



ΠΡΑΚΤΙΚΑ

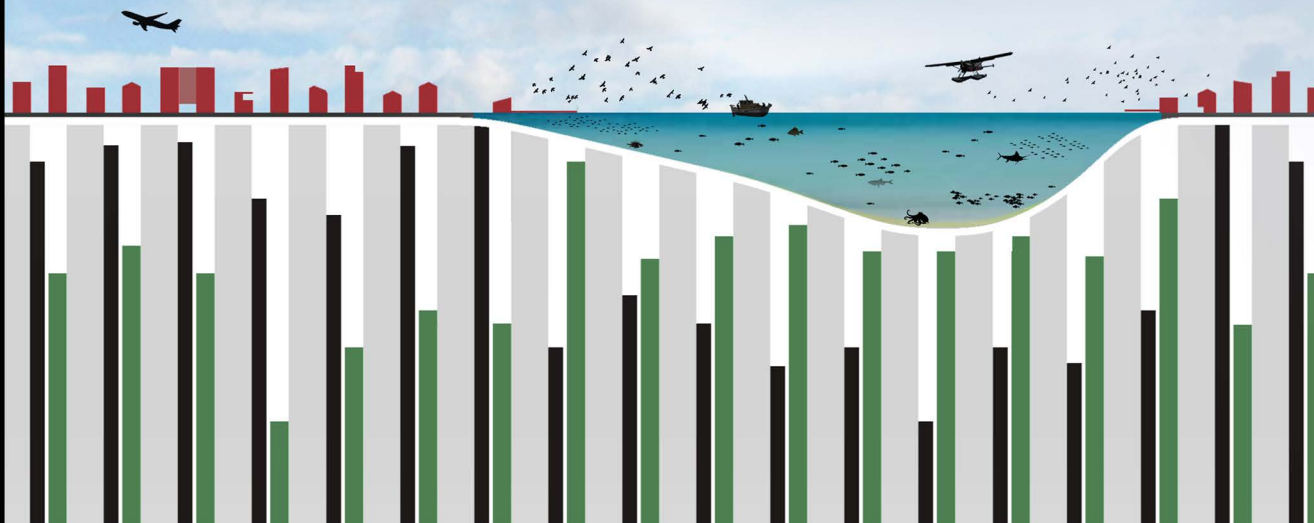
32^{ου} ΠΑΝΕΛΛΗΝΙΟΥ
ΣΥΝΕΔΡΙΟΥ ΣΤΑΤΙΣΤΙΚΗΣ

ΠΡΑΚΤΙΚΑ

32^{ου} ΠΑΝΕΛΛΗΝΙΟΥ
ΣΥΝΕΔΡΙΟΥ ΣΤΑΤΙΣΤΙΚΗΣ

PROCEEDINGS OF THE 32th PANHELLENIC STATISTICS CONFERENCE

Η ΕΠΙΣΤΗΜΗ ΤΗΣ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ Η ΣΥΜΒΟΛΗ ΤΗΣ ΣΤΗ
ΒΙΩΣΙΜΗ ΑΝΑΠΤΥΞΗ





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

Π Ρ Α Κ Τ Ι Κ Α

32^{ου} Πανελληνίου
Συνεδρίου Στατιστικής

PROCEEDINGS

of the 32nd Panhellenic
Statistics Conference

**Η Επιστήμη της Στατιστικής και η
Συμβολή της στη Βιώσιμη Ανάπτυξη**

**The Science of Statistics and its
Contribution to Sustainable
Development**

Ιωάννινα, 30 Μαΐου-1 Ιουνίου 2019



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

Π Ρ Α Κ Τ Ι Κ Α

32^ο Πανελληνίου
Συνεδρίου Στατιστικής

**Η Επιστήμη της Στατιστικής και η
Συμβολή της στη Βιώσιμη Ανάπτυξη**

Οργάνωση

ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

Ιωάννινα, 30 Μαΐου-1 Ιουνίου 2019



ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ

Σολωμού 5 (Πλατεία Εξαρχείων)

Τηλ.& Fax 210 33.03.909

Email: secretariat@esi-stat.gr

<http://www.esi-stat.gr>

ISSN: 1792-2461

ΠΕΡΙΕΧΟΜΕΝΑ

	σελ.
ΠΡΟΛΟΓΟΣ	6
ΧΟΡΗΓΟΙ.....	10
ΠΡΟΓΡΑΜΜΑ ΣΥΝΕΔΡΙΟΥ.....	12
ΕΠΙΤΡΟΠΕΣ ΣΥΝΕΔΡΙΟΥ	22

Προσκεκλημένη Εργασία

<i>Γ. Δονάτος, Γ. Χάλκος : ΣΤΑΤΙΣΤΙΚΕΣ ΜΕΘΟΔΟΙ ΚΑΙ ΠΡΑΣΙΝΗ ΑΝΑΠΤΥΞΗ</i>	<i>25</i>
---	-----------

Εργασίες στα Ελληνικά

<i>Μ. Αναστασοπούλου , Α. Ρακιτζής: ΔΙΑΓΡΑΜΜΑΤΑ ΕΛΕΓΧΟΥ ΤΥΠΟΥ ΕΥΡΩ ΓΙΑ ΤΗΝ ΠΑΡΑΚΟΛΟΥΘΗΣΗ ΔΙΕΡΓΑΣΙΩΝ ΜΕ ΠΕΠΕΡΑΣΜΕΝΟ ΕΥΡΟΣ ΤΙΜΩΝ.....</i>	<i>46</i>
<i>Ε. Αποστολίδου, Φ. Κολυβά – Μαχαίρα : ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΤΗΣ ΔΥΝΑΜΙΚΗΣ ΣΥΣΧΕΤΙΣΗΣ ΟΙΚΟΝΟΜΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΜΕ ΤΗ ΧΡΗΣΗ ΧΡΟΝΙΚΑ ΜΕΤΑΒΑΛΛΟΜΕΝΩΝ ΣΥΖΕΥΞΕΩΝ</i>	<i>61</i>
<i>Γ. Βασιλειάδης: ΜΕΛΕΤΗ ΤΗΣ ΟΥΡΑΣ Geo/Geo/1/N ΜΕ ΤΗ ΒΟΗΘΕΙΑ ΕΝΟΣ ΜΑΡΚΟΒΙΑΝΟΥ ΣΥΣΤΗΜΑΤΟΣ ΔΙΑΚΡΙΤΟΥ ΧΡΟΝΟΥ</i>	<i>77</i>
<i>Οδυσσέας Γαλάνης, Μαρία Παπουτσόγλου, Ελευθέριος Αγγελής: ΑΝΑΛΥΣΗ ΤΗΣ ΔΙΑΔΙΚΤΥΑΚΗΣ ΚΟΙΝΟΤΗΤΑΣ ΕΡΩΤΗΣΕΩΝ ΚΑΙ ΑΠΑΝΤΗΣΕΩΝ CROSS VALIDATED</i>	<i>90</i>
<i>Σ. Δαφνής, Ε. Μακρή, Μ. Κούτρας : ΝΕΕΣ ΚΑΤΑΝΟΜΕΣ ΣΧΗΜΑΤΙΣΜΩΝ</i>	<i>105</i>
<i>Χ. Ζάραγκας: ΑΝΑΛΥΣΗ ΑΓΩΝΙΣΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΠΡΟΠΟΝΗΤΙΚΗ ΚΑΘΟΔΗΓΗΣΗ ΤΕΧΝΙΚΟΤΑΚΤΙΚΗΣ ΣΤΟ ΤΖΟΥΝΤΟ: ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ.....</i>	<i>118</i>
<i>Ο. Θεοδοσιάδου, Γ. Τσακλίδης: ΜΟΝΤΕΛΟ ΧΩΡΟΥ ΚΑΤΑΣΤΑΣΕΩΝ ΜΕ ΕΚΘΕΤΙΚΟΥ ΤΥΠΟΥ ΔΙΑΝΥΣΜΑ ΚΑΤΑΣΤΑΣΗΣ ΚΑΙ ΓΡΑΜΜΙΚΗ ΕΞΙΣΩΣΗ ΠΑΡΑΤΗΡΗΣΕΩΝ</i>	<i>131</i>
<i>Β. Κούτρας, Μ. Κούτρας, Σ. Δαφνής: ΜΙΑ ΝΕΑ ΔΙΑΚΡΙΤΗ ΚΑΤΑΝΟΜΗ.....</i>	<i>145</i>
<i>Μ. Κούτρας, Δ. Λυμπερόπουλος: ΝΕΕΣ ΑΝΙΣΟΤΗΤΕΣ ΓΙΑ N-DEMIMARTINGALES ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΙΣ ΣΤΑΤΙΣΤΙΚΕΣ ΣΥΝΑΡΤΗΣΕΙΣ ΣΑΡΩΣΗΣ.....</i>	<i>155</i>
<i>Ζ. Κορανά, Α. Ρακιτζής : ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΤΩΝ ΔΙΑΤΡΟΦΙΚΩΝ ΣΥΝΗΘΕΙΩΝ ΤΩΝ ΕΦΗΒΩΝ ΣΤΗΝ ΕΛΛΑΔΑ.....</i>	<i>168</i>
<i>Κ. Λουμπόνιας, Γ. Τσακλίδης: ΕΓΧΡΩΜΟ ΤΟΒΙΤ ΚΑΛΜΑΝ ΦΙΛΤΡΟ.....</i>	<i>183</i>
<i>Ρ. Λύκου , Γ. Τσακλίδης: ΕΚ ΤΩΝ ΠΡΟΤΕΡΩΝ ΕΚΤΙΜΗΣΗ ΤΩΝ ΣΦΑΛΜΑΤΩΝ ΠΑΡΑΤΗΡΗΣΗΣ ΣΤΟ ΦΙΛΤΡΟ ΣΩΜΑΤΙΔΙΩΝ</i>	<i>195</i>

Ε.Μπαμίχα, Γ. Βερροπούλου: ΑΠΑΣΧΟΛΗΣΗ ΚΑΙ ΚΑΤΑΘΛΙΨΗ ΣΕ ΑΤΟΜΑ ΗΛΙΚΙΑΣ 50 ΕΤΩΝ ΚΑΙ ΑΝΩ ΣΤΗΝ ΕΛΛΑΔΑ ΚΑΙ ΣΤΗΝ ΕΥΡΩΠΗ ΒΑΣΕΙ ΣΤΟΙΧΕΙΩΝ ΤΟΥ 2 ^{ΟΥ} ΚΥΜΑΤΟΣ ΤΗΣ ΕΡΕΥΝΑΣ SHARE	205
Π. Μουσιάδης: ΣΤΑΤΙΣΤΙΚΗ ΜΕ ΤΗΝ ΡΥΘΜΟΝ	221
Ι. Παπατσούμα, Ν. Φαρμάκης: ΣΥΝΑΡΤΗΣΕΙΣ ΠΥΚΝΟΤΗΤΑΣ ΠΙΘΑΝΟΤΗΤΑΣ ΜΕ ΑΡΝΗΤΙΚΟ ΕΚΘΕΤΗ: ΘΕΩΡΗΤΙΚΗ ΚΑΙ ΔΕΙΓΜΑΤΟΛΗΠΤΙΚΗ ΠΡΟΣΕΓΓΙΣΗ	235
Π. Προδρομίδης: Η ΤΡΑΠΕΖΙΚΗ ΧΡΗΜΑΤΟΔΟΤΗΣΗ ΤΗΣ ΠΑΡΑΓΩΓΗΣ ΣΤΗΝ ΔΙΑΡΚΕΙΑ ΤΩΝ ΜΝΗΜΟΝΙΩΝ: ΤΑΣΕΙΣ, ΕΠΟΧΙΚΟΤΗΤΑ, ΜΕΣΟΧΡΟΝΙΕΣ ΑΥΞΟΜΕΙΩΣΕΙΣ ΚΑΤΑ ΚΛΑΔΟ (ΙΟΥΝ. 2010 – ΔΕΚ.2018)	247
Κ.Τασιάς, Γ. Νενές, Σ. Παναγιωτίδου : ΑΝΑΠΤΥΞΗ ΜΟΝΤΕΛΟΥ ΟΛΙΣΤΙΚΗΣ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ ΠΑΡΑΓΩΓΙΚΩΝ ΔΙΑΔΙΚΑΣΙΩΝ ΣΥΝΕΧΟΥΣ ΠΑΡΑΚΟΛΟΥΘΗΣΗΣ	262
Δ. Τζαφέρη, Γ. Τσακλίδης, Π. Φουσέκης : Η ΣΧΕΣΗ ΑΝΑΜΕΣΑ ΣΤΟΝ S&P 500 ΚΑΙ VIX ΑΝΑΛΥΣΗ ΜΕ ΣΥΖΕΥΞΕΙΣ	278
Α.Τσιμπάνος, Χ.Αγιακλόγλου: ΕΠΙΛΟΓΗ ΧΩΡΙΚΩΝ ΟΙΚΟΝΟΜΕΤΡΙΚΩΝ ΥΠΟΔΕΙΓΜΑΤΩΝ ΜΕ ΠΛΗΡΟΦΟΡΙΑΚΑ ΚΡΙΤΗΡΙΑ	293

Εργασίες στα Αγγλικά

A. Anastasiou, P. Hatzopoulos, A. Karagrigroriou, G. Mavridoglou: ON TIME SERIES CLUSTERING METRICS WITH APPLICATIONS IN HEALTHCARE SYSTEMS.....	308
D. Bagavos, P.N. Patil, A.T.A. Wood: AN L_2 DISTANCE BASED GOODNESS-OF-FIT TEST.....	324
A. Bozikas, G. Pitselis: A THREE-LEVEL HIERARCHICAL CREDIBILITY REGRESSION APPROACH TO MODELLING MULTI-POPULATION MORTALITY DATA.....	336
C. Charalambides: MULTIVARIATE DISCRETE q -UNIFORM DISTRIBUTIONS.....	351
L. Zangelidis: A QUANTITATIVE ANALYSIS OF AGRICULTURAL PRODUCTION IN GREECE, 2004-2016.....	361
S. Nikolakopoulos, I. Ntzoufras : EVIDENCE SYNTHESIS USING BAYES FACTORS.....	377
C. Tripodis, D. Karlis: HIERARCHICAL FORECASTING IN GREEK CONVENIENCE RETAIL INDUSTRY.....	392

ΠΡΟΛΟΓΟΣ

Το Συνέδριο πραγματοποιήθηκε στα Ιωάννινα, 30 Μαΐου – 1 Ιουνίου 2019, σε συνεργασία με το Τμήμα Μαθηματικών του Πανεπιστημίου Ιωαννίνων και τίτλο: «Η Επιστήμη της Στατιστικής και η συμβολή της στη βιώσιμη ανάπτυξη». Η πλειονότητα των εργασιών έλαβε χώρα στο Συνεδριακό Κέντρο «Κ. Παπούλιας» πλην της επίσημης τελετής έναρξης και τριών κεντρικών ομιλιών που πραγματοποιήθηκαν στην αίθουσα «Γ. Μυλωνάς» της Πανεπιστημιούπολης.

Βασικός στόχος του Συνεδρίου ήταν η ανάδειξη του ρόλου της Στατιστικής στην υλοποίηση της ατζέντας για την βιώσιμη ανάπτυξη και συνακόλουθα ο διεπιστημονικός χαρακτήρας της και οι δυνατότητες χρήσης και εφαρμογής της. Σε αυτό το πλαίσιο, επιδιώχθηκε η παρουσίαση συγχρόνων μεθόδων για τη διαχείριση μεγάλου όγκου δεδομένων, την ανάπτυξη υποδειγμάτων και την εξαγωγή αποτελεσμάτων σε διάφορους τομείς, άμεσα σχετιζόμενους με τη βιώσιμη ανάπτυξη.

Την έναρξη κήρυξε ο Πρόεδρος του ΕΣΙ, Ομότιμος Καθηγητής Πολυχρόνης Μωυσιάδης. Εκ μέρους της Επιστημονικής Επιτροπής και της Οργανωτικής Επιτροπής χαιρετισμό απηύθυνε ο Πρόεδρος του Τμήματος Μαθηματικών, Καθηγητής Κ. Ζωγράφος. Χαιρετισμό εκ μέρους των αρχών του Πανεπιστημίου Ιωαννίνων απηύθυναν, επίσης, ο Κοσμήτορας της Σχολής Θετικών Επιστημών, Καθηγητής Κ. Κοσμίδης, και ο Πρύτανης, Καθηγητής Τρ. Αλμπάνης.

Το Επιστημονικό Πρόγραμμα του Συνεδρίου ήταν ιδιαίτερα πλούσιο καθώς περιλάμβανε στις τρεις ημέρες του Συνεδρίου:

- Τρεις κεντρικές ομιλίες: των Γ. Δονάτου (ΕΚΠΑ) και Γ. Χάλκου (Παν. Θεσσαλίας) με τίτλο «Στατιστικές μέθοδοι και πράσινη ανάπτυξη»,

- Φ. Κουντούρη (ΟΠΑ) με τίτλο «Science driven implementation of sustainable development: the role of the "total economic value"» και
- Α. Κασάπη (ΕΛΣΤΑΤ) με τίτλο «Οι στόχοι βιώσιμης ανάπτυξης 2030 του Οργανισμού Ηνωμένων Εθνών και η σημασία των επιμέρους δεικτών: Η σημασία των Επίσημων Στατιστικών και οι προκλήσεις για τη στατιστική κοινότητα».
- 19 παράλληλες συνεδρίες με 94 παρουσιάσεις εργασιών, καθώς και δύο αναρτημένες ανακοινώσεις (posters) σε θέματα Αναλογισμού-Ασφαλιστικών και Χρηματοοικονομικών Μαθηματικών, Αναλύσεως Δεδομένων, Βιοστατιστικής-Ιατρικής Στατιστικής, Ελέγχου Ποιότητας-Αξιοπιστίας, Μαθηματικής Στατιστικής, Οικονομίας-Ασφαλίσεων, Πειραματικών Σχεδιασμών-Στατιστικών Πακέτων, Πιθανοτήτων, Στατιστικής, Στατιστικής Σεισμολογίας, Στατιστικής στην Οικονομία, Στοχαστικών Διαδικασιών-Φίλτρων, Στοχαστικών Μοντέλων-Επιχειρησιακής Έρευνας.
- δύο ειδικές συνεδρίες με αντικείμενο την ανάλυση πολυμεταβλητών χρονοσειρών
- δύο ερευνητικές συνεδρίες νέων στατιστικών
- μία συνεδρία για την μεθοδολογία της έρευνας και τη στατιστική ανάλυση δεδομένων στη διδασκαλία και την έρευνα στα ελληνικά ΑΕΙ
- μία ειδική συνεδρία για την πιστοποίηση του επαγγέλματος του στατιστικού
- Ένα πεντάωρο σεμινάριο με τίτλο «Στατιστικά Μοντέλα Μετά-Ανάλυσης» που συντονίστηκε από τον Επίκουρο Καθηγητή Δ. Μαυρίδη.

Στην πρωινή συνεδρία του Σαββάτου 1ης Ιουνίου 2019, παρουσιάστηκαν οι εργασίες των υποψηφίων για το Βραβείο Καλύτερης Εργασίας Νέου Στατιστικού, το οποίο θεσμοθετήθηκε το 2005 μετά από εισήγηση του επιτίμου μέλους του ΕΣΙ, Καθηγητή Ν. Balakrishnan και απονέμεται κάθε έτος στα Συνέδρια του ΕΣΙ από το 2006. Οι φετινοί υποψήφιοι ήταν ο κ. Ν. Παπαντώνης, ο οποίος παρουσίασε την εργασία με τίτλο «Μέθοδοι βελτιστοποίησης του πλήθους των κλάσεων», ο κ. Α. Μποζίκας ο οποίος παρουσίασε την εργασία με τίτλο «Multi-population mortality modelling and forecasting: A hierarchical credibility approach», η κα. Ε. Αποστολίδου, η οποία παρουσίασε την εργασία με τίτλο «Μοντελοποίηση της δυναμικής συσχέτισης οικονομικών δεδομένων με τη χρήση χρονικά μεταβαλλόμενων συζεύξεων» και ο κ. Α. Αψεμίδης ο οποίος παρουσίασε την εργασία με τίτλο «Χρήση kernel αποστάσεων στο Στατιστικό Έλεγχο Ποιότητας». Η Επιτροπή του Βραβείου Καλύτερης Εργασίας Νέου Στατιστικού κατέθεσε τη βαθμολογία της και στην τελετή λήξης του

συνεδρίου το βραβείο απονεμήθηκε στον κ. Α. Μποζικά. Η επιτροπή αποτελείται από τους Καθηγητές Α. Καραγρηγορίου και Δ. Καρλή, καθώς και τον Ερευνητή Α΄ βαθμίδος Π. Προδρομίδα.

Για το Ελένιο Βραβείο Καλύτερης Διδακτορικής Διατριβής στη Στατιστική που απονέμεται στη μνήμη της μικρής Ελένης με πρόταση και χορηγία των γονέων της, του συναδέλφου Τρύφωνα Δάρα και της συζύγου του Πολυξένης, οι φετινοί υποψήφιοι ήταν οι κ.κ. Γ. Χρόνης με θέμα «Stochastic modeling of time series with intermittency, persistence and extreme variability, with application to spatio-temporal averages of rainfall fields», Β. Κούτρα με θέμα «Designing Experiments on Networks», Χ. Μερκάτας με θέμα «Bayesian Nonparametrics and Applications», Ι. Παπατσούμα με θέμα «Συμβολή στη Δειγματοληπτική Ανάδειξη Μοντέλων Κατανομών με Χρήση του Συντελεστή Μεταβλητότητας», Σ. Τζανίνης με θέμα «Ουδέτερες κινδύνου κατανομές πιθανότητας μεμειγμένων στοχαστικών διαδικασιών και εφαρμογές». Η Επιτροπή αξιολόγησης αποτελείται από τους Ομότιμους Καθηγητές Θ. Κάκκουλο, Σ. Κουνιά, Χ. Χαραλαμπίδη και τον Καθηγητή Κ. Ζωγράφο και βάσει της βαθμολογίας που κατέθεσε, το ΔΣ του ΕΣΙ απένειμε το βραβείο στον κ. Χ. Μερκάτα και έπαινο στην κα. Β. Κούτρα.

Το Συνέδριο έκλεισε με συνεδρία στην ολομέλεια το Σάββατο 1η Ιουνίου, στην οποία επιχειρήθηκε ένας απολογισμός και τιμήθηκαν οι τοπικοί οργανωτές, και συντελεστές του συνεδρίου: Κ. Ζωγράφος, Α. Μπατσίδης, Κ. Σίμος και Κ. Τζουβάρα.

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 ΕΣΙ, αφορούν κυρίως στον τρόπο παρουσίασης της εργασίας και στην παρουσία ή απουσία τυπογραφικών και σοβαρών επιστημονικών λαθών. Οι εργασίες πρέπει να έχουν ικανό στατιστικό περιεχόμενο, να αναδεικνύουν το πρόβλημα που μελετούν και να μην περιορίζονται μόνο σε Περιγραφική Στατιστική. Για το σκοπό αυτό υπάρχουν κριτήρια δημοσίευσης εργασιών στα Πρακτικά του ΕΣΙ τα οποία είναι αναρτημένα στην ιστοσελίδα του ΕΣΙ, 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 35 εργασίες. Ως κριτές των εργασιών συνεργάστηκαν οι: Χρήστος Αγιακλόγλου, Μαλβίνα Βαμβακάρη, Βασίλειος Βασδέκης, Γεωργία Βεροπούλου, Βασίλειος Γεωργίου, Δημήτρης Ιωαννίδης, Αγλαΐα Καλαματιανού, Αλέξανδρος Καραγρηγορίου, Δημήτρης Καρλής, Δημήτριος Κουγιουμτζής, Φοίβη Κουντούρη, Σταύρος Κουρούκλης, Χρήστος Κίτσος, Φωτεινή Κολυβά Μαχαιρά, Βασίλειος Κούτρας, Σωτήριος Κωτσιαντής, Σωτηρία Μαλεφάκη, Νικόλαος Μαχαιράς, Φώτης Μηλιένος, Απόστολος Μπατσίδης, Χρήστος Νάκας, Δημοσθένης Παναγιωτάκος, Δημήτρης Παπαναστασίου, Αθανάσιος Ρακιτζής, Αθανάσιος Σαχλάς, Γεώργιος Τσακλίδης, Θεόδωρος Χατζηπαντελής, Στέλιος Ψαράκης και Γεώργιος Ψαρράκος .

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. Προηγούνται οι εργασίες στην ελληνική και έπονται οι εργασίες στην αγγλική.

Το Διοικητικό Συμβούλιο (ΔΣ) του ΕΣΙ αισθάνεται την ανάγκη να ευχαριστήσει το Πανεπιστήμιο Ιωαννίνων και την Οργανωτική Επιτροπή για την πολύ καλή οργάνωση και προσφορά τους. (Η σύνθεση της Οργανωτικής Επιτροπής παρατίθεται στη σελίδα 23). Επίσης το ΔΣ του ΕΣΙ ευχαριστεί θερμά τον κ. Δήμο Μωϋσιάδη για την ευγενική προσφορά του να δημιουργήσει την αφίσα του Συνεδρίου.

ΕΚ ΜΕΡΟΥΣ ΤΟΥ ΔΣ ΤΟΥ ΕΣΙ

Ελευθέριος Αγγελής	Δημήτριος Κωνσταντινίδης
Πολυχρόνης Μωϋσιάδης	Γεώργιος Παπαδόπουλος
Παναγιώτης Παπαϊωάννου	Πρόδρομος Προδρομίδης
Αθανάσιος Ρακιτζής	

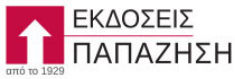
ΧΟΡΗΓΟΙ ΣΥΝΕΔΡΙΟΥ

Το Διοικητικό Συμβούλιο του ΕΣΙ ευχαριστεί θερμά τους κάτωθι Χορηγούς. Χωρίς την οικονομική ενίσχυσή τους θα ήταν εξαιρετικά δύσκολο να επιτύχουμε τους στόχους του Συνεδρίου.



ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ
ΕΥΡΩΣΥΣΤΗΜΑ





ΠΡΟΓΡΑΜΜΑ ΣΥΝΕΔΡΙΟΥ

ΠΕΜΠΤΗ 30 ΜΑΙΟΥ

8:00-10:00 Εγγραφή Συνέδρων – Διανομή συνεδριακού υλικού
ΒΙΟΣΤΑΤΙΣΤΙΚΗ - ΙΑΤΡΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

Προεδρεύων: Δ. ΜΑΥΡΙΑΔΗΣ

- 10:00 **A.-A. Βερνίκη, I.R. White, G. Schwarzer, G. Rücker, Δ. Μανρίδης, J.P. Higgins, Γ. Σαλαντή.** Αξιολόγηση της συνέπειας στη μετα-ανάλυση δικτύου πολλαπλών παρεμβάσεων
- 10:20 **M. Petropoulou, G. Salanti, G. Rücker, G. Schwarzer, I. Moustaki, D. Mavridis.** Methodological developments on detecting outliers in network meta-analysis
- 10:40 **S. Noll, R. Furrer, B. Reiser, C.T. Nakas.** Inference in ROC surface analysis via a trinormal model-based testing approach
- 11:00 **Χ. Σκιαδάς, Χ. Σκιαδάς, Κ.Ν. Ζαφείρης.** Η σχέση της παραμέτρου μορφής (shape parameter) του μοντέλου του Weibull με τα έτη υγιούς ζωής που χάνονται στον άνθρωπο: Αναλυτική παρουσίαση και εκτίμηση από έναν εκτεταμένο πίνακα επιβίωσης

ΑΝΑΛΥΣΗ ΔΕΛΟΜΕΝΩΝ Ι

Προεδρεύουσα: Κ. ΣΚΟΥΡΗ

- 10:00 **Χ. Μήτση, Ε.-Ε. Πρίσκα, Ε. Τζιντζηροπούλου.** Πώς αντιλαμβάνονται 303 Έλληνες μαθητές τον καρκίνο του μαστού
- 10:20 **Ε. Μπαμίχα, Γ. Βερροπούλου.** Απασχόληση και κατάθλιψη σε άτομα ηλικίας 50 ετών και άνω στην Ελλάδα και στην Ευρώπη βάσει στοιχείων του 2ου κύματος της έρευνας SHARE
- 10:40 **Κ. Κουτσαυτάκης, Ε.-Ε. Πρίσκα, Α.-Μ. Μιχαλόπουλος, Ν. Στρόμπλος.** Η άνθηση του παραεμπορίου στην τοπική κοινωνία του Πειραιά και η άποψη των αστυνομικών τμημάτων της Αστυνομικής Διεύθυνσης Πειραιώς και νήσων
- 11:00 **Χ.Κ. Ζάραγκας.** Ανάλυση αγωνιστικών δεδομένων και προπονητική καθοδήγηση τεχνικοτακτικής στο τζούντο
- 11:20 **ΔΙΑΛΕΙΜΜΑ - ΚΑΦΕΣ**

ΣΥΝΕΔΡΙΑ ΝΕΩΝ ΣΤΑΤΙΣΤΙΚΩΝ I

Προεδρεύων: **Ν. ΦΑΡΜΑΚΗΣ** 11:50 **Ν. Παπαντώνης**. Εκτιμητές kernel

12:10 **Θ. Εβρένογλου**. Μείωση της μεροληψίας του εκτιμητή μέγιστης πιθανοφάνειας σε περιπτώσεις σπάνων γεγονότων

12:30 **Κ. Κρικώνης, Μ. Αναστασίου, Κ. Σκούρη, Ι. Κωνσταντάρης**. Μοντέλα διαχείρισης ανθρώπινου δυναμικού εποχικής απασχόλησης σε επιχειρήσεις τουρισμού και εστίασης στην Ελλάδα: Βιβλιογραφική ανασκόπηση

ΠΙΘΑΝΟΤΗΤΕΣ I

Προεδρεύων: **Χ. ΧΑΡΑΛΑΜΠΙΔΗΣ** /

11:50 **Σ.Α. Δαφνής, Φ.Σ. Μακρή, Μ.Β. Κούτρας**. Γενικεύσεις κατανομών ροών και σχηματισμών

12:10 **C.A. Charalambides**. Multivariate discrete q-uniform distributions

12:30 **Β.Ε. Πιπερίγκου**. Μελέτη του ρυθμού αποτυχίας σε μίξεις διακριτών κατανομών

ΣΥΝΕΔΡΙΑ ΝΕΩΝ ΣΤΑΤΙΣΤΙΚΩΝ II

Προεδρεύων: **Π. ΠΡΟΔΡΟΜΙΔΗΣ**

13:00 **L. Zangelidis**. Estimating the production function in agriculture: The case of Greece, 2004-2016

13:20 **K. Bakri, C. Malakates, F. Mermigkas, A.-N. Nikolaou, E. Papadopoulou, P. Pavlou**. Patterns and shifts in employment and performance across the secondary sector in the EU during 2007-16

13:40 **Μ. Σκοτόρης**. Οι δείκτες βιώσιμης ανάπτυξης: Μια πρώτη διερεύνηση των σχέσεων μεταξύ των δεικτών στην Ελλάδα, την Κύπρο και την υπόλοιπη Ε.Ε.

ΠΙΘΑΝΟΤΗΤΕΣ II

Προεδρεύουσα: **Β. ΠΙΠΕΡΙΓΚΟΥ**

13:00 **Β.Μ. Κούτρας, Μ.Β. Κούτρας, Σ.Α. Δαφνής**. Μία νέα διακριτή κατανομή

13:20 **Μ. Κούτρας, Δ. Λυμπερόπουλος**. Νέες ανισότητες για N-demimartingales με εφαρμογές στις στατιστικές συναρτήσεις σύρωσης

13:40 **G.-J. Siouris, A. Karagrorgiou**. Calculation of moments for the distribution of the percentage changes of discrete valued stochastic processes

ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ II

Προεδρεύων: **A. ΡΑΚΙΤΖΗΣ**

13:00 **Z. Κυρανά, A. Ρακιτζής.** Στατιστική ανάλυση δεδομένων των διατροφικών συνηθειών των εφήβων στην Ελλάδα.

13:20 **A. Κοντογιάννη, A. Εμβλωτής.** Διερευνώντας διαστάσεις του στατιστικού εγγραμματισμού σε μελλοντικούς εκπαιδευτικούς: Ανησυχητική καθήλωση

13:40 **M. Τσικαλοπούλου, N. Φαρμάκης.** Διερευνητική παραγοντική ανάλυση: Μελέτη της διάθεση των φοιτητών για το μάθημα της Στατιστικής

14:00 **ΜΕΣΗΜΒΡΙΝΗ ΔΙΑΚΟΠΗ**

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ-ΦΙΑΤΡΑ

Προεδρεύων: **N. ΤΣΑΝΤΑΣ**

16:00 **Γ. Βασιλειάδης.** Μελέτη της ουράς Geo/Geo/I/N με τη βοήθεια ενός μαρκοβιανού συστήματος διακριτού χρόνου

16:20 **O. Θεοδοσιάδου, B. Πολυμένης, Γ. Τσακλίδης.** Μία ημιπαραμετρική μέθοδος για την εκτίμηση των συντελεστών βήτα των κρυφών αλμάτων στις αποδόσεις μετοχών

16:40 **K. Λουμπόνιας, Γ. Τσακλίδης.** Έγχρωμο tobit Kalman φίλτρο

17:00 **P. Λόκου, Γ. Τσακλίδης.** Πιθανότητες επιλογής σωματιδίων κατά την αναδειγματοληψία του φίλτρου σωματιδίων

ΟΙΚΟΝΟΜΙΑ - ΑΣΦΑΛΙΣΕΙΣ

Προεδρεύων: **Γ. ΧΑΛΚΟΣ**

16:00 **Z. Georganta.** Economic damage and spillovers from the giant Ponzi scheme running in Greece: An econometric approach

16:20 **A. Stavrakoudis, M. Burchard-Dziubinska.** Visualization of SMDX dataQ The case of Eurostat's sustainable development goals

16:40 **E. Beliu (Llagami), O. Zaçaj, K. Haxhi, E. Raço.** Mortality rates: Updating the most appropriate rates for the Albanian life insurance companies

17:00 **K. Haxhi, E. Raço, E. Llagami.** The predictive distribution for claims reserves in the Albanian motor insurance

ΕΛΕΓΧΟΣ ΠΟΙΟΤΗΤΑΣ - ΑΞΙΟΠΙΣΤΙΑ

Προεδρεύουσα: **Α. ΚΑΛΑΜΑΤΙΑΝΟΥ**

16:00 Μ. Αναστασοπούλου, Α. Ρακιτζής. Διαγράμματα ελέγχου τύπου EWMA για την παρακολούθηση διεργασιών με πεπερασμένο εύρος τιμών

16:20 Α. Ρακιτζής, Π. Μαραβελάκης, P. Castagliola. Διαγράμματα ελέγχου για την παρακολούθηση αναλογιών με εκτεταμένο αριθμό μηδενικών τιμών

16:40 Κ. Bourazas, P. Tsiamyrtzis. Bayesian self-starting Shiryaev statistic for bivariate Phase I data

17:00 Π. Μπομποτάς, Μ. Κατέρη. Βέλτιστος σχεδιασμός σε step-stress μοντέλα: Μερικές νέες ιδέες

17:30 Συνεδρία για θέματα νέων στατιστικών

Προεδρεύων: **Π. ΠΡΟΔΡΟΜΙΔΗΣ** Κεντρικός Εισηγητής: **Κ. ΤΑΣΙΑΣ**

18:30 Έναρξη του Συνεδρίου - Χαιρετισμοί (ΑΙΘΟΥΣΑ ΜΥΛΩΝΑ)

Προεδρεύων: **Π. ΜΩΥΣΙΑΔΗΣ**

19:15 Γ. Δονάτος, Γ. Χάλκος. Στατιστικές μέθοδοι και πράσινη ανάπτυξη

20:15 Α. Κασάπης. Οι στόχοι βιώσιμης ανάπτυξης 2030 του Οργανισμού Ηνωμένων Εθνών και η σημασία των επιμέρους δεικτών: Η σημασία των Επίσημων Στατιστικών και οι προκλήσεις για τη στατιστική κοινότητα

21:00 Καλωσόρισμα / Welcome cocktail

ΠΑΡΑΣΚΕΥΗ 31 ΜΑΙΟΥ

ΠΕΙΡΑΜΑΤΙΚΟΙ ΣΧΕΔΙΑΣΜΟΙ - ΣΤΑΤΙΣΤΙΚΑ ΠΑΚΕΤΑ

Προεδρεύων: **Σ. ΚΟΥΝΙΑΣ**

9:00 Β. Χασιώτης, Σ. Χατζόπουλος, Σ. Κουνιάς, Ν. Φαρμάκης. Η χρήση ισορροπημένων σχηματισμών σε παραγοντικούς σχεδιασμούς τριών επιπέδων

9:20 Β. Χασιώτης, Σ. Κουνιάς. Βέλτιστοι σχεδιασμοί κυκλικά επαναλαμβανόμενων μετρήσεων με δύο αγωγές σε 2,3,4 περιόδους

9:40 Χ. Μουσιάδης. Στατιστική με την Python

10:00 V.L. Georgiou, N. Perpinias, A. Veskoukis, A. Effraimis, A. Karamitrou, C. Kotopoulos. Utilizing ShinyProxy and containers to create scalable, resource efficient production web applications with R and Shiny

10:20 Ο. Γαλάνης, Μ. Παπουτσόγλου, Ε. Αγγελής. Ανάλυση της διαδικτυακής κοινότητας ερωτήσεων και απαντήσεων cross validated

ΜΑΘΗΜΑΤΙΚΗ ΣΤΑΤΙΣΤΙΚΗ Ι

Προεδρεύων: **Κ. ΠΕΤΡΟΠΟΥΛΟΣ**

9:00 L.K. Patra, S. Kumar, Κ. Πετρόπουλος. Βελτιωμένοι εκτιμητές για συναρτήσεις παραμέτρων κλίμακας σε μοντέλα μείξης

9:20 E. Castilla, N. Martin, L. Pardo, Κ. Ζωγράφος. Εκτιμητές ελάχιστης απόκλισης βασισμένοι στη σύνθετη πιθανοφάνεια και εφαρμογές σε ελέγχους υποθέσεων

9:40 Γ. Αυλογιάρης, Α. Μιχέας, Κ. Ζωγράφος. Κριτήριο τοπικής επιλογής μοντέλου

10:00 D. Bagkavos, P.N. Patil, A.T.A. Wood. A smooth goodness-of-fit test for densities

10:20 Α. Μπατσίδης, M.D. Jimenez-Gamero, A. Lemonte. Έλεγχοι καλής προσαρμογής για την κατανομή Bell

ΑΝΑΛΟΓΙΣΜΟΣ-ΑΣΦΑΛΙΣΤΙΚΑ ΚΑΙ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΑ ΜΑΘΗΜΑΤΙΚΑ

Προεδρεύων: Γ. ΠΙΤΣΕΛΗΣ

- 9:00 **Α. Κανελλόπουλος.** Εφαρμογές των στοχαστικών διατάξεων στη Θεωρία Συλλογικού Κινδύνου
- 9:20 **D.G. Konstantinides.** Ruin probabilities for a double renewal risk model with frequent premium arrivals
- 9:40 **N.Δ. Μαχαίρας, Σ.Μ. Τζανίνης.** Ένας χαρακτηρισμός martingale-ισοδύναμων κατανομών πιθανότητας για σύνθετες μεικτές ανανεωτικές διαδικασίες
- 10:00 **Γ. Ψαρράκος, M.A. Sordo.** Μία οικογένεια μέτρων μεταβλητότητας βασισμένα σε μοντέλα αναλογικών κινδύνων και ουρές κατανομών
- 10:20 **D.G. Konstantinides and G.C. Zachos.** Exhibiting abnormal returns under a risk averse strategy
- 10:45 **ΔΙΑΛΕΙΜΜΑ ΓΙΑ ΚΑΦΕ**

ΣΤΟΧΑΣΤΙΚΑ ΜΟΝΤΕΛΑ-ΕΠΙΧΕΙΡΗΣΙΑΚΗ ΕΡΕΥΝΑ

Προεδρεύων: Α. ΜΠΟΥΡΝΕΤΑΣ

- 11:10 **I. Dimitriou.** Stability and queueing analysis of adaptive cooperative wireless networks
- 11:30 **I. Dimitriou.** Stationary analysis of an ALOHA-type random-access network under the join-the-shortest-queue routing policy
- 11:50 **Κ.Α. Τασιάς, Γ. Νενές, Σ. Παναγιωτίδου.** Ανάπτυξη μοντέλου ολιστικής βελτιστοποίησης παραγωγικών διαδικασιών συνεχούς παρακολούθησης
- 12:10 **B.A. Δημητρίου, Ν. Τσάντας, Α.Κ. Γεωργίου.** Μοντελοποιώντας τη διατμηματική κινητικότητα σε χρόνο συνεχής: Η μη ομογενής περίπτωση
- 12:30 **Α. Μπουρνέτας, Ο. Καναβέτας.** Πολιτικές εκτίμησης και διαχείρισης αποθεμάτων για δύο προϊόντα με υποκατάσταση ζήτησης

ΣΤΑΤΙΣΤΙΚΗ Ι

Προεδρεύων: Δ. ΚΩΝΣΤΑΝΤΙΝΙΔΗΣ

11:10 **E. Castilla, A. Ghosh, N. Martín, L. Pardo.** A new approach to polytomous logistic regression models

11:30 **P. Tsamtsakiri, D. Karlis.** A new bivariate INGARCH model based on Sarmanov distributions

11:50 **C. Tripodis, D. Karlis.** Hierarchical forecasting in Greek convenience retail industry

12:10 **R. Dervishi, E. Cenaj.** Positivist Social Science and the role of Statistics

12:30 **O. Zacaj, O. Hoxhaj, M. Gjoni.** Growth difference in twins

ΣΤΑΤΙΣΤΙΚΗ ΣΕΙΣΜΟΛΟΓΙΑ / STATISTICAL SEISMOLOGY

Προεδρείο: Γ. ΤΣΑΚΛΙΔΗΣ-Ε.ΠΑΠΑΔΗΜΗΤΡΙΟΥ

11:10 **C. Kourouklas, V. Karakostas.** Study of the distribution of moderate to strong earthquakes interevent times in Greece

11:30 **P. Λύκου, Ε. Παπαδημητρίου.** Ανάλυση σημείων αλλαγής σε ενδιάμεσους χρόνους μεταξύ διαδοχικών σεισμών για την περιοχή των Ιονίων Νήσων κατά τη χρονική περίοδο 1970-2018

11:50 **O. Mangira, G. Vasiliadis, V. Karakostas.** A constrained-memory stress releas model for the earthquake occurrence in the Corinth gulf

12:10 **P. Bountzis, E. Papadimitriou.** Are large earthquakes triggered on a global scale?

12:30 **D. Chorozoglou, D. Kugiumtzis, E. Papadimitriou.** Investigation of the correlation of successive earthquakes preceding main shocks in the Greek territory

Προεδρεύων: Γ. ΔΟΝΑΤΟΣ

13:15 **P. Koundouri.** Science driven implementation of sustainable development: the role of the "total economic value"

14:15 **ΕΛΛΟΦΥ ΓΕΥΜΑ**

15:00 **ΑΝΑΧΩΡΗΣΗ ΕΚΔΡΟΜΗΣ**

21:30 **ΕΠΙΣΗΜΟ ΔΕΙΠΝΟ**

ΣΑΒΒΑΤΟ 1 ΙΟΥΝΙΟΥ
ΜΑΘΗΜΑΤΙΚΗ ΣΤΑΤΙΣΤΙΚΗ ΙΙ

Προεδρεύουσα: Ι. ΠΑΠΑΓΕΩΡΓΙΟΥ

- 9:00** **Γ. Τζαβελάς, Α. Μπατσίδης, Π. Οικονόμου.** Στατιστική συμπερασματολογία βασισμένη σε μεροληπτικά δείγματα
- 9:20** **Π. Οικονόμου, Α. Μπατσίδης, Γ. Τζαβελάς, Π. Αλεξόπουλος.** Σταθμισμένες κατανομές και το παράδοξο του Berkson
- 9:40** **Ι. Παπαγεωργίου, Ε. Μουστάκη.** Δειγματοληψία ζευγών για την ανά ζεύγη πιθανοφάνεια λανθανουσών μοντέλων
- 10:00** **Γ. Γιαμλόγλου, Μ. Παπαγεωργίου, Ν. Φαρμάκης.**
Δειγματοληψία και δίκτυα

ΑΝΑΛΥΣΗ ΠΟΛΥΜΕΤΑΒΛΗΤΩΝ ΧΡΟΝΟΣΕΙΡΩΝ Ι

Προεδρεύων: Δ. ΚΟΥΓΙΟΥΜΤΖΗΣ

- 9:00** **Α. Αναστασίου, Π. Χατζόπουλος, Α. Καραγρηγορίου, Γ. Μαυριδόγλου.** Τεχνικές συσταδοποίησης πολυμεταβλητών χρονοσειρών με εφαρμογές σε συστήματα υγείας
- 9:20** **Δ. Τζαφέρη, Γ. Τσακλίδης, Π. Φουσεκής.** Η σχέση ανάμεσα στον S&P 500 και VIX ανάλυση με συζεύξεις
- 9:40** **Α. Papana, Ε. Siggiridou.** Identifying causality based on Information Theory - Facing the curse of dimensionality
- 10:00** **Ε. Siggiridou, Α. Papana.** Direct Granger causality using variable selection and dimension reduction techniques

ΣΤΑΤΙΣΤΙΚΗ ΣΤΗΝ ΟΙΚΟΝΟΜΙΑ

Προεδρεύων: Θ. ΧΑΤΖΗΠΑΝΤΕΛΗΣ

- 9:00** **Π. Προδρομίδης.** Η πορεία της εγχωρίου τραπεζικής χρηματοδότησης της παραγωγής κατά κλάδο από το πρώτο μνημόνιο και μετά, Ιούν.2010-Δεκ.2018
- 9:20** **Α. Τσιμπάνος, Χ. Αγιακλόγλου.** Επιλογή χωρικών οικονομετρικών υποδειγμάτων με πληροφοριακά κριτήρια
- 9:40** **Α. Αδαμόπουλος.** Τουριστική και οικονομική ανάπτυξη. Μία συγκριτική μελέτη για αναδυόμενες χώρες με τη μέθοδο προσομοίωσης Monte Carlo
- 10:00** **Α. Αδαμόπουλος.** Ενεργειακή και οικονομική ανάπτυξη. Μία εμπειρική ανάλυση
- 10:30** **ΔΙΑΛΕΙΜΜΑ ΓΙΑ ΚΑΦΕ**

ΒΡΑΒΕΙΟ ΝΕΟΥ ΣΤΑΤΙΣΤΙΚΟΥ

Προεδρεύων: Δ. ΚΑΡΛΗΣ

- 11:00 **N. Παπαντώνης, N. Φαρμάκης.** Μέθοδοι βελτιστοποίηση του πλήθους των κλάσεων
- 11:20 **A. Bozikas, G. Pitselis.** Multi-population mortality modelling and forecasting: A hierarchical credibility regression approach
- 11:40 **E. Αποστολίδου, Φ. Κολυβά – Μαχαίρα.** Μοντελοποίηση της δυναμικής συσχέτισης οικονομικών δεδομένων με τη χρήση χρονικά μεταβαλλόμενων συζεύξεων
- 12:00 **A. Αψεμίδης, Σ. Ψαράκης.** Χρήση kernel αποστάσεων στο Στατιστικό Έλεγχο Ποιότητας

ΑΝΑΛΥΣΗ ΠΟΛΥΜΕΤΑΒΛΗΤΩΝ ΧΡΟΝΟΣΕΙΡΩΝ II

Προεδρεύων: Φ. ΜΗΛΙΕΝΟΣ

- 11:00 **D. Kugiumtzis.** Structural change detection from causality networks on multivariate time series
- 11:20 **A. Τσιμπίρης.** Πρόβλεψη πολυμεταβλητών χρονοσειρών με βαθιά νευρωνικά δίκτυα
- 11:40 **A. Φωτιάδης.** Πρόβλεψη χρονοσειρών με χρήση τοπικά χωρικά κανονικοποιημένων γραμμικών μοντέλων
- 12:00 **D. Chorozoglou, D. Kugiumtzis, E. Papadimitriou.** Testing the structure of earthquake networks from multivariate time series of successive main shocks in Greece

11:00-16:00 ΠΑΡΑΛΛΗΛΟ ΣΕΜΙΝΑΡΙΟ ΜΕΤΑ-ΑΝΑΛΥΣΗΣ

Υπεύθυνος σεμιναρίου: Δ. ΜΑΥΡΙΑΔΗΣ

ΣΤΑΤΙΣΤΙΚΗ II

Προεδρεύων: Γ. ΨΑΡΡΑΚΟΣ

12:30 **DM. Tsagris, S. Fafalios, A. Alenazi.** Computationally efficient and extremely efficient univariate filtering

12:50 **S. Nikolakopoulos, I. Ntzoufras.** Evidence synthesis for Bayes factors

13:10 **A.E. Μηλιώνης, Ν.Γ. Γαλανόπουλος.** Μελέτη της επίδρασης της "γραμματικοποίησης" και του μετασχηματισμού δεδομένων στις προβλέψεις χρονοσειρών. Μία πρακτική προσέγγιση

13:30 **X. Pedeli, D. Karlis.** Modeling multivariate surveillance data

13:50 **D. Karlis.** Clustering discrete valued time series

ΣΤΑΤΙΣΤΙΚΗ III

Προεδρεύων: Ε. ΑΓΓΕΛΗΣ

12:30 **Κ. Κρικώνης, Μ. Καφέ, Κ. Σκούρη, Ι. Κωνσταντάρας.** Μοντέλα αμοιβής εργαζομένων εποχικής απασχόλησης σε επιχειρήσεις τουρισμού και εστίασης στην Ελλάδα

12:50 **Δ. Πανάρετος, Μ. Βαμβακάρη, Γ. Τζαβελάς, Δ. Παναγιωτάκος.** Διερεύνηση του ρόλου των ορθογωνίων και μη μετασχηματισμών των αξόνων στην παραγοντική ανάλυση σε σχέση με την επαναληψιμότητα των εξαγομένων προτύπων και του επιπέδου του τυχαίου σφάλματος σε προσομοιωμένα δεδομένα

13:10 **Ν. Παπαντώνης.** Εκτίμηση της αμοιβαίας πληροφορίας

13:30 **Δ. Καλαμάρας, Α. Μαντόγλου.** Η εκτίμηση της στατιστικής σημαντικότητας των υπό συνθήκη πιθανοτήτων σε ένα μοντέλο ανάλυσης λανθάνουσών κλάσεων (latent class model). Ένα παράδειγμα βασισμένο στη μελέτη των αυτοβιογραφικών νοσταλγικών μνημών

14:10 **ΕΛΑΦΡΥ ΓΕΥΜΑ / LIGHT LUNCH**

Προεδρεύων: Γ. ΠΑΠΑΔΟΠΟΥΛΟΣ

14:30 **Ε. Αγγελής, Α.Ε. Καλαματιανού, Γ. Παπαδόπουλος, Θ. Χατζηπαντελής**

Μεθοδολογία της έρευνας και στατιστική ανάλυση δεδομένων στη διδασκαλία και την έρευνα στα ελληνικά ΑΕΙ - Υπάρχει ανάγκη πιστοποίησης του επαγγέλματος του στατιστικού;

16:20 **ΑΠΟΛΟΓΙΣΜΟΣ ΚΑΙ ΛΗΞΗ ΣΥΝΕΔΡΙΟΥ**

Προεδρεύων Π. ΜΩΥΣΙΑΔΗΣ

ΕΠΙΣΤΗΜΟΝΙΚΗ ΕΠΙΤΡΟΠΗ

- Ε. Αγγελής**, Καθηγητής Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης.
Β. Βασδέκης, Καθηγητής Οικονομικού Πανεπιστημίου Αθηνών.
Χ. Δαμιανού, Αν. Καθηγητής Πανεπιστημίου Αθηνών.
Κ. Ζωγράφος, Καθηγητής Πανεπιστημίου Ιωαννίνων.
Γ. Ηλιόπουλος, Καθηγητής Πανεπιστημίου Πειραιώς.
Χ. Κουκουβίνος, Καθηγητής Εθνικού Μετσόβιου Πολυτεχνείου.
Φ. Κολυβά-Μαχαίρα, Αν. Καθηγήτρια Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης.
Δ. Κωνσταντινίδης, Καθηγητής Πανεπιστημίου Αιγαίου.
Σ. Λουκάς, Καθηγητής Πανεπιστημίου Ιωαννίνων.
Α. Μπατσιδής, Επ. Καθηγητής Πανεπιστημίου Ιωαννίνων.
Α. Μπουρνέτας, Καθηγητής Πανεπιστημίου Αθηνών.
Π. Μωυσιάδης, Ομότιμος Καθηγητής Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης.
Π. Οικονόμου, Επ. Καθηγητής Πανεπιστημίου Πατρών.
Δ. Παναγιωτάκος, Καθηγητής Χαροκόπειου Πανεπιστημίου.
Γ. Παπαδόπουλος, Αν. Καθηγητής Γεωπονικού Πανεπιστημίου Αθηνών.
Τ. Παπαϊωάννου, Ομότιμος Καθηγητής των Πανεπιστημίων Πειραιώς και Ιωαννίνων.
Π. Προδρομίδης, Ερευνητής Α' Βαθμίδος του Κέντρου Προγραμματισμού και Οικονομικών Ερευνών.
Α. Ρακιτζής, Επ. Καθηγητής Πανεπιστημίου Αιγαίου.
Ι. Τριανταφύλλου, Επ. Καθηγητής Πανεπιστημίου Θεσσαλίας.
Γ. Τσακλίδης, Καθηγητής Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης.
Τ. Χριστοφίδης, Καθηγητής Πανεπιστημίου Κύπρου.

ΟΡΓΑΝΩΤΙΚΗ ΕΠΙΤΡΟΠΗ

- Κ. Ζωγράφος**, Καθηγητής Τμήματος Μαθηματικών Πανεπιστημίου Ιωαννίνων.
Δ. Μαυρίδης, Επ. Καθηγητής Τμήματος Π.Τ.Δ.Ε. Πανεπιστημίου Ιωαννίνων.
Φ. Μηλιένος, Λέκτορας Τμήματος Ψυχολογίας Πανεπιστημίου Ιωαννίνων.
Α. Μπατσίδης, Επ. Καθηγητής Τμήματος Μαθηματικών Πανεπιστημίου Ιωαννίνων.
Κ. Σκούρη, Αν. Καθηγήτρια Τμήματος Μαθηματικών Πανεπιστημίου Ιωαννίνων.
Τ. Παπαϊωάννου, Ομότιμος Καθηγητής Τμήματος Μαθηματικών Πανεπιστημίου Ιωαννίνων και Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης Πανεπιστημίου Πειραιώς.

ΥΠΕΥΘΥΝΟΣ ΠΡΟΓΡΑΜΜΑΤΟΣ

- Γ. Ηλιόπουλος**, Καθηγητής Πανεπιστημίου Πειραιώς.

προσκεκλημένη

εργασία



ΣΤΑΤΙΣΤΙΚΕΣ ΜΕΘΟΔΟΙ ΚΑΙ ΠΡΑΣΙΝΗ ΑΝΑΠΤΥΞΗ

Γεώργιος Δονάτος¹ και Γεώργιος Χάλκος²

¹*Τμήμα Οικονομικών Επιστημών, Πανεπιστήμιο Αθηνών
gdonat@econ.uoa.gr*

²*Τμήμα Οικονομικών Επιστημών, Πανεπιστήμιο Θεσσαλίας
halkos@econ.uth.gr*

ΠΕΡΙΛΗΨΗ

Στην εργασία αυτή παρουσιάζονται τα προβλήματα και οι δράσεις για την πραγματοποίηση της πράσινης ανάπτυξης και επιχειρείται να εξετασθεί η συμβολή των στατιστικών μεθόδων στη μελέτη της επίτευξης της διατήρησιμότητας του περιβάλλοντος. Ειδικότερα, για να δειχθεί η συνειφορά της Στατιστικής, γίνεται αναφορά σε διάφορα στατιστικά και οικονομετρικά υποδείγματα που χρησιμοποιούνται για να ερευνηθούν το μέγεθος και οι επιπτώσεις της αυξανόμενης περιβαντολογικής ζημιάς που συντελείται, η σχέση του κόστους-ωφέλειας από την εφαρμογή των μέτρων για τον περιορισμό των επιπτώσεων της κλιματικής αλλαγής και των άλλων περιβαντολογικών κινδύνων, η δυνατότητα ικανοποίησης των ανθρώπων με την ορθή χρήση των περιβαντολογικών αγαθών καθώς και άλλα σημαντικά περιβαντολογικά ζητήματα. Ακόμη, για να επιτευχθεί η αξιοπιστία της στατιστικής ανάλυσης επαρκών δεδομένων σε σχέση με τον τόπο, το αντικείμενο και το χρόνο προτείνεται η οικοδόμηση και λειτουργία ενός Γεωγραφικού Συστήματος Στατιστικής Πληροφόρησης, το οποίο θα βασίζεται σε ένα Πολύσκοπο Κτηματολόγιο.

Λέξεις κλειδιά: Στατιστικές μέθοδοι, πράσινη ανάπτυξη, κλιματική αλλαγή, διατήρηση της βιοποικιλότητας, διαχείριση φυσικών πόρων.

1. ΕΙΣΑΓΩΓΗ

Το δίλημμα «οικονομική ανάπτυξη ή προστασία του περιβάλλοντος» είναι ένα ψευτοδίλημμα που καλλιεργείται λανθασμένα για πολλά χρόνια. Η οικονομική ευημερία και ανάπτυξη όχι μόνο δεν απαιτεί εγκατάλειψη κάθε προσπάθειας για αντιμετώπιση κινδύνων που απειλούν φυσικό και ανθρώπινο περιβάλλον και μάλιστα με μη αντιστρεπτές πολλές φορές συνέπειες, αλλά αντίθετα ολοκληρωμένη οικονομική και κοινωνική ανάπτυξη επιτυγχάνεται με φροντίδα για το περιβάλλον

και μέσα από ανάπτυξη πράσινης επιχειρηματικότητας, που αναφέρεται στις οικονομικές δραστηριότητες που αναπτύσσονται με βάση την αρχή της βιωσιμότητας. Ειδικότερα, η πράσινη επιχειρηματικότητα αναφέρεται στις αντίστοιχες οικονομικές δραστηριότητες που αφορούν τους επιμέρους τομείς της οικονομίας (πρωτογενής: βιολογικά προϊόντα, δευτερογενής: παραγωγή ενέργειας, πράσινες υποδομές και εξοικονόμηση ενέργειας, ανακύκλωση, τριτογενής: εμπόριο πράσινων προϊόντων, νέες τεχνολογίες, τεταρτογενής: επικοινωνία, προβολή, πιστοποίηση). Μία οικονομική και κοινωνική ανάπτυξη που αποβλέπει στην προστασία και συντήρηση του περιβάλλοντος καλείται «ολοκληρωμένη βιώσιμη ή άλλως αειφόρος ή πράσινη ανάπτυξη». Βασική απειλή για την υλοποίηση της πράσινης ανάπτυξης είναι η κλιματική αλλαγή και κάθε άλλη πίεση που ασκείται στη διατήρηση και διαχείριση του ανθρώπινου και φυσικού περιβάλλοντος ταυτόχρονα.

Καταρχήν η επίτευξη του στόχου που αφορά την αποτελεσματική εφαρμογή της πράσινης ανάπτυξης εξαρτάται από την κοινωνική συναίνεση, σύνταξη, αποδοχή και τήρηση ενός ξεκάθਾਰου θεσμικού πλαισίου και υλοποίηση κατάλληλων πολιτικών καθώς και συμφωνία των χωρών, σε ευρωπαϊκό και διεθνές επίπεδο. Ακόμη, για την επίτευξη του στόχου της πράσινης ανάπτυξης, μεταξύ των άλλων, πρέπει να πραγματοποιηθούν τα εξής: α) Μέτρα για τον μετριασμό και την προσαρμογή της κλιματικής αλλαγής. β) Βιώσιμη διατήρηση και διαχείριση φυσικών πόρων. γ) Βιώσιμη παραγωγή και κατανάλωση προϊόντων, ανακύκλωση και ορθή διαχείριση των αποβλήτων. δ) Δημιουργία πράσινων υποδομών. ε) Κατάρτιση εθνικής στρατηγικής για τη διατήρηση της βιοποικιλότητας κάθε χώρας. στ) Καθιέρωση και λειτουργία υπό προστασία περιοχών και ορθολογική χρήση γης. ζ) Αποτελεσματική προστασία των απειλούμενων ειδών. η) Διατήρηση του γενετικού υλικού των ενδημικών φυτών και ζώων. θ) Ευαισθητοποίηση και ενημέρωση του κοινού. ι) Βιώσιμα συστήματα μεταφορών που να ανταποκρίνονται στις οικονομικές κοινωνικές και περιβαλλοντολογικές ανάγκες. κ) Σημαντική αύξηση του ρυθμού ανάπτυξης των Ανανεώσιμων Πηγών Ενέργειας. λ) Επένδυση στην πράσινη εκπαίδευση, έρευνα και καινοτομία. μ) Βελτίωση της ποιότητας υγείας και ζωής των πολιτών και σεβασμό στα δικαιώματα των μελλοντικών γενεών.

Για την κλιματική αλλαγή, η οποία αποτελεί τον κυριότερο περιβαλλοντολογικό κίνδυνο, ευθύνονται κατά κύριο λόγο οι εκπομπές των αερίων του θερμοκηπίου (διοξείδιο του άνθρακα, μεθάνιο, υποξείδιο του αζώτου και φθοριούχα αέρια) που δεν οφείλονται μόνο σε φυσικές διεργασίες, αλλά ιδιαίτερα σε ανθρωπογενείς παρεμβάσεις. Τα αέρια του θερμοκηπίου προέρχονται, κυρίως, από την παραγωγή ενέργειας από πετρέλαιο, άνθρακα και φυσικό αέριο. Οι στατιστικές και οικονομικές αναλύσεις (Nordhaus, 2007, Stern, 2008) επισημαίνουν ότι πρέπει να αναληφθεί έντονη δράση για τον «μετριασμό» των επιπτώσεων της κλιματικής αλλαγής, μέσα από την μείωση των εκπομπών αερίων του θερμοκηπίου, και την «προσαρμογή» προς την κλιματική αλλαγή, με τον περιορισμό των ζημιών που αυτή συνεπάγεται. Για τον «μετριασμό» χρησιμοποιούνται οι φόροι άνθρακα με βάση το οριακό κοινωνικό κόστος του άνθρακα (Nordhaus, 2014) και οι μέθοδοι απομάκρυνσης του άνθρακα από την ατμόσφαιρα, ενώ για την «προσαρμογή» επιλέγονται επενδύσεις σε προγράμματα προσαρμογής.

2. ΣΥΜΒΟΛΗ ΣΤΑΤΙΣΤΙΚΩΝ ΜΕΘΟΔΩΝ ΣΤΗ ΜΕΛΕΤΗ ΕΠΙΤΕΥΞΗΣ ΠΡΑΣΙΝΗΣ ΑΝΑΠΤΥΞΗΣ

2.1 Υποδείγματα κόστους – ωφέλειας

Τα ζητήματα που αφορούν τον κίνδυνο και την αβεβαιότητα εμφανίζονται συχνά στη μελέτη της κλιματικής αλλαγής και των άλλων περιβαλλοντολογικών κινδύνων σε σχέση με το χώρο ή το χρόνο ή άλλα χαρακτηριστικά που συνδέονται με την πράσινη ανάπτυξη (Heal and Millner, 2014). Αβεβαιότητα υπάρχει σε πλείστα ζητήματα, όπως στον τρόπο με τον οποίο η αύξηση της θερμοκρασίας επιδρά στα καιρικά συστήματα, αλλά και στον τρόπο που οι διάφορες κλιματικές επιπτώσεις και οι άλλοι περιβαλλοντολογικοί κίνδυνοι θα επηρεάσουν την πράσινη οικονομία και την ευημερία. Άλλες αβεβαιότητες αφορούν τους ρυθμούς οικονομικής ανάπτυξης, την αύξηση του πληθυσμού, την ικανότητα των κοινωνιών και οικοσυστημάτων να προσαρμόζονται, τους θεσμούς, την ένταση των εκπομπών αερίων, την ένταση των ζημιών λόγω των περιβαλλοντολογικών κινδύνων, τις νέες τεχνολογίες καθώς και τη μελλοντική διαθεσιμότητά τους, το κόστος τους και την αποτελεσματικότητά τους (Stern, 2007). Η συνολική αβεβαιότητα επηρεάζει τα κόστη και τα οφέλη των στόχων πολιτικής για τον περιορισμό των περιβαλλοντολογικών κινδύνων που εμποδίζουν την επίτευξη της πράσινης ανάπτυξης. Στα υποδείγματα, λαμβανομένης υπόψη της αβεβαιότητας που υπάρχει, ιδιαίτερη προσπάθεια καταβάλλεται για να εκφρασθεί κατά τον καλύτερο τρόπο η επιδίωξη ίσης μεταχείρισης και ίσων συνεπειών για διαφορετικές κοινωνικές ομάδες, διαφορετικές χώρες και διαφορετικές γενεές.

Αν και υπάρχει αβεβαιότητα σχετικά με τις κλιματικές μεταβολές και είναι ιδιαίτερα δύσκολο να βρεθεί η πιθανότητα να συμβούν καταστροφικές καταστάσεις (Weitzman, 2009, a, b, c, d, Weitzman, 2010, a, b, Dasgupta, 2008), είναι πολύ χρήσιμα τα αποτελέσματα μιας ανάλυσης «κόστους – ωφέλειας», προκειμένου να ληφθούν κατάλληλα μέτρα για τον περιορισμό των επιπτώσεων των περιβαλλοντολογικών κινδύνων. Δηλαδή, μιάς μελέτης που αποσκοπεί στο να εξετάσει κατά πόσο το όφελος από τον περιορισμό των επιπτώσεων της κλιματικής αλλαγής και των άλλων περιβαλλοντολογικών κινδύνων είναι μεγαλύτερο από το κόστος των μέτρων που πρέπει να ληφθούν για να περιορισθεί το πρόβλημα. Η ανάλυση κόστους – ωφέλειας απαιτεί την επιλογή διαχρονικής στάθμισης του κόστους και ωφέλειας. Μάλιστα, για τη διαχρονική σύγκριση χρηματικών αξιών χρησιμοποιείται σύστημα στάθμισης που καλείται «προεξόφληση» και μετατρέπει μελλοντικές αξίες σε αντίστοιχες σημερινές. Για τον προσδιορισμό ενός αποδεκτού κοινωνικά «προεξοφλητικού επιτοκίου» λαμβάνεται υπόψη η αβεβαιότητα και οι επιπτώσεις της κλιματικής αλλαγής και επιχειρούνται τα σενάρια «μη δράσης», «μετριασμού» και «προσαρμογής», προκειμένου να εκτιμηθεί σε κάθε περίπτωση η ένταση των κλιματικών αλλαγών καθώς και το κόστος της δράσης που θα επιλεγεί. Σε όσο μεταγενέστερο χρόνο τοποθετείται το κόστος και το όφελος, τόσο μικρότερη είναι η στάθμιση (προεξόφληση) που αποδίδεται. Ο Weitzman, (2001), σε αντικατάσταση

ενός σταθερού προεξοφλητικού επιτοκίου, με βάση τον τύπο του Ramsey (Ramsey, 1926), σε όλη τη διάρκεια της προσαρμογής, πρότεινε την εφαρμογή μειούμενων προεξοφλητικών επιτοκίων και προεξόφληση τύπου «Γάμμα» (Bank of Greece, 2018). Ειδικότερα, θεώρησε ότι η συνάρτηση

$$W(t) = \int_0^{+\infty} e^{-xt} f(x) dx$$

είναι αποτελεσματική προεξοφλητική συνάρτηση σε χρόνο t , η οποία εκφράζει την παρούσα τιμή ενός ευρώ σε χρόνο t , σταθμισμένο από την $f(x)$. Το στιγμιαίο αποτελεσματικό προεξοφλητικό επιτόκιο σε χρόνο t ορίζεται ως:

$$R(t) = \frac{W'(t)}{W(t)}$$

Για την γάμμα κατανομή πιθανότητας έχουμε:

$$W(t) = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^{+\infty} x^{\alpha-1} e^{-(\beta+t)x} dx$$

Όπου: α, β είναι θετικοί αριθμοί, ενώ $\mu = (\alpha/\beta)$ και $\sigma^2 = (\alpha/\beta^2)$ είναι ο μέσος και η διακύμανση, αντίστοιχα, της γάμμα κατανομής. Ο Weitzman, (2001) έδειξε ότι:

$$W(t) = \frac{1}{\left(1 + t \frac{\sigma^2}{\mu}\right)^{\mu^2/\sigma^2}} \quad \text{και} \quad R(t) = \frac{\mu}{1+t}$$

Αν $\sigma = 0$ έχουμε την περίπτωση σταθερού προεξοφλητικού επιτοκίου, το οποίο πιθανό να μην είναι ικανοποιητική παραδοχή στην περίπτωση έργων κλιματικής αλλαγής με μακροπρόθεσμα μελλοντικά οφέλη.

2.2 Άλλα Στατιστικά Υποδείγματα

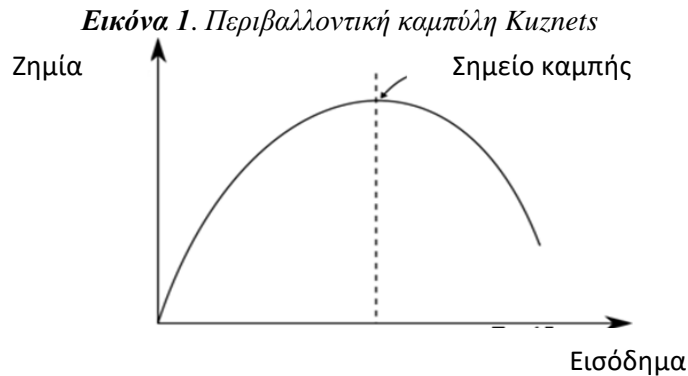
α. Στατιστικοί δείκτες

Οι «στατιστικοί δείκτες» είναι σημαντικοί στο πλαίσιο μιας συγκεκριμένης πολιτικής. Για παράδειγμα, ο Reid et al., (1993) για το World Resources Institute, πρότειναν την κατασκευή 22 δεικτών για τα άγρια είδη, τη γενετική ποικιλότητα, την ποικιλότητα των κοινωνιών και τα εξημερωμένα είδη. Επίσης, ένας άλλος δείκτης, είναι το «οριακό κοινωνικό» κόστος του άνθρακα, που δείχνει τις μελλοντικές ζημιές ανά μονάδα αύξησης του αερίου του θερμοκηπίου (Nordhaus, 2014).

β. Οικονομετρικά Υποδείγματα

Οι σύγχρονες ερευνητικές δραστηριοποιήσεις, αναφορικά με την πράσινη ανάπτυξη, συνδέθηκαν με τη μελέτη της σχέσης περιβαλλοντικής υποβάθμισης και εισοδήματος, στηριζόμενες στον έλεγχο της υπόθεσης του Kuznets (Environmental Kuznets Curve, EKC). Σύμφωνα με την υπόθεση αυτή η περιβαλλοντική ζημία στα αρχικά επίπεδα της οικονομικής ανάπτυξης αυξάνεται μαζί με το κατά κεφαλήν

ακαθάριστο εγχώριο προϊόν (ΑΕΠ/σ) μέχρι κάποιο (κρίσιμο) σημείο (το σημείο καμπής), και μετά από το οποίο αρχίζει να μειώνεται, όπως φαίνεται στην Εικόνα 1.



Τα επόμενα χρόνια, τα συμπεράσματα που προέκυψαν από την έρευνα επιβεβαιώθηκαν με τη χρήση διαφόρων οικονομετρικών μεθόδων και για μια ποικιλία ρυπαντών (Panayotou, 1993, Selden and Song, 1994, Shafik, 1994, Grossman and Krueger, 1995, Stern et al., 1996, Cole et al., 1997, Stern and Common, 2001, Halkos, 2003, Cole, 2004, 2005, Markandya et al., 2006, Halkos, 2012). Η εκτίμηση της ΕΚΚ γίνεται βάσει της εξειδίκευσης οικονομετρικού υποδείγματος σε επίπεδα ή λογάριθμους ή και άλλες πολυωνυμικές μορφές, όπως:

$$(E/c)_t = \beta_0 + \beta_1(AEP/c)_t + \beta_2(AEP/c)_t^2 + \beta_3(AEP/c)_t^3 + \beta_4Z_{1t} + \beta_5Z_{2t} + \dots + \varepsilon_t$$

όπου: i) E/c η μεταβλητή της κατά κεφαλήν περιβαλλοντικής ζημίας (π.χ. κατά κεφαλή εκπομπές ρύπων όπως CO_2 , SO_2 , NO_x κ.α.), ii) AEP/c το κατά κεφαλή ακαθάριστο εγχώριο προϊόν, iii) Z αντιπροσωπεύει άλλες ερμηνευτικές μεταβλητές που μπορούν να επιδράσουν στην περιβαλλοντική ζημία (όπως η πυκνότητα του πληθυσμού, οι θερμοκρασίες, η βιομηχανική ανάπτυξη, κ.α.) και iv) ε_t ο διαταρακτικός όρος.

Η σχέση ρύπανσης και εισοδήματος εν γένει αυξάνεται μονοτονικά για τους παγκόσμιους ρυπαντές, όπως το CO_2 , και για τους ρυπαντές αποθέματος, όπως τα αστικά απόβλητα (Χάλκος, 2016). Οι Halkos και Tsionas, (2001) σχολίασαν τα προβλήματα με κάποιες από τις υπάρχουσες στη βιβλιογραφία για την ΕΚΚ (περιβαντολογική καμπύλη) εκτιμήσεις και την ερμηνεία τους. Μεταξύ αυτών ανέφεραν: i) τα οικονομετρικά προβλήματα ii) το πρόβλημα με το μέσο-διάμεσο εισόδημα.

Ο Δονάτος, (1981) ανέπτυξε ένα οικονομετρικό υπόδειγμα, όπου η αξία ακινήτου μιάς συγκεκριμένης περιοχής σχετίζεται συναρτησιακά με το ευπρόσιτο, με την τοπογραφία, με τις υπηρεσίες για άμεση ή μελλοντική χρήση, με τα δομικά και ιδιοκτησιακά χαρακτηριστικά, με ορισμένους ιστορικούς παράγοντες που χαρακτηρίζουν τον πληθυσμό ή την περιοχή, με την έκταση του χώρου του ακινήτου (και είδος καλλιέργειας προκειμένου για αγροτική γη) και με μια σειρά επιθυμητών – περιβαντολογικών χαρακτηριστικών, όπως: απόσταση από δημοτικό πάρκο,

ακροθαλασσιά, λίμνη ή ποταμός, ασφάλεια περιοχής, θόρυβος (δρόμοι με μεγάλη κίνηση, γραμμές τραίνου, αεροδρόμιο), μόλυνση περιβάλλοντος (εργόστασια, απόβλητα, κλιματολογικές συνθήκες) κλπ. Ο Tyrväinen, (1997), με χρήση της οικονομετρίας, έλαβε υπόψη την περιβαλλοντική σημασία αστικού δάσους της περιοχής Joensuu της Φινλανδίας, για τον προσδιορισμό της αξίας των διαμερισμάτων της περιοχής. Ειδικότερα, για τον υπολογισμό της αξίας των διαμερισμάτων (P_i) χρησιμοποίησε τη συνάρτηση: $P_i = f(A_i, L_i, E_i)$, όπου: i) A_i είναι τα δομικά χαρακτηριστικά, ii) L_i είναι τα στοιχεία για τα χαρακτηριστικά της γειτονιάς, iii) E_i είναι τα περιβαλλοντικά χαρακτηριστικά.

Ο Halkos, (2011) σε δείγμα 71 χωρών μοντελοποίησε την ύπαρξη κινδύνου στη διασφάλιση διαφόρων ειδών χλωρίδας και πανίδας, χρησιμοποιώντας τις ερμηνευτικές μεταβλητές: κατά κεφαλή εκπομπή διοξειδίου του άνθρακα, ενεργειακή ένταση, αγροτική παραγωγή, κατά κεφαλή ακαθάριστο εγχώριο προϊόν, ύπαρξη προστατευόμενων περιοχών και δείκτης ανισοκατανομής εισοδημάτων (δείκτης Gini). Ως εξαρτημένη μεταβλητή χρησιμοποιήθηκε η ύπαρξη κινδύνου στη διασφάλιση διαφόρων ειδών, όπως αμφίβιων, ερπετών, θηλαστικών, ψαριών, πουλιών, δένδρων και φυτών. Τα αποτελέσματα της υποδειγματοποίησης είναι ενδιαφέροντα καθώς χρησιμοποιείται μία μη-παραμετρική παλινδρόμηση ποσοστημορίων (quantile regression) με αναφορά όχι στην επίδραση των ερμηνευτικών μεταβλητών στη μέση συμπεριφορά της εξαρτημένης μεταβλητής, αλλά σε διάφορα ποσοστημόρια. Ομοίως, για κάποιες περιπτώσεις χρησιμοποιήθηκε η μερική παλινδρόμηση (partial regression) ανάμεσα στις ανεξάρτητες και όχι σε μία, αλλά περισσότερες, ταυτόχρονα, εξαρτημένες μεταβλητές. Με τον τρόπο αυτό εξήχθησαν τα αποτελέσματα: i) ο αριθμός των δένδρων σε κίνδυνο αυξάνεται κατά 1,72, 2,9, 0,03 και 0,09, ii) ο αριθμός των φυτών σε κίνδυνο αυξάνεται κατά 1,653, 2,52, 0,022 και 0,0065, iii) ο αριθμός των θηλαστικών σε κίνδυνο αυξάνεται κατά 0,09, 0,54, 0,025 και 0,0001, ανά μονάδα αύξησης στην αγροτική παραγωγή, στην ανισοκατανομή εισοδημάτων, στο επίπεδο ρύπων CO_2 και στον πληθυσμό. Οι επιδράσεις των έξι ερμηνευτικών μεταβλητών (ενεργειακή ένταση, δείκτης Gini, ΑΕΠ/σ, πυκνότητα πληθυσμού, αγροτική Παραγωγή/σ και εκπομπές CO_2 /σ) ανά είδος σε κίνδυνο διαφέρουν. Για την περίπτωση της εξαρτημένης μεταβλητής «ψάρια σε κίνδυνο» όλες οι ερμηνευτικές μεταβλητές έχουν θετική επίδραση με τη μεταβλητή εκπομπές ρύπων CO_2 να έχει τη σημαντικότερη επίδραση, τη μεταβλητή ανισοκατανομής (Gini) να έχει τη δεύτερη σημαντικότερη επίδραση και την ενεργειακή ένταση να έχει τη χαμηλότερη επίδραση. Γενικά τα αποτελέσματα δείχνουν ότι: Αύξηση πληθυσμού, αστικοποίηση και επέκταση οικονομικών δραστηριοτήτων κάνουν ανέφικτη τη διατήρηση του φυσικού περιβάλλοντος των ζώων και των φυτών και τη βελτίωση των προβλημάτων στο μέλλον.

Από τις αρχές της δεκαετίας του '80 οι οικονομολόγοι χρησιμοποιούν μεθόδους έκφρασης αποτίμησης, όπως η αποτίμηση σε υποθετικές αγορές (Contingent valuation markets, CVM), ρωτώντας άμεσα την πρόθεση πληρωμής (Willingness to pay, WTP) των ατόμων για τη διατήρηση κάποιων σπάνιων ή σε κίνδυνο ειδών και εκτιμούν τις περιβαλλοντικές αξίες. Με τη χρήση τυχαίων και αντιπροσωπευτικών δειγμάτων και σωστών δομημένων ερωτηματολογίων μπορούμε

να εκτιμήσουμε τη συνολική αξία περιβαλλοντικών αγαθών ή υπηρεσιών, εκτιμώντας τη συνάρτηση πρόθεσης πληρωμής που περιλαμβάνει προσδιοριστικούς παράγοντες, όπως: i) εισόδημα του ατόμου (Y_i), ii) μορφωτικό επίπεδο (M_i), iii) φύλο (Φ_i), iv) άλλες κοινωνικοοικονομικές μεταβλητές (K_i), v) περιβαλλοντική ή όχι ευαισθητοποίηση (E_i) και vi) κατάσταση του περιβαλλοντικού αγαθού ή υπηρεσίας (ΠΑ). Δηλαδή, για την πρόθεση πληρωμής (WTP), χρησιμοποιείται η συνάρτηση των μεταβλητών αυτών εκφρασμένη ως:

$$WTP_i = f(Y_i, M_i, \Phi_i, K_i, E_i, ΠΑ)$$

Οι Loomis και White, (1996), ερευνώντας 18 διαφορετικά είδη, συμπεράναν ότι η αποτίμηση σε υποθετικές αγορές μπορεί να παρέχει σημαντικές εκτιμήσεις των ωφελειών που αποκομίζουν οι άνθρωποι από τη διατήρηση σπάνιων ή σε κίνδυνο ειδών. Η μέθοδος παρουσιάζει μειονεκτήματα και δυσκολίες αναφορικά με τη δημιουργία μιάς υποθετικής αγοράς καθώς και διάφορες δυνητικές μεροληψίες (Mitchell and Carson, 1989 και Bateman et al., 2002). Οι Nunes και van den Bergh, (2001) παρουσιάζουν μία βιβλιογραφική ανασκόπηση της οικονομικής αξιολόγησης της βιοποικιλότητας βάσει των προαναφερθέντων μεθόδων. Οι περισσότερες στατιστικές μελέτες έχουν γίνει για τις ΗΠΑ και εμφανίζουν την ύπαρξη θετικής διάθεσης της κοινωνίας για τη διαφύλαξη και διατήρηση της βιοποικιλότητας, αλλά ταυτόχρονα δείχνουν ότι η οικονομική και στατιστική βιβλιογραφία έχει ελλείψεις αναφορικά με την κάλυψη των ποικιλόμορφων μεθόδων και ωφελειών σχετικά με την βιοποικιλότητα. Αρκετοί ερευνητές παρουσιάζουν χρηματικές αξίες για διάφορα είδη ζώων (γκρίζα αρκούδα, λευκή φάλαινα, κάποια είδη αετών, γερακιών κά), οι οποίες είναι συνάρτηση της τοποθεσίας, του σχεδιασμού και του χρόνου διεξαγωγής της μελέτης.

Οι Desvousges et al., (1987) ρώτησαν ένα αντιπροσωπευτικό δείγμα νοικοκυριών σε τοπική περιοχή τι θα ήταν διατεθειμένοι να πληρώσουν ως επιπλέον φόρους, ώστε να διατηρήσουν ή και να βελτιώσουν την ποιότητα του νερού στον ποταμό Monongahela της Pennsylvania των ΗΠΑ. Οι αναλυτές χρησιμοποίησαν διάφορες παραλλαγές της μεθόδου υποθετικής αγοράς. Σε μία από αυτές, παρουσιάστηκαν τρία διαφορετικά σενάρια για την ποιότητα του νερού. i) Σενάριο 1: Διατήρηση της παρούσας ποιότητας του νερού αντί να επιτραπεί η χειροτέρευση της ποιότητας σε επίπεδα τα οποία θα το καθιστούσαν ακατάλληλο για οποιαδήποτε χρήση. ii) Σενάριο 2: Βελτίωση της ποιότητας σε επίπεδα που να επιτρέπεται και το ψάρεμα. iii) Σενάριο 3: Επιπλέον βελτίωση μέχρι και του σημείου που να επιτρέπεται η κολύμβηση. Μεταξύ των νοικοκυριών που μελετήθηκαν, μερικά χρησιμοποιούσαν τον ποταμό για αναψυχή, ενώ άλλα όχι. Οι αναλυτές μπορούσαν να δουν το ποσό που οι χρήστες ήταν διατεθειμένοι να πληρώσουν σε σχέση με τους μη-χρήστες, ενώ υπολογίστηκαν τα αποτελέσματα για το σύνολο του δείγματος. Ο Πίνακας 1 δείχνει την μέση προθυμία πληρωμής των χρηστών, των μη-χρηστών και ολοκλήρου του δείγματος, για κάθε προτεινόμενο σενάριο αλλαγής της ποιότητας του νερού.

Πίνακας 1. Μέση προθυμία πληρωμής της ποιότητας του νερού ποταμού

Σενάριο	Μέση WTP για όλο το δείγμα (\$)	Μέση WTP χρηστών (\$)	Μέση WTP μη-χρηστών (\$)
1	24,5	45,3	14,2
2	17,6	31,3	10,8
3	12,4	20,2	8,5

Οι Halkos και Jones, (2012) εξέτασαν την επίδραση κοινωνικών παραγόντων στην απόφαση των ατόμων να πληρώσουν για προστασία της βιοποικιλότητας, καταγράφοντας στάσεις και αντιλήψεις των τοπικών κοινοτήτων στα Εθνικά Πάρκα Δέλτα Λουδία-Αξιού-Αλιάκμονα και Δέλτα Έβρου. Στην υποδειματοποίηση και για την αντιμετώπιση των μηδενικών απαντήσεων προτάθηκαν κατάλληλες στατιστικές και οικονομετρικές μέθοδοι (Tobit first and second hurdle). Οι Halkos και Matsiori, (2017) διερεύνησαν την προθυμία πληρωμής των πολιτών για την προστασία της θαλάσσιας βιοποικιλότητας στο πλαίσιο μιας ευρύτερης έρευνας που σκοπό είχε τη σκιαγράφηση της στάσης των πολιτών απέναντι στο περιβάλλον. Σε πρωτογενή έρευνα στα δημοτικά διαμερίσματα του Βόλου και της Θεσσαλονίκης και με την εφαρμογή κατάλληλης κλίμακας εξήχθησαν συμπεράσματα, σχετικά με τη στάση των ερωτώμενων απέναντι στο περιβάλλον. Σύμφωνα με τα αποτελέσματα της έρευνας οι προσδιοριστικοί παράγοντες της προθυμίας πληρωμής των πολιτών για την προστασία της θαλάσσιας βιοποικιλότητας είναι η ηλικία, το εισόδημα των πολιτών και η οικολογική τους συμπεριφορά.

Στο πλαίσιο του υποδείγματος κόστους – ωφέλειας, μια άλλη μέθοδος είναι του κόστους «ταξιδιού», όπου συγκεντρώνουμε πληροφορίες από πολίτες που χρησιμοποιούν το αγαθό ή την υπηρεσία (π.χ. επίσκεψη σε μια λίμνη ή δάσος). Κεντρική ιδέα της μεθόδου είναι ότι το κόστος επίσκεψης συγκεκριμένου χώρου εκφράζει επιθυμία να αγοράσουν (χρησιμοποιήσουν) το περιβαλλοντικό αγαθό ή υπηρεσία. Τα κόστη, από χρήση υπηρεσιών περιβαλλοντικού αγαθού, συμπεριλαμβάνουν: i) το κόστος ταξιδιού εκφρασμένο σε κόστος των καυσίμων, ii) την καταβολή χρηματικών ποσών για την είσοδο σε μία τοποθεσία αναψυχής, iii) τα διάφορα επιτόπια έξοδα, καθώς και iv) το κόστος ευκαιρίας. Το συνολικό κόστος (K_{ij}) για κάθε άτομο i που επισκέπτεται δεδομένη περιοχή “ j ” μπορεί να εκφραστεί ως (Hanley and Spash, 1993): $K_{ij} = f(D_{ij}, T_{ij}, E_j)$. Όπου: Για κάθε άτομο i ($i = 1 \dots n$) που επισκέπτεται μία δεδομένη περιοχή j , D_{ij} είναι το κόστος μεταφοράς για κάθε άτομο, το οποίο είναι συνάρτηση της απόστασης που πρέπει να καλύψει το συγκεκριμένο άτομο και του κόστους ανά χιλιόμετρο ταξιδιού (κόστος καυσίμων ή εισιτηρίων λεωφορείων κλπ), T_{ij} είναι τα κόστη χρόνου που περιλαμβάνουν το χρόνο για να φτάσει ο επισκέπτης στην τοποθεσία και το χρόνο παραμονής του στην τοποθεσία, E_j είναι το ποσό που καταβάλλεται από τους επισκέπτες κατά την είσοδό τους στην τοποθεσία. Ο Shammin, (1999) χρησιμοποίησε τη μέθοδο κόστους ταξιδιού για να καθορίσει την προθυμία πληρωμής για επίσκεψη και προσφερόμενες υπηρεσίες στον ζωολογικό κήπο Dhaka στο Μπαγκλαντές. Έρευνα με ερωτηματολόγια διεξήχθη σε δείγμα 1000 επισκεπτών, ώστε να συλλεχθούν πληροφορίες, σχετικά με τους επισκέπτες και τα έξοδα τους, προκειμένου να επισκεφτούν τον ζωολογικό κήπο. Η ανάλυση παλινδρόμησης χρησιμοποιήθηκε για

υπόδειγμα ζωνών, με τον αριθμό επισκέψεων να αποτελεί την εξαρτημένη μεταβλητή.

Οι Kahnemann et al., (1997), οι Kahnemann και Sudgen, (2005) και ο Rehbanz, (2007) εστιάζουν την προσοχή τους στην εμπειρία και στην ικανοποίηση που απολαμβάνουν τα άτομα από τη χρήση των περιβαλλοντικών αγαθών. Στην περίπτωση αυτή οι ερωτηθέντες δεν απαντούν σχετικά με την προθυμία πληρωμής για συγκεκριμένα μέτρα πολιτικής, αλλά για το πόσο είναι ικανοποιημένοι με τη ζωή που βιώνουν, η οποία επηρεάζεται άμεσα από τις οικονομικές (εισόδημα, ανεργία, πληθωρισμός, κλπ) και περιβαλλοντικές (όχληση, αέρια ρύπανση, μόλυνση υδάτων, κλπ.) συνθήκες (Clark and Oswald, 1994, DiTella et al., 2001 και Easterlin, 2001). Οι Kountouri et al., 2009 ερευνούν την αξία της κατασκευής ενός αιολικού πάρκου χρησιμοποιώντας τη μέθοδο της υποθετικής αγοράς.

Οι Dell et al., (2014) και ο Hsiang, (2016) καταγράφουν επισκοπήσεις οικονομετρικών μελετών και των στατιστικών μεθοδολογιών που αναπτύχθηκαν για θέματα κλιματολογικών μεταβολών. Ο Fankhauser, (2017) εξετάζει οικονομετρικά τις αντιδράσεις των οικονομικών μέσων στις μεταβολές του κλίματος. Έχουν δημοσιευθεί πολλές οικονομετρικές εργασίες περιπτώσεων που αφορούν τη διαδικασία της προσαρμογής στις καιρικές συνθήκες και κλιματικές - περιβαντολογικές μεταβολές για ειδικά θέματα, όπως: Η ασφάλιση (Ranger and Surminski, 2013), η μετανάστευση (Penning – Rowseu et al., 2013), τα κατοικίδια ζώα και άλογα (Seo et al., 2008, 2009, Seo, 2011). Οι Mendelsohn και Neumann, (1999) διερεύνησαν οικονομετρικά τα κόστη και οφέλη των στρατηγικών πολλαπλής προσαρμογής στην κλιματική αλλαγή, όταν χρησιμοποιούνται ταυτόχρονα. Πολλοί ερευνητές χρησιμοποίησαν μεγάλα δείγματα για δεδομένα νοικοκυριών, αγροκτημάτων και επιχειρήσεων για να μελετήσουν τη σχέση των οικονομικών και των περιβαντολογικών μεταβολών. Στην ανάλυση χρονολογικών σειρών οι παράγοντες του χώρου παραμένουν σταθεροί και οι μεταβολές του καιρού είναι εξωγενείς παράγοντες, κάνοντας ευκολότερη την ταυτοποίηση. Προσομοιώσεις με μεγάλα δείγματα και πολλά σενάρια και στατιστικούς ελέγχους χρησιμοποιούνται για να ενισχυθεί η αξιοπιστία, όταν οι τιμές των παραμέτρων που εμπλέκονται στις μελέτες προσαρμογής στην κλιματική αλλαγή εμφανίζουν αβεβαιότητα. (Dessai and Greenstone, 2007, Lempert and Collins, 2007, Wilby and Dessai, 2010, Lempert and Kalra, 2011, Bhave et al., 2016).

γ. Panel Δεδομένα

Ο Halkos, (2003) έλεγξε εμπειρικά την υπόθεση της ΕΚC μεταξύ περιβαντολογικής ζημίας από εκπομπές θείου και ΑΕΠ με «panel δεδομένα» 73 χωρών (κράτη-μέλη του ΟΟΣΑ και μη) για 31 χρόνια (1960-1990), εφαρμόζοντας για πρώτη φορά στην ερευνητική αυτή περιοχή δύο οικονομετρικές μεθόδους: τους τυχαίους συντελεστές (random coefficients) και την Arellano-Bond Γενικευμένη Μέθοδο των Ροπών (Generalized Method of Moments, A-B GMM). Τα εμπειρικά αποτελέσματα δείχνουν ότι η υπόθεση ελέγχου (ΕΚC) δεν απορρίπτεται στην περίπτωση της Γενικευμένης Μεθόδου των Ροπών. Από την άλλη πλευρά, δεν υπάρχει υποστήριξη της υπόθεσης στην περίπτωση που χρησιμοποιούμε το

υπόδειγμα των τυχαίων συντελεστών. Επίσης, οι Deschênes και Greenstone, (2011), Hsiang και Narita, (2012), Dell et al., (2014) και Burke και Emeric, (2015) χρησιμοποίησαν panel δεδομένα στη μελέτη των καιρικών επιδράσεων και στα σενάρια προσαρμογής για την εκτίμηση της έντασης των κλιματικών αλλαγών. Ο Halkos, (2012) ερεύνησε τη σχέση μεταξύ οικονομικής ανάπτυξης και περιβαλλοντικής ζημίας με τη χρήση δυναμικών διαστρωματικών στοιχείων (panel δεδομένα) για 97 χώρες, για την περίοδο 1950-2003 και σε δείγμα χωρών της Ευρωπαϊκής Ένωσης, αλλά και στο σύνολο των υπό εξέταση χωρών. Όμως, οι Burke και Emeric, (2016) επισημαίνουν προβλήματα στην εφαρμογή των panel δεδομένων όταν συμβαίνουν ξαφνικά μεγάλες καιρικές μεταβολές, αλλά και σε θέματα κλιματικής αλλαγής.

δ. Συγχρονική Ανάλυση

Οι Mendelsohn et al., (1994), Schlenker et al., (2004) και Kurukulasuriya και Mendelsohn, (2008) εφάρμοσαν τη «Συγχρονική ανάλυση (Cross – sectional analysis)» για να μελετήσουν την επίδραση των κλιματικών αλλαγών στον αγροτικό τομέα και οι Deschênes και Greenstone, (2007) στον τομέα της υγείας.

ε. Πολυκριτηριακές Αναλύσεις

Οι Kubal et al., (2009), Viguié και Hallegatte, (2012) με «πολυκριτηριακές αναλύσεις» εξέτασαν ζητήματα προσαρμογής στην κλιματική αλλαγή για τον αστικό χώρο και οι Julius και Scheraga, (2000) για τον αγροτικό χώρο. Πολυκριτηριακές αναλύσεις με δεδομένα γεωγραφικών περιοχών, προκειμένου να διερευνήσουν θέματα κλιματικών αλλαγών, πραγματοποίησαν πολλοί ερευνητές, μεταξύ των οποίων οι Smith και Lenhar, (1996) για την Αφρική, ο Qin et al., (2008) για τον Καναδά και οι Brouwer και Van Ek, (2004) και ο de Bruin et al., (2009) για τις κάτω χώρες.

στ. Πειραματική Ανάλυση

Οι τεχνικοί εφαρμόζουν την «πειραματική ανάλυση» για να εξετάσουν πως τα μέτρα για τον μετριασμό και την προσαρμογή στην κλιματική αλλαγή επηρεάζουν τις ηλεκτρικές και αγροτικές ανάγκες, πως είναι αποτελεσματικές οι στρατηγικές προσαρμογής και πως τα κλιματικά εισάγονται στις συναρτήσεις παραγωγής (Masseti και Mendelsohn, 2015).

ζ. Παίγνια

Για τη μελέτη του κόστους – ωφέλειας της προσαρμογής στην κλιματική αλλαγή και της εφαρμογής πράσινης ανάπτυξης λαμβάνεται υπόψη ότι οι εκπομπές αερίων θερμοκηπίου συμβάλλουν στην κλιματική αλλαγή ανεξάρτητα από την περιοχή ή τις δραστηριότητες που προέρχονται, ενώ οι συνέπειες της κλιματικής αλλαγής διαρκούν για πολύ μεγάλο χρονικό διάστημα. Επειδή τα αέρια του θερμοκηπίου ρυπαίνουν χώρες πέραν των χωρών που είναι κυρίως υπαίτιες για τη ρύπανση, οι χώρες διαπραγματεύονται τους όρους και τη συμμετοχή τους σε μία συμφωνία για το κλίμα, λαμβάνοντας υπόψη το δικό τους κόστος και όφελος. Στη

μελέτη των συμφωνιών αυτών για τη μείωση των εκπομπών των αερίων του θερμοκηπίου χρησιμοποιείται η «θεωρία των παιγνίων» (Cararo and Siniscalco, 1993, Murdoch and Sandler, 1997, Diamantoudi and Sartzetakis, 2006). Στα παίγνια αυτά οι χώρες θεωρούνται παίκτες που με κάποια πιθανότητα εφαρμόζουν διαφορετικές στρατηγικές ή μεικτές στρατηγικές για αναμενόμενες πληρωμές. Δηλαδή, κάθε χώρα – παίκτης επιλέγει τη στρατηγική που ελαχιστοποιεί την μέγιστη αναμενόμενη οφειλή λόγω ζημίας. Οι παίκτες – χώρες ενεργούν σύμφωνα με τα συμφέροντά τους, αλλά συνδέονται μεταξύ τους, μέσω των ζημιών από την κλιματική αλλαγή που εξαρτώνται από τις συνολικές εκπομπές των αερίων του θερμοκηπίου. Ταυτόχρονα οι επιλογές κάθε χώρας επηρεάζουν τις αποφάσεις των άλλων και αντίστροφα (Bank of Greece, 2018). Ειδικότερα, ο σχηματισμός της συμφωνίας αντιμετωπίζεται ως παίγνιο με συνεργατική συμπεριφορά, με μη συνεργατική συμπεριφορά και με τη συνεργατική συμπεριφορά που χαρακτηρίζει μια χώρα που σκέπτεται να αποχωρήσει υπολογίζοντας μόνο τις αντιδράσεις των άλλων χωρών και το μέγεθος του συνόλου των χωρών που θα παραμείνουν (Carraro and Siniscalco, 1991, 1993, Barret, 1994, Murdoch and Sandler, 1997, Chander and Tulkens, 1995, 1997, Rubio and Casino, 2001, 2005, Rubio and Ulph, 2006, Chander, 2007, Acheim et al., 2006, Diamantoudi and Sartzetakis, 2006, 2015, 2017).

η. Ολοκληρωμένα Υποδείγματα Αξιολόγησης

Στην εμπειρική ανάλυση σημαντικά ευρήματα έχουν ληφθεί από την χρήση των υποδειγμάτων που εξετάζουν τις αλληλεπιδράσεις μεταξύ κλίματος, ενέργειας, βιομηχανικών διεργασιών και οικονομίας (Bank of Greece, 2018), τα οποία είναι γνωστά ως «ολοκληρωμένα υποδείγματα αξιολόγησης (Integrated Assessment Models (IAM))», (Hope and Helm, 2005, Fussler, 2009). Κυριότερος εκπρόσωπος των IAM είναι το «δυναμικό ολοκληρωμένο υπόδειγμα κλίματος και οικονομίας (Dynamic Integrated Model of Climate and Economy (DICE))», που παρουσίασε ο βραβευμένος με το βραβείο Νόμπελ στην οικονομική επιστήμη το 2018 William Nordhaus. Οι Nordhaus και Sztorc, 2013, επιχείρησαν μια λεπτομερή περιγραφή του υποδείγματος DICE. Στα υποδείγματα αυτά τα αέρια του θερμοκηπίου μπορεί να επηρεάζουν την θερμοκρασία, τον όγκο και την ένταση των βροχοπτώσεων, τη νέφωση, τη συμπεριφορά των ανέμων, τη στάθμη της θάλασσας κλπ. Οι μεταβολές, όμως, που προκαλούνται έχουν ως συνέπεια κάποιες φυσικές, περιβαλλοντικές και κοινωνικές επιπτώσεις, που μάλιστα μπορεί να εκτιμηθούν ως πιθανό οικονομικό κόστος της κλιματικής αλλαγής, με μορφή χρηματικών αξιών και να ανάγονται σε μέτρηση της κοινωνικής ευημερίας.

Στα υποδείγματα αυτά υπάρχει μια αμφίδρομη σχέση, αφού η οικονομία επηρεάζει την κλιματική αλλαγή και η εκπομπή αερίων του θερμοκηπίου προέρχεται, κυρίως, από την παραγωγή και κατανάλωση των προϊόντων, οπότε εμφανίζεται αλυσίδα αλληλεπιδράσεων (de Bruin et al., 2009). Ιδιαίτερο ενδιαφέρον στις αλληλεπιδράσεις που παρατηρούνται έχει η διαπίστωση ότι γεγονότα ακραία ως προς τις κλιματικές μεταβολές, αλλά με μικρή πιθανότητα να πραγματοποιηθούν (ουρές της κανονικής κατανομής που συνήθως χρησιμοποιείται από τους ερευνητές) απαιτούν πολιτικές αποφασιστικών δράσεων (Malamud, 2004). Ο Weitzman, 2010c,

λόγω αυτής της διαπίστωσης αμφισβητεί την ικανότητα των υποδειγμάτων αυτών να παρέχουν ακριβείς εκτιμήσεις κόστους και οφέλους, ώστε να υπάρξει αποτελεσματική πολιτική, αφού τα θεωρούμενα λιγότερο πιθανά ενδεχόμενα μπορεί να προκαλέσουν ακραίες ζημιές. Συνεπώς, υποστηρίζει ότι η κατανομή θα πρέπει να έχει παχιά ουρά και προτείνει να χρησιμοποιείται η κατανομή Student t ή η κατανομή Pareto, αντί της κανονικής κατανομής. Γενικά, στα υποδείγματα που ενσωματώνουν αβεβαιότητα εφαρμόζεται η χρήση μιας κατάλληλης συνάρτησης πυκνότητας πιθανότητας, η οποία συνδέει τις πιθανότητες με τα διάφορα ενδεχόμενα. Έτσι, οι πιθανές μελλοντικές ζημιές σταθμίζονται με βάση την πιθανότητα πραγματοποίησης των αντίστοιχων ενδεχομένων, ώστε να υπολογιστούν αξιόπιστα στατιστικά μέτρα που αφορούν το ύψος των ζημιών (Quiggin, 2008). Παρ' όλα αυτά ο κίνδυνος μιας αιφνίδιας μεγάλης περιβαλλοντικής καταστροφής δεν πρέπει να αξιολογείται όπως της κλιματικής αλλαγής, επειδή έχει μοναδικά χαρακτηριστικά (Posner, 2005, Sunstein, 2009, Weitzman, 2010c).

2.3 Γεωγραφικό Σύστημα Στατιστικής Πληροφόρησης

Η εφαρμογή των στατιστικών μεθόδων και η αξιοπιστία της ανάλυσης δεδομένων εξαρτώνται σε μεγάλο βαθμό από το εύρος και την ποιότητα των δεδομένων που παρέχονται. Ιδιαίτερη σημασία έχει η οικοδόμηση και λειτουργία ενός ορθά σχεδιασμένου πλήρους «Γεωγραφικού Συστήματος Στατιστικής Πληροφόρησης» που να εξασφαλίζει το συντονισμό των δημόσιων και ιδιωτικών υπηρεσιών, φορέων, ερευνητικών κέντρων κλπ., με ενιαίες προδιαγραφές για τη συλλογή καταγραφή και αποθήκευση των στοιχείων. Το ολοκληρωμένο πληροφοριακό σύστημα θα πρέπει να αποτελεί ενιαίο φορέα και να διαθέτει αυτοματοποιημένη «Βάση Δεδομένων» που θα περιέχει στοιχεία, τα οποία θα αφορούν τους νομούς, τις περιφέρειες, τη χώρα (κλιματολογικά, φυσικά, ενεργειακά, γαιολογικά ποικιλότητας και δασικών οικοσυστημάτων, γεωργικά, υδροφόρου ορίζοντα, παράκτιας ζώνης, τοπογραφικά, χρήσης γης, δομημένου περιβάλλοντος, πληθυσμιακά, οικονομικά, μετανάστευσης, τουρισμού, υγείας κλπ.). Η αναφορά σε ένα Εθνικό Σύστημα Συντεταγμένων επιτρέπει την κωδικοποίηση όλων των δεδομένων με την κωδικοποίηση των δεδομένων που αντιστοιχούν στα εδαφοτεμάχια ενός «Πολύσκοπου Κτηματολογίου» (Donatos, 1986) και παρέχει τη δυνατότητα σύνδεσης με την ευρύτερη περιοχή (πχ. της Ευρωπαϊκής Ένωσης) στην περίπτωση λειτουργίας αντίστοιχων Βάσεων Δεδομένων με ενιαίες προδιαγραφές και χρήσης κοινού γαιωδετικού συστήματος αναφοράς (Δονάτος, 1986, Δονάτος, 1996).

Με τη δημιουργία Γεωγραφικού Συστήματος Στατιστικής Πληροφόρησης με βάση το Πολύσκοπο Κτηματολόγιο (Donatos, 1986, Δονάτος, 2005) τα δεδομένα που ανήκουν στη διευρυμένη Βάση Δεδομένων χαρακτηρίζονται από ακρίβεια, πληρότητα, επάρκεια, συνέπεια και αξιοπιστία σε σχέση με τον τόπο, το αντικείμενο

και το χρόνο και αξιοποιούνται στατιστικά. Δηλαδή, συγκεντρώνονται μεθοδικά, κατατάσσονται ομοιόμορφα και συστηματικά, γίνεται κατάλληλη παρουσίασή τους και χρησιμοποιούνται στον υπολογισμό μέτρων και δεικτών και στην εκτίμηση υποδειγμάτων, που έχουν ιδιαίτερη σημασία για την αξιολόγηση καταστάσεων και τη λήψη αποφάσεων για την ορθή και ολοκληρωμένη πράσινη ανάπτυξη της χώρας.

3. ΕΠΙΛΟΓΟΣ

Την τελευταία δεκαετία, οι περιβαλλοντικοί κίνδυνοι κυριαρχούν προοδευτικά στο παγκόσμιο τοπίο των κινδύνων. Είναι χαρακτηριστικό ότι, σύμφωνα με την έκθεση του World Economic Forum για το 2019, όπως και τα δύο προηγούμενα χρόνια, οι τρεις από τους πέντε σοβαρότερους παγκόσμιους κινδύνους, από άποψη πιθανότητας σημαντικών οικονομικών επιπτώσεων, είναι περιβαλλοντικοί με κυριότερο την κλιματική αλλαγή.

Στο τέλος του 21ου αιώνα προβλέπεται ότι οι επιπτώσεις της κλιματικής αλλαγής και των άλλων απειλών για την επίτευξη της περιβαντολλογικής ισορροπίας θα είναι καθοριστικές για την εφαρμογή της πράσινης ανάπτυξης στην Ελλάδα. Ιδιαίτερα, ενδιαφέροντα, σχετικά με τη δυνατότητα πραγματοποίησης της πράσινης ανάπτυξης, είναι οι προβλέψεις και ποικίλα άλλα ευρήματα που προέκυψαν από τη στατιστική μελέτη των περιβαντολλογικών, οικονομικών και κοινωνικών επιπτώσεων της κλιματικής αλλαγής που εκπονήθηκε από μια πλειάδα ερευνητών διαφόρων επιστημονικών ειδικοτήτων, μετά από πρωτοβουλία της Τράπεζας της Ελλάδος, η οποία προέβη στη σύσταση της Επιτροπής της μελέτης (Τράπεζα Ελλάδος, 2011).

Η παγκόσμια οικονομική κρίση αποτελεί πρόκληση για τη χρήση της Στατιστικής για την μελέτη των περιβαντολλογικών και κοινωνικοοικονομικών προβλημάτων, την εκτίμηση των επιπτώσεών τους, την πρόβλεψη επικίνδυνων καταστάσεων και την ανάλυση κόστους - ωφέλειας. Έτσι, από την ανάλυση των δεδομένων θα ληφθεί ολοκληρωμένη πληροφόρηση, ώστε να παρθούν από τους αρμοδίους οι αποφάσεις εκείνες για το σχεδιασμό κατάλληλης πολιτικής και την εφαρμογή των αναγκαίων παρεμβάσεων, καθώς και την υλοποίηση στοχευμένων προγραμμάτων που οδηγούν στην πράσινη ανάπτυξη. Η αισιοδοξία για την αντιμετώπιση των περιβαντολλογικών κινδύνων που εγκυμονούν και για την επίτευξη μιας πράσινης ανάπτυξης συνεχίζει να εκφράζεται μέσα από τη φράση: «Έχουμε ακόμη το χρόνο και τη γνώση να δράσουμε. Αλλά μόνο αν δράσουμε συστηματικά, αντικειμενικά, δυναμικά και γρήγορα».

ABSTRACT

This paper presents the problems and actions for achieving green growth and attempts to examine the contribution of statistical methods to the study of achieving environmental sustainability. In particular, to illustrate the contribution of Statistics, reference is made to the various statistical and econometric models used to investigate

the magnitude and impact of the increasing environmental damage that occurs, the cost-benefit relationship of applying measures to reduce costs, the effects of climate change and other environmental risks, the ability to satisfy people with the proper use of environmental goods, and other important environmental matters. In addition, in order to achieve the reliability of the statistical analysis of sufficient data regarding the area, the object and the time it is proposed to build and operate a Geographic Statistical Information System based on a Multipurpose Cadastre.

ΑΝΑΦΟΡΕΣ

- Asheim, G.B., Froyen C.B., Hovi J. and Menz F.C. (2006). Regional versus global cooperation for climate control, *Journal of Environmental Economics and Management*, 51, 1, 93–109.
- Auffhammer, M. and Aroonruengsawat A. (2011). Simulating the impacts of climate change, prices and population on California’s residential electricity consumption, *Climatic Change*, 109, 1, 191–210.
- Auffhammer, M. and Mansur E.T. (2014). Measuring climatic impacts on energy consumption: A review of the empirical literature, *Energy Economics*, 46, 522–530.
- Auffhammer, M. and Schlenker W. (2014). Empirical studies on agricultural impacts and adaptation, *Energy Economics*, 46, 555–561.
- Banerjee, L. (2007). Effect of flood on agricultural wages in Bangladesh: An empirical analysis, *World Development*, 35, 11, 1989–2009.
- Bank of Greece (2018). *The economics of climate change*, Athens.
- Barrett, S. (1994a). Self-enforcing international environmental agreements, *Oxford Economic Papers*, 46, 878–894.
- Bateman, I.J., Carson, R.T., Day, B., Hanemann, M., Hanley, N., Hett, T., Jones-Lee, M., Loomes, G., Mourato, S., Ozdemiroglu, E., Pearce, D.W., Sudgen, R. and Swanson, J. (2002). *Economic evaluation with stated preference techniques: A manual*, Elgar, Cheltenham.
- Bhave, A.G., Conway D., Dessai S. and Stainforth D.A. (2016). Barriers and opportunities for robust decision making approaches to support climate change adaptation in the developing world, *Climate Risk Management*, 14, 1–10.
- Brouwer, R. and Van Ek R. (2004). Integrated ecological, economic and social impact assessment of alternative flood control policies in the Netherlands, *Ecological Economics*, 50, 1, 1–21.
- Burke, M. and Emerick K. (2016). Adaptation to climate change: Evidence from US agriculture, *American Economic Journal: Economic Policy*, 8, 3, 106–140.
- Carraro, C. and Siniscalco D. (1991). Strategies for the international protection of the environment, Discussion Paper 568, *Centre for Economic Policy Research*, London.

- Carraro, C. and Siniscalco D. (1993). Strategies for the international protection of the environment, *Journal of Public Economics*, 52, 3, 309–328.
- Carraro, C. and Siniscalco D. (1995). R&D cooperation and the stability of international environmental agreements, *CEPR Discussion Paper*, No. 1154.
- Chander, P. (2007). The gamma-core and coalition formation, *International Journal of Game Theory*, 35, 4, 539–556.
- Chander, P. and Tulkens H. (1995). A core-theoretic solution for the design of cooperative agreements on transfrontier pollution, *International Tax and Public Finance*, 2, 2, 279–293.
- Chander, P. and Tulkens H. (1997). The core of an economy with multilateral environmental externalities, *International Journal of Game Theory*, 36, 379–401.
- Clark, A. and Oswald, A. (1994). Unhappiness and unemployment. *Economic Journal*, 104, 648–659.
- Cole, M.A., Rayner, A.J. and Bates, J.M. (1997). The Environmental Kuznets curve: an empirical analysis. *Environment and Development Economics*, 2, 4, 401–416.
- Cole, M.A. (2004). Trade, the pollution haven hypothesis and environmental Kuznets curve: examining the linkages, *Ecological Economics*, 48, 71–81.
- Cole, M.A. (2005). Re-examining the pollution-income relationship: a random coefficients approach, *Economics Bulletin*, 14, 1–7.
- Dasgupta, P. (2008). Discounting climate change, *Journal of Risk and Uncertainty*, Vol. 37, No. 2, 141–69.
- de Bruin, K.C., Dellink R.B. and Tol R.S.J. (2009). AD-DICE: An implementation of adaptation in the DICE model, *Climatic Change*, 95, 1, 63–81.
- Dell, M., Jones B.F. and Olken B.A. (2014). What do we learn from the weather? The new climate–economy literature, *Journal of Economic Literature*, 52, 3, 740–798.
- Dell, M., Jones B.F. and Olken B.A. (2014). What do we learn from the weather? The new climate–economy literature, *Journal of Economic Literature*, 52, 3, 740–798.
- Deschênes, O. and Greenstone M. (2007). The economic impacts of climate change: Evidence from agricultural output and random fluctuations in weather, *American Economic Review*, 97, 1, 354–385.
- Deschênes, O. and Greenstone M. (2011). Climate change, mortality, and adaptation: Evidence from annual fluctuations in weather in the US, *American Economic Journal: Applied Economics*, 3, 4, 152–185.
- Desvousges, W.H., Smith, V.K. and Fisher A. (1987). Option Price Estimates for Water Quality Improvements: A Contingent Valuation Study for the Monongahela River, *Journal of Environmental Economics and Management*, 14, 3, 248–267.
- Diamantoudi, E. and Sartzetakis E.S. (2006). Stable international environmental agreements: An analytical approach, *Journal of Public Economic Theory*, 8, 2, 247–263.
- Diamantoudi, E. and Sartzetakis E.S. (2015). International environmental agreements – Coordinated action under Foresight, *Economic Theory*, 59, 527–546.
- Diamantoudi, E. and Sartzetakis E.S. (2017). International environmental agreements: The role of Foresight, *Environmental and Resource Economics*, DOI: 10.1007/s10640-017-0148-1.

- DiTella, R., MacCulloch, R.J. and Oswald, A.J. (2001). Preferences over inflation and unemployment: Evidence from surveys of happiness. *The American Economic Review*, 91, 335-341
- Δονάτος, Γ. (1981). Η εκτίμηση της αξίας της γης βασισμένη στην τράπεζα πληροφοριών γης του πολυσήμαντου κτηματολογίου, Σπουδαί, 1193-1212.
- Donatos, G. (1986). A land information system based on a multipurpose cadaster, *The Valuer*, XXIX.2, 82-84.
- Δονάτος, Γ. (1996). Πολύσκοπο κτηματολόγιο και στατιστική πληροφόρηση, *Πρακτικά 9^ο Πανελληνίου Συνεδρίου Στατιστικής*, Ε.Σ.Ι., Ξάνθη, 80-88.
- Δονάτος, Γ. (2005). Ένα ολοκληρωμένο σύστημα στατιστικής πληροφόρησης για την περιφερειακή ανάπτυξη, *Μελέτες προς τιμήν του καθηγητή Θεοδώρου Α. Σκούντζου*, Πανεπιστήμιο Πειραιώς, Πειραιάς, 423-428.
- Easterlin, R.A. (2001). Income and happiness: Towards a unified theory, *The Economic Journal*, 111, 465-484.
- Grossman, G. and Krueger, A. (1995). Economic Growth and the Environment, *Quarterly Journal of Economics*, 112, 2, 353-377.
- Fankhauser, S. (2017). Adaptation to climate change, *Annual Review of Resource Economics*, 9, 1, 209-230.
- Füssel, H. (2009). An updated assessment of the risks from climate change based on research published since the IPCC Fourth Assessment Report, *Climatic Change*, 97, 3-4, 469-482.
- Halkos, G. and Tsionas, E. (2001). Environmental Kuznets curves: Bayesian evidence from switching regime models, *Energy Economics*, 23, 2, 191-210.
- Halkos, G. (2003). Environmental Kuznets Curve for sulfur: evidence using GMM estimation and random coefficient panel data models, *Environment and Development Economics*, 8, 4, 581-601.
- Halkos, G. (2011). Nonparametric modelling of biodiversity: Determinants of threatened species, *Journal of Policy Modeling*, 33, 4, 618-635.
- Halkos, G. (2012). Environmental Pollution And Economic Development: Explaining The Existence Of An Environmental Kuznets Curve, *Journal of Applied Economic Sciences*, 6, 6(18), 148-159.
- Halkos, G. and Jones, N. (2012). Modeling the effect of social factors on improving biodiversity protection, *Ecological Economics*, 78, C, 90-99.
- Halkos, G. and Matsiori, S. (2017). Environmental attitude, motivations and values for marine biodiversity protection, *Journal of Behavioral and Experimental Economics* (formerly *Journal of Socio-Economics*), 69, C, 61-70.
- Hanley, N. and Spash, C.L. (1993). *Cost-benefit analysis and the environment*, Edward Elgar Publishing Ltd., UK.
- Heal, G. and Millner A. (2014). Reflections: Uncertainty and decision making in climate change economics, *Review of Environmental Economics and Policy*, 8, 1, 120-137.
- Hope, C. and Helm D. (2005). Integrated assessment models, in Helm D. (ed), *Climate-Change Policy*, Oxford University Press, Oxford, 77-98.

- Hsiang, S. (2016). Climate econometrics, *Annual Review of Resource Economics*, 8, 1, 43–75.
- Julius, S.H. and Scheraga J.D. (2000). The TEAM model for evaluating alternative adaptation strategies, *Research and Practice in Multiple Criteria Decision Making*, 319–330.
- Kahneman, D., Wakker, P.P. and Sarin, R. (1997). Back to Bentham? Explorations of experienced utility, *Quarterly Journal of Economics*, 112, 375-405.
- Kahneman, D. and Sudgen, R. (2005). Experienced utility as a standard of policy evaluation, *Environmental and Resource Economics*, 32, 161–181.
- Kountouri, P., Kountouris, Y. and Remoundou, K. (2009). Valuing a wing farm construction: a contingent valuation study in Greece, *Energy Policy*, 37, 1939-1944.
- Kubal, C., Haase D., Meyer V. and Scheuer S. (2009). Integrated urban flood risk assessment – adapting a multicriteria approach to a city, *Natural Hazards and Earth System Sciences*, 9, 6, 1881–1895.
- Kurukulasuriya, P. and Mendelsohn R. (2008). Crop switching as a strategy for adapting to climate change, *African Journal of Agricultural and Resource Economics*, 2, 1, 105–126.
- Lempert, R.J. and Collins M.T. (2007). Managing the risk of uncertain threshold responses: Comparison of robust, optimum, and precautionary approaches, *Risk Analysis*, 27, 4, 1009–1026.
- Lempert, R.J. and Kalra N. (2011). Managing climate risks in developing countries with robust decision making, *World Resources Report*, Washington, DC.
- Loomis, J.B. and White, D.S. (1996). Economic benefits of rare and endangered species: summary and meta-analysis, *Ecological Economics*, 18, 197-206.
- Markandya, A., Golub, A. and Pedrozo-Galinato, S. (2006). Empirical Analysis of National Income and SO₂ Emissions in Selected European Countries, *Environmental & Resource Economics*, 35, 221-257.
- Massetti, E. and Mendelsohn R. (2015). The economics of adaptation to climate change, *Working paper*, Yale University.
- Mendelsohn, R., Nordhaus W.D. and Shaw D. (1994). The impact of global warming on agriculture: A Ricardian analysis, *American Economic Review*, 84, 4, 753–771.
- Mitchell, R.C. and Carson, R.T. (1989). Using surveys to value public goods: The contingent valuation method, *Resources for the Future*, Washington, D.C.
- Murdoch, J. and Sandler T. (1997). The voluntary provision of a pure public good: The case of reduced CFC emissions and the Montreal Protocol, *Journal of Public Economics*, 63, 3, 331–349.
- Nordhaus, W. (2007). A review of the Stern Review on the economics of climate change, *Journal of Economic Literature*, XLV, 686–702.
- Nordhaus, W.D. (2014). Estimates of the social cost of carbon: Concepts and results from the DICE-2013R model and alternative approaches, *Journal of the Association of Environmental and Resource Economists*, 1, 1/2, 273–312.
- Nordhaus, W.D. and Sztorc P. (2013). *DICE 2013-R: Introduction and User's Manual*, Yale University, New Haven, CT.

- Nunes, P.A.L.D. and van den Bergh, J.C.J.M. (2001). Economic valuation of biodiversity: sense or nonsense? *Ecological Economics*, 39, 203-222.
- Panayotou, T. (1993). Empirical Tests and Policy Analysis of Environmental Degradation at Different Stages of Economic Development. *Working Paper* WP238, Technology and Employment Programme, International Labour Office, Geneva.
- Qin, X.-S., Huang G.H., Chakma A., Nie X.H. and Lin Q.G. (2008). A MCDM-based expert system for climate-change impact assessment and adaptation planning – A case study for the Georgia Basin, Canada, *Expert Systems with Applications*, 34, 3, 2164–2179.
- Quiggin, I. (2008). Uncertainty and Climate Change Policy, *Economic Analysis and Policy*, 38,2.
- Ramsey, F.P. (1926). Truth and probability, in Braithwaite, R. (ed) *The Foundations of Mathematics and Other Logical Essays*, Kegan, Paul, Trench, Trubner & Co, London.
- Ranger, N. and Surminski S. (2013). A preliminary assessment of the impact of climate change on non-life insurance demand in the BRICS economies, *International Journal of Disaster Risk Reduction*, 3, 14–30.
- Reid, W.V., McNeely J.A., Tunstall D.B., Bryant D.A. and Winograd M. (1993). *Biodiversity indicators for policy-makers*, IUCN The World Conservation Union.
- Rehbanz, K. (2007). Species diversity and human well-being: a spatial econometric approach, *Working Paper FNU-151*. Hamburg University and Centre of Marine and Atmospheric Science, Germany.
- Rubio, J.S. and Casino B. (2001). International cooperation in pollution control, *Working Paper*, Series AD 2001-21, Instituto Valenciano de Investigaciones Economicas.
- Rubio, S.J. and Ulph A. (2006). Self-enforcing international environmental agreements revisited, *Oxford Economic Papers*, 58, 2, 233– 263.
- Rubio, S.J. and Casino B. (2005). Self-enforcing international environmental agreements with a stock pollutant, *Spanish Economic Review*, 7, 2, 89–109.
- Schlenker, W., Hanemann M. and Fisher A. (2004). *Will US agriculture really benefit from global warming? Accounting for irrigation in the hedonic approach*, Department of Agricultural and Resource Economics, University of California, Berkeley.
- Seo, S.N. (2011). A geographically scaled analysis of adaptation to climate change with spatial models using agricultural systems in Africa, *The Journal of Agricultural Science*, 149, 4, 437–449.
- Seo, S.N. and Mendelsohn R. (2008). Animal husbandry in Africa: Climate change impacts and adaptations, *African Journal of Agricultural and Resource Economics*, 2, 1, 65–82.
- Seo, S.N., Mendelsohn R., Dinar A. and Kurukulasuriya P. (2009). Adapting to climate change mosaically: An analysis of African livestock management by agroecological zones, *The B.E. Journal of Economic Analysis & Policy*, 9, 2, 1–35.
- Smith, J.B. and Lenhart S.S. (1996). Climate change adaptation policy options, *Climate Research*, 6, 2, 193–201.

- Selden, T.M., and Song, D. (1994). Environmental Quality and Development: Is There a Kuznets Curve for Air Pollution Emissions? *Journal of Environmental Economics and Management*, 27, 2, 147-162.
- Shafik, N. (1994). Economic Development and Environmental Quality: An Econometric Analysis, *Oxford Economic Papers*, 46, 757-773.
- Shammin, R. (1999). *Application of the travel cost method (TCM): A case study of environmental valuation of Dhaka zoological garden*. IUCN World Conservation Union.
- Stern D.I., Common M.S. and Barbier E.B, (1996). Economic growth and environmental degradation: the environmental Kuznets curve and sustainable development, *World Development*, 24, 1151-1160.
- Stern D.I. and Common M.S. (2001). Is there an environmental Kuznets curve for sulphur? *Journal of Environmental Economics and Management*, 41, 162-178.
- Stern, N. (2007). *The Economics of Climate Change: The Stern Review*, Cambridge University Press, Cambridge, UK.
- Stern, N. (2008). The economics of climate change, *American Economic Review, Papers and Proceedings*, 98, 1–37.
- Τράπεζα της Ελλάδος (2011). *Οι περιβαλλοντικές και κοινωνικές επιπτώσεις της κλιματικής αλλαγής στην Ελλάδα*, Αθήνα.
- Tyrväinen L. (1997). The amenity value of the urban forest: an application of the hedonic pricing method. *Landscape and Urban Planning*, 37, 3–4, 211-222.
- Viguié, V. and Hallegatte S. (2012). Trade-offs and synergies in urban climate policies, *Nature Climate Change*, 2, 5, 334–337.
- Viscusi, W.K. and Aldy J.E. (2003). The value of a statistical life: A critical review of market estimates throughout the world, *Journal of Risk and Uncertainty*, 27, 1, 5–76.
- Watkiss, P.L.S., Horrocks, S., Searl, A.S. and Hunt, A.A. (2009). Impact of Climate Change in Human Health in Europe, PESETA – *Human Health Study*, EUR 24135EN.
- Weitzman, M. (2010). What is the ‘damages function’ for global warming – and what difference might it make?, *Climate Change Economics*, 1, 1, 57–69.
- Weitzman, M.L. (1998). Why the far-distant future should be discounted at its lowest possible rate, *Journal of Environmental Economics and Management*, 36, 201–208.
- Weitzman, M.L. (2001). Gamma discounting, *American Economic Review*, 91, 260–271.
- Weitzman, M.L. (2009a). Additive Damages, Fat-Tailed Climate Dynamics, and Uncertain Discounting, *Economics E-Journal*, <http://www.economics-ejournal.org/economics/journalarticles/2009-39>.
- Weitzman, M.L. (2009b). On Modeling and Interpreting the Economics of Catastrophic Climate Change, *Review of Economics and Statistics*, Vol. 91, No. 1, 1-19.
- Weitzman, M. L. (2009c). Reactions to the Nordhaus critique, *mimeo*, Harvard University.

- Weitzman, M. L. (2009d). Some Basic Economics of Extreme Climate Change, *mimeo*, Harvard University.
- Weitzman, M.L. (2010a). GHG Targets as Insurance against Catastrophic Climate Damages, *Technical Report Working Paper, 16136*, National Bureau of Economic Impacts.
- Weitzman, M.L. (2010b). What is the ‘damages function’ for global warming – and what difference might it make? *Climate Change Economics*, Vol. 1, No. 1, 57.
- Weitzman, M.L. (2010c). Fat-tailed uncertainty in the economics of catastrophic climate change, *Review of Environmental Economics and Policy*, 5, 275–292.
- Wilby, R.L. and Dessai S. (2010). Robust adaptation to climate change, *Weather*, 65, 7, 180–185.
- Χάλλκος, Γ. (2016). *Οικονομική Φυσικών Πόρων και Περιβάλλοντος*. Εκδόσεις ΔΙΣΙΓΜΑ.

εργασίες

στα ελληνικά



ΔΙΑΓΡΑΜΜΑΤΑ ΕΛΕΓΧΟΥ ΤΥΠΟΥ ΕWMA ΓΙΑ ΤΗΝ ΠΑΡΑΚΟΛΟΥΘΗΣΗ ΔΙΕΡΓΑΣΙΩΝ ΜΕ ΠΕΠΕΡΑΣΜΕΝΟ ΕΥΡΟΣ ΤΙΜΩΝ

Αναστασοπούλου Μ., Ρακιτζής Α.

Εργαστήριο Στατιστικής και Ανάλυσης Δεδομένων, Τμήμα Στατιστικής και
Αναλογιστικών-Χρηματοοικονομικών Μαθηματικών, Πανεπιστήμιο Αιγαίου
anastasopoulou@aegean.gr, arakitz@aegean.gr

ΠΕΡΙΛΗΨΗ

Οι διεργασίες οι οποίες χαρακτηρίζονται από την παρουσία συσχέτισης μεταξύ των διαδοχικών παρατηρήσεων με πεπερασμένο εύρος τιμών συχνά μοντελοποιούνται ως ένα Διωνυμικό Αυτοπαλίνδρομο μοντέλο 1ης τάξης (Binomial AR(1) ή BAR(1)). Όμως, σε πολλές πρακτικές εφαρμογές παρατηρείται υπερσκέδαση στα δεδομένα, οπότε ένα καταλληλότερο υπόδειγμα σε αυτή την περίπτωση είναι το Βήτα-Διωνυμικό Αυτοπαλίνδρομο μοντέλο 1^{ης} τάξης (Beta-Binomial AR(1) ή BBAR(1)). Στόχος της παρούσας εργασίας είναι η ανάπτυξη και μελέτη διαγραμμάτων ελέγχου τύπου EWMA για την παρακολούθηση των παραπάνω διεργασιών, προκειμένου να ανιχνευθούν τυχόν αλλαγές στις τιμές των παραμέτρων. Δίδεται ο στατιστικός σχεδιασμός των προτεινόμενων διαγραμμάτων ελέγχου καθώς και αριθμητικές συγκρίσεις με τα αντίστοιχα διαγράμματα τύπου Shewhart, με κριτήριο το μέσο μήκος ροής.

Λέξεις Κλειδιά: BAR(1), Beta-BAR(1), Μέσο μήκος ροής, Στατιστικός έλεγχος διεργασιών.

1. ΕΙΣΑΓΩΓΗ

Οι τεχνικές στατιστικής παρακολούθησης διεργασιών αποτελούν το βασικό εργαλείο του στατιστικού ελέγχου διεργασιών (ΣΕΔ). Κύριος εκπρόσωπός τους είναι το διάγραμμα ελέγχου (control chart) το οποίο μπορεί να χρησιμοποιηθεί για την έγκαιρη και έγκυρη ανίχνευση μη-φυσιολογικών (συνήθως ανεπιθύμητων) καταστάσεων. Παραδοσιακά, τα διαγράμματα ελέγχου χρησιμοποιούνται στη βιομηχανία για την παρακολούθηση παραγωγικών διεργασιών, με σκοπό την ανίχνευση της χειροτέρευσης της ποιότητας των παραγόμενων αντικειμένων (π.χ. αύξηση στο ποσοστό των ελαττωματικών). Τα συνήθη διαγράμματα ελέγχου σε αυτή την περίπτωση είναι τα διαγράμματα ελέγχου τύπου Shewhart p και np (Montgomery (2009)): το πρώτο αφορά την παρακολούθηση του ποσοστού των ελαττωματικών, ενώ το δεύτερο αφορά την παρακολούθηση του αριθμού των ελαττωματικών. Βασική

υπόθεση για την εφαρμογή των διαγραμμάτων ελέγχου p και np είναι ότι ο αριθμός των ελαττωματικών είναι μια διωνυμική τυχαία μεταβλητή (τ.μ.).

Εκτός των διαγραμμάτων p και np , οι Gan (1993), Chang and Gan (2001) και Wu et al. (2008) μελέτησαν διαγράμματα ελέγχου τύπου CUSUM και EWMA. Είναι γνωστό ότι τα διαγράμματα ελέγχου Shewhart δεν είναι ευαίσθητα στην ανίχνευση μικρών και μεσαίων μετατοπίσεων στο μέσο επίπεδο της διεργασίας (π.χ. στο μέσο αριθμό ελαττωματικών αντικειμένων). Τα διαγράμματα CUSUM και EWMA, ως διαγράμματα ελέγχου με μνήμη, ανιχνεύουν ταχύτερα τέτοιου είδους μεταβολές.

Η συνήθης υπόθεση για την κατασκευή ενός np -chart είναι ότι οι διαδοχικές μετρήσεις που λαμβάνονται αποτελούν ανεξάρτητες και ισόνομες τ.μ. Στην πράξη και ειδικά στις σύγχρονες διεργασίες, οι διαδοχικές μετρήσεις χαρακτηρίζονται από την παρουσία συσχέτισης μεταξύ τους, κάτι το οποίο μπορεί να οφείλεται στην υψηλή συχνότητα λήψης δειγμάτων από τη διεργασία (Psarakis and Papaleonida (2007)). Σε αυτή την περίπτωση τα διαγράμματα ελέγχου που αναφέρθηκαν προηγουμένως δεν μπορούν να χρησιμοποιηθούν λόγω του αυξημένου ποσοστού εσφαλμένων συναγερωμών.

Μια λύση σε αυτό το πρόβλημα αποτελεί, αρχικά, η επιλογή ενός κατάλληλου μοντέλου χρονολογικών σειρών με ακέραιες τιμές και στη συνέχεια, η ανάπτυξη διαγραμμάτων ελέγχου με βάση το μοντέλο αυτό. Ειδικότερα, στην περίπτωση που ενδιαφέρει η παρακολούθηση του αριθμού των ελαττωματικών σε ένα δείγμα μεγέθους n αντικειμένων, το μοντέλο αυτό θα πρέπει να λαμβάνει πεπερασμένο πλήθος τιμών.

Ο Weiß (2009) ήταν ο πρώτος που μελέτησε διαγράμματα ελέγχου τύπου Shewhart και κινούμενου μέσου (Moving Average, MA) για την παρακολούθηση διεργασιών που μοντελοποιούνται σύμφωνα με ένα διωνυμικό αυτοπαλίνδρομο μοντέλο $1^{ης}$ τάξης (Binomial AR(1) ή BAR(1), McKenzie (1985)). Το μοντέλο BAR(1) αποτελεί ένα απλό και κατανοητό μοντέλο για δεδομένα από διωνυμική κατανομή, τα οποία εμφανίζουν συσχέτιση $1^{ης}$ τάξης. Οι Rakitzis et al. (2017) ανέπτυξαν διαγράμματα ελέγχου τύπου Shewhart και CUSUM για την ανίχνευση αλλαγών σε μοντέλα χρονολογικών σειρών με ακέραιες τιμές και πεπερασμένο εύρος. Εκτός του μοντέλου BAR(1), οι Rakitzis et al. (2017) ανέπτυξαν διαγράμματα ελέγχου και με βάση το Βήτα-διωνυμικό Αυτοπαλίνδρομο μοντέλο $1^{ης}$ τάξης (Beta-Binomial AR(1) ή BBAR(1), Weiß and Kim (2014)). Το μοντέλο BBAR(1) εμφανίζεται σε πρακτικές εφαρμογές όπου παρατηρείται ετερογένεια μεταξύ των μονάδων (π.χ. η πιθανότητα ελαττωματικού δεν είναι η ίδια για τις n μονάδες στο δείγμα) ή η μεταβλητότητα των δεδομένων παρουσιάζει σημαντική απόκλιση από την αντίστοιχη του διωνυμικού μοντέλου, κάτι που αναφέρεται με τον όρο υπερσκέδαση (*overdispersion*).

Ως ένδειξη ύπαρξης ή όχι υπερσκέδασης στα δεδομένα, χρησιμοποιείται ο διωνυμικός δείκτης μεταβλητότητας (*binomial index of dispersion*) I_d . Έστω X τ.μ. με τιμές στο $\{0, \dots, n\}$, με μέση τιμή μ και διασπορά σ^2 . Τότε ο δείκτης I_d ισούται με

$$I_d = \frac{n\sigma^2}{\mu(n-\mu)} \in (0, \infty) \quad (1)$$

και η τιμή του αποτελεί ένδειξη του κατά πόσο διαφέρει η διακύμανση από τη μέση τιμή. Στην περίπτωση της διωνυμικής κατανομής $B(n, \pi)$, το $I_d = 1$ για κάθε $\pi \in (0, 1)$. Επομένως, οποιαδήποτε τ.μ. με πεπερασμένο εύρος τιμών και $I_d > 1$, παρουσιάζει υπερσκέδαση (*extra-binomial variation*), δηλαδή μεγαλύτερη μεταβλητότητα σε σύγκριση με τη μεταβλητότητα της κατανομής $B(n, \pi)$. Άρα σε δεδομένα που παρουσιάζουν υπερσκέδαση θα πρέπει να χρησιμοποιηθεί ένα εναλλακτικό μοντέλο από το BAR(1).

Στόχος της παρούσας εργασίας είναι η ανάπτυξη και μελέτη μονόπλευρων διαγραμμάτων ελέγχου τύπου EWMA για την παρακολούθηση διεργασιών BAR(1) και BBAR(1). Κάτι τέτοιο δεν έχει συμβεί έως τώρα στη βιβλιογραφία. Τα προτεινόμενα διαγράμματα βασίζονται σε στατιστικές συναρτήσεις που λαμβάνουν ακέραιες τιμές. Αυτό επιτρέπει τον ακριβή υπολογισμό της απόδοσής τους. Η διάρθρωση της εργασίας είναι η εξής: Στην Ενότητα 2 παρουσιάζονται συνοπτικά οι ιδιότητες των μοντέλων BAR(1) και BBAR(1). Στην Ενότητα 3 παρουσιάζεται το διάγραμμα ελέγχου s -EWMA. Στην Ενότητα 4 παρουσιάζονται τα αποτελέσματα αριθμητικής μελέτης σχετικά με την απόδοση των άνω μονόπλευρων διαγραμμάτων s -EWMA για διεργασίες BAR(1) και BBAR(1). Στην Ενότητα 5 δίδεται μια πρακτική εφαρμογή του διαγράμματος s -EWMA σε πραγματικά επιδημιολογικά δεδομένα. Τέλος, στην Ενότητα 6 συνοψίζονται τα συμπεράσματα της μελέτης.

2. ΜΟΝΤΕΛΑ ΔΙΑΚΡΙΤΩΝ ΧΡΟΝΟΣΕΙΡΩΝ ΜΕ ΠΕΠΕΡΑΣΜΕΝΟ ΕΥΡΟΣ ΤΙΜΩΝ

2.1 Το μοντέλο BAR(1)

Το μοντέλο BAR(1) προτάθηκε από τον McKenzie (1985) (δείτε επίσης Al-Osh and Alzaid (1987)) και είναι ένα υπόδειγμα κατάλληλο για διωνυμικά δεδομένα με συσχέτιση 1^{ης} τάξης.

Ορισμός 2.1: Έστω $n \in \mathbb{N}$. Οι παράμετροι του μοντέλου είναι οι π, ρ με $\pi \in (0, 1)$, $\rho \in ([\max(-\pi/(1-\pi), -(1-\pi)/\pi)], 1)$ και ορίζεται από τη σχέση:

$$X_t = \alpha \circ X_{t-1} + \beta \circ (n - X_{t-1}), \quad t \geq 1, \quad (2)$$

όπου $X_0 \sim B(n, \pi)$, $\beta = \pi(1-\rho)$ και $\alpha = \beta + \rho$. Οι τ.μ. $\alpha \circ X_{t-1}$ και $\beta \circ (n - X_{t-1})$ είναι αμοιβαία ανεξάρτητες ακολουθίες ανεξάρτητων τ.μ. Bernoulli με παραμέτρους α και β , αντίστοιχα. Επίσης, είναι ανεξάρτητες των $X_s, s < t$.

Ο τελεστής « \circ » είναι γνωστός ως τελεστής διωνυμικής λέπτυνσης (*binomial thinning operator*, Steutel and van Harn (1979)). Αν N είναι μια μη αρνητική διακριτή τ.μ. και $\alpha \in (0, 1)$, τότε για δεδομένη τιμή της N , η τ.μ. $\alpha \circ N = X_1 + \dots + X_N$, ακολουθεί τη διωνυμική κατανομή με παραμέτρους N και α . Κατά συνέπεια, οι τιμές X_t όπως ορίζονται στη (2), είναι πάντοτε ακέραιες.

Το μοντέλο BAR(1) έχει ως περιθώρια κατανομή τη $B(n, \pi)$, δηλ. η $p_x = P(X_t = x)$ υπολογίζεται μέσω της συνάρτησης πιθανότητας της $B(n, \pi)$. Άρα, ισχύει ότι $E(X_t) = n\pi$ και $V(X_t) = n\pi(1 - \pi)$. Αν συμβολίσουμε με $\rho(k) = \text{Corr}(X_t, X_{t-k})$ την αυτοσυσχέτιση k -τάξης, αυτή ισούται με ρ^k , $k \geq 0$.

Επίσης, η ακολουθία X_t , $t \geq 1$, είναι μία στάσιμη Μαρκοβιανή αλυσίδα με πιθανότητες μετάβασης $p_{kl} = P(X_t = k | X_{t-1} = l)$ οι οποίες, για $k, l \in \{0, 1, \dots, n\}$, δίνονται από τη σχέση:

$$p_{kl} = \sum_{m=\max(0, k+l-n)}^{\min(k, l)} \binom{l}{m} \binom{n-l}{k-m} \alpha^m (1-\alpha)^{l-m} \beta^{k-m} (1-\beta)^{n-l+m-k}, \quad (3)$$

Οι παράμετροι π, ρ του μοντέλου BAR(1) μπορούν να εκτιμηθούν με τη μέθοδο της μέγιστης πιθανοφάνειας (*maximum likelihood estimation*, MLE). Έστω ότι έχουμε στη διάθεσή μας ένα τμήμα (*segment*) X_1, \dots, X_T , με $T \in \mathbb{N}$ από μια στάσιμη διαδικασία BAR(1). Η συνάρτηση πιθανοφάνειας ισούται με

$$L(\pi, \rho) = \binom{n}{x_1} \pi^{x_1} (1-\pi)^{n-x_1} \prod_{t=2}^T p_{x_{t-1}x_t}, \quad (4)$$

όπου οι πιθανότητες μετάβασης $p_{x_{t-1}x_t}$ δίνονται από τη σχέση (3). Οι εκτιμητές μέγιστης πιθανοφάνειας (EMΠ) $\hat{\pi}_{ML}$, $\hat{\rho}_{ML}$ μπορούν να βρεθούν με αριθμητική μεγιστοποίηση του λογαρίθμου της συνάρτησης της πιθανοφάνειας $l(\pi, \rho) = \log L(\pi, \rho)$ ενώ τα αντίστοιχα τυπικά σφάλματα υπολογίζονται μέσω της παρατηρούμενης πληροφορίας του Fisher (Weiß and Kim (2013)).

2.2 Το μοντέλο BBAR(1)

Οι Weiß and Kim (2014) πρότειναν μια γενίκευση του τελεστή διωνυμικής λέπτυνσης, χρησιμοποιώντας τη Βήτα-διωνυμική κατανομή. Ειδικότερα, έστω τ.μ. $\alpha_\phi \sim \text{Beta}(\alpha(1-\phi)/\phi, (1-\alpha)(1-\phi)/\phi)$, ανεξάρτητη της X , με $\alpha, \phi \in (0, 1)$. Τότε, η τ.μ. $\alpha_\phi \circ X$ προκύπτει από τη X μέσω Βήτα-διωνυμικής λέπτυνσης όπου « \circ » είναι ο συνήθης τελεστής διωνυμικής λέπτυνσης και ο οποίος εφαρμόζεται ανεξάρτητα στις α_ϕ και X . Άρα, για δεδομένη τιμή της τ.μ. X , η τ.μ. $\alpha_\phi \circ X$ ακολουθεί τη Βήτα-διωνυμική κατανομή. Για $\phi \rightarrow 0$, ο τελεστής Βήτα-διωνυμικής λέπτυνσης προσεγγίζει τον συνήθη τελεστή διωνυμικής λέπτυνσης. Παρακάτω δίνονται συνοπτικά ο ορισμός και οι ιδιότητες του μοντέλου BBAR(1) (δείτε επίσης Weiß and Kim (2014)):

Ορισμός 2.2: Έστω $n \in \mathbb{N}$. Οι παράμετροι του μοντέλου είναι π, ρ, ϕ με $\pi \in (0, 1)$, $\rho \in ([\max(-\pi/(1-\pi), -(1-\pi)/\pi)], 1)$, $\phi \in (0, 1)$ και ορίζεται από τη σχέση

$$X_t = \alpha_\phi \circ X_{t-1} + \beta_\phi \circ (n - X_{t-1}), \quad t \geq 1, \quad (5)$$

όπου τα α_ϕ, β_ϕ είναι τ.μ., ανεξάρτητες της X , που ακολουθούν Βήτα κατανομή με $\text{BETA}((1-\phi)\alpha/\phi, (1-\phi)(1-\alpha)/\phi)$ και $\text{BETA}((1-\phi)\beta/\phi, (1-\phi)(1-\beta)/\phi)$, αντίστοιχα. Τα α, β ορίζονται όπως στην περίπτωση του μοντέλου BAR(1). Όλες οι τ.μ. α_ϕ, β_ϕ καθώς και οι αντίστοιχες λεπτύνσεις εκτελούνται ανεξάρτητα των υπολοίπων ενώ όλες οι τ.μ. α_ϕ, β_ϕ καθώς και οι αντίστοιχες λεπτύνσεις τη χρονική στιγμή t είναι ανεξάρτητες των X_s , για $s < t$.

Οι πιθανότητες μετάβασης $p_{k|l}$, $k, l \in \{0, 1, \dots, n\}$ για το μοντέλο BBAR(1), ισούνται με

$$p_{k|l} = \sum_{m=\max(0, k+l-n)}^{\min(k, l)} \binom{l}{m} \binom{n-l}{k-m} \frac{B\left(m + \frac{1-\phi}{\phi} \alpha, l-m + \frac{1-\phi}{\phi} (1-\alpha)\right)}{B\left(\frac{1-\phi}{\phi} \alpha, \frac{1-\phi}{\phi} (1-\alpha)\right)} \times \frac{B\left(k-m + \frac{1-\phi}{\phi} \beta, n-l-k+m + \frac{1-\phi}{\phi} (1-\beta)\right)}{B\left(\frac{1-\phi}{\phi} \beta, \frac{1-\phi}{\phi} (1-\beta)\right)} \quad (6)$$

Αν η αρχική παρατήρηση X_0 προέρχεται από τη στάσιμη περιθώρια κατανομή, τότε όλη η διαδικασία X_t , $t \geq 1$, είναι στάσιμη. Για δεδομένες τιμές των π , ρ και ϕ , οι στάσιμες περιθώριες πιθανότητες $p_x = P(X_t = x)$, $x \in \{0, 1, \dots, n\}$, προσδιορίζονται αριθμητικά, χρησιμοποιώντας τις πιθανότητες μετάβασης. Πιο συγκεκριμένα, αν $\mathbf{p} = (p_0, \dots, p_n)$ είναι το διάνυσμα των περιθωρίων πιθανοτήτων, αυτό προκύπτει ως λύση του συστήματος εξισώσεων $\mathbf{Pp} = \mathbf{p}$, όπου \mathbf{P} είναι ο $(n+1) \times (n+1)$ πίνακας πιθανοτήτων μετάβασης.

Επίσης, η μέση τιμή και η διασπορά της στάσιμης κατανομής του μοντέλου BBAR(1) δίνονται από τις σχέσεις

$$\mu = E(X_t) = E(E(X_t | X_{t-1})) = \rho \cdot \mu + n\beta, \quad (7)$$

$$\sigma^2 = n\pi(1-\pi) \cdot \frac{(1-\phi)(1+\rho) + n\phi(1-2\pi(1-\pi)(1-\rho))}{(1-\phi)(1+\rho) + \phi(1-2\pi(1-\pi)(1-\rho))}. \quad (8)$$

Δεν είναι δύσκολο να διαπιστώσουμε πως και στο μοντέλο BBAR(1), είναι $\mu = n\pi$ ενώ από τη σχέση (8), παρατηρούμε ότι η ποσότητα $n\pi(1-\pi)$ είναι η διωνυμική διασπορά και αυτό που απομένει, καθορίζει την απόκλιση μεταξύ πραγματικής μεταβλητότητας και μεταβλητότητας στο διωνυμικό μοντέλο. Τέλος, η k -τάξης αυτοσυσχέτιση στο μοντέλο BBAR(1) είναι $\rho(k) = \rho^k$.

3. ΔΙΑΓΡΑΜΜΑΤΑ ΕΛΕΓΧΟΥ ΓΙΑ ΔΙΕΡΓΑΣΙΕΣ BAR(1) ΚΑΙ BBAR(1)

Σε αυτή την ενότητα θα αναπτύξουμε διαγράμματα ελέγχου τύπου EWMA για την παρακολούθηση διεργασιών που προέρχονται από τα υποδείγματα BAR(1) και BBAR(1). Στόχος είναι τα προτεινόμενα διαγράμματα να μπορούν να ανιχνεύσουν γρήγορα, και με ακρίβεια, μεταβολές στο μέσο επίπεδο της διεργασίας, δηλαδή στη μέση τιμή $\mu \equiv \mu_X = E(X_t) = n\pi$.

Όταν η διεργασία βρίσκεται εντός ελέγχου (*in-control process, IC*) η μέση τιμή ισούται με $\mu_X = \mu_{0,X}$, ενώ όταν η διεργασία είναι εκτός ελέγχου (*out-of-control process, OoC*) ισούται με $\mu_X = \mu_{1,X}$. Όμοια, οι τιμές των παραμέτρων του μοντέλου

BAR(1) στην περίπτωση διεργασιών IC και OC, είναι αντίστοιχα ίσες με π_0, ρ_0 και π_1, ρ_1 . Ομοια, για το μοντέλο BBAR(1), έχουμε π_0, ρ_0, ϕ_0 (IC) ή π_1, ρ_1, ϕ_1 (OC).

Στην πράξη, είναι πολύ σημαντικό να ανιχνεύσουμε μια μεταβολή από $\mu_{0,X}$ σε $\mu_{1,X} > \mu_{0,X}$. Τέτοιου είδους μεταβολές συνδέονται με τη χειροτέρευση της διεργασίας αφού π.χ. αυξάνεται ο μέσος αριθμός ελαττωματικών αντικειμένων που αυτή παράγει. Αξίζει πάντως να αναφερθεί ότι μια μεταβολή από $\mu_{0,X}$ σε $\mu_{1,X} < \mu_{0,X}$ συνδέεται με τη βελτίωση της διεργασίας, αφού π.χ. μειώνεται ο μέσος αριθμός ελαττωματικών αντικειμένων που αυτή παράγει. Στην παρούσα εργασία, θα μας απασχολήσει μόνο η περίπτωση αυξήσεων στο μέσο επίπεδο της διεργασίας.

Τα διαγράμματα EWMA προτάθηκαν από τον Roberts (1959) και η α.σ.σ. είναι

$$Q_t = \lambda X_t + (1 - \lambda)Q_{t-1}, \quad Q_0 = q_0, \quad t=1, 2, \dots, \quad (9)$$

όπου $\lambda \in (0, 1]$ είναι η παράμετρος εξομάλυνσης. Για μικρές τιμές του λ , δίδεται μικρότερη βαρύτητα στην πιο πρόσφατη παρατήρηση X_t και μεγαλύτερη σε όλες τις διαθέσιμες παρατηρήσεις από την αρχή της εφαρμογής του διαγράμματος EWMA. Άρα, το διάγραμμα διαθέτει «μνήμη» (*memory-type chart*) και είναι ικανότερο ενός διαγράμματος Shewhart στην ανίχνευση μικρών ή/και μεσαίων μετατοπίσεων στο μέσο της διεργασίας.

Προτού προχωρήσουμε στην παρουσίαση των διαγραμμάτων τύπου EWMA για διεργασίες BAR(1) και BBAR(1), πρέπει να αναφέρουμε ότι ένα ιδιαίτερο χαρακτηριστικό αυτών των διεργασιών είναι ότι οι τιμές $X_t, t \geq 1$, είναι ακέραιες. Άμεσα έπεται ότι με απευθείας εφαρμογή της (9) σε δεδομένα από διεργασίες BAR(1) και BBAR(1), οι τιμές Q_t δεν παραμένουν ακέραιες και σε αυτή την περίπτωση η απόδοση του διαγράμματος EWMA μπορεί να υπολογιστεί προσεγγιστικά. Ο Weiß (2011) πρότεινε και μελέτησε άνω μονόπλευρα διαγράμματα ελέγχου τύπου EWMA για την παρακολούθηση διεργασιών PINAR(1). Προκειμένου να παραμένουν ακέραιες οι τιμές της απεικονιζόμενης στατιστικής συνάρτησης (α.σ.σ.), χρησιμοποίησε ως α.σ.σ. την

$$Q_t = \text{round}(\lambda X_t + (1 - \lambda)Q_{t-1}), \quad Q_0 = q_0, \quad t=1, 2, \dots, \quad (10)$$

όπου $\text{round}(x) = z$ αν $x \in [z - 0.5, z + 0.5]$ (στρογγυλοποίηση στον πλησιέστερο ακέραιο) και $q_0 \in \mathbb{Z}$. Θα αναφερόμαστε στο συγκεκριμένο διάγραμμα ως το 1-EWMA, στο οποίο απεικονίζονται μόνο ακέραιες τιμές και δίνει ένδειξη OC διεργασίας όταν για πρώτη φορά $Q_t > UCL_1$, όπου UCL_1 κατάλληλο άνω όριο ελέγχου. Αξίζει να σημειωθεί πως για την περίπτωση που οι X_t είναι ανεξάρτητες και ισόνομες τ.μ. από διωνυμική κατανομή, το διάγραμμα 1-EWMA ταυτίζεται με το διάγραμμα EWMA του Gan (1993).

Για τον υπολογισμό της απόδοσης του διαγράμματος 1-EWMA, απαιτείται ο υπολογισμός της κατανομής του μήκους ροής του διαγράμματος, η οποία ορίζεται ως η κατανομή της τ.μ. $L = \min\{j : Q_j > UCL_1\}$. Για τον υπολογισμό της κατανομής της L , χρησιμοποιείται η μεθοδολογία των Μαρκοβιανών αλυσίδων (Brook and Evans,

1972). Λόγω του ότι η Q_t λαμβάνει ακέραιες τιμές, είναι δυνατός ο υπολογισμός της ακριβούς κατανομής της L . Για περισσότερες λεπτομέρειες, δείτε Weiß (2011).

Ως μέτρο απόδοσης χρησιμοποιείται η $E(L)$, γνωστή και ως μέσο μήκος ροής (*average run length, ARL*). Σύμφωνα με τον Weiß (2015), για διεργασίες IC, προτείνεται η χρήση του *ARL* μηδενικής κατάστασης (*zero-state ARL* ή *zsARL*), το οποίο εκφράζει τον αναμενόμενο αριθμό σημείων που πρέπει να αναπαρασταθούν στο διάγραμμα μέχρι αυτό να δώσει για 1^η φορά (εσφαλμένα) ένδειξη *OoC* διεργασίας. Για διεργασίες *OoC*, χρησιμοποιείται το *ARL* σταθερής κατάστασης (*steady-state ARL* ή *ssARL*), το οποίο εκφράζει τον αναμενόμενο αριθμό σημείων που πρέπει να αναπαρασταθούν στο διάγραμμα μέχρι αυτό να δώσει για 1^η φορά ένδειξη *OoC* διεργασίας, δεδομένου ότι η διεργασία ευρισκόταν αρχικά σε κατάσταση *IC*.

Για την εφαρμογή του διαγράμματος 1-EWMA απαιτείται ο προσδιορισμός των τιμών (λ, UCL_1). Αυτό μπορεί να γίνει ως εξής (δείτε επίσης Weiß (2011)):

B1. Επιλέγουμε τις τιμές των παραμέτρων π_0, ρ_0 (για διεργασία *BAR(1)*) ή π_0, ρ_0, ϕ_0 (για διεργασία *BBAR(1)*) καθώς και την επιθυμητή τιμή ARL_0 για το *zsARL*.

B2. Επιλέγουμε μία αρχική τιμή για το UCL_1 έτσι ώστε το αντίστοιχο άνω μονόπλευρο διάγραμμα *Shewhart* (περίπτωση $\lambda = 1$) να έχει $zsARL < ARL_0$.

B3. Μειώνουμε την τιμή λ έτσι ώστε το *zsARL* να είναι όσο το δυνατόν κοντά στην τιμή ARL_0 . Έτσι προκύπτει το ζεύγος (λ, UCL_1)

Η διαδικασία συνεχίζεται μειώνοντας περαιτέρω την αρχική τιμή του UCL_1 (έτσι ώστε να είναι πάντοτε $UCL_1 > \mu_{0,X}$) και αναζητώντας στη συνέχεια την τιμή του λ προκειμένου το *zsARL* να είναι όσο το δυνατόν κοντά στην τιμή ARL_0 . Τελικά προκύπτει ένα σύνολο δυνατών ζευγών (λ, UCL_1) για το διάγραμμα 1-EWMA, τα οποία δίνουν απόδοση το δυνατόν κοντά στην επιθυμητή. Μεταξύ αυτών, επιλέγεται αυτό το οποίο δίνει τη μικρότερη τιμή *ssARL* για δεδομένη μετατόπιση στις παραμέτρους της διεργασίας. Επίσης, προτείνεται $q_0 = 0$ ώστε να δοθεί στο διάγραμμα η δυνατότητα Άμεσης Αρχικής Αντίδρασης (*Fast Initial Response*, Lucas and Crosier (1982)).

Εκτός από το 1-EWMA, ο Weiß (2011) πρότείνει το διάγραμμα ελέγχου *s-EWMA*, $s = 2, 3, \dots$, για την παρακολούθηση διεργασιών *PINAR(1)*. Στο διάγραμμα *s-EWMA*, η α.σ.σ. είναι η

$$Q_t^{(s)} = s - \text{round}(\lambda X_t + (1 - \lambda)Q_{t-1}^{(s)}), Q_0^{(s)} = q_0/s, t = 1, 2, \dots, \quad (11)$$

όπου $s - \text{round}(x) = z$ αν $x \in [z - 0.5 \cdot s, z + 0.5 \cdot s]$ και $q_0 \in \square$. Συνήθεις τιμές για το s είναι οι 2 και 4, ενώ για $s = 1$ προκύπτει το διάγραμμα 1-EWMA. Πλέον η στρογγυλοποίηση των τιμών της (9) δε γίνεται σε ακέραιες τιμές αλλά σε πολλαπλάσια π.χ. του 0.5 (για $s = 2$) ή του 0.25 (για $s = 4$). Με τον τρόπο αυτό παραμένει δυνατός ο ακριβής υπολογισμός της κατανομής του μήκους ροής του διαγράμματος *s-EWMA* ενώ και η εξομάλυνση των αρχικών παρατηρήσεων είναι λιγότερο έντονη, σε σύγκριση με την περίπτωση της συνήθους στρογγυλοποίησης.

Το άνω μονόπλευρο διάγραμμα *s-EWMA* δίνει ένδειξη *OoC* διεργασίας όταν για 1^η φορά $Q_t^{(s)} > UCL_s = U/s$. Πλέον, για την εφαρμογή του διαγράμματος *s-EWMA*

απαιτείται ο προσδιορισμός των τιμών (s, λ, UCL_s). Αυτό μπορεί να γίνει ως εξής (δείτε επίσης Weiß (2011)):

B1. Επιλέγουμε τις τιμές των παραμέτρων π_0, ρ_0 (για διεργασία BAR(1)) ή π_0, ρ_0, ϕ_0 (για διεργασία BBAR(1)) καθώς και την επιθυμητή τιμή ARL_0 για το $zsARL$. Επίσης επιλέγουμε την τιμή $s > 1$.

B2. Επιλέγουμε ως αρχικές τιμές, τις τιμές (λ, UCL_1) που προέκυψαν για το 1-EWMA. Για το δεδομένο s , προσδιορίζεται μία αρχική τιμή για το UCL_s έτσι ώστε το αντίστοιχο άνω μονόπλευρο διάγραμμα s -EWMA να έχει $zsARL < ARL_0$.

B3. Προσδιορίζουμε την τιμή λ έτσι ώστε το $zsARL$ να είναι όσο το δυνατόν κοντά στην τιμή ARL_0 . Έτσι προκύπτει το ζεύγος (s, λ, UCL_1)

Η διαδικασία συνεχίζεται εφαρμόζοντας την παραπάνω διαδικασία και για τα υπόλοιπα ζεύγη τιμών (λ, UCL_1) που προέκυψαν κατά τη μελέτη του διαγράμματος 1-EWMA. Τελικά προκύπτει ένα σύνολο δυνατών ζευγών (s, λ, UCL_s) για το διάγραμμα s -EWMA, για διάφορες τιμές του s , τα οποία δίνουν απόδοση το δυνατόν κοντά στην επιθυμητή. Όπως και στην περίπτωση του διαγράμματος 1-EWMA, μεταξύ αυτών επιλέγεται εκείνο που δίνει τη μικρότερη τιμή $ssARL$ για δεδομένη μετατόπιση στις παραμέτρους της διεργασίας. Και εδώ, προτείνεται $q_0 = 0$.

4. ΑΡΙΘΜΗΤΙΚΗ ΜΕΛΕΤΗ

Στην παρούσα ενότητα παρουσιάζονται τα αποτελέσματα μιας εκτεταμένης αριθμητικής μελέτης σχετικά με την απόδοση των άνω μονόπλευρων διαγραμμάτων s -EWMA για διεργασίες BAR(1) και BBAR(1). Για τις εντός ελέγχου τιμές των παραμέτρων κάθε διεργασίας, υποθέτουμε $\mu_0 \in \{5, 10\}$, $\rho_0 \in \{0.25, 0.50, 0.75\}$ και $\phi_0 \in \{0, 0.25, 0.75\}$. Το $\phi_0 = 0$ συμβολίζει την περίπτωση μιας διεργασίας BAR(1). Επίσης, $n \in \{15, 30\}$. Επιπλέον, όταν η διεργασία είναι ΟοC, υποθέτουμε ότι $\rho_1 = \rho_0$ και $\phi_0 = \phi_1$ και οι ειδικές αιτίες μεταβλητότητας επιδρούν μόνο στο π_0 , το οποίο μεταβάλλεται από π_0 σε $\pi_1 = \delta \cdot \pi_0$, $\delta > 1$. Ως επιθυμητές τιμές για τη μετατόπιση δ , θεωρούμε τις $\delta = 1.2, 1.4$. Επίσης, ως ARL_0 επιλέχθηκε η τιμή 370.4, μια τιμή που εμφανίζεται παραδοσιακά στα διαγράμματα ελέγχου και προέρχεται από την τιμή ARL_0 του δίπλευρου διαγράμματος ελέγχου \bar{X} με όρια 3σ (δείτε π.χ. Montgomery 2009).

Στους Πίνακες 1 και 2 δίνονται οι τιμές των παραμέτρων (s, λ, UCL_s) του διαγράμματος s -EWMA, για διεργασίες BAR(1) και BBAR(1), αντίστοιχα. Οι εντός ελέγχου τιμές της διεργασίας δίνονται στις στήλες « ϕ_0 », « μ_0 », « n » και « ρ_0 ». Η ακριβής εντός ελέγχου απόδοση δίδεται στη στήλη « $zsARL$ » ενώ για τον υπολογισμό της εκτός ελέγχου απόδοσης χρησιμοποιήθηκε το $ssARL$ (στήλες « $ssARL(1.2)$ » για $\delta = 1.2$ και « $ssARL(1.4)$ » για $\delta = 1.4$). Ως προς την επιλογή των τιμών των παραμέτρων (s, λ, UCL_s), επιλέχθηκε το διάγραμμα με τιμή $zsARL$ όσο το δυνατόν πιο κοντά στο 370.4. Λόγω της διακριτής φύσης των διεργασιών BAR(1), BBAR(1), δεν είναι δυνατό να επιτευχθεί τιμή $zsARL = 370.4$. Παρατηρούμε ότι μεγάλες τιμές του s , π.χ. $s = 2$ ή 4 δίνουν τιμές $zsARL$ κοντά στην επιθυμητή τιμή ARL_0 . Επίσης, καθώς μειώνεται η τιμή UCL_s , μειώνεται και η τιμή λ , κάτι το οποίο είναι συνήθως

επιθυμητό ώστε να έχουν βελτιωμένη απόδοση τα διαγράμματα EWMA έναντι των διαγραμμάτων τύπου Shewhart. Αν και συνήθως, για την περίπτωση μετρήσεων από την Κανονική κατανομή, προτείνονται τιμές $0.05 \leq \lambda \leq 0.30$, κάτι τέτοιο δε συμβαίνει εδώ.

Πίνακας 1. Τιμές (s , λ , UCL) και απόδοση του s -EWMA, Διεργασία BAR(1)

ϕ_0	μ_0	n	ρ_0	s	λ	U	UCL_s	$zsARL$	$ssARL(1.2)$	$ssARL(1.4)$	
0	5	15	0.25	1	0.16	7	7	415.71	48.42	16.03	
			0.50	2	0.10	14	14/2	328.59	43.07	16.45	
			0.75	1	0.25	9	9	371.31	88.76	34.03	
	30	15	0.25	4	0.33	33	33/4	352.94	48.55	15.09	
			0.50	1	0.12	7	7	379.56	67.34	25.93	
			0.75	2	0.08	14	14/2	374.46	81.85	35.04	
		10	15	0.25	4	0.56	53	53/4	375.92	13.70	3.03
				0.50	4	0.51	54	54/4	352.91	19.05	4.63
				0.75	4	0.22	51	51/4	332.57	27.62	8.89
30	15	0.25	2	0.26	27	27/2	353.54	20.01	6.34		
		0.50	4	0.48	61	61/4	364.46	38.90	10.99		
		0.75	2	0.07	24	24/2	328.32	41.92	16.95		

Πίνακας 2. Τιμές (s , λ , UCL) και απόδοση του s -EWMA, Διεργασία BBAR(1)

ϕ_0	μ_0	n	ρ_0	s	λ	U	UCL_s	$zsARL$	$ssARL(1.2)$	$ssARL(1.4)$	
0.025	5	15	0.25	4	0.61	37	37/4	361.06	65.50	19.40	
			0.50	1	0.6	10	10	359.41	79.24	26.65	
			0.75	4	0.41	37	37/4	385.18	103.64	41.16	
	30	15	0.25	4	0.25	33	33/4	365.14	60.07	19.73	
			0.50	2	0.31	19	19/2	372.13	88.46	32.94	
			0.75	4	0.59	45	45/4	371.75	130.84	59.26	
		10	15	0.25	4	0.8	57	57/4	366.07	25.12	4.13
				0.50	1	0.91	15	15	281.97	29.43	6.23
				0.75	4	0.41	55	55/4	372.72	37.02	10.00
30	15	0.25	4	0.58	65	65/4	370.73	41.85	10.72		
		0.50	4	0.53	66	66/4	365.93	53.27	15.56		
		0.75	4	0.6	69	69/4	373.90	78.41	27.44		
0.05	5	15	0.25	4	0.53	37	37/4	398.44	71.91	21.34	
			0.50	2	0.7	21	21/2	366.92	95.37	34.03	
			0.75	4	0.62	41	41/4	371.70	119.21	50.74	
	30	15	0.25	4	0.54	45	45/4	372.73	108.10	41.04	
			0.50	4	0.62	49	49/4	379.84	132.89	57.28	
			0.75	4	0.58	49	49/4	369.66	150.91	73.96	
		10	12	0.25	2	0.5	27	27/2	307.67	14.98	3.59
				0.50	4	0.67	57	57/4	349.83	30.64	6.47
				0.75	4	0.59	57	57/4	373.97	47.69	12.85

30	0.25	4	0.74	73	73/4	360.95	61.52	16.94
	0.50	2	0.56	35	35/2	310.10	59.00	18.85
	0.75	4	0.6	73	73/4	374.48	93.23	34.50

Στον Πίνακα 3 δίνεται η αριθμητική σύγκριση διαγραμμάτων ελέγχου τύπου Shewhart και τύπου EWMA. Το άνω μονόπλευρο διάγραμμα ελέγχου Shewhart αποτελεί ειδική περίπτωση του διαγράμματος 1-EWMA για $\lambda=1$. Ως UCL_{SH} συμβολίζεται το άνω όριο ελέγχου του διαγράμματος Shewhart, το οποίο δίνει ένδειξη εκτός ελέγχου διεργασίας όταν για πρώτη φορά $X_i > UCL_{SH}$. Για περισσότερες λεπτομέρειες, δείτε Rakitzis et al. (2017). Τα διαγράμματα έχουν συγκρίσιμη εντός ελέγχου απόδοση, αν και λόγω της διακριτής φύσης των δεδομένων, δεν ήταν εφικτό να επιτύχουμε να έχουν όλα σχεδόν την ίδια τιμή $zsARL$. Είναι σαφής η υπεροχή του s -EWMA, έναντι του διαγράμματος Shewhart ενώ φαίνεται πως καθώς το s αυξάνει, βελτιώνεται ακόμη περισσότερο η απόδοσή του.

Πίνακας 3. Αριθμητική Σύγκριση Διαγραμμάτων Ελέγχου Shewhart και s -EWMA για Διεργασίες $BAR(1)$ και $BBAR(1)$

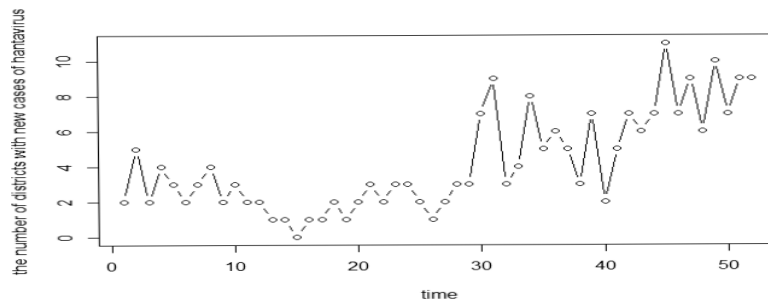
ϕ_0	μ_0	n	ρ_0	δ	Shewhart	1-EWMA	2-EWMA	4-EWMA
0	5	30	0.25	1.0	502.72	503.62	492.27	427.7
				1.1	223.27	204.62	135.39	128.25
				1.2	110.34	95.09	55.90	52.54
				1.3	59.68	49.60	30.95	26.13
				1.4	34.85	28.57	20.66	15.66
				1.5	21.74	17.91	15.46	10.64
				1.7	9.92	8.61	10.42	6.08
				2.0	4.13	4.09	7.20	3.78
UCL					11	10	14/2	31/4
λ					1	0.78	0.18	0.29
0.025	5	15	0.75	1.0	393.26	320.84	379.95	385.18
				1.1	208.46	164.18	190.31	185.98
				1.2	119.57	95.28	107.32	103.64
				1.3	73.69	58.44	65.46	61.43
				1.4	48.39	37.75	41.30	41.16
				1.5	33.60	26.68	28.96	28.20
				1.7	18.60	15.41	16.46	16.36
				2.0	9.88	8.69	9.02	9.15
UCL					10	9	18/2	36/4
λ					1	0.5	0.46	0.41
0.05	5	30	0.50	1.0	385.96	369.19	348.06	379.84
				1.1	247.82	217.53	202.43	217.99
				1.2	163.98	135.21	124.80	132.89

	1.3	111.71	87.98	80.83	85.24
	1.4	78.25	59.71	54.75	57.28
	1.5	56.30	42.11	38.64	40.16
	1.7	31.42	23.22	21.44	22.06
	2.0	15.49	11.75	11.01	11.23
<i>UCL</i>		14	12	24/2	48/4
λ		1	0.69	0.65	0.62

5. ΑΡΙΘΜΗΤΙΚΗ ΕΦΑΡΜΟΓΗ

Στην παρούσα ενότητα παρουσιάζεται μία πρακτική εφαρμογή των διαγραμμάτων ελέγχου *s*-EWMA σε επιδημιολογικά δεδομένα. Στη διάθεση μας έχουμε τον εβδομαδιαίο αριθμό X_t των περιφερειών στη Γερμανία, όπου καταγράφονται νέα περιστατικά λοίμωξης από τον ιό hantavirus το έτος 2011, για $t=1,2,\dots,T$ με $T=52$ εβδομάδες. Το πλήθος των περιφερειών είναι $n=38$. Για περισσότερες λεπτομέρειες, δείτε Ristic et al. (2015).

Εικόνα 1. Γραφική Παράσταση Δεδομένων



Στην Εικόνα 1, δίνεται η γραφική παράσταση των δεδομένων, απ' όπου διαπιστώνεται ότι οι τιμές στο δείγμα κυμαίνονται μεταξύ 0 και 11. Ο δειγματικός μέσος και η δειγματική διακύμανση ισούνται με $\bar{x} = 4.173$ και $s^2 = 7.793$, αντίστοιχα. Επίσης, $\rho(1) = 0.634$. Από την τιμή του $I_d = 2.098 > 1$ έχουμε μία ένδειξη υπερσκέδασης στα δεδομένα. Επιπλέον, είναι λογικό να θεωρηθεί ότι η πιθανότητα εμφάνισης νέας λοίμωξης δεν μπορεί να είναι ίδια σε όλες τις περιφέρειες αφού αυτές συνήθως διαφέρουν (με βάση π.χ. κοινωνικά και οικονομικά κριτήρια).

Στη συνέχεια, θα προσαρμόσουμε τα μοντέλα BAR(1), BBAR(1) στα διαθέσιμα δεδομένα, εκτιμώντας τις παραμέτρους αυτών με τη μέθοδο της μέγιστης πιθανοφάνειας. Αξίζει να σημειωθεί πως σύμφωνα με τους Weiß and Pollett (2014), το συγκεκριμένο σύνολο δεδομένων μπορεί να μοντελοποιηθεί ως ένα αυτοπαλίνδρομο μοντέλο 1^{ης} τάξης. Επιπλέον, για την επιλογή του μοντέλου με την καλύτερη προσαρμογή, χρησιμοποιούνται τα κριτήρια *AIC* και *BIC*, με

$$AIC = -2 \ell(\hat{\theta}) + \kappa \cdot 2, \quad BIC = -2 \ell(\hat{\theta}) + \kappa \cdot \log(T),$$

όπου $\ell(\theta)$ είναι ο λογάριθμος της πιθανοφάνειας και $\hat{\theta}$ είναι ο ΕΜΠ του $\theta = (\pi, \rho)$ (για το BAR(1)) ή του $\theta = (\pi, \rho, \phi)$ (για το BBAR(1)). Επίσης, το k είναι το πλήθος των υπό εκτίμηση παραμέτρων και T είναι το πλήθος των διαθέσιμων μετρήσεων. Τα αποτελέσματα από την προσαρμογή των δύο μοντέλων στα δεδομένα δίνονται στον Πίνακα 4. Στις παρενθέσεις δίνονται οι τιμές των τυπικών σφαλμάτων των εκτιμήσεων.

Αφού οι τιμές των κριτηρίων AIC , BIC είναι μικρότερες για το υπόδειγμα BBAR(1), συμπεραίνουμε ότι το μοντέλο BBAR(1) έχει καλύτερη προσαρμογή στα δεδομένα. Αξίζει να σημειωθεί πως θα μπορούσε να διεξαχθεί και στατιστικός έλεγχος της μορφής H_0 : BAR(1) έναντι H_1 :BBAR(1) (δείτε Ενότητα 4, Weiß and Kim (2014)). Για λόγους οικονομίας χώρου, στην παρούσα εργασία θα βασιστούμε μόνο στις τιμές των AIC, BIC

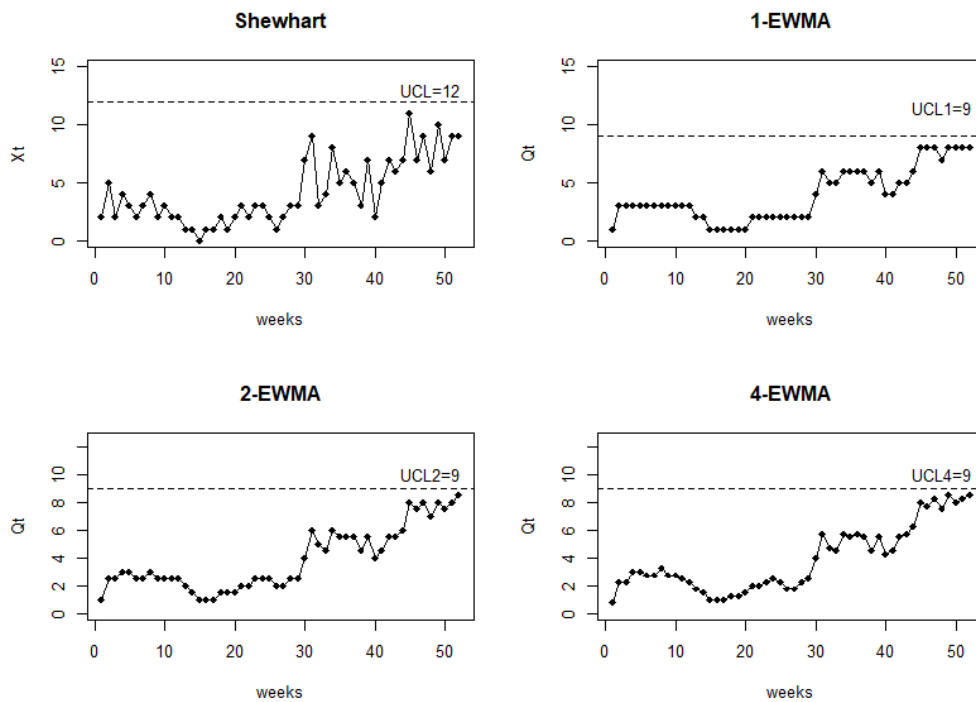
Στη συνέχεια, προχωρούμε στην ανάπτυξη άνω μονόπλευρων διαγραμμάτων Shewhart και s -EWMA για μια διεργασία BBAR(1). Οι εικόνες των διαγραμμάτων δίνονται στην Εικόνα 2.

Πίνακας 4. Προσαρμογή Μοντέλων BAR(1), BBAR(1) στα Επιδημιολογικά Δεδομένα

Μοντέλο	$\hat{\pi}$	$\hat{\rho}$	$\hat{\phi}$	AIC	BIC
BAR(1)	0.111 (0.013)	0.529 (0.070)		227.029	230.931
BBAR(1)	0.112 (0.179)	0.564 (0.074)	0.029 (0.016)	217.929	223.782

Για την ανάπτυξη των διαγραμμάτων, θεωρήθηκε ότι οι εκτιμήσεις των παραμέτρων, είναι οι πραγματικές τιμές των παραμέτρων της διεργασίας ενώ επιλέχθηκε $ARL_0 = 370.4$. Ακολουθώντας τα βήματα για τον προσδιορισμό των παραμέτρων (s, λ, UCL_s), για $s = 1, 2, 4$, καταλήξαμε στα παρακάτω διαγράμματα, με τιμή $zsARL$ κοντά στο 370.4: Διάγραμμα Shewhart με $UCL_{SH} = 12$ και $zsARL = 324.52$, 1-EWMA με $(\lambda, UCL_1) = (0.43, 9)$ και $zsARL = 352.85$, 2-EWMA με $(\lambda, UCL_2) = (0.38, 18/2)$ και $zsARL = 354.62$, 4-EWMA με $(\lambda, UCL_4) = (0.36, 36/4)$ και $zsARL = 370.60$. Τα διαγράμματα για τα διαθέσιμα δεδομένα δίνονται στην Εικόνα 3.

Εικόνα 2. Άνω μονόπλευρα διαγράμματα Shewhart και s -EWMA.



Παρατηρούμε ότι στις τελευταίες εβδομάδες του έτους υπάρχει μια έντονη διαταραχή (ξεκινάει λίγο μετά την 30^η εβδομάδα) και υπάρχει σαφής ένδειξη αυξητικής μετατόπισης (από την 41^η εβδομάδα και μετά), η οποία γίνεται αντιληπτή από όλα τα διαγράμματα. Τυπικά, σε κανένα διάγραμμα από τα επόμενα δεν έχουμε ένδειξη εκτός ελέγχου διεργασίας, αφού δεν υπάρχει σημείο πάνω από το *UCL*. Πιθανά, για το συγκεκριμένο σύνολο δεδομένων, αυτό να οφείλεται στην τιμή για το ARL_0 που επιλέχθηκε: π.χ. αν είχε επιλεγεί ως ARL_0 μια τιμή περίπου στο 200, ενδέχεται κάποιο από τα διαγράμματα να έδινε ένδειξη εκτός ελέγχου διεργασίας. Φυσικά ένα τέτοιο διάγραμμα θα είχε μεγαλύτερη πιθανότητα εσφαλμένου συναγερευμού. Συνήθως, η τελική επιλογή για την επιθυμητή τιμή του ARL_0 αφορά τον άμεσο χρήστη του διαγράμματος.

6. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα εργασία, μελετήθηκαν διαγράμματα ελέγχου τύπου EWMA με ακέραιες τιμές, για την παρακολούθηση διεργασιών BAR(1) και BBAR(1). Για τα συγκεκριμένα διαγράμματα είναι δυνατός ο ακριβής προσδιορισμός της κατανομής του μήκους ροής, με χρήση της μεθόδου των Μαρκοβιανών αλυσίδων. Δόθηκαν οι τιμές των παραμέτρων του διαγράμματος ώστε να έχει την επιθυμητή εντός ελέγχου απόδοση, για διάφορες εντός ελέγχου διεργασίες. Επιπλέον, υπολογίστηκε και η εκτός ελέγχου απόδοσή του και συγκρίθηκε με την απόδοση των αντίστοιχων άνω μονόπλευρων διαγραμμάτων Shewhart. Τέλος, παρουσιάστηκε και μια πρακτική εφαρμογή των διαγραμμάτων EWMA σε πραγματικά επιδημιολογικά δεδομένα. Για τον υπολογισμό της κατανομής του μήκους ροής των διαγραμμάτων EWMA

αναπτύχθηκαν προγράμματα στην R, τα οποία είναι διαθέσιμα, κατόπιν σχετικού αιτήματος.

ABSTRACT

In this work, we develop and study EWMA control charts that are suitable for monitoring processes with correlated counts, on a finite range. The processes are modeled according to a first order Binomial Autoregressive model (Binomial AR (1) or BAR (1)) as well as a first order Beta-Binomial Autoregressive model (Beta - Binomial AR (1) or BBAR (1)). The later model is useful in case where there is heterogeneity between the sample members of the sample or in the case where extra-binomial variation is present. Using the method of Markov chains, we provide numerical results on the statistical design and the performance of the proposed charts. A real-date example is also given.

ΑΝΑΦΟΡΕΣ

- Al-Osh MA and Alzaid AA. (1987). First-order integer-valued autoregressive (INAR (1)) process. *Journal of Time Series Analysis*, **8**(3), 261-275.
- Brook D and Evans DA. (1972) An approach to the probability distribution of CUSUM run length, *Biometrika*, **59**, 539-549.
- Chang TC and Gan FF. (2001) Cumulative sum charts for high yield processes, *Statistica Sinica*, **11**, 791-805.
- Gan FF. (1993) An optimal design of CUSUM control charts for binomial counts, *Journal of Applied Statistics*, **20**, 445-460.
- Lucas JM and Crosier RB. (1982) Fast initial response for CUSUM quality control schemes, *Technometrics*, **24**, 199-205
- McKenzie E. (1985). Some simple models for discrete variate time series, *Water Resources*, **21**, 645-650.
- Montgomery DC. (2009). Introduction to Statistical Quality Control, *6th edn. John Wiley & Sons, Inc.: New York, USA*.
- Psarakis S and Papaleonida GEA. (2007) SPC procedures for monitoring autocorrelated processes, *Quality Technology & Quantitative Management*, **4**, 501-540.
- Rakitzis AC., Weiss, CH. and Castagliola, P. (2017). Control charts for monitoring correlated counts with a finite range, *Applied Stochastic Models in Business and Industry*, **33**, 733-749.
- Ristic, M., Weiß CH. and Janjic A. (2015). A Binomial Integer-Valued ARCH Model, *The International Journal of Biostatistics*, **12**, 1-21
- Roberts S.W (1959). Control Chart Tests Based on Geometric Moving Averages, *Technometrics*, **1**, 239-250.
- Stutel F.W. and van Harn K. (1979). Discrete analogues of self-decomposability and stability, *Annals of Probability*, **7**, 893-899.
- Weiß C.H. (2009a). Monitoring correlated processes with binomial marginal, *Journal of Applied Statistics*, **36**, 399-414.

- Weiß C.H. (2011). Detecting mean increases in Poisson INAR(1) processes with EWMA control charts, *Journal of Applied Statistic*, **38**, 383-398.
- Weiß C.H. (2015) SPC methods for time-dependent processes of counts-A literature review, *Cogent Mathematics*, **2**, 1111116
- Weiß C.H. and Kim H.Y. (2013). Parameter estimation for binomial AR(1) models with applications in finance and industry. *Statistical Papers*, **54**, 563-590.
- Weiß C.H. and Kim H.Y. (2014). Diagnosing and modeling extra-binomial variation for time-dependent counts, *Applied Stochastic Models in Business and Industry*, **30**, 519-656.
- Weiß, C.H. and Pollett, P.K. (2014). Binomial autoregressive processes with density-dependent thinning. *Journal of Time Series Analysis*, **35**(2), 115-132.
- Wu Z, Jiao J and Liu Y. (2008). A binomial CUSUM chart for detecting large shifts in fraction nonconforming, *Journal of Applied Statistics*, **35**, 1267-1276.

ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΤΗΣ ΔΥΝΑΜΙΚΗΣ ΣΥΣΧΕΤΙΣΗΣ ΟΙΚΟΝΟΜΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΜΕ ΤΗ ΧΡΗΣΗ ΧΡΟΝΙΚΑ ΜΕΤΑΒΑΛΛΟΜΕΝΩΝ ΣΥΖΕΥΞΕΩΝ

Ε. Αποστολίδου, Φ. Κολυβά – Μαχαίρα

Τμήμα Μαθηματικών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
adevange@math.auth.gr, fkolyva@math.auth.gr

ΠΕΡΙΛΗΨΗ

Στη παρούσα εργασία παρουσιάζεται ένα χρονικά μεταβαλλόμενο μοντέλο συζεύξεων για τη μελέτη των δυναμικών δομών εξάρτησης του ανοίγματος του δεκαετούς κρατικού ομόλογου της Ελλάδας με των αντίστοιχων τιμών του ανοίγματος των ομολόγων των G20 χωρών, της Ισπανίας και της Πορτογαλίας, τη χρονική περίοδο 2007–2016. Το μοντέλο αυτό χρησιμοποιεί περιστραμμένες συζεύξεις για τη μοντελοποίηση της θετικής και αρνητικής εξάρτησης. Η σύζευξη που αντιπροσωπεύει τη δομή εξάρτησης των δεδομένων επιλέγεται με κριτήριο την εντροπία της. Ενώ, με τη μέθοδο του κινούμενου παραθύρου (rollingwindow) εκτιμάται η σύζευξη που περιγράφει καλύτερα τη κοινή κατανομή των δεδομένων σε κάθε χρονική περίοδο. Σύμφωνα με τα αποτελέσματα, το μοντέλο αυτό περιγράφει ικανοποιητικά τις αλλαγές στο τύπο και στη δομή της εξάρτησης των δεδομένων. Η εντροπία των συζεύξεων φαίνεται να είναι ένα σημαντικό κριτήριο για την επιλογή της καταλληλότερης σύζευξης.

Λέξεις Κλειδιά: εξάρτηση ουράς, εντροπία, περιστραμμένες συζεύξεις, δομή εξάρτησης

1. ΕΙΣΑΓΩΓΗ

Οι συζεύξεις αποτελούν ένα μαθηματικό εργαλείο που χρησιμοποιείται όλο και πιο πολύ για τη μοντελοποίηση της από κοινού κατανομής γεγονότων που συμβαίνουν ταυτόχρονα. Η ευελιξία που προσφέρουν οι συζεύξεις τις καθιστούν ιδανικές για τη μοντελοποίηση οικονομικών δεδομένων, επιτρέποντας τη μοντελοποίηση των περιθώριων κατανομών ξεχωριστά από την αντίστοιχη δομή εξάρτησης καθώς και τη μοντελοποίηση της εξάρτησης ουράς (taildependence), ενώ παράλληλα μένουν αμετάβλητες από τους μετασχηματισμούς των δεδομένων (Cherubini κ.ά. 2004, Penzer κ.ά. 2012, Wukai Lin 2014).

Τα τελευταία χρόνια η Ευρωζώνη βρίσκεται στον απόηχο της οικονομική κρίσης που ξεκίνησε από τους κρατικούς τομείς πίστωσης των οικονομιών της Ελλάδας, της Πορτογαλίας, της Ιρλανδίας και της Ιταλίας. Η κρίση αυτή ξεκίνησε το 2007/2008 από της ΗΠΑ και εξαπλώθηκε παγκοσμίως (Bekiros, 2013). Κατά συνέπεια η μεταφορά του κινδύνου (risk) στις αγορές αποτέλεσε αντικείμενο έρευνας πολλών

αναλυτών τα τελευταία χρόνια. Η ανάγκη να περιγραφεί η δυναμική εξάρτηση μεταξύ των αγορών, ιδιαίτερα σε ακραίες καταστάσεις φέρνει τα μοντέλα των συζεύξεων στο προσκήνιο όλο και περισσότερο.

Οι χρονικά μεταβαλλόμενες συζεύξεις που χρησιμοποιούνται για τη μοντελοποίηση της δομής εξάρτησης των κρατικών ομολόγων μέχρι σήμερα, δεν έχουν την δυνατότητα να περιγράψουν τις αλλαγές στο τύπο της εξάρτησης μεταξύ των ομολόγων. Γι' αυτό στην εργασία αυτή παρουσιάζεται ένα μοντέλο συζεύξεων που μοντελοποιεί με βέλτιστο τρόπο τη δομή εξάρτησης των κρατικών ομολόγων της Ελλάδας με τα κρατικά ομόλογα των χωρών της ομάδας G20, της Ισπανίας και της Πορτογαλίας. Μέσω του μοντέλου αυτού εξετάζεται και παρουσιάζεται η αλλαγή στη δομή και στο τύπο της εξάρτησης (πριν, κατά τη διάρκεια της κρίσης και μέχρι σήμερα), που εμφανίζουν χώρες της Ευρώπης, της Ασίας, της Αμερικής και της Ωκεανίας με την Ελλάδα. Οι περιστραμμένες συζεύξεις σε συνδυασμό με τη μέθοδο του κινούμενου παραθύρου που χρησιμοποιεί το μοντέλο αυτό συμβάλουν στην εύρεση και κατανόηση της κίνησης των κρατικών ομολόγων που εξετάζονται, τη τελευταία δεκαετία (2007-2016). Παράλληλα, εξετάζεται και αν η εντροπία των συζεύξεων μπορεί να αποτελέσει κριτήριο για την επιλογή της κατάλληλης σύζευξης αντί του κριτηρίου AIC που χρησιμοποιείται έως σήμερα για την επιλογή αυτή.

Η υπόλοιπη εργασία οργανώνεται ως εξής: η ενότητα 2 παρουσιάζει το θεωρητικό υπόβαθρο του μοντέλου. Η ενότητα 3 περιγράφει τη μεθοδολογία, η ενότητα 4 τα δεδομένα και τα αποτελέσματα και στην ενότητα 5 παρουσιάζονται τα συμπεράσματα.

2. ΣΥΝΝΑΡΤΗΣΕΙΣ ΣΥΖΕΥΞΕΩΝ

Σύμφωνα με τον Schweizer (1991) ο όρος σύζευξη (copula) χρησιμοποιήθηκε πρώτη φορά από τον Sklar (1959). Παρόλα αυτά η ιδέα τέτοιων δομών εξάρτησης είχε εμφανιστεί νωρίτερα σε διάφορα κείμενα κυρίως του Hoeffding (1940, 1941), ο οποίος καθιέρωσε τα καλύτερα δυνατά όρια για αυτές τις συναρτήσεις και μελέτησε μέτρα συσχέτισης τα οποία είναι αναλλοίωτα σε γνησίως μονότονους μετασχηματισμούς.

Σύζευξη ονομάζεται μια συνάρτηση που εκφράζει την αντιστοίχιση των τιμών της κοινής κατανομής σε κάθε διατεταγμένο ζεύγος τιμών των περιθώριων συναρτήσεων κατανομών (Schweizer, 1991).

Σύμφωνα με τον Sklar (1959), για μια συνάρτηση F m -μεταβλητών, η σύζευξη που σχετίζεται με την F είναι μια συνάρτηση κατανομής $C: [0,1]^m \rightarrow [0,1]$ που ικανοποιεί τη σχέση:

$$F(y_1, \dots, y_m) = C(F_1(y_1), \dots, F_m(y_m); \theta)$$

όπου θ είναι η παράμετρος της σύζευξης C και ονομάζεται εξαρτημένη παράμετρος, η οποία μετρά την εξάρτηση των περιθώριων συναρτήσεων.

Οι συζεύξεις που χρησιμοποιούνται πιο συχνά στη διεθνή βιβλιογραφία είναι η Gaussian, η t , η Clayton, η Gumbel, η Frank και η Joe. Η Gaussian και η τσζεύξεις χρησιμοποιούνται κυρίως για την περιγραφή της συμμετρικής θετικής ή αρνητικής

εξάρτησης και ανήκουν στην οικογένεια των ελλειπτικών συζεύξεων. Η Gaussian δεν παρουσιάζει εξάρτηση ουράς και η παρουσιάζει συμμετρική εξάρτηση ουράς.

Η διμεταβλητή γκαουσιανή σύζευξη προκύπτει εύκολα από τη διμεταβλητή κανονική κατανομή και η συνάρτηση κατανομής της δίνεται από τον ακόλουθο τύπο:

$$C_{Gaussian}(u, v; \vartheta) = \Phi_G(\Phi^{-1}(u), \Phi^{-1}(v); \vartheta) \\ = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi\sqrt{1-\vartheta^2}} \exp\left(-\frac{s^2 - 2\vartheta st + t^2}{2(1-\vartheta^2)}\right) ds dt$$

όπου Φ είναι η συνάρτηση αθροιστικής κατανομής της τυπικής διμεταβλητής κανονικής κατανομής και $\Phi_G(u, v)$ είναι η τυπική διμεταβλητή κανονική κατανομή με παράμετρο γραμμικής συσχέτισης ϑ που ορίζεται στο $[-1, 1]$.

Η σύζευξη t είναι η σύζευξη της πολυδιάστατης t -student κατανομής με ν βαθμούς ελευθερίας και συσχέτιση ρ και ορίζεται ως εξής:

$$C_t(u, v; \vartheta_1, \vartheta_2) = \int_{-\infty}^{t_{\vartheta_1}^{-1}(u)} \int_{-\infty}^{t_{\vartheta_2}^{-1}(v)} \frac{1}{2\pi(1-\vartheta_2^2)^{1/2}} \\ \times \left\{ 1 + \frac{(s^2 + 2\vartheta_2 st + t^2)^{-(\vartheta_1+2)/2}}{\vartheta_1(1-\vartheta_2^2)} \right\} ds dt$$

όπου $t_{\vartheta_1}^{-1}(u)$ την αντίστροφη αθροιστική συνάρτηση της τυπικής t -κατανομής με ϑ_1 βαθμούς ελευθερίας. Οι $(\vartheta_1, \vartheta_2)$ είναι οι δύο παράμετροι συσχέτισης, με τη παράμετρο ϑ_1 να ελέγχει την εξάρτηση στα άκρα της κατανομής και τη παράμετρο ϑ_2 να είναι η παράμετρος γραμμικής συσχέτισης. Καθώς η παράμετρος $\vartheta_1 \rightarrow \infty$, η t -σύζευξη προσεγγίζει τη γκαουσιανή με παράμετρο ϑ_2 .

Οι συζεύξεις Clayton, Gumbel και Joe περιγράφουν την ασύμμετρη θετική εξάρτηση ενώ η σύζευξη Frank επιτρέπει την αρνητική συμμετρία μεταξύ των περιθώριων κατανομών και ανήκουν στην οικογένεια των αρχιμήδειων συζεύξεων. Κατασκευάζονται με τη χρήση μίας γεννήτριας συνάρτησης σύζευξης φ και όχι όπως οι ελλειπτικές συζεύξεις που προκύπτουν εφαρμόζοντας το θεώρημα του Sklar στις αντίστοιχες πολυμεταβλητές κατανομές τους. Η Clayton εμφανίζει εξάρτηση κάτω ουράς, ενώ η Gumbel και η Joe συζεύξεις εμφανίζουν εξάρτηση άνω ουράς.

Αν η γεννήτρια συνάρτηση είναι $\varphi(t) = \frac{t^{-\theta}-1}{\theta}$ με $\theta \in [-1, \infty) \setminus \{0\}$ η σύζευξη που προκύπτει είναι η Clayton σύζευξη με συνάρτηση κατανομής:

$$C_{\theta}^{Clayton} = \max\left([u^{-\theta} + v^{-\theta} - 1]^{-1/\theta}, 0\right)$$

Για $\theta > 0$ η C_{θ} είναι γνήσια και έχει τη μορφή:

$$C_{\theta} = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$$

Η σύζευξη Gumbel προκύπτει από τη γεννήτρια συνάρτηση $\varphi(t) = -(\ln t)^{\theta}$ με $\theta \geq 1$ και συνάρτηση κατανομής:

$$C_{\theta}^{Gumbel}(u, v) = \exp\left(-[(-\ln u)^{\theta} + (-\ln v)^{\theta}]^{1/\theta}\right)$$

Η γεννήτρια συνάρτηση $\varphi(t) = \ln \frac{e^{-\theta t}-1}{e^{-\theta}-1}$ για $\theta \neq 1$ δίνει τη σύζευξη Frank:

$$C_{\theta}^{Frank}(u, v) = -\frac{1}{\theta} \ln \left(1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1} \right)$$

η σύζευξη αυτή επιτρέπει αρνητική συμμετρία μεταξύ των περιθώριων κατανομών και έχει ακτινική συμμετρία όπως οι ελλειπτικές συζεύξεις.

Σύμφωνα με τον Joe (1997) η γεννήτρια συνάρτηση $\varphi(t) = 1 - \ln[1 - (1 - t)^{\theta}]$ για $\theta > 1$ δίνει τη σύζευξη με συνάρτηση κατανομής:

$$C_{\theta}^{Joe} = 1 - ((1 - u)^{\theta} + (1 - v)^{\theta} - (1 - u)^{\theta}(1 - v)^{\theta})^{1/\theta}$$

Στη παρούσα εργασία, για την περιγραφή της άνω και κάτω εξάρτησης ουράς στη περίπτωση αρνητικής και θετικής ασυμμετρίας χρησιμοποιούνται οι περιστραμμένες συζεύξεις των Clayton, Gumbel και Joe κατά 90, 180 και 270 μοίρες. Οι περιστραμμένες συζεύξεις προκύπτουν ως εξής:

$$\begin{aligned} C_{90}(u_1, u_2) &= u_2 - C(1 - u_1, u_2), \\ C_{180}(u_1, u_2) &= u_1 + u_2 - 1 + C(1 - u_1, 1 - u_2) \\ C_{270}(u_1, u_2) &= u_1 - C(u_1, 1 - u_2). \end{aligned}$$

2.1 Χρήση συζεύξεων στη μοντελοποίηση της δομής εξάρτησης οικονομικών δεδομένων

Οι συζεύξεις χρησιμοποιούνται ευρέως σε οικονομικά δεδομένα κυρίως γιατί υπάρχουν εμπειρικές αποδείξεις ότι η εξάρτηση μεταξύ τέτοιων δεδομένων δεν είναι κανονική. Συγκεκριμένα, οι Erb, κ. ά. (1994), Longin και Solnik (2001), Ang και Chen (2002), Ang και Bekaert (2002) και Bae, κ. ά. (2003) απέδειξαν ότι δύο οικονομικές χρονοσειρές παρουσιάζουν ασυμμετρία και εξάρτηση ουράς απορρίπτοντας την υπόθεση ότι ακολουθούν πολυμεταβλητή κανονική κατανομή. Ταυτόχρονα έδειξαν ότι οι αποδόσεις των στοιχείων ενός χαρτοφυλακίου παρουσιάζουν μεγαλύτερη συσχέτιση σε περιπτώσεις ευμεταβλητότητας και επιδείνωσης της αγοράς και ότι υπάρχει μεγαλύτερη συσχέτιση μεταξύ μεγάλων απωλειών από ότι μεταξύ μεγάλων κερδών. Τέτοια φαινόμενα ασυμμετρίας δεν μπορούν να μοντελοποιηθούν σωστά από συμμετρικές κατανομές (Huang κ. ά., 2009). Έχοντας αυτό υπόψη και δεδομένων των μειονεκτημάτων που προκύπτουν από την μοντελοποίηση της εξάρτησης με γραμμική συσχέτιση σύμφωνα με τους Embrechts κ. ά. (2002), οι ερευνητές οδηγήθηκαν στη χρήση των συζεύξεων. Με τις συζεύξεις επιτρέπεται ο προσδιορισμός των περιθώριων κατανομών χωρίς τη χρήση της δομής εξάρτησης των μεταβλητών.

Ο Patton (2006a) επέκτεινε πρώτος τη θεωρία των συζεύξεων για τις δεσμευμένες κατανομές προτείνοντας ένα παραμετρικό μοντέλο που περιγράφει την εξέλιξη της σύζευξης στο χρόνο. Το θεώρημα του Sklar για δύο χρονοσειρές x_t, y_t επεκτείνεται:

$$H_t(x_t, y_t | \mathfrak{F}_{t-1}) = C_t(F_t(x_t | \mathfrak{F}_{t-1}), G_t(y_t | \mathfrak{F}_{t-1}), |\mathfrak{F}_{t-1}),$$

όπου C_t είναι σύζευξη σε όλες τις χρονικές στιγμές t και \mathfrak{F} μία σ -άλγεβρα που παράγεται από όλες τις προηγούμενες τιμές των μεταβλητών.

Ακολούθησαν πολλές μελέτες με την εφαρμογή και την εξέλιξη των χρονικά μεταβλητών συζεύξεων, όπως των Guegan και Zhang (2010), Cholletek κ. ά. (2009) και Hafner και Manner (2012). Ο συνδυασμός των μοντέλων πολλών μεταβλητών

GARCH και των δεσμευμένων συζεύξεων δημιούργησε μια καινούρια κατηγορία μοντέλων, τα copula GARCH μοντέλα (CGARCH).

3.ΜΕΘΟΔΟΛΟΓΙΑ

Η τιμή των κρατικών ομολόγων επηρεάζεται από πολλούς παράγοντες (όπως το πληθωρισμό του κράτους, τη προσφορά και τη ζήτηση του ομολόγου, τα επιτόκια κτλ.), καθιστώντας δυναμική τη δομή εξάρτησης μεταξύ κρατικών ομολόγων διαφορετικών χωρών. Η δυναμική εξάρτηση δεν μπορεί να περιγραφεί επαρκώς από μία μόνο σύζευξη. Συνεπώς, προτείνεται στην παρούσα εργασία ένα μοντέλο χρονικά μεταβαλλόμενων συζεύξεων που καθορίζεται από τα δεδομένα (data-driven time varying copula model) για να περιγράψει με ακρίβεια τη δομή εξάρτησης των κρατικών ομολόγων.

Τα ομόλογα εμφανίζουν αρνητική και θετική μη γραμμική εξάρτηση ταυτόχρονα με την εξάρτηση ακραίων τιμών. Για τη περιγραφή αυτών των φαινομένων χρησιμοποιούνται ο συντελεστής συσχέτισης Kendall που ορίζεται ως:

$$\tau_n = \frac{P_n - Q_n}{\binom{n}{2}} = \frac{4}{n(n-1)} P_n - 1,$$

όπου P_n και Q_n είναι ο αριθμός των εναρμονισμένων και μη εναρμονισμένων ζευγών, και οι συντελεστές εξάρτησης άνω και κάτω ουράς. Έστω ένα διάνυσμα τυχαίων συνεχών μεταβλητών (X, Y) με περιθώριες συναρτήσεις κατανομών F και G αντίστοιχα

$$\lambda_U = \lim_{u \uparrow 1} P[Y > G^{-1}(u) \mid X > F^{-1}(u)]$$

$$\lambda_L = \lim_{u \uparrow 0} P[Y \leq G^{-1}(u) \mid X \leq F^{-1}(u)]$$

όπου λ_U, λ_L , εκφράζουν την εξάρτηση άνω και κάτω ουράς αντίστοιχα. Για τη μοντελοποίηση της αρνητικής εξάρτησης των ομολόγων χρησιμοποιούνται οι περιστραμμένες συζεύξεις των Clayton, Gumbel και Joe.

Η εφαρμογή των συζεύξεων σε οικονομικά δεδομένα προϋποθέτει ότι από τα δεδομένα έχει απαλειφθεί η ετεροσκεδαστικότητα και η αυτοσυσχέτιση με τη χρήση ενός TGARCH μοντέλου. Το TGARCH(m,s) μοντέλο μιας χρονοσειράς r_t έχει τη μορφή:

$$r_t = \mu_t + \alpha_t, \quad \alpha_t = \sigma_t \varepsilon_t, \quad \sigma_t^2 = \alpha_0 + \sum_{i=1}^s (\alpha_i + \gamma_i N_{t-i}) \alpha_{t-i}^2 + \sum_{j=1}^m \beta_j \sigma_{t-j}^2,$$

όπου N_{t-i} είναι ένας δείκτης για αρνητικά α_{t-i} :

$$N_{t-i} = \begin{cases} 1 & \text{αν } \alpha_{t-i} < 0, \\ 0 & \text{αν } \alpha_{t-i} \geq 0 \end{cases}$$

μ_t η δεσμευμένη μέση τιμή της χρονοσειράς και σ_t^2 η δεσμευμένη διακύμανση. ε_t είναι μια ακολουθία ανεξάρτητων τυχαίων μεταβλητών που ακολουθούν όλες την ίδια κατανομή (i.i.d), με μέση τιμή 0 και διασπορά 1. Συχνά θεωρείται ότι η ε_t ακολουθεί κανονική κατανομή ή κανονικοποιημένη Student-t κατανομή ή τη κατανομή γενικευμένων σφαλμάτων. Οι $\alpha_0 > 0$, $\alpha_i, \gamma_i, \beta_j \geq 0$ είναι μη αρνητικοί παράμετροι του μοντέλου. Η α_i και η β_j αναφέρονται ως ARCH και GARCH

παράμετροι, αντίστοιχα. Η μέση τιμή ακολουθεί ένα ARMA(0,0) μοντέλο, δηλαδή θεωρείται μη δεσμευμένη.

Τα i.i.d κατάλοιπα που προκύπτουν μετασχηματίζονται σε μεταβλητές που ακολουθούν ομοιόμορφη κατανομή. Όταν ο ερευνητής δεν θέλει να κάνει κάποια υπόθεση για την κατανομή των καταλοίπων ο μετασχηματισμός γίνεται μέσω της εμπειρικής αθροιστικής συνάρτησης κατανομής:

$$F_i(x) = \frac{1}{T+1} \sum_{j=1}^T 1_{\{x_{i,j} \leq x\}},$$

όπου 1 δηλώνει μία συνάρτηση δείκτη που παίρνει τη τιμή 1 αν $x_{i,j} \leq x$ και διαφορετικά τη τιμή 0 (Fermanian και Scaillet, 2004).

Πρώτο βήμα για την επιλογή του μοντέλου, είναι ο προσδιορισμός του συντελεστή συσχέτισης Kendall. Αν οι δύο χρονοσειρές είναι αρνητικά συσχετισμένες τότε οι Clayton, Gumbel, Joe και οι αντίστοιχες συζεύξεις επιβίωσης (περιστραμμένες κατά 180 μοίρες), δεν λαμβάνονται υπόψη και η επιλογή γίνεται από τις υπόλοιπες συζεύξεις. Αντίστοιχα, αν οι δύο χρονοσειρές είναι θετικά συσχετισμένες δε λαμβάνονται υπόψη οι συζεύξεις Clayton, Gumbel, Joe, περιστραμμένες κατά 90 και 270 μοίρες. Οι παράμετροι όλων των διαθέσιμων συζεύξεων υπολογίζονται με τη μέθοδο μέγιστης πιθανοφάνειας και στη συνέχεια υπολογίζεται η εντροπία κάθε σύζευξης με τον εκτιμητή ChaoShen(2003) και η τιμή του AIC. Το μοντέλο με το μικρότερο AIC και τη μικρότερη εντροπία είναι το καταλληλότερο.

Παράλληλα, στη παρούσα εργασία χρησιμοποιείται η μέθοδος του κινούμενου παραθύρου (rollingwindow) προκειμένου να ερευνηθούν τα δυναμικά χαρακτηριστικά των κρατικών ομολόγων της Ελλάδας με τα αντίστοιχα ομόλογα των G20 χωρών, της Πορτογαλίας και της Ισπανίας σε τρία διαφορετικά χρονικά διαστήματα.

4. ΕΦΑΡΜΟΓΗ

Στηνενότητα αυτή προσομοιώνονται οι δομές εξάρτησης των μηνιαίων τιμών του ανοίγματος του δεκαετούς ομόλογου της Ελλάδας με των αντίστοιχων τιμών του ανοίγματος των ομολόγων των χωρών που αποτελούν την ομάδα των είκοσι G20, τη Πορτογαλία και την Ισπανία από το 2007 έως το 2016. Τα μέλη της G20 περιλαμβάνουν 19 επιμέρους χώρες - Αργεντινή, Αυστραλία, Βραζιλία, Γαλλία, Γερμανία, Ηνωμένες Πολιτείες Αμερικής, Ηνωμένο Βασίλειο, Ιαπωνία, Ιταλία, Ινδία, Ινδονησία, Καναδάς, Κίνα, Μεξικό, Νότια Αφρική, Νότια Κορέα, Ρωσία, Σαουδική Αραβία, Τουρκία και Ευρωπαϊκή Ένωση (ΕΕ). Οι χώρες που δεν εξέδωσαν όλα τα έτη (από το 2007 έως το 2016) δεκαετές ομόλογο, εξαιρούνται από τη παρούσα έρευνα. Οι χώρες αυτές είναι Τουρκία, Αργεντινή, Βραζιλία και Σαουδική Αραβία.

Οι τιμές του ανοίγματος των ομολόγων (bondspreads) προκύπτουν αν αφαιρεθεί από τη κάθε μηνιαία τιμή των ομολόγων η αντίστοιχη μηνιαία τιμή μιας χώρας που θεωρείται σημείο αναφοράς (benchmark). Στη παρούσα θέση θεωρείται η Γερμανία

ως σημείο αναφοράς λόγω της σταθερής της οικονομίας. Άρα εξετάζεται η συσχέτιση των τιμών του ανοίγματος των ομολόγων των υπόλοιπων 14 χωρών με αυτές της Ελλάδας.

Οι χρονοσειρές με οικονομικά δεδομένα είναι γνωστό ότι έχουν ετεροσκεδαστικότητα, οπότε το πρώτο βήμα για την επεξεργασία των δεδομένων είναι ο έλεγχος για την ύπαρξη ετεροσκεδαστικότητας και έπειτα η αφαίρεση της με τη προσαρμογή ενός tGARCH(1,1) μοντέλου. Η ανάλυση των δεδομένων έγινε στη γλώσσα R 3.4.3.

4.1 Περιγραφικά στατιστικά

Τη μεγαλύτερη μέση τιμή του ανοίγματος των ομολόγων, όπως παρατηρείται στο παρακάτω πίνακα, έχει η Ελλάδα με διαφορά από τις υπόλοιπες χώρες. Το ίδιο ισχύει και για τη τυπική απόκλιση της τιμής των ομολόγων, η οποία είναι 7.39 για την Ελλάδα και ακολουθεί η Πορτογαλία με 3.12. Ο Καναδάς και η Γαλλία έχουν τη μικρότερη μέση τιμή, ενώ η Γαλλία έχει και τη μικρότερη τυπική απόκλιση από όλες τις χώρες. Αναφορικά με την ελάχιστη τιμή του ανοίγματος των κρατικών ομολόγων, η Ελλάδα, ο Καναδάς, η Ρωσία, τη Νότιος Αφρική, το Μεξικό, η Γαλλία, η Ιταλία, η Ισπανία και η Πορτογαλία την εμφανίζουν τη περίοδο του 2007, λίγο πριν τη παγκόσμια κρίση του 2008. Αντίθετα οι Ηνωμένες Πολιτείες, η Ινδονησία και το Ηνωμένο Βασίλειο έχουν το μικρότερο άνοιγμα του δεκαετούς τους ομολόγου το 1/2012 οι 2 πρώτες και τον 11/2011 η τρίτη χώρα. Τα ομόλογα της Αυστραλίας, της Κίνας και της Ινδίας εμφανίζουν το μικρότερο άνοιγμα τέλη του 2008 με αρχές του 2009 ενώ το ομόλογο της Ιαπωνίας έχει το ελάχιστο άνοιγμα το 2016. Από την άλλη πλευρά, η μέγιστη τιμή του ανοίγματος του ελληνικού ομολόγου είναι το 2012 και μόνο τα ομόλογα της Γαλλίας, της Ιταλίας, της Ισπανίας και της Πορτογαλίας εμφανίζουν την ίδια περίοδο τη μέγιστη τιμή τους. Το 2016 έχουν το μεγαλύτερο άνοιγμα τα ομόλογα των Ηνωμένων Πολιτειών, του Καναδά, της Ν. Αφρικής και του Μεξικού. Η Ινδία, η Ρωσία, το Ηνωμένο Βασίλειο και η Κίνα εμφανίζουν το 2014 με αρχές του 2015 μεγαλύτερη ύφεση, η Αυστραλία το 2009 και τέλος η Ινδονησία και η Ιαπωνία το 2008.

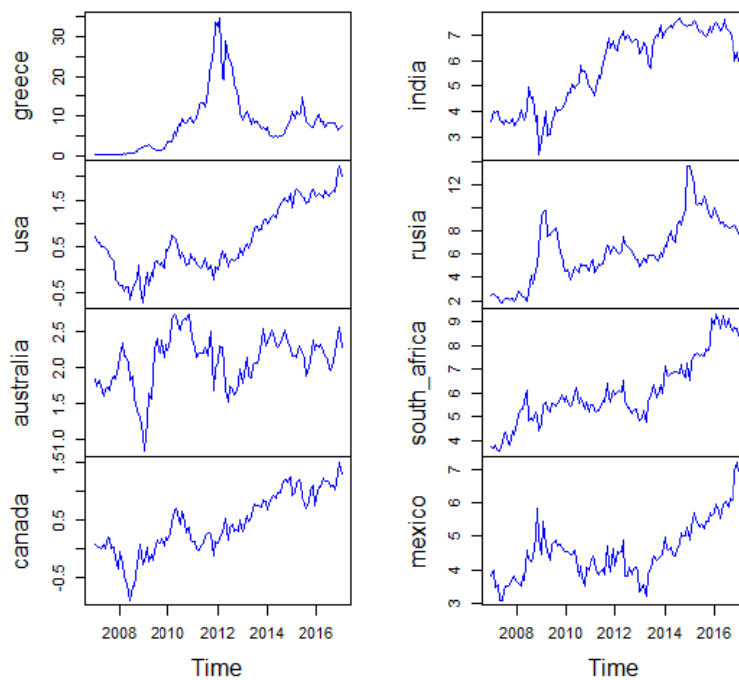
Πίνακας 1. Περιγραφικά στατιστικά

	Μέση τιμή	Τυπική απόκλιση	Ελάχιστη τιμή/ Ημερομηνία ελάχιστης τιμής	Μέγιστη τιμή/ Ημερομηνία μέγιστης τιμής
Ελλάδα	8.10	7.39	0.22 / Μαΐ-07	34.78 / Φεβ-12
ΗΠΑ	0.61	0.72	0.002 / Ιαν-12	2.24 / Δεκ-16
Αυστραλία	2.11	0.36	0.82 / Ιαν-09	2.75 / Απρ-10
Καναδάς	0.43	0.52	0.001 / Απρ-07	1.51 / Δεκ-16

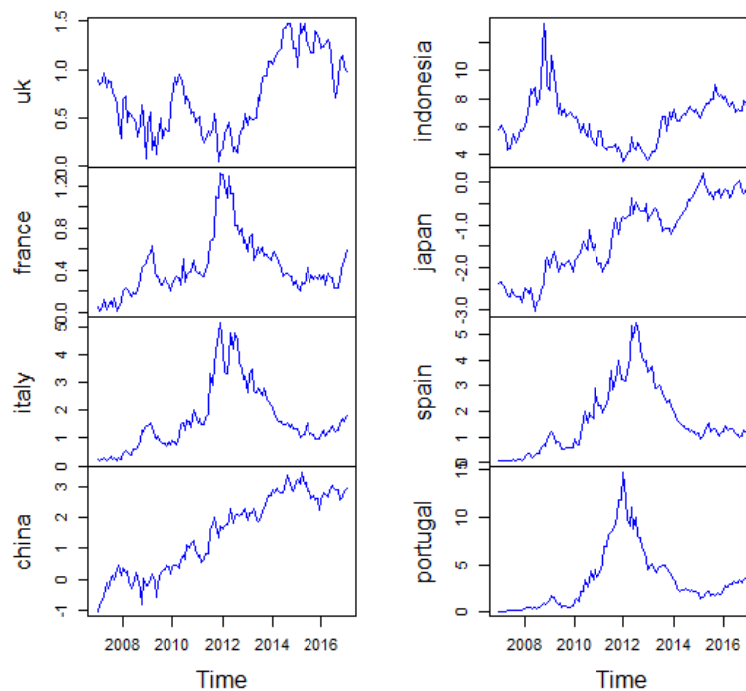
Ινδία	5.72	1.50	2.31 / Δεκ-08	7.68 / Αυγ-14
Ρωσία	6.35	2.69	1.78 / Ιουν-07	13.60 / Ιαν-15
Ν.Αφρική	6.08	1.41	3.53 / Απρ-07	9.31 / Φεβ-16
Μεξικό	4.56	0.84	3.10 / Ιουν-07	7.20 / Δεκ-16
Κίνα	1.48	1.24	-1.01 / Απρ-09	3.44 / Μαρ-15
Ινδονησία	6.44	1.75	3.51 / Ιαν-12	13.37 / Οκτ-08
Ιαπωνία	-1.25	0.88	-0.003 / Αυγ-16	-3.03 / Ιουν-08
Η.Βασίλειο	0.74	0.39	0.04 / Νοε-11	1.48 / Αυγ-14
Γαλλία	0.43	0.28	0.007 / Οκτ-07	1.33 / Δεκ-11
Ιταλία	1.69	1.17	0.16 / Οκτ-07	5.17 / Δεκ-11
Ισπανία	1.67	1.32	0.04 / Μαρ-07	5.48 / Ιουλ-12
Πορτογαλία	3.36	3.12	0.13 / Μαρ-07	14.63 / Ιαν-12

Οι συντελεστές της κυρτότητας και της λοξότητας δείχνουν ότι όλες οι τιμές είναι λοξές και παρουσιάζουν σημαντική πλατυκυρτότητα (σχετικά μικρή συγκέντρωση τιμών γύρω από τη μέση τιμή). Επιπλέον, ο έλεγχος κανονικότητας AdjustedJarque-Bera απορρίπτει σε επίπεδο σημαντικότητας $\alpha=0.05$ την αρχική υπόθεση ότι τα δεδομένα ακολουθούν κανονική κατανομή στις 5 από τις 16 χώρες. Τα αποτελέσματα που προκύπτουν από το Ljung-Box Q τάξης 4 δεν απορρίπτουν την μηδενική υπόθεση για όλες τις χρονοσειρές όμως σύμφωνα με το τεστ του πολλαπλασιαστή Lagrange υπάρχει ετεροσκεδαστικότητα σε όλες τις χρονοσειρές. Για την απαλοιφή της ετεροσκεδαστικότητας προσαρμόζουμε ένα μοντέλο tGARCH(1,1) σε όλες τις χρονοσειρές. Ακολουθούν οι γραφικές παραστάσεις του ανοίγματος των ομολόγων των 16 χωρών που μελετήθηκαν στη παρούσα εργασία, από το 2007 έως 01/2017.

Εικόνα 1.Γραφικές παραστάσεις του ανοίγματος των κρατικών ομολόγων της Ελλάδας, των ΗΠΑ, της Αυστραλίας, του Καναδά, της Ν.Αφρικής, της Ρωσίας, της Ινδίας και του Μεξικό



Εικόνα 2.Γραφικές παραστάσεις του ανοίγματος των κρατικών ομολόγων του Ηνωμένου Βασιλείου, της Γαλλίας, της Ιταλίας, της Κίνας, της Ινδονησίας, της Ιαπωνίας, της Ισπανίας και της Πορτογαλίας



4.2 Εφαρμογή του μοντέλου στα δεδομένα

Στον Πίνακα 2 παρουσιάζονται τα αποτελέσματα από τη προσαρμογή του μοντέλου στα δεδομένα της 1^{ης} χρονικής περιόδου 01/2007-03/2010, λίγο πριν την είσοδο της Ελλάδας στο Δ.Ν.Τ. Σύμφωνα με το πίνακα το AIC και η εντροπία συμφωνούν στην επιλογή του καταλληλότερου μοντέλου στη πλειοψηφία των περιπτώσεων. Ο συντελεστής συσχέτισης Kendall δείχνει ότι η Γαλλία, η Ιταλία, η Ισπανία και η Πορτογαλία έχουν την ισχυρότερη εξάρτηση με την Ελλάδα, από τις υπόλοιπες χώρες. Παράλληλα, χώρες με μέτρια εξάρτηση όπως η Ν.Αφρική, το Μεξικό, η Ιαπωνία και η Ινδονησία έχουν μέτρια προς ισχυρή εξάρτηση ουράς. Οι ΗΠΑ και το Η.Βασίλειο είναι οι μόνες χώρες με ασθενή αλλά αρνητική εξάρτηση με την Ελλάδα.

Στις δύο περιπτώσεις που το AIC επιλέγει διαφορετικό μοντέλο από την εντροπία, παρατηρείται ότι και τα δύο μοντέλα έχουν παρόμοια χαρακτηριστικά. Στη περίπτωση Ελλάδα – Η.Βασιλείου και οι δύο συζεύξεις περιγράφουν την αρνητική εξάρτηση μεταξύ των χωρών και την απουσία εξάρτησης ουράς. Στη δεύτερη περίπτωση, Ελλάδα- Ιταλία, η σύζευξη που προτείνεται από την εντροπία προσαρμόζεται καλύτερα στα δεδομένα σύμφωνα με τη γραφική τους παράσταση, καθώς παρουσιάζουν εξάρτηση κάτω ουράς.

Πίνακας 2. Προσαρμογή σύζευξης στα δεδομένα της 1^{ης} χρονικής περιόδου

	AIC - Εντροπία		Παράμετροι		Kendall 's τ	Άνω/Κάτω Εξάρτηση	
Ελλάδα - ΗΠΑ	Rotated Joe 90 degrees		-1.62		-0.17	0/0	
Ελλάδα - Αυστραλία	Joe		1.37		0.1	0.34/0	
Ελλάδα - Καναδάς	Joe		1.49		0.05	0.36/0	
Ελλάδα - Ινδία	Joe		1.41		0.11	0.37/0	
Ελλάδα - Ρωσία	Frank		7.58		0.55*	0/0	
Ελλάδα - Ν.Αφρική	Survival Joe		3.42		0.58*	0/0.78	
Ελλάδα - Μεξικό	Clayton		1.97		0.54*	0/0.7	
Ελλάδα - Κίνα	Gaussian		0.48		0.24*	0/0	
Ελλάδα - Ινδονησία	Survival Joe		2.63		0.43*	0/0.7	
Ελλάδα - Ιαπωνία	Survival Clayton		1.7		0.5*	0.66/0	
Ελλάδα - Ηνωμένο Βασίλειο	Rotated Joe 90 degrees	t	-1.86	-0.52, 2.38	-0.36*	0/0	0.04/0.04
Ελλάδα - Γαλλία	Frank		10.79		0.7*	0/0	
Ελλάδα - Ιταλία	Frank	Survival Joe	13.77	4.92	0.76*	0/0	0/0.85
Ελλάδα - Ισπανία	Frank		22.81		0.86*	0/0	
Ελλάδα - Πορτογαλία	Frank		12.49		0.76*	0/0	

Η δεύτερη χρονική περίοδος που μελετάται είναι από 04/2010 έως 02/2012 και περιλαμβάνει την υπογραφή του πρώτου και δεύτερου μνημονίου και τη μεγαλύτερη ύφεση της οικονομίας. Στη περίοδο αυτή όπως φαίνεται στο παρακάτω πίνακα η εντροπία με το AIC συμφωνούν στις 11 από τις 15 περιπτώσεις. Σε αντίθεση με τη πρώτη περίοδο η εξάρτηση της Ελλάδας με την Γαλλία, την Ιταλία, την Ισπανία και την Πορτογαλία δεν είναι το ίδιο ισχυρή και εμφανίζουν και εξάρτηση ουσίας. Ταυτόχρονα, οι ΗΠΑ, η Αυστραλία, ο Καναδάς, η Ινδονησία και το Η.Βασίλειο

έχουν μέτρια αρνητική εξάρτηση με τη Ελλάδα. Οι υπόλοιπες χώρες εμφανίζουν ασθενή εξάρτηση και ασθενή εξάρτηση ουράς. Στη περίπτωση Ελλάδας-Αυστραλίας και τα δύο μοντέλα δεν εμφανίζουν εξάρτηση ουράς και περιγράφουν την αρνητική εξάρτηση των δεδομένων. Παρόμοια, τα μοντέλα που προσαρμόζονται στα δεδομένα Ελλάδας-Ινδίας, έχουν θετική εξάρτηση και εξάρτηση άνω ουράς. Αντίθετα, η γραφική παράσταση των δεδομένων Ελλάδας-Γαλλίας δεν συμφωνεί με τις προτεινόμενες συζεύξεις καθώς τα δεδομένα φαίνεται, από τη γραφική παράσταση, να παρουσιάζουν εξάρτηση άνω και κάτω ουράς. Αναφορικά με τη περίπτωση της Ελλάδας-Ιταλίας τα δεδομένα δεν παρουσιάζουν εξάρτηση ουράς, όπως παρατηρείται στη γραφική παράσταση, οπότε η σύζευξη Frankταιριάζει περισσότερο στη δομή εξάρτησης των δεδομένων.

Πίνακας 3. Προσαρμογή σύζευξης στα δεδομένα της 2^{ης} χρονικής περιόδου

	AIC - Εντροπία		Παράμετροι		Kendall 's τ	Άνω/ΚάτωΕξάρτηση	
Ελλάδα - ΗΠΑ	Rotated Gumbel 90 degrees		-2.26		-0.55*	0/0	
Ελλάδα - Αυστραλία	Gaussian	Rotated Gumbel 270 degrees	-0.75	-2.05	-0.51*	0/0	
Ελλάδα - Καναδάς	Rotated Joe 90 degrees		-2.97		-0.44*	0/0	
Ελλάδα - Ινδία	Survival Clayton	Joe	1.69	2.54	0.48*	0.66/0	0.69/0
Ελλάδα - Ρωσία	Gaussian		0.79		0.56*	0/0	
Ελλάδα – Ν.Αφρική	Survival Clayton		0.56		0.16	0.29/0	
Ελλάδα - Μεξικό	Joe		1.35		0.17	0.33/0	
Ελλάδα - Κίνα	Gaussian		0.77		0.55*	0/0	
Ελλάδα - Ινδονησία	Frank		-10.08		-0.67*	0/0	
Ελλάδα - Ιαπωνία	Joe		2.38		0.34*	0.66/0	
Ελλάδα – Ηνωμένο Βασίλειο	Rotated Gumbel 90 degrees		-2.7		-0.62*	0/0	
Ελλάδα - Γαλλία	Gumbel	Survival Joe	3.26	3.85	0.69*	0.76/0	0/0.8

Ελλάδα - Ιταλία	Gumbel	Frank	2.93	10.13	0.72*	0.73/0	0/0
Ελλάδα - Ισπανία	Survival Gumbel		2.71		0.66*	0/0.71	
Ελλάδα - Πορτογαλία	Survival Gumbel		2.71		0.66*	0/0.71	

Στο Πίνακα 4 παρουσιάζονται τα αποτελέσματα από την εφαρμογή του μοντέλου στα δεδομένα της 3^{ης} χρονικής περιόδου που ξεκινά από 03/2012 και τελειώνει το 01/2017. Η περίοδος αυτή μελετά την τιμή του ανοίγματος των κρατικών ομολόγων το διάστημα του τρίτου μνημονίουμέχρι την εμφάνισηπρώτης ανάκαμψης. Ο πίνακας δείχνει ότι η εντροπία και το AICσυμφωνούν στην επιλογή της σύζευξης στις 11 από τις 15 περιπτώσεις. Στη περίοδο αυτή η Γαλλία, η Ιταλία, η Ισπανία και η Πορτογαλία έχουν ασθενή θετική εξάρτηση σύμφωνα με το συντελεστή συσχέτισης Kendall αλλά μέτρια εξάρτηση άνω ουράς. Παρόμοια και οι υπόλοιπες χώρες εμφανίζουν ασθενή εξάρτηση αρνητική ή θετική και ανεξαρτησία ουράς.Σε αυτή τη χρονική περίοδο, παρατηρείται πως στις περιπτώσεις που η εντροπία και το AICδεν συμφωνούν στο μοντέλο της σύζευξης που αντιπροσωπεύει καλύτερα τη δομή εξάρτησης των δεδομένων, τα μοντέλα έχουν το ίδιο είδος εξάρτησης (αρνητική, θετική) και την απουσία εξάρτησης ουράς.

Πίνακας 4. Προσαρμογή σύζευξης στα δεδομένα της 3^{ης} χρονικής περιόδου

	AIC - Εντροπία		Παράμετρο ι		Kendal I 's τ	Άνω/ΚάτωΕξάρτησ η
Ελλάδα- ΗΠΑ	Rotated Joe 270 degrees		-2.02		-0.28*	0/0
Ελλάδα- Αυστραλία	Gaussian	Rotated Gumbel 270 degrees	-0.74	- 1.97	-0.51*	0/0
Ελλάδα - Καναδάς	Rotated Clayton 90 degrees	Rotated Joe 270 degrees	-1.15	- 1.98	-0.33*	0/0
Ελλάδα- Ινδία	Rotated Clayton 270 degrees		-0.5		-0.16	0/0
Ελλάδα- Ρωσία	Frank	Survival Gumbel	-0.3	1	0.02	0/0
Ελλάδα – Ν.Αφρική	Rotated Clayton 90 degrees		-0.55		-0.2*	0/0
Ελλάδα- Μεξικό	Rotated Clayton 90 degrees		-0.48		-0.17	0/0

Ελλάδα - Κίνα	Rotated Clayton 90 degrees	Rotated Joe 270 degrees	-0.9	-1.77	-0.3*	0/0
Ελλάδα - Ινδονησία	Rotated Clayton 90 degrees		-0.65		-0.21*	0/0
Ελλάδα - Ιαπωνία	Clayton		0.3		0.1*	0/0.1
Ελλάδα – Ηνωμένο Βασίλειο	Rotated Joe 270 degrees		-2.31		-0.39*	0/0
Ελλάδα-Γαλλία	Joe		1.95		0.2	0.57/0
Ελλάδα-Ιταλία	Joe		2.12		0.25*	0.61/0
Ελλάδα-Ισπανία	Joe		2.08		0.26 *	0.6/0
Ελλάδα-Πορτογαλία	Joe		2.08		0.26*	0.6/0

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στόχος της παρούσας έρευνας ήταν η εφαρμογή ενός χρονικά μεταβαλλόμενου μοντέλου συζεύξεων που θα αναλύει τις δυναμικές δομές εξάρτησης του ανοίγματος των ομολόγων μεταξύ της Ελλάδας, των G20 χωρών, της Ισπανίας και της Πορτογαλίας. Το μοντέλο αυτό βασίστηκε στο μοντέλο των Liu και Fan (2017), με τη διαφορά ότι χρησιμοποιήθηκαν περισσότερα μοντέλα συζεύξεων και στην επιλογή του κατάλληλου για τα δεδομένα μοντέλου λήφθηκε υπόψη η εντροπία του ChaoShen της κάθε σύζευξης.

Τα εμπειρικά αποτελέσματα δείχνουν ότι η πλειοψηφία των συζεύξεων που προέκυψαν βάση της εντροπίας είναι ίδιες με τις συζεύξεις που προέκυψαν από το κριτήριο AIC. Παρατηρείται ότι ενώ τη πρώτη (01/2007 – 03/2010) και στη τρίτη (03/2012 – 01/2017) χρονική περίοδο τα ομόλογα των περισσότερων χωρών έχουν μικρή συσχέτιση με τα ομόλογα της Ελλάδας, στη δεύτερη χρονική περίοδο (04/2010 – 02/2012) έχουν μεγάλη συσχέτιση και μερικές από τις χώρες έχουν μεγάλη αρνητική συσχέτιση, όπως το Ηνωμένο Βασίλειο, η ΗΠΑ, η Ινδονησία και η Αυστραλία. Αντίθετα, τα ομόλογα της Ισπανίας, της Πορτογαλίας, της Ιταλίας και της Γαλλίας και στις τρεις χρονικές περιόδους είναι θετικά συσχετισμένα με τα ομόλογα της Ελλάδας δείχνοντας πως η οικονομική κρίση της Ελλάδας επηρέασε και τις άλλες Ευρωπαϊκές χώρες, όπως αποδείχθηκε και από τους Abadk. ά. (2010).

Σύμφωνα με τα μοντέλα των συζεύξεων στη δεύτερη χρονική περίοδο, όπου υπάρχουν και οι μεγαλύτερες συσχετίσεις, υπάρχει και εξάρτηση ουράς στη κοινή κατανομή των εξεταζόμενων ομολόγων. Υπάρχει δηλαδή, ταυτόχρονη ακραία κίνηση των ομολόγων της Ελλάδας με τα ομόλογα των περισσότερων χωρών που εξετάζονται. Στις άλλες δύο περιόδους η δομή της συσχέτισης στις περισσότερες περιπτώσεις εκφράζεται από μοντέλα που δεν εμφανίζουν εξάρτηση ουράς ή εμφανίζουν ασθενή εξάρτηση.

Εν κατακλείδι, η εξάρτηση του ανοίγματος των ομολόγων μεταξύ χωρών αλλάζει στο χρόνο για αυτό και πρέπει να περιγράφεται με ένα χρονικά μεταβαλλόμενο μοντέλο. Μάλιστα, σε αντίθεση με τα μοντέλα κοινών συζεύξεων (π.χ. Gaussian, t, Clayton και Gumbel), οι περιστραμμένες συζεύξεις φαίνεται πως περιγράφουν ικανοποιητικότερα τις δομές εξάρτησης ακραίων φαινομένων. Παράλληλα, η εντροπία των συζεύξεων φαίνεται να είναι ένα σημαντικό κριτήριο για την επιλογή της καταλληλότερης σύζευξης και δύναται να χρησιμοποιηθεί αντί του κριτηρίου AIC που χρησιμοποιείται έως σήμερα.

Ως αποτέλεσμα, το μοντέλο φαίνεται πως θα μπορούσε να αποτελέσει σημαντική προσθήκη στη εργαλειοθήκη αναλυτών κινδύνου τόσο στην έρευνα όσο και στη πράξη.

ABSTRACT

This paper introduces a time-varying copula model to study the dynamic dependence structure of economic data. Rotating copulas are used for modeling positive and negative dependence. Entropy is employed to select the most suitable copula, whereas the copula that describes best the joint data distribution during different time periods is estimated with the rolling window method. The model is applied to explore the structure of the dependence between the spread of the ten-year bond of Greece, Spain, Portugal and G20 countries, focusing on the 2007 - 2016 time period. Results indicate that the model effectively describes changes in the type and structure of the dependence. The entropy of the copula appears to be a meaningful criterion for selecting the most appropriate copula for describing the dependence structure.

ΑΝΑΦΟΡΕΣ

- Cherubini, U., Luciano, E., & Vecchiato, W. (2004). *Copula methods in finance*. John Wiley & Sons.
- Penzer, J., Schmid, F., & Schmidt, R. (2012). Measuring large comovements in financial markets. *Quantitative Finance*, 12(7), 1037-1049.
- Wu, C. C., & Lin, Z. Y. (2014). An economic evaluation of stock-bond return comovements with copula-based GARCH models. *Quantitative Finance*, 14(7), 1283-1296.
- Bekiros, S. D. (2014). Contagion, decoupling and the spillover effects of the US financial crisis: Evidence from the BRIC markets. *International Review of Financial Analysis*, 33, 58-69.

- Schweizer, B. (1991). Thirty years of copulas. In *Advances in probability distributions with given marginals* (pp. 13-50). Springer, Dordrecht.
- Hoeffding, W. (1994). Scale—invariant correlation theory. In *The collected works of Wassily Hoeffding* (pp. 57-107). Springer, New York, NY.
- Hoeffding, W. (1994). Scale—Invariant Correlation Measures for Discontinuous Distributions. In *The Collected Works of Wassily Hoeffding* (pp. 109-133). Springer, New York, NY.
- Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. Chapman and Hall/CRC.
- Erb, C. B., Harvey, C. R., & Viskanta, T. E. (1994). Forecasting international equity correlations. *Financial analysts journal*, 50(6), 32-45.
- Longin, F., & Solnik, B. (2001). Extreme correlation of international equity markets. *The journal of finance*, 56(2), 649-676.
- Ang, A., & Chen, J. (2002). Asymmetric correlations of equity portfolios. *Journal of financial Economics*, 63(3), 443-494.
- Ang, A., & Bekaert, G. (2002). International asset allocation with regime shifts. *The Review of Financial Studies*, 15(4), 1137-1187.
- Bae, K. H., Karolyi, G. A., & Stulz, R. M. (2003). A new approach to measuring financial contagion. *The Review of Financial Studies*, 16(3), 717-763.
- Huang, J. J., Lee, K. J., Liang, H., & Lin, W. F. (2009). Estimating value at risk of portfolio by conditional copula-GARCH method. *Insurance: Mathematics and economics*, 45(3), 315-324.
- Fermanian, J. D., & Scaillet, O. (2004). Some statistical pitfalls in copula modeling for financial applications.
- Chao, A., & Shen, T. J. (2003). Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and ecological statistics*, 10(4), 429-443.
- Liu, B. Y., Ji, Q., & Fan, Y. (2017). A new time-varying optimal copula model identifying the dependence across markets. *Quantitative Finance*, 17(3), 437-453.
- Abad, P., Chulla, H., Gómez-Puig, M. (2010). EMU and European government bond market integration. *Journal of Banking & Finance* 34(12), 2851–2860.



ΜΕΛΕΤΗ ΤΗΣ ΟΥΡΑΣ $Geo/Geo/1/N$ ΜΕ ΤΗ ΒΟΗΘΕΙΑ ΕΝΟΣ ΜΑΡΚΟΒΙΑΝΟΥ ΣΥΣΤΗΜΑΤΟΣ ΔΙΑΚΡΙΤΟΥ ΧΡΟΝΟΥ

Γεώργιος Βασιλειάδης

Τμήμα Πληροφορικής,

Πανεπιστήμιο Δυτικής Μακεδονίας

g.vasiliadis@kastoria.teiwm.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία παρουσιάζεται αρχικά το ομογενές Μαρκοβιανό σύστημα (ΟΜΣ) διακριτού χρόνου με χωρητικότητες στις καταστάσεις του, το οποίο αποτελεί μία γενίκευση του κλασικού ΟΜΣ, όπου η χωρητικότητα κάθε κατάστασης θεωρείται άπειρη. Για τη μελέτη της συμπεριφοράς του συστήματος στην εξέλιξη του χρόνου χρησιμοποιούνται αναδρομικές σχέσεις για τις παραγοντικές ροπές των μεγεθών των καταστάσεων. Το γενικό αυτό μοντέλο χρησιμοποιείται στη συνέχεια για να μελετήσουμε ένα σύστημα αναμονής διακριτού χρόνου, στο οποίο υπάρχει ένα σημείο εξυπηρέτησης, οι αφίξεις και οι εξυπηρετήσεις των πελατών πραγματοποιούνται σύμφωνα με τη γεωμετρική κατανομή και το πλήθος των πελατών σε κάθε χρονική στιγμή δεν μπορεί να ξεπεράσει ένα πεπερασμένο αριθμό N πελατών (ουρά $Geo / Geo / 1 / N$).

Λέξεις Κλειδιά: Μαρκοβιανά μοντέλα διακριτού χρόνου, ΟΜΣ με χωρητικότητες, Ουρές διακριτού χρόνου.

1. ΕΙΣΑΓΩΓΗ

Θεωρούμε ένα σύστημα το οποίο αποτελείται από N , $N \in \mathbb{N}$, μέλη τα οποία μπορούν να ταξινομηθούν σε k , $k \in \mathbb{N}$, καταστάσεις σύμφωνα με κάποια χαρακτηριστικά τους. Όλα τα μέλη του συστήματος έχουν τη δυνατότητα μετακίνησης από κατάσταση σε κατάσταση σύμφωνα με τη μαρκοβιανή ιδιότητα. Οι μετακινήσεις γίνονται σε χρόνο διακριτό και σε κάθε χρονική στιγμή τα μέλη του συστήματος βρίσκονται αποκλειστικά σε μία μόνο κατάσταση. Οι πιθανότητες μετάβασης από μία κατάσταση σε άλλη είναι σταθερές (δεν εξαρτώνται από το χρόνο). Επίσης, θεωρούμε ότι το πλήθος των μελών του συστήματος είναι σταθερό, δηλαδή δεν υπάρχει η δυνατότητα εισόδου νέων μελών ούτε επιτρέπεται η έξοδος στους ήδη υπάρχοντες.

Ένα σύστημα με τα παραπάνω χαρακτηριστικά ονομάζεται κλειστό ομογενές Μαρκοβιανό σύστημα (ΟΜΣ) διακριτού χρόνου (Bartholomew, 1982). Οι

συμβολισμοί που χρησιμοποιούμε για την μελέτη αυτών των συστημάτων είναι οι εξής:

- $t = 1, 2, \dots$, η παράμετρος που δηλώνει τα βήματα (χρονικές στιγμές),
- $S = \{1, 2, \dots, k\}$, ο χώρος καταστάσεων,
- p_{ij} , $i, j = 1, 2, \dots, k$, η πιθανότητα μετάβασης από την κατάσταση i στη j σ' ένα βήμα,
- $n_i(t)$, $i = 1, 2, \dots, k$, το πλήθος των μελών που βρίσκονται στην i κατάσταση τη χρονική στιγμή t ,
- N , το πλήθος των μελών του συστήματος.

Οι πιθανότητες μετάβασης από κατάσταση σε κατάσταση δίνονται ως στοιχεία ενός πίνακα διάστασης $k \times k$, $\mathbf{P} = (p_{ij})$, $i, j = 1, 2, \dots, k$, ο οποίος ονομάζεται πίνακας μετάβασης του συστήματος. Η πληθυσμιακή δομή του συστήματος σε κάθε χρονική στιγμή t δίνεται από το διάνυσμα

$$\mathbf{n}(t) = (n_1(t), n_2(t), \dots, n_k(t)),$$

το οποίο ονομάζεται διάνυσμα κατάστασης του συστήματος για τη χρονική στιγμή t ($t = 1, 2, \dots$). Για τις βασικές έννοιες στις μαρκοβιανές αλυσίδες και τη θεμελίωση του ΟΜΣ αναφέρουμε ενδεικτικά τα Bartholomew (1982), Isaacson and Madsen (1976), Vassiliou (1997), ενώ σχετικά με εφαρμογές των ΟΜΣ αναφέρουμε τις εργασίες McClean et al (1998), Tsaklidis and Soldatos (2003), Dimitriou and Tsantas (2010).

Γενίκευση του κλασικού ΟΜΣ διακριτού χρόνου αποτελεί ένα σύστημα στο οποίο κάθε κατάσταση i του χώρου καταστάσεων S παρουσιάζει πεπερασμένη χωρητικότητα $c_i \in \mathbb{N}$ (ΟΜΣ/ c). Το γεγονός ότι κάθε κατάσταση i του χώρου καταστάσεων S ενός ΟΜΣ διακριτού χρόνου παρουσιάζει πεπερασμένη χωρητικότητα $c_i \in \mathbb{N}$, έχει ως αποτέλεσμα το μέγεθος οποιασδήποτε κατάστασης i να μην μπορεί να υπερβεί την τιμή c_i , δηλαδή

$$n_i(t) \leq c_i, \quad i \in S,$$

για κάθε χρονική στιγμή t , $t = 1, 2, \dots$. Έτσι, αν το πλήθος των μελών του συστήματος που αποφασίσουν να μετακινηθούν προς μία κατάσταση i , $i \in S$, σε κάποια χρονική στιγμή t δεν υπερβαίνει τη χωρητικότητα c_i , τότε όλα τα μέλη εισέρχονται σ' αυτήν. Αν όμως σε κάποια χρονική στιγμή t , το πλήθος αυτό υπερβεί τη χωρητικότητα, τότε στην κατάσταση i εισέρχονται μόνο c_i από αυτά τα μέλη και τα μέλη αυτά αποτελούν το μέγεθος της κατάστασης για το χρονικό διάστημα $(t, t+1]$. Στην περίπτωση αυτή λέμε ότι η κατάσταση i παρουσιάζει υπερχειλίση. Θεωρούμε ότι τα μέλη του συστήματος που υπερχειλίζουν από μία κατάσταση σε μία χρονική στιγμή εισέρχονται σε ένα εικονικό αποθηκευτικό χώρο της κατάστασης και δεν μετακινούνται πλέον σύμφωνα με τον πίνακα μετάβασης \mathbf{P} . Συμβολίζουμε με $e_i(t)$, $i = 1, 2, \dots, k$, το πλήθος των μελών που βρίσκονται στον αποθηκευτικό χώρο

της i κατάστασης τη χρονική στιγμή t . Τα μέλη του συστήματος που βρίσκονται στον αποθηκευτικό χώρο μιας κατάστασης i , $i \in S$, έχουν τη δυνατότητα να επιστρέψουν στο σύστημα (να εισέλθουν στην κατάσταση i), όταν οι μετακινήσεις από και προς την κατάσταση i στις επόμενες χρονικές στιγμές δημιουργήσουν κενές θέσεις (το μέγεθος της κατάστασης δεν υπερβαίνει τη χωρητικότητα αυτής). Προφανώς, αν συμβολίσουμε με $m_i(t)$, $i=1,2,\dots,k$, το πλήθος των μελών που βρίσκονται στην κατάσταση i και στον αποθηκευτικό χώρο αυτής τη χρονική στιγμή t , τότε

$$n_i(t) = m_i(t) - e_i(t), i = 1, 2, \dots, k.$$

Αναλυτική περιγραφή και μελέτη συστημάτων με αυτό το χαρακτηριστικό παρουσιάζονται στις εργασίες Βασιλειάδης και Τσακλίδης (2010), Vasiliadis and Tsaklidis (2011).

Τα ΟΜΣ με χωρητικότητες στις καταστάσεις τους μπορούν να χρησιμοποιηθούν και στη μελέτη συστημάτων αναμονής. Στις εργασίες Βασιλειάδης (2014), Vasiliadis (2014) και Vasiliadis (2016) εξετάζεται η συμπεριφορά κλειστών συστημάτων αναμονής με πεπερασμένη πηγή με τη βοήθεια ΟΜΣ με χωρητικότητες στις καταστάσεις τους.

Στην παρούσα εργασία κατασκευάζουμε ένα κατάλληλο ΟΜΣ διακριτού χρόνου με χωρητικότητες στις καταστάσεις του, το οποίο μπορεί να χρησιμοποιηθεί για τη μοντελοποίηση ενός συστήματος αναμονής τύπου $Geo/Geo/1/N$. Εξετάζουμε τη συμπεριφορά του ΟΜΣ/ c στην εξέλιξη του χρόνου με τη βοήθεια επαναληπτικών σχέσεων για τις παραγοντικές ροπές των μεγεθών των καταστάσεών του. Οι σχέσεις αυτές χρησιμοποιούνται στη συνέχεια για τη μελέτη του αναμενόμενου μεγέθους της ουράς $Geo/Geo/1/N$ τόσο κατά την μεταβατική περίοδο (transient period) όσο και ασυμπτωτικά.

2. ΠΑΡΑΓΟΝΤΙΚΕΣ ΡΟΠΕΣ ΤΩΝ ΜΕΓΕΘΩΝ ΤΩΝ ΚΑΤΑΣΤΑΣΕΩΝ ΕΝΟΣ ΟΜΣ/ c

Το μέγεθος οποιασδήποτε κατάστασης i , $i \in S$, ενός ΟΜΣ/ c σε μία χρονική στιγμή t , δίνεται από την τ.μ. $n_i(t)$. Το μέγεθος αυτό μπορεί να προσδιοριστεί άμεσα αν γνωρίζουμε το πλήθος των μελών που βρίσκονται στην κατάσταση i , $i \in S$, και στον αποθηκευτικό χώρο αυτής, το οποίο δίνεται από την τ.μ. $m_i(t)$, $i \in S$. Έτσι, για τη μελέτη της συμπεριφοράς ενός ΟΜΣ/ c στην εξέλιξη του χρόνου μπορούμε να χρησιμοποιήσουμε τις τ.μ. $m_i(t)$, $i \in S$.

Θέλοντας να εξετάσουμε την εξέλιξη του διανύσματος κατάστασης ενός ΟΜΣ/ c χρησιμοποιούμε τις ροπές των τ.μ. $m_i(t)$, $i \in S$, με τη βοήθεια των οποίων μπορούμε να προσδιορίσουμε στη συνέχεια την κατανομή του διανύσματος κατάστασης του συστήματος. Για το σκοπό αυτό χρησιμοποιούμε την Πρόταση 1 (Βασιλειάδης, 2014) στην οποία δίνεται μία επαναληπτική σχέση για τον υπολογισμό των παραγοντικών ροπών των τ.μ. $m_i(t)$, $i \in S$. Για τη διατύπωση της πρότασης χρησιμοποιούμε ένα

γινόμενο διανυσμάτων (χρησιμοποιούμε το σύμβολο \times), το οποίο μοιάζει με το γινόμενο Kronecker. Συγκεκριμένα, για δύο διανύσματα $\mathbf{x}_1, \mathbf{x}_2$ το αποτέλεσμα του γινομένου $\mathbf{x}_1 \times \mathbf{x}_2$ είναι ένα διάνυσμα γραμμή, τα στοιχεία του οποίου προκύπτουν από το Kronecker γινόμενο $\mathbf{x}_1 \otimes \mathbf{x}_2$, με τη διαφορά ότι οι δυνάμεις αντικαθίστανται με παραγοντικά. Για παράδειγμα, αν $\mathbf{x}_1 = (a, b), \mathbf{x}_2 = (c, d)$, τότε έχουμε

$$\mathbf{x}_1 \times \mathbf{x}_2 = (ac, ad, bc, bd),$$

και

$$\begin{aligned}\mathbf{x}_1 \times \mathbf{x}_1 &= (a(a-1), ab, ba, b(b-1)), \\ \mathbf{x}_2 \times \mathbf{x}_2 &= (c(c-1), cd, dc, d(d-1)).\end{aligned}$$

Επίσης, κατά αναλογία της r -οστής δύναμης Kronecker $\mathbf{x}_1^{\otimes r}$ που ορίζεται από τις σχέσεις

$$\mathbf{x}_1^{\otimes 1} = \mathbf{x}_1, \mathbf{x}_1^{\otimes r} = \mathbf{x}_1 \otimes \mathbf{x}_1^{\otimes (r-1)}, r = 2, 3, \dots,$$

θα έχουμε

$$\mathbf{x}_1^{\times 1} = \mathbf{x}_1, \mathbf{x}_1^{\times r} = \mathbf{x}_1 \times \mathbf{x}_1^{\times (r-1)}, r = 2, 3, \dots$$

Πρόταση 1. Έστω ένα ΟΜΣ/ε με χωρητικότητες στις καταστάσεις του και με πίνακα μετάβασης $\mathbf{P} = (p_{ij})$, $i, j = 1, 2, \dots, k$. Τότε οι μεικτές παραγοντικές ροπές των τ.μ. $m_i(t+1)$, $i = 1, 2, \dots, k$, δίνονται από τη σχέση

$$\begin{aligned}E[\mathbf{m}(t+1)^{\times r}] &= E[\mathbf{n}(t)^{\times r}] \mathbf{P}^{\otimes r} \\ &+ E[\mathbf{n}(t)^{\times (r-1)} \times \mathbf{e}(t)] (\mathbf{P}^{\otimes (r-1)} \otimes \mathbf{I}) \\ &+ E[\mathbf{n}(t)^{\times (r-2)} \times \mathbf{e}(t) \times \mathbf{n}(t)] (\mathbf{P}^{\otimes (r-2)} \otimes \mathbf{I} \otimes \mathbf{P}) \\ &+ \dots \\ &+ E[\mathbf{e}(t)^{\times (r-1)} \times \mathbf{n}(t)] (\mathbf{I}^{\otimes (r-1)} \otimes \mathbf{P}) \\ &+ E[\mathbf{e}(t)^{\times (r-2)} \times \mathbf{n}(t) \times \mathbf{e}(t)] (\mathbf{I}^{\otimes (r-2)} \otimes \mathbf{P} \otimes \mathbf{I}) \\ &+ E[\mathbf{e}(t)^{\times r}] \mathbf{I}^{\otimes r},\end{aligned}\tag{1}$$

όπου $r \in \mathbb{N}$.

Σύμφωνα με την Πρόταση 1, για να υπολογίσουμε τις παραγοντικές ροπές των τ.μ. $m_i(t)$, $i = 1, 2, \dots, k$, σε μία χρονική στιγμή t , είναι απαραίτητο να γνωρίζουμε τις παραγοντικές ροπές και τις μεικτές παραγοντικές ροπές των τ.μ. $n_i(t-1)$, $e_i(t-1)$, $i = 1, 2, \dots, k$. Στο δεξί μέλος της σχέσης (1) υπάρχουν οι μικτές παραγοντικές ροπές των τ.μ. $n_i(t)$, $e_i(t)$, $i = 1, 2, \dots, k$, που προκύπτουν από τα διανύσματα $E[\mathbf{n}(t)^{\times r}]$, $E[\mathbf{e}(t)^{\times r}]$ και τα διανύσματα της μορφής

$$E[\mathbf{n}(t)^{\times w_1} \times \mathbf{e}(t) \times \mathbf{n}(t)^{\times w_2} \times \mathbf{e}(t) \times \dots \times \mathbf{e}(t) \times \mathbf{n}(t)^{\times w_{s+1}}],\tag{2}$$

όπου το διάνυσμα $\mathbf{e}(t)$ στο γινόμενο εμφανίζεται s φορές, $s = 1, 2, \dots, r-1$, με $w_1 + w_2 + \dots + w_{s+1} + s = r$, $w_j \in \mathbb{N}$, $j = 1, 2, \dots, s+1$. Όλα τα διανύσματα της μορφής (2) που έχουν το ίδιο πλήθος διανυσμάτων $\mathbf{e}(t)$, έχουν τα ίδια ακριβώς στοιχεία αλλά με διαφορετική σειρά. Για παράδειγμα τα διανύσματα

$$E[\mathbf{n}(t)^{\times(r-1)} \times \mathbf{e}(t)], E[\mathbf{n}(t)^{\times(r-2)} \times \mathbf{e}(t) \times \mathbf{n}(t)], \dots, E[\mathbf{e}(t) \times \mathbf{n}(t)^{\times(r-1)}],$$

που είναι όλα τα διανύσματα που εμφανίζουν στο γινόμενο το διάνυσμα $\mathbf{e}(t)$ μία φορά, περιέχουν όλα τα ίδια στοιχεία αλλά με διαφορετική σειρά.

Έτσι, για να μπορέσουμε να απλοποιήσουμε τον αλγόριθμο υπολογισμού των παραγοντικών ροπών των τ.μ. $m_i(t)$, $i = 1, 2, \dots, k$, που προκύπτουν από την Πρόταση 1, μπορούμε να χρησιμοποιήσουμε μία διαφορετική μορφή της σχέσης (1) (δίνεται στην Πρόταση 2) κάνοντας χρήση του μεταθετικού πίνακα (Βασιλείου και Τσακλίδης, 2001). Για κάθε $m \times n$ πίνακα \mathbf{A} ο μεταθετικός πίνακας \mathbf{K}_{mn} είναι ο $mn \times mn$ πίνακας για τον οποίο ισχύει

$$\mathbf{K}_{mn} \text{Vec}(\mathbf{A}) = \text{Vec}(\mathbf{A}^T),$$

όπου $\text{Vec}(\mathbf{A})$ το $mn \times 1$ διάνυσμα στήλη στο οποίο έχουν τοποθετηθεί όλες οι στήλες του \mathbf{A} η μία κάτω από την άλλη.

Ο μεταθετικός πίνακας \mathbf{K}_{mn} μπορεί να υπολογιστεί από τη σχέση

$$\mathbf{K}_{mn} = \sum_{i=1}^m \sum_{j=1}^n (\mathbf{e}_{m,i} \mathbf{e}_{n,j}^T) \otimes (\mathbf{e}_{n,j} \mathbf{e}_{m,i}^T),$$

όπου $\mathbf{e}_{m,i}$ το διάνυσμα στήλη διάστασης m το οποίο έχει στην i θέση 1 και όλα τα υπόλοιπα στοιχεία του 0. Κάποιες βασικές ιδιότητες του μεταθετικού πίνακα είναι οι εξής (Βασιλείου και Τσακλίδης, 2001):

- i) $\mathbf{K}_{mn}^T = \mathbf{K}_{nm}$,
- ii) $\mathbf{K}_{mn}^T \mathbf{K}_{mn} = \mathbf{K}_{mn} \mathbf{K}_{mn}^T$, ή $\mathbf{K}_{mn}^{-1} = \mathbf{K}_{mn}$,
- iii) $\mathbf{K}_{1n} = \mathbf{K}_{n1} = \mathbf{I}_n$, όπου \mathbf{I}_n ο $n \times n$ μοναδιαίος πίνακας.

Για κάθε $m \times n$ πίνακα \mathbf{A} και κάθε $r \times q$ πίνακα \mathbf{B} ισχύει

$$\mathbf{K}_{mn} (\mathbf{A} \otimes \mathbf{B}) \mathbf{K}_{nq} = \mathbf{B} \otimes \mathbf{A}.$$

Η ιδιότητα αυτή προφανώς μπορεί να εφαρμοστεί και στο γινόμενο διανυσμάτων που έχουμε ορίσει, αφού το γινόμενο αυτό προκύπτει από το γινόμενο Kronecker.

Πρόταση 2. Έστω ένα ΟΜΣ/σ με χωρητικότητες στις καταστάσεις του και με πίνακα μετάβασης $\mathbf{P} = (p_{ij})$, $i, j = 1, 2, \dots, k$. Τότε οι μεικτές παραγοντικές ροπές των τ.μ. $m_i(t+1)$, $i = 1, 2, \dots, k$, δίνονται από τη σχέση

$$\begin{aligned} E[\mathbf{m}(t+1)^{\times r}] &= E[\mathbf{n}(t)^{\times r}] \mathbf{P}^{\otimes r} \\ &+ E[\mathbf{n}(t)^{\times(r-1)} \times \mathbf{e}(t)] (\mathbf{P}^{\otimes(r-1)} \otimes \mathbf{I}) \mathbf{A}_1 \\ &+ E[\mathbf{n}(t)^{\times(r-2)} \times \mathbf{e}(t)^{\times 2}] (\mathbf{P}^{\otimes(r-2)} \otimes \mathbf{I}^{\otimes 2}) \mathbf{A}_2 \end{aligned}$$

$$\begin{aligned}
&+ \dots \dots \dots \\
&+ E[\mathbf{n}(t) \times \mathbf{e}(t)^{\times(r-1)}] (\mathbf{P} \otimes \mathbf{I}^{\otimes(r-1)}) \mathbf{A}_{r-1} \\
&+ E[\mathbf{e}(t)^{\times r}] \mathbf{I}^{\otimes r},
\end{aligned}$$

όπου $\mathbf{A}_i, i=1,2,\dots,r-1$, είναι πίνακες που προκύπτουν από ένα γινόμενο μεταθετικών πινάκων.

Απόδειξη. Για τα διανύσματα $\mathbf{n}(t)$ και $\mathbf{e}(t)$ θα έχουμε

$$(\mathbf{n}(t) \times \mathbf{e}(t)) \mathbf{K}_{kk} = \mathbf{e}(t) \times \mathbf{n}(t).$$

Άρα,

$$\begin{aligned}
\mathbf{n}(t)^{\times(r-2)} \times \mathbf{e}(t) \times \mathbf{n}(t) &= \mathbf{n}(t)^{\times(r-2)} \times ((\mathbf{n}(t) \times \mathbf{e}(t)) \mathbf{K}_{kk}) \\
&= \mathbf{n}(t)^{\times(r-2)} \times \mathbf{n}(t) \times \mathbf{e}(t) (\mathbf{I}_{k^{r-2}} \otimes \mathbf{K}_{kk}) \\
&= \mathbf{n}(t)^{\times(r-1)} \times \mathbf{e}(t) (\mathbf{I}_{k^{r-2}} \otimes \mathbf{K}_{kk}),
\end{aligned}$$

και

$$\begin{aligned}
\mathbf{P}^{\otimes(r-2)} \otimes \mathbf{I} \otimes \mathbf{P} &= \mathbf{P}^{\otimes(r-2)} \otimes (\mathbf{K}_{kk} (\mathbf{P} \otimes \mathbf{I}) \mathbf{K}_{kk}) \\
&= (\mathbf{I}_{k^{r-2}} \otimes \mathbf{K}_{kk}) (\mathbf{P}^{\otimes(r-2)} \otimes \mathbf{P} \otimes \mathbf{I}) (\mathbf{I}_{k^{r-2}} \otimes \mathbf{K}_{kk}) \\
&= (\mathbf{I}_{k^{r-2}} \otimes \mathbf{K}_{kk}) (\mathbf{P}^{\otimes(r-1)} \otimes \mathbf{I}) (\mathbf{I}_{k^{r-2}} \otimes \mathbf{K}_{kk}).
\end{aligned}$$

Έτσι για τον όρο $E[\mathbf{n}(t)^{\times(r-2)} \times \mathbf{e}(t) \times \mathbf{n}(t)] (\mathbf{P}^{\otimes(r-2)} \otimes \mathbf{I} \otimes \mathbf{P})$ που εμφανίζεται στο δεξί μέλος της (1) θα έχουμε:

$$\begin{aligned}
&E[\mathbf{n}(t)^{\times(r-2)} \times \mathbf{e}(t) \times \mathbf{n}(t)] (\mathbf{P}^{\otimes(r-2)} \otimes \mathbf{I} \otimes \mathbf{P}) = \\
&= E[\mathbf{n}(t)^{\times(r-1)} \times \mathbf{e}(t)] (\mathbf{I}_{k^{r-2}} \otimes \mathbf{K}_{kk}) (\mathbf{I}_{k^{r-2}} \otimes \mathbf{K}_{kk}) (\mathbf{P}^{\otimes(r-1)} \otimes \mathbf{I}) (\mathbf{I}_{k^{r-2}} \otimes \mathbf{K}_{kk}) \\
&= E[\mathbf{n}(t)^{\times(r-1)} \times \mathbf{e}(t)] (\mathbf{I}_{k^{r-2}} \otimes \mathbf{K}_{kk} \mathbf{K}_{kk}) (\mathbf{P}^{\otimes(r-1)} \otimes \mathbf{I}) (\mathbf{I}_{k^{r-2}} \otimes \mathbf{K}_{kk}) \\
&= E[\mathbf{n}(t)^{\times(r-1)} \times \mathbf{e}(t)] (\mathbf{I}_{k^{r-2}} \otimes \mathbf{I}_k^2) (\mathbf{P}^{\otimes(r-1)} \otimes \mathbf{I}) (\mathbf{I}_{k^{r-2}} \otimes \mathbf{K}_{kk}) \\
&= E[\mathbf{n}(t)^{\times(r-1)} \times \mathbf{e}(t)] (\mathbf{P}^{\otimes(r-1)} \otimes \mathbf{I}) (\mathbf{I}_{k^{r-2}} \otimes \mathbf{K}_{kk}).
\end{aligned}$$

Όμοια μπορούμε να δείξουμε ότι κάθε ένας όρος του αθροίσματος στο δεξί μέλος της (1), στον οποίο εμφανίζεται s φορές ($s \leq r-1$) το διάνυσμα $\mathbf{e}(t)$ και είναι της μορφής

$$\begin{aligned}
&E[\mathbf{n}(t)^{\times w_1} \times \mathbf{e}(t) \times \mathbf{n}(t)^{\times w_2} \times \mathbf{e}(t) \times \dots \times \mathbf{e}(t) \times \mathbf{n}(t)^{\times w_{s+1}}] \cdot \\
&(\mathbf{P}^{\otimes w_1} \otimes \mathbf{I} \otimes \mathbf{P}^{\otimes w_2} \otimes \mathbf{I} \otimes \dots \otimes \mathbf{I} \otimes \mathbf{P}^{\otimes w_{s+1}}),
\end{aligned}$$

με $w_1 + w_2 + \dots + w_{s+1} + s = r$, $w_1, w_2, \dots, w_{s+1}, s \in \square$, μπορεί να γραφεί ως εξής

$$E[\mathbf{n}(t)^{\times(w_1+w_2+\dots+w_{s+1})} \times \mathbf{e}(t)^{\times s}] (\mathbf{P}^{\otimes(w_1+w_2+\dots+w_{s+1})} \otimes \mathbf{I}^{\otimes s}) \mathbf{B},$$

όπου

$$\begin{aligned}
\mathbf{B} &= (\mathbf{I}_{k^{w_1}} \otimes \mathbf{K}_{k^{(w_2+w_3+\dots+w_{s+1})k}} \otimes \mathbf{I}_{k^{s-1}}) (\mathbf{I}_{k^{w_1+w_2+1}} \otimes \mathbf{K}_{k^{(w_3+w_4+\dots+w_{s+1})k}} \otimes \mathbf{I}_{k^{s-2}}) \dots \dots \\
&(\mathbf{I}_{k^{(w_1+w_2+\dots+w_{s-2}+s-3)}} \otimes \mathbf{K}_{k^{(w_{s-1}+w_s+w_{s+1})k}} \otimes \mathbf{I}_{k^2}) (\mathbf{I}_{k^{(w_1+w_2+\dots+w_{s-1}+s-2)}} \otimes \mathbf{K}_{k^{(w_s+w_{s+1})k}} \otimes \mathbf{I}_k)
\end{aligned}$$

$$(\mathbf{I}_k^{(w_1+w_2+\dots+w_s+s-1)} \otimes \mathbf{K}_{k^{w_s+1}k}).$$

Έτσι η σχέση (1) μπορεί να γραφεί ως εξής

$$\begin{aligned} E[\mathbf{m}(t+1)^{\times r}] &= E[\mathbf{n}(t)^{\times r}] \mathbf{P}^{\otimes r} \\ &+ E[\mathbf{n}(t)^{\times(r-1)} \times \mathbf{e}(t)] (\mathbf{P}^{\otimes(r-1)} \otimes \mathbf{I}) \mathbf{A}_1 \\ &+ E[\mathbf{n}(t)^{\times(r-2)} \times \mathbf{e}(t)^{\times 2}] (\mathbf{P}^{\otimes(r-2)} \otimes \mathbf{I}^{\otimes 2}) \mathbf{A}_2 \\ &+ \dots \\ &+ E[\mathbf{n}(t) \times \mathbf{e}(t)^{\times(r-1)}] (\mathbf{P} \otimes \mathbf{I}^{\otimes(r-1)}) \mathbf{A}_{r-1} \\ &+ E[\mathbf{e}(t)^{\times r}] \mathbf{I}^{\otimes r}, \end{aligned}$$

όπου $\mathbf{A}_i, i = 1, 2, \dots, r-1$, είναι πίνακες που προκύπτουν από ένα άθροισμα πινάκων της μορφής του πίνακα \mathbf{B} . □

Χρησιμοποιώντας τις παραγοντικές ροπές των τ.μ. $m_i(t)$, $i = 1, 2, \dots, k$, μπορούμε να προσδιορίσουμε τις κατανομές αυτών των τ.μ. για κάθε χρονική στιγμή t (Vasiliadis and Tsaklidis, 2011). Έτσι καταλήγουμε στο επόμενο πόρισμα, όπου δίνεται σχέση για την κατανομή των τ.μ. $m_i(t)$, $i = 1, 2, \dots, k$ και για την από κοινού κατανομή αυτών των τ.μ..

Πόρισμα 1. Έστω ένα ΟΜΣ/σ διακριτού χρόνου με χωρητικότητες στις καταστάσεις του και με πίνακα μετάβασης $\mathbf{P} = (p_{ij})$, $i, j = 1, 2, \dots, k$. Οι κατανομές των τ.μ. $m_i(t)$, $i = 1, 2, \dots, k$, για κάθε χρονική στιγμή t , δίνονται από τη σχέση

$$P[m_i(t) = n] = \frac{1}{n!} \sum_{j=0}^N \frac{(-1)^j}{j!} E[m_i^{(n+j)}(t)],$$

για $i = 1, 2, \dots, k$, $t = 1, 2, \dots$, και $n = 0, 1, \dots, N$.

Η από κοινού κατανομή των τ.μ. $m_i(t)$, $i = 1, 2, \dots, k$, για κάθε χρονική στιγμή t , δίνεται από τη σχέση

$$P[m_1(t) = r_1, m_2(t) = r_2, \dots, m_k(t) = r_k] = \frac{1}{\prod_{i=1}^k r_i!} E \left[\prod_{i=1}^k m_i^{(r_i)}(t) \right].$$

Γνωρίζοντας την κατανομή των τ.μ. $m_i(t)$, $i = 1, 2, \dots, k$, για μία χρονική στιγμή t , μπορούμε να προσδιορίσουμε και την κατανομή των τ.μ. $n_i(t)$ και $e_i(t)$, $i = 1, 2, \dots, k$, αφού η τιμή της τ.μ. μεταβλητής $m_i(t)$, καθορίζει άμεσα και τις τιμές των τ.μ. $n_i(t)$ και $e_i(t)$. Είναι

$$P[n_i(t) = n] = \begin{cases} P[m_i(t) = n], & n = 0, 1, \dots, c_i - 1, \\ \sum_{w=c_i}^N P[m_i(t) = w], & n = c_i \end{cases}$$

και

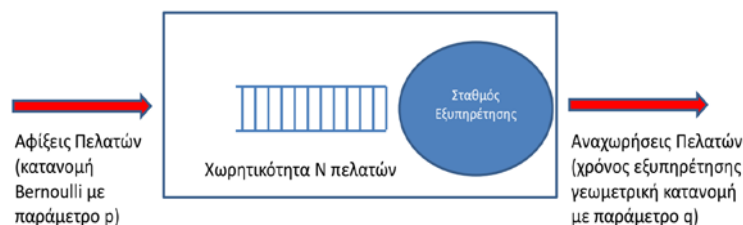
$$P[e_i(t) = n] = \begin{cases} \sum_{w=0}^{c_i} P[m_i(t) = w], & n = 0, \\ P[m_i(t) = n + c_i], & n = 1, 2, \dots, N - c_i. \end{cases}$$

Θέλοντας να υπολογίσουμε τις παραγοντικές ροπές των τ.μ. $m_i(t)$, $i = 1, 2, \dots, k$, σε μία χρονική στιγμή t χρησιμοποιώντας την Πρόταση 1, είναι απαραίτητο να γνωρίζουμε τις παραγοντικές ροπές και τις μεικτές παραγοντικές ροπές των τ.μ. $n_i(t-1)$, $e_i(t-1)$, $i = 1, 2, \dots, k$, οι οποίες προκύπτουν αν γνωρίζουμε την κατανομή του διανύσματος $\mathbf{m}(t-1)$ για τη χρονική στιγμή $t-1$. Επομένως, θεωρώντας ότι η κατανομή του αρχικού διανύσματος $\mathbf{m}(0)$ είναι γνωστή, μπορούμε να χρησιμοποιήσουμε επαναληπτικά την Πρόταση 1 και το Πόρισμα 1 έτσι ώστε να υπολογίσουμε τόσο τις παραγοντικές ροπές των τ.μ. $m_i(t)$, $i = 1, 2, \dots, k$, όσο και την κατανομή του διανύσματος $\mathbf{m}(t)$ για οποιαδήποτε χρονική στιγμή t .

3. ΟΥΡΑ *Geo/Geo/1/N*

Θεωρούμε ένα σύστημα αναμονής στο οποίο υπάρχει ένας σταθμός εξυπηρέτησης με ένα σημείο εξυπηρέτησης και το σύστημα έχει πεπερασμένη χωρητικότητα N πελατών. Οι αφίξεις των πελατών στο σταθμό εξυπηρέτησης του συστήματος καθώς επίσης και οι αναχωρήσεις απ' αυτόν πραγματοποιούνται σε χρόνο διακριτό. Υποθέτουμε ότι οι αφίξεις και οι αναχωρήσεις συμβαίνουν ταυτόχρονα στην αρχή κάθε χρονικής περιόδου t ($t = 1, 2, \dots$). Θεωρούμε επίσης ότι μια άφιξη πελάτη στο σύστημα πραγματοποιείται σύμφωνα με την κατανομή Bernoulli με παράμετρο p και ο χρόνος που απαιτείται για την εξυπηρέτηση ενός πελάτη ακολουθεί την γεωμετρική κατανομή με παράμετρο q . Αν σε κάποια χρονική στιγμή ένας πελάτης φτάσει στο σύστημα για να εξυπηρετηθεί και ο σταθμός εξυπηρέτησης είναι κατειλημμένος, τότε ο πελάτης εισέρχεται σε μία ουρά και περιμένει μέχρι να ολοκληρωθεί η εξυπηρέτηση του πελάτη που εξυπηρετείται. Αν κατά την άφιξη ενός πελάτη η χωρητικότητα του συστήματος είναι συμπληρωμένη (υπάρχουν στο σύστημα N πελάτες), τότε ο πελάτης αποβάλλεται και δεν μπορεί να εισέλθει στο σύστημα. Το σύστημα που περιγράψαμε είναι η ουρά *Geo/Geo/1/N* (Σχήμα 1).

Σχήμα 1. Ουρά *Geo/Geo/1/N*



Η μελέτη αυτού του συστήματος αναμονής μπορεί να γίνει με τη βοήθεια ενός Μαρκοβιανού συστήματος διακριτού χρόνου με χωρητικότητες στις καταστάσεις του

($OM\Sigma/c$). Θεωρούμε ένα $OM\Sigma$ διακριτού χρόνου το οποίο αποτελείται από 2 καταστάσεις και $N+1$ μέλη, τα οποία αντιστοιχούν στους πελάτες του συστήματος αναμονής και μπορούν να βρίσκονται είτε στις καταστάσεις είτε στους αποθηκευτικούς χώρους αυτών. Η πρώτη κατάσταση αντιστοιχεί στην πηγή πελατών από την οποία πραγματοποιούνται οι αφίξεις των πελατών, ενώ η δεύτερη κατάσταση αντιστοιχεί στο σύστημα αναμονής. Αφού στο σύστημα αναμονής υπάρχει ένα σημείο εξυπηρέτησης, από την δεύτερη κατάσταση μπορεί να μετακινηθεί μόνο ένα μέλος. Επίσης, σε κάθε χρονική στιγμή μπορεί να έχουμε το πολύ μία άφιξη στο σύστημα αναμονής, επομένως και από την πρώτη κατάσταση μπορεί να μετακινηθεί μόνο ένα μέλος. Έτσι, οι χωρητικότητες των καταστάσεων του $OM\Sigma$ είναι $c_i = 1$, $i = 1, 2$. Τα μέλη του συστήματος που υπερχειλίζουν από την κατάσταση i , $i = 1, 2$, λόγω του περιορισμού της χωρητικότητας εισέρχονται στον αποθηκευτικό χώρο αυτής. Προφανώς ο αποθηκευτικός χώρος της δεύτερης κατάστασης αντιστοιχεί στην ουρά του συστήματος αναμονής.

Ένα μέλος του συστήματος που σε κάποια χρονική στιγμή t είναι στην κατάσταση 1 (όχι στον αποθηκευτικό χώρο αυτής), μετακινείται στην κατάσταση 2 με πιθανότητα $p_{12} = p$, αφού η πιθανότητα να έχουμε μία άφιξη στο σύστημα αναμονής κατά τη χρονική στιγμή t είναι ίση με p . Επίσης, η πιθανότητα ένα μέλος που βρίσκεται στην κατάσταση 1 τη χρονική στιγμή t , να παραμείνει στην ίδια κατάσταση είναι $p_{11} = 1 - p$, αφού η πιθανότητα να μην έχουμε άφιξη στο σύστημα αναμονής κατά τη χρονική στιγμή t είναι ίση με $1 - p$.

Όσον αφορά τις μετακινήσεις από τη δεύτερη κατάσταση ισχύουν τα εξής: Ένα μέλος του συστήματος που σε κάποια χρονική στιγμή t είναι στην κατάσταση 2 (όχι στον αποθηκευτικό χώρο αυτής), μετακινείται στην κατάσταση 1 με πιθανότητα $p_{21} = q$, αφού η πιθανότητα να εξυπηρετηθεί ένα μέλος που βρίσκεται στο σημείο εξυπηρέτησης του συστήματος αναμονής είναι ίση με q . Επίσης, η πιθανότητα ένα μέλος που βρίσκεται στην κατάσταση 2 τη χρονική στιγμή t , να παραμείνει στην ίδια κατάσταση είναι $p_{22} = 1 - q$, αφού η πιθανότητα να μην εξυπηρετηθεί ένα μέλος που βρίσκεται στο σημείο εξυπηρέτησης του συστήματος αναμονής είναι ίση με $1 - q$.

Έτσι, οι μετακινήσεις των μελών του συστήματος από κατάσταση σε κατάσταση πραγματοποιούνται σύμφωνα με τον πίνακα μετάβασης

$$\mathbf{P} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}.$$

Η πληθυσμιακή δομή του συστήματος τη χρονική στιγμή t περιγράφεται από τα διανύσματα

$$\mathbf{m}(t) = (m_1(t), m_2(t)),$$

$$\mathbf{n}(t) = (n_1(t), n_2(t)),$$

$$\mathbf{e}(t) = (e_1(t), e_2(t)).$$

Χρησιμοποιώντας τα αποτελέσματα που δίνονται στην προηγούμενη ενότητα μπορούμε να υπολογίσουμε για οποιαδήποτε χρονική στιγμή t τόσο τις αναμενόμενες τιμές των τ.μ. $m_i(t)$, $n_i(t)$, $e_i(t)$, $i=1,2$, όσο και τις κατανομές αυτών.

Θεωρώντας ότι στο ΟΜΣ υπάρχουν $N+1$ μέλη οι δυνατές τιμές για την τ.μ. $m_2(t)$ είναι $0,1,2,\dots,N,N+1$. Στο σύστημα αναμονής $Geo/Geo/1/N$ το πλήθος των πελατών δεν μπορεί να υπερβαίνει την τιμή N . Έτσι κατά την εφαρμογή του αλγορίθμου υπολογισμού της κατανομής του διανύσματος $\mathbf{m}(t)$, για να μπορέσουμε να υπολογίσουμε την κατανομή του μεγέθους του συστήματος αναμονής, θεωρούμε την τ.μ. $\bar{\mathbf{m}}(t) = (\bar{m}_1(t), \bar{m}_2(t))$ για την οποία για κάθε χρονική στιγμή t ισχύει

$$P[\bar{\mathbf{m}}(t) = (m_1, m_2)] = P[\mathbf{m}(t) = (m_1, m_2)], m_2 \neq N+1, m_1 + m_2 = N+1,$$

και

$$P[\bar{\mathbf{m}}(t) = (1, N)] = P[\mathbf{m}(t) = (1, N)] + P[\mathbf{m}(t) = (0, N+1)].$$

Με αυτό τον τρόπο εξασφαλίζουμε ότι σε κάθε χρονική στιγμή στην πρώτη κατάσταση του ΟΜΣ που αντιστοιχεί στην πηγή πελατών του συστήματος αναμονής, θα υπάρχει τουλάχιστον ένας πελάτης ο οποίος θα μπορεί να ζητήσει να εξυπηρετηθεί από το σύστημα την επόμενη χρονική στιγμή.

Αντίστοιχα μπορούμε να θεωρήσουμε τις τυχαίες μεταβλητές $\tilde{\mathbf{n}}(t) = (\tilde{n}_1(t), \tilde{n}_2(t))$, $\tilde{\mathbf{e}}(t) = (\tilde{e}_1(t), \tilde{e}_2(t))$, η κατανομή των οποίων για οποιαδήποτε χρονική στιγμή t προκύπτει άμεσα από την κατανομή της τ.μ. $\bar{\mathbf{m}}(t) = (\bar{m}_1(t), \bar{m}_2(t))$. Έτσι, για κάθε χρονική στιγμή t , το πλήθος των πελατών του συστήματος αναμονής θα δίνεται από την τ.μ. $\bar{m}_2(t)$, το πλήθος των πελατών που εξυπηρετείται στο σημείο εξυπηρέτησης θα δίνεται από την τ.μ. $\tilde{n}_2(t)$, ενώ το πλήθος των πελατών που βρίσκονται στην ουρά του συστήματος αναμονής θα δίνεται από την τ.μ. $\tilde{e}_2(t)$. Τόσο οι αναμενόμενες τιμές όσο και η κατανομή αυτών των τ.μ. μπορούν να υπολογιστούν για κάθε χρονική στιγμή t εφαρμόζοντας τον αλγόριθμο που περιγράψαμε.

4. ΕΦΑΡΜΟΓΗ

Θεωρούμε ότι στο τηλεφωνικό κέντρο εξυπηρέτησης πελατών μιας εταιρίας δουλεύει ένας μόνο τηλεφωνητής. Εκτός από την τηλεφωνική γραμμή του πελάτη που εξυπηρετείται, το τηλεφωνικό κέντρο έχει 4 ακόμα γραμμές «αναμονής» για πελάτες που αναμένουν να εξυπηρετηθούν μέχρι να ελευθερωθεί ο τηλεφωνητής. Σε περίπτωση που και οι 5 τηλεφωνικές γραμμές (η γραμμή του τηλεφωνητή και οι 4 της «αναμονής») είναι απασχολημένες, οποιοσδήποτε πελάτης τηλεφωνεί παίρνει ένα σήμα «κατελημμένο», οπότε φεύγει για πάντα και καλεί μίαν άλλη εταιρία. Υποθέτουμε ότι οι πελάτες τηλεφωνούν στο κέντρο σε χρόνο διακριτό (κάθε 1 λεπτό) σύμφωνα με μία κατανομή Bernoulli με παράμετρο $p = 0.3$, δηλαδή κάθε χρονική

στιγμή (λεπτό) η πιθανότητα να τηλεφωνήσει ένας πελάτης είναι 0.3. Επίσης, υποθέτουμε ότι ο χρόνος που απαιτείται για την ολοκλήρωση των τηλεφωνικών συνδιαλέξεων ακολουθεί γεωμετρική κατανομή με παράμετρο $q = 0.6$.

Το σύστημα αυτό μπορεί να μελετηθεί με τη βοήθεια ενός ΟΜΣ/ c διακριτού χρόνου το οποίο αποτελείται από 2 καταστάσεις και τα μέλη του είναι τα τηλεφωνήματα των πελατών. Η πρώτη κατάσταση του συστήματος αντιπροσωπεύει την πηγή των πελατών. Σε κάθε χρονική μπορεί να πραγματοποιηθεί ένα μόνο τηλεφώνημα, δηλαδή μπορεί να πραγματοποιηθεί μετακίνηση ενός μόνο μέλους, επομένως η χωρητικότητα της πρώτης κατάστασης θα είναι $c_1 = 1$. Η δεύτερη κατάσταση του συστήματος αντιστοιχεί στο τηλεφωνικό κέντρο. Αφού υπάρχει ένας μόνο τηλεφωνητής, θα υπάρχει δυνατότητα μετακίνησης από την κατάσταση αυτή μόνο ενός μέλους, επομένως $c_2 = 1$. Τα τηλεφωνήματα που φθάνουν στο τηλεφωνικό κέντρο ενώ ο τηλεφωνητής είναι απασχολημένος, εισέρχονται στον αποθηκευτικό χώρο της κατάστασης, ο οποίος αντιπροσωπεύει την ουρά που σχηματίζεται από πελάτες που περιμένουν να εξυπηρετηθούν. Αφού το τηλεφωνικό κέντρο έχει 5 γραμμές, σε οποιαδήποτε χρονική στιγμή στην κατάσταση 2 και στον αποθηκευτικό χώρο αυτής δεν μπορούν να υπάρχουν πάνω από 5 μέλη. Επομένως το μέγεθος του ΟΜΣ/ c που θα χρησιμοποιήσουμε για την μελέτη του συστήματος αναμονής θα είναι $N = 6$ (ένα περισσότερο από τη χωρητικότητα του συστήματος αναμονής) και ο πίνακας μετάβασης \mathbf{P} των μελών του συστήματος αυτού θα είναι ο εξής

$$\mathbf{P} = \begin{pmatrix} 0.7 & 0.3 \\ 0.6 & 0.4 \end{pmatrix}.$$

Τη χρονική στιγμή $t=0$ το διάνυσμα κατάστασης είναι $\mathbf{m}(0) = (6, 0)^T$. Χρησιμοποιώντας τα θεωρητικά αποτελέσματα (Πρόταση 1 και Πόρισμα 1) και με τη βοήθεια ενός κώδικα που κατασκευάστηκε στην R, παρουσιάζουμε στον Πίνακα 1 τη συμπεριφορά (στην εξέλιξη του χρόνου) των αναμενόμενων τιμών για το πλήθος των πελατών που βρίσκεται στο τηλεφωνικό κέντρο, το πλήθος των πελατών που εξυπηρετείται και το πλήθος των πελατών που βρίσκεται στην αναμονή (ουρά του τηλεφωνικού κέντρου).

Πίνακας 1. Κατανομή - Αναμενόμενες τιμές

	$t = 1$	$t = 2$	$t = 3$		$t = 100$		$t = \infty$
$P[\mathbf{m}(t) = (6, 0)]$	0.7	0.616	0.57736		0.50048		0.50048
$P[\mathbf{m}(t) = (5, 1)]$	0.3	0.348	0.36		0.35748		0.35748
$P[\mathbf{m}(t) = (4, 2)]$	0	0.036	0.05832		0.10214		0.10214
$P[\mathbf{m}(t) = (3, 3)]$	0	0	0.00432		0.02918		0.02918
$P[\mathbf{m}(t) = (2, 4)]$	0	0	0		0.00834		0.00834
$P[\mathbf{m}(t) = (1, 5)]$	0	0	0		0.00238		0.00238
$E[\tilde{m}_1(t)]$	5.7	5.58	5.51		5.30543		5.30543
$E[\tilde{m}_2(t)]$	0.3	0.42	0.4896		0.69457		0.69457
$E[\tilde{n}_1(t)]$	1	1	1		1		1
$E[\tilde{n}_2(t)]$	0.3	0.384	0.42264		0.49952		0.49952
$E[\tilde{e}_1(t)]$	4.7	4.58	4.5104		4.30543		4.30543
$E[\tilde{e}_2(t)]$	0	0.036	0.06696		0.19505		0.19505

ABSTRACT

In this paper we consider the discrete-time homogeneous Markov system (HMS) with finite state capacities, which is a generalization of the classical HMS where the states' capacities are infinite. In order to examine the variability of the state sizes, recursive formulae for their factorial moments are derived. Next, the HMS with finite state capacities is used in order to examine the behavior of a discrete-time queue, where the arrivals and the service time are taken place according to the geometric distribution and the size of queue cannot exceed a finite number N .

ΑΝΑΦΟΡΕΣ

- Bartholomew, D.J. (1982). *Stochastic Models for Social Processes*, 3rd~edn., New York: John Wiley.
- Βασιλειάδης, Γ. και Τσακλίδης, Γ. (2010). Το ομογενές Μαρκοβιανό σύστημα διακριτού χρόνου με χωρητικότητες στις καταστάσεις και προτεραιότητες στις μετακινήσεις. *Πρακτικά 23ου Πανελληνίου Συνεδρίου Στατιστικής*, 60-67.
- Βασιλειάδης, Γ. (2014). Μελέτη ενός κλειστού δικτύου ουρών διακριτού χρόνου χρησιμοποιώντας το ομογενές Μαρκοβιανό σύστημα διακριτού χρόνου με

- χωρητικότητες στις καταστάσεις του. *Πρακτικά 27ου Πανελληνίου Συνεδρίου Στατιστικής*, 28-41.
- Βασιλείου, Π.-Χ. και Τσακλίδης, Γ. (2001). *Εφαρμοσμένη Θεωρία Πινάκων*, Εκδόσεις Ζήτη, Θεσσαλονίκη.
- Dimitriou, V.A. and Tsantas, N. (2010). Evolution of a time dependent Markov model for training and recruitment decisions in manpower planning. *Linear Algebra Appl.*, **433**, 1950-1972.
- Isaacson, D.L. and Madsen, R.W. (1976). *Markov Chains. Theory and applications*, New York: John Wiley.
- McClellan, S.I., McAlea, B. and Millard, P. (1998). Using a Markov reward model to estimate spend-down costs for a geriatric department. *J. Operat. Res. Soc.*, **10**, 1021-1025.
- Tsaklidis G. and Soldatos K.P. (2003). Modelling of continuous time homogeneous Markov system with fixed size as elastic solid. *Appl. Math. Modell*, **27**, 877-887.
- Vasiliadis, G. and Tsaklidis, G. (2011). On the distributions of the state sizes of the closed discrete-time Homogeneous Markov System with finite state capacities (HMS/c). *Markov Processes and Related Fields*, **17**, 91-118.
- Vasiliadis, G. (2014). Transient analysis of the M/M/k/N/N queue using a continuous time homogeneous Markov system with finite state size capacity. *Communications in Statistics - Theory and Methods*, **43:7**, 1548-1562.
- Vasiliadis, G. (2016). Transient analysis of a finite source discrete-time queueing system using homogeneous Markov system with state size capacities (HMS/c). *Communications in Statistics - Theory and Methods*, **43:5**, 1403-1423.
- Vassiliou, P.-C. G. (1997). The evolution of the theory of non-homogeneous Markov systems. *Stochastic Models Data Anal.*, **13**, no. 3-4, 159-176.



ΑΝΑΛΥΣΗ ΤΗΣ ΔΙΑΔΙΚΤΥΑΚΗΣ ΚΟΙΝΟΤΗΤΑΣ ΕΡΩΤΗΣΕΩΝ ΚΑΙ ΑΠΑΝΤΗΣΕΩΝ CROSS VALIDATED

Οδυσσέας Γαλάνης, Μαρία Παπουτσόγλου, Ελευθέριος Αγγελής
Τμήμα Πληροφορικής, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
{odysseus, mparouts, lef}@csd.auth.gr

ΠΕΡΙΛΗΨΗ

Μια γνωστή διαδικτυακή κοινότητα ερωτήσεων και απαντήσεων είναι η «Cross Validated». Απευθύνεται σε χρήστες που ενδιαφέρονται για θέματα στατιστικής, μηχανικής μάθησης και ανάλυσης δεδομένων. Οι χρήστες μπορούν μέσω ερωτήσεων να λάβουν απάντηση από έμπειρους χρήστες και μέσω απαντήσεών τους να συνεισφέρουν με την εμπειρία τους στην κοινότητα. Για την συνεισφορά τους βαθμολογούνται από τα άλλα μέλη της κοινότητας και η αλληλεπίδρασή τους δημιουργεί περιεχόμενο. Στόχος μας είναι η μελέτη των θεμάτων, μεθόδων και τεχνικών που ενδιαφέρουν την κοινότητα. Επίσης μελετάται η αλληλεπίδραση των χρηστών, οι βαθμολογίες τους και οι υποκοινότητες που σχηματίζουν. Για να επιτευχθούν οι στόχοι της ανάλυσης χρησιμοποιούνται τεχνικές μοντελοποίησης και ανίχνευσης θεμάτων και γραφήματα. Τα αποτελέσματα αναδεικνύουν συγκεκριμένες τάσεις της κοινότητας.

Λέξεις Κλειδιά: Q&A, Cross Validated, Stack Exchange, Κοινωνικά Δίκτυα, Γραφήματα

1. ΕΙΣΑΓΩΓΗ

Τα τελευταία χρόνια μέσα από την ανάπτυξη των τεχνολογιών Ιστού υπάρχει υψηλή παραγωγή περιεχομένου από τους χρήστες. Αυτό το περιεχόμενο συμβάλλει σημαντικά στην ανταλλαγή γνώσης, ευχρηστία, διαλειτουργικότητα και ανάπτυξη συμμετοχικότητας μεταξύ των χρηστών (Vickery and Wunsch-Vincent, 2007). Η συνεργατικότητα που έχει δημιουργηθεί με την ανταλλαγή περιεχομένου έχει συμβάλει στην ανάπτυξη δημοφιλών κοινοτήτων συνεργατικού περιεχομένου όπως είναι η Wikipedia ή ανταλλαγής γνώσεων μέσω ερωτήσεων και απαντήσεων (Q&A) με γνωστές κοινότητες όπως το Quora και τα Stack Exchange Sites (π.χ. Stack Overflow, Cross Validated κλπ.)

Οι Q&A κοινότητες συμβάλουν στην ανταλλαγή γνώσεων μέσω διαμόρφωσης των προβλημάτων των χρηστών σε μορφή ερωτήσεων. Πιο συγκεκριμένα, σε μια Q&A κοινότητα ένα μέλος απευθύνει μία ερώτηση και οι υπόλοιποι χρήστες μπορούν να διατυπώσουν μία απάντηση ανάλογα με την ειδικότητά τους. Οι κοινότητες για να κάνουν πιο θελκτική και ποιοτική την διαμόρφωση ερωτήσεων και

απαντήσεων προσφέρουν κίνητρα στους χρήστες. Αυτά τα κίνητρα δίνονται μέσω μηχανισμών αξιολόγησης τόσο των ερωτήσεων όσο και των απαντήσεων αλλά και των χρηστών. Επίσης λαμβάνονται μέτρα ώστε να μην επαναλαμβάνονται ερωτήσεις οι οποίες έχουν ήδη απαντηθεί.

Ένα χαρακτηριστικό των Q&A κοινοτήτων είναι πως η αξία τους βρίσκεται όχι μόνο στο ότι κάποιο συγκεκριμένο μέλος βρίσκει την απάντηση που ζητά τη συγκεκριμένη στιγμή, αλλά, κυρίως, στη συνεργατική δημιουργία μίας αποθήκης γνώσης, η οποία είναι χρήσιμη για την ευρύτερη κοινότητα σε βάθος χρόνου (Anderson et al. 2012, Vance 2015). Επομένως αξίζει να σημειωθεί ότι το χαρακτηριστικό των Q&A κοινοτήτων δεν είναι απλώς το ότι μπορεί ένα μέλος που θέτει μια ερώτηση να λάβει (στις περισσότερες περιπτώσεις) σε μικρό χρονικό διάστημα μια πρόταση λύσης μέσω κάποιας απάντησης, αλλά η δημιουργία μιας αποθήκης γνώσης. Η δημιουργία αυτής της γνώσης έχει θεμέλιο λίθο την συνεργατική δημιουργία που διαμορφώνεται μέσω των ερωτήσεων και των απαντήσεων.

Υπάρχουν επίσης ερευνητικά αποτελέσματα (π.χ. Movshovitz-Attias et al., 2013) τα οποία δείχνουν πως σε εξειδικευμένες κοινότητες, οι περισσότερες και ποιοτικότερες απαντήσεις δίνονται από έναν σχετικά μικρό αριθμό χρηστών. Αυτός είναι και ο λόγος για τον οποίο η μελέτη των αλληλεπιδράσεων (ερωτήσεων και απαντήσεων) αυτών των χρηστών έχει ιδιαίτερο ενδιαφέρον, αφού αυτές οι αλληλεπιδράσεις καθορίζουν σε μεγάλο βαθμό τη συνολική αξία της κοινότητας.

Η Q&A κοινότητα Cross Validated είναι μέλος του δικτύου Stack Exchange, το οποίο αποτελείται από 174 Q&A κοινότητες, κάθε μία από τις οποίες εξειδικεύεται σε κάποιο συγκεκριμένο αντικείμενο. Παράδειγμα η Stack Overflow, η οποία εξειδικεύεται σε θέματα προγραμματισμού, είναι η πιο γνωστή από αυτές τις κοινότητες. Το αντικείμενο μελέτης της παρούσας εργασίας είναι η κοινότητα Cross Validated, η οποία απευθύνεται σε χρήστες που ενδιαφέρονται για στατιστική, μηχανική μάθηση, ανάλυση δεδομένων, εξόρυξη δεδομένων και απεικόνιση δεδομένων.

Στην Cross Validated, όπως και στις υπόλοιπες κοινότητες του δικτύου Stack Exchange, ο τρόπος με τον οποίο δημιουργείται περιεχόμενο είναι με την υποβολή μίας ερώτησης. Στη συνέχεια, από τις απαντήσεις που υποβάλλονται, οι καλύτερες ανέρχονται στην κορυφή, χάρη σε ένα ποσοτικό σύστημα αξιολόγησης. Στις οδηγίες χρήσης της κοινότητας δίνεται έμφαση στο ότι η αλληλεπίδραση των χρηστών πρέπει να γίνεται όχι με διάλογο, αλλά με ερωτήσεις και απαντήσεις.

2. ΠΡΟΗΓΟΥΜΕΝΗ ΕΡΕΥΝΑ

Αν και οι Q&A κοινότητες, και ειδικά η Stack Overflow, έχουν αποτελέσει αντικείμενο ενός μεγάλου αριθμού εργασιών, δεν ισχύει κάτι τέτοιο για την Cross Validated. Για αυτό τον λόγο η παρούσα εργασία στηρίχθηκε κυρίως σε μεθόδους που χρησιμοποιήθηκαν για τη μελέτη της Stack Overflow.

Οι Anderson et al. (2012) μελέτησαν την μακροπρόθεσμη αξία των ερωτήσεων της Stack Overflow, δηλαδή το πώς οι ερωτήσεις και οι απαντήσεις τους

χρησιμοποιούνται ως γνώση σε βάθος χρόνου και όχι απλώς πως απαντώνται για να λύσουν το συγκεκριμένο πρόβλημα του συγκεκριμένου χρήστη τη συγκεκριμένη στιγμή. Επίσης, ασχολήθηκαν με το πρόβλημα του πως μπορεί να προσδιοριστεί αν μία ερώτηση έχει απαντηθεί επαρκώς. Για να απαντήσουν αυτά τα ερωτήματα, μελέτησαν τη δυναμική της κοινότητας όσον αφορά την απάντηση των ερωτήσεων.

Οι Barua et al. (2014) ασχολήθηκαν με τα θέματα που απασχολούν την κοινότητα Stack Overflow. Συγκεκριμένα, τα ερευνητικά ερωτήματα που έθεσαν ήταν ποια είναι τα κυριότερα θέματα συζήτησης στην κοινότητα, αν μία ερώτηση σε ένα θέμα προκαλεί απαντήσεις σε άλλα θέματα, πως αλλάζουν τα ενδιαφέροντα των χρηστών με το χρόνο και πως μεταβάλλεται το ενδιαφέρον για συγκεκριμένες τεχνολογίες.

Άλλοι ερευνητές που ασχολήθηκαν με τα θέματα συζήτησης της κοινότητας ήταν οι Gruetze et al. (2016) οι οποίοι χρησιμοποίησαν τις ετικέτες των ερωτήσεων και οι Johri and Bansal (2018) οι οποίοι χρησιμοποίησαν τα κείμενα των αναρτήσεων (ερωτήσεων και απαντήσεων).

Οι Pal et al. (2012) μελέτησαν την χρονική εξέλιξη των χρηστών της κοινότητας Stack Overflow. Οι Movshovitz-Attias et al. (2013) ασχολήθηκαν επίσης με τους χρήστες και συγκεκριμένα με το σύστημα βαθμολογίας/αξιολόγησης. Συγκεκριμένα μελέτησαν τους τρόπους με τους οποίους συμμετέχουν οι χρήστες με μεγάλη και με μικρή βαθμολογία. Διαπίστωσαν πως, όπως ήταν το αναμενόμενο, οι χρήστες με μεγάλη βαθμολογία ήταν η κύρια πηγή ποιοτικών απαντήσεων. Επίσης, πως αν και οι περισσότερες ερωτήσεις προέρχονται από χρήστες με χαμηλή βαθμολογία, ένας τυπικός χρήστης με υψηλή βαθμολογία θέτει περισσότερες ερωτήσεις από έναν τυπικό χρήστη με χαμηλή βαθμολογία.

2.1 Latent Dirichlet Allocation (LDA)

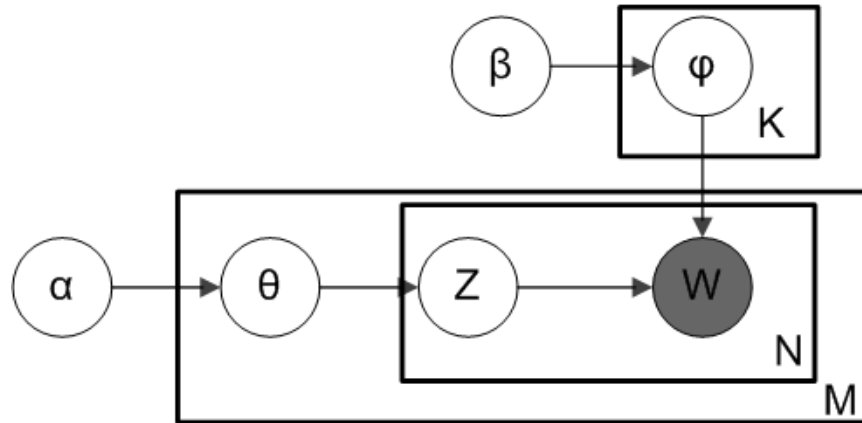
Η μέθοδος Latent Dirichlet Allocation (LDA, Blei et al. 2003) είναι ένας αλγόριθμος ο οποίος μοντελοποιεί κάθε έγγραφο ενός συνόλου εγγράφων ως μίγμα ενός πεπερασμένου αριθμού θεμάτων. Κάθε θέμα, με τη σειρά του, χαρακτηρίζεται από ένα μίγμα όρων (λέξεων).

Αναλυτικά το μοντέλο της μεθόδου LDA, φαίνεται στην Εικόνα 1 με τη μορφή plate notation. M είναι το πλήθος των κειμένων (δηλαδή των ερωτήσεων), N το πλήθος των λέξεων σε κάθε κείμενο και K το πλήθος των θεμάτων (topics). Ο λέξεις w είναι η μόνη παράμετρος που μας δίνεται. Όλες οι υπόλοιπες είναι λανθάνουσες (latent) και θεωρούμε πως παράγονται ως εξής:

- Παράγεται η κατανομή των θεμάτων για κάθε έγγραφο $\theta_i \sim \text{Dir}(\alpha)$
- Παράγεται η κατανομή των όρων για κάθε θέμα $\phi_k \sim \text{Dir}(\beta)$
- Η λέξη j του κειμένου i παράγεται ως εξής:
- Παράγεται το θέμα $z_{ij} \sim \text{Multinomial}(\theta_i)$
- Παράγεται η λέξη $w_{ij} \sim \text{Multinomial}(\phi_{z_{ij}})$

Το πλήθος των θεμάτων στο μοντέλο είναι μία από τις εισόδους του αλγορίθμου και κατά συνέπεια πρέπει να οριστεί εκ των προτέρων.

Εικόνα 1. Παραγωγή κειμένων με το μοντέλο LDA (χρήστης του Wikimedia Commons Slxu.public, διανεμημένο με άδεια CC BY-SA 3.0)



Η μέθοδος LDA δημιουργεί θέματα με περίπου ίση συμμετοχή στο σύνολο των κειμένων. Σε άλλες εφαρμογές, το πλήθος των θεμάτων επιλέγεται έτσι ώστε να μεγιστοποιεί την πιθανοφάνεια των παρατηρήσεων. Κατά κανόνα προκύπτει ένας μεγάλος αριθμός θεμάτων που εξηγεί ικανοποιητικά τις παρατηρήσεις, αλλά είναι αδύνατον να ερμηνευθεί.

2.2 PageRank

Ο αλγόριθμος PageRank (Page et al. 1999) εφαρμόζεται σε ένα κατευθυνόμενο ή μη κατευθυνόμενο γράφημα και δίνει ένα ποσοτικό μέτρο του πόσο σημαντική είναι κάθε κορυφή του. Η βασική ιδέα είναι να βρεθεί η πιθανότητα μία διαδρομή να καταλήγει στην συγκεκριμένη κορυφή ακολουθώντας μία τυχαία ακμή (με πιθανότητα d , όπου το d ονομάζεται παράγοντας απόσβεσης, damping factor), ή μεταβαίνοντας σε μία τυχαία κορυφή (με πιθανότητα $1-d$). Αυτή η διαδικασία έχει ως αποτέλεσμα να έχουν μεγάλο PageRank κορυφές στις οποίες καταλήγουν πολλές ακμές, όπως επίσης και κορυφές στις οποίες καταλήγουν ακμές που προέρχονται από κορυφές με μεγάλο PageRank.

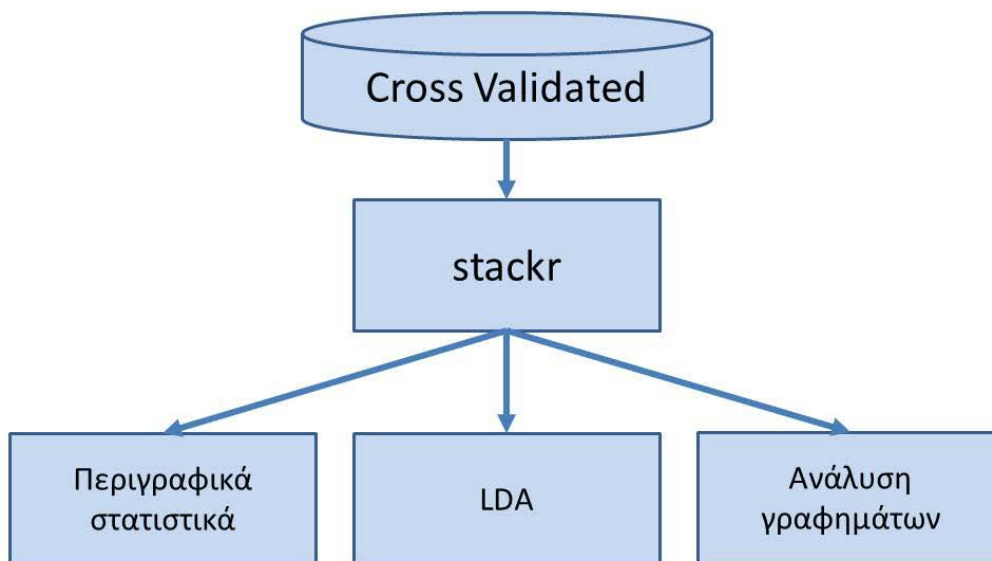
Για την εφαρμογή του αλγορίθμου PageRank σε Q&A κοινότητες, οι κορυφές αντιπροσωπεύουν τους χρήστες και οι ακμές τις μεταξύ τους αλληλεπιδράσεις. Αλληλεπίδραση θεωρείται η απάντηση σε ερώτηση. Το γράφημα μπορεί να είναι κατευθυνόμενο (με κατεύθυνση από τον χρήστη που θέτει την ερώτηση προς τον χρήστη που την απαντά), ή μη κατευθυνόμενο.

3. ΔΕΔΟΜΕΝΑ

Η εξαγωγή όλων των πληροφοριών που χρειαζόμαστε από την κοινότητα Cross Validated, γίνεται με τη χρήση της διεπαφής προγραμματισμού (API) του Stack Exchange (Stack Exchange, 2019). Για την ακρίβεια χρησιμοποιούμε τη βιβλιοθήκη

stackr της γλώσσας R (Robinson, 2018), η οποία αναλαμβάνει να απευθύνει τα σωστά αιτήματα στο API. Τα ακατέργαστα δεδομένα, δηλαδή ερωτήσεις, απαντήσεις και στοιχεία των χρηστών, αποθηκεύονται σε αρχεία κειμένου (CSV), για να μπορούν να διαβαστούν εύκολα από τα προγράμματα που τα επεξεργάζονται. Όλα αυτά περιγράφονται γραφικά στην Εικόνα 2.

Εικόνα 2. Ανάκτηση και ανάλυση δεδομένων από το Cross Validated



3.1 Μεταφόρτωση δεδομένων

Αν και όλες οι πληροφορίες που χρησιμοποιούνται είναι προσβάσιμες μέσω του Παγκόσμιου Ιστού, ο πιο αποδοτικός τρόπος να συλλεχθούν τα δεδομένα είναι μέσω της διεπαφής προγραμματισμού (API) του Stack Exchange. Με αυτή τη διεπαφή, ο χρήστης απευθύνει ένα αίτημα με τη μορφή URI στο API και παίρνει ως απάντηση ένα αρχείο σε μορφοποίηση JSON με τα δεδομένα τα οποία ζήτησε. Η βιβλιοθήκη stackr, απλοποιεί αρκετά τη διαδικασία, αφού αναλαμβάνει να συντάξει ένα έγκυρο URI και να επιστρέφει τα αποτελέσματα σε μορφή dataframe που είναι ο τύπος δεδομένων που χρησιμοποιείται στη γλώσσα R. Επίσης, φροντίζει να ικανοποιούνται οι χρονικοί περιορισμοί στη χρήση του API και διαχειρίζεται τα σχετικά σφάλματα.

3.2 Επεξεργασία δεδομένων

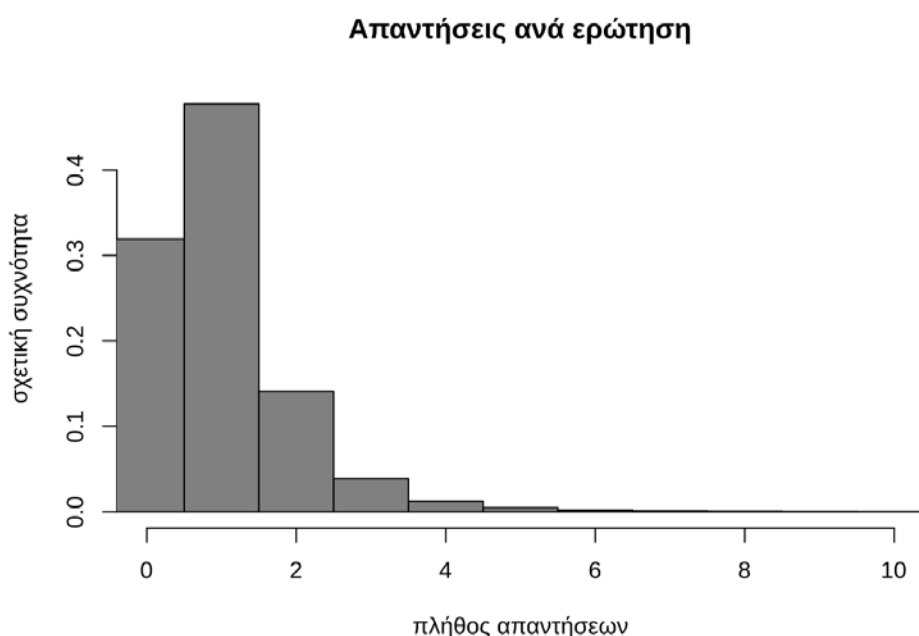
Για την επεξεργασία των δεδομένων χρησιμοποιήθηκαν προγράμματα γραμμένα στη γλώσσα R. Τα αποτελέσματα έχουν τη μορφή πινάκων και γραφικών παραστάσεων. Ήταν απαραίτητο να χρησιμοποιηθούν κάποιες εξειδικευμένες βιβλιοθήκες, κυρίως για την επεξεργασία του κειμένου των ερωτήσεων. Αυτές ήταν οι topicmodels (μοντελοποίηση θεμάτων, Grün and Kurt, 2011), tidytext (εξόρυξη κειμένου, Silge and Robinson, 2016) και SnowballC (αφαίρεση καταλήξεων,

Bouchet-Valat, 2014). Ο κώδικας που χρησιμοποιήθηκε για την επεξεργασία των δεδομένων και για την παραγωγή των γραφικών παραστάσεων είναι διαθέσιμος στο GitHub (<https://github.com/ogalanis/CrossValidatedStats>)

4. ΕΡΩΤΗΣΕΙΣ ΚΑΙ ΑΠΑΝΤΗΣΕΙΣ

Από τις ερωτήσεις που τίθενται στην κοινότητα Cross Validated απαντάται περίπου το 65%. Αυτό το ποσοστό είναι μικρότερο σε σχέση με τις συγγενέστερες κοινότητες (Stack Overflow 71%, Mathematics 77%, Physics 81%) και από τα μικρότερα σε ολόκληρο το δίκτυο Stack Exchange. Από τις ερωτήσεις που απαντώνται, οι περισσότερες έχουν μόνο μία ή δύο απαντήσεις (Εικόνα 3).

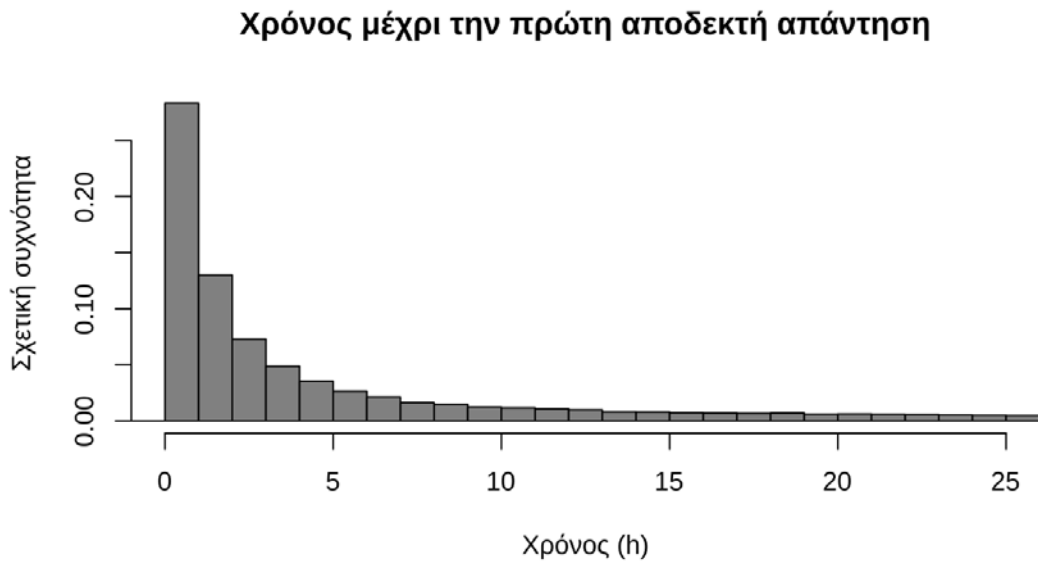
Εικόνα 3. Απαντήσεις ανά ερώτηση



Επειδή η ταχύτητα των απαντήσεων δεν συμβαδίζει πάντοτε με την ποιότητά τους (Chua and Banerjee, 2013), ο πιο αντικειμενικός τρόπος να μετρηθεί η ταχύτητα των απαντήσεων είναι ο χρόνος αναμονής μέχρι την απάντηση την οποία θεωρεί αποδεκτή ο ερωτών.

Σε συμφωνία με προηγούμενες εργασίες (Anderson et al., 2012), όταν υπάρχει αποδεκτή απάντηση, τότε αυτή έχει δοθεί τις πρώτες ώρες μετά την ερώτηση. Στην Εικόνα 4 φαίνεται ο χρόνος από την αρχική δημοσίευση της ερώτησης μέχρι την απάντηση η οποία χαρακτηρίζεται τελικά αποδεκτή. Στο ιστόγραμμα φαίνονται μόνο οι πρώτες 24 ώρες αφού η πιθανότητα να δοθεί αποδεκτή απάντηση αργότερα είναι πολύ μικρή. Οι σχετικές συχνότητες πιθανότητες που αναγράφονται υπολογίζονται στο σύνολο των ερωτήσεων που έχουν αποδεκτή απάντηση.

Εικόνα 4. Χρόνος μέχρι την πρώτη αποδεκτή απάντηση.



Εικόνα 5. Αριθμός ερωτήσεων ανά μήνα



Ένα μέτρο της δημοφιλίας της κοινότητας είναι ο αριθμός των ερωτήσεων που υποβάλλονται. Στην Εικόνα 5 δίνεται ο αριθμός των ερωτήσεων ανά μήνα από την έναρξη λειτουργίας της κοινότητας μέχρι σήμερα. Βλέπουμε πως ο αριθμός των ερωτήσεων ανά μήνα αυξάνεται σταθερά από την έναρξη λειτουργίας της κοινότητας. Διακρίνονται εποχικές μεταβολές, π.χ. σημαντική μείωση στις τελευταίες ημέρες κάθε έτους.

5. ΕΝΤΟΠΙΣΜΟΣ ΘΕΜΑΤΩΝ ΑΠΟ ΤΟ ΚΕΙΜΕΝΟ

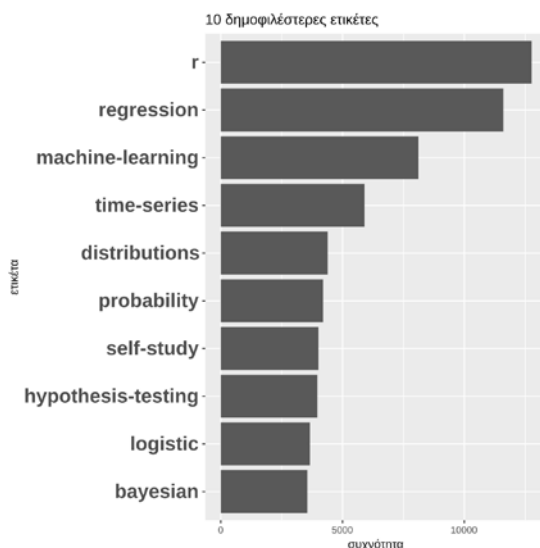
Για τη μελέτη των θεμάτων που απασχολούν την κοινότητα χρησιμοποιήθηκε τόσο το πλήρες κείμενο των ερωτήσεων, όσο και οι ετικέτες που τοποθετούν στην ερώτηση οι ερωτώντες.

5.1 Ετικέτες (tags)

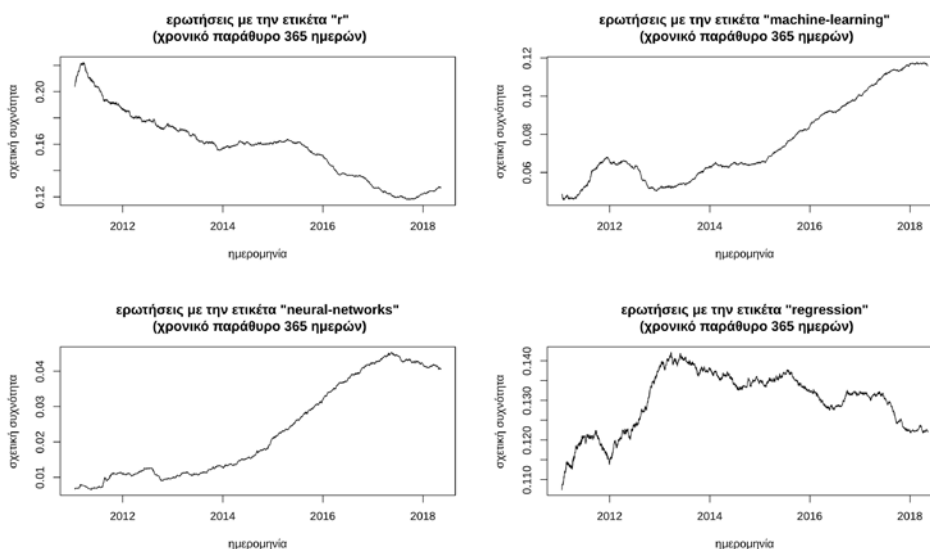
Σε κάθε ερώτηση, ο ερωτών μπορεί να τοποθετήσει μέχρι πέντε ετικέτες (tags) για να την κατηγοριοποιήσει και να διευκολύνει την αναζήτησή της. Κάθε ετικέτα έχει πολύ συγκεκριμένο ορισμό και δικαίωμα δημιουργίας νέας ετικέτας έχουν μόνο χρήστες με μεγάλη βαθμολογία. Οι πιο κοινές ετικέτες δίνονται στην Εικόνα 6. Οι ετικέτες που χρησιμοποιούνται περισσότερο, όπως είναι αναμενόμενο, έχουν σχέση με βασικές έννοιες της στατιστικής όπως κατανομές και πιθανότητες, ή με διαδοσόμενες τεχνικές όπως η μηχανική μάθηση. Η ετικέτα “r” αφορά τη γλώσσα προγραμματισμού R και όχι τον συντελεστή συσχέτισης ή οτιδήποτε άλλο μπορεί να συμβολίζεται με “r” ή “R”. Η ετικέτα “self-study” αφορά ερωτήσεις σχετικές με ασκήσεις ή εργασίες από την (τυπική ή μη) εκπαίδευση του χρήστη. Σε αυτή την περίπτωση, ασφαλώς, η κοινότητα δεν δίνει πλήρη λύση, αλλά μόνο κάποιες γενικές κατευθύνσεις.

Στην Εικόνα 7 δίνεται η συχνότητα κάποιων επιλεγμένων ετικετών σε συνάρτηση με το χρόνο. Για κάποιες ετικέτες η συχνότητά τους έχει αλλάξει σημαντικά με το πέρασμα του χρόνου. Π.χ. για την ετικέτα R (γλώσσα) έχει μειωθεί κατά 40% περίπου σε σχέση με τους πρώτους μήνες λειτουργίας της κοινότητας Cross Validated. Αυτό σίγουρα δεν οφείλεται στο ότι έχει μειωθεί η χρήση της R, δεδομένου ότι σύμφωνα με όλους τους σχετικούς δείκτες (PYPL, Carbonnelle, 2019; TIOBE, 2019), η R έχει γίνει πιο δημοφιλής σε όλο αυτό το διάστημα. Πιθανότατα οφείλεται σε ερωτήσεις οι οποίες ήταν εκτός θέματος, δηλαδή είχαν περισσότερο σχέση με προγραμματισμό παρά με στατιστική. Μάλιστα στην περιγραφή της ετικέτας αναφέρεται με έμφαση τόσο ότι πρέπει να μπαίνει σε ερωτήσεις εντός θέματος όσο και ότι οι ερωτήσεις δεν πρέπει να αφορούν αποκλειστικά την γλώσσα R. Η ετικέτα “machine-learning” (μηχανική μάθηση) έχει διπλασιάσει περίπου τη συχνότητά της στο διάστημα που λειτουργεί η κοινότητα Cross Validated. Παρομοίως η ετικέτα “neural-networks” (νευρωνικά δίκτυα) έχει σχεδόν τετραπλασιάσει τη συχνότητά της. Σε άλλες περιπτώσεις δεν υπάρχει κάποια έντονη ανοδική ή καθοδική τάση, όπως για παράδειγμα η ετικέτα “regression” (παλινδρόμηση), όπως είναι και το αναμενόμενο για έναν τόσο γενικό όρο.

Εικόνα 6. Συχνότητες εμφάνισης για τις δέκα δημοφιλέστερες ετικέτες στο Cross Validated.



Εικόνα 7. Συχνότητα εμφάνισης για τις ετικέτες “r” (επάνω αριστερά), “machine learning” (επάνω δεξιά), “neural networks” (κάτω αριστερά) και “regression” (κάτω δεξιά)



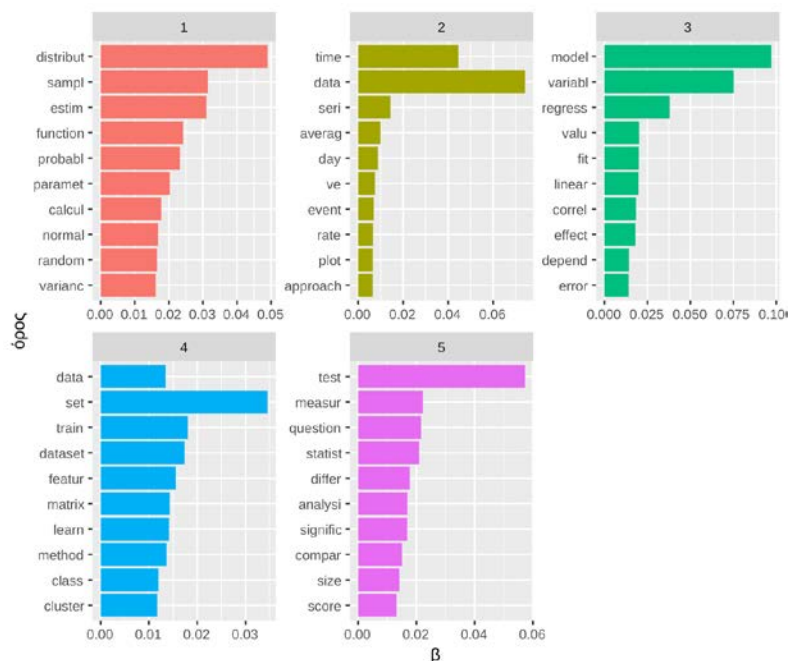
5.2 Μοντελοποίηση Θεμάτων

Για τη μοντελοποίηση των θεμάτων, εφαρμόστηκε η μέθοδος LDA στο πλήρες κείμενο των ερωτήσεων. Πριν από την κύρια επεξεργασία του κειμένου των ερωτήσεων, είναι απαραίτητη κάποια προεπεξεργασία (βλ. π.χ. Rodríguez et al. 2018)

με σκοπό την απομάκρυνση στοιχείων που δεν προσφέρουν πληροφορίες σχετικά με το θέμα της ερώτησης. Αφαιρούμε οτιδήποτε δεν είναι λέξη (μαθηματικοί τύποι, html tags κ.λπ.). Αφαιρούμε “stop words” (σύνδεσμοι, άρθρα, και γενικά πολύ κοινές λέξεις που δεν συμβάλλουν στην ανάλυση). Αφαιρούμε καταλήξεις για να ομαδοποιήσουμε τους διαφορετικούς τύπους της ίδιας λέξης (stemming). Στο τέλος αυτής της προεπεξεργασίας, αυτό που μένει είναι ένα πολυσύνολο (multiset, bag) λέξεων. Με αυτό το μοντέλο (bag-of-words) αγνοείται η γραμματική και το μόνο που χρησιμοποιείται είναι η συχνότητα της κάθε λέξης.

Επειδή, όπως αναφέρθηκε, η ερμηνεία είναι αδύνατη όταν το πλήθος των θεμάτων είναι μεγάλο, κάναμε δοκιμές με μικρό πλήθος θεμάτων. Τα αποτελέσματα του LDA για 5 θέματα φαίνονται στην Εικόνα 8. Για κάθε ένα από τα 5 θέματα δίνονται οι χαρακτηριστικότεροι όροι. Στο θέμα 1 βλέπουμε κυρίως όρους που περιγράφουν θεμελιώδεις έννοιες της στατιστικής όπως «κατανομή», «δείγμα» ή «πιθανότητα». Το θέμα 2 περιλαμβάνει όρους που έχουν σχέση με χρονοσειρές, όπως «χρόνος», «ημέρα» ή «ρυθμός». Το θέμα 3 έχει σχέση με στατιστικά μοντέλα, με όρους όπως «μοντέλο», «παλινδρόμηση» και «προσαρμογή» («fit»). Το θέμα 4 έχει σχέση με μηχανική μάθηση, αφού εκτός από τη λέξη «μάθηση», έχει τους όρους «σύνολο» και «εκπαίδευση», επειδή στις σχετικές ερωτήσεις εμφανίζονται οι όροι «σύνολο εκπαίδευσης» και «σύνολο πειραματισμού». Επίσης εμφανίζεται ο όρος «συστάδα» που χρησιμοποιείται στην μη επιβλεπόμενη μάθηση. Τέλος το θέμα 5 φαίνεται να έχει σχέση με τον έλεγχο υποθέσεων, αφού περιέχει όρους όπως «έλεγχος», «ερώτημα» και «σύγκριση».

Εικόνα 8. Οι συχνότερα χρησιμοποιούμενοι όροι για κάθε ένα από τα πέντε θέματα που προέκυψαν από τη μοντελοποίηση θεμάτων

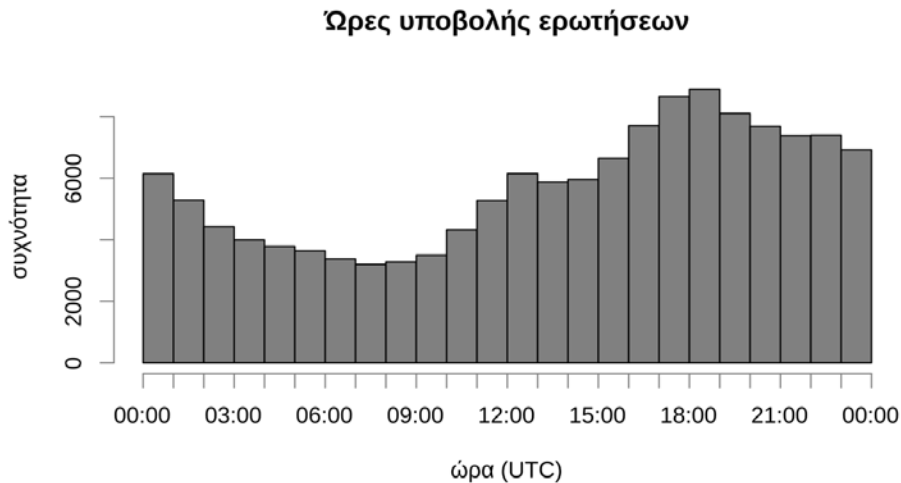


6. ΧΡΗΣΤΕΣ

Μία ιδέα για την προέλευση των χρηστών του Cross Validated μπορούμε να πάρουμε από τις ώρες της ημέρας που γίνονται οι ερωτήσεις (Εικόνα 9). Υπάρχει δραστηριότητα όλο το 24ωρο, όμως το μέγιστο εμφανίζεται τις ώρες που αντιστοιχούν σε ώρες γραφείου στη Β. Αμερική. Μερικές ώρες πριν από αυτό υπάρχει ένα μικρότερο μέγιστο που αντιστοιχεί στις ώρες γραφείου στην Ευρώπη.

Για τη μελέτη των χρηστών χρησιμοποιήθηκε τόσο η βαθμολογία τους (reputation) όπως υπολογίζεται από το σύστημα αξιολόγησης του Stack Overflow, όσο και η ανάλυση του γραφήματος των μεταξύ τους αλληλεπιδράσεων.

Εικόνα 9. Ώρες υποβολής των ερωτήσεων



6.1 Βαθμολογία Υπόληψης (reputation)

Η προσφορά κάθε χρήστη στην κοινότητα ποσοτικοποιείται με μία βαθμολογία υπόληψης (reputation). Κάθε ενέργεια του χρήστη (ερώτηση, απάντηση, σχόλιο, διόρθωση), αξιολογείται από τους υπόλοιπους χρήστες με θετική ή αρνητική ψήφο και βαθμολογείται με βάση ένα point system. Εκτός από ηθική επιβράβευση, υπάρχουν επίπεδα βαθμολογίας που ξεκλειδώνουν προνόμια μέσα στην κοινότητα, όπως δικαίωμα θετικής ψήφου, δικαίωμα αρνητικής ψήφου ή δικαίωμα σχολιασμού.

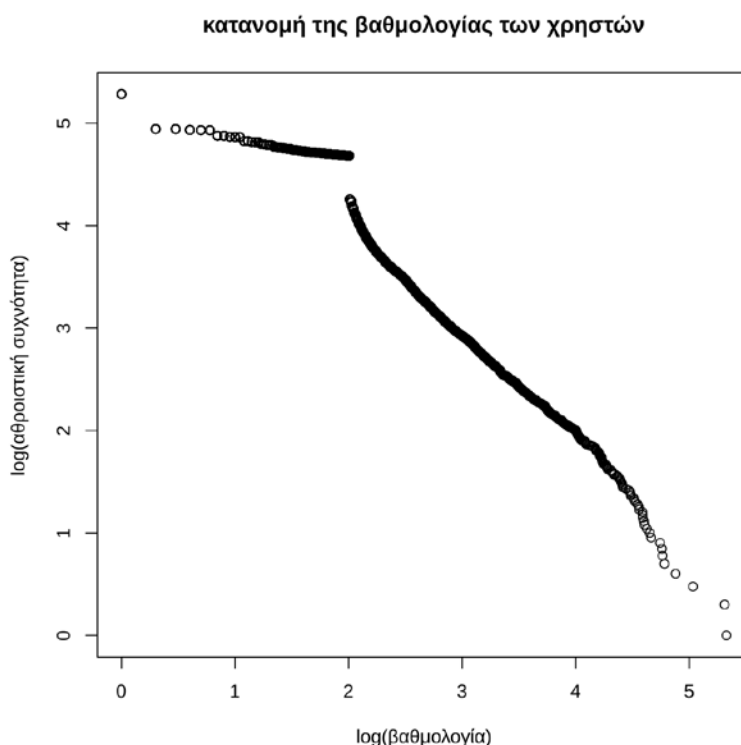
Όπως αναφέρθηκε, κάθε χρήστης αξιολογείται και βαθμολογείται για τη συνεισφορά του στην κοινότητα. Παριστάνοντας γραφικά το λογάριθμο της αθροιστικής συχνότητας σε συνάρτηση με το λογάριθμο της βαθμολογίας (reputation) του κάθε χρήστη (Εικόνα 10), διαπιστώνουμε πως η βαθμολογία ακολουθεί κατανομή power law. Υπάρχει ένα άλμα στη γραφική παράσταση, το οποίο οφείλεται σε ιδιαιτερότητες του συστήματος βαθμολόγησης. Αν κρατήσουμε τους χρήστες με βαθμολογία 100-200 ή μεγαλύτερη, τότε ο λογάριθμος της βαθμολογίας κάθε χρήστη είναι γραμμική συνάρτηση του λογαρίθμου του πλήθους των χρηστών που έχουν ίδια ή καλύτερη βαθμολογία. Με δεδομένο πως η υψηλή

βαθμολογία απορρέει από τη συχνή (και ποιοτική) δραστηριότητα, μπορούμε να συμπεράνουμε πως αυτή η δραστηριότητα οφείλεται σε έναν σχετικά μικρό αριθμό χρηστών.

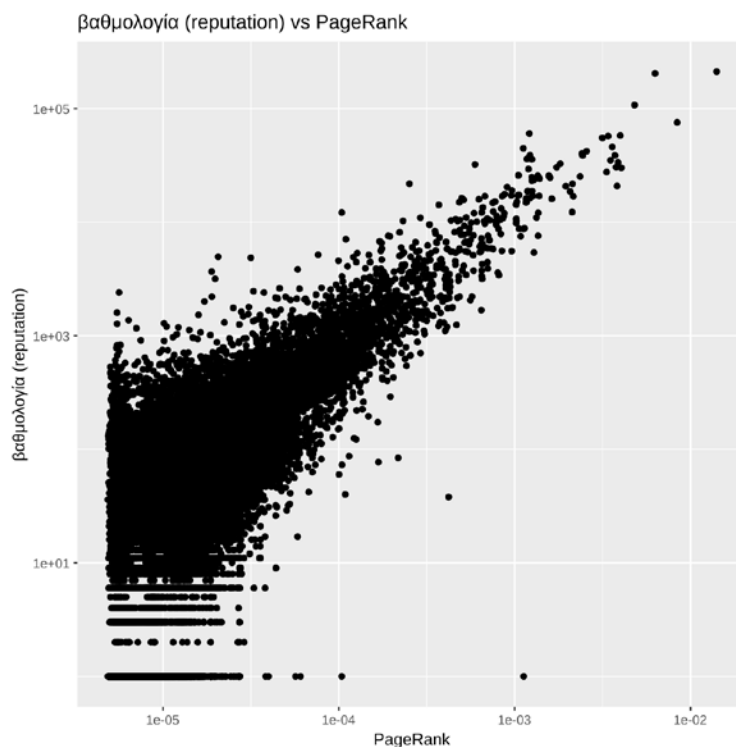
6.2 PageRank

Εκτός από το σύστημα βαθμολογίας υπόληψης στην κοινότητα Cross Validated, υπάρχουν και άλλοι τρόποι να εκτιμηθεί η σημασία ενός χρήστη. Χρησιμοποιήσαμε για αυτό τον σκοπό τον αλγόριθμο PageRank. Αυτός ο αλγόριθμος χρησιμοποιείται για να βρει τις σημαντικότερες κορυφές σε ένα γράφημα. Μπορεί να εφαρμοστεί τόσο σε κατευθυνόμενα όσο και σε μη κατευθυνόμενα γραφήματα. Η βασική ιδέα είναι πως όταν μία κορυφή έχει μεγάλο βαθμό, τότε θα πρέπει να είναι σημαντική, ειδικά αν συνδέεται με σημαντικές κορυφές. Για να εφαρμόσουμε τον αλγόριθμο PageRank θεωρούμε ένα μη κατευθυνόμενο γράφημα με κορυφές τους χρήστες και ακμές τις αλληλεπιδράσεις μεταξύ τους, δηλαδή τις ερωτήσεις και τις απαντήσεις. Στην Εικόνα 11 φαίνεται η σχέση μεταξύ των μετρικών reputation και PageRank. Υπάρχει θετική συσχέτιση μεταξύ τους, ειδικά για τους λίγους χρήστες που έχουν τις μεγαλύτερες τιμές. Κάποιες περιπτώσεις στις οποίες χρήστες με μεγάλο PageRank φαίνονται να έχουν reputation ίσο με 1, οφείλεται σε προσωρινή ποινή που δέχτηκαν από τους διαχειριστές της κοινότητας.

Εικόνα 10. Κατανομή της βαθμολογίας των χρηστών



Εικόνα 11. Σχέση μεταξύ των μετρικών reputation και PageRank



7. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η κοινότητα Cross Validated αυξάνεται σταθερά σε μέγεθος, όσον αφορά τον όγκο των ερωτήσεων και των απαντήσεων. Οι ερωτήσεις απαντώνται κυρίως τις πρώτες ώρες μετά από την ανάρτησή τους. Η συχνότητα των tags μπορεί να χρησιμοποιηθεί για τον εντοπισμό τάσεων μέσα στην κοινότητα, με προσοχή σε εξαιρέσεις όπως το tag “R”. Το topic modeling εντοπίζει τα θέματα που απασχολούν την κοινότητα, ακόμα και αν οι περισσότερες ερωτήσεις δεν κατατάσσονται ξεκάθαρα σε ένα μοναδικό θέμα. Η ανάπτυξη της κοινότητας οδηγείται κυρίως από έναν σχετικά μικρό αριθμό πολύ ενεργών χρηστών. Η βαθμολογία υπόληψης (reputation) είναι ένα σχετικά αντικειμενικό μέτρο της προσφοράς ενός χρήστη στην κοινότητα, όπως προκύπτει από τη σύγκριση με ένα αντικειμενικό μέτρο όπως το PageRank.

Η μελλοντική έρευνα επάνω στο Cross Validated μπορεί να θέσει στοχευμένα ερευνητικά ερωτήματα τα οποία ήταν πολύ εξειδικευμένα για μία πρώτη γενική επισκόπηση του θέματος όπως η παρούσα εργασία. Ένα ενδιαφέρον ερώτημα θα ήταν κατά πόσο υπάρχει εξειδίκευση στους λίγους, πολύ ενεργούς χρήστες και πως τα ενδιαφέροντά τους μεταβάλλονται με την πάροδο του χρόνου. Επίσης, αν δημιουργούνται υποομάδες χρηστών με κοινά ενδιαφέροντα και πως αυτές εξελίσσονται. Τέλος, η χρονική εξέλιξη των θεμάτων μπορεί να μελετηθεί σε μεγαλύτερο βάθος.

ABSTRACT

Cross Validated is a well-known Q and A community. It is devoted to statistics, machine learning and data analysis. Users can post questions and have them answered by experienced users, or answer questions themselves, contributing to the community. Users are quantitatively evaluated by other members of the community and through their interactions content is created. Our goal is to study the topics, methods and techniques that interest the community. We also study user interaction, their reputation scores, and the sub-communities they form. For our analysis we use topic modeling techniques and graphs. The results highlight trends in the community.

Ευχαριστίες: Το έργο συγχρηματοδοτείται από την Ελλάδα και την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) μέσω του Επιχειρησιακού Προγράμματος «Ανάπτυξη Ανθρώπινου Δυναμικού, Εκπαίδευση και Διά Βίου Μάθηση», στο πλαίσιο της Πράξης «Ενίσχυση του ανθρώπινου ερευνητικού δυναμικού μέσω της υλοποίησης διδακτορικής έρευνας» (MIS-5000432), που υλοποιεί το Ίδρυμα Κρατικών Υποτροφιών (IKY)

ΑΝΑΦΟΡΕΣ

- Anderson, A., Huttenlocher, D., Kleinberg, J. and Leskovec, J. (2012). "Discovering Value from Community Activity on Focused Question Answering Sites: A Case Study of Stack Overflow", in KDD, Beijing.
- Barua, A., Thomas, S.W. and Hassan, A.E. (2014). "What are developers talking about? An analysis of topics and trends in Stack Overflow" *Empirical Software Engineering*, **19**(3), 619-654.
- Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003). "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, **3**, pp. 993-1022, 2003.
- Bouchet-Valat, M. (2014). SnowballC: Snowball stemmers based on the C libstemmer UTF-8 library.
- Carbonnelle, P. (2019). "PYPL PopularitY of Programming Language", [Online]. Available: <http://pypl.github.io/PYPL.html>. [Accessed 4 10 2019]
- Chua, A.Y. and Banerjee, S. (2013). So fast so good: An analysis of answer quality and answer speed in community Question-answering sites, *Journal of the American Society for Information Science and Technology*, **64**(10), 2058-2068.
- Grün, B. and Kurt, H. (2011). "topicmodels: An R Package for Fitting Topic Models," *Journal of Statistical Software*, **40**(13), 1-30.
- Gruetze, T., Krestel, R. and Naumann, F. (2016). "Topic Shifts in StackOverflow: Ask it like Socrates," in International Conference on Applications of Natural Language to Information, Salford, United Kingdom.
- Johri, V. and Bansal, S. (2018). "Identifying trends in technologies and programming languages using Topic Modeling," in 12th IEEE International Conference on Semantic Computing, Laguna Hills, CA.
- Movshovitz-Attias, D., Movshovitz-Attias, Y., Steenkiste, P. and Faloutsos, C. (2013). "Analysis of the Reputation System and User Contributions on a Question Answering Website: StackOverflow," in 2013 IEEE/ACM International

- Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013), Niagara, Ontario, CAN.
- Page, L., Brin, S., Motwani, R. and Winograd, T (1999). "The PageRank Citation Ranking: Bringing Order to the Web,"
- Pal, A., Chang, S. and Konstan, J.A. (2012). "Evolution of Experts in Question Answering Communities" in Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, Dublin, Ireland.
- Robinson, D., (2018) stackr: Client for the Stack Exchange API.
- Rodríguez, L.J., Xiaoran, W. and Jilong, K. (2018). "Insights on Apache Spark Usage by Mining Stack Overflow Questions," in 2018 IEEE International Congress on Big Data, San Francisco, CA, USA.
- Silge, J. and Robinson, D. (2016). "tidytext: Text Mining and Analysis Using Tidy Data Principles in R," *JOSS*, **1**(3).
- Stack Exchange, Inc (2019). "Stack Exchange API", [Online]. Available: api.stackexchange.com. [Accessed 14 3 2019].
- TIOBE Software BV (2019). "The R Programming Language", [Online]. Available: <https://www.tiobe.com/tiobe-index/r/>. [Accessed 4 10 2019]
- Vance, E. A. (2015). Recent developments and their implications for the future of academic statistical consulting centers. *The American Statistician*, **69**(2), 127-137.
- Vickery, G. and Wunsch-Vincent, S. (2007). Participative Web and User-Created Content. Web 2.0, Wikis and Social Networking, Paris: OECD Publications, 2007, 124 p.



ΝΕΕΣ ΚΑΤΑΝΟΜΕΣ ΣΧΗΜΑΤΙΣΜΩΝ

Σ.Δ. Δαφνής¹, Φ.Σ. Μακρή², Μ.Β. Κούτρας¹

¹ Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς
dafnisspyros@gmail.com, mkoutras@unipi.gr

² Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών
makri@math.upatras.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία εισάγουμε και μελετάμε κατανομές οι οποίες σχετίζονται με το πλήθος εμφανίσεων σχηματισμών σε ακολουθίες ανεξάρτητων και ισόνομων δίτιμων δοκιμών, και γενικεύουν ή επεκτείνουν άλλες γνωστές κατανομές που έχουν μελετηθεί στο παρελθόν. Ο ακριβής υπολογισμός των συναρτήσεων πιθανότητας των κατανομών γίνεται με χρήση της μεθόδου εμφύτευσης σε Μαρκοβιανή αλυσίδα. Τέλος, προτείνουμε μια εφαρμογή των νέων αποτελεσμάτων στη βιοϊατρική μηχανική.

Λέξεις Κλειδιά: Ροές, Μαρκοβιανή αλυσίδα, κατανομές διωνυμικού τύπου, μηχανικός αερισμός

1. ΕΙΣΑΓΩΓΗ

Στο επίκεντρο του ενδιαφέροντος της Θεωρίας Πιθανοτήτων τις τελευταίες δεκαετίες είναι η μελέτη διακριτών κατανομών που σχετίζονται με την εμφάνιση σχηματισμών, λόγω της ευρείας εφαρμοσιμότητάς τους σε διάφορες επιστημονικές περιοχές (βλέπετε Balakrishnan and Koutras (2002)).

Ας θεωρήσουμε μια ακολουθία ανεξάρτητων και ισόνομων (α.ι.) δοκιμών Bernoulli Z_1, Z_2, \dots με κοινή πιθανότητα επιτυχίας $p = Pr(Z_t = 1)$ και αποτυχίας $q = Pr(Z_t = 0) = 1 - p$, $t = 1, 2, \dots$. Πρόσφατα οι Dafnis et al. (2019), μελετώντας ένα πρόβλημα στη Θεωρία Αξιοπιστίας, εισήγαγαν τις ακόλουθες οικογένειες σχηματισμών:

E : μια σειρά αποτελεσμάτων που περιέχουν k ($k \geq 2$) επιτυχίες (οι οποίες θα συμβολίζονται με S) στην οποία κάθε ζεύγος διαδοχικών επιτυχιών μπορεί να διακόπτεται από το πολύ r ($r \geq 0$) συνεχόμενες αποτυχίες (οι οποίες θα συμβολίζονται με F).

Κάθε τυπικό στοιχείο του E είναι της μορφής $S \underbrace{F \dots F}_{d_1} S \underbrace{F \dots F}_{d_2} S \dots S \underbrace{F \dots F}_{d_{k-1}} S$, όπου

$d_i \in \mathbb{N}, 0 \leq d_i \leq r$, για κάθε $i = 1, 2, \dots, k - 1$ (\mathbb{N} συμβολίζει το σύνολο των μη-αρνητικών ακεραίων).

Όμως, σε πολλές εφαρμογές, ιδιαίτερο ενδιαφέρον παρουσιάζουν τα συμπληρώματα των αρχικών σχηματισμών που μελετούνται (βλέπετε, π.χ., την περίπτωση του δυϊκού ενός πρωτεύοντος συστήματος στη Θεωρία Αξιοπιστίας). Εισάγουμε, λοιπόν, τον ακόλουθο σύνθετο σχηματισμό

E_1 : μια σειρά αποτελεσμάτων που περιέχουν k επιτυχίες στην οποία τουλάχιστον ένα ζεύγος διαδοχικών επιτυχιών διακόπτεται από τουλάχιστον r συνεχόμενες αποτυχίες (κάθε τυπικό στοιχείο του E_1 είναι της μορφής $\underbrace{SF \dots FS}_{l_1} \underbrace{SF \dots FS}_{l_2} \dots \underbrace{SF \dots FS}_{l_{k-1}}$,

όπου $l_i \in \mathbb{N}$ και $l_i \in [r, +\infty)$ για τουλάχιστον ένα $i, i = 1, 2, \dots, k - 1$). Ο E_1 είναι ένας σύνθετος σχηματισμός που περιέχει αριθμήσιμα άπειρους απλούς σχηματισμούς.

Ο νέος σχηματισμός E_1 περιέχει ακριβώς k επιτυχίες. Προκειμένου να επιτρέψουμε την ύπαρξη περισσότερων των k επιτυχιών στο σχηματισμό θα θεωρήσουμε την επόμενη οικογένεια, η οποία παράγεται επιτρέποντας την πρόσθεση ενός αριθμού επιτυχιών στο τέλος του τυπικού στοιχείου του σχηματισμού E_1 . Πιο συγκεκριμένα, το τυπικό στοιχείο της οικογένειας E_2 είναι της μορφής

$\underbrace{SF \dots FS}_{l_1} \underbrace{SF \dots FS}_{l_2} \dots \underbrace{SF \dots FSS \dots S}_{l_{k-1} \geq 0}$, όπου $l_i \in \mathbb{N}$ και $l_i \in [r, +\infty)$ για τουλάχιστον ένα

$i, i = 1, 2, \dots, k - 1$. Η αντίστοιχη λεκτική διατύπωση είναι:

E_2 : μια σειρά από τουλάχιστον k επιτυχίες διακόπτεται από τουλάχιστον r συνεχόμενες αποτυχίες μεταξύ οποιωνδήποτε δύο από τις πρώτες k επιτυχίες.

Ο σύνθετος σχηματισμός E_2 επίσης περιέχει αριθμήσιμα άπειρους απλούς σχηματισμούς.

Ας συμβολίσουμε με $N_{n,k,r}^{(i)}$ τον αριθμό των πραγματοποιήσεων μη-επικαλυπτόμενων σχηματισμών που περιλαμβάνονται στο σύνθετο σχηματισμό E_i ($i = 1, 2$) στην ακολουθία των αποτελεσμάτων Z_1, Z_2, \dots, Z_n ($n \geq k$).

Η τυχαία μεταβλητή (τ.μ.) $N_{n,k,r}^{(1)}$ αποτελεί γενίκευση μιας τ.μ. που απαριθμεί σχηματισμούς και έχει μελετηθεί εκτεταμένα στη βιβλιογραφία. Οι Dafnis et al. (2012) μελέτησαν την τ.μ. $N_{n,2,r}^{(1)}$ εφαρμόζοντας τη μέθοδο εμφύτευσης σε Μαρκοβιανή αλυσίδα και οι Sen and Goyal (2004) χρησιμοποιώντας συνδυαστικές μεθόδους. Συναφή αποτελέσματα υπάρχουν, επίσης, στους Makri and Psillakis (2012, 2013, 2014, 2017).

Στην παρούσα εργασία εφαρμόζουμε τη μέθοδο εμφύτευσης σε Μαρκοβιανή αλυσίδα για να μελετήσουμε τις κατανομές των $N_{n,k,r}^{(i)}$ ($i = 1, 2$), στην περίπτωση α.ι. δοκιμών. Προτείνουμε, επίσης, εφαρμογές και πιθανές επεκτάσεις των νέων αποτελεσμάτων.

Σε όλη την εργασία συμβολίζουμε με $[x]$ το μεγαλύτερο ακέραιο που είναι μικρότερος ή ίσος του x και με $\delta_{i,j}$ τη συνάρτηση Delta του Kronecker, δηλ.

$$\delta_{i,j} = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases}$$

2. ΠΡΟΚΑΤΑΡΚΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Ας θεωρήσουμε την ακολουθία Z_1, Z_2, \dots δίτιμων αποτελεσμάτων που παίρνουν τις τιμές 1 (επιτυχία, S) ή 0 (αποτυχία, F) και ας συμβολίσουμε με \mathcal{E} κάθε προκαθορισμένο σχηματισμό (απλό ή σύνθετο). Ο αριθμός των εμφανίσεων του \mathcal{E} στην ακολουθία Z_1, Z_2, \dots, Z_n (n σταθερός ακέραιος) θα συμβολίζεται με X_n . Σε ένα μεγάλο αριθμό εργασιών (βλέπετε, μεταξύ άλλων, τις εργασίες των Antzoulakos et al. (2003), Dafnis et al. (2010), Fu and Koutras (1994), Koutras (2003), Koutras and Alexandrou (1995)) αναπτύχθηκε μια μέθοδος εμφύτευσης σε Μαρκοβιανή αλυσίδα για τη μελέτη της ακριβούς κατανομής τ.μ. απαρίθμησης σε ακολουθίες δίτιμων δοκιμών (βλέπετε, επίσης, το βιβλίο των Fu and Lou (2003)). Πριν προχωρήσουμε στην ανάπτυξη των αποτελεσμάτων μας, θα κάνουμε μια σύντομη ανασκόπησή της, καθώς αυτή θα αποτελέσει το βασικό μας μεθοδολογικό εργαλείο.

Πρώτα εισάγουμε την έννοια της τ.μ. διωνυμικού τύπου εμφυτεύσιμης σε Μαρκοβιανή αλυσίδα. Έστω X_n (n ένας μη-αρνητικός ακέραιος) μία μη-αρνητική, πεπερασμένη τ.μ. που παίρνει ακέραιες τιμές και $\ell_n = \sup\{x : \Pr(X_n = x) > 0\}$ το ανώτερο σημείο της.

Ορισμός 1. Η τ.μ. X_n θα ονομάζεται τυχαία μεταβλητή διωνυμικού τύπου εμφυτεύσιμη σε Μαρκοβιανή αλυσίδα (MVB) αν

(a) υπάρχει Μαρκοβιανή αλυσίδα $\{Y_t, t \geq 0\}$ ορισμένη σε ένα διακριτό χώρο καταστάσεων Ω , ο οποίος μπορεί να διαμεριστεί ως εξής

$$\Omega = \bigcup_{x \geq 0} C_x, \quad C_x = \{c_{x0}, c_{x1}, \dots, c_{x,s-1}\},$$

(b) $\Pr(Y_t \in C_y | Y_{t-1} \in C_x) = 0$, για όλα τα $y \neq x, x+1$ και $t \geq 1$,

(c) το ενδεχόμενο $(X_n = x)$ είναι ισοδύναμο με το $(Y_n \in C_x)$, δηλ.

$$\Pr(X_n = x) = \Pr(Y_n \in C_x), \quad n \geq 0, x \geq 0.$$

Όπως προκύπτει από τη συνθήκη (b) του Ορισμού 1 οι μόνες εφικτές μεταβάσεις για την $\{Y_t, t \geq 0\}$ είναι αυτές που διεξάγονται μέσα στον ίδιο χώρο υποκαταστάσεων C_x ή από την υποκατάσταση C_x στη C_{x+1} . Έτσι, γεννιούνται οι επόμενοι δύο $s \times s$ πίνακες πιθανοτήτων μετάβασης

$$\mathbf{A}_t(x) = (\Pr(Y_t \in c_{xj} | Y_{t-1} \in c_{xi})), \quad \mathbf{B}_t(x) = (\Pr(Y_t \in c_{x+1,j} | Y_{t-1} \in c_{xi})).$$

Συμβολίζουμε με $\mathbf{f}_t(x)$ τα διανύσματα πιθανοτήτων

$$\mathbf{f}_t(x) = (\Pr(Y_t \in c_{x0}), \Pr(Y_t \in c_{x1}), \dots, \Pr(Y_t \in c_{x,s-1})), \quad 0 \leq t \leq n,$$

και με

$$\boldsymbol{\pi}_x = (\Pr(Y_0 \in c_{x0}), \Pr(Y_0 \in c_{x1}), \dots, \Pr(Y_0 \in c_{x,s-1})), \quad x \geq 0,$$

τις αρχικές πιθανότητες της Μαρκοβιανής αλυσίδας. Στις περισσότερες εφαρμογές στις οποίες χρησιμοποιούνται MVB ισχύει ότι $Pr(X_0 = 0) = 1$. Ως άμεση συνέπεια πηγάζει ότι

$$\boldsymbol{\pi}_0 \mathbf{1}' = (\Pr(Y_0 \in c_{00}), \Pr(Y_0 \in c_{01}), \dots, \Pr(Y_0 \in c_{0,s-1})) \mathbf{1}' = 1,$$

και $\boldsymbol{\pi}_x \mathbf{1}' = 0$ για $1 \leq x \leq \ell_n$ ($\mathbf{1} = (1, 1, \dots, 1)$ είναι το διάνυσμα γραμμή του R^s το οποίο έχει όλα τα στοιχεία του ίσα με 1).

Το ακόλουθο λήμμα (Koutras and Alexandrou (1995)) παρέχει αναδρομικές σχέσεις για τα διανύσματα πιθανότητας $\mathbf{f}_t(x)$.

Λήμμα 1. Έστω ότι η X_n είναι MVB . Η ακολουθία $\mathbf{f}_t(x)$, $t = 1, 2, \dots, n$, ικανοποιεί τις αναδρομικές σχέσεις

$$\mathbf{f}_t(0) = \mathbf{f}_{t-1}(0) \mathbf{A}_t(0)$$

$$\mathbf{f}_t(x) = \mathbf{f}_{t-1}(x) \mathbf{A}_t(x) + \mathbf{f}_{t-1}(x-1) \mathbf{B}_t(x-1), \quad 1 \leq x \leq \ell_n,$$

με αρχικές συνθήκες $\mathbf{f}_0(x) = \boldsymbol{\pi}_x$, $0 \leq x \leq \ell_n$.

Είναι προφανές ότι η συνάρτηση πιθανότητας $f_n(x)$ της X_n μπορεί να εκφραστεί ως

$$f_n(x) = \mathbf{f}_n(x) \mathbf{1}'.$$

3. ΚΑΤΑΝΟΜΕΣ ΔΙΩΝΥΜΙΚΟΥ ΤΥΠΟΥ

Σε αυτή την ενότητα θα υπολογίσουμε τις συναρτήσεις πιθανότητας των τ.μ. $N_{n,k,r}^{(i)}$, $i = 1, 2$. Πριν όμως προχωρήσουμε στην εξαγωγή των νέων αποτελεσμάτων, θα δώσουμε ένα απλό παράδειγμα για να διευκρινήσουμε περαιτέρω τις τ.μ. που εισήχθησαν νωρίτερα. Υποθέτουμε ότι $k = 3$ και $r = 2$. Στην περίπτωση αυτή ο σύνθετος σχηματισμός E_1 μπορεί να παρασταθεί ως εξής

$$E_1 = E_1^{(1)} \cup E_1^{(2)},$$

όπου

$$E_1^{(1)} = \{\underbrace{SF\dots FS}_{d_1 \geq 2} \underbrace{SF\dots FS}_{d_2 \geq 0}\}, \quad E_1^{(2)} = \{\underbrace{SF\dots FS}_{d_1 \geq 0} \underbrace{SF\dots FS}_{d_2 \geq 2}\}.$$

Ομοίως

$$E_2 = E_2^{(1)} \cup E_2^{(2)},$$

όπου

$$E_2^{(1)} = \{\underbrace{SF\dots FS}_{d_1 \geq 2} \underbrace{SF\dots FS}_{d_2 \geq 0} \underbrace{SS\dots S}_{\geq 0}\}, \quad E_2^{(2)} = \{\underbrace{SF\dots FS}_{d_1 \geq 0} \underbrace{SF\dots FS}_{d_2 \geq 2} \underbrace{SS\dots S}_{\geq 0}\}.$$

Στην παρακάτω ακολουθία από $n = 15$ δίτιμες δοκιμές: $SSFSFFSSFFSSSS$, μπορεί κάποιος εύκολα να διαπιστώσει ότι

$$N_{15,3,2}^{(1)} = 2, N_{15,3,2}^{(2)} = 1.$$

Για διευκόλυνση του αναγνώστη να επισημάνουμε ότι για την τ.μ. $N_{15,3,2}^{(1)}$ έχουν ανιχνευτεί οι σχηματισμοί $SFSFFFS$ και $SFFSS$, ενώ για την τ.μ. $N_{15,3,2}^{(2)}$ ο σχηματισμός $SFSFFFS$.

3.1 Η κατανομή της $N_{n,k,r}^{(1)}$

Το επόμενο θεώρημα δίνει τη συνάρτηση πιθανότητας της τ.μ. $N_{n,k,r}^{(1)}$.

Θεώρημα 1. Η συνάρτηση πιθανότητας της τ.μ. $N_{n,k,r}^{(1)}$ ($r \geq 1$) μπορεί να εκφραστεί ως εξής

$$P(N_{n,k,r}^{(1)} = x) = \mathbf{f}_n(x) \mathbf{1}', \quad x = 0, 1, \dots, \left\lfloor \frac{n}{k+r} \right\rfloor, \quad (1)$$

όπου $\mathbf{f}_t(x)$ είναι τα διανύσματα πιθανότητας που ικανοποιούν τις αναδρομικές σχέσεις του Λήμματος 1, $\mathbf{1} = (1, 1, \dots, 1) \in R^s$, \mathbf{A} είναι ένας πίνακας $((k-1) \times r + k) \times ((k-1) \times r + k)$ που έχει όλα του τα στοιχεία 0 εκτός από τα:

- $(1, 1)$, το οποίο ισούται με q ,
- $(2 + i, k + 1 + \max\{1, r - 1\}i)$, $i = 0, 1, \dots, k - 2$, που είναι όλα ίσα με q ,
- $(k + 1 + (r - 1)j + i, k + 2 + (r - 1)j + i)$, για $r \geq 3$, $i = 0, \dots, \max\{0, r - 3\}$, $j = 0, \dots, k - 2$, που είναι όλα ίσα με q ,
- $(k + r - 1 + (r - 1)i, (k - 1)r + 2 + i)$, για $r \geq 2$, $i = 0, \dots, k - 2$, που είναι όλα ίσα με q ,
- $((k - 1)r + 1 + i, (k - 1)r + 1 + i)$, $i = 1, \dots, k - 1$, που είναι όλα ίσα με q ,
- $(i, i + 1)$, $i = 1, \dots, k - 1$, που είναι όλα ίσα με p ,
- (k, k) το οποίο ισούται με p ,
- $(k + (r - 1)j + i, 3 + j)$ για $r \geq 2$, $j = 0, \dots, \max\{0, k - 3\}$, $i = 1, \dots, r - 1$, που είναι όλα ίσα με p ,
- $(k + (k - 2)(r - 1) + i, k)$ για $r \geq 2$ και $k \geq 3$, $i = 1, \dots, r - 1$, που είναι όλα ίσα με p ,
- $((k - 1)r + 1 + i, (k - 1)r + 2 + i)$ για $k \geq 3$, $i = 1, \dots, k - 2$, που είναι όλα ίσα με p ,

και ο \mathbf{B} είναι ένας πίνακας $((k - 1) \times r + k) \times ((k - 1) \times r + k)$ με όλα τα στοιχεία του μηδενικά εκτός από την πρώτη στήλη του η οποία ισούται με $\mathbf{1}' - \mathbf{A}\mathbf{1}'$.

Απόδειξη. Σύμφωνα με το Λήμμα 1 αρκεί να δείξουμε ότι η τ.μ. $N_{n,k,r}^{(1)}$ είναι MVB .

Θέτουμε $\ell_n = \lfloor n/(k+r) \rfloor$ και εισάγουμε το χώρο καταστάσεων $\Omega = \bigcup_{x=0}^{\ell_n} C_x$, όπου C_x , $x = 0, 1, \dots, \ell_n$ είναι ξένοι μεταξύ τους υπόχωροι με $|C_x| = (k-1)r + k$,

$x = 0, 1, \dots, \ell_n$, στοιχεία, οι οποίοι παρίστανται ως εξής

$$C_x = \{c_{x0}, c_{x1}, \dots, c_{x,k-1}, c_{x1^{(1)}}, c_{x1^{(2)}}, \dots, c_{x1^{(r-1)}}, c_{x2^{(1)}}, c_{x2^{(2)}}, \dots, c_{x2^{(r-1)}}, \\ c_{x,(k-1)^{(1)}}, \dots, c_{x,(k-1)^{(r-1)}}, c_{xL_1}, c_{xL_2}, \dots, c_{xL_{k-1}}\}.$$

Στη συνέχεια εισάγουμε μια Μαρκοβιανή αλυσίδα $\{Y_t, t \geq 0\}$ στον Ω ως εξής:

$Y_t \in c_{x,i} = \{(x, i)\}$, ή ισοδύναμα $Y_t = (x, i)$, αν στα πρώτα t αποτελέσματα Z_1, Z_2, \dots, Z_t , έχουν παρατηρηθεί x πραγματοποιήσεις σχηματισμών που περιέχονται στο E_1 και

(α) $i = 0$, αν

(1) $x = 0$ και $\prod_{j=1}^t (1 - Z_j) = 1$ (δεν έχει εμφανισθεί κανένα S μέχρι το t -οστό αποτέλεσμα) ή

(2) $x \geq 1$, $Z_t = 0$ και κανένα S δεν έχει εμφανιστεί μετά το k -οστό S της x -οστής πραγματοποίησης ενός απλού σχηματισμού που περιέχεται στο E_1 ή

(3) $x \geq 1$ και το t -οστό αποτέλεσμα είναι το k -οστό S ($Z_t = 1$) ενός απλού σχηματισμού που περιέχεται στο E_1 (δηλ. ένας απλός σχηματισμός ολοκληρώνεται στην t -οστή δοκιμή).

(β) $i = j$, $j = 1, 2, \dots, k - 1$, αν το t -οστό αποτέλεσμα είναι το τελευταίο S ($Z_t = 1$) μιας σειράς από j S μη διακοπτόμενης από τουλάχιστον r διαδοχικά F μεταξύ οποιωνδήποτε δύο συνεχόμενων S .

(γ) $i = j^{(d)}$, $1 \leq j \leq k - 1$, $1 \leq d \leq r - 1$, αν το t -οστό αποτέλεσμα είναι η d -οστή συνεχόμενη αποτυχία μιας ροής αποτυχιών (δηλ. $\prod_{i=0}^{d-1} (1 - Z_{t-i}) = 1$), που ακολουθεί μια σειρά από j S μη διακοπτόμενης από τουλάχιστον r διαδοχικά F μεταξύ οποιωνδήποτε δύο συνεχόμενων S .

(δ) $i = L_j$, $j = 1, 2, \dots, k - 1$, εάν ένας νέος απλός σχηματισμός που περιέχεται στο E_1 θα ολοκληρωθεί στις επόμενες $k - i$ S (όποτε συμβούν).

Με αυτή την προεργασία η τ.μ. $N_{n,k,r}^{(1)}$ γίνεται MVB με διάνυσμα αρχικής πιθανότητας

$$\pi_0 = (1, 0, 0, \dots, 0)_{1 \times ((k-1) \times r + k)},$$

και πίνακες \mathbf{A} και \mathbf{B} με στοιχεία όπως περιγράφονται στο θεώρημα.

Το αποτέλεσμα προκύπτει άμεσα από το Λήμμα 1 και τον Ορισμό 1.

Για την καλύτερη κατανόηση του Θεωρήματος 1, θα δώσουμε ένα παράδειγμα για την ειδική περίπτωση $k = 3$, $r = 2$ και την ακολουθία $n = 18$ δίτιμων αποτελεσμάτων: $Z_1 = 0$, $Z_2 = Z_3 = Z_4 = 1$, $Z_5 = 0$, $Z_6 = 1$, $Z_7 = Z_8 = 0$, $Z_9 = 1$, $Z_{10} = 0$, $Z_{11} = 1$, $Z_{12} = Z_{13} = Z_{14} = 0$, $Z_{15} = 1$, $Z_{16} = 0$, $Z_{17} = Z_{18} = 1$. Σε αυτή την περίπτωση προκύπτει η ακόλουθη εμφυτευμένη Μαρκοβιανή αλυσίδα: $Y_1 = (0, 0)$, $Y_2 = (0, 1)$, $Y_3 = (0, 2)$, $Y_4 = (0, 2)$, $Y_5 = (0, 2^{(1)})$, $Y_6 = (0, 2)$, $Y_7 = (0, 2^{(1)})$, $Y_8 = (0, L_2)$, $Y_9 = (1, 0)$, $Y_{10} = (1, 0)$, $Y_{11} = (1, 1)$, $Y_{12} = (1, 1^{(1)})$, $Y_{13} = (1, L_1)$, $Y_{14} = (1, L_1)$, $Y_{15} = (1, L_2)$, $Y_{16} = (1, L_2)$, $Y_{17} = (2, 0)$, $Y_{18} = (2, 1)$. Οι πίνακες \mathbf{A}

και \mathbf{B} της Μαρκοβιανής αλυσίδας είναι:

$$\mathbf{A} = \begin{pmatrix} (\cdot, 0) & (\cdot, 1) & (\cdot, 2) & (\cdot, 1^{(1)}) & (\cdot, 2^{(1)}) & (\cdot, L_1) & (\cdot, L_2) \\ q & p & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & p & q & 0 & 0 & 0 \\ 0 & 0 & p & 0 & q & 0 & 0 \\ 0 & 0 & p & 0 & 0 & q & 0 \\ 0 & 0 & p & 0 & 0 & 0 & q \\ 0 & 0 & 0 & 0 & 0 & q & p \\ 0 & 0 & 0 & 0 & 0 & 0 & q \end{pmatrix}$$

$$\mathbf{B} = \begin{pmatrix} (\cdot, 0) & (\cdot, 1) & (\cdot, 2) & (\cdot, 1^{(1)}) & (\cdot, 2^{(1)}) & (\cdot, L_1) & (\cdot, L_2) \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Ο υπολογισμός των $P(N_{5,3,2}^{(2)} = x)$, $x = 0, 1$ μπορεί να πραγματοποιηθεί εύκολα εφαρμόζοντας το Λήμμα 1 ως εξής:

$$\begin{aligned} \mathbf{f}_0(0) &= (1, 0, 0, 0, 0, 0, 0), \\ \mathbf{f}_1(0) &= \mathbf{f}_0(0) \cdot \mathbf{A} = (q, p, 0, 0, 0, 0, 0), \\ \mathbf{f}_2(0) &= \mathbf{f}_1(0) \cdot \mathbf{A} = (q^2, pq, p^2, pq, 0, 0, 0), \\ \mathbf{f}_3(0) &= \mathbf{f}_2(0) \cdot \mathbf{A} = (q^3, pq^2, p^3 + 2p^2q, pq^2, p^2q, pq^2, 0), \\ \mathbf{f}_4(0) &= \mathbf{f}_3(0) \cdot \mathbf{A} = (q^4, pq^3, p^4 + 3p^3q + 2p^2q^2, pq^3, p^3q + 2p^2q^2, 2pq^3, 2p^2q^2), \\ \mathbf{f}_5(0) &= \mathbf{f}_4(0) \cdot \mathbf{A} = (q^5, pq^4, p^5 + 4p^4q + 4p^3q^2 + 2p^2q^3, pq^4, p^4q + 3p^3q^2 + 2p^2q^3, \\ &\quad 3pq^4, p^3q^2 + 6p^2q^3), \\ \mathbf{f}_1(1) &= \mathbf{f}_0(1) \cdot \mathbf{A} + \mathbf{f}_0(0) \cdot \mathbf{B} = (0, 0, 0, 0, 0, 0, 0), \\ \mathbf{f}_2(1) &= \mathbf{f}_1(1) \cdot \mathbf{A} + \mathbf{f}_1(0) \cdot \mathbf{B} = (0, 0, 0, 0, 0, 0, 0), \\ \mathbf{f}_3(1) &= \mathbf{f}_2(1) \cdot \mathbf{A} + \mathbf{f}_2(0) \cdot \mathbf{B} = (0, 0, 0, 0, 0, 0, 0), \\ \mathbf{f}_4(1) &= \mathbf{f}_3(1) \cdot \mathbf{A} + \mathbf{f}_3(0) \cdot \mathbf{B} = (0, 0, 0, 0, 0, 0, 0), \\ \mathbf{f}_5(1) &= \mathbf{f}_4(1) \cdot \mathbf{A} + \mathbf{f}_4(0) \cdot \mathbf{B} = (2p^3q^2, 0, 0, 0, 0, 0, 0). \end{aligned}$$

Συνεπώς,

$$\begin{aligned} P(N_{5,3,2}^{(2)} = 0) &= \mathbf{f}_5(0) \cdot \mathbf{1}' = p^5 + 5p^4q + 8p^3q^2 + 10p^2q^3 + 5pq^4 + q^5, \\ P(N_{5,3,2}^{(2)} = 1) &= \mathbf{f}_5(1) \cdot \mathbf{1}' = 2p^3q^2. \end{aligned}$$

Αξιίζει να σημειωθεί ότι, στην ειδική περίπτωση $k = 2$, η τ.μ. $N_{n,2,r-2}^{(2)}$ ανάγεται στην τ.μ. $N_{n,r}^{(3)}$, η οποία έχει μελετηθεί από τους Dafnis et al. (2012) και Sen and Goyal (2004).

3.2 Η κατανομή της $N_{n,k,r}^{(2)}$

Στην υποενότητα αυτή θα μελετήσουμε την κατανομή της τ.μ. $N_{n,k,r}^{(2)}$. Η προσέγγισή μας θα είναι παρόμοια με εκείνη που εφαρμόστηκε στην προηγούμενη υποενότητα, με μικρές διαφοροποιήσεις στον ορισμό των καταστάσεων της Μαρκοβιανής αλυσίδας.

Θεώρημα 2. Η συνάρτηση πιθανότητας της τ.μ. $N_{n,k,r}^{(2)}$ ($r \geq 1$) μπορεί να εκφραστεί ως εξής

$$P(N_{n,k,r}^{(4)} = x) = \mathbf{f}_n(x)\mathbf{1}', \quad x = 0, 1, \dots, \left\lfloor \frac{n+1}{k+r+1} \right\rfloor, \quad (2)$$

όπου $\mathbf{f}_t(x)$ είναι τα διανύσματα πιθανότητας που ικανοποιούν τις αναδρομικές σχέσεις του Λήμματος 1, $\mathbf{1} = (1, 1, \dots, 1) \in R^s$, \mathbf{A} είναι ένας πίνακας $((k-1) \times r + k + 1) \times ((k-1) \times r + k + 1)$ που έχει όλα τα στοιχεία του 0 εκτός από τα:

- $(1, 1)$ και $(k+1, 1)$ που είναι και τα δύο ίσα με q ,
- $(2+i, k+2+\max\{1, r-1\}i)$, $i = 0, 1, \dots, k-2$, που είναι όλα ίσα με q ,
- $(k+1+(r-1)j+1+i, k+3+(r-1)j+i)$, για $r \geq 3$, $i = 0, \dots, \max\{0, r-3\}$, $j = 0, \dots, k-2$, που είναι όλα ίσα με q ,
- $(k+r+(r-1)i, (k-1)r+3+i)$, για $r \geq 2$, $i = 0, \dots, k-2$, που είναι όλα ίσα με q ,
- $((k-1)r+2+i, (k-1)r+2+i)$, $i = 1, \dots, k-1$, που είναι όλα ίσα με q ,
- $(i, i+1)$, $i = 1, \dots, k-1$, που είναι όλα ίσα με p ,
- (i, i) , $i = k, k+1$, που είναι όλα ίσα με p ,
- $(k+1+(r-1)j+i, 3+j)$, για $r \geq 2$, $j = 0, \dots, \max\{0, k-3\}$, $i = 1, \dots, r-1$, που είναι όλα ίσα με p ,
- $(k+1+(k-2)(r-1)+i, k)$, για $r \geq 2$ και $k \geq 3$, $i = 1, \dots, r-1$, που είναι όλα ίσα με p ,
- $((k-1)r+2+i, (k-1)r+3+i)$, για $k \geq 3$, $i = 1, \dots, k-2$, που είναι όλα ίσα με p .

και ο \mathbf{B} είναι ένας πίνακας $((k-1) \times r + k + 1) \times ((k-1) \times r + k + 1)$ με όλα τα στοιχεία του μηδενικά εκτός από την $k+1$ στήλη του η οποία ισούται με $\mathbf{1}' - \mathbf{A}\mathbf{1}'$.

Απόδειξη. Εισάγουμε μια Μαρκοβιανή αλυσίδα $\{Y_t, t \geq 0\}$ στον $\Omega = \bigcup_{x=0}^{\ell_n} C_x$ με $\ell_n = \left\lfloor \frac{n+1}{k+r+1} \right\rfloor$ και

$$C_x = \{c_{x0}, c_{x1}, \dots, c_{xk}, c_{x1(1)}, c_{x1(2)}, \dots, c_{x1(r-1)}, c_{x2(1)}, c_{x2(2)}, \dots, c_{x2(r-1)}, \\ c_{x,(k-1)(1)}, \dots, c_{x,(k-1)(r-1)}, c_{xL_1}, c_{xL_2}, \dots, c_{xL_{k-1}}\}, \quad x = 0, 1, \dots, \ell_n.$$

Ο πληθάρθρωτος του C_x ισούται με $|C_x| = (k-1)r + k + 1$, ενώ $Y_t \in c_{x,i} = \{(x,i)\}$, ή ισοδύναμα $Y_t = (x,i)$, αν στα πρώτα t αποτελέσματα Z_1, Z_2, \dots, Z_t , x πραγματοποιηθείσες σχηματισμών που περιέχονται στο E_2 έχουν παρατηρηθεί και
(α) $i = 0$, αν

(1) $x = 0$ και $\prod_{j=1}^t (1 - Z_j) = 1$ (δεν έχει εμφανισθεί κανένα S μέχρι το t -οστό αποτέλεσμα) ή

(2) $x \geq 1$, $Z_t = 0$ και κανένα S δεν έχει εμφανιστεί μετά το k -οστό S της x -οστής πραγματοποίησης ενός απλού σχηματισμού που περιέχεται στο E_2 ή

(β) $i = j$, $j = 1, 2, \dots, k-1$, αν το t -οστό αποτέλεσμα είναι το τελευταίο S ($Z_t = 1$) μιας σειράς από j S μη διακοπτόμενης από τουλάχιστον r διαδοχικά F μεταξύ οποιωνδήποτε δύο συνεχόμενων S .

(γ) $i = k$, αν το τελευταίο αποτέλεσμα είναι το τελευταίο S μια σειράς από τουλάχιστον k S , διακοπτόμενης από τουλάχιστον r διαδοχικά F μεταξύ οποιωνδήποτε δύο από τα πρώτα k συνεχόμενα S .

(δ) $i = j^{(d)}$, $1 \leq j \leq k-1$, $1 \leq d \leq r-1$, αν το t -οστό αποτέλεσμα είναι η d -οστή συνεχόμενη αποτυχία μιας ροής αποτυχιών (δηλ. $\prod_{i=0}^{d-1} (1 - Z_{t-i}) = 1$), που ακολουθεί μια σειρά από j S μη διακοπτόμενης από τουλάχιστον r διαδοχικά F μεταξύ οποιωνδήποτε δύο συνεχόμενων S .

(ε) $i = L_j$, $j = 1, 2, \dots, k-1$, εάν ένας νέος απλός σχηματισμός που περιέχεται στο E_2 θα ολοκληρωθεί στις επόμενες $k-i$ S (όποτε συμβούν).

Με αυτή την προεργασία η τ.μ. $N_{n,k,r}^{(2)}$ γίνεται MVB με διάνυσμα αρχικής πιθανότητας

$$\pi_0 = (1, 0, 0, \dots, 0)_{1 \times ((k-1) \times r + k + 1)},$$

και πίνακες \mathbf{A} και \mathbf{B} με τα στοιχεία που περιγράφονται στο θεώρημα. Το αποτέλεσμα προκύπτει άμεσα χρησιμοποιώντας το Λήμμα 1 και τον Ορισμό 1.

Θεωρούμε ξανά την ειδική περίπτωση $k = 3$, $r = 2$. Για την επόμενη ακολουθία $n = 18$ δίτιμων αποτελεσμάτων, $Z_1 = 0$, $Z_2 = Z_3 = Z_4 = 1$, $Z_5 = 0$, $Z_6 = 1$, $Z_7 = Z_8 = 0$, $Z_9 = Z_{10} = 1$, $Z_{11} = 0$, $Z_{12} = 1$, $Z_{13} = Z_{14} = Z_{15} = 0$, $Z_{16} = Z_{17} = 1$, $Z_{18} = 0$, οι καταστάσεις της εμφυτευμένης Μαρκοβιανής αλυσίδας είναι $Y_1 = (0, 0)$, $Y_2 = (0, 1)$, $Y_3 = (0, 2)$, $Y_4 = (0, 2)$, $Y_5 = (0, 2^{(1)})$, $Y_6 = (0, 2)$, $Y_7 = (0, 2^{(1)})$, $Y_8 = (0, L_2)$, $Y_9 = (1, 3)$, $Y_{10} = (1, 3)$, $Y_{11} = (1, 0)$, $Y_{12} = (1, 1)$, $Y_{13} = (1, 1^{(1)})$, $Y_{14} = (1, L_1)$, $Y_{15} = (1, L_1)$, $Y_{16} = (1, L_2)$, $Y_{17} = (2, 3)$, $Y_{18} = (2, 0)$. Οι πίνακες

που προκύπτουν είναι:

$$\mathbf{A} = \begin{pmatrix} (\cdot, 0) & (\cdot, 1) & (\cdot, 2) & (\cdot, 3) & (\cdot, 1^{(1)}) & (\cdot, 2^{(1)}) & (\cdot, L_1) & (\cdot, L_2) \\ q & p & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & p & 0 & q & 0 & 0 & 0 \\ 0 & 0 & p & 0 & 0 & q & 0 & 0 \\ q & 0 & 0 & p & 0 & 0 & 0 & 0 \\ 0 & 0 & p & 0 & 0 & 0 & q & 0 \\ 0 & 0 & p & 0 & 0 & 0 & 0 & q \\ 0 & 0 & 0 & 0 & 0 & 0 & q & p \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & q \end{pmatrix}$$

$$\mathbf{B} = \begin{pmatrix} (\cdot, 0) & (\cdot, 1) & (\cdot, 2) & (\cdot, 3) & (\cdot, 1^{(1)}) & (\cdot, 2^{(1)}) & (\cdot, L_1) & (\cdot, L_2) \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & p & 0 & 0 & 0 & 0 \end{pmatrix}$$

και εφαρμόζοντας παρόμοια υπολογιστική διαδικασία με εκείνη του παραδείγματος της προηγούμενης υποενότητας, μπορούμε να υπολογίσουμε εύκολα τις $P(N_{5,3,2}^{(2)} = x), x = 0, 1$.

4. ΕΦΑΡΜΟΓΕΣ

Οι Dafnis and Philippou (2011) υπέδειξαν ότι ο χρόνος αναμονής για την πραγματοποίηση του σχηματισμού E_1 , στην ειδική περίπτωση $k = 2$, μπορεί να χρησιμοποιηθεί ως κριτήριο απόφασης για τον απογαλακτισμό (weaning) ασθενούς από μηχανικό αερισμό (Esteban et al. (1994)) και να ενισχύσει ένα σύστημα που είναι ακόμα υπό ανάπτυξη (Buliev et al. (2002), Dermitzakis et al. (2008)). Είναι άμεση η σκέψη ότι ο γενικευμένος σχηματισμός E_1 , ο οποίος εισήχθη στην παρούσα εργασία, μπορεί να λειτουργήσει ως γενικευμένο, πιο αποτελεσματικό, κριτήριο απόφασης.

Ακολουθώντας, λοιπόν, τους Dafnis and Philippou (2011) εστιάζουμε σε μία κάρδιο-κυκλοφοριακή παράμετρο ενός ασθενούς και υποθέτουμε ότι συλλέγουμε τιμές της κάθε 20 δευτερόλεπτα. Σε κάθε τιμή που βρίσκεται σε ένα προκαθορισμένο εύρος τιμών αντιστοιχίζεται η τιμή 1, ενώ σε αντίθετη περίπτωση αντιστοιχίζεται η τιμή 0. Οι απαριθμητρίες τ.μ. $N_{n,k,r}^{(i)}, i = 1, 2$ των σχηματισμών E_1 και E_2 μπορούν να αποτελέσουν βελτιωμένα κριτήρια απόφασης που θα ενισχύσουν το προαναφερόμενο

Πίνακας 1: Η κατανομή της $N_{90,7,4}^{(1)}$

x	$\Pr(N_{90,7,4}^{(2)} = x)$				
	$p = 0.1$	$p = 0.3$	$p = 0.5$	$p = 0.7$	$p = 0.9$
0	0.192501	0.000036	0.0458621	0.611470	0.992386
1	0.744855	0.003765	0.190671	0.312359	0.007589
2	0.0624958	0.156752	0.334154	0.067787	0.000025
3	0.000148	0.655763	0.295348	0.007867	*
4	*	0.181984	0.120350	0.000502	*
5	*	0.001699	0.013518	0.000016	*
6	*	*	0.000098	*	*
7	*	*	*	*	*
8	*	*	*	*	*
$E(N_{90,7,4}^{(2)})$	0.870292	3.020990	2.294600	0.473619	0.007639
$Var(N_{90,7,4}^{(2)})$	0.238765	0.36048	1.187310	0.438418	0.007631

* : Τιμές μικρότερες από 10^{-6}

Πίνακας 2: Η κατανομή της $N_{90,7,4}^{(2)}$

x	$\Pr(N_{90,7,4}^{(4)} = x)$				
	$p = 0.1$	$p = 0.3$	$p = 0.5$	$p = 0.7$	$p = 0.9$
0	0.192501	0.000036	0.0458621	0.611470	0.992386
1	0.750021	0.004588	0.204243	0.317884	0.007594
2	0.057379	0.206825	0.362412	0.064261	0.000019
3	0.000099	0.683235	0.295470	0.006123	*
4	*	0.105068	0.088434	0.000259	*
5	*	0.000247	0.003577	0.000003	*
6	*	*	0.000002	*	*
7	*	*	*	*	*
$E(N_{90,7,4}^{(4)})$	0.865077	2.889450	2.187110	0.465828	0.007633
$Var(N_{90,7,4}^{(4)})$	0.232073	0.319340	1.034110	0.417272	0.007614

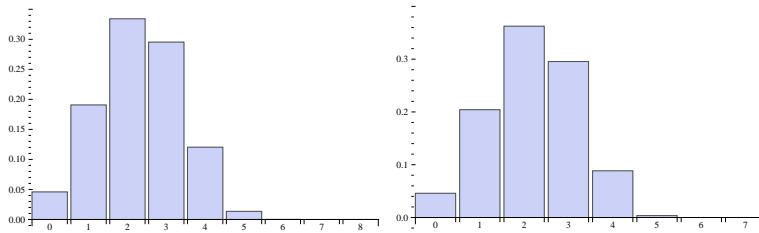
* : Τιμές μικρότερες από 10^{-6}

σύστημα (για περισσότερη πληροφορία σχετικά με τη λογική της επιλογής κριτηρίων απόφασης αυτού του είδους και, γενικότερα, για το πρόβλημα του απογαλακτισμού από μηχανικό αερισμό, ο ενδιαφερόμενος αναγνώστης παραπέμπεται στην εργασία Dafnis and Philippou (2011) και στις αναφορές που παρατίθενται).

Φυσικά, ο έλεγχος των προτεινόμενων κριτηρίων απαιτεί την εφαρμογή και έλεγχό τους σε πραγματικά κλινικά δεδομένα, διαδικασία που αποτελεί και επόμενο στόχο των συγγραφέων. Ωστόσο προκειμένου να δώσουμε κάποια αριθμητικά αποτελέσματα, θα θεωρήσουμε την ειδική περίπτωση $k = 7, r = 4, n = 90$ (σε 30 λεπτά ή 90 μετρήσεις ενδιαφερόμαστε για σχηματισμούς από (τουλάχιστον) 7 επιτυχίες (1) που διακόπτονται από τουλάχιστον 4 συνεχόμενες αποτυχίες (0) σε κάποιο σημείο τους). Χρησιμοποιώντας τα Θεωρήματα 1 και 2 υπολογίζουμε τα στοιχεία των Πινάκων 1 και 2, αντίστοιχα, τα οποία δίνουν την κατανομή, τη μέση τιμή και τη διασπορά των τ.μ. $N_{90,7,4}^{(i)}$, $i = 1, 2$ για $p = 0.1, 0.3, 0.5, 0.7$ και 0.9 .

Το Σχήμα 1 απεικονίζει τη συνάρτηση πιθανότητας των τ.μ. $N_{90,7,4}^{(i)}$, $i = 1, 2$ για $p = 0.5$.

Σχήμα 1: Η συνάρτηση πιθανότητας της $N_{90,7,4}^{(1)}$ (αριστερά) και της $N_{90,7,4}^{(2)}$ (δεξιά) για $p = 0.5$



Όλα τα αριθμητικά αποτελέσματα συνηγορούν υπέρ της ύπαρξης μοναδικής κορυφής (ότι οι νέες κατανομές είναι μονοκόρυφες).

5. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΗ ΕΡΕΥΝΑ

Στην παρούσα εργασία εισαγάγαμε και μελετήσαμε δύο νέες κατανομές διωνυμικού τύπου. Προτείνουμε πως οι νέες κατανομές μπορούν να χρησιμοποιηθούν για τη βελτίωση της αποδοτικότητας ενός, υπό έρευνα, συστήματος σχετιζόμενου με το μηχανικό αερισμό.

Τα νέα αποτελέσματα μπορούν να γενικευτούν στην περίπτωση δοκιμών που παρουσιάζουν Μαρκοβιανή εξάρτηση, ή δοκιμών που μπορούν να λάβουν περισσότερες από δύο τιμές. Τέλος, ένα ανοιχτό ερευνητικό θέμα είναι η θεωρητική απόδειξη της μονοκορυφότητας των νέων κατανομών.

ABSTRACT

In the present paper we introduce and study distributions of random variables enumerating patterns in sequences of independent and identically distributed binary trials, which generalize or extend known distributions extensively examined in the literature during the last decades. We employ the Markov chain imbedding technique to obtain exact formulas for the probability mass functions of the new distributions. Finally, we suggest an applications of our results in biomedical engineering.

Ευχαριστίες: Η παρούσα έρευνα έχει χρηματοδοτηθεί από τη Γενική Γραμματεία Έρευνας και Τεχνολογίας μέσω των κονδυλίων της Εθνικής Συμμετοχής 2016-2017 που σχετίζονται με το Ευρωπαϊκό Έργο “ISMPH: Inference for a Semi-Markov Process using Hazards Specification” (GA No 329128).

ΑΝΑΦΟΡΕΣ

- Antzoulakos, D.L., Bersimis, S. and Koutras M.V. (2003). On the distribution of the total number of run lengths. *Annals of the Institute of Statistical Mathematics* **55**, 865-884.
- Balakrishnan, N. and Koutras, M. V. (2002). *Runs and Scans with Applications*, New York: John Wiley.
- Buliev, I., Pallikarakis N. and Filos K (2006). Computer-based system for automated weaning of conscious patients from assisted mechanical ventilation: preliminary tests. *Annals of Technical University of Varna, Bulgaria*.
- Dafnis S.D., Makri F.S, and Philippou A.N. (2019). The reliability of a generalized consecutive system. *Applied Mathematics and Computation* **359**, 186-193.
- Dafnis S.D. and Philippou A.N. (2011). Distributions of patterns with applications in engineering. *IAENG International Journal of Applied Mathematics* **41**, 68-75.
- Dafnis S.D., Antzoulakos D.L. and Philippou A.N. (2010). Distributions related to (k_1, k_2) events. *Journal of Statistical Planning and Inference* **140**, 1691-1700.
- Dafnis S.D., Philippou A.N. and Antzoulakos D.L. (2012). Distributions of patterns of two successes separated by a string of $k - 2$ failures. *Statistical Papers* **53**, 323-344.
- Dermitzakis A., Buliev I. and Filos K. (2008). Fuzzy logic controller for weaning of conscious patients from mechanical ventilation: a simulation study. *6th European Symposium on Biomedical Engineering*, Chania, Greece, Code: s3.08
- Esteban A., Alia I., Ibanez J., Benito S. and Tobin MJ (1994). Modes of mechanical ventilation. A national survey of Spanish hospitals. *Chest* **106**, 1188-1193.
- Fu J.C. and Koutras M.V. (1994). Distribution theory of runs: a Markov chain approach. *Journal of the American Statistical Association* **89**, 1050- 1058.
- Fu J.C. and Lou W.Y.W. (2003). *Distribution theory of runs and patterns and its applications: a finite Markov imbedding approach*. Singapore: World Scientific Publishing.
- Koutras M.V. (2003). Applications of Markov chains to the distribution theory of runs and patterns. *Handbook of Statistics*, Amsterdam: Elsevier, 431-472.
- Koutras M.V. and Alexandrou V.A. (1995). Runs, scans and urn model distributions: a unified Markov chain approach. *Annals of the Institute of Statistical Mathematics* **47**, 743-766.
- Makri F.S. and Psillakis Z.M. (2012). Counting certain binary strings. *Journal of Statistical Planning and Inference* **142**, 908-924.
- Makri F.S. and Psillakis Z.M. (2013). Exact distributions of constrained (k, l) strings of failures between subsequent successes. *Statistical Papers* **54**, 783-806.
- Makri F.S. and Psillakis Z.M. (2014). On the expected number of limited length binary strings derived by certain urn models. *Journal of Probability* doi:10.1155/2014/646140.
- Makri F.S. and Psillakis Z.M. (2017). On limited length binary strings with an application in statistical control. *The Open Statistics and Probability Journal* **8**, 1-6.
- Sen K. and Goyal B. (2004) Distributions of patterns of two failures separated by success runs of length k . *Journal of the Korean Statistical Society* **33**, 35-58.



ΑΝΑΛΥΣΗ ΑΓΩΝΙΣΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΠΡΟΠΟΝΗΤΙΚΗ ΚΑΘΟΔΗΓΗΣΗ ΤΕΧΝΙΚΟΤΑΚΤΙΚΗΣ ΣΤΟ ΤΖΟΥΝΤΟ: ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ

Χ. Ζάραγκας

Παιδαγωγικό Τμήμα Νηπιαγωγών, Πανεπιστήμιο Ιωαννίνων
hzaragas@uoi.gr

ΠΕΡΙΛΗΨΗ

Σκοπός της παρούσας εργασίας είναι η παρατήρηση, η καταγραφή, η ανάλυση δεδομένων της αγωνιστικής δραστηριότητας των αντιπάλων αθλητριών και η διαμόρφωση τεχνικοτακτικής καθοδήγησης και εκπαίδευσης. Δημιουργήθηκε πρωτόκολλο για τα αγωνιστικά χαρακτηριστικά του τζούντο (στάση σώματος, λαβές, πιασίματα, είδος τεχνικής - τακτικής, μοχλοί χεριών, γοφού, και ποδιών επιθέσεις, άμυνες, συνδυασμοί επίθεσης, χρόνος επίθεσης, πλευρά παίξιματος, κ.ά.). Το πρωτόκολλο χρησιμοποιήθηκε επί δύο αγωνιστικά έτη σε πανελληνίους, σχολικούς και σε διεθνείς αγώνες τζούντο (βαλκανικοί) στους οποίους συμμετείχε η αθλήτρια και παράλληλα χρησιμοποιήθηκε για αγώνες της ίδιας κατηγορίας των 57 κιλών χωρίς τη συμμετοχή της, για παγκόσμια, πανευρωπαϊκά και Ολυμπιακούς Αγώνες. Πραγματοποιήθηκε ανάλυση συχνότητας περιεχομένου η οποία έδειξε τις τεχνικές ποδιών, γοφού και χεριών να εμφανίζονται με τη μεγαλύτερη συχνότητα στην επίτευξη της νίκης σε σχέση με τις τεχνικές τζούντο στο έδαφος. Τα αποτελέσματα έδειξαν στις τεχνικές εδάφους να εμφανίζονται περισσότερο οι τεχνικές ακινητοποίησης. Η στατιστική ανάλυση χ^2 φανέρωσε ότι υπάρχει σημαντική σχέση μεταξύ των τεχνικών επίτευξης της νίκης με το χρόνο εκτέλεσής τους, με τον τρόπο πιασίματος του αντιπάλου και με τη στάση του σώματος που χρησιμοποιεί ο αθλητής. Τα αποτελέσματα διαμόρφωσαν το προπονητικό – εκπαιδευτικό μοντέλο ανάπτυξης των τεχνικών και τακτικών δεξιοτήτων με σκοπό την επιτυχή αγωνιστική αντιμετώπιση των αντιπάλων αθλητριών. Το πρόγραμμα στέφθηκε με επιτυχία με την κατάκτηση της πρώτης θέσης της κατηγορία των 57 κιλών των νέων γυναικών για το 2018.

Λέξεις Κλειδιά: Τζούντο, Τεχνική – τακτική Ανάλυση, Προπονητική Καθοδήγηση

1. ΕΙΣΑΓΩΓΗ

Ο σκοπός της παρούσας έρευνας είναι: α) η μελέτη μέσω της ποσοτικής ανάλυσης δεδομένων της αγωνιστικής συμπεριφοράς (τεχνικές επίτευξης της νίκης, χρόνος επίτευξης της τεχνικής, τρόπος πιασίματος και στάση του σώματος) αθλητριών τζούντο της κατηγορίας των 57 κιλών των νέων γυναικών, τόσο σε πανελλήνιο όσο και σε διεθνές επίπεδο, καθώς επίσης και β) η διαμόρφωση τεχνικοτακτικής καθοδήγησης και εκπαίδευσης για την ανταγωνιστική επιτυχή έκβαση στο πανελλήνιο πρωτάθλημα. Ειδικά η μελέτη των τεχνικών επίτευξης της νίκης μιας πλειάδας αγώνων (πανελληνίους και άλλους) συγκεκριμένης αθλήτριας στην κατηγορία των 57 κιλών, καθώς επίσης και των τεχνικών επίτευξης της νίκης στην ίδια κατηγορία σε μεγάλες διοργανώσεις όπως Πανευρωπαϊκοί Αγώνες, (ΠΕ.Α.), Παγκόσμιοι Αγώνες, (Π.Α.), Ολυμπιακοί Αγώνες (Ο.Α.), στις οποίες η αθλήτρια μας δεν έλαβε μέρος.

Οι σημερινές συνθήκες συναγωνισμού στις αθλητικές διοργανώσεις, απαιτούν μία άκρως πιεστική και επίπονη προετοιμασία από τους αθλητές, ιδίως δε από εκείνους, οι οποίοι στοχεύουν σε υψηλές διακρίσεις. Εύλογο είναι λοιπόν, οι αθλητές να απαιτούν από τους προπονητές την ανάλογη αγωνιστική, τεχνική και τακτική γνώση, που από τη θέση τους σε μια αθλητική αγωνιστική διοργάνωση μπορεί να κρίνει την απόδοση του αθλητή.

Τεχνική είναι ένα ειδικό σύστημα, ένας τρόπος, μία ικανότητα που επιδεικνύεται από ταυτόχρονες και διαδοχικές σωματικές κινήσεις, κατευθυνόμενες προς την ορθή οργάνωση της αλληλοεπιδράσεως μεταξύ των εσωτερικών και εξωτερικών δυνάμεων, οι οποίες επιδρούν πάνω στον αθλητή /αθλήτρια, με σκοπό τη χρησιμοποίηση των τελευταίων για την επίτευξη της νίκης. Τακτική είναι όλες εκείνες οι προγραμματισμένες τεχνικές ενέργειες στα πλαίσια ενός ολοκληρωμένου σχεδίου (στρατηγικής σε άμυνα, επίθεση, αντεπίθεση), για την κατάκτηση της νίκης αξιοποιώντας τόσο τις φυσικές όσο και τεχνικές ικανότητες.

Η προπονητική και παιδαγωγική καθοδήγηση στηρίζεται τόσο στην επάρκεια γνώσεων του προπονητή όσο και στην στατιστική ανάλυση των τεχνικοτακτικών μεθόδων των αντιπάλων αθλητών αλλά και του σχετικού παγκόσμιου γίνεσθαι. Λίγες έρευνες έχουν γίνει και έχουν δημοσιευτεί για την ποσοτική ανάλυση των τεχνικοτακτικών αγωνιστικών μεθόδων στις πολεμικές τέχνες και ειδικά του τζούντο.

2. ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

Η τεχνική-τακτική ανάλυση είναι ένας τρόπος για να κατανοήσουμε τους παράγοντες που εξηγούν την επιτυχία σε αγώνες αθλητών υψηλού επιπέδου (MMA) (James et al., 2017·Miarka et al., 2016·Santos et al., 2018). Η μελέτη των Dal Belo et al., (2019) σύγκρινε και εξέτασε τις κινητικές δράσεις (τεχνική, τακτική) των αθλητών μικτών πολεμικών τεχνών (MMA) σε σχέση με τα αποτελέσματα (νίκη, νίκη στα σημεία, ήττα, νίκη με παράδοση, κ.ά.) που κατάφεραν στους αγώνες Ultimate Fighting Championship (UFC). Προς τούτο, αξιοποίησε ένα έγκυρο πρωτόκολλο τεχνικής, τακτικής ανάλυσης που είχε χρησιμοποιηθεί σε προηγούμενες μελέτες, και βασίζονταν σε πίνακες καταγραφής παρατηρήσεων των τεχνικών και της τακτικής (συνδυασμός τεχνικών) στοιχείων του αγώνα που σημειώνονταν κατά τη διάρκεια του αγώνα, καθώς επίσης και την ανάλυση απόδοσης στις μεικτές πολεμικές τέχνες (MMA). Τα αποτελέσματα δείχνουν μεγάλη εξειδίκευση στον τύπο της επιθετικής κατάστασης σύμφωνα με την τακτική για τον τερματισμό του αγώνα και τη νίκη. Αυτά τα αποτελέσματα δείχνουν επίσης ότι: α) η τακτική και οι τεχνικές αντιμετώπισης ποικίλλουν ανάλογα με τις δυνάμεις και τις αδυναμίες των αθλητών και μπορούν να χρησιμοποιηθούν από προπονητές και αθλητές για την ανάπτυξη ειδικών προγραμμάτων κατάρτισης, β) υπάρχει μεγάλη εξειδίκευση στον τύπο της επιθετικής τεχνικής, μία κατάσταση σύμφωνα με την τακτική για τον τερματισμό του αγώνα και τη νίκη, γ) η στρατηγική και οι τακτικές αντιμετώπισης ποικίλλουν ανάλογα με τις δυνάμεις και τις αδυναμίες των αθλητών και μπορούν να χρησιμοποιηθούν από προπονητές και αθλητές για την ανάπτυξη ειδικών προγραμμάτων κατάρτισης (Dal Belo et al., 2019).

Η μελέτη των Miarka et al., (2016) έδειξε ότι οι τεχνικές-τακτικές δράσεις διαφέρουν μεταξύ των νικητών και των ηττημένων, αλλά εξ όσων γνωρίζουμε ουδείς έχει έως σήμερα διερευνήσει συγκεκριμένες διαφορές τεχνικής-τακτικής δράσης και αποτελέσματα του αγώνα. Η τεχνική ικανότητα θεωρείται ένας παράγοντας υψίστης σημασίας για την απόδοση των αθλητών και το τζούντο περιέχει πληθώρα απλών, σύνθετων και πολύπλοκων τεχνικών. Η τακτική περιέχει το σύνολο των γνώσεων, των δεξιοτήτων και των ικανοτήτων ενός αγωνιζόμενου. Ο ρόλος της στον αγώνα είναι πολύ σπουδαίος διότι καλύπτονται οι αδυναμίες του αθλητή, με αποτέλεσμα να ξεπερνά τα προσόντα του αντιπάλου ανακαλύπτοντας και αξιοποιώντας τις αδυναμίες του.

Η έγκαιρη πρόγνωση των τεχνικοτακτικών πλάνων μέσα από αυτήν την εργασία, αποτελεί το «κλειδί» για την προαγωγή και ανάπτυξη του ανταγωνιστικού τομέα του αθλήματος, πρώτα από τους ίδιους τους προπονητές και το στυλ προπόνησης που επιβάλλουν, και ύστερα από τους αθλητές / τις αθλήτριες. Η ανάλυση και διάγνωση των τεχνικών-τακτικών δράσεων που έχουν σχέση με τις μεικτές πολεμικές τέχνες μπορεί να παράσχουν ουσιαστικές πληροφορίες στους προπονητές και στους αθλητές που μπορούν να χρησιμοποιηθούν για να βελτιώσουν την απόδοση κατά τη διάρκεια του ανταγωνισμού (Brandt et al., 2018 · Coswig et al., 2016 · Karpman et al., 2016 · Miarka et al., 2014). Η έρευνα πάνω στη χρήση βιντεοσκοπημένου υλικού έχει δείξει θετική αναπροσαρμογή στον προπονητή καθώς τον βοηθά να παρατηρεί με περισσότερη λεπτομέρεια τη τεχνική των αθλητών του αλλά και των αντιπάλων, δίνοντας με αυτόν τον τρόπο την απαραίτητη ανατροφοδότηση (Kajmovic et al., 2008 · Koropanovski, Jovanovic & Dopsaj, 2008 ·

Kajmovic, Radjo & Kapo, 2007 · Σφίγγου κ.ά., 2007 · Giovanni et al., 2007 · Pearce, 2005 · Lyons, 2005 · Αρματάς κ.ά., 2005 · Link & Lames, 2003 · Τσαμουρτζή, Σφίγγου και Τσιμπήρη, 2002).

Η έλλειψη ερευνών για την τεχνική και τακτική ανάλυση αγώνων τζούντο στον ελληνικό επιστημονικό χώρο μας ώθησε στη διεξαγωγή αυτής της έρευνας. Η παρούσα έρευνα μέσα από την καταγραφή και την ανάλυση όλων των τεχνικοτακτικών στοιχείων της επίθεσης και της άμυνας στο αγωνιστικό τζούντο, τη σύγκριση μεταξύ των αθλητριών, συμβάλλει σε καινούργιες τεχνικοτακτικές επιλογές, σε ουσιαστικό προγραμματισμό και αναβάθμιση της αγωνιστικής προπόνησης και γενικότερα σε νέες κατευθύνσεις και προσανατολισμούς με κύριο στόχο τη μεγιστοποίηση της απόδοσης.

3. ΥΠΟΘΕΣΕΙΣ ΤΗΣ ΕΡΕΥΝΑΣ

Οι υποθέσεις που αφορούν στα αποτελέσματα μπορούν να ταξινομηθούν ως εξής:

Υ₁) Η νίκη θα επιτυγχάνεται με διαφορετικές κατηγορίες τεχνικών της όρθιας πάλης (πόδια, χέρια, γοφός, αυτοθυσία, πλάγια αυτοθυσία) στους αγώνες της κατηγορίας των 57 κιλών των νέων γυναικών.

Υ₀₁) Η νίκη δεν θα επιτυγχάνεται με διαφορετικές κατηγορίες τεχνικών της όρθιας πάλης στους αγώνες της κατηγορίας των 57 κιλών των γυναικών.

Υ₂) Η νίκη θα επιτυγχάνεται με διαφορετικές κατηγορίες τεχνικών της πάλης εδάφους (ακινητοποίηση, πνιγμός, εξάρθρωση αγκώνα) στους αγώνες της κατηγορίας των 57 κιλών των νέων γυναικών.

Υ₀₂) Η νίκη δεν θα επιτυγχάνεται με διαφορετικές κατηγορίες τεχνικών της πάλης εδάφους στους αγώνες της κατηγορίας των 57 κιλών των νέων γυναικών.

Υ₃) Η νίκη θα επιτυγχάνεται σε διαφορετικά χρονικά διαστήματα του αγώνα της κατηγορίας των 57 κιλών των γυναικών (1°, 2°, 3°, 4° λεπτό, επιπλέον χρόνος) και αυτά θα σχετίζονται στατιστικώς σημαντικά με τις διαφορετικές κατηγορίες τεχνικών της όρθιας και πάλης εδάφους.

Υ₀₃) Η νίκη θα επιτυγχάνεται σε διαφορετικά χρονικά διαστήματα του αγώνα της κατηγορίας των 57 κιλών των νέων γυναικών και αυτά δεν θα σχετίζονται στατιστικώς σημαντικά με τις διαφορετικές κατηγορίες τεχνικών της όρθιας πάλης και πάλης εδάφους.

Υ₄) Η νίκη θα επιτυγχάνεται με διαφορετικούς τρόπους πιασίματος επί της αντιπάλου αθλήτριας στον αγώνα της κατηγορίας των 57 κιλών των νέων γυναικών και αυτοί θα σχετίζονται στατιστικώς σημαντικά με τις διαφορετικές κατηγορίες τεχνικών της όρθιας πάλης.

Υ₀₄) Η νίκη θα επιτυγχάνεται με διαφορετικούς τρόπους πιασίματος επί της αντιπάλου αθλήτριας στον αγώνα της κατηγορίας των 57 κιλών των νέων γυναικών και αυτοί δεν θα σχετίζονται στατιστικώς σημαντικά με τις διαφορετικές κατηγορίες τεχνικών της όρθιας πάλης.

Υ₅) Η νίκη θα επιτυγχάνεται με διαφορετικές στάσεις του σώματος της επιτιθέμενης αθλήτριας επί της αντιπάλου στον αγώνα της κατηγορίας των 57 κιλών των νέων γυναικών και αυτές θα σχετίζονται στατιστικώς σημαντικά με τις διαφορετικές κατηγορίες τεχνικών της όρθιας πάλης.

Υ₀₅) Η νίκη θα επιτυγχάνεται με διαφορετικές στάσεις του σώματος της επιτιθέμενης αθλήτριας επί της αντιπάλου στον αγώνα της κατηγορίας των 57 κιλών των νέων γυναικών και αυτές δεν θα σχετίζονται στατιστικώς σημαντικά με τις διαφορετικές κατηγορίες τεχνικών της όρθιας πάλης.

4. ΔΟΜΗΜΕΝΗ ΠΑΡΑΤΗΡΗΣΗ

Τα επιστημονικά συστατικά στοιχεία της παρατήρησης ως εμπειρικής τεχνικής συλλογής δεδομένων είναι εξασφαλισμένα όταν η παρατήρηση: α) εξυπηρετεί έναν συγκεκριμένο ερευνητικό σκοπό, β) σχεδιάζεται συστηματικά και δεν επαφίεται στην τύχη, γ) σχεδιάζεται συστηματικά και σχετίζεται με γενικές κρίσεις, όμως δεν αποκτά το χαρακτήρα μιας συλλογής ιδιοτροπιών ή ιδιορρυθμιών και δ) υποβάλλεται σε επαναλαμβανόμενες

εξετάσεις και ελέγχους όσον αφορά την ισχύ, την αξιοπιστία και την ακρίβεια (Schnell, Hill, Esser, 2014).

Το δείγμα βασίζεται αφενός στο υλικό που είχαμε πρόσβαση ως προπονητής της αθλήτριάς μας σε αγώνες της, αφετέρου στο γενικότερα προσπελάσιμο τηλεοπτικό υλικό των επισήμων αγώνων της ίδιας αγωνιστικής κατηγορίας (των 57 κιλών), επί των οποίων εφαρμόσαμε ακριβώς την ίδια προσέγγιση. Η λήψη επιτόπιων σημειώσεων σχετικά με τα στοιχεία τεχνικής και τακτικής αγωνιστικής στο τζούντο είναι μια σημαντική διαδικασία που ακολουθήθηκε στη μεικτή μέθοδο παρατήρησης (συμμετοχική, μη συμμετοχική) συλλογής και ανάλυσης πεδίου της παρούσας μελέτης (Bailey, 2007: 80). Η ανάλυση περιεχομένου υποβάλλει τις φωτογραφίες και τα βίντεο των αγώνων σε μια ποσοτικοποιημένη ανάλυση. Η ανάλυση συχνότητας των μονάδων καταμέτρησης που χρησιμοποιήθηκε κατηγοριοποίησε και καταμέτρησε τη συχνότητα εμφάνισής τους.

Πίνακας 1: πίνακας συχνοτήτων και αντίστοιχων ποσοστών του δείγματος των παρατηρούμενων αγώνων (διά ζώσης, βίντεο, διαδίκτυο) της συγκεκριμένης κατηγορίας των 57 κιλών των γυναικών.

Κατηγορίες Αγώνων	ΣΥΜΜΕΤΟΧΗ* ΑΘΛΗΤΡΙΑΣ ΜΑΣ		ΕΚΒΑΣΗ ΑΓΩΝΩΝ					
	f	%	ΝΙΚΗ		ΗΓΓΑ		ΣΥΝΟΛΟ	
			f	%	f	%	f	%
Διασυλλογικοί	12	40	34	47 28,6	9	19,1 7,6	43	36,2
Πανελλήνιοι	6	20	16	22 13,5	8	17 6,7	24	20,2
Διεθνείς	9	30	15	20,8 12,6	21	44,7 17,6	36	30,5
Βαλκανικοί	3	10	7	9,7 5,9	9	19,1 7,5	16	13,44
ΣΥΝΟΛΟ	30	100	72	60,5	47	39,5	119	100
ΜΗ ΣΥΜΜΕΤΟΧΗ ΑΘΛΗΤΡΙΑΣ ΜΑΣ ΣΕ ΑΓΩΝΕΣ**			ΝΙΚΗ		ΗΓΓΑ		ΣΥΝΟΛΟ	
Πανευρωπαϊκοί Αγώνες (ΠΕ.Α.)			48	32,5			48	32,5
Παγκόσμιοι Αγώνες (ΠΑ)			60	40,5			60	40,5
Ολυμπιακοί Αγώνες(ΟΑ)			40	27			40	27
ΣΥΝΟΛΟ			148	100			148	100
ΣΥΝΟΛΟ ΠΑΡΑΤΗΡΗΣΕΩΝ							267	100

Ο πίνακας (1) απεικονίζει το δείγμα παρατηρήσεων (267 αγώνες, 100%) της μελέτης και χωρίζεται σε δύο μέρη: α) το πάνω μέρος αριστερά αναφέρεται στις διοργανώσεις και το είδος αυτών που η αθλήτρια είχε συμμετοχή (*) ενώ στο δεξιό μέρος αναφέρεται το συνολικό πλήθος των αγώνων που δόθηκαν από αυτήν (119, περίπου 4 αγώνες ανά διοργάνωση), β) το κάτω μέρος αναφέρεται στις διεθνείς διοργανώσεις και σε 148 διεθνείς αγώνες που η αθλήτρια δεν συμμετείχε (**), αλλά λόγω της προπονητικής διαδικασίας πρέπει να ληφθούν υπόψη για την αγωνιστική της εκπαίδευση και εξέλιξη. Στην τελευταία γραμμή του πίνακα αριστερά διακρίνεται το σύνολο των παρατηρήσεων που εξετάζονται στη μελέτη για την προπονητική καθοδήγηση και αναφέρεται αθροιστικά σε 267 αγώνες (119 με συμμετοχή και 148 χωρίς τη συμμετοχή της αθλήτριας) της κατηγορίας των 57 κιλών.

5. ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

Η στατιστική ανάλυση πραγματοποιήθηκε με την χρήση του στατιστικού πακέτου SPSS (22) σε επίπεδο περιγραφικής και επαγωγικής στατιστικής. Αναφορικά με την αξιοποίηση στοιχείων της περιγραφικής στατιστικής, πραγματοποιήθηκε ανάλυση κατανομής συχνοτήτων εμφάνισης και σχετικών. Για τον έλεγχο των διαφορών μεταξύ των μεταβλητών επιλέχθηκαν μη παραμετρικά κριτήρια, λόγω του ότι οι εμπλεκόμενες μεταβλητές είναι κατηγορικές. Χρησιμοποιήθηκε ο έλεγχος chi-square (χ^2), η οποία είναι μια μη παραμετρική μέθοδος ανάλυσης που χρησιμοποιούμε όταν μετρούμε τη συχνότητα με την οποία τα άτομα εμφανίζονται σε κάθε κατηγορία μιας μεταβλητής. Είναι ένα τεστ ελέγχου, όσον αφορά τη

στατιστική σημαντικότητα της απόκλισης ανάμεσα στις παρατηρηθείσες τιμές (τις τιμές που συλλέξαμε) και τις αναμενόμενες τιμές (τις τιμές που θα αναμέναμε να συλλέξουμε). Στην έρευνά μας θέλουμε να διαπιστώσουμε αν το επίπεδο τεχνικής απόδοσης στο τζούντο σχετίζεται με τον τρόπο πιασίματος (Kumikata), το χρόνο εκτέλεσης της τεχνικής ή τη στάση του σώματος της αθλήτριας. Η αξιοπιστία των πινάκων παρατήρησης καταγραφής που χρησιμοποιήθηκαν στο πλαίσιο της παρούσας έρευνας ελέγχθηκε με το συντελεστή αξιοπιστίας Cronbach's Alpha, η συμφωνία μεταξύ εκτιμητών (προπονητών) ελέγχθηκε με το συντελεστή Kappa του Cohen, καθώς επίσης ελέγχθηκε και η ενδοβαθμολογική αξιοπιστία (Intra –Class – Correlation, ICC, Fischer, 1954). Η τελική βαθμολογία της έκβασης ενός αγώνα προκύπτει από τις παρατηρήσεις και τη βαθμολογία δύο εκτιμητών/ βαθμολογητών.

6. ΑΠΟΤΕΛΕΣΜΑΤΑ ΤΗΣ ΕΡΕΥΝΑΣ

Οι μεταβλητές της έρευνάς μας είναι κατηγορικές και ακολουθήθηκαν τεχνικές μη παραμετρικής στατιστικής ανάλυσης. Η παρουσίαση των αποτελεσμάτων κατηγοριοποιήθηκε σε τρεις διακριτές ενότητες, προκειμένου να καταστεί περισσότερο εύληπτη. Η πρώτη ενότητα παρουσίασης των αποτελεσμάτων (7.1) επικεντρώνεται στον έλεγχο της εγκυρότητας και της αξιοπιστίας των εργαλείων συλλογής δεδομένων που αξιοποιήθηκαν στην έρευνα, ενώ στο δεύτερο μέρος παρουσιάζονται οι αναλύσεις των μετρήσεων περιγραφικής (7.2) και επαγωγικής (7.2) στατιστικής που πραγματοποιήθηκαν.

6.1 Έλεγχος αξιοπιστίας, έλεγχος συμφωνίας μεταξύ Εκτιμητών και ενδοβαθμολογική αξιοπιστία αναφορικά με τις μεταβλητές της έρευνας

Η αξιοπιστία των μεταβλητών της έρευνας υπολογίστηκε με τον Δείκτη Εσωτερικής Συνέπειας «alpha του Cronbach» για το δείγμα των 267 (100%) παρατηρήσεων.

Πίνακας 2: Δείκτης Αξιοπιστίας (Cronbach's α) για τις μεταβλητές της έρευνας

ΜΕΤΑΒΛΗΤΕΣ ΤΗΣ ΕΡΕΥΝΑΣ	Cronbach's alpha Coefficient
ΧΡΟΝΟΣ ΕΠΙΤΕΥΞΗΣ ΝΙΚΗΣ	0,876
ΠΙΑΣΙΜΟ ΑΝΤΙΠΑΛΟΥ	0,869
ΣΤΑΣΗ ΣΩΜΑΤΟΣ	0,901
ΤΕΧΝΙΚΕΣ ΠΟΔΙΩΝ	0,898
ΤΕΧΝΙΚΕΣ ΧΕΡΙΩΝ	0,863
ΤΕΧΝΙΚΕΣ ΓΟΦΟΥ	0,874
ΤΕΧΝΙΚΕΣ ΑΥΤΟΘΥΣΙΑΣ	0,887
ΤΕΧΝΙΚΕΣ ΠΛΑΓΙΑΣ ΑΥΤΟΘΥΣΙΑΣ	0,767
ΤΕΧΝΙΚΕΣ ΑΚΙΝΗΤΟΠΟΙΗΣΗΣ	0,798
ΤΕΧΝΙΚΕΣ ΠΙΝΙΓΜΟΥ	0,912
ΤΕΧΝΙΚΕΣ ΕΞΑΡΘΡΩΣΗΣ ΑΓΚΩΝΑ	0,813

Από τον ανωτέρω πίνακα 2 διαπιστώνουμε ότι ο δείκτης αξιοπιστίας Cronbach's α για τις μεταβλητές της έρευνας όπως παρατηρήθηκαν για το χρόνο επίτευξης της νίκης (0,898), το πιάσιμο του αντιπάλου (0,869), τη στάση του σώματος (0,901), τις τεχνικές ποδιών (0,898), χεριών (0,863), γοφού (0,874), αυτοθυσία (0,887), πλάγιας αυτοθυσίας (0,767), ακινητοποίησης (0,798), πνιγμού (0,912) και εξάρθρωσης αγκώνα (0,813), είναι πολύ υψηλός, γεγονός που υποδηλώνει πολύ υψηλή εσωτερική συνέπεια των παρατηρήσεων.

Έλεγχος συμφωνίας μεταξύ Αξιολογητών-Εκτιμητών και ενδοβαθμολογική αξιοπιστία αναφορικά με τις μεταβλητές της έρευνας

Η Συμφωνία μεταξύ Αξιολογητών-Εκτιμητών (Kappa του Cohen ή Inter-observer Reliability) υπολογίστηκε για τις έντεκα μεταβλητές, όπως αποτυπώνεται στον κάτωθι πίνακα 3:

Πίνακας 3: Συμφωνία μεταξύ Κριτών- Αξιολογητών (Cohen's Kappa)

Εργαλείο	Cohen's Kappa	Αριθμός έγκυρων περιπτώσεων	Approx. Sig.
ΧΡΟΝΟΣ ΕΠΙΤΕΥΞΗΣ ΝΙΚΗΣ	0,718	267	,001
ΠΙΑΣΙΜΟ ΑΝΤΙΠΑΛΟΥ	0,789	267	,001
ΣΤΑΣΗ ΣΩΜΑΤΟΣ	0,786	267	,001
ΤΕΧΝΙΚΕΣ ΠΟΔΙΩΝ	0,901	267	,001
ΤΕΧΝΙΚΕΣ ΧΕΡΙΩΝ	0,787	267	,001
ΤΕΧΝΙΚΕΣ ΓΟΦΟΥ	0,750	267	,001
ΤΕΧΝΙΚΕΣ ΑΥΤΟΘΥΣΙΑΣ	0,852	267	,001
ΤΕΧΝΙΚΕΣ ΠΛΑΓΙΑΣ ΑΥΤΟΘΥΣΙΑΣ	0,902	267	,001
ΤΕΧΝΙΚΕΣ ΑΚΙΝΗΤΟΠΟΙΗΣΗΣ	0,871	267	,001
ΤΕΧΝΙΚΕΣ ΠΝΙΓΜΟΥ	0,813	267	,001
ΤΕΧΝΙΚΕΣ ΕΞΑΡΘΡΩΣΗΣ ΑΓΚΩΝΑ	0,905	267	,001

Από τον ανωτέρω πίνακα 3 διαπιστώνουμε ότι ο συντελεστής Kappa για τις μεταβλητές της έρευνας όπως παρατηρήθηκαν για το χρόνο επίτευξης της νίκης (0,718), το πιάσιμο του αντιπάλου (0,789), τη στάση του σώματος (0,786), τις τεχνικές ποδιών (0,901, χεριών (0,787), γοφού (0,750, αυτοθυσία (0,852), πλάγιας αυτοθυσίας (0,902), ακινητοποίησης (0,871), πνιγμού (0,813) και εξάρθρωσης αγκώνα (0,905), γεγονός που υποδηλώνει ικανοποιητική συμφωνία μεταξύ των δύο κριτών. Στις υπόλοιπες πέντε περιπτώσεις η τιμή του συντελεστή Kappa είναι ίση ή ξεπερνά το όριο του 0,75 και κατ'επέκταση υφίσταται υψηλή συμφωνία ανάμεσα στους αξιολογητές. Παράλληλα, με βάση το προσεγγιστικό επίπεδο στατιστικής σημαντικότητας (p-value) ο συντελεστής Kappa είναι σε όλες τις περιπτώσεις στατιστικά σημαντικός.

Πίνακας 4: Ενδοβαθμιολογική Αξιοπιστία (Intra-observer Reliability)

ΜΕΤΑΒΛΗΤΕΣ ΕΡΕΥΝΑΣ	Intra-class Correlation	Αριθμός έγκυρων περιπτώσεων
ΧΡΟΝΟΣ ΕΠΙΤΕΥΞΗΣ ΝΙΚΗΣ	0,992	267
ΠΙΑΣΙΜΟ ΑΝΤΙΠΑΛΟΥ	0,994	267
ΣΤΑΣΗ ΣΩΜΑΤΟΣ	0,998	267
ΤΕΧΝΙΚΕΣ ΠΟΔΙΩΝ	0,992	267
ΤΕΧΝΙΚΕΣ ΧΕΡΙΩΝ	0,997	267
ΤΕΧΝΙΚΕΣ ΓΟΦΟΥ	0,999	267
ΤΕΧΝΙΚΕΣ ΑΥΤΟΘΥΣΙΑΣ	0,952	267
ΤΕΧΝΙΚΕΣ ΠΛΑΓΙΑΣ ΑΥΤΟΘΥΣΙΑΣ	0,931	267
ΤΕΧΝΙΚΕΣ ΑΚΙΝΗΤΟΠΟΙΗΣΗΣ	0,977	267
ΤΕΧΝΙΚΕΣ ΠΝΙΓΜΟΥ	0,967	267
ΤΕΧΝΙΚΕΣ ΕΞΑΡΘΡΩΣΗΣ ΑΓΚΩΝΑ	0,975	267

Ο ανωτέρω έλεγχος εξάγει αποτελέσματα της συνέπειας ή της συμφωνίας των τιμών μεταξύ των περιπτώσεων (cases- συμμετέχοντες στην έρευνα). Με βάση τα ανωτέρω, μπορούμε να ισχυριστούμε πως, ένα ποσοστό της τάξεως ίσης ή μεγαλύτερης του 99,5% της διακύμανσης (variance) του μέσου (mean) αυτών των αξιολογητών, είναι «πραγματικό» (Shrout & Fleiss, 1979).

6.2 Αποτελέσματα περιγραφικής στατιστικής, πίνακες εμφάνισης και σχετικών συχνοτήτων

Στον πιν. 5 διακρίνονται οι συχνότητες των περιπτώσεων για το χρόνο (τον αριθμό των λεπτών) που σημειώθηκαν οι νίκες. Σημαντικό χαρακτηριστικό για την εκπαίδευση, προπόνηση διότι καθορίζει και την ποιότητα της προπονητικής διαδικασίας στους ενεργειακούς μηχανισμούς παραγωγής έργου (αερόβιο, αναερόβιο, μεικτό). Η πλειονότητα των νικών επιτυγχάνεται στον επιπλέον χρόνο, στο τέταρτο, στο δεύτερο και πρώτο λεπτό. Αυτό με βάση τη θεωρία των ενεργειακών ορίων της μυϊκής προσπάθειας (Κλεισούρας, 1989), δείχνει ότι η προπονητική διαδικασία που θα προγραμματιστεί θα στοχεύει ξεχωριστά και στους τρεις μηχανισμούς παραγωγής ενέργειας (ασκήσεις που διαρκούν: αναερόβιο γαλακτικό μέχρι 1' λεπτό, αναερόβιο γαλακτικό από 1' έως 6' λεπτά και αερόβιο >30' λεπτών). Έτσι λοιπόν, η προπόνηση της τεχνικής και τακτικής προγραμματίζεται αντίστοιχα σε ένα εβδομαδιαίο μικρόκυκλο, που θα επαναλαμβάνεται για τουλάχιστον 10 εβδομάδες. Η προγραμματισμένη προπόνηση βασιζόμενη στον αναερόβιο μηχανισμό παραγωγή ενέργειας

έχει: α) συχνότητα εφαρμογής τρεις φορές την εβδομάδα, β) ένταση που αντιστοιχεί στο 95-100% της μέγιστης καρδιακής συχνότητας της αθλήτριας, γ) διαλείμματα ξεκούρασης και αποκατάστασης εντός της προπόνησης τριών (3') λεπτών, δ) διαλείμματα μεταξύ των προπονήσεων τουλάχιστον σαράντα οκτώ (48) ώρες, και ε) διάρκεια της κάθε άσκησης θα κυμαίνεται από 10'' δευτερόλεπτα έως 10' λεπτά.

Πίνακας 5: συχνότητων και ποσοστών για το χρόνο επίτευξης της νίκης

ΧΡΟΝΟΣ ΕΠΙΤΕΥΞΗΣ ΤΗΣ ΝΙΚΗΣ	f	%
1 ^ο ΛΕΠΤΟ	49	18,3
2 ^ο ΛΕΠΤΟ	53	19,9
3 ^ο ΛΕΠΤΟ	47	17,6
4 ^ο ΛΕΠΤΟ	56	21
ΕΠΙΠΛΕΟΝ ΧΡΟΝΟΣ	62	23,2
ΣΥΝΟΛΟ	267	100

Πίνακας 6 συχνότητων και ποσοστών για τον τρόπο πιασίματος

ΤΡΟΠΟΣ ΠΙΑΣΙΜΑΤΟΣ ΑΝΤΙΠΑΛΟΥ	f	%
ΔΕΞΙΑ	85	32
ΑΡΙΣΤΕΡΑ	63	23,6
ΑΜΦΙΔΕΞΙΑ	119	44,4
ΣΥΝΟΛΟ	267	100

Πίνακας 7: συχνότητων και ποσοστών για τη στάση του σώματος

ΣΤΑΣΗ ΣΩΜΑΤΟΣ	f	%
ΟΡΘΙΟΣ	51	19,1
ΕΛΑΦΡΩΣ ΣΚΥΦΤΗ ΜΠΡΟΣΤΑ	147	55
ΣΚΥΦΤΗ ΚΑΙ ΛΥΓΙΣΜΕΝΑ ΓΟΝΑΤΑ	69	25,9
ΣΥΝΟΛΟ	267	100

Στον πιν. 6 διακρίνονται οι συχνότητες εμφάνισης των νικηφόρων τρόπων πιασίματος. Σημαντική μεταβλητή για την εκπαίδευση διότι καθορίζει και την ποιότητα της προπονητικής διαδικασίας στο πιάσιμο έτσι ώστε η αθλήτρια να πιάνει με τέτοιο τρόπο έτσι ώστε ταυτόχρονα να αμύνεται και να εμποδίζει την αντίπαλο να εκτελέσει τεχνικές ενώ αυτή να έχει το πλεονέκτημα του πιασίματος για τις τεχνικές της. Φαίνεται ότι η πλειονότητα των αθλητριών της κατηγορίας των 57 κιλών των γυναικών είναι αμφιδέξιες, γεγονός που εκφράζει το υψηλό επίπεδο εκπαίδευσής τους. Η εκπαιδευτική οδηγία που είναι απόρροια των αποτελεσμάτων του πίνακα 6, έχει να κάνει με την εκπαίδευση των τεχνικών και τακτικών τόσο με την κυρίαρχη όσο και με τη μη κυρίαρχη πλευρά της αθλήτριάς μας, με στόχο να την καταστήσει αμφιδέξια σε άμυνα, επίθεση και αντεπίθεση.

Από τον πιν. 7 διαπιστώνεται ότι η ελαφρώς σκυφτή στάση του σώματος της αθλήτριας (55%) και μάλιστα προς τα εμπρός είναι με διαφορά η συνηθέστερη όταν επιτυγχάνεται η νίκη.

Πίνακας 8: συχνότητων και αντίστοιχων ποσοστών για τεχνικές της όρθιας πάλης

ΤΕΧΝΙΚΕΣ ΠΟΔΙΩΝ	f	%	ΤΕΧΝΙΚΕΣ ΧΕΡΙΩΝ	f	%	ΤΕΧΝΙΚΕΣ ΓΟΦΟΥ	f	%
KOSOTO GAKE	16	5,9	KATA GURUMA	11	4,0	HANE GOSHI	14	5,1
KOSOTO GARI	17	6,2	MOROTESEONAGE	15	5,5	HARAI GOSHI	25	9,2
KOUCHI GARI	23	8,4	TAI OTOSHI	16	5,9	KOSHI GURUMA	15	5,5
OUCHI GARI	27	9,9	IPPON SEONAGE	22	8,1	OGOSHI	15	5,5
OSOTO GARI	21	7,7	SUKUI NAGE	11	4,0	SOSETSURIKOMIGOS	13	4,8
OSOTO OTOSHI	22	8,1	SUMI OTOSHI	12	4,4	TSUI KOMI GOSHI	11	4,0
SASAETSURIKOMIASHI	12	4,4	UKI OTOSHI	15	5,5	UKI GOSHI	11	4,0
UCHIMATA	35	12,8	SEOI OTOSHI	11	4,0	USHIRO GOSHI	11	4,0
DEASHI HARAI	16	5,9	ΣΥΝΟΛΟ	113	41,4	ΣΥΝΟΛΟ	115	42,1
ΣΥΝΟΛΟ	189	69,2	ΕΛΛΕΙΜΜΑ	154	58,6	ΕΛΛΕΙΜΜΑ	152	57,90
ΕΛΛΕΙΜΜΑ	78	30,8	ΣΥΝΟΛΟ	267	100	ΣΥΝΟΛΟ	267	100
ΣΥΝΟΛΟ	267	100						
ΤΕΧΝΙΚΕΣ	f	%	ΠΑΛΓΙΑ	f	%			
ΑΥΤΟΘΥΣΙΑΣ			ΑΥΤΟΘΥΣΙΑ					
SUMI GAESHI	15	5,5	UKI WAZA	11	5,8			
TAMOE NAGE	11	4,0	HARAI MAKIKOMI	11	5,8			
URA NAGE	13	4,8	OSOTO MAKIKOMI	11	5,8			
TANI OTOSHI	18	6,6	YOKOTOMOENAGE	14	7,4			
ΣΥΝΟΛΟ	57	20,9	UCHIMATAGAESHI	16	8,5			
ΕΛΛΕΙΜΜΑ	210	79,1	ΣΥΝΟΛΟ	63	33,3			
ΣΥΝΟΛΟ	267	100	ΕΛΛΕΙΜΜΑ	204	66,7			
			ΣΥΝΟΛΟ	267	100			

Στον πιν. 8 επάνω απεικονίζονται οι συχνότητες εμφάνισης των τεχνικών της όρθιας πάλης όσο και για τις ίδιες τεχνικές που εκτελέστηκαν στην κατηγορία των 57 κιλών των γυναικών

σε ολόκληρο το δείγμα μελέτης αγώνων. Επίσης, στον πίνακα 8, το «έλλειμμα» που εμφανίζεται σε κάθε κατηγορία τεχνικών αναφέρεται στον αριθμό των περιπτώσεων (από τις 267) που δεν εμπίπτουν στη συγκεκριμένη κατηγορία τεχνικής. Τη μεγαλύτερη συχνότητα εμφάνισης παρουσιάζουν:

- α) στην κατηγορία τεχνικής ποδιών *uchimata*, *ouchi gari* και *kouchi gari*,
- β) στην κατηγορία τεχνικής χεριών *ippon seoinage*, *tai otoshi*, *uki otoshi* και *morote seoinage*,
- γ) στην κατηγορία τεχνικών γοφού *harai goshi*, *koshi guruma*, και *ogoshi*,
- δ) στην κατηγορία τεχνικών αυτοθυσίας *tani otoshi*, *sumi gaeshi* και *ura nage*,
- ε) στην κατηγορία τεχνικών πλάγιας αυτοθυσίας *uchimata gaeshi* και *yoko tomoe nage*.

Συνάγεται ότι η δική μας αθλήτρια θα πρέπει να εκπαιδευτεί ώστε να μπορεί να ανταπεξέλθει τουλάχιστον στις συνηθέστερες ενέργειες: α) σε επίθεση για να μπορεί να νικάει, β) σε άμυνα μπλοκάροντας και αποφεύγοντας αποτελεσματικά την επίθεση της αντιπάλου, αλλά και γ) σε αντεπίθεση, δηλαδή, να χρησιμοποιεί τις αποτελεσματικά τις αντιτεχνικές της όταν η αντίπαλος επιχειρεί με αυτές να τη ρίξει. Όλα αυτά σε συνδυασμό με τον τρόπο πιασίματος και τη σωστή στάση του σώματος στην όρθια πάλη.

Στον πίν. 9 που ακολουθεί απεικονίζονται οι συχνότητες εμφάνισης και οι σχετικές τόσο για τις κατηγορίες των τεχνικών της πάλης εδάφους όσο και για τις ίδιες τεχνικές που εκτελέστηκαν στην κατηγορία των 57 κιλών των γυναικών σε ολόκληρο το δείγμα μελέτης αγώνων. Τη μεγαλύτερη συχνότητα εμφάνισης παρουσιάζουν: α) στην κατηγορία τεχνικής ακινητοποίησης *kuzure kesa gatame* και *Yoko shiho gatame*, β) στην κατηγορία τεχνικής πνιγμού *Okuri eri jime* και *Sankaku jime*, γ) στην κατηγορία τεχνικών εξάρθρωσης αγκώνα *Ude hishigi juji gatame*, και *Ude hishigi hara gatame*. Στις τεχνικές αυτές με τη μεγαλύτερη συχνότητα εμφάνισης στον πίν. 10, η δική μας αθλήτρια θα πρέπει να εκπαιδευτεί: α) σε επίθεση για να μπορεί να νικάει, β) σε άμυνα μπλοκάροντας και αποφεύγοντας αποτελεσματικά την επίθεση της αντιπάλου, αλλά και γ) σε αντεπίθεση, δηλαδή, να χρησιμοποιεί τις αποτελεσματικά τις αντιτεχνικές της όταν η αντίπαλος επιχειρεί με αυτές να τη νικήσει στην πάλη εδάφους.

Πίνακας 9: συχνότητων και αντίστοιχων ποσοστών τεχνικών στο έδαφος

ΤΕΧΝΙΚΕΣ ΑΚΙΝΗΤΟΠΟΙΗΣΗΣ	f	%	ΤΕΧΝΙΚΕΣ ΠΝΙΓΜΩΝ	f	%
KAMI SHIHO GATAME	14	7,4	HADAKA JIME	13	6,9
KESA GATAME	13	6,9	OKURI ERI JIME	18	9,5
KUZURE KAMI SHIHO GATAME	12	6,3	SANKAKU JIME	13	6,9
KUZURE TATE SHIHO GATAME	12	6,3	ΣΥΝΟΛΟ	44	23,3
KUZURE KESA GATAME	18	9,5	ΕΛΛΕΙΜΜΑ	223	76,7
TATE SHIHO GATAME	10	5,3	ΣΥΝΟΛΟ	267	100
USHIRO KESA GATAME	10	5,3	ΤΕΧΝΙΚΕΣ ΕΞΑΡΘΡΩΣΗΣ ΑΓΚΩΝΑ	f	%
YOKO SHIHO GATAME	16	8,5	UDE HISHIGI HARA GATAME	13	6,9
ΣΥΝΟΛΟ	105	55,6	UDE HISHIGI JUJI GATAME	17	9,0
ΕΛΛΕΙΜΜΑ	162	7,4	UDE HISHIGI WAKI GATAME	11	5,8
ΣΥΝΟΛΟ	267	100	ΣΥΝΟΛΟ	41	21,7
			ΕΛΛΕΙΜΜΑ	226	78,3
			ΣΥΝΟΛΟ	267	100

Πίνακας 10: συχνότητων εμφάνισης για τη σχέση δεδομένων μεταξύ του τρόπου πιασίματος και των τεχνικών εκτέλεσης με τα πόδια (N=267, 100%).

ΤΕΧΝΙΚΕΣ ΠΟΔΙΩΝ ASHI WAZA		ΔΕΞΙΑ ΠΛΕΥΡΑ	ΑΡΙΣΤΕΡΗ ΠΛΕΥΡΑ	ΑΜΦΙ- ΔΕΞΙΑ	ΣΥΝΟΛΟ
ΤΕΧΝΙΚΕΣ ΠΟΔΙΩΝ ASHI WAZA	KOSOTOGAKE	16	0	0	16
	KOSOTO GARI	17	0	0	17
	KOUCHI GARI	23	0	0	23
	OUCHI GARI	3	24	0	27
	OSOTO GARI	0	13	8	21
	OSOTOOTOSHI	0	0	22	22
	SASAETSURIKOMIAS	0	0	12	12
	UCHIMATA	0	0	35	35
	DEASHI HARAI	0	0	16	16
ΣΥΝΟΛΟ		59	37	93	189

Πίνακας 11: Συχνότητες για τη σχέση δεδομένων μεταξύ του τρόπου πιασίματος και των τεχνικών εκτέλεσης με τα χέρια (N=267, 100%)

ΤΕΧΝΙΚΕΣ ΧΕΡΙΩΝ TE WAZA		ΔΕΞΙΑ ΠΛΕΥΡΑ	ΑΡΙΣΤΕΡΗ ΠΛΕΥΡΑ	ΑΜΦΙ- ΔΕΞΙΑ	ΣΥΝΟΛΟ
ΤΕΧΝΙΚΕΣ ΧΕΡΙΩΝ TE WAZA	KATAGURUM	11	0	0	11
	MOROTSEOI	15	0	0	15
	TAI OTOSHI	16	0	0	16
	IPPONSEOING	17	5	0	22
	SUKUI NAGE	0	11	0	11
	SUMI OTOSHI	0	12	0	12
	UKI OTOSHI	0	9	6	15
	SEOI OTOSHI	0	0	11	11
	Total	59	37	17	113
ΣΥΝΟΛΟ		59	37	93	189

6.3 Αποτελέσματα επαγωγικής στατιστικής

6.3.1 Αποτελέσματα χ^2 για τις διαφορές μεταξύ δειγμάτων με δεδομένα συχνοτήτων

6.3.1.1 Χρόνος και κατηγορίες τεχνικών επίτευξης νίκης

Για να ελέγξουμε τη σχέση του χρόνου επίτευξης της νίκης (στο 1^ο, 2^ο, 3^ο, 4^ο λεπτό ή στον επιπλέον χρόνο) με ποιες τεχνικές επετεύχθη η νίκη πραγματοποιήθηκε έλεγχος χ^2 , δηλαδή, με τις τεχνικές που νίκησε και ηττήθηκε η αθλήτρια αλλά και σε εκείνες τις τεχνικές που έφεραν τη νίκη σε παγκόσμιους, πανευρωπαϊκούς και Ολυμπιακούς αγώνες της κατηγορίας των 57 κιλών, στους οποίους δεν συμμετείχε η αθλήτριά μας.

Πίνακας 12: Αποτελεσμάτων στατιστικής ανάλυσης χ^2 για τις διαφορές μεταξύ δειγμάτων με δεδομένα συχνοτήτων (χρόνος επίτευξης νίκης και κατηγορίες τεχνικών για N=267, 100%)

ΧΡΟΝΟΣ & ΚΑΤΗΓΟΡΙΕΣ ΤΕΧΝΙΚΩΝ	χ^2	df	p _{2-Tailed}
ΧΡΟΝΟΣ & ΤΕΧΝΙΚΕΣ ΠΟΔΙΩΝ	605,5	32	.001
ΧΡΟΝΟΣ & ΤΕΧΝΙΚΕΣ ΧΕΡΙΩΝ	286,4	21	.001
ΧΡΟΝΟΣ & ΤΕΧΝΙΚΕΣ ΓΟΦΟΥ	282,6	21	.001
ΧΡΟΝΟΣ & ΤΕΧΝΙΚΕΣ ΑΥΤΟΘΥΣΙΑΣ	44,6	3	.001
ΧΡΟΝΟΣ & ΤΕΧΝΙΚΕΣ ΠΛΑΓΙΑΣ ΑΥΤΟΘΥΣΙΑΣ	56,5	4	.001
ΧΡΟΝΟΣ & ΤΕΧΝΙΚΕΣ ΑΚΙΝΗΤΟΠΟΙΗΣΗΣ	210	21	.001
ΧΡΟΝΟΣ & ΤΕΧΝΙΚΕΣ ΠΝΙΓΜΩΝ	44	2	.001
ΧΡΟΝΟΣ & ΤΕΧΝΙΚΕΣ ΕΞΑΡΘΡΩΣΗΣ ΑΓΚΩΝΑ	36,1	2	.001

Σύμφωνα με τη θεωρία των ενεργειακών μηχανισμών παραγωγής ενέργειας στη βάση της φυσιολογίας του ανθρώπινου σώματος για την παραγωγή μηχανικού έργου, ο χρόνος εκτέλεσης της τεχνικής (με βάση τους πίνακες 5 και 12) καθορίζει σημαντικά στη συνέχεια την προπόνηση, δηλαδή, τα χρονικά όρια και την ένταση των ασκήσεων γενικού (όπως τρέξιμο, αντιστάσεις με βάρη, κολύμπι, κ.ά.) και ειδικού (τεχνική και τακτική) τύπου. Αυτό

σημαίνει ότι παρατηρώντας το χρόνο εκτέλεσης της αγωνιστικής τεχνικής, η αθλήτρια θα πρέπει να εκπαιδευθεί στην αναερόβια αγαλακτική και αερόβια γαλακτική (γενικού και ειδικού τύπου ασκήσεις) μορφή προπόνησης. Η αναερόβια αγαλακτική προπόνηση αναφέρεται σε γενικού και ειδικού τύπου ασκήσεις μεγάλης έντασης και μικρής χρονικής διάρκειας (το πολύ μέχρι δέκα δευτερόλεπτα), ενώ η επίπονη αναερόβια γαλακτική προπόνηση αναφέρεται σε γενικού και ειδικού τύπου ασκήσεις μεγάλης έντασης και χρονικής διάρκειας (από δέκα δευτερόλεπτα και πάνω). Στην πραγματικότητα η προπόνηση θα πρέπει να διαμορφωθεί σε ασκήσεις που θα στοχεύουν στην βελτίωση και άμεση ανταπόκριση των μηχανισμών παραγωγής ενέργειας της αθλήτριας στον αγώνα, έτσι ώστε να υπάρχει μια άνετη και αποτελεσματική αντιμετώπιση των ενεργειακών αποθεμάτων, των αγωνιστικών απαιτήσεων και της εμφάνισης της κούρασης που είναι το αποτέλεσμα των ελλείψεων απόδοσης των ενεργειακών αυτών μηχανισμών. Ακόμα η μελέτη του χρόνου εκτέλεσης τεχνικών διαμορφώνει και την ποσόστωση των ασκήσεων ειδικού τύπου στην προπόνηση, δηλαδή, σε κάθε εβδομαδιαίο μικρόκυκλο θα πρέπει να περιέχονται δύο προπονητικές μονάδες με χρονική απόσταση μεταξύ τους τουλάχιστον 48 ωρών, όπου η αθλήτρια θα εκτελεί αναερόβια γαλακτικής μορφής ασκήσεις και άλλες τέσσερις προπονητικές μονάδες, με ασκήσεις αναερόβια αγαλακτικής μορφής. Ο εβδομαδιαίος μικρόκυκλος θα περιέχει ειδικού τύπου ασκήσεις (τεχνικής και τακτικής) σε ποσοστό 60% αναερόβια αγαλακτικής και 40 αναερόβια γαλακτικής μορφής. Η ποσόστωση αυτή τον τελευταίο μήνα πριν την κρίσιμη διοργάνωση, και, όσο χρονικά πλησιάζουμε προς αυτήν αλλάζει και γίνεται από 80 – 100% αναερόβια αγαλακτική και 20 – 0% αναερόβια γαλακτική, με χαλάρωση της ποσότητας των ασκήσεων.

6.3.1.2 Τρόπος πιασίματος αντιπάλου (KUMIKATA), στάση του σώματος και κατηγορίες τεχνικών επίτευξης νίκης

Για να ελέγξουμε τη σχέση του τρόπου πιασίματος των αντιπάλων αθλητριών για την επίτευξη της νίκης (δεξιό, αριστερό, αμφιδέξιο), της στάσης του σώματος (όρθια, ελαφρώς σκυφτή, εντελώς σκυφτή) με ποιες τεχνικές επετεύχθη η νίκη πραγματοποιήθηκε έλεγχος χ^2 .

Πίνακας 13: αποτελεσμάτων στατιστικής ανάλυσης χ^2 για τις διαφορές μεταξύ δειγμάτων με δεδομένα συχνότητας (τρόπος πιασίματος αντιπάλου KUMIKATA, στάση του σώματος και κατηγορίες τεχνικών που έφεραν τη νίκη για $N=267$, 100%)

ΤΡΟΠΟΣ ΠΙΑΣΙΜΑΤΟΣ / ΣΤΑΣΗ ΣΩΜΑΤΟΣ & ΚΑΤΗΓΟΡΙΕΣ ΤΕΧΝΙΚΩΝ	χ^2	df	p2-Tailed
ΤΡΟΠΟΣ ΠΙΑΣΙΜΑΤΟΣ & ΤΕΧΝΙΚΕΣ ΠΟΔΙΩΝ	320,5	16	.001
ΣΤΑΣΗ ΣΩΜΑΤΟΣ & ΤΕΧΝΙΚΕΣ ΠΟΔΙΩΝ	317,8	16	.001
ΤΡΟΠΟΣ ΠΙΑΣΙΜΑΤΟΣ & ΤΕΧΝΙΚΕΣ ΧΕΡΙΩΝ	171,9	14	.001
ΣΤΑΣΗ ΣΩΜΑΤΟΣ & ΤΕΧΝΙΚΕΣ ΧΕΡΙΩΝ	98,2	7	.001
ΤΡΟΠΟΣ ΠΙΑΣΙΜΑΤΟΣ & ΤΕΧΝΙΚΕΣ ΓΟΦΟΥ	193,2	14	.001
ΣΤΑΣΗ ΣΩΜΑΤΟΣ & ΓΟΦΟΥ	79	7	.001
ΤΡΟΠΟΣ ΠΙΑΣΙΜΑΤΟΣ & ΤΕΧΝΙΚΕΣ ΑΥΤΟΘΥΣΙΑΣ**			
ΤΡΟΠΟΣ ΠΙΑΣΙΜΑΤΟΣ & ΤΕΧΝΙΚΕΣ ΠΛ.ΑΥΤΟΘΥΣΙΑΣ	12,6	4	.014

** Ο τρόπος πιασίματος στις τεχνικές αυτοθυσίας είναι σταθερός. Οι τεχνικές αυτοθυσίας εμφανίστηκαν ως τεχνικές νίκης μόνο με ένα συγκεκριμένο είδος πιασίματος.

Η στάση του σώματος στην κινητικότητα που αναπτύσσει το σύστημα των δύο αθλητριών που αγωνίζονται με κανονισμούς του τζούντο είναι ένας καθοριστικός παράγοντας τόσο για την άμυνα, δηλαδή, την αποφυγή των επιθέσεων της αντιπάλου όσο και της επίτευξης της νίκης, με επιθέσεις ή αντεπιθέσεις. Από τα αποτελέσματα των πινάκων 7 και 13 φαίνεται ότι η υιοθέτηση μιας στάσης, ελαφρώς σκυφτή μπροστά είναι μια ευρέως αποδεκτή σωστή στάση στην διάρκεια του αγώνα που επιτρέπει στην αθλήτρια να περνάει άμεσα τότε σε μια αμυντικογενή στάση σώματος χαμηλώνοντας το κέντρο βάρους της για να δυσχεραίνει και να αποφύγει τις επιθέσεις των αντιπάλων της και τότε σε μια επιθετικογενή πιο όρθια στάση.

Από τους πίνακες 6 και 13 φαίνεται ότι η αθλήτρια πρέπει να εκπαιδευθεί και να ενισχύσει τη μη κυρίαρχη πλευρά του σώματός της, έτσι ώστε να εκτελεί με άνεση επιθέσεις όσο και να αμύνεται αποτελεσματικά. Για παράδειγμα, στην προπόνηση, η αθλήτρια θα πρέπει να πιάνει εξίσου καλά τόσο με τη δεξιά όσο και με την αριστερή λαβή των χεριών της και ταυτόχρονα να εκτελεί αποτελεσματικά τεχνικές προς τις δύο κατευθύνσεις (αμφιδέξια). Η προπονητική διαδικασία απαιτεί χρόνο εκπαίδευσης στην επίθεση, στην αντεπίθεση και στην άμυνα εκτελώντας τεχνικές και συνδυασμούς τεχνικών με την κυρίαρχη και με τη μη κυρίαρχη πλευρά του σώματος.

7. ΕΛΕΓΧΟΣ ΥΠΟΘΕΣΕΩΝ

Τα αποτελέσματα έδειξαν ότι στην κατηγορία των 57 κιλών των γυναικών το μεγαλύτερο ποσοστό επίτευξης της νίκης έγινε στην όρθια πάλη με τις τεχνικές των ποδιών, του γοφού, των χεριών, της πλάγιας αυτοθυσίας και τελευταία της αυτοθυσίας. Στην πάλη εδάφους το μεγαλύτερο ποσοστό επίτευξης της νίκης έγινε περισσότερο με τις τεχνικές ακινητοποίησης και λιγότερο με τις τεχνικές πνιγμού και εξάρθρωσης του αγκώνα. Μετά την ποσοτική ανάλυση των δεδομένων, τη μεγαλύτερη συμμετοχή στην επίτευξη της νίκης στους αγώνες της κατηγορίας των 57 κιλών εμφάνισαν κατά σειρά οι τεχνικές: ποδιών, γοφού, χεριών, ακινητοποίησης στο έδαφος, πλάγιας αυτοθυσίας, αυτοθυσίας, πνιγμών και των εξαρθρώσεων του αγκώνα. Ο έλεγχος των υποθέσεων επαληθεύει τις υποθέσεις Y_1 και Y_2 .

Η υπόθεση Y_3 αναφέρεται στην ύπαρξη σημαντικής σχέσης μεταξύ των τεχνικών επίτευξης της νίκης και του χρόνου εκτέλεσης της τεχνικής επί της αντιπάλου αθλήτριας. Από τα αποτελέσματα της παραγράφου 6.3.1.1 και του πίνακα (12) διαπιστώνεται η σημαντική σχέση του χρόνου εκτέλεσης και τους είδους της τεχνικής. Σημαντικό, επίσης, είναι ότι αναφέρεται ο χρόνος, πότε δηλαδή επιτεύχθηκε η τεχνική νίκης σε σχέση με την κατηγορία της τεχνικής. Ο έλεγχος επαληθεύει την αρχική υπόθεση Y_3 .

Η υπόθεση Y_4 αναφέρεται στην ύπαρξη σημαντικής σχέσης μεταξύ των τεχνικών επίτευξης της νίκης και του τρόπου πιασίματος επί της αντιπάλου αθλήτριας. Από τα αποτελέσματα της παραγράφου 6.3.1.2 και του πίνακα (13) διαπιστώνεται η σημαντική σχέση του τρόπου πιασίματος επί της αντιπάλου αθλήτριας και τους είδους της τεχνικής. Ο έλεγχος επαληθεύει την αρχική υπόθεση Y_4 για όλες τις κατηγορίες των τεχνικών εκτός των κατηγοριών των τεχνικών αυτοθυσίας στην όρθια πάλη και των τεχνικών στο έδαφος.

Η υπόθεση Y_5 αναφέρεται στην ύπαρξη σημαντικής σχέσης μεταξύ των τεχνικών επίτευξης της νίκης και της στάσης του σώματος της επιτιθέμενης αθλήτριας. Από τα αποτελέσματα της παραγράφου 6.3.1.3 και του πίνακα (13) διαπιστώνεται η σημαντική σχέση της στάσης του σώματος της επιτιθέμενης αθλήτριας και τους είδους της τεχνικής που εκτελεί και νικάει. Ο έλεγχος επαληθεύει την αρχική υπόθεση Y_5 για τις κατηγορίες των τεχνικών της όρθιας πάλης (τεχνικές ποδιών, χεριών και γοφού) εκτός των τεχνικών αυτοθυσίας και πλάγιας αυτοθυσίας στην όρθια πάλη και των τριών κατηγοριών τεχνικών στο έδαφος.

8. ΣΥΜΠΕΡΑΣΜΑΤΑ, ΣΥΖΗΤΗΣΗ

Τα αποτελέσματα της έρευνας αποτελούν ένα οδηγό για τη διαμόρφωση της προπόνησης της αθλήτριας. Συγκεκριμένα, ο σχεδιασμός της προπόνησης αναφέρεται στα αποτελέσματα της παρούσας ποσοτικής ανάλυσης των τεχνικών, έτσι ώστε να παρουσιαστούν και να αναλυθούν περαιτέρω οι καταγεγραμμένες τεχνικές. Να γίνει γνωστός πρώτα ο τρόπος εφαρμογής τους ως τεχνικές επίθεσης και μετά να εφαρμοστούν και ως τεχνικές αντιμετώπισής τους (αντεπίθεση με την ίδια τεχνική) και ακόμη, να εφαρμοστούν άλλες τεχνικές εξουδετέρωσης αυτών των καταγεγραμμένων τεχνικών από την ενδιαφερόμενη αθλήτρια. Προκύπτει ένας τριπλός εκπαιδευτικός – προπονητικός στόχος. Επαναλαμβάνοντας συνοπτικά τα παραπάνω τόσο στην όρθια όσο και στην πάλη εδάφους θα πρέπει η αθλήτρια να εκπαιδευθεί σε αυτές τις τεχνικές πολύ καλά και να τις ενσωματώσει στο αγωνιστικό κινητικό της ρεπερτόριο ως: α) τεχνικές επίθεσης, β) τεχνικές αντεπίθεσης, και γ) συνδυαστικές τεχνικές.

Σε συνδυασμό με τα αποτελέσματα της παρούσας μελέτης για τον τρόπο πιασίματος της αντιπάλου αθλήτριας αλλά και τη στάση που πρέπει να υιοθετήσει για την εκτέλεση της τεχνικής υπάρχει ένα πλήθος τακτικών για την αποτελεσματική έκβαση του αγώνα που θα πρέπει να αναπτύξει η ενδιαφερόμενη αθλήτρια. Ο αμφιδέξιος τρόπος που εμφανίστηκε σε μεγάλο ποσοστό στην παρούσα μελέτη καθιστά την προπόνηση με ισόποσο φόρτο εργασίας κατά την εκτέλεση των τεχνικών και από τις δύο πλευρές. Οι τεχνικές των ποδιών για παράδειγμα (Kosoto gake, Kosoto gari, Kouchi gari, Uchi gari, Osoto gari, Osoto otoshi, Sasae tsuri komi ashi, Uchimata , Deashi harai), θα πρέπει να εκτελούνται ισοδύναμα τόσο από την δεξιά όσο και από την αριστερή πλευρά στην άμυνα και στην επίθεση. Η στάση του σώματος που πρέπει να υιοθετηθεί από την αθλήτρια, όταν αγωνίζεται είναι η ελαφρώς σκυφτή στάση μπροστά, με λυγισμένα τα γόνατα (γωνίες της άρθρωσης των γονάτων μεταξύ 90° και 150°) για χαμηλό κέντρο βάρους.

Ο χρόνος εκτέλεσης της τεχνικής νίκης είναι καθοριστικός στη διαμόρφωση της προπόνησης. Μεγάλη διαφορά έχει όταν η νίκη γίνεται στο πρώτο λεπτό και συγκεκριμένα στα πρώτα δέκα δευτερόλεπτα του πρώτου λεπτού ή όταν γίνεται στο δεύτερο, τρίτο, τέταρτο ή στον επιπλέον χρόνο του αγώνα γιατί διαφορετικοί μηχανισμοί παραγωγής ενέργειας δραστηριοποιούνται για τη συνέχιση της προσπάθειας. Σύμφωνα με τον Κλεισούρα (1989 & 2001), στην παραγωγή μυϊκής ενέργειας συμμετέχουν οι ενεργειακοί μηχανισμοί ανάλογα με την ένταση και τη διάρκεια της άσκησης. Συνδυαστικά, τα αποτελέσματα της παρούσας μελέτης για το χρόνο εκτέλεσης της τεχνικής με στόχο την επίτευξη της νίκης, η χρονική διάρκεια του αγώνα στο τζούντο και η φυσιολογία της άσκησης διαμορφώνουν το σχεδιασμό της προπόνησης της αθλήτριας σε έξι διαφορετικές κατηγορίες προπόνησης με τη συμμετοχή των ενεργειακών μηχανισμών. Κάτω υπό αυτές τις κατηγορίες άσκησης οι οποίες σχετίζονται με τους μηχανισμούς παραγωγής ενέργειας χρειάζεται η αθλήτρια να βελτιώσει τις συγκεκριμένες τεχνικές και τακτικές δεξιότητές της.

Η παρούσα εργασία ήταν μια πρωτότυπη προσπάθεια για τα ελληνικά επιστημονικά δεδομένα, η οποία συνδύασε την ποσοτική ανάλυση τεχνικών δεξιοτήτων κινήσεων - από μια σειρά παρατηρούμενων διά ζώσης και βιντεοσκοπημένων αγώνων τζούντο της κατηγορίας των 57 κιλών των γυναικών - με την εκπόνηση προπονητικού προγράμματος τζούντο με στόχο την κατάκτηση της πρώτης θέσης στην κατηγορία των 57 κιλών στο Πανελλήνιο πρωτάθλημα γυναικών 2018, όπου και επετεύχθη.

ABSTRACT

The purpose of the present work is to observe, record, analyze data on the racing activity of component athletes and to develop technical and tactical guidance and training. A protocol was developed for judo racing features (body posture, handles, grips, technique - tactics, hand, hip, and foot lever attacks, defenses, assault combinations, attack time, game side, etc.). The protocol was used for two years in national, school and international judo competitions (Balkan Championships) in which the athlete participated and was used in the same 57-Kg category without her participation in the European, World and Olympic Games. Content frequency analysis was performed this showed the foot, hip and hand techniques to be more frequent in achieving victory than judo techniques on the ground. The results showed that in the grounding techniques more immobilization techniques have appeared. Statistical analysis χ^2 and spearman p showed that there is a significant relationship between the techniques of achieving victory with the time of execution, with the way the opponent was caught and the posture used by the athlete. The results have formed the coaching-educational model of developing technical and tactical skills in order to tackle rival athletes successfully. The program was turned out to be successful with the first place in the category of 57kg women for 2018.

ΑΝΑΦΟΡΕΣ

- Bailey, C.A. (2007). *A guide to field research*, (2nd ed.). Thousand Oaks, CA: Pine Forge Press.
- Brandt, R., Bevilacqua, GG., Coimbra, DR., Pombo, LC., Miarka, B, & Lane, AM. (2018). Body Weight and Mood State Modifications in Mixed Martial Arts: An Exploratory Pilot. *Journal of Strength and Conditioning Research*, 32(9):2548 - 2554.

- Coswig, V., Fukuda, DH., Ramos, SP., Del Vecchio, FB. (2016). Biochemical Differences between Official and Simulated Mixed Martial Arts (Mma) Matches. *Asian Journal of Sports Medicine*, 7(2):e30950.
- Fischer, RA. (1954). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.
- Giovanni, M., Raffaele, S., Stefano, L., Gian Carlo, R. & Pietro Luigi, I. (2007). Technical and Tactical Differences between Regional and National Level Athletes in Kyourgi Taekwondo WTF, *12th Annual Congress of the ECSS, Jyväskylä, Finland*.
- James, LP., Robertson, S., Haff, GG., Beckman, EM., & Kelly, VG. (2017). Identifying the performance characteristics of a winning outcome in elite mixed martial arts competition. *Journal of Science Medicine and Sport*, 20(3):296-301.
- Kajmovic H., Radjo I., Mekic A. & Kapo S. (2008). Differences of Winning Techniques and Time Duration of Fight between Male and Female Judo Competitors at the Universiade - 2007, *World Congress of Performance Analysis in Sport VIII*, University Magdeburg, Germany, pp.614- 620.
- Karpman S, Reid P, Phillips L, Qin Z, Gross DP. (2016). Combative Sports Injuries: An Edmonton Retrospective. *Clinical Journal of Sport Medicine*, 26(4):332-334.
- Koropanovski N., Jovanovic S. & Dopsaj. (2008). Characteristics of pointing actions of top level female competitors in karate, *World Congress of Performance Analysis of Sport VIII, University Magdeburg, Germany*, pp.386- 392.
- Link, D. & Lames, M. (2003). Application of an Internet Based, Collaborative Software Environment in Training and Competition of Top level Beach Volleyball, *International Journal Computer Science Sport*, Vol. 2, Nr. 1.
- McArdle, w., Katch, F., & Katch, V. (2001). *Φυσιολογία της Άσκησης*. (Τόμος Β), Β., Κλεισούρας (επιμ.). Αθήνα: Ιατρικές Εκδόσεις Π.Χ. Πασχαλίδης.
- Miarka B, Coswig V, Brito CJ, Slimani M, Amtmann J, Del Vecchio, FB. (2016). Comparison of Combat Outcomes: Technical and Tactical Analysis of Female MMA. *International Journal of Performance Analysis in Sport*, 16:1.
- Miarka B, Cury R, Julianetti R, Battazza R., Julio UF, Calmet M, Franchini E. (2014). A comparison of time-motion and technical-tactical variables between age groups of female judo matches. *Journal of Sports Science*. 2014; 32 (16): 1529-1538.
- Miarka, B, Fukuda, DH, Del Vecchio, FB, & Franchini, E. (2018). Discriminant Analysis of Technical-Tactical Actions in High-Level Judo Athletes. *International Journal of Performance Analysis in Sport*, 16:30.
- Santos, DAD., Miarka, B., Bello, FD., Queiroz, ACC., de Carvalho, PHB., Brito CJ., & Beneke R. (2018). 10-Years on Time-Motion and Motor Actions of Paired Mixed Martial Arts Athletes. *International Journal of Sports Physiology Performance*, 14(3):399-402.
- Schnell, R., Hill, P., & Esser, E. (2014). *Μέθοδοι Εμπειρικής Έρευνας*. Ν., Ναγόπουλος, & Γ., Γκικόσος, Αθήνα: Προπομπός.
- Shrout, P., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420-428.
- Tsamourtzis, E. & Athanasiou, N. (2004). Registration of rebound possession zones in basketball, *International Journal of Performance Analysis in Sport, Volume 4(6)*, Number 1, pp. 34-39.
- Αρματάς, Β., Γιαννακός, Α. και Χατζημανουήλ, Δ., (2007). Καταγραφή και Αξιολόγηση των Επιτυχημένων Σημμένων Φάσεων στο Ευρωπαϊκό Πρωτάθλημα Ποδοσφαίρου στην Πορτογαλία 2004. *Αναζητήσεις στη Φυσική Αγωγή & τον Αθλητισμό*, 5(2):302-307.
- Ζάραγκας, Χ. (1993 - 1999). *Σημειώσεις μαθήματος Τζούντο*, Τμήματα Αθλητικής Διευκόλυνσης του 3^{ου} Γυμνασίου Ιωαννίνων και στα Ειδικά Τμήματα Αθλητικής Διευκόλυνσης του 5^{ου} Λυκείου Ιωαννίνων, Ιωάννινα.
- Κλεισούρας, Β. (1989). *Εργοφυσιολογία. Φυσιολογική βάση της μυϊκής προσπάθειας*. Αθήνα: Εκδ. Γ., Παρισιάνος.
- Σφίγγος, Ν., Τσιμπήρης, Α. και Τσαμουρτζής, Ε. (2001). Ψηφιακή βιντεοανάλυση στο ποδόσφαιρο, 9^ο Διεθνές Συνέδριο Φυσικής Αγωγής & Αθλητισμού, Κομοτηνή.

Μοντέλο χώρου καταστάσεων με εκθετικού τύπου διάνυσμα κατάστασης και γραμμική εξίσωση παρατηρήσεων

Ο. Θεοδοσιάδου¹, Γ. Τσακλίδης¹

¹Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
{outheod, tsaklidi}@math.auth.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία προτείνεται ένα στοχαστικό μη γραμμικό μοντέλο χώρου καταστάσεων για την εκτίμηση των κρυφών αλμάτων των αποδόσεων χρηματοοικονομικών μεγεθών. Τα θετικά και αρνητικά άλματα, που μοντελοποιούνται ως κρυφές τυχαίες μεταβλητές, υπόκεινται σε μη αρνητικούς περιορισμούς. Για τον λόγο αυτό, το θετικό και αρνητικό άλμα μοντελοποιείται ως εκθετικού τύπου συνάρτηση στην εξίσωση των καταστάσεων (άλματα), ενώ η εξίσωση των παρατηρήσεων (αποδόσεις) είναι γραμμική. Για την εξαγωγή των αναδρομικών τύπων εκτίμησης των καταστάσεων γίνεται γραμμικοποίηση των μη γραμμικών συναρτήσεων μέσω του αναπτύγματος Taylor. Ακολουθεί εφαρμογή της προτεινόμενης μοντελοποίησης σε εμπειρικά δεδομένα. Επιπλέον, με δεδομένο ότι στο προτεινόμενο μοντέλο οι διασπορές στην εξίσωση των καταστάσεων και των παρατηρήσεων είναι αμετάβλητες στον χρόνο, γίνεται εκτίμηση των σημείων αλλαγής ως προς τη διασπορά στη χρονοσειρά των δεδομένων και εκτιμώνται οι παράμετροι του μοντέλου σε καθένα από τα τμήματα που δημιουργούνται με βάση τα εκτιμώμενα σημεία αλλαγής.

Λέξεις κλειδιά: επεκτεταμένο φίλτρο Kalman, κρυφά άλματα αποδόσεων, σημεία αλλαγής

1. ΕΙΣΑΓΩΓΗ

Εμπειρικά έχει αποδειχθεί ότι μοντέλα που βασίζονται σε διαδικασίες διάχυσης δεν μπορούν να περιγράψουν επαρκώς την εξέλιξη των τιμών χρηματοοικονομικών μεγεθών. Αυτό οφείλεται στο γεγονός ότι στα μοντέλα διάχυσης θεωρείται ότι η διαδικασία εξέλιξης των τιμών συμπεριφέρεται τοπικά ως μια κίνηση Brown με αποτέλεσμα να θεωρείται εν γένει μικρή η πιθανότητα παραγωγής μιας μεγάλης κατά μέτρο (αρνητικής ή θετικής) τιμής σε μια μικρή χρονική περίοδο (Black and Scholes, 1973). Αντιθέτως, αν θεωρήσουμε ότι οι τιμές των μετοχών περιλαμβάνουν και άλματα, τότε η πιθανότητα για μια απότομη αλλαγή στη χρονική εξέλιξη των τιμών δεν είναι πλέον αμελητέα. Οι Cox and Ross, 1976 υποστηρίζουν ότι οι θεωρίες που αναπτύχθηκαν για την εξέλιξη των τιμών των μετοχών κατά συνεχή τρόπο πρέπει να τροποποιηθούν, ώστε να συμπεριλάβουν και την έκθεση στον κίνδυνο των αλμάτων.

Οι διάφορες τροποποιήσεις του μοντέλου Black-Scholes οδήγησαν σε μια πληθώρα συγγραμμάτων αναφορικά με τη μοντελοποίηση των τιμών χρηματοοικονομικών μεγεθών. Για παράδειγμα, οι Merton, 1976; Duffie et al., 2000 και Kou, 2002 προτείνουν μοντέλα που συνδυάζουν τις διαδικασίες αλμάτων με τις διαδικασίες διάχυσης, ενώ οι Schoutens, 2003; Geman et al., 2001 και Madan, 2001 εστιάζουν στη χρήση διαδικασιών Lévy.

Τα άλματα μοντελοποιούνται ως κρυφές τυχαίες μεταβλητές (τ.μ.) (βλ. π.χ. Polimenis, 2012 και Theodosiadou and Tsaklidis, 2016), καθώς δεν εντοπίζονται άμεσα στα δείγματα που λαμβάνονται σε διακριτούς χρόνους, και χωρίζονται σε δυο κατηγορίες: στα θετικά άλματα (προς τα πάνω), που δεχόμαστε ότι προκαλούνται από την άφιξη των θετικών ειδήσεων στην αγορά, και στα αρνητικά άλματα (προς τα κάτω) που προκαλούνται από την άφιξη των αρνητικών ειδήσεων.

Καθότι τα άλματα των αποδόσεων μοντελοποιούνται ως κρυφές τ.μ., στην εργασία Theodosiadou and Tsaklidis, 2017 κατασκευάζεται ένα κατάλληλο μοντέλο χώρου καταστάσεων για την εκτίμησή τους. Σε αυτήν την περίπτωση υιοθετείται ένα γραμμικό Μαρκοβιανό μοντέλο για την εξίσωση των καταστάσεων (αλμάτων), ενώ η εξίσωση των παρατηρήσεων δίνεται ως η διαφορά ανάμεσα στο θετικό και αρνητικό άλμα, υπό την επίδραση λευκού θορύβου. Με δεδομένο ότι η χρήση του γραμμικού φίλτρου Kalman (Kalman, 1960) δεν μπορεί να εγγυηθεί την ικανοποίηση των μη-αρνητικών περιορισμών για τις εκτιμήσεις των κρυφών αλμάτων κάθε χρονική στιγμή, προτείνεται για τον σκοπό αυτό μια μη-γραμμική μέθοδος, η μέθοδος 3-σταδίων, που αποτελείται από τα ακόλουθα στάδια: (α) χρήση του φίλτρου Kalman, (β) αποκοπή της διδιάστατης δεσμευμένης κατανομής του διανύσματος κατάστασης, και (γ) διαδικασία κλιμάκωσης. Επιπλέον, το προτεινόμενο μοντέλο των 3-σταδίων είναι ομογενές ως προς τον χρόνο με την έννοια ότι οι διασπορές στις εξισώσεις των καταστάσεων και των παρατηρήσεων είναι σταθερές στον χρόνο.

Στόχο της παρούσης εργασίας αποτελεί η κατασκευή ενός μη γραμμικού μοντέλου χώρου καταστάσεων, για την εκτίμηση των κρυφών θετικών και αρνητικών αλμάτων των αποδόσεων χρηματοοικονομικών μεγεθών, όπου να προκύπτουν άμεσα μη αρνητικές εκτιμήσεις, συγκριτικά με την εργασία Theodosiadou and Tsaklidis, 2017, όπου εφαρμόζεται η μέθοδος των 3-σταδίων. Για τον σκοπό αυτό, το αρνητικό και θετικό άλμα δίνονται ως εκθετικού τύπου συναρτήσεις, που ικανοποιούν τη Μαρκοβιανή ιδιότητα, ενώ στην εξίσωση των παρατηρήσεων, η απόδοση ορίζεται ως η διαφορά ανάμεσα στα άλματα των δυο κατευθύνσεων. Ωστόσο, οι αποδόσεις χρηματοοικονομικών μεγεθών εμφανίζουν εμπειρικά το χαρακτηριστικό της δεσμευμένης ως προς τον χρόνο ετεροσκεδαστικότητας (βλ. π.χ. Cont, 2001). Εν τούτοις, το προτεινόμενο μη γραμμικό μοντέλο είναι ομογενές ως προς τον χρόνο, με την έννοια ότι οι διασπορές των θορύβων στις εξισώσεις των καταστάσεων και της παρατήρησης είναι αμετάβλητες στο χρόνο. Για τον λόγο αυτό, προτού γίνει η εφαρμογή της προτεινόμενης μοντελοποίησης για την εκτίμηση των κρυφών αλμάτων, πραγματοποιείται εντοπισμός των σημείων αλλαγής ως προς τη διασπορά στα εμπειρικά δεδομένα που χρησιμοποιούνται. Ο έλεγχος που χρησιμοποιείται για τον εντοπισμό των σημείων

αλλαγής περιγράφεται στην εργασία Matteson and James, 2014. Με αυτόν τον τρόπο, το προτεινόμενο μη γραμμικό μοντέλο χώρου καταστάσεων εφαρμόζεται σε καθένα από τα τμήματα της χρονοσειράς των αποδόσεων που δημιουργείται με βάση τα σημεία αλλαγής που εκτιμώνται. Τα δεδομένα που χρησιμοποιούνται αφορούν στις αποδόσεις του δείκτη NASDAQ για την περίοδο 2006-2008 (και είναι διαθέσιμα στην ιστοσελίδα finance.yahoo.com).

Δεδομένου ότι το μοντέλο χώρου καταστάσεων που κατασκευάζεται είναι μη γραμμικό, δεν μπορεί να χρησιμοποιηθεί ο αλγόριθμος του φίλτρου Kalman για την εκτίμηση του διανύσματος κατάστασης. Για τον λόγο αυτό, οι αναδρομικές σχέσεις εκτίμησης του διανύσματος κατάστασης και του πίνακα διασπορών-συνδιασπορών των σφαλμάτων της εκτίμησης προκύπτουν μέσω γραμμικοποίησης των μη γραμμικών συναρτήσεων των εξισώσεων του μοντέλου γύρω από την εκτίμηση που προκύπτει από το κλασικό φίλτρο Kalman με τη χρήση του αναπτύγματος Taylor (βλ. για παράδειγμα Simon, 2006 και Durbin and Koopman, 2012). Επιπλέον, η προτεινόμενη μοντελοποίηση μπορεί να χρησιμοποιηθεί γενικότερα σε κάθε περίπτωση όπου οι τιμές του διανύσματος κατάστασης, που είναι προς εκτίμηση, ανήκουν σε ένα αριστερά φραγμένο χωρίο με βάση το φυσικό τους νόημα.

Η εργασία οργανώνεται ως εξής:

Στην ενότητα 2 κατασκευάζεται το μη-γραμμικό μοντέλο χώρου καταστάσεων που χρησιμοποιείται για την εκτίμηση των κρυφών αλμάτων. Στην ενότητα 3 εξάγονται οι αναδρομικές σχέσεις που θα χρησιμοποιηθούν για την εκτίμηση του διανύσματος κατάστασης. Στην ενότητα 4 γίνεται εφαρμογή της προτεινόμενης μοντελοποίησης στις αποδόσεις του δείκτη NASDAQ, όπου και ελέγχεται η προσαρμογή του μοντέλου στα εμπειρικά δεδομένα με βάση την τιμή του μέσου τετραγωνικού σφάλματος (MSE). Τέλος, στην ενότητα 5 παραθέτονται τα συμπεράσματα που προέκυψαν από την προτεινόμενη μοντελοποίηση, και γίνεται αναφορά σε μελλοντικές επεκτάσεις.

2. ΤΟ ΜΟΝΤΕΛΟ

Θεωρούμε ότι η απόδοση ενός περιουσιακού στοιχείου τη χρονική στιγμή t δίνεται από τη σχέση

$$R_t = X_t - Y_t + e_t, \quad (1)$$

όπου,

- R_t : η απόδοση τη χρονική στιγμή t ,
- X_t : το θετικό (προς τα πάνω) άλμα της απόδοσης τη χρονική στιγμή t ,
- Y_t : το αρνητικό (προς τα κάτω) άλμα της απόδοσης τη χρονική στιγμή t ,
- e_t : θόρυβος για τον οποίο θεωρούμε ότι $e_t \sim N(0, \sigma^2)$.

Στη σχέση (1), το μοναδικό παρατηρήσιμο μέγεθος είναι η απόδοση R_t . Τα άλματα X και Y μοντελοποιούνται ως κρυφές τυχαίες μεταβλητές. Συνεπώς, προκειμένου να γίνει εκτίμηση-«αποκάλυψη» της τιμής των κρυφών αλμάτων X_t, Y_t για κάθε χρονική στιγμή, κατασκευάζεται ένα κατάλληλο μοντέλο χώρου καταστάσεων, όπου το διάνυσμα κατάστασης αποτελείται από τις δυο συνιστώσες των αλμάτων. Επιπλέον, το

διάνυσμα κατάστασης υπόκειται σε ανισοτικούς περιορισμούς, καθώς τα εκτιμώμενα κρυφά άλματα που θα προκύψουν από τη μοντελοποίηση πρέπει να ικανοποιούν τη σχέση (1) υπό τον περιορισμό $X_t, Y_t \geq 0$, $t=1,2,\dots$.

Για τον λόγο αυτό, κατασκευάζεται ένα στοχαστικό μοντέλο χώρου καταστάσεων, όπου στην εξίσωση των καταστάσεων κάθε συνιστώσα του διανύσματος κατάστασης δίνεται ως μια εκθετικού τύπου συνάρτηση που ικανοποιεί τη Μαρκοβιανή ιδιότητα, ενώ η εξίσωση των παρατηρήσεων είναι γραμμική. Πιο συγκεκριμένα:

Οι εξισώσεις των καταστάσεων (άλματα) δίνονται από το σύστημα

$$\left. \begin{aligned} X_t &= e^{\alpha_1 + g_{11}X_{t-1} + g_{12}Y_{t-1} + w_{t-1,1}} \\ Y_t &= e^{\alpha_2 + g_{21}X_{t-1} + g_{22}Y_{t-1} + w_{t-1,2}} \end{aligned} \right\}$$

ή ισοδύναμα

$$z_{t,i} = e^{\alpha_i + \sum_{j=1}^2 g_{ij}z_{t-1,j} + w_{t-1,i}}, \quad i = 1, 2, \quad (2)$$

όπου,

- $\mathbf{z}_t = (z_{t,1}, z_{t,2})^T = (X_t, Y_t)^T$ το διάνυσμα κατάστασης (2×1),
- $g_{ij}, \alpha_i \in \mathbb{R}$, $i, j = 1, 2$,
- $\mathbf{w}_t = \begin{pmatrix} w_{1,t} \\ w_{2,t} \end{pmatrix}$ το διάνυσμα θορύβου για το οποίο ισχύει $\mathbf{w}_t \sim N(\mathbf{0}, \mathbf{Q} = \begin{pmatrix} q_1^2 & 0 \\ 0 & q_2^2 \end{pmatrix})$.

Η εξίσωση των παρατηρήσεων (αποδόσεων) δίνεται από τη σχέση (1) ή ισοδύναμα σε μορφή πινάκων

$$R_t = \mathbf{H}\mathbf{z}_t + e_t, \quad (3)$$

όπου $\mathbf{H} = \begin{pmatrix} 1 & -1 \end{pmatrix}$. Επιπλέον, θεωρούμε ότι οι θόρυβοι \mathbf{w}_t και e_t είναι ασυσχέτιστοι μεταξύ τους, δηλαδή $\mathbb{E}(e_k \mathbf{w}_j^T) = \mathbf{0}$.

Παίρνοντας τους λογαρίθμους στη σχέση (2) προκύπτει ότι

$$\log(z_{t,i}) = \alpha_i + \sum_{j=1}^2 g_{ij}z_{t-1,j} + w_{t-1,i}, \quad i = 1, 2,$$

και θέτοντας $\log(z_{t,i}) = \zeta_{t,i}$, $i = 1, 2$, όπου $z_{t,i} > 0$, το μοντέλο χώρου καταστάσεων (2)-(3) για την εκτίμηση των κρυφών αλμάτων X_t, Y_t των αποδόσεων μετασχηματίζεται ως εξής:

Η εξίσωση των καταστάσεων δίνεται από τη σχέση

$$\zeta_t = \begin{pmatrix} \zeta_{t,1} \\ \zeta_{t,2} \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix} \begin{pmatrix} e^{\zeta_{t-1,1}} \\ e^{\zeta_{t-1,2}} \end{pmatrix} + \begin{pmatrix} w_{1,t-1} \\ w_{2,t-1} \end{pmatrix},$$

ή ισοδύναμα με τη μορφή πινάκων

$$\zeta_t = \mathbf{A} + \mathbf{G}\mathbf{E}_{t-1}(\zeta_{t-1}) + \mathbf{w}_t \quad (4)$$

όπου,

- $\zeta_{t,1} = \log(X_t)$ και $\zeta_{t,2} = \log(Y_t)$, με $X_t, Y_t > 0$,
- $\mathbf{A} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$, $\mathbf{G} = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}$, $\mathbf{E}_t(\zeta_t) = \begin{pmatrix} e^{\zeta_{t,1}} \\ e^{\zeta_{t,2}} \end{pmatrix}$,
- $\mathbf{w}_t = \begin{pmatrix} w_{1,t} \\ w_{2,t} \end{pmatrix} \sim N(\mathbf{0}, \mathbf{Q} = \begin{pmatrix} q_1^2 & 0 \\ 0 & q_2^2 \end{pmatrix})$.

Η εξίσωση των παρατηρήσεων δίνεται από τη σχέση

$$R_t = \mathbf{H}\mathbf{E}_t(\zeta_t) + e_t, \quad (5)$$

όπου $\mathbf{H} = (1 \quad -1)$ και $e_t \sim N(0, \sigma^2)$.

Παρατήρηση 1. Σημειώνεται ότι η υιοθέτηση συναρτήσεων εκθετικού τύπου για την περιγραφή της δυναμικής εξέλιξης των κρυφών θετικών και αρνητικών αλμάτων στις αποδόσεις συμβαδίζει με τη θεώρηση της υιοθέτησης Γεωμετρικής κίνησης *Brown* για τη μοντελοποίηση των τιμών χρηματοοικονομικών μεγεθών.

3. Εκτίμηση των κρυφών καταστάσεων

Το προτεινόμενο μοντέλο χώρου καταστάσεων (4)-(5) για την εκτίμηση των κρυφών αλμάτων αποδόσεων είναι μη γραμμικό. Συνεπώς, δε μπορεί να χρησιμοποιηθεί ο αλγόριθμος του φίλτρου Kalman για την εκτίμηση του διανύσματος κατάστασης. Για τον λόγο αυτό, οι αναδρομικές σχέσεις εκτίμησης του διανύσματος κατάστασης και του πίνακα διασπορών-συνδιασπορών των σφαλμάτων της εκτίμησης προκύπτουν μέσω γραμμικοποίησης των μη γραμμικών συναρτήσεων των εξισώσεων του μοντέλου γύρω από την εκτίμηση που προκύπτει από το κλασικό φίλτρο Kalman με τη χρήση του αναπτύγματος Taylor. Για τον σκοπό αυτό, στη συνέχεια γίνεται χρήση των παρακάτω συμβολισμών:

- $\hat{\zeta}_t(-)$: η εκ των προτέρων εκτίμηση του διανύσματος κατάστασης ζ_t , θεωρώντας ότι η απόδοση R_t είναι γνωστή μέχρι (και) τη χρονική στιγμή $t - 1$,
- $\hat{\zeta}_t(+)$: η εκ των υστέρων εκτίμηση του διανύσματος κατάστασης ζ_t , θεωρώντας ότι η απόδοση R_t είναι γνωστή μέχρι τη χρονική στιγμή t ,

- $\mathbf{P}_t(-), \mathbf{P}_t(+)$: οι πίνακες διασπορών-συνδιασπορών των σφαλμάτων εκτίμησης για την εκ των προτέρων και εκ των υστέρων εκτίμηση του διανύσματος κατάστασης ζ_t αντίστοιχα τη χρονική στιγμή t ,
- $\mathbf{E}_t^-(\hat{\zeta}_t) = (e^{\hat{\zeta}_{t,1}(-)}, e^{\hat{\zeta}_{t,2}(-)})^T$ διάνυσμα στήλη (2×1),
- $diag(\mathbf{E}_t^-(\hat{\zeta}_t)) = diag(e^{\hat{\zeta}_{t,1}(-)}, e^{\hat{\zeta}_{t,2}(-)})$ διαγώνιος πίνακας (2×2) με τα στοιχεία του διανύσματος $\mathbf{E}_t^-(\hat{\zeta}_t)$ στην κύρια διαγώνιο,
- $\mathbf{E}_t^+(\hat{\zeta}_t) = (e^{\hat{\zeta}_{t,1}(+)}, e^{\hat{\zeta}_{t,2}(+)})^T$ διάνυσμα στήλη (2×1),
- $diag(\mathbf{E}_t^+(\hat{\zeta}_t)) = diag(e^{\hat{\zeta}_{t,1}(+)}, e^{\hat{\zeta}_{t,2}(+)})$ διαγώνιος πίνακας (2×2) με τα στοιχεία του διανύσματος $\mathbf{E}_t^+(\hat{\zeta}_t)$ στην κύρια διαγώνιο.

Επιπλέον, για τους εκτιμητές $\hat{\zeta}_t(-)$ και $\hat{\zeta}_t(+)$ του διανύσματος κατάστασης ζ_t θεωρούμε ότι

$$\hat{\zeta}_t(-) = \mathbb{E}[\zeta_t | \mathbf{R}_{t-1}] \quad \text{και} \quad \mathbf{P}_t(-) = \mathbb{E}[(\zeta_t - \hat{\zeta}_t(-))(\zeta_t - \hat{\zeta}_t(-))^T]$$

$$\hat{\zeta}_t(+) = \mathbb{E}[\zeta_t | \mathbf{R}_t] \quad \text{και} \quad \mathbf{P}_t(+) = \mathbb{E}[(\zeta_t - \hat{\zeta}_t(+))(\zeta_t - \hat{\zeta}_t(+))^T].$$

Για την εξαγωγή των αναδρομικών τύπων για τους εκτιμητές $\hat{\zeta}_t(-)$ και $\hat{\zeta}_t(+)$ και τους αντίστοιχους πίνακες διασπορών-συνδιασπορών $\mathbf{P}_t(-), \mathbf{P}_t(+)$ με βάση το μοντέλο (4)-(5) χρησιμοποιείται το ανάπτυγμα Taylor για τη συνάρτηση $e^{\zeta_{t,i}}$ γύρω από τα σημεία $\hat{\zeta}_{t,i}(-)$ και $\hat{\zeta}_{t,i}(+)$, $i = 1, 2$ και η γραμμική προσέγγιση οδηγεί στις σχέσεις

$$e^{\zeta_{t,i}} \approx e^{\hat{\zeta}_{t,i}(-)} + e^{\hat{\zeta}_{t,i}(-)}(\zeta_{t,i} - \hat{\zeta}_{t,i}(-)), \quad i = 1, 2, \quad (6)$$

και αντίστοιχα

$$e^{\zeta_{t,i}} \approx e^{\hat{\zeta}_{t,i}(+)} + e^{\hat{\zeta}_{t,i}(+)}(\zeta_{t,i} - \hat{\zeta}_{t,i}(+)), \quad i = 1, 2. \quad (7)$$

Επομένως, από τις σχέσεις (6) και (7) προκύπτει ότι

$$\mathbb{E}[e^{\zeta_{t,i}} | \mathbf{R}_{t-1}] \approx e^{\hat{\zeta}_{t,i}(-)}, \quad i = 1, 2, \quad (8)$$

και αντίστοιχα

$$\mathbb{E}[e^{\zeta_{t,i}} | \mathbf{R}_t] \approx e^{\hat{\zeta}_{t,i}(+)}, \quad i = 1, 2. \quad (9)$$

Μετά από υπολογισμούς βασισμένους στις σχέσεις (6)-(9) προκύπτουν ακολούθως οι αναδρομικές σχέσεις της εκ των προτέρων και υστέρων εκτίμησης του διανύσματος κατάστασης ζ_t και των αντίστοιχων πινάκων διασπορών-συνδιασπορών $\mathbf{P}_t(-), \mathbf{P}_t(+)$:

$$\hat{\zeta}_t(-) \approx \mathbf{A} + \mathbf{G}[\mathbf{M} + \mathbf{E}_{t-1}^+(\hat{\zeta}_{t-1})] \quad (\text{εκ των προτέρων εκτίμηση})$$

$$\mathbf{P}_t(-) = \mathbf{G} \text{diag}(\mathbf{E}_{t-1}^+(\hat{\zeta}_{t-1})) \mathbf{P}_{t-1}(+) \text{diag}(\mathbf{E}_{t-1}^+(\hat{\zeta}_{t-1})) \mathbf{G}^T + \mathbf{Q} \quad (\text{πίνακας διασπορών-συνδιασπορών})$$

$$\hat{\zeta}_t(+) \approx \hat{\zeta}_t(-) + \mathbf{D}_t \mathbf{S}_t^{-1} (\mathbf{r}_t - \mathbb{E}[\mathbf{r}_t | \mathbf{r}_{t-1}]) \quad (\text{εκ των υστέρων εκτίμηση})$$

$$\mathbf{P}_t(+) = \mathbf{P}_t(-) - \mathbf{D}_t \mathbf{S}_t^{-1} \mathbf{D}_t^T \quad (\text{πίνακας διασπορών-συνδιασπορών})$$

όπου,

- $\mathbb{E}[\mathbf{r}_t | \mathbf{r}_{t-1}] \approx \mathbf{B} + \mathbf{H}[\mathbf{M} + \mathbf{E}_t^-(\hat{\zeta}_t)]$
- $\mathbf{D}_t = \mathbf{P}_t(-) \text{diag}(\mathbf{E}_t^-(\hat{\zeta}_t)) \mathbf{H}^T$
- $\mathbf{S}_t = \mathbf{H} \text{diag}(\mathbf{E}_t^-(\hat{\zeta}_t)) \mathbf{P}_t(-) \text{diag}(\mathbf{E}_t^-(\hat{\zeta}_t)) \mathbf{H}^T + \sigma^2$

Η εκτίμηση του συνόλου $(\mathbf{A}, \mathbf{G}, \mathbf{Q}, \sigma^2)$ των παραμέτρων του μοντέλου (4)-(5), προκειμένου να υπολογιστούν οι αναδρομικές σχέσεις που περιγράφονται παραπάνω, γίνεται με μέθοδο της Μέγιστης Πιθανοφάνειας (ΜΠ), όπου η δεσμευμένη κατανομή της τ.μ. R_t με δεδομένη την ζ_{t-1} είναι κανονική, δηλαδή

$$R_t | \zeta_{t-1} \sim N(\mathbb{E}[R_t | R_{t-1}], \mathbf{S}_t) .$$

Επομένως, η συνάρτηση του λογαρίθμου της πιθανοφάνειας (LogL) έχει τη μορφή

$$\text{LogL}(R_1, \dots, R_n) = -n/2 \log(2\pi) - 0.5 \sum_{t=1}^n (\log(|\mathbf{S}_t|) + \mathbb{E}[R_t | R_{t-1}]^T \mathbf{S}_t^{-1} \mathbb{E}[R_t | R_{t-1}]), \quad (10)$$

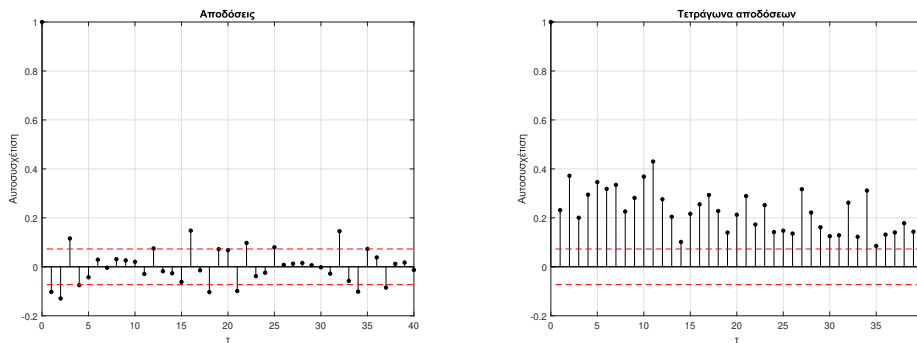
$t = 1, 2, \dots, n$, όπου n το πλήθος των παρατηρήσεων.

4. Εφαρμογή: Εκτίμηση των κρυφών αλμάτων των αποδόσεων για τον δείκτη NASDAQ

Ακολουθεί εφαρμογή της μοντελοποίησης που αναπτύσσεται στην ενότητα 3 για την εκτίμηση των κρυφών θετικών και αρνητικών αλμάτων των αποδόσεων του δείκτη NASDAQ για την περίοδο 2006-2008. Σημειώνεται ότι στο μοντέλο (4)-(5) ο πίνακας διασπορών - συνδιασπορών του θορύβου \mathbf{w}_t στην εξίσωση των καταστάσεων, καθώς και η διασπορά σ^2 του θορύβου e_t στην εξίσωση των παρατηρήσεων, δεν μεταβάλλονται στον χρόνο. Ωστόσο, έχει αποδειχθεί εμπειρικά ότι οι χρονοσειρές των αποδόσεων χρηματοοικονομικών μεγεθών εμφανίζουν το χαρακτηριστικό της

δεσμευμένης ως προς τον χρόνο ετεροσκεδαστικότητας (βλ. π.χ. Cont, 2001). Στην περίπτωση των αποδόσεων του δείκτη NASDAQ, η ύπαρξη των μεταβλητών διασπορών επιβεβαιώνεται μέσω του γραφήματος των αυτοσυσχετίσεων που δίνονται στο Σχήμα 1, καθότι τα τετράγωνα των αποδόσεων παρουσιάζουν ισχυρές αυτοσυσχετίσεις. Επομένως, η διασπορά των αποδόσεων μπορεί να θεωρηθεί ότι είναι μεταβλητή στον χρόνο και εξαρτάται από τις προηγούμενες τιμές της.

Σχήμα 1: (α') Συνάρτηση αυτοσυσχέτισης των αποδόσεων του δείκτη NASDAQ για την περίοδο 2006-2008, (β') Συνάρτηση αυτοσυσχέτισης των τετραγώνων των αποδόσεων του δείκτη NASDAQ για την περίοδο 2006-2008.



Για τον λόγο αυτό, γίνεται εντοπισμός των σημείων αλλαγής στη χρονοσειρά των αποδόσεων του δείκτη NASDAQ για την περίοδο 2006-2008. Για τον σκοπό αυτό, επιλέχθηκε η μη παραμετρική μέθοδος που περιγράφεται στην εργασία Matteson and James (2014), καθώς και στο υπολογιστικό πακέτο «ecp» της R. Η μηδενική υπόθεση αφορά στη μη ύπαρξη επιπλέον σημείων αλλαγής στα επιμέρους τμήματα της χρονοσειράς που προοδευτικά δημιουργούνται. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 1 σε στάθμη σημαντικότητας 0.05 και στο Σχήμα 2.

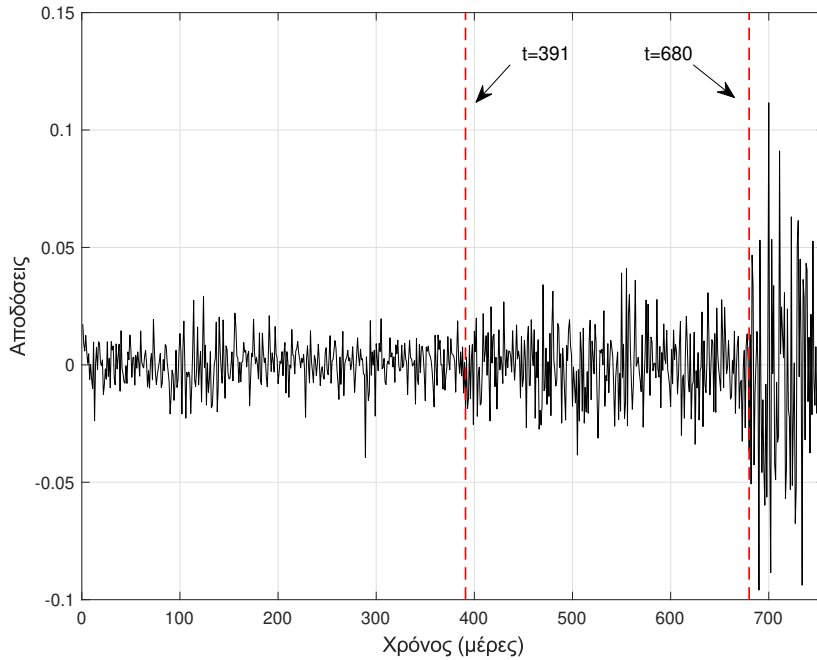
Πίνακας 1: Σημεία αλλαγής για τις αποδόσεις του δείκτη NASDAQ την περίοδο 2006-2008.

t (μέρες)	Ημερομηνία	p-τιμή
391	20/7/2007	0.0017
680	14/9/2008	0.0349

Συνεπώς, λαμβάνοντας υπόψη το Σχήμα 2, η χρονοσειρά των αποδόσεων του δείκτη NASDAQ χωρίζεται σε τρία επιμέρους τμήματα με βάση τη μεταβαλλόμενη διασπορά. Το αποτέλεσμα αυτό οδηγεί στην εκτίμηση των παραμέτρων του μοντέλου (4)-(5) για καθένα από τα τρία τμήματα στα οποία διαχωρίστηκε η χρονοσειρά.

Η εκτίμηση των παραμέτρων του μοντέλου για καθένα από τα τρία τμήματα γίνεται με τη μέθοδο ΜΠ. Πρόκειται για μια διαδικασία βελτιστοποίησης, όπου καλύπτεται ένα χωρίο τιμών για τις παραμέτρους με ένα πυκνό πλέγμα από (1771561) 9-άδες

Σχήμα 2: Σημεία αλλαγής ως προς τη διασπορά για τις αποδόσεις του δείκτη NASDAQ τη



τιμών, υπολογίζεται η τιμή της συνάρτησης (10) για κάθε σημείο (9-άδα τιμών) του πλέγματος, και επιλέγονται τα 20 πρώτα σημεία που παράγουν τη μεγαλύτερη τιμή στη συνάρτηση πιθανοφάνειας (10). Εν συνεχεία, το καθένα από αυτά τα 20 σημεία λαμβάνεται ως αρχικό σημείο σε αλγόριθμο βελτιστοποίησης για την εύρεση του μεγίστου της (10), και καταγράφεται το σημείο που αντιστοιχεί στο μέγιστο. Η παραπάνω διαδικασία πραγματοποιήθηκε για καθένα από τα τρία τμήματα της χρονοσειράς.

Το 1ο μέρος αφορά στις αποδόσεις για $t = 1$ έως $t = 391$. Η εκτίμηση του συνόλου $\psi_1 = (\mathbf{A}_1, \mathbf{G}_1, \mathbf{Q}_1, \sigma_1^2)$ (9 παράμετροι) των παραμέτρων του μοντέλου (4)-(5) για τα παραπάνω δεδομένα γίνεται μεγιστοποιώντας τη συνάρτηση πιθανοφάνειας (10), και προκύπτουν τα εξής αποτελέσματα:

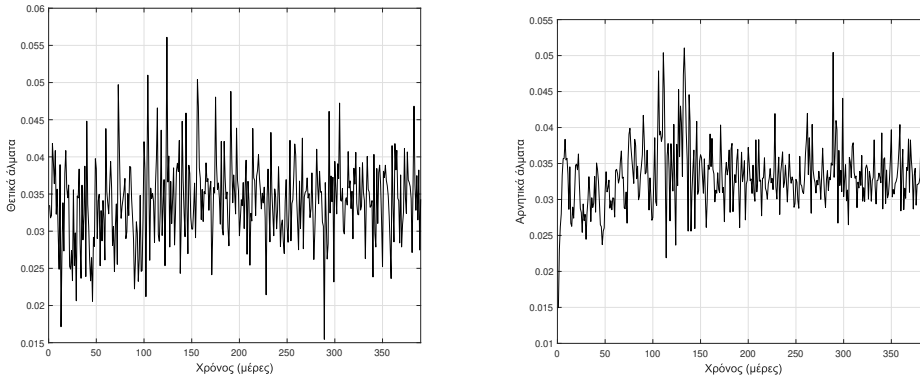
$$\mathbf{A}_1 = \begin{pmatrix} -3.87 \\ -4.55 \end{pmatrix}, \quad \mathbf{G}_1 = \begin{pmatrix} 6.85 & 7.48 \\ 14.45 & 19.36 \end{pmatrix}$$

και

$$\mathbf{Q}_1 = \begin{pmatrix} 0.051 & 0 \\ 0 & 4.64 \times 10^{-9} \end{pmatrix}, \quad \sigma_1^2 = 5.0875 \times 10^{-7}.$$

Εφαρμόζοντας τον προτεινόμενο αλγόριθμο για το μοντέλο (4)-(5), όπως περιγράφεται στην ενότητα 3, και λαμβάνοντας υπόψη στη συνέχεια τον μετασχηματισμό $\zeta_{t,1} = \log(X_t)$ και $\zeta_{t,2} = \log(Y_t)$, με $X_t, Y_t > 0$, προκύπτουν οι εκτιμήσεις των χρυ-

Σχήμα 3: (α') Οι εκτιμήσεις των θετικών αλμάτων του δείκτη NASDAQ από $t = 1$ έως $t = 391$, (β') Οι εκτιμήσεις των αρνητικών αλμάτων του δείκτη NASDAQ από $t = 1$ έως $t = 391$.



φών θετικών και αρνητικών αλμάτων στις αποδόσεις του δείκτη NASDAQ από $t = 1$ έως $t = 391$, οι οποίες παρουσιάζονται στο Σχήμα 3.

Το 2ο μέρος αφορά στις αποδόσεις από $t = 392$ έως $t = 680$. Η εκτίμηση του συνόλου $\psi_2 = (\mathbf{A}_2, \mathbf{G}_2, \mathbf{Q}_2, \sigma_2^2)$ (9 παράμετροι) των παραμέτρων του μοντέλου (4)-(5) για τα παραπάνω δεδομένα γίνεται μεγιστοποιώντας τη συνάρτηση πιθανοφάνειας (10), και προκύπτουν τα εξής αποτελέσματα:

$$\mathbf{A}_2 = \begin{pmatrix} -3.62 \\ -3.29 \end{pmatrix}, \quad \mathbf{G}_2 = \begin{pmatrix} -10.49 & 23.58 \\ -7.59 & 17.49 \end{pmatrix}$$

και

$$\mathbf{Q}_2 = \begin{pmatrix} 2.7418 \times 10^{-5} & 0 \\ 0 & 0.0132 \end{pmatrix}, \quad \sigma_2^2 = 1.1029 \times 10^{-5}.$$

Στη συνέχεια, εφαρμόζοντας τον προτεινόμενο αλγόριθμο για το μοντέλο (4)-(5) προκύπτουν οι εκτιμήσεις των κρυφών θετικών και αρνητικών αλμάτων στις αποδόσεις του δείκτη NASDAQ από $t = 392$ έως $t = 680$, οι οποίες παρουσιάζονται στο Σχήμα 4.

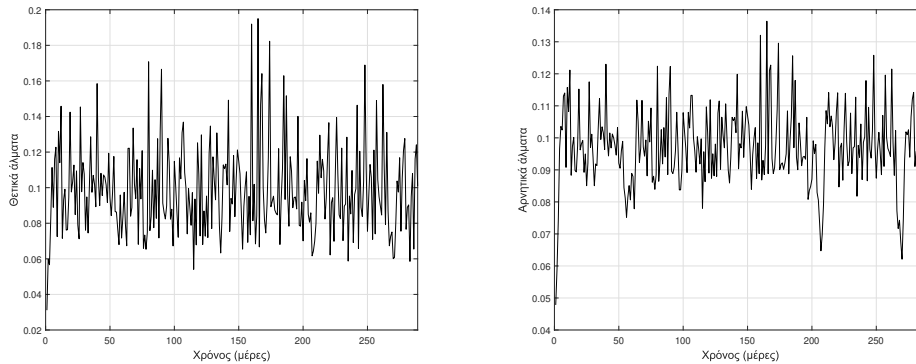
Τέλος, το 3ο μέρος αφορά στις αποδόσεις από $t = 681$ έως $t = 755$ και η εκτίμηση του συνόλου $\psi_3 = (\mathbf{A}_3, \mathbf{G}_3, \mathbf{Q}_3, \sigma_3^2)$ (9 παράμετροι) των παραμέτρων του μοντέλου (4)-(5) για τα παραπάνω δεδομένα μεγιστοποιώντας τη συνάρτηση πιθανοφάνειας (10) οδηγεί στα εξής αποτελέσματα:

$$\mathbf{A}_3 = \begin{pmatrix} -2.16 \\ -1.6 \end{pmatrix}, \quad \mathbf{G}_3 = \begin{pmatrix} -4. & 2.33 \\ -3.2 & -3.83 \end{pmatrix}$$

και

$$\mathbf{Q}_3 = \begin{pmatrix} 0.1621 & 0 \\ 0 & 4.2023 \times 10^{-9} \end{pmatrix}, \quad \sigma_3^2 = 2.9498 \times 10^{-9}.$$

Σχήμα 4: (α') Οι εκτιμήσεις των θετικών αλμάτων του δείκτη NASDAQ από $t = 392$ έως $t = 680$, (β') Οι εκτιμήσεις των αρνητικών αλμάτων του δείκτη NASDAQ από $t = 392$ έως $t = 680$.



Στη συνέχεια προκύπτουν οι εκτιμήσεις των κρυφών θετικών και αρνητικών αλμάτων στις αποδόσεις του δείκτη NASDAQ από $t = 681$ έως $t = 755$ και παρουσιάζονται στο Σχήμα 5.

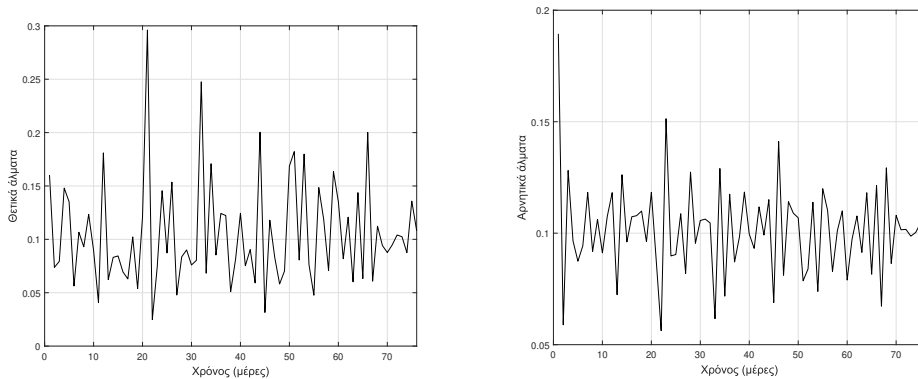
Έχοντας «αποκαλύψει» πλέον τα θετικά και αρνητικά άλματα του δείκτη NASDAQ για κάθε χρονικό βήμα, μπορούμε στη συνέχεια να εξετάσουμε την προσαρμογή των εκτιμήσεων των αποδόσεων, που προκύπτουν από τις διαφορές $\hat{X}_t - \hat{Y}_t$ των εκτιμώμενων αλμάτων, στις εμπειρικές αποδόσεις του δείκτη για καθένα από τα τρία μέρη της χρονοσειράς. Για τον σκοπό αυτό, παρατίθενται οι τιμές του MSE για τα τρία μέρη της χρονοσειράς των αποδόσεων, που υποδεικνύουν την ικανοποιητική προσαρμογή που επιτυγχάνεται.

Πίνακας 2: Τιμές MSE και RMSE για τις εκτιμήσεις των αποδόσεων του δείκτη NASDAQ για την περίοδο 2006-2008.

Μοντέλο	MSE	RMSE
1ο μέρος	5.7059×10^{-7}	7.5538×10^{-4}
2ο μέρος	1.3483×10^{-5}	0.0037
3ο μέρος	2.615×10^{-4}	0.0162

Παρατήρηση 2. Στο 3ο μέρος, όπου παρατηρείται η μεγαλύτερη διακύμανση των αποδόσεων του δείκτη NASDAQ με βάση το Σχήμα 2, η προσαρμογή είναι λιγότερο ικανοποιητική σε σχέση με τα προηγούμενα δυο μέρη -χωρίς ωστόσο να θεωρείται μη ικανοποιητική. Η λιγότερο ικανοποιητική προσαρμογή στο 3ο μέρος της χρονοσειράς ενδεχομένως να οφείλεται (και) στο γεγονός ότι ο αριθμός των δεδομένων που χρησιμοποιείται για την εκτίμηση των 9 παραμέτρων του μοντέλου (4)-(5) μέσω της

Σχήμα 5: (α') Οι εκτιμήσεις των θετικών αλμάτων του δείκτη NASDAQ από $t = 681$ έως $t = 755$, (β') Οι εκτιμήσεις των αρνητικών αλμάτων του δείκτη NASDAQ από $t = 681$ έως $t = 755$.



μεθόδου ΜΠ είναι σημαντικά μικρότερος συγκριτικά με τον αριθμό δεδομένων που χρησιμοποιείται στα άλλα δυο μέρη.

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Προτείνεται ένα μη γραμμικό μοντέλο χώρου καταστάσεων που στόχο έχει την εκτίμηση των κρυφών θετικών και αρνητικών αλμάτων των αποδόσεων χρηματοοικονομικών μεγεθών. Η εξίσωση των καταστάσεων που περιγράφει τη δυναμική εξέλιξη του διδιάστατου διανύσματος κατάστασης των αλμάτων αποτελεί εκθετικού τύπου συνάρτηση, ενώ η εξίσωση των παρατηρήσεων είναι γραμμική. Ακολουθεί εφαρμογή της προτεινόμενης μοντελοποίησης στην εκτίμηση των κρυφών θετικών και αρνητικών αλμάτων των αποδόσεων του δείκτη NASDAQ για την περίοδο 2006-2008. Καθότι οι χρονοσειρές των αποδόσεων χρηματοοικονομικών μεγεθών εμφανίζουν το χαρακτηριστικό της δεσμευμένης ως προς τον χρόνο ετεροσκεδαστικότητας, και με δεδομένο ότι το προτεινόμενο μοντέλο (2)-(3) έχει αμετάβλητους στον χρόνο πίνακες διασπορών-συνδιασπορών για τους θορύβους στην εξίσωση των καταστάσεων και των παρατηρήσεων, γίνεται εκτίμηση των σημείων αλλαγής ως προς τη διασπορά στη χρονοσειρά των δεδομένων. Με βάση τα εκτιμώμενα σημεία αλλαγής, η χρονοσειρά χωρίζεται σε τρία μέρη, και γίνεται η εκτίμηση των παραμέτρων του προτεινόμενου μοντέλου σε καθένα από αυτά. Ακολούθως της εκτίμησης -αποκάλυψης των κρυφών θετικών και αρνητικών αλμάτων, ελέγχεται και η προσαρμογή της διαφοράς τους στα εμπειρικά δεδομένα για καθένα από τα τρία μέρη οδηγώντας σε ικανοποιητικά αποτελέσματα με βάση τις τιμές του MSE στον Πίνακα 2. Τέλος, η προτεινόμενη μέθοδος μοντελοποίησης μπορεί να εφαρμοστεί γενικότερα σε κάθε περίπτωση όπου οι τιμές του διανύσματος κατάστασης ανήκουν σε ένα αριστερά φραγμένο χωρίο με βάση το φυσικό νόημα.

Σχετικά με ενδεχόμενη βελτίωση της προτεινόμενης μη γραμμικής μοντελοποίησης, θα μπορούσαν να υιοθετηθούν όροι με μεταβλητές διασπορές είτε στην εξίσωση των καταστάσεων είτε στην εξίσωση των παρατηρήσεων, προκειμένου να ενσωματωθεί στο μοντέλο χώρου καταστάσεων το χαρακτηριστικό της δεσμευμένης ως προς τον χρόνο ετεροσκεδαστικότητας για την εκτίμηση του κρυφού διανύσματος κατάστασης. Με αυτόν τον τρόπο, δεν θα είναι αναγκαίος ο έλεγχος εντοπισμού των σημείων αλλαγής ως προς τη διασπορά στα εμπειρικά δεδομένα.

ABSRTACT

An exponential state space model is presented for the estimation of the hidden two-sided jumps of asset returns, which are modeled as hidden random variables. The recursive relations of the estimation procedure are calculated. Additionally, the proposed model is applied to the empirical returns of NASDAQ index during 2006-2008 period. Taking into consideration that the proposed model is time-homogeneous in the sense that the volatilities in the state and measurement equations are time-invariant, the change points concerning the variance are detected in the empirical data, and the model is estimated at every part of the timeseries based on the estimated change points

ΑΝΑΦΟΡΕΣ

- Black, F. and Scholes, M. (1973). The Pricing of Options and Corporate Liabilities. *J. Polit. Econ.* **81**, 637-659.
- Cont, R. and Tankov, P. (2004). *Financial Modelling with Jump Processes*, Chapman & Hall/CRC.
- Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quant. Financ.* **1**, 223-236.
- Cox, J. and Ross, S. (1976). The valuation of options for alternative stochastic processes. *J. Financ. Econ.* **3**, 145-166.
- Duffie, D. and Pan, J. and Singleton, K. (2000). Transform Analysis and Asset Pricing for Affine Jump-Diffusions. *Econometrica* **68**, 1343-1376.
- Durbin, J. and Koopman, S.J. (2012). *Time Series Analysis by State Space Methods*, Oxford Univ. Press, Oxford.
- Geman, H. and Madan, D.B. and Yor, M. (2001). Time Changes for Lévy Processes. *Math. Financ.* **11**, 79-96.
- Kalman, R. (1960). A new approach to linear filtering and prediction problems. *J. Basic Engineering (Transactions of the ASME Series D)* **82**, 35-45.
- Kou, S.G.(2002). A jump-diffusion model for option pricing. *Management Science* **48**, 1086-1101.

- Madan, D. (2001). *Financial modeling with discontinuous price processes*, in Lévy Processes-Theory and Applications, Barndorff-Nielsen, O., Mikosch, T. and Resnik, S., Birkhäuser, Boston.
- Matteson, D. and James N.A. (2014). A Nonparametric Approach for Multiple Change Point Analysis of Multivariate Data. *J. Amer. Statist. Assoc.* **109**, 334-345.
- Merton, R. (1976). Option Pricing when Underlying Stock returns are discontinuous. *J. Financ. Econ.* **3**, 125-144.
- Polimenis, V. (2012). Information arrival as price jumps. *Optimization* **61**, 1179-1190.
- Schoutens, W. (2003). *Lévy Processes in Finance: Pricing Financial Derivatives*, Wiley, New York.
- Simon, D. (2006). *Optimal State Estimation*, Wiley, New York.
- Theodosiadou, O. and Polimenis, V. and Tsaklidis, G. (2016). Sensitivity analysis of market and stock returns by considering positive and negative jumps. *J. Risk Financ.* **17**, 456-472.
- Theodosiadou, O. and Tsaklidis, G. (2017). Estimating the Positive and Negative Jumps of Asset Returns via Kalman Filtering: The Case of Nasdaq Index. *J. Methodol. Comput. Appl. Probab.* **19**, 1123-1134.



ΜΙΑ ΝΕΑ ΔΙΑΚΡΙΤΗ ΚΑΤΑΝΟΜΗ

Β.Μ. Κούτρας, Μ.Β. Κούτρας, Σ.Δ. Δαφνής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς
vkoutras@icloud.com, mkoutras@unipi.gr, dafnisspyros@gmail.com

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία προτείνεται ένας τρόπος δημιουργίας μίας οικογένειας διακριτών κατανομών με χρήση της πιθανογεννήτριας μιας οποιασδήποτε διακριτής κατανομής (γεννήτορας). Γίνεται συστηματική μελέτη των ιδιοτήτων των νέων κατανομών που παράγονται και διερευνάται ιδιαίτερα η περίπτωση που ο γεννήτορας είναι μέλος της οικογένειας κατανομών Panjer. Τέλος, παρουσιάζεται μια εφαρμογή της νέας κατανομής στην αξιοπιστία συστημάτων.

Λέξεις Κλειδιά: Γεννήτορας, Πιθανογεννήτρια, Οικογένεια Panjer, Αξιοπιστία.

1. ΕΙΣΑΓΩΓΗ

Η μελέτη των ιδιοτήτων των γνωστών κατανομών, αλλά και η αναζήτηση νέων οικογενειών κατανομών, οι οποίες θα παρέχουν ευελιξία και προσαρμοστικότητα σε πραγματικά δεδομένα, αποτελούν διαχρονικά αντικείμενα έρευνας της Θεωρίας Κατανομών.

Στη διαδικασία αυτή προεξέχοντα ρόλο διαδραματίζουν οι συνεχείς κατανομές, με την κανονική κατανομή να είναι μακράν η πιο προβεβλημένη. Όμως, σε πολλές περιπτώσεις συμβαίνει οι μεταβλητές στον πραγματικό κόσμο να είναι συνεχείς στη φύση τους, αλλά διακριτές στην παρατήρηση (για παραδείγματα τέτοιων μεταβλητών παραπέμπουμε μεταξύ άλλων στη μονογραφία του Chakraborty (2015)). Για το λόγο αυτό παρατηρείται η ανάγκη μοντελοποίησης τέτοιων καταστάσεων από μια κατάλληλη διακριτή κατανομή η οποία, όμως, παράγεται από το υποκείμενο συνεχές μοντέλο και διατηρεί τα βασικά χαρακτηριστικά της συνεχούς κατανομής. Συνεπώς, η διακριτοποίηση συνεχών κατανομών έχει αποδειχθεί ένα πολύ χρήσιμο εργαλείο σε πολλές εφαρμογές. Για μια επισκόπηση των χρησιμοποιούμενων μεθόδων στις περιπτώσεις αυτές και των ιδιοτήτων των διακριτών κατανομών που προκύπτουν, ο αναγνώστης παραπέμπεται στον Chakraborty (2015).

Παράλληλα, εκτός από τη διακριτοποίηση συνεχών κατανομών, πολλές άλλες μέθοδοι έχουν αναπτυχθεί για τη δημιουργία νέων διακριτών κατανομών. Για τέτοιες μεθόδους (μείξη διακριτών κατανομών, αποκοπή τιμών διακριτής κατανομής κ.ά.) ο

αναγνώστης παραπέμπεται στους Johnson et al. (2005), Cohen (1960), Boswell (1970), Klugman et al. (2012).

Στην παρούσα εργασία, δημιουργείται μια νέα οικογένεια διακριτών κατανομών με μια νέα τεχνική. Πιο συγκεκριμένα, με δεδομένη μια διακριτή τ.μ. με πεπερασμένη πιθανογεννήτρια $P(z)$ (γεννήτορας), εισάγεται μία νέα οικογένεια χρησιμοποιώντας τις παραγώγους της $P(z)$. Στην Ενότητα 2, η νέα κατανομή ορίζεται και δίνονται κάποιες ιδιότητες της. Στην Ενότητα 3, μελετάται η κατανομή όταν ο γεννήτορας ανήκει στην οικογένεια κατανομών Panjer, καθώς στην οικογένεια αυτή ανήκουν οι κλασικές διακριτές κατανομές. Στην Ενότητα 4, δίνεται μια εφαρμογή των νέων αποτελεσμάτων στην αξιοπιστία συστημάτων.

Σε όλη την εργασία συμβολίζουμε με $(n)_x$ το μειούμενο παραγοντικό τάξης x , δηλ. $(n)_x = n!/(n-x)!$ και με $P^{(x)}(z)$, $x = 1, 2, \dots$ τις παραγώγους x τάξης της $P(z)$ ως προς z . Υπενθυμίζουμε, ότι η πιθανογεννήτρια $P(z)$ μπορεί να παρασταθεί ως μία δυναμοσειρά που συγκλίνει απολύτως και ομοιόμορφα τουλάχιστον για κάθε $-1 \leq z \leq 1$. Επομένως η $P(z)$ είναι απεριόριστα διαφορίσιμη για τουλάχιστον κάθε $-1 < z < 1$.

2. ΜΙΑ ΝΕΑ ΟΙΚΟΓΕΝΕΙΑ ΚΑΤΑΝΟΜΩΝ

Στην ενότητα αυτή θεωρούμε κάθε διακριτή τ.μ. N και παράγουμε μια οικογένεια διακριτών κατανομών οι οποίες ορίζονται μέσω της πιθανογεννήτριας της N . Στη συνέχεια παραθέτουμε μερικές ενδιαφέρουσες ιδιότητες της νέας κατανομής.

Ορισμός 1. Έστω N μια διακριτή τ.μ. με πεδίο τιμών $S_N \subseteq \{0, 1, 2, \dots\}$. Συμβολίζουμε με $f(n) = P(N = n)$ τη συνάρτηση πιθανότητας της και με $P(z) = E(z^N) = \sum_{n=0}^{\infty} P(N = n)z^n$ την πιθανογεννήτρια της. Αν $0 < \theta < 1$, τότε η κατανομή της τ.μ. X με συνάρτηση πιθανότητας

$$f_X(x; \theta) = P(X = x) = \frac{P^{(x)}(1 - \theta)}{x!} \theta^x, \quad x = 0, 1, \dots, \quad (1)$$

θα ονομάζεται f -επαγόμενη κατανομή με παράμετρο θ .

Στη συνέχεια, προκειμένου να αναφερόμαστε στην f -επαγόμενη κατανομή με παράμετρο θ , θα χρησιμοποιούμε το συμβολισμό $I(f, \theta)$; επομένως, για την τ.μ. X με συνάρτηση πιθανότητας που δίνεται από τη σχέση (1), θα γράφουμε $X \sim I(f, \theta)$.

Αφού με χρήση του αναπτύγματος Taylor προκύπτει

$$\sum_{x=0}^{\infty} f_X(x; \theta) = \sum_{x=0}^{\infty} \frac{P^{(x)}(1 - \theta)}{x!} \theta^x = P((1 - \theta) + \theta) = P(1) = 1,$$

η συνάρτηση πιθανότητας που δίνεται από τη σχέση (1) πράγματι γεννά μια κατανομή.

Ας θεωρήσουμε, ως παράδειγμα, την απλή περίπτωση που η N ακολουθεί μια διακριτή, ομοιόμορφη κατανομή με πεδίο τιμών $S_N = \{0, 1, \dots, m-1\}$. Σε αυτή την περίπτωση ισχύει ότι $P^{(x)}(z) = \frac{1}{m} \sum_{i=x}^{m-1} (i)_x z^{i-x}$, οπότε η σχέση (1) γίνεται

$$f_X(x; \theta) = P(X = x) = \frac{1}{m} \theta^x \sum_{n=x}^{m-1} \binom{n}{x} (1 - \theta)^{n-x}, \quad x = 0, 1, \dots, m - 1. \quad (2)$$

Το Σχήμα 1 απεικονίζει τη συνάρτηση πιθανότητας της κατανομής της X για διάφορες τιμές των παραμέτρων m και θ . Είναι προφανές ότι σε όλες τις περιπτώσεις η $f(x; \theta)$ είναι γνησίως φθίνουσα ως προς το x . Επίσης, τόσο η μέση τιμή, όσο και η διασπορά της κατανομής αυξάνονται καθώς το m ή το θ αυξάνεται. Πράγματι, υπολογίζοντας τη μέση τιμή και τη διασπορά της νέας κατανομής, παίρνουμε

$$E(X) = \sum_{x=0}^{m-1} x f_X(x; \theta) = \frac{(m-1)\theta}{2}, \quad E(X(X-1)) = \frac{(m-1)(m-2)\theta^2}{3}, \quad (3)$$

$$V(X) = \frac{(m^2 - 1)\theta^2}{12},$$

σχέσεις που επιβεβαιώνουν την τελευταία παρατήρηση.

Στις επόμενες δύο προτάσεις δίνονται εκφράσεις για τη συνάρτηση πιθανότητας της f -επαγόμενης κατανομής σε σχέση με τη συνάρτηση πιθανότητας της κατανομής της τ.μ. N . Είναι πολύ ενδιαφέρον ότι στην πρώτη εμφανίζεται η συνάρτηση πιθανότητας της κλασικής διωνυμικής κατανομής.

Πρόταση 1. Η συνάρτηση πιθανότητας της X μπορεί να εκφραστεί ως γραμμικός συνδυασμός της συνάρτησης πιθανότητας $f(n) = P(N = n)$, $n = 0, 1, \dots$, ως εξής

$$f_X(x; \theta) = \sum_{n=x}^{\infty} b(x; n, \theta) f(n), \quad 0 < \theta < 1, \quad (4)$$

όπου με $b(x; n, \theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}$ συμβολίζεται η συνάρτηση πιθανότητας της κλασικής διωνυμικής κατανομής με παραμέτρους n και θ .

Απόδειξη. Παραγωγίζοντας x φορές την έκφραση $P(z) = \sum_{n=0}^{\infty} P(N = n) z^n$ παίρνουμε

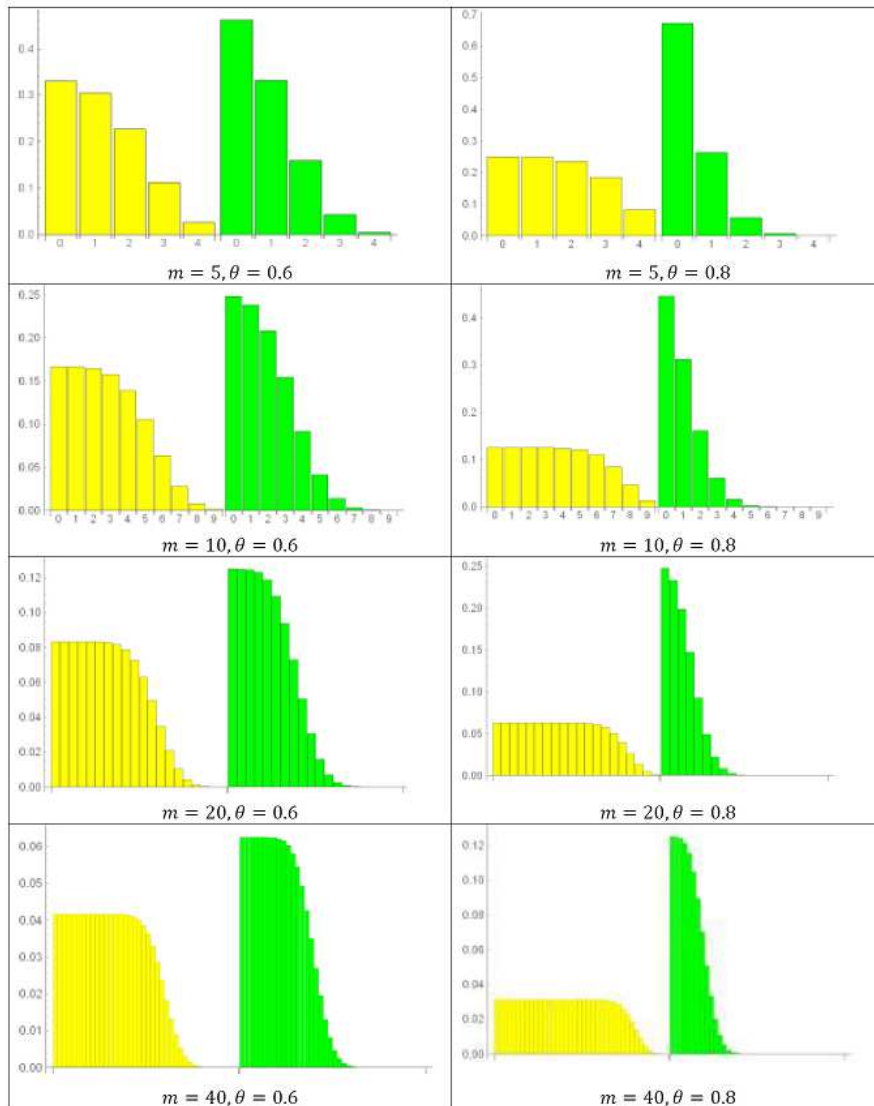
$$P^{(x)}(z) = \sum_{n=x}^{\infty} (n)_x f(n) z^{n-x}$$

και αντικαθιστώντας την τελευταία έκφραση στη σχέση (1) προκύπτει άμεσα

$$f_X(x; \theta) = \sum_{n=x}^{\infty} \binom{n}{x} \theta^x (1 - \theta)^{n-x} f(n) = \sum_{n=x}^{\infty} b(x; n, \theta) f(n), \quad 0 < \theta < 1. \quad (5)$$

Αξίζει να σημειωθεί ότι τα βάρη της $f(n)$ στο γραμμικό συνδυασμό (4) δεν αθροίζουν στο 1. Πράγματι,

Σχήμα 1: Η συνάρτηση πιθανότητας (2) για $m=5, 10, 20, 40$ και $\theta=0.2, 0.4, 0.6, 0.8$. Κάθε γράφημα απεικονίζει τη συνάρτηση πιθανότητας για την αναγραφόμενη τιμή του θ (αριστερά) και του $1 - \theta$ (δεξιά).



$$\begin{aligned}
 \sum_{n=x}^{\infty} b(x; n, \theta) &= \sum_{n=x}^{\infty} \binom{n}{x} \theta^x (1-\theta)^{n-x} = \theta^x \sum_{n=0}^{\infty} \binom{n+x}{x} (1-\theta)^n \\
 &= \theta^x (1 - (1-\theta))^{-(x+1)} = \frac{1}{\theta}.
 \end{aligned}$$

Το γεγονός αυτό δεν προκαλεί έκπληξη, καθώς οι όροι $f(n)$, $n \geq x$ δεν αθροίζονται, επίσης, στο 1.

Η επόμενη πρόταση δίνει μια εναλλακτική έκφραση για τη συνάρτηση πιθανότητας της τ.μ. X ως δυναμοσειρά του θ .

Πρόταση 2. Η συνάρτηση πιθανότητας της τ.μ. X μπορεί να εκφραστεί ως εξής

$$f_X(x; \theta) = \theta^x \sum_{j=0}^{\infty} a_{x,j} \theta^j, \quad 0 < \theta < 1,$$

όπου

$$a_{x,j} = (-1)^j \sum_{i=j}^{\infty} \binom{x+i}{x} \binom{i}{j} f(x+i).$$

Απόδειξη. Η απόδειξη προκύπτει άμεσα αν γράψουμε την (5) στη μορφή

$$f_X(x; \theta) = \theta^x \sum_{i=0}^{\infty} \binom{x+i}{x} (1-\theta)^i f(x+i) \quad (6)$$

και στη συνέχεια εφαρμόσουμε το διωνυμικό ανάπτυγμα στον όρο $(1-\theta)^i$.

Ακολουθεί ένα αποτέλεσμα για τις παραγοντικές ροπές της τ.μ. X .

Πρόταση 3. Η σχέση των παραγοντικών ροπών της τ.μ. X με τις παραγοντικές ροπές της τ.μ. N περιγράφεται από την ακόλουθη εξίσωση

$$E[(X)_r] = \theta^r E[(N)_r], \quad r = 1, 2, \dots$$

Απόδειξη. Παρατηρούμε πρώτα ότι η πιθανογεννήτρια της X , με χρήση της (1), γράφεται ως εξής

$$P_X(z) = E(z^X) = \sum_{x=0}^{\infty} f(x) z^x = \sum_{x=0}^{\infty} \frac{P^{(x)}(1-\theta)}{k!} (\theta z)^x;$$

χρησιμοποιώντας το ανάπτυγμα Taylor, καταλήγουμε ότι το δεξιό μέλος ισούται με $P((1-\theta) + \theta z)$, δηλ.

$$P_X(z) = P((1-\theta) + \theta z). \quad (7)$$

Παραγωγίζοντας την τελευταία έκφραση r φορές έχουμε

$$P_X^{(r)}(z) = E(z^X) = \theta^r P^{(r)}((1-\theta) + \theta z), \quad r = 1, 2, \dots$$

Θέτοντας $z = 1$ προκύπτει ότι $P_X^{(r)}(1) = \theta^r P^{(r)}((1-\theta) + \theta 1) = \theta^r P^{(r)}(1)$ και το αποτέλεσμα ακολουθεί άμεσα, αφού γνωρίζουμε ότι $E[(X)_r] = P_X^{(r)}(1)$ και $E[(N)_r] = P^{(r)}(1)$.

Η Πρόταση 3 για $r = 1$ και $r = 2$ δίνει άμεσα

$$E[X] = \theta E[N], \quad E[X(X-1)] = \theta^2 E[N(N-1)], \quad (8)$$

ενώ για τη διασπορά της τ.μ. X μπορούμε εύκολα να πάρουμε τον τύπο

$$\text{Var}[X] = \theta(1 - \theta)E[N] = \theta^2 \text{Var}[N].$$

Ο τύπος (3) μπορεί, επίσης, να προκύψει μέσω της (8) χρησιμοποιώντας τον ακόλουθο γνωστό τύπο για τις παραγοντικές ροπές της διακριτής ομοιόμορφης κατανομής στο $\{0, 1, \dots, m - 1\}$, $E[N]_r = \frac{(m-1)r}{r+1}$, $r = 0, 1, \dots, m - 1$.

3. Η f -ΕΠΑΓΟΜΕΝΗ ΚΑΤΑΝΟΜΗ ΠΟΥ ΓΕΝΝΑΤΑΙ ΑΠΟ ΣΥΓΚΕΚΡΙΜΕΝΕΣ ΔΙΑΚΡΙΤΕΣ ΚΑΤΑΝΟΜΕΣ

Μία ευρεία οικογένεια διακριτών κατανομών, με πολλές εφαρμογές στην Αναλογιστική Επιστήμη, είναι η αποκαλούμενη οικογένεια Panjer (δείτε π.χ. τους Panjer (1981), Guegan and Hassani (2009) ή το βιβλίο των Klugman et al. (2012)). Μια διακριτή τ.μ. N είναι μέλος της οικογένειας Panjer, ή ενάλλακτικά γνωστής ως $(a, b, 0)$ κλάσης κατανομών (συμβολισμός: $N \sim P(a, b, 0)$), αν η συνάρτηση πιθανότητας $p_n = P(N = n)$, $n = 0, 1, \dots$ της N ικανοποιεί την αναδρομική σχέση

$$p_n = \left(a + \frac{b}{n}\right)p_{n-1}, \quad n = 1, 2, \dots \quad (9)$$

με αρχική συνθήκη $p_0 = P(N = 0)$. Η οικογένεια Panjer $P(a, b, 0)$ ταυτίζεται με την οικογένεια όλων των (μη εκφυλισμένων) διωνυμικών, Poisson και αρνητικών διωνυμικών κατανομών, όπως έδειξαν οι Sundt και Jewell (1981). Στον Πίνακα 1 δίνονται οι τιμές των παραμέτρων a , b , όπως επίσης και η αντίστοιχη αρχική συνθήκη για τα 3 μέλη αυτής της οικογένειας.

Πίνακας 1: Οι διακριτές κατανομές της οικογένειας Panjer.

Κατανομή	$p_n = P(N = n)$	a	b	p_0
Διωνυμική $B(m, p)$	$\binom{m}{n} p^n (1-p)^{n-m}$	$-\frac{p}{1-p}$	$\frac{p(m-1)}{1-p}$	$(1-p)^m$
Poisson $P_0(\lambda)$	$\frac{\exp(-\lambda) \lambda^n}{n!}$	0	λ	$\exp(-\lambda)$
Αρνητική Διωνυμική $NB(m, p)$	$\frac{\Gamma(m+n)}{n! \Gamma(m)} p^m (1-p)^n$	$1-p$	$(m-1)(1-p)$	p^m

Οι Hess et al. (2002) απέδειξαν ότι η πιθανογεννήτρια $P(z)$ της $N \sim P(a, b, 0)$ ικανοποιεί τη διαφορική εξίσωση

$$(1 - az)P^{(r)}(z) = (ar + b)P^{(r-1)}(z), \quad r = 1, 2, \dots \quad (10)$$

Η επόμενη πρόταση αποδεικνύει ότι αν η τ.μ. N είναι μέλος της οικογένειας Panjer, τότε και η αντίστοιχη f -επαγόμενη κατανομή ανήκει στην ίδια κλάση κατανομών. Επομένως η οικογένεια Panjer είναι κλειστή κατά την εφαρμογή των f -επαγόμενων κατανομών.

Πρόταση 4. Εάν η τ.μ. $N \sim P(a, b, 0)$, τότε η τ.μ. X ανήκει στην οικογένεια $P(a_1, b_1, 0)$ με παραμέτρους

$$a_1 = \frac{a\theta}{1 - a(1 - \theta)}, \quad b_1 = \frac{b\theta}{1 - a(1 - \theta)}.$$

Απόδειξη. Σύμφωνα με τον ορισμό της κλάσης κατανομών $P(a, b, 0)$, ικανή και αναγκαία συνθήκη για να ανήκει η συνάρτηση πιθανότητας $p_n = P(N = n)$, $n = 0, 1, \dots$ στην κλάση, είναι να μπορεί να εκφραστεί το πηλίκο $\frac{np_n}{p_{n-1}}$ ως γραμμική συνάρτηση του n , δηλ. $\frac{np_n}{p_{n-1}} = an + b$ (βλέπετε τους Klugman et al. (2012)). Από την (1) προκύπτει

$$\frac{xf_X(x; \theta)}{f_X(x-1; \theta)} = \theta \frac{P^{(x)}(1 - \theta)}{P^{(x-1)}(1 - \theta)}$$

και λαμβάνοντας υπόψη τη (10)

$$\frac{xf_X(x; \theta)}{f_X(x-1; \theta)} = \theta \frac{ax + b}{1 - a(1 - \theta)} = \frac{a\theta}{1 - a(1 - \theta)}x + \frac{b\theta}{1 - a(1 - \theta)},$$

η οποία αποδεικνύει ότι

$$\frac{xP(X = x)}{P(X = x - 1)} = a_1x + b_1,$$

με $a_1 = a\theta/(1 - a(1 - \theta))$ και $b_1 = b\theta/(1 - a(1 - \theta))$. Επομένως, η τ.μ. X ανήκει στην οικογένεια $P(a_1, b_1, 0)$, αφού το πηλίκο $xP(X = x)/P(X = x - 1)$ είναι γραμμική συνάρτηση του x .

Η εφαρμογή της Πρότασης 4 στην περίπτωση των τριών κλασικών διακριτών κατανομών του Πίνακα 1 αποκαλύπτει ότι:

- Αν η $N \sim B(m, p)$, τότε η $X \sim B(m, p\theta)$.
- Αν η $N \sim P_0(\lambda)$, τότε η $X \sim P_0(\lambda\theta)$.
- Αν η $N \sim NB(m, p)$, τότε η $X \sim NB(m, p/(\theta + (1 - \theta)p))$.

Συνεπώς, η οικογένεια των Διωνυμικών κατανομών, η οικογένεια των κατανομών Poisson και η οικογένεια των Αρνητικών Διωνυμικών κατανομών είναι κλειστές κατά την εφαρμογή των f -επαγόμενων κατανομών.

4. ΜΙΑ ΕΦΑΡΜΟΓΗ

Ας θεωρήσουμε ένα k -από-τα- n : F σύστημα, το οποίο αποτυγχάνει όταν τουλάχιστον k συνιστώσες του συστήματος αποτύχουν. Θεωρούμε, επίσης, ένα μοντέλο σοκ (shock model) σύμφωνα με το οποίο το σύστημα δέχεται κάποια σοκ κατά τη διάρκεια της λειτουργίας του (μεταβολές στην τάση, δονήσεις, κραδασμούς κ.τ.λ.), τα οποία μπορεί να προκαλέσουν βλάβες σε κάποιες από τις συνιστώσες του. Έστω N η τ.μ. απαρίθμησης των συνιστωσών που αποτυγχάνουν όταν συμβαίνει κάποιο σοκ και αν $N \geq k$ (δηλ. αν το σύστημα μεταβεί σε κατάσταση μη-λειτουργίας) ξεκινά μια διαδικασία επισκευής για κάθε μία από τις συνιστώσες που έχουν πληγεί. Οι υπόλοιπες

(λειτουργικές) συνιστώσες δε χρησιμοποιούνται έως ότου το σύστημα επανέλθει σε κατάσταση λειτουργίας. Ας συμβολίσουμε με $Y_i, i = 1, 2, \dots, N$ το χρόνο επισκευής της i συνιστώσας που έχει αποτύχει. Προφανώς ο χρόνος T μέχρι το σύστημα να επανέλθει σε κατάσταση λειτουργίας θα περιγράφεται από τον k μεγαλύτερο χρόνο επισκευής ή ισοδύναμα, αν συμβολίσουμε με $Y_{(1)} \leq Y_{(2)} \leq \dots \leq Y_{(N)}$ το διατεταγμένο δείγμα των Y_i , ο T θα είναι η $(N - k + 1)$ διατεταγμένη παρατήρηση του τυχαίου δείγματος Y_1, Y_2, \dots, Y_N τυχαίου μεγέθους. Για να είμαστε πιο ακριβείς, ο χρόνος T ισούται με

$$T = \begin{cases} 0, & \text{αν } N < k \\ Y_{n-k+1:n}, & \text{αν } N = n \geq k. \end{cases}$$

Λαμβάνοντας υπόψη την Πρόταση 3.1 των Koutras and Koutras (2018) και τον Ορισμό 1, η αθροιστική συνάρτηση πιθανότητας της τ.μ. T μπορεί να εκφραστεί ως εξής:

$$P(T \leq t) = \begin{cases} P(N < k), & \text{για } t = 0 \\ \sum_{x=0}^{k-1} f_X(x; P(Y_i > t)), & \text{για } t > 0. \end{cases}$$

όπου η χρονική στιγμή $t = 0$ αναφέρεται στην τιμή του σοκ και $f_X(x; \theta)$ είναι η συνάρτηση πιθανότητας της f -επαγόμενης κατανομής με παράμετρο θ . Συνεπώς, η πιθανότητα το σύστημα να επανέλθει από το σοκ τη χρονική στιγμή $t > 0$ μπορεί να υπολογισθεί εύκολα, μέσω της τιμής της αθροιστικής συνάρτησης πιθανότητας της f -επαγόμενης κατανομής στο σημείο $X = k - 1$.

Σύμφωνα με το απλό μοντέλο σοκ, ένα σοκ καταστρέφει κάθε συνιστώσα με πιθανότητα p , δηλ. η $N \sim B(n; p)$. Τότε η f -επαγόμενη κατανομή με παράμετρο θ ανάγεται στη Διωνυμική κατανομή $B(n; p\theta)$ (βλέπετε το σχόλιο που ακολουθεί την Πρόταση 4) και ο προηγούμενος τύπος γίνεται

$$P(T \leq t) = \begin{cases} \sum_{x=0}^{k-1} \binom{n}{x} p^x (1-p)^{n-x}, & \text{για } t = 0 \\ \sum_{x=0}^{k-1} \binom{n}{x} (pP(Y_i > t))^x (1-pP(Y_i > t))^{n-x}, & \text{για } t > 0. \end{cases}$$

Υποθέτοντας εκθετικούς χρόνους με μέσο χρόνο επισκευής μ παίρνουμε τον ακόλουθο τύπο για την κατανομή του χρόνου T μέχρι το σύστημα να επανέλθει από το σοκ

$$P(T \leq t) = \begin{cases} \sum_{x=0}^{k-1} \binom{n}{x} p^x (1-p)^{n-x}, & \text{για } t = 0 \\ \sum_{x=0}^{k-1} \binom{n}{x} (pe^{-t/\mu})^x (1-pe^{-t/\mu})^{n-x}, & \text{για } t > 0. \end{cases}$$

Για σταθερά n, k, p, t η αθροιστική συνάρτηση πιθανότητας $P(T \leq t)$ είναι φθίνουσα συνάρτηση ως προς μ ; επομένως, αν θέλουμε να έχουμε μια προκαθορισμένη πιθανότητα $1 - \alpha$ το σύστημα να έχει επανέλθει από το σοκ τη χρονική στιγμή t , θα πρέπει να βρούμε τον κατάλληλο μέσο χρόνο επισκευής $\mu = \mu(t)$ λύνοντας την

εξίσωση $\sum_{x=0}^{k-1} \binom{n}{x} (pe^{-t/\mu})^x (1 - pe^{-t/\mu})^{n-x} = 1 - \alpha$ ως προς μ . Ο Πίνακας 2 δίνει τις θετικές, πραγματικές τιμές του μ για $1 - \alpha = 0,95$ και διαφορετικές επιλογές των παραμέτρων n, k, t, p .

Πίνακας 2: Οι τιμές του μ για $1 - \alpha = 0,95$ και διαφορετικές επιλογές των παραμέτρων n, k, t, p .

t	10			20			30			
p	0,4	0,6	0,8	0,4	0,6	0,8	0,4	0,6	0,8	
n	k									
10	1	2,29	2,10	1,98	4,59	4,20	3,96	6,88	6,30	5,94
	2	4,19	3,58	3,25	8,38	7,16	6,49	12,57	10,74	9,74
	3	6,57	5,19	4,51	13,14	10,37	9,03	19,70	15,56	13,54
	5	17,04	10,08	7,81	34,08	20,16	15,63	51,12	30,23	23,44
20	1	1,98	1,83	1,74	3,96	3,67	3,48	5,94	5,50	5,22
	3	4,44	3,77	3,38	8,89	7,53	6,80	13,33	11,30	10,19
	5	7,43	5,71	4,90	14,86	11,42	9,81	22,28	17,13	14,71
	10	35,56	14,56	10,26	71,12	29,13	20,53	106,69	43,69	30,79
50	1	1,68	1,57	1,50	3,35	3,14	3,00	5,03	4,71	4,50
	3	3,14	2,79	2,58	6,28	5,57	5,16	9,42	8,36	7,74
	5	4,35	3,70	3,34	8,71	7,40	6,69	13,06	11,10	10,03
	10	7,90	5,98	5,10	15,79	11,96	10,21	23,69	17,94	15,31
	20	28,94	13,32	9,63	57,88	26,63	19,25	86,82	39,95	28,88

ABSTRACT

In the present paper a new family of discrete distributions is introduced through the probability generating function of any discrete distribution (generator). The properties of the family are systematically studied when the generator is an arbitral discrete distribution or when it belongs to the well-known Panjer family of discrete distributions. An application of the new distributions in reliability theory is, also, presented.

Ευχαριστίες: Η παρούσα έρευνα έχει χρηματοδοτηθεί από τη Γενική Γραμματεία Έρευνας και Τεχνολογίας μέσω των κονδυλίων της Εθνικής Συμμετοχής 2016-2017 που σχετίζονται με το Ευρωπαϊκό Έργο “ISMPH: Inference for a Semi-Markov Process using Hazards Specification” (GA No 329128).

ΑΝΑΦΟΡΕΣ

- Boswell, M.T. and Patil, G.P. (1970). Chance mechanisms generating the negative binomial distributions, in: Patil, G. P. (ed.) Random Counts in Scientific Work, Vol. 1: Random Counts in Models and Structures, University Park, PA: Pennsylvania State University Press, pp. 3-22.
- Chakraborty, S. (2015). Generating discrete analogues of continuous probability distributions- A survey of methods and constructions. *Journal of Statistical Distributions and Applications* **2**, <https://doi.org/10.1186/s40488-015-0028-6>.

- Cohen, A.C. (1960). Estimating the parameter in a conditional Poisson distribution. *Biometrics* **16**, 203-211.
- Guegan, D. and Hassani, B. (2009). A modified Panjer algorithm for operational risk capital calculations. *Journal of Operational Risk*, **4**, 53-72.
- Hess, K.T., Liewald A. and Schmidt, K.D. (2002). An extension of Panjer's recursion. *ASTIN Bulletin*, **32**, 283-297.
- Johnson, N.L., Kemp A.W. and Kotz, S. (2005). *Univariate Discrete Distributions*, New York: Wiley.
- Klugman, S.A., Panjer, H.H. and Willmot, G.E. (2012). *Loss Models: From Data to Decisions*, New York: Wiley.
- Koutras, V.M. and Koutras, M.V. (2018). Exact distribution of random order statistics and applications in Risk Management. *Methodology and Computing in Applied Probability* <https://doi.org/10.1007/s11009-018-9662-z>.
- Panjer, H.H. (1981). Recursive evaluation of a family of compound distributions. *ASTIN Bulletin*, **12**, 22-26.
- Sundt, B. and Jewell, W.S. (1981). Further results of recursive evaluation of compound distributions. *ASTIN Bulletin*, **12**, 27-39.



ΝΕΕΣ ΑΝΙΣΟΤΗΤΕΣ ΓΙΑ N -DEMIMARTINGALES ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΙΣ ΣΤΑΤΙΣΤΙΚΕΣ ΣΥΝΑΡΤΗΣΕΙΣ ΣΑΡΩΣΗΣ

Μ. Κούτρας¹, Δ. Λυμπερόπουλος²

¹Πανεπιστήμιο Πειραιώς
mkoutras@unipi.gr

²Ελληνική Στατιστική Αρχή (ΕΛΣΤΑΤ)
d.lymperopoulos@statistics.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία, αρχικά αποδεικνύονται μερικές νέες ανισότητες για N -demi-(super)martingales. Βάσει αυτών, εξάγονται κάποια νέα φράγματα για τη συνάρτηση κατανομής πιθανότητας του χρόνου αναμονής μέχρι την πρώτη εμφάνιση μιας ροής τύπου r/k , σε μία ακολουθία ανεξάρτητων και ισόνομων δίτιμων δοκιμών με πιθανότητα επιτυχίας p . Τέλος, δίνονται αριθμητικά αποτελέσματα για την αποτίμηση της ποιότητας των φραγμάτων.

Λέξεις Κλειδιά: demimartingale, N -demimartingale, συνάρτηση σάρωσης, φράγμα.

1. ΕΙΣΑΓΩΓΗ

Τις τελευταίες δεκαετίες, τα demimartingales, τα N -demimartingales και σχετικές γενικεύσεις των martingales έχουν κεντρίσει το ενδιαφέρον πολλών ερευνητών (βλ. π.χ. τα πρωτοπόρα άρθρα των Newman and Wright (1982) και Christofides (2003), αντίστοιχα, όπου εισήχθησαν οι τελευταίες έννοιες), κυρίως εξαιτίας του γεγονότος ότι τα σχετικά με αυτά αποτελέσματα μπορούν να αξιοποιηθούν για την αντιμετώπιση προβλημάτων στοχαστικής εξάρτησης. Κάτι τέτοιο απαιτεί συχνά την επέκταση αποτελεσμάτων που αφορούν τα martingales κάτω από γενικότερες παραδοχές. Για περισσότερες λεπτομέρειες, παραπέμπουμε στο εξαιρετικό βιβλίο του Prakasa Rao (2012).

Τέτοιες επεκτάσεις, όμως, δεν είναι πάντα άμεσες ούτε εγγυημένες! Αντίθετα, μπορεί να αποδειχθούν ιδιαίτερα προβληματικές, όπως αποδεικνύεται και από τη σχετικά πρόσφατη εργασία των Dai et al. (2014). Εκεί παρατίθεται ένα αντιπαράδειγμα για την ισχύ μιας μεγιστικής ανισότητας για N -demimartingales τύπου Chow. Μά-

λιστα, το αντιπαράδειγμα αυτό βρίσκει εφαρμογή και σε μία άλλη γνωστή μεγιστική ανισότητα για N -demimartingales, βλ. Christofides (2003, Theorem 2.1) ή Prakasa Rao (2012, Theorem 3.2.1), που αν ίσχυε θα μπορούσε να παρουσιάζει ιδιαίτερο ενδιαφέρον για εφαρμογές σχετικές με τις στατιστικές συναρτήσεις σάρωσης. Ο κύριος στόχος της παρούσας εργασίας είναι να παράσχει κάποιες εναλλακτικές της προαναφερθείσας μεγιστικής ανισότητας.

Επισημαίνεται ότι οι στατιστικές συναρτήσεις ορίζονται ως απαριθμητήριες τυχαίες μεταβλητές (τ.μ.) των επικαλυπτόμενων κυλιόμενων παραθύρων που εμφανίζονται σε μία ακολουθία δίτιμων δοκιμών και περιέχουν έναν προκαθορισμένο αριθμό επιτυχιών. Τα σχετικά προβλήματα χρόνων αναμονής έχουν απασχολήσει εκτενώς τη διεθνή βιβλιογραφία και βρίσκουν εφαρμογή σε έναν σημαντικό αριθμό επιστημονικών περιοχών, όπως ο αναλογισμός, η θεωρία αξιοπιστίας και η μοριακή βιολογία, βλ. π.χ. Balakrishnan & Koutras (2002), Papastavridis & Koutras (1993) και Chen & Karlin (2007), αντίστοιχα. Για μια πρόσφατη επισκόπηση στις στατιστικές συναρτήσεις σάρωσης, παραπέμπουμε στον τόμο των Glaz & Koutras (2019).

Μάλιστα, τα σχετικά με τις στατιστικές συναρτήσεις σάρωσης προβλήματα, έχουν αποτελέσει αντικείμενο μελέτης κάτω από διαφορετικά πλαίσια (όπως αυτά των Fu (2001) και των Wu et al. (2013)) που αναφέρονται για παράδειγμα στις υποθέσεις που γίνονται για την ακολουθία των δίτιμων δοκιμών και για τα κυλιόμενα παράθυρα. Ιδιαίτερως, η θεωρία των martingales αποδείχθηκε ένα χρήσιμο εργαλείο για τη μελέτη χρόνων αναμονής «σύνθετων σχηματισμών» (μεταξύ των οποίων και οι σαρώσεις και οι ροές), όπως καταδεικνύεται για παράδειγμα και από τις εργασίες των Pozdnyakov et al. (2005) και Pozdnyakov and Steele (2009).

Παρόλα αυτά, υπάρχει ένα κενό στην αξιοποίηση αποτελεσμάτων που αποδείχθηκαν για τις προαναφερθείσες γενικεύσεις των martingales με σκοπό την αντιμετώπιση προβλημάτων σχετικών με τις στατιστικές συναρτήσεις σάρωσης. Στην παρούσα εργασία, παρουσιάζεται μια προσπάθεια κάλυψης του κενού αυτού για σχετικά προβλήματα χρόνων αναμονής.

Πιο συγκεκριμένα, αρχικά παρατίθενται οι απαραίτητες βασικές έννοιες και οι αναγκαίοι συμβολισμοί (βλ. Ενότητα 2), ενώ στην Ενότητα 3 εξάγονται κάποιες νέες μεγιστικές ανισότητες για N -demi(super)martingales. Εφαρμόζοντας τις ανισότητες αυτές, στην τελευταία ενότητα καθίσταται δυνατή η εξαγωγή νέων φραγμάτων για τη συνάρτηση κατανομής πιθανότητας (σ.κ.π.) του χρόνου αναμονής μέχρι την πρώτη εμφάνιση μιας σάρωσης συγκεκριμένου τύπου, η συμπεριφορά των οποίων διερευνάται και μέσω αριθμητικών αποτελεσμάτων.

2. ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ ΚΑΙ ΣΥΜΒΟΛΙΣΜΟΙ

Το σύνολο των θετικών ακεραίων συμβολίζεται με \mathbb{N} και $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. Με \mathbb{R} συμβολίζεται το σύνολο των πραγματικών αριθμών, ενώ $\mathbb{R}_+ := \{x \in \mathbb{R} : x \geq 0\}$. Για $d \in \mathbb{N}$, με \mathbb{R}^d συμβολίζεται ο Ευκλείδειος χώρος διάστασης d . Επί πλέον, θα χρησιμοποιήσουμε τους συμβολισμούς $x \wedge y := \min\{x, y\}$, $x \vee y := \max\{x, y\}$ και $x^+ := x \vee 0$ για $x, y \in \mathbb{R}$. Για κάθε $n \in \mathbb{N}$ και $i \in \{1, \dots, n\}$ 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) . 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 $T_Z := \{B \subseteq \mathcal{Y} : Z^{-1}(B) \in \Sigma\}$, 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 . Με $\mathbf{B}(n, p)$, όπου $n \in \mathbb{N}$ και $p \in (0, 1)$, σημειώνεται η διωνυμική κατανομή. Επί πλέον, η συνάρτηση πιθανότητας και η σ.χ.π. στο $x \in \mathbb{R}$ θα συμβολίζονται με $b(x; n, p)$ και $F_b(x; n, p)$, αντίστοιχα, ενώ $M_b(\alpha, n, p) := \sum_{x=\alpha}^n xb(x; n, p)$ για κάθε $\alpha \in \{1, \dots, n\}$.

Ένα σύνολο $N \in \Sigma$ με $P(N) = 0$ καλείται ένα **P -μηδενικό σύνολο**. Μια ακολουθία $\{Z_j\}_{j \in \mathbb{N}}$ από Σ -μετρήσιμες συναρτήσεις **ικανοποιεί μια ιδιότητα P -σχεδόν βέβαια (σ.β.)** αν υπάρχει ένα P -μηδενικό σύνολο, έστω O , ώστε η ιδιότητα να ισχύει για την $\{Z_j\}_{j \in \mathbb{N}}$ για κάθε $\omega \notin O$.

Με $\mathcal{L}^1(P)$ σημειώνεται η οικογένεια όλων των P -ολοκληρώσιμων συναρτήσεων $f : \Omega \rightarrow \mathbb{R}$. Συναρτήσεις που είναι P -σ.β. ίσες δεν ταυτίζονται. Η μέση τιμή της τ.μ. Z σημειώνεται με $\mathbb{E}_P[Z]$. Αν $Z \in \mathcal{L}^1(P)$ και η \mathcal{F} είναι μια σ -υποάλγεβρα της Σ , τότε κάθε συνάρτηση $\tilde{Z} \in \mathcal{L}^1(P | \mathcal{F})$ που ικανοποιεί για κάθε $F \in \mathcal{F}$ την ισότητα $\int_F \tilde{Z} dP = \int_F Z dP$ ονομάζεται μια **εκδοχή της δεσμευμένης μέσης τιμής της Z δοθείσης της \mathcal{F}** , και θα σημειώνεται με $\mathbb{E}_P[Z | \mathcal{F}]$. Επί πλέον, για κάθε $E \in \Sigma$ θέτουμε $P(E | \mathcal{F}) := \mathbb{E}_P[\chi_E | \mathcal{F}]$, όπου χ_E είναι η δείκτρια (ή χαρακτηριστική) συνάρτηση του E .

Υπενθυμίζουμε ακόμη ότι μια οικογένεια $\{\mathcal{F}_j\}_{j \in \mathbb{N}}$ από σ -υποάλγεβρες της Σ , τέτοιες ώστε να ισχύει $\mathcal{F}_j \subseteq \mathcal{F}_{j+1}$ για κάθε $j \in \mathbb{N}$, ονομάζεται μια **διύλιση** για τον μετρήσιμο χώρο (Ω, Σ) . Επί πλέον, μια ακολουθία $\{Z_j\}_{j \in \mathbb{N}}$ τ.μ. επάνω στον Ω ονομάζεται **προσαρμοσμένη σε μια διύλιση $\{\mathcal{F}_j\}_{j \in \mathbb{N}}$** αν κάθε τ.μ. Z_j είναι \mathcal{F}_j -μετρήσιμη. Αν $\mathcal{F}_j = \sigma(\bigcup_{i=1}^j \sigma(Z_i))$ για κάθε $j \in \mathbb{N}$, τότε η $\{\mathcal{F}_j\}_{j \in \mathbb{N}}$ ονομάζεται η **κανονική διύλιση** για την $\{Z_j\}_{j \in \mathbb{N}}$, και θα σημειώνεται με $\{\mathcal{F}_j^{(Z)}\}_{j \in \mathbb{N}}$.

3. ΜΕΓΙΣΤΙΚΕΣ ΑΝΙΣΟΤΗΤΕΣ ΓΙΑ N -DEMIMARTINGALES

Αρχικά ανακαλούμε κάποιες θεμελιώδεις για τους σκοπούς της παρούσας εργασίας έννοιες.

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 **(α) ένα P -martingale** (ως προς την $\{\mathcal{F}_j^{(Z)}\}_{j \in \mathbb{N}}$) αν

$$\mathbb{E}_P[(Z_{j+1} - Z_j)f(Z_1, \dots, Z_j)] = 0 \quad \text{για κάθε } j \in \mathbb{N} \quad (1)$$

και για κάθε μετρήσιμη συνάρτηση f επάνω στον \mathbb{R}^j τέτοιες ώστε να υπάρχουν οι παραπάνω μέσες τιμές.

(β) ένα P -demimartingale, αν η συνθήκη (1) αλλά με \geq στη θέση της ισότητας ικανοποιείται για κάθε f μη φθίνουσα κατά συντεταγμένες συνάρτηση επάνω στο \mathbb{R}^n για την οποία ορίζονται οι παραπάνω μέσες τιμές.

(γ) ένα P -demisubmartingale αν η συνθήκη (1) ικανοποιείται για κάθε f όπως στο (b) αλλά με $f \geq 0$.

(δ) ένα N -demimartingale κάτω από το P , αν η συνθήκη (1) αλλά με την αντίστροφη ανισότητα ικανοποιείται για κάθε f όπως στο (b). Ιδιαίτερος, αν η $f \geq 0$ τότε η $\{Z_n\}_{n \in \mathbb{N}}$ ονομάζεται ένα N -demisupermartingale κάτω από το P .

Παρατηρήσεις 1. (α) Εύκολα μπορεί να αποδειχθεί ότι η συνθήκη (1) είναι ισοδύναμη με το ότι για κάθε $n \in \mathbb{N}$ ισχύει η $\mathbb{E}_P[Z_{n+1} | \mathcal{F}_n] = Z_n$ P | \mathcal{F}_n -σ.β., δηλαδή με την κλασσική διατύπωση της martingale-ιδιότητας.

(β) 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.

Οι επόμενες δύο προτάσεις προσφέρουν εναλλακτικές στο Theorem 2.1 του Christofides (2003) για σ.β. μη αρνητικά N -demimartingales.

Πρόταση 1. Αν η ακολουθία $\{Z_j\}_{j \in \mathbb{N}}$ είναι ένα N -demimartingale κάτω από το P τέτοιο ώστε $Z_j \geq 0$ P -σ.β. για κάθε $j \in \mathbb{N}$, τότε για οποιοδήποτε σταθερό $t \in \mathbb{N}$ και για κάθε $\varepsilon > 0$ ισχύει η ακόλουθη ανισότητα:

$$\varepsilon P\left(\max_{1 \leq j \leq t} Z_j \geq \varepsilon\right) \leq \sum_{m=1}^t \mathbb{E}_P[Z_m \chi_{\{\max_{1 \leq j \leq m} Z_j \geq \varepsilon\}}].$$

Απόδειξη. Αρχικά σταθεροποιούμε $t \in \mathbb{N}$ και $\varepsilon > 0$. Στη συνέχεια ορίζουμε την τ.μ. $\tilde{\tau}_{t,\varepsilon} : \Omega \rightarrow \mathbb{R}$ μέσω του τύπου

$$\begin{aligned} \tilde{\tau}_{t,\varepsilon}(\omega) &:= \tilde{\tau}_{t,\varepsilon}(Z_1, \dots, Z_t)(\omega) \\ &:= \begin{cases} \sup\{j \in \{1, \dots, t\} : Z_j(\omega) \geq \varepsilon\}, & \text{αν } \omega \in \bigcup_{j=1}^t \{Z_j \geq \varepsilon\} \\ 1, & \text{αν } \omega \in \bigcap_{j=1}^t \{Z_j < \varepsilon\} \end{cases} \end{aligned}$$

για κάθε $\omega \in \Omega$. Προφανώς, η $\tilde{\tau}_{t,\varepsilon}$ είναι μια $\mathcal{F}_t^{(Z)}$ -μετρήσιμη συνάρτηση και $R_{\tilde{\tau}_{t,\varepsilon}} = \{1, \dots, t\}$.

Επί πλέον, η $\tilde{\tau}_{t,\varepsilon}$ είναι μια μη φθίνουσα κατά συντεταγμένες συνάρτηση των Z_1, \dots, Z_t , κάτι που αποδεικνύεται ως εξής:

(α) Έστω $\omega \in \bigcup_{j=1}^t \{Z_j \geq \varepsilon\}$. Τότε υπάρχει ένα $j_1 \in \{1, \dots, t\}$ τέτοιο ώστε $\sup\{j \in \{1, \dots, t\} : Z_j(\omega) \geq \varepsilon\} = j_1$, οπότε $\tilde{\tau}_{t,\varepsilon}(Z_1, \dots, Z_t)(\omega) = j_1$. Ας

θεωρήσουμε, τώρα, μια ακολουθία $\{\hat{Z}_j\}_{j \in \mathbb{N}}$ τ.μ. επάνω στον Ω τέτοια ώστε:

(α1) $\hat{Z}_{j_2}(\omega) \geq Z_{j_2}(\omega)$ για κάποιο $j_2 \in \{j_1, \dots, t\}$ και $\hat{Z}_j(\omega) = Z_j(\omega)$ για κάθε $j \neq j_2$. Αν $\hat{Z}_{j_2}(\omega) \geq \varepsilon$ τότε έχουμε

$$\tilde{\tau}_{t,\varepsilon}(\hat{Z}_1, \dots, \hat{Z}_{j_2}, \dots, \hat{Z}_t)(\omega) = \tilde{\tau}_{t,\varepsilon}(Z_1, \dots, \hat{Z}_{j_2}, \dots, Z_t)(\omega) = j_2, \quad (2)$$

ενώ αν $\widehat{Z}_{j_2}(\omega) < \varepsilon$, τότε η συνθήκη (2) ισχύει με j_1 στη θέση του j_2 .

($\alpha 2$) Να υπάρχει ένα $j_3 \in \{1, \dots, j_1 - 1\}$ με $\widehat{Z}_{j_3}(\omega) \geq Z_{j_3}(\omega)$ και $\widehat{Z}_j(\omega) = Z_j(\omega)$ για κάθε $j \neq j_3$. Τότε έχουμε ότι

$$\tilde{\tau}_{t,\varepsilon}(\widehat{Z}_1, \dots, \widehat{Z}_{j_3}, \dots, \widehat{Z}_t)(\omega) = \tilde{\tau}_{t,\varepsilon}(Z_1, \dots, \widehat{Z}_{j_3}, \dots, Z_t)(\omega) = j_1 > j_3.$$

(β) Έστω $\omega \in \bigcap_{j=1}^t \{Z_j < \varepsilon\}$. Ας θεωρήσουμε μια ακολουθία $\{\check{Z}_j\}_{j \in \mathbb{N}}$ τ.μ. επάνω στον Ω ώστε $\check{Z}_{j_4}(\omega) \geq Z_{j_4}(\omega)$ για κάποιο $j_4 \in \{1, \dots, t\}$ και $\check{Z}_j(\omega) = Z_j(\omega)$ για κάθε $j \neq j_4$. Αν $\check{Z}_{j_4}(\omega) \geq \varepsilon$, τότε έχουμε $\tau_{t,\varepsilon}(\check{Z}_1, \dots, \check{Z}_{j_4}, \dots, \check{Z}_t)(\omega) = j_4 > 1$, αλλιώς ισχύει η ισότητα $\tau_{t,\varepsilon}(\check{Z}_1, \dots, \check{Z}_{j_4}, \dots, \check{Z}_t)(\omega) = 1$.

Επομένως, η $\chi_{[0,t]}(\tilde{\tau}_{t,\varepsilon})$ είναι μια μη αύξουσα κατά συντεταγμένες συνάρτηση των Z_1, \dots, Z_t . Το τελευταίο γεγονός σε συνδυασμό με το ότι η $\{Z_j\}_{j \in \mathbb{N}}$ είναι ένα N -demisupermartingale κάτω από το P συνεπάγεται ότι $\mathbb{E}_P[Z_1] \geq \mathbb{E}_P[Z_{\tilde{\tau}_{t,\varepsilon} \wedge t}]$, βλ. Hu et al. (2010) ή Prakasa Rao (2012, Theorem 3.1.7), οπότε έχουμε ότι

$$\mathbb{E}_P[Z_1] \geq \mathbb{E}_P[Z_{\tilde{\tau}_{t,\varepsilon}} \chi_{\{\max_{1 \leq j \leq t} Z_j \geq \varepsilon\}}] + \mathbb{E}_P[Z_1 \chi_{\{\max_{1 \leq j \leq t} Z_j < \varepsilon\}}]$$

ή ισοδύναμα

$$\mathbb{E}_P[Z_1 \chi_{\{\max_{1 \leq j \leq t} Z_j \geq \varepsilon\}}] \geq \mathbb{E}_P[Z_{\tilde{\tau}_{t,\varepsilon}} \chi_{\{\max_{1 \leq j \leq t} Z_j \geq \varepsilon\}}].$$

Επειδή, όμως, όλες οι τ.μ. Z_n είναι P -σ.β. μη αρνητικές έπεται ότι

$$\begin{aligned} \mathbb{E}_P[Z_{\tilde{\tau}_{t,\varepsilon}} \chi_{\{\max_{1 \leq j \leq t} Z_j \geq \varepsilon\}}] &\geq \mathbb{E}_P[Z_t \chi_{\{\max_{1 \leq j \leq t} Z_j \geq \varepsilon\}} \cap \{\max_{1 \leq j \leq t-1} Z_j < \varepsilon, Z_t \geq \varepsilon\}] \\ &\geq \varepsilon P\left(\max_{1 \leq j \leq t-1} Z_j < \varepsilon, Z_t \geq \varepsilon\right), \end{aligned}$$

συνεπώς θέτοντας $\psi(t) := \psi(t; \varepsilon) := P(\max_{1 \leq j \leq t} Z_j \geq \varepsilon)$, προκύπτει ότι

$$\varepsilon[\psi(t) - \psi(t-1)] \leq \mathbb{E}_P[Z_1 \chi_{\{\max_{1 \leq j \leq t} Z_j \geq \varepsilon\}}] \quad \text{για κάθε } t \in \mathbb{N},$$

αφού το t είναι αυθαίρετο. Με επαγωγή στο $t \in \mathbb{N}$ προκύπτει η πρόταση. \square

Πρόταση 2. Αν η ακολουθία $\{Z_j\}_{j \in \mathbb{N}}$ είναι ένα N -demimartingale κάτω από το P τέτοιο ώστε $Z_j \geq 0$ P -σ.β. για κάθε $j \in \mathbb{N}$, τότε για οποιοδήποτε σταθερό $t \in \mathbb{N}$ και για κάθε $\varepsilon > 0$ ισχύει η ακόλουθη ανισότητα:

$$P\left(\max_{1 \leq j \leq t} Z_j > \varepsilon\right) \leq 1 - \frac{1}{\varepsilon} \mathbb{E}_P[Z_1] + \frac{1}{\varepsilon} \sum_{m=1}^t \mathbb{E}_P[Z_m \chi_{\{Z_m > \varepsilon\}}]. \quad (3)$$

Αν επί πλέον η $\{Z_j\}_{j \in \mathbb{N}}$ είναι άνω φραγμένη από έναν θετικό αριθμό ζ τότε

$$\zeta P\left(\max_{1 \leq j \leq t} Z_j \geq \varepsilon\right) \geq \mathbb{E}_P[Z_1] - \mathbb{E}_P[Z_t \chi_{\{\max_{1 \leq j \leq t} Z_j < \varepsilon\}}].$$

Για την απόδειξη της παραπάνω πρότασης, βλ. Koutras & Lyberopoulos (2017, Proposition 1).

4. ΜΙΑ ΕΦΑΡΜΟΓΗ ΣΤΙΣ ΣΤΑΤΙΣΤΙΚΕΣ ΣΥΝΑΡΤΗΣΕΙΣ ΣΑΡΩΣΗΣ

Στην παρούσα ενότητα θεωρούμε μια ακολουθία $\{X_n\}_{n \in \mathbb{N}}$ ανεξάρτητων και ισόνομων δίτιμων δοκιμών επάνω στον Ω με πιθανότητα επιτυχίας $p_n := P(X_n = 1) = p \in (0, 1)$ και πιθανότητα αποτυχίας $q_n := P(X_n = 0) = 1 - p_n = 1 - p =: q$ για κάθε $n \in \mathbb{N}$.

Για οποιοδήποτε σταθερό $k \in \mathbb{N}$ και για κάθε $m \in \mathbb{N}$ ώστε $m \leq k$, η ακολουθία $X_n, X_{n+1}, \dots, X_{n+m-1}$ τ.μ. επάνω στον Ω ονομάζεται ένα **κυλιόμενο παράθυρο (για τη $\{X_n\}_{n \in \mathbb{N}}$) μήκους m** . Ιδιαίτερως, 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 $X_n, X_{n+1}, \dots, X_{n+m}$ μήκους $m \leq k$ που περιέχουν με θετική πιθανότητα τουλάχιστον r -επιτυχίες.

Στο εξής θέτουμε $X_0 := 0$ και θεωρούμε κάθε άθροισμα επί κενού συνόλου δεικτών εξ ορισμού ίσο με το μηδέν. Για κάθε $n \in \mathbb{N}$ και $k \in \mathbb{N}_0$ ορίζουμε την τ.μ. $Y_{n,k}$ επάνω στο Ω μέσω της σχέσης

$$Y_{n,k} := \sum_{j=\max\{n-k+1, 1\}}^n X_j. \quad (4)$$

Για οποιοδήποτε σταθερό $k \in \mathbb{N}$ η ακολουθία $\{Y_{n,k}\}_{n \in \mathbb{N}}$ ονομάζεται η **απαριθμητρία διαδικασία σαρώσεων** μήκους k για την ακολουθία δίτιμων δοκιμών $\{X_n\}_{n \in \mathbb{N}}$. Η τ.μ. $T_r^{(k)}$ επάνω στον Ω που ορίζεται μέσω της σχέσης

$$T_r^{(k)} := \min\{n \in \mathbb{N} : Y_{n,k} \geq r\}$$

ονομάζεται ο **χρόνος αναμονής για την πρώτη εμφάνιση μιας σαρώσης τύπου r/k** .

Προφανώς, για οποιοδήποτε σταθερό $k \in \mathbb{N}$ οι τ.μ. $Y_{n,k}$, $n \in \mathbb{N}$, είναι εξαρτημένες (ως προς το μέτρο πιθανότητας P). Επειδή, όμως, εύκολα μπορεί να αποδειχθεί ότι για οποιοδήποτε σταθερό $k \in \mathbb{N}$ και για κάθε $r, k, t \in \mathbb{N}$ με $r \leq k$ ισχύει η ισότητα $\{T_r^{(k)} \leq t\} = \{\max_{1 \leq n \leq t} Y_{n,k} \geq r\}$, εύλογα προκύπτει το ερώτημα για το αν μεγιστικές ανισότητες μπορούν να χρησιμοποιηθούν για την εξαγωγή άνω και κάτω φραγμάτων για τη σ.χ.π. $F_{r:k}(t; p) := P(T_r^{(k)} \leq t)$, βλ. π.χ. Prakasa Rao (2012, Chapters 2 and 3).

Με αφορμή το παραπάνω ερώτημα, διερευνάται το αν η ακολουθία $\{Y_{n,k}\}_{n \in \mathbb{N}}$, όπου $k \in \mathbb{N}$ σταθερό, ανήκει στις κλάσεις των demi(sub)martingales και N -demi(super)martingales, ενώ στη συνέχεια δίνεται ένα αποτέλεσμα για N -demimartingales.

Το ακόλουθο αποτέλεσμα αποτελεί άμεση συνέπεια του ορισμού της δεσμευμένης μέσης τιμής και της μονοτονίας των εμπλεκόμενων σε αυτό συναρτήσεων, ενώ βάσει αυτού εύκολα διαπιστώνεται ότι η ακολουθία $\{X_n\}_{n \in \mathbb{N}}$ είναι ένα N -demimartingale

κάτω από το P αλλά δεν είναι ένα P -demisubmartingale, βλ. επίσης Koutras & Lyberopoulos (2017, Lemma 2).

Λήμμα 1. Έστω αυθαίρετο αλλά σταθερό $k \in \mathbb{N}$. Για κάθε $n \in \mathbb{N}$ με $n \geq k$ θεωρούμε επίσης μια κατά συντεταγμένες μη φθίνουσα συνάρτηση $f : \mathbb{R}^n \rightarrow \mathbb{R}$, καθώς επίσης και μια ακολουθία $\{h_{i,k}\}_{i \in \{1, \dots, n\}}$ τέτοιων συναρτήσεων επάνω στο $\mathbb{R}^{i \wedge k}$. Τότε ισχύει ότι

$$\eta_{n,k,0}(h_{i,k}; f) \leq \eta_{n,k,1}(h_{i,k}; f),$$

όπου

$$\eta_{n,k,x}(h_{i,k}; f) := \mathbb{E}_P[f(h_1(X_1), \dots, h_n(X_{n-k+1}, \dots, X_n)) \mid \{X_{n-k+1} = x\}], \quad x \in \{0, 1\}.$$

Ιδιαιτέρως,

$$\mathbb{E}_P[f(Y_{1,k}, \dots, Y_{n,k}) \mid \{X_{n-k+1} = 0\}] \leq \mathbb{E}_P[f(Y_{1,k}, \dots, Y_{n,k}) \mid \{X_{n-k+1} = 1\}].$$

Σύμφωνα με τα παραπάνω, συνάγεται ότι οι μεγιστικές ανισότητες είτε για N -demi(super)martingales ή για demi(sub)martingales δεν μπορούν να βρουν εφαρμογή στην περίπτωση της απαριθμητριάς διαδικασίας $\{Y_{n,k}\}_{n \in \mathbb{N}}$, όπου $k \in \mathbb{N}$ σταθερό. Η δυσκολία αυτή μπορεί να υπερκεραστεί αν για οποιοδήποτε σταθερό $k \in \mathbb{N}$ θεωρήσουμε την ακολουθία $\{\tilde{Y}_{n,k}\}_{n \in \mathbb{N}_0}$ τ.μ. επάνω στο Ω που ορίζεται ως εξής:

$$\tilde{Y}_{n,k} := \begin{cases} Y_{n,k}, & \text{αν } n \in \{k, k+1, \dots\} \\ Y_{k,k}, & \text{αν } n \in \{1, \dots, k-1\} \\ 0, & \text{αν } n = 0. \end{cases} \quad (5)$$

Προφανώς, $R_{\tilde{Y}_{n,k}} = \{0, \dots, k\}$ για κάθε $n \in \mathbb{N}$.

Πρόταση 3. Για οποιοδήποτε σταθερό $k \in \mathbb{N}$ η ακολουθία $\{\tilde{Y}_{n,k}\}_{n \in \mathbb{N}}$ είναι ένα N -demimartingale κάτω από το P . Επί πλέον, δεν είναι ένα P -demisubmartingale.

Απόδειξη. Αρχικά σταθεροποιούμε $k \in \mathbb{N}$. Έπειτα παρατηρούμε ότι η ακολουθία $\{\tilde{Y}_{n,k}\}_{n \in \mathbb{N}}$ είναι προσαρμοσμένη στη διύλιση $\{\mathcal{F}_n^{(X)}\}_{n \in \mathbb{N}}$, αφού το ίδιο ισχύει για την $\{Y_{n,k}\}_{n \in \mathbb{N}}$ εξαιτίας της (4). Επί πλέον, για κάθε $n \in \mathbb{N}$ και για κάθε μη φθίνουσα κατά συντεταγμένες συνάρτηση $f : \mathbb{R}^n \rightarrow \mathbb{R}$ τέτοια ώστε να υπάρχει καθεμία από τις μέσες τιμές

$$\tilde{H}_{n,k}(f) := \tilde{H}_{n,k}(\tilde{Y}_{1,k}, \dots, \tilde{Y}_{n,k}; f) := \mathbb{E}_P[(\tilde{Y}_{n+1,k} - \tilde{Y}_{n,k})f(\tilde{Y}_{1,k}, \dots, \tilde{Y}_{n,k})],$$

από την (5) έπεται ότι

$$\tilde{H}_{n,k}(f) = \begin{cases} H_{n,k}(f), & \text{αν } n \in \{k, k+1, \dots\} \\ 0, & \text{αν } n \in \{1, \dots, k-1\}, \end{cases}$$

κάτι που σε συνδυασμό από το Λήμμα 1 συνεπάγεται ότι $\tilde{H}_{n,k}(f) \leq 0$ για κάθε $n \in \mathbb{N}$. Το τελευταίο γεγονός σε συνδυασμό με το ότι $\mathbb{E}_P[\tilde{Y}_{n,k}] = kp < \infty$ για κάθε $n \in \mathbb{N}$ εξασφαλίζει το ότι η $\{\tilde{Y}_{n,k}\}_{n \in \mathbb{N}}$ είναι ένα N -demimartingale κάτω από το P .

Παρατηρούμε επίσης ότι $Y_{n,k} - Y_{n-1,k} = X_n - X_{(n-k)^+}$ για κάθε $n \in \mathbb{N}$, καθώς και ότι ισχύει η σχέση

$$P_{Y_{n,k}} = \mathbf{B}(n \wedge k, p) \quad \text{για κάθε } n \in \mathbb{N},$$

αφού εξ υποθέσεως η $\{X_n\}_{n \in \mathbb{N}}$ είναι P -ανεξάρτητη. Επομένως, έχουμε

$$H_{n,k}(\pi_n) = \mathbb{E}_P[(Y_{n+1,k} - Y_{n,k})Y_{n,k}] = \mathbb{E}_P[X_{n+1}Y_{n,k}] - \mathbb{E}_P[X_{n-k+1}Y_{n,k}] = -pq;$$

Έστω, τώρα, ότι η $\{\tilde{Y}_{n,k}\}_{n \in \mathbb{N}}$ είναι κι ένα P -demisubmartingale. Τότε έχουμε ότι $\tilde{H}_{n,k}(f) = 0$ για κάθε $n \in \mathbb{N}$ και για κάθε μη αρνητική κατά συντεταγμένες μη φθίνουσα συνάρτηση $f : \mathbb{R}^n \rightarrow \mathbb{R}$ τέτοια ώστε να υπάρχει η μέση τιμή $\tilde{H}_{n,k}(f)$. Συνεπώς, για κάθε $n \in \mathbb{N}$ με $n \geq k$ έχουμε ότι $0 = \tilde{H}_{n,k}(\pi_n) = H_{n,k}(\pi_n) = -pq$, που είναι άτοπο. Επομένως, η $\{\tilde{Y}_{n,k}\}_{n \in \mathbb{N}}$ δεν είναι ένα P -demisubmartingale. \square

Εφαρμόζοντας τις Προτάσεις 2 και 3, καθίσταται δυνατή η εξαγωγή φραγμάτων για τη σ.κ.π. του χρόνου αναμονής $T_r^{(k)}$ μέσω του ακόλουθου αποτελέσματος.

Πρόταση 4. Για οποιοδήποτε σταθερό $k \in \mathbb{N}$ και για όλα τα $r, t \in \mathbb{N}$ ώστε $r \leq k < t$, ισχύει ότι

$$\ell(r, k, p) \leq F_{r:k}(t; p) \leq u_*(r, k, t, p) := \min_{i \in \{1,2,3,4\}} u_i(r, k, t, p),$$

όπου

$$\begin{aligned} u_1(r, k, t, p) &:= 1 - \frac{kp}{r-1} + \frac{t}{r-1} M_b(r; k, p), \\ u_2(r, k, t, p) &:= \frac{t}{r} M_b(r; k, p) + \frac{1}{r} \sum_{m=k+1}^t \sum_{y=r+k-m}^{r-1} yb(y; k, p)[1 - F_b(r-1-y; m-k, p)], \\ u_3(r, k, t, p) &:= \frac{k}{r} M_b(r; k, p) + \frac{(t-k)(t-k+3)k}{2r} [1 - F_b(r-1; k, p)], \\ u_4(r, k, t, p) &:= \left[\frac{k}{r} + \frac{(t-k)(t-k+3)}{2r} \right] M_b(r; k, p) \\ &\quad + \frac{p}{r} [F_b(r-2; k-1, p) - F_b(r-1; k, p)] \sum_{m=k+1}^t \sum_{n=k}^m [(n-k) \wedge k] \\ \text{και } \ell(r, k, p) &:= \frac{1}{k} M_b(r; k, p). \end{aligned}$$

Για την απόδειξη και περισσότερες λεπτομέρειες σχετικά με την προηγούμενη πρόταση, παραπέμπουμε στους Koutras & Lyberopoulos (2017).

Παρατηρήσεις 2. Η Πρόταση 4 με μια πρώτη ματιά υποδηλώνει ότι το κάτω φράγμα $\ell(r, k, p)$ είναι μάλλον περιορισμένης πρακτικής αξίας λόγω του ότι δεν εξαρτάται από τον αριθμό των δοκιμών t . Επί πλέον, το $u_4(r, k, t, p)$ φαίνεται να υπερτερεί των άλλων τριών φραγμάτων σύμφωνα με τα παρακάτω:

(α) Η συνθήκη $\lim_{p \rightarrow 0^+} F_{r:k}(t; p) \leq \lim_{p \rightarrow 0^+} u_i(r, k, t, p) = 0$ ισχύει για $i \in \{2, 4\}$ αλλά όχι για $i \in \{1, 3\}$. Πράγματι, τα $u_1(r, k, t, p)$ και $u_3(r, k, t, p)$ τείνουν στη μονάδα και στο $(t - k)(t - k + 3)(k/2r)$, αντίστοιχα, καθώς το p φθίνει στο 0, κάτι που έρχεται σε πλήρη αντίθεση με το διαισθητικά φυσικό γεγονός ότι η πιθανότητα $P(T_r^{(k)} \leq t)$ πρέπει να τείνει στο 0, καθώς συμβαίνει το ίδιο με την πιθανότητα επιτυχίας p κάθε δίτιμης δοκιμής. Επίσης ισχύει ότι $\lim_{p \rightarrow 0^+} \ell(r, k, p) = 0$.

(β) Τα άνω φράγματα $u_1(r, k, t, p)$ και $u_3(r, k, t, p)$ μπορούν να φανούν ιδιαίτερα χρήσιμα για όλες εκείνες τις τιμές του p που δεν διαφέρουν σημαντικά από τα δύο άλλα φράγματα, αφού τα πρώτα παρουσιάζουν σαφώς μικρότερες υπολογιστικές δυσκολίες από τα υπόλοιπα και ιδιαίτερα για μεγάλες τιμές του t .

(γ) Τα άνω φράγματα $u_2(r, k, t, p)$ και $u_4(r, k, t, p)$ φαίνεται να είναι πιο κατάλληλα από τα άλλα δύο, όχι μόνο εξαιτίας του (α) αλλά και επειδή έχουν ελαχίστο «απλοϊκό» τρόπο από τα $u_1(r, k, t, p)$ και $u_3(r, k, t, p)$, βλ. Koutras & Lyberopoulos (2017).

(δ) Τα τρία πρώτα άνω φράγματα παρουσιάζουν κι ένα ακόμα πλεονέκτημα: είναι μη φθίνουσες συναρτήσεις του t , κάτι που γενικά δεν μπορούμε να ισχυριστούμε για το $u_4(r, k, t, p)$.

Παρακάτω, διερευνάται η συμπεριφορά των φραγμάτων της Πρότασης 4 και πιο συγκεκριμένα των άνω φραγμάτων, αφού όπως αναφέρθηκε και στις Παρατηρήσεις 2 το κάτω φράγμα είναι περιορισμένου ενδιαφέροντος.

Η αποτελεσματικότητα των φραγμάτων θα αξιολογηθεί βάσει των αντίστοιχων σχετικών σφαλμάτων για το (βέλτιστο) άνω και το κάτω φράγμα (σχ.σφ.α.φ. και σχ.σφ.κ.φ., αντίστοιχα), δηλαδή των λόγων

$$\frac{u_*(r, k, t, p) - \kappa(r, k, t, p)}{\kappa(r, k, t, p)} \quad \text{και} \quad \frac{|\ell(r, k, p) - \kappa(r, k, t, p)|}{\kappa(r, k, t, p)},$$

αντίστοιχα, όπου με $\kappa(r, k, t, p)$ σημειώνεται η ακριβής τιμή ή τιμή από προσομοίωση της σ.κ.π. του χρόνου αναμονής $T_r^{(k)}$. Η ακριβής σ.κ.π. υπολογίζεται μέσω των Bersimis et al. (2012, Proposition 1) για την περίπτωση των **σχεδόν τέλειων ροών**, δηλαδή σαρώσεων τύπου $(k - 1)/k$ ή $(k - 2)/k$, ενώ για τις υπόλοιπες σαρώσεις η εμπειρική σ.κ.π. υπολογίζεται μέσω προσομοίωσης ως εξής: Μια ακολουθία ανεξάρτητων και ισόνομων δίτιμων δοκιμών με πιθανότητα επιτυχίας p παράγεται κάθε φορά και καταγράφεται ο αριθμός των δοκιμών που απαιτείται μέχρι την εμφάνιση μιας σάρωσης τύπου r/k . Η προσομοιωμένη τιμή της σ.κ.π. υπολογίζεται τότε ως το πηλίκο του πλήθους των προσομοιώσεων στις οποίες εμφανίζεται ο καταγραφόμενος αριθμός δοκιμών t προς τον συνολικό αριθμό των προσομοιώσεων $s = 1000$.

Στους Πίνακες 1 έως και 3, δίνονται η ακριβής και προσομοιωμένη τιμή της σ.κ.π. του χρόνου αναμονής $T_r^{(k)}$, καθώς και οι τιμές των φραγμάτων της $F_{r:k}(t; p)$ μαζί με τα σχετικά σφάλματα των βέλτιστων φραγμάτων, τα οποία επισημαίνονται σε κάθε t με **έντονα γράμματα**.

Το κύριο συμπέρασμα που προκύπτει από την αριθμητική μελέτη των φραγμάτων της Πρότασης 4 είναι ότι η πρακτική αξία τους περιορίζεται στις περιπτώσεις όπου χρειάζεται μια, έστω και λιγότερο εκλεπτυσμένη, προσέγγιση της $F_{r:k}(t; p)$ λόγω της δυσκολίας που απαιτείται για τον υπολογισμό της τελευταίας. Σε οποιαδήποτε άλλη περίπτωση, υπάρχουν πιο αποτελεσματικά σχετικά φράγματα που είναι διαθέσιμα στη βιβλιογραφία — για παράδειγμα αναφέρουμε αυτά των Glaz & Naus (1991).

Παρόλα αυτά, παρακάτω συνοψίζουμε ορισμένες αξιοσημείωτες συμπεριφορές των φραγμάτων της Πρότασης 4:

- Ως προς τις σχεδόν τέλειες ροές, το σχετικό σφάλμα των άνω φραγμάτων επιδεικνύει συμπεριφορά πολυκόρυφης συνάρτησης του t (όταν οι άλλες τρεις παράμετροι r, k, p παραμένουν σταθερές). Από την άλλη μεριά, δεν παρατηρείται κάποιο ιδιαίτερο μοτίβο συμπεριφοράς των σχετικών σφαλμάτων συναρτήσει του p (και για σταθερά r, k, t) με εξαίρεση την περίπτωση μικρών p σε σχεδόν τέλειες ροές, όπου το σχετικό σφάλμα του βέλτιστου άνω φράγματος αυξάνει μαζί με το p (βλ. Πίνακες 1 και 2). Την αντίστροφη συμπεριφορά παρουσιάζει το σχετικό σφάλμα του κάτω φράγματος τόσο ως προς p όσο και ως προς t (και για r, k, t και r, k, p σταθερά, αντίστοιχα). Ειδικότερα, αναφέρεται ότι αυτό τείνει στη μονάδα καθώς μεγαλώνει η πιθανότητα επιτυχίας p .
- Όπως καταδεικνύεται και από τους παρακάτω πίνακες, το βέλτιστο άνω φράγμα για μικρές τιμές του p δίνεται από το $u_4(r, k, t, p)$, κάτι που αναφέρθηκε και στις Παρατηρήσεις 2. Η εν λόγω συμπεριφορά επιβεβαιώνεται και στις περιπτώσεις των σχεδόν τέλειων ροών τύπου $3/5, 4/5$ και $5/6$.
- Τα παραπάνω ευρήματα που αφορούν σχεδόν τέλειες ροές γενικά δεν ισχύουν και για άλλους τύπους σάρωσης. Μάλιστα, στην πλειονότητα των περιπτώσεων (π.χ. για μια σάρωση τύπου $4/14$ με $p = 0.1$), ακόμα και τα πιο αποτελεσματικά άνω φράγματα δεν παράγουν **αποδεκτές**, υπό την έννοια ότι παίρνουν τιμές στο $(0, 1)$, τιμές πέρα των πρώτων t . Το γεγονός αυτό εξηγείται από το ότι δεν αντιστοιχούν όλα τα φράγματα της Πρότασης 4 σε πιθανότητες. Δύναται, όμως, να υπάρξουν εξαιρέσεις όπως αυτή του Πίνακα 3.
- Τέλος, παρατηρούμε ότι για σάρώσεις που δεν είναι σχεδόν τέλειες ροές, τα άνω φράγματα με την μικρότερη υπολογιστική δυσκολία, δηλαδή τα $u_1(r, k, t, p)$ και $u_3(r, k, t, p)$, αποδεικνύονται συνήθως και τα πιο αποτελεσματικά.

ABSTRACT

We first prove some new maximal inequalities for N -demi(super)martingales. As a result, new bounds for the cumulative distribution function of the waiting time for the first occurrence of a scan of type r/k , in a sequence of independent and

identically distributed binary trials (with probability of success p), are obtained. Finally, a numerical study is carried out for assessing the quality of the derived bounds.

Ευχαριστίες: Ο Δ.Α. θα ήθελε να αφιερώσει την παρούσα εργασία στη μνήμη του πατέρα του Παναγιώτη (Τάκη). Η σχετική έρευνα πραγματοποιήθηκε στο τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς και 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 II. 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.
- Bersimis, S., Koutras, M.V. and Papadopoulos, G.K. (2014). Waiting time for an almost perfect run and applications in statistical process control. *Methodol. Comput. Appl. Probab.* **16**, 207–222.
- Chen, C. and Karlin, S. (2007). r -scans statistics of a Poisson process with events transformed by duplications, deletions and displacements. *Adv. in Appl. Probab.* **39**, 799–825.
- Christofides, T.C. (2003). Maximal inequalities for N -demimartingales. *Arch. Inequal. Appl.* **1**, 387–397.
- Dai, P., Shen, Y., Hu, S. and Yang, W. (2014). Some results for demimartingales and N -demimartingales. *J. Inequal. Appl.* **2014**:489.
- Fu, J.C. (2001). Distribution of the scan statistic for a sequence of bistate trials. *J. Appl. Probab.* **38**, 908–916.
- Glaz, J. and Koutras, M.V. (eds.) (2019). *Handbook of Scan Statistics*. Springer, New York, NY.
- Glaz, J. and Naus, J.I. (1991). Tight bounds and approximations for scan statistic probabilities for discrete data. *Ann. Appl. Probab.* **1**, 306–318.
- Hu, S.H., Yang, W.Z., Wang, X.J. and Shen, Y. (2010). A note on the inequalities for N -demimartingales and demimartingales. *J. Systems Sci. Math. Sci.* **30**, 1052–1058 (in Chinese).
- Koutras, M.V. and Lyberopoulos, D.P. (2017). New maximal inequalities for N -demimartingales with scan statistic applications. *J. Appl. Probab.* **54**, 363–378.
- Newman, C.M. and Wright, A.L. (1982). Associated random variables and martingale inequalities. *Z. Wahrsch. Theorie und Verw. Gebiete* **59**, 361–371.
- Papastavridis, S.G. and Koutras, M.V. (1993). Bounds for reliability of consecutive k -within- m -out-of- n :F systems. *IEEE Trans. Rel.* **42**, 156–160.
- Pozdnyakov, V. and Steele, J.M. (2009). *Martingale Methods for Patterns and Scan Statistics*. In J. Glaz, V. Pozdnyakov, S. Wallenstein (eds.) *Scan Statistics: Methods and Applications*, pp. 289–317. Birkhäuser, Boston.
- Pozdnyakov, V., Glaz, J., Kulldorff, M. and Steele, J.M. (2005). A martingale approach to scan statistics. *Ann. Inst. Statist. Math.* **57**, 21–37.

Prakasa Rao, B.L.S. (2012). *Associated Sequences, Demimartingales and Nonparametric Inference*. Probability and its Applications, Springer Basel AG.

Wu, T.L., Glaz, J. and Fu, J.C. (2013). Discrete, continuous and conditional multiple window scan statistics. *J. Appl. Probab.* **50**, 1089–1101.

ΠΑΡΑΡΤΗΜΑ Α: ΠΙΝΑΚΕΣ

Πίνακας 1: Φράγματα & Ακριβής σ.κ.π.: $(r, k, p) = (5, 6, 0.1)$, $\ell(5, 6, 0.1) = 4.6 \times 10^{-5}$

t	ακριβής	u_1	u_2	u_3	u_4	u_*	σχ.σφ.κ.φ.	σχ.σφ.α.φ.
7	.000096	.850483	.000484	.003631	.000434	.000434	.518	3.539
12	.000294	.850828	.006173	.044881	.001368	.001368	.843	3.659
17	.000492	.851173	.033520	.127381	.002549	.002549	.906	4.184
22	.000690	.851518	.103211	.251131	.003894	.003894	.933	4.647
27	.000887	.851863	.232137	.416131	.005405	.005405	.948	5.090
32	.001085	.852208	.429564	.622381	.007080	.007080	.958	5.524
37	.001283	.852553	.697096	.869881	.008921	.008921	.964	5.952
42	.001481	.852898	1.030550	1.158630	.010926	.010926	.969	6.378
47	.001679	.853243	1.422240	1.488630	.013097	.013097	.973	6.802
52	.001876	.853588	1.863010	1.859880	.015432	.015432	.975	7.225

Πίνακας 2: Φράγματα & Ακριβής σ.κ.π.: $(r, k, p) = (5, 6, 0.15)$, $\ell(5, 6, 0.15) = 3.3 \times 10^{-4}$

t	ακριβής	u_1	u_2	u_3	u_4	u_*	σχ.σφ.κ.φ.	σχ.σφ.α.φ.
7	.000673	.778508	.003465	.026326	.003153	.003153	.504	3.685
12	.001998	.781014	.037089	.325330	.010159	.010159	.833	4.085
17	.003320	.783520	.171040	.923338	.019508	.019508	.899	4.876
22	.004641	.786026	.455616	1.820350	.030651	.030651	.928	5.605
27	.005959	.788532	.900337	3.016370	.043588	.043588	.944	6.314
32	.007276	.791038	1.485180	4.511380	.058319	.058319	.954	7.015
37	.008592	.793544	2.178270	6.305410	.074843	.074843	.961	7.711
42	.009905	.796050	2.948110	8.398440	.093162	.093162	.966	8.406
47	.011217	.798556	3.769020	10.790500	.113275	.113275	.970	9.099
52	.012527	.801062	4.622320	13.481500	.135182	.135182	.973	9.792

Πίνακας 3: Φράγματα & Προσομοιωμένη σ.κ.π.: $(r, k, p) = (8, 14, 0.2)$, $\ell(8, 14, 0.2) = 0.0014$

t	προσ.	u_1	u_2	u_3	u_4	u_*	σχ.σφ.κ.φ.	σχ.σφ.α.φ.
16	.003	.644823	.044702	1.376750	.046113	.044702	.533	13.901
17	.007	.647624	.054032	2.450700	.055227	.054032	.800	6.719
18	.010	.650426	.067214	3.793140	.066332	.066332	.860	5.633
19	.015	.653227	.085159	5.404070	.079312	.079312	.907	4.287
20	.022	.656028	.108819	7.283480	.094052	.094052	.936	3.275
22	.032	.661631	.177140	11.847800	.128353	.128353	.956	3.011
26	.055	.672837	.422727	24.198200	.213012	.213012	.975	2.873
28	.066	.678440	.611463	31.984300	.261527	.261527	.979	2.963
32	.080	.689646	1.136690	50.778400	.368625	.368625	.982	3.608
33	.084	.692447	1.299270	56.148200	.397497	.397497	.983	3.732
35	.090	.698050	1.660980	67.693100	.457759	.457759	.984	4.086
40	.099	.712057	2.761350	101.25400	.623095	.623095	.986	5.294
41	.107	.714858	3.011290	108.77200	.658679	.658679	.987	5.156
42	.111	.717660	3.269950	116.55800	.695103	.695103	.987	5.262
43	.115	.720461	3.536770	124.612000	.732365	.720461	.988	5.265
44	.119	.723263	3.811230	132.936000	.770466	.723263	.988	5.078
64	.165	.779291	10.227700	355.780000	1.708690	.779291	.992	3.723
74	.195	.807305	13.690600	507.475000	2.303650	.807305	.993	3.140
84	.201	.835320	17.181800	686.020000	2.982520	.835320	.993	3.156

ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΤΩΝ ΔΙΑΤΡΟΦΙΚΩΝ ΣΥΝΗΘΕΙΩΝ ΤΩΝ ΕΦΗΒΩΝ ΣΤΗΝ ΕΛΛΑΔΑ

Z. Κυρανά, A. Ρακιτζής

Τμήμα Στατιστικής και Αναλογιστικών-Χρηματοοικονομικών Μαθηματικών,
Πανεπιστήμιο Αιγαίου
reniakyrana@gmail.com, arakitz@aegean.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα μελέτη διερευνάται το προφίλ των διατροφικών συνηθειών των εφήβων στην Ελλάδα, με στόχο να διαπιστωθεί η επίδραση αυτού, μαζί με άλλους διατροφικούς και γεωγραφικούς παράγοντες, στο Δείκτη Μάζας Σώματος (ΔΜΣ) και συνεπώς στην υγεία τους. Για τη συλλογή των δεδομένων χρησιμοποιήθηκε ερωτηματολόγιο εγκεκριμένο από το Παιδαγωγικό Ινστιτούτο, που σχεδιάστηκε στο Τμήμα Διατροφής και Διαιτολογίας του Αλεξάνδρειου Τ.Ε.Ι. Θεσσαλονίκης και χορηγήθηκε σε 42129 μαθητές γυμνασίων και λυκείων όλων των γεωγραφικών διαμερισμάτων της χώρας. Από τη στατιστική ανάλυση διαπιστώθηκε ότι, μεταξύ άλλων, θετικά επιδρούν στην υγεία των εφήβων οι νορμοβαρείς γονείς, η επάρκεια του ύπνου και η κατανάλωση πρωινού και δεκατιανού. Περαιτέρω διερεύνησης χρήζει η σημαντικότητα της επίδρασης της ημερήσιας κατανάλωσης νερού, στη διαμόρφωση των διατροφικών τους συνηθειών.

1. ΕΙΣΑΓΩΓΗ

Ο ρόλος της διατροφής στην ατομική υγεία, η οποία αποτελεί σημαντική προσωπική και κοινωνική αξία, είναι καθοριστικός. Η διατροφή κατά την παιδική και εφηβική ηλικία θεωρείται ιδιαίτερα κρίσιμη, επειδή συνδέεται με σοβαρές ασθένειες στην ενήλικη ζωή, ενώ πολλές από τις διατροφικές συνήθειες που διαμορφώνονται σε αυτές τις ηλικίες, διατηρούνται κατά βάση अपαράλλακτες για το υπόλοιπο της ζωής (Λουμάκου και συν. 2006). Διεθνώς, ένας από τους πιο διαδεδομένους δείκτες της σωματικής διάπλασης και ειδικότερα του υπερβάλλοντος βάρους και της παχυσαρκίας είναι ο Δείκτης Μάζας Σώματος (ΔΜΣ) (Peterson et al. 2017), ο οποίος συσχετίζει το σωματικό βάρος με το ύψος και υπολογίζεται ως:

$$\Delta\text{Μ}\Sigma = \text{ΒΑΡΟΣ (kg)} / \text{ΥΨΟΣ}^2 (m^2)$$

Με βάση την τιμή του ΔΜΣ (δείτε π.χ. Cole et al. 2000), τα άτομα χωρίζονται σε λιποβαρή ($\Delta\text{Μ}\Sigma < 18.5$), νορμοβαρή ($18.5 \leq \Delta\text{Μ}\Sigma \leq 24.9$), υπέρβαρα ($25.0 \leq \Delta\text{Μ}\Sigma \leq 29.9$) και παχύσαρκα ($\Delta\text{Μ}\Sigma \geq 30.0$).

Σε διεθνές και εθνικό επίπεδο αυξάνονται διαρκώς τα περιστατικά της εφηβικής παχυσαρκίας, με τον μεγαλύτερο επιπολασμό να εμφανίζεται στις χώρες που περιβάλλουν τη Μεσόγειο θάλασσα (Tzotzas & Krassas, 2004, Papandreou et al. 2008). Επιπρόσθετα, σε πρόσφατες μελέτες (Tzotzas et al. 2011, Patsopoulou et al. 2016) διαπιστώθηκε ότι ο επιπολασμός της εφηβικής παχυσαρκίας δεν διαφέρει σε σύγκριση με τις υπόλοιπες Ευρωπαϊκές χώρες του Νότου. Σύμφωνα με τους Ruditis et al. 2009, η παχυσαρκία των νέων στην Ελλάδα παρουσιάζει αυξητική τάση τα τελευταία 30 χρόνια. Από δύο σχετικές έρευνες του Κέντρου Ερευνών Τεχνολογικής Ανάπτυξης και Καινοτομίας (ΚΕ.Τ.Α.Κ.), τα έτη 2009 και 2013 προκύπτει ότι, τις τελευταίες δεκαετίες έχει αυξηθεί ανησυχητικά στην Ελλάδα η συχνότητα εμφάνισης νοσημάτων φθοράς που σχετίζονται με τη διατροφή σε ολοένα και μικρότερες ηλικίες. Από την άλλη, υπάρχουν και έρευνες που δείχνουν ότι, η πλειονότητα των εφήβων στην Ελλάδα εμφανίζεται ως νορμοβαρής με αυξημένο όμως κίνδυνο εμφάνισης παχυσαρκίας (Κουκ, 2002, Κυρανά, 2018).

Στόχος της παρούσας εργασίας είναι η διερεύνηση της ύπαρξης (στατιστικά) σημαντικών επιδράσεων διατροφικών και γεωγραφικών παραγόντων στις διατροφικές συνήθειες των εφήβων στην Ελλάδα, όπως αυτές αποτυπώνονται στο ΔΜΣ και κατ' επέκταση στην υγεία τους.

2. ΣΥΛΛΟΓΗ ΔΕΔΟΜΕΝΩΝ

Η συλλογή των δεδομένων της έρευνας έγινε με χρήση ερωτηματολογίου, το οποίο σχεδιάστηκε στο Τμήμα Διατροφής και Διαιτολογίας του Α.Τ.Ε.Ι. Θεσσαλονίκης και εγκρίθηκε από τη Διεύθυνση Δευτεροβάθμιας Εκπαίδευσης του Υπουργείου Παιδείας και το Παιδαγωγικό Ινστιτούτο. Από την Εθνική Στατιστική Υπηρεσία Ελλάδος (ΕΣΥΕ) αναζητήθηκε ο συνολικός πληθυσμός των εφήβων της χώρας και η κατανομή ανά φύλο στους 52 νομούς της. Η επιλογή των σχολείων έγινε από τους πίνακες της Δευτεροβάθμιας Εκπαίδευσης [Κατάλογος Σχολείων Δευτεροβάθμιας Εκπαίδευσης - e-raideia], με κριτήρια την πληθυσμιακή κατανομή ανά δημοτική ενότητα (αστική, ημιαστική και αγροτική) και την γεωφυσική της μορφολογία (παραθαλάσσια, ενδοχώρα, πεδινή & ορεινή). Στόχος ήταν η όσο το δυνατόν μεγαλύτερη αντιπροσωπευτικότητα του δείγματος στις ποικίλες κοινωνικές και οικονομικές συνθήκες διαβίωσης. Η έρευνα διήρκεσε τέσσερα έτη (2011-2015). Τα ερωτηματολόγια συμπληρώθηκαν από τους έφηβους μαθητές, υπό την επίβλεψη και καθοδήγηση ομάδων προπτυχιακών και μεταπτυχιακών φοιτητών του Τμήματος Διατροφής και Διαιτολογίας του Α.Τ.Ε.Ι. Θεσσαλονίκης. Το μέγεθος του δείγματος ήταν $N=42129$ έφηβοι. Επιχειρήθηκε η λήψη του από μαθητές γυμνασίων και λυκείων, μέσω δειγματοληψίας κατά ομάδες, ώστε η ηλικιακή κατανομή και η εκπροσώπηση των δύο φύλων στο συνολικό δείγμα, να ανταποκρίνονται (το δυνατόν) στις εκτιμήσεις της ΕΣΥΕ πανελλαδικά.

3. ΑΠΟΤΕΛΕΣΜΑΤΑ ΣΤΑΤΙΣΤΙΚΗΣ ΑΝΑΛΥΣΗΣ

Στην παρούσα ενότητα παρουσιάζονται τα αποτελέσματα της στατιστικής ανάλυσης των δεδομένων από τα διαθέσιμα ερωτηματολόγια. Οι προς μελέτη ποσοτικές

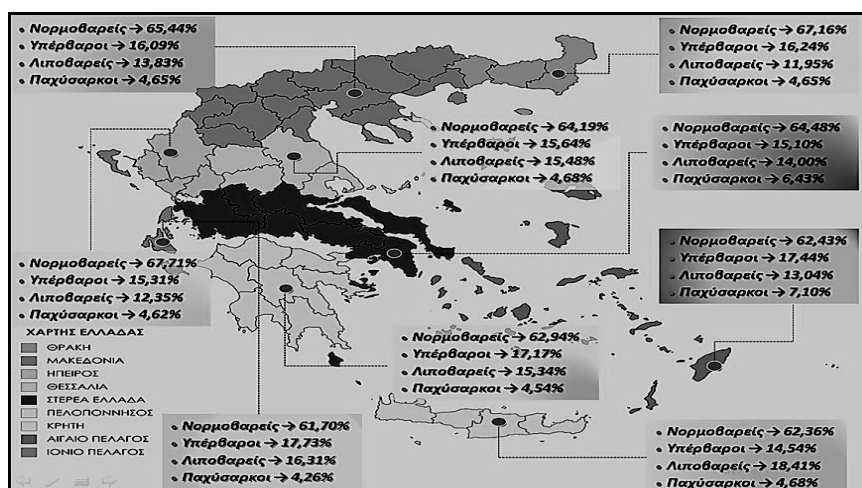
μεταβλητές είναι οι τιμές του ΔΜΣ εφήβου, πατέρα, και μητέρας, οι ώρες ύπνου/ημέρα, η ημερήσια κατανάλωση νερού (σε ποτήρια), αριθμός των ημερήσιων γευμάτων που καταναλώνουν οι έφηβοι, η εβδομαδιαία συχνότητα κατανάλωσης πρωινού, η εβδομαδιαία συχνότητα παραγγελίας φαγητού μέσω delivery καθώς και η εβδομαδιαία συχνότητα κατανάλωσης τροφίμων σε fast food, πιτσαρία, εστιατόριο και ταβέρνα.

Επίσης, οι προς μελέτη κατηγορικές μεταβλητές, μαζί με τα διαφορετικά επίπεδα κάθε μεταβλητής είναι: το Φύλο των εφήβων (δύο επίπεδα, αγόρι / κορίτσι), ο ΔΜΣ (τέσσερα επίπεδα, λιποβαρής / νορμοβαρής / υπέρβαρος / παχύσαρκος), η Μορφή οικογένειας (πέντε επίπεδα, χωρίς γονείς / με δύο γονείς / με ένα γονέα λόγω διαζυγίου / με ένα γονέα λόγω θανάτου / μονογονεϊκή), η κατανάλωση φαγητού μέσω delivery (δύο επίπεδα, ναι / όχι), ο Λόγος επιλογής τροφίμων (πέντε επίπεδα, εμφάνιση / γεύση / θρεπτική αξία / θερμίδες / κόστος) και η ημερήσια κατανάλωση νερού (δύο επίπεδα, 1-9 ποτήρια νερό ημερησίως / >9 ποτήρια νερό ημερησίως).

3.1 Περιγραφικά στατιστικά του ΔΜΣ των εφήβων

Στον παρακάτω χάρτη της Ελλάδας (Σχήμα 1) παρουσιάζονται τα ποσοστά εμφάνισης των τεσσάρων κατηγοριών του ΔΜΣ ανά γεωγραφικό διαμέρισμα.

Σχήμα 1. Χάρτης ΔΜΣ των εφήβων ανά γεωγραφικό διαμέρισμα



Από το Σχήμα 1 παρατηρείται πως, στο γεωγραφικό διαμέρισμα της Ηπείρου εμφανίζεται το μεγαλύτερο ποσοστό νορμοβαρών εφήβων (67.71%), στην Κρήτη εμφανίζεται το μεγαλύτερο ποσοστό λιποβαρών εφήβων (18.41%) ενώ στα νησιά του Αιγαίου Πελάγους εμφανίζεται το μεγαλύτερο ποσοστό υπέρβαρων και παχύσαρκων εφήβων (24.54%). Σε πανελλαδική κλίμακα προκύπτει ότι, το 65% των εφήβων είναι νορμοβαρείς, το 14.02% λιποβαρείς, ενώ το 20.93% αγγίζουν οι έφηβοι με υπερβάλλον σωματικό βάρος (υπέρβαροι και παχύσαρκοι). Αυτό σημαίνει ότι περίπου, το 35% των εφήβων αποκλίνει από τα διεθνώς καθιερωμένα πρότυπα. Συμπεράσματα ανάλογων μελετών επιβεβαιώνονται και κρίνονται ανησυχητικά,

επειδή αποτελούν την αφετηρία εκδήλωσης ποικίλων μεταβολικών διαταραχών και νοσημάτων (WHO, 2009). Επιπλέον, ανησυχητικά θεωρούνται και τα υψηλά ποσοστά εφήβων που βρίσκονται σε κατάσταση υποθρεψίας (14.02%) πανελλαδικά, αφού ανάλογες μελέτες (Birbilis et al. 2013) έδειξαν πως το ποσοστό των λιποβαρών εφήβων δεν ξεπερνά το 5% του συνολικού πληθυσμού τους.

3.2 Μη Παραμετρικοί Έλεγχοι

Για τον έλεγχο της σημαντικότητας της επίδρασης των παραγόντων όπως το φύλο, η οικογένεια και οι παραγγελίες delivery, χρησιμοποιήθηκαν οι μη-παραμετρικοί έλεγχοι Mann-Whitney και Kruskal-Wallis (δείτε π.χ. Ξεκαλάκη, 2001). Το επίπεδο σημαντικότητας (ε.σ.) για όλους τους στατιστικούς ελέγχους που ακολουθούν είναι $\alpha=0.05$. Ειδικότερα, για την επίδραση του φύλου, η H_0 του ελέγχου υποθέσεων διατυπώνεται ως εξής:

H_0 : Δεν υπάρχει διαφορά στη μέση τιμή του ΔΜΣ ως προς το «Φύλο».

Με ανάλογο τρόπο διατυπώνεται η H_0 , για τον έλεγχο της επίδρασης στη μέση τιμή του ΔΜΣ της συχνότητας κατανάλωσης φαγητού, μέσω delivery (μεταβλητή Delivery). Τα αποτελέσματα των ελέγχων (p-values) παρουσιάζονται στον Πίνακα 1, για καθένα από τα εννέα γεωγραφικά διαμερίσματα.

Πίνακας 1. Αποτελέσματα του Έλεγχου Σημαντικότητας των παραγόντων «Φύλο» και «Delivery» ανά γεωγραφικό διαμέρισμα της Ελλάδας

Γεωγραφικό Διαμέρισμα	Φύλο	Delivery
ΘΡΑΚΗ	.054	.004
ΜΑΚΕΔΟΝΙΑ	.000	.001
ΗΠΕΙΡΟΣ	.000	.571
ΘΕΣΣΑΛΙΑ	.000	.005
ΣΤΕΡΕΑ ΕΛΛΑΔΑ	.981	.449
ΠΕΛΟΠΟΝΝΗΣΟΣ	.087	.047
ΚΡΗΤΗ	.005	.348
ΑΙΓΑΙΟ ΠΕΛΑΓΟΣ	.000	.129
ΙΟΝΙΟ ΠΕΛΑΓΟΣ	.261	.667

Ως προς τη σχέση του ΔΜΣ με το φύλο των εφήβων παρατηρείται ότι, στα γεωγραφικά διαμερίσματα της Μακεδονίας, της Ηπείρου, της Θεσσαλίας, της Κρήτης και των Αιγαιοπελαγίτικων νήσων, οι τιμές των p-values $< 0.05 = \alpha$. Αυτό σημαίνει ότι, απορρίπτεται η αντίστοιχη H_0 και συνεπώς στα εν λόγω γεωγραφικά διαμερίσματα υπάρχουν σαφείς ενδείξεις για σημαντική διαφοροποίηση στη μέση τιμή του ΔΜΣ μεταξύ αγοριών και κοριτσιών. Το γεγονός κρίνεται αναμενόμενο, λόγω της διαφορετικής ανατομίας και φυσιολογίας, που κατ' επέκταση συμβάλλει στον διαφορετικό μεταβολισμό/ρυθμό ανάπτυξης των δύο φύλων.

Σε ότι αφορά τη σχέση του ΔΜΣ με την επιλογή των εφήβων να τρέφονται με φαγητό μέσω παραγγελιών delivery παρατηρείται ότι, στα γεωγραφικά διαμερίσματα της Θράκης, της Μακεδονίας, της Θεσσαλίας και της Πελοποννήσου απορρίπτεται η αντίστοιχη H_0 σε ε.σ. 5% και άρα, υπάρχουν σαφείς ενδείξεις για διαφοροποίηση στη μέση τιμή του ΔΜΣ μεταξύ των εφήβων που επιλέγουν να τρέφονται με φαγητά μέσω παραγγελιών delivery (Ομάδα Α) και αυτών που δεν τα προτιμούν (Ομάδα Β), αντίστοιχα. Συγκεκριμένα, παρατηρείται μεγαλύτερη δειγματική μέση τιμή του ΔΜΣ στην Ομάδα Α έναντι της Ομάδας Β και άρα ο ΔΜΣ των εφήβων επηρεάζεται αρνητικά (πηγή κινδύνου για εμφάνιση υπερβάλλοντος σωματικού βάρους, στους εφήβους που επιλέγουν να τρέφονται μέσω delivery). Το συμπέρασμα αυτό δικαιολογείται από την υποβαθμισμένη ποιότητα και θρεπτική αξία του φαγητού, που συνήθως διακινείται μέσω των υπηρεσιών delivery (Κουκ, 2002). Στα υπόλοιπα γεωγραφικά διαμερίσματα δε παρατηρήθηκε σημαντική επίδραση του παράγοντα «Delivery».

Στη συνέχεια, μελετήθηκε η επίδραση του παράγοντα «Μορφή της Οικογένειας» στο ΔΜΣ των εφήβων. Έστω $\mu_j, j = 0, 1, 2, 3, 4$, η μέση τιμή του ΔΜΣ για τους εφήβους των πέντε διαφορετικών επιπέδων του παράγοντα ($j = 0$: χωρίς γονείς, $j = 1$: με δύο γονείς, $j = 2$: με έναν γονέα λόγω διαζυγίου, $j = 3$: με έναν γονέα λόγω θανάτου, $j = 4$: μονογονεϊκή οικογένεια). Τότε, η H_0 διατυπώνεται ως $H_0: \mu_0 = \mu_1 = \mu_2 = \mu_3 = \mu_4$ και για τον έλεγχο της σημαντικότητας της επίδρασης χρησιμοποιήθηκε ο έλεγχος των Kruskal-Wallis.

Τα αποτελέσματα του ελέγχου (p-values), για καθένα από τα γεωγραφικά διαμερίσματα της Ελλάδας, παρουσιάζονται στον Πίνακα 2. Τα αποτελέσματα δείχνουν ότι, ως προς τη σχέση του ΔΜΣ με την μορφή της οικογένειας των εφήβων, στη Μακεδονία, τη Θράκη, την Ήπειρο, την Πελοπόννησο, την Κρήτη και τα νησιά του Ιονίου, οι τιμές p-value > 0.05. Αυτό σημαίνει ότι στα συγκεκριμένα γεωγραφικά διαμερίσματα της χώρας ο ΔΜΣ των εφήβων δεν επηρεάζεται από τη μορφή της οικογένειας των εφήβων. Αντίθετα, στα υπόλοιπα γεωγραφικά διαμερίσματα, της Θεσσαλίας, της Στερεάς Ελλάδας και των Αιγαίοπελαγίτικων νήσων φαίνεται να υπάρχει στατιστικά σημαντική διαφορά μεταξύ των εν λόγω μεταβλητών.

Στη συνέχεια, για τα γεωγραφικά διαμερίσματα όπου η H_0 απορρίφθηκε, προχωρήσαμε στη διεξαγωγή πολλαπλών συγκρίσεων (Πίνακας 3) με χρήση της μεθόδου του Dunn (1964) για μη-παραμετρικές πολλαπλές συγκρίσεις (*Dunn's Multiple comparison test*), σε επίπεδο σημαντικότητας α / κ , όπου $\alpha = 0.05$ και κ είναι το πλήθος των ανά δύο συγκρίσεων (εδώ $\kappa = 10$). Ο εν λόγω έλεγχος πραγματοποιεί μια σειρά ελέγχων των Mann-Whitney με $H_0: \mu_s = \mu_l, s \neq l$ προκειμένου να έχουμε μια εικόνα σχετικά με πιθανή ομαδοποίηση των επιπέδων του παράγοντα. Ο συνολικός βαθμός εμπιστοσύνης των πιθανών ομάδων είναι τουλάχιστον $(1 - \alpha)100\%$.

Πίνακας 2. Αποτελέσματα Ελέγχου Σημαντικότητας του παράγοντα «Μορφή της Οικογένειας» ανά γεωγραφικό διαμέρισμα της Ελλάδας

Γεωγραφικό Διαμέρισμα	Μορφή οικογένειας
ΘΡΑΚΗ	.159
ΜΑΚΕΔΟΝΙΑ	.748
ΗΠΕΙΡΟΣ	.152
ΘΕΣΣΑΛΙΑ	.007
ΣΤΕΡΕΑ ΕΛΛΑΔΑ	.000
ΠΕΛΟΠΟΝΝΗΣΟΣ	.074
ΚΡΗΤΗ	.682
ΑΙΓΑΙΟ ΠΕΛΑΓΟΣ	.023
ΙΟΝΙΟ ΠΕΛΑΓΟΣ	.093

Πίνακας 3. Αποτελέσματα Πολλαπλών Συγκρίσεων

Γεωγραφικό Διαμ.	Μορφή οικογένειας και Πιθανές Ομάδες
ΘΕΣΣΑΛΙΑ	$\mu_1 \neq \mu_3$ και $\{\mu_0, \mu_1, \mu_2, \mu_4\}$, $\{\mu_0, \mu_3, \mu_2, \mu_4\}$
ΣΤΕΡ. ΕΛΛΑΔΑ	$\mu_1 \neq \mu_3$, $\mu_1 \neq \mu_2$, $\mu_2 \neq \mu_3$ και $\{\mu_0, \mu_1, \mu_4\}$, $\{\mu_0, \mu_2, \mu_4\}$, $\{\mu_0, \mu_3, \mu_4\}$
ΑΙΓΑΙΟ ΠΕΛ.	$\mu_3 \neq \mu_4$ και $\{\mu_0, \mu_1, \mu_2, \mu_3\}$, $\{\mu_0, \mu_1, \mu_2, \mu_4\}$

Τα αποτελέσματα των πολλαπλών συγκρίσεων έδειξαν ότι στη Θεσσαλία, ο ΔΜΣ διαφέρει μεταξύ των εφήβων που ζουν σε πυρηνικές οικογένειες και αυτών που ζουν σε οικογένειες με ένα γονέα λόγω θανάτου. Στη Στερεά Ελλάδα η διαφοροποίηση αυτή συναντάται και σε οικογένειες με ένα γονέα λόγω διαζυγίου, ενώ στα νησιά του Αιγαίου Πελάγους μεταξύ αυτών που ζουν σε οικογένειες με ένα γονέα λόγω θανάτου και σε μονογονεϊκές οικογένειες. Στην υπόλοιπη Ελλάδα δεν παρατηρήθηκαν διαφοροποιήσεις. Η σημαντικότητα της επίδρασης της μορφής της οικογένειάς στο ΔΜΣ των εφήβων αποδίδεται στην επίδραση (ψυχική και κοινωνική) που έχει στους νέους η απώλεια των γονέων τους (Ardelt & Day, 2002), γεγονός που επηρεάζει τη συμπεριφορά και τις διατροφικές τους συνήθειες. Οι πιθανές ομάδες που δημιουργούνται μαρτυρούν την ομοιότητα που εμφανίζει ο ΔΜΣ των εφήβων μεταξύ των επιπέδων της μορφής της οικογένειας που απαρτίζουν την κάθε ομάδα.

Με τον ίδιο τρόπο (και με χρήση του Kruskal-Wallis test) ελέγχθηκε η σημαντικότητα των παρακάτω παραγόντων ως προς την επίδρασή τους στη μέση τιμή του ΔΜΣ:

- «ΔΜΣ πατέρα» (αντ. «ΔΜΣ μητέρας»), με 4 επίπεδα και $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ όπου $\mu_1, \mu_2, \mu_3, \mu_4$ είναι η μέση τιμή του ΔΜΣ των εφήβων με λιποβαρή, νορμοβαρή, υπέρβαρο και παχύσαρκο γονέα, αντίστοιχα.

- «Λόγος επιλογής τροφίμων», με πέντε επίπεδα και $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ όπου $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$ είναι η μέση τιμή του ΔΜΣ των εφήβων, οι οποίοι επιλέγουν τρόφιμα λόγω εμφάνισης, γεύσης, θρεπτικής αξίας, θερμίδων και κόστους, αντίστοιχα.

Επίσης, εξετάστηκε η επίδραση που έχει ο ΔΜΣ (ως κατηγορική μεταβλητή, τέσσερα επίπεδα) στο μέσο ημερήσιο αριθμό ωρών ύπνου, στο μέσο ημερήσιο αριθμό ποτηριών νερού που καταναλώνουν οι έφηβοι, στο μέσο ημερήσιο αριθμό γευμάτων και στο μέσο εβδομαδιαίο αριθμό γευμάτων μέσω delivery.

Η διεξαγωγή των παραπάνω ελέγχων έδειξε ότι ο ΔΜΣ των εφήβων επηρεάζεται από τον ΔΜΣ των γονέων πανελλαδικά, αφού εντοπίζονται διαφορές στον ΔΜΣ των εφήβων μεταξύ των τεσσάρων κατηγοριών του ΔΜΣ των γονέων τους. Επιπλέον, επηρεάζεται και από τις ώρες ύπνου/24ωρο. Συγκεκριμένα, διαπιστώθηκε ότι οι ώρες ύπνου των εφήβων διαφοροποιούνται μεταξύ των τεσσάρων κατηγοριών του ΔΜΣ για τα γεωγραφικά διαμερίσματα της Θράκης, της Μακεδονίας, της Θεσσαλίας, της Στερεάς Ελλάδας και του Αιγαίου Πελάγους. Ο ΔΜΣ των εφήβων επηρεάζεται πανελλαδικά και από την ημερήσια κατανάλωση νερού αυτών, αφού διαπιστώθηκε ότι η ποσότητα του νερού που καταναλώνουν διαφέρει ανάλογα με το αν αυτοί είναι λιποβαρείς, νορμοβαρείς, υπέρβαροι ή παχύσαρκοι. Επίσης, διαπιστώθηκε πως η ημερήσια αύξηση του αριθμού των γευμάτων σχετίζεται με τον χαμηλό ΔΜΣ των εφήβων και κυρίως των αγοριών. Οι έφηβοι δήλωσαν ότι λαμβάνουν κατά μέσο όρο τρία έως τέσσερα γεύματα την ημέρα. Όπως ήταν αναμενόμενο, ο μέσος ημερήσιος αριθμός των γευμάτων φαίνεται πως διαφοροποιείται από την κατηγορία του ΔΜΣ των εφήβων σ' ολόκληρη την επικράτεια. Τέλος, ο ΔΜΣ των εφήβων επηρεάζεται και από την εβδομαδιαία συχνότητα κατανάλωσης πρωινού πανελλαδικά, η οποία φαίνεται να διαφοροποιείται μεταξύ των τεσσάρων κατηγοριών ΔΜΣ σε όλες της περιοχές της χώρας, με εξαίρεση τη Θράκη, την Ήπειρο και την Κρήτη. Για περισσότερες λεπτομέρειες, αλλά και για όλα τα αποτελέσματα από τη διεξαγωγή των ελέγχων, δείτε Κυρανά, (2019).

3.3 Προσαρμογή Μοντέλων Πολυωνυμικής Λογιστικής Παλινδρόμησης

Η Πολυωνυμική Λογιστική Παλινδρόμηση είναι μια τεχνική σχεδιασμένη για την πραγματοποίηση ανάλυσης δεδομένων που αφορούν τη μελέτη και την πρόβλεψη τιμών κάποιας κατηγορικής εξαρτημένης μεταβλητής με περισσότερες από δύο κατηγορίες/επίπεδα (Agresti, 2001). Στην παρούσα μελέτη η μεταβλητή αυτή είναι ο ΔΜΣ των εφήβων με τις τέσσερις κατηγορίες Λ (Λιποβαρείς), Ν (Νορμοβαρείς), Υ (Υπέρβαροι) και Π (Παχύσαρκοι). Ως κατηγορία αναφοράς για την εξαρτημένη μεταβλητή επιλέχθηκε η κατηγορία Ν, αφού εκτός του ότι αποτελούν την πλειοψηφία του δείγματος, είναι και η πιο υγιεινή και επιθυμητή κατηγορία ΔΜΣ.

Το πλήρες μοντέλο αποτελείται από τρία ταυτόχρονα μοντέλα, με εξαρτημένη μεταβλητή την $\ln(p_j / p_N)$, $j = \Lambda, Y, \Pi$, δηλ. $\ln(p_\Lambda / p_N)$, $\ln(p_Y / p_N)$ και $\ln(p_\Pi / p_N)$. Ως p_j ορίζουμε την πιθανότητα ένας τυχαία επιλεγμένος έφηβος να ανήκει στην κατηγορία $j = \Lambda, N, Y, \Pi$. Επίσης, καθένα από τα τρία πλήρη μοντέλα, έχει έξι ανεξάρτητες ποσοτικές μεταβλητές (*ΔΜΣ πατέρα, ΔΜΣ μητέρας, Ώρες ύπνου/ημέρα, Ημερήσια κατανάλωση νερού (σε ποτήρια), Ημερήσιος αριθμός γευμάτων και Συχνότητα κατανάλωσης πρωϊνού/εβδομάδα*), καθώς και οκτώ ψευδομεταβλητές

$(x_1 - x_8)$ που αντιστοιχούν στα εννέα γεωγραφικά διαμερίσματα της Ελλάδας. Οι ψευδομεταβλητές $x_1 - x_8$ ορίζονται ως:

$$x_1 = \begin{cases} 1, & \text{κάτοικος Θράκης} \\ 0, & \text{διαφορετικά} \end{cases}, \dots, x_8 = \begin{cases} 1, & \text{κάτοικος Αιγίου Πελάγους} \\ 0, & \text{διαφορετικά} \end{cases}$$

ενώ αν $x_1 = x_2 = \dots = x_8 = 0$, τότε ο/η έφηβος/η είναι κάτοικος νησιών Ιονίου Πελάγους. Παρακάτω δίδεται η μορφή του πλήρους (*full*) μοντέλου, με ανεξάρτητη μεταβλητή την $\ln(p_j / p_N)$.

$$\ln(p_j / p_N) = \alpha_j + \beta_{1j} \Delta \text{ΜΣ πατέρα} + \beta_{2j} \Delta \text{ΜΣ μητέρας} + \beta_{3j} \Omega \text{ρες ύπνου} + \beta_{4j} \text{Ποτήρια νερό} + \beta_{5j} \text{Αριθμός γευμάτων} + \beta_{6j} \text{Συχνότητα κατανάλωσης πρωινού} + \beta_{7j} X_1 + \beta_{8j} X_2 + \beta_{9j} X_3 + \beta_{10j} X_4 + \beta_{11j} X_5 + \beta_{12j} X_6 + \beta_{13j} X_7 + \beta_{14j} X_8$$

Στόχος είναι η εκτίμηση ενός μοντέλου πολυωνυμικής λογιστικής παλινδρόμησης προκειμένου να διερευνηθεί η χωρική επίδραση των διαφορετικών γεωγραφικών διαμερισμάτων αλλά και η επίδραση κάθε διατροφικής συνήθειας στο ΔΜΣ των εφήβων της ελληνικής επικράτειας. Το κατάλληλο υπόδειγμα θα χρησιμοποιηθεί για την εύρεση των διατροφικών συνηθειών των εφήβων οι οποίες είναι ευεργετικές για το ΔΜΣ τους (και άρα και για την υγεία τους).

Χρησιμοποιώντας τη συνάρτηση `multinom()` της R, εκτιμήθηκαν (με τη μέθοδο της μέγιστης πιθανοφάνειας) οι συντελεστές $\alpha_j, \beta_{mj}, m = 1, 2, \dots, 14$ του μοντέλου και τα αποτελέσματα (εκτιμήσεις και τυπικά σφάλματα) δίνονται στον Πίνακα 4. Από τους μεμονωμένους ελέγχους στατιστικής σημαντικότητας για κάθε μια από τις μεταβλητές στο πλήρες μοντέλο (με χρήση του Wald's test, Wasserman, 2006) προέκυψε ότι η επίδραση των γεωγραφικών διαμερισμάτων δεν είναι στατιστικά σημαντική ($p\text{-value} > 0.05$) και στα τρία πλήρη μοντέλα. Επίσης, στατιστικά μη-σημαντικές μεταβλητές φαίνεται να είναι ο ημερήσιος αριθμός γευμάτων (στο μοντέλο Λ/N) και ο ημερήσιος αριθμός ωρών ύπνου (στο μοντέλο Y/N).

Πίνακας 4. Συντελεστές (Τυπικά Σφάλματα) και $p\text{-values}$ του πλήρους μοντέλου

Σταθερές	Λ/N	$p\text{-value}$	Y/N	$p\text{-value}$	Π/N	$p\text{-value}$
Intercepts	-1.32 (0.27)	0.000	-1.74 (0.26)	0.000	-3.76 (0.47)	0.000
ΔΜΣ πατέρ.	-0.01 (0.00)	0.000	0.01 (0.00)	0.000	0.02 (0.00)	0.000
ΔΜΣ μητέρ.	-0.02 (0.00)	0.000	0.02 (0.00)	0.000	0.04 (0.00)	0.000
Ωρες ύπνου	0.03 (0.01)	0.000	-0.01 (0.01)	0.347	-0.02 (0.01)	0.000
Ποτήρ. νερό	-0.02 (0.00)	0.000	0.03 (0.00)	0.000	0.07 (0.01)	0.000
Αρ. γευμάτ.	0.02 (0.01)	0.071	-0.12 (0.01)	0.000	-0.09 (0.02)	0.000
Κατ. πρωινό	0.07 (0.01)	0.000	-0.04 (0.01)	0.000	-0.11 (0.02)	0.000
x_1	-0.28 (0.25)	0.265	-0.41 (0.24)	0.088	-0.18 (0.44)	0.690
x_2	-0.16 (0.25)	0.532	-0.33 (0.23)	0.156	-0.05 (0.43)	0.916
x_3	-0.29 (0.25)	0.257	-0.43 (0.24)	0.068	-0.15 (0.44)	0.729
x_4	-0.01 (0.25)	0.982	-0.37 (0.23)	0.112	-0.74 (0.43)	0.864
x_5	-0.09 (0.25)	0.706	-0.44 (0.23)	0.060	0.19 (0.43)	0.659
x_6	0.03 (0.26)	0.895	-0.28 (0.24)	0.241	-0.09 (0.44)	0.844
x_7	0.19 (0.26)	0.453	-0.41 (0.24)	0.093	-0.04 (0.02)	0.958
x_8	-0.12 (0.26)	0.633	-0.04 (0.27)	0.256	0.34 (0.44)	0.441

Στη συνέχεια, αφαιρώντας τις μεταβλητές $X_1 - X_8$ (που αφορούν τη χωρική επίδραση), προχωρήσαμε στην εκτίμηση του περιορισμένου (reduced) μοντέλου, της μορφής:

$$\ln(p_j / p_N) = \alpha_j + \beta_{1j} \Delta \text{ΜΣ πατέρα} + \beta_{2j} \Delta \text{ΜΣ μητέρας} + \beta_{3j} \Omega \text{ρες ύπνου} + \beta_{4j} \text{Ποτήρια νερό} + \beta_{5j} \text{Αριθμός γευμάτων} + \beta_{6j} \text{Συχνότητα κατανάλωσης πρωινού}$$

Στον Πίνακα 5 παρουσιάζονται οι εκτιμήσεις των συντελεστών α_j , β_{mj} , $m = 1, 2, \dots, 6$ του περιορισμένου μοντέλου μοντέλων μαζί με τα αντίστοιχα τυπικά σφάλματα, καθώς και τις τιμές των p-values για τους μεμονωμένους στατιστικούς ελέγχους σημαντικότητας. Τα αποτελέσματα δείχνουν ότι, για το μοντέλο Λ/N ο αριθμός γευμάτων/ημέρα και οι ώρες ύπνου/ημέρα (για το μοντέλο Y/N) δεν είναι στατιστικά σημαντικές (p-value > 0.05).

Πίνακας 5. Συντελεστές (Τυπικά Σφάλματα) και p-values του περιορισμένου μοντέλου

Σταθερές	Λ/N	p-value	Y/N	p-value	Π/N	p-value
Intercepts	-1.43 (0.27)	0.000	-2.12 (0.26)	0.000	-3.77 (0.47)	0.000
$\Delta \text{ΜΣ πατέρ.}$	-0.01 (0.00)	0.000	0.02 (0.00)	0.000	0.02 (0.00)	0.000
$\Delta \text{ΜΣ μητέρ.}$	-0.02 (0.00)	0.000	0.02 (0.00)	0.000	0.04 (0.00)	0.000
$\Omega \text{ρες ύπνου}$	0.03 (0.01)	0.000	-0.01 (0.01)	0.357	-0.03 (0.01)	0.002
Ποτήρ. νερό	-0.02 (0.00)	0.000	0.03 (0.00)	0.000	0.07 (0.01)	0.000
Αρ. γευμάτ.	0.02 (0.01)	0.090	-0.11 (0.01)	0.000	-0.09 (0.02)	0.000
Κατ. πρωινό	0.07 (0.01)	0.000	-0.04 (0.01)	0.000	-0.11 (0.02)	0.000

Αξίζει επίσης να αναφέρουμε ότι η σύγκριση μεταξύ του πλήρους και του περιορισμένου μοντέλου, με χρήση του κριτηρίου BIC, έδωσε τιμή 81711.63 (για το περιορισμένο μοντέλο) και τιμή 81860.51 (για το πλήρες μοντέλο). Άρα, μεταξύ των μοντέλων, με βάση το κριτήριο BIC επιλέγεται το περιορισμένο μοντέλο, χωρίς την επίδραση των γεωγραφικών διαμερισμάτων.

Από την προσαρμογή των μοντέλων στα διαθέσιμα δεδομένα προκύπτει (με βάση το περιορισμένο πρότυπο) πως, όταν ο $\Delta \text{ΜΣ}$ των γονέων αυξάνεται κατά μία μονάδα, μειώνεται η πιθανότητα εμφάνισης λιποβαρών έναντι νορμοβαρών εφήβων και αυξάνεται η πιθανότητα εμφάνισης υπέρβαρων και παχύσαρκων, παρά νορμοβαρών εφήβων. Αυτή η αρνητική συμβολή του αυξημένου $\Delta \text{ΜΣ}$ των γονέων σε αποκλίνουσες καταστάσεις του $\Delta \text{ΜΣ}$ των παιδιών τους έχει διαπιστωθεί και από άλλες αντίστοιχες μελέτες (Farajian et al. 2012, Birbilis et al. 2013). Αποδίδεται στην κληρονομικότητα και τις αρνητικές διατροφικές συνήθειες των γονέων, οι οποίες οδηγούν τους νέους στο να μεγαλώνουν σε περιβάλλον που ενθαρρύνει την υπερκατανάλωση τροφών και μάλιστα πλούσιων σε θερμίδες (πλούσιων σε λιπίδια και απλά σάκχαρα).

Επίσης, όταν αυξάνεται κατά μια ώρα ο ημερήσιος χρόνος ύπνου, τότε αυξάνεται η πιθανότητα οι έφηβοι να είναι λιποβαρείς απ' ότι νορμοβαρείς και μειώνεται η πιθανότητα να είναι παχύσαρκοι απ' ότι νορμοβαρείς. Αντίστοιχες μελέτες έδειξαν ότι οι περισσότερες ώρες ύπνου/24ωρο μειώνουν τον επιπολασμό της εφηβικής παχυσαρκίας (Nielsen et al. 2010, Seegers et al. 2011).

Επιπλέον, οι έφηβοι που καταναλώνουν ένα ποτήρι νερό περισσότερο την ημέρα φαίνεται πως, μειώνουν την πιθανότητα να είναι λιποβαρείς, παρά νορμοβαρείς και αυξάνουν την πιθανότητα να είναι υπέρβαροι και παχύσαρκοι, παρά νορμοβαρείς. Η εν λόγω παρατήρηση δεν επιβεβαιώνει την άποψη των Muckelbauer et al. (2014), πως η μεγάλη κατανάλωση νερού μπορεί να βελτιώσει την κατάσταση του σωματικού βάρους των εφήβων. Λόγω του ότι η επίδραση της κατανάλωσης νερού στην πρόληψη του υπερβάλλοντος βάρους των εφήβων δεν έχει διερευνηθεί επαρκώς, το παραπάνω εύρημα χρήζει περαιτέρω μελέτης.

Διαπιστώθηκε επίσης πως, η ημερήσια αύξηση των γευμάτων κατά ένα, μειώνει την πιθανότητα εμφάνισης υπέρβαρων και παχύσαρκων έναντι νορμοβαρών εφήβων. Σε παρόμοια συμπεράσματα κατέληξαν και αντίστοιχες μελέτες, που διαπίστωσαν ότι η αύξηση της ημερήσιας κατανάλωσης γευμάτων σχετίζεται με τον χαμηλό ΔΜΣ των εφήβων, γεγονός που αποτελεί και βασική σύσταση της Μεσογειακής Διατροφής (Lagiou & Parava, 2008, Kontogianni et al. 2010).

Τέλος, διαπιστώθηκε πως, η αύξηση κατά μία ημέρα την εβδομάδα της κατανάλωσης πρωινού, αυξάνει την εμφάνιση λιποβαρών έναντι νορμοβαρών εφήβων και μειώνει την εμφάνιση υπέρβαρων και παχύσαρκων έναντι των νορμοβαρών. Το πρωινό σύμφωνα με τις συστάσεις της Μεσογειακής διατροφής αποτελεί ένα από τα βασικότερα γεύματα της ημέρας, γιατί επηρεάζει θετικά τη διαμόρφωση νορμοβαρών εφήβων (Thibault et al. 2012)

Όσον αφορά τις εκτιμώμενες πιθανότητες (δείτε π.χ. Agresti, (2001), σελ. 102-104) για τις τέσσερις κατηγορίες του ΔΜΣ των εφήβων προκύπτει ότι, ανεξαρτήτως γεωγραφικού διαμερίσματος η εκτιμώμενη πιθανότητα ένας έφηβος να είναι λιποβαρής, νορμοβαρής, υπέρβαρος και παχύσαρκος με τις εν λόγω διατροφικές συνήθειες είναι 0.158, 0.753, 0.074 και 0.015, αντίστοιχα.

3.4 Συσταδοποίηση Εφήβων ως προς τις Διατροφικές Συνήθειες

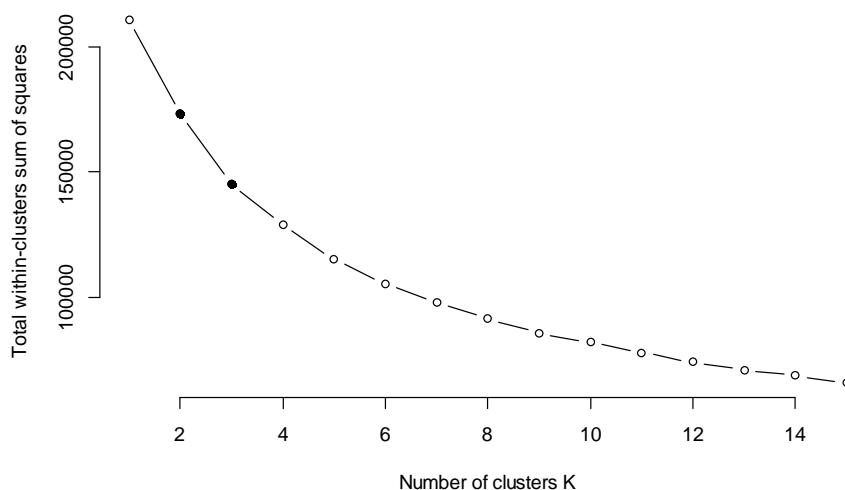
Στη συνέχεια, επιχειρείται η δημιουργία ενός γενικότερου προφίλ των διατροφικών συνηθειών των Ελλήνων εφήβων χρησιμοποιώντας την Ανάλυση κατά Συστάδες (Cluster Analysis). Η Ανάλυση κατά Συστάδες αποτελεί μια πολυμεταβλητή μέθοδο ανάλυσης, που στοχεύει στην ομαδοποίηση των δεδομένων, με βάση κάποιο μέτρο ομοιότητας (δείτε π.χ. Καρλής, 2005).

Με τη χρήση της μεθόδου Elbow δημιουργείται ένα γράφημα, το οποίο εμφανίζει μια καμπύλη (Ketchen & Shook, 1996). Το σημείο στο οποίο η καμπύλη δημιουργεί μία «γωνία» (προσθήκη έντονων κουκίδων στα επιλεγμένα σημεία), μαρτυρά τον κατάλληλο αριθμό ομάδων, που καλείται να διαλέξει ο ερευνητής, έτσι ώστε να συνεχίσει την cluster analysis εφαρμόζοντας την μέθοδο k -means. Από την εικόνα στο Σχήμα 2 φαίνεται ότι, ο ιδανικός αριθμός clusters για το εν λόγω δείγμα είναι $k = 2$ ή $k = 3$.

Για το λόγο αυτό, αποφασίστηκε η εφαρμογή Cluster Analysis για $k = 2$ και $k = 3$ συστάδες ως εξής: Αρχικά, δημιουργήθηκαν δύο clusters των διατροφικών συνηθειών των Ελλήνων εφήβων και στη συνέχεια τρία clusters, με βάση την ένδειξη

της μεθόδου Elbow. Λόγω εφαρμογής της Cluster Analysis σε πολύ μεγάλο όγκο δεδομένων, δε δίνεται γραφική αναπαράσταση των μελών σε κάθε cluster αφού αυτή δε συμβάλει περαιτέρω στην ερμηνεία τους. Στη συνέχεια, διερευνήθηκε αν κάποιο από αυτά τα υπό μελέτη χαρακτηριστικά είναι ιδιαίτερο (π.χ. εμφανίζεται σε μεγάλο ποσοστό) για τα άτομα κάθε cluster. Ενδεικτικά, για τα δύο clusters ($k=2$) και για τα τρία clusters ($k=3$) δίνονται στον Πίνακα 6 τα μεγέθη τους, καθώς και τα ποσοστά των επιπέδων (κατηγοριών) εντός κάθε cluster, για σημαντικές μεταβλητές στη μελέτη, όπως το Γεωγραφικό Διαμέρισμα (Γ.Δ.), ο λόγος επιλογής τροφίμων (Λ.Ε.Τ.), ο ΔΜΣ και η ημερήσια κατανάλωση νερού (Π.Ν.). Η τελευταία μεταβλητή, προέκυψε από την ποσοτική μεταβλητή «Ποτήρια νερό/ημέρα» αφού αυτή κατηγοριοποιήθηκε σε δύο επίπεδα: (i) 1^η κατηγορία: 1-9 ποτήρια νερό/ημέρα (Χαμηλή-Φυσιολογική κατανάλωση νερού) και (ii) 2^η κατηγορία: > 9 ποτήρια νερό/ημέρα (Υπερβάλλουσα κατανάλωση νερού). Επίσης, για τα άτομα κάθε cluster, είτε για $k=2$ είτε για $k=3$, δίνεται η δειγματική μέση τιμή του ΔΜΣ σε κάθε cluster, η διάμεσος της ημερήσιας κατανάλωσης ποτηριών νερού, του ημερήσιου αριθμού ωρών ύπνου (Ω.Υ.), της εβδομαδιαίας συχνότητας παραγγελιών delivery (Del.), του ημερήσιου αριθμού γευμάτων (Α.Γ.) και της εβδομαδιαίας συχνότητας κατανάλωσης πρωϊνού (Κ.Π.).

Σχήμα 2. Καμπύλη της μεθόδου Elbow



Από τον Πίνακα 6 διαπιστώνουμε ότι δε διαφαίνεται να είναι υπεύθυνη για τη δημιουργία των clusters. κάποια από τις υπό εξέταση μεταβλητές, εκτός ίσως από την ημερήσια κατανάλωση νερού. Συγκεκριμένα, είτε για $k=2$, είτε για $k=3$, τα ποσοστά των διαφορετικών επιπέδων των υπό εξέταση μεταβλητών είναι (περίπου) τα ίδια, π.χ. για τις διαφορετικές κατηγορίες του ΔΜΣ ή για το Λ.Ε.Τ. Επίσης, η δειγματική μέση τιμή του ΔΜΣ δε φαίνεται να διαφοροποιείται είτε μεταξύ των δύο clusters είτε μεταξύ των τριών clusters.

Σε ότι αφορά την ημερήσια κατανάλωση νερού, η διάμεσος σε κάθε cluster, για $k=2$ είναι 6 και 13 ενώ για $k=3$ είναι 6, 6 και 13. Επίσης, για $k=2$, διαπιστώθηκε

ότι κοινό χαρακτηριστικό όλων των εφήβων που ανήκουν στο δεύτερο cluster (25.7% του συνολικού δείγματος), είναι ότι καταναλώνουν περισσότερα από εννέα ποτήρια νερό/ημέρα, ενώ όλοι οι έφηβοι που ανήκουν στο πρώτο cluster (74.3% του συνολικού δείγματος) καταναλώνουν ελάχιστο έως και επαρκές νερό/ημέρα, δηλαδή από 1 έως 9 ποτήρια. Η μεγάλη ανομοιογένεια που παρατηρείται μεταξύ των δύο clusters ως προς την ημερήσια κατανάλωση νερού αποτελεί ένδειξη του ότι η συσταδοποίηση των διατροφικών συνθηκών, για $k=2$, ενδέχεται να βασίστηκε στον ημερήσιο αριθμό ποτηριών νερού που καταναλώνουν οι έφηβοι στην Ελλάδα. Σε κάθε περίπτωση, θα πρέπει να διερευνηθεί περαιτέρω το κατά πόσο η ημερήσια κατανάλωση νερού επιδρά σημαντικά στις διατροφικές συνήθειες των εφήβων όπως, στην εβδομαδιαία συχνότητα κατανάλωσης φαγητού μέσω των παραγγελιών delivery, στη εβδομαδιαία συχνότητα κατανάλωσης πρωινού, στις ώρες ύπνου/ημέρα, καθώς και στον ημερήσιο αριθμό γευμάτων.

Πίνακας 6. Αποτελέσματα Cluster Analysis

Μέγεθος	k=2		k=3		
	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 3
Γ.Α.					
Αιγαίο Πέλαγος	4.4	4.8	3.6	4.6	4.8
Ήπειρος	6.6	6.0	5.4	6.9	6.0
Θεσσαλία	19.7	22.0	20.3	19.5	21.9
Θράκη	6.2	4.8	6.4	6.1	4.8
Ιόνιο Πέλαγος	0.3	0.4	0.3	0.3	0.4
Κρήτη	3.7	4.2	3.5	3.7	4.2
Μακεδονία	40.3	37.1	40.7	40.2	37.1
Πελοπόννησος	4.5	4.6	3.8	4.8	4.6
Στερεά Ελλάδα	14.4	16.2	16.0	13.8	16.3
Α.Ε.Τ.					
Γεύση	68.3	61.8	66.4	68.8	61.9
Εμφάνιση	6.3	6.5	6.5	6.2	6.5
Θερμίδες	8.7	10.4	8.3	8.8	10.4
Θρεπτική Αξία	15.5	20.0	17.6	14.8	19.9
Κόστος	1.2	1.3	1.1	1.3	1.2
ΔΜΣ					
Νορμοβαρείς	66.1	61.6	64.5	66.6	61.6
Παχύσαρκοι	4.0	7.9	3.8	4.1	7.9
Υπέρβαροι	15.1	18.2	14.7	15.2	18.1
Λιποβαρείς	14.8	12.4	17.0	14.1	12.4
Π.Ν.					
1-9 ποτήρια	100.0	0.0	99.05	100	0
>9 ποτήρια	0.0	100.0	0.95	0	100
Μ.Ο. ΔΜΣ	22.09	22.95	21.90	22.15	22.95
Διάμεσος Π.Ν.	6	13	6	6	13
Διάμεσος Ω. Υ.	8	9	15	8	9
Διάμεσος Del.	1	1	1	1	1
Διάμεσος Α. Γ.	4	4	4	4	4

Διάμεσος Κ. Π.	4	5	4	4	5
----------------	---	---	---	---	---

Αντίθετα, για $k=3$, δε φαίνεται ξεκάθαρα αν υπάρχει κάποιο ιδιαίτερο χαρακτηριστικό σε καθένα από τα τρία clusters. Αυτό που φαίνεται, από τα αποτελέσματα της cluster analysis είναι πως υπάρχει ένα cluster μεγέθους περίπου 10800 εφήβων (cluster 2 για $k=2$ και cluster 3 για $k=3$), το οποίο διατηρεί τα χαρακτηριστικά του και για τις δύο τιμές του k .

4. ΣΥΖΗΤΗΣΗ - ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα εργασία παρουσιάζονται τα αποτελέσματα στατιστικής ανάλυσης από δεδομένα ερωτηματολογίων, σχετικά με τις διατροφικές συνήθειες των Ελλήνων εφήβων. Η ανάλυση κατέληξε σε συμπεράσματα, τα οποία επιβεβαιώνονται και από ανάλογες μελέτες που πραγματοποιήθηκαν στη χώρα μας και διεθνώς.

Αρχικά, από τους μη παραμετρικούς ελέγχους (*Ενότητα 3.2*) και την πολυωνμική λογιστική παλινδρόμηση (*Ενότητα 3.3*), φαίνεται να ασκούν θετική επιρροή στο ΔΜΣ και κατ' επέκταση την υγεία των εφήβων, οι νορμοβαρείς γονείς, ο επαρκής ύπνος (10 έως 11 ώρες/24ωρο, Iglowstein et al. 2003) και οι διατροφικές συνήθειες που υιοθετούνται από τη Μεσογειακή διατροφή (καθημερινή κατανάλωση πρωινού, συχνότερα ημερήσια γεύματα και κατανάλωση σπιτικού φαγητού). Επιπλέον, από την ανάλυση σε συστάδες (*Ενότητα 3.4*), το εύρημα του σημαντικού ρόλου που φαίνεται να διαδραματίζει η ημερήσια κατανάλωση νερού στη διαμόρφωση του προφίλ των εφήβων ως προς τις βασικές διατροφικές συνήθειες, αποτελεί νέο δεδομένο για την επιστήμη της διατροφής, που πρέπει να τύχει περαιτέρω διερεύνησης.

Τέλος, σήμα κινδύνου για τις υπηρεσίες που διαμορφώνουν τη διατροφική πολιτική της χώρας, για επιστήμονες και ερευνητές αποτελεί η ένδειξη μείωσης του ποσοστού των νορμοβαρών εφήβων στη χώρα μας σε 65% από 75% (Patsopoulou et al. 2016), με άνω όριο εμπιστοσύνης 95% το 65.3%. Αυτό οφείλεται στα υψηλά ποσοστά υπέρβαρων και παχύσαρκων εφήβων (21% συνολικά), καθώς και στη δραματική αύξηση των λιποβαρών εφήβων, από 5% (Birbilis et al. 2013) στο 14%. Η νέα αρνητική εικόνα που αφορά το προφίλ των εφήβων με χαμηλό ΔΜΣ ενδεχομένως να μπορεί να αποδοθεί και στην επιδεινωθείσα διατροφική κατάσταση των Ελλήνων, λόγω της πρόσφατης οικονομικής κρίσης. Τέλος, ως μελλοντική έρευνα, έχει ήδη ξεκινήσει η μελέτη της επίδρασης των διατροφικών συνηθειών των Ελλήνων εφήβων σε πιο σύγχρονους δείκτες όπως ο Tri-Ponderal Mass Index (Peterson et al. 2017).

ABSTRACT

This study investigates the profile of nutritional habits of adolescents in Greece. The aim is to determine the effect of these habits, along with other nutritional and geographic factors, on the Body Mass Index (BMI) and hence, on the health of adolescents in Greece. The method of questionnaire was used to collect the data. The questionnaire was designed by the Department of Nutrition and Dietetics of Alexander Technological Educational Institute of Thessaloniki and approved by the Pedagogical Institute. 42,129 high school students, from all the

geographic districts of Greece, participated in the survey. The results of the statistical analysis revealed that parent's normal BMI, sleep adequacy and breakfast consumption have a positive effect on adolescents' health. Further investigation is needed in order to determine whether or not the daily water consumption has a significant influence on the formation of their eating habits.

ΑΝΑΦΟΡΕΣ

- Καρλής, Δ. (2005). *Πολυμεταβλητή Στατιστική Ανάλυση*. Εκδόσεις Σταμούλη.
- Κουκ, Τ. (2002). *Εκτίμηση διατροφικής κατάστασης εφήβων βάσει της Μεσογειακής Πυραμίδας*. Χαροκόπειο Πανεπιστήμιο.
- Κυρανά Ζ., Γεροθανάση Κ., Ζήμερας Σ., Κυρανάς Ε. (2018). Εφαρμογή της Πολυωνυμικής Λογιστικής Παλινδρόμησης και τεχνικών ομαδοποίησης στις διατροφικές συνήθειες των Ελλήνων εφήβων. *Πρακτικά 31^{ου} Πανελληνίου Συνέδριου Στατιστικής*, Λαμία, σελ. 105-119.
- Κυρανά, Ζ. (2019). *Στατιστική Ανάλυση Δεδομένων των Διατροφικών Συνηθειών των Εφήβων στην Ελλάδα με χρήση του Στατιστικού Πακέτου R*. Μεταπτυχιακή Διατριβή. Τμήμα Στατιστικής Αναλογιστικών-Χρηματοοικονομικών Μαθηματικών. Πανεπιστήμιο Αιγαίου.
- Λουμάκου Μ., Σαραφίδου Ε., Κορδούτης Π. και Μπαρμποπούλου Δ. (2006). Σχέσεις δασκάλων και διατροφική αγωγή στο σχολείο. *Μέντορας*. **9**. 43-59.
- Ξεκαλάκη, Ε. (2001). *Μη Παραμετρική Στατιστική*. Αθήνα: Εκδόσεις Ε. Μπένου.
- Agresti, A. (2001). *Categorical Data Analysis*. 2nd edition. New Jersey: John Wiley & Sons.
- Ardelt M. and Day L. (2002). Parents, Siblings, and Peers: Close Social Relationships and Adolescent Deviance. *Journal of Early Adolescence*, **22**. 310-349.
- Birbilis M., Moschonis G., Mougios Y., Manios Y. and Healthy Growth Study group. (2013). Obesity in adolescence is associated with perinatal risk factors, parental BMI and sociodemographic characteristics. *European Journal of Clinical Nutrition*. **67**.115-121.
- Cole TJ., Bellizzi MC., Flegal KM. and Dietz WH. (2000). Establishing a standard definition for child overweight and obesity worldwide: international survey. *British Medical Journal*. **320**. 1240-1243.
- Dunn OJ. (1964). Multiple comparisons using rank sums. *Technometrics*. **6**. 241-252.
- Farajian P., Panagiotakos DB., Risvas G., Karasouli K., Bountziouka V., Voutzourakis N. and Zampelas A. (2012). Socio-economic and demographic determinants of childhood obesity prevalence in Greece: the GRECO (Greek Childhood Obesity) study. *Public Health Nutrition*. **16**. 240-247.
- Iglowstein I., Jenni OG., Molinari L. and Largo RH. (2003). Sleep duration from infancy to adolescence: reference values and generational trends. *Pediatrics*, **111**(2). 302-307.
- Ketchen DJ. and Shook CL. (1996). The application of cluster analysis in Strategic Management Research: An analysis and critique. *Strategic Management Journal*. **17**(6). 441-458.

- Kontogianni MD., Farmaki AE., Virda N., Sofrona S., Magkanari F. and Yannakoulia M. (2010). Associations between lifestyle patterns and body mass index in a sample of Greek children and adolescents. *Journal of the American Dietetic Association*. **110**. 215-221.
- Lagiou A. and Parava M. (2008). Correlates of childhood obesity in Athens. Greece. *Public Health Nutrition*. **11**. 940-945.
- Muckelbauer R., Barbosa CL., Mittag T., Burkhardt K., Mikelaishvili N. and Müller-Nordhorn J. (2014). Association between water consumption and body weight outcomes in children and adolescents: A systematic review. *Obesity*. **22**(12). 2462-2475.
- Nielsen LS., Danielsen KV. and Sørensen TIA. (2010). Short sleep duration as a possible cause of obesity: critical analysis of the epidemiological evidence. *Obesity Reviews*. **12**. 78-92.
- Papandreou D., Malindretos P. and Roussol I. (2008). Risk factors for childhood obesity in a Greek pediatric population. *Public Health Nutrition*. **13**. 1535-1539.
- Patsopoulou A., Rachiotis G., Katsioulis A., Molyvdas PA. and Hadjichristodoulou C. (2016). Overweight and obese adolescents in Greece. A study protocol of the feeding-exercise trial in adolescents (FETA). *Proceedings, 7th International Congress of Internal Medicine of Central Greece*. Larissa: 19-21 March. 2015. Greece: pp. 27.
- Peterson CM., Su H., Thomas DM., Heo M., Golnabi AH., Pietrobelli A. and Heymsfield SB. (2017). Tri-Ponderal Mass Index Vs Body Mass Index in Estimating Body Fat During Adolescence. *JAMA Pediatrics*. **171**(7). 629-636.
- Roditis ML., Parlapani ES., Tzotzas T., Hassapidou M. and Krassas GE. (2009). Epidemiology and predisposing factors of obesity in Greece: from the Second World War until today. *Journal of Pediatric Endocrinology and Metabolism*. **22**. 389-405.
- Seegers V., Petit D., Falissard B., Vitaro F., Tremblay RE. and Montplaisir J. (2011). Short sleep duration and body mass index: a prospective longitudinal study in preadolescence. *American Journal of Epidemiology*. **173**. 621-629.
- Thibault H., Carrière C., Langevin C., Déti EK., Barberger-Gateau P. and Maurice S. (2012). Prevalence and factors associated with overweight and obesity in French primary-school children. *Public Health Nutrition*, **16**(2). 193-201.
- Tzotzas T. and Krassas GE. (2004). Prevalence and trends of obesity in children and adults of South Europe. *Pediatric Endocrinology Reviews*. **3**. 448-454.
- Tzotzas T., Kapantais E., Tziomalos K., Ioannidis I., Mortoglou A., Bakatselos S., Kaklamanou M., Lanaras L. and Kaklamanou D. (2011). Prevalence of overweight and abdominal obesity in Greek children 6-12 years old: Results from the National Epidemiological Survey. *Hippokratia*. **15**. 48-53.
- Wasserman, L. (2006). *All of Nonparametric Statistics*. New York: Springer.
- WHO, ENHIS. (2009). *Prevalence of overweight and obesity in children and adolescents*. Fact sheet 2.3.

ΕΓΧΡΩΜΟ ΤΟΒΙΤ ΚΑΛΜΑΝ ΦΙΛΤΡΟ

Κ. Λουμπόνιας¹, Γ. Τσακλίδης¹

¹Τμήμα Μαθηματικών, Α.Π.Θ.
{kostikasl, tsaklidis}@math.auth.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία γίνεται εκτίμηση κρυφών καταστάσεων σε μοντέλα χώρου καταστάσεων σε συνεχή χώρο και διακριτό χρόνο, όταν οι παρατηρούμενες μετρήσεις (παρατηρήσεις) εμπεριέχουν λευκό θόρυβο με άγνωστη, αλλά σταθερή διασπορά, και είναι λογοκριμένες. Λόγω της λογοκρισίας, το μοντέλο των κρυφών καταστάσεων είναι μη γραμμικό, συνεπώς το απλό φίλτρο Kalman, δεν μπορεί να χρησιμοποιηθεί. Για την πραγματοποίηση της ζητούμενης εκτίμησης, υπολογίζονται οι ροπές της λογοκριμένης κανονικής κατανομής με τη χρήση της ροπογεννήτριας συνάρτησης. Θεωρείται ότι οι κρυφές καταστάσεις εμπεριέχουν θόρυβο με μη σταθερή διασπορά (έγχρωμο θόρυβο) και για την αντιμετώπιση της ετεροσκεδαστικότητας χρησιμοποιείται η θεωρία από τα Έγχρωμα Φίλτρα Kalman. Τα αποτελέσματα μέσω προσομοίωσης έδειξαν ότι η προτεινόμενη μέθοδος παρέχει αμερόληπτες εκτιμήσεις για τις κρυφές καταστάσεις και σε σχέση με άλλα φίλτρα έχει το μικρότερο μέσο τετραγωνικό σφάλμα.

Λέξεις κλειδιά: φίλτρο Kalman, Λογοκριμένα δεδομένα, Έγχρωμος θόρυβος.

1. ΕΙΣΑΓΩΓΗ

Ο αλγόριθμος του φίλτρου Kalman (ΚΦ) (Kalman 1960) χρησιμοποιεί μια σειρά από παρατηρούμενες μετρήσεις (οι οποίες εμπεριέχουν θόρυβο) στην πορεία του χρόνου, προκειμένου να εκτιμήσει κρυφές καταστάσεις (σε συνεχή χώρο καταστάσεων). Ο αλγόριθμος ΚΦ παρέχει βέλτιστες εκτιμήσεις για τις κρυφές καταστάσεις (αμερόληπτες με ελάχιστη διασπορά σφαλμάτων), όταν α) το μοντέλο χώρου καταστάσεων είναι γραμμικό και β) όταν ο θόρυβος των μετρήσεων και καταστάσεων είναι λευκός. Σε πολλά προβλήματα καμία από τις δύο παραπάνω συνθήκες δεν ισχύει, και αυτό έχει ως αποτέλεσμα το ΚΦ να παρέχει μεροληπτικές εκτιμήσεις.

Μία μορφή μη γραμμικότητας στο μοντέλο χώρου καταστάσεων προκύπτει λόγω της λογοκρισίας στις μετρήσεις (Buckley et al. 1979). Οι λογοκριμένες μετρήσεις, που μελετώνται στην παρούσα εργασία, περιγράφονται από το μοντέλο Tobit, τύπου I (Tobin 1958). Το μοντέλο αυτό εξετάζει την περίπτωση μετρήσεων, με τιμές σε πεπερασμένο διάστημα, που καθορίζεται από (πεπερασμένο) κάτω και άνω όριο (όρια λογοκρισίας). Στην εργασία Allik et al. (2014) παρουσιάζεται ο αλγόριθμος

Tobit Kalman Φίλτρο (TKΦ), ο οποίος χειρίζεται λογοκριμένες μετρήσεις (Tobit τύπου I) για την εκτίμηση κρυφών καταστάσεων. Από τα πειράματα σε (τεχνητά κατασκευασμένες) λογοκριμένες μετρήσεις, προκύπτει ότι το TKΦ παρέχει εκτιμήσεις για τις κρυφές καταστάσεις με μικρότερο τετραγωνικό σφάλμα έναντι του απλού ΚΦ (Allik et al. 2014). Το αποτέλεσμα αυτό είναι αναμενόμενο, αφού ως γνωστό το ΚΦ δεν μπορεί να χειριστεί λογοκριμένες μετρήσεις. Επιπλέον στις εργασίες Loumponias et al. (2016), Loumponias et al. και (2018) χρησιμοποιήθηκε μια παραλλαγή του TKΦ, με μη σταθερά όρια λογοκρισίας, για την παρακολούθηση της θέσης ανθρώπων σε εσωτερικό και εξωτερικό χώρο.

Οι παραπάνω μέθοδοι παρέχουν αμερόληπτους εκτιμητές με την προϋπόθεση ότι ο θόρυβος των καταστάσεων είναι λευκός με σταθερή διασπορά. Ωστόσο σε πολλά προβλήματα μια τέτοια συνθήκη δεν ισχύει και ο θόρυβος των καταστάσεων μπορεί να περιγραφεί από ένα αυτοπαλινδρομούμενο μοντέλο $AR(p)$ τάξης p , (Chui et al. 2017). Ο θόρυβος που περιγράφεται από ένα μοντέλο $AR(p)$ καλείται συχνά, στην βιβλιογραφία, έγχρωμος θόρυβος (Bryson et al. 1965, Koo et al. 1989). Ένας τρόπος αντιμετώπισης του έγχρωμου θορύβου στη διαδικασία του ΚΦ είναι η περιγραφή του μοντέλου χώρου καταστάσεων ως σύστημα με λευκό θόρυβο (Zimmerman 1969, Anderson 2012, Chui et al. 2017) .

Ο σκοπός της παρούσας εργασίας είναι η δημιουργία ενός φίλτρου τύπου Kalman, το οποίο α) θα μπορεί να διαχειρίζεται λογοκριμένες μετρήσεις (μονοδιάστατες) και β) ο θόρυβος των καταστάσεων (πολυδιάστατος) θα περιγράφεται από ένα μοντέλο $AR(1)$. Σε αρκετές εργασίες που χρησιμοποιείται το TKΦ, (Allik et al. 2014, Allik et al. 2015, Loumponias et al. 2016, Han et al. 2018), δεν υπολογίζεται με ακρίβεια η διασπορά των λογοκριμένων μετρήσεων και κατ' επέκταση, ούτε η ακριβής πιθανότητα μιας μέτρησης να ανήκει ή όχι στη λογοκριμένη περιοχή. Στην εργασία Loumponias et al. (2018) δίνεται ο ακριβής τύπος υπολογισμού της διασποράς των λογοκριμένων μετρήσεων, χωρίς απόδειξη. Στην παρούσα εργασία δίνεται η αναλυτική απόδειξη για τον υπολογισμό της διασποράς των λογοκριμένων μετρήσεων. Η ακρίβεια στον υπολογισμό των λογοκριμένων ροπών έχει ως επακόλουθο και καλύτερες επιδόσεις για το TKΦ. Στη συνέχεια χρησιμοποιείται η μεθοδολογία που περιγράφεται στις Anderson (2012), Chui et al. (2017) για την αντιμετώπιση του έγχρωμου θορύβου. Η μελέτη μέσω προσομοίωσης δείχνει ότι οι επιδόσεις του προτεινόμενου Έγχρωμου Tobit Kalman Φίλτρου (ETKΦ) είναι καλύτερες από εκείνες των TKΦ και ΕΚΦ.

Το υπόλοιπο της εργασίας οργανώνεται ως εξής: Στην Ενότητα 2, γίνεται ο αναλυτικός υπολογισμός της λογοκριμένης αναμενόμενης μέσης τιμής και της διασποράς των καταστάσεων. Στην Ενότητα 3, περιγράφεται η διαδικασία ETKΦ. Στην Ενότητα 4, χρησιμοποιούνται και συγκρίνονται τα ΕΚΦ, TKΦ και ETKΦ σε τεχνητά δεδομένα. Στην Ενότητα 5, παρατίθενται τα συμπεράσματα της έρευνας.

2. ΛΟΓΟΚΡΙΜΕΝΗ ΡΟΠΟΓΕΝΝΗΤΡΙΑ ΣΥΝΑΡΤΗΣΗ

Έστω ένα τυχαίο διάνυσμα (τ.δ) $\mathbf{z}^* \sim \mathbf{N}(\mathbf{m}, \mathbf{S})$ διάστασης $n + 1$, όπου $\mathbf{z}^* = (x_1, x_2, \dots, x_n, y^*)'$. Θεωρούμε ότι η τελευταία συνιστώσα του τ.δ. \mathbf{z}^* λογοκρίνεται στο διάστημα $[a, b]$, και συγκεκριμένα:

$$y = \begin{cases} y^*, & a < y^* < b \\ a, & y^* \leq a \\ b, & y^* \geq b \end{cases}. \quad (1)$$

Η κατανομή του τ.δ. $\mathbf{z} = (x_1, x_2, \dots, x_n, y)'$ δίνεται από την εξίσωση (Loumponias et al. 2016)

$$\begin{aligned} f_{\mathbf{z}}(\mathbf{z}) &= f_{\mathbf{z}^*}(\mathbf{z}) \cdot u_{(a,b)}(y^*) \\ &+ \int_{-\infty}^a f_{\mathbf{z}^*}(\mathbf{z}) dy^* \cdot \delta(y^* - a) \\ &+ \int_b^{+\infty} f_{\mathbf{z}^*}(\mathbf{z}) dy^* \cdot \delta(y^* - b), \end{aligned} \quad (2)$$

όπου δ είναι το δέλτα του *Kronecker* και $u_{[a,b]}(x)$ είναι μια συνάρτηση που είναι ίση με 1 όταν $x \in [a, b]$ και 0, αλλιώς. Από τη σχέση (2) υπολογίζεται η ροπογεννήτρια του τ.δ. \mathbf{z} , που δίδεται στην επόμενη πρόταση.

Πρόταση 1. Όταν $\mathbf{z}^* \sim \mathbf{N}(\mathbf{m}, \mathbf{S})$, τότε η ροπογεννήτρια του τ.δ. \mathbf{z} έχει τη μορφή:

$$\begin{aligned} M_{\mathbf{z}}(\mathbf{t}) &= \exp(\mathbf{t}' \mathbf{m} + \frac{1}{2} \mathbf{t}' \mathbf{S} \mathbf{t}) \cdot \\ &\left(F_{e_{n+1}}(b - m_{n+1} - \sum_{i=1}^n S_{n+1,i} t_i) - F_{e_{n+1}}(a - m_{n+1} - \sum_{i=1}^{n+1} S_{n+1,i} t_i) \right) \\ &+ \exp(t_{n+1} a + \mathbf{t}'_0 \mathbf{m} + \mathbf{t}'_0 \mathbf{S} \mathbf{t}_0 / 2) F_{e_{n+1}}(a - m_{n+1} - \sum_{i=1}^{n+1} S_{n+1,i} t_{0,i}) \\ &+ \exp(t_{n+1} b + \mathbf{t}'_0 \mathbf{m} + \mathbf{t}'_0 \mathbf{S} \mathbf{t}_0 / 2) (1 - F_{e_{n+1}}(b - m_{n+1} - \sum_{i=1}^{n+1} S_{n+1,i} t_{0,i})), \end{aligned}$$

όπου $\mathbf{t}_0 = (t_1, \dots, t_n, 0)'$, $e_{n+1} \sim N(0, S_{n+1,n+1})$ και $F_{e_{n+1}}(x)$ είναι η συνάρτηση κατανομής της τυχαίας μεταβλητής (τ.μ.) e_{n+1} .

Απόδειξη. Η ροπογεννήτρια του τ.δ. \mathbf{z} από την σχέση (2) υπολογίζεται ως εξής:

$$\begin{aligned}
M_{\mathbf{z}}(\mathbf{t}) &= E e^{\mathbf{t}'\mathbf{z}} = \int_{R^n} \int_a^b e^{\mathbf{t}'\mathbf{z}} f_{\mathbf{z}^*}(\mathbf{z}) d\mathbf{z} \\
&\quad + \int_{R^n} e^{\mathbf{t}_0'\mathbf{z} + t_{n+1}a} \left(\int_{-\infty}^a f_{\mathbf{z}^*}(\mathbf{z}) dy^* \right) d\mathbf{x} \\
&\quad + \int_{R^n} e^{\mathbf{t}_0'\mathbf{z} + t_{n+1}a} \left(\int_b^{-\infty} f_{\mathbf{z}^*}(\mathbf{z}) dy^* \right) d\mathbf{x},
\end{aligned} \tag{3}$$

όπου $\mathbf{x} = (x_1, \dots, x_n)'$. Ο πρώτος όρος στη δεξιά πλευρά της σχέσης (3) γράφεται ως εξής:

$$A_1(\mathbf{t}) = \int_{R^n} \int_a^b e^{\mathbf{t}'\mathbf{z}} \frac{1}{2\pi|\mathbf{S}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{z} - \mathbf{m})'\mathbf{S}^{-1}(\mathbf{z} - \mathbf{m})\right) d\mathbf{z},$$

όπου για $\mathbf{r} = \mathbf{z} - \mathbf{m}$ προκύπτει

$$\begin{aligned}
A_1(\mathbf{t}) &= \int_{R^n} \int_{a-m_{n+1}}^{b-m_{n+1}} e^{\mathbf{t}'(\mathbf{r}+\mathbf{m})} \frac{1}{2\pi|\mathbf{S}|^{1/2}} \exp\left(-\frac{1}{2}\mathbf{r}'\mathbf{S}^{-1}\mathbf{r}\right) d\mathbf{r} \\
&= \frac{e^{\mathbf{t}'\mathbf{m}}}{2\pi|\mathbf{S}|^{1/2}} \int_{R^n} \int_{a-m_{n+1}}^{b-m_{n+1}} \exp\left(-\frac{1}{2}\mathbf{r}'\mathbf{S}^{-1}\mathbf{r} + \mathbf{t}'\mathbf{r}\right) d\mathbf{r}.
\end{aligned} \tag{4}$$

Τότε, για $\mathbf{j} = \mathbf{S} \cdot \mathbf{t}$ προκύπτει,

$$\begin{aligned}
A_1(\mathbf{t}) &= \frac{e^{\mathbf{t}'\mathbf{m}}}{2\pi|\mathbf{S}|^{1/2}} \int_{R^n} \int_{a-m_{n+1}}^{b-m_{n+1}} \exp\left(-\frac{1}{2}(\mathbf{r} - \mathbf{j})'\mathbf{S}^{-1}(\mathbf{r} - \mathbf{j}) + \frac{1}{2}\mathbf{t}'\mathbf{S}\mathbf{t}\right) d\mathbf{r} \\
&= \frac{\exp(\mathbf{t}'\mathbf{m} + \frac{1}{2}\mathbf{t}'\mathbf{S}\mathbf{t})}{2\pi|\mathbf{S}|^{1/2}} \int_{R^n} \int_{a-m_{n+1}}^{b-m_{n+1}} \exp\left(-\frac{1}{2}(\mathbf{r} - \mathbf{j})'\mathbf{S}^{-1}(\mathbf{r} - \mathbf{j})\right) d\mathbf{r}.
\end{aligned} \tag{5}$$

Θέτοντας $\mathbf{e} = \mathbf{r} - \mathbf{j}$, η (5) γράφεται ως εξής:

$$\begin{aligned}
A_1(\mathbf{t}) &= \frac{\exp(\mathbf{t}'\mathbf{m} + \frac{1}{2}\mathbf{t}'\mathbf{S}\mathbf{t})}{2\pi|\mathbf{S}|^{1/2}} \int_{R^n} \int_{a-m_{n+1}-j_{n+1}}^{b-m_{n+1}-j_{n+1}} \exp\left(-\frac{1}{2}\mathbf{e}'\mathbf{S}^{-1}\mathbf{e}\right) d\mathbf{e} \\
&= \exp(\mathbf{t}'\mathbf{m} + \frac{1}{2}\mathbf{t}'\mathbf{S}\mathbf{t}) (F_{e_{n+1}}(b - m_{n+1} - j_{n+1}) - F_{e_{n+1}}(a - m_{n+1} - j_{n+1})),
\end{aligned} \tag{6}$$

όπου $F_{e_{n+1}}(x)$ είναι η συνάρτηση κατανομής της τ.μ. $e_{n+1} \sim N(0, S_{n+1,n+1})$ και

$$j_{n+1} = \sum_{i=1}^{n+1} S_{n+1,i} t_i.$$

Στη συνέχεια, ο δεύτερος όρος της (3) υπολογίζεται ως εξής:

$$\begin{aligned}
A_2(\mathbf{t}) &= \frac{e^{t_{n+1}a}}{2\pi|\mathbf{S}|^{1/2}} \int_{R^n} \int_{-\infty}^a \exp(-\frac{1}{2}(\mathbf{z} - \mathbf{m})' \mathbf{S}^{-1}(\mathbf{z} - \mathbf{m}) + \mathbf{t}'_0 \mathbf{z}) d\mathbf{z} \\
&= \frac{\exp(t_{n+1}a + \mathbf{t}'_0 \mathbf{m})}{2\pi|\mathbf{S}|^{1/2}} \int_{R^n} \int_{-\infty}^{a-m_{n+1}} \exp(-\frac{1}{2}\mathbf{z}^T \mathbf{S}^{-1} \mathbf{z} + \mathbf{t}'_0 \mathbf{x}) d\mathbf{z}. \quad (7)
\end{aligned}$$

Στη συνέχεια, θέτοντας $\mathbf{j}_0 = \mathbf{S} \cdot \mathbf{t}_0$, η (7) γράφεται

$$\begin{aligned}
A_2(\mathbf{t}) &= \frac{\exp(t_{n+1}a + \mathbf{t}'_0 \mathbf{m} + \mathbf{t}'_0 \mathbf{S} \mathbf{t}_0 / 2)}{2\pi|\mathbf{S}|^{1/2}} \int_{R^n} \int_{-\infty}^{a-m_{n+1}} \exp(-\frac{1}{2}(\mathbf{z} - \mathbf{j}_0)^T \mathbf{S}^{-1}(\mathbf{z} - \mathbf{j}_0)) d\mathbf{z} \\
&= \frac{\exp(t_{n+1}a + \mathbf{t}'_0 \mathbf{m} + \mathbf{t}'_0 \mathbf{S} \mathbf{t}_0 / 2)}{2\pi|\mathbf{S}|^{1/2}} \int_{R^{n-1}} \int_{-\infty}^{a-m_{n+1}-j_{0,n+1}} \exp(-\frac{1}{2}\mathbf{e}' \mathbf{S}^{-1} \mathbf{e}) d\mathbf{e} \\
&= \exp(t_{n+1}a + \mathbf{t}'_0 \mathbf{m} + \mathbf{t}'_0 \mathbf{S} \mathbf{t}_0 / 2) F_{e_{n+1}}(a - m_{n+1} - j_{0,n+1}), \quad (8)
\end{aligned}$$

όπου $j_{0,n+1} = \sum_{i=1}^{n+1} S_{n+1,i} t_{0,i}$.

Ο τρίτος όρος της (3) αποδεικνύεται, όμοια με την (7), ότι είναι ίσος με

$$A_3(\mathbf{t}) = \exp(t_{n+1}b + \mathbf{t}'_0 \mathbf{m} + \mathbf{t}'_0 \mathbf{S} \mathbf{t}_0 / 2) (1 - F_{e_{n+1}}(b - m_{n+1} - j_{0,n+1})). \quad (9)$$

Η ροπογεννήτρια της \mathbf{z} προκύπτει από την αντικατάσταση των (6), (8) και (9) στην (3). \square

Στη ροπογεννήτρια συνάρτηση της \mathbf{z} , αντικαθιστώντας $\mathbf{t} = (0, \dots, 0, t_{n+1})$, προκύπτει η ροπογεννήτρια της λογοκριμένης περιθώριας μεταβλητής, y , η οποία είναι ίση με,

$$\begin{aligned}
M_y(t_{n+1}) &= \exp(t_{n+1} - m_{n+1} + \frac{1}{2} t_{n+1}^2 S_{n+1,n+1}) \cdot \\
&\quad (F_{e_{n+1}}(b - m_{n+1} - S_{n+1,n+1} t_{n+1}) - F_{e_{n+1}}(a - m_{n+1} - S_{n+1,n+1} t_{n+1})) \\
&\quad + \exp(t_{n+1}a) F_{e_{n+1}}(a - m_{n+1}) + \exp(t_{n+1}b) (1 - F_{e_{n+1}}(b - m_{n+1})), \quad (10)
\end{aligned}$$

και είναι της ίδιας μορφής, όπως στην εργασία Loumponias et al. (2018).

Από την (10), υπολογίζονται άμεσα η λογοκριμένη αναμενόμενη τιμή και η διασπορά της τ.μ y , που είναι ίσες με:

$$\begin{aligned}
E(y) &= \left. \frac{dM_y(t)}{dt} \right|_{t=0} = aF_{e_{n+1}}(a - m_{n+1}) + b(1 - F_{e_{n+1}}(b - m_{n+1})) \\
&\quad + (F_{e_{n+1}}(b - m_{n+1}) - F_{e_{n+1}}(a - m_{n+1}))m_{n+1} \\
&\quad + S_{n+1,n+1}(f_{e_{n+1}}(a - m_{n+1}) - f_{e_{n+1}}(b - m_{n+1})), \tag{11}
\end{aligned}$$

$$\begin{aligned}
Var(y) &= \left. \frac{d^2M_y(t)}{dt^2} \right|_{t=0} - E^2(y) = \\
&= a^2F_{e_{n+1}}(a - m_{n+1})(1 - F_{e_{n+1}}(a - m_{n+1})) \\
&\quad + b^2F_{e_{n+1}}(b - m_{n+1})(1 - F_{e_{n+1}}(b - m_{n+1})) + m_{n+1}P(1 - P) + S_{n+1,n+1}^2P \\
&\quad + 2m_{n+1}S_{n+1,n+1}(f_{e_{n+1}}(a - m_{n+1}) - f_{e_{n+1}}(b - m_{n+1})) \\
&\quad + S_{n+1,n+1}^2((a - m_{n+1})f_{e_{n+1}}(a - m_{n+1}) - (b - m_{n+1})f_{e_{n+1}}(b - m_{n+1})) \\
&\quad - 2abF_{e_{n+1}}(a - m_{n+1})(1 - F_{e_{n+1}}(b - m_{n+1})) \\
&\quad - S_{n+1,n+1}^2(f_{e_{n+1}}(a - m_{n+1}) - f_{e_{n+1}}(b - m_{n+1}))^2 \\
&\quad - 2[Pm_{n+1} + S_{n+1,n+1}(f_{e_{n+1}}(a - m_{n+1}) - f_{e_{n+1}}(b - m_{n+1}))] \cdot \\
&\quad \cdot [aF_{e_{n+1}}(a - m_{n+1}) + b(1 - F_{e_{n+1}}(b - m_{n+1}))], \tag{12}
\end{aligned}$$

όπου $P = F_{e_{n+1}}(b - m_{n+1}) - F_{e_{n+1}}(a - m_{n+1})$.

Επιπλέον, από την Πρόταση 1 και την (11), μπορεί να υπολογιστεί η συνδιασπορά της περιθώριας τ.μ. $x_i \sim N(m_i, S_{i,i})$ και της λογοκριμένης τ.μ. y :

$$\begin{aligned}
cov(x_i, y) &= E(x_i y) - E(x_i)E(y) \\
&= \left. \frac{dM_{\mathbf{z}}(\mathbf{t})}{dt_i dt_{n+1}} \right|_{\mathbf{t}=\mathbf{0}} - m_i E(y) \\
&= P \cdot S_{i,n+1} \tag{13}
\end{aligned}$$

όπου $S_{i,n+1} = cov(x_i, y^*)$.

3. ΤΟΒΙΤ ΚΑΛΜΑΝ ΦΙΛΤΡΟ ΜΕ ΕΓΧΡΩΜΟ ΘΟΡΥΒΟ

Το μοντέλο χώρο καταστάσεων, σε ένα γραμμικό σύστημα με έγχρωμο θόρυβο, δίνεται από το παρακάτω σύστημα,

$$\begin{cases} \mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{u}_t, \\ \mathbf{u}_t = \mathbf{C}\mathbf{u}_{t-1} + \mathbf{w}_t, \\ y_t^* = \mathbf{H}\mathbf{x}_t + v_t, \end{cases} \tag{14}$$

όπου \mathbf{x}_t και y_t^* είναι το διάνυσμα κατάστασης και η μέτρηση τη χρονική στιγμή t , αντίστοιχα. \mathbf{A} και \mathbf{H} είναι ο πίνακας μετάβασης και παρατήρησης, αντίστοιχα.

$\mathbf{w}_t \sim N(\mathbf{0}, \mathbf{Q})$ και $v_t \sim N(0, r^2)$ είναι οι θόρυβοι των καταστάσεων και μετρήσεων, αντίστοιχα. Το \mathbf{u}_t είναι ένας έγχρωμος θόρυβος που περιγράφεται από ένα μοντέλο $AR(1)$ (ο πίνακας \mathbf{C} θεωρείται γνωστός ή εκτιμάται).

Το ΚΦ υπολογίζει αμερόληπτους εκτιμητές για τις κρυφές καταστάσεις, \mathbf{x}_t , όταν ο θόρυβος είναι λευκός. Για τον σκοπό αυτό, ο αλγόριθμος ΕΚΦ περιγράφει το σύστημα (14) ως μοντέλο χώρου καταστάσεων με λευκό θόρυβο, θέτοντας (Anderson 2012, Chui et al. 2017): $\bar{\mathbf{A}} = \begin{bmatrix} \mathbf{A} & \mathbf{I}_n \\ \mathbf{0}_n & \mathbf{C} \end{bmatrix}$, $\mathbf{v}_t = (\mathbf{x}'_t, \mathbf{u}'_t)'$, $\bar{\mathbf{H}} = [\mathbf{H} \ \mathbf{0}']$, $\bar{\mathbf{w}}_t = (\mathbf{0}', \mathbf{w}'_t)' \sim N(\mathbf{0}, \bar{\mathbf{Q}})$, όπου $\bar{\mathbf{Q}} = \begin{bmatrix} \mathbf{0}_n & \mathbf{0}_n \\ \mathbf{0}_n & \mathbf{Q} \end{bmatrix}$. Τότε το σύστημα (14) γράφεται ως εξής:

$$\begin{cases} \mathbf{v}_{t+1} = \bar{\mathbf{A}}\mathbf{z}_t + \bar{\mathbf{w}}_t \\ y_t^* = \bar{\mathbf{H}}\mathbf{v}_t + v_t \end{cases} \quad (15)$$

Ο αλγόριθμος του ΕΚΦ υπολογίζει τους βέλτιστους γραμμικούς εκτιμητές \mathbf{v}_t , όταν το μοντέλο χώρου καταστάσεων είναι της μορφής (15). Ο αλγόριθμος περιλαμβάνει δύο στάδια. Στο πρώτο στάδιο (*prediction step*) γίνεται μια εκ των προτέρων εκτίμηση του διανύσματος κατάστασης στο χρόνο t , δηλαδή, χωρίς να είναι γνωστή η μέτρηση τη δεδομένη χρονική στιγμή. Στο δεύτερο στάδιο (*updating step*) διορθώνεται η αρχική εκτίμηση, λαμβάνοντας υπόψη τη μέτρηση στο χρόνο t . Πρέπει να επισημανθεί ότι ο αλγόριθμος του ΕΚΦ μετά την αφαίρεση του έγχρωμου θορύβου ταυτίζεται με τον αντίστοιχο του ΚΦ. Παρακάτω, παρατίθενται οι εξισώσεις του αλγόριθμου ΕΚΦ.

Prediction Step:

$$\hat{\mathbf{v}}_t^- = \bar{\mathbf{A}}\mathbf{v}_{t-1} \quad (16)$$

$$\mathbf{P}_t^- = \bar{\mathbf{A}}\mathbf{P}_{t-1}\bar{\mathbf{A}}' + \bar{\mathbf{Q}} \quad (17)$$

Updating Step:

$$\mathbf{C}_{\mathbf{v}, y^*} = cov(\mathbf{v}_t, y_t^* | t-1) = \mathbf{P}_t^- \bar{\mathbf{H}}' \quad (18)$$

$$s_{y^*}^2 = var(y_t^* | t-1) = \bar{\mathbf{H}}\mathbf{P}_t^- \bar{\mathbf{H}}' + r^2 \quad (19)$$

$$m_{y^*} = E(y_t^* | t-1) = \bar{\mathbf{H}}\hat{\mathbf{v}}_t^- \quad (20)$$

$$\hat{\mathbf{v}}_t = \hat{\mathbf{v}}_t^- + \mathbf{C}_{\mathbf{z}, y^*} (s_{y^*}^2)^{-1} (y_t^* - m_{y^*}) \quad (21)$$

$$\mathbf{P}_t = \mathbf{P}_t^- - \mathbf{C}_{\mathbf{z}, y^*} (s_{y^*}^2)^{-1} \mathbf{C}'_{\mathbf{z}, y^*} \quad (22)$$

όπου $\hat{\mathbf{v}}_t^-$ και $\hat{\mathbf{v}}_t$ είναι οι εκ των προτέρων και υστέρων εκτιμήσεις (κατέπεχταση και της \mathbf{x}_t) τη χρονική στιγμή t , αντίστοιχα. \mathbf{P}_t^- και \mathbf{P}_t είναι οι πίνακες συνδιασπορών των εκ των προτέρων και υστέρων σφαλμάτων εκτίμησης, αντίστοιχα.

Οι σχέσεις (16)-(22) παρέχουν βέλτιστους εκτιμητές όταν οι μετρήσεις, y^* , ακολουθούν κανονική κατανομή. Στην περίπτωση των λογοκριμένων μετρήσεων, y , στο διάστημα $[a, b]$ (1), τροποποιούνται οι υπολογισμοί των (18)-(20) (Allik et al. 2014, 2015). Οι (16)-(17) δε μεταβάλλονται διότι χρησιμοποιούνται μόνο οι εκ των υστέρων εκτιμήσεις από την προηγούμενη χρονική στιγμή και όχι η λογοκριμένη μέτρηση, y_t . Συνεπώς, οι εκ των προτέρων εκτιμήσεις θα παραμείνουν ίδιες και στη διαδικασία ΕΤΚΦ.

Από την (13) και (18) προκύπτει ότι

$$\mathbf{C}_{\mathbf{v},y} = P\mathbf{C}_{\mathbf{v},y^*} \quad (23)$$

όπου P είναι η πιθανότητα η λογοκριμένη μέτρηση, y_t , να ανήκει μέσα στο διάστημα $[a, b]$ και υπολογίζεται από την

$$P = F_e(b - m_{y^*}) - F_e(a - m_{y^*}),$$

όπου $e \sim N(0, s_{y^*}^2)$. Για τον υπολογισμό της λογοκριμένης διασπορά s_y^2 αντικαθιστούμε στην (12), όπου $m_{n+1} = m_{y^*}$ και $S_{n+1,n+1} = s_{y^*}^2$ και προκύπτει το ζητούμενο. Θα πρέπει να σημειωθεί, ότι η διασπορά (12) διαφέρει από αυτήν που υπολογίζεται στις Allik et al. (2014, 2015) και Loumponias (2016). Στις παραπάνω περιπτώσεις δίνεται μια προσέγγιση και όχι ο ακριβής υπολογισμός. Στη συνέχεια, η λογοκριμένη αναμενόμενη τιμή, m_y , προκύπτει άμεσα από την (11) αντικαθιστώντας $m_{n+1} = m_{y^*}$ και $S_{n+1,n+1} = s_{y^*}^2$. Συνοπτικά, οι εκ των υστέρων εκτιμήσεις με λογοκριμένες μετρήσεις (παρατηρήσεις) στη διαδικασία ΕΤΚΦ είναι,

$$\hat{\mathbf{v}}_t = \hat{\mathbf{v}}_t^- + \mathbf{C}_{\mathbf{z},y}(s_y^2)^{-1}(y_t - m_y) \quad (24)$$

$$\mathbf{P}_t = \mathbf{P}_t^- - \mathbf{C}_{\mathbf{z},y}(s_y^2)^{-1}\mathbf{C}'_{\mathbf{z},y} \quad (25)$$

όπου τα m_y , s_y^2 , $\mathbf{C}_{\mathbf{z},y}$ υπολογίζονται από τις (11)-(13), αντίστοιχα.

Σε πολλά προβλήματα με λογοκριμένες μετρήσεις και με έγχρωμο θόρυβο στις καταστάσεις, η διασπορά των σφαλμάτων μέτρησης, r^2 , και ο πίνακας \mathbf{C} (14) είναι άγνωστες παράμετροι. Για την εκτίμηση των αγνώστων παραμέτρων του συστήματος (14), χρησιμοποιείται η συνάρτηση πιθανοφάνειας των λογοκριμένων μετρήσεων $\{y_t\}_{t=1}^T$ (T πλήθος των μετρήσεων) που δίνεται από την παρακάτω εξίσωση:

$$\begin{aligned} L(\mathbf{y}) = & \prod_{a < y_t < b} \frac{1}{(\overline{\mathbf{H}}\mathbf{P}_t^-\overline{\mathbf{H}}' + r^2)^{1/2}} \phi\left(\frac{y_t - \overline{\mathbf{H}}\hat{\mathbf{v}}_t^-}{(\overline{\mathbf{H}}\mathbf{P}_t^-\overline{\mathbf{H}}' + r^2)^{1/2}}\right) \\ & \times \prod_{y_t=a} \Phi\left(\frac{a - \overline{\mathbf{H}}\hat{\mathbf{v}}_t^-}{(\overline{\mathbf{H}}\mathbf{P}_t^-\overline{\mathbf{H}}' + r^2)^{1/2}}\right) \\ & \times \prod_{y_t=b} \left(1 - \Phi\left(\frac{b - \overline{\mathbf{H}}\hat{\mathbf{v}}_t^-}{(\overline{\mathbf{H}}\mathbf{P}_t^-\overline{\mathbf{H}}' + r^2)^{1/2}}\right)\right), \end{aligned} \quad (26)$$

όπου ϕ και Φ είναι η συνάρτηση πυκνότητας πιθανότητας και η αθροιστική κατανομή της τυπικής κανονικής κατανομής.

4. ΠΕΙΡΑΜΑΤΑ

Στην ενότητα αυτή συγκρίνουμε την επίδοση του προτεινόμενου ΕΤΚΦ έναντι του ΤΚΦ και ΕΚΦ. Για τον σκοπό αυτό, μελετάται ένας ταλαντωτής (χωρίς απόσβεση), όπως έχει γίνει και σε άλλες εργασίες με λογοκριμένες μετρήσεις (Allik et al. 2014, Geng et al. 2017, Han et al. 2018). Το μοντέλο του ταλαντωτή περιγράφεται από τις εξισώσεις:

- $\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{u}_{t-1}, \quad t = 1, \dots, 1000.$
- $\mathbf{u}_t = \mathbf{C}\mathbf{u}_{t-1} + \mathbf{w}_t,$
- $y_t^* = \mathbf{H}\mathbf{x}_t + v_t$

όπου $\mathbf{A} = \begin{bmatrix} \cos(w) & -\sin(w) \\ \sin(w) & \cos(w) \end{bmatrix}, \quad w = 0.005 \cdot 2\pi, \mathbf{H} = [1 \quad 0.5], \mathbf{C} = \begin{bmatrix} 0.8 & 0 \\ 0 & 0.6 \end{bmatrix},$

και τα όρια λογοκρισίας είναι $a = -1$ και $b = 1, \mathbf{w}_t \sim N(\mathbf{0}, \mathbf{Q}), \mathbf{Q} = \begin{bmatrix} 0.05^2 & 0 \\ 0 & 0.05^2 \end{bmatrix}$

και $v_k \sim N(0, r^2 = 0.5).$

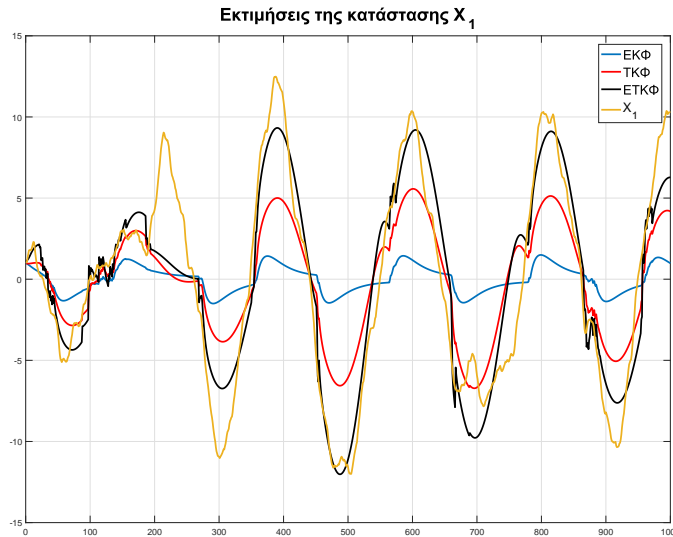
Αρχικά, ο πίνακας \mathbf{C} και η διασπορά r^2 θεωρούνται γνωστά προκειμένου να κατασκευάσουμε τις μετρήσεις y_t^* , σύμφωνα με το παραπάνω μοντέλο. Στη συνέχεια, οι μετρήσεις λογοκρίνονται στο διάστημα $[-1, 1]$. Σκοπός του παραπάνω προβλήματος είναι να εκτιμήσουμε τις καταστάσεις, $\mathbf{x}_t = (x_{1,t}, x_{2,t})$ χρησιμοποιώντας τις λογοκριμένες μετρήσεις. Για τον λόγο αυτό χρησιμοποιούνται το ΕΚΦ, ΤΚΦ και ΕΤΚΦ. Στην προτεινόμενη μέθοδο, πριν την εκτέλεση του αλγορίθμου χρησιμοποιείται η συνάρτηση πιθανοφάνειας (26) για την εκτίμηση των \mathbf{C} και r^2 , στη συνέχεια εφαρμόζεται το ΕΤΚΦ. Η παραπάνω διαδικασία επαναλαμβάνεται 100 φορές και σε κάθε επανάληψη υπολογίζεται η ρίζα του μέσου τετραγωνικό σφάλμα εκτίμησης (PMTΣΕ) για τις κρυφές καταστάσεις \mathbf{x}_t .

Φίλτρο	μ.ο των PMTΣΕ για \hat{x}_1	μ.ο των PMTΣΕ για \hat{x}_2
ΕΚΦ	7.3619	7.2980
ΤΚΦ	4.6689	4.6143
ΕΤΚΦ	2.7811	2.7490

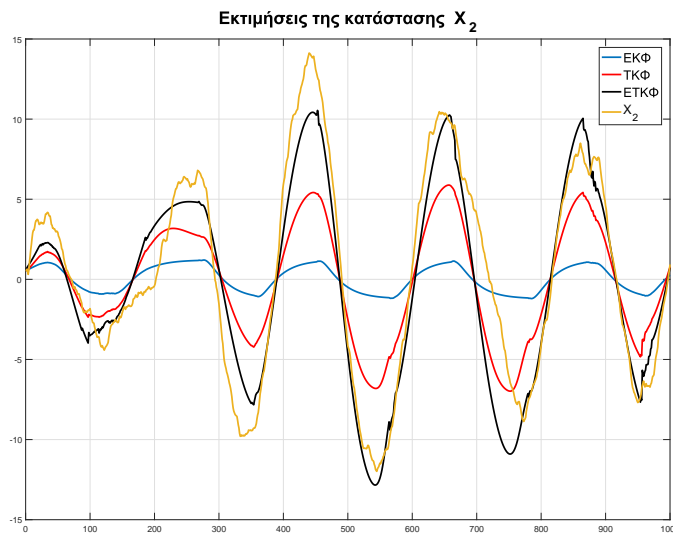
Πίνακας 1: Μέσοι όροι των PMTΣΕ για το ΕΚΦ, ΤΚΦ και ΕΤΚΦ, αντίστοιχα.

Οι μέσοι όροι (μ.ο.) των PMTΣΕ στις 100 επαναλήψεις για τα τρία φίλτρα (ΕΚΦ, ΤΚΦ, ΕΤΚΦ) παρουσιάζονται στον Πίνακα 1. Όπως γίνεται εύκολα αντιληπτό, το ΕΚΦ έχει τη μεγαλύτερη PMTΣΕ και για τις δύο καταστάσεις, αφού δεν μπορεί να διαχειριστεί λογοκριμένες μετρήσεις (Σχήματα 1-2). Το ΤΚΦ έχει καλύτερη επίδοση από το ΕΚΦ (μικρότερη PMTΣΕ) ωστόσο δεν μπορεί να διαχειριστεί την ετεροσκεδαστικότητα που προκύπτει λόγω του πίνακα \mathbf{C} . Η προτεινόμενη μέθοδος έχει την

καλύτερη επίδοση βελτιώνοντας το ΤΚΦ, διότι α) έγινε η διόρθωση της λογοκριμένης διασποράς και β) λάβαμε υπόψη ότι ο θόρυβος είναι έγχρωμος. Στα Σχήματα 1-2 παρουσιάζονται οι εκτιμήσεις των τριών φίλτρων για τις δύο κρυφές καταστάσεις (κίτρινο χρώμα).



Σχήμα 1: Οι εκτιμήσεις των φίλτρων EKΦ(μπλε), TKΦ (κόκκινο), ETKΦ (μαύρο) για την κρυφή κατάσταση x_1 (κίτρινο).



Σχήμα 2: Οι εκτιμήσεις των φίλτρων EKΦ(μπλε), TKΦ (κόκκινο), ETKΦ (μαύρο) για την κρυφή κατάσταση x_2 (κίτρινο).

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Σκοπός της εργασίας αυτής είναι α) να διορθώσει τον ήδη γνωστό αλγόριθμο ΤΚΦ (Allik et al. 2014, 2015) και β) να τον βελτιώσει έτσι ώστε να μπορεί να διαχειρίζεται έγχρωμο θόρυβο στις καταστάσεις (πολυδιάστατες). Για τον σκοπό αυτό, υπολογίστηκε αναλυτικά η ροπογεννήτρια συνάρτηση μιας λογοκριμένης (μονοδιάστατης) τυχαίας μεταβλητής. Στη συνέχεια, μέσω της ροπογεννήτριας υπολογίστηκαν η ακριβής λογοκριμένη αναμενόμενη τιμή και διασπορά. Επιπλέον χρησιμοποιήθηκε η διαδικασία μετατροπής του μοντέλου χώρου καταστάσεων από σύστημα με έγχρωμο σε σύστημα με λευκό θόρυβο (Anderson 2012, Chui et al. 2017). Η προτεινόμενη μέθοδος, ΕΤΚΦ, συγκρίθηκε με το ΕΚΦ και ΤΚΦ σε τεχνητά δεδομένα (προσημείωση ταλαντωτή). Το πρόβλημα που μελετήθηκε, περιλάμβανε μονοδιάστατες λογοκριμένες μετρήσεις και διδιάστατες κρυφές καταστάσεις. Τα αποτελέσματα έδειξαν ότι το ΕΤΚΦ υπερτερεί έναντι του ΕΚΦ και ΤΚΦ με κριτήριο τη μείωση της ΡΜΤΣΕ και στις δύο κρυφές καταστάσεις. Αυτό οφείλεται στο γεγονός ότι α) έγινε η διόρθωση της λογοκριμένης διασποράς και β) το ΕΤΚΦ είναι ικανό να διαχειριστεί έγχρωμο θόρυβο, που εμφανίζεται στις κρυφές καταστάσεις.

Μια μελλοντική έρευνα που μπορεί να πραγματοποιηθεί, είναι η επέκταση του αλγορίθμου ΕΤΚΦ στις πολυδιάστατες λογοκριμένες μετρήσεις, οι οποίες θα είναι συσχετισμένες μεταξύ τους.

ABSTRACT

This paper deals with the hidden states estimation in continuous-space and discrete-time state models, when the observed measurements (observations) contain white noise with unknown but constant variance, and are censored. Due to censoring, the model of hidden states is non-linear, thus, a simple Kalman filter, cannot be utilized. To carry out the required estimation, the moments of the censored normal distribution are calculated by using the moment generating function. The hidden states are considered to contain noise with non-constant covariance matrix (color noise) and the Kalman Color Filters theory is used to deal with heteroskedasticity. The results used on artificial data show that the proposed method provides unbiased estimates for hidden states and has a smaller mean square error than other filters.

ΑΝΑΦΟΡΕΣ

Allik, B., Miller, C., Piovoso, M. J., Zurakowski, R. (2014). Estimation of saturated data using the Tobit Kalman filter. In 2014 American Control Conference (pp. 4151-4156). IEEE.

- Allik, B., Miller, C., Piovoso, M. J., & Zurakowski, R. (2015). The Tobit Kalman filter: an estimator for censored measurements. *IEEE Transactions on Control Systems Technology*, 24(1), 365-371.
- Anderson, B. D., & Moore, J. B. (2012). *Optimal filtering*. Courier Corporation. Dover Publications INC. Mineloa, New York.
- Bryson, A., & Johansen, D. (1965). Linear filtering for time-varying systems using measurements containing colored noise. *IEEE Transactions on Automatic Control*, 10(1), 4-10.
- Buckley, J., James, I. (1979). Linear regression with censored data. *Biometrika*, 66(3), 429-436.
- Chui, C. K., & Chen, G. (2017). *Kalman filtering*. Springer International Publishing, 5th Edition.
- Geng, H., Wang, Z., Liang, Y., Cheng, Y., & Alsaadi, F. E. (2017). Tobit Kalman filter with fading measurements. *Signal Processing*, 140, 60-68.
- Han, F., Dong, H., Wang, Z., Li, G., & Alsaadi, F. E. (2018). Improved Tobit Kalman filtering for systems with random parameters via conditional expectation. *Signal Processing*, 147, 35-45.
- Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1), 35-45.
- Koo, B., Gibson, J. D., & Gray, S. D. (1989, May). Filtering of colored noise for speech enhancement and coding. In *International Conference on Acoustics, Speech, and Signal Processing*, (pp. 349-352). IEEE.
- Loumponias, K., Dimou, A., Vretos, N., Daras, P. (2018, November). Adaptive Tobit Kalman-Based Tracking. In *2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)* (pp. 70-76). IEEE.
- Loumponias, K., Vretos, N., Daras, P., Tsaklidis, G. (2016). Using kalman filter and tobit kalman filter in order to improve the motion recorded by kinect sensor ii. In *Proceedings of the 29th Panhellenic Statistics Conference*, (pp. 322-334).
- Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, 24-36.
- Zimmerman, W. (1969). On the optimum colored noise Kalman filter. *IEEE Transactions on Automatic Control*, 14(2), 194-196.



ΕΚ ΤΩΝ ΠΡΟΤΕΡΩΝ ΕΚΤΙΜΗΣΗ ΤΩΝ ΣΦΑΛΜΑΤΩΝ ΠΑΡΑΤΗΡΗΣΗΣ ΣΤΟ ΦΙΛΤΡΟ ΣΩΜΑΤΙΔΙΩΝ

Λύκου Ροδή¹, Τσακλίδης Γεώργιος¹
Τμήμα Μαθηματικών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
{lykourodi, tsaklidi}@math.auth.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία πραγματοποιείται μελέτη της κατανομής των σφαλμάτων παρατήρησης στα σωματίδια του Φίλτρου Σωματιδίων μονοδιάστατων μεταβλητών όπου η εξίσωση παρατήρησης δίνεται με γραμμική σχέση. Εκτιμάται το σφάλμα παρατήρησης που προκύπτει ένα χρονικό βήμα έπειτα από το τρέχον βήμα. Η εκτίμηση αυτή προστίθεται ως νέο βήμα στον αλγόριθμο του φίλτρου. Έτσι, προκύπτουν η εκτίμηση των βαρών για το επόμενο χρονικό βήμα, καθώς και η πιθανότητα αναπαραγωγής σωματιδίου κατά τη διάρκεια της αναδειγματοληψίας του επόμενου βήματος, ώστε να είναι δυνατόν να κατασκευαστεί πλέγμα στο χώρο καταστάσεων και να εκτιμηθεί η κατανομή των σωματιδίων στα κελιά του πλέγματος. Παρατίθενται παραδείγματα για γνωστές κατανομές καθώς και προσομοιωτικό παράδειγμα με ελλιπή δεδομένα, όπου εφαρμόζεται η προτεινόμενη μέθοδος.

Λέξεις Κλειδιά: Φίλτρο σωματιδίων, Εκτίμηση βαρών, Εκφυλισμός (Degeneracy), Φτωχοποίηση (Improverishment), Ελλιπή δεδομένα (Missing data)

1. ΕΙΣΑΓΩΓΗ

Η περιγραφή δυναμικών συστημάτων περιλαμβάνει τη μη γραμμική μοντελοποίηση, συχνά με μη Γκαουσιανούς θορύβους, που στοχεύει στην εκτίμηση κρυφών (λανθανόντων) μεταβλητών (latent variables) παρουσία παρατηρήσεων (observations-measurements) σχετιζόμενων με τις κρυφές μεταβλητές. Τέτοια μοντέλα έχουν συνήθως τη μορφή

$$X_t \square h(X_{t-1}, V_t) \quad (1)$$

$$Y_t \square g(X_t, U_t) \quad (2)$$

όπου $t = 0, 1, 2, \dots$ είναι ο δείκτης του χρόνου, $\{X_t\}_{t \geq 0}$ είναι η κρυφή n -διάστατη Μαρκοβιανή διαδικασία πρώτης τάξης, $\{Y_t\}_{t \geq 0}$ είναι η m -διάστατη ακολουθία των μεταβλητών παρατήρησης, το σύμβολο \square δηλώνει κατανομή σύμφωνα με κάποια συνάρτηση πυκνότητας πιθανότητας, οι συναρτήσεις h, g είναι συναρτήσεις πυκνότητας πιθανότητας, με V_t συμβολίζεται ο (οποιασδήποτε διάστασης) θόρυβος

της διαδικασίας (process noise) και με U_t ο (οποιασδήποτε διάστασης) θόρυβος παρατήρησης/μέτρησης (observation/measurement noise), όπου κάθε ακολουθία $\{V_t\}_{t \geq 0}$, $\{U_t\}_{t \geq 0}$ απαρτίζεται από ανεξάρτητες και ισόνομες τυχαίες μεταβλητές. Το Φίλτρο Σωματιδίων (Particle Filter) (Gordon et al., 1993) είναι ένας αλγόριθμος που μπορεί να εκτιμήσει την κατανομή του διανύσματος κατάστασης X_t τη χρονική στιγμή t βασισμένο σε Monte Carlo προσομοιώσεις της κατανομής της κρυφής μεταβλητής. Κατά την εφαρμογή του Φίλτρου Σωματιδίων, σε κάθε χρονικό βήμα εκτελείται Monte Carlo προσομοίωση της κατανομής της κρυφής μεταβλητής και ανατίθενται βάρη στα παραγόμενα δείγματα της προσομοίωσης, τα σωματίδια (particles), σύμφωνα με τη ληφθείσα παρατήρηση. Έτσι, προσεγγίζονται η εκ των προτέρων (prior) καθώς και η εκ των υστέρων (posterior) κατανομή της κρυφής μεταβλητής πριν και μετά τη λήψη της παρατήρησης.

Επειδή ο υπολογισμός των βαρών των σωματιδίων αποτελεί δομικό συστατικό του αλγορίθμου, επηρεάζει σημαντικά τις επιδόσεις του φίλτρου. Από τον υπολογισμό των βαρών προκύπτουν δυο από τα μεγαλύτερα προβλήματα του Φίλτρου Σωματιδίων, ο εκφυλισμός (degeneracy) και η φτωχοποίηση (impoverishment). Εκφυλισμός καλείται το φαινόμενο κατά το οποίο ένας πολύ μικρός αριθμός σωματιδίων συγκεντρώνει τόσο μεγάλα βάρη ώστε η παρουσία των υπολοίπων σωματιδίων καθίσταται αμελητέα. Για την αντιμετώπιση του εκφυλισμού έχει προταθεί η αναδειγματοληψία των σωματιδίων με επανατοποθέτηση, σύμφωνα με τα υπολογιζόμενα βάρη, ώστε να αυξηθεί ο αριθμός των σωματιδίων μεγάλων βαρών στο δείγμα. Ωστόσο, αυτή η διαδικασία οδηγεί προοδευτικά όλα τα σωματίδια να παίρνουν τιμές από ένα πολύ μικρό σύνολο τιμών σε σχέση με τις υπόλοιπες δυνατές τιμές. Το φαινόμενο αυτό ονομάζεται φτωχοποίηση. Αν και έχουν αναπτυχθεί ποικίλες τεχνικές για την αντιμετώπιση των προαναφερθέντων προβλημάτων, ωστόσο, με αυτές τις παρεμβάσεις, η πολυπλοκότητα του φίλτρου αυξάνεται τόσο ώστε σε κάποιες περιπτώσεις χάνεται η δυνατότητα εκτίμησης της μεταβλητής κατάστασης ταυτόχρονα με τη συλλογή των δεδομένων (Li et al., 2014). Παράλληλα, καθώς τα βάρη επηρεάζουν την τελική εικόνα της κατανομής της κρυφής μεταβλητής στο τέλος ενός χρονικού βήματος, επηρεάζουν και τις προβλέψεις για τα επόμενα βήματα μέσω της εξίσωσης κατάστασης. Συνεπώς, έχουν καίριο ρόλο στις εκ των προτέρων εκτιμήσεις για το επόμενο χρονικό βήμα. Ο ρόλος αυτός ενισχύεται ακόμη περισσότερο σε μοντέλα ελλিপών δεδομένων (missing data), οπότε η μέτρηση απουσιάζει σε κάποιο χρονικό βήμα και η εκ νέου ανάθεση βαρών μέσω της εξίσωσης παρατήρησης καθίσταται προβληματική. Για τη διαχείριση τέτοιων μοντέλων έχει προταθεί ο αλγόριθμος Μεγιστοποίησης Προσδοκίας (Expectation Maximization) είτε με γνωστή (Gopaluni, 2008) είτε με άγνωστη δομή (Gopaluni et al., 2009) των σχέσεων (1), (2), καθώς και η παραγωγή τυχαίων τιμών που αντικαθιστούν τις τιμές που απουσιάζουν (Zhang et al., 2015).

Στην παρούσα εργασία διερευνάται η δυνατότητα εκτίμησης των σφαλμάτων παρατήρησης του Φίλτρου Σωματιδίων για ένα χρονικό βήμα έπειτα από το τρέχον, για μοντέλα με μονοδιάστατες τυχαίες μεταβλητές, όπου η εξίσωση παρατήρησης δίνεται με γραμμική σχέση. Στο κεφάλαιο 2 η προτεινόμενη μέθοδος παρουσιάζεται

αναλυτικά. Στο κεφάλαιο 3 παρατίθεται προσομοιωτικό παράδειγμα. Στο κεφάλαιο 4 καταγράφονται σχετικά συμπεράσματα.

2. ΜΕΘΟΔΟΣ

2.1 Ο αλγόριθμος του φίλτρου σωματιδίων

Το Φίλτρο Σωματιδίων είναι ένα φίλτρο Bayes, ώστε για τις εκ των προτέρων και εκ των υστέρων εκτιμήσεις βασίζεται στις εξισώσεις

$$p(X_{t+1} | Y_{1:t}) = \int p(X_{t+1} | X_t) p(X_t | Y_{1:t}) dX_t \quad (\text{πρόβλεψη})$$

$$p(X_t | Y_{1:t}) = \frac{p(Y_t | X_t) p(X_t | Y_{1:t-1})}{\int p(Y_t | X'_t) p(X'_t | Y_{1:t-1}) dX'_t} \quad (\text{διόρθωση}),$$

όπου με p συμβολίζονται συναρτήσεις πυκνότητας πιθανότητας και $Y_{1:t} = \{Y_i\}_{i=1}^t$. Τα παραπάνω ολοκληρώματα είναι συχνά δύσκολο ή αδύνατο να υπολογιστούν αναλυτικά. Γι' αυτό το λόγο, αξιοποιείται η Monte Carlo ολοκλήρωση με τη βοήθεια της Monte Carlo προσομοίωσης. Το Φίλτρο Σωματιδίων απαιτεί:

- τον ορισμό του αριθμού N των σωματιδίων που προσεγγίζουν την κατανομή της κρυφής μεταβλητής,
- ένα κατώφλι N_{Thr} για την λήψη απόφασης για αναδειγματοληψία, καθώς και
- μια προτεινόμενη πυκνότητα $q(X_t)$ που προσεγγίζει συνολικά την πυκνότητα h της εξίσωσης (1), μιμούμενη τη διαδικασία μετάβασης σε ένα χρονικό βήμα από την παλαιότερη κατάσταση X_{t-1} στην X_t .

Όπως η διαδικασία περιγράφεται στον Αλγόριθμο 1, σε τυχαίο βήμα $t \geq 1$, η κατανομή της κρυφής μεταβλητής X_t προσεγγίζεται αρχικά από δείγμα $\{x_t^i(-)\}_{i=1}^N$ παραγόμενο από Monte Carlo προσομοίωση, σύμφωνα με την προτεινόμενη πυκνότητα $q(X_t)$. Με τη λήψη της παρατήρησης y_t , τα σωματίδια σταθμίζονται με βάρη $\{w_t^i\}_{i=1}^N \geq 0$ που αντιστοιχούν συνήθως στις τιμές της πυκνότητας των σφαλμάτων παρατήρησης $\{u_t^i\}_{i=1}^N$. Η εκτίμηση της συνάρτησης πυκνότητας πιθανότητας για την κρυφή μεταβλητή X_t θα είναι

$$\hat{p}(x_t) = \sum_{i=1}^N w_t^i \delta_{x_t^i}(x_t), \quad \text{με} \quad \sum_{i=1}^N w_t^i = 1,$$

όπου με δ_x συμβολίζεται η συνάρτηση μάζας δέλτα του Dirac. Τέλος, εξετάζεται αν

$$N_{Eff}(t) = \left(\sum_{i=1}^N (w_t^i)^2 \right)^{-1} > N_{Thr}.$$

Πρόκειται για ένα κριτήριο που δείχνει το βαθμό εκφυλισμού του υπάρχοντος δείγματος, προκειμένου να ληφθεί απόφαση για εκτέλεση αναδειγματοληψίας.

Αλγόριθμος 1.) Βασικός αλγόριθμος του φίλτρου σωματιδίων με αναδειγματοληψία

1) Παραγωγή δείγματος $\{x_0^i\}_{i=1}^N$ σύμφωνα με την $q(X_0)$.

2) Για $t \geq 1$

2α) Παραγωγή δείγματος $\{x_t^i(-)\}_{i=1}^N$ σύμφωνα με την $q(X_t)$.

2β) Ανάθεση βαρών $\{w_t^i\}_{i=1}^N$ σύμφωνα με την παρατήρηση y_t και κανονικοποίηση των βαρών.

2γ) Υπολογισμός της ποσότητας

$$N_{Eff}(t) = \left(\sum_{i=1}^N (w_t^i)^2 \right)^{-1}.$$

Εάν $N_{Eff}(t) > N_{Thr}$, τότε εκτελείται αναδειγματοληψία με επανατοποθέτηση βάσει των υπολογισμένων βαρών $\{w_t^i\}_{i=1}^N$ από την οποία προκύπτουν τα σωματίδια $\{x_t^i\}_{i=1}^N$ που χαρακτηρίζονται από ίσα μεταξύ τους βάρη. Διαφορετικά, τίθενται $\{x_t^i\}_{i=1}^N = \{x_t^i(-)\}_{i=1}^N$ με βάρη τα υπάρχοντα $\{w_t^i\}_{i=1}^N$.

2.2 Εκτίμηση των σφαλμάτων παρατήρησης επόμενου βήματος

Έστω ένα μοντέλο της μορφής

$$X_t = f(X_{t-1}) + V_t \quad (3)$$

$$Y_t = c + c_1 X_t + U_t \quad \text{ή} \quad Y_t - c - c_1 X_t = U_t \quad (4)$$

όπου $X_t \in \mathbb{R}^n$ είναι η κρυφή μεταβλητή, f είναι γνωστή προσδιοριστική (deterministic) συνάρτηση, $Y_t \in \mathbb{R}^m$ είναι η μεταβλητή παρατήρησης, V_t είναι ο θόρυβος της διαδικασίας, U_t είναι ο θόρυβος παρατήρησης, κάθε ακολουθία $\{V_t\}_{t \geq 0}$, $\{U_t\}_{t \geq 0}$ αποτελείται από ανεξάρτητες και ισόνομες τυχαίες μεταβλητές, και $c, c_1 \in \mathbb{R}^m$ είναι γνωστές σταθερές. Το μοντέλο αυτό αποτελεί ειδική περίπτωση του συστήματος εξισώσεων (1), (2).

Από το βασικό αλγόριθμο φαίνεται ότι ο υπολογισμός των βαρών σε επόμενο βήμα βασίζεται στα μετρούμενα σφάλματα παρατήρησης της εξίσωσης (4),

$$\{U_{t+1}^i\}_{i=1}^N = \{Y_{t+1} - c - c_1 X_{t+1}^i\}_{i=1}^N.$$

Τα βάρη θα λάβουν τις τιμές της συνάρτησης πυκνότητας που αντιστοιχίζεται στις τιμές $\{U_t^i\}_{i=1}^N$. Επομένως, από το προηγούμενο βήμα t , ζητούνται οι κατανομές των $\{U_{t+1}^i\}_{i=1}^N$, με δεδομένες τις παρατηρήσεις και τις εκτιμήσεις της κρυφής μεταβλητής μέχρι αυτό το βήμα.

Με αντικατάσταση από την εξίσωση (4),

$$U_{t+1}^i = Y_{t+1} - c - c_1 X_{t+1}^i, \quad (5)$$

όπου τα Y_{t+1}, X_{t+1}^i είναι άγνωστα τη χρονική στιγμή t . Επομένως, με αντικατάσταση του Y_{t+1} από την εξίσωση (4) έχουμε

$$U_{t+1}^i = c + c_1 X_{t+1} + U_{t+1} - c - c_1 X_{t+1}^i.$$

Με αντικατάσταση του X_{t+1} από την εξίσωση (3), προκύπτει

$$U_{t+1}^i = +c_1 f(X_t) + c_1 V_{t+1} + U_{t+1} - c_1 f(X_t^i) - c_1 V_{t+1}^i.$$

Συνεπώς,

$$U_{t+1}^i = +c_1 V_{t+1} + U_{t+1} - c_1 V_{t+1}^i + c_1 [f(X_t) - f(X_t^i)]. \quad (6)$$

Τα αθροίσματα $c_1 V_{t+1} + U_{t+1} - c_1 V_{t+1}^i$, $i = 1, \dots, N$ είναι ισόνομα με κατανομή τη συνέλιξη των κατανομών των μεταβλητών $c_1 V_{t+1}$, U_{t+1} , $(-c_1) V_{t+1}^i$, και μπορούν να υπολογιστούν. Αξίζει να σημειωθεί ότι η ζητούμενη συνέλιξη δεν οδηγεί πάντοτε σε αναλυτικές λύσεις για την κατανομή του ανωτέρω αθροίσματος. Ορισμένα χαρακτηριστικά παραδείγματα κατανομών για τους θορύβους παρατίθενται στο Παράρτημα. Η ποσότητα $f(X_t)$ είναι δυνατόν να εκτιμηθεί από την $f(\hat{X}_t)$ κατά τη χρονική στιγμή t , οπότε και εκτιμάται η κατανομή της X_t . Ως εκ τούτου, από την ανίσωση (6), τα σφάλματα παρατήρησης του επόμενου βήματος μπορούν να προσεγγιστούν από τα αθροίσματα τυχαίων μεταβλητών $c_1 V_{t+1} + U_{t+1} - c_1 V_{t+1}^i$, μετατοπισμένα κατά σταθερές $c_1 [f(\hat{X}_t) - f(x_t^i)]$, ως

$$U_{t+1}^i \square c_1 V_{t+1} + U_{t+1} - c_1 V_{t+1}^i + c_1 [f(\hat{X}_t) - f(x_t^i)].$$

Με γνωστή την κατανομή των αθροισμάτων $c_1 V_{t+1} + U_{t+1} - c_1 V_{t+1}^i$, είναι δυνατό να εκτιμηθούν οι κατανομές των U_{t+1}^i και, έτσι, οι ποσότητες U_{t+1}^i ως αναμενόμενες τιμές των εκτιμώμενων κατανομών. Εάν τα βάρη είναι μονότονες και διαφορίσιμες συναρτήσεις των σφαλμάτων παρατήρησης, όπως στην περίπτωση της πυκνότητας της Εκθετικής κατανομής, τότε μπορεί να υπολογιστεί η κατανομή και η αναμενόμενη τιμή τους. Διαφορετικά, ανατίθενται βάρη στις εκτιμήσεις των σφαλμάτων U_{t+1}^i . Η προτεινόμενη διαδικασία δύναται να τροποποιήσει την εκτίμηση της διασποράς των βαρών του βήματος 2γ στον Αλγόριθμο 1 ως εξής:

“2γ) Εκτίμηση σφαλμάτων παρατήρησης επόμενου βήματος $\{\hat{u}_{t+1}^i\}_{i=1}^N$ και των αντίστοιχων βαρών $\{\hat{w}_{t+1}^i\}_{i=1}^N$.

Υπολογισμός της ποσότητας

$$N_{Eff}(t) = \left(\sum_{i=1}^N (\hat{w}_{t+1}^i)^2 \right)^{-1}.$$

Εάν $N_{Eff}(t) > N_{Thr}$, τότε εκτελείται αναδειγματοληψία με επανατοποθέτηση βάσει των υπολογισμένων βαρών $\{w_t^i\}_{i=1}^N$ από την οποία προκύπτουν τα σωματίδια $\{x_t^i\}_{i=1}^N$ που χαρακτηρίζονται από ίσα μεταξύ τους βάρη. Διαφορετικά, τίθενται $\{x_t^i\}_{i=1}^N = \{x_t^i(-)\}_{i=1}^N$ με βάρη τα υπάρχοντα $\{w_t^i\}_{i=1}^N$.”

Η εκτίμηση των βαρών του επόμενου βήματος μπορεί να αξιοποιηθεί με πολλούς τρόπους. Για την περίπτωση του εκφυλισμού, είναι δυνατή η εκτίμηση της μελλοντικής συμπεριφοράς του δείγματος μέσα από τη διαδικασία αυτή. Καθώς η εκτίμηση της κατανομής των βαρών του επόμενου βήματος είναι η μόνη διαθέσιμη εκτίμηση για τη μελλοντική πορεία της κατανομής των βαρών, η αναδειγματοληψία, που υλοποιείται για την αντιμετώπιση του εκφυλισμού, είναι προτιμότερο να βασιστεί στην εκτίμηση των βαρών του επόμενου βήματος. Ακόμη, η πρόβλεψη της κατανομής των σωματιδίων του επόμενου χρονικού βήματος μπορεί να φτάσει μέχρι και την αναδειγματοληψία του επόμενου βήματος, καθώς με θεώρηση πλέγματος (διαμέρισης) $\Pi(X_t)$ k κελιών στο χώρο καταστάσεων και την υπόθεση ότι m_i σωματίδια θα βρίσκονται στο i -στο κελί του πλέγματος, η κατανομή των σωματιδίων στα κελιά κατά την πιθανή αναδειγματοληψία θα είναι πολυωνυμική με εκτιμώμενο διάνυσμα πιθανοτήτων $(m_1 \hat{w}_{t+1}^1, m_2 \hat{w}_{t+1}^2, \dots, m_k \hat{w}_{t+1}^k)$. Τονίζουμε ότι μια τέτοια εκτίμηση (μεταξύ άλλων) χρειάζεται για την απόφαση λήψης μέτρων κατά της φτωχοποίησης. Η εκτίμηση των σφαλμάτων παρατήρησης και των αντίστοιχων βαρών μπορεί να αξιοποιηθεί και σε προβλήματα ελλιπών δεδομένων, όταν η παρατήρηση απουσιάζει και τα σφάλματα παρατήρησης δεν μπορούν να υπολογιστούν. Έτσι, τα βάρη είναι δυνατό να εκτιμηθούν για το τρέχον βήμα απουσία παρατήρησης, όπως παρουσιάζουμε κατωτέρω.

3. ΠΡΟΣΟΜΟΙΩΣΗ

Στην παρούσα ενότητα παρατίθεται προσομοιωτικό παράδειγμα, ως ειδική περίπτωση της θεωρίας, για ελλιπή δεδομένα με μοντέλο της μορφής

$$X_t = \delta \eta \mu(X_{t-1}) + V_t, \quad V_t \square N(0, 0.5)$$

$$Y_t = X_t + U_t, \quad U_t \square N(0, 2)$$

όπου $\eta\mu(x)$ είναι η συνάρτηση ημιτόνου και οι θόρυβοι είναι Γκαουσιανοί μηδενικού μέσου, με διασπορές $\sigma_V^2 = 0.5$ και $\sigma_U^2 = 2$. Θεωρούμε γνωστή την αρχική κρυφή κατάσταση $X_0 = 0$ και ότι η παρατήρηση αποτυγχάνεται να ληφθεί με πιθανότητα $p = .15$. Το φίλτρο, όταν μια παρατήρηση απουσιάζει, δεν χρειάζεται να προβλέψει από το προηγούμενο χρονικό βήμα τα βάρη για το χρονικό βήμα με την απύσχα παρατήρηση, αλλά στο βήμα χωρίς παρατήρηση να παράγει εκ των προτέρων τα σωματίδια και να εκτιμήσει τα πιθανά σφάλματα παρατήρησης. Για τον λόγο αυτό, το βήμα 2β του Αλγορίθμου 1 τροποποιείται ως εξής:

“2β) -Εάν ληφθεί παρατήρηση y_t , στα σφάλματα $\{u_t^i = y_t - x_t^i\}_{i=1}^N$ της εξίσωσης

(4) ανατίθενται βάρη $\{w_t^i = f_{N(0,2)}(u_t^i)\}_{i=1}^N$, όπου f συμβολίζει τη συνάρτηση πυκνότητας πιθανότητας, τα οποία, έπειτα, κανονικοποιούνται.

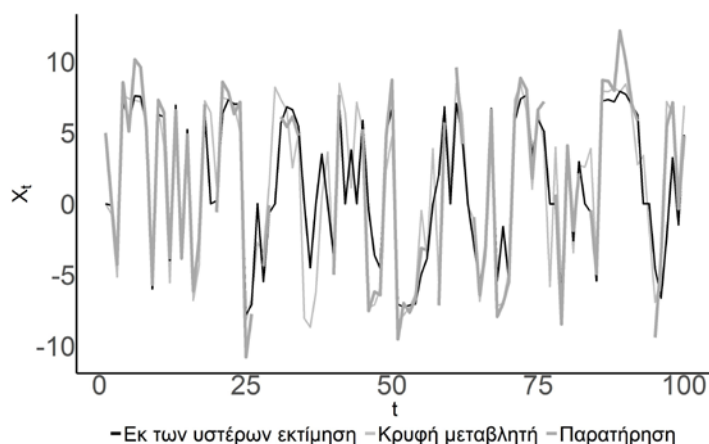
-Διαφορετικά, εκτιμώνται τα σφάλματα παρατήρησης του τρέχοντος βήματος $\{\hat{u}_t^i = 8\eta\mu(\hat{x}_t) - x_{t+1}^i\}_{i=1}^N$ και τα αντίστοιχα βάρη τους $\{\hat{w}_t^i = f_{N(0,2)}(\hat{u}_t^i)\}_{i=1}^N$. Τίθενται $\{w_t^i\}_{i=1}^N = \{\hat{w}_t^i\}_{i=1}^N$ και κανονικοποιούνται.”

Έτσι, με την εκ των προτέρων παραγωγή σωματιδίων στο νέο βήμα, στην ανίσωση (5), η μεταβλητή X_{t+1}^i είναι πλέον γνωστή και δεν χρειάζεται ούτε να αντικατασταθεί ούτε να εκτιμηθεί. Επομένως, οι διαδοχικές αντικαταστάσεις και η απομόνωση των τυχαίων μεταβλητών οδηγούν στην ανίσωση

$$U_{t+1}^i = c_1 V_{t+1} + U_{t+1} + c_1 [f(X_{0:t}) - x_{t+1}^i].$$

Προκύπτει, λοιπόν, για το παρόν σύστημα ότι $c_1 V_{t+1} + U_{t+1} \square N(0, 2.5)$, ώστε για το αναμενόμενο σφάλμα παρατήρησης για ένα σωματίδιο θα ισχύει $U_{t+1}^i \square N(8\eta\mu(\hat{x}_t) - x_{t+1}^i, 2.5)$, όπου \hat{x}_t είναι ο σταθμικός μέσος των εκ των υστέρων εκτιμήσεων. Και έτσι, κατά την ανάθεση βαρών, τα σφάλματα παρατήρησης αντικαθίστανται από την εκτίμηση της αναμενόμενης τιμής τους. Στην Εικόνα 1 παρατίθενται τα αποτελέσματα. Η λεπτή γκρι γραμμή παριστά την κρυφή μεταβλητή, η μαύρη γραμμή τη σημειακή εκτίμηση του φίλτρου με τον σταθμικό μέσο του δείγματος από το φίλτρο και η παχιά γκρι γραμμή τις παρατηρήσεις. Φαίνεται ότι όταν οι παρατηρήσεις απουσιάζουν, το φίλτρο δύναται να εκτιμήσει ικανοποιητικά την κρυφή μεταβλητή, ως εξής: Το μέσο τετραγωνικό σφάλμα για τις εκ των υστέρων εκτιμήσεις προκύπτει να είναι για τις χρονικές στιγμές με διαθέσιμες παρατηρήσεις $MSE_1=5.52$ (με αντίστοιχο τυπικό σφάλμα $SE_1=2.35$) και για τις χρονικές στιγμές με μη διαθέσιμες παρατηρήσεις $MSE_2=26.48$ (με τυπικό σφάλμα $SE_2=5.15$), ενώ, ενδεικτικά, το εύρος τιμών για τις κρυφές μεταβλητές και τις παρατηρήσεις είναι αντίστοιχα $R_1=17.18$ και $R_2=22,96$.

Εικόνα 1. Διάγραμμα χρονοσειρών για τις μεταβλητές προσομοίωσης. Η λεπτή γκρι γραμμή αντιπροσωπεύει τη χρονοσειρά της κρυφής μεταβλητής, η μαύρη γραμμή την εκ των υστέρων σημειακή εκτίμηση και η παχιά γκρι γραμμή τη χρονοσειρά των παρατηρήσεων.



4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα εργασία πραγματοποιείται εκτίμηση των σφαλμάτων παρατήρησης του Φίλτρου Σωματιδίων ένα βήμα έπειτα από το τρέχον βήμα για μια ειδική κατηγορία μοντέλων χώρου καταστάσεων, για μονοδιάστατες μεταβλητές με γραμμική σχέση μεταξύ της μεταβλητής κατάστασης και της παρατήρησης. Όπως καταδεικνύεται από τα παρατιθέμενα παραδείγματα, η εκτίμηση είναι εφικτή, συχνά, μάλιστα, από γνωστούς και αναλυτικούς τύπους για συνήθεις κατανομές. Ακόμη, σε περιπτώσεις εκφυλισμού και φτωχοποίησης είναι χρήσιμη η πρόβλεψη των τιμών των βαρών για την εκτίμηση της πορείας του δείγματος σε επόμενα βήματα. Σύμφωνα με το προσομοιωτικό παράδειγμα, η εκ των προτέρων εκτίμηση των σφαλμάτων παρατήρησης μπορεί να αξιοποιηθεί σε προβλήματα ελλειπών δεδομένων.

5. ΠΑΡΑΡΤΗΜΑ

Καθώς η εύρεση της κατανομής του αθροίσματος $c_1 V_{t+1} + U_{t+1} - c_1 V_{t+1}^i$ δεν οδηγεί πάντοτε σε ευρέως γνωστές κατανομές, παρατίθενται κάποιες από τις συνήθεις περιπτώσεις θορύβων. Χαρακτηριστικά είναι τα παραδείγματα Ομοιόμορφης, Γκαουσιανής και Εκθετικής κατανομής. Διερευνάται η δυνατότητα εύρεσης αναλυτικού τύπου για τη συνάρτηση κατανομής, διότι η ολοκλήρωση της συνάρτησης πυκνότητας σχετίζεται με τον υπολογισμό της μέσης τιμής της μεταβλητής, που αξιοποιείται στον προτεινόμενο αλγόριθμο.

Στην περίπτωση της Ομοιόμορφης κατανομής, αν $U_{t+1} \square U(\alpha, \beta)$ και $c_1 V_{t+1} \square U(\gamma, \delta)$, η συνέλιξη έχει υπολογιστεί (Killman and von Collani, 2001) ως:

$$f_{c_1 V_{t+1} + U_{t+1} - c_1 V_{t+1}^i}(x) = \begin{cases} 0, x \leq z_1 \\ \frac{\sum_{\vec{j} \in B^2} (-1)^{|\vec{j}|+1} K(x - (\gamma - \delta + \alpha) - \langle \vec{j}, (\beta - \alpha, \delta - \gamma) \rangle)}{6(\delta - \gamma)^2(\beta - \alpha)}, z_1 < x < z_2 \\ 0, x \geq z_2 \end{cases}$$

όπου $z_1 = \gamma - \delta - \alpha$ και $z_2 = \delta - \gamma + \beta$, B^2 είναι ο χώρος των διδιάστατων δίτιμων διανυσμάτων $\{0,1\}^2$, το μέτρο $|\vec{j}|$ αντιστοιχεί στο άθροισμα των συνιστωσών του διανύσματος \vec{j} και

$$K(x) = \begin{cases} 0, x < 0 \\ x^2, 0 \leq x < \delta - \gamma \\ x^2 - (x - (\delta - \gamma))^2, \delta - \gamma \leq x \end{cases}.$$

Έτσι, καθώς η συνάρτηση πυκνότητας πιθανότητας είναι πολυωνυμική, προκύπτει αναλυτικός τύπος για τη συνάρτηση κατανομής.

Αντιστοίχως, αν οι θόρυβοι είναι Γκαουσιανοί, ώστε $U_{t+1} \square N(\mu_1, \sigma_1^2)$ και $c_1 V_{t+1} \square N(\mu_2, \sigma_2^2)$, τότε $c_1 V_{t+1} + U_{t+1} - c_1 V_{t+1}^i \square N(\mu_1, \sigma_1^2 + 2\sigma_2^2)$.

Σχετικά με την Εκθετική κατανομή, αν $U_{t+1} \square \text{Exp}(\lambda_1)$ και $|c_1| V_{t+1} \square \text{Exp}(\lambda_2)$, η συνάρτηση πυκνότητας δίνεται από τον τύπο

$$f_{c_1 V_{t+1} + U_{t+1} - c_1 V_{t+1}^i}(x) = \begin{cases} \frac{\lambda_1 \lambda_2}{2(\lambda_1 + \lambda_2)} \exp(\lambda_1 x), x < 0 \\ \frac{\lambda_1 \lambda_2}{2(\lambda_1 - \lambda_2)} \exp(-\lambda_1 x) - \frac{\lambda_1^2 \lambda_2}{2(\lambda_2^2 - \lambda_1^2)} \exp(-\lambda_2 x), 0 \leq x \end{cases}$$

και η συνάρτηση κατανομής μπορεί να υπολογιστεί αναλυτικά.

Τέλος, για το συνδυασμό κανονικής κατανομής με εκθετική, όπου $U_{t+1} \square N(\mu, \sigma^2)$ και $c_1 V_{t+1} \square \text{Exp}(\lambda)$,

$$f_{c_1 V_{t+1} + U_{t+1} - c_1 V_{t+1}^i}(x) = \frac{\lambda}{2} \exp\left(\frac{\sigma^2 \lambda (2x - 2\mu + \sigma^2 \lambda)}{2\sigma^2}\right) F_{N(x - \mu + \sigma^2 \lambda, \sigma^2)}(0) \\ - \frac{\lambda}{2} \exp\left(\frac{-\sigma^2 \lambda (2x - 2\mu - \sigma^2 \lambda)}{2\sigma^2}\right) F_{N(x - \mu - \sigma^2 \lambda, \sigma^2)}(0)$$

όπου $F_{N(\mu, \sigma^2)}$ συμβολίζει τη Γκαουσιανή συνάρτηση κατανομής με παραμέτρους

μ, σ^2 . Η συνάρτηση κατανομής του ζητούμενου αθροίσματος δεν υπολογίζεται αναλυτικά. Ωστόσο, η χρήση αριθμητικών μεθόδων για τον υπολογισμό των τιμών της δεν θα επηρεάσει το υπολογιστικό κόστος του αλγορίθμου, διότι οι παράμετροι είναι σταθερές, ώστε οι τιμές της συνάρτησης μπορούν να υπολογιστούν πριν την

έναρξη του φίλτρου και να αποθηκευτούν σε πίνακα που θα είναι διαθέσιμος κατά τη λειτουργία του φίλτρου για τον υπολογισμό των βαρών.

ABSTRACT

The distribution of observation errors during the implementation of Particle Filter is studied for one-dimensional variables with linear observation equation. The observation errors are estimated for one time step after the current step. Thus, the estimation of weights for the next step can be achieved and the probability that a particle is reproduced during resampling in the next step can be evaluated, so that it is possible for a grid on the state space to be constructed and the distribution of the particles over the cells of the grid to be estimated for one step forward. Simple examples of well-known distributions, as well as a simulating example with missing data are given.

ΑΝΑΦΟΡΕΣ

- Gopaluni R. B. (2008). A particle filter approach to identification of nonlinear processes under missing observations. *The Canadian Journal of Chemical Engineering*, **86**(6).
- Gopaluni R., Schön T. B. and Wills, A. G. (2009). Particle Filter Approach to Nonlinear System Identification under Missing Observations with a Real Application. *IFAC Proceedings Volumes*, **42**(10), 810-815.
- Gordon N., Salmond D., and Smith A. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F Radar and Signal Processing*, **140**(2), 107.
- Killmann F., and Von Collani E. (2001). A Note on the Convolution of the Uniform and Related Distributions and Their Use in Quality Control. *Economic Quality Control*, **16**(1).
- Li T., Sun S., Sattar T. P., & Corchado J. M. (2014). Fight sample degeneracy and impoverishment in particle filters: A review of intelligent approaches. *Expert Systems with Applications*, **41**(8), 3944-3954.
- Zhang X., Khwaja A. S., Luo J., Housfater A. S. and Anpalagan A. (2015). Multiple Imputations Particle Filters: Convergence and Performance Analyses for Nonlinear State Estimation With Missing Data, in *IEEE Journal of Selected Topics in Signal Processing*, **9**(8), 1536-1547.



ΑΠΑΣΧΟΛΗΣΗ ΚΑΙ ΚΑΤΑΘΛΙΨΗ ΣΕ ΑΤΟΜΑ ΗΛΙΚΙΑΣ 50 ΕΤΩΝ ΚΑΙ ΑΝΩ ΣΤΗΝ ΕΛΛΑΔΑ ΚΑΙ ΣΤΗΝ ΕΥΡΩΠΗ ΒΑΣΕΙ ΣΤΟΙΧΕΙΩΝ ΤΟΥ 2^{ΟΥ} ΚΥΜΑΤΟΣ ΤΗΣ ΕΡΕΥΝΑΣ SHARE

Έλενα Μπαμίχα, Γεωργία Βερροπούλου

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς

elenabami93@gmail.com, gverrop@unipi.gr

ΠΕΡΙΛΗΨΗ

Η κατάθλιψη αποτελεί τη σημαντικότερη νόσο της σημερινής εποχής. Ενώ η ύπαρξή της είναι διαδεδομένη από αρχαιοτάτων χρόνων, σύμφωνα με ιστορικές πηγές, σήμερα ταλανίζει εκατομμύρια ανθρώπους σε όλο τον κόσμο, ενώ αποτελεί κύρια αιτία νοσηρότητας στην τρίτη ηλικία επηρεάζοντας την καθημερινή ζωή των ηλικιωμένων και οδηγώντας τους στην περιθωριοποίηση και το στιγματισμό. Ένα κρίριο ζήτημα που προκύπτει αφορά το ποιοι είναι οι παράγοντες εκείνοι που συμβάλλουν στην εμφάνιση της κατάθλιψης. Με γνώμονα λοιπόν το συγκεκριμένο ερώτημα, στην παρούσα έρευνα, μελετάται η συσχέτιση αυτής με ορισμένους επαγγελματικούς/εργασιακούς, δημογραφικούς, καθώς και κοινωνικο-οικονομικούς παράγοντες, αλλά και μεταβλητές νοσηρότητας (δείκτες υγείας) σε άτομα ηλικίας 50 ετών και άνω τόσο στην Ελλάδα όσο και στην Ευρώπη. Η ανάλυση που πραγματοποιείται βασίζεται στο δεύτερο κύμα συλλογής στοιχείων της διεπιστημονικής και διαχρονικής Έρευνας για την Υγεία, τη Γήρανση και τη Συνταξιοδότηση στην Ευρώπη (SHARE), ενώ η στατιστική μέθοδος που ακολουθείται είναι η Δίτιμη Λογιστική Παλινδρόμηση. Τα ευρήματα που προέκυψαν ανέδειξαν ως σημαντικότερους παράγοντες που επηρεάζουν την εμφάνιση κατάθλιψης: το φύλο, την ευκολία ή δυσκολία ανταπόκρισης ενός νοικοκυριού στις οικονομικές του υποχρεώσεις, την αντίληψη που έχει διαμορφωθεί από τους ίδιους τους ερωτώμενους σχετικά με το επίπεδο της σωματικής και ψυχικής τους υγείας και την ικανοποίηση ή όχι που αντλείται από την τρέχουσα απασχόληση.

Λέξεις Κλειδιά: Κατάθλιψη, Εργασιακό άγχος, Σύνδρομο επαγγελματικής εξουθένωσης, Περιγραφική Ανάλυση, Δίτιμη Λογιστική Παλινδρόμηση

1. ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

Ο όρος «κατάθλιψη» εμφανίστηκε για πρώτη φορά το 1660 μ.Χ., ονομασία, η οποία άρχισε να χρησιμοποιείται ευρύτερα τον 19^ο αιώνα. Έως τότε οι ψυχικές διαταραχές αποκαλούνταν «μελαγχολία» (Πολυκανδριώτη και Στεφανίδου, 2013). Η νοσολογική οντότητα της κατάθλιψης ακολουθεί την ανθρωπότητα ήδη από τα πρώιμα στάδια της ύπαρξής της μέχρι και σήμερα. Ειδικά στις μέρες μας, λόγω της οικονομικής κρίσης με την οποία έχουν έρθει αντιμέτωπες οι περισσότερες χώρες του κόσμου, του έντονου ρυθμού ζωής και της απαιτητικής καθημερινότητας, αποτελεί μία πολύ σοβαρή κατάσταση και αφορά ένα μεγάλο μέρος του πληθυσμού. Εκτιμάται ότι δέκα στους εκατό ανθρώπους του γενικού πληθυσμού, αναμένεται να νοσήσουν στη διάρκεια της ζωής τους από κατάθλιψη (Χριστοδούλου, 2005). Αναμφισβήτητα, η εν λόγω ψυχιατρική διαταραχή εμφανίζεται με μεγάλη συχνότητα στην εποχή μας προσελκύνοντας κατά αυτόν τον τρόπο το ενδιαφέρον ολοένα και περισσότερων επιστημών και ερευνών που έχουν ως στόχο την ολοένα και καλύτερη κατανόηση της ίδιας της νόσου, των αιτιών που οδηγούν σε αυτήν, των συμπτωμάτων, αλλά και του αντίκτυπου που αυτή έχει στους διάφορους τομείς της καθημερινής ζωής των πασχόντων.

Ιδιαίτερη έμφαση δίνεται στη μελέτη των παραγόντων εκείνων που συμβάλλουν στην εμφάνιση της κατάθλιψης στην τρίτη ηλικία, αφού η αύξηση του αριθμού των ηλικιωμένων (>65), των υπερηλικών (>90), αλλά και των αιωνόβιων (>100) είναι εξαιρετικά εμφανής τόσο στις βιομηχανικά ανεπτυγμένες όσο και στις υπό ανάπτυξη χώρες. Το 1950, 23 ηλικιωμένοι αντιστοιχούσαν σε 100 νέους, ενώ το 2050 προβλέπεται ότι 110 ηλικιωμένοι θα αντιστοιχούν σε 100 νέους. Επιπροσθέτως, στην περίοδο της τρίτης ηλικίας συμβαίνουν σημαντικές αλλαγές στη ζωή του ατόμου, οι οποίες αποτελούν σημαντικούς παράγοντες κινδύνου για την εκδήλωση κατάθλιψης (Πολίτης, 2011).

Μέσω μίας πληθώρας ερευνών, έχει αποδειχθεί ότι η παρουσία οξείας ή χρόνιας νόσου, η παρουσία αναπηρίας ή ενός κινητικού περιορισμού, καθώς και ορισμένες σωματικές παθήσεις όπως είναι επί παραδείγματι ο υπο/υπερθυρεοειδισμός, η νόσος του Parkinson και η αγγειακή νόσος, μπορούν να πυροδοτήσουν την εμφάνιση καταθλιπτικής διαταραχής στα ηλικιωμένα άτομα προκαλώντας σημαντικές διαταραχές στη φυσιολογική τους ζωή (Γουρνέλλης, 2011). Οι Verropoulou και Tsimbos (2007) μελετώντας τις ψυχικές διαταραχές σε ερωτώμενους άνω των 50 ετών στην Ελλάδα, κατέληξαν στο ότι η κακή σωματική υγεία των ατόμων αυτών συνδέεται στενά με την κακή ψυχική υγεία, τονίζοντας μάλιστα ότι το να νοσεί ένα άτομο από δύο ή περισσότερες χρόνιες ασθένειες, σχεδόν διπλασιάζει τις πιθανότητες να εμφανίσει κατάθλιψη στο μέλλον. Η έρευνα των Nelson et al. (2001) που πραγματοποιήθηκε σε άντρες και γυναίκες καρκινοπαθείς, οι οποίοι βρισκότουσαν σε μονάδα εντατικής θεραπείας αφού είχαν αξιολογηθεί ως άτομα υψηλού κινδύνου για νοσοκομειακό θάνατο, έδειξε ότι το 40% των ατόμων αυτών, ένα αξιοσημείωτα σημαντικό ποσοστό, εμφάνιζε συμπτώματα κατάθλιψης.

Πέραν, όμως, των μεταβλητών υγείας και άλλοι παράγοντες, όπως είναι το φύλο, η οικογενειακή κατάσταση και η ηλικία, η οποία, όμως, αποτελεί ένα αμφιλεγόμενο αίτιο που χρήζει ξεχωριστής αναφοράς, σχετίζονται με την εμφάνιση της καταθλιπτικής νόσου. Έχει βρεθεί ότι, οι γυναίκες έχουν σχεδόν δύο φορές περισσότερες πιθανότητες να υποφέρουν από κατάθλιψη σε σχέση με τους άντρες (Angst et al., 2002), με την κατάθλιψη να εκδηλώνεται μέσω διαφορετικών συμπτωμάτων στα δύο φύλα. Έτσι, η δακρύρροια, η κόπωση και τα προβλήματα ύπνου πρωταγωνιστούν στις γυναίκες, ενώ η ευερεθιστότητα φαίνεται να είναι το μόνο καταθλιπτικό σύμπτωμα που υπερισχύει στους άντρες (Vergopoulou and Tsimbos, 2007).

Επιπλέον, η οικογενειακή κατάσταση φαίνεται να επηρεάζει σε έντονο βαθμό την εμφάνιση της κατάθλιψης ή την αύξηση των καταθλιπτικών συμπτωμάτων, όπως υποστηρίζουν οι περισσότερες μελέτες. Σύμφωνα με την έρευνα των Vergopoulou και Tsimbos (2007) οι χήροι, καθώς και οι διαζευγμένοι, αλλά και τα άτομα που βρίσκονται σε διάσταση έχουν 43% περισσότερες πιθανότητες να εμφανίσουν συμπτώματα κατάθλιψης συγκριτικά με τους παντρεμένους και τα άτομα που δεν βρίσκονται σε σχέση. Ιδίως οι παντρεμένοι φαίνεται να ζουν πολύ καλύτερα από τους αδέσμευτους, καθότι ο γάμος αποτελεί πηγή συναισθηματικής και οικονομικής υποστήριξης και κατά συνέπεια, δρα προστατευτικά έναντι της κατάθλιψης.

Τέλος, έχουν δημοσιευθεί πολλές μελέτες που συμπεραίνουν ότι τα καταθλιπτικά συμπτώματα αυξάνονται με την πάροδο των χρόνων, χωρίς ωστόσο η συσχέτιση μεταξύ ηλικίας και κατάθλιψης να είναι πάντα ξεκάθαρη. Οι Brodaty et al. (1991) παρατήρησαν ότι άτομα μεγαλύτερων ηλικιακών ομάδων έχουν περισσότερες πιθανότητες να εμφανίσουν καταθλιπτικά συμπτώματα συγκριτικά με άτομα μικρότερων ηλικιακών ομάδων. Ωστόσο, η έρευνα των Ζήντρου και συν. (2014), η οποία έλαβε χώρα σε 200 ηλικιωμένα άτομα, έδειξε ότι η ηλικιακή ομάδα 75-84 ετών παρουσιάζει υψηλότερα ποσοστά κατάθλιψης σε σχέση με την ηλικιακή ομάδα 51-74 ετών, αλλά χαμηλότερα ποσοστά κατάθλιψης συγκριτικά με τα άτομα των οποίων οι ηλικίες κυμαίνονται από τα 85 έως τα 98 έτη. Στα ίδια περίπου συμπεράσματα κατέληξε και η έρευνα του Jorm (2000), σύμφωνα με την οποία η πιθανότητα εμφάνισης καταθλιπτικών συμπτωμάτων αυξάνεται μέχρι τα 69 έτη, ενώ αρχίζει να μειώνεται από τα 70 έτη και μετά. Τα αποτελέσματα των προαναφερθέντων ερευνών υποδεικνύουν ότι δεν υπάρχει συνέπεια αποτελεσμάτων μεταξύ κατάθλιψης και ηλικίας.

Οι Vergopoulou και Tsimbos (2007) μέσω της έρευνάς τους ανακάλυψαν ότι τα άτομα ηλικίας 65-74 ετών έχουν 40% περισσότερες πιθανότητες να εμφανίσουν καταθλιπτικά συμπτώματα συγκριτικά με τα άτομα που είναι κάτω των 65 χρόνων, ενώ τα άτομα άνω των 75 ετών φαίνεται να έχουν πολύ πιο έντονη τάση να εμφανίσουν κατάθλιψη σε σχέση με τα άτομα κάτω των 65 ετών. Εντούτοις, υποστήριξαν «Η άποψη αυτή, όμως, μπορεί να είναι ψευδής, δεδομένου ότι τα ηλικιωμένα άτομα, και πιο συγκεκριμένα τα άτομα άνω των 65 ετών, υποφέρουν σε μεγαλύτερο βαθμό, σε σχέση με τις μικρότερες ηλικιακές ομάδες, από αναπηρία,

καθώς και από διάφορα στρεσογόνα γεγονότα της ζωής, όπως είναι παραδείγματος χάριν το πένθος».

Η χηρεία, οι σωματικές παθήσεις, η αναπηρία και τα συστήματα κοινωνικής υποστήριξης που τις περισσότερες φορές υπολειπούνται, αποτελούν γεγονότα, τα οποία είναι όλο και πιο πιθανό να προκύψουν καθώς τα άτομα μεγαλώνουν και γηράσκουν. Με άλλα λόγια, αποτελούν παράγοντες που αντικατοπτρίζουν τις εμπειρίες και τις συνθήκες που επικρατούν σε υψηλό βαθμό σε άτομα μεγάλων ηλικιών (Bruce, 2002). Κατά συνέπεια, τα καταθλιπτικά συμπτώματα, τα οποία αυξάνονται με την πάροδο των χρόνων μπορεί να μην οφείλονται εξολοκλήρου στην ηλικία, αλλά να οφείλονται κυρίως στις συνθήκες που επικρατούν, καθώς και στα γεγονότα με τα οποία έρχονται αντιμέτωπα τα άτομα της τρίτης ηλικίας. Εκτός, όμως, από τη συγκεκριμένη άποψη, επικρατεί και μία άλλη άποψη κατά την οποία η μειωμένη συναισθηματική ανταπόκριση, ο αυξημένος συναισθηματικός έλεγχος και η ψυχολογική ανοσοποίηση απέναντι σε αγχωτικές εμπειρίες αποτελούν παράγοντες που συμβάλλουν στην εμφάνιση χαμηλότερων ποσοστών κατάθλιψης σε άτομα μεγαλύτερων ηλικιακών ομάδων (Jorm, 2000).

Σχετικά με την απασχόληση, έχει διαπιστωθεί ότι το επαγγελματικό περιβάλλον, και κατ' επέκτασιν η εργασία στο σύγχρονο κόσμο, αποτελούν μία από τις σημαντικότερες πηγές άγχους. Οι σχέσεις που αναπτύσσει το άτομο μέσα στο εργασιακό του περιβάλλον, η ίδια η φύση της εργασίας και τα ενδεχόμενα προβλήματα που αντιμετωπίζει σε αυτή, επηρεάζουν σε σημαντικό βαθμό την καθημερινότητα και την ποιότητα της ζωής του (Στάγια και Ιορδάνης, 2014), χωρίς ωστόσο να δύναται ο εργαζόμενος να αντιμετωπίσει αποτελεσματικά τις συγκεκριμένες καταστάσεις, καθότι συνήθως δεν υπάρχουν πολλές πράξεις στις οποίες μπορεί να προβεί προκειμένου να τροποποιήσει το περιβάλλον εργασίας του, με αποτέλεσμα να συνεχίζει να λειτουργεί υπό στρεσογόνες συνθήκες (Iacovidis et al., 2003) και να υποφέρει, σε ορισμένες περιπτώσεις, από το σύνδρομο της επαγγελματικής εξουθένωσης (job burnout). Πιο συγκεκριμένα, ως επαγγελματική εξουθένωση θεωρείται μία μορφή χρόνιου και παρατεταμένου επαγγελματικού άγχους, κατά τη διάρκεια του οποίου το άτομο προοδευτικά αποδυναμώνεται, ενώ ταυτόχρονα αισθάνεται ότι τα ψυχικά αποθέματα που διαθέτει δεν επαρκούν για να αντιμετωπίσει τους ιδιαίτερα έντονους και πειστικούς παράγοντες της εργασίας του. Αυτή η μακροχρόνια έκθεση και η εμπλοκή σε ιδιαίτερα απαιτητικές καταστάσεις συναισθηματικής συμμετοχής μπορεί να προκαλέσει φαινόμενα σωματικής, συναισθηματικής, αλλά και ψυχικής κόπωσης έχοντας αρνητικό αντίκτυπο όχι μόνο στην άσκηση της εργασίας, αλλά και στην προσωπική και την κοινωνική ζωή των εργαζομένων (Δημητρόπουλος και Φιλίππου, 2008).

Όπως έχει προκύψει από διάφορες έρευνες, η επαγγελματική εξουθένωση μπορεί να οδηγήσει στην εμφάνιση καταθλιπτικής νόσου. Η πιθανότητα αυτή, το να εμφανίσει, δηλαδή, ένα άτομο κατάθλιψη αυξάνεται βάσει της σοβαρότητας του συνδρόμου της επαγγελματικής εξουθένωσης που το εν λόγω άτομο υφίσταται. Μελέτη που έλαβε χώρα στη Φιλανδία σε 3276 εργαζόμενους ηλικίας 30-64 ετών, έδειξε ότι οι μισοί

περίπου από τους συμμετέχοντες που εμφάνιζαν σοβαρά συμπτώματα επαγγελματικής εξουθένωσης παρουσίαζαν παράλληλα και καταθλιπτική διαταραχή (Ahola et al., 2005). Κλείνοντας, ιδιαίτερο ενδιαφέρον παρουσιάζει η έρευνα των Middeldorp et al. (2005) σύμφωνα με την οποία προέκυψε ότι το να είναι ένα άτομο άνεργο, σε ορισμένες περιπτώσεις, αποδεικνύεται ότι είναι λιγότερο επιβλαβές, από το να είναι ένας δυσαρεστημένος και πιεσμένος εργαζόμενος, ο οποίος υφίσταται επαγγελματική εξουθένωση.

2. ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ – ΠΕΡΙΓΡΑΦΙΚΗ ΑΝΑΛΥΣΗ

2.1 Έρευνα για την Υγεία, τη Γήρανση και τη Συνταξιοδότηση στην Ευρώπη - Survey of Health, Ageing and Retirement in Europe (SHARE) - Δεδομένα

Η Έρευνα για την Υγεία, τη Γήρανση και τη Συνταξιοδότηση στην Ευρώπη (SHARE), αποτελεί ένα μείζον επιστημονικό εγχείρημα, το οποίο εξετάζει την υγεία (Health, Health Care), τις οικογενειακές και κοινωνικές σχέσεις (Social Networks), τις διαδικασίες αποχώρησης από την εργασία, το εισόδημα, την περιουσία και την ευημερία (Economics) των ατόμων ηλικίας 50 ετών και άνω σε 28 διαφορετικές χώρες και πιο συγκεκριμένα, σε 27 Ευρωπαϊκές χώρες, όπως και στο Ισραήλ (Börsch-Supan et al., 2013). Μέχρι στιγμής, στις έρευνες που έχουν πραγματοποιηθεί από το SHARE έχουν συμμετάσχει παραπάνω από 140000 άτομα, ενώ οι συνεντεύξεις που τα εν λόγω άτομα έχουν παραχωρήσει αγγίζουν έως και τις 380000 (SHARE, 2016). Οι συγκεκριμένες συνεντεύξεις είναι διαχρονικές και λαμβάνουν χώρα κάθε δύο χρόνια, αναδεικνύοντας κατά αυτόν τον τρόπο τη διαγενεακή – μακροσκοπική διάσταση των ερευνών του SHARE, αφού παρακολουθούν τα ίδια άτομα σε βάθος χρόνου, καθώς αυτά αποσύρονται από την εργασία τους, συνταξιοδοτούνται και γηράσκουν (Börsch-Supan et al., 2013).

Το SHARE περιλαμβάνει συνολικά 7 κύματα δεδομένων. Για τους σκοπούς της παρούσας μελέτης, το δείγμα που επιλέχθηκε αναφέρεται σε διακρατικά δεδομένα από το δεύτερο κύμα της έρευνας που πραγματοποιήθηκε το 2006-2007. Αναλυτικότερα, το δείγμα που μελετάται, περιλαμβάνει 34734 συνολικά ερωτώμενους από 14 διαφορετικές Ευρωπαϊκές χώρες. Η επιλογή των ερωτώμενων από τις Ευρωπαϊκές χώρες πραγματοποιήθηκε μέσω διαφορετικών δειγματοληπτικών σχεδίων που κρίνονται αντιπροσωπευτικά για την κάθε χώρα ξεχωριστά, αφού η βάση δεδομένων του SHARE παρέχει βάρη για το σχεδιασμό δειγματοληψίας προκειμένου οι εκτιμητές των παραμέτρων του πληθυσμού να είναι αμερόληπτοι. Πρόκειται, ουσιαστικά, για μία ισορροπημένη εκπροσώπηση των διαφόρων περιοχών της Ευρώπης, από τη Σκανδιναβία/Βόρεια Ευρώπη (Σουηδία, Δανία και Ιρλανδία) και την Κεντρική Ευρώπη (Αυστρία, Γαλλία, Γερμανία, Ελβετία, Βέλγιο και Ολλανδία) έως την Ανατολική Ευρώπη (Τσεχία και Πολωνία) και τη Μεσόγειο/Νότια Ευρώπη (Ισπανία, Ιταλία και Ελλάδα).

2.2 Αποκριτική (εξαρτημένη) μεταβλητή

Η έρευνα SHARE με σκοπό την αξιολόγηση της συναισθηματικής υγείας των ερωτηθέντων χρησιμοποιεί τη μεταβλητή EURO-D (Euro Depression Scale), η οποία αποτελεί την Ευρωπαϊκή κλίμακα μέτρησης της κατάθλιψης και περιλαμβάνει τα ακόλουθα 12 καταθλιπτικά συμπτώματα: θλίψη ή καταθλιπτική διάθεση, πεσιμιστική διάθεση, τάσεις αυτοκτονίας, ενοχές, αϋπνία ή υπερυπνία, έλλειψη ενδιαφέροντος, ευερεθιστότητα, μειωμένη ή αυξημένη όρεξη για κατανάλωση φαγητού, προβλήματα κόπωσης, έλλειψη συγκέντρωσης, έλλειψη διάθεσης για διασκέδαση και δακρύρροια (Börsch-Supan and Jürges, 2005). Τα προαναφερθέντα συμπτώματα προέκυψαν μέσα από μία σειρά ερωτήσεων που τέθηκαν στους συμμετέχοντες της έρευνας. Πρόκειται, ειδικότερα, για 21 ερωτήσεις στις οποίες καλούνται να απαντήσουν οι ερωτώμενοι με σκοπό να διευκρινίσουν το πώς αισθάνονται για πράγματα και καταστάσεις που συμβαίνουν γύρω τους.

Για τις ανάγκες της συγκεκριμένης μελέτης, η αποκριτική μεταβλητή EURO-D κωδικοποιήθηκε λαμβάνοντας δυαδική μορφή, έτσι ώστε να αντανακλά την εμφάνιση ή μη κατάθλιψης στους ερωτηθέντες. Η κωδικοποίηση αυτή έχει ως εξής:

- "0" \approx οι ερωτώμενοι παρουσιάζουν από 0 έως και 3 καταθλιπτικά συμπτώματα, χωρίς να εμφανίζουν κατά αυτόν τον τρόπο κατάθλιψη
- "1" \approx οι ερωτώμενοι παρουσιάζουν από 4 και άνω καταθλιπτικά συμπτώματα, πάσχοντας έτσι από κατάθλιψη.

Ως κατηγορία αναφοράς έχει οριστεί η πρώτη κατά σειρά, κατηγορία που λαμβάνει την τιμή 1 και της οποίας οι ερωτηθέντες νοσούν από κατάθλιψη, κατηγορία δηλαδή που χαρακτηρίζεται από την ύπαρξη (επιτυχία) του συμβάντος που ερευνάται.

Μέσω της νέας μεταβλητής που κατασκευάστηκε, προκύπτει ότι από τους 34734 (100.0%) συμμετέχοντες, η συντριπτική τους πλειοψηφία και πιο συγκεκριμένα, οι 26160 (75.3%) δεν πάσχουν από κατάθλιψη, αφού εμφανίζουν λιγότερα από 4 καταθλιπτικά συμπτώματα, ενώ μόλις οι 8574 (24.7%) φαίνεται να νοσούν από καταθλιπτική διαταραχή, καθώς εμφανίζουν από 4 και άνω καταθλιπτικά συμπτώματα. Διαφορετικά, η πρώτη κατηγορία, η κατηγορία δηλαδή, η οποία αποτελείται από άτομα που παρουσιάζουν από 0 έως και 3 συμπτώματα κατάθλιψης, έχει με διαφορά τη μεγαλύτερη συχνότητα.

2.3 Επεξηγηματικές (ανεξάρτητες) μεταβλητές

Οι ανεξάρτητες μεταβλητές περιλαμβάνουν επεξηγηματικούς παράγοντες, οι οποίοι αναφέρονται σε δημογραφικά, οικονομικά, επαγγελματικά/εργασιακά στοιχεία, αλλά και σε δείκτες υγείας. Παρακάτω, αναφέρονται ονομαστικά οι επεξηγηματικές μεταβλητές που χρησιμοποιήθηκαν και αναλύονται οι κατηγορίες από τις οποίες αποτελούνται οι εν λόγω μεταβλητές.

Πίνακας 1. Οι επεξηγηματικές (ανεξάρτητες) μεταβλητές

Δημογραφικά Χαρακτηριστικά	Περιορισμοί στις δευτερεύουσες καθημερινές δραστηριότητες
Χώρα διαμονής	Αριθμός χρόνιων παθήσεων
Φύλο	Αριθμός κινητικών περιορισμών
Ηλικία	Αριθμός σωματικών συμπτωμάτων
Οικογενειακή κατάσταση	Επαγγελματικοί/Εργασιακοί Παράγοντες
Οικονομική Μεταβλητή	Ικανοποίηση με το τρέχον επάγγελμα
Ευκολία ή δυσκολία ανταπόκρισης ενός νοικοκυριού στα έξοδά του	Διαρκής πίεση χρόνου λόγω αυξημένου φόρτου εργασίας
Δείκτες Υγείας	Επιθυμία πρόωρης συνταξιοδότησης
Αυτοαντίληψη για το επίπεδο της συνολικής υγείας	Φόβος σχετικά με το εάν η υγεία των ερωτώμενων θα περιορίσει την ικανότητά τους να εργάζονται στο τρέχον επαγγελματικό περιβάλλον πριν την κανονική συνταξιοδότηση

1) Δημογραφικά Χαρακτηριστικά: Οι δημογραφικές μεταβλητές περιλαμβάνουν βασικές δημογραφικές πληροφορίες που αφορούν το δείγμα και σχετίζονται με

- **την χώρα διαμονής (country):** όπως αναφέρθηκε στην παράγραφο 2.1, οι χώρες που περιλαμβάνονται στο δεύτερο κύμα του SHARE είναι η Αυστρία, η Γερμανία, η Σουηδία, η Ολλανδία, η Ισπανία, η Ιταλία, η Γαλλία, η Δανία, η Ελλάδα, η Ελβετία, το Βέλγιο, η Τσεχία, η Πολωνία και η Ιρλανδία. Τα δείγματα που συλλέχθηκαν από τις συγκεκριμένες χώρες είναι σχεδόν ισομερή μεταξύ τους.
- **το φύλο (gender):** οι άντρες που συμμετέχουν στην έρευνα του δεύτερου κύματος SHARE ανέρχονται στους 15359 (44.2%), ενώ ο αριθμός των γυναικών ισούται με 19375 (55.8%). Το ποσοστό των γυναικών δηλαδή είναι υψηλότερο κατά 11.6% από αυτό των αντρών, με αποτέλεσμα το γυναικείο φύλο να υπερισχύει.
- **την ηλικία (age):** υπενθυμίζεται ότι οι ηλικίες των ατόμων που συμμετέχουν στη συγκεκριμένη εργασία τοποθετούνται στα 50 έτη και άνω.
- **την οικογενειακή κατάσταση (mstat/marital status):** η μεταβλητή αυτή περιλαμβάνει τις εξής κατηγορίες: έγγαμος/η και ζώντας μαζί με την/τον σύζυγο, συμβίωση σε μόνιμη βάση με την/τον σύντροφο, παντρεμένος/η, ζώντας χωριστά από την/τον σύζυγο (σε διάσταση), άγαμος/η, διαζευγμένος/η και χήρος/α. Ωστόσο, στα πλαίσια της ανάλυσης οι κατηγορίες όπου περιλαμβάνονται τα άτομα που συμβιώνουν σε μόνιμη βάση με τους/τις συντρόφους τους και τα άτομα που είναι παντρεμένα, αλλά ζουν χωριστά από τους/τις συζύγους τους συγχωνεύθηκαν με τις κατηγορίες στις οποίες ανήκουν όσοι/ες είναι παντρεμένοι και ζουν με τις/τους συζύγους τους και όσοι/ες είναι διαζευγμένοι/ες, αντίστοιχα.

Ο λόγος που οδήγησε στο να επιτελεστεί μία τέτοια συγχώνευση είναι οι μικρές συχνότητες που αντιστοιχούν στις κατηγορίες «συμβίωση σε μόνιμη βάση με τον/την σύντροφο» και «παντρεμένος/η, ζώντας χωριστά από την/τον σύζυγο». Έτσι, η μεταβλητή της οικογενειακής κατάστασης, ενώ αρχικά, αποτελούνταν από 6 διαφορετικές κατηγορίες, κατέληξε να αποτελείται από 4. Η διαδικασία της ενοποίησης των κατηγοριών με μικρές συχνότητες είναι συνήθης διαδικασία που συχνά επιλέγεται, αφού βελτιώνει την αξιοπιστία της ανάλυσης (Linacre, 2002).

Από το σύνολο του δείγματος, το υψηλότερο ποσοστό που ισούται με 72,2% είναι οι παντρεμένοι/ες που ζουν με την/τον σύζυγό τους και όσοι/ες συμβιώνουν σε μόνιμη βάση με τις/τους συντρόφους τους, ενώ οι άγαμοι/ες αποτελούν μόλις το 5,1%, όντας το χαμηλότερο ποσοστό της μεταβλητής «οικογενειακή κατάσταση». Το ποσοστό των χήρων είναι ίσο με 14,9%.

2) Οικονομική Μεταβλητή: Οι οικονομικές μεταβλητές χρησιμοποιούνται προκειμένου να περιγράψουν το βιοτικό επίπεδο των ερωτηθέντων. Στο μοντέλο δίτιμης λογιστικής παλινδρόμησης που παρουσιάζεται στην παράγραφο 3, περιλαμβάνεται μόνο μία οικονομική μεταβλητή, ονόματι *fdistress* (*household able to make ends meet*), η οποία διακρίνεται σε 4 κατηγορίες και χρησιμοποιείται προκειμένου να περιγράψει εάν τα νοικοκυριά που συμμετέχουν στην έρευνα μπορούν να ανταπεξέλθουν με ευκολία ή όχι στις οικονομικές τους υποχρεώσεις.

Οι κατηγορίες όπου διακρίνεται η μεταβλητή *fdistress* είναι οι εξής: (1) στην πρώτη κατηγορία ανήκουν τα νοικοκυριά που δυσκολεύονται σε σημαντικό βαθμό να ανταπεξέλθουν στις οικονομικές τους υποχρεώσεις, (2) στη δεύτερη κατηγορία κατατάσσονται τα νοικοκυριά τα οποία επίσης δυσκολεύονται να ανταπεξέλθουν στις οικονομικές τους υποχρεώσεις, αλλά σε μικρότερο βαθμό συγκριτικά με εκείνον με τον οποίο έρχονται αντιμέτωπα τα νοικοκυριά της πρώτης κατηγορίας, (3) η τρίτη κατηγορία περιλαμβάνει τις οικογένειες που μπορούν να τα βγάλουν πέρα σχετικά εύκολα, όσον αφορά τα οικονομικά τους ζητήματα και (4) η τέταρτη κατηγορία αποτελείται από τις οικογένειες που μπορούν να τα βγάλουν πέρα εύκολα όσον αφορά τα οικονομικά τους καθήκοντα.

Από τα 34734 νοικοκυριά που συμμετέχουν στο δεύτερο κύμα του SHARE, στην ερώτηση που τους τέθηκε σχετικά με το πόσο εύκολα ή δύσκολα ανταπεξέρχονται στα οικονομικά τους καθήκοντα, απάντησαν 34500 νοικοκυριά, προκύπτοντας κατά αυτόν τον τρόπο 234 (0.7%) ελλείπουσες τιμές - νοικοκυριά. Από τα 34500 νοικοκυριά, τα 11559 (33.3%) απάντησαν ότι καταφέρνουν με σχετική ευκολία να τα βγάλουν πέρα οικονομικά, ενώ τα 3912 (11.3%) δήλωσαν ότι ανταπεξέρχονται με μεγάλη δυσκολία στις οικονομικές τους υποχρεώσεις. Σημειώνεται ότι όσον αφορά τις ελλείπουσες τιμές που προέκυψαν, δεν χρησιμοποιήθηκε κάποια μέθοδος υποκατάστασής τους (*imputation method*).

3) Δείκτες Υγείας: Σκοπός των μεταβλητών υγείας είναι η συλλογή πληροφοριών σχετικών με την ψυχική υγεία των ερωτώμενων, καθώς και η μελέτη της σωματικής τους ικανότητας και λειτουργίας. Οι μεταβλητές που περιλαμβάνονται σε αυτήν την κατηγορία αναλύονται ακολούθως.

- **Αυτοπροσδιοριζόμενη υγεία (*sphus/self-perceived health*):** περιγράφει την υποκειμενική αντίληψη των ερωτηθέντων για το επίπεδο της ψυχικής και σωματικής τους υγείας μέσω 5 κατηγοριών: (1) άριστο, (2) πολύ καλό, (3) καλό, (4) μέτριο και (5) κακό.

Η πλειοψηφία των ερωτώμενων (36.9%) φαίνεται ότι κρίνει το επίπεδο της συνολικής της υγείας της ως καλό, ενώ το 24.5% θεωρεί ότι το επίπεδο της υγείας του είναι μέτριο. Μόνο το 9.2% κρίνει ότι η υγεία του είναι άριστη.

- **Περιορισμοί στις δευτερεύουσες καθημερινές δραστηριότητες (iادل/limitations with instrumental activities of daily living):** ως δευτερεύουσες καθημερινές δραστηριότητες χαρακτηρίζονται: η χρήση χάρτου για εύρεση προσανατολισμού σε ένα άγνωστο μέρος, η προετοιμασία ενός ζεστού γεύματος, τα ψώνια από το μπακάλικο, η χρήση τηλεφώνου, η λήψη φαρμάκων, οι δουλειές στο σπίτι ή στον κήπο, η διαχείριση χρημάτων, όπως είναι για παράδειγμα η πληρωμή λογαριασμών.

- **Χρόνιες παθήσεις (chronic):** στα χρόνια ή μακράς διάρκειας προβλήματα υγείας που είναι πιθανό να εμφανίζουν κάποιοι από τους ερωτηθέντες περιλαμβάνονται: (1) το έμφραγμα, η καρδιακή προσβολή, η θρόμβωση στεφανιαίων ή άλλη καρδιακή νόσος, όπως η χρόνια καρδιακή ανεπάρκεια, (2) η υψηλή (αρτηριακή) πίεση αίματος, (3) τα υψηλά επίπεδα χοληστερόλης στο αίμα, (4) το εγκεφαλικό επεισόδιο ή άλλη νόσος των αγγείων του εγκεφάλου, (5) ο σακχαρώδης διαβήτης, (6) η χρόνια βρογχίτιδα ή το εμφύσημα, (7) το άσθμα, (8) η αρθρίτιδα, (9) η οστεοπόρωση, (10) ο καρκίνος, (11) το έλκος στομάχου, (12) η νόσος του Parkinson, (13) ο καταρράκτης, (14) το κάταγμα ισχίου ή μηρού, (15) άλλα κατάγματα, (16) το Αλτσχάιμερ, (17) κάποιος καλοήθης όγκος. Η συγκεκριμένη μεταβλητή αποτελείται από 2 κατηγορίες. Στη μία κατηγορία κατατάσσονται τα άτομα που εμφανίζουν λιγότερες από 2 χρόνιες παθήσεις, ενώ στην άλλη εκείνα που εμφανίζουν από 2 και άνω χρόνιες παθήσεις.

Μέσω της ανάλυσης προκύπτει ότι το 56.4% των ερωτηθέντων πάσχει από μία ή από καμία χρόνια πάθηση, με το υπόλοιπο 43.6% να ταλαιπωρείται από 2 ή περισσότερες μακράς διάρκειας ασθένειες. Τα συγκεκριμένα αποτελέσματα υποδεικνύουν ότι οι χρόνιες παθήσεις είναι ήδη αυξημένες απειλώντας κατά αυτόν τον τρόπο, σε σημαντικό βαθμό, την υγεία του δείγματος.

- **Κινητικοί περιορισμοί (mobility):** οι κινητικές δυσκολίες και περιορισμοί με τους οποίους έρχονται αντιμέτωποι ορισμένοι από τους ερωτηθέντες εντοπίζονται στα εξής: (1) περπάτημα 100 μέτρων, (2) ορθοστασία για περίπου δύο ώρες, (3) σήκωμα από καρέκλα, αφού το άτομο ήταν καθισμένο σε αυτή για μεγάλες χρονικές περιόδους, (4) ανέβασμα αρκετών σειρών σκαλοπατιών χωρίς ανάπαυση, (5) ανέβασμα μίας σειράς σκαλοπατιών χωρίς ανάπαυση, (6) σκύψιμο ή γονάτισμα, (7) έκταση των χεριών πάνω από το επίπεδο των ώμων, (8) τράβηγμα ή σπρώξιμο μεγάλων αντικειμένων, (9) σήκωμα ή σπρώξιμο βαρών άνω των 5 κιλών, (10) το να πιάσει ένα άτομο ένα μικρό νόμισμα από το τραπέζι. Πρόκειται για δίτιμη μεταβλητή, η οποία στην πρώτη της κατηγορία περιλαμβάνει τους ερωτηθέντες που αντιμετωπίζουν λιγότερους από 2 κινητικούς περιορισμούς, ενώ στη δεύτερη περιλαμβάνει τους ερωτηθέντες που παρουσιάζουν από 2 και πάνω κινητικούς περιορισμούς. Οι ερωτηθέντες που εμφανίζουν από 2 και πάνω κινητικούς περιορισμούς αγγίζουν τους 11634 (33.5%), ενώ ο αριθμός των

ερωτηθέντων που εμφανίζουν το πολύ 1 κινητική δυσκολία ισούται με 23100 (66.5%).

- **Σωματικά συμπτώματα (symptoms):** στην κατηγορία αυτή δηλώνεται εάν ο ερωτώμενος εμφανίζει λιγότερα από 2 ή από 2 και άνω φυσικά συμπτώματα με χρονική διάρκεια τουλάχιστον έξι μηνών από έναν κατάλογο με 12 συμπτώματα που περιλαμβάνει: (1) πόνο στη μέση, στα γόνατα, στα ισχία ή σε οποιαδήποτε άλλη άρθρωση, (2) καρδιακό πρόβλημα ή στηθάγχη, θωρακικό πόνο κατά την άσκηση, (3) δύσπνοια, (4) επίμονο βήχα, (5) πρησμένα πόδια, (6) προβλήματα ύπνου, (7) πτώσεις ή πεσίματα, (8) φόβο για πτώσεις, (9) ζάλη, λιποθυμίες, (10) προβλήματα στομάχου ή εντέρου, (11) ακράτεια ούρων, (12) κόπωση. Οι 20045 (57.7%) ερωτώμενοι από τους 34734, εμφανίζουν λιγότερα από 2 φυσικά συμπτώματα, ενώ οι υπόλοιποι 14689 (42.3%) ερωτώμενοι παρουσιάζουν από 2 και πάνω σωματικά συμπτώματα.

4) Επαγγελματικοί/Εργασιακοί Παράγοντες: Μέσω των εργασιακών παραγόντων παρέχονται πληροφορίες σχετικές με τις τρέχουσες εργασιακές δραστηριότητες των συμμετεχόντων. Στην κατηγορία αυτή, ανήκουν οι μεταβλητές που ακολουθούν: (1) **ο βαθμός ικανοποίησης με το τρέχον επάγγελμα (satisfied/satisfied with main job)**, ο οποίος διακρίνεται σε 4 κατηγορίες: σημαντικός, ικανοποιητικός, μέτριος, ανύπαρκτος, (2) **η διαρκής πίεση χρόνου λόγω αυξημένου φόρτου εργασίας (timepress/time pressure due to a heavy workload)**, που αποτελείται από 4 κατηγορίες: σημαντικά αυξημένη, αυξημένη, μη σημαντικά αυξημένη, καθόλου αυξημένη, (3) **η ύπαρξη ή μη επιθυμίας για πρόωρη συνταξιοδότηση (earlyretirement/ look for early retirement in main job)** και (4) **ο φόβος ή όχι σχετικά με το εάν η υγεία των ερωτώμενων θα περιορίσει την ικανότητά τους να εργάζονται στο τρέχον επαγγελματικό περιβάλλον πριν την κανονική συνταξιοδότηση (limwork/afraid health limits ability to work before regular retirement in main job)**.

Από το σύνολο των ερωτηθέντων, από τα 34734, δηλαδή, άτομα, μόνο τα 9916 (28.5%) εργάζονται, με την πλειοψηφία του δείγματος που ισούται με 17231 (49.6%) άτομα να έχει συνταξιοδοτηθεί. Από τους 9916 εργαζόμενους, οι 5361 δηλώνουν ικανοποιημένοι με την εργασία τους, ενώ μόλις 186 άτομα (0.5%) δηλώνουν ότι δεν είναι καθόλου ικανοποιημένα με την τρέχουσα επαγγελματική τους κατάσταση.

3. ΑΠΟΤΕΛΕΣΜΑΤΑ ΤΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Για τη διερεύνηση της σχέσης μεταξύ της αποκριτικής μεταβλητής και των επεξηγηματικών μεταβλητών έγινε χρήση του ακόλουθου μοντέλου δίτιμης λογιστικής παλινδρόμησης,

$$\log(\text{odds}) = \log\left(\frac{p(x)}{1-p(x)}\right) = a + \beta_1 * \text{country} + \beta_2 * \text{gender} + \beta_3 * \text{age} + \beta_4 * \text{mstat} + \beta_5 * \text{fdistress} + \beta_6 * \text{sphus} + \beta_7 * \text{iadl} + \beta_8 * \text{chronic} + \beta_9 * \text{mobility} + \beta_{10} * \text{symptoms} + \beta_{11} * \text{satisfied} + \beta_{12} * \text{timepress} + \beta_{13} * \text{earlyretirement} + \beta_{14} * \text{limwork} \quad (1)$$

όπου $odds = p(x) / 1 - p(x)$ με το $p(x)$ να εκφράζει την πιθανότητα τους συμβάντος του γεγονότος, την πιθανότητα δηλαδή να εμφανίσει κάποιο άτομο κατάθλιψη (επιτυχία) και το $1 - p(x)$ να εκφράζει την πιθανότητα να μην συμβεί το γεγονός που εξετάζεται, να μην εμφανίσει δηλαδή κάποιο άτομο κατάθλιψη (αποτυχία). Με άλλα λόγια, το $odds$ ορίζεται ως ο λόγος (ή η αναλογία) των συμπληρωματικών πιθανοτήτων. Τονίζεται ότι, το μοντέλο (*I*) εφαρμόστηκε σε μέγεθος δείγματος $N = 9916$ και περιλαμβάνει συνολικά 14 επεξηγηματικές μεταβλητές στις οποίες αναφερθήκαμε αναλυτικά στην παράγραφο 2.3, ενώ οι υπολογισμοί πραγματοποιήθηκαν χρησιμοποιώντας το SPSS, έκδοση 23.

Ο τρόπος επιλογής της κατηγορίας αναφοράς στην εκάστοτε περίπτωση, εάν δηλαδή θα ορίζεται ως κατηγορία αναφοράς η πρώτη ή η τελευταία κατηγορία από τις ήδη υπάρχουσες κατηγορίες μίας μεταβλητής, εξαρτάται από τη συχνότητα των παρατηρήσεων που διαθέτει η κάθε κατηγορία. Κατά κανόνα, αποφεύγεται να τίθεται ως κατηγορία αναφοράς, κατηγορία με μικρές συχνότητες. Στον πίνακα που ακολουθεί, παρατίθενται οι κατηγορίες αναφοράς για την καθεμία μεταβλητή.

Πίνακας 2. Κατηγορίες αναφοράς επεξηγηματικών μεταβλητών

<i>Επεξηγηματικές Μεταβλητές</i>	<i>Κατηγορίες Αναφοράς</i>
<i>Χώρα διαμονής</i>	<i>Αυστρία</i>
<i>Φύλο</i>	<i>Γυναίκες</i>
<i>Ηλικία</i>	-
<i>Οικογενειακή κατάσταση</i>	<i>Έγγαμος/η και ζώντας μαζί με την/τον σύζυγο & Συμβίωση σε μόνιμη βάση με τον/την σύντροφο</i>
<i>Ευκολία ή δυσκολία ανταπόκρισης ενός νοικοκυριού στα έξοδά του</i>	<i>Το νοικοκυριό ανταπεξέρχεται εύκολα στα έξοδά του</i>
<i>Αυτοαντίληψη για το επίπεδο της συνολικής υγείας</i>	<i>Κακό επίπεδο συνολικής υγείας</i>
<i>Περιορισμοί στις δευτερεύουσες καθημερινές δραστηριότητες</i>	-
<i>Αριθμός χρόνιων παθήσεων</i>	<i>Λιγότερες από 2 χρόνιες παθήσεις</i>
<i>Αριθμός κινητικών περιορισμών</i>	<i>Λιγότεροι από 2 κινητικοί περιορισμοί</i>
<i>Αριθμός σωματικών συμπτωμάτων</i>	<i>Λιγότερα από 2 σωματικά συμπτώματα</i>
<i>Ικανοποίηση με το τρέχον επάγγελμα</i>	<i>Άτομα ικανοποιημένα σε σημαντικό βαθμό</i>
<i>Διαρκής πίεση χρόνου λόγω αυξημένου φόρτου εργασίας</i>	<i>Άτομα που βρίσκονται σε μεγάλο βαθμό υπό διαρκή πίεση χρόνου</i>
<i>Επιθυμία πρόωρης συνταξιοδότησης</i>	<i>Άτομα που δεν θέλουν να συνταξιοδοτηθούν πρόωρα</i>
<i>Φόβος σχετικά με το εάν η υγεία των ερωτώμενων θα περιορίσει την ικανότητά τους να εργάζονται στο τρέχον επαγγελματικό περιβάλλον πριν την κανονική συνταξιοδότηση</i>	<i>Άτομα που δεν φοβούνται ότι η υγεία τους θα περιορίσει την ικανότητά τους να εργάζονται στο τρέχον επαγγελματικό περιβάλλον πριν την κανονική συνταξιοδότηση</i>

Οι *p*-τιμές των ελέγχων που διεξήχθησαν συγκρίθηκαν με επίπεδο σημαντικότητας που λαμβάνει την τιμή 0.05. Αν για μία μεταβλητή του μοντέλου (*I*) ισχύει «*p*-τιμή

του ελέγχου της μεταβλητής < 0.05 » τότε συμπεραίνουμε ότι «η μεταβλητή είναι στατιστικά σημαντική σε επίπεδο σημαντικότητας 5%». Η σημαντικότητα και η αξιοπιστία της προσαρμογής του μοντέλου (I) ελέγχθηκε με βάση τον Πίνακα Ταξινόμησης, ο οποίος υποδεικνύει το ποσοστό των περιπτώσεων της εξαρτώμενης μεταβλητής που προβλέπονται σωστά από το μοντέλο. Στη συγκεκριμένη περίπτωση, το ποσοστό του Πίνακα Ταξινόμησης που προκύπτει, ισούται με 85,4%, ποσοστό που μπορεί να κριθεί ως αρκετά ικανοποιητικό, καθιστώντας το μοντέλο μας αξιόπιστο.

Πίνακας 3. Εκτιμήσεις συντελεστών των δημογραφικών, οικονομικών μεταβλητών και δεικτών υγείας

	B	Sig	Exp(B)		B	Sig	Exp(B)
ΧΩΡΑ ΔΙΑΜΟΝΗΣ							
Γερμανία	,074	,784	1,077	Άγαμος/η	,255	,041	1,290
Σουηδία	,273	,303	1,314	Χήρος/α	,593	,000	1,810
Ολλανδία	,366	,174	1,441	ΤΟ ΝΟΙΚΟΚΥΡΙΟ ΑΝΤΑΠΕΞΕΡΧΕΤΑΙ			
Ισπανία	,270	,335	1,310	Με μεγάλη δυσκολία	,724	,000	2,063
Ιταλία	,896	,001	2,449	Με σχετική δυσκολία	,385	,000	1,470
Γαλλία	,973	,000	2,646	Με σχετική ευκολία	,285	,001	1,329
Λανία	,610	,020	1,840	ΑΥΤΟΠΡΟΣΔΙΟΡΙΖΟΜΕΝΟ ΕΠΙΠΕΔΟ ΣΩΜΑΤΙΚΗΣ ΚΑΙ ΨΥΧΙΚΗΣ ΥΓΕΙΑΣ			
Ελλάδα	-,306	,264	,737	Άριστο επίπεδο	-1,644	,000	,193
Ελβετία	,565	,042	1,760	Πολύ καλό επίπεδο	-1,404	,000	,246
Βέλγιο	,835	,001	2,305	Καλό επίπεδο	-1,087	,000	,337
Τσεχία	-,231	,395	,793	Μέτριο επίπεδο	-,622	,001	,537
Πολωνία	1,018	,000	2,769	ΠΕΡΙΟΡΙΣΜΟΙ ΣΤΙΣ ΔΕΥΤΕΡΕΥΟΥΣΕΣ ΚΑΘΗΜΕΡΙΝΕΣ ΔΡΑΣΤΗΡΙΟΤΗΤΕΣ	,329	,002	1,390
Ιρλανδία	,416	,167	1,516	ΑΡΙΘΜΟΣ ΧΡΟΝΙΩΝ ΠΑΘΗΣΕΩΝ			
ΦΥΛΟ				2 και άνω χρόνιας παθήσεις	,044	,565	1,045
Άντρες	-,832	,000	,435	ΑΡΙΘΜΟΣ ΚΙΝΗΤΙΚΩΝ ΠΕΡΙΟΡΙΣΜΩΝ			
ΗΛΙΚΙΑ				2 και άνω κινητικοί περιορισμοί	,145	,113	1,156
ΟΙΚΟΓΕΝΕΙΑΚΗ ΚΑΤΑΣΤΑΣΗ				ΑΡΙΘΜΟΣ ΣΩΜΑΤΙΚΩΝ ΣΥΜΠΤΩΜΑΤΩΝ			
Παντρεμένος/η, ζώντας χωριστά από την/τον σύντροφο & Διαζευγμένος/η	,013	,901	1,013	2 και άνω σωματικά συμπτώματα	1,050	,000	2,858

Σκιαγραφώντας τα κυριότερα ευρήματα της ανάλυσης, σημειώνονται τα ακόλουθα. Σχετικά με την χώρα διαμονής, οι κάτοικοι της Πολωνίας έχουν έως και 3, περίπου, φορές υψηλότερη πιθανότητα να εμφανίσουν καταθλιπτικά συμπτώματα συγκριτικά με τους κατοίκους της Αυστρίας, ενώ οι κάτοικοι της Ελλάδας φαίνεται ότι δεν

διαφέρουν στατιστικά σημαντικά σε επίπεδο σημαντικότητας 5%, δεν διαφοροποιούνται δηλαδή σε σημαντικό βαθμό, ως προς την εμφάνιση κατάθλιψης συγκριτικά με τους κατοίκους της Αυστρίας. Ακολούθως, το φύλο, η ηλικία και η οικογενειακή κατάσταση φαίνεται να επηρεάζουν την εμφάνιση κατάθλιψης. Πιο συγκεκριμένα, οι άντρες έχουν κατά 56,5% μειωμένη την πιθανότητα να νοσήσουν από κατάθλιψη, σε σχέση με τις γυναίκες, ενώ όσον αφορά την ηλικία, διαπιστώνεται ότι αύξηση της ηλικίας των ερωτηθέντων κατά ένα έτος οδηγεί σε μείωση της πιθανότητας να εμφανίσουν, οι ερωτηθέντες, κατάθλιψη. Επιπροσθέτως, προκύπτει ότι οι χήροι/ες διαθέτουν 2 σχεδόν φορές υψηλότερη πιθανότητα εμφάνισης καταθλιπτικής διαταραχής συγκριτικά με τους έγγαμους/ες που ζουν μαζί με τις/τους συζύγους τους και όσους/ες συμβιώνουν σε μόνιμη βάση τις/τους συντρόφους τους. Επιπλέον, οι μεταβλητές «το νοικοκυριό ανταπεξέρχεται», «περιορισμοί στις δευτερεύουσες καθημερινές δραστηριότητες» και «αριθμός σωματικών συμπτωμάτων» έχουν θετική επίδραση στην παρουσία κατάθλιψης υπό την έννοια ότι όσο πιο δύσκολα εκπληρώνει ένα νοικοκυριό τα οικονομικά του καθήκοντα, όσο περισσότεροι είναι οι περιορισμοί που ένα άτομο αντιμετωπίζει στις δευτερεύουσες καθημερινές του δραστηριότητες και η εμφάνιση από 2 και πάνω φυσικών συμπτωμάτων, αντίστοιχα, οδηγούν σε αύξηση των πιθανοτήτων εμφάνισης κατάθλιψης. Αντιθέτως, η μεταβλητή «αυτοπροσδιοριζόμενο επίπεδο σωματικής και ψυχικής υγείας» συμβάλλει αρνητικά στην εμφάνιση κατάθλιψης, καθώς όσο πιο θετική είναι η αντίληψη που έχουν διαμορφώσει οι ερωτηθέντες για το επίπεδο της συνολικής τους υγείας, τόσο χαμηλότερη είναι η πιθανότητα να νοσήσουν από κατάθλιψη.

Πίνακας 4. Εκτιμήσεις συντελεστών εργασιακών μεταβλητών

	B	Sig	Exp(B)		B	Sig	Exp(B)
ΒΑΘΜΟΣ ΙΚΑΝΟΠΟΙΗΣΗΣ ΑΠΟ ΤΟ ΤΡΕΧΟΝ ΕΠΑΓΓΕΛΜΑ				Μη σημαντικά αυξημένη	-.466	.000	.627
Ικανοποιητικός	.226	.002	1,254	Καθόλου αυξημένη	-.316	.012	.729
Μέτριος	.663	.000	1,940	ΕΠΙΘΥΜΙΑ ΠΡΟΩΡΗΣ ΣΥΝΤΑΞΙΟΔΟΤΗΣΗΣ			
Ανύπαρκτος	.652	.002	1,919	Μη επιθυμία πρόωρης συνταξιοδότησης	.152	.026	1,165
ΔΙΑΡΚΗΣ ΠΙΕΣΗ ΧΡΟΝΟΥ ΛΟΓΩ ΑΥΞΗΜΕΝΟΥ ΦΟΡΤΟΥ ΕΡΓΑΣΙΑΣ				ΦΟΒΟΣ ΣΧΕΤΙΚΟΣ ΜΕ ΤΟ ΕΝΔΕΧΟΜΕΝΟ Η ΥΓΕΙΑ ΝΑ ΠΕΡΙΟΡΙΣΕΙ ΤΗΝ ΙΚΑΝΟΤΗΤΑ ΑΠΑΣΧΟΛΗΣΗΣ ΠΡΙΝ ΤΗΝ ΚΑΝΟΝΙΚΗ ΣΥΝΤΑΞΙΟΣΟΤΗΣΗ			
Αυξημένη	-.291	.001	.748	Άτομα που φοβούνται ότι η υγεία τους θα περιορίσει την ικανότητά τους να εργάζονται στο τρέχον επαγγελματικό περιβάλλον πριν την κανονική συνταξιοδότηση	.355	.000	1,426

Τέλος, παρατηρώντας τα αποτελέσματα που παρουσιάζονται κατόπιν ανάλυσης των εργασιακών μεταβλητών, προκύπτει ότι όσα άτομα είναι όλο και λιγότερο ικανοποιημένα από το τρέχον εργασιακό τους περιβάλλον, επιθυμούν να συνταξιοδοτηθούν πρόωρα ή/και φοβούνται ότι η υγεία τους θα περιορίσει την ικανότητά τους να εργάζονται, πριν την κανονική συνταξιοδότηση, έχουν υψηλότερες πιθανότητες να υποφέρουν από καταθλιπτική νόσο. Ενδεικτικά, τα άτομα που δεν αντλούν καμία ικανοποίηση από την τρέχουσα επαγγελματική τους κατάσταση έχουν κατά 91,9% αυξημένη την πιθανότητα να εμφανίσουν κατάθλιψη σε σύγκριση με τα άτομα που είναι ικανοποιημένα σε σημαντικό βαθμό από το επάγγελμά τους. Αντίθετα, όσο λιγότερο πιέζονται, οι ερωτηθέντες, από τον χρόνο λόγω αυξημένου φόρτου εργασίας, τόσο μικρότερη είναι η πιθανότητα να νοσήσουν από κατάθλιψη.

4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Μέσω της επεξεργασίας των στοιχείων του δεύτερου κύματος της έρευνας SHARE, δόθηκαν κάποιες καίριες απαντήσεις στο βασικό ερώτημα της παρούσας εργασίας: κατά πόσον η κατάθλιψη στα άτομα ηλικίας 50 ετών και άνω που διαμένουν στην Ελλάδα και την Ευρώπη συσχετίζεται με την απασχόληση, με εργασιακούς δηλαδή παράγοντες, καθώς και με δημογραφικούς, αλλά και κοινωνικο-οικονομικούς παράγοντες, όπως επίσης και με παράγοντες νοσηρότητας.

Με μία πρώτη ματιά, διαπιστώνεται ότι τα περισσότερα άτομα του δείγματος αισθάνονται υγιή αφού κρίνουν ότι το επίπεδο τόσο της σωματικής όσο και της ψυχικής τους υγείας είναι ικανοποιητικό, γεγονός που σημαίνει ότι δεν αντιμετωπίζουν κάποιο σοβαρό χρόνιο νόσημα, κινητικό περιορισμό ή ψυχική ασθένεια που να παρεμποδίζει την αυτονομία και τη λειτουργικότητά τους. Πράγματι, όσον αφορά στην ψυχική υγεία των ατόμων ηλικίας 50 ετών και άνω φαίνεται ότι ένα συντριπτικά μεγάλο ποσοστό δεν πάσχει από κλινική κατάθλιψη, ενώ σχετικά με τη σωματική υγεία, το μεγαλύτερο ποσοστό δήλωσε ότι αντιμετωπίζει μόνο ένα ή κανένα χρόνιο νόσημα ή/και κινητικό περιορισμό, έχοντας έτσι την ικανότητα να αυτοεξυπηρετείται σε δευτερεύουσες καθημερινές δραστηριότητες. Όσον αφορά στην επαγγελματική κατάσταση των ερωτηθέντων, οι περισσότεροι από αυτούς έχουν συνταξιοδοτηθεί, ενώ η πλειονότητα των ατόμων που εργάζονται δείχνει ότι είναι ικανοποιημένη με το τρέχον επάγγελμα που ασκεί.

Κατόπιν περαιτέρω ανάλυσης και εφαρμογής δίτιμης λογιστικής παλινδρόμησης τα ευρήματα που προέκυψαν ανέδειξαν ως σημαντικότερους παράγοντες που επηρεάζουν την εμφάνιση κατάθλιψης: το φύλο, την ευκολία ή δυσκολία ανταπόκρισης ενός νοικοκυριού στις οικονομικές του υποχρεώσεις, την αντίληψη που έχει διαμορφωθεί από τους ίδιους τους ερωτώμενους σχετικά με το επίπεδο της σωματικής και ψυχικής τους υγείας και την ικανοποίηση ή όχι που αντλείται από την τρέχουσα απασχόληση.

ABSTRACT

Depression is the most important mental illness in this day and age. While its existence has been widespread since ancient times, according to historical sources, today millions of people worldwide are being afflicted with depressive disorders. In addition, it is known that depression tends to be a major cause of morbidity in the elderly, affecting their daily lives and leading them to marginalization and stigmatization. An important enough question which arises, concerns the factors that contribute to the emerge of depression. Taking into consideration this particular question, the aim of this paper is to investigate the association of certain occupational and demographic factors, socio-economic status and health related variables with depression, among people who are 50 years or more across Greece and Europe. The analysis is based on data of the second wave of the interdisciplinary and longitudinal Survey of Health, Ageing and Retirement in Europe (SHARE) and was performed using binary logistic regression models. The findings which emerged, showed that the most important factors affecting the occurrence of depression are: gender, ease or difficulty of a household to make ends meet, the individual perception of the respondents about the level of their physical and mental health and the satisfaction that a person draws from his working environment.

ΑΝΑΦΟΡΕΣ

- Γουρνέλλης, Ρ. (2011). Κατάθλιψη στους Ηλικιωμένους: Σύγχρονα Δεδομένα. *Εγκέφαλος*, **48**, 141-142.
- Δημητρόπουλος, Χ. και Φίλιππου, Ν. (2008). Η επαγγελματική εξουθένωση στο χώρο της υγείας. *Αρχαία Ελληνικής Ιατρικής*, **25**(5), 642-647.
- Ζήντρου, Γ., Βασιλοπούλου, Γ., Βακάλη, Χ., Γερογιάννη, Γ., Καμπά, Ε. και Μπαμπάτσικου, Φ. (2014). Συγκριτική μελέτη εμφάνισης κατάθλιψης στην τρίτη ηλικία σε κλειστή και ανοιχτή δομή φροντίδας ηλικιωμένων με την κλίμακα Hamilton. *Περιοδικό το βήμα του Ασκληπιού*, **13**(2), 144-159.
- Πολίτης, Α.Μ. (2011). Ψυχική υγεία στην τρίτη ηλικία, στο Κολίσης, Φ., Παπαδημητρίου, Γ., Λύκουρας, Ε., Δουζένης, Α., Θελερίτης, Χ., Κάττουλας, Ε., Γκαράνη-Παπαδάτου, Τ., Πολίτης, Α., Σκαλτσή, Π., Οικονόμου-Λαλιώτη, Μ., Καρύδη, Μ.Β., Χειλάκος, Ι., Νομίδου, Ν.Ε., Σκόρδος, Λ., Τερζίδου, Μ., Τσακλακίδου, Δ., Λιάππας, Ι.Α., Κοκκέβη, Α., Τζινιέρη-Κοκκώση, Μ., Μανουσέλης, Σ., Σάλλα, Τ. και Λεωνίδου-Σακαλή, Ε.Α., *Ψυχική Υγεία: Σύγχρονες Προσεγγίσεις - Προβληματισμοί*, 71-74. Αθήνα: Εθνικό Ίδρυμα Ερευνών.
- Πολυκανδριώτη, Μ. και Στεφανίδου, Σ. (2013). Κατάθλιψη σε μη ψυχιατρικούς ασθενείς. *Περιοδικό το βήμα του Ασκληπιού*, **12**(4), 397-408.
- Στάγια, Δ. και Ιορδανίδης, Γ. (2014). Το επαγγελματικό άγχος και η επαγγελματική εξουθένωση των εκπαιδευτικών στην εποχή της οικονομική κρίσης. *Επιστημονική Επετηρίδα Τμήματος Νηπιαγωγών Πανεπιστημίου Ιωαννίνων*, **7**, 56-82.
- Χριστοδούλου, Γ.Ν. (2005). Η κλινική έκφραση των συναισθηματικών διαταραχών, στο *Κοινωνία και υγεία IV*, 231-237. Αθήνα: Εθνικό Ίδρυμα Ερευνών.
- Ahola, K., Honkonen, T., Isometsä, E., Kalimo, R., Nykyri, E., Aromaa, A. and Lönnqvist, J. (2005). The relationship between job-related burnout and depressive

- disorders – results from the Finnish Health 2000 Study. *Journal of Affective Disorders*, **88**(1), 55-62.
- Angst, J., Gamma, A., Gastpar, M., Lupine, J.-P., Mendlewicz, J. and Tylee, A. (2002). Gender differences in depression: Epidemiological findings from the European DEPRES I and II studies. *European Archives of Psychiatry and Clinical Neuroscience*, **252**, 201-209.
- Börsch-Supan, A. and Jurges, H. (2005). *The Survey of Health, Ageing and Retirement in Europe. Methodology*. Mannheim, Germany.
- Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmayer, J., Malter, F., Schaan, B., Stuck, S. and Zuber, S. (2013). Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology*, **42**(4), 992-1001.
- Brodsky, H., Peters, K., Boyce, P., Hickie, I., Parker, G., Mitchell, P. and Wilhelm, K. (1991). Age and depression. *Journal of Affective Disorders*, **23**(3), 137-149.
- Bruce, M.L. (2002). Psychological risk factors for depressive disorders in late life. *Biological Psychiatry*, **52**(3), 175-184.
- Iacovides, A., Fountoulakis, K.N., Kaprinis, St. and Kaprinis, G. (2003). The relationship between job stress, burnout and clinical depression. *Journal of Affective Disorders*, **75**, 209-221.
- Jorm, A.F. (2000). Does old age reduce risk of anxiety and depression? A review of epidemiological studies across the adult life span. *Psychological Medicine*, **30**(1), 11-22.
- Linacre, J.M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measures*, **3**(1), 85-106.
- Middeldorp, C.M., Cath, D.C. and Boomsma, D.I. (2005). A twin-family study of the association between employment, burnout and anxious depression. *Journal of Affective Disorders*, **90**(2-3), 163-169.
- Nelson, J.E., Meier, D.E., Oer, E.J., Nierman, D.M., Senzel, R.S. Manfredi, P.L., Davis, S.M. and Morrison, R.S. (2001). Self-reported symptom experience of critically ill cancer patients receiving intensive care. *Critical Care Magazine*, **29**(2), 277-282.
- SHARE (2016). *The Survey of Health, Ageing and Retirement in Europe (SHARE)*. <http://www.share-project.org/home0.html>
- Verropoulou, G. and Tsimbos, C. (2007). Socio-demographic and health-related factors affecting depression of the Greek population in later life: an analysis using SHARE data. *European Journal of Ageing*, **4**(3), 171-181.



ΣΤΑΤΙΣΤΙΚΗ ΜΕ ΤΗΝ ΡΥΘΟΝ

Μουσιάδης Πολυχρόνης

Ομ. Καθηγητής ΑΠΘ

cmoi@math.auth.gr

ΠΕΡΙΛΗΨΗ

Η γλώσσα προγραμματισμού Python είναι μια γενική γλώσσα προγραμματισμού που μπορεί να αντιμετωπίσει με αξιοπιστία προβλήματα ανάλυσης και οπτικοποίησης από όλα τα πεδία της Επιστήμης, συμπεριλαμβανομένης της Στατιστικής. Διατίθεται δωρεάν στο διαδίκτυο, όπου υπάρχουν πολλά εγχειρίδια με οδηγίες και αρκετά βιβλία με εφαρμογές σε διάφορους τομείς της επιστήμης. Με τη βοήθεια μιας δικτυακής πλατφόρμας της Jupyter η εργασία με την Python γίνεται πολύ εύκολη, αποθηκεύεται και μεταβιβάζεται. Περιγράφουμε εδώ μερικές από τις βιβλιοθήκες που εστιάζονται σε στατιστικές αναλύσεις και γραφικές παραστάσεις και δίνουμε κάποια παραδείγματα.

ΕΙΣΑΓΩΓΙΚΑ-ΙΣΤΟΡΙΚΑ

Τα τελευταία 10-20 χρόνια η γλώσσα R, αναπτύχθηκε πάρα πολύ κατακτώντας την πρώτη θέση για την ανάλυση στατιστικών δεδομένων και την παρουσίασή τους. Ένας από τους λόγους, ίσως ο κυριότερος, είναι ότι η γλώσσα αυτή είναι ανοικτού κώδικα και δωρεάν και επιτρέπει στους εξειδικευμένους χρήστες να αναπτύξουν και να προωθούν δικές τους βιβλιοθήκες βοηθώντας τους ερευνητές απανταχού της γης. Έτσι σιγά-σιγά πολλοί χρήστες γνωστών εμπορικών πακέτων όπως τα SPSS, STATA, S-Plus, Matlab, Minitab και άλλων, στράφηκαν στην R, ώστε σήμερα οι περισσότεροι ερευνητές της Στατιστικής γνωρίζουν και χρησιμοποιούν την R για τις μελέτες τους. Η R οφείλει το όνομά της στα αρχικά των δύο ερευνητών Ross Ihaka και Robert Gentleman που την ανέπτυξαν από το 1992 έως το 2000, αλλά και σ' ένα λογοπαίγνιο με την S, την οποία εξέλιξε και αντικατέστησε.

Τελευταία, εμφανίστηκε στο προσκήνιο, ανταγωνιστικά ή/και συναγωνιστικά η γλώσσα Python, η οποία όπως θα δούμε μπορεί να κάνει και στατιστικές αναλύσεις και να οπτικοποιήσει δεδομένα εξίσου καλά με την R. Η Python αναπτύχθηκε κυρίως από τον Van Rossum (μετά το 1989) και πήρε το όνομά της από μια τηλεοπτική σειρά του BBC «Monty Python's Flying Circus». Η έκδοση 2 εμφανίστηκε το 2000 και η 3 το 2008. Εδώ θα ασχοληθούμε με την Python 3.6.

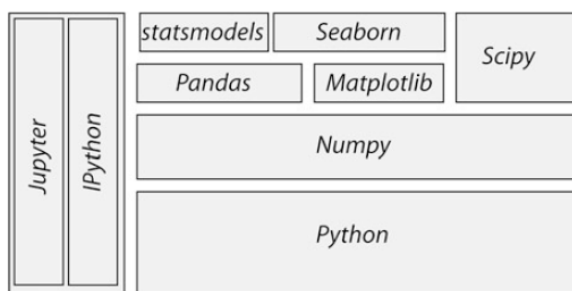
Πολλές σελίδες στο διαδίκτυο εξετάζουν τις διαφορές και ομοιότητες των δύο προγραμμάτων και προτείνουν άλλοτε το ένα και άλλοτε το άλλο για χρήση από τους Στατιστικούς. Μάλιστα, πολλές σελίδες θέτουν το δίλημμα: “R ή Python?”. Η

απάντηση θα μπορούσε να είναι «και R και Python», δεδομένου ότι η Python είναι γενική γλώσσα προγραμματισμού, η οποία εμπλουτισμένη από ένα πλήθος βιβλιοθηκών που διατίθενται δωρεάν στο διαδίκτυο μπορεί να κάνει τα πάντα, ακόμη και στατιστική. Αυτό, έχει μια ιδιαίτερη σημασία, κυρίως όταν ενδιαφερόμαστε να κάνουμε στατιστική ανάλυση ή παρουσίαση σε δεδομένα που δεν τα έχουμε, αλλά πρέπει να τα αντλήσουμε από κείμενα, ιστοσελίδες ή/και βάσεις δεδομένων. Τέτοια είναι η περίπτωση π.χ. της εξόρυξης δεδομένων ή της ανάλυσης μεγάλων δεδομένων (big data).

Η Python έχει αρκετά κοινά με την R, όπως για παράδειγμα ότι είναι ανοικτού κώδικα, έχει πολύ καλά βοηθήματα και οδηγίες στο δίκτυο, έχει βιβλιοθήκες που εμπλουτίζονται από τους χρήστες και κάνει πολύ καλές γραφικές παραστάσεις. Έχει, όμως, ένα πολύ αυστηρό τρόπο γραφής των κωδίκων της και απαιτεί, νομίζω, περισσότερο χρόνο για την εκμάθησή της. Θεωρώ όμως ότι τα βασικά της γλώσσας είναι γνωστά, και αν όχι, οι κωδικές της κατανοούνται από όποιον έχει μια σχετική εξοικείωση με γλώσσες προγραμματισμού.

ΒΙΒΛΙΟΘΗΚΕΣ ΤΗΣ PYTHON ΓΙΑ ΣΤΑΤΙΣΤΙΚΗ

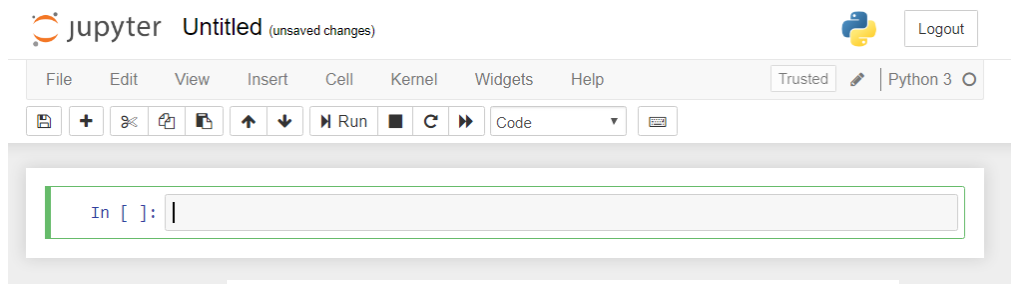
Οι βασικές βιβλιοθήκες της Python που χρειάζονται για τη Στατιστική φαίνονται στην Εικ.1. Η χρήση αυτών και η λειτουργία τους περιγράφεται σε διάφορους ιστοτόπους στο διαδίκτυο, αλλά και σε σχετικά εγχειρίδια. Ένα πολύ καλό βιβλίο για το θέμα αυτό είναι του Thomas Haslwanter του 2016. Στο βιβλίο αυτό εκτός από παραδείγματα εφαρμογής σε διάφορα προβλήματα, δίνονται και οδηγίες για το κατέβασμα των διάφορων βιβλιοθηκών για διάφορα λειτουργικά συστήματα. Επίσης ένα εισαγωγικό βιβλίο για τις δυνατότητες της Python είναι το βιβλίο του Johansson Robert του 2015.



Εικόνα 1: Δομή των βασικών βιβλιοθηκών της Python

Χωρίς να είναι υποχρεωτική, η χρήση του Jupyter κάνει την λειτουργία της Python πολύ ευκολότερη. Πρόκειται για μία εφαρμογή, που χρησιμοποιεί τους συνήθεις φυλλομετρητές (browsers), με τους οποίους πλοηγούμαστε στο Internet, για να κατασκευάσει μία πλατφόρμα-κονσόλα, στην οποία μπορούμε να γράφουμε κείμενα και να εκτελούμε εντολές και στο τέλος να αποθηκεύουμε την εργασία μας ως notebook, με επέκταση `ipynb`. Αυτό το αρχείο αποθηκεύει μόνο τις εντολές και τα κείμενα, σαν ένα script αρχείο, σε απλή μορφή κειμένου και μπορούμε στη συνέχεια να το εκτελέσουμε ή να το αποστείλουμε με ηλεκτρονικό μήνυμα, ή να το ανεβάσουμε στο δίκτυο για χρήση από τους ενδιαφερομένους.

Η αρχική σελίδα του Jupyter έχει τη μορφή της Εικ. 2, και είναι πολύ εύκολο στη χρήση. Φαίνονται οι συνήθειες σελίδες (File, Edit, κλπ), ο τίτλος, το ότι το ανοίξαμε για Python 3 και μία σειρά χρήσιμων κουμπιών. Επίσης, ανοίγει μια περιοχή εισαγωγής εντολών ή κειμένων. Αν στη θέση αυτή γράψουμε μία εντολή π.χ.



Εικόνα 2: Αρχική σελίδα του Jupyter

`plot(x, y)`, και εκτελέσουμε πατώντας το Run θα πάρουμε τη ζητούμενη γραφική παράσταση. Απαραίτητη προϋπόθεση, εκτός από τον ορισμό των απαραίτητων μεταβλητών, είναι να φορτώσουμε τις απαραίτητες βιβλιοθήκες.

Η πρώτη εντολή που τρέχουμε στο Jupyter αν πρόκειται να κάνουμε γραφικές παραστάσεις είναι η:

```
% matplotlib inline
```

Στη συνέχεια ακολουθούν οι βιβλιοθήκες που θα χρειαστούμε ή στοιχεία από βιβλιοθήκες. Για να κάνουμε χρήση μαθηματικών αλλά και αρκετών στατιστικών συναρτήσεων χρειαζόμαστε τη βιβλιοθήκη `numpy` την οποία εισάγουμε, συνήθως, με την εντολή:

```
import numpy as np
```

οπότε, για παράδειγμα, `np.sum(v)` μας δίνει το άθροισμα των στοιχείων του `v`, `np.sin(v)` μας δίνει τα ημίτονα των στοιχείων του `v` τα οποία θεωρούμε ότι μετρώνται σε ακτίνια, ενώ αν είναι σε μοίρες το `np.sin(v*np.pi/180)` μας δίνει τα ημίτονα των στοιχείων του. Δηλαδή, είτε καλούμε γνωστές συναρτήσεις είτε γνωστές σταθερές πρέπει να θέτουμε τη συντομογραφία του `numpy`.

Για να κάνουμε γραφικές παραστάσεις πρέπει να εισάγουμε τη βιβλιοθήκη `matplotlib` που συνήθως εισάγεται ως `plt`. Δίνουμε δηλαδή την εντολή:

```
import matplotlib.pyplot as plt
```

Αυτή η βιβλιοθήκη είναι πολύ πλούσια και μπορεί να κάνει πάρα πολλές γραφικές παραστάσεις, όπως θα φανεί και στα παραδείγματα που ακολουθούν.

ΠΑΡΑΔΕΙΓΜΑΤΑ

Παράδειγμα 1. Περιγραφική Στατιστική.

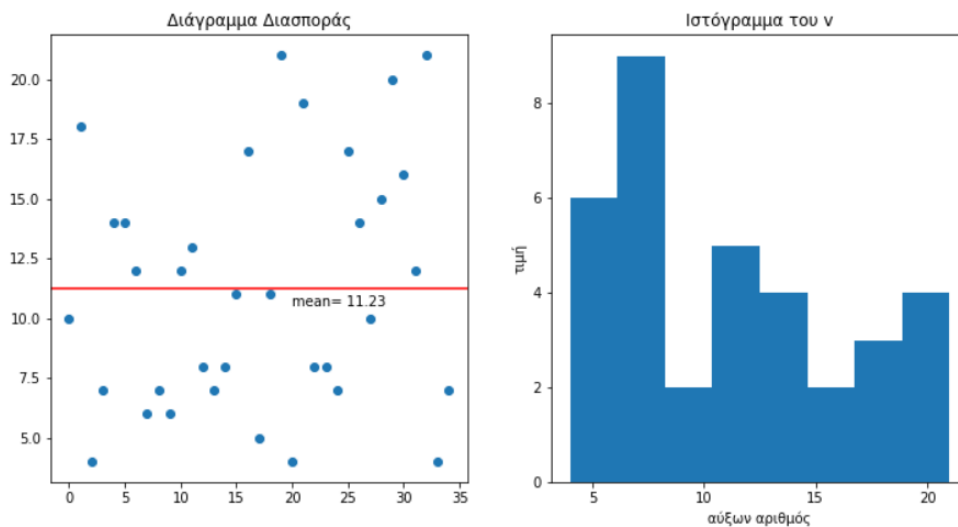
Θα σχεδιάσουμε ένα διάγραμμα διασποράς και ένα ιστόγραμμα για δοθέν σύνολο τιμών, σε ένα σχήμα. Θα μορφοποιήσουμε τα γραφήματα στολίζοντας τα με ορισμένα συνήθη στοιχεία, όπως για παράδειγμα τη μέση τιμή.

Δίνουμε τον παρακάτω κώδικα και εξηγούμε κάποιες από τις εντολές του.

Κώδικας του παραδείγματος 1.

```
% matplotlib inline
import numpy as np
import matplotlib.pyplot as plt
v = [10,18,4,7,14,14,12,6,7,6,12,13,8,7,8,11,17,5,11,21,4,
     19,8,8,7,17,14,10,15,20,16,12,21,4,7]
fig, axes = plt.subplots(1, 2, figsize=(12, 6)) # (1)
axes[0].scatter(range(len(v)),v) # (2)
axes[0].axhline(y=np.mean(v), color='r', linestyle='-') # (3)
axes[0].text(20,10.5, r'mean= %5.2f' % np.mean(v)) # (4)
axes[0].set_title("Διάγραμμα Διασποράς") # (5)
axes[1].hist(v, bins=8) # (6)
axes[1].set_xticks([5,10,15,20]) # (7)
axes[1].set_xlabel("αύξων αριθμός") # (8)
axes[1].set_ylabel("τιμή") # (9)
axes[1].set_title("Ιστόγραμμα του v") # (10)
```

Η εκτέλεση του κώδικα δίνει την Εικ. 3.



Εικόνα 3: Εκτέλεση του κώδικα του Παραδ.1

Οι τρεις πρώτες εντολές του κώδικα είναι οι βιβλιοθήκες που απαιτούνται, ενώ η τέταρτη ορίζει ένα σύνολο δεδομένων ως λίστα. Η επόμενη εντολή η (1) καθορίζει

ότι το γράφημα θα έχει δύο υπογραφήματα με συγκεκριμένες διαστάσεις. Η Python έχει μεγάλη ευελιξία στον ορισμό υπογραφημάτων μέσα σε γράφημα με μεγάλη ποικιλία στην τοποθέτησή τους. Οι εντολές (2)-(5) αφορούν το πρώτο υπογράφημα, ενώ οι υπόλοιπες το δεύτερο υπογράφημα. Οι εντολές (2) και (6) καθορίζουν τα δύο γραφήματα, ότι δηλαδή το ένα είναι διάγραμμα διασποράς και το άλλο ιστόγραμμα. Οι εντολές (5), (8), (9) και (10) προσθέτουν τίτλους στο γράφημα ή σε άξονες. Η εντολή (3) σχεδιάζει μια οριζόντια κόκκινη γραμμή που δηλώνει τη θέση της μέσης τιμής. Προσέξτε ότι η μέση τιμή υπολογίζεται μέσα στην εντολή σχηματισμού του γραφήματος, φυσικά με αναφορά στη βιβλιοθήκη `numpy` ή `np` που περιέχει τη συνάρτηση `mean()`. Τέλος, η εντολή (4) τυπώνει κείμενο στο γράφημα με την πολύ αποδοτική μέθοδο της Python να τυπώνει την τιμή μεταβλητών μέσα σε κείμενα με συγκεκριμένο `format`. Εδώ η ένδειξη `%5.2f` σημαίνει ότι η τιμή της `np.mean(v)` που βρίσκεται μετά το σύμβολο `%` θα τυπωθεί ως πραγματικός αριθμός με 2 δεκαδικά ψηφία σε 5 θέσεις.

Η σύνταξη των εντολών μπορεί να βρεθεί είτε στα εγχειρίδια που συνοδεύουν ή επεξηγούν τις βιβλιοθήκες, αλλά και με απευθείας ερωτήσεις στο διαδίκτυο.

Παράδειγμα 2: Τυχαίος περίπατος.

Θεωρούμε ένα κινητό που εκτελεί απλό τυχαίο περίπατο σε άξονα, ρίχνοντας ένα κανονικό νόμισμα. Το κινητό πηγαίνει μία θέση δεξιά όταν έρχεται Κορώνα ή μία θέση αριστερά όταν έρχεται Γράμματα. Παριστάνουμε την κίνηση με μία τροχιά σημείων (x,y) , όπου x =αύξων αριθμός ρίψης και y =θέση του κινητού.

Θα προσομοιώσουμε τον τυχαίο περίπατο με την Python.

Κώδικας του παραδείγματος 2.

```
'''      Τυχαίος περίπατος σε μία διάσταση      '''

% matplotlib inline # (1)
import matplotlib.pyplot as plt
import random
import numpy as np

def tales(p): # Κίνηση αν έλθει Γράμματα (2)
    x = p[0]
    y = p[1]
    return x + 1, y - 1

def heads(p): # Κίνηση αν έλθει Κορώνα (3)
    x = p[0]
    y = p[1]
```

```

    return x + 1, y + 1

def coin(p): # Οι δύο δυνατότητες του νομίσματος (3)
    coin_drop = [heads, tails] # Ρίψη του νομίσματος
    cd = random.choice(coin_drop)
    x, y = cd(p)
    return x, y

def trajectory(p, n): # (4)
    x = [p[0]]
    y = [p[1]]
    for i in range(n):
        p = coin(p)
        x.append(p[0])
        y.append(p[1])
    return x, y

if __name__ == '__main__': # (5)
    p = (0, 0) # Αρχικό σημείο εκκίνησης
    n = int(input('Εισάγετε επιθυμητό πλήθος ρίψεων: '))
    random.seed(9006) # (6)
    x, y = trajectory(p, n) # (7)
    perc=sum(1 for t in y if t>=0)/n # ποσοστό τροχιάς άνω του 0 (8)
    plt.plot(x, y) # Το γράφημα
    plt.hlines(0, 0, n, colors='red')
    plt.xlabel('Αύξων Αριθμός Ρίψης')
    plt.ylabel('Θέση στον περίπατο')
    plt.text(0, -40, 'percentage over 0= %5.3f' % perc)
    plt.show()

```

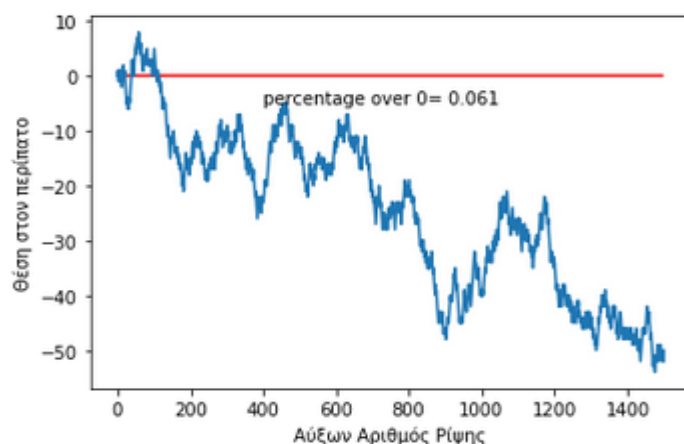
Μία εκτέλεση με 1500 ρίψεις αυτού του προγράμματος έδωσε την Εικ. 4:

Το κυρίως πρόγραμμα βρίσκεται στην τελευταία παράγραφο (σημείο(5)), όπου πρώτα δηλώνουμε ένα αρχικό σημείο εκκίνησης και ζητούμε από το χρήστη να δηλώσει πόσες ρίψεις του νομίσματος να προσομοιώσουμε. Στη συνέχεια στο (6) δηλώνουμε ένα αρχικό σπόρο για τους τυχαίους αριθμούς που θα χρησιμοποιήσουμε, ώστε να πάρουμε το γράφημα της Εικ. 4. Αν βάλουμε άλλο σπόρο ή άλλο πλήθος ρίψεων, θα έχουμε διαφορετική τροχιά, οπότε θα πρέπει να τροποποιήσουμε κατάλληλα τη θέση της λεζάντας. Στην (7) καλούμε την υπορουτίνα-συνάρτηση `trajectory()` (που δίνεται στο σημείο (4)), η οποία εκτελεί όλη τη διαδικασία. Καλεί την `coin()` η οποία με πιθανότητα $\frac{1}{2}$ καλεί την `heads()` και με άλλη τόση την `tales()`. Οι δύο τελευταίες εκτελούν τον τυχαίο περίπατο είτε δεξιά είτε αριστερά από την τρέχουσα θέση και έτσι προστίθενται τα σημεία της τροχιάς. Στην (8) υπολογίζεται με έναν πολύ ευφυή τρόπο το ποσοστό των ρίψεων κατά τις οποίες η τροχιά είναι στο θετικό ημιεπίπεδο. Τέλος, στις επόμενες εντολές σχηματίζουμε τη γραφική παράσταση της τροχιάς, τη θέση του άξονα και επιπλέον το ποσοστό της τροχιάς που είναι στο θετικό ημιεπίπεδο, που όπως είναι γνωστό είναι στις περισσότερες περιπτώσεις είτε κοντά στο 1 είτε κοντά στο 0.

Παράδειγμα 3: Κατανομή μήκους λέξεων και γραμμάτων στον Εθνικό Ύμνο.

Ψάχνοντας στο δίκτυο για τον Εθνικό μας Ύμνο, βρίσκουμε πάρα πολλές σελίδες που έχουν όλο τον Εθνικό Ύμνο, δηλαδή τα 158 τετράστιχα που έγραψε ο Εθνικός μας ποιητής ο Διονύσιος Σολωμός.

Μία από αυτές, που έχει τον ύμνο σε καλή μορφή είναι η:
https://el.wikisource.org/wiki/Υμνος_εις_την_Ελευθερίαν



Εικόνα 4: Εκτέλεση του κώδικα του Παραδ. 2

Αντιγράφουμε το ποίημα και το μεταφέρουμε σε ένα κελί στο Jupyter, όπως φαίνεται στον κώδικα (όπου είναι μόνο οι δύο πρώτοι και οι δύο τελευταίοι στίχοι).

Το ποίημα είναι σε μορφή αλφαριθμητική και αν χρειάζεται μπορούμε να το επεξεργαστούμε με διάφορους τρόπους. Κάτι χρήσιμο για τον υπολογισμό του μήκους των λέξεων και του πλήθους των γραμμάτων, είναι να απαλλάξουμε το κείμενο από τα διάφορα σημεία στίξης, κόμματα, τελείες, άνω τελείες, θαυμαστικά κλπ, που η παρουσία τους θα αλλοίωνε τα ζητούμενα μεγέθη. Για να γίνει αυτό χρησιμοποιούμε τη βιβλιοθήκη `re` (Regular Expressions) και αντικαθιστούμε όλα τα σημεία στίξης με κενό. Το νέο κείμενο το συμβολίσαμε `s` (θέσεις (1), (2) στον κώδικα).

Γνωρίζουμε ότι το ποίημα έχει $158 \cdot 4 = 632$ στίχους. Έστω, όμως, ότι θέλουμε να βρούμε πόσες λέξεις έχει.

Για το σκοπό αυτό μετατρέπουμε το κείμενο σε λίστα χρησιμοποιώντας τη μέθοδο `split` (θέση (3)). Οι πρώτες 8 και οι τελευταίες 8 λέξεις της λίστας που δημιουργήσαμε φαίνονται στη θέση (4).

Με μια απλή εντολή εκτύπωσης στην (5) βρίσκουμε ότι το πλήθος των λέξεων του ποιήματος είναι 2594.

Με τις τρεις εντολές στο (6) σχηματίζουμε μια νέα λίστα `s_words_lengths` στην οποία φορτώνουμε τα μήκη των λέξεων του ύμνου. Στο (7) φαίνονται τα μήκη των 20 πρώτων λέξεων. Με την εντολή (8) βρίσκουμε ότι τα γράμματα του εθνικού ύμνου είναι 12369.

Με την εντολή στο (9) υπολογίζουμε τον πίνακα συχνοτήτων των μηκών των λέξεων και στο (10) τον τυπώνουμε. Στο (11) κάνουμε ένα ραβδόγραμμα των μηκών των λέξεων που είναι αυτό της Εικ. 5. Αν θέλουμε να δούμε ποια είναι η λέξη με τα 15 γράμματα δίνουμε την εντολή στο (12) και μας δίνει τη λέξη 'κατασκορπιούντο'. Αν θέλουμε να καταγράψουμε όλες τις λέξεις με δύο γράμματα θα δώσουμε την εντολή στο (13) που δημιουργεί μια λίστα με όλες αυτές τις λέξεις, που όπως είδαμε είναι 366.

Για να υπολογίσουμε τώρα τις συχνότητες των γραμμάτων, πρέπει πρώτα να αντιστοιχίσουμε όλα τα γράμματα τα τονούμενα, τα κεφαλαία, αυτά που έχουν περισπωμένες, δασείες κλπ με απλά γράμματα. Αυτό γίνεται με τη δημιουργία ενός λεξικού (θέση (14) στον κώδικα) που εδώ είναι αρκετά εκτεταμένο λόγω του ότι το κείμενο ήταν σε πολυτονικό σύστημα. Στη συνέχεια στο (15) δημιουργούμε ένα άλλο λεξικό που αντιστοιχεί σε κάθε γράμμα τη συχνότητά του. Το λεξικό αυτό μπορούμε να το τυπώσουμε σαν πίνακα (16) (βλ. Πίν. 1), να το διατάξουμε σε αύξουσα σειρά γραμμάτων (17) και να πάρουμε ένα ραβδόγραμμα (18) που δίνεται στην Εικ. 6. Παρατηρούμε ότι το γράμμα που επικρατεί είναι το Α με μεγάλη διαφορά και ακολουθούν τα Ε, Ι και Ο.

Κώδικας του παραδείγματος 3.

```
nat_anthem='''
    Σὲ γνωρίζω ἀπὸ τὴν κόψι
    τοῦ σπαθιοῦ τὴν τρομερῆ,
-----
    Βασιλεῖς, ἐλάτε, ἐλάτε,
    καὶ κτυπήσετε κι ἐδῶ!»
'''

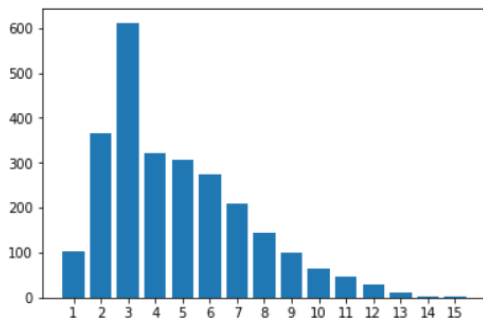
import re          # (1)
import numpy as np # (2)
s=re.sub('[\.\;:\!\\,\·\|\«\»]', ' ', nat_anthem) # (2)
s_words=s.split() # Μήκη Λέξεων (3)
print(s_words[:8])
print(30*'-' )

print(s_words[-8:]) # (4)
# ['Σὲ', 'γνωρίζω', 'ἀπὸ', 'τὴν', 'κόψι', 'τοῦ', 'σπαθιοῦ', 'τὴν']
# -----
# ['Σταυρό', 'Βασιλεῖς', 'ἐλάτε', 'ἐλάτε', 'καὶ', 'κτυπήσετε', 'κι', 'ἐδῶ']
print("To πλήθος λέξεων του Εθνικού Ὕμνου εἶναι:
{}".format(len(s_words))) # (5)
# Το πλήθος λέξεων του Εθνικού Ὕμνου εἶναι: 2594
s_words_lengths=[] # (6)
for w in s_words:
    s_words_lengths.append(len(w))
print(s_words_lengths[:20]) # (7)
# [2, 7, 3, 3, 4, 3, 7, 3, 7, 2, 7, 3, 3, 3, 3, 2, 3, 7, 3, 2]
sum(s_words_lengths) # 12369 # (8)
unique_lengths, counts_lengths = np.unique(s_words_lengths,
      return_counts=True) # (9)
print("Συχνότητες μηκῶν λέξεων του Εθνικού Ὕμνου:")
print(np.asarray((unique_lengths, counts_lengths)))
# Συχνότητες μηκῶν λέξεων του Εθνικού Ὕμνου: # (10)
# [[ 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15]
# [102 366 612 322 308 275 210 144 101 63 48 30 11 1 1]]
plt.bar(unique_lengths, counts_lengths) # (11)
plt.xticks(unique_lengths, unique_lengths)
s_words[s_words_lengths.index(15)] # 'κατασκορπιῶντο' # (12)
[s_words[i] for i, e in enumerate(s_words_lengths) if e == 2] # (13)
# ['Σὲ', 'σὲ', 'μὲ', 'γῆ', 'τὰ', 'τὰ', 'κι', '.....', 'ἐξ', 'κι'] # 366 λέξεις
# Συχνότητες Γραμμάτων
tonoi={"ά":"α", "Α":"α", "ᾶ":"α", "ᾷ":"α", "ῶ":"α", "ῷ":"α", "ᾠ":"α", "ᾡ":"α", "ᾢ":"α", "ᾣ":"α", "ᾤ":"α", "ᾥ":"α", "ᾦ":"α", "ᾧ":"α", "ᾨ":"α", "ᾩ":"α", "ᾪ":"α", "ᾫ":"α", "ᾬ":"α", "ᾭ":"α", "ᾮ":"α", "ᾯ":"α",
"E":"ε", "Ε":"ε", "ἒ":"ε", "ἓ":"ε", "ἔ":"ε", "ἕ":"ε", "἖":"ε", "἗":"ε", "Ἐ":"ε", "Ἑ":"ε", "Ἒ":"ε", "Ἓ":"ε", "Ἔ":"ε", "Ἕ":"ε", "ἦ":"ε", "ἧ":"ε", "Ἠ":"ε", "Ἡ":"ε", "Ἢ":"ε", "Ἣ":"ε", "Ἤ":"ε", "Ἥ":"ε", "Ἰ":"ι",
"Ι":"ι", "Ἢ":"ι", "Ἣ":"ι", "Ἤ":"ι", "Ἥ":"ι", "Ἰ":"ι", "Ἱ":"ι", "ἰ":"ι", "ἱ":"ι", "ἲ":"ι", "ἳ":"ι", "ἴ":"ι", "ἵ":"ι", "ἶ":"ι", "ἷ":"ι", "Ἰ":"ι", "Ἱ":"ι", "ἰ":"ι", "ἱ":"ι", "ἲ":"ι", "ἳ":"ι",
"Ο":"ο", "ό":"ο", "ο":"ο", "ό":"ο", "ο":"ο", "ό":"ο", "ο":"ο", "ο":"ο", "ό":"ο", "ο":"ο", "ο":"ο", "ο":"ο", "ο":"ο", "ο":"ο", "ο":"ο", "ο":"ο", "ο":"ο", "ο":"ο", "ο":"ο", "ο":"ο", "ο":"ο", "ο":"ο",
"ύ":"υ", "ύ":"υ", "ύ":"υ", "ύ":"υ", "ύ":"υ", "ύ":"υ", "ύ":"υ", "ύ":"υ", "ύ":"υ", "ύ":"υ", "ύ":"υ", "ύ":"υ", "ύ":"υ", "ύ":"υ", "ύ":"υ", "ύ":"υ", "ύ":"υ", "ύ":"υ", "ύ":"υ", "ύ":"υ", "ύ":"υ", "ύ":"υ", "ύ":"υ", "ύ":"υ",
"ω":"ω", "ώ":"ω", "ω":"ω", "ώ":"ω", "ώ":"ω", "ώ":"ω", "ώ":"ω", "ώ":"ω", "ώ":"ω", "ώ":"ω", "ώ":"ω", "ώ":"ω", "ώ":"ω", "ώ":"ω", "ώ":"ω"}
grammata = {}
```

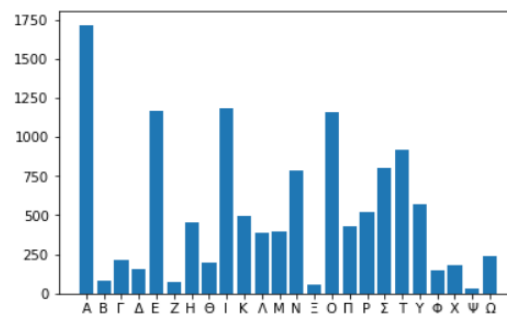
```
for c in s: # (15)
    if c.isalpha():
        if c in tonoi:
            c = tonoi[c]
        c = c.upper()
        grammata[c] = grammata.get(c,0) + 1
for c in sorted(grammata.keys()) : # (16)
    print ( c, " : ", grammata[c])
sorted_grammata= {k: grammata[k] for k in sorted(grammata)} # (17)
plt.bar(*zip(*sorted_grammata.items())) # (18)
```

Πίνακας 1. Συχνότητες γραμμάτων Εθνικού Ύμνου

Α : 1716	Η : 455	Ν : 783	Τ : 918
Β : 84	Θ : 195	Ξ : 53	Υ : 574
Γ : 211	Ι : 1183	Ο : 1162	Φ : 144
Δ : 159	Κ : 494	Π : 430	Χ : 179
Ε : 1168	Λ : 387	Ρ : 523	Ψ : 30
Ζ : 76	Μ : 394	Σ : 799	Ω : 238



Εικόνα 5: Μήκη λέξεων



Εικόνα 6: Συχνότητες γραμμάτων

Η βιβλιοθήκη `scipy`

Η βιβλιοθήκη `numpy` που χρησιμοποιήθηκε στα προηγούμενα χρειάζεται κυρίως για μαθηματικές πράξεις, γνωστές συναρτήσεις και όπως είδαμε και για απλή περιγραφική στατιστική. Αυτή όμως που εστιάζεται στη στατιστική ανάλυση είναι η βιβλιοθήκη `scipy` και ιδιαίτερα μια υπο-βιβλιοθήκη `scipy.stats` που έχει φορτωμένες όλες τις γνωστές κατανομές και μεθόδους. Μία πολύ χρήσιμη σελίδα που περιγράφει τις δυνατότητες της `scipy.stats` και δίνει παραδείγματα χρήσης της είναι η:

<https://docs.scipy.org/doc/scipy/reference/tutorial/stats.html>

Στη σελίδα αυτή έχει δύο διασυνδέσεις που μας πηγαίνουν στις διακριτές και στις συνεχείς κατανομές, αντίστοιχα. Δίνουμε παρακάτω δύο παραδείγματα από αυτές.

Παράδειγμα 4. Η κατανομή Poisson

```
%matplotlib inline
from scipy.stats import poisson
import matplotlib.pyplot as plt
import numpy as np
fig, ax = plt.subplots(1, 1)
mu = 0.6
mean, var, skew, kurt = poisson.stats(mu, moments='mvsk')
```

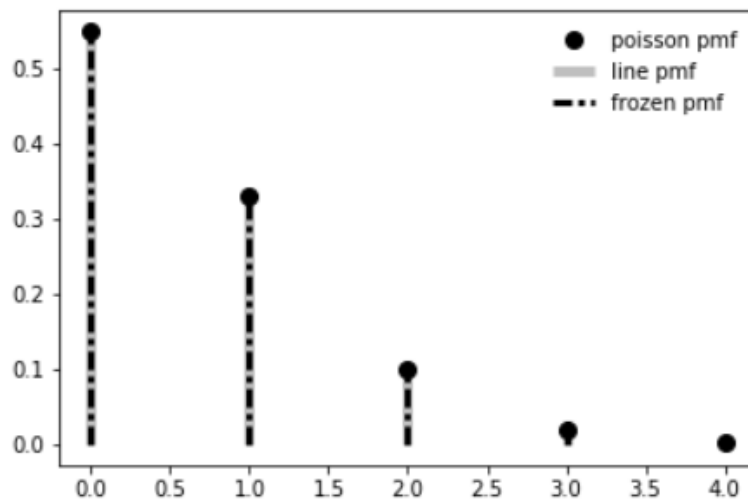


```

x = np.arange(poisson.ppf(0.01, mu),
              poisson.ppf(0.9999, mu))
ax.plot(x, poisson.pmf(x, mu), 'ko', ms=8, label='poisson pmf')
ax.vlines(x, 0, poisson.pmf(x, mu), colors='gray',
          linestyle='solid', lw=5, alpha=0.5, label='line pmf') # (1)
rv = poisson(mu)
ax.vlines(x, 0, rv.pmf(x), colors='k',
          linestyle=(0, (3,1,1,1)), lw=3, label='frozen pmf') # (2)
ax.legend(loc='best', frameon=False)
plt.show()

```

Η εκτέλεση του παραπάνω κώδικα δίνει το γράφημα της Εικ. 7. Παρατηρούμε ότι θεωρητικές πιθανότητες (σημείο (1)), που συμβολίζονται με κυκλάκι και γκρι συνεχή γραμμή, ταυτίζονται με τις σχετικές συχνότητες του τυχαίου δείγματος (σημείο (2)), που συμβολίζονται με γραμμές διακεκομμένες, όπως αναμένονταν.



Εικόνα 7: Η κατανομή Poisson

Παράδειγμα 5. Η κατανομή χ^2 με 78 β.ε.

```

%matplotlib inline
from scipy.stats import chi2
import matplotlib.pyplot as plt
import numpy as np
fig, ax = plt.subplots(1, 1)
df = 78
mean, var, skew, kurt = chi2.stats(df, moments='mvsk')
x = np.linspace(chi2.ppf(0.01, df),
                 chi2.ppf(0.99, df), 100)

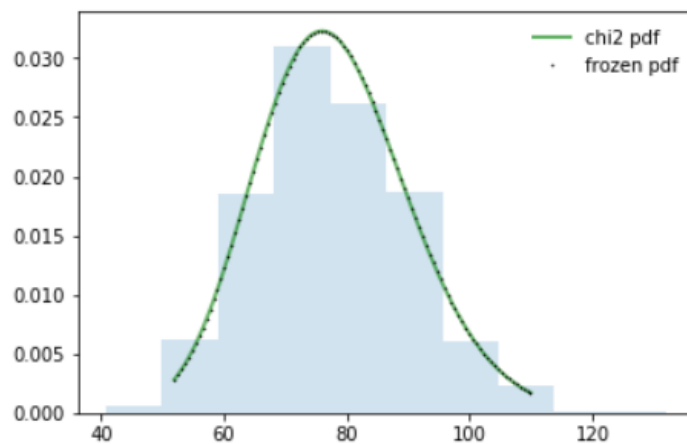
```

```

ax.plot(x, chi2.pdf(x, df),
        'g-', lw=2, alpha=0.6, label='chi2 pdf') # (1)
rv = chi2(df)
ax.plot(x, rv.pdf(x), 'k.', lw=.3, label='frozen pdf',
        markersize=1) # (2)
r = chi2.rvs(df, size=1000)
ax.hist(r, density=True, histtype='stepfilled', alpha=0.2) # (3)
ax.legend(loc='best', frameon = False)
plt.show()

```

Η εκτέλεση του παραπάνω κώδικα δίνει το γράφημα της Εικ. 8. Παρατηρούμε ότι θεωρητική συνάρτηση πυκνότητας πιθανότητας (σημείο (1)), που συμβολίζεται με συνεχή γραμμή, ταυτίζεται με την καμπύλη σχετικών συχνοτήτων του τυχαίου δείγματος (σημείο (2)), που συμβολίζεται με γραμμή από σημεία, όπως αναμένονταν. Επίσης, στο σημείο (3), σχεδιάζεται ένα ιστόγραμμα του τυχαίου δείγματος, και επιβεβαιώνεται ότι η συνάρτηση πυκνότητας πιθανότητας διέρχεται από τα μέσα των ορθογωνίων.



Εικόνα 8: Η κατανομή χ^2_{78}

Άλλες βιβλιοθήκες όπως pandas, rpy2, κλπ

Τα περιορισμένα πλαίσια μιας παρουσίασης δεν μας επιτρέπουν να ασχοληθούμε με όλες τις βιβλιοθήκες της Python. Για παράδειγμα η βιβλιοθήκη pandas δημιουργεί και χειρίζεται πλαίσια δεδομένων με πολύ αποδοτικό τρόπο, όπως περιγράφεται στο βιβλίο του Femi Anthony. Η βιβλιοθήκη rpy2 δίνει τη δυνατότητα να τρέχουμε εντολές της R στο περιβάλλον της Python, και υπάρχουν πολλές άλλες βιβλιοθήκες που δημιουργήθηκαν για να λύσουν συγκεκριμένα προβλήματα.

ΕΠΙΛΟΓΟΣ

Από τα παραπάνω, διαπιστώνουμε ότι η Python έχει τη δυνατότητα να κάνει Στατιστική. Αυτό, δεν σημαίνει ότι θα έπρεπε να εγκαταλείψουμε την R, για παράδειγμα, ή άλλα στατιστικά πακέτα και να ασχοληθούμε μόνο με την Python. Η προσωπική μας γνώμη είναι ότι βοηθά σε κάποιες περιπτώσεις η γνώση της Python, κυρίως όταν τα δεδομένα μας είναι σε κάποιες σελίδες ή ιστοτόπους στο διαδίκτυο και απαιτείται να τα αναζητήσουμε και να τα προσαρμόσουμε για την ανάλυση που πρόκειται να κάνουμε, όπως π.χ. το παράδειγμα 3 με τον Εθνικό μας ύμνο.

Καταλήγοντας, θα συνιστούσαμε στους νέους στατιστικούς, να προσθέσουν και την Python στα εργαλεία της έρευνάς τους, ώστε με έναν συμπληρωματικό τρόπο να πετυχαίνουν το στόχο τους σε όποιο πρόβλημα αντιμετωπίζουν.

SUMMARY

The Python programming language is a general programming language that can reliably encounter problems of analysis and visualization from all fields of science, including statistics. Available free on the internet, has many manuals with instructions and several books with applications in various fields of science. With the help of a Jupyter platform, working with Python becomes very easy, stored and transferred. We describe here some of the libraries that focus on statistical analyses and graphs and we give some applications.

ΑΝΑΦΟΡΕΣ

- Femi Anthony (2015). *Mastering pandas-Master the features and capabilities of pandas, a data analysis toolkit for Python*, PACKT
- Haslwanter Thomas (2016). *An Introduction to Statistics with Python-With Applications in the Life Sciences*, Springer
- Johansson Robert_(2015). *Numerical Python - A Practical Techniques Approach for Industry*, Apress



ΣΥΝΑΡΤΗΣΕΙΣ ΠΥΚΝΟΤΗΤΑΣ ΠΙΘΑΝΟΤΗΤΑΣ ΜΕ ΑΡΝΗΤΙΚΟ ΕΚΘΕΤΗ: ΘΕΩΡΗΤΙΚΗ ΚΑΙ ΔΕΙΓΜΑΤΟΛΗΠΤΙΚΗ ΠΡΟΣΕΓΓΙΣΗ

Παπατσούμα Ιωάννα, Φαρμάκης Νικόλαος

Τμήμα Μαθηματικών, Σχολή Θετικών Επιστημών, Αριστοτέλειο Πανεπιστήμιο
Θεσσαλονίκης

ioannaparatsouma@gmail.com, farmakis@math.auth.gr

ΠΕΡΙΛΗΨΗ

Η δειγματοληψία μπορεί να συμβάλει στη μελέτη των κατανομών των διαφόρων τυχαίων μεταβλητών με πολύ καλά αποτελέσματα και μικρό υπολογιστικό κόστος. Το δείγμα και οι διάφορες παραμέτροι μιας τυχαίας μεταβλητής (τ.μ.) X , μέσα από κάποια διαδικασία, οδηγούν στην εκτίμηση του τύπου της συνάρτησης πυκνότητας πιθανότητας (σ.π.π.) της τμ X . Η διαδικασία αυτή που ακολουθείται είναι συνήθως θεωρητική και βασίζεται σε μια αρχική υπόθεση για την μορφή που μπορεί να έχει η σ.π.π. ή τουλάχιστο μία προσέγγισή της που μπορεί να είναι αποδοτική και εύκολη στη διαχείρισή της συγχρόνως. Σημαντικό ρόλο στην επιλογή του τύπου της σ.π.π. που θα χρησιμοποιηθεί παίζει και ο χρόνος που θα χρειαστεί για την επεξεργασία των δεδομένων για το σκοπό αυτό. Στην παρούσα εργασία γίνεται η μελέτη της διαδικασίας που οδηγεί από το δείγμα στην εκτιμήτρια της σ.π.π. όταν η σ.π.π. έχει μορφή μονώνυμου με εκθέτη αρνητικό ακέραιο, αρχικά για τις τιμές -1 και -2 και στη συνέχεια, γενικεύεται το συμπέρασμα.

Λέξεις κλειδιά: Δειγματοληψία, συνάρτηση πυκνότητας πιθανότητας, μέση τιμή, διασπορά, συντελεστής μεταβλητότητας

AMS ταξινόμηση: 62D05, 62E17

1. ΕΙΣΑΓΩΓΗ

Η δειγματοληψία αποτελεί απαραίτητο εργαλείο για ερευνητές όλων των επιστημών, θετικών, εφαρμοσμένων, ανθρωπιστικών, καθώς και κοινωνικών επιστημών, και η ανάδειξη μοντέλων κατανομών μέσα από δειγματοληψία είναι μία αποτελεσματική διαδικασία (Παπατσούμα & Φαρμάκης, 2015, 2017; Papatsouma & Farmakis, 2020). Αξίζει μάλιστα να τονιστεί το μικρό υπολογιστικό κόστος, με το οποίο αυτή υλοποιείται, από το οποίο μπορεί να επωφεληθεί ο εκάστοτε ερευνητής. Η συνάρτηση πυκνότητας πιθανότητας (σ.π.π.), έννοια γνωστή από τη βιβλιογραφία, αναφέρεται σε μία τυχαία μεταβλητή (τ.μ.) X και μας δίνει μια καλή και πλήρη εικόνα για τη συμπεριφορά και την «ταυτότητα» της. Η γνώση της σ.π.π. είναι πολύ χρήσιμη για τη σωστή μελέτη της X και συντελεί μεταξύ άλλων στην εύρεση των ποσοστιαίων σημείων, όπως και στην κατασκευή στοχαστικών μοντέλων για διάφορες εφαρμογές (Hollands & Suehrcke, 2013; Munkhammar et al., 2014). Από την ανασκόπηση της σχετικής βιβλιογραφίας, διαπιστώθηκε ότι με την εκτίμηση της σ.π.π. γνωστών κατανομών (κανονική κατανομή, λογαριθμοκανονική κατανομή και κατανομή Weibull) έχουν ασχοληθεί, επίσης, οι (Munkhammar et al., 2017) με μια τελείως διαφορετική προσέγγιση και πολυπλοκότερο τύπο, καθώς επιλέγουν αυθαίρετα έναν εκθέτη, N , φυσικό αριθμό, με μόνη προϋπόθεση να μπορεί να υπολογιστεί η ροπή N -τάξης. Στην παρούσα εργασία η δειγματοληψία δίνει αρχικά αρκετά στοιχεία χρήσιμα ώστε να αξιοποιηθεί και η υπόλοιπη πληροφορία που βρίσκεται στα δεδομένα. Το βασικό στοιχείο είναι το εύρος της κατανομής της τ.μ. X το οποίο εκτιμάται και από τη μέγιστη τιμή των δειγματικών δεδομένων X_{max} αλλά γίνεται και η εκτίμηση της ελάχιστης τιμής της κατανομής της τ.μ. X .

2. ΜΕΛΕΤΗ ΣΥΝΑΡΤΗΣΕΩΝ ΠΥΚΝΟΤΗΤΑΣ ΠΙΘΑΝΟΤΗΤΑΣ ΜΕ ΑΡΝΗΤΙΚΟ ΕΚΘΕΤΗ

Η σ.π.π. μιας συνεχούς τ.μ. X έχει τις εξής δύο βασικές ιδιότητες:

$$f(x) \geq 0, \forall x \quad (1)$$

$$\int_{-\infty}^{+\infty} f(x) dx = 1 \quad (2)$$

Το πλήθος των μορφών των σ.π.π. είναι απείρως υπεραριθμίσσιμο. Εδώ γίνεται μελέτη των σ.π.π. μονωνυμικής μορφής με γενικό τύπο:

$$f(x) = \begin{cases} x^{-\nu}, & x \in [\varepsilon, \beta] \\ 0, & x \notin [\varepsilon, \beta] \end{cases}, 0 < \varepsilon < \beta, \nu = 1, 2, 3, \dots \quad (3)$$

όπου το β προσδιορίζεται κατάλληλα από το δείγμα δ , το οποίο επιλέγεται τυχαία από τον πληθυσμό (Φαρμάκης, 2016), και είναι πάντα $\beta > X_{max}$ και το ε είναι κατάλληλο ώστε να ισχύουν οι σχέσεις (1) και (2).

Στη συνέχεια, θα εξεταστούν οι περιπτώσεις $\nu = 1$ και $\nu = 2$ καθώς και η γενική περίπτωση για όλες τις δυνατές τιμές του $\nu > 2$.

Για $\nu = 1$, η συνάρτηση f από τη σχέση (3) γίνεται:

$$f(x) = \begin{cases} x^{-1}, & x \in [\varepsilon, \beta] \\ 0, & x \notin [\varepsilon, \beta] \end{cases}, 0 < \varepsilon < \beta \quad (4)$$

Προφανώς ισχύει η (1) επειδή $0 < \varepsilon \leq x \leq \beta$. Επίσης είναι:

$$1 = \int_{-\infty}^{+\infty} f(x) dx = \int_{\varepsilon}^{\beta} x^{-1} dx = \ln \frac{\beta}{\varepsilon} \Leftrightarrow \varepsilon = \frac{\beta}{e} \quad (5)$$

Οι παράμετροι της τ.μ. X υπολογίζονται από τις γνωστές σχέσεις ως εξής:
Μέση τιμή:

$$EX = \mu = \int_{\varepsilon}^{\beta} x \cdot f(x) dx = \beta - \varepsilon = \beta \cdot \left(\frac{e-1}{e}\right) \quad (6)$$

Και με ακρίβεια 4 δ.κ.ψ είναι $\mu = 0,6321\beta$.

Διασπορά:

$$VarX = \sigma^2 = EX^2 - (EX)^2 = \frac{\beta^2(e^2-1)}{2 \cdot e^2} - \beta^2 \cdot \left(\frac{e-1}{e}\right)^2 = \frac{\beta^2}{2 \cdot e^2} \cdot (1 - (e-2)^2) \quad (7)$$

και με ακρίβεια 4 δ.κ.ψ. είναι: $\sigma^2 = 0,0328 \cdot \beta^2$.

Συντελεστής Μεταβλητότητας (ΣΜ):

$$CV = \frac{\sqrt{VarX}}{EX} = \frac{\sqrt{2 \cdot \{1 - (e-2)^2\}}}{2 \cdot (e-1)} \quad (8)$$

και με ακρίβεια 4 δ.κ.ψ. είναι $CV = 0,2863$ ανεξάρτητος του β .

και ο αντίστροφος του Συντελεστή Μεταβλητότητας στο τετράγωνο:

$$ICV^2 = \frac{(EX)^2}{VarX} = \frac{2 \cdot (e-1)^2}{1 - (e-2)^2} \quad (9)$$

και με ακρίβεια 4 δ.κ.ψ. είναι ανεξάρτητο του β , $q = ICV^2 = 12,1999$.

Για $\nu = 2$, η συνάρτηση f από τη σχέση (3) γίνεται:

$$f(x) = \begin{cases} x^{-2}, & x \in [\varepsilon, \beta] \\ 0, & x \notin [\varepsilon, \beta] \end{cases}, 0 < \varepsilon < \beta \quad (10)$$

Προφανώς ισχύει η (1) επειδή $0 < \varepsilon \leq x \leq \beta$. Επίσης είναι:

$$1 = \int_{-\infty}^{+\infty} f(x) dx = \int_{\varepsilon}^{\beta} x^{-2} dx = -x^{-1} \Big|_{\varepsilon}^{\beta} = \frac{\beta - \varepsilon}{\beta \cdot \varepsilon} = 1 \Leftrightarrow \varepsilon = \frac{\beta}{\beta+1} \quad (11)$$

Οι παράμετροι της τ.μ. X υπολογίζονται από τις γνωστές σχέσεις ως εξής:
Μέση τιμή:

$$EX = \mu = \int_{\varepsilon}^{\beta} x \cdot f(x) dx = \int_{\varepsilon}^{\beta} \frac{dx}{x} = \ln x \Big|_{\varepsilon}^{\beta} = \ln \frac{\beta}{\varepsilon} = \ln(\beta + 1) \quad (12)$$

Διασπορά:

$$VarX = \sigma^2 = EX^2 - (EX)^2 = \frac{\beta^2}{\beta+1} - \{\ln(\beta + 1)\}^2 = \frac{\beta^2 - (\sqrt{\beta+1} \ln(\beta+1))^2}{\beta+1} \quad (13)$$

Συντελεστής Μεταβλητότητας (ΣΜ):

$$CV = \sqrt{\frac{\beta^2}{\{\sqrt{\beta+1} \ln(\beta+1)\}^2} - 1} \quad (14)$$

και ο αντίστροφος του Συντελεστή Μεταβλητότητας στο τετράγωνο:

$$q = ICV^2 = \frac{\{\sqrt{\beta+1} \ln(\beta+1)\}^2}{\beta^2 - \{\sqrt{\beta+1} \ln(\beta+1)\}^2} \quad (15)$$

v > 2 ΚΑΙ ΓΕΝΙΚΕΥΣΗ

Για $v > 2$, ισχύει η (1) επειδή $0 < \varepsilon \leq x \leq \beta$.

Αναφορικά με τη (2):

$$1 = \int_{-\infty}^{+\infty} f(x) dx = \int_{\varepsilon}^{\beta} x^{-v} dx = -\frac{x^{-v+1}}{v-1} \Big|_{\varepsilon}^{\beta} = \frac{\beta^{v-1} - \varepsilon^{v-1}}{(v-1) \cdot \beta^{v-1} \cdot \varepsilon^{v-1}}$$

$$\Leftrightarrow \frac{1}{\varepsilon^{v-1}} - \frac{1}{\beta^{v-1}} = v-1$$

Και τελικά προκύπτει:

$$\varepsilon = \frac{1}{v-1 \sqrt{(v-1) + \frac{1}{\beta^{v-1}}}} \quad (16)$$

Το ε μπορεί να θεωρηθεί ως ο v -οστός όρος μιας ακολουθίας ε_n . Τότε η (16) έχει τη μορφή:

$$\varepsilon_n = \frac{1}{v-1 \sqrt{(v-1) + \frac{1}{\beta^{v-1}}}} \quad (17)$$

και η ακολουθία των ε_n έχει όριο τη μονάδα (Μουσιάδης, 2000). Από τη (17) προκύπτει ότι εν γένει είναι $0 < \varepsilon_n < 1$.

Οι παράμετροι της τ.μ. X υπολογίζονται από τις γνωστές σχέσεις ως εξής:

Μέση τιμή:

$$EX = \mu = \int_{\varepsilon}^{\beta} x \cdot f(x) dx = \int_{\varepsilon}^{\beta} x^{-v+1} dx = -\frac{1}{v-2} \cdot x^{-v+2} \Big|_{\varepsilon}^{\beta} = \frac{1}{v-2} \cdot \frac{\beta^{v-2} - \varepsilon^{v-2}}{\beta^{v-2} \cdot \varepsilon^{v-2}} \quad (18)$$

Διασπορά:

$$VarX = \sigma^2 = EX^2 - (EX)^2 = \frac{\beta^{v-3} - \varepsilon^{v-3}}{(v-3) \cdot \beta^{v-3} \cdot \varepsilon^{v-3}} - \frac{(\beta^{v-2} - \varepsilon^{v-2})^2}{(v-2)^2 \cdot \beta^{2v-4} \cdot \varepsilon^{2v-4}}$$

$$= \frac{\beta^{2v-4} \cdot \{(v-2)^2 \cdot \varepsilon^{v-1} - v+3\} - \varepsilon^{2v-4} \cdot \{(v-2)^2 \cdot \beta^{v-1} - v+3\} + 2(v-3) \cdot (\beta\varepsilon)^{v-2}}{(v-2)^2 \cdot (v-3) \cdot (\beta\varepsilon)^{2v-4}} \quad (19)$$

Συντελεστής Μεταβλητότητας (ΣΜ):

$$CV = \sqrt{\frac{\beta^{2v-4} \cdot \{(v-2)^2 \cdot \varepsilon^{v-1} - v+3\} - \varepsilon^{2v-4} \cdot \{(v-2)^2 \cdot \beta^{v-1} - v+3\} + 2(v-3) \cdot (\beta\varepsilon)^{v-2}}{(v-3) \cdot (\beta^{2v-4} - \varepsilon^{2v-4})}} \quad (20)$$

και ο αντίστροφος του Συντελεστή Μεταβλητότητας στο τετράγωνο:

$$q = ICV^2 = \frac{(v-3) \cdot (\beta^{2v-4} - \varepsilon^{2v-4})}{\beta^{2v-4} \cdot \{(v-2)^2 \cdot \varepsilon^{v-1} - v + 3\} - \varepsilon^{2v-4} \cdot \{(v-2)^2 \cdot \beta^{v-1} + v - 3\} + 2(v-3) \cdot (\beta\varepsilon)^{v-2}} \quad (21)$$

3. ΠΑΡΑΔΕΙΓΜΑΤΑ

Παράδειγμα 1^ο: Σε τυχαίο δείγμα $n = 500$ φοιτητών τέθηκε ένα ερώτημα A , το οποίο έπρεπε να απαντηθεί μέσα από περιήγηση στο διαδίκτυο. Απάντησαν όλοι και σημειώθηκε ο χρόνος αντίδρασης τους, με μέγιστο χρόνο $X_{max} = \beta = 10$ λεπτά και ελάχιστο $X_{min} = 3,4$ λεπτά. Επειδή η ερώτηση είναι σχετικά εύκολη η κατανομή του χρόνου των απαντήσεων προέκυψε φθίνουσας μορφής. Με βήμα $w = 1,1$ και πλήθος κλάσεων $k = 6$ προκύπτει ο παρακάτω πίνακας συχνοτήτων (Πίνακας 1), συμπληρωμένος και με στοιχεία της επεξεργασίας. Για τον υπολογισμό των παραμέτρων χρησιμοποιείται η μέθοδος των ομαδοποιημένων δεδομένων σε k ομάδες (κλάσεις) (Φαρμάκης, 2001), (Κολυβά-Μαχαίρα, Μπόρα-Σέντα, 2013).

Πίνακας 1. Αποτελέσματα ερωτήματος A

Κλάσεις	Κέντρα x_i	Συχνότητες n_i	x'_i	nx'	nx^2	θ_i
3,4 – 4,5	3,95	111	-2	-222	444	101,9
4,5 – 5,6	5,05	98	-1	-98	98	108,2
5,6 – 6,7	6,15	87	0	0	0	89,7
6,7 – 7,8	7,25	76	1	76	76	76,0
7,8 – 8,9	8,35	68	2	136	272	65,9
8,9 – 10,0	9,45	60	3	180	540	58,3
ΣΥΝΟΛΑ		500		T₁ = 72	T₂ = 1430	500

Έτσι 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 ημίαθροισμα των άκρων της εκάστοτε κλάσης. Από τις κεντρικές τιμές υπολογίζονται οι τιμές $x'_i = (x_i - m)/w$, $i = 1, \dots, 6$, όπου $w = 1,1$ και m η προσωρινή μέση τιμή (κεντρική τιμή μεσαίας κλάσης συνήθως). Το κέντρο της τρίτης κλάσης ορίστηκε ως προσωρινή μέση τιμή ($m = 6,15$, βλ. εντονη (bold) γραφή). Η τελευταία στήλη περιέχει τις θεωρητικές συχνότητες, δηλαδή τα στοιχεία που θα προέκυπταν αν ίσχυε η σχέση (4), με μικρές προσαρμογές ώστε να προκύπτει άθροισμα 500 και να μπορέσει να εφαρμοστεί έπειτα η δοκιμασία X^2 .

Τίθεται τώρα το ερώτημα αν η σ.π.π. της τυχαίας μεταβλητής «Χρόνοςαντίδρασης» μπορεί να έχει τη μορφή (4), γιατί οι συχνότητες έχουν μια «ομαλά» φθίνουσα μορφή.

Απάντηση: Αρχικά και με μέγιστη τιμή της τμ $X = \text{Χρόνος αντίδρασης}$ βρίσκουμε την προβλεπόμενη για την σ.π.π. (3) ελάχιστη τιμή της X . Αυτή δίνεται από την (5) και προέκυψε $\varepsilon = 3,68$. Ήδη αυτό είναι μια καλή αρχική διαπίστωση προς την κατεύθυνση να θεωρηθεί η (3) ως σ.π.π. για την τμ X που μελετάται, επειδή $X_{\min} = 3,4 = \varepsilon$ παρατηρούμενο.

Η μέση τιμή που προκύπτει από τα πειραματικά δεδομένα είναι:

$$\bar{x} = m + w \cdot T_1/n = 6,15 + 1,1 \cdot 72/500 = 6,3084$$

Υπολογίζοντας τη μέση τιμή από την (6) προκύπτει:

$$EX = 6,3212$$

Από τον Πίνακα 1 προκύπτει η διασπορά:

$$s^2 = \frac{w^2}{n-1} \left(T_2 - \frac{T_1^2}{n} \right) = 3,4424$$

Από την (7) προκύπτει αντίστοιχα η τιμή:

$$VarX = \sigma^2 = 0,0328 \cdot \beta^2 = 3,28$$

Από την (8) υπολογίζεται η τιμή του Cv :

$$Cv = \frac{\sqrt{VarX}}{EX} = \frac{\sqrt{2 \cdot \{1 - (e-2)^2\}}}{2 \cdot (e-1)} = 0,2863$$

με τιμή από το δείγμα:

$$\hat{Cv} = \frac{\sqrt{3,4424}}{6,3084} = 0,2941.$$

Από την (9) η παράμετρος q παίρνει την τιμή (ανεξάρτητη του β) $q = ICv^2 = 12,1999$ και από το δείγμα την πειραματική τιμή:

$$q = I\hat{Cv}^2 = \frac{\bar{x}^2}{s^2} = \frac{6,3084^2}{3,4424} = 11,5601$$

Στον Πίνακα 2 παρουσιάζονται τα αποτελέσματα από το δείγμα των 500 ατόμων μαζί με τα αντίστοιχα θεωρητικά και προκύπτουν συγκριτικά στοιχεία.

Πίνακας 2. Σύγκριση δειγματικών και θεωρητικών μεγεθών παραδείγματος 1

	Μέση τιμή	Διασπορά	Τυπ. Αποκλ.	ΣΜ	q
Δείγμα	6,3084	3,4424	1,8554	0,2941	11,5601
Θεωρία	6,3212	3,2800	1,8111	0,2863	12,1999

Παρατηρούνται μικρές αποκλίσεις μεταξύ θεωρητικών και πειραματικών τιμών. Χρειάζεται όμως και η σχετική δοκιμασία για να απορριφθεί ή όχι η υπόθεση ότι τα πειραματικά δεδομένα προέρχονται από την κατανομή με τη μορφή (3) και εκθέτη -1 .

Έγινε και δοκιμασία χ^2 με υποθέσεις:

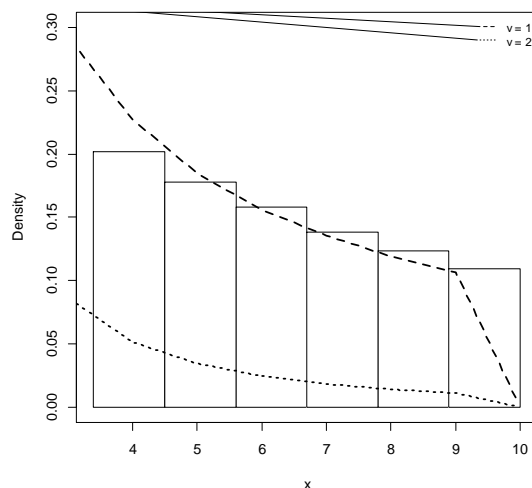
H_0 : Οι παρατηρούμενες πειραματικές τιμές προσαρμόζονται καλά στις θεωρητικές

H_1 : Όχι η H_0

Οι πειραματικές και θεωρητικές τιμές δίνονται στις στήλες 3 και 7 του Πίνακα 1. Η τιμή του στατιστικού $\chi^2 = 1,9720 < \chi_{5,0.05}^2 = 11,0705$, συνεπώς δεν υπάρχουν επαρκή στοιχεία για την απόρριψη της μηδενικής υπόθεσης σε επίπεδο σημαντικότητας 0.05.

Όπως αναφέρθηκε παραπάνω, το ερώτημα αν η σ.π. της τυχαίας μεταβλητής «Χρόνος αντίδρασης» μπορεί να έχει τη μορφή (4) γεννάται από το γεγονός ότι οι συχνότητες έχουν μια «ομαλά» φθίνουσα μορφή. Στο σχήμα 1, που ακολουθεί, η κοινή απεικόνιση και προσαρμογή των συναρτήσεων (4) και (10) καθιστά φανερή την καλή προσαρμογή της συνάρτησης της μορφής (10).

Σχήμα 1. Προσαρμογή των συναρτήσεων (4) και (10) στην κατανομή των τιμών του παραδείγματος 1



Παράδειγμα 2^ο: Σε τυχαίο δείγμα $n = 420$ φοιτητών τέθηκε ένα ερώτημα B , το οποίο έπρεπε να απαντηθεί μέσα από περιήγηση στο διαδίκτυο. Απάντησαν όλοι και σημειώθηκε ο χρόνος αντίδρασής τους, με μέγιστο χρόνο $X_{max} = \beta = 10$ min και ελάχιστο $X_{min} = 55$ sec. Επειδή η ερώτηση ήταν σχετικά εύκολη η κατανομή του χρόνου των απαντήσεων προέκυψε φθίνουσας μορφής. Με βήμα $w = 1,3$ και πλήθος κλάσεων $k = 7$ σχηματίστηκε ο παρακάτω πίνακας συχνοτήτων (Πίνακας 3),

συμπληρωμένος και με στοιχεία της επεξεργασίας. Ακολουθήθηκε η ίδια διαδικασία όπως και στον Πίνακα 1.

Πίνακας 3. Αποτελέσματα ερωτήματος Β

Κλάσεις	Κέντρα x_i	Συχνότητες n_i	x'_i	nx'	nx^2	θ_i
0,9 – 2,2	1,55	285	-3	-855	2565	271
2,2 – 3,5	2,85	63	-2	-126	252	71
3,5 – 4,8	4,15	30	-1	-30	30	33
4,8 – 6,1	5,45	22	0	0	0	19
6,1 – 7,4	6,75	9	1	9	9	12
7,4 – 8,7	8,05	6	2	12	24	8
8,7 – 10,0	9,35	5	3	15	45	6
ΣΥΝΟΛΑ		420		T₁ = -975	T₂ = 2925	420

Το κέντρο της τέταρτης κλάσης ορίστηκε ως προσωρινή μέση τιμή ($m = 5,45$, βλ. εντονη (bold) γραφή). Η τελευταία στήλη περιέχει τις θεωρητικές συχνότητες, δηλαδή τα στοιχεία που θα προέκυπταν αν ίσχυε η σχέση (10), με μικρές προσαρμογές ώστε να προκύπτει άθροισμα 420 και να μπορέσει να εφαρμοστεί έπειτα η δοκιμασία X^2 . Τίθεται και πάλι το ερώτημα αν η σ.π.π. της τυχαίας μεταβλητής «Χρόνος αντίδρασης» μπορεί να έχει τη μορφή (10), γιατί οι συχνότητες έχουν μια «απότομα» φθίνουσα μορφή.

Απάντηση: Αρχικά και με μέγιστη τιμή της τμ $X =$ Χρόνος αντίδρασης βρέθηκε η προβλεπόμενη για την σ.π.π. (11) ελάχιστη τιμή της X . Αυτή δίνεται από την (12) και προέκυψε $\varepsilon = 0,9091$. Ήδη αυτό είναι μια καλή αρχική διαπίστωση ώστε να θεωρηθεί η (11) ως σ.π.π. για την τμ X που μελετάται, επειδή $X_{min} = 0,9 = \varepsilon$ παρατηρούμενο. Η μέση τιμή που προκύπτει από τα πειραματικά δεδομένα είναι:

$$\bar{x} = m + w \cdot T_1/n = 5,45 + 1,3 \cdot (-975)/420 = 2,4321$$

Υπολογίζοντας τη μέση τιμή από την (12) προκύπτει:

$$EX = 2,3979$$

Από τον Πίνακα 3 προκύπτει η διασπορά:

$$s^2 = \frac{w^2}{n-1} \left(T_2 - \frac{T_1^2}{n} \right) = 2,6685$$

Από την (13) προκύπτει αντίστοιχα η τιμή:

$$\text{Var}X = \sigma^2 = \frac{\beta^2 \cdot \{\ln(\beta+1)\sqrt{\beta+1}\}^2}{\beta+1} = 3,3410.$$

Άρα:

$$Cv = \frac{\sqrt{Var\bar{X}}}{E\bar{X}} = 0,7623$$

με τιμή από το δείγμα:

$$\hat{Cv} = \frac{\sqrt{2,6685}}{2,4321} = 0,6717.$$

Από την (15) προκύπτει η παράμετρος $q = ICv^2 = 1,7210$ και από το δείγμα η πειραματική της τιμή:

$$q = I\hat{Cv}^2 = \frac{\bar{x}^2}{s^2} = \frac{2,4321^2}{2,6685} = 2,2166$$

Στον Πίνακα 4 παρουσιάζονται τα αποτελέσματα από το δείγμα των 420 ατομών μαζί με τα αντίστοιχα θεωρητικά και προκύπτουν συγκριτικά στοιχεία.

Πίνακας 4. Σύγκριση δειγματικών και θεωρητικών μεγεθών παραδείγματος 2

	Μέση τιμή	Διασπορά	Τυπ. Αποκλ.	ΣΜ	q
Δείγμα	2,3979	3,3410	1,8278	0,7623	1,7210
Θεωρία	2,4321	2,6685	1,6336	0,6717	2,2166

Παρατηρούνται μικρές αποκλίσεις μεταξύ θεωρητικών και πειραματικών τιμών. Χρειάζεται όμως και η σχετική δοκιμασία για να απορριφθεί ή όχι η υπόθεση ότι τα πειραματικά δεδομένα προέρχονται από την κατανομή με τη μορφή (3) και εκθέτη -2 . Έγχε και δοκιμασία X^2 με υποθέσεις:

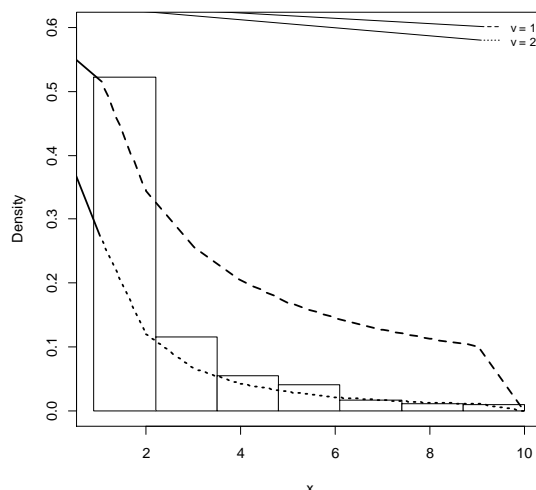
H_0 : Οι παρατηρούμενες πειραματικές τιμές προσαρμόζονται καλά στις θεωρητικές

H_1 : \neq η H_0

Οι πειραματικές και θεωρητικές τιμές δίνονται στις στήλες 3 και 7 του Πίνακα 3. Η τιμή του στατιστικού $X^2 = 4,17 < X_{6,0.05}^2 = 12,5916$, συνεπώς δεν υπάρχουν επαρκή στοιχεία για την απόρριψη της μηδενικής υπόθεσης σε επίπεδο σημαντικότητας 0.05.

Όπως αναφέρθηκε παραπάνω, το ερώτημα αν η σ.π.π. της τυχαίας μεταβλητής «Χρόνος αντίδρασης» μπορεί να έχει τη μορφή (10) γεννάται από το γεγονός ότι οι συχνότητες έχουν μια «απότομα» φθίνουσα μορφή. Στο σχήμα 2, που ακολουθεί, η κοινή απεικόνιση και προσαρμογή των συναρτήσεων (4) και (10) καθιστά φανερή την καλή προσαρμογή της συνάρτησης της μορφής (10).

Σχήμα 2. Προσαρμογή των συναρτήσεων (4) και (10) στην κατανομή των τιμών του παραδείγματος 2



4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στον χώρο των κατανομών υπαισέρχεται μια νέα συνάρτηση συνεχούς τ.μ. $X \in [\varepsilon, \beta]$, η οποία έχει μορφή μονωνύμου με εκθέτη αρνητικό ακέραιο αρνητικό εκθέτη φυσικό αριθμό. Η δειγματοληψία είναι το κλειδί που μας δίνει αρχικά αρκετά στοιχεία χρήσιμα στο να αξιοποιηθεί και η υπόλοιπη πληροφορία που βρίσκεται στα δεδομένα. Το β προσδιορίζεται κατάλληλα από το δείγμα δ και το ε εκφράζεται συναρτήσει του β . Παρατηρείται ότι για $\nu = 1$, η μέση τιμή και η διασπορά της σ.π.π. εξαρτώνται από το β , ενώ ο συντελεστής μεταβλητότητας είναι ανεξάρτητος του β , ενώ για $\nu \geq 2$, όλες οι παράμετροι εξαρτώνται από το β . Προτείνεται μία γρήγορη και εύκολη μέθοδος να εξετάζεται εάν δειγματικά δεδομένα μπορούν να περιγραφούν από τη σ.π.π. που δίνεται από τη σχέση (3) και προκύπτει, μέσω ελέγχων καλής προσαρμογής, ότι οι παρατηρούμενες πειραματικές τιμές προσαρμόζονται καλά στις θεωρητικές. Η προτεινόμενη μέθοδος θα μπορούσε να χαρακτηριστεί και ως δειγματοληψία υψηλού προϋπολογισμού (*high-budget sampling*), λόγω του μεγάλου μεγέθους δείγματος.

ABSTRACT

Sampling can help researchers to study the distributions of the various random variables and produce very satisfying results with low computational cost. The probability density function (pdf) of a random variable (rv) X can be estimated via sampling and various parameters of the rv X , e.g. the mean, the variance, the coefficient of variation, the range, etc. and then with

knowledge of the pdf, further statistical problems can be tackled. This procedure is both theoretical and empirical, based on sample data. In applied statistics, an initial hypothesis is made upon for a particular situation about the form of the pdf of X or an approximation of this pdf, that can be efficient and easy to be managed by researchers. The choice of the type of the suitable pdf and the time it takes to process the data in order to construct the pdf is vital. In the present paper we study how to get the suitable estimator from the sample, when the pdf has a polynomial form with a negative exponent. Illustrative examples are given to highlight the basic results of the theoretical approach.

Key words: Sampling, Probability density function, Coefficient of variation, Mean, Variance, Range.

AMS Classification: 62D05, 62E17

ΑΝΑΦΟΡΕΣ

- Κολυβά-Μαχαίρα Φ., Μπόρα-Σέντα Ε. (2013). *Στατιστική: Θεωρία και Εφαρμογές*. Β' Έκδοση. Εκδόσεις Ζήτη, Θεσσαλονίκη.
- Μουσιάδης, Π. (2000). *Ανώτερα Μαθηματικά*, Εκδόσεις Α&Π Χριστοδουλίδη Ο.Ε., Θεσσαλονίκη.
- Παπατσούμα, Ι., Φαρμάκης, Ν., (2015). Πολυωνυμική έκφραση συμμετρικών κατανομών: Η περίπτωση σ.π.π. τριγωνομετρικής μορφής, *Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής*, 282-294.
- Παπατσούμα, Ι., Φαρμάκης, Ν. (2017). Μελέτη αποκομμένων κατανομών: Η περίπτωση της συμμετρικής κατανομής. *Πρακτικά 9^{ου} Πανελληνίου Συνεδρίου Ανάλυσης Δεδομένων*, 167-174.
- Φαρμάκης Ν. (2001). *ΣΤΑΤΙΣΤΙΚΗ, Περιληπτική Θεωρία-Ασκήσεις*. Β' Έκδοση. Εκδόσεις Α&Π Χριστοδουλίδη, Θεσσαλονίκη.
- Φαρμάκης Ν. (2016). *Εισαγωγή στη Δειγματοληψία*. Αφοί Κυριακίδη Εκδόσεις Α.Ε., Θεσσαλονίκη.
- Balakrishnan, N., Nevzorov, V.B. (2003). *A Primer on Statistical Distributions*. Hoboken, N.J. : Wiley. DOI:10.1002/0471722227.
- Hollands, K.G.T, Suehrcke, H. (2013). A three-state model for the probability distribution of instantaneous solar radiation, with applications. *Solar Energy*, 96: 103–112. DOI: 10.1016/j.solener.2013. 07.007.
- Munkhammar, J., Mattsson, L., Rydén, J. (2017). Polynomial probability distribution estimation using the method of moments. *PLoS ONE*, 12(4): e0174573. DOI: 10.1371/ journal.pone.0174573.
- Munkhammar, J., Rydén, J., Widén, J. (2014). Characterizing probability density distributions for household electricity load profiles from high-resolution electricity use data. *Applied Energy*, 135: 382–390. DOI: 10.1016/j.apenergy.2014.08.093 3.

Papatsouma I., Farmakis, N. (2020). Approximating Symmetric Distributions via Sampling and Coefficient of Variation. *Communications in Statistics – Theory and Methods*, 49:1, 61–77, DOI: 10.1080/03610926.2018.1529244.



Η ΤΡΑΠΕΖΙΚΗ ΧΡΗΜΑΤΟΔΟΤΗΣΗ ΤΗΣ ΠΑΡΑΓΩΓΗΣ ΣΤΗΝ ΔΙΑΡΚΕΙΑ ΤΩΝ ΜΝΗΜΟΝΙΩΝ: ΤΑΣΕΙΣ, ΕΠΟΧΙΚΟΤΗΤΑ, ΜΕΣΟΧΡΟΝΙΕΣ ΑΥΞΟΜΕΙΩΣΕΙΣ ΚΑΤΑ ΚΛΑΔΟ (ΙΟΥΝ. 2010 – ΔΕΚ.2018)

Πρόδρομος Προδρομίδης

Κέντρο Προγραμματισμού και Οικονομικών Ερευνών, pjprodr@kepe.gr

ΠΕΡΙΛΗΨΗ

Εξετάζεται η πορεία της εγχώριας τραπεζικής χρηματοδότησης στους 13 κλάδους της οικονομίας μεταξύ Ιουνίου 2010 και Δεκεμβρίου 2018. Προς τούτο αξιοποιούνται μηνιαία στοιχεία της Τραπέζης της Ελλάδος και απλά οικονομετρικά εργαλεία εκτιμήσεως της τάσης, της εποχικότητας, των ενδιάμεσων περιόδων. Αν και η διαχρονική πορεία της χρηματοδότησης συνολικά συνάδει με την πορεία των καταθέσεων, η τάση ανά κλάδο ποικίλει, εποχικότητα εμφανίζει μόνον ο τουρισμός, οι μεσοχρόνιες περίοδοι (κύκλοι) αποκλίσεων από την κάθε κλαδική τάση είναι ασύγχρονες. Ο προσδιορισμός της ενάρξεως και της χρονικής διάρκειας των μεσοχρονίων αυξήσεων ή μειώσεων διευκολύνει τον εντοπισμό των συμβάντων που ίσως προξένησαν τις αυξήσεις ή μειώσεις, συμβάλλοντας έτσι στην κατανόηση (α) των πραγμάτων/καταστάσεων που επηρεάζουν την χρηματοδότηση της οικονομικής δραστηριότητας ή/και (β) της διαφοροποιημένης κλαδικής αντιδράσεως σε παρεμφερή γεγονότα.

Λέξεις Κλειδιά: τραπεζική χρηματοδότηση, κλάδοι παραγωγής, διαφοροποιημένες τάσεις, εποχικότητα, μεσοχρόνιες αυξομειώσεις.

1. ΕΙΣΑΓΩΓΗ

Το άρθρο εξετάζει την εξέλιξη της τραπεζικής χρηματοδότησης των επιμέρους κλάδων παραγωγής στην χώρα μας από τον Ιούνιο του 2010, πρώτο μήνα εφαρμογής του, λεγομένου, *πρώτου μνημονίου* της Ελλάδος με τους δανειστές της (Ευρωπαϊκή Ένωση, Διεθνές Νομισματικό Ταμείο), έως το τέλος του 2018, οπότε ψηφίστηκε από την Βουλή των Ελλήνων ο πρώτος μεταμνημονιακός ετήσιος προϋπολογισμός.

Σύμφωνα με την οικονομική θεωρία, η διάθεση από τις τράπεζες κεφαλαίων στους παραγωγούς και η χρήση τραπεζικών κεφαλαίων από τους παραγωγούς εντάσσεται σε ένα ευρύτερο σύστημα προσφοράς και ζήτησης κεφαλαίων: (α) Προσφοράς από τις χρηματοπιστωτικές αγορές (μέσω των οποίων οι αποταμιευτές απευθείας παρέχουν κεφάλαια στους δανειζομένους έναντι ομολόγων και μετοχών), τους χρηματοπι-

στωτικούς διαμεσολαβητές (δηλ., τις τράπεζες και τα αμοιβαία κεφάλαια που μεσολαβούν στην παροχή κεφαλαίων προερχομένων από τους αποταμιευτές προς τους δανειζομένους), καθώς και άλλους (λ.χ., πιστωτικές ενώσεις, συνταξιοδοτικά ταμεία, ασφαλιστικές εταιρείες, αργυραμοιβούς), βάσει της εκδιδόμενης, κατά καιρούς, ποσότητας χρήματος από τις νομισματικές αρχές. (β) Ζήτησης από καταναλωτές, παραγωγούς και την κυβέρνηση-αυτοδιοίκηση, με τις αυξήσεις (μειώσεις) των φόρων στα εισοδήματα να μειώνουν (αυξάνουν) την προσφορά και τις φοροελαφρύνσεις για νέες επενδύσεις να αυξάνουν την ζήτηση. (Βλ. ενδεικτικά, Mankiw, 2002.)

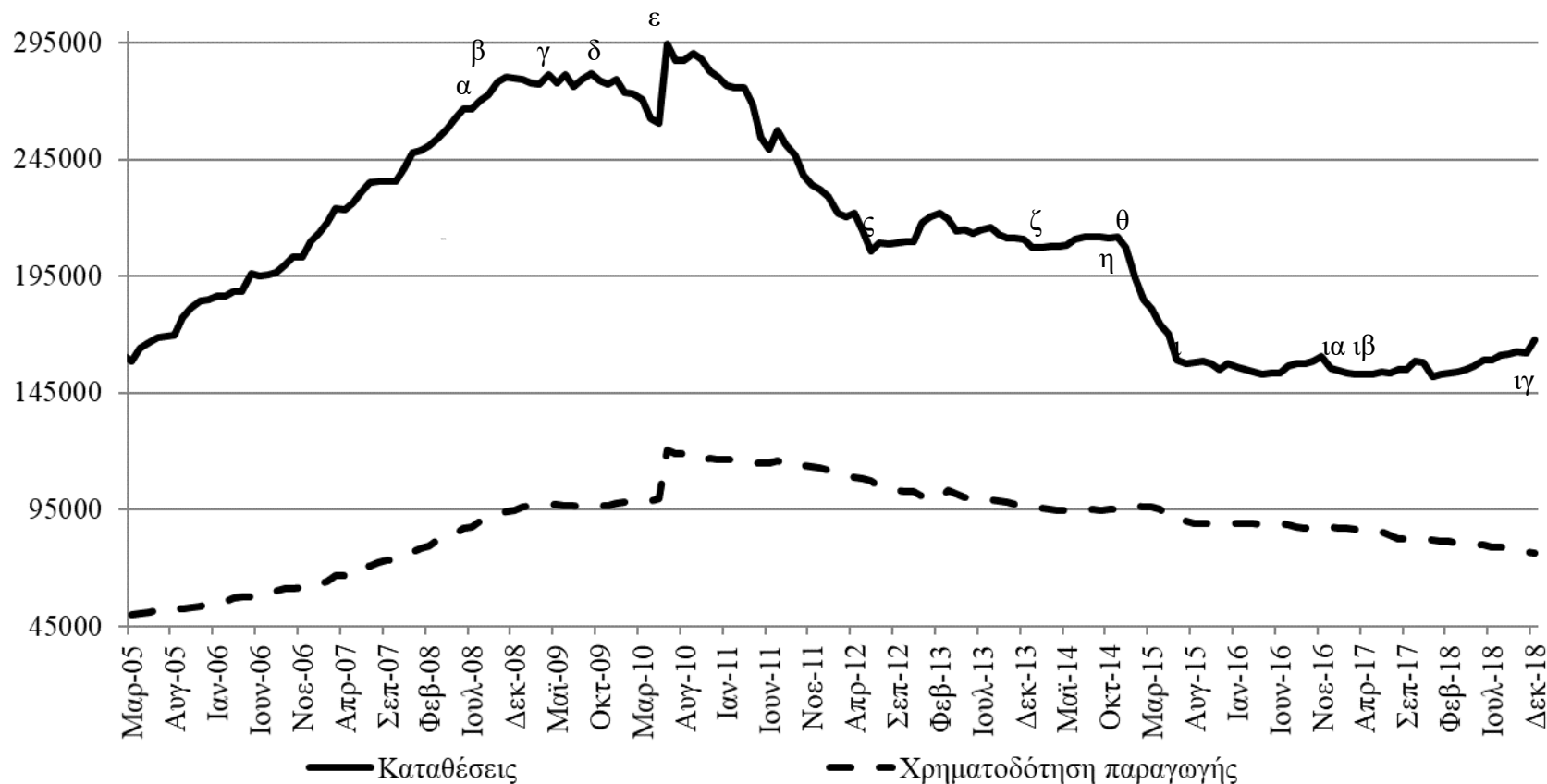
Η πρόσφατη αρθρογραφία με αντικείμενο την τραπεζική χρηματοδότηση μεμονωμένων κλάδων (βλ. ενδεικτικά Ahmed και Rashid, 2012· Chisasa και Makina, 2012· Chattoraj κ.ά., 2013· Milić και Soleša, 2017· Katan και Recio Espejo, 2018), καθώς και η περιγραφή της μειώσεώς της σε όλους τους κλάδους στην Πορτογαλία μετά την διεθνή χρηματοπιστωτική κρίση του 2007 (Rodrigues, 2016), φανερώνει ότι πρόκειται για ένα διεθνούς ενδιαφέροντος πολυδιάστατο θέμα. Επίσης, γεννά ερωτήματα για το τι συνέβη στην Ελλάδα στο ίδιο διάστημα. Έτσι, στις σελίδες που ακολουθούν, επιχειρείται η διερεύνηση των εξελίξεων στην τραπεζική χρηματοδότηση όχι σε έναν, αλλά σε όλους τους κλάδους παραγωγής στην χώρα μας, στην διάρκεια της πολυετούς υφέσεως που προέκυψε μετά την κρίση του 2007, με μηνιαία στοιχεία της Τραπεζής της Ελλάδος (ΤτΕ) και σχετικά απλά, ευρέως αποδεκτά, στατιστικά εργαλεία.

2. Η ΕΞΕΛΙΞΗ ΤΩΝ ΚΑΤΑΘΕΣΕΩΝ ΚΑΙ ΤΗΣ ΤΡΑΠΕΖΙΚΗΣ ΧΡΗΜΑΤΟΔΟΤΗΣΗΣ ΤΗΣ ΕΓΧΩΡΙΟΥ ΠΑΡΑΓΩΓΗΣ

Όπως φαίνεται στο Διάγραμμα 1, οι καταθέσεις (αποταμιεύσεις) κεφαλαίων στις εγχώριες τράπεζες: (α) Γενικώς κινήθηκαν ανοδικώς ως και τον Νοέμβριο του 2008, οπότε έγινε φανερό ότι η διεθνής χρηματοοικονομική κρίση άρχισε να επηρεάζει τις δυτικοευρωπαϊκές οικονομίες. Πιο συγκεκριμένα, από εντάξεως της Ελλάδος στην Ευρωζώνη (Ιανουάριο του 2001) ως και τον Νοέμβριο του 2008, οι περιστασιακές μειώσεις καταθέσεων για έναν μήνα ή δύο διαδοχικούς μήνες ήσαν σχετικά σπάνιες: κατά μέσον όρο μία (μηνιαία) μείωση το εξάμηνο. (β) Μειώθηκαν για τέσσερις διαδοχικούς μήνες από τον Δεκέμβριο του 2008 και ακολούθησαν αυξομειώσεις, όμως γενικώς κινήθηκαν καθοδικώς ως τον Δεκέμβριο του 2017. Η πρώτη μεγάλη μείωση σημειώθηκε μεταξύ Ιανουαρίου και Μαΐου του 2010: από ψηφίσεως του προϋπολογισμού της κυβερνήσεως Γ. Παπανδρέου και την λήψη μέτρων, ως την υπογραφή του πρώτου μνημονίου (ακριβέστερα: των πρώτων μνημονίων με τους δανειστές) και την ψήφιση νέων, συναφών μέτρων. Η δεύτερη μεγάλη μείωση σημειώθηκε μεταξύ Οκτωβρίου του 2010 και Ιουνίου του 2012: από καταρτίσεως του πρώτου μνημονιακού προϋπολογισμού της κυβερνήσεως Γ. Παπανδρέου, ως το σχηματισμό της κυβερνήσεως Σαμαρά. Η τρίτη μεγάλη μείωση σημειώθηκε μεταξύ Δεκεμβρίου 2014 και Ιουλίου 2015: από την αποτυχία εκλογής Προέδρου της Δημοκρατίας και την προκήρυξη εκλογών ως το κλείσιμο των τραπεζών και τον περιορισμό της κινήσεως κεφαλαίων —περιλαμβανομένων των αναλήψεων. (γ) Αυξήθηκαν για έξι διαδοχικούς μήνες από τον Ιανουάριο ως και τον Ιούνιο του 2018 (παράλληλα με την ανάκαμψη

Γράφημα 1

Καταθέσεις μη νομισματικών/χρηματοπιστωτικών ιδρυμάτων (NXI) σε NXI και χρηματοδοτήσεις των τελευταίων προς εγχώριες μη NX επιχειρήσεις, Μάρτιος 2005 - Δεκέμβριος 2018 (σε εκατομ. ευρώ)



Σημείωση: Στα NXI δεν υπολογίζεται η ΤτΕ. Από τον Δεκ. 2016 δεν περιλαμβάνονται τα στοιχεία του Ταμείου Παρακαταθηκών και Δανείων.
 Χρονολόγιο: ^α Πρώτη μείωση του πραγ. ΑΕΠ (2^ο τρίμ. 2008). ^β Εκλογές. Σχηματισμός νέας κυβ/σεως Κ. Καραμανλή. Στις ΗΠΑ η τράπεζα Lehman Brothers πτωχεύει, ακολουθούν άλλες (Σεπ. 2008). Ευρωπαϊκές τράπεζες πτωχεύουν. Αρκετές κυβ/σεις παρεμβαίνουν να τις σώσουν (Οκτ. 2008). Το ελληνικό κοινοβούλιο ψηφίζει πακέτο στήριξης των τραπεζών. (Νοε. 2008). ^γ 1^η μείωση του ονομ. ΑΕΠ (1^ο τρίμ. 2009). ^δ Εκλογές. Σχηματισμός κυβ/σεως Γ. Παπανδρέου. Αποκάλυψη ελλείμματος (Οκτ. 2009). ^ε Υπογραφή πρώτων μνημονίων. 3^ο πακέτο μέτρων λιτότητας κυβ/σεως Γ. Παπανδρέου (Μάιος 2010). ^ς Εκλογές. Σχηματισμός κυβ/σεως Α. Σαμαρά (Ιούν. 2012). ^ζ 1^η από 6 διαδοχικές αυξήσεις πραγ. ΑΕΠ (1^ο τρίμ. 2014). ^η Διαδοχική αύξηση ονομ. ΑΕΠ (3-4^ο τρίμ. 2014). ^θ Αποτυχία εκλογής ΠτΔ. Προκήρυξη εκλογών (Δεκ. 2014). ^ι Οι τράπεζες κλείνουν για μια εβδομάδα. Ορίζεται όριο αναλήψεων. Δημοψήφισμα. Νέο μνημόνιο. Οι περιορισμοί στην κίνηση κεφαλαίων θα αρθούν σταδιακά (Ιουν.-Ιουλ. 2015). ^{ια} Διαδοχικές αυξήσεις ονομ. ΑΕΠ (1ο τρίμ. 2017). ^{ιβ} Διαδοχικές αυξήσεις πραγ. ΑΕΠ (2^ο τρίμ. 2017). ^{ιγ} Το ελληνικό κοινοβούλιο ψηφίζει τον πρώτο μεταμνημονιακό ετήσιο προϋπολογισμό μετά το 2009 (Δεκ. 2018).

Πηγή στατιστικών στοιχείων: ΤτΕ

του ΑΕΠ) και γενικώς κινήθηκαν ανοδικώς ως το τέλος του υπό εξέταση χρονικού διαστήματος. Οι μηνιαίες μειώσεις κατέστησαν πάλι σπάνιες: κατά μέσον όρο μία μείωση το εξάμηνο.

Η εξέλιξη της χρηματοδότησης της παραγωγής από τις εγχώριες τράπεζες ομοιάζει αρκετά με την εξέλιξη των καταθέσεων· ενώ η διαφορά μεταξύ των καταθέσεων και της χρηματοδότησης της παραγωγής ανάγεται αφενός στην διακράτηση ρευστών διαθεσίμων από τις τράπεζες, αφετέρου στην χρηματοδότηση από τις τράπεζες της καταναλώσεως ή άλλων επιδιώξεων.

Στην συνέχεια αναλύεται η εξέλιξη της τραπεζικής χρηματοδότησης της εγχώριας παραγωγής ανά κλάδο παραγωγής και απομονώνεται η μακροχρόνια τάση από την εποχικότητα και από τις ενδιάμεσες (μεσοχρόνιες) αυξομειώσεις.

3. ΑΝΑΛΥΣΗ ΤΗΣ ΧΡΗΜΑΤΟΔΟΤΗΣΗΣ ΠΡΟΣ ΤΙΣ ΕΓΧΩΡΙΕΣ ΕΠΙΧΕΙΡΗΣΕΙΣ ΑΝΑ ΚΛΑΔΟ ΠΑΡΑΓΩΓΗΣ

Η ανάλυση διενεργείται στην βάση 103 μηνιαίων παρατηρήσεων τραπεζικής χρηματοδότησης για κάθε έναν εκ των δεκατριών κλάδων παραγωγής της χώρας, πλην του νομισματικού/χρηματοπιστωτικού κλάδου που παρέχει την χρηματοδότηση. Πιο συγκεκριμένα: επιχειρείται η οικονομετρική εκτίμηση και η απομόνωση των επιδράσεων της μακροχρονίου τάσεως (T), της όποιας βραχυχρονίου (μηνιαίας) εποχικότητας (E), αν υφίστανται, και των μεσοχρονίων διαφοροποιήσεων (M): (α) στην γεωργία-δασοκομία-αλιεία, (β) στα ορυχεία-λατομεία, (γ) την μεταποίηση, (δ) στην παροχή ηλεκτρισμού-φωταερίου-ύδρευσης, (ε) στις κατασκευές, (ς) στο εμπόριο, (ζ) στην παροχή καταλύματος-εστίασεως, (η) στην αποθήκευση και τις μη-πλωτές μεταφορές, (θ) στις πλωτές μεταφορές (ι) στην επικοινωνία-ενημέρωση, (ια) στην διαχείριση ακίνητης περιουσίας, (ιβ) στις επαγγελματικές-επιστημονικές-τεχνικές-διοικητικές-υποστηρικτικές δραστηριότητες και (ιγ) στους λοιπούς κλάδους.

Προς τούτο, χρησιμοποιείται μια παραλλαγή της συναρτησιακής μορφής που περιγράφουν/απεικονίζουν οι Smith και Duncan (1944), Fox (1968), Τζωρτζόπουλος (1985), Black (1992), Cameron (2005) και άλλοι:

$$Y = \beta_0 + \beta_1 \times T + \beta_2 \times T^2 + \sum \beta_{3i} \times M_i + \sum \beta_{4j} \times E_j + u, \quad (1)$$

όπου u ο λευκός θόρυβος των μεμονωμένων (μη διαδοχικών) ή πολύ μικρών αυξο-

μειώσεων.

Η διαδικασία εκτιμήσεως των παραμέτρων έχει ως εξής: Διενεργείται μια προκαταρκτική παλινδρόμηση με τα στοιχεία της τάσης και της εποχικότητας. Διατηρούνται τα T^2 και E_j με τις πλέον διαφοροποιημένες παραμέτρους ώστε να εξοικονομηθούν βαθμοί ελευθερίας. (Διαπιστώνεται ότι η απαλοιφή των περιπτώσεων με $p\text{-value} < 1\%$ επιφέρει τις μικρότερες απώλειες στο adj-R^2 .) Διενεργείται νέα παλινδρόμηση βάσει της οποίας εκτιμώνται οι τιμές των μηνιαίων καταλοίπων. Οι διαδοχικές εξάρσεις ή βυθίσεις των τελευταίων αντιστοιχίζονται σε νέες κατηγορηματικές μεταβλητές. Με τον τρόπο αυτό προσδιορίζονται δυνητικές μεσοχρόνιες περιόδους. Με διαδοχικές δοκιμαστικές αντικαταστάσεις δυαδικών τιμών στις συγκεντρώσεις των νέων αυτών μεταβλητών, καθώς και παλινδρομήσεις (επανεκτιμήσεις της σχέσεως), εξετάζεται η βελτίωση ή μη του R^2 και προσδιορίζεται το εύρος και ο συνδυασμός των

περιόδων που ταιριάζουν καλύτερα στην σχέση (δηλ., περιγράφουν καλύτερα την σχέση). Τελικώς, προκύπτει μια εκδοχή της σχέσεως (1), με στοιχεία τάσεως, εποχικότητας και κάποιες μεσοχρόνιες περιόδους, της οποίας ο αριθμός των ανεξαρτήτων μεταβλητών και παρατηρήσεων συνάδει με τις προτάσεις του Harris (1985) (βλ. και Howell, 1997). Γενικώς, ο αριθμός των ανεξαρτήτων μεταβλητών ανέρχεται στο 5,8-12,6% του αριθμού των παρατηρήσεων και εξηγεί το 92,4-99,5% της μεταβλητότητας. Επισημαίνεται ότι σε όλες τις αναλύσεις που ακολουθούν (δεκατρείς εμπειρικές αναλύσεις) τα κατάλοιπα είναι στάσιμα.

Η οικονομετρική ανάλυση της πορείας της τραπεζικής χρηματοδότησης αναδεικνύει:

- Στην γεωργία κ.ά.: Αρνητική τάση με φθίνοντα ρυθμό, απουσία εποχικότητας, καθώς και εννέα μεσοχρόνιες περιόδους: τέσσερις με πιο μειωμένο δανεισμό και πέντε με πιο αυξημένο δανεισμό. (Βλ. Πίνακα 1.)
- Στα ορυχεία-λατομεία: Αρνητική τάση με φθίνοντα ρυθμό, απουσία εποχικότητας, καθώς και επτά μεσοχρόνιες περιόδους: τέσσερις με πιο μειωμένο δανεισμό και τρεις με πιο αυξημένο δανεισμό. (Βλ. Πίνακα 2.)
- Στην μεταποίηση: Αρνητική τάση με φθίνοντα ρυθμό, απουσία εποχικότητας, καθώς και πέντε μεσοχρόνιες περιόδους: δύο με πιο μειωμένο δανεισμό και τρεις με πιο αυξημένο δανεισμό. (Βλ. Πίνακα 3)

Στην παροχή ηλεκτρισμού κ.ά.: Θετική τάση με φθίνοντα ρυθμό (ο τελευταίος προκύπτει από την συνθήκη πρώτης τάξεως για τον εντοπισμό ακροτάτου, ήτοι την εξίσωση της πρώτης παραγώγου προς τον χρόνο με το μηδέν, και έχει ως αποτέλεσμα την εμφάνιση μεγίστου και, εν συνεχεία, αρνητικής τάσης), απουσία επο-

Πίνακας 1			
Οικονομετρική ανάλυση με ανθεκτικούς εκτιμητές της πορείας της τραπεζικής χρηματοδότησης στην γεωργία-δασοκομία-αλιεία, Ιουν. 2010 – Δεκ. 2018 (σε εκατ. €)			
Ανεξάρτητες μεταβλητές	συντελεστές	p-values	
1 Σταθερός όρος	2.199,8	0	
2 Τάση	-18,4	0	
3 Τάση στο τετράγωνο (αποτυπώνει τον ρυθμό μεταβολής)	0,1	0	
Μεσοχρόνιες περίοδοι			
4 Αύγουστος – Σεπτέμβριος 2010	36,9	0,003	
5 Οκτώβριος 2010 – Οκτώβριος 2011	-49,2	0	
6 Μάρτιος 2011 – Ιούνιος 2012	90,8	0	
7 Ιούλιος 2012 – Σεπτέμβριος 2013	-239,9	0	
8 Δεκέμβριος 2013 – Μάιος 2014	49,4	0	
9 Αύγουστος 2014 – Ιούλιος 2016	53,5	0	
10 Δεκέμβριος 2016 – Φεβρουάριος 2017	54,0	0	
11 Ιούλιος – Νοέμβριος 2017	-69,3	0	
12 Απρίλιος – Σεπτέμβριος 2018	-55,9	0	
$R^2 = 99,12\%$, $adj. R^2 = 99,02\%$. $N = 103$.			
Τα κατάλοιπα είναι στάσιμα σε επίπεδο σημαντικότητας 1‰.			

Πίνακας 4

Οικονομετρική ανάλυση με ανθεκτικούς εκτιμητές της πορείας της τραπεζικής χρηματοδότησης σε ηλεκτρισμό-φωταέριο-ύδρευση, Ιουν. 2010 – Δεκ. 2018 (σε εκατ. €)

Ανεξάρτητες μεταβλητές	συντελεστές	p-values
1 Σταθερός όρος	4.719,6	0
2 Τάση	25,0	0
3 Τάση στο τετράγωνο (αποτυπώνει τον ρυθμό μεταβολής)	-0,3	0
Μεσοχρόνιες περιόδους		
4 Απρίλιος 2011 – Δεκέμβριος 2012		
5 Μάρτιος 2014 – Οκτώβριος 2016	-375,0	0
6 Δεκέμβριος 2016 – Μάρτιος 2017	-253,8	0
7 Ιούλιος – Αύγουστος 2018	-354,8	0

$R^2 = 92,36\%$, $adj. R^2 = 91,89\%$. $N = 103$.

Τα κατάλοιπα είναι στάσιμα σε επίπεδο σημαντικότητας 1‰.

Πίνακας 5

Οικονομετρική ανάλυση με ανθεκτικούς εκτιμητές της πορείας της τραπεζικής χρηματοδότησης στις κατασκευές, Ιουν. 2010 – Δεκ. 2018 (σε εκατ. €)

Ανεξάρτητες μεταβλητές	συντελεστές	p-values
1 Σταθερός όρος	10.854,7	0
2 Τάση	7,2	0
3 Τάση στο τετράγωνο (αποτυπώνει τον ρυθμό μεταβολής)	-0,3	0
Μεσοχρόνιες περιόδους		
4 Οκτώβριος 2010 – Οκτώβριος 2011	337,3	0
5 Δεκέμβριος 2011 – Φεβρουάριος 2013	-412,6	0
6 Απρίλιος – Ιούλιος 2013	-146,4	0,004
7 Οκτώβριος – Δεκέμβριος 2013	122,7	0,001
8 Φεβρουάριος – Αύγουστος 2014	-155,6	0
9 Οκτώβριος 2014 – Ιούνιος 2016	203,8	0
10 Οκτώβριος 2015 – Φεβρουάριος 2017	247,8	0
11 Μάιος – Ιούνιος 2017	126,4	0
12 Αύγουστος – Δεκέμβριος 2017	-115,0	0,001
13 Μάρτιος – Απρίλιος 2018	-102,7	0
14 Ιούνιος – Δεκέμβριος 2018	-237,5	0,003

$R^2 = 98,41\%$, $adj. R^2 = 98,18\%$. $N = 103$.

Τα κατάλοιπα είναι στάσιμα σε επίπεδο σημαντικότητας 1‰.

χικότητας, καθώς και τέσσερις μεσοχρόνιες μεσοχρόνιες περιόδους: τρεις με πιο μειωμένο δανεισμό και μια με πιο αυξημένο δανεισμό. (Βλ. Πίνακα 4.)

- Στις κατασκευές: Θετική τάση με φθίνοντα ρυθμό (ο ρυθμός έχει ως αποτέλεσμα την εμφάνιση μεγίστου σχετικά σύντομα και, εν συνεχεία, αρνητικής τάσης), απουσία εποχικότητας, καθώς και ένδεκα μεσοχρόνιες περιόδους: έξι με πιο μειω-

- ωμένο δανεισμό και πέντε με πιο αυξημένο δανεισμό. (Βλ. Πίνακα 5.)
- Στο εμπόριο: Αρνητική τάση με φθίνοντα ρυθμό, απουσία εποχικότητας, καθώς και οκτώ μεσοχρόνιες περιόδους: πέντε με πιο μειωμένο δανεισμό και τρεις με πιο αυξημένο δανεισμό. (Βλ. Πίνακα 6.)
 - Στην παροχή καταλύματος-εστίασης: Θετική τάση με φθίνοντα ρυθμό που έχει ως αποτέλεσμα την εμφάνιση μεγίστου και, εν συνεχεία, αρνητικής τάσης αύξηση κατά τον Μάρτιο και Ιούνιο (αντανακλώντας την προετοιμασία ή την εποχικώς αυξημένη δραστηριότητα των τουριστικών επιχειρήσεων), καθώς και εννέα μεσοχρόνιες περιόδους: τέσσερις με πιο μειωμένο δανεισμό και πέντε με πιο αυξημένο δανεισμό. (Βλ. Πίνακα 7.)
 - Στην αποθήκευση και τις μη-πλωτές μεταφορές: Αρνητική τάση με φθίνοντα ρυθμό που έχει ως αποτέλεσμα την εμφάνιση ελαχίστου και, εν συνεχεία, θετικής τάσης, απουσία εποχικότητας, καθώς και επτά μεσοχρόνιες περιόδους: πέντε με πιο μειωμένο δανεισμό και δύο με πιο αυξημένο δανεισμό. (Βλ. Πίνακα 8.)
 - Στις πλωτές μεταφορές: Αρνητική τάση με φθίνοντα ρυθμό που έχει ως αποτέλεσμα την εμφάνιση ελαχίστου προς το τέλος του υπό εξέταση χρονικού διαστήματος, απουσία εποχικότητας, καθώς και επτά μεσοχρόνιες περιόδους: πέντε με πιο μειωμένο δανεισμό και δύο με πιο αυξημένο δανεισμό. (Βλ. Πίν. 9.)
 - Στην επικοινωνία-ενημέρωση: Θετική τάση με φθίνοντα ρυθμό που έχει ως αποτέλεσμα την εμφάνιση μεγίστου σχετικά σύντομα και, εν συνεχεία, αρνητικής τάσης, απουσία εποχικότητας, καθώς και ένδεκα μεσοχρόνιες περιόδους: έξι με πιο μειωμένο δανεισμό και πέντε με πιο αυξημένο δανεισμό. (Βλ. Πίνακα 10.)
- Στην διαχείριση της ακίνητης περιουσίας: Θετική τάση με φθίνοντα ρυθμό που έχει ως αποτέλεσμα την εμφάνιση μεγίστου και, εν συνεχεία, αρνητικής τάσης, α-

Πίνακας 6

Οικονομετρική ανάλυση με ανθεκτικούς εκτιμητές της πορείας της τραπεζικής χρηματοδότησης στο εμπόριο, Ιουν. 2010 – Δεκ. 2018 (σε εκατ. €)

Ανεξάρτητες μεταβλητές	συντελεστές	p-values
1 Σταθερός όρος	26.963,2	0
2 Τάση	-171,9	0
3 Τάση στο τετράγωνο (αποτυπώνει τον ρυθμό μεταβολής)	0,8	0
Μεσοχρόνιες περίοδοι		
4 Οκτώβριος 2010 – Φεβρουάριος 2012	-422,1	0
5 Απρίλιος 2011 – Ιούνιος 2012	488,9	0
6 Ιούλιος 2012 – Φεβρουάριος 2013	-277,0	0
7 Μάρτιος – Απρίλιος 2013	182,0	0
8 Οκτώβριος 2013 – Μάιος 2015	-497,6	0
9 Ιούλιος – Δεκέμβριος 2015	-342,4	0
10 Ιανουάριος 2016 – Μάρτιος 2018	333,4	0
11 Μάιος – Δεκέμβριος 2018	-447,9	0,008

$R^2 = 99,43\%$, $adj. R^2 = 99,37\%$. $N = 103$.

Τα κατάλοιπα είναι στάσιμα σε επίπεδο σημαντικότητας 1‰.

Πίνακας 7

Οικονομετρική ανάλυση με ανθεκτικούς εκτιμητές της πορείας της τραπεζικής χρηματοδότησης σε καταλύματα- εστίαση, Ιουν. 2010 – Δεκ. 2018 (σε εκατ. €)

Ανεξάρτητες μεταβλητές	συντελεστές	p-values
1 Σταθερός όρος	7251,7	0
2 Τάση	11,7	0
3 Τάση στο τετράγωνο (αποτυπώνει τον ρυθμό μεταβολής)	-0,1	0
Εποχικότητας		
4 Μάρτιος	97,3	0,002
5 Ιούνιος	73,6	0
Μεσοχρόνιες περιόδους		
6 Απρίλιος 2011 – Ιανουάριος 2012	-183,8	0
7 Ιούνιος – Οκτώβριος 2012	-145,9	0
8 Μάρτιος 2013 – Φεβρουάριος 2015	181,7	0
9 Απρίλιος – Ιούνιος 2015	197,0	0
10 Ιούλιος – Σεπτέμβριος 2015	-107,0	0
11 Ιανουάριος – Φεβρουάριος 2017	96,8	0
12 Απρίλιος – Μάιος 2017	113,7	0
13 Αύγουστος – Νοέμβριος 2017	-106,4	0
14 Απρίλιος – Σεπτέμβριος 2018	139,8	0

$R^2 = 99,55\%$, $adj. R^2 = 92,65\%$. $N = 103$.

Τα κατάλοιπα είναι στάσιμα σε επίπεδο σημαντικότητας 1‰.

Πίνακας 8

Οικονομετρική ανάλυση με ανθεκτικούς εκτιμητές της πορείας της τραπεζικής χρηματοδότησης στην αποθήκευση και τις μη-πλωτές μεταφορές, Ιουν. 2010 – Δεκ. 2018 (σε εκατ. €)

Ανεξάρτητες μεταβλητές	συντελεστές	p-values
1 Σταθερός όρος	2.159,2	0
2 Τάση	-40,0	0
3 Τάση στο τετράγωνο (αποτυπώνει τον ρυθμό μεταβολής)	0,3	0
Μεσοχρόνιες περιόδους		
4 Ιούνιος – Σεπτέμβριος 2010		
5 Φεβρουάριος 2011 – Μάρτιος 2012	282,3	0
6 Απρίλιος 2012 – Μάιος 2013	-144,2	0
7 Ιανουάριος – Νοέμβριος 2014	-94,4	0
8 Αύγουστος 2015 – Σεπτέμβριος 2016	-86,9	0
9 Νοέμβριος 2016 – Ιανουάριος 2018	175,7	0
10 Ιούνιος – Δεκέμβριος 2018	-110,5	0,003

$R^2 = 97,65\%$, $adj. R^2 = 97,42\%$. $N = 103$.

Τα κατάλοιπα είναι στάσιμα σε επίπεδο σημαντικότητας 1‰.

πουσία εποχικότητας, καθώς και εννέα μεσοχρόνιες περιόδους: πέντε με πιο μει-

- ωμένο δανεισμό και τέσσερις με πιο αυξημένο δανεισμό (Βλ. Πίνακα 11.)
- Στις επαγγελματικές κ.ά. δραστηριότητες: Αρνητική τάση, απουσία εποχικότητας, καθώς και δώδεκα μεσοχρόνιες περιόδους: έξι με πιο μειωμένο και ακόμα πιο μει-

Πίνακας 9		
Οικονομετρική ανάλυση με ανθεκτικούς εκτιμητές της πορείας της τραπεζικής χρηματοδότησης στις πλωτές μεταφορές, Ιουν. 2010 – Δεκ. 2018 (σε εκατ. €)		
Ανεξάρτητες μεταβλητές	συντελεστές	p-values
1 Σταθερός όρος	19.063,9	0
2 Τάση	-206,5	0
3 Τάση στο τετράγωνο (αποτυπώνει τον ρυθμό μεταβολής)	1,0	0
Μεσοχρόνιες περιόδους		
4 Σεπτέμβριος – Οκτώβριος 2010	-686,2	0,001
5 Ιούλιος 2011 – Μάιος 2012	1.646,0	0
6 Ιούνιος 2014 – Απρίλιος 2015	2.680,3	0
7 Μάιος 2015 – Οκτώβριος 2016	-383,4	0
8 Ιούλιος – Οκτώβριος 2017	-464,2	0
9 Δεκέμβριος 2017 – Απρίλιος 2018	-951,2	0
10 Μάιος – Δεκέμβριος 2018	-741,6	0

$R^2 = 98,59\%$, $adj. R^2 = 98,46\%$. $N = 103$.
Τα κατάλοιπα είναι στάσιμα σε επίπεδο σημαντικότητας 1‰.

Πίνακας 10		
Οικονομετρική ανάλυση με ανθεκτικούς εκτιμητές της πορείας της τραπεζικής χρηματοδότησης σε επικοινωνία-ενημέρωση, Ιουν. 2010 – Δεκ. 2018 (σε εκατ. €)		
Ανεξάρτητες μεταβλητές	συντελεστές	p-values
1 Σταθερός όρος	2.904,0	0
2 Τάση	4,7	0
3 Τάση στο τετράγωνο (αποτυπώνει τον ρυθμό μεταβολής)	-0,2	0
Μεσοχρόνιες περιόδους		
4 Ιούνιος 2010 – Ιανουάριος 2011	-215,6	0
5 Ιούλιος 2011 – Δεκέμβριος 2011	155,5	0
6 Απρίλιος – Μάιος 2012	-176,1	0
7 Ιούνιος – Αύγουστος 2012	47,9	0
8 Ιανουάριος – Σεπτέμβριος 2013	197,9	0
9 Οκτώβριος – Νοέμβριος 2013	-213,3	0
10 Αύγουστος – Νοέμβριος 2014	61,8	0
11 Δεκέμβριος 2014 – Σεπτέμβριος 2015	-103,9	0
12 Ιούνιος 2016 – Ιανουάριος 2017	-84,2	0
13 Μάιος – Αύγουστος 2017	-52,6	0
14 Ιούλιος – Δεκέμβριος 2018	86,2	0

$R^2 = 98,94\%$, $adj. R^2 = 98,79\%$. $N = 103$.
Τα κατάλοιπα είναι στάσιμα σε επίπεδο σημαντικότητας 1‰.

Πίνακας 11

Οικονομετρική ανάλυση με ανθεκτικούς εκτιμητές της πορείας της τραπεζικής χρηματοδότησης στην διαχείριση της ακινήτου περιουσίας, Ιουν. 2010 – Δεκ. 2018 (σε εκατ. €)

Ανεξάρτητες μεταβλητές	συντελεστές	p-values
1 Σταθερός όρος	4.313,0	0
2 Τάση	12,5	0
3 Τάση στο τετράγωνο (αποτυπώνει τον ρυθμό μεταβολής)	-0,1	0
Μεσοχρόνιες περίοδοι		
4 Ιούνιος – Αύγουστος 2010	-778,4	0
5 Σεπτέμβριος 2010 – Αύγουστος 2011	-935,9	0
6 Σεπτέμβριος 2011 – Σεπτέμβριος 2012	119,6	0,001
7 Ιανουάριος – Φεβρουάριος 2013	-512,4	0
8 Μάιος – Νοέμβριος 2013	-141,5	0,001
9 Δεκέμβριος 2013 – Νοέμβριος 2014	315,9	0
10 Ιανουάριος – Ιούνιος 2016	330,7	0
11 Ιανουάριος – Νοέμβριος 2017	207,0	0
12 Νοέμβριος – Δεκέμβριος 2018	-389,6	0

$R^2 = 97,30\%$, $adj. R^2 = 96,97\%$. $N = 103$.

Τα κατάλοιπα είναι στάσιμα σε επίπεδο σημαντικότητας 1‰.

Πίνακας 12

Οικονομετρική ανάλυση με ανθεκτικούς εκτιμητές της πορείας της τραπεζικής χρηματοδότησης στις επαγγελματικές-επιστημονικές-τεχνικές-διοικητικές-υποστηρικτικές δραστηριότητες, Ιουν. 2010 – Δεκ. 2018 (σε εκατ. €)

Ανεξάρτητες μεταβλητές	συντελεστές	p-values
1 Σταθερός όρος	3.378,4	0
2 Τάση	-5,2	0
Μεσοχρόνιες περίοδοι		
3 Ιούνιος – Σεπτέμβριος 2010	325,7	0
4 Μάρτιος – Αύγουστος 2011	341,2	0
5 Σεπτέμβριος – Νοέμβριος 2011	113,4	0,002
6 Ιανουάριος – Σεπτέμβριος 2012	-226,1	0
7 Οκτώβριος 2012 – Νοέμβριος 2013	-568,3	0
8 Δεκέμβριος 2013 – Απρίλιος 2014	-174,9	0
9 Μάιος – Σεπτέμβριος 2014	109,8	0
10 Οκτώβριος 2014 – Σεπτέμβριος 2016	353,9	0
11 Ιούνιος – Ιούλιος 2017	95,3	0
12 Σεπτέμβριος – Νοέμβριος 2017	95,5	0
13 Δεκέμβριος 2017 – Ιούνιος 2018	-97,8	0
14 Ιούλιος – Δεκέμβριος 2018	-297,3	0

$R^2 = 95,45\%$, $adj. R^2 = 94,78\%$. $N = 103$.

Τα κατάλοιπα είναι στάσιμα σε επίπεδο σημαντικότητας 1‰.

Πίνακας 13

Οικονομετρική ανάλυση με ανθεκτικούς εκτιμητές της πορείας της τραπεζικής χρηματοδότησης στους λοιπούς κλάδους, Ιουν. 2010 – Δεκ. 2018 (σε εκατ. €)

Ανεξάρτητες μεταβλητές	συντελεστές	p-values
1 Σταθερός όρος	11.580,2	0
2 Τάση	-169,7	0
3 Τάση στο τετράγωνο (αποτυπώνει τον ρυθμό μεταβολής)	1,0	0
Μεσοχρόνιες περιόδους		
4 Ιούνιος 2010 – Φεβρουάριος 2011	1.375,4	0
5 Ιούνιος – Δεκέμβριος 2012	-317,9	0
6 Μάρτιος – Νοέμβριος 2013	337,6	0
7 Δεκέμβριος 2013 – Απρίλιος 2015	-514,3	0
8 Ιούλιος 2015 – Δεκέμβριος 2017	254,8	0,006
9 Φεβρουάριος – Δεκέμβριος 2018	315,9	0,007

$R^2 = 99,28\%$, $adj. R^2 = 99,22\%$. $N = 103$.

Τα κατάλοιπα είναι στάσιμα σε επίπεδο σημαντικότητας 6‰.

ωμένο δανεισμό και έξι με πιο αυξημένο και ακόμα πιο αυξημένο δανεισμό. (Βλ. Πίνακα 12.)

- Στους λοιπούς κλάδους: Αρνητική τάση με φθίνοντα ρυθμό που έχει ως αποτέλεσμα την εμφάνιση ελαχίστου και, εν συνεχεία, θετικής τάσης, απουσία εποχικότητας, καθώς και έξι μεσοχρόνιες περιόδους: δύο με πιο μειωμένο δανεισμό και τέσσερις με πιο αυξημένο δανεισμό. (Βλ. Πίνακα 13.)

4. ΕΞΕΤΑΣΗ ΤΗΣ ΜΑΚΡΟΧΡΟΝΙΑΣ ΤΑΣΗΣ ΚΑΙ ΤΩΝ ΜΕΣΟΧΡΟΝΙΩΝ ΑΥΞΟΜΕΙΩΣΕΩΝ

Διαπιστώνεται ότι η τραπεζική χρηματοδότηση μακροχρονίως κινήθηκε: (Α) Πτωτικώς σε πέντε κλάδους: γεωργία-δασοκομία-αλιεία, ορυχεία-λατομεία (με την μείωση της χρηματοδότησης μεταξύ αρχής και τέλους του υπό εξέταση χρονικού διαστήματος να είναι της τάξεως του 0,5 - 1,0 δισ. ευρώ), εμπόριο, μεταποίηση, επαγγελματικές κ.ά. δραστηριότητες (με την μείωση να υπερβαίνει το 1,0 δισ.). (Β) Ανοδικώς και εν συνεχεία πτωτικώς σε πέντε κλάδους: διαχείριση ακίνητης περιουσίας, κατάλυμα-εστίαση (με την άνοδο να υπερβαίνει την πτώση κατά 0,1-0,2 δισ.), κατασκευές, επικοινωνία-ενημέρωση, ηλεκτρισμό κ.ά. (με την πτώση να υπερβαίνει την άνοδο κατά 2,5, 1,3 και 0,3 δισ., αντιστοίχως). Σημειώνεται ότι όλες οι αναστροφές της ανοδικής τάσης έλαβαν χώραν κατά την πρώτη πενταετία: στις κατασκευές και την επικοινωνία-ενημέρωση μεταξύ υπογραφής πρώτου μνημονίου και σχηματισμού της κυβερνήσεως Σαμαρά, σε ηλεκτρισμό κ.ά. επί κυβερνήσεως Σαμαρά, στην διαχείριση ακινήτου περιουσίας και σε καταλύματα-εστίαση κατά το πρώτο τρίμηνο του 2015 (επί κυβερνήσεως Τσίπρα). (Γ) Πτωτικώς και εν συνεχεία ανοδικώς σε πλωτές μεταφορές, αποθήκευση-λοιπές (μη-πλωτές) μεταφορές, λοιπούς κλάδους (με την πτώση να υπερβαίνει την άνοδο κατά 10,4, 0,7 και 6,2 δισ., αντιστοίχως). Σημειώνε-

ται ότι όλες οι αναστροφές της πτωτικής τάσης έλαβαν χώραν κατά το δεύτερο μέρος του υπό εξέταση χρονικού διαστήματος (3^ο τρίμηνο 2015 ως τέλος του 2018, επί κυβερνήσεως Τσίπρα): σε αποθήκευση-μη-πλωτές μεταφορές αμέσως μετά τους περιορισμούς στις κινήσεις κεφαλαίων (περί τον Αύγουστο-Σεπτέμβριο του 2015), στις πλωτές μεταφορές και στους λοιπούς κλάδους με την σταδιακή ανάκαμψη της οικονομίας κατά τα έτη 2017-18).

Εποχικότητα διαπιστώθηκε μόνο στην περίπτωση του κλάδου των καταλυμάτων και της εστίασης (αύξηση κατά τον Μάρτιο και ακόμα περισσότερο τον Ιούνιο) σε σχέση με τους άλλους μήνες του χρόνου.

Οι μεσοχρόνιες διακυμάνσεις (δηλ. οι αυξομειώσεις επί της καθαρής τάσεως) απαιτούν την δική τους ανάγνωση. Κάποιες ήσαν της τάξεως των 600 εκατ. ευρώ και άνω (φέρουσες είτε αρνητικό είτε θετικό πρόσημο) και κάποιες (κυρίως εκ των υπολοίπων) ήσαν διετούς ή μεγαλύτερας διάρκειας (επίσης φέρουσες είτε αρνητικό είτε θετικό πρόσημο). Εν τούτοις, η πλειονότης των αυξομειώσεων ήταν μικρότερη τόσο σε μέγεθος όσο και διάρκεια. Η στατιστική επεξεργασία των εν λόγω διακυμάνσεων στην χρηματοδότηση δείχνει μέτρια θετική συσχέτιση στις περιπτώσεις: (α) της μεταποίησης και των επαγγελματικών κ.ά. δραστηριοτήτων ($\rho=75,5\%$), (β) της γεωργίας κ.ά. και των επαγγελματικών κ.ά. δραστηριοτήτων ($\rho=71,1\%$), (γ) της μεταποίησης και των κατασκευών ($\rho=69,6\%$), (δ) των κατασκευών και των επαγγελματικών κ.ά. δραστηριοτήτων ($\rho=63,5\%$), (ε) των ορυχείων-λατομείων και του εμπορίου ($\rho=62,0\%$), καθώς και μέτρια αρνητική συσχέτιση στις περιπτώσεις της: (i) παροχής ηλεκτρισμού κ.ά. και των επαγγελματικών κ.ά. δραστηριοτήτων ($\rho=-62,7\%$), (ii) διαχείρισης της ακίνητης περιουσίας και των λοιπών κλάδων ($\rho=-64,3\%$). Συνάγεται ότι ίσως: (1) Τα πέντε πρώτα ζεύγη κλάδων αντιμετώπιζαν παρεμφερείς κυκλικές καταστάσεις με αποτέλεσμα να έχουν παρεμφερείς περιστασιακές ανάγκες υψηλότερης ή χαμηλότερης τραπεζικής χρηματοδότησης. (2) Η μεσοχρόνια προσέγγιση των τραπεζών προς τα πέντε πρώτα ζεύγη κλάδων ήταν παρεμφερής. (3) Τα δύο τελευταία ζεύγη κλάδων αντιμετώπιζαν αντίθετες κυκλικές καταστάσεις με αποτέλεσμα να έχουν αντιδιαμετρικές ανάγκες τραπεζικής χρηματοδότησης. (4) Η μεσοχρόνια προσέγγιση των τραπεζών προς τα δύο τελευταία ζεύγη κλάδων αντικατοπτρίζει ανάγκες εξισορροπήσεως/διαχειρίσεως χαρτοφυλακίου.

Μια δοκιμαστική (πιλοτική) διερεύνηση των καταγεγραμμένων συμβάντων που έλαβαν χώραν μεταξύ Δεκεμβρίου 2017 και Δεκεμβρίου 2018 και, ιδίως, κατά την εποχή της κάθε μεσοχρονίου μεταβολής, αναδεικνύει πρόσθετα στοιχεία που (είτε μόνα τους είτε από κοινού με άλλα καταγεγραμμένα ή μη καταγεγραμμένα συμβάντα) ενδέχεται να επηρέασαν την ροή της τραπεζικής χρηματοδότησης προς τους κλάδους παραγωγής. (Συγχρόνως ξεπροβάλλει μια κάπως ασυνήθης χρησιμότητα της στατιστικής στον εντοπισμό κομβικών γεγονότων εκ της πλειάδος ή απειρίας των συμβάντων που λαμβάνουν χώραν.) Στο πλαίσιο αυτό σημειώνεται ότι:

ι. Η εκτός τάσεως μείωση στην χρηματοδότηση της γεωργίας κ.ά. τον Απρίλιο του 2018, φαίνεται να συμπίπτει με κυβερνητικές υποσχέσεις βελτιώσεως των όρων πληρωμής των αγροτικών χρεών. Μάλλον αναμενόμενα, η είδηση επέφερε διάθεση αναμονής/αναβολής (α) ρυθμίσεως ή αποπληρωμής χρεών σε αγρότες ή/και (β) παροχής δανείων από τράπεζες προς αγρότες καθότι προφανώς ο ηθικός κίνδυνος

- ήταν συγκυριακά αυξημένος. Ο συγκεκριμένος κύκλος μειωμένης τραπεζικής χρηματοδότησης ολοκληρώθηκε τον Σεπτέμβριο του 2018, οπότε η κυβέρνηση και ο τραπεζικός κλάδος ήλθαν σε συνεννόηση και συμφώνησαν να αυξηθεί η χρηματοδότηση στον αγροδιατροφικό τομέα.
- ii. Η εκτός τάσεως μείωση στην χρηματοδότηση των ορυχείων-λατομείων τον Ιανουάριο του 2018, φαίνεται να συμπίπτει με την ψήφιση νέου λατομικού νόμου. Ο κύκλος της μειωμένης τραπεζικής χρηματοδότησης ολοκληρώθηκε τον Οκτώβριο του 2018, οπότε η κυβέρνηση, η Ευρωπαϊκή Τράπεζα Επενδύσεων και οι διοικήσεις των εγχωρίων τραπεζών συμφώνησαν στην διευκόλυνση της τραπεζικής χρηματοδότησης για δράσεις που αφορούν στο περιβάλλον, σε υλικά, στην Δυτική Μακεδονία κ.ά.
 - iii. Η εκτός τάσεως μείωση στην χρηματοδότηση του κλάδου του ηλεκτρισμού κ.ά. τον Ιούλιο-Αύγουστο του 2018, φαίνεται να εκδηλώνεται αμέσως μετά την (α) ένεση ρευστότητας από το Δημόσιο προς την ΔΕΗ, εν είδει προπληρωμής των λογαριασμών του Δημοσίου για το υπόλοιπο του έτους, και (β) συμφωνία αναχρηματοδότησης της ΔΕΗ από τις εγχώριες τράπεζες με χαμηλό επιτόκιο.
 - iv. Η εκτός τάσεως αύξηση στην χρηματοδότηση του κλάδου της επικοινωνίας-ενημέρωσης κατά το δεύτερο ήμισυ του 2018, φαίνεται να εκδηλώνεται αμέσως μετά την επιλογή από το Εθνικό Συμβούλιο Ραδιοτηλεόρασης των ληπτών πανελλαδικών τηλεοπτικών αδειών γενικού περιεχομένου. (Ακολούθησε η διάθεση και άλλων αδειών.) Η εκτίμηση είναι ότι οι *προσωρινοί* αδειολήπτες προέβησαν σε άμεσο δανεισμό προκειμένου να πληρώσουν για τις άδειες και να καταστούν *οριστικοί* δικαιούχοι.
 - v. Η εκτός τάσεως μείωση στην χρηματοδότηση των κατασκευών τον Μάρτιο-Απρίλιο και Ιούνιο-Δεκέμβριο 2018, φαίνεται να συμπίπτει με μειωμένη ζήτηση και, κατ' επέκτασιν, μειωμένες ανάγκες των δραστηριοποιούμενων επιχειρήσεων σε ρευστά. Πρόκειται για προϊόντα, όπως η κατασκευή οδικών αξόνων, που κατά κανόνα χρειάζονται αρκετό χρόνο για να υλοποιηθούν και συνήθως εξοφλούνται με την ολοκλήρωση παράδοση, όμως απαιτούν ρευστά μέχρι να εξοφληθούν. Σε αντιδιαστολή, η αύξηση στην χρηματοδότηση του κλάδου των καταλυμάτων και της εστίασης τον Απρίλιο-Σεπτέμβριο 2018, φαίνεται να συμπίπτει με αυξημένη ζήτηση.
 - vi. Η εκτός τάσεως μείωση στην χρηματοδότηση του εμπορίου τον Απρίλιο και, εν συνεχεία, τον Μάιο-Δεκέμβριο του 2018, φαίνεται να συμπίπτει με αυξημένη ζήτηση για προϊόντα που κατά κανόνα εξοφλούνται άμεσα, οπότε εύλογα η ρευστοτητα των πωλητών βελτιώθηκε και η ανάγκη προσφυγής στο τραπεζικό δανεισμό μειώθηκε. Παρομοίως, η μείωση στην χρηματοδότηση των παρόχων υπηρεσιών: (α) Αποθήκευσης και μη-πλωτών μεταφορών τον Φεβρουάριο-Μάιο και, πάλι, τον Ιούνιο-Δεκέμβριο του 2018, φαίνεται να συμπίπτει με αυξημένη ζήτηση. (β) Πλωτών μεταφορών τον Δεκέμβριο 2017-Απρίλιο 2018 και, πάλι (σε μικρότερο βαθμό) τον Μάιο-Δεκέμβριο του 2018, φαίνεται να συμπίπτει με αυξημένη ζήτηση.
 - vii. Η εκτός τάσεως μείωση στην χρηματοδότηση του κλάδου διαχείρισεως της ακίνητης περιουσίας τον Νοέμβριο-Δεκέμβριο του 2018, φαίνεται να συμπίπτει με αύξηση της ζήτησης/αγοράς ακινήτων: (α) Από αλλοδαπούς (άρα, με εισροή κε-

φαλαίων από την αλλοδαπή) που παραλλήλως αποκτούσαν χρυσή βίτσα). (β) Λόγω νέων προοπτικών στην εκμετάλλευση ακινήτων μέσω ψηφιακών εφαρμογών βραχυχρονίου μισθώσεως. (γ) Εξ αιτίας της ανόδου του τουρισμού και, κατ' επέκτασιν, της ζήτησεως εξοχικών κατοικιών.

Οι περιπτώσεις υπό τα στοιχεία (i)-(iv) ανάγονται σε κυβερνητικές παρεμβάσεις και οι υπόλοιπες στις συνθήκες της αγοράς.

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η εκτίμηση των παραμέτρων των μακροχρονίων, των εποχικών και των ενδιάμεσων φάσεων της της τραπεζικής χρηματοδότησης προς τους κλάδους παραγωγής της οικονομίας που επιχειρήθηκε στις προηγούμενες σελίδες: (α) Παρέχει μια τεχνική ακτινογράφιση της καταστάσεως (με υψηλά R^2 και στάσιμες σειρές καταλοίπων). (β) Επιτρέπει τον εντοπισμό των κομβικών στιγμών ενάρξεως και λήξεως περιόδων αποκλίσεως από την τάση και, κατ' επέκταση, τον προσδιορισμό των συμβάντων που ενδεχομένως προξένησαν ή συνέτειναν σε αυτές τις αποκλίσεις. Εν προκειμένω, τα συμβάντα ανάγονται σε κυβερνητικές παρεμβάσεις ή σε εξελίξεις στην αγορά. Με τον τρόπο αυτό, κατά την γνώμη του γράφοντος, αντλούνται ενδιαφέροντα μαθήματα για την λειτουργία της αγοράς και την άσκηση αναπτυξιακής/κλαδικής πολιτικής, ιδίως όσον αφορά στην ανάγκη για την επανάληψη ή αποφυγή ενεργειών στο μέλλον: Αποφυγή ενεργειών που αποθαρρύνουν την αποπληρωμή δανείων διότι δυσχεραίνουν την χρηματοδότηση (ήτοι, την αιμοδότηση) των κλάδων από το τραπεζικό σύστημα και επανάληψη ενεργειών που διευκολύνουν την τραπεζική χρηματοδότηση και απρόσκοπτη λειτουργία των κλάδων παραγωγής της χώρας.

Επίσης γίνεται αντιληπτό ότι υπήρχε αρκετή ετερογένεια όσον αφορά στις ροές τραπεζικών κεφαλαίων προς τους επιμέρους κλάδους της οικονομίας. Πιο συγκεκριμένα, διαπιστώνεται ότι η μακροχρόνια τάση ήταν:

- Πτωτική σε πέντε κλάδους (γεωργία κ.ά., ορυχεία-λατομεία, εμπόριο, μεταποίηση, επαγγελματικές κ.ά. δραστηριότητες).
- Αρχικώς ανοδική και εν συνεχεία πτωτική σε πέντε κλάδους (διαχείριση ακίνητης περιουσίας, καταλύματα-εστίαση, κατασκευές, επικοινωνία-ενημέρωση, ηλεκτρισμό κ.ά.).
- Αρχικώς πτωτική και εν συνεχεία ανοδική σε τρεις κλάδους (πλωτές μεταφορές, αποθήκευση-λοιπές μεταφορές, λοιπούς κλάδους).
- Διαφοροποιείτο περαιτέρω από κλάδο σε κλάδο ως προς την κλίση, καθώς και τον χρόνο μεταβολής της κλίσης (στην περίπτωση που υφίστατο μεταβολή).
- Συνδυάζετο με εποχικότητα σε έναν μόνον κλάδο (τον κλάδο παροχής καταλυμάτων-εστίασης).

Είναι σκόπιμο να έχουμε κατά νου ότι οι επιμέρους κλάδοι της οικονομίας δεν λειτουργεί σε όλα ομοίμορφα και συγχρονισμένα.

Οι επιδράσεις των ενδιάμεσων ή μεσοχρονίων αυξομειώσεων από την τάση γενικώς ποίκιλλαν ως προς το μέγεθος και πρόσημο και ήσαν ασύγχρονες. Κάποια ομοιότητα ενδέχεται να υπήρχε στις χρηματοδοτήσεις προς την μεταποίηση και προς τις επαγγελματικές κ.ά. δραστηριότητες, ωστόσο δεν ήταν τέτοια που να επιτρέπει

μεγαλύτερη μνεία.

ABSTRACT

The paper looks into the evolution of domestic bank financing across Greece's the 13 sectors of economic activity using monthly data provided by the Bank of Greece, and simple econometric tools for estimating the impact of the long-term trend, seasonality, and intermediate fluctuations. Although the overall lending pattern generally resembles the pattern of deposits, by and large the trend parameter varies from one sector to another, the intermediate fluctuations are asynchronous, and only tourism-related funding exhibits seasonality. The timing of the intermediate deviations from the trend help identify the events that may have triggered the deviation; thus, advancing one's understanding about the operation of the economy, sectoral responses to similar events, and providing useful lessons on carrying out economic development policy

ΑΝΑΦΟΡΕΣ

- Ahmed J.U., Rashid A. (2012). "Role of Financial Institutions in Financing Small and Medium Enterprises: A Study in Garo Hills of Meghalaya." *Journal of Commerce and Accounting Research*, 1.4: 7-23.
- Black K. (1992). *Business Statistics*. St. Paul MI: West Publishing.
- Cameron S. (2005). *Econometrics*. Maidenhead UK: McGraw Hill.
- Chattoraj A.K., Mohan A., Ranjan V. (2013). "Role of Scheduled Commercial Banks in the SocioEconomic Development of India." *Anusandhanika*, 5.1-2: 199-204.
- Chisasa J., Makina D. (2012). "Trends In Credit To Smallholder Farmers In South Africa." *International Business & Economics Research Journal*, 11.7: 771-784.
- Harris R.J. (1985). *A primer of multivariate statistics*. 2nd ed. New York: Academic Press.
- Fox K.A. (1968). *Intermediate Economic Statistics*. New York: Wiley
- Howell D.C. (1987). *Statistics for psychology*. 4th ed.. Belmont CA: Duxbury.
- Katan L., Recio Espejo J.M. (2018). "Structural modeling of the financial support for the Ukrainian agrarian sector." *Investment Management & Financial Innovations*, 15.3: 199-211.
- Mankiw N.G. (2002). *Αρχές της Οικονομικής*. Μετάφρ.: Ι. Μπιλάλη, Ν. Σταματάκης. Επιστ. Επιμ.: Σ. Ζήκος. Αθήνα: Τυπωθήτω – Γιώργος Δαρδανός.
- Milić D., Soleša D. (2017). "The analysis of macroeconomic determinants of the banking sector liquidity with role in financing agricultural sector." *Ekonomika Poljoprivrede*, 64.2: 533-55.
- Rodrigues J., Santos A.C., Teles N. (2016). "Semi-peripheral financialisation: the case of Portugal." *Review of International Political Economy*, 23.3: 480-510.
- Smith J.D., Duncan A J. (1944). *Fundamentals of the Theory of Statistics*. Vol. 1. New York: McGraw Hill.
- Τζωρτζόπουλος, Π.Θ. (1985). *Ανάλυση χρονολογικών σειρών. Τάση-εποχικότητα*. 3η έκδοση. Αθήνα: Οικονομικό.



ΑΝΑΠΤΥΞΗ ΜΟΝΤΕΛΟΥ ΟΛΙΣΤΙΚΗΣ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ ΠΑΡΑΓΩΓΙΚΩΝ ΔΙΑΔΙΚΑΣΙΩΝ ΣΥΝΕΧΟΥΣ ΠΑΡΑΚΟΛΟΥΘΗΣΗΣ

Κωνσταντίνος Α. Τασιάς, Γεώργιος Νενές, Σοφία Παναγιωτίδου
Πανεπιστήμιο Δυτικής Μακεδονίας, Τμήμα Μηχανολόγων Μηχανικών
{ktasias,gnenes,span}@uowm.gr

ΠΕΡΙΛΗΨΗ

Στις σύγχρονες παραγωγικές διαδικασίες είναι πλέον εφικτή η συνεχής παρακολούθηση της παραγωγής μέσω της χρήσης πολλαπλών αισθητήρων. Ως εκ τούτου, κρίνεται ως αναγκαία η ανάπτυξη κατάλληλων στατιστικών μοντέλων, τα οποία, μέσω της επεξεργασίας της πληθώρας των διαθέσιμων δεδομένων, θα λειτουργούν ως εργαλεία λήψης αποφάσεων σε πραγματικό χρόνο. Στην παρούσα εργασία παρουσιάζεται η ανάπτυξη ενός μοντέλου συνδυαζόμενης βελτιστοποίησης τριών βασικών πτυχών κάθε επιχειρησιακής λειτουργίας: διαχείριση αποθεμάτων, συντήρηση εξοπλισμού και ποιότητα προϊόντος. Η παρακολούθηση της διαδικασίας βασίζεται στη Μπεϋζιανή ανάλυση και οι βέλτιστες παράμετροι σχεδίασης καθορίζονται με οικονομικά και στατιστικά κριτήρια. Οι ρεαλιστικές υποθέσεις κατά τη διαμόρφωση του προβλήματος οδηγούν στην ανάπτυξη ενός ευρέως εφαρμόσιμου εργαλείου στο πεδίο των σύγχρονων παραγωγικών συστημάτων.

Λέξεις Κλειδιά: Στατιστικός Έλεγχος Ποιότητας, Συντήρηση, Διαχείριση Αποθεμάτων, Industry 4.0; Θεώρημα Μπέυζ.

1. ΕΙΣΑΓΩΓΗ

Η αλματώδης τεχνολογική εξέλιξη και η ολοένα αυξανόμενη χρήση αυτής στη βιομηχανία έχει αλλάξει ριζικά το τοπίο αναφορικά με τη διοίκηση των παραγωγικών διαδικασιών. Σήμερα, είναι πλέον εφικτή η χρήση μιας πληθώρας αισθητήρων, οι οποίοι δύνανται να παρακολουθούν το σύνολο της παραγωγικής διαδικασίας, συλλέγοντας δεδομένα από όλα τα επίπεδα αυτής. Το Δίκτυο των Πραγμάτων (Internet of Things-IoT) επιτρέπει την αποτελεσματική ενσωμάτωση πραγματικών δεδομένων στο δίκτυο πληροφορίας θεμελιώνοντας τη διασύνδεση μεταξύ πραγμάτων, μηχανών και εργαζομένων.

Στο σύγχρονο υψηλά αυτοματοποιημένο παραγωγικό περιβάλλον, η συνεχής παρακολούθηση σύνθετων διαδικασιών είναι εφικτή και κατάλληλες πληροφορίες για κάθε ένα προϊόν της διαδικασίας είναι δυνατόν να συλλέγονται και να αναλύονται

περαιτέρω (Montgomery (2012)) οδηγώντας στη λήψη βέλτιστων αποφάσεων για πλήθος επιμέρους κρίσιμων πτυχών της παραγωγής.

Επιπρόσθετα, η εφαρμογή της τεχνολογίας των αισθητήρων επιτρέπει την ταυτόχρονη παρακολούθηση πολλαπλών χαρακτηριστικών ποιότητας αναφορικά με τα παραγόμενα προϊόντα (για παράδειγμα διαστάσεις, βάρος, θερμοκρασία, υγρασία κλπ), τα οποία στην πιο σύνθετη, αλλά και πιο ρεαλιστική, περίπτωση χαρακτηρίζονται από αλληλοσυσχέτιση.

Γίνεται εύκολα αντιληπτό, ότι η ταυτόχρονη παρακολούθηση πολλαπλών αλληλοσχετιζόμενων μεταβλητών για το σύνολο των προϊόντων της παραγωγικής διαδικασίας αποτελεί κοινό τόπο για τη σύγχρονη παραγωγή. Ορισμένες επιστημονικές δημοσιεύσεις έχουν ασχοληθεί με αυτό το πολύπλοκο πρόβλημα. Στις εργασίες τους οι Vargas (2003) και Variyath και Vattathoor (2014) προτείνουν διαγράμματα ελέγχου τύπου Hotelling για περιπτώσεις όπου το μέγεθος του δείγματος ισούται με μονάδα. Επιπρόσθετα, οι Yeh *et al.* (2005) μελέτησαν ένα πολυμεταβλητό διάγραμμα ελέγχου τύπου EWMA για την παρακολούθηση μεταβολών αναφορικά με τη διασπορά της διαδικασίας, μέσω μεμονωμένων μετρήσεων, ενώ, οι Zhang και Chang (2008) για την ταυτόχρονη παρακολούθηση της θέσης και της διασποράς.

Σε αυτό το πλούσιο σε διαθέσιμα δεδομένα σύγχρονο παραγωγικό περιβάλλον, οι ληφθείσες πληροφορίες μπορούν να αναβαθμίσουν σημαντικά τη διαδικασία λήψης αποφάσεων, όχι μόνο αναφορικά με την ποιότητα αυτής, όπως στις ανωτέρω εργασίες αλλά και για πρόσθετες πτυχές της παραγωγής. Δύο κρίσιμες πτυχές κάθε παραγωγικής διαδικασίας αποτελούν η συντήρηση του εξοπλισμού και η διαχείριση αποθεμάτων. Η ολιστική αντιμετώπιση ποιότητας, συντήρησης και αποθεμάτων σε βιομηχανικές εφαρμογές είναι ένα ζήτημα το οποίο δεν τυγχάνει πλούσιας έρευνας στην υπάρχουσα επιστημονική βιβλιογραφία. Ο Ben-Daya (1999) καθώς και οι Rahim και Ben-Daya (1998) μελέτησαν τη διασύνδεση μεταξύ συντήρησης, αποθεμάτων και ποιότητας σε διαδικασίες που υπόκεινται σε υποβάθμιση. Επιπρόσθετα, οι Pan *et al.* (2012) στην εργασία τους ανέπτυξαν ένα μοντέλο ταυτόχρονης βελτιστοποίησης της οικονομικής ποσότητας παραγωγής και της πολιτικής συντήρησης το οποίο αξιοποιεί τις συλλεγόμενες πληροφορίες μέσω ενός διαγράμματος ελέγχου τύπου Shewhart για ατελείς παραγωγικές διαδικασίες, πλήρως χαρακτηριζόμενες από ένα χαρακτηριστικό ποιότητας η μέση τιμή του οποίου επηρεάζεται από μια πιθανή συστηματική αιτία, είτε θετικά είτε αρνητικά.

Πιο πρόσφατες εργασίες επί του συγκεκριμένου θέματος, είναι αυτές των Nourelfath *et al.* (2016) και Fakher *et al.* (2018) οι οποίοι προσδιόρισαν κάτω από κοινό πρίσμα τις τρεις προαναφερθείσες πτυχές των παραγωγικών διαδικασιών μέσα από το κριτήριο της ελαχιστοποίησης του κόστους και της μεγιστοποίησης του κέρδους, αντίστοιχα. Οι Salmansia *et al.* (2017) ανέπτυξαν ένα ολοκληρωμένο μοντέλο για διαδικασίες που υπόκεινται σε πολλαπλές συστηματικές αιτίες, ενώ οι Salmansia *et al.* (2018) επέκτειναν το μοντέλο αυτό σε πολυμεταβλητό το οποίο,

όμως, εφαρμόζεται σε διαδικασίες που υπόκεινται σε μία συστηματική αιτία που επηρεάζει τη μέση τιμή.

Η συνεισφορά της παρούσας εργασίας έγκειται στην ανάπτυξη ενός μοντέλου ταυτόχρονης βελτιστοποίησης της ποιότητας, της συντήρησης και των αποθεμάτων το οποίο για πρώτη φορά στη βιβλιογραφία αξιοποιεί το Μπεϋζιανό θεώρημα. Το προτεινόμενο μοντέλο εφαρμόζεται για τη συνεχή παρακολούθηση πολυμεταβλητών διαδικασιών όπου πολλαπλές συστηματικές αιτίες είναι πιθανό να επιδράσουν.

2. ΜΑΘΗΜΑΤΙΚΟ ΜΟΝΤΕΛΟ

2.1 Υποθέσεις

Το προτεινόμενο μοντέλο βασίζεται στις ακόλουθες υποθέσεις:

i) Η παραγωγική διαδικασία έχει άπειρο χρονικό ορίζοντα και χαρακτηρίζεται από ρ αλληλοσχετιζόμενα ποιοτικά χαρακτηριστικά τα οποία ακολουθούν μια ρ -διάστατη κανονική κατανομή με επιθυμητό διάνυσμα μέσων μ_0 και επιθυμητό πίνακα (συν)διακυμάνσεων Σ_0 .

ii) Η διαδικασία υπόκειται σε πλήθος συστηματικών αιτιών, έστω m , οι οποίες επηρεάζουν μόνο τη θέση της κατανομής. Συγκεκριμένα, μία τυχαία συστηματική αιτία i μεταβάλλει το διάνυσμα μέσων από $\mu_0 = (\mu_{1,0}, \mu_{2,0}, \dots, \mu_{\rho,0})$ σε $\mu_i = (\mu_{1,i}, \mu_{2,i}, \dots, \mu_{\rho,i})$ ή πιο αναλυτικά σε $\mu_i = (\mu_{1,0} + \delta_{1,i} \cdot \sigma_{1,0}, \mu_{2,0} + \delta_{2,i} \cdot \sigma_{2,0}, \dots, \mu_{\rho,0} + \delta_{\rho,i} \cdot \sigma_{\rho,0})$, όπου $i = 1, 2, \dots, m$ και $\sigma_{r,0}$ η επιθυμητή τυπική απόκλιση του ποιοτικού χαρακτηριστικού r , όπου $r = 1, 2, \dots, \rho$. Αναφορικά με τα μεγέθη μεταβολής του διανύσματος μέσων από τις πιθανές συστηματικές αιτίες, υπάρχουν $\rho \times m$ πιθανές τιμές του δ , $\delta \in \{\delta_{1,1}, \dots, \delta_{1,\rho}; \dots; \delta_{\rho,1}, \dots, \delta_{\rho,m}\}$, όπου $\delta_{i,r}$ ο αριθμός τυπικών αποκλίσεων του r^{th} χαρακτηριστικού ποιότητας $(\sigma_{r,0})$ κατά τον οποίο μεταβάλλεται η μέση τιμή αυτού λόγω επίδρασης της συστηματικής αιτίας i .

iii) Το διάνυσμα μέσων παρακολουθείται μέσω του στατιστικού $T^2 = (x - \mu_0)^T \cdot \Sigma_0^{-1} \cdot (x - \mu_0)$, το οποίο είναι μια συνεχής, τυχαία μεταβλητή που ακολουθεί κατανομή χ^2 όταν η διαδικασία είναι υπό στατιστικό έλεγχο και μη κεντρική κατανομή χ^2 με ρ βαθμούς ελευθερίας και παράμετρο μη κεντρικότητας d_i^2 , όταν τελεί υπό την επίδραση της συστηματικής αιτίας i ($d_i = \sqrt{(\mu_i - \mu_0)^T \cdot \Sigma_0^{-1} \cdot (\mu_i - \mu_0)}$).

iv) Ο χρόνος μεταξύ διαδοχικών ελεύσεων κάθε συστηματικής αιτίας κατανέμεται εκθετικά. Ο ρυθμός έλευσης κάθε συστηματικής αιτίας εξαρτάται από την κατάσταση στην οποία βρίσκεται η διαδικασία, και, ως εκ τούτου, για τη μετάβαση από τη συστηματική αιτία i στην k , ο εν λόγω ρυθμός συμβολίζεται με

$\lambda_{(i \rightarrow k)}$. Γίνεται εύκολα αντιληπτό ότι ο ρυθμός με τον οποίο υποβιβάζεται η διαδικασία, όταν λειτουργεί υπό την επίδραση της συστηματικής αιτίας i , προκύπτει ως το άθροισμα όλων των ρυθμών των πιθανών μεταβάσεων από την κατάσταση i ως αρχική $\left(v_i = \sum_{k=1}^m \lambda_{(i \rightarrow k)} \right)$.

v) Πέραν των συστηματικών αιτιών, είναι πιθανό να επιδράσουν και βλάβες στη διαδικασία, η έλευση των οποίων προκαλεί διακοπή της παραγωγής. Αναφορικά με το ρυθμό έλευσης αυτών, γίνεται η ρεαλιστική υπόθεση ότι συνδέεται άμεσα με την πραγματική κατάσταση της διαδικασίας, υπό την έννοια ότι όσο περισσότερο απέχει η διαδικασία από το επιθυμητό επίπεδο λειτουργίας τόσο πιθανότερη καθίσταται η έλευση βλάβης. Για το σκοπό αυτό, ο ρυθμός έλευσης βλάβης συμβολίζεται με $\lambda_{F|i}$, όπου $i \in [0, m]$ η συστηματική αιτία υπό την επίδραση της οποίας λειτουργεί η διαδικασία.

vi) Η αποκατάσταση της διαδικασίας από μία συστηματική αιτία συνιστά ενέργεια προληπτικής συντήρησης (ΠΣ) ενώ η αποκατάσταση αυτής από μία βλάβη συνιστά ενέργεια διορθωτικής συντήρησης (ΔΣ).

vii) Μετά από K σημεία επιθεώρησης, ο κύκλος της παραγωγής περατώνεται, και στο $(K+1)$ διάστημα επιθεώρησης εκτελούνται εργασίες ΠΣ. Επισημαίνεται ότι η μεταβλητή K αποτελεί μεταβλητή απόφασης του διαμορφούμενου προβλήματος, όπως επεξηγείται και στη συνέχεια της εργασίας.

viii) Οι ρυθμοί παραγωγής (pr) και ζήτησης (dr) είναι σταθεροί και ισχύει ότι $pr \geq dr$. Λαμβάνοντας υπόψη ότι το προτεινόμενο μοντέλο εφαρμόζεται για τη συνεχή παρακολούθηση της διαδικασίας, ο χρόνος μεταξύ δύο διαδοχικών επιθεωρήσεων (h) υπολογίζεται ως ο αντίστροφος αριθμός του ρυθμού της παραγωγής ($h = 1 / pr$).

2.2. Περιγραφή Μαθηματικού Μοντέλου

Για την ολιστική βελτιστοποίηση της συντήρησης, της πολιτικής αποθεμάτων και της ποιότητας μιας διαδικασίας μέσω της συνεχούς παρακολούθησης αυτής, αναπτύσσεται ένα Μπεϋζιανό σχήμα ελέγχου. Σε κάθε σημείο επιθεώρησης, συλλέγεται ένα διάνυσμα παρατηρήσεων και η πιθανότητα να βρίσκεται η διαδικασία εκτός ελέγχου εκτιμάται βάσει του θεωρήματος Μπέυζ. Κατά συνέπεια, το Μπεϋζιανό θεώρημα λειτουργεί ως ένας μηχανισμός ανατροφοδότησης και συνεπαγόμενης μετατροπής της πιθανότητας εκτός ελέγχου στην έναρξη του διαστήματος επιθεώρησης σε πιθανότητα εκτός ελέγχου στο τέλος αυτής, δεδομένων των ληφθέντων παρατηρήσεων από την επιθεώρηση της παραγωγής.

Σε κάθε διάστημα επιθεώρησης, υπολογίζεται η εκ των υστέρων (posterior) πιθανότητα λειτουργίας εκτός ελέγχου, p' . Βάσει αυτής της τιμής και κατόπιν

σύγκρισης με το όριο ελέγχου (p_c), κατά την t -οστή επιθεώρηση, λαμβάνεται μία εκ των ακόλουθων αποφάσεων (a_t):

- Αν $p' < p_c$ τότε η διαδικασία συνεχίζει κανονικά τη λειτουργία της και η p' λειτουργεί ως εκ των προτέρων (prior) πιθανότητα λειτουργίας εκτός ελέγχου για το επόμενο διάστημα επιθεώρησης ($a_t = 0$).

- Αν $p' \geq p_c$, τότε διακόπτεται η παραγωγική διαδικασία και στην περίπτωση όπου διαπιστωθεί επίδραση συστηματικής αιτίας, εκτελούνται εργασίες ΠΣ για την αποκατάσταση της παραγωγής σε στατιστικό έλεγχο ($a_t = 1$).

Σε περίπτωση λειτουργίας της παραγωγικής διαδικασίας υπό την επίδραση της συστηματικής αιτίας i , υφίστανται τα ακόλουθα σενάρια με τις αντίστοιχες πιθανότητες:

- Παραμονή της διαδικασίας στην ίδια κατάσταση: $p_i(h) = \exp\left(-\left(v_i + \lambda_{F|i}\right) \cdot h\right)$.
- Μετάβαση της διαδικασίας στην κατάσταση k ($\forall i \neq k$):

$$p_k(h) = \int_0^h \sum_{y \in [1,m] \setminus i} \lambda_{(i \rightarrow y)} \cdot \exp\left(-\left(v_i + \lambda_{F|i}\right) \cdot t\right) \cdot p_k^y(h-t) dt$$

- Μετάβαση της διαδικασίας σε κατάσταση βλάβης F :

$$p_F(h) = \int_0^h \lambda_{F|i} \cdot \exp\left(-\left(v_i + \lambda_{F|i}\right) \cdot t\right) dt + \int_0^h \sum_{y \in [1,m] \setminus i} \lambda_{(i \rightarrow y)} \cdot \exp\left(-\left(v_i + \lambda_{F|i}\right) \cdot t\right) \cdot p_F^y(h-t) dt$$

Δεδομένης της υπόθεσης ότι το στατιστικό T^2 ακολουθεί κατανομή χ^2 όταν η διαδικασία είναι σε στατιστικό έλεγχο, η συνάρτηση πυκνότητας πιθανότητας αυτού δίδεται από την ακόλουθη σχέση:

$$f(T^2; \rho, d_0^2) = \frac{1}{\Gamma\left(\frac{\rho}{2}\right)} \cdot \frac{1}{2^{\frac{\rho}{2}}} \cdot \exp\left(-\frac{1}{2}(T^2)\right) \cdot (T^2)^{\frac{\rho}{2}-1} \quad (1)$$

Στην αντίθετη περίπτωση που η διαδικασία λειτουργεί υπό την επίδραση της συστηματικής αιτίας i , $i \in [1, m]$, τότε η συνάρτηση πυκνότητας πιθανότητας του T^2 γίνεται:

$$f(T^2; \rho, d_i^2) = \exp\left(-\left(\frac{d_i^2 + T^2}{2}\right)\right) \cdot \sum_{n=0}^{\infty} \frac{\left(\frac{d_i^2}{2}\right)^n}{n!} \cdot \frac{1}{\Gamma\left(\frac{\rho}{2} + n\right)} \cdot \frac{1}{2^{\frac{\rho}{2} + n}} \cdot (T^2)^{\frac{\rho}{2} + n - 1} \quad (2)$$

Για λόγους συντόμευσης, οι συναρτήσεις πυκνότητας πιθανότητας του T^2 (εξισώσεις (1) και (2)) θα συμβολίζονται στο υπόλοιπο της εργασίας με f_0 και f_i , αντίστοιχα.

Η αντίστοιχη συνάρτηση πυκνότητας πιθανότητας του T^2 για το επόμενο σημείο επιθεώρησης υπολογίζεται σταθμίζοντας τη συνάρτηση πυκνότητας πιθανότητας αυτού για πιθανή λειτουργία υπό κάθε πιθανή κατάσταση, εξαρτάται από την απόφαση που λήφθηκε κατά την τρέχουσα επιθεώρηση (a_{t-1}) και δίδεται από τις ακόλουθες σχέσεις:

$$f(T^2 | p_i \forall i \in [0, m], a_{t-1} = 0) = \sum_{k=0}^m \sum_{z \in [0, m] \setminus k} p_z \cdot f_k \cdot p_z(h) \quad (3)$$

$$f(T^2 | p_i \forall i \in [0, m], (a_{t-1} = 1) \cup F) = \sum_{k=0}^m f_k \cdot p_0(h) \quad (4)$$

Εφαρμόζοντας το Μπεϋζιανό θεώρημα, δοθέντων των εξισώσεων (3) και (4), η εκ των υστέρων πιθανότητα λειτουργίας της διαδικασίας υπό τη συστηματική αιτία k ($k = 0, 1, \dots, m$) ισούται με:

$$p'_k(a_{t-1} = 0) = \frac{\sum_{z=0}^m p_z \cdot f_k \cdot p_z(h)}{\sum_{v=0}^m \sum_{z=0}^m p_z \cdot f_v \cdot p_z(h)} \quad \text{και} \quad p'_k[(a_{t-1} = 1) \cup F] = \frac{f_k \cdot p_0(h)}{\sum_{v=0}^m f_v \cdot p_0(h)} \quad (5)$$

Η συνολική εκ των υστέρων πιθανότητα εκτός ελέγχου λειτουργίας προκύπτει από τις ακόλουθες εξισώσεις:

$$p'(a_{t-1} = 0) = \sum_{k=1}^m p'_k(a_{t-1} = 0) = \frac{\sum_{k=1}^m \sum_{z=0}^m p_z \cdot f_k \cdot p_z(h)}{\sum_{v=0}^m \sum_{z=0}^m p_z \cdot f_v \cdot p_z(h)}$$

$$p'[(a_{t-1} = 1) \cup F] = \sum_{k=1}^m p'_k[(a_{t-1} = 1) \cup F] = \frac{\sum_{k=1}^m f_k \cdot p_0(h)}{\sum_{v=0}^m f_v \cdot p_0(h)}$$

Η λειτουργία του προτεινόμενου σχήματος ελέγχου μοντελοποιείται μαθηματικά μέσω μιας Μαρκοβιανής αλυσίδας διακριτού χρόνου. Κάθε κατάσταση της Μαρκοβιανής αλυσίδας ορίζεται πλήρως από την πραγματική κατάσταση της διαδικασίας $Y_t = \{0, 1, \dots, m, F\}$, όπου F η κατάσταση κατά την οποία η παραγωγική διαδικασία τελεί υπό βλάβη, και την πιθανότητα λειτουργίας της παραγωγής υπό κάθε πιθανή συστηματική αιτία, η οποία συμβολίζεται ως $P = (p_1, p_2, \dots, p_m)$.

Η εκ των προτέρων πιθανότητα λειτουργίας εκτός ελέγχου μπορεί εύκολα να υπολογισθεί ως $p = \sum_{k=1}^m p_k$ και για λόγους διευκόλυνσης του αναγνώστη το διάνυσμα P διακρίνεται σε P_{cz} και P_{az} στην περίπτωση όπου $a_{t-1} = 0$ και $a_{t-1} = 1$, αντίστοιχα.

Βάσει της προαναφερθείσας μοντελοποίησης και για τη δημιουργία ενός πεπερασμένου αριθμού πιθανών καταστάσεων, οι πιθανότητες έχουν διακριτοποιηθεί στα ακόλουθα τέσσερα διαστήματα: $I_1 = [0, 0.02 ; 0.004)$, $I_2 = [0.02, 0.10 ; 0.008)$, $I_3 = [0.10, 0.60 ; 0.02)$ και $I_4 = [0.60, 1.00 ; 0.04]$, ακολουθώντας την προτεινόμενη διακριτοποίηση στη μελέτη του Tagaras (1998).

Ο πίνακας πιθανοτήτων μετάβασης που προκύπτει είναι ο ακόλουθος:

Σχήμα 1. Πίνακας πιθανοτήτων μετάβασης

		j									
		0		1		...		m		F	
i	P_{cz}	P_{az}	P_{cz}	P_{az}	P_{cz}	P_{az}	P_{cz}	P_{az}	P_{cz}	P_{az}	F
	0	P_{cz}									
P_{az}											
1	P_{cz}										
	P_{az}										
\dots	P_{cz}										
	P_{az}										
m	P_{cz}										
	P_{az}										
F	P_{cz}										
	P_{az}										

Πρόταση: Δύο διαφορετικά T^2 , T_1^2 και T_2^2 ($T_1^2 \neq T_2^2$), και συνεπώς, δύο διαφορετικά διανύσματα παρατηρήσεων, οδηγούν σε μετάβαση από ένα συγκεκριμένο διάνυσμα πιθανοτήτων $P = (p_1, p_2, \dots, p_m)$ είτε σε ένα ακριβώς ίδιο διάνυσμα πιθανοτήτων ή σε ένα εντελώς διαφορετικό διάνυσμα ($P'|T_1^2 \neq P'|T_2^2$).

Η απόδειξη της προαναφερθείσας πρότασης δεν παρατίθεται στην παρούσα εργασία για λόγους οικονομίας χώρου, αλλά τίθεται στη διάθεση του αναγνώστη κατόπιν επικοινωνίας με τους συγγραφείς.

Βάσει της ανωτέρω πρότασης, η πιθανότητα μετάβασης από ένα διάνυσμα P σε ένα P' μπορεί ισοδύναμα να υπολογιστεί μέσω της πιθανότητας μετάβασης από P σε $p'_i \in P'$, για κάθε $i \in [1, m]$. Στο υπόλοιπο της εργασίας και χωρίς οποιαδήποτε βλάβη της γενικότητας, η πιθανότητα μετάβασης από P σε P' θα υπολογίζεται μέσω της πιθανότητας $p'_i \in P'$.

Συμπερασματικά, η πιθανότητα μετάβασης από μία κατάσταση της Μαρκοβιανής αλυσίδας σε μία άλλη, μπορεί να αναλυθεί στις επιμέρους πιθανότητες: α. Πιθανότητα μετάβασης της πραγματικής κατάστασης της διαδικασίας από Y_{t-1} σε Y_t β. Πιθανότητα μετάβασης από ένα διάνυσμα P σε ένα συγκεκριμένο διάνυσμα P' , ή ισοδύναμα, όπως προαναφέρθηκε, σε ένα συγκεκριμένο $p'_i \in P'$. Ως εκ τούτου, οι πιθανότητες μετάβασης υπολογίζονται ως το γινόμενο των ανωτέρω επιμέρους πιθανοτήτων από την ακόλουθη σχέση:

$$Prob_{\substack{(Y_{t-1}, a_{t-1})^P \\ (Y_t, a_t)^{P'}}} = \begin{cases} p_k(h) \cdot P(p'_i | Y_{t-1} = i, Y_t = k, a_{t-1} = 0, P_{cz}) \\ p_k(h) \cdot P(p'_i | Y_{t-1} = i, Y_t = k, a_{t-1} = 1, p = 0) \\ p_k(h) \cdot P(p'_i | Y_{t-1} = F, Y_t = k, p = 0) \end{cases} \quad (6)$$

Οι πιθανότητες μετάβασης από P σε p'_i , με δεδομένες τιμές για τις μεταβλητές a_{t-1} , Y_{t-1} , Y_t and P , εξαρτώνται μόνο από το διάνυσμα παρατηρήσεων \bar{x}^* , και, ως εκ τούτου, από την ακριβή τιμή του $(T^2)^*$, η οποία προκύπτει, βάσει της σχέσης (5) για $k=1$, ως η μοναδική λύση των κατωτέρω εξισώσεων:

$$a_{t-1} = 0: \frac{\sum_{v=0}^m \sum_{z=0}^m p_z \cdot f_v \cdot p_z(h)}{\sum_{z=0}^m p_z \cdot f_1 \cdot p_z(h)} - \frac{1}{p_1} = 0 \quad (a_{t-1} = 1) \cup F: \frac{\sum_{v=0}^m f_v \cdot p_0(h)}{f_1 \cdot p_0(h)} - \frac{1}{p_1} = 0$$

Λόγω της δυσκολίας προσδιορισμού αναλυτικής λύσης, η τιμή του $(T^2)^*$ προσδιορίζεται αριθμητικά.

Η πιθανότητα για τη μετάβαση της διαδικασίας σε p_1' , βάσει της διακριτοποίησης των πιθανοτήτων που ακολουθήθηκε και περιγράφηκε ανωτέρω, δύναται να υπολογισθεί μέσω του εύρους όλων των πιθανών T^2 , εύρος που καθορίζεται από την τιμή του T^2 που οδηγεί στην κατώτερη τιμή του διακριτοποιημένου διαστήματος $(T_{lw}^2)^*$ και την αντίστοιχη που οδηγεί στην ανώτερη τιμή αυτού $(T_{up}^2)^*$. Κατά συνέπεια, λαμβάνοντας υπόψη την κατανομή του T^2 (Βλέπε Ενότητα 2.1, Υπόθεση (iii)), η εξίσωση (6), γίνεται:

$$Prob_{\substack{(Y_{t-1}, a_{t-1})P \\ (Y_t, a_t)P'}} = \begin{cases} p_k(h) \cdot \left[\begin{array}{l} F\left(\left(T_{up}^2\right)^*; p, d_i^2 \middle| Y_{t-1} = i, Y_t = k, a_{t-1} = 0, P_{cz}\right) - \\ F\left(\left(T_{lw}^2\right)^*; p, d_i^2 \middle| Y_{t-1} = i, Y_t = k, a_{t-1} = 0, P_{cz}\right) \end{array} \right] \\ p_0(h) \cdot \left[\begin{array}{l} F\left(\left(T_{up}^2\right)^*; p, d_0^2 \middle| Y_{t-1} = i, Y_t = k, a_{t-1} = 1, p = 0\right) - \\ F\left(\left(T_{lw}^2\right)^*; p, d_0^2 \middle| Y_{t-1} = i, Y_t = k, a_{t-1} = 1, p = 0\right) \end{array} \right] \\ p_k(h) \cdot \left[\begin{array}{l} F\left(\left(T_{up}^2\right)^*; p, d_0^2 \middle| Y_{t-1} = F, Y_t = k, p = 0\right) - \\ F\left(\left(T_{lw}^2\right)^*; p, d_0^2 \middle| Y_{t-1} = F, Y_t = k, p = 0\right) \end{array} \right] \end{cases}$$

Οι πιθανότητες σταθερής κατάστασης υπολογίζονται από την επίλυση του ακόλουθου συστήματος γραμμικών εξισώσεων:

$$\pi_{(Y_t, a_t)P'} = \sum_{Y_{t-1}=0}^{m \cup F} \sum_{a_{t-1}=0}^1 \sum_P Prob_{\substack{(Y_{t-1}, a_{t-1})P \\ (Y_t, a_t)P'}} \cdot \pi_{(Y_{t-1}, a_{t-1})P} \quad \text{και} \quad \sum_{Y_t=0}^{m \cup F} \sum_{a_t=0}^1 \sum_{P'} \pi_{(Y_t, a_t)P'} = 1$$

3. ΥΠΟΛΟΓΙΣΜΟΣ ΔΙΑΡΚΕΙΑΣ ΚΑΙ ΚΟΣΤΟΥΣ

3.1 Αναμενόμενη διάρκεια

Η αναμενόμενη διάρκεια ενός διαστήματος μεταξύ δύο διαδοχικών σημείων επιθεώρησης (expected time-ET) ισούται με το άθροισμα των ακόλουθων όρων πολλαπλασιαζόμενων με την αντίστοιχη πιθανότητα σταθερής κατάστασης κάθε περίπτωσης:

i) Αναμενόμενη διάρκεια λειτουργίας της παραγωγικής διαδικασίας σε κατάσταση στατιστικού ελέγχου: $\tau_{IC} = \int_0^h \exp\left(-\left(\nu_0 + \lambda_{F|0}\right) \cdot t\right) dt$.

ii) Αναμενόμενη διάρκεια λειτουργίας της παραγωγικής διαδικασίας σε κατάσταση εκτός στατιστικού ελέγχου όταν η διαδικασία λειτουργεί υπό έλεγχο κατά

την έναρξη του διαστήματος $\tau_{OOC/0}(h) = \int_0^h \sum_{y=1}^m \exp\left(-\left(v_0 + \lambda_{F|0}\right) \cdot t\right) \cdot \tau_{OOC/y}(h-t) dt$,

και όταν λειτουργεί υπό την επίδραση συστηματικής αιτίας ($\forall i \geq 0$)

$$\tau_{OOC/i}(h) = h \cdot \exp\left(-\left(v_i + \lambda_{F|i}\right) \cdot h\right) + \int_0^h t \cdot \lambda_{F|i} \cdot \exp\left(-\left(v_i + \lambda_{F|i}\right) \cdot t\right) dt + \int_0^h \sum_{y=i+1}^m \left[\lambda_{(i \rightarrow y)} \cdot \exp\left(-\left(v_i + \lambda_{F|i}\right) \cdot t\right) \cdot \left(t + \tau_{OOC/y}(h-t)\right) \right] dt.$$

iii) Καθυστέρηση λόγω εσφαλμένου συναγερμού T_0 .

iv) Καθυστέρηση λόγω αποκατάστασης της διαδικασίας σε κατάσταση στατιστικού ελέγχου, μέσω ενεργειών ΠΣ, η οποία για αποκατάσταση της διαδικασίας από την κατάσταση Y_i ($Y_i \in [1, m]$) συμβολίζεται με (T_{Y_i}) , ή ενεργειών $\Delta\Sigma(T_F)$.

Συμπερασματικά, η μεταβλητή ET ισούται με:

$$\begin{aligned} ET &= \sum_{P' \in P'_{az}} \pi_{(0,0)P'} \cdot \tau_{IC} + \sum_{Y_i=0}^m \sum_{P' \in P'_{az}} \pi_{(Y_i,1)P'} \cdot \tau_{IC} + \pi_F \cdot \tau_{IC} + \\ &+ \sum_{Y_i=0}^m \sum_{P' \in P'_{az}} \pi_{(Y_i,0)P'} \cdot \tau_{OOC/Y_i}(h) + \sum_{Y_i=0}^m \sum_{P' \in P'_{az}} \pi_{(Y_i,1)P'} \cdot \tau_{OOC/0}(h) + \pi_F \cdot \tau_{OOC/0}(h) + \\ &+ \sum_{P' \in P'_{az}} \pi_{(0,1)P'} \cdot T_0 + \sum_{Y_i=1}^m \sum_{P' \in P'_{az}} \pi_{(Y_i,1)P'} \cdot T_{Y_i} + \pi_F \cdot T_F \end{aligned}$$

Για έναν παραγωγικό κύκλο K σημείων επιθεώρησης, η συνολική του διάρκεια προκύπτει από τη σχέση $T = (K + 1) \cdot ET$.

3.2 Αναμενόμενο Κόστος

Το αναμενόμενο κόστος ενός διαστήματος μεταξύ δύο διαδοχικών σημείων επιθεώρησης (expected cost- EC) διακρίνεται στις ακόλουθες κατηγορίες:

- Κόστος Επιθεώρησης

Συνίσταται σε ένα σταθερό κόστος επιθεώρησης b και ένα μεταβλητό κόστος ανά επιθεωρούμενη μονάδα c : $EC_{inspection} = b + c$.

- Κόστος Ποιότητας

Το κόστος ποιότητας ενός δειγματοληπτικού διαστήματος ισούται με το αναμενόμενο κόστος εκτός ελέγχου λειτουργίας με κάθε μία πιθανή κατάσταση i να αποτελεί την αρχική κατάσταση του διαστήματος (M_i). Συνεπώς:

- Για $\forall i \in [0, m-1]$:

$$K_i(h) = M_i \cdot h \cdot \exp\left(-\left(v_i + \lambda_{F|i}\right) \cdot h\right) + \int_0^h t \cdot M_i \cdot \lambda_{F|i} \cdot \exp\left(-\left(v_i + \lambda_{F|i}\right) \cdot t\right) dt +$$

$$+ \int_0^h \sum_{k=i+1}^m \left[\lambda_{(i \rightarrow k)} \cdot \exp\left(-\left(v_i + \lambda_{F|i}\right) \cdot t\right) \cdot \left(tM_i + K_k(h-t)\right) \right] dt$$

- Για $i = m$ $K_m(h) = M_m \cdot h \cdot \exp\left(-\lambda_{F|m} \cdot h\right) + \int_0^h t \cdot M_m \cdot \lambda_{F|m} \cdot \exp\left(-\lambda_{F|m} \cdot t\right) dt$.

Κατά συνέπεια το κόστος ποιότητας προκύπτει από την ακόλουθη εξίσωση:

$$EC_{quality} = \sum_{Y_i=0}^m \sum_{P' \in P'_{cz}} \pi_{(Y_i,0)P'} \cdot K_{Y_i}(h) + \sum_{Y_i=0}^m \sum_{P' \in P'_{az}} \pi_{(Y_i,1)P'} \cdot K_0(h) + \pi_F \cdot K_0(h)$$

Για περισσότερες πληροφορίες αναφορικά με τον τρόπο εξαγωγής των ανωτέρω εξισώσεων, ο αναγνώστης παραπέμπεται στην εργασία των Nenes *et al.* (2015).

- Κόστος Συντήρησης

Τα επιμέρους συστατικά του κόστους συντήρησης είναι το κόστος των ενεργειών ΠΣ, το οποίο για αποκατάσταση της διαδικασίας από την κατάσταση Y_i ($Y_i \in [1, m]$) συμβολίζεται με (L_{Y_i}) , ΔΣ, το οποίο συμβολίζεται με L_F και το κόστος που βαρύνει την διαδικασία στην περίπτωση εσφαλμένου συναγερμού (L_0):

$$EC_{maintenance} = \sum_{Y_i=1}^m \sum_{P' \in P'_{az}} \pi_{(Y_i,1)P'} \cdot L_{Y_i} + \pi_F \cdot L_F + \sum_{P' \in P'_{az}} \pi_{(0,1)P'} \cdot L_0$$

- Κόστος Αποθεμάτων

Το κόστος αποθεμάτων αποτελείται από το μέσο κόστος προετοιμασίας ενός παραγωγικού κύκλου $EC_{setup} = \frac{D \cdot C_S}{K}$ και το μέσο κόστος διατήρησης αποθέματος

$$EC_{holding} = \left(\frac{(pr - dr)}{pr} \cdot \frac{K}{2} \cdot C_H \right) \cdot \text{Συνεπώς, } EC_{inventory} = \frac{D}{K} \cdot C_S + \left(1 - \frac{dr}{pr} \right) \cdot \frac{K}{2} \cdot C_H$$

3.3 Συνολικό Κόστος

Το συνολικό κόστος κάθε παραγωγικού κύκλου μπορεί να υπολογιστεί από το άθροισμα των προαναφερθέντων οικονομικών επιβαρύνσεων και εξάγεται από τη σχέση:

$$TC = K \cdot EC_{inspection} + (K+1) \cdot [EC_{quality} + EC_{maintenance}] + EC_{inventory}$$

4. ΠΡΟΒΛΗΜΑ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ

Το πρόβλημα βελτιστοποίησης διαμορφώνεται ως ένα πρόβλημα ελαχιστοποίησης του συνολικού κόστους του παραγωγικού κύκλου $\min_{p_c, K} TC$, μέσω της εύρεσης των βέλτιστων τιμών του ορίου ελέγχου και του αριθμού των σημείων επιθεωρήσεων ενός παραγωγικού κύκλου, ικανοποιώντας τους προφανείς περιορισμούς $0 < p_c < 1$ και $K \in Z^+$.

5. ΑΡΙΘΜΗΤΙΚΟ ΠΑΡΑΔΕΙΓΜΑ

Για την αναλυτική παρουσίαση της λειτουργίας του προτεινόμενου μοντέλου, εφαρμόζεται σε ένα αριθμητικό παράδειγμα για μία παραγωγική διαδικασία με χαρακτηριστικά που παρουσιάζονται στον παρακάτω πίνακα (Πίνακας 1):

Πίνακας 1.) Οικονομικές και Στατιστικές Παράμετροι αριθμητικού παραδείγματος

Ρυθμοί Έλευσης		Μέγεθος Μεταβολής		Χρονικές Καθυστερήσεις
$\lambda_{(0 \rightarrow 1)} = 0.01$	$\lambda_{F 0} = 0.001$	$\delta_{1,1} = 1.0$		$T_0 = 10.0$
$\lambda_{(0 \rightarrow 2)} = 0.005$	$\lambda_{F 1} = 0.002$	$\delta_{1,2} = 2.0$		$T_1 = 16.0$
$\lambda_{(1 \rightarrow 2)} = 0.0075$	$\lambda_{F 2} = 0.003$	$\delta_{2,1} = 1.5$		$T_2 = 26.0$
		$\delta_{2,2} = 3.0$		$T_F = 30.0$
Κόστη (σε Ευρώ)				
Εκτός Ελέγχου Λειτουργίας	Επιθεώρησης	Αποκατάστασης	Αποθεμάτων	
$M_1 = 150$	$c = 1$	$L_0 = 100$	$C_H = 40$	
$M_2 = 250$	$b = 2$	$L_1 = 200$	$C_S = 300$	
		$L_2 = 300$		
		$L_F = 500$		
Ρυθμός Παραγωγής $pr = 0.8$ τμχ/ώρα		Ρυθμός Ζήτησης $dr = 0.63$ τμχ/ώρα		
Διάστημα Επιθεώρησης $h = 75$ λεπτά		Ετήσια Επιθεώρηση $D = 1500$ τμχ		

Οι βέλτιστες παράμετροι σχεδίασης είναι $p_c^* = 0.058$ και $K^* = 31$, ενώ το συνολικό κόστος ανέρχεται σε $TC^* = 22131.22$ Ευρώ.

Για την εφαρμογή του προτεινόμενου μοντέλου αξιοποιείται το δισδιάστατο παράδειγμα που χρησιμοποιείται στις εργασίες των Lowry *et al.* (1992) και Crosier (1998), στο οποίο έχουμε 10 μετρήσεις από μια διαδικασία με μοναδιαίες διακυμάνσεις και συντελεστές συσχέτισης ίσους με 0.5. Μετά τις πέντε πρώτες μετρήσεις επιδρά στη διαδικασία η πρώτη συστηματική αιτία μεταβάλλοντας το διάνυσμα μέσω τιμών από $\mu_0 = (0,0)$ σε $\mu_1 = (1.0, 2.0)$.

Πίνακας 2.) Αριθμητικές Μετρήσεις και Αποτελέσματα παραδείγματος

<i>t</i>	Παρατηρήσεις		Prior Πιθανότητες			Posterior Πιθανότητες			OOC		
	X_1	X_2	T^2	p_0	p_1	p_2	p'_0	p'_1	p'_2	p'	a_t
1	-1.19	0.59	3.29	1.0	0.0	0.0	0.975	0.024	0.001	0.025	0
2	0.12	0.90	0.96	0.975	0.024	0.001	0.985	0.014	0.001	0.015	0
3	-1.69	0.40	4.92	0.985	0.014	0.001	0.921	0.075	0.004	0.079	0
4	0.30	0.46	0.22	0.921	0.075	0.004	0.982	0.017	0.001	0.018	0
5	0.89	-0.75	2.70	0.982	0.017	0.001	0.967	0.032	0.001	0.033	0
6	0.82	0.98	1.11	0.967	0.032	0.001	0.981	0.018	0.001	0.019	0
7	-0.30	2.28	7.96	0.981	0.018	0.001	0.776	0.207	0.017	0.224	1
8	0.63	1.75	3.14	-	-	-	-	-	-	-	-
9	1.56	1.58	3.29	-	-	-	-	-	-	-	-
10	1.46	3.05	9.31	-	-	-	-	-	-	-	-

Η διαδικασία ξεκινά σε στατιστικό έλεγχο και οι εκ των προτέρων πιθανότητες να λειτουργεί η διαδικασία υπό την επίδραση συστηματικής αιτίας κατά την έναρξη της πρώτης χρονικής περιόδου ισούνται με μηδέν ($p_0 = 1.0, p_1 = p_2 = 0.0$). Βάσει των παρατηρήσεων της πρώτης περιόδου υπολογίζεται το στατιστικό T^2 ίσο με 3.29 και οι εκ των προτέρων πιθανότητες μετατρέπονται στις εκ των υστέρων πιθανότητες $p'_0 = 0.975, p'_1 = 0.024, p'_2 = 0.001$, με τη συνολική πιθανότητα εκτός ελέγχου λειτουργίας να ανέρχεται σε 0.025 ($= 0.024 + 0.001$). Λόγω του γεγονότος ότι αυτή δεν ξεπερνά το όριο ελέγχου ($p_c = 0.058$), η διαδικασία συνεχίζει κανονικά τη λειτουργία της ($a_1 = 0$) και οι εκ των υστέρων πιθανότητες γίνονται εκ των προτέρων για την επόμενη χρονική περίοδο ($t=2$). Κατά το ίδιο σκεπτικό, υπολογίζονται, βάσει των παρατηρήσεων, οι εκ των υστέρων πιθανότητες της επόμενης περιόδου, προκύπτει η συνολική πιθανότητα εκτός ελέγχου λειτουργίας, συγκρίνεται με το όριο ελέγχου ($p' = 0.015 < p_c = 0.058$) και ξανά, το μοντέλο δεν επιτάσσει διακοπή της διαδικασίας. Η προαναφερθείσα αλληλουχία ενεργειών ακολουθείται χωρίς μεταβολή για τις έξι πρώτες χρονικές περιόδους. Ωστόσο, κατά την έβδομη χρονική περίοδο, το διάνυσμα παρατηρήσεων οδηγεί σε μία σχετικά μεγάλη τιμή για το στατιστικό του Hotelling ($T^2 = 7.96$) και, συνεπώς, σε μία αρκετά υψηλή πιθανότητα εκτός ελέγχου λειτουργίας, η οποία ξεπερνά το όριο ελέγχου ($p' = 0.224 \geq p_c = 0.058$). Σε αυτή την περίπτωση, το μοντέλο σημαίνει συναγερμό ($a_7 = 1$), η διαδικασία διακόπτεται για επιθεώρηση και ανακαλύπτεται η επίδραση της πρώτης συστηματικής αιτίας. Ως εκ τούτου, πραγματοποιούνται οι κατάλληλες ενέργειες ΠΣ και οι εκ των προτέρων πιθανότητες εκτός ελέγχου λειτουργίας για την όγδοη περίοδο τίθενται ίσες με μηδέν $p_0 = 1.0, p_1 = p_2 = 0.0$.

Επισημαίνεται δε συναφώς ότι στην περίπτωση όπου παρέλθουν 31 χρονικές περίοδοι ($K^* = 31$), τότε ο παραγωγικός κύκλος ολοκληρώνεται και ενέργειες ΠΣ αποκαθιστούν τη διαδικασία σε στατιστικό έλεγχο.

6. ΣΥΝΟΨΗ

Σκοπό της παρούσας εργασίας αποτελεί η ανάπτυξη ενός πρωτότυπου εργαλείου για σύγχρονες παραγωγικές διαδικασίες, όπου ένας υψηλός όγκος δεδομένων συλλέγεται μέσω μιας πληθώρας κατάλληλων αισθητήρων. Τα δεδομένα παρέχουν πληροφορίες για την κατάσταση της διαδικασίας σε πραγματικό χρόνο, γεγονός που διευκολύνει τη δημιουργία ενός αυτοματοποιημένου μηχανισμού λήψης αποφάσεων ο οποίος βασίζεται σε οικονομικά κριτήρια. Οι αποφάσεις που λαμβάνονται αφορούν την πολιτική παραγωγής, συγκεκριμένα τη βέλτιστη οικονομική ποσότητα παραγωγής, την πολιτική συντήρησης, μέσω του προσδιορισμού του ακριβή χρόνου που θα πρέπει ενέργειες ΠΣ να εφαρμοστούν και την πολιτική ποιότητας, μέσω του καθορισμού της ακριβούς τιμής των ορίων ελέγχου που θα ενεργοποιούν συναγερμό για τη διαδικασία. Οι τρεις προαναφερθείσες πολιτικές εξετάζονται σε κοινό πλαίσιο και μέσω ενός Μπεϋζιανού σχήματος ελέγχου. Η ρεαλιστική υπόθεση της ταυτόχρονης παρακολούθησης πολλαπλών αλληλοσχετιζόμενων χαρακτηριστικών ενισχύει σημαντικά την εφαρμοσιμότητα του προτεινόμενου μοντέλου.

Ευχαριστίες: Οι συγγραφείς ευχαριστούν θερμά τους κριτές για τις πολύ χρήσιμες και εποικοδομητικές τους παρατηρήσεις, οι οποίες συνέβαλαν σημαντικά στη βελτίωση της αρχικής έκδοσης της εργασίας.

ABSTRACT

In the era of smart manufacturing, a huge volume of data may be easily provided through multiple sensing devices, efficiently connected through the Internet of Things network. The challenge is to develop appropriate tools to exploit efficiently the sensing data by transforming it into optimal decisions that add real value to the monitored process. To this effect, a Bayesian model that optimizes jointly the interrelated process aspects of production, maintenance and quality through continuous, on-line monitoring in manufacturing environments is presented. The process output is fully characterized by multiple correlated quality characteristics and is subject to multiple quality shifts and failures. Optimal real-time decisions are dictated based on cost optimization criteria. These realistic assumptions lead to the development of a general, advanced, data-driven and widely applicable tool for modern production processes.

ΑΝΑΦΟΡΕΣ

- Ben-Daya, M. (1999). Integrated production maintenance and quality model for imperfect processes. *IIE Transactions*, **31**, 491-501.
- Crosier, R.B. (1998). Multivariate generalizations of cumulative sum quality-control schemes. *Technometrics*, **30** (3), 291-303.
- Fakher, H.B., Nourelfath, M., and Gendreau, M. (2018). Integrating production, maintenance and quality: A multi-period multi-product profit-maximization model. *Reliability Engineering and System Safety*, **170**, 191-201.

- Lowry, C.A., Woodall, W.H., Champ, C.W. and Rigdon, S.E. (1992). A multivariate exponentially weighted moving average control chart. *Technometrics*, **34** (1), 46-53.
- Montgomery, D.C. (2012). *Introduction to Statistical Quality Control*, New York: Wiley.
- Nenes, G., Tasiias, K.A. 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.
- Nourelfath, M., Nahas, N., and Ben-Daya, M. (2016). Integrated preventive maintenance and production decisions for imperfect processes. *Reliability Engineering and System Safety*, **148**, 21-31.
- Pan, E., Jin, Y., Wang, S., and Cang, T. (2012). An integrated EPQ model based on a control chart for an imperfect production process. *International Journal of Production Research*, **50** (23), 6999-7011.
- Rahim, A.M., and Ben-Daya, M. (1998). A generalized economic model for joint determination of production run: inspection schedule and control chart design. *International Journal of Production Research*, **36**, 277-289.
- Salmansia, A., Abdzadeh, B., and Namdar, M. (2017). A joint design of production run length, maintenance policy and control chart with multiple assignable causes. *Journal of Manufacturing Systems*, **42**, 44-56.
- Salmansia, A., Kaveie, M., and Namdar, M. (2018). An integrated production and maintenance planning model under VP-T2 Hotelling chart. *Computers and Industrial Engineering*, **118**, 89-103.
- Tagaras, G. (1996). Dynamic control charts for finite production runs. *European Journal of Operational Research*, **91** (1), 38-55.
- Variyath, A.M., and Vattathoor, J. (2014). Robust control charts for monitoring process variability in Phase-I multivariate individual observations. *Quality and Reliability Engineering International*, **30** (6), 795-812.
- Vargas, J.A.N. (2003). Robust estimation in multivariate control charts for individual observations. *Journal of Quality Technology*, **35** (4), 367-376.
- Yeh, A.B., Huwang, L., and Chien-Wei, W.U. (2005). A multivariate EWMA control chart for monitoring process variability with individual observations. *IIE Transactions*, **37** (11), 1023-1035.
- Zhang, G. and Chang, S.I. (2008). Multivariate EWMA control charts using individual observations for process mean and variance monitoring and diagnosis. *International Journal of Production Research*, **46** (24), 6855-6881.



Η ΣΧΕΣΗ ΑΝΑΜΕΣΑ ΣΤΟΝ S&P 500 ΚΑΙ VIX ΑΝΑΛΥΣΗ ΜΕ ΣΥΖΕΥΞΕΙΣ

Δ. Τζαφέρη¹, Γ. Τσακλίδης², Π. Φουσέκης¹

¹Τμήμα Οικονομικών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
{dtzaferi,fousekis}@econ.auth.gr

²Τμήμα Μαθηματικών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
tsaklid@math.auth.gr

ΠΕΡΙΛΗΨΗ

Η σχέση ανάμεσα στις αποδόσεις των μετοχών και την μεταβλητότητα τους αποτελεί ένα ιδιαίτερα ενδιαφέρον ερευνητικό ζήτημα της Χρηματοοικονομικής Ανάλυσης και συγκεκριμένα των πεδίων της βελτιστοποίησης του χαρτοφυλακίου (Markowitz, 1952), της δίκαιης τιμολόγησης και της αντισταθμιστικής πολιτικής (hedging). Στην εργασία αυτή χρησιμοποιούμε δεδομένα από το δείκτη τιμών (stock price index) S&P 500 και το σχετικό δείκτη μεταβλητότητας (implied volatility index) VIX για τον προσδιορισμό της ολικής (global) και της τοπικής (local) εξάρτησης ανάμεσα στην απόδοση και τις αντιλήψεις των επενδυτών σχετικά με το επίπεδο του κινδύνου (volatility/risk). Η ανάλυση βασίζεται στο στατιστικό εργαλείο των συζεύξεων (copulas) και σε ημερήσιες παρατηρήσεις από ορίζοντα 2/1/2004 μέχρι 31/8/2017. Τα εμπειρικά αποτελέσματα δείχνουν ότι: (α) ο S&P 500 και ο VIX παρουσιάζουν αντίρροπη ολική εξάρτηση και σημαντικές τοπικές εξαρτήσεις στα άκρα της από κοινού κατανομής τους. Επομένως, βρίσκονται σε αντίθεση με τις προβλέψεις του υποδείγματος CAPM (Sharpe, 1964) και σε συμφωνία με τις συμπεριφορικές εξηγήσεις (behavioral postulates) των Hibbert et al. (2008) και Low (2004). (β) η σύζευξη με την καλύτερη προσαρμογή είναι η στραμμένη κατά 270 μοίρες BB7.

Λέξεις Κλειδιά: αποδόσεις μετοχών, μεταβλητότητα αποδόσεων, συζεύξεις, γωνιακές πιθανότητες και όρια, στροφές συζεύξεων, Αρχιμήδειες συζεύξεις.

1. ΕΙΣΑΓΩΓΗ

Η σχέση των αποδόσεων των μετοχών με τις διακυμάνσεις αυτών απασχόλησε ιδιαίτερα την επιστημονική κοινότητα λόγω του καθοριστικού της ρόλου στη διαμόρφωση της χρηματοοικονομικής πραγματικότητας. Η εύρεση του βέλτιστου χαρτοφυλακίου, η δίκαιη τιμολόγηση και η αντισταθμιστική πολιτική είναι ορισμένα

μόνο από τα ζητήματα που επηρεάζονται άμεσα από τη σχέση ανάμεσα στις δύο εξεταζόμενες μεταβλητές. Το ενδιαφέρον μελέτης της ανωτέρω σχέσης έγκειται στο γεγονός ότι οι απόψεις των βιβλιογραφικών αναφορών δίστανται. Παρόλο που το μοντέλο μέσου διακύμανσης του Markowitz (1952) και το μοντέλο αποτίμησης των περιουσιακών τίτλων του Sharpe (1964) αποφαίνονται πως οι αποδόσεις των μετοχών σχετίζονται θετικά με τις αντίστοιχες διακυμάνσεις τους, υπάρχουν αρκετές εμπειρικές μελέτες που καταλήγουν στο ακριβώς αντίθετο συμπέρασμα, δηλαδή ότι οι δύο μεταβλητές εξελίσσονται αντίρροπα. Ο Black το 1976 με τη θεωρία του «leverage» συμπέρανε πως οι πτώσεις των τιμών των μετοχών είχαν ως επακόλουθο την αύξηση του ρίσκου που αναλάμβανε μία επιχείρηση, δηλαδή την αύξηση της αντίστοιχης μεταβλητότητας των τιμών. Όμοια οι Campbell και Hentschel το 1992 επιβεβαίωσαν την αντίρροπη σχέση των δύο μεγεθών αυτή τη φορά με αντίθετη κατεύθυνση αιτιότητας. Συγκεκριμένα η αποκαλούμενη υπόθεση του μοντέλου τους «volatility feedback» ισχυρίζονταν ότι μία μη αναμενόμενη άνοδος των διακυμάνσεων και επομένως του ρίσκου πρόβλεψη των τιμών των μετοχών θα είχε ως αποτέλεσμα την μείωση των αποδόσεων λόγω φόβου των επενδυτών για το παραπάνω ρίσκο. Ακολούθησαν και άλλες έρευνες, σύμφωνα με το Fousekis (2019), που διερεύνησαν τη σχέση ανάμεσα στις αποδόσεις και στην μεταβλητότητα των τιμών μετοχών καθώς και την κατεύθυνση εξέλιξης αυτής εξετάζοντας κάθε φορά άλλα ζεύγη οικονομικών δεικτών (Badshah, 2013, Hibbert *et al.*, 2008, Low, 2004, Bakaert και Wu, 2000, Bollerslev *et al.*, 1996) προσπαθώντας να ερμηνεύσουν την εκάστοτε σχέση με βάση τη συμπεριφορά των επενδυτών (Behavioral Finance). Επιθυμητή στρατηγική των επενδυτών είναι εκείνη που μεγιστοποιεί τις αποδόσεις των μετοχών τους ενώ παράλληλα ελαχιστοποιεί την μεταβλητότητά τους. Η συμπεριφορά αυτή αναλύθηκε εκτενώς στις εργασίες των Low (2004) και Hibbert *et al.* (2008) κατά τις οποίες οι αποδόσεις των μετοχών και η μεταβλητότητά τους εμφανίζουν ισχυρή αντίρροπη και μη γραμμική εξάρτηση (downward sloping reclined S-curve).

Στην παρούσα έρευνα εξετάζεται η αντίρροπη σχέση ανάμεσα στις ημερήσιες αποδόσεις των μετοχών (οικονομικός δείκτης S&P 500) και στις διακυμάνσεις αυτών (VIX) με τη χρήση στροφών συζεύξεων. Το πλεονέκτημα των συζεύξεων είναι ότι είναι πολυδιάστατες κατανομές. Παράλληλα, οι τιμές των περιθώριων μεταβλητών περιορίζονται στο $[0,1]^n$, n : το πλήθος των μεταβλητών, αντί του πραγματικού n -διάστατου χώρου. Τέλος, το στατιστικό εργαλείο των συζεύξεων είναι χρήσιμο γιατί μας δίνει τη δυνατότητα να μελετήσουμε τη συμπεριφορά των περιθώριων μεταβλητών οριακά (δηλ. στα άκρα των διαστημάτων ορισμού). Έτσι, στην εργασία αυτή εκτός από την εκτίμηση της σύζευξης για το ζεύγος (αποδόσεις, μεταβλητότητα αυτών) τόσο παραμετρικά όσο και εμπειρικά, γίνεται και μια εκτίμηση της οριακής από κοινού συμπεριφοράς των δύο μεταβλητών (υπολογισμός γωνιακών πιθανοτήτων). Επισημαίνεται ότι το μαθηματικό αυτό εργαλείο χρησιμοποιήθηκε πρώτη φορά για τη διερεύνηση του παρόντος προβλήματος.

Στο επόμενο κεφάλαιο παρουσιάζονται το μαθηματικό εργαλείο των σύζευξεων, η περιγραφή των δεδομένων και τα αποτελέσματα της ανάλυσης.

2. ΚΥΡΙΟ ΚΕΙΜΕΝΟ

2.1 Σύζευξη

2.1.1. Ορισμός σύζευξης και εμπειρικής σύζευξης

Με τον όρο σύζευξη δύο διαστάσεων (Sklar, 1959) καλούμε μια συνάρτηση $C : [0,1]^2 \rightarrow [0,1]^2$ για την οποία ικανοποιούνται οι ακόλουθες ιδιότητες:

1. $\forall (u, v) \in [0,1]^2, C(u, 0) = C(0, v) = 0 \wedge C(u, 1) = u, C(1, v) = v.$
2. $\forall (u_1, v_1), (u_2, v_2) \in [0,1], u_1 \leq u_2 \wedge v_1 \leq v_2, C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$

Ο πιο άμεσος μη παραμετρικός εκτιμητής μιας σύζευξης είναι η εμπειρική σύζευξη, για τον ορισμό της οποίας χρησιμοποιείται το θεώρημα του Sklar (1959) που περιγράφεται ως εξής:

Ας είναι X και Y τυχαίες μεταβλητές με κοινή συνάρτηση κατανομής $H(x, y)$ και περιθώριες συναρτήσεις κατανομών F και G , αντίστοιχα. Τότε, υπάρχει μια σύζευξη C τέτοια ώστε

$$H(x, y) = C(F(x), G(y)), \quad \forall (x, y) \in \square$$

Αντίστροφα, για κάθε συνάρτηση κατανομής $F(x)$ και $G(y)$ και κάθε σύζευξη, η συνάρτηση $C(u, v)$ που ορίστηκε παραπάνω είναι μια διδιάστατη συνάρτηση κατανομής $H(x, y)$ με περιθώριες τις $F(x)$ και $G(y)$ αντίστοιχα και είναι της μορφής

$$C(u, v) = H(F^{-1}(u), G^{-1}(v)), \quad (u, v) \in [0,1]^2$$

Έτσι, η εμπειρική σύζευξη γράφεται ως εξής

$$\begin{aligned} C_n^{(E)}(u, v) &= H_n(F_n^{-1}(u), G_n^{-1}(v)) = \frac{1}{n} \sum_{i=1}^n I\{X_i \leq F_n^{-1}(u), Y_i \leq G_n^{-1}(v)\} \\ &= \frac{1}{n} \sum_{i=1}^n I\{F_n(X_i) \leq u, G_n(Y_i) \leq v\} = \frac{1}{n} \sum_{i=1}^n I\{\hat{U}_i \leq u, \hat{V}_i \leq v\} \end{aligned}$$

όπου H_n είναι η εμπειρική από κοινού κατανομή και F_n, G_n είναι οι εμπειρικές περιθώριες κατανομές. Αντί για $(\hat{U}_i, \hat{V}_i) = (F_n(X_i), G_n(Y_i))$ ο Omelka και οι συνεργάτες του (2009) πρότειναν τη βελτίωση $(\hat{U}_i, \hat{V}_i) = (\frac{n}{n+1} F_n(X_i), \frac{n}{n+1} G_n(Y_i))$

κατά την οποία πραγματοποιείται μία μικρή μετατόπιση των παρατηρήσεων προς τα αριστερά του διαστήματος $[0,1]$, δίνοντας πιο ευσταθή αποτελέσματα.

2.1.2. Στροφή σύζευξης

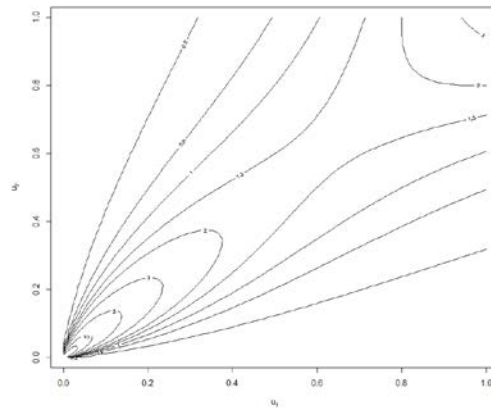
Οι συζεύξεις C^{--}, C^{+-}, C^{++} καλούνται συστραμμένες (rotated) συζεύξεις και συγκεκριμένα η $1^{\text{η}}$ καλείται συστραμμένη σύζευξη 180 μοιρών και οι άλλες δύο συστραμμένες συζεύξεις 90 μοιρών με αντίθετες κατευθύνσεις (δηλ. 90 και 270 μοιρών αντίστοιχα):

$$C^{--}(u, v) = u + v - 1 + C(1-u, 1-v), \quad C^{+-}(u, v) = v - C(1-u, v), \quad C^{++}(u, v) = u - C(u, 1-v).$$

2.1.3. Γωνιακές πιθανότητες σύζευξης και στροφών αυτής

Ορισμός, (Cech, 2006): Θεωρούμε δύο τυχαίες μεταβλητές U, V που ακολουθούν την ομοιόμορφη κατανομή σύζευξης C στο $(0,1)$. Τότε, μπορούμε να ορίσουμε τέσσερις γωνιακές πιθανότητες, δοθέντος $\alpha \in (0, 0.5]$:

Εικόνα 1. Ισοϋψείς καμπύλες του Clayton ($\theta=2.5$)



$$P_1^C := P(U \leq a | V \leq a) = \frac{P(U \leq a, V \leq a)}{P(V \leq a)} = \frac{C(a, a)}{a}$$

$$P_2^C := P(U \leq a | V > 1-a) = \frac{P(U \leq a, V > 1-a)}{1 - P(V \leq 1-a)} = \frac{a - C(a, 1-a)}{a}$$

$$P_3^C := P(U > 1-a | V > 1-a) = \frac{P(U > 1-a, V > 1-a)}{1 - P(V \leq 1-a)} = \frac{2a - 1 + C(1-a, 1-a)}{a}$$

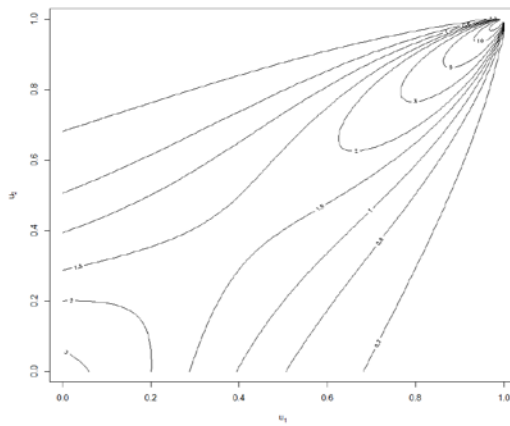
$$P_4^C := P(U > 1-a | V \leq a) = \frac{P(U > 1-a, V \leq a)}{P(V \leq a)} = \frac{a - C(1-a, a)}{a}$$

Σημειώνεται ότι οι P_1^C, P_3^C ονομάζονται και κάτω, άνω (lower, upper) γωνιακές πιθανότητες αντίστοιχα.

Πόρισμα Α

Θεωρούμε δύο τυχαίες μεταβλητές U, V , που ακολουθούν την ομοιόμορφη κατανομή σύζευξης C^{--} στο $(0,1)$. Τότε, οι τέσσερις γωνιακές πιθανότητες του C^{--} προσδιορίζονται από τις αντίστοιχες γωνιακές πιθανότητες του C , ως εξής :
(Υπενθυμίζουμε ότι C^{--} είναι η στροφή της σύζευξης κατά 180 μοίρες και δίνεται από τον τύπο: $C^{--}(u, v) = u + v - 1 + C(1-u, 1-v)$)

Εικόνα 2. Ισοϋψείς καμπύλες του Rotated 180o Clayton(2.5)



$$P_1^{C^{--}} = \frac{C^{--}(a, a)}{a} = P_3^C$$

$$P_2^{C^{--}} = \frac{a - C^{--}(a, 1-a)}{a} = \frac{a - C(1-a, a)}{a} = P_4^C$$

$$P_3^{C^{--}} = \frac{2a - 1 + C^{--}(1-a, 1-a)}{a} = \frac{C(a, a)}{a} = P_1^C$$

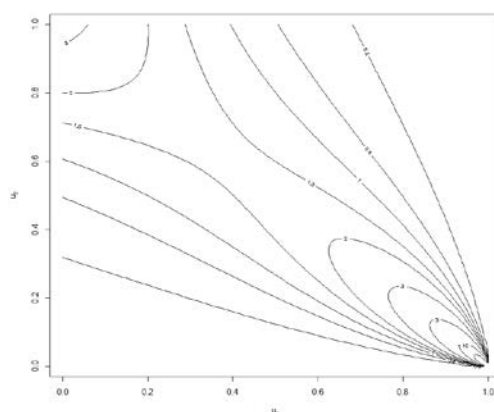
$$P_4^{C^{--}} = \frac{a - C^{--}(1-a, a)}{a} = \frac{a - C(a, 1-a)}{a} = P_2^C$$

Πόρισμα Β

Θεωρούμε δύο τυχαίες μεταβλητές U, V που ακολουθούν την ομοιόμορφη κατανομή σύζευξης C^{++} στο $(0,1)$. Τότε, οι τέσσερις γωνιακές πιθανότητες του C^{++} προσδιορίζονται από τις αντίστοιχες γωνιακές πιθανότητες του C , ως εξής :

(Υπενθυμίζουμε ότι C^{++} είναι η στροφή της σύζευξης κατά 90 μοίρες και δίνεται από τον τύπο $C^{++}(u, v) = v - C(1-u, v)$:)

Εικόνα 3. Ισοϋψείς καμπύλες του Rotated 90o Clayton(1000,-2.5)



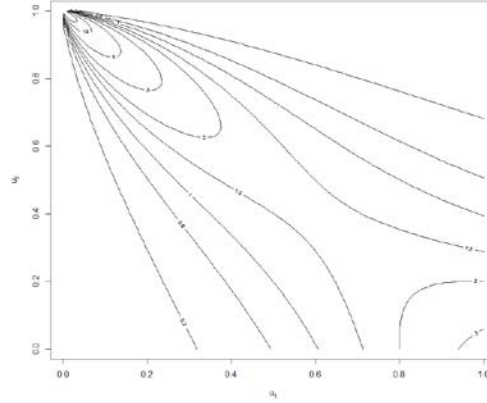
$$P_1^{C^{++}} = \frac{C^{++}(a, a)}{a} = \frac{a - C(1-a, a)}{a} = P_4^C$$
$$P_2^{C^{++}} = \frac{a - C^{++}(a, 1-a)}{a} = \frac{2a - 1 + C(1-a, 1-a)}{a} = P_3^C$$
$$P_3^{C^{++}} = \frac{2a - 1 + C^{++}(1-a, 1-a)}{a} = \frac{a - C(a, 1-a)}{a} = P_2^C$$
$$P_4^{C^{++}} = \frac{a - C^{++}(1-a, a)}{a} = \frac{C(a, a)}{a} = P_1^C$$

Πόρισμα Γ

Θεωρούμε δύο τυχαίες μεταβλητές U, V , που ακολουθούν την ομοιόμορφη κατανομή σύζευξης C^{+-} στο $(0,1)$. Τότε, οι τέσσερις γωνιακές πιθανότητες του προσδιορίζονται από τις αντίστοιχες γωνιακές πιθανότητες του C , ως εξής :

(Υπενθυμίζουμε ότι C^{+-} είναι η στροφή της σύζευξης κατά 270 μοίρες και δίνεται από τον τύπο: $C^{+-}(u, v) = u - C(u, 1-v)$)

Εικόνα 4. Ισοϋψείς καμπύλες του rotated 270ο Clayton(1000,-2.5)



$$P_1^{C^{+-}} = \frac{C^{+-}(a, a)}{a} = \frac{a - C(a, 1-a)}{a} = P_2^C$$

$$P_2^{C^{+-}} = \frac{a - C^{+-}(a, 1-a)}{a} = \frac{C(a, a)}{a} = P_1^C$$

$$P_3^{C^{+-}} = \frac{2a - 1 + C^{+-}(1-a, 1-a)}{a} = \frac{a - C(1-a, a)}{a} = P_4^C$$

$$P_4^{C^{+-}} = \frac{a - C^{+-}(1-a, a)}{a} = \frac{2a - 1 + C(1-a, 1-a)}{a} = P_3^C$$

Η χρησιμότητα των γωνιακών πιθανοτήτων έγκειται στο ότι περιγράφουν την εξάρτηση των μεταβλητών στις τέσσερις γωνίες του $[0, 1]^2$.

2.1.5. Οριακή εξάρτηση σύζευξης και στροφής αυτής

Με τον όρο οριακή εξάρτηση σύζευξης εννοούμε την ύπαρξη εξάρτησης των εκάστοτε εξεταζόμενων μεταβλητών στην πάνω δεξιά γωνία ή στην κάτω αριστερά γωνία του $[0, 1]^2$. Δοθέντων δύο τυχαίων μεταβλητών $X \square F, Y \square G$ με σύζευξη C προσδιορίζονται οι συντελεστές της οριακής εξάρτησης:

$$\lambda_L := \lim_{u \rightarrow 0} P_1^C, \lambda_U := \lim_{u \rightarrow 0} P_3^C$$

Η σύζευξη C λέγεται ότι έχει κάτω (άνω) οριακή εξάρτηση (lower (upper) tail dependence) αν και μόνο αν $\lambda_L \neq 0$ ($\lambda_U \neq 0$), δηλ. λαμβάνει θετικές τιμές. Αν πάλι το $\lambda_L = 0$ ($\lambda_U = 0$), τότε το ζεύγος (X, Y) καλείται οριακά ανεξάρτητο (tail-independent). Η σημαντικότητα υπολογισμού της οριακής εξάρτησης έγκειται στο

γεγονός ότι εξετάζεται η πιθανότητα και οι 2 μεταβλητές να λαμβάνουν ακραίες τιμές. Σε αυτό το γεγονός έγκειται και η εφαρμογή της στην εκτίμηση ρίσκου.

Με τον όρο οριακή εξάρτηση στροφής σύζευξης εννοούμε τους συντελεστές

$$\lambda_{LU} := \lim_{u \rightarrow 0} P_2^C, \lambda_{UL} := \lim_{u \rightarrow 0} P_4^C \text{ αντίστοιχα.}$$

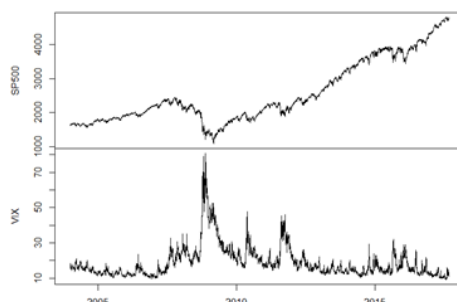
2.1.6. Αρχιμήδειες συζεύξεις

Η κατηγορία αυτή συζεύξεων αποτελεί βιβλιογραφικά μια από τις πιο εύχρηστες και εύκολες υπολογιστικά οικογένειες συζεύξεων. Μεταξύ των πιο χαρακτηριστικών και γνωστών Αρχιμήδειων συζεύξεων είναι οι Clayton, Frank, Gumbel, Joe, BB1, BB7 κ.α. (Nelsen, 2006).

2.2 Τα δεδομένα, εμπειρική ανάλυση

Η παρούσα εργασία εστιάζει το ενδιαφέρον της στην μελέτη της σχέσης ανάμεσα σε ένα ζεύγος οικονομικών δεικτών που μετρούν τις αποδόσεις των μετοχών και των διακυμάνσεων τους, το (S&P 500, VIX). Τα δεδομένα είναι ημερήσια και αφορούν το χρονικό διάστημα 2/1/2004 έως και 31/8/2017 δηλαδή επρόκειτο για ένα δείγμα 3444 παρατηρήσεων.

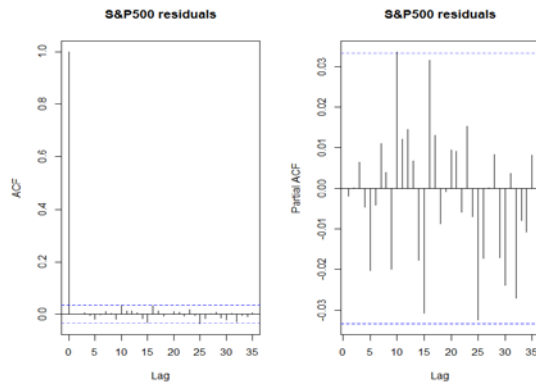
Εικόνα 5. Οι τιμές των S&P 500- VIX



Προκειμένου να ανιχνευθεί η σχέση των δύο δεικτών ακολουθείται προσαρμογή από συζεύξεις, εφόσον πρώτα τα δεδομένα μετασχηματισθούν κατάλληλα και ακολουθούν την ομοιόμορφη κατανομή (Cruz, et.al, 2015). Ο μετασχηματισμός περιλαμβάνει λογαρίθμηση, πρώτες διαφορές, απομάκρυνση αυτοσυσχετίσεων και μερικών αυτοσυσχετίσεων με μοντέλα ARCH, GARCH και τέλος η τάξη των υπολοίπων διαιρείται με το πλήθος αυτών αυξημένο κατά ένα.

Για τη χρονοσειρά S&P500 επιλέγοντας ένα μοντέλο ARMA(2,1) και GARCH(1,1) εξασφαλίζουμε την απαλλαγή της από αυτοσυσχετίσεις μιας και τόσο ο οπτικός έλεγχος όσο και ο έλεγχος Box-Pierce Portmanteau δεν απορρίπτει την μηδενική υπόθεση περί μηδενικής αυτοσυσχετίσης.

Εικόνα 6. ACF, PACF S&P 500



Πίνακας 1. Portmanteau test

Υστερήσεις	Στατιστικό ελέγχου	p-value
1	0.013	0.91
6	1.715	0.94
12	8.712	0.73
18	17.304	0.50
24	19.301	0.74

Το μοντέλο αυτό όμως μας εξασφαλίζει και την απαλλαγή από ARCH effects σύμφωνα και με τον ομώνυμο έλεγχο (Lagrange multiplier test for conditional heteroscedasticity of Engle (1982), Tsay (2005, pp. 101-102)). Τα αποτελέσματα του ελέγχου αυτού συνοψίζονται στον ακόλουθο πίνακα

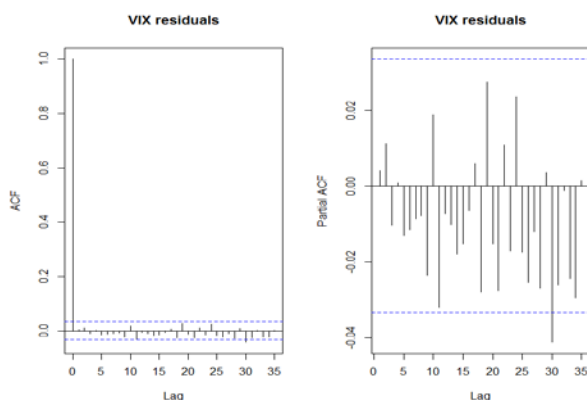
Πίνακας 2. ARCH test

Στατιστικό ελέγχου (χ^2)	Βαθμοί ελευθερίας	p-value
1.608	1	0.20
7.256	6	0.30
14.311	12	0.28
21.689	18	0.25
26.455	24	0.33

Ακολουθώντας την ίδια διαδικασία για τη χρονοσειρά VIX, επιλέγουμε ένα μοντέλο ARMA(8,2) και GARCH(1,1) εξασφαλίζοντας την απαλλαγή των δεδομένων μας

από αυτοσυσχετίσεις μιας και τόσο ο οπτικός έλεγχος όσο και ο έλεγχος Box-Pierce Portmanteau δεν απορρίπτει την μηδενική υπόθεση περί μηδενικής αυτοσυσχέτισης, αλλά και από ARCH effects, γεγονός που φανερώνεται από τον ομώνυμο έλεγχο.

Εικόνα 7. ACF, PACF VIX



Πίνακας 3. Portmanteau test

Υστερήσεις	Στατιστικό ελέγχου	p-value
1	0.056	0.81
6	1.889	0.93
12	9.110	0.69
18	13.902	0.74
24	22.818	0.53

Πίνακας 4. ARCH test

Στατιστικό ελέγχου (X^2)	Βαθμοί ελευθερίας	p-value
0.412	1	0.52
1.868	6	0.93
5.274	12	0.95
8.993	18	0.96
10.677	24	0.99

Οι περιθώριες μεταβλητές της σύζευξης λαμβάνονται παίρνοντας τις βαθμίδες των καταλοίπων των υποδειγμάτων που εφαρμόστηκαν στις χρονοσειρές S&P 500 και

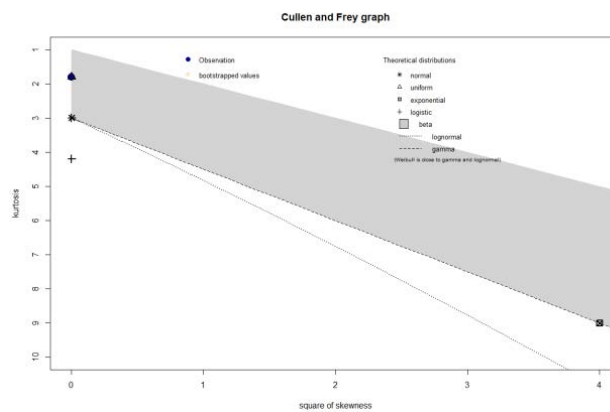
VIX. Συγκεκριμένα, προκύπτουν από το λόγο $\frac{rank(\text{καταλοίπων})}{\text{πλήθος}(\text{καταλοίπων})+1}$. Οι νέες

μεταβλητές r_{sp} , r_{vix} ακολουθούν την ομοιόμορφη κατανομή σύμφωνα με το Kolmogorov-Smirnov test (K-S test) και το γράφημα των Cullen-Frey (εκτίμηση κατανομής δεδομένων με βάση τη λοξότητα και την κύρτωση).

Πίνακας 5. *p-value* Kolmogorov Smirnov για ομοιόμορφη κατανομή

Περιθώριες μεταβλητές σύζευξης	p-value
r_{sp}	0.84
r_{vix}	0.99

Εικόνα 8. Cullen-Frey γραφήματα των r_{sp} r_{vix}



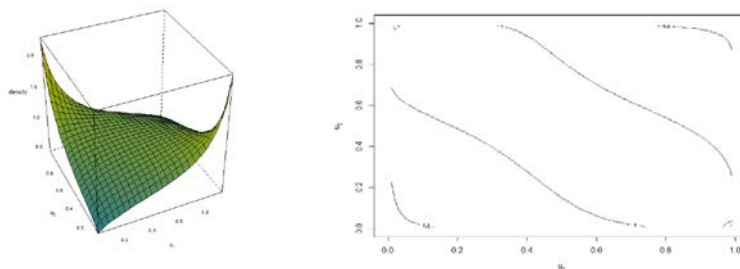
Με κριτήριο το BIC, $\frac{-2\log(L(k))}{n} + \frac{k \log(n)}{n}$ όπου $L(k)$ είναι η μέγιστη πιθανοφάνεια του μοντέλου της κάθε σύζευξης με k παραμέτρους και n το μέγεθος του δείγματος (πλήθος παρατηρήσεων) και με χρήση του πακέτου Vinetorula της R καταλήγουμε πως τα δεδομένα μας προσαρμόζονται καλύτερα στην στραμμένη κατά 270 μοίρες BB7 οικογένεια, δηλ. η αθροιστική κατανομή σύζευξης δίνεται από τον τύπο

$$C^{+-}(u, v) = u - (1 - (1 - ((1 - (1 - u)^\theta)^{-\delta} + (1 - v^\theta)^{-\delta} - 1)^{-1/\delta})^{-1/\theta})$$

όπου δ, θ παράμετροι της κατανομής. Στη συγκεκριμένη περίπτωση, οι παράμετροι εκτιμώνται να λαμβάνουν με ακρίβεια δύο δεκαδικών ψηφίων τις τιμές 0.15 και 1.11

αντίστοιχα.¹ Η προσαρμογή αυτή επιβεβαιώνει την υπόθεση περί αντίρροπης εξέλιξης των δύο οικονομικών δεικτών, δηλαδή η άνοδος του δείκτη S&P 500 συνεπάγεται την μείωση του δείκτη VIX στο χρονικό ορίζοντα 2/1/2004-31/8/2017.

Εικόνα 9. Η προσαρμοσμένη στα δεδομένα BB7 στραμμένη κατά 270 μοίρες συνάρτηση πυκνότητας πιθανότητας και οι ισοϋψείς καμπύλες



Σε επόμενο βήμα κατασκευάζεται η εμπειρική κατανομή των δεδομένων και ελέγχεται η υπόθεση προσαρμογής της ανωτέρω σύζευξης. Με άλλα λόγια κατασκευάζεται εμπειρική σύζευξη με σκοπό τη σύγκριση της με την παραμετρική που ανέδειξε το BIC. Η σύγκριση των δύο συζεύξεων γίνεται με τους στατιστικούς ελέγχους Kolmogorov-Smirnov και Anderson-Darling όσο και με τη σύγκριση των γωνιακών πιθανοτήτων τους (σχετικό σφάλμα). Τα p-value προσαρμογής είναι 0.97 και 0.88 αντίστοιχα καθιστώντας μη ικανή την απόρριψη της μηδενικής υπόθεσης του ελέγχου περί ιδίων συζεύξεων. Αντίστοιχα τα αποτελέσματα των γωνιακών πιθανοτήτων φαίνονται στους πίνακες που ακολουθούν.

Πίνακας 6. Γωνιακές πιθανότητες παραμετρικού μοντέλου.

α	P_1^C	P_2^C	P_3^C	P_4^C
0.05	0.019	0.161	0.019	0.176
0.10	0.047	0.221	0.047	0.222
0.20	0.119	0.319	0.119	0.313
0.40	0.311	0.493	0.311	0.491

¹ Σημειώνεται ότι εξετάστηκε η προσαρμογή από 40 Αρχιμήδειες συζεύξεις των δεδομένων μας (μεταξύ των οποίων Gumbel, Clayton, BB1, BB6, BB7, Joe κτλ. και η στραμμένη κατά 270° BB7 σύζευξη βρέθηκε να έχει την καλύτερη προσαρμογή. Ακόμη πέρα από το κριτήριο BIC η σύζευξη με εφαρμογή του ελέγχου καλής προσαρμογής που παρέχει το πακέτο VineCopula (Schepsmeier, 2013) έδειξε καλή προσαρμογή (p_value >0.05).

Πίνακας 7. Γωνιακές πιθανότητες εμπειρικού μοντέλου.

α	P_1^C	P_2^C	P_3^C	P_4^C
0.05	0.017	0.140	0.013	0.146
0.10	0.044	0.195	0.069	0.204
0.20	0.112	0.315	0.151	0.310
0.40	0.298	0.516	0.305	0.496

Πίνακας 8. Σχετικό σφάλμα γωνιακών πιθανοτήτων παραμετρικού και εμπειρικού μοντέλου.

α	P_1^C	P_2^C	P_3^C	P_4^C
0.05	6.01%	12.95%	27.94%	16.98%
0.10	6.53%	11.54%	32.00%	8.09%
0.20	6.30%	1.54%	21.13%	0.97%
0.40	4.07%	4.46%	1.88%	1.10%

Σε γενικές γραμμές παρατηρείται μικρό σχετικό σφάλμα εκτός από την περίπτωση όπου το α είναι πολύ μικρό γεγονός που δικαιολογείται από τις λίγες παρατηρήσεις του δείγματος.

2.3 Συμπεράσματα εμπειρικής μελέτης

Η παρούσα εργασία με μαθηματικό εργαλείο τις συζεύξεις επιβεβαίωσε ότι ο S&P 500 και ο VIX παρουσιάζουν αντίρροπη ολική εξάρτηση, εφόσον η σύζευξη με την καλύτερη προσαρμογή είναι η στραμμένη κατά 270 μοίρες BB7, και σημαντικές τοπικές εξαρτήσεις στα άκρα της από κοινού κατανομής τους και συγκεκριμένα $\lambda_{LU}, \lambda_{UL} \neq 0$, δηλαδή η πιθανότητα η μεταβλητότητα των τιμών των μετοχών να λάβει τη μικρότερη δυνατή της τιμή ενώ οι αποδόσεις των αντίστοιχων μετοχών λαμβάνουν τις μεγαλύτερες δυνατές είναι μη μηδενική και αντίστροφα. Επομένως, τα αποτελέσματα βρίσκονται σε αντίθεση με τις προβλέψεις του υποδείγματος CAPM (Sharpe, 1964) και σε συμφωνία με τις συμπεριφορικές εξηγήσεις (behavioral postulates) των Hibbert et al. (2008) και Low (2004).

Τέλος, ένα ενδιαφέρον μελλοντικό ζήτημα διερεύνησης της παρούσας εργασίας είναι ο τρόπος σύνδεσης των δύο μεταβλητών S&P 500 και VIX σε διαφορετικές μεταξύ τους χρονικές στιγμές.

ABSTRACT

The relation between the stock returns and their volatility has always been an interesting researching subject in Financial Analysis and especially in for portfolio optimization (Markowitz, 1952), hedging, and option pricing. In the present work we use data from the stock price index (S&P 500) and the implied volatility index (VIX) for the definition of the global and local dependence between the returns and the expectations of the investors for the levels of the volatility risk. The analysis is based on the statistical tool of the copulas and on daily observations counted from 2/1/2004 to 31/8/2017. The empirical results show that: (a) S&P 500 and VIX present opposite global dependence and important local tail dependencies. Consequently, they are in contrast with the predictions of the CAPM model (Sharpe, 1964) and they agree with the behavioral postulates of the Hibbert et.al. (2008) and Low (2004), (b) the copula with the best fit is the 270 rotated BB7.

ΑΝΑΦΟΡΕΣ

- Badshah, I. (2013). Quantile regression analysis of the asymmetric return-volatility relation. *The Journal of Futures Markets*, **33**, 235–265.
- Bai J. and Perron P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, **18**, 1-22.
- Bakaert, G. and Wu, G. (2000). Asymmetric volatility and risk in equity markets. *Review of Financial Studies*, **13**, 1–42.
- Black, F. (1976). Studies of stock volatility changes. Proceeding of the 1976 American Association. *Business and Economics Statistics Section*, 177-81.
- Bollerslev, T., Litvinova, J. and Tauchen, G. (2006). Leverage and volatility feedback effects in high-frequency data. *Journal of Financial Econometrics*, **4**, 353–384.
- Campbell, J. and Hentschel, L. (1992). No news is good news: An asymmetric model of changing volatility in stock returns. *Journal of Financial Economics*, **31**, 281-318.
- Cech, C. (2006). Copula-based top-down approaches in financial risk aggregation. *SSRN Electronic Journal*. doi:10.2139/ssrn.953888.
- Cruz, M., Peters, G. and Shevchenko, P. (2015). *Fundamental aspects of operational risk and insurance analytics. A handbook of operational risk*, John Wiley & Sons.
- Fousekis, P. and Grigoriadis, V. (2019). How well can investors diversify with commodities? Evidence from a flexible copula approach. *Studies in Economics and Finance*, <https://doi.org/10.1108/SEF-05-2018-0138>.
- Hibbert, A., Daigler, R. and B. Dupoyet (2008). A behavioural explanation for the negative asymmetric return-volatility relation. *Journal of Banking and Finance*, **32**, 2254–2266.
- Low, C. (2004). The fear and exuberance from implied volatility of S&P 100 index options. *Journal of Business*, **77**, 527–546.
- Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, **7**, 77-91.
- Nelsen, R. (2006). An introduction to copulas. *The Springer Series in Statistics*.

- Omelka M., Gijbels I. and Veraverbeke, N. (2009). Improved kernel estimation of copulas: Weak convergence and goodness of fit testing. *The Annals of Statistics*, **37**, 3023-3058.
- Schepsmeier, U. (2013) A goodness-of-fit test for regular vine copula models. Preprint <http://arxiv.org/abs/1306.0818>.
- Sharpe, W., (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, **19**, 425–442.
- Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris (in French)*, **8**, 229–231.
- Τζαφέρη Δ. (2018). Γωνιακές πιθανότητες και όρια για στροφές 90 και 270 μοιρών συνήθων συζεύξεων και για μίξεις αυτών. *Διπλωματική Εργασία Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης*.



ΕΠΙΛΟΓΗ ΧΩΡΙΚΩΝ ΟΙΚΟΝΟΜΕΤΡΙΚΩΝ ΥΠΟΔΕΙΓΜΑΤΩΝ ΜΕ ΠΛΗΡΟΦΟΡΙΑΚΑ ΚΡΙΤΗΡΙΑ

Απόστολος Τσιμπάνος¹, Χρήστος Αγιακλόγλου²

¹Τμήμα Στατιστικής και Αναλογιστικών-Χρηματοοικονομικών Μαθηματικών,
Πανεπιστήμιο Αιγαίου
atsibanos@aegean.gr

²Τμήμα Οικονομικής Επιστήμης, Πανεπιστήμιο Πειραιώς
agiaklis@unipi.gr

ΠΕΡΙΛΗΨΗ

Στη χωρική οικονομετρία η εμφάνιση του προβλήματος της χωρικής αυτοσυσχέτισης στα σφάλματα ενός γραμμικού υποδείγματος αντιμετωπίζεται με την επιλογή και εκτίμηση του κατάλληλου χωρικού υποδείγματος. Στην εργασία αυτή, με τη βοήθεια προσομοίωσης, εξετάζεται η δυνατότητα των τριών πιο συχνά χρησιμοποιούμενων πληροφοριακών κριτηρίων, δηλαδή του κριτηρίου AIC του Akaike, του κριτηρίου BIC του Schwarz και του κριτηρίου HQC των Hannan και Quinn να επιλέγουν το σωστό υπόδειγμα. Τα αποτελέσματα δείχνουν ότι η χρησιμοποίηση ενός πληροφοριακού κριτηρίου συνεισφέρει σημαντικά στην επιλογή του σωστού υποδείγματος.

Λέξεις Κλειδιά: Χωρικά Οικονομετρικά Υποδείγματα, Πληροφοριακά Κριτήρια

1. ΕΙΣΑΓΩΓΗ

Η εμφάνιση χωρικής αυτοσυσχέτισης στα σφάλματα είναι ένα πολύ συνηθισμένο πρόβλημα όταν εκτιμώνται υποδείγματα γραμμικής παλινδρόμησης με χωρικά δεδομένα, δηλαδή με δεδομένα που προέρχονται από διαφορετικές γεωγραφικές περιοχές. Χωρική αυτοσυσχέτιση (*spatialautocorrelation*) σημαίνει ότι οι παρατηρήσεις συσχετίζονται σύμφωνα με τις γεωγραφικές τους θέσεις και είναι αποτέλεσμα της χωρικής εξάρτησης (*spatialdependence*) η οποία μαζί με τη χωρική ετερογένεια (*spatialheterogeneity*) αποτελούν τις χωρικές επιδράσεις (*spatialeffects*) που οφείλονται στο είδος των δεδομένων (Anselin, 1988). Η ύπαρξη χωρικής αυτοσυσχέτισης παραβιάζει την υπόθεση του τυχαίου δείγματος καθιστώντας τα αποτελέσματα από την εκτίμηση ενός οικονομετρικού υποδείγματος αναξιόπιστα. Το πρόβλημα αντιμετωπίζεται από τον ερευνητή επιλέγοντας και εκτιμώντας το κατάλληλο χωρικό οικονομετρικό υπόδειγμα που λαμβάνει υπ' όψιν την παρουσία των χωρικών επιδράσεων.

Στις περισσότερες εμπειρικές εργασίες η επιλογή του χωρικού υποδείγματος γίνεται με την εφαρμογή μιας ομάδας ελέγχων που ανήκουν στην κατηγορία των ελέγχων πολλαπλασιαστή Lagrange και που γενικότερα είναι γνωστοί στη βιβλιογραφία ως έλεγχοι χωρικής εξάρτησης (Burridge, 1980, Anselin, 1988b, Anselin, Bera, Florax και Yoon, 1996). Ωστόσο, η μεθοδολογία αυτή, αν και καθιερωμένη, παρουσιάζει δυο σοβαρά μειονέκτημα. Το πρώτο μειονέκτημα είναι ότι αυτοί οι έλεγχοι είναι αρκετά περιοριστικοί ως προς τα εναλλακτικά χωρικά υποδείγματα καθώς εφαρμόζονται αποκλειστικά για την επιλογή μεταξύ του απλού οικονομετρικού υποδείγματος και ενός χωρικού υποδείγματος από το υπόδειγμα χωρικής υστέρησης ή το υπόδειγμα χωρικού σφάλματος. Το δεύτερο μειονέκτημα αυτών των ελέγχων είναι ότι τις περισσότερες φορές που εφαρμόζονται παρουσιάζουν σημαντική στατιστική ισχύ και ως προς τα δύο εναλλακτικά χωρικά υποδείγματα οδηγώντας σε αντιφατικά συμπεράσματα ως προς την επιλογή του σωστού χωρικού οικονομετρικού υποδείγματος. Το πρόβλημα αυτό έχει διερευνηθεί με τη βοήθεια προσομοιώσεων στις εργασίες των Anselin και Florax (1995) και των Τσιμπάνος, Αγιακλόγλου και Τσίμπος (2016) όπου και έχουν προταθεί στρατηγικές επιλογής των υποδειγμάτων.

Οι LeSage και Pace (2009) επισημαίνουν ότι οι έλεγχοι χωρικής εξάρτησης αναπτύχθηκαν και καθιερώθηκαν στη λογική ότι ο υπολογισμός των στατιστικών τους βασίζεται αποκλειστικά στα κατάλοιπα που προκύπτουν από την εκτίμηση του απλού οικονομετρικού υποδείγματος, με τη μέθοδο των ελαχίστων τετραγώνων, χωρίς να απαιτείται η εκτίμηση του αντίστοιχου χωρικού οικονομετρικού υποδείγματος κάτι που τα παλιότερα ήταν δύσκολο και χρονοβόρο διότι απαιτούσε εξειδικευμένο λογισμικό που δεν ήταν διαθέσιμο. Εν τούτοις, τη σημερινή εποχή τα λογισμικά χωρικής ανάλυσης έχουν ευρέως διαδοθεί τόσο σε εμπορικές όσο και σε ανοικτού λογισμικού εκδόσεις και μπορούν εύκολα να εγκατασταθούν στους περισσότερους υπολογιστές. Για το λόγο αυτό, οι LeSage και Pace (2009) θεωρούν ότι η επιλογή ενός χωρικού υποδείγματος θα πρέπει να γίνεται στα πλαίσια της σύγκρισης των πιθανοφανειών των υποδειγμάτων με τους στατιστικούς ελέγχους να περιορίζονται στις χωρικές παραμέτρους τους ενώ η αφετηρία της ανάλυσης θα πρέπει να είναι ένα γενικότερο υπόδειγμα που εμφωλεύει τόσο το υπόδειγμα χωρικής υστέρησης όσο και το υπόδειγμα χωρικού σφάλματος.

Η χρησιμοποίηση πληροφοριακών κριτηρίων αποτελεί συνηθισμένη πρακτική στην επιλογή στατιστικών υποδειγμάτων. Στην εργασία αυτή, με τη βοήθεια προσομοίωσης, διερευνάται η ικανότητα των τριών πιο συχνά χρησιμοποιούμενων πληροφοριακών κριτηρίων, δηλαδή του κριτηρίου AIC του Akaike, του κριτηρίου BIC του Schwarz και του κριτηρίου HQC των Hannan και Quinn να επιλέγουν χωρικά οικονομετρικά υποδείγματα. Τα αποτελέσματα δείχνουν ότι η σύγκριση των τιμών των πληροφοριακών κριτηρίων που προκύπτουν από την εκτίμηση των εναλλακτικών υποδειγμάτων οδηγεί με εξαιρετικά αποτελέσματα στην επιλογή του σωστού υποδείγματος κάτι που φανερώνει ότι οι έλεγχοι χωρικής εξάρτησης ίσως θα μπορούσαν να παρακαμφθούν από τον ερευνητή.

2. ΧΩΡΙΚΑ ΥΠΟΔΕΙΓΜΑΤΑ ΚΑΙ ΠΛΗΡΟΦΟΡΙΑΚΑ ΚΡΙΤΗΡΙΑ

Έστω το απλό οικονομετρικό υπόδειγμα $y = X\beta + \varepsilon$ όπου y είναι το $(n \times 1)$ διάνυσμα των παρατηρήσεων της εξαρτημένης μεταβλητής, X η $(n \times (k+1))$ μήτρα των παρατηρήσεων των k ανεξάρτητων μεταβλητών έχοντας στην πρώτη στήλη μονάδες για την εκτίμηση του σταθερού όρου, β το $((k+1) \times 1)$ διάνυσμα των συντελεστών του υποδείματος και $\varepsilon \sim N(0, \sigma^2 I)$ το $(n \times 1)$ διάνυσμα των τυχαίων σφαλμάτων. Οι χωρικές επιδράσεις ενσωματώνονται στο υπόδειγμα με τη βοήθεια της $(n \times n)$ μήτρα χωρικών σταθμίσεων W που καθορίζει την αλληλεπίδραση μεταξύ των γειτονικών περιοχών. Τα δύο πιο συχνά χρησιμοποιούμενα χωρικά οικονομετρικά υποδείγματα είναι το υπόδειγμα χωρικής υστέρησης (*SpatialLagModel* – *SLM*) που ορίζεται ως $y = \rho Wy + X\beta + \varepsilon$ και το υπόδειγμα χωρικού σφάλματος (*SpatialErrorModel* – *SEM*) που ορίζεται ως $y = X\beta + \varepsilon$ με $\varepsilon = \lambda W\varepsilon + u$ όπου u είναι ένα $(n \times 1)$ διάνυσμα σφαλμάτων. Στα υποδείγματα αυτά Wy και $W\varepsilon$ είναι οι μήτρες χωρικών υστερήσεων που ενσωματώνουν την χωρική εξάρτηση και ουσιαστικά αποτελούνται από τους σταθμισμένους μέσους όρους των τιμών των μεταβλητών στις γειτονικές περιοχές ενώ και λείναι οι συντελεστές χωρικών υστερήσεων της εξαρτημένης μεταβλητής και των σφαλμάτων αντίστοιχα. Τα υποδείγματα αυτά είναι παρόμοια με τα αντίστοιχα που εφαρμόζονται στην ανάλυση χρονοσειρών με τη διαφορά ότι εδώ η εξάρτηση είναι προς πολλές κατευθύνσεις.

Ένα χωρικό οικονομετρικό υπόδειγμα που εμφωλεύει το απλό οικονομετρικό υπόδειγμα αλλά και τα υποδείγματα χωρικής υστέρησης και χωρικού σφάλματος είναι το υπόδειγμα των χωρικών αυτοπαλίνδρομων κατανεμημένων υστερήσεων (*SpatialAutoregressivelyDistributedLagModel* – *SADLM*) που ορίζεται ως $y = \rho Wy + X\beta + WX\theta + \varepsilon$ όπου WX είναι η μήτρα χωρικών υστερήσεων των ανεξάρτητων μεταβλητών (εκτός του σταθερού όρου) και θ το $(k \times 1)$ διάνυσμα των συντελεστών τους (Bivand, 1984 και Lauridsen, 2006). Το υπόδειγμα αυτό τις περισσότερες φορές στη βιβλιογραφία αναφέρεται ως χωρικό υπόδειγμα Durbin (*SpatialDurbinModel* – *SDM*). Ωστόσο, αποτελεί ένα αυτόνομο υπόδειγμα που περιλαμβάνει τις χωρικές υστερήσεις των ανεξάρτητων μεταβλητών ενώ το χωρικό υπόδειγμα Durbin αποτελεί μια ειδική περίπτωση αυτού που προκύπτει από κατάλληλο μετασχηματισμό του υποδείματος χωρικού σφάλματος. Ειδικότερα, το υπόδειγμα χωρικού σφάλματος γράφεται ισοδύναμα ως εξής:

$$y = X\beta + (I - \lambda W)^{-1} u$$

$$(I - \lambda W)y = (I - \lambda W)X\beta + u \Leftrightarrow y = \lambda Wy + X\beta - \lambda WX\beta + u$$

δηλαδή προκύπτει το υπόδειγμα των χωρικών αυτοπαλίνδρομων κατανεμημένων υστερήσεων με τον περιορισμό $-\lambda\beta = \theta$. Ο περιορισμός αυτός θα μπορούσε να ελεγχθεί με τη διεξαγωγή στατιστικού ελέγχου που είναι γνωστός ως έλεγχος κοινού παράγοντα και που παρουσιάζεται αναλυτικά στους Muir και Angulo (2006) για να διερευνηθεί αν πρόκειται στην πραγματικότητα για το υπόδειγμα χωρικού σφάλματος

ή για το γενικότερο υπόδειγμα. Αξίζει να αναφερθεί ότι από τη διαδικασία του μετασχηματισμού και εφόσον η μήτρα χωρικών σταθμίσεων είναι τυποποιημένη ανά γραμμή η χωρική υστέρηση του σταθερού όρου προστίθεται στον σταθερό όρο του υποδείγματος και δεν μπορεί να εκτιμηθεί ξεχωριστά.

Η λογαριθμική συνάρτηση πιθανοφάνειας του υποδείγματος των χωρικών αυτοπαλίνδρομων καταταξιμένων υστερήσεων είναι:

$$\ln L = -(n/2)\ln(2\pi) - (n/2)\ln\sigma^2$$

$$- (1/2\sigma^2)(\mathbf{y} - \rho\mathbf{W}\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{W}\mathbf{X}\boldsymbol{\theta})'(\mathbf{y} - \rho\mathbf{W}\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{W}\mathbf{X}\boldsymbol{\theta}) + \ln|\mathbf{I} - \rho\mathbf{W}|$$

όπου $|\mathbf{I} - \rho\mathbf{W}|$ είναι η Ιακωβιανή ορίζουσα από το μετασχηματισμό του τυχαίου διανύσματος $\boldsymbol{\varepsilon}$ στο διάνυσμα \mathbf{y} η οποία θα πρέπει να είναι θετική ώστε να ορίζονται ο λογάριθμός της και η συνάρτηση πιθανοφάνειας. Όπως έχει δείξει ο Ord (1975) αυτό εξασφαλίζεται όταν ο συντελεστής ρ ορίζεται στο διάστημα $(1/\omega_{\min}, 1/\omega_{\max})$ όπου ω_{\min} και ω_{\max} είναι τα πραγματικά μέρη της ελάχιστης και μέγιστης ιδιοτιμής της μήτρας χωρικών σταθμίσεων. Εφόσον η μήτρα χωρικών σταθμίσεων είναι τυποποιημένη κατά γραμμή μεγαλύτερη ιδιοτιμή της είναι πάντα +1 και επομένως το επάνω όριο του διαστήματος θα είναι +1 ενώ για το κάτω όριο δεν υπάρχει κάποια συγκεκριμένη τιμή που συνήθως είναι μικρότερο του -1. Εάν στη συνάρτηση πιθανοφάνειας τεθούν $\rho = 0$ και $\boldsymbol{\theta} = \mathbf{0}$ προκύπτει η συνάρτηση πιθανοφάνειας του απλού οικονομετρικού υποδείγματος ενώ εάν τεθεί $\boldsymbol{\theta} = \mathbf{0}$ προκύπτει η συνάρτηση πιθανοφάνειας του υποδείγματος χωρικής υστέρησης.

Η απόφαση που πρέπει να ληφθεί είναι εάν με βάση το διαθέσιμο χωρικό δείγμα θα πρέπει να επιλεγεί το απλό οικονομετρικό υπόδειγμα ή κάποιο από τα χωρικά υποδείγματα. Η επιλογή αυτή θα μπορούσε να γίνει χρησιμοποιώντας τους ελέγχους χωρικής εξάρτησης λαμβάνοντας υπ' όψιν τους περιορισμούς και τα προβλήματα που έχουν αναφερθεί ή αφού τα υποδείγματα είναι εμφωλευμένα διεξάγοντας στατιστικούς ελέγχους για τη στατιστική σημαντικότητα των χωρικών παραμέτρων καθώς και λόγου πιθανοφανειών που συγκρίνουν το απλό με τα χωρικά υποδείγματα.

Εναλλακτικά, θα μπορούσαν να εκτιμηθούν με τη μέθοδο της μέγιστης πιθανοφάνειας τα τρία υποδείγματα, δηλαδή το απλό οικονομετρικό υπόδειγμα, το υπόδειγμα χωρικής υστέρησης και το υπόδειγμα των χωρικών αυτοπαλίνδρομων καταταξιμένων υστερήσεων, να υπολογιστεί για κάθε ένα υπόδειγμα η τιμή κάποιου πληροφοριακού κριτηρίου και να επιλεγεί εκείνο το υπόδειγμα για το οποίο η τιμή του κριτηρίου είναι ελάχιστη. Η φιλοσοφία των πληροφοριακών κριτηρίων βασίζεται στην ποσοτικοποίηση της καλής προσαρμογής του υποδείγματος στις διαθέσιμες παρατηρήσεις περιλαμβάνοντας κάποιο τύπο ποινής για τον αριθμό των παραμέτρων του και το μέγεθος του δείγματος. Τα τρία κριτήρια που χρησιμοποιούνται τις περισσότερες φορές σε εμπειρικές εφαρμογές εκτίμησης στατιστικών υποδειγμάτων είναι το πληροφοριακό κριτήριο AIC (*Akaike Information Criterion*) του Akaike (1973) που ορίζεται ως $AIC = -2\ln \hat{L} + 2p$, το BIC (*Bayesian Information Criterion*) του Schwarz (1978) που ορίζεται ως $BIC = -2\ln \hat{L} + p \ln n$ και το πληροφοριακό κριτήριο

HQC(*HannanQuinnInformationCriterion*) των Hannan και Quinn (1978) που ορίζεται ως $HQC = -2\ln \hat{L} + 2p \ln \ln n$. Στα κριτήρια αυτά, $\ln \hat{L}$ είναι η μέγιστη τιμή της λογαριθμικής συνάρτησης πιθανοφάνειας, p ο αριθμός των παραμέτρων κάθε υποδείγματος και n το μέγεθος του δείγματος με το οποίο έγινε η εκτίμηση. Επομένως, ο ερευνητής διαλέγει το πληροφοριακό κριτήριο που θα χρησιμοποιήσει, υπολογίζει τις τιμές του για όλα τα υποψήφια υποδείγματα και επιλέγει ως καταλληλότερο αυτό για το οποίο η τιμή του κριτηρίου είναι η ελάχιστη.

3. ΠΡΟΣΟΜΟΙΩΣΗ

Στην ενότητα αυτή εξετάζεται με τη βοήθεια προσομοίωσης η ικανότητα των τριών πληροφοριακών κριτηρίων να διακρίνουν το σωστό υπόδειγμα μεταξύ του απλού οικονομετρικού υποδείγματος, του υποδείγματος χωρικής υστέρησης και του υποδείγματος των χωρικών αυτοπαλίνδρομων κατανεμημένων υστερήσεων. Για την πραγματοποίηση της προσομοίωσης θεωρήθηκε ότι η μήτρα \mathbf{X} είναι διαστάσεων $(n \times 2)$ και αποτελείται από μια ανεξάρτητη μεταβλητή και μία στήλη με μονάδες για την εκτίμηση του σταθερού όρου ενώ στο (2×1) διάνυσμα $\boldsymbol{\beta}$ οι τιμές των συντελεστών τέθηκαν και οι δυο ίσοι με τη μονάδα. Η ανεξάρτητη μεταβλητή προέρχεται από ομοιόμορφη $U(0, 10)$ κατανομή και προσομοιώθηκε μόνο μία φορά παραμένοντας σταθερή σε όλες τις επαναλήψεις. Το διάνυσμα των τυχαίων σφαλμάτων $\boldsymbol{\varepsilon}$ προέρχεται από $N(0, \mathbf{I})$ κατανομή και προστίθεται στο διάνυσμα $\mathbf{X}\boldsymbol{\beta}$ για να παραχθεί το διάνυσμα της εξαρτημένης μεταβλητής \mathbf{y} στο απλό οικονομετρικό υπόδειγμα.

Η χωρική εξάρτηση ορίστηκε με μία $(n \times n)$ τυποποιημένη κατά γραμμή μήτρα χωρικών σταθμίσεων \mathbf{W} που κατασκευάστηκε με βάση τα κριτήρια του πύργου (*rookcontiguity*) και τη βασίλισσας (*queencontiguity*) για ένα ομαλό τετραγωνικό πλέγμα διαστάσεων 10×10 και 20×20 παρέχοντας δείγματα 100 και 400 παρατηρήσεων. Από τον υπολογισμό των ιδιοτιμών για τις τέσσερις μήτρες χωρικών σταθμίσεων προκύπτει ότι το κάτω όριο του χωρικού συντελεστή για το κριτήριο του πύργου είναι -1 τόσο στο δείγμα των 100 όσο και των 400 παρατηρήσεων ενώ για το κριτήριο της βασίλισσας -1,97 και -1,921 αντίστοιχα.

Το υπόδειγμα χωρικής υστέρησης παράγεται πολλαπλασιάζοντας τη δεξιά πλευρά του απλού οικονομετρικού υποδείγματος με το χωρικό πολλαπλασιαστή $(\mathbf{I} - \rho \mathbf{W})^{-1}$ οπότε προκύπτει $\mathbf{y} = (\mathbf{I} - \rho \mathbf{W})^{-1} \mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \rho \mathbf{W})^{-1} \boldsymbol{\varepsilon}$ ενώ το υπόδειγμα χωρικού σφάλματος παράγεται από τη σχέση $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \lambda \mathbf{W})^{-1} \mathbf{u}$ όπου το $(n \times 1)$ διάνυσμα \mathbf{u} προέρχεται από $N(0, \mathbf{I})$ κατανομή. Για τις τιμές των χωρικών συντελεστών ρ και λ θεωρήθηκαν ότι λαμβάνουν τις τιμές 0,1, 0,2, 0,3, 0,4, 0,5, 0,6, 0,7, 0,8 και 0,9 που βρίσκονται εντός των εφικτών ορίων. Στο υπόδειγμα των χωρικών αυτοπαλίνδρομων κατανεμημένων υστερήσεων για το συντελεστή ρ θεωρήθηκαν οι τιμές 0,2, 0,5 και 0,9 και για το συντελεστή θ της χωρικής υστέρησης της ανεξάρτητης μεταβλητή οι τιμές 0,1, 0,2, 0,3, 0,4, 0,5, 0,6, 0,7, 0,8 και 0,9. Πρέπει να αναφερθεί ότι για το συντελεστή θ δεν υπάρχει κάποιος περιορισμός ως προς τις τιμές που μπορεί να λάβει αλλά θεωρήθηκαν οι τιμές αυτές για λόγους

σύγκρισης με τα ανάλογα υποδείγματα της ανάλυσης χρονοσειρών. Η παραγωγή του υποδείματος γίνεται από τη σχέση $y = (I - \rho W)^{-1} (X\beta + WX\theta + \varepsilon)$.

Η διαδικασία της προσομοίωσης πραγματοποιήθηκε στο στατιστικό πρόγραμμα R με τη βοήθεια του πακέτου SPDEP (Bivand, 2015). Εκτελέστηκαν 1.000 επαναλήψεις και για κάθε προσομοιωμένο υπόδειγμα εκτιμήθηκαν το απλό οικονομετρικό υπόδειγμα, το υπόδειγμα χωρικής υστέρησης και το υπόδειγμα των χωρικών αυτοπαλίνδρομων κατανεμημένων υστερήσεων, υπολογίστηκαν οι τιμές των κριτηρίων AIC, BIC και HQC και καταγράφησαν τα ποσοστά επιλογής κάθε υποδείματος. Επισημαίνεται ότι το υπόδειγμα χωρικού σφάλματος δεν εκτιμάται άμεσα αλλά έμμεσα μέσω του χωρικού υποδείματος Durbin. Κατά συνέπεια, σε αυτή την περίπτωση τα πληροφοριακά κριτήρια θα πρέπει να επιλέγουν το υπόδειγμα των χωρικών αυτοπαλίνδρομων κατανεμημένων υστερήσεων του οποίου ειδική περίπτωση αποτελεί το υπόδειγμα χωρικού σφάλματος.

Πίνακας 1. Ποσοστά επιλογής των τριών υποδειμάτων για όταν ισχύει το απλό γραμμικό υπόδειγμα

Σταθμίσεις	Πύργος		Βασίλισσα		
	<i>N</i>	100	400	100	400
AIC-Απλό		75,7	79,4	76,4	78,5
AIC-SLM		15,3	12,0	14,5	12,9
AIC-SADLM		9,0	8,6	9,1	8,6
BIC- Απλό		94,9	98,1	95,5	97,9
BIC- SLM		4,5	1,6	3,6	1,7
BIC- SADLM		0,6	0,3	0,9	0,4
HQC- Απλό		89,0	93,1	89,0	93,1
HQC- SLM		8,2	5,2	8,2	5,2
HQC- SADLM		2,8	1,7	2,8	1,7

Τα ποσοστά που τα πληροφοριακά κριτήρια επιλέγουν τα τρία υποψήφια υποδείγματα παρουσιάζονται στον Πίνακα 1 για την περίπτωση απουσίας χωρικών επιδράσεων, δηλαδή για όταν ισχύει το απλό οικονομετρικό υπόδειγμα. Από τον πίνακα αυτόν διαπιστώνεται ότι το κριτήριο BIC παρουσιάζει εξαιρετική επίδοση στον εντοπισμό του απλού υποδείματος με ποσοστά επιλογής κοντά στο 95% στα δείγματα των 100 παρατηρήσεων και που πλησιάζουν στο 98% στα δείγματα των 400 παρατηρήσεων. Εσφαλμένα επιλέγεται κάποιο από τα δύο χωρικά υποδείγματα με ποσοστό περίπου 5%. Καθώς οι έλεγχοι χωρικής εξάρτησης έχουν ως μηδενική υπόθεση το απλό οικονομετρικό υπόδειγμα το αποτέλεσμα αυτό θα μπορούσε να συγκριθεί με το μέγεθος από την εφαρμογή τους όπως έχει παρουσιαστεί σε αντίστοιχες προσομοιώσεις στις εργασίες των Anselin και Florax (1995) και των Τσιμπάνος, Αγιακλόγλου και Τσίμπος (2016). Κατά συνέπεια, η εφαρμογή του κριτηρίου BIC ουσιαστικά οδηγεί σε παρόμοια συμπεράσματα με τους ελέγχους χωρικής εξάρτησης κάτι που σημαίνει ότι ίσως θα μπορούσαν να παρακαμφθούν από την ανάλυση. Το κριτήριο που ακολουθεί με ποσοστά σωστής επιλογής λίγο μικρότερα από αυτά του BIC είναι το κριτήριο HQC. Αντίθετα, τα χαμηλότερα ποσοστά σωστής επιλογής διαπιστώνονται για το κριτήριο AIC κάτι αναμενόμενο

καθώς είναι γνωστό ότι το κριτήριο αυτό έχει την τάση να επιλέγει υποδείγματα με μεγάλο αριθμό παραμέτρων. Τα κριτήρια ορισμού της μήτρας χωρικών σταθμίσεων αλλά και το μέγεθος του δείγματος δεν επηρεάζουν ιδιαίτερα τα αποτελέσματα.

Πίνακας 2. Ποσοστά επιλογής των τριών υποδειγμάτων για όταν ισχύει το υπόδειγμα χωρικής υστέρησης SLM για διάφορες τιμές του ρ

N	ρ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
100	AIC-Απλό	36,8	2,9	0	0	0	0	0	0	0
	AIC-SLM	49,9	80,1	82	81,4	81,4	81,7	81,2	81,1	81,6
	AIC-SADLM	13,3	17	18	18,6	18,6	18,3	18,8	18,9	18,4
	BIC- Απλό	65,9	11	0,4	0	0	0	0	0	0
	BIC- SLM	32,7	85,7	96,2	96	95,7	95,9	96,2	96,1	95,5
	BIC- SADLM	1,4	3,3	3,4	4	4,3	4,1	3,8	3,9	4,5
	HQC- Απλό	49,4	5,5	0,1	0	0	0	0	0	0
	HQC- SLM	44,2	85,6	90,7	91	90,7	90,3	90,6	90,2	90,4
	HQC- SADLM	6,4	8,9	9,2	9	9,3	9,7	9,4	9,8	9,6
400	AIC-Απλό	3,4	0	0	0	0	0	0	0	0
	AIC-SLM	80,3	83	82	82	81,9	81,5	82,1	83	84
	AIC-SADLM	16,3	17	18	18	18,1	18,5	17,9	17	16
	BIC- Απλό	21,8	0	0	0	0	0	0	0	0
	BIC- SLM	77,1	98,9	98,8	98,6	98,4	98,3	98,6	98,4	98,2
	BIC- SADLM	1,1	1,1	1,2	1,4	1,6	1,7	1,4	1,6	1,8
	HQC- Απλό	8,6	0	0	0	0	0	0	0	0
	HQC- SLM	85,1	93,3	93	93,8	93,9	93,2	93,5	93,1	93,2
	HQC- SADLM	6,3	6,7	7	6,2	6,1	6,8	6,5	6,9	6,8
100 Βασίλεισα	AIC-Απλό	57,3	18,4	1,5	0	0	0	0	0	0
	AIC-SLM	31	64,7	80,1	81,5	82	82	82,5	83	81,8
	AIC-SADLM	11,7	16,9	18,4	18,5	18	18	17,5	17	18,2
	BIC- Απλό	83,1	42	7,2	0,3	0	0	0	0	0
	BIC- SLM	15,2	54,8	88,6	95,2	96	96	96	96	96,2
	BIC- SADLM	1,7	3,2	4,2	4,5	4,3	3,9	4	4,2	3,8
	HQC- Απλό	70,6	28,9	3,5	0	0	0	0	0	0
	HQC- SLM	24,9	64	87,5	90,6	91	91	90,8	90	90,5
	HQC- SADLM	4,5	7,1	9	9,4	9,5	9,4	9,2	9,6	9,5
400 Βασίλεισα	AIC-Απλό	18,9	0	0	0	0	0	0	0	0
	AIC-SLM	65,8	83	83,3	83,2	84	83	84,4	85	84,9
	AIC-SADLM	15,3	17	16,7	16,8	16	17	15,6	16	15,1
	BIC- Απλό	55,6	1,6	0	0	0	0	0	0	0
	BIC- SLM	43,6	97,1	98,6	98,6	99	99	98,7	99	98,4
	BIC- SADLM	0,8	1,3	1,4	1,4	1,2	1,3	1,3	1,2	1,6
	HQC- Απλό	33	0,7	0	0	0	0	0	0	0
	HQC- SLM	62,4	93,5	94	94,5	94	93	93,3	94	93,5
	HQC- SADLM	4,6	5,8	6	5,5	5,9	6,7	6,7	6,3	6,5

Ο Πίνακας 2 συνοψίζει τα αποτελέσματα των προσομοιώσεων όταν στην πραγματικότητα ισχύει το υπόδειγμα χωρικής υστέρησης. Όπως προκύπτει από αυτόν τον πίνακα το κριτήριο BIC υπερτερεί στα ποσοστά σωστής επιλογής ανεξάρτητα του μεγέθους του δείγματος και της μήτρας χωρικών σταθμίσεων. Πιο συγκεκριμένα, στα δείγματα των 100 παρατηρήσεων το κριτήριο BIC επιλέγει το υπόδειγμα χωρικής υστέρησης με ποσοστά κοντά στο 96% ενώ στα δείγματα των 400 παρατηρήσεων με

ποσοστά που είναι μεγαλύτερα του 98%. Τα κριτήρια που ακολουθούν σε ποσοστά σωστής επιλογής είναι το HQC και το AIC. Γενικά, για το κριτήριο AIC τα ποσοστά σωστής επιλογής δεν ξεπερνάνε σε καμία περίπτωση το 85%. Αυτό συμβαίνει διότι

Πίνακας 3. Ποσοστά επιλογής των τριών υποδειγμάτων για όταν ισχύει το υπόδειγμα χωρικού σφάλματος SEM για διάφορες τιμές του λ

N	λ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	
100	Πύσος	AIC-Απλό	69,5	51,1	30,8	11,9	3	0,3	0,1	0	0
		AIC-SLM	14,4	13,4	10,4	5,9	2,4	0,3	0	0	0
		AIC-SADLM	16,1	35,5	58,8	82,2	94,6	99,4	99,9	100	100
		BIC- Απλό	93,8	86,1	66,8	41,4	17,6	4,1	0,4	0,1	0
		BIC- SLM	4,9	6,2	7,4	6,5	4,2	2	0,3	0	0
		BIC- SADLM	1,3	7,7	25,8	52,1	78,2	93,9	99,3	99,9	100
		HQC- Απλό	84,3	70,2	46,7	22,8	7,5	1	0,2	0	0
		HQC- SLM	9,6	9,4	9	6,8	3,5	1,2	0,2	0	0
		HQC- SADLM	6,1	20,4	44,3	70,4	89	97,8	99,6	100	100
400	Πύσος	AIC-Απλό	56,1	12,6	0,4	0,1	0	0	0	0	0
		AIC-SLM	13,2	7,3	0,4	0	0	0	0	0	0
		AIC-SADLM	30,7	80,1	99,2	99,9	100	100	100	100	100
		BIC- Απλό	93,8	63,4	13,7	0,3	0,1	0	0	0	0
		BIC- SLM	3,4	6,1	3,2	0,2	0	0	0	0	0
		BIC- SADLM	2,8	30,5	83,1	99,5	99,9	100	100	100	100
		HQC- Απλό	80,1	30,4	2,9	0,1	0	0	0	0	0
		HQC- SLM	8,3	7,7	1,6	0,1	0	0	0	0	0
		HQC- SADLM	11,6	61,9	95,5	99,8	100	100	100	100	100
100	Βασίλισσα	AIC-Απλό	74,4	64,7	46,4	27,1	12	3,2	0,7	0,1	0
		AIC-SLM	14,9	15,1	16	15,1	9,7	5,3	2,2	0,5	0
		AIC-SADLM	10,7	20,2	37,6	57,8	79	92	97,1	99	100
		BIC- Απλό	94,2	90,1	79	61,6	37	16	3,8	0,5	0
		BIC- SLM	5	6,9	9,6	11,1	12	9,8	5,1	2,1	0,1
		BIC- SADLM	0,8	3	11,4	27,3	51	74	91,1	97	99,9
		HQC- Απλό	85,9	77,4	63,9	42,7	21	6,9	1,5	0,2	0
		HQC- SLM	10,2	11,6	13	13,1	12	7	3,3	0,9	0
		HQC- SADLM	3,9	11	23,1	44,2	67	86	95,2	99	100
400	Βασίλισσα	AIC-Απλό	66,2	30,5	5,4	0,3	0	0	0	0	0
		AIC-SLM	15,2	13,1	6	0,3	0	0	0	0	0
		AIC-SADLM	18,6	56,4	88,6	99,4	100	100	100	100	100
		BIC- Απλό	95,9	80,2	43,1	8	0,5	0	0	0	0
		BIC- SLM	2,7	8,1	8,3	3,9	0,2	0	0	0	0
		BIC- SADLM	1,4	11,7	48,6	88,1	99	100	100	100	100
		HQC- Απλό	85,1	57	18,1	1,7	0,1	0	0	0	0
		HQC- SLM	8,6	10,9	8	1,7	0	0	0	0	0
		HQC- SADLM	6,3	32,1	73,9	96,6	100	100	100	100	100

για όλες τις τιμές του ρ καταγράφονται αξιόλογα ποσοστά ασφαλισμένης επιλογής του υποδείγματος των χωρικών αυτοπαλίνδρομων κατανεμημένων υστερήσεων κάτι και πάλι αναμενόμενο σε κάποιο βαθμό αφού το κριτήριο αυτό έχει τάση να επιλέγει υποδείγματα με μεγάλο αριθμό παραμέτρων. Αξιοσημείωτο είναι το γεγονός ότι το απλό υπόδειγμα δεν επιλέγεται σχεδόν ποτέ από κανένα κριτήριο ακόμη και για πολύ μικρές τιμές του ρ , δηλαδή η επιλογή γίνεται μόνο μεταξύ του υποδείγματος χωρικής

υστέρησης και του υποδείγματος των χωρικών αυτοπαλίνδρομων κατανεμημένων υστερήσεων. Αυτό σημαίνει ότι τα κριτήρια έχουν την ικανότητα να εντοπίζουν την παρουσία των χωρικών επιδράσεων και οδηγούν σε κάποιο από τα δυο χωρικά

Πίνακας 4. Ποσοστά επιλογής των τριών υποδειγμάτων για όταν ισχύει το υπόδειγμα κατανεμημένων χωρικών υστερήσεων SADLM για $\rho=0,2$ και διάφορες τιμές του θ

N	$\rho = 0,2$	θ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	
100	Πύσος	AIC-Απλό	0,1	0	0	0	0	0	0	0	0	0
		AIC-SLM	68,1	41,2	15,8	2,7	0,1	0	0	0	0	0
		AIC-SADLM	31,8	58,8	84,2	97,3	99,9	100	100	100	100	100
		BIC- Απλό	0,8	0	0	0	0	0	0	0	0	0
		BIC- SLM	88,3	67,1	38,1	12,6	1,3	0,1	0	0	0	0
		BIC- SADLM	10,9	32,9	61,9	87,4	98,7	99,9	100	100	100	100
		HQC- Απλό	0,3	0	0	0	0	0	0	0	0	0
		HQC- SLM	79	53,1	24,8	6	0,7	0,1	0	0	0	0
		HQC- SADLM	20,7	46,9	75,2	94	99,3	99,9	100	100	100	100
400	Πύσος	AIC-Απλό	0	0	0	0	0	0	0	0	0	
		AIC-SLM	48	6,3	0,1	0	0	0	0	0	0	0
		AIC-SADLM	52	93,7	99,9	100	100	100	100	100	100	
		BIC- Απλό	0	0	0	0	0	0	0	0	0	0
		BIC- SLM	82,8	30,8	1,8	0	0	0	0	0	0	0
		BIC- SADLM	17,2	69,2	98,2	100	100	100	100	100	100	
		HQC- Απλό	0	0	0	0	0	0	0	0	0	0
		HQC- SLM	65,9	14,8	0,3	0	0	0	0	0	0	0
		HQC- SADLM	34,1	85,2	99,7	100	100	100	100	100	100	100
100	Βασίλεισα	AIC-Απλό	3	0,2	0	0	0	0	0	0	0	
		AIC-SLM	70,5	57,1	39,5	24,2	11	3,8	0,8	0	0	
		AIC-SADLM	26,5	42,7	60,5	75,8	89	96	99,2	100	100	
		BIC- Απλό	13,4	1,7	0	0	0	0	0	0	0	
		BIC- SLM	79,5	81,3	68,1	50,5	31	16	5,7	1,7	0,4	
		BIC- SADLM	7,1	17	31,9	49,5	69	84	94,3	98	99,6	
		HQC- Απλό	5,8	0,3	0	0	0	0	0	0	0	
		HQC- SLM	78,4	69,7	53,1	34,1	19	7,4	2,1	0,5	0	
		HQC- SADLM	15,8	30	46,9	65,9	81	93	97,9	100	100	
400	Βασίλεισα	AIC-Απλό	0	0	0	0	0	0	0	0	0	
		AIC-SLM	61,9	25,8	4,9	0,8	0,1	0	0	0	0	
		AIC-SADLM	38,1	74,2	95,1	99,2	100	100	100	100	100	
		BIC- Απλό	0	0	0	0	0	0	0	0	0	
		BIC- SLM	91,1	67,1	28,8	5,1	0,8	0,1	0	0	0	
		BIC- SADLM	8,9	32,9	71,2	94,9	99	100	100	100	100	
		HQC- Απλό	0	0	0	0	0	0	0	0	0	
		HQC- SLM	79,3	44,5	11,5	1,7	0,2	0,1	0	0	0	
		HQC- SADLM	20,7	55,5	88,5	98,3	100	100	100	100	100	

οικονομετρικά υποδείγματα που περιλαμβάνουν τη χωρική υστέρηση της εξαρτημένης μεταβλητής.

Στον Πίνακα 3 περιλαμβάνονται τα ποσοστά επιλογής των τριών εναλλακτικών υποδειγμάτων από τα πληροφοριακά κριτήρια όταν ισχύει το υπόδειγμα χωρικού σφάλματος. Επαναλαμβάνεται ότι σε αυτή την περίπτωση η επιτυχία των κριτηρίων

έγκειται στην επιλογή της ισοδύναμης μορφής του, δηλαδή του υποδείγματος των χωρικών αυτοπαλίνδρομων κατανεμημένων υστερήσεων καθώς στην πραγματικότητα πρόκειται για το χωρικό υπόδειγμα Durbin. Το βασικό συμπέρασμα από τη μελέτη του πίνακα είναι ότι για μέτριες και μεγάλες τιμές του λ και δείγματα

Πίνακας 5. Ποσοστά επιλογής των τριών υποδειγμάτων για όταν ισχύει το υπόδειγμα κατανεμημένων χωρικών υστερήσεων SADLM για $\rho=0,5$ και διάφορες τιμές του θ

N	$\rho = 0,5$	θ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	
100	Πύσνος	AIC-Απλό	0	0	0	0	0	0	0	0	0	0
		AIC-SLM	61,8	29,3	6,9	0,6	0	0	0	0	0	0
		AIC-SADLM	38,2	70,7	93,1	99,4	100	100	100	100	100	100
		BIC- Απλό	0	0	0	0	0	0	0	0	0	0
		BIC- SLM	85,4	55,4	21,9	3,3	0,3	0	0	0	0	0
		BIC- SADLM	14,6	44,6	78,1	96,7	99,7	100	100	100	100	100
		HQC- Απλό	0	0	0	0	0	0	0	0	0	0
		HQC- SLM	73,2	41,3	12,9	1,1	0,1	0	0	0	0	0
		HQC- SADLM	26,8	58,7	87,1	98,9	99,9	100	100	100	100	100
400	Πύσνος	AIC-Απλό	0	0	0	0	0	0	0	0	0	
		AIC-SLM	37,7	2	0	0	0	0	0	0	0	0
		AIC-SADLM	62,3	98	100	100	100	100	100	100	100	
		BIC- Απλό	0	0	0	0	0	0	0	0	0	0
		BIC- SLM	75,9	16,8	0,2	0	0	0	0	0	0	0
		BIC- SADLM	24,1	83,2	99,8	100	100	100	100	100	100	100
		HQC- Απλό	0	0	0	0	0	0	0	0	0	0
		HQC- SLM	57,1	5,7	0	0	0	0	0	0	0	0
		HQC- SADLM	42,9	94,3	100	100	100	100	100	100	100	100
100	Βασίλισσα	AIC-Απλό	0	0	0	0	0	0	0	0	0	
		AIC-SLM	70,4	53,4	31,9	15,2	5,1	1	0,2	0	0	
		AIC-SADLM	29,6	46,6	68,1	84,8	95	99	99,8	100	100	
		BIC- Απλό	0	0	0	0	0	0	0	0	0	
		BIC- SLM	91,1	79,5	61,3	39,9	20	7,4	1,9	0,4	0	
		BIC- SADLM	8,9	20,5	38,7	60,1	80	93	98,1	100	100	
		HQC- Απλό	0	0	0	0	0	0	0	0	0	
		HQC- SLM	82	66	45,9	25,1	9,8	3,2	0,6	0	0	
		HQC- SADLM	18	34	54,1	74,9	90	97	99,4	100	100	
400	Βασίλισσα	AIC-Απλό	0	0	0	0	0	0	0	0	0	
		AIC-SLM	57,3	17	2	0,2	0,1	0	0	0	0	
		AIC-SADLM	42,7	83	98	99,8	100	100	100	100	100	
		BIC- Απλό	0	0	0	0	0	0	0	0	0	
		BIC- SLM	88,9	56,9	14,9	1,5	0,2	0	0	0	0	
		BIC- SADLM	11,1	43,1	85,1	98,5	100	100	100	100	100	
		HQC- Απλό	0	0	0	0	0	0	0	0	0	
		HQC- SLM	74,3	34,5	5,1	0,6	0,1	0	0	0	0	
		HQC- SADLM	25,7	65,5	94,9	99,4	100	100	100	100	100	

400 παρατηρήσεων και τα τρία κριτήρια επιλέγουν το σωστό υπόδειγμα με ποσοστά σχεδόν 100% ανεξάρτητα της μήτρας χωρικών σταθμίσεων. Για πολύ μικρές τιμές του λ το κριτήριο BIC επιλέγει εσφαλμένα το απλό υπόδειγμα και το κριτήριο AIC έχει καλύτερα ποσοστά επιλογής. Ωστόσο, πρέπει να αναφερθεί ότι στην περίπτωση του υποδείγματος χωρικού σφάλματος οι χωρικές επιδράσεις δεν είναι

ουσιαστικές αλλά εμφανίζονται ως ενόχληση (*nuisance*) κάτι που συνεπάγεται ότι η επιλογή του απλού υποδείγματος δεν αποτελεί σοβαρό πρόβλημα καθώς οι εκτιμητές του παραμένουν αμερόληπτοι. Αντίθετα, στα δείγματα των 100 παρατηρήσεων καλύτερα

Πίνακας 6. Ποσοστά επιλογήστων τριών υποδειγμάτων για όταν ισχύει το υπόδειγμα κατανομής χωρικών υστερήσεων SADLM για $\rho=0,9$ και διάφορες τιμές του θ

N	$\rho = 0,9$	θ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
100	Πύσος	AIC-Απλό	0	0	0	0	0	0	0	0	0
		AIC-SLM	44,6	8,9	0,4	0	0	0	0	0	0
		AIC-SADLM	55,4	91,1	99,6	100	100	100	100	100	100
		BIC- Απλό	0	0	0	0	0	0	0	0	0
		BIC- SLM	74,2	26,7	3	0	0	0	0	0	0
		BIC- SADLM	25,8	73,3	97	100	100	100	100	100	100
		HQC- Απλό	0	0	0	0	0	0	0	0	0
		HQC- SLM	56,7	16,8	1,2	0	0	0	0	0	0
		HQC- SADLM	43,3	83,2	98,8	100	100	100	100	100	100
400	Πύσος	AIC-Απλό	0	0	0	0	0	0	0	0	0
		AIC-SLM	14,5	0	0	0	0	0	0	0	0
		AIC-SADLM	85,5	100	100	100	100	100	100	100	100
		BIC- Απλό	0	0	0	0	0	0	0	0	0
		BIC- SLM	50,6	0,7	0	0	0	0	0	0	0
		BIC- SADLM	49,4	99,3	100	100	100	100	100	100	100
		HQC- Απλό	0	0	0	0	0	0	0	0	0
		HQC- SLM	30	0,1	0	0	0	0	0	0	0
		HQC- SADLM	70	99,9	100	100	100	100	100	100	100
100	Βασίλεισσα	AIC-Απλό	0	0	0	0	0	0	0	0	0
		AIC-SLM	63,4	33,7	10	2,7	0,2	0	0	0	0
		AIC-SADLM	36,6	66,3	90	97,3	100	100	100	100	100
		BIC- Απλό	0	0	0	0	0	0	0	0	0
		BIC- SLM	86,3	64	33,7	10	2,6	0,1	0	0	0
		BIC- SADLM	13,7	36	66,3	90	97	100	100	100	100
		HQC- Απλό	0	0	0	0	0	0	0	0	0
		HQC- SLM	74,3	48,1	18,2	4,5	0,7	0	0	0	0
		HQC- SADLM	25,7	51,9	81,8	95,5	99	100	100	100	100
400	Βασίλεισσα	AIC-Απλό	0	0	0	0	0	0	0	0	0
		AIC-SLM	36,6	2,6	0,1	0	0	0	0	0	0
		AIC-SADLM	63,4	97,4	99,9	100	100	100	100	100	100
		BIC- Απλό	0	0	0	0	0	0	0	0	0
		BIC- SLM	77,8	20,7	1	0,1	0	0	0	0	0
		BIC- SADLM	22,2	79,3	99	99,9	100	100	100	100	100
		HQC- Απλό	0	0	0	0	0	0	0	0	0
		HQC- SLM	56,7	7,2	0,2	0	0	0	0	0	0
		HQC- SADLM	43,3	92,8	99,8	100	100	100	100	100	100

ποσοστά σωστής επιλογής επιτυγχάνει το κριτήριο AIC και ακολουθούν τα κριτήρια HQC και BIC.

Οι Πίνακες 4, 5 και 6 περιέχουν τα αποτελέσματα των προσομοιώσεων για τις περιπτώσεις που ισχύει το υπόδειγμα των χωρικών αυτοπαλίνδρομων κατανομών υστερήσεων για τιμές του ρ ίσες με 0,2, 0,5 και 0,9 και για διάφορες

τιμές θ . Από αυτούς τους πίνακες είναι ξεκάθαρο ότι και τα τρία κριτήρια παρουσιάζουν εξαιρετικές επιδόσεις στην επιλογή του σωστού υποδείγματος. Ειδικότερα, η βασική διαπίστωση είναι ότι το απλό υπόδειγμα δεν επιλέγεται ποτέ και η επιλογή γίνεται μόνο μεταξύ του υποδείγματος χωρικής υστέρησης και του υποδείγματος των χωρικών αυτοπαλίνδρομων κατανεμημένων υστερήσεων κάτι που είχε εντοπιστεί και στην περίπτωση του υποδείγματος χωρικής υστέρησης. Και τα τρία κριτήρια για μέτριες και μεγάλες τιμές του θ επιλέγουν το σωστό υπόδειγμα με ποσοστά επιτυχίας 100% ανεξάρτητα από τη μήτρα χωρικών σταθμίσεων και το μέγεθος του δείγματος. Η μήτρα χωρικών σταθμίσεων και το μέγεθος του δείγματος φαίνονται ότι επηρεάζουν μόνο για πολύ μικρές τιμές του θ . Σε αυτές τις περιπτώσεις παρατηρούνται υψηλότερα ποσοστά σωστών επιλογών όταν η μήτρα χωρικών σταθμίσεων έχει οριστεί με το κριτήριο του πύργου σε σχέση με το κριτήριο της βασιλίσσας ενώ και το μεγαλύτερο δείγμα βοηθάει στην επιλογή του σωστού υποδείγματος. Το κριτήριο AIC σημειώνει τις περισσότερες επιτυχίες και ακολουθούν τα κριτήρια HQC και BIC. Η τιμή του ρ επηρεάζει τα αποτελέσματα μόνο για πολύ μικρές τιμές του θ . Συγκεκριμένα, η μεγαλύτερη τιμή του ρ βελτιώνει τα ποσοστά σωστών επιλογών. Το τελικό συμπέρασμα είναι ότι μεγάλες τιμές του ρ ταυτόχρονα με μεγάλες τιμές του θ και μεγάλος αριθμός παρατηρήσεων οδηγούν και τα τρία κριτήρια να επιλέξουν με ποσοστό 100% το υπόδειγμα των χωρικών αυτοπαλίνδρομων κατανεμημένων υστερήσεων που είναι και το σωστό υπόδειγμα.

4. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΣΧΟΛΙΑ

Η ταυτοποίηση του υποδείγματος που εκφράζει με τον πιο αντιπροσωπευτικό τρόπο τη χωρική εξάρτηση στις παρατηρήσεις αποτελεί ζήτημα πρωταρχικής σημασίας για τον ερευνητή σε κάθε χωρική οικονομετρική ανάλυση. Στις εμπειρικές εφαρμογές τα υποδείγματα που συνήθως παρουσιάζονται είναι το υπόδειγμα χωρικής υστέρησης και το υπόδειγμα χωρικού σφάλματος κάτι που οφείλεται σε μεγάλο βαθμό στα αποτελέσματα της εφαρμογής των διαδοσμένων ελέγχων χωρικής εξάρτησης ενώ αγνοούνται άλλα γενικότερα υποδείγματα που θα μπορούσαν να εκφράζουν πληρέστερα τη μορφή της χωρικής εξάρτησης. Στην εργασία αυτή εξετάστηκε με τη βοήθεια προσομοίωσης η συμπεριφορά των τριών πιο συχνών πληροφοριακών κριτηρίων, δηλαδή των AIC, BIC και HQC στο να επιλέγουν χωρικά οικονομετρικά υποδείγματα. Επομένως, παρακάμπτεται η διαδικασία των ελέγχων χωρικής εξάρτησης και συγκρίνονται αποκλειστικά οι τιμές των πληροφοριακών κριτηρίων για το απλό οικονομετρικό υπόδειγμα, το υπόδειγμα χωρικής υστέρησης και το υπόδειγμα των χωρικών αυτοπαλίνδρομων κατανεμημένων υστερήσεων. Το τελευταίο αποτελεί ένα γενικότερο υπόδειγμα που εμφωλεύει τα δύο πρώτα υποδείγματα αλλά και το υπόδειγμα χωρικού σφάλματος ως την ειδική περίπτωση που είναι γνωστή ως χωρικό υπόδειγμα Durbin. Τα αποτελέσματα των προσομοιώσεων φανερώνουν ότι τα πληροφοριακά κριτήρια αποτελούν εργαλεία που μπορούν να βοηθήσουν τον ερευνητή σε μια χωρική οικονομετρική ανάλυση να επιλέξει το σωστό υπόδειγμα συνεισφέροντας στα αποτελέσματα των ελέγχων χωρικής εξάρτησης οι οποίοι ίσως θα μπορούσαν και να παραλειφθούν από τα στάδια της ανάλυσης.

Ιδιαίτερα, πρέπει να αναφερθεί η περίπτωση που στο πραγματικό υπόδειγμα περιλαμβάνεται η χωρική υστέρηση της εξαρτημένης μεταβλητής, δηλαδή όταν ισχύει το υπόδειγμα χωρικής υστέρησης ή το υπόδειγμα των χωρικών αυτοπαλίνδρομων κατανεμημένων υστερήσεων καθώς τότε πρόκειται για ουσιαστική χωρική εξάρτηση και οι χωρικές επιδράσεις προκαλούν αποτελέσματα διάχυσης στο χώρο. Όταν υπάρχει ουσιαστική χωρική εξάρτηση και το τελικό υπόδειγμα που επιλεχθεί από τον ερευνητή δεν περιέχει τη χωρική υστέρηση της εξαρτημένη μεταβλητής δημιουργούνται σοβαρά προβλήματα αμεροληψίας των συντελεστών αλλά και ερμηνείας καθαυτών των αποτελεσμάτων. Ωστόσο, οι προσομοιώσεις έδειξαν ότι και τα τρία πληροφοριακά κριτήρια εντοπίζουν την αναγκαιότητα της χωρικής υστέρησης στην εξαρτημένη μεταβλητή και επιλέγουν μόνο μεταξύ του υποδείματος χωρικής υστέρησης και του υποδείματος των χωρικών αυτοπαλίνδρομων κατανεμημένων υστερήσεων.

Ένα ζήτημα που πρέπει να αποσαφηνιστεί αφορά την ερμηνεία των συντελεστών στα χωρικά οικονομετρικά υποδείγματα. Ειδικότερα, οι συντελεστές στα υποδείματα χωρικής υστέρησης και χωρικών αυτοπαλίνδρομων κατανεμημένων υστερήσεων δεν πρέπει να ερμηνεύονται ως μερικές μεταβολές όπως ισχύει στο απλό υπόδειγμα και στο υπόδειγμα χωρικού σφάλματος. Αυτό συμβαίνει διότι η μεταβολή της τιμής στην ανεξάρτητη μεταβλητή σε κάποια περιοχή θα επηρεάσει τις τιμές της εξαρτημένης μεταβλητής σε όλες τις περιοχές του χώρου όπως παρουσιάζουν αναλυτικά οι LeSage και Pace (2009) που προτείνουν συγκεκριμένους τρόπους ερμηνείας. Επομένως, όταν το υπόδειγμα χωρικού σφάλματος εκτιμάται στη μορφή του χωρικού Durbin υποδείματος προκύπτει πρόβλημα λανθασμένης ερμηνείας των συντελεστών. Εντούτοις, αποδεικνύεται ότι σε αυτή την περίπτωση η αντικατάσταση του περιορισμού του κοινού παράγοντα στη μήτρα του χωρικού πολλαπλασιαστή θα απαλείψει τα αποτελέσματα διάχυσης και θα οδηγήσει στη συνηθισμένη ερμηνεία από τη μεθοδολογία της ανάλυσης παλινδρόμησης. Κατά συνέπεια, ο έμμεσος τρόπος εκτίμησης του υποδείματος χωρικού σφάλματος ως εμφωλευμένου υποδείματος που χρησιμοποιήθηκε στις προσομοιώσεις έχει εγκυρότητα αφού οδηγεί στα ίδια αποτελέσματα με την άμεση εκτίμησή του.

ABSTRACT

The problem of spatial autocorrelation in the errors of a linear model is treated by selecting and estimating the appropriate spatial econometric model. In this study, using simulation techniques, we examine the ability of the three most commonly applied in practice information criteria which are Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC) and Hannan and Quinn's Criterion (HQC) to select the correct model. The results indicate that the adoption of an information criterion contributes significantly to choosing the best model.

ΑΝΑΦΟΡΕΣ

Akaike H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle, in B. N Petrov and Csaki F.(eds), *Proceedings of the Second*

- International Symposium of Information Theory*, Academia Kiado, Budapest, 267-281.
- Anselin, L. (1988). *Spatial Econometrics: Methods and Models*, Kluwer Academic Publishers, Dordrecht.
- Anselin, L. (1988b). Lagrange Multiplier Test Diagnostics for Spatial Dependence and Spatial Heterogeneity. *Geographical Analysis*, **20**, 1-17.
- Anselin L., Bera A., Florax R. and Yoon, M. (1996). Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics*, **26**, 77-104.
- Anselin, L. and Florax, R. J. (1995). Small sample properties of tests for spatial dependence in regression models: Some further results, in Anselin, L. and Florax, R. J., (eds.). *New Directions in Spatial Econometrics*, 21-74. Springer-Verlag, Berlin.
- Bivand R. (1984). Regression Modeling with Spatial Dependence: An Application of Some Class Selection and Estimation Methods. *Geographical Analysis*, **16**, 25-37.
- Bivand R. (2015). Spdep: Spatial Dependence: Weighting Schemes. Statistics and Models. R package version 0.5-82. <http://CRAN.R-project.org/package=spdep>
- Burridge, P. (1980). On the Cliff-Ord Test for Spatial Correlation, *Journal of the Royal Statistical Society B*, **42**, 107-108.
- Hannan E. J. and Quinn B. G. (1979). The Determination of the Order of an Autoregression. *Journal of the Royal Statistical Society, Series B*, **41**, 190-195.
- Lauridsen, J. (2006). Spatial Autoregressively Distributed Lag Models: Equivalent Forms, Estimation, and an Illustrative Commuting Model. *The Annals of Regional Science*, **40**, 297-311.
- LeSage, J. and Pace, R. (2009). *Introduction to Spatial Econometrics*, Chapman & Hall/CRC, Boca Raton.
- Mur J. and Angulo A. M. (2006). The Spatial Durbin Model and the Common Factor Tests. *Spatial Economic Analysis*, **1**, 207-226.
- Ord J. K (1975). Estimation Methods for Models of Spatial Interaction. *Journal of the American Statistical Association*, **70**, 120-126.
- Schwarz G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, **6**, 461-464.
- Τσιμπάνος Α., Αγιακλόγλου Χ. και Τσίμπος Κλ. (2016). Η συμπεριφορά των ελέγχων χωρικής εξάρτησης σε μεγάλα δείγματα. *Πρακτικά 29^{ου} Πανελληνίου Συνεδρίου Στατιστικής*, Ελληνικό Στατιστικό Ινστιτούτο, 183-193.

εργασίες

στα αγγλικά



On Time Series Clustering Metrics with Applications in Healthcare Systems

*A. Anastasiou¹, P. Hatzopoulos², A. Karagrigoriou¹,
G. Mavridoglou³*

¹Laboratory of Statistics and Data Analysis, Department of Statistics and Actuarial-Financial Mathematics, University of the Aegean
{sasm17001, alex.karagrigoriou}@aegean.gr

²Department of Statistics and Actuarial-Financial Mathematics, University of the Aegean
xatzopoulos@aegean.gr

³Department of Accounting and Finance, University of Peloponnese
g.mavridoglou@teipel.gr

ABSTRACT

In this article our purpose is to present and discuss Multivariate Time Series Clustering Techniques and explore how these techniques can be applied to health systems. We describe various methods for clustering and see how similarity measures affect the statistical procedures. One of the contributions of this work is the development of two new distance measure algorithms, called Causality Within Groups (CAWG) and Causality Between Groups (CABG) both of which are based on the well-known Granger Causality. The proposed distance algorithms are suitable for classification purposes for the analysis of multivariate time series data with emphasis on financial and economic data where causal relationships are frequently present. In order to explore the appropriateness of the methodology, we implement the proposed algorithms to hierarchical clustering for the classification of 19 EU countries based on 16 variables related to the health system including health resources and health outcomes. Classical metrics for time series classification, like the Autocorrelation and the Partial Autocorrelation distance are also considered for comparative purposes. The experimental results are indicative of the most efficient countries. Based on the CAWG algorithm we conclude that almost all countries are quite similar while the use of CABG algorithm implies that the extent of causality between Northern and Southern European countries is of the same magnitude.

Keywords: TimeSeries, Granger Causality, Clustering, Distances, Healthcare systems, Multivariate Analysis, Pattern Recognition, Data Mining.

1. INTRODUCTION

In time series analysis and generally in statistics we are interested in the correlation between variables which is accomplished with the aim of discovering the degree and

the extent of their association. Such statistical techniques include among others the autocovariance, the autocorrelation, the Pearson correlation coefficient etc. However, in most cases it is hard to detect whether two variables cause one another or are independent of each other or only one is causing the other. The difficulties of establishing a causal relationship between economic variables led Granger ([1]) to develop the economic concept of causality known as Granger Causality. In this work we take into consideration the causal relationship measured by the Granger Causality and propose two new metric algorithms for measuring the distance (closeness) between two time series. The proposed algorithms are suitable for classification purposes for both univariate and multivariate time series where causal relationships are frequently present. Such techniques are highly useful in cases where financial variables and/or economic indicators are measured across groups (regions, zones, countries etc) and the purpose of the analysis is the classification into clusters based on the degree of closeness or similarity among groups (see [7], [8], [9], [13]).

Univariate and multivariate time series techniques have numerous applications in various fields. A field of great significance with considerable economic as well as social impact is the dynamic study of the effectiveness of healthcare systems.

In most developed countries, the state devotes a considerable share of resources to healthcare, and healthcare costs have been steadily increasing, to current levels, over 10% of Gross Domestic Product (GDP) globally, with few signs of slowing down in the future. In many countries of the Organization for Economic Co-operation and Development (OECD), the health care system constitutes the largest service industry.

To explain cost increasing, many studies have tried to diagnose the underlying factors such as an ageing population, increased social expectations, broader insurance coverage, supplier induced demand and relative prices as well as productivity growth in healthcare that may affect the utilization and costs of healthcare services.

The results of a World Bank Health paper ([4]) and a World Health Organization study for measuring healthcare efficiency using different performance indicators, reveal enormous variability in health outcomes, despite similar income and education levels across the regions covered by the studies. This generated considerable interest in the measurement of healthcare efficiency. Another interesting issue to be considered was the comparison of the effectiveness of healthcare systems between different countries. For further reading see [3], [5], [6], [7], [10] references.

In this article we introduce two new distance measures based on Granger Causality for classification purposes. The fact that necessitates the development of new measures (metrics) is the understanding that groups should be combined into clusters according to the degree and extent of their causality. In order to explore the applicability and the appropriateness of the proposed methodology, the above new metrics are implemented to healthcare efficiency for combining and classifying via hierarchical clustering, 19 EU countries according to the variables of two datasets of a survey of OECD covering the period 1999-2016. Classical distance measures of association like autocorrelation and partial autocorrelation are also used, for comparative purposes. The presentation of the paper is as follows:

In Sections 2.2-2.3 we discuss classical distance measures and present the new distance measures based on Granger Causality while in Section 2.4 we present the data description. Furthermore, in Section 3 we provide the results of the application to healthcare efficiency data from OECD. The paper concludes with some general results and comments.

2. MATERIALS AND METHODS

In this section we introduce two new distance measures for time series based on the concept of Granger Causality. The Section ends with recalling the hierarchical clustering algorithm to be used in the following Section in conjunction with the distance measures introduced in this Section.

A useful measure of dependence, since it is dimensionless, is the autocorrelation function (ACF) usually denoted by $\rho(h)$, which measures the potential serial correlation of a time series with itself shifted in time. Another classical measure of dependence is the partial autocorrelation function which is denoted by $\pi(h)$ and gives the partial correlation of a time series with its own lagged values, controlling for the values of the time series at all shorter lags.

2.1 Granger Casuality

The concept of Granger Causality between two time series X_t and Y_t is presented below. Let us assume that X_t and Y_t are two time series and Ω_t is the probability space containing all the information up to time t . Then X_t is said not to Granger-cause Y_t if for all $h > 0$,

$$F(Y_{t+h}/\Omega_t) = F(Y_{t+h}/\Omega_t - X_t)$$

where $F(.|.)$ denotes the conditional distribution and $\Omega_t - X_t$ contains all the information except series X_t . In other words, X_t is said to not Granger-cause Y_t if X cannot help in predicting a future value of Y .

For the Granger Test for causality the following autoregressions are considered

$$Y_t = \mu_0 + \sum_{i=1}^h a_i Y_{t-i} + \sum_{j=1}^k b_j X_{t-j} + u_t \quad (1)$$

$$X_t = \phi_0 + \sum_{i=1}^k c_i Y_{t-i} + \sum_{j=1}^h d_j X_{t-j} + e_t \quad (2)$$

where $\mu_0, \phi_0, a_i, b_i, c_i, d_i, i = 1, 2, \dots, h$ are appropriate coefficients and u_t & e_t are the error sequences. Consider also the restricted autoregression associated with (1) where Y_t is regressed only on its past values excluding all X_t terms. A similar restricted autoregressions associated with (2) can be obtained for X_t .

- If in (1) the coefficients b_j are not statistically significant at a given significance level while in (2) the coefficients c_i are statistically significant then we conclude that Y_t is causing according to Granger, X_t .

- If in (1) the coefficients b_j are statistically significant at a given significance level while in (2) the coefficients c_i are not statistically significant then we conclude that X_t is causing according to Granger, Y_t .

- If all b's and c's are statistically significant at a given significance level then, there is a two-way causality.

- If the coefficients b_j in (1) are not statistically significant at a given significance level and the same applies to coefficients c_i in (2), X_t and Y_t are independent.

The hypothesis that Y_t is causing according to Granger X_t is tested by using the test statistic F defined by

$$F_{x,y} = \frac{(SSE^* - SSE)/h}{SSE/(n-k)} \quad (3)$$

where, SSE^* is the restricted sum of squares of residuals associated with (2), SSE is the unrestricted sum of squares of residuals, n is the sample size, h is the number of lags and k the number of parameters of X_t , in (2). For the hypothesis that X_t is causing Y_t the test statistic $F_{Y,X}$ is given by (3) where SSE^* is the restricted SSE associated with (1). Under the null hypothesis the test statistic F follows an F -distribution with h and $(n - k)$ degrees freedom. For further details see [1], [12].

2.2 Classical Distance Measures for Time Series

Autocorrelation Distance computes the dissimilarity between a pair of numeric time series based on their estimated autocorrelation (or partial autocorrelation) coefficients. In such settings a straightforward measure of distance is based on the computation of the autocorrelation coefficients $\rho_i = (\rho_i(1), \dots, \rho_i(h))'$ for some h such that $\rho_i(j) \approx 0$ for $j > h$ and then is used to define the measure between X_i and X_j as follows:

$$D(X_i, X_j) = (\rho_i - \rho_j)' W (\rho_i - \rho_j)$$

where W is a weighting function that can be used to assign weights to the coefficients that decrease with the lag. The distance matrix is obtained through the distance measure $D(X_i, X_j) \forall i, j$. A similar distance can be acquired with the use of partial autocorrelation function with $\pi_i = (\pi_i(1), \dots, \pi_i(h))'$ in the place of ρ_i .

The above distance like any other typical distance measure e.g. (City-Block, Minkowski, Mahalanobis etc.) can be used for classification purposes. For these purposes hierarchical clustering algorithms could be used.

2.3 Granger Causality Distance Measures For Time Series

In this section we introduce new distance measures that could be used in hierarchical clustering for multivariate Time Series which are based on the Granger Causality

Test. The proposal is based on the idea that variables should be combined as long as the causalities among the variables/elements of a multivariate Time Series are similar to the causalities of the same variables of another multivariate Time Series.

2.3.1 The CAWG Distance Algorithm

Let us consider Q groups of Multivariate Time Series. Each group has the following structure

$$X^q = (X_{t1}^q, \dots, X_{ti}^q, \dots, X_{tk}^q), q = 1, \dots, Q$$

where,

- X_{ti}^q is a univariate time series variable, $t = 1, 2, \dots$ of the group q , $q = 1, \dots, Q$.
- k : number of variables (univariate time series).

For the Causality Within Groups (CAWG) Distance the following algorithm is used:

CAWG Algorithm

1. For $q = 1$ using (2) calculate the F_{ij}^q value of the Granger Causality test which tests whether X_{ij}^q causes X_{ti}^q , $i = 1, \dots, k - 1$, $j = 2, \dots, k$, $i < j$. The total number of F values is equal to $[k(k - 1)]/2$.
2. Repeat step 1 for $q = 2$.
3. Compute the squared differences between each pair of the F_{ij}^1 and the F_{ij}^2 values and then calculate their sum.
4. Repeat steps 1-3 for the calculation of F_{ij} values and their squared differences for each pair $q_1 \& q_2$ such that $q_1, q_2 = 1, \dots, Q$, with $q_1 \neq q_2$.

According to the above algorithm, the Granger causality Within Groups (CAWG) distance between two multivariate Time Series X^{q_1} and X^{q_2} is defined as follows:

$$CAWG(X^{q_1}, X^{q_2}) = \sum_{i=1}^{k-1} \left| \sum_{j=2}^k [F_{ij}^{q_1} - F_{ij}^{q_2}]^2 \right|, i < j$$

Observe that $CAWG(X^{q_1}, X^{q_2})$ for $q_1 = q_2$ equals zero.

At the end of the algorithm the resulting Causality Distance $Q \times Q$ matrix could be useful for classification purposes and especially in hierarchical clustering for classifying the Q groups into clusters.

2.3.2 The CABG Distance Algorithm

For the calculation of CABG Distance $Q \times Q$ matrix for $Q = \#$ multivariate time series we use the following algorithm:

CABG Algorithm

1. For $q = 1 \& q = 2$ calculate the $F_i^{1,2}$ value of the Granger Causality test which tests whether X_{ti}^2 causes X_{ti}^1 . The total number of F_i values is equal to k , $i = 1, 2, \dots, k$.
2. For $q = 1 \& q = 2$ calculate the $F_i^{2,1}$ value of the Granger Causality test which tests whether X_{ti}^1 causes X_{ti}^2 .
3. Add $F_i^{1,2}$ and $F_i^{2,1}$ values.
4. Compute the converse fraction of the above sum.
5. Repeat steps 1-4 for each combination of q_1, q_2 with $q_1, q_2 = 1, \dots, Q$, $q_1 \neq q_2$.

Thus, the Granger Causality Between Groups (CABG) distance between the multivariate Time Series $X^{q_1} \& X^{q_2}$ is defined as follows:

$$CABG(X^{q_1}, X^{q_2}) = \frac{1}{\sum_{i=1}^k (F_i^{q_1, q_2} + F_i^{q_2, q_1})}$$

We define $CABG(X^{q_1}, X^{q_2}) \equiv 0$ for $q_1 = q_2$.

As in the case of CAWG algorithm, the CABG algorithm results in a Causality Distance $Q \times Q$ matrix which could be useful for classification purposes and especially in hierarchical clustering for classifying the Q groups into clusters.

CAWG distance joins multivariate time series with similar causalities (either weak or strong) among their components while in the CABG distance the series are joined together as long as elementwise the causalities are strong.

The class of pseudodistances is a general class of distances that contains measures that do not necessarily satisfy the symmetric property. In general this class requires all its members to satisfy the properties of non-negativity and equality to zero when the two arguments coincide. Additional properties (like the symmetric property) may or may not be satisfied by a member of the pseudodistance class. Observe that both proposed distances belong to the above class. Notice though that only CABG distance satisfies additionally, by construction, the symmetric property.

2.4 Hierarchical Algorithms

The hierarchical clustering technique to be used in Section 3 in conjunction with the distance measures of the present Section. Hierarchical algorithms, with a widespread use in practice, for performing cluster analysis, are based on a series of successive groupings (Agglomerative algorithms) or successive divisions (Divisive algorithms).

Having as primary goal the distance measurement between groups, standard methods like the nearest neighbour or furthest neighbour etc are used. For details see [14].

2.5 Data

In order to implement the proposed methodology we consider now two sets of variables which are associated with the health care system. The data sources used in

our study, were collected from the OECD ILibrary and The EUROSTAT Health Database and concern 19 EU countries, as shown in Table 1. The survey focused on EU countries, without taking into account the health care system of the respective jurisdictions.

Table 1. Countries analyzed in the survey.

Id	Country	Code
1/e1	Austria	AUT
2/e2	Belgium	BEL
3/e3	Czech Republic	CZE
4/e4	Denmark	DNK
5/e5	Finland	FIN
6/e6	France	FRA
7/e7	Germany	DEU
8/e8	Greece	GRC
9/e9	Hungary	HUN
10/e10	Ireland	IRL
11/e11	Italy	ITA
12/e12	Luxembourg	LUX
13/e13	Netherlands	NLD
14/e14	Poland	POL
15/e15	Portugal	PRT
16/e16	Slovakia	SVK
17/e17	Spain	ESP
18/e18	Sweden	SWE
19/e19	United Kingdom of Great Britain and Northern Ireland	GBR

For each country, 16 variables were collected spanning from 1999 to 2016, of which 7 describe health resources and health equipments (input variables, Table 2) and the rest are related to the system's efficiency (output variables, Table 3).

Table 2. Input Variables.

Name	Variable
in1	Total Health Spendings as a share of GDP
in2	Government Spendings as a share of Total Spendings
in3	Out of Pocket as a share of Total Spendings
in4	Pharmaceutical spending as a share of Total Spendings
in5	Doctors per 1000 inhabitants
in6	Nurses per 1000 inhabitants
in7	Beds per 1000 inhabitants

In order to handle missing values, we proceed with imputations using the linear interpolation technique. The percentage of missing values in the input dataset is approximately 10% while for the output dataset is 20%. The analysis of the data was conducted with R.

Table 3. Output Variables.

Name	Variable
out1	Length of Stay
out2	Hospital discharge rate 100000
out3	Life Expectancy at birth
out4	Life Expectancy at birth - Male
out5	Life Expectancy at birth - Female
out6	Life Expectancy at 65 - Male
out7	Life Expectancy at 65 - Female
out8	Healthy life expectancy at birth (years) - Both sexes
out9	Healthy life expectancy at birth (years) - At 60

3 RESULTS

Implementing the proposed methodology to the input (Table 2) and output (Table 3) datasets, we observe that both CAWG and CABG distances have been used in this Section (Subsections 3.3 and 3.4 respectively). The distance among all possible pairs of countries (Table 1) in each dataset has been calculated according to the algorithms in Subsections 2.3.1 and 2.3.2. For comparative purposes the ACF distance has been applied to the input dataset (Subsection 3.1) while the PACF distance has been applied to the output dataset (Subsection 3.2).

Note that the agglomerative algorithm and the complete linkage method have been used for illustrative purposes. Equally effective would have been the divisive algorithm in conjunction with any other of the well-known distances between clusters (single linkage, average linkage etc.).

One of the most common issues encountered in cluster analysis is the need to know in advance the required number of clusters. The most popular methods which one can use to confront this weakness are the Silhouette and Elbow methods. For the case of the two datasets under investigation, both methods suggest the creation of 2 or 3 clusters for the variables of Tables 2 and 3.

3.1 Clustering with Autocorrelation Distance

Following the method presented in Section 2.2 we obtain the autocorrelation distance between each pair of countries of Table 1 for the input dataset. Initially, each country forms a cluster of its own so that the clustering algorithm starts with 19 single-element clusters. For $k = 2$ the two clusters formed are:

Cluster 1: Hungary and Luxembourg.

Cluster 2: All the other countries.

For $k = 3$ the 2nd cluster above is divided (dotted vertical line in Figure 1) into two new clusters:

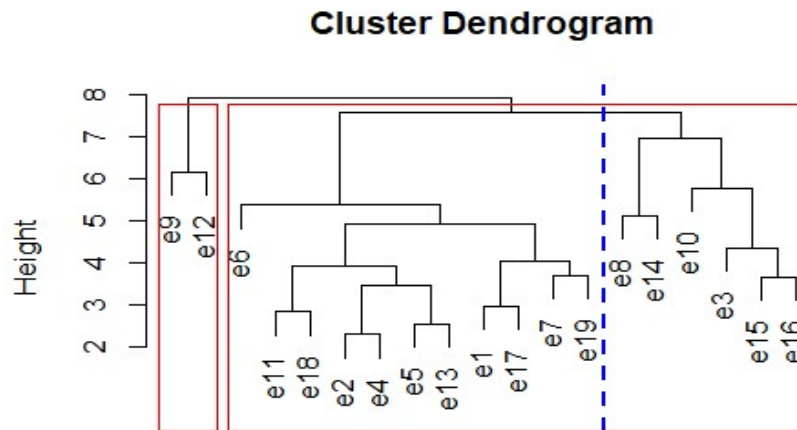
Cluster 2: France, Italy, Sweden, Belgium, Denmark, Finland, Netherlands,

Austria, Spain, United Kingdom and Germany.

Cluster 3: Greece, Poland, Ireland, Czech R., Portugal, Slovakia.

Observe that the first of the two new clusters consists of economically strong EU countries while in the other not so economically strong countries are joined together (Czech Republic, Poland, Slovakia etc).

Figure 1: Clusters using the autocorrelation distance, $k = 2,3$. The vertical dotted line divides cluster #2 into two new clusters.



The reader is referred to Table 1 for the numbering of countries.

3.2 Clustering with Partial Autocorrelation Distance

Since the purpose is an illustrative one the input dataset was chosen earlier for the ACF distance and the output dataset is chosen in this subsection for the PACF distance. By applying the above procedure to the output health dataset, Luxembourg forms one cluster by itself while all other countries form a second cluster ($k = 2$). For $k = 3$, the second cluster is divided into two new clusters (see Figure 2):

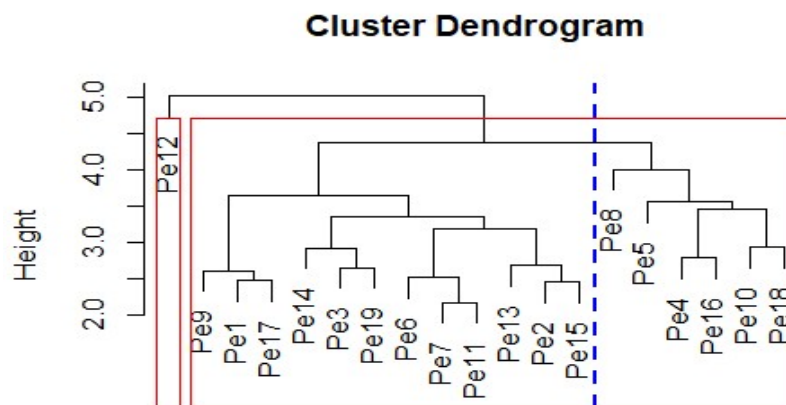
Cluster 2: Hungary, Poland, Czech republic, France, Italy, Portugal, Belgium, The Netherlands, Austria, Spain, United Kingdom and Germany.

Cluster 3: Greece, Ireland, Finland, Denmark, Slovakia and Sweden.

Observe that in this case the third cluster (see the dotted vertical line in Figure 2) consists of three not so economically strong countries with three Scandinavian countries.

Although the complete linkage has been used in this work alternative methods like the single linkage, the average linkage and Ward can also be used. Furthermore, note that the focus of this work is on the presentation and investigation of distances appropriate for time series data so that the use of a single method like the complete linkage, suffices for our purposes.

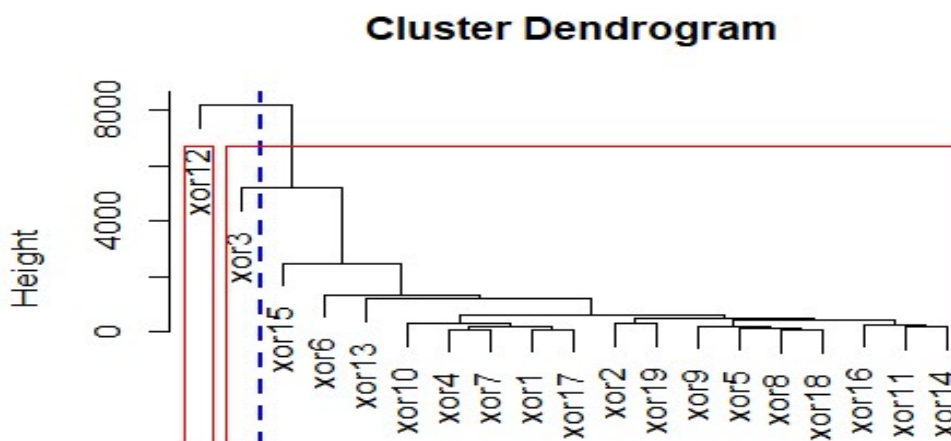
Figure 2: Clusters using the partial autocorrelation distance, $k = 2,3$. The vertical dotted line divides cluster #2 into two new clusters.



3.3 Clustering with CAWG Distance

In order to explore the capabilities of the proposed causality distance, namely CAWG distance, we proceed below with its implementation into the hierarchical clustering method. Following the algorithm which is described on section 2.3.1 we apply the CAWG distance. Consequently initially for input and afterwards for output datasets the following Experimental Results are obtained. The results for the input dataset are summarized in the hierarchical trees that are presented in Figure 3.

Figure 3: Clusters using the CAWG distance for input dataset, $k = 2,3$. The vertical dotted line divides cluster #2 into two new clusters.



In particular, for the case of two clusters, the first cluster contains only Luxembourg with all other countries staying together in a single cluster. For the case

of 3 clusters, Czech Republic is separated from the large cluster created above and creates a new cluster of its own (see the dotted vertical line in Figure 3).

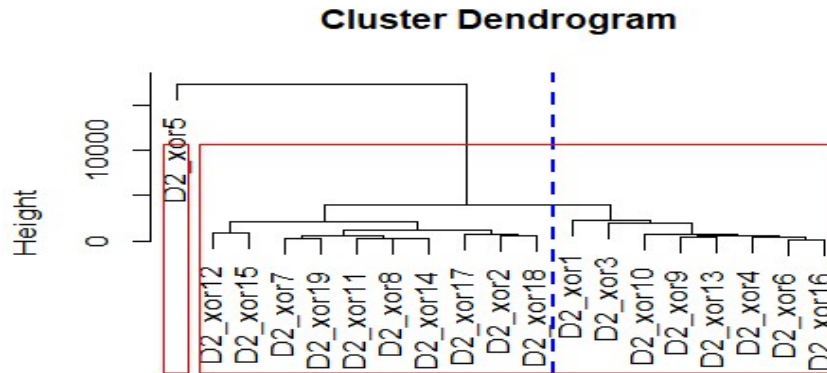
The results indicate that the causalities obtained are more or less similar among almost all the countries participated in this analysis with the exception of Luxembourg and Czech Republic. Indeed, all causalities are in the region of 100 – 2000 while for Luxembourg is equal to 5000 and for Czech Republic is equal to 3000.

For the output dataset we observe (Figure 4) that if the number of clusters is two then the first cluster consists of only one country (Finland) and all other countries build the second cluster. The situation changes dramatically for the case of $k = 3$ since the above 2nd large cluster is split into two new clusters:

Cluster 2: Greece, Belgium, Germany, Italy, Portugal, Poland, Spain, Luxembourg, Sweden and United Kingdom.

Cluster 3: Austria, Ireland, Czech Republic, Hungary, the Netherlands, Denmark, Slovakia and France.

Figure 4: Clusters using the CAWG distance for output dataset, $k = 2,3$. The vertical dotted line divides cluster #2 into two new clusters.



3.4 Clustering with CABG Distance

The Experimental Results of the application of the CABG distance defined in section 2.3.2 to the input and output datasets are presented in Figure 5 & 6 respectively. For the input dataset and for $k = 2$ the two clusters are:

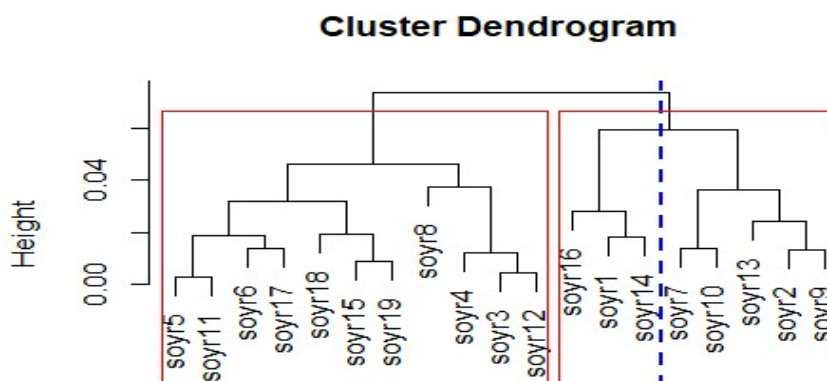
Cluster 1: Austria, Belgium, Germany, Hungary, Ireland, Netherlands, Poland and Slovakia.

Cluster 2: Czech Republic, Denmark, Finland, France, Greece, Italy, Luxembourg, Portugal, Spain, Sweden and United Kingdom.

Observe that most countries in Cluster 2 are in the outside borders of the European continent (North and South Europe) while most of the countries in Cluster 1 lie

geographically in the central part of Europe. For $k = 3$, a third cluster is created when Austria, Poland and Slovakia are separated from cluster #1 above.

Figure 5: Clusters using the CABG distance for input dataset, $k = 2,3$. The vertical dotted line divides cluster #2 into two new clusters.



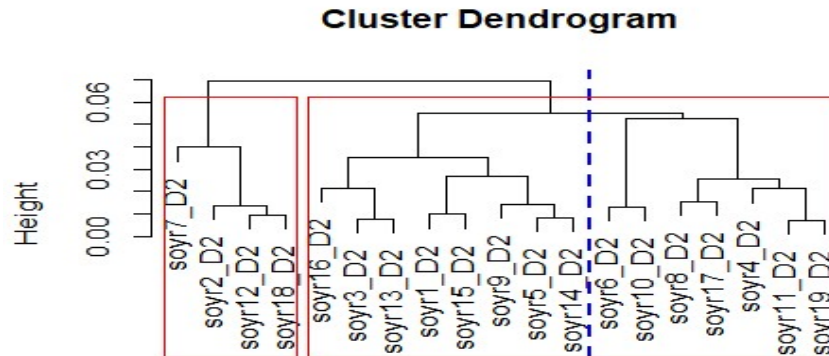
For the output dataset and for $k = 2$ the two clusters are:

Cluster 1: Belgium, Germany, Luxembourg, Sweden.

Cluster 2: Austria, Slovakia, Czech Republic, Netherlands, Portugal, Hungary, Finland, Poland, Denmark, France, Greece, Ireland, Italy, Spain and United Kingdom.

Observe that most countries in Cluster 1 have a high degree of development in areas such as the health care. For $k = 3$, a third cluster is created when Denmark, France, Greece, Ireland, Italy, Spain and United Kingdom are separated from cluster #2 above.

Figure 6: Clusters using the CABG distance for output dataset, $k = 2,3$. The vertical dotted line divides cluster #2 into two new clusters.



4.1 Specific Statements

4.1.1 Input Dataset

- For the autocorrelation method and for $k = 3$, the second cluster consists of economically strong EU countries while in the third not so economically strong countries are joined together (Czech Republic, Poland, Slovakia etc).

- For the CAWG Distance the results indicate that the causalities obtained are more or less similar among almost all the countries participated in this analysis with the exception of Luxembourg and Czech Republic (for the case of $k = 3$ clusters). Indeed, all causalities are in the region of 100-2000 while for Luxembourg is equal to 5000 and for Czech Republic is equal to 3000. This clearly shows that the overall Granger causalities among the elements of each multivariate Time Series are quite similar across the EU countries. Thus, the effect of the overall cause is more or less the same among the European countries implying that the causal relationship between the variables (the elements of the multivariate Time Series) affecting the health system, is quite similar across Europe.

- For the CABG Distance and for $k = 2$, most countries in Cluster 1 are in the outside borders of the European continent (North and South Europe) while most of the countries in Cluster 2 lie geographically, in the central part of Europe. This clearly shows that multivariate Time Series are combined as long as they have very strong overall causalities.

4.1.2 Output Dataset

- In the case of Partial Autocorrelation method and for $k = 3$, the third cluster consists of 3 not so economically strong countries together with three Scandinavian countries.

- For the CABG Distance, most countries in Cluster 1 have a high degree of development in areas such as the health care.

4.2 Global Statements

Generally all results provide an indication of the most efficient countries. Based on the Causality Within Groups we conclude that almost all countries are quite similar (only Luxembourg and Czech Republic are further away) which implies that the extend of overall causality within variables is of the same magnitude for almost all countries examined.

On the other hand the overall Causality Between Groups (i.e. between countries) results in combining together countries of North and South Europe leaving on their own, Central European countries to be combined in a separate entity. This observation implies that the extend of causality between Northern and Southern European countries is of the same magnitude with causalities lower than the causalities between central European countries.

Based on the results of the present analysis, we observe that the classification is strongly related to the distance/similarity measure considered. Different measures give rise to different classifications. It is clear from the analysis that the distance measure plays an important role so that the investigator should choose the one that is preferable according to the issue under investigation in order to incorporate into the resulted classification the special characteristics of time series data. Moreover, it is natural to explore and implement more than one distance measures in which case a comparison between the methods should be made by quantifying and evaluating the appropriateness of each method via a proper loss/error function.

5 Conclusions

Due to rapid integration of international economic markets, causal relationships are considered to be vital in international economy since the identification of a possible impact could be used to alter or void economic policies, prevent socio-economic crises or enforce the same economic or financial decisions to groups with similar causal relationships. A metric that takes into consideration any kind of causal relationship like the one introduced by Granger and used in this article for proposing CABG and CAWG metrics, is therefore recommended for classification purposes to ensure as accurate and as precise as possible decision making across groups with similar causalities. The contribution of this article lies on the proposal of the new distances CABG and CAWG which are based on the idea that groups, regions or countries should be combined into clusters as long as the causalities among the elements of two multivariate time series are similar.

Acknowledgements: The authors wish to express their appreciation to the anonymous reviewers for their comments and suggestions that greatly improved the quality and presentation of the manuscript. This work was completed as part of the activities of the Laboratory of Statistics and Data Analysis of the University of the Aegean.

ΠΕΡΙΛΗΨΗ

Σκοπός της εργασίας αυτής είναι η παρουσίαση ΤεχνικώνΣυσταδοποίησης Πολυδιάστατων Χρονοσειρών και η αξιοποίησή τους μεταξύ άλλων, σε συστήματα υγείας. Μία από τις συνεισφορές αυτής της μελέτης είναι η ανάπτυξη δύο νέων αλγορίθμων μέτρησης απόστασης, που ονομάζονται CausalityWithinGroups (CAWG) andCausalityBetweenGroups (CABG) και βασίζονται στην Αιτιότητα κατά Granger. Οι προτεινόμενοι αλγόριθμοι απόστασης είναι κατάλληλοι για σκοπούς ταξινόμησης για την ανάλυση των δεδομένων πολυμεταβλητών χρονοσειρών με έμφαση στα οικονομικά και χρηματοπιστωτικά δεδομένα όπου συχνά υπάρχουν αιτιώδεις σχέσεις. Για να διερευνήσουμε την καταλληλότητα της μεθοδολογίας, εφαρμόζουμε τους προτεινόμενους αλγόριθμους με χρήστης μεθόδου της ιεραρχικήςσυσταδοποίησης για την ταξινόμηση 19 χωρών της ΕΕ με βάση 16μεταβλητές που σχετίζονται με το σύστημα υγείας, συμπεριλαμβανομένων των πόρων υγείας και των αποτελεσμάτων της υγείας. Οι κλασσικές μέθοδοι απόστασης που βασίζονται στην αυτοσυσχέτιση και τη μερική αυτοσυσχέτιση, εξετάζονται επίσης για συγκριτικούς σκοπούς. Τα πειραματικά αποτελέσματα αναδεικνύουν ενδείξεις για την υπαρξηαποδοτικώνχωρών. Με βάση τον αλγόριθμο CAWG καταλήγουμε στο συμπέρασμα ότι σχεδόν όλες οι χώρες είναι αρκετά παρόμοιες, ενώ η χρήση του αλγορίθμου CABG υποδηλώνει ότι το εύρος της αιτιότητας μεταξύ των χωρών της Βόρειας και της Νότιας Ευρώπης έχει το ίδιο μέγεθος.

ΑΝΑΦΟΡΕΣ

1. Granger, C.W.J.(1969) Investigating causal relation by econometric and cross-sectional method. *Econometrica*,**37**, 424–438.
2. Organisation for Economic Co-operation and Development, OECD, <http://www.oecd.org/>
3. Paolucci, F. (2011) Health financing and insurance: options for design. In: Frech HE, Zweifel P, editors. *Series:Developments in Health Economics and Public Policy*, **10**.
4. Wang, J., Jamison, D., Bos, E., Preker, A., Peabody, J. (1999)Measuring country performance on health - selected 648 indicators for 115 countries. *Health, nutrition, and population series*.
5. WHO, *World Health Report 2000*, Geneva, Switzerland. WHO.
6. Varabyova, Y., Schreygg, J. (2013) International comparisons of the technical efficiency of the hospital sector: Panel 656 data analysis of OECD countries using parametric and non-parametric approaches. *Health Policy*, **112**,70–79.
7. Liao, T. W. (2005) Clustering of time series data—a survey. *Pattern recognition*, **38**(11), 1857–1874.
8. Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015) Time-series clustering—A decade review. *InformationSystems*, **53**, 16–38.
9. Fu, T. C. (2011) A review on time series data mining. *Engineering Applications of Artificial Intelligence*, **24**(1),164–181.
10. Cinaroglu, S. (2016) Clustering of OECD Countries Out of Pocket Health Expenditure Time Series Data. *Research in Applied Economics*, **8**, 23–38.

11. Lefevre, T., Rondet, C., Parizot, I., Chauvin, P. (2014) Applying Multivariate Clustering Techniques to Health Data: The 4 Types of Healthcare Utilization in the Paris Metropolitan Area. *PLOS ONE*, **19**.
12. Siggiridou, E., Kugiumtzis D. (2016) Granger Causality in Multi-variate Time Series using a Time Ordered Restricted Vector Autoregressive Model, *IEEE Transactions on Signal Processing*, **64**(7), 1759–1773.
13. Rousseeuw P. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics* **20**, 53–65.
14. Everit B. S., Landau, S. Leese, M. and Stahl, D. (2011) *Cluster Analysis*, Wiley.

AN L_2 DISTANCE BASED GOODNESS-OF-FIT TESTD. Bagkavos¹, P.N. Patil² and A.T.A. Wood³¹University of Ioannina, Department of Mathematics,
dimitrios.bagkavos@gmail.com,²Mississippi State University, Department of Mathematics and Statistics,³ University of Nottingham, Department of Mathematical Sciences.

ABSTRACT

A goodness-of-fit test is introduced based on a piecewise local linear approximation of the L_2 distance between the null density and its estimate. The null and alternative distributions of the test statistic are established and its superior performance against the classical kernel based goodness-of-fit test for Pitman alternatives is established.

Keywords: Hypothesis testing, null distribution, power.

1. INTRODUCTION

Let X_1, X_2, \dots, X_N be a random sample from an unknown absolutely continuous distribution function F having probability density function (p.d.f.) f . A substantial amount of goodness-of-fit tests for testing the hypothesis

$$H_0 : f(x) = f_0(x) \quad (1)$$

is based on measuring the distance between an estimate of f and the null model f_0 by the Integrated Squared Error (ISE) metric. The well known Pearson's χ^2 -test belongs to this category. It uses a histogram, based on the probability integral transformed data to measure the distance between the estimated and the hypothesized probability model F . Thus, it is a test of uniformity and can be viewed as a non-parametric smoothing-based test with fixed smoothing parameter. A natural improvement on the χ^2 -test is to replace the histogram with the classical kernel density estimate

$$f_N(x) = \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x - X_i}{h}\right).$$

Here, K is a kernel function and h is the bandwidth. Naturally now h depends on N , however this dependence is suppressed in what follows. Bickel and Rosenblatt (1973) were the first to investigate the resulting test statistic, see also Rosenblatt (1975), Ghosh and Huang (1991), Fan (1994), Gouriéroux and Tenreiro (2001), Chebana (2004) and Li and Racine (2007) for extensions and improvements of this approach to various directions.

The practical performance incurred by this improvement, is that for $h \rightarrow 0$ the standardized version of the test statistic converges to the normal distribution with rate $N^{-\delta}$, $\delta \in (0, 1/4)$. It was further shown that it has a nontrivial power for the 'standard' Pitman alternatives converging to the null at the rate of $N^{-\beta}$ for $\beta = \frac{1}{2} - \frac{\delta}{4}$. An important limitation here is that as δ decreases, the rate of convergence to normality suffers. In general, tests based on the ISE of f_N will distinguish alternatives converging to null with rate $N^{-4/9}$, a slightly faster rate than $N^{-2/5}$, if the bias of f_N converges to zero at a slightly faster than its mean square error optimal rate. In other words, focusing on optimal f_N estimation instead of optimal testing may diminish the efficiency of the testing procedure. This is an intrinsic problem in ISE based hypothesis tests, see e.g. Fan (1994).

In addressing this limitation, the present article starts by adopting a regression view to density estimation. Subsequently, an equivalent statistic is obtained by testing the lack of fit of the hypothesized

density, $f_0(x)$, to the estimated population density, $\hat{f}(x)$, at each of the design points x_i and combining these tests together. The resulting test statistic is the L_2 error between the density estimate and the hypothesized density function minus the diagonal terms. Full details are given in Section 2. There, the asymptotic normality under the null and fixed alternative hypotheses are established and its power properties are analyzed. Specifically, it is shown that it has nontrivial power against alternatives converging to the null at the rate of N^β where $\beta < 1/2$ and that it converges faster to the normal than the test of Bickel and Rosenblatt (1973).

The practical performance of the proposed test in comparison to similar tests in the literature for various density shapes is discussed in Section 3. The outcome verifies that the suggested modification leads to superior practical performance against kernel ISE based tests. All proofs are deferred for the Appendix.

2. L_2 DISTANCE-BASED GOODNESS-OF-FIT TESTING

Assume a sample of i.i.d. observations X_1, \dots, X_N from a probability density function $f(x)$ on \mathcal{R} . In view of the support of $f(x)$ under H_0 for each N , let $x_{1:N} < x_{2:N} < \dots < x_{n:N}$ denote a sequence of equally-spaced points such that $x_{i+1:N} - x_{i:N} = \Delta$. For notational simplicity we shall write x_1, x_2, \dots, x_n in what follows. Define

$$Y_1 = \#\{i: X_i \leq x_2 - \Delta/2\}, \quad Y_n = \#\{i: X_i > x_{n-1} + \Delta/2\}$$

and

$$Y_j = \#\{i: X_i \in (x_j - \Delta/2, x_j + \Delta/2)\}, \quad 2 \leq j \leq n-1.$$

Now, to collect the evidence against (or for) the null hypothesis, in every h -neighborhood of x_i ($h > 0$), except at x_i , compute

$$S'_i(h) = \sum_{j=1, j \neq i}^n I(|x_j - x_i| \leq h)(Y_j - Np_j),$$

where $p_j = \int_{(x_j - \Delta/2, x_j + \Delta/2)} f(x) dx$ and $I(\cdot)$ denotes the indicator function. Note under H_0 , $E[S'_i] = 0$. Then to combine the evidence across the scope of the design points x_2, \dots, x_{n-1} compute the weighted sum of $S'_i(h)$, where weights are $Y_i - Np_i$, that is,

$$\begin{aligned} S'_N(h) &= \sum_{i=1}^n S'_i(h)(Y_i - Np_i) \\ &= \sum_{i \neq j} \sum I(|x_i - x_j| \leq h)(Y_i - Np_i)(Y_j - Np_j). \end{aligned}$$

Note that, since the true density f is assumed to be a smooth function, under H_1 , if $S'_i(h) > 0$ it is expected that $Y_i - Np_i > 0$ and if $S'_i(h) < 0$ then it is expected that $Y_i - Np_i < 0$; whereas as under the null, $E[S'_i(h)] = E[Y_i - Np_i] = 0$. Thus a test based on $S'_N(h)$ will reject the null if it is too large. The actual test statistic that we use is simply a scaled version of $S'_N(h)$ and is defined as

$$S_N(h) = \frac{1}{N(N-1)h} S'_N(h). \quad (2)$$

Remark 1. In a regression view of density estimation problem, as mentioned in Fan and Gijbels (1996), the above test statistic can be thought of as a lack of fit test of the hypothesized mean function $f_0(x, \theta)$. For that set $Y'_i = \frac{Y_i}{N\Delta}$, $i = 1, 2, \dots, n$ which are then modeled as, $Y'_i = m(x_i) + \sigma(x_i)\epsilon_i$, $i = 1, 2, \dots, n$, where $m(x) = f_0(x, \theta)$ and $\sigma^2(x) = \frac{f_0(x, \theta)}{N\Delta}$. The test statistic, $S_n^*(h)$ say, for a lack of fit test is then constructed by first considering the mean of the differences between Y'_j and $f_0(x_j)$ for all Y'_j in the neighbourhood of x_i except for Y'_i corresponding to x_i , i.e. $\frac{1}{(n-1)h} \sum_{j=1, j \neq i}^n I(|x_j - x_i| \leq h)(Y'_j - f_0(x_j))$, and adding these for each x_i with weights $Y'_i - f_0(x_i)$. That is,

$$S_n^*(h) = \frac{1}{n(n-1)h} \sum_{i \neq j} I(|x_i - x_j| \leq h)(Y'_i - f_0(x_i))(Y'_j - f_0(x_j)).$$

Noting that $\Delta \sim n^{-1}$ it is easy to conclude that $|S_N(h) - S_n^*(h)| = o_p(1)$.

We now consider two sets of assumptions.

(A1) Assume $\Delta = N^{-\alpha}$, where $\frac{1}{2} < \alpha < 1$; as $N \rightarrow \infty$, $n = n(N)$ is such that $n\Delta \rightarrow \infty$, and $x_1 \rightarrow -\infty$, $x_n \rightarrow \infty$; and $h = h(N)$ is such that $h \rightarrow 0$ and $h/\Delta \rightarrow \infty$ as $N \rightarrow \infty$.

(A2) The p.d.f $f(x)$ and its first derivative is bounded and uniformly continuous on \mathcal{R} .

The asymptotic distribution of $S_N(h)$ is established in the following theorem.

Theorem 1. Under assumptions (A1) and (A2), and under $H_0 : f(x) = f_0(x)$

$$N\sqrt{h}S_N(h) \rightarrow N(0, \sigma_0^2) \quad \text{as } N \rightarrow \infty,$$

where $\sigma_0^2 = 2R(f_0(x))$.

Remark 2. Thus, asymptotically, a size α test will reject the null hypothesis if

$$S_N(h) > z_\alpha \times \frac{\sigma_0}{N\sqrt{h}} = C(\alpha, h, f_0(x)),$$

where z_α is such that for a standard normal variable Z , $P[Z > z_\alpha] = \alpha$.

Remark 3. Clearly for $h \sim N^{-\delta}$, where $0 < \delta < 1/2$, the rate of convergence to normality, $N^{1-\delta/2}$, is much faster than the test of Bickel and Rosenblatt (1973).

Let $R(g)$ is the integral of the real valued function g^2 over its domain. The asymptotic distribution of the test statistic under a given fixed alternative follows from Theorem 1 and is stated in the following corollary.

Corollary 1. Under assumptions (A1) and (A2) and under $H_1 : f(x) = f_1(x)$

$$N\sqrt{h}S_N(h) \rightarrow N(\mu_1, \sigma_1^2) \quad \text{as } N \rightarrow \infty,$$

where $\mu_1 = 2R(f_1 - f_0)$ and $\sigma_1^2 = 2R(f_1)$.

The two types of Pitman alternative considered in the literature (see e.g. Ghosh and Huang (1991)) are

$$f_{1N}(x) = f_0(x) + N^{-\beta}\eta(x) + o(N^{-\beta}), \quad (3)$$

$$f_{2N}(x) = f_0(x) + N^{-\epsilon} \sum_{j=1}^k \eta_j\left(\frac{x - c_j}{N^{-\gamma}}\right) + o(N^{-\epsilon-\gamma}), \quad (4)$$

where $\beta > 0, \gamma > 0$ are such that $hN^\gamma \rightarrow 0$, $\epsilon > 0$, $\eta(x)$ is a continuous and square integrable function and η_j 's are square integrable, band limited continuously differentiable up to second order. The power of the test based on $S_N(h)$ against the alternatives (3), (4) follows by a direct application of Corollary 1 and is given next.

Corollary 2. *Under assumptions (A1) and (A2),*

(a) *With $l = 2\sigma_0^{-1}R(\eta) - z_\alpha$ and Φ the standard normal distribution function,*

$$P[S_N(h) > C(\alpha, h, f_0(x)) | f = f_{1N}] = \begin{cases} \alpha, & \text{if } \beta > \frac{1}{4} \frac{\log h}{\log N} + \frac{1}{2} \\ \Phi(l), & \text{if } \beta = \frac{1}{4} \frac{\log h}{\log N} + \frac{1}{2} \\ 1, & \text{if } \beta < \frac{1}{4} \frac{\log h}{\log N} + \frac{1}{2} \end{cases}$$

(b) *For $\gamma > 0$ such that $hN^\gamma \rightarrow 0$ and $\epsilon > 0$,*

$$P[S_N(h) > C(\alpha, h, f_0(x)) | f = f_{2N}] = \begin{cases} \alpha, & \text{if } 2\epsilon + \gamma > \frac{1}{2} \frac{\log h}{\log N} + 1 \\ \Phi(l'), & \text{if } 2\epsilon + \gamma = \frac{1}{2} \frac{\log h}{\log N} + 1 \\ 1, & \text{if } 2\epsilon + \gamma < \frac{1}{2} \frac{\log h}{\log N} + 1 \end{cases}$$

where $l' = \frac{2}{\sigma_0} \sum_{j=1}^k R(\eta_j) - z_\alpha$.

Remark 4. *Suppose $h \sim N^{-\delta}$ where $0 < \delta < 1/2$. Then, similar to the test of Bickel and Rosenblatt (1973), the power of the test based on $S_N(h)$ improves as δ increases. But in contrast to that test, the improvement does not conflict with the rate of its convergence to normality.*

3. NUMERICAL EXAMPLES

The practical performance of $S_N(h)$ in terms of power under Pitman alternatives is investigated in this section. A thorough summary of power comparisons, indirectly connected with the present study, involving earlier and some of the most classical tests, can be found in Table 2 of Lequesne and Regnault (2018). The performance of $S_N(h)$ is benchmarked throughout, against the Kolmogorov-Smirnov (D_N) and Cramer-von Mises (C_N) tests as well as the L_2 distance and the empirical likelihood ratio approach goodness-of-fit tests. These are respectively the tests of Fan (1994) (F_N), and the test of Gurevich and

Vexler (2010) (V_N). The tests are defined by

$$F_N = N^{-1} \sum_{i=1}^N f_{-i}(X_i) + \int (K_h * f(x, \theta))^2 dx - 2N^{-1} \sum_{i=1}^N K_h * f_0(X_i, \theta),$$

$$V_N = \min_{1 \leq M \leq N^{1/2}} (2\pi e \hat{S}^2)^2 \prod_{i=1}^N \frac{2M}{N(X_{(i+M)} - X_{(i-M)})},$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$, g is the true density with parameter(s) $\hat{\theta}$ resulting by maximum likelihood estimation, \hat{S}^2 is the unbiased sample variance estimate and

$$N^{-1} \sum_{i=1}^N \hat{f}_{-i}(X_i) = (N(N-1))^{-1} \sum_{i=1}^N \sum_{j \neq i}^N K_h(X_i - X_j),$$

$$K_h * f_0(x, \theta) = \int K_h(x-v) f_0(v, \theta) dv.$$

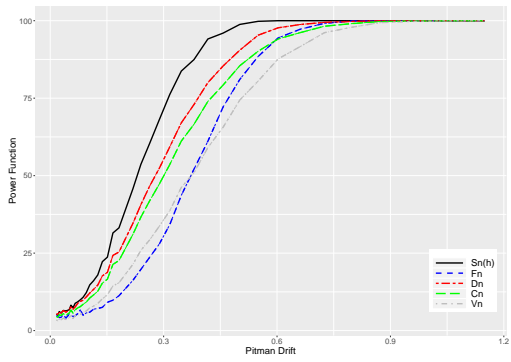
F_N is implemented by the `fan.test` routine of the R package `GoFKernel`. The package also provides a bandwidth choice method via the option `dpik` which is actually a direct plugin rule. The routine returns the bias corrected version of the test discussed in Li and Racine (2007), pages 380–381. V_N is implemented by the `dbEmpLikeGOF` routine of the R package with the same name. All R functions mentioned above have a `p.value` argument corresponding to finite samples which is used to determine acceptance or rejection of the hypothesis being tested. Particularly for $S_N(h)$ the cut-off points are determined by Theorem 1. The confidence level on all examples in this Section is set to $100(1 - \alpha)\% = 95\%$. All implementations use the Epanechnikov kernel.

The data used in the power simulations come from densities with a wide variety of interesting features such as multiple modes and asymmetry characteristics. For this reason various normal mixture (NM) densities from Marron and Wand (1994), which do not seem to be frequently used in the goodness-of-fit literature are utilized. Their shapes include skewed unimodal (NM2), bimodal (NM6), separated bimodal (NM7), asymmetric bimodal (NM8), trimodal (NM9) and asymmetric double claw (NM12) types of densities. At each level of the alternative, 10,000 replications are performed. In each replication, a random sample of size $N = 40$ is generated based on the specification of the alternative and then used in calculating simultaneously all test statistics. Their power functions are then approximated as percentage rejections of the null. The average of 100 power functions for each test, calculated in exactly the same manner is drawn on the figures.

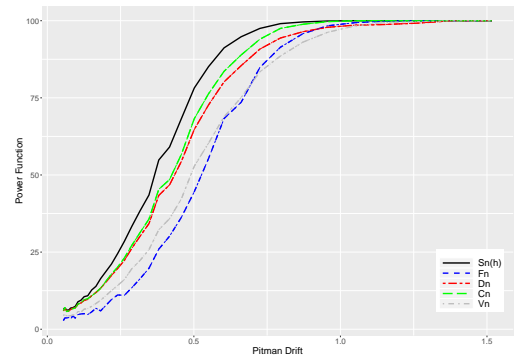
In accordance to the theoretical findings in Corollary 1, $S_N(h)$ outperforms F_N , D_N and C_N across all examples. Fig. 1 also visualizes the impact from modifying the definition of F_N to derive $S_N(h)$. Specifically, the distance between the two power curves demonstrates the significantly improved performance of $S_N(h)$ in comparison to F_N .

ΠΕΡΙΛΗΨΗ

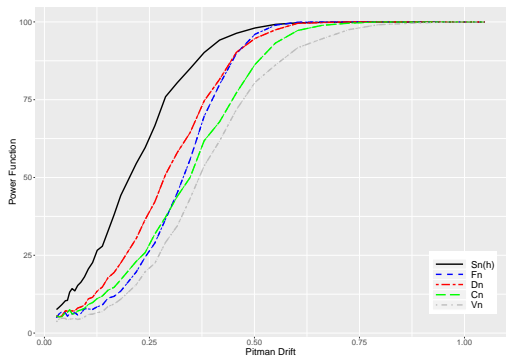
Εισάγουμε ένα καινούργιο τεστ καλής προσαρμογής το οποίο βασίζεται στην τμηματικά τοπική γραμμική



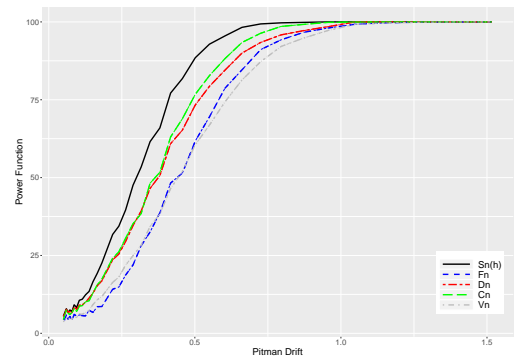
(a) NM2



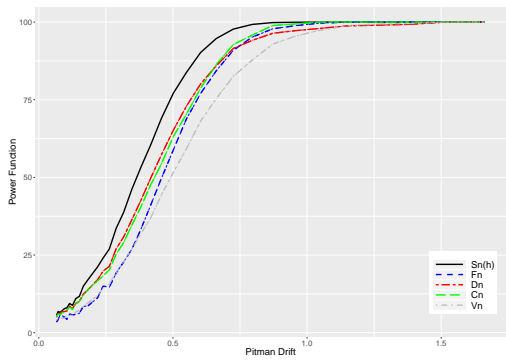
(b) NM6



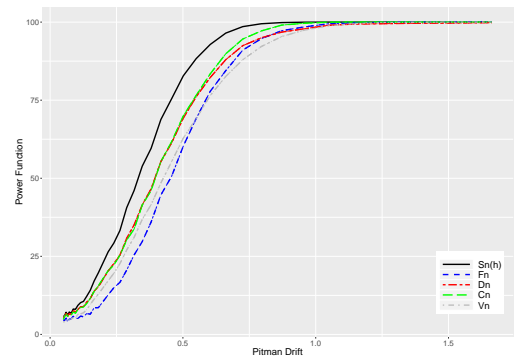
(c) NM7



(d) NM8



(e) NM9



(f) NM13

Figure 1: Empirical powers for $S_N(h)$, D_N , C_N , F_N and V_N in testing (1) with the alternative specified by $f_{1n}(x)$. $n = 40$, $\alpha = 5\%$.

προσέγγιση της L_2 μετρικής μεταξύ της συνάρτησης πυκνότητας πιθανότητας υπό τη μηδενική υπόθεση και της εκτίμησης αυτής. Μελετάμε την κατανομή του στατιστικού υπό τη μηδενική και την εναλλακτική υπόθεση καθώς και την ισχύ του για εναλλακτικές του τύπου Pitman.

REFERENCES

- Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **1**, 1071–1095.
- Cao, R. and Lugosi G. (2005). Goodness-of-fit tests based on the kernel density estimator. *Scand. Jour. Statist.* **32**, 599–616.
- Chebana, F. (2004). On the optimization of the weighted Bickel-Rosenblatt test. *Statist. Prob. Letters.* **68**, 333–345.
- M. Y. Cheng, (1994) *On boundary effects of smooth curve estimates*. PhD thesis, University of North Carolina at Chapel Hill.
- Fan, Y. (1994). Testing the goodness-of-fit of a parametric density function by kernel method. *Econ Theory* **10**, 316–356.
- Fan, J. and Gijbels, I. (1996). *Local polynomial modeling and its applications*. Chapman and Hall, London.
- Hall, P. and Wand, M. P., On the Accuracy of Binned Kernel Density Estimators *J. Mult. Anal.* **56**, 165–164.
- Gurevich G. and Vexler A. (2010). Empirical likelihood ratios applied to goodness-of-fit tests based on sample entropy. *Coputational Statist. Data Anal.* **54**, 531–545.
- González-Manteiga, W. and Crujeiras, R.M. (2013). An updated review of Goodness-of-Fit tests for regression models. *Test* **22**, 361–411.
- Gouriéroux, C. and Tenreiro, C. (2001). Local power properties of kernel based goodness of fit tests. *J. Multivar. Anal.* **78**, 161–190.
- Ghosh, B. K. and Huang, W. (1991). The power optimal kernel of the Bickel-Rosenblatt test for goodness of fit. *Ann. Statist.* **19**, 999–1009.
- Jager, L. and Wellner, J.A. (2007). Goodness-Of-Fit tests via Phi-divergences, *Ann. Statist.* **35**, 2018–2053.
- Lequesne, J. and Regnault, P. (2018). vsgoftest: An R Package for Goodness-of-Fit Testing Based on Kullback-Leibler Divergence, <https://arxiv.org/abs/1806.07244>.
- Li, O. and Racine, J.F. (2007). *Nonparametric Econometrics*. Princeton University Press, New Jersey.
- Marron, J.S. and Wand, M.P. (1994). Exact mean integrated squared error, *Ann. Statist.* **20**, 712–736.
- McCullagh, P. (1987). *Tensor Methods in Statistics*. Chapman and Hall: London.
- Sheather, S. J., and M. C. Jones. (1991). A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation. *Journal of the Royal Statistical Society. Series B* **53** pp. 683–690.
- Song, K.-S. (2002). Goodness-of-Fit Tests Based on Kullback–Leibler Discrimination Information. *IEEE Transactions On Information Theory* **48**, 1103–1117.
- Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall. New York.

Rosenblatt, M. (1975). A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *Ann. Statist.* **3**, 1–14.

A PROOF OF THEOREM 1

The proof is broken into the following four steps. First note that henceforth, \sim means that the ratio of the LHS to the RHS converges to 1 as $N \rightarrow \infty$.

Step 1. Show that for large N , $-E\{NS_N(h)\} \sim 2R(f)$, so that $E\{N\sqrt{h}S_N(h)\} \rightarrow 0$ as $N \rightarrow \infty$, using assumption (A1).

Step 2. Show that $\text{Var}\{N\sqrt{h}S_N(h)\} \sim 2R(f_0)$ as $N \rightarrow \infty$.

Step 3. Show that for each $r \geq 3$, $\mathcal{C}_r\{S'_N(h)\} = O(N^r h^{r-1})$.

Step 4. Under the assumptions of Theorem 1, $N\sqrt{h}S_N(h)\sigma_0^{-1} \rightarrow N(0, 1)$.

Proof of Step 1. Since (Y_1, \dots, Y_n) is $\text{Mult}_n(N, p)$,

$$E\{S_N(h)\} = -\frac{1}{N(N-1)h} \sum_{i \neq j} \sum I(|x_i - x_j| \leq h) N p_i p_j.$$

From the assumptions of Theorem 1, as $N \rightarrow \infty$,

$$p_1 = \int_{-\infty}^{x_1 + \Delta/2} f(x) dx \rightarrow 0, \quad p_n = \int_{x_n - \Delta/2}^{\infty} f(x) dx \rightarrow 0$$

and, for $2 \leq i \leq n-1$, $p_i \sim f(x_i)\Delta$. Also, by a simple counting argument, for all i except a negligible proportion close to 1 or n , $\#\{j: |x_i - x_j| \leq h\} \sim 2h/\Delta$. Therefore

$$\begin{aligned} \frac{1}{(N-1)h} \sum_{i \neq j} \sum I(|x_i - x_j| \leq h) p_i p_j &\sim \frac{1}{(N-1)h} \sum_{i=2}^{n-1} (2h/\Delta) f(x_i)^2 \Delta^2 \\ &= \frac{2}{N-1} \sum_{i=2}^{n-1} f(x_i)^2 \Delta \sim \frac{2}{N-1} \int_{-\infty}^{\infty} f(x)^2 dx, \end{aligned}$$

where in the final step we have approximated the Riemann sum by its integral, using assumption (A2). \square

Proof of Step 2. Direct calculation shows that

$$\text{Var}\{S_N(h)\} = \frac{1}{N^2(N-1)^2 h^2} \sum_{i_1 \neq j_1} \sum_{i_2 \neq j_2} \sum I(|x_{i_1} - x_{j_1}| \leq h) I(|x_{i_2} - x_{j_2}| \leq h) \rho^{i_1 j_1, i_2 j_2},$$

where

$$\rho^{i_1 j_1, i_2 j_2} = \text{Cov}\{(Y_{i_1} - N p_{i_1})(Y_{j_1} - N p_{j_1}), (Y_{i_2} - N p_{i_2})(Y_{j_2} - N p_{j_2})\}.$$

Expressing generalized cumulants in terms of ordinary cumulants (see (3.2) of McCullagh (1987), for this particular case), and noting that all the ordinary first cumulants, $C_1(Y_i)$, are zero, we obtain

$$\rho^{i_1 j_1, i_2 j_2} = \mathcal{C}(Y_{i_1}, Y_{j_1}, Y_{i_2}, Y_{j_2}) + \mathcal{C}(Y_{i_1}, Y_{i_2})\mathcal{C}(Y_{j_1}, Y_{j_2}) + \mathcal{C}(Y_{i_1}, Y_{j_2})\mathcal{C}(Y_{i_2}, Y_{j_1}). \quad (5)$$

Writing

$$\mathcal{C}(Y_{i_1}, Y_{i_2}) = -Np_{i_1}p_{i_2} + I(i_1 = i_2)Np_{i_1},$$

and using similar expressions involving indicator functions for the other cumulants on the right-hand side of (5), it is seen that

$$\begin{aligned} & \sum_{i_1 \neq j_1} \sum_{i_2 \neq j_2} \sum I(|x_{i_1} - x_{j_1}| \leq h)I(|x_{i_2} - x_{j_2}| \leq h) \{ \mathcal{C}(Y_{i_1}, Y_{i_2})\mathcal{C}(Y_{j_1}, Y_{j_2}) \\ & \quad + \mathcal{C}(Y_{i_1}, Y_{j_2})\mathcal{C}(Y_{i_2}, Y_{j_1}) \} \\ & = 2N^2 \sum_{i_1 \neq j_1} \sum_{i_2 \neq j_2} \sum I(|x_{i_1} - x_{j_1}| \leq h)I(|x_{i_2} - x_{j_2}| \leq h) p_{i_1} p_{i_2} p_{i_3} p_{i_4} \\ & \quad - 4N^2 \sum_{i_1 \neq j_1} \sum_{i_i \neq j_j} \sum I(|x_{i_1} - x_{j_1}| \leq h)I(|x_{i_1} - x_{j_2}| \leq h) p_{i_1} p_{j_1} p_{j_2} \\ & \quad + 2N^2 \sum_{i_1 \neq j_1} \sum I(|x_{i_1} - x_{j_1}| \leq h) p_{i_1} p_{i_2}. \quad (6) \end{aligned}$$

Using approximations similar to those in the proof of Step 1,

$$\begin{aligned} & \sum_{i_1 \neq j_1} \sum_{i_2 \neq j_2} \sum I(|x_{i_1} - x_{j_1}| \leq h)I(|x_{i_2} - x_{j_2}| \leq h) p_{i_1} p_{j_1} p_{i_2} p_{j_2} \\ & \quad \sim \sum_{i=2}^{n-1} \sum_{j=2}^{n-1} (2h/\Delta)^2 f(x_i)^2 \Delta^2 f(x_j)^2 \Delta^2 \sim 4h^2 \left(\int_{-\infty}^{\infty} f(x)^2 dx \right)^2; \quad (7) \end{aligned}$$

$$\begin{aligned} & \sum_{i_1 \neq j_1} \sum_{i_1 \neq j_2} \sum I(|x_{i_1} - x_{j_1}| \leq h)I(|x_{i_1} - x_{j_2}| \leq h) p_{i_1} p_{j_1} p_{j_2} \\ & \quad \sim \sum_{i=2}^{n-1} (2h/\Delta)^2 f(x_i)^3 \Delta^3 \sim 4h^2 \int_{-\infty}^{\infty} f(x)^3 dx; \quad (8) \end{aligned}$$

and

$$\sum_{i_1 \neq j_1} \sum I(|x_{i_1} - x_{j_1}| \leq h) p_{i_1} p_{j_1} \simeq \sum_{i=2}^{n-1} (2h/\Delta) f(x_i)^2 \Delta^2 \sim 2hR(f). \quad (9)$$

Moreover, by similar calculations,

$$\sum_{i_1 \neq j_1} \sum_{i_2 \neq j_2} \sum I(|x_{i_1} - x_{j_1}| \leq h)I(|x_{i_2} - x_{j_2}| \leq h) \mathcal{C}(Y_{i_1}, Y_{j_1}, Y_{i_2}, Y_{j_2}) = O(h). \quad (10)$$

Consequently, since each term on the RHS of (6) is multiplied by N^2 , it follows that (7), (8) and (9) should also be multiplied by N^2 , while (10) should be multiplied by N . Then finally to compute variance of $S_N(h)$ we divide each of these terms by $[N(N-1)h]^2$. Therefore (9) is the dominant term and $\text{Var}\{S_N(h)\} \sim 2N^{-2}h^{-1}R(f_0)$.

Proof of Step 4. From Step 1 and Step 2 respectively it is seen that as $N \rightarrow \infty$

$$E[S_N] \rightarrow 0 \text{ and } \text{Var} \left(\frac{N\sqrt{h}}{\sigma_0} S_N \right) \rightarrow 1.$$

Assuming that Step 3 holds for $r \geq 3$,

$$C_r(N\sqrt{h}S_N) = (N\sqrt{h})^r \times O \left\{ \frac{N^r h^{r-1}}{N^{2r} h^r} \right\} = O(h^{(r/2)-1}) \text{ as } N \rightarrow \infty,$$

which converges to 0 because $h \rightarrow 0$.

Proof of Step 3. It follows from the multilinearity property of cumulants that the r th cumulant of $S'_N(h)$ is given by

$$C_r\{S'_N(h)\} = \sum_{i_1 \neq j_1} \sum \cdots \sum_{i_r \neq j_r} \left\{ \prod_{k=1}^r I(|x_{i_k} - x_{j_k}| \leq h) \right\} \rho^{i_1 j_1, \dots, i_r j_r}, \quad (11)$$

where

$$\rho^{i_1 j_1, \dots, i_r j_r} = \mathcal{C}\{(Y_{i_1} - Np_{i_1})(Y_{j_1} - Np_{j_1}), \dots, (Y_{i_r} - Np_{i_r})(Y_{j_r} - Np_{j_r})\} \quad (12)$$

is the joint cumulant of $(Y_{i_1} - Np_{i_1})(Y_{j_1} - Np_{j_1}), \dots, (Y_{i_r} - Np_{i_r})(Y_{j_r} - Np_{j_r})$. Each $\rho^{i_1 j_1, \dots, i_r j_r}$ is a generalized cumulant which, using Theorem 3, can be expressed as a sum of the ordinary cumulants of $Y_{i_1}, Y_{j_1}, \dots, Y_{i_r}, Y_{j_r}$. However, there is a complication: as in the derivation of the asymptotic variance of $S'_N(h)$ given in the previous section, we need to take account of any equalities among the indices $i_1, j_1, \dots, i_r, j_r$ that occur (though recall that we always have $i_1 \neq j_1, \dots, i_r \neq j_r$). In fact, as in the case of the asymptotic variance, the leading terms are those with the maximum number of equalities.

Write $\Upsilon^* = \{v_1^*, \dots, v_r^*\}$ where $v_k^* = \{i_k, j_k\}$, $1 \leq k \leq r$, and let $\Upsilon = \{v_1, \dots, v_A\}$ denote any partition of $\{i_1, j_1, \dots, i_r, j_r\}$ complementary to Υ^* (in the notation of Theorem 3, $\Upsilon \vee \Upsilon^* = 1$). Then, from Theorem 3,

$$\rho^{i_1 j_1, \dots, i_r j_r} = \sum_{\Upsilon: \Upsilon \vee \Upsilon^* = 1} \mathcal{C}(v_1) \cdots \mathcal{C}(v_A).$$

Substituting into (11), we obtain

$$\begin{aligned} C_r\{S'_N(h)\} = & \sum_{i_1 \neq j_1} \sum \cdots \sum_{i_r \neq j_r} \left\{ \prod_{k=1}^r I(|x_{i_k} - x_{j_k}| \leq h) \right\} \sum_{\Upsilon: \Upsilon \vee \Upsilon^* = 1} \mathcal{C}(v_1) \cdots \mathcal{C}(v_A) = \\ & \sum_{\Upsilon: \Upsilon \vee \Upsilon^* = 1} \sum_{i_1 \neq j_1} \sum \cdots \sum_{i_r \neq j_r} \left\{ \prod_{k=1}^r I(|x_{i_k} - x_{j_k}| \leq h) \right\} \mathcal{C}(v_1) \cdots \mathcal{C}(v_A). \quad (13) \end{aligned}$$

Within each v_i we need to allow for the possibility that some of the indices are equal. Let us focus on the general case: for each i , let $v_i = \bigcup_{j=1}^{d_i} v_{ij}$, where $v_{ij} \cap v_{ik} = \emptyset$ if $j \neq k$; all indices within any v_{ij} are constrained to be equal to a common value, a_{ij} say; and for any $j \neq k$, no indices in v_{ij} are constrained to be equal to any indices in v_{ik} . Then, using Lemma 1, we have the bound

$$|\mathcal{C}(v_1) \dots \mathcal{C}(v_A)| \leq D_r N^A \prod_{i=1}^A \prod_{j=1}^{d_j} p_{a_{ij}},$$

where $D_r \geq 1$ depends only on r . In considering (13), it is helpful to define a graph Γ with vertex set $\{v_{ij} : j = 1, \dots, d_i, i = 1, \dots, A\}$ and edge structure defined as follows:

v_{lm} is connected to v_{pq} if and only if for some k , either $i_k \in v_{lm}$ and $j_k \in v_{pq}$, or $j_k \in v_{lm}$ and $i_k \in v_{pq}$.

Note that, by construction, Γ has no singleton components. Suppose that Γ has T connected components $\Gamma_1, \dots, \Gamma_T$ with m_t vertices in Γ_t , $1 \leq t \leq T$. Since Γ is such that i_k must always be connected to j_k , it follows that

$$\bigcup_{v: v \in \Gamma_t} v = \{i_{k_l}, j_{k_l} : l \in B_t\}$$

for some $B_t \subseteq \{1, \dots, r\}$. Moreover, B_1, \dots, B_T is a partition of $\{1, \dots, r\}$. It follows that, for this Υ with the assumed equality constraints among the indices,

$$\sum_{i_1 \neq j_1} \sum \dots \sum_{i_r \neq j_r} \left\{ \prod_{k=1}^r I(|x_{i_k} - x_{j_k}| \leq h) \right\} |\mathcal{C}(v_1) \dots \mathcal{C}(v_A)| = N^A \prod_{t=1}^T J_t, \quad (14)$$

where

$$J_T \leq D_r \sum_{i_k \neq j_k, k \in B_t}^* \left\{ \prod_{k \in B_t} I(|x_{i_k} - x_{j_k}| \leq h) \right\} \prod_{v: v \in \Gamma_t} p_v,$$

the summation $\sum_{i_k \neq j_k, k \in B_t}^*$ means summation over all $i_k \neq j_k$ for $k \in B_t$, but taking any equality constraints among the indices i_k, j_k , $k \in B_t$, into account; and with slight abuse of notation we have written $\prod_{v: v \in \Gamma_t} p_v$ for the product of the p 's indexed by the common value of the indices in each $v \in \Gamma_t$.

Now, following the calculation in the previous section,

$$J_t \leq D_r \sum_{i=1}^n (2c_r h / \Delta)^{m_t-1} (f(x_i) \Delta)^{m_t} \sim D_r (2c_r h)^{m_t-1} \int_{-\infty}^{\infty} f(x)^{m_t} dx,$$

where c_r is a finite constant depending only on r . Consequently, the LHS of (14) is bounded by

$$G_r(f) h^{(\sum_{t=1}^T m_t) - T} N^A,$$

where $G_r(f) > \infty$ depends only on r and f . The proof of Step 3 will be complete if we can show that

$$\left(\sum_{t=1}^T m_t \right) - T \geq A - 1 \quad (15)$$

because then

$$\begin{aligned} N^A h^{(\sum_{t=1}^T m_t) - T} &\leq N^A h^{A-1} \quad (\text{since } h \rightarrow 0) \\ &\leq N^r h^{r-1} \quad (\text{since } A \leq r \text{ and } Nh \rightarrow \infty) \end{aligned}$$

and in (13) the sum over partitions Υ and the sum over different configurations of equality constraints both only have a finite number of terms.

Proof of (15). First of all, note that

$$\sum_{t=1}^T m_t = \sum_{i=1}^A d_i, \tag{16}$$

the common value being the number of distinct indices when equality constraints are taken into account. When $d_1 = \dots = d_A = 1$, so that all the indices in each v_i , $1 \leq i \leq A$, are constrained to be equal, we must have $T = 1$. Otherwise, the requirement that Υ is complementary to Υ^* would be violated. In this case $(\sum_{t=1}^T m_t) - T = (\sum_{i=1}^A 1) - 1 = A - 1$. We can increase the number of components T by removing edges. But removing an edge corresponds to breaking at least one equality constraint, which means increasing at least one d_i by 1, which, by (16), corresponds to increasing at least one m_t by 1. Therefore for any partition Υ complementary to Υ^* , and for any configuration of equality constraints within each v_i , (15) must hold.



A THREE-LEVEL HIERARCHICAL CREDIBILITY REGRESSION APPROACH TO MODELLING MULTI-POPULATION MORTALITY DATA

Apostolos Bozikas^{1,2}, Georgios Pitselis²

1. Department of Economics, National and Kapodistrian University of Athens

2. Department of Statistics and Insurance Science, University of Piraeus

{bozikas,pitselis}@unipi.gr

ABSTRACT

In this work, we present a three-level hierarchical credibility regression approach to modelling and forecasting the mortality rates for multiple populations. The credibility estimation is described and the forecasting performances between the proposed model, the original Lee-Carter model and two extensions of the Lee-Carter model for multiple populations are compared, per gender and age, for three northern European countries. Numerical illustrations indicate that the proposed model contributes to more accurate forecasts than those produced by the Lee-Carter model and its multi-population extensions.

Keywords: Hierarchical credibility regression, multi-population mortality models, Lee-Carter.

1. INTRODUCTION

Mortality has significantly declined in most developed countries around the world, mainly due to the continuous improvement of living conditions and the evolution of medical science and technology. Eventually, the decline in mortality creates higher financial responsibilities for governments and annuity providers. Consequently, finding ways to manage the mortality dynamics of a population is a very important step in building a sustainable health and pension system.

Several methods have been proposed in order to model the mortality trends of a population. Lee and Carter (1992) developed a pioneer method to forecast the mortality of the total population in the United States. Cairns et al. (2006) proposed the Cairns-Blake-Dowd (CBD) model, particularly designed for higher ages. Furthermore, Renshaw and Haberman (2006) extended the Lee-Carter model by including a cohort effect, while Plat (2009) proposed a model which combines preferable characteristics of the Lee-Carter and the CBD models.

Credibility theory can serve as an alternative tool in mortality modelling. With applications mainly in non-life insurance, credibility techniques should also be used

in order model the period patterns of mortality data for a specific age, using information from a wider age span. Bühlmann (1967) established the theoretical foundation of modern credibility theory (also known as greatest accuracy credibility theory). Bühlmann and Straub (1970) extended the Bühlmann model by allowing unequal number of exposure units for each risk and Hachemeister (1975) introduced a credibility regression model to estimate auto-mobile bodily injury claims for various states in the USA. For an extensive review on credibility theory for non-life insurance, we also refer to Goovaerts et al. (1990) and Bühlmann and Gisler (2005).

Regarding some life insurance applications of credibility theory, Schinzinger et al. (2016) presented a multivariate evolutionary credibility model for mortality improvement rates to describe the joint dynamics of mortality through time in several populations. Tsai and Lin (2017) proposed a Bühlmann credibility mortality modelling approach and evaluated its forecasting performance, while Bozikas and Pitselis (2019) described how credibility regression can be used to model the mortality trends of a population, especially in cases where there is limited mortality experience for a specific age, but extensive experience for a wider age range.

Even if credibility regression can yield desirable results when it is applied on mortality rates of a single population, a question arises when we deal with datasets of multiple populations. Many developments in multi-population mortality modelling have been appeared in the literature. The most used extensions of the Lee-Carter model for multiple populations are the Carter and Lee (1992) method, which applies a common time-varying index to all populations and the Li and Lee (2005) method, which proposes a two-step procedure to model the mortality dynamics for multiple populations.

Alternatively, hierarchical credibility can accommodate classification schemes, where multi-population mortality data are represented in a multi-level structure. In the literature, we can find many works related with hierarchical credibility, particularly for non-life insurance applications. Jewell (1975) extended Bühlmann credibility into a hierarchical model, while Sundt (1979, 1980) generalized it into a hierarchical credibility regression model on two and multiple levels, respectively.

Based on hierarchical credibility, Tsai and Zhang (2019) presented a multi-dimensional Bühlmann credibility approach to model mortality rates for multiple populations. This work aims to extend the credibility regression mortality approach of Bozikas and Pitselis (2019) to its multi-level hierarchical counterpart, giving a detailed description for the estimation of model parameters and the derivation of mortality forecasts for multiple countries.

This work is organized as follows. Section 2 presents the three-level hierarchical credibility mortality model, provided with a detailed description of parameters estimation. Section 3 illustrates the proposed credibility model on male and female mortality data of three countries in northern Europe with small populations, the Ireland, the Norway and the Finland. Section 4 concludes this work.

2.A THREE-LEVEL HIERARCHICAL CREDIBILITY REGRESSION MORTALITY MODEL

Here, we present a three-level credibility regression approach to model mortality for multiple populations in a hierarchical form. Beginning from the bottom, mortality data are first classified by age (Level 1). Ages are then grouped by gender (Level 2) and genders by country (Level 3) to constitute a multi-country credibility structure. Each level is also associated with a random risk variable that reflects specific characteristics for each country, gender and age. More precisely, country level is reflected by a random variable Θ_c , $c = 1, \dots, C$, gender level is reflected by Θ_{gc} , $g = 1, \dots, G_c$, and age level by Θ_{xgc} , $x = x_1, \dots, x_{k_{gc}}$.

Let us denote $D_{t,xgc}$ as the observed number of deaths and $E_{t,xgc}$ as the average population (exposure to risk) in calendar year $t = t_1, \dots, t_n$, for age $x = x_1, \dots, x_k$, gender $g = 1, \dots, G$ ($G = 2$, i. e., male = 1, female = 2) and country $c = 1, \dots, C$. Then, the corresponding age-specific mortality rates $m_{t,xgc}$ are obtained by the ratio $D_{t,xgc}/E_{t,xgc}$.

Now assume that the observable variable $Y_{t,xgc}$ corresponds to the log transform of mortality rates $m_{t,x}$, for $t = t_1, \dots, t_n$, $x = x_1, \dots, x_{k_{gc}}$, $g = 1, \dots, G_c$ and $c = 1, \dots, C$. Also, denote $\mathbf{Y}_{xgc} = (Y_{t_1,xgc}, \dots, Y_{t_n,xgc})'$ as the mortality vector for age x , gender g , country c and \mathbf{Z}_{xgc} as the corresponding design matrix of explanatory variables. We note that in general, design matrix could consist of various explanatory variables that reflect mortality characteristics. Due to the lack of other explanatory variables, we assume that for each age, gender and country, mortality rates are only affected by time trends. Thus, for all x, g and c the observable variables and the design matrices for Level 1 (age), Level 2 (gender) and Level 3 (country) are denoted as

$$\mathbf{Y}_{xgc} = (\log m_{t_1,xgc}, \dots, \log m_{t_n,xgc})', \mathbf{Z}_{xgc} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 2 & \dots & n \end{pmatrix}', \text{(Level1)} \quad (1)$$

$$\mathbf{Y}_{gc} = (\mathbf{Y}'_{x_1gc}, \dots, \mathbf{Y}'_{x_{k_{gc}}gc})', \mathbf{Z}_{gc} = (\mathbf{Z}'_{x_1gc}, \dots, \mathbf{Z}'_{x_{k_{gc}}gc})' \text{(Level2)} \quad (2)$$

and

$$\mathbf{Y}_c = (\mathbf{Y}'_{1c}, \dots, \mathbf{Y}'_{G_c c})', \mathbf{Z}_c = (\mathbf{Z}'_{1c}, \dots, \mathbf{Z}'_{G_c c})'. \text{(Level3)} \quad (3)$$

Then, country c consists of the set variables $\{\Theta_{xgc}, \Theta_{gc}, \Theta_c, \mathbf{Y}_c\}$, for $x = x_1, \dots, x_{k_{gc}}$ and $g = 1, \dots, G_c$, gender g consists of the variables $\{\Theta_{xgc}, \Theta_{gc}, \mathbf{Y}_{gc}\}$, for $x = x_1, \dots, x_{k_{gc}}$ and age x consists of $\{\Theta_{xgc}, \mathbf{Y}_{xgc}\}$.

2.1 Notation and Assumptions

Henceforth we denote $(\Theta_{xgc}, \Theta_{gc}, \Theta_c) = \Theta_{xgc}^{123}$, $(\Theta_{gc}, \Theta_c) = \Theta_{gc}^{23}$ and $\Theta_c = \Theta_c^3$ under the following assumptions:

1. The countries are independent, i.e., $\{\Theta_{xgc}, \Theta_{gc}, \Theta_c, \mathbf{Y}_c\}$ is independent of $\{\Theta_{x'g'c'}, \Theta_{g'c'}, \Theta_{c'}, \mathbf{Y}_{c'}\}$.
2. For each $c = 1, \dots, C$ and for given Θ_c^3 , the genders $\{\Theta_{xgc}, \Theta_{gc}, \mathbf{Y}_{gc}\}$ are conditionally independent.
3. For each $c = 1, \dots, C$ and $g = 1, \dots, G_c$ and for given Θ_{gc}^{23} , the ages $\{\Theta_{xgc}, \mathbf{Y}_{xgc}\}$ are conditionally independent.
4. For each $c = 1, \dots, C$, $g = 1, \dots, G_c$, $x = x_1, \dots, x_{k_{gc}}$ and for given Θ_{xgc}^{123} , mortality observations \mathbf{Y}_{xgc} are conditionally independent.
5. All the vectors Θ_{xgc}^{123} for $c = 1, \dots, C$, $g = 1, \dots, G_c$ and $x = x_1, \dots, x_{k_{gc}}$ are identically distributed.
6. $E(\mathbf{Y}_{xgc} | \Theta_{xgc}^{123}) = \mathbf{Z}_{xgc} (\Theta_{xgc}^{123})$, for $t = t_1, \dots, t_n$, where \mathbf{Z}_{xgc} is a $n \times 2$ design matrix and (Θ_{xgc}^{123}) is the coefficients vector.
7. $Cov(\mathbf{Y}_{xgc} | \Theta_{xgc}^{123}) = \sigma^2 (\Theta_{xgc}^{123}) \mathbf{W}_{xgc}^{-1}$, where \mathbf{W}_{xgc} is a positive definite $(n \times n)$ matrix.

2.2 Structural Parameters

The structural parameters of the hierarchical credibility regression model are defined as

- a) $\boldsymbol{\beta} = E[\boldsymbol{\beta}(\Theta_{xgc}^{123})]$,
- b) $s^2 = E[\sigma^2(\Theta_{xgc}^{123})]$,
- c) $\mathbf{A}_1 = E\{Cov[\boldsymbol{\beta}(\Theta_{xgc}^{123}) | \Theta_{gc}^{23}]\}$,
- d) $\mathbf{A}_2 = E\{Cov[\boldsymbol{\beta}(\Theta_{gc}^{23}) | \Theta_c^3]\}$,
- e) $\mathbf{A}_3 = Cov[\boldsymbol{\beta}(\Theta_c^3)]$.

The regression coefficients for the first level are given by

$$\widehat{\boldsymbol{\beta}}_{xgc} = (\mathbf{Z}'_{xgc} \mathbf{W}_{xgc} \mathbf{Z}_{xgc})^{-1} \mathbf{Z}'_{xgc} \mathbf{W}_{xgc} \mathbf{Y}_{xgc}, \text{ (Level1)} \quad (4)$$

with conditional covariance

$$Cov(\widehat{\boldsymbol{\beta}}_{xgc} | \Theta_{xgc}^{123}) = \sigma^2 (\Theta_{xgc}^{123}) (\mathbf{Z}'_{xgc} \mathbf{W}_{xgc} \mathbf{Z}_{xgc})^{-1}, \quad (5)$$

while for the other levels by

$$\widehat{\boldsymbol{\beta}}_{gc} = (\sum_{x=x_1}^{x_{k_{gc}}} \mathbf{K}_{xgc})^{-1} \sum_{x=x_1}^{x_{k_{gc}}} \mathbf{K}_{xgc} \widehat{\boldsymbol{\beta}}_{xgc}, \text{ (Level2)} \quad (6)$$

$$\widehat{\boldsymbol{\beta}}_c = (\sum_{g=1}^{G_c} \mathbf{K}_{gc})^{-1} \sum_{g=1}^{G_c} \mathbf{K}_{gc} \widehat{\boldsymbol{\beta}}_{gc} \text{ (Level3)} \quad (7)$$

and

$$\widehat{\boldsymbol{\beta}} = (\sum_{c=1}^C \mathbf{K}_{gc})^{-1} \sum_{c=1}^C \mathbf{K}_{gc} \widehat{\boldsymbol{\beta}}_c. \quad (8)$$

The corresponding credibility factors for each level are respectively defined by

$$\mathbf{K}_{xgc} = \mathbf{A}_1 [\mathbf{A}_1 + s^2 (\mathbf{Z}'_{xgc} \mathbf{W}_{xgc} \mathbf{Z}_{xgc})^{-1}]^{-1}, \quad (\text{Level1})(9)$$

$$\mathbf{K}_{gc} = \mathbf{A}_2 [\mathbf{A}_2 + (\sum_{x=x_1}^{x_{k_{gc}}} \mathbf{K}_{xgc})^{-1} \mathbf{A}_1]^{-1} \quad (\text{Level2})(10)$$

and

$$\mathbf{K}_c = \mathbf{A}_3 [\mathbf{A}_3 + (\sum_{g=1}^{G_c} \mathbf{K}_{gc})^{-1} \mathbf{A}_2]^{-1}. \quad (\text{Level3})(11)$$

The following theorems present the conditional expectations and covariances of regression coefficients. For the proofs, we refer to Bozikas (2019).

Theorem 1. Based on the above assumptions and notation, we obtain the following expressions for the conditional expectations

$$E[\boldsymbol{\beta}(\theta_{xgc}^{123}) | \theta_{gc}^{23}] = \boldsymbol{\beta}(\theta_{gc}^{23}), \quad (12)$$

$$E[\boldsymbol{\beta}(\theta_{xgc}^{123}) | \theta_c^3] = \boldsymbol{\beta}(\theta_c^3), \quad (13)$$

$$E[\boldsymbol{\beta}(\theta_{gc}^{23}) | \theta_c^3] = \boldsymbol{\beta}(\theta_c^3), \quad (14)$$

$$E[\boldsymbol{\beta}(\theta_c^3) | \theta_c^3] = \boldsymbol{\beta}(\theta_c^3), \quad (15)$$

$$E[\boldsymbol{\beta}(\theta_c^3)] = \boldsymbol{\beta}, \quad (16)$$

$$E[\widehat{\boldsymbol{\beta}}_{xgc} | \theta_{xgc}^{123}] = \boldsymbol{\beta}(\theta_{xgc}^{123}), \quad (17)$$

$$E[\widehat{\boldsymbol{\beta}}_{xgc} | \theta_{gc}^{23}] = \boldsymbol{\beta}(\theta_{gc}^{23}), \quad (18)$$

$$E[\widehat{\boldsymbol{\beta}}_{xgc} | \theta_c^3] = \boldsymbol{\beta}(\theta_c^3), \quad (19)$$

$$E(\widehat{\boldsymbol{\beta}}_{xgc}) = E(\widehat{\boldsymbol{\beta}}_{gc}) = E(\widehat{\boldsymbol{\beta}}_c) = \boldsymbol{\beta}. \quad (20)$$

Theorem 2. Based on Theorem 2.1, we can prove the following expressions for the conditional covariances

$$\text{Cov}[\boldsymbol{\beta}(\theta_{xgc}^{123}), \widehat{\boldsymbol{\beta}}_{x'g'c'}] = \delta_{cc'} [\delta_{gg'} (\delta_{xx'} \mathbf{A}_1 + \mathbf{A}_2) + \mathbf{A}_3], \quad (21)$$

$$\text{Cov}[\boldsymbol{\beta}(\theta_{gc}^{23}), \widehat{\boldsymbol{\beta}}_{x'g'c'}] = \delta_{cc'} (\delta_{gg'} \mathbf{A}_2 + \mathbf{A}_3), \quad (22)$$

$$\text{Cov}[\boldsymbol{\beta}(\theta_c^3), \widehat{\boldsymbol{\beta}}_{x'g'c'}] = \delta_{cc'} \mathbf{A}_3, \quad (23)$$

$$\text{Cov}(\widehat{\boldsymbol{\beta}}_{xgc}, \widehat{\boldsymbol{\beta}}_{x'g'c'}) = \delta_{cc'} \{ \delta_{gg'} [\delta_{xx'} (\mathbf{A}_1 + s^2 (\mathbf{Z}'_{xgc} \mathbf{W}_{xgc} \mathbf{Z}_{xgc})^{-1}) + \mathbf{A}_2] + \mathbf{A}_3 \}, \quad (24)$$

$$\text{Cov}(\widehat{\boldsymbol{\beta}}_{gc}, \widehat{\boldsymbol{\beta}}_{x'g'c'}) = \delta_{cc'} \{ \delta_{gg'} [(\sum_{x=x_1}^{x_{k_{gc}}} \mathbf{K}_{xgc})^{-1} \mathbf{A}_1 + \mathbf{A}_2] + \mathbf{A}_3 \}, \quad (25)$$

$$\text{Cov}(\widehat{\boldsymbol{\beta}}_{xgc}, \widehat{\boldsymbol{\beta}}_{g'c'}) = \delta_{cc'} \{ \delta_{gg'} [\mathbf{A}_1 (\sum_{x'=x_1}^{x_{k_{gc}}} \mathbf{K}'_{x'g'c'})^{-1} + \mathbf{A}_2] + \mathbf{A}_3 \}, \quad (26)$$

$$\text{Cov}(\widehat{\boldsymbol{\beta}}_{gc}, \widehat{\boldsymbol{\beta}}_{g'c'}) = \delta_{cc'} \{ \delta_{gg'} [\mathbf{A}_1 (\sum_{x'=x_1}^{x_{k_{gc}}} \mathbf{K}'_{x'g'c'})^{-1} + \mathbf{A}_2] + \mathbf{A}_3 \}, \quad (27)$$

$$Cov(\widehat{\boldsymbol{\beta}}_{g'c'}, \widehat{\boldsymbol{\beta}}_{gc}) = \delta_{cc'} \{ \delta_{gg'} [(\sum_{x=x_1}^{x_k} \mathbf{K}_{xgc})^{-1} \mathbf{A}_1 + \mathbf{A}_2] + \mathbf{A}_3 \}, \quad (28)$$

$$Cov(\widehat{\boldsymbol{\beta}}_c, \widehat{\boldsymbol{\beta}}_{x'g'c'}) = \delta_{cc'} [(\sum_{g=1}^{G_c} \mathbf{K}_{gc})^{-1} \mathbf{A}_2 + \mathbf{A}_3], \quad (29)$$

$$Cov(\widehat{\boldsymbol{\beta}}_{xgc}, \widehat{\boldsymbol{\beta}}_{c'}) = \delta_{cc'} [\mathbf{A}_2 (\sum_{g=1}^{G_c} \mathbf{K}'_{g'c'})^{-1} + \mathbf{A}_3], \quad (30)$$

$$Cov(\widehat{\boldsymbol{\beta}}_c, \widehat{\boldsymbol{\beta}}_{g'c'}) = \delta_{cc'} [(\sum_{g=1}^{G_c} \mathbf{K}_{gc})^{-1} \mathbf{A}_2 + \mathbf{A}_3], \quad (31)$$

$$Cov(\widehat{\boldsymbol{\beta}}_{gc}, \widehat{\boldsymbol{\beta}}_{c'}) = \delta_{cc'} [\mathbf{A}_2 (\sum_{g=1}^{G_c} \mathbf{K}'_{g'c'})^{-1} + \mathbf{A}_3], \quad (32)$$

$$Cov(\widehat{\boldsymbol{\beta}}_c, \widehat{\boldsymbol{\beta}}_{c'}) = \delta_{cc'} [\mathbf{A}_2 (\sum_{g=1}^{G_c} \mathbf{K}'_{g'c'})^{-1} + \mathbf{A}_3], \quad (33)$$

$$Cov(\widehat{\boldsymbol{\beta}}_{c'}, \widehat{\boldsymbol{\beta}}_c) = \delta_{cc'} [(\sum_{c=1}^{G_c} \mathbf{K}_{gc})^{-1} \mathbf{A}_2 + \mathbf{A}_3], \quad (34)$$

$$Cov(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}_{c'}) = (\sum_{c=1}^C \mathbf{K}_c)^{-1} \mathbf{A}_3, \quad (35)$$

$$Cov(\widehat{\boldsymbol{\beta}}_c, \widehat{\boldsymbol{\beta}}) = \mathbf{A}_3 (\sum_{c=1}^C \mathbf{K}'_{c'})^{-1}, \quad (36)$$

$$Cov(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}) = (\sum_{c=1}^C \mathbf{K}_c)^{-1} \mathbf{A}_3. \quad (37)$$

2.3 Estimation of Model Parameters

Here, we present the linear credibility estimators for each level (country-gender-age) of our hierarchical regression model. For Level 3 (country), we consider

$$\mathbf{Y}_c^{Cred} = \mathbf{Z}_c \boldsymbol{\beta}_c^{Cred} (\boldsymbol{\theta}_c^3), \quad (38)$$

where $\mathbf{Y}_c^{Cred} = (\mathbf{Y}_1^{Cred'}, \dots, \mathbf{Y}_c^{Cred'})'$ and

$$\boldsymbol{\beta}_c^{Cred} (\boldsymbol{\theta}_c^3) = \mathbf{K}_c \widehat{\boldsymbol{\beta}}_c + (\mathbf{I} - \mathbf{K}_c) \boldsymbol{\beta}, \quad (39)$$

with \mathbf{K}_c defined by (11) and \mathbf{I} being the 2×2 identity matrix. For Level 2 (gender), we have

$$\mathbf{Y}_{gc}^{Cred} = \mathbf{Z}_{gc} \boldsymbol{\beta}_{gc}^{Cred} (\boldsymbol{\theta}_{gc}^{23}), \quad (40)$$

where $\mathbf{Y}_{gc}^{Cred} = (\mathbf{Y}_{1c}^{Cred'}, \dots, \mathbf{Y}_{G_c c}^{Cred'})'$ and

$$\boldsymbol{\beta}_{gc}^{Cred} (\boldsymbol{\theta}_{gc}^{23}) = \mathbf{K}_{gc} \widehat{\boldsymbol{\beta}}_{gc} + (\mathbf{I} - \mathbf{K}_{gc}) \boldsymbol{\beta}_c^{Cred} (\boldsymbol{\theta}_c^3), \quad (41)$$

with \mathbf{K}_{gc} obtained from (10). For Level 1 (age), the estimator is defined by

$$\mathbf{Y}_{xgc}^{Cred} = \mathbf{Z}_{xgc} \boldsymbol{\beta}_{xgc}^{Cred} (\boldsymbol{\theta}_{xgc}^{123}), \quad (42)$$

where $\mathbf{Y}_{xgc}^{Cred} = (\mathbf{Y}_{t_1, xgc}^{Cred}, \dots, \mathbf{Y}_{t_n, xgc}^{Cred})'$ and

$$\boldsymbol{\beta}_{xgc}^{Cred} (\boldsymbol{\theta}_{xgc}^{123}) = \mathbf{K}_{xgc} \widehat{\boldsymbol{\beta}}_{xgc} + (\mathbf{I} - \mathbf{K}_{xgc}) \boldsymbol{\beta}_{gc}^{Cred} (\boldsymbol{\theta}_{gc}^{23}), \quad (43)$$

with \mathbf{K}_{xgc} as it is given by (9). The next theorem reviews the optimal projection results, which imply the best linear estimators in hierarchical credibility regression [see Bühlmann and Gisler (2005), De Vylder (1976) and Sundt (1979, 1980)].

Theorem 3. Let $\boldsymbol{\beta}_c^{Cred}(\Theta_c^3)$ be the linear estimator of $\boldsymbol{\beta}(\Theta_c^3)$, $\boldsymbol{\beta}_{gc}^{Cred}(\Theta_{gc}^{23})$ be the linear estimator of $\boldsymbol{\beta}(\Theta_{gc}^{23})$ and $\boldsymbol{\beta}_{xgc}^{Cred}(\Theta_{xgc}^{123})$ be the linear estimator of $\boldsymbol{\beta}(\Theta_{xgc}^{123})$. Then, a) $\boldsymbol{\beta}_c^{Cred}(\Theta_c^3)$ is the best linear credibility estimator of $\boldsymbol{\beta}(\Theta_c^3)$ if it satisfies

$$E[\boldsymbol{\beta}_c^{Cred}(\Theta_c^3)] = E[\boldsymbol{\beta}(\Theta_c^3)] \quad (44)$$

and

$$Cov[\boldsymbol{\beta}_c^{Cred}(\Theta_c^3), \widehat{\boldsymbol{\beta}}_{x'g'c'}] = Cov[\boldsymbol{\beta}(\Theta_c^3), \widehat{\boldsymbol{\beta}}_{x'g'c'}], \quad (45)$$

b) $\boldsymbol{\beta}_{gc}^{Cred}(\Theta_{gc}^{23})$ is the best linear credibility estimator of $\boldsymbol{\beta}(\Theta_{gc}^{23})$ if it satisfies

$$E[\boldsymbol{\beta}_{gc}^{Cred}(\Theta_{gc}^{23})] = E[\boldsymbol{\beta}(\Theta_{gc}^{23})] \quad (46)$$

and

$$Cov[\boldsymbol{\beta}_{gc}^{Cred}(\Theta_{gc}^{23}), \widehat{\boldsymbol{\beta}}_{x'g'c'}] = Cov[\boldsymbol{\beta}(\Theta_{gc}^{23}), \widehat{\boldsymbol{\beta}}_{x'g'c'}], \quad (47)$$

c) $\boldsymbol{\beta}_{xgc}^{Cred}(\Theta_{xgc}^{123})$ is the best linear credibility estimator of $\boldsymbol{\beta}(\Theta_{xgc}^{123})$ if it satisfies

$$E[\boldsymbol{\beta}_{xgc}^{Cred}(\Theta_{xgc}^{123})] = E[\boldsymbol{\beta}(\Theta_{xgc}^{123})] \quad (48)$$

and

$$Cov[\boldsymbol{\beta}_{xgc}^{Cred}(\Theta_{xgc}^{123}), \widehat{\boldsymbol{\beta}}_{x'g'c'}] = Cov[\boldsymbol{\beta}(\Theta_{xgc}^{123}), \widehat{\boldsymbol{\beta}}_{x'g'c'}]. \quad (49)$$

The best linear credibility estimators of regression coefficients for our hierarchical model are given in Lemma 1, while the unbiased estimators of its structural parameters are presented in Theorem 4. For the proofs, we refer to Bozikas (2019).

Lemma 1. Based on (44)-(49), the best linear estimators for Level 3 (country), Level 2 (gender) and Level 1 (age) are given by

$$a) \boldsymbol{\beta}_c^{Cred}(\Theta_c^3) = \mathbf{K}_c \widehat{\boldsymbol{\beta}}_c + (\mathbf{I} - \mathbf{K}_c) \boldsymbol{\beta}, \quad (\text{Level3}) \quad (50)$$

$$b) \boldsymbol{\beta}_{gc}^{Cred}(\Theta_{gc}^{23}) = \mathbf{K}_{gc} \widehat{\boldsymbol{\beta}}_{gc} + (\mathbf{I} - \mathbf{K}_{gc}) \boldsymbol{\beta}_c^{Cred}(\Theta_c^3), \quad (\text{Level2}) \quad (51)$$

$$c) \boldsymbol{\beta}_{xgc}^{Cred}(\Theta_{xgc}^{123}) = \mathbf{K}_{xgc} \widehat{\boldsymbol{\beta}}_{xgc} + (\mathbf{I} - \mathbf{K}_{xgc}) \boldsymbol{\beta}_{gc}^{Cred}(\Theta_{gc}^{23}), \quad (\text{Level1}) \quad (52)$$

where \mathbf{K}_c , \mathbf{K}_{gc} and \mathbf{K}_{xgc} are defined in (11), (10) and (9), respectively.

Theorem 4. The unbiased estimators of the structural parameters s^2 , \mathbf{A}_1 , \mathbf{A}_2 and \mathbf{A}_3 are given by the following formulas

$$\widehat{s^2} = \frac{1}{\sum_c^C \sum_{g=1}^{G_c} \sum_{x=x_1}^{x_{k_{gc}}} (n-2)} \sum_c^C \sum_{g=1}^{G_c} \sum_{x=x_1}^{x_{k_{gc}}} (\mathbf{Y}_{xgc} - \mathbf{Z}_{xgc} \widehat{\boldsymbol{\beta}}_{xgc})' \mathbf{W}_{xgc} (\mathbf{Y}_{xgc} - \mathbf{Z}_{xgc} \widehat{\boldsymbol{\beta}}_{xgc}), \quad (53)$$

$$\widehat{\mathbf{A}}_1 = \frac{1}{\sum_c^C \sum_{g=1}^{G_c} (x_{k_{gc}} - x_1)} \sum_c^C \sum_{g=1}^{G_c} \sum_{x=x_1}^{x_{k_{gc}}} \mathbf{K}_{xgc} (\widehat{\boldsymbol{\beta}}_{xgc} - \widehat{\boldsymbol{\beta}}_{gc}) (\widehat{\boldsymbol{\beta}}_{xgc} - \widehat{\boldsymbol{\beta}}_{gc})', \quad (54)$$

$$\widehat{\mathbf{A}}_2 = \frac{1}{\sum_c^C (G_c - 1)} \sum_c^C \sum_{g=1}^{G_c} \mathbf{K}_{gc} (\widehat{\boldsymbol{\beta}}_{gc} - \widehat{\boldsymbol{\beta}}_c) (\widehat{\boldsymbol{\beta}}_{gc} - \widehat{\boldsymbol{\beta}}_c)', \quad (55)$$

$$\widehat{\mathbf{A}}_3 = \frac{1}{(C-1)} \sum_{c=1}^C \mathbf{K}_c (\widehat{\boldsymbol{\beta}}_c - \widehat{\boldsymbol{\beta}}) (\widehat{\boldsymbol{\beta}}_c - \widehat{\boldsymbol{\beta}})'. \quad (56)$$

3. CREDIBILITY MORTALITY FORECASTS

Here, we fit the proposed model to the $\log m_{t,xgc}$ rates for year $t = t_1, \dots, t_n$, age $x = x_1, \dots, x_{k_{gc}}$, gender $g = 1, \dots, G_c$ and country $c = 1, \dots, C$. Then, fitted rates up to year t_n are given by $\widehat{\mathbf{Y}}_{xgc} = \mathbf{Z}_{xgc} \widehat{\boldsymbol{\beta}}_{xgc}$ for $x = x_1, \dots, x_{k_{gc}}$, $g = 1, \dots, G_c$, $c = 1, \dots, C$ and mortality rates up to year $t_n + 1$ (one-year ahead) can be compactly written as

$$\mathbf{Y}_{xgc}^{(n+1)Cred} = \mathbf{Z}_{xgc}^{(n+1)} \boldsymbol{\beta}_{xgc}^{Cred} (\boldsymbol{\Theta}_{xgc}^{123}), \quad (57)$$

where $\mathbf{Y}_{xgc}^{(n+1)Cred}$ and $\mathbf{Z}_{xgc}^{(n+1)}$ denote the rates and the design matrix, respectively for one-year ahead. Substituting $\boldsymbol{\beta}_{xgc}^{Cred} (\boldsymbol{\Theta}_{xgc}^{123})$ from Lemma 1 equivalently yields

$$\mathbf{Y}_{xgc}^{(n+1)Cred} = \mathbf{Z}_{xgc}^{(n+1)} \{ \mathbf{K}_{xgc} \widehat{\boldsymbol{\beta}}_{xgc} + (\mathbf{I} - \mathbf{K}_{xgc}) [\mathbf{K}_{gc} \widehat{\boldsymbol{\beta}}_{gc} + (\mathbf{I} - \mathbf{K}_{gc}) (\mathbf{K}_c \widehat{\boldsymbol{\beta}}_c + (\mathbf{I} - \mathbf{K}_c) \boldsymbol{\beta})] \}.$$

One-year ahead mortality estimates can be extracted from (57) as

$$\log \widehat{m}_{t_n+1,xgc}^{Cred} = \beta_{1,xgc}^{Cred} (\boldsymbol{\Theta}_{xgc}^{123}) + \beta_{2,xgc}^{Cred} (\boldsymbol{\Theta}_{xgc}^{123}) (n+1), \quad (58)$$

where $Y_{t_n+1,xgc}^{Cred} = \log \widehat{m}_{t_n+1,xgc}^{Cred}$ is the credibility estimate for year $t_n + 1$ and $\boldsymbol{\beta}_{xgc}^{Cred} (\boldsymbol{\Theta}_{xgc}^{123}) = (\beta_{1,xgc}^{Cred} (\boldsymbol{\Theta}_{xgc}^{123}), \beta_{2,xgc}^{Cred} (\boldsymbol{\Theta}_{xgc}^{123}))'$ is the credibility coefficients vector. We can now employ two methods to extrapolate mortality rates over a given forecasting horizon $h = 1, 2, \dots, H$.

Method 1: Initial Fitting Span (IF)

The estimates of future mortality rates for ages $x = x_1, \dots, x_{k_{gc}}$ are given by extrapolating one-year ahead estimates (58) to $\log \widehat{m}_{t_n+h,xgc}^{Cred} = \beta_{1,xgc}^{Cred} (\boldsymbol{\Theta}_{xgc}^{123}) + \beta_{2,xgc}^{Cred} (\boldsymbol{\Theta}_{xgc}^{123}) (n+h)$, for $h = 2, \dots, H$. Under this method, forecasts are based on the mortality data of the initial fitting span.

Method 2: Moving Fitting Span (MF)

In actuarial practice, we can use the MF method to estimate future mortality rates. One-year ahead estimates $\log \widehat{m}_{t_n+1,xgc}^{Cred}$ are embedded to the existing fitting span and $\log m_{t_1,xgc}$ is simultaneously excluded from it, such that the fitting year span is moved by one year each time to $[t_2, t_n + 1]$, $[t_3, t_n + 2]$, $[t_4, t_n + 3]$ and on. Then, after estimating structural parameters, we can consecutively obtain $\log \widehat{m}_{t_n+2,xgc}^{Cred}$, $\log \widehat{m}_{t_n+3,xgc}^{Cred}$, \dots , $\log \widehat{m}_{t_n+H,xgc}^{Cred}$. Note that under the MF method forecasts are based on more recent mortality trends, while for both extrapolation methods, one-year ahead estimates $\log \widehat{m}_{t_n+1,xgc}^{Cred}$ are the same.

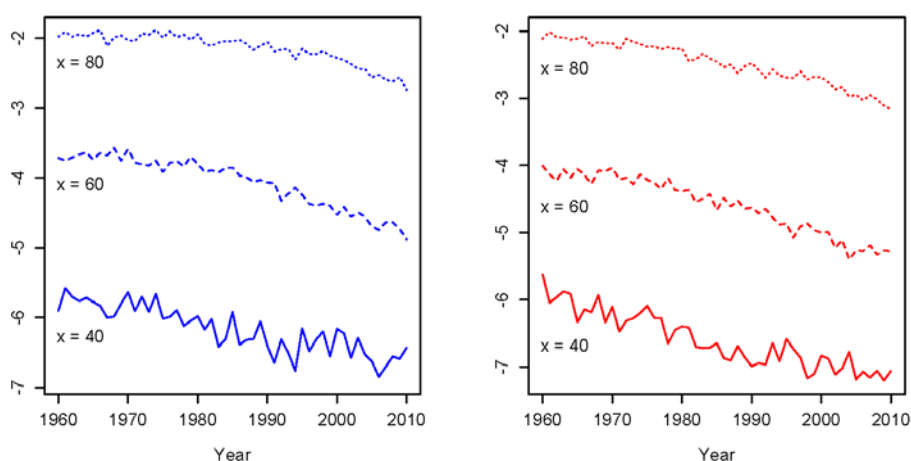
Table 1. Multi-population mortality forecasting formulas

Model	Structure	Original Papers
LC	$\log\hat{m}_{t_n+h,xgc}^{LC} = \log\hat{m}_{t_n,xgc} + (\hat{\alpha}_{xgc}^{(2)} \hat{\theta}_{gc}) h$	Lee and Carter (1992)
LC _{joK}	$\log\hat{m}_{t_n+h,xgc}^{LC_{joK}} = \log\hat{m}_{t_n,xgc} + (\hat{\alpha}_{xgc}^{(2)} \hat{\theta}) h$	Carter and Lee (1992)
LC _{acf}	$\log\hat{m}_{t_n+h,xgc}^{LC_{acf}} = \log\hat{m}_{t_n,xgc} + (\hat{\alpha}_x^{(2)} \hat{\theta} + \hat{\alpha}_{xgc}^{(3)} \hat{\theta}_{gc}) h$	Li and Lee (2005)
HCR	$\log\hat{m}_{t_n+h,xgc}^{Cred} = \beta_{1,xgc}^{Cred} (\Theta_{xgc}^{123}) + \beta_{2,xgc}^{Cred} (\Theta_{xgc}^{123})(n+h)$	in this work

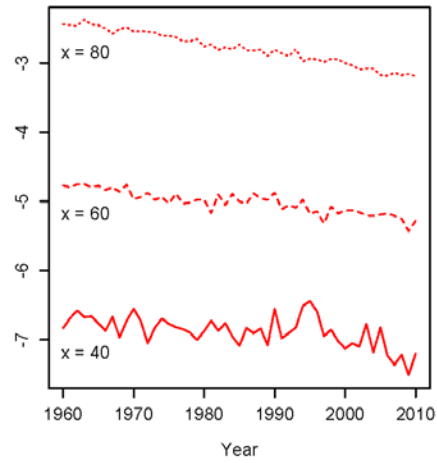
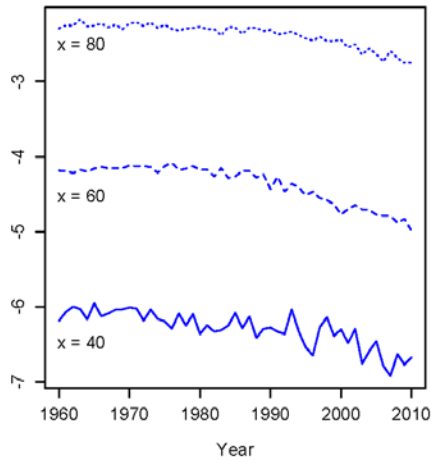
3.1 Evaluation of Forecasting Performance

We proceed by evaluating the forecasting performance of the original Lee-Carter (LC) the joint- K (LC_{joK}), the augmented common factor (LC_{acf}) and the proposed hierarchical credibility regression (HCR) models (see Table 1). Mortality data are obtained from the Human Mortality Database (2018), including the mortality rates for both genders of three northern European countries with small populations, the Ireland, the Norway and the Finland. According to the World Bank database (<https://data.worldbank.org/indicator/SP.POP.TOTL>), the total population for 2018 was 4.85 million for Ireland, 5.31 for Norway and 5.52 for Finland. Mortality evolution over the period 1960-2010 in Ireland, Norway and Finland is illustrated in Figures 1.1, 1.1 and 1.1, respectively, where the observed rates show a downward trend for discrete ages $x = 40, 60, 80$ for males in the left panels and females in the right panels.

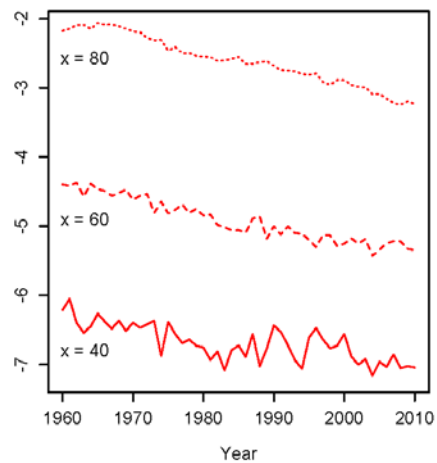
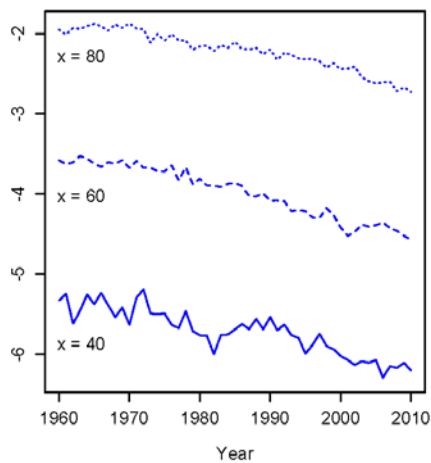
Figure 1. Observed $\log m_{t,x}$ against year at the age of 40, 60 and 80 for males (left) and females (right) in: (a) Ireland, (b) Norway and (c) Finland



(a) Ireland



(b) Norway



(c) Finland

For the numerical illustration that follows, let \mathbf{W}_{xgc} be an identity matrix and $k_{gc} = k$, which implies the same number of fitted ages for any gender g and country c . Then, the $\log m_{t,xgc}$ rates are fitted on selected recent periods of totally $n = 31$ ($[t_1, t_n] = [1970, 2000]$) years and $k = 65$ ($[x_1, x_k] = [20, 84]$) ages.

We consider two forecasting periods $[t_n + 1, t_n + H]$, $h = 1 \dots, H$, for an $H = 10$ years span (2001-2010) and an $H = 20$ years span (1991-2010). Also, for robustness (relative to changes in the fitting period of data) fitting and forecasting periods are partitioned in four sub-periods, i.e., $[1970, 1990]$, $[1970, 2000]$, $[1980, 2000]$ and $[1990, 2000]$. The HCR mortality methods were built into the R (R Core Team, 2017) statistical software.

3.2 Numerical Illustration

The LC model is fitted for selected periods, independently on six single populations (males and females of Ireland, Norway and Finland), while the LC_{joK} , the LC_{acf} , the HCR_{IF} and the HCR_{MF} models are jointly applied to six populations.

The forecasting performances of the LC, the LC_{joK} , the LC_{acf} and the HCR models are comparatively evaluated using the mean absolute percentage forecast error (MAPFE) measure. MAPFE has been used in many studies to measure forecast errors. For some of them, we refer to D'Amato et al. (2012), Lin and Tsai (2015) and Tsai and Zhang (2019). MAPFE values over the selected forecasting horizon H , for age x , gender g and country c are defined as

$$MAPFE = \frac{1}{H \times k} \sum_{h=1}^H \left| \frac{e^{\log \hat{m}_{t_n+h, xgc}^M} - m_{t_n+h, xgc}}{m_{t_n+h, xgc}} \right| \times 100,$$

where $M = LC, LC_{joK}, LC_{acf}$ for each one of the Lee-Carter models and $M = HCR$ for the hierarchical credibility regression models. Accuracy results over the 10-year forecasting period [2001,2010] are presented for both genders of Ireland, Norway and Finland in Tables 2 (a), (b) and (c) for fitting periods [1970,2000], [1980,2000] and [1990,2000], respectively. The corresponding results over the 20-year forecasting period are given in Table 3 for the [1970,1990] fitting period.

MAPFE results indicate that HCR models gave us the better forecasts in average, compared with the single and the multi-population Lee-Carter models for both genders of the three countries (six populations in total). The average values over the six populations are given in the last row of each subtable. The ranking order, based on the lowest values, is given in brackets and shows that for each one of the selected fitting and forecasting periods, the HCR_{MF} and the HCR_{IF} models outperform in average.

For consistency, the forecasting performance was also evaluated with other well-known measures, such as the mean absolute forecast error (MAFE), the root sum of squared forecast error (RSSFE) and the root mean of squared forecast error (RMSFE), leading us to the same ranking results, with the HCR_{MF} and the HCR_{IF} models being on top.

From the results of Tables 2 and 3, we can also observe that deviations in MAPFE values between the best performed HCR and Lee-Carter models range between 10% - 17%. This means that if we use the HCR_{MF} instead of the LC_{acf} to model mortality for years 1970-2000, the 10-year average forecasting performance for both genders of the three countries gets improved by 10%, while if we use the HCR_{IF} instead of the LC_{acf} for years 1990-2000, it gets improved by 17%.

Table 2. MAPFE values (%) over the 10 year forecasting period [2001,2010] for the Lee-Carter type and the proposed hierarchical credibility regression models fitted on years (a) [1970,2000], (b) [1980,2000] and (c) [1990,2000], for both genders of Ireland, Norway and Finland

MAPFE _[t₁,2001,2010]		Lee-Carter Models			Hierarchical Models	
Country	Gender	LC	LC _{joK}	LC _{acf}	HCR _{IF}	HCR _{MF}
(a) Fitting period: [1970,2000]						
Ireland	Male	26.48	27.18	25.75	23.74	22.40
	Female	25.38	21.93	25.55	18.87	18.75
Norway	Male	19.92	21.91	19.95	19.78	18.81
	Female	16.59	16.47	16.08	15.28	15.89
Finland	Male	12.42	14.29	12.24	12.08	12.38
	Female	19.31	17.32	18.60	16.40	17.19
	Average	20.02(5)	19.85(4)	19.69(3)	17.69(2)	17.57(1)
(b) Fitting period: [1980,2000]						
Ireland	Male	24.91	25.44	23.57	21.72	22.49
	Female	24.38	22.43	24.39	19.86	20.77
Norway	Male	18.75	17.11	18.54	15.77	14.13
	Female	18.15	17.64	16.88	16.45	15.28
Finland	Male	13.17	15.16	13.87	13.12	11.38
	Female	21.42	19.83	21.40	18.58	17.12
	Average	20.13(5)	19.60(3)	19.78(4)	17.58(2)	16.86(1)
(c) Fitting period: [1990,2000]						
Ireland	Male	28.39	30.26	26.13	25.14	23.13
	Female	27.52	28.06	29.05	23.06	24.79
Norway	Male	17.19	15.80	16.92	13.86	16.17
	Female	17.78	17.46	17.32	14.04	15.17
Finland	Male	12.20	12.42	10.78	9.84	9.97
	Female	20.32	18.81	21.86	15.35	15.47
	Average	20.57(5)	20.47(4)	20.34(3)	16.88(1)	17.45(2)

Table 3. MAPFE values (%) over the 20 year forecasting period [1991,2010] for the Lee-Carter type and the proposed hierarchical credibility regression models fitted on years [1970,1990], for both genders of Ireland, Norway and Finland

MAPFE $_{[1970,1]}^{[1991,2]}$		Lee-Carter Models			Hierarchical Models	
Country	Gender	LC	LC _{joK}	LC _{acf}	HCR _{IF}	HCR _{MF}
Ireland	Male	25.79	28.24	23.13	23.07	19.6116.0
	Female	23.24	23.62	23.06	17.82	7
Norway	Male	24.17	27.26	23.43	24.47	25.07
	Female	17.88	16.43	17.71	14.83	16.43
Finland	Male	19.42	16.07	19.28	12.25	13.59
	Female	18.78	16.58	17.42	16.20	17.00
	Average	21.55(5)	21.37(4)	20.67(3)	18.11(2)	17.96(1)

4. CONCLUSIONS

In this work, we presented a hierarchical credibility regression modelling approach for mortality data of multiple populations, structured by country, gender and age. Credibility estimators for the parameters of the proposed model are given and future mortality rates are derived using different extrapolation methods.

The forecasting performances between the hierarchical model, the original Lee-Carter model and two extensions of the Lee-Carter model for multiple populations are compared as well. Based on the accuracy results, the proposed credibility models gave us a better forecasting performance in comparison with the Lee-Carter models. This indicates that hierarchical credibility regression modelling can be effectively applied to mortality datasets of multiple populations. Our illustrations also show that the LC_{acf} model performs evidently better than its Lee-Carter counterparts for the selected datasets.

More specifically, accuracy results show that the HCR_{MF} is best performing model for the majority of selected fitting and forecasting periods. This indicates that hierarchical credibility regression modelling can be effectively applied to mortality datasets of multiple populations, with possible similarities in their demographic or socio-economic structure.

ΠΕΡΙΛΗΨΗ

Στη εργασία αυτή, παρουσιάζουμε ένα ιεραρχικό μοντέλο παλινδρόμησης χαρτοφυλακίου τριών επιπέδων για τη μοντελοποίηση των δεικτών θνησιμότητας πολλαπλών πληθυσμών. Η διαδικασία εκτίμησης των παραμέτρων του μοντέλου αξιοπιστίας περιγράφεται, ενώ η προβλεπτική του απόδοση συγκρίνεται ανά φύλο και ηλικία, με αυτήν του μοντέλου Lee-Carter και δύο επεκτάσεών του για τα δεδομένα θνησιμότητας τριών χωρών της Βόρειας Ευρώπης. Τα αριθμητικά αποτελέσματα δείχνουν ότι η προτεινόμενη μέθοδος συνεισφέρει σε πιο ακριβείς προβλέψεις, σε σχέση με αυτές που προκύπτουν μέσω του μοντέλου Lee-Carter και των επεκτάσεών του για πολλαπλούς πληθυσμούς.

Acknowledgments: This work has been partly supported by the University of Piraeus Research Center.

REFERENCES

- Bozikas, A. (2019). *Actuarial models in demography*. PhD thesis, Department of Statistics and Insurance Science, University of Piraeus, Greece.
- Bozikas, A. and Pitselis, G. (2019). Credible Regression Approaches to Forecast Mortality for Populations with Limited Data. *Risks*, 7(1), 27.
- Bühlmann, H. (1967). Experience Rating and Credibility. *ASTIN Bulletin*, 4(3):199–207.
- Bühlmann H, Straub E (1970). Glaubwürdigkeit für Schadensätze. *Bulletin of the Swiss Association of Actuaries* 70:111–133.
- Bühlmann, H. and Gisler, A. (2005). *A course in credibility theory and its applications*. The Netherlands: Springer.
- Cairns, A. J. G., Blake, D., and Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. *Journal of Risk and Insurance*, 73(4):687–718.
- Carter, L. R. and Lee, R. D. (1992). Modeling and forecasting US sex differentials in mortality. *International Journal of Forecasting*, 8(3):393–411.
- D’Amato, V., Haberman, S., Piscopo, G., and Russolillo, M. (2012). Modelling dependent data for longevity projections. *Insurance: Mathematics and Economics*, 51(3):694–701.
- De Vylder, F. (1976). Geometrical credibility. *Scandinavian Actuarial Journal*, pages 121–149.
- Goovaerts, M. J., Kaas, R., Van Heerwaarden, A. E., and Bauwelinckx, T. (1990). *Effective actuarial methods*. North-Holland.
- Hachemeister, C. (1975). Credibility for Regression Models with Application to Trend (Reprint). In Kahn, P., editor, *CREDIBILITY: Theory and Applications*, pages 307–348. Academic Press, Inc.
- Human Mortality Database (2018). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org.

- Jewell, W. S. (1975). The use of collateral data in credibility theory: a hierarchical model. *IIASA Research Memorandum*, (June):1–16.
- Lee, R. D. and Carter, L. R. (1992). Modeling and Forecasting U.S. Mortality. *Journal of the American Statistical Association*, 87(419):659–671.
- Li, N. and Lee, R. (2005). Coherent mortality forecasts for a group of populations: an extension of the Lee-Carter method. *Demography*, 42(3):575–594.
- Lin, T. and Tsai, C. (2015). A Simple Linear Regression Approach to Modeling and Forecasting Mortality Rates. *Journal of Forecasting*, 34(May):543–559.
- Plat, R. (2009). On stochastic mortality modeling. *Insurance: Mathematics and Economics*, 45(3):393–404.
- R Core Team (2017). R: A language and environment for Statistical Computing. Available at <https://www.r-project.org/>.
- Renshaw, A. E. and Haberman, S. (2006). A cohort-based extension to the Lee-Carter model for mortality reduction factors. *Insurance: Mathematics and Economics*, 38(3):556–570.
- Schinzinger, E., Denuit, M. M., and Christiansen, M. C. (2016). A multivariate evolutionary credibility model for mortality improvement rates. *Insurance: Mathematics and Economics*, 69:70–81.
- Sundt, B. (1979). A Hierarchical Credibility Regression Model. *Scandinavian Actuarial Journal*, 1979(2-3):107–114.
- Sundt, B. (1980). A multi-level hierarchical credibility regression model. *Scandinavian Actuarial Journal*, 1980(1):25–32.
- Tsai, C. and Lin, T. (2017). A Bühlmann credibility approach to modeling mortality rates. *North American Actuarial Journal*, 21(2):204–227.
- Tsai, C. and Zhang, Y. (2019). A multi-dimensional Bühlmann credibility approach to modeling multi-population mortality rates. *Scandinavian Actuarial Journal*, 0(0):1–26.



MULTIVARIATE DISCRETE q -UNIFORM DISTRIBUTIONS

Charalambos A. Charalambides

Department of Mathematics, University of Athens,
Panepistemiopolis, GR-15784 Athens, Greece
ccharal@math.uoa.gr

ABSTRACT

Suppose that n indistinguishable balls are randomly q -distributed, one after the other, into $k+1$ distinguishable urns (cells), with limited or unlimited capacity, and let X_j be the number of balls placed in the j th urn, for $j = 1, 2, \dots, k$. The distribution of the random vector (X_1, X_2, \dots, X_k) in both cases, which turned out to be q -analogues of multivariate discrete uniform distributions, are derived. Furthermore, the multivariate discrete q -uniform distributions of the first and second kind are obtained as the joint conditional distribution of k independent q -Bernoulli and q -geometric distributions, respectively, given their sum with another q -Bernoulli and q -geometric distribution, respectively, independent of them.

Keywords: Multivariate discrete q -distribution; q -Bernoulli distribution; q -Geometric distribution

1. INTRODUCTION

A discrete q -uniform distribution first emerged as a congruence class distribution, modulo n , of Bernoulli generated numbers, in a paper of Rawlings (1994). Kupershmidt (2000) discussed a discrete q -uniform distribution starting with a nonnegative q -function defined on the set $\{0, 1, \dots, n\}$ and summing to one. Charalambides (2014, 2016) defined a discrete q -uniform distribution on a suitable stochastic model and extensively studied it.

The most important multivariate discrete uniform distributions are the Fermi-Dirac and Bose-Einstein distributions, with probability functions

$$P(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) = 1 / \binom{k+1}{n},$$

for $x_j = 0, 1$ and $j = 1, 2, \dots, k$, with $\sum_{j=1}^k x_j \leq n$, and

$$P(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) = 1 / \binom{k+n}{n},$$

for $x_j = 0, 1, \dots, n$ and $j = 1, 2, \dots, k$, with $\sum_{j=1}^k x_j \leq n$, respectively. In both stochastic models, a mechanical system of n particles is considered and X_j is the number of particles in the j th energy level, $j = 1, 2, \dots, k$, of the system. In the Fermi-Dirac model the particles obey the Pauli exclusion principle. These distributions are obtained by assuming that any particle is equally likely to move from the zeroth energy level to any of the $k + 1$ energy levels. In the present article, allowing the probability of a particle to move from the zeroth energy level to one of the $k + 1$ energy levels to vary geometrically, with rate q , multivariate discrete q -uniform distributions are introduced and studied. Section 2 is devoted to the presentation of q -permutations and q -combinations, which pave the way to the introduction and study of multivariate discrete q -uniform distributions. In section 3, a stochastic model of a sequence of successive q -distribution of n indistinguishable balls into distinguishable urns (cells) with limited or unlimited capacity is presented. Then, multivariate discrete q -uniform distributions are defined on this model and their probability functions are derived.

2. BASIC q -COMBINATORICS

The notions of q -permutations and q -combinations, which are useful in q -combinatorics and discrete q -probability, may be introduced by first defining the notion of a q -selection of an element from a set. Thus, consider a set of n elements $W_n = \{w_1, w_2, \dots, w_n\}$, with $w_1 < w_2 < \dots < w_n$, and the permutation of its elements in ascending order, $P_n = (w_1, w_2, \dots, w_n)$.

A selection of element w_{i_1} of the set W_n , through traveling from the left to the right of P_n , with weight q^{i_1-1} , where $i_1 - 1$ is the number of elements preceded w_{i_1} in P_n , is called q -selection. Chung and Kang (1994), who introduced the notion of a q -selection, consider the weight q^{i_1-1} as the payment for the $i_1 - 1$ jumps made in traveling from the left to the right of P_n before selecting w_{i_1} .

The number of ways of q -selecting an element from a set of n elements $W_n = \{w_1, w_2, \dots, w_n\}$ equals the q -integer $[n]_q = (1 - q^n)/(1 - q)$. Indeed, the element w_i is q -selected with weight (payment) q^{i-1} , for $i = 1, 2, \dots, n$, and, according to the addition principle, either w_1 or w_2 or \dots or w_n is q -selected with weight $\sum_{i=1}^n q^{i-1} = (1 - q^n)/(1 - q) = [n]_q$.

Furthermore, an ordered q -selection $(w_{i_1}, w_{i_2}, \dots, w_{i_r})$ of r elements from the set $W_n = \{w_1, w_2, \dots, w_n\}$, with $w_1 < w_2 < \dots < w_n$, is called q -permutation, while a (non-ordered) q -selection $\{w_{i_1}, w_{i_2}, \dots, w_{i_r}\}$ is called q -combination. Note that the elements of a q -permutation or a q -combination of r out of n , may or may not be different elements of W_n . For the first case, the term q -permutation

or q -combination of r out of n is used, while for the second and when unrestricted repetitions are allowed, the term q -permutation or q -combination of r out of n , with repetition, is used.

The next theorem is concerned with the numbers of q -permutations and the numbers of q -combinations, without and with repetitions.

Theorem 1. (a) The number of q -permutations of r out of n equals $[n]_{r,q}$, while the number of q -permutations of r out of n , with repetition, is $[n]_q^r$.

(b) The number of q -combinations of r out of n equals $\begin{bmatrix} n \\ r \end{bmatrix}_q$, while the number of q -combinations of r out of n , with repetition, is $\begin{bmatrix} n+r-1 \\ r \end{bmatrix}_q$.

Proof. (a) In any q -permutation $(w_{i_1}, w_{i_2}, \dots, w_{i_r})$ of r out of n , the first element w_{i_1} is q -selected from the set W_n of n elements in $[n]_q$ ways, while after the q -selection of the first element, the second element w_{i_2} , which must be different from w_{i_1} , is q -selected from the set $W_n - \{w_{i_1}\}$ of $n-1$ elements in $[n-1]_q$ ways. Finally, after the q -selection of the first $r-1$ elements $(w_{i_1}, w_{i_2}, \dots, w_{i_{r-1}})$, the last element w_{i_r} is q -selected from the set $W_n - \{w_{i_1}, w_{i_2}, \dots, w_{i_{r-1}}\}$ of $n-r+1$ elements in $[n-r+1]_q$ ways. Thus, according to the multiplication principle, the number of q -permutations of r out of n is given by the product $[n]_q [n-1]_q \cdots [n-r+1]_q = [n]_{r,q}$.

Also, in any q -permutation $(w_{i_1}, w_{i_2}, \dots, w_{i_r})$ of r out of n , with repetition, the m th element w_{i_m} is q -selected from the set W_n of n elements in $[n]_q$ ways, for $m = 1, 2, \dots, r$. Thus, according to the multiplication principle, the number of q -permutations of r out of n , with repetition, is given by the power $[n]_q^r$.

(b) Consider a q -combination $\{w_{i_1}, w_{i_2}, \dots, w_{i_r}\}$ of r out of n , with $1 \leq i_1 < i_2 < \dots < i_r \leq n$. The element w_{i_1} is q -selected from the set W_n with weight q^{i_1-1} . After the q -selection of w_{i_1} , the element w_{i_2} is q -selected from the set $W_n - \{w_{i_1}\}$ with weight q^{i_2-2} . Finally, after the q -selection of the $r-1$ elements $\{w_{i_1}, w_{i_2}, \dots, w_{i_{r-1}}\}$, the element w_{i_r} is q -selected from the set $W_n - \{w_{i_1}, w_{i_2}, \dots, w_{i_{r-1}}\}$ with weight q^{i_r-r} . Thus, according to the multiplication principle, the r elements of the q -combination $\{w_{i_1}, w_{i_2}, \dots, w_{i_r}\}$ can be q -selected with weight

$$q^{i_1-1} q^{i_2-2} \cdots q^{i_r-r} = q^{i_1+i_2+\dots+i_r-\binom{r+1}{2}}, \quad 1 \leq i_1 < i_2 < \dots < i_r \leq n,$$

according to the addition principle, the number of q -combinations of r out of n , equals the number of q -selections of their elements, with total weight the sum

$$\sum_{1 \leq i_1 < i_2 < \dots < i_r \leq n} q^{i_1+i_2+\dots+i_r-\binom{r+1}{2}},$$

which, by Theorem 1.2 in Charalambides (2016), equals $\begin{bmatrix} n \\ r \end{bmatrix}_q$.

Consider now, a q -combination $\{w_{i_1}, w_{i_2}, \dots, w_{i_r}\}$ of r out of n , with repetition, and $1 \leq i_1 \leq i_2 \leq \dots \leq i_r \leq n$. The element w_{i_m} is q -selected from the

set W_n with weight q^{i_m-1} , for $m = 1, 2, \dots, r$. Therefore, according to the multiplication principle, the r elements of the q -combination $\{w_{i_1}, w_{i_2}, \dots, w_{i_r}\}$ can be q -selected with weight

$$q^{i_1-1}q^{i_2-1} \dots q^{i_r-1} = q^{i_1+i_2+\dots+i_r-r}, \quad 1 \leq i_1 \leq i_2 \leq \dots \leq i_r \leq n,$$

according to the addition principle, the number of q -combinations of r out of n , with repetition, equals the number of q -selections of their elements, with total weight the sum

$$\sum_{1 \leq i_1 \leq i_2 \leq \dots \leq i_r \leq n} q^{i_1+i_2+\dots+i_r-r},$$

which, by Theorem 1.2 in Charalambides (2016), equals $\left[\begin{smallmatrix} n+r-1 \\ r \end{smallmatrix} \right]_q$.

Two useful expressions of the q -binomial coefficients $\left[\begin{smallmatrix} k+1 \\ n \end{smallmatrix} \right]_q$ and $\left[\begin{smallmatrix} k+n \\ n \end{smallmatrix} \right]_q$, with k and n positive integers, as multiple sums, over all partitions into k unequal parts and k parts, respectively, none greater than n , are presented in the following corollary of Theorem 1.2 in Charalambides (2016).

Corollary 1. *Let k and n be positive integers, and q be a real number, with $q \neq 1$. Then,*

$$\sum_{\substack{r_i=0,1, i=1,2,\dots,k \\ r_1+r_2+\dots+r_k \leq n}} q^{r_1+2r_2+\dots+kr_k} \binom{n}{2} = \left[\begin{smallmatrix} k+1 \\ n \end{smallmatrix} \right]_q, \quad n \leq k+1, \quad (1)$$

and

$$\sum_{\substack{r_i=0,1,\dots,n, i=1,2,\dots,k \\ r_1+r_2+\dots+r_k \leq n}} q^{r_1+2r_2+\dots+kr_k} = \left[\begin{smallmatrix} k+n \\ n \end{smallmatrix} \right]_q. \quad (2)$$

Proof. The q -binomial coefficients $\left[\begin{smallmatrix} k+1 \\ n \end{smallmatrix} \right]_q$ and $\left[\begin{smallmatrix} k+n \\ n \end{smallmatrix} \right]_q$, according to Theorem 1.2 in Charalambides (2016) may be expressed as

$$\sum_{1 \leq i_1 < i_2 < \dots < i_n \leq k+1} q^{i_1+i_2+\dots+i_n} \binom{n+1}{2} = \left[\begin{smallmatrix} k+1 \\ n \end{smallmatrix} \right]_q \quad (3)$$

and

$$\sum_{1 \leq i_1 \leq i_2 \leq \dots \leq i_n \leq k+1} q^{i_1+i_2+\dots+i_n} = \left[\begin{smallmatrix} k+n \\ n \end{smallmatrix} \right]_q. \quad (4)$$

Let r_j be the number of the variables i_1, i_2, \dots, i_n that are equal to $j+1$, for $j = 0, 1, \dots, k$. Note that $r_j = 0, 1$, for $1 \leq i_1 < i_2 < \dots < i_n \leq k+1$ and $r_j = 0, 1, \dots, n$, for $1 \leq i_1 \leq i_2 \leq \dots \leq i_n \leq k+1$. Then,

$$i_1 + i_2 + \dots + i_n = r_0 + 2r_1 + \dots + (k+1)r_k, \quad \text{with } r_0 + r_1 + \dots + r_k = n.$$

Thus

$$i_1 + i_2 + \cdots + i_n - \binom{n+1}{2} = r_1 + 2r_2 + \cdots + kr_k - \binom{n}{2}, \text{ with } r_1 + r_2 + \cdots + r_k \leq n,$$

and

$$i_1 + i_2 + \cdots + i_n - n = r_1 + 2r_2 + \cdots + kr_k, \text{ with } r_1 + r_2 + \cdots + r_k \leq n.$$

Consequently, (3) and (4) may be expressed as (1) and (2), respectively.

3. MULTIVARIATE DISCRETE q -UNIFORM DISTRIBUTIONS

A random distribution (placement) of balls into distinguishable urns (cells) is a simple and very useful stochastic model. Among its most striking applications, the Bose-Einstein and Fermi-Dirac models (or statistics in physicist's terminology) worth mentioning.

A random q -distribution (placement) of a ball into r distinguishable urns (cells) $\{c_1, c_2, \dots, c_r\}$ may be introduced as follows. Assume that r numbered balls $\{1, 2, \dots, r\}$, representing the r urns are forced to pass through a random mechanism, one after the other, in the order $(1, 2, \dots, r)$ or in the reverse order $(r, r-1, \dots, 2, 1)$. Also, suppose that each passing ball may or may not be caught by the mechanism, with probabilities $p = 1 - q$ and q , respectively. In the case all r balls pass through the mechanism and no ball is caught, the ball passing procedure is repeated, with the same order. Then, the number on the first caught ball determines the urn (cell) in which the ball is placed. Clearly, the probability that a ball is placed in the j th in order urn is given by

$$p_j = \sum_{k=0}^{\infty} (1-q)q^{(j-1)+kr} = \frac{q^{j-1}}{[r]_q}, \quad j = 1, 2, \dots, r,$$

or by

$$p_j = \sum_{k=0}^{\infty} (1-q)q^{(r-j)+kr} = \frac{q^{r-j}}{[r]_q}, \quad j = 1, 2, \dots, r,$$

where $0 < q < 1$, according to whether the ball passing order is $(1, 2, \dots, r)$ or $(r, r-1, \dots, 2, 1)$. These probabilities, on using the expression $q^{j-1}/[r]_q = q^{-\{r-j\}}/[r]_{q^{-1}}$, may be written in a single formula as

$$p_j = \frac{q^{r-j}}{[r]_q}, \quad j = 1, 2, \dots, r, \tag{5}$$

where $0 < q < 1$ or $1 < q < \infty$. Note that this is the probability function of a Discrete q -Uniform Distribution of the set $\{1, 2, \dots, r\}$.

Furthermore, assume that n indistinguishable balls are randomly q -distributed, one after the other, into $k+1$ distinguishable urns (cells) $\{c_1, c_2, \dots, c_{k+1}\}$, each

with capacity unlimited to one ball, with $n \leq k+1$. Let X_j be the number of balls placed in urn c_j , for $j = 1, 2, \dots, k+1$. Note that $X_{k+1} = n - X_1 - X_2 - \dots - X_k$. The distribution of the random vector (X_1, X_2, \dots, X_k) is called *Multivariate Discrete q -Uniform Distribution of the first kind*, with parameters n and q . Its probability function is derived in the following theorem.

Theorem 2. *The probability function of the multivariate discrete q -uniform distribution of the first kind, with parameters n and q is given by*

$$P(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) = q^{\sum_{j=1}^k (k-j+1)x_j - \binom{n}{2}} / \left[\begin{matrix} k+1 \\ n \end{matrix} \right]_q, \quad (6)$$

for $x_j = 0, 1$ and $j = 1, 2, \dots, k$, with $\sum_{j=1}^k x_j \leq n$, and $0 < q < 1$ or $1 < q < \infty$.

Proof. A random q -distribution of n indistinguishable balls into the $k+1$ distinguishable urns, of limited capacity, may be represented by the collection of n q -selected urns $\{c_{i_1}, c_{i_2}, \dots, c_{i_n}\}$, where the q -selection of an urn x times corresponds to the placement of x balls into it, for $x = 0, 1$. Notice that, after the q -selection of an urn and the placement in it a ball, because its capacity is limited to one ball, the next q -selection is made among the remaining urns. Therefore, the probability for such a q -distribution, on using successively (5), with $r = k+1, k, \dots, k-n+2$, is given by

$$c q^{k-i_1+1} q^{k-i_2} \dots q^{k-i_n-n+2} = c q^{(k+1)n - (i_1+i_2+\dots+i_n) - \binom{n}{2}},$$

with $1 \leq i_1 < i_2 < \dots < i_n \leq k+1$. Clearly, the number x_j of balls q -distributed into urn c_j equals the number of variables i_1, i_2, \dots, i_n that are equal to j , for $j = 1, 2, \dots, k+1$, with $x_{k+1} = n - \sum_{j=1}^k x_j$. Also, the exponent of q in the expression of the preceding random q -distribution, may be written as

$$(k+1)n - \sum_{r=1}^n i_r - \binom{n}{2} = \sum_{j=1}^{k+1} (k+1)x_j - \sum_{j=1}^{k+1} jx_j - \binom{n}{2} = \sum_{j=1}^k (k-j+1)x_j - \binom{n}{2}$$

and so

$$P(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) = c q^{\sum_{j=1}^k (k-j+1)x_j - \binom{n}{2}},$$

for $x_j = 0, 1$ and $j = 1, 2, \dots, k$, with $\sum_{j=1}^k x_j \leq n$. Summing these probabilities, using (1), and equating this sum to one, we get the expression $c = 1 / \left[\begin{matrix} k+1 \\ n \end{matrix} \right]_q$, which completes the derivation of (6).

The multivariate discrete q -uniform distribution of the first kind may be obtained as the conditional distribution of k independent q -Bernoulli distributions of the first kind, given their sum with another q -Bernoulli distribution of the first kind independent of them, according to the following theorem.

Theorem 3. Consider a sequence of independent Bernoulli trials and assume that the probability of success at the i th trial is given by

$$p_i = \frac{\theta q^{i-1}}{1 + \theta q^{i-1}}, \quad i = 1, 2, \dots, \quad 0 < q < 1 \quad \text{or} \quad 1 < q < \infty.$$

Let X_j be the number of successes at the j th trial, for $j = 1, 2, \dots, k + 1$. Then, the conditional probability function of the random vector (X_1, X_2, \dots, X_k) , given that $X_1 + X_2 + \dots + X_{k+1} = n$, is the multivariate discrete q -uniform distribution of the first kind with probability function (6).

Proof. The random variables X_j , $j = 1, 2, \dots, k + 1$, are independent, with probability function, according to Theorem 2.1 in Charalambides (2016), is given by

$$P(X_j = x_j) = \frac{\theta^{x_j} q^{(j-1)x_j}}{1 + \theta q^{j-1}}, \quad x_j = 0, 1, \quad j = 1, 2, \dots, k + 1.$$

Similarly, the probability function of the sum $Y_{k+1} = X_1 + X_2 + \dots + X_{k+1}$, which is the number of successes in $k + 1$ trials, is

$$P(Y_{k+1} = n) = \binom{k+1}{n}_q \frac{\theta^n q^{\binom{n}{2}}}{\prod_{i=1}^{k+1} (1 + \theta q^{i-1})}, \quad n = 0, 1, \dots, k + 1.$$

Then, the joint conditional probability function of the random vector (X_1, X_2, \dots, X_k) , given that $Y_{k+1} = n$,

$$P(X_1 = x_1, \dots, X_k = x_k | Y_{k+1} = n) = \frac{P(X_1 = x_1) \cdots P(X_k = x_k) P(X_{k+1} = n - y_k)}{P(Y_{k+1} = n)},$$

with $y_k = \sum_{j=1}^k x_j$, on using these expressions, is obtained as

$$P(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k | Y_{k+1} = n) = q^{c_{n,k}(x_1, x_2, \dots, x_k)} \Big/ \binom{k+1}{n}_q,$$

where

$$\begin{aligned} c_{n,k}(x_1, x_2, \dots, x_k) &= \sum_{j=1}^k (j-1)x_j - \sum_{j=1}^k kx_j + nk - \binom{n}{2} \\ &= - \sum_{j=1}^k (k-j+1)x_j + \binom{n}{2} + n(k-n+1). \end{aligned}$$

Thus, since

$$q^{-n(k-n+1)} \binom{k+1}{n}_q = \binom{k+1}{n}_{q^{-1}},$$

it reduces to

$$P(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k | Y_{k+1} = n) = q^{-\sum_{j=1}^k (k-j+1)x_j + \binom{n}{2}} / \left[\begin{matrix} k+1 \\ n \end{matrix} \right]_{q^{-1}},$$

which is expression (6) with q replaced by q^{-1} .

Suppose now that n indistinguishable balls are randomly q -distributed, one after the other, into $r = k + 1$ distinguishable urns (cells) $\{c_1, c_2, \dots, c_{k+1}\}$, with unlimited capacity. Let X_j be the number of balls placed in urn c_j , for $j = 1, 2, \dots, k + 1$. Note that $X_{k+1} = n - X_1 - X_2 - \dots - X_k$. The distribution of the random vector (X_1, X_2, \dots, X_k) is called *Multivariate Discrete q -Uniform Distribution of the second kind*, with parameters n and q . Its probability function is derived in the following theorem.

Theorem 4. *The probability function of the multivariate discrete q -uniform distribution of the first kind, with parameters n and q is given by*

$$P(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) = q^{\sum_{j=1}^k (k-j+1)x_j} / \left[\begin{matrix} k+n \\ n \end{matrix} \right]_q, \quad (7)$$

for $x_j = 0, 1, \dots, n$ and $j = 1, 2, \dots, k$, with $\sum_{j=1}^k x_j \leq n$, and $0 < q < 1$ or $1 < q < \infty$.

Proof. A random q -distribution of n indistinguishable balls into the $k + 1$ distinguishable urns may be represented by the collection of n q -selected urns $\{c_{i_1}, c_{i_2}, \dots, c_{i_n}\}$, with repetition, where the q -selection of an urn x times corresponds to the placement of x balls into it. The probability for such a q -distribution, on using (5) with $r = k + 1$, is given by

$$c q^{k-i_1+1} q^{k-i_2+1} \dots q^{k-i_n+1} = c q^{(k+1)n - (i_1+i_2+\dots+i_n)},$$

with $1 \leq i_1 \leq i_2 \leq \dots \leq i_n \leq k + 1$. Clearly, the number x_j of balls q -distributed into urn c_j equals the number of variables i_1, i_2, \dots, i_n that are equal to j , for $j = 1, 2, \dots, k + 1$, with $x_{k+1} = n - \sum_{j=1}^k x_j$. Also, the exponent of q in the expression of the preceding random q -distribution, may be written as

$$(k+1)n - \sum_{r=1}^n i_r = \sum_{j=1}^{k+1} (k+1)x_j - \sum_{j=1}^{k+1} jx_j = \sum_{j=1}^k (k-j+1)x_j$$

and so

$$P(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) = c q^{\sum_{j=1}^k (k-j+1)x_j},$$

for $x_j = 0, 1, \dots, n$ and $j = 1, 2, \dots, k$, with $\sum_{j=1}^k x_j \leq n$. Summing these probabilities, using (1), and equating this sum to one, we get the expression $c = 1 / \left[\begin{matrix} k+n \\ n \end{matrix} \right]_q$, which completes the derivation of (7).

The multivariate discrete q -uniform distribution of the second kind may be obtained as the conditional distribution of k independent q -geometric distributions of the second kind, given their sum with another q -geometric distribution of the second kind independent of them, according to the following theorem.

Theorem 5. Consider a sequence of independent Bernoulli trials and assume that the conditional probability of success at a trial, given that $j - 1$ successes occur in the previous trials, is given by

$$p_j = 1 - \theta q^{j-1}, \quad j = 1, 2, \dots, \quad 0 < \theta < 1, \quad 0 < q < 1 \quad \text{or} \quad 1 < q < \infty,$$

where, for $1 < q < \infty$, the number j of successes is restricted by $j \leq m = -\log \theta / \log q$. Let W_j be the number of failures after the $(j - 1)$ th success and until the occurrence of the j th success, for $j = 1, 2, \dots, k + 1$, where $k + 1 \leq m$ in the case $1 < q < \infty$. Then, the conditional probability function of the random vector (W_1, W_2, \dots, W_k) , given that $W_1 + W_2 + \dots + W_{k+1} = n$, is the multivariate discrete q -uniform distribution of the second kind with probability function (7).

Proof. Clearly, the random variables W_j , $j = 1, 2, \dots, k + 1$, are independent, with probability function,

$$P(W_j = w_j) = (\theta q^{j-1})^{w_j} (1 - \theta q^{j-1}), \quad w_j = 0, 1, \dots, \quad j = 1, 2, \dots, k + 1.$$

Also, the probability function of the sum $U_{k+1} = W_1 + W_2 + \dots + W_{k+1}$, which is the number of failures until the occurrence of the $(k + 1)$ th success, is

$$P(U_{k+1} = n) = \begin{bmatrix} k + n \\ n \end{bmatrix}_q \theta^n \prod_{i=1}^{k+1} (1 - \theta q^{i-1}), \quad n = 0, 1, \dots$$

Then, the joint conditional probability function of the random vector (W_1, W_2, \dots, W_k) , given that $U_{k+1} = n$,

$$P(W_1 = w_1, \dots, W_k = w_k | U_{k+1} = n) = \frac{P(W_1 = w_1) \cdots P(W_k = w_k) P(W_{k+1} = n - u_k)}{P(U_{k+1} = n)},$$

with $u_k = \sum_{j=1}^k w_j$, on using these expressions, is obtained as

$$P(W_1 = w_1, \dots, W_k = w_k | U_{k+1} = n) = q^{c_{n,k}(w_1, w_2, \dots, w_k)} / \begin{bmatrix} k + n \\ n \end{bmatrix}_q,$$

where

$$c_{n,k}(w_1, w_2, \dots, w_k) = \sum_{j=1}^k (j - 1)w_j - \sum_{j=1}^k k w_j + kn = - \sum_{j=1}^k (k - j + 1)w_j + kn.$$

Thus, since

$$q^{-kn} \begin{bmatrix} k+n \\ n \end{bmatrix}_q = \begin{bmatrix} k+n \\ n \end{bmatrix}_{q^{-1}},$$

it reduces to

$$P(W_1 = w_1, \dots, W_k = w_k | U_{k+1} = n) = q^{-\sum_{j=1}^k (k-j+1)w_j} / \begin{bmatrix} k+n \\ n \end{bmatrix}_{q^{-1}}$$

which is expression (7) with q replaced by q^{-1} .

REFERENCES

- Charalambides, Ch. A. (2014). Discrete q -uniform distribution. *Proceedings of the 27th Panhellenic Statistics Conference*, 357-366.
- Charalambides, Ch. A. (2016). *Discrete q -Distributions*. John Wiley & Sons, Hoboken, New Jersey.
- Chung, W-S. and Kang, H-J. (1994). q -Permutation and q -combinations. *Intern. J. Theor. Physics* **33**, 851-856.
- Kupershmidt, B. A. (2000). q -Probability: 1. Basic discrete distributions. *J. Nonlinear Math. Phys.* **7**, 73-93.
- Rawlings, D. (1994). Bernoulli trials and number theory. *Amer. Math. Monthly* **101**, 948-952.



A QUANTITATIVE ANALYSIS OF AGRICULTURAL PRODUCTION IN GREECE, 2004-2016

Leonidas Zangelidis

University of Athens, KEPE intern

leonidaszangelidis@gmail.com

ABSTRACT

The paper looks into Greece's agricultural production in terms of labor, land, capital, energy and other inputs at the subnational level, across four regions, on the basis of annual observations for the average farm, running from 2004 to 2016, supplied by the EU Commission. It depicts the evolution of the main variables and of productivity; carries out Granger causality tests on all inputs in each region and of each input across the four regions to gain insights in the sector's mechanics; and econometrically estimates the impact of the said inputs on output and the magnitude of multifactor productivity (i.e., of entrepreneurship, technology and the impact of the factors not considered in the regression) in an extended Cobb-Douglas function framework. Translog-like features are taken into account, alternative specifications are considered in order to deal with non-stationarity, and all regressors are rendered uncorrelated to each other so as to deal with heteroskedasticity. The results suggest that labor and the cost of energy constitute the main explanatory factors. However, their impact along with the size of multifactor productivity vary across space. This implies that there is room for spatially differentiated interventions.

Keywords: Cobb-Douglas, agricultural production, productivity, regional analysis, stationarity

1. INTRODUCTION

The purpose of the paper is to study the economics of production in the agricultural sector in Greece across four regions (north Greece, west Greece, Thessaly, the rest of Greece), a few years prior to the international financial and economic crisis and during the country's long recession that followed, specifically from 2004 to 2016, in order to get a sense of the sector's mechanics and how the sector evolved. As the reader may or may not know, the agricultural sector has been seen as a fallback during the downturn; e.g., Giannakis and Bruggeman, 2017. To that end the paper uses the annual data associated with the average farm in each territorial division, as provided by the European Commission's Farm Accountancy Data Network (FADN).

Similar issues have been studied and analyses carried out in a number of countries (e.g., Nowak and Kijek, 2016; Baráth and Fertő, 2017; Fuglie et al., 2017; European Commission, 2018; and the sources cited therein –to mention some of the most recent), while many more focus on individual subsectors, products, regions or subregions. As a result, the paper is part of an identifiable strand in economic literature.

It is organized as follows: Section 2 describes the data and in the process presents some easy-to-compute figures. Section 3 discusses modeling issues for the econometric estimation of the sector's production function. Section 4 supplies the empirical findings along with the results of Granger causalities tests, while Section 5 provides the conclusions.

2. INITIAL FINDINGS

In theory, the level of production (output), Q , depends on the quantity and quality of inputs, the way inputs are combined, and the factors shaping people's demand. The inputs consist of labor (specialized and/or unskilled), L ; manmade capital (the stock of buildings, K_1 , and machines, K_2), K ; the land (earth's surface in the broad sense: area size and configuration, water, flora-fauna-minerals etc.) either in terms of value, T_a , or size (hectares) T_b ; energy, E , in monetary figures; and the other materials used in the process (all obtained from land, labor and capital), M .

The descriptive statistics of the FADN data are provided in Table 1, and suggest that during the time in question, on average, holdings in: (i) The north part of the country (running from West Macedonia to West Thrace, along with the islands of Thasos and Samothraki) featured more K_2 and T_b , a higher value of livestock, and higher costs for E , feeds, seeds, fertilizers, and such inputs. (ii) The west part of the country (spanning Epiros, the Ionian islands and the Peloponnese) were associated with higher T_a . (iii) Thessaly (in central Greece) and north Greece provided more Q . (iv) The rest of Greece (i.e., the part of the mainland situated south of Epiros and Thessaly and north of the Peloponnese, plus Crete and the other Aegean islands) relied on more L , used more K_1 , livestock (in terms of heads), and carried out more investments in buildings and machinery.

In addition, Figures 1-2, regarding the output and the inputs involved, reveal that:

- Q generally increased over time; was higher in north Greece and Thessaly, and lower in west Greece and the rest of Greece. (In the latter two regions it evolved very much the same from 2008 onwards.)
- L generally decreased over time; was lower in Thessaly in 2004-2006, in north Greece in 2007-2010, and in west Greece in 2011-6, and higher in the rest of Greece.
- K increased over time, much in the same manner across all regions, and was higher in north Greece, modest in Thessaly, and lower in west Greece during 2004-10, and in both west Greece and the rest of Greece during 2011-16. (It seems that in 2010-11, the government attempted to direct to the primary sector an exceptional amount of EU- funding (Giouroukeli, 2010). The 75% increase observed between

Table 1. Descriptive statistics of agricultural production in Greece, 2004-16 (average holding, annual data)

Variables	north country ⁱ		west country ⁱⁱ		Thessaly		Rest of Greece	
	mean	std.dev	mean	std.dev	mean	std.dev	mean	std.dev
1. Output (Q) in €	23904	1864	20731	1627	23949	1785	22201	1403
2. Labor (L) in full-time person equivalent (FTPE)	1.13	0.09	1.16	0.19	1.18	0.12	1.35	0.10
3. Terrain (T)								
a. Value in €(T _a)	51672	9370	77158	19999	68001	5196	73184	9601
b. Utilized area in hectares (T _b)	11.31	0.85	6.34	1.00	10.48	0.97	7.85	1.80
c. Value per hectare: (a)/(b)	4547	690.0	12036	1610.9	6535	682	9551	1217
4. Capital, investment, costs in €								
a. Stock of buildings (K ₁)	8086	1104	7467	3522	5474	1071	8734	1654
b. Stock of machinery (K ₂)	24868	6300	12690	4810	20189	5840	13519	3404
c. New buildings-machinery (I)	677	175	526	120	607	380	942	191
d. Costs for feeds, seeds etc. (M)	8251	935	5011	874	7470	755	6014	470
e. Breeding livestock converted in head units	6.10	0.56	5.30	0.52	6.37	0.64	7.16	0.35
f. Breeding livestock in €	4683	6459	2894	216	4161	63	4171	174
g. Cost of energy (E) in €	2371	609	1091	332	2363	559	1489	348

Notes: ⁱ The north country consists of Western, Central and Eastern Macedonia, Western Thrace, the islands of Thasos and Samothraki.

ⁱⁱ The west country consists of Epirus, the Ionian islands, and the Peloponnese.

Source: own calculations based on the annual FADN data.

Figure 1. *Agricultural inputs and output in Greece, 2004-16 (average holding, annual data)*

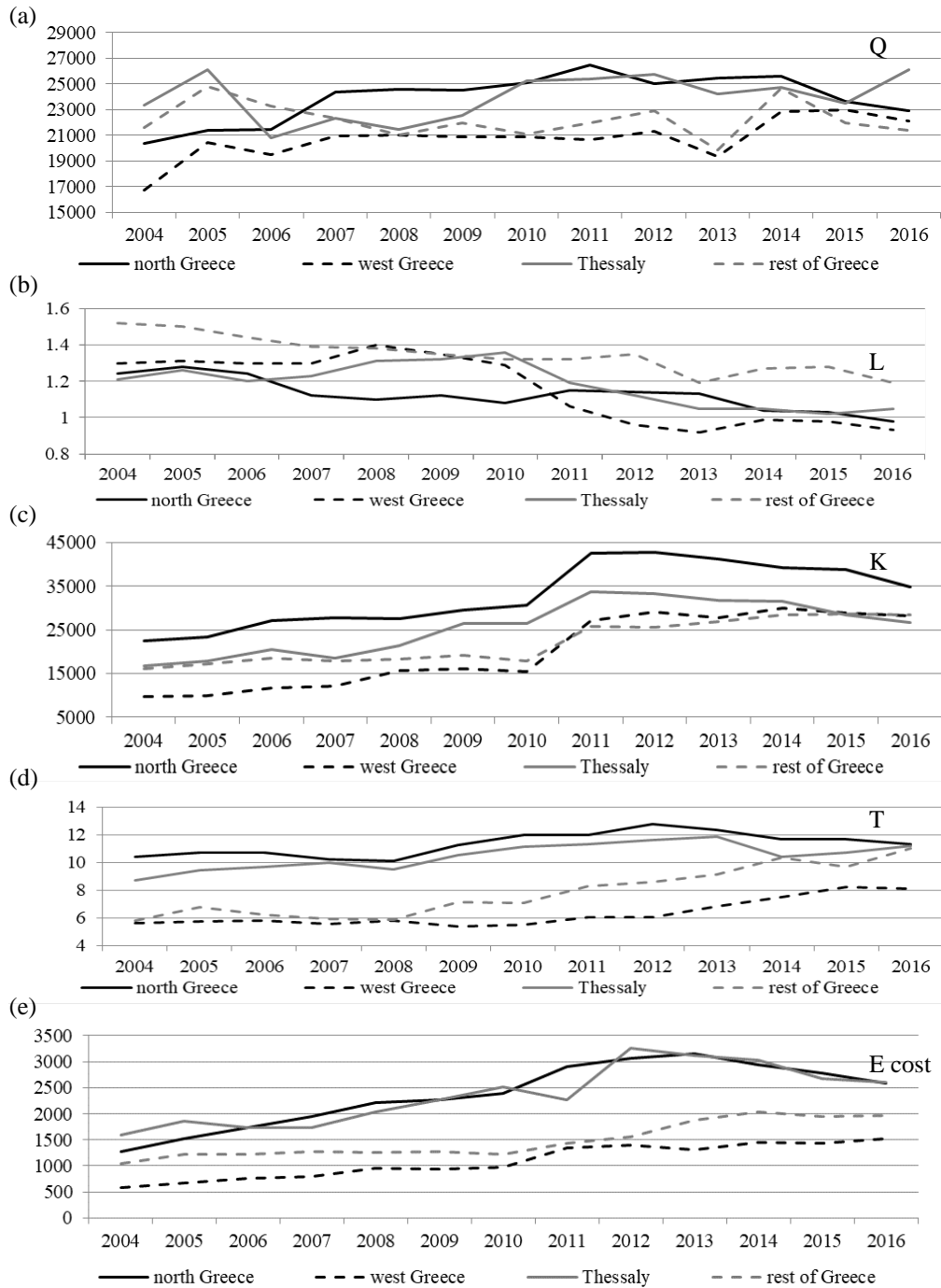
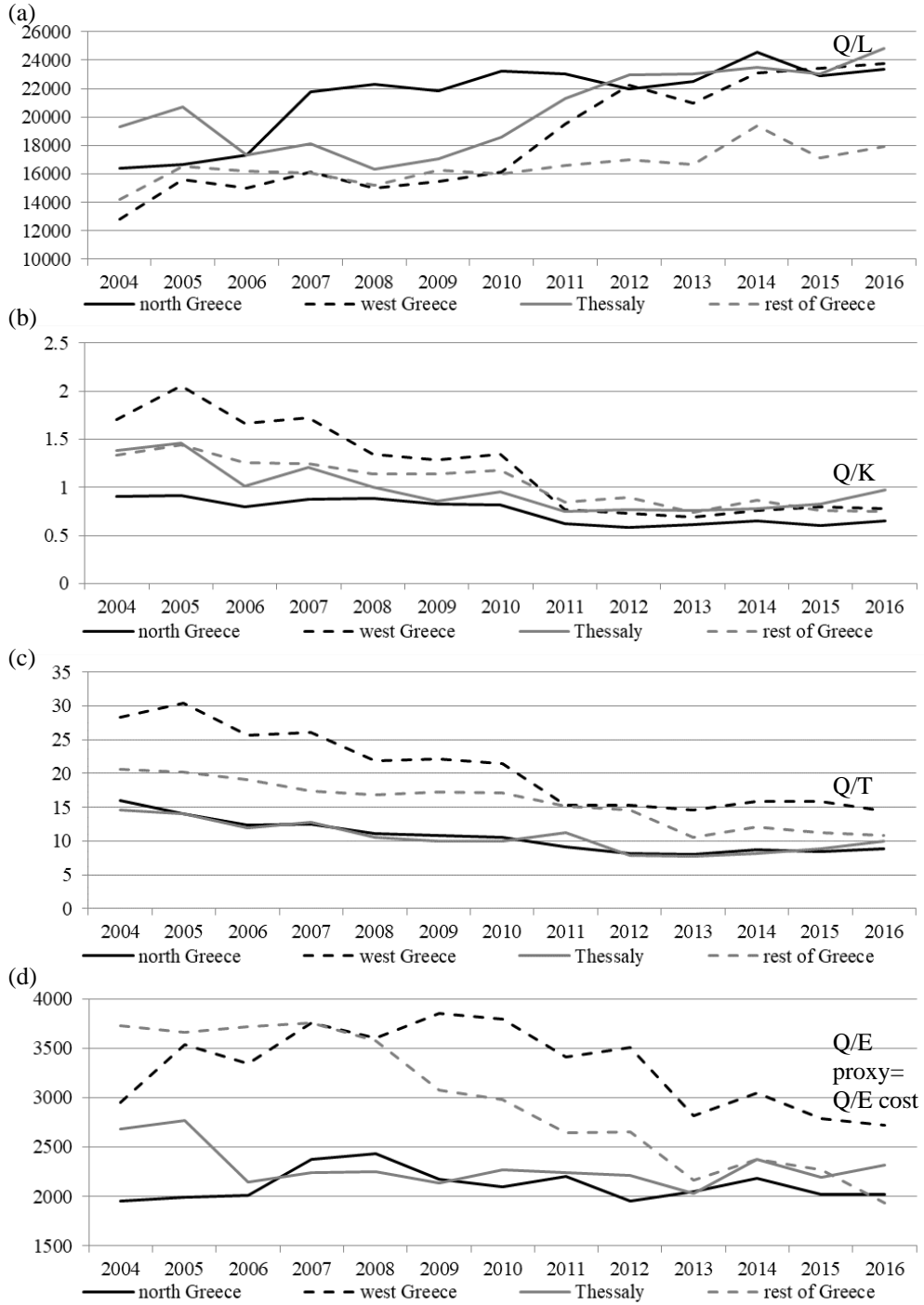


Figure 2. Productivity in agricultural across Greece, 2004-16 (average holding, annual data)



2010 and 2011 in western Greece corresponds to the largest change observed across all inputs.) Given L 's decrease and K 's increase, obviously the sector moved from a more labor-intensive (less capital-intensive) to a less labor-intensive (more capital-intensive) state.

- T_b generally increased over time; was lower in west Greece throughout the period, and in the rest of Greece during 2004-08; and higher in north Greece, Thessaly, and (at the end of the period) in the rest of Greece.
- The cost of E generally increased over time; was higher in north Greece and Thessaly, lower in west Greece, and modest in the rest of Greece.
- Labor productivity (Q/L) increased over time; was higher in north Greece, Thessaly (esp. in 2004-6, 2012-16) and west Greece (in 2012-16), modest in Thessaly (in 2007-11) and west Greece (in 2010-11), and lower in west Greece (in 2004-10) and the rest of Greece.
- Capital productivity (Q/K) decreased over time; was lower in north Greece, and higher in west Greece (up to 2010), and in Thessaly and the rest of Greece (subsequently).
- Land productivity (Q/T_b) generally decreased over time, was higher in west Greece from 2008 on, and the rest of Greece in 2004-8, modest in the rest of Greece in 2009-12, and lower in north Greece and Thessaly throughout the period, and in the rest of Greece during 2013-6.
- The proxy for energy productivity (Q/E cost) decreased over time, was higher in west Greece, lower in Thessaly and north Greece (it evolved very much the same), and modest in the rest of Greece.

3. MODELING ISSUES

To properly look in the impact of these inputs on production we turn to econometrics. First, we select a model that rules out production on the basis of a single input. The simplest approach is to employ an extended Cobb-Douglas (CD) expression (Charnes et al. 1976), for instance,

$$Q = A L^a K^b T^c E^d u, \quad (1)$$

with u denoting the error term, the other lower case letters standing for input coefficients, and A standing for multifactor productivity (MFP), i.e., the element that captures (a) the impact of entrepreneurship and technology (E&T) in combining the inputs involved, as well as (b) the impact of other factors not specified in the expression (e.g., Chiang, 1984; Duffy and Papageorgiou, 2000; Erken et al., 2016).

Next, we consider the issue of stationarity (or non-stationarity), that is, if the mean and variance of the variables involved do not vary (or vary) over time. (This is important as certain analytical tools and models employed in a number of studies –not in this one– rely on the mean and variance being fixed.) To deal with non-stationarity one may (a) incorporate in the expression a trend term, t , (b) detrend the variables in-

volved or (c) rely on successive differences (Maddala, 2001; Gujarati, 1995). All three are technically simple, though, in order to preserve degrees of freedom (DoF), we will not proceed beyond first differences. (We are agnostic as to which method is preferable in this situation, so we probe them all. The findings of the relevant tests are supplied in the notes of Tables 2-5.) To incorporate a trend, and as matters in each region may have evolved in a different way, a variant of expression (1) with factors that may vary over time ($t=1, \dots, 13$), and across space ($i=1, \dots, 4$), in order to suit the data, is considered:

$$Q_{it} = A_{it} L_{it}^a K_{it}^b T_{it}^c E_{it}^d u_{it}. \quad (2)$$

Following the example of J. Tinbergen (Nobel Prize laureate of 1969) and allowing A to vary over time (Wallis, 1973), A_{it} is specified as $A_{it} e^{r_i + \lambda_i t}$, with t commencing at 1:

$$Q_{it} = A_{it} e^{r_i + \lambda_i t} L_{it}^a K_{it}^b T_{it}^c E_{it}^d u_{it}. \quad (3)$$

Equivalently:

$$\ln Q_{it} = \ln A_{it} + (r_i + \lambda_i t) + a \ln L_{it} + b \ln K_{it} + c \ln T_{it} + d \ln E_{it} + u_{it}. \quad (4)$$

To capture cross-input effects and each input's rate of change, additional terms in multiplicative and squared value form, may be considered. In the case of incorporating the squared value of the regressor associated with labor, the expression takes the following form:

$$\ln Q_{it} = \ln A_{it} + (r_i + \lambda_i t) + a \ln L_{it} + g(\ln L_{it})^2 + b \ln K_{it} + c \ln T_{it} + d \ln E_{it} + u_{it}. \quad (5)$$

The employment of a full-blown translog formula with a good number of additional product terms, like the squared term, is not pursued so as to allow for more DoF.

Last but not least, all explanatory variables are made linearly independent of one another: L from t , K from t and L , T from t , L and K , E from t , L , K , and T ; thus, satisfying a basic assumption regarding the independence of regressors (e.g., by Economidou et. al., 2019, and the literature cited therein). In essence, instead of regressing $\ln Q$ on arguments t , $\ln L_{it}$, $\ln K_{it}$, $\ln T_{it}$, and $\ln E_{it}$, initially $\ln L_{it}$ is regressed on t , an $\ln L_{it}'$ is predicted, and an orthogonal $\ln L_{it}^\circ = \ln L_{it} - \ln L_{it}'$ is estimated; next, $\ln K_{it}$ is regressed on t and $\ln L_{it}^\circ$, an $\ln K_{it}'$ is predicted, and an orthogonal $\ln K_{it}^\circ = \ln K_{it} - \ln K_{it}'$ is estimated, and so on. Thus, we may explain $\ln Q$ in terms of t , $\ln L_{it}^\circ$, $\ln K_{it}^\circ$, and additional regressors estimated in the same manner. Consequently, in the context of Tables 2-5, the second regressor is independent of the first regressor, the third regressor is independent of the former two, and so on. Indeed, insofar as in the first column of Table 3 the second regressor (lines 6-9) is t , all other explanatory variables (regarding the inputs used) are made linearly independent of t . Obviously, the order of the regressors employed affects the size of the estimated parameter (coefficient) for if the order was different, a different number of effects would be subtracted from each explanatory variable. Labor is ordered prior to other inputs, and all orders of other inputs are considered. We provide the one associated with the highest fitness.

4. EMPIRICAL RESULTS

The analysis is carried out in Stata. Four models are considered:

- Model A is based on expression (5).
- Model B is a variant of expression (5) that employs detrended variables. (These are explanatory variables regressed on t , on the basis of which the trend element is subtracted.)
- Model C is based on the first differences of expression (2).
- Model D is based on the first differences of expression (5).

Each model is estimated in panel and pooled format. In the case of the former, the random effects (RE) variant is always preferred over its fixed effects (FE) counterpart on the basis of Hausman (1978) test (the respective p-values of the null hypotheses are 0.6138, 0.1778, 0.9830, 0.9891); and in the case of the latter, spatial dummies are employed so that individual regional effects (if any) may be spotted. To deal with heteroscedasticity, all analyses are carried out with robust standard errors (Huber, 1967; White, 1984). The best fits of the both the RE and the pooled analysis results are presented in Tables 2 and 3, respectively.

From the RE results associated with low p-values (less than 1%), the coefficients regarding L (in Models B-D), T_a , the cost of E (Models B-C), and the cost of feeds, seeds, fertilizers etc. (in Model A) are associated with a positive sign; while the coefficient regarding K_I (in Model D) is probably associated with a negative sign (the p-value = 0,011). This suggests that, *ceteris paribus*, to increase Q it might be better if K_I were reduced.

Model A is not likely to satisfy stationarity proponents, while Model C (which is based on a linear arrangement of terms, of the sort $dQ = a dL + c dT + d dE + u$) implies that the production function is linear and, hence, irreconcilable with the multiplicative CD setting of expression (1). Indeed, it suggests that inputs are perfect substitutes and, hence, allows for production entirely without L (and/or other inputs). As Model C is associated with the lowest R^2 , it turns out to be the weakest not only on theoretical grounds but also in terms of empirical fitness. Model D has both a typical CD component and a linearly added component (in particular: $Q = 1.014 L^{0.963} T^{0.227} K_I^{-0.163} + e^{4.737(\ln L)(\ln L)}$) neither of which allows for zero inputs. Overall, the estimated coefficients of Models A and B add up to less than one, while the estimated coefficients of Model D to more than one. This suggests decreasing returns to scale (DRTS) and increasing returns to scale (IRTS), respectively.

The pooled analysis (Table 3) provides more information at the regional level. The results associated with low p-values (less than 1%) reveal:

- Considerable E&T heterogeneity across space and time: The term associated with A_{it} is higher in Thessaly (in Models A-B), and the rest of Greece (in Model A), and increased over time in west Greece, Thessaly, and north Greece (in the latter case at decreasing rate) (in Model A).
- That the impact of L , the cost of E (in all Models), K (in Model A), the flow of investments (in Model B), and T_a (in Model C) varied across space.

As in the RE analysis, the model which is irreconcilable with the CD setting and, hen-

Table 2. Random effects GLS regressions with robust standard errors on holding output across Greece, 2004-16

Explanatory variables	Model A ⁱ log-log	Model B log-log detrended	Model C first differ- ences	Model D log-log first differences
1 Constant	10.056	9.808	299.261	0.014
2 Labor in FTPE (L)	-0.169	0.362	17,023.100	0.963
3 Labor in FTPE, squared			66,297.340	4.737
4 Costs for feeds, seeds etc. (M)	0.316			
5 Value of land (T _a)		0.382	0.079	0.227
6 Stock of buildings (K ₁)				-0.163
7 Cost of energy €	0.124	0.175	1.699	
St. Dev. (u)	0	0	0	0
St. Dev. (e)	0.060	0.049	1667.042	0.074
Rho (fraction of variance due to u)	0	0	0	0
Observations (N)	52	52	48	48
Number of groups	4	4	4	4
Model fitness (R ²)				
• within	41%	25%	33%	34%
• between	100%	99%	90%	31%
• overall	64%	81%	33%	34%

Notes: The second regressor is linearly independent of the first regressor, the third regressor is linearly independent of the former two regressors, and so on. The Levin et al. (2002), Harris and Tzavalis (1999), Breitung (2001), Im et al. (2003), Hardi (2000) tests are considered. All variables of Models B-D are stationary. P-values: Model A: (2nd line) 0.011, (7th line) 0.032. Model C: (5th line) 0.004, (2nd line) 0.005, (1st line) 0.020, (3rd line) 0.059. Model D: (6th line) 0.011, (5th line) 0.024, (1st line) 0.044. The rest are equal to 0.000.

ⁱ Some variables are not stationary.

Source: As in Table 1.

ce, does not fit well with economic theory (Model C) is empirically associated with the lowest R². In the other three models, the estimated input coefficients—involving *L* and the variable associated with *E* (also, a good proxy for the active use of *K*₂) plus one other factor— seem to add up to more than one in the rest of Greece; but may or may not add to less than one in the other regions. This suggests IRTS in the former region and unclear returns to scale in the other regions. Model D recalls a RE result by yielding a negative effect for *K*₁ in west Greece (the relevant p-value is 0.030). However, Model B, is associated with the highest goodness of fit (91%) and more DoF vis-a-vis the other pooled data analyses carried out, so perhaps best captures and advances our understanding on how the country's agricultural production economy operated during the period in question. Consequently, in the following paragraphs we study it more detail.

Table 3. OLS regressions with robust standard errors of pooled data on agricultural production across Greece, 2004-16

Explanatory variables		Model A log-log with a trend term	Model B log-log detrended	Model C first differences	Model D log-log first differences
1	Constant	9.898	9.547	476.387	0.023
2	Regions I-II (north Greece, west Greece)	ref.	ref.		ref.
3	Region III (Thessaly)		0.464		ref.
4	Region IV (rest of Greece)		ref.		ref.
5	Regions III-IV	0.082			
6	Time trend in region I	0.058			
7	regions II-III	0.010			
8	region IV	0.000			
9	region I squared	-0.003			
10	Labor (in work units) in region I		2.545	-1813.958	-0.059
11	regions I, III	-0.141			
12	region II	0.414			
13	regions II-III			9601.630	0.533
14	regions II, IV		1.108		
15	region III		0.037		
16	region IV	1.605		23500.970	1.394
17	region I squared		-5.538		
18	region II squared	-3.441	-0.843		
19	Cost of energy in region I			5.068	0.502
20	regions I-II, IV	0.423			
21	regions I-IV		0.344		
22	regions II-III			1.582	0.190
23	region III	0.364			
24	region IV			8.900	0.593
25	region II squared	-2.745	-4.431		
26	region IV squared	3.559			
27	Stock of capital (buildings & machinery) in regions I-II, IV	0.009			
28	region III	-0.723			
29	region III squared	3.900			
30	Investments in buildings & machinery in regions I-II		-0.158		
31	regions III-IV		-0.322		
32	region IV squared		14.990		
33	Value of terrain in region I			0.043	
34	region II			0.092	
35	region III			0.029	
36	region IV			0.091	
37	Stock of buildings in regions I, III-IV				-0.150
38	region II				-0.453
	Observations (N)	52	52	48	48
	Model fitness (R^2)	87%	91%	45%	49%

Notes: Regions that feature similar coefficients are grouped together so as to preserve DoF. All tests mentioned in Table 2 are carried out. All variables are stationary (exception: the regressands of Models A, C-D are stationary, respectively, in two, four, four out of five tests.)

Table 3 (continued)

P-values per model (the parentheses specify lines): A: (5), (23), (28) 0.001; (12), (26) 0.003; (16) 0.005; (25) 0.006; (18) 0.027; (11) 0.412; (8) 0.912; (27) 0.948. B: (32) 0.040; (25) 0.047; (30) 0.544; (15) 0.833. C: (34) 0.001; (24) 0.016; (13) 0.036; (36) 0.049; (1) 0.050; (33) 0.069; (22) 0.159; (10) 0.622; (35) 0.737. D: (16) 0.001; (24) 0.021; (13) 0.029; (1), (38) 0.030; (22) 0.141; (37) 0.262; (10) 0.751. The rest are equal to 0.000.

Source: As in Table 1.

Table 4. A variant of Model B resulting from an OLS regression with robust standard errors on pooled agricultural production data in Greece, 2004-16

Explanatory variables	coefficients	
1 Constant	9.791	Note 1: north Greece, west Greece, Thessaly, the rest of Greece correspond to regions 1-IV, respectively. Note 2: Regions that feature similar coefficients are grouped together. All tests mentioned in Table 2 are carried out. All variables are stationary. Note 3: P-values (the parentheses specify lines): (8) 0.001, (4) 0.016, (6) 0.047, (9) 0.072, (7) 0.268, (10) 0.314. The rest are equal to 0.000.
2 Regions I-II, V	ref	
3 Region III	0.193	
4 Labor (in work units) in regions I-III	0.130	
5 region IV	0.571	
6 Cost of energy in region I	0.479	
7 region II	0.294	
8 region III	0.314	
9 region IV	0.337	
10 Investments in build. & machin. in regions I-II, IV	-0.218	
11 region III	-0.315	
Observations (N)	52	Source: As in Table 1.
Model fitness (R^2)	88%	

According to the findings associated with p-values below 1%, *ceteris paribus*, a marginal increase in (a) L had a negligible output effect in Thessaly and a positive output effect in north, west and rest of Greece; (b) the energy bill had a positive output effect throughout the country; (c) capital investments for buildings and machinery had a negative output effect in Thessaly and the rest of Greece. On the other hand, as mentioned in the beginning of the paragraph, Thessaly exhibited higher levels of MFP.

Table 4 provides a variant specification of Model B in a simple multiplicative form associated both with more DoF, and a slightly lower R^2 . (The adj. R^2 is 85%. In the previous case it was 88%.) Again, the findings suggest E&T heterogeneity across space (higher in Thessaly), and that a marginal increase in capital investments for buildings and machinery had a negative output effect in Thessaly. At the same time: (i) a marginal increase in the energy bill had a positive output effect in Thessaly, (ii) a marginal increase in L had a positive output effect in the rest of Greece, and (iii) all regions operated under DRTS.

Table 5. A variant of Model B resulting from an OLS regression with robust standard errors on deflated pooled agricultural production data in Greece, 2004-16

	Explanatory variables	coefficients	
1	Constant	10.0141	
2	Regions I-II, V	ref	
3	Region III	0.366	
4	Labor (in work units) in region I	2.492	
5	region II	0.424	
6	region III	0.219	
7	region IV	1.038	
8	region I squared	-5.268	
9	Cost of energy in region I	0.296	
10	regions II-III	0.257	
11	region IV	0.390	
	Deflated stock of machines in		
12	regions I-II, IV	0.073	
13	region III	0.335	
	Observations (N)	52	
	Model fitness (R^2)	86%	

Notes 1 and 2: As in Table 4.
 Note 3: P-values (the parentheses specify lines): (4) 0.001, (13) 0.003, (8) 0.007, (5) 0.013, (10) 0.014, (11) 0.036, (9) 0.089, (6) 0.155, (12) 0.313. The rest are equal to 0.000.

Source: As in Table 1.

Table 5 provides the best alternative in terms of model fitness, when deflated values of Q and K are considered. (Its R^2 and DoF are slightly inferior to that of Table 4.) Once again, the findings suggest E&T heterogeneity across space (higher in Thessaly). However, in this setting: (a) a marginal increase in L had a positive effect, not only in the rest of Greece, but also in north Greece; (b) a marginal increase in K_2 had a positive effect in Thessaly; (c) the north part of Greece, the west part of Greece, and Thessaly operated under DRTS, while the rest of Greece operated under IRTS.

To gain additional insights we also look into the Granger (1969) causality tests among inputs, both within and across regions, based on first differences. The tests suggest that, at the 0,01 probability of error, in Thessaly changes in the stock of buildings steadily anticipated changes in the value of livestock, and changes in the size of lives-

Table 6. Granger causalities with one lag on first differences of agricultural production inputs across holdings in Greece (2004-16)

Null hypothesis	Region	p-values < 0.01%
Stock of buildings does not Granger cause livestock value	Thessaly	0.0064
Livestock (in heads) does not Granger cause labor	Thessaly	0.0090
Holding's land value does not Granger cause labor	Rest of Greece	0.0005
New buildings-machinery do not Granger cause livestock (in heads)	Rest of Greece	0.0092

Source: As in Table 1.

tock (number of heads) steadily anticipated changes in labor; while in the rest of Greece changes in land value steadily anticipated changes in labor, and changes in building and machinery investments steadily anticipated changes in livestock size (Table 6).

5. CONCLUSIONS

The paper reveals considerable heterogeneity in agricultural production across Greece and advances our understanding on how the country's holdings operated at the regional level. It turns out that during 2004-16, the average holding generally decreased the amount of labor used, and increased its output, capital, use of land, and the money paid for energy. As a result, it may have not been able to absorb workers discharged from other sectors during the recession; and as it shifted from a more to a less labor-intensive state of production, its labor productivity increased while its capital and land productivity, and the ratio of output over energy expenses decreased. In addition, the paper identifies a number of differences in the features of the average holding in each region; and finds that the model which is irreconcilable with the Cobb-Douglas (i.e., Model C) is also associated with the lowest empirical goodness of fit; while the econometric analyses which are based on three Cobb-Douglas or Cobb-Douglas-like models (i.e., Models A-B, D and two variants of B) suggest that the impact of E&T and of other factors (i.e., multifactor productivity) was, probably, higher in Thessaly (an outcome observed in model B and its variants, where the issue is considered); while a marginal increase in labor had a positive output effect in the rest of Greece (the outcome was observed in all five models), and a marginal increase in the energy bill (and, probably, the use of energy) had a positive output effect in the north part of the country (models A-B, D) and Thessaly (models A-B and one of B's variants). These imply that there is room for spatially differentiated interventions. In addition, they suggest that it might be sensible if E&T practices carried out in Thessaly were considered in other regions, and reductions (increases) in the cost of energy and in labor in agricultural production took place if they affected higher output increases (smaller output reductions) elsewhere in the economy. However, the need for more specificity regarding individual activities/products will require different data and additional analyses.

ΠΕΡΙΛΗΨΗ

Παρουσιάζονται βάσει ευρωπαϊκών στοιχείων οι εξελίξεις των εισροών, της παραγωγικότητας και της αγροτικής παραγωγής σε τέσσερις εδαφικές διαρέσεις της χώρας. Εκτιμώνται οικονομετρικά οι παράμετροι των εισροών και η πολυπαραγοντική παραγωγικότητα. Η τελευταία και οι επιδράσεις εργασίας και ενέργειας (αναδύονται ως βασικοί ερμηνευτικοί παράγοντες) διαφοροποιούνται στο χώρο, αναδεικνύοντας την ανάγκη διαφοροποιημένων παρεμβάσεων.

Acknowledgements: The paper was supervised by P. Prodrromidis (KEPE). The helpful comments of an anonymous referee are greatly appreciated.

REFERENCES

- Baráth L., Fertő I. 2017. Productivity and Convergence in European Agriculture. *Journal of Agricultural Economics*, **68**, 228-248.
- Breitung J. 2001. The local power of some unit root tests for panel data. *Advances in Econometrics*. Vol. 15: *Nonstationary Panels, Panel Cointegration, and Dynamic Panels*. Ed. by B. Baltagi, T. Fomby, R. Carter Hill. Bingley UK: Emerald, 161-178.
- Charnes A., Cooper W.W., Schinnar A.P. 1976. A theorem on homogeneous functions and extended Cobb-Douglas forms. *Proceedings of the National Academy of Sciences of the United States of America*, **73**, 3747-3748.
- Chiang A.C. 1984. *Fundamental methods of mathematical economics*. 3rd ed. New York: McGraw-Hill.
- Duffy J., Papageorgiou C. 2000. A Cross-Country Empirical Investigation of the Aggregate Production Function Specification. *Journal of Economic Growth*, **5**, 87-120.
- Economou F., Prodrromidis P., Skintzi G. 2019. Large fire disaster and the regional economy: the 2007 case of the Peloponnese. *South-Eastern Europe Journal of Economics*, **17**, 7-31.
- Erken H., Donselaar P., Thurik R. 2016. Total factor productivity and the role of entrepreneurship. *Journal of Technology Transfer*, **43**, 1493-1521.
- European Commission. 2018. *CAP Context Indicators 2014-2020*. Brussels.
- Fuglie K., Clancy M., Heisey P., Macdonand J. 2017. Research, Productivity, and output growth in U.S. Agriculture. *Journal of Agricultural and Applied Economics*, **49**, 514-554.
- Giannakis E., Bruggeman A. 2017. Economic crisis and regional resilience: Evidence from Greece. *Papers in Regional Science*, **96**, 451-477.
- Giouroukeli M. 2010. Two billion for agricultural investment by the end of 2011. *Stockwatch*. (In Greek.) Issue of Nov. 4.
- Granger C.J. 1969. Investigating Causal Relationships by Econometrics Models and Cross Spectral Methods. *Econometrica*, **37**, 424-438.
- Gujarati D.N. 1995. *Basic Econometrics*. New York: McGraw-Hill.
- Hadri K. 2000. Testing for stationarity in heterogeneous panel data. *Econometrics Journal*, **3**, 148-161.
- Harris R.D.F., Tzavalis E. 1999. Inference for unit roots in dynamic panels where the time dimension is fixed. *Journal of Econometrics*, **91**, 201-226.
- Hausman J.A. 1978. Specification tests in econometrics. *Econometrica*, **46**, 1251-1271.
- Huber P.J. 1967. The behavior of maximum likelihood estimates under nonstandard

- conditions. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol.1, 221–233.
- Im K.S., Pesaran M.H., Shin Y. 2003. Testing for unit roots in heterogeneous panels. *Journal of Econometrics*, **115**, 53–74.
- Levin A., Lin C.-F., Chu C.-S.J. 2002. Unit root tests in panel data: Asymptotic and finite-sample properties. *Journal of Econometrics*, **108**, 1–24.
- Maddala G.S. 2001. *Introduction to Econometrics*, 3rd ed. Chichester: Wiley.
- Nowak A., Kijek T. 2016. The effect of effect of human capital on labour productivity of farms in Poland. *Studies in Agricultural Economics*, **118**, 16-21.
- White H. 1984. *Asymptotic theory for econometricians*. Orlando: Academic press.



EVIDENCE SYNTHESIS USING BAYES FACTORS

Stavros Nikolakopoulos and Ioannis Ntzoufras

Department of Statistics, Athens University of Economics and Business
sknikolak@gmail.com

ABSTRACT

The problem of evidence synthesis can arise in (at least) two areas of Bayesian analysis. One is the case of (very) large sample sizes where standard Bayesian computational tools (such as MCMC) do not scale up and thus a way around it would be to split the data into sub-samples and then combine the aggregate-level statistics. The other is meta-analysis of published results where researchers do not have access to individual subjects' data. This paper investigates methods for synthesizing data evidence which are quantified in terms of Bayes Factors. By harvesting recent analytical results concerning the relationship between the Bayes Factor and the Likelihood Ratio statistic, we evaluate evidence synthesis techniques for variable selection problems. We do so for the case of linear regression when employing g-priors. We demonstrate the performance of the methods suggested in the simple case of a t-test. We also discuss more complicated problems related to variable selection.

Keywords: Bayes Factors, Evidence Synthesis, Meta Analysis, Likelihood Ratio

1. INTRODUCTION

In evidence-based empirical sciences, research results arise from a variety of sources and accumulate in an all-increasing rate. Therefore, evidence synthesis is of paramount importance. In the context of synthesizing aggregate data, the field of meta-analysis has received significant attention from the statistical community, mostly within the classical (frequentist) framework of inference.

In classical statistics, hypothesis testing related inference is typically based on p -values, and on point or interval parameter estimates when the focus is estimation. Thus, for the meta-analytic problem of having K statistics $T = \{T_1, \dots, T_K\}$ estimating the same parameter θ from K independent datasets, the meta-analytic estimate of θ is of the general form $\mathcal{J}_{MA} = f_t(T)$. Each statistic can be transformed to a p -value for a given H_0 , $p_i = Pr(T_i \geq T_i^{obs} | H_0)$ resulting to a respective vector of p -values $\mathbf{p} = \{p_1, \dots, p_K\}$. Meta-analysis concerned with hypothesis testing can be based on an

overall p -value $\mathcal{P}_{MA} = f_p(\mathbf{p})$ taking into account the synthesis of evidence arising from the individual studies.

Research focusing on frequentist meta-analysis is vast and a review is beyond the scope of this paper. Methods for combining p -values date back to [Pearson(1933)] and [Fisher (1934)]. For a recent overview and methodological developments see [Heard and Rubin-Delanchy (2018)]. For a more general overview of meta-analysis methods, see [Hedges and Olkin (1985)].

The shortcomings of p -values as a measure of evidence have long been the subject of a heated discussion among statisticians. Their inability to quantify evidence *for* a null hypothesis as well as their overstatement of evidence against it have been overly stressed in the statistical literature [Berger and Sellke (1987), Selke, Bayarri and Berger (2001) Held and Ott (2017)]. Therefore, for a meta-analysis focusing on hypothesis testing, combining p -values as a means for quantifying overall evidence suffers from the same shortcomings.

Within the Bayesian framework, hypothesis tests are simply viewed as a pairwise model comparison problem. The most prevalent Bayesian tool for comparing two hypotheses or models is the Bayes Factor (BF) [Jeffrey (1961), Kass and Raftery (1995)]. From a scientific inquiry point of view, BF s have the significant conceptual advantage of providing evidence both for and against a null hypothesis. This feature can be of paramount importance since null hypotheses often correspond to theoretically useful statements of invariance and constraint [Kass(1992)]. One notable example is a series of experiments in psychology testing the existence of so-called extrasensory perception (ESP), the claim that people can feel or sense salient events in the future that could not otherwise be anticipated [Bem (2011)]. In that case, the null hypothesis is the theoretically attractive, reasonable, and highly interpretable constraint that ESP does not exist [Rouder and Morey (2011)].

Synthesis of BF s arising from individual datasets is therefore of direct concern in the meta-analysis of published research. The difficulties arising in the computation of BF s makes meta-analytic methods also very relevant in data partitioning for Big Data when computations become *CPU-intensive* as well as *data-intensive* and cannot be handled by a single machine [Zhao et al (2013)]. P -value combination and other meta-analytic methods have been proposed in this field in order to tackle parallel processing of data partition for frequentist analyses [Tsamardinos et al (2019)].

Nevertheless, research on the methods for synthesizing BF s has been very limited. Maybe this can be attributed to the fallacious intuition that the BF resulting from a series of BF s comparing the same two models is merely the product of those BF s, by sequential updating of the BF s [Goodman (2016), Held and Ott (2017)]. This intuition that the meta-analytic BF is the product of individual BF s is not correct [Rouder and Morey (2011)].

In this paper we present a method for calculating the overall BF from a set of available BF s that compare the same two models. We focus on linear regression models and demonstrate the method for the case of a t -test. We utilize recent results that provide an analytical relationship between the BF and the Likelihood Ratio (LR) in combination with meta-analytic methodology. Section 2 provides a general framework and priors setting for the models explored, the combination methods assessed, as well the design of a simulation study for a simple t -test setting. Section 3 describes the results while the paper ends with a discussion.

2. METHODS

2.1 Bayes Factor and Likelihood Ratio

For a given set of response values \mathbf{y} and two competing models $M_0 = \{f(\cdot | \theta_0, M_0), \theta_0 \in \Theta_0\}$ and $M_1 = \{f(\cdot | \theta_1, M_1), \theta_1 \in \Theta_1\}$ the Bayes Factor B_{01} is the ratio of the marginal densities of \mathbf{y} under M_0 and M_1 given by

$$BF_{01} = \frac{f(\mathbf{y}|M_0)}{f(\mathbf{y}|M_1)} = \frac{\int_{\Theta_0} f(\mathbf{y}|\theta_0, M_0)\pi(\theta_0|M_0)d\theta_0}{\int_{\Theta_1} f(\mathbf{y}|\theta_1, M_1)\pi(\theta_1|M_1)d\theta_1},$$

where $\pi(\theta_\ell|M_\ell)$ represents the prior distribution of θ_ℓ under model M_ℓ for $\ell \in \{0,1\}$. The BF does not depend on the prior probabilities of the models and it is usually interpreted as the odds provided by the data for M_0 to M_1 . However, BF_{01} obviously depends on the prior distributions of the model parameters. Large values of BF_{kl} indicate strong posterior support of model M_k against model M_l ; for details see, e.g. [Kass and Raftery (1995)].

The likelihood ratio statistic for a null hypothesis $H_0: \theta \in \Theta_0$ and alternative $H_1: \theta \in \Theta_1$ for $\Theta_0 \subseteq \Theta_1$ is defined as

$$\Lambda_{01} = 2 \ln \left[\frac{\sup_{\theta \in \Theta_1} \mathcal{L}(\theta)}{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta)} \right]$$

where $\mathcal{L}(\cdot)$ denotes the likelihood function. Commonly, Λ_{01} is represented as a function of the MLE of θ , $\hat{\theta}$ ($\ell(\cdot)$ denoting the log-likelihood function)

$$\Lambda_{01} = -2[\ell(\theta_0) - \ell(\hat{\theta})]$$

a functional form which stresses the difference with BF s: In order to calculate the BF one has to integrate the likelihood function over the respective prior distribution for each model, while Λ calls for conditional evaluation of the likelihood function given the

MLE. For a comprehensive overview concerning the relationship as well as approximations between BF s and A see [Kass and Raftery (1995)].

2.2 Bayes Factors in Linear Regression

We consider the standard general hypothesis testing setting in linear regression with normal independent errors:

$$\begin{aligned} H_0(M_0): y|\mathbf{a}, \tau &\sim MVN(\mathbf{W}\mathbf{a}, \tau^{-1}\mathbf{I}_n) \\ H_1(M_1): y|\mathbf{a}, \boldsymbol{\beta}, \tau &\sim MVN(\mathbf{W}\mathbf{a} + \mathbf{L}\boldsymbol{\beta}, \tau^{-1}\mathbf{I}_n) \end{aligned} \quad (1)$$

where MVN stands for the multivariate normal distribution, \mathbf{I}_n is an $n \times n$ identity matrix, \mathbf{W} is a full-rank $n \times q$ matrix including the nuisance covariates, including a column of $\mathbf{1}$, \mathbf{a} is a q -vector, \mathbf{L} is an $n \times p$ matrix including the covariates of interest, $\boldsymbol{\beta}$ is a p -vector and τ^{-1} is the error variance. The null model (M_0) is nested within M_1 and represents no effect of \mathbf{L} .

For this parameterization of the linear regression problems there are three forms of conjugate priors. The normal-inverse-gamma prior, NIG, [O'Hagan and Forster (2004), chap. 9] is formulated as:

$$\begin{aligned} \mathbf{a}|\tau &\sim MVN(\mathbf{0}, \tau^{-1}\mathbf{V}_a) \\ \boldsymbol{\beta}|\tau &\sim MVN(\mathbf{0}, \tau^{-1}\mathbf{V}_\beta) \\ \tau &\sim \text{Gamma}(k_1/2, k_2/2) \end{aligned}$$

where \mathbf{V}_a and \mathbf{V}_β are some positive definite matrices, and the gamma distribution is in the shape-rate parameterization. By letting $\mathbf{V}_a^{-1} \rightarrow 0$ and $k_1, k_2 \rightarrow 0$, BF can be computed by:

$$BF_{10} = |\mathbf{V}_\beta|^{-1/2} |\mathbf{X}^t \mathbf{X} + \mathbf{V}_\beta^{-1}|^{-1/2} \times \left\{ 1 - \frac{\mathbf{y}^t \mathbf{X} (\mathbf{X}^t \mathbf{X} + \mathbf{V}_\beta^{-1})^{-1} \mathbf{X}^t \mathbf{y}}{\mathbf{y}^t \mathbf{y} - \mathbf{y}^t \mathbf{W} (\mathbf{W}^t \mathbf{W})^{-1} \mathbf{W}^t \mathbf{y}} \right\}^{-n/2} \quad (2)$$

where

$$\mathbf{X} = (\mathbf{I}_n - \mathbf{W} (\mathbf{W}^t \mathbf{W})^{-1} \mathbf{W}^t) \mathbf{L}$$

is the residuals of \mathbf{L} after regressing out \mathbf{W} , and $|\cdot|$ denotes the determinant. By imposing $\mathbf{V}_a^{-1} \rightarrow 0$ and $k_1, k_2 \rightarrow 0$ the limiting prior distributions for \mathbf{a} and τ are equivalent to Jeffrey's prior $p(\mathbf{a}, \tau) \propto \tau^{-(q-2)/2}$ [O'Hagan and Forster (2004)] and their posteriors are proper. We assume independence between effects and covariates which implies $\mathbf{V}_\beta = \sigma_\beta^2 \mathbf{I}_p$.

The second conjugate prior is Zellner's g -prior [Zellner (1986), Liang et al (2008)], where:

$$p(\mathbf{a}, \tau) \propto \frac{1}{\tau}$$

$$\boldsymbol{\beta}|\tau \sim MVN(\mathbf{0}, g\tau^{-1}(X^tX)^{-1})$$

This is a special case of the NIG prior with $V_\beta = g(X^tX)^{-1}$ and thus calculation of BF is straightforward from (2). The third conjugate prior is the normal prior which is equivalent to the NIG prior for known τ^{-1} .

The g -prior setup will be the focus of this paper. Alongside popularity, computational efficiency and understandable interpretation of g -priors, resulting BF s have been shown to be monotone functions of the F statistic for testing (1) and therefore gives the uniformly most powerful invariant test [Shively and Walker (2018)].

Equation (2) makes clear that in the case of conjugate analysis for the linear model described in (1), the BF comparing the two models is merely a statistic. As such, it is natural to explore its sampling distribution as well as its relationship with other statistics employed for testing the same hypotheses. [Zhou and Guan (2018)] derive the sampling distribution of BF and show that for the g -prior setting, there is an analytical relationship between BF and Λ :

$$2\log BF = \frac{g}{g+1}\Lambda + p\log\left(\frac{1}{g+1}\right) + o(1) \quad (3)$$

where $o(1)$ vanishes in probability as $n \rightarrow \infty$. This result has several implications. Relevant to the present work is the fact that one can compute a p -value associated with a BF . Furthermore, the task of synthesizing BF s from independent datasets boils down to synthesizing the respective Λ s. Note here that $\Lambda|H_0 \sim \chi_p^2$ [Wilks (1938)].

2.3 Synthesis of Likelihood Ratios & Bayes Factors

Suppose we have K datasets each resulting from the same data generating model. This could be a result of either a (fixed effects) meta-analysis or a data partitioning problem arising from Big Data where $n \gg p$. From each dataset k we have available $BF_{12(k)} = BF_k$ comparing the same models M_1 and M_2 . Each BF is calculated by employing the g -prior setting discussed above. We are interested in $BF_{12} = BF$, the evidence taking into account all K datasets. Analytical derivation is not possible without the raw data. In what follows, we will denote by BF, Λ the true values of the statistics calculated from all raw data, BF_k, Λ_k the calculated statistics from the k -th dataset and by $\widetilde{BF}, \widetilde{\Lambda}$ the synthesized statistics from the respective $\{BF_k\}_{k=1}^K, \{\Lambda_k\}_{k=1}^K$. We are thus interested in methods that result in $\widetilde{BF} \approx BF$.

We will further focus on simple linear regression and more specifically on the t -test as it is the most common form of testing encountered in meta-analyses. Furthermore, it

allows for both computational simplicity as well as demonstrating the problems with naïve synthesis of BF s (multiplication).

Since (3) provides an analytical relationship between the BF and Λ , our attention will be directed to established frequentist methods for synthesizing evidence. A first naïve thought could be that since the samples are independent, likelihood functions could be multiplied in order to give the likelihood of the full dataset. This approach would be analytically correct for known parameter values (i.e. under H_0 or for a point alternative), but not for likelihood functions conditional on MLEs. In terms of the LR statistics, such a synthesis would imply that $\tilde{\Lambda} = \sum_k \Lambda_k$, which is not true. In order for this to become more clear, one can consider that in simple linear regression $\Lambda_k|H_0 \sim \chi_1^2 \Rightarrow \sum_k \Lambda_k \sim \chi_K^2$, while $\Lambda \sim \chi_1^2$. Note that the same analogy holds for BF multiplication.

A detailed example of why the product of BF s resulting from partitioning a dataset is not the same as the BF computed from the pooled data is provided in [Rouder and Morey (2011)]. It is there shown, that for small sample sizes, small effects may be considered evidence for the null model, since it is a more parsimonious description given the resolution provided by the data. But as the sample size grows, small effects are more concordant with the alternative and thus BF favors the alternative. In the specific example, the product of the BF_k equals 0.092, favoring the null with evidence slightly larger than 10:1, while the BF from the whole dataset is equal to 54:1 in favor of the alternative.

2.4 Synthesis of p -values

Base on the result that for the vector of p -values $\mathbf{p} = \{p_1, \dots, p_K\}$ of K independent tests, $p_i|H_0 \sim Uniform(0,1) \forall i \in \{1, \dots, K\}$, several combination methods have been proposed so as to obtain $\tilde{p} = f(\mathbf{p})$, the overall measure of departure from H_0 . Four of the most common statistics, of which the null distribution is known are:

Fisher's method: $S_F = -2 \sum_k \log(p_i) \sim \chi_{2K}^2$ [Fisher (1934)]

Lancaster's method: $S_L = \sum_k F_{\chi_{1/K}^2}^{-1}(1 - p_i) \sim \chi_1^2$ [Lancaster (1961)]

Stouffer's method: $S_S = \sum_k \Phi^{-1}(1 - p_i) \sim N(0, K)$ [Stouffer et al (1949)]

Tippet's method: $S_T = \min\{p_1, \dots, p_K\} \sim Beta(1, K)$ [Tippet (1931)]

Since (3) provides an analytical relationship between Λ and BF , one could explore the following strategy for obtaining \tilde{BF} from BF_k : For each BF_k , calculate the respective Λ_k and the associated p -value; synthesize the resulting vector of $\mathbf{p} = \{p_1, \dots, p_K\}$ via some method of p -values combination in order to obtain the overall \tilde{p} ; Transform \tilde{p} to

an appropriate $\tilde{\Lambda}$ (i.e. for simple linear regression via $\tilde{\Lambda} = F_{\chi_1^2}^{-1}(\tilde{p})$) and this in turn to \tilde{BF} . However, such a strategy would be merely a heuristic and lacks any theoretical justification. P -value combination methods are only meant to result to a statistic that has a known distributional form under the intersection of *all* null hypotheses being tested $H_0 := \cap_{i=1}^K \{H_{0i}\}$. Thus, their purpose is to retain frequentist testing operational characteristics (type I error control) and not to synthesize test statistics. However, even though fundamentally different [Held and Ott (2017)], BF s and p -values are both measures of evidence and thus the exploration of the strategy outlined above is part of the scope of this work.

2.5 Synthesis of non-directional evidence

The general problem presented in (1) is, in hypothesis testing terms, a two-sided one. For simple linear regression and thus a single parameter β , p -values and BF s are calculated for $H_0: \beta = 0$ vs $H_1: \beta \neq 0$. Additionally to the reasons outlined in section (2.3) describing why multiplying BF s is not a valid method of synthesis, the combination of two-sided tests comes with added complications [Owen (2009)].

Consider a set of data that arises from H_0 , that is, the true $\beta=0$ and the covariate X is a dichotomous group indicator, which reduces to the standard t -test. A random split of the data in $K=2$ partitions could result in $\hat{\beta}_1 > 0$ and $\hat{\beta}_2 < 0$, or vice-versa. Depending on the magnitude of the subsamples' deviations from H_0 , this in turn could translate to a pair of small two-sided p_i 's or, equivalently, large test statistics (i.e. $|t|$, F , Λ or BF). Synthesizing any of these two-sided statistics via the above procedures, without taking into account the sign of $\hat{\beta}_i$ would result to a misleading combined estimate. [Owen (2009)] reviews methods for synthesizing two-sided p -values but they do not apply to test statistics.

One approach for circumventing this problem is to take into consideration the sign of the $\hat{\beta}_i$'s for each dataset and (by assuming that all the rest remain the same and the available statistics are in the form of BF_k), calculate the weighted Z scores via:

$$S_{wz} = \sum_{k=1}^K \frac{1}{\sqrt{K}} [\mathbb{1}_{\{\hat{\beta}_{1(k)} \geq 0\}} \times (\sqrt{\Lambda_k}) + \mathbb{1}_{\{\hat{\beta}_{1(k)} < 0\}} \times (-\sqrt{\Lambda_k})] \sim N(0,1)$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function. This formulation takes advantage of the asymptotic equivalence between the LR and the Wald statistics in linear regression [Buse (1982), Engle (1984)]. Subsequently,

$$\tilde{\Lambda} = (S_{wz})^2 \sim \chi_1^2 \tag{4}$$

will give an approximation to the original LR and \widetilde{BF} can be acquired via (3).

2.6 Simulation Study

In order to evaluate the performance of the suggested methods in synthesizing BF_k s, we conducted a simulation study. The data generating model was $y \sim N(\beta X, 1)$, $X \in \{0,1\}$, representing the simple t -test. The total sample was $N=1000$ and we explored $K = \{2,5,10\}$ balanced partitions so that $N_k=N/K$. We conducted $R=1000$ replications per scenario. We assessed the null and levels of moderate departure by varying $\beta \in \{0,0.1,0.2,0.3\}$. For each of the K samples per scenario and iteration, BF_k was calculated as well as BF from the whole dataset. The g -prior setup was used with $g_k=N_k$ and $g=N$. Sticking to the assumption that BF_k is the available statistic from each subsample, we transformed each BF_k to A_k via (3) and calculated the associated p -value. For S_F, S_L and S_T we employed the aforementioned p -value combination methods and transposed the calculated \tilde{p} on a χ_1^2 distribution in order to obtain $\tilde{\Lambda}$ which in turn was turned to \widetilde{BF} via (3). We also calculated $\tilde{\Lambda}$ via (4) for S_{WZ} , utilizing the direction of the effect in each subsample. $\tilde{\Lambda} = \sum_k \Lambda_k$ was also assessed for demonstrative purposes.

In order to assess the performance of each method, and since the p -value combination methods are not expected to perform adequately in terms of reconstructing the BF , we are mainly interested in the quantification of evidence in a discrete sense. According to [Kass and Raftery (1995)], the evidence provided by a BF_{10} can be categorized as follows:

Table 1. Strength of Evidence for M_1 relative to M_0 relative to the value of BF_{10}

$2\ln BF_{10}$	BF_{10}	Evidence against H_0
0-2	1-3	Not worth more than a bare mention
2-6	3-20	Positive
6-10	20-150	Strong
>10	>150	Very strong

Per simulation scenario and for each method, a contingency table was constructed in which the rows represented the categorized value of BF while the columns the value of \widetilde{BF} . Agreement in terms of correct classification in the discrete categories of strength of evidence was assessed via Cohen's κ statistic [Cohen (1960)]. The categories used to evaluate agreement are shown in **Table 2**.

Quantification of the continuous association between BF and \widetilde{BF} was carried out by calculating the root squared error of \widetilde{BF} through:

$$\text{RSE} = \sqrt{(\widetilde{BF} - BF)^2}$$

and its empirical average and quantiles.

Table 2. Level of agreement between BF and \widetilde{BF} for different ranges of Cohen's κ

Cohen's κ	Level of Agreement
<0	No agreement
0 - 0.2	Slight agreement
0.2 - 0.4	Fair agreement
0.4 - 0.6	Moderate agreement
0.6 - 0.8	Substantial agreement
0.8 - 1	Almost Perfect agreement

3. RESULTS

Figure 1 shows the RMSE of the several methods evaluated alongside the empirical central 90% of the sampling distribution, for different K and across true β . It is evident that p -value combination methods are not suitable for meta-analysis of BF s. Furthermore, $\tilde{\Lambda} = \sum_k \Lambda_k$ is consistently producing a high RSE and this result demonstrates why summation of LRs (and thus multiplication of BF s) is not a valid approach to BF synthesis.

We note that for all methods, RMSE increases with increasing K and β . Increasing K is straightforward to understand since more noise is introduced as all methods are either heuristic (p -value combinations) or approximate (4). With increasing β any p -value combination method is less powerful than the UMP test based on the whole dataset which translates to an overall combined statistic smaller than the original one [Bauer (1994)].

Figures 2-4 show the association plots between the estimated $2\ln\widetilde{BF}$ from the methods explored and the $2\ln BF$ calculated from the pooled data. They are depicted as $2\ln BF_{10}$, thus quantifying evidence for the alternative, except for the case when $\beta=0$, where they are shown as $2\ln BF_{01}$. On the bottom right of each figure the κ statistic is printed, with color according to the categorization in Table 2, with a gradient from red to green, where dark red signifies "No agreement" and dark green "Almost Perfect agreement". The axes are shortened at the edges, grouping values <-1 or >12 . Only results for $K=5$ are shown but results for other K ($\{2,5\}$) are very similar. It is clear from the figures that the only method from the ones explored here for synthesizing BF s is the one based on S_{wZ} .

Lancaster's and Tippet's methods perform very poorly. The rest of the naïve methods demonstrate some linearity between $2\ln\widetilde{BF}$ and $2\ln BF$ but produce considerably

Figure 1. Relative Bias of the different synthesis methods

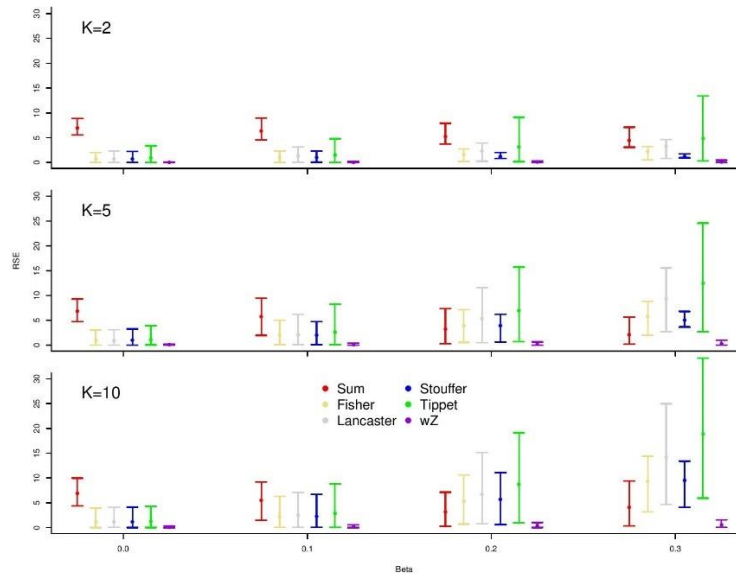


Figure 2. Association of BF and \widetilde{BF} calculated from $\tilde{\Lambda} = \sum_k \Lambda_k$, for $K=5$

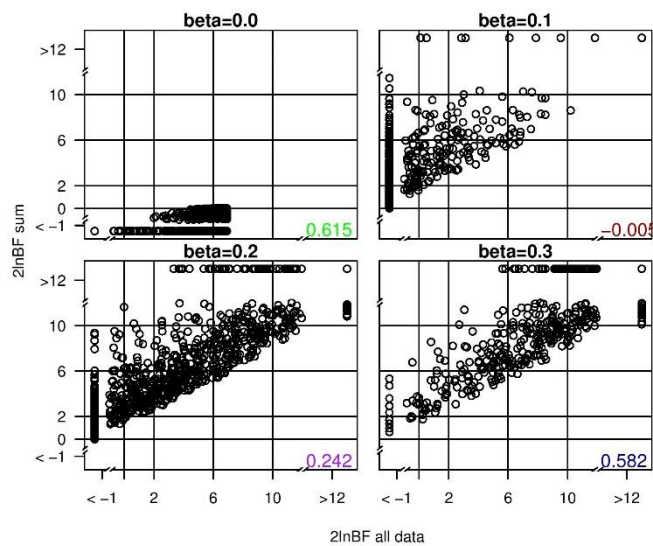
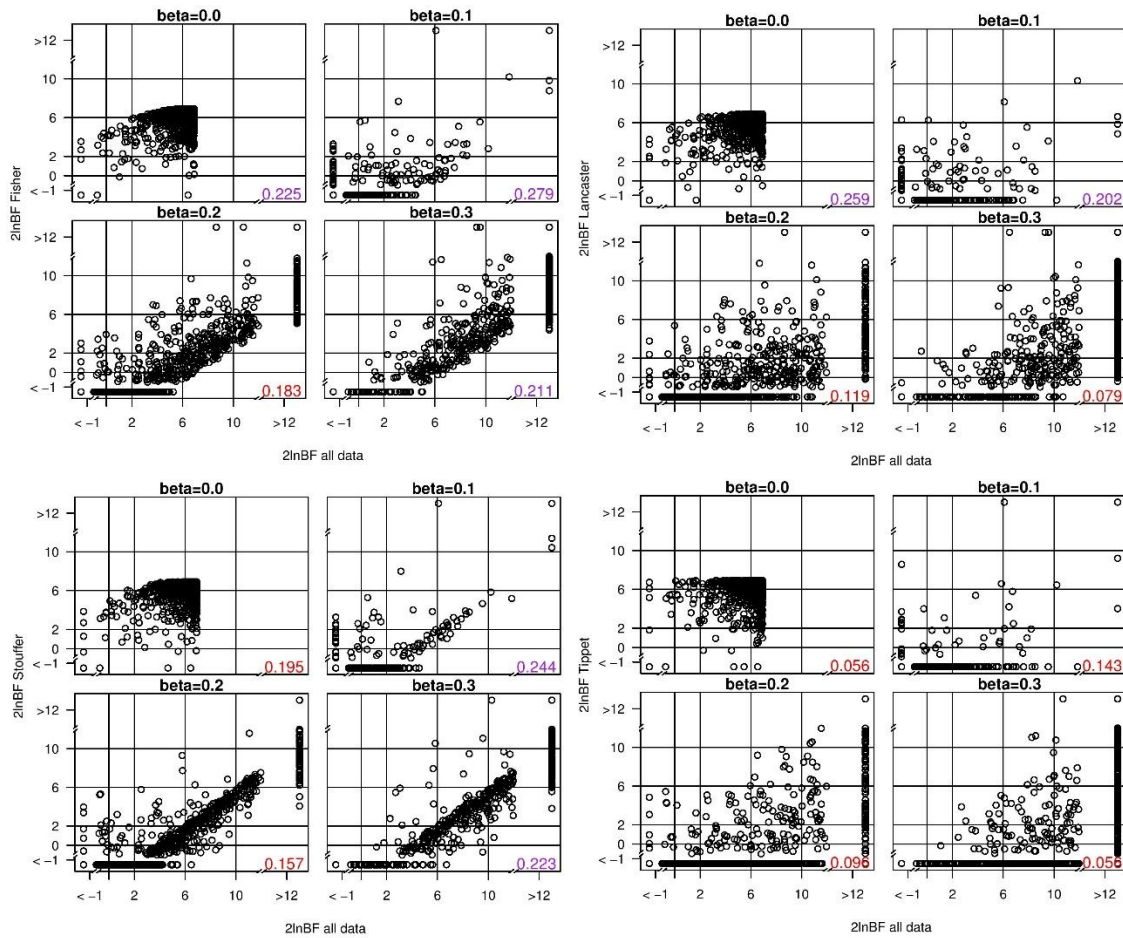
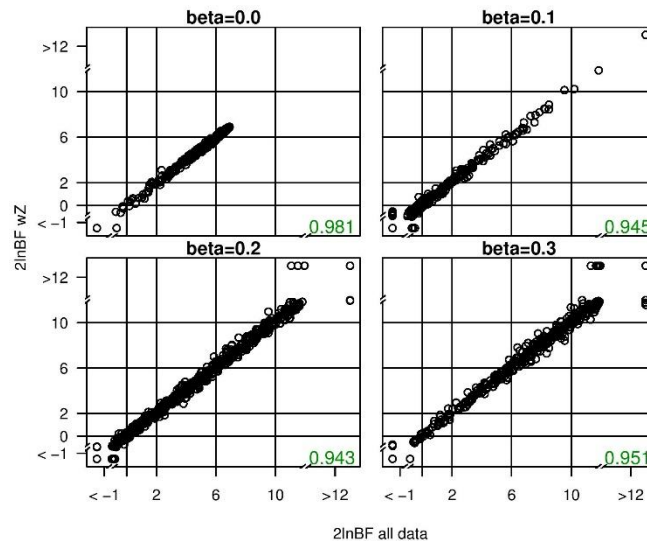


Figure 3. Association of BF and \widehat{BF} calculated from p -value combination methods for $K=5$



smaller $2\ln \widehat{BF}$ relative to $2\ln BF$. It is perhaps interesting to point out how none of the naïve methods manages to capture evidence for the null model adequately when H_0 is true, as a result of the shortcomings of synthesizing undirected p -values as discussed in section 2.5. We also notice that the largest the deviation from H_0 , the better the performance of most of the combiners as the effect becomes sharper in each sub-sample.

Figure 4. Association of BF and \widetilde{BF} calculated from S_{wZ} (weighted Zs method), for $K=5$



4. DISCUSSION

We explore methods for synthesizing evidence that are reported in the form of BF s. We do so by utilizing recent results that identify the sampling distribution of a BF in linear regression for specific prior settings and its relationship with the LR statistic. By combining them with frequentist theory for evidence synthesis we attempt to provide methods that can reasonably reproduce a BF based on the BF_k s that are calculated from the K subsamples. We show that p -value combination methods perform very poorly in such a task. This should not come as a surprise since their rationale is merely to maintain operational characteristics of significance testing. We suggest a method that is based on the weighted synthesis of normally distributed statistics, exploiting the direction of the effects alongside the values of the BF_k s. Alternative approaches could also be explored, like weighing by the inverse of the standard errors instead of the sample fractions or exploiting different approximations between the BF and the LR.

Synthesis of BF s has drawn very little attention in the statistical literature. A potential reason for this is the fallacious assumption reported in the literature that the total evidence from multiple BF s is their product. We show how this does not hold. It is also straightforward to show that the marginal likelihood involved in the calculation of BF s is not the product of the partial likelihoods as it should involve integrating each partition conditional on the posteriors arising from the rest of the data.

Our demonstration is limited to the case of a t -test. Such a task could be very useful in simple but very common settings, as is the synthesis of evidence from a series of

experiments, where testing is usually concerned with a difference between groups. The method is directly applicable to the case of simple linear regression with a continuous covariate. Specifically for the t -test a method has been provided elsewhere for a different setting of (potentially informative) prior distributions [Rouder and Morey (2011)].

For more general problems where BF s are commonly employed, like variable/model selection, the extension is not straightforward. A similar strategy could be sought where regression slopes arising from independent datasets would be synthesized, though such a task does not come without complexity [Becker and Wu (2007)]. Such complexities include not reported covariance matrices as well as systematically missing covariates. Admittedly, it is difficult to imagine a situation where such a task would be relevant for meta-analysis, though it is an interesting problem for Big Data.

The methods suggested would also be of direct application for the transformation of frequentist evidence arising in any form (t -statistics, LR or p -values) into Bayesian evidence in the form of BF s.

ΠΕΡΙΛΗΨΗ

Το πρόβλημα της σύνθεσης στατιστικών συναρτήσεων μπορεί να προκύψει σε (τουλάχιστον) δύο περιοχές της ανάλυσης κατά Μπέυζ. Το ένα είναι η περίπτωση (πολύ) μεγάλων μεγεθών δειγμάτων όπου τα υπολογιστικά εργαλεία (όπως το MCMC) δεν κλιμακώνονται και έτσι ένας τρόπος γύρω από αυτό θα ήταν να χωρίσουν τα δεδομένα σε υπο-δείγματα και στη συνέχεια να συνδυάσουν τις στατιστικές συναρτήσεις. Το άλλο είναι η μέτα-ανάλυση των δημοσιευμένων αποτελεσμάτων, όπου οι ερευνητές δεν έχουν πρόσβαση σε δεδομένα μεμονωμένων ατόμων. Η παρούσα εργασία διερευνά μεθόδους για τη σύνθεση αποτελεσμάτων που είναι ποσοτικοποιημένα με βάση τους Bayes Factors. Συγκεντρώνοντας πρόσφατα αναλυτικά αποτελέσματα σχετικά με τη σχέση μεταξύ του Bayes Factor και του λόγου πιθανοφάνειας, αξιολογούμε τις τεχνικές σύνθεσης στοιχείων για προβλήματα επιλογής μεταβλητών. Αυτό συμβαίνει για την περίπτωση της γραμμικής παλινδρόμησης κατά τη χρήση των g -priors. Επιδεικνύουμε την απόδοση των μεθόδων που προτάθηκαν στην απλή περίπτωση ενός t -test. Συζητάμε επίσης πιο πολύπλοκα προβλήματα σχετικά με τη επιλογή μεταβλητών στη γραμμική παλινδρόμηση.

Acknowledgments: This research was supported by the AUEB DRASI-II funding scheme.

REFERENCES

Bauer P, Köhne K (1994) Evaluation of experiments with adaptive interim analyses. *Biometrics* 50:1029 -1041

- Becker BJ, Wu M. (2007) The synthesis of regression slopes in meta-analysis. *Statistical Science* 2007; 3: 414
- Bem, D. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100, 407–425.
- Berger J.O. & Sellke T (1987) Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence, *Journal of the American Statistical Association*, 82:397, 112-122
- Buse, A. (1982). The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. *American Statistician*, 36, 153-157.
- Cohen, J (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 20 (1): 37–46
- Engle, R. F. (1984), Wald, Likelihood Ratio and Lagrange Multiplier Tests, in *Handbook of Econometrics* (Vol. 2), eds. Z. Griliches and M. D. Intriligator, Amsterdam: North-Holland, pp. 776-826
- Fisher, R. A. (1934). *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd, 4th ed.
- Goodman,S.N. (2016) Aligning statistical and scientific reasoning. *Science*, 352, 1180–1181.
- Heard, N.A. and Rubin-Delanchy, P. (2018). Choosing between methods of combining pp-values. *Biometrika* **105** 239–246.
- Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.
- Held, L., and Ott, M. (2018), “On p-Values and Bayes Factors,” *Annual Review of Statistics and Its Application*, 5, 393–419
- Jeffreys, H. (1961). *Theory of Probability* (3rd edition). Oxford University Press.
- Kass, R. E. (1992). Bayes factors in practice. *Journal of the Royal Statistical Society Series D (The Statistician)*, 2, 551–560.
- Kass, R. and A. Raftery (1995). Bayes factors. *Journal of the American Statistical Association* 90,773-79
- Lancaster, H.O., (1961). The combination of probabilities: an application of orthonormal functions. *Australian Journal of Statistics* 3, 20–33.
- Liang, F., Paulo, R., Molina, G., Clyde, M. A. and Berger, J. O. (2008), Mixtures of g priors for Bayesian Variable Selection, *Journal of the American Statistical Association*, 103, 410–423.
- O’Hagan, A., and Forster, J. J. (2004), Kendall’s Advanced Theory of Statistics, Volume 2B: Bayesian Inference, London:Arnold.
- Owen AB. Karl Pearson's meta-analysis revisited (2009), *Annals of Statistics*, 3867-3892
- Pearson, K. (1933). On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika* **25**, 379–410.

- Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, *18*, 682–689
- Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *American Statistician*, *55*, 62–71.
- Shively, T., Walker, S., 2018. On bayes factors for the linear model. *Biometrika* 105 (3), 739–744
- Stouffer, S. A., Suchman, E. A., Devinney, L. C., Star, S. A. Williams, R. M. (1949). *The American Soldier. Adjustment During Army Life*. Princeton, New Jersey: Princeton University Press.
- Tippet, L. H. C. (1931). *The Methods of Statistics*. London:Williams and Norgate.
- Tsamardinos I, Borboudakis G, Katsogridakis P, Pratikakis P, and Christophides V (2019) A greedy feature selection algorithm for Big Data of high dimensionality. *Machine Learning*, 108(2), 149-202.
- Wilks, S. S. (1938), "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses," *The Annals of Mathematical Statistics*, 9, 60–62.
- Zellner, A. (1986), "On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions," *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti*, 6, 233–243.
- Zhao, Z., Zhang, R., Cox, J., Duling, D., & Sarle, W. (2013). Massively parallel feature selection: An approach based on variance preservation. *Machine Learning*, 92(1), 195–220.
- Zhou, Q. and Guan, Y. (2018). On the null distribution of bayes factors in linear regression. *Journal of the American Statistical Association*, *113*, 1362– 1371.



HIERARCHICAL FORECASTING IN GREEK CONVENIENCE RETAIL INDUSTRY

Christos Tripodis, Dimitris Karlis

Athens University of Economics and Business

christostripodis@gmail.com

karlis@aueb.gr

ABSTRACT

In this study, the hierarchical structure of Greek convenience retail sales is explored, and appropriate forecasting methods are applied to estimate short-term forecasts. The data are organized under two distinct levels of hierarchy, such as, product categories and physical store locations. The underlined structure is considered as a grouped time series given that the data can be independently grouped either by categories or by stores, irrelevantly of the order of the aggregation. An optimal forecasting method has been introduced, which independently calculates the forecasts across all the levels of series and then applies a regression model to come up with combined forecasts. By such an approach the forecasts are consistent in each level and they sum up to the previous/next level which is not the case when forecasting at each level is done separately. In addition, bottom-up method is also applied to the same data and a comparison between the methods is provided.

Keywords: retail, hierarchical forecasting, bottom-up method, optimal method

1. INTRODUCTION

In retail and fast-moving-consumer-goods (FMCG) industries, the business objective of forecasting volume and value sales often arises. For example, retailers and manufacturers can use sales predictions to take business decisions in relation to stock management so that efficiently prevent stock-out and overstocking events to maximize profit. Furthermore, forecasting applications are used to estimate future sales of own and competitive portfolio products and to quantify relationships among price diversities and promotion activities. This kind of knowledge could help brand

and marketing managers drive business decisions related to price and promotional strategies.

However, the challenge of accurately forecasting sales in these industries is attributed to the nature of the data which is, commonly, characterized by time series hierarchical structures, organized at several different levels such as products or geographies. More specifically, products could be aggregated and represented as brands, brands can be further aggregated to categories and so on. Similar aggregation structures exist in geographical dimensionalities where stores can be aggregated to broader geographical areas such as regions, states or countries.

The most widespread forecasting methods applied to hierarchical time series is the bottom-up or the top-down, or a combination of both as the middle-out approach. The bottom-up method entails forecasting each of the disaggregated series at the lowest level of the hierarchy, and then, derive the forecasts at the higher levels by simply following the aggregation structure of the hierarchy [Kahn (1998)]. Conversely, the top-down method entails forecasting the time series at the top level of the hierarchy and then with the help of calculated weight factors based on historical proportions, estimate the series at the lower levels of the hierarchy. [Gross et. al. (1990)] elaborate with different ways of calculating the proportions associated to the bottom level series. Middle-out comes as a combination of the two, where the forecasts are first obtained at an intermediate level of the series, and then forecasts are derived for the higher and lower levels of hierarchy by applying the bottom-up and top-down methods respectively. Alternatively, the forecasts can be simply obtained at the individual levels of the hierarchy by ignoring the hierarchy structure, but this would result to estimations which would not be consistent and coherent, as the lower level forecasts would not appropriately add up to the higher-level forecasts [Flidner (2001)]. Throughout this study we will refer to this kind of forecast as “base” forecasts.

Rather than modeling either the top level or the top-down levels and then proceeding with post modeling adjustments, [Hyndman et. al. (2011)] proposed directly modeling the entire hierarchy of time series simultaneously. In [Hyndman et. al. (2011)] approach, the forecast of all series at all levels are first obtained and then using a regression model these forecasts are optimally combined. The resulting reconciled forecasts outweigh the forecasts of the other methods. All the information available within the hierarchy is used, allowing for correlations and interactions among the levels of the hierarchy. By such an approach, the independent base forecasts are regressed against the sum of the expected values of the future series plus an error term. Conceding that base forecasts are unbiased, and the covariance matrix is known, then the generalized least squares (GLS) comes up with unbiased and minimum variance estimates. In general, estimation of the covariance matrix is impossible to estimate due to identifiability conditions. To overcome this issue, [Hyndman et. al. (2011)] proposed that under some basic assumptions the estimation could be approached as ordinary least squares estimation (OLS) which simplifies the calculations. One of the limitations of the OLS model is that when there are

complicated hierarchy structures which involve thousand or even millions of disaggregated series the models is impossible to estimate as the design matrix has one column for each of the series at the most disaggregated level. [Hyndman et. al. (2016)] proposed a solution to this problem by exploiting the unique structure of the linear model to efficiently estimate the coefficients.

Literature has been mainly focusing on comparing the bottom-up with top-down method. [Kanh, (1998)] suggested that top-down approach performs better at the aggregated levels whilst the bottom-up method performs better at the lowest levels. One major disadvantage of the top-down method is that it assumes a common seasonal factor across all the disaggregated series which often is an unrealistic assumption. For example, as we will see in this study, chocolates have the opposite seasonal effect than ice-creams. [Lapide, (2006)] proposed that that top-down forecasting suits best when a top-level aggregated group constitutes time series that have similar variation, whilst the use of bottom-up is better for situations where the individual components have different patterns of variation. [Hyndman et. al. (2011)] first compared and showed that the optimal method outperforms top-down, bottom-up and base forecasts in Australian domestic tourism data. [Wickramasuriya et. al. (2019)] and [Van Erven and Cugliari (2014)] suggested that forecasting reconciliation should preferred against approaches that don't use the whole information provided by the hierarchical structure such as bottom-up and top-down. [Athanasopoulos et.al. (2009)] showed that top-down method based on forecasted proportions and the optimal combination approach outperform the bottom-up.

In this work we will focus on Greek convenient retail area and forecast sales of kiosks and mini markets. We will analyze value sales occurred at the point of sales (kiosk or mini market) for 9 categories of products across 52 stores located in various areas in Greece. In section 2, we present a framework for general hierarchical forecasting according to the literature and we will show how some of the methods fit and apply to this framework. Sections 3 and 4 apply the methodology to the dataset from Greek retailer and compare how the optimal method suggested by [Hyndman et. al. (2011)] compares with bottom-up and base forecasts to the domains of Greek convenience retail and FMCG. Finally, some remarks can be found in Section 5.

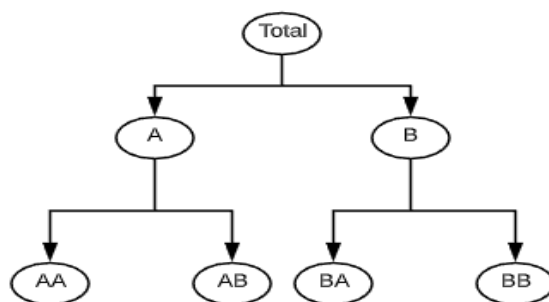
2. NOTATION

2.1 Hierarchical time series

In order to demonstrate the structure of hierarchical time series data we will use a simple hierarchy with $k = 2$ levels as shown in Figure 1 and then we will introduce the appropriate notation to describe its properties. Hierarchical time series refers to time series that are organized based on a unique hierarchical structure. At the higher level of the aggregation denoted as level 0, the total sales are represented. Total sales can be disaggregated at level 1 of the hierarchy into 2 categories of products denoted as category A and B. Finally, the final level denoted as level 2 consists of 4 time

series which represent the corresponding disaggregated brands donated as AA, AB, BA and BB. Let us denote with m_i the number of series at level i such that $m_0 = 1$, $m_1 = 2$ and $m_2 = 4$ and m denotes that total number of series $m = m_0 + m_1 + m_2$.

Figure 1. Hierarchical time series



For any given time t , the most disaggregated series will add up to series at the levels above according to hierarchy structure. Expressing this to matrix notation, the matrix S of order $m \times m_k$ is introduced which reflects the hierarchy structure. We denote as y_t the total at time t , while y_{it} , $i=\{A,B, AA,AB,BA,BB\}$ are the value at time t for each level. For the hierarchy of Figure 1 we can write the matrix notation as per the below

$$\mathbf{y}_t = \begin{bmatrix} y_t \\ y_{At} \\ y_{Bt} \\ y_{AA t} \\ y_{AB t} \\ y_{BA t} \\ y_{BB t} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} y_{AA t} \\ y_{AB t} \\ y_{BA t} \\ y_{BB t} \end{bmatrix}$$

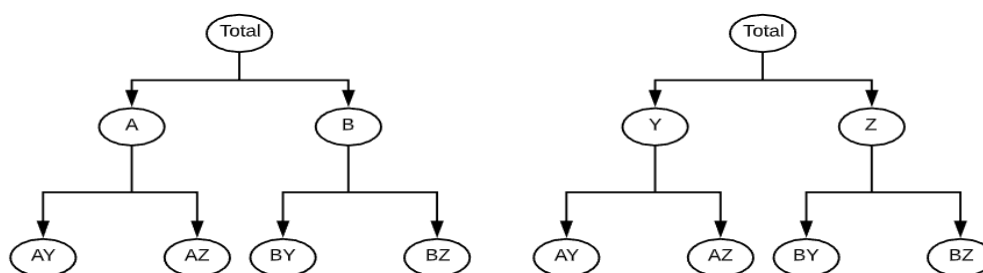
or in more compact form $\mathbf{y}_t = \mathbf{S}\mathbf{B}_t$ where \mathbf{S} stands for the summation matrix and \mathbf{B}_t for the vector of bottom level series. With that said, all the hierarchical time series could be written in this form as a function of the summation matrix and the bottom level series.

2.2 Grouped time series

When the disaggregation is not uniquely hierarchical, but also involves more than one dimensionality such as for example geography and product, then we deal with grouped time series. A common feature of a grouped time series is that the order in which the disaggregation occurs is not unique. In the following grouped time series example, the disaggregation can take place in terms of either product or geography as shown in Figure 2. We will be focusing on this this type of hierarchy, as in this study we have chosen to analyze a grouped time series. Mapping this example to the

convenience retail data, the total sales could be derived from either the hierarchy of the product (product, brand, category) or from the hierarchy of geography (city, region, country).

Figure 2. Grouped time series



The features of this hierarchy are described as that $m_0 = 1$, $m_1 = 2$ and $m_2 = 2$ and $m_3 = 4$ where the total series of the hierarchy equal $m = 9$. The mathematical formulation of this structure can be expressed with matrix notation as.

$$\mathbf{y}_t = \begin{bmatrix} y_t \\ y_{At} \\ y_{Bt} \\ y_{Yt} \\ y_{Zt} \\ y_{AYt} \\ y_{AZt} \\ y_{BYt} \\ y_{BZt} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} y_{AYt} \\ y_{AZt} \\ y_{BYt} \\ y_{BZt} \end{bmatrix}$$

2.3 Hierarchical forecasting

We just presented a notation that describes a hierarchical time series data, however, in hierarchical forecasting the focus is on the forecasted and not on the observed values of the time series. The forecasting method to predict each individual time series is irrelevant of the hierarchical forecasting which is just a combination or reconciliation according to the method applied of the already calculated forecasts. Having clarified that, we denote as h the forecasting horizon of the predictions. The forecasting values have been estimated based on of $t = 1, 2, \dots, n$ such that the forecasts correspond to time $t = n + h$. We denote with $\hat{Y}_n(h)$ the vector of all base forecasts across all levels of hierarchy stacked as \mathbf{y}_t . Using this notation all the hierarchical forecasting approaches can be expressed by the general form

$$\hat{Y}_n(h) = \mathbf{S} \mathbf{G} \hat{Y}_n(h), \tag{1}$$

where \mathbf{S} is the summation matrix order of $m \times m_k$ and \mathbf{G} is a matrix of order $m_k \times m$. The role of \mathbf{G} is to extract and combine and project the base forecasts $\hat{Y}_n(h)$ into bottom level disaggregated forecasts which are then summed by \mathbf{S} to give the final revised hierarchical forecasts, $\tilde{Y}_n(h)$ which are now consistent and coherent by construction and it varies depending on the hierarchical approach. In the following section we will fit and expand (1) for bottom-up, top-down and optimal method for the example grouped time series that we saw earlier on.

2.4 The bottom-up method

One of the most popular methods to address hierarchical forecasting problems is the bottom-up method. Bottom-up method involves estimating just the bottom level series with some forecasting method and then by summing up the bottom level forecasts calculate the forecasts in the upper level series. In order to represent this approach according to the general hierarchical form we denote.

$$\mathbf{G} = [\mathbf{0}_{m_k \times (m-m_k)} \mid \mathbf{I}_{m_k}] \quad (2)$$

where $\mathbf{0}_{m_k \times (m-m_k)}$ is a null matrix (all elements are zero) and \mathbf{I}_{m_k} is the identity matrix order of m_k . As demonstrated below, \mathbf{G} zeroes-out the effect of the forecast of the top series and picks up the bottom level forecasts. According to [Athanasopoulos et. al. (2009)] the greatest advantage of this approach is that it captures better the dynamics of the individual series, since the modeling occurs at the lowest level not losing any information due to any aggregations prior to modeling. On the contrary, when bottom level data is quite noisy is difficult to systematic capture patterns that are becoming obvious after any aggregations. The matrix notation below illustrates how the grouped time series depicted in Figure 2 fits to equation (1) according to bottom-up method.

$$\begin{bmatrix} \tilde{Y}_t \\ \tilde{Y}_{At} \\ \tilde{Y}_{Bt} \\ \tilde{Y}_{Yt} \\ \tilde{Y}_{Zt} \\ \tilde{Y}_{AYt} \\ \tilde{Y}_{AZt} \\ \tilde{Y}_{BYt} \\ \tilde{Y}_{BZt} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} \hat{Y}_t \\ \hat{Y}_{At} \\ \hat{Y}_{Bt} \\ \hat{Y}_{Yt} \\ \hat{Y}_{Zt} \\ \hat{Y}_{AYt} \\ \hat{Y}_{AZt} \\ \hat{Y}_{BYt} \\ \hat{Y}_{BZt} \end{bmatrix}$$

2.5 The top-down method

Top-down approach involves first estimating the base forecasts at the top level and then disaggregating those downwards based on pre-defined proportions. We denote with p_1, \dots, p_n a set of proportions which determine how the forecasts at the top level are distributed towards the lower level series

$$\mathbf{G} = [\mathbf{p} | \mathbf{0}_{m_k \times (m-1)}] \quad (3)$$

The proportions are defined in such way that when aggregating up the calculated bottom level forecasts by following the summation matrix we get the forecasts at the upper series. [Gross and Sohl (1990)] focused on top-down method and they suggested two promising methods based on historical proportions and proportions of historical averages. The matrix notation below expands how the grouped time series depicted in Figure 2 fits in equation (1) according to top-down method.

$$\begin{bmatrix} \widehat{Y}_t \\ \widehat{Y}_{At} \\ \widehat{Y}_{Bt} \\ \widehat{Y}_{Yt} \\ \widehat{Y}_{Zt} \\ \widehat{Y}_{AYt} \\ \widehat{Y}_{AZt} \\ \widehat{Y}_{BYt} \\ \widehat{Y}_{BZt} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} p_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} \widehat{Y}_t \\ \widehat{Y}_{At} \\ \widehat{Y}_{Bt} \\ \widehat{Y}_{Yt} \\ \widehat{Y}_{Zt} \\ \widehat{Y}_{AYt} \\ \widehat{Y}_{AZt} \\ \widehat{Y}_{BYt} \\ \widehat{Y}_{BZt} \end{bmatrix}$$

2.6 The optimal method

[Hyndman et. al. (2011)] first introduced the optimal method based on which the forecasts are initially estimated across all the time series within the hierarchy. These forecasts concern the base forecasts which don't add up according to the hierarchy structure. Afterwards, the base forecasts are optimally reconciled to a set of revised forecasts which meet the aggregation constraints imposed by the hierarchy. According to the optimal method the representation of the h-step-ahead base forecasts for the whole hierarchy can be written in the form of a linear model.

$$\widehat{Y}_n(h) = \mathbf{S}\beta_h + \varepsilon_h, \quad (4)$$

where $\widehat{Y}_n(h)$ is a vector of the h-step-ahead base forecasts for the whole hierarchy, β_h is the unknown mean of the future values of the bottom level k, and ε_h has zero mean and covariance matrix Σ_h . Treating the above as a regression equation and assuming that Σ_h is known, then β_h could be estimated with the generalized least squares estimation. However, the task of this estimating Σ_h could be very challenging especially when the hierarchy structure is very complicated. [Hyndman et. al. (2011)] showed that under the reasonable assumption that the errors satisfy the aggregation constrain of the original data, then, β_h could be estimated with the Ordinary Least Squares estimator (OLS) such that

$$\beta_h = \mathbf{S} (\mathbf{S}' \mathbf{S})^{-1} \mathbf{S}' \widehat{Y}_n(h) \quad (5)$$

$$\mathbf{G} = \mathbf{S} (\mathbf{S}' \mathbf{S})^{-1} \quad (6)$$

which leads to the reconciled forecasts given by $\hat{Y}_n(h) = S\beta_h$ and fits the general equation of hierarchical forecasting (1). A very convenient property of the OLS estimator is that the weights are solely dependent on the hierarchy structure and not on the underlined data. That said, the weights can be computed once for the hierarchy and then used again and again once new data fits in. The summation matrix and the weights of the grouped time series presented in Figure 2 can be written as

$$S = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$S(S'S)^{-1}S' = \begin{bmatrix} 0.44 & 0.22 & 0.22 & 0.22 & 0.22 & 0.11 & 0.11 & 0.11 & 0.11 \\ 0.22 & 0.44 & -0.22 & 0.11 & 0.11 & 0.22 & 0.22 & -0.11 & -0.11 \\ 0.22 & -0.22 & 0.44 & 0.11 & 0.11 & -0.11 & -0.11 & 0.22 & 0.22 \\ 0.22 & 0.11 & 0.11 & 0.44 & -0.22 & 0.22 & -0.11 & 0.22 & -0.11 \\ 0.22 & 0.11 & 0.11 & -0.22 & 0.44 & -0.11 & 0.22 & -0.11 & 0.22 \\ 0.11 & 0.22 & -0.11 & 0.22 & -0.11 & 0.44 & -0.22 & -0.22 & 0.11 \\ 0.11 & 0.22 & -0.11 & -0.11 & 0.22 & -0.22 & 0.44 & 0.11 & -0.22 \\ 0.11 & -0.11 & 0.22 & 0.22 & -0.11 & -0.22 & 0.11 & 0.44 & -0.22 \\ 0.11 & -0.11 & 0.22 & -0.11 & 0.22 & 0.11 & -0.22 & -0.22 & 0.44 \end{bmatrix}$$

3. GREEK CONVENIENCE RETAIL FORECASTS

The data has been gathered with the help of an electronic point of sales software (POS) installed at the store location. By using the POS, store owners can track any transactional information such as time of purchase, discounts, price of products which then can be transformed to consolidated reports related to the performance of the store. An average kiosk submits 300 transactions per day which significantly vary in the summer months versus the winter months where it is observed that the sales are increased during summer. Data has been selected across 52 stores for a 2 years period. Having two years of historical data will help the construction of more robust models as the concept of seasonality is also introduced. The first week of the analysis is the week ending 31/7/2016 and the last week of the analysis is the week ending 4/11/2018. In total, 119 weeks are analyzed. In order to evaluate the results of the model we will be using a train and a validation dataset. For this reason, the first 109 weeks are considered as the observed weeks where the models are trained, whilst the

forecasting results are evaluated for the period of weeks 110-119. The analyzed time series are organized in hierarchical structure as per the properties of grouped time series. There are 2 dimensionalities based on which the aggregation could occur, product categories and stores. Level 1 of product dimension consists of the following components: tobacco, beverages, alcohol, ice-creams, snacks, chocolates, water, and dairy. The level 1 of the store dimension consists of 52 time series standing for total sales per store. In total, there are 468 bottom level time series, 9 product categories series aggregated at total stores, 52 aggregated series representing the total sales of each store and one more series representing the grand total. The final number of the estimated time series is 530. This structure reflects a grouped time series because the hierarchical structure is not unique. In other words, the total sales of those 52 stores could be first disaggregated by product category and then by store, or it could be disaggregated first by store and then by product category.

In order to estimate the base forecasts for all series we used an automate process which applies a step-wise algorithm, trying a variation of ARIMA models with a different set of parameters. Then the model with the best AIC is selected as the final model. For this purpose, we used the *auto-arima* function in R software. More details about this algorithm could be found in the work of [Hyndman & Khandakar (2008)].

4. FORECASTS EVALUATION

In order to evaluate the performance of each of the hierarchical forecasting method we apply, we perform an out-of-sample forecast evaluation across product category and store hierarchies. The modeling occurs at the trained dataset estimating a 10-step-ahead forecast and then we use the estimated results against the real data to assess model performance. [Adhikari and Argawal (2013)] suggested several measures for evaluating the forecasting accuracy which could be classified into three categories:

- Scale dependent measures such as mean error (ME), mean absolute error (MAE) and root square mean square error (RMSE) which are easily interpreted and can be used for comparing different forecasting methods within the same time series.
- Percentage Error Metrics such as mean percentage error (MPE) and mean absolute percentage error (MAPE), which are scale free and can be used for comparisons across different time series.
- The scale-free error metric such as mean absolute scaled error (MASE) can be used to compare forecast methods on a single series and to compare forecast accuracy between series. It was proposed by [Hyndman (2006)], who described it as a generally applicable measurement of forecast accuracy to overcome the problems seen in the other measurements.

Table 1-3 describes all the measures for three different approaches, namely base forecast with ARIMA models (Table 1), bottom-up method (Table 2) and the optimal approach (Table 3). We report values for the levels 0 and 1. As expected, one can see

major difference between categories, since we refer to different scales and only the scale free measure MASE can be compared across categories.

In Figure 3 the forecast estimations at level 1 of product hierarchy derived from base method are demonstrated. It's obvious that all the categories are characterized by seasonal patterns except for tobacco which doesn't seem to be seasonal. This is attributed to the nature of the data where out of the 468 bottom level series, 393 have been estimated with a seasonal ARIMA which approximates 82% of total series. Seasonality is a predominant factor in the Greek convenience retail sales.

Most of the forecasts are robust indicated by the percentage error metrics where total series comes with MAPE of 1.78% ,1.76% and 1.77% in optimal, bottom-up and base methods respectively. In simple terms this means that the average absolute percentage error is very small and that the forecasting values are very close to the observed values.

For level 1 series in terms of product category, both the optimal combination approach and the bottom-up seems to be performing well whilst the base forecasts underperform. In many cases there are strong patterns in the form of trends and seasonality which eventually yields good forecasts at the aggregated levels. Furthermore, as [Athanasopoulos et al. (2009)] claimed, bottom-up method is also advantaged by the short-term forecasts that we have produced in this study

For specific product categories the results are very different between optimal and bottom-up method depending on the nature of the product category. For example, we know that the least seasonal products from the ones investigated are the tobacco products. This is the only category that all the measurements show that optimal method performs better than the bottom-up method. Another non-seasonal product category is the dairy where most of the performance measurements show that optimal is better method. On the contrary, when looking to very seasonal categories such as beverages or ice-creams we see the opposite result where bottom-up performs better than the optimal method. Interestingly, chocolate category appears with opposite seasonal pattern where the sales are increased during the winter as opposed to summer months. Even if in this study the top-down method is not examined, that would have been a case where this method would have failed to capture the opposite pattern.

Table 1. Base forecast accuracy measures for level 0 and 1 aggregation in terms of products category

Measure	Total	Alcohol	Beverages	Chocolates	Dairy
ME	-3,278	-2,416	-968	-401	-130
RMSE	14,914	2,963	2,663	842	380
MAE	11,478	2,721	2,478	645	243
MAPE	1.77	9.22	7.67	4.95	4.08
MPE	-0.52	-8.04	-3.41	-2.38	-2.36
MASE	0.22	0.77	0.76	0.78	0.70

Measure	Gum	Ice-creams	Snacks	Tobacco	Water
ME	-270	-318	-887	5,869	-621
RMSE	457	1,349	1,235	13,204	1,200
MAE	313	1,127	1,000	12,007	1,068
MAPE	5.05	11.53	5.05	2.29	11.78
MPE	-4.42	-1.43	-4.50	1.07	-7.44
MASE	0.53	0.43	0.66	0.29	0.64

Table 2. Bottom-up method accuracy measures for level 0 and 1 aggregation in terms of products category

Measure	Total	Alcohol	Beverages	Chocolates	Dairy
ME	5,254	-1,005	-76	-218	-93
RMSE	16,123	1,830	2,460	764	326
MAE	11,515	1,530	2,072	616	231
MAPE	1.76	5.45	6.25	4.81	3.87
MPE	0.75	-3.42	-0.66	-0.68	-1.79
MASE	0.22	0.43	0.64	0.74	0.66

Measure	Gum	Ice-creams	Snacks	Tobacco	Water
ME	-259	-521	-201	7,976	-350
RMSE	454	1,299	779	13,074	1,077
MAE	311	1,185	534	10,033	939
MAPE	5.03	12.83	2.70	1.91	10.10
MPE	-4.27	-5.63	-1.06	1.50	-4.54
MASE	0.52	0.45	0.35	0.24	0.56

Table 3. Optimal method accuracy measures for level 0 and 1 aggregation in terms of products category

Measure	Total	Alcohol	Beverages	Chocolates	Dairy
ME	-318	-2,167	-791	-401	-146
RMSE	14,524	2,789	2,600	787	331
MAE	11,504	2,534	2,299	592	194
MAPE	1.78	8.54	7.07	4.63	3.30

MPE	-0.10	-7.12	-2.74	-2.80	-2.57
MASE	0.22	0.72	0.70	0.71	0.56

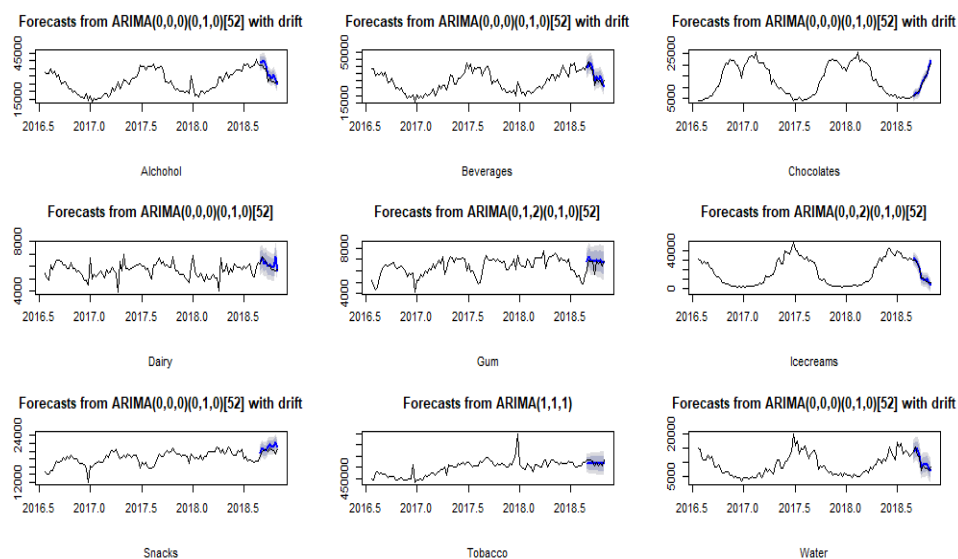
Measure	Gum	Ice-creams	Snacks	Tobacco	Water
ME	-290	-510	-758	5355	-609
RMSE	482	1382	1084	11044	1205
MAE	325	1189	853	9365	1009
MAPE	5.25	11.81	4.32	1.78	11.06
MPE	-4.73	-2.72	-3.86	0.99	-6.91
MASE	0.55	0.46	0.56	0.23	0.60

Level 1 of stores hierarchy consists of 52 time series which would be impossible to go through this paper. One way to summarize the results is to calculate average metrics of the performance measures across all 52 series. For that reason, we concentrate on the percentage errors as they are scale free and makes sense to average them. We could have also used weighted averages for the non-percentage errors so that we got a good representation of the overall performance. Table 4 is indicative where the optimal method outweighs both bottom-up and base forecasts method.

Table 4. Average percentage measures across all 52 store aggregated time series

Measure	Bottom-up	Optimal	Base
MAPE	7.10	6.88	7.09
MPE	0.65	-0.23	-0.06
MASE	0.63	0.61	0.63

Figure 3. Base forecasts on the aggregate series of products. All the series have been seasonal differenced apart from tobacco



5. CONCLUSIONS

In this paper we have applied hierarchical forecasting to the Greek convenient retail industry. The hierarchy structure involved grouped time series were products aggregate in 9 product categories and 52 stores. This means that the hierarchy structure is not unique and that there are two ways to calculate the grand total series. We considered three different forecasting methodologies, base forecasts, bottom-up and optimal hierarchical and made comparisons among them. In total, 530 time series have been estimated and the results were evaluated at the upper series. In general, all methods returned good results which is something attributed to the nature of the data. Seasonality and trends are a common characteristic of the sales in the convenience market which was something detected by the models and helped the forecasting values to be close to the actuals. Our evaluation of the forecast performance revealed that bottom-up and optimal methods perform better than the base forecasts. Particularly, the optimal method seems to outweigh bottom-up when the seasonality is not present such as in tobacco and dairy product categories. When looking at the level 1 of the stores' dimensionality, optimal method seems to consistently outperform both bottom-up and base forecasts methods.

ΠΕΡΙΛΗΨΗ

Σε αυτή την εργασία αναλύεται η ιεραρχική δομή χρονοσειρών πωλήσεων στην Ελληνική αγορά στο κανάλι της μικρής λιανικής. Συγκεκριμένα, εξετάζονται πωλήσεις που προέρχονται κυρίως από μικρά καταστήματα, όπως περίπτερα και

καταστήματα που εμπορεύονται ψιλικά. Εφαρμόζονται μεθοδολογίες κατάλληλές για πρόβλεψη ιεραρχικών δεδομένων σύμφωνα με την βιβλιογραφία. Τα δεδομένα οργανώνονται σύμφωνα με δύο ανεξάρτητες ιεραρχίες, δηλαδή την κατηγορία προϊόντος και το κατάστημα, και μπορούν να αθροιστούν ξεχωριστά ως προς αυτές. Εφαρμόζεται η «βέλτιστη» μεθοδολογία σύμφωνα με την οποία χρησιμοποιούνται όλες οι προβλέψεις όλων των χρονοσειρών μέσα στην ιεραρχία και με την βοήθεια της γραμμικής παλινδρόμησης αυτές συνδυάζονται βέλτιστα ώστε να εκτιμηθούν διορθωμένες, συνεπείς και συνεκτικές προβλέψεις. Επίσης εφαρμόζονται και άλλες μεθοδολογίες οι οποίες χρησιμοποιούνται για την πρόβλεψη ιεραρχικών χρονοσειρών και γίνεται σύγκριση μεταξύ αυτών.

REFERENCES

- Adhikari, R. and Agrawa R.K. (2013). *An Introductory Study on Time Series Modeling and Forecasting*. New Delhi: LAP Lambert Academic Publishing.
- Athanasopoulos, G. and Ahmed, A.R. and Hyndman R.J. (2009). Hierarchical forecasts for Australian domestic tourism. *International Journal of Forecasting*, **25**, 146-166.
- Fliedner, G. (2001). Hierarchical forecasting: issues and use guidelines. *Industrial Management & Data Systems*, **101**, 5-12.
- Gross C.W. and Sohl J.E. (1990). Disaggregation methods to expedite product line forecasting. *Journal of Forecasting*, **9**, 233-254.
- Hyndman R.J. and Khandakar, Y. (2008). Automatic Time Series Forecasting: the forecast Package for R. *Journal of Statistical Software* **27**.
- Hyndman, R.J. (2006). Another Look at Forecast Accuracy Metrics for Intermittent Demand. *Foresight: The International Journal of Applied Forecasting*, **4**, 43-46.
- Hyndman, R.J. and Ahmed, R.A. and Athanasopoulos G. and Shang H.L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics and Data Analysis*, **55**, 2579-2589.
- Hyndman, R.J. and Lee, A.J. and Wang, E. (2016). Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational Statistics & Data Analysis*, **97**, 16-32.
- Kahn, K.B. (1998). Revisiting top-down versus bottom-up forecasting Methods & Systems. *The Journal of Business Forecasting*. **17**, 14-19.

- Lapide, L. (2006). Top-down & bottom-up forecasting in s&op. *Journal of Business Forecasting Methods & Systems*, Vol. 25 Issue 2, 14-17
- Van Erven, T. and Cugliari, J. (2014). *Game-theoretically Optimal Reconciliation of Contemporaneous Hierarchical Time Series Forecasts*. Springer Lecture Notes in Statistics, pp. 297–317.
- Wickramasuriya, S.L. and Athanasopoulos G. and Hyndman R.J. (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of American Statistical Association*, **114**, 804-819.



ΕΥΡΕΤΗΡΙΟ ΣΥΓΓΡΑΦΕΩΝ (AUTHOR INDEX)

ΑΓΓΕΛΗΣ Ε.	90
ΑΓΙΑΚΛΟΓΛΟΥ Χ.	293
ΑΝΑΣΤΑΣΟΠΟΥΛΟΥ Μ.	308
ΑΠΟΣΤΟΛΙΔΟΥ Ε.	61
ΒΑΣΙΛΕΙΑΔΗΣ Γ.	77
ΒΕΡΟΠΟΥΛΟΥ Γ.	205
ΓΑΛΑΝΗΣ Ο.	90
ΔΑΦΝΗΣ Σ.	105
ΔΟΝΑΤΟΣ Γ.	25
ΖΑΡΑΓΚΑΣ Χ.	118
ΘΕΟΔΟΣΙΑΔΟΥ Ο.	131
ΚΟΛΥΒΑ-ΜΑΧΑΙΡΑ Φ.	61
ΚΟΥΤΡΑΣ Β. Μ.	145
ΚΟΥΤΡΑΣ Μ.Β.	105,145,155
ΚΥΡΑΝΑ Ζ.	168
ΛΟΥΜΠΟΝΙΑΣ Κ.	183
ΛΥΚΟΥ Ρ.	195
ΛΥΜΠΕΡΟΠΟΥΛΟΣ Δ.	155
ΜΑΚΡΗ Ε.	105
ΜΠΑΜΙΧΑ Ε.	205
ΜΩΥΣΙΑΔΗΣ Π.	221
ΝΕΝΕΣ Γ.	262
ΠΑΝΑΓΙΩΤΙΔΟΥ Σ.	262
ΠΑΠΑΤΣΟΥΜΑ Ι.	235
ΠΑΠΟΥΤΣΟΓΛΟΥ Μ.	90
ΡΑΚΙΤΖΗΣ Α.	46,168
ΤΑΣΙΑΣ Κ.	262
ΤΖΑΦΕΡΗ Δ.	278
ΤΣΑΚΛΙΔΗΣ Γ.	131,183,195,278
ΤΣΙΜΠΙΑΝΟΣ Α.	293
ΦΑΡΜΑΚΗΣ Ν.	235
ΦΟΥΣΕΚΗΣ Π.	278
ΑΝΑΣΤΑΣΙΟΥ Α.	308
BAGAVOS P.	324
BOZIKAS A.	336
CHARALABIDES C.	351
ZANGELIDIS L.	361
HATZOPOULOS P.	308
KARLIS D.	392
KARAGRIGORIOU A.	308
MAVRIDOGLOU G.	308
NIKOLAKOPOULOS S.	377
NTZOUFRAS I.	377
PATIL P.N.	324
PITSELIS G.	336
TRIPODIS C.	392
WOOD A.T.A.	324