

**ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ**



**ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ  
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ**

**ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ  
ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ  
ΣΤΑΤΙΣΤΙΚΗ**

**ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΣΕΙΣΜΙΚΩΝ  
ΔΕΔΟΜΕΝΩΝ ΜΕ ΤΗ ΧΡΗΣΗ ΤΗΣ  
ΘΕΩΡΙΑΣ ΑΚΡΑΙΩΝ ΤΙΜΩΝ**

Παναγιώτης Κολοκυθάς

Διπλωματική Εργασία

Που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς  
Μάρτιος 2009

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίστηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. .... συνεδρίαση του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της επιτροπής ήταν :

-.....(Επιβλέπων)

-.....

-.....

Η έγκριση της Διπλωματικής Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνώμων του συγγραφέα.

UNIVERSITY OF PIRAEUS



DEPARTMENT OF STATISTICS  
AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN  
APPLIED STATISTICS

**STATISTICAL ANALYSIS OF EARTH-  
QUAKE DATA USING EXTREME VAL-  
UE THEORY**

By

Panagiotis Kolokithas

MSc Dissertation

Submitted to the Department of Statistics and Insurance  
Science of the University of Piraeus in partial fulfillment  
of the requirements for the degree of Master of Science in  
Applied Statistics

Piraeus, Greece

March 2009

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

*Σε κάποιον...*

## ΠΕΡΙΛΗΨΗ

Κύριος στόχος της εργασίας αυτής είναι η παρουσίαση στατιστικών μεθόδων από την θεωρία ακραίων τιμών, κατάλληλων για την στατιστική ανάλυση σεισμικών δεδομένων. Με την εφαρμογή αυτών των μεθόδων σε πραγματικά δεδομένα, σκοπός μας είναι η εξαγωγή χρήσιμων συμπερασμάτων.

Αρχικά παραθέτονται τα βασικά σημεία της θεωρίας και αναλύονται οι μέθοδοι Block Maxima και Peaks Over Threshold (POT).

Στα πλαίσια της εργασίας παρουσιάζονται δύο μέθοδοι για την εκτίμηση των παραμέτρων της γενικευμένης κατανομής ακροτάτων (Generalized Extreme Value Distribution, (GEV)) όταν διαχειριζόμαστε λογοκριμένα δείγματα. Η μέθοδος της μέγιστης πιθανοφάνειας και αυτή των μερικών πιθανοτικά σταθμισμένων ροπών.

Στη συνέχεια παρουσιάζεται ο τρίτος τύπος της κατανομής *Gumbel* (GIII), ο οποίος χρησιμοποιείται σε πλήθος εργασιών για την στατιστική ανάλυση σεισμικών δεδομένων σύμφωνα με την μέθοδο Block Maxima. Ακολούθως γίνεται μια εμπειρική αξιολόγηση της μεθόδου για λογοκριμένες παρατηρήσεις.

Οι στατιστικές αναλύσεις που διεξήχθησαν αφορούν δύο σύνολα δεδομένων. Το πρώτο περιέχει δεδομένα από όλη την Ελλάδα για την περίοδο 1970 – 2007 και το δεύτερο σύνολο περιέχει δεδομένα από όλη την Ελλάδα για την περίοδο 1901 – 1996.

Στον πρώτο κατάλογο σεισμών εφαρμόστηκαν οι εξής μέθοδοι σε 27 περιοχές της Ελλάδας (35 με την μέθοδο 3)

1. Διμηνιαία ή τριμηνιαία block maxima με πλήρη δεδομένα
2. POT
3. Μηνιαία block maxima με λογοκριμένα δεδομένα

Ενώ στον δεύτερο διεξήχθησαν οι ίδιες αναλύσεις σε όλο τον Ελλαδικό χώρο, με μόνη διαφορά ότι στην πρώτη μέθοδο πήραμε ετήσια block maxima.

Σε κάθε περίπτωση με την βοήθεια του λογισμικού πακέτου R-project βρέθηκαν οι εκτιμήσεις των παραμέτρων των οριακών κατανομών. Εκτιμήθηκε η στάθμη απόδοσης από 10 έως 100 χρόνια απόδοσης και υπολογίστηκε η διασπορά της σύμφωνα με την μέθοδο Δέλτα. Επίσης έγινε σύγκριση των μεθόδων μετά την εφαρμογή τους που βασίστηκε και σε κοινά γραφήματα της στάθμης απόδοσης για κάθε περιοχή. Τέλος παρουσιάζονται ποσοτικά συμπεράσματα για την σεισμική δραστηριότητα στον γεωγραφικό χώρο της Ελλάδας.

## Abstract

The main target of this dissertation is to present and apply several statistical methods from Extreme Value Theory for the analysis of seismic data collected in the last decades from the broader area of Greece.

Initially, we present the two main methods for studying and analyzing extremes, namely the *Block Maxima* and *Peaks Over Threshold* (POT) methods.

According to the first method, the block maxima will approximately follow a *Generalized Extreme Value* (GEV) distribution. We use two methods for the estimation of the parameters of a GEV distribution that can also be applied to censored seismic samples: the method of *Maximum Likelihood Estimation* (MLE) and the method of *Partial Probability Weighted Moments* (PPWM).

We also discuss the case when the block maxima asymptotically follow Gumbel's third type (GIII or reversed Weibull) distribution, which can be considered as a special case of the GEV distribution. This distribution has been used by many authors for modeling seismic data. We empirically study the behavior of the MLE's of GIII parameters when, in addition, we have a censored sample.

Two data sets are being used for the application of the theory. The first set contains earthquakes which occurred in Greece in the period 1970-2007 (62053 observations) and the second set contains earthquakes which occurred in Greece in the period 1901-1996 (5991 observations).

Using the first data set we employ three different statistical methods for analyzing the seismic activity in 27 subregions of Greece (we have 35 subregions in the third analysis). More specifically we use:

- (a) Bimonthly or trimester block maxima (leading to full maxima samples)
- (b) Peaks over an appropriately chosen threshold (POT method)
- (c) Monthly block maxima (leading to censored maxima samples)

With the second data set we employ the same three methods, now for the whole area of Greece. Now we consider yearly block maxima instead of bimonthly or trimester maxima in (a).

For all necessary computations and numeric estimations we use the software package R ([www.R-project.org](http://www.R-project.org)). More specifically, we estimate the parameters of the asymptotic distributions (GIII or Generalized Pareto for the Block maxima and the POT method respectively), the return level of 10 to 100 years period and compute the

variance of the estimates in every case (mainly exploiting Delta method). Finally we perform a comparison between the statistical methods (a), (b), (c) and we discuss some general conclusions for Greece's seismicity.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

## Περιεχόμενα

Κατάλογος Πινάκων	xv
Κατάλογος Σχημάτων	xvii
Κατάλογος Συντομογραφιών	xix
<b>1. Εισαγωγή</b>	13
1.1 Ιστορική αναδρομή	13
1.2 Το πρόβλημα	14
1.3 Στατιστική και σεισμολογικά δεδομένα	15
<b>2. Θεωρία ακραίων τιμών – Στατιστική ανάλυση ακραίων παρατηρήσεων</b>	17
2.1 Θεωρία ακραίων τιμών	17
2.1.1 Εισαγωγή στην θεωρία ακραίων τιμών	17
2.1.2 Περιοχές έλξης	20
2.2 Στατιστική ανάλυση ακραίων παρατηρήσεων	23
2.2.1 Η μέθοδος Block Maxima	23
2.2.2 Εκτιμήσεις παραμέτρων της <i>GEV</i>	24
2.2.3 Έλεγχος καλής προσαρμογής	27
2.2.4 Στάθμη απόδοσης για $1/p$ χρονικές περιόδους απόδοσης	27
2.2.5 Η μέθοδος Peaks Over Threshold (POT)	30
2.2.6 Εκτίμηση των παραμέτρων της GPD και έλεγχος καλής προσαρμογής	33
2.2.7 Στάθμη απόδοσης για $m$ παρατηρήσεις	35
2.3 Λογοκριμένα δεδομένα	37
2.3.1 Μέθοδος Μέγιστης Πιθανοφάνειας	38
2.3.2 Μέθοδος Μερικών Πιθανοτικά Σταθμισμένων Ροπών	38
<b>3. Στατιστική ανάλυση σεισμικών δεδομένων – Εμπειρική αξιολόγηση της μεθόδου Block Maxima με λογοκριμένες παρατηρήσεις</b>	41
3.1 Ο τρίτος τύπος της οριακής κατανομής του <i>Gumbel</i>	41
3.2 Εμπειρική αξιολόγηση της μεθόδου Block Maxima με λογοκριμένες παρατηρήσεις	46
<b>4. Εφαρμογή της θεωρίας ακραίων τιμών σε σεισμολογικά δεδομένα της Ελλάδας</b>	53
4.1 Σεισμολογικά δεδομένα περιόδου 1970 – 2007	53



4.1.1 Μέθοδος Α – Πλήρη δεδομένα & <i>Gumbel</i> III	54
4.1.2 Μέθοδος Β – Peaks Over Threshold	54
4.1.3 Μέθοδος Γ – Λογοκριμένα δεδομένα & <i>Gumbel</i> III	71
4.2 Σεισμολογικά δεδομένα περιόδου 1901 – 1996	75
4.2.1 Μέθοδος Block Maxima με ετήσια τοπικά μέγιστα	76
4.2.2 Μέθοδος Peaks Over Threshold	77
4.2.3 Μέθοδος Block Maxima με μηνιαία τοπικά μέγιστα	79
<b>5. Συμπεράσματα</b>	82
5.1 Ποιοτικά συμπεράσματα	82
5.2 Ποσοτικά συμπεράσματα	91
<b>Παράρτημα</b>	96
<b>Βιβλιογραφία</b>	103

## Κατάλογος Πινάκων

3-1 Μέση απόλυτη Διαφορά για τις παραμέτρους $\omega$ και $\lambda$ .	52
4-1 Οι 27 περιοχές της Ελλάδας με τα κέντρα τους και η χρονική περίοδος που πάρθηκαν τα τοπικά μέγιστα.	55
4-2 Εκτιμήσεις παραμέτρων GIII, στάθμης απόδοσης για 10 και 50 χρόνια απόδοσης και τυπικά σφάλματα.	61
4-3 Το κατόφλι $u$ , οι εκτιμήσεις των παραμέτρων της GPD, της στάθμης απόδοσης για 10 και 50 χρόνια απόδοσης και τα τυπικά σφάλματα.	70
4-4 Οι 8 επιπλέον περιοχές για την Μέθοδο Γ.	71
4-5 Εκτιμήσεις παραμέτρων GIII, στάθμης απόδοσης για 10 και 50 χρόνια απόδοσης, τυπικά σφάλματα και πλήθος λογοκριμένων παρατηρήσεων.	74
4-6 Εκτιμήσεις παραμέτρων GIII και τυπικά σφάλματα με ετήσια block maxima.	76
4-7 Εκτιμήσεις παραμέτρων GIII και τυπικά σφάλματα με μηνιαία block maxima.	81
5-1 Εύρος 95% Διαστήματος Εμπιστοσύνης της στάθμης απόδοσης για 50 χρόνια απόδοσης με τις τρεις μεθόδους.	90

## Κατάλογος Σχημάτων

3-1 Εκτιμήσεις των παραμέτρων $\omega, \lambda$ για μέγεθος block : 1, 2, 5, 10,25, 50, 100	49
3-2 Εκτιμήσεις των παραμέτρων $\omega, \lambda$ με αρχικές συνθήκες τις πραγματικές τιμές, για μέγεθος block ίσο με 2	51
4-1 Τα κέντρα των 27 περιοχών	56
4-2 Q-Q plots της GIII για την Μέθοδο A	62
4-3 Mean Excess Plot, Mean Residual Live Plot, Q-Q Plot, διάστημα εμπιστοσύνης για την παράμετρο $\xi$ με την Profile Deviance Function, για τις 27 περιοχές της Ελλάδας	65
4-4 Τα κέντρα των 35 περιοχών	72
4-5 Διαγραμμα Gutenberg - Richter για το πρώτο σύνολο δεδομένων	73
4-6 Q-Q plots της GIII για την Μέθοδο $\Gamma$ στις 8 επιπλέον περιοχές	75
4-7 Q-Q Plot της GIII με ετήσια block maxima	76
4-8 Διάγραμμα στάθμης απόδοσης και 95% δ.ε. με ετήσια block maxima	77
4-9 Εμπειρική μέση υπερβάλλουσα συνάρτηση	78
4-10 Q-Q plot για την GPD	78
4-11 Διάγραμμα στάθμης απόδοσης για 10 ως 100 χρόνια απόδοσης και 95% δ.ε.	79
4-12 Διαγραμμα Gureberg - Richter για το δεύτερο σύνολο δεδομένων	80
4-13 Q-Q Plot της GIII με μηνιαία block maxima	80
4-14 Διάγραμμα στάθμης απόδοσης για 10 ως 100 χρόνια απόδοσης και 95% δ.ε.	81
5-1 Κοινά διαγράμματα της στάθμης απόδοσης για 10 έως 100 χρόνια απόδοσης για τις τρεις αναλύσεις A, B, $\Gamma$ .	84
5-2 Εκτιμήσεις στάθμης απόδοσης για 50 χρόνια απόδοσης με GIII και πλήρη δεδομένα.	92
5-3 Contour plot εκτιμήσεων στάθμης απόδοσης για 50 χρόνια απόδοσης με GIII και πλήρη δεδομένα.	92
5-4 Εκτιμήσεις στάθμης απόδοσης για 50 χρόνια απόδοσης με GPD.	93
5-5 Contour plot εκτιμήσεων στάθμης απόδοσης για 50 χρόνια απόδοσης με GPD	93
5-6 Εκτιμήσεις στάθμης απόδοσης για 50 χρόνια απόδοσης με GIII και λογοκριμένα δεδομένα.	94
5-7 Contour plot εκτιμήσεων στάθμης απόδοσης για 50 χρόνια απόδοσης με GIII και λογοκριμένα δεδομένα.	94

## Κατάλογος Συντομογραφιών

τ.μ.	τυχαία μεταβλητή
σ.π.π.	συνάρτηση πυκνότητας πιθανότητας
iid	independent, identically distributed (ισόνομες τυχαίες μεταβλητές)
δ.ε.	διάστημα εμπιστοσύνης
GEV	Generalized Extreme Value distribution (γενικευμένη κατανομή ακροτάτων)
POT	Peaks Over Threshold (υπερβάσεις κατωφλίου)
GPD	Generalized Pareto Distribution (γενικευμένη κατανομή Pareto)
PPWM	Partial Probability Weighted Moments (Μερικές Πιθανοτικά Σταθμισμένες Ροπές)
GumbelIII-GIII	Gumbel third type asymptotic distribution ( τρίτος τύπος της οριακής κατανομής του <i>Gumbel</i> )

# ΚΕΦΑΛΑΙΟ 1

## Εισαγωγή

Μορφοποιήθηκε

### 1.1 Ιστορική αναδρομή

Η θεωρία ακραίων τιμών ασχολείται με την στοχαστική συμπεριφορά του μέγιστου και του ελάχιστου από κάποιες τυχαίες μεταβλητές  $X_1, X_2, \dots, X_n$  οι οποίες συνήθως είναι ανεξάρτητες και ισόνομες. Οι ιδιότητες των κατανομών ακροτάτων και η υπέρβαση πάνω (κάτω) από κάποιο υψηλό (χαμηλό) κατώφλι καθορίζονται από την δεξιά και αριστερή ουρά της κατανομής των τυχαίων μεταβλητών  $X_1, X_2, \dots, X_n$ .

Ιστορικά, ο πρώτος που ασχολήθηκε με πρόβλημα ακραίων τιμών ήταν ο *Nicolas Bernouli* το 1709 όπου αναφέρθηκε στη μέση μέγιστη απόσταση μεταξύ της αρχής και  $n$  σημείων τυχαία διασκορπισμένων πάνω σε ένα ευθύγραμμο τμήμα με μήκος  $t$ . Η συστηματική ανάπτυξη της θεωρίας ουσιαστικά ξεκίνησε το 1922 όπου σε μια εργασία του ο *Von Bortkiewicz* ανέφερε για πρώτη φορά την κατανομή ακραίων τιμών. Οι *Fisher Tippet* (1928) απέδειξαν ότι η οριακή κατανομή ακροτάτων μπορεί να είναι μόνο μία από τρεις συγκεκριμένους τύπους κατανομών, επεκτείνοντας την έρευνα του *Frechet* το 1927 που είχε προσδιορίσει μια πιθανή οριακή κατανομή. Στη συνέχεια, ο *Von Mises* παρουσίασε κάποιες απλές αλλά και χρήσιμες συνθήκες για την ασθενή σύγκλιση διατεταγμένων τυχαίων μεταβλητών για κάθε έναν από τους τρεις τύπους των *Fisher Tippet* και ο *Gnedenko* (1943) έδωσε μια αυστηρή περιγραφή των συνθηκών.

Η θεωρητική ανάπτυξη που συντελέστηκε στην δεκαετία του 20 και μέχρι τα μέσα του 30 συνεχίστηκε με πολλές εφαρμογές, από τα τέλη της δεκαετίας του 30 και σε όλη την δεκαετία του 40. Οι εφαρμογές αυτές αφορούσαν αρκετά θέματα όπως την εκπομπή ραδιενέργειας (*Gumbel* (1937)), αντοχή υλικών (*Weibull* (1939)), πλημμύρες (*Gumbel* (1941)), ανάλυση σεισμών (*Nordquist* (1945)), βροχοπτώσεις κ.α. Όσον αφορά τις εφαρμογές της θεωρίας, ο *Gumbel* συνέβαλε σημαντικά στην στατιστική ανάλυση των ακραίων τιμών με τη μονογραφία του, *Statistics of Extremes* που εκδόθηκε το 1958. Από τότε μέχρι σήμερα έχουν δημοσιευθεί χιλιάδες εργασίες σχετικές με τη θεωρία ακραίων τιμών και των εφαρμογών της, κάτι που αποδεικνύει την χρησιμότητα και αναγκαιότητα της.

## 1.2 Το πρόβλημα

Παρακάτω θα αναλύσουμε εν συντομία γιατί υπήρξε αναγκαία η ανάπτυξη της συγκεκριμένης θεωρίας και ποιο είναι το πρόβλημα που μπορεί να αντιμετωπίσει καλύτερα από τις κλασσικές στατιστικές μεθόδους εκτίμησης και ελέγχων υποθέσεων.

Έστω ότι έχουμε ανεξάρτητες τυχαίες παρατηρήσεις  $X_1, X_2, \dots, X_n$  και υποθέτουμε ότι ακολουθούν μια κατανομή  $F_\theta$  όπου  $\theta$  είναι οι άγνωστες παράμετροι της κατανομής. Αν θελήσουμε να απαντήσουμε στο ερώτημα: *ποιο είναι το κατώφλι  $c_p$  το οποίο δεν θα υπερβεί καμία παρατήρηση με πιθανότητα  $1 - p$* , η κλασσική αντιμετώπιση είναι η εξής:

$$\begin{aligned} P(\max\{X_1, \dots, X_n\} < c_p) &= 1 - p \Leftrightarrow P(X_1 < c_p, \dots, X_n < c_p) = 1 - p \\ &\Leftrightarrow P(X_1 < c_p) \dots P(X_n < c_p) = 1 - p \\ &\Leftrightarrow F_\theta(c_p)^n = 1 - p \\ &\Leftrightarrow c_p = F_\theta^{-1}((1 - p)^{1/n}) \end{aligned}$$

και αφού εκτιμήσουμε το  $\theta$ , βρίσκουμε ως εκτίμηση του  $c_p$  την

$$\hat{c}_p = F_\theta^{-1}((1 - p)^{1/n}).$$

Τα προβλήματα που ανακύπτουν με την παραπάνω μεθοδολογία είναι :

- 1) Η υπόθεση ότι τα  $X_i$  ακολουθούν την  $F_\theta$  είναι συνήθως αυθαίρετη και δεν είμαστε σίγουροι ότι η δεξιά ουρά της παρουσιάζει την ίδια συμπεριφορά με την πραγματική ουρά των δεδομένων.
- 2) Οι εκτιμήσεις των παραμέτρων  $\theta$  έχουν γίνει με βάση το σύνολο των παρατηρήσεων και όχι με βάση τις ακραίες παρατηρήσεις που περιγράφουν πολύ καλύτερα την μορφή της δεξιάς ουράς της πραγματικής κατανομής των δεδομένων.
- 3) Έστω και αν δεχθούμε ότι υποθέσαμε σωστά για την  $F_\theta$ , μικρές αποκλίσεις στην εκτίμηση του  $\theta$  μπορούν να προξενήσουν μεγάλες αποκλίσεις στις εκτιμήσεις που γίνονται «κοντά» στο δεξί όριο της  $F_\theta$ .

Όλα τα παραπάνω προβλήματα αντιμετωπίζονται με την θεωρία των ακραίων τιμών αφού με αυτή μελετάμε και επικεντρωνόμαστε στην δεξιά ή αριστερή ουρά κάποιας κατανομής. Επίσης μια πρακτική δυσκολία που αντιπαρερχόμαστε είναι ότι

όταν μελετάμε ακραίες τιμές (μέγιστα ή ελάχιστα) μπορεί να έχουμε στην διάθεση μας μόνο αυτές και όχι το σύνολο των παρατηρήσεων του φαινομένου κάτι που κάνει πολύ δύσκολο να εκτιμήσουμε με ακρίβεια ολόκληρη την κατανομή  $F_{\theta}$ .

### 1.3 Στατιστική και σεισμολογικά δεδομένα

Παρόλη την εξέλιξη της επιστήμης της σεισμολογίας, δεν έχει περιγραφεί ακόμα επακριβώς πως λειτουργεί ο μηχανισμός της φύσης που ευθύνεται για το φαινόμενο εμφάνισης σεισμικών δονήσεων στην επιφάνεια της γης. Είναι λοιπόν αδύνατον με τα σημερινά δεδομένα οι σεισμολόγοι να κάνουν μια ασφαλή και ακριβή πρόβλεψη για την χρονική στιγμή, την τοποθεσία και το μέγεθος του επόμενου σεισμού. Το φυσικό αυτό φαινόμενο μπορεί συνεπώς να θεωρηθεί ως ένα τυχαίο φαινόμενο και για αυτόν τον λόγο οι στατιστικοί, σε συνεργασία με τους σεισμολόγους, προσπαθούν να το ερμηνέψουν και να εξάγουν συναφή στατιστικά συμπεράσματα. Τα αρχικά μοντέλα που προτάθηκαν βασίζονταν στην κατανομή Poisson καθώς και σε Μαρκοβιανά μοντέλα για εξαρτημένα γεγονότα. Ο πιο απλός και γνωστός εμπειρικός κανόνας που συνδυάζει το μέγεθος ενός σεισμού με τη συχνότητα εμφάνισής του δόθηκε από τους Gutenberg και Richter (1954) και περιγράφεται από τη σχέση

$$\log N_c(M) \approx a - bM,$$

όπου  $N_c$  είναι ο ετήσιος αριθμός σεισμών που έχουν μέγεθος μεγαλύτερο ή ίσο από  $M$ , και  $a, b$  κάποιες θετικές σταθερές. Προφανώς για να την χρησιμοποιήσουμε απαιτείται η γνώση όλων των σεισμών πάνω από ένα συγκεκριμένο μέγεθος. Τα μειονεκτήματα αυτού του μοντέλου είναι η έλλειψη ακρίβειας, ομοιογένειας καθώς και η μη πληρότητα των δεδομένων όσο αφορά τους μικρούς σε μέγεθος σεισμούς. Τα προβλήματα αυτά ανακύπτουν από τον τρόπο που γίνεται η καταγραφή των σεισμών από τα διαφορετικά σεισμολογικά κέντρα ανά τον κόσμο.

Η μεγάλη αλλαγή στον τρόπο της στατιστικής ανάλυσης των σεισμολογικών δεδομένων ήρθε με την εδραίωση της θεωρίας ακραίων τιμών. Βάση αυτής, αντιμετωπίστηκε το βασικότερο πρόβλημα που ισχύει για όλα τα φυσικά φαινόμενα, ότι δεν μπορούμε να έχουμε έναν πλήρη κατάλογο των γεγονότων. Οι εκτιμήσεις μέσω της θεωρίας ακραίων τιμών βασίζονται μόνο στις ακραίες παρατηρήσεις, όπου στην περίπτωση των σεισμών, είναι οι μεγάλοι σε ένταση σεισμοί. Αυτοί μπορούν να καταγραφούν από οποιοδήποτε σεισμολογικό κέντρο στον κόσμο, για οποιαδήποτε περιοχή του κόσμου και με μεγάλη ακρίβεια, σε αντίθεση με τους μικρούς σε ένταση σει-

σμούς. Το ερώτημα που μένει να απαντηθεί είναι αν τα σεισμολογικά δεδομένα καλύπτουν τις δύο βασικές συνθήκες της θεωρίας ώστε να μπορούμε να την χρησιμοποιήσουμε. Δηλαδή αν οι παρατηρήσεις είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές. Παίρνοντας υπόψη μας μόνο τους μεγαλύτερους σεισμούς ανά μια χρονική περίοδο έστω και αν αυτή είναι σχετικά μικρή π.χ. ανά τρίμηνο (πόσο μάλλον αν χρησιμοποιούμε ετήσια μέγιστα) εξαλείφουμε την πιθανότητα να έχουμε εξαρτημένες παρατηρήσεις, αφού οι προσεισμοί και μετασεισμοί ενός κύριου μεγάλου σεισμού θα είναι σίγουρα μικρότερου μεγέθους αλλά και θα συμβαίνουν χρονικά “κοντά” στον κύριο σεισμό. Επίσης μπορούμε να διασφαλίσουμε ότι είναι και ισόνομες τ.μ. αν εξετάσουμε τα αίτια γένεσης ενός σεισμού. Ανάλογα με τα αίτια, οι σεισμοί διακρίνονται σε τεκτονικούς, ηφαιστειογενείς και εγκατακρημνισιγενείς και ανάλογα με το εστιακό τους βάθος σε επιφανειακούς ( $\leq 60$  km), ενδιάμεσου βάθους ( $60\text{km} < \leq 300\text{km}$ ) και μεγάλου βάθους ( $\geq 300\text{km}$ ). Το 90% των επιφανειακών είναι τεκτονικοί, το 7% ηφαιστειογενείς και το 3% εγκατακρημνισιγενείς, όπου οι τελευταίοι είναι μικρής έντασης. Επομένως αν μελετήσουμε επιφανειακούς σεισμούς και γνωρίζοντας ότι δεν οφείλονται σε ηφαιστειακή έκρηξη δεχόμαστε ότι ο μηχανισμός που τους γεννά είναι μοναδικός και άρα μιλάμε για ισόνομες τυχαίες μεταβλητές.



## ΚΕΦΑΛΑΙΟ 2

### Θεωρία ακραίων τιμών – Στατιστική ανάλυση ακραίων παρατηρήσεων

Μορφοποιήθηκε

#### 2.1 Θεωρία ακραίων τιμών

##### 2.1.1 Εισαγωγή στη θεωρία ακραίων τιμών

Όπως θα δούμε στη συνέχεια, χρησιμοποιώντας την θεωρία ακραίων τιμών μπορεί να βρεθεί μια οριακή κατανομή για τα δειγματικά μέγιστα ανεξάρτητων και ισόνομων τυχαίων μεταβλητών. Ας αναφέρουμε όμως πρώτα κάποιες χαρακτηριστικές ιδιότητες των δειγματικών μεγίστων. Έστω  $F$  η κοινή συνάρτηση κατανομής των ανεξάρτητων και ισόνομων τυχαίων μεταβλητών (iid)  $X_1, X_2, \dots, X_n$  και  $x_F$  το δεξί άκρο του στηρίγματος της  $F$ , δηλαδή

$$x_F = \sup\{x \in R : F(x) < 1\} \leq \infty.$$

Τότε ισχύει,

$$M_n = \max\{X_1, X_2, \dots, X_n\} \rightarrow x_F \text{ με πιθανότητα } 1$$

(επειδή η ακολουθία των τ.μ.  $M_n$  είναι αύξουσα) και

$$P(M_n \leq x) = P(X_1 \leq x, X_2 \leq x, \dots, X_n \leq x) = \begin{cases} F(x)^n \rightarrow 0, & x < x_F \\ F(x)^n \rightarrow 1, & x \geq x_F \end{cases}$$

όταν το πλήθος των τ.μ. τείνει στο άπειρο δηλαδή όταν  $n \rightarrow \infty$ .

Για να βρεθεί μία μη εκφυλισμένη οριακή κατανομή που ακολουθούν τα  $M_n$  είναι αναγκαία μια «τυποποίηση» ή «κανονικοποίηση» ανάλογη με αυτήν που πραγματοποιείται στο Κεντρικό Οριακό Θεώρημα. Αναζητούμε λοιπόν ακολουθίες πραγματικών αριθμών  $c_n > 0$  και  $d_n$  τέτοιες ώστε η ποσότητα

$$\frac{M_n - d_n}{c_n}$$

να έχει μια μη εκφυλισμένη οριακή κατανομή  $G$  για  $n \rightarrow \infty$ , δηλαδή, για κάθε  $x$ ,

$$P\left(\frac{M_n - d_n}{c_n} \leq x\right) = P(\max\{X_1, X_2, \dots, X_n\} \leq c_n x + d_n) = F^n(c_n x + d_n) \xrightarrow{n \rightarrow \infty} G(x)$$

Είναι σχετικά εύκολο να διαπιστώσουμε (π.χ. μέσω αντιπαραδειγμάτων) ότι δεν μπορούμε πάντοτε να κανονικοποιήσουμε το δειγματικό μέγιστο, δηλαδή να βρούμε κατάλληλα  $c_n$  και  $d_n$ , έτσι ώστε να ισχύει η παραπάνω σύγκλιση. Όταν όμως υπάρχουν τέτοιες ακολουθίες, έχουν αποδειχθεί οι ιδιότητες και η μορφή των μη εκφυλισμένων κατανομών  $G$  που εμφανίζονται ως όριο κανονικοποιημένων μεγίστων. Η βασική τους ιδιότητα είναι ότι είναι max-stable, που σημαίνει ότι κανονικοποιημένα μέγιστα ανεξ. τ.μ. από την  $G$  θα ακολουθούν και πάλι την  $G$ . Πιο συγκεκριμένα μια κατανομή  $G$  θεωρείται max-stable, αν για  $Y, Y_1, Y_2, \dots, Y_n$  iid που ακολουθούν την  $G$ , υπάρχουν  $c_n > 0$ ,  $d_n \in \mathbb{R}$  τέτοια ώστε να ισχύει

$$P\left(\frac{\max\{Y, Y_1, Y_2, \dots, Y_n\} - d_n}{c_n} \leq x\right) = P(Y \leq x) = G(x)$$

για κάθε  $x$  και  $n = 1, 2, \dots$  ή ισοδύναμα,  $G^n(c_n x + d_n) = G(x)$ .

Ειδικότερα έχει αποδειχθεί ότι οι κατανομές που μπορούν να προκύψουν ως όρια κανονικοποιημένων μεγίστων είναι μόνο αυτές που ανήκουν στην κλάση των max-stable κατανομών. Αυτό που έχει μεγαλύτερο ενδιαφέρον είναι ότι η μορφή που μπορούν να έχουν, είναι ένας από μόλις τρεις τύπους κατανομών. Το θεώρημα των Fisher-Tippett (1928) περιγράφει αυτούς τους τρεις τύπους. Θα λέμε ότι δύο κατανομές  $G_1, G_2$  είναι του ίδιου τύπου αν υπάρχουν σταθερές  $a, b$  τέτοιες ώστε  $G_1(x) = G_2((x - b)/a)$  για κάθε  $x$ .

### Θεώρημα 1. (Fisher-Tippett)

Έστω  $X_1, X_2, \dots$  μια ακολουθία από ανεξάρτητες ισόνομες τυχαίες μεταβλητές. Αν υπάρχουν ακολουθίες  $c_n > 0$  και  $d_n \in \mathbb{R}$  και μια μη εκφυλισμένη κατανομή  $G$  ώστε

$$\frac{M_n - d_n}{c_n} = \frac{\max\{X_1, X_2, \dots, X_n\} - d_n}{c_n} \rightarrow_d G$$

τότε η  $G$  θα είναι του ίδιου τύπου με μια από τις τρεις ακόλουθες κατανομές :

$$\text{Frechet: } \Phi_a(x) = \begin{cases} 0 & , x \leq 0 \\ e^{-x^{-a}} & , x > 0 \end{cases} , a > 0$$

$$\text{Weibull: } \Psi_a(x) = \begin{cases} e^{-(-x)^a} & , x \leq 0 \\ 1 & , x > 0 \end{cases} , a > 0$$

$$\text{Gumbel: } \Lambda(x) = e^{-e^{-x}} , x \in \mathbb{R}$$

Το παραπάνω θεώρημα αποδεικνύεται δείχνοντας ότι οι μόνες (μη εκφυλισμένες) λύσεις της συναρτησιακής εξίσωσης που προκύπτει από τον ορισμό των max-stable κατανομών είναι οι τρεις παραπάνω. Οι κατανομές  $\Lambda$ ,  $\Phi_a$ ,  $\Psi_a$  καλούνται *τυπικές κατανομές ακροτάτων* ενώ κατανομές που είναι του ίδιου τύπου με αυτές καλούνται *κατανομές ακροτάτων*.

Όταν τώρα έχουμε ένα τυχαίο δείγμα που θέλουμε να αναλύσουμε, προσδιορίζουμε ποια από τις τρεις οριακές κατανομές ακολουθεί το κανονικοποιημένο μέγιστο από το τυχαίο δείγμα και βρίσκουμε με κατάλληλο τρόπο εκτιμήσεις για τις παραμέτρους της. Σε αυτή τη διαδικασία υπάρχει το μειονέκτημα ότι μια λανθασμένη επιλογή οριακής κατανομής θα έχει ως συνέπεια και λανθασμένα συμπεράσματα. Το μειονέκτημα αυτό μπορεί να εξαλειφθεί, γιατί οι τρεις τύποι των τυπικών κατανομών ακροτάτων που ορίστηκαν πριν μπορούν να ενοποιηθούν σε μια οικογένεια, η οποία καλείται *γενικευμένη κατανομή ακροτάτων GEV* (Generalized Extreme Value Distribution).

## Θεώρημα 2.

Έστω  $X_1, X_2, \dots$  μια ακολουθία από ανεξάρτητες ισόνομες τυχαίες μεταβλητές. Αν υπάρχουν ακολουθίες  $c_n > 0$  και  $d_n \in \mathbb{R}$  και μια μη εκφυλισμένη κατανομή  $G$  ώστε

$$\frac{M_n - d_n}{c_n} \stackrel{d}{=} \frac{\max\{X_1, X_2, \dots, X_n\} - d_n}{c_n} \stackrel{d}{=} G$$

τότε η  $G$  θα είναι της μορφής

$$G(x; \mu, \sigma, \xi) = e^{-\left(1 + \xi \frac{x - \mu}{\sigma}\right)^{\frac{1}{\xi}}}, \quad x : 1 + \xi \frac{x - \mu}{\sigma} > 0,$$

όπου  $\mu, \zeta \in \mathbb{R}$  και  $\sigma > 0$ . Για  $\zeta = 0$  θεωρούμε ως  $G$  το όριο της για  $\zeta \rightarrow 0$ , δηλαδή

$$G(x; \mu, \sigma, 0) = e^{-e^{-\frac{x-m}{\sigma}}}, x \in \mathbb{R}.$$

Για κατάλληλες τιμές των παραμέτρων  $\mu, \sigma, \zeta$  της  $GEV$  παίρνουμε προφανώς τους τρεις τύπους των κατανομών ακροτάτων :

- Αν  $\zeta > 0$  και θέσουμε  $\zeta = 1/\alpha, \sigma = c/\alpha, \mu = c + d$  τότε

$$G(x) = e^{-\left(1+x\frac{x-m}{\sigma}\right)^{\frac{1}{\alpha}}} = e^{-\left(\frac{x-d}{c}\right)^{\alpha}} = \Phi_{\alpha}((x-d)/c), \quad \frac{x-d}{c} > 0 \Leftrightarrow 1+x\frac{x-m}{\sigma} > 0$$

- Αν  $\zeta < 0$  και θέσουμε  $\zeta = -1/\alpha, \sigma = c/\alpha, \mu = -c + d$  τότε

$$G(x) = e^{-\left(1+x\frac{x-m}{\sigma}\right)^{\frac{1}{\alpha}}} = e^{-\left(\frac{x-d}{c}\right)^{\alpha}} = \Psi_{\alpha}((x-d)/c), \quad \frac{x-d}{c} < 0 \Leftrightarrow 1+x\frac{x-m}{\sigma} > 0$$

- Αν  $\zeta = 0$  (δηλαδή θεωρήσουμε ότι  $\zeta \rightarrow 0$ ) και θέσουμε  $\sigma = c, \mu = d$  τότε

$$G(x) = e^{-\left(1+x\frac{x-m}{\sigma}\right)^{\frac{1}{\alpha}}} \xrightarrow{x \rightarrow 0} e^{-e^{-\frac{x-m}{\sigma}}} = \Lambda((x-d)/c), x \in \mathbb{R}$$

Βασιζόμενοι στο προηγούμενο θεώρημα και στην παρατήρηση που ακολουθεί μπορούμε να προσδιορίσουμε την κατανομή που ακολουθεί οριακά το μέγιστο, προσαρμόζοντας τις παρατηρήσεις που διαθέτουμε στην  $GEV$  και εκτιμώντας τις δικές της παραμέτρους. Έτσι, είμαστε σε θέση να αποφασίσουμε ποια από τις *Frechet, Weibull, Gumbel* είναι η κατανομή που ακολουθεί οριακά το κανονικοποιημένο μέγιστο. Προφανώς όσο καλύτερη προσαρμογή έχουμε στην  $GEV$  τόσο πιο σίγουροι θα είμαστε για την μετέπειτα επιλογή μας.

### 2.1.2 Περιοχές Έλξης

Στην περίπτωση που γνωρίζουμε την κατανομή  $F$  από την οποία προέρχονται τα δεδομένα μας,  $X_1, X_2, \dots, X_n$ , τότε είμαστε σε θέση να προσδιορίσουμε επακριβώς ποια θα είναι η μη εκφυλισμένη κατανομή  $G$  που συγκλίνουν τα κανονικοποιημένα μέγιστα  $(M_n - d_n)/c_n$ , χωρίς να χρειάζεται να ενοποιήσουμε τις τρεις κατανομές και

μέσω της παραπάνω διαδικασίας να αποφασίσουμε ποια είναι. Καταρχήν ας ορίσουμε την έννοια της περιοχής έλξης.

**Ορισμός 1.**

Μια τ.μ.  $X$  (ή αντίστοιχα η κατανομή της,  $F$ ) ανήκει στην περιοχή έλξης μιας κατανομής ακροτάτων  $G$  αν υπάρχουν ακολουθίες  $c_n > 0$  και  $d_n \in \mathbb{R}$  έτσι ώστε η

$$\frac{M_n - d_n}{c_n} \xrightarrow{d} G$$

(συμβολικά  $F \in \text{MDA}(G)$ ).

Επειδή με την θεωρία ακραίων τιμών μελετάμε την μορφή που έχει η δεξιά (ή αριστερή) ουρά κάποιας κατανομής  $F$  που ακολουθούν iid τυχαίες μεταβλητές, είναι λογικό ότι το χαρακτηριστικό της  $F$  που προσδιορίζει σε ποια περιοχή έλξης ανήκει είναι η δεξιά ουρά της (αντίστοιχα η αριστερή, αν αντί για δειγματικά μέγιστα έχουμε δειγματικά ελάχιστα). Στη συνέχεια θα προσδιοριστούν οι συνθήκες όπως δόθηκαν από τους Haan-Ferreira (2006) που χρειάζεται να ισχύουν ώστε για μια σ.κ.  $F$  να υπάρχουν  $c_n > 0$  και  $d_n \in \mathbb{R}$  τέτοια ώστε

$$\lim_{n \rightarrow \infty} F^n(c_n x + d_n) = GEV.$$

Ουσιαστικά η τρεις περιπτώσεις που θα αναφερθούν αντιστοιχούν στα εξής :

1.  $F \in \text{MDA}(\Phi_a)$
2.  $F \in \text{MDA}(\Psi_a)$
3.  $F \in \text{MDA}(A)$

**Θεώρημα 3. (Haan-Ferreira)**

Η σ.κ.  $F$  ανήκει στην περιοχή έλξης της  $GEV$  αν και μόνο αν

1. Για  $\xi > 0$  είναι  $x_F = \infty$  και

$$\lim_{x \rightarrow \infty} \frac{1 - F(tx)}{1 - F(x)} = t^{-\frac{1}{\xi}}, \text{ για κάθε } t > 0$$

2. Για  $\zeta < 0$  είναι  $x_F < \infty$  και

$$\lim_{x \rightarrow 0} \frac{1 - F(x_F - tx)}{1 - F(x_F - x)} = t^{-\frac{1}{\zeta}}, \text{ για κάθε } t > 0$$

3. Για  $\zeta = 0$  είναι  $x_F = \infty$  και

$$\lim_{x \rightarrow x_F} \frac{1 - F(x + tf(x))}{1 - F(x)} = e^{-t} \text{ για κάθε } t \in \mathbb{R},$$

όπου η  $f$  είναι μια κατάλληλη θετική συνάρτηση. Αν η παραπάνω σχέση ισχύει για κάποια  $f$  τότε η  $F$  ικανοποιεί την

$$\int_x^{x_F} (1 - F(s)) ds < \infty \text{ για } x < x_F,$$

και η  $f$  θα έχει την ακόλουθη μορφή

$$f(x) := \frac{\int_x^{x_F} (1 - F(s)) ds}{1 - F(x)}.$$

Για τις δύο πρώτες περιπτώσεις επίσης μπορεί να αποδειχθούν τα ακόλουθα :

1. Έστω  $\zeta > 0$ ,  $x_F = \infty$  και έστω ότι υπάρχει η  $F'$ . Αν

$$\lim_{x \rightarrow \infty} \frac{x F'(x)}{1 - F(x)} = \frac{1}{\zeta}$$

τότε  $F \in \text{MDA}(\Phi_\alpha)$

2. Έστω  $\zeta < 0$ ,  $x_F < \infty$  και έστω ότι υπάρχει η  $F'$  για  $x < x_F$ . Αν

$$\lim_{x \rightarrow x_F} \frac{(x_F - x) F'(x)}{1 - F(x)} = -\frac{1}{\zeta}$$

τότε  $F \in \text{MDA}(\Psi_\alpha)$

## 2.2 Στατιστική ανάλυση ακραίων παρατηρήσεων

### 2.2.1 Η μέθοδος Block Maxima

Ας υποθέσουμε ότι έχουμε  $X_1, X_2, \dots, X_m$  με  $m = nk$  ανεξάρτητες τ.μ. που προέρχονται από μια άγνωστη συνάρτηση κατανομής  $F$ . Μπορούμε να θεωρήσουμε ότι οι παρατηρήσεις χωρίζονται σε  $k$  το πλήθος υποσύνολα (blocks) από  $n$  παρατηρήσεις το καθένα. Συνήθως τα blocks καθορίζονται με βάση κάποια χρονική περίοδο (ημέρα, μήνα, έτος κτλ.). Συμβολίζουμε με  $Y_1, Y_2, \dots, Y_k$  τις μέγιστες τιμές σε καθένα από αυτά τα  $k$  υποσύνολα, οι οποίες καλούνται *block maxima* (μερικά μέγιστα).

Συγκεκριμένα

$$Y_i = M_n^{(i)} = \max\{X_{(i-1)n+1}, X_{(i-1)n+2}, \dots, X_{(i-1)n+n}\}, i = 1, 2, \dots, k$$

Οι τ.μ.  $X_1, X_2, \dots, X_m$  δεν χρειάζεται να είναι γνωστές, αρκεί να ξέρουμε τα *block maxima*. Εάν το  $n$  είναι αρκετά μεγάλο τότε μπορούμε να υποθέσουμε ότι τα κανονικοποιημένα  $Y_i$  ακολουθούν την Γενικευμένη Κατανομή Ακροτάτων (*GEV*). Αυτό υποδηλώνει ότι και τα  $Y_i$  ακολουθούν την *GEV* διότι

$$P\left(\frac{Y_i - d_n}{c_n} \leq z\right) \approx G(z)$$

και άρα

$$P(Y_i \leq y) \approx G\left(\frac{y - d_n}{c_n}\right) = G^*(y)$$

όπου  $G^*$  είναι κάποια άλλη κατανομή από την οικογένεια της *GEV*.

Εκτιμώντας την παράμετρο  $\zeta$  της  $G^*(z)$  και το εύρος ενός διαστήματος εμπιστοσύνης για αυτήν, μπορούμε να προσδιορίσουμε ποια από τις τρεις κατανομές ακροτάτων ( $\Phi_\alpha, \Psi_\alpha, A$ ) ακολουθούν τα *block maxima*.

Το Θεώρημα 2 μας παρέχει ένα μοντέλο για την κατανομή των *block maxima*. Η εφαρμογή του προϋποθέτει, όπως έχουμε αναφέρει, την λήψη των μέγιστων παρατηρήσεων (*block maxima*) από ισομεγέθη υποσύνολα των δεδομένων και την προσαρμογή τους στην *GEV*. Για να εφαρμόσουμε το μοντέλο σε κάποια δεδομένα η επιλογή των blocks είναι κρίσιμη. Αν τα blocks είναι αρκετά μικρά σε μέγεθος (μικρό  $n$ ) η προσαρμογή των μεγίστων από αυτά στην *GEV* θα είναι «φτωχή» κάτι που οδηγεί

σε μεγάλα σφάλματα στις εκτιμήσεις των παραμέτρων. Από την άλλη, αν το μέγεθος των blocks είναι μεγάλο θα έχουμε λίγα *block maxima* με συνέπεια οι εκτιμήσεις των παραμέτρων  $\mu$ ,  $\sigma$ ,  $\xi$  της GEV να παρουσιάζουν μεγάλη διακύμανση.

Η επιλογή του μεγέθους των blocks πρέπει να είναι προσεκτική ανάλογα και με την φύση των δεδομένων. Αν π.χ. μελετούμε θερμοκρασίες δεν θα ήταν σωστό να παίρναμε σαν μια χρονική περίοδο μια εποχή (ένα τρίμηνο), αφού σε αυτήν την περίπτωση το μέγιστο των καλοκαιρινών μηνών θα ήταν πολύ μεγαλύτερο από αυτό των χειμερινών. Κάτι τέτοιο θα είχε ως συνέπεια τα *block maxima* να μην ακολουθούν την ίδια κατανομή που είναι μια βασική προϋπόθεση για να εφαρμόσουμε την όλη θεωρία των ακραίων τιμών. Η άλλη βασική προϋπόθεση, δηλαδή ότι τα *block maxima* είναι ανεξάρτητες τ.μ. είναι μια λογική προσέγγιση αν φυσικά και οι τ.μ. που έχουμε ως αρχικά δεδομένα είναι ανεξάρτητες.

### 2.2.2 Εκτιμήσεις των παραμέτρων της GEV

Υπάρχουν αρκετές μέθοδοι για την εκτίμηση των παραμέτρων του μοντέλου μας, κάποιες από αυτές είναι

- Μέσω γραφημάτων (βασίζονται σε κατάλληλα probability plots)
- Μέθοδος των Ροπών (method of moments)
- Μέθοδος των σταθμισμένων ροπών (method of probability weighted moments)
- Μέθοδος των L-Ροπών (method of L-moments)
- Μέθοδος μέγιστης πιθανοφάνειας (MLE)

Στη συνέχεια θα επικεντρωθούμε στη μέθοδο της μέγιστης πιθανοφάνειας (MLE) που ασυμπτωτικά και κάτω από συνθήκες ομαλότητας για την σ.κ. του δείγματος, δίνει τις καλύτερες εκτιμήτριες οι οποίες ακολουθούν γνωστή κατανομή (πολυδιάστατη κανονική). Ένα πρόβλημα που ενδέχεται να προκύψει είναι κατά πόσο η GEV ικανοποιεί τις συνθήκες ομαλότητας που απαιτούνται ώστε να ισχύουν τα γνωστά ασυμπτωτικά αποτελέσματα που σχετίζονται με τις εκτιμήτριες μέγιστης πιθανοφάνειας. Αυτές οι συνθήκες δεν ικανοποιούνται σε όλο τον παραμετρικό χώρο της GEV γιατί το στήριγμά της είναι συναρτήσεις των παραμέτρων  $\mu$ ,  $\sigma$ ,  $\xi$ . Συγκεκριμένα,



το  $\mu$ -σ/ζ είναι το άνω στήριγμα της κατανομής όταν  $x < 0$ , ενώ είναι το κάτω στήριγμα, όταν  $x > 0$ . Η παραβίαση των συνθηκών σημαίνει ότι τα ασυμπτωτικά αποτελέσματα δεν μπορούν να εφαρμοστούν αυτόματα. Ο Smith (1985) μελέτησε το συγκεκριμένο πρόβλημα και κατέληξε στα εξής :

- Όταν  $x > -0.5$  οι ε.μ.π. έχουν τις γνωστές ασυμπτωτικές τους ιδιότητες
- Όταν  $-1 < x < -0.5$  οι ε.μ.π. μπορούν να βρεθούν αλλά δεν έχουν τις γνωστές ασυμπτωτικές τους ιδιότητες
- Όταν  $x < -1$  οι ε.μ.π. πιθανόν να μην μπορούν να βρεθούν.

Έστω  $Z_1, \dots, Z_k$  ανεξάρτητες τ.μ. που ακολουθούν την *GEV*, η συνάρτηση πυκνότητας πιθανότητας της *GEV* για  $\zeta \neq 0$  είναι :

$$g(x; m, s, x) = \frac{1}{s} \left( 1 + x \frac{x-m}{s} \right)^{\frac{1}{x}-1} e^{-\left( 1 + x \frac{x-m}{s} \right)^{\frac{1}{x}}}, \quad x \in \mathbb{R} \text{ με } 1 + x \frac{x-m}{s} > 0$$

Ο λογάριθμος της πιθανοφάνειας (log-likelihood) των  $Z_i$  θα είναι

$$\begin{aligned} l(m, s, x) &= \sum_{i=1}^k \ln g(Z_i : m, s, x) = \\ &= -k \ln s - \left( \frac{1}{x} + 1 \right) \sum_{i=1}^k \ln \left( 1 + x \frac{Z_i - m}{s} \right) - \sum_{i=1}^k \left( 1 + x \frac{Z_i - m}{s} \right)^{\frac{1}{x}}, \end{aligned}$$

ενώ ικανοποιείται η

$$1 + x \frac{Z_i - m}{s} > 0 \text{ για } i = 1, \dots, k$$

Για έναν συνδυασμό των παραμέτρων που δεν ισχύει η τελευταία σχέση, σημαίνει ότι τουλάχιστον μια από τις παρατηρήσεις μας (*block maxima*) είναι πέρα από κάποιο στήριγμα της κατανομής. Επομένως η πιθανοφάνεια είναι 0 και η log-likelihood είναι  $-\infty$ .

Στην περίπτωση που  $\zeta = 0$ , θεωρούμε ότι  $\zeta \rightarrow 0$ , και η κατανομή *GEV* εκφυλίζεται στην κατανομή *Gumbel* με σ.π.π.

$$g(x; m, s, 0) = \frac{1}{s} e^{-\frac{x-m}{s}} e^{-e^{-\frac{x-m}{s}}}, \quad x \in \mathbb{R}$$

και log-likelihood των  $Z_i$ ,

$$l(\mathbf{m}, \mathbf{s}, 0) = \sum_{i=1}^k \ln g(Z_i : \mathbf{m}, \mathbf{s}, 0) = -k \ln \mathbf{s} - \sum_{i=1}^k \frac{Z_i - \mathbf{m}}{\mathbf{s}} - \sum_{i=1}^k e^{-\frac{Z_i - \mathbf{m}}{\mathbf{s}}}.$$

Στην γενική περίπτωση της GEV δεν υπάρχουν αναλυτικές εκφράσεις για τις εκτιμήτριες μέγιστης πιθανοφάνειας. Αλλά για ένα δοσμένο σύνολο δεδομένων μπορούμε να βρούμε τις τιμές που μεγιστοποιούν την  $l(\mu, \sigma, \xi)$  χρησιμοποιώντας αριθμητικές μεθόδους βελτιστοποίησης. Πρέπει όμως να είμαστε βέβαιοι ότι οι αλγόριθμοι εύρεσης του μεγίστου της πιθανοφάνειας δεν εισέρχονται σε συνδυασμό των παραμέτρων που παραβιάζουν την

$$1 + x \frac{Z_i - \mathbf{m}}{\mathbf{s}} > 0.$$

Επίσης δυσκολίες που προκύπτουν κατά την αριθμητική εκτίμηση των παραμέτρων όταν το  $\xi$  παίρνει τιμές “κοντά” στο 0, παρακάμπτονται εύκολα αν χρησιμοποιήσουμε για log-likelihood την  $l(\mu, \sigma, 0)$  αντί της  $l(\mu, \sigma, \xi)$ .

Έχοντας πλέον τις εκτιμήσεις των παραμέτρων της GEV μπορούμε να εκμεταλλευτούμε τα γνωστά ασυμπτωτικά αποτελέσματα που αφορούν τις ε.μ.π. για να δημιουργήσουμε τα αντίστοιχα διαστήματα εμπιστοσύνης. Σύμφωνα με αυτά, οι  $\hat{\mathbf{m}}, \hat{\mathbf{s}}, \hat{\mathbf{x}}$  ακολουθούν μια πολυδιάστατη κανονική κατανομή με μέση τιμή  $(\mu, \sigma, \xi)^T$  και πίνακα διασποράς, τον αντίστροφο του πίνακα πληροφορίας, ο οποίος εκτιμάται από τον

$$\hat{\mathbf{I}} = - \begin{bmatrix} \frac{\partial^2 l}{\partial \mathbf{m}^2}(\hat{\mathbf{m}}, \hat{\mathbf{s}}, \hat{\mathbf{x}}) & \frac{\partial^2 l}{\partial \mathbf{m} \partial \mathbf{s}}(\hat{\mathbf{m}}, \hat{\mathbf{s}}, \hat{\mathbf{x}}) & \frac{\partial^2 l}{\partial \mathbf{m} \partial \mathbf{x}}(\hat{\mathbf{m}}, \hat{\mathbf{s}}, \hat{\mathbf{x}}) \\ \frac{\partial^2 l}{\partial \mathbf{s} \partial \mathbf{m}}(\hat{\mathbf{m}}, \hat{\mathbf{s}}, \hat{\mathbf{x}}) & \frac{\partial^2 l}{\partial \mathbf{s}^2}(\hat{\mathbf{m}}, \hat{\mathbf{s}}, \hat{\mathbf{x}}) & \frac{\partial^2 l}{\partial \mathbf{s} \partial \mathbf{x}}(\hat{\mathbf{m}}, \hat{\mathbf{s}}, \hat{\mathbf{x}}) \\ \frac{\partial^2 l}{\partial \mathbf{x} \partial \mathbf{m}}(\hat{\mathbf{m}}, \hat{\mathbf{s}}, \hat{\mathbf{x}}) & \frac{\partial^2 l}{\partial \mathbf{x} \partial \mathbf{s}}(\hat{\mathbf{m}}, \hat{\mathbf{s}}, \hat{\mathbf{x}}) & \frac{\partial^2 l}{\partial \mathbf{x}^2}(\hat{\mathbf{m}}, \hat{\mathbf{s}}, \hat{\mathbf{x}}) \end{bmatrix}$$

Αν  $s_{i,i}$  είναι το  $i$ -διαγώνιο στοιχείο του  $\hat{\mathbf{I}}^{-1}$ , προσεγγιστικά ισχύει ότι

$$\hat{\mathbf{m}} \sim N(\mathbf{m}, s_{1,1}), \quad \hat{\mathbf{s}} \sim N(\mathbf{s}, s_{2,2}), \quad \hat{\mathbf{x}} \sim N(\mathbf{x}, s_{3,3})$$

και άρα τα

$$(\hat{m} \pm \sqrt{s_{1,1}} z_{\alpha/2}), \quad (\hat{s} \pm \sqrt{s_{2,2}} z_{\alpha/2}), \quad (\hat{x} \pm \sqrt{s_{3,3}} z_{\alpha/2}),$$

αποτελούν προσεγγιστικά δ.ε. συντελεστού  $1 - \alpha$  για τις παραμέτρους  $\mu, \sigma, \xi$ , όπου  $z_\alpha$  είναι το άνω  $\alpha$ -σημείο της  $N(0,1)$ .

### 2.2.3 Έλεγχος καλής προσαρμογής

Σύμφωνα με το Θεώρημα 2, τα κανονικοποιημένα *block maxima* ακολουθούν οριακά την κατανομή *GEV*. Αυτό συμβαίνει όταν το μέγεθος των blocks είναι πολύ μεγάλο (θεωρητικά, όταν τείνει στο άπειρο). Όπως αναφέρθηκε και στην Παράγραφο 1.2.1, η επιλογή του πλήθους των υποσυνόλων ( $k$ ) και του μεγέθους αυτών ( $n$ ), παίζει σημαντικό ρόλο στην στατιστική ανάλυση των δεδομένων μας. Επομένως όταν επιλέγουμε τα  $k, n$  θα πρέπει μετά να ελέγξουμε πόσο καλά προσαρμόζονται τα *block maxima* στην *GEV*, ώστε να εξάγουμε με ασφάλεια τα συμπεράσματά μας. Τέτοιος έλεγχος μπορεί να γίνει μέσω γραφημάτων P-P plots ή Q-Q plots, καθώς και μέσω ιστογράμματος των *block maxima* μαζί με την συνάρτηση πυκνότητας πιθανότητας της *GEV* με παραμέτρους που έχουν εκτιμηθεί από τα δεδομένα.

Έστω  $Y_1, Y_2, \dots, Y_k$  τα *block maxima* τα οποία, θα πρέπει να ακολουθούν (προσεγγιστικά) μία  $GEV(\hat{m}, \hat{s}, \hat{x})$ . Το Q-Q plot των *block maxima* έναντι της *GEV* θα αποτελείται από τα σημεία

$$(Y_{(i)}, G^{-1}(\frac{i}{k+1}; \hat{m}, \hat{s}, \hat{x})) = (Y_{(i)}, \hat{m} - \frac{\hat{s}}{\hat{x}} (1 - (\ln \frac{i}{k+1})^{-\hat{x}})), i = 1, 2, \dots, k$$

όπου  $Y_{(1)} < Y_{(2)} < \dots < Y_{(k)}$  είναι τα διατεταγμένα  $Y_1, Y_2, \dots, Y_k$ . Εάν τα σημεία αυτά βρίσκονται κοντά στην διαγώνιο τότε συμπεραίνουμε ότι η προσαρμογή των δεδομένων είναι καλή. Ανάλογο συμπέρασμα εξάγουμε και από το P-P plot.

### 2.2.4 Στάθμη απόδοσης για $1/p$ χρονικές περιόδους απόδοσης

Μέχρι στιγμής έχουμε αναφέρει πότε χρησιμοποιείται η Θεωρία των Ακραίων Τιμών, τα βασικά σημεία της, την μέθοδο των *block maxima*, καθώς και πως ελέγχουμε ότι μπορούμε να χρησιμοποιήσουμε την μέθοδο αυτή για ασφαλή συμπερασματολογία. Αυτό που μένει να αναφέρουμε είναι τι συμπεράσματα μπορούμε να ε-

ξάγουμε. Μια σημαντική ποσότητα είναι η στάθμη απόδοσης η οποία έχει τον εξής ορισμό:

### Ορισμός 2.

Το κατόφλι  $z_p$  το οποίο υπερβαίνουν τα block maxima κατά μέσο όρο κάθε  $1/p$  χρονικές περιόδους, καλείται άνω όριο απόδοσης ή στάθμη απόδοσης (return level) για  $1/p$  χρονικές περιόδους απόδοσης (return period).

Ισοδύναμα, το  $z_p$  είναι η στάθμη την οποία υπερβαίνει ένα block maximum με πιθανότητα  $p$ .

Εκτιμώντας λοιπόν το  $z_p$  είμαστε σε θέση να απαντήσουμε σε ερωτήσεις του τύπου:

- ποια η πιθανότητα ο μέγιστος σεισμός το επόμενο έτος να έχει ένταση μεγαλύτερη από 5.5 βαθμούς της κλίμακας Ρίχτερ

ή

- ποιο είναι το μέγεθος που δεν θα υπερβεί κάποιος σεισμός τα επόμενα 10 χρόνια με πιθανότητα 90% .

Η στάθμη απόδοσης  $z_p$  μπορεί εύκολα να βρεθεί από τη σχέση

$$P(M_n \leq z_p) = G(z_p) = 1 - p .$$

Χρησιμοποιώντας την αντίστροφη της  $G$ , έχουμε :

$$z_p = G^{-1}(1 - p) = \begin{cases} m - \frac{s}{x} (1 - (-\ln(1 - p))^{-x}), & x \neq 0 \\ m - s \ln(-\ln(1 - p)), & x = 0 \end{cases}$$

Προφανώς, μια εκτίμηση του  $z_p$  βρίσκεται όταν στην παραπάνω σχέση αντικαταστήσουμε τις παραμέτρους  $\mu$ ,  $\sigma$ ,  $\xi$  με τις εκτιμήσεις  $\hat{\mu}$ ,  $\hat{\sigma}$ ,  $\hat{\xi}$  που πήραμε για αυτές σύμφωνα με την μέθοδο της μέγιστης πιθανοφάνειας. Χρησιμοποιώντας την Μέθοδο Δέλτα βρίσκουμε προσεγγιστικά την κατανομή που ακολουθεί η  $\hat{z}_p$ ,

$$\hat{z}_p = g(\hat{\mu}, \hat{\sigma}, \hat{\xi}) \sim N(g(\mathbf{m}, \mathbf{s}, \mathbf{x}), \left(\frac{\partial g}{\partial \mu}, \frac{\partial g}{\partial \sigma}, \frac{\partial g}{\partial \xi}\right)^T \cdot \mathbf{I}^{-1} \cdot \left(\frac{\partial g}{\partial \mu}, \frac{\partial g}{\partial \sigma}, \frac{\partial g}{\partial \xi}\right))$$

όπου

$$\left(\frac{\partial g}{\partial m}, \frac{\partial g}{\partial s}, \frac{\partial g}{\partial x}\right) = \left(1, -\frac{1}{x}(1 - e^{-xw_p}), \frac{s}{x^2}(1 - e^{-xw_p}) + \frac{s}{x}w_p e^{-xw_p}\right)$$

είναι το διάνυσμα των μερικών παραγώγων της  $g(\mathbf{m}, \mathbf{s}, \mathbf{x})$  και

$$w_p = -\ln(-\ln(1-p)).$$

Προκειμένου να εκτιμήσουμε την διασπορά της  $\hat{z}_p$  θέτουμε όπου  $\mu, \sigma, \xi$  στο παραπάνω διάνυσμα τις ε.μ.π. τους. Αν η παράμετρος  $\xi$  εκτιμηθεί ότι είναι ίση με 0 τότε όμοια

$$g(\hat{\mathbf{m}}, \hat{\mathbf{s}}, 0) \sim N(g(\mathbf{m}, \mathbf{s}, 0), \left(\frac{\partial g}{\partial m}, \frac{\partial g}{\partial s}\right)^T \cdot \mathbf{I}^{-1} \cdot \left(\frac{\partial g}{\partial m}, \frac{\partial g}{\partial s}\right))$$

όπου αυτή τη φορά ο πίνακας πληροφορίας είναι  $2 \times 2$  και εκτιμάται θεωρώντας ότι  $\xi = 0$ .

Για την κατασκευή διαστήματος εμπιστοσύνης για το  $z_p$  μπορούμε εναλλακτικά να χρησιμοποιήσουμε την profile deviance function  $D_p(z_p)$ . Για να γίνει αυτό πρέπει να πραγματοποιήσουμε μια νέα παραμετροποίηση της GEV ώστε η  $z_p$  να είναι μια από τις παραμέτρους της GEV, π.χ. στη θέση της  $\mu$  να χρησιμοποιήσουμε την

$$\mu = z_p + \frac{s}{x}(1 - (-\ln(1-p))^{-x}).$$

Αν  $l'(z_p, \mathbf{s}, \mathbf{x})$  η πιθανοφάνεια των *block maxima* ως προς την νέα παραμετροποίηση, τότε αποδεικνύεται ότι θα ισχύει προσεγγιστικά,

$$D_p(z_p) = 2(l'(z_p, \hat{\mathbf{s}}, \hat{\mathbf{x}}) - \max_{\mathbf{s}, \mathbf{x}} l'(z_p, \mathbf{s}, \mathbf{x})) \sim c_1^2$$

Η τιμή της  $D_p(z_p)$  μπορεί να βρεθεί αριθμητικά για δεδομένες τιμές της  $z_p$ . Συνεπώς, το σύνολο

$$\{z_p : D_p(z_p) \leq c_1^2(\alpha)\}$$

είναι ένα δ.ε. συντελεστού  $1 - \alpha$  για το  $z_p$ . Όπως συνήθως εργαζόμαστε με αυτή την μέθοδο, κατασκευάζουμε το γράφημα της συνάρτησης  $\max_{\sigma, \xi} l'(z_p, \sigma, \xi)$  και προσθέτουμε την οριζόντια γραμμή με τεταγμένη  $l'(\hat{z}_p, \hat{\mathbf{s}}, \hat{\mathbf{x}}) - c_1^2(\alpha)/2$ . Οι τεταγμένες των σημείων τομής της ευθείας με την καμπύλη της συνάρτησης προσδιορίζουν τα άκρα του δ.ε. για το  $z_p$ .

### 2.2.5 Η μέθοδος Peaks Over Threshold (POT)

Μια εναλλακτική μέθοδος που αναπτύχθηκε για την μελέτη της συμπεριφοράς της δεξιάς ή της αριστερής ουράς μιας κατανομής βασίζεται στις υπερβάσεις των παρατηρήσεων ενός προκαθορισμένου κατώφλιου (Peaks Over Threshold, POT). Στην παράγραφο αυτή θα αναφέρουμε τα βασικά χαρακτηριστικά της όσον αφορά την δεξιά ουρά μιας κατανομής. Ανάλογα βέβαια είναι τα συμπεράσματα και για την αριστερή ουρά. Η μέθοδος POT μπορεί να χρησιμοποιηθεί μόνο όταν είναι διαθέσιμος ένας πλήρης κατάλογος παρατηρήσεων του φαινομένου που μελετάται (ή των παρατηρήσεων πάνω από μια στάθμη) έναντι της Block Maxima η οποία χρειάζεται μόνο τις μέγιστες παρατηρήσεις ανά χρονική περίοδο. Όταν έχουμε έναν τέτοιο κατάλογο είναι φυσικό να θέλουμε να εκμεταλλευτούμε όσο το δυνατόν περισσότερη πληροφορία μας παρέχει. Ας υποθέσουμε ότι έχουμε ετήσια δεδομένα για μια περίοδο 30 ετών. Σύμφωνα με την μέθοδο Block Maxima, επιλέγουμε την χρονική περίοδο όπου θα παίρνουμε τα τοπικά μέγιστα να είναι ένας χρόνος. Με αυτό τον τρόπο εκμεταλλευόμαστε 30 μόλις παρατηρήσεις για να αποφασίσουμε τι μορφή θα έχει η ουρά της κατανομής από την οποία προέρχονται τα δεδομένα. Αν όμως χρησιμοποιήσουμε την μέθοδο POT οι παρατηρήσεις που θα λάβουμε υπόψη μπορεί να αυξάνονται σημαντικά, διότι επιλέγουμε όλες τις παρατηρήσεις που υπερβαίνουν ένα κατώφλι  $u$ , το οποίο ορίζεται από τον ερευνητή.

Το βασικό ερώτημα είναι ποια κατανομή  $F_u$  ακολουθούν οι παρατηρήσεις πάνω από ένα συγκεκριμένο  $u$ . Αν γνωρίζαμε την αρχική κατανομή  $F$  από την οποία προέρχονται οι  $n$  παρατηρήσεις  $X_1, X_2, \dots, X_n$  τότε θα ήταν εύκολο να την προσδιορίσουμε αφού ισχύει :

$$F_u(y) = P(X - u \leq y | X > u) = 1 - \frac{1 - F(u + y)}{1 - F(u)}, y > 0.$$

Όπως και στην περίπτωση των *block maxima*, έτσι και τώρα μπορεί να βρεθεί η οριακή κατανομή της  $F_u$  όταν το  $n$  και το  $u$  είναι αρκετά μεγάλο. Το επόμενο θεώρημα μας δίνει την μορφή της οριακής κατανομής η οποία ονομάζεται *γενικευμένη κατανομή Pareto (GPD)*.

#### Θεώρημα 4.

Έστω  $X_1, X_2, \dots, X_n$  μια ακολουθία από ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με συνάρτηση κατανομής  $F$  για την οποία ισχύει ότι η κατανομή του κανονικοποιημένου μεγίστου συγκλίνει στην κατανομή  $GEV$ . Θεωρούμε επίσης ότι το  $n$  είναι αρκετά μεγάλο ώστε

$$P(M_n \leq x) \approx G(x; m, s, x),$$

όπου  $M_n = \max\{X_1, X_2, \dots, X_n\}$  και  $G$  είναι η σ.κ. της  $GEV$  με παραμέτρους  $\mu, \sigma, \xi$ . Τότε για αρκετά μεγάλο  $u$ , θα ισχύει ότι

$$F_u(y) = P(X - u \leq y | X > u) \approx H(y; \tilde{s}, x) = 1 - \left(1 + \frac{x y}{\tilde{s}}\right)^{-1/x},$$

για  $y > 0$  με  $1 + \frac{x y}{\tilde{s}} > 0$ .

Για τις δύο οριακές κατανομές που έχουμε αναφέρει, την  $GEV$  και την  $GPD$  υπάρχουν σχέσεις μεταξύ των παραμέτρων τους. Η παράμετρος  $\xi$  είναι κοινή και για τις δύο, ενώ η  $\tilde{s}$  αποτελεί συνάρτηση των  $\sigma, \mu$  και συγκεκριμένα,

$$\tilde{s} = s + x(u - m).$$

Ανάλογα με την τιμή της παραμέτρου  $\xi$  έχουν αποδειχθεί τα ακόλουθα:

- όταν  $x < 0$  τότε η  $GPD$  έχει άνω όριο στηρίγματος το  $u - \tilde{s} / x$ ,
- όταν  $x > 0$  η  $GPD$  εκτείνεται δεξιά ως το άπειρο.
- όταν  $x = 0$  τότε η  $GPD$  εκφυλίζεται στην εκθετική κατανομή με παράμετρο  $1/\tilde{s}$ .

Δύο βασικές υποθέσεις που πρέπει να ισχύουν για το Θεώρημα 4 είναι ότι, το πλήθος των τυχαίων μεταβλητών πρέπει να είναι μεγάλο (θεωρητικά το  $n$  πρέπει να τείνει στο  $\infty$ ), αλλά και το κατώφλι  $u$  πρέπει να είναι αρκετά μεγάλο. Στην μέθοδο των Block Maxima έπρεπε να διαλέξουμε  $k$  υποσύνολα  $n$  μεγέθους ώστε να υπάρχει καλή προσαρμογή στην  $GEV$  αλλά και να παίρνουμε εκτιμήσεις με μικρή μεταβλητότητα. Αντίστοιχα στη μέθοδο POT πρέπει να είμαστε προσεκτικοί όσο αφορά την επιλογή του κατωφλίου  $u$ . Δύο πράγματα πρέπει να ληφθούν υπόψη:

(i) αν το κατώφλι είναι μεγάλο τότε τα  $X_i$  που το υπερβαίνουν θα είναι λίγα με αποτέλεσμα οι εκτιμήσεις να μην είναι ασφαλείς και

(ii) αν το  $u$  είναι μικρό τότε η  $F_u$  δεν θα προσεγγίζει ικανοποιητικά την  $GPD$ .

Για την βέλτιστη επιλογή του κατώφλιου θα παρουσιάσουμε τρεις τεχνικές που βασίζονται σε κατάλληλα γραφήματα.

(1) Η πρώτη βασίζεται στο γεγονός ότι, αν υπάρχει ένα  $u_0$  τέτοιο ώστε η  $F_{u_0} \approx GPD$ , τότε για κάθε  $u > u_0$  θα ισχύει  $F_u \approx GPD$ , με συνέπεια οι αντίστοιχες μέσες τιμές των δύο κατανομών να είναι ίσες, δηλαδή

$$e(u) = E(X - u | X > u) \approx \frac{\tilde{S}}{1-x} = \frac{S + x(u - m)}{1-x}, \quad u > u_0.$$

Η παραπάνω δεσμευμένη μέση τιμή καλείται *μέση υπερβάλλουσα συνάρτηση*. Όταν λοιπόν κατασκευάσουμε το γράφημα  $(u, e(u))$ , θα επιλέξουμε ως  $u_0$  το σημείο δεξιά από το οποίο το γράφημα γίνεται γραμμικό. Στην πράξη μπορούμε να εκτιμήσουμε την  $e(u)$  από την δειγματική μέση τιμή των υπερβάσεων των  $X_i$  πάνω από το  $u$ , δηλαδή από την *εμπειρική μέση υπερβάλλουσα συνάρτηση*

$$\hat{e}(u) = \frac{1}{k(u)} \sum_{i: X_i > u} (X_i - u)$$

όπου  $k(u)$  είναι το πλήθος των  $X_i$  που υπερβαίνουν το  $u$  και να επιλέξουμε το κατώφλι  $u$  δεξιά από το οποίο το γράφημα  $(u, \hat{e}(u))$  γίνεται περίπου γραμμικό.

(2) Η δεύτερη τεχνική για τον προσδιορισμό του  $u$  είναι να εκτιμήσουμε τις παραμέτρους  $\xi, \tilde{S}$  της  $GPD$  για διάφορες τιμές του  $u$ . Αν ισχύει  $F_u \approx GPD$ , θα πρέπει η εκτίμηση του  $\xi$  να μην επηρεάζεται από το  $u$ , ενώ η εκτίμηση του  $\tilde{S} = S + x(u - m)$  να μεταβάλλεται γραμμικά ως προς το  $u$ . Από τα αντίστοιχα γραφήματα μπορούμε να επιλέξουμε το  $u$  που ικανοποιεί όλα τα παραπάνω.

(3) Τέλος η τρίτη γραφική μέθοδος η οποία προτάθηκε από τους Gertensgrabe και Werner (1989) βασίζεται στο προσδιορισμό του σημείου που ξεκινά η περιοχή των ακραίων τιμών. Πιο συγκεκριμένα έχοντας μια σειρά από διαφορές  $\Delta_i$  από ένα διατεταγμένο δείγμα  $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ , το σημείο που ξεκινά η περιοχή των ακραίων τιμών μπορεί να ανιχνευθεί προσδιορίζοντας το σημείο που αλλάζει μορφή η σειρά. Η ιδέα αυτή στηρίζεται στο ότι οι διαφορές των ακραίων τιμών θα είναι διαφορετικές από τις διαφορές των υπόλοιπων παρατηρήσεων. Για να εντοπιστεί λοιπόν το σημείο αυτό χρησιμοποιούμε την σειρά

$$U_i = \frac{U_i^* - \frac{i(i-1)}{4}}{\sqrt{\frac{i(i-1)(i-5)}{72}}}$$



όπου

$$U_i^* = \sum_{k=1}^i n_k$$

και  $n_k$  είναι ο αριθμός των  $\Delta_i$ ,  $i = 1, \dots, k$  που είναι μικρότερες ή ίσες από την  $\Delta_k$ . Επιπλέον, μια δεύτερη σειρά υπολογίζεται με την ίδια διαδικασία, αλλά από το τέλος προς την αρχή. Το σημείο που οι δύο σειρές τέμνονται καθορίζει το σημείο αλλαγής, δηλαδή το κατώφλι  $u$  που θα χρησιμοποιήσουμε.

**Παρατήρηση.** Από το γράφημα της εμπειρικής μέσης υπερβάλλουσας συνάρτησης έχουμε μια πρώτη ένδειξη για την μορφή της ουράς της κατανομής, άρα και για την τιμή της παραμέτρου  $\xi$ . Αποδεικνύεται εύκολα ότι, αν τα  $X_i$  ακολουθούν την εκθετική κατανομή με παράμετρο  $\lambda$  τότε  $e(u) = 1/\lambda$  και επομένως το γράφημα  $(u, e(u))$  θα είναι μια οριζόντια γραμμή, αφού το  $e(u)$  δεν εξαρτάται από το  $u$ . Αν στο γράφημα παρουσιάζεται μια φθίνουσα τάση τότε η ουρά της κατανομής θα είναι λεπτή και η παράμετρος  $\xi < 0$ , αντιθέτως μια αύξουσα τάση υποδεικνύει βαριά ουρά και επομένως αναμένουμε  $\xi > 0$ .

### 2.2.6 Εκτίμηση των παραμέτρων της GPD και έλεγχος καλής προσαρμογής

Εφόσον έχει καθοριστεί το κατώφλι  $u$  πάνω από το οποίο θα χρησιμοποιήσουμε τις παρατηρήσεις, μπορούμε να εκτιμήσουμε και τις παραμέτρους της γενικευμένης κατανομής Pareto. Έστω  $X_1, \dots, X_n$  ανεξάρτητες και ισόνομες τυχαίες μεταβλητές, τότε τα  $Z_1, \dots, Z_k$  είναι τα  $X_i$  που είναι μεγαλύτερα του  $u$  μείον το  $u$ , δηλαδή

$$\{Z_1, \dots, Z_k\} = \{X_i - u : X_i > u\}$$

και ακολουθούν οριακά την GPD. Η συνάρτηση πυκνότητας πιθανότητας αυτής είναι

$$h(z; \tilde{S}, x) = \frac{1}{\tilde{S}} \left(1 + \frac{xz}{\tilde{S}}\right)^{\frac{1}{x}-1} \cdot I(z, \tilde{S}, x), \quad z > 0$$

όπου  $I(z, \tilde{S}, x) = 1$  ή  $0$  ανάλογα με το αν  $1 + xz/\tilde{S} > 0$  ή όχι αντίστοιχα. Επομένως η log-likelihood των  $Z_1, \dots, Z_k$  θα είναι ( $\xi \neq 0$ )

$$l(\tilde{S}, x) = -k \ln \tilde{S} - \left(1 + \frac{1}{x}\right) \sum_{i=1}^k \ln \left(1 + \frac{xZ_i}{\tilde{S}}\right) + \sum_{i=1}^k \ln I(Z_i, \tilde{S}, x).$$

Το άθροισμα των  $\ln I(Z_i, \tilde{S}, \mathbf{x})$  είναι ίσο με μηδέν αν τα  $\tilde{S}, \zeta$  ικανοποιούν την  $1 + \mathbf{x} Z_i / \tilde{S} > 0$  για όλα τα  $Z_i$  και  $-\infty$  διαφορετικά. Στην περίπτωση που  $\zeta = 0$  αντίστοιχα έχουμε

$$l(\tilde{S}, 0) = -k \ln \tilde{S} - \frac{1}{\tilde{S}} \sum_{i=1}^k Z_i .$$

Όπως και στην περίπτωση της *GEV* έτσι και για την *GPD* δεν υπάρχουν αναλυτικές εκφράσεις για τις εκτιμήσεις των  $\tilde{S}, \zeta$  αλλά έχοντας ένα συγκεκριμένο δείγμα  $z_1, \dots, z_k$  μπορούμε να βρούμε τις τιμές των εκτιμήσεων μέσω αριθμητικών μεθόδων βελτιστοποίησης. Και σε αυτήν την περίπτωση πρέπει να διασφαλίσουμε ότι η επαναληπτική μέθοδος δεν θα εξέλθει από το σύνολο των εφικτών τιμών των παραμέτρων. Επίσης όταν το  $\zeta$  είναι κοντά στο 0 είναι καλύτερο να χρησιμοποιείται η  $l(\tilde{S}, 0)$ .

Σύμφωνα με τα γενικά συμπεράσματα που ισχύουν για τις ε.μ.π., οι  $\tilde{S}, \mathbf{x}$  θα ακολουθούν μια πολυδιάστατη κανονική κατανομή με μέση τιμή  $(\tilde{S}, \mathbf{x})^T$  και πίνακα διασποράς, τον αντίστροφο του πίνακα πληροφορίας ο οποίος εκτιμάται από τον

$$I = - \begin{pmatrix} \frac{\partial^2 l}{\partial \tilde{S}^2}(\tilde{S}, \mathbf{x}) & \frac{\partial^2 l}{\partial \tilde{S} \partial \mathbf{x}}(\tilde{S}, \mathbf{x}) \\ \frac{\partial^2 l}{\partial \mathbf{x} \partial \tilde{S}}(\tilde{S}, \mathbf{x}) & \frac{\partial^2 l}{\partial \mathbf{x}^2}(\tilde{S}, \mathbf{x}) \end{pmatrix}$$

Αν  $s_{i,i}$  είναι το διαγώνιο στοιχείο του  $I^{-1}$ , προσεγγιστικά ισχύει ότι

$$\tilde{S} \sim N(\tilde{S}, s_{1,1}), \quad \mathbf{x} \sim N(\mathbf{x}, s_{2,2})$$

και άρα τα

$$(\tilde{S} \pm \sqrt{s_{1,1}} z_{\alpha/2}), \quad (\mathbf{x} \pm \sqrt{s_{2,2}} z_{\alpha/2})$$

είναι τα προσεγγιστικά διαστήματα εμπιστοσύνης συντελεστού  $1-\alpha$  για τις παραμέτρους  $\tilde{S}, \mathbf{x}$ .

Όλα τα παραπάνω έχει νόημα να υπολογιστούν αν τα δεδομένα προσαρμόζονται καλά στο μοντέλο της *GPD*. Για τον έλεγχο καλής προσαρμογής των δεδομένων στην γενικευμένη κατανομή *Pareto* μπορούμε να χρησιμοποιήσουμε πάλι τα Q-Q και P-P plot, αλλά και να κατασκευάσουμε το γράφημα των παρατηρούμενων και θεωρη-

τικών συχνοτήτων σε μια λογαριθμική κλίμακα. Τα σημεία των Q-Q και P-P plots αντίστοιχα θα είναι τα παρακάτω,

$$(Z_{(i)}, H^{-1}(\frac{i}{k+1}, \frac{\sum_{j=1}^i Z_{(j)}}{i}, \bar{x})) = (Z_{(i)}, \frac{\sum_{j=1}^i Z_{(j)}}{i} ((1 - \frac{i}{k+1})^{-k} - 1)), \quad i = 1, 2, \dots, k$$

$$(H(Z_{(i)}, \frac{i}{k+1})) = (1 - (1 + \frac{\sum_{j=1}^i Z_{(j)}}{i})^{-1/k}, \frac{i}{k+1}), \quad i = 1, 2, \dots, k$$

όπου  $Z_{(1)} < Z_{(2)} < \dots < Z_{(k)}$  είναι τα διατεταγμένα POT. Επαναλαμβάνουμε ότι αν τα  $k$  αυτά σημεία αυτά είναι κοντά στη διαγώνιο, τότε μπορούμε να πούμε ότι έχουμε καλή προσαρμογή.

### 2.2.7 Στάθμη απόδοσης για $m$ παρατηρήσεις

Η έννοια της στάθμης απόδοσης διαφοροποιείται όταν αναλύουμε τα δεδομένα με την μέθοδο POT σε σύγκριση με την μέθοδο Block Maxima. Πιο συγκεκριμένα, όταν είχαμε στη διάθεσή μας τοπικά μέγιστα ανά χρονική περίοδο υπολογίζαμε το κατώφλι που μια παρατήρηση ξεπερνά κατά μέσο όρο κάθε  $1/p$  περιόδους. Με την μέθοδο POT υπολογίζεται το κατώφλι που ξεπερνά μια παρατήρηση κατά μέσο όρο κάθε  $m$  παρατηρήσεις. Υπενθυμίζουμε ότι την *GPD* ακολουθούν οριακά οι τ.μ.

$$\{Z_1, Z_2, \dots, Z_k\} = \{X_i - u : X_i > u\}$$

επομένως τα ερωτήματα που θα απαντήσουμε αφού αναλύσουμε τα σεισμολογικά δεδομένα μπορεί να είναι της μορφής:

- ποίο είναι το μέγεθος που δεν θα υπερβεί κάποιος σεισμός με πιθανότητα 90%
- ή
- κάθε πόσους κατά μέσο όρο (καταγεγραμμένους) σεισμούς θα συμβεί σεισμός μεγαλύτερος από 6 Ρίχτερ.

Εύκολα γίνεται αντιληπτό ότι οι απαντήσεις τέτοιου τύπου δεν έχουν ιδιαίτερη χρηστικότητα, αφού εμπλέκεται το πλήθος των καταγεγραμμένων σεισμών σε μια περιοχή. Για αυτόν τον λόγο πρέπει να γίνει μια αναγωγή από αριθμό συμβάντων σε χρονική περίοδο όπως έχουμε και στην περίπτωση των block maxima. Ο τρόπος θα αναλυθεί στην Παράγραφο 4.2, στη συνέχεια δίνεται ο ορισμός της στάθμης απόδοσης.

### Ορισμός 3.

Το κατώφλι  $x_m$  το οποίο υπερβαίνουν τα  $X_i$  κατά μέσο όρο κάθε  $m$  παρατηρήσεις καλείται άνω όριο απόδοσης ή στάθμη απόδοσης για  $m$  παρατηρήσεις. Ισοδύναμα μπορεί να θεωρηθεί ως η στάθμη που υπερβαίνει μία παρατήρηση  $X_i$  με πιθανότητα  $1/m$ .

Σύμφωνα με όσα έχουμε αναπτύξει, για  $x > u$  θα ισχύει ότι

$$P(X > x | X > u) \approx H(x - u) = \left(1 + x \frac{x - u}{\tilde{S}}\right)^{-1/x},$$

και επομένως

$$P(X > x) \approx P(X > u) \left(1 + x \frac{x - u}{\tilde{S}}\right)^{-1/x}$$

Το  $x_m$  θα πρέπει να ικανοποιεί την εξίσωση  $P(X > x_m) = 1/m$ . Θέτοντας όπου  $x$  το  $x_m$  και όπου  $P(X > x_m)$  το  $1/m$  στην παραπάνω σχέση και λύνοντας ως προς  $x_m$  προκύπτει ότι

$$x_m \approx u + \frac{\tilde{S}}{x} \left( (mP(X > u))^x - 1 \right)$$

Επομένως η εκτίμηση της στάθμης απόδοσης για  $m$  παρατηρήσεις θα είναι

$$\hat{x}_m \approx u + \frac{\hat{S}}{\hat{x}} \left( \left( \frac{mk(u)}{n} \right)^x - 1 \right),$$

όπου  $\hat{S}$ ,  $\hat{x}$  οι ε.μ.π των αντίστοιχων παραμέτρων και η  $P(X > u)$  εκτιμάται από το ποσοστό των παρατηρήσεων πάνω από το  $u$ , δηλαδή με το  $k(u)/n$ . Από τις ασυμπτωτικές ιδιότητες των ε.μ.π μπορούμε να εκτιμήσουμε τον πίνακα διασποράς των τριών εκτιμητριών των  $\tilde{S}$ ,  $x$ ,  $P(X > x)$  και σύμφωνα με την Μέθοδο Δέλτα να κατασκευάσουμε ένα διάστημα εμπιστοσύνης για το  $x_m$ . Δηλαδή το  $x_m = g(\tilde{S}, \hat{x})$  θα ακολουθεί προσεγγιστικά την

$$N\left(g(\tilde{S}, \hat{x}), \left(\frac{\partial g}{\partial \tilde{S}}, \frac{\partial g}{\partial \hat{x}}\right)^T \cdot \hat{I}^{-1} \cdot \left(\frac{\partial g}{\partial \tilde{S}}, \frac{\partial g}{\partial \hat{x}}\right)\right),$$

όπου

$$\frac{\partial g}{\partial \tilde{S}} = \frac{1}{\tilde{S}} \left( \left( \frac{mk(u)}{n} \right)^x - 1 \right), \quad \frac{\partial g}{\partial \hat{x}} = -\frac{\tilde{S}}{\hat{x}^2} \left( \left( \frac{mk(u)}{n} \right)^x - 1 \right) + \frac{\tilde{S}}{\hat{x}} \left( \frac{mk(u)}{n} \right)^x \ln\left(\frac{mk(u)}{n}\right)$$

και  $\hat{I}$  ο εκτιμημένος πίνακας πληροφορίας.

Εναλλακτικά μπορούμε να χρησιμοποιήσουμε πάλι την profile deviance function  $D_p(x_p)$  για την κατασκευή δ.ε. για το  $x_p$  όπως ακριβώς είχαμε περιγράψει για την

περίπτωση της  $z_p$ , κάνοντας δηλαδή μια νέα παραμετροποίηση ώστε η  $x_p$  να είναι μια από τις παραμέτρους της *GPD*.

### 2.3 Λογοκριμένα δεδομένα

Ένα από τα βασικότερα πλεονεκτήματα που παρέχει η χρήση της θεωρίας ακραίων τιμών είναι ότι ο κατάλογος των δεδομένων δεν χρειάζεται να είναι πλήρης. Όμως και πάλι μπορούν να ανακύψουν προβλήματα όταν έχουμε στην διάθεσή μας σχετικά λίγες παρατηρήσεις. Μπορεί να μην έχουμε καλή προσαρμογή των μοντέλων αφού μιλάμε πάντα για οριακές κατανομές, αλλά υπάρχει και η περίπτωση που δεν μπορούμε καν να χρησιμοποιήσουμε την συγκεκριμένη θεωρία όπως ακριβώς παρουσιάστηκε στις προηγούμενες παραγράφους. Με την μέθοδο Block Maxima αναφέραμε ότι συλλέγουμε τα τοπικά μέγιστα, χωρίζοντας τις παρατηρήσεις σε ισομεγέθη blocks (π.χ.  $n$  παρατηρήσεων το καθένα). Όταν σε συγκεκριμένα χρονικά διαστήματα αντιστοιχεί ίδιος (αναμενόμενος) αριθμός  $n$  παρατηρήσεων, τότε μπορεί να θεωρηθεί ότι αυτά τα χρονικά διαστήματα είναι τα blocks. Όταν τα χρονικά διαστήματα είναι σχετικά μεγάλα (π.χ. ετήσια) τότε είναι σχεδόν σίγουρο ότι θα υπάρχουν παρατηρήσεις (και άρα μέγιστο) σε κάθε διάστημα. Όταν όμως αυτά μικραίνουν, έτσι ώστε να μεγαλώσει το δείγμα, υπάρχει κίνδυνος να μην έχει καταγραφεί σε κάποια από αυτά τιμή του φαινομένου που μελετάμε. Η απουσία παρατήρησης σε μια χρονική περίοδο μπορεί π.χ. να οφείλεται στο ότι η τιμή της ήταν μικρότερη από κάποιο όριο, πάνω από το οποίο καταγράφονται οι τιμές του φαινομένου. Στην περίπτωση αυτή μπορεί να θεωρηθεί ότι έχουμε λογοκριμένα δεδομένα και πιο συγκεκριμένα αριστερή λογοκρισία τύπου I.

Το πρόβλημα λοιπόν έγκειται στο πως θα εκτιμηθούν οι παράμετροι της κατανομής από την οποία προέρχονται τα τοπικά μέγιστα, δηλαδή της *GEV*, όταν αυτά δεν είναι διαθέσιμα στο σύνολό τους. Στην βιβλιογραφία αναφέρονται δύο μέθοδοι για την αντιμετώπιση αυτού του προβλήματος, η μια είναι η μέθοδος *μέγιστης πιθανοφάνειας για λογοκριμένα δεδομένα* και η άλλη είναι η μέθοδος των *Partial Probability Weighted Moments*. Στη συνέχεια θα παρουσιαστούν οι βασικές εξισώσεις των δύο μεθόδων, μέσω των οποίων υπολογίζονται οι εκτιμήσεις των τριών παραμέτρων της *GEV*.

### 2.3.1 Μέθοδος Μέγιστης Πιθανοφάνειας

Η συνάρτηση κατανομής της Γενικευμένης Κατανομής Ακροτάτων είναι η

$$F(x) = e^{-\left(1 - b \frac{x-c}{a}\right)^{1/b}},$$

όπου οι παράμετροι  $a, b, c$  αντιστοιχούν στις  $\sigma, -\xi, \mu$  του τύπου που δόθηκε στην Παράγραφο 2.1.1 για την *GEV*. Παραγωγίζοντας ως προς  $x$  παίρνουμε την συνάρτηση πυκνότητας πιθανότητας

$$f(x) = \frac{e^{-y-e^{-y}}}{a(1-t)},$$

όπου

$$y = -\frac{1}{b} \ln(1-t), \quad t = \frac{b(x-c)}{a}.$$

Έχοντας ένα δείγμα τυχαίων μεταβλητών  $X_1, X_2, \dots, X_N$  από το οποίο οι  $m$  μικρότερες παρατηρήσεις έχουν λογοκριθεί (κάτω από το  $x_L$ ), η συνάρτηση πιθανοφάνειας είναι

$$\frac{N!}{m!} [F(x_L)]^m \prod_{i=1}^n f(x_i),$$

όπου  $n = N - m$  και το  $x_L$  είναι το αριστερό όριο λογοκρισίας. Στην λογοκρισία τύπου I το  $x_L$  είναι σταθερός αριθμός, ενώ οι παρατηρήσεις  $n$  που δεν λογοκρίνονται είναι τυχαία μεταβλητή. Η αντίστοιχη log-likelihood είναι

$$l(a, b, c) = \ln N! - \ln m! + m \ln F(x_L) + \sum_{i=1}^n \ln f(x_i)$$

Οι τιμές των παραμέτρων που μεγιστοποιούν την παραπάνω συνάρτηση δεν μπορούν να βρεθούν αναλυτικά, αλλά έχοντας ένα συγκεκριμένο δείγμα μπορούμε με επαναληπτικές μεθόδους, όπως την Newton-Raphson, να βρούμε τις ε.μ.π. των τριών παραμέτρων της *GEV*.

### 2.3.2 Μέθοδος των Μερικών Πιθανοτικά Σταθμισμένων Ροπών

Οι Πιθανοτικά Σταθμισμένες Ροπές (Probability Weighted Moments - PWM) μια τυχαίας μεταβλητής  $X$  με συνάρτηση κατανομής  $F(x) = P(X \leq x)$  είναι οι ποσότητες

$$M_{p,r,s} = \int_0^1 [x(F)]^p F^r (1-F)^s dF = \int_0^1 [F^{-1}(u)]^p u^r (1-u)^s du$$

όπου  $p, r, s$  πραγματικοί αριθμοί. Όταν  $p = 1$  και  $s = 0$  έχουμε

$$b_r = M_{1,r,0} = \int_0^1 x(F)F^r dF = \int_0^1 F^{-1}(u)u^r du$$

Έχοντας ένα τυχαίο δείγμα μεγέθους  $n$  από την κατανομή  $F$ , οι εκτιμήσεις των  $\beta_r$  βασίζονται στο διατεταγμένο δείγμα  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ . Η στατιστική συνάρτηση

$$b_r = \frac{1}{n} \sum_{i=1}^n \frac{(i-1)(i-2)\dots(i-r)}{(n-1)(n-2)\dots(n-r)} x_{(i)}$$

είναι αμερόληπτη εκτιμήτρια του  $\beta_r$ .

Έτσι όπως ορίστηκαν οι PWM μπορούν να χρησιμοποιηθούν μόνο για πλήρη δεδομένα, ωστόσο εύκολα μπορούν να επεκταθούν ώστε να εφαρμόζονται και σε λογοκριμένα. Αντί να ολοκληρώσουμε από το 0 ως το 1, ολοκληρώνουμε από ένα κάτω όριο  $F_0$  ως το 1 και παίρνουμε

$$M'_{p,r,s} = \int_{F_0}^1 [x(F)]^p F^r (1-F)^s dF = \int_{F_0}^1 [F^{-1}(u)]^p u^r (1-u)^s du,$$

όπου  $F_0 = F(x_0)$  και  $x_0$  είναι το κατώφλι κάτω από το οποίο λογοκρίνονται τα δεδομένα. Οι ποσότητες  $M'_{p,r,s}$  ονομάζονται *Μερικές Πιθανοτικά Σταθμισμένες Ροπές* (Partial Probability Weighted Moments - PPWM). Ανάλογες ποσότητες ορίζονται όταν έχουμε δεξιά ή δίπλευρη λογοκρισία. Για  $p = 1$  και  $s = 0$  οι PPWM γίνονται

$$b'_r = M'_{1,r,0} = \int_{F_0}^1 x(F)F^r dF = \int_{F_0}^1 F^{-1}(u)u^r du$$

Ομοίως για το διατεταγμένο δείγμα  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$  η παρακάτω στατιστική συνάρτηση είναι αμερόληπτη εκτιμήτρια του  $b'_r$ .

$$b'_r = \frac{1}{n} \sum_{i=1}^n \frac{(i-1)(i-2)\dots(i-r)}{(n-1)(n-2)\dots(n-r)} x_{(i)}^*,$$

όπου

$$x_{(i)}^* = \begin{cases} 0, & x_{(i)} \leq x_0 \\ x_{(i)}, & x_{(i)} > x_0 \end{cases}.$$

Όταν η κατανομή από την οποία προέρχονται τα δεδομένα είναι η *GEV* με παραμέτρους αυτές που δόθηκαν στην προηγούμενη παράγραφο και  $b \neq 0$  τα PPWM υπολογίζονται από τον επόμενο τύπο,

$$b'_r = (c + \frac{a}{b}) \frac{1}{r+1} (1 - F_0^{r+1}) - \frac{a}{b} \frac{\Gamma(1+b)}{(r+1)^{1+b}} P(1+b, -(r+1) \ln F_0),$$

όπου  $P(\dots)$  είναι η μη-πλήρης συνάρτηση Γάμμα (Incomplete Gamma Function),

$$P(1+b, -(r+1)\ln F_0) = \int_0^{-(r+1)\ln F_0} \frac{q^b e^{-q} dq}{\Gamma(1+b)}.$$

Από την εξίσωση των PPWM έχουμε

$$b'_0 = \left(c + \frac{a}{b}\right)(1 - F_0) - \frac{a}{b}\Gamma(1+b)P(1+b, -\ln F_0)$$

$$\frac{2b'_1}{1 - F_0^2} - \frac{b'_0}{1 - F_0} = -\frac{a}{b} \left[ \frac{P(1+b, -2\ln F_0)}{2^b(1 - F_0^2)} - \frac{P(1+b, -\ln F_0)}{1 - F_0} \right]$$

$$\frac{2b'_1}{1 - F_0^2} - \frac{b'_0}{1 - F_0} = \frac{P(1+b, -2\ln F_0)}{2^b(1 - F_0^2)} - \frac{P(1+b, -\ln F_0)}{1 - F_0}$$

$$\frac{3b'_2}{1 - F_0^3} - \frac{b'_0}{1 - F_0} = \frac{P(1+b, -3\ln F_0)}{3^b(1 - F_0^3)} - \frac{P(1+b, -\ln F_0)}{1 - F_0}$$

Όταν το  $F_0$  είναι γνωστό μπορούμε να εκτιμήσουμε τις παραμέτρους  $a$ ,  $b$ ,  $c$  με επαναληπτικές μεθόδους, αντικαθιστώντας στις παραπάνω τρεις εξισώσεις τα  $b'_r$  με τους αμερόληπτους εκτιμητές αυτών  $b'_r$ .



## ΚΕΦΑΛΑΙΟ 3

### Στατιστική ανάλυση σεισμικών δεδομένων – Εμπειρική αξιολόγηση της μεθόδου Block Maxima με λογοκριμένες παρατηρήσεις

Μορφοποιήθηκε

#### 3.1 Ο τρίτος τύπος της οριακής κατανομής του *Gumbel*

Τα σεισμολογικά δεδομένα, όπως αναφέρθηκε και στην εισαγωγή, έχουν τέτοια μορφή που μας επιτρέπουν αλλά και μας αναγκάζουν να τα αντιμετωπίσουμε με την θεωρία ακραίων τιμών. Πολλοί στατιστικοί έχουν στρέψει την προσοχή τους προς αυτήν την κατεύθυνση και έχουν δημοσιευτεί εργασίες με εφαρμογές της συγκεκριμένης θεωρίας σε σεισμολογικά δεδομένα για διάφορες περιοχές του κόσμου Burton, P.W, Qin G., Tselentis G.A. and Sokos E. (2004), Pisarenko V. F. And Sor-netted D.(2003). Φυσιολογικό ήταν λοιπόν να υπάρξει μια πιο εξειδικευμένη αντιμετώπιση του προβλήματος πρόβλεψης σεισμών. Μια από αυτές στηρίζεται στο γεγονός ότι το μέγεθος οποιουδήποτε σεισμού δεν μπορεί να υπερβεί κάποια συγκεκριμένη τιμή, αφού η ενέργεια που συσσωρεύεται σε μια περιοχή, η οποία είναι και είναι το αίτιο για έναν σεισμό, είναι πεπερασμένη. Ο Richter (1958) υπέδειξε ότι αυτή είναι κοντά στο 8.5 με 9 βαθμούς της ομώνυμης κλίμακας για οποιονδήποτε σεισμό στον κόσμο. Έχοντας ως δεδομένο ότι υπάρχει ένα πεπερασμένο άνω δεξιό στήριγμα, η κατάλληλη συνάρτηση κατανομής για να μελετήσουμε την ουρά της κατανομής των σεισμολογικών δεδομένων, σύμφωνα με τους Yegular and Kuo (1974) και Burton (1979) είναι ο τρίτος τύπος της οριακής κατανομής του *Gumbel*. Συγκεκριμένα, μπορούμε να θεωρήσουμε ότι τα τοπικά μέγιστα των σεισμών ανά χρονική περίοδο και σε συγκεκριμένη περιοχή (η οποίες ορίζονται από τον ερευνητή) ακολουθούν προσεγγιστικά την κατανομή GIII (*Gumbel III*) με συνάρτηση κατανομής:

$$G_{III}(x) = e^{-\left(\frac{w-x}{w-m}\right)^k}, k > 0$$

ή

$$G_{III}(x) = e^{-\left(\frac{w-x}{w-m}\right)^{\frac{1}{\lambda}}}, \lambda > 0$$

Με  $\omega, \mu, k = 1/\lambda$ , συμβολίζονται οι τρεις παράμετροι της οριακής κατανομής. Η παράμετρος  $\omega$  εκφράζει το άνω όριο της έντασης ενός σεισμού, το  $\mu$  εκφράζει την χαρακτηριστική ακραία τιμή και η  $\lambda$  είναι η παράμετρος μορφής (shape) της κατανομής. Όταν το  $x$  τείνει στο άνω όριο  $\omega$  τότε

$$G_{III}(x) \rightarrow 1,$$

ενώ όταν το  $x$  ελαττώνεται τότε

$$G_{III}(x) \rightarrow 0.$$

Είναι προφανές ότι ο τρίτος τύπος της οριακής κατανομής του *Gumbel* συνδέεται άμεσα με την κατανομή *Weibull* του θεωρήματος των *Fisher-Tippet* που παρουσιάστηκε στο δεύτερο κεφάλαιο της εργασίας. Και οι δύο τύποι έχουν τρεις παραμέτρους και με κατάλληλη παραμετροποίηση μπορούμε να μεταβούμε από τον έναν στον άλλον.

**Παρατήρηση.** Ο παραπάνω τύπος της *Gumbel III* συνήθως χρησιμοποιείται όταν έχουμε στη διάθεσή μας ετήσια μέγιστα. Φυσικά ο ερευνητής μπορεί να επιλέξει μικρότερη ή μεγαλύτερη χρονική περίοδο στην οποία λαμβάνει τα τοπικά μέγιστα. Η συνάρτηση κατανομής για  $T$  έτη ως χρονική περίοδο είναι

$$G_{III}^T(x) = e^{-T \left( \frac{w-x}{w-m} \right)^{\frac{1}{\lambda}}}, \lambda > 0$$

Αρκετοί ερευνητές που χρησιμοποιούν την συγκεκριμένη κατανομή για την μελέτη σεισμικών δεδομένων, συνήθως ακολουθούν την ακόλουθη μεθοδολογία για την εκτίμηση των παραμέτρων της. Μπορεί να θεωρηθεί ότι οι παράμετροι της *Gumbel III* εκτιμώνται έτσι ώστε να πραγματοποιείται βέλτιστη προσαρμογή των δεδομένων σε αυτήν μέσω του Q-Q plot. Υπενθυμίζεται ότι το Q-Q Plot των  $x_1, x_2, \dots, x_n$  ως προς μία σ.κ.  $G$  αποτελείται από τα σημεία

$$\left( x_{(i)}, G^{-1}\left(\frac{i}{n+1}\right) \right)$$

όπου  $x_{(1)}, \dots, x_{(n)}$  οι διατεταγμένες κατά αύξουσα σειρά τιμές των  $x_i$ . Αν τα δεδομένα προέρχονται από την  $G$  θα πρέπει  $x_{(i)} \approx G^{-1}(i/(n+1))$ . Επομένως, οι εκτιμήσεις των  $\omega, \mu, \lambda$  βρίσκονται ελαχιστοποιώντας το άθροισμα

$$c_T^2(w, m, l) = \sum_{i=1}^n w_i (x_{(i)} - G_{III}^{-1}(i/(n+1)))^2$$

όπου  $w_i$  είναι κάποια βάρη που σχετίζονται με τον αριθμό των παρατηρήσεων στην περίοδο που αντιστοιχεί το  $x_{(i)}$  και

$$GIII^{-1}(i/(n+1)) = w - (w - m)[-\ln(i/(n+1))/T]^l .$$

Μερικές φορές αντί του  $i/(n+1)$  (Gumbel (1954)) στους παραπάνω τύπους λαμβάνονται

$$\frac{i-0.44}{n+0.12}, \text{ (Gringorten (1963))} \quad \text{ή} \quad \frac{i-0.3}{n+0.4}, \text{ (Turnbull and Weatherly (2006))}$$

Με την συγκεκριμένη μεθοδολογία δεν χρησιμοποιούνται όλα τα δεδομένα. Ο ερευνητής εξ αρχής έχει θέσει ένα κατώφλι και χρησιμοποιεί τους σεισμούς που η έντασή τους το υπερβαίνει. Αυτό όμως έχει ως αποτέλεσμα ότι για κάποιες χρονικές περιόδους μπορεί να μην υπάρχει τοπικό μέγιστο που να υπερβαίνει αυτό το κατώφλι. Όσο μεγαλώνουν τα χρονικά διαστήματα τόσο λιγότερα κενά θα υπάρχουν, αλλά ταυτόχρονα θα μειώνεται ο αριθμός των τοπικών μεγίστων, δηλαδή το μέγεθος του δείγματος. Και οι δύο αυτές ποσότητες έχουν άμεσο αντίκτυπο στην καλή προσαρμογή του μοντέλου. Ο στόχος λοιπόν είναι να βρεθεί το διάστημα εκείνο που βελτιστοποιεί την προσαρμογή. Αυτό επιτυγχάνεται επαναλαμβάνοντας την παραπάνω διαδικασία για διάφορες χρονικές περιόδους  $T_1, T_2, \dots, T_z$  και επιλέγεται εκείνη που δίνει το μικρότερο  $c^2(w, m, I)$ .

Στην συγκεκριμένη εργασία δεν θα ακολουθήσουμε την προαναφερθείσα μεθοδολογία, αλλά οι εκτιμήσεις των παραμέτρων θα βρεθούν με την κλασική μέθοδο, δηλαδή τη μέθοδο μέγιστης πιθανοφάνειας. Επίσης οι χρονικές περίοδοι που θα επιλεγθούν αρχικά θα είναι τέτοιες ώστε να υπάρχουν τοπικά μέγιστα σε κάθε μια από αυτές. Η συνάρτηση πυκνότητας πιθανότητας της  $GIII(x)$  είναι

$$f(x; w, m, l) = \frac{T}{l} \left( \frac{w-x}{w-m} \right)^{\frac{1}{l}-1} \frac{1}{w-m} e^{-T \left( \frac{w-x}{w-m} \right)^{\frac{1}{l}}}$$

και η log-likelihood αντίστοιχα

$$\begin{aligned} l(\omega, \mu, \lambda) &= \sum_{i=1}^n \ln f(x; w, m, l) = \\ &= \sum_{i=1}^n \left\{ \ln T - \ln(l(w-m)) + \left(\frac{1}{l} - 1\right) \ln \left( \frac{w-x_i}{w-m} \right) - T \left( \frac{w-x_i}{w-m} \right)^{\frac{1}{l}} \right\}. \end{aligned}$$

Οι παραπάνω εξισώσεις χρησιμοποιούνται όταν υπάρχει τουλάχιστον μια παρατήρηση σε καθένα από τα χρονικά διαστήματα. Όταν δεν ισχύει κάτι τέτοιο πρέπει να εκτιμηθούν οι παράμετροι της GIII θεωρώντας ότι υπάρχουν λογοκριμένα δεδομένα. Όπως αναπτύξαμε στην τελευταία παράγραφο του Κεφαλαίου 2 την μεθοδολογία για την εκτίμηση των παραμέτρων της GEV έχοντας μη πλήρη (λογοκριμένα) δεδομένα, θα παρουσιάσουμε και εδώ τις εξισώσεις που χρησιμοποιούνται όταν τα λογοκριμένα δεδομένα προέρχονται από την τρίτη οριακή κατανομή του Gumbel (GIII). Προφανώς στην κατηγορία των σεισμικών δεδομένων η λογοκρισία που παρουσιάζεται είναι τύπου I και μάλιστα αριστερή, δηλαδή λείπουν παρατηρήσεις κάτω από κάποιο κατώφλι. Αυτό συμβαίνει διότι οι σεισμοί μικρότεροι από κάποια ένταση, είτε δεν καταγράφονται όλοι από τα σεισμολογικά κέντρα, είτε δεν εμπεριέχονται σε καταλόγους σεισμών παλαιότερων ετών. Επαναλαμβάνουμε ότι έχοντας  $N$  παρατηρήσεις από την GIII, από τις οποίες  $m$  λογοκρίθηκαν (όσες είναι μικρότερες του  $x_L$ ), η αντίστοιχη log-likelihood θα είναι

$$l(w, m, I) = \ln N! - \ln m! + m \ln F(x_L) + \sum_{i=1}^n \ln f(x_i)$$

Οι εκτιμήσεις των παραμέτρων θα βρεθούν μεγιστοποιώντας την παρακάτω ποσότητα ως προς τις  $w, \mu, \lambda$ ,

$$-mT\left(\frac{w-x_L}{w-m}\right)^{1/I} + \sum_{i=1}^n \left\{ -\ln(I(w-m)) + \left(\frac{1}{I}-1\right) \ln\left(\frac{w-x_i}{w-m}\right) - T\left(\frac{w-x_i}{w-m}\right)^{\frac{1}{I}} \right\}.$$

Οι ποσότητες  $\ln N!$ ,  $\ln m!$  και  $\ln T$  είναι σταθερές, αφού δεν εξαρτώνται από τις παραμέτρους και επομένως δεν επηρεάζουν την μεγιστοποίηση της log-likelihood και άρα μπορούν να εξαλειφθούν από την παράσταση.

**Παρατήρηση.** Ο ερευνητής δεν μπορεί να γνωρίζει επακριβώς ποιο είναι το κατώφλι κάτω από το οποίο δεν καταγράφονται όλοι οι σεισμοί, και αν το ορίσει αυθαίρετα τότε υπάρχει κίνδυνος να πάρει εκτιμήσεις που απέχουν πολύ από τις πραγματικές τιμές των παραμέτρων. Αν το κατώφλι είναι αρκετά μικρό τότε μπορεί μερικά τοπικά μέγιστα να προέρχονται από περιοχή δεδομένων που γίνεται λογοκρισία, ενώ αν είναι μεγάλο τότε θα μειωθούν σημαντικά οι διαθέσιμες (μη-λογοκριμένες) παρατηρήσεις  $n$  με συνέπεια να έχουμε μεγάλες διασπορές στις εκτιμήσεις. Για να αποφευχθούν τέτοιου είδους προβλήματα μπορούμε να βασιστούμε στον γνωστό εμπειρικό κανόνα

Gutenberg-Richter (Gutenberg and Richter, (1954)). Συγκεκριμένα έχει παρατηρηθεί ότι ισχύει (προσεγγιστικά) η σχέση

$$\log_{10} N(M) = a - b \cdot M$$

όπου  $N(M)$  είναι το πλήθος των σεισμών έντασης  $M$  (σε κάποια περιοχή και για κάποιο χρονικό διάστημα) και  $a, b$  κάποιες σταθερές. Δηλαδή, ο λογάριθμος της συχνότητας εμφάνισης των σεισμών μειώνεται γραμμικά με την έντασή τους. Είναι αξιοσημείωτο το γεγονός ότι ο εμπειρικός αυτός κανόνας φαίνεται να είναι ευσταθής χωρικά και χρονικά. Η σταθερά  $b$  είναι συνήθως ίση με 1 ενώ η σταθερά  $a$  επηρεάζεται από την σεισμικότητα της εκάστοτε περιοχής.

Κατασκευάζοντας λοιπόν το γράφημα του λογαρίθμου της συχνότητας ως προς το μέγεθος του σεισμού και παρατηρώντας που χάνεται η γραμμικότητα μπορούμε να συμπεράνουμε εμμέσως ποιο είναι το κατώφλι κάτω από το οποίο έχουμε ελλιπή δεδομένα (δεν έχουν καταγραφεί όλοι οι σεισμοί) και στη συνέχεια να λάβουμε υπόψη στην στατιστική ανάλυση μόνο τους σεισμούς που το υπερβαίνουν.

Αφού βρεθούν τα  $\hat{w}, \hat{m}, \hat{l}$  με την μέθοδο της μέγιστης πιθανοφάνειας για πλήρη ή λογοκριμένα δεδομένα, πρέπει να ελέγξουμε πόσο καλά προσαρμόζονται τα δεδομένα στο μοντέλο που έχουμε επιλέξει. Αυτό μπορεί να γίνει γραφικά με Q-Q plot. Το Q-Q plot των μεγίστων ανά μια χρονική περίοδο ως προς την  $GIII$  με πλήρη δεδομένα θα αποτελείται από τα σημεία

$$(x_{(i)}, GIII^{-1}(\frac{i}{n+1}; \hat{w}, \hat{m}, \hat{l})) = (x_{(i)}, \hat{w} - (\hat{w} - \hat{m})(-\ln \frac{i}{n+1})/T)^{\hat{l}}, i = 1, 2, \dots, n$$

όπου  $x_{(1)} < x_{(2)} < \dots < x_{(n)}$  είναι τα διατεταγμένα τοπικά μέγιστα. Ενώ αν τα δεδομένα είναι λογοκριμένα θα αποτελείται από τα σημεία

$$(x_{(i)}, \hat{w} - (\hat{w} - \hat{m})(-\ln(\frac{i}{n+1} + (1 - \frac{i}{n+1})GIII(x_L)))/T)^{\hat{l}}$$

Αν υπάρχει λοιπόν καλή προσαρμογή τότε τα  $n$  σημεία θα βρίσκονται κοντά στη διαγώνιο.

Όταν εκτιμήσουμε τις παραμέτρους του τρίτου τύπου της οριακής κατανομής *Gumbel* και επιβεβαιώσουμε την καλή προσαρμογή του μοντέλου, είμαστε σε θέση να εκτιμήσουμε τα ποσοστιαία σημεία της κατανομής. Μπορεί να βρεθεί η στάθμη απόδοσης  $z_p$  για  $1/p$  χρονικές περιόδους απόδοσης, δηλαδή το κατώφλι το οποίο η ένταση ενός σεισμού θα ξεπεράσει την επόμενη χρονική περίοδο  $T$ , με πιθανότητα  $p$ . Πιο συγκεκριμένα έχουμε

$$P(X > z_p) = p \Leftrightarrow P(X \leq z_p) = 1 - p$$

$$\Leftrightarrow GIII(z_p) = 1 - p$$

$$\Leftrightarrow z_p = GIII^{-1}(1 - p)$$

Αντιστρέφοντας την  $GIII$  και αντικαθιστώντας τις παραμέτρους της με τις εκτιμήσεις που βρέθηκαν με την μέθοδο μέγιστης πιθανοφάνειας παίρνουμε την εκτίμηση του  $z_p$  που είναι ίση με

$$\hat{z}_p = \hat{w} - (\hat{w} - \hat{m}) \left( \frac{-\ln(1-p)}{T} \right)^{\hat{l}}$$

Όπως αναφέρθηκε στην παράγραφο 2.2.4 υπάρχει η δυνατότητα να βρεθεί προσεγγιστικά η κατανομή που ακολουθεί η  $z_p$  με την με την μέθοδο Δέλτα. Βάση αυτής γνωρίζουμε ότι

$$\hat{z}_p = g(\hat{w}, \hat{m}, \hat{l}) \sim N\left(g(w, m, l), \left(\frac{\partial g}{\partial w}, \frac{\partial g}{\partial m}, \frac{\partial g}{\partial l}\right)^T \cdot \hat{I}^{-1} \cdot \left(\frac{\partial g}{\partial w}, \frac{\partial g}{\partial m}, \frac{\partial g}{\partial l}\right)\right),$$

όπου  $\hat{I}$  είναι η εκτίμηση του πίνακα πληροφορίας.

Στην περίπτωση που τα block maxima ακολουθούν την  $GIII$  δείξαμε ότι

$$g(w, m, l) = w - (w - m) \left( \frac{-\ln(1-p)}{T} \right)^l,$$

επομένως έχουμε

$$\left(\frac{\partial g}{\partial w}, \frac{\partial g}{\partial m}, \frac{\partial g}{\partial l}\right) = \left(1 - \left(\frac{w-p}{T}\right)^l, \left(\frac{w-p}{T}\right)^l, -(w-m) \left(\frac{w-p}{T}\right)^l \ln\left(\frac{w-p}{T}\right)\right),$$

όπου  $w_p = -\ln(1-p)$ .

Αντικαθιστώντας λοιπόν τις παραμέτρους  $w, m, l$  με τις ε.μ.π μπορούμε να εκτιμήσουμε την διασπορά του  $z_p$  και να κατασκευάσουμε (προσεγγιστικά) διαστήματα εμπιστοσύνης για αυτό.

### 3.2 Εμπειρική αξιολόγηση της μεθόδου *Block Maxima* με λογοκριμένες παρατηρήσεις

Πριν προχωρήσουμε στη στατιστική ανάλυση και συμπερασματολογία των σεισμολογικών δεδομένων που έχουμε στην διάθεσή μας, θα ήταν χρήσιμο πρώτα να αξιολογηθούν εμπειρικά οι εκτιμήσεις των παραμέτρων της *Gumbel III* που υπολογίζονται με την μέθοδο της μέγιστης πιθανοφάνειας, στην περίπτωση που έχουμε λογοκριμένα δεδομένα. Το βασικό ερώτημα που θα διερευνήσουμε είναι αν η επαναληπτική μέθοδος που θα χρησιμοποιήσουμε για τον υπολογισμό των εκτιμήσεων των παραμέτρων (μέσω του πακέτου R) είναι αξιόπιστη και επομένως τα συμπεράσματα που θα εξαχθούν βάσει αυτών είναι ασφαλή. Επίσης μας ενδιαφέρει να εξετάσουμε αν

βελτιώνονται ή χειροτερεύουν οι εκτιμήσεις όσο μεγαλώνει η αναλογία των λογοκριμένων δεδομένων σε σχέση με το μέγεθος του δείγματος, δηλαδή όσο μειώνεται το μέγεθος του block.

Για να απαντηθούν εμπειρικά (μέσω προσομοίωσης) τα δύο παραπάνω ερωτήματα ακολουθήθηκε η εξής διαδικασία. Παραγάγαμε 100000 αριθμούς από την *Gumbel*III με συγκεκριμένες παραμέτρους  $\omega = 9$ ,  $\mu = 5$  και  $\lambda = 0.15$  (θεωρώντας π.χ. ότι έχουμε μια παρατήρηση ανά ώρα ή ανά ημέρα) από τους οποίους λογοκρίθηκαν όσοι ήταν μικρότεροι από 6.5 (περίπου το 96% των παρατηρήσεων). Στη συνέχεια κατασκευάσαμε ένα δείγμα μεγέθους  $n = 1000$ , παίρνοντας τα τοπικά μέγιστα ανά 100 παρατηρήσεις. Λαμβάνοντας τα τοπικά μέγιστα ανά 50, 25, 10, 5, 2, 1 παρατηρήσεις δημιουργήσαμε άλλα 6 δείγματα. Τέλος, με βάση την εξίσωση πιθανοφάνειας για λογοκριμένα δεδομένα εκτιμήσαμε (αριθμητικά μέσω του R) τις παραμέτρους της κατανομής σε κάθε μία από τις 7 αυτές περιπτώσεις. Η παραπάνω διαδικασία πραγματοποιήθηκε 1000 φορές και αποθηκεύτηκαν οι εκτιμήσεις των παραμέτρων  $\omega$  και  $\lambda$  που παίρναμε για κάθε 100000 τυχαίους αριθμούς (διαφορετικούς κάθε φορά) από την *Gumbel*III(9,5,0.15) (Παράρτημα Κώδικας 1).

Ένα ερώτημα που ανακύπτει είναι τι κατανομή ακολουθούν τα δείγματα που κατασκευάσαμε όταν το μέγεθος του block είναι μεγαλύτερο από ένα. Είναι σίγουρο ότι θα είναι μια *Gumbel*III αφού ανήκει στην οικογένεια των max-stable κατανομών, αλλά δεν γνωρίζουμε ποιες θα είναι οι καινούριες παράμετροί της κάθε φορά, έτσι ώστε να μπορούμε να αξιολογήσουμε και να συγκρίνουμε τις εκτιμήσεις αυτών. Αποδεικνύεται ότι αν έχουμε  $X_1, \dots, X_k$  ανεξάρτητες τυχαίες μεταβλητές που προέρχονται από την *Gumbel*III με παραμέτρους  $\omega, \mu, \lambda$  τότε η τυχαία μεταβλητή

$$Y = \max\{X_1, \dots, X_k\}$$

ακολουθεί και αυτή την *Gumbel*III( $\omega', \mu', \lambda'$ ) όπου

$$\omega' = \omega$$

$$\mu' = (\alpha \cdot \omega - \omega + \mu) / \alpha$$

$$\lambda' = \lambda$$

και  $\alpha = k^{\lambda}$ . Επομένως η μόνη παράμετρος που αλλάζει τιμή όταν αλλάζει το μέγεθος του block είναι η  $\mu$ .

Η απόδειξη του παραπάνω προκύπτει άμεσα διότι

$$P(Y \leq x) = P(X_1 \leq x, \dots, X_k \leq x) = P(X_1 \leq x) \cdot \dots \cdot P(X_k \leq x)$$

$$= P(X_1 \leq x)^k = e^{-k \left( \frac{w-x}{w-m} \right)^{\frac{1}{\lambda}}} = e^{-\left( \frac{w'-x}{w'-m'} \right)^{\frac{1}{\lambda'}}}$$

αν  $\lambda' = \lambda$ ,  $w' = w$  και  $m' = (aw - w + m)/a$ , με  $\alpha = k^\lambda$

Στον επόμενο πίνακα δίνεται το μέγεθος του αρχικού δείγματος  $n$ , το ποσοστό των παρατηρήσεων που βρέθηκαν μικρότερες του 6.5 (λογοκριμένες) και το μέγεθος του τελικού δείγματος  $n^*$  ανάλογα με το μέγεθος του block. Πρέπει να αναφερθεί ότι οι δύο τελευταίες στήλες αναφέρονται στο μέσο όρο από τις 1000 φορές που διεξάχθηκε το πείραμα.

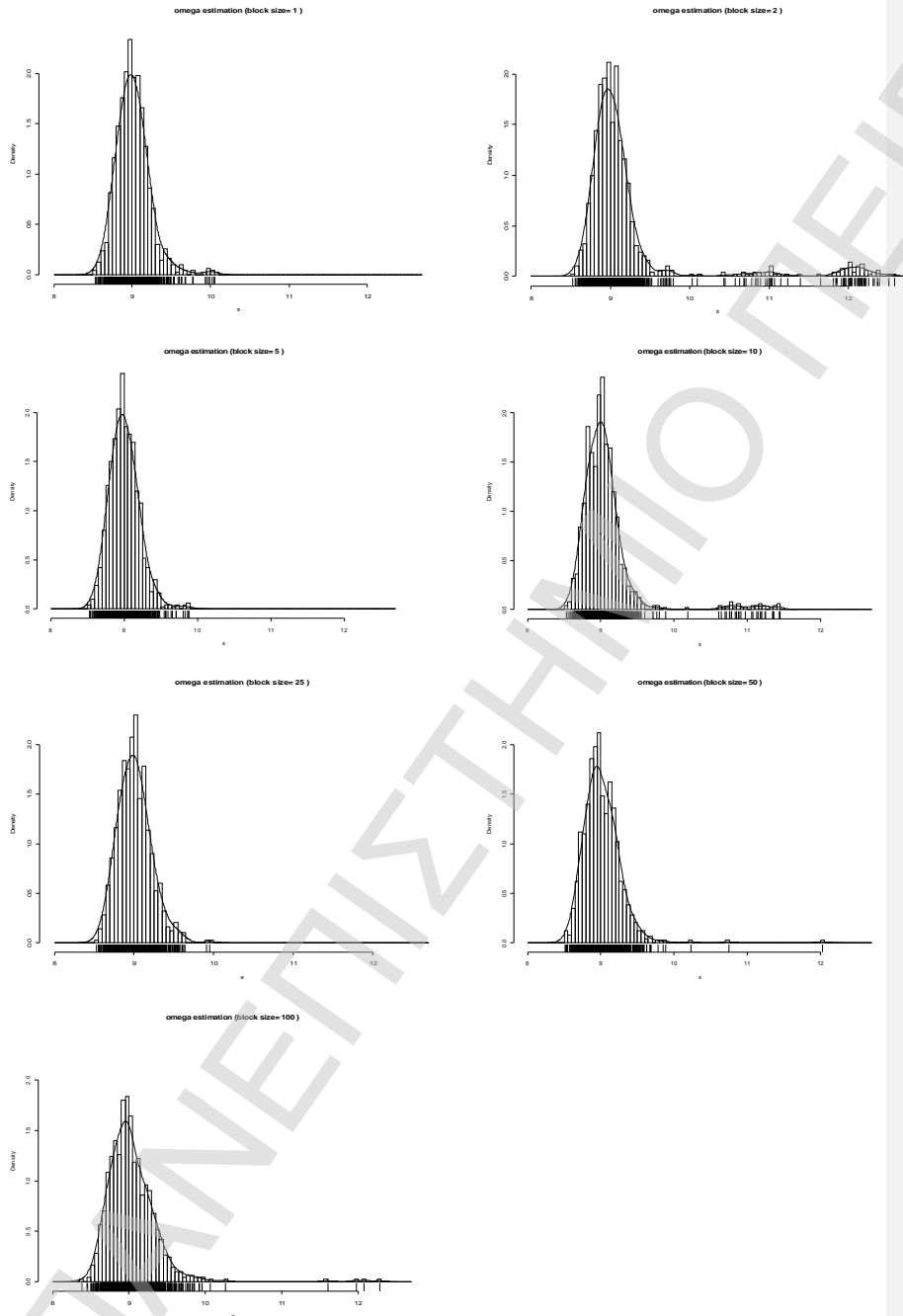
Μέγεθος Block	$n$	Ποσοστό Λογοκριμένων Παρατηρήσεων	$n^*$
1	100000	96%	4000
2	50000	92%	4000
5	20000	80%	4000
10	10000	65%	3500
25	4000	34%	2640
50	2000	11%	1780
100	1000	1%	990

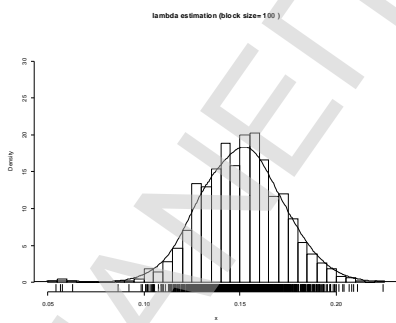
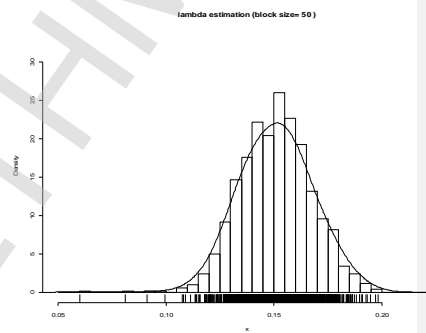
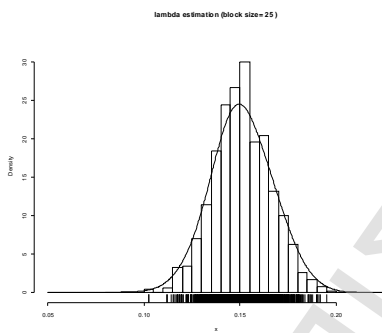
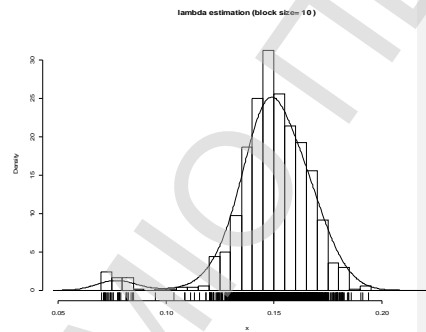
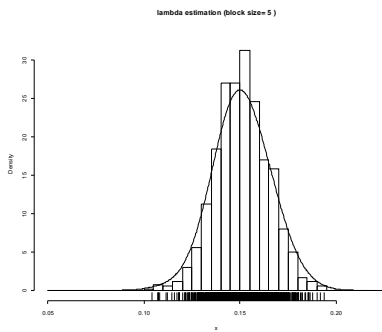
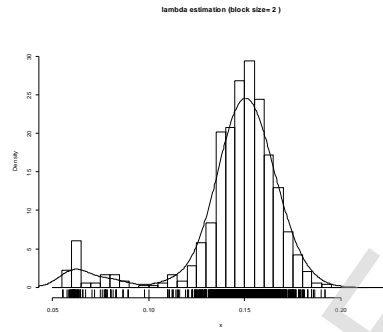
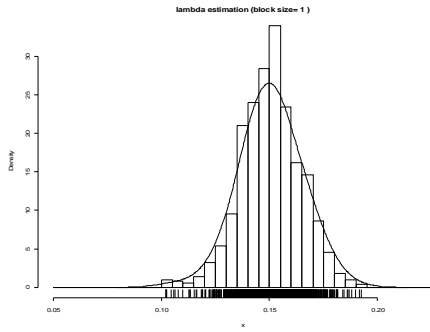
Στη συνέχεια ακολουθούν τα ιστογράμματα των εκτιμήσεων των παραμέτρων  $\omega$  και  $\lambda$  για κάθε περίπτωση από τις 7 (το  $\mu$  δεν συμπεριλήφθη διότι, όπως είδαμε, μεταβάλλεται η τιμή του σε σχέση με το μέγεθος του block). Υπενθυμίζουμε ότι οι πραγματικές τιμές είναι  $\omega = 9$  και  $\lambda = 0.15$



### Σχήμα 3-1

Εκτιμήσεις των παραμέτρων  $\omega$ ,  $\lambda$  για μέγεθος block : 1, 2, 5, 10, 25, 50, 100

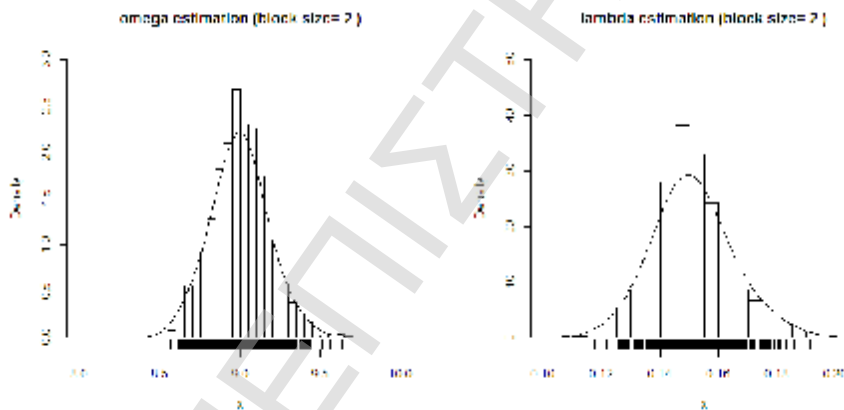




Από τα διαγράμματα παρατηρούμε ότι ο επαναληπτικός αλγόριθμος που χρησιμοποιείται για την εκτίμηση των παραμέτρων της οριακής κατανομής τις περισσότερες φορές προσεγγίζει ικανοποιητικά τις πραγματικές τιμές, ανεξαρτήτως του μεγέθους των block. Προφανώς υπάρχουν περιπτώσεις που αποτυγχάνει, και αυτό είναι πιο εμφανές όταν το μέγεθος των block είναι 2, αλλά και εκεί οι λάθος προσεγγίσεις δεν ξεπερνάνε το 2% του πλήθους των επαναλήψεων της διαδικασίας. Πρέπει να σημειωθεί ότι όταν παίρνουμε λάθος εκτιμήσεις αυτές είναι υπερεκτιμημένες για την παράμετρο  $\omega$  και υποεκτιμημένες για την παράμετρο  $\lambda$ . Στις περιπτώσεις αυτές μάλλον ο αλγόριθμος εγκλωβίζεται σε μια περιοχή όπου υπάρχει τοπικό μέγιστο και για αυτό δίνει εκτιμήσεις μακριά από τις πραγματικές τιμές. Αυτό μπορεί να οφείλεται π.χ. στις αρχικές τιμές που δόθηκαν ώστε να ξεκινήσει η επαναληπτική μέθοδος. Αν επαναλάβουμε την παραπάνω διαδικασία με την μόνη διαφορά ότι οι αρχικές τιμές του αλγορίθμου είναι ίδιες με τις πραγματικές, παίρνουμε τα επόμενα ιστογράμματα για τις εκτιμήσεις των παραμέτρων (Παράρτημα Κώδικας 2).

### Σχήμα 3-2

Εκτιμήσεις των παραμέτρων  $\omega$ ,  $\lambda$  με αρχικές συνθήκες τις πραγματικές τιμές, για μέγεθος block ίσο με 2



Παρατηρούμε ότι σε αυτήν την περίπτωση δεν υπάρχουν ακραίες εκτιμήσεις των  $\omega$ ,  $\lambda$  και αυτό επιβεβαιώνει την προηγούμενη παρατήρηση. Η διαδικασία επαναλήφθηκε για 1000 φορές. Ως γενικό συμπέρασμα μπορούμε να πούμε ότι η συγκεκριμένη επαναληπτική μέθοδος είναι αρκετά αξιόπιστη για οποιοδήποτε ποσοστό λογοκρινόμενων δεδομένων.

Όσον αφορά το δεύτερο ερώτημα που τέθηκε, αν δηλαδή υπάρχει κάποια βελτίωση στις εκτιμήσεις όσο μικραίνει το μέγεθος των block, από τα διαγράμματα φαίνεται ότι όταν τα block έχουν μέγεθος 50 ή 100 τότε οι εκτιμήσεις και του  $\omega$  και του  $\lambda$  έχουν μεγαλύτερες διασπορές από ότι στις άλλες περιπτώσεις. Ένας άλλος τρόπος για να ελέγξουμε πότε παίρνουμε τις καλύτερες εκτιμήσεις είναι να υπολογίσουμε την μέση απόλυτη διαφορά μεταξύ της εκτίμησης και της πραγματικής τιμής κάθε παραμέτρου. Ο δείκτης αυτός υπολογίστηκε σε κάθε περίπτωση και για τις δύο παραμέτρους, με την προϋπόθεση ότι αλγόριθμος δεν αποτυγχάνει στην προσέγγιση των εκτιμητριών (Παράρτημα Κώδικας 3). Τα αποτελέσματα παρουσιάζονται στον επόμενο πίνακα.

### Πίνακας 3-1

Μέση απόλυτη Διαφορά για τις παραμέτρους  $\omega$  και  $\lambda$ .

Μέγεθος block	Μέση Απόλυτη Διαφορά		
	$\omega$	$\lambda$	$n^*$
1	0.01079240	0.1488065	4000
2	0.01151020	0.1524171	4000
5	0.01082843	0.1484100	4000
10	0.01144271	0.1503826	3500
25	0.01168361	0.1571365	2640
50	0.01305818	0.1685797	1780
100	0.01629198	0.1949213	990

Παρατηρώντας τις τιμές του δείκτη, φαίνεται ότι σταθεροποιούνται όταν το μέγεθος του block είναι μικρότερο ή ίσο του 10. Αυτό είναι μια ένδειξη ότι όταν μικραίνουμε τα χρονικά διαστήματα που παίρνουμε τα τοπικά μέγιστα, βελτιώνεται και η ακρίβεια του επαναληπτικού αλγορίθμου, ακόμη και αν αυξάνεται το ποσοστό των λογοκρινμένων παρατηρήσεων. Η βελτίωση αυτή οφείλεται στο ότι μεγαλώνει το δείγμα ( $n^*$ ) βάση του οποίου γίνονται οι εκτιμήσεις. Όμως φαίνεται ότι υπάρχει κάποιο όριο για το μέγεθος των block, κάτω από το οποίο η εκτίμηση δεν βελτιώνεται. Μια στρατηγική λοιπόν για να πάρουμε τις καλύτερες εκτιμήσεις, είναι να μικρύνουμε τα χρονικά διαστήματα τόσο, όσο το  $n^*$  να πάρει την μέγιστη τιμή.

## ΚΕΦΑΛΑΙΟ 4

### Εφαρμογή της θεωρίας ακραίων τιμών σε σεισμολογικά δεδομένα της Ελλάδας

Μορφοποιήθηκε

**4.1 Σεισμολογικά δεδομένα περιόδου 1970-2007** Τα δεδομένα που θα αναλυθούν στην παρούσα παράγραφο αφορούν επιφανειακούς σεισμούς (σεισμοί με εστιακό βάθος μικρότερο από 60 χιλιόμετρα) που συνέβησαν στην περιοχή της Ελλάδας από τον Ιανουάριο του 1970 έως τον Δεκέμβριο του 2007. Ο κατάλογος των σεισμών αυτών βρίσκεται στην ιστοσελίδα του Γεωδυναμικού Ινστιτούτου του Εθνικού Αστεροσκοπείου Αθηνών (<http://www.gein.noa.gr>).

**Μέθοδος Α (δι-τριμηνιαία block maxima με πλήρη δεδομένα).** Τα σεισμικά δεδομένα αρχικά θα αναλυθούν σύμφωνα με την θεωρία που αναπτύχθηκε στο τρίτο κεφάλαιο της εργασίας και θα βρεθούν εκτιμήσεις των παραμέτρων του τρίτου τύπου της οριακής κατανομής του *Gumbel*. Επίσης θα εκτιμηθεί η στάθμη απόδοσης για 10 και 50 χρόνια σε 27 περιοχές της Ελλάδος. Οι περιοχές αυτές έχουν επιφάνεια  $2^\circ \times 3^\circ$ , δηλαδή 2 μοίρες γεωγραφικού πλάτους (Longitude) επί 3 μοίρες γεωγραφικού μήκους (Latitude), ενώ η επιλογή των περιοχών έχει γίνει ανά 1 μοίρα για το γεωγραφικό πλάτος και ανά 1.5 μοίρες για το γεωγραφικό μήκος. Η επιλογή αυτή επιλέχθηκε ώστε να έχουμε όσο το δυνατόν περισσότερες περιοχές με πλήρη δεδομένα, δηλαδή σε κάθε χρονική περίοδο να υπάρχει τουλάχιστον μια παρατήρηση. Οι χρονικές περιόδους κατά τις οποίες πήραμε τα τοπικά μέγιστα είναι είτε δίμηνο, είτε τρίμηνο ανάλογα με το πόσες παρατηρήσεις είχαμε διαθέσιμες σε κάθε περιοχή. Είναι προφανές ότι με βάση την παραπάνω επιλογή των περιοχών, τα κελιά που ορίστηκαν επικαλύπτουν το ένα το άλλο και κάποια δεδομένα θα χρησιμοποιούνται περισσότερο από μια φορά. Είναι ίσως καλύτερο τα συμπεράσματα που θα εξαχθούν να μην αντιστοιχηθούν σε όλη την περιοχή, αφού στην τομή των κελιών δεν θα είναι ξεκάθαρο ποία είναι η στάθμη απόδοσης, αλλά στο κέντρο της, δηλαδή στο σημείο τομής των διαγωνίων του κάθε κελιού.

**Παρατήρηση.** Πρέπει να αναφερθεί ότι για την 3<sup>η</sup> και την 5<sup>η</sup> περιοχή από τις 27, έχουμε μεγαλώσει την επιφάνειά τους κατά μία μοίρα γεωγραφικού μήκους ώστε να έχουμε πλήρη δεδομένα, καθώς και για την 9<sup>η</sup> περιοχή η χρονική περίοδο που συλλέγουμε τα τοπικά μέγιστα είναι 4 μήνες για τον ίδιο λόγο.

**Μέθοδος Β (Peaks Over Threshold).** Στη συνέχεια για τις 27 αυτές περιοχές που επιλέχθηκαν θα αναλυθούν τα δεδομένα σύμφωνα με την θεωρία Peaks Over Threshold. Θα βρεθεί για κάθε μια το κατώφλι  $u$  και θα εκτιμηθούν οι παράμετροι της GPD καθώς και η στάθμη απόδοσης για τόσες παρατηρήσεις, ώστε να αντιστοιχούν σε 10 και 50 χρόνια απόδοσης.

**Μέθοδος Γ (Μηνιαία block maxima με λογοκριμένα δεδομένα).** Επίσης θα χρησιμοποιήσουμε τον συγκεκριμένο κατάλογο σεισμών για να εκτιμήσουμε τις παραμέτρους του τρίτου τύπου της οριακής κατανομής Gumbel με τα τοπικά μέγιστα να είναι μηνιαία, για τις ίδιες 27 περιοχές αλλά και επιπλέον για άλλες 8 περιοχές της Ελλάδος. Στην ανάλυση αυτή, επειδή τα τοπικά μέγιστα είναι ανά μήνα θα χρησιμοποιηθούν οι εξισώσεις που αναφέρονται για λογοκριμένα δεδομένα, αφού δεν έχουν καταγραφεί σεισμοί κάθε μήνα σε καμία από τις 35 περιοχές.

Η σύγκριση των αποτελεσμάτων των τριών παραπάνω αναλύσεων του ίδιου κατάλογου σεισμών θα γίνει με την βοήθεια κοινών διαγραμμάτων απόδοσης για κάθε περιοχή.

#### 4.1.1 Μέθοδος Α - Πλήρη & Gumbel III

Καταρχήν  $α_5$  παρουσιάσουμε τις 27 περιοχές της Ελλάδος που επιλέχθηκαν για την στατιστική ανάλυση του καταλόγου των σεισμών για την περίοδο 1970 έως 2007, τα κέντρα τους και τη χρονική περίοδο που ελήφθησαν τα τοπικά μέγιστα για κάθε μια από αυτές, ώστε να έχουμε στη διάθεσή μας πλήρη δεδομένα. Πρέπει να σημειωθεί ότι έγινε προσπάθεια να διαμεριστεί όλος ο Ελλαδικός χώρος, αλλά τελικά εξαιρέθηκαν οι περιοχές της Μακεδονίας και τα νησιά του ανατολικού Αιγαίου. Σε αυτές τις περιοχές οι καταγεγραμμένοι σεισμοί ήταν λίγοι, και για να έχουμε πλήρη δεδομένα έπρεπε τα τοπικά μέγιστα να είναι ετήσια, έχοντας ως αποτέλεσμα η ανάλυση να γίνει μόνο με 38 παρατηρήσεις και τα συμπεράσματα να μην είναι ασφαλή. Ενώ έχουμε στην διάθεσή μας  $114 = 3 \times 38$ ,  $152 = 4 \times 38$  και  $228 = 6 \times 38$  παρατηρήσεις αν τα τοπικά μέγιστα πάρθηκαν ανά τετράμηνο, τρίμηνο και δίμηνο αντίστοιχα.

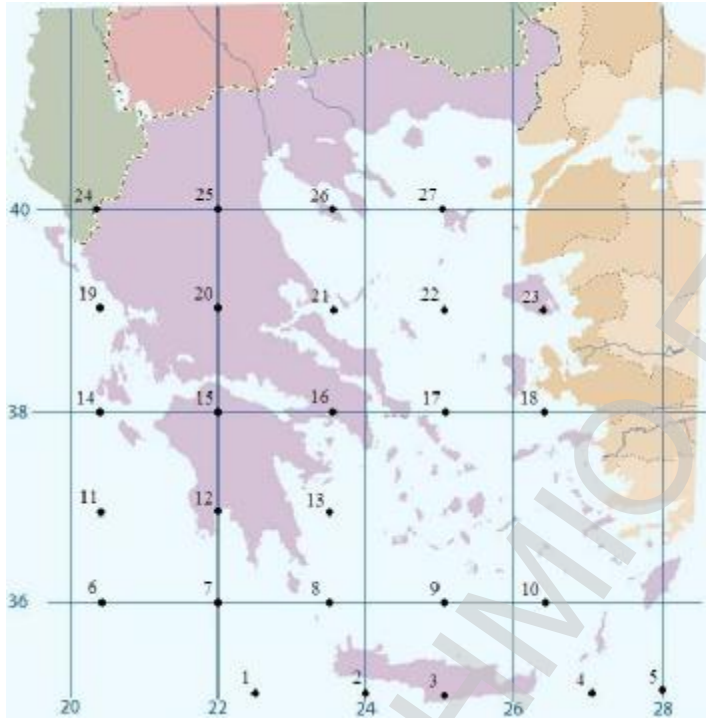
### Πίνακας 4-1

Οι 27 περιοχές της Ελλάδας με τα κέντρα τους και η χρονική περίοδος που πάρθηκαν τα τοπικά μέγιστα.

ΠΕΡΙΟΧΗ	LONGITUDE	LATITUDE	ΚΕΝΤΡΟ		ΧΡΟΝΙΚΗ ΠΕΡΙΟΔΟΣ
			LONG	LAT	
1 <sup>η</sup>	34 <sup>0</sup> – 36 <sup>0</sup>	21 <sup>0</sup> – 24 <sup>0</sup>	35 <sup>0</sup>	22.5 <sup>0</sup>	Τρίμηνο
2 <sup>η</sup>	34 <sup>0</sup> – 36 <sup>0</sup>	22.5 <sup>0</sup> – 25.5 <sup>0</sup>	35 <sup>0</sup>	24 <sup>0</sup>	Τρίμηνο
3 <sup>η</sup>	34 <sup>0</sup> – 36 <sup>0</sup>	23.5 <sup>0</sup> – 27.5 <sup>0</sup>	35 <sup>0</sup>	25 <sup>0</sup>	Τρίμηνο
4 <sup>η</sup>	34 <sup>0</sup> – 36 <sup>0</sup>	25.5 <sup>0</sup> – 28.5 <sup>0</sup>	35 <sup>0</sup>	27 <sup>0</sup>	Τρίμηνο
5 <sup>η</sup>	34 <sup>0</sup> – 36 <sup>0</sup>	26 <sup>0</sup> – 30 <sup>0</sup>	35 <sup>0</sup>	28 <sup>0</sup>	Τρίμηνο
6 <sup>η</sup>	35 <sup>0</sup> – 37 <sup>0</sup>	19 <sup>0</sup> – 22 <sup>0</sup>	36 <sup>0</sup>	20.5 <sup>0</sup>	Τρίμηνο
7 <sup>η</sup>	35 <sup>0</sup> – 37 <sup>0</sup>	20.5 <sup>0</sup> – 23.5 <sup>0</sup>	36 <sup>0</sup>	22 <sup>0</sup>	Τρίμηνο
8 <sup>η</sup>	35 <sup>0</sup> – 37 <sup>0</sup>	22 <sup>0</sup> – 25 <sup>0</sup>	36 <sup>0</sup>	23.5 <sup>0</sup>	Τρίμηνο
9 <sup>η</sup>	35 <sup>0</sup> – 37 <sup>0</sup>	23.5 <sup>0</sup> – 26.5 <sup>0</sup>	36 <sup>0</sup>	25 <sup>0</sup>	Τετράμηνο
10 <sup>η</sup>	35 <sup>0</sup> – 37 <sup>0</sup>	25 <sup>0</sup> – 28 <sup>0</sup>	36 <sup>0</sup>	26.5 <sup>0</sup>	Τρίμηνο
11 <sup>η</sup>	36 <sup>0</sup> – 38 <sup>0</sup>	19 <sup>0</sup> – 22 <sup>0</sup>	37 <sup>0</sup>	20.5 <sup>0</sup>	Δίμηνο
12 <sup>η</sup>	36 <sup>0</sup> – 38 <sup>0</sup>	20.5 <sup>0</sup> – 23.5 <sup>0</sup>	37 <sup>0</sup>	22 <sup>0</sup>	Δίμηνο
13 <sup>η</sup>	36 <sup>0</sup> – 38 <sup>0</sup>	22 <sup>0</sup> – 25 <sup>0</sup>	37 <sup>0</sup>	23.5 <sup>0</sup>	Δίμηνο
14 <sup>η</sup>	37 <sup>0</sup> – 39 <sup>0</sup>	19 <sup>0</sup> – 22 <sup>0</sup>	38 <sup>0</sup>	20.5 <sup>0</sup>	Δίμηνο
15 <sup>η</sup>	37 <sup>0</sup> – 39 <sup>0</sup>	20.5 <sup>0</sup> – 23.5 <sup>0</sup>	38 <sup>0</sup>	22 <sup>0</sup>	Δίμηνο
16 <sup>η</sup>	37 <sup>0</sup> – 39 <sup>0</sup>	22 <sup>0</sup> – 25 <sup>0</sup>	38 <sup>0</sup>	23.5 <sup>0</sup>	Δίμηνο
17 <sup>η</sup>	37 <sup>0</sup> – 39 <sup>0</sup>	23.5 <sup>0</sup> – 26.5 <sup>0</sup>	38 <sup>0</sup>	25 <sup>0</sup>	Δίμηνο
18 <sup>η</sup>	37 <sup>0</sup> – 39 <sup>0</sup>	25 <sup>0</sup> – 28 <sup>0</sup>	38 <sup>0</sup>	26.5 <sup>0</sup>	Τρίμηνο
19 <sup>η</sup>	38 <sup>0</sup> – 40 <sup>0</sup>	19 <sup>0</sup> – 22 <sup>0</sup>	39 <sup>0</sup>	20.5 <sup>0</sup>	Τρίμηνο
20 <sup>η</sup>	38 <sup>0</sup> – 40 <sup>0</sup>	20.5 <sup>0</sup> – 23.5 <sup>0</sup>	39 <sup>0</sup>	22 <sup>0</sup>	Δίμηνο
21 <sup>η</sup>	38 <sup>0</sup> – 40 <sup>0</sup>	22 <sup>0</sup> – 25 <sup>0</sup>	39 <sup>0</sup>	23.5 <sup>0</sup>	Δίμηνο
22 <sup>η</sup>	38 <sup>0</sup> – 40 <sup>0</sup>	23.5 <sup>0</sup> – 26.5 <sup>0</sup>	39 <sup>0</sup>	25 <sup>0</sup>	Δίμηνο
23 <sup>η</sup>	38 <sup>0</sup> – 40 <sup>0</sup>	25 <sup>0</sup> – 28 <sup>0</sup>	39 <sup>0</sup>	26.5 <sup>0</sup>	Τρίμηνο
24 <sup>η</sup>	39 <sup>0</sup> – 41 <sup>0</sup>	19 <sup>0</sup> – 22 <sup>0</sup>	40 <sup>0</sup>	20.5 <sup>0</sup>	Δίμηνο
25 <sup>η</sup>	39 <sup>0</sup> – 41 <sup>0</sup>	20.5 <sup>0</sup> – 23.5 <sup>0</sup>	40 <sup>0</sup>	22 <sup>0</sup>	Τρίμηνο
26 <sup>η</sup>	39 <sup>0</sup> – 41 <sup>0</sup>	22 <sup>0</sup> – 25 <sup>0</sup>	40 <sup>0</sup>	23.5 <sup>0</sup>	Δίμηνο
27 <sup>η</sup>	39 <sup>0</sup> – 41 <sup>0</sup>	23.5 <sup>0</sup> – 26.5 <sup>0</sup>	40 <sup>0</sup>	25 <sup>0</sup>	Δίμηνο

### Σχήμα 4-1

Τα κέντρα των 27 περιοχών



Στη συνέχεια παρατίθενται τα ραβδογράμματα των τοπικών μεγίστων που προέκυψαν για κάθε περιοχή.

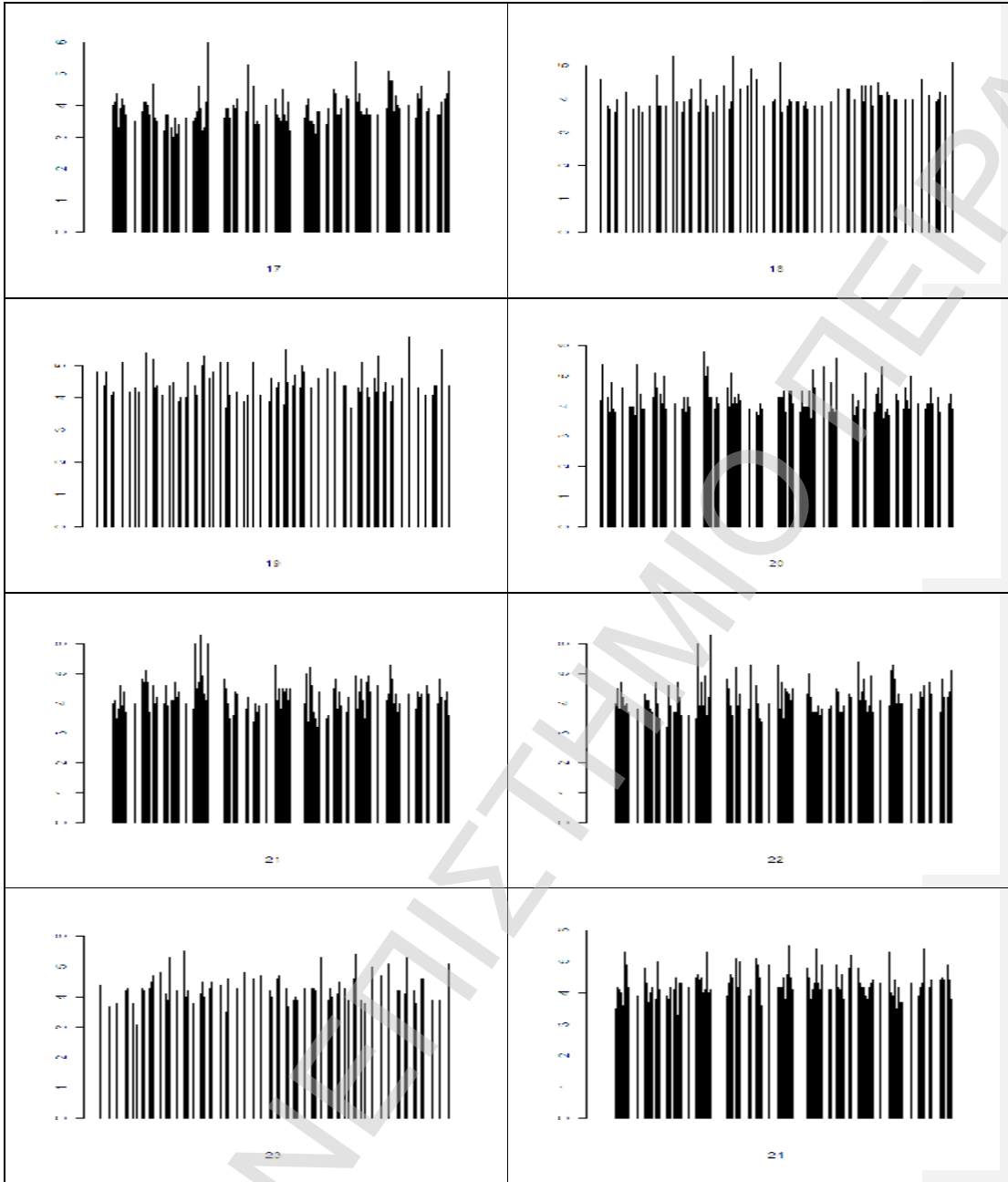


### Σχήμα 4-2

Ραβδογράμματα τοπικών μεγίστων για τις 27 περιοχές









Στον επόμενο πίνακα παρουσιάζονται οι εκτιμήσεις των παραμέτρων μέγιστης πιθανοφάνειας της *Gumbel* III και τα τυπικά σφάλματα αυτών για τα δεδομένα που παρουσιάστηκαν μέσω των προηγούμενων γραφημάτων. Επίσης εκτιμήθηκε η στάθμη απόδοσης για 10 και 50 χρόνια απόδοσης και υπολογίστηκε η διασπορά της  $\hat{z}_{p,50}$  με την μέθοδο Δέλτα (Παράρτημα Κώδικας 4 και 5).

### Πίνακας 4-2

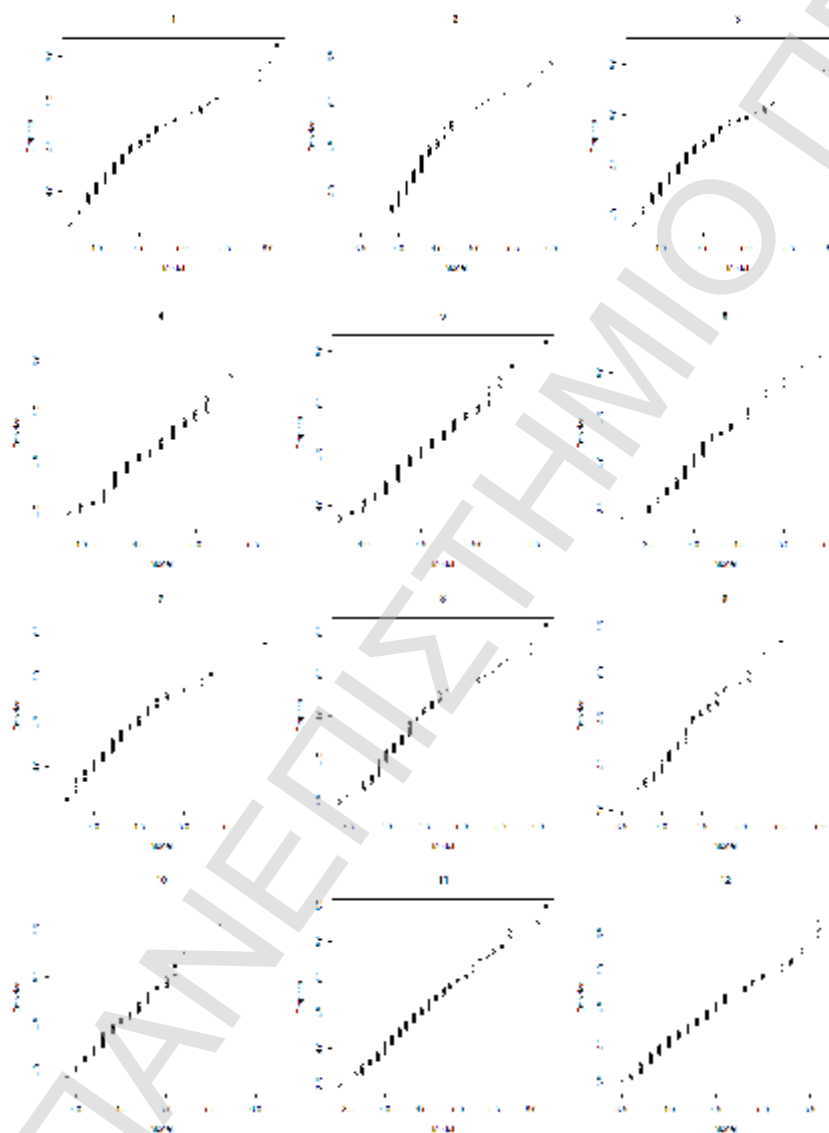
Εκτιμήσεις παραμέτρων GIII, στάθμης απόδοσης για 10 και 50 χρόνια απόδοσης και τυπικά σφάλματα.

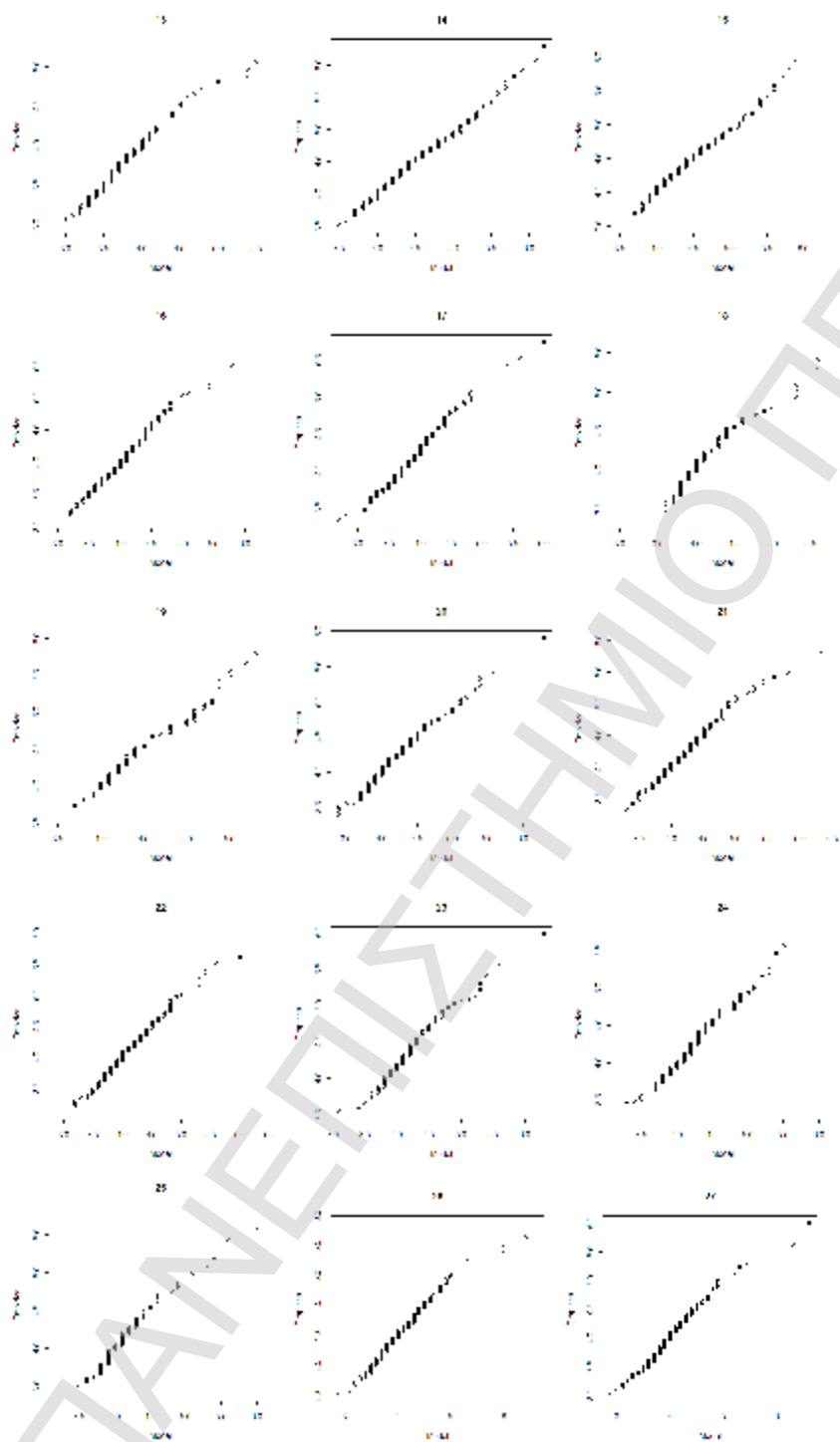
	$\hat{w}$	s.e.( $\hat{w}$ )	$\hat{h}$	s.e.( $\hat{h}$ )	$\hat{l}$	s.e.( $\hat{l}$ )	$\hat{z}_{p10}$	$\hat{z}_{p50}$	s.e.( $\hat{z}_{p50}$ ).
1	13.294	2.8042	4.601	0.0452	0.03627	0.0115	<b>5.30</b>	<b>5.77</b>	0.1168
2	11.825	2.7092	4.634	0.0414	0.04203	0.0150	<b>5.25</b>	<b>5.69</b>	0.1178
3	13.372	3.3111	4.760	0.0376	0.03090	0.0114	<b>5.34</b>	<b>5.74</b>	0.1032
4	11.393	3.9912	4.767	0.0381	0.04113	0.0234	<b>5.36</b>	<b>5.75</b>	0.1375
5	8.0908	2.1658	4.748	0.0371	0.07928	0.0474	<b>5.30</b>	<b>5.63</b>	0.1847
6	10.087	3.5154	4.333	0.0418	0.05258	0.0301	<b>4.98</b>	<b>5.40</b>	0.1687
7	14.722	3.7393	4.533	0.0424	0.02936	0.0104	<b>5.19</b>	<b>5.63</b>	0.1141
8	15.752	3.9744	4.469	0.0445	0.02830	0.0095	<b>5.17</b>	<b>5.65</b>	0.1195
9	11.841	4.4546	4.455	0.0451	0.04242	0.0246	<b>5.13</b>	<b>5.58</b>	0.1499
10	11.335	3.5300	4.767	0.0411	0.04440	0.0227	<b>5.40</b>	<b>5.81</b>	0.1368
11	12.448	2.8356	4.806	0.0470	0.04637	0.0160	<b>5.57</b>	<b>6.07</b>	0.1434
12	11.445	2.6320	4.731	0.0456	0.05063	0.0185	<b>5.46</b>	<b>5.93</b>	0.1439
13	10.717	2.3184	4.195	0.0419	0.04874	0.0161	<b>4.88</b>	<b>5.32</b>	0.1266
14	12.283	3.4100	5.001	0.0520	0.05328	0.0230	<b>5.83</b>	<b>6.37</b>	0.1915
15	12.227	3.0936	4.925	0.0499	0.05091	0.0201	<b>5.73</b>	<b>6.24</b>	0.1664
16	12.262	3.2131	4.574	0.0512	0.05039	0.0194	<b>5.41</b>	<b>5.94</b>	0.1747
17	10.067	2.8058	4.377	0.0497	0.06607	0.0293	<b>5.17</b>	<b>5.67</b>	0.2179
18	12.775	3.7283	4.464	0.0477	0.04174	0.0177	<b>5.22</b>	<b>5.71</b>	0.1483
19	12.070	4.5319	4.783	0.0517	0.05031	0.0297	<b>5.57</b>	<b>6.08</b>	0.1976
20	11.483	2.4602	4.732	0.0470	0.05231	0.0177	<b>5.49</b>	<b>5.98</b>	0.1448
21	12.493	3.2938	4.726	0.0524	0.05086	0.0200	<b>5.58</b>	<b>6.12</b>	0.1779
22	12.004	3.3601	4.667	0.0529	0.05475	0.0229	<b>5.53</b>	<b>6.08</b>	0.2010
23	12.238	4.011	4.597	0.0527	0.049778	0.0246	<b>5.42</b>	<b>5.94</b>	0.1849
24	7.7085	1.0933	4.743	0.0435	0.11097	0.0351	<b>5.41</b>	<b>5.78</b>	0.1857
25	11.819	3.5934	4.572	0.0497	0.04955	0.0232	<b>5.35</b>	<b>5.85</b>	0.1687
26	14.063	3.0876	4.546	0.0527	0.04193	0.0127	<b>5.41</b>	<b>5.98</b>	0.1469
27	14.655	3.5345	4.706	0.0607	0.036918	0.0124	<b>5.51</b>	<b>6.04</b>	0.1559

Ο έλεγχος για το αν τα δεδομένα προσαρμόζονται καλά στο μοντέλο θα γίνει μέσω Q-Q plots. Κατασκευάζουμε λοιπόν τα επόμενα γραφήματα (Παράρτημα Κώδικας 6), από τα οποία κρίνεται ότι η προσαρμογή των δεδομένων είναι μάλλον ικανοποιητική σε όλες τις περιοχές που έχουν επιλεγεί.

#### Σχήμα 4-2

Q-Q plots της GIII για την Μέθοδο Α





Με βάση τα παραπάνω Q-Q plots μπορούμε να πούμε ότι τα δεδομένα (block maxima) προσαρμόζονται ικανοποιητικά στην GumbelIII. Επομένως οι εκτιμήσεις των παραμέτρων μπορούν να χρησιμοποιηθούν ώστε να εξάγουμε ασφαλή συμπεράσματα μέσω της εκτίμησης της στάθμης απόδοσης. Για παράδειγμα, από την εκτίμηση του  $z_{50}$  για την 19<sup>η</sup> περιοχή συμπεραίνουμε ότι ένας σεισμός με ένταση πάνω από 6.08 βαθμούς της κλίμακας Richter εμφανίζεται σε αυτή την περιοχή κατά μέσο όρο κάθε 50 χρόνια.

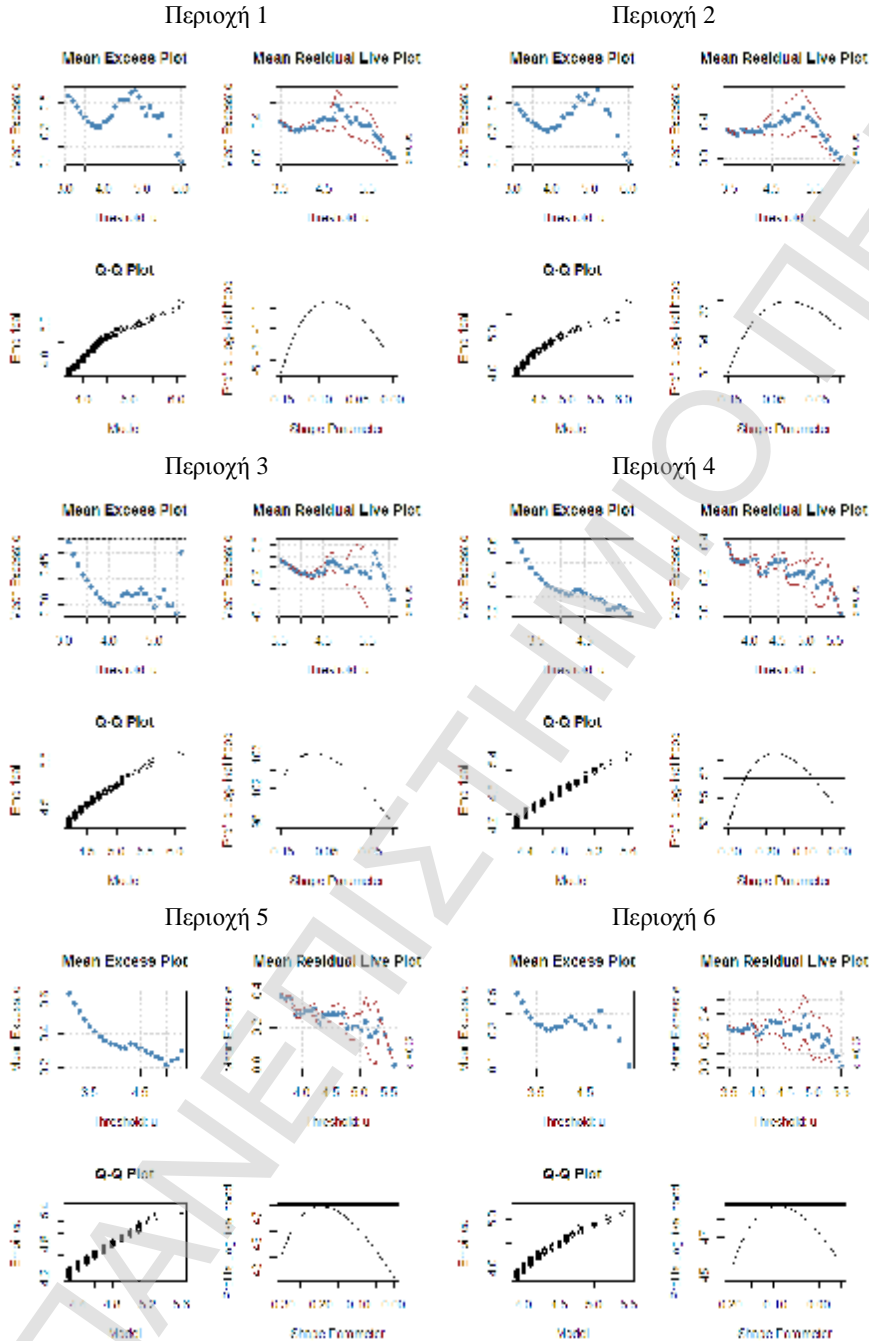
#### 4.1.2 Μέθοδος B - Peaks Over Threshold

Σύμφωνα με την μέθοδο POT το πρώτο που πρέπει να επιλέξουμε είναι το κατώφλι  $u$ , πάνω από το οποίο γνωρίζουμε ότι (προσεγγιστικά) οι παρατηρήσεις θα ακολουθούν την *Γενικευμένη Κατανομή Pareto*, και στη συνέχεια να εκτιμήσουμε τις παραμέτρους της. Η επιλογή του κατώφλιού μπορεί να γίνει με τρεις γραφικές μεθόδους όπως αναπτύχθηκε στην θεωρία (βλ. παράγραφο 2.2.5). Στην συγκεκριμένη εργασία η επιλογή του  $u$  έγινε με την βοήθεια της *εμπειρικής μέσης υπερβάλλουσας συνάρτησης* (Mean Residual Live Plot). Πρέπει να σημειωθεί ότι μετά την πρώτη επιλογή του  $u$  από τα γραφήματα και την προσαρμογή των δεδομένων στο μοντέλο, σε ορισμένες περιοχές επαναπροσδιορίστηκε το κατώφλι υπέρβασης. Αυτό έγινε διότι οι παρατηρήσεις που ξεπερνούσαν το αρχικό  $u$  ήταν ελάχιστες με αποτέλεσμα οι εκτιμήτριες να παρουσιάζουν μεγάλη διακύμανση. Για κάθε μία περιοχή κατασκευάστηκε Q-Q plot ώστε να ελεγχθεί κατά πόσο οι σεισμοί που υπερβαίνουν το  $u$  ακολουθούν την *GPD*. Επίσης κατασκευάστηκε το γράφημα της Profile Deviance Function της παραμέτρου  $\zeta$  και ορίστηκε γραφικά το διάστημα εμπιστοσύνης της συντελεστού 95% (Παράρτημα Κώδικας 7). Ακολουθούν τα παραπάνω 4 γραφήματα για κάθε μια από τις 27 περιοχές της Ελλάδας.

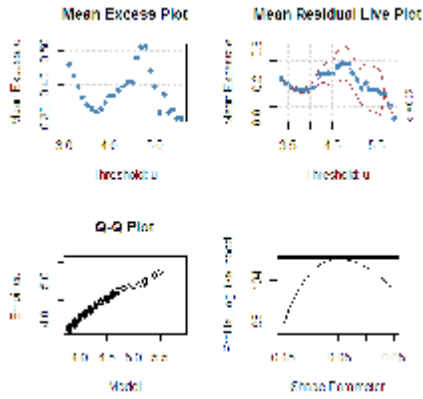


### Σχήμα 4-3

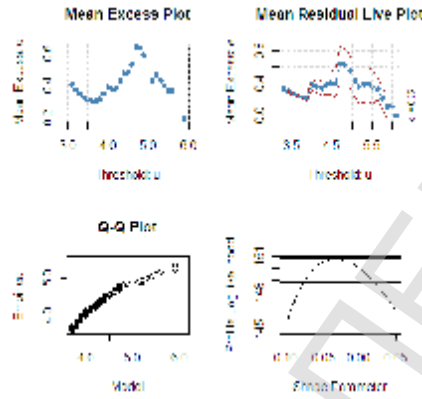
Mean Excess Plot, Mean Residual Live Plot, Q-Q Plot, διάστημα εμπιστοσύνης για την παράμετρο  $\xi$  με την Profile Deviance Function, για τις 27 περιοχές της Ελλάδας



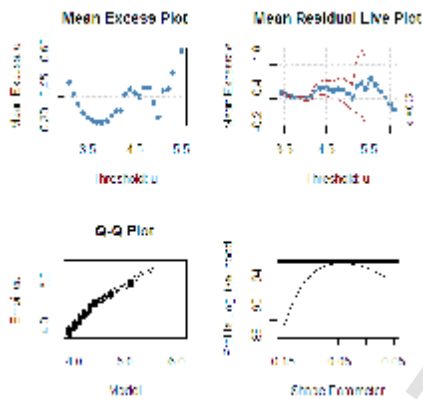
Περιοχή 7



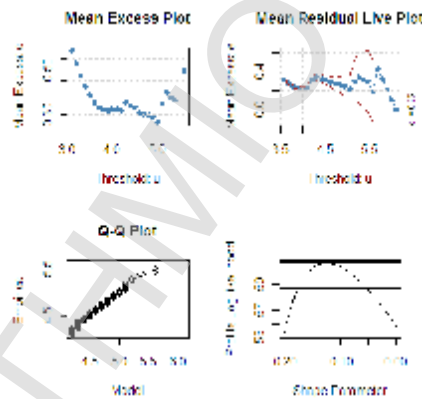
Περιοχή 8



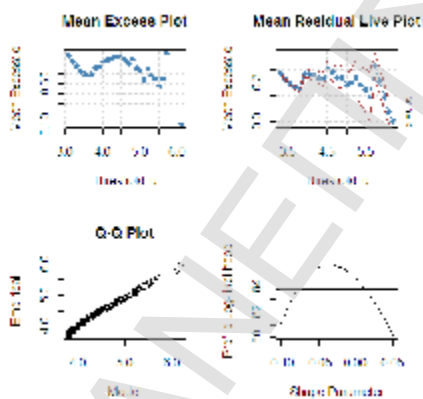
Περιοχή 9



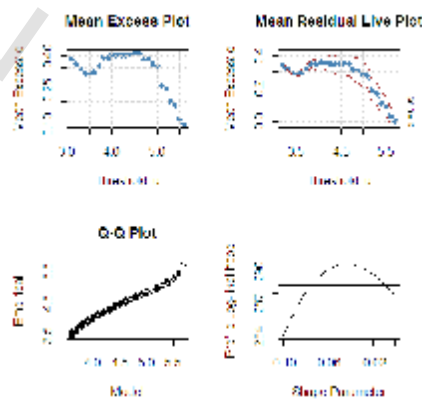
Περιοχή 10



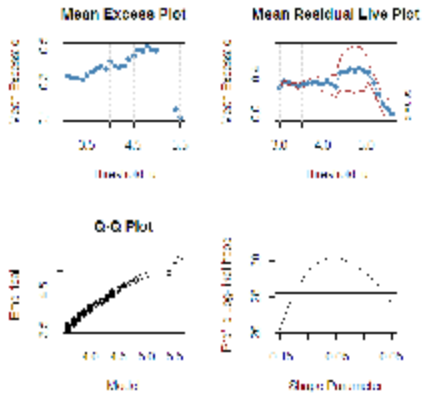
Περιοχή 11



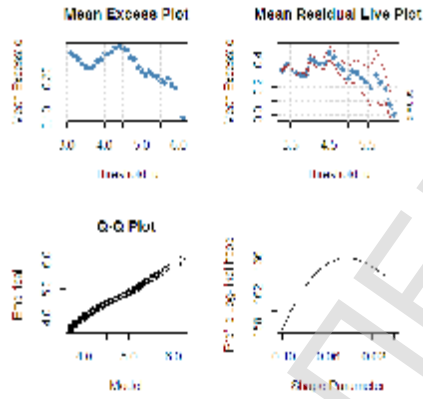
Περιοχή 12



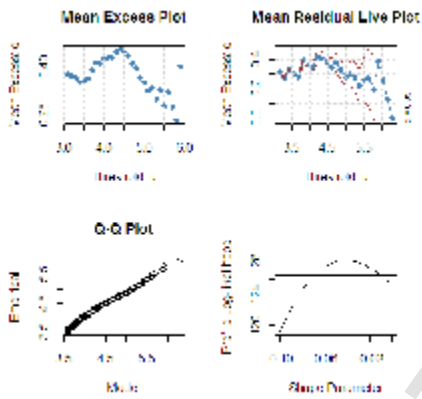
Περιοχή 13



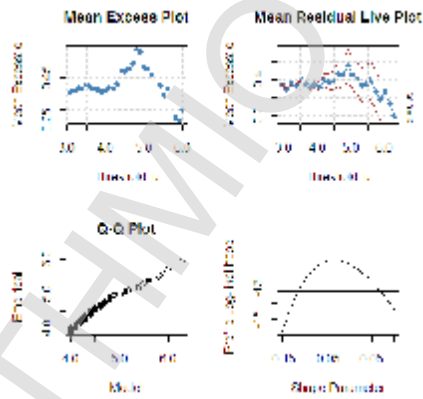
Περιοχή 14



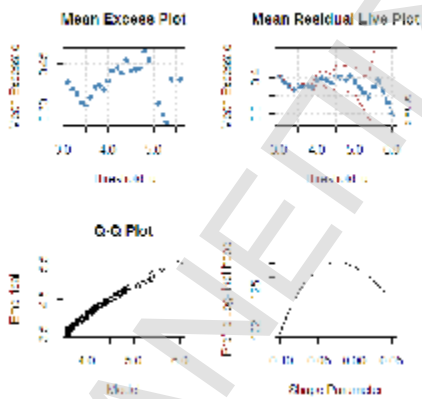
Περιοχή 15



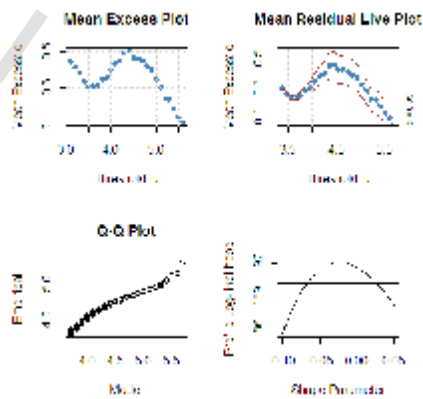
Περιοχή 16



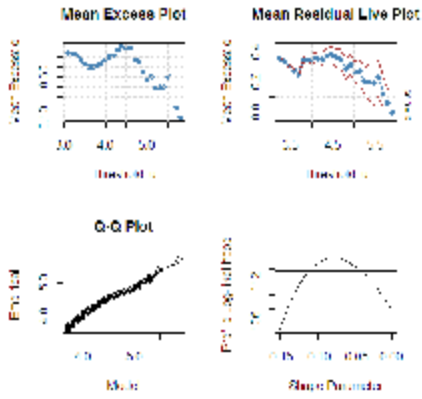
Περιοχή 17



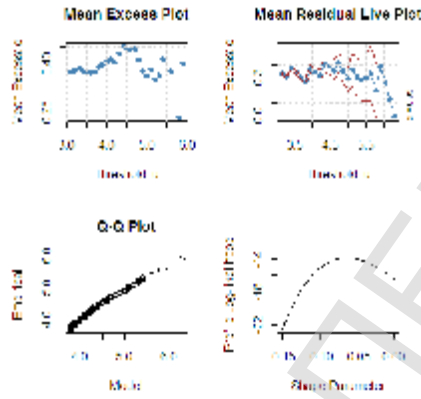
Περιοχή 18



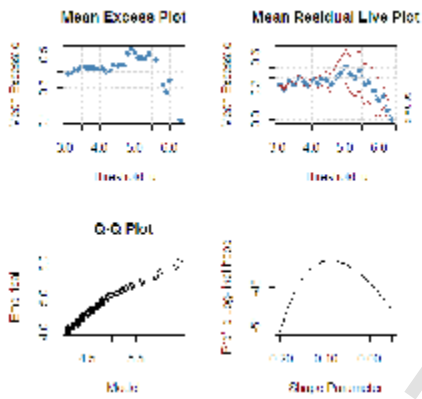
Περιοχή 19



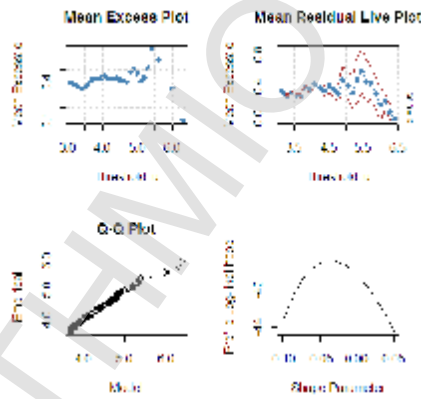
Περιοχή 20



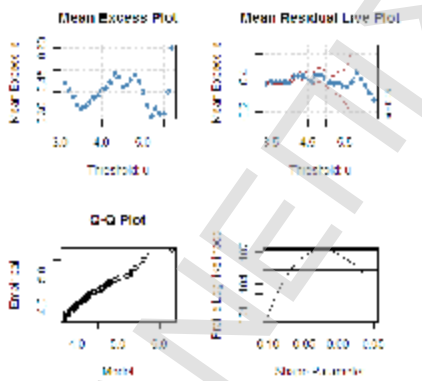
Περιοχή 21



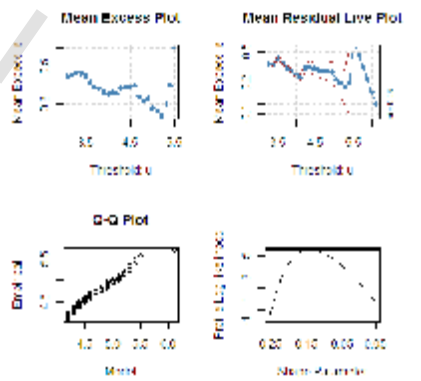
Περιοχή 22



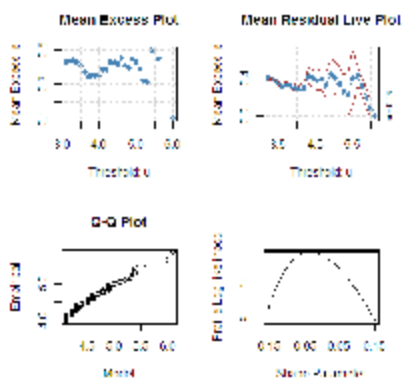
Περιοχή 23



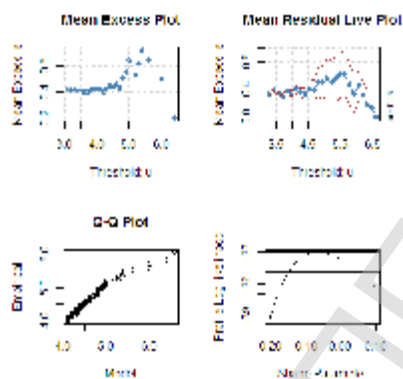
Περιοχή 24



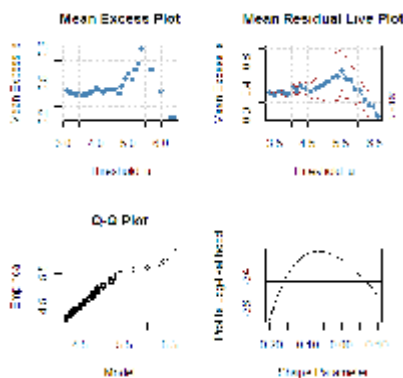
Περιοχή 25



Περιοχή 26



Περιοχή 27



Στον επόμενο πίνακα παρουσιάζονται για κάθε περιοχή το κατώφλι  $u$ , οι σημειακές εκτιμήσεις των παραμέτρων  $\xi$ ,  $\zeta$  με τις τυπικές αποκλίσεις τους και η στάθμη απόδοσης για  $m_1$ ,  $m_2$  παρατηρήσεις ώστε να αντιστοιχούν σε 10 και 50 χρόνια απόδοσης. Για να γίνει η παραπάνω αντιστοιχία μετρήθηκαν οι  $n$  καταγεγραμμένοι σεισμοί στην περιοχή για το χρονικό διάστημα 1970 με 2007 και διαιρέθηκαν με το 38, ώστε να υπολογίσουμε πόσοι σεισμοί συνέβησαν κατά μέσο όρο κάθε χρόνο στην περιοχή και να αναγάγουμε τις  $n/38$  παρατηρήσεις σε χρονική διάρκεια ενός έτους. Στην συνέχεια πολλαπλασιάστηκε το αποτέλεσμα της διαίρεσης με 10 και 50 για να εκτιμήσουμε την στάθμη απόδοσης για 10 και 50 χρόνια. Στην τελευταία στήλη δίνεται η διασπορά της στάθμης απόδοσης για 50 χρόνια που υπολογίστηκε σύμφωνα με την Μέθοδο Δέλτα (Παράρτημα Κώδικας 8).

### Πίνακας 4-3

Το κατώφλι  $u$ , οι εκτιμήσεις των παραμέτρων της GPD, της στάθμης απόδοσης για 10 και 50 χρόνια απόδοσης και τα τυπικά σφάλματα.

	$u$	$\hat{S}$	s.e.( $\hat{S}$ )	$\hat{x}$	s.e.( $\hat{x}$ )	$\hat{x}_{m_1}$	$\hat{x}_{m_2}$	s.e.( $\hat{x}_{m_2}$ )
1	3.6	0.3995	0.0157	-0.0843	0.021	<b>5.34</b>	<b>5.72</b>	0.1102
2	4.0	0.3312	0.0237	-0.0170	0.0437	<b>5.40</b>	<b>5.89</b>	0.1912
3	4.1	0.3183	0.0181	-0.0777	0.0340	<b>5.38</b>	<b>5.71</b>	0.1210
4	4.2	0.3617	0.0263	-0.1751	0.0460	<b>5.30</b>	<b>5.54</b>	0.0994
5	4.2	0.3755	0.0280	-0.1952	0.0456	<b>5.28</b>	<b>5.51</b>	0.0911
6	3.8	0.3364	0.0259	-0.0898	0.0477	<b>4.99</b>	<b>5.33</b>	0.1416
7	3.7	0.3326	0.0154	-0.0470	0.0278	<b>5.24</b>	<b>5.65</b>	0.1363
8	3.6	0.3158	0.0133	-0.0305	0.0239	<b>5.16</b>	<b>5.58</b>	0.1287
9	3.8	0.3029	0.0183	-0.0434	0.0343	<b>5.07</b>	<b>5.46</b>	0.1346
10	4.1	0.3615	0.0206	-0.1242	0.0311	<b>5.39</b>	<b>5.68</b>	0.0989
11	3.7	0.3581	0.0127	-0.0358	0.0231	<b>5.59</b>	<b>6.05</b>	0.1506
12	3.5	0.3376	0.0091	-0.0434	0.0169	<b>5.38</b>	<b>5.78</b>	0.1139
13	3.5	0.3589	0.0233	-0.0542	0.0402	<b>4.96</b>	<b>5.39</b>	0.1682
14	3.6	0.3622	0.0087	-0.0419	0.0152	<b>5.69</b>	<b>6.12</b>	0.1177
15	3.5	0.3671	0.0087	-0.0410	0.0150	<b>5.64</b>	<b>6.07</b>	0.1190
16	3.9	0.3938	0.0285	-0.0333	0.0462	<b>5.52</b>	<b>6.05</b>	0.2200
17	3.5	0.3256	0.0141	-0.0226	0.0271	<b>5.16</b>	<b>5.62</b>	0.1549
18	3.6	0.3139	0.0122	-0.0233	0.0238	<b>5.25</b>	<b>5.69</b>	0.1387
19	3.6	0.3739	0.0108	-0.0779	0.0176	<b>5.44</b>	<b>5.78</b>	0.1031
20	3.7	0.3968	0.0158	-0.0623	0.0252	<b>5.57</b>	<b>6.00</b>	0.1469
21	4.0	0.4567	0.0304	-0.0869	0.0405	<b>5.71</b>	<b>6.17</b>	0.1803
22	3.6	0.3980	0.0170	-0.0358	0.0276	<b>5.59</b>	<b>6.01</b>	0.1812
23	3.6	0.3980	0.0170	-0.0358	0.0276	<b>5.63</b>	<b>6.14</b>	0.1569
24	4.1	0.4125	0.0313	-0.1352	0.0456	<b>5.43</b>	<b>5.78</b>	0.1338
25	4.0	0.3615	0.0297	-0.0307	0.0548	<b>5.43</b>	<b>5.93</b>	0.2230
26	4.0	0.4538	0.0378	-0.0620	0.0499	<b>5.63</b>	<b>6.17</b>	0.2179
27	4.1	0.4698	0.0440	-0.0648	0.0570	<b>5.70</b>	<b>6.26</b>	0.2353

Από τα Q-Q διαγράμματα καταλήγουμε στο συμπέρασμα ότι τα δεδομένα προσαρμόζονται αρκετά καλά στο μοντέλο και μπορούμε να υποθέσουμε με ασφάλεια ότι, σε κάθε περιοχή, οι σεισμοί που υπερβαίνουν το  $\mu$  ακολουθούν την *GPD*. Επίσης παρατηρούμε ότι όλες οι σημειακές εκτιμήσεις της παραμέτρου  $\zeta$  είναι μικρότερες του μηδέν, και στο 95% διάστημα εμπιστοσύνης της, για τις μισές περιοχές δεν συμπεριλαμβάνεται το μηδέν. Αυτό αποτελεί μια σοβαρή ένδειξη ότι τα κανονικοποιημένα μέγιστα των σεισμικών δεδομένων ακολουθούν την κατανομή *Weibull* όπως ορίστηκε στο θεώρημα των *Fisher-Tippett*. Αυτό έρχεται σε συμφωνία με το γεγονός ότι χρησιμοποιούμε την *Gumbel III* στην αρχική μας ανάλυση, αφού αυτή η οριακή κατανομή ισοδυναμεί με την *Weibull*.

#### 4.1.3 Μέθοδος Γ - Λογοκρμμένα δεδομένα & *Gumbel III*

Η επόμενη ανάλυση των δεδομένων, όπως αναφέρθηκε και στην εισαγωγή του κεφαλαίου, θα γίνει πάλι με τον τρίτο τύπο της οριακής κατανομής *Gumbel* με την διαφορά ότι τώρα θα συλλέξουμε τα τοπικά μέγιστα ανά μήνα και θα επεκτείνουμε την ανάλυση και σε άλλες 8 περιοχές της Ελλάδας οι οποίες παρουσιάζονται στον επόμενο πίνακα, οι οποίες συμπληρώνουν τον Πίνακα 4-1.

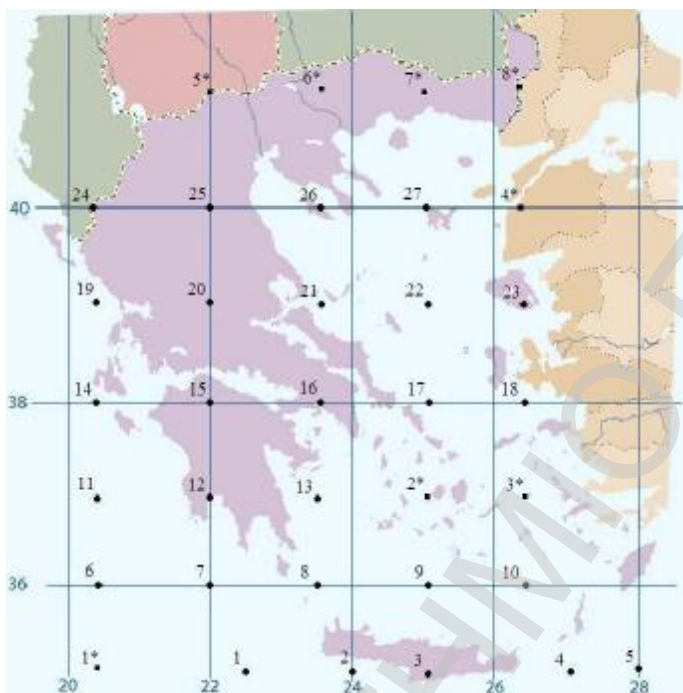
#### Πίνακας 4-4

Οι 8 επιπλέον περιοχές για την Μέθοδο Γ.

ΠΕΡΙΟΧΗ	LONGITUDE	LATITUDE	ΚΕΝΤΡΟ		ΧΡΟΝΙΚΗ ΠΕΡΙΟΔΟΣ
			LONG	LAT	
1* <sup>η</sup>	34 <sup>0</sup> – 36 <sup>0</sup>	19 <sup>0</sup> – 22 <sup>0</sup>	35 <sup>0</sup>	20.5 <sup>0</sup>	Μήνας
2* <sup>η</sup>	36 <sup>0</sup> – 38 <sup>0</sup>	23.5 <sup>0</sup> – 26.5 <sup>0</sup>	37 <sup>0</sup>	25 <sup>0</sup>	Μήνας
3* <sup>η</sup>	36 <sup>0</sup> – 38 <sup>0</sup>	25 <sup>0</sup> – 28 <sup>0</sup>	37 <sup>0</sup>	26.5 <sup>0</sup>	Μήνας
4* <sup>η</sup>	39 <sup>0</sup> – 41 <sup>0</sup>	25 <sup>0</sup> – 28 <sup>0</sup>	40 <sup>0</sup>	26.5 <sup>0</sup>	Μήνας
5* <sup>η</sup>	40 <sup>0</sup> – 42 <sup>0</sup>	20.5 <sup>0</sup> – 23.5 <sup>0</sup>	41 <sup>0</sup>	22 <sup>0</sup>	Μήνας
6* <sup>η</sup>	40 <sup>0</sup> – 42 <sup>0</sup>	22 <sup>0</sup> – 25 <sup>0</sup>	41 <sup>0</sup>	23.5 <sup>0</sup>	Μήνας
7* <sup>η</sup>	40 <sup>0</sup> – 42 <sup>0</sup>	23.5 <sup>0</sup> – 26.5 <sup>0</sup>	41 <sup>0</sup>	25 <sup>0</sup>	Μήνας
8* <sup>η</sup>	40 <sup>0</sup> – 42 <sup>0</sup>	25 <sup>0</sup> – 28 <sup>0</sup>	41 <sup>0</sup>	26.5 <sup>0</sup>	Μήνας

#### Σχήμα 4-4

Τα κέντρα των 35 περιοχών



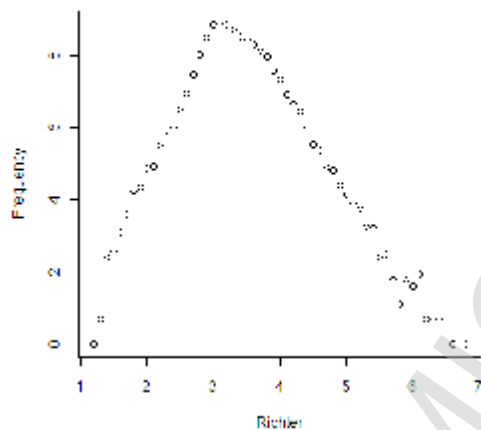
Πριν προχωρήσουμε στην ανάλυση πρέπει να επεξεργαστούμε τα δεδομένα έτσι ώστε να έχουμε έναν πλήρη κατάλογο σεισμών, δηλαδή να βρούμε το όριο πάνω από το οποίο έχουν καταγραφεί όλοι οι σεισμοί που έχουν συμβεί και να κρατήσουμε αυτούς που το υπερβαίνουν. Αυτή η διαδικασία είναι απαραίτητη για να μην υπάρχει αμφιβολία αν κάποιο τοπικό μέγιστο υπάρχει ή όχι. Παραδείγματος χάριν αν για τον Μάρτιο του 1978 βρούμε ότι ο μεγαλύτερος σεισμός ήταν έντασης 2.7 βαθμών Richter πρέπει να είμαστε σίγουροι ότι έχουν καταγραφεί όλοι οι σεισμοί της τάξεως του 2.7 για να τον χρησιμοποιήσουμε ως παρατήρηση, διαφορετικά εκείνη την χρονική περίοδο θα σημειώσουμε ότι έχουμε λογοκρισία. Για να βρούμε αυτό το όριο θα βασιστούμε στον εμπειρικό κανόνα Gutenberg and Richter που αναφέρθηκε παραπάνω (βλ. παράγραφο 3.1) σύμφωνα με τον οποίο η συχνότητα εμφάνισης των σεισμών σε λογαριθμική κλίμακα είναι αντιστρόφως ανάλογη με την έντασή τους. Κατασκευάζουμε λοιπόν το γράφημα «συχνότητα – μέγεθος» και στο σημείο που φαίνεται να



παραβιάζεται ο παραπάνω κανόνας τοποθετούμε το όριο, το οποίο θα μας επιτρέψει να θεωρήσουμε ότι ο κατάλογος των σεισμών που διαχειριζόμαστε να είναι πλήρης.

#### Σχήμα 4-5

Διάγραμμα Gutenberg - Richter για το πρώτο σύνολο δεδομένων



Από το γράφημα παρατηρούμε ότι το όριο που ζητάμε είναι περίπου 3.6 βαθμοί της κλίμακας Richter και αφορά τους σεισμούς σε όλη την Ελλάδα.

Βελτιστοποιώντας την συνάρτηση πιθανοφάνειας με την βοήθεια επαναληπτικών μεθόδων (βλ. παράγραφο 3.1) βρίσκουμε τις εκτιμήσεις των παραμέτρων της οριακής κατανομής *Gumbel III* και τα τυπικά σφάλματα για κάθε μια (Παράρτημα Κώδικας 9). Στη συνέχεια υπολογίζουμε την εκτίμηση της στάθμης απόδοσης για 10 και 50 χρόνια απόδοσης και την διασπορά της  $zp_{50}$  με την Μέθοδο Δέλτα (Παράρτημα Κώδικας 5). Τα αποτελέσματα αυτά συνοψίζονται στον Πίνακα 5. Στην τελευταία στήλη του έχει σημειωθεί το πλήθος των λογοκριμένων παρατηρήσεων,  $m$ , που είχαμε σε κάθε περιοχή από τις συνολικά  $456 = 12 \times 38$  παρατηρήσεις (μηνιαία μέγιστα για 38 έτη).

### Πίνακας 4-5

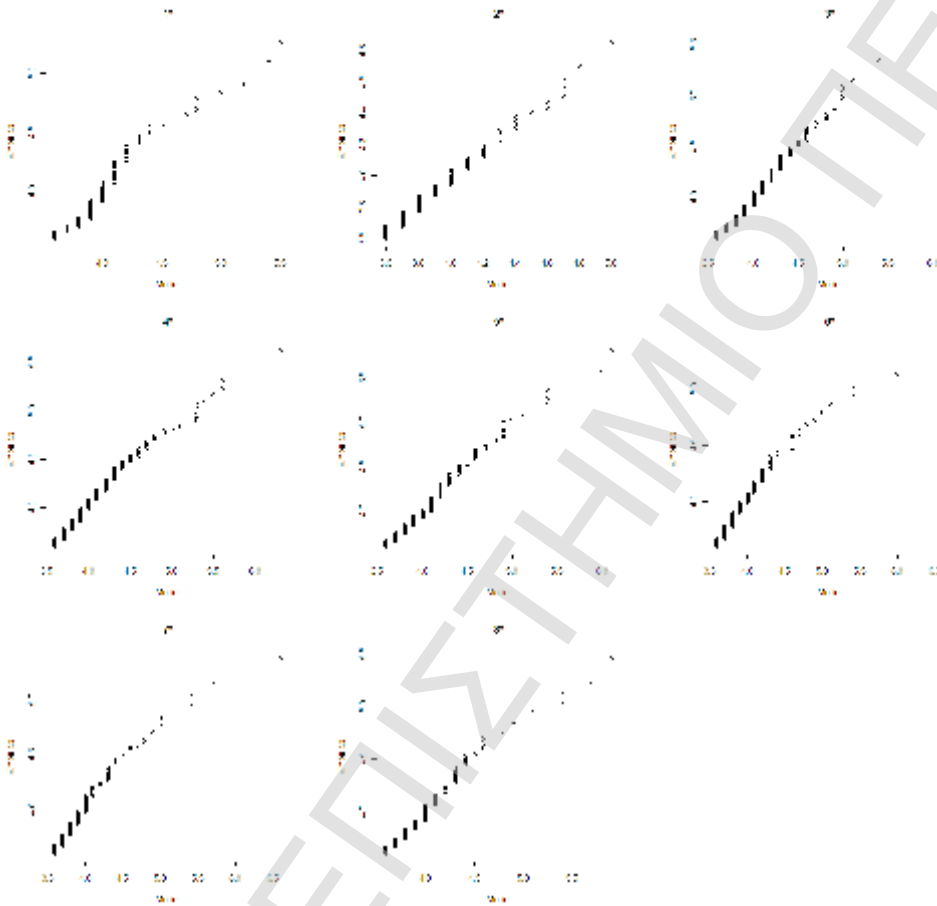
Εκτιμήσεις παραμέτρων GIII, στάθμης απόδοσης για 10 και 50 χρόνια απόδοσης, τυπικά σφάλματα και πλήθος λογοκριμένων παρατηρήσεων.

	$\hat{w}$	s.e.( $\hat{w}$ )	$\hat{m}$	s.e.( $\hat{m}$ )	$\hat{l}$	s.e.( $\hat{l}$ )	$\hat{z}_{p10}$	$\hat{z}_{p50}$	s.e.( $\hat{z}_{p50}$ )	$m$
1	12.809	1.6226	4.586	0.0369	0.036	0.0066	<b>5.24</b>	<b>5.67</b>	0.08779	66
2	11.156	1.8644	4.672	0.0377	0.047	0.0122	<b>5.33</b>	<b>5.76</b>	0.1129	63
3	11.275	3.3638	4.887	0.0428	0.048	0.0226	<b>5.56</b>	<b>6.00</b>	0.2022	31
4	5.836	0.1188	4.860	0.0302	0.252	0.0228	<b>5.29</b>	<b>5.47</b>	0.06379	66
5	5.821	0.1151	4.842	0.0308	0.255	0.0225	<b>5.28</b>	<b>5.46</b>	0.06217	73
6	10.811	1.8125	4.306	0.0325	0.038	0.0101	<b>4.86</b>	<b>5.22</b>	0.08718	132
7	11.998	1.4350	4.510	0.0335	0.036	0.0064	<b>5.10</b>	<b>5.50</b>	0.0788	55
8	11.218	1.5210	4.444	0.0344	0.041	0.0086	<b>5.05</b>	<b>5.44</b>	0.0855	96
9	11.448	2.0287	4.459	0.0369	0.041	0.0110	<b>5.09</b>	<b>5.48</b>	0.10390	113
10	6.643	0.3089	4.833	0.0338	0.152	0.0206	<b>5.37</b>	<b>5.65</b>	0.09606	51
11	13.355	1.7964	4.747	0.0391	0.037	0.0071	<b>5.45</b>	<b>5.92</b>	0.09453	23
12	12.471	1.7431	4.687	0.0383	0.039	0.0082	<b>5.37</b>	<b>5.82</b>	0.0956	31
13	12.091	2.8535	4.172	0.0401	0.036	0.0117	<b>4.81</b>	<b>5.22</b>	0.1112	255
14	12.407	1.7413	4.949	0.0427	0.047	0.0100	<b>5.71</b>	<b>6.20</b>	0.1134	14
15	13.014	1.8024	4.884	0.0429	0.043	0.0087	<b>5.65</b>	<b>6.13</b>	0.1092	23
16	10.875	1.8978	4.529	0.0455	0.054	0.0153	<b>5.28</b>	<b>5.76</b>	0.1320	157
17	13.377	2.5625	4.307	0.0407	0.033	0.0090	<b>4.97</b>	<b>5.42</b>	0.1063	196
18	13.548	2.0023	4.419	0.0361	0.031	0.0063	<b>5.04</b>	<b>5.45</b>	0.08748	106
19	11.772	1.6277	4.753	0.0401	0.046	0.0098	<b>5.47</b>	<b>5.93</b>	0.1051	42
20	12.980	1.8122	4.697	0.0410	0.039	0.0080	<b>5.42</b>	<b>5.90</b>	0.10160	56
21	11.000	2.0069	4.722	0.0490	0.060	0.0178	<b>5.54</b>	<b>6.04</b>	0.15871	126
22	15.046	2.3077	4.584	0.0447	0.033	0.0068	<b>5.35</b>	<b>5.82</b>	0.1089	113
23	13.587	2.1266	4.539	0.0401	0.034	0.0076	<b>5.23</b>	<b>5.68</b>	0.1001	101
24	8.583	1.6593	4.743	0.0419	0.084	0.0318	<b>5.42</b>	<b>5.80</b>	0.2099	77
25	12.807	1.9683	4.546	0.0407	0.038	0.0086	<b>5.25</b>	<b>5.69</b>	0.1033	108
26	15.576	2.7874	4.508	0.0475	0.032	0.0078	<b>5.30</b>	<b>5.82</b>	0.1204	168
27	16.258	2.4355	4.521	0.0451	0.029	0.0058	<b>5.29</b>	<b>5.79</b>	0.1090	142
1*	7.727	1.9805	4.228	0.0470	0.096	0.0516	<b>4.92</b>	<b>5.32</b>	0.2489	287
2*	10.261	2.5632	4.052	0.0323	0.037	0.0150	<b>4.57</b>	<b>4.91</b>	0.0998	262
3*	10.905	2.1831	4.519	0.0382	0.046	0.0146	<b>5.17</b>	<b>5.58</b>	0.1234	105
4*	13.587	2.1266	4.539	0.0401	0.035	0.0076	<b>5.23</b>	<b>5.68</b>	0.1001	101
5*	11.973	3.3071	4.542	0.0508	0.051	0.0216	<b>5.37</b>	<b>5.84</b>	0.1881	194
6*	14.256	3.2740	4.237	0.0440	0.031	0.0102	<b>4.94</b>	<b>5.41</b>	0.1209	253
7*	13.056	2.9464	4.321	0.0466	0.038	0.0125	<b>4.96</b>	<b>5.45</b>	0.1307	270
8*	12.507	4.6683	4.241	0.0502	0.043	0.0240	<b>5.03</b>	<b>5.55</b>	0.1885	287

Στη συνέχεια κατασκευάζουμε τα Q-Q plots για όλες τις περιοχές της Ελλάδας που έχουμε επιλέξει και παρατηρούμε μια ικανοποιητική προσαρμογή στο μοντέλο (Παράρτημα Κώδικας 10). Ακολουθούν τα Q-Q plots για τις 8 επιπλέον περιοχές που πήραμε σε αυτό το κεφάλαιο.

#### Σχήμα 4-6

Q-Q plots της GIII για την Μέθοδο Γ στις 8 επιπλέον περιοχές



#### 4.2 Σεισμολογικά δεδομένα περιόδου 1901-1996

Σε αυτήν την παράγραφο θα γίνει στατιστική ανάλυση επιφανειακών σεισμών που συνέβησαν στην περιοχή της Ελλάδας την χρονική περίοδο 1901 έως 1996. Ο συγκεκριμένος κατάλογος σεισμών βρίσκεται στην ιστοσελίδα του Τμήματος Γεωλογίας του Πανεπιστημίου Αθηνών ([www.geophysics.geol.uoa.gr](http://www.geophysics.geol.uoa.gr)). Οι αναλύσεις και η

συμπερασματολογία θα γίνει για όλο τον Ελλαδικό χώρο και όχι για μικρότερες περιοχές όπως στην Παράγραφο 4.1.

#### 4.2.1 Μέθοδος Block Maxima με ετήσια τοπικά μέγιστα

Τα τοπικά μέγιστα ελήφθησαν ανά έτος και έχουμε στην διάθεσή μας πλήρη δεδομένα. Από δείγμα 96 παρατηρήσεων λοιπόν εκτιμούμε τις παραμέτρους της GIII και τα τυπικά σφάλματά τους με επαναληπτικές μεθόδους και παίρνουμε τα εξής αποτελέσματα.

**Πίνακας 4-6**

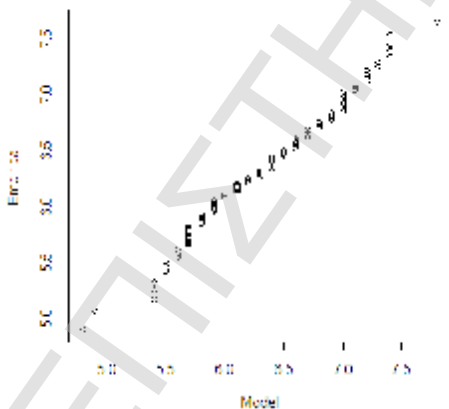
Εκτιμήσεις παραμέτρων GIII και τυπικά σφάλματα με ετήσια block maxima

$\hat{w}$	s.e. ( $\hat{w}$ )	$\hat{m}$	s.e. ( $\hat{m}$ )	$\hat{l}$	s.e. ( $\hat{l}$ )
8.0921	0.2855	6.1444	0.0708	0.3248	0.0643

Στη συνέχεια κατασκευάζουμε το Q-Q Plot

**Σχήμα 4-7**

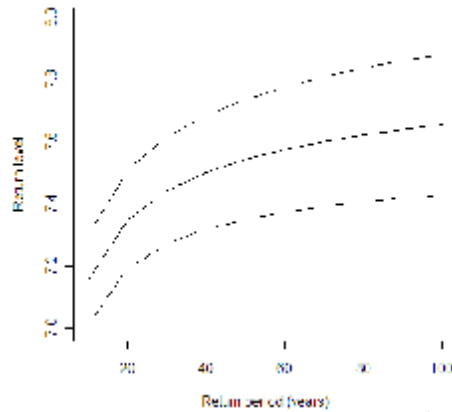
Q-Q Plot της GIII με ετήσια block maxima



Παρατηρούμε ότι τα σημεία είναι πολύ κοντά στην διαγώνιο που σημαίνει ότι έχουμε ικανοποιητική προσαρμογή. Επομένως μπορούμε να προχωρήσουμε στην εκτίμηση της στάθμης απόδοσης. Ακολουθεί το διάγραμμα της στάθμης απόδοσης για 10 έως 100 χρόνια καθώς και το 95% διάστημα εμπιστοσύνης, που βρέθηκε βάση της Μεθόδου Δέλτα.

### Σχήμα 4-8

Διάγραμμα στάθμης απόδοσης για 10 ως 100 χρόνια απόδοσης και 95% δ.ε. με ετήσια block maxima



Στον επόμενο πίνακα παρουσιάζονται κάποιες ενδεικτικές τιμές της εκτίμησης της στάθμης απόδοσης για  $1/p = 10, 50$  και  $100$ , μαζί με τα τυπικά σφάλματά τους.

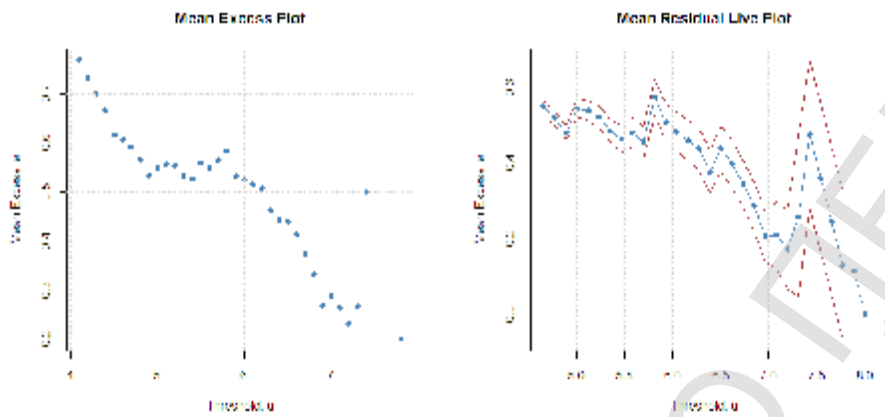
Χρόνια Απόδοσης	$\frac{1}{z_p}$	$s.e(\frac{1}{z_p})$
10	7.15	0.0742
50	7.54	0.0983
100	7.65	0.1153

#### 4.2.2 Μέθοδος Peaks Over Threshold

Το πρώτο που πρέπει να αποφασίσουμε στην μέθοδο POT είναι το μέγεθος του κατωφλίου  $u$ . Πάλι με την βοήθεια των γραφημάτων της εμπειρικής μέσης υπερβάλλουσας συνάρτησης που ακολουθούν καταλήγουμε ότι  $u = 5.7$ .

### Σχήμα 4-9

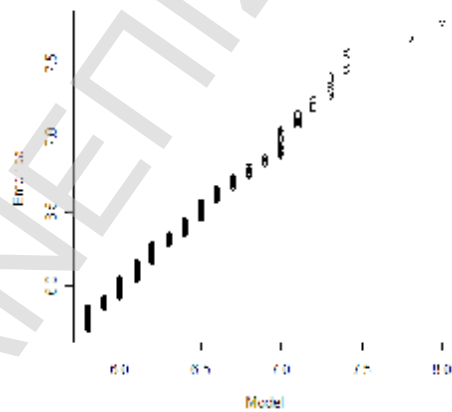
Εμπειρική μέση υπερβάλλουσα συνάρτηση



Πάνω από το κατώφλι  $u$  βρίσκονται  $k(u) = 229$  από τις 5991 παρατηρήσεις. Οι εκτιμήσεις των παραμέτρων της GPD είναι  $\hat{\xi} = 0.7064$  και  $\hat{\chi} = -0.2568$  με αντίστοιχα τυπικά σφάλματα 0.05656 και 0.04829. Ακολουθεί το Q-Q Plot βάσει του οποίου κρίνεται ικανοποιητική η προσαρμογή των δεδομένων στο μοντέλο.

### Σχήμα 4-10

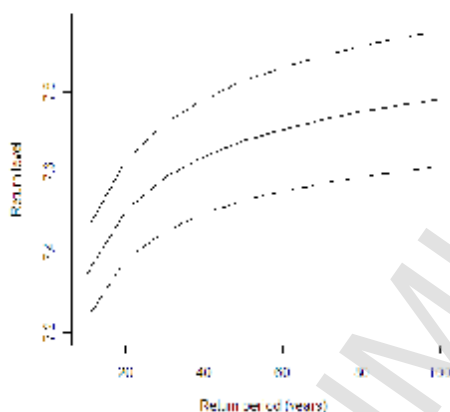
Q-Q plot για την GPD



Επίσης υπολογίστηκε η στάθμη απόδοσης για 10 έως 100 χρόνια απόδοσης καθώς και τα 95% προσεγγιστικά διαστήματα εμπιστοσύνης με την Μέθοδο Δέλτα. Τα παραπάνω παρουσιάζονται σε ένα κοινό γράφημα. Να σημειωθεί ότι η αναγωγή από παρατηρήσεις σε χρόνια έγινε με τον τρόπο που περιγράφεται στην Παράγραφο 4.1.2.

**Σχήμα 4-11**

Διάγραμμα στάθμης απόδοσης για 10 ως 100 χρόνια απόδοσης και 95% δ.ε.



Κλείνοντας την παράγραφο αυτή δίνονται τρεις ενδεικτικές τιμές του  $x_m$  και οι τυπικές τους αποκλίσεις.

Χρόνια Απόδοσης	$\hat{x}_m$	s.e( $\hat{x}_m$ )
10	7.23	0.0704
50	7.64	0.1035
100	7.77	0.1221

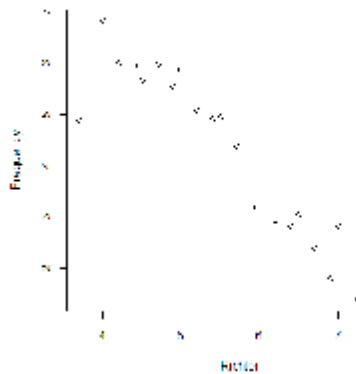
#### 4.2.3 Μέθοδος Block Maxima με μηνιαία τοπικά μέγιστα

Όπως αναφέραμε στην Παράγραφο 3.2, για να βελτιώσουμε τις εκτιμήσεις των παραμέτρων της GIII πρέπει να αυξηθεί το μέγεθος του δείγματος. Οπότε αντί να πάρουμε τα τοπικά μέγιστα ανά έτος είναι καλύτερα να τα πάρουμε ανά μήνα, με συνέπεια όμως να έχουμε και λογοκριμένες παρατηρήσεις. Για να διαχειριστούμε πλέον τα δεδομένα πρέπει να βρεθεί (μέσω γραφήματος) το όριο λογοκρισίας του συγκεκρι-

μένου καταλόγου σεισμών. Από το επόμενο γράφημα φαίνεται να είναι κοντά στους 4 βαθμούς της κλίμακας Richter.

#### Σχήμα 4-12

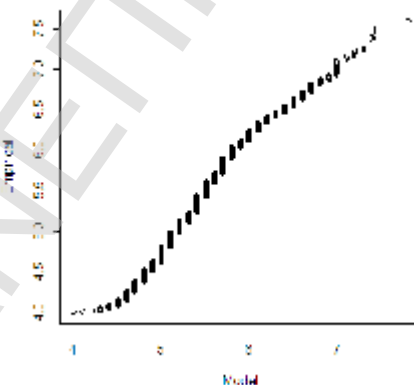
Διάγραμμα Gutenberg - Richter για το δεύτερο σύνολο δεδομένων



Με την βοήθεια επαναληπτικού αλγορίθμου εκτιμήσαμε τις παραμέτρους της GIII από ένα δείγμα 1152 παρατηρήσεων από τις οποίες οι 334 λογοκρίθηκαν. Στη συνέχεια υπολογίστηκε η στάθμη απόδοσης για 10 έως 100 χρόνια απόδοσης και η διασπορά της με την μέθοδο Δέλτα, ενώ έγινε και έλεγχος καλής προσαρμογής με Q-Q Plot. Όλα τα αποτελέσματα παρουσιάζονται ακολούθως.

#### Σχήμα 4-13

Q-Q Plot της GIII με μηνιαία block maxima





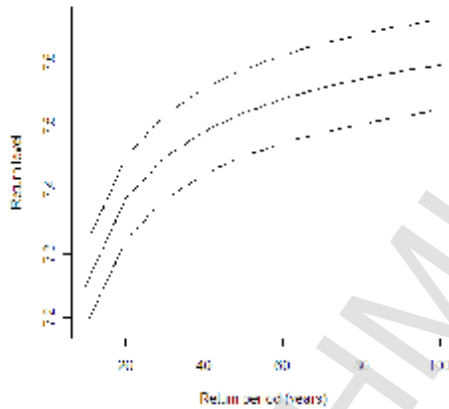
### Πίνακας 4-7

Εκτιμήσεις παραμέτρων GIII και τυπικά σφάλματα με μηνιαία block maxima

$\hat{w}$	s.e.( $\hat{w}$ )	$\hat{m}$	s.e.( $\hat{m}$ )	$\hat{l}$	s.e.( $\hat{l}$ )
7.8599	0.0464	6.4573	0.0407	0.3626	0.0134

### Σχήμα 4-14

Διάγραμμα στάθμης απόδοσης για 10 ως 100 χρόνια απόδοσης και 95% δ.ε.



Χρόνια Απόδοσης	$\hat{z}_p$	s.e.( $\hat{z}_p$ )
10	7.25	0.0323
50	7.52	0.0348
100	7.59	0.0363

# ΚΕΦΑΛΑΙΟ 5

## Συμπεράσματα

Μορφοποιήθηκε

### 5.1 Ποιοτικά συμπεράσματα

Κάθε μία από τις τρεις αναλύσεις που διεξήχθησαν (Α, Β, Γ) έχει τα μειονεκτήματα και τα πλεονεκτήματά της. Θεωρητικά, αν ο κατάλογος των σεισμών επεκτεινόταν για περισσότερα χρόνια και επομένως είχαμε μεγάλο αριθμό παρατηρήσεων οι εκτιμήσεις που θα παίρναμε θα ήταν πιο ασφαλείς και τα αποτελέσματα των τριών αναλύσεων πιο κοντά μεταξύ τους.

Όσον αφορά την μέθοδο POT, ένα από τα βασικά της μειονεκτήματα είναι ότι η επιλογή του κατωφλιού  $u$ , που παίζει πολύ σημαντικό ρόλο στην ανάλυση γίνεται με γραφικές μεθόδους, οι οποίες δεν είναι τόσο αξιόπιστες όσο θα θέλαμε. Επίσης όταν εφαρμόζουμε την μέθοδο POT σε σεισμολογικά δεδομένα και εκτιμούμε την στάθμη απόδοσης για  $m$  παρατηρήσεις πρέπει να αναγάγουμε αυτές τις  $m$  παρατηρήσεις σε χρόνο. Και αυτό γιατί ένα συμπέρασμα της μορφής ότι «*αναμένουμε έναν σεισμό μεγαλύτερο από 5 Richter κάθε 100 καταγεγραμμένους σεισμούς*» δεν έχει ιδιαίτερη χρηστικότητα, αφού η συχνότητα εμφάνισης των σεισμών μπορεί να διαφέρει αρκετά μεταξύ μιας περιοχής με έντονη σεισμική δραστηριότητα και μιας με χαμηλή. Το λεπτό λοιπόν σημείο εδώ είναι το πέρασμα από τις παρατηρήσεις σε χρόνια. Στην συγκεκριμένη εργασία αυτό έγινε βρίσκοντας τον μέσο αριθμό των συμβάντων ανά έτος σε κάθε περιοχή. Το πλεονέκτημα της μεθόδου POT είναι ότι μπορεί να θεωρηθεί πιο ασφαλής γιατί λαμβάνει υπόψη της ένα μεγαλύτερο μέρος των δεδομένων που έχουμε στη διάθεσή μας, αυτά που ξεπερνούν το  $u$  και όχι μόνο τις μεγαλύτερες παρατηρήσεις ανά χρονική περίοδο.

Μεταξύ των δύο άλλων αναλύσεων που έγιναν είναι προφανές ότι όταν διαχειριζόμαστε λογοκριμένα δεδομένα (μέθοδος Γ) έχουμε την δυνατότητα να επεκτείνουμε την ανάλυση σε περισσότερες περιοχές και να εξάγουμε συμπεράσματα για όλο τον Ελλαδικό χώρο, αφού δεν είναι αναγκαίο να υπάρχει τουλάχιστον ένας σεισμός ανά χρονική περίοδο, όπως στην περίπτωση με τα πλήρη δεδομένα. Επίσης, κατά τη μέθοδο Γ μπορούμε να μικρύνουμε τα χρονικά διαστήματα που επιλέγονται τα

τοπικά μέγιστα και να αυξήσουμε με αυτόν τον τρόπο το μέγεθος του δείγματος από το οποίο υπολογίζονται οι εκτιμήσεις της οριακής κατανομής. Αυτό έχει ως αποτέλεσμα να βελτιώνεται η ακρίβεια τους, όπως παρατηρήσαμε και στην Παράγραφο 3.2. Βέβαια πρέπει να προσέξουμε ώστε να μην παραβιάζεται η βασική συνθήκη της ανεξαρτησίας μεταξύ των παρατηρήσεων. Οπότε στην περίπτωση των σεισμολογικών δεδομένων δεν θα ήταν σωστό να λαμβάνονται οι μεγαλύτεροι σεισμοί ανά πολύ μικρά χρονικά διαστήματα, γιατί ένας κύριος σεισμός μεγάλου μεγέθους είναι γνωστό ότι έχει προσεισμούς και μετασεισμούς που το μέγεθός τους εξαρτάται από αυτόν και συμβαίνουν χρονικά κοντά σε αυτόν. Ένα μειονέκτημα που εμφανίζεται στην επεξεργασία λογοκριμένων δεδομένων είναι ότι πρέπει να οριστεί γραφικά το όριο πάνω από το οποίο έχουν καταγραφεί όλοι οι σεισμοί στον συγκεκριμένο κατάλογο. Το όριο αυτό διαδραματίζει σημαντικό ρόλο στην περαιτέρω στατιστική ανάλυση και η επιρροή του γίνεται πιο εμφανής όταν η αναλογία των λογοκριμένων προς τα πλήρη δεδομένα είναι μεγάλη. Πιο συγκεκριμένα στην εξίσωση πιθανοφάνειας

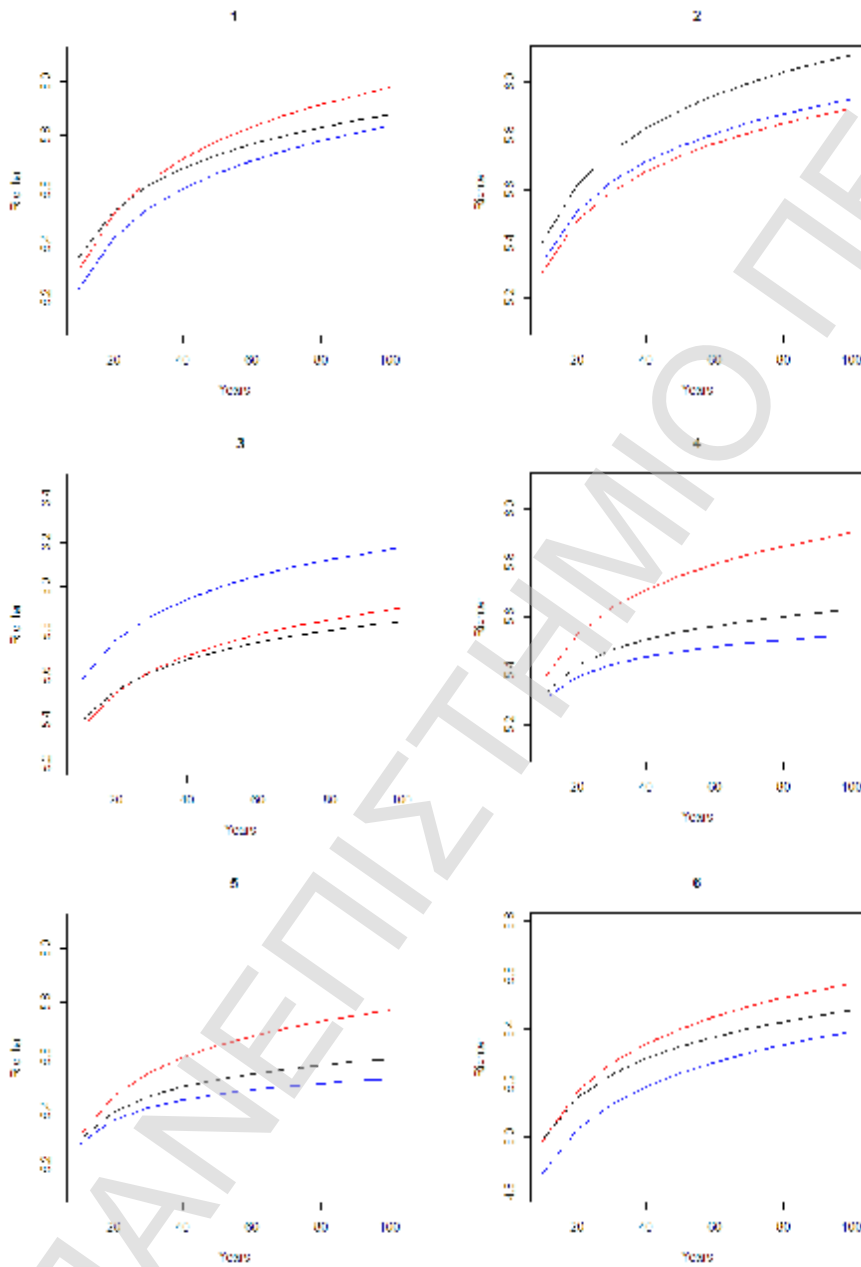
$$l(w, m, I) = -mT \left( \frac{w - x_L}{w - m} \right)^{1/I} + \sum_{i=1}^n \left\{ -\ln(I(w - m)) + \left( \frac{1}{I} - 1 \right) \ln \left( \frac{w - x_i}{w - m} \right) - T \left( \frac{w - x_i}{w - m} \right)^{\frac{1}{I}} \right\}$$

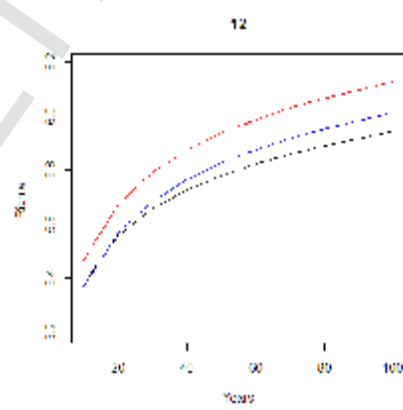
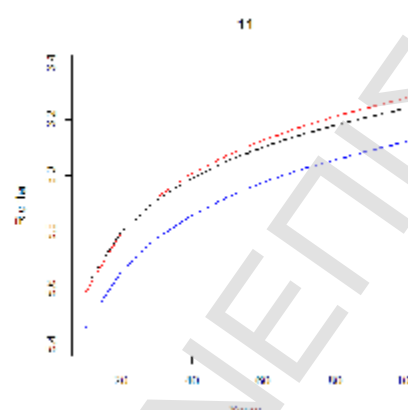
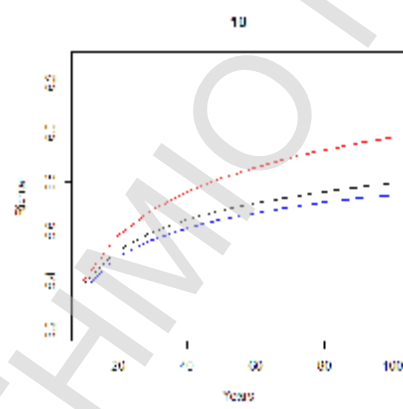
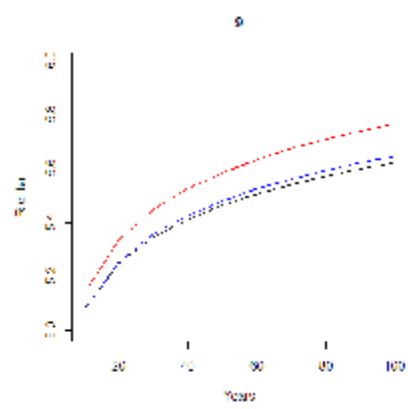
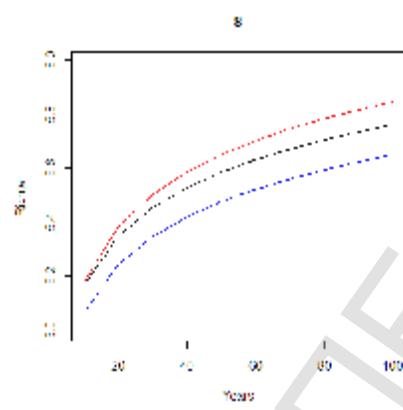
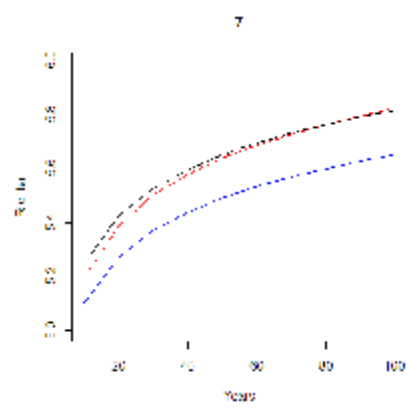
ο πρώτος όρος είναι αυτός που συμβάλει περισσότερο στην μεγιστοποίησή της, όταν το  $m$  είναι μεγάλο και συνεπώς στην εκτίμηση των τριών παραμέτρων. Αν το  $x_L$  λοιπόν δεν έχει οριστεί σωστά υπάρχει κίνδυνος οι εκτιμήσεις των  $w$ ,  $m$ ,  $\lambda$  να απέχουν από τις πραγματικές τιμές τους. Το πρόβλημα αυτό μπορεί να ξεπεραστεί αν έχουμε στην διάθεσή μας έναν κατάλογο σεισμών που ξεπερνούν ένα συγκεκριμένο μέγεθος.

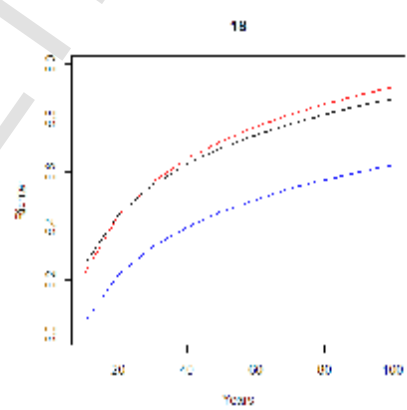
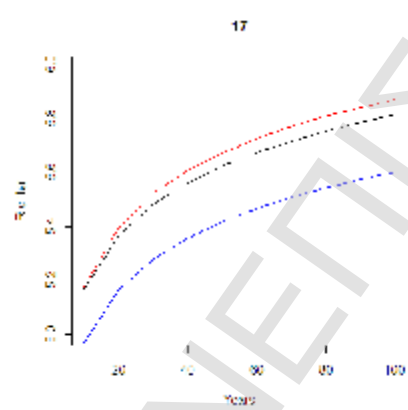
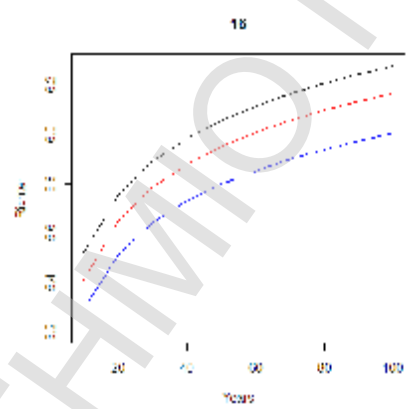
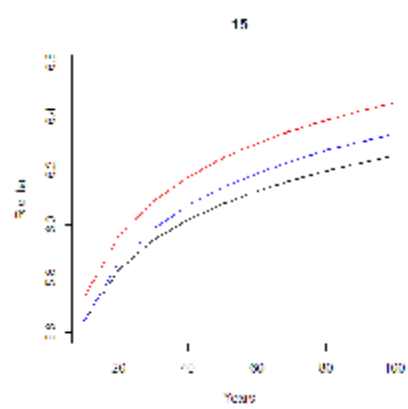
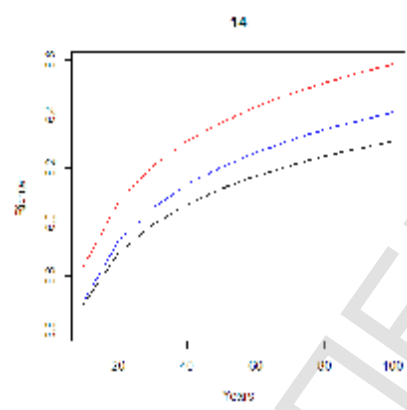
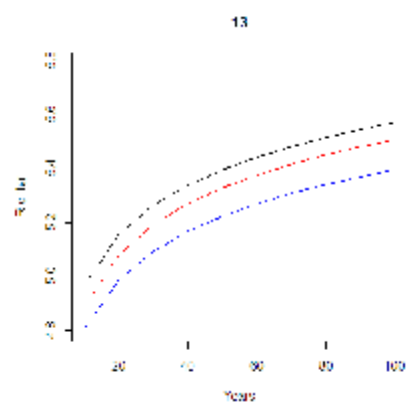
Κατασκευάστηκαν επίσης τα διαγράμματα της εκτίμησης της στάθμης απόδοσης για 10 ως 100 χρόνια απόδοσης, για τις τρεις διαφορετικές αναλύσεις που έγιναν (Μέθοδος Α, Β και Γ) και παρουσιάζονται σε ένα κοινό γράφημα για κάθε μια από τις 27 περιοχές του Πίνακα 1. Αυτά μπορούν να χρησιμοποιηθούν για να συγκρίνουμε τις τρεις μεθόδους. Με κόκκινο χρώμα είναι η εκτίμηση της στάθμης απόδοσης όταν είχαμε πλήρη δεδομένα (Α), με μπλε είναι όταν είχαμε λογοκριμένα (Γ) και με μαύρο η εκτίμηση όταν επεξεργαστήκαμε τα δεδομένα σύμφωνα με την μέθοδο POT (Β).

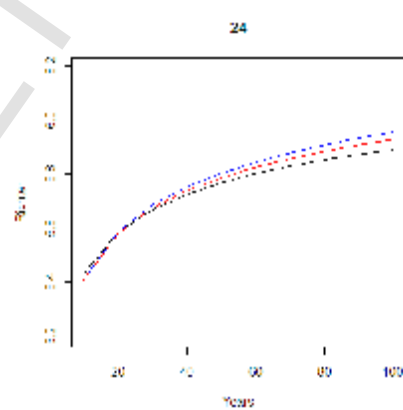
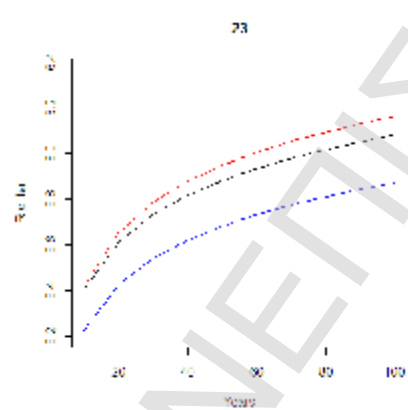
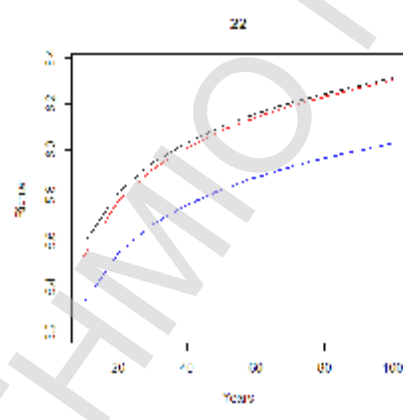
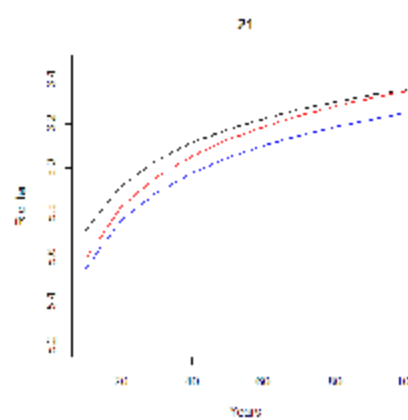
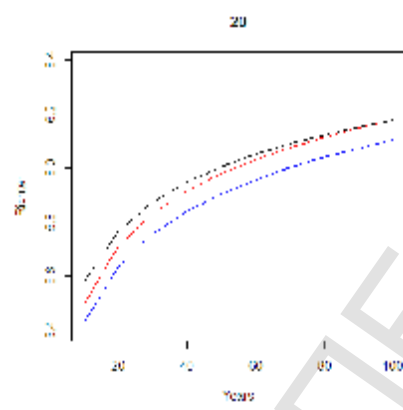
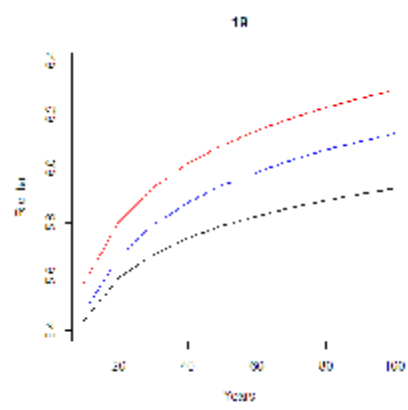
### Σχήμα 5-1

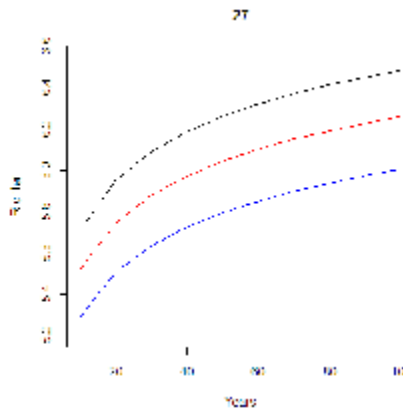
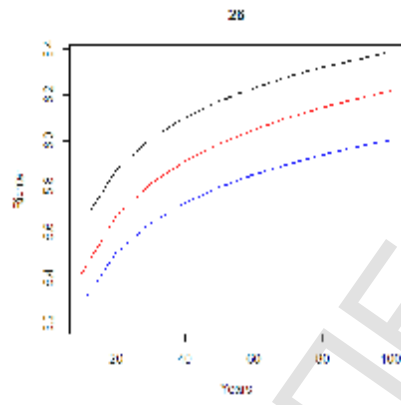
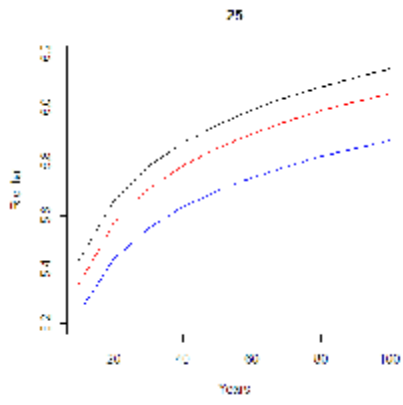
Κοινά διαγράμματα της στάθμης απόδοσης για 10 έως 100 χρόνια απόδοσης για τις τρεις αναλύσεις (Α, Β, Γ).











ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ



Παρατηρούμε ότι οι εκτιμήσεις που παίρνουμε όταν διαχειριστήκαμε τα λογοκριμένα δεδομένα (Γ) είναι τις περισσότερες φορές μικρότερες ή ίσες από αυτές που λάβαμε με την μέθοδο POT (B) και με την μέθοδο που χρησιμοποιούσε πλήρη δεδομένα (A). Χαρακτηριστικά παραδείγματα είναι τα γραφήματα των περιοχών 17, 18, 22 και 23 όπου η κόκκινη και η μαύρη γραμμή σχεδόν ταυτίζονται ενώ η μπλε είναι αρκετά χαμηλότερα. Σε αυτές τις περιπτώσεις πρέπει μάλλον να εμπιστευτούμε περισσότερο τις δύο πρώτες αναλύσεις, αφού η τρίτη φαίνεται να υποεκτιμά την πραγματική τιμή της στάθμης απόδοσης. Αυτό μάλλον οφείλεται στο γεγονός ότι οι εκτιμήσεις των παραμέτρων της οριακής κατανομής *Gumbel III* υπολογίστηκαν από ένα δείγμα 456 παρατηρήσεων και τουλάχιστον 100 ήταν λογοκριμένες με όριο λογοκρισίας τους 3.6 βαθμούς Richter, ενώ μπορεί για αυτές τις περιοχές το όριο να είναι διαφορετικό. Για να αποφευχθεί αυτό ίσως θα έπρεπε το όριο λογοκρισίας να λαμβάνεται ξεχωριστά σε κάθε περιοχή. Επίσης μια σημαντική παρατήρηση είναι ότι τις περισσότερες φορές οι εκτιμήσεις που παίρνουμε με την μέθοδο POT (B) και την πρώτη ανάλυση (A) είναι αρκετά κοντά, ενώ όταν δεν συμβαίνει αυτό δεν φαίνεται κάποια ανάλυση να δίνει πάντα μεγαλύτερη ή μικρότερη εκτίμηση. Αυτό σημαίνει ότι δεν υπάρχει κάποια τάση μια ανάλυση να υπερεκτιμά ή υποεκτιμά την στάθμη απόδοσης σε σχέση με την άλλη. Επίσης φαίνεται ότι η αναγωγή των παρατηρήσεων σε έτη που έγινε στην μέθοδο POT για την στάθμη απόδοσης δεν δημιούργησε προβλήματα. Τέλος να τονίσουμε ότι από τις 27 περιοχές μόνο σε δύο, στην 19<sup>η</sup> και 26<sup>η</sup>, οι τρεις εκτιμήσεις για 100 χρόνια απόδοσης διαφέρουν η μια από την άλλη κατά 0.3 βαθμούς Richter περίπου γεγονός που ενισχύει την αρχική πρόταση του κεφαλαίου, ότι θεωρητικά οι τρεις αναλύσεις δίνουν σχεδόν τα ίδια αποτελέσματα.

Ένας κλασικός τρόπος για να κρίνουμε ποια εκτίμηση είναι καλύτερη είναι να κοιτάξουμε τις τυπικές αποκλίσεις που παίρνουμε σε κάθε ανάλυση ή ισοδύναμα το εύρος του διαστήματος εμπιστοσύνης συντελεστού  $\alpha$  της εκτίμησης. Στην προκειμένη περίπτωση έχουν υπολογιστεί τα τυπικά σφάλματα της στάθμης απόδοσης με την Μέθοδο Δέλτα. Στον επόμενο πίνακα παρουσιάζονται τα εύρη των 95% δ.ε. της στάθμης απόδοσης για 50 χρόνια απόδοσης.

### Πίνακας 5-1

Εύρος 95% Διαστήματος Εμπιστοσύνης της στάθμης απόδοσης για 50 χρόνια απόδοσης με τις τρεις Μεθόδους

	Μεθοδος Α	Μεθοδος Β	Μεθοδος Γ
1	0.4581	0.4322	<b>0.3441</b>
2	0.4620	0.7496	<b>0.4426</b>
3	<b>0.4048</b>	0.4746	0.7928
4	0.5393	0.3898	<b>0.2500</b>
5	0.7242	0.3574	<b>0.2437</b>
6	0.6613	0.5552	<b>0.3417</b>
7	0.4474	0.5344	<b>0.3092</b>
8	0.4686	0.5048	<b>0.3351</b>
9	0.5876	0.5277	<b>0.4073</b>
10	0.5364	0.3877	<b>0.3765</b>
11	0.5622	0.5907	<b>0.3705</b>
12	0.5644	0.4468	<b>0.3749</b>
13	0.4962	0.6595	<b>0.4361</b>
14	0.7509	0.4616	<b>0.4447</b>
15	0.6526	0.4666	<b>0.4282</b>
16	0.6849	0.8625	<b>0.5177</b>
17	0.8544	0.6074	<b>0.4168</b>
18	0.5814	0.5437	<b>0.3429</b>
19	0.7749	<b>0.4044</b>	0.4120
20	0.5676	0.5761	<b>0.3982</b>
21	0.6975	0.7068	<b>0.6221</b>
22	0.7882	0.7105	<b>0.4270</b>
23	0.7250	0.6152	<b>0.3922</b>
24	0.7281	<b>0.5248</b>	0.8228
25	0.6615	0.8745	<b>0.4050</b>
26	0.5762	0.8543	<b>0.4719</b>
27	0.6113	0.9226	<b>0.4273</b>

Παρατηρώντας τον Πίνακα 8 βλέπουμε ότι τις καλύτερες εκτιμήσεις τις παίρνουμε με την Μέθοδο Γ, η οποία σε 24 από τις 27 περιπτώσεις δίνει τα στενότερα διαστήματα εμπιστοσύνης. Μεταξύ των άλλων δύο μεθόδων υπάρχει μια ισορροπία ως προς το ποια εκτιμά με μεγαλύτερη ακρίβεια την στάθμη απόδοσης.

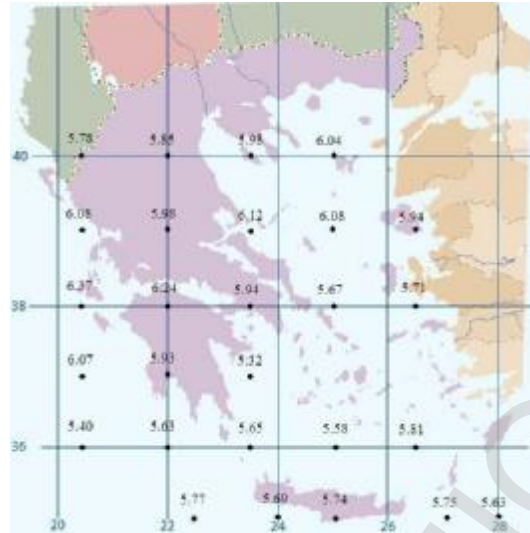
## 5.2 Ποσοτικά συμπεράσματα

Στους πίνακες 2, 3 και 5 παρουσιάστηκαν αναλυτικά για κάθε περιοχή τα μεγέθη των μεγαλύτερων σεισμών που περιμένουμε ανά 10ετία και 50ετία και από αυτές τις προβλέψεις μπορούμε να κατατάξουμε τις περιοχές της Ελλάδας ανάλογα με την επικινδυνότητά τους. Το *κεντρικό Ιόνιο πέλαγος*, η *βόρεια Πελοπόννησος*, οι *βόρειες Σποράδες* είναι οι τρεις περιοχές (14<sup>η</sup>, 15<sup>η</sup>, 21<sup>η</sup>), όπου και με τις τρεις αναλύσεις βρέθηκαν να παρουσιάζουν την πιο έντονη σεισμική δραστηριότητα και ειδικότερα η στάθμη απόδοσης για 50 χρόνια απόδοσης εκτιμήθηκε μεγαλύτερη από 6.00 βαθμούς της κλίμακας Richter.

Όπως αναφέραμε παραπάνω, η ανάλυση με λογοκριμένα δεδομένα φαίνεται σε ορισμένες περιοχές να υποεκτιμά την στάθμη απόδοσης σε σύγκριση με τις δύο πρώτες αναλύσεις. Αν βασιστούμε μόνο σε αυτές τότε ολόκληρο το *Ιόνιο πέλαγος*, η *κεντρική ηπειρωτική Ελλάδα* και το *βορειοανατολικό Αιγαίο* μπορούν να θεωρηθούν ως περιοχές με μεγάλη επικινδυνότητα. Τέλος η ανάλυση που έγινε για τις επόμενες 8 περιοχές, για τις οποίες δεν ήταν δυνατόν να έχουμε μεγάλο αριθμό παρατηρήσεων με πλήρη δεδομένα δεν υπάρχει κάποια περιοχή, όπου η στάθμη απόδοσης για 50 χρόνια εκτιμάται μεγαλύτερη από 5.85 βαθμούς Richter. Αυτό ενισχύει το συμπέρασμα ότι σε αυτές τις περιοχές η σεισμική δραστηριότητα είναι χαμηλή και γι' αυτό δεν μπορούσαμε να έχουμε έναν τουλάχιστον σεισμό ανά δίμηνο, τρίμηνο ή τετράμηνο, ώστε να συμπεριληφθούν στην πρώτη ανάλυση. Ακολουθούν οι χάρτες της Ελλάδας με σημειωμένες τις «σημειακές» εκτιμήσεις απόδοσης για 50 χρόνια απόδοσης και τα αντίστοιχα contours plots για τις τρεις αναλύσεις που έγιναν.

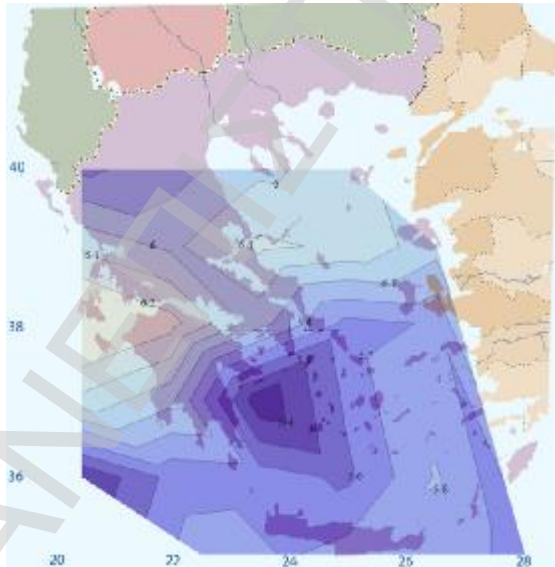
**Σχήμα 5-2**

Εκτιμήσεις στάθμης απόδοσης για 50 χρόνια απόδοσης με GIII και πλήρη δεδομένα.



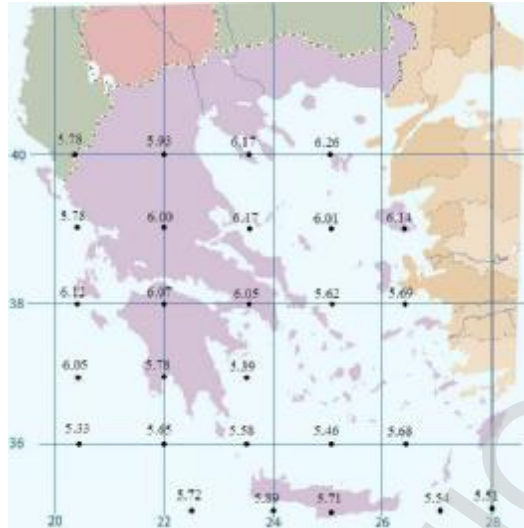
**Σχήμα 5-3**

Contour plot εκτιμήσεων στάθμης απόδοσης για 50 χρόνια απόδοσης με GIII και πλήρη δεδομένα.



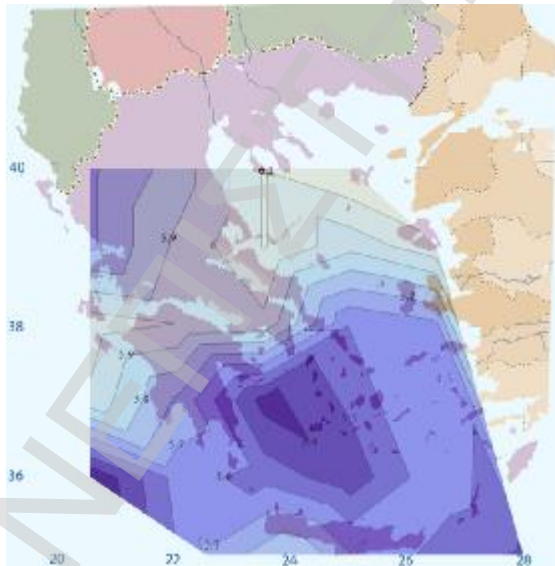
**Σχήμα 5-4**

Εκτιμήσεις στάθμης απόδοσης για 50 χρόνια απόδοσης με *GPD*.



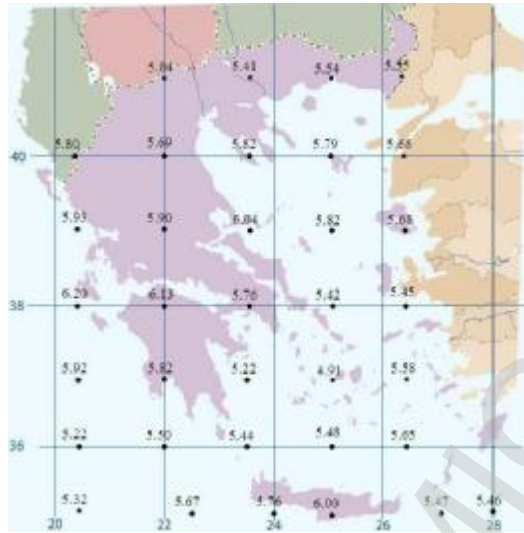
**Σχήμα 5-5**

Contour plot εκτιμήσεων στάθμης απόδοσης για 50 χρόνια απόδοσης με *GPD*.



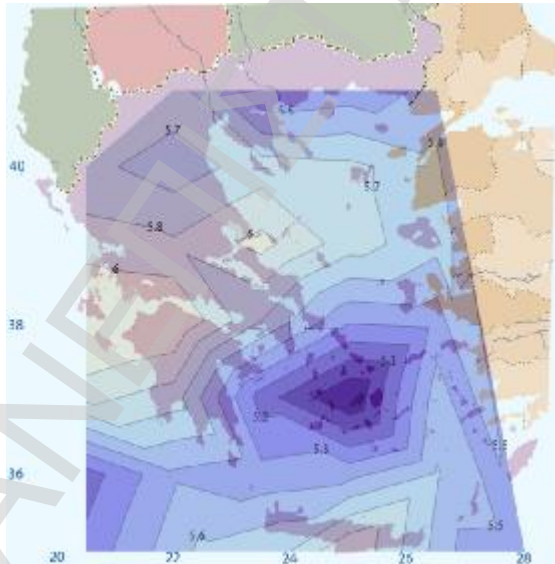
### Σχήμα 5-6

Εκτιμήσεις στάθμης απόδοσης για 50 χρόνια απόδοσης με GIII και λογοκριμένα δεδομένα.



### Σχήμα 5-7

Contour plot εκτιμήσεων στάθμης απόδοσης για 50 χρόνια απόδοσης με GIII και λογοκριμένα δεδομένα.



Όσον αφορά την ανάλυση που έγινε με τον δεύτερο κατάλογο σεισμών για όλη την Ελλάδα, τα συμπεράσματα που μπορούν να εξαχθούν βασίζόμενοι στην εκτίμηση της στάθμης απόδοσης με την μέθοδο Block Maxima και ετήσια τοπικά μέγιστα είναι.

- Αναμένεται να συμβεί ένας σεισμός στην Ελλάδα με ένταση μεγαλύτερη από 7.15 (αντ. 7.65) βαθμούς της κλίμακας Richter κάθε 10 (αντ. 100) χρόνια.

Με την μέθοδο POT

- Αναμένεται να συμβεί ένας σεισμός στην Ελλάδα με ένταση μεγαλύτερη από 7.34 (αντ. 7.78) βαθμούς της κλίμακας Richter κάθε 10 (αντ. 100) χρόνια.

Με την μέθοδο Block Maxima και μηνιαία τοπικά μέγιστα

- Αναμένεται να συμβεί ένας σεισμός στην Ελλάδα με ένταση μεγαλύτερη από 7.25 (αντ. 7.59) βαθμούς της κλίμακας Richter κάθε 10 (αντ. 100) χρόνια.

**Κώδικας 1.** Με το ακόλουθο script έγινε η εμπειρική αξιολόγηση της μεθόδου Block

Maxima με λογοκριμένα δεδομένα.

```

la1<-c();la2<-c();la5<-c();la10<-c();la25<-c();
la50<-c();la100<-c()
om1<-c();om2<-c();om5<-c();om10<-c();om25<-c();
om50<-c();om100<-c()

plot(-1:1,-1:1)
for(j in 1:1000)
{
plot(-1:1,-1:1);text(0.1,0,j)
omega0<-9; mu0<-5; lamda0<-0.15
u<-runif(100000,0,1)
x1<-omega0-(omega0-mu0)*(-log(u))^lamda0

m=1; k=100000/m; bm<-rep(0,k);
for(i in 1:k){bm[i]<-max(x1[((i-1)*m+1):(i*m)])}
mera1<-bm
n1<-length(mera1)
meralc<-mera1[mera1>=3.6]
m1<-n1-length(meralc)
ll<- function(mu=4,omega=10,lamda=0.2)
-m1*(-((omega-3.6)/(omega-mu))^(1/lamda))-sum(-
log(lamda*(omega-mu))+(1/lamda-1)*log((omega-
meralc)/(omega-mu))-((omega-meralc)/(omega-
mu))^(1/lamda))
fit1<-mle(ll)

m=2; k=100000/m; bm<-rep(0,k);
for(i in 1:k){bm[i]<-max(x1[((i-1)*m+1):(i*m)])}
mera2<-bm
n2<-length(mera2)
mera2c<-mera2[mera2>=3.6]
m2<-n2-length(mera2c)
ll<- function(mu=4,omega=10,lamda=0.2)
-m2*(-((omega-3.6)/(omega-mu))^(1/lamda))-sum(-
log(lamda*(omega-mu))+(1/lamda-1)*log((omega-
mera2c)/(omega-mu))-((omega-mera2c)/(omega-
mu))^(1/lamda))
fit2<-mle(ll)

m=5; k=100000/m; bm<-rep(0,k);
for(i in 1:k){bm[i]<-max(x1[((i-1)*m+1):(i*m)])}
mera5<-bm
n5<-length(mera5)
mera5c<-mera5[mera5>=3.6]
m5<-n5-length(mera5c)
ll<- function(mu=4,omega=10,lamda=0.2)
-m5*(-((omega-3.6)/(omega-mu))^(1/lamda))-sum(-
log(lamda*(omega-mu))+(1/lamda-1)*log((omega-

```



```

mera5c)/(omega-mu))-((omega-mera5c)/(omega-
mu)^(1/lamda))
fit5<-mle(ll)

m=10; k=100000/m; bm<-rep(0,k);
for(i in 1:k){bm[i]<-max(x1[((i-1)*m+1):(i*m)])}
mera10<-bm
n10<-length(mera10)
mera10c<-mera10[mera10>=3.6]
m10<-n10-length(mera10c)
ll<- function(mu=4,omega=10,lamda=0.2)
-m10*(-((omega-3.6)/(omega-mu))^(1/lamda))-sum(-
log(lamda*(omega-mu))+(1/lamda-1)*log((omega-
mera10c)/(omega-mu)))-((omega-mera10c)/(omega-
mu)^(1/lamda))
fit10<-mle(ll)

m=25; k=100000/m; bm<-rep(0,k);
for(i in 1:k){bm[i]<-max(x1[((i-1)*m+1):(i*m)])}
mera25<-bm
n25<-length(mera25)
mera25c<-mera25[mera25>=3.6]
m25<-n25-length(mera25c)
ll<- function(mu=4,omega=10,lamda=0.2)
-m25*(-((omega-3.6)/(omega-mu))^(1/lamda))-sum(-
log(lamda*(omega-mu))+(1/lamda-1)*log((omega-
mera25c)/(omega-mu)))-((omega-mera25c)/(omega-
mu)^(1/lamda))
fit25<-mle(ll)

m=50; k=100000/m; bm<-rep(0,k);
for(i in 1:k){bm[i]<-max(x1[((i-1)*m+1):(i*m)])}
mera50<-bm
n50<-length(mera50)
mera50c<-mera50[mera50>=3.6]
m50<-n50-length(mera50c)
ll<- function(mu=4,omega=10,lamda=0.2)
-m50*(-((omega-3.6)/(omega-mu))^(1/lamda))-sum(-
log(lamda*(omega-mu))+(1/lamda-1)*log((omega-
mera50c)/(omega-mu)))-((omega-mera50c)/(omega-
mu)^(1/lamda))
fit50<-mle(ll)

m=100; k=100000/m; bm<-rep(0,k);
for(i in 1:k){bm[i]<-max(x1[((i-1)*m+1):(i*m)])}
mera100<-bm
n100<-length(mera100)
mera100c<-mera100[mera100>=3.6]
m100<-n100-length(mera100c)
ll<- function(mu=4,omega=10,lamda=0.2)
-m100*(-((omega-3.6)/(omega-mu))^(1/lamda))-sum(-
log(lamda*(omega-mu))+(1/lamda-1)*log((omega-mera100c)
/(omega-mu)))-((omega-mera100c)/(omega-mu)^(1/lamda))
fit100<-mle(ll)

```

```

la1<-append(la1,coef(fit1)[3])
la2<-append(la2,coef(fit2)[3])
la5<-append(la5,coef(fit5)[3])
la10<-append(la10,coef(fit10)[3])
la25<-append(la25,coef(fit25)[3])
la50<-append(la50,coef(fit50)[3])
la100<-append(la100,coef(fit100)[3])

om1<-append(om1,coef(fit1)[2])
om2<-append(om2,coef(fit2)[2])
om5<-append(om5,coef(fit5)[2])
om10<-append(om10,coef(fit10)[2])
om25<-append(om25,coef(fit25)[2])
om50<-append(om50,coef(fit50)[2])
om100<-append(om100,coef(fit100)[2])

}

x<-om1
hist(x,seq(8,12.7,0.05),prob=T,ylim = c(0, 2.4)
,main = paste("omega estimation (block size=",1,"")
lines(density(x,bw=0.08))
rug(x)

x<-la1
hist(x,seq(0.05,0.225,0.005),prob=T, ylim = c(0, 33)
,main = paste("lambda estimation (block size=",1,"")
lines(density(x,bw=0.0075))
rug(x)

```

**Κώδικας 2.** Εκτιμήσεις των παραμέτρων  $\omega$ ,  $\lambda$  με αρχικές συνθήκες τις πραγματικές τιμές, για μέγεθος block ίσο με 2.

```

plot(-1:1,-1:1)
for(j in 1:1000)
{
plot(-1:1,-1:1);text(0.1,0,j)
omega0<-9; mu0<-5; lamda0<-0.15
u<-runif(100000,0,1)
x1<-omega0-(omega0-mu0)*(-log(u))^lamda0

m=2; k=100000/m; bm<-rep(0,k);
for(i in 1:k){bm[i]<-max(x1[((i-1)*m+1):(i*m)])}
mera2<-bm
n2<-length(mera2)
mera2c<-mera2[mera2>=6.5]
m2<-n2-length(mera2c)

ll<- function(mu=5, omega=9, lamda=0.15)
-m2*(-((omega-6.5)/(omega-mu))^(1/lamda))-sum(-
log(lamda*(omega-mu))+(1/lamda-1)*log((omega-

```

```

mera2c)/(omega-mu))-((omega-mera2c)/(omega-
mu))^(1/lamda))
fit2<-mle(l1)

la2<-append(la2,coef(fit2)[3])
om2<-append(om2,coef(fit2)[2])
}

x<-om2
hist(x,seq(8,10,0.05),prob=T,ylim = c(0, 3.0)
,main = paste("omega estimation (block size=",2,""))
lines(density(x,bw=0.08))
rug(x)

x<-la2
hist(x,seq(0.10,0.20,0.005),prob=T, ylim = c(0, 50)
,main = paste("lambda estimation (block size=",2,""))
lines(density(x,bw=0.0075))
rug(x)

```

**Κώδικας 3.** Υπολογισμός μέσης απόλυτης διαφοράς για τις  $\omega$ ,  $\lambda$

```

sum(abs(l1a1[l1a1>0.08]-lamda0))/length(l1a1[l1a1>0.08])
sum(abs(l1a2[l1a2>0.08]-lamda0))/length(l1a2[l1a2>0.08])
sum(abs(l1a5[l1a5>0.08]-lamda0))/length(l1a5[l1a5>0.08])
sum(abs(l1a10[l1a10>0.08]-
lamda0))/length(l1a10[l1a10>0.08])
sum(abs(l1a25[l1a25>0.08]-
lamda0))/length(l1a25[l1a25>0.08])
sum(abs(l1a50[l1a50>0.08]-
lamda0))/length(l1a50[l1a50>0.08])
sum(abs(l1a100[l1a100>0.08]-
lamda0))/length(l1a100[l1a100>0.08])

ul=9.8
sum(abs(omm1[omm1<ul]-omega0))/length(omm1[omm1<ul])
sum(abs(omm2[omm2<ul]-omega0))/length(omm2[omm2<ul])
sum(abs(omm5[omm5<ul]-omega0))/length(omm5[omm5<ul])
sum(abs(omm10[omm10<ul]-omega0))/length(omm10[omm10<ul])
sum(abs(omm25[omm25<ul]-omega0))/length(omm25[omm25<ul])
sum(abs(omm50[omm50<ul]-omega0))/length(omm50[omm50<ul])
sum(abs(omm100[omm100<ul]-
omega0))/length(omm100[omm100<ul])

```

**Κώδικας 4.** Ο υπολογισμός των εκτιμήσεων των παραμέτρων της GIII γίνεται μέσω της συνάρτησης `mle` η οποία ελαχιστοποιεί την  $-\log L$  ως προς τις παραμέτρους  $\omega$  (`omega`),  $\mu$  (`mu`),  $\lambda$  (`lamda`). Το διάνυσμα `x` αποτελείται από τα τοπικά μέγιστα και η μεταβλητή `TT` αντιστοιχεί στην χρονική περίοδο που επιλέχθηκαν. Δηλαδή για `TT = 1` έχουμε ετήσια, `TT = 4` τριμηνιαία, `TT = 6` διμηνιαία, `TT = 12` μηνιαία τοπικά μέγιστα.

```

ll<- function(mu=5,omega=12,lamda=0.1)
-sum(-log(lamda*(omega-mu))+(1/lamda-1)*log((omega-x)
/(omega-mu))-(1/TT)*((omega-x)/(omega-mu))^(1/lamda))
fit<-mle(ll)

summary(fit)

```

**Κώδικας 5.** Το επόμενο script υπολογίζει την στάθμη απόδοσης για 10 έως 10 χρόνια απόδοσης και την διασπορά της βάση της μεθόδου Δέλτα.

```

mu<-coef(fit)[1]
omega<-coef(fit)[2]
lamda<-coef(fit)[3]
z<-vcov(fit)

xronia10<-TT*10;xronia100<-TT*100
dia<-seq(xronia10,xronia100,xronia10)
p<-1/dia
zp<-omega-(omega-mu)*(-log(1-p)*TT)^lamda

gomega<-1-(-log(1-p)*TT)^lamda
gmu<-(-log(1-p)*TT)^lamda
glamda<--(omega-mu)*(-log(1-p))^lamda*log(-log(1-p)*TT)

se<-c()
for(i in 1:length(dia)){
g<-c(gmu[i],gomega[i],glamda[i])
diaspora<-t(g)**z**g
se<-append(se,diaspora^0.5)}
print(zp[1]);print(zp[5]);print(se[5])

```

**Κώδικας 6.** Το επόμενο script κατασκευάζει το Q-Q Plot για την GIII

```

Empirical<-c()
for(i in 1:length(x))
{Empirical<-append(Empirical,omega-(omega-mu)*(-
log(i/(length(x)+1))*TT)^lamda)}
Model<-sort(x)
plot(Model,Empirical,main="1")

```

**Κώδικας 7.** Στην μέθοδο POT τα γραφήματα για την εύρεση του κατωφλίου  $u$ , οι εκτιμήσεις των παραμέτρων της GPD και τα διαστήματα εμπιστοσύνης της  $\hat{x}$  βάση της Profile Deviance Function έγιναν με εντολές των πακέτων fExtremes και extRemes του R-project:

```

mePlot(data[data>3],doplot=TRUE,labels=TRUE)
mr1Plot(data,nint=25)

u<-3.6
b<-gpd.fit(data,u)

```

`gpd.profxi(b,-0.2,0.0)`  
 Ενώ τα Q-Q plot κατασκευάστηκαν με τις εξής εντολές :

```
r<-data[data>u]
Empirical<-c()
for(i in 1:k)
{
Empirical<-append(Empirical,sigma2/xi2*((1-i/(k+1))^-xi2-
1)+u)
}
Model<-sort(r)
plot(Model,Empirical,main=1)
```

**Κώδικας 8.** Εκτίμηση της στάθμης απόδοσης για 10 έως 100 χρόνια απόδοσης και υπολογισμός της διασποράς της με την Μέθοδο Δέλτα. Η μεταβλητή *xronia* παίρνει την τιμή 38 και 96 όταν διαχειριζόμαστε τον πρώτο και δεύτερο κατάλογο σεισμών αντίστοιχα.

```
sigma<-b$mle[1]
xi<-b$mle[2]
k<-b$next
z<-b$cov

n<-length(data)
m<-n/xronia
aa<-(m*k/n)
xm<-u+sigma/xi*(aa^xi-1)

gsigma<-1/xi*(aa^xi-1)
gxi<--sigma/xi^2*(aa^xi-1)+aa^xi*sigma/xi*log(aa)
se<-c()
for(i in 1:10){
g<-c(gsigma[i],gxi[i])
diaspora<-t(g)%*%z%*%g
se<-append(se,diaspora^0.5)
}
print(xm[1]);print(xm[5]);print(se[5])
```

**Κώδικας 9.** Όταν έχουμε λογοκριμένα δεδομένα η συνάρτηση `ll` γίνεται

```
ll<- function(mu=5,omega=12,lamda=0.1)
m*(1/TT)*((omega-xL)/(omega-mu))^(1/lamda)
-sum(-log(lamda*(omega-mu))+(1/lamda-1)*log((omega-x)
/(omega-mu))-(1/TT)*((omega-x)/(omega-mu))^(1/lamda))
fit<-mle(ll)

summary(fit)
```

όπου *xL* το όριο κάτω από το οποίο έχουμε λογοκρισία.

**Κώδικας 10.** Με τις ακόλουθες εντολές κατασκευάζουμε τα Q-Q plot στην περίπτωση που έχουμε λογοκριμένα δεδομένα.

```
Empirical<-c()
for(i in 1:length(x)){
Empirical<-append(Empirical,omegac-(omegac-muc)*(-
12*log(i/(length(x)+1)+(1-i/(length(x)+1))*exp(-
(1/12)*((omegac-4)/(omegac-muc))^(1/lamdac))))^lamdac)
}
Model<-sort(x)
plot(Model,Empirical)
```

## ΒΙΒΛΙΟΓΡΑΦΙΑ

### Ελληνική

Μπούτσικας Μ. (2008) *Θεωρία Ακραίων Τιμών*, Πανεπιστημιακές Σημειώσεις, ΠΜΣ στην Αναλογιστική Επιστ. και Διοικητική Κινδύνου, Παν. Πειραιώς.

### Ξένη

Burton P.W. (1979) Seismic risk in Southern Europe through to India examined using Gumbel's third distribution of extreme values, *Geophys. J.R. Ast Soc* **59**, 249-280.

Burton P.W and Makropoulos, K.C. (1985) Seismic risk of Circum-Pacific earthquakes : II Extreme values using Gumbel's third distribution and the relationship with strain energy release, *Pure Appl. Geophys.* **123**, 849-869.

Burton P.W, Qin G., Tselentis G.A. and Sokos E. (2004) Extremes earthquakes and earthquake perceptibility study in Greece and its surrounding area, *Natural Hazards* **32**, 277-312.

Coles S. (2001) *An introduction to statistical modeling of extreme values*, Springer Series in Statistics.

Embrechts P., Kluppelberg C., Mikosh T. (1997) *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag.

Fisher R.A. and Tippett LHC. (1928) Limiting forms of the frequency distribution of the largest or smallest member of a sample, *Mathematical Proceedings of the Cambridge Philosophical Society* **24**, issue 02, 180-190.

Gertensgarbe F.W. and Worner P.C. (1989) A method for the statistical definition of extreme value regions and their application to meteorological time series, *Z. Meteorol.* **39**, 224-226.

Gnedenko B.V. (1943) Sur la distribution limite du terme maximum d'une serie aleatoire, *Ann. Math.* **44**, 423-453.

Gumbel E.J. (1937) Les intervalles extremes entre les emissions radioactives, *Journal de Physique et le Radium* **8**, 11, 446-452.

Gumbel E.J. (1941) The return period of flood flows, *Annals of Mathematical Statistics*, **12**, 163-190.

Gumbel E.J. (1954) Statistical theory of extreme values and some practical applications, *Nat. Bureau of Standards. Appl. Math. Ser* 33. U.S. Govt. Washington DC.

Gumbel E.J. (1958) *Statistics of extremes*, Columbia University Press, New York.

Gutenberg B. and Richter C.F. (1954) *Seismicity of the earth and associated phenomena*, 2<sup>nd</sup> ed Princeton N.J.: Princeton University Press.

Haan Ferreira A. (2006) *Extreme value theory: An introduction*. Springer Series in Operations Research and Financial Engineering.

Kotz S. and Nadarayah S. (2000) *Extreme value theory and applications*, Imperial College Press, London.

Nordquist J.N. (1945) Theory of largest values applied to earthquake magnitude, *Eos. Trans. Amer. Geophys. Union*, **26**, 29-31.

Phien H. N., Fang and Emma T-S. (1989) Maximum likelihood estimation of the parameters and quantiles of the General Extreme Value Distribution from censored samples, *Journal of Hydrology*, **105**, 139-155.

Pisarenko V.F. and Sornette D. (2003) Characterization of the frequency of extreme earthquake events by the Generalized Pareto Distribution, *Pure and Applied Geophysics* **160**, 2343-2364.

Richter C.F. (1958) *Elementary seismology*, Freeman and Co., San Francisco.

Smith R.L. (1985) Maximum Likelihood Estimation in a Class of Non-Regular Cases, *Biometrika* **72**, 67-92.

Turnbull M.L. and Weatherly D. (2006) Validation of using Gumbel probability plotting to estimate Gutenberg-Richter seismicity parameters, *Proceedings of the AEES*, Australia.

Wang Q.J. (1990) Estimation of the GEV distribution from censored samples by method of partial probability weighted moments, *Journal of Hydrology*, **120**, 103-114.

Weibull (1939) A statistical theory of the strength of materials, *Proc. Royal Swedish Inst. Eng Res* **151**, 1-45.

Yegulap T.M. and Kuo J.T. (1974) Statistical prediction of the occurrence of maximum magnitude earthquakes, *Bull. Seismic. Soc. Am.* **64**, 393-414.