33ο ΠΑΝΕΛΛΗΝΙΟ ΣΥΝΕΔΡΙΟ ΣΤΑΤΙΣΤΙΚΗΣ

Η Στατιστική στην Οικονομία και τη Διοίκηση

ΠΡΑΚΤΙΚΑ

23-26 Σεπτεμβρίου | 2021

Λάρισα, Πανεπιστήμιο Θεσσαλίας

ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ Ε.Σ.Ι. ΗΡΑΚΛΕΙΤΟΣ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ ΕΤΟΣ ΙΔΡΥΣΗΣ 1984

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

Π Ρ Α Κ Τ Ι Κ Α

33ου Πανελληνίου
Συνεδρίου Στατιστικής

Η Στατιστική στην Οικονομία και τη Διοίκηση

Οργάνωση

ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

Λάρισα (Διαδικτυακά) 23-26 Σεπτεμβρίου 2021

ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ

Σολωμού 5 (Πλατεία Εξαρχείων)

Τηλ. 210 33.03.909

Email: secretariat@esi-stat.gr

http://www.esi-stat.gr

ISSN: 1792-2461

Περιεχόμενα – Contents

Προσκεκλημένες Εργασίες – Invited Papers

Εργασίες στα Ελληνικά – Papers in Greek

Εργασίες στα Αγγλικά – Papers in English

Πρόλογος – Preface

Ολοκληρώθηκε με επιτυχία το 33ο Πανελλήνιο Συνέδριο Στατιστικής που οργανώθηκε διαδικτυακά από 23 ως 26 Σεπτεμβρίου 2021 σε συνεργασία του ΕΣΙ με το Τμήμα Διοίκησης Επιχειρήσεων και το Τμήμα Οικονομικών Επιστημών της Σχολής Οικονομικών και Διοικητικών Επιστημών του Πανεπιστημίου Θεσσαλίας.

Η τελετή έναρξης του συνεδρίου έγινε την Πέμπτη 23 Σεπτεμβρίου 2021 με την έναρξη να κηρύσσει ο Πρύτανης του Πανεπιστημίου Θεσσαλίας Καθηγητής Ζήσης Μαμούρης. Επίσης, στην εναρκτήρια τελετή απηύθυναν χαιρετισμούς ο Πρόεδρος του Τμήματος Διοίκησης Επιχειρήσεων Καθηγητής Λεωνίδας Ανθόπουλος, ο Πρόεδρος του Τμήματος Οικονομικών Επιστημών Καθηγητής Ηλίας Κεβόρκ και ο Πρόεδρος του ΕΣΙ, Καθηγητής Αλέξανδρος Καραγρηγορίου. Το συντονισμό της εκδήλωσης είχε ο Πρόεδρος της Οργανωτικής Επιτροπής του συνεδρίου συνάδελφος Κλεάνθης Συρακούλης.

Η εναρκτήρια τελετή περιελάμβανε την κεντρική ομιλία του Καθηγητή Νικόλαου Βαρσακέλη του Τμήματος Οικονομικών του Πανεπιστημίου Θεσσαλονίκης με τίτλο «Οικονομική επιστήμη και στατιστική: Ένας έρωτας που διαρκεί χρόνια».

Το κύριο μέρος του συνεδρίου άρχισε από το μεσημέρι της Πέμπτης 23/9 με ένα Invited Session και 3 κοινές συνεδρίες. Η ομιλία του συναδέλφου Γ. Ψαρράκου με τίτλο «Από το θεώρημα της μέσης τιμής στην κατασκευή ταυτοτήτων συνδιακύμανσης τύπου Stein» ήταν αφιερωμένη στον Καθηγητή Θ. Κάκουλλο ενώ η προσκεκλημένη ομιλία του συναδέλφου Χ. Μέρκατα νικητή του Ελένειου Βραβείου 2019, στη μνήμη του Καθηγητή του, συναδέλφου Σπύρου Χατζησπύρου. Την πρώτη ημέρα εκτός από την προσκεκλημένη ομιλία του Καθηγητή L. Bermudez προσκλήθηκε και παρουσίασε μέρος της διδακτορικής του διατριβής ο νικητής του Ελένειου Βραβείου 2021, συνάδελφος Κ. Λουμπόνιας που μίλησε για «Modified Kalman Filter for Hidden States Estimation with Censored Measurements».

Οι εργασίες που παρουσιάστηκαν την Παρασκευή (μοναδική ημέρα του συνεδρίου αφιερωμένη σε ομιλίες στην Ελληνική) είχαν ιδιαίτερο ενδιαφέρον και κάλυψαν θεματικές ενότητες σε θέματα Δειγματοληψίας, Στοχαστικών Διαδικασιών, Εφαρμογών της Στατιστικής σε θέματα Οικονομίας και Διοίκησης, χρήσης Χρονοσειρών στην κεφαλαιαγορά, ανάπτυξης της ψηφιακής κοινωνίας και οικονομίας στην Ευρωπαϊκή Ένωση και άλλα.

Το Σάββατο, το συνέδριο συνεχίστηκε με την προσκεκλημένη ομιλία του κ. Νικόλαου Λημνιού του University of Technology of Compiegne της Γαλλίας και θέμα Asymptotic properties of the empirical estimator of the stationary distribution of ergodic semi-Markov processes. Οι εργασίες συνεχίστηκαν με την ειδική συνεδρία Στατιστική/Σεισμολογία υπό την προεδρία της Καθηγήτριας Σεισμολογίας του ΑΠΘ κ. Ελευθερίας Παπαδημητρίου και μια ειδική συνεδρία σε δύο μέρη για νέους Έλληνες στατιστικούς όπου 12 αξιόλογοι Έλληνες ερευνητές επιβεβαίωσαν την ενεργή δραστηριότητα των νέων επιστημόνων στην ανάπτυξη της Στατιστικής. Η τρίτη ημέρα του Συνεδρίου έκλεισε με τις ομιλίες του διεθνούς τμήματος του Συνεδρίου από τις Καθηγήτριες κ. Eugenia Stoimenova της Βουλγαρικής Ακαδημίας Επιστημών, κ. Bojana Milošević, από το Πανεπιστήμιο του Βελιγραδίου και κ. Maria Delores Jiménez-Gamero, από το Πανεπιστήμιο της Σεβίλλης. Την Κυριακή στην τελευταία ημέρα του Συνεδρίου παρουσιάστηκαν οι 5 εργασίες που διαγωνίστηκαν για το Βραβείο Καλύτερης Εργασίας Νέου Έλληνα Στατιστικού το οποίο έχει δωροθετήσει ο Καθηγητής Narayanaswamy Balakrishnan, ο οποίος παραβρέθηκε στην τελετή λήξης και στην απονομή του βραβείου. Για την ιστορία το βραβείο διεκδίκησαν 5 εξαίρετοι νέοι επιστήμονες, οι κ. Θ. Γκελσίνης (Univ. of Rouen), Κ. Γκίλλας (Παν. Πατρών), Δ. Κορδαλής (Παν. Αθηνών), Χ. Μεσελίδης (Παν. Αιγαίου) και Ι. Οικονομίδης (Παν. Αθηνών). Το πρώτο βραβείο απονεμήθηκε, μετά από απόφαση της Επιτροπής Βραβείου, στους Χ. Μεσελίδη για την εργασία του με τίτλο “The use of dual divergence statistics in multiway contingency tables” και στον Ι. Οικονομίδη για την εργασία του με τίτλο “Regression-type approaches for prediction of crop stage percentages”. Την Επιτροπή αξιολόγησης αποτελούσαν οι Καθηγητές Ν. Τσάντας, Ι. Μπασιάκος, Γ. Ψαρράκος και Α. Μπατσίδης. Την ίδια μέρα παρουσιάστηκαν 14 ακόμα ομιλίες από προσκεκλημένους ομιλητές του εξωτερικού. Παράλληλα στα πλαίσια του συνεδρίου διοργανώθηκε Εκπαιδευτικό/Επιμορφωτικό Σεμινάριο με θέμα «Οπτικοποίηση Δεδομένων: Εργαλεία και Τεχνικές για Δεδομένα από Φύλλα Excel» και εισηγήτρια την κα. Κυριακή Τσιλίκα του Παν. Θεσσαλίας. Στο συνέδριο που για πρώτη φορά είχε διεθνή διάσταση με 25 προσκεκλημένους ξένους ομιλητές, 50 συνολικά ομιλίες στην Αγγλική Ελληνική και 25 στην Ελληνική, συμμετείχαν 150 συνάδελφοι και φίλοι του ΕΣΙ. Αυτή η πρωτοβουλία φαίνεται ότι προσδίδει μια άλλη προοπτική στο συνέδριο του ΕΣΙ, με διεθνή αναγνωρισιμότητα.

Το Συνέδριο ολοκληρώθηκε με την τελετή λήξης το βράδυ της Κυριακής 26/9 με την απονομή του Χρυσού Μεταλλίου του ΕΣΙ στους νικητές του βραβείου Καλύτερου Νέου Στατιστικού κ. Χ. Μεσελίδη και Ι. Οικονομίδη. Επίσης, ο Πρόεδρος και το ΔΣ του ΕΣΙ απένειμαν σε ένδειξη εκτίμησης για την ανεκτίμητη και διαχρονική προσφορά τους

στο ΕΣΙ, αναμνηστική πλακέτα στους διατελέσαντες Προέδρους του ΕΣΙ κκ. Πολυχρόνη Μωυσιάδη (ΑΠΘ), Χαράλαμπο Δαμιανού (ΕΚΠΑ) και Χαράλαμπο Χαραλαμπίδη (ΕΚΠΑ) καθώς και στο Επίτιμο μέλος του ΕΣΙ, Καθηγητή N. Balakrishnan. Το συνέδριο έκλεισε με την ομιλία του Προέδρου της Οργανωτικής Επιτροπής συναδέλφου Κ. Συρακούλη.

Για τα πρακτικά υποβλήθηκαν συνολικά 26 εργασίες. Ως κριτές των εργασιών συνεργάστηκαν οι: Ελευθέριος Αγγελής, Σταύρος Βακερούδης, Τρύφων Δάρας, Σπύρος Δαφνής, Γιώργος Ηλιόπουλος, Βασίλειος Κούτρας, Δημήτρης Ιωαννίδης, Αναστάσιος Κατσιλέρος, Δημήτριος Κουγιουμτζής, Νικόλαος Κυριαζής, Φωτεινή Κολυβά-Μαχαίρα, Βασίλειος Κούτρας, Ευφροσύνη Μακρή, Απόστολος Μπατσίδης, Ιωάννης Ντζούφρας, Κίμων Ντότσης, Γεώργιος Νενές, Σταύρος Ντεγιαννάκης Πολυχρόνης Οικονόμου, Κωνσταντίνος Πολίτης, Παναγιώτης Παπασταμούλης, Δημήτρης Παπαναστασίου, Κώστας Τριανταφυλλόπουλος, Γεώργιος Χάλκος, Χαράλαμπος Χαραλαμπίδης και Στέλιος Ψαράκης.

Η Επιτροπή Έκδοσης Πρακτικών του ΕΣΙ εκφράζει τις ευχαριστίες της προς όλους τους κριτές για την επιμελημένη και προσεκτική αξιολόγηση των εργασιών καθώς και προς τον ΥΔ του Πανεπιστημίου Αιγαίου κ. Κίμων Ντότση για την επιμέλεια της έκδοσης.

Η σειρά παρουσίασης των εργασιών στον παρόντα τόμο είναι αλφαβητική με βάση το επώνυμο του πρώτου συγγραφέα. Προηγούνται οι προσκεκλημένες εργασίες ακολουθούν οι εργασίες στην ελληνική και έπονται οι εργασίες στην αγγλική.

ΕΚ ΜΕΡΟΥΣ ΤΟΥ ΔΣ ΤΟΥ ΕΣΙ

Μαλβίνα Βαμβακάρη, Αλέξανδρος Καραγρηγορίου,

Σωτηρία Μαλεφάκη, Σωτήριος Μπερσίμης, Πολυχρόνης Μωϋσιάδης

Γεώργιος Παπαδόπουλος, Γεώργιος Ψαρράκος

Πρόγραμμα Συνεδρίου – Conference Program

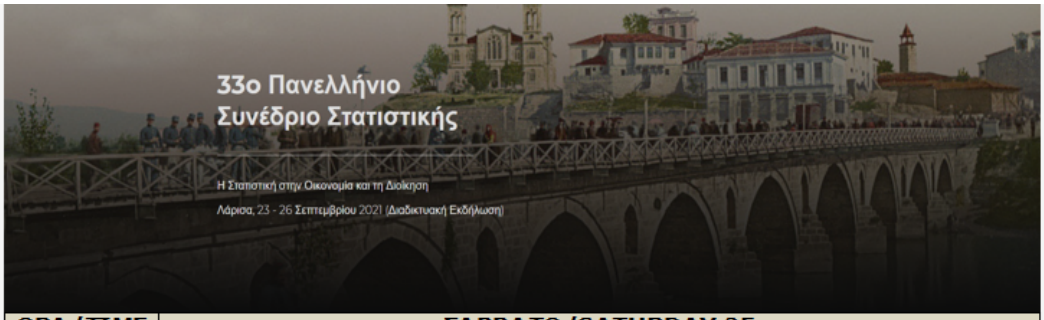
Σημειώσεις/Notes:

1. Τίτλοι στα ελληνικά αφορούν ομιλίες στα ελληνικά (*Πέμπτη πρωί και Παρασκευή*)/ Titles in Greek refer to talks to be delivered in Greek (*Thursday morning and Friday*)

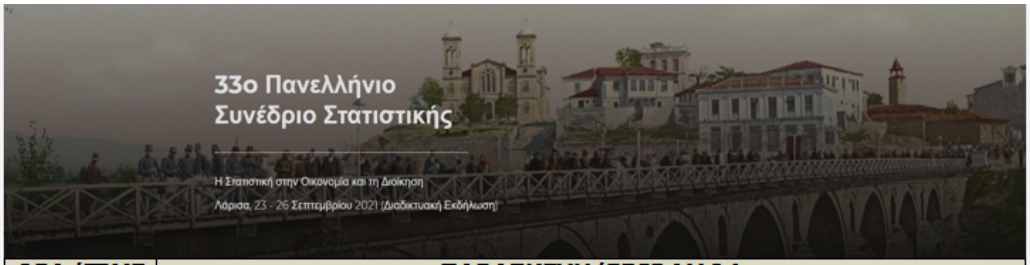
2. Τίτλοι στα αγγλικά αφορούν ομιλίες στα αγγλικά (*Πέμπτη απόγευμα, Σάββατο και Κυριακή*)/ Titles in English refer to talks to be delivered in English (*Thursday afternoon, Saturday and Sunday*)

3. Οι αναφερόμενες ωρες αφορούν στην Ωρα Ελλάδος/All times refer to the Greek timezone (CET+2)

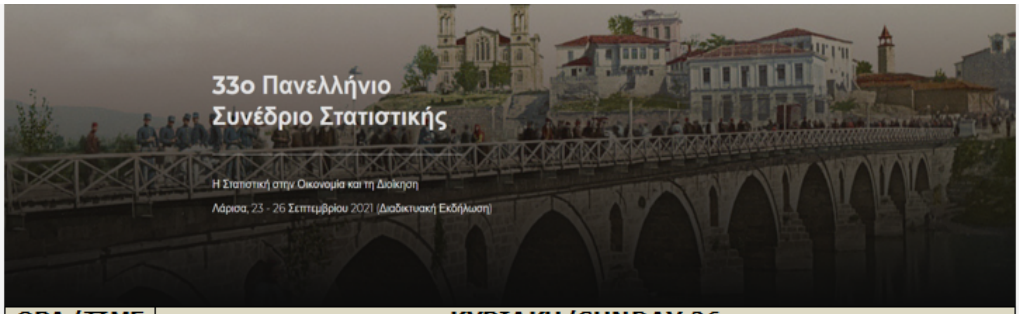
ΩΡΑ/TIME	ΠΕΜΠΤΗ/THURSDAY 23
	Τελετή Έναρξης/Opening Ceremony
12:00	Χαιρετισμοί Παν. Θεσσαλιας & ΕΣΙ
	Κεντρική Ομιλία
	Προεδρεύων: Χ. Χαραλαμπίδης
12:20	**Ν. Βαρσακέλης:** Οικονομική επιστήμη και στατιστική: Ένας έρωτας που διαρκεί χρόνια
13:00-13:10	**Short Break**
	Συνεδρία: Πιθανότητες & Στατιστική
	Προεδρεύων: Σ. Μεϊντάνης
13:10	**Γιώργος Ψαρράκος:** Από το θεώρημα της μέσης τιμής στην κατασκευή ταυτοτήτων συνδιακύμανσης τύπου Stein
13:30	**Χρήστος Ζώτος:** Διατακτική παλινδρόμηση με τη χρήση ροών
13:50	**Γιώργος Τσακλίδης:** Μοντέλο χώρου καταστάσεων με χρήση μη αρνητικής τετραγωνικής συνάρτησης στην εξίσωση κατάστασης
14:10-16:00	**Mid-day Break**
	Invited Session Probability & Statistics I
	Chairperson: D. Karlis
16:00	**L. Bermudez:** Ratemaking for insurance policies with multiple guarantees based on finite mixture of regressions
16:30	**K. Lubonias (*Eleneio PhD Award 2021*):** Modified Kalman Filter for Hidden States Estimation with Censored Measurements
17:00	**Ch. Merkatas (*Eleneio PhD Award 2019*):** Bayesian nonparametric estimation of random dynamical systems
17:30-17-45	**Short Break**
	Contributed Session: Economics & Econometrics
	Chairperson: K. Tsilika
17:45	**Z. Georganta:** Public deficit, debt and stock-flow adjustment (SFA): Statistical and econometric investigation, 1960-2017
18:05	**A. Georgakis:** Further improvements of growing stock volume estimations at Stratum-level with the application of Fay-Herriot model
18:25	**A. Milionis:** Problems and suggestions relating to the estimation of systematic risk. An empirical analysis with reference to the Athens stock exchange
18:45-19:00	**Short Break**
	Contributed Session: Statistics
	Chairperson: D.Kugiumtzis
19:00	**A. Rakitzis:** On the use of runs-based tests in phase I analysis of control charts
19:20	**G. Tzoumerkas:** Power expected posterior prior in shrinkage regression: Model formulation and preliminary results
19:40	**A. Karagrigoriou:** Exponentially vs. light and long-concavity
20:00	**E. Siggiridou:** Estimation of causality in discrete time series using partial mutual information from mixed embedding



ΩΡΑ/TIME	ΣΑΒΒΑΤΟ/SATURDAY 25
	Keynote Speech
	Chairperson: G. Tsaklidis
9:00	**N. Limnios:** Asymptotic properties of the empirical estimator of the stationary distribution of ergodic semi-Markov processes
	Special Session Statistics/Seismology
	Chairperson: E. Papadimitriou
9:45	**P. Bountzis:** Spatiotemporal analysis of microseismicity associated with the intense 2020-2021 earthquake swarm in Corinth gulf through stochastic point processes
10:05	**D.-E. Chorozoglou:** Investigation of the correlation of successive earthquakes preceding main shocks in the Greek territory
10:25	**Ch. Kourouklas:** Short-term clustering features of seismicity in Eastern Aegean Sea: Evaluation and retrospective forecast testing
10:45	**O. Mangira:** Retrospective testing of an earthquake clustering in North Aegean area (Greece)
11:05-11:20	**Short Break**
	Special Session Young (Greek) Statisticians Contribution to Probability and Statistics with Applications
	Chairperson: A. Makrides
11:20	**Ch. Parpoula**: Optimal multiple change-point detection and inference
11:40	**I. Spyroglou:** Mixed models combined with time series analysis as a tool for comparing dynamic changes in plant phenomics
12:00	**D. Pilavakis:** Block bootstrap consistency and bootstrap-based testing of equality of mean functions for functional time series
12:20	**E.-N. Kalligeris:** Continuous-time hidden semi-Markov modeling of time series incidence data
12:40	**A. Kekempanos:** Prediction with limited information: An automatic valuation system approach with incomplete data for real estate application
13:00	**E. Skamnia:** Hotspot identification method based on Andrews curves with an application in monitoring COVID-19 crisis effects on caregiver distress in neurocognitive disorder
13:20-16:00	**Mid-day Break**
	Special Session Young (Greek) Statisticians Contribution to Probability and Statistics with Applications
	Chairperson: E.-N. Kalligeris
16:00	**K. Tasias:** Optimal location of aerial firefighting resources to maximize coverage: A case study of Greece
16:20	**A. Makrides:** Reliability analysis using semi-Markov processes under general classes of distributions
16:40	**P. Vliora:** Sensitivity analysis and tail variability for the Wang's actuarial index
17:00	**K. Ntotsis:** Penalty-based multicollinearity detection criterion
17:20	**M. Anastasopoulou:** CUSUM control charts for monitoring BINARCH(1) processes
17:40	**E. Sofikitou:** Statistical distances, evidence functions and the problem of model selection
18:00-18:15	**Short Break**
	Invited Session Probability & Statistics II
	Chairperson: A. Batsidis
18:15	**Eugenia Stoimenova:** Rank tests based on precedence and exceedance statistics
18:45	**Bojana Milošević:** On Bahadur efficiency in goodness of fit testing: A review of recent results and challenges
19:15	**M. Dolores Jiménez Gamero:** The k-sample problem: A brief review and some new issues



ΩΡΑ / TIME	ΠΑΡΑΣΚΕΥΗ/FRIDAY 24
	Συνεδρία: Στατιστική & Δειγματοληψία
	Προεδρεύων: Γ. Τζαβελάς
9:00	**Πολυχρόνης Οικονόμου:** Μεροληπτική ως προς το μέγεθος δειγματοληψία από πεπερασμένους πληθυσμούς χωρίς επανατοποθέτηση
9:20	**Απόστολος Μπατσίδης:** Διδιάστατα μεροληπτικά δείγματα και σταθμισμένες κατανομές
9:40	**Μαρία Μπατσιάκα & Χριστίνα Χατζημιχαήλ:** Συνεχείς τεθλασμένες κατανομές και συντελεστής μεταβλητότητας
10:00	**Κώστας Πετρόπουλος:** Βελτιωμένοι εκτιμητές για συναρτήσεις παραμέτρων κλίμακας σε μοντέλα μείξης
10:20	**Παναγιώτης Μπομποτάς:** Βελτιωμένη εκτίμηση της μικρότερης παραμέτρου κλίμακας γάμμα κατανομών
10:40-11:00	**Short Break**
	Συνεδρία: ΕΕ & Στοχαστικές & ΣΕΠ
	Προεδρεύων: Α. Μπουρνέτας
11:00	**Ρόδη Λύκου:** Φίλτρο κρυφού ομογενούς μαρκοβιανού συστήματος
11:20	**Οδυσσέας Μπούμπουλης:** Βέλτιστες πολιτικές αποδοχής μοσχευμάτων υπό δυνατότητα συντήρησης
11:40	**Ιωάννης Τριανταφύλλου:** Μη παραμετρικό διάγραμμα ελέγχου προηγήσεων με προσδευτικά επικαιρομένο δείγμα φάσης Ι για μοντέλο αλλαγής σημείου
12:00	**Κωνσταντίνος Φουντουκίδης:** Ροή Αθροίσματος στο max διάγραμμα ελέγχου
12:20	**Αργυρώ Λαφατζή:** Διαγράμματα ελέγχου EWMA για την παρακολούθηση ποσοστών και αναλογιών: Μια συγκριτική μελέτη
12:40-16:00	**Mid-day Break**
	Συνεδρία: Εφαρμοσμένη Στατιστική
	Προεδρεύων: Α. Ρακιτζής
16:00	**Αθανάσιος Κούλης:** Επαγγελματική ανάπτυξη και εξουθένωση (burnout) των εκπαιδευτικών: Εφαρμογή του μοντέλου λογιστικής παλινδρόμησης σε εκπαιδευτικά δεδομένα
16:20	**Μαρία Διαμαντοπούλου:** Ευφυή συστήματα για την πρόβλεψη βιολογικών μεταβλητών: Εφαρμογή σε πρωτογενή στοιχεία δέντρων πεύκης
16:40	**Βίκτωρ Τραπουζανλής:** Αποδοτικοί σχεδιασμοί για πειράματα κρησαρίσματος
17:00	**Αριστείδης Γεωργάκης:** Η στρωμάτωση των δασικών συστάδων ως βάση για εκτιμήσεις μικρής έκτασης
17:20	**Βασίλης Καραγιάννης:** Το δίκτυο συνεργασίας μεταξύ επιστημόνων από το 1988 ως το 2019 στα πρακτικά των συνεδρίων του ΕΣΙ
17:40	**Μαρία Γανοπούλου:** Αιτιατά μοντέλα για την πρόβλεψη των αποτελεσμάτων αγγειοπλαστικής επέμβασης στις χρόνιες ολικές αποφράξεις
18:00-18:20	**Short Break**
	Συνεδρία: Χρονοσειρές & Οικονομετρία
	Προεδρεύων: Π. Οικονόμου
18:20	**Νικόλαος Παπαϊωάννου:** Διόρθωση μεροληψίας στην εκτίμηση της αμοιβαίας πληροφορίας και αντίστοιχων δικτύων συσχέτισης από πολυμεταβλητές χρονοσειρές
18:40	**Βασίλειος Βαρλάγκας:** Κινητός διάμεσος έναντι κινητού μέσου, θεσμικοί επενδυτές έναντι μικροεπενδυτών και αποτελεσματικότητα κεφαλαιαγορών
19:00	**Νικόλαος Γαλανόπουλος:** Πρόβλεψη αναλογιστικών χρονοσειρών: Μελέτη της επίδρασης της "γραμμικοποίησης" και του μετασχηματισμού δεδομένων
19:20	**Μελπομένη Μασούρα:** Μελέτη της ανάπτυξης της ψηφιακής κοινωνίας και οικονομίας στην Ευρωπαϊκή Ένωση
19:40	**Χρήστος Κάτρης:** Μοντέλα πρόβλεψης χρονολογικών σειρών με έντονη εποχικότητα
20:00	**Ευαγγελία Ελένη Πρίσκα:** Πιθανοθεωρητικές θεωρίες πολέμου και η δυνατότητα εφαρμογής τους στη διοίκηση και την οικονομία



ΩΡΑ/TIME	ΚΥΡΙΑΚΗ/SUNDAY 26
	BALA Young Greek Statisticians Award
	Chairperson: N. Tsantas
9:00	**Th. Gkelsinis:** Statistical inferential techniques based on the corrected weighted Kullback-Leibler (CWKL) divergence measure
9:20	**K. Gillias:** Asymmetric exceedance-time model: An optimal threshold approach
9:40	**L. Kordalis:** Time multidimensional semi-Markov chains
10:00	**Ch. Meselidis:** The use of dual divergence statistics in multiway contingency tables
10:20	**I. Oikonomidis:** Regression-type approaches for prediction of crop stage percentages
10:40-11:00	**Short Break**
	Invited Session Probability & Statistics III
	Chairperson: K. Sirakoulis
11:00	**Antonio Di Crescenzo:** Statistical analysis and applications of the multi-sigmoidal deterministic and stochastic logistic growth
11:30	**Claudia Tarantola:** Unfolding models for ordinal data for cyber risk assessment
12:00	**Juan M. Rodríguez-Díaz:** Design optimality for different covariance structures in time-depending multiresponse-multisubject models
12:30	**Nikolaos I. Stilianakis:** Epidemiological modelling of respiratory pathogen transmission
13:00	**Aylin Alin:** Robust gene network analysis
13:30-15:00	**Mid-day Break**
	Invited Session Probability & Statistics IV
	Chairperson: G. Psarrakos
15:00	**Milto Hatzikyriakou:** Some convergence results on generalized Oppenheim expansions
15:30	**Esther Frostig:** A dual risk model with additive and proportional gains: Ruin probability and dividends
16:00	**Azize Haifavi:** An improper integral representation of Linnik's probability densities
16:30-16:45	**Short Break**
16:45	**Maria Longobardi:** New measures of discrimination and their applications
17:15	**Jorge Navarro:** Distortion representations of multivariate distributions
17:45	**Miquel Sordo:** Stochastic comparisons of some distances between random variables
18:15-18:30	**Short Break**
	Invited Session Probability & Statistics V
	Chairperson: S. Malefaki
18:30	**A. Lisnianksi:** Dynamic large scale multi-state system performability. Concepts, measures, Lz-transform evaluation method
19:00	**V. Koutras:** Stochastic modelling of software rejuvenation: Recent advances and future directions
19:30	**I. Frenkel:** Operational availability and performance analysis of the multi-state multi-drive electric propulsion system of the Arctic icebreaker gas tanker "Christophe de Margerie"
20:00	**Τελετή Λήξης/Closing Ceremony**

Χορηγοί – Sponsors

Οργανωτική Επιτροπή – Local Organizing Committee

- Συρακούλης Κλεάνθης (πρόεδρος), Αναπληρωτής Καθηγητής Τμήματος Διοίκησης Επιχειρήσεων Πανεπιστημίου Θεσσαλίας sirakoul@uth.gr
- Τσέλιος Δημήτριος, Αναπληρωτής Καθηγητής Τμήματος Διοίκησης Επιχειρήσεων Πανεπιστημίου Θεσσαλίας dtselios@uth.gr
- Κουστέλιος Αθανάσιος, Καθηγητής Τμήματος Διοίκησης Επιχειρήσεων Πανεπιστημίου Θεσσαλίας, akoustel@uth.gr
- Κεβόρκ Ηλίας, Καθηγητής Τμήματος Οικονομικών Επιστημών Πανεπιστημίου Θεσσαλίας, kevork@econ.uth.gr
- Τσιλίκα Κυριακή, Επίκουρος Καθηγήτρια Τμήματος Οικονομικών Επιστημών Πανεπιστημίου Θεσσαλίας, ktsilika@econ.uth.gr
- Νάκας Χρήστος, Καθηγητής Τμήματος Γεωπονίας Φυτικής Παραγωγής και Αγροτικού Περιβάλλοντος Πανεπιστημίου Θεσσαλίας, cnakas@uth.gr
- Αγγελής Ελευθέριος, Καθηγητής Τμήματος Πληροφορικής Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης, lef@csd.auth.gr
- Μωυσιάδης Πολυχρόνης, Ομότιμος Καθηγητής Τμήματος Μαθηματικών Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης, cmoi@math.auth.gr

Τεχνική Υποστήριξη & Σχεδιασμός Ιστοσελίδας
Technical Support & Web Design

- Α. Βασιλειάδης, iED Εταιρεία Πληροφορικής Λάρισα, tasos@entre.gr
- Κ. Παρίζα, kpariza@ied.eu

Επιστημονική Επιτροπή – Scientific Committee

- 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.
- Α. Καραγρηγορίου, Καθηγητής Πανεπιστημίου Αιγαίου.
- Η. Κεβόρκ, Καθηγητής Πανεπιστημίου Θεσσαλίας.
- Φ. Κολυβά-Μαχαίρα, Αν. Καθηγήτρια Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης.
- Χ. Κουκουβίνος, Καθηγητής Εθνικού Μετσόβιου Πολυτεχνείου.
- Δ. Κωνσταντινίδης, Καθηγητής Πανεπιστημίου Αιγαίου.
- Σ. Μαλεφάκη, Επ. Καθηγήτρια Πανεπιστημίου Πατρών.
- Α. Μπατσίδης, Επ. Καθηγητής Πανεπιστημίου Ιωαννίνων.
- Α. Μπουρνέτας, Καθηγητής Πανεπιστημίου Αθηνών.
- Π. Μωυσιάδης, Ομότιμος Καθηγητής Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης.
- Π. Οικονόμου, Αν. 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.
- Κ. Συρακούλης, Αν. Καθηγητής του Πανεπιστημίου Θεσσαλίας.
- Α. Ρακιτζής, Επ. Καθηγητής Πανεπιστημίου Αιγαίου.
- Ι. Τριανταφύλλου, Επ. Καθηγητής Πανεπιστημίου Θεσσαλίας.
- Γ. Τσακλίδης, Καθηγητής Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης.
- Τ. Χριστοφίδης, Καθηγητής Πανεπιστημίου Κύπρου.
- Γ. Ψαρράκος, Αν. Καθηγητής Πανεπιστημίου Πειραιώς.

Προσκεκλημένες Εργασίες

Invited Papers

Greek Statistical Institute
Proceedings of the 33rd Panhellenic Statistics Conference (2021), pp. 17−25

ROBUST CONNECTIVITY SCORES FOR GENE CO-EXPRESSION NETWORK ANALYSIS

***A. Alin*[1], *A. Olmez*[1]**
[1]Dokuz Eylul University
aylin.alin@deu.edu.tr

ABSTRACT

With the advantage of high throughput technologies, gene network analysis became a crucial tool in bioinformatics. Large-scale microarray expression gene network studies explore the identification, functions and relations of individual genes or their products across biological conditions. One of the network-based methods is gene coexpression networks, which are used for describing associations between high-throughput expression patterns of genes. Connectivity scores, calculated using model-based approaches, can be used to build edges in networks. Herein, we introduce new model-based scores that are robust to noise and non-normality in the data.

Keywords: Gene Co-Expression Network, Partial Least Squares Regression, Partial Robust M Regression.

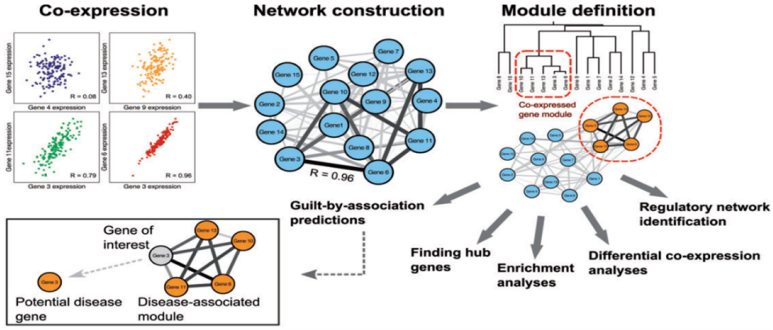
1. INTRODUCTION

Systematically identifying all molecules within a living cell and how they interact is of great interest in biological research. Networks are convenient representations of patterns of interaction between biological elements. Those networks can be used for various purposes, including candidate disease gene prioritization, functional gene annotation, and the identification of regulatory genes. Unlike experimental detection of biological networks which is an expensive and labor-intensive process, computational methods on high-throughput gene expression data allow revealing the interactions and building networks. Gene expression is the process of converting the information stored in our DNA to a functional product such as proteins or other molecules. The amount of functional product produced is referred to as the level of expression. With the advantage of high throughput technologies, gene co-expression networks have become a crucial tool in biological studies. Figure 1 provides various purposes of Gene Co-Expression networks (GCN) such as candidate disease gene prioritization, functional gene annotation and the identification of regulatory genes (Van Dam et al. (2018)).

Co-expression networks are mainly able to identify correlations, but do not normally give information about causality or distinguish between regulatory or regulated

genes. Model-based approaches can be a solution. The idea is based on how well the expression level of a given gene can be predicted from the other genes. Even though a regression model that is based on the ordinary least squares approach is a very popular modelling method, it is not effective in this scenario. A high throughput data set includes expression levels of thousands of genes that are measured at only a handful of time points or for a handful of samples where ordinary least squares estimators are not efficient. Datta (2001) and Pihur et al. (2008) showed that Partial Least Squares Regression (PLSR) scores serve as good indicators of biological relationships. The idea of PLSR was to get a smaller set of uncorrelated, so-called latent variables that would have maximum covariance with the response.

Figure 1: *Construction of Gene Co-Expression Network (Van Dam et al. (2018))*



Datta (2001) proposed PLSR for gene co-expression network construction, and used for classification into temporal groups using expression levels during sporulation of Saccharomyces cerevisiae (SC, budding yeast). There have been many studies on gene expression data using PLSR. Land Jr et al. (2011) claimed PLS could be used to discover biomarkers of colon cancer. Faria et al. (2011) investigated brain tumor classes according their biochemical changes and patterns. Huang & Pan (2003) investigated multi-classification toxicology to show some prognostic significance of breast cancer. Ding (2014) used PLS to identify prognostic genes in end-stage renal failure patients. Shokri-Kojori et al. (2017) used PLSR to expose alcohol and resting brain activity relationships. Olmez (2017) constructed gene networks for three brain regions of a developing mouse brain in the embryonic period. Chan et al. (2018) showed the identification of putative interacting partners of gene markers for frontotemporal dementia gene markers in human brain.

High-throughput gene expression data sets are subject to noise and error (Detours, V. et al. (2003)). As popular as PLSR is, it is also vulnerable to outlying data points. In this paper, we propose a robust similarity measure that is both based on causality relation between genes and robust to noise and error. In the following section we will go over PLSR-based similarity/association measures (connectivity scores) that will be used for building the GCN, and present new robust connectivitiy scores. In Section 3, the performance of both networks will be compared on simulated data sets regarding

how many correct interactions are found by networks under contamination of different levels.

2. METHODS

There are several methods used for GCN analysis. One of the most popular and widely used methods is Weighted Gene Co-Expression analysis. Apart from this, Pearson Correlation, Random Forest and Partial Least Squares Regression methods are frequently preferred. In this section, connectivity scores based on PLSR are presented, and the new robust scores are introduced.

2.1 Partial Least Squares Regression

High-throughput data generally suffer from noise, missing values, and multicollinearity that would requare methods with new variables called as latent variables. One of the well known methods is PLSR. The idea is to build a model on latent variables having maximum covariance and being linear combinations of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

$\boldsymbol{Y}$ $(n\text{x}m)$ is a matrix of response variables and $\boldsymbol{X}$ $(n\text{x}k)$ is a matrix of independent variables (Equation (1)), where n is the number of observations, m is the number of $\boldsymbol{Y}$ variables, and k is the number of $\boldsymbol{X}$ variables. PLSR decomposes $\boldsymbol{X}$ and $\boldsymbol{Y}$ matrices as the linear models in Equations (2) and (3). $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)$ is *iid* normally distributed error with $E(\boldsymbol{\varepsilon}_i) = 0$ and $Var(\boldsymbol{\varepsilon}_i) = \sigma^2$. $\boldsymbol{T}$ $(n\text{x}A)$ is the score matrix of $\boldsymbol{X}$, and $\boldsymbol{P}$ $(k\text{x}A)$ is the loading matrix of $\boldsymbol{X}$. $\boldsymbol{U}$ $(n\text{x}A)$ and $\boldsymbol{C}$ $(m\text{x}A)$ are, respectively, the score and loading matrices for $\boldsymbol{Y}$. $\boldsymbol{E}$ $(n\text{x}k)$ and $\boldsymbol{F}$ $(n\text{x}m)$ are residual matrices for $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, while A corresponds to number of latent variables.

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}' + \boldsymbol{E} \tag{2}$$

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{C}' + \boldsymbol{G} \tag{3}$$

The classical and standard algorithm for PLSR model estimation is the Non-linear Iterative Partial Least Squares (NIPALS). The NIPALS algorithm, which works with centralized and scaled $\boldsymbol{X}$ and $\boldsymbol{Y}$ variables, consists of the following steps (Wold, S. et al. (2001)). These steps (of the NIPALS algorithm) are being repeated A times, where A is the number of the predetermined/selected latent variables.

Step 1: A single $\boldsymbol{Y}$ variable column is assigned to $\boldsymbol{u}$, $\boldsymbol{u} = \boldsymbol{y}$.
Step 2: The $\boldsymbol{\omega}$ weight vector for $\boldsymbol{X}$ is obtained with $\boldsymbol{u}$, $\boldsymbol{\omega} = \boldsymbol{X'u}/\boldsymbol{u'u}$.
Step 3: The $\boldsymbol{\omega}$ is scaled, $\boldsymbol{\omega} = \boldsymbol{\omega}/\|\boldsymbol{\omega}\|$.
Step 4: The $\boldsymbol{X}$ scores, is calculated, $\boldsymbol{t} = \boldsymbol{X\omega}$.
Step 5: The $\boldsymbol{c}$ loading vector for $\boldsymbol{y}$ is calculated, $\boldsymbol{c} = \boldsymbol{yy'}/\boldsymbol{t't}$.
Step 6: The $\boldsymbol{c}$ is scaled, $\boldsymbol{c} = \boldsymbol{c}/\|\boldsymbol{c}\|$.
Step 7: The $\boldsymbol{u}$ vector is updated for next iteration, $\boldsymbol{u} = \boldsymbol{yc}/\boldsymbol{c'c}$.

Step 8: The loadings of $\boldsymbol{X}$ is computed, $\boldsymbol{p} = \boldsymbol{X't}/\boldsymbol{t't}$.
Step 9: The deflation of $\boldsymbol{X}$ variables, $\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{tp'}$.
Step 10: The deflation of $\boldsymbol{Y}$ variables, $\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{tc'}$ (No deflation is needed when there is only one response variable).

2.2 Partial Robust M Regression

Unfortunately, the PLSR algorithm (NIPALS) is not robust to outliers, leverage points or non-normal error terms. The Partial Robust M Regression (PRM) introduced by Serneels et al. (2005) based on M estimator (Huber, P., (1981)) is one of the robust alternatives. The main advantage of PRM is downweighting outliers and leverage points at the same time. The Least Squares (LS) estimator of β is defined as

$$\widehat{\boldsymbol{\beta}}_{LS} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - x_i\boldsymbol{\beta})^2. \tag{4}$$

That is known to be the optimal when error terms follow normal distribution. However, if the error terms have another distribution, such as a heavy tailed distribution, the LS estimator loses its optimality. M-estimator is obtained by replacing the squares in $\widehat{\boldsymbol{\beta}}_{LS}$ by a symmetric and non-decreasing loss function. The equation can be rewritten as

$$\widehat{\boldsymbol{\beta}}_{M} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(y_i - x_i\boldsymbol{\beta}). \tag{5}$$

To give less importance to large residuals, one chooses a bounded loss function ρ, resulting in a more robust estimator than LS. Let $r_i = y_i - x_i\boldsymbol{\beta}$ denote the residuals in the $\widehat{\boldsymbol{\beta}}_{M}$. The weight that provides robustness to residuals is $w_i^r = \rho(r_i)/{r_i}^2$. Then the formula in Equation (5) is rewritten as

$$\widehat{\boldsymbol{\beta}}_{M} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} {w_i}^r (y_i - x_i\boldsymbol{\beta})^2. \tag{6}$$

Another important step is to accommodate for leverage points. The weight in $\widehat{\boldsymbol{\beta}}_{M}$ will be multiplied by a second weight. This leads the new robust estimator (Equation (7)):

$$\widehat{\boldsymbol{\beta}}_{RM} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} {w_i}^r {w_i}^x (y_i - x_i\boldsymbol{\beta})^2. \tag{7}$$

The w_i^r weights have been computed as

$$w_i^r = f\Big(\frac{r_i}{\hat{\sigma}}, c\Big), \tag{8}$$

where, $\hat{\sigma} = MAD(r_1, ..., r_j)$, $j = 1, ..., k$ and the MAD is the median absolute deviation. The tuning constant that causes flatter weight function for larger values is taken as

c=4. The weight function f is called the fair function (Equation (9)). There are several options for the weight function. However, many experiments (Cummins et al. (1995)) showed that Fair function and the constant taken 4 are the proper compromise between robustness and efficiency.

$$f(z,c) = \frac{1}{\left(1+|\frac{z}{c}|\right)^2}. \tag{9}$$

The weights w_i^x have been computed as follows:

$$w_i^x = f\Big(\frac{||t_i - med_{L_1}(T)||}{median_i||t_i - med_{L_1}(T)||}, c\Big), \tag{10}$$

where, $||.||$ stands for the Euclidean norm and $med_{L_1}(T)$ denotes the L1-median from the X scores. The PRM algorithm is as follows:

Step 1: Robust starting values for the weights $w_i = w_i^r w_i$ using the Equation (6) with residuals calculated by subtracting the median of the response variable from each response data are obtained.
Step 2: Then PLSR is performed with weighted $\boldsymbol{X^w} = \boldsymbol{X}\sqrt{w_i}$ and $\boldsymbol{Y^w} = \boldsymbol{Y}\sqrt{w_i}$.
Step 3: The residuals and weights are recomputed and updated until convergence.
Step 4: The final estimate $\hat{\boldsymbol{\beta}}$ is obtained from the last weighted PLS step.

2.3 Connectivity Scores

Connectivity scores are useful for exploring associations in gene network structure (Pihur et al. (2008)). If there is an edge between two nodes (ith and jth genes), this edge is formed by a statistically significant connectivity score calculated in the presence of other genes:

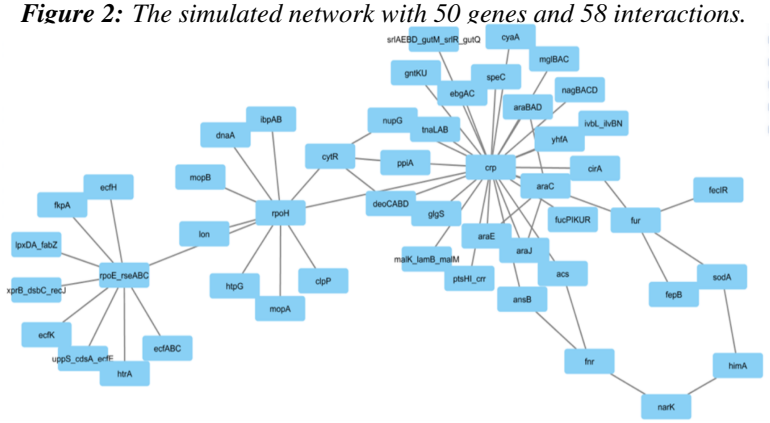
$$\hat{s}_{ij(l)} = \frac{\sum_{a=1}^{A} c_{i(l)}^a \omega_{ij(l)}^a + \sum_{a=1}^{A} c_{j(l)}^a \omega_{ji(l)}^a}{2}, \tag{11}$$

where, l represents the method (PLSR or PRM) used for calculating the loadings and weights in Equation(9). The left part of the summation, $\sum_{a=1}^{A} c_{i(l)}^a \omega_{ij(l)}^a$, is the response variable for the i^{th} gene, $c_{i(l)}^a$ is the loading of the i^{th} gene on the A-th latent variable, and $\omega_{ij(l)}^a$ is the contribution of the j^{th} gene on the A-th latent variable when the i^{th} gene is modeled by other genes. To the right of summation we have the symmetric half, $\sum_{a=1}^{A} c_{j(l)}^a \omega_{ji(l)}^a$, where gene j is the response variable. The interpretations of $c_{j(l)}^a$ and $\omega_{ji(l)}^a$ are the same, except this time the j^{th} gene is taken as response variable. Once connectivity scores are calculated for each gene pair, the gene network can be constructed with the significant scores. To decide if a connectivity score is significant some threshold values, denoted as $\boldsymbol{\epsilon}$, are used. The modular structure and the sensitivity of the gene network changes with the choice of ϵ which can be $\boldsymbol{\epsilon} \in \{0.35, 0.4, 0.45, 0.5, 0.55\}$ (Gil et al. (2010)).

3. SIMULATION STUDY

The data in the simulation study were generated with the SynTRen program (Van den Bulcke et al. (2006)) that simulates biological networks with known underlying structures. We generated a network having 50 genes and 50 sample points with 58 interactions (the default parameters' values are kept as they are). The simulated network is illustrated in Figure 2 using the Cytoscape software (Shannon et al. (2003)).

Figure 2: *The simulated network with 50 genes and 58 interactions.*



In order to demonstrate robustness against outliers and leverage points, outliers were added to the simulation data. Following Liebmann et al. (2010), the x and y variables were changed with outlying observations based on (Equations (12) and (13)):

$$x_{ij.out} = x_{j.max} U(\alpha, \beta), \tag{12}$$

where $x_{j.max}$ is the maximum of an x variable, and $U(\alpha, \beta)$ is the uniform distribution with $\alpha = 3$ and $\beta = 10$.

$$y_{i.out} = y_i \pm U(y_{max}, 2(y_{mean})), \tag{13}$$

where, $y_{i.out}$ is calculated by adding or substracting a value ranging between the maximum of y and twice the mean of y.

PLSR and PRM-based connectivity scores were calculated for the obtained data set. Our criterion is the number of true interactions obtained by the connectivity scores calculated from the PLSR and PRM methods. The absolute value of the connectivity scores calculated for each gene pair was compared with different thresholds, $\boldsymbol{\epsilon} = \{0.35, 0.45, 0.55\}$. The connectivity scores greater than the considered threshold value indicate statistically significant interactions. For the contamination free data (data with no outliers) 123 and 103 statistically significant interactions were obtained with PLSR and PRM-based connectivity scores, respectively. We identified 15 true interactions out of 123 (12.19% accuracy rate) with PLSR-based connectivity scores while PRM-based scores revealed 13 true interactions out of 103 (12.62% accuracy rate). When the data

had 5% outliers, PLSR-based scores flagged 11 true interactions out of 165 (6.66% accuracy rate), and the PRM scores identified 15 true interactions out of 114 (13,15% accuracy rate). Table 1 presents the interaction results for 0%, 5%, 10% and 15% contamination levels for three thresholds. For the scenario considered, regardless of the threshold, the more the data include abnormal values in both response and independent variables, the larger the accuracy rate of proposed PRM-based robust connectivity scores. As pointed by Gil et al. (2010), sensitivity changes with the threshold. The larger the contamination in the data, the smaller the threshold is needed for better sensitivity. At the same time, PLSR fails to retain a respectable level verifying its inability under the presence of outlying observations.

Table 1: *Ratio of true interactions identified by PLSR and PRM based scores under different contamination and threshold levels*

ϵ	Method	0%	5%	10%	15%
0.35	PLSR	6.25%	3.41%	1.62%	0.04%
	PRM	5.26%	5.3%	6.75%	3.65%
0.45	PLSR	12.19%	6.66%	6.56%	2.99%
	PRM	12.62%	13.15%	21.0%	17.98%
0.55	PLSR	9.09%	3.75%	4.05%	2.22%
	PRM	8.33%	15.38%	13.33%	11.11%

ACKNOWLEDGEMENT

We thank an anonymous referee whose comments/suggestions helped improve and clarify this manuscript. Ayca Olmez was supported by a grant from the Department of Scientific Research Projects (BAP), project no:2021.KB.FEN. 038.

ΠΕΡΙΛΗΨΗ

Με το πλεονέκτημα των τεχνολογιών υψηλής απόδοσης, η γονιδιακή ανάλυση αποτελεί πλέον ενα καθοριστικό εργαλείο στη βιοπληροφορική. Μελέτες δικτύου γονιδίων έκφρασης μικροσυστοιχιών μεγάλης κλίμακας διερευνούν την αναγνώριση, τις λειτουργίες και τις σχέσεις μεμονωμένων γονιδίων κάτω από διάφορες βιολογικές καταστάσεις. Μία από τις μεθόδους αυτές είναι τα δίκτυα συνέκφρασης γονιδίων (gene co-expression networks), τα οποία χρησιμοποιούνται για την περιγραφή συσχετίσεων μεταξύ προτύπων έκφρασης γονιδίων υψηλής απόδοσης. Τα connectivity scores (σκορ συνδεσιμότητας), που υπολογίζονται χρησιμοποιώντας προσεγγίσεις που βασίζονται στη μοντελοποίηση, μπορούν να χρησιμοποιηθούν για τη δημιουργία ακμών δικτύων. Στην εργασία αυτή, εισάγουμε νέους τύπους βαθμολόγησης (σκορς) βάσει μοντέλων που είναι ανθεκτικοί στο θόρυβο και στην απουσία κανονικότητας στα δεδομένα.

REFERENCES

Chan, S. C., Wu, H. C., Lin, J. Q. and Zhang, Z. G. (2018). A Partial least squares-based regression approach for analysis of frontotemporal dementia gene markers in human brain gene microarray data. *In 2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)* (pp. 1-5). IEEE.

Cummins, D. J. and Andrews, C. W. (1995). Iteratively reweighted partial least squares: A performance analysis by Monte Carlo simulation. *Journal of Chemometrics*, **9(6)**, 489-507.

Datta, S. (2001). Exploring relationships in gene expressions: a partial least squares approach. *Gene Expression*, **9(6)**, 249-255.

Detours, V., Dumont, J. E., Bersini, H. and Maenhaut, C. (2003). Integration and cross-validation of high-throughput gene expression data: comparing heterogeneous data sets. *FEBS Letters*, **546(1)**, 98-102.

Ding, S., Xu, Y., Hao, T. and Ma, P. (2014). Partial least squares based gene expression analysis in renal failure. *Diagnostic pathology*, **9(1)**, 1-6.

Faria, A., Macedo Jr., F., Marsaioli, A., Ferreira, M. and Cendes, F. (2011). Classification of brain tumor extracts by high resolution 1h mrs using partial least squares discriminant analysis. *Brazilian Journal of Medical and Biological Research*, **44(2)**, 149-164.

Huang, X. and Pan, W. (2003). Linear regression and two-class classification with gene expression data. *Bioinformatics*, **19(16)**, 2072-2078.

Gill, R., Datta, S. and Datta, S. (2010). A statistical framework for differential network analysis from microarray data. *BMC Bioinformatics*, **11(1)**, 95.

Huber, P., J. (1981), *Robust Statistics*, New York, Wiley.

Pihur, V., Datta, S. and Datta, S. (2008). Reconstruction of genetic association networks from microarray data: a partial least squares approach. *Bioinformatics*, **24(4)**, 561-568.

Land Jr, W., H., Ford, W., Park, J., W., Mathur, R., Hotchkiss, N., Heine, J., Eschrich, S., Qiao, X. and Yeatman, T. (2011). Partial least squares (pls) applied to medical bioinformatics. *Procedia Computer Science*, **6**, 273-278.

Liebmann, B., Filzmoser, P. and Varmuza, K. (2010). Robust and classical pls regression compared. *Journal of Chemometrics*, **24(3-4)**, 111-120.

Olmez, A. (2017). Partial least sqaures method for the analysis of gene expression data. Master Thesis, *Dokuz Eylul University*, Department of Statistics.

Serneels, S., Croux, C., Filzmoser, P. and Van Espen, P. J. (2005). Partial robust m-regression. *Chemometrics and Intelligent Laboratory Systems*, **79(1-2)**, 55-64.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, **13(11)**, 2498-2504.

Shokri-Kojori, E., Tomasi, D., Wiers, C., E., Wang, G., J. and Volkow, N., D. (2017).

Alcohol affects brain functional connectivity and its coupling with behavior: greater effects in male heavy drinkers. *Molecular Psychiatry*, **22(8)**, 1185.

Van Dam, S., Vosa, U., van der Graaf, A., Franke, L. and de Magalhaes, J., P. (2018). Gene co-expression analysis for functional classification and gene disease predictions. *Briefings in Bioinformatics*, **19(4)**, 575-592.

Van den Bulcke, T., Van Leemput, K., Naudts, B., van Remortel, P., Ma, H., Verschoren, A., De Moor, B. and Marchal, K. (2006). Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics*, **7(1)**, 43.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. *Multivariate Analysis*, 391-420.

Wold, S., Sjistrim, M. and Eriksson, L. (2001). Pls-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, **58(2)**, 109-130.

Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, **4(1)**.

Greek Statistical Institute
Proceedings of the 33rd Panhellenic Statistics Conference (2021), pp. 26−38

STATISTICAL ANALYSIS AND APPLICATIONS OF THE MULTI-SIGMOIDAL DETERMINISTIC AND STOCHASTIC LOGISTIC GROWTH

***Antonio Di Crescenzo*[1], *Paola Paraggio*[1], *Patricia Román-Román*[2], *Francisco Torres-Ruiz*[2]**
[1]University of Salerno
{adicrescenzo, pparaggio}@unisa.it
[2]University of Granada
{proman, fdeasis}@ugr.es

ABSTRACT

We consider a multi-sigmoidal generalization of the logistic growth model. The deterministic model is provided together with its stochastic counterpart. More in detail, we analyse two different birth-death processes with linear and quadratic rates, respectively. From the latter we derive a more manageable diffusive approximation by means of a suitable scaling. Furthermore, we study two possible strategies to obtain the maximum likelihood estimates of the parameters. To validate the described procedures, we conclude with a simulation study. The first-passage-time problem is also addressed.

Keywords: Logistic model, multi-sigmoidal growth models, birth-death process, diffusion process, maximum likelihood estimation, first-passage-time problem.

1. INTRODUCTION

The exponential curve is the most common basic model to describe growth of populations in ideal conditions. However, such kind of growth does not occur in nature apart from short time periods. For most living species, indeed, there exists a critical density beyond which the relative population does not find sufficient environmental factors to grow and reproduce. Mathematical models which take into account environmental factors that limit the growth rate of population are characterized by a S-shape and for this reason are called sigmoidal. The logistic model, more in detail, is a sigmoidal growth model with an initial slow growth followed by an explosion of exponential-type which finally flattens up to an equilibrium status, known as carrying capacity. The application of sigmoidal curves are various and they involve several contexts of interest, from biology to medicine, from ecology to software reliability (see, for instance, Erto *et al.* (2020)). For example, in the recent works of Rajasekar *et al.* (2020) the authors analyse a stochastic version of SIR model for the diffusion of the COVID-19 pandemic, by

considering a logistic-kind growth for the susceptible individuals.

Anyway, it is possible that a population reaches its limit value after various successive steps. This is the reason why recent investigations address their interest to a generalization of the sigmoidal models by introducing multiple inflections. Such generalizations are the so-called multi-sigmoidal models (cf. Román-Román *et al.* (2019)). The multi-sigmoidal logistic model, in particular, is appropriate to describe maturation of some fruit species as peaches or coffee berries (see Figure 1) which show a growth trend with multiple fluctuations.

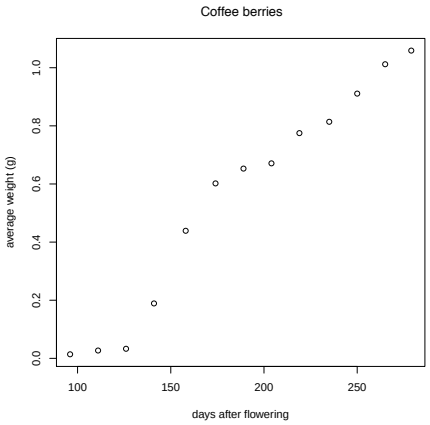


Figure 1: *The multi-sigmoidal growth of coffee berries. The data are taken from Cuhna and Volpe (2011).*

This work is a brief summary of a larger study concerning multi-sigmoidal logistic growth model. The deterministic model together with the birth-death processes and the diffusive approximation have been analysed widely in Di Crescenzo *et al.* (2021a). Whereas, the statistical analysis of the above-mentioned model is the subject of a paper submitted for publication (cf. Di Crescenzo *et al.* (2021b)).

2. THE DETERMINISTIC MODEL

The multi-sigmoidal logistic curve $l_m(t)$ satisfies a generalized version of the Cauchy problem related to the classical logistic model, i.e.

$$\frac{d}{dt}l_m(t) = h_\theta(t)l_m(t), \qquad t \geq t_0, \qquad l_m(t_0) = l_0, \tag{1}$$

where

$$h_\theta(t) = \frac{P_\beta(t)e^{-Q_\beta(t)}}{\eta + e^{-Q_\beta(t)}}, \tag{2}$$

with

$$Q_\beta(t) = \sum_{i=1}^{p} \beta_i t^i, \qquad \beta_p > 0, \qquad P_\beta(t) = \frac{d}{dt}Q_\beta(t), \tag{3}$$

for $\eta > 0$, $\beta_1, \ldots, \beta_{p-1} \in \mathbb{R}$, $\beta_p > 0$, $\theta = (\eta, \beta^T)^T$ and $\beta^T = (\beta_1, \ldots, \beta_p)$, $p \in \mathbb{N}$. Note that when $p = 1$, $Q_\beta(t)$ is linear and from Eq. (1) we come to the classical logistic equation. The solution of the initial value problem (1) is given by

$$l_m(t) = l_0 \frac{\eta + e^{-Q_\beta(t_0)}}{\eta + e^{-Q_\beta(t)}}, \qquad t \geq t_0. \tag{4}$$

In Eqs. (1) and (4) the subscript m means 'multi-sigmoidal'. Various choices of the parameters $\eta, \beta_1, \ldots, \beta_p$ lead to different kinds of shape characterized by multiple inflection points. See, for instance, Figure 2. It is possible to show that the carrying

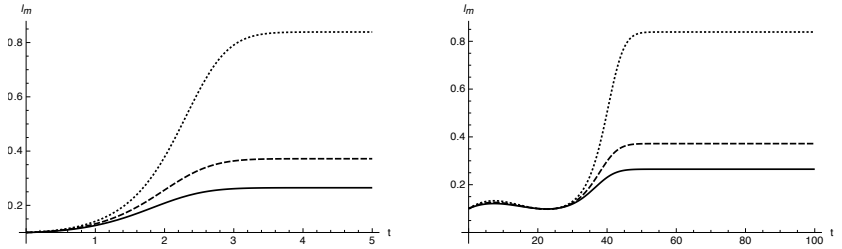


Figure 2: *The multi-sigmoidal logistic function for $p = 3$, $t_0 = 0$, $l_0 = 0.1$, $\beta_1 = 0.1$ (a) $\beta_2 = 0.2$, $\beta_3 = 0.1$, (b) $\beta_2 = -0.009$, $\beta_3 = 0.0002$. In both cases $\eta = e^{-0.5}, e^{-1}, e^{-2}$ (from bottom to top).*

capacity of the model depends on the relevant parameters θ and on the initial condition $l_m(t_0) = l_0$. More in detail, it is given by C/η with $C = C(l_0, \theta, t_0) = l_0(\eta + e^{-Q_\beta(t_0)})$. Indeed, from the assumption $\beta_p > 0$, one has the following limit

$$\lim_{t \to \infty} l_m(t) = l_0 \frac{\eta + e^{-Q_\beta(t_0)}}{\eta} = \frac{C}{\eta}.$$

A key-role in the analysis of the multi-sigmoidal logistic function is played by the inflection points which give to the curve the characteristic shape. By performing the second derivative of Eq. (4), it is possible to show that the inflection points are the solutions of the following equation (in the unknown $t \geq t_0$)

$$\frac{d^2}{dt^2} Q_\beta(t) = \left(\frac{d}{dt} Q_\beta(t) \right)^2 \frac{\eta - e^{-Q_\beta(t)}}{\eta + e^{-Q_\beta(t)}}, \tag{5}$$

with $Q_\beta(t)$ defined in Eq. (3). Due to the transcendental nature of the above-mentioned equation, one is forced to use numerical methods to solve it.

Example 1. *With reference to the real data given in Section 1, we now consider an application of the multi-sigmoidal model. To avoid numerical problems, we perform a time shifting so that $t_0 = 0$ (this does not affect the generality of the analysis, as pointed*

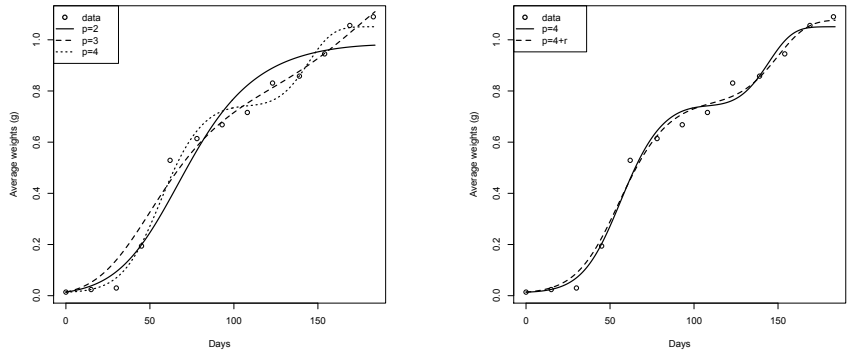


Figure 3: *Fitted multi-sigmoidal logistic curve for the coffee berries with (a) integer and (b) non-integer degrees, with* $r = -4.498494 \cdot 10^{-1}$ *in case (b).*

out in Remark 2.1 of Di Crescenzo et al. (2021a)). Specifically, we determine the values of the parameters θ *minimizing the square error* S_p *defined as follows*

$$S_p(\theta) = \sum_{i=1}^{n} (y_i - l_m(t_i))^2, \qquad \theta = (\eta, \beta^T)^T$$

where y_i *are the data,* t_i *are the shifted times for* $i = 1, \ldots, n$ *and* p *is the degree of* Q_β*. As shown in Figure 3-(a), the best fit is attained when* $p = 4$*. Note that the minimization of the function* S_p *has been performed by means of the Nelder-Mead optimization method. Since it is an iterative method, the needed initial solution can be determined as described in Section 3 of Di Crescenzo et al. (2021a). To improve the goodness-of-fit of the proposed model, the last term of the polynomial* Q_β *can be modified to have a real exponent:*

$$\tilde{l}_m(t) = l_0 \frac{\eta + e^{-\tilde{Q}_\beta(t_0)}}{\eta + e^{-\tilde{Q}_\beta(t)}}, \qquad t \geq t_0,$$

where $\tilde{Q}_\beta(t) = \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 t^{4+r}$ *and* $r \in \mathbb{R}$*. Now the aim is to find the best set of parameters* $\tilde{\theta} = (\theta^T, r)^T$*, i.e. the set which minimizes the cumulative square error defined below*

$$S_{4+r}(\tilde{\theta}) = \sum_{i=1}^{n} (y_i - \tilde{l}_m(t_i))^2,$$

where, as above, y_i *are the data,* t_i *are the time instants for* $i = 1, \ldots, n$*. As shown in Figure 3-(b), the goodness-of-fit increases since* $S_4(\theta) > S_{4+r}(\tilde{\theta})$*. In particular, we have* $S_4(\theta) = S_{4+r}(\tilde{\theta}) + 22,52\%$*.*

3. BIRTH-DEATH PROCESSES

Birth-death (BD) processes are often adopted to describe stochastic dynamics in various fields of biomathematics, being appropriate to model the random evolution of the

number of particles or individuals in a system. Let us consider a time-inhomogeneous BD process $\{N(t); t \geq 0\}$ with state space $\mathbb{N}_0$ and linear birth and death rates given by

$$\begin{aligned} &b_n(t) = n\lambda(t), \qquad n \in \mathbb{N}_0 \\ &d_n(t) = n\mu(t), \qquad n \in \mathbb{N}, \qquad d_0(t) = 0, \end{aligned} \tag{6}$$

where the individual birth and death rates λ and μ are integrable and positive functions in any set $(0, t)$ with $t \geq 0$. In the following proposition, a sufficient and necessary condition to have a conditional mean of multi-sigmoidal logistic type is provided (as done elsewhere, for example in Di Crescenzo and Paraggio (2019) and Di Crescenzo and Spina (2016)).

Proposition 1. *The BD process $N(t)$ with rates given in Eq.* (6) *has conditional mean $\mathbb{E}[X(t)|X(0) = n_0]$ of multi-sigmoidal logistic type if, and only if, the net growth rate $\xi(t) = \lambda(t) - \mu(t)$ is given by $\xi(t) = h_\theta(t)$, for $t \geq 0$.*

See Figure 4 for some plots of the conditional mean considering different choices of the parameters.

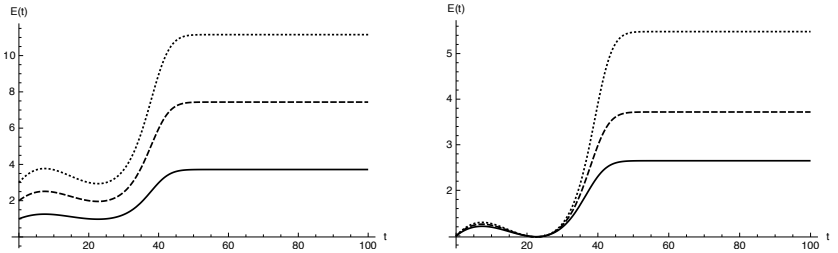


Figure 4: *The conditional mean $\mathbb{E}(t) = \mathbb{E}[X(t)|X(0) = n_0]$ for $p = 3$, $t_0 = 0$, $\beta_1 = 0.1$, $\beta_2 = -0.009$, $\beta_3 = 0.0002$ (a) $\eta = e^{-1}$ and $n_0 = 1, 2, 3$ (from bottom to top) (b) $n_0 = 1$, $\eta = e^{-0.5}, e^{-1}, e^{-1.5}$ (from bottom to top).*

The first-passage-time (FPT) problem is relevant in several applications in population dynamics since the first crossing of a critical high (low) threshold can be viewed as the rising of an overpopulation (extinction). For a fixed threshold $n \in \mathbb{N}$, the FPT of the process $N(t)$ through the state n starting from $N(0) = n_0$ is defined as follows

$$T_{n_0,n} = \inf \{t \geq 0 : N(t) = n\}, \qquad N(0) = n_0.$$

Let us denote by $g_{n_0,n}$ the corresponding probability density function, i.e.

$$g_{n_0,n}(t) = \frac{d}{dt}\mathbb{P}\left(T_{n_0,n} \leq t\right), \qquad t \geq 0.$$

Considering the same matrix-based approach adopted by Tan in Section 3 of Tan (1986), the FPT density vector $g_n := [g_{1,n}, \ldots, g_{n-1,n}]^T$ can be expressed as

$$g_n(t) = \lambda(t) \left(P_1 e^D P_1^{-1}\right)^{-\Lambda(t)} \left(P_2 e^D P_2^{-1}\right)^{-M(t)} P_1 D P_1^{-1} \mathbb{I}_{n-1,1},$$

where $A_i = P_i D P_i^{-1}$ for $i = 1, 2$, $A_1 = \left(a_{i,j}^{(1)}\right)$ and $A_2 = \left(a_{i,j}^{(2)}\right)$ defined in such a way

$$a_{i,j}^{(1)} = \begin{cases} -i, & j = i+1 \\ i, & j = i \\ 0, & \text{otherwise} \end{cases}$$

for $i = 1, \ldots, n-2$ and

$$a_{i,j}^{(2)} = \begin{cases} -i, & j = i-1 \\ i, & j = i \\ 0, & \text{otherwise} \end{cases}$$

for $i = 2, \ldots, n-1$. Moreover, $\Lambda(t) = \int_0^t \lambda(s)ds$, $M(t) = \int_0^t \mu(s)ds$ and $\mathbb{I}_{n-1,1}$ is a column of all 1 of dimension $n-1$. In Figure 5 we provide some plots of the FPT density.

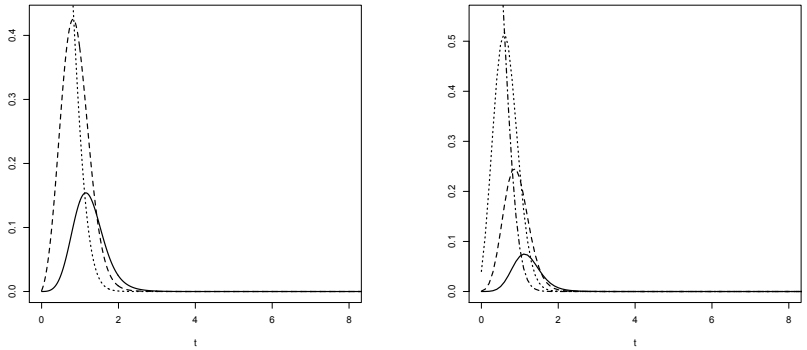


Figure 5: *The FPT density for* $\lambda(t) = 2h_\theta(t)$, $\mu(t) = h_\theta(t)$, $Q_\beta(t) = 0.1t + 0.2t^2 + 0.1t^3$, $n_0 = 1$ *(solid),* 2 *(dashed),* 3 *(dotted),* 4 *(dot-dashed) and (a)* $n = 4$ *and (b)* $n = 5$.

In various applications in biomathematics the systems under investigation are subject to dynamics regulated by transitions where rates are allowed to be nonlinear. Let $\left\{\bar{X}(t); t \geq 0\right\}$ be an inhomogeneous nonlinear BD process having $\mathbb{N}_0$ as state space and birth and death rates given by

$$\begin{aligned} &\bar{b}_n(t) = \lambda_1(t) + \lambda_2(t)n + \lambda_3(t)n^2, \qquad n \in \mathbb{N}_0, \\ &\bar{d}_n(t) = \mu_1(t) + \mu_2(t)n + \mu_3(t)n^2, \qquad n \in \mathbb{N}, \qquad d_0(t) = 0, \end{aligned} \tag{7}$$

where λ_1 and μ_1 are non-negative and integrable functions and λ_i and μ_i for $i = 2, 3$ are positive and integrable functions on any set $(0, t)$. It is easy to note that when $\lambda_1(t) = \mu_1(t) = 0$ and $\lambda_3(t) = \mu_3(t)$, then the mean $m_1(t) = \mathbb{E}[\bar{N}(t)|\bar{N}(0) = n_0]$ satisfies the differential equation

$$\frac{d}{dt} m_1(t) = (\lambda_2(t) - \mu_2(t)) m_1(t),$$

which is a differential equation of the same type of Eq. (1). Hence, assuming that $\lambda_2(t) - \mu_2(t) = h_\theta(t)$, the mean can be expressed as

$$m_1(t) = n_0 \frac{\eta + e^{-Q_\beta(t_0)}}{\eta + e^{-Q_\beta(t)}}, \qquad m_1(0) = n_0.$$

4. THE CORRESPONDING DIFFUSION PROCESS

In order to obtain a more manageable description of the growth phenomenon, we perform a diffusive approximation of the BD process with nonlinear rates given in Eq. (7). The diffusive approximation is based on the scaled BD process $N_\varepsilon(t) = \varepsilon \bar{N}(t)$ whose probability $p_n^\varepsilon(t)$, for $\varepsilon \simeq 0$, gives $p_n^\varepsilon(t) \simeq f(x,t)\varepsilon$ with $x = n\varepsilon$. Under some suitable assumptions, the limits below hold for $\varepsilon \to 0$

$$\begin{aligned} &(\mu_i(t) - \lambda_i(t))\varepsilon \to 0, \qquad i = 1,3, \qquad (\mu_2(t) - \lambda_2(t))\varepsilon \to -r(t), \\ &(\mu_i(t) + \lambda_i(t))\varepsilon^2 \to 0, \qquad i = 1,2, \qquad (\mu_3(t) + \lambda_3(t))\varepsilon^2 \to \sigma^2. \end{aligned}$$

Hence, performing the derivative of f with respect to t and expanding f as a Taylor series around x, the density function f of the approximating process satisfies the following Fokker-Plank equation

$$\frac{\partial}{\partial t} f(x,t) = -\frac{\partial}{\partial x}[r(t) x f(x,t)] + \frac{1}{2} \frac{\partial^2}{\partial x^2}[\sigma^2 x^2 f(x,t)].$$

In other terms, for $r(t) = h_\theta(t)$ with h_θ defined in Eq. (2), the BD process $\bar{N}(t)$ leads to the lognormal diffusion process $X(t)$ having infinitesimal moments

$$A_1(x,t) = h_\theta(t) x, \qquad A_2(x) = \sigma^2 x^2.$$

The initial condition $p_{n_0}(0) = 1$ becomes $\lim_{t \to 0} f(x,t) = \delta(x - x_0)$, where δ is the Dirac delta function. The resulting diffusion process $\{X(t); t \geq t_0\}$ has state space $(0, +\infty)$ and is governed by the following SDE:

$$dX(t) = h_\theta(t) X(t) dt + \sigma X(t) dW(t), \qquad X(t_0) = X_0, \tag{8}$$

where $W(t)$ is a Wiener process independent on the initial condition X_0, for any $t \geq t_0$, $\theta = (\eta, \beta^T)^T$ and $\sigma > 0$. By applying Itô's formula to Eq. (8), we obtain

$$X(t) = X_0 \exp\left[H_\xi(t_0, t) + \sigma\left(W(t) - W(t_0)\right)\right], \qquad t \geq t_0, \tag{9}$$

where $\xi = (\theta^T, \sigma^2)^T$ is the vector containing the parameters of the model and

$$H_\xi(s,t) = \log \frac{\eta + e^{-Q_\beta(s)}}{\eta + e^{-Q_\beta(t)}} - \frac{\sigma^2}{2}(t - s), \qquad t_0 \leq s < t. \tag{10}$$

It is worth to notice that if the initial state X_0 is lognormally distributed with parameters μ_0 and σ_0^2 or is degenerate, then the finite dimensional distributions of $X(t)$ are lognormal. Under the above-mentioned assumptions on X_0, the mean of the process is

$$m_1(t) = \mathbb{E}[X(t)] = \mathbb{E}[X_0]\frac{\eta + e^{-Q_\beta(t_0)}}{\eta + e^{-Q_\beta(t)}}, \qquad t \geq t_0,$$

the mode is

$$Mode[X(t)] = Mode[X_0]\frac{\eta + e^{-Q_\beta(t_0)}}{\eta + e^{-Q_\beta(t)}} \exp\left(-\frac{3}{2}\sigma^2(t-t_0)\right), \qquad t \geq t_0,$$

and, finally, the median is

$$med[X(t)] = med[X_0]\frac{\eta + e^{-Q_\beta(t_0)}}{\eta + e^{-Q_\beta(t)}} \exp\left(-\frac{\sigma^2}{2}(t-t_0)\right), \qquad t \geq t_0.$$

5. MAXIMUM LIKELIHOOD ESTIMATES

In this section we describe two different procedures to find the maximum likelihood estimates (MLEs) of the parameters. We consider a discrete sampling of $X(t)$ based on d independent sample paths, with n_i different observation instants for the i-th sample path, i.e. t_{ij} for $j = 1, \ldots, n_i$, $i = 1, \ldots, d$. For simplicity, we assume that the first time instant is the same for all the sample paths, i.e. $t_{i1} = t_0$, $i = 1, \ldots, d$. The vector $\mathbb{X} = (\mathbb{X}_1^T | \ldots | \mathbb{X}_d^T)^T$, where $\mathbb{X}_i = (X(t_{i1}), \ldots, X(t_{in_i}))^T$ for $i = 1, \ldots, d$ and $X(t_0)$ is lognormally distributed with parameters μ_1 and σ_1^2, has density

$$f_{\mathbb{X}}(x) = \prod_{i=1}^{d} \exp\left(-\frac{(\log x_{i,1} - \mu_1)^2}{x_{i,1}\sigma_1\sqrt{2\pi}}\right) \prod_{j=1}^{n_i - 1} \frac{\exp\left(-\frac{\left[\log\left(\frac{x_{i,j+1}}{x_{i,j}}\right) - m_\xi^{i,j+1,j}\right]^2}{2\sigma^2\Delta_i^{j+1,j}}\right)}{x_{i,j}\sigma\sqrt{2\pi\Delta_i^{j+1,j}}}$$

where $\Delta_i^{m,n} = t_{i,m} - t_{i,n}$, $m, n = 1, \ldots, n_i - 1$, $m > n$, $\xi = (\theta^T, \sigma^2)^T$ and $m_\xi^{i,m,n} = H_\xi(t_{i,n}, t_{i,m})$ with H_ξ defined in Eq. (10).

If (μ_1, σ_1^2) and ξ are functionally independent, the MLEs of (μ_1, σ_1^2) leads to $\hat{\mu}_1 = \frac{1}{d}\log x_{i,1}$, and $\hat{\sigma}_1^2 = \frac{1}{d}\sum_{i=1}^{d}(\log x_{i,1} - \hat{\mu}_1)^2$. The estimation of ξ is obtained from the following system

$$\begin{cases} \sigma^2\left(n + \frac{\sigma^2}{4}Z_3\right) - Z_1 - A_\theta + 2B_\theta = 0 \\ Y_l^\theta + \frac{\sigma^2}{2}W_l^\theta + X_l^\theta = 0, \quad l = 0, 1, \ldots, p, \end{cases} \tag{11}$$

where for $v_{0i} = x_{i,1}$ and $v_{i,j} = \left(\Delta_i^{j+1,j}\right)^{-1/2} \log\left(\frac{x_{i,j+1}}{x_{i,j}}\right)$, $j = 1, \ldots, n_i - 1$, $i = 1, \ldots, d$, we have set

$$W_l^\theta = \sum_{i=1}^{d} {}_l D_\theta^{i,n_i,1}, \quad Y_l^\theta = \sum_{i=1}^{d} \sum_{j=1}^{n_i-1} \frac{1}{\Delta_i^{j+1,j}} \log\left[\frac{\eta + e^{-Q_\beta(t_{i,j+1})}}{\eta + e^{-Q_\beta(t_{i,j})}}\right] {}_l D_\theta^{i,j+1,j}$$

$$X_l^\theta = \sum_{i=1}^{d} \sum_{j=1}^{n_i-1} \frac{v_{i,j}}{\left(\Delta_i^{j+1,j}\right)^{1/2}} {}_l D_\theta^{i,j+1,j}, \qquad l = 0, 1, \ldots, p,$$

$$\lambda_\theta^{i,m,n} = \log \frac{\eta + e^{Q_\beta(t_{i,n})}}{\eta + e^{Q_\beta(t_{i,m})}}, \quad m > n,\ i = 1, \ldots, d, \quad Z_1 = \sum_{i=1}^{d} \sum_{j=1}^{n_i-1} v_{ij}^2$$

$$Z_3 = \sum_{i=1}^{d} \Delta_i^{n_i,1}, \quad A_\theta = \sum_{i=1}^{d} \sum_{j=1}^{n_i-1} \frac{\left(\lambda_\theta^{i,j+1,j}\right)^2}{\Delta_i^{j+1,j}}, \quad B_\theta = \sum_{i=1}^{d} \sum_{j=1}^{n_i-1} \frac{v_{i,j} \lambda_\theta^{i,j+1,j}}{\left(\Delta_i^{j+1,j}\right)^{1/2}}.$$

From now on, we suppose, without loss of generality, that $t_0 = 0$ and that $n_i = N$ for $i = 1, \ldots, d$. The system (11) cannot be solved explicitly and it is therefore necessary to use a numerical method, such as Newton-Raphson. Hence, an initial approximation is required. An initial solution of σ^2 is calculated by performing a simple linear regression of $\sigma_i^2 = 2\log(m_i/m_i^g)$ where m_i denotes the sample mean and m_i^g the geometric sample mean. Whereas, an initial solution for the coefficients β and η is obtained by a linear regression taking as data the pairs $\left(t_i, -\log\left(\frac{m_N}{m_i} - 1\right)\right)$ where m_N is the last value of the sample mean.

Alternatively, one can obtain the estimates of ξ by maximizing the likelihood function

$$\tilde{L}(\xi) = -\frac{n}{2}\log\sigma^2 - \frac{Z_1 + \Phi_\xi - 2\Gamma_\xi}{2\sigma^2},$$

where

$$n = \sum_{i=1}^{d}(n_i - 1), \quad Z_1 = \sum_{i=1}^{d} \sum_{j=1}^{n_i-1} \frac{1}{\Delta_i^{j+1,j}} \log^2 \frac{x_{i,j+1}}{x_{i,j}}$$

$$\phi_\xi = \sum_{i=1}^{d} \sum_{j=1}^{n_i-1} \frac{(m_\xi^{i,j+1,j})^2}{\Delta_i^{j+1,j}}, \quad \Gamma_\xi = \sum_{i=1}^{d} \sum_{j=1}^{n_i-1} \frac{1}{(\Delta_i^{j+1,j})^{1/2}} \log \frac{x_{i,j+1}}{x_{i,j}} m_\xi^{i,j+1,j}.$$

To maximize the function $\tilde{L}$, we use a meta-heuristic optimization method, namely Simulated Annealing (S.A.). This algorithm (see as a reference Kirkpatrick *et al.* (1983)) is used for problems like finding $\arg\min_{\theta\in\Theta} f(\theta)$ and in recent years also in the context of parameters estimation (cf. da Luz Sant'Ana *et al.* (2018) and Román-Román and Torres-Ruiz (2015)). At any step, S.A. generates a new solution in a neighborohood

of the previous one and (i) if the new solution improves the objective function, then it replaces the previous, otherwise (ii) if the new solution does not improve the objective function, then it can replace the previous with a probability rate which depends on the increase of the objective function and on a scale factor, called temperature, in agreement with the metallurgical annealing that inspires the method. S.A. avoids in this way local minima but it needs a restriction of the parametric space Θ. In our context of interest, the set Θ contains the parameters ξ. Until now, it is continuous and unbounded, since $\Theta = \{(\eta, \beta^T, \sigma^2) : \eta > 0,\, \beta_1, \ldots, \beta_{p-1} \in \mathbb{R},\, \beta_p > 0,\, \sigma^2 > 0\}$. To bound Θ, we consider $0 < \sigma < 0.1$ so that the simulated sample paths are less variable around the sample mean and thus the multi-sigmoidal logistic profile is advisable. For the parameters $\beta = (\beta_1, \ldots, \beta_p)^T$, we consider the confidence intervals, found by using the data of the polynomial regression performed previously for the initial solutions. Specifically, for β we consider the confidence intervals of the coefficients of the polynomial regression of $-\log\left[\left(\frac{m_N}{m_j} - 1\right)\hat{\eta}\right]$ against t_j, for $j = 1, \ldots, N$, where $\hat{\eta} = \left(\frac{m_N}{m_1} - 1\right)^{-1}$. Finally, for η, we consider the interval (a, b) where

$$a = \min_{1 \le i \le d} \left(\frac{x_{i,n_i}}{x_{i,1}} - 1\right)^{-1}, \qquad b = \max_{1 \le i \le d} \left(\frac{x_{i,n_i}}{x_{i,1}} - 1\right)^{-1}.$$

Regarding the distributions of the MLEs, it is worth to notice that the exact distribution of $\hat{\mu}_1$ is Gaussian $\mathcal{N}(\mu_1, \sigma_1^2/d)$ and the one of $d\hat{\sigma}_1^2/\sigma_1^2$ is chi-square χ^2_{d-1}. Furthermore, the asymptotic distribution of $\hat{\xi}$ is a $(p+2)$-dimensional normal distribution with mean ξ and covariance matrix $I(\xi)^{-1}$, where $I(\xi) \in \mathbb{R}^{(p+2)\times(p+2)}$ is the Fisher information matrix and can be expressed as

$$I(\xi) = \frac{1}{\sigma^2} \begin{pmatrix} \Xi_\xi & -\frac{1}{2}\left(\frac{\partial}{\partial\theta}\gamma_\xi\right) \\ -\frac{1}{2}\left(\frac{\partial}{\partial\theta}\gamma_\xi\right)^T & \frac{n}{2\sigma^2} - \frac{Z_3}{4} \end{pmatrix},$$

where $\Xi_\xi \in \mathbb{R}^{(p+1)\times(p+1)}$ and $\dfrac{\partial}{\partial\theta}\gamma_\xi \in \mathbb{R}^{(p+1)\times 1}$ are defined as

$$\Xi_\xi = \sum_{i=1}^{d} \sum_{j=1}^{n_i - 1} \left(\Delta_i^{j+1,j}\right)^{-1} \left(\frac{\partial}{\partial\theta} m_\xi^{i,j+1,j}\right) \left(\frac{\partial}{\partial\theta} m_\xi^{i,j+1,j}\right)^T$$

and

$$\frac{\partial}{\partial\theta}\gamma_\xi = \sum_{i=1}^{d} \sum_{j=1}^{n_i - 1} \frac{\partial}{\partial\theta} m_\xi^{i,j+1,j}.$$

6. SIMULATIONS

A simulation study is developed to verify the validity of the two aforementioned procedures. As a case study, we use a pattern of 100 independent sample paths simulated by using the expression of the process obtained as the solution of the stochastic

differential equation (9). All the sample paths contain the same sumber of data (that is 501), being $(i-1)\cdot 0.1$ for $i = 1, \ldots, 501$ the observation times. The parameters used for the simulation are $\eta = e^{-1}$, $\beta_1 = 0.1$, $\beta_2 = -0.009$, $\beta_3 = 0.0002$, $\sigma = 0.01$. See Figure 6 for the plot of the paths. For simplicity, we have chosen a degenerate initial

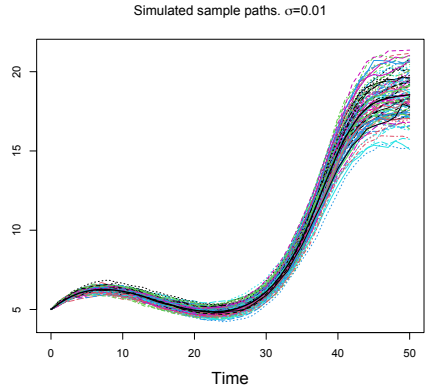


Figure 6: *100 simulated sample paths of the process* $X(t)$ *for* $p = 3$, $t_0 = 0$, $x_0 = 5$, $Q_\beta(t) = 0.1t - 0.009t^2 + 0.0002t^3$, $\eta = e^{-1}$ *and* $\sigma = 0.01$.

distribution centered in $x_0 = 5$, i.e. $\mathbb{P}(X_0 = 5) = 1$. After obtaining each trajectory, we chose 51 values from the first one and using a step equal to 1. The MLEs obtained by solving the system (11) are summarized in Table 1.

Table 1: *The MLEs obtained by solving the system* (11).

	Real	Initial	Estimations	Rel. Err.
$\boldsymbol{\eta}$	$e^{-1} = 0.3678794$	0.3695601	0.3695532	$1.673743e$-03
$\boldsymbol{\beta_1}$	0.1	−0.04606928	0.10021729	$2.172891e$-04
$\boldsymbol{\beta_2}$	−0.009	−0.003311808	−0.009030270	$2.030270e$-03
$\boldsymbol{\beta_3}$	0.0002	0.0001356439	0.0002006625	$6.625399e$-07
$\boldsymbol{\sigma}$	0.01	$9.912749e$-03	$9.982475e$-03	$3.502022e$-07

As a further case study, we use the same pattern as before by applying S.A. Moreover, since S.A. is a meta-heuristic algorithm, we apply the procedure 10 times and then we consider the mean of the resulting values. Clearly, if the number of replications increases then the goodness of the results improves but also the computational cost. The MLEs obtained in this way are summarized in Table 2.

Table 2: *The MLEs obtained via S.A.*

	Real	Range	Estimations	Abs. Err.
$\boldsymbol{\eta}$	$e^{-1} = 0.3678794$	[0.287803710, 0.405281061]	0.3862679	0.04998513
$\boldsymbol{\beta_1}$	0.1	[0.094192479, 0.155306328]	0.1099219	0.09921900
$\boldsymbol{\beta_2}$	−0.009	[−0.012033562, −0.009118551]	−0.009722108	0.08023422
$\boldsymbol{\beta_3}$	0.0002	[0.000206216, 0.000245301]	0.0002131091	0.06554550
$\boldsymbol{\sigma}$	0.01	[0.000000000, 0.010000000]	0.0001283712	0.13301015

7. CONCLUSIONS

In this paper, we considered the deterministic multi-sigmoidal logistic function and we used the presented model to describe the double-sigmoidal growth of coffee berries. In order to make the model more realistic, we analysed its stochastic counterpart. More in detail, we studied two different birth-death processes, the former with linear rates and the latter with quadratic rates. From the last one, we derive a diffusive approximation by means of a suitable scaling. Then, we found the MLEs of the parameters of the diffusion process by using two different strategies: by solving a non-linear system and by maximizing the log-likelihood function via S.A. We also studied the asymptotic distribution of the resulting MLEs. Finally, to validate the described procedures, we performed a simulation study. Future investigations will be devoted to determine the degree p of the polynomial Q_β, since it is unknown a priori and to use different meta-heuristic strategies to find nice MLEs in a shorter computational time. It will be interesting also to consider a real application based on the diffusion process.

ΠΕΡΙΛΗΨΗ

Θεωρούμε μια πολύ-σιγμοειδή γενίκευση του μοντέλου λογιστικής αύξησης. Το ντετερμινιστικό μοντέλο παρουσιάζεται μαζί με το αντίστοιχό του στοχαστικό μοντέλο. Ειδικότερα, αναλύουμε δύο διαφορετικές διαδικασίες γέννησης-θανάτου με γραμμικούς και τετραγωνικούς ρυθμούς μεταβάσεων, αντίστοιχα. Από το τελευταίο δίνουμε μια πιο εύχρηστη προσέγγιση διάχυσης μέσω μιας κατάλληλης κλιμάκωσης. Επιπλέον, μελετάμε δύο πιθανές στρατηγικές για να υπολογίσουμε τους εκτιμητές μέγιστης πιθανοφάνειας των παραμέτρων του μοντέλου. Για την επαλήθευση των περιγραφόμενων διαδικασιών, παρουσιάζουμε μία μελέτη προσομοίωσης. Διερευνάται επίσης το πρόβλημα του πρώτου χρόνου διέλευσης.

Acknowledgements: Antonio Di Crescenzo and Paola Paraggio are members of the research group GNCS of INdAM (Istituto Nazionale di Alta Matematica). This work was supported in part by the Ministerio de Economía Industria y Competitividad, Spain, under Grant MTM2017-85568-P, FEDER/Junta de Andalucía-Consejería de Economía y Conocimiento, under Grant A-FQM-456-UGR18 and by Italian MIUR-PRIN 2017, project "Stochastic Models for Complex Systems", No. 2017HJFFHSH. Paola Paraggio thanks the Department of Statistics and Operations Research, Faculty of Sciences of the University of Granada and the Institute of Mathematics of the University of Granada (IMAG) for the hospitality during the one-month visit carried out in 2019.

REFERENCES

da Cunha, A. R.; Volpe, C.A. (2011) Growth curves of coffee fruits Obatã IAC 1669-20 in different alignments planting. *Semina: Ciências Agrárias, Londrina*, **32**, 1, 49–62.

Di Crescenzo, A.; Paraggio, P. (2019) Logistic growth described by birth-death and diffusion processes. *Mathematics* **7 (6)**, 1–28, 489.

Di Crescenzo, A.; Paraggio, P.; Román-Román, P.; Torres-Ruiz, F. (2021a) Applications of the multi-sigmoidal deterministic and stochastic logistic models for plant dynamics. *Appl Math Model*, **92**, 884–904.

Di Crescenzo, A.; Paraggio, P.; Román-Román, P.; Torres-Ruiz, F. (2021b) Statistical analysis and first-passage-time applications of a lognormal diffusion process with multi-sigmoidal logistic mean (submitted for publication).

Di Crescenzo, A.; Spina, S. (2016) Analysis of a growth model inspired by gompertz and korf laws, and an analogous birth-death process. *Math Biosci*, **282**, 121–134.

Erto, P.; Giorgio, M.; Lepore, A. (2020) The Generalized Inflection S-Shaped Software Reliability Growth Model. *IEEE Trans Reliab*, **69**, 1, 228–244.

Kirkpatrick, S.; Gelatt, C.D.; Vecchi, M.P. (1983) Optimization by simulated annealing. *Science, New Series*, **220**, 4598, 671–680.

da Luz Sant'Ana, I.; Román-Román, P.; Torres-Ruiz, F. (2018) The Hubbert diffusion process: Estimation via simulated annealing and variable neighborhood search procedures - application to forecasting peak oil production. *Appl Stochastic Models Bus Ind* **34**, 376–394.

Rajasekar, S.P.; Pitchaimani, M.; Quanxin Zhu (2020) Progressive dynamics of a stochastic epidemic model with logistic growth and saturated treatment. *Physica A* **538**, 122649, 1–20.

Román-Román, P.; Torres-Ruiz, F. (2015) The nonhomogeneous lognormal diffusion process as a general process to model particular types of growth patterns. *Lecture Notes of Seminario Interdisciplinare di Matematica* **12**, 201–219.

Román-Román, P.; Serrano-Pérez, J.J.; Torres-Ruiz, F. (2019) A note on estimation of multi-sigmoidal Gompertz functions with random noise. *Mathematics* **7**, 541, 1–18.

Tan, W.Y. (1986) A stochastic Gompertz birth-death process. *Stat Prob Lett* **4**, 25–28.

Greek Statistical Institute
Proceedings of the 33rd Panhellenic Statistics Conference (2021), pp. 39−44

UNFOLDING MODELS FOR ORDINAL DATA IN CYBER RISK ASSESSMENT

***S. Facchinetti*[1], *M. Iannario*[2], *S.A. Osmetti*[1], *C. Tarantola*[3]**
[1]Università Cattolica del Sacro Cuore, Milano, Italy
{silvia.facchinetti, silvia.osmetti}@unicatt.it
[2]University of Naples Federico II, Napoli, Italy
maria.iannario@unina.it
[3]University of Pavia, Pavia Italy
claudia.tarantola@unipv.it

ABSTRACT

In an increasingly digitalized world, where organizations are affected by technological evolution, cyber attacks are multiplying rapidly. They have an impact on every class of business and no industry can consider itself immune to them. Quantitative loss data are rarely available while it is possible to obtain a qualitative evaluation, expressed on a rating scale, from experts of the sector. Hence, we focus on ordinal data models for cyber risk evaluation (rating) with particular emphasis on a mixture model taking into account the uncertainty in the process of scoring. We examine a set of data regarding cyber attacks that occurred worldwide before and during the pandemic due to Covid-19. The aim of our analysis is to investigate if Covid-19 has affected experts' uncertainty and assessment, and identify the relevant factors which influence the severity of an attack.

Keywords: Cyber risk, CUP models, Rating, Uncertainty.

1. INTRODUCTION

Throughout the last years the use of statistical modelling in the analysis of cyber risk assessment has gained a rapidly increasing interest. While quantitative loss data are rarely available, a qualitative evaluation of the level of severity of an attack, expressed on a rating scale, from experts of the sector is able to be obtained. In this way, we can identify which types of attacks are the most dangerous.

The rating assigned by the expert can be considered as the final outcome of a complex activity based on knowledge of the topic, collection of information but also instinct and feeling of the expert himself.

In this contribution we rely on ordinal mixture models to mimic the skilled decision-making process. In particular, we exploit the CUP mixture introduced by Tutz *et al.* (2017). It is based on a *C*ombination of an incertitude in the process of assessment

(*U*ncertainty component) and a deliberate choice based on the evaluation of the respondent (*P*erception component). The latter component accounts for reasoned judgments toward the attack under evaluation as well as the set of experts' perceptions and information connected with it. The uncertainty component accounts for other unreasonable elements such as the difficulty in expressing a rating regarding a specific event about which the expert has not a clear opinion or has a limited set of information. Furthermore, it is also related to the amount of time devoted to the judgment or experts' laziness, boredom or circumstances. The mixture can be considered as a combination of the distributions of a discretised version of the underlying continuous latent variables describing these different components.

We implement the CUP mixture in rating systems for the analysis of risk assessment of worldwide cyber attacks occurred in the period 2018-2020 years (before and during the Covid-19 pandemic) generalizing the main findings in Facchinetti *et al.* (2020, 2021) referred to 2017-2019 data. The proposal discussed here extends the previous contributions from two directions: the analysis of uncertainty in the experts' assessment and the evaluation of epidemic period by considering the new tools affecting the process of scoring.

The plan of the paper is as follows. Section 2. is devoted to the description of the considered model. In Section 3. we present the data and discuss the obtained results. The paper ends with some concluding remarks and avenues for future research.

2. METHODOLOGY

In this section we briefly review standard CUP models. The perception component is described via a cumulative link model under the proportional odds assumption (McCullagh, 1980) while a discrete Uniform distribution is used for the uncertainty component.

More precisely, the observed rating s (severity level) assigned to a specific cyber attack can be considered as a realisation of a random variable S with probability distribution

$$P(S_i = s|\boldsymbol{x}_i) = \pi P_M(Y_i = s|\boldsymbol{x}_i) + (1-\pi)P(U_i = s), \quad s = 1, 2, \ldots, m. \qquad (1)$$

As earlier pointed out, the preference component $P_M(Y_i = s|\boldsymbol{x}_i)$ is defined via a cumulative link model on an appropriate row vector of covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ij}, \ldots x_{ip})$. Formally, we have

$$link\left[P_M(Y_i \leq s|\boldsymbol{x})\right] = \alpha_s - \boldsymbol{x}_i\boldsymbol{\gamma} \quad i = 1, 2 \ldots, n; \quad s = 1, 2 \ldots, m-1,$$

where $\boldsymbol{\gamma}$ is the parameter vector for the preference component, whereas $-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_m = \infty$ represent the thresholds of the scale of the latent variable Y^* behind Y. Among the alternative choices for $link$ functions we focus on the logit one for easiness of interpretation and robustness properties. The uncertainty component is modelled as $P(U_i = s) = 1/m$. The two components are then combined via the parameter π eventually depending on a vector of covariates $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{ij}, \ldots w_{iq})$,

with a possible non empty intersection with $\boldsymbol{x}_i$. To model the effect of the covariates on the uncertainty component we use a logit link as well, $\pi = \pi(\boldsymbol{\beta}) = 1/(1 + e^{-\boldsymbol{w}_i\boldsymbol{\beta}})$, where $\boldsymbol{\beta}$ is the parameter vector for the related component. The standard cumulative link model is a special case of (1) with $\pi = 1$.

From an inferential point of view, a way to obtain stable estimates is to consider the mixture as a problem with incomplete data and use the EM algorithm (Dempster *et al.*, 1977); see the Appendix of Tutz *et al.* (2017) for further details.

3. APPLICATION TO CYBER RISK DATA

We consider a set of data collected by the experts of the "Hackmanac" society (`https://hackmanac.com/`). Hackmanac is a company based in Dubai that monitors the evolution of real global cyber threats with the aim to support companies and institutions to define their cyber defense strategy.

In particular, we investigate a sample of more than 5.000 statistical units regarding cyber attacks occurred worldwide during the years 2018, 2019 and 2020. For each attack we have information regarding the following "macro variables": `Type of attack` (main actors and motivations of the attack), `Attack Technique` (adversary tactics and techniques of attack), `Target Class` (victims of cyber attack), `Continent` (where the attacks took place), and `Severity` (an ordinal classification of the gravity of an attack). Indeed, experts classify the gravity of an attack on the basis of their knowledge by the ordinal variable `Severity` assuming values 1 (low severity), 2 (medium severity), 3 (high severity) and 4 (critical severity).

The evaluation of each attack on the basis of its seriousness is the outcome of a complex activity based on various aspects such as awareness and experience about the geopolitical, social, economic, and image impact on the victims, but also sensation and feeling of the expert himself. Due to the characteristics of this decision-making process that combines knowledge of the examined event and expert awareness, in this paper we rely on CUP models.

Based on a preliminary model selection analysis, we decided to include in our model the following binary covariates for each "macro variable":

- Cybercrime as the candidate `Type of attack`
- Information and communication technologies (ICT) and Government-Military-Law-Enforcement (GOV) as `Target Class`
- Target group (TG) as an indicator of those statistical units compromised by lower than 3 attacks
- Vulnerabilities as `Attack Technique`
- `Continent` as a factor variable with Africa as reference level (AF=Africa, AM=America, AS=Asia, EU=Europa, OC=Oceania, MC=Multiple continent)
- `Covid` a dummy variable representative of the pandemic period.

In Table 1 the estimated values of the parameters and the asymptotic standard errors (in brackets) for the examined CUP model are reported. The latter were computed via

(numerical) Hessian. The symbol "$\star$" indicates that the corresponding parameter is not significant at 5% level. Furthermore, the Bayesian Information Criterion (BIC) index (Schwarz, 1978) of the selected model is 12619.46 whereas the BIC index of the nested standard cumulative model is 12874.14 highlighting the added value of the proposal.

Table 1: *Estimated* CUP *model for cyber risk analysis.*

Uncertainty component	
β_0	0.973 (0.142)
Covid	2.676 (0.291)
Perception component	
α_1	-2.330 (0.496)
α_2	-0.409 (0.493)
α_3	2.211 (0.478)
`Type of attack`	
Cybercrime	-1.735 (0.134)
`Target Class`	
ICT	0.866 (0.106)
GOV	0.979 (0.116)
`Attack Technique`	
Vulnerabilities	0.898 (0.119)
`Continent`	
AM	0.972 (0.458)
AS	1.809 (0.456)
EU	1.008 (0.465)
MC	0.221 (0.470) $\star$
OC	0.942 (0.508)
`Target Group`	
TG	0.411 (0.082)
`Period`	
Covid	-0.335(0.070)

Before commenting out the obtained results, we recall that the weight of the uncertainty component in the CUP mixture is equal to $(1 - \pi)$.

`Covid` is the only covariate affecting both components of the CUP mixture; it influences negatively both the uncertainty and the perception. Thus, the fitted model indicates a low level of uncertainty in the severity evaluation during the pandemic. This result is probably due to a more accurate evaluation expressed on cyber attacks during the Covid period than before. With regards to the perception component, we observe a lower probability to obtain an elevate level of severity during the epidemic than the previous period.

Furthermore, we observe that also Cybercrime has a negative influence on the experts' risk perception. This result is consistent with the analysis of Facchinetti et al. (2020, 2021) that pointed out that even if this type of offensive is quite frequent, in terms of gravity it determines attacks of minor severity. With reference to `Target Class`, `ICT` and `GOV` are associated with a higher probability of a critical severity attack with respect to the others.

The parameter associated to the examined category of `Attack technique` is positive. This indicates that the exploitation of system vulnerabilities (weakness that can be used to gain unauthorized access to a computer system) can led attacks scored with a high level of severity.

On the subject of `Continent`, the parameter related to category MC is not significant. This could be explained by the fact that attacks directed against single continents are more effective than the ones involving more of them. Moreover, we observed a lower severity level for Africa (the baseline level) with respect to the other ones.

Finally, an attack directed to a victim belonging to a target group compromised by more than 3 attacks determines a substantial higher level of severity.

4. CONCLUSION

In this paper we illustrated how CUP models can be an useful instrument for cyber risk evaluation.

The CUP mixture allows to improve results with respect to the classical assumption of the standard models used for the analysis of ordinal (rating) data by means of the added value of the uncertainty component which also represents an advantage over classical mixture models. In comparison with the latter the components are fully specified and are not from the same class of models; see among others Greene and Hensher (2003), Grün and Leisch (2008), Breen (2010). More specifically, when the uncertainty component is neglected, the strength of covariates tends to be underestimated. In addition, when uncertainty is very high, the study of the perception component without the assessment of the uncertainty one causes a loss of information and a misspecification of the model.

Further analyses will be devoted to the probability-based measures for comparing clusters (`Target group`) on ratings, while adjusting for other explanatory variables as also reported in Iannario and Tarantola (2021a, 2021b), and to a hierarchical modelling structure taking into account the homogeneity of the clusters related to several countries.

Acknowledgements: We acknowledge support from the European Cost Action "CA19130 - Fin-tech and Artificial Intelligence in Finance - Towards a transparent financial industry".

ΠΕΡΙΛΗΨΗ

Η παρούσα εργασία εστιάζει σε μοντέλα διατακτικών δεδομένων πριν και κατά τη διάκρεια της πανδημίας του Covid-19, για την αξιολόγηση κινδύνων στον κυβερνοχώρο με έμφαση σε μοντέλα μίξης λαμβάνοντας υπόψην την αβεβαιότητα στη διαδικασία αξιολόγησης.

REFERENCES

Breen, R., Luijkx, R. (2010). Mixture models for ordinal data. *Sociological Methods and Research*, **39**, 3–24.

Dempster, A.P., Laird N.M. and Rubin, D.B.(1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Greene, W., Hensher, D. (2003). A latent class model for discrete choice analysis: Contrasts with mixed logit. *Transportation Resesearch*, Part B, **39**, 681–689.

Grün, B., Leisch, F. (2008). Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *Journal of Classification*, **25**, 225–247.

Facchinetti, S., Osmetti, S.A., Tarantola, C. (2020). How to perform cyber risk assessment via cumulative logit models, *Book of short papers - SIS 2020*, 1083-1086.

Facchinetti, S., Osmetti, S.A., Tarantola, C. (2021). A statistical approach for assessing cyber risk via ordered response models, under review for international journal.

Iannario, M. and Tarantola, C. (2021a). Effect Measures for Group Comparisons in a Two-Component Mixture Model: A Cyber Risk Analysis. In Balzano, S., Porzio, G.C., Salvatore, R., Vistocco, D., Vichi, M. (Eds.) *Statistical Learning and Modeling in Data Analysis*. Springer Nature Switzerland AG. Springer Nature Switzerland AG.

Iannario, M., Tarantola, C. (2021b). How to Interpret the Effect of Covariates on the Extreme Categories in Ordinal Data Models, *Sociological Methods & Research*, https://doi.org/10.
1177/0049124120986179.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Ser.* B, 42, 109-142.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Tutz, G., Schneider, M., Iannario, M., Piccolo, D. (2017). Mixture models for ordinal responses to account for uncertainty of choice, *Advances in Data Analysis and Classification*, **11**, 281-305.

Greek Statistical Institute
Proceedings of the 33rd Panhellenic Statistics Conference (2021), pp.45-48

AN IMPROPER INTEGRAL REPRESENTATION OF LINNIK'S PROBABILITY DENSITIES

Azize Hayfavi
Middle East Technical University, Institute of Applied Mathematics
azizeh@metu.edu.tr

ABSTRACT

In 1953 (1963 in English) Linnik introduced the probability density $p_\alpha(x)$ defined in terms of its characteristic function

$$\varphi_\alpha(t) = \frac{1}{1+|t|^\alpha}, \qquad 0 < \alpha < 2.$$

In Kotz et al. (1995a) and Kotz et al. (1995b) the expansions of probability density functions, say $p_\alpha(x)$, into convergent series in terms of $\log(|x|), |x|^{k\alpha}, (k = 0,1,2, \dots)$ are obtained and the asymptotic behavior of $p_\alpha(x)$ at 0 and ∞ is investigated.

In Hayfavi A. (1998), an improper integral representation of Linnik's probability density is presented. Also, an investigation into the exceptional set is achieved as well. This series of works provide a general overview of the research on these probability densities and provides new ideas for some important processes like Linnik Lévy processes and so on.

Keywords: Contour integration, Linnik's probability density, Liouville numbers

1. INTRODUCTION

Linnik first proved that

$$\varphi_\alpha(t) = \frac{1}{1+|t|^\alpha}, \qquad 0 < \alpha < 2$$

is a characteristic function of a probability density, say $p_\alpha(x)$. Using the inversion formula, we can write

$$p_\alpha(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-itx}}{1+|t|^\alpha}\, dt, \qquad 0 < \alpha < 2.$$

In the previous papers asymptotic behavior of the density function $p_\alpha(x)$ at 0 and ∞ was investigated and the expansions of $p_\alpha(x)$ into convergent series were obtained for almost all α's. Furthermore, it was proved that the exceptional set is a subset of

Liouville numbers and one counter example was constructed to show that the exceptional set is not empty.

In this short paper which is the outline of the presentation given in GSI (2021), I will state some of the results obtained and the importance of the distribution itself.

If we observe carefully different data like physical, financial, etc., most of them have long-tailed and fat-tailed distributions. This is why the Linnik distribution represents better such datas. Also, it can be considered as a generalization of Laplace distribution, i.e., Laplace distribution is a Linnik distribution with parameter $\alpha = 2$.

2. REPRESENTATION OF LINNIK PROBABILITY DENSITY BY AN IMPROPER INTEGRAL

To obtain the improper integral representation of Linnik's probability densities, we use the representation of Linnik probability density by a contour integral as follows:

In Kotz et al. (1995b), Theorem 13.1 p.513 it is proved that for any $\alpha \in (0,2)$ the representation

$$p_\alpha(x) = \frac{1}{x} I_\delta(x;\alpha), \qquad x > 0$$

is valid. Here δ is such that $\delta < 1/2$ and $\alpha \in [\delta, 2-\delta]$.

The integral

$$I_\delta(x;\alpha) = \frac{i}{4\alpha} \int_{L(\delta)} \frac{e^{z \log x}\, dz}{\Gamma(z)\, \sin\left(\frac{\pi}{\alpha} z\right) \cos\left(\frac{\pi}{2} z\right)}, \qquad x > 0.$$

where $L(\delta)$ is the boundary of the region:

$$G(\delta) = \left\{ z: |z| > \frac{\delta}{2}, |arg\ z| < \frac{\pi}{4} \right\}$$

described in the positive direction, and $\Gamma(z)$ that appears in the integrand is the Gamma function (see: Wittaker et al. (1962)) and proved that the integral $I_\delta(x;\alpha)$ is absolutely convergent.

Using the Cauchy residue theorem, we obtained:

$$I_\delta(x;\alpha) = \frac{1}{2} \sum_{k=1}^{\infty} \frac{(-1)^{k+1} x^{k\alpha}}{\Gamma(k\alpha) \cos\left(\frac{\pi}{2} k\alpha\right)} + \frac{1}{\alpha} \sum_{k=0}^{\infty} \frac{(-1)^{k+1} x^{k+1}}{\Gamma(2k+1) \sin\left(\frac{\pi}{\alpha}(2k+1)\right)}.$$

Next, we change the contour of the above integral $I_\delta(x;\alpha)$.

We consider the region

$$D(\delta) = \left\{z: |z| > \frac{\delta}{2}, |arg\, z| < \frac{\pi}{2}\right\}.$$

We call $\Lambda(\delta)$ the boundary of the region $D(\delta)$, where the transition on the boundary is the usual positive direction. We call

$$J_\delta(x;\alpha) = \frac{i}{4\alpha}\int_{\Lambda(\delta)} \frac{e^{zlogx}\, dz}{\Gamma(z)\, sin\left(\frac{\pi}{\alpha}z\right) cos\left(\frac{\pi}{2}z\right)}, \qquad x > 0.$$

Using once more the Cauchy residue theorem we obtain

$$J_\delta(x;\alpha) = \frac{1}{2}\sum_{k=1}^{\infty} \frac{(-1)^{k+1}x^{k\alpha}}{\Gamma(k\alpha)\, cos\left(\frac{\pi}{2}k\alpha\right)} + \frac{1}{\alpha}\sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{\Gamma(2k+1)\, sin\left(\frac{\pi}{\alpha}(2k+1)\right)}. \qquad (2.1)$$

We conclude stating some of the results obtained:

Theorem 2.1 [Hayfavi, A. (1998), p.236] For any $\alpha \in (0,2)$ we have the representation

$xp_\alpha(x) = J_\delta(x;\alpha), \qquad x > 0.$

Theorem 2.3 [Hayfavi, A. (1998), p.239] For any $\alpha \in (0,2)$ we have the formula:

$$xp_\alpha(x) = \frac{-1}{4\alpha}\int_{-\infty}^{\infty} \frac{e^{irlogx}\, dr}{\Gamma(ir)\, sinh\left(\frac{\pi}{\alpha}r\right) cosh\left(\frac{\pi}{2}r\right)}, \qquad x > 0.$$

Remark: To prove the Theorem 9.5 in Kotz et al. (1995b) we had constructed a transcendental number α to show that the exceptional set that both series in Eq. (2.1) diverge is not empty.

Now we give another result about this exceptional set:

Theorem 2.5 [Hayfavi, A. (1998), p.241] The transcendental numbers $\beta \in (0,2)$ that the series on the right-hand side of Eq. (2.1) are divergent, are dense in the interval $(0,1)$.

3. CONCLUSION

As we mentioned at the end of the introduction, Linnik distribution and Linnik Lévy processes are used to the description of heavy-tailed data. Linnik Lévy processes are processes of independent stationary increments, having Linnik distribution. To see the performance of these distributions the best is to apply them to some fat-tailed or long-tailed distributions.

ΠΕΡΙΛΗΨΗ

Το 1953 ο Linnik όρισε την πυκνότητα πιθανότητας συναρτήσει της χαρακτηριστικής συνάρτησης $\varphi_\alpha(t)$, $0 <\alpha<2$.

Ο Kotz και συνεργάτες (1995a; 1995b) έδωσαν το ανάπτυγμα της πυκνότητας πιθανότητας σε συγκλίνουσα σειρά όρων όπως $\log(|x|)$, $|x|^{k\alpha}$, $k=0,1,2,\ldots$ και μελέτησαν την ασυμπτωτική του συμπεριφορά. Η Hayfavi (1998) προχώρησε στην αναπαράσταση της πυκνότητας πιθανότητας του Linnik σε "improper" ολοκλήρωμα και μελέτησε το επονομαζόμενο «exceptional set». Η εργασία αυτή παρουσιάζει μια περιεκτική σύνοψη του ερευνητικού αυτού πεδίου.

REFERENCES

Hayfavi, A. (1998). An improper integral representation of Linnik's probability densities. *Turkish Journal of Mathematics*, **22(2),** 235-242.

Hardy, G. H., and Wrigth, E. M. (1979). *An Introduction to the Theory of Numbers*, Oxford Science Publications.

Kotz, S., Ostrovskii, I. V. and Hayfavi, A. (1995a). Analytic and asymptotic properties of Linnik's probability densities, I. *Journal of Mathematical Analysis and Applications*, **193(1)**, 353-371.

Kotz, S., Ostrovskii, I, V. and Hayfavi, A. (1995b). Analytic and asymptotic properties of Linnik's probability densities, II. *Journal of Mathematical Analysis and Applications*, **193(2)**, 497-521.

Linnik, Ju. V. (1953). Linear forms and statistical criteria, I. II. Selected Translations Mathematical Statistics and Probabability, 3, (1963), 1-90. American Mathematical Society, Providence, RI (Original paper appeared in: Ukrainskii Mat. Zhurnal, 5, 207-290.

Wittaker, E.T., Watson, G.N. (1962). *A Course of Modern Analysis*, 4^{th} edition, Cambridge University Press, Cambridge.

Εργασίες στα Ελληνικά

Papers in Greek

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 33ου Πανελληνίου Συνεδρίου Στατιστικής (2021), σελ.50-64

ΡΟΗ ΑΘΡΟΙΣΜΑΤΟΣ ΣΤΟ MAX ΔΙΑΓΡΑΜΜΑ ΕΛΕΓΧΟΥ

Δ. Αντζουλάκος[1], Α. Χ. Ρακιτζής[2], Κ. Φουντουκίδης[1]

[1]Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς

dantz@unipi.gr, k.fountoukidis@unipi.gr

[2]Τμήμα Στατιστικής και Αναλογιστικών-Χρηματοοικονομικών Μαθηματικών, Πανεπιστήμιο Αιγαίου

arakitz@aegean.gr

ΠΕΡΙΛΗΨΗ

Στο στατιστικό έλεγχο διεργασιών για μεταβλητές παρακολουθούμε με ξεχωριστά διαγράμματα ελέγχου τη μέση τιμή και τη διασπορά της κατανομής ενός χαρακτηριστικού που μας ενδιαφέρει. Τα τελευταία χρόνια έχει δοθεί μεγάλη έμφαση στην ανάπτυξη διαγραμμάτων ελέγχου τα οποία επιτρέπουν την ταυτόχρονη παρακολούθηση του μέσου και της διασποράς μιας διεργασίας με ένα ενιαίο διάγραμμα ελέγχου. Ένα τέτοιο ενιαίο διάγραμμα είναι το Max διάγραμμα ελέγχου. Στην παρούσα εργασία, προκειμένου να βελτιωθεί η απόδοση του Max διαγράμματος ελέγχου εφαρμόζεται η μέθοδος απόδοσης σκορ σε διάφορες περιοχές (ζώνες) του και παρακολουθείται η σχετική ροή αθροίσματος των σκορ η οποία καθορίζει την εμφάνιση σήματος εκτός ελέγχου διεργασίας. Παρουσιάζεται ο στατιστικός σχεδιασμός του προτεινόμενου διαγράμματος ελέγχου και γίνεται αναλυτική μελέτη των ιδιοτήτων του. Επιπρόσθετα γίνονται συγκρίσεις με άλλα ανταγωνιστικά διαγράμματα ελέγχου χρησιμοποιώντας διάφορα μέτρα απόδοσης των διαγραμμάτων ελέγχου.

Λέξεις κλειδιά: Max διάγραμμα ελέγχου, μαρκοβιανή εμφύτευση, μέσο μήκος ροής, ροή αθροίσματος σκορ, στατιστικός έλεγχος διεργασιών.

1. ΕΙΣΑΓΩΓΗ

Οι τεχνικές στατιστικής παρακολούθησης διεργασιών αποτελούν το βασικό εργαλείο του Στατιστικού Ελέγχου Διεργασιών (ΣΕΔ). Η κυριότερη και πιο διαδεδομένη

τεχνική είναι τα διαγράμματα ελέγχου που βοηθούν στην έγκυρη ανίχνευση μη-φυσιολογικών καταστάσεων. Παραδοσιακά, τα διαγράμματα ελέγχου χρησιμοποιούνται στη βιομηχανία για την παρακολούθηση παραγωγικών διεργασιών, με σκοπό την ανίχνευση της χειροτέρευσης της ποιότητας των παραγόμενων προϊόντων. Για την παρακολούθηση της κατανομής ενός (συνεχούς) χαρακτηριστικού ποιότητας ενός προϊόντος χρησιμοποιούμε συνήθως ένα διάγραμμα ελέγχου για την παρακολούθηση του μέσης τιμής και ένα (ξεχωριστό) διάγραμμα ελέγχου για την παρακολούθηση της τυπικής απόκλισης.

Τα τελευταία χρόνια έχει δοθεί αρκετή έμφαση στην ανάπτυξη διαγραμμάτων ελέγχου τα οποία παρακολουθούν ταυτόχρονα τον μέσο και την τυπική απόκλιση τα οποία αναφέρονται στη βιβλιογραφία ως ενιαία διαγράμματα ελέγχου. Το κύριο πλεονέκτημά τους είναι ότι με ένα διάγραμμα ελέγχου και απεικονίζοντας σε αυτό μία μόνο στατιστική συνάρτηση, μπορούμε να ανιχνεύσουμε ταυτόχρονα μετατοπίσεις του μέσου ή/και της τυπικής απόκλισης της κατανομής του χαρακτηριστικού που παρακολουθούμε. Το Max διάγραμμα ελέγχου είναι ένα τέτοιο διάγραμμα, ουσιαστικά τα κριτήρια που αναφέρουμε τα πληροί καθώς οι στατιστικές συναρτήσεις που συνδυάζει στην εντός ελέγχου διεργασίας έχουν την ίδια κατανομή. Αρκετοί ερευνητές έχουν αναπτύξει και μελετήσει τέτοια διαγράμματα ελέγχου. Οι Cheng & Li (1993) πρότειναν το T διάγραμμα ελέγχου, οι Chen & Chao (1996) πρότειναν το ημικυκλικό διάγραμμα ελέγχου, οι Chen & Cheng (1998) ανέπτυξαν το Max διάγραμμα ελέγχου, οι Chen et al. (2001) ανέπτυξαν το Max-EWMA διάγραμμα ελέγχου, ο Xie (1999) ανέπτυξε το SS-EWMA και ο Thaga (2003) ανέπτυξε το Max-CUSUM διάγραμμα ελέγχου. Για μια πρόσφατη επισκόπηση της περιοχής των ενιαίων διαγραμμάτων ελέγχου δείτε τους Thaga & Sivasamy (2015).

Ο Roberts (1966) εισήγαγε το διάγραμμα ελέγχου ροής αθροίσματος (Run Sum) το οποίο μελετήθηκε περεταίρω από τον Reynolds (1971). Αρχικά, το διάγραμμα αυτό προτάθηκε για την παρακολούθηση του μέσου μιας διεργασίας. Σύμφωνα με αυτή την προσέγγιση οι δύο ζώνες πάνω και κάτω από την κεντρική γραμμή χωρίζονται σε περιοχές που τους αποδίδεται ένα σκορ, θετικό ή αρνητικό. Η λειτουργία του διαγράμματος βασίζεται στην παρακολούθηση της ροής αθροίσματος των σκορ και όταν η απόλυτη τιμή της στατιστικής της ροής αθροίσματος υπερβεί μια θετική τιμή, τότε δίνεται σήμα εκτός ελέγχου διεργασίας. Σύμφωνα με τους Champ & Ridgon (1997), όσο ο αριθμός των περιοχών αυξάνεται, τόσο το διάγραμμα ελέγχου ροής αθροίσματος γίνεται ολοένα και πιο ανταγωνιστικό με τα CUSUM και EWMA διαγράμματα ελέγχου. Ωστόσο η εφαρμογή του διαγράμματος ελέγχου ροής αθροίσματος γίνεται περισσότερο περίπλοκη όσο ο αριθμός των περιοχών αυξάνεται.

Οι Champ & Ridgon (1997) ανέπτυξαν μία τεχνική βασισμένη στις Μαρκοβιανές αλυσίδες για τη μελέτη της απόδοσης των διαγραμμάτων ελέγχου ροής αθροίσματος. Ο Jaehn (1991) πρότεινε μία ειδική περίπτωση του διαγράμματος ελέγχου ροής αθροίσματος, που ονόμασε διάγραμμα ελέγχου ζωνών, οι Davis et al. (1994) πρότειναν και μελέτησαν διάφορα διαγράμματα ελέγχου ζωνών και τα σύγκριναν με διαγράμματα ελέγχου εφοδιασμένα με κανόνες ροών, και οι Acosta-Mejia & Pignatiello (2010), Aguirre-Torres & Reyes-Lopez (1999) και Rakitzis & Antzoulakos (2016) χρησιμοποίησαν διαγράμματα ελέγχου ροής αθροίσματος για την παρακολούθηση της διασποράς μιας διεργασίας. Διαγράμματα ελέγχου που χρησιμοποιούν την τεχνική ροής αθροίσματος έχουν προταθεί, μεταξύ άλλων, και από τους Khoo et al. (2013), Acosta-Meija (2013), Teoh (2016) Han and Khooo (2019) και Abubakar et al. (2020).

Στην παρούσα εργασία εφαρμόζουμε την τεχνική ροής αθροίσματος στο Max διάγραμμα ελέγχου για τη βελτίωση της ικανότητας του διαγράμματος στην ταυτόχρονη παρακολούθηση του μέσου και της μεταβλητότητας της παραγωγικής διεργασίας. Εξετάζουμε τις ιδιότητες και την αποτελεσματικότητα του διαγράμματος για διάφορα μεγέθη δείγματος και προκαθορισμένες εντός ελέγχου τιμές του μέσου μήκους ροής (Average Run Length ή ARL) χρησιμοποιώντας τεχνικές Μαρκοβιανών αλυσίδων. Εξετάζεται επίσης η συνολική αποτελεσματικότητα του προτεινόμενου διαγράμματος ελέγχου χρησιμοποιώντας ως μέτρο το αναμενόμενο μέσο μήκος ροής (Expected Average Run Length ή EARL). Επιπλέον, γίνεται σύγκριση της απόδοσης του διαγράμματος με άλλα ανταγωνιστικά διαγράμματα ελέγχου με βάση το ARL. Πιο συγκεκριμένα, στην Ενότητα 2 παρουσιάζουμε το Max διάγραμμα ελέγχου και μελετάμε την ροή αθροίσματος σε αυτό. Στην Ενότητα 3, παρουσιάζουμε τα αποτελέσματα αριθμητικής μελέτης σχετικά με την απόδοση του διαγράμματος ελέγχου $\mathrm{RSMax}_H(a_1, a_2, a_3, a_4)$. Τέλος, στη Ενότητα 4 συνοψίζονται τα συμπεράσματα της μελέτης.

2. ΡΟΗ ΑΘΡΟΙΣΜΑΤΟΣ ΣΤΟ MAX ΔΙΑΓΡΑΜΜΑ ΕΛΕΓΧΟΥ

2.1 Το Max διάγραμμα ελέγχου

Υποθέτουμε ότι η εντός ελέγχου κατανομή των παρατηρήσεων είναι η κανονική κατανομή με εντός ελέγχου μέση τιμή μ_0 και διασπορά σ_0^2. Από τη διεργασία επιλέγουμε τυχαία δείγματα μεγέθους n σε τακτά χρονικά διαστήματα. Έστω $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{in})$ το δείγμα που πήραμε στην i-οστή δειγματοληψία ($i = 1,2, \ldots$). Όταν η διεργασία λειτουργεί υπό συνθήκες φυσικής μεταβλητότητας τότε $X_{ij} \sim N(\mu_0, \sigma_0^2)$, $i = 1,2, \ldots,$ και $j = 1,2, \ldots, n$.

Μια εκτός ελέγχου τιμή για το μέσο θα εκφράζεται ως $\mu_1 = \mu_0 + a\sigma_0$ ($a \in \mathbb{R}$) και μια εκτός ελέγχου τιμή για την τυπική απόκλιση ως $\sigma_1 = b\sigma_0$ $(b \geq 0)$. Οι τιμές $a = 0$ και $b = 1$ οδηγούν στις εντός ελέγχου τιμές της μέσης τιμής και της τυπικής απόκλισης.

Έστω $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ ένα τυχαίο δείγμα μεγέθους n από την κανονική κατανομή $N(\mu, \sigma^2)$, και έστω οι στατιστικές συναρτήσεις

$$\bar{X} = \frac{(X_1 + X_2 + \cdots + X_n)}{n}, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Οι Chen και Cheng (1998) όρισαν τις ακόλουθες δύο στατιστικές συναρτήσεις

$$U = \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}}$$

$$V = \Phi^{-1}\left\{H_{n-1}\left(\frac{(n-1)S^2}{\sigma^2}\right)\right\}$$

όπου $\Phi(\cdot)$ είναι η συνάρτηση κατανομής της τυπικής κανονικής κατανομής και $H_n(\cdot)$ είναι η συνάρτηση κατανομή της χι-τετράγωνο κατανομής με n βαθμούς ελευθερίας (συμβ. χ_n^2). Οι στατιστικές συναρτήσεις U, V είναι ανεξάρτητες διότι οι στατιστικές συναρτήσεις $\bar{X}$ και S^2 είναι ανεξάρτητες στην περίπτωση τυχαίου δείγματος από κανονική κατανομή.

Η απεικονιζόμενη στατιστική συνάρτηση στο Max διάγραμμα ελέγχου είναι η

$$M = \max\{|U|, |V|\}.$$

Η στατιστική M θα παίρνει μεγάλες (θετικές) τιμές όταν ο μέσος της διεργασίας μετατοπιστεί ή/και η μεταβλητότητα της διεργασίας αυξηθεί η μειωθεί. Από την άλλη, η στατιστική συνάρτηση M θα παίρνει μικρές (θετικές) τιμές όταν ο μέσος και η μεταβλητότητα της διεργασίας παραμένουν κοντά στις εντός ελέγχου τιμές τους.

Το άνω όριο ελέγχου UCL του Max διαγράμματος ελέγχου θα υπολογιστεί με τέτοιο τρόπο έτσι ώστε να έχουμε προκαθορισμένη πιθανότητα σφάλματος τύπου I ίση με α. Δηλαδή όταν η διεργασία είναι εντός ελέγχου ($X \sim N(\mu_0, \sigma_0^2)$), ισχύει ότι

$$1 - \alpha = P(M \leq UCL).$$

Όμως, για $y \geq 0$, η εντός ελέγχου συνάρτηση κατανομής $F_M(y)$ της M είναι ίση με

$$F_M(y) = P(\max(|U|, |V|) \le y) = P(|U| \le y, |V| \le y)$$
$$= P(|U| \le y)P(|V| \le y)$$
$$= \big(\Phi(y) - \Phi(-y)\big)^2 = [P(\chi_1^2 \le y^2)]^2.$$

Συνεπώς, στο Max διάγραμμα ελέγχου έχουμε ότι

$$UCL = \sqrt{\chi^2_{1;\sqrt{1-a}}}, \quad CL = \sqrt{\chi^2_{1;a}}$$

όπου $\chi^2_{1;\gamma}$ είναι το γ ποσοστιαίο σημείο της κατανομής χ^2_1. Όταν χρησιμοποιούμε τα όρια πιθανότητας, είναι σύνηθες να χρησιμοποιούμε την διάμεσο γραμμή ως κεντρική γραμμή.

2.2 Το Run Sum Max διάγραμμα ελέγχου

Σε ένα άνω μονόπλευρο διάγραμμα ελέγχου, όπως είναι το Max διάγραμμα ελέγχου, η εφαρμογή της τεχνικής της ροής αθροίσματος απαιτεί να χωριστεί η ζώνη πάνω από την κεντρική γραμμή σε περιοχές στις οποίες αντιστοιχούμε διάφορα σκορ. Όσο οι διαδοχικές τιμές της στατιστικής συνάρτησης M πέφτουν στην πάνω πλευρά της κεντρικής γραμμής, τα διαδοχικά σκορ αθροίζονται. Ωστόσο όταν μια τιμή της στατιστικής συνάρτησης M πέσει στην κάτω πλευρά της κεντρικής γραμμής το αθροιστικό σκορ μηδενίζεται. Όταν το άθροισμα των σκορ ξεπεράσει μια κρίσιμη τιμή τότε το διάγραμμα ελέγχου δίνει σήμα εκτός ελέγχου διεργασίας.

Πιο συγκεκριμένα, έστω ότι η ζώνη πάνω από την κεντρική γραμμή χωρίζεται σε $k+1$ περιοχές που οριοθετούνται από k άνω όρια ελέγχου, τα $UCL_1 < UCL_2 < \cdots < UCL_k$. Στις $k+1$ περιοχές $[CL, UCL_1), [UCL_1, UCL_2)$, …, $[UCL_k, \infty)$ αντιστοιχούμε τα μη αρνητικά σκορ $a_1, a_2, \dots, a_{k+1}$. Η στατιστική συνάρτηση που απεικονίζεται στο διάγραμμα ελέγχου είναι η

$$CU_i = \begin{cases} CU_{i-1} + a_{j+1}, & \text{αν } UCL_j \le M_i \le UCL_{j+1} \\ 0, & \text{αν } M_i < CL \end{cases}$$

όπου M_i η τιμή της στατιστικής συνάρτησης M για το i-οστό δείγμα ($i = 1,2, \dots$) και $CU_0 \ge 0$. Το διάγραμμα ελέγχου δίνει σήμα εκτός ελέγχου διεργασίας όταν $CU_i \ge H$, όπου H είναι μία κατάλληλη κριτική τιμή. Ονομάζουμε το διάγραμμα αυτό ως Run Sum Max διάγραμμα ελέγχου κα το συμβολίζουμε ως $\text{RSMax}_H(a_1, a_2, \dots, a_{k+1})$.

Έστω $p_0 = P(M \le UCL_0)$ και

$$p_j = P\big(UCL_{j-1} \le M \le UCL_j\big), \quad j = 1,2, \dots, k+1$$

όπου $UCL_0 = CL$ και $UCL_{k+1} = \infty$. Θεωρώντας ότι οι παρατηρήσεις μας προέρχονται από κανονική κατανομή με μέση τιμή $\mu_1 = \mu_0 + a\sigma_0$ και τυπική

απόκλιση $\sigma_1 = b\sigma_0$, είναι εύκολο να διαπιστωθεί ότι ο υπολογισμός των πιθανοτήτων p_j προκύπτει από τις σχέσεις

$$p_0 = F(UCL_0)$$

$$p_j = F(UCL_j) - F(UCL_{j-1}), \quad j = 1,2,\dots,k+1,$$

όπου

$$F(y) = \left\{\Phi\left(\frac{y}{b} - \sqrt{n}\frac{a}{b}\right) - \Phi\left(-\frac{y}{b} - \sqrt{n}\frac{a}{b}\right)\right\} \times \left\{H_{n-1}\left(\frac{\chi^2_{n-1;\Phi(y)}}{b^2}\right) - H_{n-1}\left(\frac{\chi^2_{n-1;\Phi(-y)}}{b^2}\right)\right\}.$$

Τώρα σχετικά με την επιλογή των $UCL_1 < UCL_2 < \cdots < UCL_k$, προτείνουμε, σύμφωνα με τον Khoo et al. (2013), τα παρακάτω όρια ελέγχου

$$UCL_j = L \times \sqrt{\chi^2_{1;\sqrt{\Phi(j)}}}, \quad j = 1,2,\dots,k$$

όπου η τιμή της παραμέτρου L καθορίζεται έτσι ώστε να έχουμε το επιθυμητό ARL_0.

Στο παρόν άρθρο θα ασχοληθούμε αποκλειστικά με τέσσερις περιοχές, τις $[CL, UCL_1)$, $[UCL_1, UCL_2)$, $[UCL_2, UCL_3)$ και $[UCL_3, \infty)$, με αντίστοιχα σκορ a_1, a_2, a_3 και a_4, και αντίστοιχες πιθανότητες p_1, p_2, p_3 και p_4.

Για τη στατιστική σχεδίαση του $\text{RSMax}_H(a_1,a_2,a_3,a_4)$ διαγράμματος προτείνονται τα παρακάτω βήματα:

Βήμα 1: Επιλέγουμε το μέγεθος n των δειγμάτων, το διάνυσμα των σκορ (a_1,a_2,a_3,a_4), και την κρίσιμη τιμή H.
Βήμα 2: Θέτουμε $ARL_0 = c$.
Βήμα 3: Υπολογίζουμε την μοναδική τιμή της L (και συνεπώς τα όρια ελέγχου UCL_j $j = 1,2,\dots,k$) έτσι ώστε $ARL_0 = c$.
Βήμα 4: Η διαδικασία είναι εκτός ελέγχου στο i −οστό δείγμα αν $CU_i \geq H$.

3. ΑΝΑΛΥΣΗ ΤΟΥ ΔΙΑΓΡΑΜΜΑΤΟΣ $\text{RSMax}_H(a_1,a_2,a_3,a_4)$

Στην παρούσα ενότητα παρουσιάζονται τα αποτελέσματα μιας εκτεταμένης αριθμητικής μελέτης σχετικά με την στατιστική σχεδίαση και απόδοση διαφόρων $\text{RSMax}_H(a_1,a_2,a_3,a_4)$ διαγραμμάτων ελέγχου για διάφορα σύνολα σκορ και κρίσιμων τιμών. Συγκεκριμένα μελετήσαμε τα εξής διαγράμματα ελέγχου:

C1: $\text{RSMax}_5(0,1,2,3)$, C2: $\text{RSMax}_4(0,1,2,4)$, C3: $\text{RSMax}_{12}(0,1,6,12)$,

C4: $\text{RSMax}_{12}(0,3,4,12)$, C5: $\text{RSMax}_8(0,2,3,8)$, C6: $\text{RSMax}_{14}(1,2,7,14)$,

C7: $RSMax_{19}(1,3,11,19)$, C8: $RSMax_{15}(1,3,8,15)$, C9: $RSMax_{13}(1,2,7,13)$.

Τα παραπάνω σκορ προτάθηκαν από τους Reynolds (1971), Woodall (1990) και Davis et al. (1994).

Τιμές του μέσου μήκους ροής των εννέα παραπάνω διαγραμμάτων ελέγχου δίνονται στον Πίνακα 2. Για τον υπολογισμό του μέσου μήκους ροής χρησιμοποιούμε τη μέθοδο της εμφύτευσης σε Μαρκοβιανή αλυσίδα των Champ and Rigdon (1997) (δείτε επίσης τους Rakitzis & Anzoulakos (2016)). Χρησιμοποιήθηκε εντός ελέγχου μέσο μήκος ροής $ARL_0 = 200$, μέγεθος δείγματος $n = 5$ και αρχική τιμή $CU_0 = 0$. Στις δύο πρώτες στήλες δίνονται τιμές των a και b που καθορίζουν την μετατόπιση του μέσου και της τυπικής απόκλισης, αντίστοιχα, ενώ στην τρίτη στήλη Max δίνονται τιμές ARL του κλασικού Max διαγράμματος. Οι έντονες στο χρώμα τιμές δείχνουν την μικρότερη τιμή του εκτός ελέγχου μέσου μήκους ροής. Τα όρια ελέγχου και η κεντρική γραμμή των διαγραμμάτων ελέγχου δίνονται στον Πίνακα 1.

Πίνακας 1. *Τιμές των ορίων ελέγχου για τα MAX και RSMax διαγράμματα ελέγχου: $ARL_0 = 200$, $n = 5$*

Limits	*L*	UCL_3	UCL_2	UCL_1	*CL*
C1	0.900326	3.06063	2.27688	1.56201	1.051796
C2	1.00463	3.41521	2.54066	1.74297	1.051796
C3	0.932917	3.17142	2.3593	1.61855	1.051796
C4	0.953877	3.24267	2.41231	1.65492	1.051796
C5	0.962262	3.27118	2.43351	1.66947	1.051796
C6	0.971105	3.30124	2.45587	1.68481	1.051796
C7	0.979884	3.33108	2.47808	1.70004	1.051796
C8	0.992274	3.3732	2.50941	1.72154	1.051796
C9	1.00749	3.42493	2.54789	1.74793	1.051796

MAX: UCL=3.022962, CL=1.051796

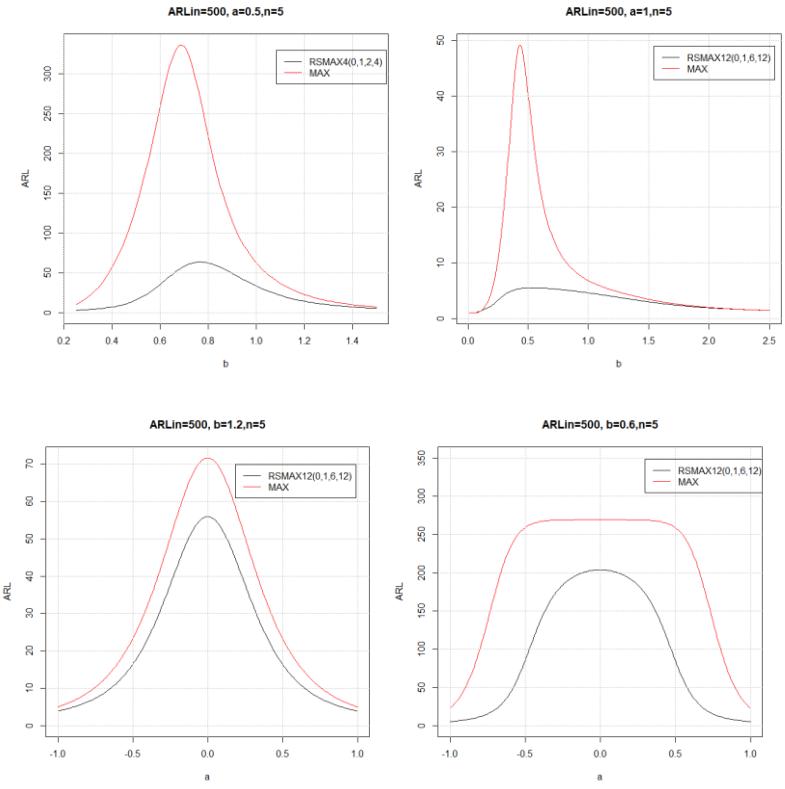
Πίνακας 2. *Τιμές ARL για τα RSMax διαγράμματα ελέγχου:* $ARL_0 = 200,\ n = 5$

a	b	Max	C1	C2	C3	C4	C5	C6	C7	C8	C9
	0.25	5.09	2.95	2.78	2.71	3.31	3.06	3.00	2.95	3.06	3.11
	0.5	55.23	16.71	18.50	27.53	19.49	18.99	20.40	20.18	19.90	18.80
0.0	1	200	200	200	200	200	200	200	200	200	200
	1.2	37.98	31.41	29.47	31.37	32.27	30.98	30.62	30.23	30.89	30.45
	1.5	7.57	7.39	6.37	6.30	6.96	6.66	6.56	6.51	6.70	6.68
	0.25	5.09	2.95	2.78	2.71	3.31	3.05	3.00	2.95	3.06	3.11
	0.5	55.02	9.62	11	17.66	10.72	10.70	11.70	11.82	11.39	10.76
0.5	1	32.42	21.73	21.32	24.18	23.65	22.59	22.67	22.28	22.69	22.14
	1.2	14.21	11.71	10.67	11.11	11.84	11.27	11.18	11.03	11.33	11.19
	1.5	5.33	5.59	4.71	4.56	5.10	4.90	4.81	4.79	4.93	4.93
	0.25	5.08	2.78	2.63	2.38	3.17	2.91	2.73	2.73	2.84	2.96
	0.5	13.36	3.50	3.46	3.96	4.03	3.76	4.10	3.94	4.08	3.97
1.0	1	4.59	4.14	3.70	3.63	4.08	3.89	3.83	3.81	3.92	3.91
	1.2	3.75	4.02	3.95	3.23	3.66	3.52	3.44	3.44	3.53	3.55
	1.5	2.74	3.47	2.73	2.54	2.86	2.78	2.71	2.72	2.79	2.82
	0.25	3.75	2.59	2.17	1.90	2.77	2.69	2.04	2.08	2.14	2.23
	0.5	3.82	2.74	2.53	2.23	3.04	2.81	2.57	2.58	2.67	2.76
1.2	1	2.71	3.07	2.57	2.39	2.76	2.66	2.57	2.58	2.65	2.69
	1.2	2.52	3.13	2.50	2.32	2.63	2.56	2.48	2.49	2.55	2.59
	1.5	2.15	2.97	2.23	2.06	2.30	2.26	2.19	2.21	2.26	2.29
	0.25	1.08	2.00	1.55	1.20	1.37	1.44	1.37	1.42	1.49	1.57
	0.5	1.33	2.09	1.59	1.36	1.58	1.61	1.48	1.51	1.55	1.61
1.5	1	1.59	2.36	1.72	1.56	1.74	1.73	1.66	1.68	1.72	1.76
	1.2	1.63	2.46	1.75	1.61	1.77	1.75	1.70	1.72	1.75	1.79
	1.5	1.60	2.49	1.72	1.59	1.72	1.71	1.67	1.69	1.71	1.75

Το γενικό συμπέρασμα που προκύπτει από τον Πίνακα 2 είναι ότι για μεγάλες ανοδικές/αυξητικές μετατοπίσεις στο μέσο ($a > 1$) και ανεξαρτήτως του μεγέθους της μετατόπισης της τυπικής απόκλισης την καλύτερη απόδοση έχει σχεδόν πάντα το διάγραμμα ελέγχου C3. Στις υπόλοιπες περιπτώσεις κάποιο εκ των C1, C2 και C3 έχει την καλύτερη απόδοση.

Παρακάτω παρουσιάζουμε ενδεικτικά διάφορες γραφικές παραστάσεις τιμών ARL των διαγραμμάτων ελέγχου RSMax C2 και C3 καθώς επίσης και του συνήθους διαγράμματος ελέγχου Max, για διάφορες μετατοπίσεις στο μέσο ή/και στην τυπική απόκλιση.

Εικόνα 1. *Διαγράμματα τιμών ARL για τα* $RSMax_4(0,1,2,4)$, $RSMax_{12}(0,1,6,12)$ *και* Max *διαγράμματα ελέγχου*



Συμπερασματικά από τα παραπάνω σχήματα για τις συγκρίσεις που κάναμε προκύπτει ότι τα RSMax διαγράμματα ελέγχου έχουν πάντα καλύτερη απόδοση σε σχέση με το Max. Ωστόσο για ακραίες μετατοπίσεις του μέσου ή/και της τυπικής απόκλισης η υπεροχή του RSMax διαγράμματος ελέγχου μετριάζεται.

Παραπάνω εξετάσαμε την αποτελεσματικότητα του RSMax διαγράμματος για γνωστές μετατοπίσεις στο μέσο και στην τυπική απόκλιση. Στην πράξη όμως δεν γνωρίζουμε την πραγματική τιμή των μετατοπίσεων, δηλαδή τις τιμές των παραμέτρων a, b. Μια πιο ρεαλιστική υπόθεση είναι ότι οι μετατοπίσεις a, b είναι τυχαίες μεταβλητές. Σε αυτήν την περίπτωση η συνολική αποτελεσματικότητα του διαγράμματος ελέγχου μπορεί να αποτιμηθεί με το αναμενόμενο σταθμισμένο μήκος ροής (expected weighted run-length ή EWRL)) που δίνεται από τον τύπο

$$EWRL = \int_{a_{min}}^{a_{max}} \int_{b_{min}}^{b_{max}} w(a,b) ARL(a,b)\, f(a,b) dbda$$

(δείτε Mukherjee & Sen (2018)), όπου $f(a,b)$, για $a_{min} \le a \le a_{max}$ και $b_{min} \le b \le b_{max}$, η από κοινού συνάρτηση πυκνότητας των τυχαίων μεταβλητών a, b, $w(a,b)$ είναι κατάλληλο βάρος συνήθως μη αρνητικό, και $ARL(a,b)$ το μέσο μήκος ροής για τα συγκεκριμένα a, b. Συνήθως $w(a,b)=1$, οπότε το $EWRL$ ονομάζεται αναμενόμενο μέσο μήκος ροής (expected average run length ή EARL), Ryu et al. (2010).

Στη συνέχεια παρουσιάζουμε την αποτελεσματικότητα των RSMax διαγραμμάτων ελέγχου χρησιμοποιώντας ως μέτρο απόδοσης το EARL. Ως από κοινού συνάρτηση πυκνότητας των τυχαίων μεταβλητών a, b θεωρούμε τη διδιάστατη ομοιόμορφη κατανομή, δηλαδή

$$f(a,b) = \frac{1}{(a_{max}-a_{min})(b_{max}-b_{min})}, \quad a_{min} \le a \le a_{max},\ b_{min} \le b \le b_{max}.$$

Προφανώς όσο μικρότερη είναι η τιμή του EARL τόσο καλύτερη απόδοση έχει το διάγραμμα ελέγχου.

Πίνακας 3. *Τιμές EARL για τα* RSMax *διαγράμματα ελέγχου,* $a \in [0.3, 2]$ *και* $b \in [0.3, 2]$

	$ARL_0 = 200$		$ARL_0 = 500$	
	$n = 5$	$n = 10$	$n = 5$	$n = 10$
Max	10.8966	3.03307	24.7173	5.8567
$RSMax_5(0,1,2,3)$	5.74659	2.97794	8.36368	3.47646
$RSMax_4(0,1,2,4)$	5.56896	2.30526	9.06091	2.84361
$RSMax_{12}(0,1,6,12)$	6.36598	2.37598	12.5429	3.33545
$RSMax_{12}(0,3,4,12)$	5.57024	2.42394	8.93554	3.11908
$RSMax_8(0,2,3,8)$	5.43125	2.33895	8.52398	3.00982
$RSMax_{14}(1,2,7,14)$	5.6397	2.3491	9.05014	3.06656
$RSMax_{19}(1,3,11,19)$	5.65234	2.34967	9.18396	2.95956
$RSMax_{15}(1,3,8,15)$	5.55136	2.42604	9.01893	2.9914
$RSMax_{13}(1,2,7,13)$	5.76251	2.34419	8.86046	3.05815

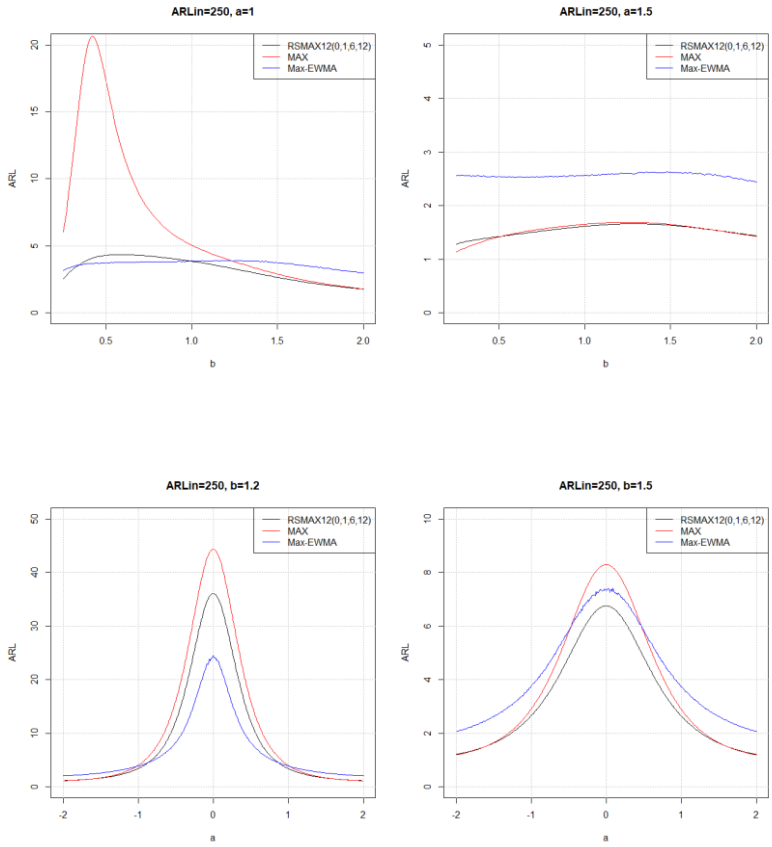
Πίνακας 4. *Τιμές EARL για τα* RSMax *διαγράμματα ελέγχου,* $a \in [0.2,1]$ *και* $b \in [0.2,1]$

	$ARL_0 = 200$		$ARL_0 = 500$	
	$n = 5$	$n = 10$	$n = 5$	$n = 10$
Max	50.8511	13.4053	127.878	29.7315
$RSMax_5(0,1,2,3)$	21.6464	6.82962	43.7367	10.6629
$RSMax_4(0,1,2,4)$	24.3814	6.80852	51.4205	10.402
$RSMax_{12}(0,1,6,12)$	32.0523	8.19683	73.2454	15.0909
$RSMax_{12}(0,3,4,12)$	24.0734	7.27154	50.7586	11.8029
$RSMax_8(0,2,3,8)$	23.6438	7.1103	50.4798	11.326
$RSMax_{14}(1,2,7,14)$	25.1218	7.32586	51.1332	11.7975
$RSMax_{19}(1,3,11,19)$	25.234	6.9566	52.6417	11.2361
$RSMax_{15}(1,3,8,15)$	24.6878	7.30896	48.9352	11.6112
$RSMax_{13}(1,2,7,13)$	23.4488	6.94337	47.3288	11.3598

Παρατηρούμε από τους παραπάνω πίνακες ότι το Max διάγραμμα ελέγχου έχει τη χειρότερη συνολική απόδοση σε κάθε περίπτωση. Μεταξύ των RSMax διαγραμμάτων, το $RSMax_5(0,1,2,3)$ έχει σχεδόν πάντα την καλύτερη απόδοση, ενώ το $RSMax_{12}(0,1,6,12)$ τη χειρότερη απόδοση.

Στη συνέχεια συγκρίνουμε το $RSMax_{12}(0,1,6,12)$ διάγραμμα ελέγχου που φαίνεται να έχει την καλύτερη απόδοση σε όρους ARL (δείτε Πίνακα 2) με το Max-EWMA διάγραμμα ελέγχου που πρότειναν οι Chen et al. (2001). Θέτουμε το εντός ελέγχου μέσο μήκος ροής να είναι 250 και για το Max-EWMA επιλέγουμε $K = 2.785$ και $\lambda = 0.10$ ώστε να έχουμε το επιθυμητό εντός ελέγχου μέσο μήκος ροής. Ακολούθως παρουσιάζουμε κάποια σχήματα τιμών *ARL* για διάφορες μετατοπίσεις στο μέσο ή/και στη μεταβλητότητα.

Εικόνα 2. *Διαγράμματα τιμών ARL για τα* RSMax, Max *και* Max-EWMA *διαγράμματα ελέγχου*



Παρατηρούμε από τις παραπάνω συγκρίσεις ότι για μικρές προς μέτριες αυξήσεις στη μεταβλητότητα και μεγάλη αύξηση στο μέσο της διεργασίας το RSMax και το Max είναι καλύτερα στον εντοπισμό των μετατοπίσεων έναντι του Max-EWMA διαγράμματος ελέγχου. Επίσης για μικρή αύξηση στη μεταβλητότητα και μεγάλη αύξηση στο μέσο τα RSMax και Max διαγράμματα ελέγχου είναι καλύτερα από το Max-EWMA. Τέλος για μεγάλες αυξήσεις στην μεταβλητότητα και μικρή αύξηση στο μέσο το RSMax είναι καλύτερο στον εντοπισμό της μετατόπισης από ότι το Max και Max-EWMA.

4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα εργασία προτείνεται η χρήση της ροής αθροίσματος στο Max διάγραμμα ελέγχου για την παρακολούθηση του μέσου και της μεταβλητότητας μιας διεργασίας. Αυτά τα διαγράμματα ελέγχου βελτιώνουν την αποτελεσματικότητα του Max διαγράμματος στον εντοπισμό μικρών προς μέτριων μετατοπίσεων στο μέσο ή/και στην μεταβλητότητα και προσφέρουν το πλεονέκτημα της επιλογής του επιθυμητού εντός ελέγχου μέσου μήκους ροής. Η αποτελεσματικότητα του $\mathrm{RSMax}_H(a_1,a_2,a_3,a_4)$ διαγράμματος ελέγχου αξιολογήθηκε χρησιμοποιώντας την τεχνική της εμφύτευσης σε Μαρκοβιανή αλυσίδα για τον ακριβή υπολογισμό του μέσου μήκους ροής. Επίσης παρουσιάσαμε το καλύτερα διαγράμματα για τον εντοπισμό συγκεκριμένων μετατοπίσεων στον μέσο ή/και στην μεταβλητότητα. Ανάλογα με το μέγεθος δείγματος *n* και το εντός ελέγχου μέσο μήκος ροής, προτείνουμε τη χρήση των $\mathrm{RSMax}_5(0,1,2,3)$ ή $\mathrm{RSMax}_4(0,1,2,4)$ διαγραμμάτων ελέγχου για (α) τον εντοπισμό μικρών αυξήσεων στο μέσο ($0 < a \leq 0.5$), (β) μεγάλων μειώσεων στη μεταβλητότητα ($b \leq 0.5$), και (γ) μικρών προς μέτριων αυξήσων στη μεταβλητότητα ($1 \leq b \leq 1.5$). Επιπρόσθετα προτείνουμε τη χρήση του $\mathrm{RSMax}_{12}(0,1,6,12)$ διαγράμματος για μεγάλες μετατοπίσεις στο μέσο ($a > 1$) σε συνδυασμό είτε με μεγάλες μειώσεις στην μεταβλητότητα ($b \leq 0.5$), είτε για μικρές προς μέτριες αυξήσεις στη μεταβλητότητα ($1 \leq b \leq 1.5$). Επιπλέον, το προτεινόμενο διάγραμμα ελέγχου είναι καλύτερο στον εντοπισμό μεγάλων αυξήσεων στον μέσο ($a > 1$) και μικρές προς μέτριες αυξήσεις στη μεταβλητότητα ($1 \leq b \leq 1.5$) από το Max-EWMA. Όπως επίσης και για μικρή αύξηση στο μέσο ($a \leq 0.8$) και για μεγάλη μετατόπιση στη μεταβλητότητα ($b \geq 1.5$).

Η συνολική αποτελεσματικότητα του προτεινόμενου διαγράμματος ελέγχου $\mathrm{RSMax}_H(a_1,a_2,a_3,a_4)$ αποτιμήθηκε και με τιμές του EARL για διάφορα σενάρια για το εύρος των μετατοπίσεων. Οι υπολογισμοί μας έδειξαν ότι τα διαγράμματα $\mathrm{RSMax}_5(0,1,2,3)$ και $\mathrm{RSMax}_4(0,1,2,4)$ έχουν την καλύτερη συνολική αποτελεσματικότητα.

ABSTRACT

In statistical process control for variables the process mean and the variability of a process are monitored separately. In recent years, it has been a great effort on developing control charts which monitor simultaneously the process mean and the variability with a single control chart. The Max chart is such a chart. In this paper, in order to improve the effectiveness of the Max chart we apply the scoring method in different areas (zones) of the chart and we calculate the run sum of these scores which

determines the occurrence of an out of control signal. The statistical design of the proposed control chart is presented and a detailed study of its properties is performed. In addition, comparisons are made with other competing control charts using various performance measures of control charts.

ΑΝΑΦΟΡΕΣ

Abubakar, S.S., Khoo, M.B.C., Saha, S. & Teoh, W.L. (2020): Run sum control chart for monitoring the ratio of population means of a bivariate normal distribution, *Communications in Statistics - Theory and Methods*,

Acosta-Mejia, C.A. (2013). Two-sided charts for monitoring nonconforming parts per million. *Quality Engineering*, 25, 34–45.

Acosta-Mejia, C.A. & Pignatello,J.J. Jr (2008). Modified R charts for improved performance. *Quality Engineering*, 20, 361-369.

Acosta-Mejia, C.A. & Pignatello,J.J. Jr (2010). The run sum R chart with fast initial response. *Communications in Statistics-Simulation and Computation*, 39, 921-932.

Aguirre-Torres, V. & Reyes-Lopes, D. (1999). Run sum charts for both $\bar{X}$ και R. *Quality Engineering*, 12, 7-12.

Antzoulakos, D.L. & Rakitzis, A.C. (2016) Run sum control charts for the monitoring of process variability. *Quality Technology & Quantitative Management*, 13:1, 58-77.

Champ, C.W. & Rigdon, S.E. (1997). Analysis of the run sum control chart. *Journal of Quality Technology*, 29, 407–417.

Cheng SW, Li GY. (1993). A single variables control chart. *Technical Report*, University of Manitoba, Winnipeg, Canada.

Chen, G. & Cheng, S. W. (1998). MAX chart: Combining X-bar chart and S chart. *Statistica Sinica*, 8:263–271.

Chen, G., Cheng, S. W. & Xie, H. (2001). Monitoring process mean and variability with one EWMA chart. *J. Qual. Technol*., 33:223–233.

Cheng, S. W. & Thaga, K. (2006). Single variables control charts: An overview. *Quality Reliability Engineering*. *Int.*, 22:811–820.

Davis, R.B., Chun, J. & Guo, Y. (1994). Improving the performance of the zone control chart. *Communications in Statistics–Theory and Methods*, 23, 3557–3565.

Davis, R.B., Homer, A. & Woodall, W.H. (1990). Performance of the zone control chart. *Communications in Statistics–Theory and Methods*, 19, 1581–1587.

Fu, J.C., Spiring, F.A. & Xie, H. (2002). On the average run lengths of quality control schemes using a Markov chain approach. *Statistics and Probability Letters*, 56, 369–380.

Han. G., Khoo, M.B.C.,Teh S.Y. & Teoh, W.L.(2019). A Study on the Median Run

Length Performance of the Run Sum S Control Chart. International Journal of Mechanical Engineering and Robotics Research Vol. 8, No. 6, 885-890.

Jaehn, A. H. (1991). Zone Control Charts: A New Tool for Quality Control. *Tappi Journal*, 70, 159-161.

Khoo, M.B.C., Sitt, C.K., Wu, Z. & Castagliola, P. (2013). A run sum Hotelling's χ^2 control chart. *Computers and Industrial Engineering*, 64, 686–695.

Mukherjee, A. & Sen, R. (2018). Optimal design of Shewhart-Lepage type schemes and its application in monitoring service quality. *European Journal of Operational Researc*, 266, 147-167.

Reynolds, J.H. (1971). The run sum control chart procedure. *Journal of Quality Technology*, 3, 23–27.

Roberts, S. W. (1966). A Comparison of Some Control Chart Procedures. *Technometrics*, 8, 411-430.

Teoh, W.L., Khoo, M.B.C., Castagliola, P., Yeong, W.C. & Teh, S.Y. (2016). Run-Sum Control Charts for Monitoring the Coefficient of Variation. *European Journal of Operational Research*, 257, Pages 144-158

Thaga, K. & Sivasamy, R. (2015). Single Variables Control Charts: A Further Overview. *Indian Journal of Science and Technology*, Vol 8(6), 518-528.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 33ου Πανελληνίου Συνεδρίου Στατιστικής (2021), σελ.65-81

ΕΥΦΥΗ ΣΥΣΤΗΜΑΤΑ ΓΙΑ ΤΗΝ ΠΡΟΒΛΕΨΗ ΒΙΟΛΟΓΙΚΩΝ ΜΕΤΑΒΛΗΤΩΝ: ΕΦΑΡΜΟΓΗ ΣΕ ΠΡΩΤΟΓΕΝΗ ΣΤΟΙΧΕΙΑ ΔΕΝΤΡΩΝ ΠΕΥΚΗΣ

Διαμαντοπούλου Ι. Μαρία
Τμήμα Δασολογίας και Φυσικού Περιβάλλοντος, Εργαστήριο Δασικής Βιομετρίας, Α.Π.Θ., ΤΚ-54124, Θεσσαλονίκη, e-mail: mdiamant@for.auth.gr

ΠΕΡΙΛΗΨΗ

Το πρόβλημα εύρεσης της κατάλληλης σχέσης εκτίμησης δύσκολα μετρούμενων βιολογικών μεταβλητών των οποίων οι τιμές διαμορφώνονται από πολλούς και ανεξέλεγκτους παράγοντες, αποτελεί πεδίο εντατικής έρευνας στη δασολογική επιστήμη. Στην εργασία αυτή διερευνάται η δυνατότητα της εφαρμογής ευφυών συστημάτων, προκειμένου να εκτιμηθεί το μέγεθος των διαμέτρων κορμών ιστάμενων δέντρων σε οποιοδήποτε ύψος, λαμβάνοντας ως δεδομένα μεταβλητές που μπορούν εύκολα να μετρηθούν στο πεδίο. Η αποτελεσματικότητα των μεθόδων διερευνάται, συγκρίνεται και αξιολογείται, με σκοπό την πρόταση της βέλτιστης μεθόδου μέσω της οποίας θα επιτευχθεί η εκτίμηση αξιόπιστης πληροφορίας, αξιοποιήσιμη στον υπολογισμό του ξυλώδους όγκου των δέντρων, κερδίζοντας χρόνο, κόστος και κόπο στις μετρήσεις πεδίου.

Λέξεις Κλειδιά: Logistic μοντέλο, Levenberg-Marquardt Artificial Neural Network μοντέλα,Support Vector Machine μοντέλα.

1. ΕΙΣΑΓΩΓΗ

Ο ποσοτικός προσδιορισμός της ανάπτυξης του κορμού του δέντρου, αποτελεί πεδίο εντατικής έρευνας στη δασολογική επιστήμη, γιατί μέσω αυτού του προσδιορισμού, δίνεται η δυνατότητα εύκολης και κατά το δυνατό ακριβούς ογκομέτρησης του ιστάμενου κορμού των δέντρων, με μεθόδους τμηματικής ογκομέτρησης. Η ανάπτυξη του κορμού μπορεί να προσδιοριστεί μέσω του μεγέθους των διαμέτρων του σε διάφορα ύψη από το έδαφος. Η εύρεση της κατάλληλης σχέσης εκτίμησης δύσκολα μετρούμενων βιολογικών μεταβλητών των οποίων οι τιμές διαμορφώνονται από πολλούς και ανεξέλεγκτους παράγοντες, όπως το κλιματεδαφικό περιβάλλον και η βιολογία του ίδιου του οργανισμού, αποτελεί κλειδί στην ακριβή εκτίμηση της ξυλώδους βιομάζας. Προς αυτή τη κατεύθυνση, η γνώση του μεγέθους των διαμέτρων ενός ιστάμενου δέντρου σε διάφορα ύψη από το έδαφος είναι απαραίτητη, γιατί συμβάλλει άμεσα και ουσιαστικά στην ακριβέστερη ογκομέτρηση του κορμού, η οποία εξαρτάται από τον αριθμό των γνωστών διαμέτρων σε διάφορα ύψη αυτού και έμμεσα συμβάλλοντας στην περιγραφή της δομής των συστάδων (Μάτης 2004,

West 2009). Γενικότερα, η γνώση της διάστασης των διαμέτρων σε διάφορα ύψη των ιστάμενων δέντρων, χωρίς να είναι απαραίτητη η υλοτόμησή τους, αποτελεί σημαντική πληροφορία για την ορθολογική διαχείριση των δασικών οικοσυστημάτων.

Σήμερα, υπάρχουν διαθέσιμες πολλές μεθοδολογίες και τεχνικές κατάρτισης μοντέλων εκτίμησης. Η περισσότερο διαδεδομένη και χρησιμοποιούμενη μεθοδολογία είναι αυτή της παλινδρόμησης (Draper και Smith 1998, Ratkowsky 1990) μέσω της οποίας καταρτίζονται πολύ καλά μοντέλα εκτίμησης με μικρά σχετικά σφάλματα. Η δυσκολία η οποία πρέπει να αντιμετωπιστεί στη διαχείριση των βιολογικών δεδομένων, είναι η προσέγγιση των προϋποθέσεων εφαρμογής της θεωρίας της παλινδρόμησης με ικανοποιητική ακρίβεια (Διαμαντοπούλου και Σταματέλλος 2013). Σε διαφορετική περίπτωση, προκειμένου να εξαχθεί ένα αξιόπιστο μοντέλο παλινδρόμησης, θα πρέπει να αναγνωριστούν και να αντιμετωπιστούν προβλήματα τα οποία προκύπτουν, συνηθέστερα εκ των οποίων είναι: α) η ασταθής εκτίμηση των συντελεστών παλινδρόμησης, β) λανθασμένες αποφάσεις ελέγχων υποθέσεων, γ) λανθασμένα πρόσημα των συντελεστών παλινδρόμησης, δ) μεροληπτική επιλογή μεταβλητών του μοντέλου, κλπ (Ratkowsky 1990, Belsley 1991, Draper και Smith 1998, Chatterjee κ.α. 2000). Η παραβίαση των προϋποθέσεων αυτών αποτελεί συχνό φαινόμενο σε βιολογικά δεδομένα, όπως είναι τα δεδομένα που προέρχονται από μετρήσεις σε δέντρα, στο δασικό περιβάλλον, αποτελώντας σοβαρό εμπόδιο στην εύρεση ενός στατιστικά αξιόπιστου και ταυτόχρονα ακριβούς μοντέλου εκτίμησης. Επιπρόσθετα, η προσπάθεια εύρεσης της κατάλληλης μορφής μοντέλου παλινδρόμησης το οποίο μπορεί να περιγράψει τα πρωτογενή δεδομένα, αποτελεί μια δύσκολη και χρονοβόρα απαίτηση, η οποία όμως πρέπει να αντιμετωπιστεί επιτυχώς. Γι' αυτούς τους λόγους, τελευταία, η επιστημονική έρευνα στο δασικό επιστημονικό πεδίο, έχει επικεντρωθεί στην εφαρμογή νέων μεθόδων μοντελοποίησης, όπως των ευφυών συστημάτων τεχνητής νοημοσύνης (Artificial Intelligence, AI) και συγκριτικής αξιολόγησής τους με τις περισσότερο κλασικές μεθόδους μοντελοποίησης οι οποίες χρησιμοποιήθηκαν ευρέως και χρησιμοποιούνται και σήμερα, όπως πχ. η θεωρία της παλινδρόμησης, προκειμένου να διαπιστωθεί η χρησιμότητά τους στην επίλυση προβλημάτων της δασικής έρευνας (Youquan κ.α. 2013, Διαμαντοπούλου και Σταματέλλος 2013, Bayat κ.α. 2020). Μεταξύ των τεχνικών ευφυών συστημάτων που μπορούν να χρησιμοποιηθούν για κατάρτιση αξιόπιστων μοντέλων εκτίμησης είναι και αυτή των τεχνητών νευρωνικών δικτύων (ANNs) καθώς και των μοντέλων «υποστηρικτικής διανυσματικής παλινδρόμησης» (SVMs). Επιπλέον, στην εξόρυξη δεδομένων, ο αλγόριθμος του τυχαίου δάσους (Random Forest regression, RFr) (Breiman 2001, Segal 2003, Prasad κ.α. 2006, Cluter κ.α. 2011), μπορεί να χρησιμοποιηθεί για τη λύση προβλημάτων εκτίμησης των τιμών συνεχών μεταβλητών, όπως είναι η τυχαία μεταβλητή της διαμέτρου των δέντρων.

Οι τεχνικές αυτές ευφυών συστημάτων διερευνώνται τα τελευταία χρόνια σε μεγάλο εύρος δασικών προβλημάτων, σχετικά με τη δυνατότητά τους να ανακαλύψουν τις σχέσεις που μπορεί να συνδέουν βιολογικές μεταβλητές έτσι ώστε να αποτελέσουν

αξιόπιστη λύση σε δασικά, και όχι μόνο, προβλήματα (Diamantopoulou 2005, Diamantopoulou κ.α. 2009, Youquan κ.α. 2012, Aschonitis 2017, Diamantopoulou κ.α. 2018, Özçelik κ.α. 2019, Bayat κ.α. 2020).

Σκοπός της εργασίας αυτής είναι αφενός μεν η κατάρτιση μοντέλων: α) μη-γραμμικής παλινδρόμησης (NLR), β) νευρωνικών δικτύων (ANN), γ) υποστηρικτικής διανυσματικής παλινδρόμησης (SVMr) και δ) τυχαίου δάσους (RFr), για την εύρεση αξιόπιστης σχέσης εκτίμησης των τιμών των διαμέτρων σε διάφορα ύψη του κορμού των δέντρων Πεύκης, αφετέρου δε, η συγκριτική αξιολόγησή τους, προκειμένου να δειχθεί η καταλληλότερη μέθοδος εκτίμησης για τα πρωτογενή δεδομένα τα οποία αναλύονται.

2. ΜΕΘΟΔΟΣ ΕΡΕΥΝΑΣ

2.1 Δεδομένα

Από το περιαστικό δάσος Θεσσαλονίκης έκτασης 3.018,84 ha, πάρθηκε συστηματικό δείγμα μεγέθους n=94 δέντρων Πεύκης (Pinus brutia). Επί του κορμού των δέντρων αυτών μετρήθηκαν η διάμετρος σε ύψος 30 εκατοστά από το έδαφος, (πρεμνική διάμετρος, $d_{0.3}$) και η διάμετρος σε ύψος 1,3 μέτρα από το έδαφος (στηθιαία διάμετρος, $d_{1.3}$) με παχύμετρο, οι διάμετροι (d_i) ανά ένα μέτρο πάνω από τη στηθιαία διάμετρο μέχρι το ολικό ύψος κάθε δέντρου με ρελασκόπιο και το ολικό ύψος (h_{total}) κάθε δέντρου του δείγματος, με το υψόμετρο Blume-Leiss (Παράρτημα, Πίνακας Α.). Μετά την ολοκλήρωση των μετρήσεων προέκυψε δείγμα μεγέθους n=445 γραμμών δεδομένων. Για τη διερεύνηση και κατάρτιση του καταλληλότερου μοντέλου, στην περίπτωση μοντελοποίησης μέσω των τεχνικών των ευφυών συστημάτων, το δείγμα των n=445 γραμμών δεδομένων χωρίστηκε κάνοντας χρήση τυχαίων αριθμών σε δύο διακριτά μέρη: α) στο δείγμα των δεδομένων κατάρτισης του κατάλληλου μοντέλου (fitting data set) το οποίο αποτελεί το 90% των συνολικών γραμμών δεδομένων (n1=401) και β) στο δείγμα των δεδομένων επαλήθευσης του μοντέλου (test data set) που καταρτίστηκε, τα οποία αποτελούνται από τις υπόλοιπες 10% γραμμές δεδομένων (n2=44). Τα δεδομένα επαλήθευσης δεν χρησιμοποιήθηκαν σε κανένα σημείο της διαδικασίας κατάρτισης των μοντέλων ευφυών συστημάτων. Στην περίπτωση της κατάρτισης των μοντέλων μη γραμμικής παλινδρόμησης, δεν χρησιμοποιήθηκε αυτός ο διαχωρισμός γιατί δεν επιδρά στην κατάρτιση του μοντέλου παλινδρόμησης (Hursch 1991).

2.2 Μοντέλο μη-γραμμικής παλινδρόμησης (non-linear regression model, NLR)

Πριν την κατάρτιση του κατάλληλου μοντέλου παλινδρόμησης το οποίο θα έχει τη δυνατότητα να εκτιμά την έμφλοια διάμετρο σε οποιοδήποτε ύψος του ιστάμενου κορμού, εφαρμόστηκε διερευνητική ανάλυση των πρωτογενών δεδομένων με τη χρήση του στατιστικού πακέτου IBM-SPSS (IBM-SPSS 19, 2016), προκειμένου να διερευνηθούν οι βασικότερες προϋποθέσεις εφαρμογής της παλινδρόμησης (κανονική κατανομή της εξαρτημένης μεταβλητής και ομοιογένεια διακύμανσης). Έγινε προσαρμογή ενός μεγάλου αριθμού εξισώσεων πολλαπλής και μη-γραμμικής

παλινδρόμησης στα δεδομένα, με τη χρήση του στατιστικού με εξαρτημένη μεταβλητή την ($d_i/d_{1.3}$) και ανεξάρτητες μεταβλητές την πρεμνική διάμετρο ($d_{0.3}$), το ολικό ύψος (h_{total}) και το ύψος εκείνο στο οποίο χρειάζεται η εκτίμηση της έμφλοιας διαμέτρου (h_{di}). Ο αλγόριθμος του Levenberg-Marquardt. χρησιμοποιήθηκε για την κατάρτιση των μη-γραμμικών εξισώσεων. Μεταξύ των μοντέλων μη-γραμμικής παλινδρόμησης τα οποία διερευνήθηκαν, το logistic (1) μοντέλο (Ratkowsky 1990), έδωσε την καλύτερη προσαρμογή στα δεδομένα:

$$\frac{d_i}{d_{1.3}} = \frac{1}{1+e^{\left[-(\theta_1+\theta_2\cdot d_{0.3}+\theta_3\cdot h_{total}+\theta_4\cdot h_{di})\right]}} \qquad (1)$$

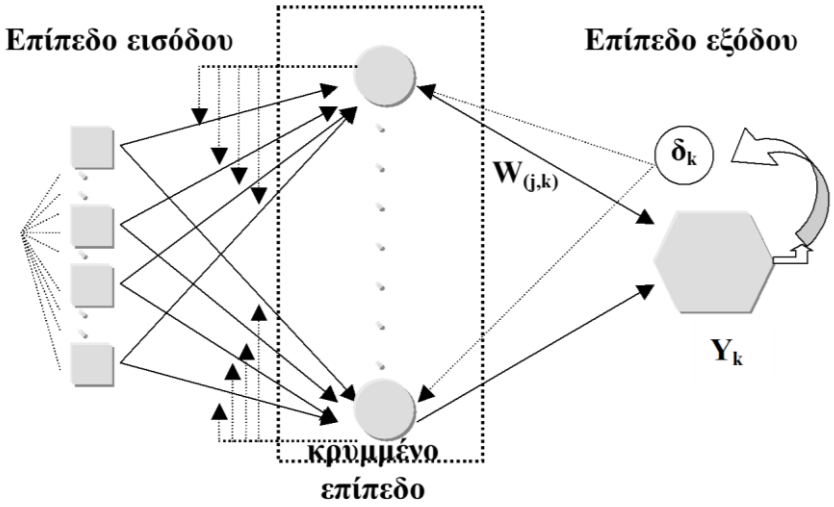
όπου $\theta_{i=1,2,34}$ είναι οι συντελεστές παλινδρόμησης.

Εξαιτίας της ανομοιογένειας της διακύμανσης που εμφανίζουν οι διάμετροι, χρησιμοποιήθηκε στάθμιση των μοντέλων παλινδρόμησης τα οποία διερευνήθηκαν.

2.3 Μοντέλο νευρωνικού δικτύου (Artificial Neural Network model, ANN)

Για την κατάρτιση του μοντέλου νευρωνικού δικτύου, δηλαδή για την εύρεση των κατάλληλων τιμών των βαρών ($w_{j,k}$) του δικτύου, χρησιμοποιήθηκε η δομή της πολυστρωματικής αντίληψης-νόησης (Multilayer perceptron, MLP) (Εικόνα 1).

Εικόνα 1. *Αρχιτεκτονική δομή της πολυστρωματικής αντίληψης-νόησης*



Ένα μοντέλο νευρωνικού δικτύου εποπτευόμενης μάθησης εκπαιδεύεται χρησιμοποιώντας κόμβους (nodes). Στο επίπεδο εισαγωγής βρίσκονται οι μεταβλητές εισόδου και στο επίπεδο εξόδου βρίσκεται η μεταβλητή εξόδου. Οι μεταβλητές οι οποίες χρησιμοποιήθηκαν στα δύο αυτά επίπεδα είναι ίδιες με τις ανεξάρτητες μεταβλητές και την εξαρτημένη μεταβλητή, οι οποίες χρησιμοποιήθηκαν για την κατάρτιση του μοντέλου μη-γραμμικής παλινδρόμησης. Ενδιάμεσα αυτών, υπάρχει

το κρυμμένο επίπεδο, όπου εκεί ο κατάλληλος αριθμός των κόμβων που χρησιμοποιήθηκαν, διερευνήθηκε αρχίζοντας από 1 κόμβο μέχρι 10. Το μοντέλο νευρωνικού δικτύου το οποίο καταρτίστηκε επιλέχθηκε να είναι αυτό το οποίο χρησιμοποιεί τον αλγόριθμο εκπαίδευσης του Levenberg-Marquardt (LMANN), ως ένας ενδιάμεσος αλγόριθμος μεταξύ του Gauss-Newton, ενός γρήγορου αλγορίθμου και του gradient descent, ενός σταθερού αλγορίθμου, σε μια προσπάθεια να «εκμεταλλευτούμε» τα πλεονεκτήματα και των δύο αλγορίθμων. Το μοντέλο καταρτίστηκε κάνοντας χρήση του software της Matlab (Beale κ.α. 2014). Η αποτελεσματικότητα και η σύγκλιση της εκπαίδευσης του LMANN μοντέλου εξαρτάται σημαντικά από την ορθή επιλογή του συντελεστή μ (combination coefficient, μ). Γι' αυτό το λόγο, κατά την κατάρτιση του LMANN μοντέλου διερευνήθηκε η κατάλληλη τιμή του μ. Δηλαδή, δοκιμάστηκαν τιμές με έναρξη από μ = 0.01 και χρησιμοποιώντας έναν συντελεστή προσαρμογής ν = 10, ο οποίος πολλαπλασιάζονταν ή διαιρούνταν με τον συντελεστή μ, μέχρι να καταρτιστεί μοντέλο με το μικρότερο άθροισμα τετραγωνικών σφαλμάτων, το οποίο οριοθετούσε τον τερματισμό της εκπαίδευσης. Λεπτομερής θεωρητική ανάπτυξη της δομής και λειτουργίας των μοντέλων νευρωνικών δικτύων μπορεί να βρεθεί στην πολυπληθή διεθνή βιβλιογραφία (Gurney, 1999, Haykin, 2009)

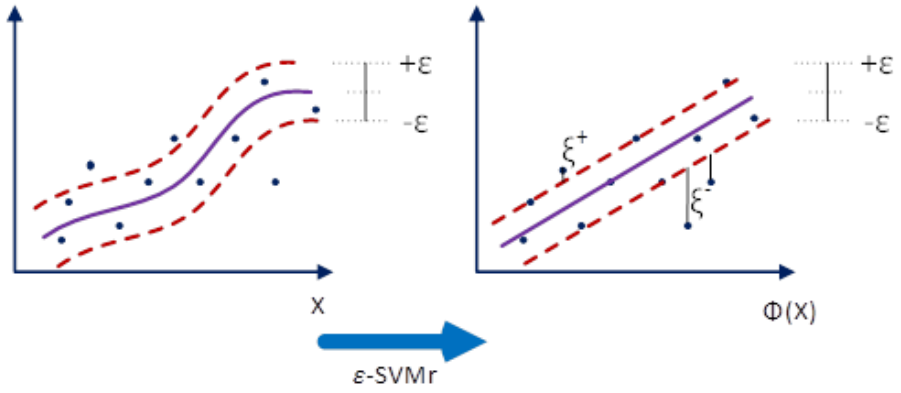
2.4 Μοντέλο υποστηρικτικής διανυσματικής παλινδρόμησης (Support Vector Machine model, SVMr)

Σε γενικές γραμμές, κατά την εφαρμογή της τεχνικής SVMr αρχικά δημιουργείται ένας χώρος πλάτους 2ε, με ε>0, έτσι ώστε τα αρχικά δεδομένα να βρίσκονται εντός του διαστήματος [-ε,+ε]. Στη συνέχεια χρησιμοποιούνται συναρτήσεις kernel προκειμένου να προβληθούν τα δεδομένα σε έναν υπερχώρο m-διαστάσεων όπου οι πολύπλοκες μη γραμμικές σχέσεις μεταξύ των πρωτογενών δεδομένων να είναι δυνατό να αναπαρασταθούν απλά μέσω της βέλτιστης ευθείας γραμμής (Williams 2011) (Εικόνα 2). Ο βέλτιστος προσανατολισμός της ευθείας επιτυγχάνεται με ελαχιστοποίηση της αντικειμενικής συνάρτησης του προβλήματος βελτιστοποίησης η οποία συμπεριλαμβάνει τη διορθωτική μεταβλητή (ξ_i) η οποία χρησιμοποιείται προκειμένου να συμπεριληφθούν όλα τα σημεία δεδομένων εντός του υπερχώρου m-διαστάσεων.

Η ικανότητα του συστήματος των υποστηρικτικών διανυσμάτων τα οποία δημιουργούνται προκειμένου το SVMr μοντέλο να έχει την δυνατότητα της ακριβούς εκτίμησης της εξαρτημένης μεταβλητής, ελέγχεται από τρεις μετα-παραμέτρους, την (C), της οποίας η τιμή καθορίζει την πολυπλοκότητα του συστήματος σε σχέση με την ακρίβεια που επιτυγχάνεται, τη (γ) η οποία καθορίζει την Ευκλείδια απόσταση μεταξύ των υποστηρικτικών διανυσμάτων (support vectors) και την (ε), της οποίας η τιμή καθορίζει το πλάτος του χώρου ο οποίος χρησιμοποιείται για την εκπαίδευση του SVMr μοντέλου, έτσι ώστε να διασφαλίζεται ότι η επίλυση της αντικειμενικής συνάρτησης βελτιστοποίησης οδηγεί σε γενικό ελάχιστο σφάλμα. Οι βέλτιστες τιμές των τριών μετα-παραμέτρων προσδιορίστηκαν μέσω της διαδικασίας grid search (Kavzoglu and Colkesen 2009) και με εφαρμογή της τεχνικής k-fold διασταυρωμένης

επικύρωσης (k-fold cross validation), με k=10. Τέλος, ο τύπος της συνάρτησης kernel ο οποίος χρησιμοποιήθηκε ήταν αυτός της ακτινικής βάσης πυρήνα (Radial Basis Function, RBF). Η εκμάθηση του μοντέλου υποστηρικτικής διανυσματικής παλινδρόμησης προγραμματίστηκε σε γλώσσα προγραμματισμού Python (Van Rossum και Drake 2011, Python Software Foundation), με χρήση των βιβλιοθηκών της scikit-learn (Pedregosa κ.α. 2011). Λεπτομερής θεωρητική ανάπτυξη της δομής και λειτουργίας των μοντέλων υποστηρικτικής διανυσματικής παλινδρόμησης μπορεί να βρεθεί στην πολυπληθή διεθνή βιβλιογραφία (Olson and Delen, 2008).

Εικόνα 2. *Εφαρμογή τεχνικής SVMr*



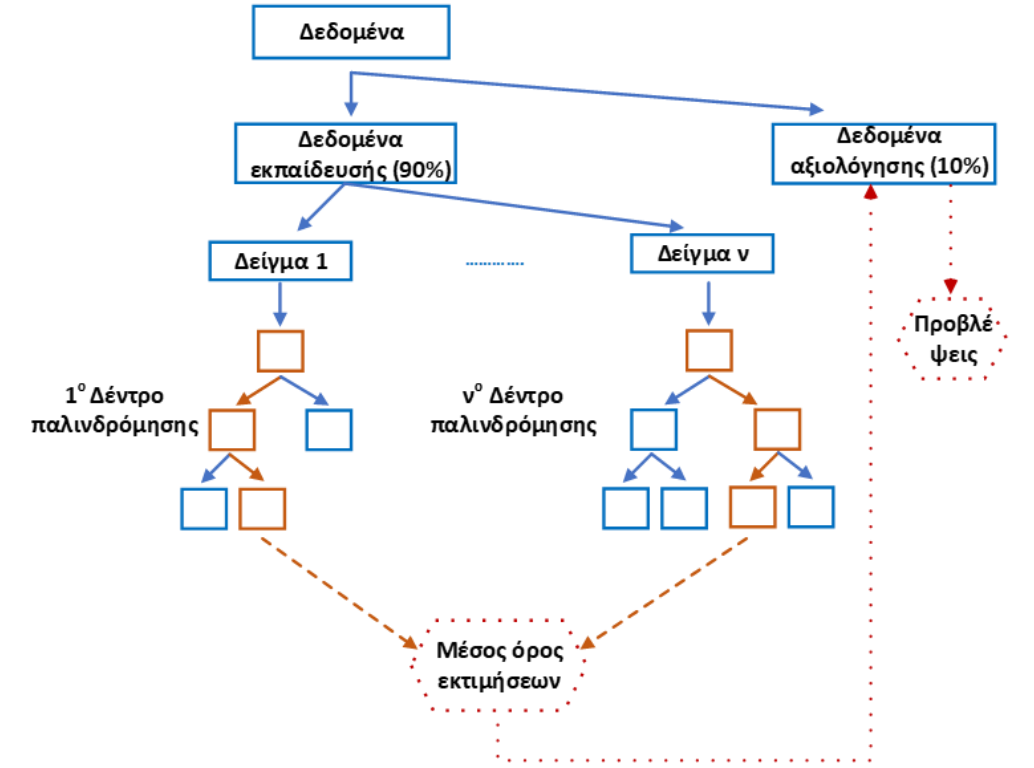
2.5 Μοντέλο τυχαίου δάσους (Random Forest regression model, RFr)

Πρόκειται για έναν εποπτευόμενο αλγόριθμο μηχανικής μάθησης. Ο Random Forest regression, RFr αλγόριθμος βασίζεται και χρησιμοποιεί περισσότερα από ένα δέντρα απόφασης (decision trees), γι' αυτό έχει επικρατήσει η ορολογία «δάσος» (Εικόνα 3).

Τα δέντρα αποφάσεων στα οποία η μεταβλητή στόχος είναι συνεχής, αποτελούν τα δέντρα παλινδρόμησης. Το τυχαίο δάσος αποτελεί σύνολο ή συλλογή των δέντρων παλινδρόμησης των οποίων την πληροφορία χρησιμοποιεί. Κατά τη διαδικασία εκπαίδευσης, οι παρατηρήσεις που περιλαμβάνονται στο σετ δεδομένων εκπαίδευσης χρησιμοποιούνται με στόχο τη δημιουργία μεγάλου πλήθους δέντρων παλινδρόμησης, καθένα από τα οποία έχει διαφορετικές παραμέτρους εκπαίδευσης και συνεπώς συμμετέχει με διαφορετικό τρόπο στη διαδικασία πρόβλεψης. Η τελική πρόβλεψη για μια παρατήρηση προέρχεται από το συνδυασμό όλων των επιμέρους προβλέψεων, γεγονός που σε συνδυασμό με τα διαφορετικά εσωτερικά χαρακτηριστικά του κάθε δέντρου, ενισχύει τη δυνατότητα γενίκευσης. Αποτελεί ένα είδος εκμάθησης διάρθρωσης αποφάσεων, βασιζόμενο σε μοντέλο πρόβλεψης, προκειμένου με δεδομένες τις τιμές των παρατηρήσεων των ανεξάρτητων μεταβλητών να είναι σε θέση να εκτιμήσει τις αντίστοιχες τιμές της εξαρτημένης μεταβλητής. Οι εκτιμήσεις της εξαρτημένης μεταβλητής εξάγονται ως μέσος όρος των αποτελεσμάτων των επιμέρους δέντρων. Κάθε επιμέρους δέντρο παλινδρόμησης

αποτελείται από ένα συνδεδεμένο διάγραμμα ροής. Σ΄αυτό, υπάρχει ένας μοναδικός αρχικός κόμβος από τον οποίο ξεκινούν αρχικά δύο ακμές (κλαδιά) και καταλήγουν σε κόμβους «παιδιά» τα οποία προέρχονται από τους γονικούς κόμβους. Για κάθε κόμβο υπάρχει μια συνθήκη ικανοποίησης. Εφόσον αυτός ο στόχος δεν επιτευχθεί, προχωράει η διαδικασία σε νέο κόμβο και νέα παιδιά (Εικόνα 3).

Εικόνα 3. *Διάρθρωση τυχαίου δάσους*



Ο αλγόριθμος του τυχαίου δάσους παρουσιάζει ένα σημαντικό πλεονέκτημα: αποφεύγει την υπερπαραμετροποίηση (overfitting) του μοντέλου. Το βασικό μειονέκτημα της μεθόδου είναι ότι ο σχεδιασμός του έχει γίνει κατά τέτοιο τρόπο ώστε να μην έχει τη δυνατότητα να προβλέπει τιμές πέραν των ορίων των μεταβλητών με τις οποίες έχει εκπαιδευτεί. Η εκμάθηση του μοντέλου τυχαίου δάσους προγραμματίστηκε σε γλώσσα προγραμματισμού Python (Van Rossum και Drake 2011, Python Software Foundation), με χρήση των βιβλιοθηκών της scikit-learn (Pedregosa κ.α. 2011). Λεπτομερής θεωρητική ανάπτυξη της δομής και λειτουργίας των μοντέλων τυχαίου δάσους μπορεί να βρεθεί στην πολυπληθή διεθνή βιβλιογραφία (Breiman 2001, Cluter κ.α. 2011).

2.6 Κριτήρια αξιολόγησης των μοντέλων

Προκειμένου να αξιολογηθούν τα μοντέλα ευφυών συστημάτων τα οποία καταρτίστηκαν για την εκτίμηση των διαμέτρων του κορμού δέντρων σε οποιοδήποτε ύψος από το έδαφος, αλλά και για τη συγκριτική αξιολόγηση των τεχνικών οι οποίες εφαρμόστηκαν, μεταξύ τους, χρησιμοποιήθηκαν τα στατιστικά μέτρα: 1) ο συντελεστής συσχέτισης (R) μεταξύ των πραγματικών τιμών και των αντίστοιχων τιμών του μοντέλου, 2) το μέγιστο απόλυτο σφάλμα (MaxAE) μεταξύ των πραγματικών τιμών και των αντίστοιχων τιμών του μοντέλου, 3) το τυπικό μέσο τετραγωνικό σφάλμα (RMSE) μεταξύ των πραγματικών τιμών και των εκτιμώμενων τιμών από το μοντέλο και 4) ο δείκτης FI του Furnival, προκειμένου να είναι εφικτή η σύγκριση των μοντέλων ευφυών συστημάτων με το καλύτερα προσαρμοζόμενο στα δεδομένα μοντέλο μη-γραμμική παλινδρόμησης. Αποτελεί έναν πολύ σημαντικό δείκτη γιατί μπορεί να χρησιμοποιηθεί σε περιπτώσεις που υπάρχει αναγκαιότητα σύγκρισης της προσαρμογής εξισώσεων οι οποίες δεν έχουν την ίδια έκφραση της εξαρτημένης μεταβλητής. Μετασχηματίζει το τυπικό σφάλμα εκτίμησης σε μονάδες της εξαρτημένης μεταβλητής. Ο δείκτης FI υπολογίζεται ως FI = RMSE·(Παράγοντας), όπου ο Παράγοντας ισούται με τον αντιλογάριθμο του μέσου της αντίστροφης πρώτης παραγώγου της εξαρτημένης μεταβλητής της εξίσωσης, ως προς το h_{total}. Όσο η τιμή του δείκτη αυτού είναι μικρότερη, τόσο καλύτερα προσαρμοσμένο είναι το μοντέλο στα δεδομένα.

3. ΑΠΟΤΕΛΕΣΜΑΤΑ ΚΑΙ ΣΥΖΗΤΗΣΗ

Η γνώση των διαμέτρων ενός ιστάμενου κορμού σε οποιοδήποτε ύψος από το έδαφος, αποτελεί μια επίπονη διαδικασία στο πεδίο. Η γνώση των τιμών αυτών των διαμέτρων με ακρίβεια, δίνει τη δυνατότητα όχι μόνο της ακριβούς τμηματικής ογκομέτρησης του κορμού, αλλά και την διεξαγωγή συμπερασμάτων για τη δομή των συστάδων με αποτέλεσμα την ορθολογική διαχείρισή τους. Γι΄αυτό το λόγο, προκειμένου να διερευνηθούν διάφορες διαδικασίες μοντελοποίησης της διαμέτρου των δέντρων σε διάφορα ύψη, τόσο κλασικές όσο και ευφυών συστημάτων, πάρθηκε συστηματικό δείγμα μεγέθους n=94 δέντρων Πεύκης (Pinus brutia), από το περιαστικό δάσος Θεσσαλονίκης, έκτασης 3.018,84 ha, Τα περιγραφικά στατιστικά στοιχεία για την πρεμνική ($d_{0.3}$), τη στηθιαία διάμετρο ($d_{1.3}$) και το ολικό ύψος (h_{total}) των δέντρων των πρωτογενών δεδομένων, δίνονται στον Πίνακα 1.

Πίνακας 1. *Περιγραφικά στατιστικά στοιχεία των πρωτογενών δεδομένων*

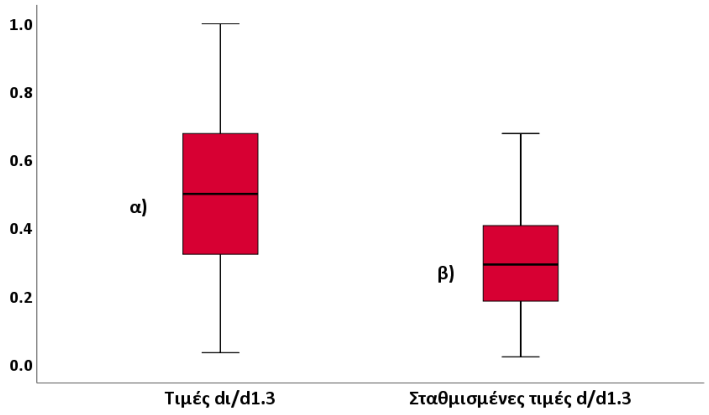
Μεταβλητή	Μέσος	Τυπικό σφάλμα μέσου	Μέγιστη τιμή	Ελάχιστη τιμή	Διακύμανση
$d_{0.3}$, εκ.	18,93	0,3045	39,0	9,0	41.25
$d_{1.3}$, εκ.	15,02	0,3191	38,5	6,0	45.31
h_{total}, μ.	6,95	0,0848	12,0	3,3	3.20

Για την κατάρτιση των εξισώσεων παλινδρόμησης, η ανομοιογένεια στη διακύμανση της εξαρτημένης μεταβλητής αντιμετωπίστηκε με στάθμιση. Η διερεύνηση του

καταλληλότερου σταθμικού έγινε με τη μέθοδο της μέγιστης πιθανοφάνειας. Το σταθμικό που προέκυψε με βάση τη διακύμανση των τιμών της διαμέτρου, μετά από διερεύνηση στο διάστημα τιμών [-3,3], ανά 0,1, ήταν $w_i = \frac{1}{(d_{1.3})^{0.2}}$.

Η διερευνητική ανάλυση των πρωτογενών δεδομένων έδειξε μικρή απόκλιση από την κανονικότητα. Από τα θηκογράμματα της Εικόνας 4, φαίνεται η διόρθωση της ανομοιογένειας της διακύμανσης της εξαρτημένης μεταβλητής μετά τη στάθμιση.

Εικόνα 4. *Θηκογράμματα εξαρτημένης μεταβλητής σε μη σταθμισμένες (α) και σε σταθμισμένες (β) τιμές.*



Από την εφαρμογή της πολλαπλής και μη-γραμμικής παλινδρόμησης, το μοντέλο που προσαρμόστηκε καλύτερα στα δεδομένα ήταν το logistic (εξ. 1) μοντέλο. Ο έλεγχος στατιστικής σημαντικότητας των συντελεστών παλινδρόμησης δίνεται στον Πίνακα 2 και έδειξε ότι όλοι οι συντελεστές παλινδρόμησης είναι στατιστικά σημαντικά διάφοροι του μηδενός.

Πίνακας 2. *Τιμές συντελεστών παλινδρόμησης και 95% ασυμπτωτικά διαστήματα εμπιστοσύνης (α=0,05), για το σταθμισμένο λογιστικό (logistic) μοντέλο (εξ. 1).*

Σταθμισμένο μοντέλο μη-γραμμικής παλινδρόμησης, με σταθμικό: $w_i = \frac{1}{(d_{1.3})^{0.2}}$

$$\frac{d_i}{d_{1.3}} \cdot w_i = \left(\frac{1}{1 + e^{[-(\theta_1 + \theta_2 \cdot d_{0.3} + \theta_3 \cdot h_{total} + \theta_4 \cdot h_{di})]}}\right) \cdot w_i$$

Συντελεστές παλινδρόμησης	Τιμή	Ασυμπτωτικό διάστημα εμπιστοσύνης	
		Κατώτερο όριο	Ανώτερο όριο
θ_1	1,184	0,982	1,387
θ_2	-0,024	-0,035	-0,014
θ_3	0,302	0,262	0,343
θ_4	-0,638	-0,676	-0,599

Προκειμένου να ελεγχθεί η συμπεριφορά πρόβλεψης των μοντέλων ευφυούς μάθησης (LMANN, SVMr και RFr) σε νέα δεδομένα, τα πρωτογενή δεδομένα

διαιρέθηκαν τυχαία σε δύο τμήματα: στο 90% των δεδομένων το οποίο χρησιμοποιήθηκε για την κατάρτιση του μοντέλου και στο υπόλοιπο 10% για την δοκιμή και επαλήθευση του εκπαιδευμένου μοντέλου σε νέα δεδομένα, τα οποία δεν χρησιμοποιήθηκαν σε καμία φάση της κατάρτισης του μοντέλων ευφυών συστημάτων.

Για την κατάρτιση του LMANN μοντέλου, για το 90% των δεδομένων κατάρτισης του μοντέλου, χρησιμοποιήθηκε η τεχνική k-fold διασταυρωμένης επικύρωσης (k-fold cross validation), με k=10, προκειμένου να διασφαλιστεί η χρήση διαδοχικά όλων των δεδομένων κατάρτισης (το 90% των αρχικών δεδομένων) αφενός για εκπαίδευση των μοντέλων (training), αφετέρου για επικύρωση (validation).

Η αποτελεσματική εκπαίδευση του LMANN μοντέλου εξαρτάται σημαντικά από την τελική τιμή του αποσβεστικού παράγοντα (μ). Αυτή η τελική τιμή προσδιορίστηκε μετά από δοκιμές, αρχίζοντας από την τιμή 0,01 και χρησιμοποιώντας το συντελεστή προσαρμογής v = 10, μέχρι να επιτευχθεί το μικρότερο σφάλμα θεωρητικών τιμών του μοντέλου. Η βέλτιστη τιμή του αποσβεστικού παράγοντα (μ) βρέθηκε ίση με 0,00001. Το μοντέλο LMANN καταρτίστηκε μετά από 502 επαναλήψεις (epochs). Ο βέλτιστος αριθμός των αριθμός των κόμβων του κρυμμένου επιπέδου επιλέχθηκε με βάση την τεχνική δοκιμασίας-λάθους (trial and error), δοκιμάζοντας από 1 έως 10 κόμβους. Ο βέλτιστος αριθμός των κόμβων ο οποίος ελαχιστοποιούσε το μέσο τετραγωνικό σφάλμα του μοντέλου ήταν ίσος με 7. Τέλος η συνάρτηση ενεργοποίησης που χρησιμοποιήθηκε ήταν η μη-γραμμική εξίσωση μεταφοράς tanh(s) (Fausett 1994).

Για την κατάρτιση του SVMr μοντέλου, χρησιμοποιήθηκε η τεχνική k-fold διασταυρωμένης επικύρωσης (k-fold cross validation), με k=10. Χρησιμοποιώντας τη συνάρτηση kernel RBF (radial basis function), ο καλύτερος συνδυασμός των τριών μετα-παραμέτρων διερευνήθηκε και βρέθηκε μετά από δοκιμή συνδυασμών των τιμών τους, μέσω της τεχνικής grid search. Για τη διερεύνηση αυτή χρησιμοποιήθηκαν τιμές της μετα-παραμέτρου (ε) στο εύρος από 0.001 μέχρι 0.1 ανά 0.01, τιμές της (C) στο εύρος από 1 μέχρι 100 ανά 1 και τιμές της (γ) στο εύρος από 0.1 μέχρι 1 ανά 0.01. Κάθε συνδυασμός των παραπάνω τιμών ελέγχθηκε όσον αφορά την προσαρμογή του μοντέλου στα δεδομένα κατάρτισης, μέσω της τιμής του μέσου τετραγωνικού σφάλματος. Τελικά, ο καλύτερος συνδυασμός των τριών μετα-παραμέτρων ήταν C=25, e=0,01 και γ=0,21.

Η εκπαίδευση του μοντέλου τυχαίου δάσους έγινε με τη χρήση 100 δέντρων παλινδρόμησης των οποίων ο αριθμός αποφασίστηκε μετά από δοκιμές αριθμών δέντρων από 2 έως και 150. Διαπιστώθηκε ότι μετά τη χρήση των 100 δέντρων δεν υπήρξε σημαντική βελτίωση του μέσου σφάλματος εκτίμησης θεωρητικών τιμών από το μοντέλο. Επίσης, επιλέχθηκε η bootstrap μέθοδος, όσον αφορά την επιλογή των δεδομένων για κάθε δέντρο παλινδρόμησης, έτσι ώστε να τυχαιοποιηθεί με επανάθεση πλήρως η επιλογή των δεδομένων κατάρτισης των δέντρων απόφασης. Τέλος, επιλέχθηκε παρέμβαση στο βάθος των κλάδων των δέντρων, στους 10 κλάδους, προκειμένου να μην υπάρχει υπερπαραμετροποίηση της εκμάθησης του

δέντρου απόφασης, μετά από επαναληπτική διαδικασία εκπαίδευσης του μοντέλου.

Τα στατιστικά κριτήρια αξιολόγησης για το συνολικό δείγμα των 445 γραμμών, για τα μοντέλα τα οποία καταρτίστηκαν, δίνονται στον Πίνακα 3. Ειδικότερα, για το σταθμισμένο μοντέλο της μη-γραμμικής παλινδρόμησης, χρησιμοποιήθηκε ο δείκτης (I) του Furnival προκειμένου να μετασχηματιστεί το σφάλμα των σταθμισμένων εξισώσεων (μετασχηματισμένες τιμές σφάλματος) οπότε να είναι συγκρίσιμο με το αντίστοιχα σφάλματα των υπολοίπων μοντέλων.

Πίνακας 3. *Κριτήρια αξιολόγησης των μοντέλων*

Μοντέλο	Εξαρτημένη μεταβλητή	R	MaxAE	RMSE	Παράγοντας	Δείκτης Furnival
NLR	$d_i/d_{1.3} \cdot w_i$	0,8928	0,3404	0,0611	1,6921	0,1034
LMANN	$d_i/d_{1.3}$	0,9328	0,2925	0,0824	1	0,0824
SVMr	$d_i/d_{1.3}$	0,9551	0,3169	0,0681	1	0,0643
RFr	$d_i/d_{1.3}$	0,9696	0,2790	0,0561	1	0,0561

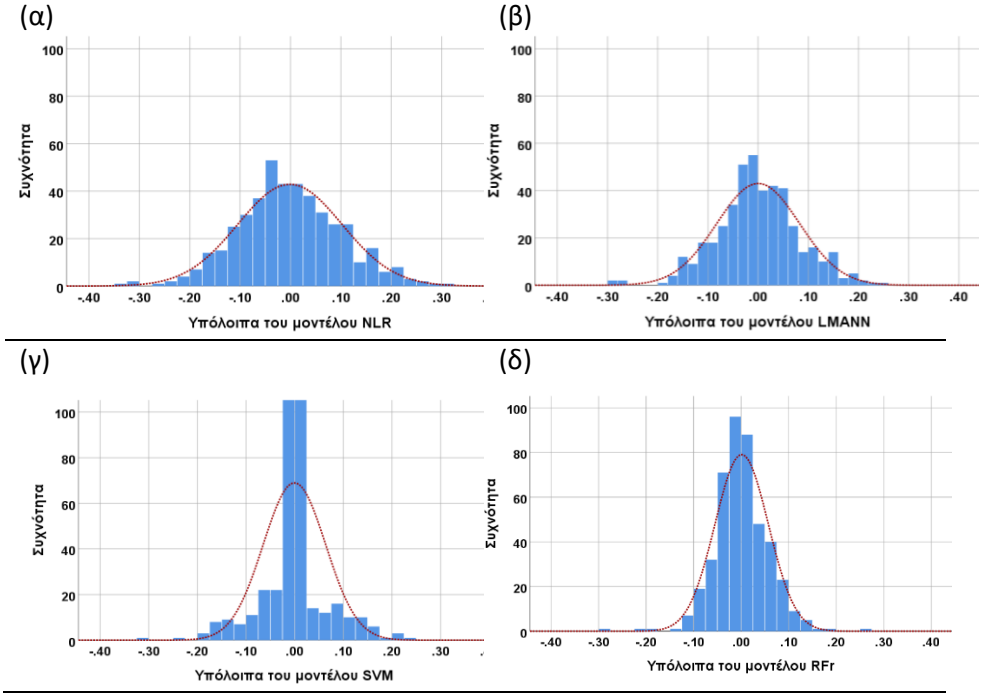
Όπως φαίνεται στον Πίνακα 3, όλα τα στατιστικά μέτρα αξιολόγησης έχουν καλύτερες τιμές για το RFr μοντέλο.

Την αμέσως καλύτερη προσαρμογή στα δεδομένα του δείγματος δίνει το μοντέλο SMVr, ακολουθεί το μοντέλο νευρωνικού δικτύου LMANN, και στη συνέχεια ακολουθεί η εξίσωση που αναφέρεται στη μη-γραμμική παλινδρόμηση. Συγκεκριμένα, το τυπικό σφάλμα εκτίμησης θεωρητικών τιμών (RMSE) για το RFr μοντέλο, είναι κατά 54,25% μικρότερο από το αντίστοιχο σφάλμα του logistic (εξ. 1) μοντέλου, 12.75% μικρότερο από το αντίστοιχο σφάλμα του SVMr μοντέλου, και 31,92% μικρότερο από το αντίστοιχο σφάλμα του LMANN μοντέλου.

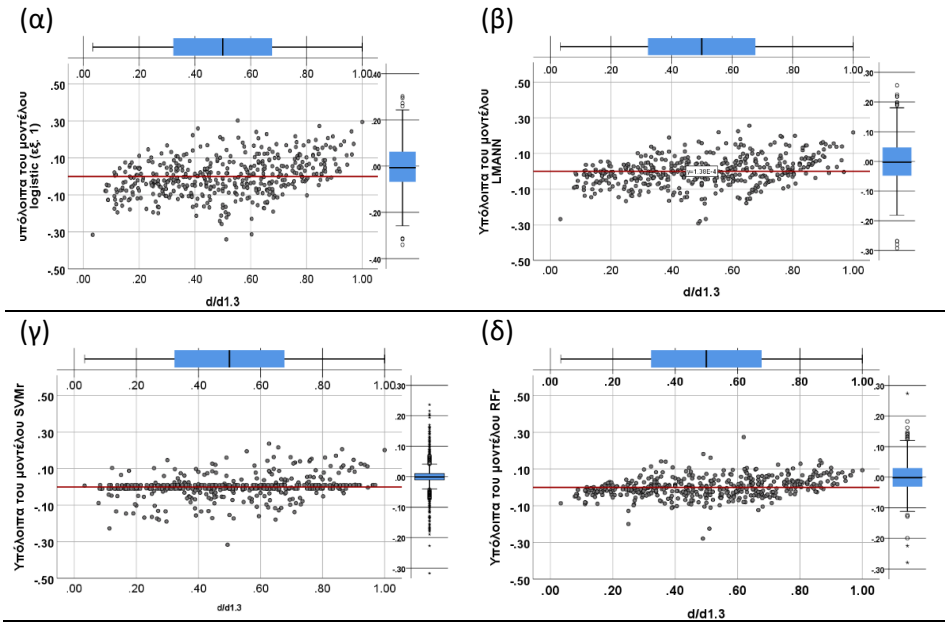
Ακολούθησε ανάλυση υπολοίπων των τεσσάρων μοντέλων, από την οποία προέκυψαν υπόλοιπα χωρίς σημαντικές αποκλίσεις από την κανονική κατανομή (Εικόνα 5), με μικρή και ομοιογενή διασπορά (Εικόνα 6), γεγονός που υποστηρίζει την στατιστικά ορθή κατασκευή των μοντέλων.

Τα παραπάνω εκπαιδευμένα μοντέλα ευφυών συστημάτων δοκιμάστηκαν σε νέο δείγμα δεδομένων με n=44 γραμμές. Τα περιγραφικά στατιστικά για τις μεταβλητές εισόδου του δείγματος κατάρτισης των μοντέλων ευφυών συστημάτων και του νέου δείγματος, δίνονται στον Πίνακα 4.

Εικόνα 5. *Ιστογράμματα υπολοίπων (α, β, γ, και δ) για τα μοντέλα NLR, LMANN, SVMr και RFr, αντίστοιχα.*



Εικόνα 6. *Στικτά διαγράμματα και θηκογραφήματα των υπολοίπων (α,β, γ, και δ) για τα μοντέλα NLR, LMANN, SVMr και RFr, αντίστοιχα.*



Πίνακας 4. *Περιγραφικά στατιστικά στοιχεία του δείγματος κατάρτισης $n_κ$=401 και του νέου δείγματος δεδομένων με $n_δ$ =44 γραμμές*

	Δεδομένα δείγματος κατάρτισης $n_κ$=401			Δεδομένα δείγματος δοκιμής $n_δ$=44		
Μεταβλητή	Μέσος	Μέγιστη τιμή	Ελάχιστη τιμή	Μέσος	Μέγιστη τιμή	Ελάχιστη τιμή
$d_{0.3}$, εκ.	19,06	39,00	9,00	17,78	39,00	10,00
h_{di}, εκ.	4,41	9,30	2,30	4,34	7,30	2,30
h_{total}, μ.	6,95	12,00	3,30	6,86	11,50	4,30
$d_i/d_{1.3}$	0,50	1,00	0,03	0,50	0,86	0,12

Τα κριτήρια αξιολόγησης για τις προβλέψεις των μοντέλων ευφυών συστημάτων για τα στοιχεία του δείγματος κατάρτισης $n_κ$=401 και του νέου δείγματος δεδομένων με $n_δ$ =44, δίνονται στον παρακάτω Πίνακα 5.

Αν γινόταν εκτίμηση μόνο των δεδομένων του δείγματος των 44 γραμμών από την καταρτισμένη εξίσωση μη-γραμμικής παλινδρόμησης, ο δείκτης I του Furnival θα ήταν ίσος 0,1023. Δηλαδή, η εξίσωση μη-γραμμικης παλινδρόμησης εκτιμά με σφάλμα κατά 21,01% μεγαλύτερο τις διαμέτρους των δεδομένων δοκιμής συγκρινόμενο με το σφάλμα πρόβλεψης του μοντέλου τυχαίου δάσους, το οποίο καταρτίστηκε χωρίς να χρησιμοποιήσει σε καμία φάση κατάρτισής του τα δεδομένα αυτά.

Πίνακας 5. *Κριτήρια αξιολόγησης των μοντέλων ευφυών συστημάτων για τα στοιχεία του δείγματος κατάρτισης $n_κ$=401 και του νέου δείγματος δεδομένων με $n_δ$ =44 γραμμές*

Δεδομένα δείγματος κατάρτισης $n_κ$=401						
Μοντέλο	Εξαρτημένη μεταβλητή	R	MaxAE	RMSE	Παράγοντας	Δείκτης Furnival
LMANN	$d_i/d_{1.3}$	0,9358	0,1893	0,0813	1	0,0813
SVMr	$d_i/d_{1.3}$	0,9585	0,3169	0,0653	1	0,0653
RFr	$d_i/d_{1.3}$	0,9736	0,2790	0,0526	1	0,0526
Δεδομένα δείγματος δοκιμής $n_δ$=44						
Μοντέλο	Εξαρτημένη μεταβλητή	R	MaxAE	RMSE	Παράγοντας	Δείκτης Furnival
LMANN	$d_i/d_{1.3}$	0,9062	0,2925	0,0916	1	0,0916
SVMr	$d_i/d_{1.3}$	0,9374	0,2325	0,0893	1	0,0893
RFr	$d_i/d_{1.3}$	0,9280	0,1994	0,0808	1	0,0808

Από τα δεδομένα του Πίνακα 5, προκύπτει ότι όλα τα εκπαιδευμένα μοντέλα έχουν τη δυνατότητα να εκτιμούν με αποδεκτή ακρίβεια τις διαμέτρους κορμών Πεύκης σε οποιοδήποτε ύψος του κορμού. Η κατάταξη των εκπαιδευμένων μοντέλων ευφυών συστημάτων, σύμφωνα με το σφάλμα πρόβλεψης των διαμέτρων κορμού σε νέο σετ δεδομένων δέντρων από το ίδιο δάσος, αρχίζοντας από το μικρότερο σφάλμα, είναι RFr μοντέλο με την υψηλότερη ακρίβεια πρόβλεψης, SVMr και LMANN μοντέλο.

4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην εργασία αυτή έγινε εφαρμογή των τεχνικών των ευφυών συστημάτων τεχνητής νοημοσύνης, αυτών α) των νευρωνικών δικτύων, με τη χρήση της αρχιτεκτονικής της πολυστρωματικής αντίληψης-νόησης (Multilayer perceptron, MLP) και με εφαρμογή του αλγορίθμου εκπαίδευσης του Levenberg-Marquardt, β) της τεχνικής της υποστηρικτικής διανυσματικής παλινδρόμησης SVMr και γ) της μεθοδολογίας του τυχαίου δάσους για κατάρτιση μοντέλων τύπου παλινδρόμησης (RFr), ως εναλλακτικές διαδικασίες μοντελοποίησης σχετικά με την κλασική εφαρμογή της μεθόδου της μη-γραμμικής παλινδρόμησης.

Η αναγκαιότητα της εφαρμογής μιας μεθόδου ευφυούς συστήματος προκύπτει από το γεγονός των δυσκολιών που πρέπει να αντιμετωπιστούν κατά την εφαρμογή της μεθόδου της παλινδρόμησης. Βασική τροχοπέδη αποτελεί η απαίτηση των στατιστικών προϋποθέσεων όσον αφορά τα πρωτογενή δεδομένα (κανονικότητα, σταθερή διακύμανση τιμών, κλπ). Προκειμένου να αντιμετωπιστούν οι αποκλίσεις από τις προϋποθέσεις της παλινδρόμησης είναι αναγκαία η εφαρμογή διορθωτικών μεθόδων, οι οποίες απαιτούν χρόνο και προσπάθεια. Η προηγούμενη προϋπόθεση, σε συνδυασμό με την απαίτηση της γνώσης της κατάλληλης μορφής της εξίσωσης, η οποία θα έχει τη δυνατότητα να περιγράψει με αξιοπιστία τα πρωτογενή δεδομένα, καθιστούν την εφαρμογή μεθόδων παλινδρόμησης δύσκολη και χρονοβόρα. Ειδικότερα, για την εφαρμογή της μη-γραμμικής παλινδρόμησης, είναι απαραίτητη η γνώση αρχικών τιμών από τον ερευνητή, έτσι ώστε να έχει τη δυνατότητα να συγκλίνει το μοντέλο σε γενικά ελάχιστα και να μην παγιδεύεται σε τοπικές λύσεις. Παρόλα αυτά, η εφαρμογή της μεθόδου της παλινδρόμησης και ως προς τη φάση δημιουργίας του μοντέλου, αλλά και ως προς τη χρησιμοποίηση του καταρτισμένου μοντέλου, είναι γνωστή και κατανοητή και δεν απαιτεί ιδιαίτερη ικανότητα προγραμματισμού από το χρήστη. Από την άλλη πλευρά, τα ευφυή συστήματα, απαιτούν βασικές προγραμματιστικές ικανότητες, ενώ έχουν τη δυνατότητα να χρησιμοποιούν τα πρωτογενή δεδομένα χωρίς καμιά προϋπόθεση και δεν απαιτούν από τον ερευνητή να γνωρίζει τη μορφή του μοντέλου το οποίο μπορεί με αξιοπιστία να περιγράψει τα δεδομένα. Το ίδιο το σύστημα το δημιουργεί. Είναι απαραίτητη όμως η επιλογή των κατάλληλων τιμών των παραμέτρων βάση των οποίων γίνεται η εκπαίδευση ενός μοντέλου ευφυούς μάθησης προκειμένου αυτό να προσαρμοστεί με ακρίβεια στα δεδομένα. Αυτή η επιλογή γίνεται με μεθόδους επαναληπτικής διαδικασίας εκμάθησης του μοντέλου με διαφορετικές παραμέτρους και τελική επιλογή της ακριβέστερης μάθησης. Συγκεκριμένα, για την εκπαίδευση του μοντέλου νευρωνικού δικτύου απαιτήθηκε αρχικά η επιλογή της κατάλληλης αρχιτεκτονικής και του αλγορίθμου εκπαίδευσης και στη συνέχεια για το LMANN μοντέλο απαιτήθηκε η ορθή επιλογή του αποσβεστικού παράγοντα (μ), του αριθμού των κρυμμένων κόμβων του κρυμμένου επιπέδου και η επιλογή της συνάρτησης μεταφοράς. του τυχαίου δάσους, απαιτήθηκε η επιλογή του κατάλληλου αριθμού δέντρων απόφασης και ο προσδιορισμός της πολυπλοκότητας του καθενός από αυτά τα δέντρα. Για την εκπαίδευση του μοντέλου υποστηρικτικής διανυσματικής παλινδρόμησης απαιτήθηκε αρχικά ο τύπος της συνάρτησης βελτιστοποίησης και η

ορθή επιλογή του συνδυασμού των μετα-παραμέτρων (C, ε και γ), ενώ για την κατάρτιση του μοντέλου τυχαίου δάσους απαιτήθηκε η ορθή επιλογή του αριθμού των δέντρων απόφασης καθώς και το βάθος των κλάδων των δέντρων.

Η ικανότητα των μοντέλων ευφυούς μάθησης να εκτιμούν με μεγαλύτερη ακρίβεια, συγκρινόμενα με το μοντέλο μη-γραμμικής παλινδρόμησης, τις διαμέτρους σε οποιοδήποτε ύψος του κορμού, έδειξε ότι οι τεχνικές αυτές ευφυούς μάθησης που διερευνήθηκαν στην παρούσα εργασία, μπορούν να εφαρμοστούν με επιτυχία σε δασικά δεδομένα και συγκεκριμένα μπορούν να χρησιμοποιηθούν με ασφάλεια για την εκτίμηση διαμέτρων κορμού δέντρων σε διάφορα ύψη, σε δέντρα Πεύκης με εύρος τιμών κορμών παρόμοιο με αυτό των δεδομένων κατάρτισης των μοντέλων. Ειδικότερα το μοντέλο τυχαίου δάσους, έδωσε τα ακριβέστερα αποτελέσματα. Αν και το συγκεκριμένο μοντέλο δεν έχει τη μορφή ενός συμβατικού μοντέλου, η εφαρμογή του στα δασικά πρωτογενή δεδομένα δέντρων Πεύκης από το περιαστικό δάσος Θεσσαλονίκης έδωσε αποτελέσματα τα οποία οδηγούν στο συλλογισμό ότι θα προσέφερε μια εναλλακτική αξιόπιστη λύση στο πρόβλημα ακριβούς εκτίμησης των διαμέτρων σε οποιοδήποτε ύψος του κορμού των ιστάμενων δέντρων γενικότερα.

ABSTRACT

This paper explores the possibility of applying the Artificial Neural Network, the Support Vector Machine for regression, and the Random Forest regression methodologies, as possible alternatives to non-linear regression models, in order to assess as accurately as possible the size of the diameters of the tree trunks at any height above the ground, taking into account data that can be easily measured in the field, since the difficulty of locating and measuring the tree trunk diameters at heights far from the ground is a serious problem in the field measurements, that need to be addressed. The effectiveness of the intelligence systems' models, that fitted to the available data, is compared with the results of the best fitted non-linear regression model to our data in hand and evaluated. This investigation has shown that each one of the models can be used for the tree trunk diameter estimation. Furthermore, the RFr model showed superior adaptation to our data in hand as compared with the other investigated models and can be considered as a reliable alternative methodology in order to achieve the accuracy of the information provided, saving time and effort in field.

ΑΝΑΦΟΡΕΣ

Aschonitis V., Diamantopoulou M. and Papamichail D. (2017). Modeling plant density and ponding water effects on flooded rice evapotranspiration and crop coefficients: critical discussion about the concepts used in current methods. *Theoretical and Applied Climatology*, **132**, 1165–1186.

Bayat M., Bettinger P., Heidari S., Henareh Khalyani A., Jourgholami M. and Hamidi S.K. (2020). Estimation of tree heights in an uneven-aged, mixed forest in northern Iran using artificial intelligence and empirical models. *Forests*, **11**, 324.

Beale M., Hagan M. and Demuth H. (2014). Neural network ToolboxTM User's Guide, R2014a. Natick, MA: The MathWorks Inc.

Belsley D.A. (1991). Conditioning diagnostics: Collinearity and Weak Data in Regression. Wiley-Interscience, New York, pp. 396.

Breiman L. (2001). Random Forests. *Machine Learning*, **45**, 5–32.
Chatterjee S., Hadi A.S. and Price B. (2000). Regression analysis by example. 3rd Ed. J. Willy and Sons, Inc., New York, pp. 601.
Cluter A, Cluter D.R. and Stevens J.R. (2011). Random Forests. *Machine Learning*, **45**, 157-176.
Diamantopoulou M. (2005). Artificial neural networks as an alternative tool in pine bark volume estimation. *Computers and Electronics in Agriculture,* **48,** 235-244.
Diamantopoulou M.J., Milios E., Doganos D., and Bistinas I. (2009). Artificial Neural Network Modeling For Reforestation Design Through The Dominant Trees Bole-Volume Estimation. *Natural Resource Modeling*, **22**, 511-543.
Διαμαντοπούλου Μ. και Σταματέλλος, Γ. (2013). Εφαρμογή νευρωνικών δικτύων στην εκτίμηση του αριθμού κορμών σε δασικές εκτάσεις. Πρακτ. 16ου Παν. Δασ. Συν. & Annual Meeting Pro Silva Europe. "Προστασία – Διαχείριση των Ελληνικών Δασών σε περίοδο οικονομικής κρίσης και η πρόκληση της Φυσικής Δασοπονίας", Θεσσαλονίκη, 6-9 (13): 388-395.
Diamantopoulou M.J. Özçelik R. and Yavuz H. (2018). Tree-bark volume prediction via machine learning: a case study based on black alder's tree-bark production. *Computers and Electronics in Agriculture*, **151**, 431–440.
Draper N.R. and Smith H. (1998). Applied Regression Analysis. Wiley, N.Y. pp. 706.
Fausett L. (1994). Fundamentals of neural networks architectures. Algorithms and Applications. NJ: Prentice Hall, Englewood Cliffs.
Gurney K. (1999). An introduction to neural networks. Prentice Hall, UK.
Haykin S. (2009). Neural networks and Learning Machines, 3rd ed. Prentice Hall, UK.
Hursch R. (1991). Validation samples. *Biometrics*, **47**, 1193–1194.
IBM-SPSS 19. (2016). Guide to Data Analysis by Marija Norusis, Inc. SPSS
Kavzoglu T. and Colkesen I. (2009). A kernel functions analysis for support vector machines for land cover classification. *International Journal of Applied Earth Observation and Geoinformation*, **11**, 352-359.
Μάτης Κ.Γ. (2004). Δασική Βιομετρία ΙΙ. Δενδρομετρία. Εκδ. Πήγασος 2000, Θεσσαλονίκη. σελ 674.
Olson D. and Delen D. (2008). Advanced Data Mining Techniques. Springer-Verlag Berlin Heidelberg, 180 pp.
Özçelik R., Diamantopoulou M.J. and Trincado G. (2019). Evaluation of potential modeling approaches for Scots pine stem diameter prediction in north-eastern Turkey. *Computers and Electronics in Agriculture*, **162**, 773-782.
Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M. and Duchesnay E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825-2830.
Prasad A.M. Iverson L.R. and Liaw A. (2006). Newer Classification and Regression Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems*, **9**, 181—199.
Python Software Foundation. Python Language Reference, version 3.9. Available at http://www.python.org

Ratkowsky D.A. (1990). Handbook of nonlinear regression models. Statistics: Textbooks and Monographs, vol. 107. Marcel Dekker Inc., N.Y. pp. 241.

Segal M.R. (2003). Machine Learning Benchmarks and Random Forest Regression. UCSF: Center for Bioinformatics and Molecular Biostatistics. Retrieved from https://escholarship.org/uc/item /35x3v9t4.

Van Rossum G. and Drake F.L. (2011). The Python Language Reference Manual. Network Theory Ltd. pp.150.

West P.W. (2009). Tree and Forest Measurement. 2nd ed. Springer-Verlag, Berlin.

Williams G. (2011). Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery, use R. Springer Science+Business Media, LLC, DOI 10.1007/9781441998-2.

Youquan J., Lixi Z., Ou D., Weiheng X. and Zhongke F. (2013). Calculation of live tree timber volume based on particle swarm optimization and support vector regression. *Transactions of the Chinese Society of Agricultural Engineering*, **29**, 160–167.

ΠΑΡΑΡΤΗΜΑ

Πίνακας Α. *Περιγραφή πρωτογενών δεδομένων*

Συμβολισμός μεταβλητής	Έννοια μεταβλητής	Μονάδες μέτρησης	Ύψος μέτρησης πάνω στον κορμό, από το έδαφος, σε μ.	Μετρητικό όργανο
$d_{0.3}$	διάμετρος (πρεμνική)	εκατοστά	0,3	παχύμερο
$d_{1.3}$	διάμετρος (στηθιαία)	εκατοστά	1,3	παχυμετρο
$d_{2.3}$	διάμετρος	εκατοστά	2,3	ρελασκόπιο[1]
$d_{3.3}$	διάμετρος	εκατοστά	3,3	ρελασκόπιο[1]
$d_{4.3}$	διάμετρος	εκατοστά	4,3	ρελασκόπιο[1]
$d_{5.3}$	διάμετρος	εκατοστά	5,3	ρελασκόπιο[1]
$d_{6.3}$	διάμετρος	εκατοστά	6,3	ρελασκόπιο[1]
$d_{7.3}$	διάμετρος	εκατοστά	7,3	ρελασκόπιο[1]
$d_{8.3}$	διάμετρος	εκατοστά	8,3	ρελασκόπιο[1]
$d_{9.3}$	διάμετρος	εκατοστά	9,3	ρελασκόπιο[1]
$d_{10.3}$	διάμετρος	εκατοστά	10,3	ρελασκόπιο[1]
$d_{11.3}$	διάμετρος	εκατοστά	11,3	ρελασκόπιο[1]
h_{total}	ύψος	μέτρα	Ύψος κορυφής	υψόμετρο[2]

[1]ρελασκόπιο του Bitterlich (Spiegel Relaskop), [2]υψόμετρο Blume-Leiss

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 33ου Πανελληνίου Συνεδρίου Στατιστικής (2021), σελ.82-92

ΔΙΑΤΑΚΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ ΜΕ ΤΗ ΧΡΗΣΗ ΡΟΩΝ

Χ.Ε. Ζώτος[1], Σ.Δ. Δαφνής[1], Γ.Κ. Παπαδόπουλος[1]
[1]Τμήμα Επιστήμης Φυτικής Παραγωγής, Γεωπονικό Πανεπιστήμιο Αθηνών
{czotos, sdafnis, gpapadop}@aua.gr

ΠΕΡΙΛΗΨΗ

Τις τελευταίες δεκαετίες μεγάλος αριθμός προβλημάτων από διάφορες ερευνητικές περιοχές μοντελοποιείται με την κατηγοριοποίηση των πειραματικών δεδομένων σε δύο κατηγορίες, τη θεώρηση δίτιμων δοκιμών (με τιμές 0 ή 1) και τη μελέτη των ακολουθιών των αποτελεσμάτων. Μεταξύ των δυνατών ακολουθιών, ιδιαίτερο ενδιαφέρον παρουσιάζουν οι ροές μονάδων μήκους k $(k \geq 1)$. Στην παρούσα εργασία προτείνεται για πρώτη φορά ένα μοντέλο διατακτικής παλινδρόμησης με χρήση ροών. Το μοντέλο εφαρμόζεται σε ένα τρέχον πρόβλημα σχετιζόμενο με τη βιολογία του εντόμου *Marchalina hellenica* και τη μελισσοκομία και αποδεικνύεται ότι ο συνδυασμός των ροών και της διατακτικής παλινδρόμησης μπορεί να προσφέρει στους ειδικούς εξαιρετικά χρήσιμα αποτελέσματα.

Λέξεις Κλειδιά: Δίτιμες δοκιμές, τρόποι μέτρησης ροών, θερμοκρασία, *Marchalina hellenica*

1. ΕΙΣΑΓΩΓΗ

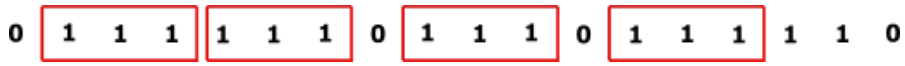
Τις τελευταίες δεκαετίες μεγάλος αριθμός προβλημάτων από διάφορες ερευνητικές περιοχές μοντελοποιείται με την κατηγοριοποίηση των πειραματικών δεδομένων σε δύο κατηγορίες, τη θεώρηση δίτιμων δοκιμών (με τιμές 0 ή 1) και τη μελέτη των ακολουθιών των αποτελεσμάτων. Μεταξύ των ακολουθιών, ιδιαίτερο ενδιαφέρον παρουσιάζουν οι ροές μονάδων μήκους k $(k \geq 1)$.

Η θεωρία των ροών εφαρμόζεται εκτεταμένα, μεταξύ άλλων, στην Αξιοπιστία Συστημάτων, στη Βιολογία, στις Γεωπονικές Επιστήμες, στον Έλεγχο Ποιότητας, στη Μετεωρολογία, στα Χρηματοοικονομικά και αλλού. Για μία ανασκόπηση των εφαρμογών της θεωρίας των ροών παραπέμπουμε στους Balakrishnan & Koutras, 2002 και Dafnis & Makri, 2021.

Στη βιβλιογραφία συναντιούνται αρκετοί, διαφορετικοί μεταξύ τους, τρόποι καταμέτρησης ροών. Η ποικιλομορφία αυτή οφείλεται στα διαφορετικά χαρακτηριστικά των εφαρμογών προς μελέτη. Στην παρούσα εργασία θα μας απασχολήσει ο τρόπος καταμέτρησης ροών που ονομάζεται τύπου Ι (Feller, 1968, Hirano, 1986, Philippou & Makri, 1986, Balakrishnan & Koutras, 2002). Σύμφωνα με τον τρόπο καταμέτρησης τύπου Ι, μία ροή μήκους k καταγράφεται όταν παρατηρηθούν k διαδοχικά 1 και το $(k+1)$ −οστό διαδοχικό 1 καταμετράται ως το πρώτο 1 της επόμενης ροής.

Έστω ότι σε $n = 18$ επαναλήψεις έχει παρατηρηθεί η ακολουθία 011111101110111110. Στο Σχήμα 1 παρουσιάζονται οι 4 ροές μήκους $k = 3$, οι οποίες καταμετρούνται με τον προαναφερόμενο τρόπο καταμέτρησης ροών.

Σχήμα 1. *Καταμέτρηση ροών τύπου Ι μήκους 3*



Στην παρούσα εργασία η εφαρμοσιμότητα των ροών θα αναδειχθεί μέσω του μοντέλου της διατακτικής παλινδρόμησης. Πιο συγκεκριμένα, είναι η πρώτη φορά που ο αριθμός ροών θα χρησιμοποιηθεί ως ανεξάρτητη μεταβλητή σε ένα τέτοιο μοντέλο και με τον τρόπο αυτό θα αναδειχθεί η ερμηνευτική του ικανότητα. Η διαδικασία αυτή θα πραγματοποιηθεί σε ένα τρέχον πρόβλημα που αφορά στη μελισσοκομία. Φυσικά, η προτεινόμενη μεθοδολογία μπορεί να προσαρμοσθεί κατάλληλα και στις ερευνητικές περιοχές που εφαρμόζεται η θεωρία των ροών. Για την καλύτερη παρουσίαση των αποτελεσμάτων (Ενότητα 3), ακολουθεί (Ενότητα 2) μία συνοπτική παρουσίαση του μοντέλου διατακτικής παλινδρόμησης που θα χρησιμοποιήσουμε, αυτού των Αναλογικών Συμπληρωματικών Πιθανοτήτων.

2. ΜΟΝΤΕΛΟ ΤΩΝ ΑΝΑΛΟΓΙΚΩΝ ΣΥΜΠΛΗΡΩΜΑΤΙΚΩΝ ΠΙΘΑΝΟΤΗΤΩΝ

Τα μοντέλα διατακτικής παλινδρόμησης χρησιμοποιούνται όταν η μεταβλητή απόκρισης είναι διατακτική με περισσότερες από δύο κατηγορίες/επίπεδα. Στην παρούσα εργασία χρησιμοποιείται το μοντέλο των Αναλογικών Συμπληρωματικών Πιθανοτήτων (Proportional Odds Model – POM, McCullagh, 1980)

$$\begin{aligned} Y_j = Pr(Y \le j|\mathbf{x}) &= \pi_1(\mathbf{x}) + \pi_2(\mathbf{x}) + \cdots + \pi_k(\mathbf{x}) \\ &= 1/(1 + e^{-(a_j + \boldsymbol{\beta}'\mathbf{x})}), \qquad j = 1, \dots, k-1, \end{aligned} \tag{1}$$

όπου, Y η μεταβλητή απόκρισης με k διατακτικές κατηγορίες/επίπεδα (εξαρτημένη μεταβλητή), $\mathbf{x}$ διάνυσμα ανεξάρτητων μεταβλητών, $\pi_1(\mathbf{x}), \pi_2(\mathbf{x}), \dots, \pi_k(\mathbf{x})$ οι πιθανότητες των διατακτικών κατηγοριών της εξαρτημένης μεταβλητής για την τιμή $\mathbf{x}$ των ανεξάρτητων μεταβλητών, a_j ο σταθερός όρος για κάθε κατηγορία j και $\boldsymbol{\beta}$ το διάνυσμα των συντελεστών παλινδρόμησης (με $\boldsymbol{\beta}'$ συμβολίζουμε τον ανάστροφο του πίνακα $\boldsymbol{\beta}$).

Σε μορφή logit το μοντέλο γίνεται

$$\begin{aligned} logit(Y_j) = \ln(Y_j/(1-Y_j)) &= \ln(Pr(Y \le j|\mathbf{x})/Pr(Y > j|\mathbf{x})) \\ &= a_j + \boldsymbol{\beta}'\mathbf{x}, \quad j = 1, \dots, k-1. \end{aligned} \tag{2}$$

Στο POM το διάνυσμα των συντελεστών παλινδρόμησης β δεν εξαρτάται από το j και επομένως η σχέση μεταξύ Y και x είναι ανεξάρτητη του j. Επίσης, ο σχετικός λόγος των συμπληρωματικών πιθανοτήτων (Odds Ratio – OR), για τιμές $\mathbf{x} = \mathbf{x_1}$ και $\mathbf{x} = \mathbf{x_2}$ του διανύσματος των ανεξάρτητων μεταβλητών $\mathbf{x}$,

$$OR = \frac{Pr(Y \le j|\mathbf{x_1})/Pr(Y > j|\mathbf{x_1})}{Pr(Y \le j|\mathbf{x_2})/Pr(Y > j|\mathbf{x_2})}$$

$$= e^{\boldsymbol{\beta}'(\mathbf{x_1}-\mathbf{x_2})}, \quad (3)$$

είναι ανεξάρτητος του j και εξαρτάται μόνο από τη διαφορά $\mathbf{x_1} - \mathbf{x_2}$ (Agresti, 2013). Στην παρούσα εργασία, το διάνυσμα των ανεξάρτητων μεταβλητών $\mathbf{x}$ αποτελείται από μια μόνο μεταβλητή, την X, η οποία εκφράζει τον αριθμό των ροών των κρύων ημερών το μήνα Φεβρουάριο. Έτσι, το διάνυσμα των συντελεστών παλινδρόμησης $\boldsymbol{\beta}$ αποτελείται από ένα μόνο συντελεστή, το β_1 και το μοντέλο γίνεται

$$Y_j = Pr(Y \leq j|x) = 1/(1 + e^{-(a_j+\beta_1 x)}) \quad (1')$$

και αντίστοιχα στη logit μορφή

$$logit(Y_j) = \ln(Y_j/(1 - Y_j)) = \ln(Pr(Y \leq j|x)/Pr(Y > j|x)) = a_j + \beta_1 x. \quad (2')$$

3. ΔΙΑΤΑΚΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ ΜΕ ΧΡΗΣΗ ΡΟΩΝ ΣΤΗ ΜΕΛΙΣΣΟΚΟΜΙΑ

Οι Gounari, et al. (2021) χρησιμοποίησαν ένα μοντέλο διατακτικής παλινδρόμησης για να προβλεφθεί το δεκαπενθήμερο ολοκλήρωσης του βιολογικού κύκλου του μελιτογόνου εντόμου *Marchalina hellenica* (*M. hellenica)* βάσει του αριθμού των κρύων ημερών του Φεβρουαρίου. Ως όριο για το χαρακτηρισμό της ημέρας ως κρύα θεωρήθηκε η μέγιστη ημερήσια θερμοκρασία των $7.3°C$.

Αυτή η δυνατότητα πρόβλεψης έχει ιδιαίτερη αξία για της μελισσοκόμους, καθώς οδηγεί σε ορθότερους χειρισμούς και αυξημένη παραγωγή μελιού (Gounari et al. (2021)). Στην παρούσα εργασία θα δώσουμε συμπληρωματικά κριτήρια πρόβλεψης. Πιο συγκεκριμένα η πρόβλεψη θα στηριχθεί της ροές κρύων ημερών του Φεβρουαρίου μήκους $k\ (k \geq 1)$. Είναι προφανές ότι για $k = 1$ καταμετρούνται συνολικά οι κρύες ημέρες, επομένως τα αποτελέσματα της παρούσας εργασίας ανάγονται, στην ειδική περίπτωση αυτή, στα αντίστοιχα αποτελέσματα των Gounari, et al. (2021). Στην παρούσα εργασία θα παρουσιάσουμε τα αποτελέσματα για ροές κρύων ημερών μήκους $k = 2$ και $k = 3$ και επιπλέον θα αναδειχθεί ότι η ερμηνευτική δυνατότητα των ροών ως ανεξάρτητη μεταβλητή επιτρέπει και την επιπλέον χαλάρωση του ανώτατου ορίου μέγιστης θερμοκρασίας για να χαρακτηριστεί μία ημέρα του Φεβρουαρίου ως κρύα (το όριο θα αυξηθεί στους $8°C$ και στους $9°C$). Τα αποτελέσματα της, λοιπόν, θα αφορούν τις εξής 4 περιπτώσεις: Περίπτωση Ι, k =2, θ=$8°C$, Περίπτωση ΙΙ, k =2, θ=$9°C$, Περίπτωση ΙΙΙ, k =3, θ=$8°C$, Περίπτωση IV, k =3, θ=$9°C$.

Στον Πίνακα 1 παρατίθεται η συχνότητα (απόλυτη, σχετική και αθροιστική) ολοκλήρωσης του βιολογικού κύκλου του υπό εξέταση εντόμου ανά δεκαπενθήμερο αναφοράς στις δύο υπό μελέτη περιοχές που βρίσκονται στο νησί της Ρόδου, στις Καλυθιές και στον Έμπωνα.

Πίνακας 1. *Συχνότητα (απόλυτη, σχετική και αθροιστική) ολοκλήρωσης του βιολογικού κύκλου του M. Hellenica ανά δεκαπενθήμερο αναφοράς*

	Κωδικοποίηση	Συχνότητα	Σχετική Συχνότητα	Αθροιστική Συχνότητα
Δεκαπενθήμερα Αναφοράς	1 = 2ο δεκαπενθήμερο Μαρτίου	1	9.1%	9.1%
	2 = 1ο δεκαπενθήμερο Απριλίου	4	36.3%	45.4%
	3 = 2ο δεκαπενθήμερο Απριλίου	2	18.2%	63.6%
	4 = 1ο δεκαπενθήμερο Μαΐου	2	18.2%	81.8%
	5 = 2ο δεκαπενθήμερο Μαΐου	2	18.2%	100%
Έγκυρα		11	100%	
Ελλείποντα		1		
Σύνολο		12		

Στον Πίνακα 2 παρουσιάζεται αναλυτικά η περιγραφή της εξαρτημένης μεταβλητής (Y) και της εκάστοτε ανεξάρτητης μεταβλητής (X) που θα χρησιμοποιηθούν στο POM, έτσι όπως αυτό παρουσιάστηκε στην Ενότητα 2. Εν συνεχεία, στον Πίνακα 3 παρατίθενται οι αντίστοιχες τιμές των μεταβλητών που περιεγράφηκαν στο Πίνακα 2, για τις δυο υπό μελέτη περιοχές της Ρόδου που αφορούν τα συναπτά έτη 2014 έως 2019.

Πίνακας 2. *Περιγραφή μεταβλητών του μοντέλου διατακτικής παλινδρόμησης*

Μεταβλητή	Σύμβολο	Περιγραφή	Μονάδα/Κλίμακα Μέτρησης
Εξαρτημένη μεταβλητή			
Δεκαπενθήμερο	Y	Δεκαπενθήμερο ολοκλήρωσης του βιολογικού κύκλου του *M. hellenica*	Διατακτική μεταβλητή μετρημένη σε 5 κατηγορίες
Ανεξάρτητη μεταβλητή			

Αριθμός ροών κρύων ημερών μήκους 2	X_1	Αριθμός ροών κρύων ημερών μήκους 2 με μέγιστη ημερήσια θερμοκρασία $< 8°C$ κατά το μήνα Φεβρουάριο	Αναλογική μεταβλητή
Αριθμός ροών κρύων ημερών μήκους 2	X_2	Αριθμός ροών κρύων ημερών μήκους 2 με μέγιστη ημερήσια θερμοκρασία $< 9°C$ κατά το μήνα Φεβρουάριο	Αναλογική μεταβλητή
Αριθμός ροών κρύων ημερών μήκους 3	X_3	Αριθμός ροών κρύων ημερών μήκους 3 με μέγιστη ημερήσια θερμοκρασία $< 8°C$ κατά το μήνα Φεβρουάριο	Αναλογική μεταβλητή
Αριθμός ροών κρύων ημερών μήκους 3	X_4	Αριθμός ροών κρύων ημερών μήκους 3 με μέγιστη ημερήσια θερμοκρασία $< 9°C$ κατά το μήνα Φεβρουάριο	Αναλογική μεταβλητή

Πίνακας 3. *Παρουσίαση πειραματικών δεδομένων*

	Φεβρουάριος	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	**Y**
Καλυθιές	2014	0	0	0	0	2
	2015	3	4	2	2	3
	2016	1	1	0	1	4
	2017	1	2	1	1	2
	2018	0	0	0	0	1
	2019	0	0	0	0	2
Έμπωνας	2014	-	-	-	-	4
	2015	7	9	4	5	5
	2016	3	4	2	2	4

2017	4	7	3	5	5
2018	1	4	0	2	2
2019	5	7	3	4	3

Πριν παρουσιαστούν τα νέα αποτελέσματα του POM με τη χρήση ροών μήκους $k > 1$, θα αναφερθεί το αποτέλεσμα του POM για $k = 1$ και ανώτατο όριο μέγιστης ημερήσιας θερμοκρασίας τους $7.3°C$ (βλέπετε Gounari, et al., 2021). Σύμφωνα με το προαναφερόμενο μοντέλο για κάθε επιπλέον κρύα μέρα (με μέγιστη ημερήσια θερμοκρασία μικρότερη των $7.3°C$) το μήνα Φεβρουάριο, ο λόγος των συμπληρωματικών πιθανοτήτων (OR) το *M. hellenica* να ολοκληρώσει το βιολογικό του κύκλο εντός ενός συγκεκριμένου δεκαπενθημέρου μειώνεται κατά 52.15%.

Στη συνέχεια θα εφαρμοστεί το POM για τις (τέσσερις) περιπτώσεις των συνδυασμών των κριτηρίων που αναφέρθηκαν στην αρχή της ενότητας. Σε όλες τις προαναφερθείσες περιπτώσεις έγινε αποδεκτή η αρχική υπόθεση περί παραλληλίας των γραμμών logit ως προς την x $(p = 0.557, p = 0.475, p = 0.483, p = 0.766)$.

Για κάθε μια από τις προαναφερθείσες περιπτώσεις, παρατίθεται ένας πίνακας με τα αποτελέσματα της διατακτικής παλινδρόμησης και ένα γράφημα αθροιστικών ποσοστών, το οποίο μας δείχνει την αθροιστική πιθανότητα το *M. hellenica* να ολοκληρώσει το βιολογικό του κύκλο έως ένα δεκαπενθήμερο αναφοράς. Η κάθε καμπύλη του γραφήματος αντιστοιχεί σε έναν αριθμό ροών κρύων ημερών (από μηδέν έως έξι) που καταγράφηκαν για το μήνα Φεβρουάριο. Όσο πιο αριστερά στο γράφημα βρίσκεται μια καμπύλη αθροιστικών πιθανοτήτων τόσο σε πιο μικρό αριθμό ροών κρύων ημερών αφορά, και αντίστοιχα η πιθανότητα το *M. hellenica* να ολοκληρώσει το βιολογικό του κύκλο σε κάποιο από τα μεταγενέστερα δεκαπενθήμερα αναφοράς είναι πιο μικρή.

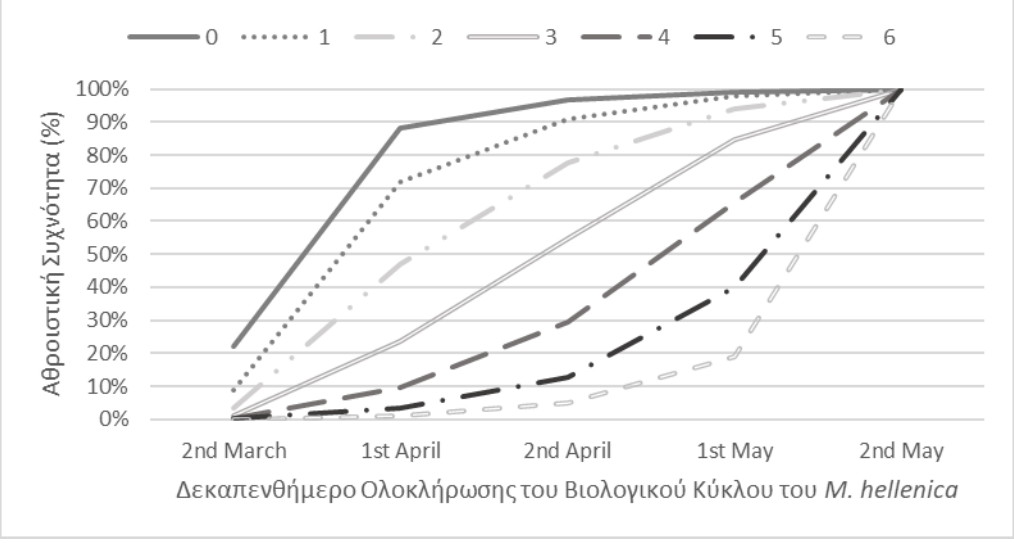
Τα στατιστικά αποτελέσματα από την εφαρμογή των προαναφερθέντων περιπτώσεων έχουν ως εξής:

Περίπτωση I $(k = 2, \text{threshold}: 8°C)$: Με την προσθήκη μίας ροής δυο ημερών με μέγιστη θερμοκρασία $< 8°C$ κατά το μήνα Φεβρουάριο, ο OR το *M. hellenica* να ολοκληρώσει το βιολογικό του κύκλο εντός ενός συγκεκριμένου δεκαπενθημέρου μειώνεται κατά 65.3% (Πίνακας 4). Επομένως, οι ροές κρύων ημερών μήκους 2 αποτελούν κρισιμότερο παράγοντα του χρόνου ολοκλήρωσης του βιολογικού κύκλου του εντόμου σε σχέση με το σύνολο των κρύων ημερών. Στο Σχήμα 2, παρουσιάζεται η αύξηση της πιθανότητας ολοκλήρωσης του βιολογικού κύκλου του *M. hellenica* σε κάποιο από τα μεταγενέστερα δεκαπενθήμερα, καθώς αυξάνεται ο αριθμός δυο συνεχόμενων κρύων ημερών με μέγιστη ημερήσια θερμοκρασία $< 8°C$.

Πίνακας 4. *Στατιστικά Αποτελέσματα POM* $(k=2, \text{threshold: } 8°C)$

		Εκτίμηση Παραμέτρων	Τυπικό Σφάλμα	p	
Σταθερά					
α_1	Δεκαπενθήμερο = [1]	-1.262	1.129	0.264	
α_2	Δεκαπενθήμερο = [2]	1.999	1.193	0.094	
α_3	Δεκαπενθήμερο = [3]	3.363	1.774	0.020	
α_4	Δεκαπενθήμερο = [4]	4.886	2.513	0.006	
Ανεξάρτητη Μεταβλητή		**β_1**			**Exp(β_1)**
X_1		-1.058	0.477	0.018	0.347

Σχήμα 2. *Γράφημα παρατηρούμενων αθροιστικών ποσοστών του δεκαπενθημέρου ολοκλήρωσης του βιολογικού κύκλου του M.hellenica* $(k=2, \text{threshold: } 8°C)$

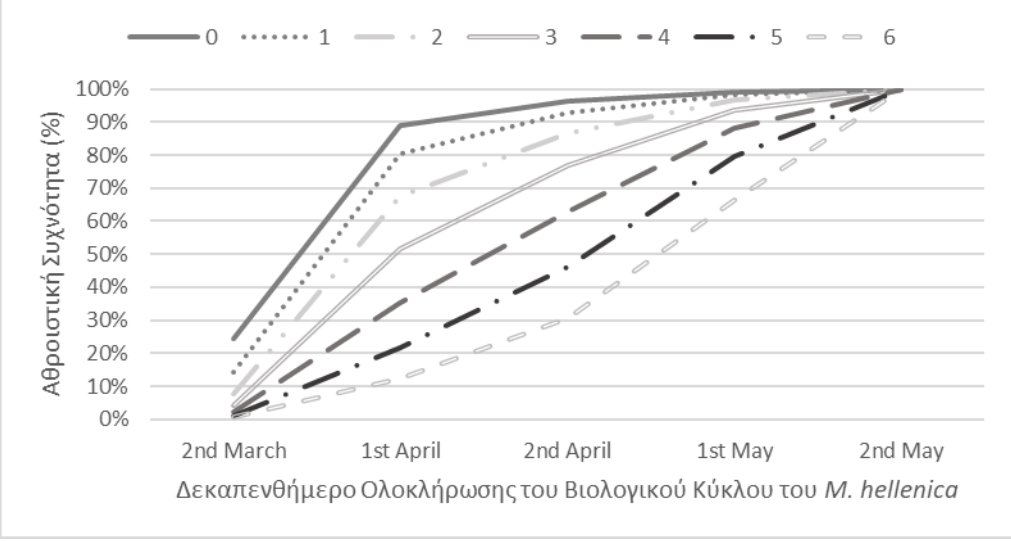


Περίπτωση ΙΙ $(k = 2, \text{threshold: } 9°C)$: Με την προσθήκη δυο συνεχόμενων ημερών με μέγιστη θερμοκρασία $< 9°C$ κατά το μήνα Φεβρουάριο, ο αντίστοιχος OR μειώνεται κατά 49% (Πίνακας 5). Παρόλο που αυξήσαμε το ανώτατο όριο θερμοκρασίας κατά ένα βαθμό $°C$, το POM συνεχίζει να είναι στατιστικά σημαντικό $(p = 0.017)$. Είναι φανερό (βλ. Σχήμα 2 και 3) ότι για τον ίδιο αριθμό δυο συνεχόμενων κρύων ημερών η πιθανότητα ολοκλήρωσης του βιολογικού κύκλου του εντόμου σε κάποιο από τα αρχικά δεκαπενθήμερα αναφοράς είναι μεγαλύτερη όταν ως μέγιστη θερμοκρασία για το χαρακτηρισμό των κρύων ημερών επιλεγούν οι 9°C.

Πίνακας 5. *Στατιστικά Αποτελέσματα POM (k=2, threshold: 9°C)*

		Εκτίμηση Παραμέτρων	Τυπικό Σφάλμα	p	
Σταθερά					
α_1	Δεκαπενθήμερο = [1]	-1.128	1.152	0.327	
α_2	Δεκαπενθήμερο = [2]	2.085	1.294	0.095	
α_3	Δεκαπενθήμερο = [3]	3.223	1.415	0.023	
α_4	Δεκαπενθήμερο = [4]	4.723	1.794	0.008	
Ανεξάρτητη Μεταβλητή		**β_1**			**Exp(β_1)**
X_2		-0.674	0.282	0.017	0.51

Σχήμα 3. *Γράφημα παρατηρούμενων αθροιστικών ποσοστών του δεκαπενθημέρου ολοκλήρωσης του βιολογικού κύκλου του M.hellenica (k=2, threshold: 9°C)*



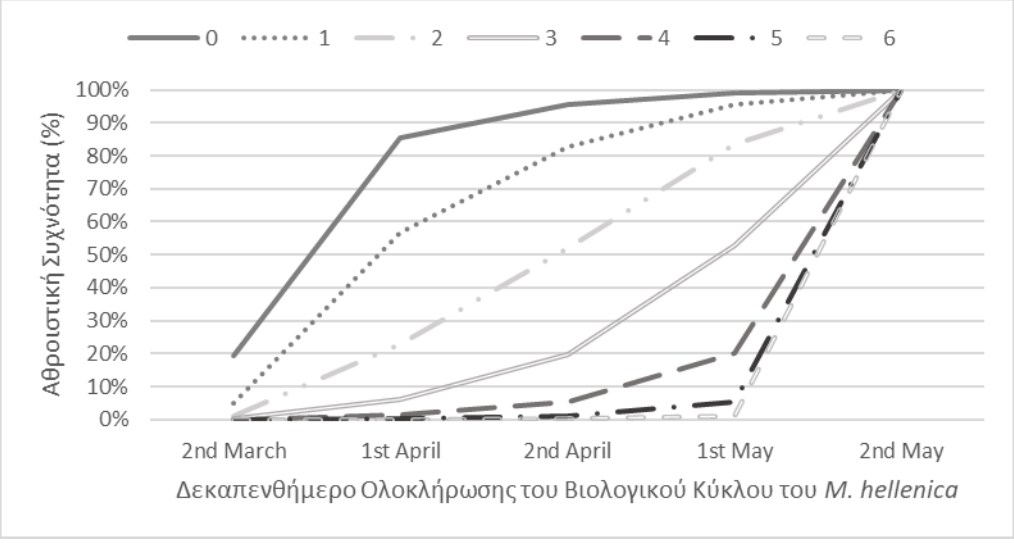
Περίπτωση III $(k = 3, \text{threshold: } 8°C)$: Με την προσθήκη τριών συνεχόμενων ημερών με μέγιστη ημερήσια θερμοκρασία $< 8°C$ κατά το μήνα Φεβρουάριο, ο αντίστοιχος OR μειώνεται κατά 77.5% (Πίνακας 6). Σε αυτή την υποπερίπτωση, αξιοσημείωτο είναι πως αυξάνοντας το μήκος της ροής σε $k = 3$ και θεωρώντας ίδιο ανώτατο όριο θερμοκρασίας τους $8°C$, ο αντίστοιχος OR να ολοκληρώσει το υπό μελέτη έντομο το βιολογικό του κύκλο μέσα σε δεδομένο δεκαπενθήμερο αναφοράς μειώνεται περαιτέρω κατά 12.2%. Συμπληρωματικά, στο Σχήμα 4, παρατηρείται πως οι καμπύλες αθροιστικών πιθανοτήτων που αντιστοιχούν σε μεγαλύτερο αριθμό ροών κρύων ημερών δείχνουν πως η πιθανότητα ολοκλήρωσης του βιολογικού κύκλου του

M. hellenica σε κάποιο από τα αρχικά δεκαπενθήμερα αναφοράς είναι σχεδόν μηδενική.

Πίνακας 6. *Στατιστικά Αποτελέσματα POM (k=3, threshold: 8°C)*

		Εκτίμηση Παραμέτρων	Τυπικό Σφάλμα	p	
Σταθερά					
α_1	Δεκαπενθήμερο = [1]	-1.422	1.115	0.202	
α_2	Δεκαπενθήμερο = [2]	1.769	1.154	0.125	
α_3	Δεκαπενθήμερο = [3]	3.070	1.398	0.028	
α_4	Δεκαπενθήμερο = [4]	4.597	1.758	0.009	
Ανεξάρτητη Μεταβλητή		**β_1**			**Exp(β_1)**
X_3		-1.493	0.616	0.015	0.225

Σχήμα 4. *Γράφημα παρατηρούμενων αθροιστικών ποσοστών του δεκαπενθημέρου ολοκλήρωσης του βιολογικού κύκλου του M.hellenica (k=3, threshold: 8°C)*

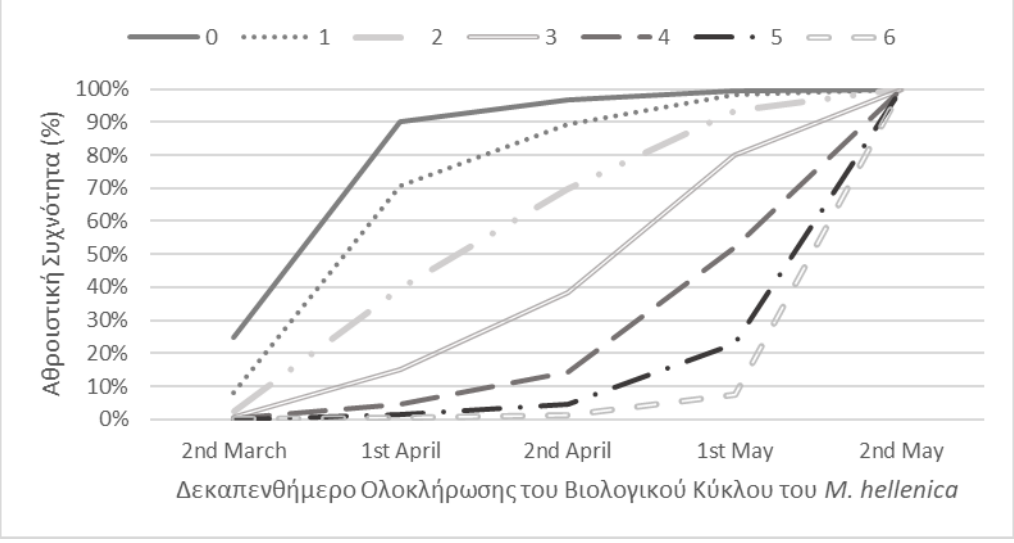


Περίπτωση IV $(k = 3, \text{threshold: } 9°C)$: Με την προσθήκη τριών συνεχόμενων ημερών με μέγιστη ημερήσια θερμοκρασία $< 9°C$ κατά το μήνα Φεβρουάριο, ο αντίστοιχος OR μειώνεται κατά 72.8% (Πίνακας 7). Σε σύγκριση με την υποπερίπτωση II, ο αντίστοιχος OR να ολοκληρώσει το *M. hellenica* το βιολογικό του κύκλο μέσα σε ένα συγκεκριμένο δεκαπενθήμερο αναφοράς μειώνεται περαιτέρω κατά 23.8%. Το αντίστοιχο γράφημα αθροιστικών ποσοστών παρουσιάζεται στο Σχήμα 5.

Πίνακας 7. *Στατιστικά Αποτελέσματα POM (k=3, threshold: 9°C)*

		Εκτίμηση Παραμέτρων	Τυπικό Σφάλμα	p	
Σταθερά					
α_1	Δεκαπενθήμερο = [1]	-1.114	1.147	0.331	
α_2	Δεκαπενθήμερο = [2]	2.189	1.213	0.071	
α_3	Δεκαπενθήμερο = [3]	3.437	1.405	0.014	
α_4	Δεκαπενθήμερο = [4]	5.289	2.030	0.009	
Ανεξάρτητη Μεταβλητή		**β_1**			**Exp(β_1)**
X_4		-1.303	0.506	0.010	0.272

Σχήμα 5. *Γράφημα παρατηρούμενων αθροιστικών ποσοστών του δεκαπενθημέρου ολοκλήρωσης του βιολογικού κύκλου του M.hellenica (k=3, threshold: 9°C)*



4. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΗ ΕΡΕΥΝΑ

Συμπερασματικά, στην παρούσα εργασία εισήχθη η χρήση του αριθμού των ροών ως ανεξάρτητη μεταβλητή σε μοντέλα διατακτικής παλινδρόμησης. Στο τρέχον πρόβλημα στη μελισσοκομία, η χρήση ροών μήκους k ($k > 1$) ανέδειξε την επιδραστικότητα των ροών κρύων ημερών στο χρόνο ολοκλήρωσης του βιολογικού κύκλου του *M. hellenica* ως πιο κρίσιμο παράγοντα σε σχέση με τον συνολικό αριθμό κρύων ημερών. Το γεγονός αυτό, υπογραμμίζει τη σημασία της χρήσης των ροών μήκους k ως συμπληρωματικό κριτήριο απόφασης σε μια πληθώρα εφαρμογών και προβλημάτων.

Μελλοντικά δύναται να χρησιμοποιηθούν και να εφαρμοστούν επιπλέον τρόποι καταμέτρησης ροών (για παράδειγμα επικαλυπτόμενες ροές) σε προβλεπτικά μοντέλα διατακτικής παλινδρόμησης. Επίσης, μπορεί να πραγματοποιηθεί η χρήση αντίστοιχων μοντέλων διατακτικής παλινδρόμησης σε άλλες ερευνητικές περιοχές που εφαρμόζεται εκτεταμένα η θεωρία ροών μήκους *k*. Τέλος, μια ακόμα πρόκληση αποτελεί η εφαρμογή και επέκταση ανάλογων μοντέλων παλινδρόμησης με την χρήση ροών αξιοποιώντας μεγαλύτερο όγκο δεδομένων και μεταβλητών.

ABSTRACT

During the last decades, a wide range of problems in several research areas has been modelled by classifying the experimental trials in two exclusive categories, considering sequences of binary trials (with values 0 or 1) and studying the sequences of outcomes. This study usually involves searching for great concentration of outcomes of a specific type. Such a concentration is traditionally measured by the enumeration of runs of k $(k \geq 1)$ ones. In the present work, an ordinal regression model employing runs is proposed. The proposed model is applied to the case of on-going research related to the biology of the insect *Marchalina hellenica* and it is demonstrated that the combination of runs, and ordinal regression may be proved beneficial to the field.

ΑΝΑΦΟΡΕΣ

Agresti, A. (2013*). Categorical data analysis* (3rd ed.). River Street, Hoboken, NJ: John Wiley & Sons.

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

Dafnis S. & Makri, F. (2021). Weak runs in sequences of binary trials. Metrika. https://doi.org/10.1007/s00184-021-00842-1

Gounari, S., Zotos, C., Dafnis, S., Moschidis, G., & Papadopoulos, G. (2021). On the impact of critical factors to honeydew production: The case of *Marchalina hellenica* and pine honey. *Journal of Apicultural Research*, https://doi.org/10.1080/00218839.2021.1999684

Feller, W. (1968). An Introduction to Probability Theory and Its Applications. Vol. I, 3rd edn., New York: John Wiley

Hirano, K. (1986). Some properties of the distributions of order k. In: Philippou A., Bergum G., Horadam A. (eds). *Fibonacci numbers and their applications*. Reidel, Dordecht, pp 43-53.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)* **42(2)**, 109-127.

Philippou, A., Makri, F. (1986). Successes, runs and longest runs. *Statistics and Probability Letters* IQ: 171-175.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 33ου Πανελληνίου Συνεδρίου Στατιστικής (2021), σελ.93-107

ΜΟΝΤΕΛΑ ΠΡΟΒΛΕΨΗΣ ΧΡΟΝΟΛΟΓΙΚΩΝ ΣΕΙΡΩΝ ΜΕ ΕΝΤΟΝΗ ΕΠΟΧΙΚΟΤΗΤΑ

Σόνια Μαλεφάκη[1], Χρήστος Κάτρης[2]
[1]Τμήμα Μηχανολόγων και Αεροναυπηγών Μηχανικών, Πανεπιστήμιο Πατρών
smalefaki@upatras.gr,
[2]Τμήμα Λογιστικής και Χρηματοοικονομικής, Οικονομικό Πανεπιστήμιο Αθηνών
chriskatris@aueb.gr

ΠΕΡΙΛΗΨΗ

Ένας από τους σημαντικότερους πυλώνες της ελληνικής οικονομίας είναι ο τουρισμός, οπότε η συλλογή, επεξεργασία και ανάλυση δεδομένων από τον συγκεκριμένο κλάδο είναι εξέχουσας σημασίας. Στην παρούσα εργασία μελετώνται οι μηνιαίες αφίξεις διεθνών τουριστών στις δεκατρείς (13) περιφέρειες της Ελλάδας. Τα συγκεκριμένα δεδομένα αποτελούν χρονοσειρές με βασικό χαρακτηριστικό τους την έντονη εποχικότητα. Για τη μοντελοποίηση τους χρησιμοποιήθηκαν οι παραδοσιακές μέθοδοι (μοντέλα SARIMA), Τεχνητά Νευρωνικά Δίκτυα (ΤΝΔ) και Υβριδικά Μοντέλα σε μία προσπάθεια να βρεθεί το καλύτερο πρότυπο το οποίο θα χρησιμοποιηθεί για την πρόβλεψη μελλοντικών τιμών τους. Εφαρμόζοντας τις συγκεκριμένες μεθόδους δε βρέθηκε ένα ενιαία βέλτιστο μοντέλο πρόβλεψης για όλες τις περιφέρειες, οπότε προτείνεται ένας τρόπος επιλογής καταλληλότερου μοντέλου με βάση την εποχικότητα των δεδομένων. Για τη μέτρηση της εποχικότητας χρησιμοποιείται ο συντελεστής Gini, ενώ μέσω της ROC ανάλυσης και του δείκτη Youden εντοπίζεται το σημείο αποκοπής για την επιλογή του καταλληλότερου μοντέλου.

Λέξεις Κλειδιά: Τουρισμός, πρόβλεψη αφίξεων, εποχικότητα, χρονοσειρές, Μοντέλα SARIMA, Τεχνητά Νευρωνικά Δίκτυα, υβριδικά μοντέλα.

1. ΕΙΣΑΓΩΓΗ

Ο τουρισμός αποτελεί ένα πολύπλοκο και δυναμικό σύστημα όπου διάφορα περιστατικά μπορούν να επηρεάσουν σοβαρά την παραγωγή ή/και την κατανάλωσή του. Είναι σαφές ότι παράγοντες όπως φυσικές καταστροφές, πολιτιστικά γεγονότα, κοινωνικές συμπεριφορές, πολιτικές μάρκετινγκ κ.λπ., μπορούν να επηρεάσουν την τουριστική ροή σε μια περιοχή, είτε με την αύξηση, είτε με τη μείωσή της. Αυτό ισχύει και για την Ελλάδα, όπου ο τουρισμός αποτελεί έναν από τους σημαντικότερους πυλώνες της οικονομίας της χώρας.

Ο τουρισμός αποτελεί μια σημαντική πηγή εσόδων για την Ελλάδα καθώς συμβάλλει σημαντικά στο ελληνικό ακαθάριστο εγχώριο προϊόν (ΑΕΠ) (από το 2006 συμμετέχει στο ΑΕΠ με ποσοστό άνω του 15.8% ενώ σημειώνεται ότι ξεπέρασε το

18% συμμετοχής στο ΑΕΠ το 2016, βάση στοιχείων από το Ινστιτούτο Συνδέσμου Ελληνικών Τουριστικών Επιχειρήσεων (ΙΝΣΕΤΕ)- https://sete.gr/el/stratigiki-gia-ton-tourismo/vasika-megethi-tou-ellinikoy-tourismoy), ειδικότερα τις τελευταίες δεκαετίες μέσω των εσόδων που προέρχονται από άμεσους και έμμεσους τουριστικούς πόρους, και οδηγεί στην αύξηση της απασχόλησης αλλά και την ανάπτυξη έμμεσων τουριστικών οικονομικών δραστηριοτήτων. Αν και τα τελευταία χρόνια, η Ελλάδα βιώνει την παγκόσμια χρηματοπιστωτική κρίση με όλες τις αρνητικές της συνέπειες, υπάρχει μεγάλη αύξηση διεθνών τουριστών (αύξηση άνω του 37% κατά την περίοδο 2012-2016 σύμφωνα με την Ελληνική Στατιστική Αρχή, www.statistics.gr) .

Η μελέτη της διεθνούς τουριστικής ζήτησης σε επίπεδο περιφέρειας στην Ελλάδα αποτελεί μεγάλη πρόκληση, λόγω των διαφορετικών χαρακτηριστικών που έχουν οι περιοχές αυτές (νησιωτικές, ορεινές, μεγάλες πόλεις, απομακρυσμένα χωριά, τουριστικές υποδομές, αρχαιολογικοί χώροι, αθλητικές εγκαταστάσεις κ.λπ.). Είναι σημαντικό να τονιστεί ότι στις περιοχές όπου τα κυρίαρχα έσοδα προέρχονται από τουριστικές δραστηριότητες, η συνεισφορά τους υπερβαίνει το 50% της συνολικής συνεισφοράς της περιοχής στο ΑΕΠ. Αυτό δείχνει ότι η ανάλυση σε επίπεδο χώρας αδυνατεί να εντοπίσει και να αναδείξει τις πραγματικές ανάγκες που υπάρχουν σε επίπεδο περιφέρειας. Βλέπουμε λοιπόν ότι η μέτρηση και η ανάλυση των τουριστικών ροών σε επίπεδο περιφέρειας αποτελεί ένα απαραίτητο βήμα για την αξιολόγηση της τουριστικής ιδιαιτερότητας αλλά και της τουριστικής συνεισφοράς κάθε περιοχής και επιτρέπει στους ειδικούς του τουρισμού να σχεδιάσουν μια σαφέστερη εικόνα της ανάπτυξης του τουρισμού σε αυτές τις περιοχές, να σχεδιάσουν πιο αποτελεσματικές πολιτικές βάσει του προφίλ κάθε περιοχής αλλά και να δημιουργήσουν ένα πλαίσιο παρακολούθησης της αποτελεσματικότητας αυτών των πολιτικών στην πάροδο του χρόνου. Κεντρικό ρόλο σε αυτό παίζει η πρόβλεψη της αντίστοιχης τουριστικής ζήτησης διαχρονικά. Επίσης για τη διαχείριση τουριστικών αφίξεων βραχυπρόθεσμα (ειδικά για την καλοκαιρινή περίοδο) αλλά και τον σχεδιασμό τουριστικής πολιτικής μακροπρόθεσμα, είναι απαραίτητα κατάλληλα μοντέλα πρόβλεψης για τους αντίστοιχους ορίζοντες πρόβλεψης. Μεγαλύτερη ακρίβεια των προβλεπτικών μοντέλων οδηγεί σε αύξηση της αποτελεσματικότητας των βραχυπρόθεσμων και μακροπρόθεσμων πολιτικών τουριστικής ανάπτυξης.

Το ερευνητικό πρόβλημα σχετικά με την τουριστική ζήτηση που έχει μελετηθεί περισσότερο τα τελευταία χρόνια είναι αυτό της πρόβλεψης. Το πρόβλημα της πρόβλεψης χρονοσειρών έχει μελετηθεί διεξοδικά και έχουν προταθεί πληθώρα τεχνικών για την αντιμετώπιση του. Τα πιο δημοφιλή μοντέλα είναι τα μονομεταβλητά μοντέλα χρονοσειρών (Gunter & Önder, 2015) όπου η πιο ευρέως χρησιμοποιούμενη τεχνική σε αυτό το πλαίσιο είναι τα (Εποχιακά) μοντέλα Αυτοπαλίνδρομης Ολοκλήρωσης Κινούμενου Μέσου (μοντέλα SARIMA) (Box & Jenkins, 1976). Τα SARIMA μοντέλα μάλιστα θεωρούνται τόσο επιτυχημένα που αποτελούν τη κύρια τεχνική πρόβλεψης της τουριστικής ζήτησης. Επίσης, έχουν χρησιμοποιηθεί ευρέως πολυμεταβλητά ή/και οικονομετρικά μοντέλα, όπως για παράδειγμα τα Αυτοπαλίνδρομα Κατανεμημένα μοντέλα υστερήσεων (autoregressive

distributed lag models (ADLs)) (Dritsakis, & Athanasiadis, 2000), μοντέλα Διόρθωσης σφάλματος (Error Correction models) (Kulendran, & Witt, 2003), Αυτοπαλίνδρομα διανυσματικά υποδείγματα (VAR) (Shan & Wilson, 2001) και μοντέλα Χρονικά μεταβαλλόμενων παραμέτρων (Time-Varying Parameter) (Song, & Witt, 2006). Ένα από τα σημαντικότερα πλεονεκτήματα των μοντέλων χρονολογικών σειρών είναι ότι είναι σε θέση να δίνουν προβλέψεις γνωρίζοντας μόνο την ιστορία της υπό μελέτη μεταβλητής, ενώ από την άλλη τα οικονομετρικά μοντέλα συμπεριλαμβάνουν στη μελέτη και άλλες μεταβλητές. Την τελευταία δεκαετία, μέθοδοι τεχνητής νοημοσύνης έχουν παρουσιαστεί στη βιβλιογραφία με σκοπό την πρόβλεψη της τουριστικής ζήτησης, όπως μοντέλα νευρωνικών δικτύων (Koutras et al., 2016; Olmedo 2016), μοντέλα Μηχανών Διανυσμάτων Υποστήριξης (Support Vector Machines (SVMs)) (Koutras et al., 2016) όπως επίσης και εργασίες που πρότειναν υβριδικές τεχνικές πρόβλεψης (Chen, 2011). Ο κύριος λόγος χρησιμοποίησης αυτών των τεχνικών είναι ότι μπορεί να υπάρχουν μη-γραμμικότητες (φανερές ή μη) που τα κλασσικά γραμμικά μοντέλα χρονοσειρών δεν μπορούν να εντοπίσουν. Μια εξαντλητική ανασκόπηση των παραπάνω τεχνικών μπορεί κανείς να βρεις στους Song & Li, (2008).

Ένα από τα βασικότερα συμπεράσματα που προκύπτει από την εργασία των Song και Li (2008), είναι ότι δεν υπάρχει ένα ενιαία βέλτιστό μοντέλο που να μπορεί να χρησιμοποιηθεί σε όλες τις περιπτώσεις όσον αφορά τη μοντελοποίηση και την πρόβλεψη της τουριστικής ζήτησης. Επομένως, η ανάλυση είναι σημαντικό να γίνεται σε επίπεδο περιφερειών και όχι ενιαία για όλη τη χώρα, προκειμένου να προκύψουν πιο ακριβείς προβλέψεις και στοχευμένα συμπεράσματα, και αυτό μπορεί να επιτευχθεί μόνο αν οι ειδικοί στον τουρισμό γνωρίζουν ποια μοντέλα πρόβλεψης θα πρέπει να εφαρμοστούν σε ποια περιοχή και για ποιον ορίζοντα πρόβλεψης ώστε να είναι σε θέση να σχεδιάσουν βέλτιστες τουριστικές στρατηγικές.

Στην παρούσα εργασία, αναπτύσσονται τρία διαφορετικά μοντέλα πρόβλεψης για τη διερεύνηση της μελλοντικής τουριστικής ζήτησης στην Ελλάδα σε δύο διαφορετικές περιόδους: τη βραχυπρόθεσμη (χρονικός ορίζοντας 3 μηνών, ο οποίος σκοπεύει να προβλέψει τη ζήτηση της περιόδου αιχμής, τους καλοκαιρινούς μήνες για την Ελλάδα) και τη μακροπρόθεσμη (χρονικός ορίζοντας 12 μηνών). Οι δύο αυτοί χρονικοί ορίζοντες επιλέχθηκαν κυρίως γιατί αντιστοιχούν στην ανάπτυξη στρατηγικών που συνεπάγονται την εφαρμογή πολιτικών με διαφορετικούς στόχους. Η έρευνα επικεντρώνεται στη μελέτη των αφίξεων διεθνών τουριστών στην Ελλάδα, όχι σε εθνικό επίπεδο, αλλά σε κάθε μία από τις 13 περιφέρειες ξεχωριστά. Πρώτα αναπτύσσονται τα κλασικά μοντέλα SARIMA, στη συνέχεια εφαρμόζονται μοντέλα τεχνητών νευρωνικών δικτύων με αρχιτεκτονική εμπρόσθιας τροφοδότησης και πλήρως συνδεδεμένα (Multi-Layer Fully Connected Feed-Forward) (MLP). Εκτός των μεμονωμένων μοντέλων, προτείνεται η υβριδοποίηση που συνδυάζει τα μοντέλα SARIMA με τα τεχνητά νευρωνικά δίκτυα MLP. Τα προαναφερθέντα μοντέλα συγκρίνονται με βάση την ευρέως χρησιμοποιούμενη μετρική αξιολόγησης

προβλέψεων, το RMSE (Root Mean Squared Error=$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2}$) όπου y_i η πραγματική τιμή και $\hat{y_i}$ η πρόβλεψη της για κάθε χρονική στιγμή i.

Ένα βασικό ζήτημα που προκύπτει είναι η επιλογή κατάλληλου μοντέλου για την πρόβλεψη των αφίξεων για τις δύο διαφορετικές χρονικές περιόδους (12 και 3 μήνες που αντιστοιχούν στην καλοκαιρινή περίοδο) και για κάθε περιφέρεια ξεχωριστά. Η επιλογή του καταλληλότερου μοντέλου γίνεται μέσω της ανάλυση ROC. Χρησιμοποιώντας τον δείκτη Youden εντοπίζεται το σημείο αποκοπής ως προς ένα δείκτη εποχικότητας των διεθνών αφίξεων της περιοχής (το συντελεστή Gini) και με βάση αυτό το σημείο μπορούμε να επιλέξουμε το καταλληλότερο μοντέλο ανάλογα με την ένταση της εποχικότητας της αντίστοιχης περιοχής.
Το υπόλοιπο της εργασίας οργανώνεται ως εξής. Στην Ενότητα 2 παρουσιάζεται η περιγραφή των δεδομένων και ορισμένοι δείκτες εποχικότητας που είναι και το βασικό τους χαρακτηριστικό. Στην Ενότητα 3 δίνεται μια σύντομη περιγραφή των χρησιμοποιούμενων προβλεπτικών μεθοδολογιών ενώ στην Ενότητα 4 αναπτύσσεται η προτεινόμενη μεθοδολογία επιλογής καταλληλότερου μοντέλου με βάση την εποχικότητα. Τέλος, η εργασία ολοκληρώνεται με μια σύντομη συζήτηση και επισημαίνοντας θέματα για περαιτέρω έρευνα.

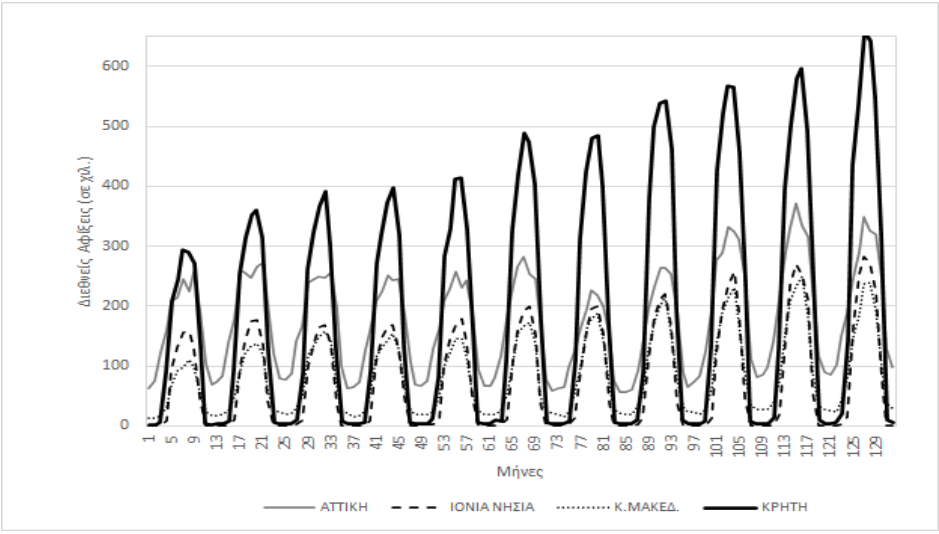
2. ΧΑΡΑΚΤΗΡΙΣΤΙΚΑΤΩΝ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΠΟΧΙΚΟΤΗΤΑ

Για κάθε περιοχή, το σύνολο δεδομένων που μελετήθηκε αποτελείται από 132 παρατηρήσεις, τις μηνιαίες διεθνείς αφίξεις για τα έτη 2006-2016. Στο Σχήμα 1, απεικονίζονται ενδεικτικά τα διαγράμματα των χρονοσειρών (time-series plot) των περιφερειών Αττικής, Κρήτης, Ιονίων Νησιών και Κεντρικής Μακεδονίας. Σε όλες τις περιφέρειες παρατηρείται εποχικότητα ανά έτος και υψηλότερες τιμές κατά την καλοκαιρινή περίοδο. Η Κρήτη και τα Ιόνια Νησιά έχουν πιο έντονη εποχικότητα σε σχέση με την Αττική και την Κεντρική Μακεδονία. Στο Σχήμα 1 αναδεικνύεται η διαφορετική τουριστικής συμπεριφοράς στις νησιωτικές και ηπειρωτικές περιοχές της Ελλάδας.

Τα βασικά περιγραφικά μέτρα του συνόλου των δεδομένων μας παρουσιάζονται στον Πίνακα 1 όπου οι περιοχές ταξινομούνται κατά φθίνουσα σειρά σε σχέση με το πλήθος αφίξεων διεθνών τουριστών τους. Φαίνεται ότι η Κρήτη, το Νότιο Αιγαίο, η Αττική, η Κεντρική Μακεδονία και τα Ιόνια Νησιά είναι οι πέντε περιοχές με την εντονότερη τουριστική κίνηση.

Η εποχικότητα αποτελεί έναν κρίσιμο τουριστικό παράγοντα, καθώς, όπως παρατήρησε ο Butler (2014), αντιμετωπίζεται ως πρόβλημα που γενικά δε βελτιστοποιεί τα ετήσια οικονομικά οφέλη της περιοχής. Ένας καλός δείκτης εποχικότητας θα μπορούσε να θεωρηθεί και ο συντελεστής μεταβλητότητας (CV) ο οποίο απεικονίζεται για κάθε περιφέρεια στην τελευταία στήλη του Πίνακα 1.

Σχήμα 1. *Διεθνείς Αφίξεις Τουριστών για τα έτη 2006-2016*



Παρατηρούμε ότι κάποιες περιοχές (όπως η Κρήτη, τα Επτάνησα, το Νότιο και το Βόρειο Αιγαίο) παρουσιάζουν υψηλή εποχικότητα. Από την άλλη πλευρά, υπάρχουν περιοχές με χαμηλές τιμές του δείκτη CV που αντιστοιχούν σε πιο ομοιόμορφες ροές αφίξεων τουριστών (π.χ. Αττική, Στερεά Ελλάδα). Παρόλο που η τιμή του συντελεστή μεταβλητότητας μπορεί να υπολογιστεί εύκολα, είναι δύσκολο να ερμηνευτεί (Þórhallsdóttir, & Ólafsson, 2017). Μελέτες για την εποχικότητα του τουρισμού στην Ελλάδα (www.itep.gr) έδειξαν τη σημαντικά πιο έντονη εποχικότητα των αφίξεων των διεθνών τουριστών της χώρας μας σε σύγκριση με τις ανταγωνιστικές χώρες, αλλά και το γεγονός ότι η εποχικότητα αποτελεί κοινό χαρακτηριστικό σε όλες τις επιμέρους περιφέρειες της με την ένταση της όμως να αλλάξει από περιφέρεια σε περιφέρεια.

Λόγω των παραπάνω, χρησιμοποιείται ένας ακόμη πολύ σημαντικός δείκτης εποχικότητας που είναι ο γνωστός συντελεστής Gini (GC), ο οποίος μετρά τη διαφορά της πραγματικής κατανομής των επισκεπτών μέσα στο έτος από την ομοιόμορφη κατανομή τους μέσα σε αυτό. Ο δείκτης GC έχει το πλεονέκτημα ότι είναι λιγότερο ευαίσθητος στις υψηλότερες τιμές εποχικότητας, αλλά επηρεάζεται αρκετά από τις διακυμάνσεις εκτός της περιόδου αιχμής (Þórhallsdóttir and Ólafsson, 2017). Ο δείκτης Gini μπορεί να οριστεί ως μαθηματικό ισοδύναμο της καμπύλης Lorenz και ο υπολογισμός του πραγματοποιείται χρησιμοποιώντας τον τύπο:

$$G_k = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_{ki} - x_{kj}\right|}{2n\sum_{i=1}^{n} x_{ki}}.$$

Πίνακας1. *Περιγραφικά στατιστικά διεθνών τουριστικών αφίξεων (2006-2016)*

Περιφέρεια	ΜΤ	ΤΑ	Ασυμ.	Κύρτ.	Ελαχ.	Μεγ.	CV
Κρήτη (CRE)	197423,8	201295,5	0.531	-1.145	2003	658890	1.0196
Νότιο Αιγαίο (SAG)	183283,4	198685,1	0.617	-1.115	936	606078	1.0840
Αττική (ATT)	176295,4	84665,4	0.188	-1.170	55810	370917	0.4802
Κεντρική Μακεδονία (CMC)	84186,8	66519,1	0.703	-0.746	12349	249481	0.7901
Ιόνια Νησιά (ION)	76330,8	86670,7	0.676	-0.987	388	282154	1.1355
Πελοπόννησος (PEL)	24127,7	17459,4	0.168	-1.377	2779	57347	0.7236
Δυτική Ελλάδα (WGR)	20288,4	15827,5	0.280	-1.390	1377	49912	0.7801
Θεσσαλία (THE)	20164,0	14832,8	0.414	-1.159	2842	54979	0.7356
Κεντρική Ελλάδα (CGR)	14752,0	9425,6	-0.011	-1.520	1813	31777	0.6389
Ανατολική Μακεδονία-Θράκη (EMT)	14121,8	13633,5	1.435	1.402	1572	56379	0.9654
Βόρειο Αιγαίο (NAG)	14030,6	15351,0	0.812	-0.583	285	56398	1.0941
Ήπειρος (EPR)	5543,2	4545,2	1.022	0.205	1103	21072	0.8200
Δυτική Μακεδονία (WMC)	1590,8	653,2	0.070	-0.973	312	3099	0.4106
ΕΛΛΑΔΑ (ΣΥΝΟΛΟ)	832138,6	714195,3	0.528	-1.131	98156	2402993	0.8580

όπου x_{ki} είναι οι αφίξεις διεθνών τουριστών της k περιοχής τον i μήνα και n το πλήθος των διαθέσιμων μετρήσεων.

Στον Πίνακα 2 παρουσιάζονται οι διεθνείς αφίξεις τουριστών σε όλες τις περιφέρειες της Ελλάδας για το έτος 2016. Εδώ, φαίνεται καθαρά η ασύμμετρη συγκέντρωση του τουρισμού στην Ελλάδα, αφού σχεδόν το 90% των ετήσιων διεθνών αφίξεων τουριστών φιλοξενούνται σε μόνο 5 περιφέρειες. Επίσης, στην τελευταία στήλη του

Πίνακας 2. *Κατανομή συχνότητας διεθνών τουριστικών αφίξεων για το έτος 2016*

Περιφέρεια	Αφίξεις	%	Αθροιστικό %	Συντελεστής GINI
CRE	3334850	25.67	25.67	0.50
SAG	2933974	22.58	48.25	0.52
ATT	2530513	19.48	67.73	0.24
CMC	1338542	10.30	78.03	0.39
ION	1295209	9.97	88.00	0.55
PEL	339781	2.62	90.62	0.35
WGR	196720	1.51	92.13	0.38
THE	267390	2.06	94.19	0.40
CGR	189844	1.46	95.65	0.33
EMT	268807	2.07	97.72	0.44
NAG	187936	1.45	99.17	0.48
EPR	93809	0.72	99.89	0.45
WMC	14510	0.11	100.00	0.25
ΕΛΛΑΔΑ (ΣΥΝΟΛΟ)	12991885			0.43

πίνακα, παρουσιάζεται η τιμή του συντελεστή Gini (GC) για το έτος 2016 για κάθε περιοχή. Παρατηρούμε ότι σε όλα τα ελληνικά νησιά ο μέσος όρος του συντελεστή είναι υψηλότερος σε σύγκριση με τις αντίστοιχες τιμές του στις υπόλοιπες περιοχές της Ελλάδας. Με βάση τον μέσο όρο του GC χαμηλότερη εποχικότητα παρουσιάζουν η Αττική και η Δυτική Μακεδονία, η οποία υποδηλώνει πιο ομοιόμορφες ροές τουριστών κατά τη διάρκεια του έτους ενώ οι περιοχές με τη μεγαλύτερη τιμή GC είναι η Κρήτη, τα Ιόνια Νησιά και το Νότιο Αιγαίο, οι οποίες παρουσιάζουν υψηλές ροές τουριστών κατά κύριο λόγο τους καλοκαιρινούς μήνες. Παρατηρούμε ότι σε όλα τα ελληνικά νησιά. Ο συντελεστής Gini είναι υψηλότερος σε σύγκριση με τις αντίστοιχες τιμές του στις υπόλοιπες περιοχές της Ελλάδας.

Επίσης, ελέγχονται οι υποθέσεις της κανονικότητας (έλεγχος Jarque-Bera), της στασιμότητας (μέσω του ελέγχου Augmented Dickey-Fuller (ADF), της αυτοσυσχέτισης (μέσω του ελέγχου Ljung-Box) και της μη γραμμικότητας (έλεγχος White Neural Network (Lee, White, & Granger, 1993) των δεδομένων. Οι μηδενικές υποθέσεις για τους παραπάνω ελέγχους είναι: για τον έλεγχο ADF η μη στασιμότητα έναντι της εναλλακτικής υπόθεσης ότι η σειρά είναι στάσιμη ή trend-στάσιμη, για τον έλεγχο Ljung-Box η ανεξαρτησία, για τον έλεγχο White Neural Network η γραμμικότητα και για τον έλεγχο Jarque-Bera η κανονικότητα των δεδομένων. Στον Πίνακα 3 παρουσιάζονται οι τιμές των στατιστικών των παραπάνω ελέγχων και οι αντίστοιχες περιοχές που βρίσκονται οι p-τιμές, σε παρένθεση στην πρώτη γραμμή του πίνακα (για κάθε έλεγχο). Στις περιπτώσεις που ισχύει κάτι διαφορετικό, σημειώνεται η p-τιμή σε παρένθεση στο αντίστοιχο κελί. Με βάση τους παραπάνω ελέγχους (βλέπε Πίνακα 3) μπορούμε να ισχυριστούμε ότι τα δεδομένα μας μπορούν

Πίνακας 3. *Στατιστικοί Έλεγχοι και αντίστοιχα p-values*

Region	Στασιμότητα (ADF test) (<0.01)	Αυτοσυσχέτιση (Ljung-Box) (<0.01)	Μη-γραμμικότητα (White Test) (>0.1)	Κανονικότητα (Jarque-Bera) (<0.01)
CRE	-12.225	92.874	3.0770	13.21
SAG	-10.315	91.080	0.6259	15.056
ATT	-6.563	94.789	1.8571	7.9682 (0.0186)
CMC	-10.175	94.864	1.3790	13.897
ION	-9.148	87.943	0.9935	15.313
PEL	-10.467	89.083	1.2413	10.679
WGR	-10.142	86.838	3.3044	12.015
THE	-8.673	90.930	3.2498.	10.895
CGR	-8.392	84.194	1.3034	12.324
EM&T	-8.436	90.032	2.925	58.245
NAG	-8.530	85.840	4.6816 (0.096)	16.483
EPR	-9.989	87.905	3.9354	23.838
WMC	-5.628	63.326	0.8601	4.9875 (0.083)
ΕΛΛΑΔΑ (ΣΥΝΟΛΟ)	-10.104	93.711	3.4791	12.966

να θεωρηθούν αυτοσυσχετιζόμενα. Επιπλέον, για τις διεθνείς αφίξεις τουριστών δεν μπορούμε να υποθέσουμε ότι ακολουθούν κανονική κατανομή με εξαίρεση τις αφίξεις της περιφέρειας Δυτικής Μακεδονίας, ενώ δεν υπάρχει σημαντική ένδειξη ύπαρξης μη-γραμμικότητας. Όλοι οι παραπάνω έλεγχοι έγιναν σε επίπεδο σημαντικότητας 5%.

3. ΜΕΘΟΔΟΛΟΓΙΕΣ ΠΡΟΒΛΕΨΗΣ

Το Autoregressive Integrated Moving Average Model (ARIMA), εισήχθη αρχικά από τους Box και Jenkins (1976) και χρησιμοποιείται ευρέως για τη μοντελοποίηση και την ανάλυση χρονοσειρών. Τα μοντέλα ARIMA εστιάζουν στη βραχυπρόθεσμη ανάλυση (Kantz, & Schreiber, 2004) και συμβολίζονται ως ARIMA(p, d, q), όπου p είναι η τάξη του αυτοσυσχετιζόμενου μέρους, d είναι η τάξη διαφοροποίησης και με q συμβολίζεται το μέρος του κινητού μέσου όρου. Γενικά, τα μοντέλα ARIMA χρησιμοποιούνται σε χρονοσειρές που δεν παρουσιάζουν εποχικότητα. Ωστόσο, αυτό το μοντέλο μπορεί να επεκταθεί σε ένα πολλαπλασιαστικό μοντέλο, που ονομάζεται SARIMA, SARIMA(p, d, q)x(P, D, Q)s, όπου ο όρος (P, D, Q) αντιστοιχεί στο εποχικό μέρος των δεδομένων και ο δείκτης s αντιστοιχεί στην περίοδο εποχικότητας. Στα πλαίσια αυτήν της εργασίας, η τάξη του μοντέλου αποφασίζεται με βάση το κριτήριο BIC και η εκτίμηση των παραμέτρων πραγματοποιείται χρησιμοποιώντας το

πακέτο πρόβλεψης `forecast` του λογισμικού `R` με τη μέθοδο Μέγιστης Πιθανοφάνειας. (Hyndman & Khandakar, 2008).

Η προσέγγιση των Τεχνητών Νευρωνικών Δικτύων (ANN) είναι μια τεχνική μηχανικής μάθησης που εκτός των άλλων χρησιμοποιείται και για πρόβλεψη (π.χ. Lippmann, 1987) και αποτελεί μεταγενέστερη τεχνική, σε σύγκριση με τα κλασικά μοντέλα χρονοσειρών. Τα μοντέλα πρόβλεψης ANN για χρονοσειρές χρησιμοποιούν ένα σύνολο από k πρόσφατες, διαθέσιμες τιμές για την πρόβλεψη της επόμενης. Υπήρξαν πολλές διαφορετικές αρχιτεκτονικές για τα δίκτυα ANN, αλλά στην παρούσα εργασία χρησιμοποιούνται τα πολυεπίπεδα πλήρως συνδεδεμένα Νευρωνικά Δίκτυα (Feed-Forward Neural Networks) (MLP), τα οποία θεωρούνται ως μία από τις πιο δημοφιλείς αρχιτεκτονικές.

Μετά την απόφαση της αρχιτεκτονικής του MLP, πρέπει να αποφασίσουμε για τις μεταβλητές εισόδου, τον αριθμό των κρυφών επιπέδων και τον αριθμό των κόμβων σε κάθε επίπεδο. Ως είσοδο, θεωρούμε δύο προγενέστερες παρατηρήσεις που είναι οι αφίξεις του αντίστοιχου μήνα του προηγούμενου έτους και οι αφίξεις του προηγούμενου μήνα του συγκεκριμένου έτους. Η λογική πίσω από αυτή την επιλογή είναι η ιδιότητα της ετήσιας εποχικότητας, την οποία γνωρίζουμε ότι υπάρχει στα δεδομένα τουριστικής ζήτησης. Διατηρούμε την είσοδο έως δύο κόμβους για λόγους απλότητας. Εμπειρική έρευνα έχει δείξει ότι ένα κρυφό επίπεδο είναι αρκετό στις περισσότερες εφαρμογές (Haykin, 2001). Έτσι, το νευρωνικό δίκτυο MLP που χρησιμοποιείται εδώ περιλαμβάνει ένα επίπεδο εισόδου με μία τιμή (υστέρηση 12) ή δύο τιμές (υστέρηση 1, 12) ως εισόδους, ένα κρυφό επίπεδο με 1 και 10 κόμβους όταν υπάρχει ένας κόμβος εισόδου ή 2, 4, 10 και 15 κόμβους όταν υπάρχουν 2 κόμβοι εισόδου και ένα επίπεδο εξόδου με έναν κόμβο. Τα δεδομένα τυποποιούνται (z-score) χρησιμοποιώντας τη μέση τιμή και τη τυπική απόκλιση του δείγματος εκπαίδευσης. Κάθε επίπεδο συνδέεται πλήρως με το επόμενο και η συνάρτηση ενεργοποίησης που χρησιμοποιείται στο κρυφό επίπεδο είναι η σιγμοειδής: $S(t) = \frac{1}{1 + e^{-t}}$. Επιπλέον, μια γραμμική συνάρτηση χρησιμοποιείται στο επίπεδο εξόδου για να μετατρέψει τις προηγούμενες εισόδους σε τελικές εξόδους. Η εκπαίδευση του δικτύου έχει γίνει με την προσαρμοστική μέθοδο backpropagation with momentum, όπου τα βάρη των συνδέσεων στο νευρωνικό δίκτυο υπολογίζονται χρησιμοποιώντας τον προσαρμοστικό αλγόριθμο βελτιστοποίησης κλίσης (Haykin, 2001). Ο κύριος στόχος της παρούσας εργασίας είναι να δώσει προβλέψεις για τις αφίξεις διεθνών τουριστών σε κάθε περιφέρεια της Ελλάδας τρία (3) και δώδεκα (12) βήματα μπροστά που αντιστοιχούν στους προσεχείς καλοκαιρινούς μήνες και στο επόμενο έτος αντίστοιχα. Η πολυβηματική πρόβλεψη ενός ANN είναι αρκετά δύσκολη και δεν υπάρχει μοναδική προσέγγιση γι' αυτό. Μια ανασκόπηση των προσεγγίσεων αυτών υπάρχει στην εργασία των Boné & Crucianu, (2002). Οι δύο κύριες προσεγγίσεις είναι η άμεση μέθοδος, όπου κατασκευάζονται διαφορετικά μοντέλα MLP, ένα για κάθε βήμα πρόβλεψης και η επαναληπτική μέθοδος, όπου η πρόβλεψη για την επόμενη περίοδο θεωρείται νέα παρατήρηση και το μοντέλο ανακατασκευάζεται προκειμένου να ληφθεί η επόμενη πρόβλεψη. Σε αυτήν την εργασία χρησιμοποιούμε την

επαναληπτική προσέγγιση και οι προβλέψεις πολλαπλών περιόδων προκύπτουν επαναληπτικά ως προβλέψεις ενός βήματος μπροστά. Η κλασική επαναληπτική μέθοδος χρησιμοποιεί την πρόβλεψη για την επόμενη περίοδο από το μοντέλο (MLP) ως πραγματική παρατήρηση, στη συνέχεια το μοντέλο επανεκτιμάται και λαμβάνουμε την πρόβλεψη για μια περίοδο μπροστά. Ωστόσο, για να βελτιώσουμε περαιτέρω την απόδοση του μοντέλου, εκτελούμε μια παραλλαγή της επαναληπτικής μεθόδου. Χρησιμοποιούμε ως εκτιμήσεις όχι μόνο τις προβλέψεις του νευρωνικού δικτύου, αλλά και άλλων μοντέλων πρόβλεψης. Σε αυτή την εργασία, μετά την κατασκευή του ANN θεωρούμε 3 τιμές: τις προβλέψεις του, τις προβλέψεις από ένα εποχικό μοντέλο Holt-Winters και τις προβλέψεις από το μοντέλο SARIMA ως πραγματικά δεδομένα της επόμενης περιόδου. Έτσι δημιουργούνται 3 σενάρια προβλέψεων και για κάθε ένα από αυτά παράγουμε προβλέψεις με το κατασκευασμένο MLP και επιλέγουμε αυτό με το μικρότερο σφάλμα.

Η υβριδοποίηση του μοντέλου SARIMA με τα μοντέλα MLP επιτρέπει σε κάποιον να αποκομίσει οφέλη και από τους δύο τύπους μοντέλων, αφού μέσω των μοντέλων SARIMA λαμβάνουμε υπόψη τη βραχυπρόθεσμη επίδραση παρατηρήσεων και την εποχικότητα ενώ μέσω των MLP τις μη γραμμικές δομές, όποτε υπάρχουν. Η στρατηγική εδώ είναι ο συνδυασμός ενός κλασσικού μοντέλου SARIMA με ένα νευρωνικό δίκτυο προσαρμόζοντας πρώτα ένα μοντέλο SARIMA στα αρχικά δεδομένα και στη συνέχεια κατασκευάζοντας ένα νευρωνικό δίκτυο για τα υπόλοιπα (Zhang, 2003; Chen, 2011). Η κατασκευή του νευρωνικού δικτύου για τα υπολείμματα γίνεται ακριβώς με την ίδια λογική που παρουσιάστηκε παραπάνω. Σε κάθε βήμα, οι προβλέψεις των δύο μοντέλων προστίθενται σε μία μόνο πρόβλεψη για την επόμενη περίοδο. Έτσι, η πρόβλεψη για κάθε t δίνεται από τη σχέση:

$$\hat{y}_t = \hat{y}_t^{MP} + \hat{y}_t^{Res}$$

όπου $\hat{y}_t^{MP}$ και $\hat{y}_t^{Res}$ οι προβλεπόμενες τιμές από τα μοντέλα SARIMA και ANN αντίστοιχα.

Η ακρίβεια πρόβλεψης των παραπάνω μοντέλων εκτιμάται χρησιμοποιώντας το RMSE. Όλα τα προαναφερθέντα μοντέλα εφαρμόζονται στις μηνιαίες αφίξεις διεθνών τουριστών σε όλες τις περιφέρειες της Ελλάδας από τον Ιανουάριο του 2006 έως τον Δεκέμβριο του 2016. Για την πρόβλεψη του έτους 2016 χρησιμοποιούνται δεδομένα από τον Ιανουάριο του 2006 έως τον Δεκέμβριο του 2015, ενώ για την πρόβλεψη των καλοκαιρινών μηνών του 2016, χρησιμοποιούνται δεδομένα από τον Ιανουάριο του 2006 έως τον Μάιο του 2016. Στους Πίνακες 4 και 5, παρουσιάζονται τα βέλτιστα μοντέλα για κάθε μία από τις τρεις μεθόδους καθώς επίσης και η ακρίβεια πρόβλεψης των προαναφερθέντων μοντέλων για όλες τις περιφέρειες της Ελλάδος και για τους δύο χρονικούς ορίζοντες, για ολόκληρο το επόμενο έτος και για τους καλοκαιρινούς μήνες αντίστοιχα.

Πίνακας 4. *Αποτίμηση προβλέψεων για το έτος 2016 (h=12)*

Regions	SARIMA	MLP	Hybrid SARIMA-MLP
CRE	***(1,0,0)×(1,1,0)***	***(1,10,1)***	***SARIMA-MLP (1,1,1)***
	RMSE=41081	***RMSE= 26440***	*RMSE= 36102*
SAG	***(1,0,0)×(1,1,0)***	***(2,4,1)***	***SARIMA-MLP (1,1,1)***
	RMSE=25387	***RMSE= 22904***	*RMSE= 24119*
ATT	***(0,1,0)×(1,1,0)***	***(1,10,1)***	***SARIMA-MLP (2,2,1)***
	RMSE=22676	*RMSE= 19560*	***RMSE= 18384***
CMC	***(1,0,0)×(0,1,0)***	***(2,4,1)***	***SARIMA-MLP (1,1,1)***
	RMSE=9150	***RMSE= 8583***	*RMSE= 8711*
ION	***(1,1,1)×(0,1,0)***	***(2,2,1)***	***SARIMA-MLP (1,1,1)***
	RMSE=9520	***RMSE= 6765***	*RMSE=8903*
PEL	***(1,0,0)×(1,1,0)***	***(2,2,1)***	***SARIMA-MLP (1,1,1)***
	RMSE=4874	*RMSE= 5165*	***RMSE=4676***
WGR	***(1,0,0)×(0,1,1)***	***(2,2,1)***	***SARIMA-MLP (1,10,1)***
	RMSE=4983	***RMSE= 3371***	*RMSE= 4128*
THE	***(1,0,0)×(1,1,0)***	***(2,2,1)***	***SARIMA-MLP (1,1,1)***
	RMSE=4220	*RMSE= 2942*	***RMSE= 2006***
CGR	***(2,0,1)×(0,0,2)***	***(2,2,1)***	***SARIMA-MLP (1,1,1)***
	RMSE=3559	***RMSE= 2275***	*RMSE= 3206*
EM&T	***(0,1,1)×(0,1,0)***	***(1,10,1)***	***SARIMA-MLP (1,1,1)***
	RMSE= 2177	*RMSE=3309*	***RMSE= 2034***
NAG	***(0,1,1)×(0,1,0)***	***(2,2,1)***	***SARIMA-MLP (1,1,1)***
	RMSE=11729	***RMSE= 6668***	*RMSE= 11678*
EPR	***(1,0,0)×(1,1,0)***	***(1,1,1)***	***SARIMA-MLP (1,1,1)***
	RMSE=1794	*RMSE= 1748*	***RMSE= 1595***
WMC	***(1,0,1)×(0,1,1)***	***(2,1,1)***	***SARIMA-MLP (1,1,1)***
	RMSE=315	*RMSE= 267*	***RMSE=222***
ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ	RMSE: 0	**RMSE: 7**	RMSE: 6

Πρόβλεψη ολόκληρης της σεζόν (ορίζοντας 12 μήνες) Η μεγαλύτερη ακρίβεια πρόβλεψης επιτυγχάνεται χρησιμοποιώντας κυρίως μοντέλα MLP και Hybrid SARIMA-MLP. Αναλυτικότερα, το MLP έχει μεγαλύτερη ακρίβεια πρόβλεψης στο Βόρειο Αιγαίο, τη Δυτική Ελλάδα, τα Ιόνια Νησιά, την Κρήτη, την Κεντρική Μακεδονία και την Κεντρική Ελλάδα και στο Νότιο Αιγαίο. Από την άλλη πλευρά, το Hybrid SARIMA-MLP υπερτερεί στην Αττική, τη Δυτική Μακεδονία, τη Θεσσαλία, την Ανατολική Μακεδονία, τη Θράκη και την Πελοπόννησο. Είναι σημαντικό να σημειωθεί ότι φαίνεται η προσέγγιση MLP να είναι πιο κατάλληλη για νησιωτικές περιοχές που παρουσιάζουν εντονότερη εποχικότητα ενώ το υβριδικό μοντέλο για περιφέρειες με λιγότερο έντονη εποχικότητα.

Πρόβλεψη καλοκαιρινών μηνών (ορίζοντας 3 μήνες) Για τους καλοκαιρινούς μήνες μεγαλύτερη ακρίβεια πρόβλεψης έχουμε με το μοντέλο Hybrid SARIMA-MLP, το οποίο υπερτερεί σε 11 (Ανατολική Μακεδονία και Θράκη, Αττική, Βόρειο Αιγαίο, Δυτική Μακεδονία, Ήπειρος, Θεσσαλία, Ιόνια Νησιά, Κρήτη, Νότος Αιγαίου, Πελοποννήσου και Στερεά Ελλάδα) από τις 13 περιφέρειες της Ελλάδας. Στις υπόλοιπες περιφέρειες υπερτερεί το MLP (Δυτικής Ελλάδας, Κεντρικής

Πίνακας 5. *Αποτίμηση προβλέψεων για τους καλοκαιρινούς μήνες του 2016 (h=3)*

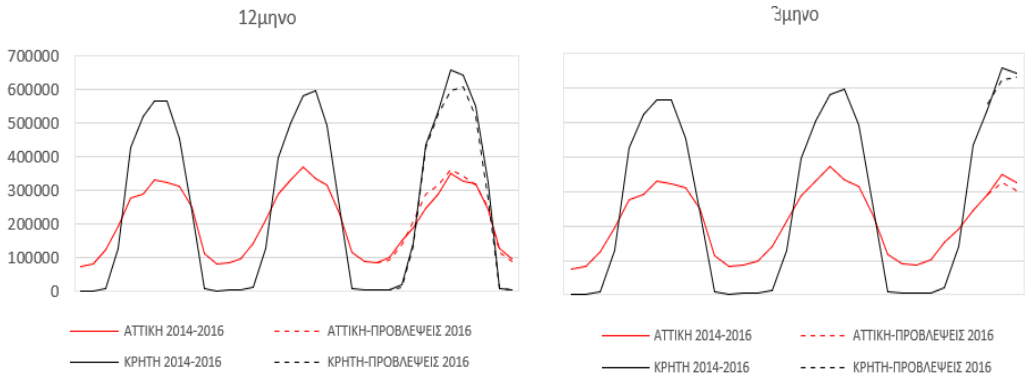
Regions	SARIMA	MLP	Hybrid SARIMA-MLP
CRE	***(1,0,0)×(1,1,0)***	***(2,15,1)***	***SARIMA-MLP (1,1,1)***
	RMSE=39628	*RMSE= 29425*	***RMSE= 19828***
SAG	***(1,0,0)×(0,1,1)***	***(1,10,1)***	***SARIMA-MLP (2,10,1)***
	RMSE=23258	*RMSE= 18934*	***RMSE= 10875***
ATT	***(1,0,0)×(2,1,1)***	***(1,1,1)***	***SARIMA-MLP (1,1,1)***
	RMSE=24361	*RMSE= 27974*	***RMSE= 20940***
CMC	***(1,0,0)×(0,1,0)***	***(1,10,1)***	***SARIMA-MLP (1,10,1)***
	RMSE=10326	***RMSE= 9678***	*RMSE=9988*
ION	***(1,1,1)×(0,1,0)***	***(2,2,1)***	***SARIMA-MLP (2,4,1)***
	RMSE=5321	*RMSE= 4273*	***RMSE= 873***
PEL	***(1,0,0)×(1,1,0)***	***(1,1,1)***	***SARIMA-MLP (1,1,1)***
	RMSE=2516	*RMSE= 5459*	***RMSE= 1217***
WGR	***(2,0,1)×(0,0,2)***	***(2,15,1)***	***SARIMA-MLP (2,2,1)***
	RMSE=8391	***RMSE= 4376***	*RMSE= 7786*
THE	***(1,0,0)×(1,1,0)***	***(2,4,1)***	***SARIMA-MLP (1,10,1)***
	RMSE=1746	*RMSE= 2016*	***RMSE= 473***
CGR	***(2,0,1)×(0,0,2)***	***(2,2,1)***	***SARIMA-MLP (1,10,1)***
	RMSE=3457	*RMSE= 2578*	***RMSE= 2469***
EM&T	***(0,1,1)×(0,1,0)***	***(2,1,1)***	***SARIMA-MLP (2,2,1)***
	RMSE= 1539	*RMSE= 3719*	***RMSE= 1490***
NAG	***(1,1,0)×(0,1,1)***	***(1,10,1)***	***SARIMA-MLP (2,10,1)***
	RMSE=4886	*RMSE= 3042*	***RMSE= 2798***
EPR	***(1,0,0)×(1,1,0)***	***(2,15,1)***	***SARIMA-MLP (2,4,1)***
	RMSE=2407	*RMSE= 2075*	***RMSE= 1328***
WMC	***(1,0,1)×(0,1,1)***	***(2,1,1)***	***SARIMA-MLP (2,10,1)***
	RMSE=480	*RMSE= 213*	***RMSE= 194***
ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ	RMSE:0	RMSE: 2	**RMSE: 11**

Μακεδονίας). Επομένως, θα μπορούσαμε να πούμε ότι η επιλογή του υβριδικού μοντέλου για πρόβλεψη των αφίξεων κατά τους καλοκαιρινούς μήνες είναι ικανοποιητική.

Στο Σχήμα 2 απεικονίζονται οι προβλέψεις από τα βέλτιστα μοντέλα μαζί με τις αντίστοιχες πραγματικές τιμές για το 2016 (αριστερά) αλλά και για μόνο τους τρεις καλοκαιρινούς μήνες του 2016 (δεξιά) ενδεικτικά για τις περιφέρειες Αττικής και Κρήτης.

Με βάση τα παραπάνω ένα ερώτημα που προκύπτει είναι αν και με ποιον τρόπο θα μπορούσαμε να αποφασίσουμε ποιο προβλεπτικό μοντέλο θα πρέπει να χρησιμοποιούμε για την πρόβλεψη της επόμενης σεζόν (12 βήματα μπροστά) γιατί δεν υπάρχει ξεκάθαρη επικράτηση κάποιου μοντέλου σε αυτή την περίπτωση. Το βασικό χαρακτηριστικό που μπορεί να μας βοηθήσει εδώ είναι η εποχικότητα που χαρακτηρίζει τα δεδομένα τουριστικών αφίξεων και η διαφοροποίηση της έντασης

Σχήμα 2. *Ενδεικτικές προβλέψεις τουριστών*



της από περιφέρεια σε περιφέρεια. Στην επόμενη ενότητα παρουσιάζεται η προτεινόμενη μεθοδολογία.

4. ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ ΜΕ ΒΑΣΗ ΤΗΝ ΕΠΟΧΙΚΟΤΗΤΑ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

Για την επιλογή μοντέλου στην περίπτωση της πρόβλεψης των αφίξεων διεθνών τουριστών για την επόμενη σεζόν (12 βήματα μπροστά) θα στηριχθούμε στην εποχικότητα που χαρακτηρίζει τα δεδομένα τουριστικών αφίξεων. Με βάση την τιμή του συντελεστή Gini (GC) (αντίστοιχα είναι τα αποτελέσματα που προκύπτουν και αν χρησιμοποιηθεί ο δείκτης CV) και χρησιμοποιώντας ανάλυση ROC, θα υπολογίσουμε το σημείο αποκοπής (μέσω του συντελεστή Youden) έτσι ώστε αν η εποχικότητα έχει τιμή κάτω από αυτό το σημείο (χαμηλή) να επιλέγεται η υβριδική προσέγγιση ενώ σε αντίθετη περίπτωση (υψηλή) να επιλέγεται ένα τυπικό τεχνητό νευρωνικό δίκτυο για πρόβλεψη. Στη συγκεκριμένη εργασία, ως πιθανές καταστάσεις θεωρούμε την επιλογή ενός από τα 2 μοντέλα (MLP και Hybrid SARIMA-MLP) και επιλέγουμε το κατάλληλο μοντέλο με βάση τον συντελεστή Youden (J) $J = \frac{a}{a+b} + \frac{b}{a+b} - 1$, όπου a σωστή επιλογή και b λάθος επιλογή μοντέλου. Χρησιμοποιώντας τον δείκτη Youden βρίσκουμε ως σημείο αποκοπής την τιμή 0.529 για τον συντελεστή Gini. Αυτό μας οδηγεί σε μια επιλογή υβριδικού μοντέλου όταν GC < 0.529 και στην επιλογή ενός καθαρού νευρωνικού δικτύου όταν GC > 0.529. Η ακρίβεια στην επιλογή μας φτάνει το 77% με λάθος επιλογή μοντέλου σε μόνο τρεις περιφέρειες (Κεντρική Μακεδονία, Δυτική Ελλάδα, Κεντρική Ελλάδα).

5. ΕΠΙΛΟΓΟΣ

Σε αυτή την εργασία μελετήθηκε το πρόβλημα της πρόβλεψης διεθνών τουριστικών αφίξεων στις περιφέρειες της Ελλάδας, με σκοπό τη ακριβέστερη πρόβλεψη της τουριστικής ζήτησης που μπορεί να οδηγήσει στον σωστότερο σχεδιασμό τουριστικών πολιτικών. Μελετήθηκαν τρία μοντέλα πρόβλεψης: τα Εποχικά ARIMA (SARIMA), τα Τεχνητά Νευρωνικά Δίκτυα (MLP) αλλά και τα υβριδικά μοντέλα τα οποία προκύπτουν από τον συνδυασμό των δύο πρώτων και αξιολογήθηκαν ως προς την προβλεπτική τους ικανότητα είτε για το επόμενο έτος είτε για την επόμενη καλοκαιρινή περίοδο Ως προς την πρόβλεψη για την επόμενη χρονιά δεν υπήρξε ένα ενιαία βέλτιστό μοντέλο για όλες τις περιφέρεις και για αυτό προτείνεται μια μεθοδολογία επιλογής καταλληλότερου μοντέλου με βάση τον συντελεστή Youden στο πλαίσιο της ανάλυσης ROC, με ιδιαίτερα ενθαρρυντικά αποτελέσματα αφού στο 77% (10/13) των περιφερειών επιλέγεται το σωστό μοντέλο.

Για την περαιτέρω ενίσχυση της προβλεπτικής ικανότητας των παραπάνω μοντέλων προτείνεται η εξέταση διαφορετικών αρχιτεκτονικών κατασκευής Τεχνητών νευρωνικών δικτύων, η χρήση διαφορετικών τεχνικών υβριδοποίησης και η ανάπτυξη οικονομετρικών μοντέλων εμπλουτισμένων με επιπλέον πληροφορίες από άλλες οικονομικές μεταβλητές σε επίπεδο περιφέρειας.

ABSTRACT

One of the most important pillars of the Greek economy is tourism, so the collection, processing, and analysis of data from this sector is of utmost importance. In the current work, the monthly occupancy of accommodation in the thirteen (13) regions of Greece is studied. A key feature of these time series is the intense seasonality. Traditional methods (SARIMA models), Artificial Neural Networks (ANNs) and Hybrid Models were used for their modeling to find the best model that will be used to predict their future prices. Applying these methods, a single optimal forecast model was not found for all regions, so a way to select a better model based on the seasonality of the data is proposed. The Gini coefficient is used to measure seasonality, while through the ROC analysis and the Youden index the cut-off point for the selection of the most suitable model is identified.

ΑΝΑΦΟΡΕΣ

Boné, R., & Crucianu, M. (2002). Multi-step-ahead prediction with neural networks: a review. *9emes rencontres internationales: Approches Connexionnistes en Sciences,* **2**, 97-106.

Box, G. & Jenkins, G. (1976). *Time Series Analysis: Forecasting and Control (2nd ed.).* San Francisco: Holden-Day.

Butler, R. (2014). Addressing seasonality in tourism: The development of a prototype. *In Conclusions and recommendations resulting from the International Conference on Managing Seasonality in Tourism, Punta del Este. UNWTO.*

Chen, K. Y. (2011). Combining linear and nonlinear model in forecasting tourism demand. *Expert Systems with Applications*, **38(8),** 10368-10376.

Dritsakis, N., & Athanasiadis, S. (2000). An econometric model of tourist demand: the case of Greece. *Journal of Hospitality and Leisure Marketing,* **7,** 39–49.

Gunter, U. & Önder, I. (2015). Forecasting international city tourism demand for Paris: Accuracy of uni- and multivariate models employing monthly data. *Tourism Management*, **46,** 123–135.

Haykin, S.S. (Ed.). (2001). *Kalman filtering and neural networks* (p. 304). New York: Wiley.

Hyndman, R., & Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software,* **27(3),** 1 - 22.

Kantz, H., & Schreiber, T. (2004). *Nonlinear time series analysis* (Vol. 7). Cambridge University Press.

Kulendran, N. & Witt, S. F. (2003). Forecasting the demand for international business tourism. *Journal of Travel Research,* **41,** 265–271.

Koutras, A., Panagopoulos, A., & Nikas, I. A. (2016). Forecasting tourism demand using linear and nonlinear prediction models. *Academica Turistica,* **9(1),** 85-102.

Lee, T. H., White, H., & Granger, C. W. (1993). Testing for neglected nonlinearity in time series models: A comparison of neural network methods and alternative tests. *Journal of Econometrics,* **56(3),** 269-290.

Lippmann, R. (1987). An introduction to computing with neural nets. *IEEE Assp magazine,* **4(2),** 4-22.

Olmedo, E. (2016). Comparison of near neighbour and neural network in travel forecasting. *Journal of Forecasting,* **35(3),** 217-223.

Shan, J., & Wilson, K. (2001). Causality between trade and tourism: empirical evidence from China. *Applied Economics Letters,* **8,** 279–283.

Song, H., & Li, G. (2008). Tourism demand modelling and forecasting – A review of recent research. *Tourism Management,* **29,** 203–220.

Song, H., & Witt, S. F. (2006). Forecasting international tourist flows to Macau. *Tourism Management,* **27**, 214–224.

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing,* **50,** 159-175.

Þórhallsdóttir, G., & Ólafsson, R. (2017). A method to analyse seasonality in the distribution of tourists in Iceland. *Journal of Outdoor Recreation and Tourism,* **19,** 17-24.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά του 33ου Πανελλήνιου Συνεδρίου Στατιστικής (2021), pp. 108–120

Συνελίξεις ακολουθιών πινάκων σε σε χρονικά πολυδιάστατες Μαρκοβιανές ανανεωτικές αλυσίδες

Λ. Κορδαλής[1]***, Σ. Τρέβεζας***[1]
[1]Τμήμα Μαθηματικών, ΕΚΠΑ
{lkordali, strevezas}@math.uoa.gr

Περίληψη

Στην παρούσα εργασία εξετάζονται οι ιδιότητες των συνελίξεων όταν ο χρόνος είναι διακριτός και πολυδιάστατος, ενώ με τη βοήθεια αλγεβρικών δομών εξετάζεται η ύπαρξη και δίνεται ο ακριβής τύπος του συνελιξιακού αντιστρόφου. Ο τελευταίος θα διαδραματίσει ιδιαίτερο ρόλο στην κατασκευή μίας επέκτασης της θεωρίας της κλάσης των Μαρκοβιανών ανανεωτικών αλυσίδων όπου η έννοια του χρόνου θα έχει πολυδιάστατη μορφή. Συγκεκριμένα, θα χρησιμοποιηθούν συνελιξιακές πράξεις για την εύρεση λύσης των Μαρκοβιανών ανανεωτικών εξισώσεων και ιδιαίτερα στην αναζήτηση αναπαραστάσεων ακολουθιών πινάκων οι οποίοι είναι σημαντικοί για τη θεμελίωση και την εξέλιξη αυτής της θεωρίας.

Λέξεις-Κλειδιά: Συνελίξεις πολυδιάστατου χρόνου, Συνελιξιακός Αντίστροφος, Χρονικά πολυδιάστατη Μαρκοβιανή ανανεωτική θεωρία, Μαρκοβιανές ανανεωτικές εξισώσεις

1. Εισαγωγή

Οι συνελίξεις διακριτού χρόνου αποτελούν θεμέλιο λίθο στις εφαρμογές της Θεωρίας πιθανοτήτων όπως είναι η θεωρία των ανανεωτικών και των Μαρκοβιανών ανανεωτικών διαδικασιών. Το σημείο εκκίνησης αυτής της μελέτης μπορεί να βρεθεί στους Barbu and Limnios (2009), οι οποίοι χρησιμοποίησαν συνελίξεις διακριτού χρόνου για ακολουθίες πινάκων μίας μεταβλητής για τη λύση ανανεωτικών και Μαρκοβιανών εξισώσεων. Ιδιαίτερο ρόλο στη συγκεκριμένη θεωρία, έχει ο αριστερός αντίστροφος μίας ακολουθίας πινάκων και ο υπολογισμός του γίνεται βάσει αλγοριθμικών τεχνικών. Στη Μαρκοβιανή ανανεωτική θεωρία χρησιμοποιείται ο συνελιξιακός αντίστροφος για τον υπολογισμό της συνάρτησης πιθανοτήτων μετάβασης μίας ημιμαρκοβιανής αλυσίδας και για την ανάπτυξη ημιμαρκοβιανών συστημάτων αξιοπιστίας.

Στην παρούσα μελέτη θα χρησιμοποιηθούν συνελίξεις πολυδιάστατου διακριτού χρόνου κάνοντας σύνδεση με συγκεκριμένες αλγεβρικές έννοιες όπως είναι ο δακτύ-

λιος, με σκοπό την ανάπτυξη νέων αναπαραστάσεων και πιο συγκεκριμενα τη μελέτη της ύπαρξης, του τύπου και των ιδιοτήτων του συνελιξιακού αντιστρόφου. Ακόμη, θα εφαρμοστεί η συγκεκριμένη θεωρία με σκοπό την ανάπτυξη νέων Μαρκοβιανών ανανεωτικών μοντέλων, όπου ο χρόνος θα είναι διακριτός αλλά και πολυδιάστατος. Επιπλέον, θα αναδειχθούν ιδιότητες και τύποι ειδικών συναρτήσεων οι οποίες είναι σημαντικές για την ανάπτυξη της θεωρίας που βασίζεται στη νέα αυτή ιδέα.

Οι Μαρκοβιανές ανανεωτικές αλυσίδες (μ.α.α) και οι επαγόμενες ημι-Μαρκοβιανές αλυσίδες (ημ.α) έχουν μελετηθεί εκτενώς στο κλασικό τους πλαίσιο και αποτελούν πεδίο διάφορων εφαρμογών (π.χ: Pyke and Schaufele (1964), Barbu and Limnios (2009), Limnios and Oprisan (2012)).

Η πιθανοτική μοντελοποίηση στο διακριτό πλαίσιο δεν είχε λάβει την αντίστοιχη προσοχή όπως η αντίστοιχη στο συνεχές, παρόλο που οι πρώτες εργασίες στο διακριτό χρόνο ξεκινάνε από τη δεκαετία του 1960. Ο Anselone (1960) ανέπτυξε ορισμένα στοιχεία εργοδικής θεωρίας των Μαρκοβιανών ανανεωτικών αλυσίδων. Στην περίπτωση μη ικανοποίησης της εργοδικότητας, μιας Μαρκοβιανής ανανεωτικής αλυσίδας η οποία απορροφάται σε μία συγκεκριμένη κατάσταση, ο Gerontidis (1994) μελέτησε την οριακή συμπεριφορά Μαρκοβιανών ανανεωτικών αλυσίδων αντικατάστασης, όπου τα ανεξάρτητα δείγματα συνδέονται όταν συμβαίνει η απορρόφηση. Εφαρμογές των Μαρκοβιανών Ανανεωτικών αλυσίδων με έμφαση στις διαδικασίες αποφάσεων δίνονται από τον Howard (1971).

Ακόμη, μπορούμε να εντοπίσουμε εφαρμογές στη Θεωρία ουρών αναμονής και στην επιδημιολογία καθώς και χρήση αλγορίθμων για την εφαρμογή των Μ.α.α σε ηλεκτρονικό υπολογιστή στις εργασίες των Mode and Pickens (1988) και Mode and Sleeman (2000) . Το πρώτο υπολογιστικό πακέτο το οποίο συνδυάζει προσομοιώσεις και εκτιμήσεις ορισμένων γνωστών παραμετρικών οικογενειών κατανομών, εντοπίζεται στην εργασία των Barbu et al. (2018). Οι πιο πρόσφατες αναφορές σε θεωρία και εφαρμογές (π.χ Θεωρία αξιοπιστίας) των Μ.α.α και ημ.α, εντοπίζονται στους Barbu and Limnios (2009), and Barbu et al. (2018).

Σε αυτή την εργασία, εισάγουμε τις χρονικά πολυδιάστατες μ.α.α. Είναι μία επέκταση της κλασικής Μαρκοβιανής ανανεωτικής θεωρίας, όπου η έννοια του χρόνου μπορεί να είναι πολυδιάστατη. Κίνητρο για τη συγκεκριμένη μελέτη αποτελούν οι μελέτες των μοντέλων ανάπτυξης φυτών, στα οποία ο θερμικός χρόνος είναι ο πλέον κατάλληλος για την περιγραφή της κατάστασης ενός φυτού. Ο συνδυασμός διαφορετικών χρόνων μπορεί να μας δώσει μία πιο βολική καταγραφή της στοχαστικής εξέλιξης μίας κατάστασης.

Η παρούσα εργασία μπορεί να θεωρηθεί επίσης ως μία διακριτή γενίκευση των διδιάστατων ανανεωτικών διαδικασιών οι οποίες αναφέρονται στις εργασίες Hunter (1974a) και Hunter (1974b), όπου ο συγγραφέας όρισε και έλυσε ανανεωτικές εξισώσεις και μελέτησε την ασυμπτωτική συμπεριφορά χρονικά πολυδιάστατων ανανεωτικών μοντέλων. Ο Mallor et al. (2007) έδειξε τις ασυμπτωτικές ιδιότητες των χρονικά πολυδιάστατων σταθμισμένων ανανεωτικών συναρτήσεων. Επίσης, εστιάζουν σε πολυμεταβλητούς μετασχηματισμούς Laplace και στην οριακή συμπεριφορά των

ανανεωτικών διαδικασιών και ειδικότερα στη συσχέτιση μεταξύ των χρονικών συνιστωσών.

Αρχικά, θα αναλυθεί το πλαίσιο των συνελίξεων ακολουθιών πινάκων διακριτού πολυδιάστατου χρόνου και θα μελετηθούν οι ιδιότητες συγκεκριμένων κλάσεων ακολουθιών με κύριο σκοπό στην αναπαράσταση του συνελιξιακού αντιστρόφου, όταν αυτός υφίσταται. Στο τελευταίο κεφάλαιο θα γίνει η εφαρμογή τους στις χρονικά πολυδιάστατες Μαρκοβιανές ανανεωτικές αλυσίδες με σκοπό τη θεωρητική τους θεμελίωση και ιδιαίτερα στη μελέτη χρονικά πολυδιάστατων μαρκοβιανών εξισώσεων και στην κατασκευή της επαγόμενης χρονικά πολυδιάστατης ημιμαρκοβιανής αλυσίδας.

2. Συνελίξεις διακριτού χρόνου

Στην παρούσα παράγραφο θα μελετήσουμε το συνελιξιακό γινόμενο μιας συγκεκριμένης κατηγορίας πινακικών συναρτήσεων και θα δώσουμε συγκεκριμένες αλγεβρικές ιδιότητες του συνελιξιακού τελεστή, δίνοντας ιδιαίτερη προσοχή στον συνελιξιακό αντίστροφο. Για την εισαγωγή των χρονικά πολυδιάστατων συνελίξεων μεταξύ πινακικών ακολουθιών, θα χρειαστούμε κάποιους ορισμούς.

Θεωρούμε ένα πεπερασμένο σύνολο $E = \{1, 2, \ldots, s\}$ και γράφουμε ως $\mathscr{M}_s := \mathbb{R}^{s\times s}$ το σύνολο όλων των πινάκων στο $E \times E$ και ως $\mathscr{M}_s(\mathbb{N}^d)$ το σύνολο των πινακικών ακολουθιών με πεδίο ορισμού το $\mathbb{N}^d$ και πεδίο τιμών το $\mathscr{M}_s$. Για κάθε $A \in \mathscr{M}_s(\mathbb{N}^d)$, θα γράφουμε $A := \big(A(k_{1:d}); k_{1:d} \in \mathbb{N}^d\big)$, όπου για κάποιο σταθερό $k_{1:d} \in \mathbb{N}^d$, παίρνουμε τον πίνακα $A(k_{1:d}) = (a_{ij}(k_{1:d}))_{i,j\in E}$. Γράφουμε ως 0_d το ουδέτερο στοιχείο του $\mathbb{R}^d$. Επίσης, συμβολίζουμε με I_s και $\mathbb{O}_s$ τον ταυτοτικό και το μηδενικό πίνακα του $\mathscr{M}_s$ αντίστοιχα.

Το στοιχείο $A = (a_{ij})_{i,j\in E}$ μπορεί να ερμηνευθεί ως πίνακας που περιέχει d-διάστατες ακολουθίες, όπου $a_{ij} \in R \triangleq \mathbb{R}^{\mathbb{N}^d}$. Με αυτό το σκεπτικό $A \in R^{s\times s}$, κι έτσι αντιστοιχεί σε έναν πίνακα με στοιχεία από κάποιον μεταθετικό δακτύλιο. Αυτή η παρατήρηση θα ελεγχθεί στη συνέχεια.

Ορισμός 1. *Έστω $A, B \in \mathscr{M}_s(\mathbb{N}^d)$. Η πινακοσυνάρτηση $A * B \in \mathscr{M}_s(\mathbb{N}^d)$, η οποία δίνεται από:*

$$[A * B]\,(k_{1:d}) := \sum_{l+l'=k_{1:d}} A(l)B(l'), \quad k_{1:d} \in \mathbb{N}^d,$$

και έχει στοιχεία

$$[A * B]_{ij}\,(k_{1:d}) := \sum_{r\in E} \sum_{l+l'=k_{1:d}} \alpha_{ir}(l)\beta_{rj}(l'), \quad i, j \in E, k_{1:d} \in \mathbb{N}^d,$$

καλείται d-διάστατο διακριτό συνελιξιακό γινόμενο των A και B.

Σημείωση 1. *Στον παραπάνω ορισμό δόθηκε έμφαση στην ερμηνεία του $A * B$ ως μία d-διάστατη ακολουθία πινάκων η οποία προκύπτει από τη συνέλιξη ακολουθιών πινάκων πολλών μεταβλητών. Παρ'όλα αυτά θα είναι το ίδιο χρήσιμη η ερμηνεία του $A * B$ ως πίνακα*

πολυδιάστατων ακολουθιών που προέρχονται από τον πολλαπλασιασμό των πινάκων A και B. Για την ακρίβεια, $A*B=([A*B]_{ij})_{i,j\in E}$, όπου

$$[A*B]_{ij}=\sum_{r\in E}\alpha_{ir}*\beta_{rj}.$$

Με αυτόν τον τρόπο, η συνέλιξη δύο πινάκων στο σύνολο $\mathscr{M}_s(\mathbb{N}^d)$ αντιστοιχεί στο συνήθη πολλαπλασιασμό πινάκων στον $R^{s\times s}$, όπου ως πολλαπλασιαστική πράξη στο R θεωρείται η συνέλιξη μεταξύ ακολουθιών d-διάστατου χρόνου.

Σημαντικό ρόλο για την ανάπτυξη της θεωρίας θα παίξει η μοναδιαία συνάρτηση (η ακολουθία που είναι ταυτοτικά 1) την οποία θα συμβολίζουμε με $\mathbb{1}$. Επίσης, ορίζουμε την ακολουθία $\mathbb{I}=\operatorname{diag}\{\mathbb{1}\}$. Για $s=1$, παίρνουμε άμεσα ότι $\mathbb{I}=\mathbb{1}$. Επιπλέον, η ακολουθία $\mathbb{I}$ αντιστοιχεί στον προσθετικό τελεστή, δηλαδή:

$$[A*\mathbb{I}](k_{1:d})=\sum_{l=0_d}^{k_{1:d}}A(l),\quad k_{1:d}\in\mathbb{N}^d.$$

Στην επόμενη πρόταση, θα δώσουμε μερικές αλγεβρικές ιδιότητες του συνελιξιακού γινομένου μεταξύ ακολουθιών πινάκων, οι οποίες μπορούν να αποδειχθούν εύκολα:

Πρόταση 1. *Ο συνελιξιακός τελεστής d-διάστατου διακριτού χρόνου είναι προσεταιριστικός, επιμεριστικός ως προς τη συνήθη πρόσθεση πινάκων και περιέχει το μοναδιαίο στοιχείο e_0, το οποίο δίνεται από:*

$$e_0(k_{1:d})=\begin{cases} I_s, & \text{αν } k_{1:d}=0_d,\\ \mathbb{O}_s, & \text{διαφορετικά.}\end{cases}$$

και είναι μεταθετικός ανν $s=1$.

Συνεπώς, ο χώρος $\mathscr{M}_s(\mathbb{N}^d)$, εφοδιασμένος με τη συνήθη πράξη της πρόσθεσης "+" και του πολλαπλασιαστικού τελεστή "*" αποτελεί ένα μη-μεταθετικό δακτύλιο με μονάδα. Επιπλέον, το ζεύγος $(\mathscr{M}_s(\mathbb{N}^d),*)$ είναι ένα μονοειδές.

Παρακάτω, εισάγουμε τις δυνάμεις ακολουθιών πινάκων μέσω της πράξης της συνέλιξης.

Ορισμός 2. *Έστω $A\in\mathscr{M}_s(\mathbb{N}^d)$ μία πινακοσυνάρτηση και $n\in\mathbb{N}$. Η n-οστή συνελιξιακή δύναμη $A^{(n)}$ είναι η ακολουθία πινάκων η οποία ορίζεται από τον ακόλουθο τύπο:*

$$\begin{aligned} A^{(0)} &:= e_0\\ A^{(n)} &:= A*A^{(n-1)},\quad n\geq 1.\end{aligned}$$

Από την παραπάνω έκφραση παίρνουμε άμεσα ότι για κάθε $n\geq 0$ και $k_{1:d}\in\mathbb{N}^d$:

$$A^{(n)}(k_{1:d}):=\sum_{\substack{l_1,\ldots,l_n\in\mathbb{N}^d\\ l_1+\ldots+l_n=k_{1:d}}}A(l_1)\cdots A(l_n).\tag{1}$$

Στην ακόλουθη πρόταση, δίνουμε μία χρήσιμη ιδιότητα για πινακικές ακολουθίες A με $A(0) = \mathbb{O}_s$, η οποία θα χρησιμοποιηθεί και στη συνέχεια της εργασίας.

Πρόταση 2. *Έστω $A \in \mathscr{M}_s(\mathbb{N}^d)$ μία ακολουθία πινάκων με $A(0_d) = \mathbb{O}_s$. Τότε, ισχύει $A^{(n)}(k_{1:d}) = \mathbb{O}_s$ για κάθε $k_1, \ldots, k_d, n \in \mathbb{N}$ με $k_1 + \cdots + k_d < n$.*

Απόδειξη. Από τη σχέση (1), έχουμε ότι

$$A^{(n)}(k_{1:d}) = \sum_{\substack{l_1,\ldots,l_n \in \mathbb{N}^d \\ l_1+l_2+\ldots+l_n = k_{1:d}}} A(l_1)A(l_2)\cdots A(l_n).$$

Επειδή, $A(0_d) = \mathbb{O}_s$, θα έχουμε τουλάχιστον ένα $l_i = 0_d$ και τότε το παραπάνω γινόμενο πινάκων θα είναι μηδενικό. Πράγματι, αυτό είναι αληθές όταν $k_1 + \cdots + k_d < n$ και συνεπώς έχουμε ότι $A^{(n)}(k_{1:d}) = \mathbb{O}_s$. □

Σημείωση 2. *Μία άμεση συνέπεια της παραπάνω πρότασης ειναι ότι για κάθε $k_{1:d}$, η ακολουθία $A^{(n)}(k_{1:d})$ είναι τελικά μηδενική. Επιπροσθέτως, η πινακική ακολουθία η οποία ορίζεται από τη σειρά $\sum_n A^{(n)}$ είναι καλά ορισμένη, απο τη στιγμή που κάθε στοιχείο της είναι ένα πεπερασμένο άθροισμα πινάκων.*

Ορισμός 3. *Έστω $A \in \mathscr{M}_s(\mathbb{N}^d)$. Αν υπάρχει $B \in \mathscr{M}_s(\mathbb{N}^d)$ τέτοιο ώστε*

$$A * B = e_0,$$

τότε ο B αποκαλείται δεξιός συνελιξιακός αντίστροφος του A και γράφεται ως $A_r^{(-1)}$. Αν υπάρχει ακολουθία $C \in \mathscr{M}_s(\mathbb{N}^d)$ ώστε

$$C * A = e_0, \tag{2}$$

τότε η ακολουθία C καλείται αριστερός συνελιξιακός αντίστροφος του A και συμβολίζεται με $A_l^{(-1)}$.

Επειδή $\left(\mathscr{M}_s(\mathbb{N}^d), *\right)$ είναι ένα μονοειδές, μπορούμε να συμπεράνουμε ότι οι πινακικές ακολουθίες $A_r^{(-1)}$ και $A_l^{(-1)}$ είτε θα υπάρχουν και οι δύο, είτε καμία. Στην περίπτωση που υπάρχουν τότε θα ταυτίζονται. Επομένως, υπάρχει μοναδικό $B \in \mathscr{M}_s(\mathbb{N}^d)$ ώστε

$$A * B = B * A = e_0, \tag{3}$$

και ο B αποκαλείται συνελιξιακός αντίστροφος του A και συμβολίζεται ως $A^{(-1)}$.

Ο αντίστροφος μίας ακολουθίας πινάκων A δεν υπάρχει πάντα. Για παράδειγμα, θεωρούμε κάποιον $A \in \mathscr{M}_s(\mathbb{N}^d)$, με $A(0_d) = \mathbb{O}_s$. Αν ο A είχε αντίστροφο, τότε θα βγάζαμε άμεσα ότι:

$$\mathbb{O}_s = B(0_d)A(0_d) = [B * A](0_d) = e_0(0_d) = I_s.$$

Επομένως, ο συνελιξιακός αντίστροφος των πινακικών ακολουθιών ορίζεται μόνο υπό συγκεκριμένες υποθέσεις, οι οποίες δίνονται παρακάτω:

Θεώρημα 1. *Ο συνελιξιακός αντίστροφος μιας πινακικής ακολουθίας $A \in \mathscr{M}_s\left(\mathbb{N}^d\right)$ υπάρχει όταν και μόνο όταν ο $A(0_d)$ είναι αντιστρέψιμος πίνακας.*

Απόδειξη. Τα σχόλια κάτω από τον Ορισμό 3 υποδεικνύουν άμεσα το γεγονός ότι αν $A \in \mathscr{M}_s(\mathbb{N}^d)$ με $A(0_d)$ να είναι μη αντιστρέψιμος, τότε δεν υπάρχει $B \in \mathscr{M}_s(\mathbb{N}^d)$ που να ικανοποιεί τη σχέση (3). Από την άλλη, αν $A(0_d)$ είναι αντιστρέψιμος, τότε εύκολα φτάνουμε στην αντιστρεψιμότητα του A. □

Θεώρημα 2. *Έστω $A \in \mathscr{M}_s(\mathbb{N}^d)$ και $A(0_d)$ αντιστρέψιμος. Ο συνελιξιακός αντίστροφος $A^{(-1)}$ δίνεται από την ακόλουθη σχέση*

$$A^{(-1)} = [A(0_d)]^{-1}\left[\sum_{m=0}^{\infty}(e_0 - A_0)^{(m)}\right] = \left[\sum_{m=0}^{\infty}(e_0 - A_0^*)^{(m)}\right][A(0_d)]^{-1} \tag{4}$$

όπου,

$$\begin{aligned} A_0(k_{1:d}) &= A(k_{1:d})\,[A(0_d)]^{-1}, \quad k_{1:d} \in \mathbb{N}^d. \\ A_0^*(k_{1:d}) &= [A(0_d)]^{-1}A(k_{1:d}), \quad k_{1:d} \in \mathbb{N}^d. \end{aligned}$$

Επιπλέον, κάθε στοιχείο $A^{(-1)}(k_{1:d})$ της (4) μπορεί να αναπαρασταθεί ως πεπερασμένο άθροισμα

$$A^{(-1)}(k_{1:d}) = [A(0_d)]^{-1}\left[\sum_{m=0}^{k_1+\ldots+k_d}(e_0 - A_0)^{(m)}(k_{1:d})\right] \tag{5}$$

Απόδειξη. Θα δείξουμε ότι το μεσαίο μέρος της (5) ισχύει. Η ισότητα η οποία πηγάζει από το δεξιό μέρος μπορεί να αποδειχθεί με αντίστοιχο τρόπο. Για να δείξουμε τη ζητούμενη σχέση, θα θεωρήσουμε χωρίς βλάβη της γενικότητας ότι $A(0) = I_s$ και συνεπώς αρκεί να αποδείξουμε ότι

$$A^{(-1)} = \sum_{m=0}^{\infty}(e_0 - A)^{(m)}, \tag{6}$$

διότι στη γενική περίπτωση θα έχουμε $A = A_0A(0_d)$ και επειδή $A_0(0_d) = I_s$ μπορούμε να πάρουμε εύκολα ότι το αποτέλεσμα ισχύει για $A_0^{(-1)}$, τότε $A^{(-1)} = (A(0_d))^{-1}A_0^{(-1)}$. Για να δείξουμε τη σχέση (6), αρκεί να αποδείξουμε ότι η συνέλιξη μεταξύ της A ικανοποιούν το αριστερό και μεσαίο μέλος της (3). Πράγματι, επειδή $[e_0 - A](0_d) = \mathbb{O}_s$ τότε η συνθήκη της Σημείωσης 2 ικανοποείται και συνεπώς παίρνουμε:

$$\begin{aligned} \left(\sum_{m=0}^{\infty}(e_0 - A)^{(m)}\right) * A &= \left(\sum_{m=0}^{\infty}(e_0 - A)^{(m)}\right) * [e_0 - (e_0 - A)] \\ &= \sum_{m=0}^{\infty}(e_0 - A)^{(m)} - \sum_{m=0}^{\infty}(e_0 - A)^{(m+1)} \\ &= (e_0 - A)^{(0)} = e_0. \end{aligned}$$

Το μεσαίο μέλος της (3) αποδεικνύεται με παρόμοιο τρόπο. Επιπλέον, από το γεγονός ότι $(e_0 - A)(0_d) = \mathbb{O}_s$, θα έχουμε από την Πρόταση 2, ότι $(e_0 - A)^{(m)}(k_{1:d}) = \mathbb{O}_s$, για κάθε $k_1 + \ldots + k_d < m$, και συνεπώς η (5) ισχύει. □

3. Χρονικά πολυδιάστατες Μαρκοβιανές ανανεωτικές αλυσίδες

Θεωρούμε ένα τυχαίο σύστημα με πεπερασμένο χώρο καταστάσεων $E = \{1, ..., s\}$. Υποθέτουμε, επίσης, ότι το σύστημα αυτό εξελίσσεται με άλματα στο d-διάστατο χρόνο και μία διαδικασία (J_n) η οποία καταγράφει τις επισκεπτόμενες καταστάσεις. Στην παρούσα εργασία θα επιτρέψουμε άλματα στην ίδια κατάσταση, επομένως οι ανανεωτικές αλυσίδες θεωρούνται ειδική περίπτωση των Μαρκοβιανών ανανεωτικών αλυσίδων για $s = 1$.

Οι χρόνοι των αλμάτων θα καταγράφονται από μία d-διάστατη διαδικασία $(S_n) = (S_n^u)_{1 \le u \le d}$ με τιμές στο $\mathbb{N}^d$ η οποία αντιστοιχεί στη d-διάστατη διακριτή χρονική περιοχή. Αρχικά, θεωρούμε ότι $S_0 = 0$. Ο χώρος $\mathbb{N}^d$ εφοδιάζεται με τη μερική διάταξη $k \le u$, όταν και μόνο όταν $k_i \le u_i$ για κάθε $1 \le i \le d$, και θα γράφουμε $k < u$ όταν η γνήσια ανισότητα ισύχει για τουλάχιστον μία από τις συνιστώσες τους. Επίσης, θεωρούμε ότι όταν πραγματοποείται κάποιο άλμα τότε η (S_n) θα αυξάνεται γνήσια, έτσι $S_n < S_{n+1}$ για κάθε $n \in \mathbb{N}$. Από την άλλη, αυτή η ανισότητα δεν ισχύει απαραίτητα για τις υπόλοιπες συνιστώσες της (μηδενικοί χρόνοι γεγονότων). Ακόμη, ορίζουμε την ακολουθία των ενδιάμεσων χρόνων γεγονότων $(X_n) = (X_n^u)_{1 \le u \le d}$. Θεωρούμε ότι $X_0 = 0$ και $X_n = S_n - S_{n-1}$, για κάθε $n \ge 1$. Ορίζουμε την απαριθμήτρια διαδικασία η οποία καταγράφει το πλήθος αλμάτων που έχουν συμβεί σε χρονικές στιγμές που ανήκουν στο καρτεσιανό γινόμενο διαστημάτων $\prod_{i=1}^d [0, k_i]$

$$N(k_{1:d}) := \sup\{n \in \mathbb{N} : S_n \le k_{1:d}\},$$

και αν η N^u είναι η περιθώρια απαριθμήτρια διαδικασία η οποία είναι εμφυτευμένη στην S^u:

$$N^u(k_u) := \sup\{n \in \mathbb{N} : S_n^u \le k_u\},$$

τότε προκύπτει εύκολα η παρακάτω αναπαράσταση της N

$$N(k_{1:d}) = \min_{1 \le u \le d} \{N^u(k_u)\}, k_{1:d} \in \mathbb{N}^d.$$

.

Μία πολυδιάστατη επέκταση του ημιμαρκοβιανού πυρήνα ορίζεται ακολούθως:

Ορισμός 4. *Η ακολουθία πινάκων $q = \left(q_{ij}(k_{1:d})\right) \in \mathscr{M}_s(\mathbb{N}^d)$ ονομάζεται χρονικά d-διάστατος ημιμαρκοβιανός πυρήνας όταν ικανοποιεί τις ακόλουθες δύο συνθήκες:*

(*i*) $q_{ij}(k_{1:d}) \ge 0,\ i, j \in E,\ k_{1:d} \in \mathbb{N}^d,$

(*ii*) $\sum_j \sum_{k_{1:d}} q_{ij}(k_{1:d}) = 1,\ i \in E.$

Αν επιπλέον, ισχύει το παρακάτω

(iii) $q_{ij}(0_d) = 0,\ i, j \in E,$
τότε οι d-διάστατοι μηδενικοί χρόνοι μεταβάσεων δε θα επιτρέπονται.

Σημείωση 3. *Αν ορίσουμε την ακολουθία πινάκων $q^u = (q^u(k))_{k\geq 0} \in \mathscr{M}_s(\mathbb{N})$, όπου*

$$q^u_{ij}(k_u) = \sum_{k_{1:u-1}} \sum_{k_{u+1:d}} q_{ij}(k_{1:u-1}, k_u, k_{u+1:d}), \tag{7}$$

τότε παίρνουμε άμεσα ότι για $1 \leq u \leq d$, η ακολουθία q^u είναι ένας ημιμαρκοβιανός πυρήνας ο οποίος πιθανόν να επιτρέπει μεταβάσεις μηδενικών χρόνων.

Επίσης, ορίζουμε τον αθροιστικό ημιμαρκοβιανό πυρήνα $Q \in \mathscr{M}_s(\mathbb{N}^d)$, μέσω του q:

$$Q = \mathbb{I} * q.$$

Παρακάτω, δίνεται ο ορισμός των χρονικά πολυδιάστατων Μαρκοβιανών ανανεωτικών αλυσίδων.

Ορισμός 5. *Η αλυσίδα $(J,S) := (J_n, S_n)_{n\in\mathbb{N}}$ ονομάζεται χρονικά πολυδιάστατη (d-διάστατη) Μαρκοβιανή ανανεωτική αλυσίδα (χ.π.μ.α.α), όταν για κάθε $n \in \mathbb{N}$, $i, j \in E$ and $k_{1:d} \in \mathbb{N}^d$, ικανοποιεί σ.β.*

$$\mathbb{P}(J_{n+1} = j, S_{n+1} - S_n = k_{1:d} \,|\, J_{0:n}, S_{0:n}) = \mathbb{P}(J_{n+1} = j, S_{n+1} - S_n = k_{1:d} \,|\, J_n). \tag{8}$$

Επίσης, αν η (8) είναι ανεξάρτητη του n, τότε η (J,S) ονομάζεται χρονικά ομογενής και τότε:

$$q_{ij}(k_{1:d}) := \mathbb{P}(J_{n+1} = j, S_{n+1} - S_n = k_{1:d} \,|\, J_n = i).$$

Η ακολουθία $q = \big(q_{ij}(k_{1:d})\big) \in \mathscr{M}_s(\mathbb{N}^d)$ ικανοποιεί τις συνθήκες του ημιμαρκοβιανού πυρήνα και θα αναφέρεται ως ο επαγόμενος ημιμαρκοβιανός πυρήνας.

Από μία χ.π.μ.α.α (J,S), μπορούμε να βρούμε χρήσιμες Μαρκοβιανές αλυσίδες.

Πρόταση 3. *Αν η (J,S) είναι μία χ.π.μ.α.α και ο q είναι ο επαγόμενος ημιμαρκοβιανός πυρήνας, τότε οι διαδικασίες (J,S), (J,S^u), (J,X) και (J,X^u) είναι Μαρκοβιανές αλυσίδες και οι αντίστοιχες πιθανότητες μεταβάσεων, είναι:*

$$\begin{aligned}
\mathbb{P}(J_{n+1} = j, S_{n+1} = s_{1:d} + k_{1:d} \,|\, J_n = i, S_n = s_{1:d}) &= q_{ij}(k_{1:d}),\\
\mathbb{P}(J_{n+1} = j, S^u_{n+1} = s_u + k_u \,|\, J_n = i, S^u_n = s_u) &= q^u_{ij}(k_u),\\
\mathbb{P}(J_{n+1} = j, X_{n+1} = k_{1:d} \,|\, J_n = i, X_n = k'_{1:d}) &= q_{ij}(k_{1:d}),\\
\mathbb{P}(J_{n+1} = j, X^u_{n+1} = k_u \,|\, J_n = i, X^u_n = k'_u) &= q^u_{ij}(k_u),
\end{aligned}$$

όπου η $q^u_{ij}(k_u)$ ορίζεται στη σχέση (7). Η αλυσίδα (J,S^u) είναι μία Μαρκοβιανή ανανεωτική αλυσίδα με επαγόμενο ημιμαρκοβιανό πυρήνα q^u. Η διαδικασία J είναι επίσης μία μαρκοβιανή αλυσίδα, η οποία αποκαλείται εμφυτευμένη μαρκοβιανή αλυσίδα, επαγόμενη στη (J,S), όπως και στη μαρκοβιανή ανανεωτική αλυσίδα (J,S^u). Οι επαγόμενες πιθανότητες μετάβασης, δίνονται από:

$$p_{ij} := \mathbb{P}(J_{n+1} = j \,|\, J_n = i) = \sum_{k_{1:d}} q_{ij}(k_{1:d}).$$

Για να μελετήσουμε το πιθανοτικό πλαίσιο των χρονικά πολυδιάστατων μ.α.α, χρειάζεται να ορίσουμε δύο τύπους κατανομών, αυτής των χρόνων παραμονής και των αντίστοιχων δεσμευμένων κατανομών όταν είναι γνωστή η επόμενη κατάσταση.

Ορισμός 6. *Για μία χ.π.μ.α.α* (J,S), $i,j \in E$, $k_{1:d} \in \mathbb{N}^d$ *και* $k_u \in \mathbb{N}$, *για* $1 \leq u \leq d$, *εισάγουμε:*

(i) *την από κοινού συνάρτηση μάζας πιθανότητας του χρόνου παραμονής:*

$$h_i(k_{1:d}) = \mathbb{P}(X_{n+1} = k_{1:d} \mid J_n = i),$$

(ii) *τη συνάρτηση κατανομής του χρόνου παραμονής:*

$$H_i = \mathbb{1} * h_i,$$

(iii) *το συμπλήρωμα της συνάρτησης κατανομής στην κατάταση* i

$$\widetilde{H}_i = \mathbb{1} - H_i = \mathbb{1} - \mathbb{1} * h_i,$$

Σημείωση 4. *Η κατανομή των χρόνων παραμονής σε μία κατάσταση* i *μπορεί να γραφεί μέσω της* Q *ως:*

$$H_i(k_{1:d}) = \sum_{j \in E} Q_{ij}(k_{1:d})$$

Στη συνέχεια αυτής της εργασίας, θα παρουσιάσουμε μία επέκταση της χρονικά μονοδιάστατης ημιμαρκοβιανής αλυσίδας σε μία πολυδιάστατη έννοια του χρόνου και θα μελετήσουμε μερικά πιθανοτικά χαρακτηριστικά. Το τελευταίο, βασίζεται σε μία φυσική επέκταση μέσω των μαρκοβιανών ανανεωτικών εξισώσεων.

Στην ανάπτυξη της θεωρίας, θα παίξει σημαντικό ρόλο η ακολουθία πινάκων $e_0 - q$ και στην ακόλουθη πρόταση δίνουμε τη μορφή του συνελιξιακού αντιστρόφου της.

Πρόταση 4. *Ο συνελιξιακός αντίστροφος της ακολουθίας* $e_0 - q$ *υπάρχει και δίνεται μέσω του τύπου*

$$u := (e_0 - q)^{(-1)} = \sum_{n \geq 0} q^{(n)}. \tag{9}$$

Απόδειξη. Επειδή το $q(0_d)$ είναι ο μηδενικός πίνακας τότε $[e_0 - q](0_d) = I_s$ και επομένως ο συνελιξιακός αντίστροφος του $e_0 - q$ υπάρχει μέσω του Θεωρήματος 2. Επιπλέον, από τη σχέση (4), παίρνουμε άμεσα ότι:

$$(e_0 - q)^{(-1)} = \sum_{n \geq 0} (e_0 - (e_0 - q))^{(l)} = \sum_{n \geq 0} q^{(n)}.$$

□

Η σχέση (9) μπορεί να γραφεί τετριμμένα: $u = e_0 + q * u$. Η εν λόγω αναπαράσταση μας οδηγεί σε μία ειδική περίπτωση πολυδιάστατης επέκτασης της γνωστής κλάσης των μαρκοβιανών ανανεωτικών εξισώσεων διακριτού χρόνου.

Ορισμός 7. *Έστω $L \in \mathscr{M}_s(\mathbb{N}^d)$ μία άγνωστη ακολουθία πινάκων και $G \in \mathscr{M}_s(\mathbb{N}^d)$ κάποια γνωστή. Η εξίσωση*

$$L = G + q * L \tag{10}$$

ονομάζεται μαρκοβιανή ανανεωτική εξίσωση πολυδιάστατου διακριτού χρόνου.

Από τη σχέση (9) παίρνουμε άμεσα τη λύση της παραπάνω εξίσωσης.

Πόρισμα 1. *Η μαρκοβιανή ανανεωτική εξίσωση πολυδιάστατου χρόνου (10) έχει μοναδική λύση η οποία δίνεται από*

$$L = u * G.$$

Η παραπάνω λύση θα διευκολύνει τη διαδικασία ανάπτυξης της θεωρίας των χρονικά πολυδιάστατων ημιμαρκοβιανών αλυσίδων αντίστοιχα με τη μονοδιάστατη περίπτωση. Κάποια από αυτά τα χαρακτηριστικά ορίζονται παρακάτω.

Ορισμός 8. *Έστω (J,S) μία χ.π.μ.α.α. Η διαδικασία $Z = (Z_{k_{1:d}})_{k_{1:d} \in \mathbb{N}^d}$ η οποία καταγράφει την κατάσταση του συστήματος σε μία δεδομένη χρονική στιγμή $k_{1:d}$,*

$$Z_{k_{1:d}} = J_{N(k_{1:d})},$$

ονομάζεται χρονικά πολυδιάστατη (d-διάστατη) ημιμαρκοβιανή αλυσίδα η οποία επάγεται από τη (J,S).

Ορισμός 9. *Η συνάρτηση μεταβάσεων της Z γράφεται ως $P \in \mathscr{M}_s(\mathbb{N}^d)$ και ο τύπος της δίνεται από:*

$$P_{ij}(k_{1:d}) := \mathbb{P}(Z_{k_{1:d}} = j \mid Z_{0_d} = i),\ i,j \in E,\ k_{1:d} \in \mathbb{N}^d.$$

Σημείωση 5. *Για $d = 1$, η $Z_{k_{1:d}}$ αντιστοιχεί στη συνήθη ημιμαρκοβιανή αλυσίδα.*

Αντίστοιχα με τη χρονικά μονοδιάστατη περίπτωση, η συνάρτηση μεταβάσεων της Z ικανοποιεί μία μαρκοβιανή ανανεωτική εξίσωση πολυδιάστατου χρόνου, η οποία μπορεί να λυθεί εύκολα

Πρόταση 5. *Η ακολουθία πινάκων $P \in \mathscr{M}_s(\mathbb{N}^d)$ υπολογίζεται ως εξής*

$$P = u * \widetilde{H}, \tag{11}$$

όπου

$$\widetilde{H} = \operatorname{diag}\left\{\widetilde{H}_i(k_{1:d})\right\} \in \mathscr{M}_s(\mathbb{N}^d).$$

Απόδειξη. Θεωρούμε το ενδεχόμενο $A_{k_{1:d}} := \{S_1 \leq k_{1:d}\}$. Τότε,

$$P_{ij} = \mathbb{P}\left(Z_{k_{1:d}} = j, A_{k_{1:d}} \mid Z_{0_d} = i\right) + \mathbb{P}\left(Z_{k_{1:d}} = j, A^c_{k_{1:d}} \mid Z_{0_d} = i\right). \tag{12}$$

Παρατηρούμε ότι

$$\mathbb{P}\left(Z_{k_{1:d}} = j, A^c_{k_{1:d}} \mid Z_{0_d} = i\right) = \mathbb{1}_{\{i=j\}} \cdot \widetilde{H}_i(k_{1:d}), \tag{13}$$

και

$$\begin{aligned}
\mathbb{P}(Z_{k_{1:d}} = j, A_{k_{1:d}} \mid Z_{0_d} = i) &= \mathbb{P}(Z_{k_{1:d}} = j, S_1 \leq k_{1:d} \mid J_0 = i) \\
&= \sum_{r\in E}\sum_{l=0_d}^{k_{1:d}} \mathbb{P}(Z_{k_{1:d}} = j, Z_{S_1} = r, S_1 = l \mid J_0 = i) \\
&= \sum_{r\in E}\sum_{l=0_d}^{k_{1:d}} \mathbb{P}(Z_{k_{1:d}} = j, J_1 = r, S_1 = l \mid J_0 = i) \\
&= \sum_{r\in E}\sum_{l=0_d}^{k_{1:d}} \mathbb{P}(Z_{k_{1:d}} = j \mid J_1 = r, S_1 = l)\,\mathbb{P}(J_1 = r, S_1 = l \mid J_0 = i) \\
&= \sum_{r\in E}\sum_{l=0_d}^{k_{1:d}} q_{ir}(l)\mathbb{P}(Z_{k_{1:d}} = j \mid J_1 = r, S_1 = l)\,.
\end{aligned}$$

Τότε, από τη χρονική ομογένεια της (J,S), παίρνουμε άμεσα ότι:

$$\mathbb{P}(Z_{k_{1:d}} = j \mid J_1 = r, S_1 = l) = \mathbb{P}(Z_{k_{1:d}-l} = j \mid J_0 = r) = P_{rj}(k_{1:d} - l)$$

και συνεπώς

$$\mathbb{P}(Z_{k_{1:d}} = j, A_{k_{1:d}} \mid Z_{0_d} = i) = \sum_{r\in E}\sum_{l=0_d}^{k_{1:d}} q_{ir}(l)P_{rj}(k_{1:d} - l) = \sum_{r\in E}[q_{ir} * P_{rj}](k_{1:d}). \qquad (14)$$

Επομένως, αντικαθιστώντας τις σχέσεις (13) και (14) στη (12) λαμβάνουμε

$$\mathbb{P}(Z_{k_{1:d}} = j \mid Z_{0_d} = i) = \mathbb{1}_{\{i=j\}} \cdot \widetilde{H}_i(k_{1:d}) + \sum_{r\in E}[q_{ir} * P_{rj}](k_{1:d}),$$

ή ισοδύναμα, τη Μαρκοβιανή ανανεωτική εξίσωση:

$$P = \widetilde{H} + q * P,$$

από την οποία προκύπτει άμεσα η (11). □

Ορίζουμε επίσης, την απαριθμήτρια διαδικασία η οποία καταγράφει τον αριθμό των επισκέψεων σε μία συγκεκριμένη κατάσταση $i \in E$ για την εμφυτευμένη Μαρκοβιανή αλυσίδα J, ως το χρόνο $k_{1:d}$:

$$\widetilde{N}_i(k_{1:d}) := \sum_{n=0}^{N(k_{1:d})} \mathbb{1}_{\{J_n=i\}} = \sum_{n=0}^{k_1+\ldots+k_d} \mathbb{1}_{\{J_n=i, S_n\leq k_{1:d}\}}.$$

Ο επόμενος ορισμός επεκτείνει την έννοια της Μαρκοβιανής ανανεωτικής συνάρτησης σε μια χρονικά πολυδιάστατη περίπτωση.

Ορισμός 10. *Η ακολουθία πινάκων $\widetilde{U} \in \mathscr{M}_s(\mathbb{N}^d)$ λεγεται μαρκοβιανή ανανεωτική συνάρτηση της χρονικά πολυδιάστατης μαρκοβιανής αλυσίδας (J,S) και δίνεται από τον τύπο*

$$\widetilde{U}_{ij}(k_{1:d}) = \mathbb{E}_i\left[\widetilde{N}_j(k_{1:d})\right], \quad i,j \in E,\ k_{1:d} \in \mathbb{N}^d.$$

Δείχνοντας ότι η $\widetilde{U}$ ικανοποιεί μία χρονικά πολυδιάστατη Μαρκοβιανή ανανεωτική εξίσωση, παίρνουμε μία απλοποιημένη αναπαράσταση της.

Πρόταση 6. *Η χρονικά πολυδιάστατη Μαρκοβιανή ανανεωτική συνάρτηση $\widetilde{U}$ ικανοποιεί τη Μαρκοβιανή ανανεωτική εξίσωση*

$$\widetilde{U} = \mathbb{I} + q * \widetilde{U},$$

και υπολογίζεται μέσω της ακόλουθης συνελιξιακής αναπαράστασης

$$\widetilde{U} = \mathbb{I} * u.$$

4. Συμπεράσματα

Τα αποτελέσματα αυτής της έρευνας ανέδειξαν συγκεκριμένες ιδιότητες που ικανοποιούνται από το συνελιξιακό γινόμενο ακολουθιών πινάκων στον πολυδιάστατο χρόνο και ειδικότερα και στο μονοδιάστατο. Αυτές συνέβαλλαν στην ακριβή αναπαράσταση του συνελιξιακού αντιστρόφου, όταν αυτός υπάρχει, και στην επακόλουθο υπολογισμό άλλων χαρακτηριστικών που υπολογίζονται με τη βοήθεια αυτού. Τέλος, η εν λόγω θεωρητική μελέτη χρησιμοποιήθηκε για τη θεμελίωση της θεωρίας των χρονικά πολυδιάστατων μαρκοβιανών ανανεωτικών αλυσίδων και των επαγόμενων ημιμαρκοβιανών αλυσίδων. Το πλαίσιο αυτό μπορεί να αναπτυχθεί περαιτέρω για μελλοντικές εφαρμογές στη θεωρία αξιοπιστίας, αλλά και σε προβλήματα βιολογίας.

ABSTRACT

This paper introduces some fundamental algebraic aspects of the discrete time convolution of sequences of matrices in several variables and examines the existence and form of the convolutional inverse. The latter will play a fundamental role for a new study of the Markov renewal theory, where the notion of time can be multidimensional. More specifically, we will present convolution approaches for solving Markov renewal equations, the construction and evolution of this new theory.

ΑΝΑΦΟΡΕΣ

P. M. Anselone. Ergodic theory for discrete semi-markov chains. *Duke Mathematical Journal*, 27(1):33--40, 1960.

V. S. Barbu and N. Limnios. *Semi-Markov chains and hidden semi-Markov models toward applications: their use in reliability and DNA analysis*, volume 191. Springer Science & Business Media, 2009.

V. S. Barbu, C. Bérard, D. Cellier, M. Sautreuil, and N. Vergne. SMM: An R Package for Estimation and Simulation of Discrete- time semi-Markov Models. *The R Journal*, 10(2): 226--247, 2018.

I. Gerontidis. Semi-markov replacement chains. *Advances in Applied Probability*, 26, 09 1994. doi: 10.2307/1427818.

R. A. Howard. *Dynamic probabilistic systems. Series in decision and control.* Wiley, New York, 1971.

J. J. Hunter. Renewal theory in two dimensions: asymptotic results. *Advances in Applied Probability*, 6(3):546--562, 1974a.

J. J. Hunter. Renewal theory in two dimensions: basic results. *Advances in Applied Probability*, 6(2):376--391, 1974b.

N. Limnios and G. Oprisan. *Semi-Markov processes and reliability.* Springer Science & Business Media, 2012.

F. Mallor, E. Omey, and J. Santos. Multivariate weighted renewal functions. *Journal of Multivariate Analysis*, 98:30--39, 01 2007. doi: 10.1016/j.jmva.2005.07.007.

C. J. Mode and G. T. Pickens. Computational methods for renewal theory and semi-markov processes with illustrative examples. *The American Statistician*, 42(2):143--152, 1988.

C. J. Mode and C. K. Sleeman. *Stochastic processes in epidemiology: HIV/AIDS, other infectious diseases and computers.* World Scientific, 2000.

R. Pyke and R. Schaufele. Limit theorems for markov renewal processes. *The Annals of Mathematical Statistics*, pages 1746--1764, 1964.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 33ου Πανελληνίου Συνεδρίου Στατιστικής (2021), σελ.121-134

ΕΠΑΓΓΕΛΜΑΤΙΚΗ ΑΝΑΠΤΥΞΗ ΚΑΙ ΕΞΟΥΘΕΝΩΣΗ (BURNOUT) ΤΩΝ ΕΚΠΑΙΔΕΥΤΙΚΩΝ: ΕΦΑΡΜΟΓΗ ΤΟΥ ΜΟΝΤΕΛΟΥ ΛΟΓΙΣΤΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΣΕ ΕΚΠΑΙΔΕΥΤΙΚΑ ΔΕΔΟΜΕΝΑ

Αθανάσιος Κ. Κούλης[1], Κωνσταντίνος Πετρόπουλος[2]
[1]Τμήμα Κοινωνικής και Εκπαιδευτικής Πολιτικής, Πανεπιστήμιο Πελοποννήσου
tnskoul@hotmail.gr
[2]Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών
costas@math.upatras.gr

ΠΕΡΙΛΗΨΗ

Η εργασιακή εξουθένωση των εκπαιδευτικών αποτελεί ένα ιδιαίτερα σημαντικό οργανωσιακό φαινόμενο, επηρεάζοντας σε μεγάλο βαθμό την ποιότητα της παρεχόμενης εκπαίδευσης αλλά και την ίδια τη λειτουργία του εκπαιδευτικού θεσμού. Η συμμετοχή των εκπαιδευτικών σε δραστηριότητες επαγγελματικής ανάπτυξης μπορεί να τους εξοπλίσει καταλλήλως ούτως ώστε να ανταπεξέρχονται στους φθοροποιούς παράγοντες που οδηγούν σταδιακά στην εξουθένωση. Η παρούσα έρευνα στόχευσε στη διερεύνηση της επίδρασης της συμμέτοχης των εκπαιδευτικών σε δραστηριότητες επαγγελματικής ανάπτυξης καθώς και διαφόρων δημογραφικών χαρακτηριστικών στα επίπεδα εξουθένωσής τους. Διεξήχθη μια ποσοτική έρευνα με χρήση ερωτηματολογίου σε δείγμα 366 εκπαιδευτικών του Νομού Αχαΐας, το οποίο επελέγη με στρωματοποιημένη μονοσταδιακή δειγματοληψία κατά συστάδες. Στα πλαίσια της ανάλυσης των δεδομένων αξιοποιήθηκε το λογιστικό μοντέλο παλινδρόμησης. Τα αποτελέσματα της έρευνας έδειξαν ότι η συμμετοχή των εκπαιδευτικών σε επιμορφωτικές δραστηριότητες καθώς και δημογραφικοί παράγοντες όπως η σχέση εργασίας και το φύλο πράγματι επιδρούν στη διαμόρφωση του βαθμού της εργασιακής τους εξουθένωσης.

Λέξεις Κλειδιά: Εργασιακή Εξουθένωση, Επαγγελματική Ανάπτυξη, Λογιστική Παλινδρόμηση, Δημογραφικοί Παράγοντες

1. ΕΙΣΑΓΩΓΗ

Πρώτιστος στόχος κάθε εκπαιδευτικού συστήματος είναι η διατήρηση ενός ικανού και άρτια καταρτισμένου εκπαιδευτικού δυναμικού (Rebore, 2010). Δεδομένης της έντονης ρευστότητας που χαρακτηρίζει τη σύγχρονη κοινωνία και δοθέντος του ότι ένας μεγάλος αριθμός εκπαιδευτικών μένουν ελλιπώς προετοιμασμένοι από τις αρχικές τους σπουδές για τις μετέπειτα ανάγκες του επαγγέλματός τους (Darling-

Hammond and Sykes, 2003), καθίσταται περισσότερο αναγκαία από ποτέ η ανάγκη για συμμετοχή των λειτουργών της εκπαίδευσης σε δραστηριότητες επαγγελματικής ανάπτυξης. Οι δραστηριότητες αυτές δύνανται να τους εξοπλίσουν με τρόπο τέτοιο ώστε να είναι σε θέση να ανταπεξέλθουν στις ποικίλες προκλήσεις και απαιτήσεις που αναδύονται ολοένα συχνότερα και εντονότερα στα πλαίσια της καθημερινής άσκησης του διδακτικού επαγγέλματος (Girvan et al., 2016). Έτσι είναι δυνατό να αποφευχθούν φαινόμενα όπως η εργασιακή εξουθένωση, που δύνανται να ζημιώσουν το ηθικό, την αφοσίωση, την απόδοση του εκπαιδευτικού δυναμικού και ασφαλώς την ποιότητα της παρεχόμενης εκπαίδευσης, με ό,τι αυτό συνεπάγεται. Η διερεύνηση ακριβώς της σχέσης μεταξύ της εργασιακής εξουθένωσης των εκπαιδευτικών από τη μια και της επαγγελματικής τους ανάπτυξης καθώς και των ιδιαίτερων δημογραφικών χαρακτηριστικών τους από την άλλη, όπως η σχέση αυτή διαμορφώνεται στο ελληνικό εκπαιδευτικό περιβάλλον, αποτελούν τους ερευνητικούς στόχους της παρούσας μελέτης.

2. ΘΕΩΡΗΤΙΚΟ ΠΛΑΙΣΙΟ

2.1 Επαγγελματική Ανάπτυξη

Η επαγγελματική ανάπτυξη με μία ευρύτερη έννοια αναφέρεται στην ανάπτυξη ενός ατόμου στα πλαίσια του επαγγελματικού του ρόλου. Στην περίπτωση του εκπαιδευτικού, συνίσταται στην επαγγελματική του βελτίωση, την ανάπτυξη που επιτυγχάνεται ως αποτέλεσμα της αποκόμισης ολοένα και περισσότερων εμπειριών και της εξέτασης, της ανάλυσης και του αναστοχασμού που ο ίδιος συστηματικά κάνει στη διδασκαλία (Glatthorn, 1995). Σε ένα παρεμφερές, αλλά και συνάμα πιο απτό πλαίσιο ο Οργανισμός Οικονομικής Συνεργασίας και Ανάπτυξης (δείτε OECD (2009)) ορίζει ως επαγγελματική ανάπτυξη τις δραστηριότητες εκείνες που αναπτύσσουν τις ατομικές δεξιότητες του εκπαιδευτικού, την εξειδικευμένη γνώση του για το επιστημονικό του αντικείμενο (Dadds, 2014), τις διδακτικές πρακτικές και στρατηγικές του. Δραστηριότητες σαν κι αυτές δύνανται να έχουν άλλοτε περισσότερο επίσημο χαρακτήρα (προγράμματα ενίσχυσης προσόντων, παρακολούθηση συνεδρίων, συμμετοχή σε δίκτυα εκπαιδευτικών) κι άλλοτε λιγότερο (προσωπική μελέτη επιστημονικών δημοσιεύσεων, παρακολούθηση ιστοτόπων εκπαιδευτικού περιεχομένου) (Ganser, 2000).

Σε κάθε περίπτωση, η συμμετοχή σε τέτοιες δραστηριότητες μπορεί να έχει καθοριστικό αντίκτυπο στις αντιλήψεις και τη γενικότερη συμπεριφορά των εκπαιδευτικών σε επίπεδο διδακτικών, και όχι μόνο, πρακτικών και μεθόδων (Supovitz et al., 2000; Griffin et al., 2017), καλλιεργώντας παράλληλα τη διδακτική τους ευελιξία και ενισχύοντας τον επαγγελματισμό, τη διδακτική ετοιμότητα και την αυτοπεποίθησή τους (Ngala and Odebero, 2010). Οι δραστηριότητες επαγγελματικής ανάπτυξης συμβάλλουν επίσης στην καλλιέργεια των αναστοχαστικών δεξιοτήτων των εκπαιδευτικών (Day, 1985; Elliott, 1991) ωθώντας τους στην κατεύθυνση του προτύπου ενός αναστοχαζόμενου επαγγελματία και οδηγώντας έτσι τελικά σε μια συνολικότερη βελτίωση της παρεχόμενης εκπαίδευσης (Avalos, 2011; Girvan et al., 2016).

Ως αποτέλεσμα της βελτίωσης αυτής ενισχύονται τα επίπεδα επίτευξης των μαθητών (Darling-Hammond, 1999; Villegas-Reimers, 2003), με την αναγνώριση στο έργο των εκπαιδευτικών εκ μέρους συναδέλφων, κηδεμόνων και ηγεσίας να αυξάνεται αισθητά. Μια τέτοια αυξημένη αναγνώριση συμβάλλει στη δημιουργία ενός καλύτερου κλίματος εντός της σχολικής μονάδας, κάτι που οδηγεί τους εκπαιδευτικούς σε αυξημένα επίπεδα ικανοποίησης και χαμηλά επίπεδα εξουθένωσης, μειώνοντας τις πιθανότητες εργασιακής φυγής (Reynolds et al., 2008; Özer and Beycioglu, 2010). Εντός ενός τέτοιου κλίματος λοιπόν φαίνεται να επιτυγχάνονται αποτελεσματικότερα οι ποικίλοι στόχοι της σχολικής μονάδας, με τη βελτίωση της ίδιας της λειτουργίας της αλλά και συνολικότερα της εκπαιδευτικής διαδικασίας να είναι θεαματική. Προκειμένου όμως η συμμετοχή των εκπαιδευτικών σε επιμορφωτικές δράσεις να έχει όλα τα προαναφερθέντα οφέλη πρέπει οι τελευταίες να είναι άρτια σχεδιασμένες και στοχευμένες στις εκάστοτε ανάγκες των εκπαιδευτικών.

Στην Ελλάδα, οι ανάγκες των εκπαιδευτικών δε διερευνώνται συστηματικά ούτε και συνυπολογίζονται στη διαμόρφωση των επιμορφωτικών προγραμμάτων (Κασσωτάκης και Αθανασοπούλου-Βραχάμη, 2011). Παρέχονται προγράμματα ακαδημαϊκού, ως επί το πλείστον, χαρακτήρα (Κατσαρού και Δεδούλη, 2008; Κασσωτάκης και Αθανασοπούλου-Βραχάμη, 2011) που δεν ανταποκρίνονται στις άμεσες και συνήθως επί του πρακτέου ανάγκες των εκπαιδευτικών Η εικόνα αυτή που διαμορφώνεται στο επίπεδο της ενδοϋπηρεσιακής επιμόρφωσης έχει ως αποτέλεσμα τη μειωμένη ικανοποίηση των εκπαιδευτικών από τα προγράμματα στα οποία έχουν ως ώρας συμμετάσχει (Βεργίδης κ.α., 2010).

2.2 Εργασιακή Εξουθένωση

Ο όρος εργασιακή εξουθένωση (burnout) διερευνήθηκε για πρώτη φορά τη δεκαετία του 1970 με αφορμή μία εκτεταμένη κρίση στο πεδίο των ανθρωπιστικών επαγγελμάτων και δη ανάμεσα στους υπέρ-φορτωμένους από υποχρεώσεις και συνάμα απογοητευμένους εργαζόμενους των σχετικών αυτών κλάδων (Freudenberger, 1974). Είναι δύσκολο να δοθεί ακριβής ορισμός της εξουθένωσης καθώς πρόκειται για μια έννοια πολυεπίπεδη (Jackson at al., 1986; Farber, 1991).

Η εξουθένωση συνήθως περιγράφεται ως ένα σύνδρομο συναισθηματικής εξάντλησης, αποπροσωποποίησης και μειωμένης προσωπικής επίτευξης (Maslach and Jackson, 1981; Maslach et al., 1996). Οι Maslach et al. (1996) αναγνωρίζουν τη συναισθηματική εξάντληση ως ένα κομβικό ζήτημα και παράγοντα σημαίνουσας σημασίας για τη διαμόρφωση της επαγγελματικής εξουθένωσης, ενώ οι Pines and Aronson (1988) συμπεριλαμβάνουν και τη φυσική εξάντληση που χαρακτηρίζεται από χαμηλά επίπεδα ενέργειας και χρόνια κούραση.

Υπάρχουν αρκετά μοντέλα για την εννοιολογική προσέγγιση της εξουθένωσης, καθένα εκ των οποίων περιγράφει υπό ένα διαφορετικό πρίσμα την ανάπτυξη και την εξελικτική πορεία του φαινομένου (Maslach and Jackson, 1984; Farber, 1991). Δημοφιλέστερο όλων είναι το μοντέλο των Maslach and Jackson (Maslach and Jackson, 1981; Maslach et al., 1996), το οποίο συνίσταται σε τρία αλληλοσυσχετιζόμενα στοιχεία. Πρόκειται για τη συναισθηματική εξάντληση

(emotional exhaustion), την αποπροσωποποίηση (depersonalization) και την προσωπική επίτευξη (personal accomplishment) (Maslach and Jackson, 1984; Jackson et al., 1986; Farber, 1991).
Η συναισθηματική εξάντληση αναφέρεται στην αίσθηση του να νιώθει κανείς ψυχολογικά υπέρ-φορτωμένος και παράλληλα αποστραγγισμένος ως προς τα ψυχικά του αποθέματα. Η αποπροσωποποίηση αναφέρεται σε μία αρνητική, σκληρή ή υπερβολικά αποστασιοποιημένη συμπεριφορά προς τους γύρω, και δη προς τους αποδέκτες του παρεχόμενου έργου. Η μειωμένη προσωπική επίτευξη αναφέρεται σε μια χαμηλή αίσθηση προσωπικής αξιοσύνης και ως εκ τούτου σε μια αυτοεκλαμβάνουσα αδυναμία επιτυχούς διεκπεραίωσης της εργασίας. Ο καθοριστικός ρόλος που διαδραματίζουν τα τρία αυτά δομικά στοιχεία του συγκεκριμένου μοντέλου στη διαμόρφωση της εργασιακής εξουθένωσης έχουν επαληθευτεί, αναφορικά με το πεδίο της εκπαίδευσης, τόσο στο πλαίσιο της πρωτοβάθμιας όσο και σε εκείνο της δευτεροβάθμιας (Jackson et al., 1986; Friesen and Sarros, 1989).
Οι εκπαιδευτικοί είναι ανάμεσα στους επαγγελματίες που βιώνουν τα υψηλοτέρα επίπεδα εργασιακού άγχους (Jennet et al., 2003; Stoeber and Rennert, 2008). Το άγχος αυτό δύναται να οδηγήσει σε συχνές απουσίες από την εργασία, σε μικρότερη αφοσίωση, σε παροδικές ή διαρκείς ασθένειες, σε διάφορες ψυχοσωματικές παθήσεις καθώς και σε έντονα επίπεδα εργασιακής πίεσης (Nias, 1989). Όταν οι εκπαιδευτικοί εργάζονται εντός ενός τέτοιου αγχογόνου περιβάλλοντος, τείνουν να μη μένουν ικανοποιημένοι από την εργασία τους και οδηγούνται προοδευτικά στην εξουθένωση (Jennet et al, 2003; Hakanen et al., 2006) και εν συνεχεία στην εγκατάλειψη της εργασίας ή ισοδύναμα στην πλήρη αποξένωση από αυτή.
Η εξουθένωση έχει βρεθεί να σχετίζεται με το φύλο των εκπαιδευτικών, τη βαθμίδα εκπαίδευσης και την προϋπηρεσία τους (Maslach et al., 1996; Antoniou et al., 2006; Antoniou et al., 2013). Οι Έλληνες εκπαιδευτικοί φαίνεται να βιώνουν θετικά συναισθήματα προσωπικής επίτευξης (Papastylianou et al., 2009) και έτσι, μολονότι τα επίπεδα άγχους είναι αρκετά υψηλά (Κάντας, 2001; Antoniou et al., 2006), φαίνεται να βιώνουν τελικά λιγότερο έντονα συναισθήματα εξουθένωσης σε σχέση με συναδέλφους τους άλλων χωρών (Κάντας, 2001; Koustelios, 2001; Antoniou et al., 2006).

3. ΜΕΘΟΔΟΛΟΓΙΑ

Το σύνολο των δεδομένων που χρησιμοποιήθηκε στα πλαίσια της παρούσας εργασίας προέρχεται από τη διδακτορική διατριβή του Κούλης (2019), η οποία εστίασε, μεταξύ άλλων, στη μελέτη της εργασιακής ικανοποίησης, της εργασιακής εξουθένωσης και της επαγγελματικής ανάπτυξης των εκπαιδευτικών. Η συλλογή των δεδομένων αυτών είχε γίνει μέσω μιας έρευνας ερωτηματολογίου (survey research), το δείγμα της οποίας αποτελούταν από 366 εκπαιδευτικούς πρωτοβάθμιας και δευτεροβάθμιας εκπαίδευσης της περιφερειακής ενότητας Αχαΐας και είχε επιλεγεί με τη μέθοδο της στρωματοποιημένης μονοσταδιακής δειγματοληψίας κατά συστάδες. Η στρωματοποίηση του πληθυσμού των σχολικών μονάδων της περιοχής

ενδιαφέροντος είχε γίνει με βάση τη βαθμίδα εκπαίδευσης στην οποία αυτές υπάγονται (πρωτοβάθμια ή δευτεροβάθμια) και το βαθμό αστικοποίησης της περιοχής στην οποία εδράζονται (αστική ή ημιαστική-αγροτική). Με τον τρόπο αυτό προέκυψαν 4 στρώματα.

Το ερωτηματολόγιο, το οποίο μοιράστηκε τόσο σε έντυπη όσο και σε ηλεκτρονική μορφή, απαρτιζόταν από τέσσερα επιμέρους τμήματα. Εξ αυτών, η παρούσα εργασία επικεντρώθηκε σε εκείνο των δημογραφικών χαρακτηριστικών των συμμετεχόντων, σε εκείνο που είχε να κάνει με τη διερεύνηση της επαγγελματικής τους ανάπτυξης, και πιο συγκεκριμένα των μορφών επιμόρφωσης στις οποίες είχαν αυτοί συμμετάσχει σε διάστημα τριών ετών πριν τη χρονική περίοδο συμπλήρωσης του ερωτηματολογίου, και τέλος σε εκείνο που αφορούσε την εργασιακή τους εξουθένωση. Στα πλαίσια του τελευταίου αυτού τμήματος του ερωτηματολογίου της έρευνας αναφοράς έγινε χρήση της κλίμακας MBI-ES (Maslach Burnout Inventory-Educators Survey). Πρόκειται για τη δημοφιλέστερη και για μια εκ των πιο αξιόπιστων και έγκυρων κλιμάκων για τη μέτρηση της εργασιακής εξουθένωσης στη διεθνή βιβλιογραφία, με τη συγκεκριμένη μάλιστα έκδοση (Educators Survey) να είναι ειδικά κατασκευασμένη και στοχευμένη στην αξιολόγηση των επιπέδων εξουθένωσης των λειτουργών της εκπαίδευσης (Maslach at al.,1996). Αξίζει να σημειωθεί ότι πριν το διαμοιρασμό του ερωτηματολογίου στο επιλεχθέν δείγμα πραγματοποιήθηκε πιλοτική μελέτη σε δείγμα 20 εκπαιδευτικών πρωτοβάθμιας και δευτεροβάθμιας εκπαίδευσης του νομού Αχαΐας.

Στη βάση της φύσης και της μορφής των δεδομένων τα οποία αποτέλεσαν το επίκεντρο του ενδιαφέροντος της παρούσας εργασίας επελέγη να χρησιμοποιηθεί το μοντέλο της Λογιστικής Παλινδρόμησης για την ανάλυσή τους. Έγινε χρήση του πακέτου στατιστικών αναλύσεων IBM SPSS Statistics v.25. Οι αναλύσεις κινήθηκαν σε δυο κύριους άξονες. Ο πρώτος αφορά μοντέλα που διερευνούν τη συσχέτιση της εργασιακής εξουθένωσης των εκπαιδευτικών με τις ποικίλες επιμορφωτικές δράσεις στις οποίες είχαν λάβει μέρος, ενώ ο δεύτερος μοντέλα που μελετούν τη συσχέτιση εργασιακής εξουθένωσης και δημογραφικών χαρακτηριστικών των συμμετεχόντων. Σε αμφότερες τις ομάδες μοντέλων, μεταβλητές απόκρισης αποτέλεσαν οι τρεις μεταβλητές που περιγράφουν την εργασιακή εξουθένωση των εκπαιδευτών στα πλαίσια της κλίμακας MBI-ES. Πρόκειται για τη συναισθηματική εξάντληση, την αποπροσωποποίηση και την προσωπική επίτευξη, όλες τους διατακτικές μεταβλητές τριών κατηγοριών έκαστη. Η κατάρτιση των εν λόγω κατηγοριών είχε ήδη γίνει στο πλαίσιο του σετ δεδομένων με βάση τον Πίνακα 1 που ακολουθεί.

Όσον αφορά τις ανεξάρτητες μεταβλητές των μοντέλων της πρώτης ομάδας, αυτές συνίστανται σε εκείνες που αφορούν τη συμμετοχή σε προγράμματα επιμόρφωσης, όλες τους ονομαστικές διχοτομικές μεταβλητές (Ναι ή Όχι). Από την άλλη, για τα μοντέλα της δεύτερης ομάδας ανεξάρτητες μεταβλητές ήταν τα έτη προϋπηρεσίας, μια συνεχής μεταβλητή, και τα λοιπά δημογραφικά χαρακτηριστικά, όλα τους ονομαστικές μεταβλητές.

Πίνακας 1. *Scoring Table μεταβλητών Εξουθένωσης*

	Συναισθηματική Εξάντληση	Αποπροσωποποίηση	Προσωπική Επίτευξη
Υψηλή	27 ή περισσότερο	13 ή περισσότερο	39 ή περισσότερο
Μέτρια	17-26	7-12	32-38
Χαμηλή	0-16	0-6	0-31

Από το σύνολο των 6 μοντέλων που μελετήθηκαν επελέγη να παρουσιαστούν στην παρούσα εργασία τα δύο που αφορούν την προσωπική επίτευξη των εκπαιδευτικών. Η επιλογή αυτή έγινε αφενός λόγω έλλειψης χώρου και αφετέρου με το σκεπτικό ότι η συγκεκριμένη μεταβλητή της εργασιακής εξουθένωσης είναι υπεύθυνη, σύμφωνα με τη βιβλιογραφία (Κάντας, 2001; Koustelios, 2001; Antoniou et al., 2006; Papastylianou et al., 2009), για τα χαμηλότερα επίπεδα εξουθένωσης των Ελλήνων εκπαιδευτικών σε σχεση με τους συναδέλφους τους άλλων χωρών. Ως εκ τούτου η διεξοδικότερη παρουσίαση των μοντέλων αυτών συγκεντρώνει ιδιαίτερο ενδιαφέρον.
Τέλος, να σημειωθεί ότι για καθεμιά εκ των ανεξάρτητων ονομαστικών μεταβλητών των μοντέλων το πακέτο στατιστικών αναλύσεων δημιούργησε αυτόματα dummy (εικονικές) μεταβλητές που αναπαριστούσαν τη σύγκριση μεταξύ κάθε κατηγορίας της μεταβλητής ενδιαφέροντος με την κατηγορία που λειτουργούσε κάθε φορά ως κατηγορία αναφοράς.

4. ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

4.1. Εξουθένωση & Επιμόρφωση

Προσωπική Επίτευξη & Επιμόρφωση

Ο Πίνακας 2 παρέχει δύο μέτρα που μπορούν να χρησιμοποιηθούν για την αξιολόγηση του πόσο καλά προσαρμόζεται το προτεινόμενο μοντέλο στα δεδομένα.

Πίνακας 2. *Έλεγχοι Καλής Προσαρμογής*

	Chi-Square	df	Sig.
Pearson	199,808	172	,072
Deviance	184,524	172	,243

Η τιμή της στατιστικής συνάρτησης Pearson chi-square είναι ίση με 199,808 με p-τιμή=0,072>0,05 ενώ η τιμή της στατιστικής συνάρτησης Deviance chi-square είναι 184,524 με p-τιμή=0,243>0,05. Συνεπώς αμφότερες οι στατιστικές συναρτήσεις καταδεικνύουν καλή προσαρμογή του μοντέλου στα δεδομένα.
Ένα γενικό μέτρο της σημαντικότητας του μοντέλου λαμβάνουμε από τον Πίνακα 3.

Πίνακας 3. *Σημαντικότητα του Μοντέλου*

Model	Model Fitting Criteria -2 Log Likelihood	Likelihood Ratio Tests		
		Chi-Square	df	Sig.
Intercept Only	343,850			
Final	293,212	50,638	16	,000

Η p-τιμή<0,001, κάτι που σημαίνει ότι το πλήρες μοντέλο προβλέπει στατιστικά σημαντικά την μεταβλητή απόκρισης καλύτερα από το μοντέλο σταθερού όρου. Ιδιαίτερα σημαντικά είναι τα αποτελέσματα που παρουσιάζονται στον Πίνακα 4 που παρατίθεται παρακάτω.

Πίνακας 4. *Έλεγχοι Λόγου Πιθανοφάνειας*

Effect	Model Fitting Criteria -2 Log Likelihood of Reduced Model	Likelihood Ratio Tests		
		Chi-Square	df	Sig.
Intercept	293,212[a]	,000	0	.
Προγράμματα Ενίσχυσης Προσόντων	294,141	,929	2	,628
Εκπαιδευτικά Συνέδρια ή Σεμινάρια	294,748	1,536	2	,464
Υποχρεωτικά Ενδοϋπηρεσιακά Προγράμματα	295,469	2,257	2	,324
Μη Υποχρεωτικά Ενδοϋπηρεσιακά Προγράμματα	313,900	20,687	2	,000
Ατομική ή Συνεργατική Έρευνα	298,103	4,891	2	,087
Δίκτυο Εκπαιδευτικών	294,094	,882	2	,644
Προγράμματα άλλων οργανισμών εκτός Δημοσίου	301,985	8,773	2	,012
Επισκέψεις παρατήρησης σε σχολικές μονάδες	297,924	4,711	2	,095

Ο εν λόγω πίνακας καταδεικνύει ποιες από τις ανεξάρτητες μεταβλητές του μοντέλου είναι στατιστικά σημαντικές. Μπορούμε να διακρίνουμε ότι από τις ανεξάρτητες μεταβλητές του μοντέλου, στατιστικά σημαντικές σε επίπεδο σημαντικότητας α=5%,

είναι οι μεταβλητές Μη Υποχρεωτικά Ενδοϋπηρεσιακά Προγράμματα (p<0,001) και Προγράμματα άλλων οργανισμών εκτός Δημοσίου (p=0,012). Σε επίπεδο σημαντικότητας α=10% στατιστικά σημαντικές είναι και οι μεταβλητές Ατομική ή Συνεργατική Έρευνα (p=0,087) και Επισκέψεις παρατήρησης σε σχολικές μονάδες (p=0,095). Ο παραπάνω πίνακας είναι ιδιαίτερα χρήσιμος για την περίπτωση των ονομαστικών ανεξάρτητων μεταβλητών διότι πρόκειται για τον μοναδικό που λαμβάνει υπόψη του την συνολική επίδραση μιας τέτοιας μεταβλητής στην απόκριση.

Περνώντας στα ευρήματα των πινάκων Εκτιμητών Παραμέτρων (Parameter Estimates), οι οποίοι παραλείπονται από την παρούσα εργασία για το σύνολο των παρουσιαζόμενων μοντέλων λόγω έλλειψης χώρου, μπορούμε να συμπεραίνουμε ότι οι εκπαιδευτικοί που έχουν συμμετάσχει σε μη υποχρεωτικά ενδοϋπηρεσιακά προγράμματα επιμόρφωσης έχουν 3,5 φορές μεγαλύτερη πιθανότητα από τους συναδέλφους τους που δεν έχουν συμμετάσχει σε τέτοιου είδους δράσεις να εμφανίζουν υψηλά, σε σχέση με χαμηλά, επίπεδα προσωπικής επίτευξης. Επιπλέον, φαίνεται ότι οι εκπαιδευτικοί που έχουν συμμετάσχει σε Ατομική ή Συνεργατική Έρευνα σε ζητήματα ενδιαφέροντος έχουν 2,1 φορές μεγαλύτερη πιθανότητα από τους συναδέλφους τους που δεν έχουν συμμετάσχει σε τέτοιου είδους δράσεις να εμφανίζουν υψηλά, σε σχέση με χαμηλά, επίπεδα προσωπικής επίτευξης. Οι εκπαιδευτικοί που έχουν συμμετάσχει σε προγράμματα επιμόρφωσης άλλων οργανισμών εκτός Δημοσίου έχουν 2,4 φορές μικρότερη πιθανότητα από τους συναδέλφους τους που δεν έχουν συμμετάσχει σε τέτοιου είδους δράσεις να εμφανίζουν υψηλά, σε σχέση με χαμηλά, επίπεδα προσωπικής επίτευξης, ενώ εκείνοι που έχουν συμμετάσχει σε μη υποχρεωτικά ενδοϋπηρεσιακά προγράμματα επιμόρφωσης έχουν 5,7 φορές μεγαλύτερη πιθανότητα από τους συναδέλφους τους που δεν έχουν συμμετάσχει σε τέτοιου είδους δράσεις να εμφανίζουν μέτρια, σε σχέση με χαμηλά, επίπεδα προσωπικής επίτευξης. Τέλος, οι εκπαιδευτικοί που έχουν συμμετάσχει σε προγράμματα επιμόρφωσης άλλων οργανισμών εκτός Δημοσίου έχουν 2 φορές μεγαλύτερη πιθανότητα από τους συναδέλφους τους που δεν έχουν συμμετάσχει σε τέτοιου είδους δράσεις να εμφανίζουν μέτρια, σε σχέση με υψηλά, επίπεδα προσωπικής επίτευξης.

4.2. Εξουθένωση & Δημογραφικά Στοιχεία

Προσωπική Επίτευξη & Δημογραφικά Στοιχεία

Στον Πίνακα 5 μπορούμε να διακρίνουμε ότι αμφότερες οι στατιστικές συναρτήσεις Pearson chi-square (τιμή 589,745 και p-τιμή 0,054>0,05) και Deviance chi-square (τιμή 456,289 και p-τιμή 0,995) καταδεικνύουν καλή προσαρμογή του μοντέλου στα δεδομένα.

Πίνακας 5. *Έλεγχοι Καλής Προσαρμογής*

	Chi-Square	df	Sig.
Pearson	589,745	536	,054
Deviance	456,289	536	,995

Ο Πίνακας 6 δείχνει ότι το πλήρες μοντέλο προβλέπει στατιστικά σημαντικά την μεταβλητή απόκρισης καλύτερα από το μοντέλο σταθερού όρου (p-τιμή<0,001).

Πίνακας 6. *Σημαντικότητα του Μοντέλου*

Model	Model Fitting Criteria -2 Log Likelihood	Likelihood Ratio Tests		
		Chi-Square	df	Sig.
Intercept Only	578,328			
Final	510,885	67,444	20	,000

Στον Πίνακα 7 μπορούμε να διακρίνουμε ότι από τις ανεξάρτητες μεταβλητές του μοντέλου, στατιστικά σημαντικές σε επίπεδο σημαντικότητας α=5% είναι το Φύλο (p=0,006), το Επίπεδο Εκπαίδευσης (p=0,001), η Βαθμίδα Εκπαίδευσης (p=0,045), η Σχέση Εργασίας (p=0,027) και ο Τόπος Εργασίας (p=0,004).

Πίνακας 7. *Έλεγχοι Λόγου Πιθανοφάνειας*

Effect	Model Fitting Criteria -2 Log Likelihood of Reduced Model	Likelihood Ratio Tests		
		Chi-Square	df	Sig.
Intercept	510,885[a]	,000	0	.
Έτη Προϋπηρεσίας	514,288	3,404	2	,182
Φύλο	521,177	10,292	2	,006
Εκπαίδευση	530,366	19,482	4	,001
Βαθμίδα Εκπαίδευσης	517,079	6,194	2	,045
Ειδικότητα	520,759	9,875	6	,130
Σχέση Εργασίας	518,078	7,193	2	,027
Τόπος Εργασίας	521,793	10,908	2	,004

Όσον αφορά τα ευρήματα των πινάκων Εκτιμητών Παραμέτρων συμπεραίνουμε ότι οι άνδρες εκπαιδευτικοί έχουν 2,7 φορές μεγαλύτερη πιθανότητα από τις γυναίκες συναδέλφους τους να εμφανίζουν υψηλά, σε σχέση με τα χαμηλά, επίπεδα προσωπικής επίτευξης. Επιπλέον φαίνεται ότι οι εκπαιδευτικοί πρωτοβάθμιας έχουν 5,6 φορές μεγαλύτερη πιθανότητα από τους συναδέλφους τους της δευτεροβάθμιας να εμφανίζουν υψηλά, σε σχέση με τα χαμηλά, επίπεδα προσωπικής επίτευξης. Οι φιλόλογοι έχουν 2,8 φορές μεγαλύτερη πιθανότητα από τους συναδέλφους τους των ειδικοτήτων και 7,5 φορές μεγαλύτερη πιθανότητα από τους δασκάλους να εμφανίζουν υψηλά, σε σχέση με τα χαμηλά, επίπεδα προσωπικής επίτευξης. Επιπρόσθετα οι καθηγητές θετικών επιστημών έχουν 6,4 φορές μεγαλύτερη πιθανότητα από τους δασκάλους να εμφανίζουν υψηλά, σε σχέση με τα χαμηλά, επίπεδα προσωπικής επίτευξης Οι εκπαιδευτικοί που εργάζονται σε αστικές περιοχές έχουν 3 φορές μεγαλύτερη πιθανότητα από τους συναδέλφους τους που εργάζονται σε ημιαστικές ή αγροτικές περιοχές να εμφανίζουν υψηλά, σε σχέση με τα χαμηλά,

επίπεδα προσωπικής επίτευξης. Οι εκπαιδευτικοί πρωτοβάθμιας έχουν 8,4 φορές μεγαλύτερη πιθανότητα από τους συναδέλφους τους της δευτεροβάθμιας να εμφανίζουν μέτρια, σε σχέση με τα χαμηλά, επίπεδα προσωπικής επίτευξης. Επίσης, οι κάτοχοι μεταπτυχιακού ή/και διδακτορικού διπλώματος έχουν 10,2 φορές μεγαλύτερη πιθανότητα από τους αποφοίτους της Παιδαγωγικής Ακαδημίας να εμφανίζουν μέτρια, σε σχέση με τα χαμηλά, επίπεδα προσωπικής επίτευξης. Οι απόφοιτοι ΑΕΙ έχουν 5,8 φορές μικρότερη πιθανότητα από τους κατόχους μεταπτυχιακού ή/και διδακτορικού διπλώματος να εμφανίζουν μέτρια, σε σχέση με τα χαμηλά, επίπεδα προσωπικής επίτευξης. Συνεχίζοντας, οι μόνιμοι εκπαιδευτικοί έχουν 6,8 φορές μεγαλύτερη πιθανότητα από τους αναπληρωτές ή ωρομίσθιους συναδέλφους τους να εμφανίζουν μέτρια, σε σχέση με τα χαμηλά, επίπεδα προσωπικής επίτευξης, ενώ οι εκπαιδευτικοί που εργάζονται σε αστικές περιοχές έχουν 5,3 φορές μεγαλύτερη πιθανότητα από τους συναδέλφους τους που εργάζονται σε ημιαστικές η αγροτικές περιοχές να εμφανίζουν μέτρια, σε σχέση με τα χαμηλά, επίπεδα προσωπικής επίτευξης. Τέλος, οι κάτοχοι μεταπτυχιακού ή/και διδακτορικού διπλώματος έχουν 7,4 φορές μεγαλύτερη πιθανότητα από τους αποφοίτους της Παιδαγωγικής Ακαδημίας και 2,7 φορές μεγαλύτερη πιθανότητα από τους αποφοίτους ΑΕΙ να εμφανίζουν μέτρια, σε σχέση με τα υψηλά, επίπεδα προσωπικής επίτευξης.

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Μέσα από τα ευρήματα των αναλύσεων που διεξήχθησαν, τα οποία και παρατέθηκαν διεξοδικά παραπάνω, αναδεικνύονται ορισμένα σημαντικά συμπεράσματα, τα οποία σκιαγραφεί η παρούσα ενότητα. Αναφορικά με τον πρώτο ερευνητικό στόχο της μελέτης, δηλαδή τη διερεύνηση της σχέσης μεταξύ εξουθένωσης και επαγγελματικής ανάπτυξης των εκπαιδευτικών, τα ευρήματα υποδεικνύουν ότι η συμμετοχή σε προγράμματα ενίσχυσης προσόντων, αλλά και σε υποχρεωτικά και μη ενδοϋπηρεσιακά προγράμματα φαίνεται πράγματι να συμβάλλει σε χαμηλότερα επίπεδα αποπροσωποποίησης και σε αυξημένα επίπεδα προσωπικής επίτευξης. Η συμμετοχή σε εκπαιδευτικά συνέδρια ή σεμινάρια βρέθηκε επίσης να επιδρά ευεργετικά στα συνολικότερα επίπεδα εξουθένωσης των εκπαιδευτικών.

Άξιο αναφοράς, στο σημείο αυτό, είναι το ότι οι εκπαιδευτικοί που στρέφονται σε ατομική η συνεργατική έρευνα σε ζητήματα ενδιαφέροντος βρέθηκαν να εμφανίζουν υψηλότερα επίπεδα αποπροσωποποίησης σε σχέση με τους συναδέλφους τους, ενώ, στο ίδιο μήκος κύματος, εκείνοι που στρέφονται σε προγράμματα επιμόρφωσης άλλων οργανισμών εκτός δημοσίου φαίνεται να βιώνουν χαμηλότερα επίπεδα προσωπικής επίτευξης και αυξημένα επίπεδα συναισθηματικής εξάντλησης σε σχέση με τους υπολοίπους. Μολονότι το εν λόγω εύρημα φαίνεται οξύμωρο, εντούτοις ίσως αυτά ακριβώς τα συναισθήματα εξουθένωσης, στα οποία ενδεχομένως οι υπόλοιπες επιμορφωτικές δράσεις στις οποίες έχουν συμμετάσχει οι εκπαιδευτικοί να μην δύναται να δώσουν απαντήσεις, να είναι που οδηγούν τους εν λόγω εξουθενωμένους εκπαιδευτικούς στο να στραφούν σε άλλες δραστηριότητες, που θίγουν διαφορετικά ζητήματα, πέρα από τις ορισμένες φορές τετριμμένες θεματικές των υποχρεωτικών

ενδοϋπηρεσιακών επιμορφώσεων, προκειμένου να φθάσουν στις λύσεις που επιζητούν. Ίσως δηλαδή στο σημείο αυτό η εξουθένωση να είναι το αίτιο και όχι το αιτιατό ως προς τη συμμετοχή στις εν λόγω επιμορφωτικές δράσεις. Περαιτέρω έρευνα, ποιοτική ή ποσοτική, δύναται να ενισχύσει αυτήν την υπόθεση.

Αναφορικά με το δεύτερο ερευνητικό στόχο, δηλαδή τη διερεύνηση της σχέσης μεταξύ εργασιακής εξουθένωσης και δημογραφικών χαρακτηριστικών των εκπαιδευτικών, φαίνεται ότι, ως προς το ανώτατο επίπεδο εκπαίδευσης, οι κάτοχοι μεταπτυχιακού ή και διδακτορικού τείνουν να εμφανίζουν υψηλότερα επίπεδα συναισθηματικής εξουθένωσης και μέτρια επίπεδα προσωπικής επίτευξης σε σχέση με τους συναδέλφους τους. Μέτρια είναι και τα επίπεδα συναισθηματικής εξάντλησης των αποφοίτων της Παιδαγωγικής Ακαδημίας, σε σχέση με τους υπολοίπους. Τα εν λόγω συμπεράσματα παρουσιάζουν ιδιαίτερο ενδιαφέρον και ενδέχεται να σχετίζονται με τις απαιτήσεις, τις προσδοκίες και τους στόχους που θέτει καθημερινά αλλά και μακροπρόθεσμα καθεμιά εκ των συναφών ομάδων εκπαιδευτικών στα πλαίσια της διδακτικής πράξης. Περαιτέρω έρευνα θα μπορούσε να ρίξει περισσότερο φως στους μηχανισμούς που οδηγούν στα παρατηρηθέντα ευρήματα.

Σε ό,τι αφορά τη σχέση εργασίας, οι αναπληρωτές εκπαιδευτικοί παρουσιάζουν υψηλότερα επίπεδα εξουθένωσης και ειδικότερα υψηλότερη αποπροσωποποίηση και χαμηλότερη προσωπική επίτευξη, σε σχέση με τους μόνιμους συναδέλφους τους.

Περνώντας στο βαθμό αστικοποίησης της περιοχής στην οποία εδράζεται η εκάστοτε σχολική μονάδα, οι εργαζόμενοι σε σχολικές μονάδες αστικών περιοχών εμφανίζουν υψηλότερα επίπεδα προσωπικής επίτευξης σε σχέση με τους συναδέλφους τους που εργάζονται σε αγροτικές ή ημιαστικές περιοχές. Αναφορικά με το φύλο, ιδιαίτερα σημαντική παράμετρο σύμφωνα με τη σχετική βιβλιογραφία, οι άνδρες εκπαιδευτικοί φαίνεται να εμφανίζουν υψηλότερα επίπεδα αποπροσωποποίησης αλλά και υψηλότερα επίπεδα προσωπικής επίτευξης από τις γυναίκες συναδέλφους τους.

Σχετικά με την βαθμίδα εκπαίδευσης, οι εκπαιδευτικοί πρωτοβάθμιας εκπαίδευσης εμφανίζουν γενικά υψηλότερα επίπεδα προσωπικής επίτευξης, αλλά και υψηλότερα επίπεδα συναισθηματικής εξάντλησης, σε σχέση με τους συναδέλφους τους της δευτεροβάθμιας. Τέλος όσον αφορά την ειδικότητα των εκπαιδευτικών, οι φιλόλογοι και οι καθηγητές θετικών επιστημών φαίνεται να βιώνουν υψηλότερα επίπεδα προσωπικής επίτευξης σε σχέση με τους συναδέλφους τους.

ABSTRACT

Teachers' burnout consists a very important organizational phenomenon, greatly affecting the quality of the education provided as well as the general operation of the educational system. Participation in professional development (PD) activities can prepare teachers adequately to cope with factors causing burnout feelings. The present study aimed to investigate the effect of teachers' participation in such activities as well as of their various demographic characteristics on their burnout levels, in the Greek context. Survey research was conducted on a sample of 366 primary and secondary education teachers, chosen through stratified single-stage

cluster sampling. Logistic Regression Model was used for the data analyses. Results demonstrated an actual effect of teachers' participation in PD activities as well as of demographic factors, such as employment relationship and gender, on their burnout feelings.

ΑΝΑΦΟΡΕΣ

Antoniou, A.-S, Ploumpi, A., and Ntalla, M. (2013). Occupational Stress and Professional Burnout in Teachers of Primary and Secondary Education: The Role of Coping Strategies. Psychology, 4 (3A), 349-355.

Antoniou, A.-S., Polychroni, F., and Vlachakis, A.-N. (2006). Gender and age differences in occupational stress and professional burnout between primary and high-school teachers in Greece. Journal of Managerial Psychology, 21(7), 682–690.

Avalos, B. (2011). Teacher professional development in Teaching and Teacher Education over ten years. Teaching and teacher education, 27(1), 10- 20.

Dadds, M. (2014). Continuing Professional Development: nurturing the expert within. Professional Development in Education, 40(1), 9–16.

Darling-Hammond, L. (1999). Target time toward teachers, Journal of Staff Development, 20(2), 31-36.

Darling-Hammond, L., and Sykes, G. (2003). Wanted: A national teacher supply policy for education: The right way to meet the "Highly Qualified Teacher" challenge. Education Policy Analysis Archives, 11(33).

Day, C. (1985). Professional Learning and Researcher Intervention: an action research perspective. British Educational Research Journal, 11(2), 133–152.

Elliott, J. (1991). Action Research for Educational Change. Buckingham: Open University Press.

Farber, B. A. (1991). Crisis in education: Stress and burnout in the American teacher. San Francisco: Jossey-Bass.

Freudenberger, H. (1974). Staff Burnout. Journal of Social Issues, 30(1), 159-165.

Friesen, D., and Sarros, J. C. (1989). Sources of burnout among educators. Journal of Organizational Behavior, 10(2), 179-188.

Ganser, T. (2000). An Ambitious Vision of Professional Development for Teachers. NASSP Bulletin, 84(618), 6-12.

Girvan, C., Conneely, C., and Tangney, B. (2016). Extending experiential learning in teacher professional development. Teaching and Teacher Education, 58, 129-139.

Glatthorn, A. (1995). Teacher development. In L. Anderson (Ed.), International encyclopedia of teaching and teacher education (2nd ed., pp. 41-45). London: Pergamon Press.

Griffin, C. C., Dana, N. F., Pape, S. J., Algina, J., Bae, J., Prosser, S. K., and League, M. B. (2017). Prime Online: Exploring Teacher Professional Development for Creating Inclusive Elementary Mathematics Classrooms. Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children, 41(2), 121–139.

Hakanen, J. J., Bakker, A. B., and Schaufeli, W. B. (2006). Burnout and work engagement among teachers. Journal of School Psychology, 43(6), 495-513.

Jackson, S. E., Schwab, R. L., and Schuler, R. S. (1986). Toward an understanding of the burnout phenomenon. Journal of Applied Psychology, 71(4), 630-640

Jennett, H. K., Harris, S. L., and Mesibov, G. B. (2003). Commitment to philosophy, teacher efficacy, and burnout among teachers of children with autism. Journal of Autism and Developmental Disorders, 33(6), 583-593.

Koustelios, A. (2001). Organizational factors as predictors of teachers' burnout. Psychological Reports, 88(3), 627-634.

Maslach, C., and Jackson, S. E. (1981). The measurement of experienced burnout. Journal of Occupational Behavior, 2, 99-113.

Maslach, C, and Jackson, S. E. (1984). Burnout in organizational settings. In S. Oskamp (Ed.), Applied social psychology annual: Applications in organizational settings (Vol. 5, pp. 133-153). Beverly Hills, CA: Sage.

Maslach, C., Jackson, S. E., and Leiter, M. P. (1996). Maslach Burnout Inventory Manual (3rd ed.). Mountain View, CA: CPP, Inc.

Maslach, C., Jackson, S. E., and Schwab, R. L. (1996). Maslach Burnout Inventory-Educators Survey (MBI-ES). In C. Maslach, S. E. Jackson, and M. P. Leiter (Eds.), Maslach Burnout Inventory Manual (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.

Ngala, F., and Odebero, S. (2010). Teachers' Perceptions of Staff Development Programmes As It Relates to Teachers' Effectiveness: A Study of Rural Primary Schools' in Kenya. Educational Research & Review, 5(1), 1-9.

Nias, J. (1989). Subjectively speaking: English primary teachers' careers. International Journal of Educational Research, 13(4), 391–402.

OECD (2009). Creating Effective Teaching and Learning Environments: First Results from TALIS. OECD Publications: Paris.

Papastylianou, A., Kaila, M., and Polychronopoulos, M. (2009). Teachers' Burnout, Depression, Role Ambiguity and Conflict. Social Psychological Education, 12(3), 295-314.

Pines, A., and Aronson, E. (1988). Career burnout: Causes and cures. New York, NY, US: Free Press.

Rebore, R. W. (2010). Human Resources Administration in Education: A Management Approach (9th ed.). Upper Saddle River, NJ: Prentice Hall.

Reynolds, A., Ross, S. M., and Rakow, J. H. (2002). Teacher retention, teaching effectiveness, and professional preparation: A comparison of professional development school and non-professional development school graduates. Teaching and Teacher Education, 18(3), 289-303

Stoeber, J., and Rennert, D. (2008). Perfectionism in school teachers: relations with stress appraisals, coping styles, and burnout. Anxiety, Stress & Coping. An International Journal, 21(1), 37-53.

Supovitz, J. A., Mayer, D. P., and Kahle, J. B. (2000). Promoting inquiry-based instructional practice: the longitudinal impact of professional development in the context of systemic reform. Educational Policy, 14(3), 331- 356.

Villegas-Reimers, E. (2003). Teacher Professional Development: An International Review of the Literature. Paris: UNESCO International Institute for Educational Planning.

Özer, N. and Beycioglu, K. (2010). The relationship between teacher professional development and burnout. Procedia-Social and Behavioral Sciences, 2(2), 4928-4932.

Βεργίδης, Δ., Αναστασιάδης, Π., Καραδήμας, Ε., Φερεντίνος, Σ., Τραντάς, Π., Καρβούνης, Λ.,…, και Συρίου, Ι. (2010). Η συμβολή της διερεύνησης επιμορφωτικών αναγκών στην επιμόρφωση των εκπαιδευτικών: Συγκριτική Ερμηνεία Αποτελεσμάτων (7ο Μέρος). Αθήνα: Παιδαγωγικό Ινστιτούτο-Μείζον Πρόγραμμα Επιμόρφωσης Εκπαιδευτικών.

Κάντας, Α. (2001). Οι παράγοντες άγχους και η επαγγελματική εξουθένωση στους εκπαιδευτικούς. Στο: Ε. Βασιλάκη, Σ. Τριλίβα, και Η. Μπεζεβέγκης (Επιμ.), Το στρες, το άγχος και η αντιμετώπισή τους (σελ. 217-230). Αθήνα: Ελληνικά Γράμματα.

Κασσωτάκης, Μ., και Αθανασοπούλου-Βραχάμη, Γ. (2011). Η συνεχιζόμενη κατάρτιση των Ελλήνων εκπαιδευτικών της Δευτεροβάθμιας Εκπαίδευσης. Στο: Β. Δ. Οικονομίδης (Επιμ.), Εκπαίδευση και επιμόρφωση εκπαιδευτικών (σσ. 673-705). Αθήνα: Πεδίο

Κατσαρού, Ε., και Δεδούλη, Μ. (2008). Επιμόρφωση και Αξιολόγηση στο χώρο της Εκπαίδευσης. Αθήνα: Παιδαγωγικό Ινστιτούτο.

Κούλης, Α. Κ. (2019). *Εργασιακή Ικανοποίηση και Επαγγελματική Ανάπτυξη των Εκπαιδευτικών* (Διδακτορική Διατριβή, Πανεπιστήμιο Πελοποννήσου, Τμήμα Κοινωνικής και Εκπαιδευτικής Πολιτικής). Διαθέσιμο από το Εθνικό Αρχείο Διδακτορικών Διατριβών (Κωδ. 46922).

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 33ου Πανελληνίου Συνεδρίου Στατιστικής (2021), σελ.135-149

ΔΙΑΓΡΑΜΜΑΤΑ ΕΛΕΓΧΟΥ EWMA ΓΙΑ ΤΗΝ ΠΑΡΑΚΟΛΟΥΘΗΣΗ ΠΟΣΟΣΤΩΝ ΚΑΙ ΑΝΑΛΟΓΙΩΝ: ΜΙΑ ΣΥΓΚΡΙΤΙΚΗ ΜΕΛΕΤΗ

Λαφατζή Α., Ρακιτζής Α. Χ
Εργαστήριο Στατιστικής και Ανάλυσης Δεδομένων, Τμήμα Στατιστικής και Αναλογιστικών-Χρηματοοικονομικών Μαθηματικών, Πανεπιστήμιο Αιγαίου
sasm19010@sas.aegean.gr, arakitz@aegean.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία προτείνουμε και μελετάμε δίπλευρα διαγράμματα ελέγχου τύπου EWMA για την παρακολούθηση ποσοστών και αναλογιών. Υποθέτουμε ότι σε κάθε στάδιο της δειγματοληψίας λαμβάνονται μεμονωμένες παρατηρήσεις, με πιθανές τιμές στο διάστημα (0, 1). Για τη στατιστική μοντελοποίηση αυτού τους είδους δεδομένων, θεωρούμε ως υποψήφια μοντέλα τις κατανομές Βήτα, Simplex και Unit Gamma και υπολογίζουμε την απόδοση των διαγραμμάτων ελέγχου EWMA για κάθε μοντέλο. Η απόδοσή τους συγκρίνεται με την απόδοση των αντίστοιχων διαγραμμάτων ελέγχου τύπου Shewhart από όπου προκύπτει η υπεροχή τους έναντι των Shewhart. Επιπλέον, εξετάζεται η επίδραση που έχει στην απόδοση των EWMA διαγραμμάτων η χρήση ορίων ελέγχου τα οποία δεν έχουν υπολογιστεί υπό το πραγματικό (σωστό) μοντέλο.

Λέξεις Κλειδιά: Κατανομή Βήτα, Κατανομή Simplex, Κατανομή Unit Gamma, Ποσοστά, EWMA Διαγράμματα Ελέγχου

1. ΕΙΣΑΓΩΓΗ

Ο στατιστικός έλεγχος διεργασιών (ΣΕΔ) είναι μια συλλογή εργαλείων που επιτρέπει την παρακολούθηση μιας διαδικασίας. Μεταξύ αυτών των εργαλείων, το πιο διαδεδομένο είναι το διάγραμμα ελέγχου το οποίο χρησιμοποιείται συνήθως στη βιομηχανία προκειμένου να ανιχνευθεί έγκαιρα οποιαδήποτε μη-φυσιολογική (συνήθως ανεπιθύμητη) κατάσταση που επηρεάζει τη διαδικασία παραγωγής. Σε περίπτωση παρουσίας αυτών των ανεπιθύμητων καταστάσεων, η ποιότητα παραγόμενων προϊόντων επιδεινώνεται και έτσι το ποσοστό των ελαττωματικών προϊόντων αυξάνεται.

Σε αρκετές περιπτώσεις, ταξινομούμε ένα προϊόν σαν ελαττωματικό ή μη συμμορφούμενο (defective, nonconforming) εάν τουλάχιστον ένα ποιοτικό χαρακτηριστικό δεν ικανοποιεί τις προδιαγραφές που έχουν τεθεί κατά τη φάση σχεδιασμού του. Με τον όρο ποσοστό ελαττωματικών προϊόντων αναφερόμαστε

στον αριθμό των ελαττωματικών προϊόντων προς το συνολικό αριθμό παραγόμενων προϊόντων. Έστω ότι το ποσοστό των ελαττωματικών προϊόντων που αποδίδει μια παραγωγική διεργασία είναι γνωστό και ίσο με p και έστω ότι επιλέγουμε ανεξάρτητα τυχαία δείγματα μεγέθους n το καθένα. Συμβολίζουμε με X_{ij} την τυχαία μεταβλητή (τ.μ.) που παίρνει τις τιμές 1 ή 0 ανάλογα αν το j-οστό προϊόν του i-οστού δείγματος είναι ελαττωματικό ή όχι. Τότε, η X_{ij} ακολουθεί την κατανομή Bernoulli με πιθανότητα επιτυχίας $p \in (0, 1)$ (δηλ. $X_{ij} \sim B(1, p)$) και η τυχαία μεταβλητή X_i που δηλώνει τον αριθμό των ελαττωματικών προϊόντων στο i-οστό δείγμα ακολουθεί τη διωνυμική κατανομή με παραμέτρους n, p (δηλ. $X_i \sim Bin(n, p)$). Επομένως, σε αυτή την περίπτωση τα διαθέσιμα δεδομένα αναφέρονται ως *αποτελέσματα δοκιμών Bernoulli*, ή γενικά ως δεδομένα χαρακτηριστικών, καθώς δεν είναι δυνατόν να ληφθεί αριθμητική τιμή από το χαρακτηριστικό που περιγράφει την ποιότητα των παραγόμενων προϊόντων· απλά καταγράφουμε την παρουσία ή απουσία ενός χαρακτηριστικού στο αντικείμενο.

Τα πιο γνωστά και ευρέως χρησιμοποιούμενα διαγράμματα ελέγχου για ιδιότητες είναι τα διαγράμματα ελέγχου τύπου Shewhart p και np (Montgomery, 2013) τα οποία χρησιμοποιούνται για την ανίχνευση αλλαγών στην αναλογία ή τον αριθμό των μη συμμορφούμενων προϊόντων που παράγει μια διεργασία, αντίστοιχα. Στην περίπτωση του διαγράμματος p, οι τιμές των σημείων που απεικονίζονται είναι τιμές, γενικά, στο διάστημα [0, 1]. Ωστόσο, υπάρχουν περιπτώσεις όπου οι τιμές του ποιοτικού χαρακτηριστικού βρίσκονται στο [0, 1] αλλά δεν είναι αποτελέσματα Bernoulli πειραμάτων. Για παράδειγμα το ημερήσιο ποσοστό σχετικής υγρασίας σε μία πόλη ή το ποσοστό λίπους στο σώμα ενός ασθενούς. Σε τέτοιες περιπτώσεις, τα συνήθη διαγράμματα p και np δεν μπορούν να εφαρμοστούν και άρα θα πρέπει να αναπτυχθούν εναλλακτικά διαγράμματα ελέγχου τα οποία θα βασίζονται σε ένα πιο κατάλληλο μοντέλο πιθανότητας.

Τα τελευταία χρόνια, παρατηρείται αυξημένο ενδιαφέρον για την ανάπτυξη μοντέλων και τεχνικών ΣΕΔ, για δεδομένα που είναι διπλά οριοθετημένα, π.χ. στο [0, 1] ή στο (0, 1). Μια γνωστή κατανομή για τη μοντελοποίηση αυτού του είδους διεργασιών είναι η κατανομή Βήτα (*Beta distribution)*, η οποία είναι μια ευέλικτη συνεχής κατανομή που μπορεί να μοντελοποιήσει δεδομένα στο (0, 1) και της οποίας η μορφή ποικίλει βάσει των τιμών των παραμέτρων της (Kieschnick and McCullough, 2003). Οι Gupta and Nadarajah (2004) παρουσίασαν αρκετές εφαρμογές της κατανομής Βήτα και φαίνεται ότι είναι οι πρώτοι που παρουσίασαν μια εφαρμογή της σε διαγράμματα ελέγχου. Οι Sant'Anna and ten Caten (2012) ανέπτυξαν και εφάρμοσαν διαγράμματα ελέγχου τύπου Shewhart με βάση την κατανομή Βήτα για την παρακολούθηση ποσοστών (αντί για το σύνηθες διάγραμμα ελέγχου p). Οι Ho et al. (2019) ανέπτυξαν διαγράμματα ελέγχου Shewhart σύμφωνα με τρία διαφορετικά μοντέλα πιθανότητας (Βήτα, Simplex και Unit Gamma) για διεργασίες διπλής οριοθέτησης με τιμές στο διάστημα (0, 1).

Σε όλες τις εργασίες που αναφέρθηκαν παραπάνω, τα προτεινόμενα διαγράμματα είναι διαγράμματα ελέγχου τύπου Shewhart. Είναι γνωστό ότι όταν υπάρχουν περιορισμοί στο μέγεθος του δείγματος που λαμβάνεται από τη διεργασία σε διαδοχικές χρονικές στιγμές, τα διαγράμματα Shewhart δεν είναι αρκετά ευαίσθητα

στην ανίχνευση μικρών ή/και μεσαίων αλλαγών στις παραμέτρους της διεργασίας. Έτσι, δεν μπορούν να τις εντοπίσουν γρήγορα. Αυτό αποδίδεται στο γεγονός ότι σε ένα διάγραμμα ελέγχου Shewhart η απόφαση εάν μια διεργασία είναι εντός ή εκτός ελέγχου βασίζεται αποκλειστικά στην πιο πρόσφατη παρατήρηση. Μια λύση σε αυτό το πρόβλημα αποτελεί η χρήση διαγραμμάτων ελέγχου με μνήμη, όπως το διάγραμμα ελέγχου τύπου EWMA. Το διάγραμμα αυτό προσφέρει αυξημένη ευαισθησία στην ανίχνευση μικρών και μεσαίων μετατοπίσεων στις παραμέτρους της διεργασίας καθώς οι τιμές που απεικονίζονται ενσωματώνουν πληροφορίες τόσο από πρόσφατες όσο και από παλαιότερες παρατηρήσεις.

Στην παρούσα εργασία, προτείνουμε και μελετάμε δίπλευρα διαγράμματα ελέγχου τύπου EWMA μεμονωμένων παρατηρήσεων με βάση τρία διαφορετικά μοντέλα πιθανότητας, με σκοπό την παρακολούθηση της κατανομής χαρακτηριστικών με τιμές στο (0, 1). Η διάρθρωση της εργασίας έχει ως εξής: Στην Ενότητα 2 δίνονται συνοπτικά οι ιδιότητες των τριών κατανομών, Βήτα, Simplex και Unit Gamma, οι οποίες χρησιμοποιούνται ως πιθανά μοντέλα για τη διεργασία. Στην Ενότητα 3, παρουσιάζονται τα διαγράμματα ελέγχου EWMA με βάση τις τρεις παραπάνω κατανομές, μαζί με την αλγοριθμική διαδικασία για το στατιστικό σχεδιασμό τους καθώς και τον υπολογισμό βασικών μέτρων για την αξιολόγηση της απόδοσής τους. Στην Ενότητα 4 παρουσιάζονται τα αποτελέσματα μιας εκτεταμένης μελέτης προσομοίωσης, σχετικά με την απόδοση των δίπλευρων EWMA διαγραμμάτων. Τα προτεινόμενα διαγράμματα EWMA συγκρίνονται με τα αντίστοιχα Shewhart των Ho et al. (2019), από όπου προκύπτει ότι η απόδοσή τους είναι καλύτερη από αυτή των Shewhart διαγραμμάτων. Στην Ενότητα 5 δίδεται μια πρακτική εφαρμογή των προτεινόμενων διαγραμμάτων. Τέλος, στην Ενότητα 6, συνοψίζονται τα συμπεράσματα της μελέτης.

2. ΚΑΤΑΝΟΜΕΣ ΓΙΑ ΤΗ ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΠΟΣΟΣΤΩΝ ΚΑΙ ΑΝΑΛΟΓΙΩΝ

Ακολουθώντας την εργασία των Ho et al. (2019), εξετάζουμε τρία μοντέλα πιθανότητας με στήριγμα το (0, 1) και παρουσιάζουμε εν συντομία τις ιδιότητές τους. Αξίζει να σημειωθεί ότι τα ακόλουθα μοντέλα πιθανότητας δεν είναι τα μοναδικά για την παρακολούθηση διεργασιών διπλής οριοθέτησης στο διάστημα (0, 1). Δείτε π.χ. Lima-Filho et al. (2020).

2.1 Κατανομή Beta

Έστω ότι η τ.μ. X ακολουθεί την κατανομή Βήτα με παραμέτρους $\alpha > 0$, $\beta > 0$ (συμβ. $X \sim Beta(\alpha, \beta)$). Τότε, η συνάρτηση πυκνότητας πιθανότητας (σ.π.π) δίνεται από τη σχέση

$$f_{Beta}(x|\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}, x \in (0,1),$$

όπου $\mathrm{B}(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta)$ είναι η συνάρτηση Βήτα και $\Gamma(y) = \int_0^\infty u^{y-1}e^{-u}du$ είναι η συνάρτηση Γάμμα. Επίσης, η αθροιστική συνάρτηση κατανομής της (α.σ.κ.) δίνεται από τη σχέση

$$F_{Beta}(x|\alpha,\beta) = \frac{\mathrm{B}_x(\alpha,\beta)}{\mathrm{B}(\alpha,\beta)},$$

όπου $\mathrm{B}_x(\alpha,\beta) = \int_0^x t^{\alpha-1}(1-t)^{\beta-1}dt$ είναι η μη πλήρης συνάρτηση Βήτα. Η μέση τιμή και η διασπορά είναι $E(X) = \alpha/(\alpha+\beta)$ και $V(X) = \alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$. Σε αυτή την εργασία, χρησιμοποιούμε την παραμετροποιημένη Βήτα κατανομή (Ferrari and Cribari-Neto, 2004) η οποία προκύπτει αν θέσουμε $\alpha = \mu\phi$ και $\beta = (1-\mu)\phi$, για $\mu \in (0, 1)$ και $\phi > 0$. Επομένως, η σ.π.π. παίρνει την ακόλουθη μορφή

$$f_{Beta}(x|\mu,\phi) = \frac{x^{\mu\phi-1}(1-x)^{(1-\mu)\phi-1}}{\mathrm{B}(\mu\phi,(1-\mu)\phi)}, x \in (0,1),$$

ενώ η μέση τιμή και η διασπορά τώρα είναι ίσες με

$$E(X) = \mu, \quad V(X) = \frac{\mu(1-\mu)}{\phi+1}. \tag{1}$$

Η παράμετρος ϕ μπορεί να ερμηνευθεί ως παράμετρος ακριβείας αφού όταν το μ παραμένει αμετάβλητο, όσο μεγαλύτερη είναι η τιμή της παραμέτρου ϕ, τόσο μικρότερη είναι η διασπορά της $Beta(\mu, \phi)$.

2.2 Κατανομή Simplex

Έστω ότι η τ.μ. X ακολουθεί την κατανομή Simplex με παραμέτρους $\mu \in (0, 1)$, $\sigma^2 > 0$ (συμβ. $X \sim S(\mu, \sigma^2)$). Τότε, η σ.π.π. δίνεται ως εξής (Barndorff-Nielsen and Jorgensen, 1991)

$$f_{Simplex}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2x^3(1-x)^3}}e^{\left(-\frac{1}{2\sigma^2}d(x;\mu)\right)}, \ x \in (0,1),$$

όπου ο όρος $d(x; \mu)$ είναι γνωστός ως συνάρτηση απόκλισης και δίνεται από τη σχέση

$$d(x;\mu) = \frac{(x-\mu)^2}{x(1-x)\mu^2(1-\mu)^2}.$$

Η μέση τιμή και η διασπορά είναι ίσες με

$$E(X) = \mu, \quad V(X) = \frac{1}{\sqrt{2\sigma^2}}e^{\left(\frac{1}{2\sigma^2\mu^2(1-\mu)^2}\right)}\Gamma\left(\frac{1}{2},\frac{1}{2\sigma^2\mu^2(1-\mu)^2}\right), \tag{2}$$

όπου $\Gamma(r,s) = \int_s^\infty u^{r-1}e^{-u}du$ είναι η πλήρης συνάρτηση Γάμμα.

2.3 Κατανομή Unit Gamma

Έστω ότι η τ.μ. X ακολουθεί την κατανομή Unit Gamma με παραμέτρους, $\theta > 0$, $\tau > 0$ (συμβ. $X \sim uGA(\theta, \tau)$). Τότε, η σ.π.π. δίνεται ως εξής (Grassia, 1977)

$$f_{uGA}(x|\theta,\tau) = \frac{\theta^\tau}{\Gamma(\tau)}x^{\theta-1}\left(\log\left(\frac{1}{x}\right)\right)^{\tau-1}, x \in (0,1),$$

ενώ η μέση τιμή και η διασπορά είναι ίσες με

$$E(X) = \left(\frac{\theta}{\theta+1}\right)^\tau, \quad V(X) = \left(\frac{\theta}{\theta+2}\right)^\tau - \left(\frac{\theta}{\theta+1}\right)^{2\tau}.$$

Σε αυτή την εργασία, χρησιμοποιούμε την παραμετροποιημένη Unit Gamma κατανομή (Mousa et al., 2016) η οποία προκύπτει θέτοντας $\theta = \mu^{1/\tau} / (1 - \mu^{1/\tau})$. Τότε, η σ.π.π. παίρνει την ακόλουθη μορφή

$$f_{uGA}(x|\mu,\tau) = \frac{\left(\frac{\mu^{1/\tau}}{(1-\mu^{1/\tau})}\right)^{\tau}}{\Gamma(\tau)} x^{\frac{\mu^{1/\tau}}{(1-\mu^{1/\tau})}-1} \left(\log\left(\frac{1}{x}\right)\right)^{\tau-1}, x \in (0,1),$$

όπου $\mu \in (0, 1)$, $\tau > 0$. Επομένως, για την παραμετροποιημένη Unit Gamma κατανομή, η μέση τιμή και η διασπορά ισούνται με

$$E(X) = \mu, \quad V(X) = \mu\left(\frac{1}{(2-\mu^{1/\tau})^{\tau}} - \mu\right). \tag{3}$$

3. ΔΙΑΓΡΑΜΜΑΤΑ ΕΛΕΓΧΟΥ EWMA ΓΙΑ ΔΙΕΡΓΑΣΙΕΣ ΔΙΠΛΗΣ ΟΡΙΟΘΕΤΗΣΗΣ

Σε αυτήν την ενότητα, θα αναπτύξουμε δίπλευρα διαγράμματα ελέγχου τύπου EWMA για την παρακολούθηση διεργασιών με δεδομένα τα οποία αποτελούν μεμονωμένες παρατηρήσεις στο (0,1). Αρχικά, όταν η διεργασία βρίσκεται εντός ελέγχου συμβολίζουμε το μέσο επίπεδο της διεργασίας της ως $\mu_{0,X} = \mu_0$, το οποίο είναι η εντός ελέγχου μέση αναλογία. Επίσης, οι παράμετροι διασποράς συμβολίζονται ως ϕ_0 (για την κατανομή Βήτα), ως σ_0 (για την κατανομή Simplex) και ως τ_0 (για την κατανομή Unit Gamma). Επομένως, δηλώνουμε την εντός ελέγχου διασπορά της διεργασίας ως $\sigma^2_{0,X}$, η οποία υπολογίζεται με τη χρήση των εξισώσεων (1)-(3) για το $V(X)$, αντικαθιστώντας τις εντός ελέγχου τιμές για το κάθε μοντέλο.

Όταν η διεργασία είναι εκτός ελέγχου, υποθέτουμε ότι η παρουσία ανεπιθύμητων αιτιών επηρεάζει μόνο το μέσο επίπεδο της διεργασίας, το οποίο μετατοπίζεται από μ_0 σε $\mu_1 \neq \mu_0$, με $\mu_1 \in (0,1)$. Συγκεκριμένα, όταν $\mu_1 > \mu_0$, το μέσο επίπεδο της διεργασίας έχει αυξηθεί ενώ όταν $\mu_1 < \mu_0$, το μέσο ποσοστό έχει μειωθεί. Στην παρούσα εργασία υποθέτουμε ότι η τιμή της παραμέτρου διασποράς παραμένει αμετάβλητη υπό την παρουσία ειδικών αιτιών στη διεργασία. Στόχος είναι η ανίχνευση μεταβολών διάφορων μεγεθών στο εντός ελέγχου μέσο επίπεδο της διεργασίας $\mu_{0,X} = \mu$, το οποίο δεν επηρεάζεται άμεσα από αλλαγές στα ϕ_0 , σ_0 και τ_0. Πιστεύουμε πως για την ανίχνευση αλλαγών σε δύο ή περισσότερες παραμέτρους πρέπει να χρησιμοποιηθεί αντίστοιχος αριθμός διαγραμμάτων ελέγχου. Αυτό μπορεί να γίνει σε μελλοντική εργασία.

Τα διαγράμματα ελέγχου τύπου EWMA προτάθηκαν από τον Roberts (1959) και τα σημεία που απεικονίζονται στο διάγραμμα αυτό υπολογίζονται από την παρακάτω στατιστική συνάρτηση

$$Z_t = \lambda X_t + (1 - \lambda) Z_{t-1}, \quad t = 1, 2, \ldots,$$

όπου με X_t συμβολίζεται η τιμή της μεμονωμένης παρατήρησης που λαμβάνεται τη χρονική στιγμή t. Επίσης, η αρχική τιμή για το $Z_0 = \mu_0$, δηλαδή ισούται με την εντός ελέγχου μέση τιμή του X. Η σταθερά $\lambda \in (0,1]$ καλείται παράμετρος εξομάλυνσης και καθορίζει τη βαρύτητα που δίνεται στις πρόσφατες παρατηρήσεις της διεργασίας. Έτσι, για μικρές τιμές του λ (πλησιέστερα στο 0) δίνεται μεγαλύτερη βαρύτητα στις

λιγότερο πρόσφατες παρατηρήσεις ενώ για μεγάλες τιμές του λ (πλησιέστερα στο 1) δίνεται μεγαλύτερη βαρύτητα στις πιο πρόσφατες παρατηρήσεις. Σύμφωνα με τον Montgomery (2013), για το λ επιλέγεται κάποια τιμή στο διάστημα [0,05, 0,25]. Επίσης, για $\lambda = 1$, ένα δίπλευρο EWMA διάγραμμα ταυτίζεται με ένα δίπλευρο διάγραμμα ελέγχου τύπου Shewhart καθώς $Z_t = X_t$. Η λειτουργία ενός EWMA διαγράμματος για την παρακολούθηση της μέσης τιμής μιας παραγωγικής διεργασίας βασίζεται στη στατιστική συνάρτηση Z_t, τιμές τις οποίας απεικονίζονται σε αυτό, με όρια ελέγχου τα

$$LCL = \mu_{0,X} - L\sigma_{0,X}\sqrt{\frac{\lambda}{2-\lambda}},\ CL = \mu_{0,X},\ UCL = \mu_{0,X} + L\sigma_{0,X}\sqrt{\frac{\lambda}{2-\lambda}}\ .$$

Τα παραπάνω όρια είναι επίσης γνωστά ως όρια ελέγχου σταθερής κατάστασης για το δίπλευρο EWMA διάγραμμα. Το διάγραμμα σηματοδοτεί για εκτός ελέγχου διεργασία ακριβώς στο *t*-οστό στάδιο της δειγματοληψίας εάν $Z_t \notin [LCL, UCL]$. Ο αριθμός των σημείων που απεικονίζονται σε ένα διάγραμμα μέχρι να εμφανιστεί για πρώτη φορά ένα σημείο εκτός των ορίων ελέγχου καλείται μήκος ροής (*run length*) και είναι μια τ.μ. Η κατανομή του μήκους ροής χρησιμοποιείται για την αξιολόγηση ενός διαγράμματος ελέγχου. Το πιο συνηθισμένο μέτρο απόδοσης είναι η μέση τιμή της κατανομής του μήκους ροής (*average run length*) ή $ARL = E(RL)$ ενώ χρησιμοποιούμε επιπλέον την τυπική απόκλιση της κατανομής του μήκους ροής (*standard deviation run-length*) ή $SDRL = \sqrt{V(RL)}$ και τη διάμεσο της κατανομής του μήκους ροής (*median run length*) ή *MRL*. Για περισσότερες λεπτομέρειες σχετικά με τη χρήση των *ARL*, *SDRL* και *MRL* ως μέτρων απόδοσης δείτε Maravelakis et al. (2005) Στην περίπτωση ενός δίπλευρου διαγράμματος ελέγχου τύπου Shewhart, η κατανομή του μήκους ροής είναι μια γεωμετρική κατανομή με παράμετρο p_{out}, όπου η παράμετρος p_{out} είναι η πιθανότητα να παρατηρηθεί σημείο εκτός των ορίων ελέγχου. Επομένως, τα *ARL*, *SDRL* και *MRL* υπολογίζονται ως εξής:

$$ARL = \frac{1}{p_{out}},\ SDRL = \frac{\sqrt{1-p_{out}}}{p_{out}},\ MRL = \left\lceil \frac{\log(0.5)}{\log(1-p_{out})} \right\rceil,$$

όπου $\lceil x \rceil$ είναι ο ελάχιστος ακέραιος ο οποίος είναι μεγαλύτερος ή ίσος του x.

Οι παράμετροι σχεδιασμού του διαγράμματος EWMA είναι το *L* που εκφράζει την απόσταση των ορίων ελέγχου από την κεντρική γραμμή και το λ που είναι η παράμετρος εξομάλυνσης. Ο σχεδιασμός του συγκεκριμένου διαγράμματος βασίζεται στην κατάλληλη επιλογή αυτών των δύο παραμέτρων ώστε το διάγραμμα να έχει την επιθυμητή τιμή για το εντός ελέγχου μέσο μήκος ροής. Είναι γνωστό επίσης ότι στην περίπτωση ενός δίπλευρου διαγράμματος ελέγχου τύπου EWMA, η κατανομή του μήκους ροής δεν είναι η γεωμετρική (Crowder, 1987). Σε αυτήν την εργασία, εφαρμόζουμε τη μέθοδο προσομοίωσης Monte Carlo για τον προσδιορισμό της παραμέτρου *L* για μια δεδομένη τιμή λ, έτσι ώστε η απόδοση του διαγράμματος να είναι η επιθυμητή. Συγκεκριμένα, τα βήματα της αλγοριθμικής διαδικασίας για τον προσδιορισμό των τιμών (λ, *L*) δίνονται παρακάτω (δείτε Alevizakos and Koukouvinos, 2019).

B1. Επιλέγουμε τις εντός ελέγχου τιμές των παραμέτρων για κάθε κατανομή. Αυτές είναι (μ_0, ϕ_0) για την κατανομή Βήτα, (μ_0, σ_0) για την κατανομή Simplex και (μ_0, τ_0) για την κατανομή Unit Gamma.

B2. Επιλέγουμε την επιθυμητή τιμή για το εντός ελέγχου *ARL*, έστω αυτή ARL_0, και προσδιορίζουμε την τιμή της παραμέτρου λ.

B3. Χρησιμοποιούμε ως αρχική τιμή $L = 0.001$ και υπολογίζουμε τα όρια ελέγχου σταθερής κατάστασης.

B4. Προσομοιώνουμε 10000 εντός ελέγχου διεργασίες με βάση τις παραμέτρους του βήματος 1 και για κάθε προσομοίωση καταγράφουμε τον αριθμό των σημείων μέχρι τον πρώτο εσφαλμένο συναγερμό.

B5. Υπολογίζουμε το εντός ελέγχου *ARL* ως τον δειγματικό μέσο όρο των 10000 τιμών *RL* που προέκυψαν από το βήμα 4. Αν $ARL \notin [ARL_0 - \xi, ARL_0 + \xi]$, όπου $\xi = 5$ είναι ένας αριθμός ανοχής, αυξάνουμε την τιμή της παραμέτρου *L* κατά 0.001 και επιστρέφουμε στο βήμα 4. Διαφορετικά, προχωράμε στο βήμα 6.

B6. Χρησιμοποιούμε την τιμή της παραμέτρου *L* η οποία έχει βρεθεί από το προηγούμενο βήμα και καθορίζουμε τα όρια ελέγχου για το δίπλευρο διάγραμμα EWMA και δηλώνουμε την διαδικασία ως εκτός ελέγχου εάν $Z_t \notin [LCL, UCL]$.

4. ΜΕΛΕΤΗ ΠΡΟΣΟΜΟΙΩΣΗΣ

Σε αυτήν την ενότητα, παρουσιάζουμε τα αποτελέσματα μιας εκτεταμένης αριθμητικής μελέτης ανθεκτικότητας σχετικά με το σχεδιασμό και την απόδοση των προτεινόμενων διαγραμμάτων ελέγχου EWMA. Στόχος είναι η σύγκριση αυτών των διαγραμμάτων με τα αντίστοιχα διαγράμματα ελέγχου τύπου Shewhart για ποσοστά και αναλογίες που μελετήθηκαν από τους Ho et al. (2019). Έπειτα, διερευνούμε πόσο επηρεάζονται τα EWMA διαγράμματα που βασίζονται στις τρεις κατανομές που αναφέραμε παραπάνω όταν τα όρια ελέγχου υπολογίζονται από διαφορετικό μοντέλο και όχι από το πραγματικό.

Όπως έχουμε ήδη αναφέρει, η παρουσία ειδικών αιτιών μεταβλητότητας επηρεάζει μόνο το μέσο επίπεδο της διεργασίας, το οποίο μετατοπίζεται από μ_0 σε $\mu_1 \neq \mu_0$. Ειδικότερα, θεωρούμε ότι η εντός ελέγχου μέση τιμή της διεργασίας είναι $\mu_0 = 0.2$ και η εκτός ελέγχου μέση τιμή είναι $\mu_1 = \mu_0 \pm \Delta$, με $\Delta \in \{0.02, 0.04, 0.06, 0.08\}$ να δηλώνει τα διάφορα επίπεδα μετατόπισης. Σχετικά με τις παραμέτρους διασποράς του κάθε μοντέλου, ακολουθώντας την εργασία των Ho et al. (2019), θεωρούμε ότι αυτές παραμένουν αμετάβλητες. Ωστόσο επιλέγονται τέσσερα επίπεδα διασποράς (δείτε Πίνακα 1), τα οποία δηλώνουν αντίστοιχα τις περιπτώσεις με τις μικρότερες και τις μεγαλύτερες διασπορές. Παρατηρούμε ότι τα τρία μοντέλα παρουσιάζουν ομοιότητες μεταξύ τους. Για λόγους οικονομίας χώρου, θα παρουσιάσουμε αποτελέσματα μόνο για τις περιπτώσεις 2 και 4.

Στη συνέχεια παρουσιάζονται τα αποτελέσματα από τη σύγκριση της απόδοσης των δίπλευρων διαγραμμάτων ελέγχου EWMA και Shewhart για κάθε ένα από τα τρία μοντέλα. Υποθέτουμε ότι η πιθανότητα εσφαλμένου συναγερμού ισούται με $\alpha = 0.0027$, ώστε να επιτύχουμε εντός ελέγχου *ARL* όσο το δυνατόν πιο κοντά στην

επιθυμητή τιμή $ARL_0 = 370.4$. Εφαρμόζοντας τα βήματα της παραπάνω αλγοριθμικής διαδικασίας προσδιορίζεται η τιμή L για κάθε μοντέλο, για κάθε περίπτωση των παραμέτρων διασποράς και για τρεις τιμές του $\lambda = 0.05, 0.10$ και 0.20.

Έπειτα, για δεδομένες τιμές των λ και L, εφαρμόζουμε ξανά τον προηγούμενο αλγόριθμο αλλά για τις μετατοπισμένες τιμές του μέσου αυτή τη φορά ($\mu_1 = \mu_0 \pm \Delta$). Πλέον, οι υπολογισμοί αφορούν την εκτός ελέγχου απόδοση των διαγραμμάτων EWMA. Τα αποτελέσματα της προσομοίωσης (για το διάγραμμα EWMA) δίνονται στους Πίνακες 2-4. Να σημειωθεί ότι για το Shewhart διάγραμμα χρησιμοποιήσαμε όρια ελέγχου πιθανότητας ίσων ουρών, δηλαδή οι τιμές των ορίων UCL, LCL, προκύπτουν ως λύσεις των εξισώσεων $P(X > UCL) = a/2$ και $P(X \leq LCL) = a/2$ όταν η διεργασία είναι εντός στατιστικού ελέγχου. Επίσης, στις γραμμές 'UCL', 'LCL' του Πίνακα 2 δίνονται τα όρια ελέγχου κάθε διαγράμματος ενώ οι τιμές με έντονη γραφή αντιστοιχούν στην ελάχιστη τιμή του ARL μεταξύ των διαφορετικών μοντέλων, για τη δεδομένη μετατόπιση στο μ_0.

Πίνακας 1. *Μεταβλητότητα κάθε μοντέλου για $\mu_0 = 0.2$, Περιπτώσεις 1-4.*

	Beta		Simplex		Unit Gamma	
Case	ϕ_0	$\sigma_{0,X}$	σ_0	$\sigma_{0,X}$	τ_0	$\sigma_{0,X}$
1	290	0.02344842	0.37	0.02355733	155	0.02582828
2	148	0.03276928	0.50	0.03170082	96	0.03279827
3	80	0.04444444	0.71	0.04460488	51	0.04493217
4	31	0.07071068	1.20	0.07309293	20	0.07138937

Σύμφωνα με τα παρακάτω αποτελέσματα, παρατηρούμε ότι η χρήση των διαγραμμάτων ελέγχου EWMA βελτιώνει την ικανότητα ανίχνευσης μεταβολών στο μέσο επίπεδο της διεργασίας συγκριτικά με το Shewhart διάγραμμα, καθώς προκύπτουν μικρότερες τιμές ARL_1. Για την κατανομή Βήτα παρατηρούμε ότι το EWMA διάγραμμα για $\lambda = 0.05$ είναι το βέλτιστο στην ανίχνευση μικρών μετατοπίσεων, δηλαδή για $\mu_1 \in [0.18,0.22]$ και όσο η τιμή του μ_1 αυξάνεται, μια μεγαλύτερη τιμή του λ, όπως $\lambda = 0.10$ ή 0.20, παρέχει μια πιο βελτιωμένη απόδοση. Σχετικά με την κατανομή Simplex παρατηρούμε ότι το EWMA διάγραμμα για $\lambda = 0.05$ επιτυγχάνει το ελάχιστο ARL_1 και για τις δύο περιπτώσεις των παραμέτρων διασποράς. Παρόμοια αποτελέσματα προκύπτουν και για την Unit Gamma κατανομή.

Στους Πίνακες 3-5 δίνονται τα αποτελέσματα της μελέτης σχετικά με την ανθεκτικότητα των διαγραμμάτων ελέγχου EWMA στη χρήση 'λανθασμένων ορίων ελέγχου'

Πίνακας 2. *Σύγκριση Δίπλευρων Διαγραμμάτων Shewhart και EWMA*

Περίπτωση Κατανομής Βήτα											
ϕ_0	μ	SH	λ=0.05	λ=0.10	λ=0.20	ϕ_0	μ	SH	λ=0.05	λ=0.10	λ=0.20
148	0.12	**2.36**	5.04	4.35	3.81	31	0.12	15.68	10.15	**9.11**	9.21
	0.14	5.72	6.42	5.52	**4.94**		0.14	36.30	13.93	**13.28**	15.30
	0.16	20.11	9.46	8.40	**8.11**		0.16	90.21	**23.24**	24.96	37.61
	0.18	100.51	**21.09**	21.40	27.07		0.18	220.61	**64.30**	87.98	188.96
	0.20	370.37	370.14	370.44	370.32		0.20	370.37	370.50	370.66	370.16
	0.22	129.21	**21.05**	21.19	24.19		0.22	289.82	**59.48**	66.36	78.97
	0.24	32.24	9.55	8.47	**8.12**		0.24	155.75	23.13	**23.07**	26.48
	0.26	10.61	6.47	5.64	**5.09**		0.26	81.39	14.17	**13.30**	13.57
	0.28	4.53	5.09	4.41	**3.88**		0.28	44.63	10.37	9.30	**9.02**
	UCL	0.3081	0.213	0.220	0.231		*UCL*	0.4518	0.228	0.244	0.268
	LCL	0.1133	0.187	0.180	0.169		*LCL*	0.0450	0.172	0.156	0.132
Περίπτωση Κατανομής Simplex											
σ_0	μ	SH	λ=0.05	λ=0.10	λ=0.20	σ_0	μ	SH	λ=0.05	λ=0.10	λ=0.20
0.50	0.12	**1.87**	3.34	4.21	3.66	1.20	0.12	35.02	**4.90**	9.63	10.69
	0.14	5.72	**4.20**	5.37	4.76		0.14	80.59	**6.17**	14.46	21.26
	0.16	24.83	**5.95**	8.16	7.83		0.16	173.86	**9.08**	29.74	74.03
	0.18	128.23	**13.34**	21.06	28.78		0.18	332.35	**20.24**	125.34	546.56
	0.20	370.48	366.76	375.08	371.43		0.20	370.40	373.08	370.88	370.23
	0.22	90.00	**13.19**	19.51	21.54		0.22	191.01	**20.17**	60.46	67.19
	0.24	21.68	**6.08**	8.26	7.72		0.24	84.09	**9.25**	23.01	24.83
	0.26	7.94	**4.28**	5.52	4.94		0.26	41.59	**6.36**	13.65	13.68
	0.28	3.94	**3.52**	4.38	3.83		0.28	23.42	**5.01**	9.75	9.26
	UCL	0.3085	0.213	0.220	0.230		*UCL*	0.4743	0.229	0.246	0.271
	LCL	0.1205	0.187	0.180	0.170		*LCL*	0.0594	0.171	0.154	0.129
Περίπτωση Κατανομής Unit Gamma											
τ_0	μ	SH	λ=0.05	λ=0.10	λ=0.20	τ_0	μ	SH	λ=0.05	λ=0.10	λ=0.20
96	0.12	**2.23**	5.05	4.35	3.79	20	0.12	14.82	10.22	**9.22**	9.47
	0.14	5.36	6.43	5.58	**4.99**		0.14	34.21	14.10	**13.56**	15.87
	0.16	18.84	9.44	8.39	**8.22**		0.16	85.15	**23.67**	25.25	40.62
	0.18	95.63	**20.95**	21.59	27.55		0.18	210.67	**65.20**	92.48	219.01
	0.20	370.37	370.50	370.30	370.97		0.20	370.37	370.05	370.49	370.41
	0.22	138.59	21.44	**20.94**	24.58		0.22	316.76	**60.87**	67.42	79.77
	0.24	35.74	9.54	8.52	**8.21**		0.24	184.20	23.67	**23.59**	27.75
	0.26	11.80	6.50	5.64	**5.11**		0.26	102.06	14.31	**13.35**	14.13
	0.28	4.97	5.12	4.42	**3.92**		0.28	58.31	10.46	9.43	**9.34**
	UCL	0.3155	0.213	0.220	0.231		*UCL*	0.4878	0.228	0.244	0.269
	LCL	0.1184	0.187	0.180	0.169		*LCL*	0.0596	0.172	0.156	0.131

Παρατηρούμε ότι από τη χρήση εσφαλμένων ορίων ελέγχου επηρεάζεται κυρίως η εντός ελέγχου απόδοση των διαγραμμάτων EWMA. Σημαντικές διαφορές σημειώνονται και στην εκτός ελέγχου απόδοση, ειδικά για μετατοπίσεις στο διάστημα [0.18, 0.22]. Όσον αφορά την εντός ελέγχου απόδοση, άλλοτε είναι

μεγαλύτερη και άλλοτε μικρότερη της επιθυμητής τιμής 370.4. Γενικά, μια μεγάλη τιμή στην εντός ελέγχου απόδοση είναι επιθυμητή, εκτός και αν επηρεάζεται η εκτός ελέγχου απόδοση. Για την περίπτωση της κατανομής Βήτα (Πίνακας 3), παρατηρούμε ότι η χρήση Simplex ορίων ελέγχου έχει ως αποτέλεσμα μεγάλες διαφοροποιήσεις στην εντός ελέγχου απόδοση για την περίπτωση 4. Από την άλλη, οι εκτός ελέγχου μετρήσεις είναι χαμηλότερες από αυτές όταν χρησιμοποιούνται τα πραγματικά όρια ελέγχου για την περίπτωση 2. Αυτό συμβαίνει επειδή και η εντός ελέγχου απόδοση στην περίπτωση αυτή είναι μικρότερη της επιθυμητής τιμής 370.4. Συνεπώς, μια αύξηση στο ρυθμό εσφαλμένου συναγερμού οδηγεί σε μείωση της εκτός ελέγχου απόδοσης. Για την περίπτωση της κατανομής Simplex (Πίνακας 4), παρατηρούμε ότι η εντός ελέγχου απόδοση είναι μεγαλύτερη από την επιθυμητή υπό τη χρήση εσφαλμένων ορίων ελέγχου, ενώ η εκτός ελέγχου απόδοση δεν επηρεάζεται σημαντικά. Τέλος, για την περίπτωση της κατανομής Unit Gamma (Πίνακας 5), παρατηρούμε ότι υπάρχει σημαντική απόκλιση στην εντός ελέγχου απόδοση όταν χρησιμοποιούνται Simplex όρια ελέγχου, ενώ σημαντική απόκλιση παρατηρείται και στην εκτός ελέγχου απόδοση.

Πίνακας 3. *Απόδοση Δίπλευρων EWMA Διαγραμμάτων, Beta Chart, λ=0.05.*

		Beta - True			Simplex			Unit Gamma		
	μ	ARL	SDRL	MRL	ARL	SDRL	MRL	ARL	SDRL	MRL
Case 2:	0.12	5.04	0.77	5	5.01	0.76	5	5.02	0.77	5
$\phi_0 = 148$	0.14	6.42	1.23	6	6.40	1.24	6	6.38	1.27	6
$\sigma_0 = 0.50$	0.16	9.46	2.58	9	9.41	2.58	9	9.40	2.55	9
$\tau_0 = 96$	0.18	21.09	9.89	19	20.95	10.01	19	20.97	9.84	19
	0.20	370.14	351.44	265	363.33	343.52	259	357.90	345.13	253
	0.22	21.05	10.73	19	21	10.59	19	21.13	10.74	19
	0.24	9.55	3.02	9	9.51	2.98	9	9.49	2.97	9
	0.26	6.47	1.56	6	6.45	1.55	6	6.46	1.56	6
	0.28	5.09	1.02	5	5.08	1	5	5.10	1.01	5
	LCL		0.187			0.187			0.187	
	UCL		0.213			0.213			0.213	
Case 4:	0.12	10.15	2.69	10	10.45	2.67	10	10.07	2.61	10
$\phi_0 = 31$	0.14	13.93	4.71	13	14.52	4.96	14	13.85	4.66	13
$\sigma_0 = 1.20$	0.16	23.24	11.01	21	24.34	11.64	22	23.09	11.03	21
$\tau_0 = 20$	0.18	64.30	46.81	52	69.74	52.14	55	63.25	46.57	50
	0.20	370.50	360.16	257	439.72	424.28	312	351.96	344.37	243
	0.22	59.48	44.65	47	64.97	50.17	50.5	59.27	44.36	47
	0.24	23.13	12.74	20	24.21	13.02	21	23.15	12.42	20
	0.26	14.17	5.92	13	14.65	6.07	13	14.04	5.92	13
	0.28	10.37	3.62	10	10.69	3.71	10	10.27	3.61	10
	LCL		0.172			0.171			0.172	
	UCL		0.228			0.229			0.228	

Πίνακας 4. *Απόδοση Δίπλευρων EWMA Διαγραμμάτων, Simplex Chart, λ=0.10.*

		Simplex - True			Beta			Unit Gamma		
	μ	ARL	SDRL	MRL	ARL	SDRL	MRL	ARL	SDRL	MRL
Case 2:	0.12	4.21	0.47	4	4.25	0.48	4	4.26	0.48	4
$\phi_0 = 148$	0.14	5.37	0.90	5	5.45	0.92	5	5.45	0.93	5
$\sigma_0 = 0.50$	0.16	8.16	2.25	8	8.29	2.29	8	8.30	2.29	8
$\tau_0 = 96$	0.18	21.06	11.61	18	21.91	12.49	19	21.92	12.26	19
	0.20	375.08	368.08	264	416.84	410.89	288	428.18	415.18	304
	0.22	19.51	12.36	16	20.69	12.97	17	20.20	12.66	17
	0.24	8.26	3.33	8	8.45	3.46	8	8.42	3.37	8
	0.26	5.52	1.73	5	5.63	1.74	5	5.64	1.77	5
	0.28	4.38	1.17	4	4.41	1.15	4	4.43	1.15	4
	LCL		0.180			0.180			0.180	
	UCL		0.220			0.220			0.220	
Case 4:	0.12	9.63	2.23	9	9.20	2.16	9	9.23	2.19	9
$\phi_0 = 31$	0.14	14.46	5.20	13	13.64	4.89	13	13.69	4.94	13
$\sigma_0 = 1.20$	0.16	29.74	17.79	25	26.85	15.51	23	26.94	15.84	23
$\tau_0 = 20$	0.18	125.34	110.46	91	104.56	88.94	77	103.40	90.40	76
	0.20	370.88	363.92	261	310.84	305.53	216	301.09	294.40	209
	0.22	60.46	51.95	45	54.62	47.05	40	55.43	47.55	42
	0.24	23.01	16.08	19	21.76	15.25	18	21.95	15.25	18
	0.26	13.65	7.94	12	12.97	7.45	11	13.05	7.60	11
	0.28	9.75	4.84	9	9.31	4.59	8	9.31	4.66	8
	LCL		0.154			0.156			0.156	
	UCL		0.246			0.244			0.244	

Αξίζει να σημειωθεί ότι τα όρια ελέγχου δίνονται με ακρίβεια τριών δεκαδικών ψηφίων. Παρόλο που οι τιμές είναι αρκετά κοντά μεταξύ τους αριθμητικά, υπάρχουν μικρές διαφοροποιήσεις οι οποίες έχουν ως αποτελέσμα τις διαφοροποιήσεις κυρίως στην εντός ελέγχου απόδοση. Για πρακτικές εφαρμογές θα προτείναμε τη χρήση ενός EWMA διαγράμματος με $\lambda = 0.05$ καθώς έχει καλύτερη απόδοση από το Shewhart διάγραμμα σε μικρές μεταβολές, ενώ η εκτός ελέγχου απόδοσή του μπορεί να θεωρηθεί ότι είναι η ίδια, είτε χρησιμοποιηθούν τα όρια ελέγχου υπό το «σωστό» μοντέλο είτε όχι. Τέλος, για περισσότερες πληροφορίες αναφορικά με όλα τα αποτελέσματα για τη μελέτη ανθεκτικότητας (για μονόπλευρα και δίπλευρα διαγράμματα) όπως επίσης και για τις υπόλοιπες περιπτώσεις των παραμέτρων διασποράς, δείτε Λαφατζή, (2021).

5. ΕΜΠΕΙΡΙΚΗ ΜΕΛΕΤΗ

Σε αυτή την ενότητα παρουσιάζεται μια πρακτική εφαρμογή των προτεινόμενων διαγραμμάτων ελέγχου EWMA. Τα διαθέσιμα δεδομένα αφορούν το ποσοστό των μολυσμένων φιστικιών (Sant' Anna and ten Caten, 2012).

Πίνακας 5. *Απόδοση Δίπλευρων EWMA Διαγραμμάτων, Unit Gamma Chart, λ=0.20.*

		Unit Gamma - True			Beta			Simplex		
	μ	ARL	SDRL	MRL	ARL	SDRL	MRL	ARL	SDRL	MRL
Case 2:	0.12	3.79	0.74	4	3.76	0.73	4	3.66	0.71	4
ϕ_0 = 148	0.14	4.99	1.36	5	4.89	1.29	5	4.75	1.23	5
σ_0 = 0.50	0.16	8.22	3.45	7	7.98	3.35	7	7.56	3.15	7
τ_0 = 96	0.18	27.55	20.99	21	26.69	20.02	21	23.99	17.46	19
	0.20	370.97	369.27	256	340.51	333.03	237	262.56	256.87	182
	0.22	24.58	19.58	19	23.68	18.46	18	21.43	16.11	17
	0.24	8.21	3.94	7	7.99	3.76	7	7.65	3.64	7
	0.26	5.11	1.69	5	4.99	1.62	5	4.83	1.55	5
	0.28	3.92	0.98	4	3.87	0.97	4	3.75	0.92	4
	LCL		0.169			0.169			0.170	
	UCL		0.231			0.231			0.230	
Case 4:	0.12	9.47	3.98	8	9.19	3.82	8	9.91	4.21	9
ϕ_0 = 31	0.14	15.87	9.37	13	15.49	8.94	13	17.23	10.51	14
σ_0 = 1.20	0.16	40.62	32.82	31	37.62	29.59	29	47.98	40.20	36
τ_0 = 20	0.18	219.01	212.44	154	192.60	187.68	134	286.12	277.19	200
	0.20	370.41	371.31	256	330.26	326.91	231	459.47	461.78	317
	0.22	79.77	74.21	57	75.79	70.80	54	91.58	86.22	65
	0.24	27.75	21.93	21	26.58	21.18	21	30.05	24.43	23
	0.26	14.13	9.25	12	13.77	9.03	11	15.09	9.85	12
	0.28	9.34	4.93	8	9.09	4.73	8	9.75	5.19	8
	LCL		0.131			0.132			0.129	
	UCL		0.269			0.268			0.271	

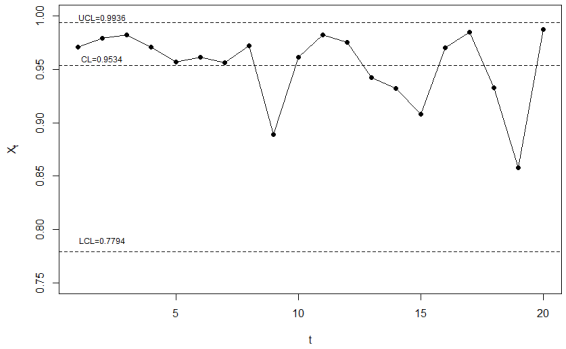
Στη διάθεσή μας έχουμε 34 μεμονωμένες μετρήσεις οι οποίες δίνουν την αναλογία μολυσμένων φιστικιών σε διαδοχικές δειγματοληψίες. Οι πρώτες 20 παρατηρήσεις χρησιμοποιούνται ως δείγματα για την ανάλυση Φάσης Ι. Στον Πίνακα 6 δίνονται οι εκτιμήσεις των παραμέτρων για τα τρία μοντέλα όπως επίσης και οι τιμές των κριτηρίων πληροφορίας AIC και BIC, από τις οποίες συμπεραίνουμε ότι η κατανομή Simplex είναι το μοντέλο με την καλύτερη προσαρμογή στα δεδομένα.

Πίνακας 6. *Εκτιμήσεις παραμέτρων για κάθε μοντέλο και κριτήρια επιλογής μοντέλου.*

Μοντέλο	MLEκτιμητές	AIC	BIC
Beta	$\hat{\mu}$ = 0.9533 $\hat{\phi}$ = 48.9438	-85.455	-83.464
Simplex	$\hat{\mu}$ = 0.9534 $\hat{\sigma}$ = 3.5742	**-88.653**	**-86.662**
Unit Gamma	$\hat{\mu}$ = 0.9534 $\hat{\tau}$ = 2.2798	-85.455	-83.463

Στη συνέχεια κατασκευάζουμε διαγράμματα ελέγχου τύπου Shewhart και EWMA σύμφωνα με την κατανομή Simplex. Αρχικά, κατασκευάζουμε το δίπλευρο διάγραμμα ελέγχου τύπου Shewhart Φάσης I προκειμένου να ελέγξουμε αν η διεργασία ήταν εντός ελέγχου όταν συλλέχθηκαν τα δεδομένα. Όλα τα διαγράμματα έχουν $ARL_0 = 370.4$. Από το Σχήμα 1 παρατηρούμε ότι όλα τα σημεία βρίσκονται εντός των ορίων ελέγχου και άρα μπορούμε να θεωρήσουμε ότι η διεργασία ήταν εντός ελέγχου όταν συλλέχθηκαν τα δεδομένα.

Σχήμα 1. *Shewhart διάγραμμα Φάσης I για την αναλογία μη-μολυσμένων φιστικιών.*

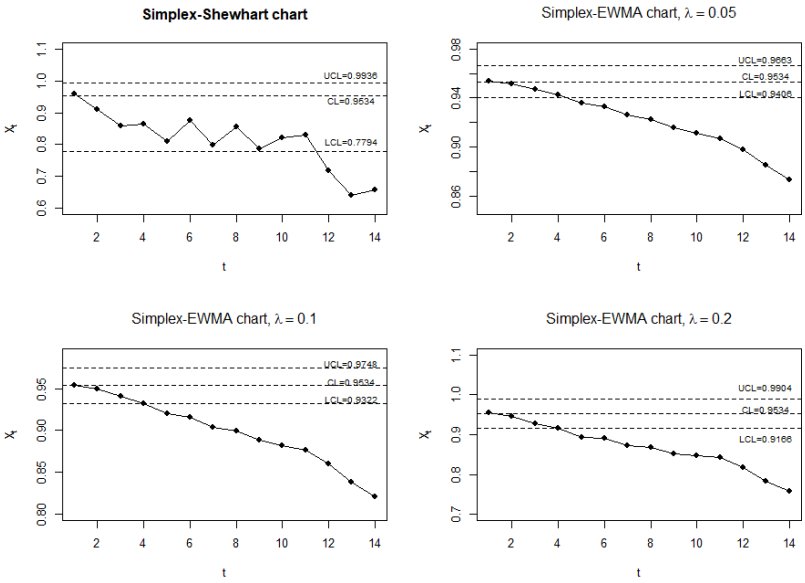


Έπειτα, προχωράμε στην κατασκευή διαγραμμάτων ελέγχου τύπου Shewhart και EWMA Φάσης II, θεωρώντας τις εκτιμήσεις των παραμέτρων μ, σ ως τις πραγματικές εντός ελέγχου τιμές των παραμέτρων. Από το Σχήμα 2 παρατηρούμε ότι το διάγραμμα Shewhart δίνει ένδειξη για εκτός ελέγχου διεργασία για πρώτη φορά στο σημείο 12. Στη συνέχεια, εφαρμόζοντας τα βήματα της αλγοριθμικής διαδικασίας για το στατιστικό σχεδιασμό του δίπλευρου διαγράμματος EWMA, προσδιορίσαμε τα όρια ελέγχου για $\lambda \in \{0.05,0.10,0.20\}$. Παρατηρούμε ότι το διάγραμμα σηματοδοτεί για πρώτη φορά για εκτός ελέγχου διεργασία στο δείγμα 5 (για $\lambda = 0.05$ και για $\lambda = 0.10$) ή στο δείγμα 4 (για $\lambda = 0.20$). Αυτό αποτελεί ένδειξη επιδείνωσης της διεργασίας, καθώς το ποσοστό των μη μολυσμένων φιστικιών μειώνεται. Τέλος, σε σύγκριση με το διάγραμμα Shewhart, το διάγραμμα EWMA ανιχνεύει τη μετατόπιση που συμβαίνει περίπου 7-8 σημεία νωρίτερα.

6. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα εργασία, μελετήσαμε διαγράμματα ελέγχου τύπου EWMA μεμονωμένων μετρήσεων για την παρακολούθηση διπλά οριοθετημένων διεργασιών στο διάστημα (0,1) χρησιμοποιώντας τρία διαφορετικά μοντέλα πιθανότητας. Με χρήση προσομοίωσης Monte Carlo υπολογίσαμε την απόδοση των διαγραμμάτων για τρεις τιμές της παραμέτρου λ και για κάθε ένα από τα τρία μοντέλα. Η απόδοσή τους συγκρίθηκε με την απόδοση των αντίστοιχων διαγραμμάτων ελέγχου τύπου Shewhart (Ho et al. (2019)), χωριστά για καθένα από τα τρία μοντέλα.

Σχήμα 2. *Διαγράμματα Φάσης ΙΙ για την αναλογία μη-μολυσμένων φιστικιών.*



Τα αποτελέσματα έδειξαν την υπεροχή των διαγραμμάτων ελέγχου EWMA στην ανίχνευση αυξήσεων και μειώσεων, ειδικά μικρής τάξης, στο μέσο επίπεδο της διεργασίας. Η βελτίωση στην ανίχνευση μικρών μετατοπίσεων κυμαίνεται από 70% έως 90%. Επιπλέον, διερευνήσαμε πόσο επηρεάζονται τα διαγράμματα ελέγχου EWMA για μεμονωμένες μετρήσεις στο (0, 1) όταν τα όρια ελέγχου υπολογίζονται υπό το «λάθος» μοντέλο. Η εκτεταμένη μελέτη προσομοίωσης έδειξε ότι σε όλες σχεδόν τις περιπτώσεις επηρεάζεται κυρίως η εντός ελέγχου απόδοση. Για τις περιπτώσεις μεγάλης διασποράς επηρεάζεται και η εκτός ελέγχου απόδοση, ειδικά για μικρές μετατοπίσεις στο μέσο επίπεδο της διεργασίας. Για την υλοποίηση των προσομοιώσεων χρησιμοποιήθηκε η R (R Core Team (2021)) και τα σχετικά προγράμματα για την αναπαραγωγή των αποτελεσμάτων είναι διαθέσιμα από τους συγγραφείς κατόπιν σχετικού αιτήματος.

ABSTRACT

In this work we propose and study two-sided EWMA type control charts for monitoring double bounded processes. Specifically, the term double bounded refers to observations in the interval (0, 1) and thus, these charts are suitable for monitoring rates, proportions and percentages. There are several models that can be used to describe this kind of data (and the respective processes, as well) such as the Beta distribution, the Simplex distribution and the Unit Gamma distribution. For each of these three models, we provide the statistical design and the performance of the proposed EWMA charts. Also, apart from providing the appropriate values for the design parameters of each chart, we investigate how much the performance of the EWMA schemes is affected by using the values of control limits which have not been calculated under the true model.

ΑΝΑΦΟΡΕΣ

Λαφατζή, Α. (2021). Διαγράμματα Ελέγχου για Ποσοστά και Αναλογίες. Μεταπτυχιακή Διατριβή. Τμήμα Στατιστικής Αναλογιστικών-Χρηματοοικονομικών Μαθηματικών. Πανεπιστήμιο Αιγαίου.

Alevizakos, V. and Koukouvinos, C. (2019). Monitoring of zero-inflated Poisson processes with EWMA and DEWMA control charts. *Quality and Reliability Engineering International*, **36**(1), 88–111.

Barndorff-Nielsen, O. E. and Jorgensen, B. (1991). Some parametric models on the Simplex. *Journal of multivariate analysis*, **39**(1), 106–116.

Crowder, S. V. (1987). Average run lengths of exponentially weighted moving average control charts. *Journal of Quality Technology*, **19**(3), 161–164.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, **31**(7), 799–815.

Grassia, A. (1977). On a family of distribution with argument between 0 and 1 obtained by transformation of the gamma and derived compound distributions. *Australian Journal of Statistics*, **19**(2), 108–114.

Gupta, A. and Nadarajah, S. (2004). *Handbook of Beta Distribution and its Applications*. CRC press.

Kieschnick, R. and McCullough, B. D. (2003). Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *Statistical Modelling*, **3**(3), 193–213.

Lee Ho, L., Fernandes, F. H. and Bourguignon, M. (2019). Control charts to monitor rates and proportions. *Quality and Reliability Engineering International*, **35**(1), 74–83.

Lima-Filho, L. M. A., Bourguignon, M., Ho, L. L. and Fernandes, F. H. (2020). Median control charts for monitoring asymmetric quality characteristics double bounded. *Quality and Reliability Engineering International*, **36**, 2285–2308.

Maravelakis, P. E., Panaretos, J. and Psarakis, S. (2005). An examination of the robustness to non normality of the EWMA control charts for the dispersion. Communications in Statistics-Simulation and Computation, **34**(4), 1069-1079.

Montgomery, D. C. (2013). Introduction to Statistical Quality Control, *6th edn. John Wiley & Sons, Inc.: New York, USA.*

Mousa, A. M., El-Sheikh, A. A,, Abdel-Fattah, M. A. (2016). A gamma regression for bounded continuous variables. *Advances and Applications in Statistics*, **49**, 305–326.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

Roberts, S. W. (1959). Control chart test based on geometric moving averages. *Technometrics*, **1**, 97–101.

Sant'Anna, A. M. O. and ten Caten, C. S. (2012). Beta control charts for monitoring fraction data. *Expert Systems with Applications*, **39**(11), 10236–10243.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 33ου Πανελληνίου Συνεδρίου Στατιστικής (2021), σελ.150-162

ΦΙΛΤΡΟ ΚΡΥΦΟΥ ΟΜΟΓΕΝΟΥΣ ΜΑΡΚΟΒΙΑΝΟΥ ΣΥΣΤΗΜΑΤΟΣ

Ρ. Λύκου, Γ. Τσακλίδης
Τμήμα Μαθηματικών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
lykourodi@math.auth.gr, tsaklidi@math.auth.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία μελετάται ένα κλειστό, διακριτού Χρόνου Κρυφό Ομογενές Μαρκοβιανό Σύστημα ως μοντέλο χώρου καταστάσεων όπου το κρυφό μοντέλο είναι το γνωστό στη βιβλιογραφία ως Ομογενές Μαρκοβιανό Σύστημα (ΟΜΣ-HMS). Αυτή η προσέγγιση θέτει τις βάσεις για την αξιοποίηση στατιστικών φίλτρων, προκειμένου να εκτιμηθεί η κατανομή των μελών στις καταστάσεις. Η περίπτωση συστημάτων μεγάλων πληθυσμών αντιμετωπίζεται με φίλτρο συνεχών μεταβλητών, το φίλτρο σωματιδίων. Η δομή του μοντέλου χώρου καταστάσεων εξαιτίας του υποκείμενου ΟΜΣ επιτρέπει την εκτίμηση των πιθανοτήτων μετάβασης με τη χρήση των σωματιδίων.

Λέξεις Κλειδιά: Κλειστό διακριτού χρόνου Ομογενές Μαρκοβιανό Σύστημα; Στατιστικά Φίλτρα

1. ΕΙΣΑΓΩΓΗ

Τα διακριτού χρόνου Ομογενή Μαρκοβιανά Συστήματα (ΟΜΣ) είναι πιθανοκρατικά μοντέλα που παρέχουν ένα γενικό πλαίσιο για να περιγραφεί η εξέλιξη πολλών τύπων πιθανοκρατικών μοντέλων, συμπεριλαμβανομένου και του κλασικού μοντέλου Μαρκοβιανής αλυσίδας. Αποτελούν μια ειδική περίπτωση των Μη Ομογενών Μαρκοβιανών συστημάτων που έχουν τις ρίζες τους στο άρθρο Bartholomew (1963) και ορίστηκαν πλήρως στη δημοσίευση Vassiliou (1982). Μια ανασκόπηση αυτών απαντάται στο άρθρο Vassiliou (1997). Οι υπάρχουσες μελέτες στα διακριτού χρόνου ΟΜΣ αφορούν μεταξύ άλλων την εξέλιξη των προσιτών δομών (Tsaklidis (1994)), την εξέλιξη της κατανομής των διανυσμάτων κατάστασης (Vasiliadis and Tsaklidis (2008)), την τάξη μεγέθους τους (Kipouridis, I. and Tsaklidis, G. (2001)), καθώς και την εξέλιξη των ροπών (Vasiliadis and Tsaklidis (2007)) και των κατανομών (Vasiliadis and Tsaklidis (2011)) των μεγεθών των καταστάσεων σε διακριτού χρόνου ΟΜΣ με πεπερασμένη χωρητικότητα στις καταστάσεις. Ακόμη, η πρόσφατη βιβλιογραφία περιλαμβάνει ευρύτερες προσεγγίσεις των μη ομογενών Μαρκοβιανών Συστημάτων που θίγουν νόμους μεγάλων αριθμών και θεωρήματα για το ρυθμό σύγκλισης και την περιοδικότητα της αναμενόμενης πληθυσμιακής δομής σε ένα

γενικό χώρο καταστάσεων (Vassiliou 2020α και b) όπως, επίσης, και αντίστοιχα συστήματα συνόλων (Vassiliou 2021), αλλά και ημι-Μαρκοβιανών Συστημάτων (Vassiliou and Papadopoulou 2014) που περιλαμβάνουν τα ΟΜΣ ως ειδική περίπτωση. Οι εφαρμογές τους αξιοποιούνται κατά κύριο λόγο στον προγραμματισμό ανθρώπινου δυναμικού (για μια εισαγωγή στις στατιστικές τεχνικές βλ. Bartholomew et al. (1991) και μια ανασκόπηση αυτών στην εργασία των Feyter and Guerry (2011)), αγγίζοντας, μεταξύ άλλων, ειδικά ζητήματα όπως ο εκπαιδευτικός σχεδιασμός (Osagiede and Ekhosuehi (2006)) και η θεωρία ουρών (Vasiliadis (2016)).

Στην περίπτωση που είναι διαθέσιμες παρατηρήσεις, αυτές αφορούν συχνά μόνο το μέγεθος των καταστάσεων του μοντέλου παρά ξεχωριστές παρατηρήσεις για την κάθε μετάβαση. Τέτοιες εφαρμογές μοντέλων μπορεί να αφορούν ανάλυση μετανάστευσης πτηνών (Sheldon et al. (2007)), προσαρμογή μοντέλων εκλογικών αποτελεσμάτων (Flaxman et al. (2015)), μοντελοποίηση πιστωτικού κινδύνου (Jones (2005)) και χωρικά στατιστικά κινητών (Terada and Nagata (2013)). Στο παρελθόν, τα σφάλματα παρατήρησης και η αβεβαιότητα έχουν παραβλεφθεί και έχει υιοθετηθεί η υπόθεση ύπαρξης ακριβών δεδομένων (χωρίς θεώρηση σφαλμάτων) εξαιτίας των αυξημένων απαιτήσεων των Κρυφών Μαρκοβιανών Μοντέλων (ΚΜΜ) (Jones (2005), p. 5). Όμως τα τελευταία χρόνια, έχει αυξηθεί το ενδιαφέρον αναφορικά με την εκμάθηση ΚΜΜ παρουσία αθροιστικών παρατηρήσεων με θόρυβο. Οι Bernstein and Sheldon (2016) χρησιμοποίησαν εκτιμητές ροπών στα πλαίσια των συλλογικών γραφικών μοντέλων (Sheldon and Dietterich (2011)) για το σκοπό αυτό. Οι Iwata and Shimizu (2019) επιστράτευσαν ένα νευρωνικό δίκτυο για να ελαττώσουν τον αριθμό των απαιτούμενων παραμέτρων στα ίδια πλαίσια. Οι Lyubchyk et al. (2019) πρότειναν μια αναδρομική μέθοδο εκτίμησης. Επίσης οι Ma et al. (2020) στράφηκαν στη μάθηση στοχαστικής συμπεριφοράς. Ο Zheng (2019) ασχολήθηκε με τις συνολικές παρατηρήσεις (ensemble observations), ενώ οι Singh et al (2022) συνέθεσαν στο πεδίο της εκτιμητικής έναν νέο αλγόριθμο Μεγιστοποίησης-Προσδοκίας (Dempster et al. (1977)).

Η παρούσα μελέτη εστιάζει στην εκτίμηση των μεγεθών των καταστάσεων ενός Κρυφού διακριτού χρόνου ΟΜΣ που είναι κλειστό, δηλαδή δεν υπάρχουν μέλη που αποχωρούν ή που εισέρχονται στο σύστημα, με παρουσία παρατηρήσεων με θόρυβο για τα μεγέθη των καταστάσεων. Αυτό επιτυγχάνεται με την αξιοποίηση στατιστικών φίλτρων συνεχών μεταβλητών, κυρίως με το Φίλτρο Σωματιδίων (ΦΣ) για τη γενική περίπτωση. Στην ενότητα 2, παρατίθενται στοιχειώδεις ορισμοί και θεμελιώνεται η ιδέα του Κρυφού ΟΜΣ διακριτού χρόνου. Στην ενότητα 3, κατασκευάζεται κατάλληλη αναπαράσταση του συστήματος με μοντέλο χώρου-καταστάσεων (state space model) προκειμένου να μπορεί να λειτουργήσει ένα φίλτρο συνεχών μεταβλητών, και συντίθεται κατάλληλο ΦΣ για το υπό μελέτη μοντέλο, καθώς, επίσης, παρατίθενται σχόλια για τις δυνατότητες εκτίμησης παραμέτρων.

2. ΤΟ ΚΛΕΙΣΤΟ ΚΡΥΦΟ ΟΜΟΓΕΝΕΣ ΜΑΡΚΟΒΙΑΝΟ ΣΥΣΤΗΜΑ ΔΙΑΚΡΙΤΟΥ ΧΡΟΝΟΥ

Δεδομένου του χρονικού δείκτη *t=0,1,2,…* ένα κλειστό, διακριτού χρόνου ΟΜΣ αποτελείται από έναν πληθυσμό $N\in\mathbb{N}$ μελών τα οποία κατανέμονται σε ένα σύνολο καταστάσεων $S=\{1,2,\ldots,k\}$ σε κάθε χρονικό βήμα. Υποτίθεται ότι σε κάθε χρονικό σημείο *t* ένα μέλος του πληθυσμού ανήκει σε μία και μόνο μια από τις καταστάσεις *1,2,…,k* . Υπάρχουν

$$\mathscr{E}_k^N = \binom{k+N-1}{N} = \frac{(k+N-1)!}{N!(k-1)!}$$

τρόποι για να κατανεμηθούν τα *N* μέλη στις *k* καταστάσεις. Θεωρείται ότι το σύστημα δεν έχει μνήμη, με την έννοια ότι η μετάβαση ενός μέλους από μια κατάσταση στην άλλη εξαρτάται στοχαστικά μόνο από την τρέχουσα κατάσταση του μέλους, ανεξάρτητα από τις μεταβάσεις σε προηγούμενα χρονικά βήματα (έλλειψη μνήμης υπό τη δέσμευση του γεγονότος της τελευταίας χρονικής στιγμής). Η πιθανότητα για ένα μέλος να μεταβεί σε ένα βήμα από μια κατάσταση $i\in S$ σε μια κατάσταση $j\in S$ συμβολίζεται με p_{ij}. Το σύστημα θεωρείται ομογενές, δηλαδή θεωρείται ότι η πιθανότητα μετάβασης p_{ij} είναι ανεξάρτητη του χρόνου για κάθε *i, j*. Ο αντίστοιχος *kxk* πίνακας πιθανοτήτων μετάβασης συμβολίζεται με $\mathbf{P}=(p_{ij})$.

Η δομή του πληθυσμού του ΟΜΣ διακριτού χρόνου, τη χρονική στιγμή *t* εκφράζεται από το τυχαίο διάνυσμα κατάστασης

$$\mathbf{n}(t) = (n_1(t), n_2(t), \ldots, n_k(t))^T$$

όπου $n_i(t)$, $i\epsilon S$, είναι ο αριθμός των μελών του πληθυσμού που βρίσκονται στη θέση *i* τη χρονική στιγμή *t* και ο εκθέτης *T* συμβολίζει το ανάστροφο ενός διανύσματος ή ενός πίνακα. Αντιστοίχως, $n_{ij}(t)$ είναι ο αριθμός των μελών που μεταβαίνουν από την κατάσταση $i\epsilon S$ στην κατάσταση $j\epsilon S$ κατά το χρονικό διάστημα *(t-1,t]*, όπου *t=1,2,…* Υποτίθεται ότι τα μέλη του πληθυσμού δεν μετακινούνται κατά το αρχικό χρονικό σημείο t_0. Έστω,

$$\mathbf{n}_{i*}(t) = (n_{i1}(t), n_{i2}(t), \ldots, n_{ik}(t))^T$$

το διάνυσμα του πλήθους των μελών που μετακινούνται από την κατάσταση *i* στις καταστάσεις *1,2,…,k* του συστήματος κατά τη διάρκεια του χρονικού διαστήματος *(t-1,t]*. Τότε προφανώς (Bartholomew (1982), p.28),

$$\mathbf{n}_{i*}(t)\,\big|\,n_i(t-1) \sim Multinomial\,(n_i(t-1), \mathbf{p}_i) \qquad (1)$$

όπου με $\mathbf{p}_i^T$ συμβολίζεται το *i*-οστό διάνυσμα-γραμμή του πίνακα **P** και το σύμβολο $\sim$ δηλώνει ότι ένα τυχαίο διάνυσμα/τυχαία μεταβλητή ακολουθεί μια συγκεκριμένη κατανομή.

Παρατήρηση 1

Καθώς ισχύει

$$\mathbf{n}(t) = \sum_{i\in S} \mathbf{n}_{i*}(t-1) \qquad (2)$$

η συνάρτηση κατανομής της τ.μ. $\mathbf{n}(t)\big|\mathbf{n}(t-1)$ είναι η συνέλιξη των κατανομών όλων των $\mathbf{n}_{i*}(t-1)$ δεδομένου του $\mathbf{n}(t-1)$. Σύμφωνα με τη σχέση (1), αυτή η συνέλιξη είναι μια ειδική περίπτωση της γενικευμένης πολυωνυμικής κατανομής (Beaulieu (1991)) ή, ισοδύναμα, μιας Poisson δυωνυμικής κατανομής (που έχει τις ρίζες της στο σύγγραμμα Poisson (1837), §14) ή ενός πολυωνυμικού ανάλογου της J-Δυωνυμικής κατανομής (Benneyan (2004)). Εάν ο πίνακας **P** αποτελείται από ίσες γραμμές, τότε **n**(*t*)~ *Multinomial*(*N*,**p**), όπου **p** είναι το κοινό διάνυσμα γραμμή του πίνακα **P**, ανεξάρτητα από το **n**(*t-1*).

Τόσο η θεμελίωση του ΟΜΣ διακριτού χρόνου, όσο και η Παρατήρηση 1 καταδεικνύουν ότι η στοχαστική διαδικασία της πληθυσμιακής δομής $\mathbf{n}(t)\in\mathbb{N}^k$ ορίζει μια νέα Μαρκοβιανή διαδικασία που εξελίσσεται σύμφωνα με έναν σταθερό $\mathcal{E}_k^N$ x $\mathcal{E}_k^N$ πίνακα πιθανοτήτων μετάβασης $\mathbf{A}_{\text{ps}}(\mathbf{P},N,k)=(\alpha_{ij})$. Οι δείκτες γραμμής του πίνακα $\mathbf{A}_{\text{ps}}$ αντιστοιχούν στις δυνατές πραγματοποιήσεις του **n**(*t-1*), ενώ οι δείκτες στήλης του στις δυνατές πραγματοποιήσεις του **n**(*t*). Στις γραμμές του πίνακα $\mathbf{A}_{\text{ps}}$ ανατίθενται οι τιμές της κατανομής του $\mathbf{n}(t)\big|\mathbf{n}(t-1)$ όπως σχολιάστηκε στην Παρατήρηση 1.

Στην παρούσα μελέτη, η πληθυσμιακή δομή **n**(*t*) θεωρείται κρυφή. Διαθέσιμες είναι διακριτές παρατηρήσεις της με θόρυβο $\mathbf{y}(t)\in O=\{\mathbf{o}_1,\mathbf{o}_2,...,\mathbf{o}_m\}$, όπου $\mathbf{o}_i\in\mathbb{N}^d$, $d\in\mathbb{N}$, $i\in\{1,2,...,m\}$, $m\in\mathbb{N}$, συμβολίζουν τις δυνατές παρατηρούμενες δομές. Ως προς τη διάσταση *d* της παρατήρησης $\mathbf{y}(t)$ είναι δυνατό να ισχύει ότι $d<k$ όταν δεν υπάρχουν διαθέσιμες πληροφορίες για τον πληθυσμό σε μια ή περισσότερες καταστάσεις. Η κατανομή της παρατήρησης **y**(t) εξαρτάται από το διάνυσμα κατάστασης **n**(*t*) σε κάθε χρονικό σημείο *t*. Έστω $D=\{\mathbf{d}_1,\mathbf{d}_2,...,\mathbf{d}_{\mathcal{E}_k^N}\}\subset\mathbb{N}^k$ το σύνολο όλων των δυνατών πραγματοποιήσεων του $\mathbf{n}(t)$. Τότε, μπορεί να καθοριστεί ένα Κρυφό Μαρκοβιανό Μοντέλο (ΚΜΜ), σχετιζόμενο με το προαναφερθέν ΟΜΣ, με την ακόλουθη τριάδα παραμέτρων

$$\lambda=(\mathbf{A}_{\text{ps}},\mathbf{B},\mathbf{p}_0),$$

όπου $\mathbf{B}=(b_{i,j})=\left(p(\mathbf{y}(t)=\mathbf{o}_j\,|\mathbf{n}(t)=\mathbf{d}_j)\right)$, $\mathbf{o}_j\in O,\mathbf{d}_j\in D$, είναι ο πίνακας πιθανοτήτων παρατήρησης και $\mathbf{p}_0=(p_i)=\left(p(\mathbf{n}(0)=\mathbf{d}_i)\right)$, $\mathbf{d}_i\in D$ είναι η κατανομή του διανύσματος **n**(*0*). Η από κοινού πιθανότητα του κρυφού διανύσματος κατάστασης και του διανύσματος παρατήρησης για μια ακολουθία ενός ΚΜΜ είναι η

$$p(\mathbf{n}(0),,...,\mathbf{n}(t),\mathbf{y}(1),...,\mathbf{y}(t))=p(\mathbf{n}(0))\prod_{k=1}^{t}p(\mathbf{n}(k)\big|\mathbf{n}(k-1))\,p(\mathbf{y}(k)\big|\mathbf{n}(k))$$

Εναλλακτικά, μπορεί να οριστεί ένα κρυφό Μαρκοβιανό σύστημα με *k* καταστάσεις και μέγεθος *N*, ως η τετράδα

$$\mu=(N,\mathbf{P},\mathbf{B},\mathbf{p}_0),$$

που αντιπροσωπεύει το προαναφερθέν ΚΜΜ. Αφού διαπιστώθηκε ότι το κρυφό Μαρκοβιανό σύστημα μπορεί να εκφραστεί ως ένα ισοδύναμο ΚΜΜ, θεωρήματα και τεχνικές των ΚΜΜ μπορούν να επιστρατευθούν γι' αυτή την ειδική περίπτωση. Παραδείγματος χάριν, ο αλγόριθμος Viterbi (Viterbi (1967), Forney (1973)) θα μπορεί να αξιοποιηθεί για την εκτίμηση του κρυφού διανύσματος.

2.1 Υπολογιστική πολυπλοκότητα του κρυφού Μαρκοβιανού συστήματος

Το πλήθος των δυνατών πληθυσμιακών δομών αυξάνεται με τον αριθμό των μελών και των καταστάσεων, έτσι ώστε τα κλειστά Κρυφά ΟΜΣ διακριτού χρόνου με N μέλη και k καταστάσεις αντιστοιχούν σε κρυφές Μαρκοβιανές αλυσίδες $\mathbf{n}(t)$ με $\mathcal{E}_k^N$ δυνατές πραγματοποιήσεις. Αυτό σημαίνει ότι η υπολογιστική πολυπλοκότητα της απλής προσέγγισης είναι $O((\mathcal{E}_k^N)^T T)$, ενώ πιο εκλεπτυσμένες τεχνικές, όπως η μέθοδος forward-backward (Stratonovich (1965)) και ο αλγόριθμος Viterbi, έχουν πολυπλοκότητα $O((\mathcal{E}_k^N)^2 T)$. Παράλληλα, η Παρατήρηση 1 καταδεικνύει ότι η εύρεση των τιμών της κατανομής είναι διαδικασία υπολογιστικά απαιτητική. Καθώς το μέγεθος της δεσμευόμενης μνήμης και το υπολογιστικό κόστος εξαρτώνται από το πλήθος των καταστάσεων k και το μέγεθος του πληθυσμού N, μια τέτοια πρακτική καθίσταται αδύνατο να εφαρμοστεί σε πολυπληθή συστήματα. Στην επόμενη ενότητα αξιοποιούνται τεχνικές για συνεχείς μεταβλητές για την αντιμετώπιση του προβλήματος αυτού.

3. ΣΤΑΤΙΣΤΙΚΑ ΦΙΛΤΡΑ ΣΥΝΕΧΩΝ ΜΕΤΑΒΛΗΤΩΝ

Το συνεχές ανάλογο των ΚΜΜ εμπλέκεται στη διαδικασία εκτίμησης του $\mathbf{n}(t)$. Για το σκοπό αυτό, κατασκευάζεται κατάλληλη αναπαράσταση μοντέλου χώρου-καταστάσεων στην οποία τα κρυφά διανύσματα $\mathbf{n}(t)$ και οι παρατηρήσεις $\mathbf{y}(t)$ αντιμετωπίζονται ως συνεχή τυχαία διανύσματα. Κατόπιν, ο αλγόριθμος του Φίλτρου Σωματιδίων (Gordon et al. (1993)) προσαρμόζεται στο δομημένο μοντέλο.

Ένα κλειστό Κρυφό ΟΜΣ διακριτού χρόνου μπορεί να εκφραστεί από τις εξισώσεις

$$\mathbf{n}(t) = f(\mathbf{n}(t-1))$$
$$\mathbf{y}(t) = g(\mathbf{n}(t)),$$

όπου f και g είναι συναρτήσεις σχετιζόμενες με τις πιθανότητες $\mathrm{p}(\mathbf{n}(t)|\mathbf{n}(t-1))$ και $\mathrm{p}(\mathbf{y}(t)|\mathbf{n}(t))$ αντίστοιχα. Αν και είθισται τα ορίσματα των συναρτήσεων f και g να περιέχουν μεταβλητές θορύβου και οι εξισώσεις να είναι αλγεβρικές, εν προκειμένω ο θόρυβος θα εκφραστεί με ειδικό τελεστή, και η τυχαιότητα θα εκφράζεται από την ίδια τη συνάρτηση, ενώ το αντίστοιχο σύμβολο της μεταβλητής θορύβου θα παραλείπεται. Για την αντιπροσώπευση του προαναφερθέντος ΚΜΜ, η εξίσωση κατάστασης μπορεί να πάρει τη μορφή

$$\mathbf{n}(t) = \mathbf{P}^T * \mathbf{n}(t-1) = \sum_{i-1}^{k} \mathbf{p}_i * n_i(t-1), \qquad (3)$$

όπου το σύμβολο $*$ δηλώνει τον τελεστή πολυωνυμικής εκλέπτυνσης (multinomial thinning operator), όπως ορίστηκε από τον McKenzie (McKenzie (2003)), δηλαδή το $\mathbf{p}_i * n_i(t-1)$ είναι ένα τυχαίο διάνυσμα που κατανέμεται σύμφωνα με την κατανομή $Multinomial\,(n_i(t-1), \mathbf{p}_i)$. Αυτή η αναπαράσταση είναι εναλλακτική των σχέσεων (1) και (2).

3.1 Το Φίλτρο Σωματιδίων

Το Φίλτρο Σωματιδίων (ΦΣ) είναι κατάλληλο για μη-γραμμικά μοντέλα χώρου-καταστάσεων με μη Γκαουσιανούς θορύβους. Είναι ένα Μπεϋζιανό φίλτρο, δηλαδή στηρίζεται στις εξισώσεις πρόβλεψης-διόρθωσης

$$\begin{aligned} p(\mathbf{n}(t)|\mathbf{y}_{1:t-1}) &= \int p(\mathbf{n}(t)|\mathbf{n}(t-1))p(\mathbf{n}(t-1)|\mathbf{y}_{1:t-1})d^k\mathbf{n}(t-1) \\ p(\mathbf{n}(t)|\mathbf{y}_{1:t}) &= \frac{p(\mathbf{y}(t)|\mathbf{n}(t))p(\mathbf{n}(t)|\mathbf{y}_{1:t-1})}{\int p(\mathbf{y}(t)|\mathbf{n}'(t))p(\mathbf{n}'(t)|\mathbf{y}_{1:t-1})d^k\mathbf{n}'(t)} \end{aligned}$$

για την αποτίμηση της εκ των υστέρων συνάρτησης πιθανότητας $p(\mathbf{n}(t)\big|\mathbf{y}_{1;t})$, όπου $\mathbf{y}_{1:t} = \{\mathbf{y}(1), \ldots, \mathbf{y}(t)\}$. Βασίζεται στην ολοκλήρωση Monte Carlo και αναπαριστά την εκ των υστέρων κατανομή του κρυφού διανύσματος $p(\mathbf{n}(t)\big|\mathbf{y}_{1;t})$ με $M \in \mathbb{N}$ σταθμισμένα σωματίδια $\mathbf{n}^{(\xi)}(t)$, βαρών $w_t^{(\xi)}$, $\xi=1,2,\ldots,M$, έτσι ώστε

$$\hat{p}(\mathbf{n}(t)|\mathbf{y}_{1:t}) = \sum_{\xi=1}^{M} w_t^{(i)} \delta(\mathbf{n}(t) - \mathbf{n}^{(\xi)}(t))$$

όπου δ είναι μια πυκνότητα Radon-Nikodým ώστε για διακριτές μεταβλητές, παίρνει την τιμή 1 στο 0 και την τιμή 0 παντού αλλού, ενώ για συνεχείς μεταβλητές είναι η συνάρτηση δέλτα του Dirac, και τα βάρη κανονικοποιούνται προκειμένου να ισχύει

$$\sum_{\xi=1}^{M} w_t^{(\xi)} = 1.$$

Στα πλαίσια του ΦΣ, η πιθανότητα $p(\mathbf{n}(t)\big|\mathbf{n}(t-1))$ μπορεί να προσομοιωθεί αντί για να υπολογιστεί επακριβώς κατά την Παρατήρηση 1. Προτείνεται η κατανομή του τυχαίου διανύσματος $\mathbf{n}_{i*}(t)$ να προσομοιωθεί με σωματίδια $\mathbf{n}_{i*}^{(\xi)}(t)$, $i \in S$, $\xi=1,2,\ldots,M$, για τα οποία ισχύει

$$\mathbf{n}_{i*}^{(\xi)}(t)\big|\mathbf{n}_{0;t-1}^{(\xi)} \sim Multinomial\,(n_i^{(\xi)}(t-1), \mathbf{p}_i),$$

όπως προκύπτει από τη σχέση (1), όπου $n_i^{(\xi)}(t)$ είναι ο αριθμός των μελών που βρίσκονται στην κατάσταση i τη χρονική στιγμή t για την ξ-οστή περίπτωση. Τότε, η εμπειρική κατανομή του

$$\mathbf{n}^{(\xi)}(t) = \sum_{i=1}^{k} \mathbf{n}_{i\bullet}^{(\xi)}(t), \xi = 1, \ldots, M$$

προσομοιώνει την κατανομή του $\mathbf{n}(t)\big|\mathbf{n}(t-1)$. Τα βάρη των σωματιδίων υπολογίζονται αναδρομικά ως

$$w_t^{(\xi)} \propto w_{t-1}^{(\xi)} p(\mathbf{y}_t|\mathbf{n}^{(\xi)}(t))$$

όπου το σύμβολο $\propto$ δηλώνει αναλογία κατά σταθερά.

3.1.1 Αναδειγματοληψία

Διαπιστώνεται, καθώς ο χρόνος περνά, ότι στα σωματίδια που βρίσκονται εγγύτερα στο κρυφό διάνυσμα ανατίθενται όλο και μεγαλύτερα βάρη, ενώ τα βάρη των υπολοίπων σωματιδίων γίνονται αμελητέα. Ως εκ τούτου, μόνο ένα πολύ μικρό τμήμα των σωματιδίων παραμένει χρήσιμο για την εκτίμηση του κρυφού διανύσματος. Αυτό το φαινόμενο ονομάζεται εκφυλισμός. Προκειμένου να αποφευχθεί ο εκφυλισμός, ενσωματώνεται στον αλγόριθμο ένα βήμα αναδειγματοληψίας οδηγώντας σε έναν αλγόριθμο Δειγματοληψίας κατά Σημαντικότητα- Αναδειγματοληψίας (Sampling Importance Resampling-SIR). Σε αυτό το βήμα, εκτελείται αναδειγματοληψία με επανατοποθέτηση σύμφωνα με τα βάρη. Ωστόσο, η διαδοχική αναδειγματοληψία αναγκάζει προοδευτικά τις λιγότερο πιθανές τιμές να αντικατασταθούν από επαναλαμβανόμενες πιθανές. Αυτό το πρόβλημα ονομάζεται φτωχοποίηση. Σε αναζήτηση μιας μέσης λύσης ανάμεσα στα δυο προβλήματα, έχει προταθεί ένα κριτήριο σχετικό με τη διασπορά των βαρών για τη λήψη απόφασης σε κάθε χρονικό βήμα αν πρέπει να τελεστεί αναδειγματοληψία των διαθέσιμων σωματιδίων ή όχι. Ο Liu (Liu (2004), p.35-36) παραθέτει το ακόλουθο μέγεθος για να εκφράσει τον βαθμό εκφυλισμού των σωματιδίων,

$$N_{eff}(t) = \frac{M}{1 + \mathrm{Var}_{p(\bullet|y_{1:t})}(w(\mathbf{n}(t)))}$$

το οποίο καλείται αποτελεσματικό μέγεθος δείγματος (effective sample size). Αυτή η ποσότητα εκτιμάται ως

$$\hat{N}_{eff}(t) = 1/\sum_{\xi=1}^{M}\left(w_t^{(\xi)}\right)^2$$

διότι δεν μπορεί να υπολογιστεί επακριβώς. Όταν $\hat{N}_{eff}(t) \le N_T$, όπου N_T=cN , c$\epsilon\mathbb{R}$, είναι ένα καθορισμένο κατώφλι, το βήμα της αναδειγματοληψίας υλοποιείται. Τα βήματα του ΦΣ παρουσιάζονται στον Αλγόριθμο 1. Το κομμάτι της δειγματοληψίας αντιστοιχεί στη φάση πρόβλεψης της Μπεϋζιανής συμπερασματολογίας, ενώ η ανάθεση βαρών και η αναδειγματοληψία συνιστούν τη φάση διόρθωσης. Με την ίδια λογική, υλοποιούνται και οι τεχνικές λείανσης (smoothing). Ωστόσο, δεν παρουσιάζονται λεπτομερώς στην παρούσα εργασία, καθώς η εφαρμογή στο προαναφερθέν μοντέλο είναι απλό ανάλογο της περίπτωσης του φίλτρου.

3.1.2 Εκτίμηση παραμέτρων και υπολογιστική πολυπλοκότητα

Η μέθοδος που περιγράφηκε παρέχει μια καινούρια προοπτική εκτίμησης παραμέτρων. Δεδομένου ότι n_{ij} είναι ο συνολικός αριθμός των μελών του μεταβαίνουν από την κατάσταση i στην κατάσταση j σε ένα χρονικό βήμα κατά τη διάρκεια μιας περιόδου παρατήρησης/εκμάθησης. Οι Anderson και Goodman (Anderson and Goodman (1957)) έδειξαν ότι η ποσότητα

$$\hat{p}_{1,ij} = n_{ij}/\sum_{j=1}^{k} n_{ij}$$

είναι εκτιμητής μέγιστης πιθανοφάνειας της πιθανότητας p_{ij}, ο οποίος είναι συνεπής και μεροληπτικός, με τη μεροληψία να τείνει στο μηδέν καθώς το μέγεθος του δείγματος αυξάνεται. Καθώς το ΦΣ προσομοιώνει τις κρυφές μεταβάσεις $\mathbf{n}_{i*}(t)$ με

Αλγόριθμος 1. *Βασικός αλγόριθμος του φίλτρου σωματιδίων με αναδειγματοληψία (SIR) για το κλειστό Κρυφό ΟΜΣ διακριτού χρόνου*

Require: M, N_T, T
  Initialize $\{\mathbf{n}^{(\xi)}(0), w_0^{(\xi)}\}$
  for $t = 1,2,...,T$ **do**
    1.Importance Sampling
    Sample
    $\tilde{\mathbf{n}}_{i\bullet}^{(\xi)}(t) \sim \text{Mult}(n_i^{(\xi)}(t-1), \mathbf{p}_i)$
    Set $\tilde{\mathbf{n}}^{(\xi)}(t) = \sum_{i=1}^{k} \tilde{\mathbf{n}}_{i\bullet}^{(\xi)}(t)$
    Set $\tilde{\mathbf{n}}_{0:t}^{(\xi)} = (\mathbf{n}_{0:t-1}^{(\xi)}, \tilde{\mathbf{n}}^{(\xi)}(t))$,
    Calculate importance weights
    $\tilde{w}_t^{(\xi)} = w_{t-1}^{(\xi)} p(\mathbf{y}_t | \tilde{\mathbf{n}}^{(\xi)}(t))$
  end for
  for $\xi = 1,2,...,M$ **do**
    Nomalize weights $w_t^{(\xi)} = \frac{\tilde{w}_t^{(\xi)}}{\sum_{i=1}^{N} \tilde{w}_t^{(\xi)}}$
    2. Resampling
    if $\hat{N}_{eff}(t) \geq N_T$ **then**
      for $\xi = 1,2,...,M$ **do**
        $\mathbf{n}_{0:t}^{(\xi)} = \tilde{\mathbf{n}}_{0:t}^{(\xi)}$
      end for
    else
      for $\xi = 1,2,...,M$ **do**
        Sample with replacement index $j(i)$ according to the discrete weight distribution $p(j(\xi) = d) = w_t^{(d)}, d = 1,...,M$
        Set $\mathbf{n}_{0:t}^{(\xi)} = \tilde{\mathbf{n}}_{0:t}^{(j(\xi))}$ and $w_t^{(\xi)} = \frac{1}{M}$
      end for
    end if
  end for

τα σωματίδια $\mathbf{n}_{i*}^{(\xi)}(t)$, τα $\hat{p}_{1,ij}$ μπορούν να προσεγγιστούν με τη βοήθεια των σωματιδίων. Ως προς το παρόν μοντέλο των κρυφών διανυσμάτων και το ΦΣ, υπό την προϋπόθεση ότι $n_{ij}^{(\xi)}(t)$ άτομα μεταβαίνουν από την κατάσταση i στην κατάσταση j κατά την ξ-οστή περίπτωση, προτείνεται οι πιθανότητες μετάβασης p_{ij} να εκτιμώνται ως

$$\hat{p}_{2,ij} = \sum_{\xi=1}^{M} w_t^{(\xi)} \frac{n_{ij}^{(\xi)}}{\sum_{j=1}^{k} n_{ij}^{(\xi)}}.$$

Καθώς το **n**(t) είναι άγνωστο, ο τρόπος με τον οποίο ορίζεται το $\hat{p}_{2,ij}$ είναι ισοδύναμος με το γεγονός ότι

$$\hat{p}_{2,ij} \xrightarrow[M\to+\infty]{P} E[\hat{p}_{1,ij}] \xrightarrow[N\to+\infty]{P} p_{ij}$$

σύμφωνα με το νόμο των μεγάλων αριθμών- που οδηγεί στη σχεδόν βέβαιη σύγκλιση και κατ' επέκταση στο πρώτο όριο, στα πλαίσια της δειγματοληψίας χωρίς την αναδειγματοληψία, με την προϋπόθεση πεπερασμένων αναμενόμενων βαρών (Chen (2003))- και τους Anderson και Goodman (Anderson and Goodman (1957)) για τα δυο όρια αντιστοίχως.

Παρατήρηση 2
Έπειτα από το βήμα της αναδειγματοληψίας οι τροχιές των σωματιδίων δεν είναι πια μεταξύ τους ανεξάρτητες, γεγονός που διαταράσσει το συμπέρασμα της σχεδόν βέβαιης σύγκλισης. Η αντίστοιχη υπάρχουσα απόδειξη της σχεδόν βέβαιης σύγκλισης της δειγματικής κατανομής κατόπιν αναδειγματοληψίας στη θεωρητική προϋποθέτει μεταξύ άλλων την ύπαρξη συνεχούς και φραγμένης συνάρτησης μετάβασης στην εξίσωση των καταστάσεων (Crisan and Doucet (2002)), μια συνθήκη που αναιρείται στη σχέση (3), στην οποία διαφαίνεται συνάρτηση φραγμένη αλλά όχι συνεχής. Στο άρθρο Crisan and Doucet (2002) φαίνεται, ωστόσο, ότι ο προταθείς εκτιμητής είναι αμερόληπτος και με τις υποθέσεις φραγμένης πιθανοφάνειας και φραγμένης προσεγγιζόμενης συνάρτησης επιτυγχάνεται σύγκλιση κατά μέσο τετράγωνο, άρα και κατά πιθανότητα.

Έτσι, οι εκτιμητές μέγιστης πιθανοφάνειας μπορούν να βρεθούν αναλυτικά (χωρίς τη χρήση επιπρόσθετων αριθμητικών μεθόδων). Το γεγονός αυτό είναι σημαντικό για την υπολογιστική πολυπλοκότητα του αλγορίθμου και επιτρέπει την εύκολη εκτίμηση παραμέτρων online- παράλληλα με τη συλλογή των δεδομένων). Μια τέτοια προσέγγιση μπορεί να αξιοποιηθεί σε ένα πλαίσιο Μεγιστοποιήσης-Προσδοκίας (Expectation-Maximization). Η υπολογιστική πολυπλοκότητα της προτεινόμενης μεθόδου είναι τουλάχιστον $O(Mk^2T)$ και κυμαίνεται ανάλογα με τις μεθόδους δειγματοληψίας που επιλέγονται κατά περίπτωση (Doucet & Johansen 2009, p. 36). Εξαρτάται, έτσι, από το πλήθος των καταστάσεων k του συστήματος, τον αριθμό M των σωματιδίων που χρησιμοποιούνται για την προσομοίωση, αλλά, σύμφωνα με τον Αλγόριθμο 1, είναι ανεξάρτητη του πλήθους των μελών του συστήματος.

4. **ΕΠΙΛΟΓΟΣ**

Στην παρούσα εργασία ορίζεται το κλειστό Κρυφό ΟΜΣ διακριτού χρόνου σε σύνδεση με ένα αντίστοιχο ΚΜΜ. Η διαδικασία εκτίμησης των κρυφών μεταβλητών με το αντίστοιχο ΚΜΜ απαιτεί υπολογιστική πολυπλοκότητα που αυξάνεται συναρτήσει του πληθυσμιακού μεγέθους N. Για αυτό το λόγο, κατασκευάζεται αναπαράσταση χώρου-καταστάσεων για το υπό μελέτη μοντέλο, και έτσι αποκτάται η δυνατότητα χρήσης φίλτρων συνεχών μεταβλητών. Κατασκευάζεται κατάλληλος αλγόριθμος για την αξιοποίηση του Φίλτρου Σωματιδίων με υπολογιστική πολυπλοκότητα ανεξάρτητη του μεγέθους του πληθυσμού, ώστε η προτεινόμενη μέθοδος καθίσταται κατάλληλη για τη μοντελοποίηση πολυμελών συστημάτων. Έτσι, αντιμετωπίζεται το πρόβλημα αύξησης της πολυπλοκότητας κατά το συνυπολογισμό των σφαλμάτων παρατήρησης στη μοντελοποίηση αυτή.

Αναφορικά με τις δυνατότητες μελλοντικής έρευνας, διακρίνονται προοπτικές ανάπτυξης εναλλακτικών μεθόδων διαχείρισης του υπό μελέτη προβλήματος, αλλά και εφαρμογής της μεθόδου σε πραγματικά δεδομένα. Η αξιοποίηση του φίλτρου Kalman (Kalman (1960)) με τη βοήθεια της προσέγγισης των πολυωνυμικών

κατανομών από κανονικές για μεγάλους πληθυσμούς, θα μπορούσε ενδεχομένως να ελαττώσει περαιτέρω το υπολογιστικό κόστος της υπό μελέτη διαδικασίας. Ως προς τις δυνατές εφαρμογές της προτεινόμενης μεθόδου, αυτή ανταποκρίνεται στο μοντέλο μετανάστευσης πτηνών, όπως σχεδιάστηκε από τους Sheldon et al. (2013), κατά το οποίο καταγράφονται μόνο τα αθροιστικά δεδομένα, για την κατανομή του πληθυσμού των πτηνών σε πλέγμα, και λαμβάνονται παρατηρήσεις με θόρυβο. Αντιστοίχως, κατά τη μοντελοποίηση του πιστωτικού κινδύνου στα πρότυπα της μελέτης Jones (2005), η μέθοδος που παρουσιάζεται συστήνεται προκειμένου ο θόρυβος των διαθέσιμων εποπτικών στοιχείων να ληφθεί υπόψη. Ακόμη, το μοντέλο που θεμελιώθηκε στην παρούσα εργασία θα μπορούσε να συμβάλει σε σεισμολογικές μελέτες με την εξέλιξη υπαρχόντων ΚΜΜ επάνω στα επίπεδα τάσης (Votsi et al. (2013)) και άλλων, αντίστοιχων μοντέλων Μαρκοβιανών διαδικασιών αφίξεων που αποσκοπούν στη σεισμική προβλεψιμότητα (Bountzis et al. (2019)) σε νέα πιο σύνθετα μοντέλα.

ABSTRACT

A closed discrete-time Hidden Homogeneous Markov System is studied as a state-space model where the hidden model is known in the literature as Homogeneous Markov System (HMS). This approach lays the foundations for the employment of statistical filters, in order for the distribution of the members in the states to be estimated. The case of systems with big populations is handled by means of a continuous variable filter, the particle filter. The importance of the particularity of the model for parameter estimation is also highlighted.

ΑΝΑΦΟΡΕΣ

Anderson, T. W. and Goodman, L. A. (1957). Statistical Inference about Markov Chains. The Annals of Mathematical Statistics **28**, 89-110.

Bartholomew, D.J. (1963). A Multi-Stage Renewal Process. Journal of the Royal Statistical Society: Series B (Methodological) **25**, 150-168.

Bartholomew, D.J. (1982). Stochastic models for social processes. 3rd edn. Wiley, Chichester (1st edn. 1967, 2nd edn. 1973)

Bartholomew, D. J., Forbes, A. F. and McClean, S. I. (1991). Statistical techniques for manpower planning. Wiley, Chichester; New York.

Beaulieu, N. (1991). On the generalized multinomial distribution, optimal multinomial detectors, and generalized weighted partial decision detectors. IEEE Transactions on Communications **39**, 193-194.

Benneyan, J. and Borgman, A. D. (2004). A useful j-binomial type distribution for non-homogeneous dichotomous events. Industrial Engineering Research Conference Proceedings, 1861-1866.

Bernstein, G. and Sheldon, D. (2016). Consistently estimating Markov chains with noisy aggregate data. In Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016. Cadiz, Spain.

Bountzis, P., Papadimitriou, E. and Tsaklidis, G. (2019). Estimating the160earthquake occurrence rates in Corinth Gulf (Greece) through Markovian arrival160process modeling. Journal of Applied Statistics **46**, 995-1020.

Chen, Z. (2003). Bayesian filtering: From Kalman filters to particle filters, and beyond. Statistics, **182**(1), 1-69.

Crisan, D., and Doucet, A. (2002). A survey of convergence results on particle filtering methods for practitioners. IEEE Transactions on signal processing, **50**(3), 736-746.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. Journal of the Royal Statistical Society: Series B (Methodological) **39**, 1-22.

Doucet, A., & Johansen, A. M. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. Handbook of nonlinear filtering, **12**(656-704), 3.

Feyter, T. and Guerry, M.-A. (2011). Markov models in manpower planning: A review. Handbook of Optimization Theory: Decision Analysis and Application 67-88.

Flaxman, S. R., Wang, Y.-X. and Smola, A. J. (2015). Who Supported Obama in 2012? In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. pp. 289-298.

Forney, G. (1973). The viterbi algorithm. Proceedings of the IEEE **61**, 268-278

Gordon, N., Salmond, D. and Smith, A. (1993). Novel approach to160nonlinear/non-Gaussian Bayesian state estimation. IEE Proceedings F Radar and160Signal Processing **140**, 107-113.

Iwata, T. and Shimizu, H. (2019). Neural Collective Graphical Models for Estimating Spatio-Temporal Population Flow from Aggregated Data. Proceedings of the AAAI Conference on Artificial Intelligence **33**, 3935-3942.

Jones, M. T. (2005). Estimating Markov Transition Matrices Using Proportions Data: An Application to Credit Risk. IMF Working Papers **05**, 1.

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. Journal of Basic Engineering **82**, 35-45.

Kipouridis, I. and Tsaklidis, G. (2001). The size order of the state vector of discrete-time homogeneous Markov systems. Journal of Applied Probability **38**, 357-368.

Liu, J. S. (2004). Monte Carlo Strategies in Scientific Computing. Springer Series.160in Statistics. Springer New York, New York, NY.

Lyubchyk, L., Grinberg, G., Dunaievska, O. and Lubchick, M. (2019). Recurrent Estimation of Hidden Markov Model Transition Probabilities from Aggregate Data. In 2019 9th International Conference on Advanced Computer Information Technologies (ACIT). IEEE, Ceske Budejovice, Czech Republic. pp. 64-67.

Ma, S., Liu, S., Zha, H. and Zhou, H. (2020). Learning Stochastic Behaviour of Aggregate Data.

McKenzie, Eddie (2003). “Ch. 16. Discrete variate time series”. Handbook of Statistics, 573–606.

Osagiede, A. and Ekhosuehi, V. (2006). Markovian approach to school enrolement projection process. Global Journal of Mathematical Sciences **5**.

Papadopoulou, A. and Vassiliou, P. C. (2014). On the variances and convariances of the duration state sizes of semi-Markov systems. Communications in Statistics-Theory and Methods, **43**(7), 1470-1483.

Poisson, S. D. (1837). Recherches sur la probabilit´e des jugements en matière criminelle et en matière civile. Bachelier

Sheldon, D. and Dietterich, T. G. (2011). Collective graphical models. In Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011. Granada, Spain.

Sheldon, D., Elmohamed, M. and Kozen, D. (2007). Collective inference on markov models for modeling bird migration. In Advances in Neural Information Processing Systems.

Sheldon, Daniel, Sun, Tao, Kumar, Akshat, and Dietterich, Thomas G. Approximate Inference in Collective Graphical Models. In proceedings of the 30th International Conference on Machine Learning, 2013.

Singh, R., Zhang, Q., & Chen, Y. (2022). Learning hidden Markov models from aggregate observations. Automatica, **137**, 110100.

Stratonovich, R. L. (1965). Conditional markov processes. In Non-linear transformations of stochastic processes. Elsevier, 427-453.

Terada, M. and Nagata, T. (2013). Population estimation technology for mobile spatial statistics. In NTT DOCOMO Technical Journal. vol. **14**, 10-15.

Tsaklidis, G. M. (1994). The evolution of the attainable structures of a homogeneous Markov system by fixed size. Journal of Applied Probability **31**, 348-361.

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 **45**, 1403-1423.

Vasiliadis, G. and Tsaklidis, G. (2007). On the moments of the state sizes of the discrete time homogeneous Markov system with a finite state capacity. In Recent Advances in Stochastic Modeling and Data Analysis. WORLD SCIENTIFIC, 190-197.

Vasiliadis, G. and Tsaklidis, G. (2008). On the Distributions of the State Sizes of Discrete Time Homogeneous Markov Systems. Methodology and Computing in Applied Probability **10**, 55-71.

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.

Vassiliou, P.-C. G. (1982). Asymptotic behaviour of Markov systems. Journal of Applied Probability **19**, 851–857.

Vassiliou, P.-C. G. (1997). The evolution of the theory of non-homogeneous Markov systems. Applied Stochastic Models and Data Analysis **13**(3-4), 159–534.

Vassiliou, PC.G (2020a). Laws of Large Numbers for Non-Homogeneous Markov Systems. Methodology and Computing in Applied Probability **22**(4), 1631–1658. https://doi.org/10.1007/s11009-017-9612-1

Vassiliou, P.-C.G. (2020b). Rate of Convergence and Periodicity of the Expected Population Structure of Markov Systems that Live in a General State Space Mathematics **8**(6), 1021. https://doi.org/10.3390/math8061021

Vassiliou, P.-C.G. (2021). Non-Homogeneous Markov Set Systems. Mathematics **9**(5), 471. https://doi.org/10.3390/math9050471

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically162optimum decoding algorithm. IEEE Transactions on Information Theory **13**,162260-269.

Votsi, I., Limnios, N., Tsaklidis, G. and Papadimitriou, E. (2013). Hidden Markov models revealing the stress field underlying the earthquake generation. Physica A: Statistical Mechanics and its Applications **392**, 2868-2885.

Zeng, S. (2019). Sample-based population observers. Automatica **101**, 166-174.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 33ου Πανελληνίου Συνεδρίου Στατιστικής (2021), σελ.163-177

ΜΕΛΕΤΗ ΤΗΣ ΑΝΑΠΤΥΞΗΣ ΤΗΣ ΨΗΦΙΑΚΗΣ ΚΟΙΝΩΝΙΑΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ ΣΤΗΝ ΕΥΡΩΠΑΙΚΗ ΕΝΩΣΗ

Μελπομένη Μασούρα[1], Σόνια Μαλεφάκη[2]
[1]Σχολή Θετικών Επιστημών και Τεχνολογίας, Ελληνικό Ανοιχτό Πανεπιστήμιο
std138386@eap.gr
[2]Τμήμα Μηχανολόγων & Αεροναυπηγών Μηχανικών, Πανεπιστήμιο Πατρών
smalefaki@upatras.gr

ΠΕΡΙΛΗΨΗ

Η ραγδαία εξέλιξη των Τεχνολογιών Πληροφορίας και Επικοινωνίας (ΤΠΕ) τα τελευταία χρόνια, έχει επιφέρει σημαντικές αλλαγές σε πολλούς κοινωνικούς τομείς όπως στην επικοινωνία, στην οικονομία, στην ψυχαγωγία και άλλους. Η Ευρωπαϊκή Ένωση (ΕΕ) αναγνωρίζοντας το βασικό ρόλο που παίζουν οι ΤΠΕ στην αναπτυξιακή της πορεία, ανέπτυξε έναν σύνθετο δείκτη, το Δείκτη Ψηφιακής Οικονομίας και Κοινωνίας (Digital Economy and Society Index - DESI) με σκοπό την αξιολόγηση του ψηφιακού επιπέδου των κρατών-μελών της. Στην παρούσα εργασία επιχειρείται να γίνει αποτίμηση της ανάπτυξης της ψηφιακής οικονομίας και κοινωνίας στην ΕΕ μελετώντας τις πέντε διαστάσεις του δείκτη DESI για τα έτη 2014-2019, χρησιμοποιώντας τις αντίστοιχες εκθέσεις DESI (DESI 2015 - DESI 2020). Σκοπός μας είναι να μελετηθούν οι πέντε διαστάσεις του δείκτη DESI και να ομαδοποιηθούν οι χώρες-μέλη της ΕΕ με βάση τις επιδόσεις τους σε αυτές χρησιμοποιώντας γνωστές τεχνικές συσταδοποίησης (clustering). Έτσι δημιουργούνται δύο ομάδες στις οποίες κατατάσσονται οι χώρες-μέλη της ΕΕ, αυτές με Υψηλές και αυτές με Χαμηλές επιδόσεις στις πέντε διαστάσεις του δείκτη DESI. Η εξέλιξη κάθε χώρας-μέλους και οι πιθανές μεταπτώσεις της από την μία ομάδα στην άλλη κατά τη διάρκεια των ετών 2014-2019 αποτελεί επίσης σημείο ενδιαφέροντος. Από την ομαδοποίηση των χωρών μελών της ΕΕ στις δύο ομάδες παρατηρήθηκε ότι κοινωνικο-οικονομικοί παράγοντες ενδεχομένως να επηρεάζουν το δείκτη DESI. Μέσω μοντέλων τυχαίων επιδράσεων (Mixed Effect Models) επιβεβαιώνεται η επίδραση του κατά κεφαλήν Ακαθάριστου Εγχώριου Προϊόντος (ΑΕΠ) και του μέσου αριθμού εβδομαδιαίων ωρών εργασίας στον δείκτη DESI, κάτι που όμως δεν επαληθεύεται για την ανεργία.

Λέξεις Κλειδιά: ψηφιακή οικονομία, ψηφιακή κοινωνία, δείκτης DESI, k-means, mixed effect models

1. ΕΙΣΑΓΩΓΗ

Τα τελευταία χρόνια η ανάπτυξη της τεχνολογίας και των δικτύων είναι ραγδαία και αυτό έχει άμεσες επιδράσεις στην εξέλιξη της κοινωνίας. Η ολοένα και συχνότερη επαφή του ανθρώπου με τις νέες ψηφιακές τεχνολογίες, έχει επιφέρει σημαντικές αλλαγές σε πολλούς τομείς της ζωής του και στο επίκεντρο όλων βρίσκεται η έννοια της ψηφιοποίησης (Castells, 2010). Με την ψηφιοποίηση και την ανάπτυξη των Τεχνολογιών Πληροφορίας και Επικοινωνιών (ΤΠΕ) αναδιαρθρώνονται πολλοί τομείς της κοινωνικής ζωής του ανθρώπου όπως η εργασία, η διοίκηση των επιχειρήσεων, η επικοινωνία, η ψυχαγωγία, η εκπαίδευση και η οικονομία (Parviainen, Tihinen, Kääriäinen, & Teppola, 2017).

Η Ευρώπη αναγνωρίζοντας το βασικό ρόλο που παίζουν οι ΤΠΕ στην αναπτυξιακή τής πορεία, ανέπτυξε τον ετήσιο Δείκτη Ψηφιακής Οικονομίας και Κοινωνίας (Digital Economy and Society Index - DESI) με σκοπό την αξιολόγηση των ψηφιακών επιδόσεων των κρατών-μελών της. Ο δείκτης DESI αποτελείται από πέντε διαστάσεις που αντιστοιχούν στους πέντε βασικούς τομείς πολιτικής της ΕΕ (Συνδεσιμότητα, Ανθρώπινο Κεφάλαιο, Χρήση Υπηρεσιών Διαδικτύου, Ενσωμάτωση Ψηφιακής Τεχνολογίας, Ψηφιακές Δημόσιες Υπηρεσίες), οι οποίες συγκεντρώνουν συνολικά 37 μεμονωμένους δείκτες και αποτυπώνουν την ψηφιακή ανάπτυξη των χωρών-μελών της ΕΕ (European Commission, Digital Economy and Society Index 2020: Methodological Note, 2020).

Οι (Bánhidi, Dobos, & Nemeslaki, 2020) μελέτησαν τις συσχετίσεις και τις μερικές συσχετίσεις των πέντε διαστάσεων του δείκτη DESI παρατηρώντας σχέσεις αιτίου-αποτελέσματος μεταξύ αυτών των διαστάσεων στα δεδομένα της έκθεσης DESI 2018. Εφαρμόζοντας ανάλυση κύριων συνιστωσών (Principal Component Analysis-PCA), κ μείωσαν τις αρχικές μεταβλητές από πέντε σε δυο και τέλος, ομαδοποίησαν τις χώρες-μέλη της ΕΕ εφαρμόζοντας την τεχνική της ιεραρχικής ομαδοποίηση (hierarchical clustering) καθώς και τεχνικές πολυδιάστατης κλιμάκωσης (Multi Dimensional Scaling - MDS). Σύμφωνα με τα αποτελέσματα της μελέτης τους, η επαρκής γνώση της σχέσης αιτίου-αποτελέσματος μεταξύ των διαστάσεων του DESI θα βοηθήσει τους εκάστοτε υπευθύνους στην ορθότερη λήψη αποφάσεων και χάραξη πολιτικής για τη βελτίωση της ανταγωνιστικότητας σε τοπικό, περιφερειακό και Ευρωπαϊκό επίπεδο.

Η μελέτη της (Huseyin, 2021) χρησιμοποίησε δεδομένα της έκθεσης DESI 2020 και εφαρμόζοντας τον αλγόριθμο K-means ομαδοποίησε τις χώρες-μέλη της ΕΕ σε 4 ομάδες, σύμφωνα με τις επιδόσεις τους στις πέντε διαστάσεις του δείκτη DESI. Επίσης εξέτασε αν υπάρχουν ομοιότητες μεταξύ της ομαδοποίησης που προέκυψε και της ταξινόμησης των καθεστώτων ευημερίας (welfare state regimes) του Esping-Andersen (Esping-Andersen, 1990), της Νότιας (Kammer, Niehues, & Peichl, 2012), της Κεντρικής και της Ανατολικής Ευρώπης (Lauzadyte-Tutliene, Balezentis, & Goculenko, 2018). Στη συγκεκριμένη μελέτη οι χώρες μέλη της ΕΕ χωρίζονται στους παρακάτω έξι τύπους καθεστώτων ευημερίας: τα Σοσιαλδημοκρατικά, τα Συντηρητικά, τα Φιλελεύθερα, τα καθεστώτα Νότιας, Κεντρικής και Ανατολικής

Ευρώπης. Ο κάθε τύπος καθεστώτος ευημερίας χαρακτηρίζεται από την πολιτική (εκλέξιμη κυβέρνηση ή όχι), την κοινωνική (αγορά εργασίας, διαφορές μεταξύ των φύλων στην αγορά εργασίας, δημογραφική κατάσταση, εκπαίδευση πληθυσμού, συνθήκες διαβίωσης) και την οικονομική κατάσταση (πραγματικό ΑΕΠ) μιας χώρας όπως επίσης και από τα κυβερνητικά προγράμματα που παρέχονται στους πολίτες. Από τη μελέτη προέκυψε πως οι χώρες που εφαρμόζουν το σοσιαλδημοκρατικό καθεστώς ευημερίας (Φιλανδία, Σουηδία, Δανία, Ολλανδία, Νορβηγία) είναι πιο προετοιμασμένες για την ψηφιοποίηση της οικονομίας και τον μελλοντικό ψηφιακό ανταγωνισμό σε σύγκριση με άλλες χώρες της ΕΕ, καθώς εμφανίζουν τις υψηλότερες τιμές σε όλες τις διαστάσεις του δείκτη DESI. Ακολουθούν οι χώρες φιλελεύθερων καθεστώτων ευημερίας (Ηνωμένο Βασίλειο, Ιρλανδία), με μικρή διαφορά από την πρώτη κατηγορία και στη συνέχεια οι χώρες συντηρητικών καθεστώτων ευημερίας (Γερμανία, Λουξεμβούργο, Βέλγιο, Αυστρία, Γαλλία). Τέλος, οι υπόλοιπες χώρες (Ιταλία, Ισπανία, Ελλάδα, Πορτογαλία, Κροατία, Πολωνία, Σλοβακία, Σλοβενία, Βουλγαρία, Εσθονία, Λετονία, Λιθουανία, Ρουμανία) που ανήκουν στις τρεις τελευταίες κατηγορίες έχουν τις χαμηλότερες τιμές σε όλες τις διαστάσεις του DESI και είναι πιο πιθανό να αντιμετωπίσουν προβλήματα στο μέλλον όσον αφορά τον τεχνολογικό μετασχηματισμό και τον ψηφιακό ανταγωνισμό με τις άλλες χώρες της ΕΕ.

Ο δείκτης DESI παρουσιάζοντας τα δυνατά και τα αδύνατα σημεία των χωρών της ΕΕ ως προς τις ΤΠΕ συμμετέχει στη λήψη αποφάσεων για τη χάραξη πολιτικών όσο αφορά στον ψηφιακό μετασχηματισμό της ΕΕ. Επιπλέον η λήψη αποφάσεων δεν μπορεί να είναι ανεξάρτητη από τα καθεστώτα ευημερίας των χωρών, καθώς αυτά επηρεάζουν άμεσα τις πολιτικές ψηφιακού μετασχηματισμού.

Από όσο γνωρίζουμε, δεν υπάρχουν εργασίες που να μελετούν την εξέλιξη του δείκτη DESI στο χρόνο και να εξετάζουν κοινωνικό-οικονομικούς παράγοντες που ενδεχομένως τον επηρεάζουν. Η εργασία μας, έρχεται να συνδράμει προς αυτή την κατεύθυνση. Η λοιπή εργασία έχει οργανωθεί ως εξής: Στη δεύτερη ενότητα παρουσιάζεται η δομή του δείκτη DESI και περιγράφονται οι πέντε διαστάσεις του. Στην τρίτη ενότητα μελετάμε την εξέλιξη στο χρόνο του δείκτη DESI ενώ στην τέταρτη ενότητα γίνεται διερεύνηση της επίδρασης των κοινωνικο-οικονομικών παραγόντων όπως το κατά κεφαλήν Ακαθάριστο Εγχώριο Προϊόν (ΑΕΠ), το μέσο αριθμό εβδομαδιαίων ωρών εργασίας και την ανεργία στον δείκτη DESI. Στην τελευταία ενότητα παρουσιάζονται τα σημαντικότερα συμπεράσματα της παρούσας εργασίας και επισημαίνονται θέματα για περαιτέρω έρευνα.

2. Ο ΔΕΙΚΤΗΣ DESI

Ο Δείκτης Ψηφιακής Κοινωνίας και Οικονομίας (Digital Economy and Society Index - DESI) είναι ένας σύνθετος δείκτης που συνοψίζει σχετικούς δείκτες που αφορούν στην ψηφιακή επίδοση της Ευρωπαϊκής Ένωσης (ΕΕ) και παρακολουθεί την εξέλιξη των κρατών μελών της στο πεδίο της ψηφιακής ανταγωνιστικότητας. Ο DESI αποτελεί από το 2014 ένα βασικό αναλυτικό εργαλείο για την ετήσια μέτρηση της

προόδου των χωρών της ΕΕ με κατεύθυνση την ψηφιακή οικονομία και κοινωνία (G20, 2018).

2.1 Η δομή του δείκτη DESI

Ο δείκτης DESI αποτελείται από πέντε διαστάσεις που αντιστοιχούν στους πέντε βασικούς τομείς πολιτικής της ΕΕ οι οποίοι με τη σειρά τους αποτελούνται συνολικά από 37 δείκτες (European Commission, Digital Economy and Society Index 2020: Methodological Note, 2020). Η δομή του δείκτη είναι τριών επιπέδων, με 5 διαστάσεις. Κάθε διάσταση αποτελείται από υποδιαστάσεις και αυτές με τη σειρά τους από δείκτες. Συνολικά οι 5 διαστάσεις αποτελούνται από 12 υποδιαστάσεις και αυτές συνολικά από 37 δείκτες.

Η διάσταση **Συνδεσιμότητα**, εξετάζει τόσο τη ζήτηση όσο και την προσφορά σταθερών και κινητών ευρυζωνικών συνδέσεων (European Commission, Digital Economy and Society Index (DESI) 2020: Connectivity, 2020).

Η διάσταση **Ανθρώπινο Κεφάλαιο**, αφορά στις ψηφιακές δεξιότητες των πολιτών της ΕΕ που κυμαίνονται από βασικές δεξιότητες χρήσης που επιτρέπουν στα άτομα να συμμετέχουν στην ψηφιακή κοινωνία και να καταναλώνουν ψηφιακά αγαθά και υπηρεσίες, έως προηγμένες δεξιότητες που ενδυναμώνουν το εργατικό δυναμικό για την ανάπτυξη νέων ψηφιακών αγαθών και υπηρεσιών (European Commission, Digital Economy and Society Index (DESI) 2020: Human capital, 2020).

Η διάσταση **Χρήση Υπηρεσιών Διαδικτύου,** μετρά το ποσοστό των ατόμων που χρησιμοποιούν υπηρεσίες διαδικτύου και ποιες από τις διαθέσιμες υπηρεσίες χρησιμοποιούν. Οι δραστηριότητες περιλαμβάνουν την κατανάλωση διαδικτυακού περιεχομένου (π.χ. ψυχαγωγία όπως μουσική, ταινίες, τηλεόραση ή παιχνίδια, λήψη πληροφοριών πλούσιων σε πολυμέσα ή συμμετοχή σε διαδικτυακή κοινωνική αλληλεπίδραση), χρήση σύγχρονων δραστηριοτήτων επικοινωνίας (π.χ. συμμετοχή σε βιντεοκλήσεις) και δραστηριότητες συναλλαγών όπως οι online αγορές και οι τραπεζικές συναλλαγές (European Commission, Digital Economy and Society Index (DESI) 2020: Use of internet services, 2020).

Η διάσταση **Ενσωμάτωση Ψηφιακής Τεχνολογίας** μετρά την ψηφιοποίηση των επιχειρήσεων και του ηλεκτρονικού εμπορίου (European Commission, Digital Economy and Society Index (DESI) 2020: Integration of digital technology, 2020).

Τέλος, η διάσταση **Ψηφιακές Δημόσιες Υπηρεσίες** αφορά τόσο στη ζήτηση όσο και στην προσφορά ψηφιακών δημόσιων υπηρεσιών, καθώς και τις πολιτικές διαχείρισης των ανοιχτών δεδομένων (European Commission, Digital Economy and Society Index (DESI) 2020: Digital public services, 2020).

2.2 Συλλογή δεδομένων - τυποποίηση

Τα δεδομένα που χρησιμοποιεί η Ευρωπαϊκή Επιτροπή για τις ετήσιες εκθέσεις του δείκτη DESI αφορούν στη χρονική περίοδο από 1η Φεβρουαρίου του προηγούμενου έτους έως και 31 Ιανουαρίου του επόμενου.

Προκειμένου να υπολογιστούν οι πέντε διαστάσεις του δείκτη DESI αλλά και ο συνολικός δείκτης DESI, οι αρχικοί 37 δείκτες τυποποιούνται καθώς εκφράζονται αρχικά σε διαφορετικές μονάδες μέτρησης. Η τυποποίηση γίνεται χρησιμοποιώντας τη μέθοδο min-max, η οποία συνίσταται σε μια γραμμική προβολή κάθε δείκτη σε μια κλίμακα μεταξύ 0 και 1.

2.3 Απόδοση βαρών - κατασκευή δείκτη

Ο συνολικός δείκτης DESI προκύπτει από τη συνένωση των δεικτών σε υποδιαστάσεις, έπειτα των υποδιαστάσεων σε διαστάσεις και τέλος των διαστάσεων στο συνολικό δείκτη, χρησιμοποιώντας βάρη. Ορισμένες διαστάσεις, υποδιαστάσεις και μεμονωμένοι δείκτες είναι πιο σημαντικοί από άλλους και γι' αυτό τους δίνεται υψηλότερο βάρος στον υπολογισμό της τελικής τιμής του δείκτη DESI (European Commission, Digital Economy and Society Index 2020: Methodological Note, 2020). Τα συνολικά βάρη που αποδίδονται στις κύριες διαστάσεις του DESI αντικατοπτρίζουν τις προτεραιότητες της ψηφιακής πολιτικής της ΕΕ. Ενδεικτικά αναφέρουμε ότι τα βάρη που χρησιμοποιούνται σε επίπεδο διάστασης είναι: **Συνδεσιμότητα**: 25%, **Ανθρώπινο Κεφάλαιο**: 25%, **Χρήση Υπηρεσιών Διαδικτύου**: 15%, **Ενσωμάτωση Ψηφιακής Τεχνολογίας**: 20% και **Ψηφιακές Δημόσιες Υπηρεσίες**: 15%.

3. ΕΞΕΛΙΞΗ ΔΕΙΚΤΗ DESI ΩΣ ΠΡΟΣ ΤΟ ΧΡΟΝΟ

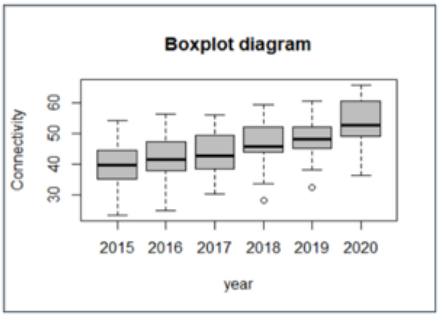
Στη συνέχεια θα μελετήσουμε την εξέλιξη των πέντε διαστάσεων του δείκτη DESI στο χρόνο. Προκειμένου να επιλεγεί κατάλληλος στατιστικός έλεγχος, αρχικά ελέγχουμε την κανονικότητα των δεδομένων μας. Για τον έλεγχο της κανονικότητας χρησιμοποιήθηκε το Shapiro-Wilk normality test (Shapiro & Wilk, 1965) και διαπιστώθηκε πως όλες οι μεταβλητές ακολουθούν κανονική κατανομή οπότε σε κάθε διάσταση χρησιμοποιείται ανάλυση διασποράς σε σχεδιασμό επαναλαμβανόμενων μετρήσεων (Repeated Measure ANOVA) (Salkind, 2010). Από τα αποτελέσματα του Repeated Measure ANOVA προκύπτει ότι πράγματι υπάρχει στατιστικά σημαντική διαφορά στη μέση επίδοση σε όλες τις διαστάσεις του δείκτη DESI κατά το χρονικό διάστημα 2014 -2019, το οποίο αποτυπώνεται στις εκθέσεις DESI 2015 – DESI 2020 (https://digital-agenda-data.eu/datasets/desi). Στη συνέχεια εφαρμόζουμε έλεγχο Tukey (Tukey test) για να δούμε ακριβώς, ανάμεσα σε ποια έτη παρατηρούνται αυτές οι στατιστικά σημαντικές διαφορές. Στις Εικόνες 3.1-3.5 παρουσιάζονται τα θηκογράμματα κάθε διάστασης ανά έτος και οι πίνακες με τις p-τιμές (p-values) των παραπάνω ελέγχων. Τα γραμμοσκιασμένα κελιά των πινάκων των p-values των ελέγχων Tukey αφορούν στα έτη μεταξύ των οποίων υπάρχει στατιστικά σημαντική διαφορά ($p\text{-value}<\alpha=0.05$).

3.1 Διάσταση Συνδεσιμότητα

Παρατηρούμε μια σταθερή ανοδική πορεία της διαμέσου της διάστασης Συνδεσιμότητας με την πάροδο του χρόνου (Εικόνα 3.1 αριστερά). Από την Εικόνα 3.1 (αριστερά), παρατηρούμε τον συγκεκριμένο δείκτη να εξελίσσεται πολύ γρήγορα

στο χρόνο, πράγμα που επιβεβαιώνεται και από τους αντίστοιχους Tukey ελέγχους (Εικόνα 3.1 δεξιά) καθώς σε όλες τις εκθέσεις DESI από το 2015 έως το 2020 υπάρχει στατιστικά σημαντική διαφορά ως προς τις μέσες επιδόσεις των χωρών μελών της ΕΕ σε αυτή τη διάσταση, από χρονιά σε χρονιά. Αυτή η διάσταση αφορά στις τεχνολογικές υποδομές αλλά και την προσφορά και ζήτηση των τεχνολογικών παροχών. Πιο συγκεκριμένα, αφορά στις υποδομές ως προς τις ταχύτητες κινητών και σταθερών συνδέσεων αλλά και στη δυνατότητα σύνδεσης 3G, 4G και στην ετοιμότητα για 5G. Αυτή η διάσταση θα λέγαμε ότι αποτελεί τη βάση που θα δώσει τον απαραίτητο χώρο και τρόπο για να μπορέσουν να εξελιχθούν και οι άλλες διαστάσεις.

Εικόνα 3.1. *Διάσταση Συνδεσιμότητα: Θηκόγραμμα (αριστερά), Πίνακας p-values των ελέγχων Tukey (δεξιά)*

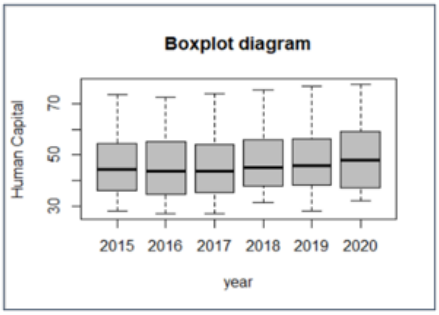


Tukey Pairwise Comparisons (p-values)					
Συνδεσιμότητα	**2015**	**2016**	**2017**	**2018**	**2019**
2016	0.0008				
2017	<.0001	0.0315			
2018	<.0001	<.0001	<.0001		
2019	<.0001	<.0001	<.0001	0.0208	
2020	<.0001	<.0001	<.0001	<.0001	<.0001

3.2 Διάσταση Ανθρώπινο Κεφάλαιο

Αρχικά θα πρέπει να επισημάνουμε ότι δεν υπάρχει στατιστικά σημαντική διαφορά ως προς τις μέσες επιδόσεις των χωρών μελών της ΕΕ στη διάσταση Ανθρώπινο κεφάλαιο ανάμεσα στις εκθέσεις DESI 2015, 2016 και 2017. Στατιστικά σημαντική διαφορά εμφανίζεται για πρώτη φορά ανάμεσα στις εκθέσεις DESI 2017 - DESI 2018 και ανάμεσα στις εκθέσεις DESI 2019 - DESI 2020 (Εικόνα 3.2 δεξιά) γεγονός που ενδεχομένως οφείλεται στο ότι ο πληθυσμός είναι πιθανό να χρειάζεται χρόνο ώστε να εξοικειωθεί με τη νέα τεχνολογία και να κατανοήσει πως μπορεί να την εκμεταλλευτεί.

Εικόνα 3.2. *Διάσταση Ανθρώπινο Κεφάλαιο: Θηκόγραμμα (αριστερά) Πίνακας p-values των ελέγχων Tukey (δεξιά)*

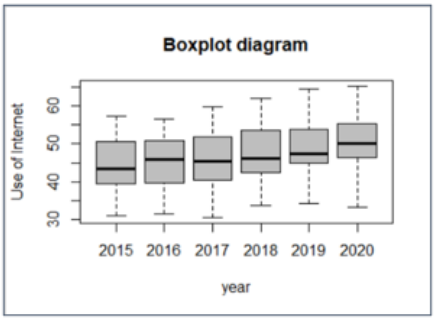


Tukey Pairwise Comparisons (p-values)					
Ανθρώπινο Κεφάλαιο	**2015**	**2016**	**2017**	**2018**	**2019**
2016	0.4031				
2017	1	0.4105			
2018	0.001	<.0001	0.001		
2019	<.0001	<.0001	<.0001	0.8341	
2020	<.0001	<.0001	<.0001	0.0001	0.0106

3.3 Διάσταση Χρήση Υπηρεσιών Διαδικτύου

Παρατηρούμε μια σταθερή ανοδική πορεία της διαμέσου της Χρήσης Υπηρεσιών Διαδικτύου με την πάροδο του χρόνου (Εικόνα 3.3 αριστερά). Από την Εικόνα 3.3 (δεξιά), παρατηρούμε ότι και στη διάσταση Χρήση Υπηρεσιών Διαδικτύου υπάρχει επίσης καθυστέρηση στην εμφάνιση στατιστικά σημαντικής διαφοράς μεταξύ των μέσων επιδόσεων των χωρών μελών της ΕΕ, καθώς αυτή εμφανίζεται από την έκθεση DESI 2018 και μετά.

Εικόνα 3.3. *Διάσταση Χρήση Υπηρεσιών Διαδικτύου: Θηκόγραμμα (αριστερά) Πίνακας p-values των ελέγχων Tukey (δεξιά)*



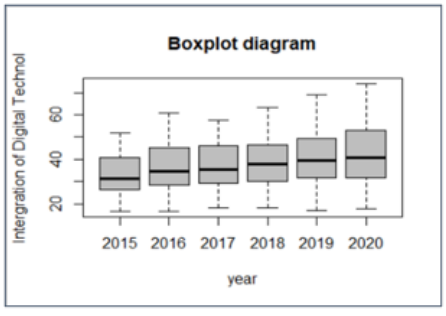
Tukey Pairwise Comparisons (p-values)					
Χρήση Υπηρεσιών Διαδικτύου	**2015**	**2016**	**2017**	**2018**	**2019**
2016	0.8109				
2017	0.1924	0.8918			
2018	<.0001	<.0001	0.0021		
2019	<.0001	<.0001	<.0001	0.0024	
2020	<.0001	<.0001	<.0001	<.0001	0.0138

Ωστόσο, όπως και στη διάσταση Ανθρώπινο Κεφάλαιο έτσι και στη διάσταση αυτή, υπάρχει κάποιο χρονικό διάστημα με μη στατιστικά σημαντικές διαφορές από χρονιά σε χρονιά. Πιθανά και εδώ να σχετίζεται με το χρόνο που πιθανά χρειάζεται ο πληθυσμός ώστε να εξοικειωθεί με τις υπηρεσίες του διαδικτύου όπως οι ηλεκτρονικές επικοινωνίες και συναλλαγές.

3.4 Διάσταση Ενσωμάτωση Ψηφιακής Τεχνολογίας

Παρατηρούμε μια σταθερή ανοδική πορεία της διαμέσου της Ενσωμάτωσης της Ψηφιακής Τεχνολογίας με την πάροδο του χρόνου (Εικόνα 3.4 αριστερά).

Εικόνα 3.4. *Διάσταση Ενσωμάτωση Ψηφιακής Τεχνολογίας: Θηκόγραμμα (αριστερά) Πίνακας p-values των ελέγχων Tukey (δεξιά)*



Tukey Pairwise Comparisons (p-values)					
Ενσωμάτωση Ψηφιακής Τεχνολογίας	**2015**	**2016**	**2017**	**2018**	**2019**
2016	0.0173				
2017	0.0018	0.985			
2018	<.0001	0.0268	0.1484		
2019	<.0001	<.0001	<.0001	0.1115	
2020	<.0001	<.0001	<.0001	<.0001	0.0445

Από τον πίνακα της Εικόνας 3.4 (δεξιά) παρατηρούμε στατιστικά σημαντική διαφορά στο μέσο επίπεδο της συγκεκριμένης διάστασης σχεδόν κάθε δεύτερη χρονιά.

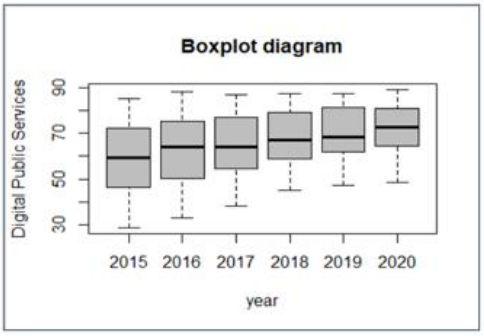
Εξαίρεση αποτελεί το έτος 2015 όπου εμφανίζεται στατιστικά σημαντική διαφορά από την επόμενη κιόλας χρονιά.

Το ηλεκτρονικό επιχειρείν και το ηλεκτρονικό εμπόριο είναι οι δύο υποδιαστάσεις που αφορούν τη διάσταση αυτή και είναι λογικό πως για να γίνει ο ψηφιακός μετασχηματισμός των επιχειρήσεων ως προς τις παραπάνω υποδιαστάσεις απαιτείται ένα εύλογο χρονικό διάστημα. Έτσι μπορούμε να δικαιολογήσουμε και τη χρονική απόσταση των δύο ετών ώστε να εμφανιστεί στατιστικά σημαντική διαφορά για τη συγκεκριμένη διάσταση.

3.5 Διάσταση Ψηφιακές Δημόσιες Υπηρεσίες

Παρατηρούμε μια σταθερή ανοδική πορεία της διαμέσου της διάστασης Ψηφιακές Δημόσιες Υπηρεσίες με την πάροδο του χρόνου (Εικόνα 3.5 αριστερά). Επίσης, από τον πίνακα της Εικόνας 3.5 (δεξιά), παρατηρείται στατιστικά σημαντική διαφορά στις επιδόσεις των χωρών μελών της ΕΕ στη συγκεκριμένη διάσταση, σε όλες τις εκθέσεις DESI από το 2017 μέχρι και το 2019 ενώ για πρώτη φορά παρατηρήθηκε στις εκθέσεις DESI από το 2015 στο 2016. Αυτή η εικόνα είναι πιθανό να οφείλεται στις προτεραιότητες που δίνει κάθε χώρα, είτε εστιάζει στον εκσυγχρονισμό των εθνικών πυλών της είτε στην πολιτική των ανοιχτών δεδομένων.

Εικόνα 3.5. *Διάσταση Ενσωμάτωση Ψηφιακής Τεχνολογίας: Θηκόγραμμα (αριστερά) Πίνακας p-values των ελέγχων Tukey (δεξιά)*



Tukey Pairwise Comparisons (p-values)					
Ψηφιακές Δημόσιες Υπηρεσίες	2015	2016	2017	2018	2019
2016	0.0001				
2017	<.0001	0.5168			
2018	<.0001	<.0001	0.0245		
2019	<.0001	<.0001	<.0001	0.0229	
2020	<.0001	<.0001	<.0001	<.0001	0.1173

Συνοψίζοντας αυτό που μπορούμε να πούμε είναι ότι υπάρχει μια ξεκάθαρη εξέλιξη των πέντε διαστάσεων του δείκτη DESI στο χρόνο με κάποιες διαστάσεις να εξελίσσονται γρηγορότερα από κάποιες άλλες.

3.6 Ανάλυση κατά συστάδες

Στη συνέχεια, ομαδοποιήσαμε τις χώρες-μέλη της ΕΕ με βάση τις επιδόσεις τους στις πέντε διαστάσεις του DESI για το διάστημα 2014-2019, χρησιμοποιώντας δεδομένα από τις αντίστοιχες ετήσιες εκθέσεις DESI 2015 έως DESI 2020.

Για την ομαδοποίηση των χωρών μελών της ΕΕ επιλέξαμε τη μέθοδο k-means και ως μέτρο απόστασης την Ευκλείδεια απόσταση. Για τον προσδιορισμό του βέλτιστου πλήθους συστάδων στις οποίες θα ομαδοποιηθούν τα δεδομένα μας, χρησιμοποιήσαμε τη βιβλιοθήκη `NbClust` της R (Charrad, Ghazzali, Boiteau, &

Niknafs, 2014). Η συγκεκριμένη βιβλιοθήκη, παρέχει 30 δείκτες και προτείνει το βέλτιστο πλήθος συστάδων για τα δεδομένα που προκύπτει από την υλοποίηση της μεθόδου k-means με μέτρο απόστασης την Ευκλείδεια και πλήθος συστάδων από min=2 έως max=15. Από την Εικόνα 3.5 παρατηρούμε πως για κάθε έτος η πλειοψηφία των δεικτών προτείνει 2 συστάδες ως βέλτιστο αριθμό συστάδων. Ωστόσο για το έτος 2019 προτάθηκαν 2 και 3 συστάδες. Έτσι δοκιμάσαμε να κατανείμουμε τα δεδομένα μας σε 3 συστάδες αλλά δεν παρατηρήσαμε στατιστικά σημαντικές διαφορές στις επιδόσεις και στις πέντε διαστάσεις του δείκτη DESI των χωρών μεταξύ της δεύτερη και της τρίτης συστάδας για όλα τα έτη. Συνεπώς καταλήξαμε στις 2 συστάδες. Έτσι προέκυψαν οι ομάδες με Υψηλές και Χαμηλές επιδόσεις στις πέντε διαστάσεις του δείκτη DESI.

Εικόνα 3.5. *Αποτελέσματα από την εφαρμογή της βιβλιοθήκης NbClust της R*

2015	2016	2017
* Among all indices: * 8 proposed 2 as the best number of clusters * 4 proposed 3 as the best number of clusters * 3 proposed 4 as the best number of clusters * 1 proposed 6 as the best number of clusters * 1 proposed 9 as the best number of clusters * 3 proposed 13 as the best number of clusters * 3 proposed 15 as the best number of clusters	* Among all indices: * 7 proposed 2 as the best number of clusters * 4 proposed 3 as the best number of clusters * 4 proposed 4 as the best number of clusters * 1 proposed 5 as the best number of clusters * 1 proposed 11 as the best number of clusters * 1 proposed 12 as the best number of clusters * 1 proposed 13 as the best number of clusters * 4 proposed 15 as the best number of clusters	* Among all indices: * 7 proposed 2 as the best number of clusters * 6 proposed 3 as the best number of clusters * 1 proposed 4 as the best number of clusters * 1 proposed 6 as the best number of clusters * 1 proposed 7 as the best number of clusters * 2 proposed 11 as the best number of clusters * 1 proposed 13 as the best number of clusters * 4 proposed 15 as the best number of clusters
2018	**2019**	**2020**
* Among all indices: * 8 proposed 2 as the best number of clusters * 4 proposed 3 as the best number of clusters * 1 proposed 4 as the best number of clusters * 1 proposed 5 as the best number of clusters * 1 proposed 8 as the best number of clusters * 1 proposed 10 as the best number of clusters * 1 proposed 12 as the best number of clusters * 3 proposed 14 as the best number of clusters * 3 proposed 15 as the best number of clusters	* Among all indices: * 7 proposed 2 as the best number of clusters * 7 proposed 3 as the best number of clusters * 1 proposed 4 as the best number of clusters * 1 proposed 8 as the best number of clusters * 1 proposed 11 as the best number of clusters * 1 proposed 12 as the best number of clusters * 2 proposed 13 as the best number of clusters * 3 proposed 14 as the best number of clusters	* Among all indices: * 9 proposed 2 as the best number of clusters * 7 proposed 3 as the best number of clusters * 1 proposed 4 as the best number of clusters * 1 proposed 5 as the best number of clusters * 2 proposed 9 as the best number of clusters * 1 proposed 12 as the best number of clusters * 3 proposed 15 as the best number of clusters

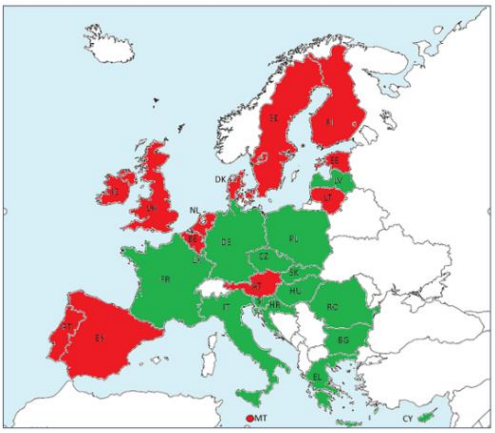
Στην Εικόνα 3.6 απεικονίζεται, ανά χρονιά, ο χάρτης της Ευρώπης όπου με κόκκινη σκιαγράφηση έχουν επισημανθεί οι χώρες με υψηλή επίδοση στον δείκτη DESI και με πράσινη αυτές με χαμηλή επίδοση. Παρατηρούμε πως υπάρχουν χώρες που στην πάροδο των ετών, έχουν διατηρήσει το υψηλό επίπεδο επιδόσεων και αυτές που παραμένουν σε σταθερά χαμηλό επίπεδο επιδόσεων. Επίσης παρατηρούμε χώρες όπως η Ισπανία, η Λιθουανία, η Πορτογαλία και η Αυστρία να διατηρούν επιδόσεις υψηλού επιπέδου για μεγάλα χρονικά διαστήματα ενώ το 2020 καταλήγουν σε χαμηλού επιπέδου τιμές. Τέλος, παρατηρούμε πως κάποιες χώρες όπως η Γαλλία, η Λετονία και η Γερμανία δεν διατηρούν σταθερή πορεία στην πάροδο των ετών και εναλλάσσονται μεταξύ των χαμηλών και υψηλών επιδόσεων.

Όλες οι χώρες εξελίσσονται ως προς το δείκτη DESI και αυτή η μετακίνηση που παρατήσαμε από τη μία κατηγορία στην άλλη, οφείλεται στο διαφορετικό ρυθμό με τον οποίο εξελίσσεται κάθε χώρα σε σχέση με τις υπόλοιπες. Κάποιες χώρες φαίνεται να εξελίσσονται πιο γρήγορα ενώ κάποιες άλλες πιο αργά. Τέλος, παρατηρούμε ότι ο διαχωρισμός σε δύο ομάδες ταιριάζει και με το οικονομικό προφίλ των χωρών.

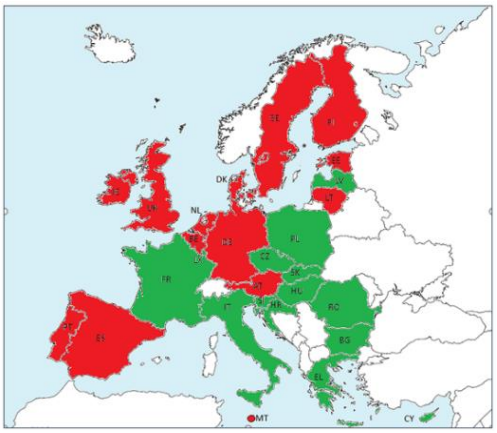
Στη συνέχεια θα μελετήσουμε αν και πως οικονομικοί και κοινωνικοί δείκτες επηρεάζουν τον δείκτη DESI.

Εικόνα 3.6. *Χάρτες Συστάδων χωρών- μελών ΕΕ βάση τις 5 διαστάσεις του DESI*

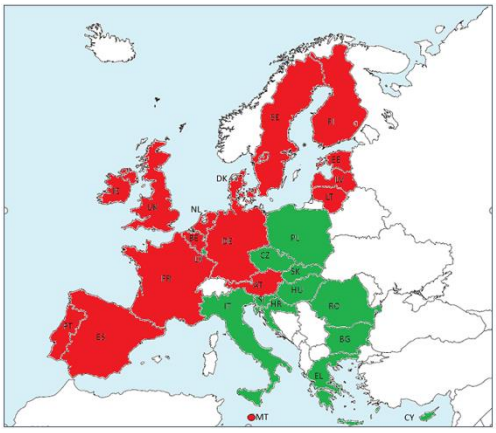
2015



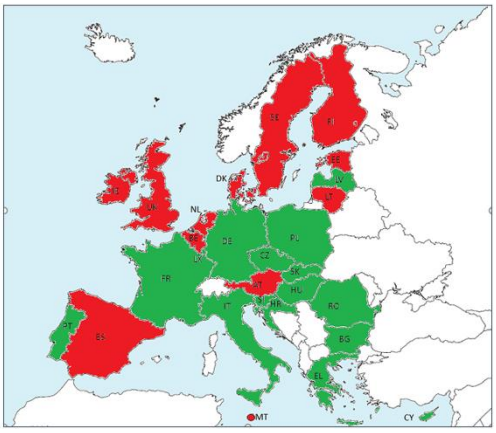
2016



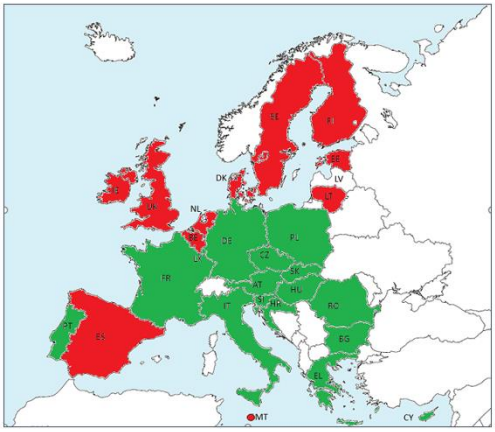
2017



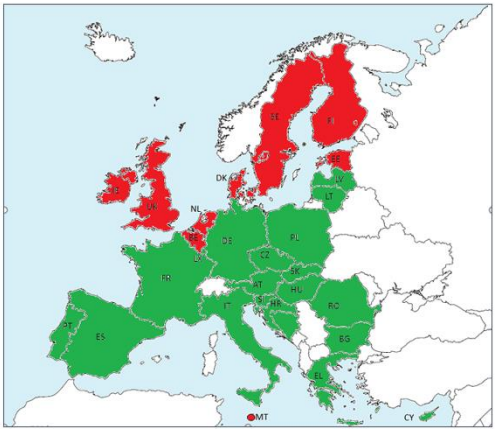
2018



2019



2020



4. ΠΑΡΑΓΟΝΤΕΣ ΠΟΥ ΕΠΗΡΕΑΖΟΥΝ ΤΟ ΔΕΙΚΤΗ DESI

Παρατηρήσαμε ότι υπάρχει μία ομοιογένεια στις ομάδες των χωρών-μελών της ΕΕ που δημιουργήθηκαν βάση των πέντε διαστάσεων του δείκτη DESI και ως προς τα κοινωνικο-οικονομικά χαρακτηριστικά τους, για αυτό θα θέλαμε να δούμε αν και πως κάποιες βασικές οικονομικές και κοινωνικές μεταβλητές όπως: το κατά κεφαλήν Ακαθάριστο Εγχώριο Προϊόν – ΑΕΠ (zRealGDP_PC),(https://ec.europa.eu/eurostat/databrowser/view/NAMA10PC/default/table?lang=en), ο μέσος αριθμός εβδομαδιαίων ωρών εργασίας (HOUR_PER_WEEK),(https://ec.europa.eu/eurostat/databrowser/view/LFSAURGAEDcustom_844619/default/table), και η ανεργία (UNEMPLOYMENT), (https://ec.europa.eu/eurostat/databrowser/view/LFSAEWHUN2custom890938/default/table) επηρεάζουν τον δείκτη DESI. Χρησιμοποιώντας δεδομένα από τις ετήσιες εκθέσεις DESI 2015-2020, κατασκευάσαμε μοντέλα τυχαίων επιδράσεων (Mixed Effect Models), τα οποία μπορούν να περιγράψουν την επίδραση των παραπάνω παραγόντων στη μεταβλητή (DESI).

Χρησιμοποιώντας το πακέτο `lme4` της R (Bates, Bolker, Mächler, & Walker, 2014), κατασκευάσαμε αρχικά δύο μοντέλα. Το μοντέλο 1, με τυχαίους σταθερούς όρους ως προς το χρόνο και τη χώρα και το μοντέλο 2 στο οποίο προσθέσαμε επιπλέον και τυχαίες κλίσεις ως προς τον χρόνο ανά χώρα. Επιλέξαμε το μοντέλο 2 ως βέλτιστο με βάση το κριτήριο Akaike (Burnham & Anderson, 2004) (βλέπε Πίνακα 4.1). Στο μοντέλο αυτό προσθέσαμε αρχικά τον παράγοντα του χρόνου (μοντέλο 3) και στη συνέχεια τους υπόλοιπους παράγοντες, το ΑΕΠ (zRealGDP_PC), το μέσο αριθμό των εβδομαδιαίων ωρών εργασίας (HOUR_PER_WEEK) και την ανεργία (UNEPMPLOYMENT) (μοντέλο 4). Στον Πίνακα 4.1 παρουσιάζεται η σύγκριση των τεσσάρων μοντέλων.

Πίνακας 4.1 *Σύγκριση μοντέλων 1,2,3,4*

Models	*AIC*	*BIC*	*Chisq Df*	*Pr(>Chisq)*
Model 1	683.51	696.00		
Model 2	673.71	692.46	13.794	0.001011**
Model 3	654.79	676.65	20.927	4.77e-06 ***
Model 4	647.58	678.82	13.201	0.004222 **

Πίνακας 4.2. *Μοντέλο 4*

Model 4	**Est**	**S.E.**	**t val.**	**p**
Intercept	75.45	13.10	5.76	0.00
aYear	1.72	0.15	11.42	0.00
UNEPMPLOYMENT	-0.07	0.09	-0.79	0.43
zRealGDP_PC	2.55	0.98	2.61	0.01
HOUR_PER_WEEK	-0.87	0.35	-2.53	0.01

Ωστόσο, στο μοντέλο 4 παρατηρούμε πως ο παράγοντας ανεργία δεν είναι στατιστικά σημαντικός (p-value = 0.43, Πίνακας 4.2) οπότε τον παραλείπουμε από το μοντέλο μας και έτσι οδηγούμαστε στο τελικό μοντέλο (μοντέλο 5). Η σύγκριση των μοντέλων 3 και 5 παρουσιάζεται στον Πίνακα 4.4.

Πίνακας 4.3. *Μοντέλο 5*

Model 5	**Est**	**S.E.**	**t val.**	**p**
Intercept	76.08	13.21	5.76	0.00
aYear	1.76	0.13	13.29	0.00
zRealGDP_PC	2.63	0.98	2.7	0.01
HOUR_PER_WEEK	-0.91	0.35	-2.62	0.01

Πίνακας 4.4*. Σύγκριση Μοντέλων 3 και 5*

Models	*AIC*	*BIC*	*Chisq Df*	*Pr(>Chisq)*
Model 3	654.79	676.65		
Model 5	646.08	674.2	12.703	0.001744 **

Στο μοντέλο 5 (Πίνακας 4.3) παρατηρούμε ότι ο μέσος δείκτης DESI αυξάνεται κατά 1.76 ανά έτος για σταθερές τιμές των μεταβλητών κατά κεφαλήν ΑΕΠ και αριθμό εβδομαδιαίων ωρών εργασίας. Επίσης αυξάνεται κατά 2.63 για κάθε μονάδα αύξησης του κατά κεφαλήν ΑΕΠ για συγκεκριμένη χρονιά και πλήθος εβδομαδιαίων ωρών εργασίας, ενώ μειώνεται κατά 0.91 για κάθε μονάδα αύξησης των εβδομαδιαίων ωρών εργασίας για συγκεκριμένη χρονιά και κατά κεφαλήν ΑΕΠ.

Σύμφωνα με το παραπάνω μοντέλο, οι περισσότερο οικονομικά ανεπτυγμένες χώρες φαίνεται να επενδύουν περισσότερο στον ψηφιακό μετασχηματισμό τόσο της οικονομίας όσο και της κοινωνίας, παρέχοντας στους πολίτες τους τις κατάλληλες τεχνολογικές υποδομές ώστε να συμμετέχουν ενεργά στη νέα ψηφιακή εποχή. Όσο αφορά στις ώρες εργασίας, παρατηρούμε επίσης αύξηση των τιμών του δείκτη DESI καθώς αυτές μειώνονται. Με την ανάπτυξη των νεών τεχνολογιών, συγκεκριμένα έργα παράγονται σε λιγότερο χρόνο συνεπώς ο χρόνος για ψηφιακές δραστηριότητες αυξάνεται όπως και ο χρόνος για χρήση κυβερνητικών πυλών.

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η παρούσα εργασία μελέτησε την εξέλιξη των πέντε διαστάσεων του δείκτη DESI για τα έτη 2014-2019 χρησιμοποιώντας τις αντίστοιχες εκθέσεις DESI 2015 έως DESI 2020. Επίσης οι χώρες-μέλη της ΕΕ ομαδοποιήθηκαν με βάση τις επιδόσεις τους σε αυτές τις πέντε διαστάσεις και μελετήθηκαν παράγοντες που επηρεάζουν τον γενικό δείκτη DESI. Για τις αναλύσεις μας σχετικά με το ΑΕΠ, την ανεργία και τις

ώρες εργασίας χρησιμοποιήθηκαν δεδομένα της Στατιστικής Υπηρεσίας της Ευρωπαϊκής Ένωσης (Eurostat). Από τη μελέτη της εξέλιξης των πέντε διαστάσεων του δείκτη DESI παρατηρήσαμε πως συνολικά υπάρχει σημαντική διαφορά των επιδόσεων των χωρών της ΕΕ στα έτη 2014-2019. Πιο συγκεκριμένα, η μέση διαφορά στη διάσταση **Συνδεσιμότητα** που αφορά στις τεχνολογικές υποδομές, από χρόνο σε χρόνο είναι στατιστικά σημαντική κάτι αναμενόμενο καθώς αυτή η διάσταση θα λέγαμε ότι αποτελεί τη βάση που θα δώσει τον απαραίτητο χώρο και τρόπο για να μπορέσουν να εξελιχθούν και οι άλλες διαστάσεις. Για τις διαστάσεις **Ανθρώπινο Κεφάλαιο** και **Χρήση Υπηρεσιών Διαδικτύου** διαπιστώσαμε πως παρότι υπάρχουν οι τεχνολογικές υποδομές, οι ψηφιακές δεξιότητες και η χρήση υπηρεσιών διαδικτύου καθυστερούν να εμφανίσουν στατιστικά σημαντική διαφορά κάτι που πιθανά να οφείλεται στο ότι χρειάζεται χρόνος ώστε ο πληθυσμός να εξοικειωθεί με τη νέα τεχνολογία και να κατανοήσει πως μπορεί να την εκμεταλλευτεί. Από τη μελέτη μας για τη διάσταση **Ενσωμάτωση Ψηφιακής Τεχνολογίας** που αφορά στο ηλεκτρονικό επιχειρείν και στο ηλεκτρονικό εμπόριο, παρατηρήσαμε στατιστικά σημαντικές διαφορές σχεδόν ανά δεύτερη χρονιά κάτι που μπορεί να αποδοθεί στο ότι για να γίνει ο ψηφιακός μετασχηματισμός των επιχειρήσεων απαιτείται ένα εύλογο χρονικό διάστημα. Τέλος, για τη διάσταση **Ψηφιακές Δημόσιες Υπηρεσίες** παρατηρήσαμε στατιστικά σημαντικές διαφορές όχι μεταξύ όλων των ετών κάτι που πιθανά να οφείλεται στις προτεραιότητες που δίνει κάθε χώρα, που είτε εστιάζει στον εκσυγχρονισμό των εθνικών πυλών της είτε στην πολιτική των ανοιχτών δεδομένων.

Στη συνέχεια, εφαρμόσαμε τη μέθοδο ομαδοποίησης K-means με σκοπό να ομαδοποιήσουμε τις χώρες-μέλη της ΕΕ με βάση τις επιδόσεις τους στις πέντε διαστάσεις του δείκτη DESI. Συμπεραίναμε ότι υπάρχουν δυο ομάδες στις οποίες κατανέμονται οι χώρες-μέλη της ΕΕ με βάση τις επιδόσεις τους στις πέντε διαστάσεις του δείκτη DESI. Έχουμε μια ομάδα που αποτελείται από χώρες με υψηλές επιδόσεις και μια ομάδα με χαμηλές επιδόσεις αντίστοιχα σε όλες τις διαστάσεις του DESI.

Τέλος μελετήσαμε τη σχέση που συνδέει το δείκτη DESI με παράγοντες που σχετίζονται με την ανάπτυξη μιας χώρας όπως το κατά κεφαλήν Ακαθάριστο Εγχώριο Προϊόν (ΑΕΠ), ο μέσος αριθμός εβδομαδιαίων ωρών εργασίας και η ανεργία. Καταλήξαμε στο ότι ο δείκτης DESI πράγματι επηρεάζεται σημαντικά από τους δύο πρώτους παράγοντες.

Ευελπιστούμε η παρούσα εργασία να αποτελέσει ερέθισμα για τη συστηματικότερη μελέτη των ψηφιακών επιδόσεων των χωρών-μελών της ΕΕ τα επόμενα έτη. Για μία διεξοδικότερη μελέτη του δείκτη DESI, θα μπορούσαν στο προτεινόμενο μοντέλο να ενσωματωθούν και άλλοι κοινωνικο-οικονομικοί παράγοντες και να μελετηθούν πιο σύνθετα μοντέλα με αλληλεπιδράσεις.

Η σύγχρονη κοινωνία και οικονομία αλλάζουν ριζικά και οδηγούνται στην ψηφιοποίηση τους καθώς η ανάπτυξη των ΤΠΕ είναι ραγδαία τα τελευταία χρόνια. Αν επιπλέον λάβουμε υπόψη και τις νέες ανάγκες που προέκυψαν από την πανδημία COVID-19 για εργασία, ψυχαγωγία, επικοινωνία, εκπαίδευση και αγορές μέσω

διαδικτύου καταλαβαίνουμε πόσο επιτακτική είναι η ανάγκη της ΕΕ να παρακολουθεί και να ακολουθεί τις εξελίξεις. Αναμένουμε την αποτύπωση της ψηφιακής εικόνας της κοινωνίας και της οικονομίας για το 2020 μέσω της έκθεσης DESI 2021.

ABSTRACT

The rapid development of Information and Communication Technologies (ICT) in recent years, has brought about significant changes in many social sectors such as communication, economy, entertainment and others. The European Union (EU) in order to define the key role that ICT will play in its development course, has developed a composite indicator, the Digital Economy and Society Index (DESI), to assess the digital performance of its member states. In the current work an attempt is made to assess the development of the digital economy and society in the EU by studying the five dimensions of the index DESI for the years 2014-2019, using the corresponding DESI reports (DESI 2015-DESI 2020). Our aim is to study the five dimensions of the DESI index and group the EU member states based on their performance in the five dimensions using well-known clustering techniques. EU member states are classified in two groups, one with High and one with Low performance in the five dimensions of the DESI index. The evolution of each member country and the possible transitions from one group to another during the years 2014-2019 is also a point of interest. The grouping of EU member states into the two groups showed that socio-economic factors may affect the overall DESI index. Mixed effect models confirm the effect of Gross Domestic Product (GDP) per capita and the average number of weekly working hours on the DESI index, which, however, is not verified for unemployment.

ΑΝΑΦΟΡΕΣ

Bánhidi, Z., Dobos, I., & Nemeslaki, A. (2020). What the overall Digital Economy and Society Index reveals: A statistical analysis of the DESI EU28 dimensions. In *Regional Statistics* (pp. 42-62). The Central and Eastern European Online Library.

Bates, Bolker, Mächler, & Walker. (2014). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*.

Burnham, & Anderson. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 261–304.

Castells, M. (2010). *The Rise of the Network Society. Malden: Wiley-Blackwell.*

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, 61(6), 1-36.

Esping-Andersen, G. (1990). *The Three Worlds of Welfare Capitalism.* Cambridge: Polity Press.

European Commission. (2020). *Digital Economy and Society Index (DESI) 2020: Connectivity.* Retrieved from https://digital-strategy.ec.europa.eu/en/library/digital-economy-and-society-index-desi-2020

European Commission. (2020). *Digital Economy and Society Index (DESI) 2020: Digital public services.* Retrieved from https://digital-strategy.ec.europa.eu/en/library/digital-economy-and-society-index-desi-2020

European Commission. (2020). *Digital Economy and Society Index (DESI) 2020: Human capital.* Retrieved from https://digital-strategy.ec.europa.eu/en/library/digital-economy-and-society-index-desi-2020

European Commission. (2020). *Digital Economy and Society Index (DESI) 2020: Integration of digital technology.* Retrieved from https://digital-strategy.ec.europa.eu/en/library/digital-economy-and-society-index-desi-2020

European Commission. (2020). *Digital Economy and Society Index (DESI) 2020: Use of internet services.* Retrieved from https://digital-strategy.ec.europa.eu/en/library/digital-economy-and-society-index-desi-2020

European Commission. (2020). *Digital Economy and Society Index 2020: Methodological Note.* Retrieved from https://digital-strategy.ec.europa.eu/en/library/digital-economy-and-society-index-desi-2020

G20. (2018). *Toolkit for measuring the digital economy.* Argentina: G20. Retrieved from https://www.oecd.org/g20/summits/buenos-aires/G20-Toolkit-for-measuring-digital-economy.pdf

Huseyin, S. (2021). Analysis of the digital economy and society index (DESI) through a cluster analysis. *Journal of Social Science*, 37-51.

Kammer, A., Niehues, J., & Peichl, A. (2012). Welfare regimes and welfare state outcomes in Europe. *Journal of European Social Policy*.

Lauzadyte-Tutliene, A., Balezentis, T., & Goculenko, E. (2018). Welfare State in Central and Eastern Europe. *Economics and Sociology*, 100-123.

Parviainen, Tihinen, Kääriäinen, & Teppola. (2017). Tackling the digitalization challenge: how to benefit f rom digitalization in practice. *International Journal of Information Systems and Project Management*, 36-67.

Salkind. (2010). Repeated Measures Design. *SAGE Research Methods.* .

Shapiro, & Wilk. (1965, December). An analysis of variance test for normality (complete samples). *Biometrika*, pp. 591–611.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 33ου Πανελληνίου Συνεδρίου Στατιστικής (2021), σελ.178-194

ΚΙΝΗΤΟΣ ΔΙΑΜΕΣΟΣ ΕΝΑΝΤΙ ΚΙΝΗΤΟΥ ΜΕΣΟΥ, ΘΕΣΜΙΚΟΙ ΕΠΕΝΔΥΤΕΣ ΕΝΑΝΤΙ ΜΙΚΡΟΕΠΕΝΔΥΤΩΝ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΙΚΟΤΗΤΑ ΚΕΦΑΛΑΙΑΓΟΡΩΝ

Μηλιώνης Ε. Αλέξανδρος [1], Βαρλάγκας Θ. Βασίλειος [2]
[1] Τράπεζα της Ελλάδος και Πανεπιστήμιο Αιγαίου, Τμήμα Στατιστικής και Αναλογιστικών – Χρηματοοικονομικών Μαθηματικών
amilionis@aegean.gr
[2] Υπουργείο Εθνικής Άμυνας, Πολεμικό Ναυτικό και Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, Τμήμα Οικονομικών Επιστημών
billbar@econ.uoa.gr

ΠΕΡΙΛΗΨΗ

Ένας τυπικός τρόπος ελέγχου της αποτελεσματικότητας κεφαλαιαγορών είναι η σύγκριση της απόδοσης τεχνικών συναλλακτικών κανόνων έναντι της στρατηγικής της αγοράς και διακράτησης. Στο παρόν προτείνεται ένας εναλλακτικός τεχνικός κανόνας που στηρίζεται στον κινητό διάμεσο και συγκρίνεται η προβλεπτική του ικανότητα στο Χρηματιστήριο Αξιών Αθηνών με αυτή του δημοφιλέστερου τεχνικού κανόνα του κινητού μέσου, με και χωρίς κόστη συναλλαγών. Από τα εμπειρικά ευρήματα προκύπτει ότι σε κάθε περίπτωση η προβλεπτική ικανότητα του κινητού διάμεσου είναι υψηλότερη από αυτή του κινητού μέσου. Χωρίς κόστη συναλλαγών και οι δύο κανόνες αποδίδουν καλύτερα της παθητικής στρατηγικής, συνεπώς η υπόθεση των αποτελεσματικών αγορών απορρίπτεται. Εισάγοντας όμως συναλλακτικά κόστη, ωσάν τα ισχύοντα στην συγκεκριμένη αγορά, προκύπτει ότι για ένα θεσμικό επενδυτή είναι δυνατό, έστω και οριακά, να υπερνικήσει την αγορά, αλλά κάτι τέτοιο δεν ισχύει για έναν τυπικό μικροεπενδυτή, λόγω υψηλότερου συναλλακτικού κόστους για τον τελευταίο. Συνεπώς, δεδομένου του κόστους συναλλαγών η απόρριψη ή μη της υπόθεσης των αποτελεσματικών κεφαλαιαγορών εξαρτάται από την ταυτότητα του επενδυτή.

Λέξεις Κλειδιά: Τεχνική Ανάλυση, Αποτελεσματικότητα Αγορών, Κινητός Μέσος, Κινητός Διάμεσος, Χρηματιστήριο Αθηνών

1. ΕΙΣΑΓΩΓΗ

Η δυνατότητα πρόβλεψης των μελλοντικών τιμών των μετοχών από τις παρελθούσες και τρέχουσες πληροφορίες, πιο φορμαλιστικά η υπόθεση των αποτελεσματικών αγορών, είναι ένα από τα πιο σημαντικά αντικείμενα της σύγχρονης χρηματοοικονομικής θεωρίας, τόσο για τη θεωρητική του αξία, όσο και για τις επιπτώσεις του στις επενδύσεις. Αν και η αποτελεσματικότητα των αγορών

ορίζεται διαφορετικά από διάφορους συγγραφείς (Black, 1986, Malkiel, 1992, Milionis, 2007), ο ορισμός που έχει καθιερωθεί οφείλεται στον Fama (1970). Σύμφωνα με αυτόν τον ορισμό, μια αγορά είναι αποτελεσματική εάν «οι τιμές "αντανακλούν" πλήρως όλες τις διαθέσιμες πληροφορίες». Η κλασσική κατηγοριοποίηση, κατατάσσει την αποτελεσματικότητα ως ασθενούς μορφής, όταν το σύνολο των πληροφοριών περιλαμβάνει τις τιμές του παρελθόντος, μέσης-ισχύος, όταν το σύνολο των πληροφοριών περιλαμβάνει όλες τις δημόσια διαθέσιμες πληροφορίες και ισχυρής μορφής, όταν το σύνολο των πληροφοριών περιλαμβάνει όλες τις δημόσια ή ιδιωτικά διαθέσιμες πληροφορίες. Στους γνωστούς ελέγχους για την προβλεψιμότητα των αποδόσεων (Fama, 1991) το διαθέσιμο πληροφοριακό σύνολο, εκτός από τις παρελθούσες τιμές, μπορεί επίσης να περιλαμβάνει ειδικά χαρακτηριστικά της επιχείρησης. Σε μια αποτελεσματική αγορά, τα αποτελέσματα των ελέγχων για την προβλεψιμότητα των αποδόσεων δεν θα πρέπει να απορρίπτουν τη μηδενική υπόθεση, στην οποία οι αποδόσεις είναι μη προβλέψιμες.

Μέχρι τις αρχές της δεκαετίας του 1990, το γενικό συμπέρασμα που ανέκυπτε από τα αποτελέσματα των περισσότερων εμπειρικών ελέγχων για την αποτελεσματικότητα της αγοράς ήταν ότι, με λίγες εξαιρέσεις, η υπόθεση των αποτελεσματικών κεφαλαιαγορών δεν απορρίπτεται, τουλάχιστον στην ασθενή και μέσης-ισχύος μορφή της (Fama, 1970, Fama, 1991, Elton and Gruber, 1995). Ωστόσο, στις πιο πρόσφατες ερευνητικές εργασίες η υπόθεση των αποτελεσματικών αγορών απορρίπτεται συχνά, ακόμη και στην ασθενή της μορφή. Για τον εμπειρικό έλεγχο της αποτελεσματικής αγοράς ασθενούς ισχύος (Weak-Form Market Efficiency (WFME) έχει χρησιμοποιηθεί μια σειρά από μεθοδολογικές προσεγγίσεις (Fama, 1970, Fama, 1991), οι οποίες μπορούν να ταξινομηθούν σε δύο κύριες κατηγορίες: (α) αμιγώς στατιστικοί-οικονομετρικοί έλεγχοι και (β) έλεγχοι που βασίζονται σε συναλλακτικούς κανόνες της τεχνικής ανάλυσης.

Στην πρώτη κατηγορία η αποτελεσματικότητα της αγοράς θα πρέπει απαραίτητα να ελέγχεται σε συνδυασμό με ένα υπόδειγμα αποτίμησης περιουσιακών στοιχείων, το οποίο θεωρούμε ότι παράγει τις υπό συνθήκη προσδοκίες των αποδόσεων των περιουσιακών στοιχείων (πρόβλημα κοινής υπόθεσης). Στη δεύτερη κατηγορία οι αποδόσεις που προκύπτουν από την εφαρμογή συναλλακτικών κανόνων συγκρίνονται άμεσα με τις αντίστοιχες αποδόσεις της παθητικής στρατηγικής (αγορά και διακράτηση). Με τον τρόπο αυτό η υπόθεση της αποτελεσματικής αγοράς ασθενούς ισχύος εξαρτάται λιγότερο από ένα υπόδειγμα αποτίμησης, καθώς η μόνη υπόθεση που υιοθετείται είναι ότι οι τιμές ακολουθούν μια διαδικασία submartingale (δηλαδή $E(R_{t+1}|\Phi_t) \geq 0$ όπου E είναι ο τελεστής της αναμενόμενης τιμής και $E(R_{t+1}|\Phi_t)$ είναι η αναμενόμενη απόδοση τη χρονική στιγμή t+1 δοθέντων των διαθέσιμων πληροφοριών μέχρι τη χρονική στιγμή t (Φ_t)). Αν και οι πρώτες εργασίες για τον έλεγχο της αποδοτικότητας με τη χρήση τεχνικών συναλλακτικών κανόνων έδειξαν μη απόρριψη της WFME (Cowles, 1934, Fama and Blume, 1966) υπήρξε μια αναζωπύρωση της έρευνας μετά την σημαίνουσα εργασία των Brock et al. (1992), στην οποία τεκμηριώθηκε η προβλεπτική ικανότητα των τεχνικών συναλλακτικών κανόνων.

Μεταξύ των κανόνων της τεχνικής ανάλυσης, οι οποίοι είναι μαθηματικά καλά ορισμένοι κατά την έννοια του Neftci (1991), αυτός που χρησιμοποιείται συχνότερα από τους ερευνητές για τον έλεγχο της αποτελεσματικότητας της αγοράς είναι ο κινητός μέσος (Moving Average (MA)). Πράγματι, ο κανόνας MA έχει χρησιμοποιηθεί εκτενώς από πολλούς ερευνητές και για πολλές αγορές κεφαλαίου και συναλλάγματος (π.χ. Brock et al., 1992, Hudson et al., 1996, Kwon and Kish, 2002, Olson, 2004, Cai et al., 2005).

Η συνηθέστερη εκδοχή του χρησιμοποιεί δύο κινητούς μέσους με διαφορετικό μήκος, οι οποίοι υπολογίζονται από την χρονοσειρά των τιμών ενός αξιογράφου ή ενός δείκτη:

$$MAS_t = \left(\frac{1}{M}\sum_{i=0}^{M-1}\theta_i B^i P_t\right)$$

$$MAL_t = \left(\frac{1}{N}\sum_{i=0}^{N-1}\theta_i B^i P_t\right) \quad \text{με } N > M,$$

όπου ο MAS_t αντιπροσωπεύει τον βραχύ MA με μήκος M υπολογιζόμενος στο χρόνο t και ο MAL_t αντιπροσωπεύει τον εκτενέστερο MA με μήκος N. P_t είναι η τιμή του αξιογράφου ή του δείκτη στο χρόνο t, θ_i είναι μη χρονικοί παράμετροι και B είναι ο τελεστής χρονικής υστέρησης. Τα σήματα αγοράς δημιουργούνται στους χρόνους τ_j^B, όπου:

$$\tau_j^B \equiv inf\{t: t > \tau_j^B, MAL_t - MAS_t > DP_{t-1}\}$$

και τα σήματα πώλησης δημιουργούνται στους χρόνους τ_j^S, όπου:

$$\tau_j^S \equiv inf\{t: t > \tau_j^S, MAS_t - MAL_t > DP_{t-1}\}$$

Οι αρχικοί χρόνοι τ_0^B και τ_0^S ορίζονται ίσοι με μηδέν και το D είναι το γνωστό εύρος ζώνης (μια προκαθορισμένη μη μηδενική σταθερά).

Σύμφωνα με την τεχνική ανάλυση, η απόδοση του συναλλακτικού κανόνα MA βελτιώνεται αν συνδυαστεί με άλλους δείκτες (Murphy, 1999). Ωστόσο, αν εστιάσουμε στον πυρήνα του κανόνα MA της τεχνικής ανάλυσης γίνεται εύκολα αντιληπτό ότι χρησιμοποιείται ο μέσος ως μια στατιστική της κεντρικής τάσης. Αναμφισβήτητα, ο μέσος έχει συγκεκριμένα πλεονεκτήματα συγκριτικά με άλλα μέτρα κεντρικής τάσης. Για παράδειγμα, για συμμετρικούς πληθυσμούς τόσο ο δειγματικός μέσος όσο και ο δειγματικός διάμεσος είναι συνεπείς εκτιμητές του πληθυσμιακού μέσου, αλλά ο δειγματικός μέσος είναι ασυμπτωτικά ο βέλτιστος εκτιμητής του πληθυσμιακού μέσου. Από την άλλη βέβαια, ο διάμεσος επηρεάζεται λιγότερο από τις ακραίες τιμές συγκριτικά με το μέσο. Ως εκ τούτου, θεωρούμε ότι είναι ενδιαφέρον να τροποποιήσουμε τον εν λόγω συναλλακτικό κανόνα, αντικαθιστώντας τον κινητό μέσο με τον κινητό διάμεσο και καθώς ο τελευταίος επηρεάζεται λιγότερο από τις ακραίες τιμές πιθανολογούμε ότι βελτιώνεται η προβλεπτική του ικανότητα.

Αναφορικά με την επιλογή του εύρους των δύο ΜΑ του κανόνα, στις περισσότερες δημοσιευμένες μελέτες χρησιμοποιούνται συγκεκριμένοι συνδυασμοί του βραχέως και του εκτενέστερου κινητού μέσου (πχ. Brock et al., 1992, Hudson et al., 1996, Mills, 1997). Οι συνδυασμοί που συνήθως επιλέγονται είναι αυτοί που χρησιμοποιούνται περισσότερο από τους αναλυτές της αγοράς και η επιλογή τους είναι, τουλάχιστον σε κάποιο βαθμό, αυθαίρετη.

Σε αυτή την εργασία, αρχικά εισάγουμε έναν εναλλακτικό συναλλακτικό κανόνα, ως μια τροποποίηση του τεχνικού κανόνα ΜΑ, ο οποίος υπολογίζει τον κινητό διάμεσο αντί του κινητού μέσου και χρησιμοποιούμε γι' αυτόν τον όρο "κινητός διάμεσος" (εφεξής ΜΜ). Εν συνεχεία, εκτιμάμε την προβλεπτική ικανότητα τόσο του κινητού διάμεσου όσο και του κινητού μέσου και συγκρίνουμε τις αποδόσεις τους, υπολογίζοντας ταυτόχρονα όλους τους συνδυασμούς του εύρους τους. Τέλος, λαμβάνουμε υπόψη τα πραγματικά συναλλακτικά κόστη που ισχύουν στο Χρηματιστήριο Αξιών Αθηνών και παρουσιάζουμε τις επιπτώσεις που επιφέρουν, αφενός στην απόδοση των υπό εξέταση συναλλακτικών κανόνων και αφετέρου στον έλεγχο της υπόθεσης της αποτελεσματικής κεφαλαιαγοράς ασθενούς ισχύος.

2. ΔΕΔΟΜΕΝΑ ΚΑΙ ΜΕΘΟΔΟΛΟΓΙΑ

Στην παρούσα μελέτη χρησιμοποιήθηκαν οι ημερήσιες τιμές στο κλείσιμο από το Γενικό Δείκτη (εφεξής ΓΔ) του Χρηματιστηρίου Αξιών Αθηνών (εφεξής ΧΑΑ) για την περίοδο 02 Ιουνίου 2014 έως και 31 Ιουλίου 2020. Το Χρηματιστήριο Αθηνών αν και εντάχθηκε το 2001 στη χορεία των αναπτυγμένων αγορών, στις 21 Μαρτίου 2016 υποβαθμίστηκε στην κατηγορία των προηγμένων αναπτυσσόμενων αγορών από τον οίκο αξιολόγησης FTSE. Η εν λόγω υποβάθμιση αποδόθηκε τόσο στον επιβεβλημένο περιορισμό κεφαλαίων του 2015 και στην παύση λειτουργίας του ΧΑΑ για πέντε εβδομάδες το καλοκαίρι του ίδιου έτους, όσο και στην αδιάκοπη αστάθεια της ελληνικής οικονομίας. Στο Σχήμα 1 φαίνεται η πορεία του ΓΔ στο συγκεκριμένο χρονικό διάστημα.

Στη συνέχεια εξετάστηκε και μια υποπερίοδος περίπου 4 ετών, από 24 Φεβρουαρίου 2016 έως και 16 Μαρτίου 2020, κατά την οποία η τιμή του ΓΔ έκλεισε πολύ κοντά στην τιμή εκκίνησής του. Η επιλογή αυτής της υποπεριόδου αποσκοπεί στην αποσύνδεση των αποτελεσμάτων από συγκεκριμένη πορεία του ΓΔ και πιθανής μεροληψίας (return metric bias) στην αξιολόγηση της απόδοσης των τεχνικών κανόνων. Συνεπώς, μελετήθηκε μια περίοδος έξι ετών κατά την οποία ο ΓΔ στη λήξη σημείωσε σημαντικές απώλειες και μια υποπερίοδος τεσσάρων ετών κατά την οποία η τιμή του ΓΔ στη λήξη παρέμεινε στο αρχικό της επίπεδο, παρά τις ενδιάμεσες διακυμάνσεις.

Σχήμα 1. *Διάγραμμα Γενικού Δείκτη του Χρηματιστηρίου Αξιών Αθηνών*

Η επιλογή της συγκεκριμένης περιόδου των έξι ετών, μικρότερης συγκριτικά με παρόμοιες μελέτες, αφενός θεωρείται ως μια πρώτη προσέγγιση του όλου εγχειρήματος και αφετέρου εξυπηρετεί την παραδοχή ότι κατά τη διάρκεια που ο τεχνικός κανόνας επιτάσσει την παραμονή του ρευστοποιημένου κεφαλαίου εκτός αγοράς, αυτό είναι άμεσα διαθέσιμο σε έναν τρεχούμενο τραπεζικό λογαριασμό χωρίς προσθήκη τόκων. Υιοθετήθηκε αυτή η θεώρηση διότι ο δικαιούμενος τόκος είναι αμελητέος και δεν επηρεάζει τα παραγόμενα αποτελέσματα, λόγω των σχεδόν μηδενικών επιτοκίων που επικρατούσαν στην υπό μελέτη περίοδο (2014-2020).

Αρχικά, υιοθετώντας τη μεθοδολογία των Milionis and Papanagiotou (2011), υπολογίζονται οι αποδόσεις που προκύπτουν από την εφαρμογή του τεχνικού κανόνα MA για όλους τους συνδυασμούς του βραχέως (MAS) και του εκτενέστερου κινητού μέσου (MAL), προκειμένου να λάβουμε υπόψη την ευαισθησία των αποτελεσμάτων από τις μεταβολές στο μήκος του κινητού μέσου. Στη συνέχεια, υπολογίζονται οι αποδόσεις του συναλλακτικού κανόνα MM με την ίδια μεθοδολογία που χρησιμοποιήθηκε στον κινητό μέσο. Συγκεκριμένα, ο βραχύς κινητός διάμεσος (MMS) ορίζεται σταθερός με μήκος ίσο με ένα, ενώ το μήκος του εκτενέστερου κινητού διάμεσου (MML) μεταβάλλεται από 5 έως 100 με βήμα ένα. Στις εξισώσεις υπολογισμού του κινητού μέσου και κινητού διάμεσου θέτουμε όλες τις παραμέτρους θ_i ίσες με τη μονάδα και το εύρος ζώνης D ίσο με μηδέν. Οι ποσοστιαίες αποδόσεις που ανακύπτουν σε κάθε εξεταζόμενη περίοδο, από τις εν λόγω τεχνικές, για κάθε μήκος του εκτενέστερου κινητού διάμεσου ή μέσου δημιουργούν μια σειρά.

Προκειμένου να δοθεί στην παρούσα μελέτη μια σφαιρική αλλά και πρακτική προσέγγιση, οι αποδόσεις από την εφαρμογή των εν λόγω τεχνικών συναλλακτικών κανόνων υπολογίστηκαν λαμβάνοντας υπόψη και τα πραγματικά συναλλακτικά κόστη που ενυπάρχουν στην πραγματική οικονομία και συγκεκριμένα στη δευτερογενή ελληνική κεφαλαιαγορά. Σε όλες τις οικονομικές οντότητες που συναλλάσσονται στο ΧΑΑ επιβάλλονται σταθερά και μεταβλητά κόστη επί της αξίας των συναλλαγών, καθώς και φορολογία. Ειδικότερα, οι εν λόγω επιβαρύνσεις περιλαμβάνουν έξοδα εκτέλεσης συναλλαγών, διακανονισμού, εκκαθάρισης, λοιπές χρεώσεις του Χρηματιστηρίου και φόρο επί των πωλήσεων, των οποίων οι ποσοστιαίες και σταθερές χρεώσεις είναι ίδιες για όλους τους συναλλασσόμενους. Όμως, πέραν των ανωτέρω χρεώσεων οι συναλλασσόμενοι στο ΧΑΑ επιβαρύνονται και με προμήθεια επί των συναλλαγών, το ύψος της οποίας κυμαίνεται από 0% έως 1% και εξαρτάται από την ιδιότητα του εντολέα της συναλλαγής. Συνεπώς, εξετάστηκαν πέντε σενάρια στις αποδόσεις κάθε συναλλακτικού κανόνα. Στο πρώτο

σενάριο αποκρυσταλλώνεται η θεωρητική προσέγγιση, όπου οι συναλλαγές δεν επιβαρύνονται με οποιοδήποτε κόστος και φορολογία. Στο δεύτερο σενάριο υπολογίζονται οι αποδόσεις για ένα μέλος του ΧΑΑ, το οποίο θα εφαρμόσει την εν λόγω τεχνική σε μέρος του κεφαλαίου του. Τα μέλη του χρηματιστηρίου, έχουν μηδενική προμήθεια επί των συναλλαγών καθώς διενεργούν τις συναλλαγές για ίδιον όφελος. Το τρίτο σενάριο αντιπροσωπεύει τους θεσμικούς επενδυτές, για τους οποίους η προμήθεια επί των συναλλαγών καθορίζεται βάσει ιδιωτικών συμφωνιών και εκτιμάται ότι κυμαίνεται σε ποσοστά από 0,1% μέχρι περίπου 0,2%. Στο τέταρτο σενάριο αντιπροσωπεύεται ένας επαγγελματίας ιδιώτης επενδυτής, ο οποίος αντιμετωπίζει μειωμένη προμήθεια συναλλαγών, ήτοι ενδεικτικά 0,5%. Το πέμπτο και τελευταίο σενάριο αντιπροσωπεύει τον τυπικό ιδιώτη μικροεπενδυτή, ο οποίος υπόκειται στη μέγιστη χρέωση επί της προμήθειας ανά συναλλαγή, η οποία ανέρχεται μέχρι και το 1%. Όλα τα αναφερόμενα κόστη συναλλαγών περιγράφονται αναλυτικά στον ακόλουθο Πίνακα 1.

Πίνακας 1. *Πάσης φύσεως συναλλακτικά έξοδα εκτελούντων συναλλαγές στο Χρηματιστήριο Αξιών Αθηνών, βάσει της ΠΟΛ.1056/28.3.2011.*

Προμήθεια συναλλαγής	Μέγιστο 1,00% επί της συναλλαγής
Έξοδα εκτέλεσης συναλλαγών	0,0125% επί της αξίας συναλλαγής, πλέον 0,06€ ανά εντολή
Έξοδα διακανονισμού, εκκαθάρισης και λοιπές χρεώσεις	0,06% επί της αξίας συναλλαγής, πλέον 0,75€ ανά κινητή αξία
Φόρος πώλησης	0,20% επί της αξίας συναλλαγής

Ακολούθως, εξετάζεται η πιθανή ύπαρξη μοναδιαίας ρίζας στις παραγόμενες σειρές των αποδόσεων κάθε σεναρίου, ώστε να εξακριβωθεί αν οι σειρές είναι στάσιμες. Σε περίπτωση ύπαρξης στασιμότητας η σειρά θα κυμαίνεται έχοντας ως σημείο αναφοράς ένα μέσο επίπεδο, βάσει μιας σταθερής διακύμανσης. Συνεπώς, μπορεί να χρησιμοποιηθεί το ύψος του μέσου από τον εκάστοτε συναλλακτικό κανόνα, ως το ύψος της αναμενόμενης απόδοσής του και στη συνέχεια να συγκριθεί με το ύψος της απόδοσης από την στρατηγική της αγοράς και διακράτησης. Προκειμένου αυτή η σύγκριση να καταστεί αξιόπιστη πραγματοποιείται ένας έλεγχος σημαντικότητας σε διάστημα εμπιστοσύνης 95%. Ειδικότερα, εκτιμάται το διάστημα εμπιστοσύνης γύρω από την αναμενόμενη απόδοση του συναλλακτικού κανόνα και εξετάζεται αν το ύψος της απόδοσης από την παθητική στρατηγική βρίσκεται μέσα ή έξω από τα όρια αυτού του διαστήματος εμπιστοσύνης. Όπως και στη μελέτη των Milionis and Papanagiotou (2011) η εν λόγω σύγκριση απολήγει σε τρείς δυνατές καταστάσεις, στις οποίες η απόδοση του εκάστοτε συναλλακτικού κανόνα είναι είτε υψηλότερη είτε χαμηλότερη είτε δε διαφέρει στατιστικά από την απόδοση της παθητικής στρατηγικής. Στις περιπτώσεις που η σειρά των αποδόσεων του συναλλακτικού κανόνα έχει μοναδιαία ρίζα, αυστηρά δεν καθίσταται εφικτή μια τέτοια σύγκριση, διότι η διακύμανση της σειράς δεν είναι σταθερή και διαγράφει μεγάλες περιπλανήσεις, επομένως η σειρά δεν έχει σταθερό επίπεδο αναφοράς.

Ωστόσο, για τον καθορισμό του διαστήματος εμπιστοσύνης γύρω από την εν λόγω αναμενόμενη απόδοση, στις περιπτώσεις όπου η σειρά των αποδόσεων είναι στάσιμη, απαιτείται ο υπολογισμός της διακύμανσης. Όμως, στην προκειμένη περίπτωση η διακύμανση δεν μπορεί να εκτιμηθεί με το γνωστό θεώρημα του δειγματικού μέσου για τυχαία δείγματα, διότι οι αποδόσεις της σειράς εμφανίζουν μεταξύ τους ισχυρές συσχετίσεις και δεν αποτελούν τυχαίο δείγμα. Προκειμένου να ξεπεραστεί το ανωτέρω πρόβλημα υιοθετήθηκε το επαυξημένο θεώρημα του δειγματικού μέσου (Augmented Sample Mean Theorem, ASMT) για τον υπολογισμό της διακύμανσης, στο οποίο λαμβάνονται υπόψη και οι περιπτώσεις ύπαρξης γραμμικών αλληλεξαρτήσεων μεταξύ των δειγματικών παρατηρήσεων (λεπτομέρειες δίνονται στους Milionis and Papanagiotou, 2013). Αν οι τυχαίες μεταβλητές X_1, X_2, …, X_N εμφανίζουν γραμμικές αλληλεξαρτήσεις, τότε αποτελούν ένα μη τυχαίο δείγμα μεγέθους N, από ένα πληθυσμό με μέσο μ και διακύμανση σ^2. Επιπρόσθετα, αν ρ_κ με κ = 1, 2, … συμβολίζει την συνάρτηση αυτοσυσχέτισης για τις X_1, X_2, …, X_N, τότε ο δειγματικός μέσος $\bar{X}$ είναι ένας αμερόληπτος εκτιμητής του μέσου μ και η διακύμανση του δειγματικού μέσου δίνεται από τον τύπο:

$$VAR(\bar{X}) = \frac{\sigma^2}{N}\left(1 + 2\frac{(N-1)}{N}\rho_1 + 2\frac{(N-2)}{N}\rho_2 + \cdots + \frac{2}{N}\rho_{N-1}\right).$$

Αναφορικά με τον έλεγχο για τυχόν ύπαρξη μοναδιαίας ρίζας στις σειρές των αποδόσεων ακολουθήθηκε η μεθοδολογία των Milionis and Papanagiotou (2008). Αρχικά, χρησιμοποιήθηκε ο πιο κοινός και δημοφιλής έλεγχος, αυτός του Augmented Dickey-Fuller (ADF), ενώ για να ενδυναμωθεί η αξιοπιστία των αποτελεσμάτων διενεργήθηκε επιπρόσθετα η τροποποίηση του ανωτέρω ελέγχου από τους Elliot et al. (1996), ο γνωστός έλεγχος ERS. Να σημειωθεί ότι ο υπολογισμός του διαστήματος εμπιστοσύνης σε κάθε συνάρτηση αυτοσυσχέτισης πραγματοποιήθηκε με την προσέγγιση του Bartlett (Bartlett, 1946) για την εκτίμηση της διακύμανσης, η οποία δίνεται από τον τύπο:

$$\text{VAR}(\rho_k) = \frac{1}{N}\left(1 + 2\sum_{i=1}^{m}\rho_i^2\right)$$

όπου k > m, N είναι το πλήθος της σειράς και ρ_k είναι ο συντελεστής αυτοσυσχέτισης για υστέρηση k.

Η ανωτέρω μεθοδολογία εφαρμόστηκε για κάθε σενάριο και στις δύο περιόδους, τόσο με την τεχνική του κινητού διάμεσου όσο και με την τεχνική του κινητού μέσου. Απώτερος στόχος είναι αφενός ο έλεγχος της προβλεπτικής ικανότητας του κινητού διάμεσου και η σύγκρισή του με τον ευρέως διαδεδομένο κανόνα του κινητού μέσου· αφετέρου η ανάδειξη τυχόν διαφοροποίησης ανάμεσα στα εξεταζόμενα σενάρια στην επίτευξη ή μη υπερνίκησης της αγοράς από την εφαρμογή

***Πίνακας 2**. Αποτελέσματα ελέγχου μοναδιαίας ρίζας και τα εκτιμώμενα στοχαστικά μοντέλα ARIMA (p, d, q).*

Σενάρια	Κινητός Διάμεσος		Κινητός Μέσος	
	Έλεγχος Στασιμότητας	Στοχαστικό Μοντέλο	Έλεγχος Στασιμότητας	Στοχαστικό Μοντέλο
Περίοδος : 2014 - 2020				
Χωρίς Κόστη & Φόρους	Στάσιμη σειρά	$R_L = 0,47R_{L-1} + 0,77R_{L-2} - 0,39R_{L-3} + 0,63\varepsilon_{L-1} + \varepsilon_L$	Στάσιμη σειρά (ADF 90%)	$R_L = 0,39R_{L-1} + 0,47R_{L-2} + 0,80\varepsilon_{L-1} + \varepsilon_L$
Μέλος Χρηματιστη-ρίου	Στάσιμη σειρά	$R_L = -29,08 + 1,28R_{L-1} - 0,38R_{L-3} - 0,55\varepsilon_{L-1} + \varepsilon_L$	Στάσιμη σειρά	$R_L = -27,76 + 1,7R_{L-1} - 0,74R_{L-2} - 0,79\varepsilon_{L-1} + 0,20\varepsilon_{L-3} + \varepsilon_L$
Επαγγελμα-τίας Επενδυτής	Στάσιμη σειρά με τάση	$R_L = -82,29 + 0,37L + 0,90R_{L-1} + \varepsilon_L$	Στάσιμη σειρά	$R_L = -57,49 + 1,65R_{L-1} - 0,67R_{L-2} - 0,83\varepsilon_{L-1} + 0,37\varepsilon_{L-3} + \varepsilon_L$
Ερασιτέχνης Επενδυτής	Στάσιμη σειρά με τάση	$R_L = -96,63 + 0,35L + 0,87R_{L-1} + \varepsilon_L$	Στάσιμη σειρά	$R_L = -70,25 + 2,2R_{L-1} - 1,7R_{L-2} + 0,48R_{L-3} - 1,51\varepsilon_{L-1} + 0,86\varepsilon_{L-2} + \varepsilon_L$
Θεσμικός Επενδυτής	Στάσιμη σειρά	$R_L = -43,68 + 0,96R_{L-1} - 0,10R_{L-7} + \varepsilon_L$	Στάσιμη σειρά	$R_L = -40,42 + 0,96R_{L-1} + \varepsilon_L$
Περίοδος : 2016 - 2020				
Χωρίς Κόστη & Φόρους	Στάσιμη σειρά	$R_L = 51,65 + 0,92R_{L-1} + \varepsilon_L$	Στάσιμη σειρά (ADF 90%)	$R_L = 53,29 + 0,93R_{L-1} + 0,33\varepsilon_{L-8} + \varepsilon_L$
Μέλος Χρηματιστη-ρίου	Τυχαίος περίπατος με εκτροπή	$R_L = 27,75 + 0,92R_{L-1} + \varepsilon_L$	Στάσιμη σειρά	$R_L = 26,29 + 1,97R_{L-1} -0,98R_{L-2} - 1,14\varepsilon_{L-1} + 0,19\varepsilon_{L-3} + \varepsilon_L$
Επαγγελμα-τίας Επενδυτής	Στάσιμη σειρά	$R_L = 0,89R_{L-1} + \varepsilon_L$	Στάσιμη σειρά με τάση	$R_L = -62,19 +1,01L + 0,95R_{L-1} + \varepsilon_L$
Ερασιτέχνης Επενδυτής	Στάσιμη σειρά	$R_L = 0,95R_{L-1} + \varepsilon_L$	Στάσιμη σειρά	$R_L = 1,93R_{L-1} - 0,94R_{L-2} - 0,84\varepsilon_{L-1} + \varepsilon_L$
Θεσμικός Επενδυτής	Στάσιμη σειρά με τάση (ERS 90%)	$R_L = 0,33L + 0,92R_{L-1} + \varepsilon_L$	Στάσιμη σειρά με τάση	$RL = 0,43L + 0,95R_{L-1} + \varepsilon_L$

των εν λόγω τεχνικών κανόνων και οι απορρέουσες επιπτώσεις στον έλεγχο της αποτελεσματικής αγοράς ασθενούς ισχύος.

3. ΑΠΟΤΕΛΕΣΜΑΤΑ ΚΑΙ ΣΧΟΛΙΑΣΜΟΣ

Στον Πίνακα 2 παρουσιάζονται τα αποτελέσματα των ελέγχων για τυχόν ύπαρξη μοναδιαίας ρίζας στις παραγόμενες σειρές των αποδόσεων, καθώς και το αντίστοιχο εκτιμώμενο στοχαστικό μοντέλο ARIMA (p, d, q). Σε κάποιες περιπτώσεις απαιτήθηκε η προσθήκη γραμμικής τάσης. Σύμφωνα με τα αποτελέσματα όλες οι σειρές είναι στάσιμες, πλην της περίπτωσης του μέλους του χρηματιστηρίου με τον κανόνα του κινητού διάμεσου στην υποπερίοδο 2016-2020. Καθώς είναι γνωστό οι έλεγχοι ADF και ERS έχουν βέβαια χαμηλή ισχύ, όμως στην περίπτωση που βρέθηκε μοναδιαία ρίζα η σύγκριση της απόδοσης του συναλλακτικού κανόνα με την παθητική στρατηγική διενεργείται με επιφύλαξη. Στις υπόλοιπες, στάσιμες, περιπτώσεις η εν λόγω σύγκριση διενεργείται σε διάστημα εμπιστοσύνης 95%.

Στη συνέχεια εξετάζεται το ύψος της αναμενόμενης απόδοσης και η τυπική απόκλιση των εν λόγω συναλλακτικών κανόνων σε κάθε σενάριο στις δύο περιόδους. Βάσει των αποτελεσμάτων, τα οποία αποτυπώνονται στον Πίνακα 3, προκύπτουν ενδιαφέροντα συμπεράσματα. Απουσία φορολογίας και πάσης φύσεως εξόδων επί των συναλλαγών, ο κανόνας του κινητού διάμεσου παρουσιάζει υψηλότερη αναμενόμενη απόδοση συγκριτικά με τον κανόνα του κινητού μέσου. Ειδικότερα, στην περίοδο 2014-2020 η αναμενόμενη απόδοση του κινητού διάμεσου είναι υψηλότερη κατά 2,12%, ενώ στην περίοδο 2016-2020 η διαφορά αυτή αμβλύνεται και διαμορφώνεται στο 0,1%. Αντίθετα, στα υπόλοιπα σενάρια, όπου υπεισέρχονται κόστη συναλλαγών και φορολογία, οι υψηλότερες αποδόσεις του κινητού διάμεσου εξαϋλώνονται. Σε αυτά τα σενάρια ο κανόνας του κινητού διάμεσου διαγράφει χαμηλότερες αποδόσεις συγκριτικά με τον κινητό μέσο, ενώ η εν λόγω διαφορά αυξάνεται αναλογικά με την αύξηση της προμήθειας επί των συναλλαγών. Η διαφορά μεταξύ των αποδόσεων των δύο συναλλακτικών κανόνων, στα συγκεκριμένα σενάρια, είναι εντονότερη στην περίοδο 2016-2020.

Η αύξηση στο ρυθμό εξασθένησης της αναμενόμενης απόδοσης του κινητού διάμεσου, όσο αυξάνονται τα συναλλακτικά κόστη, καθιστά τις αποδόσεις του κινητού μέσου συγκριτικά υψηλότερες. Αυτή η αντιστροφή στον "ηγέτη" της αναμενόμενης απόδοσης οφείλεται στον υψηλότερο αριθμό συναλλαγών που επιτάσσει ο κανόνας του κινητού διάμεσου συγκριτικά με τον κανόνα του κινητού μέσου σχεδόν για όλα τα μήκη του εκτενέστερου κινητού διάμεσου. Αυτή η διαφορά είναι εμφανής στα συγκριτικά διαγράμματα του Σχήματος 2, στα οποία υποτυπώνεται ο αριθμός των συναλλαγών κάθε κανόνα σε κάθε περίοδο.

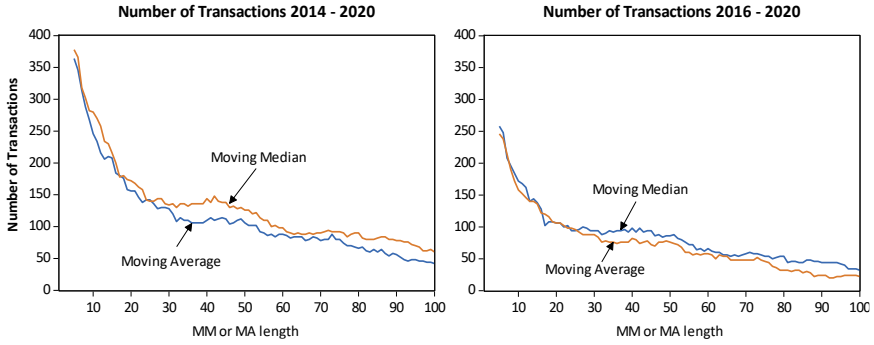
Μια άλλη σημαντική διαφορά μεταξύ των δύο συναλλακτικών κανόνων παρατηρείται στην τυπική απόκλιση της σειράς των αποδόσεων τους. Συγκρίνοντας τις τιμές του Πίνακα 3 και στις δύο εξεταζόμενες περιόδους και σε όλα τα σενάρια η τυπική απόκλιση του κινητού διάμεσου είναι αισθητά μικρότερη σε σχέση με αυτή του κινητού μέσου. Η διαφορά αυτή κυμαίνεται από 24% έως 55% και είναι πιο έντονη την περίοδο όπου η αγορά παρέμεινε σταθερή (2016-2020). Μεγιστοποίηση

της διαφοράς παρατηρείται στην περίπτωση του τυπικού μικροεπενδυτή, στην ίδια περίοδο, όπου καταγράφεται 55% υψηλότερη τυπική απόκλιση στις αποδόσεις του κινητού μέσου συγκριτικά με αυτή του κινητού διάμεσου.

Πίνακας 3. *Αποτελέσματα της αναμενόμενης απόδοσης και της τυπικής απόκλισης κάθε συναλλακτικού κανόνα ανά σενάριο καθώς και η απόδοση της παθητικής στρατηγικής. (* = αποτέλεσμα με επιφύλαξη)*

Εξεταζόμενα Σενάρια	Κινητός Διάμεσος		Κινητός Μέσος		Αγορά και Διακράτηση (%)
	E(R)	σ	E(R)	σ	
Χρονική Περίοδος : 2014 - 2020					
Χωρίς Κόστη & Φόρους	-10,99	3,78	-13,11	4,89	-50,40
Μέλος Χρηματιστηρίου	-30,17	3,12	-29,33	4,25	-50,58
Επαγγελματίας Επενδυτής	-61,98*	3,51	-57,48	5,05	-51,08
Ερασιτέχνης Επενδυτής	-78,25*	3,19	-73,05	4,76	-51,57
Θεσμικός Επενδυτής	-44,13	3,37	-41,30	4,81	-50,76
Χρονική Περίοδος: 2016 - 2020					
Χωρίς Κόστη & Φόρους	47,63	5,86	47,53	7,30	2,13
Μέλος Χρηματιστηρίου	27,01*	5,20	29,75	7,89	1,77
Επαγγελματίας Επενδυτής	-14,15	7,76	-6,31*	11,12	0,75
Ερασιτέχνης Επενδυτής	-40,17	8,42	-29,87	13,00	-0,26
Θεσμικός Επενδυτής	12,92*	5,82	18,77*	8,81	1,47

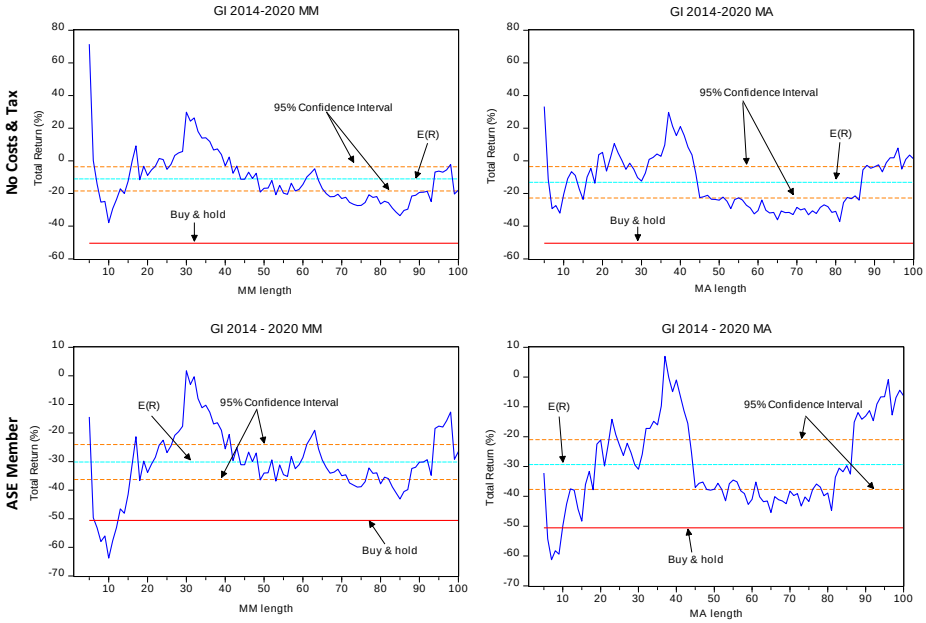
Σχήμα 2. *Αριθμός συναλλαγών κινητού διάμεσου και κινητού μέσου*



Όσον αφορά τα αποτελέσματα της σύγκρισης μεταξύ των αποδόσεων των δύο συναλλακτικών κανόνων έναντι της στρατηγικής της αγοράς και διακράτησης αυτά παρουσιάζονται γραφικά για την περίοδο 2014-2020 στα Σχήματα 3 και 5 και στα Σχήματα 4 και 6 για την περίοδο 2016-2020. Ειδικότερα, σε αυτά τα γραφήματα αποτυπώνεται η μεταβολή στις αποδόσεις των συναλλακτικών κανόνων του κινητού διάμεσου και του κινητού μέσου, η αναμενόμενη απόδοσή τους με το αντίστοιχο διάστημα εμπιστοσύνης 95% και η απόδοση της στρατηγικής της αγοράς και διακράτησης σε κάθε σενάριο. Και στις δύο περιόδους είναι ευδιάκριτο ότι χωρίς

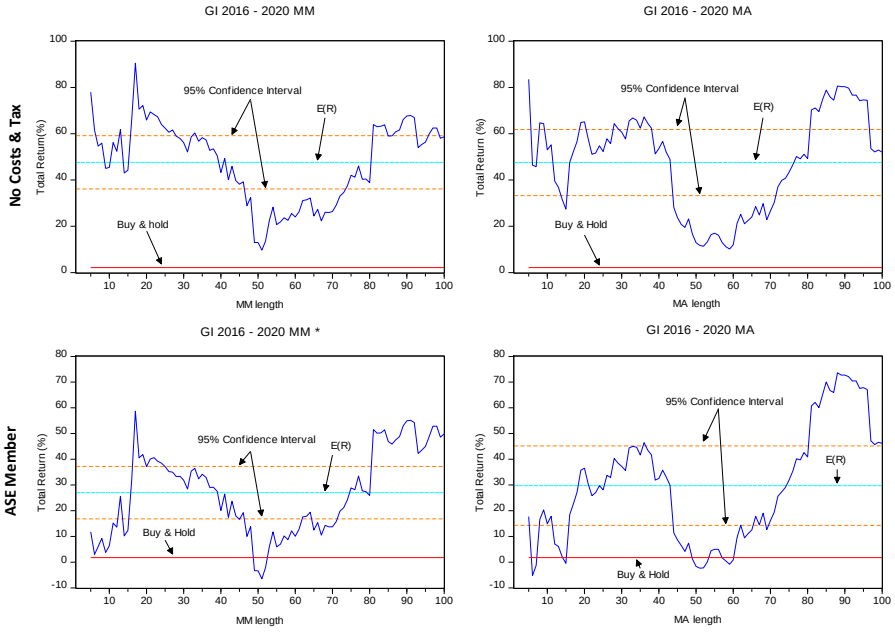
κόστη συναλλαγών και φορολογία η απόδοση της παθητικής στρατηγικής είναι χαμηλότερη από το κατώτερο όριο του διαστήματος εμπιστοσύνης της αναμενόμενης απόδοσης και των δύο συναλλακτικών κανόνων. Η μέση απόδοση του ΜΜ είναι υψηλότερη από την παθητική στρατηγική κατά 39,4% στην περίοδο 2014-2020 και κατά 45,5% στην περίοδο 2016-2020. Αντίστοιχα η μέση απόδοση του ΜΑ είναι υψηλότερη από την παθητική στρατηγική στην περίοδο 2014-2020 κατά 37,3% και στην περίοδο 2016-2020 κατά 45,4%. Συνεπώς, επιτυγχάνεται υπερνίκηση της αγοράς τόσο με τον κανόνα του κινητού διάμεσου όσο και με τον κανόνα του κινητού μέσου.

Σχήμα 3. *Έλεγχος σημαντικότητας της περιόδου 2014 – 2020 για τα σενάρια χωρίς συναλλακτικά κόστη και για μέλος του ΧΑΑ.*

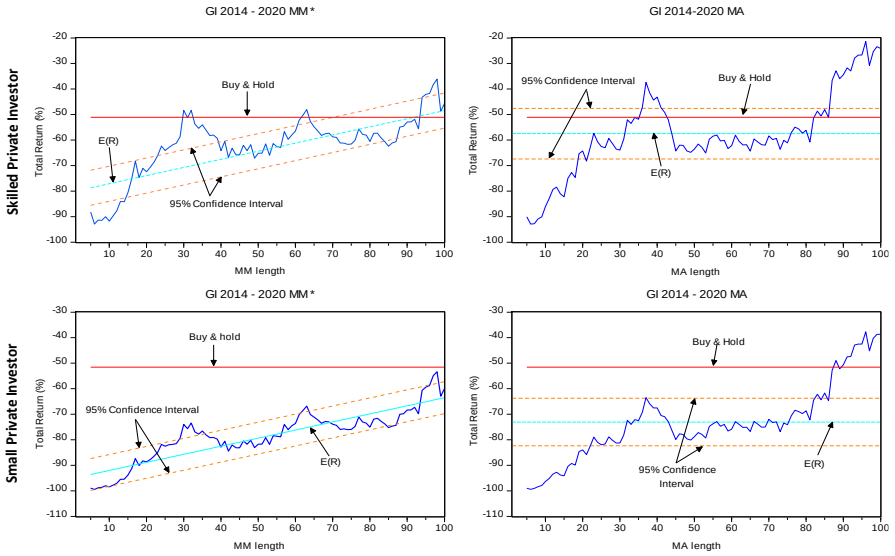


Τα αποτελέσματα δε διαφέρουν σημαντικά και στην περίπτωση που ένα μέλος του ΧΑΑ εφαρμόσει τους εν λόγω συναλλακτικούς κανόνες. Η φορολογία και τα έξοδα υπέρ του χρηματιστηρίου μειώνουν μεν το ύψος της αναμενόμενης απόδοσης των κανόνων, αλλά επιτυγχάνεται και πάλι υπερνίκηση της αγοράς. Ειδικότερα, η μέση απόδοση του ΜΜ στο συγκεκριμένο σενάριο εκτιμάται κατά 20,4% υψηλότερη από την παθητική στρατηγική στην περίπτωση της καθοδικής αγοράς, ενώ στην περίπτωση της σταθερής αγοράς η απόδοση αυτή αναμένεται υψηλότερη κατά 25,2%. Ενώ στην περίπτωση του ΜΑ, η εν λόγω διαφορά διαμορφώνεται στο 21,2% στην περίοδο 2014-2020 και στο 28% στην περίοδο 2016-2020.

Σχήμα 4. *Έλεγχος σημαντικότητας της περιόδου 2016 – 2020 για τα σενάρια χωρίς συναλλακτικά κόστη και για μέλος του ΧΑΑ. (* = αποτέλεσμα με επιφύλαξη)*



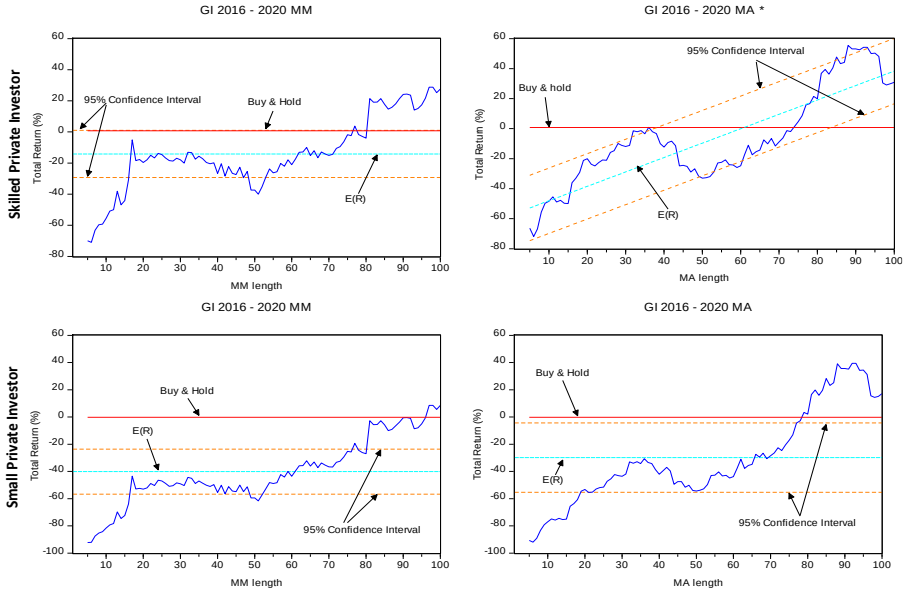
Σχήμα 5. *Έλεγχος σημαντικότητας της περιόδου 2014 – 2020 των σεναρίων για τους ιδιώτες επενδυτές. (* = αποτέλεσμα με επιφύλαξη)*



Όσον αφορά τα σενάρια για τις περιπτώσεις των ιδιωτών επενδυτών, στα οποία υπεισέρχεται προμήθεια επί των συναλλαγών επιπροσθέτως των λοιπών χρεώσεων, τα αποτελέσματα διαφέρουν ριζικά. Σε αυτές τις περιπτώσεις η απόδοση της στρατηγικής της αγοράς και διακράτησης είναι υψηλότερη από την αναμενόμενη

απόδοση και των δύο συναλλακτικών κανόνων. Εξαίρεση εν μέρει αποτελεί η περίπτωση του επαγγελματία μικροεπενδυτή στην περίοδο 2016-2020 κατά την οποία επιτυγχάνεται υπερνίκηση της αγοράς με τον MA για μήκη > 85, λόγω της ύπαρξης γραμμικής τάσης. Εν γένει στα δύο σενάρια που αφορούν τους ιδιώτες επενδυτές η απόδοση των δύο κανόνων δε διαφέρει ή είναι χειρότερη από την απόδοση της παθητικής στρατηγικής.

Σχήμα 6. *Έλεγχος σημαντικότητας της περιόδου 2016 – 2020 των σεναρίων για τους ιδιώτες επενδυτές. (* = αποτέλεσμα με επιφύλαξη)*

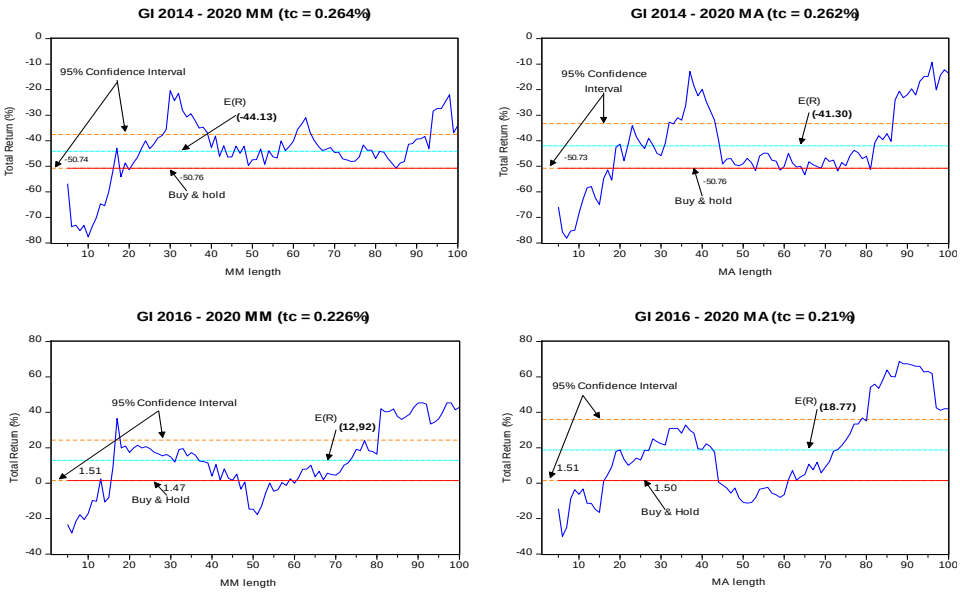


Από τα αποτελέσματα των εξεταζόμενων περιόδων φαίνεται ότι τα συναλλακτικά κόστη προκαλούν σημαντική μείωση στην αναμενόμενη απόδοση των συναλλακτικών κανόνων. Όμως, το ύψος της προμήθειας επί των συναλλαγών είναι αυτό που διαδραματίζει τον καθοριστικό ρόλο στην υπερνίκηση ή μη της ελληνικής αγοράς με τη χρήση των εξεταζόμενων συναλλακτικών κανόνων. Σε όλα τα ανωτέρω διαγράμματα επαληθεύεται και οπτικά η ουσιωδώς μικρότερη διακύμανση που διαπιστώθηκε στις σειρές των αποδόσεων του κινητού διάμεσου συγκριτικά με αυτές του κινητού μέσου στον Πίνακα 3 και ακριβώς το γεγονός αυτό αποτελεί και το συγκριτικό πλεονέκτημα του συναλλακτικού κανόνα του κινητού διάμεσου έναντι του κινητού μέσου.

Τέλος, στο εξεταζόμενο σενάριο που αντιπροσωπεύει τους θεσμικούς επενδυτές υιοθετήθηκε μια διαφορετική προσέγγιση για τον καθορισμό του ύψους της προμήθειας επί των συναλλαγών. Καθώς το εν λόγω ποσοστό καθορίζεται βάσει ιδιωτικών συμφωνιών αναζητήθηκε το μέγιστο ύψος, ως ποσοστό των συνολικών συναλλακτικών εξόδων, δεδομένης της τρέχουσας φορολογίας επί των πωλήσεων αξιογράφων στο ΧΑΑ, ώστε να επιτυγχάνεται υπερνίκηση της αγοράς από την εφαρμογή των ανωτέρω συναλλακτικών κανόνων. Συγκεκριμένα, ακολουθήθηκε μια

επαναληπτική διαδικασία δοκιμών που τερματίζεται όταν ικανοποιούνται από κοινού δύο συνθήκες: (1) μεγιστοποίηση στο ύψος του κόστους συναλλαγών και (2) οριακή υπερνίκηση της αγοράς σε διάστημα εμπιστοσύνης 95%. Το ευρεθέν μέγιστο ύψος συναλλακτικών εξόδων κυμαίνεται από 0,21% έως 0,264% και εξαρτάται τόσο από τον συναλλακτικό κανόνα όσο και από την εξεταζόμενη περίοδο (Σχήμα 7). Δεδομένης της υπόθεσης ότι οι θεσμικοί επενδυτές επιβαρύνονται με συναλλακτικά κόστη μέχρι 0,2% το πολύ, προκύπτει, βάσει των ευρημάτων, ότι δύναται να κερδίσουν την συγκεκριμένη αγορά με την χρήση τόσο του ΜΜ όσο και του ΜΑ. Συνεπώς, στο σενάριο των θεσμικών επενδυτών δύναται να επιτευχθεί υπερνίκηση της αγοράς και με τους δύο συναλλακτικούς κανόνες.

Σχήμα 7. *Σενάριο θεσμικών επενδυτών των περιόδων 2014-2020 και 2016-2020*



Από την σύγκριση μεταξύ των αποτελεσμάτων των δύο κανόνων, στα διαγράμματα του Σχήματος 7, ανακύπτει ότι και στις δύο περιόδους ο κανόνας του κινητού διάμεσου επιτυγχάνει υπερνίκηση της αγοράς με υψηλότερο κόστος συναλλαγών. Ωστόσο, βάσει του Πίνακα 3 και του Σχήματος 2, ο κανόνας του κινητού διάμεσου μειονεκτεί έναντι του κινητού μέσου όταν υπεισέρχονται έστω και τα ελάχιστα δυνατά συναλλακτικά κόστη, λόγω των περισσότερων διενεργούμενων συναλλαγών. Επομένως, είναι αναμενόμενο ο κανόνας του κινητού μέσου να επιτυγχάνει υπερνίκηση της αγοράς με υψηλότερο κόστος συναλλαγών. Όμως, η εμφανώς μικρότερη τυπική απόκλιση που εμφανίζει ο κινητός διάμεσος συγκριτικά με τον κινητό μέσο οδηγεί σαφώς σε μικρότερο εύρος στο διάστημα εμπιστοσύνης του πρώτου. Αυτή η σημαντική διαφορά στο εύρος του διαστήματος εμπιστοσύνης επαρκεί, ώστε να υπερκεραστεί το μειονέκτημα του κινητού διάμεσου (οφειλόμενο στον αριθμό των συναλλαγών) και να επιτευχθεί υπερνίκηση της αγοράς με υψηλότερο κόστος συναλλαγών. Άλλωστε, είναι ευδιάκριτο σε όλα τα διαγράμματα της παρούσας εργασίας το ευρύτερο διάστημα εμπιστοσύνης που εμφανίζει η

αναμενόμενη απόδοση του κινητού μέσου σε σχέση με αυτό του κινητού διάμεσου. Στο σημείο αυτό να σημειώσουμε ότι στο σενάριο των θεσμικών επενδυτών στην περίοδο 2016-2020 ανιχνεύτηκε η ύπαρξη ασθενούς γραμμικής τάσης στις σειρές των αποδόσεων των συναλλακτικών κανόνων. Συνεπώς, εκλαμβάνουμε τα εν λόγω αποτελέσματα με επιφύλαξη.

4. ΣΥΝΟΨΗ ΚΑΙ ΣΥΜΠΕΡΑΣΜΑΤΑ

Βασικός σκοπός της παρούσας μελέτης ήταν να προτείνει έναν εναλλακτικό συναλλακτικό κανόνα, ο οποίος στηρίζεται στον κινητό διάμεσο και να συγκριθεί η προβλεπτική του ικανότητα με αυτή του ευρέως χρησιμοποιούμενου συναλλακτικού κανόνα του κινητού μέσου με εφαρμογή στο Γενικό Δείκτη του Χρηματιστηρίου Αξιών Αθηνών. Διαπιστώθηκε ότι, ο συναλλακτικός κανόνας του κινητού διάμεσου αποδίδει καλύτερα από αυτόν του κινητού μέσου, γεγονός το οποίο αποδίδεται πρωτίστως όχι τόσο στο ύψος της αναμενόμενης απόδοσης, όσο στην εμφανώς μικρότερη διακύμανση του. Συνεπώς, το σημαντικότερο πλεονέκτημα του προτεινόμενου συναλλακτικού κανόνα ΜΜ είναι η εμφανώς μικρότερη μεταβλητότητα που εμφανίζουν οι σειρές των αποδόσεων που προκύπτουν από τα διαφορετικά μήκη του μεγάλου κινητού διάμεσου. Χωρίς φορολογία και κόστη συναλλαγής και οι δύο συναλλακτικοί κανόνες εξασφαλίζουν υψηλότερες αποδόσεις συγκριτικά με την στρατηγική της αγοράς και διακράτησης, επιτυγχάνοντας υπερνίκηση της αγοράς. Επομένως, στη συγκεκριμένη περίπτωση απορρίπτεται η υπόθεση της αποτελεσματικής κεφαλαιαγοράς ασθενούς ισχύος.

Επιπρόσθετος στόχος αυτής της μελέτης ήταν να εξετάσει την ικανότητα των συναλλακτικών κανόνων ΜΜ και ΜΑ να επιτύχουν υπερνίκηση της αγοράς, λαμβάνοντας υπόψη τα πραγματικά συναλλακτικά κόστη που επιβαρύνουν τους συναλλασσόμενους στη δευτερογενή ελληνική κεφαλαιαγορά. Από τα εμπειρικά ευρήματα προκύπτει ότι, ένα μέλος του ΧΑΑ δύναται να κερδίσει σημαντικά υψηλότερες αποδόσεις έναντι της παθητικής στρατηγικής, με την εφαρμογή αμφότερων των ανωτέρω συναλλακτικών κανόνων. Επομένως, και σε αυτήν την περίπτωση απορρίπτεται η υπόθεση της αποτελεσματικής αγοράς ασθενούς ισχύος. Όσον αφορά την περίπτωση ενός θεσμικού επενδυτή, είναι δυνατό ένας τέτοιος επενδυτής, έστω και οριακά, να υπερνικήσει την συγκεκριμένη αγορά. Αναγκαία συνθήκη για να το επιτύχει αποτελεί η ύπαρξη ικανής διαπραγματευτικής δύναμης από πλευράς του, ώστε να συνάπτει συμβάσεις με συναλλακτικά κόστη μέχρι του εκτιμώμενου μέγιστου ποσοστού. Συνεπώς, σε αυτή την περίπτωση η απόρριψη ή μη της υπόθεσης της αποτελεσματικής αγοράς ασθενούς ισχύος εξαρτάται από το ύψος του συναλλακτικού κόστους που αντιμετωπίζει η εκάστοτε θεσμική οντότητα. Αντίθετα, οι ιδιώτες επενδυτές τόσο οι επαγγελματίες όσο και οι απλοί, οι οποίοι αντιμετωπίζουν σημαντικά υψηλότερη προμήθεια επί των συναλλαγών, όχι μόνο δεν επιτυγχάνουν να υπερνικήσουν την συγκεκριμένη αγορά, αλλά οι αποδόσεις από την εφαρμογή των εν λόγω συναλλακτικών κανόνων είναι χαμηλότερες από αυτές της παθητικής στρατηγικής, επιφέροντας ζημιά στο κεφάλαιο τους. Συνεπώς, στα

σενάρια των ιδιωτών επενδυτών δεν απορρίπτεται η υπόθεση της αποτελεσματικής αγοράς ασθενούς ισχύος.

Συνοψίζοντας, στην ελληνική κεφαλαιαγορά η βέλτιστη επενδυτική στρατηγική, ανάμεσα στην εφαρμογή της τεχνικής ανάλυσης και της αγοράς και διακράτησης, εξαρτάται από την ιδιότητα του επενδυτή. Το γεγονός αυτό έχει άμεσες επιπτώσεις στον έλεγχο της υπόθεσης της αποτελεσματικότητας κεφαλαιαγορών, καθώς η απόρριψη ή μη της εν λόγω υπόθεσης εξαρτάται από την ιδιότητα της εκάστοτε επενδυτικής οντότητας. Τέλος, αξίζει να σημειωθεί ότι ήδη από τις αρχές του 2020 το ΧΑΑ, όπως και όλες οι κεφαλαιαγορές παγκοσμίως, επηρεάστηκε έντονα από την πανδημία του κορονοϊού (corona virus covid 19), κάτι που είναι φανερό και από την οπτική επισκόπηση του Σχήματος 1. Ασφαλώς, μετά τη λήξη της πανδημίας θα ήταν ενδιαφέρον να επαναληφθεί η μελέτη ειδικά για την χρονική περίοδο της πανδημίας.

ABSTRACT

A typical way of checking the market efficiency hypothesis is to compare the performance of technical trading rules with the buy and hold strategy. In this work, an alternative technical rule is proposed based on moving median, and its predictive power for the Athens Stock Exchange is contrasted with that of the most popular technical trading rule of moving average, with and without transaction costs. In the theoretical case of no transaction costs empirical findings show that the predictive power of the moving median rule is higher than that of the moving average rule, while both perform better than the passive strategy. Hence, for both trading rules the hypothesis of weak-form efficiency is rejected. However, by introducing real transaction costs, such as those existing in the particular market, it is found that it is still possible for an institutional investor to beat the market, even marginally, but this is not the case for a typical small investor, due to higher transaction costs for the latter. Consequently, the result on the testing of the hypothesis of efficient markets, given transaction costs, depends on the status of the investor.

ΑΝΑΦΟΡΕΣ

Bartlett, M. (1946). On the theoretical specification of sampling properties of autocorrelated time series, *Journal of the Royal Statistical Society Supplement*, **8**, pp.27-41.

Black, F. (1986). Noise, *The Journal of Finance*, **41(3)**, 529-543.

Brock, W., Lakonishok, J. and LeBaron, B. (1992). Simple technical trading rules and the stochastic properties of stock returns, *Journal of Finance*, **47**, 1731-1764.

Cai, B. M., Cai, C. X. and Keasey, K. (2005). Market efficiency and returns to simple technical trading rules: Further evidence from U.S., U.K., Asian and Chinese stock markets. *Asia-Pacific Financial Markets*, **12**, 45-60.

Cowles, A. (1934). Can stock market forecasters forecast?, *Econometrica*, **1**, 309-324.

Elliot, G., Rothenberg, T.J. and Stock, J.H. (1996). Efficient Tests for an Autoregressive Unit Root, *Econometrica*, **64**, pp.813-836.

Elton, E. J. and Gruber, M. J. (1995). *Modern Portfolio Theory and Investment Analysis*, 5th edition, New York: Wiley.

Fama, E. (1970). Efficient capital markets: a review of theory and empirical work, *Journal of Finance*, **25**, 383-417.

Fama, E. (1991). Efficient capital markets II, *Journal of Finance*, **46**, 1557-1617.

Fama, E. and Blume M. (1966). Filter rules and stock market trading profits, *Journal of Finance*, **39**, 226-241.

Hudson, R., Dempsey, M. and Keasy, K. (1996). A note on the weak form efficiency of capital markets: the application of simple technical trading rules to UK stock prices – 1935 to 1994, *Journal of Banking and Finance*, **20**, 1121-1132.

Kwon, K. Y. and Kish, R. J. (2002). Technical trading strategies and return predictability: NYSE, *Applied Financial Economics*, **12 (9)**, 639-653.

Malkiel, B. (1992). *Efficient Market Hypothesis*, in Newman, P., Milgate, M., and Eatwell, J. (eds). New Palgrave Dictionary of Money and Finance, London: McMillan.

Milionis, A.E. (2007). Efficient Capital Markets: A Statistical Definition and Comments. *Statistics and Probability Letters*, **77**, 607-613.

Milionis, A.E. and Papanagiotou, E., (2008). On the use of the moving average trading rule to test for weak form efficiency in capital markets, *Econ. Notes 37*, pp. 181-201.

Milionis, A.E. and Papanagiotou, E., (2011). A test of significance of the predictive power of the moving average trading rule of technical analysis based on sensitivity analysis: application to the NYSE, the Athens Stock Exchange and the Vienna Stock Exchange. Implications for weak-form market efficiency testing, *Applied Financial Economics*, **21:6**, 421-436.

Milionis, A.E. and Papanagiotou, E., (2013). Decomposing the predictive performance of the moving average trading rule of technical analysis: the contribution of linear and non-linear dependencies in stock returns, *Journal of Applied Statistics*, **40:11**, 2480-2494.

Mills, T. (1997). Technical Analysis and the London Stock Exchange: Testing Trading Rules Using the FT30, *International Journal of Finance & Economics*, **2**, pp. 319-331.

Murphy, J. (1999). *Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications*, New York Institute of Finance.

Neftci, S.N. (1991). Naive trading rules in financial markets and Wiener-Kolmogorov prediction theory: a study of technical analysis, *Journal of Business*, **64(4)**, pp. 549-571.

Olson, D. (2004). Have trading rule profits in the currency markets declined over time?, *Journal of Banking and Finance*, **28**, 85–105.

Taylor, M. and Allen, H. (1992). The use of technical analysis in the foreign exchange market, *Journal of International Money and Finance*, **11**, pp. 304-314.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 33ου Πανελληνίου Συνεδρίου Στατιστικής (2021), σελ.195-208

Συνεχείς Τεθλασμένες Κατανομές και Συντελεστής Μεταβλητότητας

Μπατσιάκα Μαρία[1], Χατζημιχαήλ Χριστίνα[1], Φαρμάκης Νικόλαος[1][1]
ΠΜΣ Τμήμα Μαθηματικών Α.Π.Θ.
xristina.k.xatzimixail@gmail.com, marmpat@hotmail.com , farmakis@math.auth.gr

ΠΕΡΙΛΗΨΗ

Η Δειγματοληψία μπορεί να συμβάλει στη μελέτη των κατανομών των διαφόρων τυχαίων μεταβλητών με πολύ καλά αποτελέσματα από άποψη ακρίβειας και κυρίως ταχύτητας. Με την βοήθεια του δείγματος υπολογίζουμε τις διάφορες παραμέτρους μιας τυχαίας μεταβλητής (τ.μ.) X και μέσα από κάποια διαδικασία τον τύπο της συνάρτησης πυκνότητας πιθανότητας (σ.π.π.) της τ.μ. X. Η διαδικασία αυτή που ακολουθούμε είναι συνήθως θεωρητική και βασίζεται σε μια αρχική υπόθεση για την μορφή που μπορεί να έχει η σ.π.π. ή τουλάχιστον μία προσέγγισή της που μπορεί να είναι αποδοτική και εύκολη στη διαχείρισή της συγχρόνως. Στην παρούσα εργασία ασχολούμαστε με συνεχείς κατανομές όπου η καμπύλη της σ.π.π. είναι τεθλασμένη γραμμή. Η καμπύλη έχει σχήμα τεθλασμένης γραμμής αν οι δύο κλάδοι της καμπύλης είναι ευθύγραμμα τμήματα με διαφορετική κλίση σε κάθε κλάδο. Η όλη μελέτη είναι θεωρητική προσέγγιση της κατανομής. Δίνονται και κατάλληλα παραδείγματα με χρήση Δειγματοληψίας για την καλύτερη κατανόηση του πώς από το δείγμα φτάνουμε στην εκτιμήτρια της σ.π.π. σε ελάχιστο χρόνο σε σχέση με κλασσικές μεθόδους προσδιορισμού της μορφής της σ.π.π.

Λέξεις κλειδιά: Δειγματοληψία, συνάρτηση πυκνότητας πιθανότητας, μέση τιμή, διασπορά, συντελεστής μεταβλητότητας, τεθλασμένη γραμμή.

MSC: 62D05; 62E17

Εισαγωγή

Ο αναλυτικός προσδιορισμός της κατανομής μιας τυχαίας μεταβλητής (τ.μ.) X είναι πάρα πολύ σημαντική υπόθεση για τη μελέτη της τ.μ. και είναι συνήθως πολύ χρονοβόρα η διαδικασία. Πολλές φορές αυτή η προσέγγιση για σημαντικές τ.μ. κράτησε έτη ή και δεκαετίες. Η προσέγγιση αυτή γινόταν με διάφορες μεθόδους και

τεχνικές που διέφεραν από περίπτωση σε περίπτωση ανάλογα και με το εκάστοτε αντικείμενο που αναφέρονταν οι τ.μ. .

Από τις αρχές της 3ης μ.Χ. χιλιετίας άρχισε να χρησιμοποιείται για τον προσεγγιστικό τουλάχιστον προσδιορισμό (εκτίμηση) των συναρτήσεων πυκνότητας πιθανότητας (σ.π.π.) της τ.μ. X η έννοια του Συντελεστή Μεταβλητότητας (ΣΜ), (Coefficient of Variation (Cv)), Farmakis (2003, 2010), Παπατσούμα (2018), Φαρμάκης (2015).

Οι σ.π.π. που δίνονται στις παραπάνω δημοσιεύσεις είναι μονώνυμο βαθμού ν>-1 και αναφέρονται συνήθως σε συνεχείς τ.μ. X. Μελετήθηκαν κάποιες βασικές μορφές των σ.π.π. όπως: *Συμμετρικές*, Farmakis (2003), Παπατσούμα (2018), Φαρμάκης (2015), *Αύξουσες*, Farmakis (2010), Παπατσούμα (2018), *Φθίνουσες*, Παπατσούμα (2018). Με τις μεθόδους αυτές αξιοποιείται ο ΣΜ και μάλιστα η παράμετρος q, το τετράγωνο του αντιστρόφου του, δηλαδή το

$$q = ICv^2 = Cv^{-2} \tag{1}$$

όπου $Cv=\sigma/\mu$, μ=μέση τιμή και σ=τυπική απόκλιση της τ.μ. X. Στις εφαρμογές χρησιμοποιούνται οι εκτιμήτριες των παραμέτρων μ, σ, Cv, q. Οι μορφές των σ.π.π. στις τρεις βασικές μορφές συμμετρική, αύξουσα και φθίνουσα δίνονται εδώ συνοπτικά:

(Α) Συμμετρική σ.π.π., Farmakis (2003)

$$f(x) = \begin{cases} hx^{\nu}, x \in [0, \frac{\beta}{2}) \\ h(\beta - x)^{\nu}, x \in [\frac{\beta}{2}, \beta] \\ 0, x \notin [0, \beta] \end{cases}, \nu = \frac{-5+\sqrt{1+8q}}{2}, h = \frac{2^{\nu}(\nu+1)}{\beta^{\nu+1}} \tag{2}$$

(Β) Αύξουσα σ.π.π., Farmakis (2010), Παπατσούμα (2018)

$$f(x) = \begin{cases} h\left(\frac{x}{\beta}\right)^{\nu}, & x \in [0, \beta] \\ 0, & x \notin [0, \beta] \end{cases}, \ \nu = -2 + \sqrt{1+q}, \ h = \frac{\nu+1}{\beta} \tag{3}$$

(Γ) Φθίνουσες Τύπου Ι σ.π.π., Παπατσούμα (2018)

$$f(x) = \begin{cases} h\left(1 - \frac{x}{\beta}\right)^{\nu}, & x \in [0, \beta] \\ 0, & x \notin [0, \beta] \end{cases}, \ \nu = \frac{3-q}{q-1} = -1 + \frac{2}{q-1}, \ h = \frac{\nu+1}{\beta} \tag{4}$$

και Φθίνουσες Τύπου ΙΙ, Παπατσούμα (2018)

$$f(x)=\begin{cases} h\left(1-\left(\dfrac{x}{\beta}\right)^{\nu}\right), & x\in[0,\beta] \\ 0 \quad , & x\notin[0,\beta] \end{cases},\ \nu=-2+\sqrt{3\cdot\frac{q+1}{3-q}},\ h=\frac{\nu+1}{\nu\cdot\beta}\ . \tag{5}$$

Η απόφαση για το ποια μορφή σ.π.π. θα χρησιμοποιήσουμε βασίζεται στο σχήμα που μας δίνει το ιστόγραμμα σε συνδυασμό με το πολύγωνο συχνοτήτων, το σχετικό με τα δεδομένα μας, που είναι δειγματικά συνήθως.

Σύνθετες μορφές της σ.π.π.: Τεθλασμένη συνεχής κατανομή

Η νέα ιδέα αυτής της εργασίας είναι να γίνει εκτίμηση της σ.π.π. με τη βοήθεια δείγματος και του ΣΜ όταν η καμπύλη της σ.π.π. της συνεχούς τμ Χ εικάζεται ότι αποτελείται από δύο κλάδους. Όταν π.χ. αυτοί οι κλάδοι είναι τμήματα ευθείας τότε το όλο σχήμα είναι το σχήμα μιας τεθλασμένης με δύο σκέλη (πλευρές), (καταχρηστικά χρησιμοποιούμε τον όρο τεθλασμένη και όταν οι κλάδοι δεν είναι ευθύγραμμα τμήματα). Τότε οι δύο κλάδοι έχουν στο κρίσιμο σημείο (με τετμημένη τιμή της τ.μ. X την τιμή $x_{κρ}$) διαφορετική κλίση δηλαδή η από αριστερά παράγωγος και η από δεξιά παράγωγος δεν είναι ίσες και άρα η καμπύλη της σ.π.π. δεν έχει παράγωγο στο σημείο αυτό, Μωυσιάδης (2000). Αυτονόητα και στους δύο κλάδους η τ.μ. X είναι συνεχής και παίρνει τιμές στο [0, β].

Η μέθοδος εργασίας έχει δύο στάδια και περιγράφεται παρακάτω:

1ο) Η εκτίμηση για την τιμή $x_{κρ}$ όπου τέμνονται οι δύο κλάδοι Α και Β αντίστοιχα της καμπύλης της σ.π.π., δηλαδή το σημείο αλλαγής κλάδου της καμπύλης της σ.π.π. .
2ο) Η εύρεση των δύο εξισώσεων που περιγράφουν αντίστοιχα τους δύο κλάδους.

Των δύο αυτών σταδίων προηγείται και προϋποτίθεται πάντα δειγματοληψία και η συλλογή και καταγραφή των δεδομένων που προκύπτουν για την τ.μ. X από το δείγμα. Η μέθοδος ή η τεχνική της δειγματοληψίας μπορεί να είναι οποιαδήποτε κριθεί κατάλληλη με βάση τα πορίσματα της διεθνούς βιβλιογραφίας, π.χ. Cochran (1977), Φαρμάκης (2009, 2015, 2016).

Από το δείγμα εκτιμάται το εύρος τιμών β της κατανομής και οι τιμές της τ.μ. X ταξινομούνται σε κ κλάσεις στις οποίες αντιστοιχούν οι κ συχνότητες τιμών της τ.μ. . Κατασκευάζεται το αντίστοιχο ιστόγραμμα συχνοτήτων της τ.μ. X απ' όπου θα εκτιμηθεί και η μορφή της σ.π.π. δηλαδή αν είναι συμμετρική, αύξουσα, φθίνουσα, τεθλασμένη, κλπ. Ενδιαφέρει η τεθλασμένη εδώ και μάλιστα θα περιοριστούμε στην

περίπτωση που έχουμε δύο κλάδους της σ.π.π. με μορφή αύξουσα αμφότεροι και με διαφορετική κλίση, φυσικά. Αν οι συχνότητες που αντιστοιχούν στους δύο κλάδους είναι N_A και N_B τότε εκτιμούμε τα δύο ποσοστά από

$$\hat{p}_A = \mathrm{N}_\mathrm{A} / \left(\mathrm{N}_\mathrm{A} + \mathrm{N}_\mathrm{B}\right) \quad \& \quad \hat{p}_\mathrm{B} = \mathrm{N}_\mathrm{B} / \left(\mathrm{N}_\mathrm{A} + \mathrm{N}_\mathrm{B}\right). \tag{6}$$

Η εύρεση των εξισώσεων των δύο κλάδων ακολουθεί τα όσα περιγράφονται σχετικά με τις αύξουσες σ.π.π. στη βιβλιογραφία, Farmakis(2010), Παπατσούμα (2018) και δίνονται εδώ από τη σχέση (3) κυρίως. Ακολουθούν παραδείγματα.

Παραδείγματα εύρεσης εκτιμήτριας της σ.π.π. σε τεθλασμένη συνεχή κατανομή

Παράδειγμα 1ο: Η επίλυση ενός προβλήματος που περιλάμβανε και χρήση μηχανής αναζήτησης διαρκεί από μερικά λεπτά έως πάνω από 100 λεπτά της ώρας. Αυτό προέκυψε από δειγματοληψία μεταξύ φοιτητών ΑΕΙ. Το δείγμα είχε μέγεθος N=573. Χωρίσαμε το εύρος της τ.μ. X=*χρόνος επίλυσης* σε κ=11 τάξεις, με ίδιο εύρος w=10min όλες. Σημειώσαμε τις αντίστοιχες συχνότητες στον Πίνακα 1. Οι αθροιστικές συχνότητες εμφανίζονται στη 4η στήλη με επικεφαλίδα N_i, η 5η στήλη είναι θεωρητικά μεγέθη που προκύπτουν από τη σ.π.π. με βάση τη σχέση (3), όπως θα φανεί κατά την πρόοδο της επίλυσης.

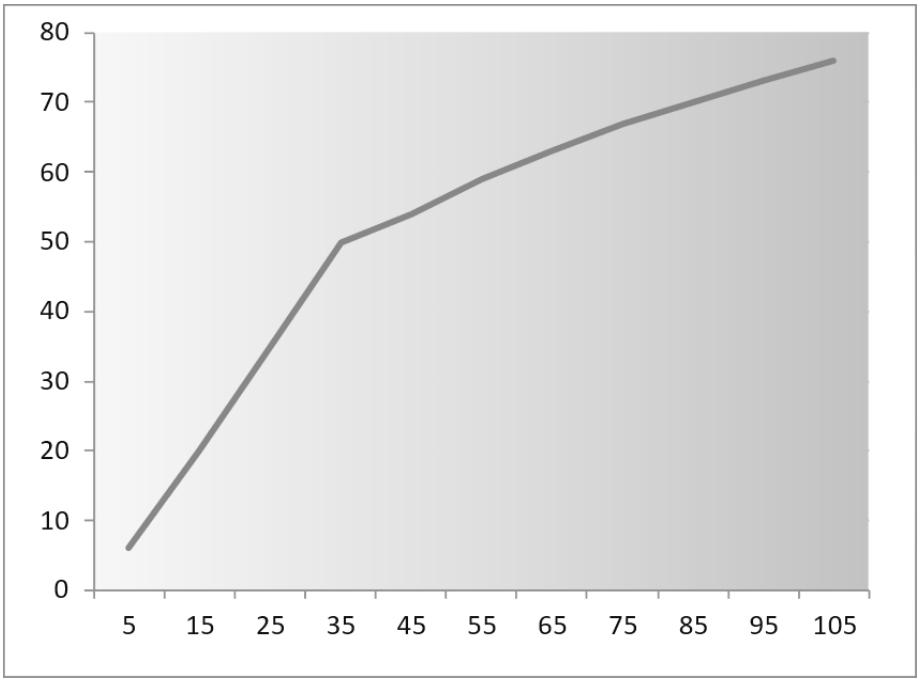
Πίνακας 1

Τάξεις	**Κέντρα τάξεων x_i**	**Συχνότητες n_i**	N_i	θ_i
[0, 10)	5	6	6	**10.11**
[10, 20)	15	20	26	**22.37**
[20, 30)	25	35	61	**31.80**
[30, 40)	35	50	111	**40.06**
[40, 50)	45	54	165	**47.58**
[50,60)	55	59	224	**54.59**
[60, 70)	65	63	287	**61.20**
[70, 80)	75	67	354	**67.49**
[80, 90)	85	70	424	**73.52**
[90, 100)	95	73	497	**79.33**
[100, 110]	105	76	573	**84.96**
Σύνολα		573=n		573.01

Να ευρεθεί η εκτιμήτρια της σ.π.π. από όλο το δείγμα. Μετά να εξεταστεί αν μπορεί να βρεθεί εκτιμήτρια σ.π.π. κατά κλάδο αφού διαπιστωθεί ότι μπορεί να θεωρηθεί ότι έχουμε δύο κλάδους. Τέλος να γίνει σύγκριση των δύο μεθόδων.

Λύση: Κατασκευάζουμε πρώτα το ιστόγραμμα συχνοτήτων της τ.μ. X με βάση τα στοιχεία του Πίνακα 1:

Σχήμα 1-Πολύγωνο Συχνοτήτων



Έτσι βλέπουμε ότι διαμορφώνονται δύο κλάδοι για το πολύγωνο συχνοτήτων: Ο ένας περιλαμβάνει τις 4 πρώτες τάξεις με άθροισμα συχνοτήτων N_A=111 και ο δεύτερος τις 7 επόμενες αντίστοιχο άθροισμα N_B=462 . Το πολύγωνο συχνοτήτων είναι μια τεθλασμένη γραμμή με δύο «πλευρές»-κλάδους.

Αρχικά εκτιμούμε την σ.π.π. της τ.μ. X από όλο το δείγμα θεωρώντας ότι η γωνία μεταξύ των δύο κλάδων είναι αρκετά μικρή και ότι η μέθοδος της σχέσης (3) θα αποδώσει αξιόπιστη σ.π.π. ενός κλάδου και με αύξουσα τάση. Η σ.π.π. που προέκυψε είναι η (7) και έπεται του Πίνακα 2 που ακολουθεί και περιέχει τις παραμέτρους της:

Πίνακας 2

Παράμετρος	**Τιμή Παραμέτρου**
w	10
Εύρος κατανομής *β*	110
Μέση τιμή	67.39

Διασπορά	732.2105
Τυπική απόκλιση	27.05
ΣΜ=Cv	0.401529
Q=ICv²	6.202501
Εκθέτης=ν	0.6837
Συντελεστής=h	0.0153068

$$f(x)=\begin{cases}0.0153068\cdot\left(\dfrac{x}{110}\right)^{0.6837}, & x\in[0,110]\\ 0, & x\notin[0,110]\end{cases} \qquad (7)$$

Με τη βοήθεια της σ.π.π. (7) υπολογίσαμε θεωρητικά μεγέθη θ_i, i=1,2,…,11 για τις 11 τάξεις του δείγματος και κάναμε τη δοκιμασία X^2 καλής προσαρμογής των n_i στα προκύψαντα θ_i και βρέθηκε συντελεστής X^2=7.6087, ενώ η κρίσιμη τιμή του για 8 βαθμούς ελευθερίας (β.ε.) και στάθμη σημαντικότητας (σ.σ.) α=0.05 είναι $X^2(8;0.05)$=15.5073. Άρα έχουμε καλή προσαρμογή και η (7) δίνει μία καλή εκτίμηση για την σ.π.π. σχετική με τα δεδομένα του Πίνακα 1.

Εξετάζουμε στη συνέχεια αν μπορούμε να βρούμε μία σ.π.π. με δύο κλάδους και με σημείο τομής που αντιστοιχεί στην τιμή της τ.μ. X (τετμημένη) $x_{κρ}$=40 όπως προκύπτει από την σύγκριση του Πίνακα 1 και του Σχήματος 1.

Ο κλάδος Α θα προκύψει με τη βοήθεια του Πίνακα 3 αντίστοιχη σ.π.π. με την (7).

Πίνακας 3

Τάξεις	**Κέντρα τάξεων x_i**	**Συχνότητες n_i**	$N_{i,A}$	θ_i
[0, 10)	5	6	6	**6.11**
[10, 20)	15	20	26	**19.93**
[20, 30)	25	35	61	**34.77**
[30, 40)	35	50	111	**50.19**
Σύνολα		111=N_A		111.00

Η σ.π.π. ου κλάδου Α είναι η

$$f(x)=\begin{cases}0.000933x^{1.0915}, & x\in[0,40]\\ 0, & x\notin[0,40]\end{cases}. \qquad (8)$$

Αντίστοιχα για τον κλάδο Β σημειώνουμε τα εξής:

1ο) Η σχέση (3) που θα βασιστούμε υποδεικνύει ότι η καμπύλη της σ.π.π. περνάει από την αρχή $O(0,0)$. Άρα θα εργαστούμε με τα στοιχεία των 7 τελευταίων τάξεων του Πίνακα 1 αλλά θα κάνουμε μετασχηματισμό και στους δύο άξονες με τμ $X'=X$-40 και με συχνότητες $n_i'=n_i$-50 (50 είναι η συχνότητα της 4ης τάξης που είναι αμέσως προηγούμενη της 1ης τάξης του κλάδου Β).

2ο) Η σ.π.π. θα προκύψει από τα στοιχεία του Πίνακα 4 και με βάση τη σχέση (3) και μετά θα παίρνεται υπόψη ο μετασχηματισμός επί των 2 αξόνων. Αυτό σημαίνει ότι η (3) θα πάρει τη μορφή:

$$f(x)=\begin{cases} h\left(\dfrac{x-40}{\beta-40}\right)^{\nu}, & x\in[40,\beta] \\ 0, & x\notin[40,\beta] \end{cases},\ \nu=-2+\sqrt{1+q},\ h=\frac{\nu+1}{\beta-40}, w=10 \qquad (9)$$

Η σχέση (9) θα πάρει τελικά τη μορφή:

$$f(x)=\begin{cases} 0.024384\cdot\left(\dfrac{x-40}{70}\right)^{0..7069}, & x\in[40,110] \\ 0, & x\notin[40,110] \end{cases}. \qquad (10)$$

Μετά θα υπολογίσουμε τα θεωρητικά μεγέθη $\theta_i=\theta_i'+50$ όπου το θ_i' θα προκύπτει από ολοκλήρωση της σ.π.π. της (10) με μέγεθος δείγματος $N_B'=112$ που προκύπτει από το Β κλάδο μετά από την μείωση όλων των συχνοτήτων (που είναι 7) κατά 50, όπου $n_i'=n_i$-50 είναι ο μετασχηματισμός. Τα θεωρητικά μεγέθη είναι στην τελευταία στήλη του Πίνακα 4.

Πίνακας 4

Τάξεις	**Κέντρα τάξεων** $\mathbf{x_i}$	**Συχνότητες** $\boldsymbol{n_i}$	**Συχνότητες** $\boldsymbol{n_i'}$	$\boldsymbol{N'_{i,B}}$	$\boldsymbol{\theta_i}$
[40, 50)	45	**54**	4	4	**54.04**
[50,60)	55	**59**	9	13	**59.16**
[60, 70)	65	**63**	13	26	**63.17**
[70, 80)	75	**67**	17	43	**66.72**
[80, 90)	85	**70**	20	63	**69.98**
[90, 100)	95	**73**	23	86	**73.02**
[100, 110]	105	**76**	26	112	**75.91**
Σύνολα		462	112=N'_B		**462.00**

Συνοπτικά τα αποτελέσματα στον ενιαίο Πίνακα 5 και για τους δύο κλάδους:

Πίνακας 5

Τάξεις	**Κέντρα τάξεων x_i**	**Συχνότητες n_i**	N_i	θ_i
[0, 10)	5	**6**	6	**6.11**
[10, 20)	15	**20**	26	**19.93**
[20, 30)	25	**35**	61	**34.77**
[30, 40)	35	**50**	111	**50.19**
[40, 50)	45	54	165	54.04
[50,60)	55	59	224	59.16
[60, 70)	65	63	287	63.17
[70, 80)	75	67	354	66.72
[80, 90)	85	70	424	69.98
[90, 100)	95	73	497	73.02
[100, 110]	105	76	573	75.91
Σύνολα		573=n		573.00

Η δοκιμασία X^2 έδωσε δειγματικό στατιστικό X^2=0,0066<< $X^2_{(8;0.05)}$=15.5073, δηλαδή έχουμε καλή προσαρμογή των παρατηρούμενων στα θεωρητικά μεγέθη συχνοτήτων. Το Στατιστικό αυτό σε σχέση με το αντίστοιχο προηγούμενο που βρήκαμε για την ενιαίου κλάδου λύση, το X^2=7.6087, είναι περίπου 1153 φορές μικρότερο.

Παράδειγμα 2ο: Η επίλυση ενός πολύ δύσκολου προβλήματος που περιλάμβανε και χρήση μηχανής αναζήτησης στο διαδίκτυο δόθηκε σε ένα δείγμα N=605 ατόμων και μετρήθηκε ο χρόνος X=Χρόνος επίλυσης (τ.μ. X). Από το δείγμα εκτιμήθηκε το εύρος β=132 min των τιμών της τ.μ. X. Το εύρος αυτό χωρίστηκε σε 11 τμήματα όλα μήκους w=12 min. Οι συχνότητες εμφάνισης των τιμών της X στα 11 τμήματα δίνονται στον παρακάτω Πίνακα 6.

Να ευρεθεί η εκτιμήτρια της σ.π.π. από όλο το δείγμα. Μετά να εξεταστεί αν μπορεί να βρεθεί εκτιμήτρια σ.π.π. κατά κλάδο αφού διαπιστωθεί ότι μπορεί να θεωρηθεί ότι έχουμε δύο κλάδους. Τέλος να γίνει σύγκριση των δύο μεθόδων.

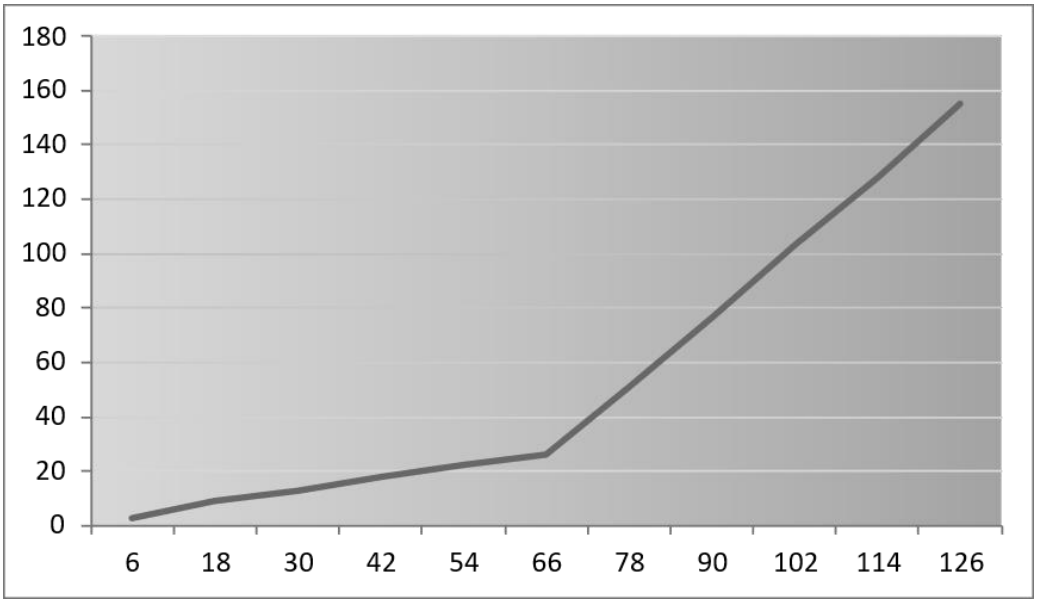
Πίνακας 6

Τάξεις	**Κέντρα τάξεων x_i**	**Συχνότητες n_i**	N_i	θ_i
[0, 12)	6	3	3	**0.82**
[12, 24)	18	9	12	**4.70**
[24, 36)	30	13	25	**11.34**
[36, 48)	42	18	43	**20.38**

[48, 60)	54	22	65	**31.63**
[60,72)	66	26	91	**44.96**
[72, 84)	78	51	142	**60.25**
[84, 96)	90	77	219	**77.44**
[96, 108)	102	103	322	**96.46**
[108, 120)	114	128	450	**117.25**
[120, 132]	126	155	605	**139.77**
Σύνολα		605=*n*		**605.00**

Λύση: Κατασκευάζουμε πρώτα το ιστόγραμμα συχνοτήτων της τ.μ. *X* με βάση τα στοιχεία του Πίνακα 6:

Σχήμα 2 – Πολύγωνο Συχνοτήτων



Οι τάξεις αντιπροσωπεύονται από τα κέντρα τους. Έτσι βλέπουμε ότι διαμορφώνονται δύο κλάδοι για το πολύγωνο συχνοτήτων: Ο ένας περιλαμβάνει τις 6 πρώτες τάξεις με άθροισμα συχνοτήτων N_A=91 και ο δεύτερος τις 5 επόμενες αντίστοιχο άθροισμα N_B=514. Το πολύγωνο συχνοτήτων είναι μια τεθλασμένη γραμμή με δύο «πλευρές»-κλάδους.

Αρχικά εκτιμούμε την σ.π.π. της τ.μ. *X* από όλο το δείγμα θεωρώντας ότι η γωνία μεταξύ των δύο κλάδων είναι αρκετά μικρή και ότι η μέθοδος της σχέσης (3) θα αποδώσει αξιόπιστη σ.π.π. ενός κλάδου και με αύξουσα τάση. Η σ.π.π. που προέκυψε δίνεται από την (11) που έπεται:

$$f(x)=\begin{cases}0.0250557\cdot\left(\dfrac{x}{132}\right)^{1.7561}, & x\in[0,132]\\ 0, & x\notin[0,132]\end{cases} \qquad (11)$$

Με τη βοήθεια της σ.π.π. (11) υπολογίσαμε θεωρητικά μεγέθη θ_i, i=1,2,…,11 των 11 τάξεων του δείγματος και η δοκιμασία X^2 καλής προσαρμογής των n_i στα προκύψαντα θ_i. Ο δειγματικός συντελεστής είναι X^2=23.6063 ενώ η κρίσιμη τιμή του για τους 7 β.ε. (μετά από σύμπτυξη των δύο πρώτων κλάσεων) και σ.σ.=α=0.05 είναι $X^2(7;0.05)$=14.0671. Άρα δεν έχουμε καλή προσαρμογή και η (11) δεν δίνει καλή εκτίμηση για την σ.π.π. σχετική με τα δεδομένα του Πίνακα 6.

Εξετάζουμε τώρα αν μπορούμε να βρούμε μία σ.π.π. με δύο κλάδους και με σημείο τομής που αντιστοιχεί στην τιμή της τ.μ. X (τετμημένη) $x_{κρ}$=72 όπως προκύπτει από συσχέτιση του Πίνακα 6 και του Σχήματος 2.

Ο κλάδος Α θα προκύψει με τη βοήθεια του Πίνακα 7 αντίστοιχη σ.π.π. με την (11).

Πίνακας 7

Τάξεις	**Κέντρα τάξεων x_i**	**Συχνότητες n_i**	$N_{i,A}$	θ_i
[0, 12)	6	3	3	**3.3**
[12, 24)	18	9	12	**8.6**
[24, 36)	30	13	25	**13.3**
[36, 48)	42	18	43	**17.7**
[48, 60)	54	22	65	**22.0**
[60, 72]	66	26	91	**26.1**
Σύνολα		91=N_A		**91.0**

Η σ.π.π. ου κλάδου Α είναι η

$$f(x)=\begin{cases}0.000673x^{0.8521}, & x\in[0,72]\\ 0, & x\notin[0,72]\end{cases}. \qquad (12)$$

Αντίστοιχα για τον κλάδο Β έχουμε τα εξής:

1ο) Η σχέση (3) που θα βασιστούμε υποδεικνύει ότι η καμπύλη της σ.π.π. περνάει από την αρχή $O(0,0)$ ή έστω εφάπτεται κύκλου (O, ρ) με ακτίνα ρ απειροστό σε σχέση με τα δεδομένα του προβλήματος (συχνότητες, εύρος κατανομής κλπ.). Άρα θα εργαστούμε με τα στοιχεία των 5 τελευταίων τάξεων του Πίνακα 6 αλλά θα

κάνουμε μετασχηματισμό και στους δύο άξονες με τμ $X'=X$-72 και με συχνότητες $n_i'=n_i$-48. Το 48 είναι ανάμεσα στις συχνότητες της 6ης και της 7ης και πιο κοντά στο 51 της 7ης τάξης ώστε το υπόλοιπο να είναι κοντά στο 0 για να μπορεί να υποτεθεί ότι η καμπύλη της σ.π.π. περνάει από την αρχή $O(0,0)$.

2ο) Η σ.π.π. θα προκύψει από τα στοιχεία του Πίνακα 8 και με βάση τη σχέση (3) και μετά θα παίρνεται υπόψη ο μετασχηματισμός επί των 2 αξόνων. Αυτό σημαίνει ότι η (3) θα πάρει τη μορφή:

$$f(x)=\begin{cases} h\left(\dfrac{x-72}{\beta-72}\right)^{\nu}, & x\in[72,\beta] \\ 0, & x\notin[72,\beta] \end{cases},\ \nu=-2+\sqrt{1+q},\ h=\frac{\nu+1}{\beta-72}, w=12 \qquad (13)$$

Η σχέση (13) θα πάρει τελικά τη μορφή:

$$f(x)=\begin{cases} 0.040228\cdot\left(\dfrac{x-72}{60}\right)^{0..7069}, & x\in[72,132] \\ 0, & x\notin[72,132] \end{cases}. \qquad (14)$$

Μετά θα υπολογίσουμε τα θεωρητικά μεγέθη $\theta_i=\theta_i'+48$ όπου το θ_i' θα προκύπτει από ολοκλήρωση της σ.π.π. της (14) με μέγεθος δείγματος $N_B'=274$ που προκύπτει από το Β κλάδο μετά από την μείωση όλων των συχνοτήτων (που είναι 5) κατά 48, όπου $n_i'=n_i$-48 είναι ο μετασχηματισμός. Τα θεωρητικά μεγέθη είναι στην τελευταία στήλη του Πίνακα 8.

Πίνακας 8

Τάξεις	**X-72**	**Κέντρα τάξεων x_i**	**Συχνότητες n_i**	n_i'	θ_i
[72, 84)	[0, 12)	6	51	3	**53.6**
[84, 96)	[12, 24)	18	77	29	**72.5**
[96, 108)	[24, 36)	*30*	103	55	**97.8**
[108, 120)	[36, 48)	42	128	80	**128,0**
[120, 132]	[48, 60]	54	155	107	**162.1**
Σύνολα			514	274	**514.0**

Συνοπτικά τα αποτελέσματα στον Πίνακα 9 ενιαίο και για τους δύο κλάδους:

Πίνακας 9

Τάξεις	**Κέντρα τάξεων x_i**	**Συχνότητες n_i**	N_i	θ_i
[0, 12)	6	3	3	**3.3**
[12, 24)	18	9	12	**8.6**
[24, 36)	30	13	25	**13.3**
[36, 48)	42	18	43	**17.7**
[48, 60)	54	22	65	**22.0**
[60,72)	66	26	91	**26.1**
[72, 84)	78	51	142	53.6
[84, 96)	90	77	219	72.5
[96, 108)	102	103	322	97.8
[108, 120)	114	128	450	128.0
[120, 132]	126	155	605	162.1
Σύνολα		605		605.0

Η δοκιμασία X^2 έδωσε δειγματικό στατιστικό X^2=1.0510<< $X^2(8;0.05)$=15.5073, δηλαδή έχουμε καλή προσαρμογή των παρατηρούμενων στα θεωρητικά μεγέθη συχνοτήτων. Το Στατιστικό αυτό σε σχέση με το αντίστοιχο προηγούμενο που βρήκαμε για την ενιαίου κλάδου λύση, το X^2=23.6063, είναι περίπου 23 φορές μικρότερο.

Συμπεράσματα σχετικά με τεθλασμένη συνεχής κατανομή.

Από τα προηγούμενα προκύπτει ότι μπορούμε μέσα από ομαδοποίηση των δειγματικών δεδομένων να έχουμε σ.π.π. σύνθετης μορφής. Η ομαδοποίηση γίνεται μετά από διαπίστωση ότι σε κάποιο σημείο $X=x_{κρ}$ η καμπύλη της κατανομής, σ.π.π., αλλάξει «απότομα» κλίση δηλαδή οι από δεξιά και από αριστερά παράγωγοι δεν έχουν την ίδια τιμή δηλαδή στο $x_{κρ}$ η καμπύλη δεν είναι λεία Μωυσιάδης (2000).

Μετά την ομαδοποίηση εργαζόμαστε χωριστά στον κάθε ένα από τους δύο κλάδους Α και Β του δείγματος. Οι δύο κλάδοι αντιπροσωπεύουν δύο διαφορετικές κατανομές. Η μορφή αυτή της κατανομής περιγράφει (εκτιμάει) πολύ καλύτερα την υφιστάμενη κατανομή και αυτό διαπιστώνεται από διάφορες δοκιμασίες όπως είναι (π.χ.) η δοκιμασία X^2. Στο δεύτερο παράδειγμα φαίνεται καθαρά η υπεροχή της μεθόδου κατά κλάδους έναντι της ενιαίας αντιμετώπισης όλου του δείγματος. Αυτό προέκυψε από το γεγονός ότι στη μεν μέθοδο κατά κλάδους η δοκιμασία X^2 δείχνει ότι έχουμε (πολύ) καλή προσαρμογή θεωρητικών και παρατηρούμενων συχνοτήτων

στην δε ενιαία αντιμετώπιση του δείγματος η ίδια δοκιμασία δείχνει μη καλή προσαρμογή, άρα δεν εφαρμόζεται στην περίπτωση αυτή η μέθοδος ενιαίας αντιμετώπισης. Στο πρώτο παράδειγμα και οι δύο διαδικασίες (ενιαία και κατά κλάδους) έδωσαν για το αντίστοιχο δείγμα καλή προσαρμογή. Η κατά κλάδους διαδικασία όμως έχει τιμή X^2 χίλιες και πλέον φορές μικρότερη, από την αντίστοιχη τιμή X^2 που παρατηρήθηκε στην ενιαία διαδικασία, δηλαδή μπορεί να ειπωθεί ότι έχουμε περίπου χίλιες και πλέον φορές μικρότερη απόκλιση από την πραγματικότητα από την απόκλιση που έχουμε κατά την ενιαία αντιμετώπιση του δείγματος.

Συζήτηση, Προοπτική

Η προσπάθεια να εξεταστούν κατανομές με τεθλασμένη μορφή της αντίστοιχης σ.π.π. αποτελεί συνέχεια της προσπάθειας που άρχισε το έτος 2000. Οι ως τώρα δημοσιεύσεις έχουν αντικείμενο συμμετρικές κατανομές ή κατανομές ενιαίου κλάδου farmakis(2003, 2010), Παπατσούμα (2018), Φαρμάκης(2015). Τώρα μπαίνουμε σε δεύτερο στάδιο όπου έχουμε τις κατανομές 2 κλάδων. Εξετάσαμε με τα παραδείγματα μας δύο περιπτώσεις, ενώ υπάρχουν έξι τουλάχιστον διαφορετικές περιπτώσεις με βάση τις κλίσεις των δύο κλάδων.

Ένας σχεδιασμός για τη μελέτη των υπολοίπων τεσσάρων περιπτώσεων της τεθλασμένης κατανομής δύο κλάδων είναι αυτονόητο καθήκον για την ομάδα των τριών υποφαινομένων συν-συγγραφέων τουλάχιστον.

ABSTRACT

Sampling can contribute to the study of the distribution of a random variable with very good results in terms of accuracy and especially speed. From the sample and through a process, we arrive from various parameters of the random variable X to the estimation of the probability density function (p.d.f.) of X. This process is usually theoretical and is based on an initial assumption about the form that p.d.f. can take or on its approach which can be efficient and easy to manage at the same time. In this paper, we deal with continuous distributions where the curve of the p.d.f is not a smooth line at one point. If the two branches of the curve are straight segments, the curve has the shape of a broken line.In general we use the term of zigzag line, even if the branches are not straight segments.In this case, the two branches have different slope at the point where the curve is not a smooth line. How we reach to the estimator of the p.d.f. in minimum time from the sample, compared to classical methods of determining the form of the p.d.f., is getting understandable with appropriate examples.

Key words: Sampling, probability density function, mean, variance, Coefficient of Variation, zigzag line

MSC: 62D05; 62E17

ΑΝΑΦΟΡΕΣ

Μωυσιάδης Π. (2000), *«Ανώτερα Μαθηματικά»,* Εκδόσεις Α & Π Χριστοδουλίδη Ο.Ε., Θεσσαλονίκη.
Παπατσούμα Ι. (2018), *«Συμβολή στη Δειγματοληπτική Ανάδειξη Μοντέλων Κατανομών με χρήση του Συντελεστή Μεταβλητότητας»*, Διδακτορική. Διατριβή, Τμ. Μαθηματικών Α.Π.Θ.
Φαρμάκης Ν. (2001), *«ΣΤΑΤΙΣΤΙΚΗ, Περιληπτική Θεωρία, Ασκήσεις»*, Εκδόσεις Α & Π Χριστοδουλίδη Ο.Ε., Θεσσαλονίκη.
Φαρμάκης Ν. (2009), *«Δημοσκοπήσεις & Δεοντολογία»*, Εκδόσεις Α & Π Χριστοδουλίδη Ο.Ε., Θεσσαλονίκη.
Φαρμάκης Ν. (2015), *«Δειγματοληψία και εφαρμογές»,* Ελληνικά Ακαδημαϊκά Ηλεκτρονικά Συγγράμματα & Βοηθήματα, Αθήνα ISBN: 978-960-603-093-2.
Φαρμάκης Ν. (2016), *«Εισαγωγή στη Δειγματοληψία»,* Αφοί Κυριακίδη, Εκδόσεις Α.Ε., Θεσσαλονίκη.
Cochran W. (1977), "*Sampling Techniques*", John Wiley & Sons, Inc, New York, London, Sydney, Toronto.
Farmakis N. (2003), "Estimation of Coefficient of Variation: Scaling of Symmetric Continuous Distributions", *STATISTICS in TRANSITION,* Vol. **6**, No 1, pp 83-96 & Website: http://www.stat.gov.pl/english/transition.htm.
Farmakis N. (2010), "Coefficient of Variation: Connecting Sampling with some Increasing Distribution Models" *Proceedings of Stochastic Modelling Techniques and Data Analysis International Conference* (*SMTDA-2010*), June 8 - 11, 2010 Chania Crete, Greece.

"Οι συγγραφείς ευχαριστούν την Επιτροπή Πρακτικών του Συνεδρίου και ειδικά τον κριτή για τις εποικοδομητικές παρατηρήσεις και προτάσεις"

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά του 33ου Πανελλήνιου Συνεδρίου Στατιστικής (2021), pp. 209–217

Τεχνικές Παλινδρόμησης για την Πρόβλεψη των Ποσοστιαίων Σταδίων Ανάπτυξης Καρπών

Ι. Οικονομίδης[1]***, Σ. Τρέβεζας***[1]
[1]Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
{goikon, strevezas}@math.uoa.gr

ΠΕΡΙΛΗΨΗ

Στην εργασία αυτή εξετάζεται το πρόβλημα της πρόβλεψης ποσοστών σταδίων ανάπτυξης σε καλλιέργειες, με εφαρμογή σε πραγματικά δεδομένα από καλλιέργειες καλαμποκιού στις ΗΠΑ. Κατασκευάζεται ένα μοντέλο λογιστικής παλινδρόμησης, το οποίο χρησιμοποιεί τρεις προβλεπτικούς παράγοντες, τον θερμικό χρόνο, τον υετό και ένα δείκτη βλάστησης. Ο θερμικός χρόνος και ο υετός υπολογίζονται με δεδομένα από μετεωρολογικούς σταθμούς σε όλη την έκταση ενδιαφέροντος, ενώ ο δείκτης βλάστησης μετράται από τον δορυφορικό αισθητήρα MODIS. Η απόδοση των μοντέλων μετράται με το RMSPE.

Λέξεις Κλειδιά: Λογιστική Παλινδρόμηση, Στάδια Ανάπτυξης Καλλιεργειών, Δείκτης Βλάστησης, Θερμικός Χρόνος, Υετός

1. Εισαγωγή

Η παγκόσμια αύξηση του ανθρώπινου πληθυσμού που παρατηρείται τις τελευταίες δεκαετίες έχει αναμφισβήτητα δημιουργήσει πολλά θέματα προς συζήτηση, το κυριότερο από τα οποία είναι η εξασφάλιση επαρκούς ποσότητας τροφής. Σύμφωνα με τον Οργανισμό Γεωργίας και Τροφίμων των ΗΠΑ, ο παγκόσμιος πληθυσμός αναμένεται να φτάσει τα 9.7 δισεκατομμύρια μέχρι το 2050 (FAO, 2018). Η αύξηση αυτή κάνει επιτακτική την ανάγκη για αυτοματοποίηση του αγροτικού τομέα προκειμένου η παραγωγή σιτηρών να καλύψει τις παγκόσμιες ανάγκες.

Η αγροτική αυτοματοποίηση και η γεωργία ακριβείας μπορούν να πάρουν πολλές μορφές, όπως η παρακολούθηση των συστατικών του εδάφους και η προσανατολισμένη καταπολέμηση των ζιζανίων. Σε αυτήν την εργασία, ερευνούμε την παρακολούθηση της ανάπτυξης των φυτών σε καλλιέργειες μεγάλης κλίμακας. Ο κύκλος ζωής ενός φυτού χωρίζεται σε στάδια ανάπτυξης, ξεκινώντας από τη φύτευση (planting) και τελειώνοντας με τη συλλογή των καρπών (harvesting). Οι ανάγκες του φυτού αλλάζουν σημαντικά από στάδιο σε στάδιο, επομένως η γνώση του σταδίου στο οποίο βρίσκεται ένα φυτό είναι ζωτικής σημασίας για την ορθή και έγκαιρη παρέμβαση (λίπανση, πότισμα κ.ο.κ.). Σε μεγάλες καλλιέργειες, τα φυτά συνυπάρχουν σε διαφορετικά στάδια. Στόχος της εργασίας αυτής είναι η ακριβής πρόβλεψη του

ποσοστού των φυτών που βρίσκονται σε κάθε στάδιο, σε πραγματικό χρόνο, χρησιμοποιώντας μετεωρολογικά δεδομένα και τεχνολογίες τηλεπισκόπησης.

1.1 Η Φαινολογία του Καλαμποκιού

Τα στάδια ανάπτυξης του καλαμποκιού μπορούν να χωριστούν σε δύο κατηγορίες, βλαστικά (vegetative) και αναπαραγωγικά (reproductive). Στην παρούσα έρευνα τα στάδια που μελετώνται είναι επτά: planted (φύτευση), emerged (φύτρωμα), silking (μετάξωμα), dough (σκλήρυνση καρπών), dented (κοίλωμα καρπών), mature (ωρίμανση) και harvested (συλλογή καρπών). Επιπλέον, εισάγουμε ένα αρχικό στάδιο (preseason) ώστε να συγχρονίσουμε δεδομένα από διαφορετικές χρονιές και τα ποσοστά να αθροίζουν στη μονάδα σε κάθε χρονική στιγμή.

Τα δεδομένα που μελετάμε στην παρούσα έρευνα προέρχονται από την Εθνική Αγροτική Στατιστική Αρχή (National Agricultural Statistical Service) του Υπουργείου Αγροτικής Ανάπτυξης των ΗΠΑ (United States Department of Agriculture). Κατά την καλοκαιρινή περίοδο, διενεργούνται εβδομαδιαίες έρευνες σχετικά με τον πρόοδο και την κατάσταση των αγροκαλλιεργειών. Η έρευνα που αφορά τη συγκεκριμένη μελέτη είναι οι Αναφορές Προόδου των Καρπών (Crop Progress Reports, CPR), στις οποίες οι γεωργοί αξιολογούν το ποσοστό των φυτών που βρίσκονται σε κάθε φαινολογικό στάδιο ανάπτυξης. Τα δεδομένα της έρευνας αφορούν το μέσο όρο σε επίπεδο πολιτείας. Τα CPR περιλαμβάνουν δεδομένα για τα επτά στάδια που αναφέραμε (USDA, 2020).

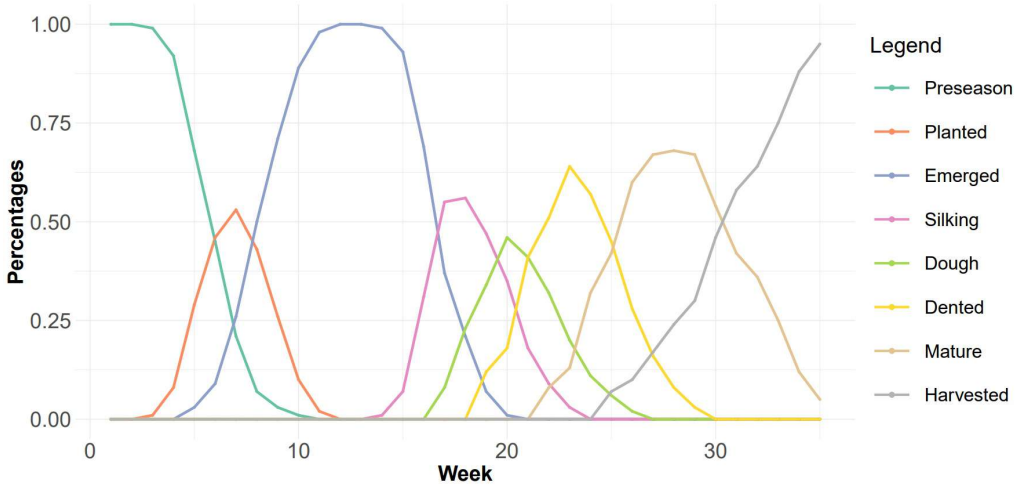
Το USDA έχει δημιουργήσει έναν αγροτικό χάρτη με τη θέση και τον καρπό που καλλιεργείται σε κάθε κτήμα (Cropland Data Layer, CDL). Ο χάρτης αυτός επιτρέπει την παρακολούθηση των καλλιεργειών μέσα από μετεωρολογικούς σταθμούς και δορυφόρους. Στην παρούσα έρευνα έχουμε κατεβάσει και επεξεργαστεί τα δεδομένα με το λογισμικό R (R Core Team, 2021) και συγκεκριμένα με το πακέτο cdlTools (Chen and Lisic, 2018; USDA-NASS, 2019).

Η παρούσα έρευνα αφορά τις καλλιέργειες καλαμποκιού στην Nebraska. Το χρονικό διάστημα αφορά τις καλοκαιρινές καλλιέργειες από την 13η έως την 47η εβδομάδα του χρόνου (Απρίλιος - Νοέμβριος), για 18 χρόνια (2002 - 2019). Το Σχήμα 1 δείχνει τα CPR για το 2002.

1.2 Δείκτες Βλάστησης

Η τηλεπισκόπηση έχει επιτρέψει την έγκυρη και έγκαιρη παρακολούθηση των αγροκαλλιεργειών σε εκτάσεις μεγάλης κλίμακας. Η τεχνολογία αυτή βασίζεται στη μέτρηση της ανακλώμενης ακτινοβολίας από την επιφάνεια της Γης, η οποία συνοψίζεται στους δείκτες βλάστησης (Vegetation Indices, VI). Οι δείκτες βλάστησης είναι συναρτήσεις της ορατής κόκκινης και υπέρυθρης ακτινοβολίας, σχεδιασμένες να εκτιμούν τα επίπεδα χλωροφύλλης των φυτών. Ο καθιερωμένος δείκτης βλάστησης είναι αυτός της κανονικοποιημένης διαφοράς (Normalized Difference Vegetation Index, NDVI) (Rouse et al., 1974).

Έστω ρ_{RED} και ρ_{NIR} το μέσο ποσοστό ανακλώμενης ακτινοβολίας στα μήκη κύ-



***Σχήμα 1**: CPR για την πολιτεία της Νεμπράσκα, εβδομάδες 13-47, 2002. Το πλήρες σύνολο δεδομένων περιλαμβάνει τα έτη 2002 - 2019. Δημιουργήθηκε με το πακέτο ggplot2 (Wickham, 2016) της R.*

ματος του ορατού κόκκινου ($620-670$ nm) και του υπέρυθρου ($841-876$ nm) φωτός, αντίστοιχα. Τότε, ο δείκτης βλάστησης κανονικοποιημένης διαφοράς (NDVI) δίνεται από τη σχέση:

$$\text{NDVI} := \frac{\rho_{\text{NIR}} - \rho_{\text{RED}}}{\rho_{\text{NIR}} + \rho_{\text{RED}}},$$

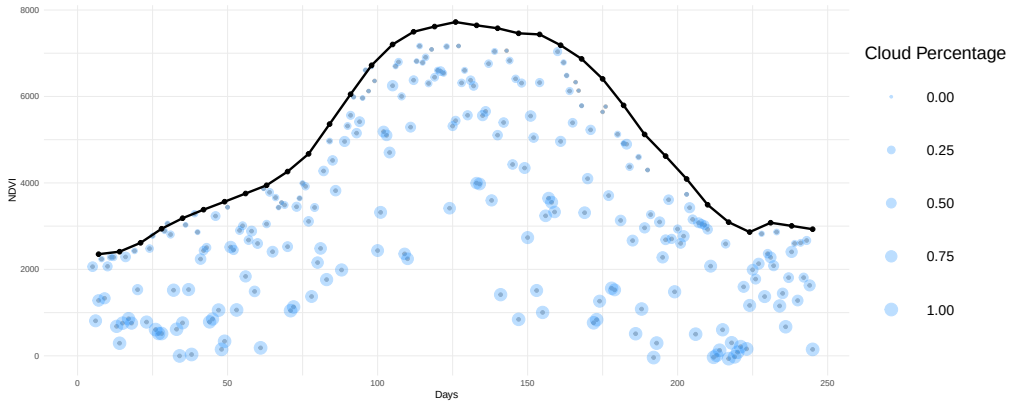
όπου $\rho_{\text{RED}}, \rho_{\text{NIR}} \in [0, 1]$. Το NDVI παίρνει τιμές από το -1 μέχρι το 1. Χαρακτηριστικό της χλωροφύλλης είναι η ανάκλαση του υπέρυθρου και η απορρόφηση του ορατού κόκκινου φωτός, κάτι που οδηγεί το NDVI να πάρει τιμές κοντά στη μονάδα. Αντίθετα, μη-καλλιεργήσιμες περιοχές (χώμα, νερό κ.ο.κ.) παίρνουν πολύ χαμηλότερες τιμές NDVI.

Το μεγάλο μειονέκτημα των δορυφορικών δεδομένων και ειδικά στο φάσμα του ορατού φωτός είναι ο θόρυβος που προκαλείται από τα σύννεφα, τα οποία εμποδίζουν την παρακολούθηση των καλλιεργειών. Για τον λόγο αυτό, τα δορυφορικά δεδομένα συχνά έρχονται με συμπληρωματικές ενδείξεις των νεφώσεων, ώστε να μπορεί να γίνει καθαρισμός των δεδομένων. Στην παρούσα έρευνα, έχοντας δεδομένα τόσο για τη θέση των υπό μελέτη κτημάτων, όσο και της νεφελότητας, είναι δυνατό να καθαρίσουμε τα δεδομένα κρατώντας πληροφορία μόνο από τα κτήματα για τα οποία ο δορυφόρος έχει καθαρή εικόνα. Η επεξεργασία των δορυφορικών δεδομένων έγινε στην R και συγκεκριμένα με τα πακέτα MODIStsp και raster (Busetto and Ranghetti, 2016; Hijmans, 2021).

Για την παρακολούθηση των αγροκαλλιεργειών χρησιμοποιήθηκε ο αισθητήρας MODIS του δορυφόρου Terra (NASA, 2002). Τα δορυφορικά δεδομένα είναι ημερήσια, ενώ τα δεδομένα για τα στάδια ανάπτυξης είναι εβδομαδιαία. Αυτό επιτρέπει έναν ακόμα τρόπο διόρθωσης των δεδομένων, επιλέγοντας τη μέγιστη τιμή του δεί-

κτη βλάστησης για κάθε κτήμα μέσα στην εβδομάδα. Τέλος υπολογίζεται η μέση τιμή του δείκτη όλων των κτημάτων για την εβδομάδα, καταλήγοντας σε μία χρονοσειρά εβδομαδιαίων παρατηρήσεων του NDVI. Στην παρούσα έρευνα, τα δεδομένα έχουν επιπλέον εξομαλυνθεί χρησιμοποιώντας το φίλτρο Savitzky-Golay (Savitzky and Golay, 1964), με παραμέτρους $d = 3$ (βαθμός πολυωνύμου) και $m = 7$ (εύρος παραθύρου εξομάλυνσης) με το πακέτο spatialEco της R (Evans, 2021). Η επιλογή των παραμέτρων βασίστηκε σε σύγκριση της απόδοσης του τελικού μοντέλου για τους συνδυασμούς των τιμών $d = 2, 3, 4$ και $m = 5, 7, 9$. Η προβλεπτική ικανότητα των 9 μοντέλων ήταν παρόμοια, με τον συνδυασμό $d = 3$ και $m = 7$ να παράγει τα καλύτερα αποτελέσματα.

Το Σχήμα 2 δείχνει το αρχικό, ακαθάριστο μέσο NDVI κάθε ημέρας (μπλε κύκλοι με ακτίνα ανάλογη της νεφελότητας) για το 2002. Το τελικό, καθαρισμένο και εξομαλυμένο NDVI συμβολίζεται με μία συνεχή μαύρη γραμμή.



Σχήμα 2: *Το αρχικό, ακαθάριστο μέσο NDVI κάθε ημέρας (μπλε κύκλοι με ακτίνα ανάλογη της νεφελότητας) για το 2002. Το τελικό, καθαρισμένο και εξομαλυμένο NDVI συμβολίζεται με μία συνεχή μαύρη γραμμή. Δημιουργήθηκε με το πακέτο ggplot2 (Wickham, 2016) της R.*

1.3 Θερμικός Χρόνος

Η θερμοκρασία είναι ο πιο σημαντικός περιβαλλοντικός παράγοντας στην ανάπτυξη των φυτών. Η εξάρτηση αυτή μπορεί να μοντελοποιηθεί εύκολα με την έννοια των θερμικών ημερών ανάπτυξης (Growing Degree Days - GDD), ένα μέτρο της ημερήσιας θερμικής συσσώρευσης (Gilmore and Rogers, 1958).

Η έννοια του GDD πηγάζει από το γεγονός ότι ο ρυθμός ανάπτυξης ενός φυτού γίνεται βέλτιστος σε μία συγκεκριμένη θερμοκρασία T_o (Temperature optimal), ενώ μηδενίζεται εκτός συγκεκριμένων ορίων που θέτουν μία κάτω και μία άνω θερμοκρασία T_b, T_c (Temperature base, ceiling), αντίστοιχα. Για το καλαμπόκι οι καθιερωμένες θερμοκρασίες είναι $T_b = 10^oC, T_o = 30^oC$ και $T_c = 50^oC$. Έτσι, διορθώνοντας τη μέση ημερήσια θερμοκρασία ώστε να βρίσκεται εντός των παραπάνω ορίων, το GDD_t

μίας ημέρας t μπορεί να υπολογιστεί. Αθροίζοντας τις θερμικές ημέρες ανάπτυξης, μπορούμε να πάρουμε το συνολικό θερμικό χρόνο ζωής ενός φυτού έως την ημέρα t, AGDD_t (Accumulated Growing Degree Days). Συμβολίζοντας με $T_{min}(t)$ και $T_{max}(t)$ την ελάχιστη και μέγιστη θερμοκρασία της ημέρας t, αντίστοιχα, έχουμε

(i) τη θερμική ημέρα ανάπτυξης για την ημέρα t, GDD_t, που ορίζεται ως

$$\text{GDD}_t := \frac{T^*_{max}(t) + T^*_{min}(t)}{2} - T_b,$$

όπου $T^*_{max}(t) = \min\{T_{max}(t), T_o\}$ και $T^*_{min}(t) = \max\{T_{min}(t), T_b\}$,

(ii) το θερμικό χρόνο ζωής μέχρι την ημέρα t, AGDD_t, που ορίζεται ως

$$\text{AGDD}_t := \sum_{i=1}^{t} \text{GDD}_i.$$

Σημειώνεται ότι από τον ορισμό του $T^*_{max}(t)$, θερμοκρασίες ανώτερες της βέλτιστης (που σπάνια παρατηρούνται στο παρόν σύνολο δεδομένων) αντιστοιχίζονται στον ίδιο θερμικό χρόνο με αυτή.

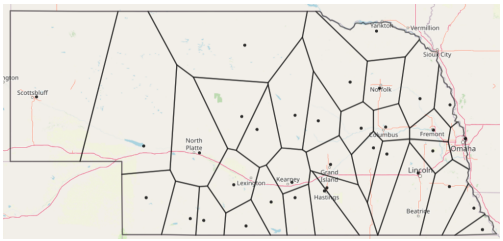
1.4 Βροχόπτωση

Ένας πολύ σημαντικός παράγοντας για την ανάπτυξη των φυτών είναι ο υετός. Ακραίες συνθήκες βροχόπτωσης και χιονόπτωσης είναι δυνατόν να βλάψουν ή και να καταστρέψουν ολόκληρες καλλιέργειες. Στην παρούσα έρευνα, ο υετός μοντελοποιείται ακριβώς όπως και ο θερμικός χρόνος. Συμβολίζουμε με P_t τον υετό της ημέρας t, εκφρασμένο σε χιλιοστά (mm) κάλυψης του εδάφους. Ο συνολικός υετός μέχρι την ημέρα t, AP_t, ορίζεται ως

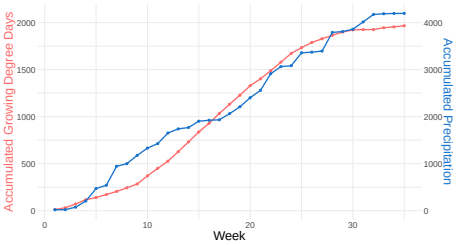
$$\text{AP}_t := \sum_{i=1}^{t} \text{P}_i.$$

1.5 Πολύγωνα Thiessen

Στην παρούσα έρευνα, τα δεδομένα της θερμοκρασίας και του υετού προέρχονται από 35 μετεωρολογικούς σταθμούς στην Νεμπράσκα. Προκειμένου να συνοψιστούν τα δεδομένα σε μία τιμή αντιπροσωπευτική για όλη την πολιτεία, χρησιμοποιούμε τα πολύγωνα Thiessen, που χωρίζουν την πολιτεία σε πολύγωνα με όρια που ισαπέχουν από δύο γειτονικούς σταθμούς. (Σχήμα 3). Κάθε κτήμα αντιστοιχίζεται στον σταθμό του πολυγώνου στο οποίο ανήκει (στον κοντινότερο μετεωρολογικό σταθμό σε αυτό). Έτσι, μπορεί να υπολογιστεί ένας σταθμισμένος μέσος, όπου τα κανονικοποιημένα βάρη είναι ανάλογα του αριθμού των κτημάτων κάθε πολυγώνου (Σχήμα 4). Η ανάλυση έγινε στην R με το πακέτο SDraw (McDonald and McDonald, 2020).



Σχήμα 3: *Πολιτικός χάρτης της Νεμπράσκα, διαμερισμένος σε πολύγωνα Thiessen. Οι 35 μετεωρολογικοί σταθμοί συμβολίζονται με μαύρα σημεία. Δημιουργήθηκε με το πακέτο leaflet (Graul, 2016) της R.*



Σχήμα 4: *AGDD (κόκκινο) και AP (μπλε) για το έτος 2002. Και οι δύο παράγοντες είναι στάθμιση 35 μετεωρολογικών σταθμών, με βάρη ανάλογα των κτημάτων κάθε πολυγώνου. Δημιουργήθηκε με το πακέτο ggplot2 (Wickham, 2016) της R.*

2. Μοντελοποίηση

Σκοπός της παρούσας έρευνας είναι η πρόβλεψη των εβδομαδιαίων ποσοστών σταδίων ανάπτυξης (CPR) μέσω του ημερολογιακού και θερμικού χρόνου (Week, AGDD), του υετού (AP) και του δείκτη βλάστησης κανονικοποιημένης διαφοράς (NDVI). Για τη μοντελοποίηση του προβλήματος κάνουμε χρήση των δεικτών $y \in Y$ για το έτος, $w \in W$ για την εβδομάδα και $s \in S$ για το στάδιο. Έτσι, το παρατηρούμενο CPR για το έτος y, την εβδομάδα w και το στάδιο s συμβολίζεται με $\pi_{yw}(s)$.

Για την αξιολόγηση των μοντέλων θα χρησιμοποιηθεί η ρίζα του μέσου τετραγωνικού σφάλματος πρόβλεψης (Root Mean Squared Prediction Error, RMSPE), το οποίο εκτιμάται με Monte Carlo cross-validation (Stone, 1974; Geisser, 1975). Συγκεκριμένα, 13 από τα 18 χρόνια (2002-2019) επενδύονται στην εκπαίδευση (training) του μοντέλου και τα υπόλοιπα 5 στην αξιολόγηση (testing). Η διαδικασία αυτή επαναλαμβάνεται $K = 500$ φορές, με διαφορετικές διαμερίσεις των 18 ετών. Ως σημείο αναφοράς, χρησιμοποιούμε το μοντέλο του ιστορικού μέσου (historic mean) που επιλέγει ως πρόβλεψη για όλα τα χρόνια το μέσο εβδομαδιαίο CPR για κάθε μία από τις εβδομάδες των 13 ετών που περιλαμβάνονται στην εκπαίδευση του μοντέλου. Το μοντέλο αυτό θα είναι η βάση για την αξιολόγηση άλλων μοντέλων.

Για την εφαρμογή αυτή, θεωρούμε τη δείκτρια e_{ywi} του i-οστού φυτού ($e_{ywis} = 1$, αν το φυτό i βρίσκεται στο στάδιο s τη δεδομένη χρονική στιγμή και $e_{ywis} = 0$ διαφορετικά). Τότε, υπό δεδομένους περιβαλλοντικούς παράγοντες x_{yw}, οι δείκτριες ακολουθούν κατηγορική κατανομή, και υπό την υπόθεση της δεσμευμένης ανεξαρτησίας, το άθροισμα τους ακολουθεί πολυωνυμική κατανομή, δηλαδή

$$n\pi_{yw}|x_{yw} = \sum_{i=1}^{n} e_{ywi}|x_{yw} \sim \mathcal{M}(n, p_{yw}),$$

όπου π_{yw} είναι τα CPR για το έτος y και την εβδομάδα w. Επιλέγουμε να μοντελοποιήσουμε κάθε στάδιο ξεχωριστά, χρησιμοποιώντας τα αθροιστικά CPR $\pi^{\star}_{yw}$ τα οποία

δηλώνουν το ποσοστό των φυτών που βρίσκονται σε ένα στάδιο ή το έχουν ξεπεράσει. Προκύπτει άμεσα ότι

$$n\pi^{\star}_{yws}|x_{yws} \sim Bin\left(n, p^{\star}_{yws}\right),\ p^{\star}_{yws} = \sum_{k=s}^{S} p_{ywk},\ \ s = 1, \ldots, S,$$

Με βάση τα παραπάνω, δημιουργούμε ένα μοντέλο λογιστικής παλινδρόμησης για κάθε στάδιο ανάπτυξης (παρατηρούμε ότι για $s = 1$, το μοντέλο μπορεί να παραληφθεί, αφού τα αθροιστικά CPR είναι σταθερά και ίσα με τη μονάδα). Κατασκευάζουμε το γενικευμένο γραμμικό μοντέλο (Nelder and Wedderburn, 1972) της μορφής

$$p^{\star}_{yws} = F\left(x^{\top}_{yws}\beta_s\right),\ s = 2, \ldots, S,$$

όπου $F(x) = 1/(1 + e^{-x})$, η συνάρτηση λογιστικής κατανομής. Οι εκτιμήσεις των β_s (τα οποία έχουν διάσταση d, όπου $d-1$ είναι το πλήθος των προβλεπτικών παραγόντων του μοντέλου) προκύπτουν από τη συνάρτηση μερικής πιθανοφάνειας, λύνοντας τις εξισώσεις Score

$$\frac{\partial \ell\left(\beta_s\right)}{\partial \beta_{sj}} = \sum_{y\in Y}\sum_{w\in W}\left(\pi^{\star}_{yws} - p^{\star}_{yws}\right)x^{\top}_{ywsj} = 0, \qquad 1 \leq j \leq d,$$

οι οποίες μπορούν να επιλυθούν αριθμητικά για $s = 2, \ldots, 8$, χρησιμοποιώντας τη μέθοδο Newton-Raphson. Επιπρόσθετα, υπό συγκεκριμένες συνθήκες κανονικότητας, (Wooldridge, 2010), οι εκτιμήσεις μερικής πιθανοφάνειας $\widehat{\beta}_s$ του β_s είναι συνεπείς και ασυμπτωτικά κανονικές.

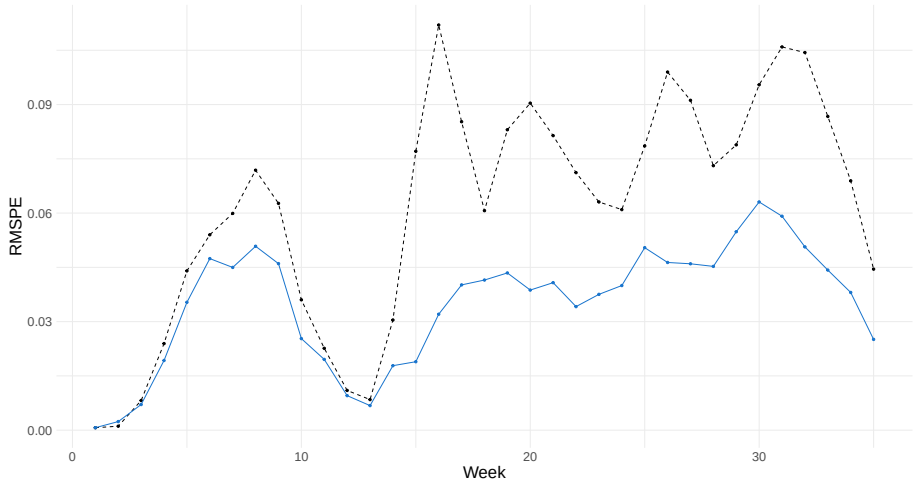
Η ξεχωριστή μοντελοποίηση κάθε σταδίου προσφέρει το πλεονέκτημα επιλογής των προβλεπτικών μεταβλητών. Συγκεκριμένα, ο ημερολογιακός και θερμικός χρόνος συμπεριλαμβάνονται σε όλα τα στάδια, ενώ ο υετός μόνο στα στάδια της φύτευσης, της φύτρωσης και της συλλογής καρπών (planted, emerged, harvested), όπου η αυξημένη βροχόπτωση αναμένεται να προκαλέσει προβλήματα στις καλλιέργειες. Τέλος, το NDVI μετρά την χλωροφύλλη, επομένως εισάγεται στα στάδια μετά την φύτρωση και πριν τη συλλογή καρπών (silking, dough, dented, mature). Έτσι, κάθε στάδιο περιλαμβάνει ακριβώς τρεις προβλεπτικούς παράγοντες.

3. Αποτελέσματα

Η εφαρμογή των μοντέλων λογιστικής παλινδρόμησης πάνω στα δεδομένα των 18 ετών (fitting) έδειξε τη στατιστική σημαντικότητα όλων των παραγόντων, σε όλα τα στάδια. Ενδιαφέρον παρουσιάζει το γεγονός ότι ο συντελεστής για τον υετό είναι αρνητικός για τα στάδια φύτευσης, φύτρωσης και συλλογής καρπών, επιβεβαιώνοντας την υπόθεση αρνητικής επίδρασης του υετού στις καλλιέργειες. Οι προβλεπτικοί παράγοντες ελέγχθηκαν για πολυσυγγραμμικότητα με τον Generalized Variance

Inflation Factor (Fox and Monette, 1992), όπου το καθιερωμένο όριο του 10 δεν παραβιάστηκε. Αναφέρουμε ότι η ίδια μεθοδολογία δοκιμάστηκε και για άλλες συνδετικές συναρτήσεις, αλλά η λογιστική έδωσε τα καλύτερα αποτελέσματα.

Η σύγκριση του RMSPE (Fig. 5) δείχνει ότι τα λογιστικά μοντέλα παράγουν καθολικά χαμηλότερα σφάλματα από το μοντέλο ιστορικού μέσου. Συγκεκριμένα, παρατηρούμε δύο περιόδους υψηλών σφαλμάτων, στην αρχή και το τέλος της καλλιεργητικής περιόδου, οι οποίες πηγάζουν σε μεγάλο βαθμό από τη φύτευση και τη συλλογή καρπών, στάδια που καθορίζονται κυρίως από ανθρώπινες δραστηριότητες.



Σχήμα 5: *Σύγκριση του RMPSE ανάμεσα στο μοντέλο ιστορικού μέσου (μαύρη διακεκκομένη γραμμή) και το μοντέλο λογιστικής παλινδρόμησης (μπλε γραμμή). Δημιουργήθηκε με το πακέτο ggplot2 (Wickham, 2016) της R.*

4. Σύνοψη

Στην παρούσα έρευνα εξετάστηκε ένα μοντέλο λογιστικής παλινδρόμησης με σκοπό την πρόβλεψη των ποσοστών σταδίων ανάπτυξης σε καλλιέργειες καλαμποκιού. Το μοντέλο αυτό έχει την ικανότητα να παράγει προβλέψεις σε πραγματικό χρόνο, κάτι ιδιαίτερα σημαντικό για την έγκαιρη παρακολούθηση των καλλιεργειών.

Σε επόμενη μελέτη θα θέλαμε να χρησιμοποιήσουμε δεδομένα από διαφορετικές πολιτείες των ΗΠΑ και να αναπτύξουμε ένα μοντέλο μεικτών επιδράσεων (mixed effects model), με σκοπό την μοντελοποίηση της μεταβλητότητας που προέρχεται από διαφορετικά έτη καθώς και διαφορετικές πολιτείες. Τέλος, ιδιαίτερο ενδιαφέρον παρουσιάζει η εφαρμογή με δεδομένα άλλων δορυφόρων και συγκεκριμένα τους δορυφόρους Sentinel του προγράμματος Copernicus της Ευρωπαϊκής Ένωσης.

ABSTRACT

This study concerns the problem of crop stage percentages estimation, presenting a case study on USA corn cultivations. A logistic regression model is developed, using three predictive

factors, thermal time, precipitation and one vegetation index. Thermal time and precipitation are calculated from meteorological stations, while the vegetation index is monitored via the satellite sensor MODIS. The performance of the model is evaluated with the RMSPE.

ΑΝΑΦΟΡΕΣ

L. Busetto and L. Ranghetti. Modistsp: an r package for preprocessing of modis land products time series. *Computers & Geosciences*, 97:40--48, 2016. ISSN 0098-3004. doi: 10.1016/j.cageo.2016.08.020.

L. Chen and J. Lisic. *Tools to Download and Work with USDA Cropscape Data*, 2018.

J. S. Evans. *spatialEco*, 2021. R package version 1.3-6.

FAO. The future of food and agriculture—alternative pathways to 2050, 2018.

J. Fox and G. Monette. Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87(417):178--183, 1992.

S. Geisser. The predictive sample reuse method with applications. *Journal of the American statistical Association*, 70(350):320--328, 1975.

J. E. Gilmore and J. Rogers. Heat units as a method of measuring maturity in corn 1. *Agronomy journal*, 50(10):611--615, 1958.

C. Graul. *leafletR: Interactive Web-Maps Based on the Leaflet JavaScript Library*, 2016. R package version 0.4-0.

R. J. Hijmans. *raster: Geographic Data Analysis and Modeling*, 2021. R package version 3.4-13.

T. McDonald and A. McDonald. *SDraw: Spatially Balanced Samples of Spatial Objects*, 2020. R package version 2.1.13.

NASA. Moderate resolution imaging spectroradiometer, 2002.

J. A. Nelder and R. W. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370--384, 1972.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

J. W. Rouse, R. H. Haas, J. A. Schell, D. W. Deering, and J. C. Harlan. Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation. *NASA/GSFC Type III Final Report, Greenbelt, Md*, 371, 1974.

A. Savitzky and M. J. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627--1639, 1964.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111--133, 1974.

USDA. Crop progress reports, 2020.

USDA-NASS. *National Agricultural Statistics Service Cropland Data Layer*, 2019.

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4.

J. M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

Εργασίες στα Αγγλικά

Papers in English

Greek Statistical Institute
Proceedings of the 33rd Panhellenic Statistics Conference (2021), pp.219-232

CUSUM CONTROL CHARTS FOR MONITORING BINARCH(1) PROCESSES

Maria Anastasopoulou, Athanasios Rakitzis
Lab of Statistics and Data Analysis, Department of Statistics and Actuarial-Financial Mathematics, University of the Aegean, 83200 Karlovasi, Samos, Greece
anastasopoulou@aegean.gr; arakitz@aegean.gr

ABSTRACT

In this work, we develop and study one-sided CUSUM control charts for monitoring correlated counts with finite range. Often in practice, data of that kind can be adequately described by a first-order binomial integer-valued ARCH model (or BINARCH(1)). The proposed charts are based on the likelihood ratio and can be used for detecting upward or downward shifts in process mean level. The general framework for the development and the practical implementation of the proposed charts is given. Using Monte Carlo simulation, we compare the performance of the proposed CUSUM charts with the corresponding one-sided Shewhart and EWMA charts for BINARCH(1) processes. A real-data application of the proposed charts in epidemiology is also discussed.

Keywords: Average run length, BINARCH(1) model, CUSUM, Statistical process control.

1. INTRODUCTION

Statistical process control (SPC) is a collection of statistical tools that focuses on the monitoring of a (stochastic) process and aims to detect changes in it. The most representative tool of this collection is the control chart, originally proposed by Walter A. Shewhart in the 1920's. The main area of the application of control charts is on the monitoring of industrial processes but nowadays, since processes become more and more complex, their use is not restricted in industry but also on several other areas of applied science (see, for example, [Bersimis et al. (2018)], [Woodall et al. (2017)].

There are two main categories of control charts: The variables charts (for continuous random variables) and the attributes charts (for discrete random variables). In this work, we focus on attributes control charts and specifically, we are interested in monitoring the number of nonconforming units, within a sample size *n*.

The usual control chart in this case is the Shewhart *np* chart [Montgomery (2009)]. These monitoring schemes are developed under the assumption that the number of

nonconforming units follows a binomial distribution B(n,π), where n is the sample size and π is the success probability. Moreover, an assumption when the *np* charts are applied in practice is that the successive counts are independent and identically distributed (i.i.d) binomial random variables (r.v).

However, when the assumption of serially independent counts of nonconforming items is violated, properly adjusted control charts have to be applied for process monitoring, instead of the abovementioned ones. According to Kim and Lee (2019) (see also [Psarakis and Papaleonida (2007)], the automation and advancement of quality in production processes, has made serially correlated processes of counts (or, integer-valued time series) very common in the modern manufacturing industry while the monitoring of these processes has received considerable attention from many researchers.

In this case, the *np* chart cannot be used because of the excessive false alarm rate. A solution to this problem is to select first an appropriate model of integer-valued time series and then to develop control charts based on this model.

In this category belongs the BAR(1) model, which was proposed by McKenzie in 1985. The BAR(1) model is a simple model and suitable for modelling autocorrelated binomial counts with structure similar to that of the usual AR(1) model. Moreover, it is suitable when the possible values of the process are finite, e.g. in the set $\{0, 1, \ldots, n\}$. Another popular choice is that of the first order beta-binomial autoregressive model (BBAR(1), Weiß and Kim (2014)). The BBAR(1) model is suitable for modeling correlated binomial counts with extra binomial variation, i.e., when the variation is larger compared to the variation under the binomial model.

Furthermore, the BBAR(1) model is suitable when the sample consists of non-homogeneous units. In the non-homogeneity case, the probability for a unit in the sample to be nonconforming is not the same for all the units in it. For example, the probability for a country to have monthly inflation rate below 2% is not the same among all the Eurozone members, because of the different socioeconomic structure and policy across the countries.

An additional model in the same category is the BINARCH(1) model. [Ristic et al. (2016)], [Weiß and Pollett (2014)]. The BINARCH(1) model has a dependence structure similar to that of the usual AR(1) model and it is appropriate when extra-binomial variation is present in the process.

Several control charts for monitoring correlated counts with finite range are available in the related literature. Weiß (2009) proposed and studied a moving average control chart and a runs based chart for monitoring a BAR(1) process. Rakitzis et al. (2017) studied one-sided (upper- and lower-sided) Shewhart charts and one-sided CUSUM charts for monitoring a BAR(1) and a BBAR(1) process. Recently, Anastasopoulou and Rakitzis (2020, 2022) proposed and studied upper and lower one-sided EWMA type charts for monitoring BAR(1) and BBAR(1) processes. To the best of our knowledge, CUSUM charts have not been studied in the case of

BINARCH(1) processes. Moreover, motivated by their well-known optimality property [Moustakides (1986)], we aim at investigating if this property is preserved in the case of BINARCH(1) processes, too.

This work is organized as follows: In Section 2, we briefly present the main properties of the BINARCH(1) model. In Section 3 we present the one-sided CUSUM charts for monitoring a BINARCH(1) process (Sections 3.1 and 3.2), as well as the performance measures for each chart (Section 3.3) and their statistical design (Section 3.4). Section 4 consists of the results of an numerical study on the performance of one-sided Shewhart, *s*-EWMA and CUSUM charts in the monitoring of BINARCH(1) processes. In Section 5, we provide an example for the practical implementation of the proposed charts in epidemiological process. Finally, conclusions are summarized in Section 6.

2. THE BINARCH(1) MODEL

The BINARCH(1) is a suitable model for autocorrelated processes of counts with a finite range. It was first proposed by Weiß and Pollett (2014) as a limiting case of the density dependent thinning models and recently, it was studied further by Ristic et al. (2016).

Let X_t, $t \geq 1$, be a sequence of serially dependent counts which take values in $\{0,1,...,n\}$, $n \in \square$. In the BINARCH(1) model the conditional distribution of $X_t \mid X_{t-1}, X_{t-2}, \ldots \sim B(n, \alpha_t)$, , $t \geq 1$, The $\alpha_t = a_0 + (a_1/n)X_{t-1}$, where $a_0 > 0$, $a_1 \geq 0$ and $a_0 + a_1 < 1$.

For the marginal distribution of X_t, the expected value $E(X_t)$, the variance $V(X_t)$ and the autocorrelation function $\rho(k) = Corr(X_t, X_{t-k})$ are, respectively, equal to

$$\mu \equiv E(X_t) = na_0/(1-a_1),\ \sigma^2 \equiv V(X_t) = \frac{\mu(1-\mu/n)}{1-(1-1/n)a_1^2},\ \rho(k) = a_1^k,\ k \geq 1.$$

Moreover, the transition probabilities $p_{j|i} = \left(X_t = j \mid X_{t-1} = i\right)$ are given by

$$p_{j|i} = \binom{n}{j}\left(a_0 + \frac{1}{n}a_1 \cdot i\right)^j \left(1 - a_0 - \frac{1}{n}a_1 \cdot i\right)^{n-j}, \tag{1}$$

for $i, j \in \{0,1,...,n\}$, while the conditional expected value $E(X_t \mid X_{t-1})$ and the conditional variance $V(X_t \mid X_{t-1})$ are, respectively, equal to

$$E(X_t \mid X_{t-1}) = n\alpha_t,\ V(X_t \mid X_{t-1}) = n\alpha_t(1-\alpha_t). \tag{2}$$

Parameters α_0 and α_1 of the BINARCH(1) model can be estimated via the method of Conditional Maximum Likelihood (CML). Let us assume that $X_1,...,X_T$ is a segment

from a stationary BINARCH(1) process. Then, by conditioning on x_1, the conditional log-likelihood function equals [see Ristic et al. (2016)]

$$l(a_0, a_1) = \sum_{t=2}^{T} \left\{ \log \binom{n}{x_t} + x_t \log \alpha_t + (n - x_t) \log(1 - \alpha_t) \right\}. \quad (3)$$

The CML estimators $\hat{a}_{0ML}$, $\hat{a}_{1ML}$ of a_0, a_1 are obtained by maximizing numerically the function $l(a_0, a_1)$ while the corresponding standard errors of the estimates can be obtained via Fisher's Information matrix. Other estimation methods of the parameters of a BINARCH(1) model can be found in Ristic et al. (2016), where the interested reader is referred to.

3. CONTROL CHARTS

3.1 Method

It is well known that Shewhart charts are control charts without memory since they make use of only the value of the most recent observation. Consequently, they are not very sensitive in small and moderate changes in the values of process parameters. On the other hand, the cumulative sum (CUSUM, [Page (1954)]) and exponentially weighted moving average (EWMA, [Roberts (1959)]) control charts, as control charts with memory, detect these types of changes more quickly than the Shewhart charts. See, for example, the works of [Gan (1990)], [Wu et al. (2008)], [Bourke (2001)], [Morais and Pacheco (2006)], [Haridy et al. (2020)] and references therein.

In this section, we focus on the development of upper and lower one-sided CUSUM control charts for monitoring a BINARCH(1) process. The aim is to detect quickly and accurately a change in the process mean level. When the process is in-control (IC), we will denote its IC process mean level as $\mu_{0,X}$ while in the out of control state (OOC), it is denoted as $\mu_{1,X}$. In a similar manner, the IC (OOC) parameter values of the BINARCH(1) model are denoted as a_{00} and a_{01} (a_{10} and a_{11}).

Usually, practitioners focus on changes in the mean level $\mu \equiv E(X_t) = na_0/(1 - a_1)$ of the process, mainly in detecting increases in the process mean level, from an IC value $\mu_{0,X}$ to an OOC value $\mu_{1,X} > \mu_{0,X}$. In practice, it is of great importance to detect an increase in the mean of the process, because it is related to a process deterioration. However, when a decrease in the mean of the process has occurred, the process has been improved. In this work we consider both cases.

3.2 CUSUM Control Charts

In the sequel, we propose one-sided CUSUM control charts for monitoring BINARCH(1) processes, by using the likelihood ratio (LR) statistic (see also [Weiß and Testik (2012)]). Specifically, using the transition probabilities given by (1), we form the following LR statistic for a BINARCH(1) process:

$$LR(a_{00},a_{01},a_{10},a_{11}) = \frac{L(a_{10},a_{11})}{L(a_{00},a_{01})} = \frac{\binom{n}{j}\left(a_{10}+\frac{1}{n}a_{11}\cdot i\right)^{j}\left(1-a_{10}-\frac{1}{n}a_{11}\cdot i\right)^{n-j}}{\binom{n}{j}\left(a_{00}+\frac{1}{n}a_{01}\cdot i\right)^{j}\left(1-a_{00}-\frac{1}{n}a_{01}\cdot i\right)^{n-j}}$$

$$= \left(\frac{a_{10}+\frac{1}{n}a_{11}\cdot i}{a_{00}+\frac{1}{n}a_{01}\cdot i}\right)^{j}\left(\frac{1-a_{10}-\frac{1}{n}a_{11}\cdot i}{1-a_{00}-\frac{1}{n}a_{01}\cdot i}\right)^{n-j} . \quad (4)$$

Then the one-sided CUSUM chart is defined as $C_t = \max(0, C_{t-1} + lR_t)$, $t \geq 2$ and give an OOC sign at sample *t* if $C_t \geq h$, where *h* is the decision interval (control limit) of the CUSUM chart and lR_t is the log-likelihood ratio which equals:

$$lR_t = j\cdot\log\left(\frac{a_{10}+\frac{1}{n}a_{11}\cdot i}{a_{00}+\frac{1}{n}a_{01}\cdot i}\right)+(n-j)\cdot\log\left(\frac{1-a_{10}-\frac{1}{n}a_{11}\cdot i}{1-a_{00}-\frac{1}{n}a_{01}\cdot i}\right),\ t\geq 2 \quad (5)$$

Assume now that a shift to a certain OOC parametrization (a_{10}, a_{11}) is of particular interest. Then, by defining (a_{10}, a_{11}) in terms of (a_{00}, a_{01}), we can derive the corresponding LR test statistic and formulate the CUSUM charts for this OOC situation. This is exemplified in the sequel by considering the following three cases:, (i) $a_{10} = \delta^* \cdot a_{00}$, ($\delta^* > 1$, for increases or $0 < \delta^* < 1$ for decreases) and $a_{11} = a_{01}$, i.e., a change only in a_{00}, (ii) $a_{11} = a_{01} + \tau^*$ (for increasing shifts) or $a_{11} = a_{01} - \tau^*$ (for decreasing shifts), where $\tau^* > 0$, i.e., a change only in a_{01} and (iii) a simultaneous change in both parameters.

Except for the suggested CUSUM chart defined previously, we also propose the use of a combined scheme, which consists of two CUSUM charts running simultaneously. Specifically, we have one CUSUM chart suitable to detect changes only in a_{00} and another CUSUM chart suitable to detect changes only in a_{01}. The combined scheme gives an OOC sign at sample *t* if $C_t \geq h_A$ or $C_t \geq h_B$ where h_A and h_B are the decision intervals (control limits) for each chart of the combined CUSUM.

3.3 Performance Measures

In order to evaluate the performance of the proposed charts, it is necessary to determine their run length (*RL*) distribution. The *RL* distribution is the distribution of the number of points plotted on the chart until it gives for the first time an OOC signal. Its expected value *E*(*RL*), also known as average run length (*ARL*), is the most common performance measure of a control chart. The *ARL* expresses the average number of points to be plotted on the chart until it gives for the first time an OOC

signal. Due to the autocorrelation among the successive counts, different performance measures must be used regarding the IC and the OOC performance of the charts.

In this work, the IC performance of the proposed schemes is evaluated in terms of the zero-state *ARL* (zs*ARL*) which is the expected number of points plotted on the chart until the first (false) alarm is given.

For an OOC process, the performance of the proposed schemes is evaluated in terms of the steady-state *ARL* (ss*ARL*) which gives an approximation of the true mean delay for detection after a change in the process, from the IC state to the OOC state. Specifically, we assume that a change in process happens at an (unknown) changepoint $\xi = 1, 2, \ldots$. That is, for $\xi < t$, the process is in the IC state while for $\xi \geq t$, the process has shifted to the OOC state. Therefore, the ss*ARL* expresses the expected number of points to be plotted on the chart until it gives for the first time an indication of an OOC process, given that the process has been operated for "sufficient time" in control. According to Weiß and Testik (2011), the zs*ARL* and the ss*ARL* are substantially different in the case of monitoring processes with correlated counts.

3.4 Statistical Design

The statistical design of the proposed CUSUM chart requires the determination of the value of h (or h_A and h_B for the combined CUSUM), such that its IC performance is the desired one. Next, we provide the steps of the algorithmic procedure that is used for the determination of the h value for a CUSUM chart with the desired IC *ARL* performance.

Step 1. Choose the values for n, a_{00}, a_{01} and the desired IC *ARL* value, say ARL_0.
Step 2. Choose the shift of interest (e.g. $a_{10} = \delta^* \cdot a_{00}$, $\delta^* > 1$, $a_{11} = a_{01} + \tau^*$, $\tau^* > 0$).
Step 3. Determine h such that the IC *zsARL* is close enough to the desired ARL_0 value.

The above steps apply for any shift (either increasing or decreasing) in process parameters.

3.5 Numerical Study

In this section, we present the results of a numerical study on the performance of one-sided Shewhart, *s*-EWMA and CUSUM charts in the monitoring of BINARCH(1) processes. The one-sided Shewhart and *s*-EWMA for BINARCH(1) processes have been studied by Anastasopoulou and Rakitzis (2020). The *s*-EWMA control chart has been proposed and studied by Weiss (2011) and it is a modification of the usual EWMA statistic. Specifically, the values plotted on a *s*-EWMA chart are given by

$$Q_t^{(s)} = s - \text{round}\left(\lambda X_t + (1-\lambda) Q_{t-1}^{(s)}\right), Q_0^{(s)} = q_0^{(s)},$$

where the initial value $q_0^{(s)} = \lfloor \mu_{0,X} \rfloor$. The *s*-round(…) function is defined as *s*-round(x) = z if-f $z - 0.5 \cdot s \leq x \leq z + 0.5 \cdot s$ and it is a generalization of the usual rounding function since it rounds a number x to the nearest fraction with denominator

s. Using Monte Carlo simulation, we compare the performance of the proposed CUSUM charts with the corresponding one-sided Shewhart and *s*-EWMA charts for BINARCH(1) processes. The aim is to detect quickly and accurately changes in process parameters.

For the IC design parameters, we consider various combinations for the values (a_{00}, a_{01}) such that $a_{00} + a_{01} < 1$. When the process is OOC, both parameters can change, either simultaneously or not. Therefore, we assume that the parameter a_{00} changes from a_{00} to $a_{10} = \delta \cdot a_{00}$, $0 < \delta < 1$, (downward shifts) or $\delta > 1$ (upward shifts), such that $a_{10} + a_{11} < 1$. In a similar manner, the changes in a_{01} are given as follows: In the case of upward shifts, we assume that $a_{11} = a_{01} + \tau$, while in the case of downward shifts, we assume that $a_{11} = a_{01} - \tau$, where $\tau > 0$ and $a_{10} + a_{11} < 1$.

Table 1 gives the results of a comparative study between the upper one-sided Shewhart, *s*-EWMA and CUSUM charts for BINARCH(1) processes. The IC parameter values of the process are given in the columns "a_{00}", "a_{01}", "$\mu_{0,X}$"and "n", while the rows "UCL", "λ", h_A, h_B consist of the values of the design parameters of each chart. The h_A line consists of the value of h for the proposed CUSUM charts whereas, h_A and h_B are the decision intervals that formulated the combined CUSUM schemes. Also, the rows d_A^*, d_B^*, τ_A^*, τ_{B}^* consist of the values of the shifts of interest (or fixed shifts), that is the shifts that we want to detect. It should be also mentioned that the design parameters for the *s*-EWMA and CUSUM charts have been determined so as their IC performance is the closest possible to the IC performance of the Shewhart chart, in order to have a fair comparison between the different charts.

Table 1: *ARL comparison, one-sided Shewhart, s-EWMA and CUSUM charts for upward shifts*

μ_0	n	a_{00}	a_{01}	δ	τ	Shewhart	*s*-EWMA			CUSUM			Combined CUSUM
							s=1	*s*=2	*s*=4				
3	30	0.05	0.5	1	0	398.78	365.27	403.34	403.14	398.63	397.67	397.08	385.23
				1.1	0	250.53	206.28	231.35	209.26	**138.12**	171.85	178	142.31
				1.3	0	112.49	81.28	91.87	74.81	**46.57**	56.35	57.02	47.47
				2	0	17.67	12.83	13.61	12.02	13.96	12.11	**11.66**	12.49
				1	0.05	191.67	148.22	164.29	149.49	127.68	126.29	131.73	**119.79**
				1	0.07	145.19	108.28	119.63	106.98	93.51	89.23	93.8	**85.33**
				1	0.1	97.56	70.67	77.47	68.55	63.4	57.74	60.54	**57.57**
				1.1	0.05	126.66	92.48	101.97	88.55	67.71	70.96	74.93	**65.69**
				1.3	0.05	62.01	43.41	47.94	39.94	32.34	33.05	33.03	**31.3**
				1.5	0.05	35.13	24.42	26.6	22.18	21.23	20.2	**19.73**	20.11
				1.1	0.07	97.73	69.68	77.32	66.75	54.41	54.88	57.29	**52.13**
				1.3	0.07	49.8	35.25	38.19	32.04	28.26	27.65	27.98	**27.24**
				1.2	0.1	48.84	34.81	37.51	32.07	30.12	28.56	28.64	**28.23**
				1.3	0.1	36.43	26.17	27.96	24.12	24.13	22.24	**21.98**	22.32
UCL						10	7	7	5.75				
λ						1	0.27	0.35	0.18				
h_A										2.11	2.05	2.9	2.33
h_B													2.26
d_A^*										1.2	1	1.2	1.2
d_B^*													1
τ_A^*										0	0.1	0.1	0
τ_B^*													0.1

Table 2: *ARL comparison, one-sided Shewhart, s-EWMA and CUSUM chart for downward shifts*

μ_0	n	a_{00}	a_{01}	δ	τ	Shewhart	s-EWMA			CUSUM			Combined CUSUM
							s=1	s=2	s=4				
18.75	30	0.05	0.92	1	0	408.78	372.98	409.45	405.09	407.41	410.33	408.17	410.53
				0.9	0	227.67	202.88	216.65	207.54	179.77	234.32	242.27	197.94
				0.8	0	137.06	122.10	128.19	123.30	105.08	144.66	149.92	118.77
				0.7	0	89.06	79.97	83.05	79.18	71.79	95.35	98.02	79.06
				1	0.05	72.48	62.90	65.33	59.37	49.19	69.19	82.44	55.00
				1	0.1	31.73	27.85	28.73	27.38	26.64	27.47	30.26	26.52
				1	0.15	18.78	17.21	17.99	18.45	18.61	17.05	16.42	16.97
				0.9	0.05	52.68	46.47	48.21	44.83	39.85	51.29	57.63	43.74
				0.8	0.05	39.99	36.10	36.79	35.36	33.55	39.82	42.88	35.85
				0.7	0.05	31.84	29.23	29.63	29.30	28.34	32.46	33.22	29.90
LCL						3	6	7	9				
λ						1	0.26	0.2	0.11				
h_A										1.69	4.22	4.34	2.01
h_B													4.47
d_A^*										0.8	1	0.8	0.8
d_B^*													1
τ_A^*										0	0.1	0.1	0
τ_B^*													0.1

The comparison of the *ARL* profiles between the Shewhart, *s*-EWMA and the CUSUM charts for BINARCH(1) processes reveals that the proposed CUSUM charts are very powerful charts in the detection of small and moderate shifts in the mean level of the process. This conclusion stems from the fact that the CUSUM control charts have the lowest value ss*ARL* for each shift.

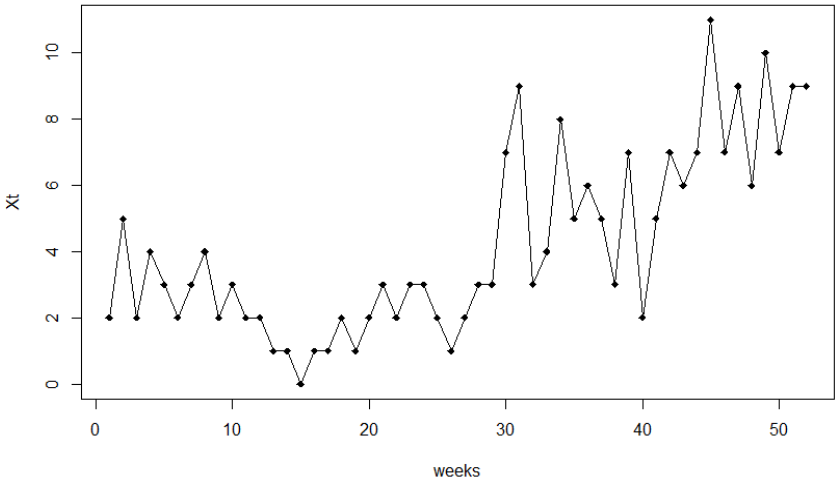
Similar conclusions are drawn for the one-sided control charts for decreasing shifts in the process mean level (see Table 2). Again, the CUSUM charts have the minimum ss*ARL* value for each shift and thus, are the charts with the best performance.

4. A REAL DATA EXAMPLE

In this section we present an application of the proposed CUSUM charts to real data for monitoring BINARCH(1) processes. These data refer to the regional spread of an infection in Germany within a year. Specifically, we have the weekly number X_t of regions in Germany, with a new case of hantavirus infection in 2011, for $T = 52$ weeks. The number of regions is $n = 38$. More details on this dataset can be found in

[Weiß and Pollett (2014)] and [Ristic et al. (2016)]. The time series plot in Figure 1 shows that the values in the sample are between 0 and 11. The sample mean is 4.173 and the sample variance is 7.793.

Figure 1. *Weekly number of regions with new cases of hantavirus*



Next, we fit the BINARCH(1) model in the data by using the method of Conditional Maximum Likelihood (CML) where the logarithm of the likelihood function (conditioned on x_1) is given in (3). Thus, the maximum likelihood (ML) estimates are obtained by maximizing it. Using `R` [R Core Team (2002)] and the function `optim()`, we obtain the ML estimates for a_{00}, a_{01} (in the parentheses we provide the respective standard errors), that is $\hat{a}_{00ML} = 0.03$ (0.011) and $\hat{a}_{01ML} = 0.748$ (0.108). In the sequel we assume these are the actual values of the IC process parameters. Moreover, for the development of all the considered control charts, we choose (for illustrative purposes) an ARL_0=100 as the desired IC *ARL* value.

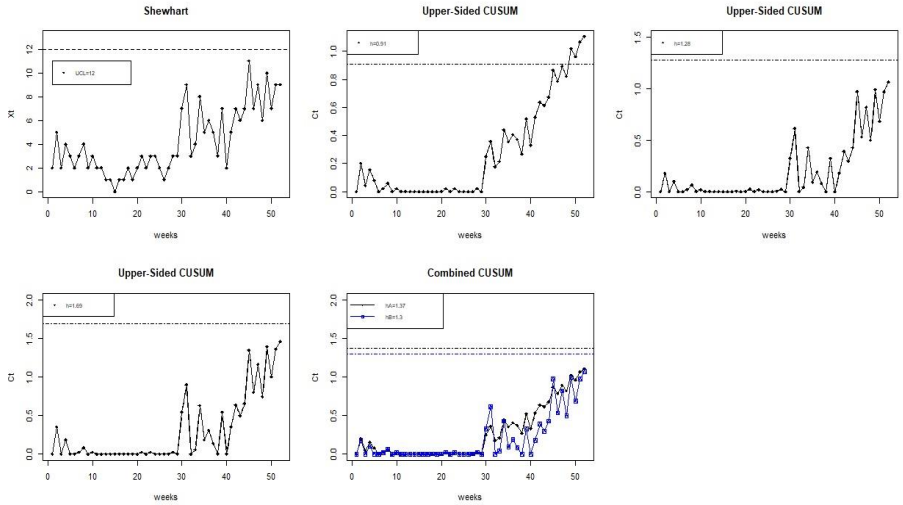
Following the steps for determining the chart's parameters (h_A for CUSUM, h_A and h_B for Combined CUSUM), we present the following one-sided control charts for upward shifts in the mean level of the process, with an IC zs*ARL* value as close as possible to 100 and with the following fixed shifts (lines d_A^*, d_B^*, t_A^*, t_B^*). Figure 2 consists of the CUSUM charts for the hantavirus data, in the case of upward shifts in the process mean level.

We observe that in the last weeks of the year, there is a strong upward movement which begins shortly after the 30th week, which is followed by a clear indication of an upward shift from the 41st week onwards. This shift is perceived by all schemes; the CUSUM chart with change only in a_{00} gives an OOC signal at the 49th week.

Table 3. *Design Table for the CUSUM charts for upward shifts in the mean*

Charts	*Shewart*	*CUSUM*	*CUSUM*	*CUSUM*	*Combined CUSUM*
zsARL	83.30	99.09	100.78	100.22	99.19
h_A		0.91	1.28	1.69	1.37
h_B					1.3
d_A^*		1.2	1	1.2	1.2
d_B^*					1
τ_A^*		0	0.1	0.1	0
τ_B^*					0.1
UCL	12				

Figure 2. *Shewhart and CUSUM charts of Table 3 for the Hantavirus data*

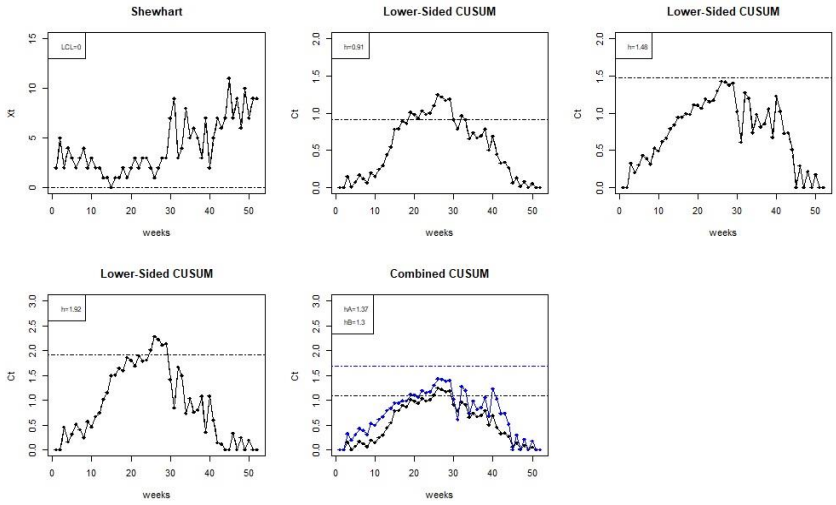


Next, we proceed with the CUSUM charts for detecting downwards shifts in the process mean level. Specifically, we consider the following schemes while the respective CUSUM charts for the hantavirus data are given in Figure 3. We notice that the lower one-sided Shewhart chart gives an OOC signal (X_{15} = 0), but it should be interpreted with caution because of its low IC zs*ARL* value. On the contrary, the proposed CUSUM charts give IC ARL_0 close to the desired value. The CUSUM chart optimized for a change only in a_1 gives an OOC signal for the first time at point 19, while the combined CUSUM chart signals at the point 25.

Table 4. *Design Table for the CUSUM charts for downward shifts in the mean*

Charts	*Shewhart*	*CUSUM*	*CUSUM*	*CUSUM*	*Combined CUSUM*
zsARL	36.91	100.1563	100.2056	100.8044	99.09152
h_A		0.91	1.48	1.92	1.09
h_B					1.69
d_A^*		0.8	1	0.8	0.8
d_B^*					1
τ_A^*		0	0.1	0.1	0
τ_B^*					0.1
UCL	0				

Figure 3. *Shewhart and CUSUM charts of Table 4 for the Hantavirus data*



6. CONCLUSIONS

In this work, we developed and studied one-sided CUSUM control charts for monitoring a BINARCH(1) process. The proposed charts are based on a likelihood ratio statistic and can be used for detecting upward or downward shifts in process mean level of the process. Our numerical analysis (based on Monte Carlo simulation) revealed that the proposed charts outperform the respective Shewhart and *s*-EWMA charts for small to moderate increasing and decreasing shifts in the parameters of the process. Finally, they can be effectively used in the monitoring of the epidemiological data.

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία, αναπτύσσονται μονόπλευρα διαγράμματα ελέγχου τύπου CUSUM για την παρακολούθηση μιας διεργασίας η οποία περιγράφεται από ένα μοντέλο χρονολογικών σειρών με ακέραιες τιμές, το BINARCH(1). Τα προτεινόμενα διαγράμματα βασίζονται στον λόγο πιθανοφάνειας και μπορούν να χρησιμοποιηθούν για την έγκυρη ανίχνευση αυξήσεων ή μειώσεων στο μέσο επίπεδο της διεργασίας. Παρουσιάζονται αριθμητικά αποτελέσματα μιας συγκριτικής μελέτης μέσω προσομοίωσης. Από τις συγκρίσεις που γίνονται διαπιστώνεται ότι τα προτεινόμενα διαγράμματα υπερέχουν των αντίστοιχων διαγραμμάτων Shewhart και *s*-EWMA για μικρές ή/και μεσαίες μετατοπίσεις στις παραμέτρους της διαδικασίας. Τέλος, δίνεται μια πρακτική εφαρμογή των προτεινόμενων διαγραμμάτων στην παρακολούθηση επιδημιολογικών δεδομένων.

REFERENCES

Anastasopoulou M. and Rakitzis A. C. (2020). EWMA control charts for monitoring correlated counts with finite range. *Journal of Applied Statistics* (*to appear*). DOI: 10.1080/02664763.2020.1820959.

Anastasopoulou M. and Rakitzis A. C. (2022). Monitoring a BAR(1) Process with EWMA and DEWMA Control Charts. In Tran K.P. (ed.) *Control Charts and Machine Learning for Anomaly Detection in Manufacturing* (pp. 77-103). Springer, Cham.

Bersimis S., Sgora A., and Psarakis S. (2018). The application of multivariate statistical process monitoring in non-industrial processes. *Quality Technology & Quantitative Management*, **15**(4), 526–549.

Bourke P. D. (2001). Sample size and the binomial CUSUM control chart: the case of 100% inspection. *Metrika*, **53**(1), 51–70.

Gan F. F. (1990). Monitoring observations generated from a binomial distribution using modified exponential weighted moving average control chart. *Journal of Statistical Computation and Simulation*, **37**, 45–60.

Haridy S., Shamsuzzaman M., Alsyouf I., and Mukherjee A. (2020). An improved design of exponentially weighted moving average scheme for monitoring attributes. *International Journal of Production Research*, **58**(3):931–946.

Kim H. and Lee S. (2009). Improved CUSUM monitoring of markov counting process with frequent zeros. *Quality and Reliability Engineering International*, **35**(7), 2371–2394.

McKenzie (1985). E. Some simple models for discrete variate time series. *Water Resources Bulletin*, **21**, 645–650.

Montgomery D. C. (2009). *Introduction to Statistical Quality Control*. John Wiley & Sons, Inc., New York, USA, 6th edition.

Morais M. C. and Pacheco A. (2006). Combined CUSUM–Shewhart schemes for binomial data. *Stochastics and Quality Control*, **21**(1), 43–57.

Moustakides G.V. (1986). Optimal stopping times for detecting changes in distributions. *Annals of Statistics*, **14**, 1379–1387.

Page E. S. (1954). Continuous Inspection Schemes. *Biometrika*, **41**, 1100–115.
Psarakis S. and Papaleonida G. E. A. (2007). SPC procedures for monitoring autocorrelated processes. *Quality Technology & Quantitative Management*, **4**(4), 501–540.
R Core Team (2022). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing. http://www.r-project.org
Rakitzis A. C., Weiß C.H., and Castagliola P. (2017). Control charts for monitoring correlated counts with a finite range. *Applied Stochastic Models in Business and Industry*, **33**(6), 733–749.
Ristic M. M., Weiß C. H., and Janjic A. D. (2016). A binomial integer-valued ARCH model. *The International Journal of Biostatistics*, **12**(2). DOI: 10.1515/ijb-2015-0051.
Roberts S. W. (1959). Control chart tests based on geometric moving averages. *Technometrics*, **1**, 239–250.
Weiß C. H. (2009). Monitoring correlated processes with binomial marginals. *Journal of Applied Statistics*, **36** (4), 399–414.
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**(5), 588–608.
Weiß C. H. and Pollett P. K. (2014). Binomial autoregressive processes with density dependent thinning. *Journal of Time Series Analysis*, **35**, 115–132.
Weiß C. H. and Testik M. C. (2011). The Poisson INAR(1) CUSUM chart under overdispersion and estimation error. *IIE Transactions*, **43**(11), 805–818.
Weiß C.H., Testik M.C (2012). Detection of Abrupt Changes in Count Data Time Series: Cumulative Sum Derivations for INARCH(1) Models, *Journal of Quality Technology* , 44:3, 249 264, DOI:10.1080/00224065.2012.11917898.
Woodall W. H., Zhao M. J., Paynabar K., Sparks R., and Wilson J. D. (2017). An overview and perspective on social network monitoring. *IISE Transactions*, **49**(3), 354–365.
Wu Z., Jiao J. and Liu Y. (2008). A binomial CUSUM chart for detecting large shifts in fraction nonconforming. *Journal of Applied Statistics*, **35**(11), 1267–1276.

Greek Statistical Institute
Proceedings of the 33rd Panhellenic Statistics Conference (2021), pp.233-247

STRATIFICATION OF FOREST STANDS AS A BASIS FOR SMALL AREA ESTIMATIONS

Georgakis Aristeidis
Aristotle University of Thessaloniki, School of Forestry and Natural Environment, Thessaloniki, Lab of Forest Biometry, PC-54124, Thessaloniki
e-mail: arisgeorg@for.auth.gr

ABSTRACT

Forest inventories provide all the necessary information for the sustainable management of forest ecosystems. This information includes the estimation of forest biometric variables, such as the growing stock volume (GSV) which is of the most interest. In the present work, an attempt was made to define post-strata after grouping homogeneous forest stands/compartments with cluster analysis. For this purpose, the combination of different parameters from past forest inventories, such as the total GSV and the tree density per compartment and hectare, were used as grouping variables. Twenty-eight different clusters were defined based on different auxiliary variables (observations) and using different clustering methods, including the proper selection of the number of clusters *k* and the aggregation algorithms that grouped the observations. The evaluation of the number of clusters is based mainly on the silhouette width. Clustering is used as a support tool for the estimation phase, for enlarging the small areas and correspondingly the sample size. Thus, the final evaluation of the clusters is based on the direct estimates, aiming at the minimum relative standard error of the mean (rRMSE%) of GSV.

Keywords: Clustering, past census data, subpopulation estimates, Silhouette, PAM

1. INTRODUCTION

The sustainability of the forests and forest products depends on the proper forest management plans. Forest inventories (FIs) provide all the basic information necessary for the sustainable management of forest ecosystems. Forest inventories can be distinguished in national and management forest inventories (MFIs). Forest growing stock volume (GSV), one of the most important forest biometric variables, includes the stem volume (m^3/ha) of all living trees from ground level or stump height up to a minimum diameter of the treetop or branches (FAO, 2004). MFI faces a big challenge because it aims to provide estimates not only in the whole forest population but rather for each management unit (forest stand or compartment) that corresponds to a geographic subpopulation (small area or domain). Recently, there

has been a growing interest of small area estimations (SAE) both in research; (Rao & Molina, 2015) and practical applications such as MFIs (Breidenbach, Magnussen, Rahlf, & Astrup, 2018; Goerndt, 2010; Magnussen, Mauro, Breidenbach, Lanz, & Kändler, 2017). SAE can be considered a technique that aims to provide estimates in small areas (geographical areas in forestry) when direct estimates are not feasible due to the small sample size. The plethora of SAE literature includes model-based and model-assisted estimations (Hill, Mandallaz, & Langshausen, 2018), using correspondingly good auxiliary covariates.

An alternative approach in case of unreliable design-based direct estimations and when auxiliary covariates are not available for model-based or model-assisted SAE is the post-stratification. With post-stratification or post-stratified sampling, the forest population is divided into larger homogenous groups, the so-called post-strata or just strata, under the existing sampling design. We can consider a pre-stratification of the existing sampling survey that was applied only to the forest area, excluding non-forest areas.

The problem of SAE remains also in strata estimations. Westfall et al. (2011) detected a disagreement between researchers relative to the minimum number of sample units per stratum, which ranges from 10-20, and also some form of ambiguity in the justification of the findings. Their research (Westfall et al., 2011) recommends *at least 10 sample* units as the minimum total sample size within-stratum, for the stability of the mean and the estimated standard error of the mean. An additional problem to the small sample size per stratum is the unequal sampling fraction that can sometimes be extremely small. Generally, stratification yields more efficient estimators under the consideration that within, the strata are homogeneous and the overall population heterogeneous (Golder & Yeomans, 1973). Stratification is used as a sampling variance reduction technique, compared to the simple random sampling (SRS), when there is smaller within-strata variance (homogeneity) and larger between-strata variance (heterogeneity) (Golder & Yeomans, 1973).

Cluster analysis was used for post-stratification in this research. With clustering, an effort has been done for the enlargement of the forest stands (sub-populations of interest) into larger groups, with the expectation to provide more accurate post-strata estimates. A prerequisite for conducting post-stratification with cluster analysis is the availability of auxiliary variables that can describe the heterogeneity across the forest stands. Such auxiliary variables can be from past census data, remote sensing data (Potapov et al., 2021), abiotic characteristics such as aspect or/and slope of forest stands, or data that describe the spatial contiguity between the stands such as centroids (X, Y coordinates). Generally, we cannot rely on the sample units (plots) for clustering, assuming that the extremely small sample size cannot describe the status of the stands. Stratification and clustering can be used interchangeably.

The delineation of the forest ecosystem into sub-populations of forest stands or compartments has been done based on the natural lines, such as streams, ridges, and roads. We assume that forest stands are not always necessarily different or there is

some degree of similarity between them, regarding the variable of interest or another auxiliary variable that can be used in cluster analysis. This proposed methodology can be used when there are no available auxiliary covariates at the unit-level (sample plots) or area-level (aggregated information in the forest stands) that can be linearly related to the target variable of GSV and for this reason, no other SAE technique can be used. Keeping in mind that the minimum within-strata sample size should be at least 10 (Westfall et al., 2011), an expected solution is to cluster around 7 stands averagely with correspondingly 10 plots per group. The direct estimates of the post-strata made use of the sample plots from the systematic sampling but were treated as they came from a simple random sampling from an infinite population.

Cluster analysis or clustering is an unsupervised machine learning technique that helps to explore the data by partitioning the data observations into meaningful groups that share similar characteristics amongst each other. Clustering methods can be distinguished into two broad categories, hierarchical and non-hierarchical methods (Giordani, Ferraro, & Martella, 2020). Standard clustering methods include the agglomerative hierarchical clustering of the well-known single linkage method, average linkage method, complete linkage method, and Ward's method. From the non-hierarchical clustering methods, the most famous are the *k*-Means algorithm and the *k*-Medoids (Giordani et al., 2020). Cluster information can be incorporated into the SAE models (Fay & Herriot, 1979) in the form of dummy/factor variable (Torkashvand, Jozani, & Torabi, 2017; You & Chapman, 2006; Zulkarnain, Jayanti, & Listianingrum, 2020) or can improve the linear relation in statum-level with the variable of interest.

The primary goal of this study was to achieve an acceptable accurate direct estimate per post-statum with as much as a possible greater number of clusters. Thus, finding the minimum sample size with acceptable accuracy for forest stands (sub-populations) in the University Forest of Pertouli, took great effort. Similar to relative research (Torkashvand et al., 2017), the optimal number of clusters is not the smallest number of clusters because the aim is to have, as much as possible, similar small areas inside a cluster. The "best" number of clusters can be considered subjective to a large extent, but with silhouette analysis (Rousseeuw, 1987) this subjectivity was minimized. The evaluation of clusters was initially based on the average silhouette criterium (Rousseeuw, 1987) on the working auxiliary data. The basic components of the clustering methods are the distance metric (ex. Euclidian), the number of clusters and the clustering algorithm**.** The statistical analysis was done with the open-source statistical software R (R Core Team, 2021) and the cluster analysis based mainly on the package 'cluster' (Maechler, Rousseeuw, Struyf, Hubert, & Hornik, 2014).

Twenty-eight different clusters were defined based on different auxiliary variables (observations) and using different clustering methods, including the proper selection of the number of clusters *k* and the aggregation algorithms that grouped the observations. Clusters were created with partitioning around medoids (PAM) (Kaufman & Rousseeuw, 1990), *k*-means (Giordani et al., 2020) and hierarchical

algorithms (Ward, 1963), while the number of clusters derived with silhouette analysis (Rousseeuw, 1987). Clustering is used as a support tool for the estimation phase, for enlarging the small areas and correspondingly the sample size. Thus, the final evaluation of the clusters is based on the direct estimates in terms of the minimum small relative standard error of the mean (rRMSE%) of GSV.

2. MATERIALS AND METHODS

2.1 Case study

This study area of interest is the University forest of Pertouli in Greece with the dominant species the *Abies borisii-regis* (Mattf.) (hybrid fir). The population consists of the total forest with adequate 174 small areas to be the forest stands/compartments. The variable of interest (target variable) is the GSV m^3/ha. The clustering will be conducted in 168 stands removing the unmanaged with no sample plots. From the 168 stands, the 160 stands are considered as planned because they have at least one plot and 8 non-sampled or unplanned because there are no sample plots. Half of the stands have only one sample plot and the other half just two, with a few stands to have non and three plots. Due to the extremely small sample size per stand (1-3 sample plots), we cannot rely on the survey sample data for clustering, instead, we use auxiliary data that cover all the small areas.

Auxiliary data that is used in clustering include area-level data from past census measurements including a complete enumeration of trees per domain, where and ForestDensity/FirTreeDensity adequate to trees/ha, diameter at breast height (DBH) and GSV m^3/ha measured per compartment. Here is the description of the auxiliary variables used in clustering:

a) Census 1988, GSV m^3/ha (GSV88, FirGSV88), trees/ha (ForestDensity, FirTreeDensity88) and distribution of DBH/GSV for 174 compartments
b) Census 1997, GSV m^3/ha (GSV97, FirGSV97), trees/ha (ForestDensity97, FirTreeDensity97) and distribution (fir_dbh_distribution97) of DBH/GSV for 173 compartments
c) Diameter at breast height was categorized in classes of 4 & 5cm DBH
d) Systematic sampling 2008 (sampling intensity 1%, sample size ≈ 250, sample unit =0,1 ha) and their estimations per stand
e) Systematic sampling 2018 (sampling intensity 1%, sample size ≈ 250, sample unit =0,1 ha) and their estimations per stand
f) Abiotic factor include
    i) slope and aspect for each stand.
    ii) Site Index (SI) also characterizes the productivity of the stands (average value)
    iii) An estimated optimum density number of trees per stand
g) Forest canopy height data (meanHeight) from Global Ecosystem Dynamics Investigation (GEDI) lidar instrument integrated with Landsat data (30m pixel size) (Potapov et al., 2021).

h) The spatial relation of the stands from the corresponding centroids-coordinates (X, Y).

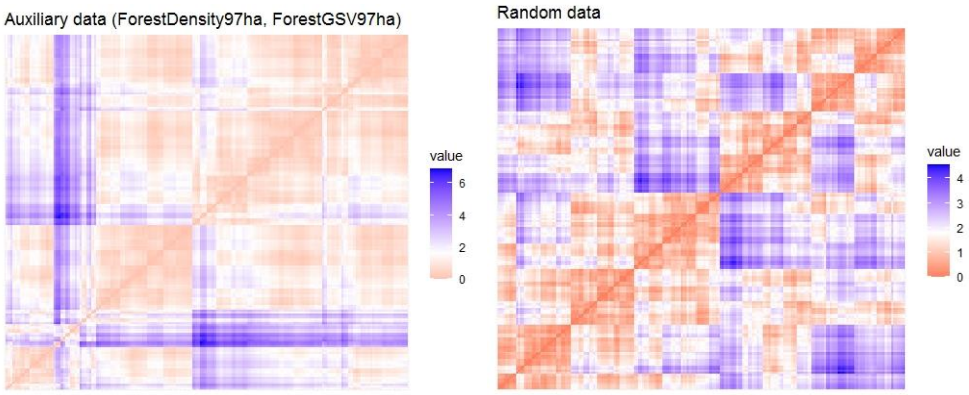
Sampling design usually is not considered in SAE in advance (Georgakis & Stamatellos, 2020). The direct estimates derived from the systematic sampling survey of 2018, having sampling fraction 1%, sample size (plots) 252 and sample unit (plot)=0,1 ha. Samples are located only in a forested area and every sample plot includes measures of the variable of interest (GSV).

2.2 Clustering methodology

First, we examine clusterability with the following two tests and optical illustration. "Multiple modes in the distance distribution suggest the presence of multiple clusters" (Adolfsson, Ackerman, & Brownstein, 2019). Using the Hartigans' Dip Test for Unimodality we proved that we have non-unimodal and have at least bimodal distribution for example for the auxiliary candidates (ForestDensity97ha, ForestGSV97ha). In the second test, we use the Hopkins' (H) statistic (Lawson & Jurs, 1990; Wright, YiLan, & RuTong, 2021) to test clusterability via spatial randomness. We test the auxiliary candidates (ForestDensity97ha, ForestGSV97ha) selecting an integer (n=27) that corresponds to the number of points selected from sample space which is also the number of points selected from the given sample(data). Random numbers from two random variables/columns give us H = 0,49 (close to 0,5 which is the uniform distribution – Null Hypothesis), while the H = 0,18 for the real data belong to the Alternative Hypothesis and means that auxiliary data are not uniformly distributed and therefore there is cluster tendency. If we want to illustrate the clusterability visually first of many variables first we scale the data and after we computed the dissimilarity matrix between observations. After we apply the algorithm of the visual assessment of cluster tendency (VAT) approach (Bezdek and Hathaway, 2002) (Figure 1) as referred by Giordani et al. Giordani et al. (2020). With red color, there is high similarity or low dissimilarity and with blue low similarity.

Regarding variable selection for clustering, there is not a clear way to find the variables that will be used in clustering. The variable selection is based on the relation to the variable of interest GSV with the auxiliary data. They used both unique and also two or more variables for clustering the small areas. In the case that we do not know which auxiliary data is valuable for clustering homogenous small areas, we can start from the combination of all the variables and can investigate which ones can partition better the population.

Figure 1. *Visual illustration of clusterability on the left figure compared to random numbers on the right (Red: high similarity and Blue: low similarity)*

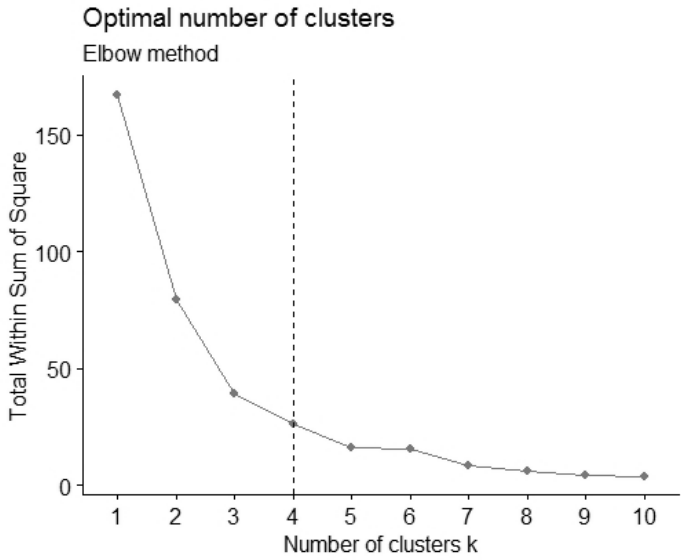


Before clustering, we have to prepare the data by doing some pre-processing. Pre-processing includes the scaling of data that are normalized/standardized for avoiding the influence of some auxiliary data on the distance measure. After we compute the dissimilarity matrix with euclidean distances (Distance = 1- Similarity), knowing that two observations are similar if they have a small distance from one to another and less similar if they have a large distance.

After pre-processing the starts the crucial task is to define the number of clusters *k*. The number of clusters is partially subjective and depends also on the research question. Therefore, in this study, the effort has been placed to find a large number of clusters using different techniques.

Initially, the elbow method was tested for providing the expected number of clusters. The elbow method relies on calculating the total within-cluster sum of squares (Giordani et al., 2020) across every cluster (usually with a *k*-means algorithm), that is the sum of Euclidean distances between each observation and the centroid (or medoid for usually) corresponding to the cluster to which the observation is assigned. *k*-means and *k*-medoids with the same number of clusters (from silhouette) do not provide substantial differences. As a usual result, *k*-means or PAM with elbow method have divided the population into a few clusters (3-4), with different variables and for this reason, the elbow method (Figure 2) was not used for the definition of the number of clusters. Therefore, we tried another method like silhouette that gives a different trend of the *k* clusters and uses the average silhouette width criterion instead of the total within-cluster sum of squares.

Figure 2. *Example of elbow method from the auxiliary data. Note the small number of clusters k*



Silhouettes or silhouette index s_i graphically can aid in the interpretation and validation of clusters (Giordani et al., 2020; Rousseeuw, 1987). Usually, Euclidean distances are used for proximities between observations. The Silhouette value is calculated for every single unit *i* where a_i is the average distance between that unit and all the units belonging to the same cluster and b_i denotes the lowest average distance of *i* to any other cluster *i.* In other words, there are two distances, the within-cluster distance α and the closest neighbor distance *b* for every observation:

$$s_i = \frac{b_i - a_i}{\max(b_i, a_i)}, \; i = 1, \ldots, n \quad . \tag{1}$$

The Silhouette takes values from -1≤ s_i ≤1. Values close to 1 mean that the unit is well assigned to the cluster and values close to −1 imply a wrong assignment of the unit to the cluster. If the s_i we can assign the unit to the neighboring cluster. An overall index of clustering goodness is the average Silhouette values of all the units S:

$$\mathrm{S} = \frac{\sum_{i=1}^{n} s_i}{n}. \tag{2}$$

If we compute the S index for different values of *k* we can select the peaks that correspond to the number of clusters *k*.

The final step for clustering is the selection of the algorithm which will partition the data. Hierarchical cluster analysis was tested with Ward's criterion (Ward, 1963) for minimum variance was applied. This method was not proved so helpful in determining the number of clusters. From non-hierarchical algorithms *k*-means and *k*-

medoids were explored. The central idea of *k*-means is to update the cluster-specific mean values clusters (centroids) and after computational iterations to converge based some criteria (Giordani et al., 2020; Xu & Tian, 2015). An evolution to *k*-means is the *k*-medoids and the algorithm which is frequently called PAM. While there is no big difference with normally distributed data, *k*- medoids represent real data units (on the contrary with centroids), can handle better outliers than *k*-means and finally can discretize better the data in specific clusters (Xu & Tian, 2015), which is desirable. An additional step in the cluster analysis could be the post-processing of the cluster results. For instance in the case of the clusters which are not significantly different between them (Tukey's test), then they could have been merged. But in this is study, this is not the case, because we aim for a large number of clusters.

After clustering, we need to do the direct estimates in the formed clusters or post-strata. Twenty-eight direct estimates have been done in post-strata with at least k=10 clusters (except one with k=4). The selection of k is based mainly on the average silhouette width. In the case, that the first peak of suggesting k was less than 10 (usually 2-4) the second big number of k was selected for the estimations. The direct estimates include the estimation of i) the average of the GSV18 for each stratum per ha for the year 2018 and ii) the relative standard error of the mean rRMSE% of within each stratum for the evaluation of the performance of the estimator. The variance of the direct estimator corresponds to the infinite approach (SRS), assuming that are infinite positions for the center of the sample plot, therefore there is no finite population correction. The evaluation in the classical stratification is done with one total error, in our study we have errors for every single cluster. For this reason, the error of the clusters is described in the form of "distributional" errors such as mean and percentiles (median, P_{90} and max) in terms of rRMSE% for the post-strata. P_{90}rRMSE% is considered to be crucial because evaluates the first 90% of the direct estimates. Secondly, to the relative standard error, we expect additionally a large number of k clusters for a given small relative standard error.

3. RESULTS

The clusters were a result of the partitioning of one or two variables. If we use more variables for the cluster analysis we don't necessarily have better clusters, but on the contrary, if we incorporate different variables, then the objects tend to be similar, thus we cannot find clear patterns. The results indicate that a single variable of GSV can give better average silhouette width that ranges mainly from 0,56-0,67, while two combined variables for clustering are 0,38-0,41. Our data can be very well clustered, without this being necessary to also provide good direct estimates. For this reason, we did not select the final clusters based on the criterion of the best partitioning but on the ability to predict the target variable accurately in the clusters. It is worth mentioning that we derive a similar number of clusters with the same variable but different years, for this reason. The censuses of 1988 and 1997 give us almost the same number of clusters when using the same variable such as FirTreeDensity or

FirGSV for clustering. We distinguished five different clusters for their performance. Tables 1-3 give details of the evaluation of the clusters.

Table 1. *Evaluation of clusters with spesific auxiliary data*

			Cluster Evaluation			
General type of Auxiliary variables	**Auxiliary variables for clustering**	Unit of measure	Algorithm for clustering	Method defines clusters	*k* clusters (for direct estimates)	average silhouette width S_i
Census (8)	Forest Density97ha, Forest GSV97ha	m^3/ha & trees/ha	PAM	silhouette	27	0,38
Census (7)	FirTree Density97ha	trees/ha	PAM	silhouette	23	0,57
Census (5)	Forest Density97ha	trees/ha	PAM	silhouette	21	0,61
Census (4)	Forest GSV97ha	m^3/ha	PAM	silhouette	22	0,60
Abiotic factors (13)	Aspect, MeanSlope	degrees & ratio %	PAM	silhouette	31	0,41

From the results we distinguish Cluster 8. The clusters that came from the use of two variables (ForestDensity97ha and ForestGSV97ha), incorporating both the GSV and the tree density per stand. The forest population is partitioned into 27 clusters (Figure 3), with an average silhouette width $S_i=0,38$ (Figure 4), an average of 7,95 stands per cluster and 9,3 plots/cluster. The distribution of direct estimations per cluster/stratum can be characterized by low relative standard errors, with 9,52 mean rRMSE%, 14,07 P90 rRMSE% and 18,24 max rRMSE%, that are acceptable in forest inventories. Additionally, all the clusters had more than two sample plots that is desirable for the estimation of direct estimations for each one. Another great asset of these estimates is the linear relation of the auxiliary variable GSV97 that was estimated for every single cluster with the direct estimated of GSV18. This is very useful later on when we have to use this information in the areal level Fay-Herriot models (Fay & Herriot, 1979).

Figure 3. *k=27 is the optimum number of clusters*

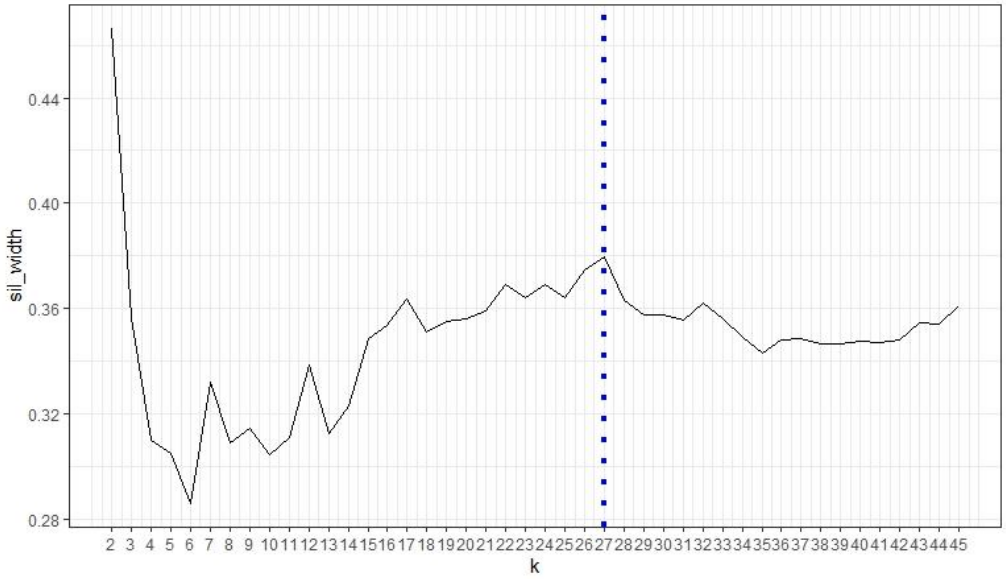


Table 2. *Evaluation of direct estimates regarding different types of rRMSE%s*

	Evaluation of Direct Estimation of GSV18 per stratum/cluster in rRMSE%					
Clusters	min rRMSE %	median rRMSE %	mean rRMSE %	P90 rRMSE %	max rRMSE %	**Corelation with GSV18**
8	4,06	8,24	9,52	14,07	18,24	0,623 GSV (removing 2 cluster outliers)
7	5,17	8,78	11,09	17,08	24,31	0,4 density(removing an outlier), 0,42 GSV (removing an outlier)
5	5,92	8,45	10,60	17,38	24,31	0,61
4	2,12	9,06	9,72	17,43	25,21	0,396 (after removing outliers)
13	1,49	10,55	11,55	17,72	27,85	(-)

Table 3. *Evaluation of direct estimates with the number of N plots/cluster and the number of i stands per cluster*

	N plots/cluster				i stands per cluster			
Clusters	min	median	mean	max	min	median	mean	max
8	2	8	9,30	20	1	5	6,22	13
7	2	8	10,83	26	1	6	7,22	16
5	2	13	11,90	24	1	5	7,95	16
4	2	12	11,41	22	1	8	7,64	15
13	2	8	8,03	19	1	5	5,35	12

Using the stand centroids for describing spatial and temporal contiguity between the stands such as centroids (X, Y coordinates) and after applying *k*-means or *k*-medoids (Figure 5).

Figure 4. *Silhouette widths (average S_i=0,38) for the cluster8 and direct distributions*

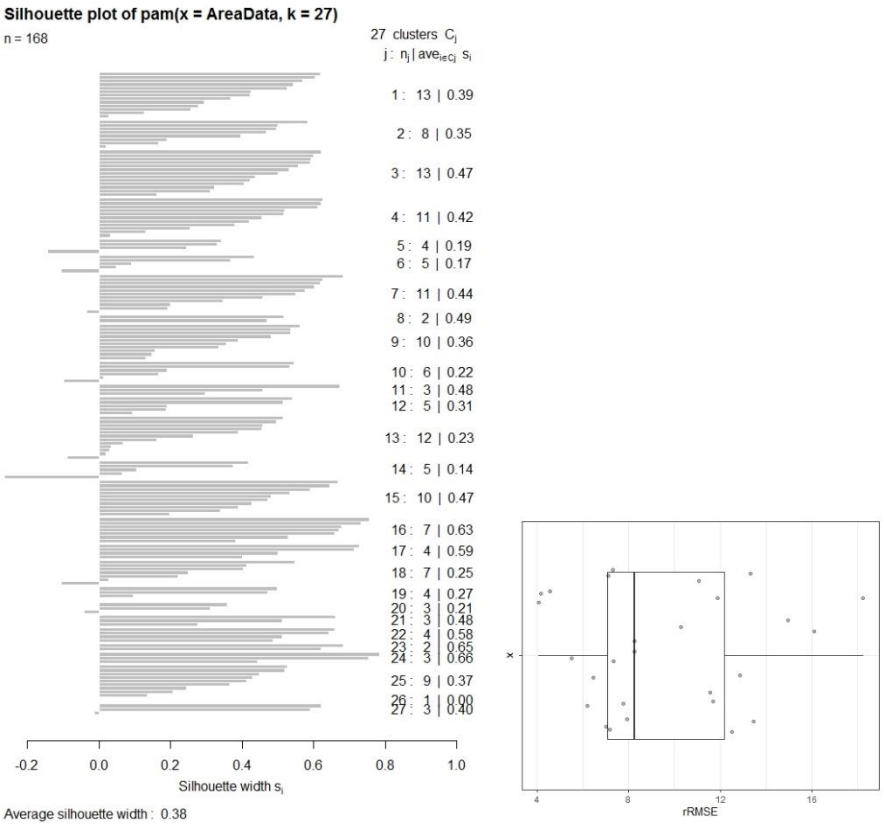
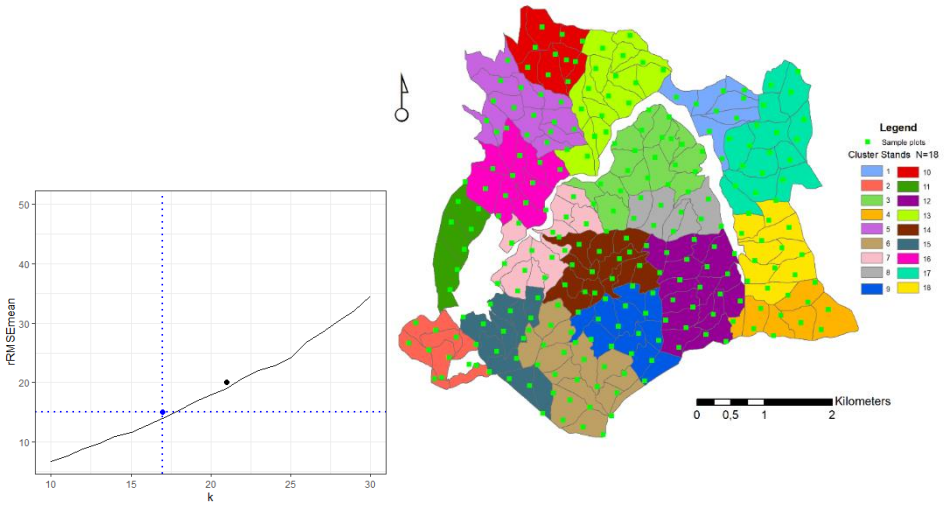


Figure 5. *k=18 clusters selected for direct estimations from stand centroids (note the steady trend of rRMSE%)*



4. DISCUSSION

The clusters were created using, mainly one or two sample plots per stand. The smaller the strata created, the more likely there are groups with only one sample plot. In the case of one or no sample plot, we cannot have error estimations since the variance cannot be calculated with one sample plot. One research question in forest stratification is how are we proving that the clusters/strata are stable through time. The more stable results can be those that came from unchanged auxiliary variables like abiotic factors. If we consider that the forest is managed sustainably, we can consider it "almost" stable through the time the clusters came from census data, assuming that the forest stands will have a similar treatment and for this reason similar forest structure. One important advantage of cluster analysis algorithms comparing some empirical clustering methods that use only one variable is correspondingly the ability to use more than one variable for clustering in the first case instead of just one in the second case. We don't have a substantial problem if we have clusters with similar direct estimates, these could have been merged but our initial aim is to preserve as much as possible many clusters or groups that are close to small areas. Even if the number of clusters is far from the optimum value, using Tukey's method, we can always combine clusters that are not significantly different. It needs to be mentioned that when we use many variables the number of clusters increases and there is the heterogeneity between clusters is decreasing.

5. CONCLUSION

This research indicates that the existence of auxiliary variables in small area levels can be used in cluster analysis and can partition the population into homogenous sub-populations. The advantage of this proposed methodology relies on the ability to provide reliable direct estimates, in clusters of small areas (with one, two sample plots, or even no plots) that are called post-strata as long as the post-strata include at least two sample plots. Cluster analysis can be used to define new larger sub-populations for direct estimates. Cluster information can be incorporated into the SAE models (Fay & Herriot, 1979) in the form of dummy/factor variable or as a covariate when there is linear relation with the variable of interest (Torkashvand et al., 2017; You & Chapman, 2006; Zulkarnain et al., 2020). Additionally, we conclude that silhouette outperformed to indicate a larger number of clusters N than the elbow method. The elbow method is not suitable for estimating the number of clusters *k* because distinguishes around three-four clusters only.

ΠΕΡΙΛΗΨΗ

Οι δασικές απογραφές παρέχουν όλες τις βασικές πληροφορίες οι οποίες είναι απαραίτητες για την αειφόρο διαχείριση των δασικών οικοσυστημάτων. Οι πληροφορίες αυτές περιλαμβάνουν την εκτίμηση δασοβιομετρικών μεταβλητών, όπως του ξυλώδους όγκου (ξυλαπόθεμα) που είναι η σημαντικότερη. Στην παρούσα εργασία καταβλήθηκε προσπάθεια ορισμού στρωμάτων μετά από ομαδοποίηση (συσταδοποίηση, clustering) ομοιογενών δασικών συστάδων/τμημάτων με την ανάλυση κατά συστάδες. Είκοσι οχτώ διαφορετικές ομάδες προέκυψαν χρησιμοποιώντας διαφορετικές βοηθητικές πληροφορίες και χρησιμοποιόντας διαφορετικές μεθόδους ομαδοποίησης, περιλαμβάνοντας την κατάλληλη επιλογή του πλήθος των ομάδων *k* και τους κατάλληλους αλγορίθμους ομαδοποίησης. Η αξιολόγηση των πλήθους των όμάδων βασίστηκε κυρίως στο πλάτος silhouette. Η ομαδοποίηση χρησιμποιήθηκε υποστηρικτηκά ως εργαλείο για τη φάση των εκτιμήσεων, για την μεγένθηση του μεγέθους των μικρών εκτάσεων και του δείγματος και αντίστοιχα. Κατ' αυτόν τον τρόπο η τελική αξιολόγηση των άμεσων μετρήσεων των ομάδων (μετα-στρωμάτων) βασίστηκε στις άμεσες εκτιμήσεις σκοπεύοντας στην έλαχιστη σχετική τυπική απόκλιση (rRMSE%) του μέσου όρου του ξυλαποθέματος.

REFERENCES

Adolfsson, A., Ackerman, M., & Brownstein, N. C. (2019). To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognition*, **88**, 13-26. https://doi.org/10.1016/j.patcog.2018.10.026

Breidenbach, J., Magnussen, S., Rahlf, J., & Astrup, R. (2018). Unit-level and area-level small area estimation under heteroscedasticity using digital aerial photogrammetry data. *Remote Sensing of Environment*, **212**, 199-211. https://doi.org/10.1016/j.rse.2018.04.028

FAO. (2004). Global forest resources assessment update 2005: terms and definitions. In: FAO Rome, Italy.

Fay, R. E., & Herriot, R. A. (1979). Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association,* **74** (366), 269-277. doi:10.2307/2286322

Georgakis, A., & Stamatellos, G. (2020). Sampling Design Contribution to Small Area Estimation Procedure in Forest Inventories. *Modern Concepts & Developments in Agronomy, 7*(1). doi:10.31031/MCDA.2020.07.000654

Giordani, P., Ferraro, M. B., & Martella, F. (2020). *An Introduction to Clustering with R*: Springer.

Goerndt, M. E. (2010). *Comparison and analysis of small area estimation methods for improving estimates of selected forest attributes.* (Doctor of Philosophy (Ph.D.) Doctoral Dissertation), Oregon State University, Retrieved from Explorer Site::Forest Explorer Available from Oregon State University ScholarsArchive@OSU database.

Golder, P. A., & Yeomans, K. A. (1973). The Use of Cluster Analysis for Stratification. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **22**(2), 213-219. doi:10.2307/2346922

Hill, A., Mandallaz, D., & Langshausen, J. (2018). A Double-Sampling Extension of the German National Forest Inventory for Design-Based Small Area Estimation on Forest District Levels. *Remote Sensing,* **10**(7), 1052.

Kaufman, L., & Rousseeuw, P. J. (1990). Partitioning around medoids (program pam). In *Finding groups in data: an introduction to cluster analysis* (Vol. 344, pp. 68-125).

Lawson, R. G., & Jurs, P. C. (1990). New index for clustering tendency and its application to chemical problems. *Journal of chemical information computer sciences,* **30**(1), 36-41.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2014). Package 'cluster'.

Magnussen, S., Mauro, F., Breidenbach, J., Lanz, A., & Kändler, G. (2017). Area-level analysis of forest inventory variables. *European Journal of Forest Research,* **136**(5), 839-855. doi:10.1007/s10342-017-1074-z

Potapov, P., Li, X., Hernandez-Serna, A., Tyukavina, A., Hansen, M. C., Kommareddy, A., . . . Hofton, M. (2021). Mapping global forest canopy height through integration of GEDI and Landsat data. *Remote Sensing of Environment,* **253**, 112165. https://doi.org/10.1016/j.rse.2020.112165

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org/index.html

Rao, J. N., & Molina, I. (2015). *Small area estimation*: John Wiley & Sons, Inc.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

Torkashvand, E., Jozani, M. J., & Torabi, M. (2017). Clustering in small area estimation with area level linear mixed models. *Journal of the Royal Statistical*

Society: Series A (Statistics in Society), **180**(4), 1253-1279. doi:10.1111/rssa.12308

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association,* **58**(301), 236-244. doi:10.1080/01621459.1963.10500845

Westfall, J. A., Patterson, P. L., & Coulston, J. W. (2011). Post-stratified estimation: within-strata and total sample size recommendations. *Canadian Journal of Forest Research, 41*(5), 1130-1139. doi:10.1139/x11-031

Wright, K., YiLan, L., & RuTong, Z. (2021). Package 'clustertend'.

Xu, D., & Tian, Y. (2015). A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science,* **2**(2), 165-193. doi:10.1007/s40745-015-0040-1

You, Y., & Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology,* **32**(1), 97.

Zulkarnain, R., Jayanti, D., & Listianingrum, T. (2020). Improving the quality of disaggregated SDG indicators with cluster information for small area estimates. *Statistical Journal of the IAOS,* **36**, 955-961. doi:10.3233/SJI-200741

Greek Statistical Institute
Proceedings of the 33rd Panhellenic Statistics Conference (2021), pp.248-261

FURTHER IMPROVEMENTS OF GROWING STOCK VOLUME ESTIMATES AT STRATUM-LEVEL WITH THE APPLICATION OF FAY-HERRIOT MODEL

Georgakis Aristeidis
Aristotle University of Thessaloniki, School of Forestry and Natural Environment, Thessaloniki, Lab of Forest Biometry, PC-54124, Thessaloniki,
e-mail: arisgeorg@for.auth.gr

ABSTRACT

The sustainability of the forests is preserved by the principles of forest management and from the information derived from the management forest inventories (FIs). One of the most important target variables of FIs is the growing stock volume. While the reliability of direct estimates is achievable for the total population, the estimates for small areas (geographic subpopulations), with very small sample sizes is a challenge. This problem can be overcome by small area estimation (SAE) models, which "borrow strength" from related areas and auxiliary covariates. This work explores the effectiveness of the Fay-Herriot (FH) area-level model to produce small area statistics utilizing past census covariates in the forest strata subpopulations. The results suggest the effectiveness of the FH model to provide reliable estimates at the stratum-level, with an average 58% CV reduction in comparison to the direct estimates.

Keywords: EBLUP area-level model, post-strata, past census covariates, small area estimation

1. INTRODUCTION

Small area estimation (SAE) is a method or set of techniques that aim to produce reliable small area statistics for subpopulations with small sample sizes (few sample plots) when design-based direct estimates cannot be obtained (Goerndt, Monleon, & Temesgen, 2011; Rao & Molina, 2015). In forest inventories (FIs) a sample plot or sample unit is a portion of forest land, consisting of a group of measured trees. SAE applications received much attention in the last decades in various scientific fields such as poverty mapping in the economy, epidemiology mapping, crop yields in agriculture and forest attributes estimates in forestry (Battese, Harter, & Fuller, 1988; Breidenbach, Magnussen, Rahlf, & Astrup, 2018; Goerndt et al., 2011; Rao & Molina, 2015). Growing stock volume (GSV) and aboveground forest biomass are the most important variables of interest. FIs can be further distinguished in national and management. Management FIs aim to provide reliable (precise and accurate) biometrical estimations at small areas (subpopulations, geographical domains) or

management units (forest stands or compartments). SAE can contribute to that direction, in the cases that the sample size cannot be increased, due to time and budgetary constraints, we can "borrow strength" by existing linearly related auxiliary variables (covariates or predictors) for producing model-based or model-assisted statistics. SAE does not rely solely on the data of the domain but puts into account the sample information outside of the specific target domain. Depending on the model, SAE can provide estimations even in unplanned domains with no sample plots.

Model selection is the most crucial part of the SAE techniques, after the definition of the variable of interest, the area of interest (small area), and the auxiliary information that will be used, under (usually) the existing sampling design that generally is not taken into consideration (Georgakis & Stamatellos, 2020). When direct design-based small area estimates are not feasible due to the small sample size, the model-based SAE approach is commonly used (Rao & Molina, 2015). The model-based selection of SAE techniques is inseparable from the availability and type of auxiliary data at different levels.

Model-based estimators in SAE literature can be distinguished in two broad categories, based on the type of auxiliary covariates they use, i) the unit-level models where covariates are available at the unit-level or the sample/pixel level in forestry (Mauro, Molina, García-Abril, Valbuena, & Ayuga-Téllez, 2016), and ii) the area-level models, first described by Fay-Herriot (FH) (1979), that use aggregated area-level covariates in the small area of interest or at forest stand/stratum level in forestry (Goerndt et al., 2011; Magnussen, Mauro, Breidenbach, Lanz, & Kändler, 2017). In FIs the most common auxiliary variables come from remote sensing data. Unit-level models usually are applied in FIs with the area-based approach (ABA) by using fine-resolution 3D remote sensing data such as light detection and ranging (LiDAR) or 3D point clouds acquired by digital aerial photogrammetry (Breidenbach et al., 2018; Ver Planck, Finley, Kershaw, Weiskittel, & Kress, 2018).

Unit-level models generally, provide more accurate estimates than area-level models, but with the condition of precise sample plot positioning that links correctly to the available auxiliary data (Breidenbach et al., 2018; Mauro, Monleon, Temesgen, & Ford, 2017). Area-level models, on the other hand, have not been explored extensively, but there is a recent increment in FIs applications (Chandra & Chandra, 2020; Green, Burkhart, Coulston, & Radtke, 2019), having the advantage to make use of the variable radius plots (Temesgen et al., 2021) that typically is not used in unit-level models, and give better estimates than post-stratified estimators (Coulston et al., 2021).

The aim of this research is the further improvements of the direct GSV estimates at the stratum-level (Georgakis, to appear), with the application of the FH area-level model. Past census data that linearly relate to the average means (direct estimates) of the strata (groups of compartments) were used as covariates for the application of the FH model.

2. MATERIALS AND METHODS

2.1 Data

The study area is the University forest of Pertouli in Greece with the dominant species the *Abies borisii-regis* (Mattf.) (hybrid fir) that form an uneven-aged forest ecosystem. The population consists of 27 large post-strata that were constructed after clustering the smaller forest compartments (Georgakis, to appear). The area of the sample plots/units were one hectare (*ha*) and the variable of interest (target variable) is the GSV m^3/ha. There are no out-of-sample domains (unplanned) and all the domains have at least two sample plots for variance estimates. Every stratum constitutes on average from 9,3 sample plots, in 6,22 aggregated (clustered) forest compartments, with an average forest area of 81,83 ha. The direct estimates of GSV were derived at the stratum-level (cluster-level). Past census forest attributes were used as covariates in the FH models for deriving area-level (stratum-level) statistics. We use the following abbreviations for the auxiliary data

- *DirectClusterVOl18:* estimates of GSV (m^3/ha) at post-stratum level (systematic sampling 2018 with 1% sampling intensity and sample unit =0,1)
- *ClusterVOl97:* mean GSV /ha (m^3/ha) per stratum (census 1997)
- *ClusterDensity97:* density trees/ha of all the trees per stratum (census 1997)
- *ClusterDensity88:* density trees/ha of all the trees per stratum (census 1988)
- *DirectClusters_2:* categorical variable derived from the clustering of *DirectClusterVOl18*

The FH model needs covariates that are linearly related. To meet this assumption, three "outliers" were removed and thus the estimations have been done for 24 strata. The correlation of the *DirectClusterVOl18* and *ClusterVOl97* was 0,62. The SAE statistics were produced with the statistical language R (R Core Team, 2021) and the emdi package (Kreutzmann et al., 2019). One important reason for the use of emdi package is the additional feature of model variance estimation in the case of a small number of domains, compared to other existing packages (Yoshimori & Lahiri, 2014).

2.2 SAE notation

In this section, the SAE notation is described and based on Rao & Molina (2015). First of all, the direct estimates of domains (post-strata) are described, secondly the FH model estimates, and third the additional improvement of the FH estimates by incorporating dummy/categorical auxiliary variables derived from cluster analysis on direct domain estimates.

2.2.1 Direct estimates

Suppose the population U consists of U_i of $i = 1, \dots, D$ subpopulations, the so-called domains or small areas (strata), and N_i distinct units and sample data s_i with sample size $n_i \geq 2$ in every domain U_i. For each domain only, a small number of sample units received $j = 1, \dots, n_i$, where *j*-th is the population unit in domain *i*. The variable

of interest θ_i (eq. 1) are the means of GSV $\bar{y}_i$ per domain, and y_{ij} is the measurement of the variable of interest for individual *j* (sample plot) within area *i*.

$$\theta_i = \hat{\bar{Y}}_i^{DIR} = \bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} \quad (1)$$

The variance of the direct estimator of the mean (eq. 2), assumed to be equal to that of simple random sampling with replacement without taking into account the finite population correction, due to extremely small sample size, or as a more appropriate approach in FIs (Mandallaz, 2008) and is design-unbiased

$$\hat{V}\left(\hat{\bar{Y}}_i^{DIR}\right) = \frac{S_i^2}{n_i} \quad (2)$$

with sample variance

$$S_i^2 = \sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_i\right)^2/(n-1). \quad (3)$$

2.2.2 Fay-Herriot model

The FH model, firstly described by Fay and Herriot (1979), is a special case of a linear mixed-effects model. The FH model can be applied when aggregated area-level auxiliary data related to the direct estimates of the areas of interest (domains). The basic FH model is a two-stage estimator. Since true values θ_i are not observable, our data will be the direct estimates $\hat{\theta}_i^{DIR}$ (left-hand side of eq.4). These estimates have an error (right-hand side of eq.4) and might be different for each area because of different sample sizes in the areas of interest. In the first stage, the following "sampling" model represents the sampling error e_i of direct estimates, where sampling variance $\sigma_{e_i}^2$ of the direct estimator $\hat{\theta}_i^{DIR}$ given for θ_i, assumed to be known for all i domains.

$$\hat{\theta}_i^{DIR} = \theta_i + e_i, \quad e_i \overset{ind}{\sim} N\left(0, \sigma_{e_i}^2\right),\ i = 1, \dots, D \quad (4)$$

In the second stage, we assume that the true values $\theta_i = g(\bar{Y}_i)$, for some specified $g(.)$, are assumed to be linearly (cross-sectionally) related with a vector of area-level auxiliary variables $\boldsymbol{x_i} = \left(x_{1i}, \dots, x_{pi}\right)'$, through a linear model (left-hand side of eq.5). This part is called the linking model because it links all the small areas (*borrows strength*) through the common model parameter $\boldsymbol{\beta} = \left(\beta_1, \dots, \beta_p\right)'$ which is a $p \times 1$ vector of regression coefficients (D > p) and the $\boldsymbol{x_i}$ area-level covariates. On the right-hand side of eq. 5 the area-specific random effects υ_i assumed to be independent and identically distributed (iid) with model variance $V_i(\upsilon_i) = \sigma_\upsilon^2\ (\geq 0)$ and also independent of the sampling errors e_i. The b_i's are known positive constants.

$$\theta_i = \boldsymbol{x_i'\beta} + b_i\upsilon_i, \quad \upsilon_i \overset{iid}{\sim} N(0, \sigma_\upsilon^2),\ i = 1, \dots, D, \quad (5)$$

Combining eq. 4 and 5, we obtain

$$\hat{\theta}_i^{DIR} = \boldsymbol{x_i'\beta} + b_i\upsilon_i + e_i\,, \quad (6)$$

which involves both design, e_i, as well as model errors, v_i. The best linear unbiased predictor (BLUP) of θ_i is given by

$$\tilde{\theta}_i = \boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \gamma_i\left(\hat{\theta}_i^{DIR} - \boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}\right) = \gamma_i\hat{\theta}_i^{DIR} + (1-\gamma_i)\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} \quad (7)$$

with $\tilde{\boldsymbol{\beta}}$ is the best linear unbiased estimator of regression coefficients $\boldsymbol{\beta}$ estimated by the weighted least squares method. One good property of the BLUP is the minimization of the MSE. The BLUP is based on known model variance $\sigma_v^2 b_i^2$ that is unknown in practice, with the further estimation of the unknown variance component $\hat{\sigma}_v^2 b_i^2$ we obtain the Empirical BLUP (EBLUP)

$$\hat{\theta}_i^{FH} = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + b_i\hat{v}_i = \hat{\gamma}_i\hat{\theta}_i^{DIR} + (1-\hat{\gamma}_i)\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}, \qquad \hat{\gamma}_i = \frac{\hat{\sigma}_v^2 b_i^2}{\hat{\sigma}_v^2 b_i^2 + \sigma_{\hat{e}_i}^2} \quad (8)$$

with estimations of $\hat{\gamma}_i$ and $\widehat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\hat{\sigma}_v^2)$ for area i. From eq.8 it is visible the weighted average of the direct estimator $\hat{\theta}_i^{DIR}$ and the regression-synthetic estimator $\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$, depends on the single component $\hat{\gamma}_i$ or shrinkage factor $\hat{\gamma}_i (0 \leq \hat{\gamma}_i \leq 1$. When the sampling variance is small in an area i, more weight is given to the direct estimator and $\hat{\gamma}_i$ is comparatively high. Conversely, if the sampling variance is large for this area, the direct estimator is considered to be unreliable, thus more weight is given to the synthetic part, with comparatively low $\hat{\gamma}_i$.

2.2.3 MSE

The model parameters σ_v^2 and $\boldsymbol{\beta}$ are usually estimated by Maximum Likelihood methods (ML/RELM), based on the normal likelihood. The accuracy or uncertainty of an EBLUP $\hat{\theta}_d$ can be assessed in the form of estimation error by estimating the Mean Square Error (MSE). The MSE of the BLUP for known model variance σ_v^2 is given by:

$$\mathrm{MSE}\left(\tilde{\theta}_i\right) = E\left(\tilde{\theta}_i - \theta_i\right) = g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2) \quad (9)$$

where $g_{1i}(\sigma_v^2) = \gamma_i\sigma_{e_i}^2$ is the leading term of $\mathrm{MSE}\left(\tilde{\theta}_i\right)$, incorporates the prediction of the random effect v_i and it is of $O(1)$ for large *D*. If we use the restricted maximum likelihood (REML) estimator $\hat{\sigma}_{v\,REML}^2$, instead of ML, a second-order unbiased estimator of $\mathrm{MSE}(\hat{\theta}_i)$ is given by

$$mse_{REML}(\hat{\theta}_i) = g_{1i}(\hat{\sigma}_{v\,REML}^2) + g_{2i}(\hat{\sigma}_{v\,REML}^2) + 2g_{3i}(\hat{\sigma}_{v\,REML}^2) \quad (10)$$

(Prasad & Rao, 1990), which includes the uncertainty arising from the estimation of σ_v^2. In the case of a few small areas (domains) the adjusted ML or/and adjusted REML (Li & Lahiri, 2010; Yoshimori & Lahiri, 2014) can be used. In this study, the adjusted REML of Yoshimori and Lahiri (2014) ("amrl_yl") was used as a new and improved version of Li and Lahiri (2010), characterized by less bias than the adjusted ML, in estimating both the model variance of the random effects and the shrinkage factors. The amrl_yl solve the practical problem of zero estimated $\hat{\sigma}_v^2$ compared to the

maximum likelihood BLUP estimator that cannot be taken into account the heterogeneity among the areas in this case (Sugasawa & Kubokawa, 2020).

2.3 Clustering the post-strata

Strata (groups of compartments) were created with modern cluster analysis techniques by the aggregation of homogenous forest compartments which are characterized by extremely small size with 1-3 sample plots per compartment (Georgakis, to appear). Clustering analysis aggregated the small areas into larger ones called (post-)strata, using available auxiliary variables that can describe the heterogeneity of these small areas. The suggested method for this purpose includes the following process steps: a) select the auxiliary variable(s), b) preprocess the auxiliary data, c) select the algorithm for clustering (usually K-medoids or PAM (Kaufman & Rousseeuw, 1990)), d) find the number of clusters (*k*) (Rousseeuw, 1987), e) evaluate the direct estimates based on relative standard error, and f) explore the linear correlation of the aggregated auxiliary variables with direct estimates, if this is feasible.

The same philosophy of clustering can be applied additionally for deriving auxiliary information in the stratum-level. A categorical variable (dummy, factor) derived from the clustered domains (larger homogenous strata) can be incorporated as auxiliary information in the FH model for further reduction of the MSE (You & Chapman, 2006; Zulkarnain, Jayanti, & Listianingrum, 2020). Further improvements of the FH model can be achieved by the incorporation of the dummy variable (You & Chapman, 2006), which classifies direct estimates of small areas into larger ones. Herein the 24 small areas were grouped into two large homogenous domains, after clustering the direct estimates with a k-means (or PAM) algorithm. The number of clusters was derived visually from the elbow graph that illustrates the total within-cluster sum of squares (Giordani, Ferraro, & Martella, 2020) across every cluster.

2.4 Model Performance

The Mean Square Error (MSE) = bias^2 (systematic error) + variance (random error), is the most common uncertainty measure for area-specific prediction (Tzavidis, Zhang, Luna, Schmid, & Rojas-Perilla, 2018). The main reason for using indirect estimators, like FH, is the reduction in MSE (Rao & Molina, 2015). More interpretable measures of uncertainty reports are the relative square root of MSE ($\mathrm{RRMSE} = \sqrt{\mathrm{MSE}/\bar{y}_i}$), the (percentage) coefficient of variation (CV%) of means $\bar{y}_i$ or the relative standard error, that is commonly used in FIs applications. SAE techniques are based on model assumptions, and therefore evaluation of the violation of model assumptions is crucial.

The evaluation of the models initially was based on the model variance, the normality of the standardized residuals, and the normality of random effects by applying the Shapiro-Wilks test. Additionally, model selection was based on information criteria such as the Akaike (AIC), the Bayesian (BIC), the biased corrected Kullback

Information Criterion (KICb2), and on the explanatory measures of Adjusted R^2 (Lahiri & Suntornchost, 2015). All the above diagnostics give a picture of the overall performance but do not provide the whole picture concerning individual forest strata. The best way to assess the performance of FH-EBLUP estimators is to observe the bias for each small area (Goerndt et al., 2011). Last, but not least, the FH model was checked for a potential bias by visual examination of the Brown et al. (2001) graphical diagnostic. An estimator is considered unbiased when the illustrated regression line of direct estimates on the *X-axis* and the model estimates on the *Y*-axis are close to the identity line with *Y=X* (Brown, Chambers, Heady, & Heasman, 2001; Kreutzmann et al., 2019). All the SAE model-based estimators inherently have some bias for the sake of a smaller variance of the estimator, but this should be checked to be as much as unbiased, otherwise, it will lead to misleading estimates.

3. RESULTS

Four different FH-EBLUP estimators (*fh1*, *fh2*, *fh3*, *fh4*) were the result of different area-level auxiliary covariates having the following form

fh1: *DirectClusterVOl18 ~ ClusterVOl97*
fh2: *DirectClusterVOl18 ~ ClusterVOl97 + ClusterDensity97*
fh3: *DirectClusterVOl18 ~ ClusterVOl97 + ClusterDensity97 + ClusterDensity88*
fh4: *DirectClusterVOl18 ~ ClusterVOl97 + ClusterDensity97 + DirectClusters_2*

Interpreting the results of *Table 1*, we understand that the *fh1* model, with one covariate (*ClusterVOl97*), does not perform well mainly due to large model variance ($\hat{\sigma}_v^2 = 171{,}03$) and the assumption of normality of the random effects are not met. By adding the second covariate (*ClusterDensity97*) the *fh2* model decreases the model variance but the problem of not normally distributed random area effects (heteroscedasticity) remains. The *fh3* has dramatic improvement compared to the previous models by adding one more covariate (*ClusterDensity88*). Both *fh3* and *fh4* have substantially lower model variance compared to *fh1* and *fh2*. The *fh3* has the minimum model variance, $\hat{\sigma}_v^2 = 4{,}41$, additionally the auxiliary variables *ClusterDensity88* and *ClusterDensity97* acted positively in "normalizing" the random area effects. But still, two negative properties for *fh3* are the model bias (explained below) and the decrease of $AdjR^2$.

Table 1. *Model performance*

	Variance estimation				**Shapiro-Wilks test p-value**	**Explanatory measures**			
model	method	MSE estimation	Estimated model variance	Standardized residuals	Random effects	AIC	BIC	KICb2	$AdjR^2$
fh1	REML	Prasad & Rao, 1990	171,03	0,34	0,0001	232,74	236,27	239,20	0,782
fh2	REML	Prasad & Rao, 1990	114,99	0,67	0,0012	233,37	238,09	240,19	0,734

fh3	amrl_yl	Prasad & Rao, 1990	**4,41**	0,73	0,8067	226,91	232,80	236,82	0,674
fh4	amrl_yl	Prasad & Rao, 1990	6,36	0,68	0,4994	**223,81**	**229,70**	**233,92**	**0,998**

The best model performance was found in *fh4*, after the incorporation of the dummy variable derived from clustering. The *fh4* has normally distributed standardized residuals and random area effects (*Figure 1*), has a very small model variance, minimum information criteria (AIC, BIC, KICb2), and best explanatory measure of AdjR^2 (*Table 1*).

Figure 1. *Normal distribution of standardized residuals & random area effects of fh4*

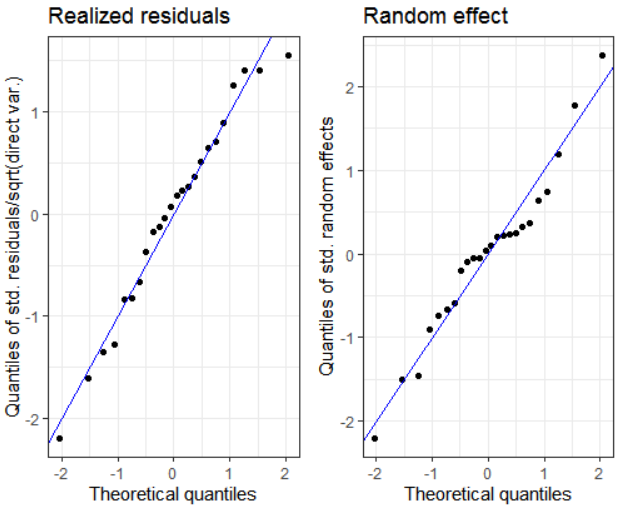
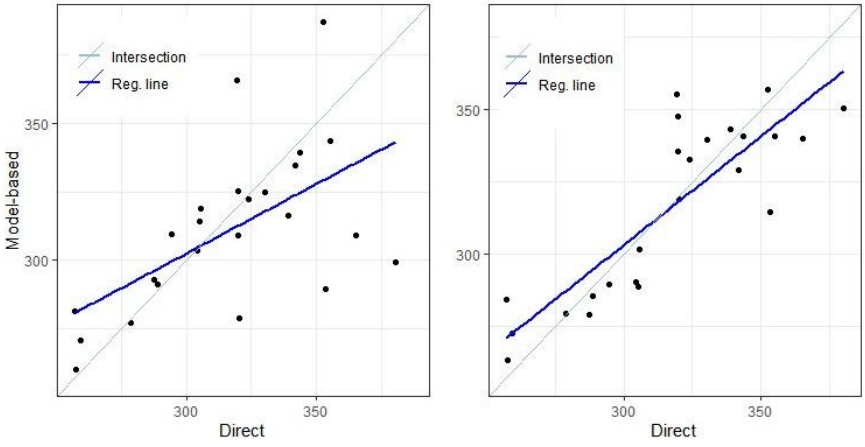
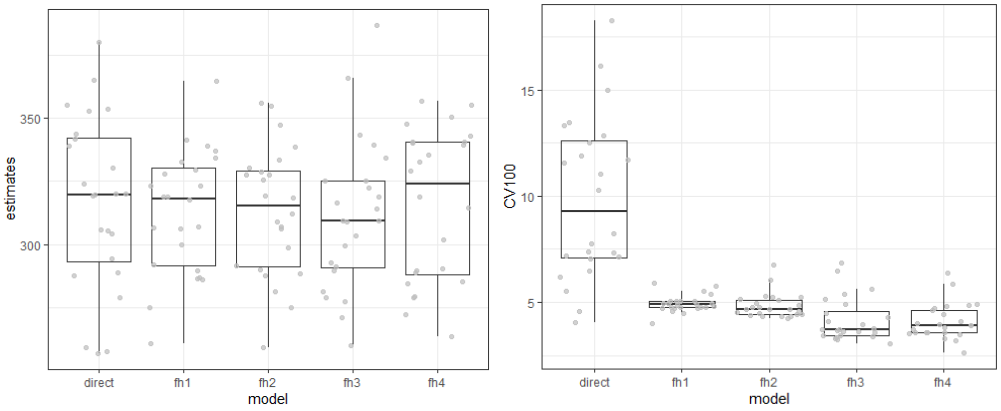


Figure 2. *The fh4 model (right plot) is nearly unbiased with the incorporation of the dummy variable compared to fh3 without (left plot). Nearly unbiased estimates when the blue regression line is close to the "intersection" line 1:1 and the points close to the blue line.*



Further quality assessment was the comparison of the model-based FH estimates with the direct estimates based on a goodness-of-fit test proposed by Brown et al. (2001) and the correlation coefficient of the synthetic part and the direct estimator. All the models passed the Brown test, by not rejecting the null hypothesis that the FH-EBLUP estimates do not differ significantly from the direct estimates. The scatter plot in *Figure 2* shows the direct and model-based *fh3* and *fh4* point estimates, the fitted regression line, and the "intersection" line ($Y = X$). The regression line of *fh4* estimates (right-side of *Figure 2*) is closer to the intersection line which is translated to small deviations from the direct estimates. The correlation coefficient between *fh4* model-based and direct estimates is 0,85. The incorporation of a dummy variable seems to partition the data into two groups in *fh4 estimates*. On the other hand, the *fh3* model-based estimates (left-side of *Figure 2*) deviate more from direct estimates. The vast majority of the model-based values are close to each other with shorter range, are closer to the average model estimates, and thus model bias can be assumed. Additionally, the equality between the total (grand) means of the direct estimates and model estimates was tested. Assuming that the grand mean of direct estimates can be predicted with high accuracy, due to the large sample size, no significant differences to model predictions have been found. The mean & median of GSV direct estimates were 316,78 m^3/ha & 319,80m^3/ha and respectively the *fh4* model-based estimates were 315,79 m^3/ha & 323,90 m^3/ha.

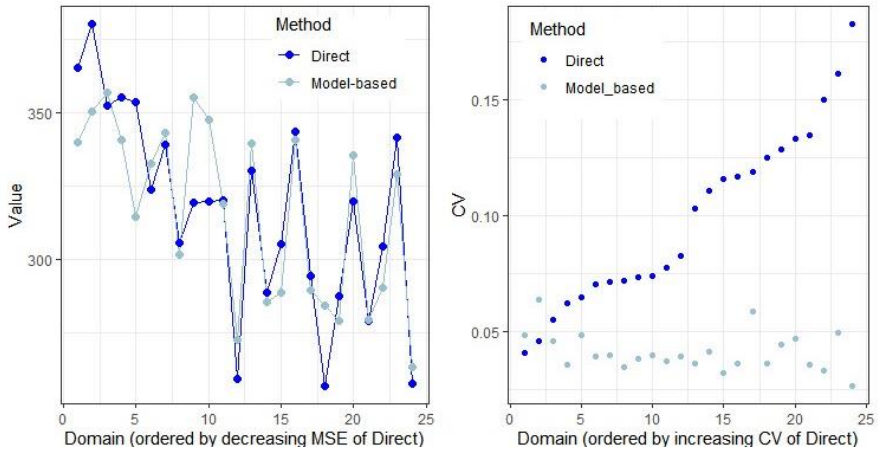
Figure 3. *Distribution of direct and Fay-Herriot model estimates (fh1, fh2, fh3, fh4) on the left side, and the corresponding error in percentage coefficient of variation (CV%) on the right side (each point represents a domain estimation)*



The results on the right side of *Figure 3* illustrate the substantial uncertainty reduction of estimates, the average direct 9,86 CV% reduced to 4,10 CV% in the FH model (*fh4*), which is translated to 58% relative reduction of CV. *Figure 4* illustrates the comparison of the direct and the *fh4* model-based estimates and the corresponding CV. On the left side of *Figure 4*, we see how the extreme direct estimates are smoothed with the application of the FH model. Additionally, on the right side of *Figure 3*, we see a substantial reduction of uncertainty, having all the domains (strata)

estimates CV smaller of 10% which is a highly desirable property in FIs, where this threshold can be up to 15% for stand-level estimates (Mauro et al., 2016).

Figure 4. *Line plot of direct and model-based FH point estimates per domain and scatter plot of the CV% on the right.*



4. DISCUSSION

FH area-level modeling that relates the direct estimates (means, totals) with area-specific auxiliary variables is a standard SAE technique for obtaining small area statistics (Rao & Molina, 2015). Unit-level models have been expected to lead to more accurate estimates, but these require good unit-level auxiliary covariates to link with the sample plots. In practice, this is not always feasible for the following reasons, i. auxiliary covariates are not available at unit-level, ii, co-registration problems exist because coordinates of fixed area plot locations are inaccurate or do not exist, or iii. linking limits arise from the type of sample plots in the case of variable radius plots (Magnussen et al., 2017; Mauro et al., 2016; Namazi-Rad & Steel, 2015; Temesgen et al., 2021). The mentioned problems can be overcome with the area-level models.

Typically, the FH models can be applied if there are at least two sample plots per domain necessary for the estimation of the sampling variance $\sigma_{e_i}^2$. Basic problems that arise with the extremely small sample size are a) low accuracy of direct estimates per domain, b) rough estimation of sampling variance of the mean from the unit level sample plots, c) bad correlation with area-level auxiliary covariates, d) FH models cannot be applied, typically, in domains (stands/compartments) with one sample plot per domain due to lack of sampling error estimation, and e) FH model cannot give predictions for non-sample/unplanned areas because direct estimates are necessary. In the case of unplanned areas with no sample units only regression synthetic estimators $\theta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ can provide estimates, using only auxiliary covariates $\boldsymbol{x}_i'$ that assuming that all the local variation is explained by these (Rao & Molina, 2015, p. 77).

Most of the above problems can be overcome if the small areas will be increased after clustering/grouping similar ones to larger ones (post-strata). In parallel work (Georgakis, to appear) domains with no or extremely small sample size aggregated to larger homogenous groups or strata via cluster analysis. Further improvements of GSV estimates at stratum-level with the application of the FH area-level model have been done. The FH substantially decreases the relative standard error for each stratum as illustrated on the right side of *Figure 3*. The model-based approach that was used strongly depends on the fulfillment of the model assumptions. For this reason, all the important assumptions were met. First, there is a linear correlation of the target variable with the auxiliary covariates (Rao & Molina, 2015). Secondly, there is a small model variance, third, the normality of the standardized residuals and random effects are met. Fourth, the model selection is based on the minimum information criteria (AIC, BIC, KICb2) and high $AdjR^2$ (Harmening, Kreutzmann, Pannier, Salvati, & Schmid; Lahiri & Suntornchost, 2015). Fifth, almost unbiased model-based estimates were achieved (Brown et al., 2001). Sixth, while differences between direct and model-based estimates exist at the small area level, there are no significant differences between the grand means of direct and model-based estimates. Lastly, the incorporation of a categorical variable (*fh4*), by grouping the direct estimates with cluster analysis into two larger homogenous groups of strata, gave the best model performance.

The estimation of sampling variance is based on the average 9 sample units for each homogenous stratum and is close to 10 minimum within-strata sample size that suggested in previous research (Westfall, Patterson, & Coulston, 2011). Assuming that sampling variance estimations were stable no further smoothing sampling variance was applied such as the *generalized variance function* approach (Wolter, 2007) or a weighted mean variance based on the size of the domain (Goerndt et al., 2011).

5. CONCLUSION

Past census data was used successfully in FH-EBLUP estimates, having good linear relation with direct estimates of sample plots in the stratum-level, and used also to produce categorical variables that further improved the efficiency of the FH model. Finally, the prediction of GSV was achieved by decreasing the uncertainties in terms of CV% for the FH estimates compared to the direct estimates. The results suggest the effectiveness of the FH model to provide reliable estimates at the stratum-level, with an average 58% CV reduction in comparison to the direct estimates. The uncertainty of average direct 9,86 CV% reduced substantially to 4,10 CV% in the FH model (*fh4*), which is translated to 58% relative reduction of CV. This threshold of uncertainty is very good and meets the highest FI demands.

Acknowledgments: The author would like to thank the unknown reviewer for the important and substantial comments and suggestions that helped to improve this paper.

ΠΕΡΙΛΗΨΗ

Η αειφορία των δασών πραγματώνεται από τις αρχές της διαχείρισης των δασών και από τις πληροφορίες που προέρχονται από τις απογραφές των διαχειριζόμενων δασών. Μια από τις σημαντικότερες μεταβλητές των δασικών απογραφών είναι το ξυλαπόθεμα (ξυλώδης όγκος). Ενώ η αξιοπιστία των άμεσων εκτιμήσεων είναι επιτεύξιμη για το σύνολο του πληθυσμού, οι εκτιμήσεις σε μικρές περιοχές (γεωγραφικούς υποπληθυσμούς) με πολύ μικρό μέγεθος δείγματος είναι πρόκληση. Αυτό το πρόβλημα μπορεί να ξεπεραστεί με μοντέλα μικρής έκτασης (SAE), τα οποία «δανείζονται δύναμη» από σχετικές περιοχές και βοηθητικές συμμεταβλητές. Αυτή η εργασία διερευνά την αποτελεσματικότητα του μοντέλου σε επίπεδο περιοχής Fay-Herriot (FH) για την παραγωγή στατιστικών μικρών περιοχών, χρησιμοποιώντας συμμεταβλητές προηγούμενων απογραφών στους υποπληθυσμούς των δασικών στρωμάτων. Τα αποτελέσματα υποδεικνύουν την αποτελεσματικότητα του μοντέλου FH για την παροχή αξιόπιστων εκτιμήσεων σε επίπεδο στρώματος, με μείωση κατά μέσο όρο 58% του ποσοστιαίου συντελεστή κύμανσης σε σύγκριση με τις άμεσες μετρήσεις.

REFERENCES

Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. *Journal of the American Statistical Association,* **83**(401), 28-36.

Breidenbach, J., Magnussen, S., Rahlf, J., & Astrup, R. (2018). Unit-level and area-level small area estimation under heteroscedasticity using digital aerial photogrammetry data. *Remote Sensing of Environment,* **212**, 199-211.

Brown, G., Chambers, R., Heady, P., & Heasman, D. (2001). *Evaluation of small area estimation methods—an application to unemployment estimates from the UK LFS.* Paper presented at the Proceedings of Statistics Canada Symposium.

Chandra, H., & Chandra, G. (2020). Small Area Estimation for Total Basal Cover in the State of Maharashtra in India. In G. Chandra, R. Nautiyal, & H. Chandra (Eds.), *Statistical Methods and Applications in Forestry and Environmental Sciences* (pp. 255-266). Singapore: Springer Singapore.

Coulston, J. W., Green, P. C., Radtke, P. J., Prisley, S. P., Brooks, E. B., Thomas, V. A., . . . Burkhart, H. E. (2021). Enhancing the precision of broad-scale forestland removals estimates with small area estimation techniques. *Forestry: An International Journal of Forest Research,* **94**(3), 427-441.

Georgakis, A. (to appear). *Stratification of forest stands as a basis for small area estimations.* Paper presented at the 33rd PanHellenic statistics conference. Statistics in the Economy and Administration, Larissa, Greece.

Georgakis, A., & Stamatellos, G. (2020). Sampling Design Contribution to Small Area Estimation Procedure in Forest Inventories. *Modern Concepts & Developments in Agronomy,* **7**(1) 694-697.

Giordani, P., Ferraro, M. B., & Martella, F. (2020). *An Introduction to Clustering with R*: Springer.

Goerndt, M. E., Monleon, V. J., & Temesgen, H. (2011). A comparison of small-area estimation techniques to estimate selected stand attributes using LiDAR-derived auxiliary variables. *Canadian Journal of Forest Research,* **41**(6), 1189-1201.

Green, P. C., Burkhart, H. E., Coulston, J. W., & Radtke, P. J. (2019). A novel application of small area estimation in loblolly pine forest inventory. *Forestry: An International Journal of Forest Research*.

Harmening, S., Kreutzmann, A.-K., Pannier, S., Salvati, N., & Schmid, T. *A Framework for Producing Small Area Estimates Based on Area-Level Models in R*. Retrieved from

Kaufman, L., & Rousseeuw, P. J. (1990). Partitioning around medoids (program pam). In *Finding groups in data: an introduction to cluster analysis* (Vol. 344, pp. 68-125).

Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., & Tzavidis, N. (2019). The R package emdi for estimating and mapping regionally disaggregated indicators. *Journal of Statistical Software, 91*.

Lahiri, P., & Suntornchost, J. (2015). Variable Selection for Linear Mixed Models with Applications in Small Area Estimation. *Sankhya B,* **77**(2), 312-320.

Li, H., & Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis, 101*(4),

Magnussen, S., Mauro, F., Breidenbach, J., Lanz, A., & Kändler, G. (2017). Area-level analysis of forest inventory variables. *European Journal of Forest Research,* **136**(5), 839-855.

Mandallaz, D. (2008). *Sampling techniques for forest inventories*: CRC Press.

Mauro, F., Molina, I., García-Abril, A., Valbuena, R., & Ayuga-Téllez, E. (2016). Remote sensing estimates and measures of uncertainty for forest variables at different aggregation levels. *Environmetrics,* **27**(4), 225-238.

Mauro, F., Monleon, V. J., Temesgen, H., & Ford, K. R. (2017). Analysis of area level and unit level models for small area estimation in forest inventories assisted with LiDAR auxiliary information. *PloS one,* **12**(12), 14.

Namazi-Rad, M. R., & Steel, D. (2015). What Level of Statistical Model Should We Use in Small Area Estimation? *Australian & New Zealand Journal of Statistics,* **57**(2), 275-298.

Prasad, N. G. N., & Rao, J. N. K. (1990). The Estimation of the Mean Squared Error of Small-Area Estimators. *Journal of the American Statistical Association, 85*(409), 163-171.

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org/index.html

Rao, J. N., & Molina, I. (2015). *Small area estimation*: John Wiley & Sons, Inc.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics,* **20**, 53-65.

Sugasawa, S., & Kubokawa, T. (2020). Small area estimation with mixed models: a review. *Japanese Journal of Statistics and Data Science*.

Temesgen, H., Mauro, F., Hudak, A. T., Frank, B., Monleon, V., Fekety, P., . . . Bryant, T. (2021). Using Fay–Herriot Models and Variable Radius Plot Data to Develop a Stand-Level Inventory and Update a Prior Inventory in the Western Cascades, OR, United States. **4**(157), 17.

Tzavidis, N., Zhang, L.-C., Luna, A., Schmid, T., & Rojas-Perilla, N. (2018). From start to finish: a framework for the production of small area official statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **181**(4), 927-979.

Ver Planck, N. R., Finley, A. O., Kershaw, J. A., Weiskittel, A. R., & Kress, M. C. (2018). Hierarchical Bayesian models for small area estimation of forest variables using LiDAR. *Remote Sensing of Environment*, **204**, 287-295.

Westfall, J. A., Patterson, P. L., & Coulston, J. W. (2011). Post-stratified estimation: within-strata and total sample size recommendations. *Canadian Journal of Forest Research*, **41**(5), 1130-1139.

Wolter, K. M. (2007). *Introduction to variance estimation* (Vol. 53): Springer.

Yoshimori, M., & Lahiri, P. (2014). A new adjusted maximum likelihood method for the Fay–Herriot small area model. *Journal of Multivariate Analysis*, **124**, 281-294.

You, Y., & Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, **32**(1), 97.

Zulkarnain, R., Jayanti, D., & Listianingrum, T. (2020). Improving the quality of disaggregated SDG indicators with cluster information for small area estimates. *Statistical Journal of the IAOS*, **36**, 955-961.

Greek Statistical Institute
Proceedings of the 33rd Panhellenic Statistics Conference (2021), pp.262-280

PUBLIC DEFICIT, DEBT AND STOCK-FLOW ADJUSTMENT (SFA): STATISTICAL AND ECONOMETRIC INVESTIGATION, 1960-2017

Zoe Georganta[1] and Nickolas Logothetis[2]
[1]University of Macedonia – Economic and Social Studies
zoe.georganta@gmail.com and zoe@uom.gr
[2]TQM Hellas S.A, info@tqmhellas.gr

ABSTRACT

Considering the surprisingly long-lasting and unprecedented debt crisis in Greece, the purpose of this paper is to go beyond political and other folklore explanations and suppositions and focus on official data in search of probable irregularities in their statistical behavior during the period of almost six decades, 1960-2017. For this purpose we use annual official data of government debt (DEBT), government deficit (DEF) and of the so called Stock-Flow Adjustment (SFA) which is theoretically a small residual or accounting item "adjusting" differences originated from the use of different data types (stock vs. flows). Within our purpose we first investigate whether our sample period is characterized by structural breaks; second, by using a recursive econometric model we estimate the relative degree of impact that SFA and DEF had on DEBT; third, we make an effort to capture other features of our fiscal variables. Among our findings are the following: (a) Three structural breaks are identified, 1975, 1994 and 2007; (b) only during 2008-2017, the SFA, both had a stronger than DEF positive effect on DEBT, and it also revealed a remarkable deviation from the best statistical practice as defined by Eurostat (the European Statistical Agency); (c) both DEBT and SFA had significant negative effects on GDP only during 2008-2017. Our findings are novel, thought-provoking and useful by signalling to international statistical agencies the importance of building legitimate official statistics. Further research on this issue is going on.

Key Words: Time Series and Structural Change, Government Debt and Deficit, SFA, Recursive (Causal) Models, Greece 1960-2017.

1. INTRODUCTION

The Greek recession started openly in November 2009 as an unexpected shock following the October announcement by Greek authorities of the revision of the nation's projected 2009 public deficit from 6% of GDP to 12.5% of GDP. The news stunned the financial markets thus triggering a non-stop rise of borrowing costs for Greece, successive increases of the government bond spread over the German bund, and a series of ratings' downgrades of Greek sovereign bonds and banks.

Consequently, Greece was forced to withdraw from international bonds markets, while the country was subjected to extraordinary austerity rescue packages and it was placed under the joint custody of the European Commission, the European Central Bank and the IMF, the so called Troika.

The increase in spreads, coupled with exceptional austerity programs, brought about a sharp contraction of GDP, which had a negative effect on debt dynamics, leading to a vicious cycle of rating downgrades, further rises in spreads and further worsening of public debt, and, eventually, a deepening of recession, which evolved into a long depression, comparable only to the US Great Depression of 1929-30, which, worth-noting, had a much shorter duration. The escalation of the Greek financial crisis since November 2009 seems to be the result of a dramatic shift in market expectations for Greece. As Trebesch and Zettelmeyer (2018, Appendix F, p.61) report in their study, the 10 year Greek bond yields from less than 5% in October 2009 escalated to 46% two years later. As a consequence, other euro area countries experienced contagion from Greece (Arghyrou and Kontonikas, 2011).

In view of the above developments, this paper focuses on letting the data give us some grains of interpretation and understanding of what happened beyond political and other, usually folklore suppositions. In other words, having in mind the question "what happened all of a sudden in 2008-09", this paper follows a directly empirical approach searching the official data for an answer. Thus, our purpose is to examine whether the Greek depression must be attributed to official fiscal data deficiencies. In this sense, this paper is at least thought-provoking.

Within the set framework, we study the evolution of the relationships between government debt, DEBT, government deficit, DEF1, and the residual accounting item SFA (Stock-Flow Adjustment), or DDA (Debt-Deficit Adjustment), which is the observed difference between the two theoretically equivalent concepts of government deficit in practice. As of its definition as a small accounting item, the SFA should have on DEBT a much weaker effect than the DEF1. This is econometrically examined. We also calculate the total (direct and indirect) estimated effects of DEBT and SFA on Greece's economic activity as expressed by her GDP. For our purpose, we use official annual data for the period 1960-2017.

The statistical and econometric examination of our time series for the entire sample period, 1960-2017 showed three structural break points, 1975, 1994 and 2007 leading to the consideration of the following four sub-periods: 1960-1975, 1976-1994, 1995-2007 and 2008-2017. Our main findings include the following: (1) in sharp contrast to the first three sub-periods, during 2008-2017 the SFA was found, first, to have a much stronger positive effect on DEBT than the DEF1 (a clear symptom of the SFA illegitimacy), and second, both, the SFA and DEBT had a significant negative effect on GDP. (2) The SFA as % of GDP exhibited extreme degrees of variation ranging from -0.6% to +13% and -42.5%, indicating a remarkable deviation from the best statistical practice of the acceptable limits, which have been described by Eurostat as below ±2% (see for example Eurostat, 2012).

The rest of the paper is structured as follows: the next section presents the definitional relationship between DEBT, DEF1 and SFA. It also describes our data

and discusses their structure. Section 3 presents the econometric model used and our empirical estimates which are also discussed. The last section presents our concluding remarks and discusses further research.

2. DEFINITIONS AND DATA DESCRIPTION

2.1 Basic Definitions

The European System of Accounts (ESA) gives a basic definition of government or public debt as the total sum of budget liabilities. This definition is wider than the Maastricht definition of debt. A concise description of the various accounting definitions of public debt and deficit is given by the ECB(2011) and Irwin(2012, 2015) among others. Further, according to theory, the relationship between debt and deficit is given by the equation:

$$DEBT_t - DEBT_{t-1} = BB_t \quad (1)$$

Equation (1) means that debt in time t equals the inter-temporally accumulated deficits. In addition, equation (1) implies the twin, theoretically equivalent definitions of government deficit: first, as the difference between the outstanding government debt ($DEBT_t$) and the government debt of the previous period ($DEBT_{t-1}$); second, as the difference between government expenditures and government revenue or Budget Balance (BB_t). Apparently, DEBT is a stock variable and BB (Budget Balance, or expenditures EXPEND – revenues REV) is a flow one. The BB as deficit determines the borrowing needs of the country. In practice equation (1) becomes the following equation:

$$DEBT_t - DEBT_{t-1} = BB_t + K_t \quad \text{(a)}$$
$$or \quad (2)$$
$$DEBT_t - DEBT_{t-1} = DEF_t + SFA_t \quad \text{(b)}$$

K in equation (2a) is evidently a residual variable correcting or adjusting small errors of data happening in practice because of use of different data sources and types (stock and flow type of data). K has been called "Stock-Flow Adjustment" (SFA) or "Debt Deficit Adjustment" (DDA). So, equation (2b) can be rewritten as follows:

$$SFA_t = (DEBT_t - DEBT_{t-1}) - DEF_t \quad (3)$$

If we denote $DEBT_t$ - $DEBT_{t-1} = DEF2_t$ and $DEF_t = DEF1_t$, equation (3) can be rewritten as follows:

$$SFA_t = DEF2_t - DEF1_t \quad (3a)$$

Equation (3a) is rather a "fiscal identity" giving the two definitions of BB_t in practice.

According to the SFA Reviews regarding EU member-states' (for example, see Eurostat 2012, pp.1-2), Eurostat reports the following: "A positive SFA means that the government debt increases more than the annual deficit (or decreases less than implied by the surplus). On the contrary, a negative SFA means that the government

debt increases less than the annual deficit (or decreases faster than implied by the surplus)." Also, "The importance of the SFA has been emphasized many times, as an efficient statistical monitoring of fiscal performance requires understanding the coherence between the two key fiscal indicators, government deficit and debt. It has been argued that since great attention is paid to the deficit under the EU multilateral fiscal surveillance (EDP *Excessive Deficit Procedures* and Stability and Growth Pact), governments may have an incentive in underreporting their deficits by reporting transactions under the SFA *instead under the debt and/or deficit*. SFAs generally have legitimate explanations … however it is important that they are closely monitored because they can highlight data quality problems." (Our own clarifications in italics)

Thus, the SFA may be an indication of either bad data quality or illegitimate transfers of economic exchanges from the budget deficit to the SFA so that the various EU member-states could present low deficits. According to its "legitimate" definition, SFA should include only minor statistical errors and minor adjustments due to the use of different data sources. Indeed, a legitimate SFA has been prescribed by Eurostat as such only if it lies below the limits of ±2% of GDP (see Eurostat 2012, p.2). We have also to emphasize that the SFA is not subjected to ESA or any other regulation regarding its composition as are the debt and deficit, so facilitating the above illegitimate transfers.

In the light of the above flexibility regarding the items included in the SFA, a plethora of research findings have been reported in literature showing that the SFA has been used as a policy tool to manage debt and deficit so that the various EU countries could go safe along the requirements of the Maastricht criteria (upper limit of 60% of GDP for the debt and 3% for the deficit – e.g. EDP Glossary 2021). It has been found that this has especially happened by almost all EU countries during their preparations to join the euro-area (among others, see Koen and Noord, 2005; Hagen and Wolff, 2006; Weber, 2012; Alt et al., 2012, 2014).

2.2 Data and Structural Change

Our data are annual and cover the period 1960-2017. Our sources are the State Budgets, the National Accounts of Greece, the OECD and the Eurostat Data Base AMECO. In order to obtain continuously compatible, accurate, objective and reliable time series we have received valuable advice from the Statisticians at the Ministry of Finance, as well as from colleagues at the Hellenic Statistics Authority (ELSTAT) who are experts in debt and deficit measurement according to ESA. Our data refer to Central Government as it is defined in ESA. Debt and deficit data, where appropriate, are measured on EDP or Maastricht basis which is different from their ESA measurement, the difference mainly centred on the exchange items included and on valuation issues (see Manual on Government Deficit and Debt, ESA95 and ESA2010).

Regarding GDP data in current market prices, we have used the OECD (2018) series which is the only one consistently compatible over-time. The OECD series is based on the system of the UN National Accounts (SNA93 and SNA2008). ESA95

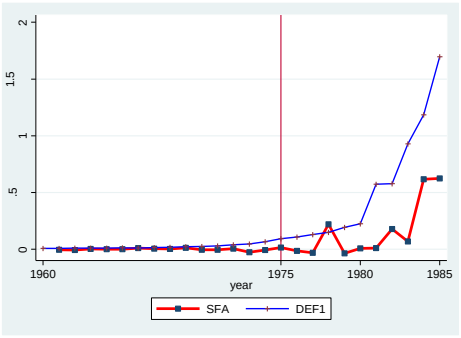
and ESA2010 are "broadly consistent" with SNA93 and SNA2008 respectively. It is noted that the differences between ESA95 and ESA10 are very small (see Eurostat, 2014) especially for Greece. In particular, the most important difference between ESA95 and ESA2010 is R&D expenditures which are considered as fixed capital formation in ESA2010, but not in ESA95. For Greece R&D expenditures have been very low during our sample period. The summary statistics of our variables are presented in ***Table 1***:

Table 1. *Summary Statistics*

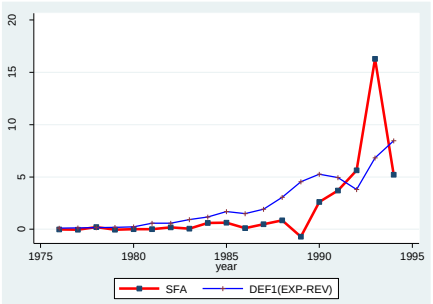
Variable	Obs	Mean	Std.Dev	Min	Max	CV
GDP in bill€	58	76.894	83.198	0.364	233.198	108.2
REV in bill€	58	19.826	22.458	0.046	60.352	113.3
EXPEND in bill€	58	24.595	27.629	0.053	86.355	112.3
DEBT in bill€	58	98.351	124.985	0.029	377.904	127.1
DEBT/GDP in %	58	6.001	9.109	-17.539	33.760	151.8
DEF1=EXPEND-REV	57	4.769	6.391	0.007	30.872	134.0
DEF2=$DEBT_t$ - $DEBT_{t-1}$	58	5.766	14.325	-66.518	60.763	248.4
SFA=DEF2-DEF1	57	0.914	12.370	-82.206	29.891	1353.4
SFA/GDP in %	57	1.456	7.251	-42.517	23.327	498.0
DEF1/GDP in %	58	5.307	3.298	1.467	13.360	62.1

According to ***Table 1***, our series exhibit a wide range of values. This is true mainly for SFA. In the ***DP1*** below we present four diagrams to visually get an idea of whether our variables exhibit any structural breaks.

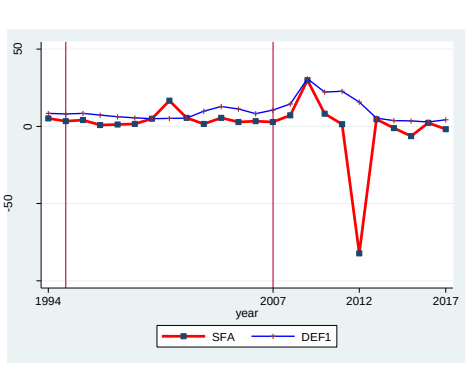
Diagramatic Presentations 1(DP1). *Time Paths of the Variables of Interest*



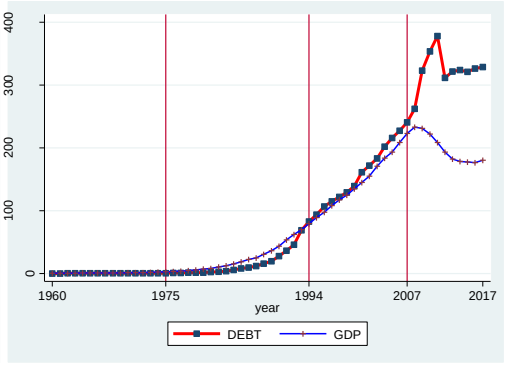
(a)



(b)



(c)



(d)

According to the four diagrams presented in ***DP1*** we visually discern three structural breaks, namely 1975, 1994 and 2007. The next step is to test econometrically whether we have to apply different regression models for the four different data sets indicated in ***DP1*** or it is preferable to consider one model for the entire sample period. In literature this problem is usually treated by the Chow test, or the dummy-variable approach. We adopt the latter because it is more informative in the sense that it will give us information about differences of both the position and the slope of the regression equation for our four sets of data (Gujarati, 2004, 306-310). Thus, we generate three dummies D6075, D6094 and D6007 differentiating their values as follows:

Table 2. *Values of Dummy Variables*

Dummy	**1960-1975**	**1976-1994**	**1995-2007**	**2008-2017**
D6075	1	0	0	0
D6094	1	1	0	0
D6007	1	1	1	0

Under the Dummy procedure we will estimate the following basic equation:

$$DepVar = \alpha + \beta_m D_m + \gamma_k IndVar_k + \delta_{mk}(D_m \times IndVar_k) + \varepsilon \quad (4)$$

D denotes the dummy variable, m is the number of dummy variables and k is the number of independent variables. If any of the coefficients β_m is statistically significant, then there is a change in the position of the empirical regression equation at the point indicated by the corresponding dummy variable. Similarly, if any of the coefficients δ_{mk} is statistically significant, then there is a change in the slope of the empirical regression equation at the point indicated by the relevant dummy.

Within the purpose of this paper, we first test the regression regarding the relative strength of the effects of SFA and DEF1 on DEBT, and, second, we test the regression involving the effect of DEBT, DEF1 and SFA on GDP. The estimated empirical formulation of equation (4) gives the following regression equations:

$$\log DEBT_t = \underset{t-value:}{5.764} - \underset{(22.0)}{3.173}D6075 - \underset{(-12.0)}{4.263}D6094 - \underset{(-7.8)}{1.038}D6007 + \underset{(-1.8)}{0.002}DEF1_t + \underset{(0.1)}{0.001}SFA_t +$$
$$+ \underset{(0.1)}{24.798}(D6075 \times DEF1)_t - \underset{(4.5)}{3.911}(D6075 \times SFA)_t + \underset{(-0.3)}{0.552}(D6094 \times DEF1)_t + \qquad \bar{R}^2 = 0.9749 \quad (5)$$
$$+ \underset{(5.9)}{0.0002}(D6094 \times SFA)_t + \underset{(0.0)}{0.051}(D6007 \times DEF1)_t - \underset{(0.9)}{0.025}(D6007 \times SFA)_t \quad (-0.9)$$

$$\log DEBT_t = \underset{t-value:}{5.795} - \underset{(14.7)}{3.623}D6075 - \underset{(-9.6)}{4.883}D6007 + \underset{(-10.9)}{0.00003}DEF1_t + \underset{(0.0)}{0.004}SFA_t +$$
$$+ \underset{(0.5)}{24.977}(D6075 \times DEF1)_t - \underset{(3.0)}{3.914}(D6075 \times SFA)_t + \qquad \bar{R}^2 = 0.9749 \quad (6)$$
$$+ \underset{(-0.2)}{0.443}(D6007 \times DEF1)_t + \underset{(9.4)}{0.061}(D6007 \times SFA)_t \quad (1.7)$$

$$\log GDP = \underset{t-value:}{5.627} - \underset{(5.0)}{2.812}D6075 - \underset{(-14.0)}{2.260}D6094 - \underset{(-6.3)}{1.455}D6007 - \underset{(-1.2)}{0.001}DEBT_t + \underset{(-0.4)}{0.011}DEF1_t + \underset{(1.0)}{0.0007}SFA_t + \underset{(0.2)}{7.425}(D6075 \times DEBT)_t -$$
$$- \underset{(2.5)}{15.022}(D6075 \times DEF1)_t - \underset{(-1.0)}{4.594}(D6075 \times SFA)_t - \underset{(-0.6)}{0.008}(D6094 \times DEBT)_t + \underset{(-0.6)}{0.436}(D6094 \times DEF1)_t - \quad (7)$$
$$- \underset{(4.4)}{0.005}(D6094 \times SFA)_t + \underset{(-0.1)}{0.008}(D6007 \times DEDT)_t - \underset{(2.0)}{0.036}(D6007 \times DEF1)_t - \underset{(-0.8)}{0.028}(D6007 \times SFA)_t \quad \bar{R}^2 = 0.9808 \quad (-1.3)$$

$$GDP = \underset{t-value:}{272.543} - \underset{(16.2)}{6.076}D6075 - \underset{(-2.0)}{5.468}D6094 - \underset{(-1.0)}{260.750}D6007 - \underset{(-14.8)}{0.309}DEBT_t + \underset{(-5.9)}{2.130}DEF1_t + \underset{(13.9)}{0.154}SFA_t - \underset{(3.0)}{6.076}(D6075 \times DEBT)_t +$$
$$+ \underset{(-2.0)}{6.183}(D6075 \times DEF1)_t - \underset{(0.03)}{5.131}(D6075 \times SFA)_t - \underset{(-0.04)}{0.039}(D6094 \times DEBT)_t + \underset{(-0.2)}{3.666}(D6094 \times DEF1)_t - \quad (8)$$
$$- \underset{(2.5)}{0.670}(D6094 \times SFA)_t + \underset{(-1.3)}{1.218}(D6007 \times DEDT)_t - \underset{(19.8)}{2.853}(D6007 \times DEF1)_t - \underset{(-4.4)}{0.892}(D6007 \times SFA)_t \quad \bar{R}^2 = 0.9971 \quad (-2.8)$$

Some Remarks: DEF1 and SFA are linearly independent over the whole sample period (1960-2017) with a sample correlation coefficient r(DEF1,SFA)=0.0694 and p-value=0.6080. The same holds for the variables DEBT and SFA: r(DEBT, SFA)= -0.0647 with p-value=0.6326. However, the correlation coefficient between DEBT and DEF1 is statistically significant, r(DEBT, DEF1)=0.7450 with p-value=0.0, but the relationship is of medium strength. Taking into account, <u>first</u>, that equations (5) to (8) are simply instrumental in the sense that we will not use them for either policy or forecasting purposes and, <u>second</u>, the highly significant values of both, the adjusted R^2 and the coefficients of DEBT, SFA and DEF1 (see equ. 8), we have decided to ignore the rather low degree of multicollinearity between DEBT and DEF1.

Equations (5) to (8) show that there are indeed three time points of structural change, 1975, 1994 and 2007, regarding either the position of the regression lines (statistical significance of the coefficients of dummies), or the slope of the regression lines (statistical significance of the coefficient of the product: D_m x $IndVar_k$), or both, position and slope. This finding means that the estimation of one regression equation for the entire sample period will give misleading results. This is also supported by the

following ***Table 2***, in which we can see the large differences in the mean and standard deviation of our variables during the different time periods.

Table 2. *Mean and Standard Deviation of our Main Variables by Time-period*

period	obs	SFA		DEF1		DEBT		GDP	
		mean	std	mean	std	mean	std	mean	std
1960-75	16	0.00007	0.0099	0.0262	0.0241	0.1568	0.1212	1.0327	0.6549
1976-94	19	1.8872	3.9396	2.4287	2.5291	18.1011	24.0857	26.8034	24.0663
1995-07	13	4.1615	4.0735	8.0127	2.5414	162.0917	48.9097	150.0839	43.2563
2008-17	10	-3.7869	29.2534	12.5865	10.1298	325.0747	29.3828	198.2975	23.3005
1960-17	**58**	**0.9140**	**12.3700**	**4.7690**	**6.3910**	**98.3510**	**124.985**	**76.8940**	**83.1980**

3. THE ECONOMETRIC MODEL AND ITS EMPIRICAL ESTIMATES

3.1 The Econometric Model

We use the following recursive or causal model:

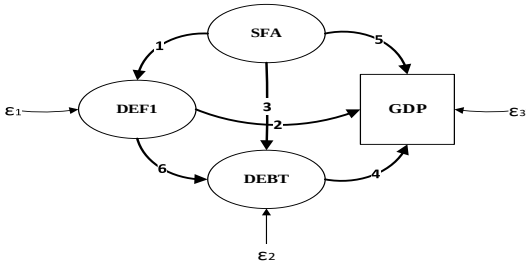
$$\begin{aligned} DEF1_t &= \alpha_1 + \gamma_1 SFA_t + \varepsilon_{1t} \quad \text{(a)} \\ DEBT_t &= \alpha_2 + \beta_2 DEF1_t + \gamma_2 SFA_t + \varepsilon_{2t} \quad \text{(b)} \\ GDP_t &= \alpha_3 + \beta_3 DEF1_t + \beta_3' DEBT_t + \gamma_3 SFA_t + \varepsilon_{3t} \quad \text{(c)} \end{aligned} \tag{9}$$

Variables DEF1, DEBT and GDP are endogenous and SFA is treated as exogenously determined. For the error terms the following equalities hold: $cov(\varepsilon_{1t}, \varepsilon_{2t})=cov(\varepsilon_{1t}, \varepsilon_{3t})=cov(\varepsilon_{2t}, \varepsilon_{3t})=0$. Model (9) has been specified within the following macroeconomic theory: Equation 9(a) is the econometric expression of the fiscal identity (3a) described in section 2.1: given the value of the sum (DEF1+SFA), a change in the value of SFA causes DEF1 to change. Thus, SFA determines DEF1 but it is not determined by it, as this is evident from the discussion in section 2.1, where we have reported Eurostat's views, as well as independent research results showing that the SFA has been used as a policy tool by EU member states, Greece included, in order to present lower deficits. So, SFA is predetermined by fiscal policy considerations.

Equation 9(b) expresses the widely known thesis that deficits contribute to an increase in debt levels, while surpluses reduce them. In addition, the SFA's predetermined fluctuations affect government debt as it is also explained in section 2.1. Equation 9(c) is the synopsis of one of the basic macroeconomic tenets postulating that management of government expenditures and revenues may encourage or discourage economic activity (GDP). For example, increasing government spending and/or decreasing tax revenue tends to encourage economic activity by increasing individuals' disposable income, but this sort of expansionary fiscal policy may lead to sizable deficits and debts which may have a negative impact on economic activity. It is evident that we can apply OLS for each equation of the causal system (9) (Gujarati, 2004, 764-766).

The following ***Schema 1*** presents the causal structure of model (9) in terms of a path diagram (see Carey, 1998). The estimated paths (1 to 6) are shown in ***Table 8*** of section 3.2.

***Schema 1**. Direct and Indirect Effects between SFA, DEBT, DEF1, GDP*



3.2 Empirical Estimates

Before presenting our estimates, we note the following:

(1) Issues of best fit may require our empirical equations being expressed in log-linear form and also we may need to include additional exogenous variables like the time trend and/or other variables like government expenditures and revenue.

(2) Some of our models (M) may have better fit if we make use of the partial adjustment process of Marc Nerlove (Gujarati, 2004, 673-675). We note that the partial adjustment models can be estimated with OLS (Gujarati, 2004, 677-678).

(3) The autocorrelation problem is detected by using the general test of Breusch-Godfrey, known also as LM test, denoted by B-G LM (Lagrange Multiplier) test, which is also appropriate for the partial adjustment model and it seems it is statistically more powerful than the h test in our small-sample cases (Gujarati, 2004, 472-474, 681). If the error term follows the autoregressive scheme of order one, the B-G test is known as the Durbin's M test. Because the number of observations in each of the four sub-periods is small, we use the B-G test for small samples and we have also considered lags of second order.

(4) The sample size of the four identified data sets is 16, 19, 13 and 10 observations corresponding to each sub-period in chronological order. We note the following: First, in macro-econometrics we usually have much smaller sample sizes than in micro, financial or sociological fields. OLS Monte Carlo simulations with 6, 10, 20 and 500 sample sizes and six independent variables have shown that the gains are minute (see MPRA, 2019, Table 1, p.8: "This tends to indicate that the affirmation of Speed (1994) regarding the relationships found in small (10 oservations) sample size tend to remain as the size of the sample increases"). We note that in our cases we have a small number of independent variables, 2 and 3.

(5) In general, the size of the regression coefficients depends partially on the mean and variance of the independent variables. For comparison purposes we have estimated the standardized beta coefficients (Gujarati, 2004, 173-175, 215) which are expressed in standard deviation units. We note that the standardized coefficients are dependent upon the variations in independent and the dependent variables observed in the particular data set. Thus, we expect that samples with great variations in one of the explanatory variables will exhibit large beta coefficients for that variable. We also report the corresponding short-run elasticities at mean values (η) as well. However this measure (η) is dependent on the mean values of the respective variables.

Our estimates are presented in the following four tables. In all four Tables we use the following notation: r is the sample Pearson's linear correlation coefficient. The parenthesis next to beta coefficients includes its standard error. The last line of the Tables presents the Pearson's r and the Variance Inflation Factors (VIF), both for checking the existence of a probable multicollinearity problem.

Table 3. *OLS Estimates of System (9): 1960-1975, 16 obs*

Regressant→	$DEF1_t$	$DEBT_t$	$logGDP_t$	$logGDP_t$
Regressors ↓	M1	M2	M3.1	M3.2
$DEF1_t$ Prob>|t| Short-run elasticity (η) beta (std)		2.8993 0.000 0.472 0.640(0.020)	4.9272 0.000 0.130 0.198(0.032)	
$DEBT_t$ Prob>|t| η beta (std)				4.6623 0.000 0.863 1.056(0.045)
SFA_t Prob>|t| η beta	0.3875 0.004 0.001 0.157(0.044)			
SFA_{t-2} Prob>|t| η beta				8.5456 0.010 -0.005 0.158(0.043)
SFA_{t-3} Prob>|t| η beta		1.0144 0.001 0.008 0.050(0.01)		
TIME Prob>|t| beta	-0.0078 0.000 -1.442(0.223)	0.012 0.000 0.376(0.029)	0.1037 0.000 0.824(0.032)	
$TIME^2$ Prob>|t| beta	0.0007 0.000 2.365(0.212)			
CONS (Prob>|t|)	0.0284 (0.000)	-0.0238 (0.001)	-1.1506 (0.000)	-0.8223(0.000)
B-G LM test small lag(1) (Prob>F) lag(2) (Prob>F)	 0.092 (0.7673) 0.435 (0.6604)	 2.612 (0.1501) 1.307 (0.3380)	 0.173 (0.6852) 1.427 (0.2811)	 0.093(0.7671) 3.088(0.1015)
Adj R^2	0.9746	0.9990	0.9957	0.9766
$r(DEBT_t, DEF1_t)$=0.9702(0.0000), $r(SFA_{t-3}, DEF1_t)$=0.0769(0.8123), $r(SFA_{t-2}, DEBT_t)$= -0.4782(0.0984) VIF M1: 1.04 (SFA_t), VIF M2: 1.18 (SFA_{t-3}), 5.80 ($DEF1_t$), VIF M3.1: 3.64 ($DEF1_t$) VIF M3.2: 1.30 (SFA_{t-2}), 1.30 ($DEBT_t$)				

Table 4. *OLS Estimates of System (9): 1976-1994, 19 obs*

Regressant→ Regressors↓	$DEF1_t$ M1	$logDEBT_t$ M2	GDP_t M3
$DEF1_t$		0.0787	5.0386
Prob>|t|		0.004	0.001
η		0.191	0.457
beta		0.124(0.037)	0.529(0.132)
$logDEBT_t$			5.1750
Prob>|t|			0.013
η			0.193
beta			0.345(0.122)
SFA_t		0.0172	1.0437
Prob>|t|		0.072	0.028
η		0.032	0.074
beta		0.042(0.022)	0.171(0.070)
SFA_{t_1}	0.2133		
Prob>|t|	0.004		
η	0.142		
beta	0.327(0.097)		
TIME	0.3228		
Prob>|t|	0.000		
Beta	0.718(0.097)		
logTIME		6.1862	
Prob>|t|		0.000	
Beta		0.861(0.033)	
CONS	-6.3091	-18.3411	2.7930
Prob>|t|	0.000	0.000	0.167
B-G LM test small			
lag(1)	2.475	2.813	0.202
Prob>F	0.1365	0.1157	0.6599
lag(2)	1.640	1.412	1.101
Prob>F	0.2291	0.2787	0.3553
Adj R^2	0.8907	0.9956	0.9536
Prob>F			

$r(SFA_t, DEF1_t)$=0.6934(0.0010), $r(DEF1_t, logDEBT_t)$=0.9098(0.0),
$r(SFA_t, logDEBT_t)$=0.6311(0.0038)
VIF M1: $1.55(SFA_{t-1})$
VIF M2:$5.58(DEF1_t)$, $1.94(SFA_t)$
VIF M3: $6.72(DEF1_t)$, $5.8(logDEBT_t)$, $1.93(SFA_t)$

Table 5. *OLS Estimates of System (9): 1995-2007, 13 obs*

Regressant→ Regressors↓	$logDEF1_t$ M1	$logDEBT_t$ M2	GDP_t M3.1	$logGDP_t$ M3.2	GDP_t M3.3
$logDEF1_t$ Prob>|t| η beta		0.0513 0.030 0.051 0.052(0.020)	9.0249 0.020 0.0002 0.066(0.024)		
$logDEBT_t$ Prob>|t| η beta				0.9512 0.000 0.951 0.997(0.024)	
SFA_t Prob>|t| η beta		0.0034 0.038 0.014 0.045(0.019)			
SFA_{t_1} Prob>|t| η beta					-0.2160 0.038 -0.006 -0.020(0.008)
SFA_{t_3} Prob>|t| η beta	0.0433 0.003 0.244 0.706(0.185)				
TIME Prob>|t| beta			10.7406 0.000 0.967(0.024)		-11.536 0.001 -1.039
logTIME Prob>|t| beta	1.7428 0.019 0.514(0.185)	3.2219 0.000 0.971(0.019)			
$TIME^2$ Prob>|t| beta					0.2693 0.000 2.038(0.211)
CONS Prob>|t|	-4.7127 0.073	-7.1034 0.000	-319.3859 0.000	0.1730 0.161	156.7538 0.011
B-G LM test small lag(1) Prob>F lag(2) Prob>F	 0.072 0.7943 0.753 0.5016	 0.398 0.5457 3.058 0.1110	 4.442 0.643 2.535 0.1404	 0.815 0.3879 1.892 0.2061	 0.611 0.4569 3.172 0.1046
Adj R^2 Prob>F	0.5973 0.0043	0.9966	0.9945	0.9931	0.9992

r (SFA_t, $logDEF1_t$)=-0.3326(0.2669), r(SFA_t, $logDEBT_t$)=0.1156(0.7067)

VIF M1: 1.02 (SFA_{t-3})

VIF M2: 1.46 ($logDEF1_t$), 1.21 (SFA_t)

VIF M3.1: 1.24 ($logDEF1_t$)

VIF M3.3:1.04 (SFA_{t-1})

Table 6. *OLS Estimates of System (9): 2008-2017, 10 obs*

Regressant→ Regressors↓	$logDEF1_t$ M1	$logDEBT_t$ M2	$logGDP_t$ M3
$logDEF1_t$		0.0352	0.1156
Prob>|t|		0.000	0.000
η		0.035	0.116
beta		0.348(0.043)	0.923(0.072)
$logDEBT_t$			-0.3831
Prob>|t|			0.005
η			-0.383
beta			-0.310(0.072)
$logDEBT_{t\_1}$		0.9357	
Prob>|t|		0.000	
beta		1.325(0.056)	
SFA_t	-0.0067	0.0030	0.0012
Prob>|t|	0.012	0.000	0.005
η	-0.025	-0.012	-0.005
beta	-0.213	0.954(0.055)	0.305(0.072)
$EXPEND_t$	0.0730		
Prob>|t|	0.000		
η	4.922		
beta	1.014		
CONS	-2.7656	0.3354	7.2507
Prob>|t|	0.000	0.193	0.000
B-G LM small			
lag(1)	0.157	2.902	1.031
Prob>F	0.7058	0.1492	0.3565
lag(2)	1.183	2.824	0.522
Prob>F	0.3797	0.1719	0.6291
Adj R^2	0.9634	0.9845	0.9545
$r(SFAt, EXPEND_t)$=0.2345(0.5143), $r(SFA_t, logDEF1_t)$=0.0253(0.9448), $r(logDEF1_t, logDEBT_t)$=0.1149(0.7520), $r(SFA_t, logDEBT_t)$=0.1062(0.7702) VIF M1: 1.06 (SFA_t), 1.06 ($EXPEND_t$) VIF M2:1.82 ($logDEBT_{t-1}$), 1.75 (SFA_t), 1.06 ($logDEF1_t$) VIF M3: 1.02($DEF1_t$), 1.01($DEBT_t$), 1.01(SFA_t)			

The next ***Table 7*** presents in a compact way the estimated beta coefficients of the variables SFA and DEF1 as reported in column M2 of the previous ***Tables 3-6***. It also presents the elasticities (η) of DEBT with respect to DEF1 and SFA.

Table 7. *Relative Strength and % Size of the effects of SFA and DEF1 on DEBT Beta Coefficients and Elasticities of DEBT with Respect to SFA and DEF1*

Model M2 from Tables (T.)	SFA		DEF1		beta:SFA/DEF1 5:(1/3)	η:SFA/DEF1 6:(2/4)
	beta (1)	η (2)	beta (3)	η (4)		
1960-1975:**T.3**	0.050	0.008	0.640	0.472	0.078	0.017
1976-1994:**T.4**	0.042	0.032	0.124	0.191	0.339	0.168
1995-2007:**T.5**	0.045	0.014	0.052	0.051	0.865	0.275
2008-2017:**T.6**	0.954	-0.012	0.348	0.035	2.741	-0.343

Table 8 presents the direct and indirect effects of SFA on GDP (see ***Schema 1***).

Table 8. *Estimated Strength of Direct, Indirect and Total Effects by Time-period*

Degree of strength of effects based on beta coefficients as estimated in ***Tables 3*** to ***6***	1960-1975	1976-1994	1995-2007	2008-2017
1. Direct Effect of SFA on DEF1	0.157	0.327	0.706	-0.213
2. Direct Effect of DEF1 on GDP	0.198	0.529	0.066	0.923
Indirect Effect of SFA on GDP through DEF1	**0.031**	**0.173**	**0.047**	**-0.197**
3. Direct Effect of SFA on DEBT	0.050	0.042	0.045	0.954
4. Direct Effect of DEBT on GDP	1.056	0.345	0.997	-0.310
Indirect Effect of SFA on GDP through DEBT	**0.053**	**0.015**	**0.045**	**-0.296**
Total Indirect Effect of SFA on GDP	**0.084**	**0.188**	**0.092**	**-0.493**
5. Direct Effect of SFA on GDP	0.158	0.171	-0.020	0.305
Total Effect of SFA on GDP	**0.242**	**0.359**	**0.072**	**-0.188**
6. Direct Effect of DEF1 on DEBT	0.640	0.124	0.052	0.348

3.3 Discussion on Findings

As can easily be seen, all estimated models (columns M) in ***Tables 3*** to ***6*** have a very good fit, they do not suffer from the problem of autocorrelation as it is shown by the Breusch-Godfrey LM test for small samples, and the regression coefficients are highly significant in their overwhelming majority. M1, M2 and M3 columns in ***Tables 3*** to ***6*** correspond to equations 9(a), 9(b) and 9(c) of the recursive system (9). We note that the three variables, DEF1, DEBT and SFA could not be accurately estimated together in one regression equation for the sub-periods 1960-1975 and 1995-2007 (***Tables 3*** and ***5*** respectively); so, column M3 of ***Table 3*** is split in two columns (M3.1 and M3.2) and in ***Table 5,*** M3 is split into 3 columns (M3.1, M3.2 and M3.3). Regarding multicollinearity, the values of all VIFs are very low in all equations of all four Tables. Although some r show the existence of linear relationship between some of our variables, we ignore the probable multicollinearity issue in such cases on the basis of the low VIFs and the high significance of both, the corresponding regression coefficients and the adjusted R^2. We categorize the discussion of our findings in the following four groupings:

a) Comparative Effects of SFA and DEF1 on DEBT

In sharp contrast to the sub-periods before 2008, during the period 2008-2017 the SFA has increased DEBT to a much stronger degree than government deficit, DEF1. In particular and ceteris paribus, the SFA effect was 2.741(=0.954/0.348) times stronger during 2008-2017 (***Table 7,*** column 5). According to its definition SFA is a corrective residual item created because DEF1 and DEF2 are not equal in practice (see equation 3 and 3a, as well as the discussion in section 2.1). Therefore, a legitimate SFA is by definition small, thus expecting it to have a weaker than DEF1 effect on DEBT. It follows that if SFA has a stronger than DEF1 effect on DEBT, then SFA is illegitimate in the sense of including governmental economic exchanges belonging in fact to DEF1. Consequently, our finding shows that SFA is illegitimately constructed for the time period after 2007. This result is also corroborated by previous

empirical studies for other countries for which it was found that the SFA increased DEBT more than the deficit (DEF1). A well founded econometric analysis is presented in Campos et al. (2006), who, among other researchers, have found that in contrast to its definition, SFA is not a small residual but a very important factor determining the accumulation of public debt, especially during bank crises.

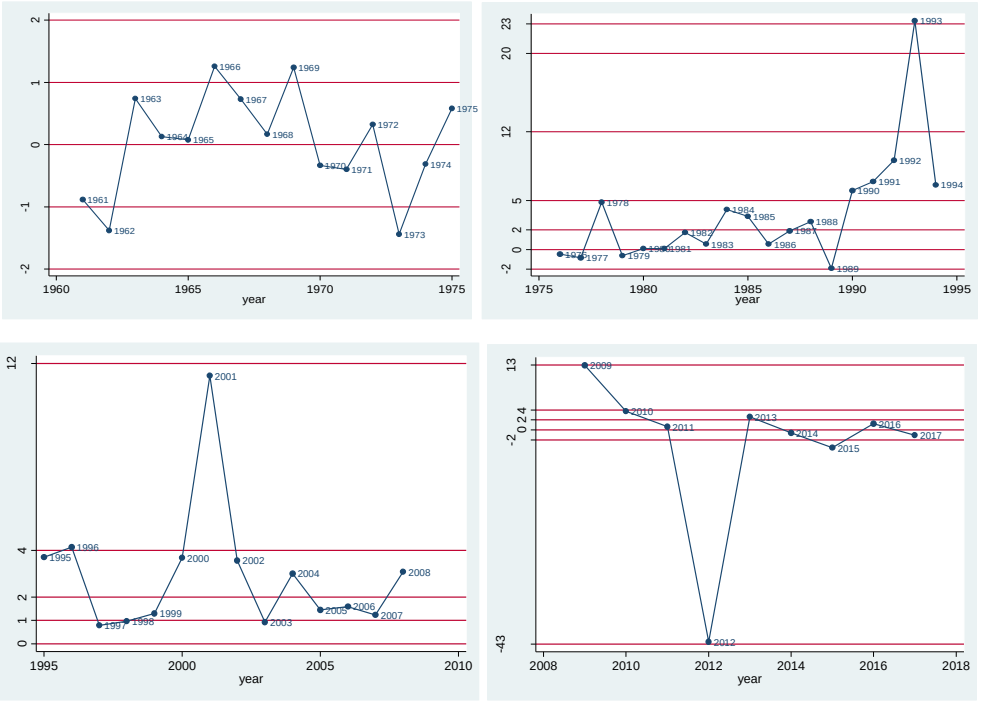
On the other hand, the elasticities of DEBT with respect to SFA and DEF1, at the mean values, show that the SFA causes lower than the DEF1 % changes in DEBT during the entire sample period, but they get higher in absolute value in comparison with the elasticities of DEBT with respect to DEF1 as column 6 of ***Table 7*** shows. Of course the elasticities at the mean values are heavily influenced by the size of the mean and, as we can see in ***Table 2***, the mean values of SFA are much lower than the mean values of DEF1. Quite the opposite happens with the standard deviation (SD). Of course the SD is a statistical measure reflecting the reliability of our official data. As we can also see, the negative elasticity of DEBT with respect to SFA during 2008-2017 is attributed not to the estimate of the regression coefficient, but to a negative mean value of SFA during this period. The main conclusion from ***Table 7*** is that during 2008-2017 the SFA has an almost three times stronger than the DEF1 effect on DEBT, and this is a fact signalling the illegitimacy of SFA during 2008-2017.

b) Effects of SFA on DEBT, DEF1 and GDP

Table 8 has been constructed according to ***Schema 1*** which shows the direct and indirect effects of SFA on GDP. The calculations are based on the estimated beta coefficients which are presented in ***Tables 3*** to ***6***. We can see that in sharp contrast to the period before 2008, the total indirect effect of SFA on GDP is negative and much stronger during 2008-2017 compared with the previous period. In particular, it is 2.6 (0.493/0.188) to 5.9 (0.493/0.084) times stronger in 2008-2017 in contrast with the previous periods. But also, the total effect (including the direct effect) of SFA on GDP is negative, showing that SFA is responsible for Greece's depression during more than a decade now. ***Table 8*** also shows that SFA reduces DEF1 (case no. 1) and increases DEBT (case no. 3) during 2008-2017 to a much stronger degree compared with the previous years. So, SFA has proved to be not a small residual, but a very important factor determining the accumulation of the Greek public debt after 2007, which is a result in agreement with Campos et al. (2006), among other researchers.

c) The Unexpected Statistical and Econometric Behaviour of the SFA in Greece. As we have already mentioned, the SFA has been defined as a variable corrective for the stock-flow (debt-deficit) data which come from different sources, and may also include other minor errors, thus essentially being a kind of residual. We remind the reader that the SFA is legitimate only if it lies below, or at least within the limits of ±2% of GDP (see reference in section 1 "***INTRODUCTION***"). We can see the time path of the SFA item as a percentage of GDP during our sample period in the following Diagrams:

Diagrammatic Presentations 2 (DP2). *Realized Value Limits of SFA/GDP*



As the ***DP2*** shows, during 1960-1975 the SFA was absolutely legitimate. During the next 33 years, 1976-2008, the SFA as a percentage of GDP exhibits six positive peaks over the value 6% of GDP, including 1993 (23.3%) and 2001(11.5%). After 2008, and only during nine years, we observe the SFA as % of GDP being wildly fluctuating, in the range of -0.6% (2014) to 12.9% (2009) and even to -42.5% (2012), year of the infamous PSI (Private Sector Involvement) or "haircut", which reduced Greek government debt only by 11.9 €bill (see Bruegel, 2012). As we have discussed in section 2.1, according to Eurostat, "A positive SFA means that the government debt increases more than the annual deficit (or decreases less than implied by the surplus). On the contrary, a negative SFA means that the government debt increases less than the annual deficit (or decreases faster than implied by the surplus)". This statistical behaviour of SFA in such a short time period raises a number of illegitimacy issues.

d) <u>The Effect of DEBT on GDP</u>.

There is an extensive literature about the effects of government debt on economic activity. In general, an increasing public debt, especially as a percentage of GDP, has been considered responsible for all phenomena which affect economic growth negatively. Although there is no agreement, older theories and empirical research have supported that a high public debt will in the long-term result in the reduction of productive capital and investments. According to Modigliani (1961), the public debt will increase economic growth only for the present generation. More recently,

literature on growth models supports that if public borrowing finances productive investments, public debt will have a positive effect on economic growth (see among others Checherita and Rother 2012). According to other researchers, if public debt becomes too large to pay it back, investments and economic growth will be affected negatively (Krugman, 1988). In fact, empirical research during the decades before the 2007 economic crisis had not reached agreement about the effects of public debt on economic activity, for example, see Ussher (1998). In our case, according to ***Table 8***, as illustrated by the data shown, DEBT has a negative effect on GDP during 2008-2017 in sharp contrast to the three previous sub-periods. It seems that during 1960-2007 government debt has financed productive investments, which have lead to increased GDP as the growth rates suggest in the following ***Table 9.***

Table 9*. Gross Fixed Capital Formation (INV), Annual Growth rates (%)*

	1960-70	1970-80	1980-95	1995-2007	2007-2017	2009-2017
INV	14.0	20.7	9.8	17.7	-4.9	-5.5
GDP	11,8	19,4	18,9	8,1	-1.6	-3,2
INV/GDP	22,8	27,9	22,4	21,5	14,7	12,9
GDP in current market prices, INV in current prices, not including dwellings. *Source:* (a) National Accounts of Greece. (b) [NAIDA_10_GDP] ESA10: selected international annual data						

The above momentum of viable debt through viable investment projects seems broken in 2007-2008 as it is also suggested in diagram (d) of ***DP1*** in section 2.2.

4. CONCLUDING REMARKS AND FURTHER RESEARCH

Considering the long duration, more than a decade, of Greek economic recession, this paper's purpose was to let data 'speak' during 1960-2017. Traditionally, empirical economists take official data for granted. This paper has focussed on data themselves, in particular, on fiscal data, taking notice that the recession started as a fiscal and financial phenomenon. Our research question was to identify probable issues relating to the measurement and composition of this data. Thus, we have examined statistically and econometrically four fiscal variables: government debt, DEBT (stock variable) and deficit, DEF1 (flow variable), as well as the Stock-Flow adjustment (SFA) accounting item, which theoretically is a kind of small residual as of its definition as the difference between two theoretically, but not empirically, equivalent definitions of government deficit. We have used statistical and econometric techniques and we have come up with the following novel findings: It was found that in spite of its definition, SFA is not a small residual, but an important illegitimate factor, which has hugely contributed to the last decade's accumulation of Greek government debt, as previous research has found for other countries (e.g. Campos et al., 2006). It has also caused the deep recession by reducing GDP through its adverse effects on government debt and deficit. A second important finding is that government debt had a significantly positive effect on GDP during 1960-2007; as data in ***Table 9*** illustrate, debt was financing productive investments. This way, debt was viable through wealth creation, this process being interrupted in 2008-09.

By showing the importance of official data legitimacy, this research is useful for contributing to understanding the causes of the Greek crisis beyond political or other,

mainly folklore explanations. This paper is at least thought-provoking. Further research is ongoing, by expanding our sample period beyond 2017 for another two years before the creation of further economic complications because of covid-19, and second by 'digging' deeply into identifying specific illegitimacies in the assembling, composition, valuation and measurement of the crucial fiscal variables examined here. Our conclusions are novel for the Greek case and in agreement with the conclusions of previous similar international research for other European countries as we have made appropriate references in sections 1, 2.1 and 3.3.

ΠΕΡΙΛΗΨΗ

Λαμβάνοντας υπ'όψιν την βαθειά και χωρίς προηγούμενο Ελληνική κρίση που ξεκίνησε το 2008-09 ως ένα μη-αναμενόμενα υψηλό ποσοστό του δημοσίου ελλείμματος ως προς το ΑΕΠ, ο σκοπός αυτού του άρθρου είναι να διερευνήσουμε πιθανές παρατυπίες όσον αφορά την μέτρηση βασικών δημοσιονομικών δεδομένων κατά το κρίσιμο έτος και κατά την περίοδο που ακολούθησε, σε σύγκριση με την περίοδο πριν την κρίση. Για τον σκοπό αυτό χρησιμοποιούμε επίσημα ετήσια στοιχεία του κυβερνητικού χρέους, του ελλείμματος και του κονδυλίου που αποκαλείται Ρύθμιση Αποθέματος-Ροής (SFA) κατά την διάρκεια έξι σχεδόν δεκαετιών από το 1960 μέχρι το 2017. Μεταξύ των οικονομετρικών και στατιστικών ευρημάτων μας είναι τα εξής: (α) Κατά την περίοδο 2008-2017, σε αντίθεση με την περίοδο πριν το 2008, το SFA αφενός αυξάνει το χρέος εντονώτερα από ό,τι το έλλειμμα, και αφετέρου παρουσιάζει εντονώτατες διακυμάνσεις. Και τα δύο φαινόμενα αποτελούν ένδειξη παρατυπιών κατά την συλλογή και μέτρηση του SFA. (β) Αντίθετα με την περίοδο 2008-2017, το χρέος έχει θετική και στατιστικά σημαντική επίδραση επί του ΑΕΠ κατά την περίοδο πριν την κρίση. Τα ευρήματά μας είναι πρωτότυπα και χρήσιμα καθόσον μηνύουν ότι η νομότυπη μέτρηση και σύνθεση των στατιστικών στοιχείων είναι πολύ σημαντική για την διαμόρφωση αποτελεσματικών κοινωνικοοικονομικών πολιτικών.

REFERENCES

Alt, J., Lassen, D. D. and Wehner, J. (2012). Moral Hazard in an Economic Union: Politics, Economics, and Fiscal Gimmickry in Europe. *Weatherhead Center for International Affairs, Harvard University*, Working Paper No. **12**-0001.

Alt, J., Lassen, D. D. and Wehner, J. (2014). It isn't Just about Greece: Domestic Politics, Transparency and Fiscal Gimmickry in Europe. *British Journal of Political Science*, **44**, **4**, 707-716.

Arghyrou, G. M. and Kontonikas, A. (2011). The EMU sovereign-debt crisis: Fundamentals, expectations and contagion. *European Economy, Economic Papers* **436**, Belgium: European Commission.

Bruegel (2012). The Greek Debt Trap:An Escape Plan, by Zsolt Darvas. *Bruegel Policy Contribution*, November, **2012/19**.

Campos, C. F. S., Jaimovich, D. and Panizza, U. (2006). The unexplained Part of Public Debt. *Inter-American Development Bank, Research Department working paper No. 554*. Also in *Emerging Markets Review*, Vol. **7**, 228-243.

Carey, G. (1998). Regression and Path Analysis, University of Colorado, USA.

Checherita-Westphal, C. and Philipp Rother (2012). The Impact of Government Debt on Growth and its Channels. An Empirical Investigation for the Euro-area, *European Economic Review*, **56, 7**, 1392-1405.

ECB (2011). *The Size and Composition of Government Debt in the Euro Area*, European Central Bank Occasional Paper Series, **132**.
EDP Glossary (2021). (https://ec.europa.eu/eurostat/statisticsexplained/), 06/07/2021.
Eurostat (2012). Stock-flow adjustment (SFA) for the Member States, the euro area and the EU27 for the period 2008-2011, as reported in the April 2012 EDP notification.
Eurostat (2014). *Manual on the changes between ESA 95 and ESA 2010.*
Gujarati, D. (2004). *Basic Econometrics*, 4th Edition on line.
Hagen, J. von and Wolff, G. B. (2006). What do deficits tell us about Debt? Empirical Evidence on creative accounting with fiscal rules in the EU. *GESY (Government and the Efficiency of Economic Systems)* Discussion paper No. **148**, Jan. 2006. 33 pages. Also, in *Journal of Banking and Finance*, **30, 12**, 3259-79.
Irwin, T. C. (2012). Accounting Devices and Fiscal Illusions, *IMF Staff Discussion Note*, SDN/**12/02**, 28 March.
Irwin, T. C. (2015). Defining the Government's Debt and Deficit. *Journal of Economic Surveys*, **29**, **4**, 711-732. Also in *IMF Working Papers*, **15/238**.
Koen, V. and van de Noord, P. (2005). Fiscal Gimmickry in Europe: One-Off Measures and Creative Accounting. *OECD Economics Department Working Paper* No. **417**.
Krugman, P. (1988). *End This Depression Now!* W.W. Norton & Co. Inc. New York.
Manual on Government Deficit and Debt ESA95 (2002 Part V, 2010 Part VIII).
Manual on Government Deficit and Debt ESA2010 (2014, 2016 Part VIII).
Modigliani, F. (1961). Long-run implications of alternative fiscal policies and the burden of national debt, *Economic Journal*, **71(284)**, 730-755.
MPRA (2019). Low sample size and regression: A Monte Carlo approach. Accessed at https://mpra.ub.uni-muenchen.de/97017/.
OECD (2018). Data extracted on 07 Jan 2018 20:58 UTC (GMT) from *OECD.Stat*.
Speed, R. (1994). Regression Type Techniques and Small Samples: A Guide to Good Practice. *Journal of Marketing Management*, **10, 89-104**.
Trebesch, C. and Zettelmeyer, J. (2018). ECB interventions in distressed sovereign debt markets: the case of Greek bonds. *Kiel Working paper*, Nr.**2101**.
Ussher, J. L. (1998). Do Budget Deficits Raise Interest Rates? A Survey of Empirical Literature, *Transformational Growth and Full Employment Project*, W.P. **3**, New School for Social Research.
Weber, A. (2012). Stock-Flow Adjustments and Fiscal Transparency: A Cross-Country Comparison. *IMF Working Paper* No. **39**.

Greek Statistical Institute
Proceedings of the 33rd Panhellenic Statistics Conference (2021), pp.281-295

MARKET RISK (BETA) ESTIMATION AND THE ROLE OF CAPITALIZATION DURING A SYSTEMIC CRISIS. THE CASE OF GREECE

A. E. Milionis
Bank of Greece and University of the Aegean
amilionis@bankofgreece.gr

ABSTRACT

The purpose of this empirical study is to investigate the possible effect of the acute systemic economic crisis that stroke Greece on the estimation of systematic risk coefficients (betas) for securities traded at the Athens Stock Exchange, emphasizing on the role of securities capitalization. To this end and amongst others, betas estimated in a business-as-usual before crisis period, were compared with those in a period under the unusual circumstances caused by the acute systemic economic crisis. The investigation focuses on the possible effect of capitalization on the explanatory power of the market model and the comparison of systematic risk estimates (betas) during crisis and non-crisis periods. In sharp contrast to the existing stylized facts, beta values are found to be positively related to capitalization. In addition, capitalization is also positively related to the explanatory power of the market model. Furthermore, crisis caused a substantial drop in beta values except for some of the larger capitalization stocks, predominantly banks.

Keywords: Systematic risk (beta coefficient), Market Model, Firm size effect, Systemic crisis, Beta subsidence, Athens Stock Exchange

1. INTRODUCTION

The connection of return with risk is unquestionable in finance, as is its importance. The most celebrated quantitative expression of this connection is through the so-called (standard) Capital Asset Pricing Model (CAPM), according to which:

$$E(R_j) = R_F + \beta_j[E(R_m) - R_F] \quad (1)$$

where, E represents the mathematical expectation operator, R_j is the return on the security j, R_F is the risk-free rate of return, R_m is the return on the market and β_j is the systematic risk coefficient for security j, also known as beta coefficient. Beta is a measure of the risk arising from exposure to general market movements which cannot be avoided or reduced through diversification.

It is important to note that the standard CAPM not only predicts what should predict returns, but also what should not. For instance, firm characteristics (other than beta) should not predict returns. However, intensive analysis of a large amount of historical

data suggests the existence of several market anomalies (Elton et. al. (2007)). Indicatively, stocks of small firms do better than stocks of large firms, even with risk-adjusted returns (the so-called firm size effect); value stocks do better than growth stocks (the so-called book to market effect).

Furthermore, multifactor models such as the APT, and/or conditional versions of the CAPM have been proposed by several researchers to incorporate additional risk premia (see amongst others Fama and French (1992, 1993, 1996), Milionis (2011)). In sequence, there is a voluminous literature on the ability of CAPM to explain the observed returns and there are serious specification issues about the explanatory power of a single – factor model as CAPM to explain expected returns. In practice however, for several reasons (Jaganathan and Wang (1996), Elton et al. (2007), Milionis (2010, 2011)) the beta coefficient is still the most widely used measure of systematic risk.

As far as beta estimation in practice is concerned, it is noted that CAPM, as expressed in Equation (1) is a general equilibrium model; hence, both the return of the each time particular asset and the market return are expected future returns, while the relevant coefficient of systematic risk is the future beta of the each time particular asset. What is feasible is to perform an ex-post analysis of ex-ante expected returns. In order to replace ex-ante with ex-post realized returns, it has to be assumed that realized returns, over a long period of time, are on average equal to the corresponding expected ones. Hence, the common approach for the estimation of the standard betas is to use the so-called Market Model (henceforth MM) and OLS estimators. In that way, a beta is estimated as the slope coefficient in the regression:

$$R_j = \alpha_j + \beta_j R_m + u_j \quad (2)$$

where u_j is the stochastic disturbance and α_j is a constant.

It is noticeable that the risk-free rate is not included in this simplified version of MM as expressed in Equation (2). For realistic changes in the value of the risk-free rate, there is very little difference in the estimated beta values.

Not taking into consideration phenomena related to market microstructure, one would expect that beta estimates would be invariant to changes in the length of the differencing interval over which index and security returns are calculated. However, a plethora of empirical findings suggest that this is not the case. Indeed, it has been found that beta estimates change systematically as the differencing interval over which they are estimated is lengthened (Cohen et al. (1983a)). This phenomenon, which has been found to exist in both mature and emerging capital markets (Corhay (1991), Oprea, (2015)) is known as "intervalling effect" and results in a biased beta estimate and consequently in a false perception of a security's risk.

The purpose of this work is to examine the possible effect of a systemic economic crisis on the stylized facts regarding the measurement of systematic risk (beta coefficients) in conjunction with securities' capitalization (henceforth cap). In a market under unusual conditions, such as those during an acute systemic crisis,

different stocks may react differently to changes in macroeconomic fundamentals and modify their market risk character. Hence, stylized facts with respect to the measurement of systematic risk may need revising. The case of Greece during the period of its acute systemic economic crisis is a good example of such conditions and its stock market, specifically daily stocks prices traded at the Main Market of the Athens Stock Exchange, will serve as our dataset for analysis. Some preliminary results on this investigation were presented in the 31rd Panhellenic Statistics Conference (Milionis and Karagiota, 2018). Hence, this work may be considered as both self-contained and a continuation of the work of Milionis and Karagiota (2018).

The structure of the paper is as follows: section 2 provides a skeleton review of the basic theory and describes the methodological steps to be followed; section 3 describes the market and the data; section 4 presents the results and commenting; section 5 concludes the paper.

2. BRIEF REVIEW OF THEORY AND METHODOLOGY

2.1 Sampling and smoothing periods

Two time parameters of great importance for the empirical determination of betas are the following: (a) The sampling period (T) that is the total time span covered by the data; (b) The smoothing period that is the differencing interval over which returns are calculated (L). Any price variations within time distance of L time units are ignored in the estimation of a beta which correspond to differencing interval L and, to emphasize that, we use the term "smoothing period". From the statistics point of view, it is apparent that more accurate estimates of betas are obtained the larger the sampling period and the smaller the smoothing period. In contrast, from the point of view of finance it is the other way round. Indeed, large T increases the possibility that significant changes in firm characteristics may have occurred, which directly affect the risk associated with this firm (e.g., increased, or decreased gearing, merger, acquisition, etc.). However, small L entails greater bias in the estimation of beta due to the possible existence of the intervaling effect, as discussed already. So, there must be a tradeoff. For the most developed financial markets, in the recent past, common practice of many financial houses offering beta services (e.g. RMS in the UK, Merrill Lynch and Salomon Brothers in USA) is to use a sampling period of five years and a smoothing period of a month.

2.2 Firm size effect

Firm size effect has attracted much attention by researchers (Banz (1981), Reinganum (1981), Fama and French (1992, 1993, 1996). It is well documented that low capitalization stocks have on average higher returns than high capitalization stocks and even with returns adjusted for risk via CAPM the difference is mitigated but does not disappear. Indeed, Banz (1981), in the probably most often cited work on the

subject, provided evidence that the size effect was almost as important as did beta. Several possible explanations have been stated for these excess returns of small cap stocks. Lower survival probabilities, higher trading costs and misestimated betas due to non-synchronous trading for low cap stocks are the most dominant explanations (Elton et. al. (2007)). However, the fact itself that the difference between average returns of low and high capitalization stocks is reduced after adjustment for risk entails that low capitalization stocks have higher betas than high capitalization stocks. Therefore, one might expect low cap stocks to be more sensitive to the economy than large cap stocks and to do worse than the latter during economic hard times. The main subject this work is to try to investigate, whether or not, this stylized fact of the betas-capitalization connection has been affected, and if so in what direction, by the systemic crisis.

2.3 Intervalling effect

The characteristic of market microstructure responsible for the effect of the differencing interval on beta estimation is friction in the trading process and the entailed price adjustment delays (Cohen et al. (1983b)). Due to friction in the trading process, true returns, which would be observed in a frictionless market, are not observable. Changes in true returns are faster than in observed returns, so in each period only a part of the true returns is incorporated in the observed ones. The abovementioned can be expressed quantitatively by the following model which connects the observed returns (R_j^o) with the true returns (R_j) and was proposed by Cohen et al. (1983b):

$$R_{jt}^o = \sum_{n=0}^{N} \gamma_{j,t-n,n} R_{j,t-n}$$

with $$E(\sum_{n=0}^{\infty} \gamma_{j,t,n}) = 1 \quad \forall j, t$$

Where, $\gamma_{j,t-n,n}$ are random variables representing the portion of true returns of security j at time $t-n$ which is reflected in observed returns at time t.
Therefore, the MM for true and observed returns respectively is written as:

$$R_{jt} = a_j + \beta_j R_{mt} + u_{jt}$$
$$R_{jt}^o = a_j^o + \beta_j^o R_{mt}^o + u_{jt}^o$$

where β_j and β_j^o express the systematic risk for true and observed returns respectively. β_j and β_j^o are related through the following equation (Cohen et al. (1983b)):

$$\beta_j^o = \beta_j(1 + 2\sum_{n=1}^{N} \beta_{m+n}^o) - \sum_{n=1}^{N}(\beta_{j+n}^o + \beta_{j-n}^o) \qquad (3)$$

where,

$$\beta_{m+n}^o = \frac{Cov(R_{m,t+n}^o, R_{m,t}^o)}{Var(R_{m,t}^o)}, \quad \beta_{j\pm n}^o = \frac{Cov(R_{j,t\pm n}^o, R_{m,t}^o)}{Var(R_{m,t}^o)}$$

From Equation (3) it is obvious that betas for the observed returns depend on betas for the true returns, on the autocorrelations of stocks and index observed returns up to

order N and on the cross-correlations among the observed returns of the security j and the market index m up to order N. Consequently, the OLS estimator of beta for observed returns is a biased estimator of the systematic risk. From Equation (3) it is evident that the bias of the beta estimator will tend to decrease as the differencing interval is lengthened.

In view of the above, it is necessary to examine the role of capitalization on beta estimates not only using daily returns, but also taking into account the possible influence of the differencing interval on beta estimates, i.e., examining the role of capitalization using returns with larger than one-day smoothing periods.

3. DATA AND THE MARKET

3.1 Data description

The data to be used in this work are daily closing prices of all stocks traded in the Main Market of the Athens Stock Exchange (ASE). As in most similar studies the sampling period *T* will be a five-year period from 25/09/2012 to 22/09/2017. Returns are expressed as the logarithmic difference of price relatives, using the unit differencing interval. The smoothing period will vary from one to thirty days. The Athens General Index (GEN) will be used as the market index. For the given sampling period, totally there are 122 stocks traded continuously in the Main Market of ASE.

As mentioned, the period under study was very unusual and, as such, very condensed in terms of important economic – political events. This is evidenced in Table 1, which presents the most important of these events.

3.2 The market

ASE started trading in 1876. From May 31, 2001, the ASE joined the Mature Financial Markets according to the classification of Morgan Stanley Capital International (MSCI). Until that date ASE belonged to the European Emerging Markets. In November 2013 the same institution downgraded ASE to the Emerging Markets status. ASE was also downgraded by FTSE on March 21, 2016, in the category of Advanced Developing Markets.

It is remarkable that until the late 1980s ASE was a small market with low capitalization value and a thin volume of trade, while participants were almost exclusively local investors. However, the changes in the legislative and regulatory framework, during the 1990s, towards harmonization with the standards of the more developed financial markets have brought significant changes for ASE (for details, see for instance Milionis et al. (1998)). As a consequence of these reforms the total capitalization of ASE increased from 13.6 million USD by the end 1993 to 145 million USD by the end 2005 (source: World Federation of Exchanges). Furthermore, foreign investors owned less than 10% of the total capitalization in 1993

(approximate indirect assessment, as no exact official figure exists,) but more than 37% of the total capitalization at end of April 2005, while their share in the average daily volume of transactions in April 2005 exceed 54% (source: Central Securities Depository of ASE, Monthly Statistical Bulletin, April 2005).

The capitalization of ASE during the period under consideration, as compared to that of 2005, had been significantly decreased, as a consequence of the economic crisis, to approximately 45 billion USD. Since then, the capitalization is gradually recovering and nowadays, as at end of first half of 2021, has exceeded 60 billion USD, while about two thirds of total share capital is now owned by foreign investors. Further details about the ASE and its institutional and operational evolution are given in many papers (see for instance Milionis and Papanagiotou (2009, 2011)).

Table 1. *The most important economic – political events during the sampling period*

Events during the sampling period *	
Date	Event
7/11/2012	Approval of the medium-term rescue program for Greece 2013-2016
11/2013	Downgrade of the Athens Stock Exchange by MSCI (Emerging market)
25/5/2014	European elections (win of SYRIZA party)
17/12/2014	Greek Parliament failed to elect President of the Republic (First time)
23/12/2014	Greek Parliament failed to elect President of the Republic (Second time)
29/12/2014	Greek Parliament failed to elect President of the Republic (Third and last time)
25/1/2015	Early national parliamentary elections
20/2/2015	First agreement between the new (SYRIZA-ANEL) government and the lenders
29/6/2015	Imposition of bank holiday and restrictions on capital movements
5/7/2015	Referendum, Closure of banks and imposition of a take-over cash limit. Closure of the Athens Stock Exchange
14/8/2015	Resolution of the third memorandum
20/11/2015	Recapitalization of the banks
21/3/2016	Downgrade of the Athens Stock Exchange by FTSE (Advanced Developing market)

*The list is indicative not exhaustive.

4. RESULTS AND DISCUSSION

4.1 MM model

The empirical results of previous studies have shown that the estimates of beta coefficients resulting for the same smoothing period *L*, but for a different starting day of returns calculation, may differ (Corhay (1992), Milionis (2011)). Therefore, to obtain all estimates of betas of each stock it is necessary to estimate *L* regressions for every smoothing period *L*, with *L* varying from one to thirty days. For the same *L* the data for each regression will start each time one day later. Subsequently, the average value of beta estimates for every *L* is calculated. The results show that from the whole sample of 122 stocks 93 of them have statistically significant beta values, at 5% level of significance, for returns calculated using $L=1$. The results, regarding the statistical significance of betas for $L>1$ are similar (space limitations do not allow full presentation of statistical results, further details are available by the author).

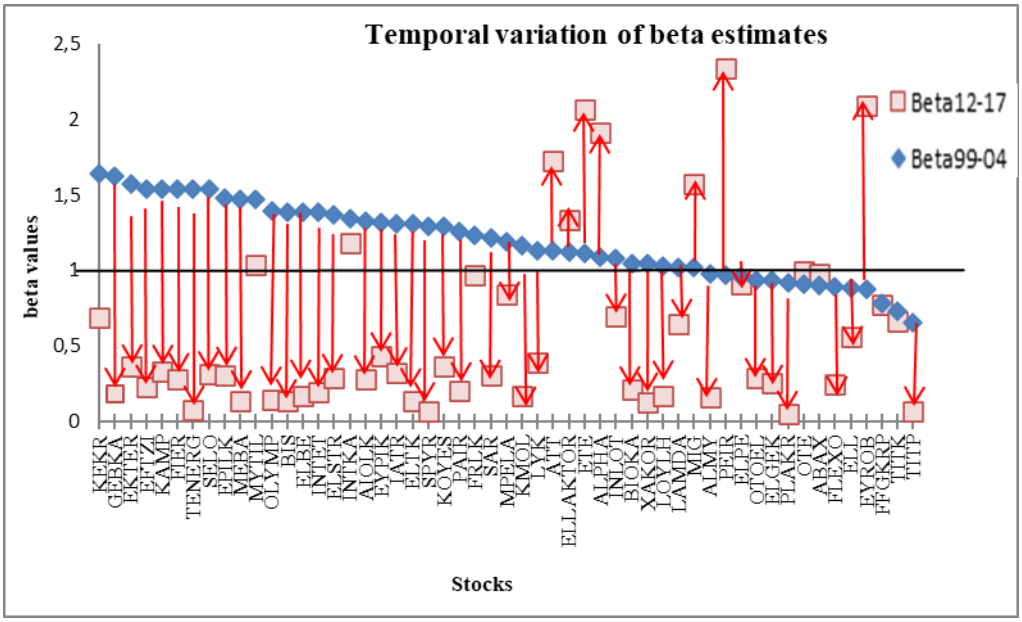
4.2 Temporal variation of beta estimates

Temporal variation of beta estimates has attracted much interest in the literature. As found by Blume (1971, 1975) and Vasicek (1973), who estimated betas over two adjacent sampling periods, beta estimates far from unity in one sampling period tend towards unity in the next sampling period. In this study it is important to compare beta values estimated in a sampling period amidst the economic crisis, with beta values estimated in a sampling period outside the crisis period. In that way, it may be possible to identify possible effects of economic crisis on systematic risk estimation. Methodologically, for such a comparison to be meaningful the following assumptions should be met: (a) only stocks which were continuously traded in both periods should be considered; (b) the same statistical estimation procedure should be used for both sampling periods; (c) the two sampling periods should not be adjacent (in order to avoid the convergence to one phenomenon).

For ASE published beta values based on OLS estimation and daily smoothing period for all stocks are available for the period 1999 – 2004 (Milionis (2010)). This period is obviously not adjacent to the sampling period used thus far, so beta values in both periods are not affected by the convergence to one effect mentioned previously. Moreover the 1999-2004 period is a business-as-usual period, as it is well before the Greek economic crisis. Hence, in view of the fulfilment of the aforementioned assumptions, it is a period suitable for our purposes.

Figure 1 exhibits the results of this comparison. The 1999-2004 beta values are presented in descending order in the figure. The horizontal axis shows the corresponding stocks' symbols. The arrows show the corresponding 2012-2017 beta values, as well as, graphically, the deviation from the original values.

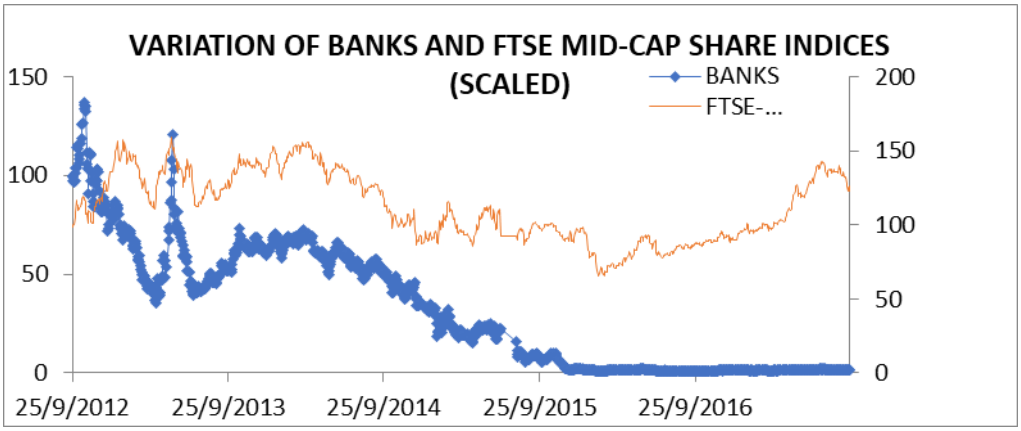
Figure 1. *Temporal variation of beta estimates*



From Figure 1 it is immediately apparent that there is a very conspicuous drop in the 2012-2017 betas. Indeed, the values of beta estimates for the second period for the great majority of securities are much smaller than their values of the first sampling period. This empirical finding will be called beta "subsidence" henceforth. Exceptions to this general tendency are the betas of some of the higher capitalization stocks, the most prominent of which are those for the betas of the securities of the banking sector, specifically National Bank of Greece (symbol ETE), ALPHA BANK (symbol ALPHA), Bank of PEIRAEUS (symbol PEIR), EUROBANK (symbol EUROB), and ATTIKA BANK (symbol ATT). This reflects the fact that the monetary-financial institutions were most responsive and those affected the most by the turbulence caused by the crisis. To further support this argument, the next figure (Figure 2) presents the movement of the Banks Stock Index in conjunction with the FTSE mid-capitalization Index of ASE, both scaled with common starting value equal to 100.

As is evident, the two indices share very little in common in terms of their progression. This is most conspicuous in the latest part of the figure. It is remarkable that the Banks Index lost almost all its value after the imposed shutdown of the banks and ASE (July 5, 2015, details in Table 1), while at the same time the movement of the FTSE mid-capitalization Index was vastly different. This provides further evidence that stocks did not react in the same manner given the crisis and that a one-factor model may not be sufficient to acceptably describe securities' risk for all stocks. Hence, further investigation is necessary.

Figure 2. *Variation of Banks and Mid-Capitalization Indices (scaled)*



4.3 Further investigation: an insight into the role of capitalization

As will be shown shortly capitalization is very crucial in the behaviour of stocks in relation to the above findings and examination of its role will provide a further insight.

Capitalization and MM

Analyses regarding beta estimates are in most cases conducted using portfolios of securities rather than individual securities. This approach is preferred because econometrics-wise errors in estimating individual securities' betas would cancel out when securities are grouped together, resulting in a smaller error in measuring the portfolio beta. Indeed, it is easily proved that with a portfolio of securities in equal proportion the standard error of the portfolio beta is equal to the average beta of individual stocks divided by their number. In sequence, historical portfolio betas are better predictors of future betas than are historical betas of individual securities. Thus far the existing evidence shows that the explanatory power of the market model using portfolios of securities does not depend on the beta value or the capitalization of the portfolio (see for instance the very influential work of Black et. al. (1972)). However, in grouping securities some subtle, yet important information may be missed out. Therefore, in this work our investigation will be conducted on the individual securities' level.

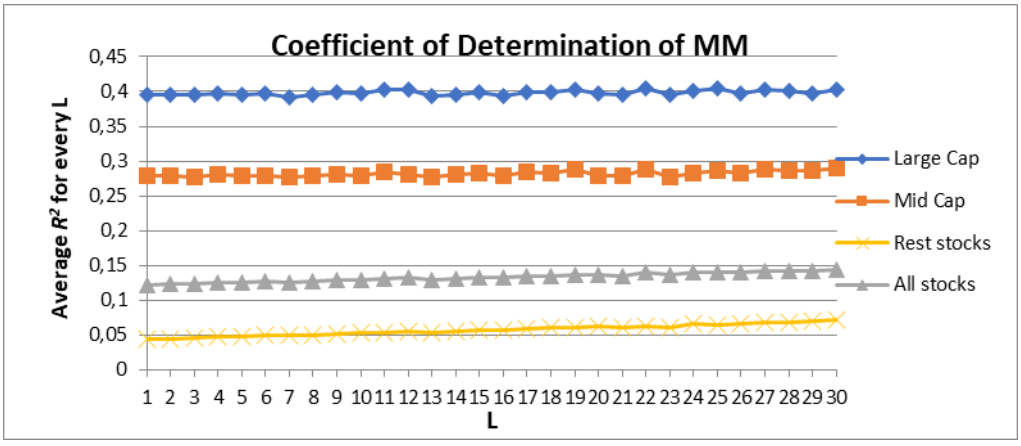
The allegation about the sufficiency of a one factor model such as the Market Model to capture securities' risk and the entailed question regarding the credibility of betas, especially those of the comparatively lower capitalization securities, as stated in previous sections, will now be examined further. At first, stocks were taxonomized according to capitalization value. To this end, the FTSE Large Capitalization and Mid Capitalization Stock Indices of ASE were used to differentiate securities according to

capitalization (see the official website of ASE: https://www.athexgroup.gr the for the composition and further details about the FTSE indices of ASE). Securities not belonging to these two indices are those of smaller capitalization value and were grouped together as "rest of stocks". Then, the average value of the Coefficient of Determination of the MM model for the stocks belonging to each category was calculated.

Figure 3 presents the variation according to L of the average explanatory power of the MM model, expressed with the value of the Coefficient of determination (R^2), for the various stock categories. As is immediately seen, there is a very sharp distinction in the explanatory power of the MM model for the various categories, so there is no doubt that capitalization is responsible for this distinction. Indeed, MM for stocks with the higher capitalizations (i.e., those stocks belonging to the Large-Cap Index) fits much better than in the other two categories, and MM for stocks with the medium capitalizations fits much better than MM for stocks with the relatively lower capitalization.

A second conclusion that is derived from Figure 3 is that the mean Coefficient of Determination (mean R^2 value) (mildly) increases with L for all categories. This is consistent with the existing stylized facts regarding the microstructure of MM (Cohen et al. (1980), Cohen et al. (1983)).

Figure 3. *Coefficient of determination of MM versus capitalization and smoothing period*

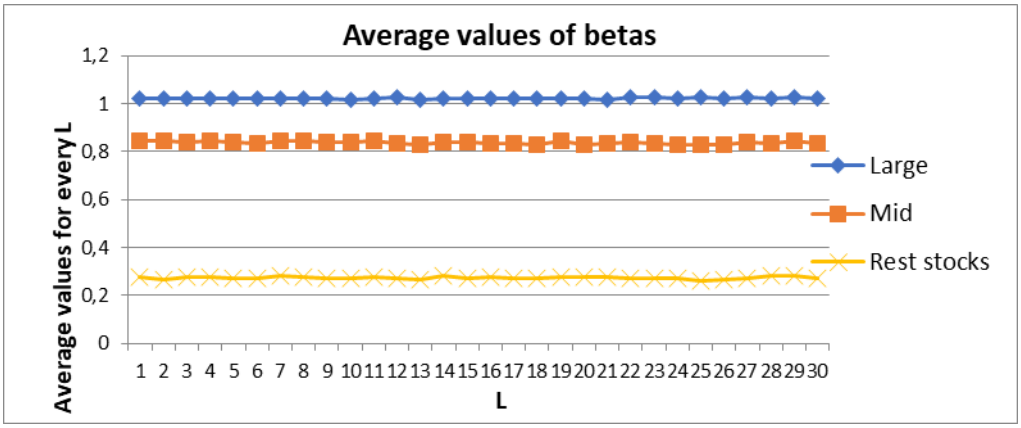


Capitalization and beta estimates

From the results presented thus far, especially the beta "subsidence" documented in subsection 4.2, there is evidence that stock capitalization seems to be related with the corresponding beta value and it is of interest to search this potential influence further. Taxonomizing again the securities in the same way as previously, Figure 4 shows the evolution of average betas for each category as a function of L. From Figure 4 at first it is evident that there is a sharp discrimination in the average beta among the three

categories. Indeed, the larger capitalization stocks have clearly the highest beta values (average beta 1.02), the lowest capitalization stocks clearly the lowest beta values (average beta only 0.28), while stocks of the Mid-Cap Index have an average beta of 0.87. Further, the characteristics of this discrimination is unaffected by the smoothing time, in contrast to other studies in which smoothing time affects low capitalization stocks more than large capitalization stocks (Brzeszczynski et al., 2011).

Figure 4. *Average beta versus capitalization and smoothing period*



Temporal variation, Capitalization and beta estimates

The results presented previously in section 4.2 clearly show that from the observed phenomenon of generalized substantial drop in beta values for the majority of stocks between the two sampling periods there was (inevitably) an exception which refers exclusively to some of the larger capitalization stocks. Therefore, it makes sense to examine these empirical findings at a higher measurement level. Figure 5 shows the scatterplot of the percentage change of beta values between the two sampling periods (1999-2004 and 2012-2017) and the corresponding logarithm of mean capitalization of the second period for all stocks traded continuously in both sampling periods. As is evident, indeed, smaller capitalization stocks have negative percentage change of betas, while positive percentage changes of betas, which occur to a minority of stocks, are associated predominantly with larger capitalization stocks.

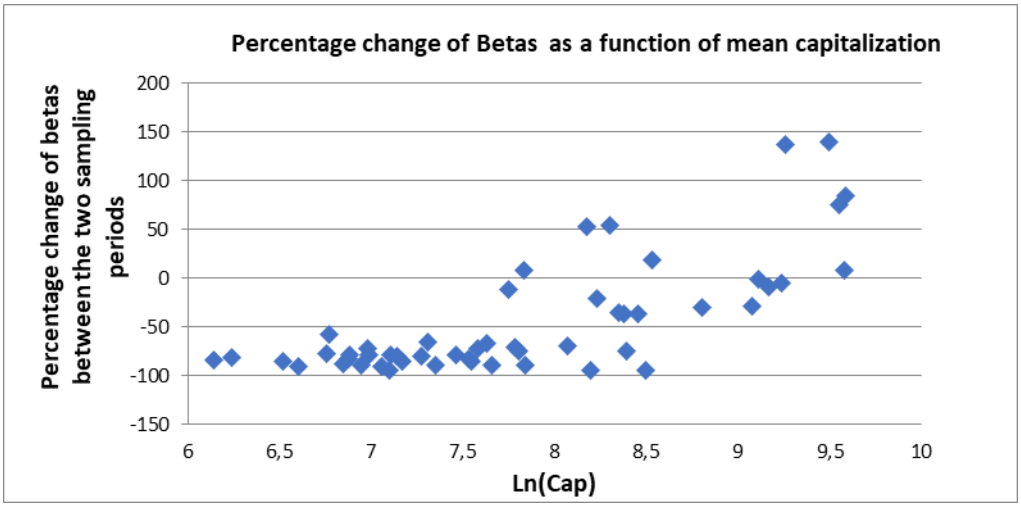
To express the dependence of percentage change in beta values on capitalization, after some experimentation we reached to the following parabolic model (in passing, it is noted that selecting among models with different dependent variable, such as *PCB* and Ln(*PCB*) is not straightforward and the procedure suggested by Gujarati and Porter (2009) was followed).

$$PCB = 727.40 - 242.86\, ln(cap) + 18.18\, ln(cap)^2 + e \qquad R^2 = 0.64$$

where *PCB* is the percentage change in beta between the two sampling periods, ln(*cap*) is the (natural) logarithm of mean capitalization for the sampling period 2012-2017, and *e* represents the residuals. As the value of the Coefficient of Determination

implies, again, capitalization plays an important role, as the model can explain almost two thirds of the variation in the percentage changes of betas.

Figure 5. *Percentage change of betas as a function of mean 2012-2017 capitalization*



4.4 Further discussion

From what we know form finance theory thus far, a low beta value for a company's stock may be attributed to such factors as the character of capital structure (e.g., low leverage), the character of company's activities (e.g., public utility companies, or companies from the food and beverages sector, have traditionally low betas) etc. In the present study, given the observed beta "subsidence", we very much doubt that this is the main reason for the estimated low beta values in the sampling period under investigation. Instead, the main mechanism which is responsible for the observed low betas seems to have a different origin. Indeed, as documented in this work, there are sharp differences in the explanatory power of the MM for the various stocks for the particular sampling period. Furthermore, low explanatory power of the MM model is associated with low betas. Clearly, this association is not unidirectional: low explanatory power in MM is the cause and low corresponding betas is the response (outcome). Further, low explanatory power in MM is related with low capitalization. However, the original cause in this cause-and-effect chain is the systemic economic crisis, which affected ununiformly the stocks traded in ASE. Banks, and some other companies, most of which have a large weight on the construction of the General Index of ASE, were those with the higher explanatory power of the MM, and at the same time those most affected by the crisis, and attracted most of the volume of trade (it is noted that during the sampling period approximately 90% of the daily value of transactions is attributed to the stocks of the FTSE Large Capitalization Index, source: Statistical Bulletin of Central Securities Depository of ASE). As a result of the crisis effect, the average explanatory power in the "rest of stocks" category is only 4,3%. Inevitably, even though this value of average explanatory power is statistically

significant, this means that MM for such stocks does not seem to adequately determine a stock's systematic risk. It must be noted that a significant beta value from the statistics point of view, as confirmed for the majority of stocks in this work (see section 4.1), does not necessarily have the same importance from the finance point of view, especially in view of the fact that the explanatory power of the MM model may vary substantially among stocks. In the present case, during the crisis, low R^2 values in MM caused low betas mainly for low capitalization stocks. For these low beta-low R^2 of MM stocks, additional risk premia may exist, which cannot be captured by a one factor model such as the MM.

The above discussion offers the opportunity to remark that a low explanatory power of MM as a reason for low betas may not be true exclusively for the circumstances of this study but have a wider applicability. Nevertheless, a detailed analysis on the relationship of R^2 of MM with beta estimates and capitalization has not been given the required attention thus far, partly because empirical studies on systematic risk very often use portfolios of stocks (Elton et. al. (2007)).

5. SUMMARY AND CONCLUSIONS

The purpose of this empirical study was to compare betas estimated in a business- as-usual period, with those in a period under unusual circumstances and identify possible effects of such circumstances on beta estimation. The Greek acute systemic economic crisis certainly represents vividly such circumstances. The results provide counterevidence for some stylized facts about beta estimation.

Specifically, the key role of capitalization in the character of systematic risk estimation was evident. For low capitalization stocks MM has a much lower explanatory power than for larger capitalization stocks and further low capitalization stocks have lower betas. These findings sharply contradict the existing stylized facts according to which not only the explanatory power of MM model does depend on capitalization, but also the larger capitalization securities are those associated with lower betas. As documented, these results have their origin to the fact that the systemic economic crisis affected ununiformly the stocks traded in ASE. The economic crisis was also the original cause of what we called beta "subsidence", i.e., the generalized drop in beta values between the sampling periods 1999-2004 and 2012-2017, except for some large capitalization stocks, predominantly banks.

It is also noted that for low capitalization stocks, beta estimates, even though were found to be significant in the statistical sense, should be seen with caution by the financial analysts and portfolio managers as they are derived by a MM with very low explanatory power. It is speculated that over the particular sampling period a single factor model, as MM, may not sufficiently capture the structure of the security returns, and a multifactor model, such as the APM, may be more appropriate.

ΠΕΡΙΛΗΨΗ

Σκοπός του παρόντος είναι να εξετάσει την ενδεχόμενη επίδραση της οξείας κρίσης που έπληξε την Ελληνική Οικονομία στις εκτιμήσεις συστηματικού κινδύνου (συντελεστές βήτα), ιδιαίτερα δε την πιθανή αναθεώρηση του ρόλου της κεφαλαιακής αξίας των αξιογράφων στη μέτρηση του συστηματικού κινδύνου των αξιογράφων που διαπραγματεύονται στο Χρηματιστήριο Αξιών Αθηνών. Για το σκοπό αυτό χρησιμοποιήθηκαν και εκτιμήσεις συστηματικού κινδύνου κατά τη διάρκεια χρονικής περιόδου προ κρίσης. Ορισμένα από τα εμπειρικά ευρήματα βρίσκονται σε οξεία αντίθεση με τα ισχύοντα στη διεθνή βιβλιογραφία. Πράγματι, στοιχειοθετήθηκε ότι η επεξηγηματική ικανότητα του Μοντέλου Αγοράς σχετίζεται θετικά με την κεφαλαιοποίηση των αξιογράφων, ενώ σύμφωνα με τα μέχρι σήμερα ισχύοντα τέτοια σχέση δεν υφίσταται. Περαιτέρω, βρέθηκε ότι οι υψηλές τιμές συντελεστών συστηματικού κινδύνου σχετίζονται με αξιόγραφα υψηλής κεφαλαιοποίησης, ενώ σύμφωνα με τα μέχρι σήμερα γνωστά εμπειρικά ευρήματα ισχύει ακριβώς το αντίθετο. Επιπλέον οι εκτιμήσεις συστηματικού κινδύνου κατά την περίοδο της κρίσης ήταν ουσιωδώς μικρότερες από τις αντίστοιχες προ κρίσης για τη μεγάλη πλειοψηφία των αξιογράφων με εξαίρεση μερικές μετοχές υψηλής κεφαλαιοποίησης, κυρίως μετοχές τραπεζών. Όπως επιχειρηματολογείται, η οξεία συστημική οικονομική κρίση είναι η γενεσιουργός αιτία αυτών των φαινομενικά απρόσμενων αποτελεσμάτων.

REFERENCES

Banz R. W. (1981). The relationship between return and market value of common stock. *Journal of Financial Economics*, **9**, 3-18.

Black F., Jensen M. C., and Scholes M. (1972). The Capital Asset Pricing Model: some Empirical Tests, in Jensen, (ed) *Studies in the theory of capital markets*, New York, Praeger.

Blume M. (1971). On the assessment of risk. *Journal of Finance*, **VI(**1), 1-10.

Blume M. (1975). Betas and their regression tendencies. *Journal of Finance*, **X**(3), 785-795.

Brzeszczynski J., Gajdka J., and Schabek T. (2011). The Role of Stock Size and Trading Intensity in the Magnitude of the "Interval Effect" in Beta Estimation: Empirical Evidence from the Polish Capital Market. *Emerging Markets and Trade*, 47(1), 28-49.

Cohen C., Hawawini G., Maier S., Schwartz R. and Whitcomb D. (1980). Implications of Microstructure Theory for Empirical Research on Stock Price Behavior. *Journal of Finance*, **XXXV**(2), 249-247.

Cohen C., Hawawini G., Maier S., Schwartz R. and Whitcomb D. (1983a). Estimating and adjusting for the intervalling effect bias in beta. *Management Science*, **29**, 1.

Cohen C., Hawawini G., Maier F., Schwartz R. and. Whitcomb D. (1983b). Friction in the trading process and the estimation of systematic risk. *Journal of Financial Economics*, **12**, 263-278.

Corhay A. (1992). The intervalling effect bias in beta: A note. *Journal of Banking and Finance*, **16**, 61-73.

Dimson E. (1979). Risk measurement when shares are subject to infrequent trading. *Journal of Financial Economics*, **7**, 197-226.

Elton E., Gruber M. and Brown S., Goetzmann W. (2007*). Modern portfolio theory and investment analysis,* 7^{th} edition, Wiley, NY.

Fama E. and French K. (1992). The cross section of expected stock returns. *Journal of Finance,* **47***(*1), 427-466.

Fama E. and French K. (1993). Common risk factors in the returns on bonds and stocks. *Journal of Finance*, **33**(1), 3-56.

Fama E. and French K. (1996). Multifactor explanations of asset pricing anomalies. *Journal of Finance*, **51**(1), 55-84.

Gujarati D. N. and Porter D. C. (2009). *Basic Econometrics*, Prentice Hall.

Jaganathan R. and Wang Z. (1996). The conditional CAPM and the cross section of expected returns. *Journal of Finance,* **51**(1), 3-53.

Milionis A. E. (2010). Unbiased estimates of systematic risk for the Athens Stock Exchange, *in Essays for the Greek monetary-financial system, Athens University of Business and Economics*, 295- 308, in Greek with an English summary.

Milionis A. E. (2011). A conditional CAPM; implications for systematic risk estimation, *The Journal of Risk Finance*, **12**(4), 306-314.

Milionis A. E., Moschos D. and Xanthakis M. (1998). The influence of foreign markets on the Athens Stock Exchange. *Spoudai Journal of Economics and Business*, **14**(1-4), 140-156.

Milionis A. E. and Papanagiotou E. (2009). A study of the predictive performance of the moving average trading rule as applied to NYSE, the Athens Stock Exchange and the Vienna Stock Exchange: sensitivity analysis and implications for weak-form market efficiency testing. *Applied Financial Economics,* **19**(14), 1171-1186.

Milionis A. E. and Papanagiotou E. (2011). A test of significance of the predictive power of the moving average trading rule of technical analysis based on sensitivity analysis: application to the NYSE, the Athens Stock Exchange and the Vienna Stock Exchange. Implications for weak-form market efficiency testing, *Applied Financial Economics,* **21**(6), 421-436.

Milionis A. E. and Karagiota V. E. (2018). Systematic risk and market microstructure in a market under conditions of economiccrisis. the case of the Athens Stock Exchange, *Proceedings of the the 31rd Panhellenic Statistics Conference*, 314-328.

Oprea D. S. (2015). The interval effect in estimating beta: empirical evidence from the Romanian Stock Market, *Review of Finance and Banking*, 7(2), 16-25.

Reinganum M., R. (1981) Misspeification of Capital Asset Pricing: Empirical anomalies based on earnings yields and market values, *Journal of Financial Economics*, **9**, 19-46.

Vasicek O., (1973). A note on using cross sectional information in Bayesian estimation of security betas. *Journal of Finance,* **VIII**(5), 1233-1239.

Greek Statistical Institute
Proceedings of the 33rd Panhellenic Statistics Conference (2021), pp.296-312

FORECASTING ACTUARIAL TIME SERIES: A STUDY OF THE EFFECT OF "LINEARIZATION" AND DATA TRANSFORMATION

A. Milionis[1], N. Galanopoulos[2], P. Hatzopoulos[2], A. Sagianou[2]
[1]Bamk of Greece and University of the Aegean
amilionis@bankofgreece.gr
[2]University of the Aegean
(sasd18001, xatzopoulos, asagianou)@aegean.gr

SUMMARY

The accurate prediction of mortality rates plays a crucial role in the management of longevity risk, one of the most important risks in the actuarial industry. In this paper the effect of time series "linearization", as well as that of possible transformations of actuarial time series are examined. The time series we considered are the period indices uncovering the mortality trend for England-Wales. The results show only a marginal improvement in terms of point forecasts, but a substantial improvement in terms of confidence interval forecasts. In practice, this improvement in forecasts confidence intervals can redefine the Solvency Capital Requirements, and subsequently the Solvency Ratio for pension funds, thus, putting pension providers at a competitive advantage as they have less capital locked in their liabilities.

Key Words: mortality rates – forecasts, time series "linearization", time series transformation

1. INTRODUCTION

In recent years, due to the developments in technology and medicine, insurance companies have to deal with a significant risk, namely the longevity risk, which is raised due to the uncertainty in the future movement of mortality rates. In this direction, several regulations have been introduced with the aim of maintaining the balance of an organization's reserve funds and control the inherent risks. Solvency II (Directive 2009/138/EC) introduces a uniform system of calculating capital requirements among all EU Members to ensure organizations' solvency and risk management ability. This capital requirement is called Solvency Capital Requirement (SCR) and covers all the risks that an insurer faces. Considering the above, Solvency II requires that an insurer holds enough reserves to cover 99.5% of scenarios which might occur over a one-year time horizon. According to the literature, it has been observed that the risk of longevity unfolds over many years and hence, it does not naturally fit into the one-year time horizon required by regulators in banking and

insurance (Kleinow, T. and Richards, SJ, (2017)). Thus, the analysis of longevity risk requires the projection of longevity related data in a time horizon of several years.

An approach to deal with the aforementioned problem is to model the mortality rates by using the mortality models reported in the literature and predict the trend of the mortality in the future. By doing so, an insurance company can reinforce the capital requirements determination process. The methodical and accurate prediction of mortality rates plays a crucial role in the management of longevity risk.

To do so, a number of stochastic models have been developed to analyze and model the mortality improvements (Lee and Carter (1992), Renshaw and Haberman (2006), Currie (2006), Cairns et al. (2006), Hyndman and Ullah (2007), Plat (2009), Hatzopoulos and Haberman (2011), Hatzopoulos and Sagianou (2020)). By utilizing historical mortality data, a stochastic mortality model can uncover the trend of mortality rates and provide a deeper understanding to the mortality dynamics. The extracted mortality rates can be used to extrapolate mortality behaviour in the future, with the aim to uncover the future behaviour of mortality trends.

In this paper, we use the multiple-component stochastic mortality model of Hatzopoulos-Sagianou (hereafter called HS), as developed in Hatzopoulos and Sagianou (2020), to model the mortality dynamics. The HS model uses a semi parametric estimation method. This method adopts Generalized Linear Models (GLMs) and Sparse Principal Components Analysis (SPCA). The SPCA requires the definition of a sparsity factor (s value) in order to pinpoint the optimal and most informative age–period and age–cohort components. To do so, the definition of the sparsity factor is based on a methodology tailored for the HS model and can measure the Unexplained Variance (UVR) of each of the age–period and age–cohort components that are incorporated in the proposed model. For more details about the novel dynamic structure and estimation method of the HS model see Hatzopoulos and Sagianou (2020).

In the family of age-period-cohort stochastic mortality models the dynamics of mortality are driven by the period and the cohort indices. Therefore, the forecasting of mortality rates requires the modelling of these indices using time series techniques. For the period indices, we assume the random walk with drift (henceforth RWD) model, the standard approach in the actuarial literature (Lee and Carter (1992), Cairns et al. (2006, 2011), Pitacco et al. (2009), Haberman and Renshaw (2011), Lovász (2011), Villegas et al. (2018)): $Y_t = d + Y_{t-1} + u_t$ where Y_t is a stochastic time series, u_t is a white noise process, and d is a constant. However, other stochastic models have also been employed in order to enhance forecasts (Lee and Miller (2001), Plat (2009), Villegas et al. (2018), Hatzopoulos and Sagianou (2020)). One strand of such models are the AutoRegressive Integrated Moving Average (henceforth ARIMA) models. An ARIMA model is a concise quantitative summary of the internal dynamics of a time series in a linear framework and it is useful, among several other reasons, for forecasting. However, time series from the real world need to undergo some statistical preparation and pre-adjustment as they are not usually "ready" to be

used for forecasting purposes in the aforementioned context. This is so because in time series of raw data variance non-stationarity may be present. Furthermore, very often there exist outliers and other causes (such as calendar effects, etc.) that disrupt the underlying stochastic process. Their treatment is known as "linearization". Variance non-stationarity not only increases time series variance, as does the existence of outliers, but also affects the character of the ARIMA model and the selection and character of outliers (Milionis, 2003; Milionis, 2004; Milionis and Galanopoulos, 2019). Inevitably, both variance non-stationarity and outliers affect point and interval forecasts. Hence, forecasts are less accurate with increased confidence intervals. In the actuarial context, this can adversely affect the longevity risk management process. More specifically, outliers and possible instability in the variance in longevity data, entail an increase in time series variance which in turn is expected to affect, amongst others, the uncertainty about the solvency capital requirements in a pension fund or insurance institution. However, if outliers in the actuarial time series data reflect events in the real world that occur or repeat themselves very rarely (e.g. the Spanish influenza of 1917, world wars etc), then it may be possible to improve the quality of forecasts by properly controlling for their influence. A similar fact that exists nowadays, could prove to be the Covid-19 with the increasing number of deaths due to this pandemic. In a possible future study of mortality rates, we need to consider the existence of the pandemic and its impact on mortality in order to possibly achieve a more accurate and qualitative prediction.

However, despite the importance of point, as well as of interval forecasts, on actuarial time series, in particular on mortality rates, the possible existence and character of variance instability and outliers, their importance and influence on such forecasts, and the possible implications for actuarial models' performance have not been systematically studied thus far. This is indeed the scope of this work. To this end, the RWD model will be used as a benchmark. Furthermore, stochastic models will be employed with and without preadjustments in order to assess the effect of the later on the quality of forecasts. The intention is clearly towards a practical approach.

The manuscript is structured as follows: in section 2 we briefly discuss the statistical treatment of variance non-stationarity and outliers in conjunction with an ARIMA model. In section 3 we describe the data and the software to be employed. In section 4 we present and comment upon our results. In section 5 we summarize and conclude.

2. STATSTICAL PREADJUSTMENTS

By and large, statistical preadjustments for time series modelling comprise: (i) data transformation; (ii) data linearization. When variance is non-stationary the principal problem is the misspecification of the ARIMA model and, in sequence the possibly false identification of the various types of outliers. However, if variance is somehow functionally related to the mean level it is possible to select a transformation to stabilize the variance. Widely used transformations to tackle this problem belong to

the class of the power Box and Cox transformation (Box and Cox, 1964). For example, very often used transformations are given by:

$$f(z_t) = \begin{cases} \frac{z_t^{\lambda}-1}{\lambda} \; if \; \lambda \neq 0 \\ lnz_t \; if \; \lambda = 0 \end{cases} \quad (1)$$

After the examination for a data transformation of the original data, time series itself or some variance stabilizing transformation of it, is "linearized" according to the general framework (Kaiser and Maraval (2001)):

$$y_t = w_t' b + C_t' \eta + \sum_{j=1}^{m} \alpha_j \mu_j(B) I_t(t_j) + x_t \quad (2)$$

Where: $y_t = f(z_t)$, f is some transformation of the raw series z_t, which may be necessary to stabilize the variance;
$b = (b_1, \ldots, b_n)$ is a vector of regression coefficients;
$w_t' = (w_{1t}, \ldots, w_{nt})$ denotes n regression or intervention variables;
C_t' denotes the matrix with columns possible calendar effect variables (e.g. trading day) and η the vector of associated coefficients;
$I_t(t_j)$ is an indicator variable for the possible presence of an outlier at period t_j;
$\mu_j(B)$ captures the transmission of the j-th effect and α_j denotes the coefficient of the outlier in the multiple regression model with m outliers;
x_t follows in general a multiplicative seasonal ARIMA(p,d,q) model:

$$\varphi(B)\nabla^d x_t = \theta(B)\varepsilon_t \quad (3)$$

where:

- $\varphi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$ is the so–called autoregressive polynomial of order p;
- $\theta(B) = 1 - \theta_1 B - \ldots - \theta_q B^q$ is the so–called moving average polynomial of order q;
- $\nabla^d \equiv (1-B)^d$ is the arithmetic difference operator of order d;
- ε_t is the stochastic disturbance.

In our case, seasonality is out of context, as annual data will be used. Hence, a non-seasonal ARIMA model will be sought for. Moreover, $b_1, \ldots, b_n$, as well as $C_t'\eta$ in equation (2) will all be set equal to zero.

3. DATA AND SOFTWARE-COMPUTATIONAL DETAILS

In this work, we use the HS multiple-component stochastic mortality model in order to model the mortality dynamics. By utilizing mortality models, we estimate the death rates and, in turn, the mortality trends in terms of time series, which reveal the behaviour of mortality over time. In the family of age-period-cohort stochastic mortality models the dynamics of mortality are driven by the period and the cohort indices. The data used, in order to estimate the time series of the period and cohort indices, consist of the number of deaths, $D_{t,x}$, and the corresponding central

exposures to risk, $E_{t,x}$, which are defined in rectangular arrangement (t, x) over a unit range of individual calendar years t $(t_1, \dots, t_n)$, and individual ages, x, last birthday $(x_1, \dots, x_n)$. Thus, we calculate the crude (unsmoothed) central death rate for any age x and calendar year t as $m_{t,x} = D_{t,x}/E_{t,x}$. $E_{t,x}$ is usually approximated by an estimate of the population aged x last birthday in the middle of the calendar year t or by an estimate of the average population aged x last birthday of the beginning and the end of the calendar year t. We model the number of deaths as independent Poisson realizations; that is, $D_{t,x}$ follow Poisson distribution with mean $E_{t,x} \cdot m_{t,x}$ (Brillinger (1986), Brouhns et al. (2002)).

Hatzopoulos and Sagianou (2020) proposed a dynamic multiple-component model that includes δ_1 age-period and δ_2 age-cohort effects. The HS model can be represented by the following generic formula:

$$\log(\widetilde{m}_{t,x}) = a_x + \sum_{i=1}^{\delta_1} \beta_x^{(i)} \kappa_t^{(i)} + \sum_{j=1}^{\delta_2} \beta_x^{c(j)} \gamma_c^{(j)} \qquad (4)$$

The term a_x reflects the main age profile of mortality by age, $\beta_x^{(i)}$ and $\beta_x^{c(j)}$ represent the age effect for each period and cohort component, respectively. The terms $\kappa_t^{(i)}$ reflect period-related effects and determine the mortality trend. The terms $\gamma_c^{(j)}$ represent the cohort-related effects, where $c = t - x$. The parameters δ_1 (>=1) and δ_2 (>=0) are indices for the number of period and cohort components included in the model structure, respectively. The number of period and cohort components vary depending on the experimental dataset, i.e., the intrinsic mortality peculiarities of the examined population in a given time frame. For the England and Wales dataset, for the period 1841-2006, δ_1 = 5 and δ_2 = 2 and for the period 1961-2006, δ_1 = 4 and δ_2 = 1. For full details of the estimation methodology, see Hatzopoulos and Sagianou (2020).

Therefore, these κ values must be projected. These period indices $\kappa_t^{(i)}$ reveal the mortality trends of unique age clusters and can be used by a time series analysis technique in order to forecast future mortality trends.

In this spirit, the approach adopted in this paper is the traditional two-stage process: firstly, we fit the stochastic mortality model in order to estimate κ values (see Hatzopoulos and Sagianou (2020)) and then we fit a projection model to the estimated κ values for forecasting.

Therefore, considering the aforementioned and according to Hatzopoulos and Sagianou (2020) results, the dataset for the time series analysis consists of nine annual time series of period indices $\kappa_t^{(i)}$ for England and Wales dataset, of which five are “long” time series, while four are “short” time series. The long time series data cover the period from 1841 to 2016 and consist of one hundred and seventy-six (176) observations. The short time series cover the period from 1961 to 2016 (55

observations). The graphical representations of the nine time series are shown in the Appendix.

To assess the effect of statistical preadjustments on forecasts two statistical software approaches will be employed, namely the AUTOARIMA command and its extensions of the well-known programming software "R" and the module TRAMO of the TSW statistical package. TSW stands for TRAMO-SEATS for Windows, a Windows version of the DOS programmes TRAMO (Time Series Regression with ARIMA Noise, Missing Observations and Outliers) and SEATS (Signal Extraction in ARIMA Time Series), and it is a software for applied time series analysis specializing, amongst others, on time series preadjustments, ARIMA modelling, model-based seasonal adjustment, and forecasting (Gomez and Maravall, 1996). TSW routines are also incorporated in other widely used econometric software products such as EVIEWS.

The AUTOARIMA command of "R" allows for the automatic selection of an ARIMA model. Moreover, forecasts based on the selected model may be obtained. On the other hand, TRAMO pre-tests for time series transformation to tackle with variance non-stationarity. Moreover, it offers several options for the treatment of outliers within the frames of the more general preadjustment procedure known as "linearization". However, transformation-wise, TSW allows for the logarithmic transformation as the only option. For this reason and for a wider selection of transformations (e.g., the square root transformation) the statistical approach and recommendations suggested by Milionis will also be used (Milionis, (2003); Milionis, 2004; Milionis and Galanopoulos, 2019). It is further noted that TSW identifies three types of outliers utilizing the Chen and Liu (1993) computational approach. These three types of outliers are discriminated according to their effect in a time series as follows: i) Additive Outliers (AO), which affect only a single observation of time series, ii) Transitory Changes (TC), in which the effect of one observation that is extremely high or low is not extinguished in the next observations but damps out gradually over a few periods, iii) Level Shifts (LS), which reflect a major change in the stochastic process and have a permanent effect as all observations subsequent to the outlier move to a new level. Outlier parameters are estimated together with the ARIMA parameters using maximum likelihood estimation. An observation is considered as an outlier according to the critical value of the appropriate statistic $\boldsymbol{\tau}$ (see Gomez and Maravall, 1996, Milionis and Galanopoulos, 2020 for details). As theory cannot predict the critical value of $\boldsymbol{\tau}$, a usual practice is to relate the critical value of $\boldsymbol{\tau}$ to the length of a time series (Fischer and Planas, 2000). In the present study the default options of TSW for the identification of outliers will be used.

4. RESULTS AND DISCUSSION

4.1 Data transformation

Firstly, it is important to note that the effect of a transformation is meant both in a direct and an indirect way. The direct way is self-explanatory and refers to the

transformation itself. The indirect way refers to the influence of the transformation on outlier detection. Indeed, regarding the later, it has been shown that data transformation affects the number and the character of outliers in a time series (Milionis, 2004; Milionis and Galanopoulos, 2019).

Table 1 presents the results on the decision about, transforming or not, the original time series data using both TSW and Milionis approaches. From the analysis of the nine time series examined, both methodologies suggest no transformation of the original data in seven cases. In two cases time series need to be transformed. Both methodologies suggest the log transformation.

Table 1. *Decision about data transformation*

Time series	**TSW**	**Milionis (2004)**
E&W L.KT1	Levels	Levels
E&W L.KT2	Levels	Levels
E&W L.KT3	Logs	Logs
E&W L.KT4	Logs	Logs
E&W L.KT5	Levels	Levels
E&W S.KT1	Levels	Levels
E&W S.KT2	Levels	Levels
E&W S.KT3	Levels	Levels
E&W S.KT4	Levels	Levels

For further analysis on statistical forecasting three different approaches will be examined. More specifically, these three approaches are the following: (a) Forecasts based on the random walk with drift model, which, for simplicity, is a widely used model in relevant actuarial studies, as already mentioned, and will be used as a benchmark; (b) Forecasts using the automatic selection and forecasting procedure of the programming software "R", specifically the "AUTOARIMA" command, as in Hatzopoulos and Sagianou (2020); (c) Forecasts based on ARIMA models after statistical preadjustments. The later implies Variance Reduction and will be called "VR" forecasts henceforth.

Table 2 shows the type of ARIMA models of methodologies (b) and (c). It is noted that the differences in the ARIMA models for those time series where no transformation was needed, should be attributed to the existence of outliers adjusted by linearization, the absence or presence of a drift, and possible differences in the algorithms of software (R and TSW).

It should be remarked that in the results of Table 2 there are only two cases with AUTOARIMA and three cases with VP where a statistically significant drift was found. The absence of a drift, which is atypical for such kind of actuarial series, obviously weakens the forecasting capacity especially for long term forecasting.

Table 2. *ARIMA models*

Time series	"Autoarima"	VR
E&W L.KT1	(0,1,0) WITH MEAN	(0,1,1) WITH MEAN
E&W L.KT2	(0,1,3) WITH MEAN	(3,1,0) WITHOUT MEAN
E&W L.KT3	(3,0,0) WITHOUT MEAN	(0,1,0) WITHOUT MEAN
E&W L.KT4	(1,0,3) WITHOUT MEAN	(0,1,1) WITHOUT MEAN
E&W L.KT5	(1,1,4) WITHOUT MEAN	(1,2,1) WITHOUT MEAN
E&W S.KT1	(0,2,2) WITHOUT MEAN	(1,1,0) WITH MEAN
E&W S.KT2	(0,1,0) WITHOUT MEAN	(0,1,0) WITHOUT MEAN
E&W S.KT3	(0,1,0) WITHOUT MEAN	(0,1,0) WITHOUT MEAN
E&W S.KT4	(0,1,0) WITHOUT MEAN	(1,0,0) WITH MEAN

4.2 The effect of "Linearization"

Outliers are major changes in values that stand out in time series. Inspecting visually the time series included in our dataset (*E&W L.KT3* and *E&W L.KT4* in natural logarithms), it is obvious that in some cases the increased variance can be attributed to outliers. TSW was used (in default options) to identify outliers in all time series.

The type of outlier and the order of observation in which they appear are presented in Table 3. The first number refers to the order of observation followed by the type of outlier. By way of an example 9 AO in E&W L.KT1 time series (see first row of the Table 3) shows that the order (9) of the observations (years) of detected outlier is the year 1849, as the initial observation is the year 1841, and the type of outlier (AO) is Additive Outlier.

4.3 The combined effect of Data Transformation and Linearization

To evaluate the combined effect of data transformation and linearization on the quality of point forecasts some typical statistics will be used. Firstly, the Mean Square Forecast Error (MSFE) which measures the average squared difference between the forecasting

Table 3. *Detected outliers and their type*

Time series	Temporal Position and Type of outliers
E&W L.KT1	9 AO, 74 LS, 75 LS, 78 TC, 79 LS, 100 LS, 106 LS
E&W L.KT2	80 LS, 100 AO
E&W L.KT3	74 LS, 79 LS, 80 LS, 89 AO, 100 LS, 102 LS, 106 LS, 111 LS, 113 TC, 116 TC, 118 AO, 124 TC, 128 TC, 130 TC, 133 LS
E&W L.KT4	9 TC, 18 AO, 74 LS, 79 LS, 88 AO, 95 AO, 100 LS, 106 LS
E&W L.KT5	9 AO, 18 AO, 23 TC, 50 TC, 74 TC, 77 TC, 78 AO, 104 AO, 157 AO
E&W S.KT1	NO OUTLIERS DETECTED
E&W S.KT2	NO OUTLIERS DETECTED
E&W S.KT3	37 AO
E&W S.KT4	NO OUTLIERS DETECTED

values (F_t) and the actual values (A_t), i.e. $MSFE = \frac{1}{n}\sum_{t=1}^{n}(A_t - F_t)^2$. It is well known that optimal forecasts are those with the minimum MSFE (Hamilton, 1994). Auxiliary the following statistics will also be used:

i) the Mean Absolute Percentage Error (MAPE) statistic given by:
$MAPE = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$, and

ii) the Mean Absolute Error (MAE) statistic given by:
$MAE = \frac{1}{n}\sum_{t=1}^{n}|A_t - F_t|$.

Furthermore, as far as interval forecasts are concerned, the forecast standard error will be utilized.

In addition, the Akaike Information Criterion (AIC) will be used as a probabilistic statistical measure that attempts to quantify the model performance on the training dataset in conjunction with the complexity of the model.

Best forecast will obviously be perceived the one with the minimum value of the each time utilized statistic from the ones mentioned above.

Table 4 presents the number of best forecasts, in terms of minimization of the corresponding statistic, when VR model is compared with the RWD model (Table is read as follows: for each statistic, in the second and third column the cases with the minimum value of the statistic (i.e., the best forecasts) out of the total number of cases (i.e. the nine time series of the dataset) are presented).

From these results it is apparent that forecasts are better in every single case in terms of the width of the forecast confidence interval with VR methodology. Also, VR methodology is superior based on the minimum value of the Akaike information

criterion. Point forecasts with VR methodology are slightly better in terms of all three statistics (MSFE, MAPE, MAE).

Table 4. *Summary table - Number of best forecasts (VR versus RWD)*

Point Forecasts	**RWD**	**VR**
MSFE	3/9	6/9
MAPE	4/9	5/9
MAE	3/9	6/9
AIC	1/9	8/9
Interval Forecasts	**RWD**	**VR**
Forecast Standard Error (SE)	0/9	9/9

The results of the examination of the forecasting performance between VR methodology and "Autoarima" are presented in Table 5.

Table 5. *Summary table - Number of best forecasts (VR versus "Autoarima")*

Point Forecasts	**"Autoarima" with further analysis in TSW**	**VR**
MSFE	3/9	6/9
MAPE	4/9	5/9
MAE	3/9	6/9
AIC	2.5/9	6.5/9
Interval Forecasts	**"Autoarima" with analysis further in TSW**	**VR**
Forecast Standard Error (SE)	1.5/9	7.5/9

Table 5 is read in the same manner as Table 4, explaining further that when the calculated values of a statistic are found to be equal, then the arithmetic value 0.5 is assigned in both methodologies. For instance, the AIC values 2.5/9 and 6.5/9 of the fourth row of the Table 5 indicate that in two out of the nine time series the corresponding statistic value is minimum with the AUTOARIMA methodology, in six out of the nine time series the corresponding statistic value is minimum with TSW methodology, and in one time series the estimated statistic value is equal in both methodologies.

From the results of Table 5 it is seen that the VR methodology outperforms "Autoarima" in terms of the interval forecasts and is better in terms of the Akaike information criterion. However, point forecasts with the VR methodology are only slightly better in terms of the three statistics (MSFE, MAE and MAPE).

4.4 An Ad-Hoc Evaluation of the overall Models' Forecasting Performance

The skill of a forecast can be assessed by comparing the relative proximity of both the forecast and a benchmark to the observations. The presence of a benchmark makes it easier to compare approaches and for this reason a benchmark is proposed to establish a common ground for comparison. In the present case an obvious benchmark is the Random Walk Model with Drift (RWD) as already mentioned.

A crude, yet very simple and transparent ad-hoc forecasting evaluation for both point and interval forecasts will be used. More specifically, for the point forecasts for each time series and for each model an arithmetic value is assigned in ascending order based on the corresponding value of the MSFE statistic (i.e., 1 for the best (minimum) MSFE value, 2 for the second best MSFE value, 3 for the worst (maximum) MSFE value). Then, adding up the arithmetic values for all series for a particular model their sum will represent the performance of the model. Models will be ranked according to the value of the corresponding sum. Apparently, the model with the lowest sum will be considered as the best one. For interval forecasts the same procedure will be followed replacing the value of the MSFE statistic with the value of the corresponding standard error around point forecasts. The results are presented in Tables 6 and 7.

Table 6. *Ranking of forecasting performance according to MSFE (points forecasts)*

Time series	RWD	"Autoarima"	VR
E&W L.KT1	*1.5*	*1.5*	*3*
E&W L.KT2	*1*	*3*	*2*
E&W L.KT3	*2*	*3*	*1*
E&W L.KT4	*2*	*3*	*1*
E&W L.KT5	*2*	*3*	*1*
E&W S.KT1	*3*	*1*	*2*
E&W S.KT2	*1*	*2.5*	*2.5*
E&W S.KT3	*3*	*1.5*	*1.5*
E&W S.KT4	*2*	*3*	*1*
Total	*17.5*	*21.5*	*15*

From the results of Tables 6 it is evident that the performance of VR methodology for point forecast is better than that of RWD model and "Autoarima". It is noteworthy that the RWD model outperforms the "Autoarima" one.

As far as interval forecasts are concerned, the results of Table 7 show that the VR methodology has a clearly better performance than RWD model and "Autoarima". In this case "Autoarima" clearly outperforms RWD model.

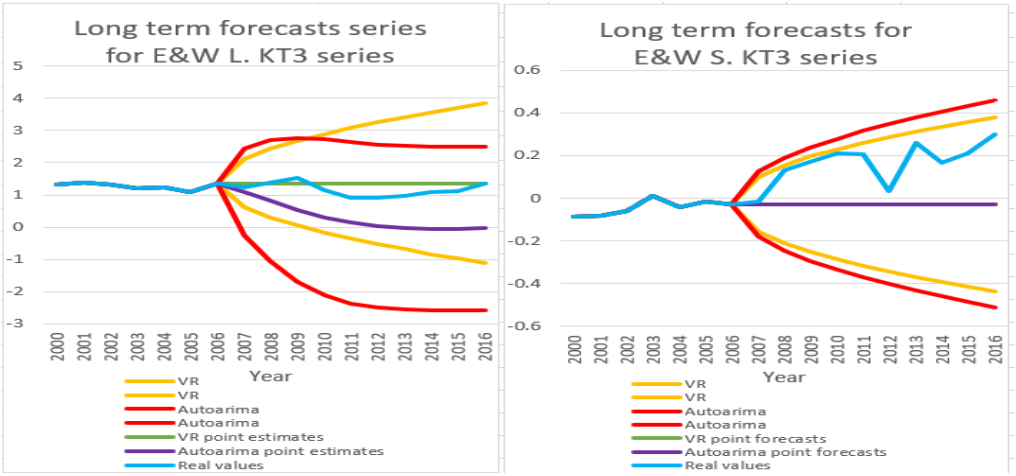
It is worthy to present some representative cases in more detail. This is done with the aid of Figure 1. More specifically, the left figure refers to the case of series in which the VR methodology deploys its full strength, as the series not only is log-transformed

Table 7. *Ranking of forecasting performance according to SE (intervals forecasts)*

Time series	RWD	"Autoarima"	VR
E&W L.KT1	*2.5*	*2.5*	*1*
E&W L.KT2	*3*	*2*	*1*
E&W L.KT3	*3*	*2*	*1*
E&W L.KT4	*3*	*2*	*1*
E&W L.KT5	*3*	*2*	*1*
E&W S.KT1	*3*	*1*	*2*
E&W S.KT2	*3*	*1.5*	*1.5*
E&W S.KT3	*3*	*2*	*1*
E&W S.KT4	*3*	*2*	*1*
Total	*26.5*	*17*	*10.5*

but also there exist several outliers (see Table 3). As is evident from left figure, with the method of VR there is a substantial reduction in forecasts confidence interval and a clear improvement in point forecasts.

Figure 1. *Forecasts and Confidence intervals with both methods for the series E&W L.KT3 and E&W S.KT3*



The right figure shows the point, as well as the interval forecasts, for the series E&W S.KT3. As presented in Table 2 the stochastic model with both methods is a simple random walk without drift. Therefore, both methods give exactly the same point forecasts, which are actually trivial, as they are all equal to the last observation. However, as presented in Table 3, an additive outlier is detected in the 37th observation with the VR method. Even in this case, forecasts quality is improved with the VR methodology. It is exactly due to the detected outlier that the forecast confidence interval is reduced with the VR method. As explained, this reduction is very important in actuarial science.

5. SUMMARY AND CONCLUSIONS

In this work we examined the effect of statistical preadjustments (data transformation and linearization) on the quality of time series forecasts of mortality rate data. It was found that there is a substantial improvement in interval forecasts which on average are shortened by approximately 35.4% when comparing VR and RWD and 20.4% when comparing VR and "Autoarima" (detailed results are not presented but are available by the authors on request). Moreover, there was a less clear improvement in point forecasts.

The above statistical findings have important implication for the actuarial science. More specifically, the improvement in interval forecasts can significantly affect the Solvency Capital Requirement, and subsequently the Solvency Ratio for a pension fund. Such an improvement might put some pension providers at a competitive advantage as they have less capital locked in their liabilities.

As a prospect for further research, we intend to explore the effect of statistical preadjustments to the financial impact on Solvency Capital Requirement, under different model structures and forecast methods. As has been noted previously, the most useful tool for investigating uncertainty over longevity risk is a stochastic mortality projection model. Since, there is a wide choice of such models in the literature, the choice of model can lead to material changes in the best-estimate reserves, while even within a model family there can be major differences (Richards & Currie 2009). For those models we aim to study the uncertainty over future mortality rates, which is measured as the variance of the mortality forecast values. In particular, we will investigate their respective contributions to the capital requirements for longevity trend risk. Our investigation will be based on the Hatzopoulos and Sagianou (2020) model, but the basic conclusions will apply to any model which uses time-series methods to project a mortality index. We will also quantify the respective contributions to capital requirements using VaR calculations.

ΠΕΡΙΛΗΨΗ

Η ακριβής πρόβλεψη ρυθμών θνησιμότητας παίζει καθοριστικό ρόλο στη διαχείριση του κινδύνου μακροζωίας, ενός εκ των σημαντικότερων κινδύνων στον αναλογιστικό κλάδο. Στο παρόν, μελετάται αφενός η επίδραση της «γραμμικοποίησης» χρονοσειρών, όταν υφίστανται αιτίες που διαταράσσουν τη στοχαστική διαδικασία, και αφετέρου ενδεχόμενων μετασχηματισμών των αρχικών χρονοσειρών, στην ποιότητα προβλεπόμενων ρυθμών θνησιμότητας. Οι χρονοσειρές που εξετάζονται είναι οι δείκτες «περιόδου» (period indices) που αποκαλύπτουν την τάση θνησιμότητας στα δεδομένα Αγγλίας-Ουαλίας. Από τα εμπειρικά αποτελέσματα προκύπτει οριακή μεν βελτίωση στις σημειακές προβλέψεις, αλλά σαφώς σημαντικότερη βελτίωση στο διάστημα εμπιστοσύνης των προβλέψεων. Στην πράξη, αυτή η βελτίωση του διαστήματος εμπιστοσύνης των προβλέψεων μπορεί να επανακαθορίσει τις κεφαλαιακές απαιτήσεις φερεγγυότητας, κατά συνέπεια το συντελεστή φερεγγυότητας ενός συνταξιοδοτικού ταμείου, δίνοντας ανταγωνιστικό πλεονέκτημα σε παρόχους συντάξεων, καθώς απαιτείται η δέσμευση μικρότερων κεφαλαίων για το αποθεματικό.

REFERENCES

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, **26**, 211-243.

Brillinger, D. (1986). The natural variability of vital rates and associated statistics. *Biometrics*, **42**, 693-734.

Brouhns, N., Denuit, M. and Vermunt, J.K. (2002). A poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, **31**, 373-393.

Cairns, A. J., 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**, 687–718.

Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D. and Khalaf-Allah, M. (2011). Mortality density forecasts: An analysis of six stochastic mortality models. *Insurance: Mathematics and Economics*, **48**, 355–367.

Chen, C. and Liu, L.-M. (1993). Forecasting Time Series With Outliers. *Journal of Forecasting*, **12**, 13-35.

Currie, I. (2006). *Smoothing and forecasting mortality rates with P-splines.* Talk given at the Institute of Actuaries, June 2006.

Fischer, B. and Planas, C. (2000). Large scale fitting of regression models with ARIMA errors. *Journal of Official Statistics*, **16**, 173-184.

Gomez, V. and Maravall, A. (1996). Programmes SEATS and TRAMO: instructions for the User'. *Bank of Spain Working Paper* 9628.

Hamilton, J. (1994). *Time Series Analysis*, Princeton University Press, New Jersey.

Hatzopoulos, P. and Haberman, S. (2011). A dynamic parameterization modelling for the age-period-cohort mortality. *Insurance: Mathematics and Economics*, **49**, 155–174.

Hatzopoulos, P. and Sagianou, A. (2020). Introducing and Evaluating a New Multiple-Component Stochastic Mortality Model. *North American Actuarial Journal*, **24**, 393-445.

Hyndman, R. and Ullah, M. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, **51**, 4942–4956.

Jarque C. and Bera, A. (1980). Efficient tests for normality homoscedasticity and serial independence of regression residuals. *Economics Letters*, **6**, 255-259.

Kaiser, R. and Maravall, A. (2001). *Measuring business cycles in economic time series*, Springer.

Kleinow, T. and Richards, SJ. (2017). Parameter risk in time-series mortality forecasts. *Scandinavian Actuarial Journal*, **2017**, 804-828.

Lee, R. D. and Carter, L. R. (1992). Modelling and forecasting U.S. mortality. *Journal of the American Statistical Association*, **87**, 659–671.

Lee, R.D. and Miller, T. (2001). Evaluating the performance of the Lee–Carter method for forecasting mortality. *Demography*, **38**, 537–549.

Lovász, E. (2011). Analysis of Finnish and Swedish mortality data with stochastic mortality models. *European Actuarial Journal*, **1**, 259–289.

Milionis, A. (2003). Modelling economic time series in the presence of variance non-stationarity. *Bank of Greece Working Paper* No 7.

Milionis, A. (2004). The importance of variance stationarity in economic time series modelling; A practical approach. *Applied Financial Economics*, **14**, 265-278.

Milionis, A. and Galanopoulos, N. (2019). Forecasting economic time series in the presence of variance instability and outliers. *Theoretic Economic Letters*, **9**, 2940-2964.

Milionis, A. and Galanopoulos, N. (2020). A study of the effect of data transformation and "linearization" on time series forecasts. A practical approach. *Bank of Greece Working Paper* No 280.

Pitacco, E., Denuit, M., Haberman, S. and Olivieri, A. (2009). *Modelling Longevity Dynamics for Pensions and Annuity Business,* Oxford University Press.

Plat, R. (2009). On stochastic mortality modelling. *Insurance: Mathematics and Economics*, **45**, 393–404.

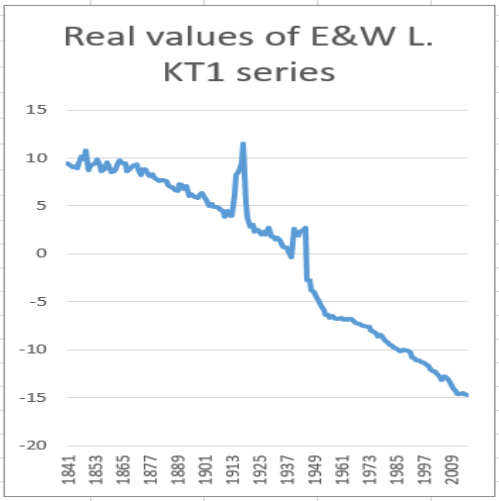
Renshaw, A. and Haberman, S. (2006). A cohort-based extension to the Lee-Carter model for mortality reduction factors. *Insurance: Mathematics and Economics*, **38**, 556–570.

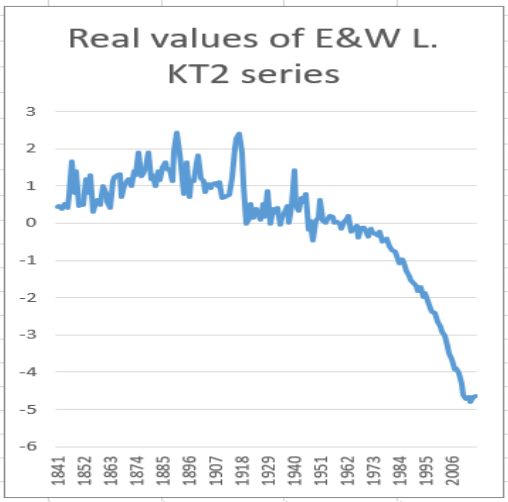
Richards, S. J., and Currie, I. D. (2009). Longevity risk and annuity pricing with the Lee-Carter model. *British Actuarial Journal*, **15**, 317-343.

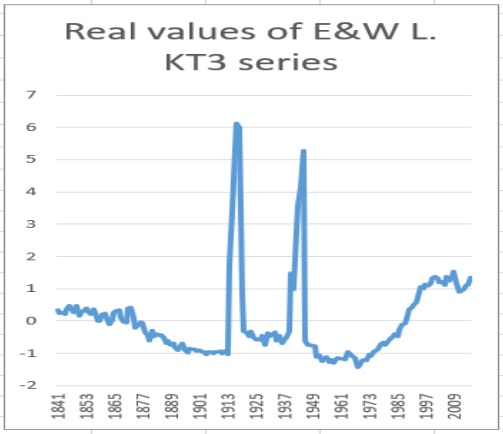
Solvency II (Directive 2009/138/EC). Directive 2009/138/EC of the European Parliament and of the Council of 25 November 2009 on the taking-up and pursuit of the business of Insurance and Reinsurance (Solvency II) (recast) (Text with EEA relevance).

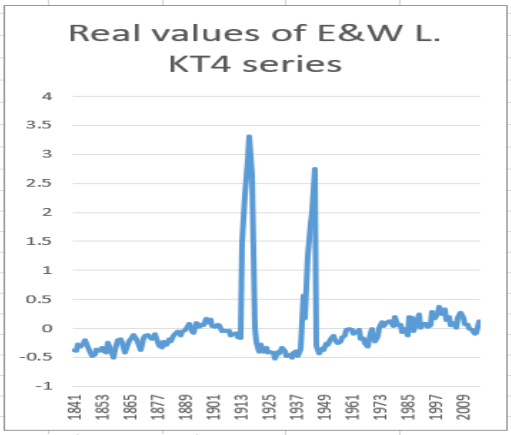
Villegas, A., Millossovich, P. and Kaishev, V. (2018). StMoMo: An R Package for Stochastic Mortality Modelling. *Journal of Statistical Software*, **84**, 1-38.

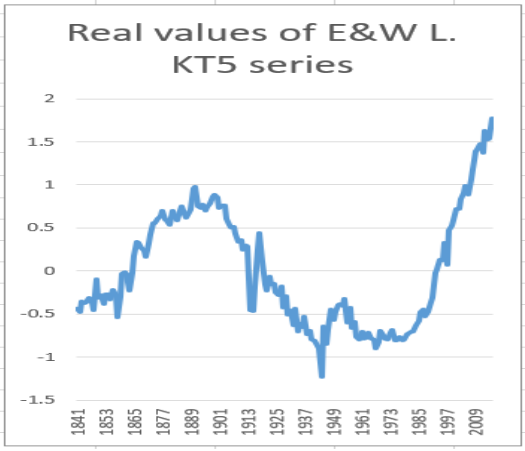
APPENDIX

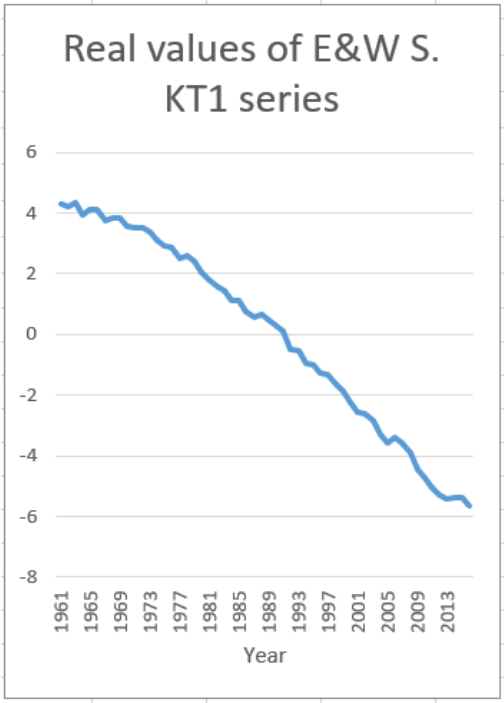

Real values of E&W L.
KT1 series
15
10
5
0
-5
-10
-15
-20
1841
1853
1865
1877
1889
1901
1913
1925
1937
1949
1961
1973
1985
1997
2009

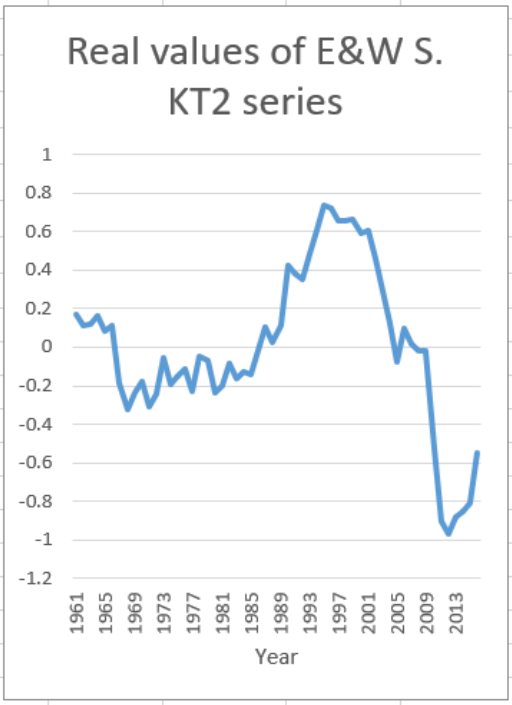

Real values of E&W L.
KT2 series
3
2
1
0
-1
-2
-3
-4
-5
-6
1841
1852
1863
1874
1885
1896
1907
1918
1929
1940
1951
1962
1973
1984
1995
2006

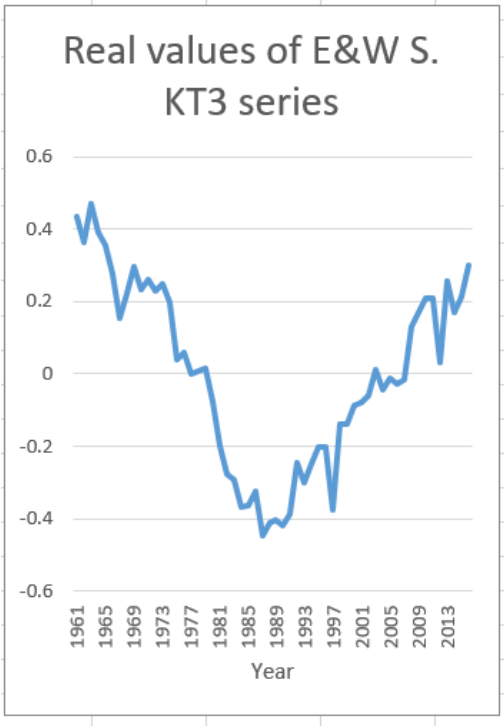

Real values of E&W L.
KT3 series
7
6
5
4
3
2
1
0
-1
-2
1841
1853
1865
1877
1889
1901
1913
1925
1937
1949
1961
1973
1985
1997
2009

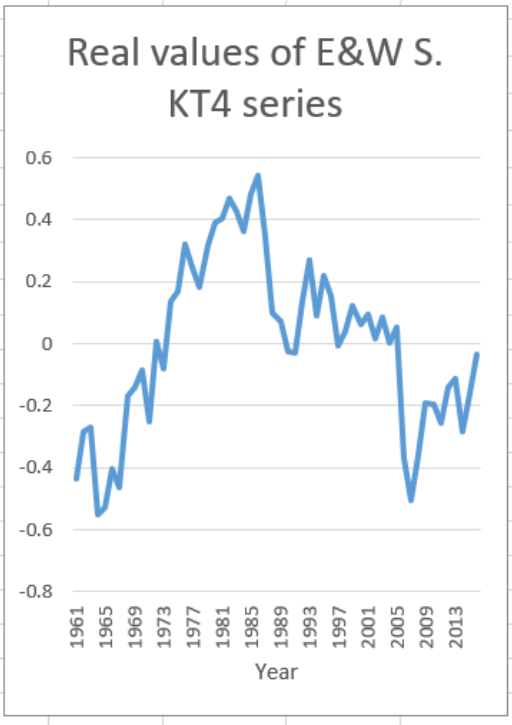

Real values of E&W L.
KT4 series
4
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
-1
1841
1853
1865
1877
1889
1901
1913
1925
1937
1949
1961
1973
1985
1997
2009


Real values of E&W L.
KT5 series
2
1.5
1
0.5
0
-0.5
-1
-1.5
1841
1853
1865
1877
1889
1901
1913
1925
1937
1949
1961
1973
1985
1997
2009


Real values of E&W S. KT1 series
6
4
2
0
-2
-4
-6
-8
1961
1965
1969
1973
1977
1981
1985
1989
1993
1997
2001
2005
2009
2013
Year


Real values of E&W S. KT2 series
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
1961
1965
1969
1973
1977
1981
1985
1989
1993
1997
2001
2005
2009
2013
Year


Real values of E&W S. KT3 series
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
1961
1965
1969
1973
1977
1981
1985
1989
1993
1997
2001
2005
2009
2013
Year


Real values of E&W S. KT4 series
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
1961
1965
1969
1973
1977
1981
1985
1989
1993
1997
2001
2005
2009
2013
Year

Greek Statistical Institute
Proceedings of the 33rd Panhellenic Statistics Conference (2021), pp.313-329

ESTIMATION OF CAUSALITY IN DISCRETE TIME SERIES USING PARTIAL MUTUAL INFORMATION FROM MIXED EMBEDDING

Elsa Siggiridou, Maria Papapetrou and Dimitris Kugiumtzis
Department of Electrical and Computer Engineering, Faculty of Engineering, Aristotle University of Thessaloniki
esingiri@auth.gr, mariapap@auth.gr, dkugiu@auth.gr

ABSTRACT

In the analysis of multivariate time series, different methods have been developed for the estimation of causal relationships among the observed variables. A subcategory of these methods employs dimension reduction to estimate direct causal effects, used to determine the connections of the complex network that forms the structure of the underlying dynamical system or stochastic process. The partial conditional mutual information from mixed embedding (PMIME) using conditional mutual information (CMI) implements such an approach and is found to be appropriate for direct causality estimation from continuous-valued time series. In this study, the interest is in discrete-valued multivariate time series, and we adapt appropriately the PMIME, called discrete PMIME (DPMIME). Appropriate estimation of discrete probability distributions and CMI for discrete variables is implemented in DPMIME. Further, asymptotic distribution of the estimated CMI is derived allowing for a parametric significance test for CMI in DPMIME, whereas in PMIME there is only resampling significance test for CMI. The parametric significance test for CMI in the progressive algorithm of DPMIME is compared favorably to the corresponding resampling significance test. Monte Carlo simulations on multivariate symbol sequences derived by discretization of continuous-valued multivariate time series showed that the accuracy of DPMIME in the estimation of direct causality converges with the time series length to the accuracy of PMIME.

Keywords: Granger causality, conditional mutual information, mixed embedding, discrete-valued time series, symbol sequences

1. INTRODUCTION

The estimation of interaction of observed variables in multivariate time series, termed as Granger causality (Granger 1969), has turned out to a challenging and yet prominent research topic in many domains of science and engineering. Granger causality has a dominant role for the identification of directional interactions among the observed variables on the basis solely on their time series data and it has been in

the focus in diverse fields ranging from economics (Billio et al 2012) and climatology (Dijkstra et al 2019) to physiology and neuroscience (Porta and Faes 2016).

Granger causality is related to the study of complex systems (Fieguth 2017, Thurner et al 2018), where each observable represents a constituent subsystem and corresponds to a node of the complex network, and the directed connection between the nodes is given by the estimated Granger causality (Newman 2010, Kirchgässner et al 2013). To-date the study of Granger causality estimation regards complex systems observed by multivariate continuous-valued time series. The main contribution of this work is to extent the study of Granger causality estimation to multivariate discrete-valued time series. Any type of discrete-valued time series can be considered (e.g., binary, categorical, count etc.) because the method relies on information measures on symbol data, and thus, there is no requirement on scale or order of the data. Further, we assume that the set of discrete values S has a relatively small cardinality (time series of few distinct values), so that the underlying Markov chain can be assumed irreducible, i.e., there is a positive transition probability from any symbol i of S to any symbol j of S in finite number of steps, and subsequently stationary (Weiss 2018).

Moreover, it extends the setting of bivariate or trivariate time series to multivariate time series where the curse of dimensionality affects the estimation of direct Granger causality. By direct Granger causality is meant the direct causal effect one observed variable (representing possibly a subsystem) has on another variable (or subsystem) accounting for the presence of the other observed variables (or subsystems).

In the analysis of discrete-valued multivariate time series, there have been developed probability distribution models based on strong assumptions about the structure of the underlying system, typically assumed to be a multivariate Markov chain. Most prominent are the Poisson distribution (Fokianos et al 2009, Neumann 2011) and negative binomial distribution (Davis and Wu 2009, Christou and Fokianos 2015). In the category of autoregressive models, in analogy to the vector autoregressive models for continuous-valued time series, there are two prominent model approximations, the Pegram's autoregressive models (Song et al 2013, Angers et al 2017) and the multivariate integer-valued autoregressive models (MINAR) (Pedeli and Karlis 2013, Scotto et al 2015). These classes of models can only be used when the dimension K of the multivariate time series, i.e. the number of observed variables, is relatively small, simply because the probability distribution estimation requires exponentially larger length of time series N with the increase of K, which cannot be met in real-world applications. Another important factor that worsens the effect of curse of dimensionality is the memory of the system, i.e. the order L of the Markov chain. One approach to reduce the dimension of the parameter space is done by the mixture transition distribution (MTD) (Raftery, 1985), which restricts the initially large number of parameters assuming that there is only an effect of each of the L lag variables separately (Ching eta al 2004, Ching et al 2008, Nicolau 2014, Zhou 2017, Tank et al 2017).

Here, we follow a model-free approach and instead of attempting to build up a model for the underlying system from the discrete-valued multivariate time series, we

estimate the interactions of the lagged variables using information theory measures, mutual information (MI) and conditional mutual information (CMI), developing a method for estimating the connectivity structure of the underlying system. Recently, we have worked this approach for continuous-valued multivariate time series, termed partial mutual information from mixed embedding (PMIME) (Vlachos and Kugiumtzis 2010, Kugiumtzis 2013). We name the adaptation of this approach to the discrete-valued multivariate time series as discrete partial mutual information from mixed embedding (DPMIME).

The basic idea in PMIME and DPMIME is as follows. In the general setting, we want to explain the future of an observed variable *Y* (response) from the presence and past up to a lag *L* of all the *K* observed variables. Even for moderate *L* and *K* it is hard to estimate the direct causal effect of the lag variables, especially when instead of linear correlation measures, information theory measures are used. The proposed algorithm, DPMIME in our case of the discrete-valued setting, finds a subset of the most relevant lag variables to the future of *Y* with cardinality much lower than *LK*, the number of all candidate lag variables. To evaluate whether one variable *X* Granger causes *Y*, we simply compute the (normalized) information of the lag components of *X* found in this subset to the future of *Y*. This constitutes the DPMIME measure of Granger causality from *X* to *Y*. If there are no lag components of *X* in the subset the DPMIME is zero.

The structure of the paper is as follows. First, in Section 2, we present the proposed measure of DPMIME and a parametric and randomization significance test used as termination criterion in the building of the subset of relevant lag components. In Section 3, we assess the efficiency of the DPMIME with the randomization and parametric test and compare them with PMIME for the continuous-valued time series. Finally, in section 4, the results are discussed, and the main conclusions are drawn.

2. DISCRETE PARTIAL MUTUAL INFORMATION ON MIXED EMBEDDING

The measure of discrete partial mutual information on mixed embedding (DPMIME) is similarly defined to the partial mutual information on mixed embedding (PMIME), but while PMIME is designed for continuous-valued time series the DPMIME is designed for discrete-valued time series. Thus, DPMIME inherits the main properties of PMIME. It is an information-based measure, and thus nonlinear and parameter-free, and by selecting only the most relevant lag variables for explaining the response it addresses the curse of dimensionality and therefore it can be reliably applied to time series of many variables (Kugiumtzis, 2013).

2.1 ITERATIVE ALGORITHM FOR THE COMPUTATION OF DPMIME

Let $\{x_{1,t}, x_{2,t}, \ldots, x_{K,t}\}$, $t = 1, 2, \ldots, n$, be the observations of a multivariate stochastic process $\{X_{1,t}, X_{2,t}, \ldots, X_{K,t}\}$ and denote the quantity (variable) each process $\{X_{i,t}\}$ refers to as X_i for simplicity, whereas when the variable of the

process is given with respect to time *t*, it is denoted $X_{i,t}$. The stochastic process is typically a multivariate Markov chain. The discrete variables can be nominal or ordinal and for convenience hereafter we refer to the data as multivariate symbol sequence.

We are interested in defining a measure for the direct causality from *X* to *Y*, where *X* and *Y* are any of the *K* observed discrete variables. For a sufficiently large maximum lag *L*, we formulate the set W_t of candidate lag variables that may have information explaining the response *Y* at one time ahead, Y_{t+1}. The set W_t has *KL* components, $X_{i,t-\tau}$, $i=1,\ldots,K$, $\tau=0,\ldots,L-1$. The algorithm DPMIME aims to build up progressively a so-called mixed embedding vector, i.e., a subset $\mathbf{w}_t$ of W_t of the most informative lag variables explaining Y_{t+1}.

In the first step, the first lag variable to enter $\mathbf{w}_t$ is the one that maximizes the MI with Y_{t+1},

$$w_1 = \arg\max_{w_t \in W_t} I(Y_{t+1}; w_t) \tag{1}$$

and $\mathbf{w}_t = \mathbf{w}_t^1 = [w_1]$ (the superscript denotes the iteration, i.e., the cardinality of the set). The MI for two variables *X* and *Y* (with regard to (1) *X* is Y_{t+1} and *Y* is w_t) defined in terms of entropy and probability mass functions (pmfs) as (Cover and Thomas, 1991)

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = \sum_{x,y} p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x)\,p_Y(y)},$$

where *H*(*X*) is the entropy of *X*, the sum is over all values of *x* and *y* of *X* and *Y*, $p_{X,Y}(x,y)$ is the joint pmf of (*X*, *Y*) and $p_X(x)$ is the pmf of *X*. The pmfs are estimated by the maximum likelihood estimate of the probabilities of each value or pair of values given simply by the relative frequency of occurrence in the sample (the multivariate symbol sequence).

In the subsequent steps, the conditional mutual information (CMI) instead of the MI is used to find the new component to enter $\mathbf{w}_t$, where CMI is defined as MI but now conditioning on the components already in $\mathbf{w}_t$, as explained below. Suppose that at step *j*, the *j* most relevant lag variables to Y_{t+1} are found and set in $\mathbf{w}_t = \mathbf{w}_t^j$. The next candidate component in $W_t \setminus \mathbf{w}_t^j$ (the *KL* components except the *j* components already selected) is the one that maximizes the CMI to Y_{t+1}, i.e., the MI of the candidate w_t and Y_{t+1} conditioned on the components in $\mathbf{w}_t^j$

$$w_{j+1} = \arg\max_{w_t \in W_t \setminus \mathbf{w}_t^j} I(Y_{t+1}; w_t \mid \mathbf{w}_t^j)\,. \tag{2}$$

The CMI for two variables X and Y given a third variable Z (with regard to (2) X is Y_{t+1}, Y is w_t and Z is $\mathbf{w}_t^j$) is defined in terms of entropy and pmfs as (Cover and Thomas, 1991)

$$I(X;Y|Z) = -H(X,Y,Z) + H(X,Z) + H(Y,Z) - H(Z)$$

$$= \sum_{x,y,z} p_{X,Y,Z}(x,y,z) \log \frac{p_{X,Y,Z}(x,y,z)\, p_Z(z)}{p_{X,Z}(x,z)\, p_{Y,Z}(y,z)}.$$

As for the MI, the pmfs are estimated by the maximum likelihood estimate of the probabilities of each value, pair of values or triple of values given by the relative frequency of occurrence in the multivariate symbol sequence. Note, however, that for the computation of CMI, the relative frequencies of words of length larger than two have to be computed (the length depends on the size of Z), which complicates the estimating procedure.

At each step, when the lag variable is selected, using (1) for the first step and (2) for the subsequent steps, a significance test is run for the MI in (1) or the CMI in (2). The significance test is performed assuming a known asymptotic null distribution of the test statistic (parametric test) and by forming it empirically using resampling (randomization test), and these two tests are presented in detail later in this Section. If the null hypothesis is rejected the selected component is added to the mixed embedding vector $\mathbf{w}_t$. For the step j+1, where w_{j+1} is found in (2), if the CMI $I(Y_{t+1}; w_{j+1} | \mathbf{w}_t^j)$ is found statistically significant by the parametric or resampling test the $\mathbf{w}_t$ is augmented as $\mathbf{w}_t = \mathbf{w}_t^{j+1} = [\mathbf{w}_t^j, w_{j+1}]$. Otherwise, there is no significant lag variable to add to the mixed embedding vector and the algorithm terminates giving the mixed embedding vector $\mathbf{w}_t = \mathbf{w}_t^j$.

The components of the mixed embedding vector $\mathbf{w}_t$ obtained upon termination of the algorithm are grouped in lag variables of the driving variable X, $\mathbf{w}_t^X$, the response variable Y, $\mathbf{w}_t^Y$, and all other K-2 variables, $\mathbf{w}_t^Z$, expressed as $\mathbf{w}_t = [\mathbf{w}_t^X, \mathbf{w}_t^Y, \mathbf{w}_t^Z]$. If $\mathbf{w}_t^X$ is empty, i.e. no lag variable $X_{t-\tau}$ has information to explain Y_{t+1} in view of the other lag variables, there is no direct causality from X to Y. Otherwise, we quantify the direct causality from X to Y as the proportion of the information of Y_{t+1} explained by the lag variables of X. The measure DPMIME is thus defined as

$$\text{DPMIME}_{X\to Y} = \begin{cases} 0 & \mathbf{w}_t^X = \varnothing \\ \dfrac{I\left(Y_{t+1}; \mathbf{w}_t^X \mid \mathbf{w}_t^Y, \mathbf{w}_t^Z\right)}{I\left(Y_{t+1}; \mathbf{w}_t\right)} & \text{otherwise} \end{cases} \quad (3)$$

In the following, we present the resampling test and the parametric test for the significance of the CMI of the response Y_{t+1} and the selected component w_{j+1} given the components already selected in $\mathbf{w}_t^j$, $I(Y_{t+1}; w_{j+1} \mid \mathbf{w}_t^j)$.

2.2 RANDOMIZATION TEST FOR THE SIGNIFICANCE OF CMI

Let us first denote *I*(.) the theoretical information measure (MI or CMI) and *I*(.) its estimate. First, we suppose that there is no asymptotic distribution of the estimate $I(Y_{t+1}; w_{j+1} \mid \mathbf{w}_t^j)$ under the null hypothesis H_0: $I(Y_{t+1}; w_{j+1} \mid \mathbf{w}_t^j) = 0$. We follow the procedure for the randomization test for the significance of CMI on univariate symbol sequences, as proposed in Papapetrou and Kugiumtzis (2013). The empirical distribution of $I(Y_{t+1}; w_{j+1} \mid \mathbf{w}_t^j)$ is formed by resampling on the initial sample of the variables Y_{t+1}, w_{j+1} and $\mathbf{w}_t^j$, as follows. The resampling is actually applied only to w_{j+1}. To retain both the marginal distribution and intra-dependence (autocorrelation) of w_{j+1} and destroy any inter-dependence on Y_{t+1} and $\mathbf{w}_t^j$, we shift cyclically the symbol sequence of w_{j+1} by a random step *k*. Thus, for the original symbol sequence $\{w_{j+1,1}, w_{j+1,2}, \ldots, w_{j+1,n}\}$ of w_{j+1} one randomized (surrogate) symbol sequence for the random step *k* is

$$\{w_{j+1,1}^{*1}, w_{j+1,2}^{*1}, \ldots, w_{j+1,n}^{*1}\} = \{w_{j+1,k+1}, \ldots, w_{j+1,n}, w_{j+1,1}, \ldots, w_{j+1,k}\}.$$

We derive a number *M* of such randomized symbol sequences and compute for each of them the corresponding CMI $I(Y_{t+1}; w_{j+1} \mid \mathbf{w}_t^j)$ under the H_0, denoted

$$I(Y_{t+1}; w_{j+1}^{*1} \mid \mathbf{w}_t^j), I(Y_{t+1}; w_{j+1}^{*2} \mid \mathbf{w}_t^j), \ldots, I(Y_{t+1}; w_{j+1}^{*M} \mid \mathbf{w}_t^j).$$

These M values form the empirical null distribution of $I(Y_{t+1}; w_{j+1} \mid \mathbf{w}_t^j)$. The H_0 is rejected if $I(Y_{t+1}; w_{j+1} \mid \mathbf{w}_t^j)$ on the original data is at the right tail of the empirical null distribution. To assess this we use rank ordering, where r^0 is the rank of $I(Y_{t+1}; w_{j+1} \mid \mathbf{w}_t^j)$ in the ordered list of the *M*+1 values, assuming ascending order. The *p*-value of the one-sided test is $1-(r^0-0.326)/(M+1+0.348)$ (using the correction in Yu and Huang (2001)). The resampling test is denoted as DPMIMErt.

2.3 PARAMETRIC TEST FOR THE SIGNIFICANCE OF CMI

The randomization test for the significance of CMI is considered to be the most rigorous test considering the lack of established results for an appropriate parametric null distribution. However, entropy and MI are well studied quantities and there is rich literature about the statistical properties and distribution of their estimates. In Papapetrou and Kugiumtzis (2014), the most prominent of these parametric distribution approximations are worked out for CMI under H_0, namely the Gaussian and Gamma distributions. For the Gamma null distribution, following the work in Goebel et al (2005), it turns out that $I(X,Y)$ follows approximately the Gamma distribution

$$I(X,Y) \sim \Gamma\left(\frac{1}{2}(D-1)^2, \frac{1}{n_f \ln 2}\right),$$

where n_f is the length of future vector of response and D is the number of different words of it.

Further, it follows that $I(X,Y|Z)$ is also approximately Gamma distributed

$$I(X,Y|Z) \sim \Gamma\left(\frac{D_Z}{2}(D-1)(D-1), \frac{1}{n_f \ln 2}\right)$$

where n_f is the length of future vector of response, D is the number of different words of it and D_Z is the number of different words of the selected lagged symbol sequence.

Replacing X to Y_{t+1}, Y to w_{j+1} and Z to $\mathbf{w}_t^j$, we derive the parametric approximation for the distribution of $I(Y_{t+1}; w_{j+1} | \mathbf{w}_t^j)$ and use it as null distribution for the null hypothesis H_0: $I(Y_{t+1}; w_{j+1} | \mathbf{w}_t^j) = 0$. Given that it always holds $I(Y_{t+1}; w_{j+1} | \mathbf{w}_t^j) \geq 0$, the parametric test is one-sided. We compute the *p*-value from the cumulative function of the null Gamma distribution of the observed $I(Y_{t+1}; w_{j+1} | \mathbf{w}_t^j)$, and we reject H_0 if the *p*-value is less than the nominal significance level $a = 0.05$. The parametric test is denoted as DPMIMEpt.

2.4 STATISTICAL EVALUATION OF METHOD ACCURACY

For a system of K variables there are $K(K\text{-}1)$ ordered pairs of variables to estimate causality. In the simulations of known systems, we know the true coupling pairs and thus we can compute performance indices to rate the causality measures as for their overall matching of the original connections in the network. Here, we consider the indices of specificity, sensitivity, Matthews correlation coefficient, F-measure and Hamming distance.

The sensitivity is the proportion of the true causal effects (true positives, TP) correctly identified as such, given as sens=TP/(TP+FN), where FN (false negatives) denotes the number of pairs having true causal effects but have gone undetected. The specificity is the proportion of the pairs correctly not identified as having causal effects (true negatives, TN), given as spec=TN/(TN+FP), where FP (false positives)

denotes the number of pairs found falsely to have causal effects. An ideal causality measure would give values of sensitivity and specificity at one. To weigh sensitivity and specificity collectively we consider the Matthews correlation coefficient (MCC) (Matthews, 1975) which ranges from -1 to 1. If MCC=1 there is perfect identification of the pairs of true and no causality, if MCC=-1 there is total disagreement and pairs of no causality are identified as pairs of causality and vice versa, whereas MCC at the zero level indicates random assignment of pairs to causal and non-causal effects.

Similarly, we consider the F-measure that combines precision and sensitivity. The precision, called also positive predictive value, is the number of detected true causal effects divided by the total number of detected casual effects and the F-measure (FM) ranges from 0 to 1. If FM=1 there is perfect identification of the pairs of true causality, whereas if FM=0 no true coupling is detected.

The Hamming distance is the sum of false positives (FP) and false negatives (FN). Thus, HD gets non-negative integer values bounded below by zero (perfect identification) and above by $K(K\text{-}1)$ if all pairs are misclassified.

3. SIMULATIONS

3.1 THE SIMULATION SETUP

In the simulation study, we compare the two versions of DPMIME to PMIME. We considered three simulation systems of which two are nonlinear and one is a linear stochastic process.

S1: The first system is the coupled Hénon map (Politi and Torcini ,1992) for K=3 and defined as,

$$X_{1,t+1} = 1.4 - X_{1,t}^2 + 0.3X_{1,t-1}$$
$$X_{2,t+1} = 1.4 - (0.5C(X_{1,t} + X_{3,t}) + (1 - C)X_{2,t})^2 + 0.3X_{2,t-1}$$
$$X_{3,t+1} = 1.4 - X_{3,t}^2 + 0.3X_{3,t-1}$$

The connectivity structure of **S1** is shown in Figure 1 a).

S2: The first system is the coupled Hénon map (Politi and Torcini ,1992) for K=5 and defined as,

$$X_{1,t+1} = 1.4 - X_{1,t}^2 + 0.3X_{1,t-1}$$
$$X_{2,t+1} = 1.4 - (0.5C(X_{1,t} + X_{3,t}) + (1 - C)X_{2,t})^2 + 0.3X_{2,t-1}$$
$$X_{3,t+1} = 1.4 - (0.5C(X_{2,t} + X_{4,t}) + (1 - C)X_{3,t})^2 + 0.3X_{3,t-1}$$
$$X_{4,t+1} = 1.4 - (0.5C(X_{3,t} + X_{5,t}) + (1 - C)X_{4,t})^2 + 0.3X_{4,t-1}$$
$$X_{5,t+1} = 1.4 - X_{5,t}^2 + 0.3X_{5,t-1}$$

The connectivity structure of **S2** is shown in Figure 1 b).

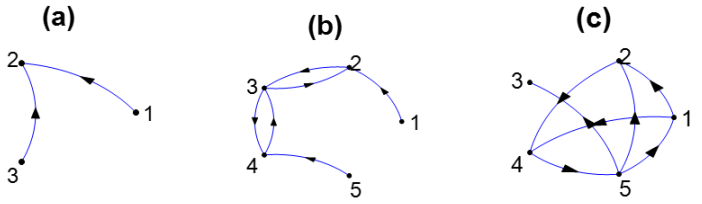
In these systems the first and the last variable in the chain of K variable drive their adjacent variable and the other variables drive the adjacent variable to their left and right. The coupling strength C is set to 0.2.

S3: A linear VAR(4) process on K=5 variables (Schelter et al, 2006) defined as,

$$X_{1,t+1} = 0.4X_{1,t} - 0.5X_{1,t-1} + 0.4X_{5,t} + \varepsilon_{1,t+1}$$
$$X_{2,t+1} = 0.4X_{2,t} - 0.3X_{1,t-3} + 0.4X_{5,t-1} + \varepsilon_{2,t+1}$$
$$X_{3,t+1} = 0.5X_{3,t} - 0.7X_{3,t-1} - 0.3X_{5,t-2} + \varepsilon_{3,t+1}$$
$$X_{4,t+1} = 0.8X_{4,t-2} + 0.4X_{1,t-1} + 0.3X_{2,t-1} + \varepsilon_{4,t+1}$$
$$X_{5,t+1} = 0.7X_{5,t} - 0.5X_{5,t-1} - 0.4X_{4,t} + \varepsilon_{5,t+1}$$

where the input errors $\varepsilon_{i,t+1}$, i=1,…,5, are uncorrelated, have unit variance and follow the Gaussian distribution. The connectivity structure of S3 is shown in Figure 1 c).

Figure 1. *The graphs of the connectivity structure of the simulated systems: (a) S1, (b) S2 and (c) S3.*



We make 100 realizations for each system and for different time series length n. For the discretization of the continuous-valued multivariate time series derived by the above systems, we used two and four symbols (ns=2,4). For example, when ns=2, for each time series the values above median are set to 0 and all other values are set to 1. The respective procedure is applied when ns=4 but using the quartiles of time series.

3.2 AN ILLUSTRATIVE EXAMPLE FOR DISCRETE PMIME

The performance of DPMIME is first illustrated with a specific example, focusing on the first two equations of system S1 for n=1024, L=5 and ns=2 using parametric test for the significance (DPMIMEpt). Table 1 shows the frequency of the inclusion of any of the 15 candidate lag terms with DPMIMEpt and PMIME in 100 Monte Carlo realizations.

Table 1. *Each cell presents the frequency of occurrence of the lag variable in the mixed embedding vector for DPMIMEpt and PMIME where the response is the first variable (columns 2 and 3) and the second variable (columns 4 and 5) of* ***S1****. The frequency is calculated over 100 realizations and for n=1024, L=5 and ns=2.*

	$X_{1,t}$		$X_{2,t}$	
	DPMIMEpt	PMIME	DPMIMEpt	PMIME
$X_{1,t}$	**100**	**100**	**10**	**1**
$X_{1,t-1}$	**100**	**100**	80	99
$X_{1,t-1}$	100	0	8	1
$X_{1,t-3}$	100	0	4	0
$X_{1,t-4}$	88	1	1	0
$X_{2,t}$	0	0	**100**	**100**
$X_{2,t-1}$	0	0	**98**	**100**
$X_{2,t-2}$	0	0	83	21
$X_{2,t-3}$	0	0	15	0
$X_{2,t-4}$	1	0	12	0
$X_{3,t}$	0	0	**13**	**5**
$X_{3,t-1}$	0	0	76	96
$X_{3,t-2}$	0	0	12	4
$X_{3,t-3}$	0	0	5	0
$X_{3,t-4}$	0	0	3	0

According to the first equation of S1, the X_1 variable is affected by the terms $X_{1,t-1}$ and $X_{1,t-2}$ which are found with 100% from both algorithms (DPMIMEpt and PMIME), with the difference that in DPMIMEpt the terms $X_{1,t-3}$, $X_{1,t-4}$ and $X_{1,t-5}$ are also selected with percentage, but since these lag terms are of the true causal variable their detection by DPMIMEpt does not affect the estimation of causality. The lag terms of the non-causal to X_1 variables have percentage at zero or one.

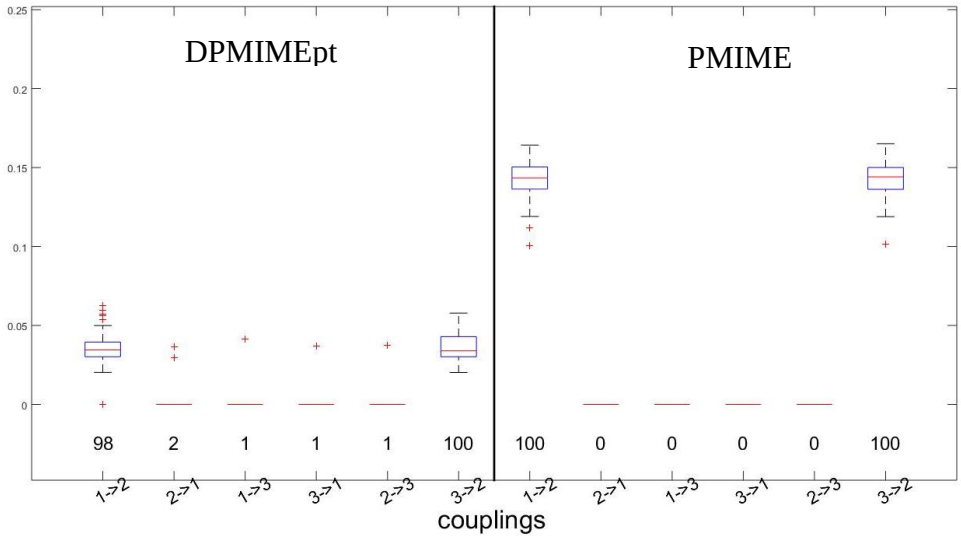
The X_2 variable depends on $X_{1,t}$, $X_{2,t}$, $X_{2,t-1}$ and $X_{3,t}$. Both algorithms do not include the lag term $X_{1,t}$ in the mixed embedding vector but $X_{1,t-1}$ so that the variable X_1 is represented in the mixed embedding vector and the correct causal effect from X_1 to X_2 is established. The lag terms $X_{2,t}$ and $X_{2,t-1}$ are always present in the mixed embedding vector for both algorithms. The representation of X_3 in the mixed embedding vector is spread to the two first lags for PMIME and to the first four lags

for DPMIMEpt, and the causal effect from X_3 to X_2 is established. So, the algorithms have similar performance.

3.3 SYSTEM 1

In the example above, we focused on the exact lag terms in the mixed embedding vector. Here, we report the causal effect as evaluated by DPMIMEpt and PMIME from each variable to the response. The causal effect for DPMIMEpt (as well as for DPMIMErt and PMIME) is given by eq.(3), evaluating the contribution of all the lag terms of the driving variable in explaining the response. The distribution of DPMIMEpt and PMIME and the corresponding rate of detection of causality for all variable pairs are shown in Figure 2.

Figure 2. *Boxplots of DPMIMEpt (ns=2) and PMIME for all variable pairs of system S1 for 100 realizations, L=5 and n=1024. At each panel the number of detections of causal effect (the measure is non-zero) is displayed below each boxplot. The true causality effects are:* $X_1 \rightarrow X_2$ *and* $X_3 \rightarrow X_2$.



The boxplots display the distribution of DPMIMEpt and PMIME over the 100 realizations for each ordered pair of variables and the number below each boxplot is the number of times the measure was found non-zero. Both measures identify almost perfectly the non-existent causal effects with a percentage of false detection less than 2%. Similarly, the two measures detect almost perfectly the true causal effects at a percentage 100 (98 for pair $X_1 \rightarrow X_2$ and DPMIMEpt).

To summarize the performance of DPMIMEpt and PMIME at each realization of S1, we calculate the performance indices sens, spec, MCC, FM and HD on the totally six binary directed connections where two of them are true. In Table 2, the average

indices on the 100 realizations of system S1 for n =1024 are shown for both measures. For both sensitivity and specificity there are presented very high values and the rest of the overall indices are following. The performance of DPMIMEpt is close to the perfect performance of PMIME.

Table 2. *Sensitivity (sens), specificity (spec), MCC, F-measure (FM) and Hamming distance (HD) (average value over 100 realizations) of the causality measures PMIME and DPMIMEpt (ns=2) for system S1, L=5 and n=1024.*

	DPMIMEpt	**PMIME**
sens	0.99	1
spec	0.99	1
MCC	0.98	1
FM	0.98	1
HD	0.07	0

In order to compare the DPMIME measures, using parametric and randomization tests, with PMIME we run simulations for different time series lengths n and number of symbols ns, and the MCC index that weighs sensitivity and specificity is presented in Table 3. Focusing on the DPMIMEpt and DPMIMErt measures, we observe a similar performance for all scenarios. The DPMIMEpt and DPMIMErt score similarly in MCC and at a lower level than PMIME, converging to the highest level with the increase of n. For ns=4 symbols, the performance of DPMIMEpt and DPMIMErt is worse than that of PMIME and the difference decreases with n, indicating that for a larger number of symbols longer time series is needed.

Table 3. *MCC (average value over 100 realizations) for DPMIMEpt, DPMIMErt and PMIME and system S1, L=5, ns=2,4 and n = 512,1024, 2048, 4096.*

	***ns*=2**		***ns*=4**		
	DPMIMEpt	**DPMIMErt**	**DPMIMEpt**	**DPMIMErt**	**PMIME**
n=512	0.84	0.77	0.24	0.13	1
n=1024	0.98	0.97	0.63	0.63	1
n=2048	1	1	1	0.96	1
n=4096	1	1	1	1	1

3.4 SYSTEM 2

Similar results for coupled Hénon maps with K=5 variables are presented in Table 4. The MCC index is presented again for the DPMIME with parametric and randomization test as well as PMIME for different n and ns. Here, again the accuracy in detecting the true causal effects is better for smaller ns and larger length n, converging to the highest performance level with n, obtained by PMIME. Again the DPMIME with the parametric test performs equally well as when randomization test is used instead.

Table 4. *MCC (average value over 100 realizations) for DPMIMEpt, DPMIMErt and PMIME and system S2, L=5, ns=2,4 and n = 512,1024, 2048, 4096.*

	***ns*=2**		***ns*=4**		
	DPMIMEpt	**DPMIMErt**	**DPMIMEpt**	**DPMIMErt**	**PMIME**
n=512	0.78	0.76	0.49	0.39	0.98
n=1024	0.96	0.95	0.70	0.70	1
n=2048	1	1	0.99	0.99	1
n=4096	1	1	1	1	1

3.5 SYSTEM 3

The system S3 is a linear VAR(4) with K=5 variables and it is included to assess the performance of information measures in the presence of solely linear interactions. We observe from the MCC in Table 4 that the performance of the DPMIME with the parametric and randomization test is very close to that of PMIME even for small time series length. In this system the raise of the number of symbols for the same n affects positively the performance of the measures.

Table 4. *MCC (average value over 100 realizations) for DPMIMEpt, DPMIMErt and PMIME for system S3, L=4, ns=2,4 and n = 512,1024, 2048, 4096.*

	***ns*=2**		***ns*=4**		
	DPMIMEpt	**DPMIMErt**	**DPMIMEpt**	**DPMIMErt**	**PMIME**
n=512	0.66	0.71	0.83	0.76	0.85
n=1024	0.71	0.76	0.98	0.97	0.87
n=2048	0.70	0.73	0.88	1	0.88
n=4096	0.69	0.72	0.87	0.99	0.88

4. DISCUSSION

In this study, we propose a causality measure for discrete multivariate time series based on partial mutual information from mixed embedding named DPMIME. The DPMIME is developed similarly to the PMIME, a causality measure for continuous-valued time series that uses information measure and dimension reduction, so that it can be applied for a large number of observed variables. The idea in PMIME and DPMIME is to search for the most relevant lag variables that can explain the response evaluated by information measures, i.e. mutual information (for the first step when the information of the response and a lag variable is evaluated) and the conditional mutual information (for the subsequent steps when the information of the response and a lag variable is evaluated accounting from the information on the response from the subset of lag variables already selected). For the termination of the iterative algorithm the significance of the conditional mutual information is tested using a parametric test, denoted DPMIMEpt, and a randomization test, denoted DPMIMErt (there is only randomization test for PMIME).

For the simulation study, three systems were used, coupled Hénon maps system with ring structure on three variables (S1) and five variables (S2) and linear stochastic process of five variables (S3). The PMIME was computed directly on the continuous-valued multivariate time series of systems S1, S2 and S3. To test the DPMIME on the same systems, first each continuous-valued time series was discretized to a predefined number of symbols *ns* and the DPMIME was computed on the discrete-valued time series. The performance of all measures was quantified by five performance indices (sensitivity, specificity, Matthews correlation coefficient, F-measure and Hamming distance). In all simulated systems, we observed that DPMIMEpt and DPMIMErt had similar performance. This is an important finding that allows for the use of the parametric test in DPIME (DPMIMEpt) and save computation time. In this way DPMIME is much faster than PMIME and can even be preferred to save time for investigating direct causality in continuous-valued time series (applying first data discretization).

Compared to PMIME, DPMIME with both tests were less accurate by their performance converged to that of PMIME with the increase of time series length *n*. In the simulations, we observed that for n=2048 both measures attained good performance close to that of the PMIME. The increase of number of symbols *ns* (the level of discretization) worsened the performance of DPMIME for small time series length, and in particular for the chaotic systems S1 and S2, implying that the inadequate estimation of probabilities due to small size increased the variability of the estimation of the information measures resulting in inaccurate detection of the true causal effects. For larger time series, however, the opposite observed, i.e. for larger *ns* there is a more detailed representation of the relationships and as the estimation of probabilities and information measures is less variable due to larger *n* the true causal effects could be better detected. Overall, the results on systems S1, S2 and S3 showed that the DPMIME measures (with parametric and randomization tests) attain high performance and for longer time series close to the high performance of PMIME.

There is an intrinsic issue of high dimensionality in the estimation of the lag causal relationships by DPMIME. This setting occurs when K is large, not treated in this study as the largest number of observed time series is K=5, but also when the maximum lag L is large or in general the product KL is large, as this defined the number of candidate components for the mixed embedding vector $\mathbf{w}_t$. If there are only few causal lag-relationships so that $\mathbf{w}_t$ is relatively small, then the entropies can be estimated and the DPMIME will provide reliable estimation of the direct causal relationships. Still, if the mixed embedding vector gets large, the estimation of the entropies from the relative frequencies of large words will be problematic and DPMIME will not perform well. The latter is a general problem in causality estimation and no remedy or solution is available to-date, to the best of our knowledge.

ΠΕΡΙΛΗΨΗ

Στην ανάλυση πολυ-διάστατων χρονοσειρών διάφορες μέθοδοι έχουν αναπτυχθεί για την εκτίμηση σχέσων αιτιότητας ανάμεσα στις παρατηρούμενες μεταβλητές. Για τις υψηλής διάστασης χρονοσειρές, προσεγγίσεις για μείωση διάστασης έχουν αναπτυχθεί για την εκτίμηση των άμεσων σχέσεων αιτιότητας, με σκοπό τον ορισμό των συνδέσεων ενός πολύπλοκου δικτύου το οποίο αναπαριστά τη δομή του υποκείμενου δυναμικού συστήματος ή της στοχαστικής διαδικασίας. Η μερική δεσμευμένη αμοιβαία πληροφορία από μικτή εμβύθινση (PMIME) χρησιμοποιώντας τη δεσμευμένη αμοιβαία πληροφορία (CMI) εφαρμόζει μια τέτοια προσέγγιση και διαπιστώνεται ότι είναι κατάλληλη για την εκτίμηση άμεσων σχέσεων αιτιότητας από υψηλής διάστασης χρονοσειρές συνεχών τιμών. Σε αυτήν τη μελέτη, το ενδιαφέρον είναι για τις διακριτές πολυδιάστατες χρονοσειρές και προσαρμόζεται κατάλληλα το PMIME για να οριστεί το διακριτό PMIME (DPMIME). Η κατάλληλη εκτίμηση διακριτών κατανομών πιθανοτήτων και CMI για διακριτές μεταβλητές εφαρμόζεται στο DPMIME. Περαιτέρω, η ασυμπτωτική κατανομή του εκτιμώμενου CMI επιτρέπει τον παραμετρικό έλεγχο σημαντικότητας του CMI στο DPMIME, ενώ στο PMIME υπάρχει μόνο έλεγχος σημαντικότητας μέσω τυχαιοποίησης για το CMI. Ο παραμετρικός έλεγχος σημαντικότητας για το CMI στον αλγόριθμο του DPMIME έχει σχεδόν την ίδια απόδοση με τον αντίστοιχο έλεγχο τυχαιοποίησης. Οι Monte Carlo προσομοιώσεις σε πολυδιάστατες ακολουθίες συμβόλων που παράγονται από τη διακριτοποίηση χρονοσειρών συνεχών τιμών έδειξαν ότι η ακρίβεια του DPMIME στην εκτίμηση της άμεσης αιτιότητας συγκλίνει, καθώς αυξάνεται το μήκος των χρονοσειρών, στην ακρίβεια του PMIME.

REFERENCES

Angers JF, Biswas A and Maiti R (2017). Bayesian Forecasting for Time Series of Categorical Data, *Journal of Forecasting*, 36: 217-229

Billio M, Getmansky M, Lo AW and Pelizzon L (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics* ,104, 535–559

Biswas A and Guha A (2009). Time series analysis of categorical data using auto-mutual information, *Journal of Statistical Planning and Inference*, 139:3076-3087

Budhathoki K and Vreeken J (2018). Causal Inference on Event Sequences, *Proceedings of the 2018 SIAM International Conference on Data Mining*, https://doi.org/10.1137/1.9781611975321.7

Ching WK, Fung ES and Ng MK (2004). Higher-order Markov chain models for categorical data sequences, *Naval Research Logistics*, 51(4)

Ching WK, Ng MK and Fung ES (2008). Higher-order multivariate Markov chains and their applications, *Linear Algebra and its Applications*, 428(2-3):492-507

Cover TM and Thomas JA (1999). *Elements of Information* Theory, Wiley, London.

Christou V and Fokianos K (2015). On count time series prediction, *Journal of Statistical Computation and Simulation*, 85(2):357-373

Davis AR and Wu R (2009). A negative binomial model for time series of counts, *Biometrika*, 96(3):735-749

Dijkstra H, Hernández-García E, Masoller C, Barreiro M (2019) *Networks in Climate*, Cambridge University Press: Cambridge, UK

Fieguth P (2017). A*n introduction to Complex Systems: Society, Ecology and Nonlinear Dynamics*, Springer, Switzerland

Fokianos K, Rahbek A and Tjøstheim D (2009). Poisson Autoregression, *Journal of the American Statistical Association*, 104(488):1430-1439

Goebel B, Dawy, Z, Hagenauer, J, Mueller, J (2005). An approximation to the distribution of finite sample size mutual information estimates, *IEEE* 2, 1102–1106

Granger CWJ (1969). Investigating causal relations by econometric models and cross-spectral Methods. *Econometrica* 37, 424–438

Kirchgässner G, Wolters J and Hassler U (2013). *Granger Causality (Chp 3)*. In: Introduction to Modern Time Series Analysis. Springer Texts in Business and Economics. Springer, Berlin, Heidelberg

Kugiumtzis D (2013). Direct-coupling information measure from non-uniform embedding, *Physical Review E*, 87(6): 062918

Matthews BW (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *Biochimica et Biophysica Acta*, vol. 405, no. 2, pp. 442–451

Newman M (2010). *Networks, an introduction*. Oxford University Press

Neumann M (2011). Absolute regularity and ergodicity of Poisson count processes, *Bernulli*, 17(4):1268-1274

Nicolau J (2014). A New Model for Multivariate Markov Chains, *Scandinavian Journal of Statistics*, 41: 1124-1135

Papapetrou M and Kugiumtzis D (2013). Markov chain order estimation with conditional mutual information, *Physica A* (392): 1593–1601

Papapetrou M and Kugiumtzis D (2015). Markov chain order estimation with parametric significance tests of conditional mutual information, *Simulation Modelling Practice and Theory*, 61: 1-13

Pedeli X and Karlis D (2013). Some properties of multivariate INAR(1) processes, *Computational Statistics and Data Analysis*, 67: 213-225

Politi A and Torcini A (1992). Periodic orbits in coupled henon maps: Lyapunov and multifractal analysis, *Chaos* 2, 293

Porta A and Faes L (2016). Wiener-Granger causality in network physiology with applications to cardiovascular control and neuroscience. *Proceedings IEEE* 104, 282–309

Raftery AE (1985). A Model for High-order Markov chains, *Journal of the Royal Statistical Society*, 47(3): 528-539

Schelter B, Winterhalder M, Hellwig B, Guschlbauer B, Lucking C.H, Timmer J (2006). Direct or indirect? Graphical models for neural oscillators. *J. Physiol. Paris*, 99:37–46.

Scotto MG, Weiss CH and Gouveia S (2015). Thinning-based models in the analysis of integer-valued time series: a review, *Statistical Modelling*, 15(6): 590–618.

Song PXK, Freeland RK, Biswas A and Zhang S (2013). Statistical analysis of discrete-valued time series using categorical ARMA models, *Computational Statistics & Data Analysis*, 57(1):112-124

Tank A, Fox E and Shojaie A (2017). Granger Causality Networks for Categorical Time Series, arXiv:1706.02781

Thurner S, Hanel R and Klimek P (2018). *Introduction to the Theory of Complex Systems*, Oxford University Press.

Weiss C (2018) *An Introduction to Discrete-Valued Time Series*, Wiley & Sons

Yu GH and Huang CC (2001). A distribution free plotting position, *Stochast. Environ. Res. Risk Assess*, 15: 462–476.

Zhou G and Ye X (2017). High-order interacting multiple model filter based on mixture transition distribution, *International Conference on Radar Systems*, DOI: 10.1049/cp.2017.0523

Greek Statistical Institute
Proceedings of the 33rd Panhellenic Statistics Conference (2021), pp.330-344

OPTIMAL LOCATION OF AERIAL FIREFIGHTING RESOURCES TO MAXIMIZE COVERAGE: A CASE STUDY OF GREECE

Konstantinos A. Tasias, Paraskevi Kosti

University of Western Macedonia, Department of Mechanical Engineering

ktasias@uowm.gr, paraskeuhkosti@gmail.com

ABSTRACT

Wildfires affect many different regions of the world having many negative consequences on humans, wildlife, and the economy. Aerial resources play a crucial role in combating wildfires and an effective resource location plan is of paramount importance for enhancing their utilization. To this effect, several factors should be considered such as the fire hazard levels of each area, the specificities of the aerial firefighting fleet, the availability of airbases (airports and heliports), etc. In the present study, a decision support model is developed, under realistic assumptions and driven by practical considerations, to dictate the optimal assignment of aerial firefighting resources to the available facilities, in order to maximize the expected coverage. To demonstrate the applicability of the model, Greece is employed as a case study, considering multi-year wildfire statistics, the Greek aerial firefighting fleet, and the licensed airbases in Greek territory. Finally, the effectiveness of the model is evaluated through a comparison between the optimal solution and an existing location plan of Greek aerial firefighting resources.

Key Words: Wildfires; Aerial firefighting resources; Location model

1. INTRODUCTION

Wildfires are uncontrollable fires that destroy forests, grasslands, savannas, and other ecosystems, being significant threats to wildlife, human lives, and property (Pausas and Keeley, 2009). Wildfires occur every year in many regions around the world necessitating effective wildfire management techniques, whose development requires more attention, funds, and effective solutions (Xanthopoulos, 2008).

The process of wildfire management consists of three phases: prevention, suppression, and restoration (Goldammer *et al*., 2019). This study is focused on the suppression aspect of wildfire management, a demanding task that prerequisites proper coordination of the forest firefighting mechanism, i.e., the aerial and terrestrial firefighting resources, and the personnel supporting the operations. However, the suppression aspect of wildfires has not been adequately studied in the existing research (Finney *et al*. (2009), Pacheco *et al*. (2015)).

Aerial firefighting is a vital ally to wildfire suppression. Aerial means started assisting firefighting at the beginning of the 20th century (Kal'avský *et al*., 2019) and till the present time, their contribution to wildfire suppression is very important. Two key features of the aerial firefighting resources are the following: a. their ability to access almost every area, a very crucial advantage, especially whenever a fire is inaccessible by land; b. their instant response, an important characteristic due to the rapid growth of fires and the necessity to be confronted as soon as possible. The aerial firefighting means are often used as an initial attack on wildfires to slow their progression, thereby facilitating the flame extinguishment for the ground-based operations. Aerial resources' allocation policy plays a key role in the fight against wildfires and an optimal location plan can significantly enhance the effectiveness and efficiency of firefighting operations.

Nevertheless, despite the crucial role of aerial firefighting resources to combat wildfires, their management has not received adequate research attention. Martell *et al*. (1984) and Maclellan and Martell (1996) were the first to study the aerial firefighting resources' location problem through an operational research approach. Islam *et al*. (2009) developed a heuristic procedure to dictate optimal operational decisions associated with the deployment of the aerial means. Zeferino (2020) developed a combinatorial optimization model for the location of aerial resources to combat wildfires and demonstrated the methodology in the case study of Portugal.

The key contribution of this study lies in the development of a quantitative and effective decision-support tool that dictates the optimal location of the aerial firefighting resources to combat wildfires under a maximal coverage criterion. The decisions are made based on the statistical analysis of historical data regarding the fire incidents and the aerial means operations. Furthermore, the proposed tool considers many realistic parameters that enhance the applicability and result in better performance. Specifically, the statistical evaluation of historical data to compute the unavailability of aerial resources adds a stochastic component to the problem that addresses the impact of multiple fire ignitions on the availability of resources. Moreover, the required time to take-off and the cruise speed of each type of resource are considered due to their direct effect on the celerity of each type to reach a fire incident, and so, to the effectiveness of the resource's wildfire response. Finally, the proposed methodology is demonstrated in the case of Greece, and a comparison with an existing location plan is performed to evaluate the model's performance.

2. PROBLEM SETTING

Let us assume a specific region of interest, such as a country or a continent, potentially affected by wildfires. The exact place and time of a wildfire initiation cannot be a priori known. However, there are subregions within the area of interest more prone to wildfires compared to others, due to several reasons (e.g., ground characteristics, climate conditions). Consequently, the wildfire ignition risk, and so, the demand for firefighting resources is different for subregions within the specific area.

The area of interest has an available fleet of aerial firefighting resources to combat wildfires. These aerial resources have different characteristics, such as their water capacity, their speed, their operational range, etc., which should be considered to evaluate the effectiveness of their wildfire response.

Furthermore, the region has a specific number of designated airbases, i.e., airports and heliports, where the aerial firefighting resources may be located to operate whenever needed.

The problem lies in the derivation of the optimal assignment of the aerial firefighting resources to the available airbases at the beginning of the fire season, to maximize the wildfire risk coverage of the region, while complying with realistic constraints.

3. MODEL FORMULATION

3.1 Notations

In principle, the aerial firefighting allocation problem is formulated as an integer linear programming problem, whose objective is the maximization of the expected coverage of the area of interest by the aerial firefighting resources.

To facilitate the model formulation, the area of interest is discretized into equal square shapes, i.e., nodes. The size of each square/node, or equivalently the level of how coarse or fine the area discretization is, does not affect the model formulation, whatsoever. Nevertheless, the area should be discretized into squares with a uniform wildfire risk to the greatest possible extent. Subsequently, the square size should be selected on a case-by-case basis, by considering that a coarse area discretization and so, a larger number of nodes, results in closer to the optimum solution, but at the expense of heavier computational requirements.

For the model formulation, the following parameters are considered:

a. The set of different types of aerial firefighting resources (T).

b. The available number of each type of aircraft and helicopter $t \in T$, denoted by AV_t.

c. The key features of each type of aerial firefighting resource:

(1) cruise speed $\left(cs_t\right)$;

(2) water capacity $\left(wc_t\right)$.

(3) time required to take-off $\left(tot_t\right)$;

d. The set of nodes representing the wildfire risk in each subregion within the area of interest (*I*).

e. The wildfire risk level of each node $i \in I$, denoted by r_i. Each level of wildfire risk is associated with a minimum number of aerial firefighting resources needed to provide coverage to nodes characterized by the specific level of fire hazard, denoted by M_i.

f. The set of available airbases (*J*).

g. The maximum capacity of each airbase *j* to support each type of resource, denoted by N_{jt}.

h. The availability of an aerial resource to cover a node $i \in I$, or equivalently, the probability of a concurrent operation to another fire incident for a specific aerial resource, denoted by *p*.

3.2. Assumptions

1. Both aircrafts and helicopters can be located at airports.

2. Aircrafts cannot be assigned to heliports. This assumption is justified by the lack of runways in facilities designed only for helicopter operations.

3. Airbases have a limited capacity regarding: a. the total number of aerial firefighting resources they can support; b. the maximum number of each type of resource. This is due to several constraints including most likely the available ground handling equipment, the available number of authorized personnel, and the airbase fuel storage capacity.

4. A motherbase, i.e., the airbase where pilot training operations and maintenance actions of a specific type of resource are implemented prior to and following the fire season, can support all the aerial means of that type. This assumption is justified by the fact that motherbases are designed for fully supporting this resource type and so, they are adequately equipped both with the required equipment and personnel to handle many resources of that type. Nevertheless, this is not a restrictive assumption since a limited capacity may also be set for motherbases.

5. THE OPTIMIZATION PROBLEM

5.1 Objective Function

The optimization purpose of the problem is to maximize the coverage of the total set of nodes *I*, depending on the values of the fire ignition risk level of each node r_i and by considering the diversion of the available firefighting resources to structure protection.

To this effect, the cruise speed (cs_t) and the water capacity (wc_t) of each type, are normalized to the largest value, thus, formulating a cruise speed indicator (CS_t) and a water capacity indicator (WC_t), as follows: $CS_t = \frac{cs_t}{\max\limits_t cs_t}$, $WC_t = \frac{wc_t}{\max\limits_t wc_t}$.

Subsequently, the cruise speed and water capacity of the aerial firefighting resources are evaluated by the objective function through their average values: $CS = \frac{\sum_{t=1}^{T} CS_t \cdot AV_t}{\sum_{t=1}^{T} AV_t}$ and $WC = \frac{\sum_{t=1}^{T} WC_t \cdot AV_t}{\sum_{t=1}^{T} AV_t}$.

Each node $i \in I$ is considered as being covered by an aerial resource $t \in T$, based on the radius of action of the latter (R_t). In specific, R_t represents the distance a specific type of aerial firefighting resource can cover within a prespecified time interval (TI), thus, being the direct analog of the service level in customer service. Regarding the variable TI, it should be set to a relatively low value (e.g., 20 minutes, 30 minutes), being an immediate response to a fire occurrence crucial for mitigating its impact. The radius of action of each type of resource depends both on its cruise speed (cs_t) and its required time to take off (tot_t) and is computed by the following expression: $R_t = cs_t \cdot (TI - tot_t)$.

Consequently, if R_t is greater than the distance between the location of the airbase $j \in J$, where an aerial resource of type *t* is assigned and the center of a node *i*, then, the node is considered as covered by this aerial resource. The two alternatives regarding the coverage of node *i* are classified into a nominal variable as follows:

$$s_{ijt} = \begin{cases} 1, & \textit{if a resource of type } t, \textit{ located at airbase } j, \textit{ can cover node } i \\ 0, & \textit{otherwise} \end{cases}$$

However, to consider a node *i* as fully covered by aerial firefighting resources, it must lie within the radius of action of a set of aerial resources, denoted by *D*.

The full coverage of a node *i* is introduced as a nominal variable as follows:

$$S_{iD} = \begin{cases} 1, & \textit{if node } i \textit{ is covered by } D \textit{ aerial means} \\ 0, & \textit{otherwise} \end{cases}$$

The objective is to find the optimal number of aerial resources of type *t* that should be located at each airbase *j*, denoted by X_{jt}, to maximize the expected coverage of the area of interest, formulated as follows:

$$\max \sum_{i \in I} r_i \cdot WC \cdot CS \cdot \sum_{d \in D} (1-p) \cdot p^{d-1} \cdot S_{iD} \tag{1}$$

5.2 Constraints

As already stated, a node *i* is considered as covered provided that the number of aerial means that can cover the node exceeds a minimum number of *D* means placed at facility *j*. To satisfy this presumption, the following constraint is considered:

$$\sum_{d \in D} S_{id} \le \sum_{j \in J} \sum_{t \in T} s_{ijt} \cdot X_{jt} \quad \forall i \in I \tag{2}$$

A minimum number of aerial means that must provide coverage to a given wildfire risk level is set, and so:

$$\sum_{d \in D} S_{id} \ge M_i \quad \forall i \in I \tag{3}$$

Apparently, the number of aerial resources of type *t*, located at an airbase $j \in J$, cannot be greater than the total number of available resources of this specific type:

$$\sum_{j \in J} X_{jt} \le AV_t \quad \forall t \in T \tag{4}$$

Furthermore, to ensure that the number of the aerial firefighting resources assigned to an airbase *j* would exceed neither the total capacity of the airbase nor its capacity regarding each type of resource *t*, the following constraints are considered:

$$X_{jt} \le N_{jt} \quad \forall j \in J,\ \forall t \in T \tag{5}$$

$$\sum_{t \in T} X_{jt} \le N_j \quad \forall j \in J,\ \forall t \in T \tag{6}$$

It is often in practice, a specific type of aircraft and/or helicopter to deploy to airbases and used in firefighting operations, in pairs. This policy is motivated by several

reasons such as the facilitation of concurrent transportation of personnel and equipment, thereby, avoiding the use of a cargo plane, the better utilization of the available resources, and other operational-related reasons. The set of types of aerial resources which deploy in pairs is denoted by TP, and variable XP_{jt} represents the number of pairs of type *t* located at airbase *j*. Consequently:

$$X_{jt} - 2 \cdot XP_{jt} = 0 \quad \forall j \in J \ , \ \forall t \in TP \tag{7}$$

Finally, the following constraints (Equations (8)-(10)) set the boundaries of the decision and nominal variables of the problem:

$$X_{jt} \in Z^{+} \quad \forall j \in J \ , \ \forall t \in T \tag{8}$$

$$XP_{jt} \in Z^{+} \quad \forall j \in J \ , \ \forall t \in TP \tag{9}$$

$$s_{ijt,} S_{id} = \{0,1\} \quad \forall i \in I \ , \ \forall j \in J \quad \forall d \in D \tag{10}$$

A computer program developed in MATLAB R2017a dictates the optimum solution to the optimization problem. Moreover, the open-source Geographic Information System (*GIS*) program software *QGIS*, was used to illustrate the geographical data and create maps portraying the results of the optimization.

6. CASE STUDY

6.1 Greece's Case

To demonstrate the applicability and evaluate the proposed model's performance, the case study of Greece is employed. Greece is severely affected by wildfires, which, apart from threatening forests, exurban areas, archeological sites, and world cultural heritage monuments, lead to the loss of human lives, as well. Specifically, in Greece on average, six people lose their lives every year due to wildfires, or equivalently wildfires claim one human life per month of the fire season (Gourbatsis, 2021).

Some recent examples, indicative of the harsh effect of wildfires in Greece, are a. the wildfires at Evia and Peloponnese, which burnt and damaged almost 670.000 acres of land and killed 63 people back in 2007; b. the wildfires at Mati and N. Voutzas, being the second deadliest wildfire worldwide for the 21st century, killing 102 people, and burning over 14.000 acres in 2018. c. the multiple wildfires in Attica, Peloponnese, and Evia in 2021 killed 3 people and burnt more than 287.000 acres.

In Greece, the forest area as a share of the land area is estimated approximately equal to 30%, considered as highly flammable due to the dry climate, undulated topography, poor soil quality, and flammable vegetation (e.g., pine trees, shrublands). More factors affecting the high risk of fire occurrence in Greece are related to climate change, socioeconomic issues, neglection and poor forest management (Koutsias *et*

al., 2012a, 2012b; Moreira *et al*., 2011; Mavrakis and Salvati, 2015). Consequently, the damages and losses, along with the fact that wildfires are an ongoing concern for Greece, necessitate an effective wildfire suppression plan (Gourbatsis, 2021; Kampouris, 2021).

The wildfire risk level in each of the 51 prefectures of Greece is presented in Figure 1 (Goldammer *et al*. 2019), illustrating a wildfire risk map based on the Greek Geodetic Reference System GGRS 87, derived from historical data covering a period of 15 years. In specific, the wildfire risk has five possible levels: low, medium, high, very high, extremely high, each highlighted based on a white to black gradient color scheme. A 33.3% of the total number of prefectures of Greece are considered as low risk, a 37.3% medium risk, 17.7% as high risk, 3.9% as very high risk, and 7.8% as extremely high risk.

To incorporate the wildfire ignition risk data into the mathematical model, the abovementioned map was converted into a map grid of 10 km squares (Figure 2), thus, forming a total number of 2048 squares/nodes for the Greek territory. Subsequently, a value ranging from 1 (low risk) to 5 (extremely high risk) is assigned at every square, signifying the level of wildfire ignition risk of each node. In case a square lies on the border of multiple prefectures, the wildfire hazard value is assigned based on the risk of the prefecture where the largest part of the square belongs.

Figure 1 *Wildfire risk map of Greece*

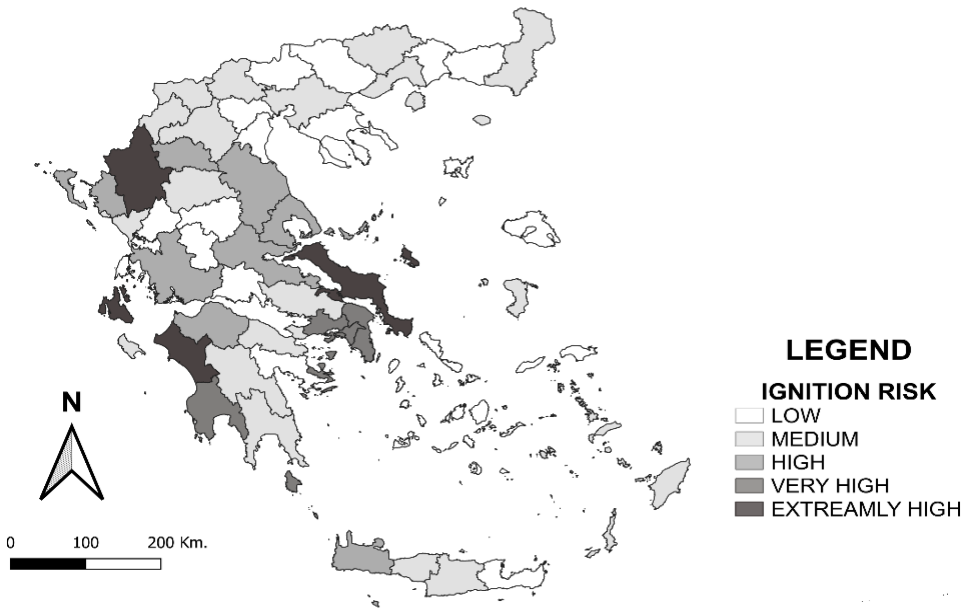
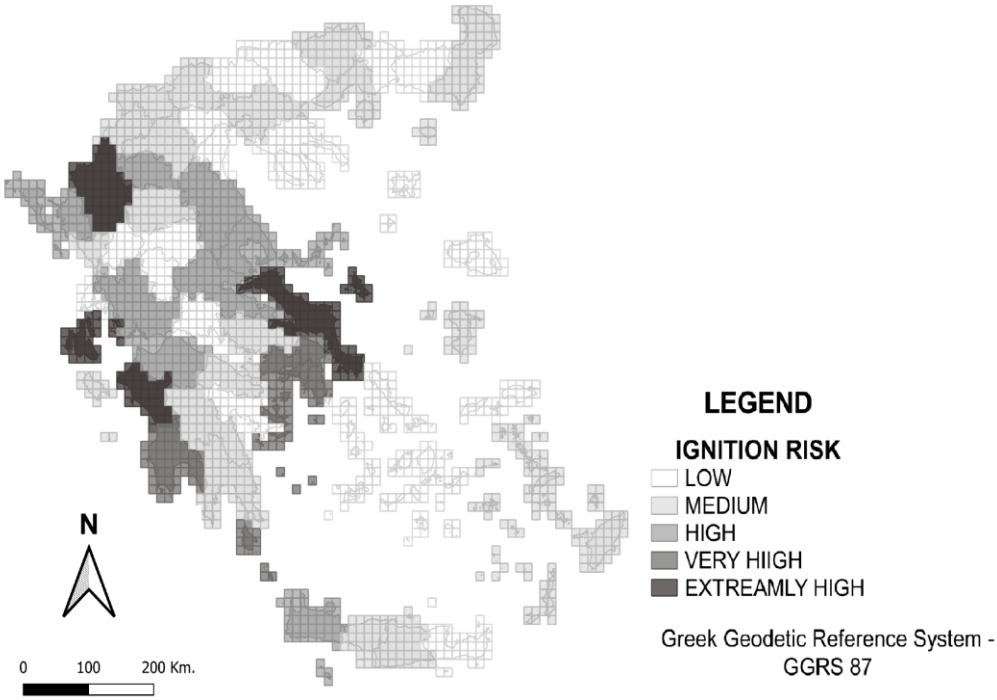


Figure 2 *Wildfire risk map of Greece in grid format*



In Greece, a total of 99 airbases (52 airports and 47 heliports) are potentially available to support aerial firefighting resources. Figure 3 presents the exact location of the airbases on the map of Greece. The motherbase of each resource type (Elefsis airbase (LGEL), Thessaloniki airbase (LGTS), and Tatoi airbase (LGTT)) can support all the available aerial means of that type, but they have a maximum capacity of 4 aircrafts and 4 helicopters regarding the other types. The largest airbases (Larissa airbase (LGLR), Volos airbase (LGBL), Tanagra airbase (LGTG), Andravida airbase (LGAD), Araxos airbase (LGRX), and Souda airbase (LGSA)) can support a maximum of 4 aircrafts, while the rest airbases a maximum of 2 aircrafts. Furthermore, heliports can support only 1 helicopter at a time.

The Greek aerial firefighting fleet consists of resources belonging to the Hellenic Fire Service and the Hellenic Air Force, and additional resources hired on a contract basis for each fire season, thereby, resulting in a slightly different fleet every year. The available firefighting fleet and its allocation in airbases is decided annually and is effective for the respective fire season, i.e., from 1st May to 31st October.

In this study, the Greek aerial firefighting fleet of 2019, consisting of 33 aircrafts and 25 helicopters is used and the types of aerial resources and their characteristics are presented in Table 2.

Figure 3 *Map of available airports and heliports in Greece*

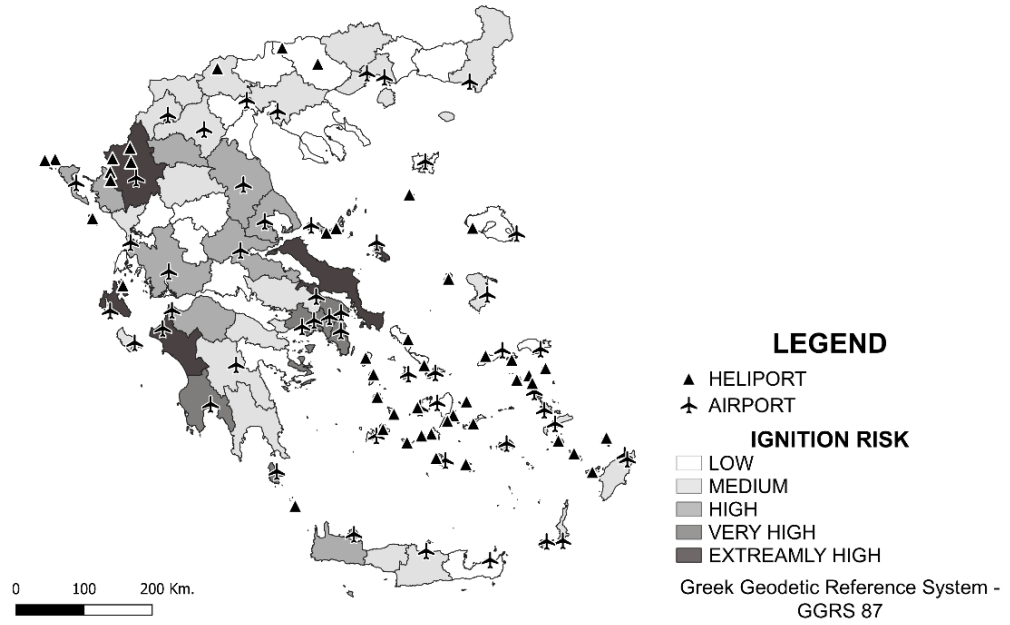


Table 2 *Types of Greek aerial firefighting resources and their characteristics*

t	AV_t	cs_t (*Km/h*)	CS_t	R_t (*Km*)	wc_t (*lt*)	WC_t
CL-415	6	333	1.0	166.5	6137	0.6
CL-215	10	291	0.9	145.5	5000	0.5
PZL	17	200	0.6	100	2200	0.2
BK-117	3	240	0.7	120	910	0.1
AS332	2	222	0.7	111	3500	0.4
Ka-32	12	230	0.7	115	3000	0.3
S-64	8	169	0.5	84.5	10000	1.0
			$CS = 0.71$			$WC = 0.43$

To facilitate the comparison between the proposed model and the existing allocation of aerial firefighting resources in 2019, the radius of action of each type of resource (R_t), is computed in accordance with the policy followed by Hellenic Fire Service, i.e., as the distance covered by each resource within 30 minutes $(TI = 0.5h)$ after its takeoff $(tot_t = 0)$.

Furthermore, the minimum number of aerial firefighting means for each possible wildfire risk level is presented in Table 3.

Table 3 *Minimum Number of Firefighting Resources for each wildfire risk level*

	Low	Medium	High	Very High	Extremely High
M	1	2	3	4	5

Finally, the probability of an aerial resource to be unavailable to combat a wildfire, i.e., simultaneously involved in another fire incident, is statistically evaluated through the annual reports published by the Hellenic Fire Service, where the fire occurrences and the aerial firefighting resources activated, are issued. The value of the above-mentioned probability is computed equal to 8% $(p = 0.08)$, a relatively high probability, indicative of the large number of parallel fire ignitions that Greece often confronts.

6.2 Results

The solution of the optimization problem for the case of Greece yielded the following findings. A 100% expected coverage through the optimal allocation of the available firefighting resources in Greece, presented in Figure 4, is achieved. The different size circles represent the radius of action of the assigned aerial firefighting resources.

To evaluate the effectiveness of the proposed model, a comparison with the existing allocation in the year 2019, illustrated in Figure 5, is made. The year 2019 is selected as the most recent year with reliable information regarding the available aerial firefighting fleet and the existing location plan of resources.

The existing expected coverage provided by the existing allocation in 2019 is computed equal to 85%. Consequently, the expected coverage is significantly lower compared to the expected coverage by the proposed allocation, thus, indicating a suboptimal utilization of the available aerial firefighting resources.

Furthermore, through the comparison of the location plans, i.e., the proposed and the existing, it is concluded that the aerial means were predominately concentrated around Attica and the Peloponnese in 2019, resulting in rather insufficient coverage of Northern Greece and the Aegean Sea islands. In particular, the existing allocation activated 22 airbases, with most of the aerial resources gathered either at the motherbases (LGEL, LGTS, and LGTT) or at the largest airbases (LGLR, LGBL, LGTG, LGAD, LGRX, and LGSA). On the other hand, the optimal allocation proposes a more even spatial distribution of the aerial firefighting fleet, activating 40 airbases (31 airports and 9 heliports). The almost doubled number of airbases compared to the respective number of 2019 demonstrates that increased expected coverage of Greece prerequisites a sporadic allocation of the available aerial firefighting resources.

Furthermore, the optimal location plan highlights the crucial role of heliports in achieving high expected coverage. As already stated, 22.5% of the activated airbases are heliports serving as locating points for the 15.5% of the available aerial firefighting means. Therefore, heliports provide coverage at areas with a limited number of surrounding airbases, such as the Greek islands, which were not sufficiently covered by the existing distribution in 2019.

Finally, another conclusion made by comparing the proposed location plan to the existing location plan is that, in the latter, the aerial means with relatively large water capacity were located near the largest cities in Greece, while the aerial means with relatively low water capacity were deployed to the peripheral airbases. Nevertheless, the proposed optimal allocation plan does not satisfy this spatial distribution pattern.

Figure 4 *Coverage map of Greece based on the optimal aerial resources' distribution*

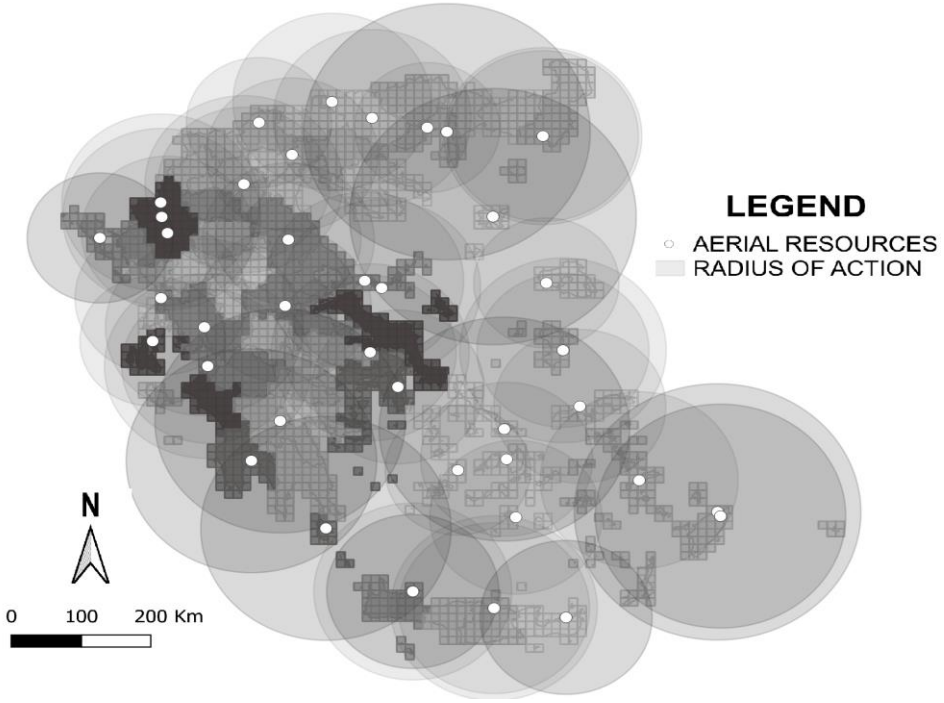
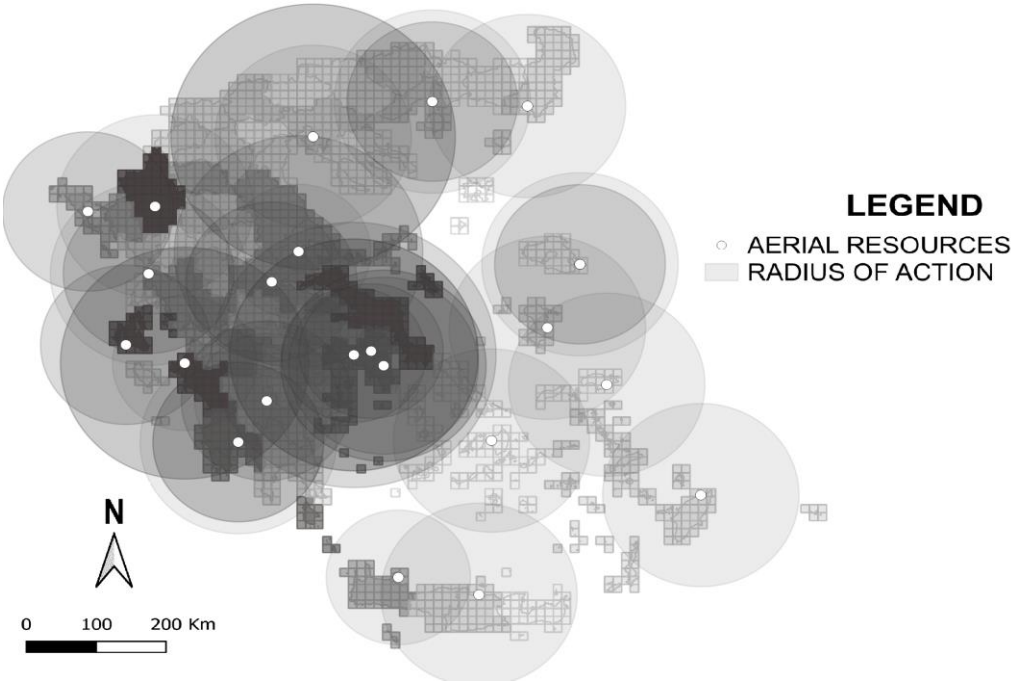


Figure 5 *Coverage map of Greece based on the 2019 aerial resources' distribution*



7. SUMMARY AND CONCLUSIONS

In this study, a decision-making tool on wildfire management is developed, which determines the aerial firefighting resources' location plan that maximizes the expected coverage of a specific area. The model incorporates many important and realistic factors to dictate the optimal aerial firefighting fleet's location plan, such as the characteristics of each type of resource, the capacity of the available airbases, and the wildfire ignition risk of each region within the area of interest. The model was evaluated in the case study of Greece, verifying a significant improvement in the efficiency of the proposed location plan to combat wildfires compared to the existing location plan and the subsequent coverage of Greece in 2019.

ΠΕΡΙΛΗΨΗ

Οι πυρκαγιές επηρεάζουν πολλές διαφορετικές περιοχές του πλανήτη με πλήθος αρνητικών επιπτώσεων στους ανθρώπους, στην άγρια ζωή και στην οικονομία. Για το λόγο αυτό, η ύπαρξη ενός αποτελεσματικού σχεδίου αντιμετώπισης των πυρκαγιών είναι καίριας σημασίας, ιδιαίτερα κατά τους κρίσιμους μήνες της αντιπυρικής περιόδου. Τα εναέρια μέσα πυρόσβεσης αποτελούν σημαντικό σύμμαχο στην προσπάθεια καταστολής των πυρκαγιών. Η παρούσα εργασία μελετά την κατανομή των κλιμακίων των πυροσβεστικών αεροσκαφών και ελικοπτέρων σε αεροπορικές βάσεις, αναπτύσσοντας ένα μοντέλο λήψης βέλτιστων

αποφάσεων σε τακτικό επίπεδο, με στόχο τη βέλτιστη αεροπορική κάλυψη μιας περιοχής από τις πυρκαγιές. Το μοντέλο λαμβάνει υπόψη ρεαλιστικές παραμέτρους και δεδομένα, όπως η επικινδυνότητα έναυσης πυρκαγιάς ανά περιοχή, τα χαρακτηριστικά των διαθέσιμων εναέριων μέσων, τη μέγιστη διαθεσιμότητα των αεροπορικών βάσεων κ.α. Η εφαρμογή και η αποτελεσματικότητα του μοντέλου αναδεικνύονται μέσω της μελέτη περίπτωσης της Ελλάδας, για την οποία εξάγεται η βέλτιστη κατανομή εναέριων μέσων για τη μέγιστη κάλυψη της επικράτειας. Στο τέλος, γίνεται η αξιολόγηση των αποτελεσμάτων και η εξαγωγή χρήσιμων συμπερασμάτων.

Acknowledgments: The authors would like to express their gratitude to Fire Major Nikolaos Papadelis of the Hellenic Air Force, for his guidance, insightful comments, and support during this study.

REFERENCES

Finney, M., Grenfell, I.C., & McHugh, C. W. (2009). Modeling containment of large wildfires using generalized linear mixed-model analysis. *Forest Science*, **55**, 249-255.

Goldammer, J. G., Xanthopoulos, G., Eftixidis, G., Mallinis, G., Mitsopoulos, I., & Dimitrakopoulos, A. (2019). Report of the Independent Committee tasked to Analyze the Underlying Causes and Explore the Perspectives for the Future Management of Landscape Fires in Greece.

Gourbatsis, A. (2021). Deadly and destructive wildfires of Greece (1981-2020). Athens. Retrieved from https://www.eglimata-emprismou.gr/

Islam, K.S, Martell, D.L., Posner, M.J. (2009). A time-dependent spatial queueing model for the daily deployment of airtankers for forest fire control. INFOR: Information Systems and Operational Research, **47**, 319-333.

Kal'avský, P., Petríček, P., Kelemen, M., Rozenberg, R., Jevčák, J., Tomaško, R., & Mikula, B. (2019). The Efficiency of Aerial Firefighting in Varying Flying Conditions. *International Conference on Military Technologies (ICMT)*, 1-5.

Kampouris, N. (2021). Greece Fires Continue in Peloponnese; 287.049 Acres Burned in 2021. Retrieved from https://www.greekreporter.com.

Koutsias, N., Arianoutsou, M., Kallimanis, A., Mallinis, G., Halley, J., & Dimopoulos, P. (2012a). Where did the fires burn in Peloponnisos, Greece the summer of 2007? Evidence for a synergy of fuel and weather. *Agricultural and Forest Meteorology*, **156**, 41-53.

Koutsias, N., Xanthopoulos, G., Founda, D., Xystrakis, F., Nioti, F., Pleniou, M., Arianoutsou, M. (2012b). On the relationships between forest fires and weather conditions in Greece from long-term national observations (1894-2010). *International Journal of Wildland Observations*, **22**, 493-507.

Maclellan, J.I., & Martell, D.L. (1996). Basing airtankers for forest fire control in Ontario. *Operations Research*, **44**, 677-686.

Martell, D.L, Drysdale, R.J., Doan, G.E., & Boychuk, D. (1984). An evaluation of forest fire initial attack resources. *Interfaces*, **14**, 20-32.

Mavrakis, A., & Salvati, L. (2015). Analyzing the behaviour of selected risk indexes during the 2007 Greek forest fires. *International Journal of Environmental Research*, **9**, 831-840.

Moreira, F., Viedma, O., Arianoutsou, M., Curt, T., N., K., Rigolot, E., Bilgili, E. (2011). Landscape-wildfire interactions in southern Europe: Implications for landscape management. *Journal of Environmental Management*, **92**, 2389-2402.

Pacheco, A.P., Claro, J., Fernandes, P.M., Neufville, R.d., Oliveira, T.M., Borges, J. G., & Rodrigues, J.C. (2015). Cohesive fire management within an uncertain environment: A review of risk handling and decision support systems. *Forest Ecology and Management*, **347**, 1–17.

Pausas, J. G., & Keeley, J. E. (2009). A Burning Story: The Role of Fire in the History of Life. *Bioscience*, **59** (7), 593-601.

Xanthopoulos, G. (2008). Who Should Be Responsible for Forest Fires? Lessons From the Greek Experience. *2nd International Symposium on Fire Economics, Planning, and Policy: A Global View*, 189-201.

Zeferino, J. A. (2020). Optimizing the location of aerial resources to combat wildfires: a case study of Portugal. *Natural Hazards*, **100**, 1195-1213.

Greek Statistical Institute
Proceedings of the 33rd Panhellenic Statistics Conference (2021), pp.345-355

USING THE POWER-EXPECTED-POSTERIOR PRIOR IN SHRINKAGE REGRESSION: A SIMULATION STUDY

G. Tzoumerkas and D. Fouskakis
Department of Mathematics, National Technical University of Athens, Athens, Greece
tzoumg@mail.ntua.gr, fouskakis@math.ntua.gr

ABSTRACT

The Power-Expected-Posterior (PEP) prior gives us a convenient method to deal with Bayesian variable selection problems in normal linear regression models. Under the PEP prior methodology, an initially chosen baseline prior is updated using imaginary data. When dealing with sparse regression scenarios, the selection of the baseline prior is crucial. Shrinkage priors share notable theoretical properties and can be used in regression problems when the number of observations n is smaller than the number of explanatory variables p. By using a shrinkage prior as a baseline prior under the PEP prior methodology, an objective Bayesian prior is created, suitable for n < p problems. In this work we briefly set the formation of the proposed PEP-Shrinkage methodology. In simulated datasets we test the performance of our new class of priors using different shrinkage priors as baseline priors and comparing their results. Additionally, comparisons are made between the PEP-Shrinkage priors and the shrinkage priors without the use of the PEP methodology and we discuss our findings.

keywords: Bayesian variable selection, imaginary training sample, objective priors, shrinkage priors, sparse datasets.

1. Introduction-Bayesian Variable Selection

In this work we consider the variable selection problem under the normal linear regression setup. For every model M_ℓ in model space S_M the sampling distribution is given by

$$f_\ell(\mathbf{y} \mid X_\ell, \boldsymbol{\beta}_\ell, \sigma^2) = f_{N_n}(\mathbf{y}; X_\ell \boldsymbol{\beta}_\ell, \sigma^2 I_n),$$

where $f_{N_d}(\mathbf{y}; \boldsymbol{\mu}, \Sigma)$ is denoting the d-dimensional normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ. Additionally, $\mathbf{y} = (y_1, ..., y_n)^T$ denotes the response data, X_ℓ is the $n \times p_\ell$ design matrix; with p_ℓ denoting the number of explanatory variables under model M_ℓ, $\boldsymbol{\beta}_\ell$ is a vector of length p_ℓ representing the effects of each

covariate on the response variable, I_n is the $n \times n$ identity matrix and σ^2 is the error variance. We assume that $\mathbf{y}$ and the columns of the design matrix of the full model (including all available explanatory variables) have been centered on zero, so there is no intercept in our model. Furthermore, we assume that the number of observations n is smaller than the number of all available explanatory variables p.

Under the Bayesian perspective, we can use posterior odds (Jeffreys, 1961) in order to compare the models M_1 and M_2**:**

$$PO_{12} = \frac{\pi(M_1 \mid \mathbf{y})}{\pi(M_2 \mid \mathbf{y})} = \frac{m_1(\mathbf{y})}{m_2(\mathbf{y})} \times \frac{\pi(M_1)}{\pi(M_2)},$$

where $\pi(M_\ell \mid y)$ is the posterior probability of model M_ℓ, $\pi(M_\ell)$ is the prior probability of M_ℓ and $m_\ell(\mathbf{y})$ is the marginal likelihood of M_ℓ given by

$$m_\ell(\mathbf{y}) = \int f_\ell(\mathbf{y} \mid \boldsymbol{\beta}_\ell, \sigma, M_\ell)\pi(\boldsymbol{\beta}_\ell, \sigma \mid M_\ell)d\boldsymbol{\beta}_\ell d\sigma.$$

In the last expression $\pi(\boldsymbol{\beta}_\ell, \sigma \mid M_\ell)$ denotes the prior distribution of the parameters $(\boldsymbol{\beta}_\ell, \sigma)$ of model M_ℓ. The ratio of any two marginal likelihoods is called Bayes factor (BF_{12}), i.e. $BF_{12} = \frac{m_1(\mathbf{y})}{m_2(\mathbf{y})}$. Then the posterior probability of any model M_ℓ is given by

$$\pi(M_\ell \mid \mathbf{y}) = \frac{m_\ell(\mathbf{y})\pi(M_\ell)}{\sum_{M_k \in S_M} m_k(\mathbf{y})\pi(M_k)}.$$

The model with the highest posterior probability (maximum a posteriori model) is often chosen as the optimal one under the Bayesian model choice problem. For large model spaces, we often use MCMC methods to estimate $\pi(M_\ell \mid \mathbf{y})$. These estimates have the disadvantage of convergence to the true quantities at a slow rate. As an alternative strategy we could use the posterior inclusion probabilities (George et. al., 1993). For each covariate X_j, $j = 1,...,p$, the posterior inclusion probability is defined as

$$\pi(\gamma_j = 1 \mid \mathbf{y}) = \sum_{M_\xi \in S_{M_j}} \pi(M_\xi \mid \mathbf{y}),$$

where γ_j is a binary indicator that takes the value 1 if the covariate X_j belongs to a model and 0 otherwise. In the above expression $S_{M_j} = \{M_\ell \in S_M : \gamma_j = 1\} \subset S_M$ is defined as the set containing all models which include X_j. The significance of the posterior inclusion probabilities can be found in Barbieri et. al. (2004), where it is proven that the median probability model, which is the model containing only the covariates with posterior inclusion probability above 0.5, has better predictive properties than the maximum a posteriori model.

As it is now clear, we must set priors for both the model space and the parameter space of each model. In this work we will focus on the parameter space. Regarding the prior on the model space, for sparsity reasons, we consider the uniform prior on model size which is a special case of the beta-binomial prior (Scott et al., 2010). With respect to the prior distribution on the coefficients in each model little prior information on their regression coefficients can be expected since we are not confident about any given set of regressors as explanatory variables. This argument alone justifies the need for an objective model choice approach in which vague prior information is assumed. Furthermore, we need to use a prior capable to deal with the n < p scenario. Finally, regarding the (common across models) error variance, the reference prior will be used, i.e. $\pi(\sigma^2) \propto \sigma^{-2}$.

2. Power-Expected-Posterior Prior for Sparse Regression

One way to define objective priors is using the Power-Expected-Posterior (PEP) methodology (Fouskakis et al., 2015 and 2016). The PEP prior manages to diminish the effects of the use of random imaginary training data of the Expected-Posterior-Prior (EPP) (Pérez et al., 2002) and simultaneously produces a minimally informative prior, by combining ideas from the power-prior approach of Ibrahim et al. (2000) and the unit-information approach of Kass et al. (1995). Under the normal linear model, the PEP prior is defined as

$$\pi^{\text{PEP}}(\boldsymbol{\beta}_\ell \mid \sigma^2, \delta, \mathrm{X}_\ell^*) = \int \pi^{\text{N}}(\boldsymbol{\beta}_\ell \mid \mathbf{y}^*, \sigma^2, \delta, \mathrm{X}_\ell^*) \mathrm{m}_0^{\text{N}}(\mathbf{y}^* \mid \sigma^2, \delta, \mathrm{X}_0^*) \mathrm{d}\mathbf{y}^* \quad (1)$$

$$\pi^{\text{PEP}}(\sigma^2) = \pi^{\text{N}}(\sigma^2) \propto 1/\sigma^2$$

with

$$\pi^{\text{N}}(\boldsymbol{\beta}_\ell \mid \mathbf{y}^*, \sigma^2, \delta, \mathrm{X}_\ell^*) \propto \mathrm{f}_\ell(\mathbf{y}^* \mid \boldsymbol{\beta}_\ell, \sigma^2, \delta, \mathrm{X}_\ell^*) \pi^{\text{N}}(\boldsymbol{\beta}_\ell \mid \sigma^2, \mathrm{X}_\ell^*) \quad (2)$$

and

$$\mathrm{f}_\ell(\mathbf{y}^* \mid \boldsymbol{\beta}_\ell, \sigma^2, \delta, \mathrm{X}_\ell^*) = \frac{\mathrm{f}_\ell(\mathbf{y}^* \mid \boldsymbol{\beta}_\ell, \sigma^2, \mathrm{X}_\ell^*)^{1/\delta}}{\int \mathrm{f}_\ell(\mathbf{y}^* \mid \boldsymbol{\beta}_\ell, \sigma^2, \mathrm{X}_\ell^*)^{1/\delta} \mathrm{d}\mathbf{y}^*}. \quad (3)$$

In the above equations, we have set $\mathbf{y}^*$ to be the imaginary observations of size n^* and X_ℓ^* the imaginary design matrix of model M_ℓ. In equation (3) the sampling distribution of the imaginary observations is raised to the power of $1/\delta$ and normalized. By doing so, we decrease the effect of the imaginary data. For $\delta = 1$ we have the EPP prior. With the aim of gaining unit information interpretation, we could set $\delta = \mathrm{n}^*$ and to avoid any effect of the choice of imaginary design matrices, we set $\mathrm{n}^* = \mathrm{n}$. By $\pi^{\text{N}}(\boldsymbol{\beta}_\ell \mid \mathbf{y}^*, \sigma^2, \delta, \mathrm{X}_\ell^*)$ we denote the conditional on σ^2 posterior of $\boldsymbol{\beta}_\ell$ using a baseline prior $\pi^{\text{N}}(\boldsymbol{\beta}_\ell \mid \sigma^2, \mathrm{X}_\ell^*)$ and data $\mathbf{y}^*$. In equation (1), $\mathrm{m}_0^{\text{N}}(\mathbf{y}^* \mid \sigma^2, \delta, \mathrm{X}_0^*)$ is the marginal likelihood, evaluated at $\mathbf{y}^*$, of the reference model M_0, given σ^2.

For parsimony reasons, we consider as reference the model with only the intercept. Finally, under model M_ℓ, the marginal likelihood using the baseline prior is given by

$$m_\ell^N(\mathbf{y}^* \mid \sigma^2, \delta, X_\ell^*) = \int f_\ell(\mathbf{y}^* \mid \boldsymbol{\beta}_\ell, \sigma^2, \delta, X_\ell^*)\pi^N(\boldsymbol{\beta}_\ell \mid \sigma^2, X_\ell^*)d\boldsymbol{\beta}_\ell.$$

It is distinct that the choice of the baseline prior is crucial, when applying the PEP methodology. Under our setup, these baseline priors should be capable to deal with the n < p problem. Under the Bayesian perspective, shrinkage priors are often used when n < p. The term shrinkage indicates that the non-true effects are going to shrink towards zero. In Table 1, we mention some often used, shrinkage priors, where τ_j denotes the 'j-th' local shrinkage hyperparameter $(j = 1,..., p_\ell)$ of model M_ℓ and λ denotes the global shrinkage hyperparameter. The global shrinkage hyperparameter, determines the overall sparsity in the whole parameter vector, while the local shrinkage hyperparameter controls the shrinkage of each individual effect. In all cases except the last (ridge g-prior) independent priors for the coefficients of model M_ℓ are used. Furthermore, $C^+(0,\gamma)$ denotes the truncated Cauchy distribution $C(0,\gamma)$, with support $(0,\infty)$ and $IG(\alpha,\beta)$ the Inverse-Gamma distribution with parameters α,β.

By choosing a shrinkage prior, as a baseline prior $\pi^N(\boldsymbol{\beta}_\ell \mid \sigma^2, X_\ell^*)$ in the PEP methodology, a PEP-Shrinkage prior is created. This is an objective prior capable of dealing with $n < p$ scenario, which has a nice interpretation and could manage to combine the advantages of both methodologies mentioned. Furthermore, it offers compatibility across models (Consonni et al., 2008). Finally, impropriety of baseline priors does not cause indeterminacy of Bayes factors.

Under equation (3) the sampling distribution of the imaginary data is given by

$$f_\ell(\mathbf{y}^* \mid X_\ell^*, \boldsymbol{\beta}_\ell, \sigma^2, \delta) = f_{N_{n^*}}(\mathbf{y}^*; X_\ell^*\boldsymbol{\beta}_\ell, \delta\sigma^2 I_{n^*}).$$

We assume a $p_\ell \times p_\ell$ diagonal matrix $\Omega_\ell \equiv \Omega_\ell(\boldsymbol{\theta}_\ell)$, where its 'i-th' main diagonal element is written as an equation of the global shrinkage parameter and the 'i-th' local shrinkage parameter. By $\boldsymbol{\theta}_\ell$ we denote the vector containing all shrinkage hyperparameters of model M_ℓ and by $\pi(\boldsymbol{\theta}_\ell)$ its prior distribution. From Table 1, it is obvious that the shrinkage priors, considered in this paper, can be written as a scale mixture prior of the form:

$$\pi^N(\boldsymbol{\beta}_\ell \mid \boldsymbol{\theta}_\ell, \sigma^2) = f_{N_{p_\ell}}(\boldsymbol{\beta}_\ell; \mathbf{0}, \sigma^2\Omega_\ell).$$

Table 1. *A list of Shrinkage priors*

#	Name	Conditional prior of β_j	Shrinkage hyperparameters
1	Lasso (Park et al., 2008)	$\beta_j \mid \tau_j^2, \sigma^2 \sim N(0, \sigma^2\tau_j^2)$	$\tau_j^2 \mid \lambda^2 \sim \text{Exp}(\frac{\lambda^2}{2})$ $\lambda \sim C^+(0,1)$
2	Horseshoe (Carvalho et al., 2010)	$\beta_j \mid \tau_j, \lambda, \sigma^2 \sim N(0, \sigma^2\lambda^2\tau_j^2)$	$\tau_j \sim C^+(0,1)$ $\lambda \sim C^+(0,1)$
3	Ridge (Hsiang, 1975)	$\beta_j \mid \lambda, \sigma^2 \sim N(0, \sigma^2/\lambda)$	$\lambda \sim C^+(0,1)$
4	Local Student's t (Tipping, 2001)	$\beta_j \mid \tau_j^2, \sigma^2 \sim N(0, \sigma^2\tau_j^2)$	$\tau_j^2 \mid \lambda \sim \text{IG}(\frac{k}{2}, \frac{k}{2\lambda})$ $\lambda \sim C^+(0,1)$ k fixed
5	Elastic Net (Kyung et al., 2010)	$\beta_j \mid \lambda_2, \tau_j^2, \sigma^2 \sim N(0, \frac{\sigma^2}{\lambda_2 + \tau_j^2})$	$\tau_j^2 \mid \lambda_1 \sim \text{Exp}(\frac{\lambda_1^2}{2})$ $\lambda_1, \lambda_2 \sim C^+(0,1)$
6	Beta prime (Bai et al., 2021)	$\beta_j \mid \tau_j^2, \sigma^2 \sim N(0, \sigma^2\tau_j^2)$	$\tau_j^2 \sim \text{Inv} - \text{Beta}(a,b)$ a,b fixed
7	Ridge g-prior (Gupta et al., 2009)	$\boldsymbol{\beta} \mid \lambda, \sigma^2 \sim N_{p_\ell}(0, \sigma^2 V_\ell)$ $V_\ell = g(X_\ell^T X_\ell + \lambda I_{p_\ell})^{-1}$	$g = \max\{n, p_\ell^2\}$ λ fixed

Using equations (1) and (2), we get that the conditional PEP-Shrinkage prior, given σ^2 and $\boldsymbol{\theta}_\ell$, is a multivariate normal distribution

$$\pi^{\text{PEP}}(\boldsymbol{\beta}_\ell \mid \sigma^2, \delta, X_\ell^*, \boldsymbol{\theta}_\ell) = f_{N_{p_\ell}}(\boldsymbol{\beta}_\ell; \mathbf{0}, \sigma^2 V_\ell),$$

with

$$V_\ell = [W_\ell^{-1} - \delta^{-2} X_\ell^{*T} Z_\ell X_\ell^*]^{-1},$$

$$Z_\ell = [\delta^{-2} X_\ell^* W_\ell X_\ell^{*T} + \delta^{-1} I_n]^{-1}$$

and

$$W_\ell = [\delta^{-1} X_\ell^{*T} X_\ell^* + \Omega_\ell^{-1}]^{-1}.$$

Conditionally on the shrinkage hyperparameters $\boldsymbol{\theta}_\ell$ the marginal likelihood of model M_ℓ under the PEP-Shrinkage prior is then given by

$$m_\ell^{PEP}(\mathbf{y} \mid \delta, X_\ell^*, X_\ell, \boldsymbol{\theta}_\ell) = \int \pi^{PEP}(\boldsymbol{\beta}_\ell \mid \sigma^2, \delta, X_\ell^*, \boldsymbol{\theta}_\ell)\pi^N(\sigma^2) f_\ell(\mathbf{y} \mid X_\ell, \boldsymbol{\beta}_\ell, \sigma^2) d\boldsymbol{\beta}_\ell d\sigma^2$$
$$\propto (\det(I_n + X_\ell V_\ell X_\ell^T))^{-1/2} (y^T [I_n + X_\ell V_\ell X_\ell^T]^{-1} y)^{-n/2}.$$

Thus, in cases where the shrinkage hyperparameters of the baseline prior are fixed (e.g. ridge g-prior), the marginal likelihood can be calculated in closed form. The unknown normalizing constant, in the above expression, comes from the improper prior of the error variance, which is common in all compared models and therefore we do not face any indeterminacy issues when calculating the Bayes factor. In all other cases the marginal likelihood is given by

$$m_\ell^{PEP}(\mathbf{y}) \equiv m_\ell^{PEP}(\mathbf{y} \mid \delta, X_\ell^*, X_\ell) = \int m_\ell^{PEP}(\mathbf{y} \mid \delta, X_\ell^*, X_\ell, \boldsymbol{\theta}_\ell)\pi(\boldsymbol{\theta}_\ell) d\boldsymbol{\theta}_\ell.$$

If the dimension of $\boldsymbol{\theta}_\ell$ is one (e.g. ridge prior) we can easily evaluate the above integral numerically. If the dimension of $\boldsymbol{\theta}_\ell$ is greater than one (e.g. horseshoe prior), we perform an MC^3 procedure (Madigan et al., 1995), conditionally on $\boldsymbol{\theta}_\ell$, as in Algorithm 3 of the Appendix of Fouskakis et. al. (2021), where each component of $\boldsymbol{\theta}_\ell$ is generated from its full conditional posterior distribution using a Metropolis-Hastings step.

3. Simulation Study

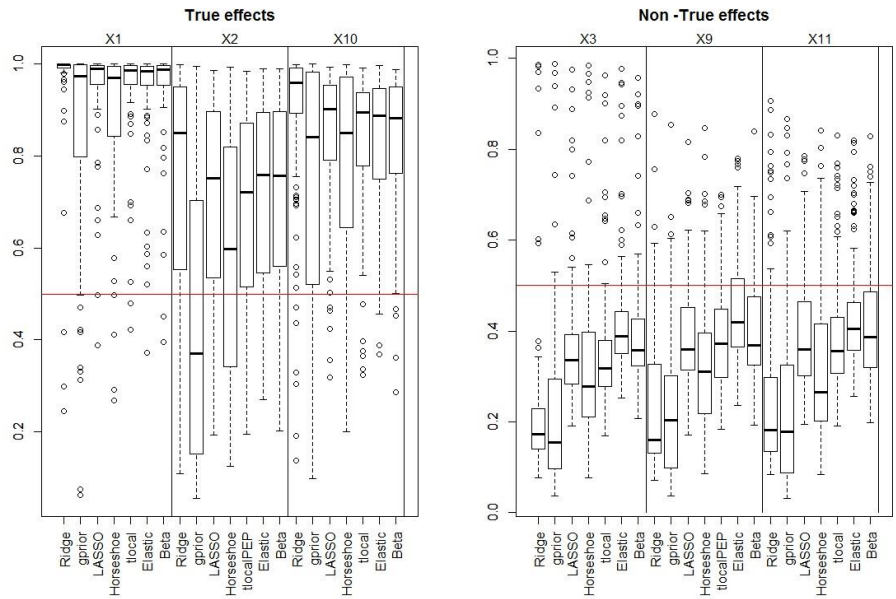
With the intention of testing the PEP-Shrinkage methodology (with $\delta = n = n^*$, $X_\ell^* = X_\ell$ and the reference model to be the null one) we perform multiple simulations, where we use as a baseline prior every shrinkage prior from Table 1 and compare the results. In addition, we contrast the results for five of the PEP-Shrinkage priors with the results obtained by using those shrinkage priors without the PEP-Shrinkage methodology.

We have simulated 100 different samples of length $n = 25$ with $p = 50$ predictors. The values of the explanatory variables have been generated from $N_{50}(\mathbf{0}, \Sigma)$, where the symmetrical matrix Σ has elements $\Sigma_{i,j} = (0.75)^{|i-j|}$, i, j $= 1, \ldots, 50$. For the predictor effects we have set $(\beta_1, \beta_2, \beta_{10})^T = (2, 0.8, 1.5)^T$ and for all of the rest, we set to be equal to 0. For the intercept term we have assumed that $\beta_0 = 0.6$ and finally we have set $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim N_{25}(\mathbf{0}, \sigma^2 I_{25})$, for $\sigma^2 = 1.5$. Finally, we have centered the values of the response variable, as well as the columns of the design matrix on zero.

In Figure 1 (left) we present the boxplots of the marginal posterior inclusion probabilities for the true effects of the 100 different samples for the seven different

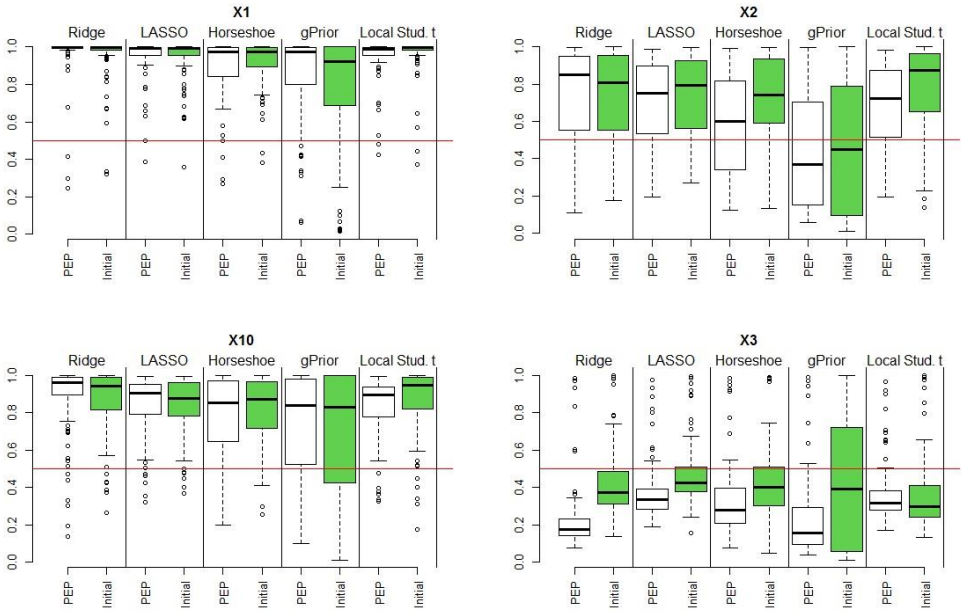
PEP-Shrinkage priors. With respect to the two most influential variables, X_1 and X_{10}, for every baseline prior we obtained high median posterior inclusion probabilities with most of their values to be above 0.5. Furthermore, for these two effects PEP-ridge seems to outperform every other PEP-Shrinkage prior. On the contrary, PEP-g-prior (PEP-ridge g-prior) seems to produce the less satisfactory results. For predictor X_2 the median marginal posterior inclusion probabilities are over the desirable value (0.5) in all the cases except once. As before, PEP-ridge produces the best results, while PEP-g-prior (PEP-ridge g-prior) fails to give posterior inclusion probabilities above 0.5 for most of the occasions.

Figure 1. *Boxplots of posterior inclusion probabilities, across 100 simulated datasets, for the true effects - variables* X_1, X_2, X_{10} *(left) and for some of the non-true effects - variables* X_3, X_9, X_{11} *(right) using the PEP-Shrinkage methodology, for different baseline prior (X-axis).*



For the non-true effects, for brevity reasons, we present results in Figure 1 (right) only for a subset of them; specifically only for variables X_3, X_9 and X_{11}. For every PEP-Shrinkage prior the median marginal posterior inclusion probabilities are below 0.5. It is distinct that, regardless the PEP-Shrinkage prior we have chosen, the non-true effects would have been accepted as effects of the true model only in a very small percentage of occasions. We notice that the PEP-ridge manages to give the smaller posterior inclusion probabilities for most of the occasions with less variability. For the rest of non-true effects, that are not presented here, we get similar results.

Figure 2. *Boxplots of posterior inclusion probabilities, across 100 simulated datasets, for the true effects - variables* X_1, X_2, X_{10} *and for one non-true effect - variable* X_3, *using five PEP-Shrinkage priors and the shrinkage priors without the PEP methodology.*



In Figure 2 we present the boxplots of the posterior inclusion probabilities of the true main effects (X_1, X_2 and X_{10}) as well as for variable X_3. We compare the performance of five different PEP-Shrinkage priors with the performance of the shrinkage priors without applying the PEP methodology. For variable X_1 we get similar results, among all pairwise comparisons. All methods (with and without applying the PEP methodology) correctly identify this variable as a true main effect. For variable X_2 the performance is again similar. When applying the PEP methodology we obtain slightly lower posterior inclusion probabilities under the horseshoe and the local Student's t prior. It is also evident that when using the PEP-ridge g-prior (with and without the PEP methodology) the majority of the values of the posterior inclusion probabilities are below 0.5. Finally, regarding variable X_{10} all methods correctly identify it as a true main effect. Under the PEP methodology we obtain slightly better results under the ridge prior and the ridge g-prior and slightly worse results under the horseshoe and the local Student's t prior. Regarding variable X_3 (non-true effect), the PEP methodology in all cases outperforms its competitors producing more parsimonious answers. Marginal posterior inclusion probabilities, under the PEP-Shrinkage methodology, are in general much lower than the corresponding ones under the shrinkage priors without the PEP methodology, with

also much lower variability. The results for the remaining non-true effects are similar and are omitted for brevity reasons.

5. Discussion

The main goal of the research presented here is to briefly show the model formulation and offer some preliminary results, on simulated datasets, of an objective prior capable of dealing with variable selection problems in normal regression models when the number of observations is smaller than the number of explanatory variables. The proposed PEP-Shrinkage prior manages to combine two methods that often used: the PEP priors and the shrinkage priors. The resulting PEP-Shrinkage prior is an objective Bayesian prior, suitable for $n < p$ problems, that has a nice interpretation, based on imaginary data, and is compatible across models. From the simulated study, presented here, it is evident that under the PEP methodology the true model is correctly identified in the majority of the cases. Furthermore, the PEP methodology seems to produce more parsimonious answers, a property that is desirable on sparse regression problems.

There are several directions of future work. The main aim is to create a unified approach, i.e. a new class of PEP-Shrinkage priors, that includes all of the above mentioned cases (shrinkage baseline priors) as special ones. To achieve this goal we will attempt to write the final PEP prior as a scale mixture of normal distribution with the mixing distribution denoting the different baseline prior. This representation will offer several advantages; faster evaluation of posterior distributions and Bayes factors, under all approaches considered, as well as computational tractability. In addition, mathematical properties of this new class of prior distributions will be checked. Finally, we will investigate the effect of the size of the imaginary data in the proposed methodology, as well as, the effect of the parameter δ. A way to deal with this parameter is by setting a prior distribution on it, as in Fouskakis et al. (2018).

Acknowledgment
This work has received funding from the Research Program PEVE 2020 of the National Technical University of Athens.

ΠΕΡΙΛΗΨΗ

Η Δυναμικά-Μεταγενέστερη-Αναμενόμενη εκ των προτέρων κατανομή (Power-Expected-Posterior (PEP) prior) μας παρέχει μια κατάλληλη μεθοδολογία, στο πρόβλημα της μπεϋζιανής επιλογής μεταβλητών σε μοντέλα γραμμικής παλινδρόμησης, με κανονικά κατανεμημένα σφάλματα. Η μεθοδολογία των PEP εκ των πρότερων κατανομών χρησιμοποιεί διδακτικά δεδομένα για να ανανεώσει μια αρχικά επιλεγμένη πρότερη (baseline prior) κατανομή. Όταν το μέγεθος του δείγματος n είναι μικρότερο του πλήθους των επεξηγηματικών μεταβλητών p, η επιλογή της αρχικής πρότερης κατανομής είναι σημαντική. Οι πρότερες κατανομές συρρίκνωσης (shrinkage priors) έχουν αξιοσημείωτες θεωρητικές ιδιότητες και μπορούν να χρησιμοποιηθούν σε τέτοιες περιπτώσεις. Χρησιμοποιώντας πρότερες κατανομές συρρίκνωσης, ως αρχικές πρότερες κατανομές στην PEP μεθοδολογία, δημιουργείται μια νέα κλάση μπεϋζιανών αντικειμενικών πρότερων κατανομών (PEP-Shrinkage), κατάλληλων για προβλήματα παλινδρόμησης με n < p. Στην εργασία αυτή

παρουσιάζουμε συνοπτικά τη μεθοδολογία των PEP-Shrinkage πρότερων κατανομών. Σε προσομοιωμένα δεδομένα συγκρίνουμε τα αποτελέσματα που παίρνουμε όταν εφαρμόζουμε την PEP-Shrinkage μεθοδολογία, με διαφορετικές πρότερες κατανομές συρρίκνωσης ως αρχικές πρότερες κατανομές. Επιπρόσθετα, γίνονται συγκρίσεις στα αποτελέσματα των PEP-Shrinkage εκ των πρότερων κατανομών με αυτά που λαμβάνουμε από τις πρότερες συρρίκνωσης χωρίς την χρήση της PEP μεθοδολογίας.

REFERENCES

Bai, R. and Ghosh, M. (2021). On the Beta Prime Prior for Scale Parameters in High-Dimensional Bayesian Regression Models. *Statistica Sinica*, **31**, 843-865.

Barbieri, M. and Berger, J. (2004). Optimal predictive model selection. *The Annals of Statistics*, **32**, 870-897.

Carvalho, C.M., Polson, N.G., and Scott, J.G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, **97**, 465-480.

Consonni,G. and Veronese, P. (2008). Compatibility of prior specifications across linear models. *Stat. Sci.*, **23**, 332-353.

Fouskakis, D., Ntzoufras, I. and Draper, D. (2015). Power-expected-posterior priors for variable selection in Gaussian linear models. *Bayesian Analysis*, **10**, 75-107.

Fouskakis, D. and Ntzoufras, I. (2016). Power-conditional-expected priors. Using g-priors with random imaginary data for variable selection. *Journal of Computational and Graphical Statistics*, **25**, 647-664.

Fouskakis, D., Ntzoufras I. and Perrakis K. (2018). Power-expected-posterior priors in generalized linear models. *Bayesian Analysis*, **13**, 721-748.

Fouskakis, D. and Ntzoufras, I. (2021). Power-Expected-Posterior Priors as Mixtures of g-Priors. *Bayesian Analysis* (accepted).

George, E. and McCulloch, R. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881-889.

Gupta, M. and Ibrahim, J. (2009). An information matrix prior for Bayesian analysis in generalized linear models with high dimensional data. *Statistica Sinica*, **19**, 1641-1663.

Hsiang, T. C. (1975). A Bayesian View on Ridge Regression. *The Statistician*, **24**, 267-268.

Ibrahim, J.G. and Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science*, **15**, 46-60.

Jeffreys, H. (1961). Theory of Probability. 3rd Edition, Clarendon Press, Oxford.

Kass, R.E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, **90**, 928-934.

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, **5**, 369 - 411.

Madigan, D., and York, J. (1995). Bayesian Graphical Models for Discrete Data. *International Statistical Review*, **63**, 215-232.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, **103**, 681-687.

Pérez, J.M. and Berger, J.O. (2002). Expected - posterior prior distributions for model selection. *Biometrika*, **89**, 491-511.

Scott, J.G. and Berger, J.O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, **38**, 2587-2619.

Tipping, M.E. (2001). Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning*, **1**, 211-244.