



ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ
ΙΝΣΤΙΤΟΥΤΟ



ΠΑΝΕΠΙΣΤΗΜΙΟ
ΜΑΚΕΔΟΝΙΑΣ



ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

29° ΠΑΝΕΛΛΗΝΙΟ ΣΥΝΕΔΡΙΟ ΣΤΑΤΙΣΤΙΚΗΣ

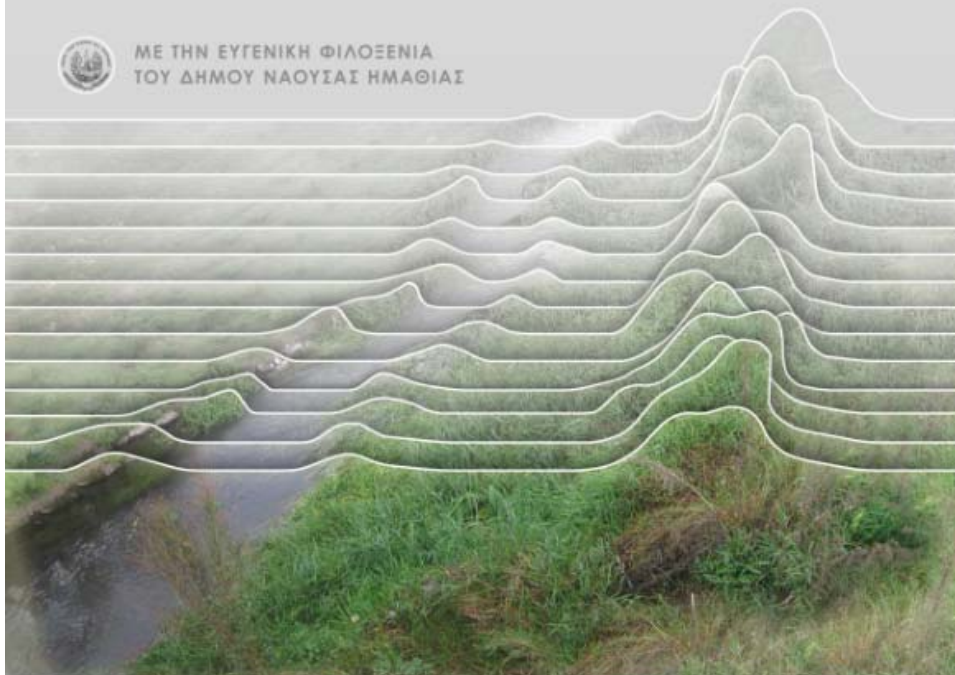
ΑΝΑΛΥΣΗ ΚΙΝΔΥΝΟΥ ΣΤΟ ΠΕΡΙΒΑΛΛΟΝ ΚΑΙ ΤΗΝ ΟΙΚΟΝΟΜΙΑ

4 ΜΑΪΟΥ 2016 - 7 ΜΑΪΟΥ 2016

ΥΠΟΒΟΛΗ ΕΡΓΑΣΙΩΝ ΕΩΣ 31 ΜΑΡΤΙΟΥ (esi-stat@hol.gr)



ΜΕ ΤΗΝ ΕΥΓΕΝΙΚΗ ΦΙΛΟΞΕΝΙΑ
ΤΟΥ ΔΗΜΟΥ ΝΑΟΥΣΙΑΣ ΗΜΑΘΙΑΣ



Περιεχόμενα



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

Π Ρ Α Κ Τ Ι Κ Α

29^{ου} Πανελληνίου
Συνεδρίου Στατιστικής

PROCEEDINGS

of the 29th Panhellenic
Statistics Conference

***Ανάλυση Κινδύνου στο Περιβάλλον και
την Οικονομία***

**Risk Analysis in Environment and
Economics**

Θεσσαλονίκη & Νάουσα, 4-7 Μαΐου 2016



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

Π Ρ Α Κ Τ Ι Κ Α

29^{ου} Πανελληνίου
Συνεδρίου Στατιστικής

Ανάλυση Κινδύνου στο Περιβάλλον και
την Οικονομία

Οργάνωση

ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ

ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ

ΔΗΜΟΣ ΝΑΟΥΣΑΣ

Θεσσαλονίκη & Νάουσα, 4-7 Μαΐου 2016



ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ

Σολωμού 5 (Πλατεία Εξαρχείων)

Τηλ. & Fax 210 33.03.909

Email: esi-stat@hol.gr

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

ISBN: 978-618-80672-9-5

ISSN: 1792-2461

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΡΟΛΟΓΟΣ	σελ. 7
ΧΟΡΗΓΟΙ.....	10
ΠΡΟΓΡΑΜΜΑ ΣΥΝΕΔΡΙΟΥ.....	11
ΕΠΙΤΡΟΠΕΣ ΣΥΝΕΔΡΙΟΥ	19

Εργασίες στα Ελληνικά	
ΑΘΑΝΑΣΙΑΣΗΣ Ι., ΙΩΑΝΝΙΔΗΣ Δ.: Στατιστική ανάλυση ενός μεγάλου συνόλου δεδομένων δεικτών οίνων και εφαρμογή στατιστικών μοντέλων πρόβλεψης δείκτη ποιότητας.....	21
ΑΡΑΠΗΣ Α.Ν., ΜΑΚΡΗ Φ.Σ., ΨΥΛΛΑΚΗΣ Ζ.Μ.: Απόσταση μεταξύ k-ροών επιτυχιών σε αλυσίδα Markov.....	35
ΔΗΜΗΤΡΙΟΥ Ι., ΖΗΣΙΜΟΠΟΥΛΟΥ Ε.-Α.: Ένα μοντέλο ουράς επαναλαμβανόμενων αφίξεων με εφαρμογή στην αποτίμηση απόδοσης του πρωτοκόλλου ελέγχου μετάδοσης πακέτων δεδομένων στο διαδίκτυο.....	43
ΔΗΜΗΤΡΙΟΥ Ι., ΚΑΛΤΣΑΣ Α.-Φ.: Μια ουρά αναμονής με πολλαπλούς υπαλλήλους και ασύγχρονες διακοπές με εφαρμογή στην διαχείριση ενέργειας σε υποδομές φιλοξενίας υπολογιστικών συστημάτων.....	57
ΔΗΜΗΤΡΙΟΥ Ι., ΜΑΡΚΟΥ Μ.: Ένα σύστημα αναμονής με ανυπόμονους πελάτες, σήματα και ομαδικές αναχωρήσεις σε ετερογενές περιβάλλον.....	67
ΘΕΟΔΟΣΙΑΔΟΥ Ο., ΤΣΑΚΛΙΔΗΣ Γ.: Χρήση φίλτρου Kalman με μεταβλητές στο χρόνο διασπορές για την εκτίμηση των αλμάτων αποδόσεων χρηματοοικονομικών μεγεθών.....	76
ΚΕΤΖΑΚΗ Ε., ΡΑΛΛΑΚΗΣ Σ.: Συμβολή των δεικτών περιβαλλοντικής ανομοιότητας στον έλεγχο εκπομπών αερίων ρύπων.....	91
ΚΟΥΝΙΑΣ Σ.: Ιωάννης Παπαδάκης (Jean Papadakis), (1903-1997).....	101
ΜΠΟΖΙΚΑΣ Α., ΠΙΤΣΕΛΗΣ Γ.: Συγκριτική ανάλυση μοντέλων θνησιμότητας για τον ελληνικό πληθυσμό.....	108
ΠΑΠΑΪΩΑΝΝΟΥ Τ.: Απαγόρευση των p-values και των ελέγχων σημαντικότητας.....	123

ΠΡΟΔΡΟΜΙΔΗΣ Π.-Ι., ΠΑΠΑΘΑΝΑΣΙΟΥ Ε.-Β.: Εκτιμήσεις των επιδράσεων εργασίας, κεφαλαίου, επενδύσεων και δημοσίων δαπανών στο περιφερειακό προϊόν: 2001-2011.....	138
ΣΑΧΛΑΣ Α., ΜΠΕΡΣΙΜΗΣ Σ., CHAKRABORTI S.: Ένας ακολουθιακός έλεγχος για την αξιολόγηση της συμφωνίας μεταξύ βαθμολογητών.....	153
ΣΥΜΕΩΝΙΔΟΥ Α., ΠΑΠΑΔΟΠΟΥΛΟΥ Α.: Διατηρησιμότητα και επιτευξιμότητα στο ημιμαρκοβιανό σύστημα σε στοχαστικό περιβάλλον....	168
ΤΣΙΜΠΑΝΟΣ Α., ΑΓΙΑΚΛΟΓΛΟΥ Χ., ΤΣΙΜΠΟΣ Κ.: Η συμπεριφορά των ελέγχων χωρικής εξάρτησης σε μεγάλα δείγματα.....	183
ΦΑΡΜΑΚΗΣ Ν., ΠΑΠΑΤΣΟΥΜΑ Ι.: Κατανομές τριγωνομετρικής έκφρασης: Η περίπτωση της εφαπτομένης.....	194
ΧΑΡΑΛΑΜΠΙΔΗΣ Χ. Α.: Αναθεώρημα του ισοδύναμου των δύο λύσεων των Pascal και Fermat στο κλασικό πρόβλημα της διανομής μεριδίων.....	203
ΧΑΡΙΤΙΔΟΥ Ο.: Ο Μικρός Θαλής και τα Στοχαστικά Μαθηματικά.....	210
ΧΑΤΖΗΠΑΝΤΕΛΗΣ Θ., ΣΩΤΗΡΟΓΛΟΥ Μ., ΓΚΟΥΡΑΜΑΝΗ Ε.: Τι είναι και τι εκτιμά το σύστημα ECTS; Ανάλυση του φόρτου εργασίας και της επίδοσης των φοιτητών του έτους 2014-2015 του τμήματος Πολιτικών Επιστημών Α.Π.Θ.....	219
Εργασίες στα Αγγλικά	
BASSIAKOS Y. C.: Estimation of simple linear regression coefficients with stratified data.....	235
BATZIOU E., GRAVANIS M., KARADIMOS P., KOULIKAS D., MPATZIAKAS A., PARAGKAMIAN S., TONIDIS E., XANTHOPOULOU G.: Comparison of correlation measures for detection of brain network connectivity alterations during epileptiform discharges.....	245
CHASIOTIS V., KOUNIAS S., FARMAKIS N.: Fractional factorial designs and orthogonal arrays.....	260
CHATZOPOULOS ST. A., KOLYVA- MACHERA F., CHARALAMBOUS H.: Analysis on the gender factor and the academic performance of students of the faculty of Sciences of Aristotle University of Thessaloniki	271
CHOROZOGLU D., D. KUGIUMTZIS, E. PAPADIMITRIOU: Seismic hazard assessment of earthquakes in Greece using directed earthquake networks.....	283
KOUNIAS S. : The usefulness of optimum experimental designs.....	299

KOUROUKLAS CH., PAPADIMITRIOU E., KARAKOSTAS V., TSAKLIDIS G.: Long range correlations and probability distributions of earthquake interevent time in Greece.....	307
LOUMPONIAS K., VRETOS N., DARAS P., TSAKLIDIS G.: Using Kalman filter and Tobit Kalman filter in order to improve the motion recorded by Kinect Sensor II.....	322
MANGIRA O., VASILADIS G., PAPADIMITRIOU E., TSAKLIDIS G.: Using a Weibull-type conditional intensity function for the linked stress release model- application for the Corinth Gulf.....	335
MANTZOUNI A., NTZOUFRAS I., KATERI M.: Bayesian Inference and Model Selection for Association models in Contingency tables.....	345
MESIMERI M., KARAKOSTAS V.: Earthquake inter-event times distribution: Application to Corinth Gulf.....	360
TASIAS K. A., NENES G.: Optimal adaptive maintenance and quality control model for processes subject to failure and multiple quality shifts affecting both location and scale.....	373

ΠΡΟΛΟΓΟΣ

Το 29^ο Πανελλήνιο Συνέδριο Στατιστικής διοργανώθηκε από το Ελληνικό Στατιστικό Ινστιτούτο (Ε.Σ.Ι.) στην Θεσσαλονίκη και τη Νάουσα την περίοδο 4-7 Μαΐου 2016 σε συνεργασία με το Αριστοτέλειο Πανεπιστήμιο, το Πανεπιστήμιο Μακεδονίας και το Δήμο Νάουσας. Το θέμα του Συνεδρίου ήταν *Ανάλυση Κινδύνου στο Περιβάλλον και την Οικονομία*.

Έλαβαν μέρος 150 Σύεδροι, από τους οποίους 88 ήταν φοιτητές και 10 συνοδεύοντα μέλη από την Ελλάδα και την Κύπρο. Η συμμετοχή των προπτυχιακών και των μεταπτυχιακών φοιτητών υπήρξε ικανοποιητική, συγκρίσιμη με το παρελθόν.

Στο Συνέδριο προσκεκλημένοι ομιλητές ήταν ο κ. Ιάκωβος Γκανούλης, Ειδικός Γραμματέας Υδάτων του Υπουργείου Περιβάλλοντος & Ενέργειας (Υ.ΠΕ.Ν.) και Ομότιμος Καθηγητής της Πολυτεχνικής Σχολής του Τμήματος Πολιτικών Μηχανικών του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης, και ο *Ευτύχιος Σαρτζετάκης*, Αναπληρωτής Πρύτανης και Καθηγητής του Τμήματος Οικονομικών Επιστημών του Πανεπιστημίου Μακεδονίας.

Το Επιστημονικό Πρόγραμμα ήταν ιδιαίτερα πλούσιο και περιελάμβανε ογδόντα (80) ανακοινώσεις κ δεκαέξι (16) σε παράλληλες Συνεδρίες σε θέματα Πειραματικών σχεδιασμών, Βιοστατιστικής, Πρόβλεψης-Χρονοσειρών, Στατιστικής στην Πληροφορική και την Οικονομία, Εφαρμοσμένης Στατιστικής, Εφαρμοσμένων Πιθανοτήτων, Στοχαστικών Διαδικασιών, Χρονοσειρών κτλ.

Η εναρκτήρια τελετή του Συνεδρίου πραγματοποιήθηκε στο αμφιθέατρο «ΕΡΙΑ» του πολυχώρου πολιτισμού «ΒΕΤΛΑΝΣ» του Δήμου Νάουσας. Την έναρξη του Συνεδρίου κήρυξε ο Πρόεδρος του ΕΣΙ τέως Αναπληρωτής Καθηγητής Χαράλαμπος Δαμιανού. Χαιρετισμούς απηύθυναν η Πρόεδρος της Οργανωτικής Επιτροπής Αν. Καθηγήτρια Φωτεινή Κολυβά-Μαχαίρα, ο Δήμαρχος της Νάουσας κ. Νικόλαος Κουτσογιάννης, ο Αν. Πρύτανης του Πανεπιστημίου Μακεδονίας κ. Ευτύχιος Σαρτζετάκης κι ο Αν. Πρόεδρος του Τμήματος Μαθηματικών του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης κ. Γεώργιος Τσακλίδης. Η τελετή έναρξης συνεχίστηκε με την ομιλία σε ολομέλεια του κ. Στρατή Κουνιά ο οποίος μίλησε για τον Ιωάννη Παπαδάκη (Jean Papadakis, 1903-1997) και τις προσκεκλημένες ομιλίες διαλέξεις των προσκεκλημένων ομιλητών *ων Ιάκωβου Γκανούλη* με τίτλο «Ανάλυση Επικινδυνότητας και Στατιστική: Μαθηματικές θεωρίες ή χρήσιμα επιχειρησιακά εργαλεία;» και του *Ευτύχιου Σαρτζετάκη*, με τίτλο «Οικονομικά της κλιματικής αλλαγής: Ο σημαντικός ρόλος της διαχείρισης του κινδύνου και της αβεβαιότητας».

Στα πλαίσια του 29^{ου} Πανελληνίου Συνεδρίου Στατιστικής πραγματοποιήθηκαν πέντε (5) ειδικές συνεδρίες με θέματα: *Στατιστική στην κλιματολογία, Στατιστική στη Σεισμολογία, Στοχαστικά Αναθεωρήματα, Στατιστικός Έλεγχος Ποιότητας και Στατιστική στη Δευτεροβάθμια Εκπαίδευση* και μία στρογγυλή τράπεζα με καθηγητές της Ε.Μ.Ε. του παραρτήματος Ημαθίας με θέμα «Το μάθημα της Στατιστικής στη Δευτεροβάθμια εκπαίδευση, σημερινή κατάσταση-μελλοντικές εξελίξεις-προτάσεις». Προσφέρθηκε, τέλος, Σεμινάριο δέκα (10) ωρών με θέμα «**Βιομηχανική Στατιστική - Στατιστικός Έλεγχος Διεργασιών**» από τους Σωτήριο Μπερσίμη, Επ. Καθηγητή του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς, τον Αθανάσιο Ρακιτζή Λέκτορα του Τμήματος Μαθηματικών του Πανεπιστημίου του Αιγαίου και τον Αθανάσιο Σαχλά, επιστημονικό συνεργάτη του τμήματος Νοσηλευτικής του Πανεπιστημίου Πελοποννήσου και εργαστηριακό διδάσκοντα στο Μεταπτυχιακό Πρόγραμμα Σπουδών «Εφαρμοσμένη Στατιστική» του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς.

Οι κοινωνικές εκδηλώσεις του συνεδρίου ήταν πλούσιες και ποικίλες. Τα συνοδεύονται μέλη πραγματοποίησαν περίπατο στην πόλη της Νάουσας με ξενάγηση από την κ. Τζωρτζίνα Αθανασίου, μέλος της Τοπικής Οργανωτικής Επιτροπής και επισκέψεις στην Εθνική Βιβλιοθήκη Αργυρουπόλεως της Ευξείνου Λέσχης Ποντίων Νάουσας, όπου φυλάσσονται σπάνια χειρόγραφα του ποντιακού ελληνισμού και στο Μουσείο Παραδοσιακών Δρώμενων της Αποκρίας στη Νάουσα «Γενίτσαροι και Μπούλες» με ξενάγηση από τον Πρόεδρο του Ομίλου κ. Άρη Τόσιο. Το απόγευμα της Πέμπτης 5 Μαΐου 2016 όλοι οι σύνεδροι και τα συνοδεύονται μέλη πραγματοποίησαν επίσκεψη στο οινοποιείο VAENI, όπου γνώρισαν τα μυστικά της οινοποιείας. Την επόμενη μέρα οι κοινωνικές εκδηλώσεις του Συνεδρίου περιελάμβαναν επίσκεψη και ξενάγηση στους μακεδονικούς τάφους των Λευκαδίων και συγκεκριμένα στον τάφο των Ανθεμίων και εκδρομή στα λουτρά Πόζαρ. Το επίσημο δείπνο του Συνεδρίου έγινε στο άλσος του Αγίου Νικολάου στη Νάουσα.

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

Οι παρατηρήσεις και τα σχόλια των κριτών, σύμφωνα με την πάγια πολιτική που ακολουθεί το Ε.Σ.Ι., αφορούν κυρίως στον τρόπο παρουσίασης της εργασίας και στην παρουσία ή όχι τυπογραφικών και σοβαρών επιστημονικών λαθών. Οι εργασίες πρέπει να έχουν ικανό στατιστικό περιεχόμενο, να αναδεικνύουν το πρόβλημα που μελετούν, να μην περιορίζονται μόνο σε Περιγραφική Στατιστική κτλ. Για το σκοπό αυτό

υπάρχουν κριτήρια δημοσίευσης εργασιών στα Πρακτικά του Ε.Σ.Ι. τα οποία είναι αναρτημένα στην ιστοσελίδα του Ε.Σ.Ι.: www.esi-stat.gr. Όλες οι εργασίες, για τις οποίες ζητήθηκε αναθεώρηση, κρίθηκαν εκ νέου από τους κριτές ή από τους υπεύθυνους έκδοσης των πρακτικών.

Συνολικά υποβλήθηκαν τριάντα τέσσερις (34) εργασίες, από τις οποίες δύο (2) εργασίες ανακλήθηκαν από τους συγγραφείς μετά την πρώτη αξιολόγηση και δύο (2) απορρίφθηκαν, σύμφωνα με απόφαση του Δ.Σ. και ύστερα από σχετική πρόταση των κριτών και των υπευθύνων της έκδοσης των Πρακτικών. Ως κριτές των εργασιών συνεργάστηκαν οι: Δ. Αντζουλάκος, Φ. Βαλλιανάτος, Ι. Βόντα, Β. Γεωργίου, Χ. Δαμιανού, Σ. Δασκαλάκη, Γ. Δονάτος, Χ. Ευαγγελάρας, Δ. Ιωαννίδης, Α. Καλαματιανού, Α. Καραγρηγορίου, Δ. Καρλής, Φ. Κολυβά-Μαχαίρα, Δ. Κουγιουμτζής, Χ. Κουκουβίνος, Β. Κούτρας, Μ. Κούτρας, Δ. Λέκας, Σ. Μαλεφάκη, Π. Μαραβελάκης, Λ. Μελικοτσίδου, Α. Μπουρνέτας, Ν. Παπαδάτος, Γ. Παπαδόπουλος, Τ. Παπαϊωάννου, Δ. Παπαναστασίου, Α. Σαχλάς, Γ. Τζαβελάς, Σ. Τρέβεζας, Γ. Τσακλίδης, Α. Τσιμπάνος, Ν. Φαρμάκης, Χ. Χαραλαμπίδης. Η Επιτροπή Έκδοσης Πρακτικών του ΕΣΙ εκφράζει τις ευχαριστίες της προς όλους τους κριτές για την επιμελημένη και προσεκτική αξιολόγηση των εργασιών.

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

Το Διοικητικό Συμβούλιο του Ε.Σ.Ι αισθάνεται την ανάγκη να ευχαριστήσει το Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, το Πανεπιστήμιο Μακεδονίας, το Δήμο Νάουσας και την Οργανωτική Επιτροπή για την πολύ καλή οργάνωση και προσφορά τους.

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

ΕΚ ΜΕΡΟΥΣ ΤΟΥ Δ.Σ. ΤΟΥ Ε.Σ.Ι

Οι υπεύθυνοι Έκδοσης Πρακτικών 28^{ου} Συνεδρίου

Γεώργιος Ηλιόπουλος	Χαράλαμπος Δαμιανού	Σόνια Μαλεφάκη
Τάκης Παπαϊωάννου	Χαράλαμπος Χαραλαμπίδης	

ΧΟΡΗΓΟΙ ΣΥΝΕΔΡΙΟΥ



ΠΡΟΓΡΑΜΜΑ ΣΥΝΕΔΡΙΟΥ

	Τετάρτη 4 Μαΐου 2016
	ΘΕΣΣΑΛΟΝΙΚΗ (ΑΙΘΟΥΣΑ Μ2 ΤΜΗΜΑΤΟΣ ΜΑΘΗΜΑΤΙΚΩΝ ΑΠΘ)
9:15	Έναρξη του Συνεδρίου-Χαιρετισμοί
	ΣΤΑΤΙΣΤΙΚΗ ΣΤΗΝ ΕΚΠΑΙΔΕΥΣΗ (Προεδρεύων: Π. Μωυσιάδης)
9:30	St. A. Chatzopoulos, F. Kolyva-Machera, H. Charalambous: A meta-analysis on the gender factor and the academic performance of students of the Faculty of Sciences of Aristotle University of Thessaloniki
10:00	Δ.-Ε. Γκόγκος, Μ. Ιασωνίδου, Χ. Παλικρούσης, Ε. Φαρμακιώτη, Θ. Χήρας: Ομαδοποίηση των πτυχιούχων του Τμήματος Μαθηματικών του Α.Π.Θ. ως προς διάφορες παραμέτρους
10:15	ΔΙΑΛΕΙΜΜΑ
	ΣΤΑΤΙΣΤΙΚΗ ΣΤΗΝ ΚΑΙΜΑΤΟΛΟΓΙΑ (Προεδρεύουσα: Φ. Κολυβά-Μαχαίρα)
10:30	Χ. Αναγνωστοπούλου: Ανάλυση καμπυλών ROC και εφαρμογή τους σε δεδομένα βροχόπτωσης στην περιοχή της Ελλάδας
10:50	Κ. Τολίκα, Π. Μαχαίρας: Ημερήσια, μηνιαία και ετήσια στατιστική ανάλυση του βροχομετρικά ακραίου έτους 2014
11:10	Π. Βενετσάνου: Ημερήσια, μηνιαία και ετήσια στατιστική ανάλυση του βροχομετρικά ακραίου έτους 2014
11:30	Γ. Λάζογλου: Περιοχική μελέτη της επίδρασης των κλιματικών αλλαγών στην ελληνική αμπελουργία
11:50	Κ. Βελίκου: Στατιστική αξιολόγηση των αλλαγών των παραμετροποιήσεων του δυναμικού μοντέλου RegCM4 για τη θερμοκρασία στην ευρύτερη περιοχή της Μεσογείου
12:10	ΔΙΑΛΕΙΜΜΑ
	ΣΤΑΤΙΣΤΙΚΗ ΣΤΗ ΣΕΙΣΜΟΛΟΓΙΑ (Προεδρεύων: Γ. Τσακλίδης)
12:30	Δ. Χορόζογλου: Εκτίμηση σεισμικής επικινδυνότητας των σεισμών στην Ελλάδα με τη χρήση κατευθυνόμενων γραφημάτων
12:50	Ο. Μαγγίρα: Η χρήση της κατανομής Weibull για τα χρονικά διαστήματα μεταξύ σεισμών στον Κορινθιακό κόλπο με βάση το συζευγμένο μοντέλο απελευθέρωσης τάσης
13:10	Χ. Κουρούκλας: Συσχετίσεις μακράς διάρκειας των χρονικών διαστημάτων μεταξύ ισχυρών σεισμών
13:30	Μ. Μεσημέρι: Στατιστική διερεύνηση της εξέλιξης σεισμικών εξάρσεων στον Κορινθιακό κόλπο
14:00	ΔΙΑΚΟΠΗ

ΝΑΟΥΣΑ (ΑΜΦΙΘΕΑΤΡΟ "ΕΡΙΑ")	
16:00-17:00	Εγγραφή Συνεδρών – Διανομή συνεδριακού υλικού
17:00	Έναρξη του Συνεδρίου - Χαιρετισμοί
	ΟΜΙΛΙΕΣ ΣΕ ΟΛΟΜΕΛΕΙΑ (Προεδρεύων: Χ. Δαμιανού)
17:30	Σ. Κουνιάς: Ιωάννης Παπαδάκης (Jean Papadakis) (1903-1997)
	ΠΡΟΣΚΕΚΛΗΜΕΝΟΙ ΟΜΙΛΗΤΕΣ
18:00	Ιάκ. Γκανούλης: Ανάλυση Επικινδυνότητας και Στατιστική: Μαθηματικές θεωρίες ή χρήσιμα επιχειρησιακά εργαλεία;
19:00	Ε. Σαρτζετάκης: Οικονομικά της κλιματικής αλλαγής: Ο σημαντικός ρόλος της διαχείρισης του κινδύνου και της αβεβαιότητας
20:00	WELCOME COCKTAIL

Πέμπτη 5 Μαΐου 2016		
	ΑΜΦΙΘΕΑΤΡΟ «ΕΡΙΑ»	ΑΙΘΟΥΣΑ Β
	ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΑΕΙ (Προεδρεύουσα: Σ. Μαλεφάκη)	ΘΕΩΡΙΑ ΟΥΡΩΝ - ΜΑΡΚΟΒΙΑΝΑ ΜΟΝΤΕΛΑ (Προεδρεύων: Σ. Κουνιάς)
9:00	Ρ. Λύκου, Μ. Μπαϊράμι, Δ. Νάστου, Μ. Τσικαλοπούλου: Μελέτη των επιδόσεων των πτυχιούχων του Τμήματος Μαθηματικών ΑΠΘ από το 1992 έως το 2015	Ι. Δημητρίου, Μ. Μάρκου: Ένα σύστημα αναμονής με ανυπόμονους πελάτες, σήματα και ομαδικές αναχωρήσεις σε ετερογενές περιβάλλον
9:20	Κ. Ασλάνογλου, Β. Ελευθεριάδης, Ν. Κολώκας, Θ. Κωστακίδης, Α. Φωτιάδης: Αξιολόγηση αποφοίτων μαθηματικών Α.Π.Θ. με ανάλυση σε κύριες συνιστώσες και αντίστοιχη συσταδοποίηση	Ι. Δημητρίου, Ε.Α. Ζησιμοπούλου: Ένα μοντέλο ουράς επαναλαμβανόμενων αφίξεων με εφαρμογή στην αποτίμηση απόδοσης του πρωτοκόλλου ελέγχου μετάδοσης πακέτων δεδομένων στο διαδίκτυο

9:40	Α. Καράκος, Δ. Καράκος: Διαχρονική έρευνα πρωτοετών φοιτητών του Τμήματος ΗΜ&ΜΥ του ΔΠΘ όσον αφορά στην επιλογή κατεύθυνσης σπουδών	Ι. Δημητρίου, Α.-Φ. Καλτσάς: Μια ουρά αναμονής με πολλαπλούς υπαλλήλους και ασύγχρονες διακοπές με εφαρμογή στην διαχείριση ενέργειας σε υποδομές φιλοξενίας υπολογιστικών συστημάτων
10:00	Δ.Α. Καλαμάρας, Α.Γ. Καλαματιανού: Ο ρόλος των προσδοκιών των φοιτητών και της ενσωμάτωσής τους στο πανεπιστημιακό περιβάλλον στην ικανοποίηση από τις σπουδές και στο χρόνο αποφοίτησης: Μια προσέγγιση με μοντέλα δομικών εξισώσεων	Α.Ν. Αράπης, Φ.Σ. Μακρή, Ζ.Μ. Ψυλλάκης: Απόσταση μεταξύ k -ροών επιτυχιών σε αλυσίδα Markov
10:20	Θ. Χατζηπαντελής, Μ. Σωτήρογλου, Ε. Γκουραμάνη: Τι είναι και τι εκτιμά το σύστημα ECTS; Ανάλυση του φόρτου εργασίας και της επίδοσης των φοιτητών του έτους 2014-2015 του Τμήματος Πολιτικών Επιστημών ΑΠΘ	Α. Συμεωνίδου, Α. Παπαδοπούλου: Διατηρησιμότητα και επιτευξιμότητα σε μη ομογενή ημιμαρκοβιανά συστήματα σε στοχαστικό περιβάλλον
09:00-10:45	ΣΕΜΙΝΑΡΙΟ Βιομηχανική Στατιστική-Στατιστικός Έλεγχος Διεργασιών	
10:40	ΔΙΑΛΕΙΜΜΑ - ΚΑΦΕΣ	
	ΑΜΦΙΘΕΑΤΡΟ «ΕΡΙΑ»	ΑΙΘΟΥΣΑ Β
	ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ ΚΑΙ ΕΦΑΡΜΟΓΕΣ (Προεδρεύων: Χ. Χαράλαμπιδης)	ΣΤΑΤΙΣΤΙΚΗ (Προεδρεύων: Δ. Ιωαννίδης)
11:10	L.A. Wright: Dependence in stock and option model pricing	Χ. Μέρκατας, Σ. Χατζησπύρος: Εξαρτημένες μίξεις από geometric weights priors
11:30	D.G. Konstantinides, J. Li: Asymptotic ruin probabilities for a multidimensional renewal risk model with multivariate regularly varying claims	Α. Μπατσίδης, Μ.Δ. Jiménez-Gamero: Εκτιμητές πληθυσμιακών παραμέτρων απαριθμητών κατανομών που στηρίζονται στην εμπειρική πιθανογεννήτρια συνάρτηση: Ιδιότητες και εφαρμογές
11:50	N.D. Macheras, S.M. Tzaninis: The problem of the preservation of the distribution of compound mixed renewal process under a change of measure	Ι.Α. Koutrouvelis: Estimation for compound Poisson-gamma and Poisson-inverse Gaussian distributions
12:10	Ο. Θεοδοσιάδου, Γ. Τσακλίδης: Χρήση κατάλληλου φίλτρου Kalman με μεταβλητές στο χρόνο διασπορές για την εκτίμηση των αλμάτων χρηματιστηριακών αποδόσεων	Ο. Davidov, Γ. Πιopoulos: Semiparametric ordered inference for conditional distributions
12:30	Κ. Λουμπόνιας, Ν. Βρέτος, Π. Δάρας, Γ. Τσακλίδης: Χρήση φίλτρου Kalman και Tobit Kalman για τη βελτίωση της κίνησης της κάμερας Microsoft Kinect	V.L. Georgiou, S. Pouloupoulou, A. Karamitrou, C. Kotopoulos: Challenges and solutions using R for business analytics – The IRI Assortment Optimization case
09:00-10:45	ΣΕΜΙΝΑΡΙΟ Βιομηχανική Στατιστική-Στατιστικός Έλεγχος Διεργασιών	

13:00	ΜΕΣΗΜΒΡΙΝΗ ΔΙΑΚΟΠΗ Ελαφρύ γεύμα	
	ΠΕΙΡΑΜΑΤΙΚΟΙ ΣΧΕΔΙΑΣΜΟΙ (Προεδρεύων: Θ. Χατζηπαντελής)	
15:00	Χ. Ευαγγελάρας: Κατασκευή βέλτιστων υπερκορεσμένων σχεδιασμών δύο επιπέδων	
15:20	Φ. Κολυβά-Μαχαίρα, Στ. Α. Χατζόπουλος: Φ-βελτιστοποίηση s^m παραγοντικών σχεδιασμών με $N^p \bmod s^2$ παρατηρήσεις	
15:40	V. Chasiotis, S. Kounias, N. Farmakis: Fractional factorial designs and orthogonal arrays	
	ΣΤΑΤΙΣΤΙΚΗ ΣΤΗ ΣΕΙΣΜΟΛΟΓΙΑ (Προεδρεύων: Γ. Τσακλίδης)	ΧΡΟΝΟΣΕΙΡΕΣ (Προεδρεύουσα: Φ. Κολυβά-Μαχαίρα)
16:00	D. Chorozoglou, D. Kugiumtzis, E. Papadimitriou: Seismic hazard assessment of earthquakes in Greece using directed earthquake networks	Π. Προδρομίδης, Ε. Παπαθανασίου: Εκτιμήσεις των επιδράσεων εργασίας, κεφαλαίου, επενδύσεων και δημοσίων δαπανών στο περιφερειακό προϊόν: 2001-2011
16:20	O. Mangira, G. Vasileiadis, E. Papadimitriou: The use of Weibull distribution for the earthquake interevent times in the Corinth Gulf based on the Linked Stress Release Model.	C. Koutlis and D. Kugiumtzis: The effect of hidden source on the estimation of connectivity networks from time series
16:40	Ch. Kourouklas, E. Papadimitriou, G. Tsaklidis, V. Karakostas: Long term correlations of earthquakes inter-event times in Greece	Κ. Καλούδης, Σ. Χατζησπύρος: Μια μπεύζιανή προσέγγιση στη μείωση δυναμικού θορύβου
17:00	Μ. Μεσημέρη, Β. Καρακώστας: Κατανομές των χρόνων μεταξύ διαδοχικών σεισμών: Εφαρμογή στον Κορινθιακό κόλπο	
15:00-17:15	ΣΕΜΙΝΑΡΙΟ Βιομηχανική Στατιστική-Στατιστικός Έλεγχος Διαδικασιών	
17:20	ΔΙΑΛΕΙΜΜΑ - ΚΑΦΕΣ	
	ΣΤΟΧΑΣΤΙΚΑ ΑΝΑΘΕΩΡΗΜΑΤΑ (Προεδρεύων: Ι. Κουτρουβέλης)	
17:50	Χ. Χαραλαμπίδης: Διανομή μεριδίων στοιχήματος σε διακοπείσα σειρά παιγνιδιών	

18:10	T. Παπαϊωάννου: Απαγόρευση των p-values και των ελέγχων σημαντικότητας	
18:30	S. Kounias: The usefulness of optimum experimental designs	
18:50	T. Cacoullos: Distributions of random points on spheres: Arcsine and arctangent (Cauchy) distributions	

	Παρασκευή 6 Μαΐου 2016	
	ΝΑΟΥΣΑ (ΑΙΘΟΥΣΕΣ "ΕΡΙΑ")	
	ΑΜΦΙΘΕΑΤΡΟ «ΕΡΙΑ»	ΑΙΘΟΥΣΑ Β
	ΣΤΑΤΙΣΤΙΚΟΣ ΕΛΕΓΧΟΣ ΠΟΙΟΤΗΤΑΣ (Προεδρεύων: Σ. Μπερσίμης)	ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ (Προεδρεύων: Ν. Φαρμάκης)
9:30	Π. Οικονόμου, Σ. Μπερσίμης: Διαγράμματα ελέγχου στην περίπτωση μεταβολής του δειγματοληπτικού σχεδιασμού μεταξύ των φάσεων I και II	Γ. Λιάγκας, Κ. Μαμμάς, Δ.Φ. Λέκκας: Σύγκριση μεθόδων εκτίμησης παραμέτρων για την ανάλυση ακραίων τιμών σε δεδομένα βροχόπτωσης
9:50	Α. Ρακιτζής, Π. Μαραβελάκης, P. Castagliola: Διαγράμματα ελέγχου για την παρακολούθηση διογκωμένων διεργασιών	Ε. Κετζάκη, Σ. Ραλλάκης: Συμβολή της περιβαλλοντικής ανομοιοτητας στον έλεγχο προβλεπόμενων εκπομπών αέριων ρύπων
10:10	K.A. Tasias, G. Nenes: Optimal adaptive maintenance and quality control model for processes subject to failure and multiple quality shifts affecting both location and scale	Α. Τσιμπάνος, Χ. Αγιακλόγλου, Κ. Τσίμπος: Η συμπεριφορά των ελέγχων χωρικής εξάρτησης σε μεγάλα δείγματα
10:30	K. Bourazas, P. Tsiamyrtzis: Bayesian self-starting methods to detect persistent shifts in statistical process control & monitoring	
10:50	ΔΙΑΛΕΙΜΜΑ - ΚΑΦΕΣ	
	ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ (Προεδρεύων: Δ. Κουγιουμτζής)	ΣΤΑΤΙΣΤΙΚΑ ΜΟΝΤΕΛΑ ΚΑΙ ΕΦΑΡΜΟΓΕΣ (Προεδρεύων: Ε. Αγγελής)
11:20	Ν. Φαρμάκης, Ε. Κετζάκη: Προσδιορισμός διμεταβλητής συνάρτησης με χρήση συστηματικής δειγματοληψίας	Ι. Αθανασιάδης, Δ. Ιωαννίδης: Χρήση στατιστικών μοντέλων σε υπερμεγέθη δεδομένα (big data) για πρόβλεψη ποιοτικών δεικτών οίνων

11:40	Y.C. Bassiakos: Estimation of simple linear regression coefficients with stratified data	Α. Μανάτος, Σ. Μαλεφάκη, Β. Κούτρας: Μοντελοποίηση και βελτιστοποίηση μέτρων διαθεσιμότητας και απόδοσης συστημάτων με πολλαπλά στάδια υποβάθμισης και εφεδρεία
12:00	Ε. Μήκα: Η πολλαπλή παλινδρόμηση ως εργαλείο εκτίμησης καταγιδόφρου δραστηριότητας	Α. Μποζίκας, Γ. Πιτσέλης: Συγκριτική ανάλυση μοντέλων θνησιμότητας για τον ελληνικό πληθυσμό
11:15-12:30	ΣΕΜΙΝΑΡΙΟ Βιομηχανική Στατιστική-Στατιστικός Έλεγχος Διεργασιών	
12:30	Ελαφρύ γεύμα	
13:00	ΕΚΔΡΟΜΗ	
21:00	ΕΠΙΣΗΜΟ ΔΕΙΠΝΟ	

	Σάββατο 7 Μαΐου 2017	
	ΝΑΟΥΣΑ (ΑΙΘΟΥΣΕΣ "ΕΡΙΑ")	
	ΒΙΟΣΤΑΤΙΣΤΙΚΗ (Προεδρεύων: Τ. Παπαϊωάννου)	ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ (Προεδρεύων: Δ. Παπαναστασίου)
10:00	D. Stogiannis, F. Siannis, L. Meligkotsidou: Meta-analysis of time-to-event studies with small treatment effects	Ν. Φαρμάκης, Ι. Παπατσούμα: Κατανομές τριγωνομετρικής έκφρασης: Η περίπτωση της εφαπτομένης
10:20	Σ. Μπερσίμης, Α. Σαγλάς, S. Chakraborti: Ένας ακολουθιακός έλεγχος για την αξιολόγηση της συμφωνίας μεταξύ βαθμολογητών	Α. Χατζοπούλου, Β. Καραγιάννης, Π. Μωυσιάδης: Η εξέλιξη του οικονομικού δικτύου του Χρηματιστηρίου Αξιών Αθηνών από το 2007 ως το 2013
10:40	Μ. Γραβάνης, Π. Καραδήμος, Δ. Κούλικας, Α. Μπατζιάκας, Ε. Μπάτζιου, Γ. Ξανθοπούλου, Σ. Παραγκαμιάν, Ε. Τονίδης: Σύγκριση μέτρων συσχέτισης για τον εντοπισμό αλλαγών συνδεσιμότητας των δικτύων του εγκεφάλου στις επιληπτικές εκφορτίσεις	Γ. Μανώλα, Β. Καραγιάννης, Π. Μωυσιάδης: Η εξέλιξη του οικονομικού δικτύου των γενικών δεικτών των Χρηματιστηρίων Αξιών για επιλεγμένες χώρες από την Αμερική, την Ευρώπη, την Ασία και την Ωκεανία, από το 2005 ως το 2015
11:00	Ν. Χαλκίδης, Μ. Τσαγιοπούλου, Ν. Παπακωνσταντίνου, Θ. Μωυσιάδης: Στατιστική ανάλυση επιγενετικών δεδομένων ασθενών που πάσχουν από χρόνια λεμφοκυτταρική λευχαιμία	Α. Κηπουρός: Ο πολιτικός κίνδυνος και η επίδρασή του στην αξία των μετοχών
10:00-12:00	ΣΕΜΙΝΑΡΙΟ Βιομηχανική Στατιστική-Στατιστικός Έλεγχος Διεργασιών	

11:30	ΔΙΑΛΕΙΜΜΑ - ΚΑΦΕΣ	
	Η ΣΤΑΤΙΣΤΙΚΗ ΣΤΗ ΔΕΥΤΕΡΟΒΑΘΜΙΑ ΕΚΠΑΙΔΕΥΣΗ (Προεδρεύων: Π. Μουσιάδης)	ΠΟΣΤΕΡ
10:00	Ο. Χαριτίδου: Ο Μικρός Θαλής και τα Στοχαστικά Μαθηματικά	D. Ioannidis: Estimating the survival function under random right censoring
10:20	Μ. Κασκαντάμης: Νέες οπτικές στη διδασκαλία της Στατιστικής στη δευτεροβάθμια εκπαίδευση	K. Mantzouni, I. Ntzoufras, M. Kateri: Bayesian inference for association models in contingency tables
10:40	Κ. Νάκος: Στατιστική και κοινωνική έρευνα	Θ.-Ε. Μοσχονά: Λογαριθμογραμμικά μοντέλα διαγενεακής επαγγελματικής κινητικότητας
11:00	Α. Καρακώστα, Π. Αναγνωστάκης: Περιγράφοντας στατιστικά τη σχολική μας κοινότητα σχολικού έτους 2015-2016	Δ. Παπαναστασίου: Έλεγχος για έναν μέσο σε μετασχηματισμένα δεδομένα
	Γ. Θωμαΐδης: Μία ανάλυση των θεμάτων Στατιστικής στις Πανελλαδικές Εξετάσεις τα τελευταία 15 χρόνια:Τί είδους γνώσεις και αντιλήψεις αποκτούν οι μαθητές;	
11:10-13:00	ΣΕΜΙΝΑΡΙΟ Βιομηχανική Στατιστική-Στατιστικός Έλεγχος Διεργασιών	
13:40	ΜΕΣΗΜΒΡΙΝΗ ΔΙΑΚΟΠΗ Ελαφρύ γεύμα	
	ΝΑΟΥΣΑ (ΣΧΟΛΗ ΑΡΙΣΤΟΤΕΛΟΥΣ, ΜΙΕΖΑ)	
	ΣΤΑΤΙΣΤΙΚΗ ΣΤΗΝ ΕΚΠΑΙΔΕΥΣΗ (Προεδρεύων: Τ. Παπαϊωάννου)	
15:40	Χ.-Α. Βογιατζή, Ε. Ασβεστόπουλος: Αντιλήψεις μαθηματικών για τη διδασκαλία της Στατιστικής στην Δευτεροβάθμια (Μία στατιστική προσέγγιση)	
	ΣΤΡΟΓΓΥΛΗ ΤΡΑΠΕΖΑ (Συντονιστής: Τ. Παπαϊωάννου)	
16:00	Ι. Θωμαΐδης, Μ. Κασκαντάμης, Γ. Πούλος: Το μάθημα της Στατιστικής στη Δευτεροβάθμια Εκπαίδευση, σημερινή κατάσταση-μελλοντικές εξελίξεις-προτάσεις	
17:00	ΟΛΟΜΕΛΕΙΑ: ΛΗΞΗ ΣΥΝΕΔΡΙΟΥ	

ΠΡΟΓΡΑΜΜΑ ΚΟΙΝΩΝΙΚΩΝ ΕΚΔΗΛΩΣΕΩΝ

Τετάρτη 4-5-2016

Welcome Cocktail: 20: 00

Παρασκευή 5-5-2016

Εκδρομή: 13:00

Επίσημο δείπνο: 21:00

ΠΡΟΓΡΑΜΜΑ ΓΙΑ ΤΑ ΣΥΝΟΔΕΥΟΝΤΑ ΜΕΛΗ

Πέμπτη 5-5-2016: Αστικός περίπατος

Ωρα συνάντησης 10:00 π.μ. στο χώρο του συνεδρίου

Ξενάγηση από εθελοντή Ξεναγό του Δήμου Νάουσας

Πέμπτη 5-5-2016: Επίσκεψη στο Οινοποιείο Vaeni

Ωρα συνάντησης 18:00 μ.μ. στο χώρο του συνεδρίου

Παρασκευή 6-5-2016: Επίσκεψη στο Μουσείον των Παραδοσιακών Δρώμενων της Αποκριάς στη Νάουσα «Γενίτσαροι και Μπούλες»

Ωρα συνάντησης 10:00 π.μ. στο χώρο του συνεδρίου

Ξενάγηση από Πρόεδρο του Ομίλου, κ. Άρη Τόσιο

Διοικητικό Συμβούλιο Ε.Σ.Ι.

Πρόεδρος: Δαμιανού Χαράλαμπος

Αντιπρόεδρος: Κουτροβέλης Ιωάννης

Γενικός Γραμματέας: Κίτσος Χρήστος

Ειδικός Γραμματέας: Παναγιωτάκος Δημοσθένης

Ταμίας: Μπερσίμης Σωτήρης

Έφορος Βιβλιοθήκης: Παπαϊωάννου Τάκης

Μέλος: Χαραλαμπίδης Χαράλαμπος

Επιστημονική Επιτροπή

Δαμιανού Χ.	Κίτσος Χ.	Νάκας Χ.
Δάρας Τ.	Κουγιουμτζής Δ.	Ντζούφρας Ι.
Δονάτος Γ.	Κουκουβίνος Χ.	Παναγιωτάκος Δ.
Ζωγράφος Κ.	Κουνιάς Σ.	Παπαϊωάννου Τ.
Ηλιόπουλος Γ.	Κούτρας Μ.	Σύψας Π.
Κάκουλλος Θ.	Κουτροβέλης Ι.	Τσακλίδης Γ.
Καλαματιανού Α.	Μπερσίμης Σ.	Φωκιανός Κ.
Καραγρηγορίου Α.	Μπουρνέτας Α.	Χριστοφίδης Τ.
Κατσή Α.	Μουσιάδης Π.	

Οργανωτική Επιτροπή

Φ. Κολυβά-Μαχαίρα, *Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης*
Ε. Αγγελής, *Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης*
Ι. Αντωνίου, *Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης*
Δ. Ιωαννίδης, *Πανεπιστήμιο Μακεδονίας*
Δ. Κουγιουμτζής, *Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης*
Π. Μουσιάδης, *Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης*
Α. Παπαδοπούλου, *Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης*
Δ. Παπαναστασίου, *Πανεπιστήμιο Μακεδονίας*
Κ. Τολικά, *Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης*
Γ. Τσακλίδης, *Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης*
Ν. Φαρμάκης, *Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης*
Θ. Χατζηπαντελής, *Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης*
Στ. Χατζόπουλος, *Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης*

Γ. Αθανασίου, *Νάουσα*
Σ. Ασβεστόπουλος, *Νάουσα*
Σ. Μιόγλου, *Νάουσα*

εργασίες

στα ελληνικά



ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΕΝΟΣ ΜΕΓΑΛΟΥ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ ΔΕΙΚΤΩΝ ΟΙΝΩΝ ΚΑΙ ΕΦΑΡΜΟΓΗ ΣΤΑΤΙΣΤΙΚΩΝ ΜΟΝΤΕΛΩΝ ΠΡΟΒΛΕΨΗΣ ΔΕΙΚΤΗ ΠΟΙΟΤΗΤΑΣ

Ιωάννης Αθανασιάδης, Δημήτριος Ιωαννίδης
Τμήμα Οικονομικών, Πανεπιστήμιο Μακεδονίας
{athang, dimioan}@uom.edu.gr

ΠΕΡΙΛΗΨΗ

Η αγορά του διαδικτύου σήμερα έρχεται αντιμέτωπη με το φαινόμενο των μεγάλων συνόλων δεδομένων. Η τεράστια ανάπτυξη των συναλλαγών ανέπτυξε αντίστοιχα και τις συναλλαγές στο ηλεκτρονικό εμπόριο μέσω των e-shops και των πελατών. Πολλές πλέον επιχειρηματικές αποφάσεις εξαρτώνται σε μεγάλο βαθμό από την στατιστική ανάλυση και πρόβλεψη αποτελεσμάτων έχοντας στην διάθεσή τους πολλά τέτοια σύνολα δεδομένων και μεταβλητών. Το κύριο εργαλείο μας για την ανάλυση και τα αποτελέσματα είναι η ανοικτού κώδικα γλώσσα στατιστικής R.

Στην εργασία μας η οποία έχει ένα μεγάλο σύνολο δεδομένων των πραγματικών φυσικοχημικών ιδιοτήτων των κρασιών μιας γνωστής ελληνικής επιχείρησης ερευνούμε και μελετούμε μερικά στατιστικά αποτελέσματα προσπαθώντας να προβλέψουμε το τελικό ποιοτικό αποτέλεσμα (quality score) εξαρτώμενοι από αυτές τις ιδιότητες πλέον εφαρμόζουμε διάφορα μοντέλα στατιστικής όπως τα "δέντρα απόφασης" (decision trees) καθώς και τα "νευρωνικά δίκτυα" (neural networks) υλοποιώντας διάφορες προβλέψεις του ποιοτικού αποτελέσματος και τέλος συγκρίνουμε αυτά τα αποτελέσματα.

Λέξεις κλειδιά: υπερ-μεγέθη δεδομένα, γλώσσα R, στατιστικά μοντέλα, ηλεκτρονικό εμπόριο, παλινδρόμηση, ποιότητα κρασιού, προβλέψεις, τεχνικές μοντελοποίησης, δέντρα απόφασης, decision trees, νευρωνικά δίκτυα, neural networks

1. ΕΙΣΑΓΩΓΗ

Το αναδυόμενο φαινόμενο των μεγάλων συνόλων δεδομένων σήμερα αυξάνεται παράλληλα με τα μεγάλα δεδομένα αγοράς Ιστού (web market) και έχει επιπτώσεις συνεπώς σε οποιαδήποτε επιχείρηση θέλει να διατηρηθεί σε αυτήν την περιοχή, ερευνώντας και ελέγχοντας τις μεθόδους ώστε να επιτύχει χρήσιμες πληροφορίες από μεγάλα σύνολα δεδομένων (big datasets). (Athanasiadis I. και Ioannides D., 2014)

Όπως όλοι γνωρίζουμε οι μεγάλες επιχειρήσεις στον Ιστό, όπως η Amazon, η Google ή η Facebook κλπ. οδηγούν την εποχή της "άντλησης γνώσης από τα δεδομένα". Η απέραντη ροή δεδομένων τους δείχνει ότι με μια κατάλληλη στατιστική ανάλυση και μία αναλυτική πρόβλεψη μεταξύ αυτών των δεδομένων είναι δυνατό να παραχθούν στρατηγικές επιχειρησιακές αποφάσεις.

Αυτή η πρόοδος της "ανάλυσης δεδομένων" παρήγαγε πολλά σημαντικά εργαλεία που ακόμη και σήμερα και μια μικρή επιχείρηση μπορεί να χρησιμοποιήσει με ασφάλεια και να ερευνήσει τα αποτελέσματα για όφελός της. Η γλώσσα R είναι ένα τέτοιο σημαντικό εργαλείο που παράγει με πολλά στατιστικά μοντέλα πολύ ενδιαφέροντα αποτελέσματα, ειδικά στην περιοχή πρόβλεψης (predictive analytics).

Μια τέτοια επιχείρηση κρασιού στη Βόρεια Ελλάδα παράγοντας κρασιά και έχοντας μια μεγάλη περιοχή αγοράς όχι μόνο στην Ελλάδα αλλά και στην Ευρώπη, μας έδωσε ένα μεγάλο σύνολο δεδομένων των κρασιών της που παρήχθησαν, προκειμένου να γίνει στατιστική έρευνα χρησιμοποιώντας διάφορα μοντέλα και αναλυτική πρόβλεψη. Αυτό το σύνολο δεδομένων περιλαμβάνει περισσότερα από 5600 εγγραφές κρασιών με μετρήσεις από φυσικοχημικές ιδιότητες των , συμπεριλαμβανομένων των τριών τύπων κρασιών (λευκό, κόκκινο και roze) και καλύπτει περίπου 9 έτη παραγωγής (από το 2004 ως το 2013).

Η ιστορία του κρασιού στην Ελλάδα ξεκινά από τους αρχαίους χρόνους και θεωρείται σε πολλές χώρες σε όλο τον κόσμο ως ένα κορυφαίας κατηγορίας κρασί. Υπάρχουν περίπου 300 ποικιλίες και τρεις κύριοι είδη ή τύποι: λευκό, κόκκινο και roze. (Galanos V., 2013). Η σημασία του κρασιού στην οικονομία είναι τεράστια δεδομένου ότι ως παραγωγός χώρα είναι 12^η στην παγκόσμια κατάταξη (Karlssoon B., 2012). Οι εξαγωγές των ελληνικών κρασιών είναι πάντα μέσα στα πρώτα των εξαγόμενων προϊόντων στην Ελλάδα και μεταξύ των πρώτων 10 χωρών στον κόσμο, που ποικίλλει από 600.000 έως 270.000 HL και 79 έως 52 εκατομμύρια σε €/έτος ανάμεσα στα έτη 2000 – 2008, (Πανελλήνια Ένωση των Οινολόγων Πτυχιούχων, ΠΑΝΕΠΟ, 2014).

Οι φυσικοχημικές ιδιότητες των κρασιών με την δοκιμαστική αισθητήρια γεύση των είναι απαραίτητες για την πιστοποίηση του κρασιού, με πρώτη δοκιμασία από τον έλεγχο ορισμένων χημικών δεικτών και δεύτερη, κυρίως από τον ειδικό γευστιγνώστη. Η προτίμηση είναι μια δύσκολη έως και ελάχιστα αντιληπτή ιδιότητα από τον άνθρωπο και οι σχέσεις μεταξύ των φυσικοχημικών αποτελεσμάτων και της ανθρώπινης αισθητήριας ανάλυσης είναι δύσκολο να προσδιοριστούν.

Συχνά ιδιαίτερα σύνθετα και μεγάλα σε όγκο σύνολα δεδομένων είναι δύσκολο να συλλεχθούν, να αποθηκευθούν και να επεξεργαστούν. Οι πρόοδοι στην τεχνολογία πληροφορικής και ειδικά με τη στατιστική γλώσσα R (ανοικτού κώδικα και ελεύθερη με πολλές βιβλιοθήκες διαθέσιμες) έχουν κάνει πιθανή την επίλυση αυτών των στόχων. (Turban E., και λοιποί, 2007)

Οι τεχνικές εξόρυξης δεδομένων (DM)(data mining) στοχεύουν στην εξαγωγή υψηλού επιπέδου γνώσης από ακατέργαστα δεδομένα. Υπάρχουν διάφοροι αλγόριθμοι DM, καθένας με τα πλεονεκτήματα και μειονεκτήματά του (Witten I.H. και λοιποί, 2005) .

Οι μεταβλητές και οι επιλογές μοντέλων στο μεγάλο σύνολο δεδομένων μας είναι κρίσιμο ζήτημα για την εφαρμογή αυτών των μεθόδων DM. Η επιλογή των σημαντικών μεταβλητών (Guyon I. and Elisseeff A. , 2003) είναι χρήσιμη ώστε να απορρίψει τις περιττές εισαγωγές κάτι που οδηγεί σε απλούστερα και ευκολότερα ερμηνεύσιμα μοντέλα. Συνήθως τα σύνθετα μοντέλα υπερπροσαρμόζουν τα δεδομένα μας, χάνοντας την ικανότητα γενίκευσης, ενώ ένα πολύ απλό μοντέλο παρουσιάζει μεγάλη μεροληψία . Ο στατιστικός καλείται να επιλέξει την καλύτερη μεθοδολογία ανάμεσα σε αυτές τις ακραίες επιλογές.

Η χρήση των συστημάτων υποστήριξης απόφασης (decision support systems) από τη βιομηχανία κρασιού στρέφεται κυρίως στη φάση παραγωγής κρασιού (Ferrer J. και λοιποί, 2008) αλλά στην περίπτωση μας έχουμε και τους δύο τρόπους ποιοτικής κατηγοριοποίησης του : τη φυσικοχημική δοκιμασία και την ανθρώπινη αισθητήρια δοκιμασία από το εκπαιδευμένο ανθρώπινο δυναμικό της επιχείρησης. Επομένως η μελέτη μας για να προβλέψουμε την ποιοτική κατάταξη του κρασιού σε μια τακτική κλίμακα, πχ ακεραίων από 1 έως 10 , είναι πολύ σημαντική για την επιχείρηση και συνεπώς για οποιαδήποτε επιχείρηση παραγωγής στην αγορά. Αναζητούμε λοιπόν έναν τρόπο σε ένα περιβάλλον συνεχώς αυξανόμενων υπερόγκων δεδομένων που δημιουργούνται από την συνεχή παραγωγή τρόπους για ασφαλέστερες προβλέψεις .

Ανατρέχοντας στην ιστορία αυτών των δοκιμών, το 1991 το σύνολο δεδομένων "Wine" δόθηκε στην αποθήκη δεδομένων (dataset repository) UCI (Asuncion A. και λοιποί, 2007). Τα δεδομένα περιέχουν 178 παραδείγματα με τις μετρήσεις 13 χημικών συστατικών (π.χ. οινόπνευμα, MG κλπ.) και ο στόχος είναι να ταξινομηθούν τρεις ποικιλίες από την Ιταλία. Αυτό το σύνολο δεδομένων είναι πολύ εύκολο να γίνει διακριτό και έχει χρησιμοποιηθεί κυρίως ως συγκριτική μέτρηση επιδόσεων για νέους ταξινομητές DM.

Στην εργασία (Cortez και λοιποί, 2009) έγιναν προβλέψεις γευσιγνωσίας του "vinho verde" κρασιού (από την περιοχή του Minho της Πορτογαλίας), παρουσιάζοντας ιδιαίτερο αντίκτυπο σε αυτήν την περιοχή. Σε αντίθεση με τις προηγούμενες μελέτες, εδώ εξετάζεται ένα μεγάλο σύνολο δεδομένων, με συνολικά 4898 λευκά και 1599 κόκκινα δείγματα και 11 φυσικοχημικές ιδιότητες (ή μεταβλητές). Σε εκείνη την εργασία χρησιμοποιήθηκε η μέθοδος νευρωνικών

δικτύων και των μηχανών διανυσματικής υποστήριξης για να ερευνηθεί το σύνολο δεδομένων. (Cortez και λοιποί, 2009)

Στην εργασία μας έχουμε ένα μεγάλο σύνολο δεδομένων κρασιών, που έχουν περισσότερα από 5600 αρχεία, με 2795 κόκκινα, 2315 λευκά και 494 ροζέ. Οι μεταβλητές μας είναι 13 συμπεριλαμβανομένης της ημερομηνίας της παραγωγής και του είδους κρασιού (ή ποικιλίας). Αυτές οι δύο μεταβλητές σε αυτήν την μελέτη είναι αποκλεισμένες για λόγους συμβατότητας αλλά πολύ χρήσιμες σε μελλοντικές μελέτες.

Η εργασία αυτή αποτελείται από τα εξής μέρη : Η παράγραφος 2 παρουσιάζει τα δεδομένα, την προκαταρκτική γραφική ανάλυση, τη διερευνητική γραφική ανάλυση, την ανάλυση κύριων συνιστωσών , τα μοντέλα DM και τη μεταβλητή προσέγγιση επιλογής. Στην παράγραφο 3, το μοντέλο δέντρων απόφασης και πρόβλεψης και στην παράγραφο 4, το νευρωνικό μοντέλο δικτύων και την αντίστοιχη πρόβλεψή τους. Όλα τα παραπάνω περιγράφονται με τα επιτευχθέντα αποτελέσματα και αναλύονται αντίστοιχα. Τέλος, τα βασικά συμπεράσματα της εργασίας δίνονται στην παράγραφο 5 με τις συγκρίσεις των αποτελεσμάτων.

2. ΔΕΔΟΜΕΝΑ ΚΑΙ ΜΕΘΟΔΟΙ

2.1 Δεδομένα κρασιών

Αυτή η μελέτη εξετάζει περίπου 117 ποικιλίες κρασιών που παράγονται από μία γνωστή επιχείρηση της Βόρειας Ελλάδας. Τα περισσότερα από αυτά τα κρασιά εξάγονται κυρίως στην Ευρώπη και σε άλλες χώρες. Αναλύουμε τρεις παραλλαγές κρασιού: το λευκό, το κόκκινο και το ροζέ. Τα δεδομένα συλλέχθηκαν από 4/2004 μέχρι το 10/2013 με εξέταση από επίσημη πιστοποίηση του τμήματος ποιότητας της επιχείρησης. Κάθε εγγραφή δείχνει μια δεδομένη δοκιμή (αναλυτική και αισθητήρια) και η τελική βάση δεδομένων εξήχθη σε ένα ενιαίο φύλλο (.csv).

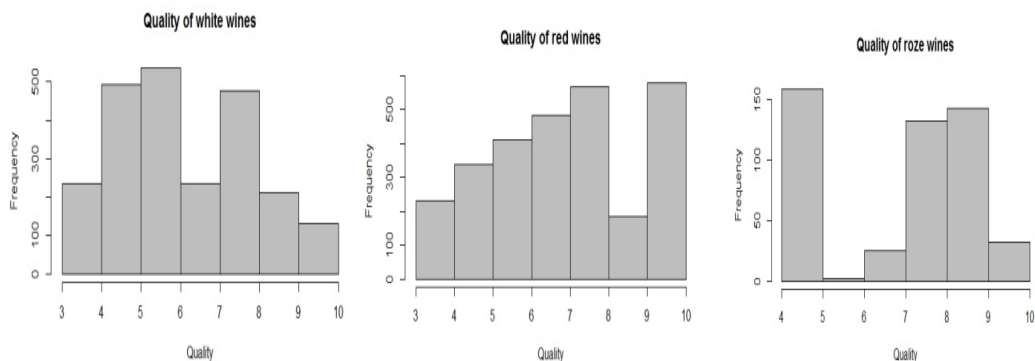
Ενδιαφερόμαστε να καταλάβουμε με την βοήθεια στατιστικών μεθόδων , π.χ. αν ένα κρασί που έχει, περισσότερη αλκοόλη ή διοξείδιο του θείου θα δίνει καλύτερη γεύση. Κατά τη διάρκεια του σταδίου της προετοιμασίας των μεταβλητών των συνόλων δεδομένων ένας μεγάλος αριθμός εγγραφών έχει ελλείπουσες τιμές σε δύο μεταβλητές: την μεταβλητή "Fe" (σίδηρος) και την "απόχρωση" (shade), κάτι που μας οδήγησε στην διαγραφή τους . Ένα άλλο στοιχείο είναι ότι αγνοήσαμε την μεταβλητή "ημερομηνία" (date) δεχόμενοι ότι ο χρόνος δεν επηρεάζει την ποιότητα , πιστεύουμε σε μια μελλοντική μελέτη μας να αξιοποιήσουμε την μεταβλητή χρόνο δίνοντας αποτελέσματα με την βοήθεια χρονικών σειρών (time series).

Δεδομένου ότι έχουμε τρεις διαφορετικούς τύπους κρασιού θα τους μελετήσουμε σύμφωνα με αυτόν τον διαχωρισμό. Έτσι έχουμε τρία σύνολα δεδομένων , τα κόκκινα με 2795 περιπτώσεις, τα λευκά με 2315, και τέλος με 494

τα ροζέ. Ο πίνακας 1 δίνει τις μέσες , ελάχιστες και μέγιστες τιμές από τις μετρήσεις των φυσικοχημικών ιδιοτήτων των τριών τύπων . Οι ποιοτικές κατατάξεις για κάθε τύπο δίνεται σε μια τακτική κλίμακα που κυμαίνεται από το 1(κακό) ως το 10 (άριστο). Στην εικόνα 1 δίνονται τα ιστογράμματα για τις κατατάξεις τριών τύπων παραλλαγών. Ο κύριος σκοπός μας είναι να προσδιορίσουμε ποιες από αυτές τις φυσικοχημικές μας μεταβλητές έχουν την σημαντικότερη επίδραση στην ποιότητα κρασιού.

	White wine			Red wine			Roze wine		
Attribute (units)	min	max	mean	min	max	mean	min	max	mean
alcohol (% vol.)	9	13.65	11.71	9.05	14.85	12.55	9	14.70	12.25
pH	0	3.98	3.38	3.02	4.07	3.50	2.98	3.68	3.29
total acidity (g(tartaric acid)/dm ³)	0	8.93	5.01	4.01	10.13	5.51	3.71	9.38	5.56
volatile acidity (g(acetic acid)/dm ³)	0.100	1.430	0.306	0.120	1.060	0.381	0.140	0.500	0.317
sugar (g/dm ³)	0.50	36	5.14	1.80	47	6.98	1.50	40	5.94
colour intensity	0.030	3.080	0.077	0.06	36.16	6.79	0.08	5.21	0.93
free sulphur dioxide (mg/dm ³)	6	65	38.5	16	67	38.21	14	68	38.32
total sulphur dioxide (mg/dm ³)	26	248	138.7	12	234	96.62	45	172	111.9

Πίνακας 1. Αριθμητικά χαρακτηριστικά φυσικοχημικών ιδιοτήτων

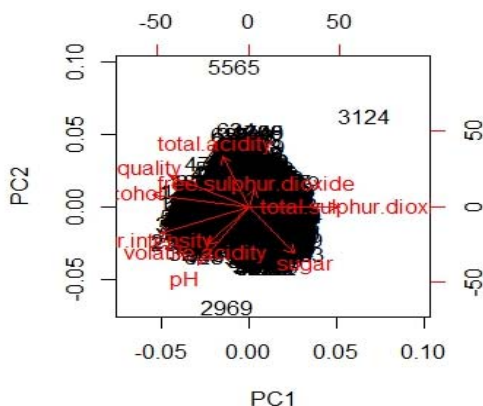


Εικ.1 Ιστογράμματα ποιότητας

2.2 Ανάλυση κυρίων συνιστωσών

Πριν προχωρήσουμε στην στατιστική ανάλυση, χρησιμοποιούμε ανάλυση κυρίων συνιστωσών για να δημιουργήσουμε τις κύριες συνιστώσες (principal

component) που δημιουργούν διακύμανση όσο και οι αρχικές μας μεταβλητές και επί πλέον θα είναι μεταξύ τους ασυσχέτιστες .



Εικ. 2. Διάγραμμα φόρτισης (loading plot)

Κάθε συνιστώσα είναι γραμμικός συνδυασμός των αρχικών μας μεταβλητών και φυσικά ανεξάρτητες προς τις υπόλοιπες. Στην βάση δεδομένων με τα ροζέ κρασιά έχουμε δυο κύριες συνιστώσες, όπου στο νέο αυτό σύστημα συντεταγμένων απεικονίζονται οι μεταβλητές μας με μορφή διανυσμάτων. Οι μεταβλητές με την ίδια φορά διανυσμάτων είναι θετικά συσχετισμένες και με συντελεστή συσχέτισης ανάλογο του συνημίτονου της γωνίας που σχηματίζουν .

Έτσι η μεταβλητή ποιότητα και η αλκοόλη (alcohol) είναι θετικά συσχετισμένες , ενώ τα σάκχαρα και το συνολικό διοξείδιο του θείου με την μεταβλητή ποιότητα έχουν αρνητική συσχέτιση

3. ΜΟΝΤΕΛΟ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ (DECISION TREES MODEL)

Τα δέντρα απόφασης δημιουργούνται χρησιμοποιώντας έναν ευρετικό (heuristic) διαχωρισμό αποκαλούμενο αναδρομικό διαχωρισμό. Αυτή η προσέγγιση είναι γενικά γνωστή ως "διαίρει και βασίλευε" επειδή χρησιμοποιεί τις τιμές χαρακτηριστικών γνωρισμάτων για να χωρίσει τις μετρήσεις από τις μεταβλητές σε μικρότερα και μικρότερα υποσύνολα σε όσο το δυνατόν ομογενοποιημένα μεταξύ τους.

Αρχίζοντας από τον κόμβο ρίζας, που αντιπροσωπεύει ολόκληρο το σύνολο δεδομένων, ο αλγόριθμος επιλέγει ένα χαρακτηριστικό του οποίου η περίπτωση μπορεί και περιγράφει κατηγορίες άλλων ανεξαρτήτων περιπτώσεων . Οι περιπτώσεις χωρίζονται έπειτα σε κατηγορίες διακριτών τιμών αυτού του χαρακτηριστικού γνωρίσματος. Αυτή η απόφαση διαμορφώνει το πρώτο σύνολο

κλάδων δέντρων. Ο αλγόριθμος συνεχίζει να διαιρεί και να κατακτά τους κόμβους, επιλέγοντας το καλύτερο χαρακτηριστικό γνώρισμα των επιλογών κάθε φορά μέχρι να ικανοποιηθεί μια συνθήκη τερματισμού. Περισσότερα ο αναγνώστης μπορεί να συμβουλευθεί (Breiman L. και λοιποί, 1984) και (Lantz B., 2013) .

3.1 Προετοιμασία του συνόλου δεδομένων λευκού κρασιού

Χρησιμοποιούμε το σύνολο του 75% από τα δεδομένα μας και το 25%, για την εκπαίδευση και δοκιμή του μοντέλου μας, αντίστοιχα. Θα αξιολογήσουμε την απόδοση των δέντρων απόφασης στα δεδομένα δοκιμής για να τα συγκρίνουμε με αντίστοιχα αποτελέσματα των (Cortez P. και άλλοι, 2009).

Έχουμε από 1736 και 579 περιπτώσεις των συνόλων εκπαίδευσης και δοκιμής, αντίστοιχα.

> my1.rpart

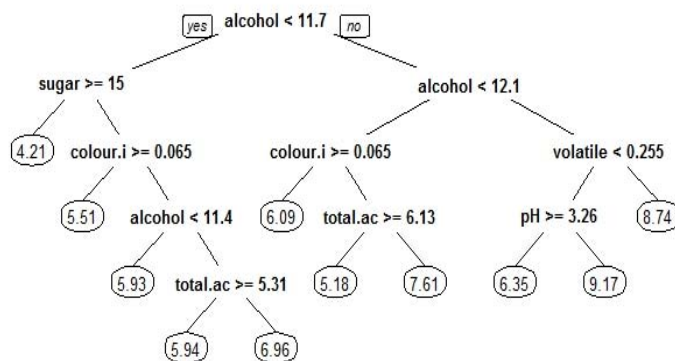
n= 1736 node), split, n, deviance, yval

* denotes terminal node

1) root	1736	5156.843000	6.681452		
2) alcohol< 11.725		921	1798.638000	5.801303	
4) sugar>=15.03		155	25.393550		4.206452 *
5) sugar< 15.03		766	1299.218000	6.124021	
10) colour.intensity>=0.065		334	345.473100	5.508982 *	
11) colour.intensity< 0.065		432	729.719900	6.599537	
22) alcohol< 11.375		90	111.600000	5.933333 *	
23) alcohol>=11.375		342	567.663700	6.774854	
46) total.acidity>=5.31		63	11.746030		5.936508 *
47) total.acidity< 5.31		279	501.641600	6.964158 *	
3) alcohol>=11.725		815	1838.483000	7.676074	
6) alcohol< 12.075		462	760.218600	7.101732	
12) colour.intensity>=0.065		137	197.766400	6.094891 *	
13) colour.intensity< 0.065		325	365.027700	7.526154	
26) total.acidity>=6.135		11	1.636364		5.181818 *
27) total.acidity< 6.135		314	300.818500	7.608280 *	
7) alcohol>=12.075		353	726.407900	8.427762	
14) volatile.acidity< 0.255		67	216.268700	7.104478	
28) pH>=3.265		49	87.102040		6.346939 *
29) pH< 3.265		18	24.500000		9.166667 *
15) volatile.acidity>=0.255		286	365.332200	8.737762 *	

Έτσι στον παραπάνω πίνακα, πχ στο κόμβο 29) έχουμε : pH < 3.265, ανήκει στην περίπτωση που είναι volatile.acidity<0.255 και alcohol >=12.075, και θα πάρουμε δείκτη ποιότητας 9.166 με 18 τέτοια κρασιά.

Γραφικά ο προηγούμενος πίνακας αποδίδεται, με το επόμενο σχήμα :



3.2 Πρόβλεψη και αξιολόγηση του μοντέλου με την γλώσσα R

Χρησιμοποιώντας την συνάρτηση `predict()` της R μπορούμε να δημιουργήσουμε τις προβλέψεις για την μεταβλητή “quality” στο σύνολο δοκιμής των 579 περιπτώσεων. Αν το αντικείμενο του συνόλου αυτού ήταν τό `dfwnew2_test` τότε χρησιμοποιώντας την συνάρτηση `predict` λαμβάνουμε το αντικείμενο διάνυσμα `pr1.rpart` (`> pr1.rpart <- predict (my1.rpart, dfwnew2_test)`). Εφαρμόζοντας την εντολή `summary()`, έχουμε

```
> summary(pr1.rpart)
```

Min. 1st Ou. Median Mean 3rd Ou. Max.

4.206 5.509 6.095 6.474 7.608 9.167

Εφαρμόζοντας την ίδια εντολή αλλά στην μεταβλητή quality της dfwnew2_test , έχουμε :

```
> summary(dfwnew2 test$quality)
```

Min. 1st Qu. Median Mean 3rd Qu. Max.

3.00 5.00 6.00 6.23 8.00 10.00

Από τα παραπάνω βλέπουμε ότι έχουμε διαφορά περίπου μια μονάδα στα άριστα και χειρότερα κρασιά, ενώ στα μεσαία συμφωνούν οι προβλέψεις με τα δεδομένα. Αν εξετάσουμε την συσχέτιση μεταξύ της προβλεπομένης ποιότητας και λαμβανόμενης ποιότητας πρέπει να κάνουμε χρήση την συνάρτηση $\text{cor}()$ της R, που δίνει

```
> cor(pr1.rpart, dfwnew2$test$quality)
```

[1] 0.5972786.

Η συσχέτιση είναι αρκετά μεγάλη περίπου 0.59 και μας δείχνει πόσο ισχυρά συσχετισμένες είναι οι προβλεπόμενες τιμές με τις λαμβανόμενες.

Ένας άλλος τρόπος αξιολόγησης του μοντέλου μας είναι μέσω του υπολογισμού της απόλυτης διακύμανσης των σφαλμάτων , δηλ. της κατά μέσον όρο απόλυτης διαφοράς της προβλεπόμενης από την αληθινή τιμή. Ο υπολογισμός της στην R γίνεται , αν ορίσουμε το αντικείμενο MAE ως εξής

```
> MAE <- function(actual, predicted) {  
  mean(abs(actual - predicted))  
}
```

Έτσι στα δικά μας δεδομένα έχουμε :

```
> MAE(pr1.rpart, dfwnew2_test$quality)
```

```
[1] 1.14525.
```

Το τελευταίο σημαίνει ότι κατά μέσο όρο η διαφορά των προβλεπόμενων τιμών από τις αληθινές είναι περίπου 1.14, που σε μια κλίμακα από 1 έως 10 είναι μια καλή τιμή. Ο Cortez βρήκαν απόλυτη διακύμανση περίπου 0.6, για διαφορετικά δεδομένα.

4. ΜΟΝΤΕΛΟ ΝΕΥΡΩΝΙΚΩΝ ΔΙΚΤΥΩΝ

4.1 Εισαγωγή

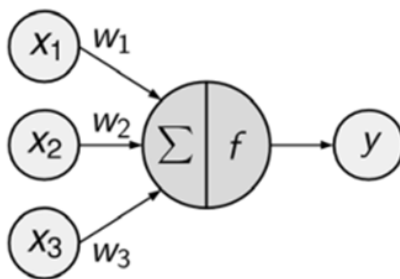
Τα νευρωνικά δίκτυα αποτελούν μέθοδο παράστασης γνώσης εμπνευσμένης από το νευρικό σύστημα, και δανείζονται έννοιες που χρησιμοποιούνται για την κατανόηση των ανθρώπινου εγκεφάλου. Παρά την πολυπλοκότητά τους, αυτά τα μοντέλα μπορούν να εφαρμοστούν εύκολα σε πραγματικά προβλήματα όπως στη διαμόρφωση της δύναμης του σκυροδέματος ή στη ανάγνωση τυπωμένου κειμένου (Lantz B., 2013)

Γενικά είναι μια μέθοδος "μαύρου κουτιού" που χρησιμοποιείται για στατιστικές προβλέψεις . Είναι "μαύρο κουτί" υπό την έννοια ότι αυτές οι μέθοδοι μετασχηματίζουν τις εισαγόμενες ανεξάρτητες μεταβλητές σε εξαγόμενες εξαρτημένες μεταβλητές με άγνωστη σχέση ως προς τις ανεξάρτητες . Σε πολλές περιπτώσεις υπάρχουν μέθοδοι ή αλγόριθμοι εκπαίδευσης που είναι αόρατες στο χρήστη.

4.2 Κατανόηση του μοντέλου ΝΔ

Το μοντέλο ενός ενιαίου τεχνητού νευρώνα μπορεί να γίνει κατανοητό με όρους αντίστοιχους του βιολογικού μοντέλου. Όπως απεικονίζεται στην ακόλουθη εικόνα, ένα κατευθυνόμενο νευρωνικό διάγραμμα καθορίζει μια σχέση μεταξύ των σημάτων εισαγωγής που παραλαμβάνονται από τους δενδρίτες (μεταβλητές x) και του σήματος εξαγωγής (μεταβλητή y). Ακριβώς όπως με το βιολογικό νευρώνα, το σήμα κάθε

δενδρίτη είναι σταθμισμένο (τιμές w) ανάλογα με τη σημασία του. Τα σήματα εισαγωγής αθροίζονται σε έναν κόμβο που τον διαπερνούν σύμφωνα με την συνάρτηση ενεργοποίησης f , $y(x) = f(\sum_{i=1}^n w_i x_i)$ (βλ.εικ. 3). Μια τέτοια συνάρτηση παριστάνει έναν τεχνητό νευρώνα με n δενδρίτες εισαγωγής. Τα βάρη w επιτρέπουν σε κάθε μια από τις n εισαγωγές, (x) , να συνεισφέρουν με έναν μεγαλύτερο ή μικρότερο ποσό στο άθροισμα των σημάτων εισαγωγής. Το ποσό εισαγωγής από τη συνάρτηση ενεργοποίησης $f(x)$, και το προκύπτον σήμα, $y(x)$, είναι το εξαγόμενο



Εικ. 3. Ο νευρώνας του ΝΑ

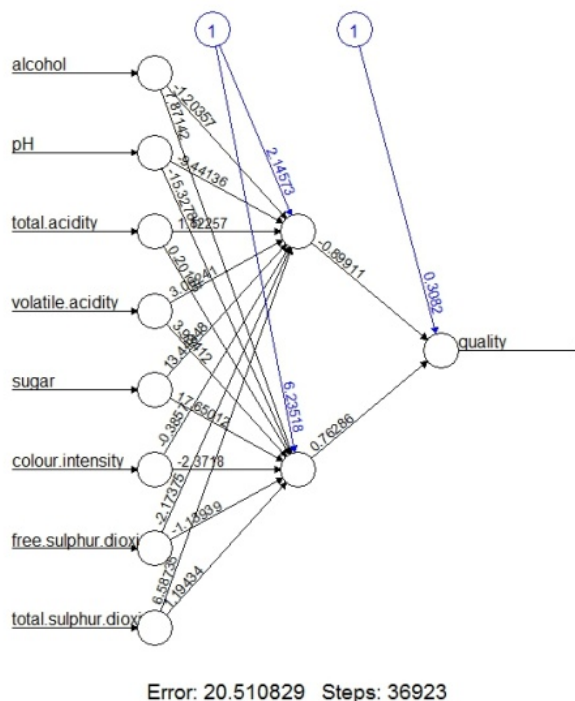
Τα νευρωνικά δίκτυα χρησιμοποιούν τους νευρώνες που καθορίζονται κατ' αυτό τον τρόπο ως δομικές μονάδες για να κατασκευάσουν σύνθετα μοντέλα δεδομένων. Αν και υπάρχουν πολυάριθμες παραλλαγές των νευρωνικών δικτύων, κάθε ένα καθορίζεται σύμφωνα με τα επόμενα χαρακτηριστικά :

- Μια **συνάρτηση ενεργοποίησης**, που μετασχηματίζει σήματα εισαγωγής ενός νευρώνα σε σήματα εξόδου και το οποίο συνεχίζει να μεταδίδεται στο δίκτυο .
- Μια **τοπολογία δικτύων** (ή αρχιτεκτονική), η οποία περιγράφει τον αριθμό νευρώνων στο μοντέλο καθώς επίσης και τον αριθμό στρωμάτων και τον τρόπο με τον οποίο συνδέονται
- Ο **αλγόριθμος εκπαίδευσης** που διευκρινίζει πώς τα βάρη σύνδεσης τίθενται ως στόχος προκειμένου να εμποδιστούν ή να διεγερθούν οι νευρώνες ανάλογα προς το σήμα εισαγωγής (Lantz B., 2013)

4.3 Διαγράμματα των ΝΑ

Χρησιμοποιούμε τα ίδια δεδομένα αφού τα έχουμε τυποποιήσει, και δημιουργούμε το σύνολο εκπαίδευσης και δοκιμής όπως προηγούμενα. Υπάρχουν διάφορα πακέτα της R με τα οποία μπορούμε να εκπαιδεύσουμε το νευρωνικό δίκτυο μας , εμείς προτείνουμε το “neuralnet” . Με την βοήθεια της συνάρτησης αυτής και ορίζοντας δυο νευρώνες εξόδου, γραμμικού τύπου, λαμβάνουμε το επόμενο διάγραμμα (εικ.4)

Αν στην συνάρτηση μας “neuralnet” είχαμε ορίσει δυο κρυφούς κόμβους τότε θα είχαμε το ΝΔ της εικόνας 4. Στην εικόνα 4 αναγράφεται ότι το τετραγωνικό σφάλμα είναι 20.510829 που επιτυγχάνεται σε 36923 βήματα .



Εικ. 4. Το ΝΔ με 2 νευρώνες/επίπεδο

Μπορούσαμε να είχαμε τα διαγράμματα των ΝΔ με 3 , 4 και 5 κρυφούς κόμβους, και θα διαπιστώναμε ότι τα αντίστοιχα τετραγωνικά σφάλματα θα ήταν 20.12476, 17.358897 και 17.358897 . Αντιλαμβανόμαστε ότι τα δίκτυα με 4 και 5 κόμβους έχουν το μικρότερο τετραγωνικό σφάλμα. Άρα θα πρέπει να αποφασίσουμε ποιο από τα δυο να επιλέξουμε. Το συμπέρασμα μας το δίνουμε στην επόμενη παράγραφο.

4.4 Πρόβλεψη και αξιολόγηση απόδοσης μοντέλου με την γλώσσα R

Εύκολα μπορούμε να υπολογίσουμε τους συντελεστές συσχέτισης μεταξύ των προβλεπομένων τιμών της μεταβλητής quality στο σύνολο των 579 τιμών του συνόλου δοκιμής και τις τιμές της ίδιας μεταβλητής στο σύνολο δοκιμής. Να τονίσουμε, όμως, ότι πριν υπολογίσουμε τις προβλεπόμενες τιμές πρέπει ενδιάμεσα να χρησιμοποιήσουμε την συνάρτηση compute(), που στην συνέχεια θα λάβουμε τις συνιστώσες της, \$neurons και \$net.results. Με την βοήθεια της τελευταίας λαμβάνουμε τις προβλεπόμενες τιμές. Για λόγους οικονομίας χώρου δεν αναφέρουμε τον κώδικα, ο αναγνώστης μπορεί να συμβουλευθεί σχετικά το βιβλίο του (Lantz B., 2013) . Δίνουμε τους συντελεστές συσχέτισης στο

πίνακα 2 μεταξύ προβλεπομένων τιμών της ποιότητας και των λαμβανομένων. Διαπιστώνουμε ότι με 4 νευρώνες/επίπεδο έχουμε τον μεγαλύτερο συντελεστή συσχέτισης: 0.93 καθώς επίσης και ένα αποδεκτό MAPE (*mean absolute percentage error*) : 0.22. Αξιολογώντας συνολικά το δίκτυο προτιμούμε το δίκτυο με 4 νευρώνες επειδή έχει τον μεγαλύτερο συντελεστή συσχέτισης, και γενικά μικρό σφάλμα τετραγωνικό και απόλυτο.

Νευρώνες/επίπεδο	1 n/l	2 n/l	3 n/l	4 n/l	5 n/l
Συντελεστής συσχέτισης	0.738	0.866	0.892	0.932	0.755
MAPE	0.3302	0.1967	0.2214	0.2291	0.2193

Πίν. 2 .Συντελεστές συσχέτισης μεταξύ προβλεπομένων τιμών και αυτών από το σύνολο δοκιμής της μεταβλητής quality με διαφορετικό αριθμό κρυφών νευρώνων και τα mean absolute percentage errors.

5. ΠΕΡΙΛΗΨΗ – ΣΥΜΠΕΡΑΣΜΑΤΑ

Σε αυτήν την εργασία από μια πραγματική βάση δεδομένων λευκών κρασιών μελετήσαμε την επίδραση των φυσικοχημικών ιδιοτήτων του κρασιού στην ποιότητα του. Η μελέτη έγινε με δυο μεθόδους: τα "δέντρα απόφασης" και τα "νευρωνικά δίκτυα".

Τρέχοντας το μοντέλο δέντρων απόφασης είδαμε ότι προσαρμόζεται καλύτερα για να μας καθοδηγήσει σε ορισμένα ποιοτικά αποτελέσματα μετά από ορισμένους όρους. Κατ' αυτό τον τρόπο είμαστε ικανοί να εξηγήσουμε τις επιρροές μιας μεταβλητής στο τελικό αποτέλεσμα και στην πορεία προς εκείνο το αποτέλεσμα.

Το επόμενο μοντέλο ήταν το νευρωνικό μοντέλο δικτύων. Αυτό το μοντέλο θεωρείται ως μοντέλο "μαύρου κουτιού" επειδή υπάρχουν αλγόριθμοι και τεχνικές ή τεχνάσματα που δεν είναι διαφανείς στο χρήστη. Από την άλλη πλευρά παρουσιάζουν μια ξεχωριστή υψηλή ικανότητα πρόβλεψης του τελικού ποιοτικού αποτελέσματος. Το ΝΔ έχει αρκετά μεγάλη εξάρτηση από τον αριθμό των νευρώνων ανά επίπεδο που χρησιμοποιούνται για την κατασκευή του δικτύου.

Όλες οι δοκιμές μας έχουν δύο κύρια σύνολα δεδομένων: της εκπαίδευσης με 75% και της δοκιμής με το υπόλοιπο 25%.

Στην έρευνα ΝΔ διαπιστώσαμε ότι το καλύτερο αποτέλεσμα δίνεται με 4 νευρώνες/επίπεδο με συντελεστή συσχέτισης μεταξύ προβλεπομένων και λαμβανομένων τιμών 0.932, ενώ με 5 νευρώνες/επίπεδο θα έχουμε 0.755.

ABSTRACT

In our work having a big dataset of real physicochemical properties of wines of a known Greek company we investigate and research some statistical results trying to predict the quality score depending on these properties. Furthermore we apply various statistic models such as “Decision Trees” and “Neural Networks” making various predictions of quality score and comparing these results.

ΑΝΑΦΟΡΕΣ

- Asuncion A. and Newman D., 2007, UCI Machine Learning Repository, University of California, Irvine [<http://www.ics.uco.edu/~mlearn/MLRepository.html>].
- Athanasiadis, I and Ioannides ,D. , 2014, Proceedings of the 27th Greek statistical conferences, A statistical analysis of big web market data structure using R and the direct e-commerce social phenomenon of producer to consumer transaction pp. 333-347.
- Breiman L., Friedman J.H., Stone C.J., and Olshen R.A, Classification and Regression Trees, Chapman & Hall, 1984.
- Cortez Paulo, António Cerdeirab, Fernando Almeida, Telmo Matosb, José Reis (2009), Modeling wine preferences by data mining from physicochemical properties, Journal Decision Support Systems, Volume 47, Issue 4, November 2009, Pages 547–553.
- Ferrer J., MacCawley A., Maturana S., Toloza S., and Vera J., 2008, An optimization approach for scheduling wine grape harvest operations. International Journal of Production Economics, 112(2):985–999.
- Galanos V., 2013, <http://antikleidi.com/2013/03/23/wine-3/>.
- Greene, W. H. ,2000, Econometric Analysis, 4th ed., Prentice-Hall, Upper Saddle River, NJ.
- Guyon I. and Elisseeff A.,2003, An introduction to variable and feature selection. Journal of Machine Learning Research, 3(7–8):1157–1182.
- Karlsson B., 2012, <http://www.krasiagr.com/?p=52263> .
- Lantz Brett, 2013, Machine Learning with R, Packt Publishing.
- Πανελλήνια Ένωση των Οινολόγων Πτυχιούχων, PANEPO, 2014, <http://www.panepo.gr/Statistics.htm>
- Ramsay, J.O. and Silverman, B.W. ,2005, Functional Data Analysis, New York: Springer- Verlag .
- Smith D. and Margolskee R.,2006, Making sense of taste. Scientific American, Special issue, 16(3):84–92.

Witten H.I. and Frank E.,2005, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco, CA,2nd edition.

ΑΠΟΣΤΑΣΗ ΜΕΤΑΞΥ k -ΡΟΩΝ ΕΠΙΤΥΧΙΩΝ ΣΕ ΑΛΥΣΙΔΑ MARKOV

A.N. Αράπης¹, Φ.Σ. Μακρή¹, Ζ.Μ. Ψυλλάκης²

¹ Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών
tasosarp@hotmail.com, makri@math.upatras.gr

² Τμήμα Φυσικής, Πανεπιστήμιο Πατρών
psillaki@physics.upatras.gr

ΠΕΡΙΛΗΨΗ

Έστω $\{X_t\}_{t \geq 1}$ μια ομογενής αλυσίδα Markov δύο καταστάσεων (αποτυχία, 0 - επιτυχία, 1). Θεωρούμε ροές επιτυχιών μήκους τουλάχιστον k (k -ροές) στην $\{X_t\}_{t=1}^n$, $n \geq 3$, ο αριθμός των οποίων έστω ότι είναι $G_{n,k}$, $1 \leq k \leq n$. Ορίζουμε την τυχαία μεταβλητή $D_{n,k}$, $3 \leq 2k+1 \leq n$, η οποία παριστάνει την απόσταση μεταξύ της πρώτης - τελευταίας επιτυχίας της πρώτης - τελευταίας k -ροής. Στην εργασία δίνονται αναδρομικές σχέσεις για τον υπολογισμό της $P(D_{n,k} = d \mid G_{n,k} \geq 2)$, $3 \leq 2k+1 \leq d \leq n$. Ένα αριθμητικό παράδειγμα διευκρινίζει περαιτέρω τα θεωρητικά αποτελέσματα.

Λέξεις Κλειδιά: Ακριβής κατανομή, Ροές, Αλυσίδα Markov.

1. ΕΙΣΑΓΩΓΗ

Ροές και γενικότερα σχηματισμοί εμφανίζονται συχνά σε εφαρμογές που μπορούν να κωδικοποιηθούν μέσω ακολουθιών δοκιμών με δύο ή περισσότερα αποτελέσματα. Ο αριθμός των ροών/σχηματισμών, σύμφωνα με διάφορους τρόπους απαρίθμησης τους (π.χ. κατά Mood (1940) ή Feller (1968)), το μήκος και οι αποστάσεις τους δίνουν σημαντικές πληροφορίες για τη δομή της ακολουθίας. Ιδιαίτερα, οι ροές επιτυχιών ή αποτυχιών σε δυαδικές ακολουθίες, αποτελούν αντικείμενο μελέτης με στόχο την εξαγωγή αποτελεσμάτων στα διάφορα πεδία εφαρμογής τους όπως: έλεγχος ποιότητας, αξιοπιστία μηχανικών συστημάτων, συμπίεση και μετάδοση πληροφορίας και μοριακή βιολογία.

Η μελέτη των ροών απαιτεί τον προσδιορισμό των κατανομών των στατιστικών συναρτήσεων που χρησιμοποιούνται για την περιγραφή τους υπό διάφορες υποθέσεις για την εγγενή με τη δομή των ακολουθιών τυχαιότητα. Τέτοιες υποθέσεις είναι, μεταξύ άλλων, η υπόθεση της ανεξαρτησίας, η υπόθεση της ανταλλαξιμότητας και η υπόθεση της Μαρκοβιανής εξάρτησης μεταξύ των στοιχείων της ακολουθίας που εξετάζεται.

Στη μοριακή βιολογία και ειδικότερα στη μελέτη του DNA θεωρούμενο ως μια ακολουθία στοιχείων από το σύνολο των νουκλεοτιδίων (βάσεων) $\{A, C, G, T\}$, έχουν χρησιμοποιηθεί τυχαίες μεταβλητές (τ.μ.) απαρίθμησης ροών ταύτισης σε διάφορα τμήματά του. Η μετατροπή (κωδικοποίηση) μιας ακολουθίας DNA μήκους m σε δυαδική ακολουθία (αποτυχία, 0 - επιτυχία, 1) μήκους n , $n \leq m$, μπορεί να γίνει με διάφορους τρόπους οι οποίοι έχουν προταθεί (π.χ. Avery and Henderson (1999)). Η απαρίθμηση ή ανίχνευση ροών με μήκος μεγαλύτερο ή ίσο με ένα προκαθορισμένο αριθμό k , $1 \leq k \leq n$, παρουσιάζει ιδιαίτερο ενδιαφέρον. Ο Benson (1999) εισήγαγε την τ.μ. $D_{n,k}$ η οποία παριστάνει την απόσταση μεταξύ της πρώτης και της τελευταίας εμφάνισης ροής επιτυχιών μήκους τουλάχιστον k σε μια δυαδική ακολουθία μήκους n . Η τ.μ. αυτή χρησιμοποιήθηκε για την ανίχνευση επαναλαμβανόμενων αλληλουχιών νουκλεοτιδίων που πιθανώς συνδέονται με την εμφάνιση ασθενειών.

Δοθέντος ότι, σε μια ακολουθία n ανεξάρτητων ή ανταλλάξιμων δυαδικών πειραμάτων $\{X_t\}_{t=1}^n$, $n \geq 3$, υπάρχουν τουλάχιστον δύο ροές επιτυχιών μήκους τουλάχιστον k , $1 \leq k \leq n$, οι Makri et al. (2015) προσδιόρισαν την ακριβή δεσμευμένη συνάρτηση πιθανότητας της τ.μ. $D_{n,k}$ η οποία παριστάνει το μήκος της μικρότερης υποακολουθίας της $\{X_t\}_{t=1}^n$ που περιέχει όλες τις ροές επιτυχιών που υπερβαίνουν ένα συγκεκριμένο μήκος. Προσδιόρισαν επίσης, μεταξύ άλλων, την από κοινού συνάρτηση πιθανότητας της τ.μ. αυτής με δύο άλλες τ.μ. απαρίθμησης ροών και απαρίθμησης επιτυχιών στην υποακολουθία. Ιδιαίτερα για ακολουθίες ανεξάρτητων και ισόνομων δυαδικών τ.μ. επιλεγμένα καθώς και επιπλέον αποτελέσματα παρουσιάζονται στην εργασία Αράπης κ.ά. (2015).

Στην παρούσα εργασία προσδιορίζεται η (δεσμευμένη) συνάρτηση πιθανότητας της $D_{n,k}$ θεωρώντας Μαρκοβιανή εξάρτηση μεταξύ των στοιχείων της δυαδικής ακολουθίας $\{X_t\}_{t=1}^n$, $n \geq 3$. Με τον τρόπο αυτόν αφ' ενός εμπλουτίζεται η μελέτη της $D_{n,k}$ και αφ' ετέρου προτείνονται εναλλακτικές σχέσεις για την κατανομή της για ακολουθίες ανεξάρτητων και ισόνομων δυαδικών πειραμάτων.

2. ΔΙΑΤΥΠΩΣΗ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ

Ένα αρκετά γενικό πρότυπο που περιγράφει την εσωτερική δομή μιας ακολουθίας 0–1 τ.μ. είναι της ομογενούς αλυσίδας Markov πρώτης τάξης (OAM1). Το πρότυπο αυτό θα χρησιμοποιηθεί για την ανάπτυξη και παρουσίαση των κύριων αποτελεσμάτων μας.

Έστω $\{X_t\}_{t \geq 1}$ μια OAM με χώρο καταστάσεων $\mathcal{A} = \{0, 1\}$, πιθανότητες μετάβασης ενός βήματος

$$p_{ij} = P(X_t = j \mid X_{t-1} = i) \geq 0, \quad i, j \in \mathcal{A}, \quad \sum_{j \in \mathcal{A}} p_{ij} = 1, \quad i \in \mathcal{A}, \quad t \geq 2, \quad (1)$$

και κατανομή πιθανότητας

$$p_i^{(t)} = P(X_t = i) \geq 0, \quad i \in \mathcal{A}, \quad \sum_{i \in \mathcal{A}} p_i^{(t)} = 1, \quad t \geq 1. \quad (2)$$

Οι πιθανότητες p_{ij} και $p_i^{(t)}$ είναι στοιχεία αντίστοιχα του 2×2 πίνακα $\mathbf{P} = (p_{ij}) = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}$ και του διανύσματος (γραμμής) $\mathbf{p}^{(t)} = (p_0^{(t)}, p_1^{(t)})$. Ισχύουν οι σχέσεις

$$\mathbf{p}^{(t)} = \mathbf{p}^{(t-1)}\mathbf{P} = \mathbf{p}^{(1)}\mathbf{P}^{t-1}, t \geq 2, \quad p_i^{(t)} = \mathbf{p}^{(t)}\mathbf{e}_i', i \in \mathcal{A}, t \geq 1,$$

$$p_{ij}^{(t-1)} = P(X_{t-1+m} = j \mid X_m = i) = \mathbf{e}_i\mathbf{P}^{t-1}\mathbf{e}_j', i, j \in \mathcal{A}, t \geq 2, m \geq 1, \quad (3)$$

όπου $\mathbf{e}_j'$ είναι το ανάστροφο διάνυσμα (στήλη) του $\mathbf{e}_j$ και $\mathbf{e}_0 = (1, 0)$, $\mathbf{e}_1 = (0, 1)$.

Θεωρούμε την ακολουθία $\{X_t\}_{t=1}^n$, $n \geq 3$, η οποία ορίζεται από τα πρώτα n στάδια εξέλιξης της $\{X_t\}_{t \geq 1}$ και ενδιαφερόμαστε για τον σχηματισμό ροών από μονάδες (1) σε αυτήν. Ως ροή από 1 ορίζεται μια (μερική) ακολουθία της $\{X_t\}_{t=1}^n$ η οποία αποτελείται από συνεχόμενες μονάδες των οποίων προηγούνται και έπονται μηδενικά (0) ή τίποτα. Ο αριθμός των 1 σε μια ροή αναφέρεται ως μήκος (ή μέγεθος) της ροής.

(a) Το μήκος της μεγαλύτερης ροής από 1 στην $\{X_t\}_{t=1}^n$ είναι μια τ.μ. η οποία συμβολίζεται με L_n (Balakrishnan and Koutras (2002)).

Θεωρώντας ένα θετικό ακέραιο k , $1 \leq k \leq n$, ενδιαφερόμαστε ιδιαίτερα για την εμφάνιση ροών από 1 μήκους k ή μεγαλύτερου. Οι ροές αυτές θα ονομάζονται στο εξής k -ροές από 1 ή απλώς k -ροές. Για τη μελέτη των ροών αυτών, ορίζουμε:

(b) Την (απαριθμητρία) τ.μ. $G_{n,k}$, $1 \leq k \leq n$, η οποία μετράει τον αριθμό των k -ροών από 1 στην $\{X_t\}_{t=1}^n$ (Balakrishnan and Koutras (2002)).

(c) Την τ.μ. $D_{n,k}$, $3 \leq 2k+1 \leq n$, η οποία παριστάνει την απόσταση μεταξύ πρώτης - τελευταίας 1 της πρώτης - τελευταίας k -ροής από 1 στην $\{X_t\}_{t=1}^n$ (Makri et al. (2015)).

Συνδυάζοντας τις (a) - (c) παρατηρούμε ότι η $D_{n,k}$ παριστάνει το μήκος του μικρότερου τμήματος (ή υποακολουθίας) της $\{X_t\}_{t=1}^n$, $n \geq 3$, στο οποίο περιέχονται οι $G_{n,k}$ το πλήθος k -ροές από 1 με μήκος η κάθε μία το πολύ ίσο με L_n . Επίσης σημειώνουμε ότι μεγάλες τιμές της $D_{n,k}$ αποτελούν ένδειξη ότι οι $G_{n,k}$ k -ροές από 1 διασπείρονται κατά μήκος μεγάλου τμήματος του διαστήματος μεταξύ της αριστερής και της δεξιάς πλευράς της $\{X_t\}_{t=1}^n$ ενώ αντίθετα μικρές τιμές της $D_{n,k}$ υποδεικνύουν ότι οι $G_{n,k}$ k -ροές από 1 είναι συγκεντρωμένες σε μια υποακολουθία της $\{X_t\}_{t=1}^n$ με μικρό μέγεθος, αφήνοντας τα υπόλοιπα τμήματα της $\{X_t\}_{t=1}^n$ άδεια από τέτοιες ροές. Μεταξύ των τ.μ. L_n , $G_{n,k}$ και $D_{n,k}$ ισχύουν οι σχέσεις: $D_{n,k} = L_n$, εάν $G_{n,k} = 1$ και $D_{n,k} > L_n$ εάν $G_{n,k} > 1$.

Ως παράδειγμα θεωρούμε τις δυαδικές δοκιμές 011111011001 αριθμημένες από 1 έως 12. Τότε έχουμε ότι: $L_{12} = 5$, $G_{12,1} = 3$, $D_{12,1} = 11$, $G_{12,2} = 2$, $D_{12,2} = 8$ και $G_{12,k} = 1$, $D_{12,k} = L_{12}$ για $k = 3, 4, 5$.

Σκοπός της εργασίας είναι ο προσδιορισμός της δεσμευμένης συνάρτησης πιθανότητας

$$f_{n,k}(d) = P(D_{n,k} = d \mid G_{n,k} \geq 2), 3 \leq 2k+1 \leq d \leq n, \quad (4)$$

όταν οι τ.μ. $G_{n,k}$ και $D_{n,k}$ ορίζονται σε 0 – 1 OAM1 $\{X_t\}_{t=1}^n$, $n \geq 3$, με γνωστή αρχική κατανομή πιθανότητας $\mathbf{p}^{(1)}$ και πίνακα μετάβασης ενός βήματος $\mathbf{P}$. Οι εκφράσεις για την $f_{n,k}(d)$ δίνονται για πρώτη φορά για $2 \leq k \leq \lfloor (n-1)/2 \rfloor$ και είναι νέες και εναλλακτικές αυτών που δίνονται από τους Arapis et al. (2016) για $k = 1$. Η $D_{n,1}$ παριστάνει το μέγεθος της υποακολουθίας στην οποία συγκεντρώνονται όλες οι επιτυχίες που εμφανίζονται στην ακολουθία.

Συγκεκριμένα στην ενότητα 3 αναπτύσσονται για OAM1 αναδρομικά σχήματα για τον υπολογισμό της $f_{n,k}(d)$. Το πρότυπο της OAM1 καλύπτει ως ειδική περίπτωση και ακολουθίες $\{X_t\}_{t=1}^n$ ανεξάρτητων και ισόνομων 0 – 1 τ.μ. Η περίπτωση αυτή συζητείται στην ενότητα 4. Ένα αριθμητικό παράδειγμα παρουσιάζεται στην ενότητα 5 διευκρινίζοντας περαιτέρω τα θεωρητικά αποτελέσματα των ενοτήτων 3 και 4.

Στην εργασία συμβολίζουμε με $\delta_{i,j}$ τη δέλτα συνάρτηση του Kronecker, με $\lfloor x \rfloor$ το (κάτω) ακέραιο μέρος ενός πραγματικού αριθμού x και με $I(\Gamma)$ τη δείκτρια συνάρτηση του Γ , $I(\Gamma) = 1$, αν το Γ συμβαίνει και 0, διαφορετικά.

3. ΚΥΡΙΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Αρχικά δίνονται σχέσεις υπολογισμού της δεσμευμένης συνάρτησης πιθανότητας της τ.μ. $G_{n,k}$ ορισμένης στην OAM1 $\{X_t\}_{t=1}^n$. Οι σχέσεις αυτές προκύπτουν με κατάλληλη τροποποίηση των αποτελεσμάτων του Λήμματος 2.2 της εργασίας Makri and Psillakis (2011) για ομογενή Μαρκοβιανή αλυσίδα πρώτης τάξης.

Λήμμα 1. Έστω $\theta_{n,k}(x) = P(G_{n,k} = x \mid X_1 = 0)$, $n \geq k \geq 1$. Τότε:

- (I) Για $x < 0$ ή $x > \lfloor \frac{n}{k+1} \rfloor$, $\theta_{n,k}(x) = 0$.
- (II) Για $1 \leq n < k+1$, $\theta_{n,k}(x) = \delta_{x,0}$.
- (III) Για $n = k+1$, $\theta_{n,k}(0) = 1 - p_{01}p_{11}^{k-1}$, $\theta_{n,k}(1) = p_{01}p_{11}^{k-1}$.
- (IV) Για $n = k+2$, $\theta_{n,k}(0) = 1 - p_{01}p_{11}^{k-1}(1 + p_{00})$, $\theta_{n,k}(1) = p_{01}p_{11}^{k-1}(1 + p_{00})$.
- (V) Για $n \geq k+3$, $0 \leq x \leq \lfloor \frac{n}{k+1} \rfloor$,

$$\theta_{n,k}(x) = p_{00}\theta_{n-1,k}(x) + p_{01} \left[p_{11}^{n-2} \delta_{x,1} + p_{10} \sum_{i=1}^{n-2} p_{11}^{i-1} \theta_{n-i-1,k}(x - I(i \geq k)) \right]. \quad \square \quad (5)$$

Λαμβάνοντας υπόψιν τη σχέση $\theta_{n,k}(0) = P(L_n < k \mid X_1 = 0)$, δίνεται στη συνέχεια η δεσμευμένη συνάρτηση κατανομής της τ.μ. L_n .

Πόρισμα 1. Έστω $\lambda_n(k) = P(L_n < k \mid X_1 = 0)$. Τότε:

- (I) Για $1 \leq n < k+1$, $\lambda_n(k) = 1$.
- (II) Για $n = k+1$, $\lambda_n(k) = 1 - p_{01}p_{11}^{k-1}$.
- (III) Για $n = k+2$, $\lambda_n(k) = 1 - p_{01}p_{11}^{k-1}(1 + p_{00})$.
- (IV) Για $n \geq k+3$,

$$\lambda_n(k) = p_{00}\lambda_{n-1}(k) + p_{01}p_{10} \sum_{i=1}^{k-1} p_{11}^{i-1} \lambda_{n-i-1}(k). \quad \square \quad (6)$$

Επίσης, μέσω της $\theta_{n,k}(x)$ υπολογίζεται άμεσα (Θεώρημα 2.6 της εργασίας Makri and Psillakis (2011)) η συνάρτηση πιθανότητας της τ.μ. $G_{n,k}$.

Πρόταση 1. Έστω $\psi_{n,k}(x) = P(G_{n,k} = x)$, $0 \leq x \leq \lfloor \frac{n+1}{k+1} \rfloor$, $n \geq k \geq 1$. Τότε,

$$\psi_{n,k}(x) = p_1^{(1)} p_{11}^{n-1} \delta_{x,1} + \sum_{i=0}^{n-1} \theta_{n-i,k}(x - I(i \geq k)) [p_0^{(1)} \delta_{i,0} + p_1^{(1)} p_{10} p_{11}^{i-1} (1 - \delta_{i,0})]. \quad \square \quad (7)$$

Εναλλακτικές εκφράσεις για την συνάρτηση πιθανότητας της $G_{n,k}$ έχουν δοθεί και από άλλους ερευνητές (βλ. π.χ. Balakrishnan and Koutras (2002)).

Πόρισμα 2. Έστω $\alpha_{n,k} = P(G_{n,k} \geq 2)$, $n \geq k \geq 1$. Τότε,

$$\alpha_{n,k} = 1 - \psi_{n,k}(0) - \psi_{n,k}(1). \quad \square \quad (8)$$

Στη συνέχεια προχωρούμε στην ανάπτυξη αποτελεσμάτων σχετικών με την τ.μ. $D_{n,k}$. Για τον σκοπό αυτόν ορίζουμε αρχικά μια χρήσιμη πιθανότητα.

Ορισμός 1. Για $m \geq 3$, ορίζουμε

$$\phi(m) = P(X_m = 1, (X_2, X_3, \dots, X_{m-1}) \in B_{m-2} \mid X_1 = 1), \quad (9)$$

όπου $B_1 = \{0\}$ και $B_{m-2} = \{0, 1\}^{m-2} - \underbrace{\{(1, 1, \dots, 1)\}}_{m-2}$, $m > 3$. $\square$

Λόγω της ομογένειας της Μαρκοβιανής αλυσίδας, ισχύει ότι

$$\phi(m) = p_{11}^{(m-1)} - p_{11}^{m-1}, \quad (10)$$

η οποία, για $1 + p_{01} - p_{11} \neq 0$, δηλ. για $\mathbf{P} \neq \mathbf{I} \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, γίνεται

$$\phi(m) = (1 + p_{01} - p_{11})^{-1} [p_{01} + (1 - p_{11})(p_{11} - p_{01})^{m-1}] - p_{11}^{m-1}. \quad (11)$$

Το υλικό που ακολουθεί (Λήμμα 2 και Πρόταση 2) αποτελεί τον κορμό των νέων αποτελεσμάτων για τη μελέτη της $D_{n,k}$ ορισμένης σε OAM1.

Λήμμα 2. Έστω $h_{n,k}(d) = P(D_{n,k} = d, G_{n,k} \geq 2 \mid X_1 = 0)$, $3 \leq 2k + 1 \leq d \leq n$. Τότε:

(I) Για $d > n - 1$, $h_{n,k}(d) = 0$.

(II) Για $d = n - 1$, $h_{n,k}(d) = p_{01} p_{11}^{2k-2} \phi(n - 2k + 1)$.

(III) Για $d = n - 2$, $h_{n,k}(d) = p_{01} p_{11}^{2k-2} (p_{00} + p_{10}) \phi(n - 2k)$.

(IV) Για $d \leq n - 3$,

$$h_{n,k}(d) = p_{10} h_{d+1,k}(d) \lambda_{n-d-1}(k) + \sum_{i=0}^{k-1} h_{n-i-1,k}(d) [p_{00} \delta_{i,0} + p_{01} p_{10} p_{11}^{i-1} (1 - \delta_{i,0})]. \quad (12)$$

Απόδειξη Θεωρούμε τις τ.μ. $D_{n,k}(i)$, $G_{n,k}(i)$ και $L_n(i)$, $i \geq 1$, αντίστοιχες των $D_{n,k}$, $G_{n,k}$ και L_n , ορισμένες στην υποακολουθία $\{X_t\}_{t=i}^n$ της $\{X_t\}_{t=1}^n$. Προφανώς

ισχύει ότι, $D_{n,k}(1) = D_{n,k}$, $G_{n,k}(1) = G_{n,k}$, $L_n(1) = L_n$. Για $2k+1 \leq d \leq n-3$ παρατηρούμε ότι

$$\begin{aligned}
h_{n,k}(d) &= P(D_{n,k}(2) = d, G_{n,k}(2) \geq 2, X_2 = 0 \mid X_1 = 0) \\
&\quad + \sum_{i=1}^{k-1} P(D_{n,k}(i+2) = d, G_{n,k}(i+2) \geq 2, (1 - X_{i+2}) \prod_{j=2}^{i+1} X_j = 1 \mid X_1 = 0) \\
&\quad + P(D_{d+1,k} = d, G_{d+1,k} \geq 2, X_{d+2} = 0, L_n(d+2) < k \mid X_1 = 0) \\
&= P(X_2 = 0 \mid X_1 = 0) P(D_{n,k}(2) = d, G_{n,k}(2) \geq 2 \mid X_2 = 0) \\
&\quad + \sum_{i=1}^{k-1} P(X_2 = 1 \mid X_1 = 0) \prod_{j=2}^i P(X_{j+1} = 1 \mid X_j = 1) \\
&\quad \times P(X_{i+2} = 0 \mid X_{i+1} = 1) \\
&\quad \times P(D_{n,k}(i+2) = d, G_{n,k}(i+2) \geq 2 \mid X_{i+2} = 0) \\
&\quad + P(D_{d+1,k} = d, G_{d+1,k} \geq 2 \mid X_1 = 0) \\
&\quad \times P(X_{d+2} = 0 \mid X_{d+1} = 1) P(L_n(d+2) < k \mid X_{d+2} = 0).
\end{aligned}$$

Το αποτέλεσμα προκύπτει από την ομογένεια της Μαρκοβιανής αλυσίδας. Οι εκφράσεις της $h_{n,k}(d)$ για $d \geq n-2$ προκύπτουν με ανάλογα επιχειρήματα. $\square$

Πρόταση 2. Έστω $g_{n,k}(d) = P(D_{n,k} = d, G_{n,k} \geq 2)$, $3 \leq 2k+1 \leq d \leq n$. Τότε:

- (I) Για $d = n$, $g_{n,k}(d) = p_1^{(1)} p_{11}^{2k-2} \phi(n-2k+2)$.
- (II) Για $d = n-1$, $g_{n,k}(d) = p_{11}^{2k-2} (p_1^{(1)} p_{10} + p_0^{(1)} p_{01}) \phi(n-2k+1)$.
- (III) Για $d \leq n-2$,

$$g_{n,k}(d) = p_{10} g_{d,k}(d) \lambda_{n-d}(k) + \sum_{i=0}^{k-1} h_{n-i,k}(d) [p_0^{(1)} \delta_{i,0} + p_1^{(1)} p_{10} p_{11}^{i-1} (1 - \delta_{i,0})]. \quad (13)$$

Απόδειξη Για $d \leq n-2$ παρατηρούμε αψ' ενός ότι

$$P(D_{n,k} = d, G_{n,k} \geq 2, X_1 = 0) = P(X_1 = 0) h_{n,k}(d)$$

και αψ' ετέρου ότι

$$\begin{aligned}
&P(D_{n,k} = d, G_{n,k} \geq 2, X_1 = 1) \\
&= P(D_{d,k} = d, G_{d,k} \geq 2, X_{d+1} = 0, L_n(d+1) < k) \\
&\quad + \sum_{i=1}^{k-1} P\left(\prod_{j=1}^i X_j = 1, X_{i+1} = 0, D_{n,k}(i+1) = d, G_{n,k}(i+1) \geq 2\right),
\end{aligned}$$

όπου οι $D_{n,k}(i)$, $G_{n,k}(i)$ και $L_n(i)$, $i \geq 1$, είναι όπως στην απόδειξη του Λήμματος 2. Στη συνέχεια, το αποτέλεσμα (III) της Πρότασης προκύπτει από τη σχέση

$$g_{n,k}(d) = P(D_{n,k} = d, G_{n,k} \geq 2, X_1 = 0) + P(D_{n,k} = d, G_{n,k} \geq 2, X_1 = 1)$$

και τη Μαρκοβιανή ιδιότητα της ακολουθίας. Τα αποτελέσματα (I) και (II) της Πρότασης προκύπτουν με ανάλογα επιχειρήματα. $\square$

Θεώρημα 1. Η δεσμευμένη συνάρτηση πιθανότητας $f_{n,k}(d) = P(D_{n,k} = d \mid G_{n,k} \geq 2)$, $2k + 1 \leq d \leq n$, $1 \leq k \leq \lfloor (n-1)/2 \rfloor$ δίνεται από τον τύπο

$$f_{n,k}(d) = g_{n,k}(d)/\alpha_{n,k}, \quad (14)$$

όπου οι πιθανότητες $\alpha_{n,k}$ και $g_{n,k}(d)$ προσδιορίζονται από το Πρόρισμα 2 και την Πρόταση 2, αντίστοιχα. $\square$

4. ΕΠΑΓΟΜΕΝΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Το πρότυπο της $0-1$ OAM1 καλύπτει επίσης και την περίπτωση μιας ακολουθίας ανεξάρτητων και ισόνομων $0-1$ τ.μ. Αυτό συμβαίνει διότι μια τέτοια ακολουθία $\{X_t\}_{t=1}^n$, $n \geq 3$, με $P(X_t = 1) = p_1$, $1 \leq t \leq n$, είναι μια ειδική OAM1 με πίνακα μετάβασης ενός βήματος

$$\mathbf{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} = \begin{pmatrix} 1-p_1 & p_1 \\ 1-p_1 & p_1 \end{pmatrix} \quad (15)$$

και κατανομή πιθανότητας

$$\mathbf{p}^{(t)} = (p_0^{(t)}, p_1^{(t)}) = (1-p_1, p_1), \quad 1 \leq t \leq n. \quad (16)$$

Για ακολουθίες της μορφής αυτής, η σχέση (11) γίνεται

$$\phi(m) = p_1(1-p_1^{m-2}), \quad 0 < p_1 < 1. \quad (17)$$

Επομένως το Θεώρημα 1 μπορεί να εφαρμοσθεί για ακολουθίες ανεξάρτητων και ισόνομων $0-1$ τ.μ. με χρήση των σχέσεων (15) - (17).

5. ΑΡΙΘΜΗΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Στην ενότητα αυτή δίνουμε ένα αριθμητικό παράδειγμα το οποίο βοηθά στην περαιτέρω κατανόηση των αποτελεσμάτων των ενοτήτων 3 και 4. Τα (αριθμητικά) αποτελέσματα αφορούν και στα δύο είδη ακολουθιών $\{X_t\}_{t=1}^n$, $n \geq 3$, που αναφέρονται στην εργασία. Αυτές είναι: (I) Μια ακολουθία από ανεξάρτητες και ισόνομες $0-1$ τ.μ. με $P(X_t = 1) = p_1 = 0.5$, $1 \leq t \leq n$. (II) Μια $0-1$ OAM1 με $p_{00} = p_{11} = 0.9$ και $p_1^{(1)} = 0.5$.

Στον Πίνακα 1 δίνονται οι πιθανότητες $f_{n,k}(d)$, $2k + 1 \leq d \leq n$, οι μέσες τιμές $\mu_{n,k} = E(D_{n,k} \mid G_{n,k} \geq 2)$ και οι διασπορές $\sigma_{n,k}^2 = Var(D_{n,k} \mid G_{n,k} \geq 2)$ για $k = 1, 2$ και $n = 7$. Οι $f_{n,k}(d)$ υπολογίστηκαν μέσω του Θεωρήματος 1. Το μήκος της ακολουθίας επιλέχτηκε μικρό, $n = 7$, έτσι ώστε οι απαιτούμενοι υπολογισμοί, ιδιαίτερα για $k = 2$, να μπορούν να γίνουν και με υπολογιστή τσέπης.

Πίνακας 1: $f_{7,k}(d)$, για $2k + 1 \leq d \leq 7$, $\mu_{7,k}$ και $\sigma_{7,k}^2$ για $k = 1, 2$.

d	3	4	5	6	7	$\mu_{7,k}$	$\sigma_{7,k}^2$
$k = 1$							
(I)	0.050505	0.121212	0.212121	0.303031	0.313131	5.7071	1.3788
(II)	0.013073	0.037352	0.071114	0.112746	0.765715	6.5807	0.7667
$k = 2$							
(I)			0.277778	0.333333	0.388889	6.1111	0.6543
(II)			0.037367	0.096085	0.866548	6.8292	0.2164

ABSTRACT

Let $\{X_t\}_{t \geq 1}$ be a time-homogeneous Markov chain of first order with state space $\{0, 1\}$. We consider in the first n , $n \geq 3$, trials runs of 1s of length greater than or equal to a fixed number k , $1 \leq k \leq n$. The number of such runs, called k -runs of 1s, is a random variable denoted as $G_{n,k}$. Let $D_{n,k}$, $3 \leq 2k + 1 \leq n$, be a random variable representing the distance between the first - last 1 of the first - last k -run of 1s in $\{X_t\}_{t=1}^n$. The study provides recursive schemes for $P(D_{n,k} = d \mid G_{n,k} \geq 2)$, $2k + 1 \leq d \leq n$, $1 \leq k \leq \lfloor (n - 1)/2 \rfloor$. A numerical example is presented.

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

ΑΝΑΦΟΡΕΣ

- Αράπης, Α.Ν., Μακρή, Φ.Σ. και Ψυλλάκης, Ζ.Μ. (2015). Μήκος και θέση της μικρότερης ακολουθίας που περιέχει ροές επιτυχιών. *ΕΣΙ: Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής*, 22-31.
- Arapis, A.N., Makri, F.S. and Psillakis, Z.M. (2016). On the length and the position of the minimum sequence containing all runs of ones in a Markovian binary sequence. *Statist. Probab. Lett.* **116**, 45-54.
- Avery, P.J. and Henderson, D.A. (1999). Detecting a changed segment in DNA sequences. *Appl. Statist.* **48**, Part 4, 489-503.
- Balakrishnan, N. and Koutras, M. V. (2002). *Runs and Scans with Applications*, New York: John Wiley.
- Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-580.
- Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Vol. I, 3rd Ed., New York: John Wiley.
- Makri, F.S. and Psillakis, Z.M. (2011). On success runs of length exceeded a threshold. *Methodol. Comput. Appl. Probab.* **13**, 269-305.
- Makri, F.S., Psillakis, Z.M. and Arapis, A.N. (2015). Length of the minimum sequence containing repeats of success runs. *Statist. Probab. Lett.* **96**, 28-37.
- Mood, A.M., (1940). The distribution theory of runs. *Ann. Math. Stat.* **11**, 367-392.

Ένα μοντέλο ουράς επαναλαμβανόμενων αφίξεων με εφαρμογή στην αποτίμηση απόδοσης του πρωτοκόλλου ελέγχου μετάδοσης πακέτων δεδομένων στο διαδίκτυο

Ι. Δημητρίου, Ε.-Α. Ζησιμοπούλου

Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών
idimit@math.upatras.gr, eelli1712@gmail.com

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία μελετάται ένα μοντέλο ουράς επαναλαμβανόμενων αφίξεων με εφαρμογή στην αποτίμηση απόδοσης του πρωτοκόλλου ελέγχου μετάδοσης (Transmission Control Protocol, TCP) πακέτων δεδομένων στο διαδίκτυο. Η μετάδοση αυτή πραγματοποιείται μέσω ενός δικτύου δρομολογητών (router), που μπορούν να αποθηκεύσουν πεπερασμένο αριθμό πακέτων. Σε περίπτωση «μπλοκαρίσματος» ενός αφιχθέντος πακέτου απαιτείται η επαναπροώθηση του, η οποία μοντελοποιείται μέσω μιας ουράς αναμονής επαναλαμβανόμενων αφίξεων. Θεωρούμε ένα σύστημα με έναν υπάλληλο και πεπερασμένο χώρο αναμονής. Τα πακέτα που μπλοκάρονται, προωθούνται στην ουρά επαναλαμβανόμενων αφίξεων, πεπερασμένης χωρητικότητας, που είναι υπεύθυνη για την επαναπροώθησή τους. Υποθέτουμε ότι οι τιμές των βασικών παραμέτρων του συστήματος καθορίζονται από μια στοχαστική διαδικασία πεπερασμένου χώρου καταστάσεων (random environment), ενώ ταυτόχρονα λαμβάνουμε υπόψη τους χρονικούς περιορισμούς όσον αφορά την μετάδοση των μπλοκαρισμένων πακέτων, που ενδεχομένως να οδηγήσει και σε απώλεια αυτών (deadlines). Το σύστημα μελετάται με μια τριδιάστατη Μαρκοβιανή διαδικασία (QBD process) και με χρήση πινακοαναλυτικών μεθόδων υπολογίζουμε τα κύρια μέτρα απόδοσης του συστήματος. Με χρήση αυτών, χρήσιμα αριθμητικά αποτελέσματα μπορούν να εξαχθούν.

Λέξεις κλειδιά: Ουρά επαναλαμβανόμενων αφίξεων; Πρωτόκολλο ελέγχου μετάδοσης (TCP); Μαρκοβιανή διαδικασία; Μέτρα απόδοσης; Πινακοαναλυτικές μέθοδοι; QBD.

1. ΕΙΣΑΓΩΓΗ

Το Internet χρησιμοποιεί την τεχνολογία μεταγωγής πακέτων για τη μεταφορά των δεδομένων. Κατά την περιήγησή μας στο διαδίκτυο στην ουσία πραγματοποιούμε ανταλλαγή δεδομένων η οποία γίνεται με την μετάδοση πακέτων δεδομένων που αποτελεί τη βασική μονάδα πληροφορίας που μεταφέρεται μέσα στα δίκτυα, με τη μορφή μιας τυχαίας ακολουθίας από συνεχόμενα bit. Χρειάζεται ένα σύνολο από

πρωτόκολλα που να καθορίζουν το πώς ανταλλάσσουν μεταξύ τους οι τερματικοί υπολογιστές τα πακέτα δεδομένων. Αυτό το σύνολο πρωτοκόλλων προσφέρει το πρωτόκολλο ελέγχου μετάδοσης (Transmission Control Protocol, TCP).

Η δρομολόγηση των πακέτων δεδομένων μέσα από τα διάφορα δίκτυα επιτυγχάνεται μέσω ενός δικτύου δρομολογητών (routers). Η δουλειά των routers είναι να δρομολογούν τα πακέτα των δεδομένων μέσα από τα διάφορα δίκτυα που αποτελούν το Internet μέχρις ότου τα επιδώσουν στον προορισμό τους. Το πρωτόκολλο ελέγχου μετάδοσης εγγυάται ότι τα πακέτα θα παραδοθούν στον προορισμό τους, ότι θα φτάσουν με τη σειρά με την οποία στάλθηκαν και ότι τα περιεχόμενά τους θα φτάσουν αναλλοίωτα. Σε περίπτωση που χαθεί κάποιο πακέτο δεδομένων κατά τη διάρκεια της μετάδοσης, το πρωτόκολλο αυτό υποθέτει πως η απώλεια οφείλεται στην συμφόρηση και το ξαναστέλνει.

Στην παρούσα εργασία θα μελετήσουμε ένα μοντέλο ουράς επαναλαμβανόμενων αφίξεων με εφαρμογή στην αποτίμηση απόδοσης του πρωτοκόλλου ελέγχου μετάδοσης πακέτων δεδομένων στο διαδίκτυο. Θεωρούμε ότι η μετάδοση αυτή πραγματοποιείται μέσω ενός δρομολογητή (router), που μπορεί να αποθηκεύσει πεπερασμένο αριθμό πακέτων. Στην περίπτωση που το buffer του δρομολογητή είναι γεμάτο, θεωρούμε ότι «μπλοκάρονται» και μεταφέρονται προσωρινά σε μια «εικονική» ουρά πεπερασμένης χωρητικότητας, από όπου πραγματοποιείται η επαναπροώθηση τους στον δρομολογητή. Η ουρά αυτή καλείται και ουρά επαναλαμβανόμενων αφίξεων (Fallin and Templeton (1997), Artalejo and Gomez-Corral (2008)). Αν όμως και αυτός είναι γεμάτος, τότε τα πακέτα χάνονται για πάντα από το σύστημα. Επιπλέον, έχουμε λάβει υπόψιν δεδομένους χρονικούς περιορισμούς (deadlines) σχετικά με την μετάδοση των πακέτων τόσο από την ουρά του δρομολογητή, όσο και στην ουρά επαναλαμβανόμενων αφίξεων, καθώς και η ριπαία κίνηση των πακέτων που οδηγούν στην θεώρηση μη στάσιμων τιμών των βασικών παραμέτρων ώστε να προσεγγίσουμε όσον το δυνατό γίνεται το πραγματικό σύστημα. Παρόμοια μοντέλα έχουν μελετηθεί στο πρόσφατο παρελθόν και από τους Avrachenkov and Yechiali (2008), Dudin and Klimenok (2012), χωρίς όμως την απαραίτητη λεπτομέρεια των (deadlines) και των μή στάσιμων χαρακτηριστικών. Το σύστημα μελετάται με μια τριδιάστατη Μαρκοβιανή διαδικασία (QBD process) και με χρήση πινακοαναλυτικών μεθόδων (Latouche and Ramaswami (1999), Neuts (1981), Gaver et al. (1984), Essia and Molle (2007)) υπολογίζουμε τα κύρια μέτρα απόδοσης του συστήματος.

Η εργασία δομείται ως εξής. Στην ενότητα 2. περιγράφουμε λεπτομερώς το μαθηματικό μοντέλο, ενώ στην ενότητα 3. προχωρούμε στην μελέτη του συστήματος με χρήση πινακοαναλυτικών μεθόδων. Έχοντας υπολογίσει την στάσιμη κατανομή, την χρησιμοποιούμε στην ενότητα 4. για να προσδιορίσουμε ορισμένα σημαντικά μέτρα απόδοσης όπως ο αναμενόμενος αριθμός πακέτων στην ουρά του δρομολογητή και στην ουρά επαναλαμβανόμενων αφίξεων. Τέλος, στην ενότητα 5. παράθετουμε ένα αριθμητικό παράδειγμα για να παρατηρήσουμε την επίδραση των αλλαγών των τιμών των παραμέτρων στην απόδοση του συστήματος.

2. Το μοντέλο

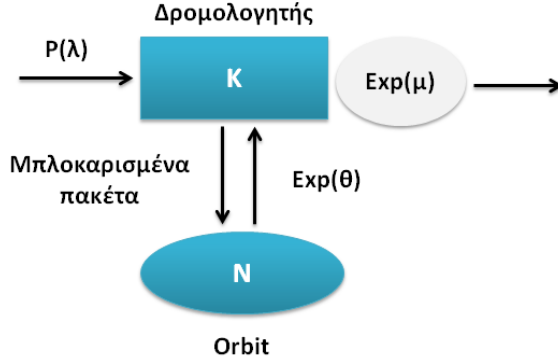
Ορίζουμε ένα σύστημα αναμονής με έναν υπάλληλο και πεπερασμένο χώρο αναμονής K (buffer), που παριστά τον υπό μελέτη δρομολογητή (Σχήμα 1). Λόγω της ριπαίας κίνησης (bursty traffic) των δεδομένων στο διαδίκτυο, οι παράμετροι του συστήματος καθορίζονται από μια στοχαστική διαδικασία πεπερασμένου χώρου καταστάσεων (random environment). Όταν η κατάσταση της διαδικασίας που περιγράφει το τυχαίο περιβάλλον είναι $i = 1, 2, \dots, M$, τα πακέτα φθάνουν σύμφωνα με την κατανομή Poisson παραμέτρου λ_i , ενώ τα πακέτα μεταδίδονται από τον δρομολογητή σύμφωνα με την εκθετική κατανομή παραμέτρου μ_i . Όταν ένα πακέτο φτάσει βρίσκοντας γεμάτο το buffer του δρομολογητή, μεταφέρεται προσωρινά στην ουρά επαναλαμβανόμενων αφίξεων (orbit), πεπερασμένης χωρητικότητας N , η οποία επαναπροωθεί τα «μπλοκαρισμένα» πακέτα στον δρομολογητή μετά από χρονικό διάστημα που ακολουθεί εκθετική κατανομή παραμέτρου θ_i .

Στην πράξη, μόλις σταλλεί ένα πακέτο, ο αποστολέας ενεργοποιεί ένα χρονόμετρο για να εξασφαλίσει την μεταφορά του πακέτου από τον παραλήπτη. Ο υπολογισμός του χρόνου επαναπροσπάθειας αποτέλεσε αντικείμενο μελέτης από τους Jakobson and Karrel (1988), Paxson and Allman (2000) και αποφανθήκαν ότι ένας από τους αποτελεσματικότερους τρόπους ήταν η επιλογή της εκθετικής ως κατανομής του χρόνου επαναπροσπάθειας (exponential back-off timer). Έχουν προταθεί και άλλοι τρόποι βασιζόμενοι στην εκτίμηση του χρόνου μετάδοσης μετέπιστροφή (round-trip time: η χρονική διάρκεια που απαιτείται για να μεταδοθεί ένα σήμα από τον αποστολέα στον παραλήπτη, συν την χρονική διάρκεια που απαιτείται για να σταλεί πίσω στον αποστολέα ή επιβεβαίωση της λήψης του σήματος από τον παραλήπτη).

Αν η ουρά επαναλαμβανόμενων αφίξεων είναι γεμάτη, τότε το εν λόγω πακέτο θα χαθεί. Κάθε πακέτο που τοποθετείται στο buffer του δρομολογητή πρέπει να μεταδοθεί εντός χρονικού διαστήματος που ακολουθεί την εκθετική κατανομή παραμέτρου ζ_i και αν αυτό παρέλθει, η μετάδοση του εν λόγω πακέτου αναβάλλεται, ενώ αυτό προωθείται στην ουρά επαναλαμβανόμενων αφίξεων. Σε πολλές πρακτικές εφαρμογές, το χρονικό όριο (delay bound) μετάδοσης των πακέτων είναι ώρες ή και ημέρες. Το μήκος του χρονικού διαστήματος καθορίζει και την σπουδαιότητα του εν λόγω πακέτου στην εκάστοτε εφαρμογή. Αν δεν μεταδοθούν εντός χρονικού διαστήματος που καθορίζεται από τον διαχειριστή (MAC protocol), παύουν να είναι χρήσιμα. Συνήθως, αυτό το χρονικό διάστημα είναι σταθερό. Για λόγους μοντελοποίησης κάνουμε την υπόθεση ότι προέρχεται από την εκθετική κατανομή. Τα μπλοκαρισμένα πακέτα πρέπει επίσης να επαναπροωθηθούν στο buffer του δρομολογητή εντός χρονικού διαστήματος που ακολουθεί την εκθετική κατανομή παραμέτρου γ_i . Αν αυτό παρέλθει τότε το εν λόγω πακέτο χάνεται.

Ορίζουμε $N_1(t)$ να είναι ο αριθμός των πελατών στο χώρο αναμονής buffer και στην εξυπηρέτηση και $N_2(t)$ να είναι ο αριθμός των πελατών στην ουρά επαναλαμβανόμενων αφίξεων τη χρονική στιγμή t . Το επίπεδο (level) αντιστοιχεί στον αριθμό των πελατών στο χώρο αναμονής. Επίσης συμβολίζουμε με $I(t)$ την

Σχήμα 1: Το μοντέλο.



κατάσταση του περιβάλλοντος την χρονική στιγμή t . Τότε η διαδικασία $X(t) = \{N_1(t), N_2(t), I(t) \mid t \geq 0\}$ είναι μια τριδιάστατη Μαρκοβιανή διαδικασία (QBD process) στο χώρο καταστάσεων $S = \{0, 1, \dots, K\} \times \{0, 1, \dots, N\} \times \{1, \dots, M\}$.

Θεωρούμε ότι οι παράμετροι του συστήματος καθορίζονται από μια στοχαστική διαδικασία πεπερασμένου χώρου καταστάσεων (random environment). Συμβολίζουμε τις καταστάσεις με $i = 1, 2, \dots, M$. Η διαδικασία που περιγράφει το τυχαίο περιβάλλον παραμένει για εκθετικό χρονικό διάστημα στην κατάσταση i , παραμέτρου η_i . Μόλις αυτό παρέλθει, με πιθανότητα q_{ij} αυτή μεταβαίνει στην $j \in \{1, 2, \dots, M\}$.

Συμβολίζουμε με $\eta_i q_{ij}$ τον ρυθμό μετάβασης από την κατάσταση i στην κατάσταση j , με $\sum_{j=1, j \neq i}^M q_{ij} = 1$. Ο γεννήτορας διάστασης $M \times M$ που μας δίνει τους ρυθμούς μετάβασης των καταστάσεων είναι ο εξής,

$$\bar{D} = \begin{pmatrix} -\eta_1 & \eta_1 q_{12} & \cdots & \eta_1 q_{1M} \\ \eta_2 q_{21} & -\eta_2 & \cdots & \eta_2 q_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \eta_M q_{M1} & \eta_M q_{M2} & \cdots & -\eta_M \end{pmatrix}.$$

Συγκεντρωτικά, οι πιθανές μεταβάσεις και οι αντίστοιχοι ρυθμοί φαίνονται στον Πίνακα 1.

Πίνακας 1: Ρυθμοί Μετάβασης Συστήματος

Από	Σε	Ρυθμός	
(n_1, n_2, i)	$(n_1 + 1, n_2, i)$	λ_i	για $0 \leq n_1 \leq K - 1, 0 \leq n_2 \leq N$
(n_1, n_2, i)	$(n_1 - 1, n_2, i)$	μ_i	για $1 \leq n_1 \leq K, 0 \leq n_2 \leq N$
(n_1, n_2, i)	$(n_1 + 1, n_2 - 1, i)$	θ_i	για $0 \leq n_1 \leq K - 1, 1 \leq n_2 \leq N$
(n_1, n_2, i)	$(n_1, n_2 - 1, i)$	$n_2 \gamma_i$	για $0 \leq n_1 \leq K, 1 \leq n_2 \leq N$
(n_1, n_2, i)	$(n_1 - 1, n_2 + 1, i)$	ζ_i	για $1 \leq n_1 \leq K, 0 \leq n_2 \leq N - 1$
(n_1, n_2, i)	(n_1, n_2, j)	$\eta_i q_{ij}$	για $0 \leq n_1 \leq K, 0 \leq n_2 \leq N, i \neq j$

Ο απειροστός γεννήτορας Q της διαδικασίας, είναι σύνθετος-τριδιαγώνιος πίνακας διάστασης $(K + 1)M(N + 1) \times (K + 1)M(N + 1)$ και έχει την ακόλουθη μορφή,

$$Q = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & \cdots & K-2 & K-1 & K \\ A_0 & A_1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ B_0 & B_1 & A_1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & B_0 & B_1 & A_1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & B_0 & B_1 & A_1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & B_0 & B_1 & A_1 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & B_0 & B_2 \end{pmatrix} \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \\ \vdots \\ K-1 \\ K \end{matrix}$$

Οι υποπίνακες A_0, A_1, B_0, B_1 και B_2 είναι τετραγωνικοί πίνακες διάστασης $(N + 1)M \times (N + 1)M$. Ο υποπίνακας A_1 περιγράφει τους ρυθμούς μετάβασης που αφορούν άφιξη στο buffer του δρομολογητή, είτε από έξω, είτε από το orbit και ο υποπίνακας B_0 περιγράφει τους ρυθμούς μετάβασης που αναφέρονται σε αναχώρηση πελάτη, είτε λόγω εξυπηρέτησης, είτε λόγω ανυπομονησίας. Συγκεκριμένα,

$$A_1 = \begin{pmatrix} E_1 & & & & \\ Z_1 & E_1 & & & \\ & Z_1 & E_1 & & \\ & & \ddots & \ddots & \\ & & & Z_1 & E_1 \end{pmatrix},$$

όπου,

$$E_1 = \underset{M \times M}{diag}(\lambda_1, \dots, \lambda_M), \quad Z_1 = \underset{M \times M}{diag}(\theta_1, \dots, \theta_M).$$

$$B_0 = \begin{pmatrix} K_1 & \Lambda_1 & & & \\ & K_1 & \Lambda_1 & & \\ & & K_1 & \Lambda_1 & \\ & & & \ddots & \ddots \\ & & & & K_1 & \Lambda_1 \\ & & & & & K_1 \end{pmatrix},$$

όπου,

$$K_1 = \underset{M \times M}{diag}(\mu_1, \dots, \mu_M), \Lambda_1 = \underset{M \times M}{diag}(\zeta_1, \dots, \zeta_M).$$

$$A_0 = \begin{pmatrix} \hat{A}_{00} & & & & \\ \hat{A}_{10} & \hat{A}_{11} & & & \\ & \hat{A}_{21} & \hat{A}_{22} & & \\ & & \ddots & \ddots & \\ & & & \hat{A}_{NN-1} & \hat{A}_{NN} \end{pmatrix},$$

όπου,

$$\hat{A}_{nn-1} = n \cdot \underset{M \times M}{diag}(\gamma_1, \dots, \gamma_M), n = 1, \dots, N$$

και

$$\hat{A}_{nn} = \begin{pmatrix} \hat{a}_{11}^{(n)} & \eta_1 q_{12} & \cdots & \eta_1 q_{1M} \\ \eta_2 q_{21} & \hat{a}_{22}^{(n)} & \cdots & \eta_2 q_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \eta_M q_{M1} & \eta_M q_{M2} & \cdots & \hat{a}_{MM}^{(n)} \end{pmatrix},$$

όπου,

$$\hat{a}_{ii}^{(n)} = \begin{cases} -(\eta_i + \lambda_i), & \text{όταν } n = 0. \\ -(\eta_i + \lambda_i + \theta_i + n\gamma_i) & \text{όταν } n = 1, \dots, N. \end{cases}, \quad i = 1, 2, \dots, M.$$

$$B_1 = \begin{pmatrix} \hat{B}_{00} & & & & \\ \hat{B}_{10} & \hat{B}_{11} & & & \\ & \hat{B}_{21} & \hat{B}_{22} & & \\ & & \ddots & \ddots & \\ & & & \hat{B}_{NN-1} & \hat{B}_{NN} \end{pmatrix},$$

όπου,

$$\hat{B}_{nn-1} = n \cdot \underset{M \times M}{diag}(\gamma_1, \dots, \gamma_M), \quad n = 1, \dots, N$$

και

$$\hat{B}_{nn} = \begin{pmatrix} \hat{b}_{11}^{(n)} & \eta_1 q_{12} & \cdots & \eta_1 q_{1M} \\ \eta_2 q_{21} & \hat{b}_{22}^{(n)} & \cdots & \eta_2 q_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \eta_M q_{M1} & \eta_M q_{M2} & \cdots & \hat{b}_{MM}^{(n)} \end{pmatrix},$$

όπου,

$$\hat{b}_{ii}^{(n)} = \begin{cases} -(\eta_i + \lambda_i + \mu_i + \zeta_i), & \text{όταν } n = 0. \\ -(\eta_i + \lambda_i + \mu_i + \zeta_i + \theta_i + n\gamma_i) & \text{όταν } n = 1, \dots, N-1. \\ -(\eta_i + \lambda_i + \mu_i + \theta_i + n\gamma_i) & \text{όταν } n = N. \end{cases}, \quad i = 1, 2, \dots, M.$$

Τέλος,

$$B_2 = \begin{pmatrix} \hat{\hat{B}}_{00} & & & & \\ \hat{\hat{B}}_{10} & \hat{\hat{B}}_{11} & & & \\ & \hat{\hat{B}}_{21} & \hat{\hat{B}}_{22} & & \\ & & \ddots & \ddots & \\ & & & \hat{\hat{B}}_{N-1,N-1} & \hat{\hat{B}}_{N-1,N} \\ & & & & \hat{\hat{B}}_{NN} \end{pmatrix},$$

όπου,

$$\hat{\hat{B}}_{nn-1} = n \cdot \text{diag}_{M \times M}(\gamma_1, \dots, \gamma_M), \quad n = 1, \dots, N$$

και

$$\hat{\hat{B}}_{nn} = \begin{pmatrix} \hat{\hat{b}}_{11}^{(n)} & \eta_1 q_{12} & \cdots & \eta_1 q_{1M} \\ \eta_2 q_{21} & \hat{\hat{b}}_{22}^{(n)} & \cdots & \eta_2 q_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \eta_M q_{M1} & \eta_M q_{M2} & \cdots & \hat{\hat{b}}_{MM}^{(n)} \end{pmatrix},$$

όπου,

$$\hat{\hat{b}}_{ii}^{(n)} = \begin{cases} -(\eta_i + \mu_i + \zeta_i), & \text{όταν } n = 0. \\ -(\eta_i + \mu_i + \zeta_i + n\gamma_i) & \text{όταν } n = 1, \dots, N-1. \\ -(\eta_i + \mu_i + N\gamma_i) & \text{όταν } n = N. \end{cases}, \quad i = 1, 2, \dots, M.$$

3. Εύρεση στάσιμης κατανομής

Εστω ότι $\tilde{\pi} \equiv (\tilde{\pi}_0, \tilde{\pi}_1, \dots, \tilde{\pi}_K)$ είναι το διάνυσμα των στάσιμων πιθανοτήτων το οποίο θέλουμε να υπολογίσουμε. Ορίζουμε,

$$\tilde{\pi}_j = (\tilde{\pi}_{j0}, \tilde{\pi}_{j1}, \dots, \tilde{\pi}_{jN}), \quad j = 0, 1, \dots, K,$$

να είναι το διάνυσμα των πιθανοτήτων να υπάρχουν j πελάτες στο buffer και στην εξυπηρέτηση όπου,

$$\tilde{\pi}_{jn} = (\pi_{jn1}, \pi_{jn2}, \dots, \pi_{jnM}), \quad j = 0, 1, \dots, K, \quad n = 0, 1, \dots, N,$$

είναι το διάνυσμα των πιθανοτήτων να υπάρχουν j πελάτες στο buffer και στην εξυπηρέτηση και n πελάτες στο orbit.

Έτσι λοιπόν συμβολίζουμε με,

$$\pi_{jnm} = \lim_{t \rightarrow \infty} P(N_1(t) = j, N_2(t) = n, I(t) = m)$$

τις οριακές πιθανότητες να υπάρχουν j πελάτες στο buffer και στην εξυπηρέτηση και n πελάτες στο orbit, στην κατάσταση m .

Για την εύρεση των πιθανοτήτων επιλύουμε το σύστημα,

$$\begin{cases} \tilde{\pi}Q = 0 \\ \tilde{\pi}\tilde{e} = 1 \end{cases},$$

όπου με $\tilde{e}$ συμβολίζουμε το διάνυσμα στήλη που αποτελείται από μονάδες.

Επομένως, επιλύοντας την εξίσωση,

$$\tilde{\pi}Q = 0,$$

προκύπτουν οι παρακάτω πινακικές εξισώσεις.

Για $j = 0$,

$$\tilde{\pi}_0 A_0 + \tilde{\pi}_1 B_0 = 0 \tag{1}$$

Για $1 \leq j \leq K - 1$,

$$\begin{cases} \tilde{\pi}_0 A_1 + \tilde{\pi}_1 B_1 + \tilde{\pi}_2 B_0 = 0 \\ \tilde{\pi}_1 A_1 + \tilde{\pi}_2 B_1 + \tilde{\pi}_3 B_0 = 0 \\ \tilde{\pi}_2 A_1 + \tilde{\pi}_3 B_1 + \tilde{\pi}_4 B_0 = 0 \\ \vdots \end{cases}$$

Δηλαδή για $1 \leq j \leq K - 1$, αναδρομικά προκύπτει η σχέση,

$$\tilde{\pi}_{j-1} A_1 + \tilde{\pi}_j B_1 + \tilde{\pi}_{j+1} B_0 = 0 \tag{2}$$

Τέλος, για $j = K$ έχουμε,

$$\tilde{\pi}_{K-1} A_1 + \tilde{\pi}_K B_2 = 0 \tag{3}$$

Ο πίνακας B_0 είναι άνω τριγωνικός με τα στοιχεία της κυρίας διαγωνίου να αποτελούν οι ρυθμοί εξυπηρέτησης σε κάθε φάση λειτουργίας. Επομένως, θεωρώντας ότι αυτοί είναι πάντα διάφοροι του μηδενός, ο B_0 είναι αντιστρέψιμος. Επομένως μπορούμε να λύσουμε την εξίσωση (1) ως προς $\tilde{\pi}_1$,

$$\tilde{\pi}_1 = -\tilde{\pi}_0 A_0 B_0^{-1} = \tilde{\pi}_0 R_1, \quad \text{όπου } R_1 = -A_0 B_0^{-1}. \tag{4}$$

Για $j = 1$ η σχέση (2) γίνεται,

$$\tilde{\pi}_0 A_1 + \tilde{\pi}_1 B_1 + \tilde{\pi}_2 B_0 = 0$$

και αντικαθιστώντας την σχέση (4) έχουμε,

$$\tilde{\pi}_0 A_1 + (-\tilde{\pi}_0 A_0 B_0^{-1}) B_1 + \tilde{\pi}_2 B_0 = 0$$

ή

$$\tilde{\pi}_2 = -\tilde{\pi}_0 (A_1 - A_0 B_0^{-1} B_1) B_0^{-1} = \tilde{\pi}_0 R_2, \quad \text{όπου } R_2 = -(A_1 - A_0 B_0^{-1} B_1) B_0^{-1}. \quad (5)$$

Για $j = 2$ η σχέση (2) γίνεται,

$$\tilde{\pi}_1 A_1 + \tilde{\pi}_2 B_1 + \tilde{\pi}_3 B_0 = 0$$

και αντικαθιστώντας τις σχέσεις (4) και (5) έχουμε,

$$-\tilde{\pi}_0 A_0 B_0^{-1} A_1 - \tilde{\pi}_0 (A_1 - A_0 B_0^{-1} B_1) B_0^{-1} B_1 + \tilde{\pi}_3 B_0 = 0.$$

ή

$$\begin{aligned} \tilde{\pi}_3 &= \tilde{\pi}_0 (A_0 B_0^{-1} A_1 + (A_1 - A_0 B_0^{-1} B_1) B_0^{-1} B_1) B_0^{-1} = \tilde{\pi}_0 R_3, \\ \text{όπου } R_3 &= (A_0 B_0^{-1} A_1 + (A_1 - A_0 B_0^{-1} B_1) B_0^{-1} B_1) B_0^{-1} \end{aligned} \quad (6)$$

Γενικεύοντας για $j = K - 1$ έχουμε ότι,

$$\tilde{\pi}_{K-2} A_1 + \tilde{\pi}_{K-1} B_1 + \tilde{\pi}_K B_0 = 0$$

ή

$$\tilde{\pi}_K = \tilde{\pi}_0 R_K \quad (7)$$

Άρα, από τις σχέσεις (4), (5), (6) και (7) προκύπτει ότι,

$$\tilde{\pi}_j = \tilde{\pi}_0 R_j, \quad \text{για } j = 1, 2, \dots, K. \quad (8)$$

Για $1 \leq j \leq K - 1$, η γενική μορφή των πινάκων R_j μπορεί να βρεθεί αντικαθιστώντας την σχέση (8) στην σχέση (2),

$$\tilde{\pi}_0 R_{j-1} A_1 + \tilde{\pi}_0 R_j B_1 + \tilde{\pi}_0 R_{j+1} B_0 = 0 \Leftrightarrow R_{j+1} = -(R_{j-1} A_1 + R_j B_1) B_0^{-1}.$$

Συγκεντρώνοντας, ο πίνακας R_j υπολογίζεται από τον παρακάτω αλγόριθμο.

Επομένως γίνεται αντιληπτό ότι για να μπορέσουμε να βρούμε τις ζητούμενες πιθανότητες, θα πρέπει να υπολογίσουμε το υποδιάγραμμα πιθανοτήτων $\tilde{\pi}_0$.

Αντικαθιστώντας την σχέση (8) στην (3) προκύπτει ότι,

$$\tilde{\pi}_0 (R_{K-1} A_1 + R_K B_2) = 0. \quad (9)$$

Algorithm 1 Υπολογισμός R_j

```
1: if  $j = 0$  then
2:   return  $R_0 \leftarrow I$ 
3: end if
4: if  $j = 1$  then
5:   return  $R_1 \leftarrow -A_0 B_0^{-1}$ 
6: end if
7: if  $j \geq 2$  then
8:   for  $j = 2 \rightarrow K$  do
9:     return  $R_j \leftarrow -(R_{j-2} A_1 + R_{j-1} B_1) B_0^{-1}$ 
10:  end for
11: end if
```

Επίσης, από τη συνθήκη κανονικότητας $\sum_{j=0}^K \tilde{\pi}_j \tilde{e} = 1$ έχουμε ότι,

$$\tilde{\pi}_0 \sum_{j=0}^K R_j \tilde{e} = 1. \quad (10)$$

Επιλύοντας την σχέση (9) μαζί με την (10) μπορούμε να βρούμε το διάνυσμα πιθανοτήτων $\tilde{\pi}_0 = (\tilde{\pi}_{00}, \tilde{\pi}_{01}, \dots, \tilde{\pi}_{0N})$, αντικαθιστώντας μία εξίσωση από το σύστημα (9) με την εξίσωση της κανονικότητας. Στη συνέχεια αντικαθιστώντας το $\tilde{\pi}_0$ στην σχέση (8) μπορούμε να υπολογίσουμε τις πιθανότητες $\tilde{\pi}_j$, για $j = 1, 2, \dots, K$ και άρα και το ζητούμενο διάνυσμα των στάσιμων πιθανοτήτων.

4. Μέτρα απόδοσης

Αφού υπολογίσουμε το διάνυσμα αυτό, μπορούμε να βρούμε κάποια σημαντικά μέτρα απόδοσης, όπως για παράδειγμα τον μέσο αριθμό των πελατών στο buffer και στην εξυπηρέτηση,

$$E(N_1) = \sum_{j=1}^K j \sum_{n=0}^N \sum_{m=1}^M \pi_{jnm}$$

και τον μέσο αριθμό των πελατών στο orbit,

$$E(N_2) = \sum_{n=1}^N n \sum_{j=0}^K \sum_{m=1}^M \pi_{jnm}.$$

5. Αριθμητικά αποτελέσματα

Έστω ότι $K = 3$, $M = 2$ και $N = 4$.

Πίνακας 2: Παράμετροι Συστήματος

Παράμετροι	Φάση 1	Φάση 2
λ πακέτα/min	$\lambda_1 = 0.5, \lambda_1 = 1.5$	$\lambda_2 = 0.1$
μ min/πακέτο	$\mu_1 = 2$	$\mu_2 = 5$
θ (σε min)	θ_1	$\theta_2 = 4$
ζ (σε min)	$\zeta_1 = 1$	$\zeta_2 = 0.5$
γ (σε min)	$\gamma_1 = 2$	$\gamma_2 = 1$
η (σε min)	$\eta_1 = 1$	$\eta_2 = 2$

Ο πίνακας 2 περιέχει τις τιμές των παραμέτρων του συστήματος, για τις δύο φάσεις λειτουργίας του. Εφόσον θέλουμε να μελετήσουμε την μεταβολή του μέσου αριθμού των πελατών στην κανονική και στην εικονική ουρά (orbit) για αυξανόμενες τιμές του ρυθμού των επαναπροσπαθειών στην φάση 1, υποθέτουμε ότι ο ρυθμός αυτός μεταβάλλεται. Παράλληλα μελετάμε πώς επηρεάζονται οι μέσοι αριθμοί των πελατών στην κανονική και στην εικονική ουρά, όταν ο ρυθμός αφίξεων λ_1 αυξάνεται.

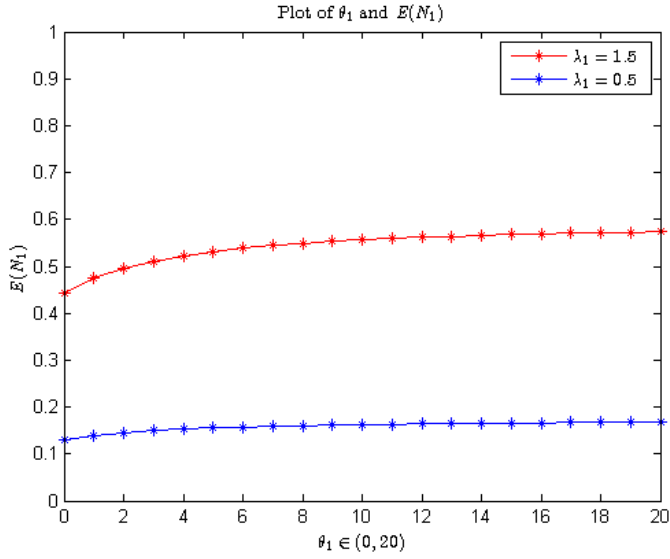
Στο Σχήμα 2 παρατηρούμε πως ο μέσος αριθμός πελατών στην κανονική ουρά $E(N_1)$ αυξάνεται για αυξανόμενες τιμές του ρυθμού επαναπροσπαθειών στην φάση 1. Όπως αναμενόταν, αυξάνοντας το θ_1 ο αριθμός των πελατών στην κανονική ουρά αυξάνει, αφού πελάτες από την εικονική ουρά προστίθενται στην κανονική ουρά. Μάλιστα η αύξηση αυτή είναι ακόμα περισσότερο αισθητή αν αυξήσουμε τον ρυθμό των νέων πελατών που φθάνουν στο σύστημα λ_1 . Επιπλέον, παρατηρούμε ότι μετά από κάποια τιμή του θ_1 ο μέσος αριθμός στην κανονική ουρά σταθεροποιείται. Αυτό είναι επίσης αναμενόμενο αφού έχουμε πεπερασμένο χώρο αναμονής στην κανονική ουρά και επομένως μετά από κάποιο σημείο όσο και να αυξάνουμε το θ_1 δεν θα έχουμε περαιτέρω μεταβολή στον μέσο αριθμό πελατών στην ουρά.

Στο Σχήμα 3, παρατηρούμε πως ο μέσος αριθμός πελατών στην εικονική ουρά $E(N_2)$ μειώνεται για αυξανόμενες τιμές του ρυθμού επαναπροσπαθειών στην φάση 1. Μάλιστα η μείωση αυτή είναι ακόμα περισσότερο αισθητή αν αυξήσουμε τον ρυθμό των νέων πελατών που φθάνουν στο σύστημα λ_1 . Επιπλέον, παρατηρούμε ότι και σ' αυτή την περίπτωση μετά από κάποια τιμή του θ_1 ο μέσος αριθμός στην εικονική ουρά σταθεροποιείται αφού έχουμε πεπερασμένο χώρο αναμονής.

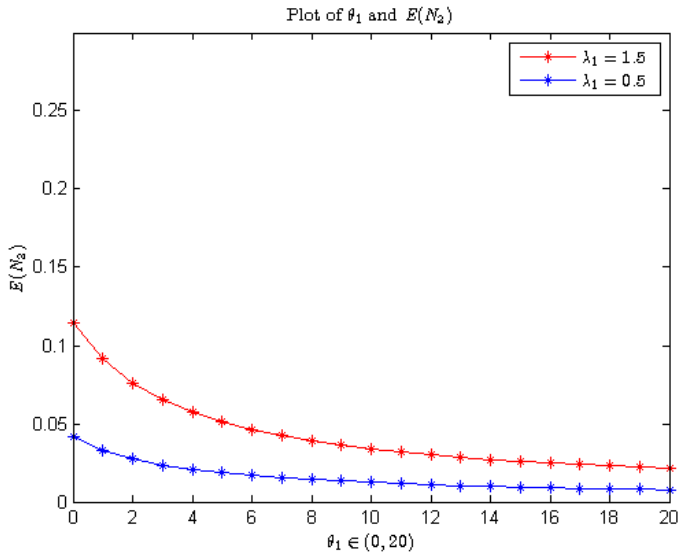
6. ΣΥΜΠΕΡΑΣΜΑ

Στην παρούσα εργασία, μελετήθηκε ένα νέο μοντέλο ουράς επαναλαμβανόμενων αφίξεων. Η καινοτομία αφορά κυρίως στην προσπάθεια να συμπεριληφθούν σε μια εργασία τα ρεαλιστικά χαρακτηριστικά της ανυπομονησίας και της ετερογένειας των βασικών χαρακτηριστικών (αφίξεων) σε ένα μοντέλο επαναλαμβανόμενων αφίξεων με άμεση εφαρμογή στο πρωτόκολλο ελέγχου TCP που χρησιμοποιείται στο διαδίκτυο. Το μοντέλο μελετήθηκε με μια τριδιάστατη Μαρκοβιανή διαδικασία (QBD process)

Σχήμα 2: Μεταβολή του μέσου αριθμού πελατών στην κανονική ουρά.



Σχήμα 3: Μεταβολή του μέσου αριθμού πελατών στην εικονική ουρά.



και με χρήση πινακοαναλυτικών μεθόδων υπολογίστηκε η στάσιμη κατανομή και στη συνέχεια τα κύρια μέτρα απόδοσης του συστήματος.

ABSTRACT

In the present paper we present a new model of retrial queue with application in the performance evaluation of the Transmission Control Protocol (TCP). The transmission takes place through the routers that can store a finite number of packets. In case of "blocking" a mover package requires the return of which is modeled by a retrial queue. We consider a queueing system with a single-server and finite buffer, which represents the router under study. Because of bursty traffic of data, the parameters of the system are determined by a stochastic process with finite state space (random environment). When the state of the process which describes the random environment is $i = 1, 2, \dots, M$, the packets arrive according to the Poisson distribution of parameter λ_i while the packets transmitted by the router according to the exponential distribution of parameter μ_i . When an arriving packet finds the router's buffer full, temporarily transferred to the orbit with finite state space N , which returns the "blocked" packets to the router after a period of time that follows the exponential distribution of parameter θ_i . If it is full, then this packet will be lost. Each packet placed in the router's buffer must be transmitted within a period of time following the exponential distribution of parameter ζ_i and if it expires, the transmission of the packet is postponed while it is moved to the orbit. Also, the blocked packets must be returned to the buffer within a period of time that follows the exponential distribution of parameter γ_i . If it is exceeded then this packet is lost. We study this system using a three-dimensional Markov QBD process and we calculate the system's main performance measures using matrix-analytic methods.

Keywords: Retrial queueing system; Transmission Control Protocol (TCP); Markov process; Stationary performance measures; Matrix-analytic methods; QBD process.

ΑΝΑΦΟΡΕΣ

- Artalejo, J.R. and Gomez-Corral, A. (2008). *Retrial Queueing Systems, A Computational Approach*. Berlin Heidelberg: Springer-Verlag.
- Avrachenkov, K. and Yechiali, U. (2008). Retrial Networks with finite buffers and their application to internet Data traffic. *Probability in the Engineering and Informational Sciences* **22**, 519 – 536.
- Dudin, A. and Klimenok, V. (2012). *Investigation of Queueing System Suitable for Mathematical Modelling of TCP Short Transfer*. Berlin Heidelberg: Springer-Verlag, 122 – 133.
- Essia, H.E. and Molle, M. (2007). On the Solution to QBD Processes with Finite State Space. *Stochastic Analysis and Applications* **25**, 763 – 779.
- Falin, G.I. and Templeton, J.G.C. (1997). *Retrial Queues*. London: Chapman and Hall.
- Gaver, D.P., Jacobs, P.A. and Latouche, G. (1984). Finite birth-and-death models in randomly changing environments. *Advances in Applied Probability* **16**, 715 – 731.
- Jacobson, V. 1988. Congestion Avoidance and Control. *Computer Communication Review* **18** 314 – 329.

- Latouche, G. and Ramaswami, V. (1999). *Introduction to Matrix Analytic Methods in Stochastic Modeling*. Philadelphia: ASA-SIAM.
- Neuts, M.F. (1981). *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*, Baltimore: The Johns Hopkins University Press.
- Paxson, V. and Allman, M. (2000). Computing TCP's Retransmission Timer. RFC 2988.

Μια ουρά αναμονής με πολλαπλούς υπαλλήλους και ασύγχρονες διακοπές με εφαρμογή στην διαχείριση ενέργειας σε υποδομές φιλοξενίας υπολογιστικών συστημάτων

Ι. Δημητρίου, Α.-Φ. Καλτσάς

Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών, 26500, Πάτρα
idimit@math.upatras.gr, kaltsasan@hotmail.gr

ΠΕΡΙΛΗΨΗ

Η παρούσα εργασία αναφέρεται στο πρόβλημα της διαχείρισης μιας υποδομής φιλοξενίας υπολογιστικών συστημάτων (data centers) με απότερο στόχο την μείωση της κατανάλωσης ενέργειας, ενώ ταυτόχρονα να παρέχεται ένα αποδεκτό επίπεδο παροχής υπηρεσιών. Το μεγαλύτερο ποσοστό της καταναλούμενης ενέργειας οφείλεται σε εξυπηρετητές (servers), οι οποίοι βρίσκονται σε πλήρη λειτουργία χωρίς να υπάρχουν εργασίες (jobs) προς εξυπηρέτηση. Μια τεχνική για την μείωση της κατανάλωσης ενέργειας μιας τέτοιας δομής είναι η απενεργοποίηση αυτών των εξυπηρετητών. Στην παρούσα εργασία μοντελοποιούμε μια τέτοια δομή με ένα γενικό σύστημα ουράς πεπερασμένου χώρου αναμονής με πολλαπλούς εξυπηρετητές, ανυπόμονους πελάτες και ασύγχρονες διακοπές υπαλλήλων, που η λειτουργία του εναλλάσσεται σε ένα τυχαίο περιβάλλον πολλαπλών φάσεων. Ορίζουμε μια τρι-διάστατη Μαρκοβιανή διαδικασία που ανήκει στην κλάση των QBD διαδικασιών και με χρήση πινακο-αναλυτικών μεθόδων προσδιορίζουμε την στάσιμη κατανομή και τα βασικά μέτρα απόδοσης. Τέλος, παρουσιάζονται αριθμητικά αποτελέσματα, που παρέχουν πληροφορία σχετικά με την λειτουργία του όλου συστήματος.

Λέξεις - Κλειδιά: Ουρές αναμονής, διακοπές εξυπηρετητών, Διαχείριση ενέργειας, Data centers, Cloud computing.

1. Εισαγωγή

Οι υποδομές φιλοξενίας υπολογιστικών συστημάτων (data centers) αποτελούνται από ένα μεγάλο αριθμό Η/Υ, υποστηρίζουν την λειτουργία μεγάλων οργανισμών (π.χ. Google) και αποτελούν την βάση επιχειρηματικών μοντέλων παροχής υπηρεσιών (Cloud Computing). Η λειτουργία τους απαιτεί μεγάλες ποσότητες ενέργειας και επομένως, είναι αναγκαίος ο καθορισμός τεχνικών μείωσης της διατηρώντας σε επιτρεπτά επίπεδα την παρεχόμενη ποιότητα εξυπηρέτησης.

Το κόστος της κατανάλωσης ενέργειας των (data centers) αυξάνεται συνεχώς κάθε χρόνο, ξεπερνώντας το 2010 τα 19 δις \$ και δημιουργώντας 11 εκ. τόνους

αερίων CO_2 (βλέπε Green Grid (2010)). Έχει παρατηρηθεί ότι με την υπάρχουσα τεχνολογική εξέλιξη, ένας άεργος εξυπηρετητής καταναλώνει περίπου το 60% της κατανάλωσης ενέργειας που αντιστοιχεί στην λειτουργία του σε περιόδους αιχμής (Greenberg et al. (2009)). Μια τεχνική σημαντικής μείωσης της κατανάλωσης ενέργειας αποτελεί η απενεργοποίηση εξυπηρετητών (H/Y) που δεν χρησιμοποιούνται.

Εκάθαρα, η απενεργοποίηση των H/Y θα συμβάλλει στην μείωση της κατανάλωσης ενέργειας, αλλά θα προκαλέσει καθυστερήσεις στην εξυπηρέτηση, που με την σειρά τους θα προκαλέσουν επιπλέον κόστος, που προέρχεται από την ενδεχόμενη παραβίαση όρων του συμβολαίου (Service Level Agreement (SLA)), που συνάπτεται μεταξύ παρόχου και πελατών. Από την άλλη πλευρά, η συνεχής απενεργοποίηση-ενεργοποίηση των H/Y ενδέχεται να προκαλέσει περαιτέρω αύξηση της κατανάλωσης ενέργειας αν δεν υλοποιηθεί σύμφωνα με συγκεκριμένους κανόνες, μιας και η μετάβαση ενός H/Y από την κατάσταση ύπνου (sleep mode) σε κατάσταση πλήρους λειτουργίας (busy mode) απαιτεί επίσης μεγάλη ποσότητα ενέργειας αλλά και χρόνου (Greenberg et al. (2009)).

Λαμβάνοντας υπόψιν τα παραπάνω και για να αποτιμήσουμε την απόδοση μιας τέτοιας δομής προτείνουμε και μελετάμε ένα σύστημα αναμονής με πολλαπλούς εξυπηρετητές, πεπερασμένο χώρο αναμονής και ασύγχρονες διακοπές, που περιγράφουν την κατάσταση ύπνου των H/Y. Υποθέτουμε ότι η λειτουργία του συστήματος εναλλάσσεται μεταξύ διαφορετικών φάσεων (random environment), που περιγράφουν το ρεαλιστικό δυναμικό περιβάλλον λειτουργίας μιας τέτοιας δομής (π.χ. λόγω της διαφορετικής ροής των εργασιών προς εξυπηρέτηση σε διαφορετικές χρονικές περιόδους (π.χ. peak hours)). Επομένως, οι βασικές παράμετροι του μοντέλου θα εξαρτώνται από την φάση λειτουργίας παρέχοντας την δυνατότητα μελέτης σε δυναμικό περιβάλλον. Κάθε εξυπηρετητής (server) που μένει άεργος (idle), ενεργοποιεί ένα χρονόμετρο που καθορίζει το μέγιστο χρονικό διάστημα που παραμένει διαθέσιμος προτού περιέλθει σε κατάσταση ύπνου (vacations). Αν κάποια εργασία φτάσει στην διάρκεια αυτής της περιόδου θα εξυπηρετηθεί άμεσα. Αν δεν υπάρχει κάποια άφιξη στην διάρκεια αυτής της περιόδου, ο H/Y εισέρχεται σε κατάσταση ύπνου (βλέπε Tian and Zhang (2006), Takagi (1991)), ώστε να περισταλεί η κατανάλωση ενέργειας. Με το πέρας αυτής ο H/Y επανέρχεται και αν υπάρχει εργασία προς εκτέλεση, θα αρχίσει άμεσα να εξυπηρετεί. Σε αντίθετη περίπτωση, ενεργοποιεί ξανά το χρονόμετρο και παραμένει άεργος όπως περιγράφηκε παραπάνω. Αν μια εργασία δεν βρει διαθέσιμο H/Y κατά την άφιξή της, θα παραμείνει στο σύστημα για χρονικό διάστημα με γνωστή κατανομή (deadline)· (βλέπε Altman and Yechiali (2006)). Αν δεν εξυπηρετηθεί σε αυτό το διάστημα θα αναχωρήσει, προκαλώντας επιπρόσθετο κόστος για τον πάροχο. Παρόμοια μοντέλα έχουν μελετηθεί στο παρελθόν και από τους Mitrani (2011, 2013), Phung-Duc (2015a,2015b).

Η εργασία δομείται ως εξής. Στην ενότητα 2. περιγράφεται λεπτομερώς το μαθηματικό μοντέλο, ενώ στην ενότητα 3. προχωρούμε στην μελέτη του συστήματος με χρήση πινακοαναλυτικών μεθόδων (βλέπε Latouche and Ramaswami (2001)). Έχοντας υπολογίσει την στάσιμη κατανομή, προσδιορίζονται στην ενότητα 4. ορισμένα

σημαντικά μέτρα απόδοσης, όπως ο αναμενόμενος αριθμός εργασιών στο σύστημα και η μέση κατανάλωση ενέργειας. Στην ενότητα 5. παρουσιάζεται ένα αριθμητικό παράδειγμα, ενώ ορισμένες σκέψεις για μελλοντική έρευνα παραθέτονται στην ενότητα 6..

2. Το μαθηματικό μοντέλο

Θεωρούμε ένα σύστημα αναμονής με c υπαλλήλους (servers), πεπερασμένο χώρο αναμονής L και ασύγχρονες διακοπές. Η λειτουργία του συστήματος εναλλάσσεται μεταξύ R διαφορετικών φάσεων (random environment), που περιγράφουν το δυναμικό περιβάλλον λειτουργίας μιας τέτοιας δομής και επηρεάζουν τις τιμές των βασικών της παραμέτρων. Στην διάρκεια της i φάσης, $i = 1, \dots, R$, που ακολουθεί την εκθετική κατανομή παραμέτρου η_i , οι εργασίες φθάνουν σύμφωνα με την κατανομή Poisson παραμέτρου λ_i , ενώ κάθε server παρέχει εξυπηρέτηση που ακολουθεί την εκθετική κατανομή παραμέτρου μ_i .

Όταν κάποιος εξυπηρετητής ολοκληρώσει μια εξυπηρέτηση και μείνει άεργος (idle), ενεργοποιεί ένα χρονόμετρο που καθορίζει τον μέγιστο χρόνο παραμονής του σε αυτή την κατάσταση. Το εν λόγω τυχαίο χρονικό διάστημα ακολουθεί εκθετική κατανομή παραμέτρου θ_i . Αν κάποια εργασία φτάσει στην διάρκεια αυτής της περιόδου θα εξυπηρετηθεί άμεσα. Αν δεν υπάρξει κάποια άφιξη στην διάρκεια του εν λόγω χρονικού διαστήματος, ο εξυπηρετητής εισέρχεται σε κατάσταση ύπνου (vacations), η διάρκεια της οποίας ακολουθεί την εκθετική κατανομή παραμέτρου γ_i . Στην διάρκεια αυτής της περιόδου ο εξυπηρετητής δεν είναι διαθέσιμος. Με το πέρας αυτής, ο εξυπηρετητής ενεργοποιείται και αν υπάρχει εργασία προς εκτέλεση, θα αρχίσει άμεσα να εξυπηρετεί. Σε αντίθετη περίπτωση, παραμένει άεργος όπως περιγράφηκε παραπάνω.

Αν μια εργασία δεν βρει διαθέσιμο εξυπηρετητή κατά την άφιξή της, τοποθετείται στην ουρά (αν υπάρχει διαθέσιμη θέση) και παραμένει στο σύστημα για χρονικό διάστημα που ακολουθεί εκθετική κατανομή παραμέτρου ζ_i (deadline). Αν δεν εξυπηρετηθεί στην διάρκεια του παραπάνω χρονικού διαστήματος αποχωρεί χωρίς να εξυπηρετηθεί.

Έστω $N(t)$ ο αριθμός των πελατών στο σύστημα, $V(t)$ ο αριθμός των εξυπηρετητών σε διακοπές και $J(t)$ η φάση λειτουργίας των εξυπηρετητών. Η διαδικασία $Z(t) = \{N(t), V(t), J(t)\}$ αποτελεί μια τρι-διάστατη Μαρκοβιανή διαδικασία σε συνεχή χρόνο με χώρο καταστάσεων $E = \{(i, j, k), i \geq 0, j = 0, 1, \dots, c, k = 1, \dots, R\}$. Γράφοντας τις καταστάσεις σε λεξικογραφική διάταξη, δηλαδή,

$$\begin{aligned} & (0, 0, 1), \dots, (0, 0, R), \dots, (0, c, 1), \dots, (0, c, R), \\ & (1, 0, 1), \dots, (1, 0, R), \dots, (1, c, 1), \dots, (1, c, R), \\ & \dots \\ & (L, 0, 1), \dots, (L, 0, R), \dots, (L, c, 1), \dots, (L, c, R), \end{aligned}$$

η Μαρκοβιανή διαδικασία $Z(t)$ είναι μια Quasi-Birth-Death process (QBD) (Latouche and Ramaswami (1999)).

3. Ανάλυση

Ο απειροστός γεννήτορας Q της $Z(t)$ είναι ένας τετραγωνικός πίνακας διάστασης $[(L+1) \times (c+1) \times R]$ και δίνεται από

$$Q = \begin{pmatrix} B_{00} & A & & & \\ B_1 & C_1 & A & & \\ & B_2 & C_2 & A & \\ & & \ddots & \ddots & \ddots \\ & & & B_{L-1} & C_{L-1} & A \\ & & & & B_L & C_L \end{pmatrix}.$$

Ακολουθώντας, περιγράφουμε τους υποπίνακες του Q : Ο πίνακας A περιέχει τους ρυθμούς μεταβάσεων της διαδικασίας που αντιστοιχούν σε άφιξη ενός πελάτη, η διάσταση του είναι $[(c+1) \times R] \times [(c+1) \times R]$ και δίνεται από

$$A = I_{(c+1) \times (c+1)} \otimes \text{diag}_{R \times R}(\lambda_1, \lambda_2, \dots, \lambda_R),$$

όπου $\otimes$ το γινόμενο του Kronecker, και $I_{(c+1) \times (c+1)}$ ο μοναδιαίος $(c+1) \times (c+1)$ πίνακας.

Ο πίνακας B_i , $i = 1, \dots, L$ περιέχει τους ρυθμούς μεταβάσεων της διαδικασίας που αντιστοιχούν σε αναχώρηση ενός πελάτη όταν υπάρχουν i πελάτες στο σύστημα και έχει διάσταση $[(c+1) \times R] \times [(c+1) \times R]$. Η μορφή του πίνακα B_i είναι:

$$B_i = \begin{pmatrix} \hat{B}_{i0} & & & & \\ & \hat{B}_{i1} & & & \\ & & \hat{B}_{i2} & & \\ & & & \ddots & \\ & & & & \hat{B}_{ic} \end{pmatrix},$$

όπου,

$$\hat{B}_{ij} = \begin{cases} (c-j)M + (i-c+j)Z & , c-j \leq i \\ iM & , c-j > i \end{cases},$$

και $M = \text{diag}_{R \times R}(\mu_1, \dots, \mu_R)$, $Z = \text{diag}_{R \times R}(\zeta_1, \dots, \zeta_R)$.

Ο πίνακας B_{00} περιέχει τους ρυθμούς μεταβάσεων της διαδικασίας που αντιστοιχούν σε αποχωρήσεις και επιστροφές εξυπηρετητών σε διακοπές καθώς και τους ρυθμούς αλλαγής φάσεων όταν στο σύστημα δεν υπάρχουν πελάτες. Η διάσταση του B_{00} είναι $[(c+1) \times R] \times [(c+1) \times R]$ και δίνεται από

$$B_{00} = \begin{pmatrix} H_0 & \Theta_0 & & & \\ \Gamma_1 & H_1 & \Theta_1 & & \\ & \Gamma_2 & H_2 & \Theta_2 & \\ & & \ddots & \ddots & \ddots \\ & & & \Gamma_{c-1} & H_{c-1} & \Theta_{c-1} \\ & & & & \Gamma_c & H_c \end{pmatrix},$$

όπου $\Theta_j = (c - j)\Theta$, $j = 0, \dots, c - 1$ με $\Theta = \text{diag}_{R \times R}(\theta_1, \theta_2, \dots, \theta_R)$, $\Gamma_j = j\Gamma$ με $\Gamma = \text{diag}_{R \times R}(\gamma_1, \gamma_2, \dots, \gamma_R)$, $j = 1, \dots, c$ και

$$H_j = \begin{pmatrix} -(\lambda_1 + \eta_1 + (c - j)\theta_1 + j\gamma_1) & \dots & \eta_1 \\ & \ddots & \\ \eta_R & \dots & -(\lambda_R + \eta_R + (c - j)\theta_R + j\gamma_R) \end{pmatrix}.$$

Κάθε πίνακας C_i περιέχει τους ρυθμούς μεταβάσεων της διαδικασίας που αντιστοιχούν σε αποχωρήσεις και αφίξεις εξυπηρετητών σε διακοπές καθώς και τους ρυθμούς αλλαγής φάσεων όταν στο σύστημα υπάρχουν i πελάτες. Η διάσταση κάθε πίνακα C_i είναι $[(c + 1) \times R] \times [(c + 1) \times R]$. Η μορφή του πίνακα C_i είναι:

$$C_i = \begin{pmatrix} T_0 & \Theta_0 & & & \\ \Gamma_1 & T_1 & \Theta_1 & & \\ & \Gamma_2 & T_2 & \Theta_2 & \\ & & \ddots & \ddots & \ddots \\ & & & \Gamma_{c-1} & T_{c-1} & \Theta_{c-1} \\ & & & & \Gamma_c & T_c \end{pmatrix},$$

όπου ο πίνακας T_j , $j = 0, 1, \dots, c$ είναι της ίδιας μορφής με τον H_j με την διαφορά ότι τα στοιχεία της κυρίας διαγωνίου είναι $-(\lambda_k + \eta_k + (c - j)\theta_k + j\gamma_k + K_k)$, $k = 1, \dots, R$ και

$$K_k = \begin{cases} (c - j)\mu_k + (i - c + j)\zeta_k & , c - j \leq i \\ i\mu_k & , c - j > i \end{cases}.$$

3.1 Η Γενική Λύση

Ο υπολογισμός του στάσιμου διανύσματος πιθανοτήτων $\tilde{\pi} = (\pi_0, \pi_1, \dots, \pi_L)$ με $\pi_i = (\underline{\pi}_{i0}, \dots, \underline{\pi}_{ic})$ και $\underline{\pi}_{ij} = (\pi_{ij1}, \dots, \pi_{ijR})$ προκύπτει από την επίλυση του

$$\tilde{\pi}Q = \tilde{0}, \quad \tilde{\pi}e = 1, \quad (1)$$

όπου e ένα $(L + 1)(c + 1)R \times 1$ διάνυσμα στήλη με όλα τα στοιχεία του 1. Αν αναπτύξουμε την (1) θα προκύψουν οι εξισώσεις:

$$\pi_0 B_{00} + \pi_1 B_1 = \tilde{0} \quad (2)$$

$$\pi_0 A + \pi_1 C_1 + \pi_2 B_2 = \tilde{0} \quad (3)$$

$$\pi_1 A + \pi_2 C_2 + \pi_3 B_3 = \tilde{0} \quad (4)$$

$$\vdots$$

$$\pi_{L-2} A + \pi_{L-1} C_{L-1} + \pi_L B_L = \tilde{0} \quad (5)$$

$$\pi_{L-1} A + \pi_L C_L = \tilde{0} \quad (6)$$

Αξίζει να σημειωθεί ότι υπάρχουν ειδικές πινακοαναλυτικές τεχνικές επίλυσης της (1) (βλέπε Latouche and Ramaswami (1999)). Όμως αυτές περιλαμβάνουν την αντιστροφή πινάκων, που αυξάνει σημαντικά το υπολογιστικό κόστος. Στην παρούσα εργασία αντιμετωπίζουμε το εν λόγω πρόβλημα, επιλύοντας με ελαχιστο υπολογιστικό κόστος εκμεταλευόμενοι το γεγονός ότι ο πίνακας A είναι διαγώνιος (με μή μηδενικά τα στοιχεία της κυρίας διαγωνίου). Επομένως, Επιλύουμε την τελευταία εξίσωση (6) ως προς το π_{L-1} και θα την εκφράσουμε ως μια ποσότητα της π_L . Αντικαθιστώντας στην (5) θα εκφράσουμε το π_{L-2} και αυτή ως μια ποσότητα της π_L . Συνεχίζοντας, θα εκφράσουμε τα π_i , $i = 0, \dots, L$ ως συναρτήσεις του π_L . Συγκεκριμένα θα έχουμε

$$\pi_{L-1} = -\pi_L C_L A^{-1} = \pi_L R_{L-1} \quad (7)$$

Στο επόμενο βήμα με την βοήθεια της (7) η εξίσωση (5) θα γίνει:

$$\pi_{L-2} A + \pi_{L-1} C_{L-1} + \pi_L B_L = \tilde{0} \Rightarrow \quad (8)$$

$$\pi_{L-2} = -\pi_L (R_{L-1} C_{L-1} + B_L) A^{-1} = \pi_L R_{L-2}$$

Ομοίως,

$$\begin{aligned} \pi_{L-3} A + \pi_{L-2} C_{L-2} + \pi_{L-1} B_{L-1} &= \tilde{0} \Rightarrow \\ \pi_{L-3} A + \pi_L R_{L-2} C_{L-2} + \pi_{L-1} B_{L-1} &= \tilde{0} \Rightarrow \\ \pi_{L-3} &= -\pi_L (R_{L-2} C_{L-2} + R_{L-1} B_{L-1}) A^{-1} = \pi_L R_{L-3} \end{aligned} \quad (9)$$

Συνεχίζοντας με τον ίδιο τρόπο, κάθε π_i , $i = 0, 1, \dots, L$ γράφεται την μορφή:

$$\pi_i = \pi_L R_i, \quad i = 0, \dots, L, \quad (10)$$

όπου,

$$R_L = I$$

$$R_{L-1} = -C_L A^{-1}$$

$$R_i = -(R_{i+1} C_{i+1} + R_{i+2} B_{i+2}), \quad i = L-2, L-3, \dots, 0$$

Τελός το π_L υπολογίζεται χρησιμοποιώντας την (2) και την συνθήκη κανονικοποίησης:

$$\left. \begin{aligned} \pi_0 B_{00} + \pi_1 B_0 &= \tilde{0} \\ \sum_{i=0}^L \pi_i e &= 1 \end{aligned} \right\} \Rightarrow \left\{ \begin{aligned} \pi_L (R_0 B_{00} + R_1 B_0) &= \tilde{0} \\ \pi_L \sum_{i=0}^L R_i e &= 1 \end{aligned} \right.$$

4. Μέτρα απόδοσης και διαχείριση ενέργειας

Τα βασικά μέτρα απόδοσης του συστήματος αποτελούν ο αναμενόμενος αριθμός πελατών στο σύστημα, $E(N)$, ο αναμενόμενος αριθμός εξυπηρετητών σε κατάσταση ύπνου, $E(V)$ και η αναμενόμενη κατανάλωση ενέργειας AEC (Average Energy Consumption), τα οποία και δίνονται από, η οποία υπολογίζεται από τον παρακάτω τύπο:

$$\begin{aligned} E(N) &= \sum_{i=0}^L i \sum_{j=0}^c \sum_{k=1}^R \pi_{ijk}, \\ E(V) &= \sum_{j=0}^c j \sum_{i=0}^L \sum_{k=1}^R \pi_{ijk}, \\ AEC &= \sum_{l=0}^c \bar{P}_l \bar{W}^{(l)}, \end{aligned} \quad (11)$$

όπου,

$$\bar{P}_l = (P_{l1}, P_{l2}, \dots, P_{lR}), \bar{W}^{(l)} = (W_1^{(l)}, W_2^{(l)}, \dots, W_R^{(l)})^T$$

με $\bar{W}^{(l)}$ το $R \times 1$ διάνυσμα στήλη με στοιχεία $W_k^{(l)}$ να είναι η κατανάλωση ενέργειας όταν έχω l εξυπηρετητές σε λειτουργία (busy or idle) και το σύστημα βρίσκεται στην k φάση, $k = 1, \dots, R$ και $\bar{P}^{(l)}$ η στάσιμη κατανομή του αριθμού των εξυπηρετητών σε λειτουργία, όπου

$$P_{lk} = \pi_{\bullet \ c-l \ k} = \sum_{i=0}^L \pi_{i \ c-l \ k}.$$

5. Αριθμητικά αποτελέσματα

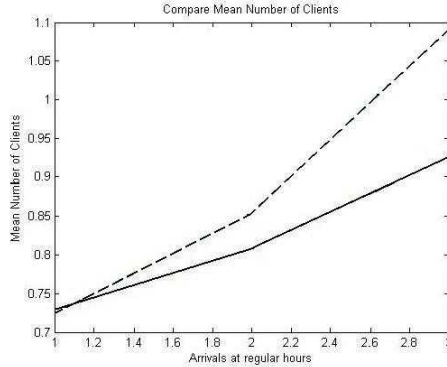
Στη συνέχεια παραθέτουμε ένα αριθμητικό παράδειγμα για να ελέγξουμε την εγκυρότητα των θεωρητικών αποτελεσμάτων. Έστω, $L = 5$, $c = 3$, ενώ το σύστημα εναλλάσσεται μεταξύ δύο φάσεων λειτουργίας: 1, όταν βρίσκεται σε τυπική φάσης κανονικού ρυθμού και 0 όταν βρίσκεται σε ώρα αιχμής. Συγκεκριμένα θα μελετήσουμε τον τρόπο με τον οποίο επηρεάζεται η μέση κατανάλωση ενέργειας και ο αναμενόμενος αριθμός εργασιών στο σύστημα ως προς λ_0 , θ_0 . Στον ακόλουθο πίνακα δίνονται τιμές για ορισμένα βασικά χαρακτηριστικά:

	0	1
λ_i	—	3
μ_i	2	5
η_i	1	2
γ_i	2	1
θ_i	—	4
ζ_i	1	5

Για την υλοποίηση του αριθμητικού παραδείγματος, θα υποθέσουμε ότι ο μέγιστος χρόνος που παραμένει διαθέσιμος ένας εξυπηρετητής μετά από μια εξυπηρέτηση θα

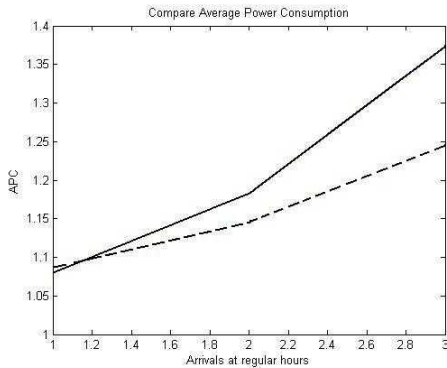
είναι μεγαλύτερος κατά την διάρκεια της περιόδου αιχμής (δηλ. $\theta_1 > \theta_0$) θα παραμένει για περισσότερο χρόνο άεργος πριν φύγει σε διακοπές, ώστε να είναι άμεσα διαθέσιμος, μιας και σε εκείνη την περίοδο είναι περισσότερο πιθανό να φτάσει κάποιος πελάτης. Επιπλέον, εξαιτίας αυτής της αυξημένης ροής πελατών στην διάρκεια της περιόδου αιχμής, είναι αναμενόμενο η διάρκεια της περιόδου διακοπών (κατάσταση ύπνου), να είναι μικρότερη σε σχέση με αυτή της τυπικής περιόδου, ώστε ο εξυπηρετητής να επιστρέψει γρηγορότερα και να παρέχει ένα επαρκές επίπεδο παροχής εξυπηρέτησης (δηλ. $\gamma_1 < \gamma_0$).

Σχήμα 1: Αναμενόμενος αριθμός πελατών στο σύστημα για κυμαινόμενες τιμές του λ_0 . Gray line: $\theta_0 = 1$, Dashed line: $\theta_0 = 3$.



Παρατηρούμε στην εικόνα 1 ότι για αυξανόμενες τιμές του ρυθμού προσέλευσης πελατών (λ_0), ο αναμενόμενος αριθμός πελατών στο σύστημα αυξάνει, γεγονός απόλυτα φυσιολογικό. Όμως η αύξηση αυτή είναι περισσότερο αισθητή καθώς ελλοτώνεται ο μέσος μέγιστος χρόνος αεργίας ($1/\theta_0$) ενός *server* στο σύστημα.

Σχήμα 2: Μέση κατανάλωση ενέργειας για κυμαινόμενες τιμές του λ_0 . Dashed line: $\theta_0 = 3$, Black line: $\theta_0 = 1$.



Στην εικόνα 2 παρατηρούμε την επίδραση του λ_0 για διαφορετικές τιμές του μέσου χρόνου αεργίας ενός *server* προτού αναχωρήσει για διακοπές. Προφανώς η αύξηση

του λ_0 θα προκαλέσει την αύξηση στην μέση κατανάλωση ενέργειας. Η αύξηση αυτή είναι λιγότερο αισθητή καθώς μειώνουμε τον μέσο μέγιστο χρόνο αεργίας μιας και σε αυτή την περίπτωση, ο server θα αναχωρήσει γρηγορότερα σε διακοπές και επομένως θα εξοικονομηθεί ενέργεια.

6. Μελλοντική Έρευνα

Μελλοντικά θα ήταν αναγκαία η μελέτη ενός γενικότερου μοντέλου που θα συμπεριλάβει περισσότερα ρεαλιστικά χαρακτηριστικά καθώς και η ανάπτυξη προβλημάτων βελτιστοποίησης. Σε αυτά απότερος στόχος θα είναι η εύρεση τιμών των βασικών παραμέτρων του συστήματος που θα ελαχιστοποιούν το συνολικό κόστος του παρόχου, το οποίο προέρχεται τόσο από την παραβίαση του συμβολαίου SLA, λόγω του αυξημένου χρόνου αναμονής των πελατών, όσο και από το ύψος της κατανάλωσης ενέργειας. Προφανώς, η βελτιστοποίηση θα γίνει κάτω από δεδομένους περιορισμούς της λειτουργίας του συστήματος. Επιπλέον, ενδιαφέρον θα παρουσίαζε και η εισαγωγή ρεαλιστικότερων διαδικασιών αφίξεων-εξυπηρέτησεων (Markovian Arrival Process, PH-type distribution) καθώς και η χρήση εναλλακτικών πηγών ενέργειας.

ABSTRACT

This work deals with the problem of managing a data center in a way to reduce its power consumption without significantly affecting the offered quality of service. A major source of power consumption is caused by servers that are powered up when they are not in use. We propose a finite capacity multiserver queue with impatient customers and asynchronous vacations, operating in a Markovian random environment. The system is studied by a three-dimensional Markov process that belongs to the class of QBD processes. We obtain the stationary distribution and its basic performance metrics. Useful numerical results are also obtained and show insights into the system behavior.

REFERENCES

- Green Grid. (2010) *Unused Servers Survey Results Analysis*, www.thegreengrid.org
- E. Altman, U. Yechiali. (2006) *Analysis of customers' impatience in queues with server vacations*, *Queueing Systems* 52(4) 261-279.
- A. Greenberg, J. Hamilton, D. A. Maltz, P. Patel. (2009) *The Cost of a Cloud: Research Problems in Data Center Networks*, *SIGCOMM Comp. & Comm. Rev.*, 39(1) 68-73.
- G. Latouche, V. Ramaswami. (1999) *Introduction to Matrix Analytic Methods in Stochastic Modeling*, *ASA-SIAM Series on Statistics and Applied Mathematics*, Philadelphia, PA.

- I. Mitrani. (2011) *Service center trade-offs between customer impatience and power consumption*, *Performance Evaluation* 68(11) 1222-1231.
- I. Mitrani. (2013) *Managing performance and power consumption in a server farm*, *Annals of Operations Research* 202(1) 121-134.
- T. Phung-Duc. (2015a) *Single-Server Systems with Power-Saving Modes*, in: *Gribaudo, M., Manini, D., Remke, A. (Eds.) ASMTA 2015, LNCS 9081*, pp. 158-172.
- T. Phung-Duc. (2015b) *Controllable Setup Queue for Energy-Aware Server*, in: *Proc. UKPEW 2015*, pp. 96-109.
- H. Takagi. (1991) *Queueing Analysis: A Foundation of Performance Evaluation. Volume 1: Vacation and Priority Systems, Part 1*, North-Holland, Amsterdam.
- N. Tian, Z.-G. Zhang. (2006) *Vacation Queueing Models - Theory and Applications*, Springer, New York.

Ένα σύστημα αναμονής με ανυπόμονους πελάτες, σήματα και ομαδικές αναχωρήσεις σε ετερογενές περιβάλλον

I. Δημητρίου, Μ. Μάρκου

Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών
idimit@math.upatras.gr, patramariajj@gmail.com

ΠΕΡΙΛΗΨΗ

Θεωρούμε ένα σύστημα εξυπηρέτησης με έναν υπάλληλο, η λειτουργία του οποίου εναλλάσσεται μεταξύ R διαφορετικών φάσεων (random environment). Στην i φάση, $i = 1, \dots, R$, η διάρκεια της οποίας ακολουθεί την εκθετική κατανομή παραμέτρου η_i , οι πελάτες καταφθάνουν σύμφωνα με την κατανομή Poisson παραμέτρου λ_i και τοποθετούνται σε ένα πεπερασμένο χώρο αναμονής N . Ο χρόνος της παρεχόμενης εξυπηρέτησης ακολουθεί την εκθετική κατανομή παραμέτρου μ_i . Κάθε πελάτης που εισέρχεται στο σύστημα είναι ανυπόμονος και αν δεν εξυπηρετηθεί εντός χρονικού διαστήματος που ακολουθεί την εκθετική κατανομή παραμέτρου θ_i αναχωρεί από το σύστημα χωρίς να εξυπηρετηθεί. Επιπλέον, στο σύστημα αναπτύσσεται μια δευτερεύουσα ροή «αρνητικών» πελατών σύμφωνα με την κατανομή Poisson παραμέτρου r_i , η οποία προκαλεί την άμεση αναχώρηση μιας ομάδας «τυπικών» πελατών τυχαίου μεγέθους (random batch removal). Το σύστημα περιγράφεται με μια διδιάστατη Μαρκοβιανή διαδικασία και υπολογίζονται η στάσιμη κατανομή και τα κύρια μέτρα απόδοσής του. Εφαρμογές του παραπάνω συστήματος εμφανίζονται τόσο στην βιομηχανία όσο και στα σύγχρονα τηλεπικοινωνιακά συστήματα.

Λέξεις - Κλειδιά: Ανυπομονησία, Αρνητικοί πελάτες (negative customers), Ετερογένεια, $M/M/1$, Ομαδικές Αναχωρήσεις, Μέτρα Απόδοσης.

1. Εισαγωγή

Τα σύγχρονα βιομηχανικά και τηλεπικοινωνιακά συστήματα χαρακτηρίζονται από την ετερογένεια (Yechiali and Naor (1971)) και τη μη στασιμότητα των βασικών παραμέτρων που τα περιγράφουν. Επιπλέον, η παροχή εξυπηρέτησης περιλαμβάνει τον καθορισμό συγκεκριμένων όρων και προϋποθέσεων σχετικά με την ολοκλήρωσή της, ενώ οι απρόβλεπτες βλάβες πρέπει να ληφθούν υπόψη για τον αποτελεσματικό σχεδιασμό αυτών των συστημάτων.

Στην παρούσα εργασία προτείνεται ένα μοντέλο με ανυπόμονους πελάτες και ομαδικές αναχωρήσεις με εφαρμογή τόσο στη βιομηχανία όσο και στα τηλεπικοινωνιακά συστήματα. Θεωρούμε μια μηχανή εξυπηρέτησης (διακομιστής), η οποία εναλλάσσεται

μεταξύ διαδοχικών φάσεων λειτουργίας. Στην μηχανή αυτή καταφθάνουν προϊόντα (τυπικοί πελάτες) προς επεξεργασία, κατόπιν κάποιας παραγγελίας, η οποία θα πρέπει να ολοκληρωθεί σε δεδομένο χρονικό διάστημα (ανυπομονησία). Η παραβίαση αυτού του χρονικού διαστήματος προκαλεί την άμεση ακύρωση της παραγγελίας. Επιπλέον, το χαρακτηριστικό των τυχαίων βλαβών ή των μαζικών ακυρώσεων παραγγελιών (αρνητικοί πελάτες) που προκαλεί την αναχώρηση μιας ομάδας πελατών του συστήματος λαμβάνεται επίσης υπόψιν. Οι παράμετροι που χαρακτηρίζουν και καθορίζουν την λειτουργία του συστήματος δεν είναι σταθερές καθιστώντας το περιβάλλον λειτουργίας ως δυναμικό και ετερογενές (random environment). Το σύστημα περιγράφεται με μια διδιάστατη Μαρκοβιανή διαδικασία και με χρήση πινακοαναλυτικών μεθόδων (Latouche and Ramaswami (1999), Essia and Molle (2007)) υπολογίζονται η στάσιμη κατανομή και τα κύρια μέτρα απόδοσης του.

Το χαρακτηριστικό της ανυπομονησίας, που προκαλεί την αναχώρηση πελατών χωρίς εξυπηρέτηση είναι ρεαλιστικό και έχει περιληφθεί και σε άλλες αντίστοιχες μελέτες στην βιβλιογραφία (π.χ. Altman and Yechiali (2006), Perel and Yechiali (2010), Yechiali (2007)). Επιπλέον, η έννοια του αρνητικού πελάτη εισήχθει απο τον Gelenbe (1991) για να περιγράψει την παρουσία ιών σε συστήματα Η/Υ που καταστρέφουν αρχεία προς εκτέλεση. «Αρνητικοί πελάτες» καλούνται οι πελάτες οι οποίοι μειώνουν το μήκος της ουράς στην οποία καταφθάνουν, χωρίς να έχουν την απαίτηση να εξυπηρετηθούν και στην πράξη αποτελούν εντολές ελέγχου του συστήματος. Εάν το σύστημα είναι άδειο δεν επιδρούν σε αυτό.

Η εργασία δομείται ως εξής. Στην ενότητα 2. περιγράφουμε λεπτομερώς το μαθηματικό μοντέλο, ενώ στην ενότητα 3. προχωρούμε στην μελέτη του συστήματος με χρήση πινακοαναλυτικών μεθόδων. Έχοντας υπολογίσει την στάσιμη κατανομή, την χρησιμοποιούμε στην ενότητα 4. για να προσδιορίσουμε ορισμένα σημαντικά μέτρα απόδοσης όπως ο αναμενόμενος αριθμός πελατών στο σύστημα. Τέλος, στην ενότητα 5. παράθετουμε ένα αριθμητικό παράδειγμα για να παρατηρήσουμε την επίδραση των αλλαγών των τιμών των παραμέτρων στην απόδοση του συστήματος.

2. Το μοντέλο

Θεωρούμε ένα Μαρκοβιανό σύστημα εξυπηρέτησης με έναν διακομιστή ($M/M/1$) σε τυχαίο περιβάλλον και πεπερασμένο χώρο αναμονής N . Θεωρούμε οτι το τυχαίο περιβάλλον αποτελείται από i διαδοχικές καταστάσεις, με $i = 1, \dots, R$ και καθορίζουν την λειτουργία του συστήματος. Συγκεκριμένα, η λειτουργία του συστήματος εναλλάσσεται μεταξύ R διαφορετικών φάσεων με διάρκεια που ακολουθεί την εκθετική κατανομή με ρυθμό η_i .

Οι πελάτες καταφθάνουν στο σύστημα κατά τη διάρκεια της i φάσης, $i = 1, 2, \dots, R$, σύμφωνα με την κατανομή Poisson με ρυθμό λ_i και σχηματίζουν μια ενιαία γραμμή αναμονής ανάλογα με τη σειρά άφιξης τους. Η εξυπηρέτηση κάθε πελάτη στην i φάση χαρακτηρίζεται από εκθετικό χρόνο μ_i . Κατά την άφιξή του κάθε πελάτης ενεργοποιεί το δικό του χρονόμετρο ανυπομονησίας εκθετικής κατανομής παραμέτρου θ_i . Εάν παρέλθει το χρονικό αυτό διάστημα και ο πελάτης δεν έχει εξυπηρετηθεί, τότε

αναχωρεί από το σύστημα και χάνεται. Συγχρόνως στο σύστημα δημιουργείται μια ροή «αρνητικών» πελατών (σήματα, «negative customers») σύμφωνα με την κατανομή Poisson παραμέτρου r_i , η οποία προκαλεί την άμεση αναχώρηση k από τους n «τυπικών πελατών» με πιθανότητα $p_k^{(n,i)}$.

Ορίζουμε $N(t)$ τον αριθμό πελατών στο σύστημα τη χρονική στιγμή t και $J(t)$ την φάση λειτουργίας του διακομιστή τη στιγμή αυτή. Η διαδικασία $\{N(t), J(t)\}$ είναι μια συνεχής διδιάστατη Μαρκοβιανή διαδικασία με χώρο καταστάσεων $X = \{(k, i), 1 \leq k \leq N, 1 \leq i \leq R\}$.

Για να απλοποιήσουμε την όλη διαδικασία, υποθέτουμε ότι το σύστημα μεταβαίνει από τη φάση i στην φάση $i + 1$ με ρυθμό η_i και μόνο από την φάση R μπορεί να μεταβεί στην φάση 1 με ρυθμό η_R . Ως εκ τούτου, ο γεννήτορας πίνακας $\bar{D}$ των ρυθμών εναλλαγής λειτουργίας του συστήματος είναι ένας πίνακας διάστασης $R \times R$ της μορφής:

$$\bar{D} = \begin{bmatrix} -\eta_1 & \eta_1 & 0 & \cdots & 0 \\ 0 & -\eta_2 & \eta_2 & \cdots & 0 \\ 0 & 0 & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & -\eta_{R-1} & \eta_{R-1} \\ \eta_R & 0 & \cdots & 0 & -\eta_R \end{bmatrix}$$

Ο γεννήτορας πίνακας Q της διαδικασίας είναι ένας πίνακας διάστασης $(1 + N) \times R$, συντιθέμενος από τις παραμέτρους τους συστήματος. Συγκεκριμένα, ορίζεται ως:

$$Q = \begin{bmatrix} A_{00} & B_0 & 0 & \cdots & 0 \\ A_{10} & A_{11} & B_0 & \cdots & 0 \\ A_{20} & A_{21} & A_{22} & \ddots & 0 \\ \vdots & \vdots & \vdots & \ddots & B_0 \\ A_{N0} & A_{N1} & \cdots & \cdots & A_{NN} \end{bmatrix}.$$

Τα στοιχεία του πίνακα Q είναι υποπίνακες διάστασης $R \times R$. Η κύρια διαγώνιος του αποτελείται από τους υποπίνακες A_{00}, A_{jj} ($j = 0, 1, \dots, N - 1$), και A_{NN} που εκφράζουν την παραμονή k πελατών στο σύστημα, ορίζονται αντίστοιχα ως :

$$A_{00} = \left[\begin{array}{c|cccc} & (0, 1) & (0, 2) & \cdots & (0, R) \\ \hline (0, 1) & -(\lambda_1 + \eta_1) & \eta_1 & \cdots & 0 \\ (0, 2) & 0 & -(\lambda_2 + \eta_2) & \ddots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (0, R) & \eta_R & 0 & \cdots & -(\lambda_R + \eta_R) \end{array} \right]$$

$$A_{jj} = \begin{bmatrix} a_{11} & \eta_1 & 0 & \cdots & 0 \\ 0 & a_{22} & \eta_2 & \ddots & 0 \\ 0 & 0 & a_{33} & \ddots & 0 \\ \vdots & \vdots & \vdots & \ddots & \eta_{R-1} \\ \eta_R & 0 & 0 & \cdots & a_{RR} \end{bmatrix},$$

όπου $a_{jj} = -(\lambda_i + \eta_i + \mu_i + k\theta_i + r_i)$, $j = 0, 1, \dots, n-1$

$$A_{NN} = \begin{bmatrix} c_{11} & \eta_1 & 0 & \cdots & 0 \\ 0 & c_{22} & \eta_2 & \cdots & 0 \\ 0 & 0 & c_{33} & \ddots & 0 \\ \vdots & \vdots & \vdots & \ddots & \eta_{R-1} \\ \eta_R & 0 & 0 & 0 & c_{RR} \end{bmatrix},$$

όπου $c_{ii} = -(\mu_i + \eta_i + r_i + n\theta_i)$, $i = 1, 2, \dots, R$.

Επιπλέον, οι υποπίνακες $A_{j+1,j}$, $j = 0, 1, \dots, n-1$ και B_0 περιγράφουν τους ρυθμούς μεταβάσεων της διαδικασίας που αντιστοιχούν σε αναχώρηση και άφιξη ενός πελάτη αντίστοιχα και έχουν την ακόλουθη μορφή :

$$A_{j+1,j} = \text{diag}(\mu_1 + r_1 p_1^{(N,1)} + (j+1)\theta_1, \dots, \mu_R + r_R p_1^{(N,R)} + (j+1)\theta_R), j = 0, 1, \dots, N-1$$

και

$$B_0 = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \cdots & \lambda_R \end{bmatrix}$$

Εκτός από τους παραπάνω πίνακες ορίζουμε τους A_{j0} , ($j \geq 1$) και A_{jk} , ($k = 0, 1, \dots, j-2$) που εκφράζουν τους ρυθμούς μετάβασης που αντιστοιχούν σε αναχώρηση όλων των πελατών. Οι πίνακες αυτοί έχουν την μορφή:

$$A_{j0} = \text{diag}((\mu_1 + \theta_1)\delta_{(j=1)} + r_1 p_n^{(n,1)}, \dots, (\mu_R + \theta_R)\delta_{(j=1)} + r_R p_n^{(n,R)}), j \geq 1,$$

όπου δ_j η δείκτρια συνάρτηση Kronecker's delta:

$$\delta_j = \begin{cases} 1 & , j = 1 \\ 0 & , j \neq 1 \end{cases}$$

και

$$A_{jk} = \text{diag}(r_1 p_{j-k}^{(j,1)}, \dots, r_R p_{j-k}^{(j,R)}), k = 0, 1, \dots, j-2$$

αντίστοιχα.

3. Ανάλυση

Το διάνυσμα στάσιμων πιθανοτήτων $\underline{\pi}$ που αντιστοιχεί στον πίνακα Q κατανομείται ως $\underline{\pi} = [\pi_0, \pi_1, \dots, \pi_N]$, όπου τα υποδιανύσματα του $\underline{\pi}$ ορίζονται με τη σειρά τους ως :

$$\underline{\pi}_j = (\pi_{j1}, \pi_{j2}, \dots, \pi_{jR})$$

Σύμφωνα με τη γεωμετρική μέθοδο επίλυσης πινάκων από την σχέση $\underline{\pi}Q = \mathbf{0}$ προκύπτουν οι εξισώσεις ισορροπίας :

$$\pi_0 A_{00} + \pi_1 A_{10} + \dots + \pi_N A_{N0} = 0 \quad (1)$$

$$\pi_0 B_0 + \pi_1 A_{11} + \dots + \pi_N A_{N1} = 0 \quad (2)$$

$$\pi_1 B_0 + \pi_2 A_{22} + \dots + \pi_N A_{N2} = 0 \quad (3)$$

⋮

$$\pi_{N-2} B_0 + \pi_{N-1} A_{N-1,N-1} + \pi_N A_{N,N-1} = 0 \quad (4)$$

$$\pi_{N-1} B_0 + \pi_N A_{N,N} = 0 \quad (5)$$

Ο υπολογισμός των υποδιανυσμάτων του διανύσματος $\underline{\pi}$ επιτυγχάνεται με την επίλυση του συστήματος

$$\begin{cases} \underline{\pi}Q = \mathbf{0} \\ \underline{\pi}\mathbf{1} = 1 \end{cases},$$

όπου $\underline{\pi}\mathbf{1}^T = 1$ είναι η συνθήκη κανονικότητας.

Ο πίνακας B_0 είναι ένας διαγώνιος πίνακας με μη μηδενικά διαγώνια στοιχεία και επομένως είναι αντιστρέψιμος πίνακας. Εφόσον η επίλυση του συστήματος στηρίζεται σε πράξεις πινάκων, μια πολύπλοκη και με υψηλό υπολογιστικό κόστος διαδικασία, χρησιμοποιούμε τον B_0^{-1} για την γενική επίλυση. Έτσι εξασφαλίζουμε απλές πράξεις κατά την επίλυση και χαμηλότερο υπολογιστικό κόστος.

Αρχικά επιλύουμε την (6) ως προς π_{N-1} και προκύπτει:

$$\pi_{N-1} = -\pi_N A_{NN} B_0^{-1} = \pi_N (-A_{NN} B_0^{-1}) = \pi_N R_{N-1}, \quad (6)$$

όπου $R_{N-1} = (-A_{NN} B_0^{-1})$. Επαναλαμβάνοντας τη διαδικασία για $k = N - 2$ έχουμε:

$$\begin{aligned} \pi_{N-2} B_0 + \pi_{N-1} A_{N-1,N-1} + \pi_N A_{N,N-1} &= 0 &\Rightarrow \\ \pi_{N-2} B_0 + \pi_N R_{N-1} A_{N-1,N-1} + \pi_N A_{N,N-1} &= 0 &\Rightarrow \\ \pi_{N-2} B_0 &= \pi_N [-(R_{N-1} A_{N-1,N-1} + A_{N,N-1})] &\Rightarrow \\ \pi_{N-2} &= \pi_N [-(R_{N-1} A_{N-1,N-1} + A_{N,N-1})] B_0^{-1} &\Rightarrow \end{aligned}$$

Συνεπώς η σχέση που προκύπτει για την π_{N-2} , η οποία εκφράζει την πιθανότητα να υπάρχουν $N - 2$ πελάτες στο σύστημα, είναι :

$$\pi_{N-2} = \pi_N R_{N-2}. \quad (7)$$

Συνεχίζοντας επαγωγικά για $j = 0, 1, \dots, n$ καταλήγουμε στο συμπέρασμα πως η σχέση που δίνει το διάνυσμα πιθανοτήτων $\underline{\pi}_j$ είναι η

$$\underline{\pi}_j = \underline{\pi}_N R_j \quad (8)$$

για $j = 1, \dots, k-1$, όπου,

$$R_j = \begin{cases} I & , j = N \\ -A_{NN}B_0^{-1} & , j = N-1 \\ -\sum_{k=j+1}^N R_k A_{k,j+1} B_0^{-1} & , j = 0, 1, \dots, N-2 \end{cases}$$

Η σχέση (2) λαμβάνοντας υπόψη την (9) μπορεί να γραφεί εναλλακτικά ως:

$$\underline{\pi}_N \sum_{j=0}^N R_j A_{j0} = 0. \quad (9)$$

Με παρόμοιο τρόπο η συνθήκη κανονικοποίησης $\underline{\pi} \underline{1}^T = 1$ αντικαθιστώντας την (9) σε αυτή, αποκτά την μορφή

$$\underline{\pi}_N \sum_{j=0}^N R_j \underline{1} = 1. \quad (10)$$

Συνεπώς ο υπολογισμός των υποδιανυσμάτων, δεδομένων των πινάκων R_j και A_{j0} , επιτυγχάνεται εναλλακτικά με την επίλυση του συστήματος:

$$\begin{cases} \underline{\pi}_N = \sum_{j=0}^N R_j A_{j0} = 0 \\ \underline{\pi}_N \sum_{j=0}^N R_j \underline{1} = 1 \end{cases}$$

4. Μέτρα απόδοσης

Ο υπολογισμός των μέτρων απόδοσης είναι απαραίτητος για την εκτίμηση της συμπεριφοράς του προς μελέτη συστήματος. Για τον λόγο αυτό καθορίζεται η σχέση που δίνει τον μέσο αριθμό πελατών στο σύστημα, και αυτών που εξυπηρετούνται αλλά και όσων βρίσκονται στην ουρά αναμονής, ως :

$$E(N) = \sum_{k=0}^N k \sum_{i=1}^R \pi_{ki}. \quad (11)$$

5. Αριθμητικά αποτελέσματα - Συμπεράσματα

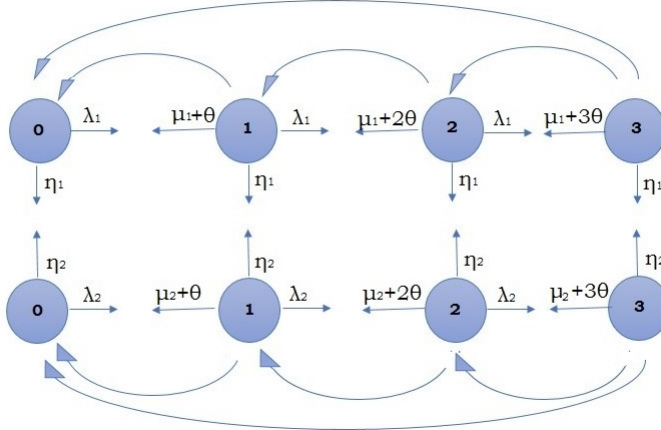
Στην ενότητα αυτή παρουσιάζονται κάποια αριθμητικά αποτελέσματα με στόχο την μελέτη του μέσου αριθμού πελατών στο σύστημα συναρτήσει του ρυθμού αφίξεων μιας φάσης λ_1 .

Θεωρούμε ένα σύστημα εξυπηρέτησης με πεπερασμένο χώρο αναμονής χωρητικότητας $N = 3$. Επιπλέον υποθέτουμε πως ο διακομιστής του συστήματος εναλλάσσεται μεταξύ 2 διαδοχικών φάσεων λειτουργίας (υψηλής και χαμηλής ζήτησης). Οι παράμετροι για τις δύο αυτές φάσεις καθορίζονται να είναι οι ακόλουθες:

Παράμετροι	Φάση 1	Φάση 2
$\bar{\lambda} = (\lambda_1, \lambda_2)$	λ_1	$\lambda_2 = 3$
$\bar{\mu} = (\mu_1, \mu_2)$	$\mu_1 = 2$	$\mu_2 = 5$
$\bar{\theta} = (\theta_1, \theta_2)$	$\theta_1 = 1$	$\theta_2 = 0.5$
$\bar{r} = (r_1, r_2)$	r_1	$r_2 = 4$
$\bar{\eta} = (\eta_1, \eta_2)$	$\eta_1 = 1$	$\eta_2 = 2$

Παρατηρούμε πως η φάση 1 χαρακτηρίζεται από μεγαλύτερη ανυπομονησία πελατών σε σχέση με την φάση 2, αλλά από μικρότερο ρυθμό εξυπηρέτησης, εναλλαγής λειτουργίας του διακομιστή αλλά και δημιουργίας ροής αρνητικών πελατών. Επίσης σημειώνουμε πως ο ρυθμός αφίξεων της φάσης 1 δεν είναι καθορισμένος, διότι μελετάμε την μεταβολή του μέσου αριθμού πελατών στο σύστημα για τις διάφορες τιμές αυτού. Το γράφημα 1 παρουσιάζει τις δυνατές μεταβάσεις μεταξύ των καταστάσεων και των φάσεων για την καλύτερη κατανόηση της λειτουργίας του συστήματος.

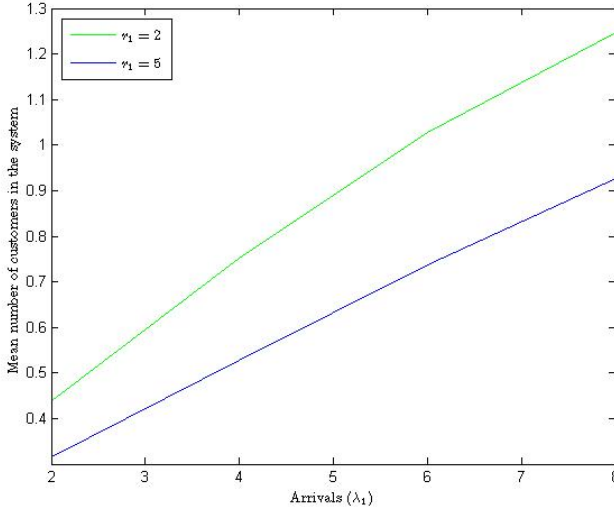
Σχήμα 1: Διάγραμμα μεταβάσεων



Φαίνεται πως η μετάβαση σε ένα ανώτερο επίπεδο (level) επιτυγχάνεται με την άφιξη ενός αριθμού πελατών με ρυθμό λ_i , ενώ η μετάβαση σε μια κατάσταση κατώτερου επιπέδου επιτυγχάνεται με την αναχώρηση ενός πελάτη ή μιας ομάδας πελατών λόγω της ολοκλήρωσης της εξυπηρέτησης του με ρυθμό μ_i ή εξαιτίας της ανυπομονησίας παραμέτρου θ_i . Επίσης η αναχώρηση μπορεί να οφείλεται και στην επιρροή αρνητικών πελατών με παράμετρο r_i .

Εφαρμόζοντας τον αλγόριθμο επίλυσης στο παρόν σύστημα, υπολογίστηκε η στάσιμη κατανομή και ο μέσος αριθμός πελατών στο σύστημα για τις διάφορες τιμές του ρυθμού αφίξεων της φάσης 1, λ_1 . Η μεταβολή του μέσου αριθμού πελατών συναρτήσει του ρυθμού αφίξεων για ροές αρνητικών πελατών $r_1 = 2$ και $r_1 = 5$ απεικονίζεται στο παρακάτω γράφημα.

Σχήμα 2: $E(N)$ as function of λ_1



Στο σχήμα 2 παρατηρούμε ότι αυξάνοντας τις τιμές του λ_1 , αυξάνονται οι τιμές του $E(N)$, όπως αναμενόταν. Η αύξηση αυτή είναι περισσότερο αισθητή καθώς μειώνεται ο ρυθμός των αρνητικών αφίξεων.

6. Συμπέρασμα

Στην παρούσα εργασία, μελετήθηκε ένα νέο σύστημα που διακρίνεται από ετερογένεια, ανυπομονησία και ομαδικές αναχωρήσεις που προκαλούνται από «αρνητικούς» πελάτες. Αφού αναλύθηκε το μαθηματικό μοντέλο του συστήματος, χρησιμοποιώντας την πινακοαναλυτική μέθοδο QBD δόθηκε η γενική επίλυση του συστήματος για την εύρεση της στάσιμης κατανομής και των μέτρων απόδοσης. Τέλος, δίνονται τα αριθμητικά αποτελέσματα για την επεξήγηση της μεταβολής του μέσου αριθμού πελατών στο σύστημα.

ABSTRACT

Consider a Markovian single server queueing system, the operation of which alternates between R different phases (random environment). During the i phase, $i = 1, \dots, R$, which follows the exponential distribution with parameter η_i , cus-

tomers arrive according to a Poisson arrival process with rate λ_i and if they find the server unavailable join a buffer of finite capacity N . Service times are exponentially distributed with rate μ_i . Every customer who enters the system is impatient, and upon arrival activates a timer which is exponentially distributed with rate θ_i . If this timer expires, the customer departs from the system without being served. Moreover, a secondary flow of "negative" customers (signals) is generated according to a Poisson process with rate r_i , which causes the immediate departure of a group "of typical" random sized customers (random batch removal). Specifically, when there are n typical customers in the system, the arrival of a signal will cause the departure of k typical customer with probability $p_k^{(n,i)}$, $\sum_{k=1}^n p_k^{(n,i)} = 1, i = 1, \dots, R$. The system is described in terms of a two-dimensional Markov process and the stationary distribution and the main performance measures are calculated. Applications of the system can be found both in industry and in telecommunications systems. Specifically, consider the server as a machine, where products arrive to be processed (typical customers) after an order. The processing of each product should be completed in a certain time otherwise it is not accepted (deadlines on the delivery of products, customer impatience). Furthermore, there may be massive cancellations of orders (negative clients). Obviously, the system operates in a dynamic and heterogeneous environment caused by the deficiency of raw materials, damage to the machine, etc., and therefore the basic engine parameters, such as the arrival rate orders and the rate of such service are not stable (random environment).

Keywords: Impatience, Negative Customers Heterogeneity, $M/M/1$, Group Departures, Efficiency Measures.

REFERENCES

- E.Altman, U.Yechiali (2006). Analysis of customers' impatience in queues with server vacations. *Queueing systems* **52**, 261 – 279
- Essia H. Elhafsi, Mart Molle (2007). On the Solution to QBD Processes with Finite State Space. *Stochastic Analysis and Applications* **25**, 763 – 779
- E.Gelenbe (1991). Product-Form Queueing Networks with Negative and Positive customers. *Journal of Applied Probability* **28**, 656 – 663
- G.Latouche, V.Ramaswami (1999). Introduction to Matrix Analytic Methods in Stochastic Modeling. Philadelphia, ASA-SIAM
- Nir Perel, U.Yechiali (2010). Queues with slow servers and impatient customers. *European Journal of Operational Research* **201**, 247 – 258
- U.Yechiali, P.Naor (1971). Queueing Problems with heterogeneous arrivals and service. *Operational Research* **19**, 722 – 734

ΧΡΗΣΗ ΦΙΛΤΡΟΥ KALMAN ΜΕ ΜΕΤΑΒΛΗΤΕΣ ΣΤΟ ΧΡΟΝΟ ΔΙΑΣΠΟΡΕΣ ΓΙΑ ΤΗΝ ΕΚΤΙΜΗΣΗ ΤΩΝ ΑΛΜΑΤΩΝ ΑΠΟΔΟΣΕΩΝ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΩΝ ΜΕΓΕΘΩΝ

Ο. Θεοδοσιάδου, Γ. Τσακλίδης
Τμήμα Μαθηματικών Α.Π.Θ.
{outhead, tsaklidi}@math.auth.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία γίνεται εκτίμηση των θετικών και αρνητικών αλμάτων στις ημερήσιες αποδόσεις του δείκτη Nasdaq και μετοχών του. Στην προτεινόμενη μεθοδολογία τα άλματα θεωρούνται κρυφές τυχαίες μεταβλητές και για το λόγο αυτό, υιοθετείται η χρήση στοχαστικών μοντέλων χώρου καταστάσεων σε διακριτό χρόνο, ενώ για την εκτίμησή τους χρησιμοποιείται ο επαναληπτικός αλγόριθμος του φίλτρου Kalman. Αρχικά, γίνεται χρήση ενός ομογενούς μοντέλου χώρου καταστάσεων με σταθερές διασπορές, ενώ στη συνέχεια τροποποιείται το μοντέλο υιοθετώντας στην εξίσωση των παρατηρήσεων μεταβλητή διασπορά, η οποία περιγράφεται από μια διαδικασία GARCH. Τα δυο μοντέλα συγκρίνονται μεταξύ τους με βάση την εφαρμογή τους σε πραγματικές χρονοσειρές αποδόσεων. Με δεδομένο ότι η ημερήσια απόδοση ορίζεται ως η διαφορά ανάμεσα στο θετικό και αρνητικό άλμα, κάτω από την επίδραση λευκού γκαουσιανού θορύβου, τα εκτιμώμενα άλματα που προκύπτουν από το φίλτρο Kalman, πρέπει να είναι μη αρνητικά σε κάθε περίπτωση. Για το σκοπό αυτό χρησιμοποιείται, με βάση τους περιορισμούς, η μέθοδος της αποκοπής της κατανομής των αλμάτων στη χρονοσειρά που εκτιμήθηκε με το φίλτρο Kalman.

Λέξεις κλειδιά: Θετικά-αρνητικά άλματα απόδοσης, Φίλτρο Kalman, μεταβλητή διασπορά, διαδικασία GARCH, αποκοπή κατανομής.

1. ΕΙΣΑΓΩΓΗ

Εμπειρικά έχει δείχθει ότι οι αποδόσεις χρηματοοικονομικών παραγώγων (π.χ. μετοχών) δεν εξελίσσονται συνεχώς στο χρόνο, όπως προτάσσει το μοντέλο των Black-Scholes (1973), αλλά εμφανίζουν και άλματα. Οι Cox and Ross (1975) και Merton (1976), υποστηρίζουν ότι οι θεωρίες που αναπτύχθηκαν για την τιμολόγηση δικαιωμάτων βασίζόμενες στην εξέλιξη των αποδόσεων κατά συνεχή τρόπο, πρέπει να τροποποιηθούν προκειμένου να συμπεριλάβουν την έκθεση στον κίνδυνο του άλματος. Στη συνέχεια ακολούθησε ένα πλήθος εργασιών, όπου ενισχύθηκε η άποψη

για την υιοθέτηση μοντέλων με άλματα προκειμένου να υπάρχει καλή προσαρμογή σε πραγματικά δεδομένα τιμών χρηματοοικονομικών μεγεθών (βλ. για παράδειγμα, Anderson et al., 2002; Pan, 2002; Eraker et al., 2003).

Ένα δεύτερο σημαντικό χαρακτηριστικό που έχει προκύψει από εμπειρικές μελέτες είναι ότι οι χρονοσειρές των αποδόσεων εμφανίζουν το φαινόμενο της δεσμευμένης -ως προς το χρόνο- ετεροσκεδαστικότητας. Με άλλα λόγια, η διασπορά των αποδόσεων δεν διατηρείται σταθερή στο χρόνο. Προκειμένου να ενσωματωθεί το φαινόμενο αυτό στην περιγραφή των αποδόσεων και την τιμολόγηση δικαιωμάτων, χρησιμοποιήθηκαν μοντέλα στοχαστικής μεταβλητότητας (βλ. για παράδειγμα, Hull and White, 1987; Fouque et al., 2000) και διαδικασίες GARCH (Bollerslev, 1986).

Στην εργασία Polimenis (2012), θεωρείται ότι οι αποδόσεις εκτίθενται τόσο σε μεταβλητότητα τύπου κίνησης Brown, όσο και σε άλματα, που είναι κρυφές τυχαίες μεταβλητές, δηλαδή δεν αποτελούν παρατηρήσιμες ποσότητες, και χωρίζονται σε δυο κατηγορίες: Τα θετικά άλματα (άλματα προς τα πάνω), τα οποία οφείλονται στην άφιξη θετικών ειδήσεων στην αγορά και τα αρνητικά άλματα (άλματα προς τα κάτω), τα οποία οφείλονται στην άφιξη αρνητικών ειδήσεων στην αγορά. Στην εργασία Θεοδοσιάδου κ.α. (2013) έγινε μια ανάλυση ευαισθησίας του μοντέλου Polimenis (2012), ενώ στο Θεοδοσιάδου κ.α. (2014) εξετάστηκε η προσαρμογή του σε πραγματικά δεδομένα.

Στόχος στην παρούσα εργασία είναι να εκτιμηθούν τα θετικά και αρνητικά άλματα στις ημερήσιες αποδόσεις πραγματικών χρονοσειρών. Με δεδομένο ότι στη μεθοδολογία που προτείνεται, τα θετικά και αρνητικά άλματα θεωρούνται κρυφές τυχαίες μεταβλητές, υιοθετείται η χρήση στοχαστικών μοντέλων χώρου καταστάσεων. Αρχικά, χρησιμοποιείται ένα γραμμικό μαρκοβιανό μοντέλο χώρου καταστάσεων (Θεοδοσιάδου και Τσακλίδης, 2015), όπου οι διασπορές θεωρούνται σταθερές τόσο στην εξίσωση των καταστάσεων (άλματα), όσο και στην εξίσωση των παρατηρήσεων (αποδόσεις). Ωστόσο, στη συνέχεια τροποποιείται το μοντέλο προκειμένου να ενσωματώσει το φαινόμενο της ετεροσκεδαστικότητας που παρατηρείται στις χρονοσειρές αποδόσεων. Για το λόγο αυτό, υιοθετείται μεταβλητή διασπορά στην εξίσωση των παρατηρήσεων, η οποία μοντελοποιείται από μια διαδικασία GARCH. Ακολουθεί σύγκριση των μοντέλων με βάση την προσαρμογή τους σε πραγματικά δεδομένα. Για τον σκοπό αυτό, χρησιμοποιούνται οι αποδόσεις του δείκτη Nasdaq και των μετοχών Google, Intel και Microsoft τη χρονική περίοδο 2006-2008 (τα δεδομένα αντλήθηκαν από την ιστοσελίδα finance.yahoo.com).

Σε κάθε περίπτωση η εκτίμηση των αλμάτων γίνεται χρησιμοποιώντας τον επαναληπτικό αλγόριθμο του φίλτρου Kalman (Kalman, 1960; Harvey, 1989; Durbin and Koopman, 2012). Ωστόσο, με δεδομένο ότι η ημερήσια απόδοση ορίζεται ως η διαφορά ανάμεσα στο θετικό και αρνητικό άλμα (υπό την επίδραση θορύβου), οι εκτιμώμενες τιμές των αλμάτων πρέπει να είναι μη αρνητικές για κάθε χρόνο t . Προκειμένου να εξασφαλιστεί αυτό, γίνεται αποκοπή στην κατανομή των εκτιμώμενων αλμάτων με βάση τους περιορισμούς και ως νέες εκτιμήσεις των αλμάτων λαμβάνονται οι αναμενόμενες τιμές των αποκομμένων κατανομών.

2α. ΤΟ ΒΑΣΙΚΟ ΜΟΝΤΕΛΟ

Στην εργασία Θεοδοσιάδου και Τσακλίδης (2015), κατασκευάστηκε ένα γραμμικό μαρκοβιανό μοντέλο χώρου καταστάσεων για την εκτίμηση των θετικών και αρνητικών αλμάτων στις ημερήσιες αποδόσεις χρηματοοικονομικών μεγεθών, όπου τα άλματα είναι οι καταστάσεις του μοντέλου, δηλαδή, οι κρυφές τ.μ. που πρέπει να εκτιμηθούν. Συμβολίζοντας με

X_t : το θετικό άλμα (προς τα πάνω) στην απόδοση ενός χρηματοοικονομικού μεγέθους στο χρόνο t ,

Y_t : το αρνητικό άλμα (προς τα κάτω) στην απόδοση ενός χρηματοοικονομικού μεγέθους στο χρόνο t ,

το ομογενές γραμμικό μοντέλο χώρου καταστάσεων έχει εξίσωση καταστάσεων

$$\left. \begin{aligned} X_t &= g_{11}X_{t-1} + g_{12}Y_{t-1} + w_{1,t} \\ Y_t &= g_{21}X_{t-1} + g_{22}Y_{t-1} + w_{2,t} \end{aligned} \right\}, \quad t=1,2,\dots, \quad (1)$$

ή σε μορφή πινάκων

$$\mathbf{z}_t = \mathbf{G}\mathbf{z}_{t-1} + \mathbf{w}_t, \quad t=1,2,\dots,$$

όπου,

$$\mathbf{z}_t = \begin{pmatrix} X_t \\ Y_t \end{pmatrix}, \quad \mathbf{G} = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}, \quad \mathbf{w}_t = \begin{pmatrix} w_{1,t} \\ w_{2,t} \end{pmatrix}$$

η δε εξίσωση παρατηρήσεων του μοντέλου δίνεται από τη σχέση

$$R_t = X_t - Y_t + e_t, \quad t=1,2,\dots, \quad (2)$$

ή σε μορφή πινάκων

$$R_t = \mathbf{H}\mathbf{z}_t + e_t, \quad t=1,2,\dots,$$

όπου $\mathbf{H}=(1,-1)$, και R_t είναι η απόδοση στο χρόνο t . Επιπλέον, θεωρούμε ότι

$$\mathbf{w}_t \sim N\left(\mathbf{0}, \mathbf{Q} = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}\right), \quad e_t \sim N(0, V),$$

και

$$E[\mathbf{w}_k \mathbf{w}_j^T] = \mathbf{Q}_k \delta_{kj}, \quad E[e_k e_j^T] = V_k \delta_{kj}, \quad E[e_k \mathbf{w}_j^T] = \mathbf{0}.$$

Στόχος μας είναι να εκτιμήσουμε τα άλματα X_t και Y_t με την υιοθέτηση του ανωτέρω μοντέλου. Σημειώνεται ότι εκφράζοντας την απόδοση R_t ως διαφορά ανάμεσα στο θετικό και αρνητικό άλμα, υπό την επίδραση λευκού γκαουσιανού

θορύβου, για τα εκτιμώμενα άλματα X_t και Y_t που ικανοποιούν τη σχέση (2) πρέπει να ισχύει $X_t, Y_t \geq 0$, για κάθε $t=1,2,\dots$.

Η εκτίμηση των αλμάτων γίνεται χρησιμοποιώντας τον επαναληπτικό αλγόριθμο του φίλτρου Kalman, ο οποίος δίνει τους βέλτιστους γραμμικούς εκτιμητές, δηλαδή εκτιμητές ελάχιστης διασποράς. Το φίλτρο λειτουργεί σε δυο στάδια. Στο πρώτο στάδιο (*prediction step*) γίνεται μια εκ των προτέρων εκτίμηση της κατάστασης στο χρόνο t , δηλαδή, χωρίς να είναι γνωστή η παρατήρηση τη δεδομένη χρονική στιγμή. Στο δεύτερο στάδιο (*updating step*) επικαιροποιείται-διορθώνεται η αρχική εκτίμηση, λαμβάνοντας υπόψη αυτή τη φορά την παρατήρηση στο χρόνο t .

Συμβολίζοντας με $\hat{z}_t^-$ την εκ των προτέρων (*a priori*) εκτίμηση της κατάστασης z_t , και με $\hat{z}_t^+$ την εκ των υστέρων (*a posteriori*) εκτίμηση, στον Πίνακα 1 παρατίθενται οι αναδρομικές εξισώσεις του φίλτρου Kalman.

Πίνακας 1. Εξισώσεις του φίλτρου Kalman σε διακριτό χρόνο

Εκ των προτέρων εκτίμηση του z_t	$\hat{z}_t^- = G\hat{z}_{t-1}^+$,
Πίνακας συνδιακύμανσης των εκ των προτέρων σφαλμάτων εκτίμησης του z_t	$P_t^- = GP_{t-1}^+G^T + Q$
Σφάλμα πρόβλεψης	$u_t = R_t - H\hat{z}_t^-$
Διασπορά του σφάλματος πρόβλεψης	$\Omega_t = HP_t^-H^T + V$
Πίνακας Kalman (<i>Kalman gain matrix</i>)	$K_t = P_t^-H^T(HP_t^-H^T + V)^{-1}$
Εκ των υστέρων εκτίμηση του z_t	$\hat{z}_t^+ = \hat{z}_t^- + K_t(u_t - H\hat{z}_t^-)$,
Πίνακας συνδιακύμανσης των εκ των υστέρων σφαλμάτων εκτίμησης του z_t	$P_t^+ = (I - K_tH)P_t^-$

Υποθέτουμε (κοινή αποδοχή), για την έναρξη της επαναληπτικής διαδικασίας του φίλτρου, ότι

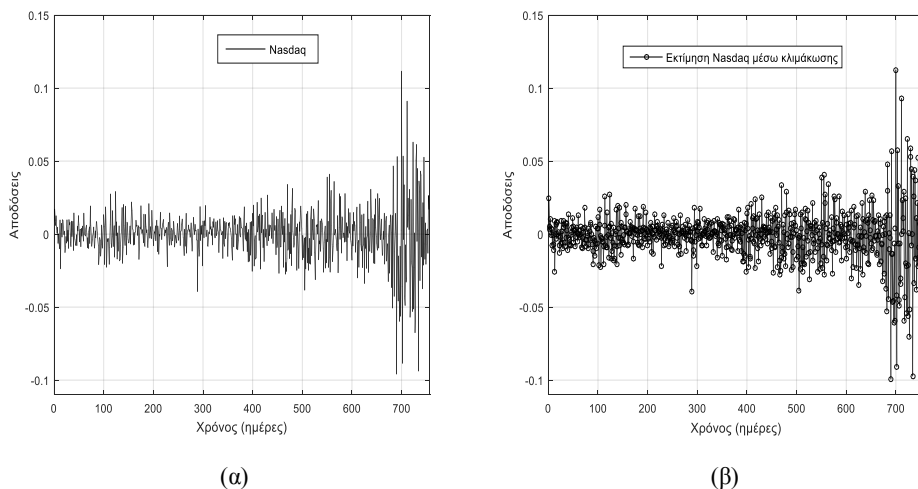
$$\hat{z}_0^+ = E(z_0) \text{ και } P_0^+ = E\left[(z_0 - \hat{z}_0^+)(z_0 - \hat{z}_0^+)^T\right].$$

2β. ΑΠΟΚΟΠΗ - ΚΛΙΜΑΚΩΣΗ ΣΤΟ ΒΑΣΙΚΟ ΜΟΝΤΕΛΟ

Στην εργασία Θεοδοσιάδου και Τσακλίδης (2015), χρησιμοποιώντας το φίλτρο Kalman εκτιμήθηκαν τα θετικά και αρνητικά άλματα του δείκτη Nasdaq για τη χρονική περίοδο 2006-2008 με βάση το μοντέλο (1)-(2). Ωστόσο, επειδή τα εκτιμώμενα άλματα που προέκυψαν δεν ικανοποιούσαν κατά κανόνα τον περιορισμό της μη αρνητικότητας, οι εκτιμήσεις των αλμάτων διορθώθηκαν αποκόπτοντας τη διδιάστατη κατανομή των αλμάτων στο μη αρνητικό επίπεδο.

Η προκύπτουσα χρονοσειρά των αποδόσεων μετά τη διαδικασία αποκοπής υποεκτιμά τις αποδόσεις του δείκτη σε κάθε χρόνο t (βλ. Σχήμα 5, Θεοδοσιάδου κ.α., 2015). Για το λόγο αυτό, ακολουθεί διόρθωση στη χρονοσειρά που προέκυψε μετά το βήμα της αποκοπής μέσω κλιμάκωσής της, όπου ο συντελεστής κλιμάκωσης υπολογίζεται μέσω γραμμικής παλινδρόμησης (Theodosiadou and Tsaklidis, 2016). Στο γράφημα που ακολουθεί φαίνεται η καλή προσαρμογή που επιτυγχάνεται μετά και τη διαδικασία της κλιμάκωσης.

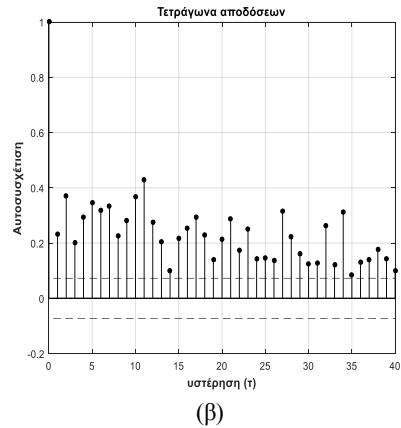
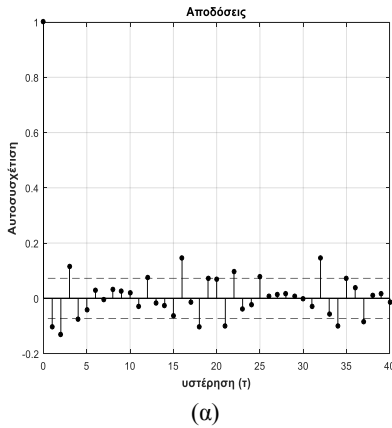
Σχήμα 1. (α) Η χρονοσειρά των αποδόσεων του δείκτη Nasdaq τη χρονική περίοδο 2006-2008, (β) Εκτίμηση των αποδόσεων του δείκτη Nasdaq μέσω κλιμάκωσης (Theodosiadou and Tsaklidis, 2016).



3. ΤΡΟΠΟΠΟΙΗΣΗ ΤΟΥ ΜΟΝΤΕΛΟΥ

Εμπειρικά έχει αποδειχθεί ότι οι χρονοσειρές των αποδόσεων εμφανίζουν το φαινόμενο της δεσμευμένης -ως προς το χρόνο- ετεροσκεδαστικότητας. Στο μοντέλο όμως που αρχικά προτάθηκε στην ενότητα 2, οι διασπορές των θορύβων τόσο στην εξίσωση των καταστάσεων, όσο και στην εξίσωση των παρατηρήσεων θεωρούνται σταθερές στο χρόνο. Προκειμένου να διερευνηθεί η υπόθεση της σταθερής διασποράς, εξετάζεται η συνάρτηση αυτοσυσχέτισης των αποδόσεων και των τετραγώνων των αποδόσεων για τον δείκτη Nasdaq.

Σχήμα 2. (α) Συνάρτηση αυτοσυσχέτισης των αποδόσεων για τον δείκτη Nasdaq, (β) Συνάρτηση αυτοσυσχέτισης των τετραγώνων των αποδόσεων για τον δείκτη Nasdaq.



Με βάση το Σχήμα 2(α)-(β) προκύπτει ότι τα τετράγωνα των αποδόσεων του δείκτη Nasdaq εμφανίζουν πολύ ισχυρές αυτοσυσχετίσεις σε σύγκριση με τις αποδόσεις. Δεδομένου ότι τα τετράγωνα των αποδόσεων μετρούν τη διακύμανση της χρονοσειράς, συμπεραίνουμε ότι η εκάστοτε διακύμανση εξαρτάται από τις προηγούμενες τιμές της. Το αποτέλεσμα του Engle's test για ARCH επιδράσεις στη χρονοσειρά των καταλοίπων, επιβεβαιώνει το παραπάνω συμπέρασμα καθώς,

$$\text{στατ.}=40.0426 > \text{κρ. τιμή}=3.8415, \text{ p-τιμή}=2.48 \times 10^{-10}.$$

Προκειμένου λοιπόν να ενσωματώσουμε το χαρακτηριστικό της ετεροσκεδαστικότητας στο μοντέλο, εισάγουμε έναν νέο όρο e_t^* στην εξίσωση των παρατηρήσεων με μεταβλητή διασπορά. Πιο συγκεκριμένα, γίνεται διάσπαση του θορύβου e_t στην εξίσωση των παρατηρήσεων (εξίσωση (2)), σύμφωνα με τη σχέση

$$e_t = e_t^* + e_t^+,$$

όπου e_t^* και e_t^+ είναι ανεξάρτητες τ.μ. για κάθε t και ισχύει

$$e_t^+ \sim N(0, \sigma_{e^+}^2) \text{ και } e_t^* \sim N(0, h_t).$$

Επιπλέον, μοντελοποιούμε τη μεταβλητή διασπορά h_t του e_t^* ως μια διαδικασία GARCH, δηλαδή,

$$h_t = \omega + \alpha (e_{t-1}^*)^2 + \beta h_{t-1}, \quad (3)$$

με $\omega, \alpha, \beta > 0$ και $\alpha + \beta < 1$.

Με βάση τη σχέση (3), η τιμή της διασποράς h_t εξαρτάται από τις προηγούμενες τιμές του e_t^* . Ωστόσο, ο όρος e_t^* αποτελεί μη παρατηρήσιμη ποσότητα για κάθε t . Προκειμένου να τον εκτιμήσουμε για να μπορέσουμε να υπολογίσουμε τη διασπορά h_t , θεωρούμε ότι είναι μια κρυφή τ.μ. και τον εισάγουμε ως τρίτη συνιστώσα στο διάνυσμα των καταστάσεων z_t . Έτσι, το νέο μοντέλο χώρου καταστάσεων για την εκτίμηση των αλμάτων αποδόσεων έχει εξίσωση καταστάσεων

$$\begin{pmatrix} X_t \\ Y_t \\ e_t^* \end{pmatrix} = \begin{pmatrix} \mathbf{G} & \mathbf{0}_{2 \times 1} \\ \mathbf{0}_{1 \times 2} & 0 \\ & & 1 \end{pmatrix} \begin{pmatrix} X_{t-1} \\ Y_{t-1} \\ e_{t-1}^* \end{pmatrix} + \begin{pmatrix} w_{1,t} \\ w_{2,t} \\ e_t^* \end{pmatrix}$$

(4)

και εξίσωση παρατηρήσεων

$$R_t = (1 \quad -1 \quad 1) \begin{pmatrix} X_t \\ Y_t \\ e_t^* \end{pmatrix} + e_t^+ \quad (5)$$

όπου $e_t^+ \sim N(0, \sigma_{e^+}^2)$, $e_t^* \sim N(0, h_t)$ με $h_t = \omega + \alpha (e_{t-1}^*)^2 + \beta h_{t-1}$ ($\omega, \alpha, \beta > 0$ και $\alpha + \beta < 1$) και

$$\begin{pmatrix} w_{1,t} \\ w_{2,t} \end{pmatrix} \sim NWN \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix} \right)$$

Σημειώνεται ότι και σε αυτή την περίπτωση τα άλματα που θα εκτιμηθούν πρέπει να ικανοποιούν τον περιορισμό $X_t, Y_t \geq 0$, για κάθε $t=1, 2, \dots$.

4. ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ

Προκειμένου να εκτιμήσουμε τις άγνωστες παραμέτρους του μοντέλου (4)-(5), δηλαδή τον πίνακα μετάβασης G , τους συντελεστές ω, α, β της διαδικασίας GARCH και τις διασπορές $\sigma_{e^+}^2, \sigma_x^2, \sigma_y^2$, εφαρμόζουμε τη μέθοδο της μέγιστης πιθανοφάνειας. Η τ.μ. R_t με δεδομένη την R_{t-1} ακολουθεί κανονική κατανομή, $R_{t|t-1} \sim N(Hz_t^-, \Omega_t)$. Επομένως, ο λογάριθμος της συνάρτησης πιθανοφάνειας έχει τη μορφή

$$\text{Log}L(R_1, \dots, R_n) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{t=1}^n \left(\log(|\Omega_t|) + u_t^T \Omega_t^{-1} u_t \right), \quad n=755.$$

Στην περίπτωση του δείκτη Nasdaq για τη χρονική περίοδο 2006-2008, προκύπτουν οι παρακάτω εκτιμήσεις

$$G = \begin{pmatrix} 0.2158 & 0.6658 \\ 0.0775 & 0.5450 \end{pmatrix},$$

$$\sigma_x^2 = 2.4062 \times 10^{-8}, \sigma_y^2 = 3.0088 \times 10^{-5}, \sigma_{e^+}^2 = 1.2663 \times 10^{-9}$$

και οι συντελεστές της διαδικασίας GARCH παίρνουν τις τιμές

$$\omega = 1.1337 \times 10^{-6}, \alpha = 0.1678 \text{ και } \beta = 0.8240.$$

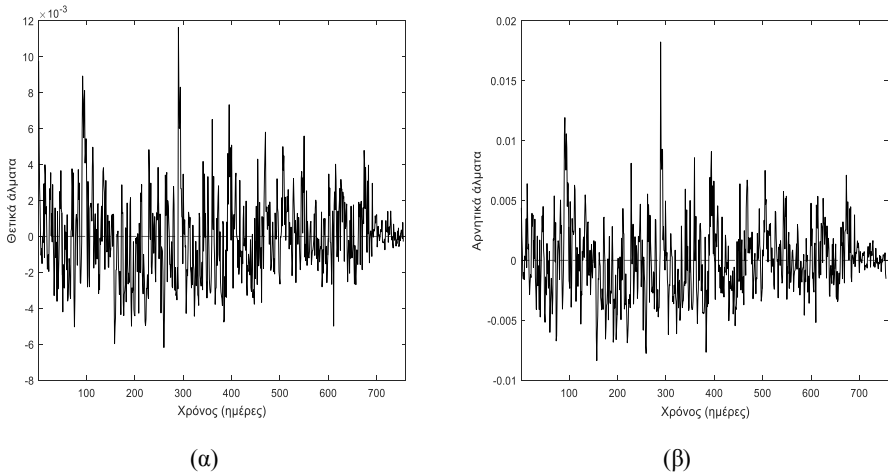
Λαμβάνοντας επομένως ως αρχικό διάνυσμα καταστάσεων για τον αλγόριθμο του φίλτρου Kalman το (Θεοδοσιάδου και Τσακλίδης, 2015)

$$\hat{z}_0(+) = \begin{pmatrix} X_0 \\ Y_0 \\ e_0^* \end{pmatrix} = \begin{pmatrix} 0.0092 \\ 0.0108 \\ 0 \end{pmatrix} \quad 82$$

$$\text{με } \mathbf{P}_0^+ = \begin{pmatrix} 0.03 & 0 & 0 \\ 0 & 0.3 & 0 \\ 0 & 0 & h_0 \end{pmatrix},$$

όπου $h_0 = \omega \times (1 - \alpha - \beta)^{-1}$ (μη δεσμευμένη διασπορά της διαδικασίας GARCH), προκύπτουν οι (εκ των υστέρων) εκτιμήσεις των αλμάτων που δίνονται στο Σχήμα 3 που ακολουθεί.

Σχήμα 3. (α) Τα εκτιμώμενα θετικά άλματα του δείκτη Nasdaq με βάση το μοντέλο (4)-(5), (β) Τα εκτιμώμενα αρνητικά άλματα του δείκτη Nasdaq με βάση το μοντέλο (4)-(5).



Όπως διαπιστώθηκε, το φίλτρο Kalman δεν μπορεί να εγγυηθεί τη μη αρνητικότητα στα εκτιμώμενα άλματα για κάθε χρόνο t . Για το λόγο αυτό, οι εκτιμήσεις που προέκυψαν με βάση το φίλτρο θα διορθωθούν, χρησιμοποιώντας τη μέθοδο της αποκοπής της κατανομής των αλμάτων. Σε αυτήν την περίπτωση, το διάνυσμα κατάστασης $\mathbf{z}_t$ ακολουθεί μια τριδιάστατη κανονική κατανομή με δεδομένη την τιμή $\mathbf{R}_t$, που δίνεται από τη σχέση

$$\mathbf{z}_t = \begin{pmatrix} X_t \\ Y_t \\ e^*_t \end{pmatrix} \sim N \left(\begin{pmatrix} \hat{X}_t(+) \\ \hat{Y}_t(+) \\ \hat{e}^*_t(+) \end{pmatrix}, \mathbf{\Lambda}_t = \begin{bmatrix} \mathbf{P}_t(+) & 0 \\ 0 & h_t \end{bmatrix} \right).$$

Επομένως, η δεσμευμένη συνάρτηση πυκνότητας της $\mathbf{z}_t$ για κάθε t , έχει τη μορφή

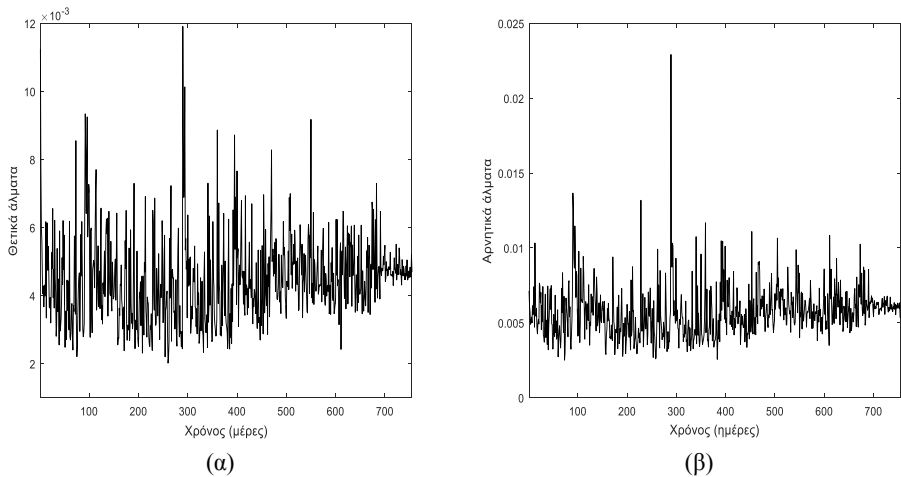
$$f(\mathbf{z}_t; \hat{\mathbf{z}}_t^+, \Lambda_t) = \frac{1}{(2\pi)^{d/2} |\Lambda_t|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{z}_t - \hat{\mathbf{z}}_t^+)^T (\Lambda_t)^{-1} (\mathbf{z}_t - \hat{\mathbf{z}}_t^+) \right\}, \quad \mathbf{z}_t \in \mathbb{R}^2,$$

και η αντίστοιχη συνάρτηση για την αποκομμένη κανονική κατανομή με άκρα $\alpha, \beta \in \mathbb{R}^3$ δίνεται από τη σχέση

$$f(\mathbf{z}_t; \hat{\mathbf{z}}_t^+, \Lambda_t, \alpha, \beta) = \begin{cases} \frac{f(\mathbf{z}_t; \hat{\mathbf{z}}_t^+, \Lambda_t)}{P(\alpha \leq \mathbf{Z}_t \leq \beta)}, & \alpha \leq \mathbf{z}_t \leq \beta \\ 0, & \text{αλλιού.} \end{cases}$$

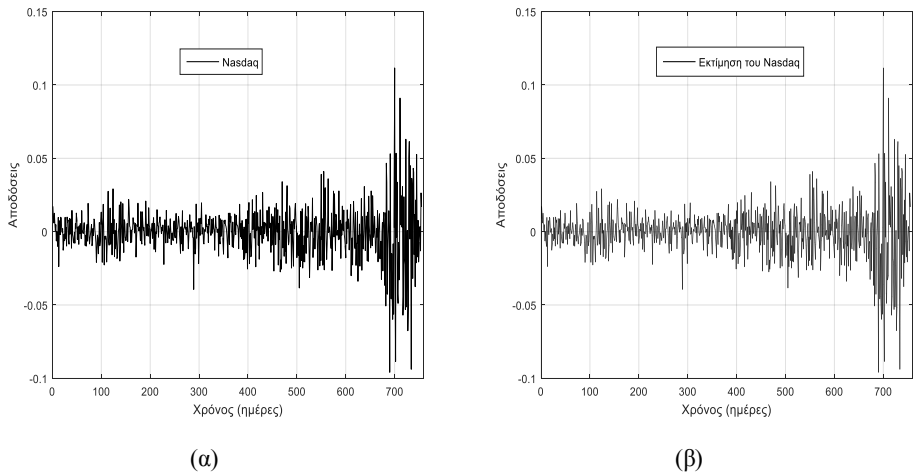
Στην περίπτωση μας, η κατανομή αποκόπτεται στο σημείο $\alpha = (0, 0, -\infty)^T$. Για το σκοπό αυτό χρησιμοποιήθηκε το υπολογιστικό πακέτο “tmvtnorm” της R (Wilhelm and Manjunath, 2010), και λήφθηκαν ως νέες εκτιμήσεις των αλμάτων οι αναμενόμενες τιμές της αποκομμένης κανονικής κατανομής. Τα εκτιμώμενα άλματα που προκύπτουν με τη μέθοδο της αποκοπής δίνονται στο Σχήμα 4 που ακολουθεί.

Σχήμα 4. (α) Τα εκτιμώμενα θετικά άλματα του δείκτη Nasdaq με βάση την αποκομμένη κατανομή, (β) Τα εκτιμώμενα αρνητικά άλματα του δείκτη Nasdaq με βάση την αποκομμένη κατανομή.



Πλέον, όπως παρατηρείται στο Σχήμα 4(α)-(β), τα εκτιμώμενα άλματα είναι μη αρνητικά για κάθε χρόνο t . Η προσαρμογή του τροποποιημένου μοντέλου (4)-(5), στη χρονοσειρά του δείκτη που προκύπτει με βάση τη σχέση (5) και τα άλματα που εκτιμήθηκαν μετά τη διαδικασία αποκοπής, φαίνεται στο Σχήμα 5 που ακολουθεί.

Σχήμα 5. (α) Αποδόσεις του δείκτη Nasdaq για τη χρονική περίοδο 2006-2008, (β) Εκτίμηση των αποδόσεων του δείκτη Nasdaq με βάση τη σχέση (5).



Παρατηρούμε ότι τα γραφήματα στο Σχήμα 5 φαίνονται σχεδόν ίδια μεταξύ τους και έχει έννοια η εκτίμηση της προσαρμογής.

5. ΣΥΓΚΡΙΣΗ ΜΟΝΤΕΛΩΝ

Χρησιμοποιώντας το μέσο τετραγωνικό σφάλμα (MSE) (ισοδύναμα, την τετραγωνική του ρίζα (RMSE)) και το τεστ προσαρμογής Kolmogorov-Smirnov για να συγκρίνουμε το βασικό μοντέλο (1)-(2) και το τροποποιημένο (4)-(5) με βάση την προσαρμογή τους στη χρονοσειρά του δείκτη Nasdaq, προκύπτει ο παρακάτω πίνακας.

Πίνακας 2. Μέσο τετραγωνικό σφάλμα και τεστ προσαρμογής Kolmogorov-Smirnov για την εκτίμηση των αποδόσεων του δείκτη Nasdaq τη χρονική περίοδο 2006-2008.

Μοντέλο	MSE	RMSE	K.S. (p-τιμή)
Ομογενές	1.41×10^{-6}	0.0012	0.0609 (0.1169)
Μη ομογενές	2.17×10^{-14}	1.47×10^{-7}	0.0013 (0.99)

Με βάση τον Πίνακα 2, συμπεραίνουμε ότι η εκτίμηση του δείκτη που επιτυγχάνεται με το μοντέλο (4)-(5), όπου προστίθεται ένας νέος όρος (νέα μεταβλητή), είναι πολύ πιο ικανοποιητική σε σχέση με αυτήν που επιτυγχάνεται με

τη χρήση του ομογενούς μοντέλου (1)-(2). Επιπλέον, διαπιστώνεται ότι η πολύ ικανοποιητική προσαρμογή στα δεδομένα, που προκύπτει στην περίπτωση όπου υιοθετείται μεταβλητή διασπορά τύπου GARCH στην εξίσωση των καταστάσεων στο μοντέλο (4)-(5), επιτυγχάνεται αποφεύγοντας το βήμα της κλιμάκωσης που χρησιμοποιείται στην αρχική μεθοδολογία μετά τη διαδικασία αποκοπής. Τέλος, με τη χρήση του τεστ Shapiro-Wilk, διαπιστώνεται ότι τα σφάλματα εκτίμησης $\varepsilon_t = R - \hat{R}_t$ της χρονοσειράς του δείκτη Nasdaq χρησιμοποιώντας τις εκτιμήσεις που προκύπτουν μετά τη διαδικασία της αποκοπής στο μη ομογενές μοντέλο και το βήμα της κλιμάκωσης στο ομογενές μοντέλο, δεν ακολουθούν πλέον στην αποκομμένη χρονοσειρά την κανονική κατανομή, όπως φαίνεται στον Πίνακα 3, που ακολουθεί.

Πίνακας 3. *Shapiro-Wilk τεστ για την κανονικότητα των σφαλμάτων εκτίμησης.*

Μοντέλο	Shapiro-Wilk (p-τιμή)
Ομογενές	0.9418 ($< 2.2 \times 10^{-16}$)
Μη ομογενές	0.9766 (1.22×10^{-9})

Προκειμένου να διαπιστωθεί αν τα παραπάνω συμπεράσματα ισχύουν και για άλλες χρονοσειρές αποδόσεων, η παραπάνω διαδικασία ακολουθήθηκε και για 3 άλλες βασικές μετοχές του δείκτη Nasdaq κατά την ίδια χρονική περίοδο, τις μετοχές Google, Intel και Microsoft. Εκτιμήθηκαν οι παράμετροι των μοντέλων τόσο για το ομογενές, όσο και για το μη ομογενές. Στη συνέχεια εκτιμήθηκαν τα άλματα με τη χρήση του φίλτρου Kalman και στις 2 περιπτώσεις μοντέλων. Κατόπιν, διορθώθηκαν τα εκτιμώμενα άλματα εφαρμόζοντας τη μέθοδο της αποκοπής προκειμένου να είναι μη αρνητικά. Αναφορικά με το ομογενές μοντέλο (1)-(2), μετά την αποκοπή των εκτιμώμενων αλμάτων, ακολούθησε κλιμάκωση των εκτιμήσεων των αποδόσεων. Τέλος, εξετάστηκε η προσαρμογή στις χρονοσειρές των μετοχών.

Στους πίνακες που ακολουθούν δίνονται οι παράμετροι που εκτιμήθηκαν με τη μέθοδο της μέγιστης πιθανοφάνειας για το μοντέλο (1)-(2), στην περίπτωση των μετοχών.

Πίνακας 4. *Εκτίμηση του πίνακα μετάβασης G στο μοντέλο (1)-(2).*

Μετοχή	g_{11}	g_{12}	g_{21}	g_{22}
Google	0.2959	0.4396	0.1466	0.2360
Intel	8.2780	-8.9901	8.6001	-9.2657
Microsoft	6.3131	-6.9034	6.5867	-7.1254

Πίνακας 5. *Εκτίμηση διασπορών στο μοντέλο (1)-(2) .*

Μετοχή	σ_x^2	σ_y^2	σ_e^2
Google	3.9009×10^{-5}	1.7301×10^{-4}	4.0115×10^{-4}
Intel	2.4050×10^{-8}	1.0696×10^{-4}	3.8919×10^{-4}
Microsoft	1.3176×10^{-8}	1.3380×10^{-4}	2.5682×10^{-4}

Κατόπιν, ακολουθούν οι εκτιμήσεις των παραμέτρων για το μοντέλο (4)-(5).

Πίνακας 6. Εκτίμηση του πίνακα μετάβασης G στο μοντέλο (4)-(5).

Μετοχή	g_{11}	g_{12}	g_{21}	g_{22}
Google	-0.1205	0.1401	-0.2142	0.4070
Intel	2.4408	-1.4344	2.3847	-1.3435
Microsoft	0.8406	-1.8320	0.9355	-1.9146

Πίνακας 7. Εκτίμηση παραμέτρων της διαδικασίας GARCH στο μοντέλο (4)-(5) .

Μετοχή	ω	α	β
Google	1.5152×10^{-6}	0.0603	0.9384
Intel	3.3289×10^{-6}	0.1428	0.8554
Microsoft	3.6240×10^{-6}	0.2588	0.7271

Πίνακας 8. Εκτίμηση διασπορών στο μοντέλο (4)-(5) .

Μετοχή	σ_x^2	σ_y^2	σ_{e+}^2
Google	1.1071×10^{-9}	6.3202×10^{-5}	8.0748×10^{-10}
Intel	1.2301×10^{-9}	8.6121×10^{-5}	1.8977×10^{-6}
Microsoft	4.4482×10^{-10}	1.1225×10^{-4}	2.3764×10^{-10}

Τέλος, ακολουθεί ο πίνακας με το μέσο τετραγωνικό σφάλμα που προκύπτει με βάση τη προσαρμογή των μοντέλων στις χρονοσειρές των μετοχών.

Πίνακας 9. Το μέσο τετραγωνικό σφάλμα για την εκτίμηση αποδόσεων μετοχών την χρονική περίοδο 2006-2008.

Μοντέλο	Σφάλμα	Google	Intel	Microsoft

Ομογενές	MSE	1.85×10^{-4}	1.68×10^{-4}	6.88×10^{-5}
	RMSE	0.0136	0.0130	0.0083
Μη ομογενές	MSE	3.1×10^{-15}	5.98×10^{-9}	2.58×10^{-15}
	RMSE	5.57×10^{-8}	7.73×10^{-5}	1.61×10^{-8}

Παρατηρώντας τον Πίνακα 9, διαπιστώνουμε ότι το μοντέλο (4)-(5), στο οποίο έχουμε εισαγάγει έναν νέο όρο με μεταβλητή διασπορά στην εξίσωση των παρατηρήσεων (διαδικασία GARCH), έχει καλύτερη προσαρμογή στις χρονοσειρές των αποδόσεων, με βάση την τιμή του μέσου τετραγωνικού σφάλματος.

6. ΣΥΜΠΕΡΑΣΜΑΤΑ

Τα θετικά και αρνητικά άλματα στις ημερήσιες αποδόσεις χρηματοοικονομικών μεγεθών, καθότι είναι μη παρατηρήσιμες τυχαίες μεταβλητές, μπορούν να εκτιμηθούν με τη βοήθεια στοχαστικών μοντέλων χώρου καταστάσεων. Για την εκτίμησή τους χρησιμοποιείται ο επαναληπτικός αλγόριθμος του φίλτρου Kalman.

Με δεδομένο ότι η ημερήσια απόδοση εκφράζεται ως η διαφορά ανάμεσα στο θετικό και αρνητικό άλμα (υπό την επίδραση λευκού γκαουσσιανού θορύβου), τα εκτιμώμενα άλματα που προκύπτουν μέσω του φίλτρου Kalman, πρέπει να είναι μη αρνητικά. Ωστόσο, το απλό (γραμμικό) φίλτρο Kalman δεν μπορεί να εγγυηθεί την ικανοποίηση αυτού του περιορισμού και προκειμένου να γίνει διόρθωση της χρονοσειράς των αλμάτων που προκύπτουν από το φίλτρο, υιοθετήθηκε η μέθοδος της αποκοπής των κατανομών τους.

Συγκρίνοντας το ομογενές και το μη ομογενές μοντέλο χώρου καταστάσεων, όπου προστίθεται στην εξίσωση των παρατηρήσεων του δεύτερου ένας νέος όρος με μεταβλητή διασπορά που περιγράφεται από μια διαδικασία GARCH, διαπιστώθηκε ότι το μη ομογενές μοντέλο έχει καλύτερη προσαρμογή στα εμπειρικά δεδομένα. Επιπλέον, διαπιστώθηκε ότι στην περίπτωση του μη ομογενούς μοντέλου υπήρξε πολύ καλή προσαρμογή στις χρονοσειρές των αποδόσεων, χωρίς να είναι απαραίτητη η διαδικασία της κλιμάκωσης μετά το στάδιο της αποκοπής, όπως έγινε στην περίπτωση του ομογενούς μοντέλου.

Μελλοντικά, προς την κατεύθυνση της ενσωμάτωσης του φαινομένου της ετεροσκεδαστικότητας στο μοντέλο χώρου καταστάσεων για την εκτίμηση των αλμάτων στις αποδόσεις, θα μπορούσαν να εισαχθούν δυο νέοι όροι με μεταβλητές διασπορές στο σύστημα των εξισώσεων των καταστάσεων, δηλαδή των αλμάτων, και να διαπιστωθεί ποιο μοντέλο τελικά επιτυγχάνει την καλύτερη προσαρμογή. Διαφορετικά, μπορούν να κατασκευαστούν μη γραμμικά μοντέλα χώρου καταστάσεων και να υιοθετηθεί η χρήση γραμμικών αλγορίθμων (Extended Kalman filter, Particle filtering) για την εκτίμηση των αλμάτων.

ABSTRACT

The daily asset log return is considered to consist of two parts, the positive and negative jump. These jumps are unobservable and for that purpose we use stochastic state space models in discrete time so as to estimate them. Two state space models are compared based on their fitting to real data. In the first model, the variances in the measurement and state equations are assumed to be time invariant, whereas in the second model a new random variable is added to the measurement equation with time variant variance described by a GARCH process. In both models, the recursive algorithm of Kalman filter is used in order to estimate the hidden jumps. However, due to the fact that the return is described as the difference between the positive and negative jump, the estimated jumps must be non negative. For that purpose, the pdf truncation method, according to the non negativity constraints, is used in the time series of jumps derived by Kalman filter.

ΑΝΑΦΟΡΕΣ

- Θεοδοσιάδου, Ο., Πολυμένης Β. και Τσακλίδης, Γ. (2013). Διερεύνηση ενός μοντέλου για τον διαχωρισμό του θορύβου από την άφιξη πληροφορίας στην εξέλιξη της τιμής μιας μετοχής. *Πρακτικά 26^{ου} Πανελλήνιου Συνεδρίου Στατιστικής*, σελ.76-86.
- Θεοδοσιάδου, Ο., Τσακλίδης Γ. και Πολυμένης Β. (2014). Διερεύνηση των αποδόσεων χρηματιστηριακών δεικτών ως προς τα θετικά και αρνητικά άλματα των αποδόσεων. Η περίπτωση της μετοχής Google. *Πρακτικά 27^{ου} Πανελλήνιου Συνεδρίου Στατιστικής*, σελ. 84-96.
- Θεοδοσιάδου, Ο. και Τσακλίδης, Γ. (2015). Προσδιορισμός φίλτρου Kalman για την εκτίμηση των θετικών και αρνητικών αλμάτων των αποδόσεων χρηματιστηριακών δεικτών. *Πρακτικά 28^{ου} Πανελλήνιου Συνεδρίου Στατιστικής*, σελ. 133-144.
- Anderson, T., Luca, B. and Jesper, L. (2002). An empirical investigation of continuous-time equity return models. *Journal of Finance*, **57**(3), 1239–1284.
- Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *J. Polit. Econ.*, **81**, 637–659 .
- Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *J. of Econometrics*, **31**(3), 307-327
- Cox, J. and Ross, S. (1976). The valuation of options for alternative stochastic processes. *Journal of Financial Economics*, **3**, pp. 145–166.
- Durbin, J., and Koopman, S.J. (2012). *Time Series Analysis by State Space Methods* (Second Edition). Oxford University Press, Oxford (UK).
- Eraker, B., Michael, S.J., Nicholas, P. (2003). The impact of jumps in returns and volatility. *Journal of Finance*, **58**, 1269–1300.

- Fouque, J.P., Papanicolaou, G. and Sircar, R. (2000). Derivatives in Financial Markets with Stochastic Volatility. Cambridge University Press, Cambridge (UK).
- Hull, J. and White, A. (1987). The pricing of Options on Assets with Stochastic Volatilities. *Journal of Finance*, **42**(2), 281-300
- Kalman, R. (1960). A new approach to linear filtering and prediction problems. *J. Basic Engineering (Transactions of the ASME, Series D)*, **82**, 35-45.
- Pan, J. (2002). The jump-risk premia implicit in options: Evidence from an integrated time-series study. *J. Financ. Econ.*, **63**, 3–50.
- Polimenis, V. (2012). Information arrival as price jumps. *Optimization: A Journal of Mathematical Programming and Operations Research*, **61**(10), 1179-1190.
- Theodosiadou, O. and Tsaklidis, G. (2016). Estimating the positive and negative jumps of asset returns via Kalman filtering. The case of Nasdaq index. *J. of Methodology and Computing in Applied Probability*, DOI: 10.1007/s11009-016-9532-5 (δεκτή προς δημοσίευση).
- Wilhelm S, Manjunath B.G. (2010). tmvtnorm: A Package for the Truncated Multivariate Normal Distribution. *The R Journal*, **2**(1), 25-29.
- Merton, R. C. (1973). Theory of rational option pricing, *Bell Journal of Economics and Management Science*, **4**(1), 141-183.

ΣΥΜΒΟΛΗ ΤΩΝ ΔΕΙΚΤΩΝ ΠΕΡΙΒΑΛΛΟΝΤΙΚΗΣ ΑΝΟΜΟΙΟΤΗΤΑΣ ΣΤΟΝ ΕΛΕΓΧΟ ΕΚΠΟΜΠΩΝ ΑΕΡΙΩΝ ΡΥΠΩΝ

Ε. Κετζάκη¹, Σ. Παλλάκη²

¹Τμήμα Μαθηματικών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
eketzaki@yahoo.gr

²Τμήμα Διοίκησης Επιχειρήσεων, Πανεπιστήμιο Μακεδονίας
srallakis@yahoo.gr

ΠΕΡΙΛΗΨΗ

Η αύξηση της εκπομπής αέριων ρύπων στην ατμόσφαιρα αποτελεί παγκοσμίως θέμα μείζονος σημασίας. Παράλληλα με την συστηματική προσπάθεια καταγραφής του φαινομένου οι διεθνείς οργανισμοί προσπαθούν να κυρώσουν συμφωνίες που θα καθορίζουν το μέγιστο επιτρεπόμενο όριο εκπομπής σωματιδίων ανά χώρα. Στην συγκεκριμένη εργασία θα αναπτυχθεί μια μέθοδος υπολογισμού της περιβαλλοντικής ανομοιότητας αναφορικά με την εκπομπή αέριων ρύπων. Η μέθοδος αυτή θα εξετάσει όχι μόνο την μέτρηση της ανομοιότητας αλλά θα συμβάλλει και στον εντοπισμό του προβλήματος του «λαθρεπιβάτη», το οποίο αποτελεί ένα συνηθισμένο αίτιο αποτυχίας διακρατικών συμφωνιών. Στην πρώτη ενότητα της εργασίας παρουσιάζεται το πρόβλημα που θα εξεταστεί, στην συνέχεια αναπτύσσεται η μέθοδος που προτείνεται και στο τέλος περιγράφεται μια εφαρμογή της συγκεκριμένης μεθόδου.

Λέξεις Κλειδιά: περιβαλλοντική ανομοιότητα, πρόβλημα του «λαθρεπιβάτη», κλιματική αλλαγή, δείκτης Gini, πρωτόκολλο του Kyoto.

1. ΕΙΣΑΓΩΓΗ

Η αύξηση των εκπομπών των αέριων ρύπων που ευθύνονται για το φαινόμενο του θερμοκηπίου, προκαλείται κυρίως από την ανθρώπινη δραστηριότητα, τόσο σε βιομηχανικό όσο και σε οικιακό επίπεδο. Το φαινόμενο του θερμοκηπίου έχει σαν άμεση συνέπεια την διατάραξη του φυσικού κύκλου του κλίματος, προκαλώντας γρηγορότερη αύξηση της μέσης θερμοκρασίας του πλανήτη. Η ακαθόριστη εναλλαγή των εποχικών φαινομένων, η αύξηση της στάθμης της θάλασσας που οφείλεται στο λιώσιμο των πολικών πάγων, οι έντονες ξηρασίες και πλημμύρες και ο κίνδυνος αφανισμού που αντιμετωπίζουν πολλά είδη της πανίδας και της χλωρίδας σε παγκόσμιο επίπεδο, είναι μόνο μερικές από τις συνέπειες της κλιματικής διαταραχής.

Ο μοναδικός εφαρμόσιμος μέχρι σήμερα τρόπος, για να περιοριστούν και κατά το δυνατόν να εξομαλυνθούν οι συνέπειες της κλιματικής αλλαγής, είναι η άμεση μείωση των εκπομπών των αερίων του θερμοκηπίου (CO₂, CH₄, N₂O και άλλα φθοριούχα αέρια) κυριότερο από τα οποία θεωρείται το διοξείδιο του άνθρακα (CO₂). Το CO₂ εκπέμπεται σε μεγάλες ποσότητες από την καύση ορυκτών καυσίμων, όπως ο άνθρακας και το πετρέλαιο. Η αντικατάσταση της παρούσας ρυπογόνου παραγωγής καθίσταται δύσκολη, αφενός εξαιτίας της αποκλειστικής χρησιμοποίησης των παραγώγων του άνθρακα στην βιομηχανική παραγωγή και στις κάθε είδους μεταφορές ως καύσιμο και αφετέρου λόγω της απουσίας μιας οικολογικής, ευρείας κλίμακας και αντίστοιχα αποδοτικής με την καύση των οργανικών ενώσεων ενέργειας.

Το πρόβλημα των ατμοσφαιρικών ρύπων έχει διασυνοριακό και όχι τοπικό χαρακτήρα. Εκτός από τις συνέπειες των δικών της αερίων εκπομπών, η κάθε χώρα επιβαρύνεται και από τις εκπομπές ρύπων των άλλων χωρών. Από τα τέλη του περασμένου αιώνα όπου το πρόβλημα της αύξησης της θερμοκρασίας άρχισε να απασχολεί όλο και περισσότερο την επιστημονική κοινότητα αλλά και την διεθνή πολιτική σκηνή, έχουν καταβληθεί προσπάθειες από πολλούς φορείς, ώστε να ληφθούν αυστηρές δεσμεύσεις από τις χώρες για μείωση των εκπομπών τους. Τα εμπόδια που προκύπτουν αφορούν την υιοθέτηση των στρατηγικών μείωσης από τις αναπτυγμένες χώρες, οι οποίες προσπαθούν να μην χάσουν τα εμπορικά και εξαγωγικά κεκτημένα τους, καθώς και από τις αναπτυσσόμενες χώρες, που προσπαθούν μέσω της παραδοσιακής δυτικής βιομηχανοποίησης να αυξήσουν το εγχώριο προϊόν τους.

Η κύρωση και εφαρμογή Διεθνών Περιβαλλοντικών Συμφωνιών (Δ.Π.Σ.) θα μπορούσε να αποτελέσει λύση στο πρόβλημα των ατμοσφαιρικών ρύπων. Τα κράτη μέλη μίας Δ.Π.Σ. έχουν ως στόχο να μειώσουν το σύνολο των αερίων εκπομπών τους, δημιουργώντας παράλληλα ένα περιβάλλον με υψηλή μέση ευημερία εντός της συνθήκης. Παρόλα αυτά η μέχρι τώρα απουσία μιας υπερεθνικής αρχής που μπορεί να επιβάλλει κυρώσεις για την μη τήρηση των συμφωνηθέντων όρων από τα μέλη της συμφωνίας ή για την μη συμμετοχή κρατών σε αυτή, καθιστά εθελοντική ή αυτοεπιβαλλόμενη την επικύρωση μιας τέτοιου είδους συμφωνίας (D' Aspremont, et al, 1983). Αυτό έχει ως συνέπεια, κάθε ανεξάρτητη χώρα δεδομένου ότι θέλει να μεγιστοποιήσει την κοινωνική της ευημερία, να επιλέγει να μην υπογράψει μια συμφωνία, από την οποία το όφελος που θα λάβει θα είναι μικρότερο από ότι με την μη συμμετοχή της σε αυτή (Σαρτζετάκης και Ραλλάκης, 2011). Επιπλέον, δρώντας μια χώρα σαν «λαθρεπιβάτης» (free rider, Finus, 2008), εκμεταλλεύεται την μείωση των ρύπων των άλλων χωρών με διττό τρόπο. Αφενός μειώνει τις συνέπειες που προκαλούνται από τους ατμοσφαιρικούς ρύπους στο εγχώριο σύστημα και αφετέρου αυξάνει παράλληλα την δική της κατανάλωση ορυκτών καυσίμων, με στόχο την αύξηση της εγχώριας παραγωγής και την βελτίωση της ανταγωνιστικότητας της οικονομίας της. Το φαινόμενο αυτό ονομάζεται «διαρροή άνθρακα» (Heugues, 2014). Οι χώρες οι οποίες δρουν ως «λαθρεπιβάτες», εκμεταλλεύονται ένα δημόσιο αγαθό, όπως είναι το περιβάλλον με άνισο, προς το σύνολο τρόπο.

Το πρωτόκολλο που πραγματοποιήθηκε στο Kyoto της Ιαπωνίας το 1997 είναι μία από τις πιο σημαντικές διαχρονικά προσπάθειες, για την επίτευξη συμφωνίας προς τον περιορισμό των αερίων του θερμοκηπίου. Με την υπογραφή του πρωτοκόλλου που ακολούθησε, τα 40 μέλη (39 χώρες και η Ευρωπαϊκή Ένωση), συμφώνησαν σε μείωση των εκπομπών των αερίων ρύπων τους, που προέρχονται από βιομηχανικές και οικιακές πηγές, θέτοντας άμεσους και απώτερους στόχους. Το πρώτο ορόσημο για την εξέταση επίτευξης του στόχου συμφωνήθηκε να είναι το 2012. Έτσι τα κράτη μέλη είχαν μια περίοδο 15 χρόνων ώστε να βελτιώσουν τους τρόπους παραγωγής τους, να εφαρμόσουν νέες τεχνολογίες και να καταφέρουν να μειώσουν τις εκπομπές τους. Οι στόχοι που τέθηκαν ήταν διαφορετικοί για κάθε χώρα, συνυπολογίζοντας το μέγεθος της παραγωγής της, την εξάρτηση της παραγωγής της από τα ορυκτά καύσιμα και την φάση της οικονομίας της. Πιο συγκεκριμένα οι χώρες που συμμετείχαν συμφώνησαν να μεταβάλλουν, μέχρι την πρώτη χρονιά-στόχο που ορίστηκε να είναι το 2012, τους αέριους ρύπους CO₂ κατά ένα ποσοστό ως προς το έτος βάσης που αποφασίστηκε για την κάθε μία ξεχωριστά. Στον Πίνακα 1 παρουσιάζονται οι χώρες που συμμετείχαν στο πρωτόκολλο του Kyoto (Annex I χώρες), το αντίστοιχο ποσοστό μείωσης ή αύξησης που προβλέπεται από την συνθήκη, καθώς και το έτος βάση για κάθε μια από τις χώρες.

Πίνακας 1. Προβλεπόμενα ποσοστά και έτος βάσης των χωρών που συμμετείχαν στο πρωτόκολλο του Kyoto.

Annex I Kyoto Χώρες	p _i (CO ₂) έως 2012	Έτος βάση
Αυστρία, Λευκορωσία, Βέλγιο, Τσεχία, Δανία, Εσθονία, Φιλανδία, Γαλλία, Γερμανία, Ελλάδα, Ιρλανδία, Ιταλία, Λετονία, Λίχτενσταϊν, Λιθουανία, Λουξεμβούργο, Μονακό, Ολλανδία, Πορτογαλία, Σλοβακία, Ισπανία, Σουηδία, Ελβετία, Ηνωμένο Βασίλειο	-8%	1990
Βουλγαρία	-8%	1988
Ουγγαρία	-6%	(1987-1985)/3
Σλοβενία	-8%	1986
Ρουμανία	-8%	1989
Πολωνία	-6%	1988
Ηνωμένες Πολιτείες Αμερικής	-7%	1990
Καναδάς, Ιαπωνία	-6%	1990
Κροατία	-5%	1990
Νέα Ζηλανδία, Ρωσία, Ουκρανία	0%	1990
Νορβηγία	+1%	1990
Αυστραλία	+8%	1990
Ισλανδία	+10%	1990

Το πρωτόκολλο του Kyoto δεν κατάφερε να δημιουργήσει μια ευνοϊκή σχέση κόστους-οφέλους για τα μέλη του, να μοιράσει εξίσου τα οικονομικά βάρη που προέκυψαν μεταξύ των χωρών ή να αποτρέψει την «λαθρεπιβίβαση» και να τιμωρήσει την μη συμμόρφωση στους στόχους (Barrett, 1998, Nordhaus and Boyer,

1998). Είναι όμως η πρώτη προσπάθεια προς την κατεύθυνση της αναγνώρισης του προβλήματος και της επίλυσης του.

2. ΠΡΟΤΕΙΝΟΜΕΝΗ ΜΕΘΟΔΟΣ

Η ανομοιότητα στην εκπομπή των ρύπων δημιουργεί εμπόδιο στην επίτευξη συμφωνίας και συνεπώς απειλεί την βιωσιμότητα του περιβάλλοντος (Duro and Padilla, 2006). Η μέτρηση της συμπεριφοράς των χωρών-μελών σε σχέση με τους στόχους, με τους οποίους έχουν δεσμευτεί, επιτυγχάνεται πολύ καλά με την χρήση των μέτρων ανομοιότητας. Ο δείκτης Gini αποτελεί ένα μέτρο το οποίο μπορεί να χρησιμοποιηθεί για την μέτρηση τέτοιου είδους ανομοιότητας (Piketty, 2015). Επιπλέον η χρήση του μέτρου αυτού δίνει την δυνατότητα, να οπτικοποιηθεί το περίγραμμα της μεταβολής συμπεριφοράς των χωρών σε σχέση με τον στόχο. Αυτό μπορεί να αποδειχθεί ιδιαίτερα χρήσιμο, στη λήψη αποφάσεων σε παρόμοιες περιπτώσεις και να είναι ένα αξιόπιστο και ασφαλές εργαλείο στα χέρια των φορέων χάραξης πολιτικής για την αποφυγή τέτοιων κινδύνων στο μέλλον.

Στην παρούσα ενότητα θα προτείνουμε μια μέθοδο υπολογισμού της περιβαλλοντικής ανομοιότητας. Ο όρος περιβαλλοντική ανομοιότητα αναφέρεται στην ύπαρξη ανομοιομορφου τρόπου κατανομής των μετρήσεων, που αφορούν περιβαλλοντικά θέματα, ανάμεσα σε διαφορετικές κοινωνικές ομάδες. (Stephens, Wills and Walker, 2007) Στην συγκεκριμένη εργασία θεωρείται ότι το περιβαλλοντικό θέμα που θα εξετασθεί είναι οι εκπομπές αερίων ρύπων και οι κοινωνικές ομάδες τα κράτη. Ο υπολογισμός της περιβαλλοντικής ανομοιότητας μπορεί να φανεί χρήσιμος είτε για να ελεγχθεί η ανομοιότητα στην εκπομπή ρύπων μια συγκεκριμένη χρονική στιγμή, είτε κατά την διάρκεια περισσότερων ετών όπου μπορούμε να εξάγουμε συμπεράσματα για την πολιτική των χωρών κατά την διάρκεια ενός χρονικού διαστήματος. Εάν κατά την διάρκεια των ετών παρατηρηθεί ότι η ανομοιότητα είναι σταθερή παρά τους στόχους που έχουν τεθεί υπάρχουν σοβαρές ενδείξεις να εμφανίζεται το φαινόμενο του «λαθρεπιβάτη».

Δεν είναι πρώτη φορά που ο συγκεκριμένος δείκτης χρησιμοποιείται για την μέτρηση τέτοιου είδους ανομοιότητας (Heil and Wodon, 2000, Duro, 2012, Boyce, Zwickl and Ash, 2016). Η διαφορά της μεθόδου που προτείνεται συγκριτικά με τις υπάρχουσες βρίσκεται στον τρόπο με τον οποίο εφαρμόζεται ο δείκτης στα δεδομένα. Στις μέχρι τώρα έρευνες ο δείκτης Gini μετρούσε την ανομοιότητα των τιμών των εκπομπών ρύπων που προέρχονται από τις χώρες. Στην προτεινόμενη μέθοδο ο δείκτης εφαρμόζεται στην διαφορά που έχουν από τον προβλεπόμενο στόχο οι εκπομπές ρύπων. Πιο αναλυτικά η μέθοδος που προτείνεται είναι η ακόλουθη.

Έστω e_i οι εκπομπές αερίων ρύπων CO₂ σε κιλοτόνους (kt) της χώρας i μια δεδομένη χρονική στιγμή, e_{ti} οι εκπομπές αερίων ρύπων CO₂ σε kt που η χώρα i έχει ως στόχο να πετύχει σε ορισμένο χρονικό διάστημα (*target emissions*), e_{bi} οι εκπομπές αερίων ρύπων CO₂ σε kt που έχουν καταγραφεί για ένα συγκεκριμένο έτος το οποίο ορίστηκε ως έτος βάση της χώρας i από την συμφωνία του Kyoto (*base year emissions*) και p_i το προβλεπόμενο ποσοστό μείωσης ή αύξησης των ρύπων που αφορά την χώρα i . Το p_i θα έχει αρνητικό πρόσημο σε περίπτωση που το ποσοστό αφορά στόχο που έχει η i χώρα για μείωση των ρύπων και θετικό πρόσημο σε

περίπτωση που το ποσοστό αφορά στόχο που έχει η i χώρα για αύξηση των εκπεμπόμενων ρύπων. Προφανώς ισχύει,

$$e_{ii} = e_{bi} \cdot (1 + p_i). \quad (1)$$

Θεωρούμε,

$$y_i = e_i - e_{ii}, \quad (2)$$

να είναι η διαφορά των εκπομπών CO₂ της χρονιάς που μελετάμε και του έτους στόχου της χώρας i .

Για να εκφράσουμε ένα μέσο μέτρησης των πολιτικών που έχουν ακολουθήσει οι χώρες προς την συμμόρφωση σχετικά με τα επίπεδα εκπομπών, προτείνεται η μέτρηση της ανομοιοτήτας των διαφορών y_i . Θα εξεταστεί επομένως εάν υπάρχει ανομοιοτήτα στην απόσταση που έχουν από τον στόχο τους οι χώρες.

Για n χώρες που συμμετέχουν σε μια συμφωνία η τιμή του δείκτη Gini προκύπτει από τον τύπο (Sen, 1973),

$$I_G = [1 / (2n^2 \mu)] \cdot \sum_{i=1}^n \sum_{j=1}^n |y_i - y_j|, \quad (3)$$

όπου μ είναι η μέση τιμή των ποσοτήτων y_i . Είναι, $0 \leq I_G \leq 1$, όπου $I_G = 0$, όταν δεν υπάρχει καθόλου ανομοιοτήτα μεταξύ των δεδομένων και $I_G = 1$, όταν υπάρχει η μέγιστη ανομοιοτήτα.

Ο δείκτης Gini όπως προκύπτει από την σχέση (3), δεν αξιοποιεί τις αρνητικές τιμές. Κάτι τέτοιο είναι λογικό, αφού πρόκειται για έναν δείκτη χρηματοοικονομικών, ο οποίος μετράει την ανομοιοτήτα μεταξύ των εισοδημάτων. Θεωρούμε ότι, όταν $y_i = e_i - e_{ii} \leq 0$, τότε η χώρα έχει πετύχει ήδη από την χρονιά που μελετάμε την επιθυμητή μείωση. Οπότε έχει πετύχει τον στόχο της. Έτσι, θέτουμε, $y_i = 0$. Με αυτό τον τρόπο θέτουμε ένα όριο μείωσης εκπομπών y_i ίσο με την τιμή στόχο. Το ζητούμενο είναι να πετύχει η κάθε χώρα τον δικό της στόχο μείωσης ρύπων και όχι να συνεισφέρει στο σύνολο των εκπομπών.

Σε περίπτωση που θέλουμε να εξεταστεί κατά πόσο οι αρνητικές τιμές που έχουν οι διαφορές επηρεάζουν την τιμή του δείκτη Gini προτείνεται αυτός να υπολογιστεί από την ακόλουθη σχέση (Raffinetti, Siletti and Vernizzi, 2015),

$$G_p = \Delta / 2\mu_y^p. \quad (4)$$

Όπου τα Δ και μ_y^p δίνονται από τις ακόλουθες σχέσεις

$$\Delta = (1 / n^2) \sum_{i=1}^N \sum_{j=1}^N |Y_i - Y_j|, \quad (5)$$

$$\mu_y^p = (1 / 2) \Delta_p, \quad (6)$$

και το Δ_p προκύπτει ως εξής:

$$\Delta_p = 2 \left[(n-1) / n^2 \right] (T^+ + T^-), \quad (7)$$

$$\text{με } T^+ = \sum_{i=1}^n \max(0, Y_i) \text{ και } T^- = \left| \sum_{i=1}^n \min(0, Y_i) \right|.$$

3. ΕΦΑΡΜΟΓΗ ΤΗΣ ΜΕΘΟΔΟΥ

Στην συγκεκριμένη ενότητα θα εφαρμοστεί η μέθοδος που περιγράφηκε στην δευτέρα ενότητα για τις 39 χώρες που συμμετείχαν στην κύρωση του πρωτοκόλλου του Kyoto. Αρχικά θα υπολογιστεί η ανομοιότητα με βάση το δείκτη Gini εφαρμόζοντας τον τύπο που περιγράφεται από την σχέση (3) και έπειτα από τον τύπο που περιγράφεται από την σχέση (4). Η διαδικασία αυτή θα πραγματοποιηθεί για 8 έτη από το 2005 που τέθηκε σε ισχύ το πρωτόκολλο του Kyoto μέχρι και το 2012.

Πίνακας 2. Οι (Annex I) χώρες και οι αντίστοιχες τιμές εκπομπών τους στο έτος που θεωρήθηκε ως βάση

<i>i</i>	Χώρες	$I + p_i$	e_{bi}	<i>i</i>	Χώρες	$I + p_i$	e_{bi}
1	Αυστραλία	1,08	277802,53	21	Λίχτενσταϊν	0,92	203,06
2	Αυστρία	0,92	61932,64	22	Λιθουανία	0,92	36168,80
3	Λευκορωσία	0,92	103500,00	23	Λουξεμβούργο	0,92	12219,20
4	Βέλγιο	0,92	118684,50	24	Μονακό	0,92	105,37
5	Βουλγαρία	0,92	98815,11	25	Ολλανδία	0,92	159389,50
6	Καναδάς	0,94	457534,00	26	Νέα Ζηλανδία	1,00	25462,57
7	Κροατία	0,95	23080,45	27	Νορβηγία	1,01	34766,97
8	Τσεχία	0,92	163864,20	28	Πολωνία	0,94	469143,82
9	Δανία	0,92	53342,45	29	Πορτογαλία	0,92	40261,95
10	Εσθονία	0,92	37677,86	30	Ρουμανία	0,92	192407,79
11	Φινλανδία	0,92	56767,66	31	Ρωσία	1,00	2500352,09
12	Γαλλία	0,92	392627,00	32	Σλοβακία	0,92	60221,70
13	Γερμανία	0,92	1032776,20	33	Σλοβενία	0,92	16281,84
14	Ελλάδα	0,92	84313,57	34	Ισπανία	0,92	228511,44
15	Ουγγαρία	0,94	85795,50	35	Σουηδία	0,92	56301,08
16	Ισλανδία	1,10	2158,64	36	Ελβετία	0,92	44553,30
17	Ιρλανδία	0,92	32559,50	37	Ουκρανία	1,00	714310,07
18	Ιταλία	0,92	434781,95	38	Ηνωμένο Βασίλειο	0,92	590319,32
19	Ιαπωνία	0,94	1144129,51	39	Η. Π. Α	0,93	5100000,00
20	Λετονία	0,92	18622,93				

Με τον τρόπο αυτό, θα υπολογιστεί η ανομοιότητα για κάθε ένα έτος και επιπλέον θα ελεγχθεί εάν η επιλογή του τύπου οφείλεται για την διαφοροποίηση στις τιμές μέτρησης της ανομοιότητας και κατ' επέκταση και στα συμπεράσματα που προκύπτουν από αυτήν.

Στον Πίνακα 2 παρουσιάζονται οι χώρες που συμμετείχαν στην κύρωση του πρωτοκόλλου του Kyoto. Οι χώρες που συμμετείχαν στην κύρωση του πρωτοκόλλου αναφέρονται σε αυτό ως "Annex I" χώρες. Στις δύο επόμενες στήλες του Πίνακα 2 καταγράφονται για κάθε χώρα *i* αντίστοιχα οι ποσότητες $(I + p_i)$ και e_{bi} με βάση τις οποίες υπολογίζεται για κάθε χώρα η ποσότητα-στόχος e_{ii} (βλ. σχέση (1)).

Για κάθε μία από τις χώρες του Πίνακα 2 υπολογίστηκαν οι εκπομπές που έχει ως στόχο να πετύχει η κάθε χώρα e_{it} , χρησιμοποιώντας την σχέση (1). Στην συνέχεια χρησιμοποιώντας τα διαθέσιμα αριθμητικά δεδομένα 8 ετών (2005-2012), που αφορούν τις εκπομπές CO₂, όπως είναι καταγεγραμμένα από τον ΟΗΕ [1], υπολογίστηκε για κάθε ένα από αυτά τα έτη η ανομοιότητα των διαφορών που έχουν οι χώρες

Τα αποτελέσματα αυτά συνοψίζονται στον Πίνακα 3. Παρατηρούμε ότι υπάρχει πολύ μεγάλη ανομοιότητα, στον τρόπο που μειώνουν τις εκπομπές τους οι χώρες, διότι σε όλα τα έτη η τιμή του δείκτη είναι πολύ κοντά στην τιμή 1. Όσο πλησιάζουμε προς το 2012 που αποτελεί και το έτος-στόχο, ενώ θα έπρεπε να μειώνεται η ανομοιότητα, αφού όλες οι χώρες θα έπρεπε να μεριμνήσουν ώστε να έχουν μικρές διαφορές από τον στόχο, εντούτοις η ανομοιότητα αυξάνεται.

Πίνακας 3. Υπολογισμός της ανομοιότητας για τα έτη 2005 έως 2012.

Έτος	2005	2006	2007	2008	2009	2010	2011	2012
Τιμή ανομοιότητας	0,87	0,87	0,88	0,88	0,88	0,88	0,90	0,90

Παρατηρώντας τα αποτελέσματα της έρευνας μας σε μια προσπάθεια να εξηγηθεί αυτή η συμπεριφορά διαπιστώθηκε ότι το 1/3 των χωρών έχουν πετύχει τους στόχους που είχαν θέσει, ήδη από το 2005. Το 2005 είναι το έτος όπου αρχίζει η εφαρμογή του πρωτοκόλλου του Kyoto. Οι υπόλοιπες χώρες συνεχίζουν να δείχνουν μια "εγωιστική" συμπεριφορά, αυξάνοντας τις εκπομπές τους και αδιαφορώντας για τους στόχους που έχουν συμφωνήσει. Μέχρι το 2012 οι 22/39 χώρες έχουν καταφέρει να πετύχουν τους στόχους τους. Οι χώρες που πέτυχαν τον τελικό στόχο είναι: Λευκορωσία, Βέλγιο, Βουλγαρία, Κροατία, Φινλανδία, Λουξεμβούργο, Τσεχία, Εσθονία, Γερμανία, Δανία Ουγγαρία, Ιταλία, Λετονία, Λιθουανία, Μονακό, Πολωνία, Ρουμανία, Ρωσία, Σλοβακία, Σουηδία, Ουκρανία και Μεγάλη Βρετανία. Η Ελλάδα βρισκόταν το 2012, 12903,92 kt μακριά από τον στόχο της. Οι χώρες με την μεγαλύτερη απόσταση από τον στόχο ήταν το 2012: ο Καναδάς (120464,64 kt), η Ιαπωνία (200128,96 kt) και οι ΗΠΑ (632002.9 kt).

Στη συνέχεια υπολογίστηκε η ανομοιότητα στην περίπτωση που συμπεριλαμβάνονται και οι αρνητικές τιμές των διαφορών εφαρμόζοντας τη σχέση (4). Τα αποτελέσματα των υπολογισμών αυτών παρουσιάζονται στον Πίνακα 4.

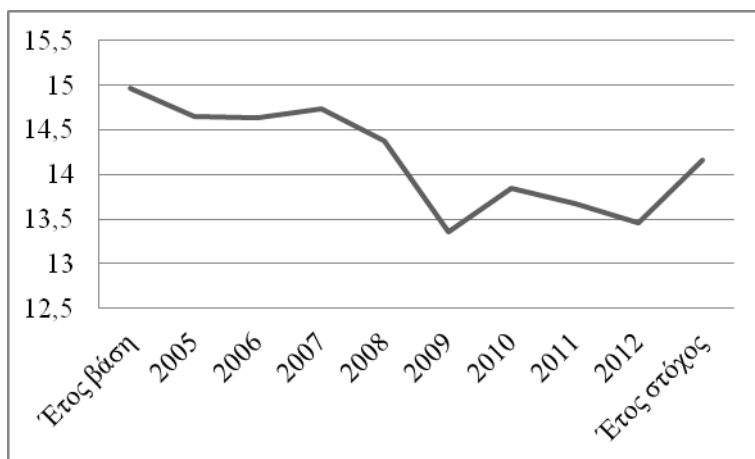
Πίνακας 4. Υπολογισμός της ανομοιότητας για τα έτη 2005 έως 2012 συμπεριλαμβανομένων και των αρνητικών τιμών των διαφορών.

Έτος	2005	2006	2007	2008	2009	2010	2011	2012
Τιμή ανομοιότητας	0,91	0,90	0,91	0,91	0,90	0,90	0,90	0,90

Στην ανάλυση αυτή συνεχίζουμε να λαμβάνουμε υπόψη, τις χώρες που υπερκαλύπτουν τον στόχο τους και θεωρούμε ότι συνεχίζουν να συνεισφέρουν στο σύνολο των εκπομπών της ομάδας των 39 χωρών. Παρατηρούμε ακολούθως μεγάλη

ανομοιότητα, όπως και στην πρώτη περίπτωση. Δεδομένου ότι οι συνολικές εκπομπές για τις 39 χώρες έχουν μειωθεί όπως προκύπτει από το σχήμα 1ν πίνακα 5 διαπιστώνεται ότι οι χώρες που δρουν ως "λαθρεπιβάτες εκμεταλλεύονται πολύ περισσότερο την μείωση των συνολικών εκπομπών και συνεχίζουν την "ρυπογόνο" δράση τους.

Σχήμα 1 Πίνακας 5. Συνολικές τιμές εκπομπών CO₂ σε Gt κατά την διάρκεια των ετών 2005-2012



4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η μέτρηση της περιβαλλοντικής ανομοιότητας, αναφορικά με την εκπομπή αερίων ρύπων, αποτελεί θέμα μείζονος σημασίας διότι συνδέεται άμεσα με τις επιπτώσεις που προέρχονται από την κλιματική αλλαγή. Οι επιπτώσεις της κλιματικής αλλαγής δεν έχουν μόνο επίδραση στο φυσικό περιβάλλον αλλά επιβαρύνουν και τους προϋπολογισμούς των κρατών. Η μέθοδος που προτείνουμε μπορεί να μετρήσει την περιβαλλοντική ανομοιότητα που σχετίζεται με την εκπομπή αερίων ρύπων για κάθε έτος. Συγκρίνοντας την προτεινόμενη μέθοδο με την μέθοδο που χρησιμοποιείται ως τώρα συμπεραίνουμε διαφοροποίηση στον τρόπο που ελέγχεται η ανομοιότητα των εκπομπών ρύπων. Οι υπάρχουσες μέθοδοι υπολογίζουν την ανομοιότητα των τιμών των εκπομπών ρύπων για μία συγκεκριμένη χρονική στιγμή. Λαμβάνοντας υπόψη μας όμως ότι κάθε χώρα έχει να πετύχει διαφορετικό στόχο, λόγω των διαφορετικών δυνατοτήτων και αναγκών που έχει, είναι αναμενόμενο να υπάρχει διαφοροποίηση στην εκπομπή των ρύπων. Για τον λόγο αυτό προτάθηκε η μέτρηση της ανομοιότητας στην διαφορά εκπομπής ρύπων που έχουν οι χώρες από το έτος-στόχο τους. Με τον τρόπο αυτό ελέγχεται εάν υπάρχει διαφορά στη μείωση των εκπομπών των ρύπων γεγονός που συνεπάγεται την έλλειψη συμμόρφωσης με την συνθήκη. Διαπιστώθηκε ότι παρόλο που οι εκπομπές ρύπων τα τελευταία χρόνια έχουν μειωθεί συνολικά η ανομοιότητα παρέμεινε σταθερή. Η τελευταία διαπίστωση αναδεικνύει το πρόβλημα «λαθρεπιβίβασης» του πρωτοκόλλου του Kyoto. Η προτεινόμενη μέθοδος φαίνεται να είναι αποδοτική στον

εντοπισμό τέτοιου είδους προβλημάτων. Επιβεβαιώθηκαν οι ενδείξεις αποτυχίας της συμφωνίας του Kyoto ότι κάποιες χώρες όχι μόνο δεν μείωσαν τις εκπομπές τους αλλά εκμεταλλεύτηκαν τον ζήλο άλλων χωρών με αποτέλεσμα να εμφανίζεται σταθερή και μεγάλη ανομοιότητα. Τα συμπεράσματα που εξήχθησαν είναι τα ίδια είτε με χρησιμοποίηση μόνο των θετικών τιμών, είτε και με χρήση θετικών ή αρνητικών τιμών, που αφορούν στο σύνολο των διαφορών.

ABSTRACT

The increase of the gas pollutant emissions into the atmosphere is considered as a major worldwide issue. Along with the systematic attempt to record the phenomenon, international organizations try to ratify agreements that will determine the maximum allowable emission limit per country. In this study, a method for calculating the environmental dissimilarity will be developed. This method will examine not only the measurement of dissimilarity but it will also contribute to the identification of the 'free rider' problem, which is a common cause of failure of interstate agreements. In the first section the problem to be tested is described, then the correspondent method is proposed and finally real data are used to examine the efficiency of the method.

Keywords: environmental diversity, 'free rider' problem, climate change, Gini index, Kyoto protocol.

ΑΝΑΦΟΡΕΣ

- d'Aspremont C., Jacquemin A., Gabszewicz J. J., and Weymark J. A. (1983). On the stability of collusive price leadership. *Canadian Journal of economics*, 17-25.
- Barrett S. (1998). Political economy of the Kyoto Protocol. *Oxford review of economic policy*, **14(4)**, 20-39.
- Boyce, J. K., Zwickl, K., and Ash, M. (2016). Measuring environmental inequality. *Ecological Economics*, **124**, 114-123.
- Duro J. A., and Padilla E. (2006). International inequalities in per capita CO₂ emissions: a decomposition methodology by Kaya factors. *Energy Economics*, **28(2)**, 170-187.
- Duro J. A. (2012). On the automatic application of inequality indexes in the analysis of the international distribution of environmental indicators. *Ecological Economics*, **76**, 1-7.
- Finus M. (2008). Game theoretic research on the design of international environmental agreements: Insights, critical remarks, and future challenges. *International Review of Environmental and Resource Economics*, **2(1)**, 29-67.
- Heil M. T., Wodon Q. T. (2000). Future inequality in CO₂ emissions and the impact of abatement proposals. *Environmental and Resource Economics*, **17(2)**, 163-181.
- Heugues M. (2014). International environmental cooperation: a new eye on the greenhouse gas emissions' control. *Annals of Operations Research*, **220**, 239-262.

- Nordhaus W. D. and Boyer J. G. (1999). Requiem for Kyoto: an economic analysis of the Kyoto Protocol. *The Energy Journal*, 93-130.
- Piketty T. (2015). Carbon and inequality: from Kyoto to Paris. Trends in the global inequality of carbon emissions (1998-2013) and prospects for an equitable adaptation fund. Paris School of Economics.
- Raffinetti E., Siletti E., and Vernizzi A. (2015). On the Gini coefficient normalization when attributes with negative values are considered. *Statistical Methods & Applications*, **24(3)**, 507-521.
- Sen A. (1973). On economic inequality. New York: Oxford University Press.
- Stephens C., Wills R., and Walker G. (2007). Addressing environmental inequalities: cumulative environmental impacts. Environment Agency.
- Σαρτζετάκης Ε. και Παλλάκης Σ. (2011). Κλιματική αλλαγή: Ένα κοινωνικό δίλημμα συλλογικού και υψηλού κινδύνου. Κοπεγχάγη 2009. Το περιβάλλον στη δίνη μιας παγκόσμιας κρίσης. Εκδόσεις Ι. Σιδέρης.
- Ιστοσελίδες:
[1] mdgs.un.org/unsd/mdg

ΙΩΑΝΝΗΣ ΠΑΠΑΔΑΚΗΣ (JEAN PARADAKIS) (1903-1997)

Στρατής Κουνιάς
Πανεπιστήμιο Αθηνών
skounias@math.uoa.gr

ΠΕΡΙΛΗΨΗ

Η παρούσα εργασία αναφέρεται στο έργο και τη συμβουλή του γεωπόνου- μηχανικού Ιωάννη Παπαδάκη, ο οποίος από το 1930 ανέπτυξε μέθοδο βελτίωσης της εκτίμησης των στατιστικών σφαλμάτων, στα γεωργικά πειράματα που έκανε για την καλυτέρευση των φυτών.

Λέξεις κλειδιά: Στατιστικά σφάλματα, εκτίμηση παραμέτρων, εξαρτημένες παρατηρήσεις.

1. ΒΙΟΓΡΑΦΙΚΑ ΣΤΟΙΧΕΙΑ

Ο Ιωάννης Παπαδάκης γεννήθηκε το 1903 στη Χώρα της Νάξου, πρώτευσε στο Γυμνάσιο της Σύρου και το 1919, μετά από εξετάσεις του υπουργείου Γεωργίας, έλαβε κρατική υποτροφία. Σπούδασε γεωπονία στο Ζαμπλού (Gembloux) του Βελγίου, το 1922 αποφοίτησε πρώτος. Συμπλήρωσε τις σπουδές του στη Γαλλία με ειδίκευση στη βελτίωση των φυτών, επέστρεψε στην Ελλάδα το 1923 και διορίστηκε στο Υπουργείο Γεωργίας.



Ο Ιωάννης Παπαδάκης

Στην αρχή αποσπάσθηκε ως επιμελητής στη νεοσύστατη Α.Γ.Σ.Α. και συνεργάσθηκε με τον Σπ. Χασιώτη διευθυντή της σχολής, ο οποίος τον ενημέρωσε για τα προβλήματα της γεωργίας στην Ελλάδα, η παραγωγή σίτου τότε κάλυπτε το 30% των αναγκών της χώρας μας.

Με πρωτοβουλία του τότε πρωθυπουργού Αλέξανδρου Παπαναστασίου ιδρύθηκε το 1925, με έδρα τη Λάρισα, το πρώτο Ερευνητικό Γεωπονικό Ίδρυμα της χώρας, ο Σταθμός Καλυτέρευσης των Φυτών υπό τη διεύθυνση του Ι. Παπαδάκη 1925-1927. Το 1927 ο Σταθμός μεταφέρθηκε στη Θεσσαλονίκη και μετονομάστηκε σε Ινστιτούτο Καλυτέρευσης Φυτών (Σήμερα Ινστιτούτο Σιτηρών), με διευθυντή τον Ι. Παπαδάκη (1927-1946).

Βραβεύτηκε από την Ακαδημία Αθηνών το 1927 και δίδασκε ως υφηγητής τα μαθήματα καλυτέρευσης των φυτών στη νεοσύστατη ΦΜΣ Πανεπιστημίου Θεσσαλονίκης, που τότε περιελάμβανε και τμήμα Γεωπονίας.

Με τη μαζική εισαγωγή από το Μεξικό ποικιλιών σίτου υψηλής στρεμματικής απόδοσης και δημιουργίας νέων ποικιλιών, αυξανόταν διαρκώς η σιτοπαραγωγή, το 1930-1940 τριπλασιάστηκε η παραγωγή σίτου και το 1957 η Ελλάδα έγινε σιτο-αυτάρκης.

Η αύξηση της παραγωγής οφείλεται και στη θέσπιση του μέτρου της κρατικής συγκέντρωσης σιταριού με καλές τιμές που εξασφάλιζε το εισόδημα των αγροτών.

Λίγους μήνες μετά τη Μικρασιατική Καταστροφή, ο Ιωάννης Παπαδάκης, εγκαθιστά τους πρώτους πειραματικούς αγρούς αξιολόγησης εγχώριων και ξένων ποικιλιών σιταριών σε τέσσερα σημεία της χώρας.

Από την αρχή της δημιουργίας του σταθμού, σκοπός ήταν η βελτίωση και η δημιουργία νέων ποικιλιών, ανθεκτικότερων στις ασθένειες, με υψηλότερες στρεμματικές αποδόσεις, δηλαδή μεγαλύτερη και ποιοτικότερη παραγωγή σιτηρών. Ασχολήθηκε κυρίως με την έρευνα σε σιτάρι μαλακό και σκληρό, κριθάρι, βρόμη, σίκαλη, βίκο, καλαμπόκι και ρύζι. Στα 90 χρόνια λειτουργίας δημιουργήθηκαν πάνω από 100 νέες ποικιλίες σιτηρών και υβρίδια καλαμποκιού, ενώ 69 απ' αυτές είναι εγγεγραμμένες στον Εθνικό Κατάλογο Ποικιλιών.

Παράλληλα δημοσίευσε το βιβλίο *Ecology Agricole* (1938) και στατιστικές εργασίες, η πλέον γνωστή είναι *Methode Statistique pour des experiences en champ*. *Thessaloniki Pl. Breed Ins. Bull.* 23, 1937.

Στην εργασία αυτή μείωσε το σφάλμα της εκτίμησης για την επίδραση των αγωγών (improvement of the estimation of treatment effects), χρησιμοποιώντας την πληροφορία που δίνουν οι αποδόσεις σε γειτονικά αγροτεμάχια. Χρησιμοποίησαν αρκετοί ερευνητές τη μέθοδο Παπαδάκη, [M. Atkinson A.C. (1969)], [Cox D. (1952)], [Bartlett 1938, 1978].

Ο «Ειδικός Σταθμός Καλυτερεύσεως των Φυτών», εξελίχθηκε σε Ινστιτούτο Σιτηρών Θέρμης Θεσσαλονίκης και συνέβαλε στο να τραφεί ο πληθυσμός της Ελλάδας σε δύσκολες εποχές.

Απεσταλμένος της Ελλάδας στον ΟΗΕ το 1946 έλαβε τηλεγράφημα από την Ελληνική κυβέρνηση με πρωθυπουργό τον Παναγή Τσαλδάρη που έγραφε: «Απολύεσαι από τη θέση σου και να επιστρέψεις στην Ελλάδα».

Με την αποχώρηση των κατακτητών από την Ελλάδα τον Οκτώβρη 1944, ήλθε η εξόριστη Ελληνική κυβέρνηση με πρωθυπουργό τον Γεώργιο Παπανδρέου, τον οποίο στήριζε η Βρετανική κυβέρνηση. Μετά τη διαφωνία του ΕΑΜ με την κυβέρνηση Γ. Παπανδρέου στο πως θα δημιουργηθεί ο νέος Ελληνικός Στρατός, ακολούθησαν τα Δεκεμβριανά, υπογράφηκε η συμφωνία της Βάρκιζας (12/2/1945) και οι συμπαθούντες του ΕΑΜ τέθηκαν υπό διωγμό.

Ένας από αυτούς ήταν και ο Ι. Παπαδάκης ως «εμφορούμενος υπό αντεθνικών φρονημάτων», εκδιώχθηκε από το Ινστιτούτο και αντί να επιστρέψει στην Ελλάδα πήγε στο Περού και μετά στην Αργεντινή όπου εργάστηκε στο Υπουργείο Γεωργίας. Του ανατέθηκε και εκπόνησε το γεωργικό οικολογικό χάρτη της χώρας. Το 1953 προσκλήθηκε από τον ΟΗΕ για να συμβάλει στη γεωργική ανάπτυξη του Περού και της Βολιβίας (1954-57).

Ο FAO (Food and Agricultural Organization of the United Nations) του ανέθεσε να κάνει τους οικολογικούς χάρτες στις χώρες της Δυτικής Αφρικής, της Ινδίας, της Χιλής, επίσης τον κάλεσε στη Ρώμη για να συμβάλει στην μελέτη της παραγωγής τροφίμων στις αναπτυσσόμενες χώρες (1964-1968).

Έγινε παγκόσμια γνωστός για τους οικολογικούς του χάρτες. Το 1972 εκλέχτηκε καθηγητής στο Πανεπιστήμιο του Μπουένος Άιρες (1972-1978), αν και είχε ξεπεράσει το όριο ηλικίας των 65 ετών.

Το 1974 εκλέχτηκε αντεπιστέλλον μέλος και το 1982 εκλέχτηκε τακτικό μέλος της Ακαδημίας Αθηνών.

Στις 6/6/1986 έγινε στην αίθουσα τελετών του ΕΣΙ η ανακήρυξη του σε επίτιμο μέλος του ΕΣΙ. Ο καθηγητής Θεόφιλος Κάκουλλος έκανε την παρουσίαση και ο τιμώμενος ανέπτυξε την εμπειρία του και τις σκέψεις του. Μετείχε στο Συνέδριο του ΕΣΙ στη Θεσσαλονίκη το 1987, όπου και τον τιμήσαμε.

Στις 14/4/1992 αναγορεύτηκε επίτιμος καθηγητής του Γεωργικού Πανεπιστημίου Αθηνών.

Είχε συγγράψει πολλά βιβλία, μονογραφίες και άρθρα σχετικά με την οικολογική γεωργία και με τα πειράματά του ανέπτυξε τη μέθοδο βελτίωσης των σφαλμάτων, γνωστή ως μέθοδος Παπαδάκη.

Απεβίωσε στην Αργεντινή το 1997 και ενταφιάστηκε εκεί.

2. Η ΜΕΘΟΔΟΣ ΠΑΠΑΔΑΚΗ

Στο απλό γραμμικό μοντέλο $y_i = \beta_0 + \beta_1 x_i + e_i$ $i = 1, \dots, N$, συνήθως θεωρούμε ότι οι παρατηρήσεις έχουν ανεξάρτητα σφάλματα με σταθερή διασπορά

σ^2 . Στην περίπτωση αυτή η εκτίμηση των παραμέτρων β_0, β_1 με τη μέθοδο ελαχίστων τετραγώνων and του σ^2 μας δίνει,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N - 2}$$

$y_i - \hat{y}_i = \hat{e}_i$ είναι τα κατάλοιπα ή εκτιμημένα σφάλματα.

Οι εκτιμήτριες των β_0, β_1 , που είναι γραμμικές συναρτήσεις των παρατηρήσεων $y_i, i = 1, \dots, N$, έχουν την ελάχιστη διασπορά από όλες τις αμερόληπτες εκτιμήτριες αυτών των παραμέτρων (Gauss-Markov) .

Αν γνωρίζουμε το νόμο της εξάρτησης των σφαλμάτων εκτιμούμε τις παραμέτρους μεγιστοποιώντας την πιθανοφάνεια των παρατηρήσεων.

Τι κάνουμε όταν οι παρατηρήσεις είναι εξαρτημένες και δεν γνωρίζουμε ποια εξάρτηση έχουν;

Η μέθοδος Παπαδάκη δίνει τις εκτιμήσεις των παραμέτρων σε πειράματα που οι γειτονικές παρατηρήσεις είναι εξαρτημένες και έχουμε επαναλαμβανόμενες παρατηρήσεις σε κάθε σημείο.

Πρόβλημα: Έχουμε p αγωγές (treatments) και $N = mp$ πειραματικές μονάδες, αυτές μπορεί να είναι τεμάχια αγροκτημάτων. Κάθε αγωγή εφαρμόζεται σε m πειραματικές μονάδες και οι παρατηρημένες τιμές είναι $y_1, \dots, y_N$.

Αν στη μονάδα i εφαρμόζεται η αγωγή s και $\alpha_s, s = 1, \dots, p$ είναι η επίδραση (effect) της αγωγής s , το μοντέλο είναι,

$$y_i = \alpha_s + u_i, \quad i = 1, \dots, N, \quad s = 1, \dots, p \quad (1)$$

όπου u_i είναι το σφάλμα, υποθέτουμε ότι $E(u_i) = 0, \text{var}(u_i) = \sigma^2$.

$$\text{Ορίζουμε:} \quad y'_i = y_i - (1/m) \sum_{\{i\}=s} y_i, \quad i = 1, \dots, N \quad (2)$$

Δηλαδή παίρνουμε όλες τις μονάδες στις οποίες εφαρμόστηκε η ίδια αγωγή και το σύμβολο

$\{i\}=s$ εννοεί όλες τις παρατηρήσεις i όπου έχει εφαρμοστεί η αγωγή s . Δηλαδή διορθώνουμε την παρατήρηση y_i .

Παρατηρούμε ότι το y_i' δεν περιέχει το a_s $s = 1, \dots, p$ αλλά περιέχει μόνο τα σφάλματα u_i $\{i\} = s$ και $\sum_{i=1}^N y_i' = 0$, $E(y_i') = 0$, $i = 1, \dots, N$.

$$\text{Ορίζουμε:} \quad x_i = (y_{i-1}' + y_{i+1}') / 2 \quad i = 1, \dots, N \quad (3)$$

Αν οι μονάδες σχηματίζουν περιφέρεια κύκλου θα είναι $y_0 = y_N, y_{N+1} = y_1$, αν οι μονάδες είναι σε ευθεία γραμμή, τότε είναι $x_1 = y_2', x_N = y_{N-1}'$.

$$\text{Ορίζουμε:} \quad \hat{\beta} = \left(\sum_{i=1}^N x_i y_i' \right) / \left(\sum_{i=1}^N x_i^2 \right) \quad (4)$$

Παρατηρούμε ότι το $\hat{\beta}$ είναι συνάρτηση μόνο των σφαλμάτων.

Συμβολίζουμε με α_s^P την εκτιμήτρια Παπαδάκη για την επίδραση της αγωγής s , τότε ορίζουμε:

$$\alpha_s^P = \frac{1}{m} \left(\sum_{\{i\}=s} y_i - \frac{\hat{\beta}}{2} \sum_{\{i\}=s} \left(y_{i\pm 1} - \frac{1}{m} \sum_{i\pm 1=s} y_i \right) \right) \quad (5)$$

Οπότε υπολογίζουμε τις εκτιμήτριες $(\alpha_s^P - \alpha_t^P)$ των παραμέτρων α_s, α_t .

Η σχέση (5) είναι εκτίμηση της παραμέτρου α_s με τη μέθοδο Παπαδάκη.

3. ΣΧΟΛΙΑ ΕΡΕΥΝΗΤΩΝ ΓΙΑ ΤΗ ΜΕΘΟΔΟ ΠΑΠΑΔΑΚΗ.

Ο Αυστραλός P.M. Williams (1952) μελέτησε το πρόβλημα όταν p αγωγές $A_1, \dots, A_p$ τοποθετούνται στη σειρά (row designs) και τον ενδιέφερε να βρει το βέλτιστο σχεδιασμό ώστε να εκτιμά την επίδραση κάθε αγωγής a_s και τις αντιθέσεις $(\alpha_s - \alpha_t)$.

Τα σφάλματα u_i είναι εξαρτημένα και ακολουθούν ένα αυτοπαλινδρομούμενο μοντέλο πρώτης τάξης (AR(1)),

$$u_i = \rho u_{i-1} + e_i, \quad i = 1, \dots, N$$

όπου $|\rho| < 1$ και τα e_i είναι ανεξάρτητα με κατανομή $N(0, \sigma^2)$.

Εξέτασε δύο σχεδιασμούς $\Pi(a), \Pi(b)$ για p αγωγές $A_1, \dots, A_p$.

$p = 3: II(a) N = 7 \quad A_1 A_2 A_3 A_1 A_3 A_2 A_1 \quad II(b) N = 10 \quad A_1 A_1 A_3 A_2 A_2 A_1 A_2 A_3 A_3 A_1$
 $p = 4, II(a) N = 13: A_2 A_1 A_2 A_3 A_4 A_2 A_3 A_1 A_4 A_3 A_1 A_4 A_2$

$II(b) N = 17 \quad A_1 A_1 A_2 A_3 A_4 A_4 A_1 A_3 A_2 A_2 A_4 A_1 A_3 A_3 A_4 A_2 A_1$

Στο σχεδιασμό $II(a)$ κάθε αγωγή είναι δίπλα σε κάθε άλλη αγωγή ίσο πλήθος φορών, στο σχεδιασμό $II(b)$ κάθε ζεύγος διαφορετικών αγωγών εμφανίζεται διπλάσιες φορές από κάθε ζεύγος επαναλαμβανόμενων αγωγών.

Ο Williams (1952) εκτίμησε με τη μέθοδο της πιθανοφάνειας τις $(\hat{\alpha}_i - \hat{\alpha}_j)$ και διαπίστωσε, αλλά δεν το απέδειξε, ότι αν $0 < \rho < 1$ ο σχεδιασμός $II(a)$ είναι ο βέλτιστος. Οι Cox (1952) και Atkinson (1969) απέδειξαν ότι ο σχεδιασμός $II(a)$ είναι A βέλτιστος και οι Kunert and Martin (1967) έδειξαν ότι είναι και D βέλτιστος.

Ο Atkinson (1969) εφάρμοσε τη μέθοδο Παπαδάκη για την εκτίμηση των α_i , $(\alpha_i - \alpha_j)$.

Διαπίστωσε ότι οι εκτιμήτριες των Williamson και Παπαδάκη:

The two estimators for $(\hat{\alpha}_i - \hat{\alpha}_j)$ are nearly identical. The variances of the two estimators are thus nearly equal, that of Papadakis estimator being slightly the larger. Both variances were less than half the variance of the crude estimate.

Να σημειώσουμε ότι η Μέθοδος Παπαδάκη είναι απλούστερη και δεν αναφέρεται στον τρόπο εξάρτησης των σφαλμάτων. Επίσης ο Bartlett (1938) αναφέρει:

The method suggested by Papadakis of using covariance with the yields of neighboring plots to reduce the error of replicated field experiments is illustrated on two large-scale cotton experiments. In a discussion on the validity and value of the method, it is concluded that for such experiments, where the number of plots per block is large, the method should be approximately valid and sometimes useful.

Spatial formulations were first proposed in the 1920's by the distinguished Greek agronomist, J. S. Papadakis, but received scant attention from statisticians until the 1980s, apart from Bartlett (1938, 1978). However, the past twenty years has seen a flurry of activity and now overtly spatial analyses are used quite widely; for example, in some 5000 field experiments annually in Australia.

Η εργασία του [Bartlett 1978] συζητήθηκε σε ειδική συνεδρίαση της London Statistical Society, έλαβαν μέρος όλοι σχεδόν οι καθηγητές Στατιστικής της Βρετανίας και όλοι σχολίασαν τη συμβολή της μεθόδου Παπαδάκη. Ο καθηγητής D.

R. Cox είπε: Papadakis's method is simple, flexible and general και αιτιολόγησε την άποψή του.

ABSTRACT

This paper refers to the work and contribution of the agricultural engineer Jean Papadakis who from 1930 developed a method, that improved the estimation of the statistical errors, in his field experiments.

ΑΝΑΦΟΡΕΣ

- Atkinson A.C. (1969). The use of residuals as a concomitant variable. *Biometrika*, 56, 1, 33-41.
- Bartlett M. S.(1938). The approximate recovery of information from replicated field experiments with large blocks. *The Journal of Agricultural Science*, 28, pp 418-427.
- Bartlett, M. S. (1978).Nearest neighbor models in the analysis of field experiments. *J. Roy. Statist. Society B*, 40, 147-158.
- Cox, D. R.(1952). Some recent work for systematic experimental designs. *J. Roy. Statist. Society, Ser. B* 14, 211-229.
- Kunert, J and Martin R. J. (1987). On the optimality of Williams II(a) designs. *The Annals of Statistics*, Vol. 15, No. 4, 1604-1608.
- Papadakis, J. S. (1937). Methode statistique pour des experiences sur champ. Institut d'Amélioration des Plantes a Thessaloniki (Grece), *Bulletin Scientifique*, No. 23.
- Papadakis, J. S. (1970). *Agricultural Research: Principles, Methodology, Suggestions*. Edicion Argentina, Buenos Aires.
- Papadakis, J. S. (1984). *Advances in the analysis of field experiments*. Proceedings of the Academy of Athens, 59, 326-342.
- Papadakis, J. S. (1931). Some considerations on the technique of field experiments Bull. Assoc.Intern. Select. plants, 4,1.
- Papadakis, J. S. (1935). Varieties experiments in pots. Thessaloniki Pl. Breed Ins. Bull. 20.
- Williams, R. M. (1952). Experimental designs for serially correlated observations. *Biometrika* 39, 151-167.
- Καραμάνος Α. (1982) Αναγόρευση του Ακαδημαϊκούκ. Ιωάννη Παπαδάκη ως επίτιμου καθηγητή του Γ.Π.Α.
- Πουλοβασίλης Αλέξανδρος (2013). *Η Ανωτάτη Γεωπονική Σχολή Αθηνών στα πρώτα μεταπολεμικά χρόνια. σ. 100-103*, Εκδόσεις ΕΜΒΡΥΟ, 2013, ISBN 978-960-8002-77-7

ΣΥΓΚΡΙΤΙΚΗ ΑΝΑΛΥΣΗ ΜΟΝΤΕΛΩΝ ΘΝΗΣΙΜΟΤΗΤΑΣ ΓΙΑ ΤΟΝ ΕΛΛΗΝΙΚΟ ΠΛΗΘΥΣΜΟ

Α. Μποζίκας, Γ. Πιτσέλης

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς
{bozikas, pitselis}@unipi.gr

ΠΕΡΙΛΗΨΗ

Τις τελευταίες δεκαετίες έχει παρατηρηθεί σημαντική αύξηση του προσδόκιμου ζωής στις περισσότερες αναπτυγμένες χώρες σε όλο τον κόσμο, ανάμεσά τους και στην Ελλάδα. Το άρθρο αυτό έχει ως στόχο τη διερεύνηση των σημαντικότερων μοντέλων θνησιμότητας στη βιβλιογραφία με στόχο την εξαγωγή πιο έγκυρων και αποτελεσματικών προβλέψεων θνησιμότητας που αφορούν τις ασφαλιστικές εταιρίες και τα συνταξιοδοτικά ταμεία. Πιο συγκεκριμένα, συγκρίνουμε τα μοντέλα των Lee and Carter (1992), Renshaw and Haberman (2006) και Cairns et al. (2006) ως προς την προσαρμογή τους στα δεδομένα της Ελλάδας, αλλά και την προβλεπτική τους ικανότητα. Ο έλεγχος καλής προσαρμογής των μοντέλων στα ελληνικά δεδομένα αξιολογείται με τα κριτήρια πληροφορίας AIC και BIC, ενώ η ακρίβεια των βραχυπρόθεσμων προβλέψεων εκτιμάται από τα μέτρα υπολογισμού σφάλματος MAE, RMSE και MAPE.

Λέξεις Κλειδιά: Στοχαστικά Μοντέλα, Προβολή Θνησιμότητας, Προσδόκιμο Ζωής, Ελλάδα.

1. Εισαγωγή

Κατά τις προηγούμενες δεκαετίες παρατηρήθηκε σημαντική αύξηση στο προσδόκιμο ζωής σε παγκόσμιο επίπεδο (McNicol, 2002). Η αλλαγή αυτή οφείλεται κυρίως στη βελτίωση των συνθηκών διαβίωσης και στην ανάπτυξη της ιατρικής επιστήμης. Την περίοδο 1961-2010, το προσδόκιμο ζωής κατά τη γέννηση στην Ελλάδα αυξήθηκε από 70.2 σε 78 έτη για τους άνδρες και από 73.8 σε 83.3 για τις γυναίκες, περίπου δηλαδή 9 έτη κατά μέσο όρο για τα δύο φύλα, τα τελευταία 50 χρόνια. Η αύξηση όμως του αριθμού των ηλικιωμένων, σε συνδυασμό με τα χαμηλά ποσοστά γονιμότητας δημιουργούν ένα σοβαρό δημογραφικό πρόβλημα και εγείρει σημαντικούς προβληματισμούς για τη μελλοντική δομή του ελληνικού συστήματος συνταξιοδότησης.

Στη διεθνή βιβλιογραφία υπάρχει μεγάλο ερευνητικό ενδιαφέρον για τη μελέτη των τάσεων θνησιμότητας ενός πληθυσμού. Μια από τις πιο δημοφιλείς μεθόδους μοντελοποίησης των δεικτών θνησιμότητας προτάθηκε από τους Lee and Carter (1992). Εκτός από τις διαφοροποιήσεις της μεθόδου που ακολούθησαν, το μοντέλο των Lee-Carter (LC) ενέπνευσε πολλούς ερευνητές για την κατασκευή μοντέλων με

επιπρόσθετους παράγοντες, όπως αυτά των Renshaw and Haberman (2006) και των Cairns et al. (2006) και πιο συγκριμένα για την περίπτωση της Ελλάδας, μια σχετική συμβολή στο πλαίσιο μελέτης των γενικευμένων γραμμικών μοντέλων περιγράφηκε από τους Hatzopoulos and Haberman (2009).

Στην εργασία αυτή, θα γίνει για πρώτη φορά (όσο γνωρίζουμε) εφαρμογή των μοντέλων Lee and Carter (1992), Renshaw and Haberman (2006) και Cairns et al. (2006) στα ελληνικά δεδομένων, χρησιμοποιώντας τη γλώσσα στατιστικού προγραμματισμού R, ακολουθώντας την προσέγγιση των Villegas et al. (2016). Μέσα από τη συγκριτική ανάλυση που θα ακολουθήσει, εξάγουμε συμπεράσματα για το πόσο καλά προσαρμόζονται τα παραπάνω μοντέλα στα ελληνικά δεδομένα, αλλά και προβλέψεις, οι οποίες θα αξιολογηθούν.

Το πλάνο της εργασίας είναι το εξής: Στην Ενότητα 2 παρουσιάζονται τα μοντέλα που θα χρησιμοποιηθούν, ενώ στην Ενότητα 3 περιγράφεται αναλυτικά η διαδικασία προσαρμογής των μοντέλων στα ελληνικά δεδομένα, καθώς και αντίστοιχοι έλεγχοι που απαιτούνται. Στην Ενότητα 4 δίνονται συγκριτικά αποτελέσματα που αφορούν τις προβλέψεις με τη χρήση του κάθε μοντέλου, ενώ τα συμπεράσματα και οι παρατηρήσεις των ευρημάτων μας δίνονται στην Ενότητα 5.

2. Μοντελοποίηση Δεδομένων Θνησιμότητας

Σύμφωνα με τους Booth and Tickle (2008), η ανάπτυξη των μοντέλων θνησιμότητας βασίστηκε σε τρεις μεθόδους: της προσδοκίας (expectation), της επεξήγησης (explanation) και της παρέκτασης (extrapolation). Στις μεθόδους προσδοκίας, η πρόβλεψη θνησιμότητας βασίζεται στην άποψη ενός ειδικού που ενσωματώνει συγκεκριμένη δημογραφική ή άλλη σχετική γνώση, κάποιες φορές όμως μπορεί να οδηγήσει σε υποκειμενικότητα ή σφάλματα μεροληψίας. Οι επεξηγηματικές μέθοδοι βασίζονται σε δομικά ή επιδημιολογικά μοντέλα συγκεκριμένων αιτιών θανάτου και περιορίζονται κυρίως σε βραχυπρόθεσμες προβλέψεις. Η μέθοδος παρέκτασης αποτελεί την πιο σύγχρονη ερευνητική μέθοδο και στηρίζεται στην υπόθεση ότι οι προηγούμενες τάσεις βελτίωσης της θνησιμότητας αναμένεται να συνεχιστούν στο μέλλον. Όλα τα μοντέλα που θα συζητηθούν στις επόμενες ενότητες ενσωματώνουν την μέθοδο παρέκτασης.

2.1 Δεδομένα και Συμβολισμοί

Ο αριθμός των παρατηρηθέντων θανάτων ατόμων ηλικίας $x = x_1, x_2, \dots, x_k$ κατά το έτος $t = t_1, t_2, \dots, t_n$ συμβολίζεται με $d_{x,t}$ και η αντίστοιχη κεντρική έκθεση στον κίνδυνο¹ με $E_{x,t}^c$. Τα δεδομένα αυτά αντλήθηκαν για τον ελληνικό πληθυσμό από την Human Mortality Database (2015). Η αντίστοιχη αρχική έκθεση στον κίνδυνο²

¹Ως κεντρική έκθεση στον κίνδυνο ορίζουμε τον αριθμό των ατόμων του πληθυσμού ηλικίας x που εκτίθενται σε κίνδυνο στο μέσον του έτους t .

²Ως αρχική έκθεση στον κίνδυνο ορίζουμε τον αριθμό των ατόμων του πληθυσμού ηλικίας x που εκτίθενται σε κίνδυνο στην αρχή του έτους t .

συμβολίζεται με $E_{x,t}^0$ και προσεγγίζεται από τη σχέση: $E_{x,t}^0 \approx E_{x,t}^c + (1/2)d_{x,t}$. Ως εκ τούτου, η κατ' έτος πιθανότητα θανάτου ατόμου ηλικίας x πριν φτάσει την ηλικία $x + 1$ κατά το ημερολογιακό έτος t συμβολίζεται με $q_{x,t}$ και υπολογίζεται από τη σχέση: $q_{x,t} = d_{x,t}/E_{x,t}^0$.

Η τυχαία μεταβλητή του αριθμού των θανάτων συμβολίζεται με $D_{x,t}$. Το $D_{x,t}$ μπορεί να έχει, μεταξύ άλλων, κατανομή (i) $\text{Poisson}(E_{x,t}^c \mu_{x,t})$, όπου $\mu_{x,t} = E(D_{x,t}/E_{x,t}^c)$ λέγεται και ένταση θνησιμότητας³ και (ii) $\text{Binomial}(E_{x,t}^0, q_{x,t})$, όπου $q_{x,t} = E(D_{x,t}/E_{x,t}^0)$ η πιθανότητα θανάτου, όπως έχει οριστεί πιο πάνω. Έτσι, τα μοντέλα θνησιμότητας χρησιμοποιούν είτε $\eta_{x,t} = \log \mu_{x,t}$, είτε το logit του $q_{x,t}$, δηλαδή $\eta_{x,t} = \text{logit } q_{x,t} = \log \frac{q_{x,t}}{1 - q_{x,t}}$.

Οι τιμές της έντασης θνησιμότητας $\mu_{x,t}$ ατόμων ηλικίας x κατά το έτος t αντλήθηκαν από την Human Mortality Database (HMD). Διευκρινίζουμε ότι η HMD περιλαμβάνει μόνο τους αντίστοιχους κεντρικούς δείκτες θνησιμότητας $m_{x,t}$ ⁴, ωστόσο υποθέτοντας ότι η ένταση θνησιμότητας παραμένει σταθερή για κάθε ηλικία και έτος (κάτι που γενικά ισχύει για ηλικίες κάτω των 100), ο κεντρικός δείκτης και η ένταση θνησιμότητας συμπίπτουν.

Στην HMD, τα ελληνικά δεδομένα διατίθενται ανά φύλο και ηλικία για την περίοδο από το 1981 έως το 2013. Στην εργασία αυτή, θα ληφθούν υπόψη οι ηλικίες $x = 60, 61, \dots, 93, 94$, καθώς παρουσιάζουν ιδιαίτερο ασφαλιστικό ενδιαφέρον. Επιπλέον, προκειμένου να καταλήξουμε σε πιο αξιόπιστα συμπεράσματα ως προς την προσαρμογή και τις προβλέψεις των μοντέλων, χρησιμοποιούνται τα δεδομένα της περιόδου $t = 1981, 1982, \dots, 2009, 2010$, αφήνοντας τα τελευταία 3 έτη (2011, 2012, 2013) για τους ελέγχους της ακρίβειας των προβλέψεων (backtesting).

2.2 Στοχαστικά Μοντέλα

Στη συνέχεια περιγράφονται τα μοντέλα που θα προσαρμόσουμε στα ελληνικά δεδομένα.

2.2.1 M_1 : The Lee-Carter Model (LC)

Ένα από τα δημοφιλέστερα μοντέλα στη διεθνή βιβλιογραφία είναι το μοντέλο των Lee and Carter (1992), το οποίο δίνεται από τη σχέση:

$$\eta_{x,t} = \alpha_x + \beta_x^{(1)} \kappa_t^{(1)}, \quad (1)$$

όπου α_x και $\beta_x^{(1)}$ είναι οι παράμετροι ηλικίας, με $x = x_1, x_2, \dots, x_k$ και $\kappa_t^{(1)}$ η παράμετρος περιόδου που εκφράζει το γενικό επίπεδο θνησιμότητας κατά το έτος

³Ως ένταση θνησιμότητας ορίζουμε την οριακή τιμή του κεντρικού δείκτη θνησιμότητας $m_{x,t}$ στο διάστημα $[x + \Delta(x))$ κατά το έτος t και δίνεται από τη σχέση: $\mu_{x,t} = \lim_{\Delta(x) \rightarrow 0} \Delta(x) m_{x,t}$.

⁴Ως κεντρικό δείκτη θνησιμότητας ορίζουμε το λόγο του αριθμού των θανάτων ($d_{x,t}$) ατόμων ηλικίας x κατά το έτος t προς τον αντίστοιχο πληθυσμό ($E_{x,t}^c$) των ατόμων της ίδιας ηλικίας στο μέσον του ιδίου έτους.

$t = t_1, t_2, \dots, t_n$. Στο μοντέλο (1) χρησιμοποιείται η υπόθεση της Poisson κατανομής για τους θανάτους με $\eta_{x,t} = \log \mu_{x,t}$. Οι τιμές των παραμέτρων εκτιμώνται χρησιμοποιώντας τη μέθοδο εκτίμησης μέγιστης πιθανοφάνειας (MLE). Παρατηρούμε επίσης ότι το $\eta_{x,t}$ στην (1) παραμένει αμετάβλητο αν για τυχαία σταθερά $c \neq 0$ χρησιμοποιήσουμε τους παρακάτω μετασχηματισμούς:

$$\{\alpha_x, \beta_x^{(1)}, \kappa_t^{(1)}\} \rightarrow \{\alpha_x - \beta_x^{(1)}c, \beta_x^{(1)}, \kappa_t^{(1)} + c\},$$

$$\{\alpha_x, \beta_x^{(1)}, \kappa_t^{(1)}\} \rightarrow \{\alpha_x, \beta_x^{(1)}c, \frac{\kappa_t^{(1)}}{c}\}.$$

Για την ταυτοποιησιμότητα (identifiability) του μοντέλου (1), οι Lee and Carter (1992) πρότειναν τους παρακάτω περιορισμούς:

$$\sum_t \kappa_t^{(1)} = 0, \quad \sum_x \beta_x^{(1)} = 1.$$

Η παράμετρος περιόδου κ_t , που όπως θα δούμε χρησιμοποιείται για τις προβλέψεις, μοντελοποιείται μέσω μιας χρονολογικής σειράς τυχαίου περιπάτου με μετατόπιση (random walk with a drift) της μορφής: $\kappa_t = \kappa_{t-1} + \delta + \epsilon_t$, όπου κάθε τιμή κ_t συνδέεται με την προηγούμενή της κ_{t-1} μέσω μιας σταθεράς μετατόπισης δ και ενός σφάλματος ϵ_t .

2.2.2 M_2 : The Renshaw-Haberman Model (RH)

Οι Renshaw and Haberman (2006) εξέλιξαν το μοντέλο Lee-Carter με την προσθήκη μιας επιπλέον παραμέτρου γ_{t-x} που συλλαμβάνει ειδικά αποτελέσματα που αφορούν στο έτος γέννησης $t - x$, καταλήγοντας στην παρακάτω μορφή για το $\eta_{x,t}$:

$$\eta_{x,t} = \alpha_x + \beta_x^{(1)} \kappa_t^{(1)} + \gamma_{t-x}, \quad (2)$$

όπου πάλι χρησιμοποιείται η υπόθεση της Poisson κατανομής θανάτων με $\eta_{x,t} = \log \mu_{x,t}$. Όπως στο προηγούμενο μοντέλο, έτσι και εδώ, η ταυτοποιησιμότητα του μοντέλου (2) για την εκτίμηση των παραμέτρων επιτυγχάνεται με τη χρήση των ακόλουθων περιορισμών:

$$\sum_x \beta_x^{(1)} = 1, \quad \sum_t \kappa_t^{(1)} = 0, \quad \sum_{x,t} \gamma_{t-x} = 0.$$

Για τις προβλέψεις, οι εκτιμήσεις που προέκυψαν με τη μέθοδο μέγιστης πιθανοφάνειας των παραμέτρων $\kappa_t^{(1)}$ και γ_{t-x} μοντελοποιήθηκαν με τη χρήση μοντέλων χρονολογικών σειρών ARIMA⁵, υποθέτοντας ανεξαρτησία μεταξύ των παραμέτρων περιόδου $\kappa_t^{(1)}$ και γενεάς γ_{t-x} .

⁵Ως Autoregressive Integrated Moving Average (ARIMA) ορίζεται ένα αυτοπαλίνδρομο μοντέλο κινούμενου μέσου, η γενική μορφή του οποίου δίνεται από τη σχέση (10) στην Ενότητα 4.

2.2.3 M_3 : The Cairns-Blake-Dowd Model (CBD)

Με σκοπό τη μείωση των παραμέτρων προς εκτίμηση, οι Cairns et al. (2006) πρότειναν ένα μοντέλο με απλούστερη δομή που ενσωματώνει μόνο δύο παραμέτρους περιόδου και το αντίστοιχο $\eta_{x,t}$ δίνεται από τη σχέση:

$$\eta_{x,t} = \kappa_t^{(1)} + (x - \bar{x})\kappa_t^{(2)}, \quad (3)$$

όπου $\kappa_t^{(1)}$ και $\kappa_t^{(2)}$ είναι οι παράμετροι περιόδου, $\bar{x}$ είναι η μέση ηλικία στα διαθέσιμα δεδομένα και $\eta_{x,t} = \text{logit } q_{x,t}$. Ο logit μετασχηματισμός για την κατ' έτος πιθανότητα χρησιμοποιείται λόγω της υπόθεσης ότι οι θάνατοι ακολουθούν διωνυμική κατανομή.

Οι προβλέψεις για το μοντέλο αυτό εξήχθησαν χρησιμοποιώντας έναν διδιάστατο τυχαίο περίπατο με μετατόπιση για τη μοντελοποίηση των παραμέτρων περιόδου, χωρίς μάλιστα την ανάγκη ύπαρξης περιορισμών, όπως στα προηγούμενα μοντέλα.

3. Προσαρμογή των Μοντέλων

Η εκτίμηση των παραμέτρων πραγματοποιείται με τη μέθοδο μέγιστης πιθανοφάνειας. Ο λογάριθμος της συνάρτησης πιθανοφάνειας (log-likelihood function) στην περίπτωση της Poisson κατανομής των θανάτων δίνεται από τη σχέση (Hunt and Blake (2015)):

$$\mathcal{L} = \sum_{x,t} \omega_{x,t} \left\{ d_{x,t} \log(E_{x,t}^c \mu_{x,t}) - E_{x,t}^c \mu_{x,t} - \log(d_{x,t}!) \right\}, \quad (4)$$

όπου τα βάρη $\omega_{x,t}$ ⁶ παίρνουν τις τιμές $\{0, 1\}$, $d_{x,t}$ είναι ο αριθμός των παρατηρούμενων θανάτων, $E_{x,t}^c$ η κεντρική έκθεση στον κίνδυνο όπως έχει οριστεί στην Υποενότητα (2.1) και

$$\mu_{x,t} = \begin{cases} \exp(\alpha_x + \beta_x^{(1)} \kappa_t^{(1)}), & \text{για το μοντέλο } M_1, \\ \exp(\alpha_x + \beta_x^{(1)} \kappa_t^{(1)} + \gamma_{t-x}), & \text{για το μοντέλο } M_2. \end{cases}$$

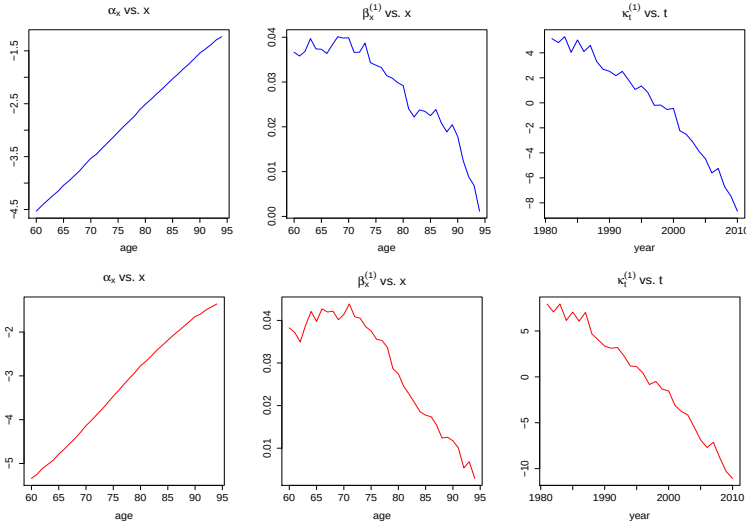
Υπό την υπόθεση της διωνυμικής κατανομής, ο λογάριθμος της συνάρτησης πιθανοφάνειας για τον αριθμό των θανάτων δίνεται από τη σχέση (Hunt and Blake, 2015):

$$\mathcal{L} = \sum_{x,t} \omega_{x,t} \left\{ d_{x,t} \log q_{x,t} + (E_{x,t}^0 - d_{x,t}) \log(1 - q_{x,t}) + \log \left(\frac{E_{x,t}^0}{d_{x,t}} \right) \right\}, \quad (5)$$

όπου $E_{x,t}^0$ η αρχική έκθεση στον κίνδυνο όπως έχει οριστεί στην Υποενότητα (2.1) και η $q_{x,t}$ για το μοντέλο M_3 υπολογίζεται από τη σχέση $\frac{q_{x,t}}{1 - q_{x,t}} = \exp(\kappa_t^{(1)} + (x - \bar{x})\kappa_t^{(2)})$.

⁶ Ακολουθώντας τη λογική των Cairns et al. (2009) για να είναι πιο αξιόπιστη η εκτίμηση της παραμέτρου γενεάς γ_{t-x} , θέτουμε $\omega_{x,t} = 0$ για τα έτη γέννησης που περιλαμβάνουν λιγότερες από 5 παρατηρήσεις, δηλαδή τα έτη 1887-1891 και 1946-1950.

Οι εκτιμήσεις των παραμέτρων για τα μοντέλα M_1 και M_2 έγιναν μεγιστοποιώντας την Εξίσωση (4), ενώ για το M_3 εκτιμήθηκαν από την Εξίσωση (5). Τα Σχήματα 1-3 παρουσιάζουν τις εκτιμήσεις των παραμέτρων των μοντέλων $M_1 - M_3$ αντίστοιχα, για τους άνδρες και τις γυναίκες, ηλικίας 60-94 κατά την περίοδο 1981-2010.



Σχήμα 1: M_1 α_x , $\beta_x^{(1)}$ και $\kappa_t^{(1)}$ εκτιμώμενες παράμετροι για τους άνδρες (άνω διαγράμματα) και τις γυναίκες (κάτω διαγράμματα).

3.1 Έλεγχος Καλής Προσαρμογής

Στη συνέχεια, θα αξιολογηθεί η προσαρμογή των μοντέλων στα δεδομένα, μέσω της μελέτης των αποκλίσεων των καταλοίπων (residuals deviance $\text{dev}(x, t)$) μεταξύ των παρατηρούμενων τιμών ($d_{x,t}$) από τα ιστορικά δεδομένα μας και των εκτιμώμενων ($\hat{d}_{x,t}$) από το κάθε μοντέλο. Η έλλειψη τυχαιότητας στην εικόνα των καταλοίπων αποτελεί μια πρώτη ένδειξη της αδυναμίας του μοντέλου να συλλάβει την επίδραση αποτελεσμάτων ηλικίας, περιόδου και γενεάς.

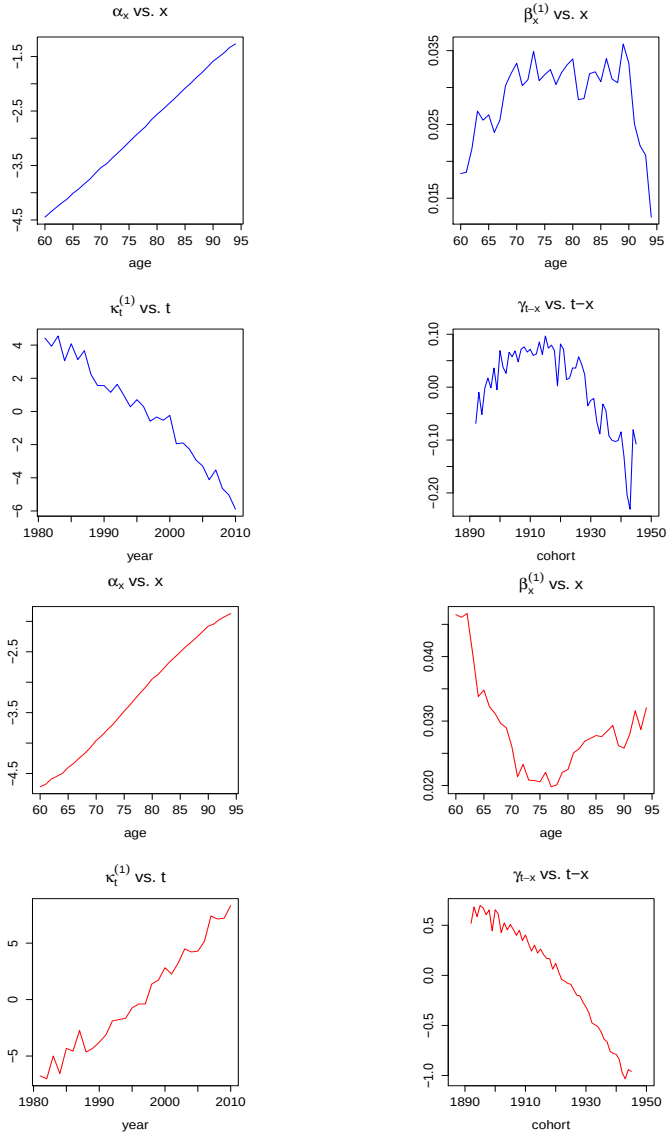
Έτσι, για την περίπτωση της Poisson κατανομής για τον αριθμό των θανάτων, η απόκλιση $\text{dev}(x, t)$ εκφράζεται από τη σχέση (Pitacco et al., 2009):

$$\text{dev}(x, t) = 2d_{x,t} \log \left(\frac{d_{x,t}}{\hat{d}_{x,t}} \right) - (d_{x,t} - \hat{d}_{x,t}),$$

και η συνολική απόκλιση για τα μοντέλα M_1 , M_2 δίνεται από τη σχέση:

$$D(d_{x,t}, \hat{d}_{x,t}) = \sum_{x,t} \text{dev}(x, t) = \sum_{x,t} 2\omega_{x,t} \left\{ d_{x,t} \log \left(\frac{d_{x,t}}{\hat{d}_{x,t}} \right) - (d_{x,t} - \hat{d}_{x,t}) \right\}. \quad (6)$$

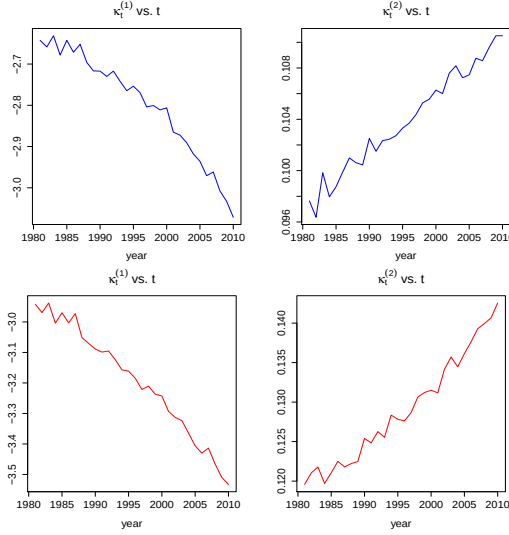
Για την περίπτωση της διωνυμικής κατανομής, η απόκλιση $\text{dev}(x, t)$ δίνεται από τη



Σχήμα 2: M_2 α_x , $\beta_x^{(1)}$, $\kappa_t^{(1)}$ και γ_{t-x} εκτιμώμενες παράμετροι για τους άνδρες (άνω διαγράμματα) και τις γυναίκες (κάτω διαγράμματα).

σχέση (Debón et al., 2010):

$$\text{dev}(x, t) = 2d_{x,t} \log \left(\frac{d_{x,t}}{\hat{d}_{x,t}} \right) + (E_{x,t}^0 - d_{x,t}) \log \left(\frac{E_{x,t}^0 - d_{x,t}}{E_{x,t}^0 - \hat{d}_{x,t}} \right),$$



Σχήμα 3: M_3 $\kappa_t^{(1)}$ και $\kappa_t^{(2)}$ εκτιμώμενες παράμετροι για τους άνδρες (άνω διαγράμματα) και τις γυναίκες (κάτω διαγράμματα).

και η συνολική απόκλιση για το μοντέλο M_3 υπολογίζεται από τη σχέση:

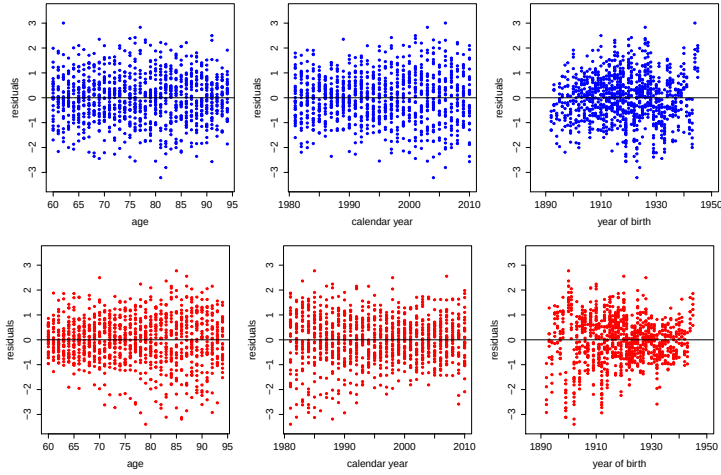
$$D(d_{x,t}, \hat{d}_{x,t}) = \sum_{x,t} \text{dev}(x,t) = \sum_{x,t} 2\omega_{x,t} \left\{ d_{x,t} \log \left(\frac{d_{x,t}}{\hat{d}_{x,t}} \right) + (E_{x,t}^0 - d_{x,t}) \log \left(\frac{E_{x,t}^0 - d_{x,t}}{E_{x,t}^0 - \hat{d}_{x,t}} \right) \right\}. \quad (7)$$

Έτσι, τα τυποποιημένα (standardized) κατάλοιπα δίνονται από τη σχέση:

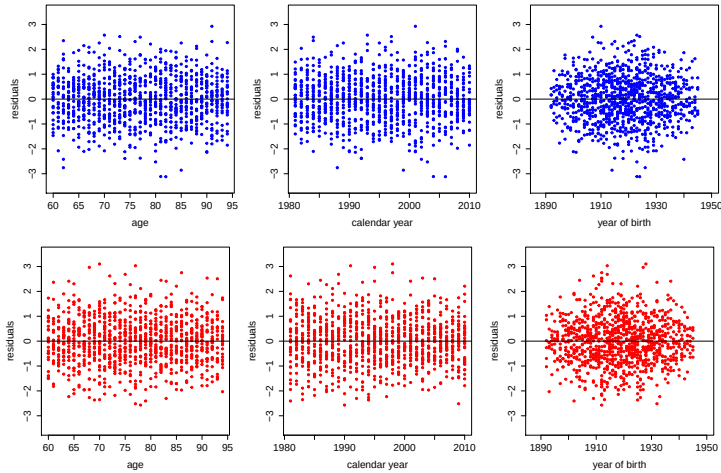
$$r_{x,t} = \text{sign}(d_{x,t} - \hat{d}_{x,t}) \left(\frac{\text{dev}(x,t)}{\hat{\phi}} \right)^{1/2}. \quad (8)$$

Τα βάρη $\omega_{x,t}$ στις Εξισώσεις (6), (7) ορίζονται όπως και στην (4) και η παράμετρος $\hat{\phi} = \frac{D(d_{x,t}, \hat{d}_{x,t})}{\nu}$, όπου ν εκφράζει τους βαθμούς ελευθερίας του μοντέλου (ο αριθμός των παρατηρήσεων μείον τον αριθμό των παραμέτρων του κάθε μοντέλου). Τα Σχήματα 4 - 6 απεικονίζουν τις τιμές των καταλοίπων ανά ηλικία, περίοδο και γενεά για τα M_1 - M_3 που προσαρμόστηκαν για την περίοδο 1981-2010 και τις ηλικίες 60-94 για τους άνδρες και τις γυναίκες αντίστοιχα.

Στα Σχήματα 4 και 6 παρατηρούμε ότι τα πιο απλά μοντέλα M_1 και M_3 φαίνονται ανεπαρκή στο να συλλάβουν την επίδραση της γενεάς (δεξιά διαγράμματα). Επιπλέον, στο Σχήμα 6 βλέπουμε την αδυναμία του μοντέλου M_3 να συλλάβει την επίδραση της ηλικίας (αριστερά διαγράμματα), ιδιαίτερα για τις γυναίκες. Εντούτοις, όλα τα μοντέλα μπορούν να συλλάβουν την επίδραση της περιόδου.



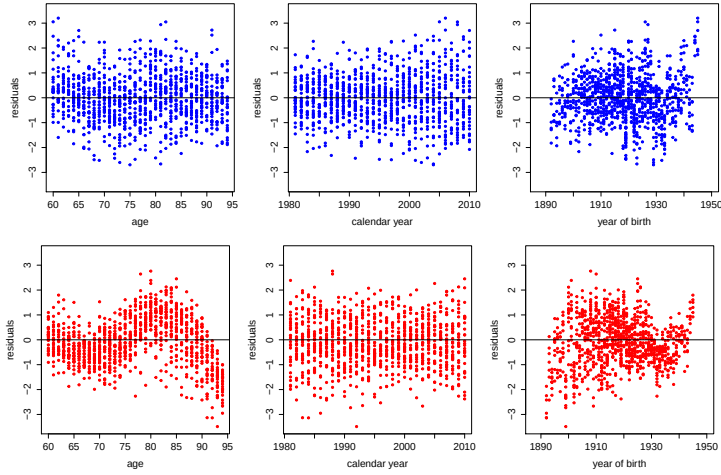
Σχήμα 4: Αποκλίσεις καταλοίπων του μοντέλου M_1 για τους άνδρες (άνω διαγράμματα) και τις γυναίκες (κάτω διαγράμματα) στην Ελλάδα. Περίοδος 1981-2010, ηλικίες 60-94.



Σχήμα 5: Αποκλίσεις καταλοίπων του μοντέλου M_2 για τους άνδρες (άνω διαγράμματα) και τις γυναίκες (κάτω διαγράμματα) στην Ελλάδα. Περίοδος 1981-2010, ηλικίες 60-94.

3.2 Κριτήρια Πληροφορίας

Γενικά, τα μοντέλα με περισσότερες παραμέτρους αναμένεται να έχουν καλύτερη προσαρμογή στα δεδομένα. Για το λόγο αυτό χρησιμοποιούνται τα κριτήρια πληροφορίας AIC και BIC για το κάθε μοντέλο. Οι Hurvich and Tsai (1989) κατέληξαν σε μια διορθωμένη έκδοση του AIC, η οποία θα αναφέρεται ως AIC(c) και προτείνεται για χρήση σε μικρότερα δείγματα. Τα κριτήρια πληροφορίας AIC(c) και BIC



Σχήμα 6: Αποκλίσεις καταλοίπων του μοντέλου M_3 για τους άνδρες (άνω διαγράμματα) και τις γυναίκες (κάτω διαγράμματα) στην Ελλάδα. Περίοδος 1981-2010, ηλικίες 60-94.

ορίζονται ως εξής:

$$AIC(c) = AIC + \frac{2k(k+1)}{n-k-1}, \text{ με } AIC = 2k - 2\log \hat{\mathcal{L}} \text{ και}$$

$$BIC = (\log n)k - 2\log \hat{\mathcal{L}},$$

όπου $\hat{\mathcal{L}}$ είναι η εκτιμήτρια μέγιστης πιθανοφάνειας, k ο αριθμός των ελεύθερων παραμέτρων⁷ και n ο αριθμός των παρατηρήσεων. Μικρότερες τιμές $AIC(c)$ και BIC αποτελούν ένδειξη καλύτερης προσαρμογής του μοντέλου. Ο Πίνακας 1 παρουσιάζει τις εκτιμήσεις μέγιστης πιθανοφάνειας, τον αριθμό των ελεύθερων παραμέτρων μαζί με τις αντίστοιχες τιμές των κριτηρίων $AIC(c)$ και BIC για τα 3 μοντέλα θνησιμότητας με σειρά κατάταξης (τιμές στις παρενθέσεις), για τους άνδρες και τις γυναίκες.

Σύμφωνα με τα αποτελέσματα του κριτηρίου AIC_c για τους άνδρες, το M_2 είναι πρώτο στη σχετική κατάσταση, το M_1 είναι δεύτερο και το M_3 τρίτο, ενώ για το BIC , το M_2 είναι πρώτο, το M_3 ακολουθεί και το M_1 είναι τρίτο. Πρέπει να σημειωθεί ότι επειδή το κριτήριο BIC είναι πιο αυστηρό από το AIC_c , αναμένεται να επιλέξει μοντέλα με λιγότερες παραμέτρους. Για τις γυναίκες, τα αποτελέσματα AIC_c και BIC είναι τα ίδια. Το M_2 είναι πρώτο, ακολουθεί το M_1 και το M_3 είναι τρίτο. Τέλος, οι τιμές των κριτηρίων για τα μοντέλα M_1 και M_3 δείχνουν ότι η επίδραση των αποτελεσμάτων γενεάς πρέπει να ληφθούν υπόψιν για την αποτελεσματικότερη μοντελοποίηση της θνησιμότητας για τον ελληνικό πληθυσμό.

⁷Ως αριθμό των ελεύθερων παραμέτρων (ή effective parameters) ορίζουμε το άθροισμα των εκτιμώμενων παραμέτρων μείον αυτές που χρησιμοποιούνται για τους περιορισμούς του κάθε μοντέλου.

Άνδρες				
Μοντέλο	Μέγιστη Πιθανοφάνεια	Ελεύθερες Παράμετροι	AIC(c)	BIC
M_1	-5.472,27	98	11.162(2)	11.623(3)
M_2	-5.080,40	151	10.516(1)	11.207(1)
M_3	-5.573,81	60	11.275(3)	11.563(2)
Γυναίκες				
M_1	-6.234,97	98	12.687(2)	13.149(2)
M_2	-5.171,16	151	10.697(1)	11.388(1)
M_3	-7.923,67	60	15.975(3)	16.263(3)

Πίνακας 1: Οι εκτιμήσεις μέγιστης πιθανοφάνειας και ο αριθμός των ελεύθερων παραμέτρων μαζί με τις αντίστοιχες τιμές των κριτηρίων AICc (σε σειρά κατάταξης) και BIC για τους άνδρες και τις γυναίκες.

4. Προβολή Θνησιμότητας

Στην ενότητα αυτή θα γίνει προβολή των δεικτών θνησιμότητας χρησιμοποιώντας τα 3 μοντέλα που αναλύθηκαν πριν. Το κλειδί για την εξαγωγή προβλέψεων είναι η μοντελοποίηση των παραμέτρων περιόδου και γενεάς, χρησιμοποιώντας μοντέλα χρονολογικών σειρών. Ακολουθώντας τους Haberman and Renshaw (2011), για τις χρονολογικές σειρές των παραμέτρων $\kappa_t^{(i)}$, $i = 1, 2$ των μοντέλων M_1 - M_3 , χρησιμοποιείται ένας πολυδιάστατος τυχαίος περίπατος με μετατόπιση (ARIMA(0,1,0) with drift) που στη γενική του μορφή δίνεται από τη σχέση:

$$\kappa_t = \delta + \kappa_{t-1} + \epsilon_t, \quad \kappa_t = (\kappa_t^{(1)}, \dots, \kappa_t^{(N)})^T, \quad \epsilon_t \sim \mathcal{N}(\mathbf{0}, \Sigma), \quad (9)$$

όπου κ_t και δ συμβολίζει τα διανύσματα N -διαστάσεων των δεικτών περιόδου και της παραμέτρου μετατόπισης αντίστοιχα, ενώ ως Σ ορίζεται ο πίνακας διασποράς - συνδιασποράς (variance - covariance matrix) μιας πολυδιάστατης διαδικασίας λευκού θορύβου (white noise) ϵ_t . Να σημειωθεί ότι τα μοντέλα M_1 και M_2 ενσωματώνουν μόνο μία παράμετρο περιόδου και συνεπώς γι' αυτά χρησιμοποιείται ο αντίστοιχος μονοδιάστατος τυχαίος περίπατος με μετατόπιση.

Ο δείκτης γενεάς $\gamma_{t-x} \equiv \gamma_c$ θεωρούμε ότι είναι ανεξάρτητος από τον δείκτη περιόδου κ_t για κάθε μοντέλο και ακολουθεί μια μονοδιάστατη χρονολογική σειρά ARIMA (Αυτοπαλίνδρομο Υπόδειγμα Κινούμενου Μέσου) της μορφής:

$$(1 - \phi_1 L - \dots - \phi_p L^p) (1 - L)^d \gamma_c = \delta_0 + (1 + \theta_1 L + \dots + \theta_q L^q) e_c, \quad (10)$$

όπου L^k είναι ο τελεστής χρονικής υστέρησης που μετατοπίζει τα δεδομένα k περιόδους πίσω, δ_0 είναι μια σταθερή παράμετρος μετατόπισης, $\phi_1, \dots, \phi_p$ αποτελεί το αυτοπαλίνδρομο (autoregressive) μέρος με $\phi_p \neq 0$, η $\theta_1, \dots, \theta_q$ είναι η παράμετρος κινούμενου μέσου (moving average) με $\theta_q \neq 0$ και e_c μια διαδικασία λευκού θορύβου.

Ανάμεσα σε ένα εύρος υποψήφιων μοντέλων, το καταλληλότερο υπόδειγμα ARI-MA για το μοντέλο M_2 , επιλέχθηκε σύμφωνα με τους ελέγχους μοναδιαίας ρίζας

ADF (Dickey and Fuller, 1979) και PP (Phillips and Perron, 1988), καθώς και τα κριτήρια πληροφορίας AIC(c) και BIC. Ο Πίνακας 2 παρουσιάζει τα επιλεγμένα μοντέλα ARIMA για τους δείκτες γενεάς γ_{t-x} του μοντέλου M_2 για τους άνδρες και τις γυναίκες.

<i>Μοντέλο</i>	<i>γ_{t-x} για τους άνδρες</i>	<i>γ_{t-x} για τις γυναίκες</i>
M_1		
M_2	ARIMA(0,1,1) με μετατόπιση	ARIMA(2,1,0) με μετατόπιση
M_3		

Πίνακας 2: Επιλεγμένα μοντέλα ARIMA για τους δείκτες γενεάς γ_{t-x} του μοντέλου M_2 για τους άνδρες και τις γυναίκες.

Με τη βοήθεια της R, οι χρονολογικές σειρές (9) και (10) παρήγαγαν 1000 μονοπάτια προσομοίωσης (simulated trajectories) για τις μελλοντικές τιμές των δεικτών περιόδου $\hat{\kappa}_t = (\hat{\kappa}_t^{(1)}, \dots, \hat{\kappa}_t^{(N)})^T$ και γενεάς $\hat{\gamma}_{t_n+s-x}$, όπου με $s = 1, 2, \dots, 20$ συμβολίζονται τα έτη του ορίζοντα πρόβλεψης. Έπειτα, οι μελλοντικές τιμές θνησιμότητας εξήχθησαν χρησιμοποιώντας την εκτιμήτρια:

$$\hat{\eta}_{x,t_n+s} = \alpha_x + \sum_{i=1}^N \beta_x^{(i)} \hat{\kappa}_{t_n+s}^{(i)} + \hat{\gamma}_{t_n+s-x}, \quad (11)$$

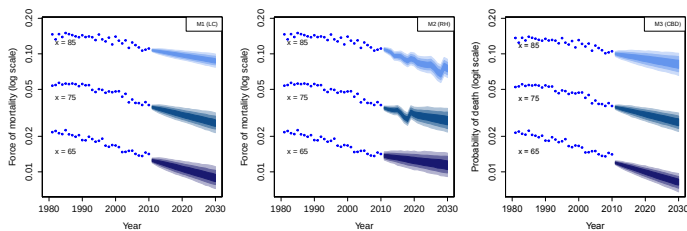
όπου t_n είναι το τελευταίο έτος της περιόδου προσαρμογής (το έτος 2011) και με $\hat{\eta}_{x,t_n+s}$ συμβολίζονται οι μελλοντικές τιμές της εκτιμήτριας, οι οποίες αντιπροσωπεύουν είτε τις μελλοντικές τιμές έντασης θνησιμότητας $\hat{\mu}_{x,t_n+s}$ για τα μοντέλα M_1 και M_2 , είτε την κατ' έτος πιθανότητα θανάτου $\hat{q}_{x,t_n+s}$ για το μοντέλο M_3 .

Ένας τρόπος για να συγκρίνουμε την προβλεπτική ικανότητα ενός μοντέλου θνησιμότητας είναι μέσω του υπολογισμού των σφαλμάτων που προκύπτουν από τη διαφορά μεταξύ της πραγματικής και της προβαλλόμενης τιμής για την ίδια περίοδο. Για τα τρία πρώτα έτη πρόβλεψης εκτός δείγματος (out-of-sample), η ακρίβεια πρόβλεψης για τα μοντέλα M_1 - M_3 μπορεί να αξιολογηθεί χρησιμοποιώντας τα μέτρα υπολογισμού σφάλματος MAE (Mean Absolute Error), RMSE (Root Mean Squared Error) και MAPE (Mean Absolute Percentage). Τα σφάλματα των βραχυπρόθεσμων προβλέψεων της περιόδου 2011-2013 για τους άνδρες και τις γυναίκες συγκρίθηκαν για τα μοντέλα M_1 - M_3 . Οι τιμές του Πίνακα 3 δείχνουν ότι το μοντέλο M_2 δίνει καλύτερες προβλέψεις σε σχέση με τα M_1 και M_3 , λόγω κυρίως της επίδρασης της παραμέτρου γενεάς στους δείκτες θνησιμότητας ανδρών και γυναικών.

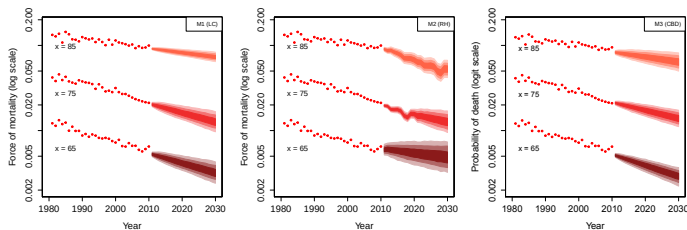
Οι προβλέψεις θνησιμότητας για τα επόμενα 20 έτη (έως και το έτος 2030) εξήχθησαν από τη σχέση (11) για τα 3 μοντέλα θνησιμότητας, χρησιμοποιώντας 1000 μονοπάτια προσομοίωσης για τους δείκτες περιόδου και γενεάς. Τα αποτελέσματα της προσομοίωσης της έντασης θνησιμότητας για τα μοντέλα M_1 και M_2 , καθώς και οι κατ' έτος πιθανότητες θανάτου για το μοντέλο M_3 απεικονίζονται στα Σχήματα 7 και 8 για τους άνδρες και τις γυναίκες αντίστοιχα.

Άνδρες			
Μέτρο Σφάλματος	M_1	M_2	M_3
MAE	0,006	0,004	0,006
RMSE	0,011	0,007	0,010
MAPE	9,139	5,455	10,180
Γυναίκες			
MAE	0,004	0,003	0,005
RMSE	0,007	0,006	0,009
MAPE	8,845	5,213	12,113

Πίνακας 3: Τιμές μέτρων σφάλματος MAE, RMSE και MAPE των βραχυπρόθεσμων προβλέψεων της περιόδου 2011-2013 για τους άνδρες και τις γυναίκες.



Σχήμα 7: Αποτελέσματα μακροπρόθεσμων προβλέψεων (2011-2030) στην ηλικία $x = 65$ (κάτω γραμμές), $x = 75$ (μεσαίες γραμμές) και $x = 85$ (πάνω γραμμές) που εξήχθησαν για τα 3 μοντέλα που προσαρμόστηκαν στα δεδομένα ανδρών για τις ηλικίες 60-94 της περιόδου 1981-2010 στην Ελλάδα. Οι σκιασμένες περιοχές στην περίοδο πρόβλεψης 2011-2030 υποδηλώνουν τα 50%, 80% και 95% αντίστοιχα διαστήματα εμπιστοσύνης.



Σχήμα 8: Αποτελέσματα μακροπρόθεσμων προβλέψεων (2011-2030) στην ηλικία $x = 65$ (κάτω γραμμές), $x = 75$ (μεσαίες γραμμές) και $x = 85$ (πάνω γραμμές) που εξήχθησαν για τα 3 μοντέλα που προσαρμόστηκαν στα δεδομένα γυναικών για τις ηλικίες 60-94 της περιόδου 1981-2010 στην Ελλάδα. Οι σκιασμένες περιοχές στην περίοδο πρόβλεψης 2011-2030 υποδηλώνουν τα 50%, 80% και 95% αντίστοιχα διαστήματα εμπιστοσύνης.

5. Συμπεράσματα

Στην εργασία αυτή επιχειρήσαμε να προσδιορίσουμε τα μοντέλα που παρουσιάζουν την καταλληλότερη προσαρμογή στα δεδομένα σε συνδυασμό με μια ικανοποιητική προβλεπτική ικανότητα. Τα αποτελέσματα της ανάλυσης έδειξαν ότι το μοντέλο M_2

(Renshaw and Haberman, 2006) ξεχώρισε και ως προς την καλύτερη προσαρμογή του στα δεδομένα θνησιμότητας ανδρών και γυναικών για τον ελληνικό πληθυσμό, αλλά και για την ακρίβεια των προβλέψεων του. Από την ανάλυση των στοχαστικών μοντέλων θνησιμότητας καταλήξαμε στα παρακάτω βασικά συμπεράσματα.

Παρότι σε μια τέτοια μελέτη είναι εξαιρετικά σημαντική η επίτευξη «καλής» προσαρμογής των μοντέλων στα ιστορικά δεδομένα, δεν ισχύει πάντα ότι το μοντέλο που προσαρμόζει καλύτερα στα δεδομένα θα δίνει και τα καλύτερα αποτελέσματα πρόβλεψης. Αυτό το σημείο τονίζεται και σε αντίστοιχες έρευνες που έχουν διεξαχθεί για άλλες χώρες στη διεθνή βιβλιογραφία. Ειδικότερα τώρα για την περίπτωση της Ελλάδας, η επίδραση της παραμέτρου γενεάς κατά τη μοντελοποίηση των δεδομένων έπαιξε καθοριστικό ρόλο στην επιλογή του καταλληλότερου μοντέλου.

Τα αποτελέσματα της προσαρμογής στα δεδομένα δείχνουν ότι όλα τα μοντέλα μπορούν να συλλάβουν αποτελεσματικά την επίδραση που έχει η παράμετρος περιόδου και για τα δύο φύλα. Παρατηρούμε επίσης ότι τα πιο απλά μοντέλα M_1 και M_3 , λόγω έλλειψης της παραμέτρου γενεάς δεν μπορούν να συλλάβουν αντίστοιχα αποτελέσματα, όπως αυτό απεικονίζεται στα δεξιά διαγράμματα των Σχημάτων 4 και 6. Επιπλέον, το μοντέλο M_3 φαίνεται ανεπαρκές στο να συλλάβει αποτελέσματα ηλικίας όπως και αναμέναμε, ειδικά για τις γυναίκες (αριστερά διαγράμματα του Σχήματος 6). Οι τιμές των κριτηρίων AIC(c) και BIC συμφωνούν στο ότι το μοντέλο M_2 είναι το καλύτερο όσον αφορά στην προσαρμογή του στα δεδομένα του ανδρικού και του γυναικείου πληθυσμού. Το γεγονός ότι τα μοντέλα M_1 και M_3 έχουν χειρότερη κατάταξη και για τα δύο φύλα αποτελεί ένδειξη της ανάγκης εισαγωγής μιας παραμέτρου γενεάς για την επίτευξη καλύτερης προσαρμογής στα ελληνικά δεδομένα. Επίσης, ο έλεγχος ακρίβειας των βραχυπρόθεσμων προβλέψεων που αξιολογήθηκε από τα αποτελέσματα των μέτρων σφάλματος MAE, RMSE και MAPE για τα έτη 2011-2013 δείχνει ότι το M_2 δίνει τις καλύτερες προβλέψεις (Πίνακας 3).

Τέλος, η περιορισμένη διαθεσιμότητα ιστορικών δεδομένων για τον ελληνικό πληθυσμό ήταν ένα επιπλέον ζήτημα που αντιμετωπίσαμε κατά τη μοντελοποίηση της θνησιμότητας. Στο μέλλον, με μεγαλύτερες περιόδους διαθέσιμων δεδομένων, ευελπιστούμε ότι θα είμαστε σε θέση να καταλήξουμε σε πιο εκτενή αποτελέσματα.

ABSTRACT

In the last decades, life expectancy has risen significantly in the most developed countries all over the world, including Greece. In this paper, we investigate the most important mortality models in the literature, aiming to generate more accurate and efficient forecasts for the insurance industry and the pension funds. More precisely, we incorporate the Greek population data into the Lee and Carter (1992), Renshaw and Haberman (2006) and Cairns et al. (2006) stochastic mortality models in order to compare their fitting behaviour and their forecasting power. The goodness of fit for each model is evaluated using the AIC(c) and BIC information criteria, while forecast accuracy is estimated by the MAE, RMSE and MAPE error measures.

REFERENCES

- Booth, H. and Tickle, L. (2008). Mortality Modelling and Forecasting: a Review of Methods. *Annals of Actuarial Science*, 3(1-2):3–43.
- Cairns, A. J. G., Blake, D., and Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. *Journal of Risk and Insurance*, 73(4):687–718.
- Cairns, A. J. G., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A., and Balevich, I. (2009). A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal*, 13(1):1–35.
- Debón, a., Martínez-Ruiz, F., and Montes, F. (2010). A geostatistical approach for dynamic life tables: The effect of mortality on remaining lifetime and annuities. *Insurance: Mathematics and Economics*, 47(3):327–336.
- Dickey, D. and Fuller, W. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366a):427–431.
- Haberman, S. and Renshaw, A. (2011). A comparative study of parametric mortality projection models. *Insurance: Mathematics and Economics*, 48(1):35–55.
- Hatzopoulos, P. and Haberman, S. (2009). A parameterized approach to modeling and forecasting mortality. *Insurance: Mathematics and Economics*, 44(1):103–123.
- Human Mortality Database (2015). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org.
- Hunt, A. and Blake, D. (2015). On the Structure and Classification of Mortality Models. Pension Institute. Working Paper. Available at <http://www.pensions-institute.org/workingpapers/wp1506.pdf>.
- Hurvich, C. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307.
- Lee, L. R. and Carter, R. D. (1992). Modeling and Forecasting US Mortality. *American Statistical Association*, 87(419):659–671.
- McNicoll, G. (2002). World Population Ageing 1950-2050. *Population and Development Review*, 28(4):814–816.
- Phillips, P. and Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, 75(2):335–346.
- Pitacco, E., Denuit, M., Haberman, S., and Olivieri, A. (2009). *Modelling longevity dynamic for pensions and annuity business*. Oxford University Press.
- Renshaw, A. and Haberman, S. (2006). A cohort-based extension to the Lee-Carter model for mortality reduction factors. *Insurance: Mathematics and Economics*, 38(3):556–570.
- Villegas, A. M., Millossovich, P., and Kaishev, V. (2016). StMoMo: An R Package for Stochastic Mortality Modelling. Working Paper. Available at <https://cran.r-project.org/web/packages/StMoMo/vignettes/StMoMoVignette.pdf>.

ΑΠΑΓΟΡΕΥΣΗ ΤΩΝ P-VALUES ΚΑΙ ΤΩΝ ΕΛΕΓΧΩΝ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ*

Τάκης Παπαϊωάννου

Πανεπιστήμια Πειραιώς και Ιωαννίνων

takrap@unipi.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία μελετάμε το θέμα της απαγόρευσης των p – values που προτάθηκε (και υιοθετήθηκε) από το περιοδικό *Basic and Applied Social Psychology* τον Φεβρουάριο του 2015. Το θέμα έχει κάνει αίσθηση στη στατιστική κοινότητα και έχουν υπάρξει πολυάριθμα σχόλια και αντιδράσεις. Παρουσιάζουμε τις θέσεις του περιοδικού και την προϊστορία του θέματος. Ανέκαθεν οι Ψυχολόγοι είχαν αμφιβολίες και διατύπωναν επιφυλάξεις για τη χρήση της στατιστικής συμπερασματολογίας (έλεγχοι σημαντικότητας και κατά συνέπεια p – values και διαστήματα εμπιστοσύνης) σε ότι αφορά την έρευνα στο χώρο της ψυχολογίας και της συμπεριφοράς.

Κάνουμε μια ανασκόπηση του θέματος και παρουσιάζουμε τις απόψεις υποστηρικτών και μη, μαζί με τη δική μας θέση και καταλήγουμε με την πολύ αναμενόμενη δήλωση/ θέση της American Statistical Association. Το κλειδί σε όλα αυτά είναι η ορθή χρήση και εφαρμογή και ερμηνεία της στατιστικής μεθοδολογίας.

1. ΕΙΣΑΓΩΓΗ

$$p\text{-value} = P(T \geq T_{\text{obs}} | H_0).$$

Τον Φεβρουάριο του 2014 το περιοδικό *Journal of Basic and Applied Social Psychology* δημοσίευσε κύριο άρθρο (editorial) με το οποίο προανήγγειλε την απαγόρευση των p -values και των ελέγχων σημαντικότητας για τα προς δημοσίευση άρθρα (David Trafimow (2014) Editorial, *Journal of Basic and Applied Psychology*, 36:1,1-2) . «Από τώρα και στο εξής το περιοδικό δεν θα απαιτεί, εκ των προτέρων, διαδικασίες συμπερασματολογικής στατιστικής. Οι συγγραφείς είναι ελεύθεροι να αποφασίσουν αν και ποια από αυτές τις διαδικασίες θα χρησιμοποιήσουν» έγραψε ο Υπεύθυνος Έκδοσης (Editor) David Trafimou. Και οι κύριοι λόγοι είναι

* Έρευνα και δημοσίευση με μερική υποστήριξη του Κέντρου Ερευνών Πανεπιστημίου Πειραιώς

α) Η διαδικασία ελέγχου σημαντικότητας μηδενικών υποθέσεων έχει αποδειχθεί ότι δεν είναι λογικά έγκυρη και παρέχει λιγοστή πληροφόρηση για την πραγματική πιθανοφάνεια της μηδενικής ή της πειραματικής υπόθεσης και

β) Επί του παρόντος δεν υπάρχει διαδικασία στατιστικής συμπερασματολογίας η οποία να έχει τύχει ευρείας αποδοχής.

Τα πειράματα, γράφει το editorial, (δηλ. οι δειγματοληψίες, τα ερωτηματολόγια, κλπ) θα πρέπει να περιλαμβάνουν επαρκή αριθμό ατόμων έτσι ώστε να υπάρχει κάποια εμπιστοσύνη στα μεγέθη των επιδράσεων - ευρημάτων (effect sizes). Με τον όρο effect size εννοείται κάθε ποσοτικό μέγεθος της δύναμης ενός φαινομένου, όρος γνωστός ιδιαίτερα στην μεταανάλυση, όπως π.χ., το r , ο εκτιμητής του β στην παλινδρόμηση, η διαφορά δυο δειγματικών μέσων, κλπ. Ας σημειωθεί ότι υπάρχουν 50-100 effect sizes (Wikipedia).

Στη συνέχεια το κύριο άρθρο του περιοδικού παραθέτει μερικά ενδιαφέροντα στοιχεία της εκδοτικής του πολιτικής όπως ότι θα δημοσιεύει μη στατιστικά σημαντικά αποτελέσματα (null effects), έρευνα η οποία φέρνει στο φως αποτελέσματα αντίθετα προς ήδη δημοσιευμένα, σχολιάζει και ουσιαστικά απορρίπτει την ανάλυση διαμεσολάβησης (mediation analysis) και σχολιάζει αρνητικά τη μεγάλη διάδοση και ουσιαστικά απορρίπτει τους μηχανισμούς αιτιότητας

Ένα χρόνο μετά, τον Μάρτιο του 2015, (David Trafimow & Michael Marks (2015) Editorial, *Basic and Applied Social Psychology*, 37:1,1-2) το περιοδικό στο κύριο άρθρο του διεκήρυξε

- Η περίοδος χάριτος έληξε:
- Από τώρα και στο εξής το BASP εργασίες που στηρίζονται σε διαδικασίες ελέγχου σημαντικότητας των μηδενικών υποθέσεων ή/και σε p-values δε θα δημοσιεύονται και επαναλαμβάνει ότι . Η διαδικασία των τεστ σημαντικότητας μηδενικών υποθέσεων χαρακτηρίστηκε ως «μη λογικά έγκυρη». Συγγραφείς που υποβάλλουν εργασίες για δημοσίευση στο περιοδικό θα πρέπει να αφαιρούν όλα τα καίρια σημεία (vestiges) της διαδικασίας του στατιστικού ελέγχου σημαντικότητας μηδενικών υποθέσεων (p-values, t-values, F-values, statements about “significant” differences or lack thereof, κλπ)

Αυτό έγινε γνωστό ως **απαγόρευση των p-values** και κατά συνέπεια εγκατάλειψη των ελέγχων (τεστ) σημαντικότητας μηδενικών υποθέσεων.

Η απόφαση ήταν και είναι εντυπωσιακή και προκάλεσε και προκαλεί αίσθηση και θύελλα συζητήσεων αντιδράσεων άρθρων και σχολίων. Χτύπησε το νεύρο των Στατιστικών. Το editorial το γνωστοποίησε στο ΕΣΙ ο συνάδελφος Γ. Παπαδόπουλος από το Γεωπονικό Πανεπιστήμιο.

Στην ουσία οι editors του περιοδικού ισχυρίζονται ότι με τα p-values υπολογίζουμε την πιθανότητα να παρατηρήσουμε την τιμή ή πιο ακραία τιμή του στατιστικού που παρατηρήθηκε δοθείσης της μηδενικής υπόθεσης (πράγμα που είναι

ορθό) ενώ θα έπρεπε να υπολογίζουμε την αντίθετη πιθανότητα. Αυτό ακούγεται ως μπεϋζιανή προσέγγιση, αν και οι editors δεν τάσσονται καθαρά υπέρ της μπεϋζιανής λύσης

Είναι γνωστή από καιρό η διάσταση απόψεων μεταξύ Ψυχολόγων και Στατιστικών σε ό,τι αφορά στον έλεγχο στατιστικών υποθέσεων και τα p-values. Βλέπε το άρθρο του D.H. Krantz με τίτλο 'The null hypotheses testing controversy in Psychology' στο περιοδικό *JASA*, Dec 1999, 1372-1381, το οποίο σε μεγάλο βαθμό αποτελεί ανασκόπηση του βιβλίου *What if there were no significance tests?* των Lisa L. Harlow, Stanley A. Mulaik και James A. Staiger, 40 χρόνια ανημιαχίες στα Ψυχολογικά Περιοδικά.

Τα τελευταία χρόνια συχνά διαβάζαμε στην επιστημονική βιβλιογραφία αμφισβητήσεις για την 'επιστημονική μέθοδο' ελέγχου υποθέσεων δια των στατιστικών αναλύσεων (βλέπε Fisher) ή ότι υπάρχουν πολλές αδυναμίες στη διαδικασία ελέγχων σημαντικότητας μηδενικών υποθέσεων. Επιπλέον ότι με 'μεγάλα δεδομένα' (large data sets ή big data) και μεγάλη ακρίβεια (precision) τα p-values είναι σχεδόν άχρηστα και ακόμη ότι μπορείς να απορρίψεις κάθε μηδενική υπόθεση με μεγάλο μέγεθος δείγματος.

Το θέμα και πρόβλημα έχει επεκταθεί τα τελευταία χρόνια με την πρόοδο της τεχνολογίας και την ευρεία, ευρύτατη χρήση των υπολογιστών και των στατιστικών προγραμμάτων. Σε κάθε έκφανση της ζωής, της επιστήμης, της ανθρώπινης δραστηριότητας, κλπ, όπου υπάρχουν δεδομένα τα οποία συλλέγονται με οποιοδήποτε τρόπο, τυχαίο ή μη με δειγματοληψία ή μη, με πείραμα ή κλινική δοκιμή, με παρατηρητική μελέτη ή μη, διατυπώνονται κάθε είδους υποθέσεις, ρεαλιστικές ή μη, μερικές φορές ανόητες, εισάγονται το δεδομένα στον υπολογιστή και με κάποιο πρόγραμμα καταλήγουμε σε p-value. Θα αναφέρω δυο παραδείγματα:

α) Υπολογίζεται ο συντελεστής συσχέτισης (του Pearson) μεταξύ της παραγωγής ενός αριθμού μικρών αγροτεμαχίων για μια καλλιεργητική περίοδο και του συνολικού αριθμού υπέρ πτήσεων πουλιών πάνω από τα αγροτεμάχια κατά την ίδια περίοδο και δίνεται το σχετικό p-value

β) Σε επιστημονικό περιοδικό καταγράφεται το ακόλουθο συμπέρασμα : "the earth is round ($p < 0.05$)".

Το θέμα συνδέεται και με την κρίση αναπαραγωγισιμότητας στις επιστήμες, θέμα με το οποίο προτιθέμεθα να ασχοληθούμε προσεχώς. Στην παρούσα εργασία κάνουμε μια ανασκόπηση του θέματος, παρουσιάζουμε τις απόψεις υποστηρικτών και μη, μαζί με τη δική μας θέση και καταλήγουμε με την πολύ αναμενόμενη δήλωση/ θέση της American Statistical Association. Το κλειδί σε όλα αυτά είναι η ορθή χρήση και εφαρμογή και ερμηνεία της στατιστικής μεθοδολογίας.

2. Η ΑΠΑΝΤΗΣΗ ΤΩΝ ΣΤΑΤΙΣΤΙΚΩΝ.

Μία απάντηση των Στατιστικών μπορεί να βρεθεί σε σύντομο σχόλιο του περιοδικού *Significance* τον Απρίλιο του 2015. Η αντίδραση ήταν μικτή. Σταχυολογώ δύο απαντήσεις: Ο Andrew Gelman του Columbia University έγραψε «Κορυφαία περιοδικά της ψυχολογίας καταλήγουν να δημοσιεύουν γελοίους επιστημονικά ανεύθυνους ισχυρισμούς οι οποίοι δικαιολογούνται με ένα $p < 0,05$ » και επιπλέον «Πιστεύω τυπικά σφάλματα, p -values και διαστήματα εμπιστοσύνης μπορεί να είναι πολύ χρήσιμα εργαλεία έρευνας, όταν θεωρούνται ως «εξυπηρετικά» (convenient) συστατικά της ανάλυσης δεδομένων.....Το πρόβλημα προκύπτει όταν τα p -values θεωρούνται ως το αποκορύφωμα της ανάλυσης ως εάν το p είναι μικρότερο του 0,05 αντιπροσωπεύει ένα είδος απόδειξης».

Η **Αμερικανική Στατιστική Εταιρεία (ASA)** δήλωσε ότι η στατιστική κοινότητα γνωρίζει από καιρό τα προβλήματα που συνδέονται με τη χρήση και ερμηνεία των επαγωγικών (συμπερασματολογικών) μεθόδων. Δε συμφωνεί με τη θέση του Περιοδικού και πιστεύει ότι η εναλλακτική πρόταση του περιοδικού να στηρίζεται σε περιγραφική στατιστική και στη «χρήση μεγάλων δειγμάτων που συνήθως χρησιμοποιούνται στις ψυχολογικές έρευνες» θα οδηγήσει σε αρνητικές συνέπειες. Η ASA ήδη επεξεργάζεται μία δήλωση-πρόταση για το θέμα με τη βοήθεια ομάδας κορυφαίων Στατιστικών. Η δήλωση παρατίθεται πιο κάτω

Η **Βασιλική Στατιστική Εταιρεία της Αγγλίας (RSS)** αντέδρασε δημοσιοποιώντας το θέμα και αναρωτήθηκε τί θα έλεγαν οι R. Fisher, J.Neyman και E.Pearson για τον τρόπο με τον οποίο η σύγχρονη επιστήμη χρησιμοποιεί τη στατιστική θεωρία που ανέπτυξαν. Κάλεσε τους Peter Diggle, Stephen Senn, Andrew Grelman, Geoff Cumming και Robert Grant να διατυπώσουν τις απόψεις και θέσεις τους. Από τους πέντε οι τρεις πρώτοι είναι επιφανείς στατιστικοί και απορρίπτουν την ακραία θέση του περιοδικού. Ο τέταρτος είναι ομότιμος Καθηγητής Ψυχολογίας και τάσσεται υπέρ της εκτιμητικής και των μεταναλύσεων και απορρίπτει τα p -values. Ο πέμπτος επικεντρώνεται στη δυσκολία της ερμηνείας των αποτελεσμάτων όταν η έρευνα περιλαμβάνει πολλά (πολλαπλά) p -values και καταλήγει λέγοντας ότι η απαγόρευση των p -values θα δράσει ως μεγάλο φύλο συκής καλύπτοντας ότι το πρόβλημα είναι το πώς σκέπτονται οι ερευνητές. Μπορείτε να δείτε τις θέσεις της RSS στο: <http://www.statslife.org.uk/opinion/2114-journal-s-ban-on-null-hypothesis>.

Η παραπάνω θέση του JBASP βρήκε υποστηρικτές λίγο καιρό αργότερα στο περιοδικό *Science News* όπου δημοσιεύτηκε η φράση « p -values are at the root of all (well, most) scientific evil». Η Καθηγήτρια Regina Nuzzo, στατιστικός και δημοσιογράφος, τον Φεβρουάριο του 2014 γράφει στο περιοδικό *Nature*: Statistical Errors – p values, the ‘gold standard’ of statistical validity, are not as reliable as many scientists assume. Η Καθηγήτρια Nuzzo, διδακτορικό από το Stanford στη Στατιστική, είναι και κάτοχος του Βραβείου Εξαίρετης Στατιστικής Δημοσιογραφίας για το 2014 (Excellence in Statistical Reporting- 2014 ESRA).

Σε όλα τα παραπάνω παραβλέπεται το επίκαιρο και πάλι θέμα της ορθής η υπερβολικής χρήσης (misuse and overuse) της στατιστικής μεθοδολογίας από μη επαγγελματίες και εκπαιδευμένους Στατιστικούς.

Το θέμα είναι ενδιαφέρον και ο συγγραφέας κάλεσε δια του *Στατιστικού Περισκοπίου* τους Έλληνες Στατιστικούς να διατυπώσουν τις θέσεις τους και τα πιστεύω τους για τα p-values με σχετική επιστολή.

3. Η ΘΕΣΗ ΤΗΣ ASA ΓΙΑ ΤΑ P – VALUES.

Στις 7 Μαρτίου 2016 η ASA εξέδωσε τη δήλωση σχετική με την απαγόρευση των p-values, ως δελτίο τύπου, το οποίο αναρτήθηκε στο διαδίκτυο, εστάλη σε πολλούς αποδέκτες και παρατίθεται στο Παράρτημα. Ταυτόχρονα έδωσε στη δημοσιότητα και κύριο άρθρο (editorial) των Wasserstein and Lazar (2016) με τίτλο ‘The ASA’s statement on p-values: context, process, and purpose’. Το άρθρο δημοσιεύθηκε στο ευρύτατης κυκλοφορίας *American Statistician*. Συνοδεύεται και με είκοσι άρθρα συζήτησης (online discussion articles) επιφανών Στατιστικών οι οποίοι σχολιάζουν το editorial.

Το άρθρο ξεκινά με ένα κυκλικό λογοπαίγνιο:

E. Γιατί τόσα πολλά Πανεπιστήμια διδάσκουν $p=0.05$;

A. Διότι το χρησιμοποιεί η επιστημονική κοινότητα και οι editors των περιοδικών.

E. Γιατί τόσοι πολλοί χρησιμοποιούν $p=0.05$;

A. Διότι αυτό είναι που διδάσκονται στα Πανεπιστήμια.

Η δήλωση της American Statistical Association για τα p-values ξεκινά δίνοντας τον συνήθη ορισμό του p-value και στη συνέχεια διατυπώνει εξι αρχές-προτάσεις οι οποίες επεξηγούν και περιγράφουν το ρόλο τη σημασία και τη χρήση των p-values. Κάθε αρχή συμπληρώνεται με επεξηγηματικά σχόλια. Οι **αρχές** αυτές είναι*

1. Τα p-values μπορεί να υποδηλώνουν πόσο ασυμβίβαστα τα δεδομένα είναι με ένα συγκεκριμένο στατιστικό μοντέλο (P-values can indicate how incompatible the data are with a specified statistical model).

2. Τα p-values δεν δίνουν (μετρούν) την πιθανότητα η υπό μελέτη (διερεύνηση) υπόθεση να είναι αληθινή ή την πιθανότητα τα δεδομένα να προέκυψαν μόνο στην τύχη (P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone).

Οι ερευνητές συχνά θέλουν να μετατρέψουν το p-value σε μία πρόταση-δήλωση για την αλήθεια (ορθότητα) μιας μηδενικής υπόθεσης, ή για την πιθανότητα η τύχη να παρήγαγε τα δεδομένα που παρατηρήθηκαν. **Το p-value δεν είναι κανένα από αυτά.**

* Για λόγους πληρέστερης διατύπωσης παρατίθεται και το αρχικό αγγλικό κείμενο

Είναι μία πρόταση για τα δεδομένα σε σχέση με κάποια υποθετική επεξήγηση και δεν είναι μια πρόταση για την ίδια την επεξήγηση (It is a statement about data in relation to a specific hypothetical explanation, and it is not a statement about the explanation itself).

3. Επιστημονικά συμπεράσματα και επιχειρηματικές αποφάσεις ή αποφάσεις πολιτικής δεν πρέπει να στηρίζονται μόνο στο εάν ένα p-value περνά κάποιο συγκεκριμένο κατώφλι (Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold).

Πραγματικοί λόγοι συχνά απαιτούν δίτιμες ναι ή όχι αποφάσεις, αλλά αυτό δεν σημαίνει ότι τα p-values από μόνα τους μπορούν να εξασφαλίσουν ότι μία απόφαση είναι ορθή ή λανθασμένη (Pragmatic considerations often require binary, yes-no decisions, but this does not mean that p-values alone can insure that a decision is correct or incorrect).

4. Η κατάλληλη συμπερασματολογία απαιτεί πλήρη έκθεση και παρουσίαση αλλά και διαφάνεια (Proper inference requires full reporting and transparency).

5. Το p-value ή η στατιστική σημαντικότητα δεν μετρούν το μέγεθος οποιασδήποτε επίδρασης (ευρήματος) ή τη σημασία (αξία) ενός αποτελέσματος (A p-value, or statistical significance, does not measure the size of any effect or the importance of a result).

6. Από μόνο του, το p-value δεν παρέχει ένα καλό μέτρο μαρτυρίας ή ένδειξης για το κατά πόσο ισχύει ένα μοντέλο ή υπόθεση (By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis).

Συμπέρασμα: Η καλή στατιστική πρακτική ως ουσιαστικό συστατικό της καλής επιστημονικής πρακτικής δίνει έμφαση σε αρχές καλού σχεδιασμού και εκτέλεσης της μελέτης-έρευνας, στη χρήση μίας ποικιλίας αριθμητικών και γραφικών περιλήψεων των δεδομένων, στην κατανόηση του υπό μελέτη φαινομένου, στην ερμηνεία των αποτελεσμάτων εντός πλαισίου, στην πλήρη έκθεση-παρουσίαση και κατάλληλη-ορθή λογική και ποσοτική κατανόηση της έννοιας και σημασίας των περιλήψεων των δεδομένων. Κανένας δείκτης από μόνος του δεν μπορεί να αντικαταστήσει την επιστημονική λογική και επιχειρηματολογία. (Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. No single index should substitute for scientific reasoning).

Άλλες Προσεγγίσεις

Λαμβάνοντας υπόψη την επικρατούσα λανθασμένη χρήση ή αντίληψη σχετικά με τα p-values, η δήλωση της ASA αναφέρει εναλλακτικές λύσεις σε περίπτωση που ο ερευνητής θέλει να αποφύγει ή συμπληρώσει τα p-values:

- Εκτίμηση σε αντιδιαστολή με έλεγχο υποθέσεων όπως π. χ. διαστήματα εμπιστοσύνης, αξιοπιστίας (credibility), πρόβλεψης
- Μπεϋζιανές μέθοδοι
- Άλλα μέτρα μαρτυρίας (evidence) όπως λόγους πιθανοφάνειας, Bayes Factors κλπ
- Decision-theoretic modelling
- False discovery rates, δηλαδή ποσοστά λανθασμένης απόρριψης της μηδενικής υπόθεσης

Όλα αυτά στηρίζονται σε περαιτέρω παραδοχές αλλά μπορεί να μετρούν καλύτερα το μέγεθος μιας επίδρασης (ευρήματος) (effect size) και την συνυπάρχουσα αβεβαιότητα ή το αν η υπόθεση είναι ορθή ή λανθασμένη

Είναι προφανές ότι η ASA έκανε μια πολύ χρήσιμη αλλά και πολύ προσεκτική και διπλωματική δήλωση προσπαθώντας να συγκεράσει όλες τις τάσεις και απόψεις που συνυπάρχουν στον κλάδο και το επάγγελμα μας. Αυτό είναι φανερό και από την διατύπωση της δήλωσης αλλά και από την καθυστέρηση και την πολυπληθή ομάδα που συγκρότησε.

Η ομάδα της ASA

Naomi Altman, Jim Berger, Yoav Benjamini, Don Berry, Brad Carlin, John Carlin, George Cobb, Marie Davidian, Steve Fienberg, Andrew Gelman, Steve Goodman, Sander Greenland, Guido Imbens, John Ioannidis, Valen Johnson, Michael Lavine, Michael Lew, Rod Little, Deborah Mayo, Chuck McCulloch, Michele Millar, Sally Morton, Regina Nuzzo, Hilary Parker, Kenneth Rothman, Don Rubin, Stephen Senn, Uri Simonsohn, Dalene Stangl, Philip Stark, Steve Ziliak.

Από αυτούς οι περισσότεροι έγραψαν και άρθρα συζήτησης (σχόλια) (discussion articles) τα οποία συνοδεύουν τη δήλωση της ASA στο *American Statistician* ως on line supplementary material. Τα σχόλια επικροτούν ως επί το πλείστον τη δήλωση, σχολιάζουν και αναλύουν τις έξι αρχές και διατυπώνουν διάφορες παρατηρήσεις ακόμα και επιφυλάξεις. Η δήλωση προέκυψε από συγκερασμό ή/και συμβιβασμό πολλών απόψεων.

4.ΤΟ BASP ΕΝΑ ΧΡΟΝΟ ΜΕΤΑ ΤΗΝ ΑΠΑΓΟΡΕΥΣΗ.

Ο David Trafimou, ένα χρόνο μετά την απαγόρευση, τον Φεβρουάριο του 2016, λέει στο περιοδικό *Significance*:

- α) «Το editorial είχε θετικό αντίκτυπο και πολύ ερευνητές κατανοούν τώρα το βασικό θέμα της ‘αντίστροφης συμπερασματολογίας’
- β) Υπήρξε μείωση των ‘υποβολών’ άρθρων για το 2015 σε σχέση με το 2014 αλλά είναι δύσκολο να την αποδώσεις στη απαγόρευση.
- γ) Οι αντιδράσεις ήταν από άκρως θετικές μέχρι άκρως αρνητικές. Πριν την απαγόρευση, τα περισσότερα σχόλια των κριτών των άρθρων αναφέρονταν σε

θέματα στατιστικής συμπερασματολογίας. Τώρα εστιάζονται περισσότερο στην αξία της θεωρίας, την ορθότητα του ερευνητικού σχεδιασμού, την πιθανότητα η έρευνα να καταλήξει σε πρακτικό όφελος κλπ

δ) Επί πλέον αναφέρει ότι είχε ενθάρρυνση από διάφορες ‘στατιστικές και μαθηματικές αρχές’ που διατύπωναν ότι αντιλαμβάνονται και κατανοούν την μη εγκυρότητα της διαδικασίας ελέγχου σημαντικότητας μηδενικών υποθέσεων και ότι υποστηρίζουν την απαγόρευση ή τουλάχιστον τη δραματική μείωση της εξάρτησης σ’ αυτές τις διαδικασίες.

5. MARLOS VIANA

Ζητήσαμε τη γνώμη και άλλων συναδέλφων. Ο Καθηγητής Ψυχολογίας Robert W. Lissitz από το University of Maryland της Αμερικής γράφει ότι είχε και έχει προβλήματα με τον κλασικό έλεγχο υποθέσεων αλλά ποτέ δεν θα πρότεινε την απαγόρευση των p-values. Υπάρχουν και άλλα μέτρα που μπορούν να χρησιμοποιηθούν συμπληρωματικά όπως τα effect sizes και λόγοι πιθανοφανειών διαφόρων υποθέσεων. Ο Καθηγητής Marlos Viana από το University of Illinois at Chicago προβάλλει δυο ενδιαφέροντα σημεία:

1. *Αβεβαιότητα και πιθανότητα*: Αν και είναι σαφές ότι οι πειραματικές υποθέσεις στις κοινωνικές/ συμπεριφοριστικές/ κλινικές (και άλλες) επιστήμες διέπονται από αβεβαιότητες, είναι λιγότερο σαφές ότι οι αβεβαιότητες αυτές μπορούν να εκφραστούν αριθμητικά ως πιθανότητες. Και ιδιαίτερα, είναι λιγότερο σαφές αν οι πιθανότητες που συνδέονται με στατιστικές παραμετρικές υποθέσεις μπορούν να ερμηνευθούν κατανοητά σε κλασικό πλαίσιο συχνοτήτων (frequentist framework). Η χρήση των p-values, με πιθανοτική ερμηνεία, θα ήταν αδικαιολόγητη σ’ αυτούς τους κλάδους

Αυτό οφείλεται στο ότι στις παραπάνω επιστήμες είναι σχεδόν αδύνατο να εκτιμηθούν τα πειραματικά σφάλματα πέρα από τα εντός του πειράματος δειγματοληπτικά σφάλματα. Π.χ. μία 5-ετής κλινική δοκιμή με χιλιάδες ασθενείς είναι ένα πείραμα μεγέθους ένα σε παρόμοια πειράματα. Εντός του πειράματος αυτού η δειγματοληπτική μεταβλητότητα τερματικών σημείων (study outcomes) μπορεί βέβαια να εκτιμηθεί.

2. *Χρησιμότητα-ωφελιμότητα και πιθανότητα* (utility and probability). Πιθανοτικές προτάσεις (διατυπώσεις), όπως τα p-values, δεν περιλαμβάνουν συνήθως ωφελιμότητα, κίνδυνο και κόστος και έτσι δεν υπόκεινται σε δημόσιο εξέταση-έλεγχο (scrutiny).

Σε άλλη του παρέμβαση δίνει ένα χαρακτηριστικό παράδειγμα λανθασμένης παρουσίασης της έννοιας της στατιστικής σημαντικότητας προς το ευρύτερο κοινό: Διαβάζει κανείς στην εφημερίδα New York Times « Αν και παρατηρήθηκε διαφορά, δεν ήταν στατιστικά σημαντική, πράγμα που σημαίνει ότι μπορεί να οφείλονταν στην τύχη και μόνο» Μερικές φορές το και μόνο παραλείπεται. Υπάρχουν πάρα πολλά τέτοια παραδείγματα κακής και λανθασμένης στατιστικής πληροφόρησης.

6. Η ΔΙΚΗ ΜΑΣ ΘΕΣΗ

Η επιστημονική μέθοδος μαζί με τους ελέγχους σημαντικότητας που είναι το κύριο εργαλείο της, αποτελεί μια από τις μεγάλες κατακτήσεις του 20^{ου} αιώνα. Η αξία και η λογική τους έχουν θεμελιωθεί ακλόνητα στην επιστήμη.

Το p-value ή το επίπεδο σημαντικότητας α για ένα έλεγχο σημαντικότητας έχει αδιαμφισβήτητη αξία και χρησιμότητα. Το p-value είναι το ελάχιστο α που απαιτείται για να απορριφθεί μία μηδενική υπόθεση. Πρόβλημα υπάρχει με τη χρήση και εφαρμογή πολλαπλών τεστ (ελέγχων σημαντικότητας) όπου αναπόφευκτα το τελικό, γενικό (overall) α είναι μεγαλύτερο του αρχικού.

Συχνά αναφέρεται ότι ο Στατιστικός μπορεί να απορρίψει οποιαδήποτε μηδενική υπόθεση αρκεί το μέγεθος του δείγματος n να είναι μεγάλο. Για πολλούς κοινούς και συνήθεις ελέγχους αυτό ισχύει. Δε γνωρίζουμε όμως κάποια γενική πρόταση ή θεώρημα που να επιβεβαιώνει την παρατήρηση. Το θέμα συνδέεται και με τα «μεγάλα δεδομένα» και τις νέες τεχνολογίες, όπου εκεί όπως φαίνεται τα p-values έχουν μικρή αξία (Little, 2016).

Φαίνεται, όμως, να υπάρχει κάποιο πρόβλημα με τα p-values, όχι ως προς τον μαθηματικό ή στατιστικό ορισμό τους αλλά ως προς τη χρήση τους κυρίως στη μη στατιστική βιβλιογραφία. Και όχι μόνο με τη χρήση αλλά και την προέλευση τους. Ο Στατιστικός διαβάζοντας μια εργασία σε μη στατιστικό περιοδικό και βλέποντας τα λίγα ή πολλά p-values που παρατίθενται συχνά αναρωτιέται για την προέλευση τους, την αιτιολογία τους, την επιβεβαίωση τους ή/και η τεκμηρίωσή τους. Το στατιστικό πακέτο ή πρόγραμμα δεν είναι και τόσο έξυπνο ώστε να γνωρίζει ή να «λέει» αν το p-value που δίνει είναι ορθό ή χρησιμοποιείται ή ερμηνεύεται ορθά.

Θα καταλήξουμε με τις ακόλουθες επιγραμματικές επισημάνσεις:

- Εξακολουθεί να υπάρχει κακή χρήση ή/και υπερβολική χρήση των p-values κυρίως από μη εκπαιδευμένους ή πιστοποιημένους Στατιστικούς
- Τα επιχειρήματα των Ψυχολόγων Ερευνητών δεν είναι πειστικά ή/και δεν έχουν μαθηματική ή στατιστική θεμελίωση.
- Βρήκα τη δήλωση της ASA χλιαρή. Περιμένα κάτι πιο δυνατό.
- Τι απαντάμε όταν το z test του Fisher για τον συντελεστή συσχέτισης r με $n=1000$ δίνει $p<0.05$ και το r είναι 0,01;
- Τα p-value δεν μπορεί να είναι σωσίβιο της μεροληψίας ή των μεροληπιών που μπορεί να έχει μια μελέτη.

ABSTRACT

In this paper we study the topic of the ban on p-values proposed and adopted by the Journal of Basic and Applied Social Psychology in Feb. 2015. The ban has impressed the statistical community and stirred many reactions and comments. We present the

history of the controversy and the journals positions. Since many years ago psychologist expressed doubts, reservations and concerns about the use of statistical inference (tests of significance, p-values and confidence intervals) in psychological and behavioural research.

We present a review of the topic and the views and arguments of those who support and do not support the ban together with our position. We also present the very important Statement of the American Statistical Association on p-values. The key in all these is the correct use and application of statistical methodology.

ΑΝΑΦΟΡΕΣ

Τα άρθρα που μνημονεύονται στο κείμενο και το παράρτημα.

Effect size. *Wikipedia*

Little, R. J. (2016) Online discussion paper in supplementary material. The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, 70:2, 129-133.

Wasserstein, R.L. and Lazar, N.A.(2016) The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, 70:2, 129-133.

ΠΑΡΑΡΤΗΜΑ

1. Editorial, *JBASP*, Feb 10, 2014

David Trafimow

New Mexico State University

As the new Editor of *Basic and Applied Social Psychology (BASP)*, I would like to thank my predecessor, Leonard S. Newman, for his admirable leadership during his tenure as Editor. I also thank the Editorial Board for their valuable services and ask that they maintain their commitment to *BASP*. Under my stewardship, the basic mission of the journal—to publish articles that bear on the importance of social psychology research for understanding social problems—will not change. My Associate Editor, Michael Marks, whom I am pleased to take this opportunity to introduce, shares these views.

However, I emphasize several policies. First, *BASP* will no longer require inferential statistics procedures on an a priori basis. For the most part, authors are free to decide whether to use any of these procedures and, if so, which ones. There is, I believe, a sound reason for this change. The null hypothesis significance testing procedure has been shown to be logically invalid and to provide little information about the actual likelihood of either the null or experimental hypothesis (see Trafimow, 2003; Trafimow & Rice, 2009a, for my position on the issue). At present, there exists no inferential statistical procedure that has elicited widespread agreement. However, *BASP* will expect authors to provide complete descriptive statistical information, including sample effect sizes. Experiments should include sufficient participants so as to instill some confidence in the stability of obtained effect sizes.

Second, *BASP* welcomes the submission of null effects. The obvious criticism of experiments that do not obtain a previously published effect is “You didn’t do it right or you would have gotten the effect!” The problem, of course, is that once an effect is published, it becomes sacred, gains inertia, and thinking to the contrary transforms into heresy. In contrast, I would like *BASP* to become an outlet for null effects, so long as (a) the researcher has tried sufficiently hard to obtain the effect *and* (b) the null effect is sufficiently important. Lest the reader doubt the potential importance of null effects, consider the glorious “failure” by Michelson and Morley (1887) to detect the theorized luminiferous ether that eventually earned a Nobel Prize.

Third, *BASP* welcomes research that contradicts already-published work. We all know from personal experience that it is often difficult to publish research, whether theoretical or empirical, that contradicts an already-published theory, particularly if that theory is considered to be a good basis for publishing research or gaining grants. *BASP* encourages manuscripts that violate accepted wisdom, in the anticipation that this will balance out the bias in the reviewing system against such manuscripts. Among other things, this emphasis pushes in the direction of overruling negative reviews if the criticisms are of the usual variety used to defend the status quo. In the spirit of generating scientific discussion, if you wish to disagree with published work, we recommend *BASP* as a good place to do it.

A fourth issue concerns the popularity of mediation analyses for establishing causal mechanisms. Just as correlation does not mean causation, mediation, which is based on a set of correlations, does not mean causation either (Spencer, Zanna, & Fong, 2005). Even in the context of an experiment, mediation analysis fails to provide strong support that the alleged mediator is a cause of the dependent variable. Nor does mediation analysis provide a valid way to test between alternative causal mechanisms unless unrealistic assumptions are made (Bullock, Green, & Ha, 2010). Put simply, mediation analyses provide poor quality tests of hypothesized mediation. Therefore, manuscripts that depend on mediation analyses as the main source of evidence for a causal mechanism are unlikely to be published in *BASP*, though we may allow mediation analyses that are “supplemental” or “auxiliary.”

Speaking of causal mechanisms, there is a fifth issue—the proliferation in social psychology of causal mechanisms. Although causal mechanisms matter, there is more to science than simply compiling a long list of hypothesized causal mechanisms. Therefore, although *BASP* will continue to publish research on causal mechanisms, *BASP* also encourages the submission of advances in integrative theorizing (mild encouragement) and in unifying theory (much encouragement) that depend on the proposal of more basic principles (see Trafimow, 2012, for my perspective on mechanisms, integration, and unifying theory). Further, *BASP* encourages the submission of philosophically or methodologically oriented manuscripts, even if no data are presented. A manuscript that does not provide strong support for a causal mechanism will not necessarily be precluded from publication in *BASP* if it has other virtues.

In conclusion, although the basic mission of *BASP* remains unchanged, I emphasize the following policies: (a) no requirement that inferential statistics procedures be used; (b) openness to publishing null effects; (c) openness to research that contradicts already-published work; (d) discouragement of manuscripts that depend on mediation analyses; and (e) openness toward theoretically, philosophically, or methodologically oriented manuscripts. Obviously, the space in an editorial is not sufficient to provide a complete explanation or justification for editorial policy, nor have I attempted this. The interested reader can consult the cited articles. A more complete view of my philosophy of journal editing can be obtained by consulting Trafimow and Rice (2009b). Future *BASP* editorial policy will be as consistent with that

article as possible. In the long run, *BASP* will thrive only if authors continue to provide a steady stream of high-quality submissions.

Therefore, it is important to provide a positive experience for authors, which includes careful attention to manuscript strengths as well as to limitations, useful feedback about how to render as publishable work that might not yet be so, and increased freedom to break with entrenched social psychology traditions.

REFERENCES

- Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (Don't expect an easy answer). *Journal of Personality and Social Psychology*, 98, 550–558. doi:10.1037/a0018933
- Michelson, A. A., & Morley, E. W. (1887). On the relative motion of the earth and the luminiferous ether. *American Journal of Science*, 34, 333–345.
- Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than meditational analyses in examining psychological processes. *Journal of Personality and Social Psychology*, 89, 845–851. doi:10.1037/0022-3514.89.6.845
- Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem. *Psychological Review*, 110, 526–535. doi:10.1037/0033-295X.110.3.526
- Trafimow, D. (2012). The role of mechanisms, integration and unification in science and psychology. *Theory & Psychology*, 22, 696–703. doi:10.1177/0959354311433929
- Trafimow, D., & Rice, S. (2009a). A test of the NHSTP correlation argument. *Journal of General Psychology*, 136, 261–269. doi:10.3200/GENP.136.3.261-270
- Trafimow, D., & Rice, S. (2009b). What if social scientists had reviewed great scientific works of the past? *Perspectives in Psychological Science*, 4, 65–78. doi:10.1111/j.1745-6924.2009.01107.x

2. Editorial, *JBASP*, Feb, 2015

David Trafimow and Michael Marks
New Mexico State University

The Basic and Applied Social Psychology (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. From now on, *BASP* is banning the NHSTP. With the banning of the NHSTP from *BASP*, what are the implications for authors? The following are anticipated questions and their corresponding answers.

Question 1. Will manuscripts with p-values be desk rejected automatically?

Answer to Question 1. No. If manuscripts pass the preliminary inspection, they will be sent out for review. But prior to publication, authors will have to remove all vestiges of the NHSTP (p-values, t-values, F-values, statements about “significant” differences or lack thereof, and so on).

Question 2. What about other types of inferential statistics such as confidence intervals or Bayesian methods?

Answer to Question 2. Confidence intervals suffer from an inverse inference problem that is not very different from that suffered by the NHSTP. In the NHSTP, the problem is in traversing the distance from the probability of the finding, given the null hypothesis, to the probability of the null hypothesis, given the finding. Regarding confidence intervals, the problem is that, for example, a 95% confidence interval does not indicate that the parameter of interest has a 95% probability of being within the interval. Rather, it means merely that if an infinite number of samples were taken and confidence intervals computed, 95% of the confidence intervals would capture the population parameter. Analogous to how the NHSTP fails to provide the probability of the null hypothesis, which is needed to provide a strong case for rejecting it, confidence intervals do not provide a strong case for concluding that the population parameter of interest is likely to be within the stated interval. Therefore, confidence intervals also are banned from BASP. Bayesian procedures are more interesting. The usual problem with Bayesian procedures is that they depend on some sort of Laplacian assumption to generate numbers where none exist. The Laplacian assumption is that when in a state of ignorance, the researcher should assign an equal probability to each possibility. The problems are well documented (Chihara, 1994; Fisher, 1973; Glymour, 1980; Popper, 1983; Suppes, 1994; Trafimow, 2003, 2005, 2006). However, there have been Bayesian proposals that at least somewhat circumvent the Laplacian assumption, and there might even be cases where there are strong grounds for assuming that the numbers really are there (see Fisher, 1973, for an example). Consequently, with respect to Bayesian procedures, we reserve the right to make case-by-case judgments, and thus Bayesian procedures are neither required nor banned from BASP.

Question 3. Are any inferential statistical procedures required?

Answer to Question 3. No, because the state of the art remains uncertain. However, BASP will require strong descriptive statistics, including effect sizes. We also encourage the presentation of frequency or distributional data when this is feasible. Finally, we encourage the use of larger sample sizes than is typical in much psychology research, because as the sample size increases, descriptive statistics become increasingly stable and sampling error is less of a problem. However, we will stop short of requiring particular sample sizes, because it is possible to imagine circumstances where more typical sample sizes might be justifiable.

We conclude with one last thought. Some might view the NHSTP ban as indicating that it will be easier to publish in BASP, or that less rigorous manuscripts will be acceptable. This is not so. On the contrary, we believe that the $p < .05$ bar is too easy to pass and sometimes serves as an excuse for lower quality research. We hope and anticipate that banning the NHSTP will have the effect of increasing the quality of submitted manuscripts by liberating authors from the stultified structure of NHSTP thinking thereby eliminating an important obstacle to creative thinking. The NHSTP has dominated psychology for decades; we hope that by instituting the first NHSTP ban, we demonstrate that psychology does not need the crutch of the NHSTP, and that other journals follow suit.

REFERENCES

- Chihara, C. S. (1994). The Howson-Urbach proofs of Bayesian principles. In E. Eells & B. Skyrms (Eds.), *Probability and conditionals: Belief revision and rational decision* (pp. 161–178). New York, NY: Cambridge University Press.

- Fisher, R. A. (1973). *Statistical methods and scientific inference* (3rd ed.). London, England: Collier Macmillan.
- Glymour, C. (1980). *Theory and evidence*. Princeton, NJ: Princeton University Press.
- Popper, K. R. (1983). *Realism and the aim of science*. London, England: Routledge.
- Suppes, P. (1994). Qualitative theory of subjective probability. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 17–38). Chichester, England: Wiley.
- Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem. *Psychological Review*, 110, 526–535.
- Trafimow, D. (2005). The ubiquitous Laplacian assumption: Reply to Lee and Wagenmakers. *Psychological Review*, 112, 669–674.
- Trafimow, D. (2006). Using epistemic ratios to evaluate hypotheses: An imprecision penalty for imprecise hypotheses. *Genetic, Social, and General Psychology Monographs*, 132, 431–462.
- Trafimow, D. (2014). Editorial. *Basic and Applied Social Psychology*, 36(1), 1–2.

3.Δελτίο Τύπου της ASA

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND P-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

The American Statistical Association (ASA) has released a “Statement on Statistical Significance and P-Values” with six principles underlying the proper use and interpretation of the p-value[<http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN>]. The ASA releases this guidance on p-values to improve the conduct and interpretation of quantitative science and inform the growing emphasis on reproducibility of science research. The statement also notes that the increased quantification of scientific research and a proliferation of large, complex data sets has expanded the scope for statistics and the importance of appropriately chosen techniques, properly conducted analyses, and correct interpretation.

Good statistical practice is an essential component of good scientific practice, the statement observes, and such practice “emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean.”

“The p-value was never intended to be a substitute for scientific reasoning,” said Ron Wasserstein, the ASA’s executive director. “Well-reasoned statistical arguments contain much more than the value of a single number and whether that number exceeds an arbitrary threshold. The ASA statement is intended to steer research into a ‘post $p < 0.05$ era.’”

“Over time it appears the p-value has become a gatekeeper for whether work is publishable, at least in some fields,” said Jessica Utts, ASA president. “This apparent editorial bias leads to the ‘file-drawer effect,’ in which research with statistically significant outcomes are much more likely to get published, while other work that might well be just as important

scientifically is never seen in print. It also leads to practices called by such names as ‘p-hacking’ and ‘data dredging’ that emphasize the search for small p-values over other statistical and scientific reasoning.”

The statement’s six principles, many of which address misconceptions and misuse of the p-value, are the following:

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

The statement has short paragraphs elaborating on each principle.

In light of misuses of and misconceptions concerning p-values, the statement notes that statisticians often supplement or even replace p-values with other approaches. These include methods “that emphasize estimation over testing such as confidence, credibility, or prediction intervals; Bayesian methods; alternative measures of evidence such as likelihood ratios or Bayes factors; and other approaches such as decision-theoretic modeling and false discovery rates.”

“The contents of the ASA statement and the reasoning behind it are not new—statisticians and other scientists have been writing on the topic for decades,” Utts said. “But this is the first time that the community of statisticians, as represented by the ASA Board of Directors, has issued a statement to address these issues.”

“The issues involved in statistical inference are difficult because inference itself is challenging,”

Wasserstein said. He noted that more than a dozen discussion papers are being published in the ASA journal *The American Statistician* with the statement to provide more perspective on this broad and complex topic. “What we hope will follow is a broad discussion across the scientific community that leads to a more nuanced approach to interpreting, communicating, and using the results of statistical methods in research.”

About the American Statistical Association

The ASA is the world’s largest community of statisticians and the oldest continuously operating professional science society in the United States. Its members serve in industry, government and academia in more than 90 countries, advancing research and promoting sound statistical practice to inform public policy and improve human welfare. For additional information, please visit the ASA website at www.amstat.org.

For more information:

Ron Wasserstein

(703) 302-1859

ron@amstat.org

ΕΚΤΙΜΗΣΕΙΣ ΤΩΝ ΕΠΙΔΡΑΣΕΩΝ ΕΡΓΑΣΙΑΣ, ΚΕΦΑΛΑΙΟΥ, ΕΠΕΝΔΥΣΕΩΝ ΚΑΙ ΔΗΜΟΣΙΩΝ ΔΑΠΑΝΩΝ ΣΤΟ ΠΕΡΙΦΕΡΕΙΑΚΟ ΠΡΟΪΟΝ: 2001-2011

Πρόδρομος-Ιωάννης Προδρομίδης¹, Ελένη-Βασιλική Παπαθανασίου²

¹Κέντρο Προγραμματισμού και Οικονομικών Ερευνών
pjprodr@kepe.gr

²Πάντειο Πανεπιστήμιο Κοινωνικών και Πολιτικών Επιστημών

ΠΕΡΙΛΗΨΗ

Επιχειρείται μια εμπειρική ανάλυση της Ελληνικής Οικονομίας με αξιοποίηση διαστρωματικών στοιχείων χρονολογικών σειρών βασικών μακροοικονομικών μεταβλητών σε περιφερειακό επίπεδο από το 2000 ως το 2011. Πιο συγκεκριμένα, (α) εξετάζονται οι επιδράσεις (προηγήσεις και ανταποκρίσεις) μεταξύ του κεφαλαίου, της εργασίας και των επενδύσεων ανά παραγωγικό τομέα, καθώς και των δημοσιονομικών δαπανών των κοινοτικών και αμιγών εθνικών προγραμμάτων και (β) εκτιμώνται (απομονώνονται) οι διαφοροποιημένες επιδράσεις αυτών στο περιφερειακό προϊόν της χώρας. Έτσι αναδεικνύονται στοιχεία που ενδέχεται να χρησιμεύουν στην άσκηση αναπτυξιακής πολιτικής σε τομεακό και περιφερειακό επίπεδο.

Λέξεις Κλειδιά: Παραγωγή, Επενδύσεις, Δημοσιονομικές επιδράσεις, Κλάδοι, Περιφέρειες, Ανάπτυξη.

1. ΕΙΣΑΓΩΓΗ

Το άρθρο εξετάζει οικονομετρικά την περιφερειακή και τομεακή διάσταση ορισμένων βασικών μακροοικονομικών μεταβλητών στην χώρα μας κατά την δεκαετία 2001-11, προκειμένου να αποτυπώσει (α) τις μεταξύ των σχέσεις και (β) τις επιδράσεις των εν λόγω μεταβλητών στο Ακαθάριστο Εγχώριο Προϊόν (ΑΕΠ) από την εποχή που η χώρα εισήλθε στην Οικονομική και Νομισματική Ένωση (ΟΝΕ), δηλ. το 2001, ως την εποχή που υπέστη την μεγαλύτερη μείωση του ΑΕΠ στον μισό αιώνα που η Εθνική Στατιστική Υπηρεσία της Ελλάδος (νυν ΕΛΣΤΑΤ) παρέχει στατιστικά στοιχεία, δηλ. το 2011.

Στον πυρήνα της ανάλυσεως βρίσκεται το βασικό υπόδειγμα μιας μακροοικονομίας που ευρίσκεται λίγο ή πολύ μακράν του σημείου της πλήρους απασχολήσεως, με αποτέλεσμα το ΑΕΠ να προσδιορίζεται από παράγοντες που συνήθως διαμορφώνουν τόσο την συνολική προσφορά, όσο και την συνολική ζήτηση (βλ. ενδεικτικά Romer, 2001). Η προσφορά εξαρτάται από τις μεταβλητές που εισέρχονται στην παραγωγή (όπως η απασχόληση, το κεφάλαιο, η γη, η επιχειρηματικότητα και η τεχνολογία) και η ζήτηση από τις μεταβλητές που καθορίζουν τις ισορροπίες των αγορών προϊόντων (όπως η ιδιωτική κατανάλωση ή αποταμίευση, οι επενδύσεις, οι δημόσιες δαπάνες, οι φόροι, οι εξαγωγές-εισαγωγές) και χρήματος, καθώς και από και τον τρόπο με τον οποίο η οικονομία ενσωματώνεται στο διεθνές σύστημα. Εκ των παραγόντων αυτών, η αγορά χρήματος ίσως είναι δευτερευούσης σημασίας όταν η οικονομία λειτουργεί με σταθερές συναλλαγματικές ισοτιμίες (ή διατηρεί το ίδιο νόμισμα) με τους βασικούς της εταίρους, καθότι τότε η ισορροπία της απλώς προσαρμόζεται στις αλλαγές ισορροπίας που διενεργούνται στις αγορές προϊόντων.

Το εγχείρημα διενεργείται σε επίπεδο διοικητικών περιφερειών (η Ελλάς διαθέτει 13), με αποτέλεσμα ο αριθμός των ετησίων παρατηρήσεων να αυξάνει από ένδεκα σε 143. Το «ανυάτισμα» των παρατηρήσεων επιτρέπει την εμπειρική υλοποίηση της ανάλυσεως εντός κάποιου ανεκτού περιθωρίου βαθμών ελευθερίας.

2. ΔΙΑΘΕΣΙΜΑ ΣΤΑΤΙΣΤΙΚΑ ΣΤΟΙΧΕΙΑ

Η απουσία σε περιφερειακό επίπεδο (α) χρονοσειρών για τις εκτάσεις και χρήσεις γης, τις εισαγωγές-εξαγωγές, τα νομισματικά, καθώς και (β) καταναλωτικών-αποταμιευτικών και φορολογικών στοιχείων, αν και περιορίζει τη δυνατότητα εξετάσεως του ρόλου αυτών, ενδεχομένως απλουστεύει τα πράγματα, καθότι λόγω του περιορισμένου αριθμού παρατηρήσεων, αδυνατούμε να προσφύγουμε στη χρήση μεγάλου αριθμού ερμηνευτικών μεταβλητών (Tabachnick και Fidell, 1989· Bartlett κ.ά., 2001). Από την άλλη μεριά, τα μεγέθη υπό το στοιχείο (β) ενδέχεται να μπορούν να προσεγγιστούν σε κάποιο βαθμό από τα πληθυσμιακά μεγέθη και το συσχετιζόμενο μέγεθος του εργατικού δυναμικού· ενώ τα μεγέθη της επιχειρηματικότητας και τεχνολογίας συνήθως προκύπτουν από την εμπειρική ανάλυση, όταν είναι γνωστά τα υπόλοιπα μεγέθη (βλ. στοιχείο «Α» της εξισώσεως (1)).

Για τις ανάγκες της παρούσης ανάλυσεως, θα στηριχτούμε στα μεγέθη:

- Του ΑΕΠ (Y), των ακαθαρίστων επενδύσεων παγίου κεφαλαίου (I) και της απασχολήσεως (L) ανά παραγωγικό τομέα, όπως εκτιμώνται και παρέχονται από την ΕΛΣΤΑΤ.
- Των δημοσίων δαπανών (G), όπως εκτιμώνται και παρέχονται από την Γενική Γραμματεία Δημοσίων Επενδύσεων και Εθνικού Στρατηγικού Πλαισίου Αναφοράς (ΕΣΠΑ).
- Του διαθέσιμου συσσωρευμένου κεφαλαίου (K) ανά παραγωγικό τομέα όπως υπολογίστηκε κατά το έτος 1999 και από τις προσαυξήσεις των ποσών των I των επομένων ετών μείον τις αποσβέσεις. (Μετά από διαδοχικούς πειραματισμούς μεταξύ των πιθανών ποσοστών αποσβέσεων ανά παραγωγικό τομέα επιλέγησαν

οι εκτιμήσεις που συνάδουν με την επίτευξη του υψηλότερου R^2 .) Στη συνέχεια παρατίθεται η διαδικασία υπολογισμού του ποσού του 1999: Οι ακαθάριστες ιδιωτικές επενδύσεις των ετών 1988-96 και οι ακαθάριστες δημόσιες επενδύσεις των ετών 1988-95 παρέχονται από την περιοδική έκδοση *Επιλογή* της Alpha Bank, οι ακαθάριστες ιδιωτικές επενδύσεις του έτους 1997 εκτιμώνται από συνάρτηση τύπου $I_t=f(I_{t-1})$ που υπολογίζεται από τις ιδιωτικές επενδύσεις των προηγούμενων ετών, οι ακαθάριστες δημόσιες επενδύσεις του έτους 1997 παρέχονται από το ΚΕΠΕ (2000) και οι ακαθάριστες δημόσιες του έτους 1996 τίθενται στο επίπεδο του μέσου όρου των ετών 1995 και 1997. Η άθροιση των προαναφερομένων δημοσίων και ιδιωτικών μεγεθών επιτρέπει την προσέγγιση των ακαθαρσίων επενδύσεων των ετών 1998 και 1999: Οι πρώτες τίθενται στο επίπεδο του μέσου όρου των ετών 1996 και 2000 και οι δεύτερες στο επίπεδο του μέσου όρου των ετών 1998 και 2000.

Στο σημείο αυτό σημειώνεται ότι υπό κάποια πιο σύνθετα θεωρητικά υποδείγματα, οι διαθέσιμες μεταβλητές ίσως να θεωρούνται επαρκείς για τα θέματα που θα εξετάσουμε, ενώ σε εμπειρικό επίπεδο διαπιστώνεται ότι η ερμηνευτική ικανότης είναι μάλλον υψηλή. Συμπληρωματικά διευκρινίζεται ότι:

- Όλα τα στοιχεία είναι σε επίπεδο διοικητικής περιφέρειας, ήτοι: Ανατολικής Μακεδονίας και Δυτικής Θράκης (ΑΜΔΘ), Κεντρικής Μακεδονίας (ΚΜ), Δυτικής Μακεδονίας (ΔΜ), Θεσσαλίας (Θ), Ηπείρου (Η), Ιονίων νήσων (Ιο), Δυτικής Ελλάδος (ΔΕ), λοιπής (ή νοτίου-κεντρικής-ανατολικής) Πελοποννήσου (ΛΠ) (ήτοι, άνευ της Αχαΐας, Ηλείας και Τροιζηνίας, αλλά με τα δυτικά Γεράνεια), Αττικής (Ατ), λοιπής (ή κεντρικής) Στερεάς Ελλάδος (ΛΣΕ), νήσων Βορείου Αιγαίου (ΒΑ), νήσων Νοτίου Αιγαίου (ΝΑ), Κρήτης (Κρ).
- Τα στοιχεία που αφορούν στις μεταβλητές I , L , K παρέχονται ανά τομέα παραγωγικών δραστηριοτήτων. Οι τομείς αυτοί είναι: (1) ο πρωτογενής, (2α) ο κατασκευαστικός, (2β) ο λοιπός δευτερογενής, (3α) ο εφοδιαστικός-διανεμητικός (περιλαμβάνει και τα καταλύματα, την εστίαση, τις επικοινωνίες), (3β) ο επιχειρηματικό-υποστηρικτικός (περιλαμβάνει και την χρηματοπιστωτική διαμεσολάβηση), (3γ) ο λοιπός τριτογενής.
- Οι δημόσιες δαπάνες διενεργούνται μέσω (i) συγχρηματοδοτούμενων ευρωπαϊκών προγραμμάτων (G_{EE}), (ii) δράσεων που κρίνονται μη επιλέξιμες (δηλ., απεντάσσονται) από τα ευρωπαϊκά προγράμματα και ως εκ τούτου τελικώς καλύπτονται από εθνικούς πόρους (G_{EE}), καθώς και (iii) δράσεων που εξ αρχής σχεδιάστηκαν και υλοποιούνται αποκλειστικώς με εθνικούς πόρους. Οι τελευταίες παρακολουθούνται περιφερειακά μόνο κατά το ήμισυ και μάλιστα και από το 2002 ως το 2011. Ως εκ τούτου η ένταξη των στις αναλύσεις (ήτοι, στα εμπειρικά υποδείγματα που ακολουθούν) αναγκαστικώς επιφέρει μείωση του αριθμού των παρατηρήσεων. Για τον λόγο αυτό η προσθήκη των δαπανών του αμιγούς εθνικού προγράμματος εξετάζεται σε εναλλακτική εκδοχή.

3. ΔΙΑΠΙΣΤΩΣΕΙΣ ΕΝΔΟΠΕΡΙΦΕΡΕΙΑΚΩΝ ΚΑΙ ΔΙΑΠΕΡΙΦΕΡΕΙΑΚΩΝ ΕΠΙΔΡΑΣΕΩΝ

Αρχικώς διερευνάται μια δυνητικά χρήσιμη διάσταση της εσωτερικής λειτουργίας των περιφερειακών οικονομιών της χώρας. Πιο συγκεκριμένα, εξετάζονται σε επίπεδο στατιστικής σημαντικότητας 0,1% με ελέγχους αιτιότητας (συστηματικής προηγέσεως) κατά Granger (1969) μέσω του στατιστικού λογισμικού EViews οι σχέσεις των ροών των διαθεσίμων μεταβλητών. Βλ. Πίνακα 1. (Ο αριθμός των συνδυασμών από κοινού με τον αριθμό των παρατηρήσεων δεν επιτρέπει την εφαρμογή της πολυμεταβλητής εκδοχής, του λεγομένου *Vector Autoregressive (VAR) model*.) Από τα 32 υψηλής σημαντικότητας ζεύγη ευρημάτων που προκύπτουν, διαπιστώνεται ότι:

- Η διενέργεια επενδύσεων στον επιχειρηματικό-υποστηρικτικό τομέα σε πέντε περιπτώσεις προηγείται της μεταβολής του κεφαλαίου στον ίδιο τομέα άλλης περιφέρειας. Πρόκειται για μια ένδειξη της υπάρξεως ανταποκρίσεων ή/και διαχύσεων της οικονομικής δραστηριότητας.
- Αρκετές προηγήσεις εμφανίζονται στις επενδύσεις του επιχειρηματικό-υποστηρικτικού τομέα (σε οκτώ περιπτώσεις), καθώς και στην Θεσσαλία, στην Κεντρική Μακεδονία και στην Δυτική Μακεδονία (σε πέντε, πέντε, έξι περιπτώσεις, αντιστοίχως), ενώ αρκετές ανταποκρίσεις (πιθανές προσαρμογές σε προηγήσεις) εμφανίζονται στις μεταβολές κεφαλαίου του πρωτογενούς και του επιχειρηματικό-υποστηρικτικού τομέα (σε επτά και εννέα περιπτώσεις, αντιστοίχως), στην Δυτική Μακεδονία, στην Θεσσαλία και στις νήσους του Νοτίου Αιγαίου (έξι, πέντε και πέντε περιπτώσεις, αντιστοίχως).
- Υπάρχουν ενδείξεις διαδοχικών διαχύσεων (α) από την Θεσσαλία στην Ανατολική Μακεδονία και Δυτική Θράκη και εν συνεχεία στην Κρήτη και (β) από την Κρήτη στις νήσους του Νοτίου Αιγαίου και εν συνεχεία στην Δυτική Μακεδονία.

Οι μεταβολές της εργασίας (ιδίως στον πρωτογενή, δευτερογενή και επιχειρηματικό-υποστηρικτικό τομέα), καθώς και οι ροές των δημοσίων δαπανών (ιδίως των αμιγών εθνικών) είναι μάλλον άκαμπτες προς τα άλλα εξεταζόμενα μεγέθη, με αποτέλεσμα οι επιδράσεις και ανταποκρίσεις (προσαρμογές) αυτών σε άλλα στοιχεία της συνολικής προσφοράς και ζήτησης εντός ή εκτός περιφέρειας, να συνιστούν μάλλον την εξαιρέση παρά τον κανόνα.

4. ΕΜΠΕΙΡΙΚΗ ΑΝΑΛΥΣΗ ΤΟΥ ΠΕΡΙΦΕΡΕΙΑΚΟΥ ΑΕΠ

Εν συνεχεία επιχειρείται, μέσω του στατιστικού λογισμικού STATA, η εμπειρική εξήγηση του Y . Στο βαθμό που:

- η παραγωγή του (κλαδικού ή συνολικού) προϊόντος εκφράζεται ως μη γραμμική συνάρτηση των εισροών K και L , λαμβάνουσα την μορφή, $Y=AL^aK^b$, γνωστής και ως συνάρτηση των Cobb και Douglas (1928), όπου «a» και «b» οι συντελεστές της L και του K , αντιστοίχως, και «A» ο δείκτης τεχνολογίας και επιχειρηματικότητας που καθορίζει την συνολική παραγωγικότητα των λοιπών

συντελεστών και ο οποίος συχνά δύναται να διαφοροποιείται διαχρονικά (Frasier, 2002).

- η πλευρά της ζήτησης λαμβάνει την μορφή μιας συνάρτησης $Y=Y(I,G)$, με κάποιο αυτόνομο όρο «ε» και συντελεστές «α» και «γ» για τις δύο μεταβλητές, μια συναρτησιακή έκφραση μεταξύ της εξαρτημένης και των ανεξαρτήτων μεταβλητών θα μπορούσε να είναι της μορφής:

$$Y=A L^a K^b e^{\varepsilon+\iota I+\gamma G} \quad (1)$$

$$\text{ή } Y=AL^a K^b I^\iota G^\gamma, \quad (2)$$

αν, για την ενσωμάτωση των όρων I και G υπό την μορφή Cobb-Douglas, δεχθούμε μια παραλλαγή του υποδείγματος του Barro (1990). Σημειώνεται ότι σε αμφότερες περιπτώσεις επιτρέπεται οι τρέχουσες I να έχουν διαφορετική επίδραση στο Y από ότι το K (δηλ., από ότι ο αθροιστικός μετασχηματισμός των παρελθουσών I). Επίσης, αν και στα απλά υποδείγματα τα «ι» και «γ» τίθενται ως ισομεγέθη, εν προκειμένω, στο βαθμό που εξετάζουμε ποικίλες τομεακές επενδύσεις και διαφορετικούς τύπους δημοσίων δαπανών, το ζήτημα του σχετικού μεγέθους αφήνεται να απαντηθεί από τα εμπειρικά ευρήματα.

Προς διευκόλυνση των παλινδρομήσεων, οι ανωτέρω σχέσεις μετασχηματίζονται, αντιστοίχως, σε:

$$\ln Y = \ln A + \varepsilon + a \times \ln L + b \times \ln K + \iota \times I + \gamma \times G \quad (1')$$

$$\ln Y = \ln A + a \times \ln L + b \times \ln K + \iota \times \ln I + \gamma \times \ln G \quad (2')$$

και, προκειμένου η επίδραση κάθε ερμηνευτικής μεταβλητής να είναι όσο σαφέστερη γίνεται, το K καθίσταται ορθογώνιο της L , η μεταβλητή των I καθίσταται ορθογώνια της L και του K και η μεταβλητή των G ορθογώνια όλων των προηγούμενων. Οι δε σειρές του ΑΕΠ και των άλλων διαθεσίμων ερμηνευτικών μεταβλητών θεωρούνται στάσιμες εφόσον το ΑΕΠ παλινδρομείται με την μέθοδο των ελαχίστων τετραγώνων με κάποια άλλη σειρά και τα κατάλοιπα αποδεικνύονται στάσιμα, με αποτέλεσμα οι υπό εξέταση σειρές να εμφανίζονται ως συνολοκληρωμένες. Στο Παράρτημα εξετάζονται τα ζεύγη σειρών ανά περιφέρεια μέσω του ελέγχου των Dickey και Fuller (1981).

Τα εμπειρικά ευρήματα που προκύπτουν με την μέθοδο των σταθερών επιδράσεων (*fixed effects*) στο χώρο και το χρόνο παρατίθενται στον Πίνακα 2. Η συγκεκριμένη μέθοδος διενεργείται βάσει της υποθέσεως ότι τα στοιχεία που αφορούν στην ίδια περιοχή ή χρονική περίοδο διαφοροποιούνται κάπως από τα άλλα στοιχεία. Εξ αυτών διαπιστώνεται ότι, *ceteris paribus*:

- α) Η επιχειρηματικότητα, η τεχνολογία και οι άλλοι παράγοντες που διαθέτουν αυτόνομα στοιχεία και συγκροτούν τον σταθερό όρο της συναρτήσεως σχηματισμού του Y (ήτοι, $\ln A + \varepsilon$, σε όρους της σχέσης (1')), συνολικά εμφανίζουν θετικό μέγεθος.
- β) Η ποσοστιαία μεταβολή τόσο της L στον (i) πρωτογενή, (ii) κατασκευαστικό και (iii) λοιπό δευτερογενή τομέα, όσο και του K στον (iv) επιχειρηματικό-υποστηρικτικό και (v) λοιπό τριτογενή τομέα, φαίνεται να επιφέρει υποδεέστερη ποσοστιαία μεταβολή του Y .
- γ) Η ποσοστιαία μεταβολή τόσο της L στον (i) επιχειρηματικό-υποστηρικτικό και (ii) λοιπό τριτογενή τομέα όσο και του K στον (iii) πρωτογενή, (iv)

- κατασκευαστικό, (v) λοιπό δευτερογενή και (vi) εφοδιαστικό-διανεμητικό τομέα, φαίνεται να επιφέρει υπέρτερη ποσοστιαία μεταβολή του Y.
- (δ) Η ποσοστιαία μεταβολή της L στον εφοδιαστικό-διανεμητικό τομέα φαίνεται να επιφέρει διαφοροποιημένες ποσοστιαίες μεταβολές του Y στα δύο υποδείγματα: υποδεέστερη στην περίπτωση της σχέσης 1' και υπέρτερη στην περίπτωση της σχέσης 2'.
- (ε) Βάσει της σχέσης 1', η διενέργεια (i) I στον πρωτογενή τομέα, δευτερογενή και τριτογενή τομέα, καθώς και (ii) G μέσω κοινοτικών προγραμμάτων και (iii) G μέσω δράσεων που (τελικά) κρίνονται μη επιλέξιμες για κοινοτική χρηματοδότηση (με αποτέλεσμα να καλύπτονται από εθνικούς πόρους), φαίνεται να έχει μεγάλη θετική ποσοστιαία επίδραση στο Y (η προσθήκη μίας μονάδας I ή G επιφέρει ποσοστιαία αύξηση του Y μεγαλύτερη του 100%)· ενώ (iv) G μέσω του αμιγούς (μη συγχρηματοδοτούμενου) εθνικού προγράμματος δημοσίων δαπανών φαίνεται να έχει μικρή θετική ποσοστιαία επίδραση (η προσθήκη μίας μονάδας I ή G επιφέρει ποσοστιαία αύξηση του Y μικρότερη του 100%).
- (ς) Βάσει της σχέσης 2', η ποσοστιαία μεταβολή των (i) I στον κατασκευαστικό τομέα και (ii) G μέσω του αμιγούς (μη συγχρηματοδοτούμενου) εθνικού προγράμματος δημοσίων δαπανών φαίνεται να επιφέρει υποδεέστερη ποσοστιαία μεταβολή του Y, ενώ η ποσοστιαία μεταβολή των (iii) I στον πρωτογενή τομέα, λοιπό δευτερογενή και τριτογενή τομέα, καθώς και (iv) G μέσω κοινοτικών προγραμμάτων, καθώς και δράσεων που κρίνονται μη επιλέξιμες για κοινοτική χρηματοδότηση φαίνεται να επιφέρει υπέρτερη ποσοστιαία αύξηση του Y.

Ωστόσο, από τον έλεγχο του Hauseman (1978), που συγκρίνει τα αποτελέσματα με αυτά που προκύπτουν μέσω της μεθόδου των τυχαίων επιδράσεων (*random effects*), συνάγεται ότι η προηγηθείσα προσέγγιση χαρακτηρίζεται από μειωμένη καταλληλότητα, με αποτέλεσμα η περαιτέρω ή διεξοδικότερη διερεύνηση της κατάστασης να αποκαλύπτει ενδιαφέρουσα ετερογένεια όσον αφορά στους συντελεστές, καθώς και στο «Α» και στο «ε» ανά παραγωγικό τομέα ή είδος δαπάνης από περιφέρεια σε περιφέρεια. Ειδικότερα, από τον Πίνακα 3 διαπιστώνεται ότι η επιχειρηματικότητα, η τεχνολογία και άλλοι παράγοντες που διαθέτουν αυτόνομα στοιχεία και συγκροτούν τον σταθερό όρο της συνάρτησης στις περιοχές αναφοράς (Αττική, Α. Μακεδονία και Δ. Θράκη, Ιονίους νήσους και νήσους του Β. Αιγαίου) κατά την περίοδο αναφοράς (2001-02) αθροίζουν σε θετικό αριθμό μικρότερο της μονάδας, με αποτέλεσμα ο φυσικός λογάριθμος να λαμβάνει αρνητική τιμή. Το μέγεθος προοδευτικά αυξάνει (α) κατά την διάρκεια των ετών 2003-6 και ακόμα περισσότερο κατά την διάρκεια των ετών 2007-11, (β) στην Ήπειρο, (γ) κατά μήκος μιας γεωγραφικής ζώνης που εκτείνεται από την Κ. και Δ. Μακεδονία στην Θεσσαλία, Δ. Ελλάδα και λοιπή Πελοπόννησο ως την Κρήτη, καθώς και (δ) στην κεντρική Στ. Ελλάδα και στις νήσους του Β. Αιγαίου. (Οι δύο τελευταίες συνιστούσαν τις λεγόμενες περιφέρειες σταδιακής εισόδου στον επιχειρησιακό σχεδιασμό του 2007-13.)

Η ποσοστιαία μεταβολή της L επιφέρει υπέρτερη ποσοστιαία μεταβολή του Y, προοδευτικά μεγαλύτερη, όταν εκδηλώνεται: (α) στον κατασκευαστικό τομέα της Ηπείρου, (β) στον επιχειρηματικό-υποστηρικτικό τομέα και στον τομέα των λοιπών υπηρεσιών της Α. Μακεδονίας και Δ. Θράκης ή στον τομέα των λοιπών υπηρεσιών της Αττικής και των νήσων του Β. Αιγαίου ή στον εφοδιαστικό-διανεμητικό τομέα των Ιονίων νήσων, (γ) στους λοιπούς συνδυασμούς παραγωγικών τομέων και περιοχών. Συγχρόνως, στο υφιστάμενο σύστημα λειτουργίας της οικονομίας η ποσοστιαία μεταβολή του K επιφέρει (i) υπέρτερη ποσοστιαία μεταβολή του Y όταν εκδηλώνεται στον πρωτογενή τομέα της Κ. Μακεδονίας και (ακόμα περισσότερο) στον κατασκευαστικό τομέα σε όλη την επικράτεια, (ii) υποδεέστερη ποσοστιαία μεταβολή του Y όταν εκδηλώνεται στους λοιπούς συνδυασμούς παραγωγικών τομέων και περιοχών.

Από την άλλη πλευρά, βάσει της:

- Σχέσης 1' η διενέργεια I στον πρωτογενή τομέα και G μέσω κοινοτικών προγραμμάτων (ακόμα και όταν τελικώς οι συγκεκριμένες G κρίνονται μη επιλέξιμες) φαίνεται ότι έχει μεγάλη θετική ποσοστιαία επίδραση στο Y, ιδίως στις περιπτώσεις δράσεων που χρηματοδοτούνται από το 3^ο Κοινοτικό Πλαίσιο Στήριξης (ΚΠΣ), το 1^ο ΕΣΠΑ ή άλλα προγράμματα στις νήσους του Β. Αιγαίου ή τυχαίνει να απεντάσσονται (ως μη επιλέξιμες) και να καλύπτονται εξ ολοκλήρου από εθνικούς πόρους στην Δ. Ελλάδα· ενώ η διενέργεια I στον δευτερογενή και τριτογενή τομέα φαίνεται να έχει είτε μεγάλη είτε μικρή θετική ποσοστιαία επίδραση στο Y και η διενέργεια G μέσω του αμιγούς (μη συγχρηματοδοτούμενου) εθνικού προγράμματος δημοσίων δαπανών φαίνεται να έχει μικρή θετική ποσοστιαία επίδραση στο Y.
- Σχέσεως 2' η ποσοστιαία μεταβολή των I στον πρωτογενή τομέα, στους λοιπούς τομείς, καθώς και των G μέσω κοινοτικών προγραμμάτων (ακόμα και αν τελικώς καταστούν μη επιλέξιμες) φαίνεται να επιφέρει υπέρτερη ποσοστιαία μεταβολή του Y, ενώ η ποσοστιαία μεταβολή των G μέσω του αμιγούς εθνικού προγράμματος φαίνεται να επιφέρει υποδεέστερη ποσοστιαία μεταβολή του Y.

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Οι περιφέρειες της χώρας λειτουργούν ετερόκλητα ως προς την απόδοση αρκετών βασικών παραγόντων που καθορίζουν το ΑΕΠ. Στο βαθμό που τόσο οι περιφέρειες όσο και οι παραγωγικοί τομείς συνιστούν κρίσιμα στοιχεία (τμήματα) της οικονομίας, οι ιδιαιτερότητες των οποίων συχνά συσκοτίζονται στα συνολικά μεγέθη (σε επίπεδο επικρατείας), θα ήταν σκόπιμο να εξετάζονται και να λαμβάνονται υπ' όψη στη χάραξη αναπτυξιακής πολιτικής, ιδίως τώρα που: (α) οι διαθέσιμοι πόροι δεν αρκούν για εκτεταμένες οριζόντιες παρεμβάσεις και (β) αποτελεσματικό θα είναι το χωρικά στοχευμένο και κλαδικά προσαρμοσμένο. Απώτεροι στόχοι πρέπει να είναι: η βελτίωση της αποδοτικότητας των εισροών (με εκπαίδευση και κατάρτιση, καλύτερη ή πληρέστερη αξιοποίηση των εισροών), η προώθηση των εκροών, η ανάληψη δράσεων αξιολόγησης-ιεράρχησης-ενίσχυσης των δυνητικών πλεονεκτημάτων σε τοπικό-περιφερειακό-διαπεριφερειακό επίπεδο, η διεύρυνση της αλυσίδας αξίας και η

προσθήκη αξίας εγχώρια. Τέλος, διαφαίνεται ότι το αμιγές εθνικό πρόγραμμα δημοσίων έργων χρήζει βελτιώσεων, προκειμένου να συμβάλλει καλύτερα στην οικονομική ανάπτυξη.

ABSTRACT

The paper studies (a) Granger causalities involving labor, capital and investment variables across six sectors, and public spending variables at the regional level, along with (b) the effects of the aforementioned factors on regional GDP from 2000 to 2011 in Greece, in order to aid in development policy planning.

ΑΝΑΦΟΡΕΣ

Alpha Bank. *Επιλογή*. Τεύχη διαφόρων ετών.

Barro R.J. 1990. «Government Spending in a Simple Model of Endogenous Growth». *Journal of Political Economy*, 98(5.2), 103-125.

Bartlett J.E. II, Kotrlik J.W. και Higgins C.C. 2001. «Organizational Research: Determining Appropriate Sample Size in Survey Research». *Information Technology, Learning, and Performance Journal*, 19 (1), 43-50.

Cobb C.W. και Douglas P.H. 1928. «*A Theory of Production*». *American Economic Review* 18 (supplement): 139–165.

Dickey D. A. και Fuller W. A. 1981. «Likelihood Ratio Tests for Autoregressive Time Series with a Unit Root». *Econometrica*, 49.4, 1057-1072.

Frasier I. 2002. «The Cobb-Douglas Production Function: An Antipodean Defense?». *Economic Issues* 7 (1), 39-58.

Granger C. W. J. 1969. «Investigating Causal Relations by Econometric Models and Cross-spectral Methods». *Econometrica* 37 (3): 424–438.

Hausman J. A. 1978. «Specification tests in econometrics». *Econometrica* 46: 1251–1271.

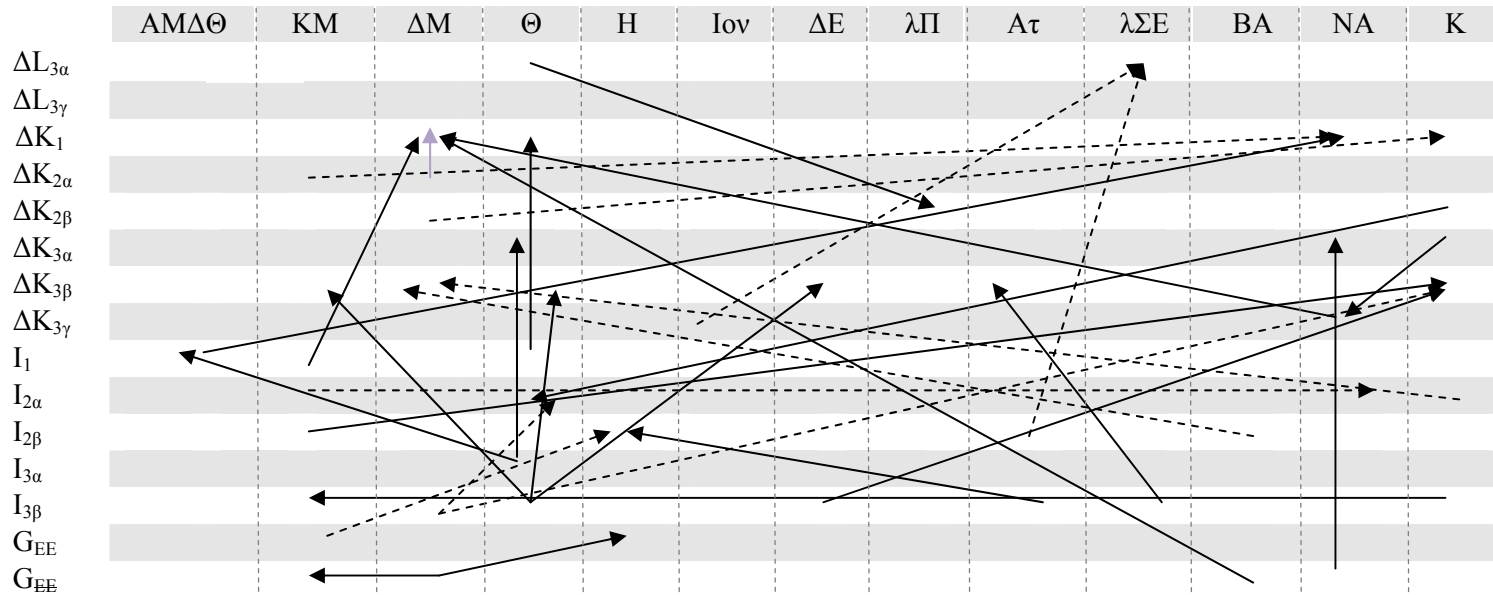
ΚΕΠΕ. 2000. *Η κοινωνικοοικονομική ταυτότητα των υδατικών διαμερισμάτων*. Υπό Β.Δεδεγιάν, Δ.Κατοχιανού και Α.Λαμπροπούλου. Αθήνα: Σειρά Εκθέσεων, 30.

Romer D. 2001. *Advanced Macroeconomics*. 2^η έκδοση. New York: McGraw Hill.

Tabachnick B.G. και Fidell L.S. 1989. *Using multivariate statistics*, 2^η έκδοση. New York: Harper Collins.

Πίνακας 1

Παράθεση ελέγχων αιτιότητας κατά Granger των μεταβολών (Δ) σε απασχόληση (L), κεφάλαιο (K), ΑΕΠ (Y) και των ροών επενδύσεων (I) και δημοσίων δαπανών (G) μεταξύ 2001 και 2011, με μια και δύο χρονικές υστερήσεις (βλ. συνεχή και διακεκομμένα βέλη, αντιστοίχως) και πιθανότητα σφάλματος 0,1%



Υπόμνημα:

- Τομείς: (1) πρωτογενής, (2α) κατασκευαστικός, (2β) λοιπός δευτερογενής, (3α) εφοδιαστικός-διανεμητικός, (3β) επιχειρηματικό-υποστηρικτικός, (3γ) λοιπός τριτογενής.
- Δαπάνες: (EE) κοινοτικές, (EE) αρχικός κοινοτικός όμως τελικός μη επιλέξιμες (απενταγμένες).

Πίνακας 2

Εκτίμηση με σταθερές επιδράσεις (*fixed effects*) χώρου και χρόνου των επιδράσεων εργασίας, κεφαλαίου, επενδύσεων, δημ. δαπανών στο περιφερειακό ΑΕΠ, 2001-11

ΑΕΠ ^{αγ}	Υπόδειγμα 1 (εξίσωση 1')				Υπόδειγμα 2 (εξίσωση 2')			
	βασική παλινδρόμηση		εναλλακτική εκδοχή		βασική παλινδρόμηση		εναλλακτική εκδοχή	
	συντελ.	p val.	συντελ.	p val.	συντελ.	p val.	συντελ.	p val.
Σταθ. όρος	10,04	0,000	7,77	0,002	9,29	0,000	7,42	0,000
Απασχόληση ^{βγ}								
Τομέας 1	-0,22	0,000	-0,16	0,017	-0,18	0,000	-0,14	0,002
Τομέας 2α	-0,14	0,005	-0,04	0,020	-0,04	0,000	-0,03	0,020
Τομέας 2β	-0,04	0,016	-0,08	0,160	-0,12	0,000	-0,07	0,074
Τομέας 3α	-0,14	0,157	-0,07	0,492	0,10	0,146	0,00	0,991
Τομέας 3β	0,05	0,158	0,04	0,195	0,05	1,045	0,06	0,010
Τομέας 3γ	0,66	0,000	0,63	0,000	0,60	0,000	0,54	0,000
Κεφάλαιο ^{αγε}								
Τομέας 1	0,20	0,012	0,21	0,018	0,19	0,001	0,21	0,001
Τομέας 2α	0,07	0,346	0,27	0,004	0,23	0,000	0,17	0,013
Τομέας 2β	0,28	0,001	0,03	0,725	0,02	0,702	0,02	0,663
Τομέας 3α	0,13	0,140	0,01	0,909	0,10	0,084	0,14	0,061
Τομέας 3β	-1,23	0,000	-0,95	0,000	-1,15	0,000	-0,96	0,000
Τομέας 3γ	-0,42	0,000	-0,39	0,001	-0,32	0,000	-0,27	0,001
Ακαθ. επενδ. ^{αδε}								
Τομέας 1	477,39	0,004	443,68	0,015	0,09	0,000	0,09	0,000
Τομέας 2α					-0,02	0,000	-0,02	0,000
Τομέας 2β					0,04	0,002	0,03	0,005
Τομέας 3					0,05	0,017	0,04	0,082
Τομείς 2-3	12,93	0,011	7,51	0,191				
Δημ. Δαπάνες ^{αδε}								
Κοινοτικές	141,37	0,006	258,64	0,000	0,07	0,000	0,10	0,000
Απενταγμ.	84,10	0,133	47,95	0,427	0,01	0,570	0,01	0,696
Αμειγείς εθν.			-70,70	0,153			-0,00	0,818
R ²	47,24%		78,82%		78,82%		78,82%	
N	143		130		142		129	

Σημειώσεις: Λόγω του λογαριθμικού μετασχηματισμού του ΑΕΠ τα αρνητικά πρόσημα δεν υποδηλώνουν αρνητική, αλλά μικρή ποσοστιαία επίδραση στο ΑΕΠ.

Για ονομασίες τομέων βλ. Πίνακα 1.

α Σε εκατ. ευρώ, τρέχουσες τιμές.

β Σε χιλιάδες μονίμους κατοίκους.

γ. Σε λογαριθμική μορφή.

δ. Σε λογαριθμική μορφή στο υπόδειγμα 2.

ε. Μετασχηματισμένη μεταβλητή ορθογώνια προς τις μεταβλητές που παρατίθενται πιο πάνω.

Πίνακας 3

Εκτίμηση με ανθεκτικούς (*robust*) εκτιμητές των επιδράσεων εργασίας, κεφαλαίου, επενδύσεων, δημ. δαπανών, χωρικών και ετησίων παραγόντων στο περιφερειακό ΑΕΠ, 2001-11

ΑΕΠ ^{αγ}	Υπόδειγμα 1 (εξίσωση 1')				Υπόδειγμα 2 (εξίσωση 2')			
	βασική παλινδρόμηση συντελ.	εναλλακτική εκδοχή p val.	βασική παλινδρόμηση συντελ.	εναλλακτική εκδοχή p val.	βασική παλινδρόμηση συντελ.	εναλλακτική εκδοχή p val.	βασική παλινδρόμηση συντελ.	εναλλακτική εκδοχή p val.
Σταθερός όρος	-4,43	0,000	-4,41	0,000	-4,43	0,000	-4,41	0,000
Περιφερειακές διαφοροποιήσεις								
Ήπειρος	2,72	0,000	2,77	0,000	2,72	0,000	2,78	0,000
Κ. και Δ. Μακεδονία, Θεσσαλία Δ.Ελλάς, λοιπή Πελοπόννησος, Κρήτη	3,36	0,000	3,42	0,000	3,36	0,000	3,43	0,000
Κεντρική Στ. Ελλάς, Ν. Αιγαίο Λοιπή χώρα (περιοχή αναφοράς)	3,52	0,000	3,56	0,000	3,52	0,000	3,57	0,000
Χρονικές διαφοροποιήσεις								
2001-02 (περίοδος αναφοράς)								
2003-06	0,17	0,000	0,15	0,000	0,17	0,000	0,16	0,000
2007-11	0,30	0,000	0,28	0,000	0,30	0,000	0,28	0,000
Απασχόληση ^{βγ}								
Τομέας 2α σε Ήπειρο	0,05	0,000	0,05	0,000	0,05	0,000	0,05	0,000
Τομέας 3β-γ σε Α. Μακεδονία και Δ. Θράκη	0,32	0,000	0,32	0,000	0,32	0,000	0,32	0,000
Τομέας 3γ σε Αττική και Ν. Αιγαίο	0,34	0,000	0,34	0,000	0,34	0,000	0,35	0,000
Τομέας 3α σε Ιονίους νήσους	0,36	0,000	0,36	0,000	0,36	0,000	0,36	0,000
Αλλού	0,81	0,000	0,80	0,000	0,81	0,000	0,81	0,000

Πίνακας 3 (συνέχεια)									
	Υπόδειγμα 1 (εξίσωση 1')				Υπόδειγμα 2 (εξίσωση 2')				
	βασική παλινδρόμηση συντελ.	εναλλακτική εκδοχή p val.	βασική παλινδρόμηση συντελ.	εναλλακτική εκδοχή p val.	βασική παλινδρόμηση συντελ.	εναλλακτική εκδοχή p val.	βασική παλινδρόμηση συντελ.	εναλλακτική εκδοχή p val.	
ΑΕΠ ^{αγ}									
Κεφάλαιο ^{αγε}									
Τομέας 1 στην Κ.Μακεδονία	0,03	0,000	0,03	0,000	0,03	0,000	0,03	0,000	
Τομέας 2α	0,12	0,002	0,11	0,002	0,12	0,001	0,11	0,001	
Αλλού	-0,15	0,012	-0,17	0,004	-0,15	0,007	-0,17	0,003	
Ακαθ. επενδύσεις ^{αδε}									
Τομέας 1	427,42	0,001	364,87	0,001	0,05	0,001	0,04	0,005	
Αλλού	-0,97	0,721	1,52	0,483	0,03	0,089	0,03	0,107	
Δημόσιες δαπάνες ^{αδε}									
• Κοινοτικές σε νήσ. Β.Αιγαίου	685,14	0,008	844,03	0,000					
• Απεντασσόμενες των ανωτέρω σε Δ. Ελλάδα	155,67	0,003	217,30	0,000					
• Κοινοτικές και απεντασσόμενες (πλην των δύο ανωτέρω στο υπόδειγμα 1)	11,79	0,487	44,58	0,012	0,01	0,253	0,02	0,063	
• Αμιγούς εθνικού προγράμματος			-61,15	0,000			-0,02	0,179	
R ²	99,73%		99,77%		99,74%		99,76%		
N	143		130		143		130		

Παράρτημα

Έλεγχοι Dickey-Fuller στις χρονοσειρές του ΑΕΠ και των ερμηνευτικών μεταβλητών

ΑΕΠ	ΑΜΔΘ	ΚΜ	ΔΜ	Θ	Η	Ιο	ΔΕ	ΛΠ	ΣΕ	Α	ΒΑ	ΝΑ	Κ
p-value: 5%													
ΔL ₁	0,03	0,05	0,11	0,01	0,24	0,12	0,08	0,14	0,12	0,08	0,26	0,06	0,03
ΔL _{2α}	0,04	0,08	0,26	0,07	0,11	0,09	0,11	0,18	0,08	0,05	0,10	0,12	0,13
ΔL _{2β}	0,15	0,09	0,17	0,08	0,12	0,12	0,11	0,13	0,11	0,13	0,11	0,16	0,14
ΔL _{3α}	0,14	0,15	0,14	0,08	0,12	0,06	0,11	0,14	0,05	0,18	0,10	0,06	0,09
ΔL _{3β}	0,15	0,14	0,12	0,08	0,10	0,10	0,10	0,15	0,12	0,21	0,13	0,12	0,11
ΔL _{3γ}	0,04	0,11	0,06	0,04	0,23	0,17	0,08	0,02	0,03	0,13	0,10	0,24	0,05
ΔK ₁	0,45	0,07	0,22	0,17	0,01	0,14	0,15	0,13	0,11	0,06	0,03	0,05	0,30
ΔK _{2α}	0,00	0,02	0,12	0,10	0,14	0,20	0,33	0,15	0,09	0,01	0,10	0,13	0,18
ΔK _{2β}	0,11	0,04	0,05	0,06	0,14	0,14	0,18	0,06	0,03	0,09	0,07	0,08	0,20
ΔK _{3α}	0,15	0,19	0,12	0,07	0,12	0,05	0,33	0,03	0,06	0,06	0,03	0,03	0,10
ΔK _{3β}	0,07	0,02	0,04	0,04	0,07	0,10	0,17	0,01	0,02	0,01	0,05	0,05	0,03
ΔK _{3γ}	0,10	0,05	0,04	0,01	0,09	0,27	0,08	0,04	0,03	0,09	0,11	0,04	0,09
I ₁	0,11	0,16	0,10	0,12	0,15	0,04	0,06	0,05	0,04	0,17	0,15	0,09	0,08
I _{2α}	0,07	0,06	0,06	0,06	0,17	0,10	0,12	0,10	0,06	0,15	0,03	0,05	0,07
I _{2β}	0,09	0,05	0,12	0,13	0,19	0,14	0,20	0,10	0,13	0,05	0,11	0,11	0,13
I _{3α}	0,21	0,12	0,12	0,06	0,17	0,05	0,11	0,10	0,09	0,11	0,16	0,17	0,12
I _{3β}	0,11	0,04	0,09	0,05	0,06	0,04	0,10	0,07	0,02	0,06	0,04	0,05	0,08
I _{3γ}	0,12	0,18	0,06	0,06	0,11	0,09	0,06	0,06	0,12	0,13	0,11	0,08	0,02
G _{EE}	0,26	0,36	0,26	0,05	0,16	0,03	0,18	0,02	0,08	0,06	0,49	0,32	0,22
G _{EE}	0,28	0,41	0,07	0,06	0,04	0,03	0,03	0,12	0,01	0,06	0,12	0,10	0,04
G _N	0,16	0,15	0,13	0,19	0,24	0,23	0,20	0,18	0,21	0,03	0,09	0,06	0,12

p-value: 15%													
ΔL ₁	0,03	0,05	0,11	0,01	0,24	0,12	0,08	0,14	0,12	0,08	0,26	0,06	0,03
ΔL _{2α}	0,04	0,08	0,26	0,07	0,11	0,09	0,11	0,18	0,08	0,05	0,10	0,12	0,13
ΔL _{2β}	0,15	0,09	0,17	0,08	0,12	0,12	0,11	0,13	0,11	0,13	0,11	0,16	0,14
ΔL _{3α}	0,14	0,15	0,14	0,08	0,12	0,06	0,11	0,14	0,05	0,18	0,10	0,06	0,09
ΔL _{3β}	0,15	0,14	0,12	0,08	0,10	0,10	0,10	0,15	0,12	0,21	0,13	0,12	0,11
ΔL _{3γ}	0,04	0,11	0,06	0,04	0,23	0,17	0,08	0,02	0,03	0,13	0,10	0,24	0,05
ΔK ₁	0,45	0,07	0,22	0,17	0,01	0,14	0,15	0,13	0,11	0,06	0,03	0,05	0,30
ΔK _{2α}	0,00	0,02	0,12	0,10	0,14	0,20	0,33	0,15	0,09	0,01	0,10	0,13	0,18
ΔK _{2β}	0,11	0,04	0,05	0,06	0,14	0,14	0,18	0,06	0,03	0,09	0,07	0,08	0,20
ΔK _{3α}	0,15	0,19	0,12	0,07	0,12	0,05	0,33	0,03	0,06	0,06	0,03	0,03	0,10
ΔK _{3β}	0,07	0,02	0,04	0,04	0,07	0,10	0,17	0,01	0,02	0,01	0,05	0,05	0,03
ΔK _{3γ}	0,10	0,05	0,04	0,01	0,09	0,27	0,08	0,04	0,03	0,09	0,11	0,04	0,09
I ₁	0,11	0,16	0,10	0,12	0,15	0,04	0,06	0,05	0,04	0,17	0,15	0,09	0,08
I _{2α}	0,07	0,06	0,06	0,06	0,17	0,10	0,12	0,10	0,06	0,15	0,03	0,05	0,07
I _{2β}	0,09	0,05	0,12	0,13	0,19	0,14	0,20	0,10	0,13	0,05	0,11	0,11	0,13
I _{3α}	0,21	0,12	0,12	0,06	0,17	0,05	0,11	0,10	0,09	0,11	0,16	0,17	0,12
I _{3β}	0,11	0,04	0,09	0,05	0,06	0,04	0,10	0,07	0,02	0,06	0,04	0,05	0,08
I _{3γ}	0,12	0,18	0,06	0,06	0,11	0,09	0,06	0,06	0,12	0,13	0,11	0,08	0,02
G _{EE}	0,26	0,36	0,26	0,05	0,16	0,03	0,18	0,02	0,08	0,06	0,49	0,32	0,22
G _{EE}	0,28	0,41	0,07	0,06	0,04	0,03	0,03	0,12	0,01	0,06	0,12	0,10	0,04
G _N	0,16	0,15	0,13	0,19	0,24	0,23	0,20	0,18	0,21	0,03	0,09	0,06	0,12

Σημείωση: Με σκίαση οι περιπτώσεις για τις οποίες η υπόθεση μη στασιμότητας των καταλοίπων απορρίπτεται, με αποτέλεσμα η χρονοσειρά να θεωρείται συνολοκληρωμένη με το περιφερειακό ΑΕΠ.

ΕΝΑΣ ΑΚΟΛΟΥΘΙΑΚΟΣ ΕΛΕΓΧΟΣ ΓΙΑ ΤΗΝ ΑΞΙΟΛΟΓΗΣΗ ΤΗΣ ΣΥΜΦΩΝΙΑΣ ΜΕΤΑΞΥ ΒΑΘΜΟΛΟΓΗΤΩΝ

Α. Σαχλάς¹, Σ. Μπερσίμης², S. Chakraborti³

¹Τμήμα Στατιστικής & Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς
asachlas@uop.gr

²Τμήμα Στατιστικής & Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς
sbersim@unipi.gr

³Department of Information Systems, Statistics and Management Science, University of
Alabama
schakrab@cba.ua.edu

ΠΕΡΙΛΗΨΗ

Η αξιολόγηση της συμφωνίας μεταξύ δύο ή περισσότερων βαθμολογητών είναι ιδιαίτερα σημαντική στην ιατρική επιστήμη. Οι υπάρχουσες τεχνικές, οι οποίες ασχολούνται με κατηγορικά δεδομένα, βασίζονται σε πίνακες συνάφειας. Αυτό αποτελεί συχνά εμπόδιο στην πράξη, καθώς θα πρέπει να περιμένουμε μεγάλο χρονικό διάστημα για να συλλέξουμε το κατάλληλο μέγεθος του δείγματος για την κατασκευή του πίνακα συνάφειας. Στην εργασία αυτή, προτείνουμε έναν μη παραμετρικό ακολουθιακό έλεγχο για την αξιολόγηση της συμφωνίας, ο οποίος μπορεί να εφαρμοστεί ενώ συλλέγονται σταδιακά τα δεδομένα και δεν απαιτεί πίνακα συνάφειας, διευκολύνοντας έτσι την γρήγορη αξιολόγηση της συμφωνίας. Ο προτεινόμενος έλεγχος βασίζεται στο σωρευτικό άθροισμα του αριθμού των διαφωνιών μεταξύ των δύο βαθμολογητών και μια κατάλληλη στατιστική συνάρτηση που αντιπροσωπεύει τον χρόνο αναμονής μέχρι το άθροισμα να υπερβεί ένα προκαθορισμένο όριο. Αντιμετωπίζουμε τις περιπτώσεις του ελέγχου της συμφωνίας δύο βαθμολογητών ως προς ένα ή περισσότερα χαρακτηριστικά χρησιμοποιώντας δύο ή περισσότερες κατηγορίες ταξινόμησης και την περίπτωση όπου οι δύο βαθμολογητές διαφωνούν σε μεγάλο βαθμό. Η αριθμητική διερεύνηση έδειξε ότι ο προτεινόμενος έλεγχος έχει εξαιρετική απόδοση. Σε σύγκριση με τις υπάρχουσες μεθόδους, η προτεινόμενη μέθοδος φαίνεται να απαιτεί σημαντικά μικρότερο μέγεθος δείγματος με ισοδύναμη ισχύ. Επιπλέον, η προτεινόμενη μέθοδος είναι εύκολα γενικεύσιμη και θέτει το πρόβλημα της εκτίμησης της συμφωνίας μεταξύ δύο ή περισσότερων βαθμολογητών ως προς ένα ή περισσότερα χαρακτηριστικά υπό ένα ενιαίο πλαίσιο, παρέχοντας έτσι ένα εύκολο στη χρήση εργαλείο για τους ιατρούς.

Keywords: Ακολουθιακός έλεγχος, Αξιολόγηση συμφωνίας, Έλεγχος υποθέσεων, Μέθοδος εμφύτευσης Μαρκοβιανής αλυσίδας, Cohen's k .

1. ΕΙΣΑΓΩΓΗ

Δύο βασικές έννοιες στην ιατρική και κοινωνική έρευνα είναι η *εγκυρότητα* (*validity*) και η *αξιοπιστία* (*reliability*). Σύμφωνα με τον Uerbersax (1988), η εγκυρότητα είναι η ακρίβεια των αξιολογήσεων ενώ η αξιοπιστία είναι ο βαθμός στον οποίο οι αξιολογήσεις συμφωνούν.

Αρκετοί δείκτες συμφωνίας και επεκτάσεις τους έχουν προταθεί στη βιβλιογραφία. Ένα κοινό στοιχείο αυτών είναι ότι όλοι εφαρμόζονται σε πίνακες συνάφειας. Αυτό αποτελεί εμπόδιο στη λήψη μια έγκαιρης απόφασης, καθώς η καταγραφή όλων των απαραίτητων αξιολογήσεων σε έναν πίνακα συνάφειας είναι συνήθως χρονοβόρα. Επιπλέον, στην κλινική και επιδημιολογική έρευνα, όπου εξετάζονται μικρού μεγέθους επιδράσεις, απαιτούνται μεγάλα δείγματα. Ωστόσο, μεγάλες μελέτες μπορεί να συνεπάγονται σπατάλη πόρων, έκθεση των ασθενών σε πιθανές βλάβες και πιθανή καθυστέρηση της έκβασης της μελέτης (Olivier and Bell, 2013). Ο Donner (1998) παρουσίασε πίνακες με τα κατάλληλα μεγέθη δειγμάτων για τη σύγκριση των συντελεστών συμφωνίας. Ωστόσο, αυτά τα μεγέθη των δειγμάτων συχνά υπερβαίνουν τα μεγέθη που χρησιμοποιούνται στην πράξη (όχι περισσότερα από 50 άτομα, στις περισσότερες περιπτώσεις). Έτσι, στην πράξη, οι περισσότερες από τις μελέτες είναι μικρής ισχύος. Από στατιστικής άποψης, ακόμη και για μέτρια μεγέθη δειγμάτων, μια μακρά περίοδος παρατήρησης είναι υποχρεωτική για να συλλεχθούν τα στοιχεία των πινάκων συνάφειας και αυτό μπορεί επίσης να αποτελεί ηθικό ζήτημα. Δύο ενδεικτικά παραδείγματα, στα οποία παρήλθε μεγάλο χρονικό διάστημα για να συγκεντρωθεί ένας μάλλον μικρός αριθμός μαστογραφιών, είναι οι μελέτες των Lobbes et al. (2015), όπου μόνο 87 περιπτώσεις καρκίνου μαστού, συγκεντρώθηκαν σε μια περίοδο 15 μηνών και των Ansell et al. (2009) όπου παρήλθε διάστημα 9 μηνών για την συγκέντρωση μόνο 102 περιπτώσεων.

Υπό το πρίσμα αυτό, προτείνουμε έναν μη παραμετρικό ακολουθιακό έλεγχο της συμφωνίας μεταξύ δύο ή περισσότερων βαθμολογητών που δεν απαιτεί πίνακα συνάφειας. Η συμφωνία αξιολογείται μέσω μιας τυχαίας μεταβλητής χρόνου αναμονής που σχετίζεται με το σωρευτικό άθροισμα των διαφωνιών των βαθμολογητών. Μελετάμε τις ιδιότητες της κατανομής της τυχαίας μεταβλητής αυτής και για το σκοπό αυτό χρησιμοποιούμε την τεχνική της εμφύτευσης της Μαρκοβιανής αλυσίδας.

Η εργασία οργανώνεται ως εξής: Στην Ενότητα 2 παρουσιάζουμε τις πιο γνωστές τεχνικές για την αξιολόγηση της συμφωνίας για κατηγορικά δεδομένα. Στην Ενότητα 3 προτείνουμε έναν ακολουθιακό έλεγχο για την αξιολόγηση του επιπέδου της συμφωνίας μεταξύ δύο ή περισσότερων βαθμολογητών. Η Ενότητα 4 περιγράφει τις περιπτώσεις που μπορεί να εφαρμοστεί ο νέος έλεγχος, ενώ η Ενότητα 5 παρουσιάζει το θεωρητικό υπόβαθρο του ελέγχου. Στην Ενότητα 6 παρουσιάζουμε αριθμητικά αποτελέσματα, ενώ ολοκληρώνουμε την εργασία με ορισμένα συμπεράσματα στην Ενότητα 7.

2. ΑΞΙΟΛΟΓΗΣΗ ΤΗΣ ΣΥΜΦΩΝΙΑΣ

Στη βιβλιογραφία έχουν προταθεί διάφορες τεχνικές για την αξιολόγηση της συμφωνίας μεταξύ βαθμολογητών.

Στην περίπτωση των κατηγορικών δεδομένων, ο πιο απλός δείκτης συμφωνίας είναι η παρατηρημένη συμφωνία ή απλή συμφωνία ή αδρή συμφωνία, η οποία είναι η συχνότητα των φορών που οι βαθμολογητές συμφωνούν μεταξύ τους (Hrippsaka and Heitjan, 2002). Στην περίπτωση δύο βαθμολογητών και διχοτομικών αποκρίσεων, χρησιμοποιώντας τον συμβολισμό ενός τυπικού 2×2 πίνακα συνάφειας της μορφής του Πίνακα 1 η παρατηρημένη συμφωνία p_o ορίζεται ως

		Βαθμολογητής 2		
		Θετικό	Αρνητικό	Σύνολο
Βαθμολογητής 1	Θετικό	O_{11}	O_{12}	$O_{1\cdot}$
	Αρνητικό	O_{21}	O_{22}	$O_{2\cdot}$
		$O_{\cdot 1}$	$O_{\cdot 2}$	N

Πίνακας 1: Ένας 2×2 πίνακας συνάφειας

$$p_o = \frac{O_{11} + O_{22}}{O_{11} + O_{12} + O_{21} + O_{22}}. \quad (1)$$

Συνεπώς, η p_o αντιπροσωπεύει το ποσοστό του συνόλου των διαγνώσεων που κατατάσσονται με τον ίδιο τρόπο από τους δύο βαθμολογητές.

Ένας άλλος δείκτης συμφωνίας είναι η ειδική συμφωνία που ποσοτικοποιεί τον βαθμό συμφωνίας για καθεμία από τις κατηγορίες χωριστά (Hrippsaka and Heitjan, 2002; Fleiss, 1981). Στην περίπτωση δύο κατηγοριών, έχουμε την θετική (p_{pos}) και την αρνητική ειδική συμφωνία (p_{neg}). Αυτές είναι οι αναλογίες των περιπτώσεων όπου οι βαθμολογητές συμφωνούν σε μια κατηγορία. Για έναν τυπικό 2×2 πίνακα συνάφειας η θετική και η αρνητική ειδική συμφωνία ορίζονται ως

$$p_{pos} = \frac{2O_{11}}{2O_{11} + O_{12} + O_{21}} \quad \text{και} \quad p_{neg} = \frac{2O_{22}}{O_{12} + O_{21} + 2O_{22}},$$

αντίστοιχα.

Ωστόσο, υπάρχει πιθανότητα μερικές από τις ίδιες ταξινομήσεις να γίνουν τυχαία. Ένα σημαντικό μειονέκτημα της παρατηρημένης και της ειδικής συμφωνίας είναι ότι δεν λαμβάνουν υπόψιν το ποσό της συμφωνίας που οφείλεται στην τύχη. Το πιο γνωστό εργαλείο για την αξιολόγηση της συμφωνίας είναι το *Cohen's κ* (Cohen, 1960), που ορίζεται ως

$$\begin{aligned} \kappa &= \frac{P(X_1 = X_2) - P(X_1 = X_2 | X_1 \text{ και } X_2 \text{ ανεξάρτητα})}{1 - P(X_1 = X_2 | X_1 \text{ και } X_2)} \\ &= \frac{\pi_{11} + \pi_{22} - (\pi_{1+}\pi_{+1} + \pi_{2+}\pi_{+2})}{1 - (\pi_{1+}\pi_{+1} + \pi_{2+}\pi_{+2})}, \end{aligned}$$

όπου το π_{ii} συμβολίζει την πιθανότητα ότι το (X_1, X_2) εμφανίζεται στο κελί στη γραμμή i και τη στήλη i ενώ τα π_{i+} και π_{+j} είναι οι αντίστοιχες περιθώριες κατανομές. Το X_1 είναι μια δίτιμη τυχαία μεταβλητή που αντιπροσωπεύει την αξιολόγηση του πρώτου βαθμολογητή ενώ το X_2 είναι επίσης μια δίτιμη τυχαία μεταβλητή που αντιπροσωπεύει το αποτέλεσμα της αξιολόγησης του δεύτερου βαθμολογητή.

Το κ εκτιμάται από τη σχέση

$$k = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

όπου το p_o είναι η παρατηρημένη συμφωνία που δίνεται στην εξίσωση (1) και

$$p_e = \sum_i p_{i+} p_{+i} = \frac{(O_{11} + O_{12})(O_{11} + O_{21}) + (O_{21} + O_{22})(O_{12} + O_{22})}{(O_{11} + O_{12} + O_{21} + O_{22})^2},$$

με $p_{i+} = \frac{O_{i+}}{O_{11}+O_{12}+O_{21}+O_{22}}$ και $p_{+i} = \frac{O_{+i}}{O_{11}+O_{12}+O_{21}+O_{22}}$.

Από την εμφάνιση του Cohen's κ , έχουν προταθεί πολλές γενικεύσεις. Ο Cohen (1968) πρότεινε μια σταθμισμένη έκδοση του κ για να χρησιμοποιηθεί στην περίπτωση διατάξιμων δεδομένων.

Είναι προφανές ότι, για να εκτιμήσουμε τη συμφωνία με τις παραπάνω τεχνικές χρειαζόμαστε έναν πίνακα συνάφειας. Αυτό σημαίνει ότι, σε πολλά πρακτικά προβλήματα, θα πρέπει να περιμένουμε μεγάλο χρονικό διάστημα για να συγκεντρώσουμε επαρκή στοιχεία για να σχηματίσουμε τον πίνακα. Παρά το γεγονός ότι ένας πρακτικός κανόνας είναι να μην χρησιμοποιούμε δείγματα μεγέθους μικρότερου από 30 για να αξιολογήσουμε τη συμφωνία μέσω του k , όσο μικρότερο είναι το μέγεθος του δείγματος, τόσο λιγότερο πιθανό είναι να έχουμε μια ακριβή εκτίμηση της συμφωνίας. Οι Donner and Eliasziw (1992) προτείνουν δείγματα άνω του 50.

Συνεπώς, απαιτείται μια μέθοδος για την ταχύτερη αξιολόγηση της συμφωνίας καθώς γίνονται διαθέσιμα τα δεδομένα. Υπό το πρίσμα αυτό, στην επόμενη ενότητα προτείνουμε και μελετάμε μια μη παραμετρική μέθοδο αξιολόγησης της συμφωνίας μεταξύ δύο βαθμολογητών που μπορεί να εφαρμοστεί ακολουθιακά.

3. Ο ΑΚΟΛΟΥΘΙΑΚΟΣ ΕΛΕΓΧΟΣ

Στην ενότητα αυτή προτείνουμε έναν ακολουθιακό έλεγχο για την αξιολόγηση της συμφωνίας δύο βαθμολογητών ως προς ένα ή δύο χαρακτηριστικά και r κατηγορίες ταξινόμησης, υπό ένα ενιαίο πλαίσιο.

Η μηδενική υπόθεση του ελέγχου είναι

$$H_0 : \pi = \sum_i \pi_{ii} = \pi_0,$$

για $i = 1, 2, \dots, r$, δηλαδή η πιθανότητα συμφωνίας μεταξύ των βαθμολογητών ισούται με μια προκαθορισμένη πιθανότητα π_0 .

Η εναλλακτική υπόθεση είναι

$$H_1 : \pi < \pi_0$$

και δηλώνει ότι η συμφωνία μεταξύ των δύο βαθμολογητών είναι μικρότερη από ένα υποτιθέμενο αποδεκτό επίπεδο συμφωνίας π_0 .

Ας υποθέσουμε ότι δύο ακτινολόγοι A και B εξετάζουν και ταξινομούν έναν αριθμό μαστογραφιών είτε ως φυσιολογικές είτε ως μη φυσιολογικές. Η αξιολόγηση κάθε ακτινολόγου την χρονική στιγμή t ορίζεται ως

$$Y_t^\ell = \begin{cases} 1, & \text{εάν ο ακτινολόγος } \ell \text{ ταξινομεί τη μαστογραφία ως φυσιολογική} \\ 2, & \text{εάν ο ακτινολόγος } \ell \text{ ταξινομεί τη μαστογραφία ως μη φυσιολογική,} \end{cases}$$

$$\ell = A, B, t = 1, 2, \dots$$

Οι πιθανότητες που σχετίζονται με τις Y_t^A, Y_t^B είναι

$$p_{ij} = P(Y_t^A = i, Y_t^B = j), \quad i, j = 1, 2.$$

Προκειμένου να αξιολογήσουμε τη συμφωνία, ορίζουμε μια τυχαία μεταβλητή T_{AD} η οποία μετρά τον αριθμό των αξιολογήσεων μέχρι το σωρευτικό άθροισμα (έως την χρονική στιγμή t_0) των απόλυτων διαφορών των αξιολογήσεων των δύο βαθμολογητών, δηλ.

$$AD = \sum_{t=1}^{t_0} |Y_t^A - Y_t^B|,$$

να ξεπεράσει για πρώτη φορά ένα κατώφλι f . Συνεπώς, η T_{AD} ισούται με t_0 εάν και μόνο εάν μετά την ολοκλήρωση της t_0 -οστής αξιολόγησης έχουμε $AD > f$ για πρώτη φορά. Μαθηματικά, μπορούμε να γράψουμε $T_{AD} = \min \{t_0 : AD > f\}$. Επειδή η $|Y_t^A - Y_t^B|$ μπορεί να πάρει μόνο τις τιμές 0 ή 1, η AD παίρνει τιμές στο $\mathbb{Z}^+$.

Ο Πίνακας 2 παρουσιάζει δύο περιπτώσεις αξιολόγησης της συμφωνίας μεταξύ δύο βαθμολογητών. Στην *Περίπτωση A* οι δύο βαθμολογητές συμφωνούν τις περισσότερες των περιπτώσεων. Έτσι, ο συνολικός αριθμός των απόλυτων διαφορών των διαφωνιών AD έχει μια μικρή τιμή (μόλις 2 μετά από 19 αξιολογήσεις). Στην *Περίπτωση B*, όπου οι βαθμολογητές διαφωνούν αρκετά συχνά, η τιμή της AD είναι υψηλή (10 μετά από 19 αξιολογήσεις). Και στις δύο περιπτώσεις, η απόλυτη διαφορά μετρά ακριβώς τον αριθμό των διαφωνιών.

Είναι προφανές ότι εάν οι δύο βαθμολογητές συμφωνούν απόλυτα, η τιμή του $|Y_t^A - Y_t^B|$ θα είναι ίση με 0, ενώ όσο περισσότερες φορές διαφωνούν οι βαθμολογητές, τόσο μεγαλύτερη θα είναι η AD και τόσο το γρηγορότερο θα ξεπεράσει την τιμή f . Έτσι, συγκρίνοντας διαδοχικά τις αξιολογήσεις των δύο ειδικών, εάν η AD είναι μεγαλύτερη από ένα όριο f γρήγορα, δηλαδή πριν από την c -οστή αξιολόγηση, αυτό είναι ένδειξη ότι υπάρχει σημαντική διαφωνία μεταξύ των βαθμολογητών.

Αξιολόγηση (t)		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	...
Περίπτωση Α	Y_t^A	1	2	2	1	2	1	1	2	1	2	1	2	1	2	2	2	1	1	1	...
	Y_t^B	1	2	2	1	2	1	2	2	1	2	2	2	1	2	2	2	1	1	1	...
	$ Y_t^A - Y_t^B $	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	...
	$AD = \sum_{t=1}^{t_0} Y_t^A - Y_t^B $	0	0	0	0	0	0	1	1	1	1	2	2	2	2	2	2	2	2	2	...
Περίπτωση Β	Y_t^A	1	1	2	2	2	2	1	2	1	2	1	2	1	2	1	2	1	1	1	...
	Y_t^B	1	2	1	2	1	1	2	2	1	2	2	2	1	1	2	1	2	1	1	...
	$ Y_t^A - Y_t^B $	0	1	1	0	1	1	1	0	0	0	1	0	0	1	1	1	1	0	0	...
	$AD = \sum_{t=1}^{t_0} Y_t^A - Y_t^B $	0	1	2	2	3	4	5	5	5	5	6	6	6	7	8	9	10	10	10	...

Πίνακας 2: Το σωρευτικό άθροισμα των απόλυτων διαφορών των διαφωνιών στην περίπτωση συμφωνίας (Περίπτωση Α) και στην περίπτωση διαφωνίας (Περίπτωση Β)

Απορρίπτουμε τη μηδενική υπόθεση με πιθανότητα σφάλματος τύπου Ι, $\alpha = P(T_{AD} \leq c|f, H_0)$. Μετά την απόρριψη, δηλώνουμε ότι η συμφωνία μεταξύ των βαθμολογητών έχει πέσει κάτω από το επιθυμητό επίπεδο, εάν για πρώτη φορά, στην t_0 -οστή αξιολόγηση, με $t_0 \leq c$, η AD είναι μεγαλύτερη από το f .

Συνεπώς, προκειμένου να ελέγξουμε ακολουθιακά την ισχύ της H_0 , επιλέγουμε πρώτα την επιθυμητή πιθανότητα σφάλματος τύπου Ι, α . Στη συνέχεια, χρησιμοποιώντας την εξίσωση $\alpha = P(T_{AD} \leq c|f, H_0)$ υπολογίζουμε διάφορες υποψήφιες τιμές του c για διάφορες τιμές του f . Η επιλογή της f εξαρτάται από το επιθυμητό επίπεδο ισχύος γ , για την υπό εξέταση εναλλακτική υπόθεση και αυτό καθορίζει την τιμή του c .

4. ΓΕΝΙΚΕΥΣΕΙΣ

Στην ενότητα αυτή θα παρουσιάσουμε τρεις γενικεύσεις. Σε όλες τις περιπτώσεις, όπως στην περίπτωση που παρουσιάστηκε στην προηγούμενη ενότητα, χρησιμοποιούμε το παράδειγμα των ακτινολόγων, οι οποίοι αξιολογούν ανεξάρτητα τις ίδιες μαστογραφίες. Ωστόσο, είναι προφανές ότι ο προτεινόμενος ακολουθιακός έλεγχος μπορεί να εφαρμοστεί και σε άλλες διαγνωστικές διαδικασίες.

Περίπτωση 1: Αξιολόγηση της συμφωνίας δύο βαθμολογητών ως προς ένα χαρακτηριστικό και δύο ή περισσότερες κατηγορίες ταξινόμησης. Ας θεωρήσουμε την περίπτωση κατά την οποία δύο ακτινολόγοι (A και B) κατατάσσουν κάθε μαστογραφία ως προς ένα χαρακτηριστικό σε μια από τις r (≥ 2) κατηγορίες. Η αξιολόγηση κάθε ακτινολόγου ορίζεται ως εξής

$$Y_t^\ell = \begin{cases} 1, & \text{εάν ο ακτινολόγος } \ell \text{ ταξινομεί τη μαστογραφία στην κατηγορία } 1, \\ 2, & \text{εάν ο ακτινολόγος } \ell \text{ ταξινομεί τη μαστογραφία στην κατηγορία } 2, \\ \vdots & \vdots \\ r, & \text{εάν ο ακτινολόγος } \ell \text{ ταξινομεί τη μαστογραφία στην κατηγορία } r, \end{cases}$$

$\ell = A, B$.

Οι πιθανότητες που σχετίζονται με τις Y_t^A, Y_t^B είναι

$$p_{ij} = P(Y_t^A = i, Y_t^B = j), \quad i, j = 1, 2, 3, \dots, r.$$

Η «στατιστική συνάρτηση» που χρησιμοποιούμε είναι το σωρευτικό άθροισμα των απόλυτων διαφορών των διαφωνιών μεταξύ των δύο βαθμολογητών, δηλ.

$$AD = \sum_{t=1}^{t_0} |Y_t^A - Y_t^B|.$$

Στην περίπτωση αυτή, η $|Y_t^A - Y_t^B|$ μπορεί να πάρει μόνο τις τιμές $0, 1, 2, \dots, r-1$, οπότε πάλι η AD παίρνει τιμές στο $\mathbb{Z}^+$. Εάν οι δύο βαθμολογητές συμφωνούν τότε η AD θα βρίσκεται κοντά στο 0. Αντίθετα, όσο περισσότερες διαφωνούν οι βαθμολογητές, τόσο υψηλότερη θα είναι η τιμή του AD . Συνεπώς, εάν η AD ξεπεράσει γρήγορα ένα κατώφλι f , ας πούμε πριν την c -οστή αξιολόγηση, τότε μπορεί να υπάρχει σημαντική διαφωνία μεταξύ των βαθμολογητών.

Ως παράδειγμα, ο Πίνακας 3 παρουσιάζει τις αξιολογήσεις κάθε βαθμολογητή και το σωρευτικό άθροισμα των διαφωνιών, AD , για $r = 3$.

Αξιολόγηση (t)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	...
Y_t^A	1	2	3	3	1	3	1	3	3	1	3	1	2	3	1	2	1	1	1	...
Y_t^B	1	3	3	1	1	3	1	3	2	2	1	2	2	3	1	1	1	1	1	...
$ Y_t^A - Y_t^B $	0	1	0	2	0	0	0	0	1	1	2	1	0	0	0	1	0	0	0	...
$AD = \sum_{t=1}^n Y_t^A - Y_t^B $	0	1	1	3	3	3	3	3	4	5	7	8	8	8	8	9	9	9	9	...

Πίνακας 3: Το σωρευτικό άθροισμα των απόλυτων διαφορών των διαφωνιών για την αξιολόγηση της συμφωνίας δύο βαθμολογητών ως προς ένα χαρακτηριστικό

Περίπτωση 2: Αντιμετώπιση της περίπτωσης ακραίας διαφωνίας.

Εδώ θεωρούμε μια ειδική, αλλά όχι τόσο ασυνήθιστη περίπτωση ταξινόμησης σε r κατηγορίες. Για παράδειγμα, ας υποθέσουμε ότι οι δύο ακτινολόγοι αξιολογούν με βάση μια 7-βάθμια κλίμακα Likert. Τι θα συμβεί αν ο πρώτος βαθμολογητής επιλέξει την κατηγορία 1 και ο δεύτερος βαθμολογητής επιλέξει την κατηγορία 7; Αυτό θα μπορούσε να σημαίνει ότι οι βαθμολογητές διαφωνούν ισχυρά και υπάρχουν αρκετές ενδείξεις για να απορρίψουμε την H_0 . Σε αυτή την περίπτωση, προτείνουμε να σταματά αμέσως η σύγκριση δηλώνοντας ότι οι δύο βαθμολογητές διαφωνούν ισχυρά.

Οι μεταβλητές Y_t^A, Y_t^B οι οποίες συμβολίζουν την αξιολόγηση κάθε βαθμολογητή ορίζονται όπως στην προηγούμενη περίπτωση. Στην περίπτωση αυτή, εκτός από τις πιθανότητες που σχετίζονται με τις Y_t^A, Y_t^B ,

$$p_{ij} = P(Y_t^A = i, Y_t^B = j), \quad i, j = 1, 2, 3, \dots, r,$$

χρειαζόμαστε επιπλέον την πιθανότητα εμφάνισης μιας ακραίας διαφωνίας (ED), p_ξ , θεωρώντας ως ακραία διαφωνία την περίπτωση όπου $|i-j| \geq \xi$ (ο ορισμός της ακραίας διαφωνίας και η επιλογή του ξ αφήνεται στον ενδιαφερόμενο επαγγελματία υγείας).

Η «στατιστική συνάρτηση» παραμένει η ίδια. Ωστόσο, εδώ παίρνουμε μια άμεση απόφαση, όταν εμφανίζεται μια ακραία διαφορά. Έτσι, δηλώνουμε ότι η υπάρχει

διαφωνία μεταξύ των βαθμολογητών εάν η AD είναι μεγαλύτερη από ένα κατώφλι f , ή εάν εμφανιστεί αρκετά νωρίς μια ακραία διαφωνία.

Ο Πίνακας 4 παρουσιάζει τις αξιολογήσεις σε μια 7-βάθμια κλίμακα Likert για κάθε βαθμολογητή και τον αθροιστικό αριθμό των διαφωνιών. Στον 19^ο ασθενή, ο βαθμολογητής A επέλεξε 7 ενώ ο B επέλεξε 1, το οποίο σημαίνει ότι οι οι βαθμολογητές διαφωνούν ακραία και παίρνουμε μια άμεση απόφαση.

ρατινγ (t)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	...
Y_t^A	1	2	3	3	1	3	4	3	3	6	3	5	2	3	7	7	3	5	7	...
Y_t^B	1	3	3	1	1	3	4	3	2	7	1	6	2	3	7	6	3	5	1	...
$ Y_t^A - Y_t^B $	0	1	0	2	0	0	0	0	1	1	2	1	0	0	0	1	0	0	ξ	...
$AD = \sum_{t=1}^n Y_t^A - Y_t^B $	0	1	1	3	3	3	3	3	4	5	7	8	8	8	8	9	9	9	ξ	...

Πίνακας 4: Το σωρευτικό άθροισμα των απόλυτων διαφορών των διαφωνιών για την αξιολόγηση της συμφωνίας δύο βαθμολογητών ως προς ένα χαρακτηριστικό και την εμφάνιση ακραίας διαφωνίας

Περίπτωση 3: Αξιολόγηση της συμφωνίας δύο βαθμολογητών ως προς δύο χαρακτηριστικά. Ας εξετάσουμε τώρα την περίπτωση που οι δύο ακτινολόγοι (A και B) ταξινομούν τις μαστογραφίες ως προς δύο χαρακτηριστικά χρησιμοποιώντας δύο κατηγορίες ταξινόμησης για κάθε χαρακτηριστικό. Για παράδειγμα, οι ακτινολόγοι αξιολογούν το σχήμα και την πυκνότητα ενός όγκου. Η αξιολόγηση κάθε βαθμολογητή ορίζεται ως εξής

$$Y_{i,t}^\ell = \begin{cases} 1, & \text{εάν ο ακτινολόγος } \ell \text{ ταξινομεί τη μαστογραφία ως φυσιολογική ως προς το χαρακτηριστικό } i \\ 2, & \text{εάν ο ακτινολόγος } \ell \text{ ταξινομεί τη μαστογραφία ως μη φυσιολογική ως προς το χαρακτηριστικό } i, \end{cases}$$

$$\ell = A, B, i = 1, 2.$$

Για κάθε αξιολόγηση t , $t = 1, 2, \dots$, έχουμε ένα αποτέλεσμα σχετικά με την ταξινόμηση. Οι πιθανότητες που σχετίζονται με τις $Y_{1,t}^A$, $Y_{1,t}^B$, $Y_{2,t}^A$, $Y_{2,t}^B$ είναι

$$p_{i_1 i_2 j_1 j_2} = P(Y_{1,t}^A = i_1, Y_{2,t}^A = i_2, Y_{1,t}^B = j_1, Y_{2,t}^B = j_2), \quad i_1, j_1, i_2, j_2 = 1, 2.$$

Για να αξιολογήσουμε τη συμφωνία των δύο βαθμολογητών, μπορούμε να χρησιμοποιήσουμε ως «στατιστική συνάρτηση» το σωρευτικό άθροισμα των απόλυτων διαφορών των διαφωνιών για κάθε χαρακτηριστικό, δηλ.

$$AD = \sum_{t=1}^{t_0} (|Y_{1,t}^A - Y_{1,t}^B| + |Y_{2,t}^A - Y_{2,t}^B|).$$

Εδώ το $|Y_{1,t}^A - Y_{1,t}^B| + |Y_{2,t}^A - Y_{2,t}^B|$ μπορεί να πάρει τις τιμές 0, 1 ή 2, οπότε η AD παίρνει τιμές στο $\mathbb{Z}^+$. Δηλώνουμε ότι υπάρχει διαφωνία μεταξύ των βαθμολογητών εάν η AD είναι μεγαλύτερη από ένα κατώφλι f . Η παραπάνω διαδικασία μπορεί να χρησιμοποιηθεί αποτελεσματικά για να χειριστεί περιπτώσεις με περισσότερα από δύο χαρακτηριστικά.

Ο Πίνακας 5 παρουσιάζει τον αθροιστικό αριθμό των διαφωνιών.

Αξιολόγηση (t)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	...
$Y_{1,t}^A$	1	1	2	2	2	2	1	2	2	1	2	2	1	2	1	2	1	1	1	...
$Y_{2,t}^A$	1	1	2	2	2	2	1	1	2	1	2	1	1	2	1	1	1	1	1	...
$Y_{1,t}^B$	1	1	1	2	2	1	1	1	1	1	2	2	1	1	1	1	1	1	1	...
$Y_{2,t}^B$	1	1	2	2	1	2	1	2	1	1	2	1	1	2	1	2	1	1	1	...
$Z_{1,t} = Y_{1,t}^A - Y_{1,t}^B $	0	0	1	0	0	1	0	1	1	0	0	0	0	1	0	1	0	0	0	...
$Z_{2,t} = Y_{2,t}^A - Y_{2,t}^B $	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	1	0	0	0	...
$Z_{1,t} + Z_{2,t}$	0	0	1	0	1	1	0	2	2	0	0	0	0	1	0	2	0	0	0	...
$AD = \sum_{t=1}^n (Z_{1,t} + Z_{2,t})$	0	0	1	1	2	3	3	5	7	7	7	7	7	8	8	10	10	10	10	...

Πίνακας 5: Το σωρευτικό άθροισμα των απόλυτων διαφορών των διαφωνιών για την αξιολόγηση της συμφωνίας δύο βαθμολογητών ως προς δύο χαρακτηριστικά

5. Η ΑΚΡΙΒΗΣ ΚΑΤΑΝΟΜΗ ΤΗΣ T_{AD}

Στην Ενότητα 3, προτείναμε τη χρήση της στατιστικής συνάρτησης χρόνου αναμονής T_{AD} που απαιτείται για το σωρευτικό άθροισμα AD να πάρει τιμή μεγαλύτερη του f , προκειμένου να ελέγξουμε ακολουθιακά την μηδενική υπόθεση. Για τον καθορισμό της ακριβούς κατανομής της T_{AD} , χρησιμοποιούμε την τεχνική εμφύτευσης της Μαρκοβιανής αλυσίδας. Η τεχνική εμφύτευσης χρησιμοποιείται ευρέως σήμερα για την εύρεση της κατανομής διαφόρων τύπων στατιστικών συναρτήσεων που ορίζονται σε μονοδιάστατες ή πολυδιάστατες ακολουθίες διακριτών τυχαίων μεταβλητών. Μεταξύ άλλων, οι Antzoulakos et al. (2003), Koutras et al. (2008) και Bersimis et al. (2015) έχουν χρησιμοποιήσει αυτή την τεχνική υπό διάφορα πλαίσια.

Για να βρούμε την ακριβή αθροιστική κατανομή μιας διακριτής τυχαίας μεταβλητής, όπως η T_{AD} , που ορίζεται σε μια ακολουθία πολυσταδιακών δοκιμών χρειαζόμαστε μια Μαρκοβιανή αλυσίδα $\{R_t, t = 0, 1, 2, \dots\}$ ορισμένη σε έναν χώρο καταστάσεων Ω και να επωφεληθούμε πλήρως από την εξίσωση

$$P(T_{AD} \leq t_0) = P(R_{t_0} = \alpha_b) = \pi'_0 \mathbf{\Lambda}_0^{t_0} \mathbf{e}_b.$$

Με α_b συμβολίζουμε την απορροφητική κατάσταση της Μαρκοβιανής αλυσίδας, π'_0 είναι το διάνυσμα γραμμή των αρχικών πιθανοτήτων, $\mathbf{\Lambda}_0$ είναι ο πίνακας πιθανοτήτων μετάβασης και $\mathbf{e}_b$ είναι ένα $1 \times b$ διάνυσμα της μορφής $(0, 0, \dots, 1)$. Η κατανομή πιθανότητας της T_{AD} μπορεί να εξαχθεί χρησιμοποιώντας τη σχέση

$$P(T_{AD} = t_0) = \pi'_1 (\mathbf{\Lambda}_1)^{t_0-1} \mathbf{h}, \quad (2)$$

όπου π_1 είναι ένα $(b-1) \times 1$ διάνυσμα στήλη που είναι το διάνυσμα π_0 χωρίς την τελευταία γραμμή και ο $\mathbf{\Lambda}_1$ είναι ο $\mathbf{\Lambda}_0$ χωρίς την τελευταία γραμμή και την τελευταία στήλη. Στη συνέχεια, αφού ορίσουμε το $\mathbf{1}$ ως ένα $(b-1) \times 1$ διάνυσμα στήλη με μονάδες και χρησιμοποιήσουμε το γεγονός ότι ο $\mathbf{\Lambda}_0$ είναι στοχαστικός, προκύπτει ότι $\mathbf{h} = \mathbf{1} - \mathbf{\Lambda}_1 \mathbf{1}$ (Balakrishnan et al., 2009).

Ο χώρος καταστάσεων της αλυσίδας $\{R_t, t = 0, 1, 2, \dots\}$ είναι ο $\Omega = \{i : i = 0, \dots, f-1, f\} \cup \{\alpha_b\}$ με το i να ακολουθεί τη διαφορά AD (προφανώς ο Ω αποτελείται από $f+2$ στοιχεία) ενώ η $\{R_t, t = 0, 1, 2, \dots\}$ ορίζεται ως

$$R_t = \begin{cases} i, & \text{εάν } AD = i \\ \alpha_b, & \text{εάν } AD > f \text{ ή εμφανιστεί μια } ED, \end{cases}$$

$t = 1, 2, \dots$ και $R_0 = 0$ από τη φύση του προτεινόμενου μοντέλου. Η κατάσταση α_b αντιπροσωπεύει όλες τις πιθανές απορροφητικές καταστάσεις της αλυσίδας. Συγκεκριμένα, $\alpha_b = \{f + 1, \dots, f + r\} \cup \{ED\}$.

Ένα βασικό πλεονέκτημα της προτεινόμενης μεθόδου είναι ότι ο Λ_1 έχει την ίδια απλή μορφή για όλες τις περιπτώσεις που παρουσιάζονται.

Περίπτωση 1: Αξιολόγηση της συμφωνίας δύο βαθμολογητών ως προς ένα χαρακτηριστικό και δύο ή περισσότερες κατηγορίες ταξινόμησης. Για δύο βαθμολογητές, ένα χαρακτηριστικό και r διαφορετικές πιθανές τιμές της απόλυτης διαφοράς των αξιολογήσεων, η γενική μορφή του Λ_1 είναι

$$\Lambda_1 = \begin{bmatrix} \begin{matrix} 0 \\ 1 \\ 2 \\ \vdots \\ r-1 \\ r \\ r+1 \\ r+2 \\ \vdots \\ f-r \\ f-(r-1) \\ \vdots \\ f-1 \\ f \end{matrix} & \begin{matrix} 0 & 1 & 2 & \dots & r-1 & r & r+1 & r+2 & \dots & f-r & f-(r-1) & \dots & f-1 & f \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \\ \vdots \\ r-1 \\ r \\ r+1 \\ r+2 \\ \vdots \\ f-r \\ f-(r-1) \\ \vdots \\ f-1 \\ f \end{matrix} & \begin{matrix} q_0 & q_1 & q_2 & \dots & q_{r-1} & & & & & & & & & \\ & q_0 & q_1 & \dots & q_{r-2} & q_{r-1} & & & & & & & & \\ & & q_0 & \dots & q_{r-3} & q_{r-2} & q_{r-1} & & & & & & & \\ & & & \ddots & \ddots & \ddots & \ddots & \ddots & & & & & & \\ & & & & q_0 & q_1 & q_2 & q_3 & \dots & & & & & \\ & & & & & q_0 & q_1 & q_2 & \dots & & & & & \\ & & & & & & q_0 & q_1 & \dots & & & & & \\ & & & & & & & q_0 & \dots & & & & & \\ & & & & & & & & \ddots & & & & & \\ & & & & & & & & & q_0 & q_1 & \dots & q_{r-2} & q_{r-1} \\ & & & & & & & & & q_0 & \dots & q_{r-3} & q_{r-2} \\ & & & & & & & & & & \ddots & \ddots & \ddots \\ & & & & & & & & & & & q_0 & q_1 \\ & & & & & & & & & & & & q_0 \end{matrix} \end{bmatrix}$$

όπου

$$\begin{aligned} q_0 &= \sum_{i,j} p_{ij}, & \text{για } i = j \text{ και } i, j = 1, 2, \dots, r, \\ q_1 &= \sum_{i,j} p_{ij}, & \text{για } |i - j| = 1 \text{ και } i, j = 1, 2, \dots, r, \\ q_2 &= \sum_{i,j} p_{ij}, & \text{για } |i - j| = 2 \text{ και } i, j = 1, 2, \dots, r, \\ &\vdots & \\ q_{r-1} &= \sum_{i,j} p_{ij}, & \text{για } |i - j| = r - 1 \text{ και } i, j = 1, 2, \dots, r. \end{aligned}$$

Ο παραπάνω πίνακας περιλαμβάνει τις πιθανότητες των μεταβάσεων της αλυσίδας από κατάσταση σε κατάσταση. Πιο συγκεκριμένα, η αλυσίδα κινείται από την κατάσταση i στην κατάσταση i , $i = 1, \dots, f$ με πιθανότητα q_0 , από την κατάσταση i στην κατάσταση $i + 1$, $i = 1, \dots, (f - 1)$ με πιθανότητα q_1 , από την κατάσταση i στην κατάσταση $i + 2$, $i = 1, \dots, (f - 2)$ με πιθανότητα q_2 , ..., από την κατάσταση i στην κατάσταση $i + r$, $i = 1, \dots, (f - r)$, με πιθανότητα q_{r-1} . Για οποιαδήποτε άλλη περίπτωση (συμπεριλαμβανομένης και της περίπτωσης της εμφάνισης πλήρους διαφωνίας), η αλυσίδα κινείται προς την απορροφητική κατάσταση. Τα κενά κελιά του Λ_1 είναι μηδέν.

Επειδή ο Λ_1 είναι ένας άνω διαγώνιος πίνακας και κάνοντας χρήση της γενικής μορφής της πιθανογεννήτριας συνάρτησης της AD , η οποία δίνεται από τη σχέση

$$G(z) = E(z^{T_{AD}}) = \sum_{t_0=1}^{\infty} P(T_{AD} = t_0) z^{t_0} = z \pi_0 (\mathbf{I} - z \Lambda_1)^{-1} \mathbf{h}$$

(Balakrishnan et al., 2009) καταλήγουμε στο

$$\sum_{t_0=1}^{\infty} P(T_{AD} = t_0) z^{t_0} = \frac{\sum_{i=1}^{f+1} v_i z^i}{1 - \sum_{i=1}^{f+1} (-1)^{i-1} \binom{f+1}{i} q_0^i z^i}. \quad (3)$$

Οι συντελεστές v_i μπορούν να υπολογιστούν εύκολα για συγκεκριμένες τιμές του f . Ωστόσο, φαίνεται ότι είναι δύσκολο να πάρει μια κλειστή μορφή. Έτσι, μπορούμε να επωφεληθούμε πλήρως από την πιθανογεννήτρια συνάρτηση παίρνοντας έναν αναδρομικό τύπο για την πιθανογεννήτρια T_{AD} .

Επιπλέον, παίρνουμε το ακόλουθο αναδρομικό σχήμα

$$P(T_{AD} = t_0) = \sum_{i=1}^{f+1} (-1)^{i-1} \binom{f+1}{i} q_0^i P(T_{AD} = t_0 - i), \quad t_0 \geq f + 1.$$

Για $t_0 \leq f$ μπορούμε να χρησιμοποιήσουμε την εξίσωση (2) ή να χρησιμοποιήσουμε τον αριθμητή της εξίσωσης (3), όπως οι Antzoulakos et al. (2003).

Περίπτωση 2: Αντιμετώπιση της περίπτωσης ακραίας διαφωνίας. Η ίδια μορφή του Λ_1 ισχύει και στην περίπτωση που έχουμε ακραία διαφωνία, δηλαδή στην Περίπτωση 2 της Ενότητας 4. Η μόνη διαφορά είναι ότι για $|i - j| \geq \xi$, $i, j = 1, 2, 3, \dots, r$ οι πιθανότητες μετάβασης μηδενίζονται.

Περίπτωση 3: Αξιολόγηση της συμφωνίας δύο βαθμολογητών ως προς δύο χαρακτηριστικά. Για την Περίπτωση 3 της Ενότητας σεστιον4, δηλαδή την αξιολόγηση της συμφωνίας δύο βαθμολογητών ως προς δύο χαρακτηριστικά και δύο κατηγορίες ταξινόμησης, έχουμε

$$q_0 = p_{1111} + p_{2222} + p_{2121} + p_{1212},$$

$$q_1 = p_{1112} + p_{1121} + p_{1211} + p_{2111} + p_{2221} + p_{2212} + p_{2122} + p_{1221},$$

$$q_2 = p_{2211} + p_{1122} + p_{1221} + p_{2112}.$$

6. ΑΡΙΘΜΗΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Παράδειγμα 1: Αξιολόγηση της συμφωνίας δύο βαθμολογητών ως προς ένα χαρακτηριστικό και δύο κατηγορίες ταξινόμησης.

Ας υποθέσουμε ότι οι δύο βαθμολογητές ταξινομούν μια μαστογραφία ως φυσιολογική με πιθανότητα $p_{11} = 0,800$ και ως μη φυσιολογική με πιθανότητα $p_{22} = 0,160$. Επιπλέον, ας υποθέσουμε ότι ο πρώτος βαθμολογητής κατατάσσει μια μαστογραφία ως φυσιολογική και ο δεύτερος βαθμολογητής ταξινομεί τη μαστογραφία ως μη φυσιολογική με πιθανότητα $p_{12} = 0,020$, ενώ η πιθανότητα ο πρώτος βαθμολογητής να ταξινομήσει μια μαστογραφία ως μη φυσιολογική και ο δεύτερος βαθμολογητής ως φυσιολογική με πιθανότητα $p_{21} = 0,020$. Το κ του Cohen ισούται με 0,864.

						f									
π_{11}	π_{12}	π_{21}	π_{22}	$\pi_{11} + \pi_{22}$	κ	3	4	5	6	7	8	9	10		
0,8000	0,0200	0,0200	0,1600	0,960	0,864	0,046 (100,000)	0,049 (125,000)	0,048 (150,000)	0,049 (175,000)	0,049 (200,000)	0,049 (225,000)	0,049 (250,000)	0,049 (275,000)		
0,7500	0,0250	0,0250	0,2000	0,950	0,857	0,088 (80,000)	0,104 (100,000)	0,111 (120,000)	0,122 (140,000)	0,133 (160,000)	0,142 (180,000)	0,149 (200,000)	0,159 (220,000)		
0,7000	0,0300	0,0300	0,2400	0,940	0,848	0,145 (66,667)	0,179 (83,333)	0,204 (100,000)	0,230 (116,667)	0,260 (133,333)	0,286 (150,000)	0,308 (166,667)	0,350 (183,333)		
0,6500	0,0350	0,0350	0,2800	0,930	0,838	0,212 (57,143)	0,271 (71,429)	0,315 (85,714)	0,362 (100,000)	0,412 (114,286)	0,455 (128,571)	0,493 (142,857)	0,535 (157,143)		
0,6000	0,0400	0,0400	0,3200	0,920	0,826	0,288 (50,000)	0,371 (62,500)	0,435 (75,000)	0,500 (87,500)	0,564 (100,000)	0,619 (112,500)	0,664 (125,000)	0,711 (137,500)		
0,5500	0,0450	0,0450	0,3600	0,910	0,813	0,367 (44,444)	0,472 (55,556)	0,552 (66,667)	0,627 (77,778)	0,698 (88,889)	0,754 (100,000)	0,798 (99,999)	0,839 (122,222)		
0,5000	0,0500	0,0500	0,4000	0,900	0,798	0,446 (40,000)	0,569 (50,000)	0,657 (60,000)	0,736 (70,000)	0,803 (80,000)	0,852 (90,000)	0,888 (100,000)	0,919 (110,000)		
0,4500	0,0550	0,0550	0,4400	0,890	0,780	0,523 (36,364)	0,656 (45,455)	0,747 (54,546)	0,821 (63,636)	0,878 (72,727)	0,917 (81,818)	0,943 (90,909)	0,963 (100,000)		
0,4000	0,0600	0,0600	0,4800	0,880	0,758	0,595 (33,333)	0,732 (41,667)	0,819 (50,000)	0,883 (58,333)	0,929 (66,667)	0,956 (75,000)	0,973 (83,333)	0,984 (91,667)		
0,3500	0,0650	0,0650	0,5200	0,870	0,732	0,661 (30,769)	0,796 (38,462)	0,874 (46,154)	0,927 (53,846)	0,960 (61,539)	0,978 (69,231)	0,988 (76,923)	0,994 (84,615)		
				c		34	50	66	83	101	119	137	156		

Πίνακας 6: Η ισχύς, ο αναμενόμενος αριθμός αξιολογήσεων και το c για $f = 3 - 10$ στην περίπτωση ενός χαρακτηριστικού και δύο κατηγοριών ταξινόμησης και υψηλό Cohen's k

Θεωρώντας το παραπάνω σύνολο πιθανοτήτων ως μηδενική υπόθεση, ο Πίνακας 6 παρουσιάζει την ισχύ $\gamma = P(T_{AD} \leq c | f, H_0)$, τον αναμενόμενο αριθμό των μαστογραφιών (σε παρένθεση) και την τιμή του c (μέγιστος αριθμός των μαστογραφιών) για $f = 3$ έως 10 για εννέα εναλλακτικές υποθέσεις, οι οποίες αντιπροσωπεύουν μείωση στο κ . Ας εξετάσουμε την εναλλακτική υπόθεση $H_1 : p_{11} = 0,450, p_{12} = 0,055, p_{21} = 0,055, p_{22} = 0,440$ που αντιστοιχεί σε $\kappa = 0,780$ (σχεδόν μια μείωση κατά 10%). Αν θέλουμε να ανιχνεύσουμε αυτή τη μείωση στο κ με $\gamma \geq 0,90$, θα πρέπει να επιλέξουμε $f = 8$, καθώς αυτό είναι το χαμηλότερο f που δίνει $\gamma \geq 0,90$ (συγκεκριμένα $\gamma = 0,917$). Η τιμή του κρίσιμου σημείου c (δηλαδή ο μέγιστος αριθμός των μαστογραφιών που απαιτούνται να αξιολογηθούν) ισούται με 119 που αντιστοιχεί σε $\alpha = 0,049$. Ο αναμενόμενος αριθμός των μαστογραφιών είναι 82.

Παρατηρούμε ότι για όλες τις εναλλακτικές υποθέσεις, καθώς αυξάνει το f αυξάνει και η ισχύς. Επιπλέον, όσο περισσότερο μειώνεται το κ , τόσο αυξάνεται η ισχύς.

Παράδειγμα 2: Αξιολόγηση της συμφωνίας δύο βαθμολογητών ως προς ένα χαρακτηριστικό και τρεις κατηγορίες ταξινόμησης.

Για την περίπτωση της αξιολόγησης της συμφωνίας δύο βαθμολογητών ως προς ένα χαρακτηριστικό και τρεις κατηγορίες ταξινόμησης, θεωρούμε, ως μηδενική υπόθεση, την περίπτωση όπου $\pi_{11} + \pi_{22} + \pi_{33} = 0,880$. Αυτό δίνεται στην πρώτη σειρά του Πίνακα 7 και αντιστοιχεί σε μια μέτρια τιμή του κ του Cohen ($\kappa = 0,571$). Κάθε πιθανότητα διαφωνίας ισούται με $(1 - \pi_{11})/2$.

Ο Πίνακας 7 παρουσιάζει την ισχύ γ , τον αναμενόμενο αριθμό μαστογραφιών (σε παρένθεση) και τον μέγιστο αριθμό των μαστογραφιών c για $f = 5$ έως 12 για εννέα εναλλακτικές υποθέσεις, οι οποίες αντιπροσωπεύουν μείωση του κ . Σε όλες τις περιπτώσεις, κάθε πιθανότητα διαφωνίας ισούται με $(1 - \pi_{11})/2$. Ας εξετάσουμε την εναλλακτική υπόθεση με $\pi_{11} + \pi_{22} + \pi_{33} = 0,505$ που δίνεται στην όγδοη σειρά

				f									
π_{11}	π_{22}	π_{33}	$\pi_{11} + \pi_{22} + \pi_{33}$	κ	5	6	7	8	9	10	11	12	
0,8000	0,0700	0,0100	0,880	0,571	0,047 (39,060)	0,048 (45,313)	0,048 (51,562)	0,046 (57,813)	0,044 (64,063)	0,049 (70,313)	0,046 (76,562)	0,049 (82,811)	
0,7500	0,0650	0,0350	0,775	0,558	0,090 (31,248)	0,099 (36,251)	0,105 (41,250)	0,109 (46,250)	0,111 (51,250)	0,123 (56,250)	0,126 (61,250)	0,139 (66,250)	
0,7000	0,0600	0,0600	0,730	0,543	0,150 (26,040)	0,172 (30,209)	0,190 (34,375)	0,203 (38,542)	0,215 (42,708)	0,248 (46,875)	0,256 (51,042)	0,284 (55,208)	
0,6500	0,0550	0,0850	0,685	0,525	0,224 (22,320)	0,263 (25,893)	0,295 (29,464)	0,323 (33,036)	0,347 (36,607)	0,400 (40,179)	0,417 (43,750)	0,462 (47,321)	
0,6000	0,0500	0,1100	0,640	0,505	0,308 (19,530)	0,365 (22,657)	0,413 (25,781)	0,454 (28,906)	0,490 (32,031)	0,557 (35,156)	0,583 (38,281)	0,636 (41,406)	
0,5500	0,0450	0,1350	0,595	0,482	0,399 (17,360)	0,472 (20,139)	0,533 (22,917)	0,584 (25,695)	0,628 (28,472)	0,699 (31,250)	0,728 (34,028)	0,779 (36,806)	
0,5000	0,0400	0,1600	0,550	0,452	0,491 (15,624)	0,577 (18,125)	0,645 (20,250)	0,701 (23,125)	0,746 (25,625)	0,811 (28,125)	0,839 (30,625)	0,880 (33,125)	
0,4500	0,0350	0,1850	0,505	0,429	0,580 (14,204)	0,673 (16,478)	0,743 (18,750)	0,797 (21,023)	0,838 (23,295)	0,891 (25,568)	0,913 (27,841)	0,941 (30,114)	
0,4000	0,0300	0,2100	0,460	0,398	0,663 (13,020)	0,756 (15,104)	0,822 (17,187)	0,870 (19,271)	0,904 (21,354)	0,942 (23,438)	0,957 (25,521)	0,974 (27,604)	
0,3500	0,0250	0,2350	0,415	0,363	0,737 (12,019)	0,826 (13,943)	0,883 (15,865)	0,921 (17,789)	0,947 (19,712)	0,972 (21,635)	0,981 (23,558)	0,990 (25,481)	
					c	14	18	22	26	30	35	39	44

Πίνακας 7: Η ισχύς, ο αναμενόμενος αριθμός αξιολογήσεων και το c για $f = 5 - 12$ στην περίπτωση ενός χαρακτηριστικού και τριών κατηγοριών ταξινόμησης

του πίνακα και αντιστοιχεί σε $\kappa = 0,429$ (σχεδόν 25% μείωση). Ας υποθέσουμε ότι θέλουμε να ανιχνεύσουμε αυτή τη μείωση με $\gamma \geq 0,80$. Επιλέγουμε το $f = 9$, καθώς αυτό είναι το μικρότερο f που δίνει $\gamma \geq 0,80$ (στην πραγματικότητα $\gamma = 0,838$). Ο μέγιστος αριθμός μαστογραφιών που απαιτούνται να αξιολογηθούν είναι $c = 30$ που αντιστοιχεί σε $\alpha = 0,044$. Ο αναμενόμενος αριθμός μαστογραφιών που πρέπει να αξιολογηθούν είναι 24.

Παράδειγμα 3: Αξιολόγηση της συμφωνίας δύο βαθμολογητών ως προς δύο χαρακτηριστικά.

Ας εξετάσουμε τώρα την περίπτωση της αξιολόγησης της συμφωνίας δύο βαθμολογητών ως προς δύο χαρακτηριστικά. Ως μηδενική υπόθεση, θεωρούμε την περίπτωση όπου $\pi_{1111} + \pi_{2222} + \pi_{2121} + \pi_{1212} = 0,892$. Αυτό δίνεται στην πρώτη σειρά του Πίνακα 8. Από όσο γνωρίζουμε, δεν υπάρχει μέτρο τύπου κ για τη συμφωνία ως προς δύο χαρακτηριστικά. Κάθε πιθανότητα διαφωνίας είναι ίση με $(1 - \pi_{1111})/50$.

Στον Πίνακα 8 παρουσιάζεται η ισχύς γ , ο αναμενόμενος αριθμός των μαστογραφιών (σε παρένθεση) και ο μέγιστος αριθμός των μαστογραφιών c για $f = 18$ έως 25 για εννέα εναλλακτικές υποθέσεις, οι οποίες αντιπροσωπεύουν μείωση του Cohen κ . Σε όλες τις περιπτώσεις, κάθε πιθανότητα διαφωνίας ισούται με $(1 - \pi_{1111})/50$. Ας εξετάσουμε την εναλλακτική υπόθεση με $\pi_{1111} + \pi_{2222} + \pi_{2121} + \pi_{1212} = 0,796$. Ας υποθέσουμε ότι θέλουμε να ανιχνεύσουμε αυτή τη μείωση με $\gamma \geq 0,80$. Επιλέγουμε το $f = 21$, καθώς αυτό είναι το μικρότερο f που δίνει $\gamma \geq 0,80$ (στην πραγματικότητα $\gamma = 0,802$). Ο μέγιστος αριθμός των μαστογραφιών που απαιτούνται να αξιολογηθούν ισούται με $c = 97$ που αντιστοιχεί σε $\alpha = 0,049$. Ο αναμενόμενος αριθμός των μαστογραφιών είναι 82.

				f								
π_{1111}	π_{2222}	π_{2121}	π_{1212}	$\pi_{1111} + \pi_{2222} + \pi_{2121} + \pi_{1212}$	18	19	20	21	22	23	24	25
0.5500	0.1500	0.1500	0.0420	0.892	0.048 (133,681)	0.049 (140,625)	0.048 (147,569)	0.049 (154,514)	0.048 (161,458)	0.049 (168,403)	0.048 (175,347)	0.049 (182,292)
0.5000	0.1400	0.1400	0.1000	0.880	0.091 (120,313)	0.096 (126,562)	0.098 (132,813)	0.099 (139,062)	0.100 (145,313)	0.103 (151,562)	0.104 (157,813)	0.105 (164,062)
0.4500	0.1300	0.1300	0.1580	0.868	0.153 (109,375)	0.163 (115,057)	0.164 (120,739)	0.173 (126,420)	0.174 (132,102)	0.183 (137,784)	0.185 (143,466)	0.191 (149,148)
0.4000	0.1200	0.1200	0.2160	0.856	0.235 (100,260)	0.252 (105,469)	0.255 (110,677)	0.271 (115,885)	0.273 (121,094)	0.288 (126,302)	0.290 (131,510)	0.303 (136,719)
0.3500	0.1100	0.1100	0.2740	0.844	0.331 (92,548)	0.354 (97,356)	0.361 (102,163)	0.383 (106,971)	0.389 (111,779)	0.409 (116,587)	0.414 (121,394)	0.433 (126,202)
0.3000	0.1000	0.1000	0.3320	0.832	0.435 (85,938)	0.464 (90,402)	0.475 (94,866)	0.501 (99,330)	0.510 (103,795)	0.535 (108,259)	0.542 (112,723)	0.565 (117,188)
0.2500	0.0900	0.0900	0.3900	0.820	0.540 (80,208)	0.573 (84,375)	0.586 (88,542)	0.616 (92,708)	0.627 (96,875)	0.653 (101,042)	0.663 (105,208)	0.686 (109,375)
0.2000	0.0800	0.0800	0.4480	0.808	0.640 (75,195)	0.674 (79,102)	0.688 (83,008)	0.717 (86,914)	0.730 (90,820)	0.755 (94,727)	0.765 (98,633)	0.787 (102,539)
0.1500	0.0700	0.0700	0.5060	0.796	0.728 (70,772)	0.760 (74,449)	0.775 (78,125)	0.802 (81,802)	0.814 (85,478)	0.836 (89,154)	0.845 (92,831)	0.863 (96,507)
0.1000	0.0600	0.0600	0.5640	0.784	0.802 (66,840)	0.831 (70,313)	0.845 (73,785)	0.867 (77,257)	0.878 (80,729)	0.896 (84,201)	0.904 (87,674)	0.918 (91,146)
c					80	86	91	97	102	108	113	119

Πίνακας 8: Η ισχύς, ο αναμενόμενος αριθμός αξιολογήσεων και το c για $f = 18 - 25$ στην περίπτωση των δύο χαρακτηριστικών

7. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην εργασία αυτή, προτείνουμε έναν μη παραμετρικό ακολουθιακό έλεγχο για την αξιολόγηση της συμφωνίας μεταξύ βαθμολογητών στην περίπτωση μεμονωμένων παρατηρήσεων. Συγκεκριμένα, εκτιμούμε τη συμφωνία με το άθροισμα των απόλυτων διαφορών των διαφωνιών των δύο βαθμολογητών ή το άθροισμα των μέγιστων απόλυτων διαφορών των διαφωνιών περισσότερων από δύο βαθμολογητών. Προκειμένου να μελετήσουμε την κατανομή του σωρευτικού αθροίσματος, χρησιμοποιούμε την τεχνική της εμφύτευσης της Μαρκοβιανής αλυσίδας.

Η νέα μέθοδος έχει τέσσερα βασικά πλεονεκτήματα. Κατ' αρχάς, δεν απαιτείται να συλλέγουμε δεδομένα σε έναν πίνακα συνάφειας, όπως απαιτούν οι περισσότεροι από τους διαθέσιμους σήμερα δείκτες συμφωνίας. Αυτό σημαίνει ότι δεν χρειάζεται να περιμένουμε πολύ για να συγκεντρώσουμε επαρκή στοιχεία για να κατασκευάσουμε έναν πίνακα συνάφειας. Το δεύτερο πλεονέκτημα είναι ότι βασίζεται σε μια απλή, εύκολα ερμηνεύσιμη «στατιστική συνάρτηση» - το άθροισμα των απόλυτων διαφορών των διαφωνιών των δύο βαθμολογητών. Το τρίτο πλεονέκτημα της προτεινόμενης μεθόδου είναι ότι ανιχνεύει ακόμη και μικρές μειώσεις στη συμφωνία με μεγάλη ισχύ και λιγότερα δεδομένα (όπως προέκυψε από την αριθμητική διερεύνηση). Τέλος, η προτεινόμενη μέθοδος μπορεί εύκολα να γενικευθεί στις περιπτώσεις των περισσότερων κατηγοριών κατάταξης, περισσότερων από δύο βαθμολογητών και περισσότερων χαρακτηριστικών.

ΑΝΑΦΟΡΕΣ

Ansell, D., Grabler, P., Whitman, S. et al. (2009). A community effort to reduce the black/white breast cancer mortality disparity in chicago. *Cancer Causes & Control*, 20, 1681–1688.

- Antzoulakos, D., Bersimis, S. and Koutras, M. (2003). On the distribution of the total number of run lengths. *Annals of the Institute of Statistical Mathematics*, 55 (4), 865–884.
- Balakrishnan, N., Bersimis, S. and Koutras, M. (2009). Run and frequency quota rules in process monitoring and acceptance sampling. *Journal of Quality Technology*, 41 (1), 66–81.
- Bersimis, S., Sachlas, A. and Papaioannou, T. (2015). Flexible designs for phase ii comparative clinical trials involving two response variables. *Statistics in Medicine*, 34 (2), 197–214.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 (1), 37–46.
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Educational and Psychological Measurement*, 70 (4), 213–220.
- Donner, A. (1998). Sample size requirements for the comparison of two or more coefficients of inter-observer agreement. *Statistics in Medicine*, 17 (10), 1157–1168.
- Donner, A. and Eliasziw, M. (1992). A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation. *Statistics in Medicine*, 11 (11), 1511–1519.
- Fleiss, J. (1981). *Statistical methods for rates and proportions*. 2 ed. New York: John Wiley & Sons.
- Hripcsaka, G. and Heitjan, D. (2002). Measuring agreement in medical informatics reliability studies. *Journal of Biomedical Informatics*, 35, 99–110.
- Koutras, M., Bersimis, S. and Antzoulakos, D. (2008). Bivariate markov chain embeddable variables of polynomial type. *Annals of the Institute of Statistical Mathematics*, 60 (1), 173–191.
- Lobbes, M., Lalji, U., Nelemans, P. et al. (2015). The quality of tumor size assessment by contrast-enhanced spectral mammography and the benefit of additional breast mri. *Journal of Cancer*, 6 (2), 144–150.
- Olivier, J. and Bell, M. (2013). Effect sizes for 2×2 contingency tables. *PLOS One*, 8 (3), e58777.
- Uebersax, J. (1988). Validity inferences from interobserver agreement. *Psychological Bulletin*, 104 (3), 405–416.

ΔΙΑΤΗΡΗΣΙΜΟΤΗΤΑ ΚΑΙ ΕΠΙΤΕΥΞΙΜΟΤΗΤΑ ΣΤΟ ΗΜΙΜΑΡΚΟΒΙΑΝΟ ΣΥΣΤΗΜΑ ΣΕ ΣΤΟΧΑΣΤΙΚΟ ΠΕΡΙΒΑΛΛΟΝ

Α. Συμεωνίδου, Α. Παπαδοπούλου

Τμήμα Μαθηματικών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
annaassymeon2@hotmail.com, apapado@math.auth.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία μελετούμε το πρόβλημα της επιτευξιμότητας και της διατηρησιμότητας μίας επιθυμητής δομής στόχου σε ένα μη ομογενές Ημιμαρκοβιανό σύστημα (ΜΟΗΜΣ) σε στοχαστικό περιβάλλον. Για την παραπάνω μελέτη είναι απαραίτητο να γνωρίζουμε την εξίσωση της αναμενόμενης πληθυσμιακής δομής του συστήματος. Για το λόγο αυτόν και εξαιτίας της πολυπλοκότητας του Ημιμαρκοβιανού συστήματος μετασχηματίζουμε το Ημιμαρκοβιανό μοντέλο σε ισοδύναμο μη ομογενές Μαρκοβιανό (ΜΟΜΣ) ορίζοντάς το σε σχέση με τις βασικές παραμέτρους του αρχικού συστήματος. Κατόπιν, προχωρούμε στη μελέτη της επίτευξης και διατήρησης μιας δομής στόχου για το Μαρκοβιανό σύστημα και τα αποτελέσματα που προκύπτουν χρησιμοποιούνται για την επίλυση των ερωτημάτων που αφορούν στο αρχικό. Τέλος, τα θεωρητικά αποτελέσματα εικονογραφούνται με αριθμητικά παραδείγματα.

Λέξεις κλειδιά: Μαρκοβιανά συστήματα, Ημιμαρκοβιανά συστήματα, Διατηρησιμότητα, Επιτευξιμότητα, Στοχαστικό περιβάλλον

1. ΕΙΣΑΓΩΓΗ

Θεωρούμε ένα σύστημα τα μέλη του οποίου ταξινομούνται σε k βαθμίδες, με βάση κάποιο χαρακτηριστικό τους. Οι βαθμίδες αυτές συνιστούν το χώρο καταστάσεων $S_H = \{1, 2, \dots, k\}$ του συστήματος, ο οποίος είναι διακριτός και πεπερασμένος. Οι καταστάσεις είναι ασυμβίβαστες ανά δύο, με την έννοια ότι κάθε μέλος δεν μπορεί να ανήκει σε περισσότερες από μία, σε συγκεκριμένη χρονική στιγμή. Τα μέλη του συστήματος που βρίσκονται σε μία συγκεκριμένη κατάσταση μπορούν, σε χρόνο διακριτό, είτε να μετακινηθούν σε κάποια άλλη, είτε να παραμείνουν στη κατάσταση αυτή για κάποιες χρονικές μονάδες.

Το σύστημα κάθε χρονική στιγμή περιγράφεται από το διάνυσμα $N(t) = (N_1(t), N_2(t), \dots, N_k(t))$, όπου $N_i(t)$ είναι το πλήθος των μελών του συστήματος τη χρονική στιγμή t ενώ το συνολικό μέγεθος του συστήματος

συμβολίζεται με $T(t)$. Στην παρούσα εργασία θεωρούμε ότι το σύστημα είναι ανοιχτό, επιτρέπεται δηλαδή η έξοδος μελών από το σύστημα αλλά και η είσοδος νέων σε αυτό καθώς και ότι η προσαύξηση του πληθυσμού δίνεται από τη σχέση $\Delta T(t) = T(t) - T(t-1) \geq 0$, $t = 1, 2, \dots$. Η έξοδος θεωρείται μία επιπλέον κατάσταση, διαχωρισμένη από το χώρο καταστάσεων S_H . Είτε όμως λόγω αύξησης του συνολικού μεγέθους του συστήματος είτε λόγω αποχωρήσεων μελών από αυτό, δημιουργούνται κενές θέσεις οι οποίες καλύπτονται από νέα μέλη. Οι εισοδοί των νεοεισερχόμενων στο σύστημα πραγματοποιούνται βάσει του στοχαστικού διανύσματος $p_{H_0}(t) = (p_{H_{01}}(t), p_{H_{02}}(t), \dots, p_{H_{0k}}(t))$, όπου $p_{H_{0k}}(t)$ είναι η πιθανότητα εισόδου ενός νέου μέλους στο σύστημα στην κατάσταση i τη χρονική στιγμή t .

Επειδή η έξοδος δεν ανήκει στο χώρο καταστάσεων, εξυπηρετεί να εισάγουμε τις ταυτότητες μέλους (Vassiliou & Papadopoulou 1992). Συγκεκριμένα, θεωρούμε ότι κάθε μέλος του συστήματος έχει στην κατοχή του μία ταυτότητα η οποία μετακινείται μαζί του μέσα στο σύστημα. Όταν το μέλος αποχωρήσει, παραδίδει ταυτόχρονα την ταυτότητα που έχει στην κατοχή του στο νεοεισερχόμενο μέλος που θα πάρει την θέση του. Η ταυτότητα μέλους θα συνεχίσει να μετακινείται με το νέο μέλος όπως ακριβώς με τον προκάτοχό της. Όταν το μέγεθος του συστήματος αυξάνει, δημιουργούνται νέες ταυτότητες μέλους τις οποίες παραλαμβάνουν τα νεοεισερχόμενα μέλη και συμπεριφέρονται ακριβώς όπως οι προηγούμενες.

Για την μελέτη του συστήματος χρησιμοποιούμε τους παρακάτω συμβολισμούς:

- $\mathbf{P}(t) = (p_{ij}(t))$, $i, j = 1, 2, \dots, k$: ο πίνακας πιθανοτήτων μετάβασης της εμβαπτισμένης Μαρκοβιανής αλυσίδας για το χρονικό διάστημα $[t, t+1)$,
- $\mathbf{H}(m) = (h_{ij}(m))$, $i, j = 1, 2, \dots, k$, $m = 1, 2, \dots$: ο πίνακας πιθανοτήτων χρόνων παραμονής,
- $\mathbf{C}(t, m) = \mathbf{P}(t) \diamond \mathbf{H}(m)$, $t = 0, 1, 2, \dots$, $m = 1, 2, \dots$: Hadamard γινόμενο των πινάκων $\mathbf{H}(m)$ και $\mathbf{P}(t)$, με στοιχεία: $c_{ij}(t, m) = p_{ij}(t)h_{ij}(m)$.
- ${}^>\mathbf{W}(n, t) = ({}^>w_i(n, t))$, $i = 1, 2, \dots, k$: διαγώνιος πίνακας με τα διαγώνια στοιχεία του ίσα με τις συναρτήσεις επιβίωσης των χρόνων παραμονής. Το i στοιχείο της διαγωνίου ισούται με:

$$\sum_{m=n+1}^{\infty} \sum_{j=1}^k p_{ij}(t)h_{ij}(m) = 1 - \sum_{m=1}^n \sum_{j=1}^k p_{ij}(t)h_{ij}(m).$$

Στόχος της εργασίας αυτής είναι η μελέτη της διατηρησιμότητας μιας δομής αλλά και της επιτευξιμότητας ενός στόχου σε ένα ΜΟΗΜΣ σε στοχαστικό περιβάλλον.

Πρώτα, πρέπει να εξετάσουμε την εξίσωση που περιγράφει την αναμενόμενη πληθυσμιακή δομή του συστήματος κάθε χρονική στιγμή. Για το λόγο αυτό μετασχηματίζουμε το ΜΟΗΜΣ σε ΜΟΜΣ (Papadopoulou & Vassiliou 1993). Στη συνέχεια εισάγουμε το ΜΟΗΜΣ σε στοχαστικό περιβάλλον (Tsantas & Vassiliou 1993) το οποίο είναι ένα σύστημα που τα μέλη του έχουν τη δυνατότητα κάθε χρονική στιγμή να επιλέξουν από ένα σύνολο πιθανοτήτων μετάβασης τον τρόπο με τον οποίο θα κινηθούν και βρίσκουμε την εξίσωση της αναμενόμενης πληθυσμιακής δομής του. Τέλος, ορίζουμε τη σχετική δομή, υπολογίζουμε την πιθανότητα διατήρησής της σε διαστήματα μιας χρονικής μονάδας και μελετάμε την επιτευξιμότητα της δομής στόχου στο ΜΟΗΜΣ σε στοχαστικό περιβάλλον. Σύμφωνα με τα προηγούμενα, η πρωτοτυπία της εργασίας ευρίσκεται στη σύνδεση γνωστών από τη βιβλιογραφία μεθοδολογιών: (α) μετατροπή του Ημιμαρκοβιανού συστήματος σε Μαρκοβιανό (Papadopoulou & Vassiliou, 1993), (β) μελέτη του Μαρκοβιανού συστήματος σε στοχαστικό περιβάλλον (Tsantas & Vassiliou, 1993).

2. ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΣ ΤΟΥ ΜΟΗΜΣ ΣΕ ΜΟΜΣ

Για την εύρεση της επιθυμητής δομής και τη διατήρησής της, είναι απαραίτητο να μετασχηματίσουμε το ΜΟΗΜΣ σε ΜΟΜΣ σύμφωνα με τα παρακάτω.

Για κάθε μια από τις καταστάσεις δημιουργούμε ένα σύνολο υποκαταστάσεων με χαρακτηριστικό τον παρόντα χρόνο παραμονής. Έτσι οι ταυτότητες μέλη που βρίσκονται στην i και έχουν συγκεκριμένο χρόνο παραμονής κάθε χρονική στιγμή ανήκουν στην αντίστοιχη υποκατάσταση. Ορίζουμε από την αρχή της διαδικασίας το μέγιστο δυνατό χρόνο παραμονής b_j για κάθε κατάσταση i , $i = 1, 2, \dots, k$. Οι μέγιστες αυτές τιμές μπορεί να είναι ίδιες για όλο το σύστημα ή να διαφοροποιούνται από κατάσταση σε κατάσταση. Οι υποκαταστάσεις μίας συγκεκριμένης κατάστασης i κατά το χρονικό διάστημα $(t-1, t]$ θα είναι όσες και οι δυνατοί χρόνοι παραμονής σε αυτήν, δηλαδή $b_j, j = 1, 2, \dots, k$. Το σύνολο των υποκαταστάσεων του συστήματος θα είναι σε πλήθος $b_1 + b_2 + \dots + b_k$, και ο νέος χώρος καταστάσεων του ΜΟΜΣ θα είναι

$$S = \{(1,0), (1,1), \dots, (1, b_1 - 1), (2,0), (2,1), \dots, (2, b_2 - 1), \dots, (k,0), \dots, (k, b_k - 1)\},$$

όπου για κάθε ζεύγος (i, j) η πρώτη συνιστώσα δηλώνει την κατάσταση i στην οποία βρίσκεται η ταυτότητα μέλος στον χρόνο t , και η δεύτερη τον χρόνο παραμονής στην κατάσταση i έως το χρόνο t .

Ορίζουμε το διάνυσμα αναμενόμενης πληθυσμιακής δομής για κάθε κατάσταση

$$\bar{N}(t) = \{\bar{N}_1(t,0), \bar{N}_1(t,1), \dots, \bar{N}_1(t, b_1 - 1), \bar{N}_2(t,0), \dots, \bar{N}_k(t,0), \dots, \bar{N}_k(t, b_k - 1)\},$$

όπου $\bar{N}_j(t,s)$ είναι ο αναμενόμενος αριθμός των ταυτοτήτων μελών που βρίσκονται στη κατάσταση i στο χρόνο t με s χρονικές μονάδες παραμονής σε αυτήν.

Οι προσλήψεις νέων μελών τη χρονική στιγμή t , για την κάλυψη των θέσεων που δημιουργούνται λόγω αύξησης του μεγέθους του συστήματος, θα γίνονται μόνον στις καταστάσεις με χρόνο παραμονής 0. Συνεπώς, το διάνυσμα πιθανοτήτων εισόδου τη χρονική στιγμή t θα έχει μη μηδενικά στοιχεία μόνον στις συνιστώσες που αντιστοιχούν στις καταστάσεις αυτές, δηλαδή θα είναι της μορφής

$$p_0(t) = (p_{(1,0)}(t), 0, \dots, 0, p_{(2,0)}(t), \dots, 0, p_{(k,0)}(t), 0, \dots, 0).$$

Στο μετασχηματισμένο πλέον MOMΣ ορίζουμε τις πιθανότητες μετάβασης και παραμονής για κάθε χρονική στιγμή $t = 0, 1, 2, \dots$.

Θεωρούμε την πιθανότητα

Prob { μία ταυτότητα μέλους που εισήλθε στην κατάσταση i τη χρονική στιγμή $t-s$, έπειτα από s χρονικές μονάδες παραμονής παραμένει για μία ακόμα χρονική μονάδα στην i και τη χρονική στιγμή $t+1$ κάνει μία μετάβαση στην j , δεδομένου ότι κατά το χρονικό διάστημα $(t-s, t]$ δεν πραγματοποίησε καμία μετάβαση } = $c_{ij}(t-s, s+1) / \sum_j w_j(s, t-s)$, $s = 0, 1, \dots, b_i - 1$, $i, j = 1, 2, \dots, k$.

Για λόγους συντομίας συμβολίζουμε

$$cw_{ij}(t-s, s+1) = \frac{c_{ij}(t-s, s+1)}{\sum_j w_j(s, t-s)}.$$

Οι πιθανότητες παραμονής στην ίδια κατάσταση προκύπτουν ως το συμπλήρωμα των πιθανοτήτων κατά τις οποίες μία ταυτότητα μέλους πραγματοποιεί μετάβαση. Έτσι,

$$\begin{aligned} & \text{Prob \{ μία ταυτότητα μέλους που τη χρονική στιγμή } t \text{ ανήκει στην κατάσταση} \\ & \quad (j, s-1) \text{ παραμένει στην κατάσταση } j \text{ στο χρόνο } t+1 \} = \\ & 1 - \sum_{r=1}^k c_{jr}(t-s+1, s) / \sum_j w_j(s-1, t-s+1) = 1 - \sum_{r=1}^k cw_{jr}(t-s+1, s), \\ & \quad s = 1, 2, \dots, b_j - 1, \quad i, j = 1, 2, \dots, k. \end{aligned}$$

Στη συνέχεια, ορίζουμε τον πίνακα πιθανοτήτων παραμονής $\sum C_i$, $i = 1, 2, \dots, k$, διάστασης $b_i \times (b_i - 1)$ για το χρονικό διάστημα $[t, t+1)$

$$\sum C_i = \begin{bmatrix} 1 - \sum_{h=1}^k \alpha w_{ih}(t, 1) & 0 & \dots & 0 \\ 0 & 1 - \sum_{h=1}^k \alpha w_{ih}(t-1, 2) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 - \sum_{h=1}^k \alpha w_{ih}(t-b_i+2, b_i-1) \\ 0 & 0 & \dots & 0 \end{bmatrix}.$$

Για κάθε μία κατάσταση ορίζουμε το διάνυσμα πιθανοτήτων μετάβασης, διάστασης $b_j \times 1$, κατά το χρονικό διάστημα $[t, t+1)$

$$\hat{c}_{ij} = (cw_{ij}(t,1), cw_{ij}(t-1,2), \dots, cw_{ij}(t-b_i+1, b_i))', \quad i, j = 1, \dots, k.$$

Από τα παραπάνω προκύπτει ο νέος πίνακας πιθανοτήτων μετάβασης, διάστασης $\left(\sum_{i=1}^k b_i \times \sum_{i=1}^k b_i \right)$ όπως παρακάτω:

$$\mathbf{M}(t) = \begin{bmatrix} \hat{c}_{11} & \Sigma \mathbf{C}_1 & \hat{c}_{12} & 0 & \dots & \hat{c}_{1k} & 0 \\ \hat{c}_{21} & 0 & \hat{c}_{22} & \Sigma \mathbf{C}_2 & \dots & \hat{c}_{2k} & 0 \\ \hat{c}_{31} & 0 & \hat{c}_{32} & 0 & \dots & \hat{c}_{3k} & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \hat{c}_{k1} & 0 & \hat{c}_{k2} & 0 & \dots & \hat{c}_{kk} & \Sigma \mathbf{C}_k \end{bmatrix}.$$

Η i γραμμή του πίνακα περιέχει τις πιθανότητες των δυνατών μεταβάσεων από την i σε όλες τις καταστάσεις (ακόμα και στην ίδια) έπειτα από όλους του δυνατούς χρόνους παραμονής, καθώς και τις πιθανότητες παραμονής σε αυτήν. Στη γενική περίπτωση οι πίνακες $\Sigma \mathbf{C}_i, \hat{c}_{ij}$ έχουν διαφορετική διάσταση λόγω της

ασυμβατότητας των b_i . Αν θεωρήσουμε $b = \max\{b_i\}$ τότε οι πίνακες $\Sigma \mathbf{C}_i, \hat{c}_{ij}$ μπορούν να είναι διάστασης $b \times (b-1)$ και $b \times 1$ συμπληρώνοντας μηδέν στις θέσεις που χρειάζεται αντίστοιχα. Αθροίζοντας τα στοιχεία της i γραμμής, εύκολα προκύπτει ότι ο πίνακας $\mathbf{M}(t)$ είναι στοχαστικός όπως είναι και το αναμενόμενο.

3. ΤΟ ΜΟΗΜΣ ΣΕ ΣΤΟΧΑΣΤΙΚΟ ΠΕΡΙΒΑΛΛΟΝ

Στην ενότητα αυτή, θεωρούμε ότι ο πίνακας πιθανοτήτων μετάβασης $\mathbf{M}(t)$ που προέκυψε σύμφωνα με τον μετασχηματισμό της παραγράφου 2, επιλέγεται για κάθε χρονική στιγμή $t = 1, 2, 3, \dots$ από ένα σύνολο πινάκων

$$\mathbf{M}_J = \{ \mathbf{M}_1(t), \mathbf{M}_2(t), \dots, \mathbf{M}_V(t) \}.$$

Η επιλογή του επόμενου πίνακα ορίζουμε να γίνεται σύμφωνα με τις δεσμευμένες πιθανότητες

$$z_{ij}(t) = \Pr \left\{ \mathbf{M}(t) = \mathbf{M}_j(t) \mid \mathbf{M}(t-1) = \mathbf{M}_i(t-1) \right\}, \quad i, j = 1, 2, \dots, V,$$

και θεωρούμε ότι η διαδικασία ξεκινάει με το πρώτο στοιχείο του συνόλου $\mathbf{M}_J$. Ο πίνακας των πιθανοτήτων αυτών έχει την μορφή

$$\mathbf{Z}(t) = \begin{bmatrix} z_{11}(t) & z_{12}(t) & \dots & z_{1V}(t) \\ z_{21}(t) & z_{22}(t) & \dots & z_{2V}(t) \\ \dots & \dots & \dots & \dots \\ z_{V1}(t) & z_{V2}(t) & \dots & z_{VV}(t) \end{bmatrix},$$

είναι στοχαστικός και ορίζει μία μη ομογενή Μαρκοβιανή αλυσίδα, εφόσον η επιλογή του πίνακα πιθανοτήτων μετάβασης τη χρονική στιγμή t εξαρτάται μόνον από την επιλογή του πίνακα που έγινε τη στιγμή $t-1$. Η Μαρκοβιανή αλυσίδα $\mathbf{Z}(t) = \{z_{ij}(t)\}_{i,j \in \mathbf{J}}$ ονομάζεται *συμβιβαστική μη ομογενή αλυσίδα*.

Καθώς η επιλογή του πίνακα πιθανοτήτων μετάβασης γίνεται με τον στοχαστικό μηχανισμό που μόλις περιγράψαμε, έχει νόημα να μελετήσουμε την μέση τιμή των στοιχείων του συνόλου $\mathbf{M}_{\mathbf{J}}$, την οποία συμβολίζουμε με $E(\mathbf{M}(t))$. Η ζητούμενη μέση τιμή δίνεται από το παρακάτω θεώρημα, η απόδειξη του οποίου ακολουθεί ανάλογα τα βήματα της απόδειξης του θεωρήματος 2.1 στην εργασία των Tsantas & Vassiliou (1993).

Θεώρημα 1. Θεωρούμε ΜΟΗΜΣ σε στοχαστικό περιβάλλον το οποίο ορίζεται από το σύνολο $\mathbf{M}_{\mathbf{J}}$, τις ακολουθίες $\{\mathbf{Z}(t)\}_{t=1}^{\infty}$, $\{T(t)\}_{t=0}^{\infty}$, $\{\mathbf{p}_0(t)\}_{t=0}^{\infty}$ και τις αρχικές τιμές $\mathbf{M}_1(0)$, $\mathbf{N}(0)$. Ο αναμενόμενος πίνακας πιθανοτήτων μετάβασης για το χρονικό διάστημα $[t, t+1)$ δίνεται από τη σχέση

$$E(\mathbf{M}(t)) = \sum_{h \in \mathbf{J}} \mathbf{e}_h \left[\mathbf{e}_1 \prod_{n=1}^t \mathbf{Z}(n) \right]^T \mathbf{M}_h(t) = \left\{ \sum_{h \in \mathbf{J}} \mathbf{e}_h \left[\mathbf{e}_1 \prod_{n=1}^t \mathbf{Z}(n) \right]^T m_{h,jj}(t) \right\}_{i,j \in \mathbf{S}}, \quad \text{όπου}$$

$\mathbf{e}_h$: διάνυσμα διάστασης $1 \times V$ με μονάδα στο στοιχείο h και μηδέν οπουδήποτε αλλού,

$m_{h,jj}(t)$: το (i, j) στοιχείο του πίνακα $\mathbf{M}_h(t)$.

Οι πιθανότητες μετάβασης και παραμονής, κάνοντας χρήση του Θεωρήματος 1 δίνονται από τις αντίστοιχες παραστάσεις

$$\sum_{h \in \mathbf{J}} \mathbf{e}_h \left[\mathbf{e}_1 \prod_{n=1}^t \mathbf{Z}(n) \right]^T cw_{h,jj}(t-s, s+1), \quad (1)$$

$$\sum_{h \in \mathbf{J}} \mathbf{e}_h \left[\mathbf{e}_1 \prod_{n=1}^t \mathbf{Z}(n) \right]^T \left(1 - \sum_{r=1}^k cw_{h,jr}(t-s, s+1) \right). \quad (2)$$

Επόμενο ερώτημα είναι η εύρεση της εξίσωσης της αναμενόμενης πληθυσμιακής δομής του ΜΟΗΜΣ σε στοχαστικό περιβάλλον συναρτήσει των βασικών παραμέτρων του συστήματος $\{\mathbf{M}_j(t)\}_{t=1}^{\infty}, \{\mathbf{Z}(t)\}_{t=1}^{\infty}, \{T(t)\}_{t=0}^{\infty}, \{\mathbf{p}_0(t)\}_{t=0}^{\infty}$ και των αρχικών τιμών $\mathbf{M}_1(0), \mathbf{N}(0)$.

Συμβολίζουμε με $N_j(t,s)$ τη τυχαία μεταβλητή που δηλώνει το πλήθος των ταυτοτήτων μελών που βρίσκονται στην κατάσταση j , με χρόνο παραμονής s τη χρονική στιγμή t και με $N_{ji}(t,s)$ τη τυχαία μεταβλητή που δηλώνει το πλήθος των ταυτοτήτων μελών που έχουν παραμονή s χρονικών μονάδων στη κατάσταση i στο χρόνο t και μεταβαίνουν από τη κατάσταση i στη j κατά το διάστημα $[t, t+1)$. Εύκολα αντιλαμβάνεται κανείς ότι οι ροές αυτές ακολουθούν τη διωνυμική κατανομή

$$B\left(N_j(t,s), \sum_{h \in \mathbf{J}} \mathbf{e}_h \left[\mathbf{e}_1 \prod_{n=1}^t \mathbf{Z}(n) \right]^T cw_{h,jj}(t-s, s+1)\right).$$

Έστω ότι $N_{0j}(t+1)$ ορίζει το πλήθος των προσλήψεων που πραγματοποιούνται στη κατάσταση j , κατά το διάστημα $(t, t+1]$.

Οι ταυτότητες μέλη που βρίσκονται στη κατάσταση $(j,0)$, $j=1,2,\dots,k$ τη χρονική στιγμή $t+1$ προέρχονται είτε από μεταβάσεις από οποιαδήποτε κατάσταση, είτε από προσλήψεις και το πλήθος τους δίνεται από τη σχέση

$$N_j(t+1,0) = \sum_{i=1}^k \sum_{s=0}^{b_i-1} N_{ji}(t,s) + N_{0j}(t+1), \quad j=1,2,\dots,k. \quad (3)$$

Αντίστοιχα, το πλήθος των ταυτοτήτων μελών που βρίσκονται στην κατάσταση (j,s) , $j=1,\dots,k$, $s=1,\dots,b_j-1$ την ίδια χρονική στιγμή είναι

$$N_j(t+1,s) = N_j(t,s-1) - \sum_{i=1}^k N_{ji}(t,s-1), \quad j=1,\dots,k, s=1,\dots,b_j-1. \quad (4)$$

Κατά τον υπολογισμό της μέσης τιμής των δομών αυτών οι σχέσεις (3), (4) γίνονται

$$E\left(N_j(t+1,0)\right) = \sum_{i=1}^k \sum_{s=0}^{b_i-1} E\left(N_{ji}(t,s)\right) + \Delta T(t+1)p_{0(j,0)}(t+1), \quad (5)$$

$$E\left(N_j(t+1,s)\right) = E\left(N_j(t,s-1)\right) - \sum_{i=1}^k E\left(N_{ji}(t,s-1)\right), \quad (6)$$

$$j=1,2,\dots,k, \quad s=1,\dots,b_j-1.$$

Καθώς οι ροές $N_{ij}(t)$, $i, j = 1, \dots, k$ ακολουθούν πολυωνυμική κατανομή, ισχύει ότι

$$\begin{aligned} E(N_{ij}(t, s)) &= E(E(N_{ij}(t, s) | N_j(t, s), \mathbf{M}_j)) \\ &= E(N_j(t, s)) \cdot \sum_{h \in \mathbf{J}} \mathbf{e}_h \left[\mathbf{e}_1 \prod_{n=1}^t \mathbf{Z}(n) \right]^T cw_{h,ij}(t-s, s+1). \end{aligned}$$

Συμβολίζουμε με $\bar{\mathbf{N}}(t)$ την μέση τιμή. Οι σχέσεις (5), (6) γίνονται

$$\begin{aligned} \bar{N}_j(t+1, 0) &= \sum_{i=1}^k \sum_{s=0}^{b_i-1} \bar{N}_i(t, s) \sum_{h \in \mathbf{J}} \mathbf{e}_h \left[\mathbf{e}_1 \prod_{n=1}^t \mathbf{Z}(n) \right]^T cw_{h,ij}(t-s, s+1) \\ &\quad + \Delta T(t+1) p_{0(j,0)}(t+1) \end{aligned} \quad (7)$$

$$\begin{aligned} \bar{N}_j(t+1, s) &= \bar{N}_j(t, s-1) \left(1 - \sum_{r=1}^k \sum_{s=0}^{b_r-1} \sum_{h \in \mathbf{J}} \mathbf{e}_h \left[\mathbf{e}_1 \prod_{n=1}^t \mathbf{Z}(n) \right]^T cw_{h,ir}(t-s, s+1) \right) \\ &\quad j = 1, 2, \dots, k, \quad s = 1, \dots, b_j - 1. \end{aligned} \quad (8)$$

Από τα παραπάνω, προκύπτει ότι η εξίσωση της αναμενόμενης πληθυσμιακής δομής σε μορφή πινάκων δίνεται από τη σχέση

$$\bar{\mathbf{N}}(t+1) = \bar{\mathbf{N}}(t) E(\mathbf{M}(t)) + \Delta T(t+1) \mathbf{p}_0(t+1), \quad (9)$$

όπου $\bar{\mathbf{N}}(t)$ το διάνυσμα της αναμενόμενης πληθυσμιακής δομής του ΜΟΗΜΣ σε στοχαστικό περιβάλλον.

4. ΔΙΑΤΗΡΗΣΙΜΟΤΗΤΑ

Σε συστήματα ανθρώπινου δυναμικού η διατήρηση μίας (αναμενόμενης) δομής αποτελεί σημαντικό πρόβλημα. Στην εργασία αυτή μελετούμε την διατηρησιμότητα του ΜΟΗΜΣ σε στοχαστικό περιβάλλον, ασκώντας έλεγχο στα διανύσματα πιθανοτήτων εισόδου. Στις περισσότερες περιπτώσεις, οι προσλήψεις αποτελούν το ελκυστικότερο μέσο ελέγχου, καθώς η είσοδος νέων μελών έχει μικρότερο δυνατό αντίκτυπο στα ήδη υπάρχοντα μέλη του συστήματος. Η μελέτη του προβλήματος είναι στοχαστική, καθώς οι μη ελεγχόμενες ροές (απώλειες, μετακινήσεις) αντιμετωπίζονται ως τυχαίες μεταβλητές. Σκοπός μας είναι να βρούμε εκείνα τα διανύσματα πιθανοτήτων εισόδου ώστε με τις ανάλογες προσλήψεις να διατηρηθεί η ήδη υπάρχουσα δομή, για την επόμενη χρονική στιγμή.

Επειδή το σύστημα είναι ανοιχτό και το συνολικό του μέγεθος μπορεί να μεταβάλλεται κάθε χρονική στιγμή, δεν έχει νόημα να μελετήσουμε τη διατήρηση της πληθυσμιακής του δομής. Ορίζουμε τη σχετική δομή και ανάγουμε το πρόβλημα σε

πρόβλημα διατήρησης της σχετικής δομής. Ορίζουμε το στοχαστικό διάνυσμα γραμμής

$$\mathbf{q}(t) = (q_{(1,0)}(t), q_{(1,1)}(t), \dots, q_{(1,b_1-1)}(t), q_{(2,0)}(t), \dots, q_{(k,b_k-1)}(t)) ,$$

όπου η συνιστώσα $q_{(i,s)}(t) = \frac{N_{ij}(t,s)}{T(t)}$, $i = 1, \dots, k$, $s = 0, 1, \dots, b_i - 1$ εκφράζει το σχετικό μέγεθος της κατάστασης (i, s) τη χρονική στιγμή t .

Ορισμός 1. Η σχετική δομή $\mathbf{q}(t)$ ενός MOMΣ σε στοχαστικό περιβάλλον με μεταβλητές ελέγχου τις προσλήψεις είναι διατηρήσιμη κατά το χρονικό διάστημα $[t, t+1]$ εάν και μόνον εάν υπάρχει διάνυσμα πιθανοτήτων εισόδου $\mathbf{p}_0(t+1)$ τέτοιο ώστε να ισχύει $\mathbf{q} = \mathbf{q}(t) = \mathbf{q}(t+1)$.

Στη μελέτη του προβλήματος σε στοχαστικό περιβάλλον, δεν είμαστε σίγουροι για την διατήρηση μιας σχετικής δομής $\mathbf{q}(t)$. Κάποιες δομές μπορούν να διατηρηθούν με μεγαλύτερη ευκολία από ότι κάποιες άλλες. Προκειμένου να διακρίνουμε την ευκολία της διατήρησης μιας σχετικής δομής, εισάγουμε την αντίστοιχη πιθανότητα:

$$P(t) = \text{Prob} \{ \text{η σχετική δομή } \mathbf{q}(t) \text{ διατηρείται κατά το χρονικό διάστημα } [t, t+1] \} .$$

Βρίσκουμε την πιθανότητα αυτή, με την βοήθεια της παρακάτω πρότασης (Bartholomew, 1982).

Πρόταση 1. Η σχετική δομή μιας κατάστασης $(j, 0)$, $j = 1, 2, \dots, k$ διατηρείται κατά το διάστημα $[t, t+1]$ εάν και μόνον εάν ισχύει

$$T(t+1)q_{(j,0)}(t+1) \geq \sum_{i=1}^k \sum_{s=0}^{b_i-1} N_{ij}(t,s) .$$

Η πιθανότητα διατήρησης είναι

$$P(t) = \text{Pr ob} \left[T(t+1)q_{(j,0)}(t+1) - \sum_{i=1}^k \sum_{s=0}^{b_i-1} N_{ij}(t,s) \geq 0; j = 1, \dots, k \mid \mathbf{q} \right] , \quad t = 0, 1, 2, \dots$$

Θεωρούμε την τυχαία μεταβλητή

$$S_j(t) = T(t+1)q_{(j,0)}(t+1) - \sum_{i=1}^k \sum_{s=0}^{b_i-1} N_{ij}(t,s) \geq 0 , \quad j = 1, 2, \dots, k ,$$

η οποία ακολουθεί την πολυωνυμική κατανομή, με

$$E(S_j(t)) = T(t+1)q_{(j,0)}(t+1) - T(t) \sum_{i=1}^k \sum_{s=0}^{b_i-1} q_{(i,s)}(t) \sum_{h \in \mathbf{J}} \mathbf{e}_h \left[\mathbf{e}_1 \prod_{n=1}^t \mathbf{Z}(n) \right]^T cw_{h,jj}(t-s, s+1),$$

$$Var(S_j(t)) = T(t) \sum_{i=1}^k \sum_{s=0}^{b_i-1} \sum_{h \in \mathbf{J}} \mathbf{e}_h \left[\mathbf{e}_1 \prod_{n=1}^t \mathbf{Z}(n) \right]^T cw_{h,jj}(t-s, s+1) \cdot \left(1 - \sum_{h \in \mathbf{J}} \mathbf{e}_h \left[\mathbf{e}_1 \prod_{n=1}^t \mathbf{Z}(n) \right]^T cw_{h,jj}(t-s, s+1) \right) q_{(i,s)}(t),$$

Τέλος, κάνοντας χρήση του Κ.Ο.Θ. η ζητούμενη πιθανότητα γίνεται

$$P(t) = \Pr \left[S_j(t) \geq 0; j = 1, \dots, k \right] = \Pr \left[Z_j \geq \frac{-E(S_j(t))}{\sqrt{Var(S_j(t))}}; j = 1, \dots, k \right],$$

όπου η τυχαία μεταβλητή Z_j ακολουθεί την τυποποιημένη κανονική κατανομή. Συμβολίζουμε με $\mathcal{U}_t$ το σύνολο των σχετικών δομών που θα μπορούσαν να διατηρηθούν κατά το χρονικό $[t, t+1]$ με πιθανότητα 1, δηλαδή $\mathcal{U}_t = \{\mathbf{q} : P(t) = 1\}$.

Οι σχετικές δομές που αποτελούν το σύνολο $\mathcal{U}_t$ δίνονται από το παρακάτω θεώρημα (Tsantas & Vassiliou, 1993).

Θεώρημα 2. Θεωρούμε ένα ΜΟΜΣ σε στοχαστικό περιβάλλον με τις ακολουθίες $\{T(t)\}_{t=0}^{\infty}$, $\{\mathbf{p}_0(t)\}_{t=0}^{\infty}$, $\{\mathbf{Z}(t)\}_{t=1}^{\infty}$, αρχικό πίνακα πιθανοτήτων μετάβασης $\mathbf{M}_1(0)$ και $\mathbf{N}(0)$. Τότε καθώς ο συνολικός αριθμός των ταυτοτήτων μελών τείνει στο άπειρο το σύνολο $\mathcal{U}_t$ αποτελείται από όλες τις σχετικές δομές για τις οποίες ισχύει η ανίσωση

$$q_{(j,0)} > \frac{T(t)}{T(t+1)} \sum_{i=1}^k \sum_{s=0}^{b_i-1} \sum_{h \in \mathbf{J}} \mathbf{e}_h \left[\mathbf{e}_1 \prod_{n=1}^t \mathbf{Z}(n) \right]^T cw_{h,jj}(t-s, s+1) q_{(i,s)}(t), j = 1, \dots, k.$$

Θεωρούμε $\bar{\mathbf{q}}(t)$ την αναμενόμενη τιμή της σχετικής δομής $\mathbf{q}(t)$. Ισχύει

$$\bar{\mathbf{q}}(t) = E(\mathbf{q}(t)) = E\left(\frac{\mathbf{N}(t)}{T(t)}\right) = \frac{1}{T(t)} E(\mathbf{N}(t)) = \frac{\bar{\mathbf{N}}(t)}{T(t)}.$$

Από την (9) η εξίσωση της αναμενόμενης σχετικής δομής σε μορφή πινάκων δίνεται από την

$$\bar{\mathbf{q}}(t+1) = \alpha(t) \bar{\mathbf{q}}(t) E(\mathbf{M}(t)) + \beta(t+1) \mathbf{p}_0(t+1),$$

όπου $\alpha(t) = T(t)/T(t+1)$, $\beta(t) = \Delta T(t+1)/T(t+1)$.

5. ΕΠΙΤΕΥΞΙΜΟΤΗΤΑ

Στην προηγούμενη παράγραφο βρήκαμε την πιθανότητα διατήρησης μιας σχετικής δομής. Ένα ερώτημα που προκύπτει είναι πώς το σύστημα μπορεί να επιτύχει την προαποφασισμένη δομή $\mathbf{q}^*$ (δομή στόχος). Ο σχεδιασμός στρατηγικών για την επίτευξη της δομής στόχου, περιλαμβάνει την εύρεση των κατάλληλων διανυσμάτων πιθανοτήτων εισόδου $\mathbf{p}_0(t)$.

Στην παράγραφο αυτή μελετούμε το πρόβλημα με στρατηγική ορισμένου χρόνου, επιδιώκουμε δηλαδή να προσεγγίσουμε όσο το δυνατόν περισσότερο τον στόχο, σε προαποφασισμένο αριθμό βημάτων t^* . Ο έλεγχος είναι του ανοιχτού βρόγχου, δηλαδή η διαθέσιμη πληροφορία κάθε χρονική στιγμή είναι το αρχικό συνολικό μέγεθος $T(0)$, η διαφορά του συνολικού μεγέθους $\Delta T(\tau+1)$, $\tau = t, \dots, t+t^*-1$ και ο πίνακας πιθανοτήτων μετάβασης $E(\mathbf{M}(\tau))$, $\tau = t, \dots, t+t^*-1$. Οι λύσεις του προβλήματος είναι τα διανύσματα $\mathbf{p}_0(\tau), \bar{\mathbf{q}}(\tau+1)$, $\tau = t, \dots, t+t^*-1$.

Για να μετράμε σε κάθε χρονικό βήμα την απόσταση της σχετικής δομής από την δομή στόχο και στη συνέχεια να την ελαχιστοποιούμε βρίσκοντας το κατάλληλο διάνυσμα πιθανοτήτων εισόδου, εισάγουμε τη συνάρτηση

$$E(CF(\tau)) = \sum_{i \in \mathbf{S}} c(\tau) \cdot W_i(\tau) \left[\left| q_{(i,0)}^* - \bar{q}_{(i,0)}(\tau+1) \right|^a \right],$$

όπου $W_1(\tau), W_2(\tau), \dots, W_k(\tau)$ είναι σύνολο μη αρνητικών βαρών επιλεγμένα έτσι ώστε να αντικατοπτρίζουν τη σημασία που αποδίδεται στο σωστό χειρισμό για κάθε μία κατάσταση, $c(\tau)$ το κόστος ανά μονάδα απόστασης $[\tau, \tau+1]$ και $a > 0$.

Για την επίλυση του προβλήματος προτείνουμε τον παρακάτω αλγόριθμο (Bartholomew, 1982).

Αλγόριθμος 1. Για $\tau = t$ έως $t+t^*-1$ επανάλαβε

Ελαχιστοποίησε τη συνάρτηση

$$E(CF(\tau)) = \left\{ \sum_{i \in \mathbf{S}} c(\tau) \cdot W_i(\tau) \left[\left| q_{(i,0)}^* - \bar{q}_{(i,0)}(\tau+1) \right|^2 \right] \right\},$$

με περιορισμούς

$$\bar{\mathbf{q}}(t+1) = \alpha(t) \bar{\mathbf{q}}(t) E(\mathbf{M}(t)) + \beta(t+1) \mathbf{p}_0(t+1),$$

$$\alpha(t) = T(t)/T(t+1), \quad \beta(t) = \Delta T(t+1)/T(t+1),$$

$$0 \leq p_{0(i,0)}(\tau+1) \leq 1, \quad \bar{q}_{(i,s)}(\tau+1) \geq 0, \quad s = 0, 1, \dots, b_i - 1, \quad i = 1, \dots, k,$$

$$\sum_{i=1}^k \sum_{s=0}^{b_i-1} \bar{q}_{(i,s)}(\tau+1)=1, \quad \sum_{i=1}^k p_{0(i,0)}(\tau+1)=1.$$

$$\text{Αποτελέσματα αλγορίθμου} \quad \left\{ \bar{\mathbf{q}}(\tau), \mathbf{p}_0(\tau) \right\}_{\tau=t+1}^{t+t^*}.$$

Ο αλγόριθμος αυτός χρησιμοποιείται επιπλέον για την εύρεση των πιθανοτήτων εισόδου στο πρόβλημα της διατηρησιμότητας, αντικαθιστώντας την δομή στόχο με την δομή που θέλουμε να διατηρηθεί.

Ενώ θεωρητικά η διατηρησιμότητα έπεται της επιτευξιμότητας, πρακτικά η μελέτη τους γίνεται με την αντίστροφη σειρά. Κι' αυτό επειδή δεν έχει νόημα η επίτευξη μίας δομής εφόσον δεν μπορεί να διατηρηθεί.

6. ΕΦΑΡΜΟΓΗ

Στην παράγραφο αυτή εικονογραφούμε τα προηγούμενα αποτελέσματα που προέκυψαν από τη μελέτη του ΜΟΗΜΣ σε στοχαστικό περιβάλλον. Θεωρούμε σύστημα με χώρο καταστάσεων $\mathbf{S}_H = \{1, 2, 3, 4\}$ και διάνυσμα μέγιστων χρόνων παραμονής $\mathbf{b} = (3, 3, 2, 2)$. Το αρχικό συνολικό μέγεθος είναι $T(0)=1010$ ενώ για τις επόμενες 8 χρονικές στιγμές αυτό δίνεται από το διάνυσμα $\mathbf{T}=(1050, 1125, 1225, 1335, 1445, 1500, 1600, 1680)$. Η αρχική πληθυσμιακή και σχετική δομή δίνονται από τα διανύσματα $\mathbf{N}_H(0)=(500, 360, 100, 50)$ και $\mathbf{q}_H(0)=(0.495, 0.356, 0.099, 0.050)$ αντίστοιχα. Οι ταυτότητες μέλη πραγματοποιούν κάθε χρονική στιγμή μεταβάσεις επιλέγοντας έναν από τους επόμενους τρεις πίνακες πιθανοτήτων μετάβασης

$$\mathbf{P}_1(t) = \begin{bmatrix} 0.70 - 1/10t & 0.20 + 1/20t & 0.08 - 1/20t & 0.02 + 1/10t \\ 0 & 0.80 - 2/15t & 0.15 + 1/15t & 0.05 + 1/15t \\ 0 & 0 & 0.90 - 1/5t & 0.10 + 1/5t \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\mathbf{P}_2(t) = \begin{bmatrix} 0.60 - 1/5t & 0.30 + 1/15t & 0.09 - 1/15t & 0.01 + 1/5t \\ 0 & 0.70 - 2/10t & 0.20 + 1/10t & 0.10 + 1/10t \\ 0 & 0 & 0.85 - 1/10t & 0.15 + 1/10t \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\mathbf{P}_3(t) = \begin{bmatrix} 0.65 - 1/15t & 0.25 + 1/15t & 0.08 - 1/15t & 0.02 + 1/15t \\ 0 & 0.90 - 2/20t & 0.08 + 1/20t & 0.02 + 1/20t \\ 0 & 0 & 0.95 - 1/15t & 0.05 + 1/15t \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\mathbf{P}(0) = \begin{bmatrix} 0.80 & 0.15 & 0.03 & 0.02 \\ 0 & 0.75 & 0.20 & 0.05 \\ 0 & 0 & 0.90 & 0.10 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Οι πίνακες πιθανοτήτων χρόνων παραμονής είναι οι ακόλουθοι

$$\mathbf{H}(1) = \begin{bmatrix} 0.35 & 0.19 & 0.30 & 0.35 \\ 0.00 & 0.40 & 0.16 & 0.20 \\ 0.00 & 0.00 & 0.55 & 0.40 \\ 0.00 & 0.00 & 0.00 & 0.50 \end{bmatrix}, \quad \mathbf{H}(2) = \begin{bmatrix} 0.30 & 0.26 & 0.45 & 0.40 \\ 0.00 & 0.20 & 0.34 & 0.55 \\ 0.00 & 0.00 & 0.45 & 0.60 \\ 0.00 & 0.00 & 0.00 & 0.50 \end{bmatrix},$$

$$\mathbf{H}(3) = \begin{bmatrix} 0.35 & 0.55 & 0.25 & 0.25 \\ 0.00 & 0.40 & 0.50 & 0.25 \\ 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix}.$$

Ο πίνακας πιθανοτήτων της συμβιβαστικής αλυσίδας είναι ο εξής

$$\mathbf{Z} = \begin{bmatrix} 0.300 & 0.200 & 0.500 \\ 0.278 & 0.389 & 0.333 \\ 0.421 & 0.368 & 0.211 \end{bmatrix}.$$

Μετασχηματίζουμε το ΜΟΗΜΣ σε ΜΟΜΣ με νέο χώρο καταστάσεων $\mathbf{S} = \{(1,0), (1,1), (1,2), (2,0), (2,1), (2,2), (3,0), (3,1), (4,0), (4,1)\}$, αρχική σχετική δομή $\mathbf{q}(0) = (0.495, 0, 0, 0.356, 0, 0, 0.099, 0, 0.050, 0)$ και πίνακες πιθανοτήτων μετάβασης $\mathbf{E}(\mathbf{M}(t))$. Αρχικά, μελετούμε την επιτευξιμότητα σε 7 χρονικά βήματα. Εξετάζουμε δηλαδή πόσο ικανοποιητικά μπορούμε να προσεγγίσουμε στόχο κατά το διάστημα $[0, 7]$, ξεκινώντας από την αρχική δομή $\mathbf{q}(0)$. Οι λύσεις που προκύπτουν παρουσιάζονται στους επόμενους πίνακες.

Πίνακας 1. Αναμενόμενη σχετική δομή για $t = 1, 2, \dots, 7$.

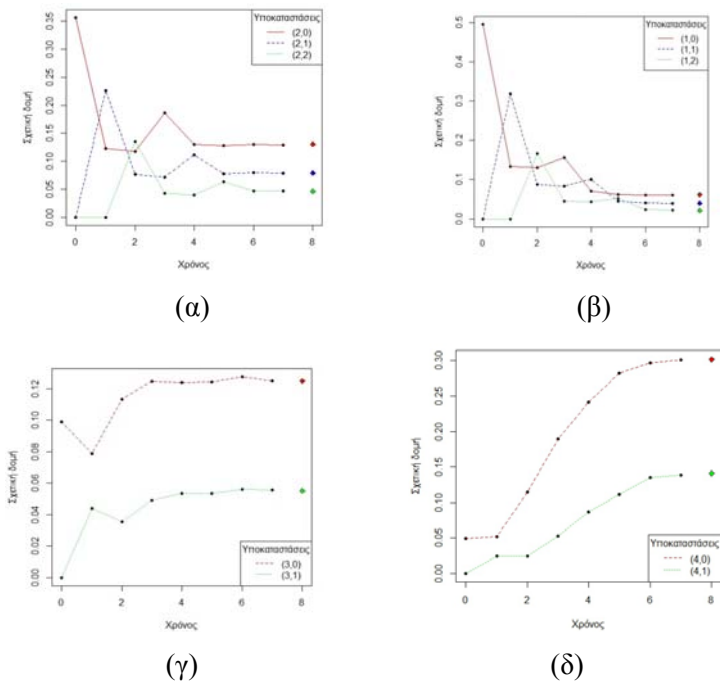
	(1,0)	(1,1)	(1,2)	(2,0)	(2,1)	(2,2)	(3,0)	(3,1)	(4,0)	(4,1)
$\bar{\mathbf{q}}(1)$	0.133	0.319	0.000	0.123	0.226	0.000	0.079	0.044	0.052	0.024
$\bar{\mathbf{q}}(2)$	0.130	0.088	0.166	0.118	0.077	0.135	0.113	0.035	0.115	0.024
$\bar{\mathbf{q}}(3)$	0.157	0.083	0.045	0.186	0.071	0.042	0.125	0.049	0.189	0.053

$\bar{q}(4)$	0.070	0.100	0.043	0.130	0.111	0.040	0.124	0.054	0.241	0.087
$\bar{q}(5)$	0.062	0.045	0.052	0.127	0.078	0.064	0.124	0.054	0.282	0.111
$\bar{q}(6)$	0.060	0.041	0.024	0.130	0.080	0.047	0.128	0.056	0.296	0.136
$\bar{q}(7)$	0.061	0.039	0.022	0.129	0.079	0.047	0.125	0.056	0.301	0.139

Πίνακας 2. Διανύσματα πιθανοτήτων εισόδου για $t = 1, 2, \dots, 7$.

	(1,0)	(1,1)	(1,2)	(2,0)	(2,1)	(2,2)	(3,0)	(3,1)	(4,0)	(4,1)
$\mathbf{p}_0(1)$	0.000	0.000	0.000	0.144	0.000	0.000	0.428	0.000	0.428	0.000
$\mathbf{p}_0(2)$	0.000	0.000	0.000	0.226	0.000	0.000	0.218	0.000	0.557	0.000
$\mathbf{p}_0(3)$	0.000	0.000	0.000	0.000	0.000	0.000	0.069	0.000	0.931	0.000
$\mathbf{p}_0(4)$	0.000	0.000	0.000	0.036	0.000	0.000	0.089	0.000	0.874	0.000
$\mathbf{p}_0(5)$	0.025	0.000	0.000	0.137	0.000	0.000	0.034	0.000	0.804	0.000
$\mathbf{p}_0(6)$	0.225	0.000	0.000	0.015	0.000	0.000	0.000	0.000	0.760	0.000
$\mathbf{p}_0(7)$	0.401	0.000	0.000	0.376	0.000	0.000	0.020	0.000	0.203	0.000

Σχήμα 1. Επιτευξιμότητα σε 7 χρονικά βήματα (α) για την κατάσταση 1 (β) για την κατάσταση 2 (γ) για την κατάσταση 3 (δ) για την κατάσταση 4



Στη συνέχεια, θεωρούμε ότι το σύστημα βρίσκεται στο στόχο $\mathbf{q}^*$ τη χρονική στιγμή $t = 7$. Προκειμένου να διατηρηθεί η δομή για ένα χρονικό βήμα χρησιμοποιούμε τον Αλγόριθμο 1. και βρίσκουμε το διάνυσμα πιθανοτήτων εισόδου $\mathbf{p}_0(8) = (0.415, 0.000, 0.000, 0.394, 0.000, 0.000, 0.060, 0.000, 0.130, 0.000)$ και την αντίστοιχη αναμενόμενη σχετική δομή $\bar{\mathbf{q}}(8) = (0.062, 0.040, 0.021, 0.130, 0.078, 0.046, 0.125, 0.054, 0.302, 0.141)$.

ABSTRACT

In the present paper we study the problem of maintainability and attainability of the target structure in non-homogeneous semi Markov systems. For this purpose we find the equation of the expected population structure at any time t , which is essential in order to study these problems. The complexity of analysis of the non-homogeneous semi Markov system's expected population structure lead us to transform the non-homogeneous semi Markov system to a non-homogeneous Markov one in relation with the basic parameters of the semi Markov system. Then, we study the problem of maintaining a desired structure and attaining the target in the system. The obtained results answer the questions concerning the semi Markov system. Last, we illustrate the theory of the paper to provide numerical results with synthesized data.

ΑΝΑΦΟΡΕΣ

- Bartholomew, D.J. (1977). *Maintaining a grade or age structure in stochastic environment*. Adv. Appl. Prob. 9, 1-17.
- Bartholomew, D.J. (1982). *Stochastic models for social processes*. 3rd Ed. Wiley, Chichester (1st Ed. 1967, 2nd Ed. 1973).
- Howard, R.A. (1971). *Dynamic Probabilistic Systems*, Vol. I, II. Wiley, Chichester.
- Παπαδοπούλου, Α.Α. (1993). *Θεωρία μη ομογενών Ημιμαρκοβιανών συστήματων*. Διδακτορική Διατριβή, Α.Π.Θ., Θεσσαλονίκη.
- Συμεωνίδου, Α.Σ. (2014). *Διατηρησιμότητα και Επιτευξιμότητα σε μη ομογενή Ημιμαρκοβιανά συστήματα*. Διπλωματική εργασία, Α.Π.Θ., Θεσσαλονίκη.
- Tsantas, N. and P. –C. G. Vassiliou (1993). *The nonhomogeneous Markov system in a stochastic environment*. J. Appl. Prob. 30 (2), 285-301
- Τσάντας, Ν.Δ. (1984). *Στοχαστικές Μέθοδοι Ελέγχου σε μη ομογενή Μαρκοβιανά συστήματα*. Διδακτορική Διατριβή, Α.Π.Θ, Θεσσαλονίκη.
- Vassiliou, P. –C. G. and Papadopoulou, A.A. (1992). *Non-homogeneous semi-Markov systems and maintainability of the state sizes*. J. Appl. Prob. 29, 519-534.

Η ΣΥΜΠΕΡΙΦΟΡΑ ΤΩΝ ΕΛΕΓΧΩΝ ΧΩΡΙΚΗΣ ΕΞΑΡΤΗΣΗΣ ΣΕ ΜΕΓΑΛΑ ΔΕΙΓΜΑΤΑ

Απόστολος Τσιμπάνος¹, Χρήστος Αγιακλόγλου² και Κλέων Τσίμπος¹

¹Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς

²Τμήμα Οικονομικής Επιστήμης, Πανεπιστήμιο Πειραιώς

{atsiban, agiaklis, cleon }@unipi.gr

ΠΕΡΙΛΗΨΗ

Στη χωρική οικονομετρία οι έλεγχοι χωρικής εξάρτησης χρησιμοποιούνται για τον εντοπισμό της αιτίας εμφάνισης του προβλήματος της χωρικής αυτοσυσχέτισης στα σφάλματα ενός γραμμικού υποδείγματος και την επίλυσή του μέσω της επιλογής του κατάλληλου χωρικού οικονομετρικού υποδείγματος. Στην εργασία αυτή, με τη βοήθεια προσομοίωσης, επιβεβαιώνονται παλιότερα συμπεράσματα για την αποτελεσματικότητα των ελέγχων αυτών σε μικρά και μέτρια δείγματα και προστίθενται καινούριες πληροφορίες για τη συμπεριφορά τους σε μεγάλα δείγματα. Επιπρόσθετα, διαπιστώνεται ότι η σύγκριση των τιμών των δύο ανθεκτικών στατιστικών προσφέρει σημαντική βοήθεια στην επιλογή του σωστού υποδείγματος στις περιπτώσεις που η εφαρμογή των ελέγχων οδηγεί σε ασάφειες.

Λέξεις Κλειδιά: Χωρική Αυτοσυσχέτιση, LM Έλεγχοι Χωρικής Εξάρτησης, Υπόδειγμα Χωρικής Υστέρησης, Υπόδειγμα Χωρικού Σφάλματος

1. ΕΙΣΑΓΩΓΗ

Η εκτίμηση υποδειγμάτων γραμμικής παλινδρόμησης με χωρικά δεδομένα, δηλαδή με δεδομένα που προέρχονται από διαφορετικές γεωγραφικές περιοχές, πολλές φορές υποκρύπτει τον κίνδυνο εμφάνισης του προβλήματος της χωρικής αυτοσυσχέτισης στα σφάλματα. Η χωρική αυτοσυσχέτιση (*spatial autocorrelation*) είναι αποτέλεσμα της χωρικής εξάρτησης (*spatial dependence*), η οποία μαζί με τη χωρική ετερογένεια (*spatial heterogeneity*) αποτελούν τις χωρικές επιδράσεις (*spatial effects*) που οφείλονται στο είδος των δεδομένων (Anselin, 1988). Η ύπαρξη χωρικής αυτοσυσχέτισης, δηλαδή συσχέτισης των παρατηρήσεων σύμφωνα με τις γεωγραφικές τους θέσεις, καθιστά τα αποτελέσματα από την εκτίμηση ενός οικονομετρικού υποδείγματος αναξιόπιστα. Η αντιμετώπιση αυτού του προβλήματος γίνεται με την εκτίμηση χωρικών οικονομετρικών υποδειγμάτων που λαμβάνουν υπ' όψιν την παρουσία των χωρικών επιδράσεων.

Εάν θεωρηθεί το απλό οικονομετρικό υπόδειγμα $y = X\beta + \varepsilon$ όπου y είναι το $(n \times 1)$ διάνυσμα των παρατηρήσεων της εξαρτημένης μεταβλητής, X η $(n \times (k+1))$ μήτρα των παρατηρήσεων των k ανεξάρτητων μεταβλητών έχοντας στην πρώτη στήλη μονάδες για την εκτίμηση του σταθερού όρου, β το $((k+1) \times 1)$ διάνυσμα των συντελεστών του υποδείγματος και $\varepsilon \sim N(0, \sigma^2 I)$ το $(n \times 1)$ διάνυσμα των τυχαίων σφαλμάτων, η ενσωμάτωση των χωρικών επιδράσεων στο υπόδειγμα γίνεται με τον ορισμό της $(n \times n)$ μήτρα χωρικών σταθμίσεων W που καθορίζει την αλληλεπίδραση μεταξύ των n γειτονικών περιοχών. Τα δύο πιο συχνά χρησιμοποιούμενα χωρικά οικονομετρικά υποδείγματα είναι το υπόδειγμα χωρικής υστέρησης (*Spatial Lag Model – SLM*) που ορίζεται ως $y = \rho Wy + X\beta + \varepsilon$ και το υπόδειγμα χωρικού σφάλματος (*Spatial Error Model – SEM*) που ορίζεται ως $y = X\beta + \varepsilon$ με $\varepsilon = \lambda W\varepsilon + u$ όπου u είναι ένα $(n \times 1)$ διάνυσμα σφαλμάτων, Wy και $W\varepsilon$ είναι οι μήτρες χωρικών υστερήσεων που ενσωματώνουν την χωρική εξάρτηση στα δυο υποδείγματα και ουσιαστικά αποτελούνται από τους σταθμισμένους μέσους όρους των τιμών των μεταβλητών στις γειτονικές περιοχές ενώ ρ και λ είναι οι συντελεστές χωρικών υστερήσεων της εξαρτημένης μεταβλητής και των σφαλμάτων αντίστοιχα. Τα υποδείγματα αυτά είναι παρόμοια με τα αντίστοιχα που εφαρμόζονται στην ανάλυση χρονοσειρών με τη διαφορά ότι εδώ η εξάρτηση είναι προς πολλές κατευθύνσεις. Στις εμπειρικές εφαρμογές, συνήθως ο ερευνητής εφαρμόζει έλεγχο χωρικής αυτοσυσχέτισης στα σφάλματα του απλού οικονομετρικού υποδείγματος με τη στατιστική I του Moran όπως έχει προταθεί από τους Cliff και Ord (1972) για την περίπτωση καταλοίπων παλινδρόμησης και εφόσον έχει ενδείξεις εμφάνισής της χρησιμοποιεί τους ελέγχους χωρικής εξάρτησης για να επιλέξει το κατάλληλο χωρικό οικονομετρικό υπόδειγμα.

Στην εργασία αυτή, με τη βοήθεια προσομοίωσης, διερευνάται η συμπεριφορά των ελέγχων χωρικής εξάρτησης με μεγάλα δείγματα. Τα αποτελέσματα δείχνουν ότι αρκετές φορές η εφαρμογή των ελέγχων μπορεί να οδηγήσει σε ασάφειες σχετικά με το χωρικό οικονομετρικό υπόδειγμα που πρέπει να επιλεγεί αλλά η σύγκριση των τιμών των ανθεκτικών στατιστικών οδηγεί στο σωστό υπόδειγμα.

2. ΟΙ ΕΛΕΓΧΟΙ ΧΩΡΙΚΗΣ ΕΞΑΡΤΗΣΗΣ

Οι έλεγχοι χωρικής εξάρτησης (*Spatial Dependence Tests*) ανήκουν στην κατηγορία των ελέγχων σκορ ή πολλαπλασιαστή Lagrange (*Score test* ή *Lagrange Multiplier test*) και εφαρμόζονται για την επιλογή μεταξύ του απλού οικονομετρικού υποδείγματος και ενός χωρικού υποδείγματος από το υπόδειγμα χωρικής υστέρησης ή το υπόδειγμα χωρικού σφάλματος. Το βασικό τους πλεονέκτημα είναι ότι ο υπολογισμός των στατιστικών τους βασίζεται αποκλειστικά στα κατάλοιπα που προκύπτουν από την εκτίμηση του απλού οικονομετρικού υποδείγματος, με τη μέθοδο των ελαχίστων τετραγώνων, χωρίς να απαιτείται η εκτίμηση του αντίστοιχου χωρικού οικονομετρικού υποδείγματος.

Η περιγραφή των ελέγχων βασίζεται στο γενικό χωρικό οικονομετρικό υπόδειγμα, το οποίο παρουσιάζεται λεπτομερώς στον Anselin (1988), και που αποτελείται από το ταυτόχρονο σύστημα δύο εξισώσεων $y = \rho W y + X\beta + \varepsilon$ και $\varepsilon = \lambda W \varepsilon + u$ με $u \sim N(0, \sigma^2 I)$ όπου οι μήτρες και οι συντελεστές έχουν οριστεί στην προηγούμενη ενότητα. Εάν στο υπόδειγμα αυτό τεθούν $\rho = \lambda = 0$ προκύπτει το απλό οικονομετρικό υπόδειγμα, εάν τεθεί $\lambda = 0$ προκύπτει το υπόδειγμα χωρικής υστέρησης και εάν τεθεί $\rho = 0$ το υπόδειγμα χωρικού σφάλματος. Ο πρώτος έλεγχος αυτής της κατηγορίας εξετάζει την αναγκαιότητα εκτίμησης του υποδείγματος χωρικού σφάλματος σε σχέση με το απλό υπόδειγμα και παρουσιάστηκε αρχικά από τον Burridge (1980). Δηλαδή, υποθέτοντας ότι $\rho = 0$ η μηδενική και η εναλλακτική υπόθεση ορίζονται αντίστοιχα ως $H_0: \lambda = 0$ και $H_1: \lambda \neq 0$ και υπολογίζεται η στατιστική:

$$LM-ERR = [\hat{\varepsilon}' W \hat{\varepsilon} / S^2]^2 / T$$

όπου με $\hat{\varepsilon}$ συμβολίζεται το διάνυσμα των καταλοίπων από την εκτίμηση, $S^2 = \hat{\varepsilon}' \hat{\varepsilon} / n$ είναι η εκτιμηθείσα διακύμανση της παλινδρόμησης και $T = tr[W^2 + W'W]$. Ο έλεγχος για την παρουσία στο υπόδειγμα χωρικής υστέρησης στην εξαρτημένη μεταβλητή προτάθηκε από τον Anselin (1988b) και υποθέτοντας ότι $\lambda = 0$ έχει ως μηδενική και εναλλακτική υπόθεση τις $H_0: \rho = 0$ και $H_1: \rho \neq 0$ και πραγματοποιείται με βάση τη στατιστική:

$$LM-LAG = [\hat{\varepsilon}' W y / S^2]^2 / \left(T / \left[(W X \hat{\beta})' M (W X \hat{\beta}) / S^2 \right] \right)$$

όπου $\hat{\beta}$ είναι το διάνυσμα με τις εκτιμηθείσες τιμές των παραμέτρων του απλού υποδείγματος και $M = I - X(X'X)^{-1}X'$.

Οι στατιστικές LM-ERR και LM-LAG δεν είναι έγκυρες όταν αντίστοιχα $\rho \neq 0$ και $\lambda \neq 0$. Οι Anselin, Bera, Florax, και Yoon (1996) ανέπτυξαν ανθεκτικές (*robust*) μορφές των ελέγχων που διορθώνουν για την ύπαρξη τοπικής λανθασμένης εξειδίκευσης (*local misspecification*) και μπορούν να χρησιμοποιηθούν για την επιλογή του σωστού χωρικού οικονομετρικού υποδείγματος όταν υπάρχει ασάφεια από την εφαρμογή των απλών ελέγχων, σημειώνοντας ότι ως τοπική λανθασμένη εξειδίκευση εννοείται ότι οι τιμές των λ και ρ δεν είναι μηδέν αλλά πολύ κοντά στο μηδέν. Ωστόσο, πρέπει να αναφερθεί πως και οι ανθεκτικές μορφές των ελέγχων αντιμετωπίζουν προβλήματα όταν οι τιμές των λ και ρ είναι πολύ μακριά από το μηδέν. Ο ανθεκτικός έλεγχος της υπόθεσης $H_0: \lambda = 0$ που λαμβάνει υπ' όψιν την παρουσία χωρικής υστέρησης της εξαρτημένης μεταβλητής πραγματοποιείται με τη στατιστική:

$$LM-EL = \left[\frac{\hat{\varepsilon}' W \hat{\varepsilon}}{S^2} - T \frac{S^2}{S^2 T + (W X \hat{\beta})' M (W X \hat{\beta})} \frac{\hat{\varepsilon}' W y}{S^2} \right] / \left(T - T^2 \left[\frac{S^2}{S^2 T + (W X \hat{\beta})' M (W X \hat{\beta})} \right] \right)$$

και για τον έλεγχο της υπόθεσης $H_0: \rho = 0$ που λαμβάνει υπ' όψιν την παρουσία χωρικής αυτοσυσχέτισης στα σφάλματα με τη στατιστική:

$$LM-LE = \left[\frac{\hat{\mathbf{e}}' \mathbf{W} \mathbf{y}}{S^2} - \frac{\hat{\mathbf{e}}' \mathbf{W} \hat{\mathbf{e}}}{S^2} \right]^2 / \left(\frac{1}{S^2} \left[(\mathbf{W} \mathbf{X} \hat{\boldsymbol{\beta}})' \mathbf{M} (\mathbf{W} \mathbf{X} \hat{\boldsymbol{\beta}}) + S^2 T \right] - T \right)$$

Οι στατιστικές αυτές υπό τις μηδενικές τους υποθέσεις ακολουθούν ασυμπτωτικά την κατανομή χ^2 με ένα βαθμό ελευθερίας.

Οι έλεγχοι χωρικής εξάρτησης μερικές φορές οδηγούν σε αντιφατικά συμπεράσματα ως προς την επιλογή του σωστού χωρικού οικονομετρικού υποδείγματος. Σε αρκετές εμπειρικές εφαρμογές, όταν υπάρχει ασάφεια από τα αποτελέσματα των ελέγχων, εφαρμόζεται η ακόλουθη στρατηγική εντοπισμού του σωστού υποδείγματος. Αρχικά διεξάγονται οι έλεγχοι με τις στατιστικές LM-ERR και LM-LAG. Εάν κανένας από τους δύο ελέγχους δεν απορρίπτει τη μηδενική του υπόθεση τότε πρέπει να εξεταστούν τα κατάλοιπα του απλού οικονομετρικού υποδείγματος για την πιθανότητα ύπαρξης άλλων προβλημάτων που μπορεί να οδήγησαν στην εσφαλμένη απόρριψη του ελέγχου χωρικής αυτοσυσχέτισης με τη στατιστική I του Moran όπως η ετεροσκεδαστικότητα και η έλλειψη κανονικότητας. Εάν στον έναν από τους δύο ελέγχους απορρίπτεται η μηδενική υπόθεση, τότε επιλέγεται και εκτιμάται το αντίστοιχο χωρικό υπόδειγμα. Στην περίπτωση που απορρίπτεται η μηδενική υπόθεση και στους δύο απλούς ελέγχους διεξάγονται οι έλεγχοι με τις στατιστικές LM-EL και LM-LE και επιλέγεται το χωρικό υπόδειγμα του ελέγχου που απορρίπτεται η μηδενική του υπόθεση. Εάν και στους δύο ανθεκτικούς ελέγχους απορρίπτεται η μηδενική υπόθεση τότε επιλέγεται και εκτιμάται το υπόδειγμα που υποδεικνύει η ανθεκτική στατιστική με τη μεγαλύτερη τιμή.

3. Η ΙΣΧΥΣ ΤΩΝ ΕΛΕΓΧΩΝ ΧΩΡΙΚΗΣ ΕΞΑΡΤΗΣΗΣ

Η αποδοχή των ελέγχων χωρικής εξάρτησης από τους ερευνητές και η προτεινόμενη διαδικασία επιλογής του χωρικού οικονομετρικού υποδείγματος βασίζονται σε μεγάλο βαθμό στα αποτελέσματα των προσομοιώσεων που περιγράφονται στις εργασίες των Anselin και Rey (1991), Anselin και Florax (1995), καθώς και στους Anselin, Bera, Florax και Yoon (1996). Στην ενότητα αυτή εξετάζεται με τη βοήθεια προσομοίωσης η συμπεριφορά και η επίδοση των ελέγχων χωρικής εξάρτησης για μεγαλύτερα μεγέθη δείγματος και μεγαλύτερο αριθμό επαναλήψεων σε σχέση με τις προηγούμενες εργασίες. Στόχος είναι να διερευνηθεί περαιτέρω η στατιστική ισχύς των ελέγχων αυτών και η δυνατότητά τους να υποδεικνύουν το σωστό χωρικό υπόδειγμα.

Για λόγους σύγκρισης με τις προαναφερθείσες εργασίες διατηρείται ο ίδιος σχεδιασμός στην προσομοίωση που θεωρεί υπό τη μηδενική υπόθεση της απουσίας χωρικής αυτοσυσχέτισης το απλό οικονομετρικό υπόδειγμα $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, όπου $\mathbf{X}$ είναι μία $(n \times 3)$ μήτρα που αποτελείται από δύο ανεξάρτητες μεταβλητές και μία στήλη με μονάδες για την εκτίμηση του σταθερού όρου και $\boldsymbol{\beta}$ το (3×1) διάνυσμα των συντελεστών που θέτονται και οι τρεις ίσοι με τη μονάδα. Οι δύο ανεξάρτητες μεταβλητές προέρχονται από ομοιόμορφη $U(0, 10)$ κατανομή και παρήχθησαν μόνο

μία φορά παραμένοντας σταθερές σε όλες τις επαναλήψεις. Το διάνυσμα των τυχαίων σφαλμάτων ε προέρχεται από $N(\mathbf{0}, \mathbf{I})$ κατανομή και προστίθεται στο διάνυσμα $\mathbf{X}\beta$ για να παραχθεί το διάνυσμα της εξαρτημένης μεταβλητής $\mathbf{y}$.

Η χωρική εξάρτηση ορίστηκε με μία $(n \times n)$ τυποποιημένη κατά γραμμή μήτρα χωρικών σταθμίσεων $\mathbf{W}$ που κατασκευάστηκε με βάση τα κριτήρια του πύργου (*rook contiguity*) και τη βασίλισσας (*queen contiguity*) για ένα ομαλό τετραγωνικό πλέγμα διαστάσεων 7x7, 10x10, 20x20 και 30x30 παρέχοντας δείγματα 49, 100, 400 και 900 παρατηρήσεων αντίστοιχα. Τα χωρικά οικονομετρικά υποδείγματα της εναλλακτικής υπόθεσης είναι το υπόδειγμα χωρικής υστέρησης και το υπόδειγμα χωρικού σφάλματος. Το υπόδειγμα χωρικής υστέρησης παράγεται πολλαπλασιάζοντας τη δεξιά πλευρά του απλού οικονομετρικού υποδείγματος με το χωρικό πολλαπλασιαστή οπότε προκύπτει $\mathbf{y} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\beta + (\mathbf{I} - \rho\mathbf{W})^{-1}\varepsilon$. Για το υπόδειγμα χωρικού σφάλματος θεωρείται ότι τα σφάλματα ακολουθούν μια στάσιμη χωρική αυτοπαλίνδρομη διαδικασία πρώτης τάξης, SAR(1), και το διάνυσμα της εξαρτημένης μεταβλητής παράγεται ως $\mathbf{y} = \mathbf{X}\beta + (\mathbf{I} - \lambda\mathbf{W})^{-1}\mathbf{u}$ όπου το $(n \times 1)$ διάνυσμα $\mathbf{u}$ προέρχεται από $N(0, \mathbf{I})$ κατανομή. Οι συντελεστές ρ και λ των χωρικών υποδειγμάτων θεωρήθηκε ότι λαμβάνουν τις τιμές 0,0, 0,2, 0,5, 0,8 και 0,9.

Η διαδικασία της προσομοίωσης πραγματοποιήθηκε στο στατιστικό πρόγραμμα R με τη βοήθεια του πακέτου SPDEP. Εκτελέστηκαν 10.000 επαναλήψεις και για κάθε χωρικό υπόδειγμα εκτιμήθηκε με τη μέθοδο των ελαχίστων τετραγώνων το απλό οικονομετρικό υπόδειγμα και κατεγράφησαν τα ποσοστά απόρριψης των LM ελέγχων χωρικής εξάρτησης που χρησιμοποιούνται για τον εντοπισμό της αιτίας της αυτοσυσχέτισης και την επιλογή των χωρικών οικονομετρικών υποδειγμάτων θεωρώντας επίπεδο σημαντικότητας 5%.

Τα αποτελέσματα της προσομοίωσης παρουσιάζονται στον Πίνακα 1 για το υπόδειγμα χωρικής υστέρησης και στον Πίνακα 2 για το υπόδειγμα χωρικού σφάλματος. Όταν $\rho=0$ και $\lambda=0$ τότε δεν εμφανίζεται καμία μορφή χωρικής εξάρτησης στο υπόδειγμα, δηλαδή ισχύει το απλό οικονομετρικό υπόδειγμα. Κατά συνέπεια, τα αντίστοιχα ποσοστά απόρριψης αποτελούν τα μεγέθη των ελέγχων. Τα ποσοστά αυτά είναι όλα κοντά στο επίπεδο σημαντικότητας του 5% κάτι που φανερώνει την απουσία αυτοσυσχέτισης και υποδεικνύει την επιλογή του απλού υποδείγματος. Εξαίρεση υπάρχει μόνο στα μικρά δείγματα των 49 και 100 παρατηρήσεων στα οποία τα ποσοστά απόρριψης των στατιστικών LM-LAG και LM-LE είναι λίγο μεγαλύτερα από το 5%, ενώ των στατιστικών LM-ERR και LM-EL είναι λίγο μικρότερα. Επομένως, η συμπεριφορά των ελέγχων, όπως εμφανίζεται με το μέγεθός τους όταν ισχύει το απλό οικονομετρικό υπόδειγμα και απουσιάζει το πρόβλημα της χωρικής αυτοσυσχέτισης, κρίνεται ότι είναι ικανοποιητική.

Όπως διαπιστώνεται από τον Πίνακα 1, οι συμπεριφορές της στατιστικής LM-LAG για την παρουσία χωρικής υστέρησης στην εξαρτημένη μεταβλητή που έχει αγνοηθεί από τον ερευνητή εκτιμώντας το απλό οικονομετρικό υπόδειγμα και της ανθεκτικής της εκδοχής LM-LE κρίνονται εξαιρετικές, καθώς απορρίπτουν τη μηδενική υπόθεση της ύπαρξης χωρικής υστέρησης στην εξαρτημένη μεταβλητή σε

ποσοστό σχεδόν 100% ακόμη και στα μικρά δείγματα και για τιμές του ρ ίσες με 0,2. Ταυτόχρονα, διαπιστώνεται ότι η στατιστική LM-ERR για τον έλεγχο του υποδείγματος χωρικού σφάλματος έχει σημαντική στατιστική ισχύ, ιδιότητα που δεν θα έπρεπε να συμβαίνει, με ποσοστά απόρριψης που ειδικά σε μεγάλα δείγματα προσεγγίζουν το 100% ακόμη και για μικρά ρ , προκαλώντας στον ερευνητή σύγχυση ως προς τον εντοπισμό της αιτίας της αυτοσυσχέτισης. Αξιοσημείωτο είναι το γεγονός ότι η ανθεκτική της εκδοχή, δηλαδή η στατιστική LM-EL, παρουσιάζει ιδιάζουσα συμπεριφορά και ειδικότερα, όταν η χωρική εξάρτηση ορίζεται με το κριτήριο του πύργου το ποσοστό απόρριψης της στατιστικής LM-EL για μικρές τιμές του ρ είναι κοντά στο 5% αλλά για μεγάλες τιμές μειώνεται και τελικά μηδενίζεται. Το πρόβλημα όμως αυτό στην πράξη δεν προκαλεί δυσκολία στην επιλογή του υποδείγματος χωρικής υστέρησης.

Πίνακας 1. Ποσοστά απόρριψης ελέγχων χωρικής εξάρτησης για όταν ισχύει το Υπόδειγμα Χωρικής Υστέρησης σε επίπεδο σημαντικότητας 5%

N	Σταθμίσεις	Πύργος					Βασίλισσα				
	ρ	0,0	0,2	0,5	0,8	0,9	0,0	0,2	0,5	0,8	0,9
49	LM-LAG	5,91	80,33	100	100	100	5,56	57,91	100	100	100
	LM-LE	6,28	78,77	100	100	100	5,86	55,39	99,98	100	100
	LM-ERR	4,78	6,50	23,71	92,73	99,8	3,86	13,45	85,69	100	100
	LM-EL	5,13	8,74	16,86	5,18	0,8	4,36	5,09	13,30	14,09	2,94
100	LM-LAG	4,95	98,71	100	100	100	5,12	85,98	100	100	100
	LM-LE	5,39	97,76	100	100	100	5,37	82,04	100	100	100
	LM-ERR	4,59	21,39	92,86	100	100	4,24	31,67	99,81	100	100
	LM-EL	4,81	5,23	3,70	0,28	0,03	4,51	6,46	49,43	80,28	56,8
400	LM-LAG	4,99	100	100	100	100	4,82	100	100	100	100
	LM-LE	5,01	100	100	100	100	4,91	100	100	100	100
	LM-ERR	4,75	77,50	100	100	100	4,97	90,28	100	100	100
	LM-EL	5,14	5,29	1,65	0,03	0,00	5,03	23,92	99,98	100	100
900	LM-LAG	4,95	100	100	100	100	4,78	100	100	100	100
	LM-LE	4,94	100	100	100	100	4,97	100	100	100	100
	LM-ERR	4,89	98,95	100	100	100	5,01	99,7	100	100	100
	LM-EL	5,04	5,30	8,75	2,09	0,00	5,02	48,68	100	100	100

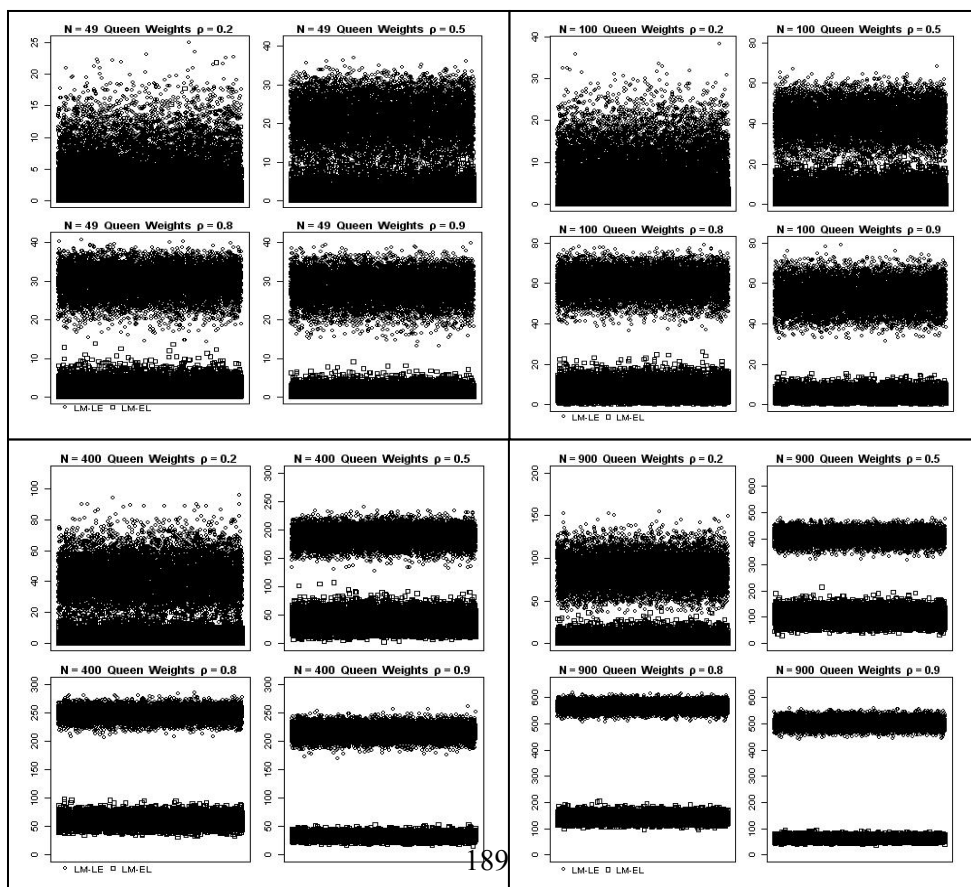
Πίνακας 2. Ποσοστά απόρριψης ελέγχων χωρικής εξάρτησης για όταν ισχύει το Υπόδειγμα Χωρικού Σφάλματος σε επίπεδο σημαντικότητας 5%

N	Σταθμίσεις	Πύργος					Βασίλισσα				
	λ	0,0	0,2	0,5	0,8	0,9	0,0	0,2	0,5	0,8	0,9
49	LM-LAG	5,91	6,22	16,68	64,45	86,82	5,56	7,65	20,29	60,83	80,17
	LM-LE	6,28	5,67	5,46	5,44	4,60	5,86	7,58	11,47	17,69	19,78
	LM-ERR	4,78	10,42	61,28	98,23	99,75	3,86	9,43	48,82	93,35	98,39
	LM-EL	5,13	9,49	54,77	95,78	99,07	4,36	9,09	44,51	90,09	97,09
100	LM-LAG	4,95	7,44	32,22	89,12	98,72	5,12	8,73	32,63	85,84	97,42
	LM-LE	5,39	5,35	6,61	10,40	11,84	5,37	7,29	12,56	20,98	24,43
	LM-ERR	4,59	25,27	94,84	100	100	4,24	18,97	83,60	99,93	100
	LM-EL	4,81	23,07	91,9	100	100	4,51	17,58	78,93	99,74	99,99
40	LM-LAG	4,99	15,18	76,82	99,98	100	4,82	12,44	65,02	99,87	100
	LM-LE	5,01	5,24	7,67	14,63	19,19	4,91	6,80	12,06	25,80	34,54

	LM-ERR	4,75	81,18	100	100	100	4,97	58,33	100	100	100
	LM-EL	5,14	76,61	100	100	100	5,03	54,57	99,98	100	100
900	LM-LAG	4,95	31,29	98,36	100	100	4,78	22,52	92,91	100	100
	LM-LE	4,94	5,17	7,58	15,55	21,51	4,97	7,15	13,64	28,15	36,32
	LM-ERR	4,89	99,01	100	100	100	5,01	89,52	100	100	100
	LM-EL	5,04	98,19	100	100	100	5,02	86,14	100	100	100

Η ιδιότροπη συμπεριφορά της στατιστικής LM-EL με το κριτήριο του πύργου είχε αναφερθεί και στις εργασίες των Anselin και Florax (1995) καθώς και Anselin, Bera, Florax, και Yoon (1996). Αντίθετα, σύγχυση προκαλείται από τη στατιστική LM-EL όταν χρησιμοποιείται το κριτήριο της βασίλισσας σε μεγάλα δείγματα και για τιμές του ρ μεγαλύτερες του 0,5, καθώς το ποσοστό απόρριψης είναι 100% κάτι που επιφέρει ασάφεια για την επιλογή μεταξύ του υποδείγματος χωρικής υστέρησης και του υποδείγματος χωρικού σφάλματος. Το αποτέλεσμα αυτό δεν είχε εντοπιστεί στις δύο προηγούμενες εργασίες καθώς το μεγαλύτερο δείγμα στις προσομοιώσεις τους αφορούσε 127 παρατηρήσεις και επιπρόσθετα μόνο δείγμα 81 παρατηρήσεων είχε οριστεί με βάση ένα ομαλό τετραγωνικό πλέγμα. Στις περιπτώσεις αυτές έχει ενδιαφέρον να μελετηθούν οι τιμές των στατιστικών LM-LE και LM-EL που προέκυψαν από την προσομοίωση. Στο Διάγραμμα 1 απεικονίζονται οι αριθμητικές τιμές των δύο ανθεκτικών στατιστικών για τα τέσσερα διαφορετικά δείγματα και για ρ ίσα με 0,2, 0,5, 0,8 και 0,9.

Διάγραμμα 1 Προσομοιωμένες τιμές των στατιστικών LM-LE και LM-EL για το υπόδειγμα χωρικής υστέρησης με το κριτήριο της βασίλισσας



Η παρατήρηση του Διαγράμματος 1 αποκαλύπτει ότι για τιμές του ρ μεγαλύτερες ή ίσες του 0,5 οι τιμές της στατιστικής LM-LE είναι πάντα μεγαλύτερες από τις τιμές της στατιστικής LM-EL. Επιπλέον, καθώς αυξάνεται το μέγεθος του δείγματος η απόσταση μεταξύ των τιμών των δύο στατιστικών αυξάνεται σημαντικά. Επομένως, εάν η τιμή της στατιστικής LM-LE είναι μεγαλύτερη από τη τιμή της LM-EL, στην περίπτωση που και οι δύο απλές στατιστικές απορρίπτουν τις μηδενικές τους υποθέσεις και συμβαίνει το ίδιο και με τις ανθεκτικές τους μορφές, τότε υπάρχουν σαφείς ενδείξεις για την επιλογή του υποδείγματος χωρικής υστέρησης που είναι το σωστό υπόδειγμα.

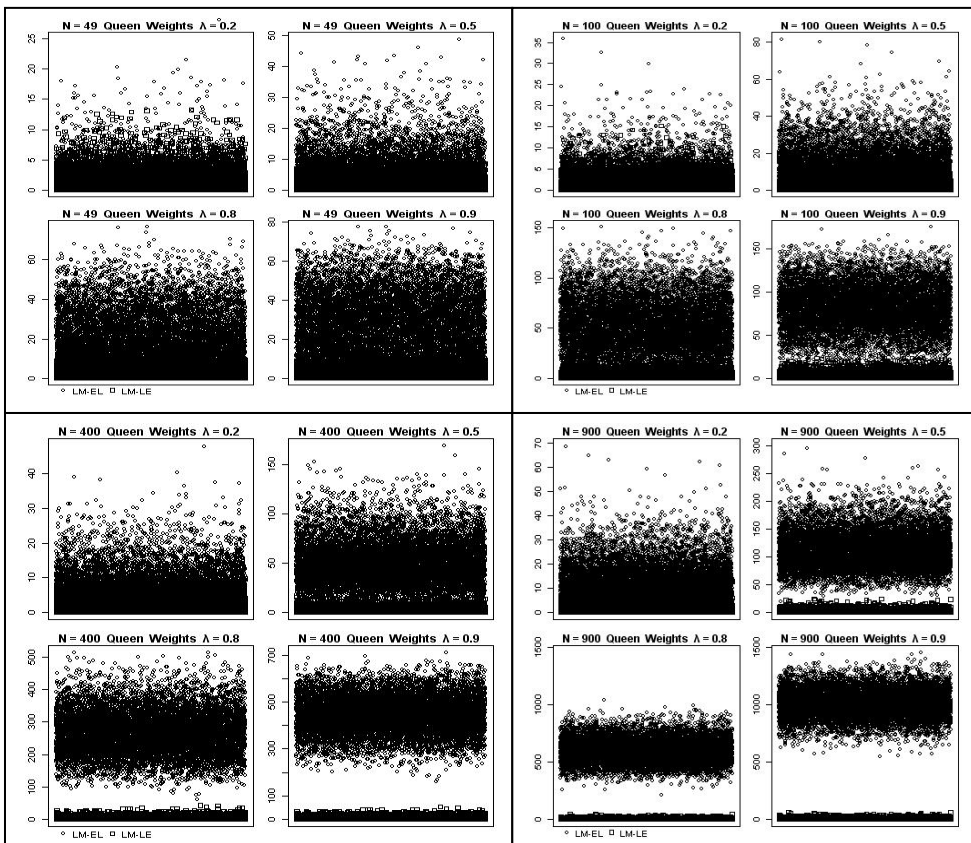
Ο Πίνακας 2 περιέχει τα αποτελέσματα της προσομοίωσης για όταν το πρόβλημα της χωρικής αυτοσυσχέτισης οφείλεται στην ύπαρξη του υποδείγματος χωρικού σφάλματος. Από τον πίνακα αυτό φαίνεται ότι τα ποσοστά απόρριψης της στατιστικής LM-ERR αυξάνονται καθώς αυξάνεται η τιμή του λ ως συνέπεια της ύπαρξης χωρικής αυτοσυσχέτισης. Ωστόσο, στα δείγματα των 49 και 100 παρατηρήσεων και για μικρές τιμές του λ διαπιστώνεται ότι η στατιστική δεν έχει υψηλή ισχύ στον εντοπισμό του προβλήματος ενώ στα μεγάλα δείγματα η ισχύς της είναι ικανοποιητική. Γενικά, συμπεραίνεται ότι η ισχύς της στατιστικής LM-ERR είναι χαμηλή στα μικρά δείγματα σε αντίθεση με τη στατιστική LM-LAG που, όπως παρουσιάστηκε στον Πίνακα 1, έχει ικανοποιητική συμπεριφορά και στα μικρά δείγματα.

Η στατιστική LM-LAG παρουσιάζει και αυτή αξιόλογα ποσοστά απόρριψης ιδιαίτερα για υψηλές τιμές του λ σε μεγάλα δείγματα προκαλώντας σύγχυση ως προς τον εντοπισμό της σωστής αιτίας της χωρικής αυτοσυσχέτισης κάτι, που όπως προηγουμένως έχει αναφερθεί, ισχύει και με τη στατιστική LM-ERR όταν η αιτία της αυτοσυσχέτισης είναι η χωρική υστέρηση της εξαρτημένης μεταβλητής. Στην περίπτωση αυτή ο ερευνητής χρησιμοποιεί τα αποτελέσματα των ανθεκτικών στατιστικών. Ειδικότερα, από τον Πίνακα 2 φαίνεται ότι η ανθεκτική στατιστική LM-EL έχει ποσοστά απόρριψης που προσεγγίζουν το 100% ακόμη και σε μετρίου μεγέθους δείγματα και για τιμές του λ μεγαλύτερες ή ίσες του 0,5. Αντίθετα, τα ποσοστά απόρριψης της ανθεκτικής στατιστικής LM-LE είναι χαμηλά, αν και ξεπερνούν το επίπεδο σημαντικότητας του 5%, στα μικρά δείγματα και μέτρια στα δείγματα των 400 και 900 παρατηρήσεων. Λίγο πιο αυξημένα ποσοστά απόρριψης εντοπίζονται όταν έχει χρησιμοποιηθεί το κριτήριο της βασιλίσσας.

Όπως και στο υπόδειγμα χωρικής υστέρησης, έτσι και στην περίπτωση του υποδείγματος χωρικού σφάλματος όταν τα αποτελέσματα των ελέγχων των δύο ανθεκτικών στατιστικών είναι αντικρουόμενα υπάρχει μεγάλη πιθανότητα το χωρικό υπόδειγμα που υποδεικνύει η ανθεκτική στατιστική με τη μεγαλύτερη αριθμητική τιμή να είναι το σωστό. Αυτό παρουσιάζεται για τα αποτελέσματα της προσομοίωσης που προέκυψαν από την εφαρμογή της μήτρας χωρικών σταθμίσεων που

κατασκευάστηκε σύμφωνα με το κριτήριο της βασίλισσας καθώς εκεί εμφανίζεται πιο σοβαρό το πρόβλημα. Οι τιμές των δύο ανθεκτικών στατιστικών για όλα τα δείγματα και τιμές του λ ίσες με 0,2, 0,5, 0,8 και 0,9 απεικονίζονται στο Διάγραμμα 2. Από το Διάγραμμα αυτό διαπιστώνεται ότι οι τιμές που προέκυψαν από την προσομοίωση για τη στατιστική LM-EL είναι μεγαλύτερες από τις τιμές της στατιστικής LM-LE. Το χαρακτηριστικό αυτό φαίνεται πιο έντονα στα δείγματα των 400 και 900 παρατηρήσεων και για τιμές των λ ίσες με 0,8 και 0,9. Πρέπει όμως να αναφερθεί ότι, παρόλο που δεν διακρίνεται καθαρά στο διάγραμμα, η ιδιότητα αυτή ισχύει για τις περισσότερες περιπτώσεις της προσομοίωσης και στα μικρά δείγματα και για μικρές τιμές του λ . Εμφανίζονται μόνο μερικές μεμονωμένες εξαιρέσεις. Επομένως, και στο υπόδειγμα χωρικού σφάλματος, όταν η τιμή της στατιστικής LM-EL είναι μεγαλύτερη από την τιμή της στατιστικής LM-LE υπάρχουν ισχυρές ενδείξεις που οδηγούν στην εκτίμηση του σωστού χωρικού υποδείγματος.

***Διάγραμμα 2** Προσομοιωμένες τιμές των στατιστικών LM-LE και LM-EL για το υπόδειγμα χωρικού σφάλματος με το κριτήριο της βασίλισσας*



4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Από την ανάλυση των αποτελεσμάτων της προσομοίωσης γίνεται φανερό ότι ενώ οι εξεταζόμενοι έλεγχοι χωρικής εξάρτησης προσφέρουν σημαντική βοήθεια στον εντοπισμό της ύπαρξης του προβλήματος της χωρικής αυτοσυσχέτισης στα σφάλματα ενός γραμμικού υποδείγματος, η συμπεριφορά τους ως προς τη διάκριση της αιτίας του προβλήματος εξαρτάται από το μέγεθος του δείγματος και σε κάποιες περιπτώσεις από τον ορισμό της χωρικής γειτνίασης μέσω της μήτρας των σταθμίσεων καθώς μπορεί να οδηγήσει σε τελείως διαφορετικά αποτελέσματα. Ασυνέπεια στα αποτελέσματα προέκυψε για τις ανθεκτικές στατιστικές LM-LE και LM-EL όταν η αιτία της αυτοσυσχέτισης είναι η ύπαρξη χωρικής υστέρησης στην εξαρτημένη μεταβλητή που έχει παραληφθεί από το υπόδειγμα όπου τα ποσοστά απόρριψης των ελέγχων με τις δύο στατιστικές παρουσιάζουν δραματικές διαφοροποιήσεις μεταξύ των κριτηρίων του πύργου και της βασίλισσας. Επομένως, ο ερευνητής θα πρέπει να εξετάζει τη συνέπεια των αποτελεσμάτων από τους ελέγχους ορίζοντας και χρησιμοποιώντας διαφορετικά κριτήρια χωρικής γειτνίασης πριν καταλήξει στο τελικό του συμπέρασμα.

Η χρησιμοποίηση μεγαλύτερων δειγμάτων δεν βελτιώνει απαραίτητα τα αποτελέσματα αλλά αυξάνει τα ποσοστά απόρριψης των ελέγχων προκαλώντας μερικές φορές ασάφεια ως προς τον εντοπισμό της αιτίας της αυτοσυσχέτισης κάτι που δεν είχε επισημανθεί στις παλιότερες γνωστές εργασίες διότι οι προσομοιώσεις τους χρησιμοποιούσαν μικρότερα δείγματα. Στις περιπτώσεις που δεν μπορεί να εντοπιστεί με σαφήνεια η αιτία του προβλήματος η σύγκριση των τιμών των στατιστικών LM-LE και LM-EL μπορεί να οδηγήσει στο σωστό υπόδειγμα. Η στρατηγική αυτή έχει αναφερθεί σε παλιότερες εργασίες χωρίς να έχει τεκμηριωθεί και χρησιμοποιείται σε πολλές εμπειρικές εφαρμογές.

ABSTRACT

In spatial econometrics, spatial dependence tests are applied for detecting the reason of the spatial autocorrelation problem in the errors of a linear model and for selecting the appropriate spatial econometric model. In this study, using simulation techniques, we confirm results presented in the literature concerning these tests performance in small and moderate samples and we contribute to further information for their behavior in large samples. Moreover, it seems that the comparison between the values of the two robust statistics is helpful for choosing the correct model in the cases where the tests produce confusing results.

ΑΝΑΦΟΡΕΣ

- Anselin, L. (1988). *Spatial Econometrics: Methods and Models*, Kluwer Academic Publishers, Dordrecht.
- Anselin L. (1988b). Lagrange Multiplier Test Diagnostics for Spatial Dependence and Spatial Heterogeneity. *Geographical Analysis*, **20**, 1-17.
- Anselin L., Bera A., Florax R. and Yoon, M. (1996). Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics*, **26**, 77-104.

- Anselin, L. and Florax, R. J. (1995). Small sample properties of tests for spatial dependence in regression models: Some further results, in Anselin, L. and Florax, R. J., (eds.). *New Directions in Spatial Econometrics*, 21-74. Springer-Verlag, Berlin.
- Anselin L. and Rey S. (1991). Properties of Tests for Spatial Dependence in Linear Regression Models. *Geographical Analysis*, **23**, 112-131.
- Burridge P. (1980). On the Cliff-Ord Test for Spatial Correlation. *Journal of the Royal Statistical Society B*, **42**, 107-108.
- Cliff A. and Ord J. K. (1972). Testing for Spatial Autocorrelation among Regression Residuals. *Geographical Analysis*, **4**, 267-284.

ΚΑΤΑΝΟΜΕΣ ΤΡΙΓΩΝΟΜΕΤΡΙΚΗΣ ΕΚΦΡΑΣΗΣ: Η ΠΕΡΙΠΤΩΣΗ ΤΗΣ ΕΦΑΠΤΟΜΕΝΗΣ

Φαρμάκης Νικόλαος, Παπατσούμα Ιωάννα

Τμήμα Μαθηματικών, Σχολή Θετικών Επιστημών, Αριστοτέλειο Πανεπιστήμιο
Θεσσαλονίκης

farmakis@math.auth.gr, ioannapapatsouma@gmail.com

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία εξετάζονται κατανομές με συνάρτηση πυκνότητας πιθανότητας (σ.π.π.) αύξουσας μορφής, οι οποίες εκφράζονται αναλυτικά με χρήση τριγωνομετρικών συναρτήσεων. Ειδικά εδώ ενδιαφέρουν οι σ.π.π. που εκφράζονται με χρήση της εφκ. Πιο συγκεκριμένα, επιχειρείται (για τις σ.π.π. που εκφράζονται με τέτοιες τριγωνομετρικές μορφές) μία προσεγγιστική έκφραση τους, με χρήση πολωνυμικής συνάρτησης προσέγγισης βαθμού n , από θεωρητική πρώτα σκοπιά και με χρήση αριθμητικών προσεγγιστικών μεθόδων. Η τιμή του εκθέτη n είναι αρκετά σταθερή σε βάθος αρκετών δεκαδικών κλασματικών ψηφίων (δ.κ.ψ.). Στη συνέχεια, επιχειρείται να προσδιοριστεί η αύξουσας μορφής σ.π.π. με βάση δεδομένα που προέρχονται από συστηματική δειγματοληψία κυρίως. Προσδιορίζεται ο εκθέτης n της πολωνυμικής έκφρασης καθώς στην όλη διαδικασία υπεισέρχεται ο συντελεστής μεταβλητότητας. Με κίνηση αντίθετης φοράς επιχειρείται να χρησιμοποιηθεί ο εκθέτης της πολωνυμικής έκφρασης ως στοιχείο ταυτοποίησης της αύξουσας σ.π.π. Αν προκύψει δηλαδή εκθέτης με συγκεκριμένη τιμή από δεδομένα να συμπεραίνεται ότι η αναλυτική μορφή της σ.π.π. με χρήση τριγωνομετρικών όρων ή παραγόντων να είναι συγκεκριμένη και άρα προβλέψιμη.

Λέξεις Κλειδιά: δειγματοληψία, συντελεστής μεταβλητότητας, τριγωνομετρική συνάρτηση, τυχαία μεταβλητή, κατανομή

AMS ταξινόμηση: 62D05, 62E17

1. ΣΥΝΑΡΤΗΣΗ ΠΥΚΝΟΤΗΤΑΣ ΠΙΘΑΝΟΤΗΤΑΣ

Θεωρούμε το διάστημα $I_1=[0, \alpha]$ της ευθείας των πραγματικών αριθμών $\mathbf{R}$ και την τυχαία μεταβλητή X τμ X , που παίρνει τιμές στο I_1 . Στο I_1 ορίζουμε τη συνάρτηση

$$f(x)=\begin{cases} h \cdot \varepsilon\phi\lambda x, & x \in [0, \alpha] \\ 0 & , x \notin [0, \alpha] \end{cases} \quad (1)$$

και αξιώνουμε η (1) να είναι συνάρτηση πυκνότητας πιθανότητας (σ.π.π.) της τμ X . Αυτό σημαίνει ότι το ολοκλήρωμα της συνάρτησης (1) στο I_1 θα πάρει την τιμή 1, ήτοι $\int_0^\alpha f(x) dx = 1$, οπότε μετά από πράξεις, έχουμε τον προσδιορισμό του

συντελεστή ως $h = \frac{\lambda}{\ln \frac{1}{\sin(\lambda\alpha)}}$. Στην περίπτωση που πάρουμε τον περιορισμό $\lambda\alpha = \pi/3$

για τον προσδιορισμό του I_1 , τότε ο συντελεστής h γίνεται $h = \frac{\pi}{3 \cdot \alpha \cdot \ln 2}$ και το

διάστημα I_1 ισοδυναμεί με το διάστημα $I_2 = [0, \pi/3]$. Αυτή η τελευταία περίπτωση παρουσιάζει ιδιαίτερο ενδιαφέρον για το πρόβλημα που μελετούμε στην παρούσα εργασία. Θα επικεντρώσουμε το ενδιαφέρον στις σ.π.π. που ορίζονται στο διάστημα $I_2 = [0, \pi/3]$. Υιοθετούμε ήδη τη συνάρτηση με την ειδική μορφή

$$f(x) = \begin{cases} \frac{\pi}{3 \cdot \alpha \cdot \ln 2} \cdot \varepsilon \varphi \frac{\pi x}{3\alpha}, & x \in [0, \alpha], \alpha > 0 \text{ και } \int_0^\alpha f(x) dx = 1. \\ 0, & x \notin [0, \alpha] \end{cases} \quad (2)$$

Για $\alpha = \pi/3$, η παραπάνω σχέση μπορεί να γραφεί απλούστερα (και λειτουργικότερα) ως

$$f(x) = \begin{cases} \frac{1}{\ln 2} \cdot \varepsilon \varphi x, & x \in [0, \frac{\pi}{3}] \text{ και } \int_0^{\frac{\pi}{3}} f(x) dx = 1. \\ 0, & x \notin [0, \frac{\pi}{3}] \end{cases} \quad (3)$$

Για τη μελέτη, λοιπόν, της τμ X με σ.π.π. την (1), αρκεί να εργαστούμε πάνω στη σχέση (3) και τα αποτελέσματα να επεκτείνονται με τους κατάλληλους μετασχηματισμούς αυτόματα στις (1) και (2).

2. ΥΠΟΛΟΓΙΣΜΟΙ ΠΑΡΑΜΕΤΡΩΝ ΤΥΧΑΙΑΣ ΜΕΤΑΒΛΗΤΗΣ X

Η μέση τιμή της τμ X που έχει σ.π.π. την (3) θα δίνεται από τη γνωστή σχέση

$$EX = \int_0^{\frac{\pi}{3}} x \cdot f(x) dx = -\frac{x \cdot \ln(\sin x)}{\ln 2} \Big|_0^{\frac{\pi}{3}} + \int_0^{\frac{\pi}{3}} \frac{\ln(\sin x)}{\ln 2} dx = \frac{\pi}{3} + \int_0^{\frac{\pi}{3}} \frac{\ln(\sin x)}{\ln 2} dx. \quad (4)$$

Η ποσότητα, όμως, $\int_0^{\frac{\pi}{3}} \frac{\ln(\sin x)}{\ln 2} dx$ δεν μπορεί να υπολογιστεί με πεπερασμένου πλήθους μετασχηματισμούς και θα χρειαστούν αριθμητικές μέθοδοι που θα λύσουν προσεγγιστικά το πρόβλημα.

Σε ανάλογο φυσικά πρόβλημα καταλήγει και η προσπάθεια να υπολογιστεί και η διασπορά της τμ X με βάση τον ορισμό της, δηλαδή από τη σχέση

$$VarX = EX^2 - (EX)^2 = \int_0^{\frac{\pi}{3}} x^2 \cdot f(x) dx - \left(\int_0^{\frac{\pi}{3}} x \cdot f(x) dx \right)^2. \quad (5)$$

Η προσεγγιστική μέθοδος για τον υπολογισμό των ανωτέρω ολοκληρωμάτων μπορεί να είναι μια από τις γνωστές από τη βιβλιογραφία, όπως η μέθοδος *τραπεζίου*, η μέθοδος *Simpson*, η μέθοδος $3/8$, κλπ., Χατζηδήμος (1980). Για τη σ.π.π. της σχέσης (3) ορισμένη στο διάστημα $I_2=[0, \pi/3]$ είναι καλύτερη η μέθοδος $3/8$, γιατί παρουσιάζει το μικρότερο προσεγγιστικό σφάλμα. Προσεκτικότερη παρατήρηση αναδεικνύει ότι η συνάρτηση (3) στο διάστημα $I_2=[0, \pi/3]$ συμπεριφέρεται ως τριτοβάθμιο τριώνυμο (περίπου). Η εφαρμογή της μεθόδου ολοκλήρωσης $3/8$ έχει τα εξής βήματα:

1^ο Βήμα: Επιλέγεται ένας αριθμός πολλαπλάσιο του 3, ο $3n$. Σε $3n$ ισομήκη διαστήματα θα διαιρεθεί το διάστημα ολοκλήρωσης $I_2=[0, \pi/3]$. Το μήκος (εύρος) καθενός από αυτά τα διαστήματα είναι προφανώς $w=\pi/(9n)$. Στο διάστημα $I_2=[0, \pi/3]$ ορίζονται έτσι (μαζί με τα άκρα του) $3n+1$ σημεία με συντεταγμένες $x_i=i\pi/(9n)$, $i=0,1,2,\dots,3n$. Στα σημεία αυτά ορίζεται αντίστοιχα η τιμή της συνάρτησης προς ολοκλήρωση $f(x_i)=y_i$.

2^ο Βήμα: Υπολογίζεται η τιμή του ολοκληρώματος κατά $3/8$ με βάση τη σχέση ολοκλήρωσης

$$Ig_{3n}=(3w/8)\{y_0+3y_1+3y_2+2y_3+3y_4+\dots+2y_{3n-3}+3y_{3n-2}+3y_{3n-1}+y_{3n}\} \quad (6)$$

Η ολοκληρωτική σχέση (6) βοηθάει στον υπολογισμό των παραμέτρων (4) και (5). Μετά υπολογίζεται η τιμή του συντελεστή μεταβλητότητας CV (ΣΜ) καθώς και της βοηθητικής ενδιάμεσης παραμέτρου $\lambda=CV^{-2}$. Είναι:

$$CV = \frac{\sqrt{VarX}}{EX} \text{ και } \lambda = CV^{-2} = \frac{(EX)^2}{VarX}. \quad (7)$$

Ο προσδιορισμός των τιμών των παραμέτρων (7) μπορεί όμως να γίνει και με άλλη καθαρά στατιστική μέθοδο που είναι και ακριβέστερη, τις περισσότερες φορές. Αυτό επειδή η μέθοδος $3/8$ είναι σχετικά ακριβής για τον υπολογισμό του ολοκληρώματος $\int_0^a \varphi x dx$ αλλά για τους υπολογισμούς των ολοκληρωμάτων $\int_0^a x \cdot \varphi x dx$ και $\int_0^a x^2 \cdot \varphi x dx$ χρειάζονται αριθμητικές μέθοδοι που είναι κατάλληλες για πολυώνυμα τέταρτου και πέμπτου βαθμού (ή συναρτήσεις που να προσομοιάζουν με αυτά). Αυτές, όμως, οι αριθμητικές μέθοδοι είναι πολύπλοκες και μάλλον περισσότερο χρονοβόρες από τη στατιστική μέθοδο που προαναφέρουμε. Κατά τη μέθοδο αυτή υιοθετούμε ένα σχετικά ευάριθμο δείγμα (π.χ. μεγέθους $n=800$) και φροντίζουμε να το κατανείμουμε σε k κλάσεις ίσου εύρους $w=\pi/3k$, επιλέγοντας ως $k=5, 6, 8, 10$ και 12 και με κατανομή των ατόμων του δείγματος βασισμένη στη συνάρτηση (3). Τα πλήθη των υποσυνόλων του δείγματος εκφράζονται με προσέγγιση ακέραιας μονάδας για την κάθε τάξη. Από αυτήν την κατανομή υπολογίζονται οι παράμετροι που παρουσιάστηκαν στη σχέση (7).

3. ΠΟΛΥΩΝΥΜΙΚΗ ΜΟΡΦΗ Σ.Π.Π. ΤΥΧΑΙΑΣ ΜΕΤΑΒΛΗΤΗΣ X

Οι σ.π.π. των σχέσεων (1), (2) και (3) είναι αύξουσες, συνεπώς η προσέγγισή τους μέσα από δειγματικά δεδομένα μπορεί να γίνει από το πολυωνυμικό μοντέλο της μορφής

$$f(x) = \begin{cases} h \cdot \left(\frac{x}{a}\right)^\nu, & x \in [0, a] \\ 0 & , x \notin [0, a] \end{cases} \quad (8)$$

γενικά, Farmakis (2010), Farmakis & Makris (2011) και ειδικότερα η (3) που ενδιαφέρει περισσότερο εδώ αντιστοιχεί στο πολυωνυμικό μοντέλο προσέγγισης

$$f(x) = \begin{cases} h \cdot \left(\frac{3x}{\pi}\right)^\nu, & x \in [0, \frac{\pi}{3}] \\ 0 & , x \notin [0, \frac{\pi}{3}] \end{cases} \quad (9)$$

Στη σχέση (9) οι παράμετροι ν (εκθέτης) και h (συντελεστής) προσδιορίζονται από τον ΣΜ μέσα από την ενδιάμεση βοηθητική παράμετρο λ της σχέσης (7), ως εξής Farmakis (2010):

$$\nu = -2 + \sqrt{1 + \lambda} \quad \text{και} \quad h = \frac{\nu + 1}{\frac{\pi}{3}} \quad (10)$$

ανεξάρτητα από την πηγή πληροφορίας για τον προσδιορισμό της τιμής του ΣΜ και του λ . Η ανεξαρτησία αυτή σημαίνει προφανώς ότι αν ο εκθέτης ν και ο συντελεστής h προσδιοριστούν θεωρητικά, τότε αποτελούν και μια μορφή ταυτότητας για τη σ.π.π. της (3). Γνωρίζοντας ήδη τις τιμές των ν και h , αν έχουμε δειγματικά δεδομένα που δίνουν τέτοιες τιμές για την πολυωνυμική προσέγγιση της σ.π.π. μιας τμ X , τότε μπορούμε να δεχτούμε ότι η αυθεντική έκφραση της σ.π.π. είναι η (3).

Για τη σ.π.π. έγιναν οι υπολογισμοί που περιγράφονται παραπάνω αλλά και στην προηγούμενη παράγραφο. Οι υπολογισμοί αυτοί έγιναν με βάση διάφορες τιμές του εύρους $w = \pi/(3\kappa)$ και του αντίστοιχου πλήθους σημείων n .

Τα αποτελέσματα φαίνονται στον παρακάτω Πίνακα 1. Για κάθε περίπτωση εύρους τάξεως $w = \pi/(3\kappa)$ υπολογίζονται 7 παράμετροι: η μέση τιμή, η διασπορά, η τυπική απόκλιση, ο ΣΜ, η ενδιάμεση βοηθητική παράμετρος λ και τέλος οι δύο κρισιμότερες τιμές του εκθέτη ν και του συντελεστή h για την πολυωνυμική προσέγγιση της έκφρασης της αντίστοιχης σ.π.π. της τμ X που μελετούμε. Τα σύμβολά τους φαίνονται στην πρώτη στήλη του πίνακα που ακολουθεί.

Πίνακας 1. Υπολογισμοί παραμέτρων για τις διάφορες τιμές του κ

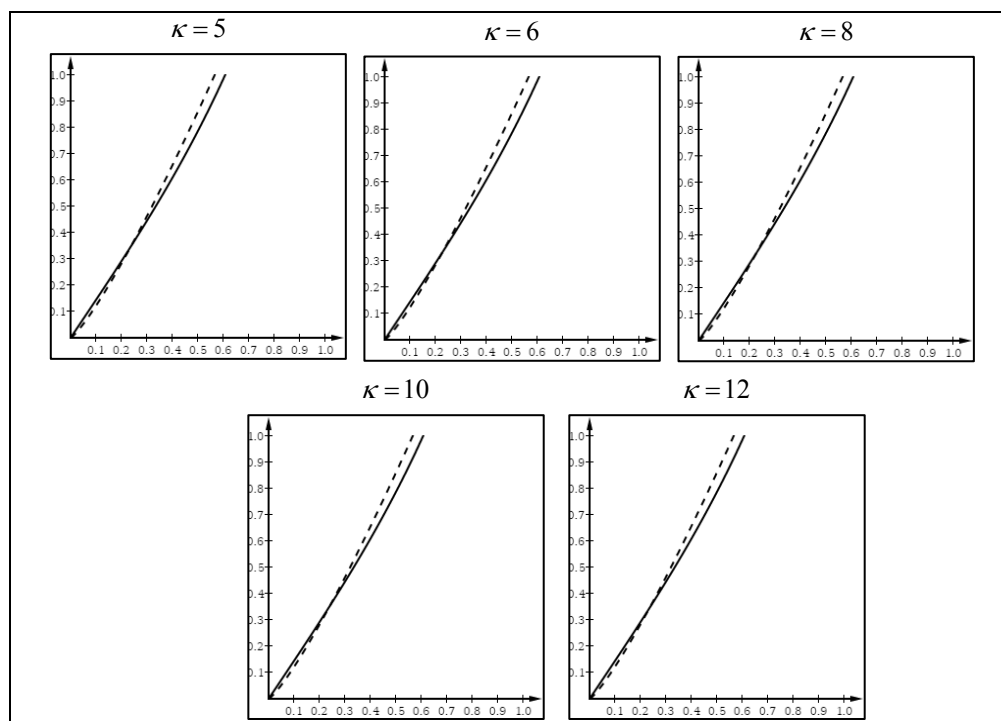
κ	5	6	8	10	12	Μέσες Τιμές
$w = \pi/(3\kappa)$	0.209440	0.174533	0.130900	0.104720	0.087266	

μ	0.723091	0.725621	0.728620	0.730682	0.731293	
σ^2	0.056079	0.056736	0.057205	0.057173	0.057421	
σ	0.236809	0.238192	0.239175	0.239108	0.239626	
ΣM	0.327496	0.32826	0.328258	0.327239	0.327674	0.327648
λ	9.323691	9.280324	9.280462	9.338334	9.313550	9.307272
ν	1.213050	1.206294	1.206316	1.215328	1.211472	1.210501
h	2.113307	2.106856	2.106877	2.115482	2.111800	2.110864
$\ln 2 =$	0.69314718					

Η πολύ καλή προσαρμογή των σ.π.π. πολωνυμικής μορφής (διακεκομμένη γραμμή) και τριγωνομετρικής μορφής (συνεχής γραμμή) για τις διάφορες τιμές του κ (Σχήμα 1) μας οδήγησε στην προσέγγιση των παραμέτρων ΣM , λ , ν και h με τις μέσες τιμές τους (τελευταία στήλη του Πίνακα 1).

Σημείωση 1^η: Για τον υπολογισμό των 7 παραμέτρων έγινε χρήση της μεθόδου των ομαδοποιημένων δεδομένων σε κ ομάδες (κλάσεις), με τιμές του $\kappa=5,6,8,10$ και 12, Φαρμάκης (2001), Κολυβά-Μαχαίρα & Μπόρα-Σέντα (2013).

Σχήμα 1. Γραφικές παραστάσεις των σ.π.π. πολωνυμικής και τριγωνομετρικής μορφής για τις διάφορες τιμές του κ



Τα σχέδια δειγματοληψίας, Φαρμάκης (2009^α και 2009^β), στηρίζονται στην υπόθεση ότι υπάρχει η σ.π.π. με έκφραση βασισμένη στη τριγωνομετρική μορφή εφχ, και έχουμε θεωρητική προσέγγιση των συχνοτήτων και άρα και της τιμής του εκθέτη ν, για διάφορα σχέδια δειγματοληψίας. Η τιμή του εκθέτη ν είναι αρκετά σταθερή και φαίνεται να είναι ανεξάρτητη από το σχέδιο, τείνοντας σε συγκεκριμένη τιμή. Με ακρίβεια 4 δεκαδικών κλασματικών ψηφίων μια καλή προσέγγιση είναι η ν=1.2105. Με κίνηση αντίθετης φοράς επιχειρείται να χρησιμοποιηθεί ο εκθέτης της πολυωνυμικής έκφρασης ως στοιχείο ταυτοποίησης της αύξουσας σ.π.π. Αν προκύψει δηλαδή εκθέτης με συγκεκριμένη τιμή από δεδομένα να συμπεραίνεται ότι η αναλυτική μορφή της σ.π.π. με χρήση τριγωνομετρικών όρων ή παραγόντων να είναι συγκεκριμένη. Δηλαδή να έχουμε σ.π.π. με τη μορφή $f(x) = \begin{cases} h \cdot \varepsilon \varphi ax, & x \in [0, \beta] \\ 0, & x \notin [0, \beta] \end{cases}$ για συγκεκριμένη τιμή του εκθέτη που θα προκύψει από τα δειγματοληπτικά δεδομένα.

4. ΠΑΡΑΔΕΙΓΜΑΤΑ ΠΟΥ ΑΝΑΔΕΙΚΝΥΟΥΝ ΤΡΙΓΩΝΟΜΕΤΡΙΚΗ ΜΟΡΦΗ Σ.Π.Π.

Είναι ενδεχόμενο από ένα δείγμα να προκύπτουν ενδείξεις ότι έχουμε σ.π.π. των σχέσεων (1), (2) και (3). Επειδή είναι αύξουσες η προσέγγισή τους μέσα από δειγματικά δεδομένα μπορεί να γίνει από το πολυωνυμικό μοντέλο της μορφής (8) ή καλύτερα με το προσαρμοσμένο μοντέλο της μορφής (9) και μέσα από τον υπολογισμό των παραμέτρων με βάση τη σχέση (10). Αν οι τιμές των παραμέτρων προκύψουν να είναι αυτές ή πολύ κοντινές με τις μέσες τιμές των ν και h, όπως φαίνονται στην τελευταία στήλη του Πίνακα 1, τότε είναι σωστό να υποθέσουμε ότι η σ.π.π. μπορεί να προσεγγιστεί και από την (3). Η εγγύτητα των τιμών που προκύπτουν από την επεξεργασία του δείγματος και των μέσων τιμών της τελευταίας στήλης του Πίνακα 1 εξαρτάται και από τις απαιτήσεις του προβλήματος. Μία μέτριας απαίτησης προσέγγιση εδώ μπορεί να αναφέρεται σε ακρίβεια 2 ή 3 δ.κ.ψ., ήτοι οι τιμές των 2 παραμέτρων πρέπει να είναι κοντά στις ν=1.211 και h=2.111. Δεν είναι γνωστή ανάλυση για δημιουργία διαστημάτων εμπιστοσύνης (δ.ε.) επί του παρόντος και επί αυτού το θέματος. Κάποια μορφή διαστήματος μπορεί να προκύψει μετά από επαναληπτική εφαρμογή της μεθόδου στο διάστημα $[0, \frac{\pi}{3}]$, όπου μεταφέρεται το οποιοδήποτε διάστημα $[0, \alpha]$ μέσω του μετασχηματισμού $x' = \pi x / 3\alpha$.

Σημειώνεται ότι για να γίνει η δειγματοληψία με βάση την κατανομή που προδικάζει η (3) υπολογίζουμε θεωρητικά την αναλογία των συχνοτήτων που εμφανίζονται στις κ κλάσεις $[i\omega, (i+1)\omega]$, $i=0,1,2,\dots,k-1$. Η αναλογία καθορίζεται από την ολοκλήρωση της (3) στα διαστήματα που αντιστοιχούν στις προηγούμενες κλάσεις. Με δεδομένο το μέγεθος του δείγματος, παίρνουμε από το κάθε στρώμα (κλάση) με απλή τυχαία δειγματοληψία τόσες παρατηρήσεις όσες προβλέπεται να είναι με βάση την αναλογία των συχνοτήτων που προαναφέραμε. Μικρές παρεκκλίσεις στον καθορισμό των μεγεθών των στρωματικών δειγμάτων μπορεί να προκύψουν λόγω στρογγυλεύσεων

στην εκάστοτε κοντινότερη ακέραια μονάδα. Έχουμε δηλαδή Στρωματοποιημένη Τυχαία Δειγματοληψία (ΣτΔ), Φαρμάκης (2009^α).

Παράδειγμα 4.1: Πήραμε ένα δείγμα μεγέθους $n=400$ με τη μέθοδο της στρωματοποιημένης δειγματοληψίας (ΣτΔ), Φαρμάκης (2009^{α,β}). Τα στρώματα είναι 5 και όλα έχουν το ίδιο εύρος $w=0.209440=\pi/15$. Η ένωσή τους καλύπτει επακριβώς το εύρος της κατανομής, το $[0, \frac{\pi}{3}]$. Το διάστημα $[0, \frac{\pi}{3}]$ έχει τη θέση του πληθυσμού, που χωρίζεται σε 5 ισομήκη στρώματα. Τα υποσύνολα του δείγματος που προέρχονται από τα 5 στρώματα έχουν αναλογία μεγέθους **24:80:140:220:336** που διαπιστώθηκε και θεωρητικά στην παράγραφο 1 με ολοκλήρωση της σχέσης (3). Μετά εφαρμόζεται η μέθοδος των ομαδοποιημένων δεδομένων, Φαρμάκης (2001), Κολυβά-Μαχαίρα & Μπόρα-Σέντα (2013). Με την εφαρμογή της στατιστικής μεθόδου έγινε ο προσδιορισμός των παραμέτρων στη σχέση (7). Πιο συγκεκριμένα στο κάθε στρώμα έγινε απλή τυχαία δειγματοληψία (ΑΤΔ) και τα υλοποιημένα μεγέθη στα στρώματα ήταν αντίστοιχα **12:40:70:110:168**. Συνολικά είχαμε 400 παρατηρήσεις. Πάρθηκαν 33 διαφορετικά τέτοια δείγματα. Οι τιμές των παραμέτρων ν και h πήραν τιμές στα παρακάτω αντίστοιχα διαστήματα: $[1.20021, 1.21485]$ και $[2.10105, 2.11503]$.

Παράδειγμα 4.2: Πήραμε ένα δείγμα μεγέθους $n=120$ με τη μέθοδο της στρωματοποιημένης δειγματοληψίας (ΣτΔ), Φαρμάκης (2009^{α,β}). Τα στρώματα είναι 5 και όλα έχουν το ίδιο εύρος $w=0.209440$. Η ένωσή τους καλύπτει επακριβώς το εύρος της κατανομής, το $[0, \frac{\pi}{3}]$. Το διάστημα $[0, \frac{\pi}{3}]$ έχει τη θέση του πληθυσμού, που χωρίζεται σε 5 ισομήκη στρώματα. Τα υποσύνολα του δείγματος που προέρχονται από τα 5 στρώματα έχουν αναλογία μεγέθους όπως αυτά στο Παράδειγμα 1^ο. Στο κάθε στρώμα έγινε απλή τυχαία δειγματοληψία (ΑΤΔ) και τα υλοποιημένα μεγέθη στα στρώματα ήταν αντίστοιχα **4:12:21:33:50**, μετά από στρογγυλοποίηση στη μονάδα. Ήτοι 120 παρατηρήσεις. Πάρθηκαν 43 διαφορετικά τέτοια δείγματα. Έγινε το ίδιο με το προηγούμενο παράδειγμα. Οι τιμές των παραμέτρων ν και h πήραν τιμές στα αντίστοιχα διαστήματα $[1.1566, 1.2274]$ ο εκθέτης ν και $[2.0594, 2.1270]$.

Σημείωση 2^η: Οι τιμές των παραμέτρων ν και h που προβλέπονται από την τελευταία στήλη του Πίνακα 1 βρίσκονται περίπου στο κέντρο των διαστημάτων του 1^{ου} παραδείγματος και με ελαφρά μετατόπιση προς τα δεξιά. Στο 2^ο παράδειγμα η αντίστοιχη μετατόπιση προς τα δεξιά είναι πολύ εντονότερη.

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Είναι ενδεχόμενο από ένα δείγμα να προκύπτουν ενδείξεις ότι έχουμε σ.π.π. των σχέσεων (1), (2) και (3). Δηλαδή να διαπιστώνεται ότι έχουμε μία κατανομή που η αντίστοιχη σ.π.π. να είναι αύξουσα και ο χαρακτήρας της αύξησης να είναι εντονότερος καθώς πλησιάζουμε το δεξιό άκρο του πεδίου ορισμού (διάστημα της ευθείας των πραγματικών αριθμών). Αν ορίσουμε την αρχή του άξονα στο αριστερό άκρο του πεδίου ορισμού όπως αυτό προκύπτει από την αρχική ανάλυση των δειγματοληπτικών δεδομένων και αποκαταστήσουμε την αναλογία ώστε το πεδίο ορισμού $[0, \alpha]$ της σ.π.π. να ισοδυναμεί με το διάστημα $[0, \frac{\pi}{3}]$, τότε προσπαθούμε να ορίσουμε τη σ.π.π. με βάση το μοντέλο που περιγράφουν οι σχέσεις (9) και (10). Αν οι παράμετροι εκθέτης ν και συντελεστής h παίρνουν τιμές 1.211 και 2.111 ή έστω συγκρίσιμες με αυτές, τότε μπορούμε να υποθέσουμε βάσιμα ότι η σ.π.π. που προσεγγίστηκε με βάση τις σχέσεις (9) και (10) είναι στην πραγματικότητα η συνάρτηση (3). Το συγκρίσιμο των τιμών του εκθέτη και του συντελεστή με τις τιμές 1.211 και 2.111 αντίστοιχα μπορεί να προκύψει από επαναλαμβανόμενη δειγματοληψία ή έστω με εξαγωγή υποσυνόλων του δείγματος (επαναληπτικά) ώστε να διαπιστωθεί ότι οι τιμές 1.211 και 2.111 παρεμβάλλονται στο εύρος των παρατηρούμενων αντίστοιχων τιμών που θα προκύψουν από τα επαναληπτικά δείγματα. Το εύρος των τιμών που προκύπτει από τα επαναλαμβανόμενα δείγματα ενεργεί δίκην διαστήματος εμπιστοσύνης, χωρίς να έχει πάντως τη σχετική συμμετρία και την αξιοπιστία του διαστήματος εμπιστοσύνης κλασσικής μορφής.

Σημείωση 3^η: Η παρούσα εργασία είναι η 3^η κατά σειρά και αποτελεί μία ενότητα μαζί με τις εργασίες Παπατσούμα, Φαρμάκης (2015) και Φαρμάκης, Παπατσούμα (2016). Όλες προέκυψαν σε μια προσπάθεια να έχουμε πολυωνυμικές προσεγγίσεις για σ.π.π. που έχουν συγκεκριμένη μορφή σχετιζόμενη με κλασικές μαθηματικές εκφράσεις, όπως είναι οι τριγωνομετρικές συναρτήσεις.

Ευχαριστίες: Οι συγγραφείς ευχαριστούν τους κριτές για τις πολύ χρήσιμες και εποικοδομητικές παρατηρήσεις.

ABSTRACT

This paper examines distributions with increasing probability density function (pdf), expressed with the trigonometric function $\tan x$. In particular, a polynomial approach of pdf is attempted, using systematic sampling and the coefficient of variation. The use of polynomial expression's exponent as identifier of the increasing pdf is also attempted, in order to assess the analytical form using trigonometric conditions or factors, when the exponent is obtained from data.

ΑΝΑΦΟΡΕΣ

- Κολυβά-Μαχαίρα Φ., Μπόρα-Σέντα Ε. (2013) "Στατιστική: Θεωρία και Εφαρμογές", Β' Έκδοση, Εκδόσεις Ζήτη, Θεσσαλονίκη.
- Παπατσούμα Ι., Φαρμάκης Ν. (2015) "Πολυωνυμική έκφραση συμμετρικών κατανομών: Η περίπτωση σ.π.π. τριγωνομετρικής μορφής", *Πρακτικά 28^{ου} Πανελληνίου Συνεδρίου Στατιστικής*, σελ. 282-294.
- Φαρμάκης Ν. (2001) "ΣΤΑΤΙΣΤΙΚΗ, Περιληπτική Θεωρία-Ασκήσεις", Β' Έκδοση. Εκδόσεις Α&Π Χριστοδουλίδη, Θεσσαλονίκη.
- Φαρμάκης Ν. (2009^α) "Εισαγωγή στη Δειγματοληψία", Εκδόσεις Α&Π Χριστοδουλίδη, Θεσσαλονίκη.
- Φαρμάκης Ν. (2009^β) "Δημοσκοπήσεις και Δεοντολογία", Β' Έκδοση. Εκδόσεις Α&Π Χριστοδουλίδη, Θεσσαλονίκη.
- Φαρμάκης Ν., Παπατσούμα Ι. (2016) "Αναλυτική έκφραση πολυωνυμικής μορφής της συνάρτησης πυκνότητας πιθανότητας τυχαίας μεταβλητής: Η περίπτωση του συνημιτόνου", *Τετράδια Ανάλυσης Δεδομένων (υπό δημοσίευση)*
- Χατζηδήμος Α., (1980) "Εισαγωγή στην Αριθμητική Ανάλυση" Β' Έκδοση. Πανεπιστήμιο Ιωαννίνων.
- Farmakis, N., (2003). "Estimation of Coefficient of Variation: Scaling of Symmetric Continuous Distributions", *Statistics in Transition*, Vol. 6, No 1, pp 83-96.
- Farmakis, N., (2010). "Coefficient of Variation: Connecting Sampling with some Increasing Distribution Models", *Proceedings of Stochastic Modeling Techniques & Data Analysis International Conference (SMTDA-2010)*, June 8-11, 2010, Chania, Crete, GREECE.
- Farmakis, N., Makris G., (2011). "Web-sampling: Probabilities of Specific Information Achievement", *Proceedings of Applied Stochastic Models Data Analysis (ASMDA2011)*, XIV ASMDA2011, Rome, Italy.

ΑΝΑΘΕΩΡΗΜΑ ΤΟΥ ΙΣΟΔΥΝΑΜΟΥ ΤΩΝ ΔΥΟ ΛΥΣΕΩΝ ΤΩΝ PASCAL ΚΑΙ FERMAT ΣΤΟ ΚΛΑΣΙΚΟ ΠΡΟΒΛΗΜΑ ΤΗΣ ΔΙΑΝΟΜΗΣ ΜΕΡΙΔΙΩΝ

Χαράλαμπος Α. Χαραλαμπίδης
Τμήμα Μαθηματικών, Πανεπιστήμιο Αθηνών

ΠΕΡΙΛΗΨΗ

Το κλασικό πρόβλημα της διανομής μεριδίων στοίχηματος σε διακοπείσα σειρά παιγνιδιών, μεταξύ δύο παικτών, πριν την ανάδειξη νικητή αποτελεί το αντικείμενο του παρόντος άρθρου. Συγκεκριμένα, παρουσιάζονται αρχικά οι λύσεις του προβλήματος στις περιπτώσεις (α) ισοδυνάμων και (β) όχι κατ' ανάγκη ισοδυνάμων παικτών. Ακολούθως, αποδεικνύεται συνολοθεωρητικά αλλά και συνδυαστικά ότι οι δύο προταθείσες από τους Pascal και Fermat λύσεις, οι οποίες βασίζονται στα γνωστά σήμερα ως διωνυμικό και αρνητικό διωνυμικό στοχαστικά πρότυπα, οδηγούν στο ίδιο αποτέλεσμα.

Λέξεις Κλειδιά: Πρόβλημα σημείων, Διωνυμική κατανομή, Αρνητική διωνυμική κατανομή, Pascal, Fermat, Roberval, D' Alembert.

1. ΕΙΣΑΓΩΓΙΚΑ

Το κλασικό πρόβλημα της διανομής μεριδίων στοίχηματος σε διακοπείσα σειρά παιγνιδιών, το οποίο είναι το παλαιότερο γνωστό πιθανοθεωρητικό πρόβλημα, αφορά μια σειρά παιγνιδιών μεταξύ δύο παικτών και νικητής αναδεικνύεται εκείνος που κερδίζει πρώτος προκαθορισμένο αριθμό παιγνιδιών. Η σειρά των παιγνιδιών διακόπτεται, για κάποιο λόγο, πριν κάποιος παίκτης αναδειχθεί νικητής. Στην περίπτωση αυτή πώς πρέπει να μοιρασθεί δίκαια το στοίχημα;

Ο Luca de Burgo Pacioli, Φραγκισκανός καλόγερος, το 1494 στο βιβλίο του *Summa de Arithmetica, Geometria, Proportioni et Proportionalità* πρότεινε το στοίχημα να μοιρασθεί ανάλογα με τα παιγνίδια που κάθε παίκτης κέρδισε μέχρι την διακοπή. Σημειώνεται ότι οι αριθμοί των υπολειπόμενων νικών των δύο παικτών δεν εισέρχονται στην προταθείσα διανομή των μεριδίων.

Ο Girolamo Cardano, Ιταλός μαθηματικός και άπληστος χαρτοπαίκτης, το 1539 στο βιβλίο του *Practica Arithmetica* απέρριψε την λύση του Pacioli ως περιλαμβάνουσα λάθος "που ακόμα και ένα παιδί αναγνωρίζει" και πρότεινε το στοίχημα να μοιρασθεί

ανάλογα με το άθροισμα των θετικών ακεραίων μέχρι τον αριθμό των υπολειπόμενων νικών του κάθε παίκτη.

Ο Nicolo Tartaglia, Ιταλός μαθηματικός και ανταγωνιστής του Cardano, το 1556 στο βιβλίο του *General Trattato di Numeri e Measure*, επίσης απέρριψε την λύση του Pacioli ως "ούτε δικαιολογημένη ούτε καλή". Η προταθείσα απ' αυτόν λύση βασίζεται στο ηλίκο της διαφοράς των νικών των δύο παικτών και του διπλασίου του προκαθορισμένου αριθμού των νικών που απαιτούνται για την ανάδειξη του νικητή της σειράς. Οι επίσης Ιταλοί μαθηματικοί Francesco Peverone το 1581 στο βιβλίο του *Arithmetica e Geometria* και ο Lorenzo Forestani το 1603 στο βιβλίο του *Practica d'Arithmetica e Geometria* ασχολήθηκαν με το πρόβλημα αυτό αλλά οι λύσεις που πρότειναν ήταν λανθασμένες.

Ο Antoine Gombaud, Γάλλος αριστοκράτης και παθολογικός παίκτης τυχερών παιγνιδιών, ο οποίος είναι περισσότερο γνωστός ως Chevalier de Merè, το 1654 έθεσε υπόψη του διάσημου Γάλλου μαθηματικού Blaise Pascal ένα πρόβλημα διανομής μεριδίων, όμοιο με το αρχικό πρόβλημα του Pacioli. Ο Pascal αλληλογράφησε για το πρόβλημα αυτό με τον επίσης διάσημο Γάλλο μαθηματικό Pierre de Fermat. Και οι δύο έδωσαν διαφόρους τρόπους λύσης του προβλήματος. Δυστυχώς, δεν έχει βρεθεί αντίγραφο του προβλήματος διανομής μεριδίων του de Merè και επίσης κάποιες από τις επιστολές που αντάλλαζαν οι Pascal και Fermat έχουν χαθεί. Όμως οι ιστορικοί των μαθηματικών προσπάθησαν να ανακατασκευάσουν το πρόβλημα διανομής μεριδίων, το οποίο σε ελαφρά πιο γενική διατύπωση έχει ως εξής:

Έστω ότι δύο παίκτες α και β αγωνίζονται σε μια σειρά παιγνιδιών και νικητής αναδεικνύεται εκείνος που κερδίζει πρώτος ν παιγνίδια. Ας υποθέσουμε ότι για κάποιο λόγο διακόπτεται η σειρά των παιγνιδιών όταν ο α χρειάζεται κ ακόμα νίκες για να αναδειχθεί νικητής της σειράς, ενώ ο β χρειάζεται r νίκες. Στην περίπτωση αυτή πώς πρέπει να μοιρασθεί δίκαια το στοίχημα;

Στο άρθρο αυτό παρουσιάζονται στα δύο πρώτα εδάφια οι λύσεις του προβλήματος στις περιπτώσεις (α) ισοδυνάμων και (β) όχι κατ' ανάγκη ισοδυνάμων παικτών. Στο τελευταίο εδάφιο αποδεικνύεται η ισότητα των πιθανοτήτων που προκύπτουν από τις δύο προτεινόμενες λύσεις. Σημειώνουμε ότι τα στοιχεία της ανωτέρω σύντομης παρουσίασης της διαχρονικής εξέλιξης του προβλήματος της διανομής μεριδίων αντλήθηκαν από τα βιβλία ιστορίας των πιθανοτήτων των David (1962) και Todhunter (1865) και την ελεύθερη στο διαδίκτυο εγκυκλοπαίδεια Wikipedia στο λήμμα Problem of Points.

2. ΙΣΟΔΥΝΑΜΟΙ ΠΑΙΚΤΕΣ

Ας θεωρήσουμε αρχικά την περίπτωση ισοδυνάμων παικτών. Η λύση που πρότεινε ο Fermat βασίζεται στην παρατήρηση ότι η σειρά των παιγνιδιών κρίνεται με βεβαιότητα σε $\kappa + r - 1$ πρόσθετα παιγνίδια, με $2^{\kappa+r-1}$ ισοπίθανα στοιχειώδη ενδεχόμενα. Το σύνολο $\Omega^{\kappa+r-1}$ των στοιχειωδών ενδεχομένων (δειγματικών σημείων) περιλαμβάνει τις διατάξεις, $(\omega_1, \omega_2, \dots, \omega_{\kappa+r-1})$, των δύο συμβόλων $\Omega = \{\alpha, \beta\}$ ανά $\kappa + r - 1$ με επανάληψη, όπου α παριστάνει νίκη του παίκτη α και β παριστάνει νίκη

του παίκτη β . Το ενδεχόμενο A_i ο παίκτης α να κερδίσει i από τα $\kappa + r - 1$ πρόσθετα παιχνίδια περιλαμβάνει τις διατάξεις, $(\omega_1, \omega_2, \dots, \omega_{\kappa+r-1})$, των δύο συμβόλων $\{\alpha, \beta\}$ ανά $\kappa + r - 1$ με επανάληψη οι οποίες περιέχουν i ακριβώς α και $\kappa + r - i - 1$ ακριβώς β και είναι πλήθους $\binom{\kappa+r-1}{i}$, οπότε

$$P(A_i) = \binom{\kappa+r-1}{i} \frac{1}{2^{\kappa+r-1}}, \quad i = 0, 1, \dots, \kappa + r - 1.$$

Περαιτέρω $A = \bigcup_{i=\kappa}^{\kappa+r-1} A_i$, με $A_i \cap A_j = \emptyset$ για $i \neq j$, είναι το ενδεχόμενο ο παίκτης α να αναδειχθεί νικητής της σειράς, επιτυγχάνοντας κ τουλάχιστο νίκες, και έτσι

$$P(A) = \frac{1}{2^{\kappa+r-1}} \sum_{i=\kappa}^{\kappa+r-1} \binom{\kappa+r-1}{i} = \frac{1}{2^{\kappa+r-1}} \sum_{i=0}^{r-1} \binom{\kappa+r-1}{\kappa+i}.$$

Σημειώνουμε ότι ο Fermat δεν έδωσε εκπεφρασμένα την πιθανότητα αυτή, αλλά διατύπωσε με λόγια τον γενικό κανόνα υπολογισμού και υπολόγισε για κάθε παίκτη και για μικρά κ και r τις πιθανότητες συνολικής νίκης και πρότεινε ως δίκαιη την διανομή του στοιχήματος ανάλογα με τις πιθανότητες αυτές.

Οι Roberval και D' Alembert διατύπωσαν την αντίρρησή τους στο ότι η υπόθεση των $\kappa + r - 1$ πρόσθετων παιχνιδιών δεν μπορεί να γίνει εφ' όσον η σειρά των παιχνιδιών μπορεί λήξει με νικητή σε λιγότερα παιχνίδια. Ο Pascal απάντησε στην αντίρρηση αυτή υποστηρίζοντας ότι, αν και είναι δυνατό η σειρά των παιχνιδιών να κριθεί σε λιγότερα από $\kappa + r - 1$ πρόσθετα παιχνίδια, τα επιπλέον παιχνίδια δεν παίζουν κανένα ρόλο στην ανάδειξη του νικητή. Ο Fermat έδωσε μια δεύτερη λύση, η οποία παρακάμπτει την επίμαχη υπόθεση και έχει ως εξής:

Έστω B_j το ενδεχόμενο ο παίκτης α να αναδειχθεί νικητής της σειράς στο j -οστό πρόσθετο παιχνίδι, για $j = \kappa, \kappa + 1, \dots, \kappa + r - 1$. Τότε $B = \bigcup_{j=\kappa}^{\kappa+r-1} B_j$, με $B_i \cap B_j = \emptyset$ για $i \neq j$, είναι το ενδεχόμενο ο παίκτης α να αναδειχθεί νικητής της σειράς. Επίσης

$$P(B_j) = \binom{j-1}{\kappa-1} \frac{1}{2^j}, \quad j = \kappa, \kappa + 1, \dots, \kappa + r - 1,$$

και έτσι

$$P(B) = \sum_{j=\kappa}^{\kappa+r-1} \binom{j-1}{\kappa-1} \frac{1}{2^j} = \sum_{j=0}^{r-1} \binom{\kappa+j-1}{j} \frac{1}{2^{\kappa+j}}.$$

Ομοίως, η πιθανότητα του ενδεχομένου B' ο παίκτης β να αναδειχθεί νικητής της σειράς υπολογίζεται ως

$$P(B') = \sum_{j=r}^{\kappa+r-1} \binom{j-1}{r-1} \frac{1}{2^j} = \sum_{j=0}^{\kappa-1} \binom{r+j-1}{j} \frac{1}{2^{r+j}}.$$

Εναλλακτικά, και επειδή $P(B') = 1 - P(B)$,

$$P(B') = 1 - \sum_{j=0}^{r-1} \binom{\kappa + j - 1}{j} \frac{1}{2^{\kappa+j}}.$$

Σημειώνουμε ότι εξισώνοντας τις δύο τελευταίες εκφράσεις της πιθανότητας $P(B')$,

$$\sum_{j=0}^{\kappa-1} \binom{r + j - 1}{j} \frac{1}{2^{r+j}} = 1 - \sum_{j=0}^{r-1} \binom{\kappa + j - 1}{j} \frac{1}{2^{\kappa+j}},$$

και θέτοντας $r = \kappa$ παίρνουμε τη σχέση

$$\sum_{j=0}^{\kappa-1} \binom{\kappa + j - 1}{j} \frac{1}{2^{\kappa+j}} = \frac{1}{2}.$$

Εισάγοντας αυτή στην πιθανότητα $P(B)$, συνάγουμε την έκφραση

$$P(B) = \begin{cases} \frac{1}{2} + \sum_{j=\kappa}^{r-1} \binom{\kappa + j - 1}{j} \frac{1}{2^{\kappa+j}}, & r > \kappa, \\ \frac{1}{2}, & r = \kappa, \\ \frac{1}{2} - \sum_{j=r}^{\kappa-1} \binom{\kappa + j - 1}{j} \frac{1}{2^{\kappa+j}}, & r < \kappa. \end{cases}$$

3. ΑΝΙΣΟΔΥΝΑΜΟΙ ΠΑΙΚΤΕΣ

Ας θεωρήσουμε στο εδάφιο αυτό την περίπτωση ανισοδυνάμων παικτών. Συγκριμένα, σε κάθε παιγνίδι της σειράς, έστω p η πιθανότητα να κερδίσει ο παίκτης α και $q = 1 - p$ η πιθανότητα να κερδίσει ο παίκτης β . Τότε, χρησιμοποιώντας τη διωνυμική κατανομή (κατ' αναλογία με την αρχική λύση του Fermat), συνάγουμε διαδοχικά τις πιθανότητες

$$P(A_i) = \binom{\kappa + r - 1}{i} p^i q^{\kappa+r-i-1}, \quad i = 0, 1, \dots, \kappa + r - 1,$$

και

$$P(A) = \sum_{i=\kappa}^{\kappa+r-1} \binom{\kappa + r - 1}{i} p^i q^{\kappa+r-i-1} = \sum_{i=0}^{r-1} \binom{\kappa + r - 1}{\kappa + i} p^{\kappa+i} q^{r-i-1}.$$

Επίσης, χρησιμοποιώντας την κατανομή Pascal ή την αρνητική διωνυμική κατανομή (κατ' αναλογία με την δεύτερη λύση του Fermat), συνάγουμε διαδοχικά τις πιθανότητες

$$P(B_j) = \binom{j-1}{\kappa-1} p^{\kappa} q^{j-\kappa}, \quad j = \kappa, \kappa + 1, \dots, \kappa + r - 1,$$

και

$$P(B) = \sum_{j=\kappa}^{\kappa+r-1} \binom{j-1}{\kappa-1} p^{\kappa} q^{j-\kappa} = \sum_{j=0}^{r-1} \binom{\kappa+j-1}{j} p^{\kappa} q^j.$$

Οι ανωτέρω εκφράσεις της πιθανότητας ο παίκτης α να αναδειχθεί νικητής της σειράς δόθηκαν για πρώτη φορά από τον Montmort (1713).

4. ΙΣΟΔΥΝΑΜΟ ΤΩΝ ΔΥΟ ΠΡΟΤΕΙΝΟΜΕΝΩΝ ΛΥΣΕΩΝ

Η αρχική διατύπωση του προβλήματος της διανομής μεριδιών αφορούσε τη μερική περίπτωση δύο ισοδυνάμων παικτών και την υπόθεση ότι κατά την διακοπή της σειράς των παιγνιδιών ο α χρειαζόταν $\kappa = 2$ ακόμα νίκες για να αναδειχθεί νικητής της σειράς, ενώ ο β χρειαζόταν $r = 3$ νίκες. Σύμφωνα με την πρώτη λύση του Fermat, αν συνεχιζόταν η σειρά των παιγνιδιών θα αναδείκνυε οπωσδήποτε νικητή σε $\kappa+r-1 = 4$ πρόσθετα παιγνίδια. Τότε

$$\begin{aligned} \Omega^4 = \{ & (\beta, \beta, \beta, \beta), (\alpha, \beta, \beta, \beta), (\beta, \alpha, \beta, \beta), (\beta, \beta, \alpha, \beta), (\beta, \beta, \beta, \alpha), \\ & (\alpha, \alpha, \beta, \beta), (\alpha, \beta, \alpha, \beta), (\alpha, \beta, \beta, \alpha), (\beta, \alpha, \alpha, \beta), (\beta, \alpha, \beta, \alpha), (\beta, \beta, \alpha, \alpha), \\ & (\alpha, \alpha, \alpha, \beta), (\alpha, \alpha, \beta, \alpha), (\alpha, \beta, \alpha, \alpha), (\beta, \alpha, \alpha, \alpha), (\alpha, \alpha, \alpha, \alpha) \}, \end{aligned}$$

με $N(\Omega^4) = 2^4 = 16$. Επίσης

$$A_2 = \{(\alpha, \alpha, \beta, \beta), (\alpha, \beta, \alpha, \beta), (\alpha, \beta, \beta, \alpha), (\beta, \alpha, \alpha, \beta), (\beta, \alpha, \beta, \alpha), (\beta, \beta, \alpha, \alpha)\},$$

$$A_3 = \{(\alpha, \alpha, \alpha, \beta), (\alpha, \alpha, \beta, \alpha), (\alpha, \beta, \alpha, \alpha), (\beta, \alpha, \alpha, \alpha)\}, \quad A_4 = \{(\alpha, \alpha, \alpha, \alpha)\},$$

και

$$\begin{aligned} A &= A_2 \cup A_3 \cup A_4 \\ &= \{(\alpha, \alpha, \beta, \beta), (\alpha, \beta, \alpha, \beta), (\alpha, \beta, \beta, \alpha), (\beta, \alpha, \alpha, \beta), (\beta, \alpha, \beta, \alpha), (\beta, \beta, \alpha, \alpha) \\ &\quad (\alpha, \alpha, \alpha, \beta), (\alpha, \alpha, \beta, \alpha), (\alpha, \beta, \alpha, \alpha), (\beta, \alpha, \alpha, \alpha), (\alpha, \alpha, \alpha, \alpha)\}, \end{aligned}$$

οπότε

$$P(A) = P(A_2) + P(A_3) + P(A_4) = \frac{6}{16} + \frac{4}{16} + \frac{1}{16} = \frac{11}{16}.$$

Σύμφωνα με την δεύτερη λύση του Fermat, αν συνεχιζόταν η σειρά των παιγνιδιών ο παίκτης α μπορούσε να αναδειχθεί νικητής της σειράς (κερδίζοντας δύο ακόμα νίκες) είτε στο δεύτερο, είτε στο τρίτο, είτε στο τέταρτο πρόσθετο παίγνι. Τότε

$$B_2 = \{(\alpha, \alpha)\}, \quad B_3 = \{(\alpha, \beta, \alpha), (\beta, \alpha, \alpha)\}, \quad B_4 = \{(\alpha, \beta, \beta, \alpha), (\beta, \alpha, \beta, \alpha), (\beta, \beta, \alpha, \alpha)\}$$

και

$$P(B) = P(B_2) + P(B_3) + P(B_4) = \frac{1}{2^2} + 2 \cdot \frac{1}{2^3} + 3 \cdot \frac{1}{2^4} = \frac{4}{16} + \frac{4}{16} + \frac{3}{16} = \frac{11}{16}.$$

Η ισότητα των δύο εκφράσεων της πιθανότητας ο παίκτης α να αναδειχθεί νικητής της σειράς, $P(A) = P(B)$, ισχύει όχι μόνο για ισοδύναμους αλλά και για ανισοδύναμους παίκτες, όπως επίσης και για παίκτες με μεταβαλλόμενες πιθανότητες νίκης σε κάθε παιχνίδι. Τούτο συνάγεται από την ακόλουθη συνολοθεωρητική απόδειξη:

Ας θεωρήσουμε τα σύνολα (καρτεσιανά γινόμενα)

$$\Gamma_2 = B_2 \times \Omega^2 = \{(\alpha, \alpha, \alpha, \alpha), (\alpha, \alpha, \alpha, \beta), (\alpha, \alpha, \beta, \alpha), (\alpha, \alpha, \beta, \beta)\},$$

$$\Gamma_3 = B_3 \times \Omega = \{(\alpha, \beta, \alpha, \alpha), (\alpha, \beta, \alpha, \beta), (\beta, \alpha, \alpha, \alpha), (\beta, \alpha, \alpha, \beta)\},$$

και

$$\Gamma_4 = B_4 = \{(\alpha, \beta, \beta, \alpha), (\beta, \alpha, \beta, \alpha), (\beta, \beta, \alpha, \alpha)\},$$

οπότε

$$\begin{aligned} \Gamma &= \Gamma_2 \cup \Gamma_3 \cup \Gamma_4 \\ &= \{(\alpha, \alpha, \alpha, \alpha), (\alpha, \alpha, \alpha, \beta), (\alpha, \alpha, \beta, \alpha), (\alpha, \alpha, \beta, \beta), (\alpha, \beta, \alpha, \alpha), (\alpha, \beta, \alpha, \beta), \\ &\quad (\beta, \alpha, \alpha, \alpha), (\beta, \alpha, \alpha, \beta), (\alpha, \beta, \beta, \alpha), (\beta, \alpha, \beta, \alpha), (\beta, \beta, \alpha, \alpha)\} = A. \end{aligned}$$

Τότε

$$P(\Gamma_2) = P(B_2)P(\Omega^2) = P(B_2), \quad P(\Gamma_3) = P(B_3)P(\Omega) = P(B_3), \quad P(\Gamma_4) = P(B_4)$$

και

$$P(A) = P(\Gamma) = P(\Gamma_2) + P(\Gamma_3) + P(\Gamma_4) = P(B_2) + P(B_3) + P(B_4) = P(B).$$

Η συνολοθεωρητική αυτή απόδειξη επεκτείνεται άμεσα σε οποιουδήποτε θετικούς ακεραίους κ και r . Σημειώνουμε ότι στη σχετική αντίρρηση των Roberval και D' Alembert, ο Fermat σε επιστολή του στον Pascal υποστήριξε την αρχική του λύση γράφοντας ότι η επέκταση του πλήθους των παιχνιδιών σε συγκεκριμένο αριθμό υπηρετεί μόνο την ευκολία του κανόνα υπολογισμού και καθιστά τα στοιχειώδη ενδεχόμενα ισοπίθανα. Επισημαίνουμε ότι, σύμφωνα την ανωτέρω συνολοθεωρητική απόδειξη, η ισοδυναμία των δύο λύσεων ισχύει και για μη ισοπίθανα δειγματικά σημεία. Δεν πρέπει όμως να παραβλέψουμε ότι οι Pascal και Fermat συζητούσαν τη μερική περίπτωση, $p = 1/2$, ισοδυνάμων παικτών.

Στις περιπτώσεις ισοδυνάμων παικτών, $p = q = 1/2$, αλλά και ανισοδυνάμων παικτών, $p \neq q$, η ισότητα των δύο εκφράσεων της πιθανότητας ο παίκτης α να αναδειχθεί νικητής της σειράς, δύναται να αποδειχθεί και συνδυαστικά ως εξής: Πολλαπλασιάζοντας την πρώτη έκφραση της πιθανότητας $P(B)$ με

$$\sum_{i=0}^{\kappa+r-j-1} \binom{\kappa+r-j-1}{i} p^i q^{\kappa+r-j-i-1} = (p+q)^{\kappa+r-j-1} = 1$$

και εναλλάσσοντας τη σειρά της άθροισης, συνάγουμε διαδοχικά

$$\begin{aligned} P(B) &= \sum_{j=\kappa}^{\kappa+r-1} \binom{j-1}{\kappa-1} p^{\kappa} q^{j-\kappa} \sum_{i=0}^{\kappa+r-j-1} \binom{\kappa+r-j-1}{i} p^i q^{\kappa+r-j-i-1} \\ &= \sum_{i=0}^{r-1} p^{\kappa+i} q^{r-i-1} \left\{ \sum_{j=\kappa}^{\kappa+r-i-1} \binom{j-1}{\kappa-1} \binom{\kappa+r-j-1}{i} \right\}. \end{aligned}$$

Χρησιμοποιώντας τη συνδυαστική ταυτότητα (βλ. Charalambides (2002), σελ. 122)

$$\sum_{j=\kappa}^{\kappa+r-i-1} \binom{j-1}{\kappa-1} \binom{\kappa+r-j-1}{i} = \binom{\kappa+r-1}{\kappa+i},$$

συμπεραίνουμε ότι

$$P(B) = \sum_{i=0}^{r-1} \binom{\kappa+r-1}{\kappa+i} p^{\kappa+i} q^{r-i-1} = P(A).$$

ABSTRACT

The classical problem of division of stakes after a series of games between two players is interrupted before neither player has achieved victory, is the subject of the present article. Specifically, the solutions of the problem in the particular cases (a) of equally skilled and (b) not necessarily equally skilled players are presented first. Subsequently, the two proposed solutions by Fermat and Pascal, which are based on the nowadays known as binomial and negative binomial stochastic models, are set-theoretically and combinatorially shown to lead to the same result.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- Charalambides, Ch. A. (2002) *Enumerative Combinatorics*, Chapman & Hall/CRC, Boca Raton, Fl.
 David, F. N. (1962) *Games, God, and Gambling*, Hafner Press, New York.
 Montmort, P. R. (1713) *Essai d'Analyse sur les Jeux de Hazards*, Paris.
 Todhunter, I. (1865) *A History of the Mathematical Theory of Probability*, MacMillan, Cambridge and London. (Reprinted Chelsea, New York, 1949)



Ο ΜΙΚΡΟΣ ΘΑΛΗΣ ΚΑΙ ΤΑ ΣΤΟΧΑΣΤΙΚΑ ΜΑΘΗΜΑΤΙΚΑ*

Χαριτίδου Ολυμπία

Μαθηματικός στο 2^ο Γυμνάσιο Κιλκίς

olympia.cha@gmail.com

ΠΕΡΙΛΗΨΗ

Τον Σεπτέμβρη του 2011 το 2ο Γυμνάσιο Κιλκίς συμμετείχε σε μια πιλοτική εφαρμογή του Υπουργείου Παιδείας σε ένα νέο Αναλυτικό Πρόγραμμα σε συνεργασία με δύο σχολικούς συμβούλους.

Το Νέο Αναλυτικό Πρόγραμμα (ΝΑΠ) εισήγαγε ορισμένες καινοτομίες στην ύλη των Μαθηματικών και μία από αυτές ήταν και τα Στοχαστικά Μαθηματικά. Βασικός σκοπός της διδασκαλίας τους είναι να αναπτύξει την ικανότητα του μαθητή -μελλοντικού πολίτη- να αξιολογεί, να κρίνει, να εξάγει συμπεράσματα και να κάνει προβλέψεις. Το ΝΑΠ έδινε στον εκπαιδευτικό την ελευθερία να είναι δημιουργικός και να εφαρμόσει διδακτικές δραστηριότητες.

Έτσι και «Ο Μικρός Θαλής» μια βιωματική δραστηριότητα εφαρμόστηκε στην πιλοτική τάξη της Γ' Γυμνασίου τον Μάρτη του 2012. Οι μαθητές είχαν την ευκαιρία να βγουν έξω από τους τέσσερις τοίχους της τάξης τους και μέσα από μια καλά οργανωμένη δραστηριότητα με κατάλληλο φύλλο εργασίας κατάφεραν, να συνεργαστούν, να οργανώσουν δεδομένα, να διατυπώσουν υποθέσεις και ερωτήματα, να συνδυάσουν Γεωμετρία με Στατιστική και τέλος να εξάγουν συμπεράσματα.

Το πρόβλημα του Μικρού Θαλή-μαθητή ήταν να μετρήσει το -απρόσιτο γι' αυτόν- ύψος ενός δέντρου στην αυλή του σχολείου του έχοντας ως εργαλεία του τη γεωμετρία και τη στατιστική.

Ποια προβλήματα αντιμετώπισαν οι μαθητές; Τι ήταν αυτό που τους γοήτευσε; Πόσο εύκολη ήταν η συνεργασία μεταξύ τους; Κατάφεραν να βρουν το πραγματικό ύψος του δέντρου της αυλής του σχολείου τους;

Λέξεις Κλειδιά: βιωματική διδασκαλία, δραστηριότητα, Γεωμετρία, Στατιστική, ομαδοσυνεργατική διδασκαλία, πιλοτικό σχολείο.

*Η παρούσα εργασία παρουσιάστηκε στην Ειδική Συνεδρία «Η Στατιστική στη Δευτεροβάθμια Εκπαίδευση» που διοργάνωσε η ΕΜΕ Ημαθίας στα πλαίσια του Συνεδρίου.

1. ΕΙΣΑΓΩΓΗ

Παρακάτω παρουσιάζεται η περιγραφή μιας δραστηριότητας σε σχολική τάξη της Γ' Γυμνασίου στο πλαίσιο πιλοτικής εφαρμογής «ΝΕΟ ΣΧΟΛΕΙΟ (Σχολείο 21ου αιώνα) του Υπουργείου Παιδείας. Τη συγκεκριμένη δραστηριότητα την παρακολούθησαν δύο Σχολικοί Σύμβουλοι ο κ. Ιωάννης Θωμαΐδης και η κ. Σταματία Σταφυλίδου και ο Διευθυντής του σχολείου Ευάγγελος Αραμπατζής. Στόχος της παρούσας εργασίας ήταν να αξιολογηθεί από τους συμβούλους

- ως προς το σχεδιασμό της η βιωματική διδασκαλία και κατά πόσο αξιοποιούνται δραστηριότητες και διδακτικά εργαλεία που προτείνονταν από το (ΝΑΠ)
- ο βαθμός συμμετοχής των μαθητών στη συγκεκριμένη δραστηριότητα
- ο σχεδιασμός της διδασκαλίας και η σύνδεση της Στατιστικής με τη Γεωμετρία
- τα αποτελέσματα που μπορεί να είχε στη μαθηματική ανάπτυξη των μαθητών.

Η συγκεκριμένη εργασία έχει δημοσιευτεί στα πρακτικά της 5^{ης} Διεθνούς Μαθηματικής Εβδομάδας 2013 στη Θεσσαλονίκη με τίτλο *"Η αναγκαιότητα της δραστηριότητας & ο ρόλος του μικροπειράματος στην κατανόηση μιας μαθηματικής έννοιας - Παραδείγματα διδασκαλίας με βάση το Αναλυτικό Πρόγραμμα των Πιλοτικών Σχολείων"*

2. ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΔΡΑΣΤΗΡΙΟΤΗΤΑΣ

Τον Σεπτέμβρη του 2011 το 2ο Γυμνάσιο Κιλκίς συμμετείχε σε μια πιλοτική εφαρμογή του Υπουργείου Παιδείας σε ένα νέο Πρόγραμμα Σπουδών (ΠΣ) σε συνεργασία με δύο σχολικούς συμβούλους τον κ. Ιωάννη Θωμαΐδη και την κ. Σταματία Σταφυλίδου.

Στόχος της πιλοτικής εφαρμογής του νέου ΠΣ ήταν η μελέτη της εφαρμογής του στη διδακτική πράξη. Ειδικότερα, μέσα από την πιλοτική εφαρμογή επιδιωκόταν:

α) Να μελετηθεί η καταλληλότητα του ΠΣ και του οδηγού του εκπαιδευτικού και τα πιθανά προβλήματα που προκύπτουν κατά την εφαρμογή του.

β) Να εντοπιστούν αλλαγές οι οποίες απαιτούνται ώστε να γίνουν οι αναγκαίες τροποποιήσεις πριν την εφαρμογή του σε όλα τα σχολεία.

γ) Να προσδιοριστούν πλήρως τα μεταβατικά προβλήματα που θα προκύψουν στην πρώτη χρονιά της εφαρμογής του και να μελετηθεί ο τρόπος αντιμετώπισης τους.

δ) Να παραχθεί πρόσθετο αναγκαίο διδακτικό υλικό σε ψηφιακή μορφή.

Για την προσέγγιση των παραπάνω καθοριστικός ήταν ο ρόλος των εκπαιδευτικών των πιλοτικών σχολείων.

Έτσι λοιπόν το νέο ΠΣ εισήγαγε ορισμένες καινοτομίες στην ύλη των Μαθηματικών και μία από αυτές ήταν και τα Στοχαστικά Μαθηματικά. Βασικός σκοπός της διδασκαλίας τους είναι να αναπτύξει την ικανότητα του μαθητή - μελλοντικού πολίτη- να αξιολογεί, να κρίνει, να εξάγει συμπεράσματα και να κάνει προβλέψεις. Το νέο ΠΣ του 2011 σε σχέση με το παλιό ΠΣ πριν το 2011 έδινε στον εκπαιδευτικό την ελευθερία να είναι δημιουργικός και να εφαρμόσει διδακτικές δραστηριότητες.

Πολλές από τις έννοιες και τις δεξιότητες που αναπτύσσονται στη Στατιστική συνδέονται με άλλες περιοχές περιεχομένου, όπως η Άλγεβρα, η Γεωμετρία, η μέτρηση και ο αριθμός. Τα Μαθηματικά και η Στατιστική είναι ανθρώπινες δημιουργίες. Έχουν διακεκριμένες, αλλά συνδεδεμένες μεθόδους έρευνας.

Μεγάλη αξία έχει μία δραστηριότητα όταν γίνεται με βιωματική διδασκαλία. Η θεωρία των όμοιων τριγώνων διδάχτηκε στους μαθητές με το εξής σενάριο :

Μια φορά ο Μικρός Θαλής στην προσπάθεια του να μετρήσει το ύψος του δέντρου στην αυλή του δοκίμασε πολλές μεθόδους. Η καλύτερη όμως από όλες ήταν αυτή που βασίστηκε στην παρατήρησή του ότι ένας πάσσαλος στην αυλή και το δέντρο δημιουργούσαν ανάλογες σκιές στο έδαφος. Έτσι το απρόσιτο ύψος του δέντρου το μετέφερε σε ανθρώπινα μέτρα.

Η παραπάνω δραστηριότητα υπάρχει ως άσκηση μέσα στο βιβλίο της Γ' Γυμνασίου. Όταν παρουσιάζεται στο πλαίσιο μιας δραστηριότητας ως σενάριο παίρνει διαφορετική διάσταση στα μάτια των παιδιών και τα μαθηματικά γίνονται πιο ενδιαφέροντα. Μετά την διδασκαλία των ομοίων τριγώνων οργάνωσα την συγκεκριμένη δραστηριότητα στην τάξη παρακινώντας τους μαθητές της Γ' Γυμνασίου (Μάρτη του 2012) να ακολουθήσουν τα βήματα του Μικρού Θαλή.

Δόθηκε στους μαθητές ένα πρώτο φύλλο εργασίας με την ονομασία «Ο Μικρός Θαλής» .

ΦΥΛΛΟ ΕΡΓΑΣΙΑΣ

ΟΝΟΜΑ ΟΜΑΔΑΣ

« Ο Μικρός Θαλής»

**ΘΕΜΑ: ΜΕΤΡΗΣΗ ΤΟΥ
ΥΨΟΥΣ ΕΝΟΣ ΔΕΝΤΡΟΥ**

ΟΝΟΜΑΤΕΠΩΝΥΜΟ

- Τι εργαλεία χρησιμοποιήσατε;
- Τι μοντέλο χρησιμοποιήσατε για την μέτρηση του δέντρου;
- Συμπληρώστε τον παρακάτω πίνακα:

Ύψος μοντέλου (σε εκατοστά/cm)	
Μήκος σκιάς μοντέλου (σε εκατοστά/cm)	

- Αποτύπωση στο χαρτί και χάραξη των τριγώνων :



- Σύγκριση όμοιων τριγώνων:
- Αναλογίες που προκύπτουν από τα όμοια τρίγωνα:
- Εύρεση του ύψους του δέντρου: 
- Σύγκριση αποτελεσμάτων με τις άλλες ομάδες:
- Τι δυσκολίες αντιμετώπισε η ομάδα σου κατά την γνώμη σου;

Κάθε μέλος της ομάδας είχε μία φωτοτυπία του φύλλου εργασίας, κάθε ομάδα είχε το δικό της όνομα, είχε τα δικά της εργαλεία όπως μετροταινία, γαλλικό μέτρο, μεζούρα, κτλ, επέλεξε το δικό της μοντέλο όπως π.χ. ένα πάσσαλο, έναν συμμαθητή τους, μία σκούπα.

Εικόνα 1. 2^ο Γυμνάσιο Κιλκίς



Βγήκαν στην αυλή του σχολείου και αφού έκαναν τις μετρήσεις τους τις κατέγραψαν στον πρώτο πίνακα και στην συνέχεια έκαναν την αποτύπωση στο χαρτί. Η συνεργασία ήταν άψογη, τα παιδιά συνεργάσιμα μεταξύ τους και για πρώτη φορά το μάθημα ξέφυγε από τους τέσσερεις τοίχους της τάξης.

Εικόνα 2. Στην αυλή του 2^{ου} Γυμνασίου Κιλκίς



Απάντησαν σε όλα τα ερωτήματα και το μάθημα συνεχίστηκε στην αίθουσα για να συγκρίνουν οι ομάδες τα αποτελέσματά τους. Την διαδικασία αυτή παρακολούθησαν εκείνη την ημέρα οι Σχολικοί Σύμβουλοι Μαθηματικών που είχαν την ευθύνη παρακολούθησης του Πιλοτικού Προγράμματος στο 2^ο Γυμνάσιο Κιλκίς. Στο τέλος υπήρχε και μία ερώτηση αυτοαξιολόγησης η οποία κατά την άποψη μου κρίθηκε απαραίτητη.

Εικόνα 3. Ομαδοσυνεργατική διδασκαλία



Τα παιδιά κατέγραψαν τις δυσκολίες τους, μερικές από τις οποίες ήταν η δυσκολία στην μέτρηση της σκιάς του δέντρου, η έλλειψη ακρίβειας εξαιτίας του εργαλείου μέτρησης που είχαν επιλέξει. Σε όλη την διαδικασία ο δικός μου ρόλος ήταν αποκλειστικά αυτός του συντονιστή.

Μετά την ανακοίνωση των αποτελεσμάτων (που διέφεραν μεταξύ τους) μέσα στην τάξη οι μαθητές εξέφρασαν μία εύλογη απορία:

«Ποιο είναι το πραγματικό ύψος του δέντρου;».

Το στοιχείο αυτό αξιοποιήθηκε σε μια επόμενη δραστηριότητα που αφορούσε τη διδασκαλία βασικών εννοιών της στατιστικής όπως διάμεσος, μέση τιμή και τυπική απόκλιση.

Η θεματική ενότητα της Στατιστικής αναπτύσσεται σε τρεις τροχιές βάση του ΠΣ του 2011: δεδομένα, μέτρα θέσης και μεταβλητότητα. Η πρώτη τροχιά «Δεδομένα» αναφέρεται στη συλλογή, οργάνωση, αναπαράσταση και ερμηνεία διαφορετικού είδους δεδομένων και εξελίσσεται για να συμπεριλάβει: α) κατηγορικά δηλαδή ποιοτικά δεδομένα, ποσοτικά διακριτά δεδομένα, ποσοτικά συνεχή δεδομένα. Οι τρόποι αναπαράστασης των δεδομένων που χρησιμοποιούν οι μαθητές εξελίσσονται από τους πιο απλούς σε πιο σύνθετους: διαγράμματα με υλικά, εικονογράμματα, ραβδογράμματα, σημειογράμματα. Η δεύτερη τροχιά «Μέτρα θέσης» αφορά στη χρήση αριθμητικών εκφράσεων, οι οποίες επιτρέπουν τη συνοπτική περιγραφή δεδομένων και τη σύγκριση ομάδων δεδομένων π.χ. μέση τιμή (arithmetic mean or average), η διάμεσος (median) και η κορυφή ή επικρατούσα τιμή (mode). Η τρίτη τροχιά «Μεταβλητότητα» αναφέρεται σε αριθμητικές εκφράσεις που χαρακτηρίζουν τη διασπορά των δεδομένων π.χ. το εύρος, η μέση απόλυτη απόκλιση.

Η δραστηριότητα αυτή περιγράφεται στο παρακάτω Φύλλο Εργασίας

ΦΥΛΛΟ ΕΡΓΑΣΙΑΣ - Ποιο είναι το πραγματικό ύψος;

Παρακάτω σας δίνονται οι μετρήσεις του ύψους ενός δέντρου από τις 6 ομάδες του μμήματος Γ2 και τις 6 ομάδες του Γ3 :

ΟΜΑΔΕΣ Γ2	ΜΕΤΡΗΣΕΙΣ ΔΕΝΤΡΟΥ
ΠΟΛΥΩΝΥΜΟΙ	617
ΧΙΑΣΤΙΔΕΣ	656
UNDEFEATED	675
ΤΕΤΡΑΓΩΝΟ	675
MATHCRASH	713
(ΑΣΤΕΡΙΑ)2	1151

ΟΜΑΔΕΣ	ΜΕΤΡΗΣΕΙΣ ΔΕΝΤΡΟΥ	
ΠΟΛΥΩΝΥΜΟΙ	617	
ΧΙΑΣΤΙΔΕΣ	656	
UNDEFEATED	675	
TETΡΑΓΩΝΟ	675	
MATHCRASH	713	
(ΑΣΤΕΡΙΑ) ²	1151	

***Πίνακας 3.** Μετρήσεις του τμήματος Γ2*

ΟΜΑΔΕΣ	ΜΕΤΡΗΣΕΙΣ ΔΕΝΤΡΟΥ	
ΣΥΝΗΜΙΤΟΝΕΣ	406	
DARK NUMBERS	596	
ΤΕΡΑΓΩΝΙΚΗ PIZA	653	
(ΚΟΡΙΤΣΙΑ) ²	653	
LOGARITHM	706	
ΠΥΘΑΓΟΡΕΙΟΙ	808	

***Πίνακας 4.** Μετρήσεις του τμήματος Γ3*

3. ΣΥΜΠΕΡΑΣΜΑ

Μια συνθετική εργασία η οποία ξεκίνησε με ένα απλό σενάριο, εξελίχθηκε σε βιωματικό μάθημα στην αυλή του σχολείου και κατέληξε σε έρευνα με την βοήθεια της στατιστικής. Οι μαθητές όχι μόνο χρησιμοποίησαν τα μαθηματικά σε πραγματικές συνθήκες, αλλά κατανόησαν την έννοια της μέτρησης και την χρησιμότητα της μέσω στατιστικών εργαλείων.

Ευχαριστίες: Κλείνοντας να εκφράσω τις βαθειά θερμές μου ευχαριστίες στους Συμβούλους Θωμαΐδη Ιωάννη και Σταφυλίδου Σταματία για την πολύτιμη και ουσιαστική βοήθειά τους και τον πραγματικά συμβουλευτικό τους ρόλο καθώς επίσης και τον Διευθυντή μου Αραμπατζή Ευάγγελο που με ενθάρρυνε, με διευκόλυνε και στήριξε κάθε μου προσπάθεια.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Σχολικό Βιβλίο ΓΤγμνασίου

<http://digitalschool.minedu.gov.gr/modules/ebook/show.php/DSGYM-C104/68/542,2184/>

Οδηγός εκπαιδευτικού

<http://digitalschool.minedu.gov.gr/info/newps/%CE%9C%CE%B1%CE%B8%CE%B7%CE%BC%CE%B1%CF%84%CE%B9%CE%BA%CE%AC/%CE%9C%CE%B1%CE%B8%CE%B7%CE%BC%CE%B1%CF%84%CE%B9%CE%BA%CE%AC%20%E2%80%94%20%CE%9F%CE%B4%CE%B7%CE%B3%CF%8C%CF%82%20%CE%93%CF%85%CE%BC%CE%BD%CE%B1%CF%83%CE%AF%CE%BF%CF%85.pdf>

Πρόγραμμα Σπουδών για τα Μαθηματικά στην Υποχρεωτική Εκπαίδευση

<http://digitalschool.minedu.gov.gr/info/newps/%CE%9C%CE%B1%CE%B8%CE%B7%CE%BC%CE%B1%CF%84%CE%B9%CE%BA%CE%AC/%CE%9C%CE%B1%CE%B8%CE%B7%CE%BC%CE%B1%CF%84%CE%B9%CE%BA%CE%AC%20%E2%80%94%20%CE%94%CE%B7%CE%BC%CE%BF%CF%84%CE%B9%CE%BA%CF%8C.pdf>

ΤΙ ΕΙΝΑΙ ΚΑΙ ΤΙ ΕΚΤΙΜΑ ΤΟ ΣΥΣΤΗΜΑ ECTS; ΑΝΑΛΥΣΗ ΤΟΥ ΦΟΡΤΟΥ ΕΡΓΑΣΙΑΣ ΚΑΙ ΤΗΣ ΕΠΙΔΟΣΗΣ ΤΩΝ ΦΟΙΤΗΤΩΝ ΤΟΥ ΕΤΟΥΣ 2014- 2015 ΤΟΥ ΤΜΗΜΑΤΟΣ ΠΟΛΙΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ Α.Π.Θ.

Θ. Χατζηπαντελής¹, Μ. Σωτήρογλου², Ε. Γκουραμάνη³

¹Καθηγητής, ²Υποψηφία Διδάκτωρ, ³Προπτυχιακή Φοιτήτρια

Τμήμα Πολιτικών Επιστημών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

¹chadjji@polsci.auth.gr, ²misotiro@polsci.auth.gr, ³egkouram@polsci.auth.gr

ΠΕΡΙΛΗΨΗ

Η παρούσα εργασία πραγματεύεται την ανάλυση του φόρτου εργασίας και της επίδοσης των φοιτητών του τμήματος Πολιτικών Επιστημών Α.Π.Θ. για το ακαδημαϊκό έτος 2014-2015. Συγκεντρώθηκε, σε πιλοτική εφαρμογή, ημερολογιακό αρχείο των ωρών απασχόλησης δείγματος φοιτητών σχετικά με τις σπουδές τους ανά εξάμηνο για το σύνολο των μαθημάτων του τμήματος. Η καταγραφή έγινε με σκοπό να εκτιμηθεί ο φόρτος εργασίας από τους ίδιους τους φοιτητές, ώστε να υπολογιστεί ο πραγματικός φόρτος εργασίας που αντιστοιχεί στις πιστωτικές μονάδες ECTS του κάθε μαθήματος.

Αναλυτικότερα, οι φοιτητές του τμήματος κατέγραφαν σε ένα τυποποιημένο φύλλο excel, τις ώρες που αφιέρωναν στις ακαδημαϊκές τους υποχρεώσεις κατά τη διάρκεια κάθε εξαμήνου. Καταγράφηκε η ημερομηνία της δραστηριότητας, ο κωδικός του μαθήματος καθώς και το είδος και οι ώρες που αφιέρωσαν για κάθε δραστηριότητα. Τα δεδομένα που έχουν προκύψει αναλύονται με τη μέθοδο της ανάλυσης διασποράς ώστε να διαπιστωθεί εάν υπάρχει αντιστοιχία των ωρών απασχόλησης με τις πιστωτικές μονάδες (ECTS) που έχουν οριστεί από το τμήμα για την επιτυχή ολοκλήρωση του μαθήματος. Επιπλέον, κάνοντας χρήση της βαθμολογίας σε κάθε μάθημα, αξιολογείται κατά πόσο η απασχόληση των φοιτητών με το μάθημα συσχετίζεται με την επιτυχή πορεία τους στις εξετάσεις.

Από την παρούσα μελέτη προκύπτει ένα ιδιαίτερα χρήσιμο εργαλείο για κάθε ακαδημαϊκή μονάδα, προκειμένου να εναρμονιστεί με τις απαιτήσεις του νόμου για την αντικειμενική αναθεώρηση του προγράμματος σπουδών της.

Λέξεις Κλειδιά: ECTS, ANOVA, ακαδημαϊκά στατιστικά.

1. ΕΙΣΑΓΩΓΗ

Το Ευρωπαϊκό Σύστημα Μονάδων Κατοχύρωσης Μαθημάτων (ECTS) και η βαθμολογική κατάταξη που αυτό περιλαμβάνει είναι ένα εργαλείο για τον σχεδιασμό, την περιγραφή και την εφαρμογή προγραμμάτων σπουδών, καθώς και τη χορήγηση τίτλων σπουδών τριτοβάθμιας εκπαίδευσης που στοχεύει στη συμβατότητα των εθνικών συστημάτων. Τα προγράμματα σπουδών και οι περιγραφές των μαθημάτων περιέχουν τους μαθησιακούς σκοπούς/στόχους καθώς και το φόρτο εργασίας που απαιτείται για την επιτυχή ολοκλήρωση του μαθήματος, που αποτυπώνεται σε μονάδες ECTS. Σύμφωνα με τον γενικό κανονισμό 1 ECTS αντιστοιχεί σε 25 ώρες εξαμηνιαίας απασχόλησης.

Σύμφωνα με το ισχύον Πρόγραμμα Σπουδών του Τμήματος Πολιτικών Επιστημών, κάθε εξαμηνιαίο υποχρεωτικό μάθημα αντιστοιχεί σε 5 πιστωτικές μονάδες του ECTS και τα επιλεγόμενα σε 4 πιστωτικές μονάδες. Για την απόκτηση του πτυχίου απαιτούνται τουλάχιστον 240 μονάδες ECTS που αντιστοιχούν σε πενήντα (51) μαθήματα συνολικά, εκ των οποίων τα τριάντα εννέα (39) είναι υποχρεωτικά και τα υπόλοιπα δώδεκα (12) επιλεγόμενα (επιλογής, σεμιναριακά, ελεύθερης επιλογής). Ειδικότερα, ο φόρτος εργασίας για κάθε υποχρεωτικό μάθημα αντιστοιχεί σε 5 ECTS δηλαδή σε 125 ώρες απασχόλησης, εκτός από ένα μάθημα «Εισαγωγή στην Επιστημονική Εργασία» στο οποίο ο φόρτος εργασίας αντιστοιχεί σε 2 ECTS. Αντίστοιχα, στα μαθήματα επιλογής ο φόρτος εργασίας αντιστοιχεί σε 4 ECTS ο οποίος υπολογίζεται σε 100 ώρες απασχόλησης.

2. ΒΑΣΗ ΔΕΔΟΜΕΝΩΝ

Η ανάλυση του φόρτου εργασίας και της επίδοσης των φοιτητών του τμήματος στηρίχθηκε σε τρεις βάσεις δεδομένων: στο ημερολογιακό αρχείο, στο βαθμολόγιο των φοιτητών και στην έρευνα κοινής γνώμης φοιτητών του τμήματος. Στο σύνολο τους πίνακες (1-12) και γραφήματα (1-6) είναι αποτέλεσμα επεξεργασίας δεδομένων.

2.1 Ημερολογιακό αρχείο

Οι φοιτητές του τμήματος κατέγραφαν σε καθημερινή βάση τις ώρες που αφιέρωναν στις ακαδημαϊκές τους υποχρεώσεις για το εαρινό και χειμερινό εξάμηνο. Η καταγραφή γινόταν σε ένα τυποποιημένο φύλλο εργασίας excel. Καταγράφηκε η ημερομηνία της συγκεκριμένης δραστηριότητας, ο κωδικός του μαθήματος, το είδος και οι ώρες που αφιέρωσαν για τη συγκεκριμένη δραστηριότητα. Όσον αφορά το κωδικό του μαθήματος, αυτός αντιστοιχεί στον κωδικό του μαθήματος στο πρόγραμμα σπουδών του συγκεκριμένου προπτυχιακού κύκλου σπουδών. Στο είδος της απασχόλησης η κωδικοποίηση έχει ως εξής: «1: διαλέξεις/σεμινάρια», «2: εργαστηριακή άσκηση», «3: άσκηση πεδίου», «4: μελέτη και ανάλυση της βιβλιογραφίας», «5: φροντιστήριο», «6: πρακτική», «7: διαδραστική διδασκαλία», «8: εκπαιδευτικές επισκέψεις», «9: εκπόνηση project» και «10: συγγραφή εργασίας».

Από το σύνολο των φοιτητών του τμήματος που συμμετείχε έστω και σε μία εξέταση μαθήματος το 2014-2015, το 26% κατέγραψε ημερολόγιο στη διάρκεια του ακαδημαϊκού έτους 2014-2015. Στην πλειονότητα τους πρόκειται για φοιτητές που παρακολούθησαν τα μαθήματα «Ποσοτικές Μέθοδοι Ανάλυσης στις Κοινωνικές Επιστήμες» και «Κοινωνική Στατιστική». Από το ηλεκτρονικό αρχείο της γραμματείας, συνολικά 890 βαθμολογήθηκαν σε 9.918 μαθήματα τουλάχιστον σε μια από τις τρεις εξεταστικές περιόδους. Από αυτά οι φοιτητές πέρασαν το 80,7% σε κάποια από τις τρεις περιόδους. Από τους εγγεγραμμένους φοιτητές του τμήματος, το ποσοστό αποτυχίας κυμάνθηκε στο 23,6% τον Ιανουάριο, 22,6% τον Ιούνιο και 30,9% τον Σεπτέμβριο.

Στον παρακάτω πίνακα, μετά από επεξεργασία, παρουσιάζεται η κατανομή του χρόνου του μέσου φοιτητή ανά δραστηριότητα.

Πίνακας 1. Κατανομή χρόνου ανά Δραστηριότητα

Κωδικός	Δραστηριότητα	Ποσοστό %
Δ1	διαλέξεις/σεμινάρια	60,7
Δ2	εργαστηριακή άσκηση	3,4
Δ3	άσκηση πεδίου	1,2
Δ4	μελέτη και ανάλυση της βιβλιογραφίας	22,0
Δ5	φροντιστήριο	1,0
Δ6	πρακτική	2,0
Δ7	διαδραστική διδασκαλία	1,0
Δ8	εκπαιδευτικές επισκέψεις	0,2
Δ9	εκπόνηση project	3,0
Δ10	συγγραφή εργασίας	6,0

Ο πίνακας 1 απεικονίζει την κατανομή του χρόνου του μέσου φοιτητή ανά δραστηριότητα στην διάρκεια του ακαδημαϊκού έτους 2014-2015. Σύμφωνα με τον παραπάνω πίνακα, 60,7% του χρόνου του αφιερώθηκε στην παρακολούθηση διαλέξεων και σεμιναρίων. Ακολούθως, 22,0% του χρόνου του αφιερώθηκε στην μελέτη και ανάλυση βιβλιογραφίας και 6,0% στη συγγραφή εργασίας.

2.2 Βαθμολόγιο Φοιτητών

Αξιοποίηση του ηλεκτρονικού αρχείου βαθμολογίας των φοιτητών του τμήματος για το ακαδημαϊκό έτος 2014-2015.

2.3 Έρευνα Κοινής Γνώμης

Το Εργαστήριο Εφαρμοσμένης Πολιτικής Έρευνας του Τμήματος Πολιτικών Επιστημών, διεξάγει τα τελευταία χρόνια έρευνα κοινής γνώμης σε φοιτητές του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης. Η δειγματοληπτική έρευνα επιδιώκει

την καταγραφή αντιλήψεων και στάσεων των φοιτητών για τις σπουδές τους στο πανεπιστήμιο, προσδοκίες και προοπτικές καθώς και επιμέρους θέματα λειτουργίας των τμημάτων και του ιδρύματος. Δίνεται έτσι η δυνατότητα να αποτυπωθούν προβλήματα και ελλείψεις τόσο στο επίπεδο της ακαδημαϊκής μονάδας όσο και του ιδρύματος όπως επίσης η δημιουργία μιας βάσης δεδομένων και συγκριτικών στοιχείων ώστε να προγραμματιστούν παρεμβάσεις.

Η έρευνα διενεργείται το εαρινό εξάμηνο στα πλαίσια του μαθήματος «Ποσοτικές Μέθοδοι Ανάλυσης στις Κοινωνικές Επιστήμες» με τη μέθοδο της «πρόσωπο με πρόσωπο» συνέντευξης σε ένα τυχαίο δείγμα φοιτητών. Το δείγμα ανά τμήμα προκύπτει με τυχαία στρωματοποιημένη δειγματοληψία με κριτήριο το διανυόμενο εξάμηνο σπουδών των φοιτητών και το φύλο ανά τμήμα. Η παρούσα εργασία ασχολείται με το Τμήμα Πολιτικών Επιστημών και ειδικότερα με την ανάλυση των ωρών απασχόλησης των φοιτητών του τμήματος σε διάφορες δραστηριότητες όπως καταγράφηκαν από την έρευνα.

3. ΑΠΟΤΕΛΕΣΜΑΤΑ

Στην παρούσα εργασία πραγματοποιήθηκε Ανάλυση Διασποράς (ANOVA) των ωρών που απασχολήθηκαν οι φοιτητές με ανεξάρτητες μεταβλητές το εξάμηνο [semes (χειμερινό, εαρινό)], το είδος μαθήματος [CATEGO (υποχρεωτικό, επιλογής)] και την επιτυχία στις εξετάσεις [success (ναι, όχι)]. Στη συνέχεια, αναλύθηκαν τα είδη δραστηριοτήτων και το ποσοστό επιτυχίας και αποτυχίας στα μαθήματα Α' και Β' έτους. Τέλος, αναλύθηκε η εβδομαδιαία απασχόληση των φοιτητών του τμήματος σε σχέση με την παρακολούθηση μαθημάτων και την επιπλέον απασχόληση τους με τις σπουδές πέραν των ακαδημαϊκών τους δραστηριοτήτων όπως καταγράφεται στην έρευνα κοινής γνώμης.

Πίνακας 2. Συσχέτιση Μεταβλητών

Tests of Between-Subjects Effects

Dependent Variable: ECTS

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	204166,361 ^a	7	29166,623	42,028	,000
Intercept	960998,840	1	960998,840	1384,754	,000
CATEGO	27390,660	1	27390,660	39,469	,000
semes	10230,719	1	10230,719	14,742	,000
success	79074,944	1	79074,944	113,943	,000
CATEGO * semes	12,076	1	12,076	,017	,895
CATEGO * success	7,546	1	7,546	,011	,917
semes * success	11102,322	1	11102,322	15,998	,000
CATEGO * semes * success	28,327	1	28,327	,041	,840
Error	1764804,192	2543	693,985		
Total	4508744,000	2551			
Corrected Total	1968970,553	2550			

a. R Squared = ,104 (Adjusted R Squared = ,101)

Από τον πίνακα 2 φαίνεται ότι είναι σημαντικές οι κύριες επιδράσεις των ανεξάρτητων μεταβλητών και η αλληλεπίδραση εξαμήνου και επιτυχίας.

Γενικότερα, παρατηρείται ότι όσο αυξάνονται οι ώρες απασχόλησης τόσο αυξάνεται το ποσοστό επιτυχίας στις εξετάσεις. Σύμφωνα με τον πίνακα 3, στο σύνολο του δείγματος, οι επιτυγχόντες απασχολήθηκαν κατά μέσο όρο 35,5 ώρες ανά μάθημα ενώ οι αποτυγχόντες απασχολήθηκαν κατά μέσο όρο 20,4 ώρες ανά μάθημα.

Πίνακας 3. Ώρες απασχόλησης/σύνολο τμήματος

	Σύνολο	N
Επιτυχία	35,5	1,877
Αποτυχία	20,4	674
Σύνολο	31,5	2,551

Στο σύνολο του τμήματος, το χειμερινό εξάμηνο η μέση απασχόληση καταγράφηκε στις 28,4 ώρες ανά μάθημα ενώ στο εαρινό εξάμηνο στις 23 ώρες ανά μάθημα. Ως προς το είδος μαθήματος, η μέση απασχόληση στα υποχρεωτικά μαθήματα είναι 30 ώρες ανά μάθημα και για τα μαθήματα επιλογής είναι 21,4 ώρες ανά μάθημα. Υπολογίζοντας το μέσο φόρτο απασχόλησης ανά μάθημα για το σύνολο των μαθημάτων, το μέγεθος αυτό αντιστοιχεί σε 22,65 ώρες (αθροιστικά καταγράφηκαν 81.749 ώρες για 3.609 μαθήματα).

3.1 Είδος μαθήματος & εξάμηνο

Στα υποχρεωτικά αλλά και στα επιλεγόμενα μαθήματα υπάρχει διαφοροποίηση των ωρών απασχόλησης ως προς το εξάμηνο. Αναλυτικότερα, στα υποχρεωτικά μαθήματα οι ώρες απασχόλησης για το χειμερινό εξάμηνο είναι 32,6 ώρες και για το εαρινό εξάμηνο είναι 27,5 ώρες. Ακολούθως, στα μαθήματα επιλογής οι ώρες απασχόλησης για το χειμερινό εξάμηνο είναι 24,1 ώρες ενώ για το εαρινό εξάμηνο 18,6 ώρες. Παρατηρείται ότι στο χειμερινό εξάμηνο οι φοιτητές απασχολούνται περισσότερες ώρες ανεξάρτητα από το είδος του μαθήματος.

Σύμφωνα με τον πίνακα 4, στα υποχρεωτικά μαθήματα χειμερινού εξαμήνου ο επιτυχών απασχολήθηκε 42,7 ώρες κατά μέσο όρο και ο αποτυχών απασχολήθηκε 22,5 ώρες κατά μέσο όρο. Στα υποχρεωτικά μαθήματα εαρινού εξαμήνου αντίστοιχα, ο επιτυχών απασχολήθηκε 32,3 ώρες κατά μέσο όρο και ο αποτυχών απασχολήθηκε 22,6 ώρες κατά μέσο όρο.

Πίνακας 4. Ώρες απασχόλησης/Υποχρεωτικό μάθημα & εξάμηνο

Υ	Χειμερινό εξάμηνο	N	Εαρινό Εξάμηνο	N
Επιτυχία	42,7	743	32,3	687
Αποτυχία	22,5	195	22,6	309
Σύνολο	38,5	938	29,3	996

Σύμφωνα με τον πίνακα 5, στα επιλεγόμενα μαθήματα χειμερινού εξαμήνου ο επιτυχών απασχολήθηκε 34,3 ώρες κατά μέσο όρο και ο αποτυχών απασχολήθηκε 13,9 ώρες κατά μέσο όρο. Στα επιλεγόμενα μαθήματα εαρινού εξαμήνου αντίστοιχα, ο επιτυχών απασχολήθηκε 23 ώρες κατά μέσο όρο και ο αποτυχών απασχολήθηκε 14,2 ώρες κατά μέσο όρο.

Πίνακας 5. Ώρες απασχόλησης/Επιλεγόμενο μάθημα & εξάμηνο

E	Χειμερινό εξάμηνο	N	Εαρινό Εξάμηνο	N
Επιτυχία	34,3	211	23	236
Αποτυχία	13,9	97	14,2	73
Σύνολο	27,9	308	20,9	309

Ο μέσος χρόνος απασχόλησης ανά μάθημα για όσους απέτυχαν δεν παρουσιάζει διαφορά κατά εξάμηνο όσον αφορά τα υποχρεωτικά και επίσης τα μαθήματα επιλογής. Αντίθετα, ο μέσος χρόνος απασχόλησης είναι μεγαλύτερος για το χειμερινό εξάμηνο και για τις δύο κατηγορίες μαθημάτων.

3.2 Ώρες απασχόλησης συνολικά ανά εξάμηνο/ επιτυχία και αποτυχία

Στο σύνολο του τμήματος ο μέσος όρος απασχόλησης όπως φαίνεται στον πίνακα 6, για την επιτυχία στις εξετάσεις είναι 40,8 ώρες για το χειμερινό εξάμηνο και 29,9 ώρες για το εαρινό εξάμηνο. Αντιθέτως, ο μέσος όρος απασχόλησης με αρνητική έκβαση στις εξετάσεις είναι 19,7 ώρες για το χειμερινό εξάμηνο και 21 ώρες για το εαρινό εξάμηνο.

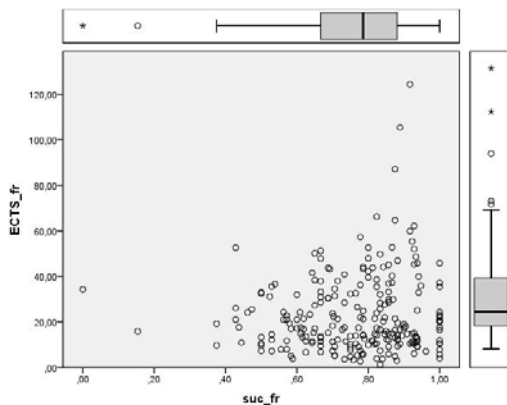
Πίνακας 6. Ώρες απασχόλησης/ εξάμηνο

	Χειμερινό εξάμηνο	N	Εαρινό Εξάμηνο	N
Επιτυχία	40,8	954	29,9	923
Αποτυχία	19,7	292	21	382
Σύνολο	35,9	1,246	27,3	1,305

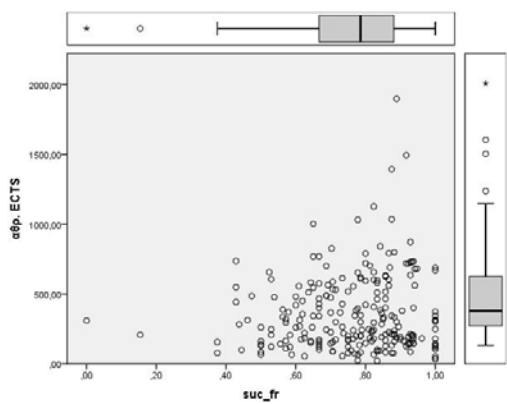
Στη συνέχεια, υπολογίστηκε ανά φοιτητή το ποσοστό των μαθημάτων στα οποία επέτυχε και ο μέσος αριθμός ωρών απασχόλησης ανά μάθημα (για το σύνολο των μαθημάτων) καθώς και το σύνολο των ωρών απασχόλησης. Στόχος ήταν η μελέτη της συσχέτισης του μέσου όρου και του συνολικού αριθμού της απασχόλησης με την επιτυχία των φοιτητών στις εξετάσεις όπως επίσης και η ανάδειξη ακραίων τιμών των μεταβλητών. Τα παρακάτω γραφήματα παρουσιάζουν το μέσο όρο των ωρών απασχόλησης ανά μάθημα (γράφημα 1), το άθροισμα των ωρών απασχόλησης ανά μάθημα (γράφημα 2), την συσχέτιση του μέσου όρου και του αθροίσματος των ωρών απασχόλησης (γράφημα 3) και την εξαμηνιαία απασχόληση του μέσου φοιτητή.

Όπως φαίνεται και από τα γραφήματα 1 & 2 οι μεταβολές των τιμών της μίας μεταβλητής δεν συνδέονται με τις μεταβολές της άλλης. Τα σημεία του νέφους είναι διασκορπισμένα χωρίς κάποια συσχέτιση.

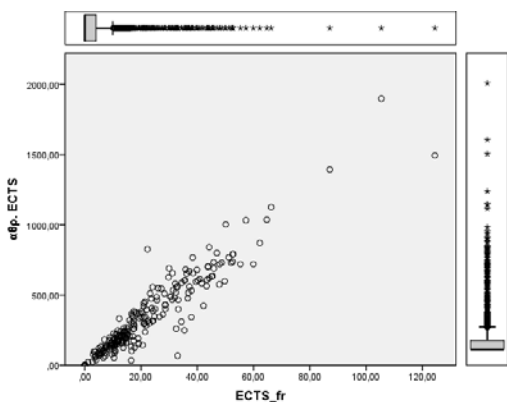
Γράφημα 1. Μέσος όρος ωρών απασχόλησης/ επιτυχία



Γράφημα 2. Άθροισμα ωρών απασχόλησης/ επιτυχία

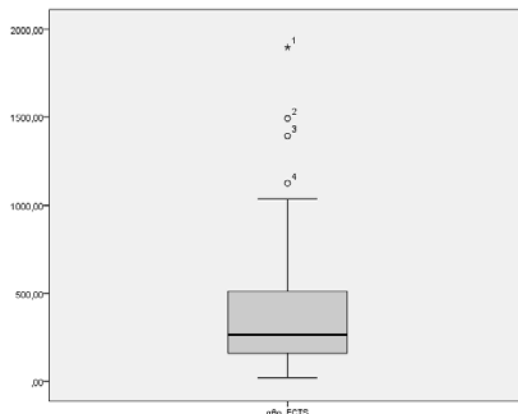


Γράφημα 3. Μέσος όρος ωρών απασχόλησης /Άθροισμα ωρών απασχόλησης



Στο γράφημα 3 φαίνεται η ισχυρή συσχέτιση του μέσου όρου απασχόλησης με τον αριθμό απασχόλησης.

Γράφημα 4. Ώρες απασχόλησης μέσου φοιτητή



Σύμφωνα με το γράφημα 4 ο μέσος φοιτητής απασχολείται με τις σπουδές του 354 ώρες δηλαδή 177 ώρες το εξάμηνο.

4. ΜΑΘΗΜΑΤΑ Α΄ ΕΤΟΥΣ

Επόμενο βήμα, η μελέτη των ωρών απασχόλησης των φοιτητών ως προς την επίδοση τους στις εξετάσεις για τα μαθήματα του Α΄ έτους σπουδών. Ο παρακάτω πίνακας (7) παρουσιάζει το μέσο όρο των ωρών απασχόλησης ανά μάθημα, την επίδοση τους στις εξετάσεις καθώς και τον αριθμό επιτυχόντων και αποτυχόντων από αυτούς που συμμετείχαν στη καταγραφή του ημερολογιακού αρχείου.

Σύμφωνα με τον παρακάτω πίνακα, το μάθημα για το οποίο εμφανίζεται ο μεγαλύτερος μέσος όρος απασχόλησης για την επιτυχία σε αυτό είναι το «Μαθηματικά στην πολιτική επιστήμη: εισαγωγή» με 45,7 ώρες απασχόλησης και ακολουθεί το «Εισαγωγή στη συγκριτική πολιτική: τα πολιτικά συστήματα» με 24,7 ώρες απασχόλησης.

Επισημαίνεται ότι ο αριθμός (N) των φοιτητών που κατέγραψαν ημερολόγιο δημιουργεί ακραίες τιμές λόγω περιορισμένης ή αυξημένης συμμετοχής των φοιτητών στα αντίστοιχα μαθήματα.

Πίνακας 7. Ώρες απασχόλησης/μαθήματα Α΄ Έτους/ Επίδοση στις εξετάσεις

		Ώρες		N	
Κωδικός	Τίτλος μαθήματος	Αποτυχία	Επιτυχία	Αποτυχία	Επιτυχία
KY0000	Εισαγωγή στην επιστημονική εργασία στις Πολιτικές Επιστήμες	9.00		3	
KY0101	Εισαγωγή στην πολιτική επιστήμη I	6,67	9,00	2	2
KY0102	Πολιτική φιλοσοφία στην αρχαία Ελλάδα		3,50	2	2

KY0103	Νεότερη ελληνική πολιτική ιστορία (1821-1909)	3,50	11,8	8	2
KY0104	Νεότερη πολιτική φιλοσοφία Ι: από τον Μακιαβέλι στις θεωρίες	3,50	1,00	2	1
KY0105	Διεθνείς σχέσεις	6,75	7,40	4	5
KY0106	Σύγχρονη ευρωπαϊκή ιστορία	19,2	16,0	7	12
KY0201	Εισαγωγή στη συγκριτική πολιτική: τα πολιτικά συστήματα	13,5	24,7	9	8
KY0203	Σύγχρονη ελληνική πολιτική ιστορία (1910-1974)	18,3	11,6	3	8
KY0204	Νεότερη πολιτική φιλοσοφία ΙΙ: 19ος αιώνας	9,2	5,6	4	3
KY0205	Συνταγματική συγκρότηση του κράτους: οργανωτικές βάσεις του	12,5	10,5	4	2
KY0206	Μαθηματικά στην πολιτική επιστήμη: εισαγωγή	12,1	45,7	42	21
KY0208	Εισαγωγή στην πολιτική επιστήμη ΙΙ	8,29	8,58	7	12

5. ΜΑΘΗΜΑΤΑ Β' ΕΤΟΥΣ

Στη συνέχεια, μελετήθηκαν οι ώρες απασχόλησης των φοιτητών ως προς την επίδοσή τους στις εξετάσεις για τα μαθήματα του Β' έτους σπουδών. Ο παρακάτω πίνακας (8) παρουσιάζει το μέσο όρο των ωρών απασχόλησης ανά μάθημα, την επίδοσή τους στις εξετάσεις καθώς και τον αριθμό επιτυχόντων και αποτυχόντων από αυτούς που συμμετείχαν στη καταγραφή του ημερολογιακού αρχείου.

Σύμφωνα με τον παρακάτω πίνακα, το μάθημα για το οποίο εμφανίζεται ο μεγαλύτερος μέσος όρος απασχόλησης για την επιτυχία σε αυτό είναι το «Πολιτικό σύστημα και κομματικοί σχηματισμοί στη σύγχρονη Ελλάδα» με 77,0 ώρες απασχόλησης, ακολουθεί το «Κοινωνική στατιστική» και το «Ποσοτικές μέθοδοι ανάλυσης στις κοινωνικές επιστήμες» με 53,2 ώρες απασχόλησης αντίστοιχα.

Πίνακας 8. Ώρες απασχόλησης/μαθήματα Β' Έτους/ Επίδοση στις εξετάσεις

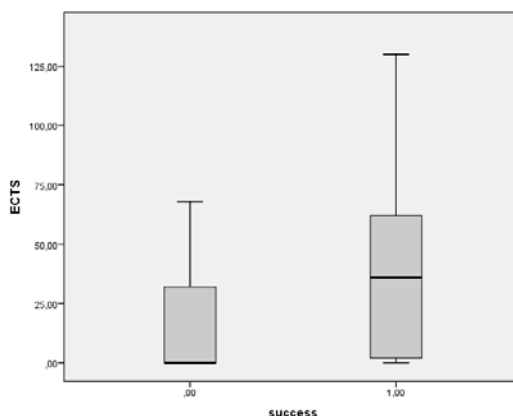
		Ώρες		N	
Κωδικός	Τίτλος μαθήματος	Αποτυχία	Επιτυχία	Αποτυχία	Επιτυχία
KY0301	Εισαγωγή στην Πολιτική Κοινωνιολογία	3,60	35,3	5	78
KY0302	Σύγχρονες πολιτικές	28,1	44,4	6	78

	θεωρίες I				
KY0303	Διαφωτισμός και διαμόρφωση των πολιτικών ιδεολογιών στην Ελλάδα	22,8	37,1	21	67
KY0304	Πολιτικό σύστημα και κομματικοί σχηματισμοί στη σύγχρονη Ελλάδα	27,4	77,0	28	66
KY0305	Κοινωνική στατιστική	18,5	53,2	4	173
KY0401	Το Ελληνικό σύστημα διακυβέρνησης	20,7	23,1	4	73
KY0402	Κλασσική κοινωνική θεωρία	28,5	33,5	34	53
KY0403	Ιστορία, θεωρίες και θεσμοί της ευρωπαϊκής ολοκλήρωσης	24,4	37,8	46	46
KY0404	Εισαγωγή στην πολιτική οικονομία	27,1		74	
KY0405	Ποσοτικές μέθοδοι ανάλυσης στις κοινωνικές επιστήμες	44,2	53,2	10	155

5.1 Το μάθημα «Ποσοτικές μέθοδοι ανάλυσης στις κοινωνικές επιστήμες»

Κρίθηκε επίσης απαραίτητη η λεπτομερέστερη ανάλυση για το μάθημα «Ποσοτικές Μέθοδοι Ανάλυσης στις Κοινωνικές Επιστήμες» διότι η ημερολογιακή καταγραφή πραγματοποιήθηκε στο πλαίσιο αυτού του μαθήματος και σχεδόν όλοι οι συμμετέχοντες κατέγραψαν ημερολόγιο.

Γράφημα 5. Επίδοση φοιτητών/ ώρες απασχόλησης για το μάθημα «Ποσοτικές Μέθοδοι Ανάλυσης στις Κοινωνικές Επιστήμες»



Το παραπάνω boxplot παρουσιάζει την διάμεσο, τις ακραίες τιμές, την ελάχιστη τιμή (min) και την μέγιστη (max) των μεταβλητών για την επίδοση των φοιτητών του μαθήματος στις εξετάσεις ως προς τις ώρες απασχόλησης. Η διάμεσος των ωρών απασχόλησης για όσους επέτυχαν είναι 30, το 75% σημείο είναι 65 ώρες απασχόλησης και το 25% σημείο είναι 5 ώρες απασχόλησης. Ωστόσο, παρατηρείται ότι το 75% των αποτυχόντων ασχολήθηκε λιγότερες ώρες από τη διάμεσο ωρών απασχόλησης των επιτυχόντων.

6. ΕΡΕΥΝΑ ΚΟΙΝΗΣ ΓΝΩΜΗΣ

Η έρευνα κοινής γνώμης πραγματοποιήθηκε σε δείγμα περίπου 2000 φοιτητών του Α.Π.Θ. Ειδικά για το τμήμα Πολιτικών Επιστημών καταγράφηκαν οι απαντήσεις 149 φοιτητών του τμήματος (ποσοστό 22,3% των ευρισκόμενων στα κανονικά εξάμηνα σπουδών σύμφωνα με τα στοιχεία της Ελληνικής Στατιστικής Αρχής για το έτος 2014-2015). Ποσοστό 40% του δείγματος ήταν άνδρες και 60% ήταν γυναίκες. Επειδή πρόκειται για μια τυχαία στρωματοποιημένη δειγματοληψία, το δειγματοληπτικό πλαίσιο διαμορφώθηκε με κριτήρια το φύλο και το εξάμηνο σπουδών. Υπολογίστηκε ο μέσος όρος των φοιτητών των κανονικών εξαμήνων και θεωρήθηκε ότι οι φοιτητές πέραν των κανονικών εξαμήνων (20^ο εξάμηνο) αντιστοιχούν σε ένα επιπλέον έτος σπουδών για λόγους αντιπροσωπευτικότητας. Το δειγματοληπτικό πλαίσιο της έρευνας παρουσιάζεται στον παρακάτω πίνακα.

Πίνακας 9. Δείγμα φοιτητών σύμφωνα με το διανυόμενο εξάμηνο και φύλο

Διανυόμενο εξάμηνο	Άνδρες	Γυναίκες
2 ^ο Εξάμηνο	5,6%	9,8%
4 ^ο Εξάμηνο	7,6%	10,0%
6 ^ο Εξάμηνο	7,4%	12,7%
8 ^ο Εξάμηνο	8%	10,3%
20 ^ο Εξάμηνο	8,6%	12,2%

Από τα στοιχεία της έρευνας αξιοποιήθηκαν οι ερωτήσεις κλειστού τύπου που αφορούν την απασχόληση των φοιτητών με τις σπουδές τους. Συγκεκριμένα, αναλύθηκε η εβδομαδιαία απασχόληση των φοιτητών του τμήματος ως προς την παρακολούθηση μαθημάτων και η επιπλέον απασχόληση με τις σπουδές πέραν των ακαδημαϊκών τους δραστηριοτήτων ώστε να διαπιστωθεί αν είναι συμβατές οι δύο καταγραφές (δηλαδή το ημερολογιακό αρχείο των φοιτητών και η δειγματοληπτική έρευνα).

Πίνακας 10. Εβδομαδιαία συχνότητα παρακολούθησης μαθημάτων ανά διανυόμενο εξάμηνο σπουδών

Ώρες Απασχόλησης	Συχνότητα	Σχετική Συχνότητα (%)	Αθροιστική συχνότητα

1-5	19	12,8	12,8
6-10	47	31,5	44,3
11-15	43	28,9	73,2
16-20	25	16,8	89,9
21-25	10	6,7	96,6
26-30	5	3,4	100,0
Σύνολο	149	100,0	

Από τον πίνακα 10 φαίνεται ότι η διάμεσος βρίσκεται στο διάστημα 11-15 ώρες εβδομαδιαίας παρακολούθησης μαθημάτων.

Πίνακας 11. Εβδομαδιαία συχνότητα επιπλέον απασχόλησης με τις σπουδές ανά διανυόμενο εξάμηνο

Ωρες Απασχόλησης	Συχνότητα	Σχετική Συχνότητα (%)	Αθροιστική Συχνότητα
1-5	64	43,0	43,0
6-10	46	30,9	73,8
11-15	24	16,1	89,9
16-20	10	6,7	96,6
21-25	2	1,3	98,0
26-30	3	2,0	100,0
Σύνολο	149	100,0	

Από τον πίνακα 11 φαίνεται ότι η διάμεσος βρίσκεται στο διάστημα 6-10 ώρες εβδομαδιαίας επιπλέον απασχόλησης με τις σπουδές.

Στον πίνακα 12 υπολογίστηκε το άθροισμα των δύο παραπάνω αναφερόμενων μεγεθών και προέκυψε ότι η διάμεσος των ωρών απασχόλησης με τις σπουδές είναι περίπου 16 ώρες την εβδομάδα.

Πίνακας 12. Μέσος όρος ανά εβδομάδα (παρακολούθηση μαθημάτων και επιπλέον απασχόληση με τις σπουδές)

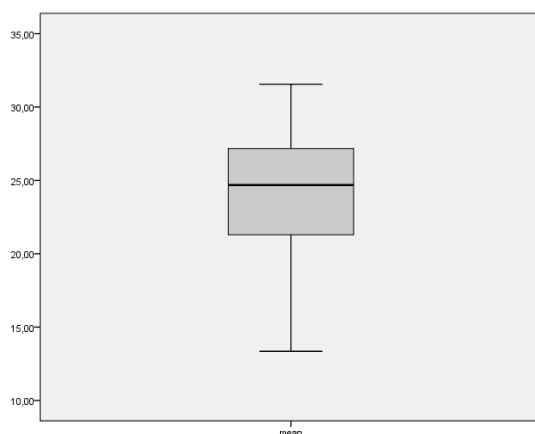
Ωρες Απασχόλησης	Συχνότητα	Σχετική Συχνότητα (%)	Αθροιστική Συχνότητα
6	14	9,3	9,4
11	23	15,3	24,8
16	40	26,7	51,7
21	25	16,7	68,5
26	19	12,7	81,2
31	14	9,3	90,6
36	10	6,7	97,3
41	1	0,7	98,0
51	1	0,7	98,7

56	2	1,3	100,0
Σύνολο	149	99,3	

Για το σύνολο του δείγματος από τα διάφορα τμήματα του Α.Π.Θ. η ανάλυση για την συχνότητα απασχόλησης ανά εβδομάδα των φοιτητών ως προς την παρακολούθηση μαθημάτων και η επιπλέον απασχόληση με τις σπουδές πέραν των ακαδημαϊκών τους δραστηριοτήτων δίνει τα εξής:

όσον αφορά την εβδομαδιαία συχνότητα παρακολούθησης μαθημάτων κατά το διανυόμενο εξάμηνο, η διάμεσος βρίσκεται στο διάστημα 11-15 ώρες εβδομαδιαίας παρακολούθησης μαθημάτων, ενώ όσον αφορά την περαιτέρω απασχόληση με τις σπουδές η διάμεσος βρίσκεται στο διάστημα 11-15 ώρες εβδομαδιαίας απασχόλησης. Βάσει του boxplot (βλ. γράφημα 6) η διάμεσος των ωρών απασχόλησης των φοιτητών του ιδρύματος με τις σπουδές τους είναι 24 ώρες, το 75% σημείο είναι 27 ώρες και το 25% σημείο είναι 21 ώρες.

Γράφημα 6. Μέσος όρος των ωρών απασχόλησης των φοιτητών του ιδρύματος



7. ΣΥΜΠΕΡΑΣΜΑΤΑ

Σύμφωνα με την ανάλυση του ημερολογίου των φοιτητών η επίδοση στις εξετάσεις (επιτυχία, αποτυχία) εξαρτάται από το εξάμηνο και το είδος μαθήματος. Παρατηρήθηκε ότι ο μέσος χρόνος απασχόλησης είναι μεγαλύτερος για το χειμερινό εξάμηνο ανεξάρτητα από το είδος του μαθήματος και ότι οι επιτυχόντες έχουν μεγαλύτερο μέσο χρόνο μελέτης από τους αποτυχόντες. Καταδείχθηκε ότι είναι σημαντικές οι κύριες επιδράσεις των ανεξάρτητων μεταβλητών (είδος μαθήματος, εξάμηνο και επιτυχία) και η αλληλεπίδραση του εξαμήνου με την επιτυχία. Επίσης δεν βρέθηκε συσχέτιση του μέσου όρου απασχόλησης και του συνολικού αριθμού ωρών απασχόλησης με την επιτυχία στις εξετάσεις.

Με βάση την ημερολογιακή καταγραφή οι ώρες απασχόλησης για υποχρεωτικό μάθημα είναι 35,5 ώρες για την επιτυχία στις εξετάσεις. Κατά τον κανονισμό ο φόρτος εργασίας για κάθε υποχρεωτικό μάθημα αντιστοιχεί σε 125 ώρες απασχόλησης. Ενώ, επίσης, ο φόρτος εργασίας σε κάθε μάθημα επιλογής αντιστοιχεί σε 100 ώρες, ο μέσος όρος απασχόλησης ανέρχεται σε 20,4 ώρες. Επιπλέον ο μέσος φοιτητής, σύμφωνα με την ημερολογιακή καταγραφή, απασχολείται 177 ώρες το εξάμηνο με το σύνολο των μαθημάτων του. Αυτό είναι συμβατό και με τα αποτελέσματα της έρευνας κοινής γνώμης όπου η διάμεσος της απασχόλησης ανά εβδομάδα είναι περίπου 16 ώρες, και κατά συνέπεια ο μέσος φοιτητής απασχολείται 160-190 ώρες με τις σπουδές του. Σύμφωνα με τον κανονισμό η μέση εβδομαδιαία απασχόληση θα έπρεπε να είναι 35 ώρες που αντιστοιχεί σε 350-420 ώρες ανά εξάμηνο.

Επομένως δεν υπάρχει αντιστοιχία μεταξύ του πραγματικού και του θεωρητικού μεγέθους όσον αφορά τις ώρες απασχόλησης που απαιτούνται για την επιτυχία στις εξετάσεις.

Από την παραπάνω ανάλυση της απασχόλησης δίνεται η δυνατότητα ελέγχου της συμβατότητας των δύο μεθόδων καταγραφής. Η έρευνα κοινής γνώμης είναι αντικειμενική εκτίμηση που επιβεβαιώνει ότι η καταγραφή μέσω φύλλου καταγραφής είναι ακριβής, καθώς ο μέσος φοιτητής απασχολείται σύμφωνα με την έρευνα με 175 ώρες το εξάμηνο και σύμφωνα με την ημερολογιακή καταγραφή απασχολείται 177 ώρες το εξάμηνο.

Τα παραπάνω αναφέρονται στο σύνολο των φοιτητών του τμήματος Πολιτικών Επιστημών. Για να προχωρήσουμε σε συγκριτική ανάλυση θα χρειαζόμασταν ανάλογες καταγραφές ημερολογίου από φοιτητές άλλων τμημάτων του Α.Π.Θ. ή/και άλλων Ιδρυμάτων. Πάντως, από τα αναλυτικά δεδομένα του φοιτητολογίου του Α.Π.Θ. γνωρίζουμε ότι ο μέσος φοιτητής κατοχυρώνει 23 μονάδες ECTS ανά εξάμηνο σπουδών (παρουσιάζει ποσοστό επιτυχίας στις εξετάσεις 76,6%). Αυτό αντιστοιχεί κατά τον ορισμό της μονάδας ECTS σε 575 ώρες απασχόλησης συνολικά ανά εξάμηνο σπουδών, ενώ απασχολείται σύμφωνα με την έρευνα κοινής γνώμης περίπου 200-240 ώρες ανά εξάμηνο σπουδών. Σημειώνουμε ότι μια αντίστοιχη καταγραφή για τα ελληνικά ΑΕΙ θα παρουσίαζε ιδιαίτερο ενδιαφέρον, ενώ δεν υπάρχουν τουλάχιστον δημοσιευμένες αντίστοιχες έρευνες για άλλα ευρωπαϊκά πανεπιστήμια.

Από την παρούσα μελέτη προκύπτει ένα ιδιαίτερα χρήσιμο εργαλείο για κάθε ακαδημαϊκή μονάδα, προκειμένου να εναρμονιστεί με τις απαιτήσεις του νόμου για την αντικειμενική αναθεώρηση του προγράμματος σπουδών της.

ABSTRACT

The department of Political Sciences, on a pilot study (project), has collected time sheet of its undergraduated students' workload courses and learning achievements for the academic year 2014-2015. The aim of this study was to evaluate students'

workload in order to estimate whether the real course workload corresponds to the European Credit Transfer and Accumulation System credits of the semesters' courses (winter and spring). More specifically, students have recorded, on a standardized excel document, the hours that have dedicated to their academic responsibilities.

They recorded the date of activity, the course code (according to the study guide), the work-based learning activity and the hours dedicated for every course (core or elective). The collected data were analyzed with ANOVA for establishing the ECTS credits correspondence to the department's credits and estimate learning achievements during semesters' exams. Furthermore, all students' grades were used to assess how learning achievements reciprocate to ECTS credits and students' success or failure at the exams.

This study emerges a useful tool for every faculty in order to comply with the ECTS, evaluate and if necessary reconsider study programmes.

Keyword: ECTS, credits, workload, ANOVA.

ΑΝΑΦΟΡΕΣ

Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης. *Ευρωπαϊκό Σύστημα Πιστωτικών Μονάδων* (<https://www.auth.gr/ects>).

Bartholomew D., Steele F., Moustaki I., Galbraith J. (επιμ.). (2007). *Ανάλυση Πολυμεταβλητών Δεδομένων για Κοινωνικές Επιστήμες*, Θεσσαλονίκη: Επίκεντρο. Ελληνική Στατιστική Αρχή, *Εκπαίδευση. Ανώτατα Εκπαιδευτικά Ιδρύματα*. (<http://www.statistics.gr/el/statistics/-/publication/SED33/->)

Ευρωπαϊκή Επιτροπή. *Εκπαίδευση και Κατάρτιση, Μονάδες ECTS* (http://ec.europa.eu/education/ects/ects_el.htm).

Lomax G. Richard. (1998). *Statistical concepts a second course for education and the behavioral sciences*, Mahwah.N.J: L. Erlbaum Associates.

Μπόρα- Σέντα Ε. (1990). *Εφαρμοσμένη στατιστική: πολλαπλή παλινδρόμηση, ανάλυση διασποράς, χρονοσειρές*, Θεσσαλονίκη: Ζήτη.

Τμήμα Πολιτικών Επιστημών, *Πρόγραμμα Προπτυχιακών Σπουδών* (<http://www.polsci.auth.gr/index.php?lang=el&rm=2&mn=185>).

Χατζηπαντελής Θ. (επιστημονικά υπεύθυνος), Παρατηρητήριο Ροής Α.Π.Θ. 1986-1996, (<http://www.polres.gr/gr/observ.htm>).

Roberts J. Maxwell & Russo R. (1999). *A student's guide to analysis of variance*, London; New York: Routledge.

εργασίες

στα αγγλικά

ESTIMATION OF SIMPLE LINEAR REGRESSION COEFFICIENTS WITH STRATIFIED DATA

Yiannis C. Bassiakos

Department of Economic Sciences,
National & Kapodistrian University of Athens
ybassiak@econ.uoa.gr

Abstract

The theoretical problem of how to estimate the regression coefficient in linear regression, when the data come from a stratified sample has been addressed in the statistical literature since the 80's. The proposed solutions are rather complicated. In the present work a comparison is made – via simulations – of some simple procedures to address this problem. Three regression models are used, with normal, uniform over $[-1,1]$ and double exponential errors, to ascertain the influence of the error distribution on the accuracy of the estimation methods. The theoretical value of the regression coefficient and the variance of error term are calculated using the distribution properties, to serve as benchmarks of the estimates. The estimation is carried through the standard procedures available in SASTM software using the weight option, and the use of bootstrap to avoid weighting. Three different weights are used: (a) the typical inverse-probability-of-selection weight, (b) the standard stratified sampling

weight $W_h = \frac{N_h}{N}$, and (c) no weight at all. In the bootstrapping process, the bootstrap samples are selected with replacement so that each sample is distributed into the strata in the same proportions as the population and the sample size remains constant. Finally, the estimation process is applied in each stratum separately and the resulting parameter estimates are combined using standard stratification weights in order to obtain the final estimates. Mean squared error and the differences between the estimated and true values for the regression coefficients are used as the main measures of quality of the estimates, along with the theoretical and estimated variances of the estimators and their differences. The results indicate that using the inverse-probability-of-selection weight produces better results and bootstrapping is a close second. The standard stratified weight lags behind – as expected from theory – since it is affected by the structure of the estimating equations which are closer to ratio estimators. Even though the independent variable X is treated as a constant (it contributes 0 to the overall variance of Y), thus one could claim that the typical OLS estimator is not a ratio but is just multiplied by a constant, apparently the algorithms are affecting the actual solution leading to variance underestimation.

Keywords: stratified sampling, weighting, linear regression, bootstrap.

1. INTRODUCTION

The classical approach to population parameter estimation is based on i.i.d. samples. Yet, in general practice many sampling schemes do not follow this assumption. Instead other sampling methods are followed, most common among them, the stratified random sampling method. In a previous work [Bassiakos 2014], we investigated the shortcomings of some approaches to remedy the situation in the estimation of common population parameters such as the population mean and the population ratio. We also partially addressed the estimation of simple linear regression coefficients.

In the present work we address the issue of estimating the simple linear regression coefficients at depth. The classical approach to parameter estimation in linear regression involves i.i.d random variables. This assumption forms the basis for the derivation of the distribution of the estimators, given the distribution of the sample.

The typical simple linear regression model is of the form, where the ε_i 's are normally distributed with $\mu=0$ and (common) variance σ^2 for $i=1 \dots n$ and $Cov(\varepsilon_i, \varepsilon_j) = 0$, for $i \neq j$. These conditions are satisfied when the data are selected through simple random sampling (SRS).

When a different sampling scheme is used – for instance stratified random sampling with SRS from each stratum – the assumptions are met within each stratum, but the common model form for the total population is not true. It is – in other words – possible to have different models in each stratum. This problem is discussed by Jewell [1985], Quesenberry & Jewell [1986] and Pringle & Rayner [1971] among others. The solutions proposed in these works are valid, but difficult to implement in the absence of specialized software. It begs the question if it could be shown that simpler procedures can yield adequate results. For instance, can a method based on a simple EXCEL worksheet be proposed? The present work examines such an approach using simulations to estimate the mean squared error of estimation of the regression coefficient in such a setting.

2. METHODOLOGY

The problem is investigated through simulations. The general population is comprised of two strata of sizes 40000 and 60000 observations and the relation between the dependent variable Y and the independent variable X is described by two linear relations (one for each stratum). For the first stratum the relation is $Y=3+5X$, and the second $Y=3+4X$. In a second step three populations are generated by adding to the equation an error term following (a) normal distribution (with mean zero and variance 4 and 16 in each stratum, (b) uniform distribution over $[-1, 1]$ for both strata and (c) double exponential distribution with $\lambda=1$ for both strata as well, to ascertain the influence of the error distribution on the accuracy of the estimation methods. Five methods of approaching the problem are investigated. The 1st method (INVW) is to

perform a typical regression using the full sample and weighting the observations by

the inverse probability of selection $\left(\frac{N_h}{n_h}\right)_i$ the 2nd method (STRW) is also a typical regression using the full sample, but with weights the stratification weights

$\left(n_h - n \cdot \frac{N_h}{N}\right)_i$ the 3rd method (BS) is bootstrap: B bootstrap samples are selected with replacement from the initial sample, but the distribution of the observations into the strata matches the population structure and not the initial sample one (B=500); the 4th method (COMB) is to perform the regression separately within each sample stratum and combine the estimate using the stratification weights, (i.e.

$\hat{\beta}_1 = \sum_{h=1}^H W_h \cdot \hat{\beta}_{1h}$; and $\sigma_{\hat{\beta}_1}^2 = \sum_{h=1}^H W_h^2 \cdot \sigma_{\hat{\beta}_{1h}}^2$) and finally – for reference reasons – simple unweighted regression was performed (UNW).

The theoretical value of the regression coefficient and the variance of error term are calculated using the distribution properties, to serve as benchmarks of the estimates. The estimation is carried through the standard procedures available in SASTM software using the weight option, and the use of bootstrap to avoid weighting. Two different weights are used: (a) the typical inverse-probability-of-selection weight

$\left(\frac{N_h}{n_h}\right)$, (b) the standard stratified sampling weight $W_h = \frac{N_h}{N}$, and (c) no weight at all. In the bootstrapping process, the (B=500) bootstrap samples are selected with replacement so that each sample is distributed into the strata in the same proportions as the population. Finally, the estimation process is applied in each stratum separately and the resulting parameter estimates are combined using standard stratification weights [Cochran, 1977, and Zairis, 1991] in order to obtain the final estimates. Mean squared error is used as a measure of quality of the estimates, along with the estimate theoretical and estimated variances. To evaluate the five methods 1000 simulations were carried out using SAS v9.2 software.

3. IMPLEMENTATION

In order to create the population the following method was used. The values of X were simulated using a uniform distribution over [5, 9]. The values of Y were simulated in the following manner: in stratum 1 $Y=3+5X+e$, and in stratum 2 $Y=3+4X+e$. The error terms formed the three cases examined: (a) Normal, with $e \sim \text{Normal}(0, 4)$ in stratum 1 and $e \sim \text{Normal}(0, 16)$, (b) $e \sim \text{Uniform}[-1, 1]$ in both

strata, and (c) $e \sim \text{Double Exponential}$ $\left(f(x) = \frac{1}{2\beta} e^{-\frac{|x-\alpha|}{\beta}}\right)$ with $\alpha=0$ and $\beta=1$, (i.e. $\mu=0$ and $\sigma^2=2$), in both strata. From these three populations 1000 samples of size $n=1000$ where drawn using the procedure SURVEYSELECT of SAS to derive stratified samples with sizes 100 and 900 respectively for the two strata. This

arbitrary distribution of the sample into the strata allowed to evaluate the performance of the method when the distribution is at odds with the population distribution. Then the true values of the parameters and the true variances of their estimators were calculated in each case. The procedure was repeated for sample sizes with $n=100$ distributed in the same way as the larger samples (10 and 90, respectively into the two strata). The calculations follow.

3.1 The normal error case

The model for the full population was calculated using the procedure REG of SAS and yielded $\beta_0 = 3.02289$, and $\beta_1 = 4.3962$. The error variance σ^2 was calculated using the following formula

$$: \sigma^2 = \frac{\sum_{h=1}^2 (N_h - 1) S_h^2 + \sum_{h=1}^2 N_h (\bar{Y}_h - \bar{Y})^2}{N - 1},$$

where

$$N_1 = 40000, \quad N_2 = 60000, \quad N = 100000, \quad \bar{Y}_1 = 3 + 5 \cdot 7 = 38, \quad \bar{Y}_2 = 3 + 4 \cdot 7 = 31, \\ \hat{S}_1^2 = 4, \quad S_2^2 = 16, \quad \text{and} \quad \bar{Y} = 0.4 \cdot 38 + 0.6 \cdot 31 = 33.8, \text{ thus.}$$

3.2 The uniform error case

The model for the full population was calculated using the procedure REG of SAS and yielded $\beta_0 = 3.02289$, and $\beta_1 = 4.3962$. The error variance σ^2 was calculated using the following formula

$$: \sigma^2 = \frac{\sum_{h=1}^2 (N_h - 1) S_h^2 + \sum_{h=1}^2 N_h (\bar{Y}_h - \bar{Y})^2}{N - 1},$$

where

$$N_1 = 40000, \quad N_2 = 60000, \quad N = 100000, \quad \bar{Y}_1 = 3 + 5 \cdot 7 = 38, \quad \bar{Y}_2 = 3 + 4 \cdot 7 = 31, \\ S_1^2 = 0.333, \quad S_2^2 = 0.333, \quad \text{and} \quad \bar{Y} = 0.4 \cdot 38 + 0.6 \cdot 31 = 33.8, \text{ thus.}$$

3.3 The double exponential error case

The model for the full population was calculated using the procedure REG of SAS and yielded $\beta_0 = 2.9953$, and $\beta_1 = 4.4004$. The error variance σ^2 was calculated using the following formula

$$: \sigma^2 = \frac{\sum_{h=1}^2 (N_h - 1) S_h^2 + \sum_{h=1}^2 N_h (\bar{Y}_h - \bar{Y})^2}{N - 1},$$

where $S_1^2 = 2$, $S_2^2 = 2$ and $\bar{Y} = 0.4 \cdot 38 + 0.6 \cdot 31 = 33.8$, thus.

4. SIMULATION RESULTS

For each of the three populations, the five methods of estimation were applied. First a sample of size 1000 was drawn from the respective population using stratified random sampling with strata sizes 100 and 900. Then the regression using the inverse probability weighting was performed (INVW) and the variance of $\hat{\beta}_1$ was calculated using the standard formula $\sigma_{\hat{\beta}_1}^2 = \frac{\sigma^2}{S_{XX}^2}$, with $S_{XX}^2 = 999 \cdot (0.16 \cdot S_{1X}^2 + 0.16 \cdot S_{2X}^2)$, for n=1000 and $S_{XX}^2 = 99 \cdot (0.16 \cdot S_{1X}^2 + 0.36 \cdot S_{2X}^2)$, for n=100, where S_{1X}^2 and S_{2X}^2 were the variances of X in stratums 1 and 2, respectively. This true variance was common for all five methods in each sample size group separately, since it is a function of the population error variance and the sample X's. Next, the regression using the stratification weights was performed (STRW), followed by the bootstrap process (BS), and then by the classical stratified estimation calculations (COMB). Finally the unweighted estimation (UNW), was implemented. The results are presented in tables 1, 2, 3 & 4 below. All estimation was carried out using PROC REG of SASTM, [Pringle & Rayner1971].

Table 1. Parameter estimates, their standard deviations, mean squared errors for $\hat{\beta}_1$, along with the difference from the true value for $\hat{\beta}_1$ and $\sigma_{\hat{\beta}_1}$, sample size n=1000.

Error distribution	method	$\hat{\beta}_1$	$\sigma_{\hat{\beta}_1}$	MSE	$\hat{\beta}_1 - \beta_1$	$\sigma_{\hat{\beta}_1} - \sigma_{\beta_1}$
Normal ($\beta_1 = 4.3962$, $\sigma_1(\beta_1, 1) = 0.1826$)	INVW	4.3883	0.1837	0.0338	-0.0134	0.0014
	STRW	4.0803	0.1098	0.1118	-0.3202	-0.0813
	BS	4.3884	0.1841	0.0339	-0.0131	0.0003
	COMB	4.4076	0.0984	0.0099	0.0114	-0.0842
	UNW	4.1080	0.1201	0.0975	-0.2882	-0.0625
Uniform [$\beta_1, 1 = 4.4031$, $\sigma_1(\beta_1, 1) = 0.1322$]	INVW	4.4058	0.1582	0.0250	0.0027	0.0260
	STRW	4.0699	0.0443	0.1130	-0.3332	-0.0879
	BS	4.4058	0.1586	0.0251	0.0027	0.0264
	COMB	4.4016	0.0225	0.0005	-0.0015	-0.1097
	UNW	4.1012	0.0597	0.0943	-0.3019	-0.0725
Double Exponential	INVW	4.4020	0.1671	0.0279	0.0016	0.0262
	STRW	4.0661	0.0578	0.1151	-0.3343	-0.0831

$[\beta_1 = 4.4004,$ $\sigma_1(\sigma_1) = 0.1409)$	BS	4.4020	0.1674	0.0280	0.0016	0.0265
	COMB	4.3978	0.0555	0.0034	-0.0026	-0.0854
	UNW	4.0974	0.0722	0.0945	-0.3030	-0.0687

Table 2. Parameter estimates and their standard deviations for β_0 , sample size $n = 1000$.

Error distribution	method	$\hat{\beta}_0$	$\hat{\sigma}_{\hat{\beta}_0}$
Normal ($\beta_0 = 3.0223$)	INVW	3.0754	1.2650
	STRW	2.9203	0.7740
	BS	3.0748	1.2679
	COMB	2.9434	0.7013
	UNW	2.9372	0.8472
Uniform ($\beta_0 = 2.9786$)	INVW	2.9540	1.0770
	STRW	2.9941	0.3031
	BS	2.9542	1.0796
	COMB	2.9882	0.1592
	UNW	2.9871	0.4078
Double Exponential ($\beta_0 = 2.9953$)	INVW	2.9795	1.1477
	STRW	3.0173	0.4011
	BS	2.9791	1.1504
	COMB	3.0098	0.4118
	UNW	3.0159	0.4987

Table 3. Parameter estimates, their standard deviations and mean squared errors for $\hat{\beta}_1$, along with the difference from the true value for $\hat{\beta}_1$ and $\hat{\sigma}_{\hat{\beta}_1}$, sample size $n=100$.

Error distribution	method	$\hat{\beta}_1$	$\hat{\sigma}_{\hat{\beta}_1}$	MSE	$\hat{\beta}_1 - \beta_1$	$\hat{\sigma}_{\hat{\beta}_1} - \sigma_{\hat{\beta}_1}$
Normal ($\beta_1 = 4.3962,$ $\sigma_1(\sigma_1) = 0.5625$)	INVW	4.3752	0.5885	0.3464	-0.0210	0.0060
	STRW	4.0654	0.3665	0.2436	-0.3308	-0.2160
	BS	4.3760	0.5918	0.3502	-0.0202	0.0093
	COMB	4.4187	0.3279	0.1107	0.0225	-0.2546
	UNW	4.0978	0.3924	0.2428	-0.2984	-0.1901
Uniform ($\beta_1 = 4.4031,$	INVW	4.3805	0.4981	0.2484	-0.0226	0.0764

$\sigma_1(\beta^*_{11}) = 0.4217$	STRW	4.0683	0.1388	0.1314	- 0.3348	-0.2829
	BS	4.3808	0.5027	0.2530	- 0.0223	0.0810
	COMB	4.3970	0.0755	0.0058	- 0.0061	-0.3462
	UNW	4.0888	0.1872	0.1338	- 0.3143	-0.2345
Double Exponential $\beta_{11} = 4.4004$ $\sigma_1(\beta^*_{11}) = 0.4509$	INVW	4.3948	0.5554	0.3082	- 0.0056	0.1045
	STRW	4.0754	0.1922	0.1425	- 0.3250	-0.2587
	BS	4.3945	0.5595	0.3128	- 0.0059	0.1086
	COMB	4.3990	0.1792	0.0363	- 0.0014	-0.2717
	UNW	4.1066	0.2218	0.1355	- 0.2938	-0.2291

Table 4. Parameter estimates and their standard deviations for $\hat{\beta}_0$.
sample size $n = 100$.

Error distribution	method	$\hat{\beta}_0$	$\hat{\sigma}_{\hat{\beta}_0}$
Normal ($\beta_{10} = 3.0229$)	INVW	3.0992	4.0453
	STRW	3.0206	2.5918
	BS	3.0950	4.0691
	COMB	2.8667	0.3279
	UNW	2.9912	2.7701
Uniform ($\beta_0 = 2.9786$)	INVW	3.0806	3.3957
	STRW	3.0066	0.9487
	BS	3.0797	3.4286
	COMB	3.0220	0.5420
	UNW	3.0801	1.2799
Double Exponential ($\beta_0 = 2.9953$)	INVW	2.9744	3.8180
	STRW	2.9605	1.3480
	BS	2.9765	3.8470
	COMB	3.0001	1.3637
	UNW	2.9591	1.5339

5. CONCLUSIONS

The inverse weighting method (INVW) and the bootstrap method (BS) give the – overall – best results; the estimate of β_1 is consistent (as measured by the difference from the true value) and has a small MSE. The classical stratification approach (COMB) yields the second best results; the estimator of β_1 is closer to its true value and its MSE is smaller in the case of normal and uniform errors. The other two methods (classical weighting – STRW and no weighting – UNW) fall behind both in consistency and the size of MSE. What makes the first two methods optimal is the fact that they also estimate the variance of the estimator consistently, whereas the third method (COMB) – even though it appears to be more consistent and better in the sense of MSE – it clearly underestimates the variance of the estimate. This underestimation – in a hypothesis testing setting – will result in erroneously statistically significant conclusions.

As far as the rate of convergence is concerned, looking at the results for $n=100$, the pattern is very similar. The inverse weighting (INVW), the bootstrap (BS) and the classical stratification (COMB) methods give best results, with the first two being better in the sense of variance estimation, while the latter seems more consistent, in the cases of the Uniform and Double Exponential distributions. The other two methods give similarly poor results as in the case on $n=1000$.

Regarding the constant term estimation, the results are mixed. The inverse weighting (INVW), the bootstrap (BS) and the classical stratification (COMB) methods give best results, in the normal and double exponential error cases, but not in the uniform case.

The overall conclusion can be summarized in the following way: the methods of choice should be either inverse weighting (INVW) or bootstrap (BS), since they give consistent estimators and estimate the estimator variance correctly. The application of these methods presupposes the use of a statistical package, which can incorporate the weighted estimation in regression or the derivation of bootstrap samples (with sampling with replacement). The classical stratification (COMB) method can also be used when such statistical packages are not available (e.g. when using EXCEL), but caution should be exercised in hypothesis testing (for instance require smaller α levels to declare statistical significance).

ΠΕΡΙΛΗΨΗ

Το θεωρητικό πρόβλημα του υπολογισμού εκτιμητριών των συντελεστών στην απλή γραμμική παλινδρόμηση, όταν τα δεδομένα προέρχονται από ένα στρωματοποιημένο δείγμα έχει αντιμετωπιστεί στη στατιστική βιβλιογραφία από τη δεκαετία του '80. Οι προτεινόμενες λύσεις είναι μάλλον περίπλοκες. Στην παρούσα εργασία γίνεται μια σύγκριση - μέσω προσομοιώσεων - ορισμένων απλών διαδικασιών για την αντιμετώπιση του προβλήματος. Χρησιμοποιούνται τρία υποδείγματα παλινδρόμησης με σφάλματα που ακολουθούν την κανονική, την ομοιόμορφη $[-1,1]$ και διπλή εκθετική κατανομή, για να εξακριβωθεί η επίδραση της κατανομής των σφαλμάτων στην ακρίβεια των μεθόδων εκτίμησης. Η θεωρητική τιμή του συντελεστή παλινδρόμησης και τη διακύμανσή της υπολογίζεται χρησιμοποιώντας τις ιδιότητες των κατανομών, για να χρησιμεύσουν ως σημεία αναφοράς των εκτιμήσεων. Η εκτίμηση γίνεται μέσω προγραμμάτων διαθέσιμων στο λογισμικό SASTM χρησιμοποιώντας την επιλογή συντελεστή στάθμισης, καθώς και δειγματοληψία με επανάθεση (bootstrap) για να αποφευχθεί η στάθμιση. Χρησιμοποιούνται τρεις διαφορετικοί συντελεστές στάθμισης: (α) ο τυπικός συντελεστής ίσος με την αντίστροφη-πιθανότητα-επιλογής (β) ο κλασσικός συντελεστής ίσος με τον συντελεστή στάθμισης

$$W_h = \frac{N_h}{N}$$

στρωματοποιημένης δειγματοληψίας, και (γ) κανένας συντελεστής. Κατά τη διαδικασία δειγματοληψίας με επανάθεση τα δείγματα επιλέγονται ούτως ώστε, αφ' ενός μεν το μέγεθος του δείγματος να είναι το ίδιο με το αρχικό, αφ' ετέρου δε η κατανομή εντός των στρωμάτων να ακολουθεί τις ίδιες αναλογίες με τον πληθυσμό. Τέλος η διαδικασία εκτίμησης εφαρμόζεται ξεχωριστά σε κάθε στρώμα και οι προκύπτουσες εκτιμήτριες συνδυάζονται για να υπολογιστεί η τελική εκτίμηση με χρήση των τύπων στη στρωματοποιημένης δειγματοληψίας. Το μέσο τετραγωνικό σφάλμα και η διαφορά εκτιμώμενων από τις πραγματικές τιμές χρησιμοποιούνται ως τα κύρια μέτρα ακρίβειας των εκτιμήσεων, μαζί με την θεωρητική και εκτιμώμενη διακύμανση των εκτιμητριών και την αντίστοιχη διαφορά. Τα αποτελέσματα δείχνουν ότι η χρήση της αντίστροφης-πιθανότητας-επιλογής ως συντελεστή στάθμισης παράγει τα καλύτερα αποτελέσματα και αμέσως μετά η δειγματοληψία με επανάθεση. Ο τυπικός συντελεστής στάθμισης υστερεί - όπως αναμενόταν από τη θεωρία - δεδομένου ότι, τα αποτελέσματα επηρεάζονται από την δομή των εξισώσεων εκτίμησης που είναι πιο κοντά στην εκτίμηση λόγου. Ακόμη και αν η ανεξάρτητη μεταβλητή X αντιμετωπίζεται ως μη στοχαστική, δηλαδή θα μπορούσε κανείς να ισχυριστεί ότι οι τυπικοί OLS εκτιμητές δεν είναι εκτιμήτριες λόγου, αλλά απλώς πολλαπλασιάζονται με μια σταθερά, οι αλγόριθμοι επηρεάζουν την λύση οδηγώντας σε υπο-εκτίμηση της διακύμανσης των σφαλμάτων.

Acknowledgements: The author wishes to thank the anonymous reviewer for the useful and constructive comments.

ΑΝΑΦΟΡΕΣ

- Bassiakos Y., (2014), Estimation Problems in Stratified sampling, Proceedings of the 27th PanHellenic Statistical Congress, Athens, Greece.
- Cochran, W. G. (1977), *Sampling Techniques*, Third Edition, New York: John Wiley & Sons.
- Jewell, N. P. (1985), *Least squares regression with data arising from stratified samples of the dependent variable*, Biometrika, **72**, 1, 11-21.
- Hidiroglou, M. A., Fuller, W. A., and Hickman, R. D. (1980), *SUPER CARP*, Ames: Statistical Laboratory, Iowa State University.
- Pringle, R. M. and Rayner, A. A. (1971), *Generalized Inverse Matrices with Applications to Statistics*, New York: Hafner Publishing.
- Quesenberry C. P. & Jewell, N. P. (1986), *Regression analysis based on stratified samples*, Biometrika, **73**, 3, 605-614.
- Zairis Poseidon (1991), *Sampling techniques*, Athens (in Greek).



COMPARISON OF CORRELATION MEASURES FOR DETECTION OF BRAIN NETWORK CONNECTIVITY ALTERATIONS DURING EPILEPTIFORM DISCHARGES

**Batziou E., Gravanis M., Karadimos P., Koulikas D., Mpatziakas A.,
Paragamian S., Tonidis E., Xanthopoulou G.**

Inter-Faculty Master Program on Complex Systems and Networks, Aristotle
University of Thessaloniki

{empatziou, mgravanis, pkaradim, demkoul, ampatziakas, savkalpar, etonidis,
georgixv}@math.auth.gr

ABSTRACT

The research on epileptiform discharges (ED) in electroencephalographic recordings of patients with epilepsy has shown that changes are observed in connectivity of the brain area correlation networks. The data processed describe various epileptiform discharge situations. In some cases, the technique of transcranial magnetic stimulation (TMS), a noninvasive stimulation technique applicable in the field of focal epilepsy treatment, is used. In the current study, we examine the connectivity change of networks, formed by analyzing the correlation of selected EEG channels using cross correlation and cross mutual information networks. The time evolution of these networks and various metrics, relative to their structure, is studied, in order to identify four distinct states in the recording of epileptiform discharges: resting state, before and during epileptiform discharge (ED) and epileptiform discharge under transcranial magnetic stimulation (TMS).

Keywords: Epileptiform discharges; electroencephalogram; transcranial magnetic stimulation; brain connectivity; directed networks, entropy, correlation networks.

1. INTRODUCTION

Human brain is one of the most complex systems in nature. Functions of the brain lay on its connectivity and interaction dynamics of its neurons [Schindler *et al.* (2008)]. Brain connectivity consists of the anatomical connectivity (e.g connectome), the functional connectivity (e.g the neuronal activation patterns) and the effective connectivity (e.g the causal relationships between neurons) [Lang *et al.* (2012); Sakkalis (2011)]. Continuous development of neurophysiological techniques such as

Electroencephalography (EEG) and Magnetoencephalography (MEG) has provided us an insight of some emergent phenomena of the brain such as neuronal synchronization [Pereda (2005)]. Studies using data derived from these techniques and then followed by multivariate time series and nonlinear dynamical analysis, showed important results in various aspects of functional and effective connectivity of the brain. Some important results come from research areas of sleep, epilepsy, psychiatric disease and normal and disturbed cognition [Stam (2005)].

Epilepsy is one of the most common neurological disorders and is associated with changes of neuronal synchronization [Lehnertz (2008)]. Apart from epilepsy seizures a lot of people are experiencing epileptiform discharges which are defined by International Federation of Societies for Electroencephalography and Clinical Neurophysiology [Chatrian (1974)] as distinctive waves or complexes, distinguished from background activity, and resembling those recorded in a proportion of human subjects suffering from epileptic disorders. And therefore epileptiform discharges (ED) can also be detected by EEG and mediated by a non-invasive technique, transcranial magnetic stimulation (TMS) [Kimiskidis *et al.* (2013)]. Also there is now growing evidence that an improved understanding of seizure dynamics can be achieved when considering epileptic seizures as network phenomena [Schindler *et al.* (2008)].

Complex systems, like the brain, can be modeled as complex networks of many interacting components [Strogatz (2001)]. In the case of EEG recordings, nodes represent different channels and links between the channels are represented by their relationship. The possible relationship between channels can be obtained from a similarity measure using multivariate time series analysis [Kolaczyk (2009)]. A link exists if the similarity measure is above a threshold and/or is statistically significant; otherwise a weighted network is the result. The similarity measure can be defined as the linear correlation between two variables X and Y, namely cross-correlation. In order to escape the assumption of linearity one can use a nonlinear measure like cross mutual information. Cross mutual information calculates mutual dependence between two signals by quantifying the amount of information gained about one signal from measuring the other. Both measures incorporate a time lag which can reflect a causal relationship and hence time directed interactions. But this is not always true because there may exist internal delay loops in the system and change the causal interpretation [Pereda (2005)].

Complex network analysis has emerged as an important tool for characterization of brain connectivity [Rubinov *et al.* (2010)]. A large number of global network properties have been defined, providing valuable information about the system [Newman (2010)]. Global measures such as clustering coefficient, degree distribution, average path length, diameter, entropy of the degree distribution can determine properties such as small worldness [Watts *et al.* (1998); Bassett *et al.*

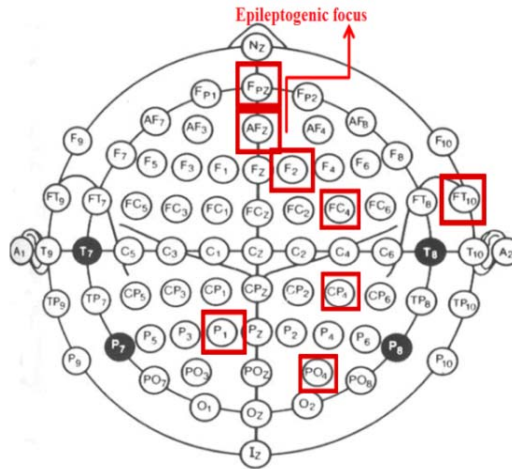
(2006)] and scale free distribution [Barabasi *et al.* (1999)], both very important for the detection and analysis of synchronization phenomena [Stam *et al.* (2007)].

In this work, we investigate differences in the network structure and its properties estimated from the aforementioned time-domain connectivity measures of the signal of selected EEG channels, before and during ED, as well as with the application of TMS. Also, this work is based on the work of [Siggiridou *et al.* (2014)], who designed the EEG experiments and examined the changes in network structure as derived from cross correlation measure.

2. DATA

The data used in this study are obtained from a single patient, whose EEG contained frequent epileptiform discharges (T~10sec). Acquisition procedure and preprocessing method are extensively described in [Siggiridou *et al.* (2014)] and the EEG channels selected for analysis are shown in Fig.1. Selection of channels was based on topographical criteria (epileptogenic area is the center of attention) and signal quality (channels free of various artifacts).

Figure 1: Positions of electrodes on the scalp surface according to 10-10 system. The channels used in this study and the epileptogenic zone are depicted. Reproduced with permission from [Siggiridou *et al.* (2014)].



3. METHODOLOGY

In order to detect possible alterations in network topology and the resulting network metrics across various states and the transitions between them, directed temporal networks are studied using two correlation metrics. In the following definitions X and Y represent two channels of the EEG, τ is the time lag and σ the standard deviation. The following are textbook definitions:

1. Cross correlation coefficient is:

$$r_{XY}(\tau) = \frac{\gamma_{XY}(\tau)}{\sigma_X \sigma_Y} \quad (1)$$

where

$$\gamma_{XY}(\tau) = \frac{1}{n-\tau} \sum_{t=1}^{n-\tau} (x_t - \bar{x})(y_{\tau+t} - \bar{y}) \quad (2)$$

2. Cross Mutual Information is:

$$J(X;Y)(\tau) = \sum_x \sum_y p(x_t, y_{\tau+t}) \log \frac{p(x_t, y_{\tau+t})}{p(x_t) p(y_{\tau+t})} \quad (3)$$

In order to eliminate any statistically significant autocorrelations, an autoregressive model $AR(p)$, of appropriate order p , is used for each time-series. The latter $AR(p)$ model is used to produce the residuals of the pre-whitening process, which have no significant autocorrelation.

The aforementioned correlation metrics $r_{xy}(\tau)$ and $J(X;Y)(\tau)$ are calculated for each pair of channels for lag values of $\tau = [0, 3]$, as larger lag values computed produced trivial results. Using a Portmanteau randomization test with $n=1000$ repetitions, p-values for the correlation measures were computed. This process results in the formation of networks in which every channel is a node and statistically significant correlations ($p < 0.05$) between pairs of channels correspond to edges between the nodes. The networks are binary except for the ones used to calculate strength, in which the weight of the edges is the absolute value of the correlation indices. For each of the networks we estimate:

- Directed Network Density [DND]:

$$D = \frac{|E|}{|V|(|V|-1)} \quad (4)$$

where $|E|$ and $|V|$ are the number of edges and nodes respectively.

- Average Path Length [APL]:

$$l_G = \frac{1}{n(n-1)} \sum_{i \neq j} d(v_i, v_j) \quad (5)$$

where n is the number of nodes and $d(v_i, v_j)$ the shortest distance between two nodes.

- Network Strength [NS]:

$$S_G = \sum_{i \neq j}^n w_{i,j} \quad i, j = (1, 2, \dots, n) \quad (6)$$

where $w_{i,j}$ is the weight of the link from node i to node j and $n = |V|$ is the number of nodes.

Finally, two different Shannon entropy metrics were estimated, each one on a different variable:

- Degree Distribution Entropy [DDE]:

$$E_{G_1} = - \sum_{i=1}^N p_i \log(p_i) \quad (7)$$

where p_i is the relative frequency of the degree i and N is the number of observed degrees on the network.

- In-Degree Entropy or “node transition probability” [IDE]:

$$E_{G_2} = - \sum_{i=1}^n \frac{\deg^-(v_i)}{|E(g)|} \log_2 \left(\frac{\deg^-(v_i)}{|E(g)|} \right) \quad (8)$$

where $E(g)$ is the number of edges in the network, $\deg^-(v_i)$ is the in-degree of a node, i.e. the number of incoming links.

4. RESULTS

In the following section we present three typical cases of the computed results for different states observed in the EEGs. Results of the observed metrics for all the datasets are presented in Tables 1 and 2 for Cross Correlation and Mutual Information respectively.

Table 1: Network metrics values for the networks of nine (9) datasets created using Cross Correlation.

Time (sec)	1	2	3	4	5	6	7	8	9	10	Dataset 1
DND	0,018	0,000	0,036	0,054	0,000	0,054	0,429	0,125	0,304	0,357	
APL	7,875	8,000	7,750	7,625	8,000	7,518	1,875	5,786	4,232	2,607	
NS	0,309	0,000	0,650	1,047	0,000	0,946	9,471	2,542	6,562	7,516	
DDE	0,811	0,000	1,000	0,811	0,000	1,406	1,750	1,906	1,406	2,250	
IDE	0,000	0,000	1,000	1,585	0,000	1,585	2,924	2,236	2,772	2,822	
STATE	RS	RS	RS	RS	RS	ED	ED	ED	ED	ED	

DND	0,161	0,214	0,214	0,107	0,000	0,018	0,036	0,089	0,107	0,054	Dataset 2
APL	6,661	5,464	5,964	7,250	8,000	7,875	7,750	7,054	7,143	7,625	
NS	3,290	5,045	4,866	2,226	0,000	0,388	0,641	1,692	1,923	1,071	
DDE	2,156	2,250	2,156	1,561	0,000	0,811	1,000	1,549	1,811	1,500	
IDE	2,503	2,459	2,396	1,792	0,000	0,000	1,000	2,322	1,792	1,585	
STATE	ED	ED	ED	ED	RS	RS	RS	RS	RS	RS	
DND	0,036	0,000	0,000	0,018	0,018	0,161	0,339	0,179	0,125	0,393	Dataset 3
APL	7,750	8,000	8,000	7,875	7,875	6,357	4,071	6,071	6,821	3,500	
NS	0,694	0,000	0,000	0,314	0,319	3,225	7,182	3,995	2,628	8,881	
DDE	1,299	0,000	0,000	0,811	0,811	1,500	2,156	2,156	1,906	1,811	
IDE	1,000	0,000	0,000	0,000	0,000	2,503	2,676	2,446	2,236	2,936	
STATE	RS	RS	RS	RS	ED	ED	ED	ED	ED	ED	
DND	0,018	0,036	0,143	0,018	0,018	0,054	0,054	0,000	0,000	0,000	Dataset 4
APL	7,875	7,643	6,571	7,875	7,875	7,625	7,625	8,000	8,000	8,000	
NS	0,400	0,949	3,472	0,409	0,436	1,387	1,295	0,000	0,000	0,000	
DDE	0,811	1,299	2,156	0,811	0,811	1,500	0,811	0,000	0,000	0,000	
IDE	0,000	1,000	2,156	0,000	0,000	1,585	1,585	0,000	0,000	0,000	
STATE	ED	ED	ED	ED	ED	ED	ED	RS	RS	RS	
DND	0,000	0,054	0,000	0,000	0,089	0,214	0,268	0,036	0,107	0,268	Dataset 5
APL	8,000	7,625	8,000	8,000	7,161	4,214	5,393	7,750	6,929	4,679	
NS	0,000	1,027	0,000	0,000	1,676	4,467	6,199	0,709	1,980	5,651	
DDE	0,000	1,406	0,000	0,000	1,406	2,156	2,250	1,000	1,299	1,906	
IDE	0,000	0,918	0,000	0,000	2,322	2,626	2,257	1,000	2,252	2,923	
STATE	RS	RS	RS	ED	ED	ED	ED	ED	ED	ED	
DND	0,018	0,036	0,036	0,054	0,268	0,036	0,143	0,286	0,071	0,018	Dataset 6
APL	7,875	7,750	7,750	7,625	4,446	7,643	6,571	5,357	7,500	7,875	
NS	0,357	0,687	0,665	1,004	5,777	0,717	3,240	6,702	1,299	0,424	
DDE	0,811	1,299	1,000	0,811	2,000	1,299	2,000	2,500	1,500	0,811	
IDE	0,000	1,000	1,000	1,585	2,790	1,000	2,500	2,555	1,500	0,000	
STATE	RS	RS	ED	ED	ED	ED	ED	ED	ED	RS	

DND	0,143	0,125	0,071	0,054	0,000	0,000	0,000	0,000	0,000	0,000	Dataset 7
APL	6,196	6,714	7,286	7,518	8,000	8,000	8,000	8,000	8,000	8,000	
NS	3,823	3,268	1,797	1,293	0,000	0,000	0,000	0,000	0,000	0,000	
DDE	2,156	2,000	1,750	1,406	0,000	0,000	0,000	0,000	0,000	0,000	
IDE	2,500	2,522	1,500	1,585	0,000	0,000	0,000	0,000	0,000	0,000	
STATE	ED	ED	ED	ED	ED	RS	RS	RS	RS	RS	
DND	0,071	0,179	0,018	0,036	0,054	0,196	0,071	0,089	0,143	0,161	Dataset 8
APL	7,500	5,536	7,875	7,750	7,518	6,107	7,393	7,161	6,464	6,250	
NS	1,281	3,660	0,330	0,620	0,972	4,269	1,395	1,666	2,773	3,408	
DDE	1,500	2,156	0,811	0,811	1,549	2,156	1,750	1,906	2,250	2,000	
IDE	1,500	2,446	0,000	1,000	1,585	2,118	1,500	1,522	1,906	2,281	
STATE	ED	TMS	TMS	TMS	TMS	TMS	TMS	TMS	TMS	TMS	
DND	0,018	0,000	0,018	0,000	0,143	0,018	0,036	0,036	0,000	0,018	Dataset 9
APL	7,875	8,000	7,875	8,000	6,018	7,875	7,750	7,643	8,000	7,875	
NS	0,430	0,000	0,405	0,000	3,709	0,491	0,850	0,898	0,000	0,409	
DDE	0,811	0,000	0,811	0,000	2,250	0,811	1,000	1,299	0,000	0,811	
IDE	0,000	0,000	0,000	0,000	2,250	0,000	1,000	1,000	0,000	0,000	
STATE	ED	TMS	TMS	TMS	TMS	TMS	TMS	TMS	TMS	TMS	

Table 2: Network metrics values for the networks of nine (9) datasets created using Cross Mutual Information

Time (sec)	1	2	3	4	5	6	7	8	9	10	Dataset 1
DND	0,018	0,036	0,000	0,036	0,000	0,071	0,125	0,089	0,107	0,196	
APL	7,875	7,643	8,000	7,643	8,000	7,393	6,804	7,375	6,232	5,500	
NS	0,227	0,430	0,000	0,438	0,000	0,864	1,662	1,089	1,287	2,387	
DDE	0,811	1,299	0,000	1,299	0,000	1,500	2,250	1,406	1,811	2,750	
IDE	0,000	1,000	0,000	1,000	0,000	2,000	1,842	1,922	2,252	2,482	
STATE	RS	RS	RS	RS	RS	ED	ED	ED	ED	ED	

DND	0,107	0,107	0,054	0,018	0,000	0,054	0,018	0,018	0,018	0,000	Dataset 2
APL	6,929	7,143	7,625	7,875	8,000	7,518	7,875	7,875	7,875	8,000	
NS	1,295	1,328	0,728	0,217	0,000	0,632	0,255	0,222	0,216	0,000	
DDE	1,811	1,750	0,954	0,811	0,000	1,406	0,811	0,811	0,811	0,000	
IDE	1,918	1,918	0,918	0,000	0,000	1,585	0,000	0,000	0,000	0,000	
STATE	ED	ED	ED	ED	RS	RS	RS	RS	RS	RS	
DND	0,054	0,018	0,036	0,018	0,036	0,036	0,089	0,071	0,036	0,179	Dataset 3
APL	7,625	7,875	7,643	7,875	7,643	7,750	7,071	7,393	7,643	5,911	
NS	0,617	0,206	0,425	0,218	0,442	0,425	1,131	0,851	0,403	2,115	
DDE	0,811	0,811	1,299	0,811	1,299	1,000	1,500	1,811	1,299	1,906	
IDE	1,585	0,000	1,000	0,000	1,000	1,000	1,922	2,000	1,000	2,446	
STATE	RS	RS	RS	RS	ED	ED	ED	ED	ED	ED	
DND	0,089	0,036	0,214	0,054	0,018	0,071	0,036	0,036	0,018	0,036	Dataset4
APL	7,161	7,643	5,000	7,518	7,875	7,393	7,750	7,750	7,875	7,750	
NS	1,101	0,429	2,623	0,711	0,238	0,850	0,451	0,478	0,209	0,469	
DDE	1,906	1,299	1,750	1,406	0,811	1,811	1,000	1,000	0,811	1,000	
IDE	1,922	1,000	2,459	1,585	0,000	2,000	1,000	1,000	0,000	1,000	
STATE	ED	ED	ED	ED	ED	ED	ED	RS	RS	RS	
DND	0,000	0,054	0,000	0,000	0,089	0,054	0,179	0,054	0,036	0,054	Dataset 5
APL	8,000	7,518	8,000	8,000	6,964	7,518	5,375	7,518	7,750	7,625	
NS	0,000	0,631	0,000	0,000	1,094	0,634	2,294	0,679	0,437	0,648	
DDE	0,000	1,406	0,000	0,000	1,549	1,406	2,500	1,500	1,000	1,406	
IDE	0,000	1,585	0,000	0,000	2,322	1,585	2,446	1,585	1,000	0,918	
STATE	RS	RS	RS	ED	ED	ED	ED	ED	ED	ED	
DND	0,125	0,018	0,018	0,036	0,107	0,071	0,054	0,125	0,054	0,000	Dataset 6
APL	6,911	7,875	7,875	7,750	7,250	7,196	7,625	6,607	7,625	8,000	
NS	1,463	0,204	0,212	0,473	1,258	0,819	0,661	1,566	0,662	0,000	
DDE	2,000	0,811	0,811	1,299	1,561	1,561	1,406	2,156	0,811	0,000	
IDE	1,842	0,000	0,000	0,000	1,792	2,000	0,000	1,449	1,585	0,000	
STATE	RS	RS	ED	ED	ED	ED	ED	ED	ED	RS	

DND	0,286	0,071	0,107	0,089	0,000	0,054	0,036	0,018	0,018	0,036	Dataset 7
APL	3,339	7,393	6,929	7,161	8,000	7,625	7,643	7,875	7,875	7,750	
NS	3,540	0,833	1,307	1,108	0,000	0,660	0,454	0,205	0,208	0,429	
DDE	2,406	1,500	1,906	1,406	0,000	0,954	1,299	0,811	0,811	1,299	
IDE	2,733	2,000	2,252	1,922	0,000	0,918	1,000	0,000	0,000	0,000	
STATE	ED	ED	ED	ED	ED	RS	RS	RS	RS	RS	
DND	0,036	0,071	0,036	0,036	0,089	0,089	0,054	0,054	0,071	0,107	Dataset 8
APL	7,750	7,196	7,750	7,750	6,571	7,161	7,518	7,625	7,393	6,554	
NS	0,409	0,897	0,450	0,448	1,039	1,111	0,676	0,623	0,837	1,290	
DDE	1,000	1,500	1,000	0,811	1,906	1,906	1,406	1,406	1,061	1,750	
IDE	1,000	1,500	1,000	1,000	1,922	1,371	1,585	0,918	2,000	2,252	
STATE	ED	TMS	TMS	TMS	TMS	TMS	TMS	TMS	TMS	TMS	
DND	0,018	0,000	0,018	0,054	0,143	0,089	0,089	0,036	0,071	0,018	Dataset 9
APL	7,875	8,000	7,875	7,625	6,089	6,875	7,375	7,750	7,286	7,875	
NS	0,207	0,000	0,266	0,640	1,742	1,113	1,075	0,438	0,834	0,209	
DDE	0,811	0,000	0,811	1,406	1,500	1,906	1,500	1,000	1,500	0,811	
IDE	0,000	0,000	0,000	1,585	2,156	1,922	1,922	1,000	2,000	0,000	
STATE	ED	TMS	TMS	TMS	TMS	TMS	TMS	TMS	TMS	TMS	

Dataset 2 describes a situation in which the subject is initially experiencing epileptic discharges (ED) (seconds 0-4) and resting state (RS) condition follows after the 5th second until the end of the observations (10th second).

Figure 2: Network metrics of dataset 2 using Cross Correlation

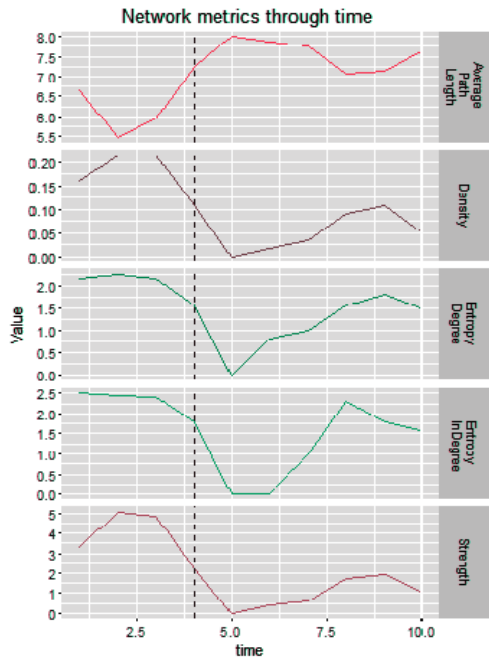
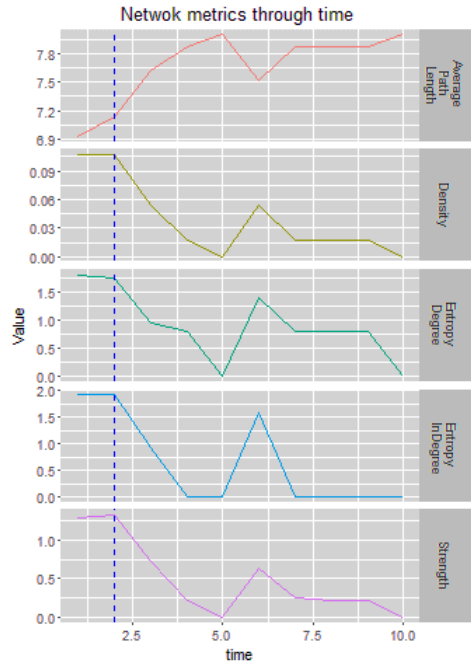


Figure 3: Network metrics of dataset 2 using Cross Mutual Information



As one can observe in Figures 2-3, the ED state is depicted on the behavior of the metric values, both in CMI (Cross Mutual Information) and CC (Cross Correlation) cases: APL abruptly increase while Strength (NS), Density (DND) and both Entropy measures (DE and IDE) decline from the 5th second and thereafter. Values of the latter metrics also gradually increase and reach a value slightly smaller than their initial.

Dataset 3 presents a case in which the subject is initially in resting state condition (seconds 0-5) and ED follows after the 6th second (seconds 6-10). In Figures 4-5, one can observe the transition effect from resting state to ED. This shift is well reflected on the CC network metrics: network APL abruptly declines while Strength (NS), Density (DND) and both Entropy measures (DE and IDE) increase from the 6th second and thereafter. In the case of the CMI networks, the image is similar: Metric values change in a way resembling the CC case, however the changes after the 6th second are less abrupt.

Figure 4: Network metrics of dataset 3 using Cross Correlation

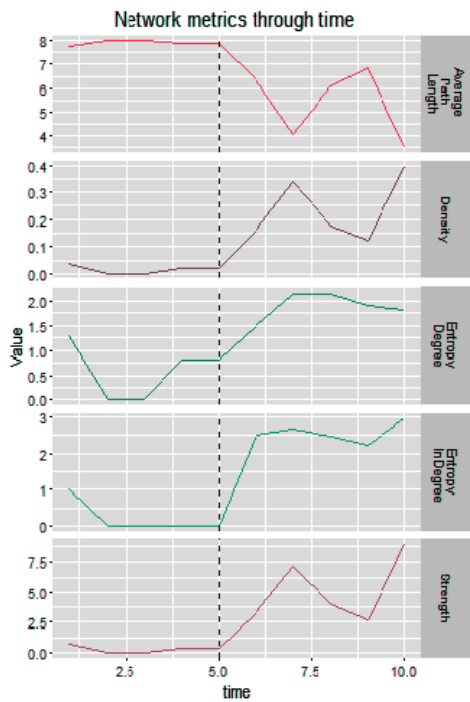
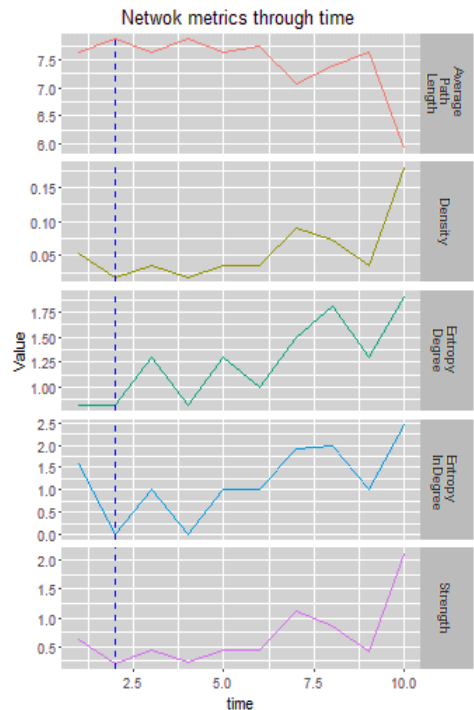


Figure 5: Network metrics of dataset 3 using Cross Mutual Information



Dataset 8 describes a situation in which ED occurs throughout the 10 seconds of the observation, while TMS (state TMS) is applied on the 3rd second. As one can observe in Figures 6-7, TMS application while in an ED state, has an effect that resembles the transition from ED to resting state (RS): This effect lasts for a limited time period of five (5) seconds and then the metrics regress to their initial values. In the case of the CC networks, APL abruptly declines while Strength (NS), Density (DND) and both Entropy measures (DE and IDE) increase after the 3rd second. CMI networks behave similarly, but as observed also in dataset 3, alterations are smoother. Density, Strength and both Entropies display values higher than their initial in the last seconds of the observation.

Figure 6: Network metrics of dataset 8 using Cross Correlation

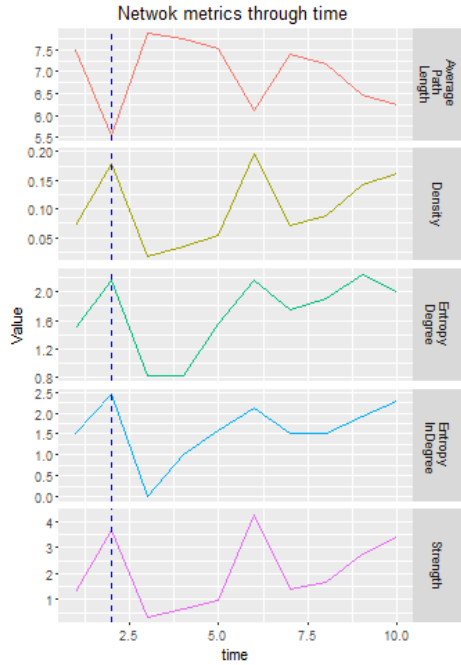
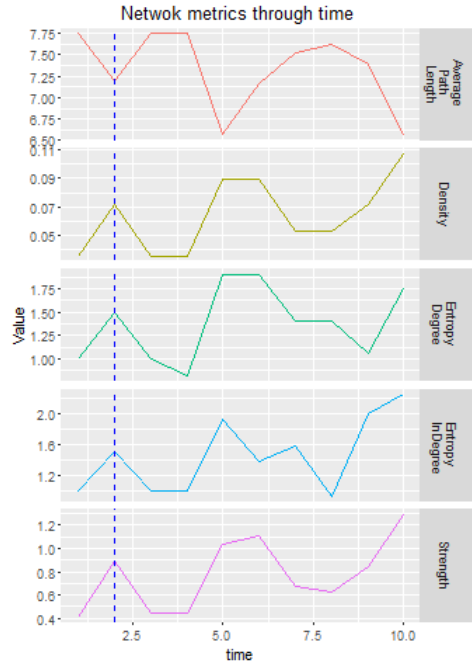


Figure 7: Network metrics of dataset 8 using Cross Mutual Information



5. CONCLUSIONS

We have elaborated on the existing literature about the connectivity analysis of the brain during ED and TMS. Our results confirm already reported findings, i.e. that significant correlations exist between different channels during ED. Most significant ones are identified for time lags in the interval $\tau = [0, 3]$, a time range which allows a brain response time of less than 4ms to be inferred. There are indications that cross correlation is more effective than cross mutual information in terms of using the alterations in connectivity to identify either a discharge or TMS application. This result is not surprising since nonlinear approaches sometimes are less robust to noise than linear ones [Pereda *et al.* (2005); Stam (2005)]. Also, these authors note that linear and nonlinear approaches access different parts of interdependence between channels. In the case of EEG channels, cross correlation and mutual information have been both successfully used to extract information for epilepsy seizures diagnosis and prediction.

The network properties chosen seem to have a consistent behavior in all datasets examined, which renders them suitable for the identification of brain network

connectivity changes as well as a possible recognition of the current state (Resting, ED or TMS). The findings are summarized in the below table.

Table 3: Network Metrics Properties

	<u>APL</u>	<u>DND</u>	<u>DE/IDE</u>	<u>NS</u>
<u>Resting State</u>	Near maximum value	Near minimum value	Abrupt decrease (in entrance)	Low value
<u>Epileptiform Discharge</u>	Abrupt decrease	Increase	Abrupt increase (in the transition)	High value
<u>Epileptiform Discharge with TMS</u>	Temporary increase	Temporary decrease	Temporary decrease	Temporary decrease

In particular, degree distribution entropy seems to be a key identifier for transitions between states. Density and APL are able to reflect both transitions and the duration of each state. Strength has the same advantages but apparently demonstrates the weight of the links between the different channels as well.

ΠΕΡΙΛΗΨΗ

Η μελέτη των επιληπτικών εκφορτίσεων (ED) σε ηλεκτροεγκεφαλογραφικές (EEG) καταγραφές ασθενών με επιληψία, έχει δείξει ότι προκαλούνται αλλαγές στη συνδεσιμότητα των δικτύων που προκύπτουν από την συσχέτιση διαφορετικών περιοχών του εγκεφάλου. Τα δεδομένα τα οποία εξετάζονται αφορούν καταγεγραμμένες επιληπτικές εκφορτίσεις, σε κάποιες από τις οποίες γίνεται χρήση της τεχνικής του διακρανιακού μαγνητικού ερεθισμού (TMS), μιας μη επεμβατικής τεχνικής διέγερσης του Κεντρικού Νευρικού Συστήματος (ΚΝΣ) που εφαρμόζεται στην αντιμετώπιση της εστιακής επιληψίας. Στην παρούσα εργασία ερευνούμε την αλλαγή της συνδεσιμότητας στα δίκτυα που σχηματίζονται μέσω ανάλυσης της συσχέτισης επιλεγμένων καναλιών EEG. Η ανάλυση που πραγματοποιείται περιλαμβάνει τη χρήση της διασυσχέτισης (cross correlation) και της αμοιβαίας πληροφορίας (cross mutual information) μεταξύ των αντίστοιχων καναλιών. Μελετάται η εξέλιξη στο χρόνο των δικτύων αυτών και παρουσιάζεται η μεταβολή μιας σειράς δεικτών σχετικών με τη δομή τους, με σκοπό να ταυτοποιηθούν τέσσερις διακριτές καταστάσεις στην καταγραφή των επιληπτικών εκφορτίσεων: πριν την κρίση (preED), κατά τη διάρκεια της κρίσης ED, μετά την κρίση (postED), καθώς και η κατάσταση μετά τη χορήγηση TMS (postTMS).

Acknowledgements: We would like to thank Dimitris Kugiumtzis (Assistant Professor, Aristotle University of Thessaloniki) for providing us both the data used and constant guidance.

REFERENCES

Chatrian, G. E., Bergmini, L., Dondey, M., Klass, M., & Lennox-Buththal, M. Petersen. I.(1974). A glossary of terms most commonly used by clinical electroencephalographers. *Electroencephalography and Clinical Neurophysiology*, 37, 538-548.

Barabasi, A.-L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286(October), 509–512.

Bassett, D. S., & Bullmore, E. (2006). Small-World Brain Networks. *The Neuroscientist*, 12(6), 512–523. <http://doi.org/10.1177/1073858406293182>

Kimiskidis, V. K., Kugiumtzis D., Papagiannopoulos S., & Vlaikidis N. (2013). Transcranial Magnetic Stimulation (Tms) Modulates Epileptiform Discharges in Patients With Frontal Lobe Epilepsy: a Preliminary Eeg-Tms Study. *International Journal of Neural Systems*, 23(01), 1250035. <http://doi.org/10.1142/S0129065712500359>

Kolaczyk, E. (2009). *Statistical analysis of network data*. Springer series in Statistics. <http://doi.org/10.1007/978-0-387-88146-1>

Lang, E. W., Tomé, A. M., Keck, I. R., Górriz-Sáez, J. M., & Puntonet, C. G. (2012). Brain connectivity analysis: A short survey. *Computational Intelligence and Neuroscience*, 2012(iii). <http://doi.org/10.1155/2012/412512>

Lehnertz, K. (2008). Epilepsy and nonlinear dynamics. *Journal of Biological Physics*, 34(3-4 SPEC. ISS.), 253–266. <http://doi.org/10.1007/s10867-008-9090-3>

Newman, M. E. J. (2010). *Networks: An Introduction*. New York: Oxford University Press. <http://doi.org/10.1093/acprof:oso/9780199206650.001.0001>

Pereda, E., Quiroga, R. Q., & Bhattacharya, J. (2005). Nonlinear multivariate analysis of neurophysiological signals. *Progress in Neurobiology*, 77(1-2), 1–37. <http://doi.org/10.1016/j.pneurobio.2005.10.003>

Rubinov, M., & Sporns, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52(3), 1059–1069. <http://doi.org/10.1016/j.neuroimage.2009.10.003>

Sakkalis, V. (2011). Review of advanced techniques for the estimation of brain connectivity measured with EEG/MEG. *Computers in Biology and Medicine*, 41(12), 1110–1117. <http://doi.org/10.1016/j.compbiomed.2011.06.020>

Schindler, K. A., Bialonski, S., Horstmann, M. T., Elger, C. E., & Lehnertz, K. (2008). Evolving functional network properties and synchronizability during human epileptic seizures. *Chaos*, 18(3). <http://doi.org/10.1063/1.2966112>

Siggiridou, E., Kugiumtzis, D. & Kimiskidis, V. K. (2014). Correlation networks for identifying changes in brain connectivity during epileptiform discharges and transcranial magnetic stimulation. *Sensors (Switzerland)*, 14(7), 12585–12597. <http://doi.org/10.3390/s140712585>

Stam, C. J. (2005). Nonlinear dynamical analysis of EEG and MEG: Review of an emerging field. *Clinical Neurophysiology*, 116(10), 2266–2301. <http://doi.org/10.1016/j.clinph.2005.06.011>

Stam, C. J., Jones, B. F., Nolte, G., Breakspear, M., & Scheltens, P. (2007). Small-world networks and functional connectivity in Alzheimer's disease. *Cerebral Cortex*, 17(1), 92–99. <http://doi.org/10.1093/cercor/bhj127>

Strogatz, S. H. (2001). Exploring complex networks. *Nature*, 410(6825), 268–276. <http://doi.org/10.1038/35065725>

Watts, D., & Strogatz, S. H. (1998). Collective Dynamics of “Small-World” Networks. *Nature*, 393(6684), 440–442. <http://doi.org/10.1038/30918>



FRACTIONAL FACTORIAL DESIGNS AND ORTHOGONAL ARRAYS

Vasilis Chasiotis¹, Stratis Kounias¹, Nikolaos Farmakis¹

¹Department of Mathematics, Aristotle University of Thessaloniki, Greece

²Department of Mathematics, University of Athens, Panepistemiopolis, Zografou, Greece

chasiotisv@math.auth.gr, skounias@math.uoa.gr, farmakis@math.auth.gr

ABSTRACT

In 3^k fractional factorial designs, where there are k factors, each at 3 levels, the interest is to find the optimal designs for estimating the linear and quadratic contrasts of factor level effects. We are working with resolution III designs, where two and higher order interactions are negligible. If the number N of observations (runs) is $N \equiv 0 \pmod 9$, the ϕ -optimal design is an Orthogonal Array with three symbols and strength two, $OA(N, k, 3, 2)$. The notion of Balanced Arrays is introduced and we develop algorithms for the construction of optimal designs for many values of N . For $N=18$ the $OA(18, k, 3, 2)$ exists for $k = 2, 3, \dots, 7$ but not for the maximum number of factors which is $k=8$. We present the D-optimal design for $N=18$ and $k=8$ factors.

Key words and phrases: Balanced Arrays, ϕ -optimality.

1. INTRODUCTION AND THE MODEL

In 3^k fractional factorial designs there are k factors $F_1, \dots, F_k$ each at three levels named: low, middle and high level and denoted by 0, 1, 2. The effect of factor F_i at level 0 is l_i , at level 1 is u_i and at level 2 is q_i . In resolution III designs the second and higher order interactions are negligible and only main effects are involved. We are interested in estimating the standardized linear and quadratic contrasts $(q_i - l_i) / \sqrt{2}$ and $(q_i - 2u_i + l_i) / \sqrt{6}$ $i = 1, \dots, k$ respectively.

For estimating these contrasts [Kiefer (1961)] proved that if $N \equiv 0 \pmod 9$ the ϕ -optimal design is the Orthogonal Array $OA(N,k,3,2)$, [Kolyva-Machera (1989)] and [Mukerjee (1999)] worked on the optimality for $N \equiv 1 \pmod 9$.

[Pericleous et al (2013,2014)] obtained the optimal designs if $N \equiv 0 \pmod 3$ and constructed the optimal designs for $N=12,15,21,24,27,30,33$. They also prove that the optimal designs for estimating linear and quadratic contrasts are Balanced Arrays. For the properties and applications of orthogonal Arrays see the book by [Hedayat, Sloane and Tufken (1999)].

The model for estimating the linear and quadratic contrasts is, [Pericleous et al (2012,2013)],

$$y = \mu \mathbf{1}_N + \sum_i^k \left(\frac{(q_i - l_i)}{\sqrt{2}} g_i + \frac{(l_i - 2u_i + q_i)}{\sqrt{6}} 3\mathbf{h}_i \right) + \mathbf{e} \quad (1)$$

where $\mathbf{g}_i = (g_{i1}, \dots, g_{iN})'$, $\mathbf{h}_i = (h_{i1}, \dots, h_{iN})'$,

$$\mathbf{g}_{i,s} = \frac{1}{\sqrt{2}} \begin{cases} -1 & \text{if } d(i,s) = 0 \\ 0 & \text{if } d(i,s) = 1 \\ 1 & \text{if } d(i,s) = 2 \end{cases}, \quad \mathbf{h}_{i,s} = \frac{1}{\sqrt{2}} \begin{cases} 1 & \text{if } d(i,s) = 0 \\ 0 & \text{if } d(i,s) = 1 \\ 1 & \text{if } d(i,s) = 2 \end{cases}$$

$d(i,s)$ is the level at which factor F_i participates in the experimental run s .

The total number of parameters is $2k+1$, here we are interested in estimating only the $2k$ linear and quadratic contrasts.

The main problem with discrete optimal designs is their construction.

Section 2 is devoted to the definition and construction of Balanced Arrays, when $N \equiv 1, 2 \pmod 3$, the case $N \equiv 0 \pmod 3$ was presented by [Pericleous et al (2013)]. In section 3 we present the construction of new $OA(N,k,3,2)$.

2. BALANCED ARRAYS

The Balanced Arrays $BA(N,k,m,2)$ were introduced by [Chakravarty (1961)] because designs derived from Balanced Arrays require a reduced number of runs in order to accommodate a given number of factors.

The following definition is an extension of Chakravarty's BA and in a number of cases gives better results.

Definition 2.1 A Balanced Array $BA(N,k,3,2)$ is a $N \times k$ structure with $k = r + s$ and the properties,

a) If $s=0$, then

(ia) Every column of the array has n^0, n^1, n^2 number of 0,1,2 respectively with $n^0 + n^1 + n^2 = N$, $n^0 = n^2$, $N - n^1 \equiv 0 \pmod{2}$.

(iia) In every two columns of the array the pairs 00,01,02,10,11,12,20,21,22 appear $n^{00}, n^{01}, n^{02}, n^{10}, n^{11}, n^{12}, n^{20}, n^{21}, n^{22}$ times respectively satisfying,

$$n^1 - n^{11} \equiv 0 \pmod{2} > 0,$$

$$n^{10} = n^{01} = n^{12} = n^{21} = (n^1 - n^{11}) / 2, n^{00} = n^{22}, n^{02} = n^{20},$$

$$n^{00} + n^{02} = (N - 2n^1 + n^{11}) / 2, \quad 0 \leq n^{00} - n^{02} \leq 1.$$

b) $0 < s < k$

(iiib) The BA consists of two blocks. The first block has r columns, the symbols are n_1^p, n_{11}^{pq} , $p, q = 0, 1, 2$ satisfying (ia) and (iia) above. The second block has s columns, the symbols are n_2^p, n_{22}^{pq} , $p, q = 0, 1, 2$ satisfying (ia) and (iia) above. For the relation between one column of the first block and one column of the second block define n_{12}^{11} satisfying: $n_1^1 - n_{12}^{11} \equiv 0 \pmod{2}$, $n_2^1 - n_{12}^{11} \equiv 0 \pmod{2}$, $n_{12}^{01} = n_{12}^{21} = (n_2^1 - n_{12}^{11}) / 2$, $n_{12}^{10} = n_{12}^{12} = (n_1^1 - n_{12}^{11}) / 2$, $n_{12}^{00} = n_{12}^{22}, n_{12}^{02} = n_{12}^{20}$, $0 \leq n_{12}^{00} - n_{12}^{02} \leq 1$, $n_{12}^{00} + n_{12}^{02} = (N - n_1^1 - n_2^1 + n_{12}^{11}) / 2$. •

From the above definition it is clear that if $s = 0$, the $BA(N, k, 3, 2)$ is completely defined if we know the values of N, n^1, n^{11} and if $0 < s < k$ the $BA(N, k, 3, 2)$, $k=r+s$, is defined if we know $N, n_1^1, n_{11}^{11}, n_2^1, n_{22}^{11}, n_{12}^{11}$.

Example 2.1 We give four designs which are $BA(N, k, 3, 2)$ for $N=7$ and $N=8$.

(a) $N = 7, n^1 = 5, n^{11} = 3$,

(b) $N = 7, n^1 = 3, n^{11} = 1$, (c) $N = 8, n^1 = 4, n^{11} = 2$, (d) $N = 8, n^1 = 2, n^{11} = 0$.

The values of the parameters n_{ij}^{pq} are determined from definition 2.

Table 1. Four Balanced Arrays

(a) $BA(7,3,3,2)$	(b) $BA(7,3,3,2)$	(c) $BA(8,3,3,2)$	(d) $BA(8,3,3,2)$
$N = 7, n^I = 5, n^{II} = 3$	$N = 7, n^I = 3, n^{II} = 1$	$N = 8, n^I = 4, n^{II} = 2$	$N = 8, n^I = 2, n^{II} = 0$
F_1 F_2 F_3	F_1 F_2 F_3	F_1 F_2 F_3	F_1 F_2 F_3
0 1 1	0 0 1	0 0 1	0 0 0
1 0 1	0 1 0	0 1 0	0 1 2
1 1 0	1 0 0	1 0 0	0 2 1
1 1 1	1 1 1	1 1 1	1 0 2
1 1 2	1 2 2	1 1 2	1 2 0
1 2 1	2 2 1	1 2 1	2 0 1
2 1 1	2 1 2	2 1 1	2 1 0
		2 2 2	2 2 2

To every design corresponds an information matrix Q for the linear and quadratic contrasts. The information matrix has the form, [Pericleous at al (2012,2013)]:

$$\text{If } s = 0, \quad Q = \begin{bmatrix} Q_1 & Q_{12} \\ \text{sym} & Q_2 \end{bmatrix},$$

$$\text{if } 0 < s < k, \quad Q_1 = \begin{bmatrix} Q_{1,11} & Q_{1,12} \\ \text{sym} & Q_{1,22} \end{bmatrix}, \quad Q_2 = \begin{bmatrix} Q_{2,11} & Q_{2,12} \\ \text{sym} & Q_{2,22} \end{bmatrix}.$$

Q_1 Q_2 are $k \times k$ matrices, Q_1 refers to linear and Q_2 refers to quadratic contrasts. The definition of the $BA(N,k,3,2)$ guarantee that $Q_{12} = \mathbf{0}$.

For the elements of the $k \times k$ matrices Q_1 Q_2 see [Pericleous at al (2012,2013)].

For every N and k there are more than one $BA(N,k,3,2)$, to decide which $BA(N,k,3,2)$ is the optimal design we find the latent roots $\lambda_1, \dots, \lambda_k, \mu_1, \dots, \mu_k$ of Q_1, Q_2 respectively and calculate the values of the A and D criteria, that is,

$$C: \frac{1}{\lambda_1} + \dots + \frac{1}{\lambda_k} + \frac{1}{\mu_1} + \dots + \frac{1}{\mu_k}, D: \prod_{i=1}^k (\lambda_i) \cdot \prod_{i=1}^k (\mu_i) \quad \text{respectively. The design}$$

minimizing C is A -optimal and the design maximizing D is D -optimal. In most of the cases the same design is A and D optimal.

In Table1 if $N=7$ the design (b) is A and D optimal and for $N=8$ the design (c) is A and D optimal.

2.1 Construction of the BA(N,k,3,2) for $N \equiv 1, 2 \pmod{3}$, $7 \leq N \leq 32$.

From (1) the number of parameters is $1 + 2k$ and the number of observations is N , then $k \leq (N - 1) / 2$, this gives the maximum number of factors that can be accommodated in a BA(N,k,3,2).

For $N=9m$ there exist OA, for $N=9m+1$ the A and D optimal design for all values of k is constructed by adjoining a row of $\underbrace{1 \dots 1}_k$ to an OA(N-1,k,3,2), see Table 2a.

For $N=11$ the D- optimal design is the BA(11,4,3,2), $4=3+1$, given in Table 2b. This consists of two blocks, the first block has the first 3 columns from the BA with $n_1^1 = 3$, $n_{11}^{11} = 1$, the second block consists of the last column with $n_2^1 = 5$, $n_{22}^{11} = 3$, $n_{12}^{11} = 1$. For $k=2$ take any 2 from the first three columns, for $k=3$ take two columns from the first three and the last column, for $k=4$ take the four columns in Table (2b). For $N = 11, k = 5$, the D-optimal design has not being constructed yet.

Table 2a BA(10,4,3,2)

$$n_1^1 = 4, n_{11}^{11} = 2$$

0	0	0	0
0	1	1	2
0	2	2	1
1	0	2	2
1	1	0	1
1	2	1	0
2	0	1	1
2	1	2	0
2	2	0	2
1	1	1	1

Table 2b BA(11,4,3,2)

$$n_1^1 = 3, n_{11}^{11} = 1, n_2^1 = 5, n_{12}^{11} = 1$$

0	0	1	0
1	1	2	0
2	2	0	0
0	1	0	1
2	2	2	1
0	0	0	1
1	2	1	1
2	0	2	1
1	0	0	2
2	1	1	2
0	2	2	2

The D optimal designs for $N=13,14,16,17,20$ were constructed using an algorithm for each case. In Table 3 we give the D optimal design for $N = 16, n^1 = 6, n^{11} = 2$.

Most of the designs are constructed using an algorithm and few are constructed by adjoining to an OA some rows.

Table 3. BA (16,7,3,2)

$$n^1 = 6, n^{11} = 2$$

0	0	0	0	0	0	0
0	0	0	0	1	1	1
0	1	1	2	0	1	2
0	1	2	1	2	2	1
0	2	1	1	1	0	0
1	0	1	2	2	0	1
1	0	2	1	0	1	0
1	1	2	0	1	0	2
1	1	0	2	1	2	0
1	2	1	0	0	2	1
1	2	0	1	2	1	2
2	0	1	1	1	2	2
2	1	1	0	2	1	0
2	1	0	1	0	0	1
2	2	2	2	2	2	2
2	2	2	2	1	1	1

We give here an algorithm, the other algorithms follow similar lines. This is the algorithm for the construction of the optimal BA(16,7,3,2) with $n^1 = 6, n^{11} = 2$, then $n^0 = n^2 = 5$,

$$n^{01} = n^{10} = n^{12} = n^{21} = (6 - 2) / 2 = 2, n^{00} + n^{02} = 3, n^{00} = n^{22} = 2, n^{02} = n^{20} = 1 \quad (2)$$

Step 1: The first column has $n^0 = n^2 = 5, n^1 = 6$

Step 2: The second column is (0 0 1 1 2 0 0 1 1 2 2 0 1 1 2 2) transposed.

Step 3: The remaining columns have in the first 5 rows two 0, two 1 and one 2, in the next six rows two 0, two 1, two 2 and in the last 5 rows one 0, two 1 and two 2. In

total there are $\binom{5}{2}^2 \times \binom{3}{1}^2 \times \binom{6}{2} \times \binom{4}{2} = 81000$ columns, this is set S1.

Step 4: Find with how many columns of S1, the second column satisfies relations (2). This is set S11.

Step 5: Take one column from S11 as the third column.

Step 6: The third column with how many columns of S11 satisfy relations 2? This is set S111.

Step 7: Take a column from S111 as fourth column.

Step 8: Continue until the set S111... is the empty set. If you have $(16-2)/2=7$ columns you have finished, if not go to the previous step and choose another column and go to step 7.

Step 9: Continue until all cases are exhausted.

3. CONSTRUCTION OF NEW ORTHOGONAL ARRAYS.

Definition 3.1 The $OA(N,k,3,2)$ is an $N \times k$ matrix with elements 0,1,2 and the properties,

- (i) Every column has every one of 0,1, 2, appearing $N/3$ times,
- (ii) Every two columns have each of the pairs 00, 01, 02, 10, 11,12, 20, 21, 22 appearing $N/9$ times.

Example 3.1 The first 9 rows of the table 2a is an $OA(9,4,3,2)$.

The maximum number $k_{\max}$ of columns satisfies $k_{\max} \leq (N-1)/2$ because if the $OA(N,k,3,2)$ is used as a design in a linear model of resolution III (second and higher order interactions are negligible), then the parameters are $2k+1$, hence $2k+1 \leq N$.

In this section we create new Orthogonal Arrays from a given $OA(N,k,3,2)$.

If A is an $OA(N,k,3,2)$, then B is formed from A with $0 \rightarrow 1 \rightarrow 2 \rightarrow 0$ and C is formed from A with $0 \rightarrow 2 \rightarrow 1 \rightarrow 0$. This is equivalent to adding 1 modulo 3 to all elements of A and adding 2 modulo 3 to every element of A.

Definition 3.2 $Q_1 : OA(N = 3,1,3,2) = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$, then $A = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$, $B = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}$, $C = \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix}$.

Construction 1: If A is the orthogonal array $Q_N : OA(N, k_1, 3, 2)$ and B, C are as defined above,

$$Q_{3N} = \left[\begin{array}{ccc|c} A & A & A & 0 \\ A & B & C & 1 \\ A & C & B & 2 \end{array} \right]$$

is an orthogonal array $Q_{3N} : OA(3 \times N, 3 \times k_1 + 1, 3, 2)$.

If $N = 3^n$, then $k_1 = k_{\max} = (3^n - 1) / 2$.

In this way there are constructed all Q_n $n = 1, 2, \dots$, that is, $OA(9, 4, 3, 2)$, $OA(27, 13, 3, 2), \dots$

Construction 2. If A is an orthogonal array $Q_N : OA(N, k_1, 3, 2)$ and B and C are as above,

$$Q_{6N} = \left[\begin{array}{cccccc|c} A & A & A & A & A & A & 0 \\ A & B & C & B & A & C & 0 \\ A & C & C & A & B & B & 1 \\ A & B & A & C & C & B & 1 \\ A & A & B & C & B & C & 2 \\ A & C & B & B & C & A & 2 \end{array} \right] \quad (3)$$

is an $OA(6N, 6k_1 + 1, 3, 2)$

In this way are constructed: $OA(18, 7, 3, 2)$, $OA(54, 25, 3, 2)$, $OA(108, 43, 3, 2), \dots$

Similarly are constructed orthogonal arrays: $OA(9N, 9k_1 + 1, 3, 2)$, $OA(12N, 12k_1 + 1, 3, 2)$, $OA(15N, 15k_1 + 1, 3, 2), \dots$

The proofs are omitted.

3.1 The D optimal design for N=18 observations and k=8 factors

In fractional factorial designs with N observations and k factors, each at three levels, when $N \equiv 0 \pmod{9}$ the ϕ -optimal design is the $OA(N, k, 3, 2)$ provided this OA exists. In the case $N=18$ the maximum value of k for estimating linear and quadratic contrasts is $\lfloor (N-1)/2 \rfloor = 8$. It is proved in Theorem 3.1 below that the $OA(18, k, 3, 2)$ has $k_{\max} = 7$ and such $OA(18, 7, 3, 2)$ exists.

The optimal $BA(18, 8, 3, 2)$, $8=r+s$, with $s=0$ is the $OA(18, 8, 3, 2)$ which does not exist. The D optimal BA has $s=2$ and is the following. The proof is omitted.

Table 4. BA(18,8,3,2), r=6,s=2

0	0	0	0	0	0	0	0
0	0	1	1	1	1	0	1
0	1	1	2	0	2	1	0
0	1	2	1	2	0	1	2
0	2	0	2	2	1	2	1
0	2	2	0	1	2	2	2
1	1	0	0	2	2	0	1
1	1	2	2	1	1	0	0
1	2	0	1	1	0	1	0
1	2	1	0	0	1	1	2
1	0	1	2	2	0	2	2
1	0	2	1	0	2	2	1
2	2	1	1	2	2	0	0
2	2	2	2	0	0	0	1
2	0	0	2	1	2	1	2
2	0	2	0	2	1	1	0
2	1	0	1	0	1	2	2
2	1	1	0	1	0	2	1

There are 8 non equivalent such designs and the determinant of the information matrix equals 1.586874×10^{12} .

Theorem 3.1 In the OA(18,k,3,2) the maximum value of k is $k_{\max} = 7$.

Proof. The OA(18,k,3,2) has in every column six 0, and in every pair of columns the number of 00 is two. Also in the first line can put k zero. If k=8 and there are x_i , $i = 0, \dots, 8$ rows with i zero, then in the last 17 rows we have: the number of 0 is

$$5 \times 8 = 40, \quad \text{the number of 00 is } \binom{8}{2} \times 1 = 28, \quad \text{hence}$$

$\sum_{i=0}^8 x_i = 17, \sum_{i=0}^8 ix_i = 40, \sum_{i=0}^8 i(i-1)x_i / 2 = 28$. If c is an integer, write the last relation as follows,

$$\sum_{i=0}^8 (i-c)(i-1-c)x_i = \sum_{i=0}^8 i(i-1)x_i - c \sum_{i=0}^8 (2i-1)x_i + c^2 \sum_{i=0}^8 x_i = 56 - 63c + 17c^2$$

Since the product of two successive integers is non negative we have,

$56 - 63c + 17c^2 \geq 0$. This is minimized for $c = 2$, hence

$56 - 63 \times 2 + 17 \times 2^2 = -2$, this is a contradiction, hence $k_{\max} \leq 7$.

ΠΕΡΙΛΗΨΗ

Σε 3^k κλασματικούς σχεδιασμούς, όπου υπάρχουν k παράγοντες, ο καθένας σε τρεις στάθμες, το ενδιαφέρον είναι στην εκτίμηση των γραμμικών και τετραγωνικών αντιθέσεων των επιδράσεων των παραγόντων. Εργαζόμαστε με σχεδιασμούς αναλυτικής τάξης III, όπου αλληλεπιδράσεις δύο ή περισσότερων παραγόντων είναι αμελητέες. Αν το πλήθος των παρατηρήσεων είναι $N \equiv 0 \pmod{9}$, ο φ- βέλτιστος σχεδιασμός είναι ένας ορθογώνιος σχηματισμός με τρία σύμβολα και ισχύ ίση με 2, $OA(N, k, 3, 2)$. Εισάγεται η έννοια των ισορροπημένων σχηματισμών και δημιουργούμε αλγόριθμους για βέλτιστους σχεδιασμούς για πολλές τιμές του N .

Για $N=18$ το $OA(18, k, 3, 2)$ υπάρχει για $k=2, 3, \dots, 7$ αλλά όχι για η μέγιστη τιμή του $k=8$. Δίνουμε το D-βέλτιστο σχεδιασμό για $N=18$ και $k=8$ παράγοντες.

REFERENCES

- Bush, K. A. 1952. Orthogonal arrays of index unity. *Ann. Math. Statist.* 23: 426–434.
- Chakravarti I.M. (1961). On some methods of construction of partially balanced arrays. *Ann. Math. Statist.* 32:1181–1185.
- Hedayat A. S., Sloane N.J.A., Stufken J. (1999). *Orthogonal arrays: theory and applications*. Springer, Berlin Heidelberg New York.
- Kolyva-Machera F. (1989). D-Optimality in 3^k Designs for $M \equiv 1 \pmod{9}$ observations. *J. Statist. Plann. Inference* 22, 95-103.
- Kiefer J. (1961). Optimum experimental designs V, with applications to systematic and rotatable designs. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* 1, 381405. Univ. California Press.
- Mukerjee R. (1999). On the optimality of orthogonal array plus one run plans *Ann. Statist.* 27:1256–1271.

Mukhopadhyay, A. C. (1981). *Construction of some series of orthogonal arrays*. *Sankhya B* 43: 81–92.

Περικλέους Κ., Χατζόπουλος Στ., Κολυβά-Μαχαίρα Φ., Κουνιάς Σ. 3^k Βέλτιστοι σχεδιασμοί για την εκτίμηση Γραμμικών και Τετραγωνικών αντιθέσεων, όταν $N \equiv 0 \pmod{3}$. *Πρακτικά 25^{ου} Πανελλήνιου Συνεδρίου Στατιστικής ΕΣΙ 2012*, σελ. 179-187.

Περικλέους Κ., Χατζόπουλος Στ., Κολυβά-Μαχαίρα Φ., Κουνιάς Σ. 3^k Βέλτιστοι σχεδιασμοί για την εκτίμηση Γραμμικών και Τετραγωνικών αντιθέσεων, όταν $N \equiv 0 \pmod{3}$. *Πρακτικά 26^{ου} Πανελλήνιου Συνεδρίου Στατιστικής ΕΣΙ 2013*, σελ. 180-189.



ANALYSIS ON THE GENDER FACTOR AND THE ACADEMIC PERFORMANCE OF STUDENTS OF THE FACULTY OF SCIENCES OF ARISTOTLE UNIVERSITY OF THESSALONIKI

St. A. Chatzopoulos, F. Kolyva – Machera, H. Charalambous

School of Mathematics, Aristotle University of Thesssaloniki

{cstavros, fkolyva, hara}@math.auth.gr

ABSTRACT

We study the academic performance of the student population of the Faculty of Sciences (F.Sc.) of Aristotle University of Thessaloniki (A.U.TH), Greece, in terms of the gender factor. In particular, we study the degree duration, the academic ranking and the success examination rate of the students of the Schools of Biology, Chemistry, Geology, Informatics, Mathematics and Physics. Moreover, we performed an analysis on the data provided by the Quality Assurance Unit (Q.A.U.) of A.U.TH.

Keywords: Chi-Square tests, Z-tests.

1. INTRODUCTION

For the present study we use the data of ten academic years (from 2005-06 to 2015-16). The data was provided by Q.A.U. for the six schools of the F.Sc. of A.U.TH. First of all, the data included, for each school and over the course of the decade, the totals for the entering classes (called inputs) and for the graduates. The mean degree duration and the corresponding standard deviation were provided while the graduates were categorized into three main categories: one category for those who obtained their degree in 4 years, one for those who needed 5–6 years to graduate and lastly all who received their degree after 6 years (> 6). In terms of their grade point average (GPA), the mean and standard deviation were provided and the graduates were categorized into three main categories: “Good” (5.0 – 6.49), “Very Good” (6.5 – 8.49) and “Excellent” (8.5-10). Finally the data included the overall number of the students enrolled in a course (and thus being eligible to participate in the examination of the course), the overall number of the students that participated in the examination and the success rate. Since the data was already tabulated, we chose to analyze it using simple statistical analysis, such as Chi-Square tests and z-tests of means. We also performed an

analysis for the quantitative characteristics of the graduates, namely the degree duration and their GPA.

The article is structured as follows: in Section 2, we present the data and the results of the descriptive statistical analysis in relation to gender. The variables we consider are the following: inputs, number of graduates, categories of degree duration together with mean and standard deviation, categories of GPA together with mean and standard deviation and the categories for the examinations, meaning the enrollment, the participation and the success rate. In Section 3, we present the results of the analysis for the graduates of all schools concerning the duration of degree and their GPA. In Section 4, we study the ratios of females versus males for the various variables and compare it to the original ratios of the incoming classes. In Section 5 we compare the results of Sections 2 and 4.

A p-value < 0.05 was considered significant in all statistical procedures. We used the Statistical Package for Social Sciences (IBM SPSS Statistics for Windows, Version 22.0, Armonk, NY: IBM Corp) and the Review Manager (RevMan Computer program, Version 5.3. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2014).

2. DESCRIPTIVE STATISTICS AND RESULTS

Tables 1-6 contain the data provided by Q.A.U. for each school of the F.Sc. The data is organized according to the variables described in the introduction. Following each table, we present the results of the statistical analysis.

Table 1: *School of Biology (2005–2015)*

variable	levels	Biology		
		Female	Male	Total
inputs		949	455	1,404
number of graduates		785	286	1071
degree duration	4 years	63	23	86
	5 – 6 years	524	161	685
	> 6 years	198	102	300
	mean	5.59	6.11	5.73
	st. dev.	0.72	0.84	0.79
GPA	good	126	56	182
	very good	638	213	851
	excellent	21	17	38
	mean	7.13	7.12	7.13
	st. dev.	0.25	0.29	0.25
examinations	enrollment	81,227	46,662	127,889
	participation	32,800	13,650	46,450
	success	25,718	10,539	36,257

For the **School of Biology**, the statistical analysis revealed that:

- (i) Gender and degree duration are independent ($\chi^2(2) = 11.744$, $p = 0.000 < 0.05$). Applying Bonferroni correction we get that in the category degree duration of 5 – 6 years, the percentage of female graduates is higher than the percentage of males ($p = 0.001 < 0.017$), while in the category degree duration of more than 6 years, the percentage of males is higher than the percentage of females ($p = 0.000 < 0.017$).
- (ii) For the graduates, the average duration of study is greater for males ($z = 9.298$, $p = 0.000 < 0.05$).
- (iii) There is no statistically significant difference in the average GPA between genders ($z = 0.517$, $p = 0.605 > 0.05$).
- (iv) Gender and categorization of the graduates' GPA are independent ($\chi^2(2) = 9.069$, $p = 0.011 < 0.05$). Applying Bonferroni correction we get that in the “Very Good” category (6.5-8.49), the percentage is higher for the females ($p = 0.007 < 0.017$), while in the “Excellent” category (8.5-10.0) the percentage is higher for the males ($p = 0.005 < 0.017$).
- (v) The percentage of participation in the examinations is higher for females ($z = 39.833$, $p = 0.000 < 0.05$).
- (vi) The percentage of success in the examinations is higher for females ($z = 2.846$, $p = 0.002 < 0.05$).

Table 2: *School of Chemistry (2005–2015)*

variable	levels	Chemistry		
		Female	Male	Total
inputs		1,220	576	1,796
number of graduates		870	422	1292
degree duration	4 years	61	22	83
	5 – 6 years	462	165	627
	> 6 years	347	235	582
	mean	6.41	7.25	6.68
	st. dev.	0.81	1.00	0.92
GPA	good	319	201	520
	very good	539	219	758
	excellent	12	2	14
	mean	6.75	6.61	6.70
	st. dev.	0.36	0.32	0.35
examinations	enrollment	164,115	149,507	313,622
	participation	51,497	26,039	77,536
	success	35,410	16,496	51,906

For the **School of Chemistry**, the statistical analysis revealed that:

- (i) Gender and degree duration are independent ($\chi^2(2) = 28.666$, $p = 0.000 < 0.05$). Applying Bonferroni correction we get that in the subcategory degree duration of 5 – 6 years, the percentage of females is higher than the percentage of males ($p = 0.000 < 0.017$), while in the subcategory degree duration of more than 6 years, the percentage of males is higher than the percentage of females ($p = 0.000 < 0.017$).
- (ii) For the graduates, the average duration of study is greater for males ($z = 15.029$, $p = 0.000 < 0.05$).
- (iii) The average GPA is higher for the females ($z = 7.074$, $p = 0.000 < 0.05$).
- (iv) Gender and categorization of the graduates' GPA are independent ($\chi^2(2) = 15.537$, $p = 0.000 < 0.05$). Applying Bonferroni correction we get that in the “Good” category (5.0-6.49), the percentage is higher for the males ($p = 0.007 < 0.017$), while for the “Very Good” category (6.5-8.49), the percentage is higher for the females ($p = 0.007 < 0.017$).
- (v) The percentage of participation in the examinations is higher for females ($z = 90.525$, $p = 0.000 < 0.05$).
- (vi) The percentage of success in the examinations is higher for females ($z = 15.125$, $p = 0.000 < 0.05$).

Table 3: *School of Geology (2005–2015)*

variable	levels	Geology		
		Female	Male	Total
inputs		641	624	1265
number of graduates		393	379	772
degree duration	4 years	20	24	44
	5 – 6 years	153	126	279
	> 6 years	220	229	449
	mean	6.81	7.46	7.13
	st. dev.	0.96	1.25	1.10
GPA	good	199	217	416
	very good	192	160	352
	excellent	2	2	4
	mean	6.59	6.50	6.54
	st. dev.	0.32	0.30	0.31
examinations	enrollment	55,613	63,003	11,8616
	participation	29,106	30,095	59,201
	success	19,842	18,318	38,160

For the **School of Geology**, the statistical analysis revealed that:

- (i) Gender and degree duration are not independent ($\chi^2(2) = 2.904, p = 0.234 > 0.05$).
- (ii) For the graduates, the average duration of study is greater for males ($z = 8.082, p = 0.000 < 0.05$).
- (iii) The average GPA is higher for the females ($z = 4.033, p = 0.000 < 0.05$).
- (iv) Gender and categorization of the graduates' GPA are not independent ($\chi^2(2) = 3.435, p = 0.179 > 0.05$).
- (v) The percentage of participation in the examinations is higher for females ($z = 15.706, p = 0.000 < 0.05$).
- (vi) The percentage of success in the examinations is higher for females ($z = 18.563, p = 0.000 < 0.05$).

Table 4: *School Informatics (2005–2015)*

variable	levels	Informatics		
		Female	Male	Total
inputs		340	1,079	1,419
number of graduates		316	654	970
degree duration	4 years	67	103	170
	5 – 6 years	183	333	516
	> 6 years	66	218	284
	mean	5.49	5.92	5.78
	st. dev.	0.75	0.84	0.82
GPA	good	67	139	206
	very good	240	473	713
	excellent	9	42	51
	mean	7.05	7.09	7.07
	st. dev.	0.25	0.28	0.27
Examinations	enrollment	27,595	88,257	115,852
	participation	15,466	43,326	58,792
	success	10,877	28,772	39,649

For the **School of Informatics**, the statistical analysis revealed that:

- (i) Gender and degree duration are independent ($\chi^2(2) = 16.849, p = 0.000 < 0.05$). Applying Bonferroni correction we get that in the category degree duration of in the category degree duration of more than 6 years, the percentage of males is higher than the percentage of females ($p = 0.000 < 0.017$).

- (ii) For the graduates, the average duration of study is greater for males ($z = 8.042, p = 0.000 < 0.05$).
- (iii) The average GPA is higher for the males ($z = 2.24, p = 0.025 < 0.05$).
- (iv) Gender and categorization of the graduates' GPA are not independent ($\chi^2(2) = 5.557, p = 0.062 > 0.05$).
- (v) The percentage of participation in the examinations is higher for females ($z = 20.172, p = 0.000 < 0.05$).
- (vi) The percentage of success in the examinations is higher for females ($z = 8.931, p = 0.000 < 0.05$).

Table 5: *School of Mathematics (2005–2015)*

variable	levels	Mathematics		
		Female	Male	Total
inputs		1,315	1,318	2,641
number of graduates		1,081	920	2,001
degree duration	4 years	180	110	290
	5 – 6 years	542	340	882
	> 6 years	359	470	829
	mean	6.07	7.25	6.61
	st. dev.	1.17	1.65	1.47
GPA	good	694	596	1,290
	very good	365	278	643
	excellent	22	46	68
	mean	6.43	6.44	6.44
	st. dev.	0.33	0.39	0.36
Examinations	enrollment	143,965	132,589	276,554
	participation	53,298	63,579	116,877
	success	30,795	37,158	67,953

For the **School of Mathematics**, the statistical analysis revealed that:

- (i) Gender and degree duration are independent ($\chi^2(2) = 65.492, p = 0.000 < 0.05$). Applying Bonferroni correction we get that in the subcategories degree duration of 4 and 5 – 6 years the percentage of females is higher than the percentage of males ($p = 0.000, 0.000 < 0.017$), while in the category degree duration of more than 6 years, the percentage of males is higher than the percentage of females ($p = 0.000 < 0.017$).
- (ii) For the graduates, the average duration of study is greater for males ($z = 18.153, p = 0.000 < 0.05$).

- (iii) There is no statistically significant difference in the average GPA between genders ($z = 0.613$, $p = 0.540 > 0.05$).
- (iv) Gender and categorization of the graduates' GPA are independent ($\chi^2(2) = 14.829$, $p = 0.001 < 0.05$). Applying Bonferroni correction we get that in the "Excellent" category (8.5-10.0), the percentage is higher for males ($p = 0.000 < 0.017$).
- (v) The percentage of participation in the examinations is higher for males ($z = 58.133$, $p = 0.000 < 0.05$).
- (vi) The percentage of success in the examinations is higher for females ($z = 2.295$, $p = 0.011 < 0.05$).

Table 6: *School of Physics (2005–2015)*

variable	levels	Mathematics		
		Female	Male	Total
inputs		857	1,433	2,299
number of graduates		650	1,076	1,726
degree duration	4 years	28	41	69
	5 – 6 years	252	314	566
	> 6 years	370	721	1091
	mean	7.11	7.76	7.52
	st. dev.	1.01	1.28	1.19
GPA	good	183	336	519
	very good	453	694	1,147
	excellent	14	46	60
	mean	6.87	6.91	6.89
	st. dev.	0.41	0.44	0.43
Examinations	enrollment	110,271	235,095	348,938
	participation	44,136	74,072	122,567
	success	26,967	45,381	75,141

For the **School of Physics**, the statistical analysis revealed that:

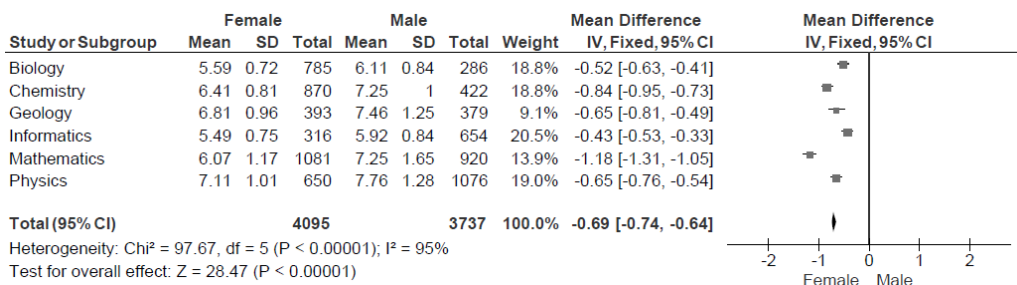
- (i) Gender and degree duration are independent ($\chi^2(2) = 18.127$, $p = 0.000 < 0.05$). Applying Bonferroni correction we get that in the subcategory degree duration of 5–6 years the percentage of females is higher than the percentage of males ($p = 0.000 < 0.017$), while in the category degree duration of more than 6 years, the percentage of males is higher than the percentage of females ($p = 0.000 < 0.017$).
- (ii) For the graduates, the average duration of study is greater for males ($z = 11.689$, $p = 0.000 < 0.05$).

- (iii) There is no statistically significant difference in the average GPA between genders ($z = 1.91, p = 0.056 > 0.05$).
- (iv) Gender and categorization of the graduates' GPA are independent ($\chi^2(2) = 8.163, p = 0.017 < 0.05$). Applying Bonferroni correction we get that in the "Very Good" category (6.5-8.49) the percentage is higher for females ($p = 0.013 < 0.017$), while in the "Excellent" category (8.5-10.0), the percentage is higher for males ($p = 0.010 < 0.017$).
- (v) The percentage of participation in the examinations is higher for females ($z = 49.185, p = 0.000 < 0.05$).
- (vi) The percentage of success in the examinations does not statistically differ relative to gender ($z = 0.567, p = 0.285 > 0.05$).

3. ANALYSIS OF THE DURATION OF DEGREE AND GPA

In this section we present the results of the analysis for the graduates of all schools concerning the duration of degree and their GPA. Figure 1 gives the results of the analysis performed for the degree duration between genders. The forest plots display effect estimates and confidence intervals for the schools of the F.Sc. of A.U.TH. and for the total. The area of the block indicates the weight assigned to the study in the analysis while the horizontal line depicts the 95% confidence interval.

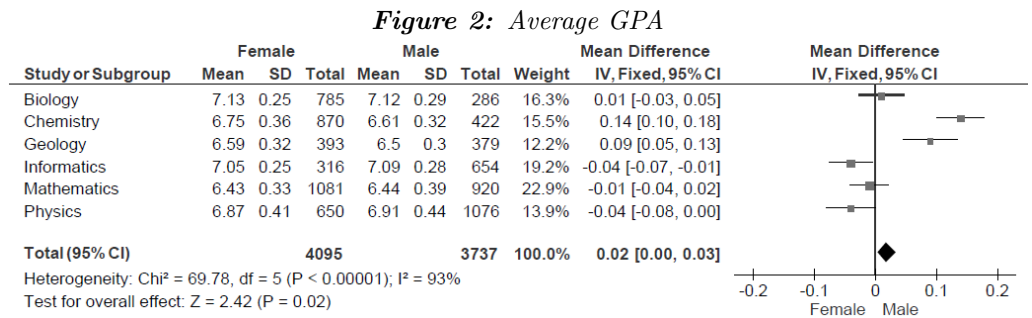
Figure 1: Degree duration for the graduates of the schools of the Faculty of Sciences.



The classical measure of heterogeneity is Cochran's Q statistic, which reflects the observed dispersion. The I^2 statistic describes the percentage of variation across studies that is due to heterogeneity rather than chance. Our analysis shows that $I^2 = 95\%$, which means that 95% of the observed variance between studies is due to real differences in the effect size. Only about 5% of the observed variance would have been expected based on random error. Moreover, $Z = 28.47, p = 0.000 < 0.05$, that is the null hypothesis that there is no overall effect is rejected. Finally, as becomes clear from Figure 1, even though the analysis is statistically significant, there is significant heterogeneity.

Figure 2 gives the results of the analysis performed for the average GPA be-

tween genders. Our analysis shows that $I^2 = 93\%$, which means that 93% of the observed variance between studies is due to real differences in the effect size. Only about 7% of the observed variance would have been expected based on random error. Moreover, $Z = 2.42$, $p = 0.02 < 0.05$, that is the null hypothesis that there is no overall effect is rejected.



4. RATIO ANALYSIS

In order to weight the results of our research we compute and compare the ratio females/males for the variables of Section 2. We call the ratio for the incoming students *iratio* and compare the rest of the ratios to it. Whenever the *iratio* is higher, an increased male presence is detected for the parameter. On the other hand, if the *iratio* is lower, then an increased female presence is detected for the parameter. Table 7 gives the results of our analysis. The abbreviations “G”, “VG”, “E” stand for “Good”, “Very Good” and “Excellent” respectively. When the symbol ** is placed next to the ratio, the significance is less than 0.01, while the symbol “xx” placed next to the ratio means that the test was not feasible. We note that for the Schools of Biology, Chemistry and Geology the tests were performed on the reverse ratios (i.e. males/females).

Table 7: Ratios and Significances (females/males) for all the Schools of the Faculty of Sciences of A.U.TH.

variable	levels	BIOL	CHEM	GEOL	INFO	MATH	PHYS
iratio		2.088	2.118	1.028	0.315	0.998	0.598
degree duration # years	4	2.739	2.773	0.833**	0.650**	1.636**	0.683
	5 – 6	3.255**	2.800**	1.214**	0.550**	1.594**	0.803**
	>6	1.941	1.477**	0.961**	0.303	0.764**	0.513**
GPA	G	2.250	1.587**	0.917**	0.482**	1.164**	0.545**
	VG	2.995**	2.461**	1.200**	0.507**	1.313**	0.653**
	E	1.235**	6.000 xx	1.000 xx	0.214	0.478**	0.304**
exam participation		2.403**	1.978**	0.967**	0.357**	0.838**	0.596
exam succes		2.440**	2.147**	1.083**	0.378**	0.829**	0.594

We summarize the conclusions of Table 7:

- (i) For the School of Biology we conclude that the iratio is higher in the category “Excellent” GPA and lower in the categories of degree duration of 5–6 years, “Very Good” GPA, exam participation and success in examinations.
- (ii) For the School of Chemistry we conclude that the iratio is higher in the categories degree duration of more than 6 years, “Good” GPA and participation in the examinations, while it is lower in the categories degree duration of 5 – 6 years, “Very Good” GPA and success in the examinations.
- (iii) For the School of Geology we conclude that the iratio is higher in the categories degree duration of 4 and more than 6 years, “Good” GPA and exam participation, while it is lower in the categories degree duration of 5 – 6 years, “Very Good” GPA and success in examinations.
- (iv) For the School of Informatics we conclude that the iratio is lower in the categories degree duration of 4 and 5 – 6 years, “Good” and “Very Good” GPA, exam participation and success.
- (v) for the School of Mathematics we conclude that the iratio is lower in the categories degree duration of 4 and 5 – 6 years and in the categories of “Good” and “Very Good” GPA, while it is higher in the categories degree duration of more than 6 years, in the exam participation and success and in the category of “Excellent” GPA.
- (vi) for the School of Physics we conclude that the iratio is lower in the category degree duration of 5 – 6 years and in the category of “Very Good” GPA. The iratio is higher in the category degree duration of more than 6 years and in the categories of ‘Good’ and “Excellent” GPA.

5. DISCUSSION

In this section, we compare the results of the previous sections and report on the conclusions for each school separately.

- (i) For the School of Biology, no difference between the results of Sections 2 and 4 was observed.
- (ii) For the School of Chemistry, in the category exam participation the ratio is lower than the iratio, even though no statistically significant difference due to gender was observed.
- (iii) For the School of Geology, the ratio is lower from the iratio in the degree duration categories of 4 and more than 6 years and “Good” GPA, even though no statistically significant difference due to gender was observed. In the categories 5 – 6 years degree duration and “Very Good” GPA the ratio is higher from the iratio, even though no statistically significant difference due to gender was observed. Moreover in the category exam participation,

the ratio is higher from the iratio while in section 2 we observed that the percentage of females is higher.

- (iv) For the School of Informatics, the ratio is higher than the iratio in the categories of 4 and 5 – 6 years degree duration, “Good” and “Very Good” GPA while in Section 2 no statistically significant difference due to gender was observed. Moreover, in the category of more than 6 years, the ratio is lower from the iratio while in Section 2 we observed that the percentage of males is higher.
- (v) For the School of Mathematics, in the categories of “Good” and “Very Good” GPA the ratio is higher than the iratio while in Section 2 we did not notice a difference relative to gender. In the category succes in the examinations, the percentage of female was higher than the percentage of males while the ratio is lower than the iratio.
- (vi) For the School of Physics, in the category of “Good” GPA the ratio is lower than the iratio even though we did not notice a difference relative to gender.

We assemble our conclusions in Tables 8 and 9. Table 8 depicts (for each variable) the gender with increased presence suggested by the analysis of Section 2, while Table 9 does the same using the analysis of Section 4. We leave blank whenever no conclusion could be drawn. The abbreviations “F” and “M” stand for “Female” and “Male”, respectively.

Table 8: *Academic performance and gender in a glance. Results from Section 2.*

		BIOL	CHEM	GEOL	INFO	MATH	PHYS
degree	4					F	
duration	5-6	F	F			F	F
# years	>6	M	M		M	M	M
duration of degree		M	M	M	M	M	M
GPA score			F	F	M		
GPA	good		M				
	very good	F	F				F
	excellent	M				M	M
exam participation		F	F	F	F	M	F
exam success		F	F	F	F	F	

Table 9: Academic performance and gender in a glance. Results from Section 4.

		BIOL	CHEM	GEOL	INFO	MATH	PHYS
degree	4			M	F	F	
duration	5-6	F	F	F	F	F	F
# years	>6		M	M		M	M
GPA	good		M	F	M	F	M
	very good	F	F	F	F	F	F
	excellent	M				M	M
exam participation		F	M	M	F	M	
exam success		F	F	F	F	M	

ΠΕΡΙΛΗΨΗ

Στην εργασία μελετώνται οι επιδόσεις των φοιτητών των έξι τμημάτων της Σχολής Θετικών Επιστημών του Α.Π.Θ., δηλαδή των τμημάτων: Βιολογίας, Γεωλογίας, Μαθηματικών, Πληροφορικής, Φυσικής και Χημείας. Παρουσιάζονται στοιχεία σχετικά με τη διάρκεια φοίτησης, την κατάταξη βαθμολογίας, τον μέσο όρο του βαθμού πτυχίου, την συμμετοχή και την επιτυχία στις εξετάσεις. Αυτές οι παράμετροι αναλύονται σε σχέση με το φύλο. Να σημειωθεί ότι για τις στατιστικές αναλύσεις που πραγματοποιήθηκαν χρησιμοποιήθηκαν τα ήδη επεξεργασμένα στοιχεία της Μονάδας Διασφάλισης Ποιότητας (ΜΟ.ΔΙ.Π.) του Α.Π.Θ.

Acknowledgements: This research is part of the initiatives undertaken by the Dean and the Committee for Gender Equality and Equity of the Faculty of Sciences of A.U.TH.

REFERENCES

- Abdi, H. (2007). Bonferroni and Sidak corrections for multiple comparisons. *In Salkind, N. J. Encyclopedia of Measurement and Statistics*, CA: Sage.
- Field, A. P. (2005). *Discovering Statistics Using SPSS (2nd edition)*, London: Sage Publications.
- Norusis, M. J. (2011). *IBM SPSS Statistics 19 Statistical Procedures Companion (1st edition)*. Upper Saddle River, N.J. Prentice Hall.

SEISMIC HAZARD ASSESSMENT OF EARTHQUAKES IN GREECE USING DIRECTED EARTHQUAKE NETWORKS

D. Chorozoglou¹, D. Kugiumtzis², E. Papadimitriou¹

¹Department of Geophysics, Aristotle University of Thessaloniki,
{chorozod, ritsa}@geo.auth.gr,

²Department of Electrical and Computer Engineering, Aristotle University of
Thessaloniki,
dkugiu@auth.gr.

ABSTRACT

The objective of this study is the seismic hazard evaluation in the area of Greece based on modeling of the transition probabilities of seismicity as a Markov and semi-Markov chain model. The data consist of strong earthquakes with magnitude $M_{thr} \geq 5.5$ that occurred during the period 1911-2015 are considered. The study area is divided into 5 subareas (seismic zones) that are homogenous from the seismotectonic point of view and the seismic catalog is divided into subsets for four magnitude ranges. Two Markov chains are defined with respect to predefined time window, one having as states the occurrence or not of strong earthquakes in any of the 5 subareas ($2^5 = 32$ states), and one having as states the occurrence or not of strong earthquakes anywhere in Greece at any of the four magnitude ranges ($2^4 = 16$ states). The states constitute the nodes of a network with weighted directed connections defined by the transition probabilities of the Markov chain. The null hypothesis that the Markov chain has no memory is rejected using test statistics for three memoryless models (uniform, Poissonian and fixed Markov chain). It is confirmed that the degree (strength) distribution of the network matches well the limiting state distribution of the Markov chain. In a different approach, two semi-Markov models are developed, one for subareas (5 states) and one for magnitudes (4 states), for the sequence of strong earthquakes using appropriate time step and core matrices. The semi-Markov model on the subareas and magnitudes is found to give satisfactory aftcast (estimation of the next transition considering the data until the time of forecast), which is regarded an estimate of seismic hazard. Finally, a new approach that combines the Markov and semi-Markov models is attempted in order to estimate the occurrence probability of the next strong earthquake assuming that the previous strong earthquake (for semi-Markov model) and the previous state (for Markov chain) are given.

Keywords: Complex networks, Seismic zones, Magnitude ranges, Markov chain, Semi-Markov chain.

1. INTRODUCTION

The seismic hazard assessment is one of the main targets of seismological research aiming to contribute in reducing the catastrophic consequences from strong earthquakes occurrence. By seismic hazard assessment we mean the probability of occurrence of strong earthquakes within a given space, time, and magnitude ranges. The most known model, which is referred for seismic hazard assessment, is the Poisson model for random series of events (earthquakes). The main conditional of this model is that the earthquake occurrences are independent in space and time. Therefore, the Poisson model is frequently applied for statistical analysis of seismicity (Lomnitz 1974, Bath 1978, Brillinger 1982, Lomnitz and Nava 1983). The Markov chain model was introduced as a suitable means for earthquake probability estimation, which contrary to the Poisson model, assumes that all events are dependent on one another in space and time (Tsapanos and Papadopoulou 1999, Console 2001, Console et al. 2002, Nava et al. 2005). The semi-Markov model employed in order to estimate the waiting time and magnitudes of strong earthquakes (Altınok 1991, Altınok and Kolçak 1999). According to the semi-Markov model, the next strong earthquake depends on the previous one and the time elapsed between them. A different emerging field for seismic hazard assessment is based on network theory. The complex network analysis was introduced by Abe and Suzuki (2004) in order to study seismicity as a spatiotemporal complex system. Considerable research work was accomplished on network theory and its applications in different disciplines ranging from communication and economics to biology and neuroscience (Wang and Chen 2003, Emmert-Streib and Dehmer 2010, Rubinov and Sporns 2010).

The main purpose of this study is to provide earthquake estimates using earthquake data for the Greek area. It is shown that the seismicity can be modeled as a Markov and semi-Markov chain. Earthquake network is formed on the basis of the transition probability matrix of the Markov chain model and the core matrix of the semi-Markov chain model. A new approach that combines the Markov and semi-Markov models is attempted in order to identify the space, time, and magnitude of the next strong earthquake.

2. METHODOLOGY

The section of methodology is divided in 7 subsections. In the first subsection the magnitude threshold of data and the states of systems are defined, in the second subsection the time interval of chains is determined, in the next three subsections the models (Markov, semi-Markov and new approach) are presented, in the sixth subsection the evaluation scheme for the results is presented and in the seventh subsection the network measures used in the study are briefly described.

2.1 The definitions of data and states

The data must satisfy the completeness requirements, namely, to contain all the earthquakes. Thus, the data include earthquakes from seismic catalog with magnitude $M \geq M_{thr}$, where M_{thr} the magnitude of completeness. Concerning the definition of states, we divide the study area into R seismic zones, which are

homogenous from the seismotectonic point of view (faulting type, seismic moment rate) and taking into account previous results (Papaioannou and Papazachos 2000). In addition, to estimate the magnitude we create K magnitude ranges.

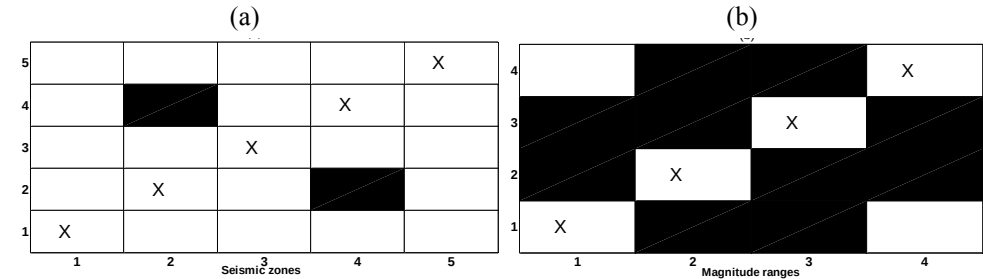
Regarding the Markov chain model, for each time interval Δt (the way it is defined is shown below, see section 2.2), we create two chains one for the regions and one for the magnitudes. Each state of seismic zones (regions) or magnitude ranges takes the value 0 or 1, corresponding to the absence or presence of earthquakes in the respective seismic zones or magnitude ranges. The chain for the regions is created by the earthquakes which are occurred in the seismic zones regardless of the magnitude ranges. In the other case the chain for the magnitudes is created by the magnitude ranges regardless of the seismic zones in which are occurred the earthquakes. The total number of states of the system, for the seismic zones or the magnitude ranges, are 2^R (R denotes the number of seismic zones) or 2^K (K denotes the number of magnitude ranges). In the binary form, each state can be denoted $z_R \dots z_2 z_1$ ($z_1, z_2, \dots, z_R$ the seismic zones) or $m_R \dots m_2 m_1$ ($m_1, m_2, \dots, m_R$ the magnitude ranges) is simply the right to left concatenation of the binary seismic zones or magnitude ranges states. The disadvantage of this approach is that the succession of earthquakes in each Δt is not to be taken into account. Two successive states formed at two subsequent time intervals Δt , define the transition between of the states.

If the seismic activity within the time window Δt were to be independent with respect to the magnitude ranges or seismic zones, the construction of states presented above would not be required and the analysis could be done separately at each magnitude range or seismic zone. The correlation analysis showed that both magnitude ranges and seismic zones are indeed correlated, as shown in Fig. 1. Specifically, for a step $\Delta t=0.5$ year, we consider the five series of earthquake occurrence frequency of the five seismic zones as well as the four series of earthquake occurrence frequency of the four magnitude ranges, having 210 data points per series. For each pair of seismic zone series the Pearson correlation coefficient is computed and the parametric significance test using the t-statistic is performed at the significance level $\alpha = 0.05$. The results are shown in matrix form in Fig. 1a and the same results are shown for the magnitude ranges in Fig.1b. The black color ($t < 0.05$) in the pairs of seismic zones and magnitude ranges reveals the correlation between them (Fig. 1).

In case of a semi-Markov model we also create two chains where the state space S is simply the R seismic zones or K magnitude ranges, respectively, as the focus is on the region or magnitude of the next strong earthquake. Thus, the total number of states of the system, for the seismic zones and the magnitude ranges, are R and K , respectively. For the transitions we take into account the time units (holding time), integer multiple of Δt , where the process of semi-Markov chain may remain at state i before made the transition in state j . Thus, two successive earthquakes define

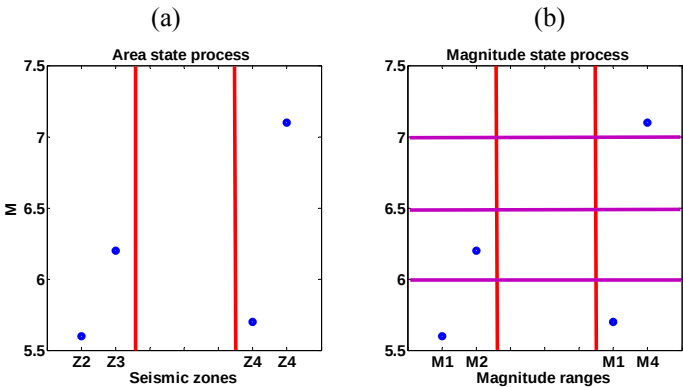
the transition between the respective states for the semi-Markov chain taking into account the holding time.

Figure 1. The value of the test statistic t among of seismic zones in (a) and magnitude ranges in (b). The black color in cells shows that the corresponding pairs are correlated ($t < 0.05$).



For example, if $R = 4$ and $M = 4$ we define the two next transitions within the three predefined time intervals $\Delta\tau$, separated by red lines in Figure 2, as described below. For the Markov chain about regions for each time interval $\Delta\tau$, we have $z_4z_3z_2z_1 = 0110$ (state 7), $z_4z_3z_2z_1 = 0000$ (state 1) and $z_4z_3z_2z_1 = 1000$ (state 15), the occurrence of earthquake represented by 1, so the transitions are $7 \rightarrow 1 \rightarrow 15$. For the semi-Markov chain the respective transitions are $2 \xrightarrow{1} 3 \xrightarrow{2} 4 \xrightarrow{1} 4$, where the numbers above the arrows represent the holding time. The transitions in case of magnitudes are defined similarly. Therefore, for a Markov chain we have $m_4m_3m_2m_1 = 0011$ (state 4), $m_4m_3m_2m_1 = 0000$ (state 1) and $m_4m_3m_2m_1 = 1001$ (state 10), so the transitions are $4 \rightarrow 1 \rightarrow 10$. For the semi-Markov the respective transitions are $1 \xrightarrow{1} 2 \xrightarrow{2} 1 \xrightarrow{1} 4$ (Fig. 2).

Figure 2. An example about the definition of states systems (Markov and semi-Markov) about the seismic zones in (a) and the magnitude ranges in (b). The red vertical lines divide the time intervals $\Delta\tau$ and the purple horizontal lines create the magnitude ranges.



2.2 Determination of time interval of chains

For the threshold time chosen, $\Delta\tau$ should be large enough to allow interaction among seismic zones or magnitude ranges in order the transition probabilities to be robust. On the other hand, $\Delta\tau$ should be small enough such that the hazard assessment be useful. If $\Delta\tau$ is too small, the most frequently occurring transition is from state 1 to state 1 (no earthquake occurs), whereas if it is too large state 2^R or 2^K to state 2^R or 2^K (earthquakes occur in all possible zones or magnitudes ranges) transitions are dominant. We consider three criteria given by the following three functions in order to determine the time interval $\Delta\tau$ for the Markov chain model.

Function 1 is the difference between the number of transitions from state 1 to state 1 and the number of transitions from state 2^R or 2^K to state 2^R or 2^K , respectively.

Function 2 is the difference between the total number of transitions from state 1 and the total number of transitions from state 2^R or 2^K .

Function 3 is given by $F = -\sum_i \pi_i \sum_j p_{ij} \log_2 p_{ij}$. This function is based on the maximum entropy principle as it is applied to finite Markov chains, where p_{ij} are elements of the transition probability matrix P and $\pi = \{\pi_i\}$ is the stationary distribution of the Markov chain.

To satisfy the functions 1-3, $\Delta\tau$ is chosen so that both Functions 1 and 2 show a value close to 0 and Function 3 is maximized.

2.3 Markov model

The Markov model is a probabilistic one useful in analysing stochastic phenomena. Suppose $S = \{1, 2, \dots, 2^R \text{ or } 2^K\}$ is the state space of a Markov chain. Let us define $\{X_t\}$, where t is a time index (at multiples of $\Delta\tau$), be a Markov chain formed by the time succession of states with values from the state space S . The Markov chain is defined in terms of a transition probability matrix: $P = P(j|i) = \{p_{ij}\}$, where p_{ij} is the probability that the state j follows state i with $i, j = 1, 2, \dots, 2^R \text{ or } 2^K$. From an observed Markov chain over n time units, the transition probability p_{ij} is estimated by the ratio of successions of the ordered pair $(i, j), \{i_{tj}\}$, over all observed successions starting at state i , $\{p_{ij}\} = \frac{\{i_{tj}\}}{\sum_{j=1}^N \{i_{tj}\}}$.

The transition probabilities p_{ij} satisfy $p_{ij} \geq 0$ and $\sum_{j=1}^N p_{ij} = 1$. Given p_{ij} and the system is in state i , we express the conditional probability of an earthquake occurring in seismic zone z_b , with $b = 1, \dots, R$, or regional activity probability, $p_{i z_b}$: $p_{i z_b} = \Pr\{z_b|i\} = \sum_{j \ni z_b} p_{ij}$, where $j \ni z_b$ means that state j includes seismicity in zone z_b (Nava et al. 2005). The regional activity probability implies that the probability of a Markov process which has entered state i will enter zone z_b on its next transition

depends only upon the current state. The magnitude activity probability, p_{lm_b} , with $b = 1, \dots, K$ is defined in a similar way.

2.4 Discrete semi-Markov model

The behaviour of a discrete semi-Markov model is similar with that of a pure Markov model. A discrete time semi-Markov process is defined completely by the transition probabilities, the holding time mass functions and the core matrices or discrete semi-Markov kernels (Pyke 1961, Barbu and Limnios 2009). The transition probability $\{p_{ij}\}$ is the probability that a semi-Markov chain that entered state i on its last transition will enter state j on its next transition, allowing for a holding time m for the transition to take place (Trezvas and Limnios 2011). The standard definition of a semi-Markov chain thus excludes the transition from one state to the same state. Applying directly the semi-Markov model to our setting, having the earthquakes at any seismic zone or magnitude range as the R or K states, respectively, would require to define the time steps by the running index in the sequence of successive earthquakes, which is not a natural time parameter. In our setting, we have defined the time steps in terms of the interval $\Delta\tau$. Since more than one earthquake at a different seismic zone or magnitude range can occur in the interval $\Delta\tau$ we modify the definition of the semi-Markov model and assume that transitions from state (earthquake) i to j are allowed within the interval $\Delta\tau$ and assign for such transitions the holding time $t_{ij}=0$. Accordingly, $t_{ij}=1$ regards the transition from a state (earthquake at a seismic zone or magnitude range) occurring at a time $t_k = k\Delta\tau$, where this is actually the whole interval $((k-1)\Delta\tau, k\Delta\tau)$, to a state occurring in the next time $t_{k+1} = (k+1)\Delta\tau$. In this case, the state i refers to the last earthquake occurring at the time interval $((k-1)\Delta\tau, k\Delta\tau)$. According to this definition of the semi-Markov model, transitions from one state to the same state are allowed. The probability mass function T_{ij} is called the holding time mass function for a transition from state i to j and is given as $Pr\{t_{ij} = m\} = T_{ij}(m)$, where m is the time unit (0 within $\Delta\tau$, and otherwise multiples of $\Delta\tau$). The final step is to define the core matrices (discrete semi-Markov kernels). The $c_{ij}(m)$ element of the core matrix $C(m)$ is the probability of the joint event that a system that entered state i makes its next transition to state j and this takes place after a holding time m (Altinok and Kolcak 1999). The core matrix is given by: $C(m) = c_{ij}(m) = p_{ij} \otimes T_{ij}(m)$, where $\otimes$ denotes multiplication of corresponding elements.

2.5 New approach combining the Markov and semi-Markov chain

Let us first concentrate on the seismic zones (the approach for the magnitude ranges is similar). The interest here is to determine the probability of having a strong earthquake at the next time interval $\Delta\tau$ in one of the seismic zones given the

information of earthquake occurrences at the present time interval. The states of the Markov chain do not contain the information about the order of occurrences of earthquakes at a time interval, and particularly the last earthquake in this interval (referring to the states representing more than one earthquake occurrence in the time interval). On the other hand, as the semi-Markov model is defined, its state carries the information about the last earthquake occurrence regarding the present time interval. In the proposed approach, we attempt to combine the information of the state of the Markov model and the semi-Markov model for the same time interval, and further use it to predict the occurrence of an earthquake at a specific seismic zone within the next time interval. First, using the states of the Markov chain for the seismic zones, we calculate the regional activity probability matrix P_{reg} of size $2^R \times R$ with components $p_{i|z_j}$, where $i = 1, 2, \dots, 2^R$, $j = 1, 2, \dots, R$, and $z_1, z_2, \dots, z_R$, are the R seismic zones. The $p_{i|z_j}$ denotes the probability that given the Markov model state i at time $t_k = k\Delta\tau$, a strong earthquake occurs at seismic zone z_j in the time interval $t_{k+1} = (t_k, t_k + \Delta\tau)$. We further involve the semi-Markov model defined in Section 2.4, and particularly the part of it that regards holding times $t_{ij}=1$, in order to use the information about the last earthquake in the current interval and predict that a strong earthquake occurs at seismic zone z_j in the time interval $(t_k, t_k + \Delta\tau)$. This information lies in the core matrix $c_{r,f}(1)$ of size $R \times R$ with $r, f = 1, 2, \dots, R$. The $c_{r,f}(1) = c_{z_r, z_f}(1)$, denoted for simplicity c_{z_r, z_f} is the probability that given the last strong earthquake is at seismic zone z_r in the current $\Delta\tau$ at time $t_k = k\Delta\tau$, a strong earthquake occurs at seismic zone z_j in the time interval $t_{k+1} = (t_k, t_k + \Delta\tau)$. The two transition probabilities $p_{i|z_j}$ and c_{z_r, z_j} from the Markov and the semi-Markov model, respectively, target in predicting the seismic zone an earthquake occurs in the next time interval $t_{k+1} = (t_k, t_k + \Delta\tau)$, the former on the basis of the state of seismic zones in $t_k = k\Delta\tau$, and the latter on the basis of the seismic zone of the last earthquake in t_k . We merge the two probabilities to the probability $q_{i, z_r, z_j} = p_{i|z_r} c_{z_r, z_j}$, which approximates the conditional probability $P(z_j | t, z_r)$, i.e. the occurrence of an earthquake in seismic zone z_j within the interval t_{k+1} when in the last interval t_k the state of seismic zones is i and the last earthquake in t_k is z_r . The probability q_{i, z_r, z_j} is an operationally suitable approximation rather than an exact expression of $P(z_j | t, z_r)$ using the probabilities obtained from the Markov and semi-Markov model. The probabilities q_{i, z_r, z_j} can be estimated by the corresponding frequencies of $p_{i|z_r}$ and c_{z_r, z_j} . The probability $q_{i, m_r, m_j} = p_{i, m_r} c_{m_r, m_j}$ is defined similarly for the magnitude ranges.

2.6 Results Evaluation

For evaluating the performance of the Markov chain, we will compare the aftcast probability of the Markov model to a threshold formed according to the hypothesis that the system has no memory, and the seismicity cannot be modeled as a Markov chain. Three memoryless models are considered: uniform, Poisson and fixed Markov chain. The uniform model has transition probabilities $P_{ij}^U = P^U = S^{-1} = \{p_{ij}\}$, where the probabilities $\{p_{ij}\}$ correspond to purely random guessing. The Poisson model has transition probabilities $P_{ij}^P = P_j^P = \prod_{z_b \in j} (1 - e^{-\lambda_{z_b} \Delta t}) \prod_{z_b \notin j} (e^{-\lambda_{z_b} \Delta t})$, where λ_{z_b} is the mean number of earthquakes per unit time in zone z_b and $z_b \in j$ means that state j includes seismicity occurrence in zone z_b . The third model is the fixed Markov chain model and is given by $p_{ij} = P_j^Q = \frac{\xi_j}{\sum_i \xi_i}$, where $\xi_i = \sum_j i_{ij}$. For magnitude states the three models are defined accordingly. For the uniform model we consider a threshold probability $P^H = (1 + \mu)P^U$, where μ is an arbitrary non-negative constant. An aftcasted transition is defined as successful if $p_{ij} > P^H$. For the Poisson or fixed Markov chain model we have a successful aftcasted transition when $p_{ij} > P_j^P$ or $p_{ij} > P_j^Q$. Aftcasted, means that the probabilities p_{ij} are evaluated based on all available information considering the data until the time of forecast.

For evaluating the performance of the semi-Markov chain, it is tested whether the next predicted transition $c_{ij}(m)$ of system having the maximum probability among of the others pair of states is in agreement with the observed transition.

2.7 Network measures

Generally with the term «network» we mean the graph $G = (N, E)$ that is defined by the nodes and the connections between them, where N is the set of nodes and E the set of connections. In our analysis, the nodes are represented by the states of system and the directed weighted connections are defined by the elements of the transition probability matrix P (for Markov model) and core matrix C (for semi-Markov model), respectively. The network properties are quantified with a number of characteristics (network measures) computed on P and C .

The simplest and most known network measure is the degree (for binary connections) or strength (for weighted connections). This characteristic measures the number of connections or the sum of weights at each node i as $k_i^P = \sum_{j \in N} p_{ij}$. Essentially, the measure reveals whether a node is active in the network. Another well-known characteristic is the average clustering coefficient, which estimates the tendency of any node i to form connected triads given as $C^* = \frac{1}{n} \sum_{i \in N} \frac{\sum_{j,k \in N} p_{ij} p_{ik} p_{jk}}{(k_i^{in} + k_i^{out})(k_i^{in} + k_i^{out} - 1) - 2 \sum_{j \in N} p_{ij} p_{ji}}$, where k_i^{in} is the sum of connection weights directed to the node and k_i^{out} is the sum of connection weights leaving the node. A high value of the average clustering coefficient indicates higher likelihood

existence of "clique" (clique is a group of fully connected nodes). The third characteristic is the betweenness centrality defined as the number of shortest paths between pairs of nodes that pass through a specific node given as $b_i = \frac{1}{(N-1)(N-2)} \sum_{h \neq i, j \neq i} \frac{\rho_{hi}(i)}{\rho_{hj}^{(i)}}$, where ρ_{hj} is the number of the shortest paths between h and j nodes and $\rho_{hi}(i)$ the number of the shortest paths between h and j nodes through the node i . Nodes with a high betweenness have a high probability to occur on a randomly chosen shortest path between two randomly chosen nodes. Such nodes are critical to the network since their removal would destroy many short paths in the network.

3. DATA AND APPLICATION

The data are obtained from the earthquake catalog compiled in the Geophysics Department of the Aristotle University of Thessaloniki (<http://geophysics.geo.auth.gr/ss/>). They comprise crustal earthquakes (focal depth less than 40 Km) that occurred in 1911-2015, and are divided in four subsets of magnitude $M_{chr} \geq 5.5$ (574 events), $M_{chr} \geq 5.5$ (444 events, after declustering), $M_{chr} \geq 6.0$ (188 events) and $M_{chr} \geq 6.0$ (154 events, after declustering). The data are complete for the study period and declustering was performed for testing subsets of the entire catalog or after removing dependent events with respect to foreshocks and aftershocks. Seismicity declustering is the identification and the separation of seismicity catalogs into main shocks (independent events), foreshocks and aftershocks (dependent events), so as to eliminate the interference of the already dense occurrence and strong dependence of the events which belong to an aftershock seismic excitation. The Reasenbergs's algorithm (1985) is used, here, for the declustering procedure. The algorithm is used for identifying aftershock clusters based on a two-parameter earthquake interaction model producing a Poissonian declustered earthquake catalog which is deprived of correlated events.

The study area is divided in 7 subareas and we define 5 seismic zones to reduce complexity, which are homogenous from the seismotectonic point of view, and are shown in Figure 3 along with the epicentral distribution of the earthquakes used for the analysis. In addition, the seismic catalog is divided in 4 magnitude ranges for each of the 4 data subsets with the purpose of distinguishing the different levels of earthquake magnitudes (moderate, strong, major and great). In case of magnitude $M_{chr} \geq 5.5$ we define the following magnitude ranges: $5.5 \leq M_1 \leq 5.9$, $6.0 \leq M_2 \leq 6.4$, $6.5 \leq M_3 \leq 6.9$ and $M_4 \geq 7.0$. Then, with magnitude $M_{chr} \geq 6.0$ we define: $6.0 \leq M_1 \leq 6.2$, $6.3 \leq M_2 \leq 6.5$, $6.6 \leq M_3 \leq 6.9$ and $M_4 \geq 7.0$.

For the five seismic zones and for the four magnitude ranges there are $2^5 = 32$ and $2^4 = 16$ states, respectively. For example in the case of seismic zones, the state 1 (denoted 00000 in binary format) corresponds to no earthquake occurrence in all five seismic zones in the chosen time interval, Δt , unlike the state 32 (denoted 11111 in binary format) that represents earthquake occurrence in all five seismic. The

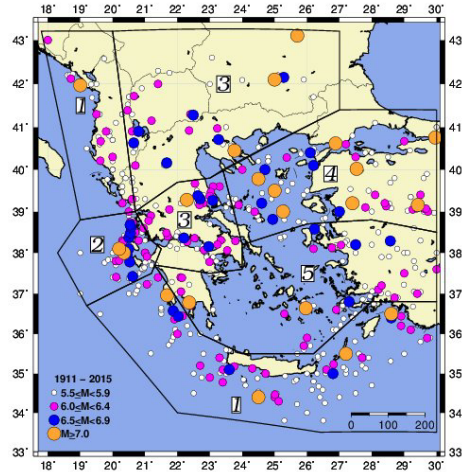
combination of seismic zones or magnitude ranges gives the nodes of the network and the states of a Markov chain system.

In the case of semi-Markov chain, the nodes of the network and the states of the system correspond to the five seismic zones (5 states) and the four magnitude ranges (4 states), respectively. The directed weighted connections of a network are given by the transition probabilities p_{ij} of the Markov chain and the elements of core matrix $c_{ij}(m)$ of the semi-Markov chain.

For the selection of the time interval Δt of chains, the optimal Δt should give the largest entropy (Function 3) and the value of Functions 1 and 2 closes to zero (see Subsection 2.2). So, in our analysis the time window that best meet the three criteria is 6 months for $M_{thr} \geq 5.5$ and 12 months for $M_{thr} \geq 6.0$ as evidenced in Figure 4.

We created the earthquake network based on the seismic zones and the magnitude ranges for each model (Markov, semi-Markov) using the different data settings. The construction of network allows us to check which distribution of network measures (strength, clustering coefficient, betweenness centrality) matches well with the limiting state distribution of the Markov chain.

Figure 3. Epicentral distribution of earthquake magnitudes $M_{thr} \geq 5.5$, that occurred in 1911-2015 in the broader area of Greece. The division of the area in seismic zones is also shown.



Having defined the data, the system states and nodes, the time interval of chains and the earthquake network, we now estimate the seismic hazard assessment for the next strong earthquake occurrence. Thus, we compute the transition probability and the regional or magnitude activity probability matrices of Markov chain and the core matrices of semi-Markov chain for all cases (data settings, models using as states the seismic zones and magnitude ranges).

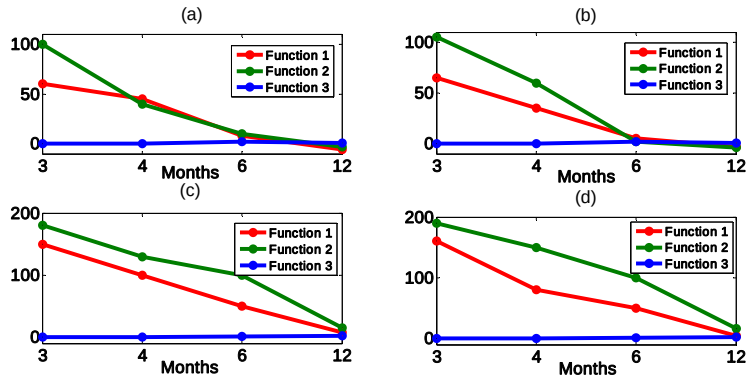
The evaluation of the model performance is done by counting the number of successes of aftcasted regions or magnitude ranges occurrences, respectively. The aftcasts are used when the seismic catalog is not large and the number of transitions is

not big in order to both obtain a robust transition probability matrix and have a statistically significant sample of forecasts.

4. RESULTS

With the chosen time interval Δt , which is 6 months for $M_{thr} \geq 5.5$ and 12 months for $M_{thr} \geq 6.0$, we compute the transition probability, the regional or magnitude activity and the core matrices for all cases. In particular, in the case of $M_{thr} \geq 5.5$ and seismic zones as states, it is evident from Figure 5a that the states of low occurrence of earthquakes in seismic zones have higher transition probabilities than the states of high occurrence. In addition, from the estimated regional activity probability matrix we notice that when the system is in state $i = 1, \dots, 2^R$, the probability of a strong earthquake occurring in seismic zone 1 in the next Δt is very high for all cases (see Fig. 5b). Thus, we can assume that the seismic zone 1 has a highest risk for a strong earthquake than the other seismic zones. This is the reason that the states which include the seismicity of seismic region 1 have high transition probabilities in Figure 5a. Also, from the core matrix (see Fig. 5c) the key finding is that the main shocks have the tendency to recur in a short time in the same seismic zone that caused the earthquake. This fact is confirmed from the values at the main diagonal (transitions among the same seismic zones) which shows higher probabilities than the other pairs of states. This is in accordance with the clustering of strong earthquakes observed in many others cases in Greece and worldwide (Kagan and Jackson 1991, Papadimitriou 2002).

Figure 4. Determination of time interval Δt when $M_{thr} \geq 5.5$ in (a) the entire catalog and (b) after declustering, and respectively for $M_{thr} \geq 6.0$ in (c) and (d).



Then, the limiting state distribution corresponding to the Markov chain model in relation with the distribution of network measures (strength, clustering coefficient, betweenness centrality) is shown in Figure 6. It can be observed that the limiting state distribution of the underlying directed network with weighted connections is in full agreement with the distribution of strength where the states (nodes) are represented by seismic zones and magnitude ranges in Figure 6a and 6d, respectively. This is due to the way that the weighted directed networks were constructed. Matching is also

observed in the distributions between of the limiting state distribution and the betweenness centrality as shown in Figures 6c and 6f unlike the case of clustering coefficient (see Fig. 6b and 6e).

The last step in our analysis is the evaluation of the performance of each model with respect to the null model of no memory (see Fig. 7). The evaluation results for Markov and semi-Markov model are obtained by quantifying the aftcast success rate from the last 20 transitions corresponding to 20% of the total transitions in the case of $M_{thr} \geq 6.0$ and 10% of $M_{thr} \geq 5.5$. The transitions that are under testing come from the transition probability and core matrix for Markov and semi-Markov model, respectively (see Fig. 7a, 7b and 7c). The success rate is better for the magnitude states than the zoning states, the latter going beyond 90% for the Markov model and 80% for the semi-Markov model, respectively. On the other hand, the success rate for zoning states is about 50% for both cases (Markov, semi-Markov) if we exclude the 85% success rate when the null model is based on the uniform distribution constituting the most random scenario. The performance evaluation for the new approach is done on the basis of the latest 10 transitions for $M_{thr} \geq 6.0$ (without and after declustering) and the time step Δt at 12 months, for data sets in 2005-2015.

Figure 5. The transition probability matrix and the regional activity probability matrix, shown as color maps (black for zero and white for 1) in (a) and (b), respectively, which come from Markov chain and the core matrix for $m = 1$ in (c) which come from semi-Markov chain in all cases with magnitude $M_{thr} \geq 5.5$ and states referring to seismic zones.

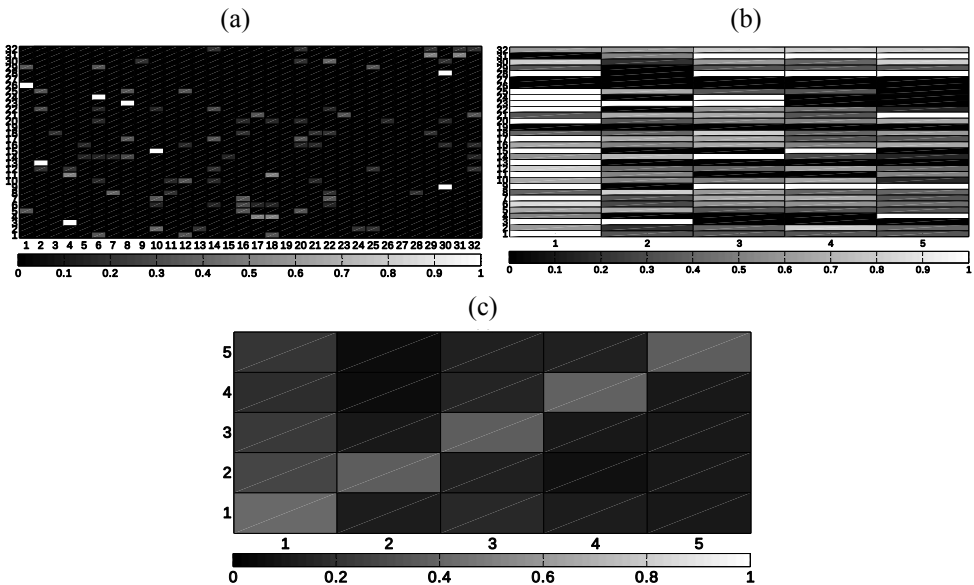
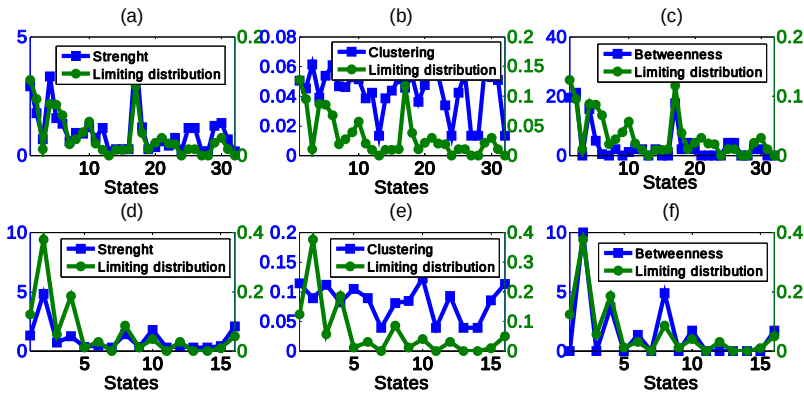


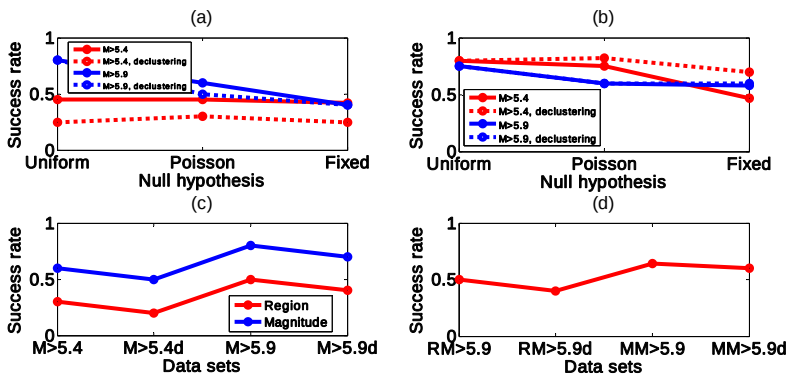
Figure 6. The comparison of the limiting distribution of Markov chains with $M_{thr} \geq 5.5$ on the basis of the strength in (a) and (d), clustering coefficient in (b) and (e), and betweenness

centrality in (c) and (f), for the states of seismic zones (2³) and magnitude ranges (2⁴), respectively.



In these 10 years (10 transitions), 14 strong earthquakes with $M_{\text{max}} \geq 6.0$ have occurred, four of them being removed after declustering. Similarly to the previous cases (see Fig. 7a, 7b and 7c) the results of evaluation are better for the magnitude states, the highest success rate being 65%, than the zoning states where the highest success rate is 50% (see Fig. 7d). In particular, the results are better when the entire catalog is used. This remark can be substantiated by the fact that when declustering algorithm is applied, main shocks considered dependent are removed from the catalog, therefore the estimation of the next transition (the next strong earthquake) becomes more unpredictable.

Figure 7. The performance evaluation in (a) with Markov model for seismic zones, in (b) with Markov model for magnitude ranges both of cases with $\mu = 1$ for the uniform null model, in (c) with semi-Markov model and (d) with new approach. The suffix ‘d’ in (c) and (d) means that the data come from declustering algorithm and the suffix ‘R’ and ‘M’ in (d) means that the data are referred in seismic zones and magnitude ranges, respectively.



5. CONCLUDING REMARKS

The seismic hazard assessment performed by means of network and analysis and Markov models shows that the occurrence of strong earthquakes could have been predicted to some extent. Application of the models (Markov, semi-Markov, new

approach) to the Greek area yields significant results that have negligible probability of being obtained by purely random guessing. The transition probabilities, or for the case of semi-Markov models the transition probability and core matrices, between the states (seismic zones, magnitude ranges) of the system generate useful seismic hazard estimates, on which the forecast of transitions with high probability may be based. The results of model performance evaluation in all cases are better for the estimation of the next magnitude states than the zoning states. In addition, the results of evaluation are better when we do not use the declustering algorithm.

The latter is attributed to the occurrence of dependent strong earthquakes, but given the fact that strong earthquakes cause huge disasters, it is preferable not to exclude these main shocks from the seismic catalog. The network theory can contribute to improve our understanding of these complex systems, such as the seismicity, as the limiting distribution of Markov chain is in agreement with the distribution of some network measures (strength, betweenness centrality). An open issue arising from this study is the application of the same analysis using synthetic seismic catalogs in order to provide more data (simulation data), i.e. more transitions, and thus the results become more robust.

ΠΕΡΙΛΗΨΗ

Ο σκοπός της παρούσας μελέτης είναι η αξιολόγηση της σεισμικής επικινδυνότητας στον χώρο της Ελλάδας με βάση τον υπολογισμό των πιθανοτήτων μετάβασης από ένα Μαρκοβιανό και ημι-Μαρκοβιανό μοντέλο, αντίστοιχα. Θεωρούμε δύο σύνολα δεδομένων τα οποία περιλαμβάνουν ισχυρούς σεισμούς με μέγεθος $M_{thr} \geq 5.5$ και $M_{thr} \geq 6.0$, αντίστοιχα, που συνέβησαν κατά την περίοδο 1911-2015. Η περιοχή μελέτης χωρίζεται σε 5 υποπεριοχές (σεισμικές ζώνες) που είναι ομοιογενείς από την σεισμοτεκτονική άποψη και ο σεισμικός κατάλογος διαιρείται σε υποσύνολα για τέσσερα εύρη μεγέθους. Δύο Μαρκοβιανές αλυσίδες ορίζονται σε σχέση με ένα κατάλληλο επιλεγμένο χρονικό παράθυρο, η μία έχει ως καταστάσεις την εμφάνιση ισχυρών σεισμών σε καμία, μία ή σε περισσότερες υποπεριοχές ($2^5=32$ καταστάσεις), και η δεύτερη έχει ως καταστάσεις την εμφάνιση ισχυρών σεισμών οπουδήποτε στην Ελλάδα από καθένα από τα τέσσερα εύρη μεγέθους ($2^4=16$ καταστάσεις). Οι καταστάσεις είναι οι κόμβοι του κατευθυνόμενου δικτύου και οι σταθμισμένες συνδέσεις ορίζονται από τις πιθανότητες μετάβασης της Μαρκοβιανής αλυσίδας οι οποίες εκτιμώνται από τα δεδομένα του σεισμικού καταλόγου. Η εκτίμηση της επόμενης μεταβατικής κατάστασης με την χρήση της Μαρκοβιανής αλυσίδας έχει βρεθεί στατιστικά σημαντική τόσο για τις υποπεριοχές όσο και για τα μεγέθη. Ειδικότερα, η μηδενική υπόθεση ότι η Μαρκοβιανή αλυσίδα δεν έχει μνήμη απορρίπτεται με χρήση ως στατιστικού ελέγχου τρία μοντέλα μνήμης (ομοιόμορφο, Poisson και καθορισμένη αλυσίδα Markov). Επιβεβαιώνεται, ότι η κατανομή του βαθμού (της ισχύος) του δικτύου ταιριάζει με την οριακή κατανομή των καταστάσεων της Μαρκοβιανής αλυσίδας. Σε μια διαφορετική προσέγγιση θεωρούνται δύο ημι-Μαρκοβιανές αλυσίδες, μία για τις υποπεριοχές (5 καταστάσεις) και μία για τα μεγέθη (4 καταστάσεις), για την ακολουθία των ισχυρών σεισμών με χρήση κατάλληλου χρονικού παραθύρου και πινάκων του πυρήνα. Η ημι-Μαρκοβιανή αλυσίδα και για την περίπτωση των υποπεριοχών και για αυτή των μεγεθών έχει βρεθεί να δίνει αξιόπιστη εκτίμηση της επόμενης μετάβασης λαμβάνοντας υπόψη τα δεδομένα μέχρι τη στιγμή της πρόβλεψης. Τέλος, μια νέα προσέγγιση, που συνδυάζει την Μαρκοβιανή και την ημι-Μαρκοβιανή αλυσίδα, επιχειρείται προκειμένου να εκτιμηθεί ο επόμενος ισχυρός σεισμός αν υποθεθεί ότι ο προηγούμενος

ισχυρός σεισμός (από την ημι-Μαρκοβιανή αλυσίδα) και η προηγούμενη κατάσταση (από την Μαρκοβιανή αλυσίδα) δίνονται.

REFERENCES

- Abe, S. and Suzuki, N. (2004). *Scale-free network of earthquakes*, Europhysics Letters, **65**, 581-586.
- Altınok, Y. (1991). *Evaluation of earthquake risk in West Anatolia by semi-Markov model*, Jeofizik, **5**, 135-140.
- Altınok, Y. and Kolçak, D. (1999). *An application of the semi-Markov model for earthquake occurrences in North Anatolia, Turkey*, Balkan Geophysical Society, **2**, 90-99.
- Barbu, V. S. and Limnios, N. (2009). *Semi-Markov chains and hidden semi-Markov models toward applications: their use in reliability and DNA analysis*, Vol. 191, Springer Science & Business Media.
- Bath, M. (1978). *Seismic risk in Fennoscandia*, Tectono-physics, **57**, 285-295.
- Brillinger, D. (1982). *Some Bounds for Seismic Risk*, Bull Seismological Society of America, **72**, 1403-1410.
- Console, R. (2001). *Testing earthquake forecast hypotheses*. Tectonophysics, **338**, 261-268.
- Console, R., Pantosti, D. and D'Addezio, G. (2002). *Probabilistic approach to earthquake prediction*. Ann Geophys, **45(6)**, 723-731.
- Emmert-Streib, F. and Dehmer, M. (2010). *Influence of the Time Scale on the Construction of Financial Networks*, Plos one **5(9)**: e12884.
- Kagan, Y. and Jackson, D. (1991). *Seismic Gap Hypothesis: Ten years after*, Journal of geophysical research, **96**, 21419-21431.
- Lomnitz, C. (1974). *Global Tectonics and Earthquake Risk*, Elsevier Scientific Publishing Co, Amsterdam-London-New York.
- Lomnitz, C. and Nava, F. (1983). *The Predictive Power of Seismic Gaps*, Bull Seismological Society of America, **73**, 1815-1824.
- Nava, F.A., Herrera, C., Frez, J. and Glowacka, E. (2005). *Seismic hazard evaluation using Markov chains: Application to the Japan area*, Pure and applied geophysics, **162**, 1347-1366.
- Papaoiannou, Ch. A. and Papazachos, B. C. (2000). *Time-independent and time-dependent seismic hazard in Greece based on seismogenic sources*, Bulletin Seismology Society of America, **90**, 22-33.
- Pyke, R. (1961). *Markov renewal processes: definitions and preliminary properties*, The Annals of Mathematical Statistics, 1231-1242.
- Reasenber, P. (1985). *Second-order moment of central California seismicity, 1969-82*, J. Geophys Research, **90**, 5479-5495.

- Rubinov, M. and Sporns, O. (2010). *Complex network measures of brain connectivity: Uses and interpretations*, NeuroImage, **52(3)**, 1059-1069.
- Trevezas, S. and Limnios, N. (2011). *Exact MLE and asymptotic properties for nonparametric semi-Markov models*, Journal of Nonparametric Statistics, **23(3)**, 719-739.
- Tsapanos, T. M. and Papadopoulou, A. A. (1999). *A discrete Markov model for earthquake occurrence in Southern Alaska and Aleutian Islands*, Balkan Geophysical Society, **2(3)**, 75-83.
- Wang, X. F. and Chen, G. (2003). *Complex Networks: Small-World, Scale-Free and Beyond*, Feature, 6-20.



THE USEFULNESS OF OPTIMUM EXPERIMENTAL DESIGNS

Stratis Kounias

Emeritus professor, University of Athens, Greece

skounias@math.uoa.gr

ABSTRACT

A preview is presented for different aspects of optimum experimental design. Linear regression, discrete optimal designs and examples, optimality criteria, factorial designs, repeated optimal designs.

Keywords and phrases: Factorial designs, repeated measurements designs, optimality criteria.

1. INTRODUCTION

In most physical processes there exists a functional relationship which is too complicated to describe in simple terms. In this case we approximate this relationship by some simple mathematical function which contains the appropriate variables and approximates the true function over a region. In most of the cases we use linear functions of the parameters plus an error term, that is, linear regression $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{e}$,

$\boldsymbol{\theta}$ is the vector of the unknown parameters, $\mathbf{e}$ the vector of errors, $\mathbf{X}$ is called the design matrix.

If the errors are uncorrelated with constant variance σ^2 , the least squares gives,

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\theta}} = \mathbf{X}'\mathbf{Y} \Rightarrow \text{cov}(\hat{\boldsymbol{\theta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1},$$

$$\text{var}(\mathbf{c}'\hat{\boldsymbol{\theta}}) = \sigma^2\mathbf{c}'\mathbf{Q}^{-1}\mathbf{c}$$

where $\hat{\boldsymbol{\theta}}$ are the estimated parameters, $\mathbf{Q} = \mathbf{X}'\mathbf{X}$ is called *information matrix*.

If the covariance matrix of the errors is $\text{Cov}(\mathbf{e}) = E(\mathbf{e}\mathbf{e}') = \sigma^2\mathbf{V}$

$\mathbf{V}$ is a known symmetric, positive definite matrix, the least squares gives

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\hat{\boldsymbol{\theta}}_w = \mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} \Rightarrow \mathbf{Q} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}, \text{cov}(\hat{\boldsymbol{\theta}}_w) = \sigma^2\mathbf{Q}^{-1}$$

2. DISCRETE OPTIMAL DESIGNS

Example 2.1 Take the simple linear model

$$Y_i = \theta_0 + x_i \theta_1 + e_i \quad i = 1, \dots, n$$

with $-1 \leq x_i \leq 1$ and the errors are uncorrelated with variance σ^2 . If we take n_i observations at the points $x_i \quad i = 1, \dots, k$, the design d is

$$d = \begin{pmatrix} x_1 & x_2 & \cdots & x_k \\ n_1 & n_2 & \cdots & n_k \end{pmatrix}, \quad n_1 + \cdots + n_k = n$$

The design matrix is $\mathbf{X} = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix}^T$, the information matrix is

$$\mathbf{X}'\mathbf{X} = \mathbf{Q} = \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n x_i^2 \end{bmatrix}.$$

Let us consider the two designs, both with 3 observations,

$$d_1 = \begin{pmatrix} -1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad d_2 = \begin{pmatrix} -1 & 1 \\ 1 & 2 \end{pmatrix}$$

Choose the design d so that,

i) $\text{var}(\hat{\theta}_1)$ is minimized, ii) $\text{var}(\hat{y}(z))$ at $z = 0$ is minimized.

where $\hat{y}(z) = \hat{\theta}_0 + \hat{\theta}_1 z$ is the estimated response surface at the point z .

Take for convenience $\sigma^2 = 1$, then

$$i) d_1 : \bar{x} = 0, \quad Q = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}, \quad \text{var}(\hat{\theta}_1) = \frac{1}{2}, \quad d_2 : \bar{x} = \frac{1}{3}, \quad Q = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}, \quad \text{var}(\hat{\theta}_1) = \frac{3}{8} < \frac{1}{2}$$

Hence d_2 is better than d_1 .

$$(ii) d_1 : \text{var}(\hat{y}(z) = \hat{\theta}_0 + \hat{\theta}_1 z) = (1 - z)Q^{-1} \begin{pmatrix} 1 \\ z \end{pmatrix} = \frac{2 + 3z^2}{6} \Rightarrow \text{var}(\hat{y}(0)) = 1/3.$$

$$d_2 : \frac{1}{8}(1-z)Q^{-1}\begin{pmatrix} 1 \\ z \end{pmatrix} = \frac{3z^2 - 2z + 3}{8} \Rightarrow \text{var}(\hat{y}(0)) = 3/8 \text{ and } 1/3 < 3/8$$

Hence d_1 is better than d_2 at $z=0$.

Proposition 2.1 If the errors are uncorrelated the D-optimal design which minimizes $\text{var}(\hat{\theta}_1)$, when $-1 \leq x_i \leq 1$ $i = 1, \dots, n$ is:

If n is even, half of the observations are at $x_i = -1$ and half at $x_i = 1$ and $\min_{x_i} \text{var}(\hat{\theta}_1) = \sigma^2 / n$.

If n is odd, $(n \pm 1) / 2$ observations are at $x_i = -1$ and $(n \mp 1) / 2$ observations are at $x_i = 1$ and $\min_{x_i} \text{var}(\hat{\theta}_1) = \sigma^2 / (n - (1/n))$

$$\text{Proof. } Q = \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n x_i^2 \end{bmatrix}, \text{ if } n \text{ is even } \text{var}(\hat{\theta}_1) = \frac{\sigma^2}{\left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right)} \geq \frac{\sigma^2}{n} \text{ with}$$

equality only if $\bar{x} = 0$ and $n/2$ observations are at $x_i = 1$ and $n/2$ are at $x_i = -1$. If n is odd, all the observations are at $x_i = \pm 1$, otherwise there exists at least one observation at $-1 < x_j < 1$. If we take a design having this observation at

$-1 \leq x_j + \varepsilon \leq 1$ for some ε , then $\bar{x}$ becomes $\bar{x} + \varepsilon / n$ and $\sum_{i=1}^n x_i^2 - n\bar{x}^2$ becomes

$$\sum_{i=1}^n x_i^2 - n\bar{x}^2 + \varepsilon^2(1 - 1/n) + 2(x_j - \bar{x})\varepsilon \leq \sum_{i=1}^n x_i^2 - n\bar{x}^2 + \varepsilon^2(1 - 1/n) \text{ for a proper}$$

value of ε , which gives a smaller $\text{var}(\hat{\theta}_1)$. If $x_i = 1$ for r values of x_i and

$x_i = -1$ for $n-r$ values of x_i , then $\bar{x} = \frac{2r-n}{n}$ and

$$\sum_{i=1}^n x_i^2 - n\bar{x}^2 \leq n - \frac{(2r-n)^2}{n} = \frac{4r(n-r)}{n} \leq n - (1/n) \text{ which is attained for}$$

$r = (n \pm 1), n-r = (n \mp 1) / 2$. •

If instead of $-1 \leq x_i \leq 1$ we have $a \leq x_i \leq b$ the observations minimizing $\text{var}(\hat{\theta}_1)$ are taken at the two end points a and b .

Example 2.2 If $y_i = \theta_0 + \theta_1 x_{i1} + \dots + \theta_k x_{ik} + e_i \quad i = 1, \dots, n$ and $x_{i1}, \dots, x_{ik}$ are in a convex polyhedron in $\mathfrak{R}^k$, the D- optimal design maximizing $\det(Q)$ has as points of support (observations) on the boundary of the polyhedron.

2.1 Optimality criteria

If we have two or more parameters in the model $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\theta}$, then we have the information matrix $\mathbf{Q} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$, where $\sigma^2\mathbf{V}$ is the covariance matrix of the errors. For uncorrelated errors $\mathbf{V} = \mathbf{I}_n$.

The most used criteria are,

D-optimality: The D optimal design d^* maximizes the determinant of Q: $\max_d \det(Q)$, for these designs the 95% confidence region for the unknown parameters has smaller volume.

A-optimality: The A optimal design d^* minimizes the sum of the variances of the estimated parameters, that is, $\min_d (\text{trace}(Q^{-1}))$

G- optimality: If $E(Y_i) = f'(x)\boldsymbol{\theta} = \sum_{j=1}^p f(x_{i,j})\theta_j \quad i = 1, \dots, n, \quad x_{i,j} \in F$ The G

optimal design d^* minimizes the maximum variance of the estimated response surface,

$$\min_d \max_{x \in F} f'(x)Q^{-1}f(x)$$

or $\text{var}(\hat{y}(x)) = \text{var}\left(f(x_0)\hat{\theta}_0 + f(x_1)\hat{\theta}_1 + \dots + f(x_k)\hat{\theta}_k\right)$, [Kounias, Lefkopoulou, Bagiatis (1983)].

3. 2^k FACTORIAL DESIGNS

We have k factors $F_1, \dots, F_k$, each at two levels, low level and high level denoted 0 1. If the effect of F_i at level 0 is l_i and at level 1 is q_i , we are interested to estimate the k contrasts $\theta_i = q_i - l_i$. The model is,

$$y_i = (x_{i,1}l_1 + z_{i,1}q_1) + \dots + (x_{i,k}l_k + z_{i,k}q_k) + e_i \quad i = 1, \dots, n$$

where $x_{i,j} = 1, z_{i,j} = 0$ if F_j enters at low level and $x_{i,j} = 0, z_{i,j} = 1$ if F_j enters at high level.

Set $z_{i,j} + x_{i,j} = 1, z_{i,j} - x_{i,j} = w_{i,j}$, $j = 1, \dots, k$, then $z_{i,j} = (1 + w_{i,j}) / 2$, and $x_{i,j} = (1 - w_{i,j}) / 2$ and the model in vector form becomes,

$$\mathbf{y} = \mu \mathbf{1}_n + \sum_{j=1}^k \mathbf{w}_j v_j + \mathbf{e}$$

Where $\mu = \sum_{j=1}^k (l_j + q_j) / 2$, $(q_j - l_j) / 2 = v_j$ and $w_{i,j} = \pm 1$. The design matrix is,

$$\mathbf{X} = [\mathbf{1}_n \quad \mathbf{w}_1 \quad \dots \quad \mathbf{w}_k], \quad \mathbf{w}_j = (w_{1,j}, \dots, w_{n,j})'$$

$$\mathbf{Q} = \mathbf{X}'\mathbf{X} = \begin{bmatrix} n & q_{12} & \dots & q_{1k} \\ & n & \ddots & \vdots \\ & & \ddots & q_{k-1,k} \\ sym & & & n \end{bmatrix}$$

If $n \equiv 0 \pmod{4}$ the ϕ -optimal design consists $(k+1)$ columns of an Hadamard matrix $\mathbf{H}_n$ with the property $\mathbf{H}_n^T \mathbf{H}_n = n \mathbf{I}_n$, that is with orthogonal $n \times 1$ columns.

If $n \equiv 1, 2, 3 \pmod{4}$ the problem is also solved [Kounias and Farmakis (1984)], [Kounias et al.(1996)].

4. REPEATED MEASUREMENTS DESIGNS

In repeated measurement or crossover designs, every unit is exposed to one out of t treatments in every one of p consecutive periods. The notation is RMD(t, n, p), where t is the number of treatments, n is the number of experimental units and p is

the number of periods. [Hedayat and Afsarinejad (1975, 1978)], [Cheng and Wu (1980)], [Shah, K.R. and Sinha, B. K. (1989)], [Carriere and Huang (2000)].

Here we have two treatments denoted A, B, and three periods, the model is:

$$Y_{ij} = \sum_{j=1}^p \tau_{d(ij)} + \pi_j + \sum_{j=1}^p \delta_{d(i, j-1)} + \gamma_i + e_{ij}$$

$$i = 1, \dots, n, \quad j = 1, \dots, p, \quad \delta_{i,0} = 0$$

where i refers to the unit, j to the period; $\tau_{d(ij)} \in \{\tau_A, \tau_B\}$ is the direct effect of the treatment applied the j th period of the i th unit; $\delta_{i, j-1} \in \{\delta_A, \delta_B\}$ is the residual effect of the treatment applied at the $(j-1)$ th period of the i th unit, as defined by the design d , π_j is the j th period effect, and γ_i is the i th unit effect. The errors e_{ij} are independent within each unit and among units, and have a mean of 0 and constant variance σ^2 ; e_{ij} follow a continuous distribution.

After a transformation the model in vector form becomes,

$$E(\mathbf{Y}) = (\boldsymbol{\tau}_A - \boldsymbol{\tau}_B) \boldsymbol{\tau} + (\boldsymbol{\delta}_A - \boldsymbol{\delta}_B) \boldsymbol{\delta} + \boldsymbol{\pi}_1 \tilde{\pi}_1 + \dots + \boldsymbol{\pi}_p \tilde{\pi}_p + \boldsymbol{\gamma}_1 \gamma_1 + \dots + \boldsymbol{\gamma}_n \gamma_n$$

Where

$$\begin{aligned} \boldsymbol{\tau} &= (\tau_A - \tau_B) / 2, \boldsymbol{\delta} = (\delta_A - \delta_B) / 2, \\ \tilde{\pi}_j &= \pi_j + (\tau_A + \tau_B + \delta_A + \delta_B) / 2, \quad j = 2, \dots, p, \quad \tilde{\pi}_1 = \pi_1 + (\tau_A + \tau_B) / 2, \end{aligned}$$

We are interested in estimating the parameters $(\boldsymbol{\tau}, \boldsymbol{\delta})$.

If we are interested in some and not in all the parameters of a linear model $E(\mathbf{Y}) = \mathbf{X}\mathbf{b}$, then we can write the vector of parameters as $\mathbf{b} = (\mathbf{b}_1', \mathbf{b}_2')$, where $\mathbf{b}_1$ contains the r parameters of interest and $\mathbf{b}_2$ contains the $s-r$ remaining parameters. So we have $E(\mathbf{Y}) = \mathbf{X}_1 \mathbf{b}_1 + \mathbf{X}_2 \mathbf{b}_2$

From least squares the information matrix is, $\mathbf{Q} = \mathbf{X}_1' (\mathbf{I}_{pn} - P(\mathbf{X}_2)) \mathbf{X}_1$

where $P(\mathbf{X}_2) = \mathbf{X}_2 (\mathbf{X}_2' \mathbf{X}_2)^{-1} \mathbf{X}_2'$ and the variance matrix of the estimated parameters is $\text{var}(\hat{\mathbf{b}}_1) = \sigma^2 \mathbf{Q}^{-1}$.

4.1 φ optimal designs for 3 periods

Definition 4.1 A design is φ optimal if $\sum_{i=1}^k \varphi(\lambda_i) \leq \sum_{i=1}^k \varphi(\mu_i)$ for all continuous decreasing convex functions φ , where $(\lambda_1, \lambda_2, \dots, \lambda_p)$ are the latent roots of the optimal information matrix $\mathbf{Q}$ and $(\mu_1, \mu_2, \dots, \mu_k)$ are the latent roots of any other information matrix [Marshall and Olkin p. 10 (1979)].

Table 4.1 Number of units for each treatment

AAA	AAB	ABA	ABB	BAA	BAB	BBA	BBB
u_0	u_1	u_2	u_3	u_4	u_5	u_6	u_7

Here u_i is the number of units where the corresponding sequence is applied,
 $u_0 + u_1 + \dots + u_7 = n$.

Theorem 4.1 For estimating jointly the difference of treatment effects and residual effects, the φ optimal design for three periods is to take the sequences ABB and BAA:

- (i) In equal numbers if n is even .
- (ii) In numbers $(n \pm 1) / 2, (n \mp 1) / 2$ respectively if n is odd.

The estimated parameters are uncorrelated and the variances of the estimated parameters are:

$$\frac{1}{\sigma^2} \text{var}(\hat{\tau}_A - \hat{\tau}_B) = \begin{cases} 3 / (2n) & \text{if } n \text{ is even} \\ 3n / 2(n^2 - 1) & \text{if } n \text{ is odd} \end{cases}$$

$$\frac{1}{\sigma^2} \text{var}(\hat{\delta}_A - \hat{\delta}_B) = \begin{cases} 2 / n & \text{if } n \text{ is even} \\ 2n / (n^2 - 1) & \text{if } n \text{ is odd} \end{cases}$$

[Kounias and Chalikias (2014)].

Theorem 4.2. The optimal design for estimating only the treatment effects or only the residual effects is the same as in the previous theorem 4.1.

ΠΕΡΙΛΗΨΗ

Δίνεται μια συνοπτική περιγραφή βέλτιστων πειραματικών σχεδιασμών. Το γραμμικό μοντέλο, διακριτοί βέλτιστοι σχεδιασμοί με παραδείγματα, παραγοντικοί σχεδιασμοί, βέλτιστοι σχεδιασμοί επαναλαμβανόμενων μετρήσεων.

REFERENCES

- Hedayat, A & Afsarinejad, K. (1975). Repeated measurements designs I. *A survey of Statistical Design and Linear Models*, Srivastava, J.N. (ed.), pp. 229-242.
- Hedayat, A & Afsarinejad, K. (1978). Repeated measurements designs II. *The Annals of Statistics*, 18, 1805-1816.
- Cheng, C. S. & Wu, C. F. (1980). Balanced repeated measurements designs. *The Annals of Statistics*,. 11, 29-50. *Correction (1983) 11, 349*.
- Carriere, K. C. & Huang, R. (2000). Crossover designs for two-treatment clinical trials. *Journal of Statistical Planning and Inference*. 87, 125-134.
- Chalikias and Kounias (2014). Optimal two Treatment Repeated Measurement Designs for three Periods, *Communications in Statistics, Theory and Methods*, accepted for publication
- Kounias, S., Lefkopoulou, M., Bagiatis, C. (1983). G-optimal N-observation first order 2^k designs. *Discrete Math*. 46, no. 1, 21-31.
- Kounias, Stratis; Farmakis, Nikos (1984). A construction of D-optimal weighing designs when $N \equiv 3 \pmod{4}$. *J. Statist. Plann. Inference* 10, no. 2, 177-187.
- Kounias, S., Koukouvinos, C., Nikolaou, N., Kakos, A. (1996). The non-equivalent circulant D-optimal designs for $n=90$. *J. Statist. Plann. Inference* 53, no. 2, 253-259.
- Marshall A, Olkin I.(1979) *Inequalities: Theory of Majorization and Its Applications*. Academic Press. New York.
- Shah,K.R. and Sinha, B. K.(1989).*Theory of Optimal Designs*. Lecture Notes in Statist. 54, Springer-Verlag, NewYork



LONG RANGE CORRELATIONS AND PROBABILITY DISTRIBUTION OF THE EARTHQUAKE INTEREVENT TIME IN GREECE

Kourouklas Ch¹., Papadimitriou E¹., Karakostas V.¹ and Tsaklidis G.²

¹ Aristotle University of Thessaloniki, Geophysics Department, 54124, Thessaloniki, Greece

² Aristotle University of Thessaloniki, School of Mathematics, 54124, Thessaloniki, Greece

ckouroukl, ritsa, vkarak@geo.auth.gr
tsaklidi@math.auth.gr

ABSTRACT

Temporal properties of earthquake occurrence constitute an important component in the development of earthquake probabilistic models. Since the earthquake occurrence is neither periodic nor completely random but often clustered in time, the detection of any correlations between earthquake interevent times, T_r , namely the existence of long range memory between them is required for the application of stochastic processes. The Autocorrelation and Partial Autocorrelation functions along with the Ljung–Box Q–test are efficient in the detection of these correlations. The methodology is applied on multiple samples by dividing the earthquake catalogues with $M \geq 4.1$ since 1975 for three different regions of Greek territory (Central Ionian Islands, Corinth Gulf and North Aegean Sea) into subsets with different thresholds, M_c , aiming to detect M_c above which the earthquake occurrence can be considered to be random in time. Four of the most popular statistical distributions (Weibull, Gamma, lognormal and exponential) are applied on those subsets that the T_r exhibits independent behavior in order to select the one with the best performance according to the Anderson–Darling goodness of fit test and the Akaike and Bayesian Information Criteria.

Keywords: earthquake interevent time, long range correlations, distribution fitting

1. INTRODUCTION

The study of temporal properties of earthquake occurrence contributes in analyzing the seismicity of a specific region, in order to develop earthquake probabilistic models, which constitute an indispensable component of seismic hazard assessment. This kind of models requires assuming the time between successive earthquakes as a point process. For such applications four characteristics must be qualified: 1) the

number of events occurring in two different time intervals are independent of each other, 2) the probability distribution of the number of events falling in one time interval depends solely on the length of the time interval, 3) there is never an occasion when two or more events occur simultaneously and 4) especially for renewal models the occurrence of the next event depends only on the time of the last event but not on the full history [Zhuang et al. (2009)].

Because the earthquakes occur neither periodically nor completely randomly but often clustered in time, the detection of any correlations between their interevent times or else the existence of long range memory between them is required before applying any stochastic procedure (modeling etc). Recent studies on interevent times of large earthquakes ($M \sim 8$) on a global scale, which might be considered as rare events, evidence that their occurrence times are characterized by lack of memory and can be described by a Poisson process [Parsons and Geist (2012)].

In contrast, interevent times of small to moderate events are characterized by existence of memory [Gkarlaoui et al. (2014)]. Consequently, it is important to detect at which magnitude threshold the sample of interevent times can be considered random in time before investigating the probability distribution by which they can modeled. For this purpose, many statistical distributions are proposed such as the Weibull [Abaimov et al. (2008)], the lognormal [Nishenko and Buland (1987); Werner et al. (2011)], the stretched exponential [Altmann and Kanz (2005)] and the Gamma distribution [Corral (2003)]. Investigation of T_r for earthquakes above various threshold magnitudes, M_c , in different regions show that they follow a universal scaling law independently of the region and the M_c [Corral (2004)].

In this study, the detection of the threshold magnitude, M_c , above which the earthquake interevent times might be considered uncorrelated, is performed by means of the Autocorrelation and Partial Autocorrelation functions along with the Ljung–Box Q–test in three of the most active regions of Greek territory, the Central Ionian Islands, the Corinth Gulf and the North Aegean Sea. Furthermore, four statistical distributions (Weibull, lognormal, Gamma and exponential) are applied to the samples of interevent times with magnitude greater than the M_c , which arise from the application of the previous methodology, in order to investigate the distribution which performs better in each region.

2. DATA AND PROCESSES

In order to detect the threshold magnitude above which T_r may be considered to be random in time for each region of interest (in other words they appear to be uncorrelated) we compiled three different catalogues of crustal earthquakes with $M \geq 4.1$ since 1975 from the regional catalogue of Geophysics Departments of the Aristotle University of Thessaloniki (<http://geophysics.geo.auth.gr/ss>). The catalogues concerning the areas of Central Ionian Islands, the Corinth Gulf and the North Aegean Sea, include 1362, 450 and 1547 events, respectively (Fig. 1 – 3).

Figure 1. Spatial distribution of earthquakes with $M \geq 4.1$ in Central Ionian Islands since 1975. White, green and red circles depict the $4.0 \leq M < 5.0$, $5.0 \leq M < 6.0$ and $6.0 \leq M < 7.0$ earthquakes, respectively. Yellow star depicts the $M=7.0$ earthquake of 1983.

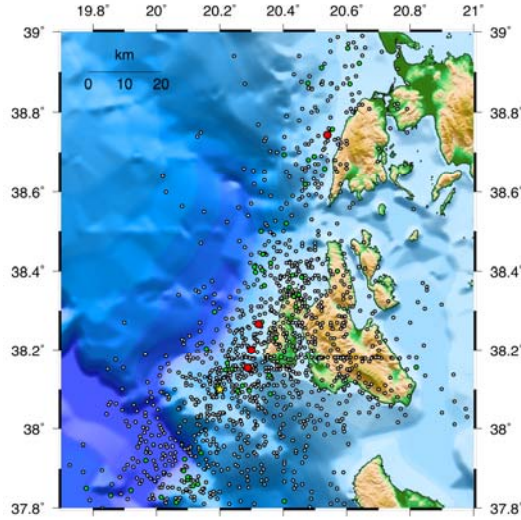


Figure 2. Spatial distribution of earthquakes with $M \geq 4.1$ in Corinth Gulf since 1975. White, green and red circles depict the $4.0 \leq M < 5.0$, $5.0 \leq M < 6.0$ and $6.0 \leq M < 7.0$ earthquakes, respectively.

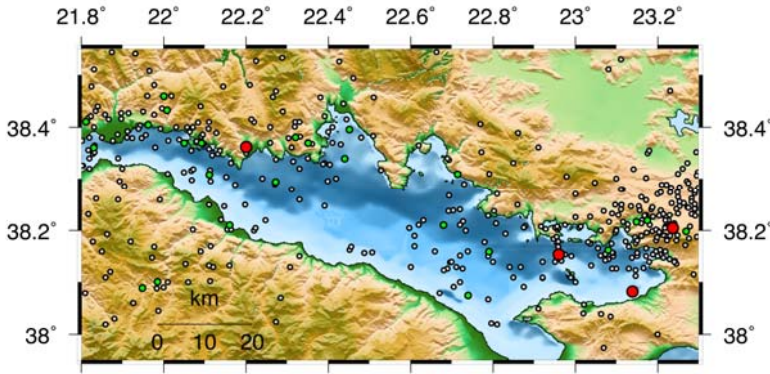
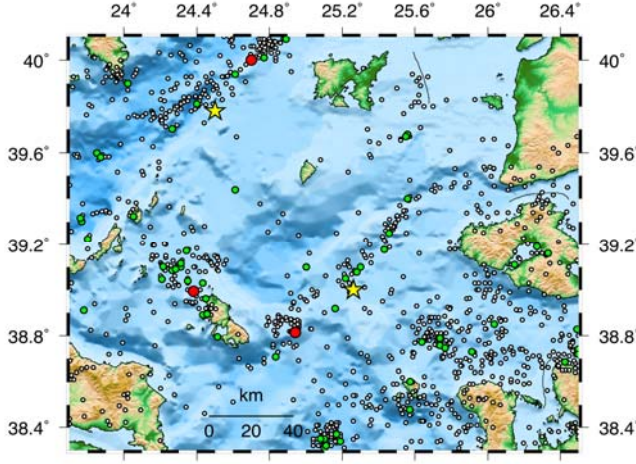


Figure 3. Spatial distribution of earthquakes with $M \geq 4.1$ in North Aegean Sea since 1975. White, green and red circles depict the $4.0 \leq M < 5.0$, $5.0 \leq M < 6.0$ and $6.0 \leq M < 7.0$ earthquakes, respectively. Yellow stars depict the $M \geq 7.0$ earthquakes.



From these data samples the interevent times $T_{r,i}$ in days are estimated as

$$T_{r,i} = T_{i+1} - T_i, \quad i = 1, 2, \dots, \quad (1)$$

for earthquakes with magnitude thresholds of 0.1 bin increment, starting from $M \geq 4.1$.

3. METHODOLOGY

3.1 Autocorrelation of interevent times samples

The threshold magnitude at which the interevent times become memoryless is detected through the investigation of the long range correlation between the values of each subset of T_r , which are created as mentioned above. This correlation is examined by calculating the Autocorrelation (ACF) and Partial Autocorrelation (PACF) functions and by applying the Ljung–Box Q–test. The Autocorrelation function,

$$\rho_k = \frac{\sum_{i=1}^{n-k} (x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (2)$$

examines correlations between past and future values of a given sample of time series where n is the number of the observations, k is the number of the lags and $\bar{x}$ is the mean value. Then, the correlations detected by the ACF can be confirmed through the values of the Partial Autocorrelation function

$$r_k = \frac{\rho_k - \sum_{j=1}^{k-1} r_{k-1,j} \rho_{k-j}}{1 - \sum_{j=1}^{k-1} r_{k-1,j} \rho_j} , \quad (3)$$

where, $r_{k,j} = r_{k-1,j} - r_{k,k} r_{k-1,k-j}$, for $j = 1, 2, \dots, k$.

Another quantitative way to examine autocorrelation of a time series is the Ljung–Box Q–test, which is a modified version of the Box–Pierce Portmanteau test. This is implemented under the null hypothesis that the time series exhibits no autocorrelation for a fixed number of lags L , against the alternative that some (statistically significant) autocorrelation exists. The test statistic is given by

$$Q = n(n+2) \sum_{k=1}^L \left(\frac{\rho_k^2}{n-k} \right). \quad (4)$$

while, L takes one by one all the possible values until the proper lag that no autocorrelation is identified. Under the null hypothesis, the asymptotic distribution of Q is chi-square with L degrees of freedom; if the statistic of the test for given lag L is less than the critical value of the chi-square ($Q_{\text{stat}} < Q_{\text{critical}}$) the null hypothesis cannot be rejected.

3.2 Distribution fitting and selection

After the detection of the threshold magnitude above which the interevent times become uncorrelated, four of the most popular statistical distributions are applied to each data set, namely the Weibull, the Gamma, the lognormal and the exponential. The parameter estimation for each distribution is achieved via the application of the Maximum Likelihood Estimation (MLE) method using the respective log-likelihood formulae (Johnson *et al.*, 1994).

In order to compare the distributions and select the best one fitting the data, the Anderson–Darling non-parametric goodness of fit test (A–D test) is applied. The A–D test is implemented by calculating the distance, A^2 , between the empirical cumulative distribution function (ecdf) and the cumulative distribution function (cdf) for the distributions applied to the data, where

$$A^2 = -n - \sum_{i=1}^n \frac{2i-1}{n} [\ln F(x_i) + \ln(1 - F(x_{n+1-i}))] , \quad (5)$$

and $\{x_1, x_2, \dots, x_n\}$ are the ordered sample data points, n is the number of observations and F is the cdf of the distribution under study. The test compares the factor A^2 with a critical value, c , under the null hypothesis that the data are distributed according to F . If the factor A^2 is less than or equal to the critical value, then the null hypothesis cannot be rejected. Also, the Akaike and Bayesian Information Criteria (AIC and BIC

respectively) are calculated. From these calculations the distribution that displays the best performance is the one with the minimum value of the criterion in both cases. The values of AIC and BIC are given by

$$AIC = -2 \ln L + 2k \quad (6)$$

$$BIC = -2 \ln L + k \ln n \quad (7)$$

where $\ln L$ stands for the log-likelihood function, k is the number of parameters of the distribution and n is the number of the observations.

4. APPLICATION AND RESULTS

The methodology described in Section 3.1 is applied to each data subset in order to detect for which threshold M_c the earthquake interevent times correlations become statistically insignificant.

The values of ACF for the Central Ionian Islands (Fig. 4, left) exhibit strong positive autocorrelation of interevent times until magnitude reaches the threshold $M_c=4.4$. From this M_c onwards, the autocorrelation becomes weaker until $M_c \geq 5.0$ when no significant autocorrelations of any lag are recorded, and the relevant samples of T_r can be considered as uncorrelated. The values of PACF (Fig. 4, right) confirm this behavior for $M_c \geq 5.0$, except for the case $M_c \geq 5.2$, for which significant partial correlations are recorded at lags 14, 15, 16. The results of the Ljung-Box Q-test (Fig. 5) show that the null hypothesis can be rejected for the samples of T_r till to $M_c = 4.9$ (without any exception for $M_c \geq 5.2$).

For the Corinth Gulf both the values of ACF and PACF show that the subsets of T_r exhibit significant correlations till $M_c = 4.7$ (Fig. 6). These results are in agreement with those of the Ljung-Box test that are shown in Figure 7. An exception from the results is detected both from ACF/PACF values and the Q-test for the subset of $M_c \geq 5.0$ which might not be considered as statistically significant due to the sample shortage (28 observations). The values of ACF and PACF values for North Aegean Sea subsets indicate that $M_c \geq 5.0$, and this threshold is the one where the interevent times start to be uncorrelated. The Ljung-Box test confirms these results (Fig. 8 and 9).

Figure 4. ACF (left), PACF (right) values and the 95% confidence intervals (blue lines) for the Central Ionian Islands, for magnitude thresholds varying from 4.1, 4.2, ... to 5.2.

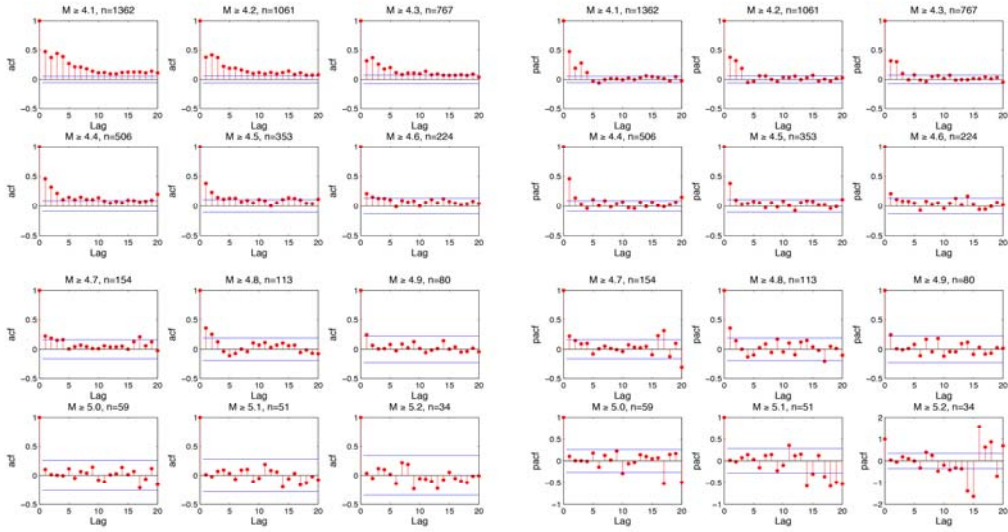


Figure 5. Q statistics of Ljung–Box test (red lines) and the critical values for the χ^2 distribution (blue lines) for the Central Ionian Islands, for $M_c=4.1, 4.2, \dots, 5.2$.

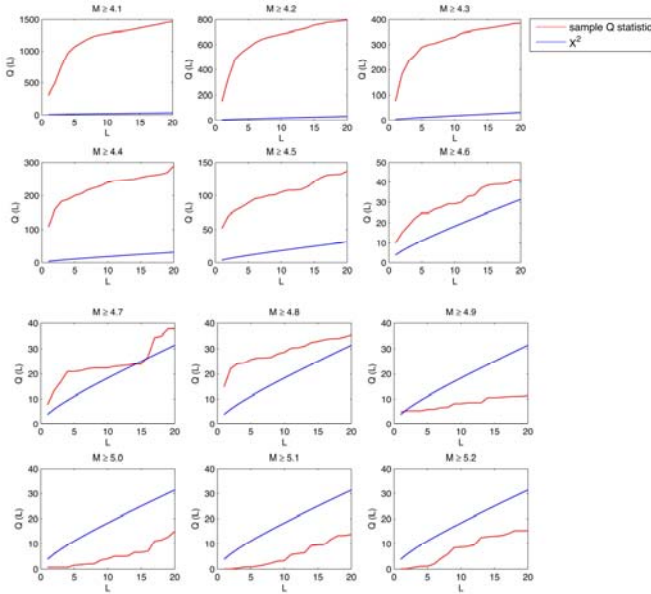


Figure 6. ACF (left), PACF (right) values and the 95% confidence intervals (blue lines) for the Corinth Gulf, for $M_c=4.1, 4.2, \dots, 5.2$.

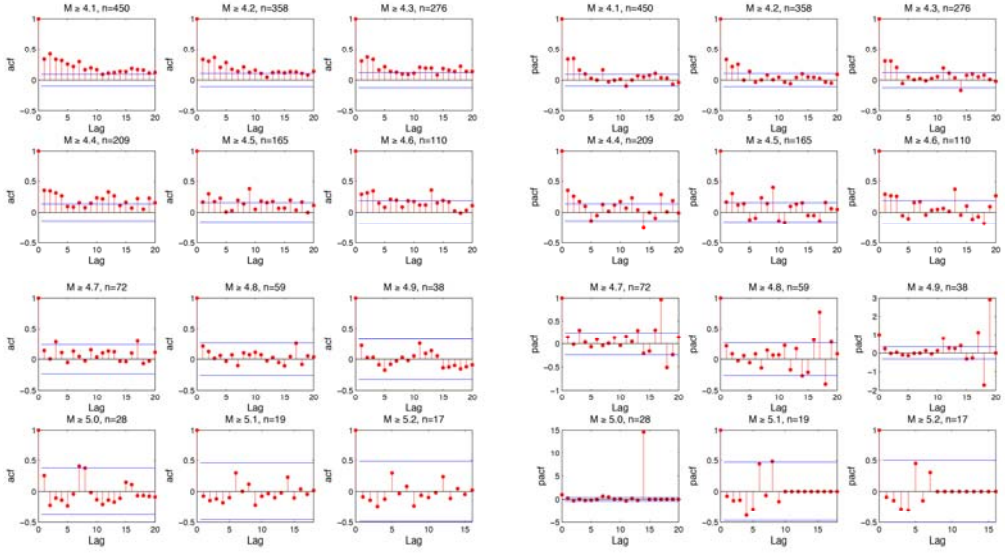


Figure 7. Q statistics of Ljung–Box test (red lines) and the critical values from the χ^2 distribution (blue lines) for the Corinth Gulf, for $M_c=4.1, 4.2, \dots, 5.2$.

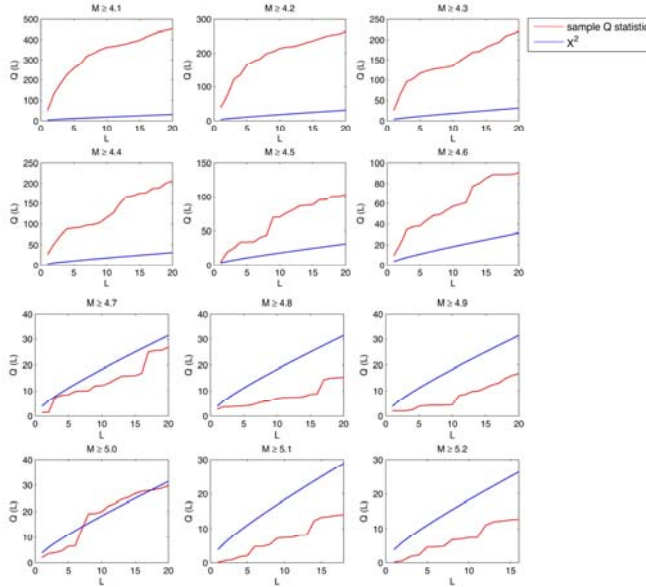


Figure 8. ACF (left), PACF (right) values and the 95% confidence intervals (blue lines) for the North Aegean Sea, for $M_c=4.1, 4.2, \dots, 5.2$.

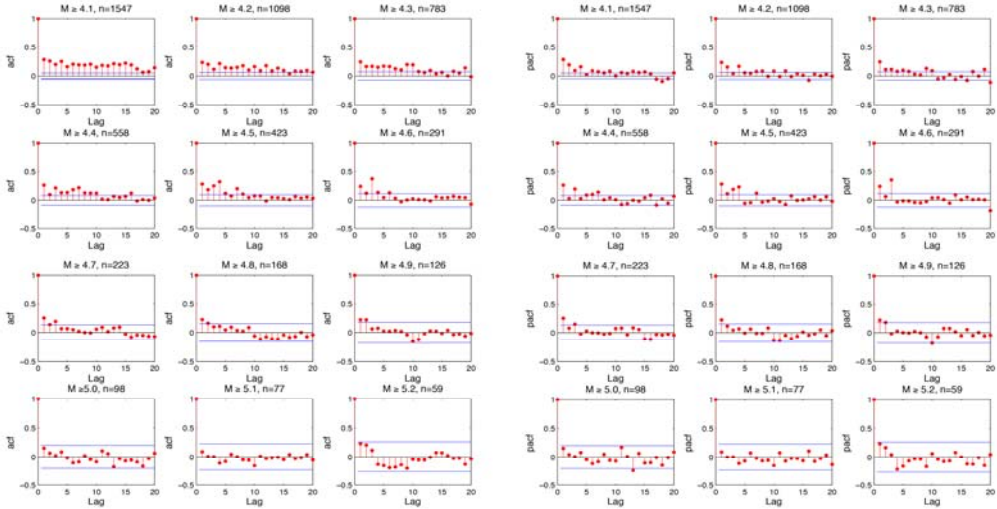
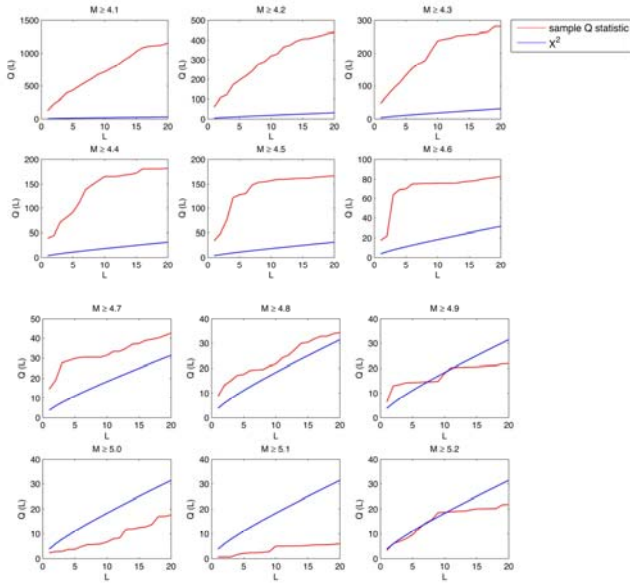


Figure 9. Q statistics of Ljung–Box test (red lines) and the critical values from the χ^2 distribution (blue lines) for the North Aegean Sea, for $M_c=4.1, 4.2, \dots, 5.2$.



Summing up the results of correlation analysis, the magnitude thresholds above which the earthquake interevent times appear not to be correlated, equals $M_c=5.0$ with 59

observations for the Central Ionian Islands, while $M_c=4.7$ with 72 observations for the Corinth Gulf, and $M_c=5.0$ with 98 observations for the North Aegean Sea.

In these three data sets, the Weibull, Gamma, lognormal and the exponential distributions are applied. The parameters of each distribution are estimated by means of the log likelihood functions, and the values of $\log L$ for each one of the distributions are recorded. From the parameter estimation, it comes out that the confidence intervals in the first two cases (Central Ionian Islands and Corinth Gulf) have considerably larger range than those of the third case, most probably due also to the data shortage in these two cases (59 and 72 observations, respectively, versus 98 in the third case).

The Anderson-Darling test is then applied to each sample in order to compare the distributions derived via the MLE parameter estimates and empirical cdf (Fig. 10). In Central Ionian Islands, the null hypothesis cannot be rejected for the Weibull, the Gamma and the lognormal distribution, while the exponential should be rejected. The Corinth Gulf sample, evidence that the statistic of the test is less than the critical values ($c=2.49$) and consequently the null hypothesis cannot be rejected for the Weibull ($A^2=2.01$) and the Gamma ($A^2=0.92$) distribution. For the North Aegean Sea sample, the null hypothesis cannot be rejected only for the Gamma distribution, which is the one that exhibits the best fitting to the data.

The Akaike and Bayesian Information Criteria values are also calculated for each distribution and for the three data sets and are given in Tables 1 through 3. The distribution with the lower values of both criteria in all cases is the Gamma distribution.

Summarizing the results of the comparison of the distributions, we get that, although more than one of them can be accepted according to the A-D test in the first two cases (i.e., the Central Ionian Islands and Corinth Gulf), the values of the statistic are considerable smaller for the Gamma distribution. On the other hand, in the case of the North Aegean Sea sample, only the Gamma distribution can be accepted. The values of information criteria further confirm that Gamma distribution best fits the data, since the relevant values are the lowest of the four. Combining the results of A–D test with those of the information criteria it was found that the Gamma distribution fits better than the other three distributions (all) the data sets.

Table 1. MLE parameters estimates, 95% confidence intervals, log likelihood calculation, A-D test statistics along with the critical value of the test and AIC, BIC calculations for the Central Ionian Islands.

Distribution	Parameters	Conf. Intervals	Log L	A-D test	AIC	BIC
				$c=2.45$		
Weibull	$\alpha=64.71$ $b=0.33$	[28.96,144.61] [0.27, 0.41]	-300.7	2.20	605	610

Gamma	$\kappa=0.23$ $\theta=1026$	[0.18, 0.31] [564.98, 1864.39]	-298.8	1.82	601	606
Lognormal	$\mu=2.41$ $\sigma=3.66$	[1.46, 3.36] [3.09, 4.47]	-302.2	2.21	608	613
Exponential	$\mu=240.24$	[189.04, 315.59]	-382.4	53.37	766	769

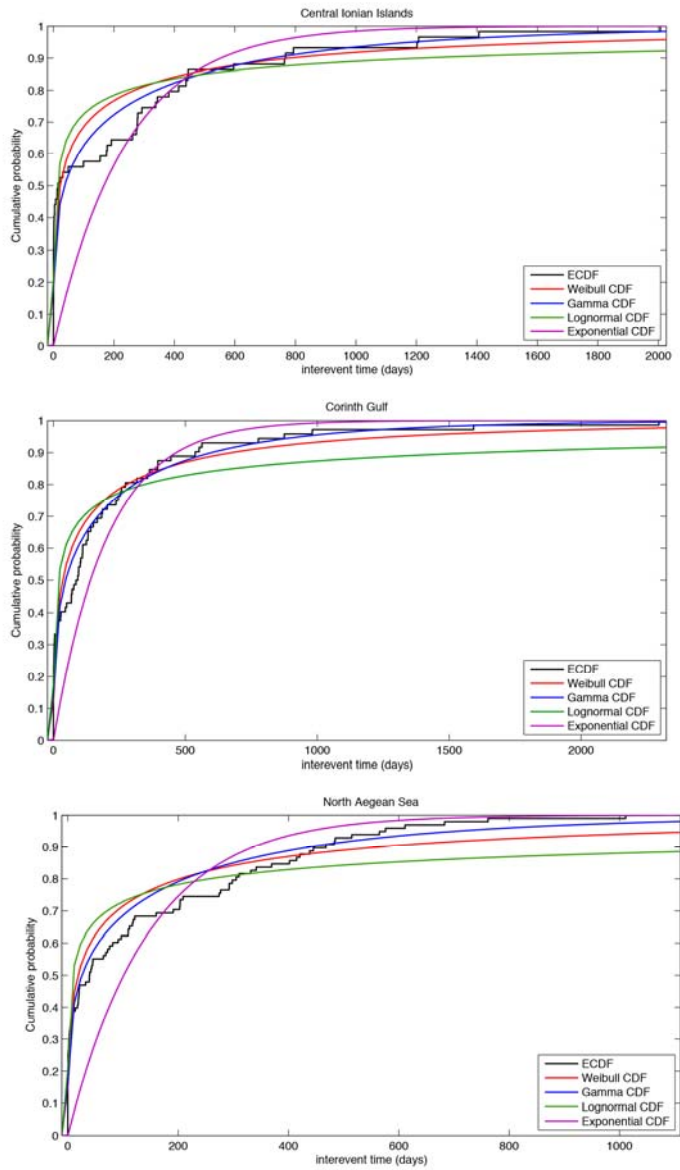
Table 2. MLE parameters estimates, 95% confidence intervals, log likelihood calculation, A-D test statistics along with the critical value of the test and AIC, BIC calculations for the Corinth Gulf.

Distribution	Parameters	Conf. Intervals	Log L	A-D test	AIC	BIC
				$c=2.49$		
Weibull	$\alpha=85.36$ $b=0.40$	[46.88, 155.45] [0.33, 0.49]	-390.4	2.01	785	789
Gamma	$\kappa=0.29$ $\theta=704.84$	[0.22, 0.37] [426.59, 1164.60]	-386.5	0.92	777	782
Lognormal	$\mu=2.88$ $\sigma=3.51$	[2.06, 3.71] [3.02, 4.20]	-399.9	3.48	804	808
Exponential	$\mu=202.05$	[162.44, 258.24]	-454.2	30.71	910	913

Table 3. MLE parameters estimates, 95% confidence intervals, log likelihood calculation, A-D test statistics along with the critical value of the test and AIC, BIC calculations for North Aegean Sea.

Distribution	Parameters	Conf. Intervals	Log L	A-D test	AIC	BIC
				$c=2.49$		
Weibull	$\alpha=52.24$ $b=0.35$	[28.94, 94.30] [0.30, 0.41]	-472.9	3.12	950	955
Gamma	$\kappa=0.25$ $\theta=584.14$	[0.20, 0.32] [371.72, 917.93]	-464.3	1.65	933	938
Lognormal	$\mu=2.14$ $\sigma=4.04$	[1.33, 2.95] [3.55, 4.71]	-485.58	4.89	975	980
Exponential	$\mu=146.05$	[120.95, 179.90]	-586.4	58.75	1175	1177

Figure 10. Comparison of the empirical cdf (ECDF) and the estimated (theoretical) cdf for each distribution applied to the Central Ionian Islands (top), Corinth Gulf (center) and North Aegean Sea (bottom) interevent time samples.



5. CONCLUSIONS

The detection of the threshold magnitude, M_c , above which the earthquake interevent times, T_r , might be considered as uncorrelated, is investigated through the ACF and PACF values along with the application of the Ljung–Box Q–test in three regions of the Greek territory. The results of this analysis reveal that the M_c characterizes the seismotectonic regions and is equal to 5.0 for the Central Ionian, to 4.7 for the Corinth Gulf and to 5.0 for the North Aegean Sea. In the derived data samples, the Weibull, the Gamma, the lognormal and the exponential distributions are applied for the interevent times and compared in order to model the behavior of T_r . The A–D goodness of fit test is used for examining which of them performs better.

In the first two cases, the test allows acceptance of more than two distributions (i.e., the Weibull, the Gamma and the lognormal for Central Ionian Islands and the Weibull and the Gamma for the Corinth Gulf). In contrast, in the third case all but the Gamma distribution are rejected. For further evaluation, the values of AIC and BIC criteria are calculated, and they have indicated that the Gamma distribution exhibits the best performance. This result is in agreement with Corral’s analysis (2004) where the earthquake interevent times follow a universal scaling law independently of the region and the M_c , and can be modeled by the Gamma distribution.

ΠΕΡΙΛΗΨΗ

Οι χρονικές ιδιότητες της γένεσης των σεισμών συνιστούν σημαντική παράμετρο για τη δημιουργία πιθανοκρατικών μοντέλων σεισμικότητας. Καθώς η γένεση των σεισμών δεν είναι ούτε περιοδική αλλά ούτε και εντελώς τυχαία στο χρόνο αλλά με εμφάνιση συχνά συσταδοποίησης, η εξακρίβωση των πιθανών φαινομένων συσχέτισης των χρονικών διαστημάτων μεταξύ των σεισμών, δηλαδή της ύπαρξης μακράς μνήμης μεταξύ τους, είναι σημαντική για τη διαδικασία της σεισμογένεσης καθώς και (ειδικότερα) για τη χρήση στοχαστικών διαδικασιών στη Σεισμολογία (μοντελοποίηση, εκτίμηση σεισμικής επικινδυνότητας κλπ). Οι συναρτήσεις Αυτοσυσχέτισης και Μερικής Αυτοσυσχέτισης σε συνδυασμό με τον έλεγχο Ljung – Box αποτελούν εργαλεία για την εξακρίβωση τέτοιου είδους συσχετίσεων. Η μεθοδολογία αυτή εφαρμόστηκε σε πολλαπλά δείγματα χωρίζοντας τον κατάλογο σεισμών με μέγεθος $M \geq 4.1$ από το 1975 για τρεις διαφορετικές περιοχές του ελληνικού χώρου (Κεντρικά Ιόνια Νησιά, Κορινθιακός κόλπος και Βόρειο Αιγαίο) σε μικρότερα δείγματα με βάση διαφορετικά κατώφλια μεγέθους, M_c , ($M_c = 4.1, 4.2, \dots, 5.2$) με στόχο την εξακρίβωση του κατωφλίου πάνω από το οποίο η γένεση των σεισμών μπορεί να θεωρηθεί τυχαία στο χρόνο. Τέσσερις από τις πιο διαδεδομένες στατιστικές κατανομές (Weibull, Γάμμα, λογαριθμοκανονική και εκθετική) εφαρμόστηκαν στα δείγματα των χρονικών διαστημάτων που διαπιστώθηκε ότι είναι στατιστικά ανεξάρτητα μεταξύ τους, με σκοπό την εξακρίβωση της κατανομής που εμφανίζει την καλύτερη προσαρμογή, σύμφωνα με τον έλεγχο καλής προσαρμογής Anderson–Darling και τα κριτήρια πληροφορίας AIC και BIC.

Acknowledgments: The maps were made using the Generic Mapping Tool (www.soest.hawaii.edu/gmt, Wessel and Smith 1998).

REFERENCES

- Abaimov, S. G., D. L. Turcotte, R. Shcherbakov, J. B. Rundle, G. Yakovlev, C. Goltz, and W. I. Newman (2008). Earthquakes: Recurrence and interoccurrence Times, *Pure Appl. Geophys.*, **165**, 777-795. doi:/10.1007/s00024-008-0331-y.
- Altmann E. G. and H. Kantz (2005). Recurrence time analysis, long-term correlations, and extreme events, *Phys. Rev. E*, **71**, 1-9, doi:10.1103/PhysRevE.71.056106.
- Box, G. E. P. and D. Pierce (1970). Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models. *J. Am. Stat. Assoc.*, **65**, 1509–1526.
- Box, G. E. P., G. M. Jenkins, and G. C. Reinsel (1994). *Time Series Analysis: Forecasting and Control*, pp. 575, 3rd ed., Prentice Hall, Englewood Cliffs, NJ.
- Bunde A., J. F. Eichner, S. Halvin and J. W. Kantelhardt (2003). The effect of long-term correlations on the return period of rare events, *Physica A*, **330**, 1-7, doi:10.1016/j.physa.2003.08.004.
- Corral, A. (2003). Local distribution and rate fluctuations in a unified scaling law for earthquakes, *Phys. Rev. E.*, **68**, 035102, doi: 10.1103/PhysRevE.68.035102.
- Corral, A. (2004). Long-term Clustering, Scaling, and Universality in the Temporal Occurrence of Earthquakes, *Phys. Rev. Lett.*, **92**, 108501, doi:10.1103/PhysRevLett.92.10501.
- Gkarlaouni, C., S. Lasocki, G. Tsaklidis, and V. Karakostas (2014). Investigation of short-term and long-term memory of earthquake occurrence, *PROCEEDINGS of the 27th Panhellenic Statistics Conference*, 53-67.
- Gkarlaouni, C., S. Lasocki, and E. Papadimitriou (2015). Investigation of earthquake magnitude and interevent time distribution in Corinth Gulf and Mygdonia Basin with the use of stochastic tools, *PROCEEDINGS of the 28th Panhellenic Statistics Conference*, 385-399.
- Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions, vol.1*, 784 pp, Wiley.
- Ljung, G. and G. E. P. Box (1978). On a Measure of Lack of Fit in Time Series Models, *Biometrika*, **66**, 67–72.
- Nishenko, S.P and Buland, R. (1987). A generic recurrence interval distribution for earthquake forecasting. *Bull. Seismol. Soc. Am.*, **77** (4), 1382-1399.
- Parsons, T. and E. L. Geist (2012). Were Global $M \geq 8.3$ Earthquake Time Intervals Random between 1900 and 2011?, *Bull. Seism. Soc. Am.*, **102**, 1583-1592, doi:10.1785/0120110282.
- Permanent Regional Seismological Network operated by the Aristotle University of Thessaloniki, doi:10.7914/SN/HT
- Saichev, A., and D. Sornette (2007). Theory of earthquake recurrence times, *J. Geophys. Res.*, **112**, B04313, doi:10.1029/2006JB004536.

- Weglarczyk, S. and Lasocki, S. (2009). Studies of short and long memory in mining-induced seismic processes, *Acta Geophysica*, special section: Triggered and Induced Seismicity, **57**, 3, 696-715.
- Werner, M. J., Ide, K., and Sornette, D.: Earthquake forecasting based on data assimilation: sequential Monte Carlo methods for renewal point processes, *Nonlin. Processes Geophys.*, **18**, 49-70, doi:10.5194/npg-18-49-2011, 2011
- Wessel, P. and Smith, W. H. F., 1998. New, improved version of the generic mapping tools released, *Trans. Am. Geophys. Union*, **79**, 579.
- Zhuang, J., D. Harte, M. J. Werner, S. Hainzl, and S. Zhou (2012). Basic models of seismicity: Temporal models, *Community Online Resource for Statistical Seismicity Analysis*, doi:10.5078/corssa-79905851.

USING KALMAN FILTER AND TOBIT KALMAN FILTER IN ORDER TO IMPROVE THE MOTION RECORDED BY KINECT SENSOR II

K. Loumponias^{1,2}, N. Vretos¹, P. Daras¹, G. Tsaklidis²

¹Centre for Research and Technology Hellas, Information Technologies Institute, Thessaloniki, Greece.

{loumponias, vretos, daras}@iti.gr

²Department of Mathematics, Aristotle University of Thessaloniki, Thessaloniki, Greece.
{kostikasl, tsaklidi}@math.auth.gr

ABSTRACT

In this paper, two methods are proposed to analyse skeleton data recorded by the Kinect v2 sensor using the Kalman filter and the Tobit Kalman filter in order to minimize the noise of the acquisition device due to occlusions, self occlusions e.t.c. The skeleton data are three-dimensional spatial coordinates that record movements of an individual's joints. The variance of the noise process is estimated using the likelihood function. In order to include into the model restrictive conditions based on the joints displacements per frame, we apply the Tobit Kalman Filter. Experiments on skeleton data show that the Tobit Kalman filter corrects better the noise than the standard Kalman filter.

Keywords: Kinect Sensor, Skeleton Data, Standard Kalman Filter, Tobit Kalman Filter.

1. INTRODUCTION

Skeleton tracking motion is a scientific area studied by means of depth sensors (i.e., sensors using the depth coordinate as basic coordinate), proved to be very useful in many applications, such as monitoring of activity recognition [Zhen et al (2015), Zhang et al (2013)] and health tracking [Galna et al (2014)]. In the present paper it is shown that the Kinect 2 sensor is able to achieve skeleton tracking performance in a low-cost manner for activity recognition; this sensor is able to track at most 25 joints.

However, Kinect sensor generates a lot of noise because of self-occlusion and lack of accuracy in fast movements. Especially when the skeleton's joints are occluded, they often appear to be shifted in a no reasonable manner.

The method presented in this paper is based on appropriately smoothing the joints' spatial coordinates. In order to smooth the coordinates, various common stochastic filters are used, e.g. filters based on moving average, Savitzky-Golay

filter e.t.c. [Yong et al (2015)], without any restrictions in joints' movements. Now we want to develop a model which will not allow the joints to move abnormally, and without affect the real movements. For this reason, we studied the joints' speeds by carrying out various experiments using groundtruth sensors. Then we applied the Tobit Kalman filter by taking into account the speeds restrictions.

In [Sungphil et al (2016)] a method based on the use of multiple Kinect sensors for skeleton tracking is proposed. They achieve in determining the reliability of each 3D joint position by employing a data fusion method based on Kalman filtering using multiple Kinect sensors. They take into account the measurement variance of noise for determining the contribution of an observation to the fused measurement. Additionally, they explain how to estimate the measurement variance for each one of the measurements. Finally, they present the average 3D position error of ten activities produced by their method, by a single Kinect and the average derived by multiple Kinect sensors, respectively. Almost in all cases, their method appears to give better results than the standard Kalman filter.

In [Berti et al (2014)] Kalman filtering is applied for robotic arms tracked by Kinect sensors. They denoise only the depth coordinate using Kalman filter methodology but they do not explain clearly the estimation process concerning the matrices involved. In their results presented through figures, it is obvious that the data are denoised, however the error reduction is not evaluated. In [Kong et al (2013)] a Kalman filter is described briefly to smooth 2D movements of a joint. In contrast with the aforementioned methods, the video data are derived via CCTV (Closed Circuit Television). Thus, they describe in which way they construct their joints' body model. Other scientists who are dealing with activity recognition via neural networks, use a simple Savitzky-Golay smoothing filter in order to correct the data [Yong et al (2015), Wentao et al (2016)]. This method is based on the previous, as well as the current and the two following observations.

In Section 2 the Kalman filter procedure along with the related likelihood function is provided. In Section 3 we describe the Tobit Kalman filter. In Section 4 the Kalman filter and Tobit Kalman filter approach for skeleton tracking is established. Finally, in Section 5, conclusions are presented.

2. KALMAN FILTER

In this section we describe briefly the Kalman filter (KF) [Peter (1979)], abbreviated as KF, which is used for estimating the unknown state vector $\mathbf{x} \in \mathbb{R}^n$, of a discrete-time stochastic process that is governed by the linear stochastic difference equation

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{w}_k \quad (1)$$

with a measurement (observation) $\mathbf{y} \in \mathbb{R}^m$ given by

$$\mathbf{y}_k = \mathbf{H}\mathbf{x}_k + \mathbf{v}_k, \quad (2)$$

where $\mathbf{w}_k \sim N(\mathbf{0}, \mathbf{Q}_k)$, $\mathbf{v}_k \sim N(\mathbf{0}, \mathbf{R}_k)$, and $N(\mu, \Sigma)$ denotes the normal distribution with mean value μ and covariance matrix Σ . The matrices $\mathbf{A}, \mathbf{H}$ are the transition and observation matrices, respectively.

We define by $\hat{\mathbf{x}}_k^-$ the a priori state estimate at step k by assuming knowledge of the process history prior to step k and $\hat{\mathbf{x}}_k$ the a posteriori state estimate at step k by assuming that the measurement $\mathbf{y}_k$ is given.

The process of KF evolves in two stages: the predict stage and the update stage, determined by the associated equations:

The Predict Stage

$$\hat{\mathbf{x}}_k^- = \mathbf{A}\hat{\mathbf{x}}_{k-1}, \quad (3)$$

$$\mathbf{P}_k^- = \mathbf{A}\mathbf{P}_{k-1}\mathbf{A}^T + \mathbf{Q}_k. \quad (4)$$

where $\mathbf{P}_k^-$ and $\mathbf{P}_{k-1}$ are the covariance matrices of the errors of the a priori and a posteriori state estimates, respectively.

The Update Stage

$$\mathbf{K}_k = \mathbf{P}_k^- \mathbf{H}^T (\mathbf{H} \mathbf{P}_k^- \mathbf{H}^T + \mathbf{R})^{-1}, \quad (5)$$

where $\mathbf{K}_k$ stands for the Kalman Gain (matrix), and

$$\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_k^- + \mathbf{K}_k(\mathbf{y}_k - \mathbf{H}\hat{\mathbf{x}}_k^-), \quad (6)$$

$$\mathbf{P}_k = (\mathbf{I} - \mathbf{K}_k \mathbf{H}) \mathbf{P}_k^-. \quad (7)$$

The error of the estimation for one step ahead and its variance are given by

$$\mathbf{u}_k = \mathbf{y}_k - \mathbf{H}\hat{\mathbf{x}}_k^- \text{ and } \mathbf{F}_k = \mathbf{H}\mathbf{P}_k^- \mathbf{H}^T + \mathbf{R}_k.$$

In applications of the filter, the measurement noise covariance matrix $\mathbf{R}_k$ is usually constant and measured prior to the filter operation or it is known. In order to estimate $\mathbf{Q}_k$, the Maximum Likelihood Estimation (MLE) method can be used. The associated log-likelihood function for n measurements has the form [Tusell (2011), Durbin J. et al (2012)]

$$\text{Log}L(\mathbf{y}_1, \dots, \mathbf{y}_n) = -\frac{n}{2} \log 2\pi - \frac{1}{2} \sum_{k=1}^n (\log(|\mathbf{F}_k|) + \mathbf{u}_k^T \mathbf{F}_k^{-1} \mathbf{u}_k), \quad (8)$$

where n denotes the number of measurements and $|\mathbf{F}_k|$ the determinant of the matrix $\mathbf{F}_k$. The maximum likelihood estimators (MLEs) are most attractive because of their asymptotic properties. Under regularity conditions [Green (2002)], the maximum likelihood estimator has the following asymptotic properties: Consistency, Asymptotic normality, Asymptotic efficiency and Invariance.

3. TOBIT KALMAN FILTER

In this section we describe the Tobit Kalman filter [Allik (2014)], abbreviated as TKF, which provides a classification scheme for censored models [Allik (2014), Tobin J (1958)]; these classes depend on the type of censoring and include also the cases of censoring, that depends on other variables. In the applications cases of censoring, the censored measurement model provides a measurement, either in knowing the exact value (it belongs to the uncensored region), or in knowing that the value lies into an interval.

In the general case of scalar measurements, the Tobit model is referred to as the censored regression model determined by (9),

$$y_k^* = hx_k + v_k,$$

$$y_k = \begin{cases} y_k^*, & T_{min} < y_k^* < T_{max} \\ T_{min}, & y_k^* < T_{min} \\ T_{max}, & y_k^* > T_{max}, \end{cases} \quad (9)$$

where y_k^* is the latent (hidden) variable, y_k is the measurement, h is an arbitrary scalar, T_{min} , T_{max} are the lower and upper thresholds-limits respectively and v_k is a Gaussian random variable with mean 0 and variance σ_v^2 . By (9) it is obvious that the Tobit Kalman Filter determines a non-linear process.

In order to face the problem of censored measurements, we propose the TKF defined by Allik (2014)]

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{w}_k,$$

$$\mathbf{y}_k^* = \mathbf{H}\mathbf{x}_k + \mathbf{v}_k,$$

with

$$y_{k,i} = \begin{cases} y_{k,i}^*, & T_{min,i} < y_{k,i}^* < T_{max,i} \\ T_{min,i}, & y_{k,i}^* < T_{min,i} \\ T_{max,i}, & y_{k,i}^* > T_{max,i}. \end{cases} \quad i = 1, 2, \dots, m \quad (10)$$

where the noises $\mathbf{w}_k$ and $\mathbf{v}_k$ are defined by equations (1), (2) and $\mathbf{y}_k = (y_{k,i})_{i=1,\dots,m}$, $\mathbf{y}_k^*$ are defined as the saturated observation and latent variable respectively. Next we prove analytically for the case $\mathbf{H} = \text{diag}(h_1, \dots, h_m)$ and $\mathbf{R} = \text{diag}(r_1^2, \dots, r_m^2)$ the following useful Lemma 3.1.

Lemma 3.1. *The probability function of the i^{th} component of the measurement given the state vector is*

$$f(y_{k,i}|x_{k,i}) = \frac{1}{r_i} \phi\left(\frac{y_{k,i} - h_i x_{k,i}}{r_i}\right) u(y_{k,i} - T_{min,i}) u(T_{max,i} - y_{k,i})$$

$$+ \Phi\left(\frac{T_{min,i} - h_i x_{k,i}}{r_i}\right) \delta(T_{min,i} - y_{k,i})$$

$$+ \left(1 - \Phi \left(\frac{T_{max,i} - h_i x_{k,i}}{r_i} \right) \right) \delta(T_{max,i} - y_{k,i}),$$

where ϕ and Φ are the probability and cumulative distribution function of standard normal distribution respectively, δ stands for the Kronecker delta function and u for the Heavyside function.

Proof. When the i^{th} component of the latent variable belongs to the uncensored region, (T_{min}, T_{max}) , we get by (10) that

$$y_{k,i} = h_i x_{k,i} + v_{k,i}, \quad (11)$$

where $v_{k,i} \sim N(0, r_{k,i}^2)$. Thus by (11), the cumulative distribution for the measurement $Y_{k,i}$ is

$$\begin{aligned} F(y_{k,i}|x_{k,i}) &= P(Y_{k,i} \leq y_{k,i}), \\ &= P(h_i x_{k,i} + v_{k,i} \leq y_{k,i}) \\ &= P\left(\frac{v_{k,i}}{r_{k,i}} \leq \frac{y_{k,i} - h_i x_{k,i}}{r_{k,i}}\right) \\ &= \Phi\left(\frac{y_{k,i} - h_i x_{k,i}}{r_{k,i}}\right) \end{aligned}$$

and by differentiating we get

$$f(y_{k,i}|x_{k,i}) = \frac{1}{r_{k,i}} \phi\left(\frac{y_{k,i} - h_i x_{k,i}}{r_{k,i}}\right). \quad (12)$$

The probability of the i^{th} component of the latent variable to belong into the censored region from below is

$$\begin{aligned} P(y_{k,i} = T_{min,i}|x_{k,i}) &= P(y_{k,i}^* \leq T_{min,i}|x_{k,i}) \\ &= P(h_i x_{k,i} + v_{k,i} \leq T_{min,i}) \\ &= P\left(\frac{v_{k,i}}{r_{k,i}} \leq \frac{T_{min,i} - h_i x_{k,i}}{r_{k,i}}\right) \end{aligned}$$

and thus,

$$P(y_{k,i} = T_{min,i}|x_{k,i}) = \Phi\left(\frac{T_{min,i} - h_i x_{k,i}}{r_{k,i}}\right). \quad (13)$$

In the same way it is proved that

$$P(y_{k,i} = T_{max,i}|x_{k,i}) = 1 - \Phi\left(\frac{T_{max,i} - h_i x_{k,i}}{r_{k,i}}\right). \quad (14)$$

By (12)-(14) the probability distribution function for the measurements $\mathbf{y}_k$ given the state vector $\mathbf{x}_k$ is derived. $\square$

We denote by $\mathbf{P}_{un,k}$, $\mathbf{P}_{min,k}$, $\mathbf{P}_{max,k}$ the probabilities of a measurement to be uncensored, or censored from below or censored from above, respectively, at time k . Then by *Lemma 3.1* it is derived that

$$\mathbf{P}_{un,k} = \text{diag} \begin{bmatrix} \Phi\left(\frac{T_{max,1}-h_1\hat{x}_{k,1}^-}{r_{k,1}}\right) - \Phi\left(\frac{T_{min,1}-h_1\hat{x}_{k,1}^-}{r_{k,1}}\right) & \dots \\ \Phi\left(\frac{T_{max,m}-h_m\hat{x}_{k,m}^-}{r_{k,m}}\right) - \Phi\left(\frac{T_{min,m}-h_m\hat{x}_{k,m}^-}{r_{k,m}}\right) \end{bmatrix}, \quad (15)$$

$$\mathbf{P}_{min,k} = \text{diag} \begin{bmatrix} \Phi\left(\frac{T_{min,1}-h_1\hat{x}_{k,1}^-}{r_{k,1}}\right) & \dots \\ \Phi\left(\frac{T_{min,m}-h_m\hat{x}_{k,m}^-}{r_{k,m}}\right) \end{bmatrix}, \quad (16)$$

$$\mathbf{P}_{max,k} = \text{diag} \begin{bmatrix} 1 - \Phi\left(\frac{T_{max,1}-h_1\hat{x}_{k,1}^-}{r_{k,1}}\right) & \dots \\ 1 - \Phi\left(\frac{T_{max,m}-h_m\hat{x}_{k,m}^-}{r_{k,m}}\right) \end{bmatrix}. \quad (17)$$

By taking into account the matrices (15)-(17) and the properties of truncated normal distribution [Burkardt (2014), Tobin (1958)], we get that the expected value of the measurement when censored and uncensored measurements are included given the a priori estimation of the state vector has the form

$$\mathbf{E}(\mathbf{y}_k) = \mathbf{P}_{un,k}(\mathbf{H}\hat{\mathbf{x}}_k^- + \mathbf{R}^{\frac{1}{2}}l_k) + \mathbf{P}_{min,k}\mathbf{T}_{min} + \mathbf{P}_{max,k}\mathbf{T}_{max} \quad (18)$$

where $\mathbf{T}_{max} = (T_{max,i})_{i=1,\dots,m}$, $\mathbf{T}_{min} = (T_{min,i})_{i=1,\dots,m}$ and the parameter l_k at time k is the inverse Mill ratio [Burkardt (2014)],

$$l_k = \mathbf{P}_{un,k}^{-1} \begin{bmatrix} \phi\left(\frac{T_{max,1}-h_1\hat{x}_{k,1}^-}{r_{k,1}}\right) - \phi\left(\frac{T_{min,1}-h_1\hat{x}_{k,1}^-}{r_{k,1}}\right) & \dots \\ \phi\left(\frac{T_{max,m}-h_m\hat{x}_{k,m}^-}{r_{k,m}}\right) - \phi\left(\frac{T_{min,m}-h_m\hat{x}_{k,m}^-}{r_{k,m}}\right) \end{bmatrix}.$$

The covariance matrix of the measurement is given by

$$\mathbf{R}_k^* = \mathbf{R}_k \left(\mathbf{I} + \mathbf{P}_{un}^{-1} \text{diag}(c_k) - \text{diag}(l_k)^2 \right) \quad (19)$$

where the parameter c_k [Burkardt (2014)] is given by

$$c_k = \text{diag} \left[\begin{array}{c} \frac{T_{min,1}-h_1\hat{x}_{k,1}^-}{r_{k,1}}\phi\left(\frac{T_{min,1}-h_1\hat{x}_{k,1}^-}{r_{k,1}}\right) - \frac{T_{max,1}-h_1\hat{x}_{k,1}^-}{r_{k,1}}\phi\left(\frac{T_{max,1}-h_1\hat{x}_{k,1}^-}{r_{k,1}}\right) \\ \dots \\ \frac{T_{min,m}-h_m\hat{x}_{k,m}^-}{r_{k,m}}\phi\left(\frac{T_{min,m}-h_m\hat{x}_{k,m}^-}{r_{k,m}}\right) - \frac{T_{max,m}-h_m\hat{x}_{k,m}^-}{r_{k,m}}\phi\left(\frac{T_{max,m}-h_m\hat{x}_{k,m}^-}{r_{k,m}}\right) \end{array} \right].$$

The Tobit Kalman Filtering process is defined as [Allik (2014)] :

The Predict Stage:

$$\hat{\mathbf{x}}_k^- = \mathbf{A}\hat{\mathbf{x}}_{k-1}, \quad (20)$$

$$\mathbf{P}_k^- = \mathbf{A}\mathbf{P}_{k-1}\mathbf{A}^T + \mathbf{Q}_k. \quad (21)$$

The Update Stage:

$$\mathbf{R}_1 = \mathbf{P}_k^- \mathbf{H}^T \mathbf{P}_{un,k},$$

$$\mathbf{R}_2 = \mathbf{P}_{un,k} \mathbf{H} \mathbf{P}_k^- \mathbf{H}^T \mathbf{P}_{un,k} + \mathbf{R}_k^*,$$

$$\mathbf{K}_k = \mathbf{R}_1 \mathbf{R}_2^{-1}, \quad (22)$$

$$\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_k^- + \mathbf{K}_k (\mathbf{y}_k - \mathbf{E}(\mathbf{y}_k | \hat{\mathbf{x}}_k^-)), \quad (23)$$

$$\mathbf{P}_k = (\mathbf{I} - \mathbf{K}_k \mathbf{P}_{un,k} \mathbf{H}) \mathbf{P}_k^-. \quad (24)$$

The likelihood function for the i^{th} component of the censored measurement $\mathbf{y}_k$, is:

$$\begin{aligned} L_i(y_{1,i}, \dots, y_{n,i}) &= \prod_{y_{k,i}=T_{below}^i} \Phi(Tl_k^i) \prod_{y_{k,i}=T_{above}^i} (1 - \Phi(Th_k^i)) \\ &\times \prod_{T_{below}^i < y_{k,i} < T_{above}^i} \frac{1}{\sqrt{h_i^2 P_{k,i}^- + r_i^2}} \phi\left(\frac{y_{k,i} - h_i \hat{x}_{k,i}^-}{\sqrt{h_i^2 P_{k,i}^- + r_i^2}}\right), \end{aligned} \quad (25)$$

4. IMPLEMENTATIONS

In the present paper we use the Microsoft Kinect v2 sensor to record 3D point sequences of a human skeleton in motion and our aim is to denoise the coordinates for every joint in order to improve the representation of the movements. For this reason, in our first approach, we use a KF for each one of the joints separately; the input includes the joints' coordinates $[x, y, z]$ (measurement) and the outputs the denoised coordinates (state vectors).

Thus, we define the transition matrix $\mathbf{A}$ and the observation matrix of the model, $\mathbf{H}$ as

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Next, we have to estimate the covariance matrix for the process noise, $\mathbf{Q}_k$ at time step k . Firstly we assume that the covariance matrix of the measurement noise, $\mathbf{R}_k$, is constant and its entries are of the order 0.01 m^2 . We chose to initialize the matrix $\mathbf{R}_k$ in that way under the assumption that the Kinect sensor exhibits significant errors on each axes. To second our claim, we have conducted various experiments by Kinect v2 sensor, showing that even if an individual is at rest and in front of the Kinect, the RMSE in the displacement estimation between measurement and groundtruth data is almost 0.02m while in the case where the human skeleton is occluded the RMSE is bigger (0.03m - 0.20m), thus a variance of 0.01 m^2 seems a good choice.

In order to create a general model for denoising Kinect's measurements, (in which we will not estimate the matrix $\mathbf{Q}_k$ for every time-window, because this is time consuming) we can assume that $\mathbf{Q}_k$ is constant. Then, by the likelihood function (8), the entries of the matrix $\mathbf{Q}$ can be derived. Interestingly we noticed by various joints' movements, that the entries of $\mathbf{Q}$ appeared (were estimated) to be smaller than those of matrix $\mathbf{R}$, and generally they depend on the accuracy of the Kinect v2 sensor and the joints' speed. Concerning slow motions or the human skeleton at rest, the values are experimentally found to be smaller than 10^{-4} m^2 and for faster motions they lie between 10^{-3} m^2 and 10^{-2} m^2 . We have to notice that in some cases where the entries of $\mathbf{Q}$ were found to be quite large (10^{-2} m^2), the human skeleton moved too quickly in an abnormal manner due to occlusions and self-occlusions. Thus, we assume that the covariance matrix of the noise process is

$$\mathbf{Q} = 0.002 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

otherwise, if we assume smaller or bigger values, the Kalman filter will over-smooth or it will not denoise the Kinect's measurements, respectively. Thus, the above assumption seems to provide a good approximation in order to smooth the Kinect's v2 sensor measurements.

In our first experiment, a man throws a ball with his right hand, and this movement is recorded by Kinect; the video consists of 266 frames (almost 8.8667 sec.). Many joints and especially the joints on the right side were self-occluded. Therefore, we used the KF as described in Section 2 in order to denoise the data, i.e., to reveal the hidden coordinates due to the occlusion. It is obvious by *Fig.1* and

2 that the KF smooths the spatial coordinates without alienating the movement, i.e., it does not provide oversmoothing of the motion.

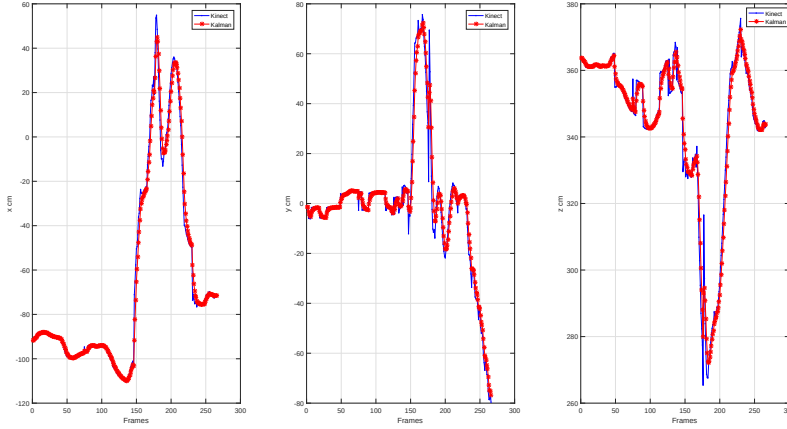


Figure 1: The spatial coordinates for the right hand as they result by Kinect and the Kalman filter.

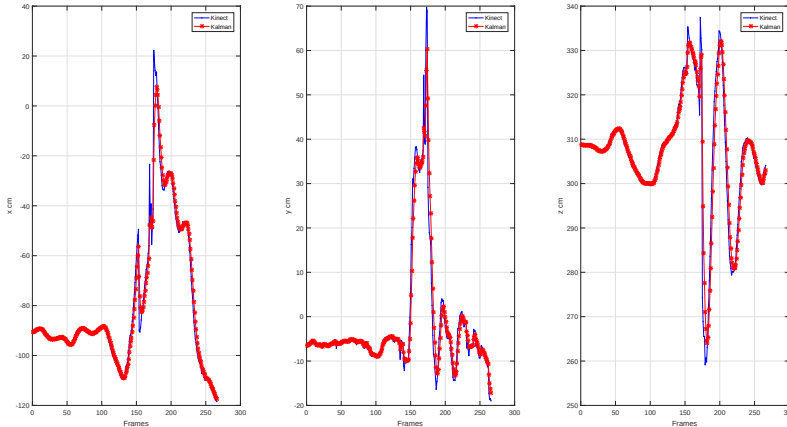


Figure 2: The spatial coordinates for the left hand as they result by Kinect and the Kalman filter.

In other experiments, an individual sits in front of the Kinect v2 sensor and the human skeleton appears to fall down 0.30-0.60 m (see Fig. 3) for only one frame (apparently this is due to some noise of Kinect). Then, we used Kalman filtering, but the noise could not be corrected satisfactorily. In order to correct the noise, we studied many recordings by the ground-truth sensor, Vicon; we observed that the velocity of spatial coordinates x and z did not exceed 31 cm per two consecutive frames for every joint, while the coordinate y did not exceed 18

cm respectively. Thus we took these restrictions into account, in order to correct the data. Therefore, we constructed a TKF with limits $\mathbf{T}_{min}$ and $\mathbf{T}_{max}$ for the spatial coordinates $[x, y, z]$ as follows,

$$\mathbf{T}_{max,k} = (\hat{x}_{k-1} + 0.31, \hat{y}_{k-1} + 0.18, \hat{z}_{k-1} + 0.31),$$

$$\mathbf{T}_{min,k} = (\hat{x}_{k-1} - 0.31, \hat{y}_{k-1} - 0.18, \hat{z}_{k-1} - 0.31),$$

where $\mathbf{T}_{max,k}$ and $\mathbf{T}_{min,k}$ are the limits of the TKF at time k which depend on the previous estimation of spatial coordinates. Then, by (10), for the measurement $\mathbf{y}_k = [x_k, y_k, z_k]$ at time k we get

$$y_{k,i} = \begin{cases} y_{k,i}^*, & T_{min,k}^i < y_{k,i}^* < T_{max,k}^i \\ T_{min,k}^i, & y_{k,i}^* < T_{min,k}^i \\ T_{max,k}^i, & y_{k,i}^* > T_{max,k}^i. \end{cases} \quad i = 1, 2, 3$$

The aforementioned TKF model can appropriately smooth big aberrant movements due to Kinect's errors. Apparently, if $T_{min,k}^i \rightarrow -\infty$ and $T_{max,k}^i \rightarrow \infty$ (i.e., the range of TKF's state values becomes too big) the TKF becomes the standard KF.

The covariance matrix for the noise measurement $\mathbf{R}$ is defined as in the KF, while the covariance matrix for the noise process $\mathbf{Q}$ can be estimated by the likelihood function (25). By the experiments on joints' movements that we mentioned in KF, we get almost the same results for the values of $\mathbf{Q}$ as in KF, so we can use the same matrix $\mathbf{Q}$ as in KF.

As can be seen in *Fig.3* the skeleton motion of the TKF (green) does not exhibit any unexplainable fall. Notice that the standard KF can correct the noise but not as good as the TKF. In our experiments, the TKF exhibits skeleton falls till almost 4 cm, on the other hand the standard KF exhibits skeleton falls of more than 4 and till 10 cm per (two) consecutive frames, which is not realistic. This conclusion is more clear in *Fig.4*, where the head's spatial coordinate y_k for each frame k is illustrated.

5. CONCLUSION

The aim of this paper is to improve skeleton tracking, using a single Kinect v2 sensor, which generates error in recordings due to occlusion, self-occlusion e.t.c.. Thus, we propose to use a Tobit Kalman filter for skeleton tracking in real time. In this approach we have to define the limits $\mathbf{T}_{max,k}$ and $\mathbf{T}_{min,k}$ in a reasonable manner for every time k . For that purpose we considered a lot of skeleton data, with various joints' movements, which were obtained by means of the groundtruth sensor, Vicon.

The covariance matrix of the noise process $\mathbf{Q}$, using Kalman filtering procedure was estimated via maximum likelihood estimation. Between the two filters, i.e.,

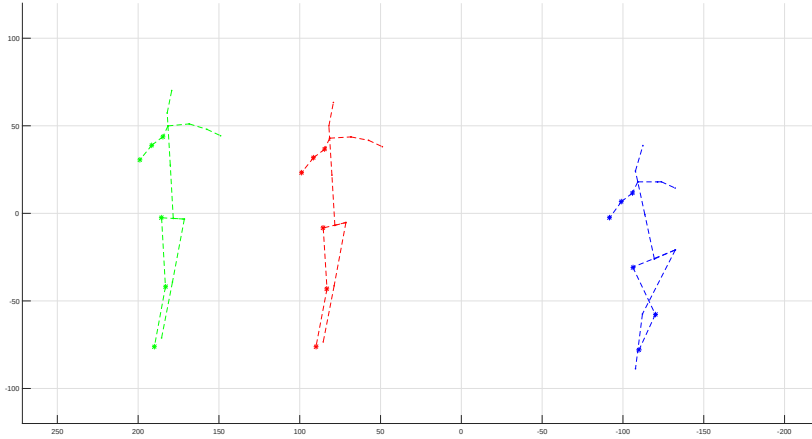


Figure 3: From the left to the right: TKF's, KF's and Kinect's skeleton respectively at the frame of the fall.

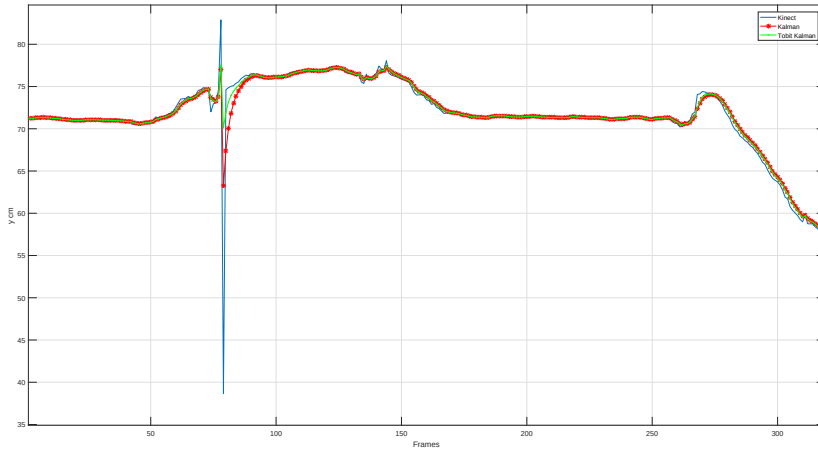


Figure 4: The head's spatial coordinate y_k for each frame k of Tobit Kalman filter, Kalman filter and Kinect v2 sensor.

the standard Kalman filter and the Tobit Kalman filter, the later was more accurate performing a better skeleton tracking. Furthermore, in some frames when the skeleton collapsed due to occlusion, the method of the Tobit Kalman filter proposed, corrected better the error in recordings than the standard Kalman filter.

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία, αναλύουμε δεδομένα από την κάμερα Microsoft Kinect 2 χρησιμοποιώντας φίλτρα Kalman και Tobit Kalman για την ελαχιστοποίηση του θορύβου που εμφανίζεται στα δεδομένα. Τα δεδομένα αφορούν τριδιάστατες χωρικές συντεταγμένες που καταγράφουν κινήσεις των αρθρώσεων ενός ανθρώπου, στις οποίες εμφανίζονται σφάλματα στην ακρίβεια των μετρήσεων. Χρησιμοποιούμε πέρα από το κλασικό φίλτρο Kalman και το Φίλτρο Tobit Kalman, προκειμένου να συμπεριλάβουμε στο μοντέλο περιοριστικές συνθήκες με βάση τα ανθρωπομετρικά στοιχεία, και συγκεκριμένα τις αποστάσεις μεταξύ διαφόρων αρθρώσεων. Στο τέλος παρουσιάζουμε προσομοιώσεις για την κίνηση του σκελετού πριν και μετά τη χρήση των φίλτρων.

Acknowledgements: This work was supported by the European Project (Horizon2020) ICT4Life under the GA 690090.

REFERENCES

- Allik B. (2014). *Tobit Filter Kalman*, University of Delaware: Department of Electrical and Computer Engineering.
- Berti E. M., Salmeron A. J. S., and Benimeli F. (2012). Kalman Filter for Tracking Robotic Arms Using low cost 3D Vision Systems. *The Fifth International Conference on Advances in Computer-Human Interactions*. 236-240.
- Burkardt J. (2014). *The Truncated Normal Distribution*, Florida State University: Department of Scientific Computing.
- Du Y., Wang W. and Wang L., (2015). Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition. *Conference on Computer Vision and Pattern Recognition* 1110-1118.
- Durbin J. and Koopman S. J. (2012). Time series analysis by state space methods. *Oxford University Press* No 38.
- Galna B., Barry, G., Jackson, D., et. al. (2014). Accuracy of the Microsoft Kinect sensor for measuring movement in people with Parkinson's disease. *Gait and Posture* Vol 39. Issue 4. 1062-1068.
- Green W. H. (2002). *Econometric Analysis*, 5th edition, Chapter 17, New York: Prentice Hall.
- Kong W. A., Hussain H. and Md Saad M. H. (2013). Essential Human Body Points Tracking Using Kalman Filter. *Proceedings of the World Congress on Engineering and Computer Science* Vol I.
- Li Z., Wei Z., Yue Y., et al. (2015). An Adaptive Hidden Markov Model for Activity Recognition Based on a Wearable Multi-Sensor Device. *Journal of Medical Systems-Springer*, 39-57.
- Peter S. (1979.) Stochastic models, estimation, and control. *Mathematics in Science and Engineering* Vol 141.
- Sungphill M., Youngbin P., Dong W. K., et. al. (2016). Multiple Kinect Sensor Fusion for Human Skeleton Tracking Using Kalman Filtering. *International Journal of Advanced Robotic Systems*.

- Tobin J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: Journal of the Econometric Society* 24-36.
- Tusell F. (2011). Kalman Filtering in R. *Journal of Statistical Software* Vol 39. Issue 2.
- Wentao Z., Cuiling L., Junlian, X., et. al. (2016). Co-occurrence Feature Learning for Skeleton based Action Recognition using Regularized Deep LSTM Networks. *The 30th AAAI Conference on Artificial Intelligence*.
- Zhang C. and Zhang L. (2013). Activity Recognition in Smart Homes Based on Second-Order Hidden Markov Model. *International Journal of Smart Home* Vol 7, No 6, 237-244.



USING A WEIBULL-TYPE CONDITIONAL INTENSITY FUNCTION FOR THE LINKED STRESS RELEASE MODEL- APPLICATION FOR THE CORINTH GULF

O. Mangira¹, G. Vasiliadis² and E. Papadimitriou¹

¹Department of Geology, Aristotle University of Thessaloniki
{omangira, ritsa}@geo.auth.gr

²Department of Informatics Engineering, TEI of West Macedonia
gvasil@math.auth.gr

ABSTRACT

Spatio-temporal stress changes on adjacent fault segments consist the most important components in earthquake generation, and consequently in seismic hazard assessment, as they can alter the occurrence probability onto them. The investigation of the interactions between adjacent areas by means of the Linked Stress Release Model (LSRM) is attempted for strong earthquakes ($M \geq 5.0$) in the Corinth Gulf, which is divided into its western and eastern parts, based on seismotectonic criteria. The behavior of this point process is determined by the conditional intensity function. In the analysis of the «classical» Stress Release Model (SRM), the conditional intensity function, has the exponential form. Alternatively, a conditional intensity function of Weibull form is used for comparing the results and identifying the most appropriate model for the interpretation of the earthquake generation process function. According to the estimated interaction coefficients of the LSRM we conclude that there is no evidence using this model that an earthquake occurrence in a subarea of the Corinth Gulf provokes damping or excitation to the neighboring subarea.

Key words: stress transfer, conditional intensity function, Weibull

1. INTRODUCTION

Isham & Westcott (1979) were the first to examine self-correcting point processes as an example of a process that automatically corrects a deviation from its mean. The Stress Release Model (SRM) introduced by Vere-Jones (1978), is a stochastic version of the elastic rebound theory [Reid (1910)]. This deterministic model indicates that elastic stress is accumulated due to the long-term tectonic loading, and is released

during earthquake when the stress exceeds a certain threshold, i.e., the strength of the medium. According to this theory, the occurrence of an earthquake and consequently the energy release should allow a certain time period to elapse until the re-accumulation of the energy and the genesis of a subsequent earthquake. Besides, in a self-correcting process the points have an inhibitory effect.

It has been noticed very often that shortly after a strong earthquake, a second one follows. Kagan and Jackson (1991) suggest a long-term weak clustering which characterizes all the mainshocks. Gabrielov and Newman (1994) also support that a period of activation rather than a period of quiescence is observed and sometimes other earthquakes of comparable magnitude follow the main event. This kind of behavior could be explained through stress transfer between adjacent areas. These ideas are incorporated in the Linked Stress Release Model (LSRM), introduced by Liu et al. (1999). In that way, interactions between different parts of an area are investigated in order to estimate the effect on the seismic hazard in each part. The interactions could provoke either damping or excitation of an adjacent subarea.

Aiming to investigate the strong earthquake occurrence in the area of Corinth Gulf, an area that very often accommodates disastrous earthquakes with $M \geq 6.0$, the LSRM is applied. The earthquake interaction is examined between two subareas in which the study area was subdivided, the western and eastern part. The application is performed by means of the conditional intensity function. Specifically, we examine the LSRM where the conditional intensity function is of the Weibull type.

2. FORMULATION OF THE MODELS

2.1 Simple Stress Release Model (SSRM)

In the SSRM, the key variable is the stress level in a region, which determines the probability of an earthquake occurrence [Vere-Jones and Deng (1988)]. The evolution of stress $X(t)$ as a function of time, could be written as

$$X(t) = X(0) + \rho t - S(t),$$

where $X(0)$ is the initial stress value, ρ is the loading rate and $S(t)$ is the accumulated stress release from earthquakes within the area over the period $(0, t)$, i.e.,

$$S(t) = \sum_{i: t_i < t} S_i$$

where t_i , S_i are the time and the stress release, respectively, associated with the i -th earthquake.

The amount of stress released by an earthquake can be estimated by empirical relations as a function of its magnitude. Bufe and Varnes (1993) suggest the cumulative Benioff strain as a measure of the total energy released, i.e.,

$$S(t) = \sum_{i=1}^{N(t)} E_i^{1/2}, \text{ where } E_i \text{ is the seismic energy of the } i\text{-th earthquake and } N(t) \text{ stands}$$

for the number of earthquakes till time t . Kanamori and Anderson (1975) show that the seismic energy released is given by the relation $E = 10^{3/2M} + \text{const}$. Then we have that

$$S = 10^{0.75(M - M_{\min})}, \quad (1)$$

where M denotes the earthquake magnitude and $M_{\min}$ the minimum magnitude appeared in the dataset.

The stochastic behavior of the model is characterized by the conditional intensity function $\lambda^*(t)$ [Daley and Vere-Jones (2003)], which is assumed to have the exponential form

$$\begin{aligned} \lambda^*(t) &= \Psi(X(t) = \exp(\mu + \nu(X(0) + \rho t - S(t))) \\ &= \exp[a + \nu(\rho t - S(t))] \\ &= \exp\{a + b[t - cS(t)]\}, \end{aligned}$$

where a, b, c are the parameters to be estimated.

The parameters are estimated numerically by maximizing the log-likelihood function

$$\log L = \sum_{i=1}^{N(t)} \log \lambda^*(t) - \int_0^T \lambda^*(u) du,$$

where $(0, T)$ is the observational interval and $N(T)$ is the total number of events in $(0, T)$.

2.2 Linked Stress Release Model

From the above description, it can be seen that stress transfer is not taken into account in the SSRM. In order to incorporate interactions between different subareas, the LSRM is proposed. In that case, the evolution of stress in the i -th area can be rewritten as

$$X_i(t) = X_i(0) + \rho_i t - \sum_j \theta_{ij} S(t, j),$$

where $S(t, j)$ stands for the accumulated stress release in the j subarea over the period $(0, t)$ and the coefficient θ_{ij} measures the fixed proportion of stress release, which is transferred from the j subarea to the i subarea. It is plausible to set $\theta_{ii}=1$, since we assume that at least a significant amount of the accumulated energy is released. Positive and negative values of θ_{ij} indicate damping and excitation, respectively.

In the “classical” SRM it is assumed that each area has a conditional intensity function of the form

$$\lambda_i^*(t) = \Psi(X_i(t)) = \exp \left\{ \alpha_i + \nu_i \left[\rho_i t - \sum_j \theta_{ij} S(t, j) \right] \right\},$$

where i denotes the specific subarea and $\alpha_i = \mu_i + \nu_i X_i(0)$, ν_i, ρ_i and θ_{ij} are the parameters to be fitted. A simpler parameterization [Liu et al. (1998)] is achieved by setting $b_i = \nu_i \rho_i$ and $c_{ij} = \theta_{ij} / \rho_i$, yielding

$$\lambda_i^*(t) = \exp \left\{ a_i + b_i \left[t - \sum_j c_{ij} S(t, j) \right] \right\},$$

where a_i , b_i , c_{ij} are the parameters to be fitted for each subarea i .

For the parameters to be estimated some restrictions exist that should be taken into account. The parameters $b_i = v_i \rho_i$, και $c_{ii} = 1/\rho_i$ should be positive, since the loading rate ρ_i and the sensitivity to stress change, v_i , take (only) positive values. On the contrary, c_{12} and c_{21} could take either positive or negative values.

Instead of using an exponential conditional intensity function we propose in the present paper a -more flexible- Weibull-type conditional intensity function. For each subarea i we assume that

$$\lambda_i^*(t) = \lambda_i \gamma_i X_i(t)^{\gamma_i-1} = \lambda_i \gamma_i \left(X_i(0) + \rho_i t - \sum_j \theta_{ij} S(t, j) \right)^{\gamma_i-1}.$$

The parameters to be estimated are λ_i , γ_i , $X_i(0)$, ρ_i and θ_{ij} , where $\lambda_i > 0$ and $\gamma_i > 0$. The conditional intensity function behaves in a different way depending on the value of the shape parameter γ . If $\gamma < 1$ the failure rate decreases over time, whereas if $\gamma > 1$ there is an increasing hazard over time. The value of $\gamma > 1$ is consistent with the geophysical point of view relative to the accumulation and release of the energy. We impose thus the restriction of $\gamma > 1$. A γ smaller than 1 would mean that the probability of a new earthquake to occur would decrease monotonically as a function of the time elapsed since the last one; this fact cannot be accepted from the physical point of view.

3. STUDY AREA AND DATA

The study area (Figure 1a) was extensively investigated during the last decades since the Corinth rift consists one of the most prominent active structures and has experienced several destructive earthquakes. The high level of seismicity is testified by the historical as well as the instrumental records, which include several moderate to large earthquakes ($M \geq 5.0$), with the maximum magnitude observed or ever reported hardly greater than 6.8, a fact that probably reflects the lack of continuity of fault segments [Jackson and White (1989)]. One more feature that should be taken into account in seismic hazard studies is that earthquakes may occur very close in time (e.g. the sequence of three ($M \geq 6.4$) 1981 earthquakes in the eastern part of the Corinth Gulf).

Due to the large number of parameters of the models used, as it was shown in the methodology section (section 2), a large data sample is required. We need though to compromise the fact that we are interested in the strong earthquakes in the study area.

For this reason the magnitude threshold was set as low as $M_{th}=5.0$ since 1911, in order to obtain the largest and longest possible complete earthquake catalog; the completeness magnitude, M_{th} , is defined as the lowest magnitude at which 100% of the earthquakes in a certain space-time volume are detected [Rydelek and Sacks (1989)]. The data used for the current study are taken from the catalog compiled from the Geophysics Department of Aristotle University of Thessaloniki (<http://geophysics.geo.auth.gr/ss/>). For the LSRM, the area was divided into two subareas according to seismotectonic criteria.

Figure 1a. Seismicity map for the Corinth Gulf showing earthquakes with magnitude $M \geq 5.0$ that occurred in the area since 1911. The events occurred in the western part are represented with red stars, whereas the ones occurred in the eastern part are represented with yellow hexagons. Their size is proportional to the earthquake's magnitude. **b.** Temporal distribution of the earthquakes with $M \geq 5.0$ that occurred in the Corinth Gulf since 1911.

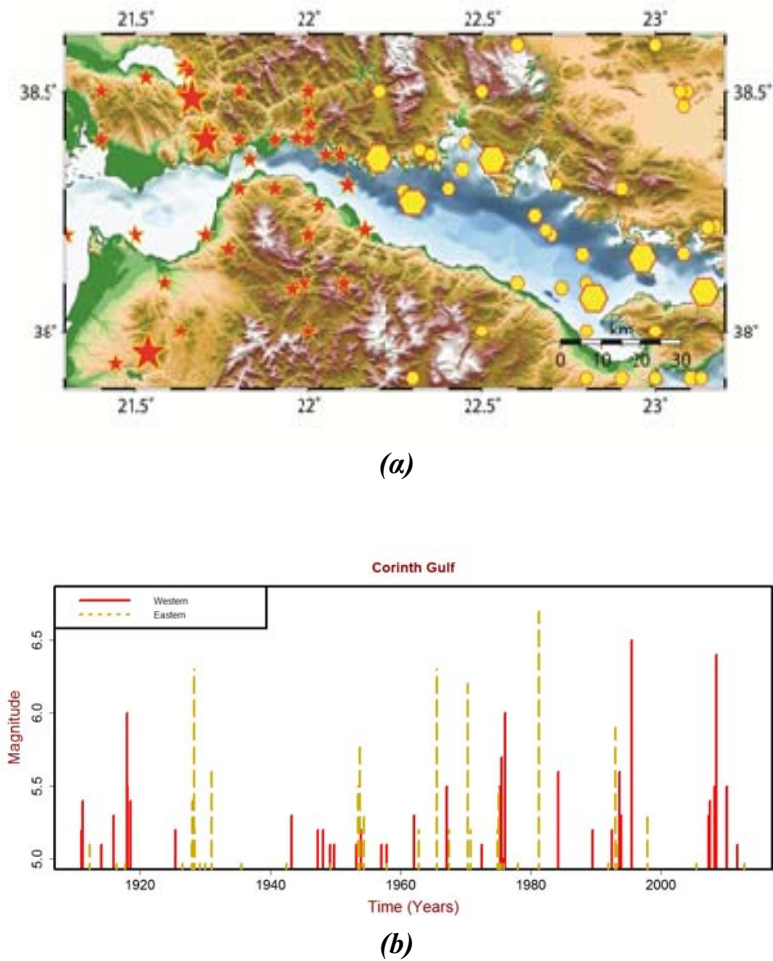
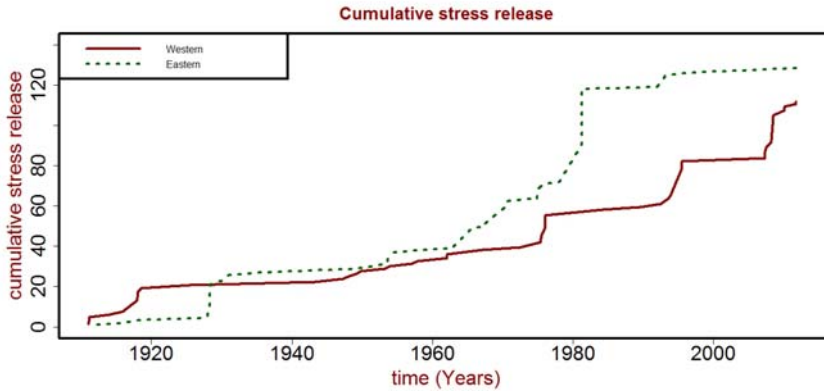


Figure 1b shows the temporal distribution of the dataset used, with the same color discrimination as in Figure 1a depending on the part of the Corinth Gulf where each event occurred. The magnitudes of the events for each subarea are transformed into cumulative stress release (Figure 2) according to the relation (1).

Figure 2. Cumulative stress release due to earthquakes occurred in each subarea since 1911.



4. APPLICATION OF THE LSRM USING A WEIBULL-TYPE CONDITIONAL INTENSITY FUNCTION

4.1 Parameter estimation

When estimating the parameters, in order to restrict their space and get results reflecting the real physical process, we expect for the loading rates the values of ρ_i to be close to the mean stress release over the time period (0, 102), i.e., $\rho_1 \sim 0.92$ and $\rho_2 \sim 1.44$.

Under this restriction we applied the LSRM, using a Weibull-type conditional intensity function. In order to find the maximum value of the likelihood function of the model, the 10-dimensional parameter space was scanned numerically ($216,8 \times 10^6$ points) using a dense grid (with step varying from 10^{-5} to 1), by also taking into account the restrictions of the model. The maximum value of the log-likelihood function of the model is found to be -178.024 (Table 1). The standard errors were calculated using the Fisher information matrix. The confidence intervals were constructed assuming (asymptotic) normality.

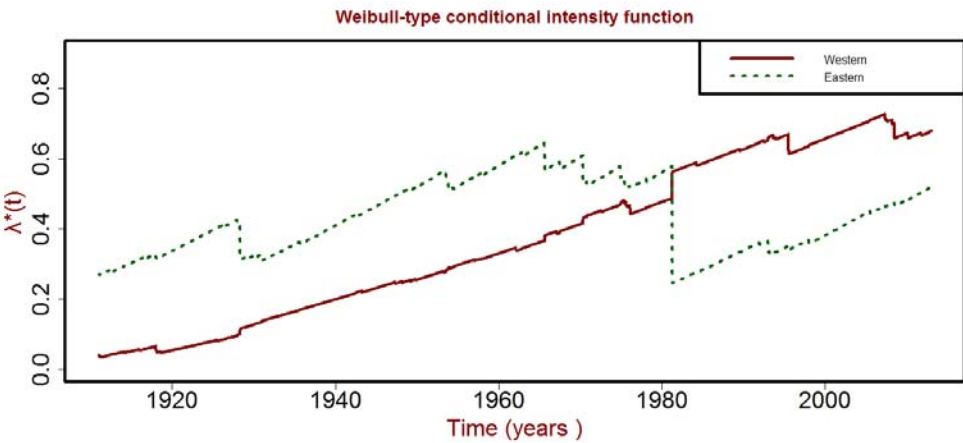
Table 1. *LSRM's parameters estimated, standard errors and 90% confidence intervals when using a Weibull-type conditional intensity function*

Parameter	Estimation	Standard errors	90% Confidence Intervals
λ_1	0.00019	$6.03 \cdot 10^{-5}$	$(9.07 \cdot 10^{-5}, 0.0003)$
λ_2	0.00015	$9.11 \cdot 10^{-5}$	$(1.76 \cdot 10^{-5}, 0.0003)$
γ_1	2.32	0.1897	(2.0080, 2.6917)
γ_2	2.56	0.1486	(2.3156, 2.8512)
$X_1(0)$	32	11.8562	(12.4965, 55.2381)
$X_2(0)$	67	26.1305	(24.0153, 118.2159)
ρ_1	2.7	1.6268	(0.0240, 5.8884)
ρ_2	1.7	0.8968	(0.2248, 3.4576)
θ_{12}	-0.5	2.3609	(-4.3836, 3.3837)
θ_{21}	0.1	1.1272	(-1.7542, 1.9542)

The model determined by the parameters given in Table 1 is consistent with our knowledge regarding the study period and area. We notice that the model overestimates the loading rates, particularly in the subarea 1, i.e., the western part of the Corinth Gulf, in comparison with the assumptions we have made. The confidence intervals of the parameters θ_{ij} concerning the interactions between the two subareas reveal that there is not enough evidence to claim that the eastern part causes excitation or damping to the western part and vice versa (since 0 belongs to these confidence intervals).

Figure 3 that follows shows the conditional intensity function for the parameters presented in Table 1.

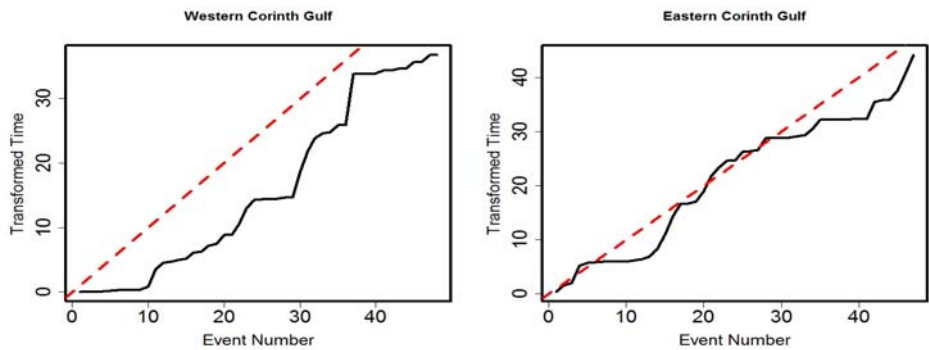
Figure 3. Conditional intensity functions for the two subareas of the Corinth Gulf using the Weibull distribution.



4.2 Model fitting

In order to evaluate the model we can perform residual analysis. Ogata (1988) provides a paradigm for assessment of the utility of a stochastic model. This technique uses the time transformation in testing the goodness-of-fit of a point process model. If the compensator used for the transformation is that of the true model, then the transformed process will be unit-rate Poisson, whereas systematic deviation of the data from the fitted model would mean that an important feature is not yet included in the model.

Figure 4. Residual analysis of the model fitted for the two parts of the Corinth Gulf.



The residual analysis for the model fitted is performed in Figure 4. As shown clearly, the fitting is poor in the western part of the Corinth Gulf. The transformed process is under the line of the unit-rate Poisson, which means that the model

underestimates the true process since the real number of events does not coincide with the number of events that emerges from the model.

5. CONCLUSIONS

The Linked Stress Release Model, incorporating a slow buildup of stress within a seismic area as well as its stochastic release through earthquakes and transfer between seismic areas, is applied to fit data from the Corinth Gulf. The conditional intensity function chosen, to determine the behavior of the point process, has a Weibull type form, as a more flexible alternative than the exponential type form. The results by adopting this model indicate that there is not enough evidence to claim that the interactions between the two subareas are robust, since the confidence intervals of the parameters concerning the interactions have both positive and negative values. Interesting results though can be extracted as it regards the behavior of the curves. The hazard in the western part of the Corinth Gulf exhibits a rising trend while in the eastern part the curve manifests more fluctuations.

ΠΕΡΙΛΗΨΗ

Οι χωρο-χρονικές μεταβολές των τάσεων σε γειτονικές ομάδες ρηγμάτων αποτελούν μία από τις πιο σημαντικές συνιστώσες στη διαδικασία της γένεσης των σεισμών, και κατ' επέκταση στην εκτίμηση της σεισμικής επικινδυνότητας, καθώς μπορούν να μεταβάλλουν τις πιθανότητες γένεσης σεισμών σε αυτές. Η διερεύνηση των αλληλεπιδράσεων μεταξύ γειτονικών περιοχών με τη βοήθεια του Συζευγμένου Μοντέλου Απελευθέρωσης Τάσης (ΣΜΑΤ) επιχειρείται για ισχυρούς σεισμούς ($M \geq 5.0$) στον Κορινθιακό Κόλπο, ο οποίος διακρίνεται στο δυτικό και το ανατολικό κομμάτι με βάση σεισμοτεκτονικά κριτήρια. Η συμπεριφορά της σημειακής διαδικασίας καθορίζεται από την υπό συνθήκη συνάρτηση του θετικού ρυθμού των αφίξεων. Στην ανάλυση του «κλασικού» Μοντέλου Απελευθέρωσης Τάσης η υπό συνθήκη συνάρτηση του θετικού ρυθμού των αφίξεων έχει εκθετική μορφή. Εναλλακτικά, προτείνεται στην παρούσα εργασία μια τύπου Weibull υπό συνθήκη συνάρτηση του θετικού ρυθμού των αφίξεων, στην προσπάθεια εύρεσης του καταλληλότερου μοντέλου που θα ερμηνεύει καλύτερα τη διαδικασία της σεισμογένεσης. Με βάση την υιοθέτηση του ανωτέρω μοντέλου συνάγεται ότι δεν υπάρχει αρκετή μαρτυρία για να υποστηριχθεί ότι η γένεση ενός σεισμού στο ένα τμήμα του Κορινθιακού Κόλπου προκαλεί σημαντική διέγερση ή αποδιέγερση στο άλλο.

Acknowledgment: The authors would like to express their sincere appreciation to Prof. Tsaklidis for his valuable comments and fruitful discussions

REFERENCES

- Bebbington, M., Harte, D. (2001). On the Statistics of the Linked Stress Release Process, *J. Appl. Prob.*, 38, 176-187.
- Bebbington, M. and Harte, D. (2003). The Linked Stress Release Model for Spatio-Temporal Seismicity: Formulations, Procedures and Applications. *Geophys. J. Int.*, 154, 925-946.
- Bebbington, M. (2005). Information gains for stress release models, *Pure Appl. Geophys.*, 162, 2229-2319.
- Console, R., Falcone, G., Karakostas, V., Murru, M., Papadimitriou E., Rhoades, D. (2013). Renewal models and coseismic stress transfer in the Corinth Gulf, Greece, fault system, *J. Geophys. Res.*, 118, 3655–3673.
- Console, R., Carluccio, R., Papadimitriou, E., Karakostas, V. (2015). Synthetic earthquake catalogs simulating seismic activity in the Corinth Gulf, Greece, fault system, *J. Geophys. Res.*, 120, 326-343.
- Daley, D., Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes*, Vol. 1, 2nd edition, Springer, New York, pp 211-287.
- Harte, D.S. and Vere-Jones, D. (2004). The Entropy Score and its Uses in Earthquake Forecasting, *Pure Appl. Geophys.*, 162, 1229–1253.
- Jiang, M., Zhu, S., Chen, Y., Ai, Y. (2011). A new multidimensional stress release statistical model based on coseismic stress transfer, *Geophys. J. Int.*, 187, 1479–1494
- Lu, C., Vere-Jones, D. (2000). Application of linked stress release model to historical earthquake data: comparison between two kinds of tectonic seismicity. *Pure Appl. Geophys.*, 157, 2351–2364.
- Lu, C., 2005. The degree of predictability of earthquakes in several regions of China: statistical analysis of historical data. *J. As. Earth Sciences*, 25, 379-385.
- Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *J. Amer. Statist. Assoc.* 83, 9-27.
- Rotondi, R., Varini, E. (2006). Bayesian analysis of marked stress release models for time-dependent hazard assessment in the western Gulf of Corinth, *Tectonophysics*, 107-113.
- Rydelek P.A. and Sacks I.S. (1989). Testing the completeness of earthquake catalogs and the hypothesis of self-similarity. *Nature*, 337, 251–253.
- Votsi, I., Tsaklidis, G., Papadimitriou, E. (2011). Seismic Hazard Assessment in Central Ionian Islands Area Based on Stress Release Models, *Acta Geophys.* , 59, 701-727.
- Zhuang, J., Harte, D., Werner, M.J., Hainzl S., Zhou S. (2012). Basic models of seismicity: temporal models, Community Online Resource for Statistical Seismicity.
- Zhuang, J., Touati, S. (2015). Stochastic simulation of earthquake catalogs, Community Online Resource for Statistical Seismology.



Bayesian Inference and Model Selection for Association Models in Contingency Tables

A. Mantzouni¹, I. Ntzoufras¹, M. Kateri²

¹Department of Statistics, Athens University of Economics and Business, Greece
{aikmantz, ntzoufras}@aueb.gr

²Department of Mathematics, RWTH Aachen University, Germany
maria.kateri@rwth-aachen.de

ABSTRACT

In this work we present Bayesian hypothesis tests for the independence between two categorical variables in contingency tables. Initially we implement conjugate analysis based on the Multinomial-Dirichlet setup. We compute the Bayes factor and assess the sensitivity of the results to the prior distribution. Then we focus on log-linear models. We compare the saturated and the model of independence with the utilization of Bayes factor compared with several model selection criteria. Finally, we focus on Bayesian inference in association models, a class of models describing the structure of association between two categorical variables. MCMC procedures are facilitated with the help of WinBUGS in order to estimate the posterior distributions of model parameters. Laplace approximation methods are applied so as to approximately compute the marginal likelihood. We illustrate all methods using real datasets.

Keywords: Ordinal classification variables, Contingency tables, Bayes Factor, Independence, Association Models, MCMC, Laplace Approximation.

1. Introduction

A contingency table is essentially a way to display cross-classification of two or more categorical variables. The term contingency table was first used by Karl Pearson (1904). Contingency tables are simple frequency tabulations, which present the frequencies for each combination of the levels (or categories) of all variables under consideration. In this work we deal with two-way tables.

In two-way contingency tables, we focus on the underlying association between the two classification variables and, therefore, on testing for their independence. From the Bayesian perspective, there are several ways to test for the independence depending on the model (likelihood and prior) you assume. Under conjugacy, we may use the Multinomial-Dirichlet or the Poisson-Gamma models. In such cases, the Bayes factor between the independence and the full (saturated) model is used

to evaluate independence between the two factors under examination.

A popular alternative can be obtained by using the Poisson log-linear models, where the Bayes factor is not analytically available and can be calculated only by using Markov chain Monte Carlo (MCMC) methods. In log-linear models there are just two options for modeling the dependence structure in two-way contingency tables. The parsimonious but restrictive and hardly fulfilled in practice model of independence and the saturated one. Association models (Goodman, 1979) fill the gap between these two extreme cases by imposing a special structure on the association of the two factors. By this way, the number of parameters involved in the interaction terms (parameters capturing dependence) is considerably reduced leading to models of intermediate complexity and high interpretational value. Thus, they are primarily applicable in case of ordinal classification variables. Within this framework, parametric scores are assigned to the rows and/or the columns of the contingency table.

2. Bayesian Inference for $I \times J$ Contingency Tables

In Section 2. we present the Bayesian independence tests based on the conjugate Multinomial-Dirichlet models (Section 2.1) and the Poisson log-linear models (Section 2.2). We also describe the well-known sensitivity of the Bayes factor on the prior parameters (Section 2.3), other Bayesian measures of model comparison (Section 2.4) and we conclude with illustrative examples, where we implement and compare these approaches. In Section 3. we proceed to Bayesian inference for association models through the Laplace approximation (Section 3.1). In last Section we summarize our results and discuss future research directions.

Let $\mathbf{n}$ be a $I \times J$ contingency table of observed cell counts obtained by the cross-classification of two categorical variables A and B . The elements of $\mathbf{n}$ are denoted by n_{ij} , $\mathbf{n} = (n_{ij})$ and correspond to the joint frequency of the i^{th} row (level of factor A) and j^{th} column (level of factor B) of the table, $i = 1, 2, \dots, I$, $j = 1, 2, \dots, J$. The row (column) marginals and the grand total of the contingency table are denoted by $n_{i+} = \sum_{j=1}^J n_{ij}$, $i = 1, 2, \dots, I$ ($n_{+j} = \sum_{i=1}^I n_{ij}$, $j = 1, 2, \dots, J$) and $N = \sum_{i=1}^I \sum_{j=1}^J n_{ij}$, respectively.

Bayesian independence tests $I \times J$ contingency tables are based on posterior model odds and Bayes factor. Bayesian comparison of two competing models, in our case the model of independence M_0 and the saturated model M_1 , is performed via the posterior model probabilities and their corresponding ratio

$$PO_{01} = \frac{f(M_0 | \mathbf{n})}{f(M_1 | \mathbf{n})} = \frac{f(\mathbf{n} | M_0)}{f(\mathbf{n} | M_1)} \times \frac{f(M_0)}{f(M_1)} = B_{01} \times \frac{f(M_0)}{f(M_1)},$$

which is the posterior model odds of model M_0 versus model M_1 ; $f(M)$ and $f(M | \mathbf{n})$ are the prior and the posterior probabilities of a model M and $f(\mathbf{n} | M)$ is the (Bayesian) marginal likelihood or prior predictive distribution of model M .

The latter is given by

$$f(\mathbf{n} \mid M) = \int f(\mathbf{n} \mid \boldsymbol{\vartheta}_M, M) f(\boldsymbol{\vartheta}_M \mid M) d\boldsymbol{\vartheta}_M;$$

where $\boldsymbol{\vartheta}_M$ is the parameter vector of model M , $f(\boldsymbol{\vartheta}_M \mid M)$ is its prior and $f(\mathbf{n} \mid \boldsymbol{\vartheta}_M, M)$ is the likelihood of the model under consideration. The Bayes factor of model M_0 versus model M_1 is defined as the ratio of the marginal likelihoods of M_0 versus M_1 (Jeffreys, 1961):

$$B_{01} = \frac{f(\mathbf{n} \mid M_0)}{f(\mathbf{n} \mid M_1)}. \quad (1)$$

The Bayes factor quantifies the evidence for and against the hypothesis of independence in contingency tables when $f(M_0) = f(M_1)$.

2.1 Conjugate Multinomial-Dirichlet Setup

Under this setup and under the model of dependence M_1 (saturated), we assume a Multinomial distribution for $\mathbf{n}$, $\mathbf{n} \sim Mult(N, \boldsymbol{\pi})$. Hence, the likelihood for model M_1 is given by

$$f(\mathbf{n} \mid \boldsymbol{\pi}, M_1) = \frac{n!}{\prod_{i=1}^I \prod_{j=1}^J n_{ij}} \prod_{i=1}^I \prod_{j=1}^J \pi_{ij}^{n_{ij}} \quad \text{with} \quad \sum_{i=1}^I \sum_{j=1}^J \pi_{ij} = 1. \quad (2)$$

The parameter of interest is the probability table $\boldsymbol{\pi}$, with entries π_{ij} , $\pi_{ij} = P(A = i, B = j)$ for the i^{th} row and j^{th} column of the table. Hence, the parameter matrix is denoted by $\boldsymbol{\pi} = (\pi_{ij}; i = 1, \dots, I, j = 1, \dots, J) = \boldsymbol{\vartheta}_1$.

The conjugate prior for the multinomial parameter $\boldsymbol{\pi}$ is the Dirichlet distribution, a multivariate generalization of the Beta distribution. Therefore, we a-priori assume that

$$\boldsymbol{\pi} \sim Dirichlet(\mathbf{k}) \quad \text{with} \quad \mathbf{k} = (k_{ij} > 0; i = 1, \dots, I, j = 1, \dots, J)$$

with density function given by

$$f(\boldsymbol{\pi}) = \frac{\Gamma(K)}{\prod_{i=1}^I \prod_{j=1}^J \Gamma(k_{ij})} \prod_{i=1}^I \prod_{j=1}^J \pi_{ij}^{k_{ij}-1},$$

where $K = \sum_{i=1}^I \sum_{j=1}^J k_{ij}$.

The posterior distribution is $Dirichlet(\mathbf{n} + \mathbf{k})$. Moreover, the marginal likelihood under M_1 is now given by

$$f(\mathbf{n} \mid M_1) = c \times \frac{\Gamma(K)}{\Gamma(n + K)} \prod_{i=1}^I \prod_{j=1}^J \frac{\Gamma(n_{ij} + k_{ij})}{\Gamma(k_{ij})}, \quad (3)$$

where $c = \frac{n!}{\prod_{i=1}^r \prod_{j=1}^c n_{ij}!}$.

Under the independence model M_0 we have that $\pi_{ij} = \pi_{i+}\pi_{+j}$ for all $i = 1, \dots, I$ and $j = 1, \dots, J$. So now the parameters of interest are the marginal probabilities vectors $\boldsymbol{\pi}^R = (\pi_{i+}; i = 1, \dots, I)$ and $\boldsymbol{\pi}^C = (\pi_{+j}; j = 1, \dots, J)$, that is $\boldsymbol{\vartheta}_0 = (\boldsymbol{\pi}^R, \boldsymbol{\pi}^C)$. By substituting $\pi_{ij} = \pi_{i+}\pi_{+j}$ in (2) we obtain

$$f(n | \boldsymbol{\pi}^R, \boldsymbol{\pi}^C) = c \times \prod_{i=1}^I \pi_{i+}^{n_{i+}} \prod_{j=1}^J \pi_{+j}^{n_{+j}}.$$

Assuming Dirichlet priors for $\boldsymbol{\pi}^R$ and $\boldsymbol{\pi}^C$ of the following form

$$\boldsymbol{\pi}^R \sim \text{Dirichlet}(k_1^R, \dots, k_I^R) \text{ and } \boldsymbol{\pi}^C \sim \text{Dirichlet}(k_1^C, \dots, k_J^C),$$

we obtain the marginal likelihood under M_0 model given by

$$f(\mathbf{n} | M_0) = c \times \frac{\Gamma(K^R)}{\Gamma(N + K^R)} \frac{\Gamma(K^C)}{\Gamma(N + K^C)} \prod_{i=1}^I \frac{\Gamma(n_{i+} + k_i^R)}{\Gamma(k_i^R)} \times \prod_{j=1}^J \frac{\Gamma(n_{+j} + k_j^C)}{\Gamma(k_j^C)} \quad (4)$$

From (3) and (4), the Bayes factor is given by

$$B_{10} = \frac{\mathcal{B}(\mathbf{n} + \mathbf{k})}{\mathcal{B}(\mathbf{k})} \times \frac{\mathcal{B}(\mathbf{k}^R)}{\mathcal{B}(\mathbf{n}^R + \mathbf{k}^R)} \times \frac{\mathcal{B}(\mathbf{k}^C)}{\mathcal{B}(\mathbf{n}^C + \mathbf{k}^C)} \quad (5)$$

where $\mathbf{n}^R = (n_{1+}, \dots, n_{I+})$, $\mathbf{n}^C = (n_{+1}, \dots, n_{+J})$ and $\mathcal{B}$ is the normalizing constant of the Dirichlet distribution (also known as multivariate beta function) given by

$$\mathcal{B}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{|\boldsymbol{\alpha}|} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{|\boldsymbol{\alpha}|} \alpha_i\right)} ;$$

for any vector $\boldsymbol{\alpha} = (\alpha_i; i = 1, \dots, |\boldsymbol{\alpha}|)$, with $|\boldsymbol{\alpha}|$ denoting the number of elements of $\boldsymbol{\alpha}$. The conjugate analysis of this section is summarized in Table 1.

Table 1: Summary of the Bayesian testing of the independence in the conjugate Multinomial-Dirichlet setup

H_0 : there is no association between the two variables	H_1 : there is association between the two variables
$M_0 : \mathbf{n} \boldsymbol{\pi}^R, \boldsymbol{\pi}^C \sim \text{Multinomial}(\mathbf{n}, \boldsymbol{\pi})$	$M_1 : \mathbf{n} \boldsymbol{\pi} \sim \text{Multinomial}(\mathbf{n}, \boldsymbol{\pi})$
$\boldsymbol{\pi} = \boldsymbol{\pi}^R [\boldsymbol{\pi}^C]^T$	$\boldsymbol{\pi} = (\pi_{ij})$
$\boldsymbol{\pi}^R = (\pi_{i+}) \sim \text{Dirichlet}(k_1^R, \dots, k_I^R)$	$\boldsymbol{\pi} \sim \text{Dirichlet}(\mathbf{k})$
$\boldsymbol{\pi}^C = (\pi_{+j}) \sim \text{Dirichlet}(k_1^C, \dots, k_J^C)$	$\mathbf{k} = (k_{ij})$
$i = 1, \dots, I$ and $j = 1, \dots, J$	$i = 1, \dots, I$ and $j = 1, \dots, J$

2.2 Log-linear Models

Poisson log-linear models are generalized linear models (GLMs) having as a response the cell frequencies of a contingency table. Associations between factors are expressed via interaction terms of the model. In the two-way contingency table, the model takes the form

$$\begin{aligned} n_{ij} &\sim \text{Poisson}(\lambda_{ij}), \quad i = 1, \dots, I, \quad j = 1, \dots, J \\ \log(\lambda_{ij}) &= \lambda_0 + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \end{aligned} \quad (6)$$

with identifiability constraints

$$\sum_{i=1}^I \lambda_i^A = \sum_{j=1}^J \lambda_j^B = \sum_{i=1}^I \lambda_{ij}^{AB} = \sum_{j=1}^J \lambda_{ij}^{AB} = 0.$$

Parameter λ_0 is an overall measure of average log-counts, while λ_i^A and λ_j^B are marginal effect terms for factors A (rows) and B (columns), respectively. The interaction term λ_{ij}^{AB} represents the association between A and B and reflects the departure or deviations from independence. In the frequentist framework, this model has 0 degrees of freedom and thus fits the data perfectly (all residuals are equal to zero).

If λ_{ij}^{AB} is equal to zero for all i and j , then we obtain the independence model with linear predictor

$$\log(\lambda_{ij}) = \lambda_0 + \lambda_i^A + \lambda_j^B. \quad (7)$$

Here we illustrate the comparison between the model of independence M_0 , given by the log-linear predictor (7) versus the saturated model M_1 with log-linear predictor (6). By this comparison we test for

$$H_0 : \lambda_{ij}^{AB} = 0 \quad \text{versus} \quad H_1 : \lambda_{ij}^{AB} \neq 0, \quad i = 1, \dots, I, \quad j = 1, \dots, J \quad (8)$$

which is equivalent to testing for independence.

For all log-linear parameters we consider normal prior distributions with zero means. The prior variances of λ_0 , λ_i^A (for $i = 2, \dots, I$) and λ_j^B (for $j = 2, \dots, J$) are set equal to large values in order to express prior ignorance. Nevertheless, the prior variance for the interaction parameters λ_{ij}^{AB} needs to be specified with caution due to the well-known sensitivity of the Bayes factors to this value for the parameters under testing. Therefore, the specification of the prior variance of λ_{ij}^{AB} is very important for avoiding the activation of the Lindley-Barlet paradox (Lindley, 1957; Bartlett, 1957). For this comparison, in the illustration of Section 3.2 we have used two different prior setups: (a) a Unit Information Empirical prior (UIE) proposed by Ntzoufras(2009, Chapter 9) and (b) a prior similar to the one proposed by Dellaportas and Foster (1999) (DF). For the first we set the prior variance of the interaction parameters equal to their posterior variances of

the full model multiplied by the sample size, in order to minimize the information introduced by the double use of the data. By this way, the contribution of the prior information is approximately equal to one data point; see for details Verdinelli and Wasserman (1995). This is an empirical Bayes approach and it cannot be considered as an orthodox Bayesian method since we use information from the data to specify the prior. Nevertheless, it can be used as a rough approximation of the actual onjective Bayesian procedure since the effect of the prior information on the posterior is minimized. For the second prior setup we use prior variance equal to two. This prior choice is the same to the one proposed by Dellaportas and Foster (1999) for the case of $I = J = 2$.

We calculate the posterior model probabilities of the two models under consideration by using the Gibbs variable selection sampler of Dellaportas, Forster and Ntzoufras (2002); also see in Ntzoufras, Forster and Dellaportas (2000) for an implementation in contingency tables using the Stochastic Search Variable Sampler. In order to implement any Gibbs variable selection, we introduce the latent binary indicator γ which takes the value of one when $\lambda_{ij}^{AB} \neq 0$ and zero otherwise. Then the log-linear predictors (6) and (7) are jointly summarized by

$$\log(\lambda_{ij}) = \lambda_0 + \lambda_i^A + \lambda_j^B + \gamma\lambda_{ij}^{AB}.$$

The posterior probabilities $P(\gamma = 1|\mathbf{n})$ and $P(\gamma = 0|\mathbf{n})$ will indicate wether the saturated ($\gamma = 1$) or the independence model ($\gamma = 0$), is more suitable for the data in hand. Gibbs variable selection can be implemented in a straightforward manner using standard Bayesian statistical packages, such as WinBUGS (see Ntzoufras, 2009, Chapter 11).

2.3 Sensitivity of the Bayes Factor

Posterior model probabilities and Bayes factor are highly sensitive to the prior specification of the model parameters. This result was reported after the publication of Lindley (1957), who reported a surprising behavior of the Bayes factor. When the sample size N increases, then the Bayes factor also increases and tends to infinity, fully supporting the simpler hypothesis in contrast to standard frequentist hypothesis test, which support the more complex hypothesis. This behavior is known as the Lindley's Paradox. Subsequently, motivated by the work of Lindley, Barlett (1957) extended this paradox by observing that the prior variance used in the Bayes factor also massively affects the supporting hypothesis as it tends to infinity. For the model comparison of Section 2.2 the Lindley-Bartlett paradox takes the form

$$Var(\lambda_{ij}) \rightarrow \infty \quad \Rightarrow \quad B_{01} \rightarrow \infty \quad \text{for any dataset } \mathbf{n}.$$

For the model comparison of Section 2.1, it is more difficult to express the Lindley-Bartlett paradox since the two models are not nested.

Therefore, the specification of the prior variance for the parameters under testing are of primary importance. When we do not have a clear indication about this value, we can perform sensitivity analysis of the Bayes factor in relation to the value of the prior variance for the parameter under testing. This procedure is similar to the regularization plots used for the selection of the shrinkage parameter in Lasso-type methods; see for example Lykou and Ntzoufras (2013).

2.4 Other Bayesian Model Selection Criteria

A variety of information criteria is available in the literature of model comparison and adequacy. The most popular information criteria are Akaike information criterion (AIC; Akaike, 1973) and the Bayes information criterion (BIC; Schwarz, 1978). More recently the deviance information criterion (DIC) was introduced by Spiegelhalter *et al.* (2002), as an extension of AIC. All these criteria are penalized likelihood measures giving different weight to complexity and goodness of fit. Lower values indicate better fitted models.

The BIC value for a model M is defined as

$$BIC(M) = D(\hat{\boldsymbol{\vartheta}}_M, M) + d_M \log N$$

where $D(\boldsymbol{\vartheta}_M, M)$ is the deviance measure of model M evaluated at $\boldsymbol{\vartheta}_M$, $\hat{\boldsymbol{\vartheta}}_M$ are the maximum likelihood estimates under model M and d_M is the dimension (number of parameters) of the model M . For large N and for specific prior families

$$-2\log B_{01} \approx BIC(M_0) - BIC(M_1) = \Delta BIC_{01},$$

see, for details Kass and Wasserman (1995). From this form we can obtain approximate posterior model probabilities, assuming the uniform prior distribution for all candidate models.

The Akaike information criterion is defined as

$$AIC(M) = D(\hat{\boldsymbol{\vartheta}}_M, M) + 2d_M$$

and is a penalized deviance measure with penalty equal to two for each estimated parameter. AIC supports less parsimonious models than BIC for $N > 7$.

Spiegelhalter *et al.* (2002) proposed the deviance information criterion (DIC) which is considered as the Bayesian analogue of AIC. DIC can be directly estimated via the MCMC output and can be applied in a much wider variety of models such as hierarchical, random effects and latent variable models.

2.5 Illustrative Examples

Here we implement the Bayesian tests of independence presented in this Section using data from several real examples. First, we implement the conjugate

analysis in two datasets: (a) a 5×3 cross-classification of the change in the clinical condition by degree of infiltration (i.e. the skin damage) from leprosy at the start of the experiment (Source: Agresti, 2013) and (b) a 4×2 contingency table of snoring by heart disease (Source: Norton and Dunn, 1985).

Table 2: Datasets (a) and (b) used for testing independence via conjugate analysis

Clinical Change	Degree of Infiltration		Snoring	Heart Disease	
	High	Low		Yes	No
Worse	1	11	Never	24	1355
Stationary	13	53	Occasionally	35	603
Slight improvement	16	42	Nearly every night	21	192
Moderate improvement	15	27	Every night	30	224
Marked improvement	7	11			

(a) Leprosy dataset (Agresti, 2013)

(b) Snoring dataset
(Norton and Dunn, 1985)

For the first dataset, the independence assumption is supported by the likelihood ratio test (LRT p -value = 0.122), AIC ($\Delta AIC_{01} = -0.72$) and BIC ($\Delta BIC_{10} = -13.84$) measures, while for the second dataset all the corresponding measures support the dependence between the two factors under consideration (LRT p -value $\ll 0.001$, $\Delta AIC_{01} = 59.9$ and $\Delta BIC_{01} = 42.45$).

Here we implement the Bayesian independence tests based on the conjugate analysis of Section 2.1. We consider prior values $k_{ij} = k_i^R = k_j^C = k$ and we perform sensitivity analysis over different values of k . From Figure 1 we observe that in dataset (a), the independence model is supported for a wide range of values, while for large values of k (accumulated prior information) the log-Bayes factor is stabilized in value just above zero. In contrast, in the second example (where the dependence is strongly supported by other methods) the Bayes factor offers conflicting evidence against M_0 for all values of k . Note that in both cases,

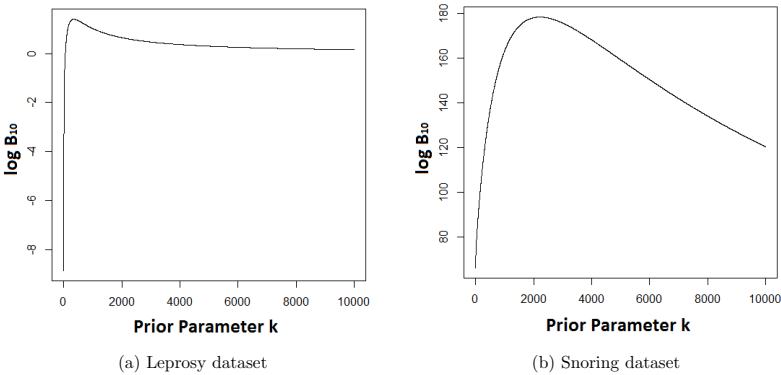


Figure 1: Sensitivity of Bayes factor to the Dirichlet prior parameter k for datasets (a) and (b) of Table 2.

the Bayes factor decreases for values of k close to zero. This is due to Lindley-Barlett paradox since in this case the prior variance of the probability parameters (induced by the imposed Dirichlet prior) is increasing.

We now proceed illustrating the Bayesian tests via log-linear models in three different 2×2 contingency tables: The first table is a cross-classification of 100 individuals randomly sampled from a large population as part of a study of sex differences by handedness (dataset c ; Source: <http://en.wikipedia.org>). The second contingency table (dataset d) refers to the number of ant and vertebrate dispersed plant species by seed and vegetative regeneration (Source: French and Westoby, 1996), while the third one consists of 200 randomly selected cancer patients examined for the cell phone usage by the brain cancer diagnosis (dataset e ; Source: <http://wiki.stat.ucla.edu>).

Table 3: Datasets (c)–(e) used for testing independence via log-linear models

	Right Handed	Left Handed		Ant	Vertebrate		Brain cancer yes	no
Males	43	9	Seed only	25	6	Cell yes	18	80
Females	44	4	Vegetative	36	21	Phone no	7	95

Dataset (c) Dataset (d) Dataset (e)
 Sex by Handness Dispersal by Regeneration Brain Cancer by Cell Phone Usage

Table 4 presents the estimated odds ratio, the p-values for testing independence based on the likelihood ratio test (LRT) with $\alpha = 0.05$ and the information criteria: DIC and BIC. For the two first examples H_0 is not rejected based on LRT in contrast to the third contingency table. In comparison with the information criteria, for the second contingency table both DIC and BIC indicate the saturated model as the best model. Therefore, the three selected datasets cover a variety of different levels of dependence which make them appropriate to illustrate model uncertainty via the Bayesian paradigm.

Table 4: Bayesian information criteria for datasets (c)–(e)

Dataset	Dependence	$\widehat{OR}$	LRT	DIC		BIC	
				M_0	M_1	M_0	M_1
(c)	Weak	0.43	Not rejected, p-value=0.30	26.38	26.73	24.53	29.85
(d)	Medium	2.43	Not rejected, p-value=0.14	28.07	26.91	30.54	30.34
(e)	Strong	3.05	Rejected, p-value=0.02	33.36	28.96	36.68	32.46

LRT: Decision concerning the independence (H_0) based on the likelihood ratio test with $\alpha = 0.05$

M_0 : independence model; M_1 : saturated model

Using Gibbs variable selection, we have estimated the posterior model probability $P(M_1|\mathbf{n}) = P(\gamma = 1|\mathbf{n})$ for a range of prior variance values for the interaction term λ_{22}^{AB} , which are depicted in Figure 2. From these figures two points are evident: (i) the surface between the line of the posterior model probability and 0.5 increases as we move from dataset (c) to dataset (e) and therefore with $\widehat{OR}$, and (ii) the posterior model probability of M_1 decreases as the prior variance increases after a threshold value. Concerning point (i), the posterior probability for

dataset (c) reaches its maximum value (which is just above 0.5) for prior variance lower than two. In dataset (d), the maximum of the posterior model probability is about 0.65, while for (e) is around 0.8. The main characteristic of the latter is that the posterior model probability of M_1 remains above 0.5 for all values of the prior variance presented here, indicating the strong association between brain cancer and cell phone usage. Point (ii) clearly depicts a realization of the Lindley-Bartlett paradox. For datasets with weak association (as in dataset c), the posterior probability of M_1 shrinks fast towards zero as the prior variance of the interaction term increases, while for strong associations (as in dataset e) the paradox is delayed and only appears for unrealistically large values of the prior variance.

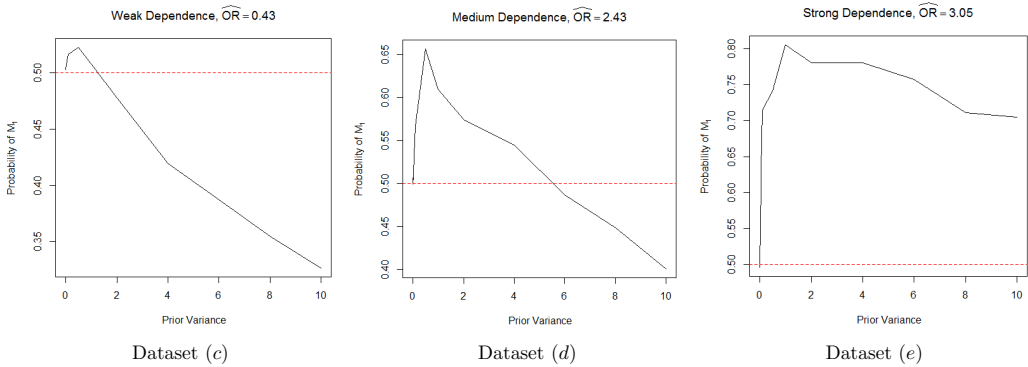


Figure 2: Posterior model probability of M_1 versus the prior variance of the interaction for datasets (c)–(e) of Table 3.

3. Bayesian Inference for Association Models

Association models describe the structure of the association between the two categorical variables in two-way contingency tables. These models assign scores to the classification variables. We denote by $\{\mu_1, \dots, \mu_I\}$ the row scores and by $\{\nu_1, \dots, \nu_J\}$ the column scores. Under this approach the interaction term is now written as $\lambda_{ij}^{AB} = \phi \mu_i \nu_j$ and therefore the linear predictor (6) is substituted by

$$\log(\lambda_{ij}) = \lambda_0 + \lambda_i^A + \lambda_j^B + \phi \mu_i \nu_j ;$$

where ϕ measures the correlation between row and column scores and is called intrinsic association parameter. For a comprehensive and detailed description of association models, along with their frequentists implementation in R, see Kateri (2014).

There are three types of association models, depending on the type of the row and column parameter score:

- Linear by Linear (LL) with fixed row and column scores. The most characteristic LL model is the Uniform (U), in case the scores are equidistant for the successive category.
- Row effect association model (R) with unknown row and fixed column scores.
- Column effect association (C) with unknown column and fixed row scores.
- Row-Column association model (RC) with both row and column scores to be parameters under estimation.

RC model is log-not-linear, because the predictor is a multiplicative in its parameters μ_i and ν_j . It treats both categorical variables as nominal and the model fit is invariant to any permutations of rows and/or columns. When one set of parameter scores is fixed, the RC model simplifies to the R or C model, which are linear function to its parameters. Hence, we are interested in comparing models $M \in \mathcal{M} = \{I, U, R, C, RC, S\}$; where I and S denote the independence and the saturated models, respectively.

Estimation of parameters of the RC model can be arduous since the likelihood may not be concave and therefore multiple local maxima may exist. For this reason, MCMC methods can be used to explore the whole posterior space. Goodman (1979) suggested an iterative model-fitting algorithm. An empirical Bayes analysis of the RC model has been considered by Chuang (1982) and Evans *et al.* (1993). In their approach, they set a prior distribution for parameters of the saturated log-linear model and then they estimated the posterior distribution of the parameters of the RC model through minimization of the Euclidean squared distance between the interaction terms of the two models. Kateri *et al.* (2005) provided Bayesian inference for the general RC(K) association model, with $RC(1)=RC$. Iliopoulos *et al.* (2007) and Iliopoulos *et al.* (2009) adopted alternative parametrizations for the RC model. Here, we use the parametrization of Iliopoulos *et al.* (2009). Hence we set

$$\mu_1 = \nu_1 = 0 \quad \text{and} \quad \mu_I = \nu_J = 1.$$

For the remaining of the parameters we assign normal prior distributions with mean zero and variance equal to two to be (roughly) in the lines of DF prior.

3.1 Marginal Likelihood Computation via the Laplace Approximation

Historically, a barrier for establishment of the Bayesian approach as a standard way to analyse data has been the difficulty in the computation of the posterior distribution when the prior is not conjugate. This problem has been solved in the early 90s with the development of MCMC methods for sampling from the posterior distribution of interest. Nevertheless, the computation of the marginal likelihood remains cumbersome motivating many alternative estimation methods; see, for example Perrakis *et al.* (2014). One of the most popular methods for the computation of the marginal likelihood is the Laplace approximation based

on the normal distribution. The Laplace method has been used from the late 80s for approximating Bayes factors in GLMs. It first appeared in 1988, in the technical report 121 of the Statistics Department of the University of Washington (Raftery, 1988). Leonard, Hsu, and Tsui (1989) facilitated Laplace approximations for approximating the marginal posterior density of summary measures of interest in contingency tables. Raftery (1996) employed the method to obtain approximate Bayes factors for GLMs. Under this approach, the marginal likelihood for contingency tables can be approximated by

$$\log f(\mathbf{n}|M) \approx \frac{d_m}{2} \log(2\pi) + \frac{1}{2} \log |H^*| + \log f(\boldsymbol{\vartheta}_M^* | M) + \log f(\mathbf{n} | \boldsymbol{\vartheta}_M^*, M),$$

where $\boldsymbol{\vartheta}_M^*$ is the posterior mode of model M , and H^* is minus the inverse Hessian evaluated at the posterior mode. This approximation works efficiently when the posterior distribution is relatively symmetric. To avoid analytic calculation of H^* and $\boldsymbol{\vartheta}_M^*$, we may use the Laplace-Metropolis estimator proposed by Lewis and Raftery (1997). Using this approach, we estimate the θ^* by the posterior mean and H^* by the posterior variance-covariance matrix from the MCMC output of the simulated values.

3.2 Illustrated Example (continued)

In this section, we use the data of a survey conducted at the University of Ioannina (Greece) in 1995 which studies the association between the use of cannabis and the alcohol consumption among students (Marselos *et al.*, 1997). The frequency of alcohol consumption is measured in a four-level scale while the use of cannabis through a three-level factor (never tried, once or twice, more frequently). These two ordinal factors are cross-classified in a 4×3 table presented in Table 5.

Table 5: Cannabis dataset (Marselos, 1997).

Alcohol consumption	Cannabis Use			
	Never	Once or twice	More often	Total
At most once/month	204	6	1	211
Twice/month	211	13	5	229
Twice/week	357	44	38	439
More often	92	34	49	175
Total	864	97	93	1054

Using this example, we illustrate the Bayesian model comparison of association models for two way contingency tables by approximating all marginal likelihoods using the Laplace-Metropolis estimator. Two normal prior distributions with zero means were used: (a) the unit empirical prior (UIE) where the prior variance of each interaction parameter is set equal to the corresponding posterior variance

multiplied by the sample size, and (b) a prior similar to the one proposed by Dellaportas and Foster (1999) with prior variance equal to two (DF).

Table 6: Bayesian measures of model evaluation for all models $M \in \mathcal{M}$ in the Cannabis dataset.

M_j	Model	DIC	BIC	Dif. BIC	$-2\log BF_{j2}$		Posterior Probabilities		
					UIE*	DF†	BIC	UIE*	DF†
1	Independence (I)	228.9	258.7	144.1	139.4	145.5	<0.01	<0.01	<0.01
2	Uniform (U)	81.1	114.8	0.0	0.0	0.0	0.956	0.973	0.801
3	Row (R)	82.8	128.1	13.4	17.9	9.1	<0.01	<0.01	<0.01
4	Column (C)	81.5	120.9	6.2	7.2	2.9	0.043	0.027	0.191
5	Row-Column (RC)	84.9	134.4	19.6	21.8	20.1	<0.01	<0.01	<0.01
6	Saturated (S)	88.2	147.7	32.9	40.4	20.5	<0.01	<0.01	<0.01

* UIE: Information Empirical prior

† DF: a prior similar to Dellaportas and Foster (1999)

All model comparison measures indicate that the Uniform model is the best. From Table 6 we observe that the DIC and BIC values for this model are clearly lower than those of remaining models. Similar is the picture from the posterior model probabilities where the U model is the maximum a-posteriori model (MAP) with probabilities 0.956, 0.973 and 0.801 for BIC, UIE prior and DF prior, respectively. Results under the two different prior setups differ due to the effect of the Lindley-Bartlett paradox. The posterior model probabilities under the empirical prior are closer to the corresponding BIC based probabilities due to the approximate unit information property of UIE. The column effect association model is supported as the second best one but with considerably lower probabilities (0.043, 0.027 and 0.191 for BIC, UIE prior and DF prior, respectively). These results are in agreement with the frequentists analysis (Kateri 2014; Chapter 6).

4. Conclusion

In this work, we have presented, reviewed and implemented Bayesian independence hypothesis tests in two-way contingency tables. We have also explored the sensitivity of Bayes factor and model selection on prior variances of model parameters. We have presented Bayesian model comparisons based both on conjugate and log-linear analysis. Furthermore, we have implemented Bayesian methods of model comparison via Laplace approximation for association models, which is broader and more flexible class of models describing association between categorical factors in a more parsimonious way. All methods have been illustrated on real datasets.

In the future, we will focus on developing compatible priors for model comparison in two-way tables. The use of power-priors seems to be appealing for this reason. Moreover, we will work for the construction of efficient algorithms for the computation of the marginal likelihood both in small and large scale problems. The mathematical properties of the Bayes factor for the comparison of association models will also be explored.

ΠΕΡΙΛΗΨΗ

Στην εργασία αυτή παρουσιάζονται έλεγχοι υποθέσεων για την ανεξαρτησία δύο κατηγορικών μεταβλητών σε πίνακες συνάφειας. Αρχικά εφαρμόζουμε συζυγή Μπεϋζιανή ανάλυση για το πολυωνυμικό-Dirichlet μοντέλο. Επικεντρωνόμαστε στα λογαριθμο-γραμμικά μοντέλα και συγκρίνουμε το κορεσμένο μοντέλο με το μοντέλο της ανεξαρτησίας με τη χρήση του παράγοντα Μπέυς, καθώς και άλλων κριτηρίων επιλογής μοντέλων. Τέλος αναπτύσσουμε τη Μπεϋζιανή συμπερασματολογία για τα μοντέλα συνάφειας, μια δημοφιλή κλάση μοντέλων που περιγράφουν τη δομή της σχέσης μεταξύ δύο κατηγορικών μεταβλητών. Εφαρμόζουμε MCMC μεθόδους με τη βοήθεια του στατιστικού πακέτου WinBUGS, με σκοπό να εκτιμηθούν οι παράμετροι των μοντέλων. Η προσέγγιστική μέθοδος Laplace χρησιμοποιείται για τον υπολογισμό της Μπεϋζιανής περιθώριας πιθανοφάνειας. Όλες οι μέθοδοι εφαρμόζονται σε μια σειρά πραγματικών δεδομένων.

Acknowledgements: This research was funded by the Research Center of the Athens University of Economics and Business (Funding program for the research publications of the AUEB Faculty members) and by the exchange program IKYDA 2014 of Greek State Scholarships Foundation (IKY) and the German Academic Exchange Service (DAAD).

REFERENCES

- Agresti, A. (2013). *Categorical Data Analysis*, 3rd edition, Wiley & Sons.
- Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. *2nd International Symposium on Information Theory*, Budapest: Akademia Kiado; p. 267–281.
- Bartlett, M. (1957). Comment on D.V. Lindley’s Statistical Paradox. *Biometrika*, **44**, 533–534.
- Chuang, C. (1982). Empirical Bayes methods for a two-way multiplicative-interaction model. *Communications in Statistics: Theory and Methods* **11**, 2977–2989.
- Dellaportas, P., Forster, J. and Ntzoufras, I. (2002). On Bayesian Model and Variable Selection Using MCMC. *Statistics and Computing*, **12**, 27–36.
- Evans, M., Gilula, Z. and Guttman, I. (1993). Computational Issues in the Bayesian Analysis of Categorical Data: Log-Linear and Goodman’s RC Model. *Statistica Sinica* **3**, 391–40
- Forster, J. and Dellaportas, P. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika*, **3**, 615–633.
- French K. and Westoby M. (1996). Vertebrate-dispersed species in a fire-prone environment. *Australian Journal of Ecology* **21**, 379–385.
- Goodman L. (1979). Simple models for the analysis of in cross-classifications having ordered categories. *Journal of American Statistical Association* **74** (367), 537–552.
- Gunel, E. and Dickey, J. (1974). Bayes factors for independence in contingency tables. *Biometrika* **61** (3), 545–557.
- Iliopoulos G., Kateri M. and Ntzoufras I. (2007). Bayesian estimation of unrestricted and order-restricted association model for a two-way contingency table. *Computational Statistics and Data Analysis* **51**, 4643–4655.

- Iliopoulos G., Kateri M. and Ntzoufras I. (2009). Bayesian model comparison for the order restricted RC association model. *Psychometrika* **74**, 561–587.
- Jeffreys, H. (1961). *Theory of Probability*, 3rd edition. Classic Texts in the Physical Sciences. Oxford University. Press, Oxford.
- Kass, R.E. and Wasserman, L. (1996). The Selection of Prior Distributions by Formal Rules. *Journal of the American Statistical Association*, **91**, 1343–1370.
- Kateri M. (2014). *Contingency Table Analysis: Methods and Implementation Using R*, Birkhäuser.
- Kateri, M., Nicolaou, A. and Ntzoufras, I. (2005). Bayesian Inference for the RC(m) Association Model. *Journal of Computational and Graphical Statistics*, **14**, 116–138.
- Leonard T., Hsu, J. and Tsui K. (1989). Bayesian marginal inference. *Journal of the American Statistical Association* **84**, 1051–1057.
- Lewis S. and Raftery A. (1997). Estimating Bayes Factors via Posterior Simulation with the Laplace-Metropolis Estimator. *Journal of the American Statistical Association* **92**, 648–655.
- Lindley, D.(1957). A Statistical Paradox. *Biometrika*, **44**, 187–192.
- Lykou, A. and Ntzoufras, I. (2013). On Bayesian Lasso Variable Selection and the Specification of the Shrinkage Parameter. *Statistics and Computing* **23**, 361–390.
- Marselos M., Boutsouris K., Liapi H., Malamas M., Kateri M. and Papaioannou T. (1997). Epidemiological Aspects on the Use of Cannabis among University Students in Greece. *European Addiction Research* **3**, 184–191.
- Norton, P. and Dunn, E. (1985). Snoring as a risk factor for disease: an epidemiological survey. *British Medical Journal*, **291**, 630–632.
- Ntzoufras I., Forster J. and Dellaportas P. (2000) Stochastic search variable selection for log-linear models. *Journal of Statistical Computation and Simulation* **68**, 23–37.
- Ntzoufras I. (2009). *Bayesian Modeling Using WinBUGS*, 1st ed. New York: John Wiley & Sons.
- Pearson K. (1904). *On the Theory of Contingency and its Relation to Association and Normal Correlation*, London: Dulau & Co.
- Perrakis, K., Ntzoufras, I. and Tsionas, E. (2014). On the Use of Marginal Posteriors in Marginal Likelihood Estimation via Importance-Sampling. *Computational Statistics and Data Analysis*, **77**, 54–69.
- Raftery, A. (1988). Approximate Bayes Factors for Generalized Linear Models. Technical Report **121**, Department of Statistics, University of Washington.
- Raftery A. (1996). Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Model. *Biometrika* **2**(83), 251–266.
- Schwarz G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, **6** 461–464.
- Spiegelhalter, D. and Smith, A. (1982). Bayes Factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society, Series B* **44**, 377–387.
- Verdinelli, I. and Wasserman, L. (1995) Computing Bayes factor using a generalization of the Savage-Dickey density ratio, *Journal of the American Statistical Association*, **90**, 614–618.



EARTHQUAKE INTER-EVENT TIMES DISTRIBUTION: APPLICATION TO CORINTH GULF

Mesimeri Maria and Karakostas Vassilios

Geophysics Department, School of Geology, Aristotle University of Thessaloniki
{mmesimer, vkarak}@geo.auth.gr

ABSTRACT

The earthquake inter-event times distribution reveals the triggering mechanism of an area or a seismic excitation. In the case of independent events the hypothesis that the earthquake inter-event times follow the exponential distribution is not rejected; then also other distributions considered as generalizations of the exponential can be examined, e.g., gamma and Weibull. Especially for seismic excitations, it is interesting to examine if/to what extent earthquake inter-event times can be accepted to follow gamma, Weibull or even lognormal distribution, whereas the exponential distribution is rejected in most cases. This indicates that these earthquakes are interacting, leading to an earthquake–earthquake triggering mechanism. In the present study the inter-event times distribution of the earthquakes occurred in the western part of Corinth Gulf is examined. The area is chosen due to high seismic activity of low and intermediate magnitude earthquakes. The complete data for the period 2008-2014 are used due to the adequate density of the national seismological network which contributes to a better estimation of focal coordinates of the earthquakes. The declustered catalog, which is a product of declustering algorithms, approaches the time series of independent events whereas the clusters are the seismic excitations. The probability density function (pdf) for the earthquake inter-event times is calculated for the declustered catalog and for particular clusters. The distribution that best describes the empirical pdf is chosen in order to relate the properties of the distribution with temporal evolution of seismicity.

Keywords: earthquake interevent times, declustering, probability density function

1. INTRODUCTION

An earthquake catalog provides information regarding the origin time, the focal coordinates (longitude, latitude and depth) and the magnitude of each individual earthquake. When we are referring to an earthquake catalog, we consider all the recorded earthquakes in an area, including all types of seismic excitations, i.e., mainshock-aftershock sequences and earthquake swarms, along with background seismicity. The difference between the background seismicity of an area and seismic excitations is that background seismicity is considered to be homogeneous in time

following a stationary Poisson process. Seismic excitations are considered as dependent events (non-homogeneous Poisson processes) triggered by each other. The background seismicity could be extracted by an earthquake catalog after applying a proper declustering algorithm. This procedure is resulting to two different catalogs, the one containing the background seismicity for a certain region in a certain time and the other comprising all the seismic excitations. The declustered catalog is then tested in order to find if the occurrence of background seismicity is a Poisson process (eg. Luen and Stark, 2012; Jacobs et al., 2013).

Several distributions have been proposed for describing the inter-event time distribution of background seismicity (Bak et al., 2002; Corral, 2004; Touati et al., 2009; Langenbruch et al., 2011; Chen et al., 2013; Davidsen and Kwiitek, 2013). It is found that the occurrence of the background seismicity could be described by power law, gamma distribution, Weibull distribution or an exponential distribution. Another approach concerns the use of a generalization of Boltzmann-Gibbs statistical physics, referred as non-extensive statistical physics (e.g. Vallianatos et al., 2014; 2016 and references therein). The aim of the present study is to test the performance of declustered algorithms and find whether or not provide us with Poissonian declustered catalogs. In addition, the inter-event times distribution for individual clusters (seismic excitations) is also searched and compared to the one describing the background seismicity.

2. METHODS

2.1 Declustering algorithms

Seismicity declustering is the process of separating an earthquake catalog into seismic excitations (clusters) and the background seismicity (declustered catalog). Several declustering algorithms have been proposed over the years (vanStiphout et al., 2012 and references therein). In the present study three different algorithms, namely Gardner and Knopoff, (1974) (GK), Reasenber, (1985) (RSB) and CURATE (Jacobs et al. 2013), are examined.

Gardner and Knopoff (1974) proposed a simple declustering algorithm which identifies aftershocks in a certain space (d) time (t) window depending on the magnitude (M) of the mainshock (eq. 1 and 2).

$$d(M) = 10^{0.1238 \cdot M + 0.983} \text{ (km)} \quad (1)$$

$$t(M) = \begin{cases} 10^{0.032 \cdot M + 2.7389}, & M \geq 6.5 \\ 10^{0.5409 \cdot M - 0.547}, & M < 6.5 \end{cases} \text{ (days)} \quad (2)$$

Reasenber (1985) proposed a method for identifying aftershocks by linking earthquakes to clusters according to spatial and temporal interaction zones (van

Stiphout et al., 2012). The algorithm requires as input the minimum ($\tau_{\min}$) and maximum ($\tau_{\max}$) duration of the earthquake sequence in days, the interaction zone in km multiplied by a factor (R_{fact}), the probability of an earthquake (p_1) being part of a seismic excitation and the magnitude of completeness. These three parameters are associated according to equation 3 where ΔM is the difference in magnitude between the largest shock and the magnitude of completeness. Equation 4 defines the interaction zone in kms.

$$\tau = \frac{-\ln(1-p_1)t}{10^{\frac{2(\Delta M-1)}{3}}} \text{ (days)} \quad (3)$$

$$r = 0.011 \cdot 10^{0.4 \cdot M} \cdot R_{fact} \text{ (km)} \quad (4)$$

The CURATE method uses seismicity rate as the main link between earthquakes (Jacobs et al., 2013). Equations 5 and 6 describe the calculation carried out for each day in the catalog, where D is the daily average number of earthquakes, t_s is the time of the first earthquake in the catalog, t_i is the time in days since the beginning of the earthquake catalog, t_f is the time of the last earthquake in the catalog, and $N(t_s - t)$ is the number of earthquakes observed between time t_s and t :

$$D = \frac{N(t_s, t_f)}{(t_f - t_s)} \quad (5)$$

$$CURATE = N(t_s - t_i) - D(t_i - t_s), \quad t_s \leq t_i \leq t_f \quad (6)$$

The (consecutive) events with inter-event times less than $1/\text{mean}(\text{rate})$ are marked as potential sequences. Then a distance rule is applied, which limits the area size of the potential sequences, along with a time window rule.

2.2 Inter-event times

The inter-event, i.e., time between consecutive events, is described by equation

$$\tau_i \equiv t_i - t_{i-1}, \quad (7)$$

where t_i and t_{i-1} denote the time of the i^{th} and $i-1^{\text{th}}$ earthquakes in the selected space-time-magnitude window. Corral (2006) defined the probability density function ($D(\tau)$) of the inter-event times as the number of inter-event times within a small interval values, normalized to the total number of inter-event times and divided by the length of the small interval,

$$D(\tau) \equiv \frac{\Pr[\tau \leq \text{rec.time} < \tau + d\tau]}{d\tau}, \quad (8)$$

where $\Pr$ denotes probability, τ refers to a precise value of inter-event time, and $d\tau$ is the length of the interval.

2.3 Statistical Analysis

Four distributions (lognormal, Weibul, gamma and exponential) were tested for finding the one which best describes the inter-event time distribution in each data set. The probability density functions (pdf) for the four distributions are given by equations 9-12, respectively,

$$f(x|\mu, \sigma) = \frac{1}{\chi\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, x > 0, \quad (9)$$

where μ is the mean and σ is the standard deviation of the random variable's natural logarithm,

$$f(x|a, b) = \left(\frac{b}{a}\right) \left(\frac{x}{a}\right)^{b-1} e^{-\left(\frac{x}{a}\right)^b}, x > 0 \quad (10)$$

where a is the scale parameter and b is the shape parameter,

$$f(x|a, b) = \frac{1}{b^a \Gamma(a)} x^{a-1} e^{-\frac{x}{b}} \quad (11)$$

where $\Gamma(\cdot)$ denotes the Gamma function.

$$f(x|\mu) = \frac{1}{\mu} e^{-\frac{x}{\mu}}, \quad (12)$$

and μ is the mean. The parameter estimation for each distribution is calculated by applying the Maximum Likelihood Estimation (MLE) method (Johnson et al., 1994).

For the sake of comparison Kolmogorov-Smirnov test (K-S test) is applied (eq. 13)

$$D = \max \left| \hat{F}(x) - G(x) \right| \quad (13)$$

where $\hat{F}(x)$ is the empirical cumulative density function (cdf) and $G(x)$ is the cdf of the hypothesized distribution. Additionally, the Akaike and Bayesian Information Criteria are also considered (AIC and BIC, respectively). The two criteria are given by the following equations:

$$AIC = -2\ln(L) + 2k \quad (14)$$

where $\ln(L)$ is the log-likelihood function and k the number of parameters, and

$$BIC = -2\ln(L) + k \ln(n) \quad (15)$$

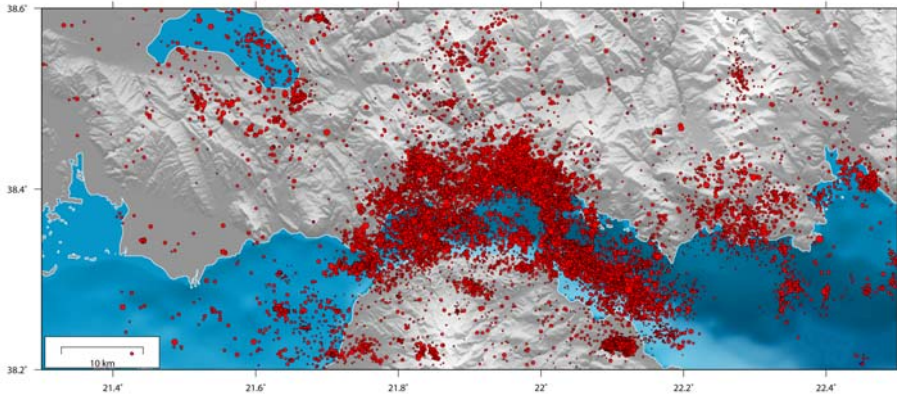
where $\ln(L)$ is the log-likelihood function, k the number of parameters and n the number of observations.

3. REGION

Corinth gulf and especially its western part (Fig. 1) is one of the most seismically active areas in Greece. An earthquake catalog is compiled using all the available earthquake phases from Geophysics Department of Aristotle University of Thessaloniki (<http://geophysics.geo.auth.gr/ss>) and these from National Observatory of Athens (<http://bbnet.gein.noa.gr/HL/>) for the period 2008-2014. In order to obtain highly accurate focal coordinates the initial catalog is relocated using the double difference technique (Waldhauser and Ellsworth, 2000). The final catalog is consisted of approximately 22,000 events.

A very important parameter, which has to be considered when temporal properties of seismicity are examined, is the magnitude of completeness. The completeness magnitude M_c is theoretically defined as the lowest magnitude at which 100% of the earthquakes in a space-time volume are detected (Rydelek and Sacks, 1989). After applying the goodness of fit method (Wiemer and Wyss, 2000) it is found that the region should be divided into two periods due to differences in M_c . Particularly, for 2008-2010 (A) M_c is found equal to 2.0 whereas for 2011-2014 (B) $M_c=1.4$. All the calculations, from now on, are made for the two distinct periods (A, B).

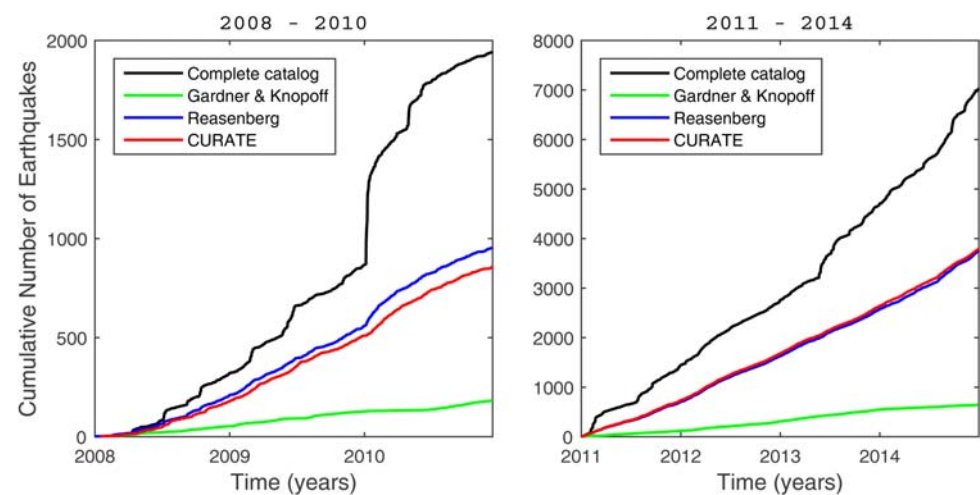
Figure 1 Seismicity map of western Corinth gulf



4. RESULTS

The aforementioned declustering algorithms are applied for each period (A: 2008-2010 and B: 2011-2014). Several sets of parameters regarding space and time are used for RSB and CURATE algorithm. Particularly for RSB, $\tau_{\min}$ is set to 1 whereas for $\tau_{\max}$ values between 3-5 days are tested. R_{fact} is set to 5 and 10. The distance and day rule for CURATE are 3-5 km and 1-3 days. The calculations are performed for all possible combinations. In this study we present the results for $\tau_{\min} = 1$, $\tau_{\max} = 3$ and $R_{\text{fact}} = 5$ for RSB algorithm for both periods. Regarding the CURATE algorithm we present the results obtained for 3 km and 2 days for period A and 2 km and 2 days for period B. In Figure 2 is shown the cumulative number of earthquakes in relation with time for the complete datasets along with the declustered catalogs for each algorithm. The complete catalogs show that the rate of occurrence differs with time and particularly there are periods presenting very high seismicity rate. On the other hand, declustered catalogs' seismicity rate could be consider as almost constant but not the same for the three cases of declustering algorithms.

Figure 2 Cumulative number of events in relation with time for the complete datasets and the declustered catalogs for the two distinct periods.



The GK declustering algorithm provides a declustered catalog with considerably less events in comparison to the RSB and CURATE algorithms (Table 1). It recognizes almost all events as clusters due to the large time and space windows that are used in the declustering process. RSB and CURATE provide declustered catalogs with similar number of events but they are differing significantly in the number of events characterized as clusters and the number of clusters.

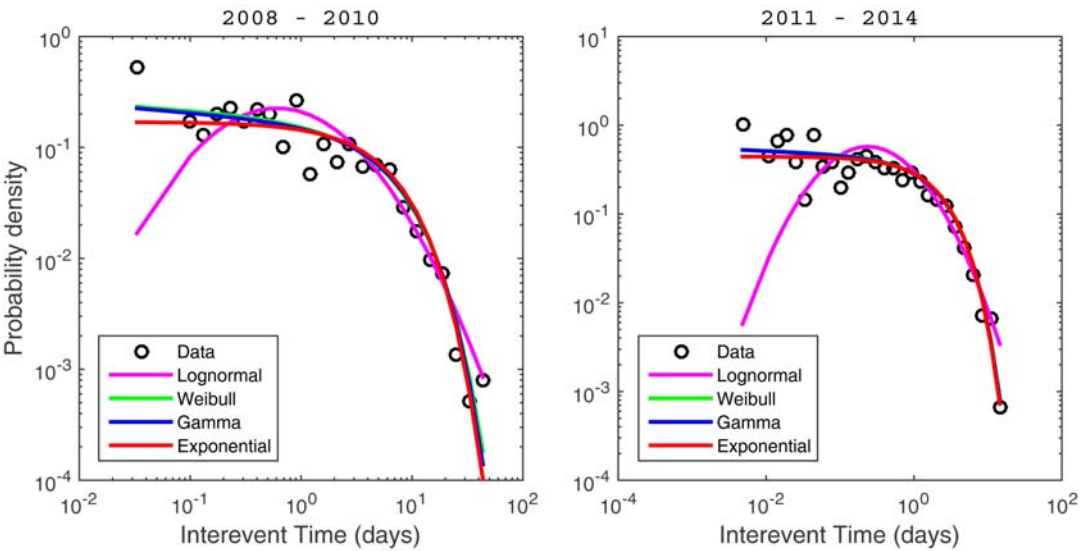
Table 1 Number of events for different datasets for the two distinct periods.

Algorithms/	A	B

Catalogs	GK	RSB	CURATE	GK	RSB	CURATE
All events	1943	1943	1943	7032	7032	7032
Declustered	183	955	866	644	3763	3797
Events in clusters	1855	1117	1444	6720	3724	4833
N of clusters(N>10)	27	17	29	69	58	70

Then for each algorithm the probability density function of the inter-event times for the two periods is calculated and four distributions, namely lognormal, Weibull, gamma, exponential are fitted to the data. The K-S test along with AIC and BIC criteria were used for comparing the distributions. As best fitted distribution is considered the one with the lowest value of K-S, AIC and BIC.

Figure 3 Probability density function for the inter-event times of the declustered catalog after the application of GK algorithm for the two distinct periods.



In Figure 3 the plots of the probability density function for the GK algorithm for the two periods are shown. The estimated parameters for the fitted distributions are shown on Table 2.

Table 2 MLE parameters estimates, 95% confidence intervals and log likelihood calculation for each period.

Alg.	Distr.	A			B		
		Parameters	Confidence Intervals	-lnL	Parameters	Confidence Intervals	-lnL

Alg.	Distr.	A			B		
		Parameters	Confidence Intervals	-lnL	Parameters	Confidence Intervals	-lnL
GK	Logn.	$\mu=1.13$ $\sigma=1.27$	[0.95, 1.32] [1.15, 1.42]	509	$\mu=0.20$ $\sigma=1.28$	[0.11, 0.30] [1.21, 1.35]	1207
	Weibull	$a=5.68$ $b=0.92$	[4.81, 6.70] [0.83, 1.03]	503	$a=2.21$ $b=0.96$	[2.04, 2.41] [0.91, 1.02]	1164
	Gamma	$a=0.91$ $b=6.40$	[0.76, 1.10] [5.06, 8.09]	504	$a=0.96$ $b=2.34$	[0.87, 1.05] [2.06, 2.65]	1165
	Exp.	$\mu=5.88$	[5.11, 6.84]	504	$\mu=2.25$	[2.08, 2.43]	1165
RSB	Logn.	$\mu=-0.66$, $\sigma=1.49$	[-0.75, -0.56] [1.42, 1.56]	1105	$\mu=-1.59$ $\sigma=1.34$	[-1.64, -1.55] [1.31, 1.37]	430
	Weibull	$a=1.01$ $b=0.83$	[0.93, 1.10] [0.79, 0.87]	1038	$a=0.37$ $b=0.92$	[0.35, 0.38] [0.90, 0.95]	164
	Gamma	$a=0.76$ $b=1.47$	[0.70, 0.82] [1.32, 1.63]	1043	$a=0.90$ $b=0.42$	[0.86, 0.93] [0.40, 0.45]	169
	Exp.	$\mu=1.12$	[1.05, 1.20]	1068	$\mu=0.38$	[0.37, 0.39]	182
CURATE declustered	Logn.	$\mu=-0.388$ $\sigma=1.25$	[-0.47, -0.30] [1.19, 1.31]	1085	$\mu=-1.52$ $\sigma=1.26$	[-1.56, -1.48] [1.24, 1.29]	492
	Weibull	$a=1.20$ $b=0.94$	[1.12, 1.29] [0.90, 0.99]	1048	$a=0.38$ $b=1.00$	[0.37, 0.39] [0.98, 1.03]	149
	Gamma	$a=0.96$ $b=1.29$	[0.88, 1.04] [1.15, 1.43]	1050	$a=1.01$ $b=0.37$	[0.97, 1.05] [0.35, 0.39]	149
	Exp.	$\mu=1.23$	[1.16, 13.2]	1051	$\mu=0.38$	[0.37, 0.39]	149
CURATE clusters	Logn.	$\mu=-4.34$ $\sigma=1.57$	[-4.54, -4.14] [1.44, 1.73]	-600	$\mu=-4.27$ $\sigma=1.90$	[-4.49, -4.04] [1.75, 2.07]	-625
	Weibull	$a=0.029$ $b=0.59$	[0.023, 0.036] [0.54, 0.64]	-568	$a=0.03$ $b=0.55$	[0.02, 0.04] [0.50, 0.60]	-611
	Gamma	$a=0.45$ $b=0.11$	[0.39, 0.52] [0.09, 0.14]	-539	$a=0.41$ $b=0.16$	[0.36, 0.47] [0.13, 0.20]	-586
	Exp.	$\mu=0.053$	[0.047, 0.061]	-468	$\mu=0.068$	[0.060, 0.076]	-476

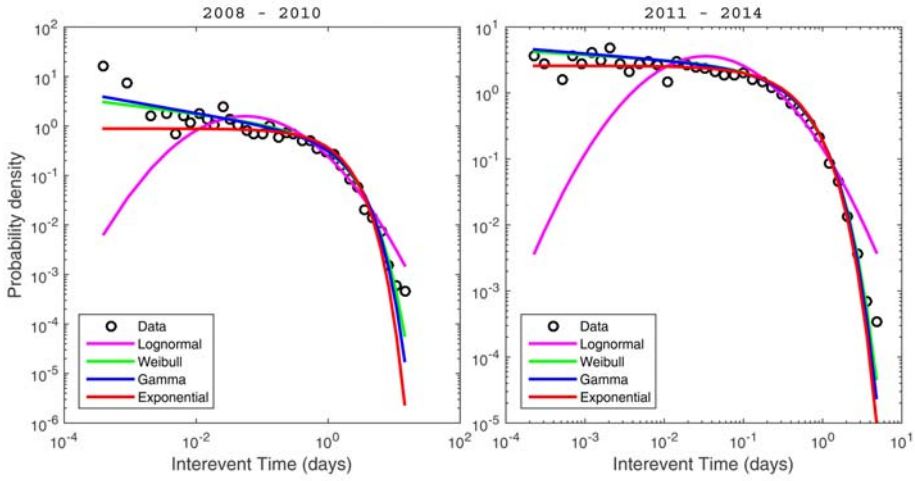
According to the tests applied for each distribution (Table 3) the Weibul, gamma and exponential are well fitted to the data whereas the least fitted, but not rejected, is the lognormal one.

Table 3 Results of Kolmogorov - Smirnov test along with the AIC, BIC and p values. The critical values for period A are 0.277, 0.2329, 0.2439, 0.2522 and for period B 0.2567, 0.2264, 0.2296, 0.2329 for each dataset, respectively. $\alpha=0.05$.

Alg.	Distr.	A				B			
		AIC	BIC	K-S	p-value	AIC	BIC	K-S	p-value
GK	Logn.	1023	1025	0.114	0.87	2419	2422	0.103	0.89
	Weibull	1011	1014	0.079	0.99	2333	2336	0.058	0.99
	Gamma	1012	1014	0.085	0.98	2334	2336	0.060	0.99
	Exp.	1011	1012	0.089	0.98	2332	2334	0.064	0.99
RSB	Logn.	2215	2218	0.11	0.76	865	869	0.102	0.80
	Weibull	2080	2083	0.062	0.99	333	336	0.054	0.99
	Gamma	2091	2094	0.075	0.98	342	346	0.060	0.99
	Exp.	2139	2141	0.105	0.80	367	369	0.074	0.97
CURE declusterd	Logn.	2175	2178	0.120	0.71	988	991	0.127	0.57
	Weibull	2101	2104	0.093	0.92	303	306	0.063	0.99
	Gamma	2105	2108	0.097	0.90	303	306	0.063	0.99
	Exp.	2104	2105	0.100	0.88	301	303	0.062	0.99
CURE clusters	Logn.	-1197	-1194	0.078	0.98	-1247	-1244	0.058	0.99
	Weibull	-1132	-1130	0.135	0.61	-1219	-1216	0.101	0.83
	Gamma	-1075	-1072	0.205	0.14	-1169	-1166	0.168	0.26
	Exp.	-934	-932	0.375	0.0003	-951	-949	0.321	0.0012

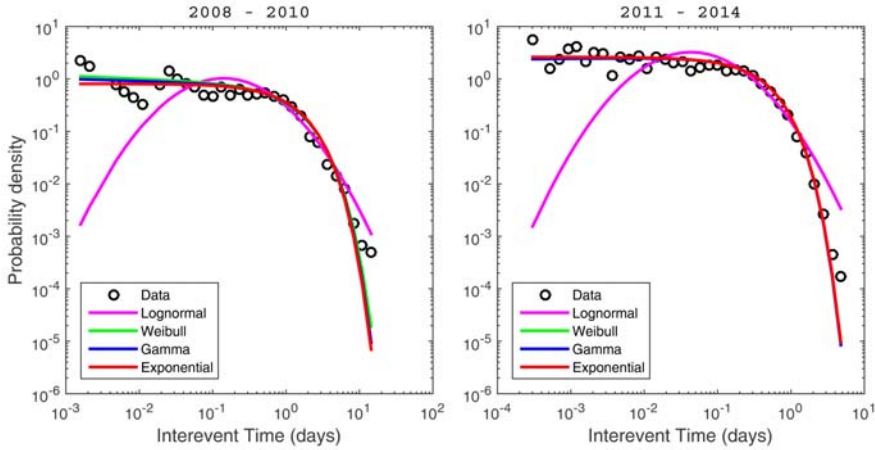
For the RSB algorithm (Fig. 4) the best fitted distribution is the Weibull distribution with $b < 1$. Considering the other distributions, the Gamma and exponential fit well to the data. The lognormal distribution is again least fitted.

Figure 4 Same as Figure 3 for RSB algorithm



For the CURATE algorithm (Fig. 5) the best fitted distribution is Weibull, with $b > 1$, for period A and the exponential for period B. However, all the three distributions are very close and could describe the data (Weibull, gamma, exponential). The lognormal distribution is far from being the best model for both periods.

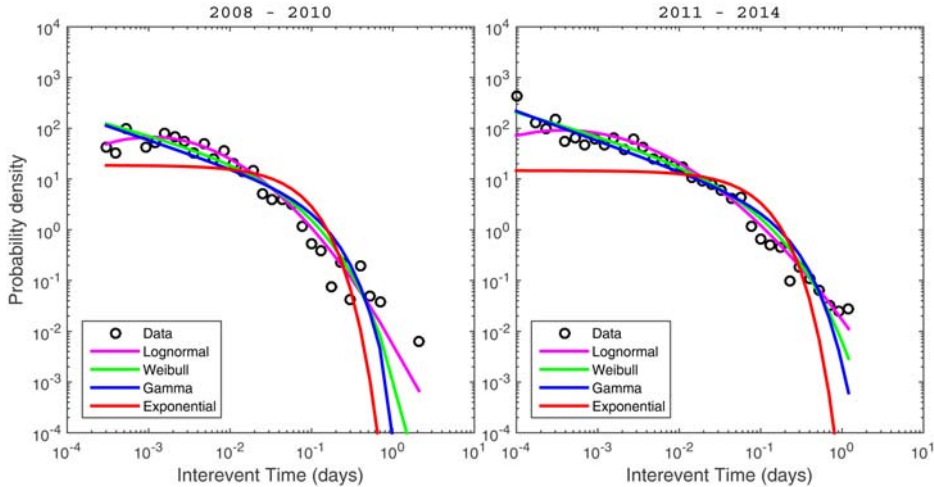
Figure 5 Same as Figure 3 for CURATE algorithm



After visual inspection of the spatial distribution of the clusters identified by the three algorithms, we concluded that the most reliable results, in means of space time properties, are given by CURATE algorithm. Therefore, the two largest clusters identified by CURATE algorithm are used in order to find the probability density distribution of their inter-event times (Fig. 6). The distribution best describing the

data is the lognormal one. The exponential distribution is rejected (Table 3) whereas the gamma distribution is far from the best model. Weibull distribution on the other hand, is not the best model but is near to be according to K-S test.

Figure 6 Probability density function for the inter-event time of two clusters (seismic excitations) after the application of CURATE algorithm.



5. DISCUSSION - CONCLUSIONS

The inter-event times of the declustered catalogs, produced after the application of three different declustering algorithms, were tested in order to find which distribution best describes the data. It was found that GK produces declustered catalogs which could be described by the exponential distribution but the events remained in these are very few in comparison with the initial one. The few events remained on the declustered catalog is a consequence of the time-space windows that the method applies. This leads to the identification of a large fraction of seismicity as clusters with a smaller one characterized as background seismicity. RSB's declustered catalogs are fitted by Weibull distribution with $b < 1$, which means that a fraction of clustered seismicity still remained in the declustered catalog and not all the excitations were extracted. CURATE could produce declustered catalogs which are fitted by exponential distribution and having also an adequate number of events remaining. Finally, the inter-event times distributions of seismic excitations are best fitted by lognormal distribution, revealing a totally different occurrence mechanism. It is noteworthy that the exponential distribution has been rejected for the seismic excitations which lead to the conclusion that the occurrence of the earthquakes in a seismic excitation is a non-homogeneous Poisson process.

ΠΕΡΙΛΗΨΗ

Η κατανομή των χρόνων μεταξύ διαδοχικών σεισμών αποτυπώνει το μηχανισμό διέγερσης της περιοχής ή της εξεταζόμενης σεισμικής εξάρσης και την αλληλεπίδραση μεταξύ διαδοχικών σεισμών. Στην περίπτωση ανεξαρτήτων μεταξύ τους σεισμών η υπόθεση της εκθετικής κατανομής για τους ενδιαμέσους χρόνους δεν απορρίπτεται, ενώ αυτοί μπορούν να περιγραφούν από διάφορες κατανομές (εκθετική, gamma, Weibull). Σε περιπτώσεις σεισμικών ή μικροσεισμικών εξάρσεων οι χρόνοι μεταξύ διαδοχικών σεισμών μπορεί να θεωρηθεί ότι ακολουθούν κατανομές gamma, Weibull, ή lognormal, ενώ η εκθετική κατανομή απορρίπτεται. Αυτό υποδεικνύει αλληλεπίδραση μεταξύ των σεισμών και διέγερση των επόμενων από τους προηγούμενους. Η κατανομή των χρόνων μεταξύ διαδοχικών σεισμών χρησιμοποιείται στην παρούσα εργασία για τη μελέτη των ιδιοτήτων της χρονικής εξέλιξης της σεισμικότητας στο Δυτικό Κορινθιακό κόλπο, μιας από τις πιο ενεργές σεισμικά περιοχές στον ελληνικό χώρο, με την εκδήλωση ισχυρών αλλά και ενδιαμέσου και μικρού μεγέθους σεισμών. Χρησιμοποιούνται τα πλήρη δεδομένα σεισμών από το 2008 ως το 2014 επειδή το σεισμολογικό δίκτυο στο χώρο έγινε αρκετά πυκνό με αποτέλεσμα τον ακριβέστερο προσδιορισμό των εστιακών παραμέτρων μεγάλου αριθμού σεισμών. Ο κατάλογος σεισμών που προκύπτει με την εφαρμογή αλγορίθμων αποσυσταδοποίησης προσεγγίζει τη χρονοσειρά των ανεξάρτητων μεταξύ τους σεισμών, ενώ οι συστάδες αποτελούν τις σεισμικές εξάρσεις. Γίνεται υπολογισμός της συνάρτησης πυκνότητας πιθανότητας των χρόνων μεταξύ διαδοχικών σεισμών του αποσυσταδοποιημένου καταλόγου αλλά και των σεισμικών εξάρσεων και επιλογή της κατανομής που προσαρμόζεται καλύτερα στα δεδομένα, με τελικό σκοπό τη συσχέτιση τους με τις ιδιότητες της χρονικής εξέλιξης της σεισμικότητας.

Acknowledgments: We would like to thank Prof. G. Tsaklidis for his useful remarks and suggestions in writing this paper.

REFERENCES

- Bak P., Christensen K., Danon L. and Scanlon T. (2002). Unified Scaling law for Earthquakes. *Physical Review Letters*, **88**, 10–13. doi: 10.1103/PhysRevLett.88.178501
- Chen C.-H., Wang J.-P., Wu Y.-M., Chan C.-H. and Chang C.-H. (2013). A study of earthquake inter-occurrence times distribution models in Taiwan. *Natural Hazards*, **69**, 1335–1350.
- Corral A. (2004). Long-Term Clustering, Scaling, and Universality in the Temporal Occurrence of Earthquakes. *Physical Review Letters*, **92**, 108501–1. doi: 10.1103/PhysRevLett.92.108501
- Corral A. (2006). Dependence of earthquake recurrence times and independence of magnitudes on seismicity history. *Tectonophysics*, **424**, 177–193. doi: 10.1016/j.tecto.2006.03.035
- Davidson J. and Kwiatak G. (2013). Earthquake interevent time distribution for induced micro-, nano-, and picoseismicity. *Physical Review Letters*, **110**, 1–5. doi: 10.1103/PhysRevLett.110.068501
- Gardner J. and Knopoff L. (1974). Is the sequence of earthquakes in southern

- California, with aftershocks removed, Poissonian? *Bulletin of Seismological Society of America*, **64**, 1363–1367.
- Jacobs K.M., Smith E.G.C., Savage M.K. and Zhuang J. (2013). Cumulative rate analysis (CURATE): A clustering algorithm for swarm dominated catalogs. *Journal of Geophysical Research: Solid Earth*, **118**, 553–569. doi: 10.1029/2012JB009222
- Johnson N.L., Kotz S. and Balakrishnan N. (1994). *Continuous Univariate Distributions*, New York: John Wiley.
- Langenbruch C., Dinske C. and Shapiro S.A. (2011). Inter event times of fluid induced earthquakes suggest their Poisson nature. *Geophysical Research Letters*, **38**, 1–6. doi: 10.1029/2011GL049474
- Luen B. and Stark P.B. (2012). Poisson tests of declustered catalogues. *Geophysical Journal International*, **189**, 691–700. doi: 10.1111/j.1365-246X.2012.05400.x
- Reasenber P. (1985). Second-Order Moment of Central California Seismicity, 1969–1982. *Journal of Geophysical Research*, **90**, 5479–5495.
- Rydelek P.A. and Sacks I.S. (1989). Testing the completeness of earthquake catalogs and the hypothesis of self-similarity. *Nature*, **337**, 251–253.
- Touati S., Naylor M. and Main I.G. (2009). Origin and nonuniversality of the earthquake interevent time distribution. *Physical Review Letters*, **102**, 1–4. doi: 10.1103/PhysRevLett.102.168501
- van Stiphout T., Zhuang J. and Marsan D. (2012). Seismicity Declustering. *Community Online Resource for Statistical Seismicity Analysis*, 1–25. doi: 10.5078/corssa-52382934.
- Vallianatos F., Karakostas V. and Papadimitriou E. (2014). A Non-Extensive statistical physics view in the spatiotemporal properties of the 2003 (M_w 6.2) Lefkada, Ionian Island Greece, aftershock sequence. *Pure and Applied Geophysics*, **171**, 1343–1354. doi: 10.1007/s00024-013-0706-6
- Vallianatos F., Michas G. and Papadakis G. (2016). A description of seismicity based on non-extensive statistical physics: A review. *Earthquakes and their impact on society*, 1–41. doi: 10.1007/978-3-319-21753-6_1
- Waldhauser F. and Ellsworth W.L. (2000). A Double-difference Earthquake location algorithm: Method and application to the Northern Hayward Fault, California. *Bulletin of Seismological Society of America*, **90**, 1353–1368. doi: 10.1785/0120000006
- Wiemer S. and Wyss M. (2000). Minimum magnitude of completeness in earthquake catalogs: Examples from Alaska, the Western United States, and Japan. *Bulletin of Seismological Society of America*, **90**, 859–869. doi: 10.1785/0119990114



OPTIMAL ADAPTIVE MAINTENANCE AND QUALITY CONTROL MODEL FOR PROCESSES SUBJECT TO FAILURE AND MULTIPLE QUALITY SHIFTS AFFECTING BOTH LOCATION AND SCALE

Konstantinos A. Tasias and George Nenes

University of Western Macedonia, Department of Mechanical Engineering

ktasias@uowm.gr, gnenes@uowm.gr

ABSTRACT

This paper presents a new model for the economic-statistical optimization of a Variable-Parameter (VP) Shewhart control scheme. The proposed model is utilized for monitoring processes where apart from multiple independent assignable causes, affecting both the mean and variance, failures can also occur. Each time the control scheme signals an alarm, preventive maintenance actions are initiated which are obviously preferable to corrective maintenance actions, required after a failure. The realistic assumption of imperfect preventive maintenance actions has been considered. The optimal design parameters of the scheme are selected through an optimization problem formulated by the long-run average cost per time unit minimization, subject to statistical constraints.

Keywords: Quality Control, Condition-Based Maintenance, Markov Processes, Multiple Assignable Causes, Cost minimization

1. INTRODUCTION

A main objective of contemporary industrial processes is to achieve ongoing high reliability along with low maintenance costs. Several general types of maintenance philosophies have been developed in the literature to reduce unexpected breakdowns which have a twofold negative effect: they reduce the efficiency of the process and they incur high restoration costs. The most sophisticated maintenance policy is known as condition-based maintenance (CBM), where preventive maintenance (PM) actions are carried out based on condition monitoring techniques. CBM is normally preferable to corrective maintenance (CM) actions, required after a breakdown. The first condition-based preventive maintenance models were proposed by Derman (1962, 1963).

It is apparent that an intuitive relationship between *CBM* and *SPC* exists in real applications. Both of these interrelated scientific fields are based on condition monitoring techniques and effective integration of them can definitely improve the process performance. The first joint optimization of *CBM* and *SPC* can be found in Paté-Cornell *et al.* (1987) and Tagaras (1988). Moreover, Rahim and Banerjee (1993) presented an age-dependent preventive policy to tackle the problem of equipment failure for processes subject to deterioration. Ben-Daya and Duffuaa (1995), Duffuaa and Ben-Daya (1995), Linderman *et al.* (2005) and Ivy and Nembhard (2005) also recognized the strong link between maintenance and quality. Furthermore, Panagiotidou and Tagaras (2010) concluded that the benefits from *CBM* and *SPC* integration can be substantial.

However, there are few studies in the literature that monitor both the process location and scale through the formulation of a quality and maintenance model, even though, quality shifts that affect both the mean and the variability of a process are commonly found in practical applications. Chiu and Huang (1995) presented the economic design of joint $\bar{X}$ and S^2 control charts for preventive maintenance. The economic-statistical design of control charts that monitor jointly the mean and variability of the process was studied by Caballero Morales (2013) who considered both static and *VSI* control charts for the integration of *PM* with general failure rates. Yin *et al.* (2015) also pointed out the need of joint monitoring of mean and variance in multi-device integrated *SPC* and maintenance models.

In this paper we propose an integrated *SPC* and maintenance model by employing a *VP* control scheme for joint monitoring of mean and variance of processes where multiple independent assignable causes, affecting both the mean and variance, are possible to occur. Moreover, in this paper we have assumed the process restoration by *PM* actions not to be necessarily perfect.

2. ASSUMPTIONS AND PROBLEM SETTING

2.1 Model Assumptions

- The process is monitored through a critical quality characteristic X , which is normally distributed with target mean μ_T and target standard deviation σ_T .
- The process starts its operation in state of statistical control ($\mu_0 = \mu_T, \sigma_0 = \sigma_T$).
- Multiple independent assignable causes may occur at random times affecting the mean and/or the standard deviation and shifting the process to out-of-control operating states.
 - Each quality shift increases the probability of an inferior shift to occur and scales-up the failure rate, in proportion to the magnitude of the shift to the process.
 - Failures may occur any time within the production cycle.

- Out-of-control operation is not directly observable, in contrast to the failure state, which seizes the process operation.
- The process is not self-correcting, so, only transitions to inferior states may occur.
- An out-of-control signal triggers an inspection to examine whether a quality shift has actually occurred or not.
- After a confirmed quality shift upon inspection, the process may be restored to any non-inferior state with known (estimated) probabilities. Therefore, an imperfect preventive maintenance action is considered.
- Whenever a failure occurs, a corrective maintenance action is initiated and the process is perfectly restored to the *IC* state.
- A *PM* action is preferable compared to a *CM* action.
- The time horizon is infinite.

2.2 Problem Setting

The process is subject to a multiple assignable cause mechanism, which consists of m quality shifts of the mean and r of the standard deviation ($m \geq 1, r \geq 1$), plus a failure. The assignable causes may shift the mean from μ_0 to $\mu_i = \mu_0 + \delta_i \sigma_0$ ($i \in [1, m], \delta_i > \delta_{i-1} \geq 0$) and the standard deviation from σ_0 to $\sigma_j = \gamma_j \cdot \sigma_0$ ($j \in [1, r], \gamma_j > \gamma_{j-1} \geq 1$). The actual state of the process is defined by the assignable causes that may have occurred and is denoted by (i, j) .

The time to the occurrence of each quality shift (and failure) follows a non-negative exponential distribution. The occurrence rates of the assignable causes are denoted by $\lambda_{x(i \rightarrow k)}$ ($\lambda_{s(j \rightarrow l)}$) for a transition of the mean (standard deviation) from i to $k, k > i$ (j to $l, l > j$). Due to the dependence of the process state on the failure rate, the latter is denoted by $\lambda_{Fx|i}$ ($\lambda_{Fs|j}$) when the process mean (standard deviation) is under the effect of i (j). As already mentioned, $\lambda_{Fx|k} > \lambda_{Fx|i}$ for $k > i$ ($\lambda_{Fs|l} > \lambda_{Fs|j}$ for $l > j$). The overall failure rate when the process operates under the effect of any state (i, j) equals the sum of the failure rate of the process mean $\lambda_{Fx|i}$ and the failure of the standard deviation $\lambda_{Fs|j}$ ($\lambda_{F|(i,j)} = \lambda_{Fx|i} + \lambda_{Fs|j}$).

The transition rates to any inferior state, either for the mean or for the standard deviation of the process, when operating under quality shifts i and j , respectively, are denoted by $v_{x,i}$ and $v_{s,j}$ $\left(v_{x,i} = \sum_{k=i+1}^m \lambda_{x(i \rightarrow k)}, v_{s,j} = \sum_{l=j+1}^r \lambda_{s(j \rightarrow l)} \right)$.

Due to the assumption of imperfect *PM*, the process may be restored from any state $(i, j) \neq (0, 0)$ to any other superior state (u, v) according to given probabilities, which are denoted by $q_{(i,j)}^{(u,v)}$. Apparently, $q_{(i,j)}^{(u,v)} = 0$ for

$$(u > i) \cup (v > j), \sum_{u=0}^i \sum_{v=0}^j q_{(i,j)}^{(u,v)} = 1 \text{ and } q_{(0,0)}^{(0,0)} = 1.$$

This general assumption for imperfect *PM* is realistic in many production processes and the restoration probabilities may be estimated from statistical data upon a Phase I analysis of the monitored process. A perfect *PM* policy can be easily accounted by setting $q_{(i,j)}^{(0,0)} = 1$ and $q_{(i,j)}^{(u,v)} = 0$ for any $(u, v) \neq (0, 0)$.

The process operation under an *OOC* state increases the production cost per time unit by $M_{(i,j)}$ ($M_{(k,l)} > M_{(i,j)}$ for $k > i, l > j$). The cost of the investigation of the process in case of a false alarm is denoted by L_0 . Moreover, the detection of a quality shift is usually preferable to precede a failure because the cost of a *PM* action, $L_{(i,j)}$, is normally considerably lower compared to the cost of a *CM* action, L_F ($L_F > L_{(k,l)} > L_{(i,j)} > L_{(0,0)} = 0$ for $k > i > 0, l > j > 0$). Apparently, sampling costs, including fixed b and variable c costs, are, also, considered.

In addition to the aforementioned costs, the restoration of the process, either due to a quality shift or a failure, imposes time delays to the process. Similarly to the restoration costs, the expected time of the investigation in case of a false alarm is denoted by T_0 . Furthermore, the time delays are higher in case of a failure T_F than an *OOC* signal and the expected time for the process restoration from an assignable cause is positively correlated to the effect of the assignable cause to the process ($T_F > T_{(k,l)} > T_{(i,j)} > T_{(0,0)} = 0$ for $k > i > 0, l > j > 0$).

3. MODEL FORMULATION

The process is monitored through a two-chart $VP \bar{X}$ -s control scheme determined by the values of the sample size n , sampling interval h , warning and control limit coefficients for the chart that monitors the mean (w_x, k_x) and the respective ones for the standard deviation's chart (w_s, k_s) . All the aforementioned variables $\{n_q, h_q, w_{x,q}, w_{s,q}, k_{x,q}, k_{s,q}\}$ have two possible values, a relaxed ($q=1$) and a tightened one ($q=2$).

For easiness of model presentation, the process mean is monitored through the standardized sample mean, denoted by z_t , which is compared to the respective warning and control limit coefficients (w_x, k_x) . On the other hand, each sample's standard deviation, denoted by s_t , is compared to the upper warning and control limits of the control chart that monitors the standard deviation (UWL_s, UCL_s) , which are derived from $UWL_{s,q} = (c_{4,q} + w_{s,q} \sqrt{1 - c_{4,q}^2}) \cdot \sigma_0$ and $UCL_{s,q} = (c_{4,q} + k_{s,q} \sqrt{1 - c_{4,q}^2}) \cdot \sigma_0$, where $c_{4,q} = (\sqrt{2 / (n_q - 1)}) \cdot (\Gamma(n_q / 2) / \Gamma((n_q - 1) / 2))$.

At each sampling instance, z_t and s_t of the collected sample are computed and one of the following three decisions, is made:

(a) If $z_t > k_{x,q}$ and/or $s_t > UCL_{s,q}$, there is an out-of-control signal and the process is halted for investigation which reveals either a false alarm and no action is taken or an *OOC* operation; the latter dictates a *PM* action which restores the process either perfectly to the *IC* state or partially to a non-inferior state with known probabilities. Without loss of generality, we utilize the relaxed scheme, i.e. $\{n_1, h_1, w_{x,1}, w_{s,1}, k_{x,1}, k_{s,1}\}$, after a *PM* action. The model can be easily modified to consider tightened control as a protection against imperfect process restoration, especially if the restoration probabilities to *OOC* states are high enough to justify the extra cost imposed by considering a large size after the shortest interval for the first sample after a *PM* action.

(b) If $w_{x,q} < z_t \leq k_{x,q}$ and/or $UWL_{s,q} < s_t \leq UCL_{s,q}$, there is no interruption, but there is a warning and the tightened set of parameters should be used $\{n_2, h_2, w_{x,2}, w_{s,2}, k_{x,2}, k_{s,2}\}$.

(c) If $z_t \leq w_{x,q}$ and $s_t \leq UWL_{s,q}$, the process continues its operation and relaxed parameters are utilized.

As already mentioned, if at any time the process operation is seized because of a failure, the process will be restored to the *IC* state through a *CM* action.

The operation of the proposed scheme is stochastically modeled through a three-dimensional discrete time Markov chain. The model state is fully defined by the actual state of the process, which at sampling instance t is denoted by Y_t , and its possible values are $Y_t = (0, 0), \dots, (m, r)$, plus the failure state F , combined with the decision made at each transition step a_t ($a_t = 0$ if the statistic lies within the central zone, $a_t = 1$, in case of a warning and $a_t = 2$, in case of an

OOC signal). Therefore, the state space includes $\left[(m+1) \times (r+1) \times 3 + 1\right]$ possible states.

In order to compute the transition probabilities of the above matrix, we need to first calculate the probability of the process moving from one state to another, denoted by p . Given that the mean and the standard deviation of the process are affected independently by the assignable causes, the probability of a process transition from state (i, j) to (k, l) $\forall i \neq k$ and $\forall j \neq l$ within an interval of duration h_q , which prerequisites a failure-free interval, is derived from the following expression:

$$p_{(i,j),(k,l)}(h_q) = \int_0^{h_q} \sum_{y=i+1}^k \lambda_{x(i \rightarrow y)} \cdot \exp\left(-\left(v_{x,i} + v_{s,j} + \lambda_{Fx|i} + \lambda_{Fs|j}\right) \cdot t\right) \cdot p_{(y,j),(k,l)}(h_q - t) dt + \int_0^{h_q} \sum_{z=j+1}^l \lambda_{s(j \rightarrow z)} \cdot \exp\left(-\left(v_{x,i} + v_{s,j} + \lambda_{Fx|i} + \lambda_{Fs|j}\right) \cdot t\right) \cdot p_{(i,z),(k,l)}(h_q - t) dt \quad (1)$$

In the non-desirable case, where a failure occurs within a sampling interval of duration h_q , the process transits to state F , the production cycle ends, and afterwards, the process is restored to the IC state. A failure, when the process operates under (i, j) , may either occur without the advent of any quality shift or after the deterioration of the process to an inferior state. The probability of a process transition to the failure state (state F), is:

$$p_{(i,j),F}(h_q) = \int_0^{h_q} \left(\lambda_{Fx|i} + \lambda_{Fs|j}\right) \cdot \exp\left(-\left(\lambda_{Fx|i} + \lambda_{Fs|j}\right) \cdot t\right) \cdot \exp\left(-\left(v_{x,i} + v_{s,j}\right) \cdot t\right) dt + \int_0^{h_q} \sum_{y=i+1}^m \lambda_{x(i \rightarrow y)} \cdot \exp\left(-\left(v_{x,i} + v_{s,j} + \lambda_{Fx|i} + \lambda_{Fs|j}\right) \cdot t\right) \cdot p_{(y,j),F}(h_q - t) dt + \int_0^{h_q} \sum_{z=j+1}^r \lambda_{s(j \rightarrow z)} \cdot \exp\left(-\left(v_{x,i} + v_{s,j} + \lambda_{Fx|i} + \lambda_{Fs|j}\right) \cdot t\right) \cdot p_{(i,z),F}(h_q - t) dt \quad (2)$$

Additionally, the transition probabilities depend on the probability of each decision to be made at each sampling interval, which can easily be derived from the scheme's operation. Consequently, the exact expressions for the transition probabilities for a process moving from any state, i.e. (i, j) , to any other state, i.e. (k, l) , are (for $a_{t-1} = 0(1) \quad q=1(2)$):

$$Prob_{(i,j)a_{t-1}}^{(i,j)a_t}(h_q) = \begin{cases} p_{(i,j)}^{(i,j)}(h_q) \cdot \Phi\left(\frac{w_{x,q} - \delta_k \sqrt{n_q}}{\gamma_l}\right) \cdot P\left(\chi_{n_q-1}^2 < \left(\frac{UWL_{s,q}}{\gamma_l \cdot \sigma_0}\right)^2 \cdot (n_q - 1)\right) & a_t = 0 \\ p_{(i,j)}^{(i,j)}(h_q) \cdot \left[\Phi\left(\frac{k_{x,q} - \delta_k \sqrt{n_q}}{\gamma_l}\right) \cdot P\left(\chi_{n_q-1}^2 < \left(\frac{UCL_{s,q}}{\gamma_l \cdot \sigma_0}\right)^2 \cdot (n_q - 1)\right) - \right. \\ \left. - \Phi\left(\frac{w_{x,q} - \delta_k \sqrt{n_q}}{\gamma_l}\right) \cdot P\left(\chi_{n_q-1}^2 < \left(\frac{UWL_{s,q}}{\gamma_l \cdot \sigma_0}\right)^2 \cdot (n_q - 1)\right) \right] & a_t = 1 \quad (3) \\ p_{(i,j)}^{(i,j)}(h_q) \cdot \left[1 - \Phi\left(\frac{k_{x,q} - \delta_k \sqrt{n_q}}{\gamma_l}\right) \cdot P\left(\chi_{n_q-1}^2 < \left(\frac{UCL_{s,q}}{\gamma_l \cdot \sigma_0}\right)^2 \cdot (n_q - 1)\right) \right] & a_t = 2 \end{cases}$$

In the special case of a process transition from any state (i, j) to failure state (state F): $Prob_{(i,j)a_{t-1}}^{(i,j)a_t}(h_q) = p_{(i,j)}^{(i,j)}(h_q)$. On the other hand, whenever a failure occurs the process resumes its operation in the IC state and the transition probabilities are $Prob_{(k,l)a_t}^{(0,0)0}(h_1) = Prob_{(k,l)a_t}^{(0,0)0}(h_1)$.

Finally, in case an alarm is signaled ($a_{t-1} = 2$) and the process was actually operating OOO ($Y_t \neq (0,0)$) the transition probabilities equal the product of multiplying all the non-negative restoration probabilities to the respective transition probabilities after PM is performed.

$$Prob_{(i,j)a_t}^{(i,j)2}(h_q) = \sum_{i'=0}^i \sum_{j'=0}^j q_{(i',j')}^{(i,j)} \cdot Prob_{(k,l)a_t}^{(i',j')0}(h_1) \quad (4)$$

The limiting probability for the process transition to each possible state $Y_t = (k, l) \cup F$ ($k, l = (0,0), \dots, (m, r)$), denoted by $\pi_{Y_t a_t}$ (in case of failure ($Y_t = F$) $\pi_{Y_t a_t}$ is denoted by π_F) can be easily computed by solving the following linear system:

$$\pi_{Y_t a_t} = \sum_{Y_{t-1}=(0,0) \cup F}^{(m,r) \cup F} \sum_{a_{t-1}=0}^2 Prob_{Y_{t-1} a_{t-1}}^{Y_t a_t}(h_q) \cdot \pi_{Y_{t-1} a_{t-1}} \quad \text{and} \quad \sum_{Y_t=(0,0) \cup F}^{(m,r) \cup F} \sum_{a_t=0}^2 \pi_{Y_t a_t} = 1 \quad (5)$$

4. EXPECTED COST PER TIME UNIT

The expected cost per time unit, denoted by ECT , based on the renewal theory, may be expressed as the ratio of the expected cycle cost, denoted by EC , over the expected cycle time length, denoted by ET , ($ECT=EC/ET$).

An important parameter of the computation of both the aforementioned measures is the definition of all the possible scenarios regarding the chronological sequence of the assignable causes for any process transition, because different permutations result to different costs and time delays for the process. So, we first implement a Depth First Search (DFS) algorithm to define the set of the possible scenarios for a specific process transition and, then, the computation of ET and EC is presented.

4.1 Depth First Search (DFS) Algorithm

This brute-force algorithm is employed to define the set of the possible scenarios for a process transition from state (i, j) (root) to another specific state (k, l) , if and only if $i \neq k$ and $j \neq l$, denoted by $SC_{(i,j) \rightarrow (k,l)}$. In case either only the process mean is affected by a transition ($j = l$) or only the standard deviation ($i = k$), there is a unique scenario as regards the occurrence of the assignable causes, based on the assumption that only transitions to inferior states may occur.

The DFS algorithm is utilized in order to traverse a directed graph made by the set of process states (nodes) V and the edges between them whenever a transition is possible. The adjacency matrix E can be easily created from the occurrence rates matrix by substituting every occurrence rate for a process transition from state (i, j) to (k, l) greater than zero with: 0, if $i \neq k$ and $j \neq l$ or $i = k$ and $j = l$; 1, if $i = k$ or $j = l$.

An easy way to present the algorithm and its implementation to our model is through the following steps:

Step 1: Every node in V is considered as unmarked. Set $SC=0$

Step 2: Select an unvisited (unmarked) node w_1 that is adjacent to the root node $w_0 = (i, j)$ ($w_1 = Adj(w_0)$).

Step 3: Mark node w_1 and check whether it is the final state (k, l) .

Step 4a: If yes, store this path to $SC_{(i,j) \rightarrow (k,l)}$, go to the predecessor node w_0 and choose another unmarked and adjacent one ($w_1 = Adj(w_0)$).

Step 4b: If not, apply the procedure (Steps 2-4) with all the adjacent to w_1 , unmarked nodes ($w_2 = Adj(w_1)$) and so on.

Step 5: Repeat until all adjacent to w_0 nodes have been visited (marked).

It is apparent that a process transition from state (i, j) to (k, l) may occur through multiple different ways as regards the number of assignable causes. If we denote by fn_1 (fn_2) the number of the assignable causes that occur within a sampling interval and affect the process mean (standard deviation), then $fn_1 \leq k - i$ ($fn_2 \leq l - j$), $fn_1, fn_2 \geq 0$.

The probability of each scenario to occur within an interval of duration h_q , when a total number of $(fn_1 + fn_2)$ assignable causes occur, is denoted by $\Pr(h_q, (fn_1 + fn_2))_{(i,j),(k,l)}$ and can be computed from the following expression:

$$\begin{aligned} \Pr(h_q, (fn_1 + fn_2))_{(i,j),(k,l)} &= \int_0^{h_q} \lambda_1 \cdot \exp(-(v_{x,i} + v_{s,j} + \lambda_{Fx|i} + \lambda_{Fs|j}) \cdot t_1) \dots \\ &\dots \int_{t_1}^{h_q} \lambda_2 \cdot \exp(-(v_{x,(1)} + v_{s,(1)} + \lambda_{Fx|(1)} + \lambda_{Fs|(1)}) \cdot (t_2 - t_1)) \dots \\ &\dots \int_{t_{[(fn_1+fn_2)-1]}}^{h_q} \lambda_{(fn_1+fn_2)} \cdot \exp\left(-\left(\begin{aligned} &v_{x,[(fn_1+fn_2)-1]} + v_{s,[(fn_1+fn_2)-1]} + \\ &+ \lambda_{Fx|[(fn_1+fn_2)-1]} + \lambda_{Fs|[(fn_1+fn_2)-1]} \end{aligned}\right) \cdot (t_{(fn_1+fn_2)} - t_{[(fn_1+fn_2)-1]})\right) \cdot \\ &\cdot \exp(-(v_{x,k} + v_{s,l} + \lambda_{Fx|k} + \lambda_{Fs|l}) \cdot (h_q - t_{(fn_1+fn_2)})) dt_{(fn_1+fn_2)} \dots dt_2 dt_1 \end{aligned} \quad (6)$$

In the above equation, λ_1 notates the occurrence rate of the first assignable cause of the scenario to occur, λ_2 of the second one, etc.

Similarly, the expected time of the occurrence of the g^{th} assignable cause of a specific scenario within an interval of duration h_q can be computed from equation (8).

It should be mentioned that the product of the multiplication of the adjacency matrix E by itself, i.e. E^2 , defines the possible 2-length paths $((fn_1 + fn_2) = 2)$, E^3 the possible 3-length paths $((fn_1 + fn_2) = 3)$, etc., for a process transition from one state to another specific state.

$$\begin{aligned}
\tau^{(g)}(h_q) = & \left[\int_0^{h_q} t_1 \lambda_1 \cdot \exp\left(-(\nu_{x,i} + \nu_{s,j} + \lambda_{Fxl|i} + \lambda_{Fsl|j}) \cdot t_1\right) \dots \right. \\
& \dots \int_{t_1}^{h_q} t_2 \lambda_2 \cdot \exp\left(-(\nu_{x,(1)} + \nu_{s,(1)} + \lambda_{Fxl|(1)} + \lambda_{Fsl|(1)}) \cdot (t_2 - t_1)\right) \dots \\
& \dots \int_{t_{((f_{n_1} + f_{n_2}) - 1)}}^{h_q} t_{(f_{n_1} + f_{n_2})} \lambda_{(f_{n_1} + f_{n_2})} \cdot \exp\left(-\left(\nu_{x,[(f_{n_1} + f_{n_2}) - 1]} + \nu_{s,[(f_{n_1} + f_{n_2}) - 1]} + \right.\right. \\
& \left. \left. + \lambda_{Fxl|[(f_{n_1} + f_{n_2}) - 1]} + \lambda_{Fsl|[(f_{n_1} + f_{n_2}) - 1]}\right) \cdot \left(t_{(f_{n_1} + f_{n_2})} - t_{[(f_{n_1} + f_{n_2}) - 1]}\right) \right) \cdot \\
& \left. \exp\left(-(\nu_{x,k} + \nu_{s,l} + \lambda_{Fxl|k} + \lambda_{Fsl|l}) \cdot (h_q - t_{(f_{n_1} + f_{n_2})})\right) \right] dt_{(f_{n_1} + f_{n_2})} \dots dt_2 dt_1 \\
& \left[\int_0^{h_q} \lambda_1 \cdot \exp\left(-(\nu_{x,i} + \nu_{s,j} + \lambda_{Fxl|i} + \lambda_{Fsl|j}) \cdot t_1\right) \dots \right. \\
& \dots \int_{t_1}^{h_q} \lambda_2 \cdot \exp\left(-(\nu_{x,(1)} + \nu_{s,(1)} + \lambda_{Fxl|(1)} + \lambda_{Fsl|(1)}) \cdot (t_2 - t_1)\right) \dots \\
& \dots \int_{t_{((f_{n_1} + f_{n_2}) - 1)}}^{h_q} \lambda_{(f_{n_1} + f_{n_2})} \cdot \exp\left(-\left(\nu_{x,[(f_{n_1} + f_{n_2}) - 1]} + \nu_{s,[(f_{n_1} + f_{n_2}) - 1]} + \right.\right. \\
& \left. \left. + \lambda_{Fxl|[(f_{n_1} + f_{n_2}) - 1]} + \lambda_{Fsl|[(f_{n_1} + f_{n_2}) - 1]}\right) \cdot \left(t_{(f_{n_1} + f_{n_2})} - t_{[(f_{n_1} + f_{n_2}) - 1]}\right) \right) \cdot \\
& \left. \exp\left(-(\nu_{x,k} + \nu_{s,l} + \lambda_{Fxl|k} + \lambda_{Fsl|l}) \cdot (h_q - t_{(f_{n_1} + f_{n_2})})\right) \right] dt_{(f_{n_1} + f_{n_2})} \dots dt_2 dt_1
\end{aligned} \tag{7}$$

where, $t_u = 1$, for $u = \{1, 2, \dots, (f_{n_1} + f_{n_2})\} \setminus \{g\}$

4.2 Expected Cycle Time Length

The expected duration of a transition step may be evaluated as the weighted average of the expected duration of each of the possible states of the process.

The components of the expected cycle time length, for the proposed model, are the following:

- (a) the expected production time in the *IC* state $\tau_{IC/(0,0)}(h_q)$.
- (b) the expected production time in the *OO*C state, when the process operates under the effect of state (i, j) $i, j \geq 0$ at the beginning of the interval $\tau_{OO/(i,j)}(h_q)$.
- (c) the restoration time from any possible state (k, l) $T_{(k,l)}$.
- (d) the restoration time from failure T_F .

Specifically, the process operates in the *IC* state until either an assignable cause or a failure occurs. So, the expected *IC* time of the process within an interval of duration h_q , is derived from the following expression:

$$\tau_{IC/(0,0)}(h_q) = \int_0^{h_q} \exp\left(-(\nu_{x,0} + \nu_{s,0} + \lambda_{Fx|0} + \lambda_{Fs|0}) \cdot t\right) dt \quad (8)$$

Furthermore, in order to compute the *OOC* time when the process operates under the effect of state $(i, j) \neq (0, 0)$ at the beginning of an interval of duration h_q , three alternative scenarios should be considered:

- (a) The process remains under the effect of (i, j) until a failure occurs.
- (b) The process mean is further deteriorated from i to $y=i+1, \dots, m$.
- (c) The standard deviation of the process is deteriorated from j to $z=j+1, \dots, r$.

In cases (b) and (c), the process operates under the effect of initial state (i, j) for t time units and continues its *OOC* operation under an inferior state being the effect of the assignable cause that occurred.

Consequently, $\tau_{OOC/(i,j)}(h_q)$ can be derived from the following expression:

$$\begin{aligned} \tau_{OOC/(i,j)}(h_q) &= h_q \cdot \exp\left(-(\nu_{x,i} + \nu_{s,j} + \lambda_{Fx|i} + \lambda_{Fs|j}) \cdot h_q\right) + \\ &+ \int_0^{h_q} t \cdot (\lambda_{Fx|i} + \lambda_{Fs|j}) \cdot \exp\left(-(\nu_{x,i} + \nu_{s,j} + \lambda_{Fx|i} + \lambda_{Fs|j}) \cdot t\right) dt + \\ &+ \int_0^{h_q} \sum_{y=i+1}^m \left[\lambda_{x(i \rightarrow y)} \cdot \exp\left(-(\nu_{x,i} + \nu_{s,j} + \lambda_{Fx|i} + \lambda_{Fs|j}) \cdot t\right) \cdot (t + \tau_{OOC/(y,j)}(h_q - t)) \right] dt + \\ &+ \int_0^{h_q} \sum_{z=j+1}^r \left[\lambda_{s(j \rightarrow z)} \cdot \exp\left(-(\nu_{x,i} + \nu_{s,j} + \lambda_{Fx|i} + \lambda_{Fs|j}) \cdot t\right) \cdot (t + \tau_{OOC/(i,z)}(h_q - t)) \right] dt \end{aligned} \quad (9)$$

Obviously, the *OOC* time when the process starts *IC* is simplified to the following equation:

$$\begin{aligned} \tau_{OOC/(0,0)}(h_q) &= \int_0^{h_q} \sum_{y=1}^m \left[\lambda_{x(0 \rightarrow y)} \cdot \exp\left(-(\nu_{x,0} + \nu_{s,0} + \lambda_{Fx|0} + \lambda_{Fs|0}) \cdot t\right) \cdot \tau_{OOC/(y,0)}(h_q - t) \right] dt + \\ &+ \int_0^{h_q} \sum_{z=1}^r \left[\lambda_{s(0 \rightarrow z)} \cdot \exp\left(-(\nu_{x,0} + \nu_{s,0} + \lambda_{Fx|0} + \lambda_{Fs|0}) \cdot t\right) \cdot \tau_{OOC/(0,z)}(h_q - t) \right] dt \end{aligned} \quad (10)$$

The expected duration of a transition step *ET* equals the sum of the *IC* time, in case: the process starts *IC*; an alarm is issued; a failure has occurred, plus the *OOC* production time, plus the time delay in case of a false alarm, plus the

restoration times from an *OOO* state and a failure, all these multiplied by the long-term probability for each decision.

$$\begin{aligned}
 ET = & \pi_{(0,0)0} \cdot \tau_{IC/(0,0)}(h_1) + \pi_{(0,0)1} \cdot \tau_{IC/(0,0)}(h_2) + \sum_{k=0}^m \sum_{l=0}^r \pi_{(k,l)0} \cdot \tau_{OOO/(k,l)}(h_1) \\
 & + \sum_{k=0}^m \sum_{l=0}^r \pi_{(k,l)1} \cdot \tau_{OOO/(k,l)}(h_2) + \sum_{k=0}^m \sum_{l=0}^r \pi_{(k,l)2} \cdot q_{(k,l)}^{(0,0)} \cdot \tau_{IC/(0,0)}(h_1) + \\
 & + \sum_{k=0}^m \sum_{l=0}^r \sum_{u=0}^k \sum_{v=0}^l \pi_{(k,l)2} \cdot q_{(u,v)}^{(k,l)} \cdot \tau_{OOO/(u,v)}(h_1) + \pi_F \cdot \left(\tau_{IC/(0,0)}(h_1) + \tau_{OOO/(0,0)}(h_1) \right) + \\
 & + \pi_{(0,0)2} \cdot T_0 + \sum_{k=0}^m \sum_{l=0}^r \pi_{(k,l)2} \cdot T_{(k,l)} + \pi_F \cdot T_F
 \end{aligned} \tag{11}$$

4.3 Expected Cycle Cost

As regards the expected cost of a transition step, it can be computed as the weighted average of the expected cost of each of the possible states of the process being the initial state of an interval.

The components of the expected cycle cost are the following:

- (a) the sampling costs (*b* and *c*).
- (b) the expected *OOO* operation cost per time unit for state (i, j) being the initial state $K_{(i,j)}(h_q)$.
- (c) the cost of a false alarm L_0 .
- (d) the cost for a process restoration from an *OOO* state (k, l) (*PM* action) $L_{(k,l)}$.
- (e) the cost for a process restoration from a failure (*CM* action) L_F .

Additional clarification should be made for the expected *OOO* operation cost, when the process operates under the effect of state (i, j) at the beginning of an interval of duration h_q . In specific, $K_{(i,j)}(h_q)$ equals the sum of the following three terms:

(a) the cost per time unit if neither an assignable cause nor a failure occurs within the interval, multiplied by the probability of this scenario.

(b) the cost per time unit in case either only the mean or only the standard deviation of the process is affected but no failure occurs, multiplied by the duration that the process is under the effect of each assignable cause, times the probability of each possible scenario.

(c) the cost per time unit when both the mean and the standard deviation of the process are affected, and, so, the probability of all the possible scenarios for a

process transition from (i, j) to any state (y, z) , $y=i+1, \dots, m$ and $z=j+1, \dots, r$ and the expected occurrence time for each assignable cause should be considered.

Therefore, the expected OOC operation cost, for state (i, j) being the initial state of a sampling interval of duration h_q , can be derived from the following expression:

$$\begin{aligned}
K_{(i,j)}(h_q) = & M_{(i,j)} \cdot h_q \cdot \exp\left(-\left(v_{x,i} + v_{s,j} + \lambda_{F_{x|i}} + \lambda_{F_{s|j}}\right) \cdot h_q\right) + \\
& + \int_0^{h_q} \left(\lambda_{F_{x|i}} + \lambda_{F_{s|j}}\right) \cdot \exp\left(-\left(v_{x,i} + v_{s,j} + \lambda_{F_{x|i}} + \lambda_{F_{s|j}}\right) \cdot t\right) \cdot t \cdot M_{(i,j)} dt \\
& + \exp\left(-\left(v_{s,j} + \lambda_{F_{x|i}} + \lambda_{F_{s|j}}\right) \cdot h_q\right) \cdot \int_0^{h_q} \sum_{y=i+1}^m \left[\lambda_{x(i \rightarrow y)} \cdot \exp(-v_{x,i} \cdot t) \cdot \left(t \cdot M_{(i,j)} + K_{(y,j)}(h_q - t)\right) \right] dt + \\
& + \exp\left(-\left(v_{x,i} + \lambda_{F_{x|i}} + \lambda_{F_{s|j}}\right) \cdot h_q\right) \cdot \int_0^{h_q} \sum_{z=j+1}^r \left[\lambda_{s(j \rightarrow z)} \cdot \exp(-v_{s,j} \cdot t) \cdot \left(t \cdot M_{(i,j)} + K_{(i,z)}(h_q - t)\right) \right] dt + \\
& + \sum_{y=i+1}^m \sum_{z=j+1}^r \sum_{SC_{(i,j)}(y,z)} \Pr(h_q, (f_{n_1} + f_{n_2}))_{(i,j)} \left[\begin{aligned} & M_{(i,j)} \cdot \tau^{(1)}(h_q) + M^{(1)} \cdot \left(\tau^{(2)}(h_q) - \tau^{(1)}(h_q)\right) + \dots \\ & \dots + M_{(y,z)} \cdot \left(\tau_{OOC/(i,j)}(h_q) - \tau^{(f_{n_1} + f_{n_2})}(h_q)\right) \end{aligned} \right] \quad (12)
\end{aligned}$$

Consequently, the expected cycle cost can be computed from the following expression:

$$\begin{aligned}
EC = & b + \sum_{k=0}^m \sum_{l=0}^r \pi_{(k,l)0} \cdot \left(c \cdot n_1 + K_{(k,l)}(h_1)\right) + \sum_{k=0}^m \sum_{l=0}^r \pi_{(k,l)1} \cdot \left(c \cdot n_2 + K_{(k,l)}(h_2)\right) + \\
& + \sum_{k=0}^m \sum_{l=0}^r \sum_{u=0}^k \sum_{v=0}^l \pi_{(k,l)2} \cdot q_{(k,l)}(u,v) \cdot \left(c \cdot n_1 + K_{(u,v)}(h_1)\right) + \pi_{(0,0)2} \cdot L_0 + \\
& + \sum_{k=0}^m \sum_{l=0}^r \pi_{(k,l)2} \cdot L_{(k,l)} + \pi_F \cdot \left(c \cdot n_1 + K_{(0,0)}(h_1) + L_F\right) \quad (13)
\end{aligned}$$

6. OPTIMIZATION PROBLEM

In this paper we consider an optimization problem where the goal is to find the optimal $DP_q = \{n_q, h_q, w_{x,q}, w_{s,q}, k_{x,q}, k_{s,q}\}$ in order for the objective function, i.e. minimize ECT , to be optimized subject to some statistical constraints. The general optimization problem is formulated as follows:

$$\begin{aligned}
& \min_{DP_q} ECT \\
& \text{s.t.} \quad DP_q > 0 \\
& \quad h_2 \leq h_1, \quad n_2 \geq n_1 \\
& \quad w_{x,2} \leq w_{x,1}, k_{x,2} \leq k_{x,1}, \quad w_{s,2} \leq w_{s,1}, k_{s,2} \leq k_{s,1} \\
& \quad n_1, n_2 \in \mathbb{Z}^+ \\
& \quad ARL_0 \geq ARL_{\min}, \quad WARL \leq WARL_{\max}
\end{aligned} \tag{14}$$

It should be noted that the statistical constraints can be placed on any other statistical measure.

7. CONCLUSIONS

In this paper, a fully adaptive integrated maintenance and quality control scheme is developed. The proposed model can monitor processes subject to a multiple assignable cause mechanism which may affect both the process location and scale. Apart from the quality shifts that deteriorate the process performance, failures that seize the process operation may also occur. The detection of a quality shift from the control scheme triggers a *PM* action which is preferable to the *CM* action triggered in case of a failure.

The operation of the proposed control scheme is stochastically modeled by means of a three-dimensional Discrete Time Markov Chain. The values of the design parameters are selected so as to achieve economic optimization and assure acceptable statistical performance through specific statistical constraints.

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία προτείνεται ένα πλήρως δυναμικό σχήμα ελέγχου για την παρακολούθηση διαδικασιών, στις οποίες πέραν πολλαπλών συστηματικών αιτιών οι οποίες επηρεάζουν τη μέση τιμή αλλά και την τυπική απόκλιση του χαρακτηριστικού ποιότητας, είναι πιθανό να επιδράσουν και βλάβες. Κάθε συναγερμός του σχήματος ελέγχου ενεργοποιεί εργασίες προληπτικής συντήρησης, οι οποίες είναι προτιμητέες έναντι των εργασιών διορθωτικής συντήρησης που ενεργοποιούνται σε περιπτώσεις βλάβης. Μια ρεαλιστική υπόθεση του μοντέλου είναι ότι η προληπτική συντήρηση είναι ατελής, δηλαδή δεν επαναφέρει απαραίτητα τη διαδικασία στην επιθυμητή κατάσταση. Η ανάγκη για διασφάλιση αποδεκτής στατιστικής συμπεριφοράς του σχήματος ελέγχου συνδυαζόμενη με την απαίτηση για ελαχιστοποίηση του συνολικού κόστους ποιότητας, οδηγεί στη διαμόρφωση ενός προβλήματος οικονομο-στατιστικής βελτιστοποίησης.

REFERENCES

Ben-Daya, M. and Duffuaa, S. O. (1995). Maintenance and quality: the missing link. *Journal of Quality in Maintenance Engineering*, **1**, 20–26.

- Caballero Morales, S. O. (2013). Economic Statistical Design of Integrated Xbar-S Control Chart with Preventive Maintenance and General Failure Distribution. *PLoS One* 8 (3): e59039 doi:10.1371/journal.pone.0059039
- Chiu, H.-N., Huang, B.-S. (1995). The economic design of $\bar{x}$ and S^2 control charts with preventive maintenance and increasing hazard rate. *Journal of Quality in Maintenance Engineering*, **1**, 17-40.
- Derman, C. (1962). On sequential decisions and Markov chains. *Management Science*, **9**, 16–24.
- Derman, C. (1963). On optimal replacement rules when changes of state are Markovian. Bellman R, editor. Mathematical optimization techniques, vol. R-396-PR. Berkeley: University of California Press; p. 201–10 chapter 9.
- Duffuaa, S. O. and Ben-Daya, M. (1995). Improving maintenance quality using SPC tools. *Journal of Quality in Maintenance Engineering*, **1**, 25-33.
- Ivy, J. S. and Nembhard, H. B. (2005). A modeling approach to maintenance decisions using statistical quality control and optimization. *Quality and Reliability Engineering International*, **21**, 355–366.
- Linderman, K., McKone-Sweet, K. E. and Anderson, J. C. (2005). An integrated systems approach to process control and maintenance. *European Journal of Operational Research*, **164**, 324–340.
- Panagiotidou, S. and Tagaras, G. (2010). Statistical Process Control and Condition-based Maintenance: A Meaningful Relationship through Data Sharing. *Production and Operations Management*, **19**, 156–171.
- Paté-Cornell, M. E., Lee, H. L. and Tagaras, G. (1987). Warnings of Malfunction: the Decision to Inspect and Maintain Production Process on Schedule or on Demand. *Management Science*, **33**, 1277-1290.
- Rahim, M. A. and Banerjee, P. K. (1993). A generalized model for the economic design of $\bar{x}$ control charts for production systems with increasing failure rate and early replacement. *Naval Research Logistics*, **40**, 789–809.
- Tagaras, G. (1988). An integrated cost model for the joint optimization of process control and maintenance. *Journal of the Operational Research Society*, **39**, 757–766.
- Yin, H., Zhang, G., Zhu, H., Deng, Y. and He, F. (2015). An integrated model of statistical process control and maintenance based on the delayed monitoring. *Reliability Engineering and System Safety*, **133**, 323-333.



ΕΥΡΕΤΗΡΙΟ ΣΥΓΓΡΑΦΕΩΝ (INDEX)

BASSIAKOS Y.	235	NENES G.	360
BATZIOU E.	235	NTZOUFRAS I.	345
CHAKRABORTI S.	153	PAPADIMITRIOU E.	283, 307, 335
CHARALAMBOUS H.	271	PARAGKAMIAN S.	245
CHASIOTIS V.	260	TASIAS K.A.	360
CHATZOPOULOS ST.	271	TONIDIS E.	245
CHOROZOGLOU D.	283	TSAKLIDIS G.	307, 322
DARAS P.	322	VASILADIS E.	335
FARMAKIS N.	260	VRETOS N.	322
GRAVANIS M.	245	XANTHOPOULOU G.	245
KARADIMOS P.	245	ΑΓΙΑΚΛΟΓΛΟΥ Χ.	183
KARAKOSTAS V	307, 360	ΑΘΑΝΑΣΙΑΔΗΣ Ι.	21
KATERI M.	345	ΑΡΑΠΗΣ Α.Ν.	35
KOLYVA-MACHERA F.	271	ΓΚΟΥΡΑΜΑΝΗ Ε.	219
KOULIKAS D.	245	ΔΗΜΗΤΡΙΟΥ Ι.	43, 57, 67
KOUNIAS S.	260, 299	ΖΗΣΙΜΟΠΟΥΛΟΥ Ε.-Α.	43
KOUROUKLAS CH.	307	ΘΕΟΔΟΣΙΑΔΟΥ Ο.	76,
KUGIUMTZIS D.	283	ΙΩΑΝΝΙΔΗΣ Δ.	21
LOUMPNIAS K.	322	ΚΑΛΤΣΑΣ Α.-Φ.	57
MANGIRA O.	335	KETZAKH E.	91
MANTZOUNI A.	345	ΚΟΥΝΙΑΣ Σ.	101,
MESIMERI M.	360	ΜΑΚΡΗ Φ.Σ.	35
MPATZIAKAS A.	245	ΜΑΚΡΟΥ Μ.	67

ΜΠΕΡΣΙΜΗΣ Φ.Γ.	153
ΜΠΟΖΙΚΑΣ Α.	108
ΠΑΠΑΔΟΠΟΥΛΟΥ Α.	168
ΠΑΠΑΘΑΝΑΣΙΟΥ Ε.-Β.	138
ΠΑΠΑΪΩΑΝΝΟΥ Τ.	123
ΠΑΠΑΤΣΟΥΜΑ Ι.	194
ΠΙΤΣΕΛΗΣ Γ.	108
ΠΡΟΔΡΟΜΙΔΗΣ Π.-Ι.	138
ΡΑΛΛΑΚΗΣ Σ.	91
ΣΑΧΛΑΣ Α.	153
ΣΥΜΕΩΝΙΔΟΥ Α.	168
ΣΩΤΗΡΟΓΛΟΥ Μ.	219
ΤΣΑΚΛΙΔΗΣ Γ.	76
ΤΣΙΜΠΑΝΟΣ Α.	183
ΤΣΙΜΠΟΣ Κ.	183
ΦΑΡΜΑΚΗΣ Ν.	194
ΧΑΡΑΛΑΜΠΙΔΗΣ Χ.	203
ΧΑΡΙΤΙΔΟΥ Ο	210
ΧΑΤΖΗΠΑΝΤΕΛΗΣ Θ.	219
ΨΥΛΛΑΚΗΣ Ζ.Μ.	35