Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ
ΤΟΜΕΑΣ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΕΠΙΧΕΙΡΗΣΙΑΚΗΣ ΕΡΕΥΝΑΣ

Χωρικά Κρυμμένα Μαρκοβιανά Μοντέλα Poisson, με εφαρμογή στη Χαρτογράφηση Ασθενειών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του φοιτητή

Παναγιώτη Ανδρεόπουλου

Επιβλέπουσα: Λουκία Μελιγκοτσίδου

ΑΘΗΝΑ 2012

Η παρούσα Διπλωματική Εργασία
εκπονήθηκε στα πλαίσια των σπουδών
για την απόκτηση του
Μεταπτυχιακού Διπλώματος Ειδίκευσης
στη
Στατιστική & Επιχειρησιακή Έρευνα
που απονέμει το
Τμήμα Μαθηματικών
Του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών

Εγκρίθηκε στις …….. από εξεταστική επιτροπή
αποτελούμενη από τους:

Μελιγκοτσίδου Λουκία	Λέκτορας (επιβλέπουσα)
Μπουρνέτας Απόστολος	Καθηγητής
Σιάννης Φώτιος	Επίκουρος Καθηγητής

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω θερμά την επιβλέπουσα δρ. Λουκία Μελιγκοτσίδου για το ότι κατ' αρχήν δέχτηκε να με αναλάβει στην εκπόνηση της παρούσας διπλωματικής μου, αν και αρκετά πιεσμένη. Γνωρίζω την έλλειψη χρόνου που είχε, όμως το ότι δέχτηκε τελικώς, με τιμά ιδιαίτερα. Την ευχαριστώ για τη βοήθεια που μου προσέφερε, αν και υπήρξαν στιγμές που ένιωσα ότι έφτασα σε αδιέξοδο ως προς τη μελέτη μας, παρόλα αυτά ήταν κι αυτός ένας από τους τρόπους διδασκαλίας της για να με κάνει να το παλέψω μόνος μου, όπου και τελικώς δικαιώθηκε. Σας ευχαριστώ πραγματικά, γιατί μέσα από τα δικά σας επιστημονικά ενδιαφέροντα με κάνατε να δω από άλλη οπτική τα Μαθηματικά και κυρίως την εφαρμόσιμη πλευρά τους στην καθημερινότητά μας. Κίνητρο που θα με οδηγεί στη μετέπειτα επιστημονική πορεία μου. Σας ευχαριστώ!

Δε θα μπορούσα επίσης να μην αναφερθώ στον καθηγητή μου Απόστολο Μπουρνέτα, όπου καθόλα τα έτη των σπουδών μου ήταν παρών σε ότι και αν τον χρειάστηκα. Από επιστημονικές συμβουλές γύρω από τα μαθήματα μέχρι επαγγελματικές για τη συνέχιση της επιστημονικής μου καριέρας. Σας ευχαριστώ πάρα πολύ!

Θα ήθελα να ευχαριστήσω και τους συμφοιτητές μου για την βοήθεια που μου προσέφεραν όποτε τη χρειάστηκα και για το ωραίο περιβάλλον που έχει δημιουργηθεί ανάμεσά μας. Γεγονός το οποίο θεωρώ πολύ σημαντικό για έναν σπουδαστή και όχι μόνο.

Τέλος δε θα μπορούσα να αφήσω στην άκρη την οικογένειά μου και τους δικούς μου ανθρώπους, που στέκονται δίπλα μου σε προσωπικό επίπεδο όλα αυτά τα χρόνια με υπομονή και επιμονή και στηρίζουν τις προσπάθειές μου. Σας ευχαριστώ!

Πίνακας Περιεχομένων

Εισαγωγή

Σε αυτή τη διπλωματική εργασία, γίνεται μία παρουσίαση της Μπεϋζιανής Συμπερασματολογίας για Χωρικά Κρυμμένα Μαρκοβιανά Μοντέλα Poisson διακριτού χρόνου και πεπερασμένου χώρου καταστάσεων, με εφαρμογή στη Χωρική Στατιστική και συγκεκριμένα στη Χαρτογράφηση Ασθενειών. Παρακάτω, γίνεται μία συνοπτική παρουσίαση του περιεχομένου της εργασίας.

Το κεφάλαιο 1 αποτελεί την εισαγωγή μας στον κλάδο της Χωρικής Στατιστικής, παραθέτοντας ορισμούς, έννοιες και αναφορές απαραίτητες για την καλύτερη κατανόηση του εδαφίου, αλλά και του κλάδου αυτού καθ' αυτού.

Στο κεφάλαιο 2, γίνεται αναφορά στη Μπεϋζιανή στατιστική. Περιλαμβάνει τα κύρια στοιχεία της Μπεϋζιανής προσέγγισης, εφαρμογές πάνω στη μίξη Poisson κατανομών, την τεχνική της αύξησης δεδομένων, καθώς και την περιγραφή του αλγορίθμου Gibbs που αποτελεί 'μέρος' της οικογένειας των MCMC αλγορίθμων. Εν συνεχεία, το κεφάλαιο 3, περιγράφει τα Κρυμμένα Μαρκοβιανά Μοντέλα (HMMs), στη συνεχή και τη διακριτή περίπτωση, καθώς γίνεται και μία προσπάθεια να περιγραφούν και να λυθούν τα τρία βασικά προβλήματα που αφορούν στη στατιστική συμπερασματολογία για τα μοντέλα αυτά. Ο αλγόριθμος «εμπρός-πίσω», ο αλγόριθμος EM (κλασσική στατιστική) και ο αλγόριθμος MCMC είναι μέθοδοι εκτίμησης που μας δίνουν λύσεις στα βασικά προβλήματα συμπερασματολογίας για τα κρυμμένα Μαρκοβιανά Μοντέλα.

Όλα τα παραπάνω θα συνδεθούν στο κεφάλαιο 4, όπου γίνεται αναφορά σε μία, ίσως καινούργια έννοια, τη Χαρτογράφηση Ασθενειών. Σε αυτό το κεφάλαιο κάνουμε μία εκτενή αναφορά στην επιστήμη της Χαρτογραφίας και της Ιατρικής και στο πως αυτή εφαρμόζεται και εξελίσσεται σήμερα με τη βοήθεια της Μπεϋζιανής στατιστικής στον ευρύτερο κλάδο των Μαθηματικών. Τα παραπάνω συμπεράσματα συνοψίζονται στο 5^0 και 6^0 κεφάλαιο, που είναι στην ουσία τα κύρια κεφάλαια της εργασίας. Συνδυάζουμε τη θεωρία των Κρυμμένων Μαρκοβιανών Μοντέλων στον κλάδο της Χαρτογράφησης των ασθενειών. Δίνοντας έτσι μία «Μπεϋζιανή» λύση στην εξάπλωση των ασθενειών, παραθέτοντας αποτελέσματα από τις εφαρμογές που έλαβαν χώρα στην περιοχή της Γαλλίας. Γίνεται δηλαδή μία προσπάθεια πρόβλεψης μιας νόσου, καθώς και στις πιθανότητες εξάπλωσής της στις γειτονικές περιοχές.

Κεφάλαιο 1^{ο}

Χωρική Στατιστική ή Χωρική Ανάλυση (Spatial Statistics)

Η Χωρική Στατιστική, περιλαμβάνει στατιστικές τεχνικές που μελετούν οντότητες οι οποίες περιγράφουν τοπολογικές, γεωμετρικές ή γεωγραφικές ιδιότητες. Με τη φράση αυτή αναφερόμαστε σε μία ποικιλία από στατιστικές τεχνικές, οι οποίες εφαρμόζονται σε διάφορους επιστημονικούς τομείς όπως για παράδειγμα η αστρονομία (δηλαδή σε μελέτες για την τοποθέτηση των γαλαξιών στο σύμπαν, με τσιπ μηχανικής κατασκευής). Επίσης χρησιμοποιείται για να περιγράψει τεχνικές που εφαρμόζονται σε δομές στην ανθρώπινη κλίμακα, κυρίως στην ανάλυση γεωγραφικών δεδομένων - τομέας όπου και θα αναπτύξουμε - και για να περιγράψει δεδομένα της γεωστατιστικής.

Το βασικότερο πρόβλημα που αντιμετωπίζουμε στη χωρική ανάλυση, είναι το πρόβλημα του ορισμού της χωρικής θέσης των φορέων που μελετώνται. Για παράδειγμα, μία μελέτη σχετικά με την υγεία του ανθρώπου, θα μπορούσε να περιγράψει τη χωρική θέση των ανθρώπων στο χώρο σε σχέση με ένα σημείο, το σημείο αυτό θα μπορούσε να απεικονίζει τα μέρη όπου ζουν, ή όπου εργάζονται και ενώνοντας τα σημεία αυτά με μία γραμμή θα μπορούσαν να περιγραφούν τα εβδομαδιαία ταξίδια τους. Κάθε επιλογή οδηγεί σε διαφορετικές τεχνικές ανάλυσης οι οποίες καταλήγουν στα ανάλογα συμπεράσματα.

Η χωρική στατιστική μπορεί ίσως να θεωρηθεί ότι έχει προκύψει από τις πρώτες απόπειρες χαρτογράφησης και αποτύπωσης γεγονότων σε πολλούς τομείς, οι οποίοι με τη σειρά τους συνέβαλαν στην περεταίρω εξέλιξή της.

Για παράδειγμα, η *Βιολογία* χρησιμοποιώντας τη χωρική στατιστική, συνέβαλε με τη βοτανική στις μελέτες της παγκόσμιας κατανομής των φυτών, στις τοπικές

παραγωγικές μονάδες, στις μετακινήσεις των ζώων, στις μελέτες βλάστησης και κυρίως στη μελέτη της δυναμικής του πληθυσμού ανά περιοχή.

Η *επιδημιολογία* χαρτογράφησε και μελέτησε τις ασθένειες και ιδίως ο John Snow με το έργο του γύρω από τη χαρτογράφηση σχετικά με το ξέσπασμα της χολέρας συνέβαλε στην εξάπλωση της στατιστικής στην καθημερινότητα.

(Επάνω δεξιά: Παρουσιάζεται ο χάρτης από το Δρ John Snow του Λονδίνου, όπου δείχνει συστάδες των κρουσμάτων χολέρας στην ευρεία περίοδο του 1854 -ξέσπασμα χολέρας . Αυτή ήταν μία από τις πρώτες χρήσεις του χάρτη με βάση τη χωρική ανάλυση).

Άλλοι τομείς όπου έχει συμβάλει η χωρική στατιστική είναι στην οικονομετρία, στο γεωγραφικό σύστημα πληροφοριών, στην επιστήμη των υπολογιστών μέσα από τη μελέτη των αλγορίθμων – κυρίως στην υπολογιστική γεωμετρία.

1.1 Θεμελιώδη ζητήματα στη χωρική στατιστική

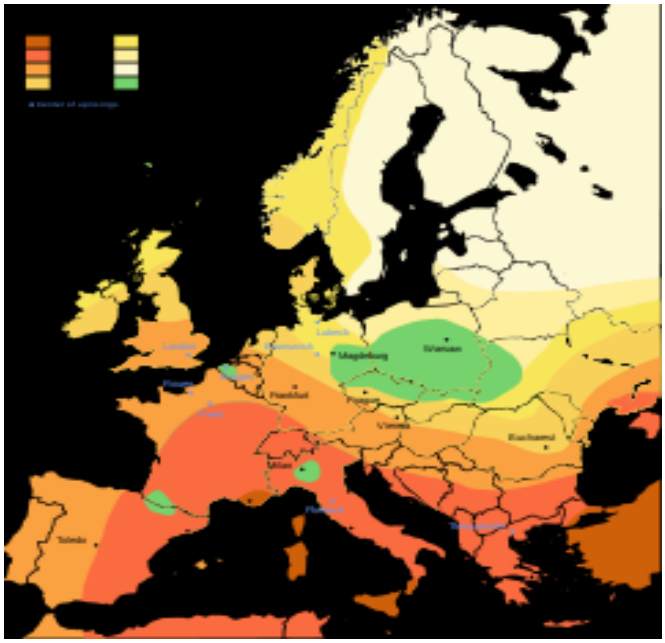
Τα θεμελιώδη ζητήματα που αντιμετωπίζει η χωρική ανάλυση, αναφέρονται στον ορισμό των εννοιών της μελέτης ενός θέματος, εν συνεχεία στην κατασκευή των αναλυτικών εργασιών που πρόκειται να χρησιμοποιηθούν, στη χρήση των υπολογιστών για την ανάλυση, ως προς τους περιορισμούς και τις ιδιαιτερότητες των αναλύσεων τα οποία είναι γνωστά, καθώς και στην παρουσίαση των αναλυτικών αποτελεσμάτων.

Συχνά, στη χωρική στατιστική προκύπτουν σφάλματα, κάποια λόγω των μαθηματικών του χώρου, ενίοτε δε και λόγω των ιδιαίτερων δεδομένων που λαμβάνουν χώρα στη χωροταξία και κάποια λόγω των εργαλείων που είναι διαθέσιμα. Συγκεκριμένα, η επιθυμία μας να γνωρίζουμε ακριβώς το μήκος μιας ακτογραμμής σε ένα νησί απαιτεί ακριβείς μετρήσεις του μήκους της, πράγμα δύσκολο αν όχι αδύνατο! Η επίλυση αυτού του προβλήματος, ήρθε από ένα λογισμικό υπολογιστή, όπου τοποθέτησε ευθείες γραμμές με την καμπύλη της ακτογραμμής και μπόρεσε να υπολογίσει εύκολα τα μήκη των γραμμών που καθορίστηκαν. Ωστόσο, αυτές οι ευθείες γραμμές δεν μπορούν να έχουν καμία

εγγενή έννοια στον πραγματικό κόσμο, όπως φάνηκε για την ακτογραμμή της Μεγάλης Βρετανίας όπου και έγινε αυτή η μελέτη.

Τα προβλήματα αυτά αποτελούν έναν από τους μεγαλύτερους κινδύνους στη χωρική στατιστική λόγω της εγγενούς δύναμης των χαρτών ως μέσα παρουσίασης. Όταν δηλαδή, τα αποτελέσματα παρουσιάζονται ως χάρτες, η παρουσίαση συνδυάζει τα χωρικά δεδομένα τα οποία είναι απολύτως ακριβή, με αναλυτικά αποτελέσματα που μπορεί να είναι ανακριβή.

Βασικό χαρακτηριστικό της χωρικής στατιστικής, είναι ότι οι στατιστικές τεχνικές που ευνοούν τον χωρικό ορισμό των αντικειμένων όπως τα σημεία, είναι πολύ λίγες, διότι υπάρχουν πολύ λίγες στατιστικές τεχνικές που μπορούν να λειτουργήσουν με μία απευθείας γραμμή, περιοχή ή με στοιχεία όγκου. Με τη βοήθεια της πληροφορικής κάποιες από αυτές τις ιδιαιτερότητες αντιμετωπίζονται.



Στον παραπάνω χάρτη παρουσιάζεται η εξάπλωση της βουβωνικής πανώλης στη μεσαιωνική Ευρώπη. Τα χρώματα δείχνουν την χωρική κατανομή των κρουσμάτων πανώλης στην πάροδο του χρόνου.

Αξίζει στο σημείο αυτό να αναφέρουμε για το πρόβλημα της χωρικής εξάρτησης που μπορεί να προκύψει στη μελέτη και ανάλυση καθ'όλη τη διαδικασία της χωρικής στατιστικής/έρευνας. Συγκεκριμένα, μερικές ιδιότητες στο γεωγραφικό χώρο, φαίνεται να σχετίζονται, είτε θετικά είτε αρνητικά. Η χωρική εξάρτηση λοιπόν, οδηγεί στη χωρική αυτοσυσχέτιση γεγονός που στατιστικά μπορεί και να 'ενοχλεί'. Συγκεκριμένα, δημιουργεί δυσκολίες στη στατιστική μοντελοποίηση και ανάλυση χωρικών δεδομένων. Για παράδειγμα, οι αναλύσεις παλινδρόμησης που δεν αντισταθμίζουν χωρική εξάρτηση μπορεί να δώσουν ασταθείς εκτιμήσεις των παραμέτρων οι οποίες με τη σειρά τους θα δώσουν αναξιόπιστες ερμηνείες.

Είναι σκόπιμο επομένως να επεξηγούνται τα παραπάνω προβλήματα με κατάλληλα χωρικά μοντέλα παλινδρόμησης, τα οποία θα μοντελοποιήσουν αυτές τις αδυναμίες και θα δίνουν πιο ορθά και κατάλληλα ερμηνευμένα αποτελέσματα.

Ένα άλλο σημείο που χρήζει ιδιαίτερης προσοχής είναι η χωρική ετερογένεια, μία διαδικασία που αφορά την τοποθεσία στον γεωγραφικό χώρο. Η περίπτωση αυτή απαλείφεται ή πιο ορθά περιορίζεται σημαντικά αν ένας χώρος είναι ενιαίος και απεριόριστος, τότε κάθε θέση θα είχε κάποια μοναδικότητα σε σχέση με τις άλλες περιοχές. Αυτό επηρεάζει τις χωρικές σχέσεις εξάρτησης και, επομένως, τη χωρική διαδικασία. Χωρική ετερογένεια επομένως, σημαίνει ότι οι συνολικές παράμετροι που υπολογίστηκαν για ολόκληρο το σύστημα δεν μπορούν να περιγράψουν επαρκώς τη διαδικασία, για κάθε δεδομένη τοποθεσία.

1.2 Λίγα λόγια για τη μέθοδο Δειγματοληψίας

Η Χωρική δειγματοληψία συνίσταται στον καθορισμό ενός περιορισμένου αριθμού θέσεων στο χώρο για τη μέτρηση φαινομένων που αποτελούν αντικείμενο της εξάρτησης και της ετερογένειας.

Με τον όρο *εξάρτηση* εννοούμε, ότι από ένα σημείο παρατήρησης μπορούμε να προβλέψουμε την αξία του σε μία άλλη θέση, χωρίς να χρειαζόμαστε τις παρατηρήσεις και στα δύο σημεία.

Με τον όρο *ετερογένεια* υποδηλώνουμε ότι αυτή η σχέση μπορεί να αλλάξει στο χώρο, και ως εκ τούτου δεν μπορούμε να εμπιστευτούμε μία παρατηρηθείσα, ως προς το βαθμό εξάρτησης, περιοχή που μπορεί να είναι μικρή.

Τα βασικά συστήματα δειγματοληψίας μπορούν να εφαρμοστούν σε πολλαπλά επίπεδα σε καθορισμένο ιεραρχικά χώρο (π.χ.: αστική περιοχή, πόλη, γειτονιά). Επίσης κατά την ανάλυσή μας μπορούμε να αξιοποιήσουμε και διάφορα βοηθητικά δεδομένα, για παράδειγμα, βλέποντας τις αξίες των ακινήτων ως οδηγό σε ένα χωροταξικό σχέδιο δειγματοληψίας για τη μέτρηση του επιπέδου της εκπαίδευσης και του εισοδήματος.

Στη συνέχεια αξιοποιώντας διάφορα χωρικά μοντέλα, όπως τα στατιστικά μέτρα αυτοσυσχέτισης ή παλινδρόμησης κλπ. υπαγορεύεται και ο κατάλληλος σχεδιασμός του δείγματος.

Σε επόμενο βήμα, κατά τη διαδικασία χωροθέτησης μίας περιοχής, ο ερευνητής μπορεί να υποπέσει σε κάποια χωροταξικά σφάλματα και κατ’επέκταση η έρευνα και το αποτέλεσμα που θα ληφθεί από τη χωρική στατιστική να διαφέρει σημαντικά. Συγκεκριμένα, το θέμα αυτό έφερε στην επιφάνεια ο Benoit Mandelbrot, όπου σε ένα χαρτί σχεδίασε την ακτή της Βρετανίας, θέλοντας να δείξει ότι είναι εκ των πραγμάτων παράλογο να συζητούν ορισμένα σύνολα χωρικών εννοιών. Εξηγώντας ότι το μήκος της ακτογραμμής στην οικολογία εξαρτάται άμεσα από την κλίμακα στην οποία έχει μετρηθεί. Έτσι, ενώ οι μελετητές συνήθως μετρούν

το μήκος ενός ποταμού, το μήκος αυτό έχει νόημα μόνο στο πλαίσιο της σημασίας της τεχνικής μέτρησης για το υπό μελέτη θέμα.

Βρετανία: μέτρηση χρησιμοποιώντας ένα μακρύ κριτήριο (1)

Βρετανία: μέτρηση χρησιμοποιώντας ένα μέσο κριτήριο (2)

Βρετανία: μέτρηση χρησιμοποιώντας ένα μικρό κριτήριο (3)

Ένα άλλο εξίσου σημαντικό στοιχείο που θα πρέπει να επισημανθεί, είναι η πλάνη που μπορεί να υποπέσει ο ερευνητής σε ότι έχει να κάνει με τη χωροθέτηση. Η *γεωγραφική πλάνη (μεροληψία)*, όπως λέγεται αλλιώς, αναφέρεται σε σφάλμα που οφείλεται σε συγκεκριμένο χωρικό χαρακτηρισμό που επιλέχθηκε για τα στοιχεία της μελέτης.

Ένας χωρικός χαρακτηρισμός μπορεί να είναι απλοϊκός ή ακόμα και τελείως λάθος. Οι διάφορες μελέτες συχνά περιορίζουν την χωρική ύπαρξη των ανθρώπων σε ένα μόνο σημείο, για παράδειγμα, τη διεύθυνση κατοικίας τους. Αυτό μπορεί να οδηγήσει σε κακή ανάλυση, για παράδειγμα, κατά την εξέταση της μετάδοσης της νόσου που μπορεί να συμβεί στη δουλειά ή στο σχολείο και, συνεπώς, μακριά από το σπίτι.

Επίσης, ο χωρικός χαρακτηρισμός μπορεί να περιορίσει έμμεσα το αντικείμενο της μελέτης. Για παράδειγμα, η χωρική ανάλυση των δεδομένων της εγκληματικότητας έχει πρόσφατα καταστεί δημοφιλής, αλλά οι μελέτες αυτές μπορούν να περιγράψουν μόνο τα συγκεκριμένα είδη των εγκλημάτων που μπορούν να περιγραφούν χωρικά. Αυτό οδηγεί σε πολλούς χάρτες παρουσίασης των επιθέσεων, αλλά σε κανένα χάρτη παρουσίασης της υπεξαίρεσης με πολιτικές συνέπειες στη

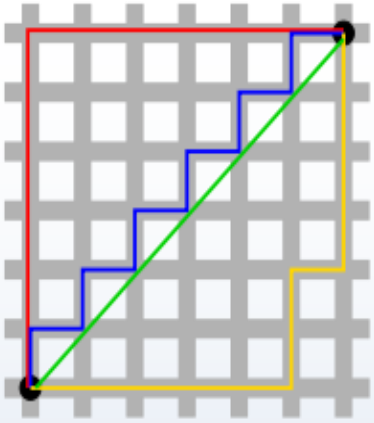
σύλληψη του εγκλήματος και το σχεδιασμό των πολιτικών για την αντιμετώπιση του ζητήματος.

Ένα άλλο ζήτημα που χρειάζεται την προσοχή του ερευνητή, είναι η *οικολογική μεροληψία,* η οποία περιγράφει σφάλματα που οφείλονται σε πραγματικές αναλύσεις σχετικά με τα συγκεντρωτικά δεδομένα όταν προσπαθούν να καταλήξουν σε συμπέρασμα σχετικά με τις μεμονωμένες μονάδες. Για παράδειγμα, ένα pixel αντιπροσωπεύει τις μέσες θερμοκρασίες της επιφάνειας εντός της ίδιας περιοχής. Οικολογική μεροληψία θα είναι να υποθέσουμε ότι όλα τα σημεία εντός της περιοχής έχουν την ίδια θερμοκρασία.

1.3 Λύσεις για τα προβλήματα που προκύπτουν

Ένας μαθηματικός χώρος υπάρχει κάθε φορά που έχουμε ένα σύνολο παρατηρήσεων και ποσοτικές μετρήσεις των χαρακτηριστικών τους. Για παράδειγμα, ο χώρος μπορεί να αντιπροσωπεύει το εισόδημα των φυσικών προσώπων ή τα χρόνια της εκπαίδευσης μέσα σε ένα σύστημα συντεταγμένων όπου η θέση του κάθε ατόμου μπορεί να καθορίζεται σε σχέση με τις δύο διαστάσεις. Οι αποστάσεις μεταξύ των ατόμων σε αυτό το διάστημα είναι ένα ποσοτικό μέτρο των διαφορών τους όσον αφορά το εισόδημα και την εκπαίδευση. Ωστόσο, στην χωρική ανάλυση ασχολούμαστε με συγκεκριμένους τύπους μαθηματικών χώρων, δηλαδή το λεγόμενο: *γεωγραφικό χώρο*.

Στο γεωγραφικό χώρο, οι παρατηρήσεις που αντιστοιχούν σε θέσεις μέσα σε ένα χωροταξικό πλαίσιο μέτρησης αποτελούν συχνά θέσεις στην επιφάνεια της Γης, αλλά αυτό δεν είναι απολύτως αναγκαίο. Ένα χωροταξικό πλαίσιο μέτρησης μπορεί ας πούμε, να λαμβάνει χώρα, στο διαστρικό διάστημα ή μέσα σε βιολογική οντότητα, όπως ένα συκώτι. Όπου αποτελεί και τη βασική αρχή στον πρώτο νόμο της Tobler Γεωγραφίας: αν η αλληλεπίδραση μεταξύ των φορέων αυξάνεται με εγγύτητα στον πραγματικό κόσμο, τότε η εκπροσώπηση στο γεωγραφικό χώρο και η εκτίμηση αυτής με χρήση τεχνικών χωρικής ανάλυσης, είναι κατάλληλες.



Απόσταση Manhattan έναντι της Ευκλείδειας απόστασης: Οι κόκκινες, μπλε και κίτρινες γραμμές έχουν το ίδιο μήκος (12) τόσο στην Ευκλείδεια γεωμετρία, όσο και στη Manhattan. Στην Ευκλείδεια γεωμετρία, η πράσινη γραμμή έχει μήκος 6 X √ 2 ≈ 8,48, και είναι η μοναδική συντομότερη διαδρομή. Στη Manhattan, το μήκος της πράσινης γραμμής εξακολουθεί να είναι 12, που την καθιστά όχι μικρότερη από ό, τι φαίνεται, έναντι κάθε άλλου μονοπατιού.

Τώρα με τη σειρά τους και οι κοντινές αποστάσεις μεταξύ των θέσεων στο χώρο, αντιπροσωπεύονται συχνά από την *Ευκλείδεια απόσταση,* αν και αυτό είναι μόνο μία δυνατότητα. Υπάρχει ένας άπειρος αριθμός αποστάσεων πέραν της Ευκλείδειας που μπορούν να υποστηρίξουν την ποσοτική ανάλυση. Για παράδειγμα, το *Manhattan ή ταξί* , όπου η κίνηση περιορίζεται σε μονοπάτια παράλληλα με τους άξονες και μάλιστα μπορεί να είναι και πιο ουσιαστική από τις Ευκλείδειες αποστάσεις σε αστικές περιοχές.

Εκτός από τις αποστάσεις, και άλλες γεωγραφικές σχέσεις, όπως η συνδεσιμότητα (π.χ.: η ύπαρξη ή ο βαθμός των κοινών συνόρων) και η κατεύθυνση μπορούν επίσης να επηρεάσουν τις σχέσεις μεταξύ των παρατηρήσεων.

Είναι επίσης δυνατόν, να υπολογιστεί το ελάχιστο κόστος μονοπατιών σε όλη την επιφάνεια εξέτασης.

1.4 Χρήσιμοι όροι για την κατανόηση της κάθε έννοιας

Χωρική αυτοσυσχέτιση: Με τη χωρική αυτοσυσχέτιση οι στατιστικοί, μετρούν και αναλύουν τον βαθμό της εξάρτησης μεταξύ των παρατηρήσεων σε ένα γεωγραφικό χώρο. Αυτά τα ζητήματα απαιτούν μία χωρική μέτρηση (μήτρα) που αντικατοπτρίζει την ένταση της γεωγραφικής σχέσης μεταξύ των παρατηρήσεων σε μία γειτονιά, π.χ. οι αποστάσεις μεταξύ των γειτόνων, τα μήκη των κοινών συνόρων που μπορεί να εμπίπτουν σε μία συγκεκριμένη κατεύθυνση. Το κλασικό της χωρικής στατιστικής συσχέτισης, είναι η σύγκριση των χωρικών βαρών στη σχέση συνδιακύμανσης με τα διάφορα ζεύγη θέσεων.

Η χωρική αυτοσυσχέτιση όταν είναι πιο θετική, απ’ ότι αναμενόταν, έχει ως αποτέλεσμα να δείχνει την ομαδοποίηση παρόμοιων τιμών σε γεωγραφικό χώρο, ενώ αν είναι πιο αρνητική δείχνει ότι οι γειτονικές τιμές είναι πιο ανόμοιες από ότι αναμενόταν, πράγμα που θα μπορούσαμε να το προσομοιώσουμε ως ένα πρότυπο μιας σκακιέρας (το παράδειγμα αυτό αναπτύχθηκα από τον Moran’s I).

Τα στατιστικά μέτρα της χωρικής αυτοσυσχέτισης, όπως δόθηκαν από τους Moran’s I και Geary C είναι παγκόσμια, με την έννοια ότι εκτιμούν το συνολικό βαθμό της χωρικής αυτοσυσχέτισης για ένα σύνολο δεδομένων. Η δυνατότητα της χωρικής ετερογένειας, υποδηλώνει ότι η εκτίμηση του βαθμού αυτοσυσχέτισης μπορεί να ποικίλλει σημαντικά μεταξύ των γεωγραφικών χώρων.

Συγκεκριμένα, η ονομαζόμενη και *Τοπική χωρική στατιστική αυτοσυσχέτιση* παρέχει εκτιμήσεις οι οποίες αναλύονται στο επίπεδο των χωρικών μονάδων, επιτρέποντας έτσι την αξιολόγηση των σχέσεων εξάρτησης σε όλο το χώρο.

Χωρική παρεμβολή: Οι μέθοδοι εκτίμησης της χωρικής παρεμβολής των μεταβλητών στις θέσεις, είναι απαραίτητες στη μελέτη του γεωγραφικού χώρου με βάση τις τιμές που παρατηρούνται στις τοποθεσίες. Οι βασικές μέθοδοι,

περιλαμβάνουν τη στάθμιση της αντίστροφης απόστασης: με την οποία μειώνεται η μεταβλητή, με την μείωση της απόστασης από την παρατηρούμενη θέση.

Η μέθοδος Kringing είναι μία πιο σύνθετη μέθοδος που παρεμβάλλεται στο χώρο σύμφωνα με μία χωρική σχέση υστέρησης, που έχει τόσο συστηματικά όσο και τυχαία στοιχεία.

Χωρική παλινδρόμηση: Οι μέθοδοι χωρικής παλινδρόμησης περιλαμβάνουν μία μεταβλητή χωρική εξάρτηση στην ανάλυση παλινδρόμησης, αποφεύγοντας έτσι προβλήματα αστάθειας και αναξιόπιστων εκτιμητών και παρέχουν σημαντικές πληροφορίες σχετικά με τις χωρικές σχέσεις μεταξύ των εμπλεκόμενων μεταβλητών. Ανάλογα με την τεχνική που χρησιμοποιούμε, η χωρική εξάρτηση στο μοντέλο παλινδρόμησης, μπορεί να εισέλθει όπως οι σχέσεις μεταξύ των εξαρτημένων μεταβλητών.

Η Γεωγραφικά σταθμισμένη παλινδρόμηση (GWR) είναι μία χωρική παλινδρόμηση που προκαλεί διαχωρισμό στις παραμέτρους των χωρικών μονάδων ανάλυσης. Αυτό επιτρέπει την εκτίμηση της χωρικής ετερογένειας των εκτιμώμενων σχέσεων μεταξύ των ανεξάρτητων και εξαρτημένων μεταβλητών. Η χρήση των *Markov Chain Monte Carlo (MCMC)* δίνουν τη δυνατότητα εκτίμησης πιο πολύπλοκων λειτουργιών, όπως Poisson-Gamma-CAR, της Poisson λογαριθμικής-SAR, ή Overdispersed μοντέλων logit.

Χωρική αλληλεπίδραση: Η χωρική αλληλεπίδραση ή τα 'μοντέλα βαρύτητας' ουσιαστικά εκτιμούν τη ροή ανθρώπων, υλικών ή πληροφοριών μεταξύ των θέσεων στο χώρο. Υπάρχουν επίσης, παράγοντες που μπορεί να περιλαμβάνονται στην αλληλεπίδραση των μεταβλητών όπως: το πλήθος μετακινούμενων ροών σε κατοικημένες περιοχές, όπως το ύψος των γραφειακών χώρων στις περιοχές απασχόλησης, όπως ακόμα και η απόσταση οδήγησης ή χρόνος ταξιδιού. Επιπλέον, οι τοπολογικές σχέσεις μεταξύ των περιοχών, πρέπει να προσδιορίζονται λαμβάνοντας ιδίως υπόψην τις συχνά αντικρουόμενες σχέσεις μεταξύ της τοπολογικής απόστασης. Για παράδειγμα, δύο χωρικά κοντινές γειτονιές δεν μπορεί να εμφανίσουν κάποια σημαντική αλληλεπίδραση όταν χωρίζονται από έναν αυτοκινητόδρομο.

Αφού λοιπόν, προσδιοριστούν οι λειτουργικές μορφές των σχέσεων αυτών, ο αναλυτής, μπορεί να εκτιμήσει τις παραμέτρους του μοντέλου χρησιμοποιώντας τα δεδομένα των ροών – παρατηρήσεων και τις αντίστοιχες τυποποιημένες τεχνικές εκτίμησης, όπως είναι η απλή μέθοδος των ελαχίστων τετραγώνων ή η εκτίμηση μέγιστης πιθανοφάνειας.

Υπάρχουν όμως και ανταγωνιστικές εκδόσεις χωρικών μοντέλων, οι οποίες με τη σειρά τους αντιμετωπίζουν τα όποια προβλήματα εξετάζοντας κατά κύριο λόγο την απόσταση μεταξύ των προορισμών, σε συνδυασμό με την εγγύτητα της προέλευσης. Αυτό οδηγεί σε μία μέθοδο ομαδοποίησης των ροών. Αυτές οι υπολογιστικές μέθοδοι, όπως είναι τα 'τεχνητά νευρωνικά δίκτυα' μπορούν επίσης

να υπολογίσουν τις χωρικές σχέσεις αλληλεπίδρασης μεταξύ των διαφόρων θέσεων και μπορούν να χειριστούν επίσης τα αντίστοιχα «θορυβώδη» και ποιοτικά δεδομένα.

Προσομοίωση και Μοντελοποίηση: Τα χωρικά μοντέλα αλληλεπίδρασης παρουσιάζονται (μοντελοποιούνται) από πάνω προς τα κάτω: καθορίζεται με αυτό τον τρόπο μία συνολική σχέση που διέπει τη ροή ανάμεσα στις τοποθεσίες. Αυτό είναι άλλωστε και το χαρακτηριστικό για τα κοινά αστικά μοντέλα, όπως εκείνα που βασίζονται στον μαθηματικό προγραμματισμό, στις ροές μεταξύ των οικονομικών τομέων ή στη θεωρία της προσφοράς/ενοικίασης.

Τα προσαρμοστικά συστήματα θεωρίας όπως εφαρμόζονται στη χωρική ανάλυση δείχνουν ότι οι απλές αλληλεπιδράσεις μεταξύ των εγγύς παρατηρήσεων μπορεί να οδηγήσουν σε περίπλοκες, επίμονες χωρικές ενότητες σε συνολικό επίπεδο.

Δύο είναι οι βασικές μέθοδοι προσομοίωσης: 1) τα κυτταρικά αυτόνομα μοντέλα που επιβάλλουν ένα χωροταξικό πλαίσιο. Τα λεγόμενα και 'κύτταρα δικτύου' όπου προσδιορίζουν τους κανόνες που υπαγορεύουν την κατάσταση ενός κελιού που βασίζεται στις γειτονιές των γειτονικών κυττάρων του. Με το πέρασμα του χρόνου όμως, προκύπτουν νέα χωρικά μοντέλα, αφού τα κύτταρα διαμορφώνουν τις γειτονιές με βάση τους γείτονές τους, γεγονός που αλλάζει τις προϋποθέσεις για μελλοντικές χρονικές περιόδους. Για παράδειγμα, ως κύτταρα μπορούμε να αναπαραστήσουμε τις θέσεις σε μία αστική περιοχή και τα κράτη τους μπορεί να είναι διάφορα είδη χρήσης γης.

2) Τα μοτίβα, που μπορεί να προκύψουν από τις απλές αλληλεπιδράσεις των τοπικών χρήσεων γης. Συπεριλαμβάνοντας τις λεγόμενες περιοχές office και την εξάπλωση των πόλεων.

Να επισημάνουμε ότι, σε αντίθεση με τα κύτταρα σε 'κυτταρικά δίκτυα', διάφοροι άλλοι παράγοντες μπορεί να είναι κινητοί σε σχέση με το χώρο. Για παράδειγμα, θα μπορούσε κανείς να διαμορφώσει τη ροή της κυκλοφορίας και τη δυναμική που χρησιμοποιούν οι εκπρόσωποι μεμονωμένων οχημάτων που προσπαθούν να ελαχιστοποιήσουν το χρόνο ταξιδιού καθορίζοντας αφετηρίες και προορισμούς. Γενικά, επιδίωξή μας είναι ο ελάχιστος χρόνος ταξιδιού, ο καθορισμός των παραγόντων που θα μας βοηθήσει να αποφύγουμε μία πιθανή σύγκρουση με άλλα οχήματα κλπ.

Τα 'κυτταρικά δίκτυα' και τα διάφορα μοντέλα είναι οι λεγόμενες στρατηγικές μοντελοποίησης. Οι οποίες μπορούν να ενταχθούν σε ένα κοινό γεωγραφικό σύστημα όπου κάποιοι παράγοντες είναι σταθεροί και άλλοι κινητοί.

Σημείο Γεωστατιστικής (MPS): Η χωρική ανάλυση του κάθε εννοιολογικού γεωλογικού μοντέλου, είναι ο κύριος σκοπός του κάθε αλγορίθμου MPS. Η μέθοδος δηλαδή, που αναλύει στατιστικά στοιχεία των χωρικών γεωλογικών μοντέλων, ονομάζεται ως «κατάρτιση εικόνας».

1.5 Η Γεωγραφική επιστήμη των πληροφοριών σε συνδυασμό με τη Χωρική Στατιστική – Βασικό πλαίσιο

Τα Γεωγραφικά συστήματα πληροφοριών (GIS) και η βασική γεωγραφική επιστήμη των πληροφοριών, έχουν μία ισχυρή επιρροή στην χωρική ανάλυση. Η μεγάλη ικανότητα που έχουν να συλλαμβάνουν και να χειρίζονται γεωγραφικά δεδομένα σημαίνει ότι η χωρική ανάλυση συμβάλλει όλο και περισσότερο στην τροφοδοσία των δεδομένων για τις επιστήμες αυτές. Τα Γεωγραφικά συστήματα συλλογής δεδομένων περιλαμβάνουν τις τηλεσκοπικές απεικονίσεις, τα συστήματα περιβαλλοντικής παρακολούθησης, όπως είναι τα ευφυή συστήματα μεταφορών και η θέση-επίγνωση του αντικειμένου, ή όπως είναι οι φορητές συσκευές που μπορούν να αναφέρουν την τοποθεσία του αντικειμένου σε σχεδόν πραγματικό χρόνο.

Τα GIS επομένως, παρέχουν πλατφόρμες για τη διαχείριση αυτών των δεδομένων, τον υπολογισμό των χωρικών σχέσεων, όπως η απόσταση, η συνδεσιμότητα και η κατεύθυνση των σχέσεων μεταξύ των χωρικών μονάδων. Και όλα αυτά δίνοντας τη δυνατότητα απεικόνισης σε ένα χαρτογραφικό περιβάλλον απεικονίζοντας τόσο τα πρωτογενή δεδομένα, όσο κυρίως τα χωρικά αποτελέσματα.

Ορισμοί & έννοιες:

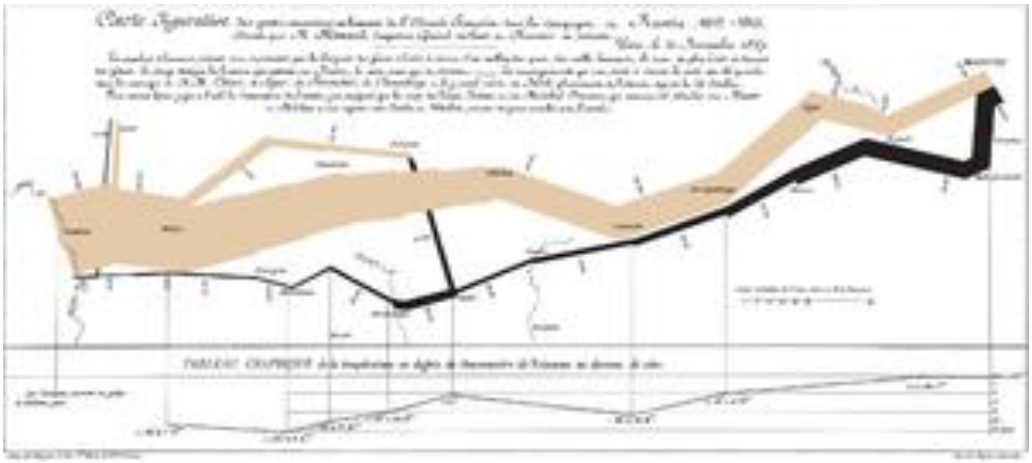
Χωροθέτηση: Μεταφορά της θέσης/πληροφορίας των διαστημικών αντικειμένων με τη βοήθεια του χώρου στις συντεταγμένες του συστήματος.

Χωρική κατανομή: Παρόμοιες ομάδες χωρικών πληροφοριών τοποθέτησης αντικειμένων, συμπεριλαμβανομένης της διανομής, των τάσεων κλπ.

Χωρική μορφή: Το γεωμετρικό σχήμα των χωρικών αντικειμένων.

Χωρικές σχέσεις: Η σχέση μεταξύ χωροαντικειμένων, συμπεριλαμβανομένων των τοπογραφικών, του προσανατολισμού, των ομοιοτήτων κτλ.

Geovisualization (GVis): Συνδυάζει την επιστημονική οπτικοποίηση, με ψηφιακές χαρτογραφήσεις για την υποστήριξη της έρευνας και της ανάλυσης των γεωγραφικών δεδομένων και πληροφοριών, συμπεριλαμβανομένων των αποτελεσμάτων της χωρικής ανάλυσης και προσομοίωσης. Το GVis αξιοποιεί τον ανθρώπινο προσανατολισμό κατευθύνοντάς τον προς την οπτική επεξεργασία των πληροφοριών στον τομέα της έρευνας, ανάλυσης και επικοινωνίας των γεωγραφικών δεδομένων και πληροφοριών. Σε αντίθεση με τις παραδοσιακές μορφές χαρτογραφίας, το GVis έχει συνήθως τρείς ή τέσσερις διαστάσεις (η οποία περιλαμβάνει χρονικό διάστημα) και τον διαδραστικό χρήστη.



Αυτός ο χάρτης απεικονίζει τη ροή της πορείας του Ναπολέοντα στην Μόσχα. Είναι ένα πρώιμο και γνωστό παράδειγμα geovisualization. Δείχνει την κατεύθυνση του Ναπολέοντα καθώς ταξίδευε, με τα στρατεύματα του, από πού πέρασαν, το μέγεθος του στρατού, πόσοι στρατιώτες πέθαναν από την πείνα, τις πληγές, και τις χαμηλές θερμοκρασίες που βίωσαν.

Γεωγραφική ανακάλυψη γνώσης (GKD): Είναι η ανθρωποκεντρική διαδικασία εφαρμογής αποτελεσματικών υπολογιστικών εργαλείων για την εξερεύνηση μαζικών χωρικών βάσεων δεδομένων. Η GKD περιλαμβάνει τη γεωγραφική εξόρυξη δεδομένων, αλλά περιλαμβάνει και πιο συναφείς δραστηριότητες, όπως η επιλογή των δεδομένων, ο καθορισμός των δεδομένων, καθώς και την επεξεργασία και ερμηνεία των αποτελεσμάτων.

Η GVis έχει κεντρικό ρόλο στη διαδικασία της GKD. Η GKD βασίζεται στην παραδοχή ότι οι μαζικές βάσεις δεδομένων περιέχουν ενδιαφέροντα πρότυπα τυποποιημένων αναλυτικών τεχνικών.

Η GKD μπορεί να χρησιμεύσει και ως μια διαδικασία υπόθεσης δημιουργίας για τη χωρική ανάλυση, αφού παράγει δοκιμαστικά σχέδια και επιβεβαιώνει τις σχέσεις που θα πρέπει να έχουν οι χωρικές αναλυτικές τεχνικές.

Χωρικά Συστήματα Υποστήριξης Αποφάσεων (SDSS): Λαμβάνουν υφιστάμενα χωρικά δεδομένα και χρησιμοποιούν μια ποικιλία από μαθηματικά μοντέλα για να κάνουν προβλέψεις για το μέλλον. Αυτό ουσιαστικά επιτρέπει σε αστικό και περιφερειακό σχεδιασμό, την παρέμβαση στις αποφάσεις, πριν από την εφαρμογή.

Κεφάλαιο 2^{ο}

Μπεϋζιανή Στατιστική (Bayesian Inference)

Υπάρχουν διάφορες πιθανοθεωρητικές προσεγγίσεις. Η κάθε μία από αυτές εξηγεί και περιγράφει με διαφορετικό τρόπο την τυχαιότητα. Οι βασικές θεωρίες είναι:

1. Αυτή η οποία αναφέρεται σε ισοπίθανα ενδεχόμενα (Laplace).
2. Εκείνη η οποία υπολογίζει παρατηρηθείσες σχετικές συχνότητες (Von Misses).
3. Αυτή που στηρίζεται στην υποκειμενική αξιολόγηση καταστάσεων με ταυτόχρονη χρήση προσωπικής γνώμης (Μπεϋζιανή).

Το βασικό μειονέκτημα της προσέγγισης ***Laplace*** είναι ότι αναφέρεται μόνο σε ισοπίθανα ενδεχόμενα. Πολλά όμως από τα φαινόμενα αβεβαιότητας που μας απασχολούν δεν είναι δυνατόν να αντιμετωπισθούν με την προσέγγιση αυτή. Δεν έχουμε, δηλαδή, δεδομένα για συχνότητα εμφάνισης καταστάσεων που δεν έχουν συμβεί στο παρελθόν. Για παράδειγμα, δεν μπορούμε να μιλάμε για την σχετική συχνότητα σε περιπτώσεις που θέλουμε να προβλέψουμε το αποτέλεσμα των επόμενων εκλογών. Δεν έχει έννοια, επομένως, να προσπαθήσουμε να τις προβλέψουμε χρησιμοποιώντας τις σχετικές συχνότητες από την απόδοση των κομμάτων που θα πάρουν μέρος στις επόμενες εκλογές σε σχέση με τις προηγούμενες εκλογικές αναμετρήσεις. Το ίδιο συμβαίνει όταν αναφερόμαστε στην επιλογή αγοράς ή πώλησης μετοχών, κλπ.

Οι θεωρίες που βασίζονται σε ισοπίθανα ενδεχόμενα ή στην έννοια της σχετικής συχνότητας δεν είναι χρήσιμες στις παραπάνω περιπτώσεις. Πρόκειται για περιπτώσεις όπου υπάρχει ενδιαφέρον για τη μελέτη της αβεβαιότητας. Οι πιθανότητες που εκφράζουν την αβεβαιότητα αυτή φαίνεται να είναι εκ των πραγμάτων, υποκειμενικές. Το καλύτερο που μπορεί να κάνει κάποιος σε τέτοιες περιπτώσεις, είναι να συγκεράσει τις διαθέσιμες ενδείξεις και τα λογικά επιχειρήματα και να καταλήξει σε μία προσωπική πιθανότητα ύπαρξης του κάθε ενδεχομένου. Η ιδέα της προσωπικής πιθανότητας έχει επεκταθεί, διαμορφωθεί και βελτιωθεί από πολλούς Στατιστικούς που χρησιμοποιούν την Μπεϋζιανή προσέγγιση ως βάση για την ανάπτυξη των απόψεών τους. Όσοι υποστηρίζουν την Μπεϋζιανή προσέγγιση ισχυρίζονται ότι η προσωπική πιθανότητα για ένα ενδεχόμενο καθορίζεται από τη διάθεση κάθε ατόμου να στοιχηματίσει για το ενδεχόμενο αυτό.

2.1 Παράθεση της Κλασσικής με την Μπεϋζιανή Στατιστική

Σε αντίθεση με την Κλασσική Στατιστική, η οποία αποτελεί τη βάση της στατιστικής θεωρίας από την εποχή που ο Fisher παρουσίασε τις πρώτες στατιστικές

έννοιες, η Μπεϋζιανή Στατιστική προσέγγιση στην συμπερασματολογία έγινε πρόσφατα ιδιαίτερα δημοφιλής. Ένας λόγος της ανάπτυξης αυτής οφείλεται στις εξελίξεις γύρω από τις υπολογιστικές μεθόδους (κυρίως μεθόδους Markov Chain Monte Carlo) που έχουν επιτρέψει σε αρκετούς επιστήμονες να χρησιμοποιούν Μπεϋζιανές μεθόδους στην ανάλυση δεδομένων. Παραδείγματα τέτοιων εφαρμογών παρουσιάζονται αναλυτικά στα βιβλία των Gelman et al. (1995) και Carlin and Louis (1996) κυρίως σε πρακτικές εφαρμογές των MCMC.

Στη συνέχεια παραθέτουμε περιληπτικά τη σύγκριση ανάμεσα στις δύο μεθόδους θεωρώντας για το σκοπό αυτό ένα παράδειγμα πάνω σε δείγμα από *n* ανεξάρτητες μεταβλητές:

Κλασσική Στατιστική

i. ***Κλασσική Συμπερασματολογία***: Έστω $Y_1, Y_2, \ldots, Y_n$ είναι n ανεξάρτητες τυχαίες μεταβλητές καθεμία από τις οποίες ακολουθεί την κανονική κατανομή με μέση τιμή μ και διακύμανση σ^2, όπου σ^2-γνωστό. Στόχος της μεθοδολογίας είναι η συμπερασματολογία για την δεδομένη αλλά άγνωστη παράμετρο μ βασισμένη πάνω σε ένα δείγμα παρατηρηθέντων τιμών $y_1, y_2, \ldots, y_n$.

ii. ***Σημειακή Εκτίμηση***: Ο δειγματικός μέσος, $\overline{Y} = \left(\frac{1}{n}\right)\sum_i Y_i$ είναι μία φυσική εκτιμήτρια για την μέση τιμή του πληθυσμού. Η κλασσική προσέγγιση αξιολογεί κάθε εκτιμήτρια βασιζόμενη σε ιδιότητες οι οποίες ισχύουν κάτω από επαναλαμβανόμενη δειγματοληψία από το ίδιο μοντέλο με σταθερές τιμές των αγνώστων παραμέτρων. Να αναφέρουμε πως ένα πολύ σημαντικό εργαλείο είναι η αμεροληψία, σύμφωνα με την οποία $E(\bar{y}) = \mu$, που ιδιότυπα ενισχύει την άποψη υπέρ της χρήσης του δειγματικού μέσου ως εκτιμήτριας για την άγνωστη μέση τιμή μ.

iii. ***Εκτίμηση με Διαστήματα Εμπιστοσύνης***: Η δειγματική κατανομή του μέσου $\overline{Y} \sim \mathcal{N}\left(\mu, {\sigma^2}/{n}\right)$. Η κατανομή αυτή είναι εκείνη που θα παρατηρούσαμε σε επαναλαμβανόμενη δειγματοληψία με δείγματα μεγέθους n από ένα κανονικό πληθυσμό μέσης τιμής μ και διακύμανσης σ^2. Από την δειγματική αυτή κατανομή, όπως είναι γνωστό, κατασκευάζεται το διάστημα εμπιστοσύνης $\left(\overline{Y} - 1.96\,\sigma/\sqrt{n}\ ,\ \overline{Y} + 1.96\sigma n\right.$ το οποίο περιέχει την πραγματική μέση τιμή του πληθυσμού στις 95% των περιπτώσεων επαναλαμβανόμενης δειγματοληψίας. Το διάστημα αυτό ονομάζεται συνήθως το 95% διάστημα εμπιστοσύνης για την μέση τιμή μ του πληθυσμού.

iv. ***Έλεγχοι Υποθέσεων***: Η συνήθης προσέγγιση στην κλασσική θεώρηση ξεκινά με μια μηδενική υπόθεση και μια εναλλακτική υπόθεση. Στη συνέχεια, με την χρήση μιας κατάλληλης στατιστικής συνάρτησης ελέγχου,

διαμορφώνουμε μια διαδικασία που μας επιτρέπει να καθορίσουμε την καταλληλότητα της μηδενικής υπόθεσης. Η καταλληλότητα αυτή μετριέται με την τιμή *p-value*, η οποία αναφέρεται στην πιθανότητα ότι, σε επαναλαμβανόμενη δειγματοληψία, θα οδηγηθούμε σε μια τιμή της στατιστικής συνάρτησης ελέγχου τόσο ακραία, από την παρατηρηθείσα τιμή της υποθέτοντας ότι ισχύει η μηδενική υπόθεση. Μικρές τιμές *p-value* αποτελούν ένδειξη ότι τα δεδομένα τα οποία χρησιμοποιήσαμε δεν είναι συνήθη υπό την μηδενική υπόθεση, το οποίο αποτελεί ένδειξη ότι η μηδενική υπόθεση ίσως δεν είναι σωστή. Να σημειώσουμε, ότι ο έλεγχος δεν μεταχειρίζεται την μηδενική και εναλλακτική υπόθεση συμμετρικά και ότι το p-value υπολογίζεται κάτω από την παραδοχή ότι ισχύει η μηδενική υπόθεση.

<u>Μπεϋζιανή Στατιστική</u>

i. ***Μπεϋζιανή Συμπερασματολογία***: α) Όλες οι άγνωστες ποσότητες αντιμετωπίζονται ως τυχαίες μεταβλητές, ενώ χρησιμοποιούνται κατανομές πιθανότητας για να περιγράψουν την κατάσταση της γνώσης μας για τις άγνωστες αυτές ποσότητες. β) Η συμπερασματολογία για τις άγνωστες ποσότητες γίνεται με βάση τον κανόνα του Bayes, που επιτρέπει την χρήση πιθανοτήτων δεσμευμένων επί των τιμών που παρατηρήθηκαν.
Ποιοτικά, η Μπεϋζιανή προσέγγιση ξεκινά με μια κατανομή πιθανότητας η οποία περιγράφει το επίπεδο της γνώσης μας αναφορικά με τις άγνωστες ποσότητες πριν συλλεγούν δεδομένα και στην συνέχεια χρησιμοποιεί τα παρατηρηθέντα δεδομένα για να επανακαθορίσει την κατανομή αυτή.
Έστω λοιπόν, $Y_1, Y_2, \dots, Y_n$ τ.δ. από την $\mathcal{N}(\mu, \sigma^2)$ κατανομή. Η περιθώρια κατανομή του μ, συμβολίζεται ως p(μ) και ονομάζεται *εκ των προτέρων κατανομή (prior distribution)* του μ. Περιγράφει την κατάσταση της γνώσης μας για το μ πριν παρατηρήσουμε οποιαδήποτε δεδομένα.
Η συμπερασματολογία μας, προκύπτει από τους νόμους των πιθανοτήτων και την χρησιμοποίηση του θεωρήματος του Bayes

$$\mathrm{p}(\mu|y_1, y_2, \dots, y_n) = \frac{p(y_1,\dots,y_n|\mu)p(\mu)}{p(y_1,\dots,y_n)}$$

όπου $\mathrm{p}(y_1, \dots, y_n)$ είναι η περιθώρια κατανομή των δεδομένων, η οποία προκύπτει από την κατανομή $\mathrm{p}(y_1, \dots, y_n|\mu)$ και p(μ).
Το αποτέλεσμα που προκύπτει από την χρήση του κανόνα του Bayes, είναι γνωστό ως *εκ των υστέρων κατανομή (posterior distribution)* του μ και περιγράφει την κατάσταση της γνώσης μας για το μ αφού παρατηρήσουμε τα $y_1, \dots, y_n$.
Να αναφέρουμε ότι στη μέθοδο αυτή, επικεντρώνουμε την προσοχή μας στο συγκεκριμένο δείγμα που έχουμε διαθέσιμο.

ii. ***Σημειακή Εκτίμηση***: Η εκ των υστέρων κατανομή περιγράφει τη γνώση μας για το μ αφού παρατηρηθούν τα δεδομένα καθορίζοντας ποιες τιμές είναι περισσότερο εύλογες και πόσο πιθανή είναι καθεμία από αυτές. Για να επιλεγεί μία μόνο σημειακή εκτίμηση κάτω από την Μπεϋζιανή προσέγγιση,

χρησιμοποιούμε μία συνάρτηση απώλειας (loss function) που καθορίζει το κόστος ενός λάθους στην εκτίμηση. Στη συνέχεια, επιλέγουμε ως εκτίμησή μας την τιμή που ελαχιστοποιεί το αναμενόμενο κόστος (expected loss) κάτω από την εκ των υστέρων κατανομή.

iii. ***Εκτίμηση με Διαστήματα Εμπιστοσύνης***: Η εκ των υστέρων κατανομή μας επιτρέπει να προσδιορίσουμε διαστήματα που περιέχουν το μ με οποιαδήποτε καθορισμένη πιθανότητα. Τα διαστήματα αυτά ονομάζονται *σύνολα αξιοπιστίας (credible sets)*. Η Μπεϋζιανή προσέγγιση κάνει σαφή χρήση της θεωρίας πιθανοτήτων προκειμένου να καταλήξει σε πιθανοθεωρητικά συμπεράσματα για την άγνωστη παράμετρο δοθέντος ενός μοναδικού συγκεκριμένου δείγματος.

iv. ***Έλεγχοι Υποθέσεων***: Για τον έλεγχο μιας υπόθεσης έναντι μιας άλλης στην Μπεϋζιανή προσέγγιση υπάρχει μια πιο τυπική διαδικασία η οποία είναι γνωστή ως ο παράγοντας Bayes (the Bayes factor).

v. ***Εκ των Προτέρων Κατανομές (prior distributions)***: Η Μπεϋζιανή προσέγγιση αποφεύγει μέρος της εννοιολογικής δυσκολίας η οποία σχετίζεται με την ερμηνεία των διαστημάτων εμπιστοσύνης και των παρατηρούμενων επιπέδων σημαντικότητας (p-values). Το 'κόστος' που υφίσταται κανείς για να αποκτήσει αυτά τα πλεονεκτήματα είναι εκείνο που αναφέρεται στον καθορισμό της εκ των προτέρων κατανομής για την άγνωστη παράμετρο. Αξίζει να παραθέσουμε ορισμένα χαρακτηριστικά της, όπως ότι σε μεγάλα δείγματα, η εκ των προτέρων κατανομή καθίσταται ανευ σημασίας. Δηλαδή, μετά από ένα αρκετά μεγάλο αριθμό επαναλήψεων των μετρήσεών μας, στο όριο, η εκ των υστέρων κατανομή θα συμπεριφέρεται ως εάν δεν υπήρχαν εκ των προτέρων πληροφορίες.

   Η Μπεϋζιανή προσέγγιση απαιτεί ότι η εκ των προτέρων κατανομή πιθανότητας, που θα θεωρήσουμε είναι μια ειλικρινής αποτίμηση των εκ των προτέρων απόψεών μας σχετικά με τις παραμέτρους του μοντέλου. Για το λόγο αυτό η επιλογή μιας εκ των προτέρων κατανομής γίνεται ευκολότερη αν εκμεταλλευτούμε την ύπαρξη συζυγών οικογενειών, οικογενειών δηλαδή εκ των προτέρων κατανομών οι οποίες, συνδυαζόμενες με μία κατανομή δεδομένων, οδηγούν σε εκ των υστέρων κατανομές της ίδιας οικογένειας (απλοποιώντας με αυτό τον τρόπο τους υπολογισμούς μας).

2.2 Μπεϋζιανή Στατιστική

Με βάση το θεώρημα του Bayes αναπτύχθηκε μία ιδιαίτερη προσέγγιση της στατιστικής, που ονομάστηκε *στατιστική κατά Bayes ή Μπεϋζιανή στατιστική*.Ενώ στην κλασσική στατιστική η συμπερασματολογία βασίζεται στην πιθανοφάνεια (likelihood), στην Μπεϋζιανή στατιστική βασίζεται στην εκ των υστέρων κατανομή πιθανότητας (posterior distribution) των παραμέτρων. Στην ουσία, η εκ των υστέρων κατανομή πιθανότητας είναι η συμπερασματολογία. Με βάση το θεώρημα

Bayes η εκ των υστέρων κατανομή πιθανότητας του διανύσματος παραμέτρων θ, δίνεται ως:

$$f(\theta|x) = \frac{f(\theta)f(x|\theta)}{\sum_{\theta\in\Theta} f(\theta)f(x|\theta)},$$

όπου x οι παρατηρήσεις μας (διάνυσμα)

Θ ο χώρος από τον οποίο λαμβάνει τιμές η παράμετρος θ, ο οποίος μπορεί να είναι διακριτός ή συνεχής (στην περίπτωση αυτή το άθροισμα γίνεται ολοκλήρωμα)

$f(\theta)$ η εκ των προτέρων κατανομή της θ, η οποία εκφράζει τις αρχικές μας πεποιθήσεις ή γνώσεις μας για τις παραμέτρους και

$f(x|\theta)$ η πιθανοφάνεια.

Για κάθε τιμή της θ παρατηρούμε ότι ο παρονομαστής παραμένει σταθερός, άρα μπορούμε να γράψουμε:

$$f(\theta|x) = \frac{f(\theta)f(x|\theta)}{c} \quad \text{ή}$$

$$f(\theta|x) = \sigma\tau\alpha\theta\varepsilon\rho\acute{\alpha} \times f(\theta) \times f(x|\theta) \quad \text{ή}$$

$$f(\theta|x) \propto f(\theta) \times f(x|\theta)\,,$$

όπου c η λεγόμενη σταθερά κανονικοποίησης και $'\propto'$ το σύμβολο της αναλογίας των δύο ποσοτήτων. Με τον τρόπο αυτό αποφεύγουμε τον υπολογισμό της σταθεράς, που σε πολύπλοκα στατιστικά προβλήματα είναι δύσκολος αν όχι αδύνατος $(a \propto b \Leftrightarrow c \in \Re: a = c \cdot b \; \gamma\iota\alpha \; b \neq 0)$.

Τελειώνοντας, στην Μπεϋζιανή στατιστική όλες οι παράμετροι θεωρούνται τυχαίες μεταβλητές και η συμπερασματολογία γίνεται από το γινόμενο

$$(\pi\iota\theta\alpha\nu o\varphi\acute{\alpha}\nu\varepsilon\iota\alpha) \times (\varepsilon\kappa \; \tau\omega\nu \; \pi\rho o\tau\acute{\varepsilon}\rho\omega\nu \; \kappa\alpha\tau\alpha\nu o\mu\acute{\eta} \; \pi\iota\theta\alpha\nu\acute{o}\tau\eta\tau\alpha\varsigma)$$

και όχι μόνο από την πιθανοφάνεια. Στο σημείο αυτό να σημειώσουμε ότι η προσοχή που πρέπει να δίνουμε στην εκ των προτέρων κατανομή, είναι μεγάλη. Όμως η επιρροή της μειώνεται όσο το πλήθος των παρατηρήσεων μεγαλώνει.
Επομένως, κινούμενοι με τον Μπεϋζιανό τρόπο ακολουθούμε τα εξής βήματα:

1. Υπολογισμός πιθανοφάνειας (likelihood).
2. Ορισμός της εκ των προτέρων κατανομής πιθανότητας (prior).
3. Υπολογισμός της εκ των υστέρων κατανομής πιθανότητας (posterior).
4. Τέλος, εξάγουμε συμπερασματολογία από την ποσότητα αυτήν.

Παράδειγμα 2.1

Ας υποθέσουμε ότι έχουμε ένα δείγμα από *n* ανεξάρτητες τυχαίες μεταβλητές $x = (x_1, x_2, \dots, x_n)$ που ακολουθούν κατανομή Poisson με παράμετρο ϑ.
Για μία παρατήρηση x_i του δείγματος η πιθανοφάνεια του ϑ είναι:

$$f(x|\theta) = \frac{\theta^x e^{-\theta}}{x!}\ , \qquad \theta \geq 0$$

Ο μέσος και η διακύμανση της κατανομής αυτής είναι:

$$E(x) = \theta$$

$$Var(x) = \theta \quad \text{αντίστοιχα.}$$

Ενώ για όλο το δείγμα η πιθανοφάνεια του ϑ είναι:

$$f(x|\theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!}$$

$$= \frac{e^{-n\theta}\theta^{\Sigma x_i}}{\prod x_i!}$$

$$\propto e^{-n\theta}\theta^{\Sigma x_i}$$

Γενικά, η επιλογή της εκ των προτέρων (prior) κατανομής της παραμέτρου ϑ μπορεί να ποικίλει από πρόβλημα σε πρόβλημα, και εξ ορισμού θα εξαρτάται από την έκταση της εκ των προτέρων γνώσης μας σχετικά με την κατάσταση. Ωστόσο, η διαδικασία την οποία ακολουθούμε, βασίζεται σε μία πιθανή οικογένεια prior κατανομών οι οποίες, οδηγούν σε απλούς μαθηματικούς υπολογισμούς.

Το σημείο το οποίο θίγεται εδώ, είναι πως δεδομένου ότι η οικογένεια είναι αρκετά μεγάλη και κατά συνέπεια καλύπτει επαρκώς ένα μεγάλο εύρος από πιθανές μορφές κατανομών, μπορούμε να χρησιμοποιήσουμε την εκ των προτέρων πεποίθησή μας μέσα από αυτήν την οικογένεια η οποία είναι αρκετά κοντά στις prior πεποιθήσεις μας. Εάν, παρόλα αυτά, δεν υπάρχει prior κατανομή μέσα στην οικογένεια αυτή η οποία ανακλά τι πραγματικά πιστεύουμε, τότε θα πρέπει να αποφύγουμε την προσέγγιση αυτή.

Έτσι, στο παράδειγμά μας, ας υποθέσουμε ότι μπορούμε να εκφράσουμε τις πεποιθήσεις μας σχετικά με την παράμετρο ϑ μέσω της κατανομής *Gamma*, δηλαδή $\theta \sim Gamma(p, q)$, έτσι ώστε:

$$f(\theta) = \frac{q^p}{\Gamma(p)}\theta^{p-1}exp\{-q\theta\}\ \ \gamma\iota\alpha\ \theta > 0$$

τότε εξάγουμε ως εκ των υστέρων κατανομή του $\vartheta|x$ την

$$f(\theta|x) \propto f(\theta) \times f(x|\theta)$$

$$\propto \frac{q^p}{\Gamma(p)}\theta^{p-1}exp\{-q\theta\} \times exp\{-n\theta\}\theta^{\sum x_i}$$

$$\propto \theta^{(p+\sum x_i-1)}exp\{-(q+n)\theta\}$$

και επομένως:

$$\theta \,|\, x \sim Gamma\left(p + \sum_{i=1}^{n} x_i\,, q+n\right).$$

Η οποία είναι μια νέα Γάμμα κατανομή της οποίας οι παράμετροι έχουν καθοριστεί με βάση τα δεδομένα του $\sum x_i$ και του *n*.

Κατ' αυτόν τον τρόπο, με προσεκτική επιλογή, έχουμε βρει την εκ των υστέρων (posterior) κατανομή η οποία ανήκει στην ίδια οικογένεια όπως και η εκ των προτέρων κατανομή, και μέσω αυτής της διαδικασίας έχουμε αποφύγει να κάνουμε υπολογισμούς ολοκληρωμάτων. Η επίδραση των δεδομένων είναι η αναπροσαρμογή των παραμέτρων της κατανομής Γάμμα από την αρχική τιμή του *p* στις posterior τιμές των $(p + \sum_{i=1}^{n} x_i\,, q+n)$.

2.3 Επιλογή της εκ των Προτέρων Κατανομής

Το παράδειγμα 2.1, ήταν μια καλή ευκαιρία, για να δούμε πρακτικά πόσο σημαντική είναι η εκ των προτέρων κατανομή (prior) και πόσο αυτή επηρεάζει τα αποτελέσματά μας. Γι'αυτό το λόγο, η επιλογή της πρέπει να γίνεται προσεχτικά και με συγκεκριμένα κριτήρια. Ένα από αυτά είναι η σχέση της με την εκ των υστέρων κατανομή (posterior).

Η επιλογή μιας συζυγούς εκ των προτέρων κατανομής, εφόσον αυτή είναι δυνατή και ταιριάζει με τις αρχικές μας πεποιθήσεις, μας διευκολύνει πολύ στην τελική μας συμπερασματολογία και απλοποιεί πολλές φορές ακόμη και τον υπολογισμό της σταθεράς κανονικοποίησης, που είναι εν γένει δύσκολος. Σε περίπτωση που δεν υπάρχουν αρχικές πεποιθήσεις, δεν έχουμε δηλαδή καμία αρχική πληροφόρηση, τότε μπορούμε να πάρουμε μία μη πληροφοριακή εκ των προτέρων κατανομή, όπως είναι η $f(\theta) \propto 1$.

Συνήθως η παράμετρος είναι διάνυσμα, της μορφής $\theta = (\theta_1, \theta_2, \dots, \theta_v)$, οπότε πρέπει εδώ να επιλέξουμε μία πολυδιάστατη εκ των προτέρων κατανομή. Αυτή μπορεί να δηλώνει συσχέτιση ανάμεσα στις συνιστώσες του ϑ ή να υποθέτει εκ των προτέρων ανεξαρτησία των παραμέτρων. Οπότε παίρνουμε εκ των προτέρων κατανομή:

$$f(\theta) = f(\theta_1) \cdot f(\theta_2) \cdot \dots \cdot f(\theta_v)$$

όπου $f(\theta_i)$ η εκ των προτέρων κατανομή για την *i* συνιστώσα του ϑ. Ως επί το πλείστον, επιλέγουμε μία εκ των προτέρων κατανομή που είναι συζυγής ή 'βολική' για υπολογισμούς. Για παράδειγμα, αν η παράμετρος είναι της μορφής $\vartheta=(\vartheta_1,\vartheta_2)$, τότε μπορούμε να πάρουμε

$$f(\theta) = f(\theta_1) \cdot f(\theta_2) \quad \text{ή} \quad f(\theta) = f(\theta_1) \cdot f(\theta_2|\theta_1).$$

2.4 Μπεϋζιανή Στατιστική και MCMC

Στην Μπεϋζιανή Στατιστική όλα τα συμπεράσματα εξάγονται από την εκ των υστέρων κατανομή των παραμέτρων. Γενικά, για τον πλήρη υπολογισμό της εκ των υστέρων κατανομής απαιτείται ιδιαίτερη προσπάθεια, εκτός από κάποιες πολύ απλές μορφές. Το πρόβλημα γίνεται ακόμα πιο περίπλοκο αν η παράμετρος είναι πολυδιάστατη (σε ορισμένες περιπτώσεις είναι αδύνατος). Στις περιπτώσεις αυτές εφαρμόζουμε μεθόδους προσομοίωσης, διαδικασίες δηλαδή που μας δίνουν τυχαίες τιμές από την από κοινού posterior κατανομή.

Μια οικογένεια τέτοιων μεθόδων είναι οι Μαρκοβιανές αλυσίδες Μόντε Κάρλο *(Markov chain Monte Carlo - MCMC)*. Μία μέθοδος MCMC στηρίζεται στην κατασκευή μιας Μαρκοβιανής αλυσίδας με στάσιμη κατανομή την από κοινού εκ των υστέρων κατανομή των παραμέτρων. Μέλος αυτής της οικογένειας είναι και ο αλγόριθμος Metropolis – Hastings, ειδική περίπτωση του οποίου αποτελεί ο δειγματολήπτης Gibbs *(Gibbs Sampler).*

2.4.1 Ο Αλγόριθμος Metropolis-Hastings

Η Monte Carlo μεθοδολογία για την δημιουργία αριθμητικών προσεγγίσεων διαφόρων τιμών της εκ των υστέρων κατανομής, όπως του μέσου και της τυπικής απόκλισης, στηρίζεται στους Ασθενείς νόμους των Μεγάλων αριθμών: σε *iid* (*iid*: ανεξάρτητα & ισόνομα) δείγμα, οι Monte Carlo εκτιμητές είναι συνεπείς, που σημαίνει ότι είναι πολύ κοντά στις πραγματικές τιμές με υψηλή πιθανότητα, καθώς ο αριθμός των επαναλήψεων m τείνει στο άπειρο. Οι Metropolis et al. προσπάθησαν να επιτύχουν το ίδιο αποτέλεσμα για ένα μη *iid* Monte Carlo δείγμα.

Συγκεκριμένα, με τη μέθοδο απόρριψης στο χρονικό σημείο t προσομοιώνουμε τιμή θ^* από την κατανομή εισήγησης $g(\theta|y)$ και την αποδεχόμαστε ή την απορρίπτουμε σύμφωνα με την πιθανότητα αποδοχής $p(\theta^*|y)$. Αν τη δεχτούμε μετακινούμαστε στο θ^*, αλλιώς προσομοιώνουμε νέα τιμή. Με τον τρόπο αυτό δημιουργούμε μία *iid* σειρά προσομοιωμένων τιμών από την εκ των υστέρων κατανομή $p(\theta|y)$.

Οι Metropolis et al. γενίκευσαν την παραπάνω ιδέα σε περιπτώσεις όπου το *iid* δείγμα είναι δύσκολο. Επέτρεψαν στην κατανομή εισήγησης στον χρόνο t να

εξαρτάται από την τωρινή κατάσταση θ_t της αλυσίδας και εν συνεχεία, για να επιτύχουν την ζητούμενη στάσιμη κατανομή, όταν μία προτεινόμενη τιμή απορρίπτονταν ανάγκαζαν την αλυσίδα να μείνει στην κατάσταση που βρισκόταν άλλη μία επανάληψη.

Η αλυσίδα που καταλήγουμε τότε είναι Μαρκοβιανή, αφού (α) οι τιμές είναι εξαρτημένες αλλά (β) από όλο το 'παρελθόν' της μόνο η πιο πρόσφατη κατάσταση καθορίζει το ΄μέλλον'.

Με την παραπάνω μέθοδο υπάρχει μεγάλη ελευθερία στην επιλογή της κατανομής εισήγησης $g(\theta^*|\theta_t, y)$, όπου με θ^* συμβολίζουμε την προτεινόμενη τιμή και με θ_t την τωρινή κατάσταση. Η αρχική ιδέα των Metropolis et al. ήταν η χρησιμοποίηση συμμετρικής κατανομής εισήγησης, δηλαδή $g(\theta^*|\theta_t, y) = g(\theta_t|\theta^*, y)$, αλλά ο Hastings το 1970 γενίκευσε την ιδέα αυτή για μη συμμετρικές κατανομές εισήγησης, δημιουργώντας τον αλγόριθμο Metropolis – Hastings. Βασιζόμενος στις ιδέες των Metropolis et al. ο Hastings απέδειξε ότι καταλήγουμε στη σωστή στάσιμη κατανομή αρκεί να χρησιμοποιήσουμε ως πιθανότητα αποδοχής την ακόλουθη:

$$p(\theta^*|\theta_t, y) = min\left\{1, \frac{\frac{p(\theta^*|y)}{g(\theta^*|\theta_t, y)}}{\frac{p(\theta_t|y)}{g(\theta_t|\theta^*, y)}}\right\}. \qquad (1)$$

Έχει αρκετό ενδιαφέρον να συγκρίνουμε τη συγκεκριμένη πιθανότητα αποδοχής με αυτή από τη μέθοδο απόρριψης. Η ουσιαστική διαφορά τους είναι πως η κατανομή εισήγησης τώρα δεν είναι σταθερή αλλά αλλάζει κάθε φορά.

Επίσης παρατηρούμε, πως η σχέση (1) είναι μία γενίκευση της πιθανότητας αποδοχής της μεθόδου απόρριψης: η νέα πιθανότητα αποδοχής μπορούμε να πούμε πως είναι το πηλίκο δύο πιθανοτήτων αποδοχής της μεθόδου απόρριψης, μίας που έχει να κάνει με το που είσαι τώρα και μίας που έχει να κάνει με το που σκέφτεσαι να πας (είναι επιπλέον ισοδύναμο να δουλεύουμε με τη g μιας και στην περίπτωση αυτή η σταθερά κανονικοποίησης θα απαλειφθεί στο πηλίκο).

Αξιοσημείωτο είναι το γεγονός ότι για οποιαδήποτε κατανομή εισήγησης η στάσιμη κατανομή θα είναι η εκ των υστέρων p. Όταν η κατανομή εισήγησης είναι συμμετρική τότε η πιθανότητα αποδοχής ισούται με $p(\theta^*|y)\big/p(\theta_t|y)$, που σημαίνει πως θέλουμε να επισκεφτούμε σημεία με μεγαλύτερη συχνότητα πιο συχνά. Η επιλογή της g δεν επηρεάζει την σύγκλιση στην εκ των υστέρων κατανομή. Επηρεάζει όμως την ταχύτητα σύγκλισης και πόσο καλά η αλυσίδα αναμιγνύεται (μίξη).

Έχει δειχθεί ότι σε απλά προβλήματα με προσεγγιστικά κανονικές εκ των υστέρων κατανομές η βέλτιστη πιθανότητα αποδοχής είναι περίπου 44%. Ενώ πέρα από μονοδιάστατο Θ, ο αλγόριθμος δουλεύει και για πολυδιάστατο Θ. Το μεγάλο

προτέρημα είναι ότι για να εφαρμόσουμε τον αλγόριθμο αρκεί να γνωρίζουμε την εκ των υστέρων χωρίς την σταθερά κανονικοποίησης.

Συγκεντρωτικά, όταν το $\Theta = (\theta_1, \theta_2, \dots, \theta_k)$ τότε μπορούμε να εφαρμόσουμε τον αλγόριθμο M-H και να προσομοιώσουμε διανύσματα Θ ή μπορούμε να δημιουργήσουμε ξεχωριστά τις συνιστώσες βάση διαφορετικών κατανομών εισήγησης. Κάθε δηλαδή επανάληψη του αλγορίθμου αποτελείτε από k βήματα, στην αρχή δηλαδή της επανάληψης t, ανανεώνουμε πρώτα το θ_1, μετά το θ_2 και στο τέλος το θ_k. Εν συνεχεία:

- Καλούμε $\theta_{t,i}$ την τιμή της i συνιστώσας στο χρόνο t και $\Theta_{t,-i} = (\theta_{t+1,1}, \theta_{t+2,2}, \dots, \theta_{t+1,i-1}, \theta_{t,i+1}, \dots, \theta_{t,k})$.
- Τέλος, το υποψήφιο σημείο θ_i^* παράγεται από την κατανομή εισήγησης $g(\theta_i^* | \theta_{t,i}, \theta_{t,-i}, y)$ με πιθανότητα αποδοχής:

$$p(\theta_i^* | \theta_{t,i}, \theta_{t,-i}, y) = min\left[1, \frac{p(\theta_i^* | \theta_{t,-i}, y) g(\theta_{t,i} | \theta_i^*, \theta_{t,-i}, y)}{p(\theta_{t,i} | \theta_{t,-i}, y) g(\theta_i^* | \theta_{t,i}, \theta_{t,-i}, y)}\right].$$

2.4.2 Ο Δειγματολήπτης Gibbs

Ο Δειγματολήπτης Gibbs βασίζεται στον ομώνημο αλγόριθμο, ενός από τους μεγαλύτερους Αμερικανούς επιστήμονες όλων των εποχών, του Josiah Willard Gibbs (1983 – 1904). Τη σημερινή του όμως τελειοποίηση, την οφείλει στους αδερφούς Geman, που την παρουσίασαν το 1984.

Κατά τη οποία, αν έχουμε ένα τυχαίο δείγμα *n* παρατηρήσεων, $x = (x_1, x_2, \dots, x_n)$, από μία κατανομή με παράμετρο $\theta = (\theta_1, \theta_2, \dots, \theta_d)$, με συνάρτηση πιθανοφάνειας $f(x|\theta)$, εκ των προτέρων κατανομή για το ϑ την $f(\theta)$ και γνωστές τις δεσμευμένες κατανομές $f(\theta_i | x, \theta_{-i})$, όπου θ_{-i} το διάνυσμα των παραμέτρων χωρίς την *i* συντεταγμένη, για *i=1,2,…,n* τότε αν δεν μπορούμε ή αν δε θέλουμε να υπολογίσουμε την εκ των υστέρων κατανομή, προσομοιώνουμε τιμές από αυτήν.

Ο Δειγματολήπτης Gibbs καθορίζεται από τα ακόλουθα βήματα:

- Παίρνουμε αρχικές τιμές για το θ, έστω $\theta^{(0)} = \left(\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_d^{(0)}\right)$,
- Για *j=1* έως *s* επαναλαμβάνουμε:
  - προσομοιώνουμε το $\theta_1^{(j)}$ από την $f\left(\theta_1 \middle| x, \theta_2^{(j-1)}, \theta_3^{(j-1)}, \dots, \theta_d^{(j-1)}\right)$,
  - προσομοιώνουμε το $\theta_2^{(j)}$ από την $f\left(\theta_2 \middle| x, \theta_1^{(j)}, \theta_3^{(j-1)}, \dots, \theta_d^{(j-1)}\right)$,
    ….
  - προσομοιώνουμε το $\theta_d^{(j)}$ από την $f\left(\theta_d \middle| x, \theta_1^{(j)}, \theta_2^{(j)}, \dots, \theta_{d-1}^{(j)}\right)$,

  λαμβάνουμε επομένως το διάνυσμα $\theta^{(j)} = \left(\theta_1^{(j)}, \theta_2^{(j)}, \dots, \theta_d^{(j)}\right)$.

Με τον τρόπο αυτό, έχουμε δημιουργήσει μία Μαρκοβιανή αλυσίδα,

$$\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(k)}, \theta^{(k+1)}, \dots, \theta^{(s)}$$

η οποία κάτω από τις κατάλληλες συνθήκες ομαλότητας και μετά από μία περίοδο «ζεστάματος» *(burn in period)*, π.χ. k βημάτων, μπορεί να αποδειχθεί ότι συγκλίνει στην εκ των υστέρων κατανομή των παραμέτρων. Τα $\theta^{(i)}$ δηλαδή, για *i>k,* μπορούν να θεωρηθούν ως πραγματοποιήσεις της εκ των υστέρων κατανομής. Έτσι, έχοντας ένα δείγμα από την κατανομή μας μπορούμε να προχωρήσουμε σε οποιαδήποτε συμπερασματολογία επιθυμούμε, προσεγγίζοντας όσο το δυνατόν περισσότερο την ποσότητα που εξετάζουμε.

Ενώ να σημειώσουμε, ότι οι επιμέρους τιμές $\theta_i^{(1)}, \theta_i^{(2)}, \dots, \theta_i^{(k)}, \theta_i^{(k+1)}, \dots, \theta_i^{(s)}$ για $i = 1,2, \dots d$ μπορούν να θεωρηθούν δείγμα από την περιθώρια κατανομή $f(\theta_i|x)$, πάλι μετά από μία περίοδο «ζεστάματος».

2.5 Τεχνική Αύξησης Δεδομένων

Η αύξηση δεδομένων είναι μία τεχνική που εφαρμόζεται όταν κάποια δεδομένα λείπουν *(missing data problems)* ή όταν η συνάρτηση πιθανοφάνειας είναι δύσκολη στο χειρισμό και δε μας επιτρέπει να εφαρμόσουμε το δειγματολήπτη Gibbs. Στην περίπτωση αυτή, εισάγουμε νέες άγνωστες ποσότητες τις οποίες αντιμετωπίζουμε ως τυχαίες μεταβλητές. Με τον τρόπο αυτό, πετυχαίνουμε να γίνει πιο εύκολη στο χειρισμό η συνάρτηση πιθανοφάνειας, επιτρέποντάς μας την εφαρμογή του δειγματολήπτη Gibbs.

Έτσι, ενώ η πιθανοφάνεια $f(y|\theta)$, με *y* το διάνυσμα των παρατηρήσεων και ϑ το διάνυσμα των παραμέτρων, μας δυσκολεύει, εισάγοντας z μη παρατηρούμενα *(unobserved)* ή χαμένα *(missing)* δεδομένα, η από κοινού πιθανοφάνεια $f(z, y|\theta)$ είναι πιο εύκολη στο χειρισμό. Στη συνέχεια, θα παρουσιάσουμε την εφαρμογή της τεχνικής αύξησης δεδομένων στην πιο χαρακτηριστική κατηγορία προβλημάτων που αυτή χρησιμοποιείται: στις πεπερασμένες μίξεις κατανομών.

2.5.1 Μίξη Poisson Κατανομών

Η εφαρμογή των μιγμάτων κατανομών σε πραγματικά προβλήματα, έχει οδηγήσει σε ένα μεγάλο αριθμό από επιστημονικές εργασίες στη βιβλιογραφία της Στατιστικής. Λόγω της ευελιξίας τους τα μικτά μοντέλα χρησιμοποιούνται ολοένα και περισσότερο ως ένας βολικός και εύχρηστος τρόπος να μοντελοποιήσουμε άγνωστες δομές κατανομών. Μπορούν να μοντελοποιήσουν καταστάσεις τις οποίες ένα απλό μοντέλο δεν μπορεί να περιγράψει. Για παράδειγμα, παρόλο που αν υποθέσουμε θεωρητικά μια συγκεκριμένη κατανομή $F(.|\theta)$ για ένα σύνολο δεδομένων, η σχέση του δειγματικού μέσου και της διασποράς είναι δεδομένη για

αυτή την κατανομή, στην πραγματικότητα όμως η σχέση αυτή μπορεί να μην ισχύει. Σε αυτές τις περιπτώσεις είναι προφανής η ανάγκη για μία πιο γενική οικογένεια κατανομών. Μία τέτοια ευέλικτη οικογένεια προκύπτει αν θεωρήσουμε την παράμετρο (ή τις παραμέτρους) θ από την αρχική κατανομή η οποία θα παίρνει τιμές σύμφωνα με μία κατανομή με συνάρτηση πυκνότητας πιθανότητας ας πούμε $g(\theta)$.

Τα μίγματα κατανομών χρησιμοποιούνται ευρύτατα σε ποικίλα και διαφορετικά πεδία εφαρμογών της Στατιστικής. Στη Στατιστική μοντελοποίηση δεδομένων, έχοντας στη διάθεσή μας ένα σύνολο παρατηρήσεων, συνήθως χρειάζεται να 'ταιριάξουμε' σε αυτό μια κατανομή, η οποία θα μπορεί να περιγράψει δεδομένα. Συχνά στην προσπάθεια αυτή επιστρατεύουμε μίγματα κατανομών. Η αρνητική διωνυμική κατανομή η οποία αποτελεί μίξη της κατανομής Poisson με τη Γάμμα ως κατανομής μίξης, αποτελεί ένα τυπικό παράδειγμα.

Στην ανάλυση κατά ομάδες, η ιδέα είναι να περιγραφεί ολόκληρος ο πληθυσμός σαν ένα μικτό μοντέλο το οποίο να περιλαμβάνει διάφορους υποπληθυσμούς (clusters). Τότε προσαρμόζοντας ένα μικτό μοντέλο μπορούμε να πάρουμε μία εκ των υστέρων πιθανότητα που να ανήκει η κάθε παρατήρηση σε κάθε έναν από τους υποπληθυσμούς.

Στην ανάλυση διασποράς, η τεχνική είναι μία συγκεκριμένη εφαρμογή των μιγμάτων κατανομών. Το μοντέλο υποθέτει ότι η μέση τιμή της κανονικής κατανομής ολόκληρου του πληθυσμού, κυμαίνεται από υποπληθυσμό σε υποπληθυσμό. Τότε θεωρούμε ότι η μέση τιμή είναι από μόνη της μία κανονική μεταβλητή και η ανάλυση της συνολικής διασποράς γίνεται βάση της τυχαιότητας και της μίξης. Γενικά, όλα τα μοντέλα που επηρεάζονται με μικτό τρόπο είναι μίγματα κατανομών.

Στην κατανομή Poisson, που εξετάζουμε κυρίως, υποθέτουμε ότι έχουμε έναν ομοιογενή πληθυσμό, στην πράξη, θα ήταν πιο ρεαλιστικό να υποθέσουμε ότι ο πληθυσμός είναι ανομοιογενής. Αυτή η ανομοιογένεια θα εκφράζεται μέσω της παραμέτρου λ (ρυθμός) την οποία πλέον θα θεωρούμε τυχαία μεταβλητή και θα παίρνει τιμές σύμφωνα με μία δική της κατανομή ανάλογα σε πιο στοιχείο του πληθυσμού αναφερόμαστε. Προκύπτει ότι, από το νόμο της ολικής πιθανότητας, η κατανομή του ολόκληρου πληθυσμού θα είναι ένα μίγμα κατανομών Poisson.

Ανάλογα με την επιλογή της κατανομής μίξης, ένας τεράστιος αριθμός από μικτές κατανομές μπορεί να προκύψει. Δυστυχώς, πολύ λίγες από αυτές, έχουν χρησιμοποιηθεί στην πράξη και η μια ελάχιστη μειοψηφία αυτών έχουν μελετηθεί σε βάθος.

Από την έρευνα των Greenwood και Yule (1920) για τα μίγματα κατανομών Poisson και έκτοτε, ένας μεγάλος αριθμός μιγμάτων κατανομών Poisson, έχει εμφανιστεί στην βιβλιογραφία. Παρόλα αυτά, πολύ λίγα από αυτά έχουν προσελκύσει το ενδιαφέρον των ερευνητών της εφαρμοσμένης στατιστικής. Ο

κύριος λόγος είμαι ότι συχνά η μορφή τους είναι πολύπλοκη και επομένως αποθαρρυντική για έναν ερευνητή να τη χρησιμοποιήσει.

Από τον αρκετά μεγάλο αριθμό μίξεων Poisson που υπάρχουν, αναφερόμαστε μόνο σε αυτές με συνεχή κατανομή μίξης. Οι διακριτές κατανομές μίξης έχουν εξεταστεί στη βιβλιογραφία, αλλά όχι με λεπτομέρεια. Η πρακτική τους αξία οφείλεται σε δύο κυρίως παράγοντες. Ο πρώτος είναι ότι μπορούν να περιγράψουν έναν πληθυσμό, ο οποίος περιέχει πεπερασμένο αριθμό υποπληθυσμών. Ο δεύτερος είναι ότι, θεωρώντας ένα μικτό μοντέλο Poisson, μπορούμε να εκτιμήσουμε μόνο μία πεπερασμένη κατανομή μίξης, δηλαδή ακόμα και αν η κατανομή μίξης είναι συνεχής, μπορούμε να την εκτιμήσουμε μέσω μιας κατανομής πεπερασμένου βήματος.

2.5.2 Πεπερασμένες μίξεις κατανομών Poisson

Έστω ότι γνωρίζουμε ότι ο πληθυσμός αποτελείται από k υποπληθυσμούς κάθε ένας από τους οποίους έχει συνάρτηση πυκνότητας πιθανότητας ή συνάρτηση πιθανότητας μιας παραμετρικής μορφής με διαφορετικές παραμέτρους έστω $f(x|\theta_j),\ j = 1,2,\dots,k$. Τότε η τυχαία μεταβλητή x έχει συνάρτηση πυκνότητας πιθανότητας (ή συνάρτηση πιθανότητας αν η x είναι διακριτή τυχαία μεταβλητή) της μορφής

$$f(x) = f_P(x) = \sum_{j=1}^{k} \pi_j f(x|\theta_j)$$

όπου $0 \leq \pi_j \leq 1, j = 1,2,\dots,k$ με $\sum_{j=1}^{k} \pi_j = 1$ και το ϑ μπορεί να είναι είτε μονοδιάστατο (όπως στην περίπτωση της Poisson) είτε διάνυσμα παραμέτρων (όπως στην περίπτωση της κανονικής κατανομής). Τα π_j καλούνται βάρη μίξης και μπορούν να θεωρηθούν ως πιθανότητες μια τυχαία επιλεγμένη παρατήρηση να ανήκει στον j-οστό υποπληθυσμό. Η κατανομή μίξης δίνει θετική μάζα πιθανότητας π_j στα σημεία $\theta_j, j = 1,2,\dots,k$ και μηδέν αλλού. Αν υποθέσουμε ότι $f(x|\theta)$ είναι η κατανομή Poisson, το διάνυσμα των παραμέτρων ϑ είναι απλά η παράμετρος λ της κατανομής Poisson. Ενώ η *κ*-πεπερασμένη μίξη Poisson με P την κατανομή μίξης, δίνεται ακολούθως, ως:

$$f_P(x) = \sum_{j=1}^{k} \pi_j \frac{\exp(-\lambda_j)\lambda_j^x}{x!}$$

με $x = 0,1,\dots$ και $\lambda_j > 0, j = 1,2,\dots,k$

Για να είναι τα πεπερασμένα μίγματα Poisson προσδιορίσιμα πρέπει να επιβάλλουμε τον περιορισμό $0 < \lambda_1 < \lambda_2 \ldots < \lambda_k$ και ο λόγος είναι ότι διαφορετικά μεταθέτοντας τις συνιστώσες θα οδηγούμασταν στο ίδιο μίγμα.

Τέλος, μέση τιμή και διασπορά του μίγματος Poisson είναι:

$$E(x) = \sum_{j=1}^{k} \pi_j \lambda_j \qquad \kappa\alpha\iota \qquad V(x) = \sum_{j=1}^{k} \pi_j \lambda_j + \sum_{j=1}^{k} {\pi_j}^2 \lambda_j$$

Παράδειγμα 2.2 *(Αύξηση Δεδομένων για πεπερασμένα μίγματα Poisson)*

Η τεχνική αύξησης δεδομένων είναι μία τεχνική που μπορεί να είναι χρήσιμη, αφού κάνει τα προβλήματα ικανά να επιλυθούν από το δειγματολήπτη Gibbs. Βασικά αυξάνει τις παραμέτρους του μοντέλου, με τυχαίες μεταβλητές, κάνοντάς το πιο εύκολο υπολογιστικά.

Ένα παράδειγμα (περίπτωση) είναι μία πιθανότητα ενός μοντέλου σε πρώτη φάση ανάγνωσης να είναι δύσκολο να υπολογιστεί, αλλά αν προσθέσεις μία συλλογή από τυχαίες μεταβλητές αυτή απλοποιείται υπολογιστικά, υπό όρους.

- Υποθέτουμε ένα δείγμα $\{y_1, y_2, \ldots, y_k\}$ από ανεξάρτητες και ισόνομες παρατηρήσεις από τη μίξη κατανομών

$$f(y_i|\lambda_1, \lambda_2) = \frac{1}{2}\left(\frac{e^{-\lambda_1}{\lambda_1}^{y_i}}{y_i!} + \frac{e^{-\lambda_2}{\lambda_2}^{y_i}}{y_i!}\right).$$

Αυτή η μίξη κατανομών μπορεί να θεωρηθεί ως μίξη μιας τυχαίας μεταβλητής της οποίας οι τιμές λαμβάνονται από την ακόλουθη διαδικασία.

Έχουμε ένα δίκαιο νόμισμα, τότε το δείγμα δοκιμών που έφερε κεφάλι ακολουθεί $y_i \sim Poisson(\lambda_1)$ ενώ αν το δείγμα δοκιμών δώσει γράμματα ακολουθεί $y_i \sim Poisson(\lambda_2)$. Για k ανεξάρτητες παρατηρήσεις, η πιθανοφάνεια του $\lambda = (\lambda_1, \lambda_2)$ θα δίνεται από

$$f(y|\lambda_1, \lambda_2) = \prod_{i=1}^{k} f(y_i|\lambda_1, \lambda_2)$$

$$\propto \prod_{i=1}^{k} \frac{(e^{-\lambda_1}{\lambda_1}^{y_i} + e^{-\lambda_2}{\lambda_2}^{y_i})}{y_i!}.$$

Αν τώρα θελήσουμε να υπολογίσουμε την εκ των υστέρων κατανομή του λ, αυτός ο υπολογισμός γίνεται αρκετά δύσκολος (ακόμα και στην περίπτωση της κλασσικής στατιστικής). Ακόμα δηλαδή και αν υπολογίσουμε τους k παράγοντες, έναν-έναν ξεχωριστά, θα καταλήξουμε σε ένα άθροισμα παραγόντων της μορφής 2^k, όπου είναι σχεδόν αδύνατο να υπολογιστούν.

Υποθέτουμε τώρα, ότι ξέρουμε ποια είναι η ακολουθία 'κεφαλών' και 'γραμμάτων' που δίνουν οι ρίψεις. Δηλαδή ξέρουμε, ποιες παρατηρήσεις προήλθαν από την κατανομή $Poisson(\lambda_1)$ και ποιες από την $Poisson(\lambda_2)$. Στο σημείο αυτό, παρατηρούμε, ότι η διάκριση μεταξύ των δεδομένων και των παραμέτρων είναι κάπως «θολή» σε μία Μπεϋζιανή προσέγγιση: και οι δύο αντιμετωπίζονται ως τυχαίες μεταβλητές. Με την τεχνική αύξησης δεδομένων, εισάγουμε νέες τυχαίες μεταβλητές z στο μοντέλο, οι οποίες παριστάνουν τα δεδομένα που λείπουν αλλά αντιμετωπίζονται ως παράμετροι. Γενικά, σε ένα πρόβλημα ελλειπών δεδομένων το z περιλαμβάνει τις πρόσθετες μεταβλητές και μπορεί να είναι μία μόνο μεταβλητή ή διάνυσμα.

Στην πεπερασμένη μίξη, η από κοινού εκ των υστέρων κατανομή του $(\lambda_1, \lambda_2, z)$ είναι:

$$f(\lambda_1, \lambda_2, z|y) \propto f(y, z|\lambda_1) f(y, z|\lambda_2) f(\lambda_1) f(\lambda_2)$$

Εδώ το z είναι μία ακολουθία από k κεφαλές ή γράμματα, με ένα στοιχείο ανά παρατήρηση. Ως εκ τούτου, $z = \{z_1, z_2, \dots, z_k\}$ και $z_i = 1$ αν είναι η i ρίψη κεφάλι και $z_i = 2$ αν η i ρίψη είναι γράμματα.

Υποθέτουμε ότι η εκ των υστέρων πεποίθησή μας για τα $\lambda_i \sim Gamma(1,1)$

$$f(y_i, z_i|\lambda_1, \lambda_2) = f(y_i|z_i, \lambda) p(z_i)$$

όπου

$$f(y_i|z_i, \lambda_1, \lambda_2) = \begin{cases} \dfrac{e^{-\lambda_1} {\lambda_1}^{y_i}}{y_i!}, & z_i = 1 \quad [Poisson(\lambda_1)] \\ \dfrac{e^{-\lambda_2} {\lambda_2}^{y_i}}{y_i!}, & z_i = 2 \quad [Poisson(\lambda_2)] \end{cases}$$

$$p(z_i = 1|\lambda_1, \lambda_2, y) = \frac{\dfrac{e^{-\lambda_1} {\lambda_1}^{y_i}}{y_i!}}{\dfrac{e^{-\lambda_1} {\lambda_1}^{y_i}}{y_i!} + \dfrac{e^{-\lambda_2} {\lambda_2}^{y_i}}{y_i!}} = \frac{e^{-\lambda_1} {\lambda_1}^{y_i}}{e^{-\lambda_1} {\lambda_1}^{y_i} + e^{-\lambda_2} {\lambda_2}^{y_i}} = p_1$$

$$p(z_i = 2|\lambda_1, \lambda_2, y) = \frac{\dfrac{e^{-\lambda_2} {\lambda_2}^{y_i}}{y_i!}}{\dfrac{e^{-\lambda_1} {\lambda_1}^{y_i}}{y_i!} + \dfrac{e^{-\lambda_2} {\lambda_2}^{y_i}}{y_i!}} = \frac{e^{-\lambda_2} {\lambda_2}^{y_i}}{e^{-\lambda_1} {\lambda_1}^{y_i} + e^{-\lambda_2} {\lambda_2}^{y_i}} = p_2$$

Οπότε, λαμβάνουμε

$$f(\lambda_1, \lambda_2|y, z) = \frac{f(\lambda_1, \lambda_2, z|y)}{f(z)} = \frac{\left\{\prod_{i=1}^{k} (y_i|z_i, \lambda_1, \lambda_2) \cdot p(z_i)\right\} \cdot f(\lambda_1, \lambda_2)}{\prod_{i=1}^{k} p(z_i)}$$

$$= \frac{\prod_{i=1}^{k} p(z_i) \cdot \prod_{i:z_i=1} \frac{e^{-\lambda_1} {\lambda_1}^{y_i}}{y_i!} \cdot \prod_{i:z_i=2} \cdot \frac{e^{-\lambda_2} {\lambda_2}^{y_i}}{y_i!} \cdot f(\lambda_1, \lambda_2)}{\prod_{i=1}^{k} p(z_i)}$$

$$\propto e^{-\lambda_1 n_1} {\lambda_1}^{\sum_{i;z_i=1} y_i} \cdot e^{-\lambda_2 n_2} {\lambda_2}^{\sum_{i;z_i=2} y_i} \cdot e^{-(\lambda_1+\lambda_2)} {\lambda_1}^{y_i} {\lambda_2}^{y_i}$$

$$\propto e^{-\lambda_1 n_1} e^{-\lambda_2 n_2} e^{-(\lambda_1+\lambda_2)} {\lambda_1}^{y_i+\sum_{i;z_i=1} y_i} {\lambda_2}^{y_i+\sum_{i;z_i=2} y_i}$$

$$\propto e^{-\lambda_1 (n_1+1)} e^{-\lambda_2 (n_2+1)} {\lambda_1}^{y_i+\sum_{i;z_i=1} y_i} {\lambda_2}^{y_i+\sum_{i;z_i=2} y_i}$$

$$\text{ή } \equiv \frac{{\lambda_1}^{y_i}}{e^{(\lambda_1+\lambda_2)}} e^{-\lambda_1 n_1} {\lambda_1}^{\sum_{i;z_i=1} y_i} \times \frac{{\lambda_2}^{y_i}}{e^{(\lambda_1+\lambda_2)}} e^{-\lambda_2 n_2} {\lambda_2}^{\sum_{i;z_i=2} y_i}$$

$$\text{ή } \propto Gamma(n_1, \sum_{i;z_i=1} y_i + 1) \times Gamma(n_2, \sum_{i;z_i=2} y_i + 1)$$

2.6 Μπεϋζιανή επιλογή μοντέλου σε μίξη Poisson κατανομών

Την παράγραφο αυτή, θα την παρουσιάσουμε ως ένα παράδειγμα, για την καλύτερη κατανόηση της Μπεϋζιανής προσέγγισης μιας και περαιτέρω αναφορά πάνω στην επιλογή μοντέλου ξεφεύγει από την ανάλυσή μας.

Θεωρούμε *k* τυχαία ανεξάρτητα δείγματα

$$y_{ij} = \left(y_{i1}, y_{i2}, \dots, y_{in_j}\right)$$

μεγέθους n_j, για $j = (1,2, \dots, k)$ από κατανομή Poisson. Θέλουμε να ελέγξουμε την υπόθεση ότι

$$H_0: \lambda_1 = \cdots = \lambda_\kappa = \lambda \ \ vs \ \ H_1: \lambda_i \neq \lambda_j, \forall (i,j)$$

Τα αντίστοιχα μοντέλα που αντιπροσωπεύουν οι παραπάνω υποθέσεις μας είναι:

$$M_0: y_{ij} \sim Poisson(\lambda) \quad M_1: y_{ij} \sim Poisson\left(\lambda_j\right)$$

Για το μηδενικό μοντέλο M_0 ορίζουμε $\lambda \sim Gamma(p, q)$ εκ των προτέρων κατανομή για την κοινή παράμετρο λ. Η εκ των υστέρων που προκύπτει είναι $f(\lambda|\boldsymbol{y}) \sim Gamma(p + N\bar{y}, q + N)$, όπου $N = \sum_{j=1}^{k} n_j$ και $\bar{y}$ είναι ο μέσος όλων των δεδομένων και $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_k)$. Οπότε,

$$f(y|M_0) = \frac{\Gamma(p + N\bar{y}) q^p}{\Gamma(p)(q + N)^{p+N\bar{y}} \prod_{i=1}^{N} y_i!}$$

Για το εναλλακτικό μοντέλο M_1 ορίζουμε $\lambda_j = Gamma\left(p_j, q_j\right)$ εκ των προτέρων κατανομές για τις παραμέτρους λ_j, όπου $j = 1,2, \dots, k$. Οπότε

$$f(\boldsymbol{y}|M_1) = \int_0^\infty \dots \int_0^\infty f(\boldsymbol{y}|\lambda_1, \dots, \lambda_k, M_1) f(\lambda_1, \dots, \lambda_k)\, d\lambda_1 \dots d\lambda_k$$

$$= \int_0^\infty f(\boldsymbol{y}|\lambda_1) f(\lambda_1)\, d\lambda_1 \times \dots \times \int_0^\infty f(\boldsymbol{y}|\lambda_k) f(\lambda_k)\, d\lambda_k$$

$$= \prod_{j=1}^{k} \frac{\Gamma(p_j + n_j\bar{y}_j) {q_j}^{p_j}}{\Gamma(p_j)(q_j + n_j)^{p_j + n_j\bar{y}_j} \prod_{i=1}^{n_j} y_{ij}!}$$

Παράγοντας Bayes:

Στο σημείο αυτό υπολογίζουμε τον Bayes Factor για τη σύγκριση του μοντέλου M_1 έναντι του M_0.

$$BF_{10} = \frac{\Gamma(p)(q+N)^{p+N\bar{y}}}{\Gamma(p + N\overline{y})q^p} \prod_{j=1}^{k} \frac{\Gamma(p_j + n_j\bar{y}_j) {q_j}^{p_j}}{\Gamma(p_j)(q_j + n_j)^{p_j + n_j\bar{y}_j}}$$

Στο αμέσως επόμενο βήμα, θα επιλέξουμε την εκ των προτέρων κατανομή:

Για να εξάγουμε συμπεράσματα θα βασιστούμε στην εκ των προτέρων κατανομή με βάρος ίσο με ένα σημείο δεδομένων, $\text{Gamma}(\delta n\overline{y_0}, \delta n)$. Οπότε για $\delta = \frac{1}{N}$ οι κατανομές που θα πάρουμε θα είναι για τη μηδενική και εναλλακτική υπόθεση αντίστοιχα οι:

$$H_0: \lambda \sim Gamma(\bar{y}, 1) \quad vs \quad H_1: \lambda_j \sim Gamma\left(\frac{n_j}{N}\bar{y}, \frac{n_j}{N}\right)$$

Η παραπάνω σχέση μπορεί πιο γενικά να πάρει την μορφή

$$H_0: \lambda \sim Gamma(\alpha, 1) \quad vs \quad H_1: \lambda_j \sim Gamma(w_j\alpha_j, w_j)$$

Όπου

$$-\alpha = w_1\alpha_1 + \dots + w_k\alpha_\kappa \qquad -w_j = \frac{n_j}{N}$$

$$-N = n_1 + \dots + n_k \qquad -j = 1, \dots, k$$

Από την γενική μορφή της εκ των υστέρων κατανομής, διακρίνουμε τις εξής περιπτώσεις:

- Αν $\alpha_j = \alpha$ τότε $\lambda \sim Gamma(\alpha, 1)$ και $\lambda_j \sim Gamma\left(\frac{n_j}{N}\alpha, \frac{n_j}{N}\right)$, ενώ για α=1 προκύπτουν οι εκ των προτέρων κατανομές
$$\lambda \sim Gamma(1,1) \quad \& \quad \lambda_j \sim Gamma\left(\frac{n_j}{N}, \frac{n_j}{N}\right)$$

- Αν $\alpha_j = \alpha$ και επιπλέον $n_1 = \cdots = n_k$ τότε $\lambda \sim Gamma(\alpha, 1)$ και $\lambda_j \sim Gamma\left(\frac{1}{k}, \frac{1}{k}\right)$,
  ενώ για α=1 προκύπτουν οι εκ των προτέρων κατανομές
  $$\lambda \sim Gamma(1,1) \quad \& \quad \lambda_j \sim Gamma\left(\frac{1}{k}, \frac{1}{k}\right)$$
  με $E(\lambda) = Var(\lambda) = 1$ και $E(\lambda_j) = 1$ & $Var(\lambda_j) = k$.
  Για α=$\bar{y}$ προκύπτουν οι εκ των προτέρων κατανομές
  $$\lambda \sim Gamma(\bar{y}, 1) \quad \& \quad \lambda_j \sim Gamma\left(\frac{\bar{y}}{k}, \frac{1}{k}\right)$$
  με $E(\lambda) = Var(\lambda) = \bar{y}$ και $E(\lambda_j) = \bar{y}$ & $Var(\lambda_j) = k\bar{y}$ που είναι *Empirical Bayes Priors* διότι βασίζονται στα παρατηρούμενα δεδομένα.

Κεφάλαιο 3^{ο}

Κρυμμένα Μαρκοβιανά Μοντέλα (HMMs) – Hidden Markov Models

Σε αυτό το κεφάλαιο θα ασχοληθούμε με το είδος των στατιστικών μοντέλων που θα χρησιμοποιήσουμε για τη μοντελοποίηση του προβλήματός μας. Έτσι θα κάνουμε μία παρουσίαση των Κρυμμένων Μαρκοβιανών Μοντέλων και τους λόγους που μας οδήγησαν στο να καταλήξουμε σε αυτό το είδος μοντέλων.

Η μοντελοποίηση με βάση τα Κρυμμένα Μαρκοβιανά Μοντέλα (HMMs) είναι μία ισχυρή στατιστική τεχνική, κατάλληλη για τη μοντελοποίηση ακολουθιακών δεδομένων (sequential data), η οποία έχει αντικρύσει μεγάλο εύρος εφαρμογής στις περιοχές του γραπτού και προφορικού λόγου, καθώς και σε εφαρμογές μηχανικής μετάφρασης. Αφού παρουσιάσουμε τα Κρυμμένα Μαρκοβιανά Μοντέλα διακριτού χρόνου και πεπερασμένου χώρου καταστάσεων, στη συνέχεια παραθέτουμε και ένα απλό παράδειγμα που θα μας βοηθήσει να κατανοήσουμε καλύτερα την τεχνική αυτή.

3.1 Κρυμμένα Μαρκοβιανά Μοντέλα

3.1.1 Μαρκοβιανή Αλυσίδα

Μία στοχαστική διαδικασία $\{X_t\}$, *t=1,2,…,* με χώρο καταστάσεων S αριθμήσιμο σύνολο (άπειρο ή πεπερασμένο), λέγεται Μαρκοβιανή αλυσίδα διακριτού χρόνου αν έχει τη Μαρκοβιανή ιδιότητα, δηλαδή αν

$$Pr(X_{t+1} = j|X_1 = i_1, X_2 = i_2, \dots, X_t = i) = Pr(X_{t+1} = j|X_t = j).$$

Η πιθανότητα η X_{t+s} να πάρει την τιμή j δεδομένου ότι η X_t έχει την τιμή i, γράφεται

$$p_{ij}(t, t+s) = Pr(X_{t+s} = j|X_t = i)$$

και ονομάζεται πιθανότητα μετάβασης *s* τάξης. Αν η πιθανότητα αυτή δεν εξαρτάται από το t, αλλά μόνο από τα βήματα *s,* τότε η αλυσίδα λέγεται ομογενής και συμβολίζουμε την πιθανότητα μετάβασης *s* τάξης απλά

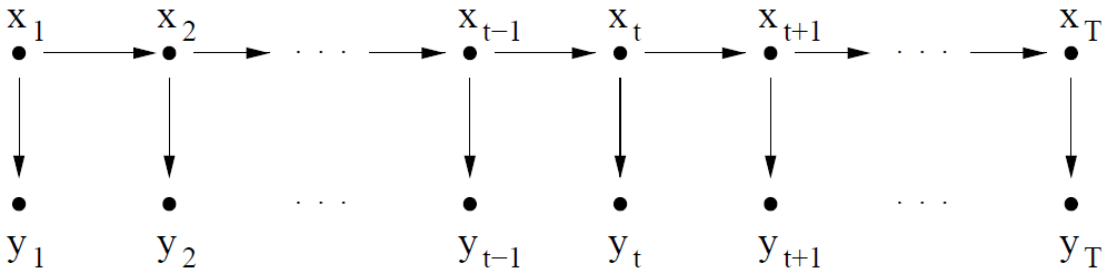
$$p_{ij}{}^{(s)} = Pr(X_{t+s} = j|X_t = i).$$

3.1.2 Κρυμμένα Μαρκοβιανά Μοντέλα Δακριτού χρόνου και Πεπερασμένου χώρου καταστάσεων

Έστω λοιπόν $\{X_t\}$, *t=1,2,…* μία ομογενής Μαρκοβιανή αλυσίδα διακριτού χρόνου, πεπερασμένου χώρου καταστάσεων, $S = \{1,2,\dots,N\}$, με πίνακα πιθανοτήτων μετάβασης $1^{ης}$ τάξης $P = (p_{ij})_{i,j\in S}$, όπου $p_{ij} = p_{ij}^{(1)}$ και στάσιμη κατανομή $\pi = (\pi_1, \pi_2, \dots, \pi_N)$. Έστω ακόμα μία στοχαστική διαδικασία $\{Y_t\}$ διακριτού χρόνου, δηλαδή *t=1,2,…* που εξαρτάται από την παραπάνω Μαρκοβιανή αλυσίδα $\{X_t\}$ με κάποιο μηχανισμό, ο οποίος θα μπορούσε να γραφεί ως

$Y_t \,|\, X_t = g(X_t) + \varepsilon_t$ για *t=1,2,…*

Οι δύο παραπάνω στοχαστικές διαδικασίες «δουλεύουν» ταυτόχρονα. Είναι δηλαδή μία διπλά στοχαστική διαδικασία, η πραγματοποίηση της οποίας σχηματικά, παρουσιάζεται ως εξής:



Σχήμα: Γραφική απεικόνιση ενός κρυμμένου Μαρκοβιανού μοντέλου διακριτού χρόνου πεπερασμένου χρόνου καταστάσεων

Το σύστημα αυτό, που περιλαμβάνει μία τέτοια στοχαστική διαδικασία $\{X_t\}$, που δεν είναι παρατηρήσιμη (είναι κρυφή), αλλά παράγει μία παρατηρήσιμη ακολουθία εξόδων $\{Y_t\}$, λέγεται κρυμμένο Μαρκοβιανό μοντέλο, διακριτού χρόνου πεπερασμένου χώρου καταστάσεων (Discrete-Time State-Space Hidden Markov Model). Έτσι, ενώ σε ένα απλό Μαρκοβιανό μοντέλο άγνωστες είναι οι πιθανότητες μετάβασης, σε ένα κρυμμένο Μαρκοβιανό μοντέλο είναι άγνωστες και οι καταστάσεις.

Δηλαδή, έχουμε παρατηρήσεις (γνωστές σε μας) οι οποίες παράγονται σε διακριτό χρόνο και εξαρτώνται από τις καταστάσεις (άγνωστες σε μας) μιας Μαρκοβιανής αλυσίδας.

Παράδειγμα 3.1

Έστω *N* δοχεία, Δ_1, Δ_2, … Δ_N. Το κάθε ένα είναι γεμάτο με έγχρωμες σφαίρες *M* διαφορετικών χρωμάτων (το κάθε δοχείο έχει πάνω από *M* σφαίρες). Διαλέγουμε ένα δοχείο, σύμφωνα με μία αρχική κατανομή και κατόπιν διαλέγουμε τυχαία μία σφαίρα. Επαναλαμβάνουμε τη διαδικασία *T* φορές. Τότε θα έχουμε μία σειρά παρατηρήσιμων τυχαίων μεταβλητών $Y_1, Y_2, \dots Y_T$, όπου Y_l το χρώμα της σφαίρας στην *i* δοκιμή.

Το μοντέλο αυτό, αποτελεί Κρυμμένο Μαρκοβιανό Μοντέλο αφού:

- «τρέχουν» δύο στοχαστικές διαδικασίες ταυτόχρονα, η πρώτη είναι η επιλογή δοχείου και η δεύτερη είναι η επιλογή σφαίρας από το επιλεγμένο δοχείο
- η δεύτερη εξαρτάται από την πρώτη και
- η δεύτερη είναι παρατηρήσιμη ενώ η πρώτη «κρυφή».

3.2 Διάκριση Κρυμμένων Μαρκοβιανών Μοντέλων

(ως προς την παρατηρούμενη διαδικασία)

Στο παράδειγμα 2.1 η κατανομή της Y_t είναι διακριτή, ενώ είναι δυνατόν η κατανομή της Y_t να είναι συνεχής (που όμως με την περίπτωση αυτή δε θα ασχοληθούμε). Με βάση λοιπόν τη μορφή της παρατηρούμενης διαδικασίας Y_t, έχουμε τη διακριτή περίπτωση, τη συνεχή περίπτωση ή ακόμα και συνδυασμό αυτών των δύο.

3.2.1 Διακριτά Κρυμμένα Μαρκοβιανά Μοντέλα

Στην περίπτωση αυτή τα κύρια στοιχεία που χαρακτηρίζουν ένα κρυμμένο Μαρκοβιανό μοντέλο είναι:

i. Ο αριθμός *N* των καταστάσεων. Θα συμβολίζουμε με $S = \{1,2, \dots, N\}$ το χώρο καταστάσεων και με X_t την κατάσταση σε χρόνο *t*.

ii. Ο αριθμός *M* των τιμών που παίρνει η διακριτή τυχαία μεταβλητή Y_t. Ορίζουμε το σύνολο των τιμών αυτών $V = \{v_1, v_2, \dots, v_M\}$.

iii. Ο πίνακας πιθανοτήτων μετάβασης $1^{ης}$ τάξης $P = \left(p_{ij}\right)_{i,j \in S}$.

iv. Οι κατανομές πιθανότητας της Y_t σε κάθε κατάσταση, δηλαδή οι $b_j(k) = Pr(Y_t = v_k | X_t = j)$ για *k=1,2,…M και j=1,2,…N*. Ορίζουμε το σύνολό τους $B = \left\{b_j(k)\right\}_{j=1}^{N}$.

v. Η στάσιμη κατανομή $\pi = (\pi_1, \pi_2, \dots, \pi_N)$, όπου $\pi_j = Pr[X_t = j]$ για *j=1,2,…N.*

Δίνοντας τώρα, κάποιες τιμές στα χαρακτηριστικά που μόλις αναφέραμε, το μοντέλο μας θα δίνει μία ακολουθία εξόδων

$$Y_1, Y_2, \dots, Y_T$$

Όπου κάθε παρατήρηση Y_t, παίρνει κάποια από τις τιμές του *V* και *T* το πλήθος των παρατηρήσεων. Επομένως, συγκρίνοντας και με τα παραπάνω, καταλήγουμε ότι ο πλήρης προσδιορισμός ενός Κρυμμένου Μαρκοβιανού Μοντέλου απαιτεί τον προσδιορισμό:

- των *N και M*,
- των τιμών v_k και
- των τριών στοχαστικών μεγεθών *P, B* και *π*, τα οποία συμβολίζονται με *λ=(P, B, π)*

Στο σημείο αυτό για την καλύτερη κατανόηση των παραπάνω, ακολουθεί η συνέχεια του παραδείγματος 2.1.

Παράδειγμα 3.1 *(συνέχεια)*

Χρησιμοποιώντας τον ίδιο συμβολισμό με τα παραπάνω, λαμβάνουμε:

$$\begin{matrix} \Delta_1 & & \Delta_N \\ Pr[\sigma\varphi\alpha\acute{\iota}\rho\alpha\ \chi\rho\acute{\omega}\mu\alpha\tau o\varsigma\ 1] = b_1(1) & & Pr[\sigma\varphi\alpha\acute{\iota}\rho\alpha\ \chi\rho\acute{\omega}\mu\alpha\tau o\varsigma\ 1] = b_N(1) \\ \dots & \dots, & \dots \\ Pr[\sigma\varphi\alpha\acute{\iota}\rho\alpha\ \chi\rho\acute{\omega}\mu\alpha\tau o\varsigma\ M] = b_1(M) & & Pr[\sigma\varphi\alpha\acute{\iota}\rho\alpha\ \chi\rho\acute{\omega}\mu\alpha\tau o\varsigma\ M] = b_N(M) \end{matrix}$$

με

- Μαρκοβιανή αλυσίδα $\{X_t\}$ t=1,2,…,T, χώρο καταστάσεων $S = \{1,2, \dots, N\}$, όπου *i* είναι το Δ_i δοχείο, *i=1,2,…,N,* πίνακα πιθανοτήτων μετάβασης $P = (p_{ij})_{i,j \in S}$ με $p_{ij} = Pr(X_{t+1} = j | X_t = i)$ για κάθε i,*j*∈*S*και στάσιμη κατανομή $\pi = (\pi_1, \pi_2, \dots, \pi_N)$.
- Πιθανότητα επιλογής σφαίρας δεδομένης της επιλογής δοχείου $Pr[\epsilon\pi\iota\lambda o\gamma\acute{\eta}\varsigma\ \sigma\varphi\alpha\acute{\iota}\rho\alpha\varsigma\ \chi\rho\acute{\omega}\mu\alpha\tau o\varsigma\ k \mid \acute{\epsilon}\chi\omega\ \epsilon\pi\iota\lambda\acute{\epsilon}\xi\epsilon\iota\ \delta o\chi\epsilon\acute{\iota}o\ i] = b_1(k)$ για *i=1,2,..,N και k=1,2,…,M*.

Έτσι σύμφωνα με την αρχική κατανομή π, διαλέγουμε δοχείο, έστω το 4, με πιθανότητα π_4. Κατόπιν διαλέγουμε σφαίρα, έστω χρώματος 2, με πιθανότητα $b_4(2)$ και το καταγράφουμε. Έπειτα διαλέγουμε πάλι δοχείο, έστω το 7, με πιθανότητα p_{47} και από σφαίρα χρώματος 3 με πιθανότητα $b_7(3)$ κ.ο.κ.

3.2.2 Συνεχή Κρυμμένα Μαρκοβιανά Μοντέλα

Στην περίπτωση αυτή τα κύρια στοιχεία ενός Κρυμμένου Μαρκοβιανού Μοντέλου είναι:

- ✓ Ο αριθμός *N* των καταστάσεων. Θα συμβολίζουμε με *S* το χώρο καταστάσεων και με X_t την κατάσταση σε χρόνο *t*.
- ✓ Ο πίνακας πιθανοτήτων μετάβασης 1ης τάξης $P = \left(p_{ij}\right)_{i,j\in S}$
- ✓ Η κατανομή πυκνότητας πιθανότητας των παρατηρούμενων τυχαίων μεταβλητών Y_t στην κατάσταση j, δηλαδή η
$$f(y_t|X_t = j),\ \ j = 1,2,\dots,N$$
- ✓ Η στάσιμη κατανομή $\pi = (\pi_1, \pi_2, \dots, \pi_N)$, όπου $\pi_j = Pr(X_t = j)$ για *j=1,2,…,N*

Παράδειγμα 3.2 *(μίξη Poisson κατανομών)*

Έστω η Μαρκοβιανή αλυσίδα $\{X_\tau\}$ για *t=1,2,…,T* με χώρο καταστάσεων $S = \{1,2,\dots,N\}$, στάσιμη κατανομή $\pi = (\pi_1, \pi_2, \dots, \pi_N)$ και πίνακα πιθανοτήτων μετάβασης 1ης τάξης $P = \left(p_{ij}\right)_{i,j\in S}$ με

$$P = \begin{bmatrix} p_1 & p_2 & \dots & p_N \\ p_1 & p_2 & & p_N \\ \vdots & \vdots & \dots & \vdots \\ p_1 & p_2 & \dots & p_N \end{bmatrix}.$$

Αν η $Y_t \,|\, X_t = j$ ακολουθεί την $Poisson(\lambda_j)$, κάτι που συμβαίνει με πιθανότητα p_j, για $t = 1,2,\dots,T$ και $j = 1,2,\dots,N$ τότε έχουμε ένα Κρυμμένο Μαρκοβιανό μοντέλο. Παρατηρώντας τη δομή του πίνακα *P* (όλες οι γραμμές είναι ίδιες, δηλαδή η πιθανότητα μετάβασης σε μία κατάσταση είναι η ίδια, από όποια κατάσταση κι αν εκκινήσουμε) αντιλαμβανόμαστε ότι το μοντέλο αυτό είναι η περίπτωση της μίξης Poisson κατανομών. Έτσι είναι πλέον φανερό ότι τα Κρυμμένα Μαρκοβιανά Μοντέλα μπορούν να θεωρηθούν και ως μία γενίκευση των μοντέλων πεπερασμένων μίξεων κατανομών.

3.3 Τα Τρία Βασικά προβλήματα

Προτού περάσουμε σε εφαρμογές, θα συζητήσουμε τρία βασικά προβλήματα που προκύπτουν κατά τη διαδικασία ανάπτυξης μεθοδολογίας στατιστικής συμπερασματολογίας για Κρυμμένα Μαρκοβιανά Μοντέλα.

✓ **Πρόβλημα 1^{ο}** *(Πρόβλημα αποτίμησης)*

Σε αυτό το πρόβλημα, εκείνο που θα πρέπει να εξετάσουμε και εντέλει να αντιμετωπίσουμε, είναι ο υπολογισμός της πιθανότητας μιας συγκεκριμένης ακολουθίας παρατηρήσεων, έστω $Y = (Y_1, Y_2, \dots, Y_T)$, να προέρχεται από ένα συγκεκριμένο Κρυμμένο Μαρκοβιανό Μοντέλο, έστω αυτό που παράγεται από τη διανυσματική παράμετρο ϑ, δηλαδή της πιθανοφάνειας $f(y|\theta)$.

Εξηγώντας καλύτερα, η πιθανότητα αυτή μας δίνει ένα μέτρο του πόσο καλά ταιριάζει το μοντέλο με τις παρατηρήσεις *Y*. Κυρίως χρησιμεύει για την περίπτωση που έχουμε πολλά μοντέλα υποψήφια, αφού μας δείχνει ποιο ταιριάζει περισσότερο.

✓ **Πρόβλημα 2^{ο}** *(Πρόβλημα αποκρυπτογράφησης)*

Εδώ στην ουσία, προσπαθούμε να ανακαλύψουμε το «κρυφό» κομμάτι του μοντέλου, δηλαδή το χώρο καταστάσεων της Μαρκοβιανής Αλυσίδας, με δεδομένη την ύπαρξη μιας συγκεκριμένης ακολουθίας παρατηρήσεων $Y = (Y_1, Y_2, \dots, Y_T)$. Όπως είναι γνωστό σε κάθε πρόβλημα στατιστικού περιεχομένου, δεν υπάρχει 'σωστή' επιλογή, ούτε και 'μοναδική', ως απάντηση στο πρόβλημα αυτό, εκτός από κάποιες εκφυλισμένες περιπτώσεις. Για την εκλογή της βέλτιστης ακολουθίας καταστάσεων, υπάρχουν αρκετά κριτήρια που χρησιμοποιούμε ανάλογα με την περίπτωση και τους σκοπούς μας.

✓ **Πρόβλημα 3^{ο}** *(Πρόβλημα εκτίμησης)*

Σκοπός μας στην περίπτωση αυτή, είναι να προσδιορίσουμε τις βέλτιστες παραμέτρους του μοντέλου μας, έτσι ώστε με δεδομένη την ακολουθία του παρατηρήσεων *Y*, να μεγιστοποιείται η πιθανοφάνεια αν δουλεύουμε με την *Κλασσική Στατιστική* ή να προσομοιώσουμε αν δουλεύουμε με την *Μπεϋζιανή Στατιστική*. Επιλέγοντας λοιπόν το μοντέλο μας, προσπαθούμε να βελτιστοποιήσουμε τις παραμέτρους του, ώστε να ανταποκρίνονται όσο το δυνατόν καλύτερα στον τρόπο με τον οποίο προέκυψε η ακολουθία *Y*. Συγκεκριμένα, δεν κάνουμε τίποτα περισσότερο από το προσαρμόζουμε τις παραμέτρους του μοντέλου που έχουμε ήδη επιλέξει, ώστε να μεγιστοποιείται η $Pr(Y|\theta)$. Είναι το πιο σημαντικό κομμάτι για σχεδόν όλες τις εφαρμογές, καθότι επιτρέπει να προσαρμόζουμε τις παραμέτρους κατά τέτοιο τρόπο, ώστε να περιγράφει το μοντέλο μας όσο το δυνατόν καλύτερα το φαινόμενο που καλείται να ερμηνεύσει.

3.4 Μέθοδοι Εκτίμησης

3.4.1 Ο Αλγόριθμος «Εμπρός-Πίσω» για τα HMMs Πεπερασμένου Χώρου καταστάσεων

Ο αλγόριθμος «εμπρός-πίσω» (*The Forward – Backward Algorithm, Baum 1972*) είναι μια διπλή ακολουθιακή διαδικασία. Βασίζεται στις αρχές του δυναμικού προγραμματισμού και αποτελείται από δύο κύρια μέρη. Την προς τα εμπρός διαδικασία και την προς τα πίσω διαδικασία. Οι δύο αυτές διαδικασίες θα μας δώσουν τις προς τα εμπρός και τις προς τα πίσω μεταβλητές αντίστοιχα, οι οποίες θα χρησιμοποιηθούν για τη λύση των τριών βασικών προβλημάτων.

3.4.2 Η προς τα Εμπρός Διαδικασία

Ορίζουμε την προς τα εμπρός μεταβλητή $\alpha_t(i)$ ως εξής

$$a_t(i) = Pr(y_1, y_2, \dots, y_t, X_t = i|\theta) = Pr(X_t = i|\theta) \cdot Pr(y_1, y_2, \dots, y_t|X_t = i, \theta),$$

για $t = 1,2, \dots, T$ και $i = 1,2, \dots, N$.

Είναι προφανές ότι η αρχική προς τα εμπρός μεταβλητή θα είναι

$$a_1(i) = \pi_i \cdot f_i(y_1) \text{ για } i = 1,2, \dots, N$$

ενώ το επαγωγικό βήμα θα δίνεται από

$$a_{t+1}(j) = Pr(y_1, y_2, \dots, y_t, y_{t+1}, X_{t+1} = j|\theta) =$$

$$\sum_{i=1}^{N} Pr\,(y_1, y_2, \dots, y_t, y_{t+1}, X_t = i, X_{t+1} = j|\theta) =$$

$$\sum_{i=1}^{N} Pr\,(y_1, y_2, \dots, y_t, X_t = i, y_{t+1}, X_{t+1} = j|\theta) =$$

$$\sum_{i=1}^{N} Pr(y_1, y_2, \dots, y_t, X_t = i|\theta) \cdot Pr(y_{t+1}, X_{t+1} = j|\theta, y_1, y_2, \dots, y_t, X_t = i) =$$

$$\sum_{i=1}^{N} Pr(y_1, \dots, y_t, X_t = i|\theta) \cdot Pr(X_{t+1} = j|\theta, y_1, \dots, y_t, X_t = i)$$

$$\cdot Pr(y_{t+1}|\theta, y_1, \dots, y_t, X_t = i, X_{t+1} = j) =$$

$$\left[\sum_{i=1}^{N} a_t(i) \cdot Pr(X_{t+1} = j|\theta, X_t = i)\right] \cdot f_j(y_{t+1}),$$

δηλαδή τελικά

$a_{t+1}(j) = \left[\sum_{i=1}^{N} a_t(i) \cdot p_{ij}\right] \cdot f_j(y_{t+1})$ για $j = 1,2,\dots,N$ και $t = 1,2,\dots,T-1$

Το τελευταίο βήμα της διαδικασίας, για *t=T* δηλαδή, θα δώσει τις πιθανότητες

$a_T(i) = Pr(y_1, y_2, \dots, y_T, X_T = i|\theta)$ για $i = 1,2,\dots,N$

Οι οποίες αθροιζόμενες για όλες τις καταστάσεις *i*, θα μας δώσουν την πιθανοφάνεια $f(y|\theta)$ για δεδομένη τιμή του θ.
Έτσι,

$$f(y|\theta) = Pr(y_1, y_2, \dots, y_T|\theta) = \sum_{i=1}^{N} a_T(i),$$

Το υπολογιστικό κόστος είναι της τάξεως του $N^2 \cdot T$ (δηλαδή $N \cdot (N+1) \cdot (T-1) + N$ πολλαπλασιασμοί και $N \cdot (N-1) \cdot (T-1)$ προσθέσεις) και όχι της τάξεως του $2 \cdot T \cdot N^T$ που είχαμε με την προηγούμενη μέθοδο της εκτίμησης μέγιστης πιθανοφάνειας. Για να καταλάβουμε καλύτερα τη διαφορά, οι πράξεις που θα χρειαστούμε για το ίδιο μοντέλο που είχαμε ως παράδειγμα και πριν (με $N = 5$ και $T = 100$ δηλαδή), είναι τώρα περίπου 3000, ενώ πριν ήταν περίπου 10^{72}.

Συνοψίζοντας, η προς τα εμπρός διαδικασία μας δίνει την πιθανοφάνεια των παρατηρήσεων με τον εξής αλγόριθμο,

1. Έναρξη
   $a_1(i) = \pi_i \cdot f_i(y_1)$ για $i = 1,2,\dots,N$
2. Επαναληπτικό βήμα
   $a_{t+1}(j) = \left[\sum_{i=1}^{N} a_t(i) \cdot p_{ij}\right] \cdot f_j(y_{t+1})$ για $j = 1,2,\dots,N$ και $t = 1,2,\dots,T-1$
3. Τερματισμός
   $f(y|\theta) = \sum_{i=1}^{N} a_T(i)$

3.4.3 Η προς τα Πίσω Διαδικασία

Αντίστοιχα με τον ορισμό των προς τα εμπρός μεταβλητών, έχουμε και τον ορισμό των προς τα πίσω μεταβλητών ως εξής:

$$b_t(i) = Pr(y_{t+1}, y_{t+2}, \dots, y_T|X_t = i, \theta)$$

για $t = T-1, T-2, \dots, 1$ και $i = 1,2, \dots, N$, ενώ για $t = T$ ορίζουμε $b_T(i) = 1$.

Το $b_t(i)$ εκφράζει την πιθανότητα να πραγματοποιηθεί το μέρος των παρατηρήσεων από $t+1$ έως T, δεδομένου ότι η κατάσταση στο χρόνο t είναι η i και το μοντέλο έχει παράμετρο θ. Ο υπολογισμός θα γίνει πάλι με τον ίδιο τρόπο, πηγαίνοντας πίσω αυτή τη φορά και το επαγωγικό βήμα υπολογίζεται ως εξής:

$$b_{t-1}(j) =$$

$$= Pr(y_1, y_{t+1}, \dots, y_T | X_{t-1} = j, \theta)$$

$$= \sum_{i=1}^{N} Pr(y_t, y_{t+1}, \dots, y_T, X_t = i | X_{t-1} = j, \theta)$$

$$= \sum_{i=1}^{N} Pr(y_t | X_{t-1} = j, X_t = i, y_{t+1}, \dots, y_T, \theta) \cdot Pr(y_{t+1}, \dots, y_T, X_t = i | X_{t-1} = j, \theta)$$

$$= \sum_{i=1}^{N} f_i(y_t) \cdot Pr(X_t = i | X_{t-1} = j, \theta) \cdot Pr(y_{t+1}, \dots, y_T | X_{t-1} = j, X_t = i, \theta)$$

$$= \sum_{i=1}^{N} f_i(y_t) \cdot p_{ji} \cdot b_t(i)$$

Όπως παρατηρούμε, ο υπολογισμός του $b_t(i)$ για $t = 1,2, \dots, T$ και $i = 1,2, \dots, N$ είναι και πάλι $N^2 \cdot T$ τάξεως υπολογισμών, δηλαδή το ίδιο «οικονομικός».

Συνοψίζοντας η προς τα πίσω διαδικασία περιλαμβάνει τα εξής:

1. Έναρξη
   $b_T(i) = 1$ για $i = 1,2, \dots, N$
2. Επαναληπτικό βήμα
   $b_{t-1}(j) = \sum_{i=1}^{N} b_t(i) \cdot p_{ji} \cdot f_i(y_t)$ για $j = 1,2, \dots, N$ και $t = T, T-1, \dots, 1$.

3.4.4 Εφαρμογή του αλγορίθμου στα Προβλήματα 1 και 2

Αφού υπολογίσουμε τις μεταβλητές $a_t(i)$ και $b_t(i)$ του αλγορίθμου, θα τις χρησιμοποιήσουμε για να λύσουμε τα βασικά μας προβλήματα. Έχει ήδη αναφερθεί ότι μία λύση του προβλήματος 1, δίνεται από τον τύπο

$$f(y|\theta) = \sum_{i=1}^{N} a_T(i)$$

Για το Πρόβλημα 2, θα προσπαθήσουμε να βρούμε την πιο πιθανή ακολουθία καταστάσεων, δεδομένης της ύπαρξης κάποιων παρατηρήσεων. Ακολούθως,

$$Pr(x_1, x_2, \dots, x_T | y_1, y_2, \dots, y_T, \theta) = Pr(x_T | y_1, y_2, \dots, y_T, \theta) \times$$
$$Pr(x_{T-1} | X_T, y_1, y_2, \dots, y_T, \theta) \times$$
$$Pr(x_{T-2} | X_{T-1}, X_T, y_1, y_2, \dots, y_T, \theta) \times$$
$$\dots$$
$$Pr(x_1 | x_2, \dots, x_{T-1}, x_T, y_1, y_2, \dots, y_T, \theta)$$

Όμως

$$Pr(x_t | x_t, x_{t+1}, x_{t+2}, \dots, x_T) = Pr(x_t | x_{t+1})$$

γιατί

$$Pr(x_t | x_{t+1}, x_{t+2}, \dots, x_T) = Pr(x_t | x_{t+1}) \Leftrightarrow$$
$$\Leftrightarrow \frac{Pr(x_t, x_{t+1}, \dots, x_T)}{Pr(x_{t+1}, \dots, x_T)} = \frac{Pr(x_t, x_{t+1})}{Pr(x_{t+1})}$$
$$\Leftrightarrow \frac{Pr(x_t, x_{t+1}, \dots, x_T)}{Pr(x_t, x_{t+1})} = \frac{Pr(x_{t+1}, \dots, x_T)}{Pr(x_{t+1})}$$
$$\Leftrightarrow Pr(x_{t+2}, \dots, x_T | x_t, x_{t+1}) = Pr(x_{t+2}, \dots, x_T | x_{t+1})$$

ισότητα η οποία ισχύει, λόγω της Μαρκοβιανής ιδιότητας.

Έτσι,

$$Pr(x_1, x_2, \dots, x_T | y_1, y_2, \dots, y_T, \theta) =$$
$$= Pr(x_T | y_1, y_2, \dots, y_T, \theta) \times$$
$$\dots$$
$$Pr(x_t | x_{t+1}, y_1, y_2, \dots, y_T, \theta) \times$$
$$\dots$$
$$Pr(x_1 | x_2, y_1, y_2, \dots, y_T, \theta).$$

Η κατάσταση σε χρόνο t δεδομένου του μοντέλου και της σειράς των παρατηρήσεων, έχει κατανομή

$$\gamma_t(i) = Pr(X_t = i|y_1, y_2, \dots, y_T, \theta) = \frac{Pr(X_t = i, y_1, y_2, \dots, y_T|\theta)}{Pr(y_1, y_2, \dots, y_T|\theta)}$$
$$= \frac{a_t(i) \cdot b_t(i)}{Pr(y_1, y_2, \dots, y_T|\theta)}$$

για $t = 1,2, \dots, T$ και $i = 1,2, \dots, N$

Έτσι για $t = T$ θα έχουμε

$$\gamma_t(i) = Pr(X_t = i|y_1, y_2, \dots, y_T, \theta) = \frac{a_T(i)}{\sum_{j=1}^{N} a_T(j)}$$

Η δεσμευμένη κατανομή της X_t δεδομένης της τιμής x_{t+1} της κατάστασης X_{t+1} και των παρατηρήσεων y, χρησιμοποιώντας το θεώρημα του Bayes είναι,

$$Pr(X_t = i|X_{t+1} = x_{t+1}, y_1, y_2, \dots, y_T, \theta) =$$

$$\frac{Pr(X_t = i|\theta) \cdot Pr(X_{t+1} = x_{t+1}, y_1, y_2, \dots, y_T|X_t = i, \theta)}{\sum_{j=1}^{N} Pr(X_t = j|\theta) \cdot Pr(X_{t+1} = x_{t+1}, y_1, y_2, \dots, y_T|X_t = j, \theta)} =$$

$$\frac{Pr(X_t = i|\theta) \cdot Pr(X_{t+1} = x_{t+1}, y_1, y_2, \dots, y_t, y_{t+1}, \dots, y_T|X_t = i, \theta)}{\sum_{j=1}^{N} Pr(X_t = j|\theta) \cdot Pr(X_{t+1} = x_{t+1}, y_1, y_2, \dots, y_t, y_{t+1}, \dots, y_T|X_t = j, \theta)} =$$

$$\frac{Pr(X_t = i|\theta) \cdot Pr(X_{t+1} = x_{t+1}, y_1, y_2, \dots, y_t|X_t = i, \theta) \cdot Pr(y_{t+1}, \dots, y_T|X_t = i, \theta)}{\sum_{j=1}^{N} Pr(X_t = j|\theta) \cdot Pr(X_{t+1} = x_{t+1}, y_1, y_2, \dots, y_t|X_t = j, \theta) \cdot Pr(y_{t+1}, \dots, y_T|X_t = j, \theta)} =$$

$$\frac{Pr(X_t = i|\theta) \cdot Pr(X_{t+1} = x_{t+1}, y_1, y_2, \dots, y_t|X_t = i, \theta)}{\sum_{j=1}^{N} Pr(X_t = j|\theta) \cdot Pr(X_{t+1} = x_{t+1}, y_1, y_2, \dots, y_t|X_t = j, \theta)} =$$

$$\frac{Pr(X_t = i|\theta) \cdot Pr(y_1, y_2, \dots, y_t|X_t = i, \theta) \cdot Pr(X_{t+1} = x_{t+1}|X_t = i, \theta)}{\sum_{j=1}^{N} Pr(X_t = j|\theta) \cdot Pr(y_1, y_2, \dots, y_t|X_t = j, \theta) \cdot Pr(X_{t+1} = x_{t+1}|X_t = j, \theta)} =$$

$\frac{a_t(i) \cdot p_{i,x_{t+1}}}{\sum_{j=1}^{N} a_t(j) \cdot p_{j,x_{t+1}}}$ για $i = 1,2, \dots, N$.

3.5 Εκτιμήσεις Μέγιστης Πιθανοφάνειας με τον Αλγόριθμο ΕΜ στην Κλασσική Στατιστική

Προκειμένου να χρησιμοποιηθεί ο αλγόριθμος ΕΜ για την εύρεση του εκτιμητή μέγιστης πιθανοφάνειας και ακόμα και αν δεν έχουμε ελλιπή δεδομένα, είναι σύνηθες να εκφράζουμε το πρόβλημα ως πρόβλημα ελλειπών δεδομένων.

Να αναφέρουμε επιγραμματικά ότι ένα πρόβλημα ελλιπών δεδομένων μπορεί να περιγραφεί ως εξής: έστω ένα στατιστικό μοντέλο καθορισμένο για το x με πυκνότητα $f(x|\theta)$, όπου θ οι παράμετροι που μας ενδιαφέρουν. Αν $h(\cdot)$ είναι μία

συνάρτηση που δεν είναι $1-1$, τότε στο πρόβλημα μας μπορούμε μόνο να παρατηρήσουμε το $y = h(x)$ και επομένως έχουμε ελλιπή πληροφόρηση σχετικά με το x. Η πυκνότητα των παρατηρήσιμων δεδομένων θα συμβολίζεται με $f(y|\theta)$.

Μπορεί να αποδειχθεί ότι, κάτω από το συγκεκριμένο πλαίσιο των ελλιπών δεδομένων, το μικτό μοντέλο προκύπτει επειδή τα z_i είναι άγνωστα και πρέπει να εκτιμήσουμε την κατανομή μίξης με τα δεδομένα που έχουμε στη διάθεσή μας από την περιθώρια κατανομή του y μόνο, παρά από την από κοινού κατανομή των (y, z). Ο αλγόριθμος ΕΜ παίρνει την από κοινού αυτή κατανομή ως ευκολία προκειμένου να υπολογίσει τους ΕΜΠ υπό την βάση των παρατηρούμενων δεδομένων y.

Αν είχαμε στη διάθεσή μας όλα τα δεδομένα (και τα μη παρατηρήσιμα z), τότε η εκτίμηση της κατανομής του δείγματος θα γινόταν με άμεσο τρόπο. Που όμως έχει και μεγάλο υπολογιστικό φόρτο.

3.5.1 Ο Αλγόριθμος ΕΜ

Ο αλγόριθμος ΕΜ είναι ένα πολύ ισχυρό εργαλείο για την εκτίμηση της ΜΠ με ευρεία χρήση. Έχει άπειρες εφαρμογές. Τα αρχικά ΕΜ προέρχονται από τις λέξεις *estimation* (αναμενόμενη τιμή) και *maximization* (μεγιστοποίηση) οι οποίες αποτελούν τα κύρια βήματα του επαναληπτικού αυτού αλγορίθμου.

Ο αλγόριθμος αυτός εγγυάται την αύξηση της πιθανοφάνειας σε κάθε επανάληψη και άρα συγκλίνει σε ένα τοπικό μέγιστο της συνάρτησης πιθανοφάνειας κάτω από πολύ ήπιες συνθήκες. Επιπλέον, σε πολλές περιπτώσεις είναι πολύ εύκολο να βελτιωθεί ο αλγόριθμος ΕΜ. Αυτά τα δύο χαρακτηριστικά συχνά κάνουν τον αλγόριθμο ΕΜ μια πολύ ελκυστική επιλογή σε σχέση με άλλες μεθόδους βελτιστοποίησης.

Η εμφάνιση του αλγορίθμου ΕΜ διευκόλυνε τους απαιτούμενους υπολογισμούς, ενώ πριν από αυτόν, απαιτούνταν κλασικές αριθμητικές τεχνικές για τον υπολογισμό των εκτιμητών ΜΠ. Στη βασική του μορφή είναι ένας ντετερμινιστικός αλγόριθμος σχεδιασμένος για εκτιμητές μέγιστης πιθανοφάνειας σε περιπτώσεις όπου υπάρχουν ελλιπή δεδομένα.

Αν τα δεδομένα αυτά είχαν παρατηρηθεί, η εκτίμηση θα ήταν εύκολη. Άρα, ο αλγόριθμος ΕΜ προχωρά λαμβάνοντας ως αναμενόμενες τιμές, τις τρέχουσες εκτιμήσεις των παραμέτρων και τότε χρησιμοποιεί αυτές τις προβλέψεις για να μεγιστοποιήσει την πιθανοφάνεια του δείγματος και να βελτιώσει τις εκτιμήσεις των παραμέτρων.

Ένα χαρακτηριστικό κλειδί του αλγορίθμου είναι ότι παρουσιάζει αργό και γραμμικό ρυθμό σύγκλισης ο οποίος εξαρτάται από την ποσότητα της πληροφορίας στη μερίδα των δεδομένων που είναι άγνωστη. Αν οι συνιστώσες είναι παρόμοιες στις πυκνότητές τους, τότε η σύγκλιση είναι εξαιρετικά αργή. Η σύγκλιση θα είναι

επίσης αργή όταν η λύση μέγιστης πιθανοφάνειας απαιτεί κάποιες από τις παραμέτρους-βάρη να είναι μηδέν, επειδή ο αλγόριθμος δεν μπορεί να φτάσει σε ένα τέτοιο φράγμα. Παρόλο που αυτό δεν αποτελεί τόσο σοβαρό πρόβλημα για τον υπολογισμό των εκτιμητών δεδομένων των υπολογιστών τελευταίας τεχνολογίας, μπορεί να κάνει τη μελέτη προσομοίωσης αρκετά κουραστική.

Ένα πρόσθετο και σχετικό πρόβλημα είναι αυτό της απόφασης του πότε θα σταματήσει ο αλγόριθμος. Ένας κανόνας θα ήταν να σταματήσει ο αλγόριθμος ανάλογα με τις προτιμήσεις μας για τις τιμές των εκτιμήσεων των παραμέτρων ή το πόσο μεγάλη θέλουμε να είναι η πιθανοφάνεια.

Ένα άλλο στοιχείο που αφορά τον αλγόριθμο ΕΜ και συγκεκριμένα τα μίγματα είναι ότι υπάρχουν μερικές φορές και άλλοι τρόποι να καθορίσουμε τα ελλιπή δεδομένα, οπότε δεν υπάρχει ένας και μοναδικός αλγόριθμος ΕΜ. Μπορεί δηλαδή να υπάρχουν και καλύτεροι ως προς την απόδοσή τους από αυτούς που μπορεί ένας ερευνητής να χρησιμοποιήσει.

Τέλος, να αναφέρουμε και την ιδέα της επιτάχυνσης του αλγορίθμου. Βέβαια όσο το πλήθος των παραμέτρων μεγαλώνει, η μέθοδος αυτή γίνεται λιγότερο εύχρηστη και αξιόπιστη.

3.6 Κρυμμένα Μαρκοβιανά Μοντέλα Poisson

Εφαρμογή

Θεωρούμε το Κρυμμένο Μαρκοβιανό Μοντέλο που αποτελείται από:

- την ομογενή Μαρκοβιανή αλυσίδα $\{X_t\}$, διακριτού χρόνου $t = 1,2,\dots,T$ πεπερασμένου χώρου καταστάσεων $S = \{1,2\}$, με πίνακα πιθανοτήτων μετάβασης $1^{ης}$ τάξης $P = \left(p_{ij}\right)_{i,j\in S}$ και στάσιμη κατανομή $\pi = \{\pi_1, \pi_2\}$. (Η $\{X_t\}$ είναι η μη παρατηρούμενη διαδικασία)
- την παρατηρήσιμη στοχαστική διαδικασία $\{Y_t\}$, διακριτού χρόνου $t = 1,2,\dots,T$ όπου

$$(Y_t|X_t = i) \sim Poisson(\lambda_i), \quad \text{για } i = 1,2$$

Η παράμετρος είναι το $(\theta) = (P, \lambda_1, \lambda_2)$.

Αυτό που μας ενδιαφέρει είναι, έχοντας μία σειρά T συγκεκριμένων παρατηρήσεων, έστω $y = \{y_1, y_2, \dots, y_T\}$ και προσομοιώνοντας με τον αλγόριθμο «εμπρός-πίσω» μία σειρά καταστάσεων, $x = \{x_1, x_2, \dots, x_T\}$, από την οποία θα μπορούσε να είχε προέλθει η y, να υπολογίσουμε την εκ των υστέρων κατανομή των παραμέτρων $f(\lambda_1, \lambda_2|y, x)$ με Μπεϋζιανή μεθοδολογία. Έτσι,

$$f(\lambda_1, \lambda_2|y, x) \propto f(y|\lambda_1, \lambda_2, x) \cdot f(\lambda_1, \lambda_2)$$

όπου,

$f(\lambda_1, \lambda_2|y, x)$ η από κοινού εκ των υστέρων κατανομή των παραμέτρων δοθέντων των παρατηρούμενων και των μη παρατηρούμενων δεδομένων, $f(y|x, \lambda_1, \lambda_2)$ η πιθανοφάνεια των παρατηρούμενων δεδομένων, $f(P, \lambda_1, \lambda_2)$ η από κοινού εκ των προτέρων κατανομή και θ ο παραμετρικός χώρος των λ_1, λ_2.

- Η εκ των προτέρων κατανομή

Ως εκ των προτέρων κατανομή, ορίζουμε την

$$f(P, \lambda_1, \lambda_2) = f(P) \cdot f(\lambda_1) \cdot f(\lambda_2)$$

Η $f(P)$ είναι η εκ των προτέρων κατανομή για τα στοιχεία του πίνακα P. Αν συμβολίσουμε με p_i την i γραμμή του πίνακα για $i = 1,2$ τότε ο P γράφεται

$$P = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix}.$$

Παίρνοντας λοιπόν, ως εκ των προτέρων κατανομή για την i γραμμή του πίνακα την κατανομή $Dirichlet$ με παράμετρο $w_i = (w_{i1}, w_{i2})$ έχουμε ότι $p_i \sim Dir(w_i)$ με w_i γνωστή υπερπαράμετρο για $i = 1,2$ δηλαδή

$$f(p_i) \propto p_{i1}^{w_{i1}-1} \cdot p_{i2}^{w_{i2}-1} \quad \text{για } i = 1,2$$

Άρα

$$P \sim \prod_{i=1}^{2} Dir(w_i),$$

δηλαδή

$$f(P) = \prod_{i=1}^{2} f(p_i) \propto \prod_{i=1}^{2} p_{i1}^{w_{i1}-1} \cdot p_{i2}^{w_{i2}-1}.$$

$$f(\lambda_1) = \frac{\lambda_1^{p-1} \cdot e^{-\lambda_1/q}}{q^p \cdot \Gamma(p)} \propto \lambda_1^{p-1} \cdot e^{-\lambda_1/q},$$

η εκ των προτέρων κατανομή για το λ_1.
Αν θεωρήσουμε ως εκ των προτέρων κατανομή την $\lambda_1 \sim Gamma(p, q)$.

Ομοίως, $f(\lambda_2) = \frac{\lambda_2^{p-1} \cdot e^{-\lambda_2/q}}{q^p \cdot \Gamma(p)} \propto \lambda_2^{p-1} \cdot e^{-\lambda_2/q}$ αν θεωρήσουμε ως εκ των προτέρων κατανομή την $\lambda_2 \sim Gamma(p, q)$.

- Για την εκ των υστέρων κατανομή των παραμέτρων

Πρώτα θα υπολογίσουμε τη δεσμευμένη πιθανοφάνεια των παρατηρούμενων, για δεδομένη σειρά καταστάσεων $x = (x_1, x_2)$, δηλαδή την

$$f(y|x,\lambda_1,\lambda_2) = Pr(Y = y|X = x, P, \lambda_1, \lambda_2).$$

Έτσι,

$$f(y|x,\lambda_1,\lambda_2) = \pi_{x_1} \cdot p_{x_1,x_2} \cdot f_{x_1}(y_1) \cdot f_{x_2}(y_2)$$

$$= \pi_{x_1} \cdot \prod_{i=1}^{2}\prod_{j=1}^{2} p_{ij}^{n_{ij}} \cdot \prod_{i=1}^{2}\prod_{t:x_t=i} f_i(y_t)$$

$$= \prod_{i=1}^{2}\left[\pi_{x_1} \cdot \prod_{j=1}^{2} p_{ij}^{n_{ij}} \cdot \prod_{t:x_t=i} f_i(y_t)\right]$$

όπου $f_i(y_t)$, για $i = 1,2$ είναι η σ.π.π. της $Gamma$ κατανομής με παράμετρο λ_i και n_{ij} το πλήθος των φορών που η Μαρκοβιανή αλυσίδα $\{X_t\}$ μεταπήδησε από την κατάσταση i στην κατάσταση j, δηλαδή

$$n_{ij} = \sum_{t=1}^{T} I\{x_t = i, x_{t+1} = j\}.$$

Τελικά η από κοινού εκ των υστέρων κατανομή των παραμέτρων

$$f(\lambda_1,\lambda_2|y,x) \propto$$

$$\propto f(y|x,\lambda_1,\lambda_2) \cdot f(\lambda_1,\lambda_2) = f(y|x,\lambda_1,\lambda_2) \cdot f(P) \cdot f(\lambda_1) \cdot f(\lambda_2)$$

$$= \prod_{i=1}^{2}\left[\pi_{x_1}\prod_{j=1}^{2} p_{ij}^{n_{ij}} \prod_{t:x_t=i} f_i(y_t)\right]\prod_{i=1}^{2} p_{i1}^{w_{i1}-1} p_{i2}^{w_{i2}-1}\prod_{i=1}^{2}[Gamma(p,q)Gamma(p,q)]$$

$$\propto \prod_{i=1}^{2}\left[\pi_{x_1} \cdot \prod_{j=1}^{2} p_{ij}^{n_{ij}} \cdot \prod_{t:x_t=i} f_i(y_t) \cdot p_{i1}^{w_{i1}-1} \cdot p_{i2}^{w_{i2}-1} \cdot \lambda_1^{p-1} e^{-\lambda_1/q} \cdot \lambda_2^{p-1} e^{-\lambda_2/q}\right].$$

Ενώ οι δεσμευμένες εκ των υστέρων κατανομές των επιμέρους παραμέτρων είναι:

- Όσων αφορά το p_i, για $i = 1,2$ είναι

$$f(p_i|y,x,\lambda_1,\lambda_2) \propto \prod_{j=1}^{2} p_{ij}^{n_{ij}} \times p_{i1}^{w_{i1}-1} \cdot p_{i2}^{w_{i2}-1}$$

$$\propto p_{i1}^{n_{i1}} \cdot p_{i2}^{n_{i2}} \cdot \ldots \cdot p_{ik}^{n_{ik}} \times p_{i1}^{w_{i1}-1} \cdot p_{i2}^{w_{i2}-1}$$

$$\propto \prod_{j=1}^{2} p_{ij}^{n_{ij}+w_{ij}-1}$$

δηλαδή $Dir(n_i + w_i)$, όπου $n_i = (n_1, n_2)$.

- Για το λ_1,

$$f(\lambda_1|y, x, P, \lambda_2) \propto Gamma\left(n_1 + p - 1, \sum_{t:x_t=1} y_t + q\right) \times \prod_{t:x_t=1} f_i(y_t)$$

- Για το λ_2,

$$f(\lambda_2|y, x, P, \lambda_1) \propto Gamma\left(n_2 + p - 1, \sum_{t:x_t=2} y_t + q\right) \times \prod_{t:x_t=2} f_i(y_t)$$

Κεφάλαιο 4ο

Χαρτογράφηση Ασθενειών (Disease Mapping)

Κατά την τελευταία δεκαετία, έχουμε δει μία έκρηξη ενδιαφέροντος για τη χαρτογράφηση ασθενειών, με τις πρόσφατες εξελίξεις στον τομέα των τεχνικών της χωρικής στατιστικής και την αύξηση της διαθεσιμότητας ηλεκτρονικών γεωγραφικών τεχνολογιών του συστήματος πληροφοριών (GIS) - Geographic Information System.

- Η Χαρτογράφηση ασθενειών είναι ένας από τους τομείς της εφαρμοσμένης στατιστικής που αναπτύσσεται πάρα πολύ γρήγορα και έχει μεγάλη ζήτηση.
- Περιλαμβάνει χωρικά στοιχεία και μεθόδους που κυμαίνονται από την απλή απεικόνιση ως τις προηγμένες στατιστικές.

4.1 Εισαγωγή

Ο όρος «χαρτογράφηση ασθενειών» (disease mapping) ή «ιατρική γεωγραφία» όπως σε πολλές αναφορές ονοματίζεται, είναι δυνατόν να εξετασθεί υπό διαφορετικές θεωρήσεις, ανάλογα με το που απευθύνεται. Σύμφωνα με ορισμένες απόψεις, η μελέτη της γεωγραφίας των ασθενειών διατυπώνεται στενά ως ένα πεδίο της ιατρικής ή της γεωγραφίας, όπου και στα δύο θεμέλιο λίθο αποτελεί η Στατιστική Ανάλυση. Στην πραγματικότητα, η «χαρτογράφηση ασθενειών» είναι ένα πολυδιάστατο γνωστικό αντικείμενο που προσεγγίζει την κατεύθυνση της κατανόησης των μηχανισμών μέσω των οποίων συνδέονται τα προβλήματα της ανθρώπινης υγείας στο χώρο.

Η «χαρτογράφηση ασθενειών», απεικονίζει τα αποτελέσματα των ερευνών για την ανθρώπινη υγεία σε χάρτες, οι οποίοι αφορούν σε μία ειδική κατηγορία θεματικών χαρτογραφικών απεικονίσεων. Ένας παράγοντας, που εξειδικεύει τις μεθόδους της χαρτογράφησης στην ιατρική γεωγραφία, είναι οι ιδιαιτερότητες, που άπτονται στην προσέγγιση της σχέσης γεωγραφικός χώρος και ανθρώπινη υγεία.

Ορισμένοι ερευνητές προσεγγίζουν τη «χαρτογράφηση ασθενειών» μελετώντας αρκετά συχνά τις σχέσεις μεταξύ φυσικού περιβάλλοντος και μεταδοτικών ασθενειών, ενώ άλλοι αναπτύσσουν διάφορες μεθοδολογίες για την ακριβή χαρτογράφηση διαφόρων τύπων ασθενειών, με σκοπό να προσδιορίσουν ή να κατανοήσουν καλύτερα απλές ή πολλαπλές συσχετίσεις των ασθενειών αυτών με φυσικά ή/και πολιτισμικά φαινόμενα. Υπάρχουν επίσης άλλοι που προσεγγίζουν την εξάπλωση και τη διάχυση των ασθενειών με εργαλεία γεωγραφικής ανάλυσης, ώστε να διαγνώσουν τους χωρικούς και χρονικούς μηχανισμούς επίδρασης στη μετεξέλιξη μίας ενδημικής ασθένειας σε επιδημία, ενώ ακόμα άλλες χωρικές

μελέτες που αφορούν σε ανθρώπινα προβλήματα υγείας έχουν να κάνουν με τη διαχείριση πόρων για την πρόνοια και τη θεραπεία.

Ένας σημαντικός ρόλος που παίζουν οι τεχνολογίες αιχμής για τη χαρτογράφηση των ασθενειών είναι προς την κατεύθυνση της αντιμετώπισης των περίπλοκων μοντέλων που απαιτούνται για τους χάρτες μελλοντικής πρόβλεψης της επίπτωσης των ασθενειών. Οι περισσότερες προσεγγίσεις για την παραγωγή χωρικών επιδημιολογικών μοντέλων ασχολούνται με την κατανόηση του πώς οι ασθένειες εξαπλώνονται από μία γεωγραφική περιοχή σε μία άλλη και πώς τα πρότυπα των ασθενειών αλλάζουν με το χρόνο (Cliff and Hagget 1996, Hagget 2000). Οι δυναμικοί χάρτες και η εξέλιξή τους είναι θεμελιώδεις σε αυτόν τον τομέα.

Αλλά, για στοιχεία βασισμένα σε συγκεκριμένη γεωγραφική περιοχή, οι εντοπισμένες χωρικές ενότητες (έτσι αναφέρονται οι περιοχές μελέτης) σπανίως έχουν όρια που να είναι σταθερά για την οποιαδήποτε λογική χρονική περίοδο, περιπλέκοντας έτσι και την απεικόνιση και την ανάλυση των στοιχείων μέσω Συστημάτων Γεωγραφικών Πληροφοριών-GIS (Σ.Γ.Π.).

Επί του παρόντος, ελάχιστα Σ.Γ.Π. μπορούν να δώσουν χωρικά επιδημικά μοντέλα. Ενόσω η πλήρης ολοκλήρωση των χωρικών μοντέλων σε ένα διαμορφωμένο για της ανάγκες της εργασίας Σ.Γ.Π. τελικά θα επιτυγχάνεται, άλλες ενδιάμεσες λύσεις είναι δυνατές μέχρι ο στόχος να υλοποιηθεί.

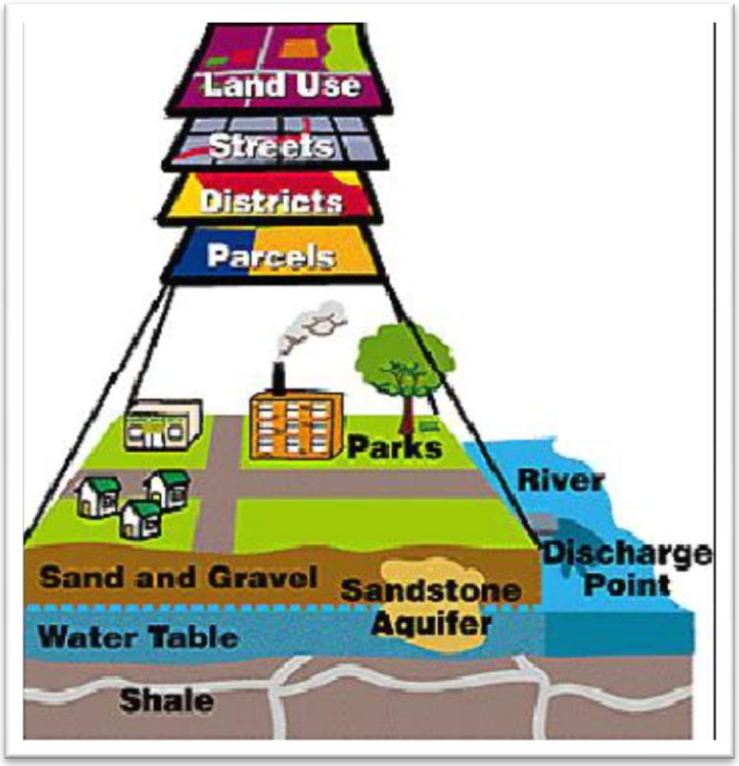
4.2 Συστήματα Γεωγραφικών Πληροφοριών (GIS)

Σύμφωνα με τον Goodchild 1985, ένα Σύστημα Γεωγραφικών Πληροφοριών (Σ.Γ.Π. - Geographic Information System, GIS), είναι ένα ολοκληρωμένο σύστημα συλλογής, αποθήκευσης, διαχείρισης, ανάλυσης και απόδοσης πληροφορίας σχετικής με φαινόμενα που εξελίσσονται στο χώρο (Goodchild, 2011).

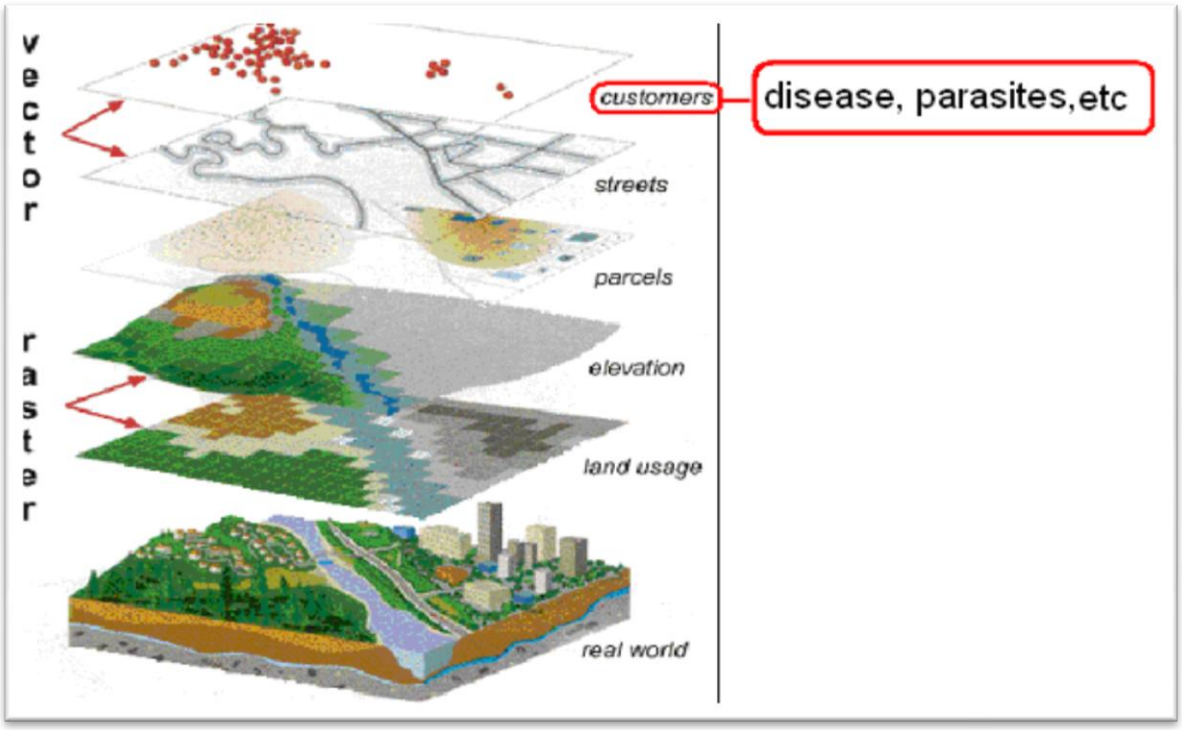
Ο όρος «ολοκληρωμένος» σημαίνει ότι το Γ.Σ.Π αντιμετωπίζεται όχι μόνο ως ένα άθροισμα μηχανημάτων και προγραμμάτων, αλλά ως μία νέα, διαφορετική τεχνολογία. Σύμφωνα με τον Burrough ένα Σ.Γ.Π αντιπροσωπεύει ένα ισχυρό σύνολο εργαλείων για τη συλλογή, αποθήκευση, ανάληψη, ανά πάσα στιγμή μετασχηματισμό και απεικόνιση χωρικών στοιχείων του πραγματικού κόσμου.

Επίσης, το Σύστημα Πληροφοριών Γης είναι ένα εργαλείο για λήψη αποφάσεων νομικής, διοικητικής, κοινωνικής και οικονομικής υφής και ένα όργανο για την ανάπτυξη και το σχεδιασμό, το οποίο αποτελείται από μία Βάση Δεδομένων που περιέχει για μία έκταση στοιχεία προσδιορισμένα στο χώρο και τα οποία σχετίζονται με τη γη και από την άλλη (αποτελείται) από διαδικασίες και τεχνικές για τη συστηματική συλλογή, ενημέρωση, επεξεργασία και διανομή των στοιχείων. Η βάση ενός Σ.Γ.Π. είναι ένα ενιαίο σύστημα (γεωγραφικής) αναφοράς, το οποίο επίσης διευκολύνει τη σύνδεση των στοιχείων μεταξύ τους καθώς και με άλλα

συστήματα που περιέχουν στοιχεία για τη γη. Συνδυάζει ισότιμα τη γεωγραφική (χαρτογραφική) και την αναλυτική (στατιστική) πληροφορία (Κουτσόπουλος, 2002).



Εικόνα 4$^{\alpha}$: Μεταφορά από τον πραγματικό κόσμο στα Σ.Γ.Π.



Εικόνα 4$^{\beta}$: Χαρτογραφικά Επίπεδα – Τελικό αποτέλεσμα

Στην παραπάνω εικόνα, φαίνεται ο τρόπος απεικόνισης των διαφόρων επιπέδων για να παραχθεί το τελικό αποτέλεσμα. Το επίπεδο ΄πελάτες΄ (customers) περιέχει σημεία, το επίπεδο "δρόμοι" (streets) περιέχει γραμμές και το επίπεδο "δέματα" (parcels) περιέχει πολύγωνα. Όλα αυτά είναι διανυσματικά στρώματα. Από την άλλη πλευρά, τα στρώματα "υψόμετρο" (elevation) και "χρήση γης" (land use) είναι στρώματα σε πλέγμα και προέρχονται από την επεξεργασία εικόνων τηλεπισκόπισης. Ουσιαστικά, ο χρήστης προσθέτει ότι είδος πληροφορίας επιθυμεί, όπως επίπτωση, συχνότητα μιας ασθένειας και οποιοδήποτε επιδημιολογικό δείκτη ή μέτρο κρίνεται απαραίτητο για την απεικόνιση, συσχέτιση και ευρύτερη ανάλυση.

Οι επιστήμες οι οποίες συνεισφέρουν στα Σ.Γ.Π. είναι η Χαρτογραφία, η Τηλεπισκόπηση/Φωτογραμμετρία, η Τοπογραφία, η Πληροφορική και τα Μαθηματικά/Στατιστική. Παράλληλα, οι εφαρμογές των Σ.Γ.Π. είναι πολλές. Για παράδειγμα στη Γεωργία, Αρχαιολογία, Περιβάλλον, Επιδημιολογία και Υγεία, δασολογία, Υπηρεσίες εκτάκτου ανάγκης, Πλοήγηση, Κτηματαγορά, Περιφερειακός και Τοπικός Σχεδιασμός, Κοινωνικές επιστήμες, Τουρισμός, Υπηρεσίες Ύδρευσης, ΔΕΗ, ΟΤΕ, Ιατρική, Επιδημιολογία, Βιολογία και άλλα.

Στο σύγχρονο κόσμο της επιστήμης τα Σ.Γ.Π. έχουν αρχίσει να αποτελούν ένα σημαντικό, ανταγωνιστικό και πρωτοποριακό εργαλείο πολλών εργασιών σχετικών με τη δημόσια υγεία και την επιδημιολογία. Ουσιαστικά, προσφέρουν νέες και επαναστατικές δυνατότητες στην επιδημιολογία διότι επιτρέπουν στον χρήστη να επιλέξει ανάμεσα σε ποικιλία επιλογών όταν ο παράγοντας «χώρος» είναι μέρος

του προβλήματος. Προσδίδουν στην ασθένεια την χωρική διάσταση, προσπαθώντας να μεταφέρουν την ασθένεια με την πληρότητα των διαστάσεων της από το πραγματικό στο ψηφιακό περιβάλλον-όσο αυτό είναι εφικτό.

Κλείνοντας, όσον αφορά τα δομικά στοιχεία ενός Σ.Γ.Π. δεν περιλαμβάνουν απλά μια βάση δεδομένων της οποίας τα στοιχεία εισάγουμε εμείς αναλόγως την περίπτωση, αλλά και μία χωρική βάση ή χαρτογραφική πληροφορία καθώς και ένα μηχανισμό ο οποίος συνδέει αυτές τις βάσεις δημιουργώντας ένα πολυδιάστατο δίκτυο (Goodchild, 1992; Clarke , 1995).

4.3 Λειτουργικές δυνατότητες Σ.Γ.Π. στη Χαρτογράφηση Ασθενειών

Οι λειτουργικές δυνατότητες ενός Σ.Γ.Π. πρακτικά μπορούν να περιγραφούν αναλύοντας τον ορισμό του είτε προσπαθώντας να απαντήσουμε στο τι είναι ένα Σ.Γ.Π. (όπως προαναφέρθηκαν στην ενότητα 4.2).

Ένα απλό παράδειγμα εντολής που μπορεί να δοθεί από το σύστημα είναι η ανάδειξη περιοχών υψηλής επικινδυνότητας για κάποιο πρόβλημα. Τα πάντα οργανώνονται από την «καρδιά» του συστήματος μέσω του πίνακα περιγραφικών δεδομένων (attribute data table). Εκεί υπάρχει δυνατότητα εισαγωγής νέας πληροφορίας, επεξεργασίας, δημιουργίας ερωτημάτων και εκτέλεσης τους, βασισμένη σε αρχές γεωμετρίας, μαθηματικών και στατιστικής (όταν υπάρχει συνδυασμός εντολών) (GIS, 1995).

4.3.1 Σύντομη Ιστορική Αναδρομή

Οι υπολογιστές εφαρμόστηκαν για πρώτη φορά στη γεωγραφία σαν αναλυτικά εργαλεία παρουσίασης κατά τη διάρκεια του 1960 (Tobler, 1959). Τα Σ.Γ.Π. ορίστηκαν ως ένα πολυδιάστατο και πολυχρηστικό πεδίο το 1970. Οι αρχές και οι παραδοχές του συστήματος βασίζονται στην θεωρία της χαρτογραφίας η οποία ακολουθεί μαθηματικά μονοπάτια. Ένα απλό παράδειγμα είναι οι αστικοί χάρτες σχεδιασμού και προγραμματισμού οι οποίοι προκύπτουν μέσω της επιλογής περιοχών και σημείων βάση πολλαπλών παραγόντων (πολυπαραγοντικά μαθηματικά μοντέλα) (Steinitz, et al., 1976).

Στόχος αυτής της διαδικασίας είναι η επαναστατική ποσοτικοποίηση ακολουθώντας τις αρχές της γεωγραφίας και κυρίως των μαθηματικών, ενώ παράλληλα συνεισφέρουν στην οργάνωση δεδομένων στην επιστήμη των υπολογιστών. Στα τέλη του 1970 με το συνδυασμό πολλών παραγόντων προέκυψε καλύτερη ποιότητα και ταχύτερη ανάπτυξη. Έτσι το σύστημα ενός Σ.Γ.Π. γίνεται διαδραστικό, συσχετίζεται με άλλα πληροφοριακά συστήματα και εφαρμόζεται σε μεγάλο εύρος επιστημών.

4.4 Εφαρμογές Σ.Γ.Π. στη Χωρική Ανάλυση

Οι επιδημιολόγοι παραδοσιακά χρησιμοποιούν χάρτες όταν αναλύουν σχέσεις μεταξύ χώρου, περιβάλλοντος, ανθρώπου και ασθένειας (Gesler, 1986). Τα Σ.Γ.Π. είναι κατάλληλα για την μελέτη αυτών των σχέσεων λόγω των δυνατοτήτων τους για χωρική ανάλυση και προσομοίωση δυναμικών (dynamics) καταστάσεων. Πρόσφατα χρησιμοποιούνται στην ανίχνευση και αποτύπωση διανυσματικών ασθενειών (Glass et al., 1995; Richards, 1993; Beck et al. 1994), ασθενειών οι οποίες έχουν το υγρό στοιχείο ως επιβαρυντικό παράγοντα (Clarke, 1991), στην περιβαλλοντική υγεία (Braddock et al., 1994; Wartenberg et al., 1993), στη μοντελοποίηση της έκθεσης σε ηλεκτρομαγνητικά πεδία (Wartenberg, 1992), στον υπολογισμό σοβαρών κινδύνων σε επίπεδο χώρας/ περιφέρειας/ νομού/ γειτονιάς (Tempalski, 1994). Επίσης στη πρόβλεψη παιδικής κακοποίησης (Barnes and Peck, 1994), και στην ανάλυση ασθενειών, πολιτικών υγείας καθώς και στο σχεδιασμό τους (Roger and Williams, 1993).

Σε πρόσφατη μελέτη, στη Βαλτιμόρη τα Σ.Γ.Π. και οι επιδημιολογικές μέθοδοι συνδυάστηκαν για να ταυτοποιήσουν και να χωροθετήσουν περιβαλλοντικούς παράγοντες κινδύνου που σχετίζονταν με λιμούς. Οικολογικά δεδομένα όπως χρήσεις γης, υγρότοποι, τύποι εδάφους, γεωλογία και δασικές εκτάσεις συλλέχθηκαν στις περιοχές κατοικίας των ασθενών και συγκρίθηκαν με δεδομένα τα οποία συγκεντρώθηκαν μέσω τυχαιοποίησης των ήδη γνωστών διευθύνσεων. Έτσι γεννήθηκε ένα μοντέλο κινδύνου (risk model) από το συνδυασμό των Σ.Γ.Π. και της ανάλυσης λογιστικής παλινδρόμησης (logistic regression analysis), ώστε να οριοθετήσει περιοχές όπου τα περιστατικά ασθενών είναι πιθανότερο να εμφανιστούν (Glass G.E., et al., 1995).

Τα Σ.Γ.Π. επιτρέπουν ανάλυση δεδομένων που προέρχονται από το Global Positioning Systems (GPS). Έτσι, αποτελούν ισχυρό εργαλείο ταυτοποίησης περιοχών ή θέσπισης προγραμμάτων παρέμβασης και ελέγχου σε περιοχές όπως η Γουατεμάλα για τη νόσο Ογκοκέρκωση (Onchocerciasis) και η Αφρική για την Τρυπανοσωμίαση (WHO, 1990).

Στο Ισραήλ τα Σ.Γ.Π. χρησιμοποιήθηκαν στο σχεδιασμό του εθνικού συστήματος υγείας για τον εντοπισμό και τον έλεγχο της Ελονοσίας. Το σύστημα συνδύασε δεδομένα των περιοχών αναπαραγωγής των κουνουπιών *Anopheles*, των νέων περιστατικών ελονοσίας, του πληθυσμού και ιδιαιτέρως του αστικού πληθυσμού. Η παρούσα γεω-βάση παρείχε τρόπους για συνεργασίες διοίκησης και ένα δίκτυο (network) κινητοποίησης της τοπικής κοινωνίας σε περίπτωση ξαφνικής εκδήλωσης της ασθένειας (Wood et al., 1994).

Η Ν.Α.Σ.Α. (National Aeronautics and Space Administration) ίδρυσε το Global Monitoring and Disease Prediction Program στο Ames Research Center, ώστε να εντοπίσει περιβαλλοντικούς παράγοντες που λειτουργούν επιβαρυντικά στην ασθένεια. Τελικό στόχο αποτέλεσε η δημιουργία ενός μοντέλου πρόβλεψης της

διανυσματικής δυναμικής του πληθυσμού και της ασθένειας (Ahearn and De Rooy, 1996).

Πιο πρόσφατες εφαρμογές παρουσιάζονται στη Νότιο Αφρική όπου στα πλαίσια της οργάνωσης, διαχείρισης και πρόληψης του HIV/AIDS δημιουργήθηκε ένα λογισμικό πρόβλεψης και διαχείρισης των δεδομένων το οποίο στο τελικό στάδιο εντοπίζει κρούσματα και προχωρεί σε μέτρα θεραπείας και δράσης (Busgeeth and Ulrike, 2004). Επίσης, στην Ιταλία έγινε οργανωμένη και συνεχής χρήση του ίδιου μοντέλου (Furlanello, et al., 2003).

Μια ακόμα χρήσιμη εφαρμογή καταγράφεται στην αντιμετώπιση ορισμένων ειδών καρκίνων, η οποία αρχικά εστιάζεται στη δημιουργία χαρτών έκθεσης (είδη, χωρική κατανομή, μεγέθη, συχνότητες, παρούσα κατάσταση). Εν συνεχεία χρησιμοποιώντας τη μεθοδολογία του Kriging προέκυψαν τέσσερα είδη χαρτών: χάρτες πρόβλεψης, χάρτες τυπικού σφάλματος, χάρτες πιθανοτήτων, και χάρτες συνυπολογισμού των τριών πρώτων. Στα πλαίσια της ίδια μελέτης προστίθεται και ένα αντίστοιχο project για την αντιμετώπιση του θυροειδή στα παιδιά της Σουηδίας, συσχετίζοντας ιατρικά αρχεία και περιβαλλοντικά δεδομένα (Krivoruchko, and Gotway, 2004). Αντίστοιχες ενέργειες γίνονται και στην Ανατολική Αφρική με τη νόσο σχιστοστομίαση (Malon, et al., 2001) καθώς και σε ποικίλα προβλήματα δημόσιας υγείας όπως το AIDS και η Φυματίωση (Tanser, et al., 2002).

Κλείνοντας, αξίζει να αναφερθεί ένα άρθρο των Rytkonen και Mika JP, όπου αναλύεται διεξοδικά η σχέση των Σ.Γ.Π. και της χωρικής ανάλυσης στην επιδημιολογία. Με το συνδυασμό αυτών των τριών μπορούμε να εξετάσουμε όλα τα φαινόμενα και τις περιπτώσεις σε μίκρο ή μάκρο – κλίμακα, καθώς και να εισάγουμε τη χρήση των GPS και της Μπεϋζιανής Στατιστικής. Σημαντικό σημείο όμως σε αυτές τις εφαρμογές είναι η σωστή χρήση και ο ακριβής έλεγχος εγκυρότητας τους διότι σε αντίθετη περίπτωση ο στόχος αντιστρέφεται και αντί να υπάρχει όφελος για τη δημόσια υγεία, επιδρά επιβαρυντικά μέσω της διεξαγωγής εσφαλμένων συμπερασμάτων (Rytkonen and Mika, 2004).

4.5 Χαρτογράφηση Ασθενειών και Σ.Γ.Π.

Συνοψίζοντας, τα Σ.Γ.Π. έχουν τη δύναμη να κατευθύνουν θέματα δημόσιας υγείας σε παγκόσμιο, εθνικό και τοπικό επίπεδο. Αυτή του η δυνατότητα, προέρχεται σε μεγάλο βαθμό από τις δυνατότητες του προγράμματος να υλοποιεί χωρική ανάλυση. Με τον όρο «χωρική ανάλυση» αναφερόμαστε στην ικανότητα να χειριζόμαστε χωρικά δεδομένα ποικίλων μορφών για να αποσπούμε την κατάλληλη πληροφορία (Κεφάλαιο 1^0, της παρούσας εργασίας). Οι χωρικές σχέσεις που προκύπτουν βασίζονται σε σχέσεις γειτνιάσεις, συσχέτισης της θέσης και σχέσεις μαθηματικού αποτελέσματος. Οι Gatrell and Bailey περιγράφουν τρεις γενικούς τύπους χωρικής ανάλυσης (Tomlin, 1990). Ξεκινούν με την οπτικοποίηση

(visualization), ανάλυση δεδομένων για διερεύνηση (exploring data analysis) και μοντελοποίηση (model building). Αυτά ποικίλουν σε πολυπλοκότητα (δηλαδή μπορεί να ξεκινούν από ένα απλό χάρτη απεικόνισης – αναπαράστασης και να φτάνουν σε στατιστικά μοντέλα όπως χωρική αλληλεπίδραση και μοντέλα διάχυσης –diffusion models). Η χρησιμότητα των χαρτών στην δημόσια υγεία υποδηλώνεται εύκολα κάνοντας αναφορά στο απλό παράδειγμα του John Snow με τους κλασσικούς χάρτες των περιστατικών χολέρας (βλ. κεφάλαιο 1^0).

Παρόλα ταύτα, τα Σ.Γ.Π. και η χωρική ανάλυση μπορούν να προσφέρουν περισσότερο από αυτό. Πολλαπλά επίπεδα πληροφορίας (layers), εξωτερικά δεδομένα και πληροφορίες, πολυκριτηριακές αναλύσεις συνθέτουν την εικόνα των δυνατοτήτων του (Gould, 1993). Εάν για παράδειγμα θέλουμε να εντοπίσουμε περιοχές αυξημένης επικινδυνότητας ή περιοχές όπου αναμένεται να εκδηλωθεί μια ασθένεια, δημιουργούμε ένα πολυπαραγοντικό μοντέλο με τους κατάλληλους συντελεστές βαρύτητας το οποίο θα αναδείξει τις ζώνες επιρροής – κινδύνου (buffers). Έτσι, θα προκύψει έγκυρη πληροφορία η οποία θα βοηθήσει στις στρατηγικές πρόληψης και αντιμετώπισης.

Μια ακόμα προσέγγιση όσον αφορά στην οπτικοποίηση, είναι να επιστρατεύσουμε αναλύσεις παλινδρόμησης (regression analysis) ώστε να παράγουμε τη γραμμική σχέση μεταξύ παραγόντων οι οποίοι εξηγούν καλύτερα την χωρική ποικιλομορφία μιας ασθένειας. Τα βάρη από το μοντέλο παλινδρόμησης μπορούν να χρησιμοποιηθούν για να δημιουργήσουν μια ολοκληρωμένη σύνθεση κινδύνων η οποία αποτυπώνεται σε ένα τελικό χάρτη βάση του οποίου θα διεξαχθούν τα συμπεράσματα (Glass et al., 1995). Ένα παράδειγμα τέτοιας εφαρμογής είναι στις Ηνωμένες Πολιτείες, η προσπάθεια αντιμετώπισης του AIDS και της ξαφνικής εξάπλωσης του στα αστικά κέντρα.

Η δεύτερη κατηγορία (ανάλυση δεδομένων για διερεύνηση) δίνει την δυνατότητα διερεύνησης άγνωστων πτυχών για τη διαμόρφωση μιας υπόθεσης εργασίας η οποία θα κατευθύνει την μελλοντική έρευνα (Tomlin, 1990). Μεταξύ των πιο διαδεδομένων μεθόδων αυτής της κατηγορίας είναι αυτές που στοχεύουν στην ταυτοποίηση και ανάδειξη ομάδων χώρου-χρόνου (space-time clusters) ή σημαντικών σημείων της ασθένειας (hot spots).

Το Openshaw’s geographic analysis machine (GAM) αποτέλεσε μια τέτοια μέθοδο που λειτούργησε μέσω ενός υβριδικού περιβάλλοντος Σ.Γ.Π. Χρησιμοποιήθηκε στην Μεγάλη Βρετανία για περιπτώσεις παιδικής λευχαιμίας, όπου συνδύασε ποικίλες μορφές δεδομένων για να καταλήξει σε ομάδες περιοχών (clusters) υψηλού κινδύνου, σε αντιμετώπιση του κινδύνου γειτονικών περιοχών και στην ανάδειξη της στατιστικά σημαντικής πληροφορίας (Openshaw et al., 1987). Παράλληλα, σε αυτή την κατηγορία χωρικής ανάλυσης εντάσσονται και οι χάρτες πιθανοτήτων και πρόβλεψης. Εδώ τα αποτελέσματα δεν εμφανίζονται ποτέ ως απόλυτοι αριθμοί αλλά ως εκτιμήσεις με διαστήματα εμπιστοσύνης. Βέβαια, υπάρχει η δυνατότητα εξομάλυνσης αυτών των αριθμών και η συγκεκριμενοποίηση τους μέσω των

μεθόδων του Bayes. Αυτός ο συνδυασμός θα μπορούσε να δώσει ακριβή αποτελέσματα τα οποία όμως προέρχονται από πολύπλοκες μεθόδους.

Η μοντελοποίηση αποτελεί την τρίτη κατηγορία χωρικής ανάλυσης. Περιλαμβάνει διαδικασίες διαμόρφωσης και ελέγχου μιας υπόθεσης σχετικής με την πορεία μια ασθένειας και της μετάδοσης της. Χρησιμοποιεί κυρίως μαθηματικούς αλγορίθμους τους οποίους έχει ήδη αποθηκευμένους το πρόγραμμα (ως εντολές). Λαμβάνει υπ' όψιν και στατιστικά μοντέλα παλινδρόμησης καθώς και διαδικασίες ψηφιακής μοντελοποίησης. Στηρίζεται σε συσχετίσεις μεταξύ χωρικών δεδομένων καθώς και χρονικών, στοιχεία για την ασθένεια και τον άνθρωπο, προσπαθώντας να καλύψει όλες τις πτυχές του πραγματικού κόσμου. Μοντέλα χωρικής διάχυσης (spatial diffusion models) αναλύουν και προβλέπουν την εξάπλωση ενός φαινομένου στο πέρας του χρόνου και του χώρου, ενώ χρησιμοποιούνται επανειλημμένως στην εκτίμηση της χωρικής διάχυσης μιας ασθένειας (Thomas, 1990).

Συνυπολογίζοντας λοιπόν το χώρο και το χρόνο μαζί με κατάλληλα επιδημιολογικά στοιχεία, τα μοντέλα μπορούν να προβλέψουν πόσο και πως θα εξαπλωθεί μια ασθένεια, χωρικά και χρονικά από τους μολυσματικούς παράγοντες σε μια περιοχή (Haggett, 1994). Έτσι, στοχεύει στη πρόληψη και κατανόηση μεταδοτικών και άλλων ασθενειών (National Center for Geographic Information and Analysis, 1990).

4.6 Τα Σ.Γ.Π. σήμερα και η Δημόσια Υγεία

Τα Σ.Γ.Π. σε συνδυασμό με τη Χωρική Στατιστική, αποτελούν μια νέα επιστημονική περιοχή στην οποία πραγματοποιείται έρευνα (Rytkonen and Mika, 2004), η οποία μπορεί να προσφέρει πολλά στη Δημόσια Υγεία και στην οργάνωση της σε μια χώρα. Σε πολλές όμως χώρες, ανάμεσα τους και η δική μας, θα ήταν δύσκολα να εφαρμοστεί συστηματικά και να προσαρμοστεί απόλυτα στο εθνικό σύστημα, λόγω των μέχρι τώρα διαφορετικών και μη οργανωμένων δομών του. Για να συμβεί αυτό απαιτείται αναδιοργάνωση και συντονισμός, συστηματική καταγραφή αρχείων και δεδομένων από κεντρικές υπηρεσίες της χώρας και την ενσωμάτωση του σε ένα δίκτυο αλληλεξαρτώμενων δομών. Δεν επιβαρύνει οικονομικά, ούτε προσδίδει στις δημόσιες δαπάνες για την υγεία, αποτελώντας έτσι ένα προσιτό πρόγραμμα έρευνας και οργάνωσης.

Η εφαρμογή των Σ.Γ.Π. και της μετέπειτα Στατιστικής Ανάλυσης δεδομένων στις πρακτικές Δημόσιας Υγείας θα πρέπει να προβληθεί ως ένας μηχανισμός αναβάθμισης και προαγωγής της υγείας. Το βέλτιστο σημείο αυτής της εφαρμογής θα ήταν ο συνδυασμός μιας πολύ καλής επιδημιολογικής μελέτης με μια καλοσχεδιασμένη χρήση και ανάλυση δεδομένων με Σ.Γ.Π., που θα στηρίζονται σε αξιόπιστα στοιχεία (National Center for Geographic Information and Analysis, 1990).

Κεφάλαιο 5^0

Τα «κρυμμένα» Μαρκοβιανά Μοντέλα στη Χαρτογράφηση ασθενειών (HMMs for Disease Mapping)

5.1 Εισαγωγή

Στα πρώτα 4 κεφάλαια της εργασίας μας, αναπτύξαμε και παρουσιάσαμε βασικές αρχές και ορισμούς της χωρικής ανάλυσης, υπό το πρίσμα της Μπεϋζιανής συμπερασματολογίας. Παρουσιάσαμε και αναδείξαμε τις έννοιες και τη σημασία της Χαρτογράφησης των ασθενειών και μελετήσαμε το βασικό αλγόριθμο προσομοίωσης της κρυμμένης αλυσίδας καταστάσεων.

Ήρθε λοιπόν η στιγμή, στο παρόν εδάφιο να ασχοληθούμε με την εφαρμογή των «κρυμμένων» Μαρκοβιανών Μοντέλων στη Χαρτογράφηση των ασθενειών, όταν οι πληροφορίες μας προέρχονται από Poisson δεδομένα.

5.2 Potts μοντέλα *(Potts models)*

Στο σημείο αυτό και πριν ξεκινήσουμε την περαιτέρω ανάλυσή μας, θα ήταν χρήσιμο να αναφέρουμε λίγες πληροφορίες για τα $Potts\ models$ και πως αυτά παίζουν σημαντικό ρόλο στην ανάγνωση και τελικά στη μελέτη ενός πίνακα, μιας χωρικής περιοχής. Παρουσιάζουμε, επομένως, ένα μαθηματικό εργαλείο παρουσίασης - μελέτης για πολυσύνθετα χωρικά συστήματα, βασισμένο στις κοντινότερες αλληλεπιδράσεις γειτόνων.

Το μοντέλο *Potts* είναι σε θέση να ερευνήσει πώς τα εσωτερικά στοιχεία μιας ομάδας (αναφερόμαστε κυρίως σε ομάδες στο χώρο π.χ. γεωγραφικά διαμερίσματα, κτιριακές εγκαταστάσεις κλπ) αντιδρούν με τον έναν ή τον άλλον τρόπο, βασισμένο σε ορισμένα χαρακτηριστικά που κάθε στοιχείο έχει. Δεδομένου ότι αυτές οι αντιδράσεις πραγματοποιούνται, αναμένουμε οι μακροσκοπικές ιδιότητες της ομάδας να εξελιχθούν. Τα μοντέλα Potts έχουν αποδειχθεί ένα πολύ χρήσιμο εργαλείο, με μια ευρεία ποικιλία διαφορετικών εφαρμογών στους τομείς όπως η βιολογία, η γεωγραφία, η κοινωνιολογία, η φυσική και η χημεία.

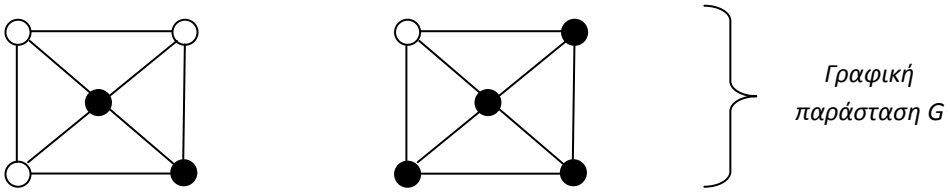
Η προέλευση των μοντέλων Potts, χρονολογείται από τις μέσες δεκαετίες του $20^{ου}$ αιώνα. Δύο μαθηματικοί, οι Julius Ashkin και Edward Teller, προσπάθησαν να πειραματιστούν με ένα μαθηματικό μοντέλο που μιμούνταν τη συμπεριφορά των διαφόρων στοιχείων μέσα σε μία ομάδα. Ένας άλλος μαθηματικός ο Cyril Domb, ανέθεσε στα πλαίσια μιας διδακτορικής διατριβής, στο μαθητή του Renfrey B. Potts, την επίλυση του παραπάνω προβλήματος. Το 1952 ο Potts δημοσίευσε τη διδακτορική διατριβή του στην οποία περιέγραφε αυτό το ιδιαίτερο πρόβλημα και τελικώς η μέθοδος αυτή πήρε και το όνομά του.

Οι μαθηματικοί, αλλά και οι διάφοροι άλλοι επιστημονικοί κλάδοι, χρησιμοποιούν τα μοντέλα Potts στις μελέτες τους. Όπου μπορούν, με τον τρόπο αυτό, να προβλέψουν τις πιθανολογικές εκβάσεις των πολυσύνθετων ομάδων και υποομάδων. Για αυτόν τον λόγο, τα μοντέλα Potts έχουν πολλές εφαρμογές στον τομέα των στατιστικών εφαρμογών. Ο στατιστικός μελετά και συνδυάζει δεδομένα και παρατηρήσεις καταλήγοντας στις ανάλογες στατιστικές μελέτες. Οι στατιστικές μελέτες με τη σειρά τους, χρησιμοποιούνται για να μελετηθούν οι πολυάριθμες μεταβλητές και να προβλεφτούν οι διάφορες εκβάσεις των μοντέλων, ενώ οι μελέτες των μηχανικών, για παράδειγμα, πώς τα εσωτερικά μόρια της ομάδας αντιδρούν σε ορισμένες εξωτερικές δυνάμεις. Τα μοντέλα Potts χρησιμοποιούνται κυρίως στις εσωτερικές αντιδράσεις μελέτης μέσα σε μία ομάδα (γειτονιά), ώστε με τον τρόπο αυτό να μπορεί να προβλεφθεί ποιές μακροπρόθεσμες εκβάσεις είναι πλέον πιθανές (π.χ. πρόβλεψη καρκίνου στα γειτονικά διαμερίσματα μίας χώρας).

5.2.1 Βασικό πλαίσιο εφαρμογής της μεθόδου

Τα μοντέλα Potts όπως αναφέραμε, είναι ένα μαθηματικό εργαλείο πρόβλεψης που οι μαθηματικοί χρησιμοποιούν για να μελετήσουν τη συμπεριφορά των διαφόρων συγκροτημάτων (ομάδων/γειτονιών). Η δομή των μοντέλων Potts δίνει τη δυνατότητα στους ερευνητές να ερευνήσουν τα εσωτερικά στοιχεία ενός σύνθετου προβλήματος γειτνίασης και να προβλέψουν πώς αυτά θα αλληλεπιδράσουν μεταξύ τους, ώστε εντέλει να μπορέσουν να καθορίσουν τη γενική συμπεριφορά της ομάδας - γειτονιάς. Δηλαδή, το μοντέλο μελετά τα μικροσκοπικά εσωτερικά στοιχεία και με βάση τις αλληλεπιδράσεις τους, μας δίνει μια πιθανή μακροσκοπική έκβαση που μπορεί να παρατηρηθεί με την πάροδο του χρόνου.

Ορισμός 1: Έστω Q ένα σύνολο από κάποιες ιδιότητες και G μία γραφική παράσταση που αποτελείται από σημεία (κόμβους) και ευθύγραμμα τμήματα που ενώνουν κάποια από τα σημεία αυτά. Μία κατάσταση b, είναι μία ανάθεση ενός στοιχείου του Q σε κάθε κόμβο του G.



Σχήμα 5.1: Οι δύο ιδιότητες για το σύνολο $Q = \{μαύρο, άσπρο\}$

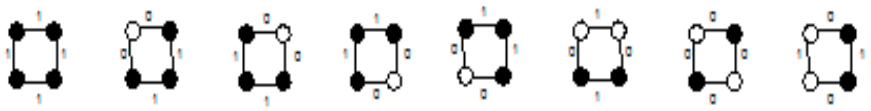
Για κάθε σημείο της γραφικής παράστασης, ορίζεται και μία στάσιμη κατάσταση. Όταν λέμε στάσιμη κατάσταση, δεν αναφερόμαστε μόνο στην ίδια θέση, αλλά και στη συνεχή εναλλαγή μεταξύ δύο ή περισσότερων θέσεων και πάντα σε κυκλική μορφή. Ο συνδυασμός στάσιμης κατάστασης και γειτνίασης καθορίζει ποια στοιχεία θα αλληλεπιδράσουν μεταξύ τους. Μερικά παραδείγματα αυτών είναι η θερμοκρασία (ζεστό ή κρύο), ο μαγνητισμός (θετικός ή αρνητικός), η κατεύθυνση (επάνω, κάτω ή λοξά), η υγεία (υγιής, άρρωστος ή πεθαμένος) και χρώμα (μπλέ, πράσινος, κόκκινος κλπ).

Δεδομένου ότι στα σημεία ορίζονται οι διαφορετικές καταστάσεις οι οποίες αλληλεπιδρούν μεταξύ τους, ανάλογα με τη θέση τους στο δικτυωτό πλέγμα, θα υπάρξει κάποια αντίδραση της γενικής ενέργειας του συστήματος. Η 'λειτουργία' που μετρά αυτή τη γενική ενέργεια μιας ομάδας ονομάζεται $Hamiltonian$. Η $Hamiltonian$ μετρά την ενέργεια μιας ιδιαίτερης κατάστασης μιας γραφικής παράστασης και πώς αυτή επηρεάζει τα άλλα σημεία μέσα στην ομάδα. Οι γραμμές δηλαδή, που ενώνουν τα σημεία σε ένα δικτυωτό πλέγμα, θα μπορούσαμε να πούμε πως δείχνουν την αλληλεπίδραση/ενέργεια/επιρροή μεταξύ των σημείων. Αυτή η επιρροή θα ποικίλει, ανάλογα με την εφαρμογή (π.χ. αν αναφερόμαστε σε γεωγραφικούς πληθυσμούς ή βιολογικούς πληθυσμούς ή ακόμα και σε δίκτυα στον τομέα των συστημάτων τηλεπικοινωνιών).

Να σημειώσουμε ότι, στη βιβλιογραφία τα γραφήματα G περιγράφονται, χρησιμοποιώντας τη συνάρτηση $Δέλτα\ του\ Kronecker$,

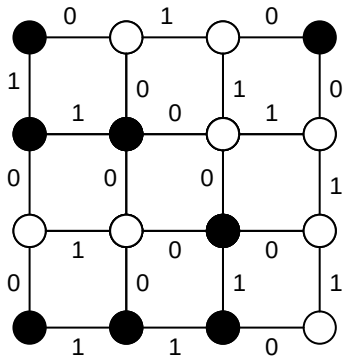
$$\delta_{\sigma_i,\sigma_j} = \begin{cases} 1 & \alpha\nu \quad \sigma_i = \sigma_j \\ 0 & \alpha\nu \quad \sigma_i \neq \sigma_j \end{cases},$$

όπου σ ορίζεται το σημείο σε ένα δικτυωτό πλέγμα και δ είναι η ενέργεια επιρροής στις διάφορες θέσεις του δικτύου. Η τιμή της ενέργειας επιρροής, αναγράφεται στις γραμμές που ενώνουν τα σημεία/κόμβους σε ένα δικτυωτό πλέγμα, όπως φαίνεται ενδεικτικά στα σχήματα που ακολουθούν, αλλά και στο παράδειγμα 5.1:



Παράδειγμα 5.1

Θεωρούμε το παρακάτω σχήμα, όπου απεικονίζει τις σχέσεις γειτνίασης, σε ένα τετράγωνο πλέγμα διαστάσεων 4×4.



Τις τιμές 1 τοποθετούμε στα τμήματα που συνδέουν τους κόμβους (edges) μεταξύ των γειτόνων με τις ομοειδείς ιδιότητες (μαύρο χρώμα) και 0 στα τμήματα που συνδέουν τους κόμβους μεταξύ των γειτόνων που έχουν τις διαφορετικές ιδιότητες (λευκό χρώμα), πάντα με βάση την εκάστοτε εφαρμογή που εξετάζουμε.

5.3 Ένα Κρυμμένο Μαρκοβιανό Μοντέλο βασισμένο στο Potts Model για τη χαρτογράφηση ασθενειών

Έφτασε πλέον η στιγμή να συνδυάσουμε όλα τα παραπάνω, για να ασχοληθούμε με τη βασική εφαρμογή της εργασίας. Στο κεφάλαιο αυτό, παρουσιάζουμε ένα Κρυμμένο Μαρκοβιανό Μοντέλο βασισμένο στο Potts Model για τη χαρτογράφηση ασθενειών.

Οι μετρήσεις μας, που είναι το πλήθος των κρουσμάτων γύρω από μία ασθένεια για μία συγκεκριμένη περιοχή, θεωρούμε ότι είναι τυχαία μεταβλητή που ακολουθεί κατανομή Poisson. Το συμπέρασμα που θα προκύψει από την επεξεργασία, εκτελείται μέσα σε ένα Μπεϋζιανό πλαίσιο εφαρμογής, με βάση τη μέθοδο Markov Chain Monte Carlo.

Η μελέτη μας εστιάζεται σε επιδημιολογικές εφαρμογές, πάνω σε δεδομένα για μία σπάνια μορφή καρκίνου στη Γαλλία. Στην έρευνά μας ενδιαφερόμαστε για τα χαρακτηριστικά γνωρίσματα των Κρυφών Μαρκοβιανών Μοντέλων και πως αυτά υιοθετούνται στον τομέα της χαρτογράφησης ασθενειών. Οδηγός για την έρευνά μας αυτή ήταν το επιστημονικό άρθρο των Peter J. Green και Sylvia Richardson που δημοσιεύθηκε το 2002 από την Αμερικάνικη Στατιστική Υπηρεσία, με τίτλο: «*Hidden Markov Models and Disease Mapping*» για την ευρύτερη περιοχή της Γαλλίας.

5.3.1 Επιλογή του Χωρικού Μοντέλου

Πριν ξεκινήσουμε, να επισημάνουμε ότι η ορολογία μας αναφέρεται στη χαρτογράφηση ασθενειών, αλλά οι ‘έννοιες’ που χρησιμοποιούμε, μπορούν εύκολα

να χρησιμοποιηθούν και σε άλλα πλαίσια στα οποία η χωρική ανάλυση είναι ενδιαφέρουσα, παραδείγματος χάριν, στην οικολογία ή τη γεωγραφική επιστήμη.

Στην ανάλυσή μας, έχουμε λάβει υπόψη όλους τους σχετικούς παράγοντες κινδύνου που μπορούν να αξιολογηθούν σε επίπεδο περιοχής. Αλλά είναι εύλογο ότι όλοι οι παράγοντες που ενεργούν στον ελλοχεύοντα κίνδυνο ασθενειών μπορούν να προσδιοριστούν ή να μετρηθούν μόνο στο απαραίτητο γεωγραφικό επίπεδο. Να σημειώσουμε στο σημείο αυτό, ότι οι επιδημιολογικές μελέτες προκύπτουν από τη φύση, με αποτέλεσμα να μην υπάρχει κανένας έλεγχος ή κι αν υπάρχει αυτός θα είναι ελάχιστος και περιορισμένος, ως προς την πηγή μεταβλητότητας τους.

Γενικά, η χωρική ετερογένεια που μπορεί να προκύψει στα διάφορα μοντέλα, εξετάζεται ιεραρχικά. Εδώ, στην ανάπτυξη της Μπεϋζιανής προσέγγισής μας, εξετάζουμε ένα μοντέλο Poisson στην απλή μορφή:

$$y_i \sim Poisson(\lambda_i E_i) \text{ για } i = 1,2,\dots,n \quad (1)$$

όπου,

i ο δείκτης της κάθε περιοχής,

y_i είναι το παρατηρούμενο πλήθος επιπτώσεων ή θανάτων ασθενειών στην περιοχή i,

E_i είναι μία σταθερά, που καθορίζει το μέγεθος πληθυσμών και

λ_i είναι μια τιμή σχετική με τη μεταβλητή κινδύνου, όπου είναι και ο κύριος στόχος του συμπεράσματός μας (ζητούμενο).

Να πούμε ότι χρησιμοποιούμε τον όρο «κίνδυνος» για να αναφερθούμε στο ρυθμό γεγονότων – κρουσμάτων ανά πληθυσμιακή μονάδα E_i.

Στη συνέχεια, εξετάζουμε τα λ_i για $i = 1,2,\dots,n$ της σχέσης (1) και τι επίδραση μπορεί να έχει αυτό στα αποτελέσματα των Poisson δεδομένων. Ενώ για να αντιμετωπίσουμε πιθανές ασυνέχειες λόγω της επίδρασης των παραμέτρων προχωρούμε στην από κοινού αντικατάσταση της τυχαίας μεταβλητής λ_i ως ακολούθως, $\lambda_i = \lambda_{z_i}$ όπου $\{\lambda_j, j = 1,2,\dots,k\}$ χαρακτηρίζει τα διαφορετικά k τμήματα επικινδυνότητας και $\{z_i, i = 1,2,\dots,n\}$ είναι οι κρυφές μεταβλητές της κατανομής που λαμβάνουν τιμές $1,2,\dots,k$. Ακολουθούμε δηλαδή τη μέθοδο της αύξησης δεδομένων (data augmentation) που προαναφέραμε.

Επισημαίνουμε ότι οι ιδιαιτερότητες της εκ των προτέρων δεν επιβάλλονται στην εκ των υστέρων συμπερασματολογία μας, υπό την έννοια ότι οποιοδήποτε μοντέλο μπορεί να μας παρέχει μία ομαλή εκτίμηση κινδύνου.

Ως ακολούθως, οι μίξεις των μοντέλων, έρχονται ως φυσική προέκταση των όσων προαναφέραμε. Το μοντέλο που προτείνουμε δηλαδή, δίνει την ευελιξία επεξεργασίας του αριθμού των κατηγοριών της κατανομής και της δύναμης της χωρικής αλληλεπίδρασης ως άγνωστα, ώστε να υπολογίζονται μαζί με τις παραμέτρους Poisson.

Ακολούθως, παρουσιάζουμε το συσχετισμένο μοντέλο κατανομής στο χώρο και περιγράφουμε την MCMC εφαρμογή μας.

5.3.2 Επιλογή του Potts μοντέλου με κατανομή Poisson

Το νέο στοιχείο της προσέγγισής μας βρίσκεται στη διαμόρφωση των μεταβλητών z_i της $\lambda_i = \lambda_{z_i}$. Όπου ο αριθμός των k μεταβλητών (ως k ορίζουμε τον αριθμό ομάδων που χωρίζονται οι n περιοχές) αντιμετωπίζεται ως σταθερός. Αυτό μας βοηθάει να ορίσουμε μία γραφική παράσταση, η οποία διαδραματίζει το ρόλο της εκ των προτέρων γραφικής παράστασης των κρυμμένων τυχαίων z_i. Ενώ να πούμε ότι ορίσαμε ως i και i', δύο περιοχές που όταν συνδέονται με τις σχέσεις $i \sim i'$ με $i \in \vartheta i'$ ή $i' \in \vartheta i$ τότε θα τις λέμε 'γειτονικές'.

Στο μοντέλο μας τα z_i διαμορφώνονται με την από κοινού κατανομή (με k σταθερό), ως εξής:

$$p(z|\psi) = e^{\psi U(z) - \theta_k(\psi)}$$

$$= \frac{e^{\psi U(z)}}{e^{\theta_k(\psi)}} = \frac{e^{\psi U(z)}}{\sum_{z \in \{1,2,\dots,k\}^n} e^{\psi U(z)}}$$

όπου

$$U(z) = \sum_{i \sim i'} I[z_i = z_{i'}]$$

$$\kappa\alpha\iota \quad \theta_k(\psi) = log\left(\sum_{z \in \{1,2,\dots,k\}^n} e^{\psi U(z)}\right) \Leftrightarrow$$

$$\Leftrightarrow e^{\theta_k(\psi)} = e^{log\left(\sum_{z \in \{1,2,\dots,k\}^n} e^{\psi U(z)}\right)} \Leftrightarrow$$

$$\Leftrightarrow e^{\theta_k(\psi)} = \sum_{z \in \{1,2,\dots,k\}^n} e^{\psi U(z)}.$$

Με την παράμετρο αλληλεπίδρασης ψ να είναι μη αρνητική, είναι προφανές ότι για θετικό ψ, η $p(z|\psi)$ ευνοεί πιθανοθεωρητικά την κατανομή όπου οι περιοχές είναι γειτονικές.

Για να ολοκληρώσουμε το μοντέλο μας, προσδιορίζουμε τα εκ των προτέρων δεδομένα μας για τις παραμέτρους λ, ψ. Εισάγουμε μία εκ των προτέρων κατανομή $\lambda_j \sim Gamma(a, b)$, όπου $\lambda_j, j = 1,2, \dots, k$. Επίσης, έχουμε υποθέσει ότι η $b \sim Gamma(b_1, b_2)$ από όπου και παίρνουμε τις υπερπαραμέτρους a και b.

Να τονίσουμε ότι επιλέγουμε για $a = 1$ και $b = {\sum_i E_i}/{\sum_i y_i}$, ενώ για τις εφαρμογές που ασχολούνται με την επιδημιολογία έχουμε $\sum_i E_i = \sum_i y_i$ για $b = 1$. Μάλιστα, προτιμάμε να κρατάμε μία ιεραρχική μορφή στα λ_j, δηλαδή $\lambda_1 < \lambda_2 < \cdots < \lambda_k$. Συνεπώς, η δεσμευμένη, ως προς λ από κοινού εκ των προτέρων κατανομή γίνεται:

$$p(\lambda|k,a,b) = k!\, I[\lambda_1 < \lambda_2 < \cdots < \lambda_k] \prod_{j=1}^{k} \frac{b^a \lambda_j^{a-1} e^{-b\lambda_j}}{\Gamma(a)}$$

$$= k! \prod_{j=1}^{k} e^{-\lambda_j}$$

$$= k!\, e^{-\sum_j \lambda_j} \qquad \text{για } a = b = 1 \ \& \ k \ \sigma\tau\alpha\theta\varepsilon\rho\acute{o}$$

$$\propto e^{-\sum_j \lambda_j}$$

Θεωρούμε την υπερπαράμετρο σταθερή και $\psi = 0.2$. Το βήμα αυτό είναι απαραίτητο χάριν υπολογιστικής ευκολίας, επειδή η κανονικοποίηση της $\theta_k(\psi)$ μπορεί να αποθηκεύει σε έναν πίνακα, μη παρατηρούμενο, τα μη χρησιμοποιούμενα δεδομένα.

Τελικά, η από κοινού κατανομή μας, που μοντελοποιεί όλες τις μεταβλητές που αντιστοιχούν στο μοντέλο μας είναι η

$$p(k)p(\psi)p(\lambda|k,a,b)p(z|k,\psi)p(y|\lambda,z)$$

όπου $p(k)$ σταθερό

$p(\psi)$ με ψ να είναι μία παράμετρος

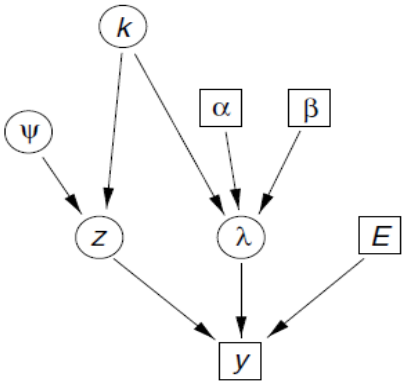
$p(\lambda|k,a,b) \propto p(\lambda|k)$ για $a = b = 1$

$p(z|k,\psi) \propto p(z|\psi) = e^{\psi U(z) - \theta_k(\psi)} = \frac{e^{\psi U(z)}}{e^{\theta_k(\psi)}} = \frac{e^{\psi U(z)}}{\sum_{z \in \{1,2,\ldots,k\}^n} e^{\psi U(z)}}$

για $U(z)$ στον αριθμητή 'δικό μας' προς όλα τα δυνατά z

και $p(y|\lambda,z)$ το ζητούμενο.

Διαγραμματικά οι παραπάνω υποθέσεις μας, παριστάνονται ως ακολούθως στο παρακάτω σχήμα, όπου με τη βοήθεια μίας γραφικής απεικόνισης, απομονώνουμε τα δεδομένα, τις σταθερές και τα ζητούμενα, λαμβάνοντας έτσι το λεγόμενο *'DAG for the Potts Spatial Mixture Model'*:



5.3.3 Μπεϋζιανή Συμπερασματολογία

Στο τελικό αυτό στάδιο χρησιμοποιούμε τις MCMC μεθόδους που έχουμε αναπτύξει, για να γίνει συμπερασματολογία για τα Χωρικά Κρυμμένα Μαρκοβιανά μοντέλα Poisson. Το μοντέλο μας προερχόμενο από τις μίξεις ακολουθεί έναν αριθμό από διαφορετικές κινήσεις. Κάθε κίνηση ενημερώνει το σύνολο των μεταβλητών της εκ των υστέρων κατανομής, η οποία με τη σειρά της παράγει μία αλυσίδα Markov.

Ο αριθμός των παραμέτρων που προκύπτει ενημερώνει την κατανομή z, δηλαδή τις παραμέτρους της κατανομής λ_j, ενώ η χωρική παράμετρος αλληλεπίδρασης είναι σταθερή και $\psi = 0.2$. Το z επομένως ενημερώνεται αυτόματα από ένα δειγματολήπτη Gibbs.

Για την καλύτερη κατανόηση του τύπου να διευκρινίσουμε, ότι θεωρούμε n_{ij} το πλήθος των γειτόνων της περιοχής i που ανήκουν στην $j-$ οστή συνιστώσα της μίξης (δηλαδή $j \in \{1,2,\dots,k\}$ δείχνει πόσοι γείτονες ανήκουν στη $j-$ οστή ομάδα), και έχουμε

$$p(z_i = j|\dots) \propto e^{-\lambda_j E_i} \lambda_j^{y_i} e^{\psi n_{ij}}.$$

Επίσης να σημειώσουμε ότι, σε αντίθεση με τις απλές μίξεις κατανομών, εδώ τα z_i δεν είναι ανεξάρτητα δεδομένα όλων των άλλων μεταβλητών, άρα και δεν μπορούν να ενημερωθούν ταυτόχρονα. Η full Conditional της παραμέτρου ψ ως προς τις άλλες παραμέτρους, θα δίνεται ως ακολούθως:

$$p(\psi|\dots) \propto p(\psi) e^{\psi U(z) - \theta_k(\psi)}$$

$$\propto p(\psi) p(z|\psi)$$

Εφαρμόζοντας αλγόριθμο Metropolis δεχόμαστε απόκλιση της τάξης του ± 0.1 για την παράμετρο ψ. Τέλος, χρειάζεται κάθε φορά να 'αναβαθμίζουμε' την παράμετρο

λ_j. Αυτό που κάνουμε, είναι να υπολογίζουμε πρώτα τον μετασχηματισμό του $log\lambda_j$ και σε κάθε τιμή του προτείνουμε διαταραχή βάσης κανονικής κατανομής με μέση τιμή μηδέν και οι νέες τιμές επανακαθορίζονται στην αυξημένη τιμή της νέας πλέον παραμέτρου λ_j^c. Αυτό έχει ως αποτέλεσμα όσο και να πυκνώνουν οι προτάσεις που κάνουμε για τη νέα τιμή, αυτές να αθροίζονται στο $k!$ επομένως, όσοι όροι και να εμφανίζονται στον αριθμητή και τον παρονομαστή της αναλογίας Hasting-Metropolis να θεωρούνται σταθερές και να απαλοίφονται.

Τελικώς, η πιθανότητα αποδοχής για τη συμπλήρωση όλων των ανανεώσεων, διαμορφώνεται από την εκ των προτέρων αναλογία, τη likelihood ratio, όπου

$$p(\lambda|k,a,b,z,\psi) = \frac{p(\lambda^c|k,a,b)p(z|\dots)}{p(\lambda|k,a,b)p(z|\dots)}, \qquad \text{και άρα}$$

$$p = $$

$$= min\left\{1, \frac{\prod_{j=1}^{k} k!\, I[\lambda_1 < \lambda_2 < \cdots < \lambda_k] \prod_{j=1}^{k} \frac{b^a \lambda_j^{c\,a-1} e^{-b\lambda_j^c}}{\Gamma(a)} \prod_{j=1}^{k} e^{-\lambda_j^c E_i} \lambda_j^{c\,y_i} e^{\psi n_{ij}}}{\prod_{j=1}^{k} k!\, I[\lambda_1 < \lambda_2 < \cdots < \lambda_k] \prod_{j=1}^{k} \frac{b^a \lambda_j^{a-1} e^{-b\lambda_j}}{\Gamma(a)} \prod_{j=1}^{k} e^{-\lambda_j E_i} \lambda_j^{y_i} e^{\psi n_{ij}}}\right\}$$

$$= min\left\{1, \prod_{j=1}^{k} \left(\frac{\lambda_j^c}{\lambda_j}\right)^{a-1} \cdot e^{-b\left(\lambda_j^c - \lambda_j\right)} \cdot \frac{\left(\lambda_j^c\right)^{\sum_{i:z_i=j} y_i}}{\left(\lambda_j\right)^{\sum_{i:z_i=j} y_i}} \cdot \frac{e^{-\lambda_j^c \sum_{i:z_i=j} E_i}}{e^{-\lambda_j \sum_{i:z_i=j} E_i}}\right\}$$

$$= min\left\{1, \prod_{j=1}^{k} \left(\frac{\lambda_j^c}{\lambda_j}\right)^{a-1+\sum_{i:z_i=j} y_i} \cdot exp\left\{-\left(\lambda_j^c - \lambda_j\right) \cdot \left[\sum_{i:z_i=j} E_i + b\right]\right\}\right\}$$

και εφαρμόζοντας το μετασχηματισμό του λογαρίθμου λαμβάνουμε τελικώς

$$p = min\left\{1, \prod_{j=1}^{k} \left(\frac{\lambda_j^c}{\lambda_j}\right)^{a+\sum_{i:z_i=j} y_i} \cdot exp\left\{-\left(\lambda_j^c - \lambda_j\right) \cdot \left[b + \sum_{i:z_i=j} E_i\right]\right\}\right\}.$$

Κεφάλαιο 6^{ο}

Εφαρμογές σε προσομοιωμένα και πραγματικά δεδομένα

6.1 Εισαγωγή

Με βάση τα όσα θεωρητικά παρουσιάσαμε στα προηγούμενα κεφάλαια, ερχόμαστε τώρα να μελετήσουμε μερικά συγκεκριμένα παραδείγματα εφαρμογής της θεωρίας των χωρικών κρυμμένων Μαρκοβιανών μοντέλων Poisson, στη χαρτογράφηση ασθενειών. Οι εφαρμογές μας, αφορούν σε προσομοιωμένα και στη συνέχεια σε πραγματικά δεδομένα.

6.2 Εφαρμογές σε Προσομοιωμένα δεδομένα

Στο κομμάτι αυτό, ερευνούμε τα ιδιαίτερα χαρακτηριστικά γνωρίσματα του μοντέλου μας και της απόδοσής του. Για το σκοπό αυτό, χρησιμοποιούμε το χωρικό σχεδιάγραμμα των 94 γαλλικών διαμερισμάτων στο σύνολο της ηπειρωτικής χώρας. Για να εξετάσουμε τα διαφορετικά χαρακτηριστικά του μοντέλου μας, παράγουμε τρία σύνολα δεδομένων που αντιστοιχούν στα αντιπαραβαλλόμενα γεωγραφικά χαρακτηριστικά γνωρίσματα, υπό τους προσομοιωμένους κινδύνους.

Ειδικότερα, ως «Block4» αναφέρεται η χωρική δομή όπου η μεταβλητή λ λαμβάνει τιμή 0.7 γενικά, ενώ τέσσερις ομάδες περιοχών λαμβάνουν τιμή του λ ίση με 1.5. Αυτές οι τέσσερις ομάδες περιοχών αποτελούνται από μία ενιαία πληθυσμιακά περιοχή, από μία ομάδα πέντε αγροτικών διαμερισμάτων και από δύο περιοχές εκ των οποίων η μία είναι στο βορρά και η άλλη στο νότο (σχήμα 6^{α}). Ως «Βορράς – Νότος» αναφέρεται η χωρική δομή όπου η μεταβλητή λ που λαμβάνει τιμή 0.7 περιορίζεται στις περιοχές του βόρειου τμήματος της Γαλλίας, ενώ για τις περιοχές του νότιου τμήματος η μεταβλητή λ λαμβάνει τιμή 1.5 (σχήμα 6^{β}). Τέλος, χωρίζουμε την ευρύτερη περιοχή της Γαλλίας σε τρεις οριζόντιες ζώνες επικινδυνότητας, όπου η μεταβλητή λ που λαμβάνει τιμή 0.7 περιορίζεται στο βόρειο τμήμα, η μεταβλητή λ που λαμβάνει τιμή 1.1 περιορίζεται στο κεντρικό τμήμα και η μεταβλητή λ που λαμβάνει τιμή 1.5 περιορίζεται στο νότιο τμήμα ($3_{η}$ ζώνη).

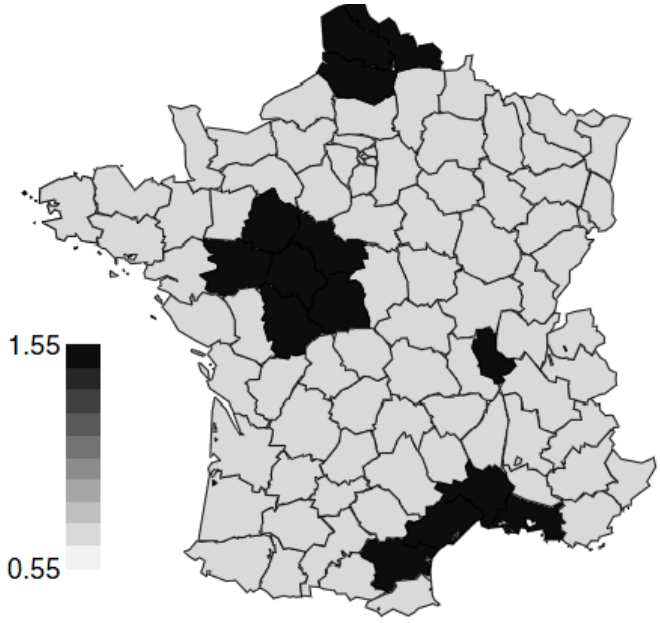
Για κάθε σύνολο δεδομένων, οι παρατηρούμενες τιμές προσομοιώθηκαν από την κατανομή

$$y_i \sim Poisson(\lambda_i E_i) \quad \text{για } i = 1,2,\dots,n$$

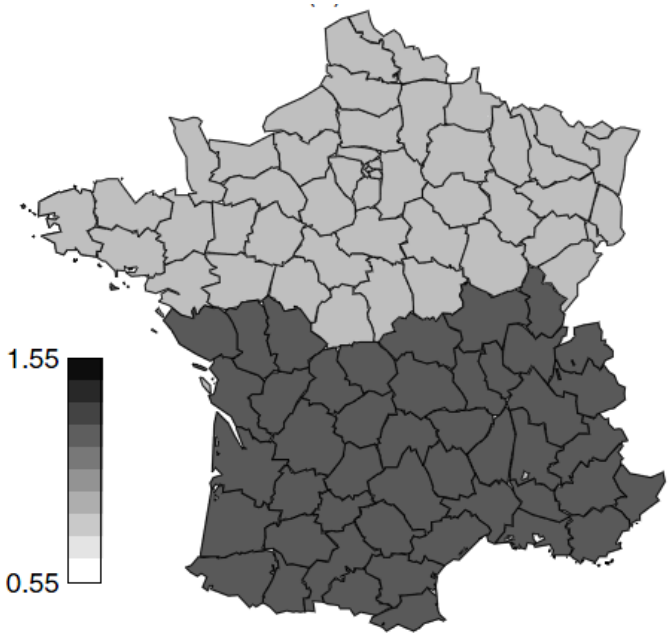
όπου οι αναμενόμενοι αριθμοί πληθυσμών επιλέχτηκαν βάσει της γαλλικής δομής πληθυσμών και αντιστοιχούν στα πραγματικά δεδομένα που θα αναλύσουμε στη συνέχεια (παράγραφος 6.3). Για το σχήμα μας «Block4» (σχήμα 6^{α}) και για το επόμενο σχήμα «Βορράς-Νότος» (σχήμα 6^{β}), αυτοί οι αριθμοί αντιστοιχούν στον

αναμενόμενο αριθμό θανάτων από τον καρκίνο του λάρυγγα στις γυναίκες για την περίοδο 1986 – 1993 και για το ηλικιακό εύρος από 2 έως 58 έτη.

Ο λόγος θνησιμότητας στον τομέα της επιδημιολογίας (SMR) παριστάνεται μαθηματικά ως το πηλίκο: $SMR = {}^{y_i}/_{E_i}$.



Σχήμα 6$^{\alpha}$: Block4



Σχήμα 6$^{\beta}$: Βορράς-Νότος

Στα παραπάνω σχήματα 6^{α} και 6^{β} απεικονίζεται η κατανομή των τιμών του λ στην περιοχή της Γαλλίας σύμφωνα με τις τιμές που προαναφέραμε. Φαίνονται καθαρά και τα 94 γεωγραφικά διαμερίσματα, ενώ οι αποχρώσεις του γκρι και του μαύρου δείχνουν με λεπτομέρεια τις διαβαθμίσεις των τιμών του λ.

6.2.1 Ανάλυση προσομοιωμένων δεδομένων

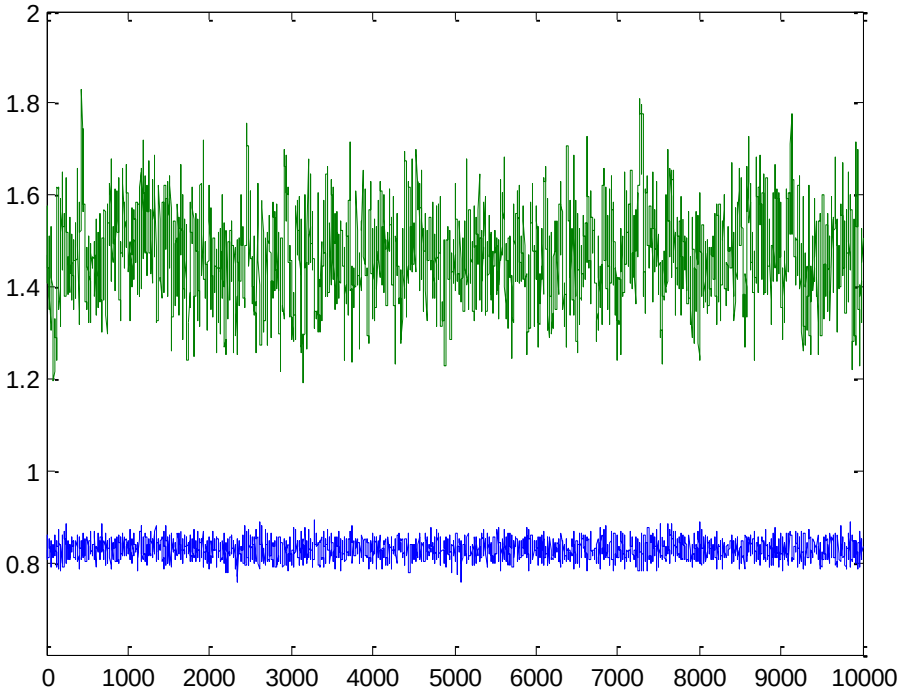
Όλα τα αποτελέσματα αντιστοιχούν σε τρέξιμο 100.000 ΄κύκλων' του αλγορίθμου όταν $\lambda_1 = 0.7$ και $\lambda_2 = 1.5$. Η απόδοση της μίξης ήταν ικανοποιητική, με τα ποσοστά περίπου 10% αποδοχής, εκτός από περιπτώσεις όπου τα στοιχεία υποστηρίζουν ένα χαμηλό αριθμό δεδομένων. Το κύριο ενδιαφέρον μας είναι στη χωρική μεταβλητή λ_{z_i} με βάση τις εκ των υστέρων πιθανότητες.

Η κεντρική ιδέα της προσομοίωσης ήταν για τα συγκεκριμένα λ_1 και λ_2 και έχοντας τους συντελεστές πληθυσμού για κάθε περιοχή E_i, να προσομοιώσουμε τα Y_i για $i = 1,2,\dots,94$ της κάθε περιοχής, δηλαδή τις παρατηρούμενες τιμές κρουσμάτων, από την κατανομή $Poisson(\lambda_i E_i)$, όπου τα παρατηρούμενα SMRs θα δίνονται από το πηλίκο $^{y_i}/_{E_i}$.

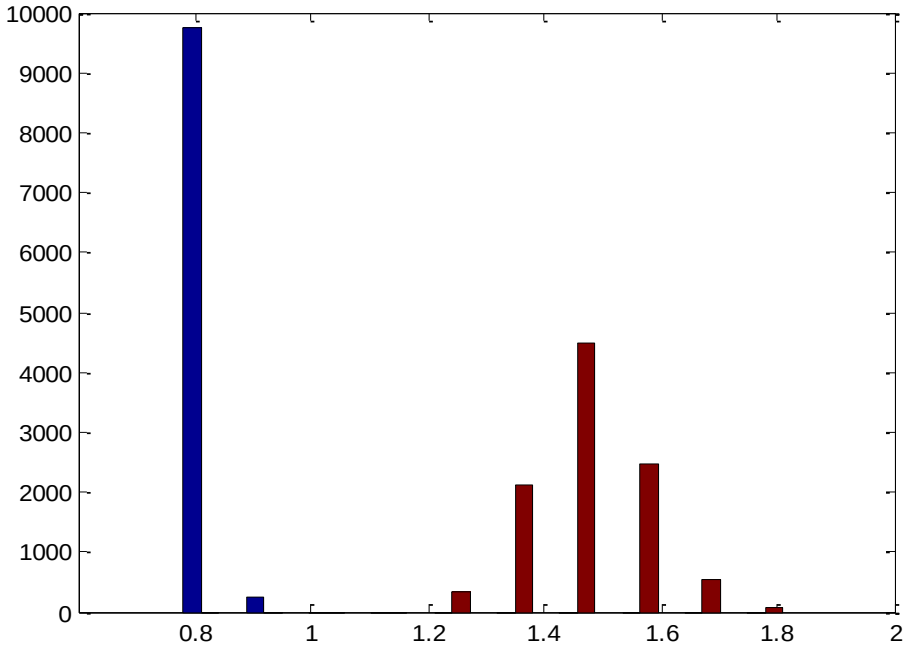
A) Για την απεικόνιση «Block4» λαμβάνουμε τα ακόλουθα αποτελέσματα: παίρνουμε μία εκ των υστέρων μέση τιμή για το $\lambda_1 \cong 0.8294$ με τυπική απόκλιση 0.0177 και για το $\lambda_2 \cong 1.4666$ με τυπική απόκλιση 0.0942, αντίστοιχα.

Να αναφέρουμε ότι για τον αλγόριθμο προσομοίωσης, οι 83 από τις 94 περιοχές που εξετάσαμε κατατάχθηκαν στην πρώτη κατηγορία με $\lambda_1 \cong 0.8294$ και οι υπόλοιπες στη δεύτερη με $\lambda_2 \cong 1.4666$. Το συμπέρασμα αυτό προέκυψε από τον πίνακα των z που λάβαμε κατά την εφαρμογή του αλγορίθμου. Δηλαδή το ποσοστό κατανομής των z είναι 88.29% στη συνιστώσα του μίγματος με $\lambda_{z_i} = 1$, όπως και αναμενόταν από την κατανομή των περιοχών στον παραπάνω χάρτη 6^{α}.

Στο $\sigma\chi\acute{\eta}\mu\alpha$ 6.1 παρουσιάζεται το χρονόγραμμα για τον διαχωρισμό των λ στο δείγμα των 94 περιοχών, ενώ στο $\sigma\chi\acute{\eta}\mu\alpha$ 6.2 παρουσιάζεται το αντίστοιχο ιστόγραμμα των λ που κράτησε ο αλγόριθμος μας δείχνοντας σε ποια κατηγορία ανήκει το κάθε ένα από αυτά προσεγγιστικά.

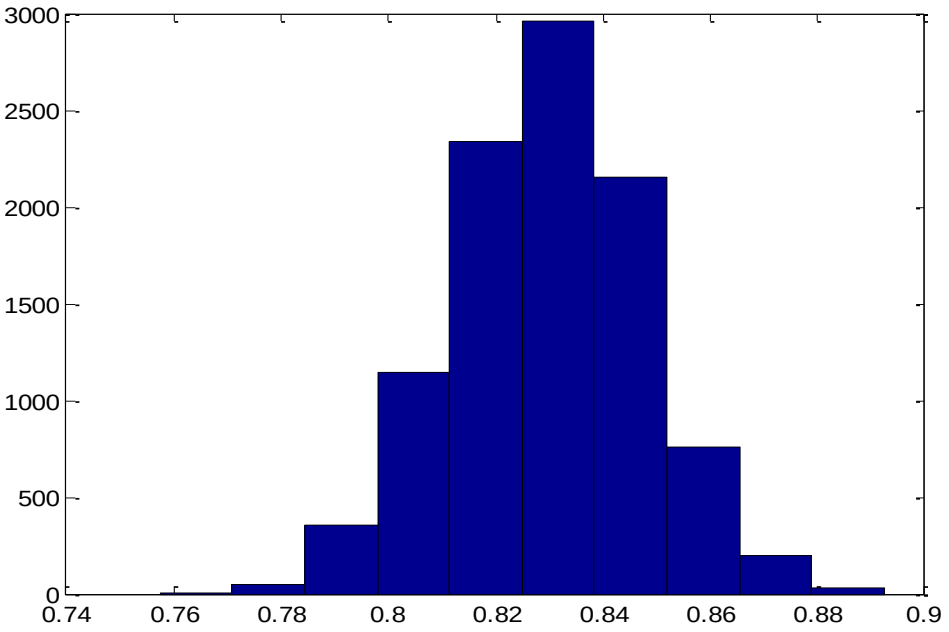


Σχήμα 6.1: Το χρονόγραμμα από τις εκ των υστέρων τιμές των λ_1 (με μπλέ) και λ_2 (με πράσινο), αντίστοιχα.

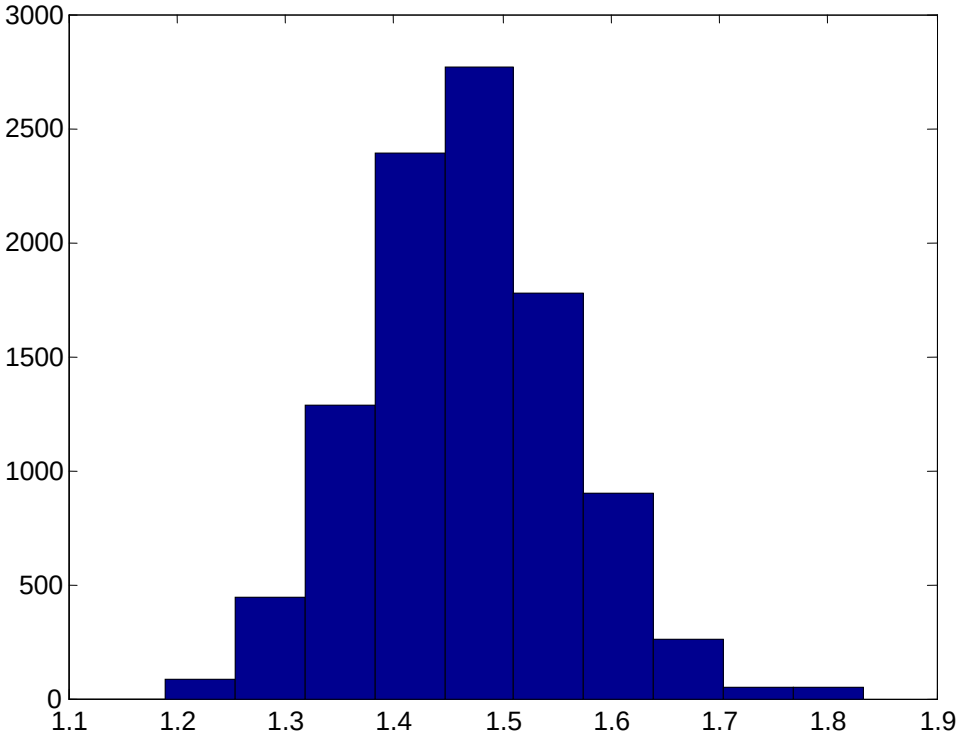


Σχήμα 6.2: Το ιστόγραμμα των λ_1 (μπλέ) και λ_2 (κόκκινο) που φανερώνει πως αυτά κατανέμονται στις μέσες τιμές σύγκλισης, είναι δηλαδή τα λ που κράτησε ο αλγόριθμος *(draws).*

Τα αντίστοιχα ιστογράμματα για τα λ_1 και λ_2, όπως αυτά απεικονίζονται κατά την ανάλυσή μας παρουσιάζονται ακολούθως:



Σχήμα 6.3: Το ιστόγραμμα του δείγματος από την εκ των υστέρων κατανομή του λ_1



Σχήμα 6.4: Το ιστόγραμμα του δείγματος από την εκ των υστέρων κατανομή του λ_2

Σύμφωνα με τη διάταξη του χάρτη 6^{α}, όπου με σκούρο χρώμα απεικονίζονται οι περιοχές της Γαλλίας με υψηλό δείκτη επικινδυνότητας και με γκρι απόχρωση οι περιοχές με μικρότερο δείκτη επικινδυνότητας, μπορούμε να εξάγουμε συμπεράσματα για την κάθε περιοχή ξεχωριστά.

Για παράδειγμα, επιλέγουμε τυχαία την περιοχή 12 *(Aveyron)*, όπου βρίσκεται στο Νότιο άκρο της Γαλλίας και εξετάζουμε με τον παραπάνω αλγόριθμο σε τι βαθμό επικινδυνότητας μπορεί να καταταχθεί.

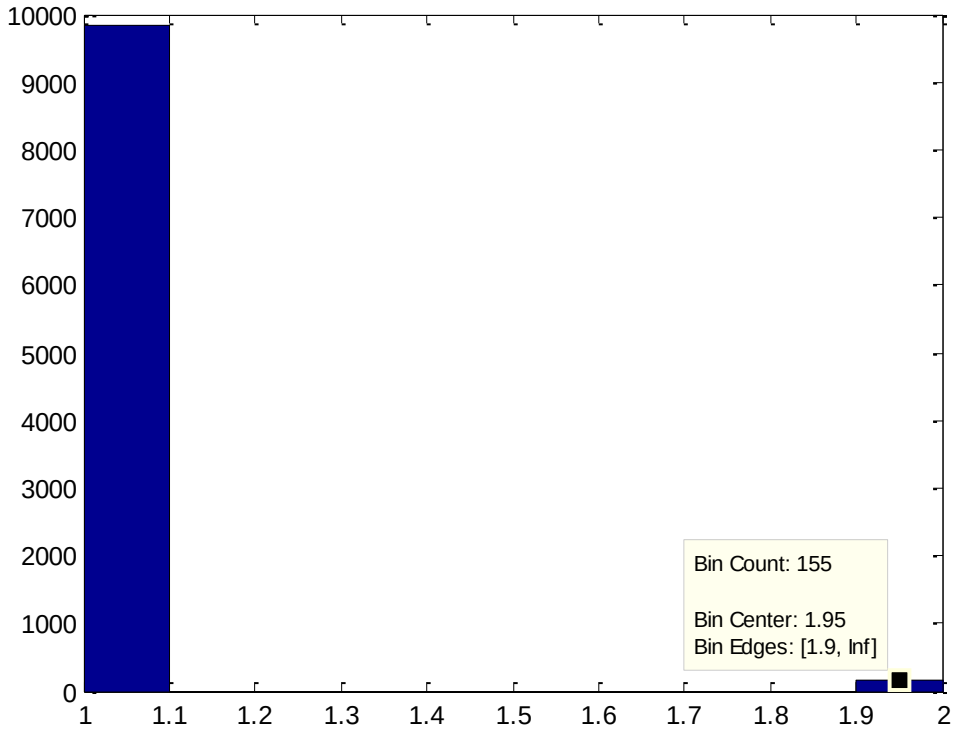
Πριν την παρουσίαση του ιστογράμματος για τη συγκεκριμένη περιοχή, είναι σημαντικό να γίνει γνωστό, πως η αρίθμηση των γεωγραφικών διαμερισμάτων της Γαλλίας έχει γίνει με βάση κριτήρια που υπάρχουν στο Εθνικό σύστημα υγείας της Γαλλίας και όπου έχουν μετρηθεί και καταχωρηθεί το 1990, σχήμα 6^{γ}. Κατά την ανάλυσή μας, δεν έχουμε λάβει υπόψη την Κορσική (αρ. 20 ή 2_{α} & 2_{β}).

Σχήμα 6^{γ}: Αριθμητική απεικόνιση των γεωγραφικών διαμερισμάτων της Γαλλίας,

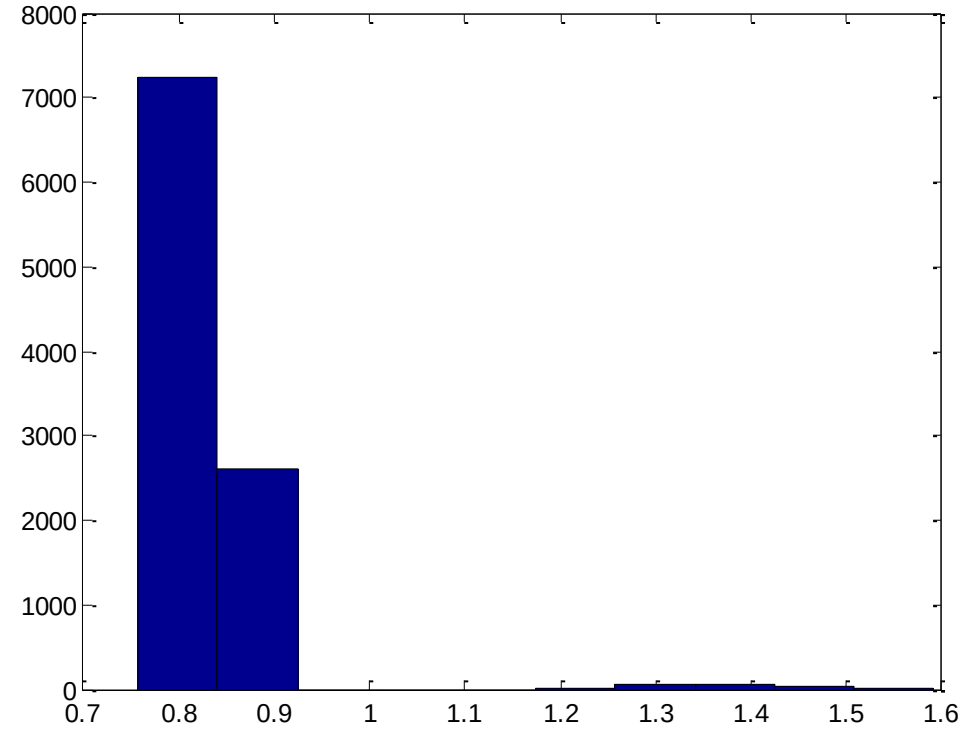
France métropolitaine		
01 Ain	32 Gers	64 Pyrénées-Atlantiques
02 Aisne	33 Gironde	65 Hautes-Pyrénées
03 Allier	34 Hérault	66 Pyrénées-Orientales
04 Alpes-de-Haute-Provence	35 Ille-et-Vilaine	67 Bas-Rhin
05 Hautes-Alpes	36 Indre	68 Haut-Rhin
06 Alpes-Maritimes	37 Indre-et-Loire	69 Rhône
07 Ardèche	38 Isère	70 Haute-Saône
08 Ardennes	39 Jura	71 Saône-et-Loire
09 Ariège	40 Landes	72 Sarthe
10 Aube	41 Loir-et-Cher	73 Savoie
11 Aude	42 Loire	74 Haute-Savoie
12 Aveyron	43 Haute-Loire	75 Paris
13 Bouches-du-Rhône	44 Loire-Atlantique	76 Seine-Maritime
14 Calvados	45 Loiret	77 Seine-et-Marne
15 Cantal	46 Lot	78 Yvelines
16 Charente	47 Lot-et-Garonne	79 Deux-Sèvres
17 Charente-Maritime	48 Lozère	80 Somme
18 Cher	49 Maine-et-Loire	81 Tarn
19 Corrèze	50 Manche	82 Tarn-et-Garonne
2A Corse-du-Sud	51 Marne	83 Var
2B Haute-Corse	52 Haute-Marne	84 Vaucluse
21 Côte-d'Or	53 Mayenne	85 Vendée
22 Côtes-d'Armor	54 Meurthe-et-Moselle	86 Vienne
23 Creuse	55 Meuse	87 Haute-Vienne
24 Dordogne	56 Morbihan	88 Vosges
25 Doubs	57 Moselle	89 Yonne
26 Drôme	58 Nièvre	90 Territoire de Belfort
27 Eure	59 Nord	91 Essonne
28 Eure-et-Loir	60 Oise	92 Hauts-de-Seine
29 Finistère	61 Orne	93 Seine-Saint-Denis
30 Gard	62 Pas-de-Calais	94 Val-de-Marne
31 Haute-Garonne	63 Puy-de-Dôme	95 Val-d'Oise

Σχήμα 6[δ]: Αριθμητική αντιστοιχία των γεωγραφικών διαμερισμάτων της Γαλλίας
Η περιοχή 2A & 2B που αντιστοιχούν στην Κορσική, δεν ελήφθησαν υπόψη κατά την ανάλυση

Επανερχόμενοι στην ανάλυσή μας για την περιοχή 12 *(Aveyron)* της Γαλλίας, λαμβάνουμε το παρακάτω ιστόγραμμα, σχήμα 6.5_a, όπου μπορούμε να συμπεράνουμε ότι από τις 10.000 φορές τρεξίματος του αλγορίθμου μας, μόνο τις 155 φορές κατατάχθηκε στη δεύτερη ζώνη επικινδυνότητας ($\lambda_2 \cong 1.4666$) με τον υψηλό συντελεστή. Αποτέλεσμα που μας δείχνει ότι η περιοχή αυτή *(Aveyron)* παρουσιάζει γενικά χαμηλό συντελεστή θνησιμότητας.



Σχήμα 6.5$_a$: Ιστόγραμμα της περιοχής 12 (Aveyron) της Γαλλίας που φαίνεται ότι ο λόγος θνησιμότητας είναι χαμηλός, τα δεδομένα κατατάσσονται στην πρώτη κατηγορία του *z* (δηλ. στο λ_1)

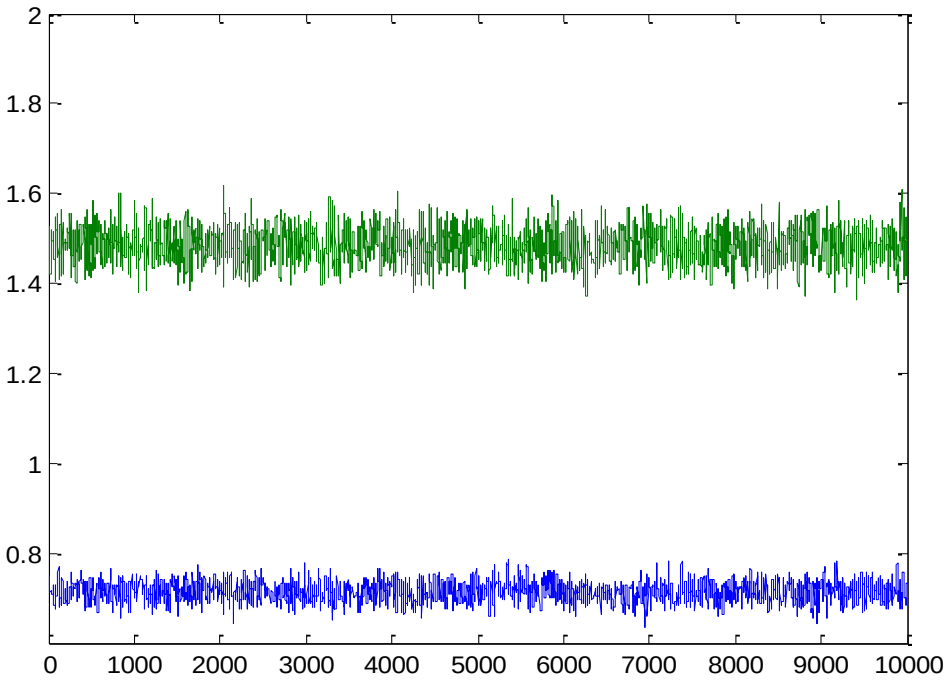


Σχήμα 6.5$_b$: Ιστόγραμμα της περιοχής 12 (Aveyron) της Γαλλίας που φαίνεται πως κατανέμεται το $\lambda_{z_{12}}$, δείχνοντας και με αυτό τον τρόπο ότι ο λόγος θνησιμότητας είναι χαμηλός.

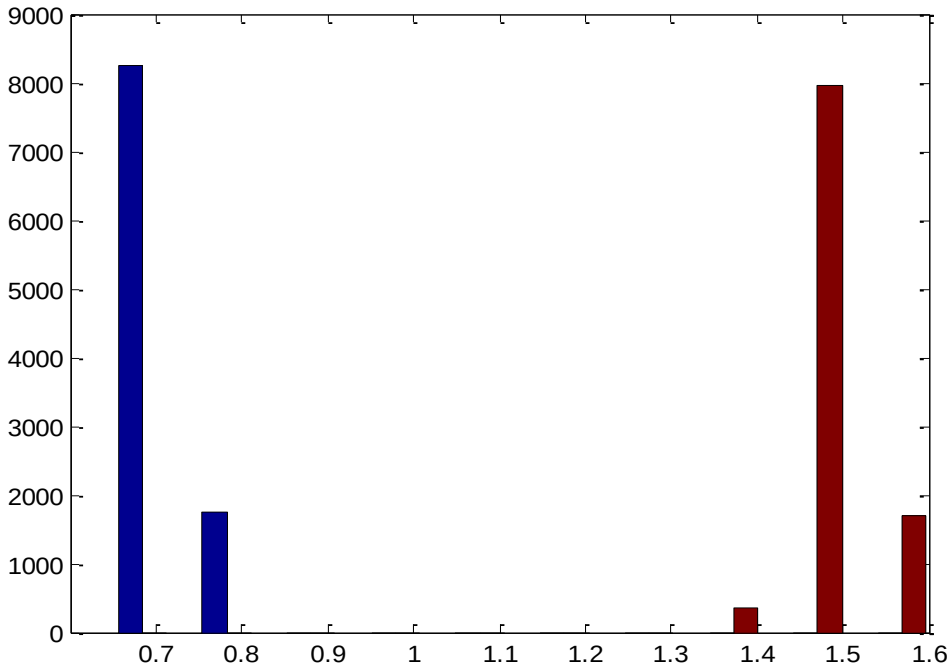
B) Εν συνεχεία, για την απεικόνιση «Βορράς – Νότος» λαμβάνουμε τα ακόλουθα αποτελέσματα: παίρνουμε μία εκ των υστέρων μέση τιμή για το $\lambda_1 \cong 0.7143$ με τυπική απόκλιση 0.0201 και για το $\lambda_2 \cong 1.4852$ με τυπική απόκλιση 0.0347, αντίστοιχα.

Να αναφέρουμε ότι οι 49 από τις 94 περιοχές που εξετάσαμε κατατάχθηκαν στην πρώτη κατηγορία με $\lambda_1 \cong 0.7143$ *(Βορράς)* και οι υπόλοιπες στη δεύτερη με $\lambda_2 \cong 1.4852$ *(Νότος)*. Το συμπέρασμα αυτό προέκυψε από τον πίνακα των z που λάβαμε κατά την εφαρμογή του αλγορίθμου. Δηλαδή το ποσοστό κατανομής των z είναι 52.13% στη συνιστώσα του μίγματος με $\lambda_{z_i} = 1$, όπως και αναμενόταν από την κατανομή των περιοχών στον παραπάνω χάρτη 6^{β}.

Στο $\sigma\chi\acute{\eta}\mu\alpha$ 6.6 παρουσιάζεται το χρονόγραμμα για τον διαχωρισμό των λ στο δείγμα των 94 περιοχών, ενώ στο $\sigma\chi\acute{\eta}\mu\alpha$ 6.7 παρουσιάζεται το αντίστοιχο ιστόγραμμα των λ που κράτησε ο αλγόριθμος μας δείχνοντας σε ποια κατηγορία ανήκει το κάθε ένα από αυτά προσεγγιστικά.



Σχήμα 6.6: Το χρονόγραμμα από τις εκ των υστέρων τιμές των λ_1 (με μπλέ) και λ_2 (με πράσινο), αντίστοιχα.

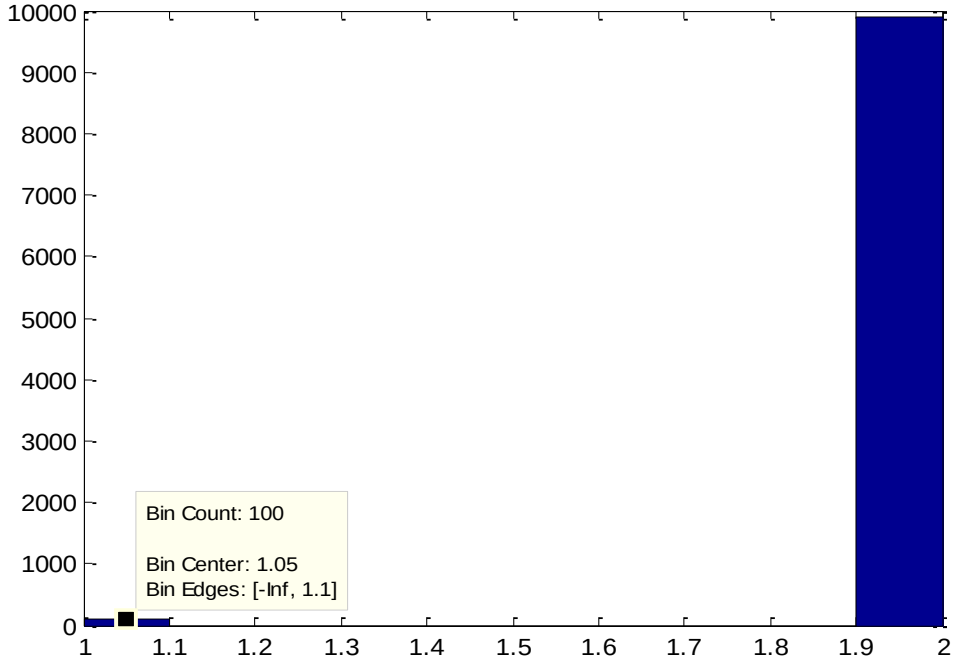


Σχήμα 6.7: Το ιστόγραμμα των λ_1 (μπλέ) και λ_2 (κόκκινο) που φανερώνει πως αυτά κατανέμονται στις μέσες τιμές σύγκλισης, είναι δηλαδή τα λ που κράτησε ο αλγόριθμος *(draws).*

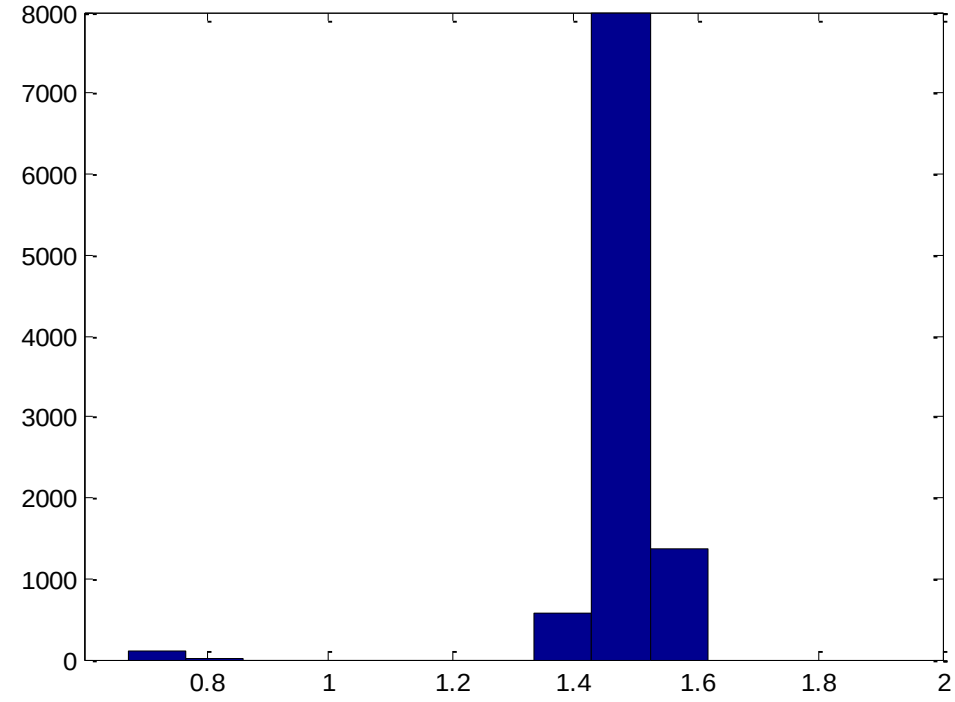
Σύμφωνα με τη διάταξη του χάρτη 6^{β}, όπου με σκούρο χρώμα απεικονίζονται οι περιοχές της Γαλλίας με υψηλό δείκτη επικινδυνότητας και με γκρι απόχρωση οι περιοχές με μικρότερο δείκτη επικινδυνότητας, μπορούμε να εξάγουμε συμπεράσματα για την κάθε περιοχή ξεχωριστά.

Για παράδειγμα, επιλέγουμε τυχαία την περιοχή 12, όπου βρίσκεται στο Νότιο άκρο της Γαλλίας και εξετάζουμε με τον παραπάνω αλγόριθμο σε τι βαθμό επικινδυνότητας μπορεί να καταταχθεί.

Για την περιοχή 12 *(Aveyron)* της Γαλλίας λοιπόν, λαμβάνουμε το παρακάτω ιστόγραμμα, σχήμα 6.8_a, όπου μπορούμε να συμπεράνουμε ότι από τις 10.000 φορές τρεξίματος του αλγορίθμου μας, μόνο τις 100 φορές κατατάχθηκε στη πρώτη ζώνη επικινδυνότητας $\lambda_1 \cong 0.7143$ *(Βορράς)* με τον χαμηλό συντελεστή. Αποτέλεσμα που μας δείχνει ότι η περιοχή αυτή (Aveyron) παρουσιάζει γενικά υψηλό συντελεστή θνησιμότητας.



Σχήμα 6.8$_a$: Ιστόγραμμα της περιοχής 12 (Aveyron) της Γαλλίας που φαίνεται ότι ο λόγος θνησιμότητας είναι υψηλός, τα δεδομένα κατατάσσονται στη δεύτερη κατηγορία του *z* (δηλ. στο λ_2)

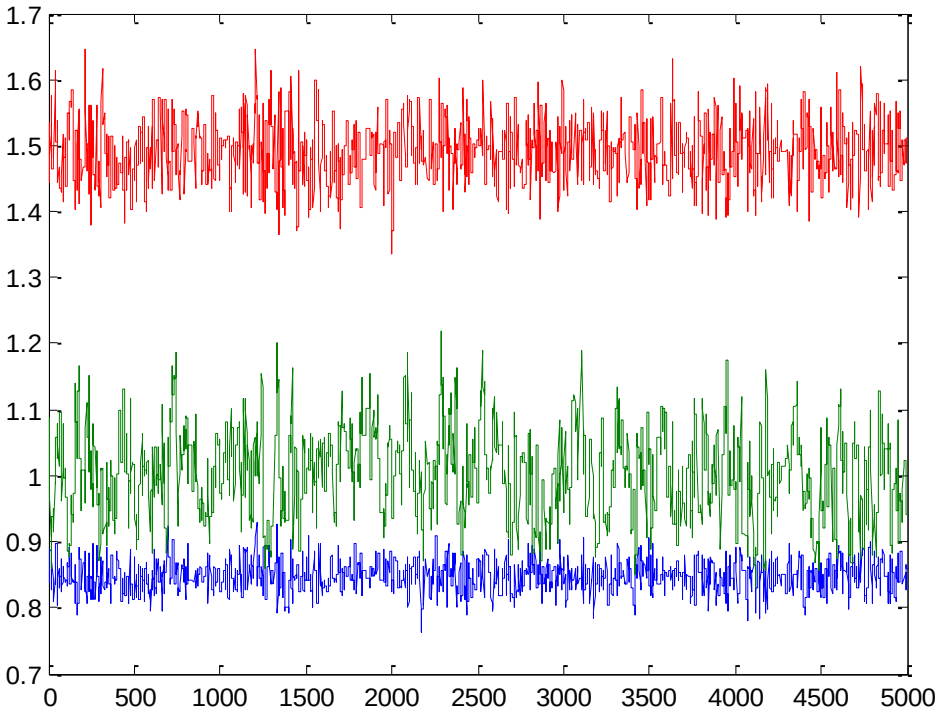


Σχήμα 6.8$_b$: Ιστόγραμμα της περιοχής 12 (Aveyron) της Γαλλίας που φαίνεται πως κατανέμεται το $\lambda_{z_{12}}$, δείχνοντας και με αυτό τον τρόπο ότι ο λόγος θνησιμότητας είναι υψηλός.

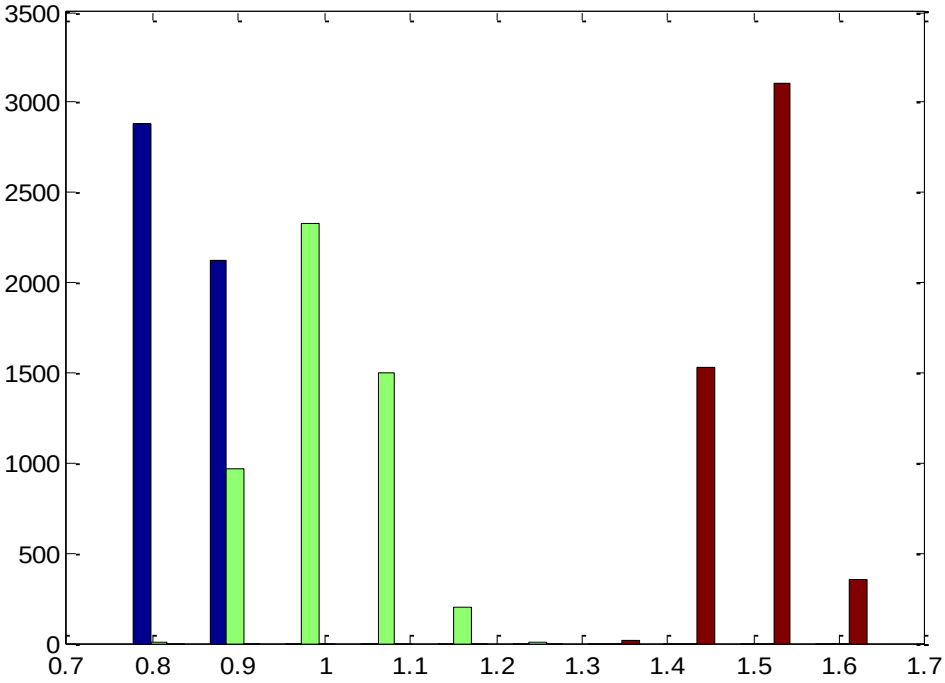
Γ) Τέλος, χωρίζουμε το πεδίο εφαρμογής της έρευνάς μας σε τρεις ζώνες, θεωρούμε δηλαδή το $k = 3$ και λαμβάνουμε τα ακόλουθα αποτελέσματα: παίρνουμε μία εκ των υστέρων μέση τιμή για το $\lambda_1 \cong 0.8479$ με τυπική απόκλιση 0.0222, για το $\lambda_2 \cong 1.0023$ με τυπική απόκλιση 0.0652 και $\lambda_3 \cong 1.4912$ με τυπική απόκλιση 0.0436, αντίστοιχα.

Να αναφέρουμε ότι για τον αλγόριθμο προσομοίωσης, οι 53 από τις 94 περιοχές που εξετάσαμε κατατάχθηκαν στην πρώτη κατηγορία με $\lambda_1 \cong 0.8479$ *(Βορράς),* οι 10 από τις 94 στη δεύτερη κατηγορία $\lambda_2 \cong 1.0023$ *(Κεντρικό τμήμα)* και οι υπόλοιπες στην τρίτη με $\lambda_3 \cong 1.4912$ *(Νότος)*. Το συμπέρασμα αυτό προέκυψε από τον πίνακα των z που λάβαμε κατά την εφαρμογή του αλγορίθμου.

Στο $\sigma\chi\acute{\eta}\mu\alpha$ 6.9 παρουσιάζεται το χρονόγραμμα για τον διαχωρισμό των λ στο δείγμα των 94 περιοχών, ενώ στο $\sigma\chi\acute{\eta}\mu\alpha$ 6.10 παρουσιάζεται το αντίστοιχο ιστόγραμμα των λ που κράτησε ο αλγόριθμος μας δείχνοντας σε ποια κατηγορία ανήκει το κάθε ένα από αυτά προσεγγιστικά.



Σχήμα 6.9: Το χρονόγραμμα από τις εκ των υστέρων τιμές των λ_1 (με μπλέ) και λ_2 (με πράσινο) και λ_3 (με κόκκινο), αντίστοιχα

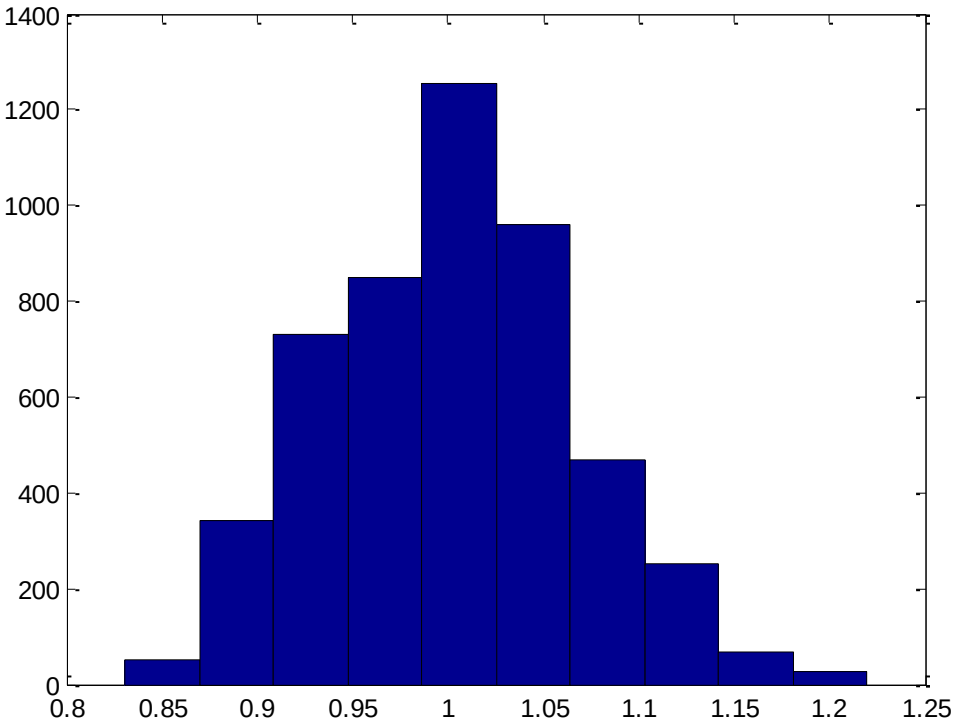


Σχήμα 6.10: Το ιστόγραμμα των λ_1 (μπλέ), λ_2 (πράσινο) και λ_3 (κόκκινο) που φανερώνει πως αυτά κατανέμονται στις μέσες τιμές σύγκλισης, είναι δηλαδή τα λ που κράτησε ο αλγόριθμος *(draws).*

Σύμφωνα με τη διάταξη του χάρτη που κατασκευάσαμε εμείς, δηλαδή σε τρία οριζόντια επίπεδα, όπου με το πιο σκούρο χρώμα απεικονίζονται οι περιοχές της Γαλλίας με υψηλό δείκτη επικινδυνότητας και όσο πάμε προς το Βορρά με γκρι απόχρωση οι περιοχές με μικρότερο δείκτη επικινδυνότητας, μπορούμε να εξάγουμε συμπεράσματα για την κάθε περιοχή ξεχωριστά.

Για παράδειγμα, επιλέγουμε τυχαία την περιοχή 37, όπου βρίσκεται στο Κεντρικό τμήμα της Γαλλίας και εξετάζουμε με τον παραπάνω αλγόριθμο σε τι βαθμό επικινδυνότητας μπορεί να καταταχθεί.

Για την περιοχή 37 *(Indre - et - Loir)* της Γαλλίας λοιπόν, λαμβάνουμε το παρακάτω ιστόγραμμα, σχήμα 6.11, όπου μπορούμε να συμπεράνουμε ότι από τις 10.000 φορές τρεξίματος του αλγορίθμου μας, κατατάχθηκε στη δεύτερη ζώνη επικινδυνότητας $\lambda_2 \cong 1.0023$ *(Κεντρικό τμήμα)* με τον μεσαίο συντελεστή. Αποτέλεσμα που μας δείχνει ότι η περιοχή αυτή *(Indre - et - Loir)* παρουσιάζει γενικά έναν μέσο συντελεστή θνησιμότητας ($2^{η}$ κατηγορία).



Σχήμα 6.11: Ιστόγραμμα της περιοχής 37 *(Indre - et - Loir)* της Γαλλίας που φαίνεται πως κατανέμεται το $\lambda_{z_{37}}$, δείχνοντας και με αυτό τον τρόπο ότι ο λόγος θνησιμότητας εμφανίζεται στη δεύτερη κατηγορία.

6.3 Μοντέλα εφαρμογής σε πραγματικά δεδομένα

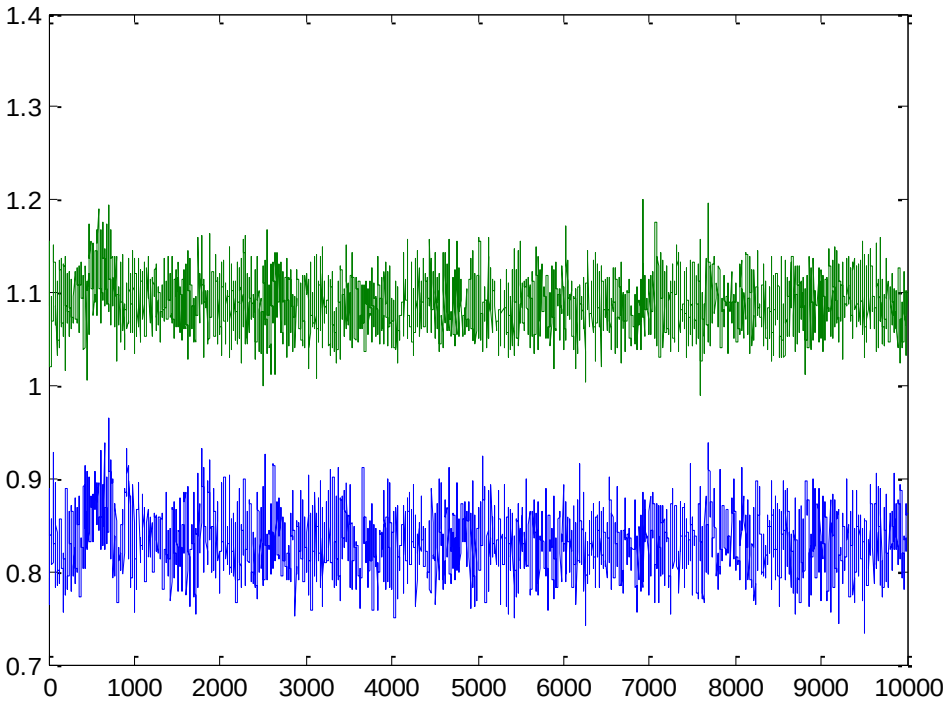
6.3.1 Εφαρμογή 1η

Στη μελέτη μας αυτή εξετάζουμε τον αλγόριθμο και τις προβλέψεις που μπορούμε να πάρουμε από αυτόν σε πραγματικά δεδομένα. Συγκεκριμένα, τα δεδομένα μας, προέρχονται από παρατηρήσεις – κρούσματα θανάτων από καρκίνο της στοματικής κοιλότητας και η ερευνά μας εστιάζεται στην ευρύτερη περιοχή της Γαλλίας.

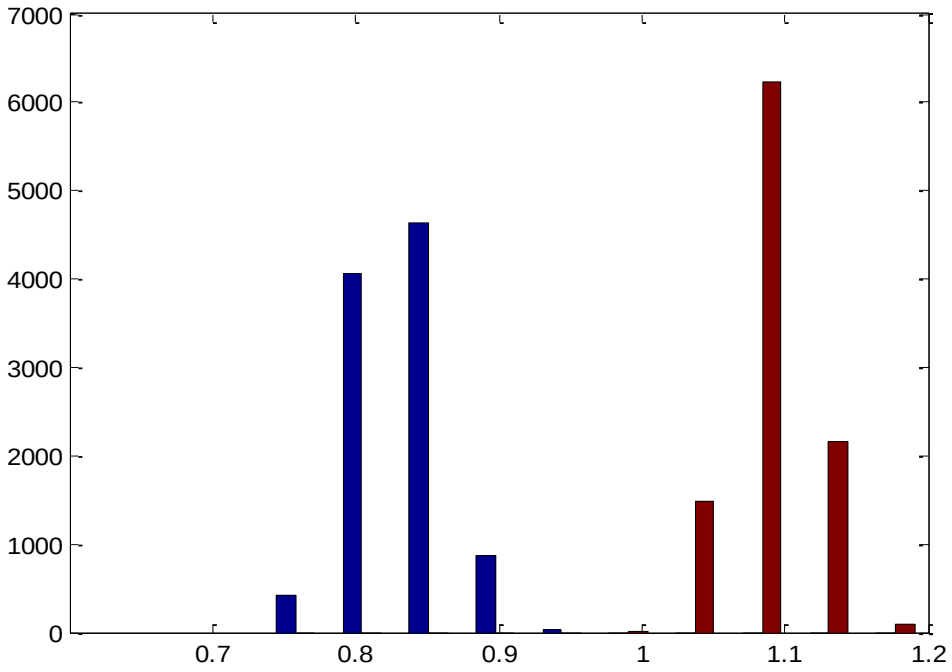
Τα αποτελέσματά μας αντιστοιχούν σε ένα τρέξιμο 100.000 'κύκλων' του αλγορίθμου μας όταν το $k = 2$ (όπου k το πλήθος των κατηγοριών που χωρίζουμε το δείγμα μας, ανάλογα με το βαθμό επικινδυνότητας). Η διαδικασία που ακολουθήσαμε είναι ίδια όπως περιγράψαμε και στην παράγραφο 6.2.

Μετά από έναν ικανό αριθμό επαναλήψεων παίρνουμε μία εκ των υστέρων μέση τιμή για το $\lambda_1 \cong 0.8325$ και για το $\lambda_2 \cong 1.0877$, αντίστοιχα. Διαγραμματικά το συμπέρασμα αυτό παρουσιάζεται στο σχήμα 6.12, ενώ στο *σχήμα* 6.13 παρουσιάζεται το αντίστοιχο ιστόγραμμα των λ που κράτησε ο αλγόριθμος μας δείχνοντας σε ποια κατηγορία ανήκει το κάθε ένα από αυτά προσεγγιστικά.

Ενώ στο τέλος της παραγράφου 6.3 παρατίθεται και ο σχετικός πίνακας με τις τυπικές αποκλίσεις των $\lambda_i, i = 1,2$, των 94 περιοχών, όταν $k = 2$.



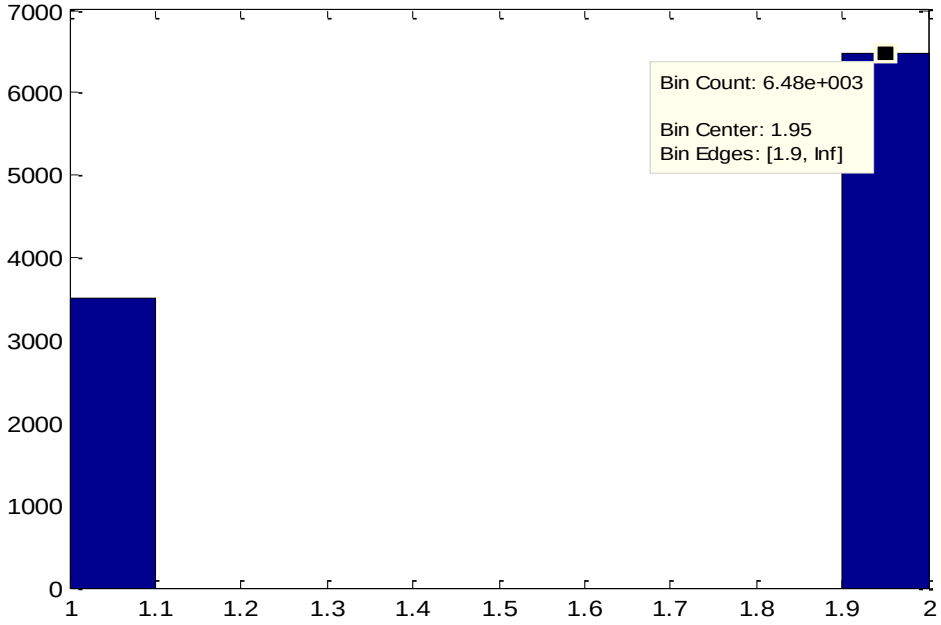
Σχήμα 6.12: Το χρονόγραμμα από τις εκ των υστέρων τιμές των λ_1 (με μπλέ) και λ_2 (με πράσινο), αντίστοιχα



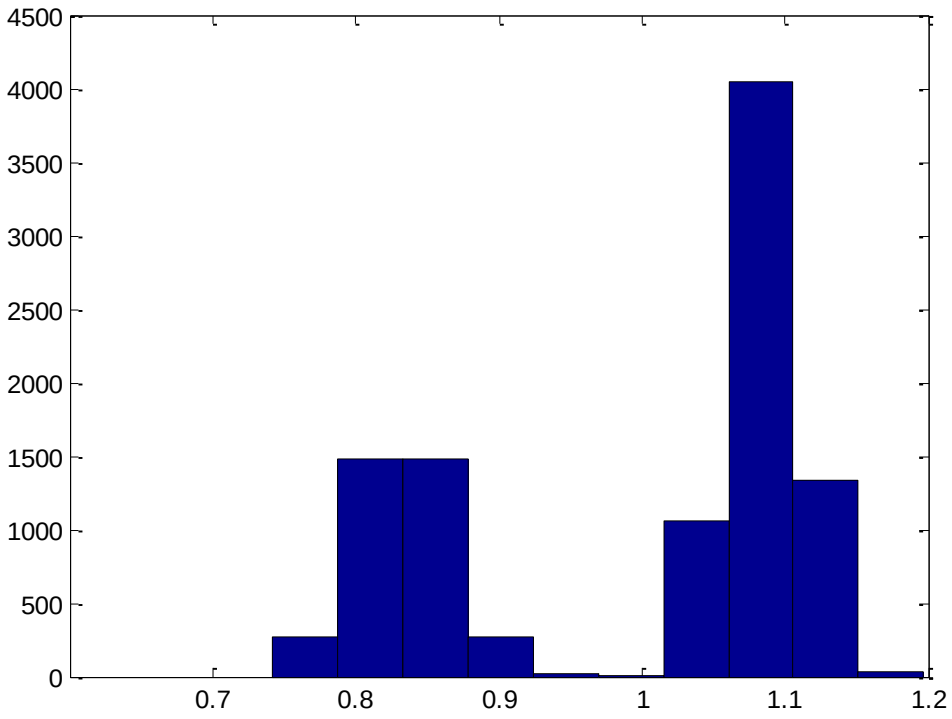
Σχήμα 6.13: Το ιστόγραμμα των λ_1 (μπλέ) και λ_2 (κόκκινο) που φανερώνει πως αυτά κατανέμονται στις εκ των υστέρων μέσες τιμές, είναι δηλαδή τα λ που κράτησε ο αλγόριθμος *(draws).*

Εξετάζουμε τώρα για ορισμένες περιοχές της Γαλλίας και προσπαθούμε να προβλέψουμε, με βάση τα παραπάνω στοιχεία, σε τι βαθμό επικινδυνότητας μπορούμε να την κατατάξουμε. Επιλέγουμε τυχαία την περιοχή 37 *(Indre - et - Loir)* της Γαλλίας και λαμβάνουμε το παρακάτω ιστόγραμμα, σχήμα 6.14_a, όπου μπορούμε να συμπεράνουμε ότι από τις 10.000 φορές τρεξίματος του αλγορίθμου μας, τις 6480 φορές (64.8%) κατατάχθηκε στη δεύτερη κατηγορία επικινδυνότητας, $(\lambda_2 \cong 1.0877)$ με τον υψηλό συντελεστή. Αποτέλεσμα που μας δείχνει ότι η περιοχή αυτή παρουσιάζει γενικά υψηλό συντελεστή θνησιμότητας.

- ***Περιοχή Indre - et - Loir***



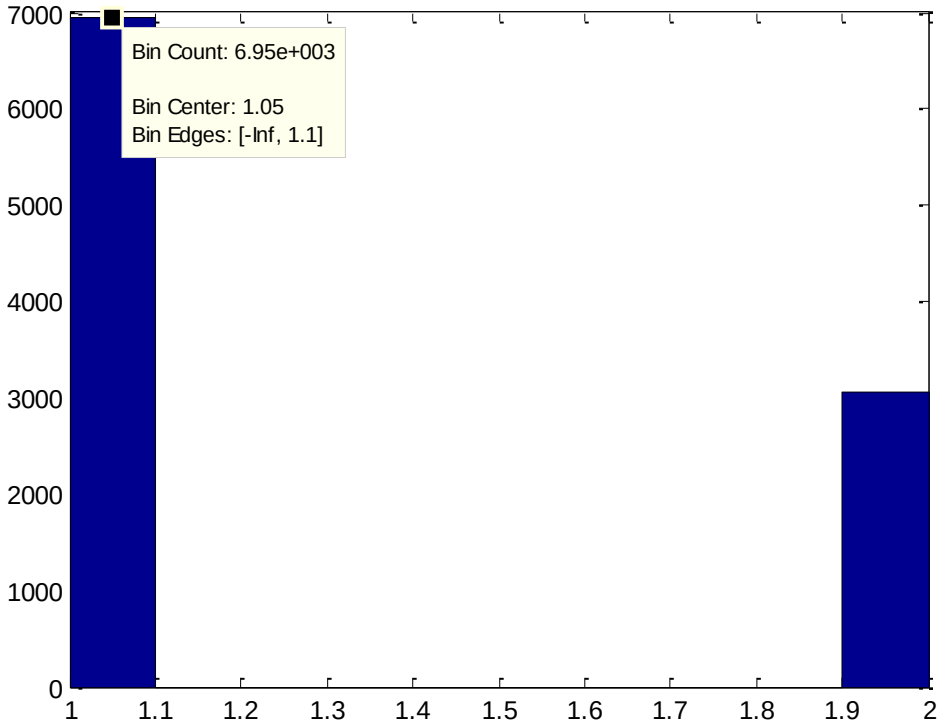
Σχήμα 6.14$_a$: Ιστόγραμμα της περιοχής 37 *(Indre - et - Loir)* της Γαλλίας, που φαίνεται ότι ο λόγος θνησιμότητας είναι υψηλός, τα δεδομένα κατατάσσονται στη δεύτερη κατηγορία του *z* (δηλ. στο λ_2)



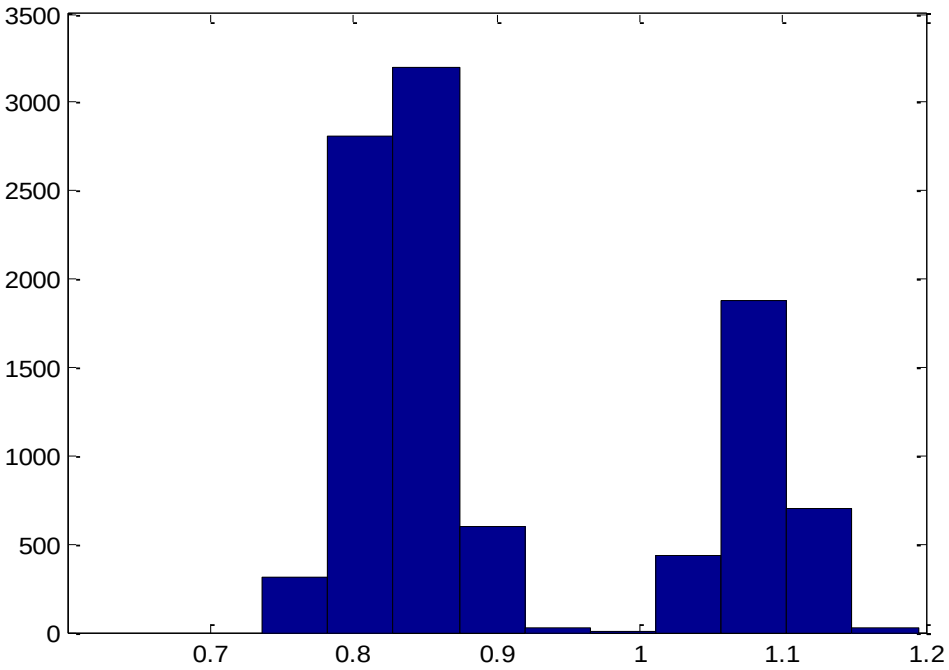
Σχήμα 6.14$_b$: Ιστόγραμμα της περιοχής 37 *(Indre - et - Loir)* της Γαλλίας που φαίνεται πως κατανέμεται το $\lambda_{z_{37}}$, δείχνοντας και με αυτό τον τρόπο ότι ο λόγος θνησιμότητας είναι υψηλός. Το τυπικό σφάλμα για το $\lambda_{z_{37}}$ είναι 0.0263.

Στη συνέχεια παρουσιάζουμε και άλλα ιστογράμματα, όπου εύκολα μπορούμε να διαπιστώσουμε για την κάθε περιοχή ξεχωριστά σε ποια κατηγορία επικινδυνότητας κατατάσσεται. Οι περιοχές που εξετάζουμε και παρουσιάζουμε είναι, αριθμητικά (η αντιστοιχία των αριθμών με τις ονομασίες των περιοχών μπορεί να γίνει από τον σχετικό πίνακα 6^{δ}), οι 12, 41, 30, 22.

- ***Περιοχή Aveyron***

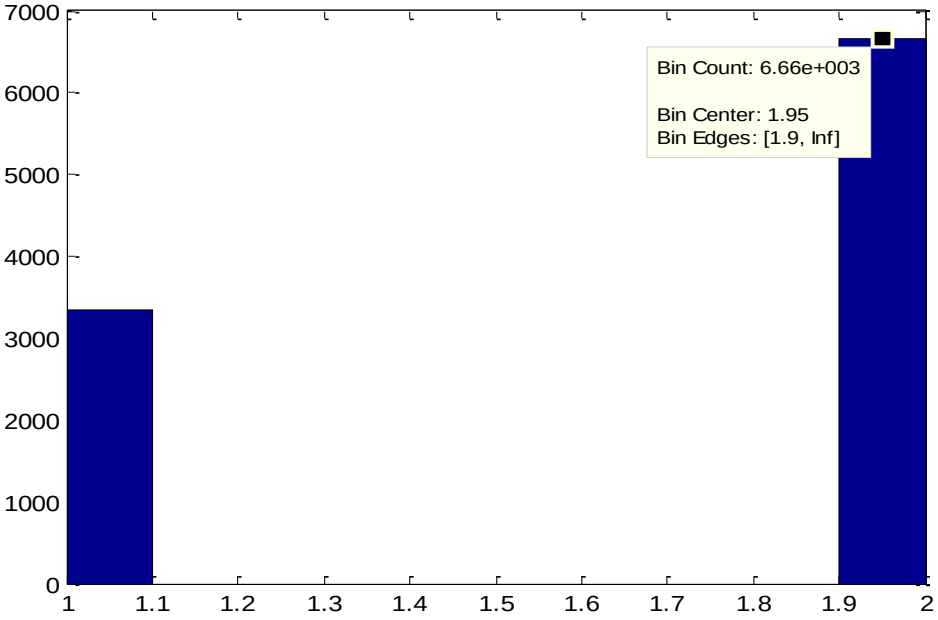


Σχήμα 6.15$_{\alpha}$: Ιστόγραμμα της περιοχής 12 *(Aveyron)* της Γαλλίας, που φαίνεται ότι ο λόγος θνησιμότητας είναι χαμηλός, τα δεδομένα κατατάσσονται στην πρώτη κατηγορία του *z* (δηλ. στο λ_1) σε ποσοστό 69.5%

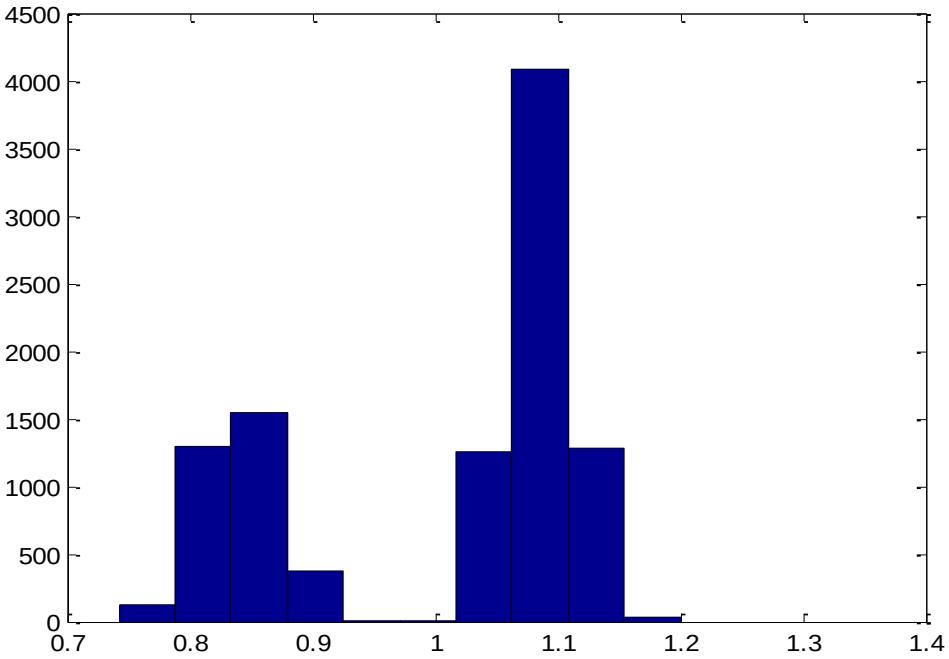


Σχήμα 6.15$_b$: Ιστόγραμμα της περιοχής 12 *(Aveyron)* της Γαλλίας που φαίνεται πως κατανέμεται το $\lambda_{z_{12}}$, δείχνοντας και με αυτό τον τρόπο ότι ο λόγος θνησιμότητας είναι χαμηλός. Το τυπικό σφάλμα για το $\lambda_{z_{12}}$ είναι 0.1200.

- ***Περιοχή Loir - et - Cher***

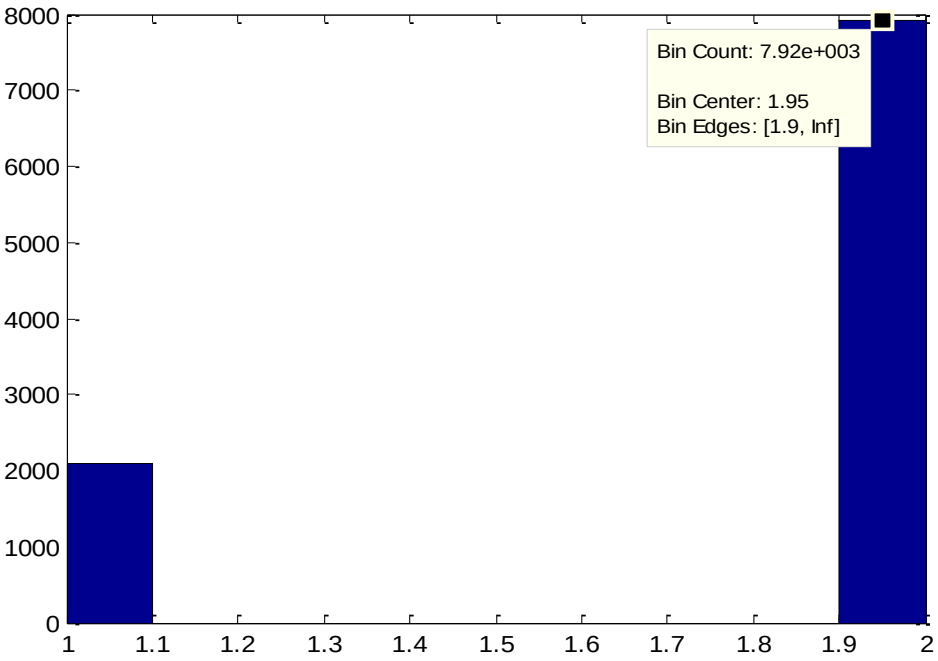


Σχήμα 6.16$_a$: Ιστόγραμμα της περιοχής 41 (*Loir – et – Cher)* της Γαλλίας, που φαίνεται ότι ο λόγος θνησιμότητας είναι υψηλός, τα δεδομένα κατατάσσονται στη δεύτερη κατηγορία του *z* (δηλ. στο λ_2) σε ποσοστό 66.6%.

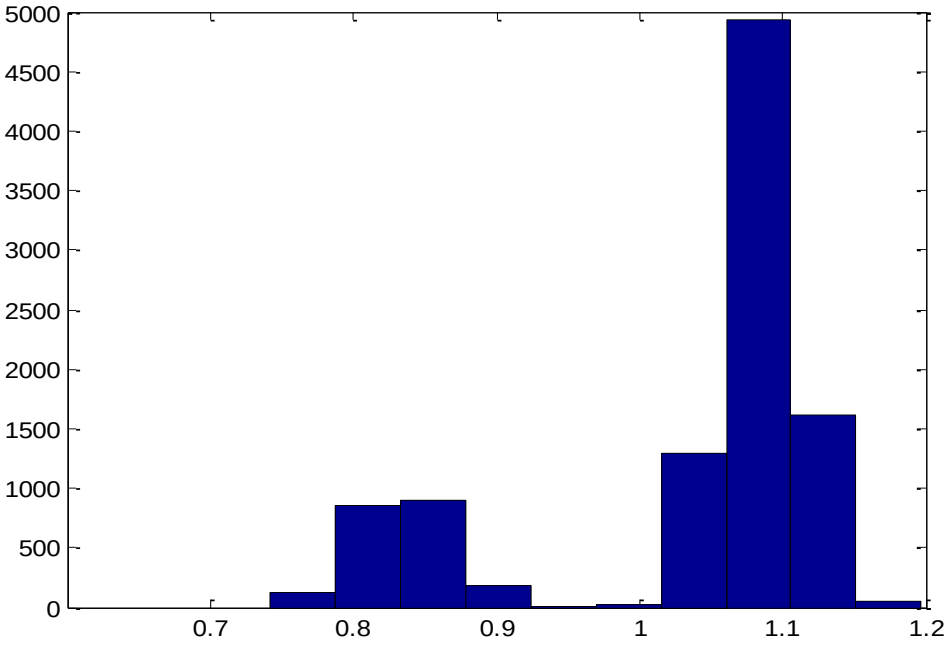


Σχήμα 6.16$_b$: Ιστόγραμμα της περιοχής 41 *(Loir – et – Cher)* της Γαλλίας που φαίνεται πως κατανέμεται το $\lambda_{z_{41}}$, δείχνοντας και με αυτό τον τρόπο ότι ο λόγος θνησιμότητας είναι υψηλός. Το τυπικό σφάλμα για το $\lambda_{z_{41}}$ είναι 0.0417.

- ***Περιοχή Gard***

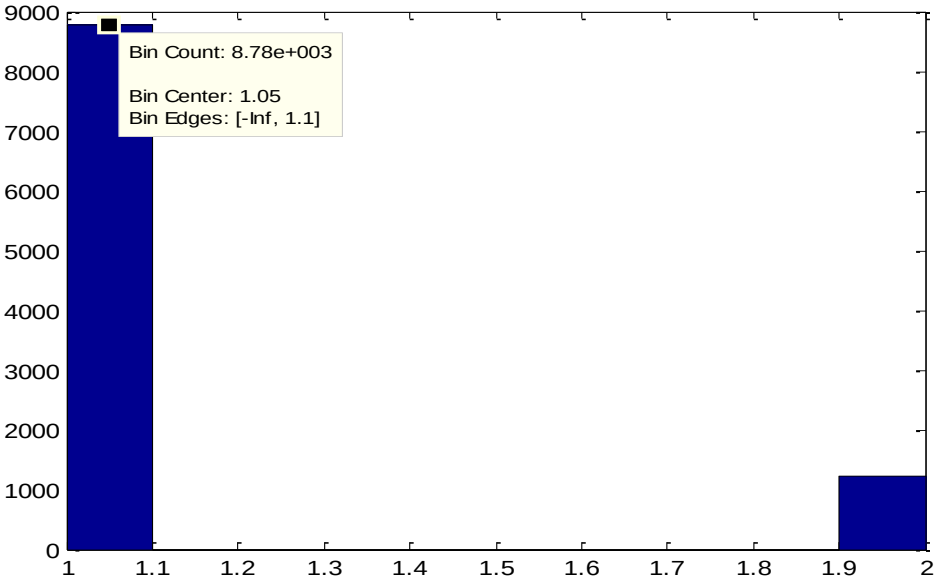


Σχήμα 6.17$_a$: Ιστόγραμμα της περιοχής 30 *(Gard)* της Γαλλίας, που φαίνεται ότι ο λόγος θνησιμότητας είναι υψηλός, τα δεδομένα κατατάσσονται στη δεύτερη κατηγορία του *z* (δηλ. στο λ_2) σε ποσοστό 79.2%

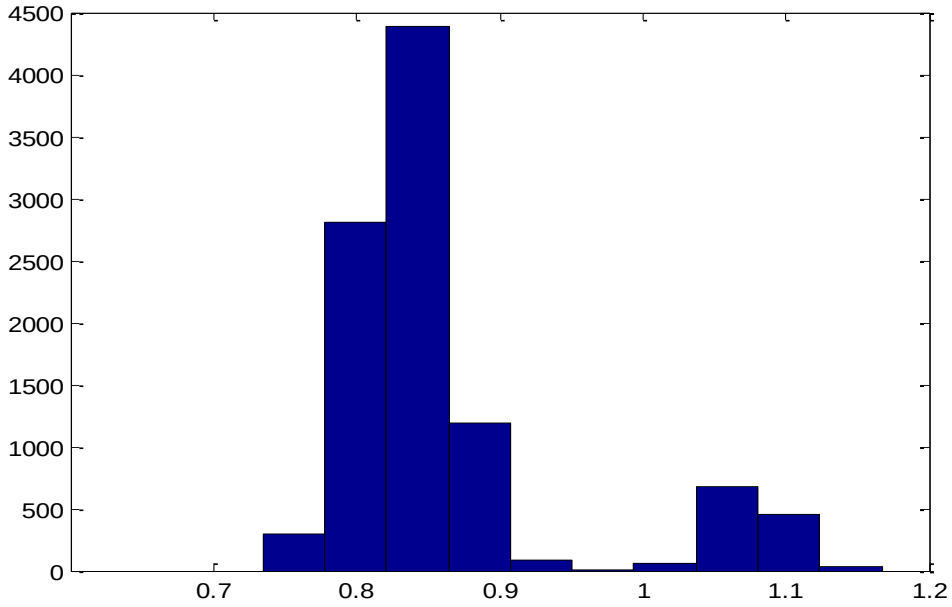


Σχήμα 6.17$_b$: Ιστόγραμμα της περιοχής 30 *(Gard)* της Γαλλίας που φαίνεται πως κατανέμεται το $\lambda_{z_{30}}$, δείχνοντας και με αυτό τον τρόπο ότι ο λόγος θνησιμότητας είναι υψηλός. Το τυπικό σφάλμα για το $\lambda_{z_{30}}$ είναι 0.0308.

- ***Περιοχή Cotes – d'Armor***



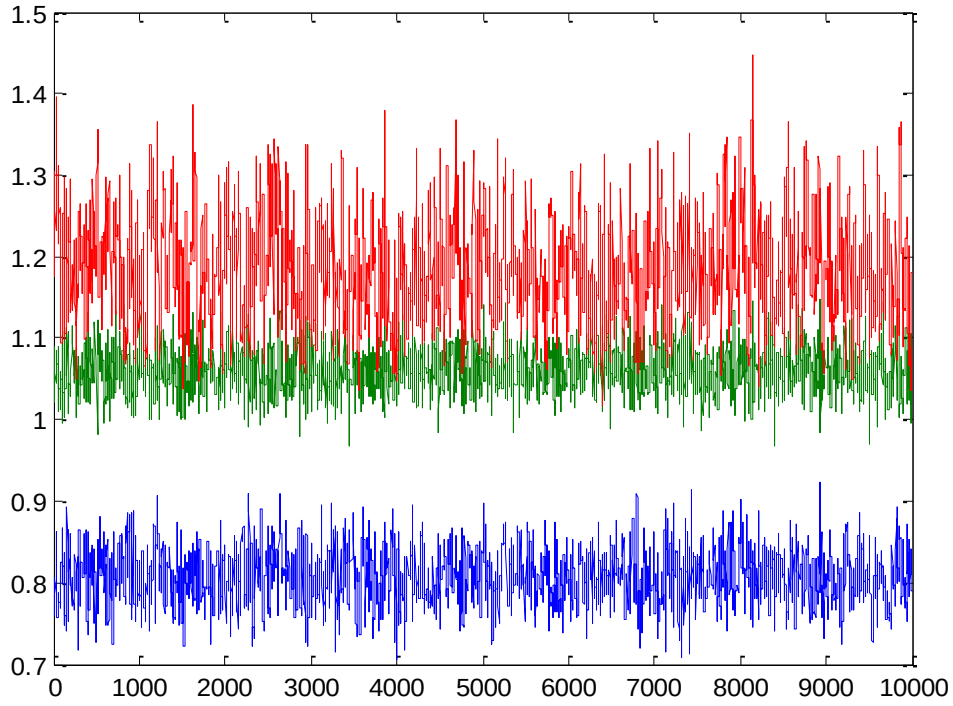
Σχήμα 6.18$_a$: Ιστόγραμμα της περιοχής 22 *(Cotes – d'Armor)* της Γαλλίας, που φαίνεται ότι ο λόγος θνησιμότητας είναι χαμηλός, τα δεδομένα κατατάσσονται στην πρώτη κατηγορία του *z* (δηλ. στο λ_1) σε ποσοστό 87.8%



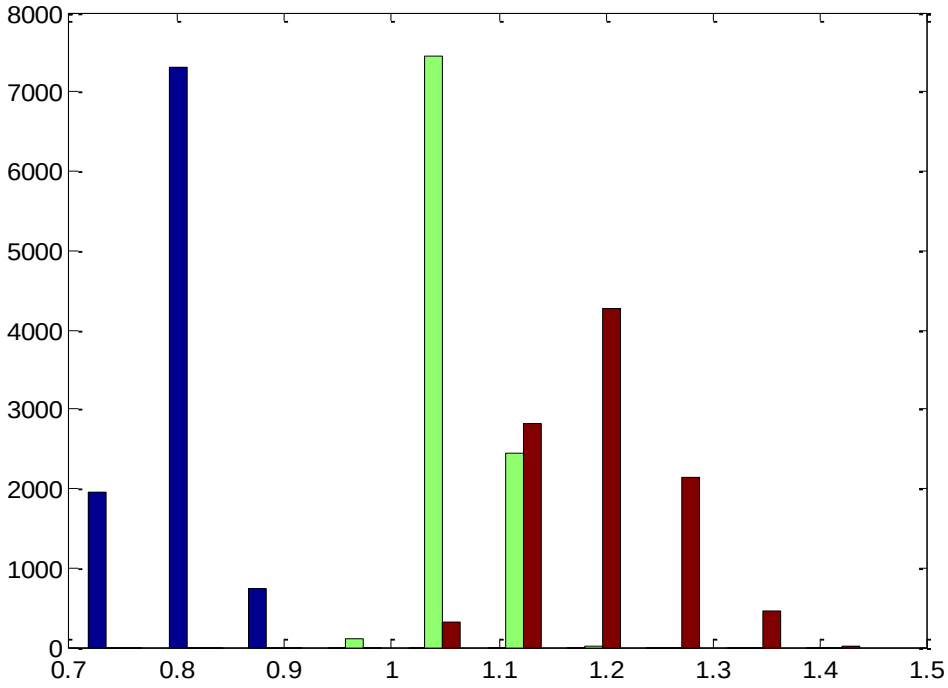
Σχήμα 6.18$_b$: Ιστόγραμμα της περιοχής 22 *(Cotes – et –d'Armor)* της Γαλλίας που φαίνεται πως κατανέμεται το $\lambda_{z_{22}}$, δείχνοντας και με αυτό τον τρόπο ότι ο λόγος θνησιμότητας είναι χαμηλός. Το τυπικό σφάλμα για το $\lambda_{z_{22}}$ είναι 0.0849.

Τα όσα συμπεράσματα παρουσιάσαμε παραπάνω, μπορούν πολύ εύκολα να διευρυνθούν αν αυξήσουμε τις κατηγορίες που κατατάσσουμε το λόγο θνησιμότητας. Εδώ εξετάζουμε την περίπτωση που το $k = 3$. Τα αποτελέσματά μας, αντιστοιχούν σε ένα τρέξιμο 100.000 'κύκλων' του αλγορίθμου μας. Μετά δηλαδή από έναν ικανό αριθμό επαναλήψεων παίρνουμε μία εκ των υστέρων μέση τιμή για τα $\lambda_1 \cong 0.8072$, $\lambda_2 \cong 1.0609$ και $\lambda_3 \cong 1.1854$ αντίστοιχα.

Διαγραμματικά το συμπέρασμα αυτό παρουσιάζεται στο *σχήμα 6.19*, ενώ στο *σχήμα* 6.20 παρουσιάζεται το αντίστοιχο ιστόγραμμα των λ που κράτησε ο αλγόριθμος μας δείχνοντας σε ποια κατηγορία ανήκει το κάθε ένα από αυτά προσεγγιστικά. Παρατηρούμε στο *σχήμα 6.19* ότι οι δύο κατηγορίες λ_2 και λ_3 βρίσκονται πολύ κοντά μεταξύ τους. Ενώ πάλι στο τέλος της παραγράφου 6.3 παρατίθεται και ο σχετικός πίνακας με τις τυπικές αποκλίσεις των $\lambda_i, i = 1,2,3$, των 94 περιοχών, όταν $k = 3$.



Σχήμα 6.19: Το χρονόγραμμα από τις εκ των υστέρων τιμές των λ_1 (με μπλέ) και λ_2 (με πράσινο) και λ_3 (με κόκκινο), αντίστοιχα



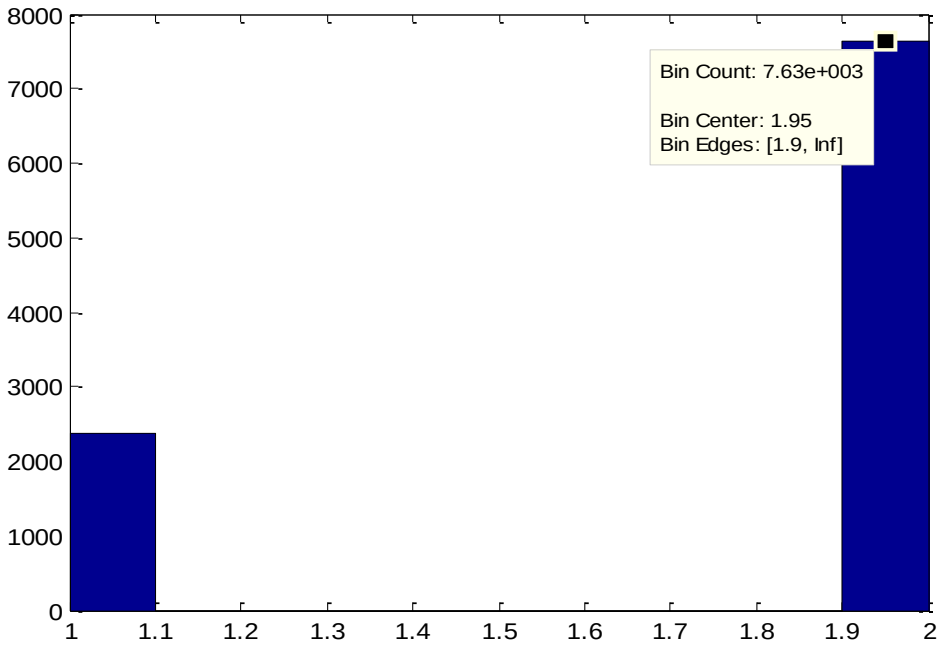
Σχήμα 6.20: Το ιστόγραμμα των λ_1 (μπλέ), λ_2 (πράσινο) και λ_3 (κόκκινο) που φανερώνει πως αυτά κατανέμονται στις εκ των υστέρων μέσες τιμές, είναι δηλαδή τα λ που κράτησε ο αλγόριθμος *(draws).*

Επιλέγουμε τυχαία την περιοχή 40 *(Landes)* της Γαλλίας και προσπαθούμε να προβλέψουμε, με βάση τα παραπάνω στοιχεία, σε τι βαθμό επικινδυνότητας μπορούμε να την κατατάξουμε όταν ο αριθμός κατηγοριών έχει αυξηθεί σε τρεις. Παρατηρούμε στο σχήμα 6.21_a ότι πάλι η περιοχή αυτή δεν παρουσιάζει καταμερισμό σε τρεις κατηγορίες αλλά σε δύο, πράγμα πολύ πιθανό.

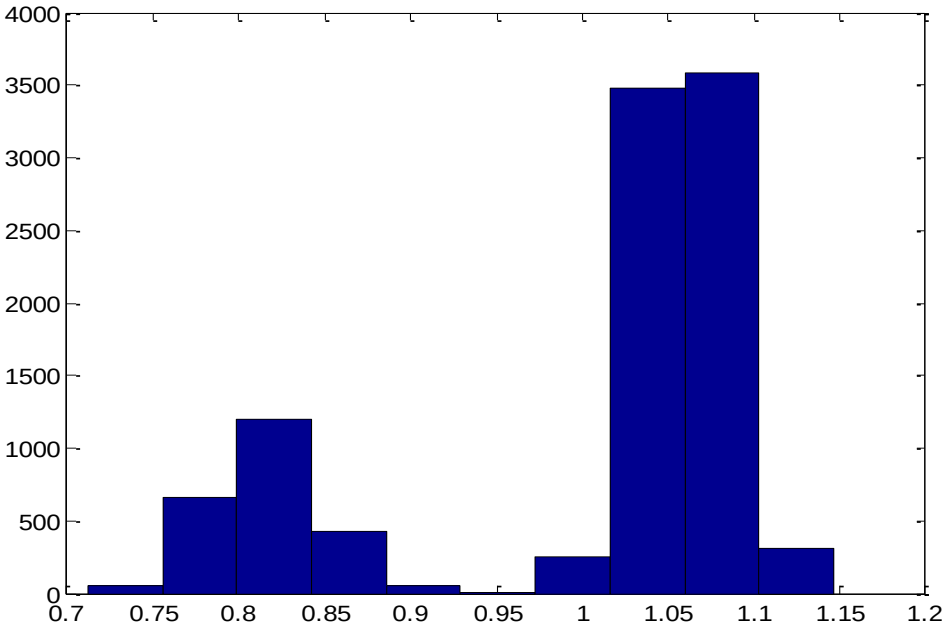
Αν όμως επιλέξουμε μία άλλη περιοχή, για παράδειγμα την 89 *(Yonne)*, σχήμα 6.22_a, παρατηρούμε ότι ο λόγος θνησιμότητας για τον καρκίνο της στοματικής κοιλότητας «μοιράζεται» και στις τρεις κατηγορίες, με μία τάση μετακίνησης στη $2^η$, σε ποσοστό 53.3%.

Αντίστοιχα τα ιστογράμματα 6.21_b και 6.22_b παρουσιάζουν την κατανομή των αντίστοιχων λ για τις περιοχές 40 και 89, ενώ από κάτω (στις λεζάντες των διαγραμμάτων) γράφουμε και τις αντίστοιχες τιμές των τυπικών σφαλμάτων.

- ***Περιοχή Landes***

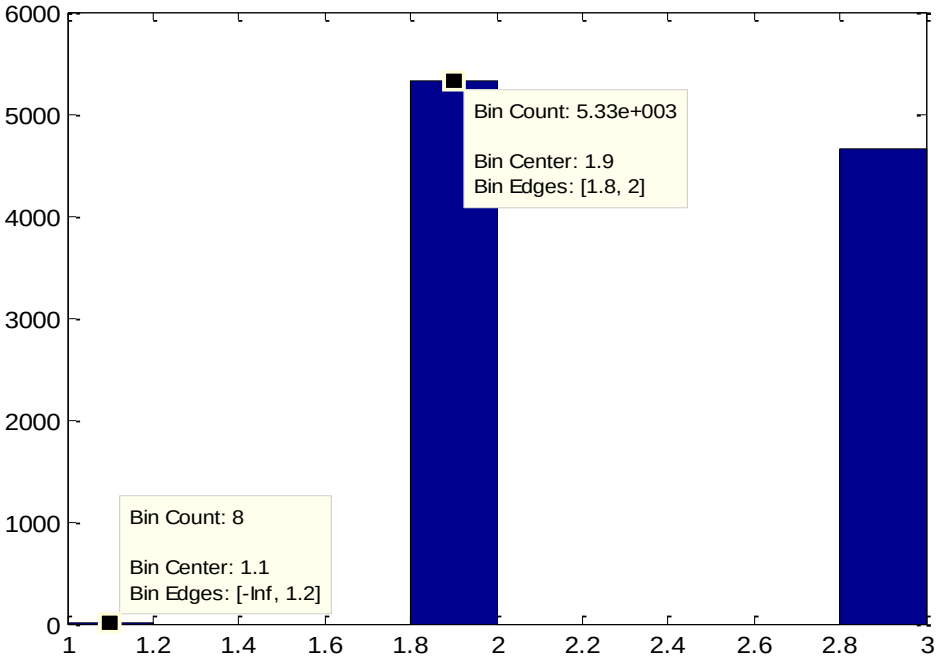


Σχήμα 6.21_a: Ιστόγραμμα της περιοχής 40 *(Landes)* της Γαλλίας, που φαίνεται ότι ο λόγος θνησιμότητας παρουσιάζει μία μέση επικινδυνότητα, τα δεδομένα κατατάσσονται στη δεύτερη κατηγορία του *z* (δηλ. στο λ_2) σε ποσοστό 76.3%

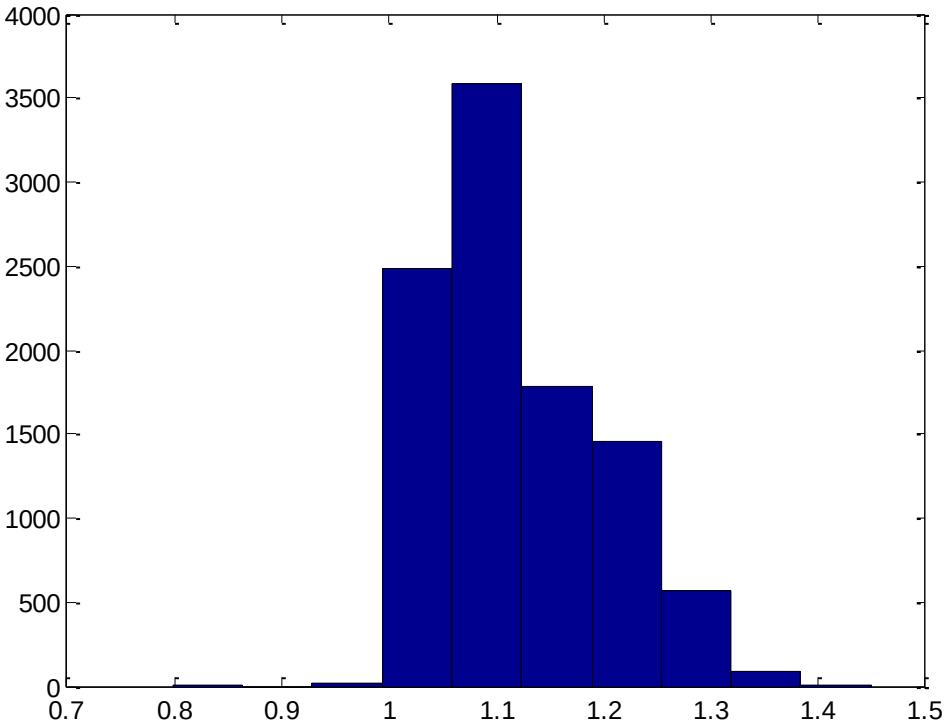


Σχήμα 6.21$_b$: Ιστόγραμμα της περιοχής 40 *(Landes)* της Γαλλίας που φαίνεται πως κατανέμεται το $\lambda_{z_{40}}$, δείχνοντας και με αυτό τον τρόπο ότι ο λόγος θνησιμότητας εντοπίζεται κυρίως στη δεύτερη κατηγορία. Το τυπικό σφάλμα για το $\lambda_{z_{40}}$ είναι 0.1069.

- ***Περιοχή Yonne***



Σχήμα 6.22$_a$: Ιστόγραμμα της περιοχής 89 *(Yonne)* της Γαλλίας, που φαίνεται ότι ο λόγος θνησιμότητας είναι σχετικά υψηλός, τα δεδομένα κατατάσσονται οριακά στη δεύτερη κατηγορία του z (δηλ. στο λ_2) σε ποσοστό 53.3%



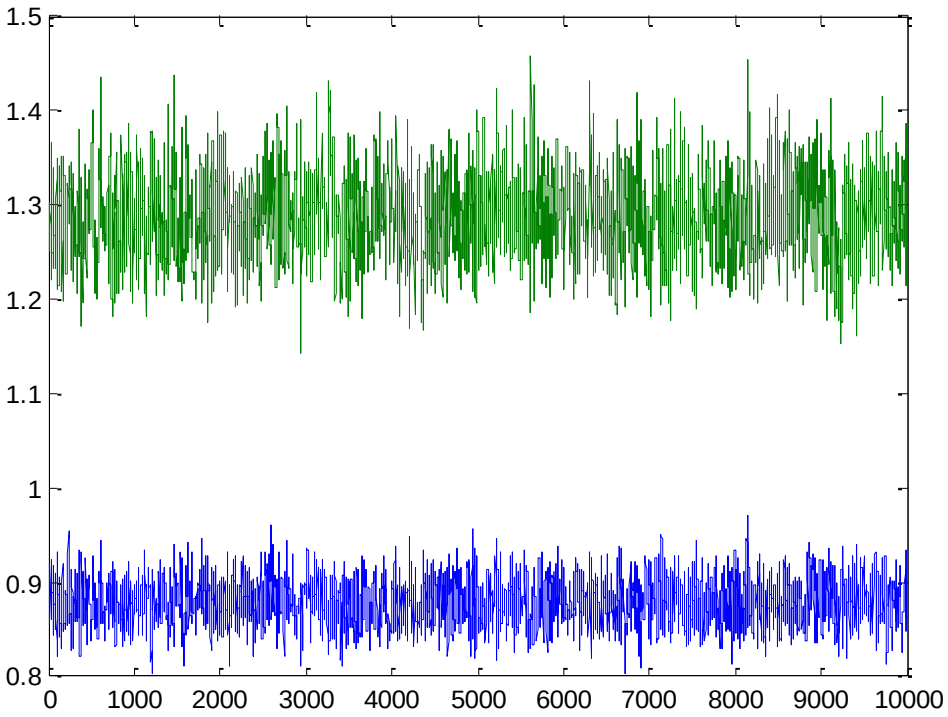
Σχήμα 6.22$_{\text{b}}$: Ιστόγραμμα της περιοχής 89 *(Yonne)* της Γαλλίας που φαίνεται πως κατανέμεται το $\lambda_{z_{89}}$, δείχνοντας και με αυτό τον τρόπο ότι ο λόγος θνησιμότητας είναι σχετικά υψηλός. Το τυπικό σφάλμα για το $\lambda_{z_{89}}$ είναι 0.0775.

6.3.2 Εφαρμογή 2^{η}

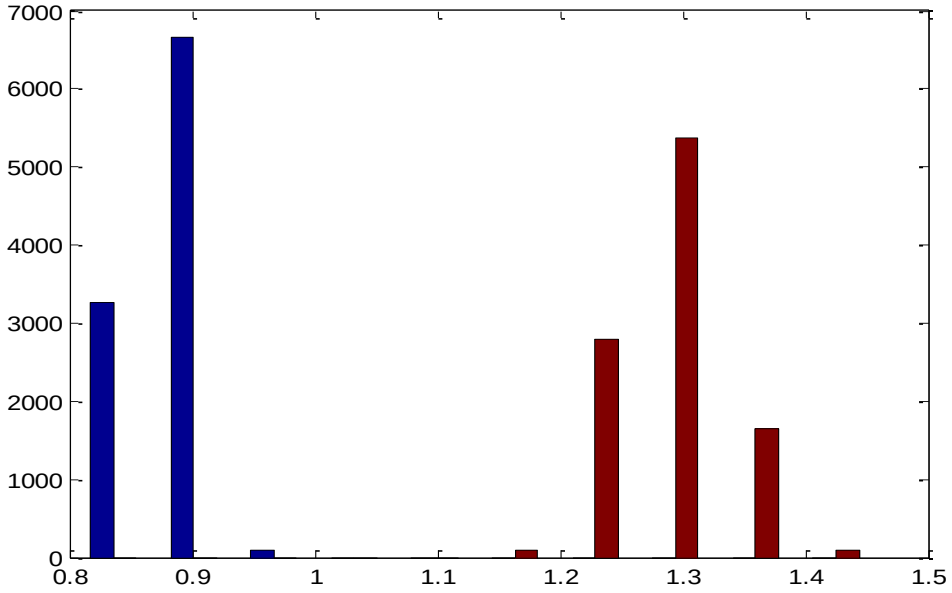
Στην περίπτωση αυτή εξετάζουμε τον αλγόριθμο και τις προβλέψεις που μπορούμε να πάρουμε από αυτόν πάλι σε πραγματικά δεδομένα. Συγκεκριμένα, αυτή τη φορά τα δεδομένα μας, προέρχονται από παρατηρήσεις – κρούσματα θανάτων από καρκίνο του φάρυγγα και η ερευνά μας εστιάζεται στην ευρύτερη περιοχή της Γαλλίας.

Τα αποτελέσματά μας αντιστοιχούν σε ένα τρέξιμο 100.000 'κύκλων' του αλγορίθμου μας με $k = 2$ (όπου k το πλήθος των κατηγοριών που χωρίζουμε το δείγμα μας, ανάλογα με το βαθμό επικινδυνότητας). Η διαδικασία που ακολουθήσαμε είναι η ίδια όπως έχουμε ήδη περιγράψει, για το λόγο αυτό θα αρκεστούμε στον έλεγχο που το $k = 2$, όπως προείπαμε.

Μετά από έναν ικανό αριθμό επαναλήψεων παίρνουμε μία εκ των υστέρων μέση τιμή για το $\lambda_1 \cong 0.8790$ και για το $\lambda_2 \cong 1.2865$, αντίστοιχα. Διαγραμματικά το συμπέρασμα αυτό παρουσιάζεται στο *σχήμα 6.23*, ενώ στο *σχήμα* 6.24 παρουσιάζεται το αντίστοιχο ιστόγραμμα των λ που κράτησε ο αλγόριθμος μας δείχνοντας σε ποια κατηγορία ανήκει το κάθε ένα από αυτά προσεγγιστικά. Μάλιστα στο τέλος της παραγράφου 6.3 παρατίθεται και ο σχετικός πίνακας με τις τυπικές αποκλίσεις των $\lambda_i, i = 1,2$, των 94 περιοχών, όταν $k = 2$.

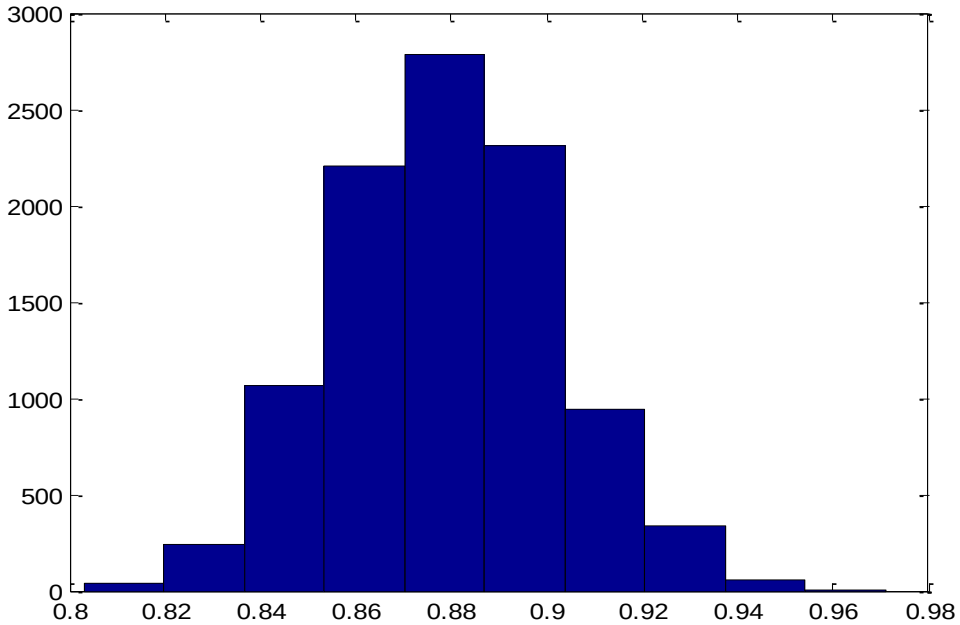


Σχήμα 6.23: Το χρονόγραμμα από τις εκ των υστέρων τιμές των λ_1 (με μπλέ) και λ_2 (με πράσινο), αντίστοιχα

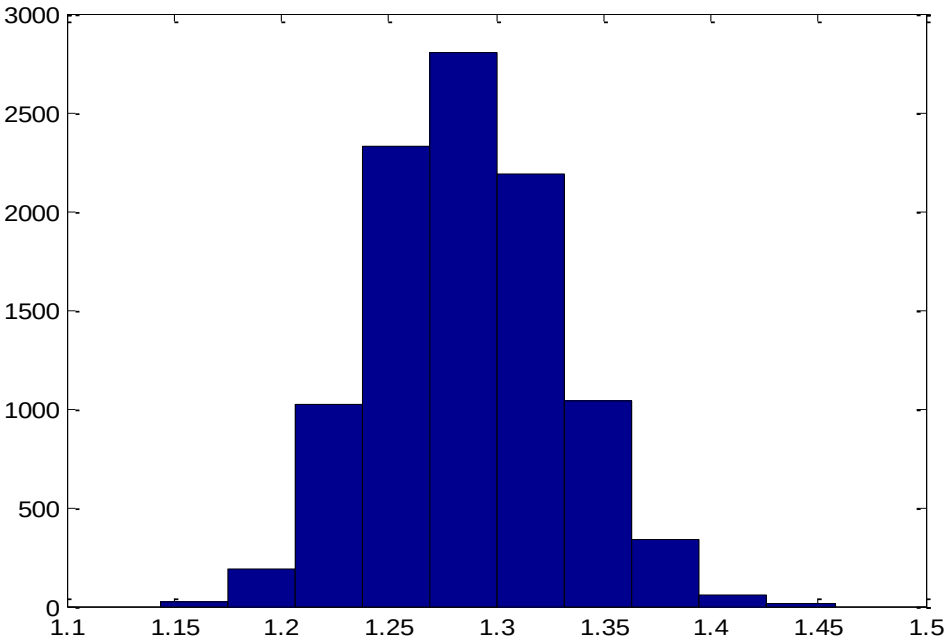


Σχήμα 6.24: Το ιστόγραμμα των λ_1 (μπλέ) και λ_2 (κόκκινο) που φανερώνει πως αυτά κατανέμονται στις εκ των υστέρων μέσες τιμές, είναι δηλαδή τα λ που κράτησε ο αλγόριθμος *(draws).*

Στα σχήματα 6.25_{α} και 6.25_{β}, αξίζει να παρατηρήσουμε με λεπτομέρεια πως τα λ_1 και λ_2 κατανέμονται στα ιστογράμματα, σε μεγαλύτερη ευκρίνεια.



Σχήμα 6.25_{α}: Ιστόγραμμα του δείγματος από την εκ των υστέρων κατανομή του λ_1



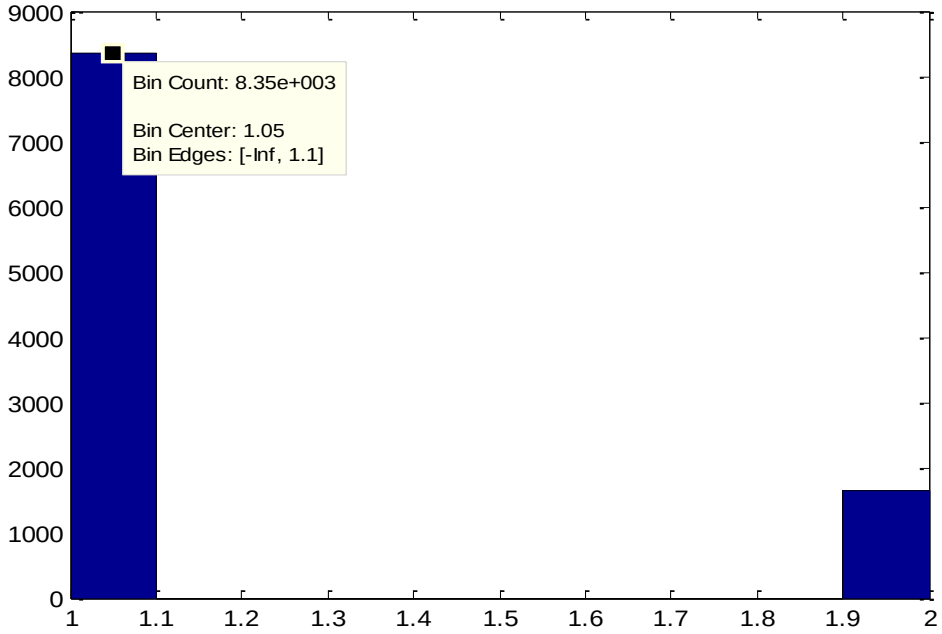
Σχήμα 6.25$_{\beta}$: Ιστόγραμμα του δείγματος από την εκ των υστέρων κατανομή του λ_2

Με βάση λοιπόν, τα παραπάνω συμπεράσματά μας, επιλέγουμε τυχαία μία περιοχή της Γαλλίας, ώστε να δούμε σε ποια από τις δύο κατηγορίες μπορούμε να την κατατάξουμε, με βάση πάλι τον βαθμό επικινδυνότητας για τον καρκίνο του φάρυγγα. Εξετάζουμε την περιοχή 12 *(Aveyron)* (σχήμα 6.26$_a$) και λαμβάνουμε ότι σε ποσοστό 83.5% τα δεδομένα μας κατανέμονται στην πρώτη κατηγορία (λ_1), με το χαμηλό δείκτη θνησιμότητας.

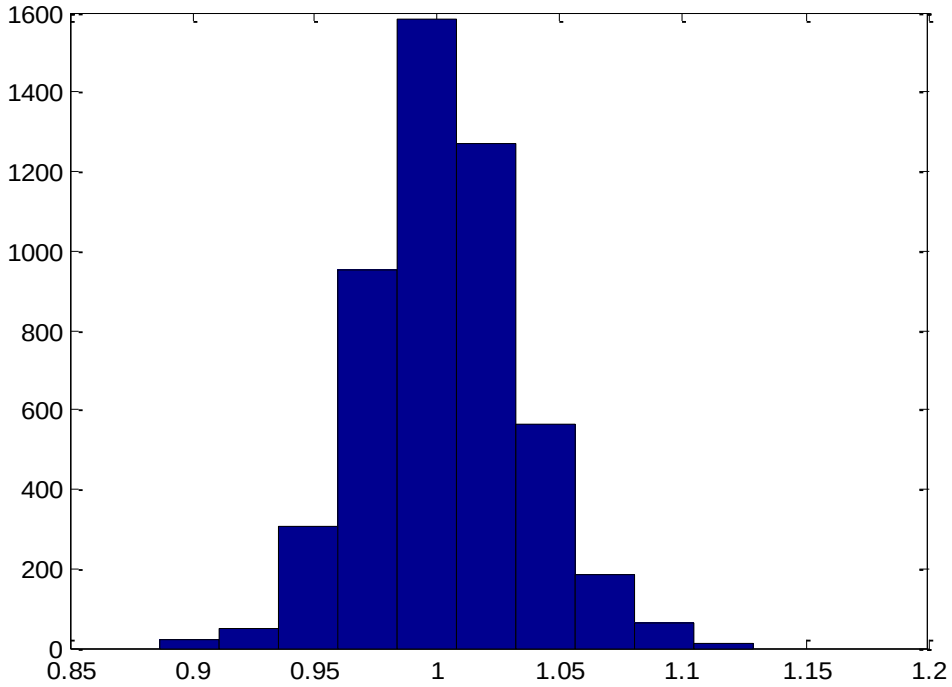
Αντίστοιχα αποτελέσματα, λαμβάνουμε και στην περιοχή 50 *(Manche)* της Γαλλίας, όπου και εκεί ο λόγος θνησιμότητας είναι χαμηλός, σχήμα 6.27$_a$. Ανάλογα συμπεράσματα μπορούμε να εξάγουμε και από τις 94 περιοχές της Γαλλίας.

Αντίστοιχα τα ιστογράμματα 6.26$_b$ και 6.27$_b$ παρουσιάζουν την κατανομή των αντίστοιχων λ για τις περιοχές 12 και 50, ενώ από κάτω (στις λεζάντες των διαγραμμάτων) γράφουμε και τις αντίστοιχες τιμές των τυπικών σφαλμάτων.

- ***Περιοχή Aveyron***

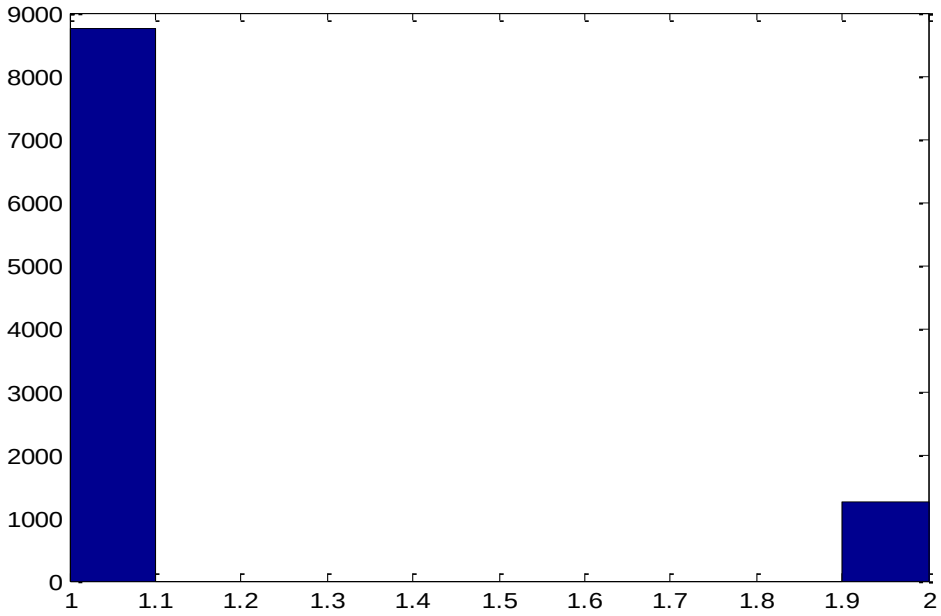


Σχήμα 6.26$_a$: Ιστόγραμμα της περιοχής 12 *(Aveyron)* της Γαλλίας, που φαίνεται ότι ο λόγος θνησιμότητας είναι χαμηλός, τα δεδομένα κατατάσσονται στην πρώτη κατηγορία του *z* (δηλ. στο λ_1) σε ποσοστό 83.5%

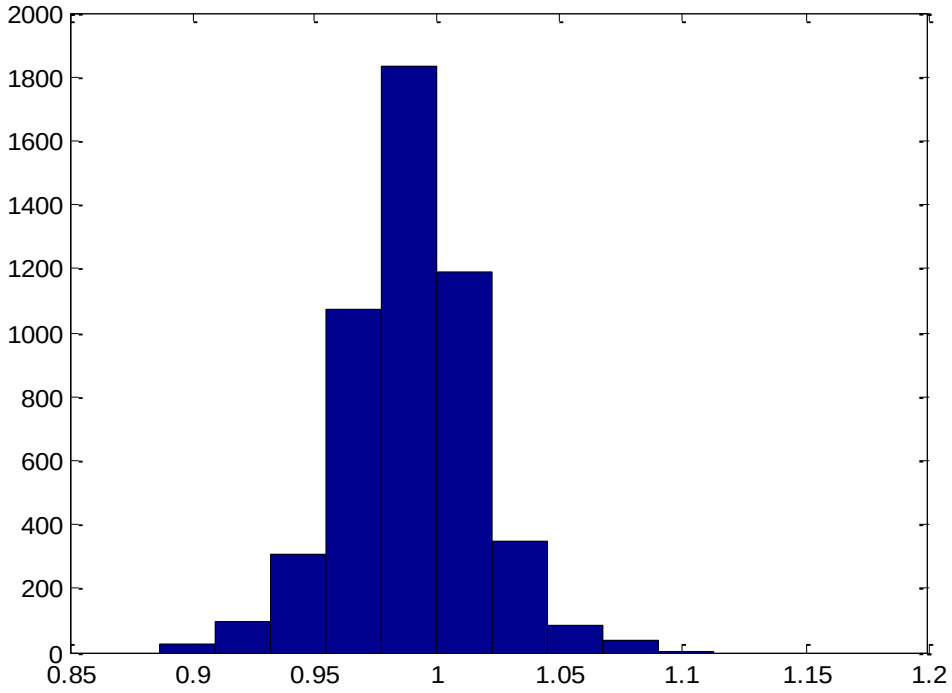


Σχήμα 6.26$_b$: Ιστόγραμμα της περιοχής 12 *(Aveyron)* της Γαλλίας που φαίνεται πως κατανέμεται το $\lambda_{z_{12}}$, δείχνοντας και με αυτό τον τρόπο ότι ο λόγος θνησιμότητας είναι σχετικά χαμηλός. Το τυπικό σφάλμα για το $\lambda_{z_{12}}$ είναι 0.0317.

- ***Περιοχή Manche***



Σχήμα 6.27ₐ: Ιστόγραμμα της περιοχής 50 *(Manche)* της Γαλλίας, που φαίνεται ότι ο λόγος θνησιμότητας είναι χαμηλός, τα δεδομένα κατατάσσονται στην πρώτη κατηγορία του *z* (δηλ. στο λ_1) σε ποσοστό 88.5%



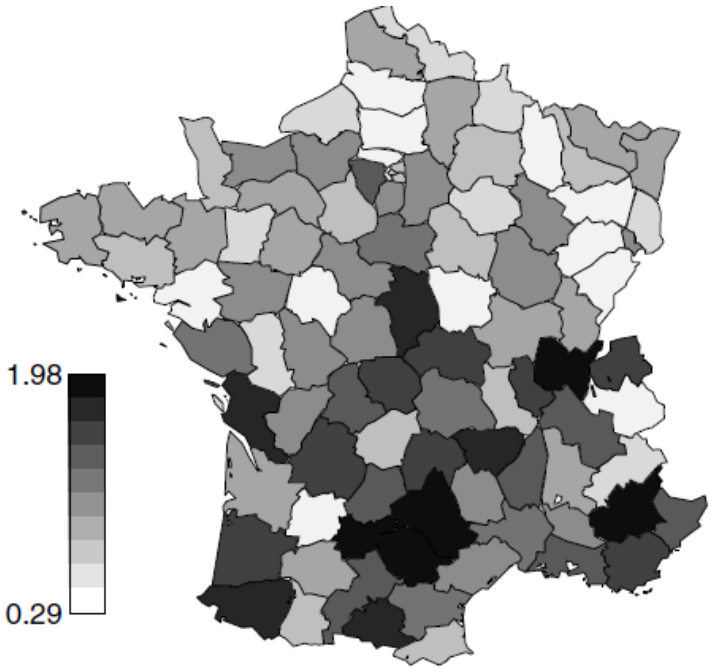
Σχήμα 6.27ᵦ: Ιστόγραμμα της περιοχής 50 *(Manche)* της Γαλλίας που φαίνεται πως κατανέμεται το $\lambda_{z_{50}}$, δείχνοντας και με αυτό τον τρόπο ότι ο λόγος θνησιμότητας είναι σχετικά χαμηλός. Το τυπικό σφάλμα για το $\lambda_{z_{50}}$ είναι 0.0271.

- Πίνακας με τις ***εκ των υστέρων μέσες τιμές & τυπικές αποκλίσεις*** των λ για τις 94 περιοχές της Γαλλίας, όταν $k = 2$ (1^{η} *στήλη*) και $k = 3$ (2^{η} *στήλη*) αντίστοιχα, στην περίπτωση του καρκίνου της στοματικής κοιλότητας.

Περιοχές	*Εκ των υστέρων μέσες τιμές & τυπικές αποκλίσεις των λ για* ***k=2***	*Εκ των υστέρων μέσες τιμές & τυπικές αποκλίσεις των λ για* ***k=3***
1	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
2	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
3	1.0669 με τ.α: (0.0704)	1.0537 με τ.α: (0.0467)
4	1.0790 με τ.α: (0.0485)	1.0609 με τ.α: (0.0248)
5	1.0838 με τ.α: (0.0389)	1.0609 με τ.α: (0.0248)
6	1.0801 με τ.α: (0.0457)	1.0609 με τ.α: (0.0248)
7	1.0809 με τ.α: (0.0435)	1.0609 με τ.α: (0.0248)
8	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
9	0.8329 με τ.α: (0.0324)	0.8072 με τ.α: (0.0318)
10	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
11	0.8330 με τ.α: (0.0326)	0.8072 με τ.α: (0.0318)
12	0.9094 με τ.α: (0.1200)	0.9334 με τ.α: (0.1501)
13	1.0807 με τ.α: (0.0440)	1.0609 με τ.α: (0.0248)
14	0.8458 με τ.α: (0.0629)	0.8072 με τ.α: (0.0318)
15	0.8724 με τ.α: (0.0957)	0.9093 με τ.α: (0.1251)
16	0.8325 με τ.α: (0.0308)	0.8072 με τ.α: (0.0318)
17	0.8325 με τ.α: (0.0308)	0.8072 με τ.α: (0.0318)
18	1.0332 με τ.α: (0.1041)	1.0293 με τ.α: (0.0846)
19	0.8451 με τ.α: (0.0610)	0.8444 με τ.α: (0.0919)
20	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
21	0.8325 με τ.α: (0.0308)	1.1854 με τ.α: (0.0629)
22	0.8622 με τ.α: (0.0849)	0.8797 με τ.α: (0.1150)
23	0.8325 με τ.α: (0.0308)	0.8072 με τ.α: (0.0318)
24	1.0877 με τ.α: (0.0265)	1.0609 με τ.α: (0.0248)
25	1.0813 με τ.α: (0.0428)	1.0609 με τ.α: (0.0248)
26	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
27	1.0868 με τ.α: (0.0303)	1.0609 με τ.α: (0.0248)
28	0.8325 με τ.α: (0.0308)	1.1854 με τ.α: (0.0629)
29	1.0335 με τ.α: (0.1044)	1.0516 με τ.α: (0.0531)
30	0.8325 με τ.α: (0.0308)	0.8072 με τ.α: (0.0318)
31	0.8325 με τ.α: (0.0308)	0.8072 με τ.α: (0.0318)
32	0.8325 με τ.α: (0.0308)	0.8072 με τ.α: (0.0318)

33	0.9033 με τ.α: (0.1160)	0.8850 με τ.α: (0.1217)
34	0.8325 με τ.α: (0.0308)	0.8072 με τ.α: (0.0318)
35	0.8494 με τ.α: (0.0683)	0.8486 με τ.α: (0.0953)
36	0.8383 με τ.α: (0.0471)	0.8072 με τ.α: (0.0318)
37	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
38	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
39	0.8325 με τ.α: (0.0308)	0.8072 με τ.α: (0.0318)
40	1.0047 με τ.α: (0.1191)	1.0024 με τ.α: (0.1069)
41	1.0818 με τ.α: (0.0417)	1.0609 με τ.α: (0.0248)
42	0.9843 με τ.α: (0.1260)	1.0444 με τ.α: (0.0669)
43	0.8325 με τ.α: (0.0308)	0.8072 με τ.α: (0.0318)
44	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
45	0.8348 με τ.α: (0.0384)	0.8129 με τ.α: (0.0487)
46	0.8325 με τ.α: (0.0308)	0.8072 με τ.α: (0.0318)
47	0.9970 με τ.α: (0.1230)	1.0431 με τ.α: (0.0698)
48	0.8325 με τ.α: (0.0308)	0.8072 με τ.α: (0.0318)
49	0.8362 με τ.α: (0.0426)	0.8072 με τ.α: (0.0318)
50	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
51	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
52	0.8355 με τ.α: (0.0405)	0.8072 με τ.α: (0.0318)
53	1.0877 με τ.α: (0.0263)	1.1854 με τ.α: (0.0629)
54	1.0877 με τ.α: (0.0264)	1.0612 με τ.α: (0.0255)
55	0.8325 με τ.α: (0.0308)	1.1854 με τ.α: (0.0629)
56	1.0877 με τ.α: (0.0263)	1.1854 με τ.α: (0.0629)
57	1.0877 με τ.α: (0.0266)	1.0609 με τ.α: (0.0248)
58	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
59	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
60	0.8457 με τ.α: (0.0627)	0.8072 με τ.α: (0.0318)
61	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
62	0.8909 με τ.α: (0.1089)	0.9402 με τ.α: (0.1268)
63	0.8325 με τ.α: (0.0308)	0.8072 με τ.α: (0.0318)
64	0.8325 με τ.α: (0.0308)	0.8072 με τ.α: (0.0318)
65	0.8328 με τ.α: (0.0320)	0.8072 με τ.α: (0.0318)
66	1.0877 με τ.α: (0.0263)	1.1854 με τ.α: (0.0629)
67	1.0877 με τ.α: (0.0263)	1.1854 με τ.α: (0.0629)
68	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
69	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0249)
70	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
71	0.8415 με τ.α: (0.0543)	0.8072 με τ.α: (0.0318)

72	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
73	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
74	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
75	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
76	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
77	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
78	0.8325 με τ.α: (0.0308)	0.8072 με τ.α: (0.0318)
79	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
80	0.8418 με τ.α: (0.0559)	0.8148 με τ.α: (0.0534)
81	0.8328 με τ.α: (0.0320)	0.8073 με τ.α: (0.0324)
82	1.0807 με τ.α: (0.0440)	1.0609 με τ.α: (0.0248)
83	1.0807 με τ.α: (0.0440)	1.0609 με τ.α: (0.0248)
84	0.8325 με τ.α: (0.0308)	0.8072 με τ.α: (0.0318)
85	0.8326 με τ.α: (0.0310)	0.8072 με τ.α: (0.0318)
86	0.8373 με τ.α: (0.0448)	0.8236 με τ.α: (0.0672)
87	1.0877 με τ.α: (0.0263)	1.1854 με τ.α: (0.0629)
88	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
89	1.0856 με τ.α: (0.0354)	1.1182 με τ.α: (0.0775)
90	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
91	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
92	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
93	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)
94	1.0877 με τ.α: (0.0263)	1.0609 με τ.α: (0.0248)



Ενδεικτικά στο διπλανό χάρτη φαίνονται οι διαβαθμίσεις των $\lambda_i, i = 1,2$ όταν το $k = 2$ όπως απεικονίζονται ανά περιοχή, σύμφωνα με την ανάλυση του αλγορίθμου μας για την περίπτωση θανάτων από καρκίνο της στοματικής κοιλότητας.

- Πίνακας με τις ***εκ των υστέρων μέσες τιμές & τυπικές αποκλίσεις*** των λ για τις 94 περιοχές της Γαλλίας, όταν $k = 2$ για την περίπτωση καρκίνου του φάρυγγα.

Περιοχές	*Εκ των υστέρων μέσες τιμές & τυπικές αποκλίσεις των λ για* ***k=2***
1	0.9856 με τ.α: (0.0244)
2	1.0214 με τ.α: (0.0285)
3	0.9856 με τ.α: (0.0244)
4	0.9857 με τ.α: (0.0244)
5	0.9861 με τ.α: (0.0248)
6	0.9858 με τ.α: (0.0245)
7	0.9856 με τ.α: (0.0244)
8	0.9917 με τ.α: (0.0283)
9	0.9860 με τ.α: (0.0248)
10	0.9859 με τ.α: (0.0247)
11	0.9856 με τ.α: (0.0244)
12	1.0031 με τ.α: (0.0317)
13	0.9856 με τ.α: (0.0244)
14	0.9858 με τ.α: (0.0244)
15	0.9857 με τ.α: (0.0244)
16	0.9856 με τ.α: (0.0244)
17	0.9856 με τ.α: (0.0244)
18	0.9856 με τ.α: (0.0244)
19	0.9856 με τ.α: (0.0244)
20	0.9856 με τ.α: (0.0244)
21	0.9856 με τ.α: (0.0244)
22	0.9856 με τ.α: (0.0244)
23	0.9856 με τ.α: (0.0244)
24	0.9856 με τ.α: (0.0244)
25	0.9856 με τ.α: (0.0244)
26	1.0201 με τ.α: (0.0292)
27	0.9873 με τ.α: (0.0257)
28	0.9856 με τ.α: (0.0244)
29	0.9856 με τ.α: (0.0244)
30	0.9856 με τ.α: (0.0244)
31	0.9856 με τ.α: (0.0244)
32	0.9856 με τ.α: (0.0244)

33	0.9857 με τ.α: (0.0244)
34	0.9856 με τ.α: (0.0244)
35	0.9856 με τ.α: (0.0244)
36	0.9856 με τ.α: (0.0244)
37	0.9856 με τ.α: (0.0244)
38	0.9856 με τ.α: (0.0244)
39	0.9856 με τ.α: (0.0244)
40	0.9856 με τ.α: (0.0244)
41	0.9856 με τ.α: (0.0244)
42	0.9856 με τ.α: (0.0244)
43	0.9856 με τ.α: (0.0244)
44	0.9856 με τ.α: (0.0244)
45	0.9856 με τ.α: (0.0244)
46	0.9856 με τ.α: (0.0244)
47	0.9881 με τ.α: (0.0259)
48	0.9856 με τ.α: (0.0244)
49	0.9856 με τ.α: (0.0244)
50	0.9892 με τ.α: (0.0271)
51	0.9857 με τ.α: (0.0244)
52	0.9856 με τ.α: (0.0244)
53	0.9856 με τ.α: (0.0244)
54	0.9862 με τ.α: (0.0249)
55	0.9856 με τ.α: (0.0244)
56	0.9856 με τ.α: (0.0244)
57	0.9856 με τ.α: (0.0244)
58	1.0214 με τ.α: (0.0285)
59	1.0214 με τ.α: (0.0285)
60	0.9856 με τ.α: (0.0244)
61	1.0214 με τ.α: (0.0285)
62	0.9856 με τ.α: (0.0244)
63	0.9856 με τ.α: (0.0244)
64	0.9857 με τ.α: (0.0244)
65	0.9859 με τ.α: (0.0246)
66	0.9856 με τ.α: (0.0244)
67	0.9856 με τ.α: (0.0244)
68	0.9856 με τ.α: (0.0244)
69	0.9856 με τ.α: (0.0244)
70	0.9856 με τ.α: (0.0244)
71	0.9856 με τ.α: (0.0244)

72	0.9856 με τ.α: (0.0244)
73	0.9856 με τ.α: (0.0285)
74	1.0214 με τ.α: (0.0289)
75	1.0207 με τ.α: (0.0289)
76	1.0214 με τ.α: (0.0285)
77	1.0214 με τ.α: (0.0285)
78	0.9856 με τ.α: (0.0244)
79	1.0214 με τ.α: (0.0285)
80	0.9857 με τ.α: (0.0244)
81	0.9856 με τ.α: (0.0244)
82	0.9856 με τ.α: (0.0243)
83	0.9856 με τ.α: (0.0244)
84	0.9856 με τ.α: (0.0244)
85	0.9856 με τ.α: (0.0244)
86	0.9856 με τ.α: (0.0244)
87	0.9856 με τ.α: (0.0244)
88	0.9857 με τ.α: (0.0244)
89	0.9861 με τ.α: (0.0248)
90	1.0214 με τ.α: (0.0285)
91	1.0214 με τ.α: (0.0285)
92	1.0214 με τ.α: (0.0285)
93	1.0214 με τ.α: (0.0285)
94	1.0214 με τ.α: (0.0285)

Σχόλιο $1^{ο}$: Από τα παραπάνω παραδείγματα - εφαρμογές, σε προσομοιωμένα και σε πραγματικά δεδομένα, γίνεται σαφές ότι το λ ναι μεν δείχνει την «επικινδυνότητα» μιας περιοχής από μόνο του – να υπενθυμίσουμε ότι οι τιμές του καθορίζουν το δείκτη επικινδυνότητας της κάθε περιοχής – δε μας εξασφαλίζει όμως αναλογικά το πλήθος των παρατηρήσεων/αναμενόμενα κρούσματα για την περιοχή αυτή. Δηλαδή την πρόβλεψη πως όσο μεγαλύτερη είναι η τιμή του τόσο πιο μεγάλη θα είναι και η τιμή του πλήθους των κρουσμάτων της περιοχής, αν και σε πρώτη ανάγνωση προς τα εκεί οδηγείται ο ερευνητής. Αλλά αντίθετα, ο συνδυασμός του λ με τα E (όπου E ο συντελεστής αναλογίας πληθυσμού) μας δίνουν τελικά τη ζητούμενη πρόβλεψη.

Συγκεκριμένα το γινόμενο $\lambda_i E_i$ μας δίνει τον ρυθμό εμφάνισης κρουσμάτων μέσω της $Poisson$ κατανομής και από εκεί εξάγονται τα διάφορα συμπεράσματά μας. Το αξιοσημείωτο είναι, πως θα μπορούσαμε να έχουμε χαμηλό συντελεστή λ για μια περιοχή, αλλά η αναλογία πληθυσμού (E) για τη συγκεκριμένη περιοχή να είναι αρκετά μεγάλη με αποτέλεσμα το γινόμενο των δύο τιμών και τελικώς ο ρυθμός $Poisson$ να μας δώσουν ένα πλήθος κρουσμάτων αρκετά μεγαλύτερο από μία άλλη περιοχή σύγκρισης που θα συνέβαινε το αντίστροφο (υψηλό λ και χαμηλή αναλογία πληθυσμού E).

Παρόλα αυτά, στη χωρική στατιστική και πιο ρεαλιστικά στην καθημερινότητά μας, ένας υψηλός δείκτης θνησιμότητας σε μία γεωγραφική περιοχή δε θεωρείται καθόλου αμελητέος παράγοντας. Έχει επιπτώσεις σε διάφορους κοινωνικούς, οικονομικούς και τεχνικούς τομείς, που χρήζουν ιδιαίτερης προσοχής από τους διάφορους φορείς που δραστηριοποιούνται στην κάθε περιοχή που εξετάζεται.

Σχόλιο 2^{0}: Για όλες τις δοκιμές, τις αναλύσεις και τα συμπεράσματά μας χρησιμοποιήθηκε κώδικας Matlab.

Βιβλιογραφία

- Baum (1972) The Forward – Backward Algorithm.
- Bernado, J.M. and Smith, A.F.M. (1994). Bayesian Theory. Wiley, Chichester.
- Ferrnhead, P. and Meligkotsidou, L. (2004). Exact filtering for partially – observed continuous – time models. Journal of the Royal Statistical Society, series B.
- Benoit Mandelbrot, «πατέρας» της γεωμετρίας των Φρακτάλ, Πανεπιστήμιο Γέιλ, 2000.
- Koutsourelis Antonios, Bayesian Inference for Hidden Markov Models with Applications in Financial Econometrics, Lancaster 2006.
- Gelman et al. (1995) & Carlin and Louis (1996), practical applications on MCMC.
- Greenwood and Yule (1920), "The Statistics of Anti-Typhoid and Anti-Cholera Inoculations, and the Interpretation of such Statistics in General", *Proceedings of the Royal Society of Medicine (Epidemiology).*
- Meligkotsidou Loukia, Bayesian Inference, Athens 2008.
- Scott, S.L. (2002) Bayesian Methods for Hidden Markov Models: Recursive computing in the 21th century.
- Θεωρία Πιθανοτήτων και Στατιστικής, Τ. Παπαϊωάννου (1997), Ιωάννινα.
- Στατιστικά Δείγματα και Χαρτομετρία, Βύρωνας Νάκος
- Graham J. Upton & Bernard Fingelton: Spatial Data Analysis by Example Volume 1: Point Pattern and Quantitative Data. John Wiley & Sons, New York. 1985.
- Hidden Markov Models and Disease Mapping: American Statistical Association (2002), Peter J. Green and Sylvia Richardson
- Potts, R. B., "Wilton, John Raymond (1884–1944), Mathematician", in John Ritchie (ed.), Australian Dictionary of Biography, vol. 12, Melbourne University Press, Melbourne, 1990.
- Spatial Statistic, Journal of the Royal Statistical Society, 1996.

- Spatial Analyses of Spatial Point Patterns, Peter J. Diggle, Professor of Statistics Lancaster University.
- Goodchild (1985), Geographic Information System (GIS).
- Κουτσόπουλος (2002), "Γεωγραφικά συστήματα πληροφοριών & ανάλυση χώρου".
- One of the first examples being the mapping of the city extent of Düsseldorf, Steinitz, et al. 1976.
- Medical Geography, Gesler (1986).
- Μπεϋζιανή Στατιστική και MCMC, Δημ. Φουσκάκης.
- American Journal of Epidemiology (Disease Mapping).
- Glass G.E., 1995, «Μοντέλα κινδύνου από τον συνδυασμό των G.I.S και της ανάλυσης λογιστικής παλινδρόμησης».
- Tomlin, (1990) Geographic Information systems and cartographic modeling.
- Public health research, John Snow (1978).
- Haggett, University of Cambridge: University Demonstrator in Geography (1994).
- Analysing Spatial Data in R, Roger Bivard.
- Γεωγραφικός χώρος & Ασθένειες (Έρευνες – Διατριβές), Α. Τσάτσαρης, Ι. Κάτσιος, Μ. Δημητράκης, Ρ. Ντούβαλη.
- Απεικονίσεις χαρτών, Ινστιτούτα Γεωγραφικής Ανάλυσης & Μελέτης Γαλλίας (5/2007).
- Administration territorial de la France (2009).
- National Center for Geographic Information and Analysis, 1990.