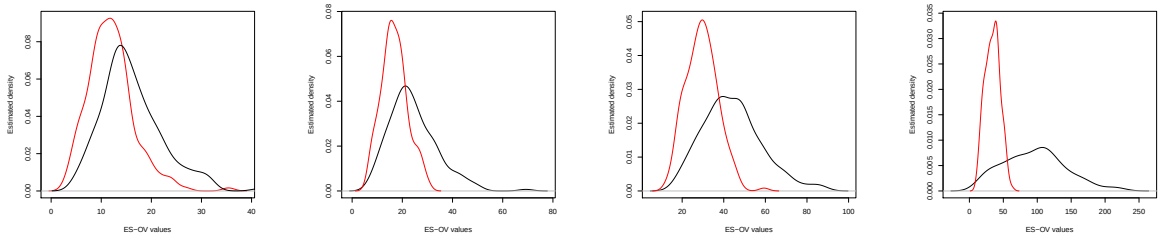
4. MISCELLANEA

When $D = 2$, we end up with proportional data which can be analysed using a beta distribution. kateri and Agresti (2010) used a ϕ-divergence regression model for binary responses. We can say, that in this cases the ES-OV is a special case of that model, since the ES-OV belongs to the class of ϕ-divergence statistics as we saw before.

Can we use another ϕ-divergence statistic for regression. Of course we can, the χ^2 distance and the Hellinger distance have been used and the results were very good. Again, the issue of consistency needs to be checked. Until this is done, they can be used for prediction purposes. In both of these cases, statistical properties about the parameters in the discrete case has been established. What about the compositional data case?

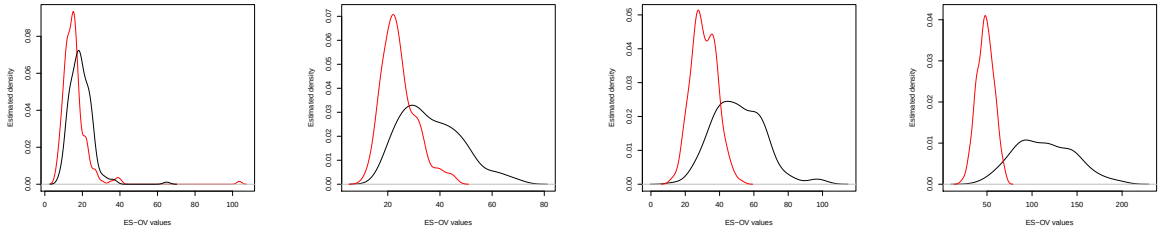
Figure 5: *All graphs present the kernel estimated densities of the* 200 *Kullback-Leibler divergences of the ES-OV (red line) and the Aitchison regression (black line).*

6 components



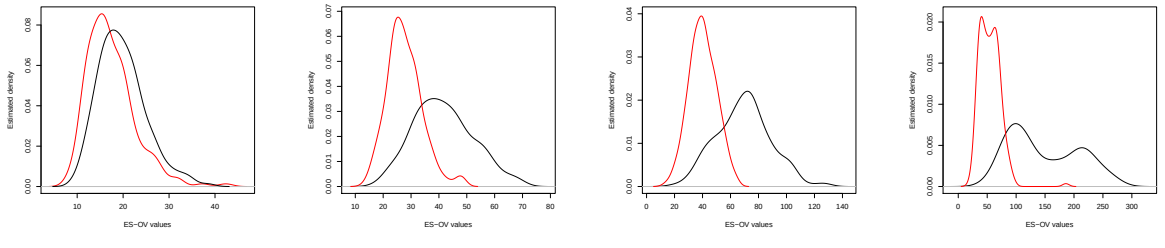
$n = 25$ $n = 50$ $n = 75$ $n = 100$

11 components



$n = 25$ $n = 50$ $n = 75$ $n = 100$

11 components



$n = 25$ $n = 50$ $n = 75$ $n = 100$

What happens when compositional data exist in the independent variables side? We request only for one condition, no zeros to exist in the independent variables compositional data. If this is satisfied, one can apply the centred log-ratio transformation (Atchison, 2003) to the independent side compositional data

$$z_i = \log \frac{x_i}{g(\mathbf{x})}, \quad \text{for} \quad i = 1, \ldots, D,$$

where $g(\mathbf{x}) = \prod_{i=1}^{1/D}$ is the geometric mean of the components. Then, multiply the transformed data with the Helmert sub-matrix (Lee and Small, 1999), which is

an orthonormal $D \times D$ matrix (Lancaster, 1965), whose first row, proportional to the vector of $1s$, is deleted. This removes the singularity problem. Half of the job is done. Now, one can perform PCA on $\mathbf{Z}^T\mathbf{Z}$ and use the scores of the principal components, which is essentially a principal component regression (Jolliffe 2005). The number of principal components to be used can be tuned via cross validation by minimizing the Kullback-Leibler divergence.

5. CONCLUSION

We cannot say that we have proposed a new regression model that always outperforms the Aitchison regression. In fact we did not even see how it compares to the other methods mentioned in Section 2. We saw an example though where Aitchison regression fails. We also saw that in the case of many zeros, Aitchison regression is not very robust, but we believe this is a problem of the EM algorithm for the zero value replacements (Templ 2011) and not of the regression itself. For these cases, the ES-OV regression does better than Aitchison regression. The goal was not to find a model that outperforms always all the other existing ones, but to propose one which is equally good or better in some cases. In the case of no zero values, Aitchison regression seems to do a better job, but in the case of many zeros, ES-OV regression is suggested.

In addition, since we have not examined the asymptotic properties of the ES-OV regression, this method is currently advisable to be used for prediction purposes. So, if the interest is to estimate as accurately as possible the compositions of new observations, given a set of observed data containing a lot of zeros, we suggest the use of the ES-OV regression.

ΠΕΡΙΛΗΨΗ

Στα δεδομένα σύστασης, μια παρατήρηση είναι ένα διάνυσμα με μη αρνητικά στοιχεία τα οποία αθροίζουν σε μία σταθερά, συνήθως 1. Δεδομένα τέτοιου τύπου συναντώται σε πολλές περιοχές, όπως γεωλογία, αρχαιολογία, βιολογία, οικονομικά και πολιτικές επιστήμες. Ο στόχος του παρόντος άρθρου είναι η πρόταση ενός νέου μοντέλου παλινδρόμησης, βασισμένο σε μέτρα απόκλισης, για δεδομένα σύστασης. Για το λόγο αυτό, μία προσφάτως αποδεδειγμένη μετρική, η οποία είναι ειδική περίπτωση της απόκλισης των Jensen και Shannon θα χρησιμοποιηθεί. Α πλεονέκτημα αυτής της νέας τεχνικής παλινδρόμησης είναι ότι οι μηδενικές τιμές είναι εύκολο να χειριστούν. Ένα παράδειγμα με πραγματικά δεδομένα καθώς και μελέτες προσομοίωσης θα παρουσιαστούν και σε κάθε περίπτωση τα αποτελέσματα θα συγκριθούν με τα αποτελέσματα από την παλινδρόμηση που βασίζεται στο μετασχηματισμό του λογάριθμου του πηλίκου που πρότεινε ο Aitchison το 1986.

Acknowledgements: The author would like to express his gratitude to Aziz Alenazi (PhD student in statistics at the university of Sheffield) and to Theo Kypraios (Assistant pro-

fessor in statistics at the university of Nottingham) for their help with the computations. The comments of the anonymous reviewer improved the image of this paper and for this he or she is greatly appreciated.

APPENDIX: R CODES

```
### Aitchison regression

compreg <- function(y, x) {
  ## y is dependent variable, the compositional data
  ## x is the independent variable(s)
  ## the next two commands make sure the data are matrices
  y <- as.matrix(y)   ## makes sure y is a matrix
  y <- y/rowSums(y)  ## makes sure y is compositional data
  z <- log( y[, -1] / y[, 1] )  ## additive log-ratio (alr) transformation
  ## with the first component being the base
  n <- nrow(z)  ## sample size
  d <- ncol(z)  ## dimensionality of z
  p <- ncol(x)  ## dimensionality of x
  X <- as.matrix( cbind(1, x) )  ## the design matrix
  beta <- solve(t(X) %*% X) %*% t(X) %*% z  ## the parameters
  est1 <- X %*% beta ## fitted values
  est2 <- cbind(1, exp(est1))
  est <- est2 / rowSums(est2)
  list(beta = beta, fitted = est)
}


### ES-OV regression

esov.compreg <- function(y, x) {
  ## y is dependent variable, the compositional data
  ## x is the independent variable(s)
  y <- as.matrix(y)
  y <- y/rowSums(y)  ## makes sure y is compositional data
  x <- as.matrix( cbind(1, x) )
  d <- ncol(y) - 1  ## dimensionality of the simplex
  n <- nrow(y)  ## sample size
  z <- list(y = y, x = x)
  reg <- function(para, z = z){
    y <- z$y  ;  x <- z$x
    be <- matrix(para,byrow = T,ncol = d)
    mu1 <- cbind(1, exp(x %*% be))
    mu <- mu1/rowSums(mu1)
    M <- (mu + y)/2
    f <- sum( y * log(y / M) + mu * log(mu / M), na.rm = T )
```

```
    f
  }
  ## the next lines minimize the reg function and obtain the estimated betas
  ini <- as.vector( t( coef( lm(y[, -1] ~ x[, -1]) ) ) )  ## initial values
  val <- NULL
  qa <- nlm(reg, ini, z = z)
  val[1] <- qa$minimum
  qa <- nlm(reg, qa$estimate, z = z)
  val[2] <- qa$minimum
  i <- 2
  while (val[i-1] - val[i] > 0.00001) {
    i <- i + 1
    qa <- nlm(reg, qa$estimate, z = z)
    val[i] <- qa$minimum
  }
  val <- min(val)
  beta <- matrix(qa$estimate, byrow = T, ncol = d)
  mu1 <- cbind(1, exp(x %*% beta))
  mu <- mu1 / rowSums(mu1)
  list(beta = beta, val = val, fitted = mu)
}
```

REFERENCES

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B* **44**, *139-177.*

Aitchison, J. (2003). *The statistical analysis of compositional data, New Jersey: Reprinted by The Blackburn Press.*

Endres, D. M. and Schindelin, J. E. (2003). A new metric for probability distributions. *Information Theory, IEEE Transactions on* **49**, *1858-1860.*

Gueorguieva, R., Rosenheck, R., and Zelterman, D. (2008). Dirichlet component regression and its applications to psychiatric data. *Computational statistics & data analysis* **52**, *5344-5355.*

Jolliffe, I. T. (2005). *Principal component analysis, New York: Springer-Verlag.*

Kateri, M. and Agresti, A. (2010). A generalized regression model for a binary response. *Statistics & Probability Letters* **80**, *89-95.*

Kent, J. T. (1982). The Fisher-Bingham distribution on the sphere. *Journal of the Royal Statistical Society. Series B* **44**, *71-80.*

Kullback, S. (1997). *Information theory and statistics, New York: Dover Publications.*

Lancaster, H. (1965). The Helmert matrices. *American Mathematical Monthly* **72**, *4-12.*

Le, H. and Small, C. G. (1999). Multidimensional scaling of simplex shapes. *Pattern Recognition* **32**, *1601-1613.*

Maier, M. J. (2014). DirichletReg: Dirichlet Regression in R. *R package version 0.5-2 .*

Martín-Fernández, J., Hron, K., Templ, M., Filzmoser, P., and Palarea-Albaladejo, J. (2012). Model-based replacement of rounded zeros in compositional data: Classical and robust approaches. *Computational Statistics & Data Analysis* **56**, *2688-2704.*

Murteira, J. M. and Ramalho, J. J. (2014). Regression analysis of multivariate fractional data. *Econometric Reviews (ahead-of-print) 1-38.*

Neocleous, T., Aitken, C., and Zadora, G. (2011). Transformations for compositional data with zeros with an application to forensic evidence evaluation. *Chemometrics and IntelligentLaboratory Systems* **109***, 77-85.*

Österreicher, F. and Vajda, I. (2003). A new class of metric divergences on probability spaces and its applicability in statistics. *Annals of the Institute of Statistical Mathematics* **55***, 639-653.*

Scealy, J. L. and Welsh, A. H. (2011). Regression for compositional data by using distributions defined on the hypersphere. *Journal of the Royal Statistical Society. Series B* **73***, 351-375.*

Stephens, M. A. (1982). Use of the von Mises distribution to analyse continuous proportions. *Biometrika* **69***, 197-203.*

Templ, M., Hron, K., and Filzmoser, P. (2011). robCompositions: Robust estimation for compositional data. *R package version 0.8-4.*

Theil, H. (1967). *Economics and information theory. Amsterdam: North-Holland publishing company.*

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής (2015), σελ.445-446

ΕΥΡΕΤΗΡΙΟ ΣΥΓΓΡΑΦΕΩΝ (INDEX)