ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Διπλωματική Εργασία

ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΑΝΕΡΓΙΑΣ ΜΕ ΧΡΗΣΗ ΤΗΣ R

Ραμαντάνη Αικατερίνη

Τριμελής Επιτροπή: Φουσκάκης Δημήτρης (Επιβλέπων Καθηγητής)
Κολέτσος Ιωάννης
Παπανικολάου Βασίλειος

Αθήνα, Ιούλιος 2015

Ευχαριστίες

Με την ολοκλήρωση αυτής της διπλωματικής εργασίας φτάνουν στο τέλος οι προπτυχιακές σπουδές μου στη σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών. Θα ήθελα λοιπόν να ευχαριστήσω όλους εκείνους που με βοήθησαν, άλλοι σε μεγαλύτερο και άλλοι σε μικρότερο βαθμό στην ολοκλήρωση των σπουδών μου αυτά τα πέντε έτη.

Περίληψη

Η παρούσα αυτή διπλωματική εργασία ασχολείται με τη μελέτη δεδομένων για την ανεργία και το απασχολούμενο εργατικό δυναμικό στην Ελλάδα που προέρχονται από την Ελληνική Στατιστική Αρχή για τα έτη (και τρίμηνα) από το 2001 ως το 2014. Η μελέτη των δεδομένων έγινε για διάφορες κατηγορίες όπως ο βαθμός αστικότητας, το φύλο, η ηλικιακή ομάδα, το επίπεδο εκπαίδευσης, ο επαγγελματικός τομέας και ο οικονομικός κλάδος δραστηριότητας. Για την μελέτη των πινάκων συχνοτήτων και σχετικών συχνοτήτων των δεδομένων χρησιμοποιήθηκαν μέθοδοι και διαγράμματα πολυμεταβλητής στατιστικής που υλοποιήθηκαν στο στατιστικό πακέτο R με τη βοήθεια διαφόρων πακέτων.
Αρχικά έγινε χρήση της μεθόδου της ανάλυσης αντιστοιχιών, η οποία παρουσιάστηκε με κάθε λεπτομέρεια και εφαρμόστηκε σε πίνακες σχετικών συχνοτήτων για τους ανέργους ανά βαθμό αστικότητας και ανά ηλικιακή. Η εφαρμογή της μεθόδου έγινε με τη βοήθεια τριών πακέτων της R που πραγματοποιεί τη μέθοδο αυτή, ενώ έγινε αναφορά και σε όλα τα δυνατά πακέτα που εφαρμόζουν τη μέθοδο της ανάλυσης αντιστοιχιών. Ταυτόχρονα μελετήθηκε με τη χρήση διαφόρων διαγραμμάτων που προσφέρει η R και η σχέση των ποσοστών των ανέργων σε συγκεκριμένα αστικά κέντρα καθώς και η διαφοροποίηση των ποσοστών ανά φύλο. Στη συνέχεια ακολούθησε η μελέτη των σχετικών συχνοτήτων των ανέργων ανά κατηγορία μορφωτικού επιπέδου με τη βοήθεια διαγραμματικών απεικονίσεων του πίνακα των ποσοστών. Τέλος ο αριθμός των απασχολούμενων ανά βασικούς επαγγελματικούς τομείς και ανά κλάδο οικονομικής δραστηριότητας μελετήθηκε με τη βοήθεια διαγραμμάτων πολυμεταβλητής στατιστικής με την κάθε κατηγορία ανά πίνακα συχνοτήτων να αντιστοιχεί σε μία μεταβλητή. Για τη δημιουργία και την παρουσίαση των διαγραμμάτων έχουν χρησιμοποιηθεί πολλά πακέτα της R για τα οποία γίνεται αναλυτική παρουσίαση των δυνατοτήτων που προσφέρουν αλλά και πλήρης εξήγηση των αποτελεσμάτων τους για τα δεδομένα της εργασίας.

Abstract

The present thesis deals with the study of data on unemployment and number of employed in Greece, collected from the Hellenic Statistical Authority, for years and quarters of years from 2001 until 2014. The study of data refers to various categories such as the degree of urbanity, gender, age group, level of education, the professional sector and the sector of economical activity. For the study of frequency tables and relevant frequency tables methods and diagramms of multivariate statistics were used. The statistical package R was used for the study of all different categories of unemployed as well as for the frequency of employed.

Firstly we used the method of correspondence analysis, which was presented in detail and applied to the relevant frequency tables for the unemployed by degree of urbanization and by age. At the same time the percentage of unemployed in certain urban areas was studied as well as the difference in percentages of male and female unemployed. This was followed by the study of the percentances of the unemployed by category of educational level. Finally the number of employees per main proffesions and by branch of economic activity was studied with the help of diagrams for multivariate statistics, with each category of profession and economic sector to correspond to one variable. For the creation and presentation of diagrams many packages of R were used for which detailed presentation was made.

Περιεχόμενα

Περίληψη **4**
Abstract **5**

Κεφάλαιο 1 Εισαγωγή
1.1 Η οικονομική ύφεση στην Ελλάδα 10
1.2 Αντικείμενο και σκοπός της διπλωματικής εργασίας 11
1.3 Στάδια υλοποίησης της εργασίας 11
1.4 Σχολιασμός των μεταβλητών που θα χρησιμοποιηθούν 12

Κεφάλαιο 2 Το ποσοστό των ανέργων ανά βαθμό αστικότητας 15
2.1 Η μέθοδος της ανάλυσης αντιστοιχειών 15
2.1.1 Χρήση και προϋποθέσεις της μεθόδου 15
2.1.2 Χρήσιμες έννοιες και ορισμοί 16
2.1.3 Η μέθοδος SVD 17
2.2 Παρουσίαση των δεδομένων 17
2.3 Εισαγωγή των δεδομένων στην R και η περιγραφή τους 19
2.3.1 Εισαγωγή των δεδομένων μέσω του πακέτου "gdata" 19
2.3.2 Περιγραφικά στοιχεία των δεδομένων 20
2.4 Η ανάλυση αντιστοιχειών στην R με το πακέτο "ca" 22
2.5 Δημιουργία διαγραμμάτων 24
2.5.1 Κατασκευή στην R και παρουσίαση του συμμετρικού διαγράμματος 25
2.5.2 Κατασκευή στην R και παρουσίαση του ασύμμετρου διαγράμματος 27
2.5.3 Κατασκευή στην R και παρουσίαση του συμμετρικού διαγράμματος για τις στήλες του πίνακα συχνοτήτων 29
2.6 Η σχετική συχνότητα σε τρεις κατηγορίες αστικών κέντρων 30

Κεφάλαιο 3 Το ποσοστό των ανέργων ανά ηλικιακή ομάδα και ανά φύλο 37
3.1 Παρουσίαση των δεδομένων για την σχετική συχνότητα των ανέργων ανά φύλο και η περιγραφή τους 37
3.2 Παρουσίαση των δεδομένων για τη σχετική συχνότητα των ανέργων ανά ηλικιακή ομάδα 41
3.3 Περιγραφικά στοιχεία για τα δεδομένα για την ηλικιακή ομάδα των ανέργων 42
3.4 Η ανάλυση αντιστοιχειών με το πακέτο "anacor" 45
3.4.1 Εφαρμογή του πακέτου 46
3.4.2 Τα διαγράμματα του πακέτου 47
3.5 Η ανάλυση αντιστοιχειών με το πακέτο "languageR" 53
3.6 Αναφορά στην ανάλυση αντιστοιχειών με άλλα πακέτα 55

Κεφάλαιο 4 Το ποσοστό των ανέργων ανά επίπεδο εκπαίδευσης 56
4.1 Παρουσίαση του πίνακα σχετικών συχνοτήτων 56
4.2 Εισαγωγή των δεδομένων στην R 58
4.3 Βασικά περιγραφικά στοιχεία των δεδομένων 59

4.4 Διαγράμματα των κατηγοριών μορφωτικού επιπέδου 61
4.4.1 Θηκόγραμμα (boxplot) 61
4.4.2 Θηκόγραμμα με εγκοπές (boxplot with notches) 63
4.4.3 Διαγράμματα με τη μέθοδο των πυρήνων για κάθε κατηγορία (density plot) 67
4.4.4 Το διάγραμμα σε σχήμα βιολιού (violin plot) 77
4.4.5 Το διάγραμμα σε σχήμα φασολιού (beanplot) 79

Κεφάλαιο 5 Απασχολούμενοι ανά επαγγελματικό τομέα και ανά κλάδο οικονομικής δραστηριότητας 85
5.1 Παρουσίαση των δεδομένων 85
5.1.1 Οι συχνότητες απασχολουμένων ανά επαγγελματικό τομέα 85
5.1.2 Οι συχνότητες απασχολουμένων ανά κλάδο οικονομικής δραστηριότητας 87
5.2 Εισαγωγή στην R και περιγραφή των πινάκων δεδομένων 88
5.3 Διαγράμματα των δεδομένων ανά επαγγελματικό τομέα με το πακέτο "aplpack" 92
5.3.1 Το δισδιάστατο κυτιογράφημα (bagplot) 92
5.3.2 Το διάγραμμα αποτέλεσμα της εντολής "plotsummary" 96
5.3.3 Τα πρόσωπα του Chernoff 99
5.4 Το διάγραμμα-αστέρι (starplot) για τους διάφορους επαγγελματικούς τομείς 101
5.5 Οι καμπύλες του Andrews για τους επαγγελματικούς τομείς 103
5.6 Διαγράμματα για τα δεδομένα των απασχολούμενων ανά κλάδο οικονομικής δραστηριότητας 106
5.5.1 Το διάγραμμα "matrix plot" 106
5.5.2 Το κοινό διάγραμμα διασπορών (matplot) 109
5.5.3 Το τρισδιάστατο διάγραμμα διασπορών για τη συχνότητα απασχολούμενων ανά κλάδο οικονομικής δραστηριότητας ανά χρονιά με τον συνολικό αριθμό απασχολούμενων 110

Κεφάλαιο 6 Τελική ανασκόπηση των αποτελεσμάτων 114
6.1 Παρουσίαση των αποτελεσμάτων για κάθε μία μεταβλητή 114
6.1.1 Η σχετική συχνότητα των ανέργων ανά βαθμό αστικότητας 114
6.1.2 Η σχετική συχνότητα των ανέργων ανά ηλικιακή ομάδα και ανά φύλο 114
6.1.3 Η σχετική συχνότητα των ανέργων ανά επίπεδο εκπαίδευσης 115
6.1.4 Η συχνότητα των ανέργων ανά επαγγελματικό τομέα και ανά κλάδο οικονομικής δραστηριότητας 116
6.2 Τελικό συμπέρασμα 117

Βιβλιογραφία 118

1. Εισαγωγή

1.1 Η οικονομική ύφεση στην Ελλάδα και η ανεργία

Την τελευταία δεκαετία η Ελλάδα εισήλθε σε μία κρίσιμη καμπή της σύγχρονης ιστορίας της καθώς βρέθηκε αντιμέτωπη με μια βαθιά οικονομική ύφεση που άρχισε στις ΗΠΑ (Reinhart, 2008) ως χρηματοοικονομική κρίση το 2007 και οδήγησε σε μια παγκόσμια ύφεση που απειλεί τώρα πολλές υπερχρεωμένες χώρες της Ευρώπης. Οι αποκαλύψεις για το γεγονός ότι το δημοσιονομικό έλλειμμα της Ελλάδας το 2009 βρέθηκε σε επίπεδο πολύ μεγαλύτερο από αυτό που θα καθιστούσε το δημοσιονομικό χρέος βιώσιμο έκαναν αδύνατο το δανεισμό με λογικά επιτόκια από τις αγορές για τη χρηματοδότηση του τρέχοντος δημοσιονομικού ελλείμματος και την αναχρηματοδότηση του χρέους. Ως αποτέλεσμα υπήρξε κίνδυνος χρεοκοπίας της χώρας και στάσης πληρωμών του δημοσίου τομέα. Η τότε ελληνική κυβέρνηση για να ανακτήσει την αξιοπιστία της χώρας στις διεθνείς αγορές και να πετύχει μείωση των επιτοκίων οδηγήθηκε σε λήψη μέτρων μείωσης των δημοσίων δαπανών που όμως δεν ήταν αρκετές και τελικά η Ελλάδα οδηγήθηκε στο να ζητήσει βοήθεια από το Διεθνές Νομισματικό Ταμείο, την Ευρωπαϊκή Ένωση και την Ευρωπαϊκή Κεντρική Τράπεζα. Η χώρα από τότε εισήλθε σε περίοδο μεγάλης οικονομικής ύφεσης, η οποία έχει επηρεάσει σε μεγάλο βαθμό τη καθημερινότητα των πολιτών και έχει ως επακόλουθο την αδυναμία απορρόφησης του εργατικού δυναμικού. Έτσι τα ποσοστά ανεργίας κατά την ύφεση αυτή έχουν αυξηθεί ραγδαία και η κίνηση του χρήματος όλο και λιγοστεύει.

Ανεργία είναι η κατάσταση κατά την οποία το άτομο ενώ είναι πρόθυμο και ικανό να απασχοληθεί δεν δύναται να βρει εργασία. Ως αποτέλεσμα μειώνεται το εισόδημά του ή και μηδενίζεται. Η μορφή ανεργίας που υπάρχει σήμερα στην Ελλάδα είναι η λεγόμενη κυκλική μορφή, που οφείλεται στη μειωμένη ζήτηση προϊόντων και υπηρεσιών λόγω της έλλειψης ανάπτυξης.

Ενδεικτικά παρατίθεται στον Πίνακα 1.1 το ποσοστό ανεργίας για τις χρονιές 2009 και 2013 και η αλλαγή του ποσοστού αυτού. Ως ποσοστό ανεργίας έχει οριστεί το πηλίκο των ανέργων διαιρούμενο με το σύνολο του εργατικού δυναμικού, δηλαδή το σύνολο των ατόμων που έχουν δηλώσει ότι επιθυμούν και είναι ικανοί να εργασθούν. Ο πίνακας προέρχεται από το σύνολο δεδομένων της Ελληνικής Στατιστικής Αρχής.

Έτος	Ποσοστό ανέργων
2009	9.60%
2013	27.50%
Αύξηση	17.90%

Πίνκακας 1.1 (Πηγή Ελληνική Στατιστική Υπηρεσία)

1.2 Αντικείμενο και σκοπός της διπλωματικής εργασίας

Η εργασία αυτή έχει ως σκοπό τη μελέτη δεδομένων σχετικά με την ανεργία στην Ελλάδα, την εξέλιξη της από το 2001 και μετά και την παρατήρηση παραγόντων που έχουν επηρεαστεί αλλά και κατηγορίες ατόμων που έχουν πληγεί περισσότερο από την οικονομική ύφεση. Επίσης θα μελετηθούν και οι οικονομικοί κλάδοι ή τα επαγγέλματα που έχουν υποστεί τη μεγαλύτερη έλλειψη απορρόφησης εργατικού δυναμικού.

Τα δεδομένα προέρχονται από την Ελληνική Στατιστική Αρχή η οποία συνήθως αναφέρεται με την συντομογραφία ΕΛ.ΣΤΑΤ., που είναι ο επίσημος δημόσιος στατιστικός φορέας της χώρας και ασχολείται κατά κύριο λόγο με τη συγκέντρωση στατιστικών πληροφοριών για ζητήματα πληθυσμού και κατοικιών της χώρας καθώς και ζητήματα οικονομίας, κοινωνικών προβλημάτων αλλά και θέματα καθημερινότητας και εκπροσωπεί την Ελλάδα ως «εθνική στατιστική υπηρεσία» στις υπηρεσίες της Ευρωπαϊκής Ένωσης και σε κάθε άλλο αρμόδιο διεθνή οργανισμό. Επίσης συνεργάζεται με την Ευρωπαϊκή Στατιστική Υπηρεσία (Eurostat) και με άλλες υπηρεσίες της Ευρωπαϊκής Επιτροπής καθώς και με τις εθνικές στατιστικές υπηρεσίες των άλλων κρατών μελών για στατιστικά θέματα.

1.3 Στάδια υλοποίησης της εργασίας

Η εργασία αυτή πραγματοποιήθηκε στο διάστημα Μάρτιος-Ιούλιος 2015 και η εκπόνηση της έγινε σε 4 στάδια.

Στάδιο 1^{ο}

Αυτό το στάδιο περιλαμβάνει την ανάληψη δίμηνης πρακτικής άσκησης στην Ελληνική Στατιστική Αρχή. Κατα την παραμονή στην υπηρεσία πέρα από τις άλλες αρμοδιότητες που χρειάστηκαν έγινε η συλλογή των δεδομένων από το σύνολο της βιβλιοθήκης των δεδομένων της ΕΛ.ΣΤΑΤ. Επίσης έγινε και μία επαφή με τον τρόπο συλλογής των στοιχείων αυτών από απογραφές και έρευνες σε δείγματα από ερωτηματολόγια σχετικά με τον πληθυσμό.

Στάδιο 2^{ο}

Στο δεύτερο στάδιο έγινε η καταγραφή των μεταβλητών που θα χρησιμοποιηθούν τελικά για την εργασία και ο διαχωρισμός των μεταβλητών σε αυτές που παρουσιάζουν ενδιαφέρον και σε αυτές που δεν χρησιμοποιήθηκαν. Επίσης συζητήθηκαν όλες οι πολυμεταβλητές τεχνικές που θα μπορούσαν να φανούν χρήσιμες και αντιστοιχήθηκαν οι ομάδες δεδομένων με συγκεκριμένες τεχνικές ανάλυσής τους.

Στάδιο 3^{ο}

Μετά από την παραπάνω ανάλυση ακολούθησε η επαφή με τις τεχνικές που θα

χρησιμοποιηθούν και η πλήρης κατανοησή όχι μόνο της χρήσης τους αλλά και της πλήρης λειτουργίας και μεθοδολογίας τους. Ύστερα έγινε η επεξεργασία των δεδομένων, η εφαρμογή των μεθόδων και η δημιουργία διαγραμμάτων για την καλύτερη κατανόηση της εξάρτησης των μεταβλητών και της εξέλιξης τους ανά κατηγορία με τη βοήθεια του στατιστικού πακέτου R.

Στάδιο $4^{ο}$

Στο τελικό στάδιο έγινε η καταγραφή των μεθόδων και των συμπερασμάτων, η οργάνωση του υλικού σε κεφάλαια και η εξαγωγή τελικών συμπερασμάτων που βοηθούν στην κατανόηση της ανάλυσης που έγινε στο σύνολο της εργασίας και ο σχολιασμός τους.

1.4 Σχολιασμός των μεταβλητών που θα χρησιμοποιηθούν

Η μεταβλητή που μελετάται για όλα τα κεφάλαια της εργασίας θα είναι το ποσοστό της ανεργίας, που έχει υπολογιστεί ως το πηλίκο του αριθμού των ανεργων προς τον αριθμό του εργατικού δυναμικού (όχι του πληθυσμού), ή ο αριθμός των ανέργων ανά κατηγορία μελέτης.
Οι κατηγορίες που θα μελετηθούν είναι:

- το φύλο που είναι μια δίτιμη κατηγορική μεταβλητή
- η ηλικιακή ομάδα που μπορεί να είναι μία από τις εξής κατηγορίες 15-19, 20-24, 25-29, 30-44, 45-64, 64+. Τα διαστήματα αυτά ξεκινούν από την ηλικία των 15 όποτε είναι νόμιμη η εργασία και χωρίζονται σε διαστήματα. Τελευταίο διάστημα είναι η ηλικία από 64 και πάνω, οπότε και η πλειοψηφία των Ελλήνων πολίτων δικαιούνται σύνταξη. Η πλειοψηφία αυτής της κατηγορίας ατόμων δε θεωρείται ότι είναι πλέον οικονομικά ενεργοί
- ο βαθμός αστικότητας που δείχνει τον αριθμό των ανέργων σε αστικές, ημιαστικές και αγροτικές περιοχές. Για τις αστικές περιοχές χρησιμοποιείτε ο αριθμός των ανέργων στην περιφέρεια της πρωτεύουσας Αθήνας, στην περιφέρεια της Θεσσαλονίκης και σε υπόλοιπες αστικές περιοχές, ώστε να παρατηρήσει κανείς την πορεία της ανεργίας στα σημαντικότερα αστικά κέντρα της χώρας
- το επίπεδο εκπαίδευσης, που προσφέρει μεγάλο ενδιαφέρον καθώς προσφέρει πληροφορία για το τι είδους ζήτηση υπάρχει σε σχέση με την εκπαίδευση στην σύγχρονη Ελλάδα και είναι ενδεικτικό του είδους της εργασίας που έχει πληγεί περισσότερο από την κρίση (δουλειά με περισσότερες γνωστικές απαιτήσεις ή με γνώσεις πιο πρακτικές και καθημερινές)
- ο κλάδος οικονομικής δραστηριότητας όπου μπορεί να δει κανείς ποιος κλάδος απασχολεί το μεγαλύτερο ποσοστό ατόμων και στη συνέχεια γίνεται μελέτη του τομέα που απασχολεί το μεγαλύτερο κομμάτι του πληθυσμού ανάμεσα στον πρωτογενή, δευτερογενή και τριτογενή και ποιος από τους τρεις έχει πληγεί λιγότερο και ποιος περισσότερο από την οικονομική ύφεση.

Όλες οι μεταβλητές θα μελετηθούν ως χαρακτηριστικό στο σύνολο των ανέργων της χώρας για τις χρονιές από 2001 ως και το 2014. Σε κάποιες περιπτώσεις θα γίνει ξεχωριστή μελέτη των χαρακτηριστικών για τα τρίμηνα της κάθε χρονιάς.

Το σύνολο των δεδομένων για τις χρονιές 2001 και 2011 προέρχονται από τις εθνικές απογραφές πληθυσμού που έλαβαν χώρα στην Ελλάδα αυτές τις χρονιές, ενώ για τις υπόλοιπες χρονιές προέρχονται από ερευνητική μελέτη σε ένα σύνολο δειγμάτων από διάφορες περιοχές της χώρας και εξαγωγή των τελικών αριθμών με μεθόδους δειγματοληψίας.

Τα νούμερα στο σύνολο των δεδομένων έχουν προέλθει από την καταμέτρηση των απαντήσεων για την κάθε μεταβλητή στα ερωτηματολόγια που είχαν χρησιμοποιηθεί και για την πλήρη εικόνα των αποτελεσμάτων έχουν χρησιμοποιηθεί και μέθοδοι εύρεσης ελλιπουσών τιμών, συγκεκριμένα η μέθοδος "hot deck procedure imputation"[1], για τις περιπτώσεις που οι απαντήσεις δεν ήταν ξεκάθαρες σε κάποια μεταβλητή ενώ οι υπόλοιπες μεταβλητές ήταν σωστά συμπληρωμένες.

1 Πρόκειται για μία αρκετά διαδεδομένη μέθοδο συμπλήρωσης ελλιπουσών τιμών. Σε αυτή τη μορφή imputation θεωρούνται γνωστές οι τιμές για τις άλλες μεταβλητές του ερωτηματολογίου και γίνεται χρήση τους για να δωθεί τιμή στη μεταβλητή στην οποία δεν έχει εκχωρηθεί τιμή αναζητώντας στο σύνολο των παρατηρήσεων κάποια που έχει τις υπόλοιπες τιμές πολύ κοντά σε σχέση με αυτή που πρέπει να συμπληρωθεί.

2. Το ποσοστό των ανέργων ανά βαθμό αστικότητας

Σε αυτό το κεφάλαιο θα γίνει χρήση της μεθόδου της ανάλυσης αντιστοιχιών για τη δημιουργία χαρακτηριστικών διαγραμμάτων που θα δώσουν μία εικόνα για την κατανομή των ανέργων ανά βαθμό αστικότητας. Για την εφαρμογή της μεθόδου και την δημιουργία των διαγραμμάτων που θα παρουσιαστούν χρησιμοποιήθηκε το στατιστικό πακέτο R.

2.1 Η μέθοδος της ανάλυσης αντιστοιχιών

2.1.1 Χρήση και προϋποθέσεις της μεθόδου

Η μέθοδος της ανάλυσης αντιστοιχειών χρησιμοποιείται για να μετατρέψει έναν πίνακα δεδομένων σε γραφική παράσταση έτσι ώστε να γίνονται εμφανείς οι τυχόν συσχετίσεις των μεταβλητών μεταξύ τους και αφορά κατηγορικές μεταβλητές και όχι ποσοστικές.Λόγω της πολυδιάστατης φύσης των δεδομένων που διαχειρίζεται η Παραγοντική Ανάλυση των Αντιστοιχιών μπορεί να θεωρηθεί ως μια Πολυμεταβλητή Στατιστική Μέθοδος μείωσης των διαστάσεων του αρχικού χώρου, στον οποίο περιγράφεται το υπό εξέταση φαινόμενο (Dimensionality Reduction Method). Με τη μέθοδο της ανάλυσης αντιστοιχειών μπορεί κανείς να δει ενδιαφέρουσες σχέσεις μεταξύ των μεταβλητών που θα ήταν δύσκολο να τις εντοπίσει κανείς στις αριθμητικές τιμές των πινάκων, ειδικά όταν πρόκειται για μεγάλους πίνακες. Στην περίπτωση της κατάταξης σε περισσότερες από δύο διαστάσεις γίνεται χρήση της πολλαπλής ανάλυσης αντιστοιχειών.
Η βασική ιδέα είναι να μετατραπεί το νέφος των σημείων σε λιγότερες διαστάσεις ώστε η νέα απεικόνιση να έχει περισσότερη πληροφορία. Επίσης σκοπός είναι να αναπαρασταθούν στο ίδιο διάγραμμα τόσο οι γραμμές όσο και οι στήλες του πίνακα των δεδομένων.
Ο ερευνητής στη μέθοδο αυτή δεν υποθέτει κάποιο συγκεκριμένο μοντέλο αλλά προσπαθεί να εντοπίσει τη δομή πίσω από τα δεδομένα. Γενικά η μέθοδος αυτή δεν χρειάζεται αρχικές υποθέσεις για να εφαρμοστεί και χαρακτηρίζεται ως “model-free” καθώς δε θέτει εξαρχής κάποιο μοντέλο και προσπαθεί να υπολογίσει τις παραμέτρους του όπως κάνουν άλλες μέθοδοι.
Οι κλάσεις κάθε μεταβλητής (ερώτησης) ονομάζονται και ιδιότητες της αντίστοιχης μεταβλητής. Κάθε γραμμή χαρακτηρίζεται από μία και μόνο ιδιότητα για κάθε μεταβλητή (αντίστοιχα μπορούμε να ορίσουμε και για κάθε στήλη).
Για την εφαρμογή της ανάλυσης αντιστοιχειών μπορεί κανείς να δουλέψει με πίνακες συνάφειας οι οποίοι περιλαμβάνουν είτε απόλυτες συχνότητες είτε σχετικές συχνότητες.
Η εφαρμογή της ανάλυσης αντιστοιχειών γίνεται μέσω της διάσπασης ιδιόμορφων τιμών (σε συντομογραφία SVD) η οποία είναι σαν γενίκευση της φασματικής μεθόδου για τους μη συμμετρικούς πίνακες.
Η ανάλυση αντιστοιχειών τελικά είναι το αποτέλεσμα της SVD του πίνακα που έχει ως

στοιχεία το αποτέλεσμα που προκύπτει αν από την πραγματική συχνότητα αφαιρέσουμε την αναμενόμενη (υπό την προυπόθεση της ανεξαρτησίας), υψώσουμε στο τετράγωνο και στη συνέχεια διαιρέσουμε με την θεωρητική συχνότητα. Όπου πραγματική συχνότητα είναι οι τιμές που έχουν προκύψει από την έρευνα διεξαγωγής του πειράματος, αναμενόμενη συχνότητα είναι οι τιμές των συχνοτήτων για τον πίνακα δεδομένων σε περίπτωση που ισχύει η μηδενική υπόθεση ότι οι διάφορες κατηγορίες του πίνακα δεδομένων δεν έχουν διαφορετική συχνότητα και θεωρητική συχνότητα είναι αυτή που αντιστοιχεί σε κάθε κελί του πίνακα δεδομένων όταν δεν υπάρχει εξάρτηση μεταξύ των γραμμών και των στηλών του.

2.1.2 Χρήσιμες έννοιες και ορισμοί

Για τη χρήση της μεθόδου της ανάλυσης αντιστοιχειών θα πρέπει να οριστούν κάποιες σημαντικές ποσότητες που θα χρειαστούν.
Αυτές είναι:

- Προφίλ γραμμών

Είναι οι σχετικές συχνότητες ανά γραμμή που προκύπτουν από τη διαίρεση της συχνότητας ενός κελιού ανά τη το σύνολο της γραμμής. Το προφίλ γραμμών μας επιτρέπει την άμεση σύγκριση μεταξύ τους.

- Κεντροειδές/Μέσο προφίλ γραμμής

Πρόκειται για τον σταθμισμένο μέσο των προφίλ γραμμής με σταθμίσεις το σύνολο των παρατηρήσεων κάθε γραμμής και είναι το προφίλ γραμμής για ολόκληρο τον πίνακα.

- Απόσταση μεταξύ παρατηρήσεων

Ένα μέτρο που δείχνει τη διαφορά 2 τυχαίων παρατηρήσεων. Υπάρχουν πολλά είδη απόστασης και στην ανάλυση αντιστοιχειών η απόσταση μπορεί να μετρηθεί με τη βοήθεια των προφίλ γραμμών.

- Μάζες

Η μάζα κάθε κελιού είναι τα αντίστοιχα περιθώρια προφίλ γραμμών. Δηλαδή είναι το αποτέλεσμα της διαίρεσης του συνόλου των παρατηρήσεων της κάθε γραμμής με το σύνολο των παρατηρήσεων γενικά.

- Αδράνεια

Είναι ένα μέτρο ομοιογένειας ή ετερογένειας των προφίλ και δείχνει πόσο διαφέρουν τα προφίλ μεταξύ τους. Συμβολίζεται με I και δίνεται από τον τύπο

$$I = \sum_{i=1}^{c} (\mu \acute{\alpha} \zeta \alpha \, \tau \eta \varsigma \, \gamma \rho \alpha \mu \mu \acute{\eta} \varsigma \, i) * d_i^2$$

όπου d_i είναι η απόσταση των παρατηρήσεων.
Η έννοια της αδράνειας είναι παρόμοια με την έννοια της διακύμανσης.

Ανάλογα μπορούν να οριστούν οι ίδιες ποσότητες για στήλες αντί για γραμμές.

2.1.3 Η μέθοδος SVD

Αν θεωρήσουμε έναν πίνακα Α διαστάσεων $I \times J$ τότε γνωρίζουμε ότι με βάση την φασματικη μέθοδο μπορεί να γραφτεί στη μορφή $A = U\Sigma V'$ όπου Σ ο διαγώνιος πίνακας των ιδιόμορφων τιμών, οι οποίες είναι οι ρίζες των ιδιοτιμών και είναι θετικές και τοποθετημένες στη διαγώνιο του Σ κατά αύξουσα σειρά.
Οι πίνακες U, V είναι οι πίνακες που έχουν στήλες τα αριστερά και τα δεξιά ιδιόμορφα διανύσματα του Α αντίστοιχα και είναι ορθομοναδιαίοι (δηλαδή ισχύει η σχέση $UU' = V'V = I$). Τα αριστερά ιδιόμορφα διανύσματα είναι τα ιδιοδιανύσματα του ΑΑ' και τα δεξιά τα ιδιοδιανύσματα του Α'Α.

2.2 Παρουσίαση των δεδομένων

Τα δεδομένα αυτού του κεφαλαίου αφορούν τη σχετική συχνότητα των ανέργων ανά βαθμό αστικότητας στο σύνολο του εργατικού δυναμικού της κάθε κατηγορίας βαθμού αστικότητας και παρουσιάζονται ανά τρίμηνο για τις χρονιές 2001 εώς και 2014. Σε κάθε τρίμηνο παρουσιάζεται το ποσοστό σε Αστικές, Ημιαστικές και Αγροτικές περιοχές. Στις αστικές περιοχές περιλαμβάνεται το ποσοστό των ανέργων τόσο στην περιφέρεια της Αττικής, στα πολεοδομικά συγκροτήματα της Θεσσαλονίκης αλλά και λοιπές αστικές περιοχές με μεγάλο αριθμό κατοίκων.

Ο Πίνακας 2.1 που ακολουθεί παρουσιάζει τις σχετικές συχνότητες του βαθμού αστικότητας επί 100.

Χρονιά και αριθμός τριμήνου	Βαθμός αστικότητας		
	Αστικές	Ημιαστικές	Αγροτικές
2001a	12,3	11,7	7,9
2001b	11,6	10,7	7
2001c	11,5	10,2	6,4
2001d	12	12,4	8,4
2002a	12	13	8,7
2002b	11	10,4	6,7
2002c	11	9,8	6,3
2002d	11	11,1	7,4
2003a	10,8	11,7	8,5
2003b	10,3	9,5	6,6
2003c	10,4	9	6,4
2003d	10,8	10	7,5
2004a	11,4	12,1	11
2004b	10,7	10,5	8,8
2004c	10,9	9,9	8
2004d	10,8	10,6	9,1
2005a	10,8	10,8	9,6
2005b	10,2	9,1	8,5
2005c	10,6	8,9	7,9
2005d	10,2	9,4	8,9
2006a	9,9	9,6	9,9
2006b	9,4	8	7,7
2006c	9,1	7,8	6,4
2006d	9,6	8,3	6,8
2007a	9,4	9,8	8,4
2007b	8,7	8,2	6,6
2007c	8,7	7,2	5,8
2007d	8,5	7,9	7,3
2008a	8,4	9	7,9
2008b	7,7	7,2	6,2
2008c	7,9	6,5	5,5
2008d	8,5	7,2	7,1
2009a	9,5	10,2	9
2009b	9,4	9,4	7,5
2009c	10,5	8,7	6,4
2009d	11,1	10,9	8,3
2010a	12	12,8	11
2010b	12,6	12,2	9,9
2010c	13,7	11,3	9,5
2010d	15,4	14,2	11,2
2011a	16,7	16,4	14
2011b	17,6	15,6	13,2
2011c	19,7	16,7	12,6
2011d	22,6	20,5	15,5
2012a	24,1	22,8	18,3
2012b	25,3	23	18,9
2012c	27,1	24	18,2
2012d	28,2	25,6	19,7
2013a	29,4	26,6	22,3
2013b	29,3	25,5	21,9
2013c	29,4	25	21,3
2013d	29,3	27,5	22,5
2014a	29	28,1	23,6
2014b	28	25,5	22,2
2014c	27,3	24,4	20,3
2014d	27,5	25,8	21,1

Πίνακας 2.1: Ο πίνακας σχετικών συχνοτήτων επί 100 για τους ανέργους ανά βαθμό αστικότητας για τα τρίμηνα των ετών 2001 ως 2014

2.3 Εισαγωγή των δεδομένων στην R και περιγραφή τους

2.3.1 Εισαγωγή δεδομένων στην R μέσω του πακέτου “gdata”

Για να εισαχθούν τα δεδομένα στο στατιστικό πακέτο R θα πρέπει να βρίσκονται σε μορφή txt ή excel ή csv. Για αυτό και τα σώζουμε ανάλογα ή μετατρέπουμε το αρχείο με τη βοήθεια ιστοσελίδων ή αντίστοιχων προγραμμάτων.
Για τα αρχεία excel (document.xls) η εντολή που τα ανοίγει στην R είναι η εντολή read.xls η οποία είναι μέρος του πακέτου gdata, το οποίο και χρειάζεται να εγκατασταθεί. Πολλές από τις εντολές αυτού του πακέτου όμως δεν είναι εγκατεστημένες στην R και καλούνται μέσω μιας διαπροσωπείας από κάποια άλλη γλώσσα προγραμματισμού,για τη συγκεκριμένη εντολή η γλώσσα αυτή είναι η “perl”. Η διαπροσωπεία αυτή συνδέει το πακέτο με τη γλώσσα και δε χρειάζεται να προγραμματίσει κανείς εκτός του περιβάλοντος της R αλλά δεν είναι εγκατεστημένη εξ' ολοκλήρου στην R και χρειάζεται να ενεργοποιηθεί εκτός του περιβάλοντος της ώστε να μπορούν να λειτουργήσουν οι συγκεκριμένες εντολές του πακέτου gdata. Αυτό γίνεται με 2 τρόπους.
Ο πρώτος περιλαμβάνει να κατεβάσει κανείς, εκτός απο την R φυσικά, σε συμπιεσμένη μορφή το πακέτο RSPerl-0.92-1 το οποίο δίνει τη δυνατότητα στο χρήστη να καλεί ρουτίνες και συναρτήσεις από την γλώσσα perl σαν να ήταν μέρος του περιβάλοντος της R και αντίστροφα. Για την εγκατάσταση και χρήση εντολών της perl στην R χρειάζεται η εγκατάσταση του πακέτου, αφού έχει αποσυμπιεστεί, στο περιβάλλον της R σαν όρισμα μέσα στην εντολή και συγκεκριμένα του αρχείου “perl.exe” που βρισκόταν μέσα στο πακέτο. Το όρισμα περιλαμβάνει την ακριβή θέση του πακέτου στον υπολογιστή.
Για παράδειγμα

read.xls(“document.xls”,perl=”C://Desktop//Perl//bin//perl.exe”)

Ο δεύτερος τρόπος ενεργοποίησης της διαπροσωπείας είναι το πρόγραμμα “ActivePerl” που διατίθεται δωρεάν στο διαδίκτυο. Το μόνο που χρειάζεται να κάνει κανείς είναι να το κατεβάσει και να το εγκαταστήσει και αυτό στη συνέχεια ενεργοποιεί από μόνο του τις εντολές της perl που βρίσκονται μέσα στην R ως Rtools αλλά δεν λειτουργούν και χρειάζονται ενεργοποιήση.

Αφού λοιπόν ενεργοποιηθεί η διαπροσωπεία πληκτρολογούμε στην R τις εντολές για την εγκατάσταση του πακέτου gdata.

```
install.packages("gdata")
library(gdata)
```

Αν η διαπροσωπεία λειτουργεί με επιτυχία κάτω από την εντολή library(gdata) η R θα δείξει αν λειτουργούν όλα τα μέρη του πακέτου με το ακόλουθο αποτέλεσμα.

```
gdata: read.xls support for 'XLS' (Excel 97-2004) files ENABLED.
gdata: read.xls support for 'XLSX' (Excel 2007+) files ENABLED.
```

Attaching package: ‘gdata’

Πλέον μπορεί κανείς να περάσει τα δεδομένα στην R από ένα αρχείο excel πληκτρολογώντας την εντολή

```
data<-read.xls("../Desktop/diplwmatikh-domh/kef2/astik.xls",sheet=1,skip=1)
```

η οποία περιέχει την ακριβή θέση του αρχείου στον υπολογιστή, το υπολογιστικό φύλο που θέλουμε να διαβάσει κα στην περίπτωσή μας είναι το πρώτο και το όρισμα skip που δείχνει πόσες γραμμές θέλουμε να παραβλέψει από το αρχείο η R πριν ξεκινήσει να διαβάζει τα κελιά ως δεδομένα. Ως προεπιλογή η εντολή θεωρεί την πρώτη γραμμή σαν τίτλους των στηλών και όχι σαν παρατήρηση αν περιέχει γραφικούς χαρακτήρες και όχι αριθμητικούς. Άμα καλέσουμε λοιπόν τη μεταβλητή εκχώρησης data μας εμφανίζει τον πίνακα με τα δεδομένα σε 3 στήλες με τίτλους σε κάθε στήλη στα αγγλικά που δείχνουν τον αριθμό των ανέργων στις 3 κατηγορίες περιοχών.
Στη συνέχεια θα οριστεί η κάθε στήλη ως διάνυσμα και θα δημιουργηθεί ο πίνακας των στηλών διανυσμάτων που δε θα περιέχει τη στήλη με τον αριθμό των τριμήνων για την ευκολότερη εφαρμογή των παρακάτω πακέτων. Ο κώδικας για αυτή τη διαδικασία είναι:

```
Ast<-(data)[,2]
Hm<-(data)[,3]
Agr<-(data)[,4]
xdata<-data.frame(Ast, Hm, Agr)
```

2.3.2 Περιγραφικά στοιχεία των δεδομένων

Για την καλύτερη εξοικείωση με τα δεδομένα μπορεί κανείς να προβεί σε μερικά περιγραφικά στοιχεία αυτής της μεταβλητής ανά χρονιά. Πληκτρολογούμε λοιπόν στην R:

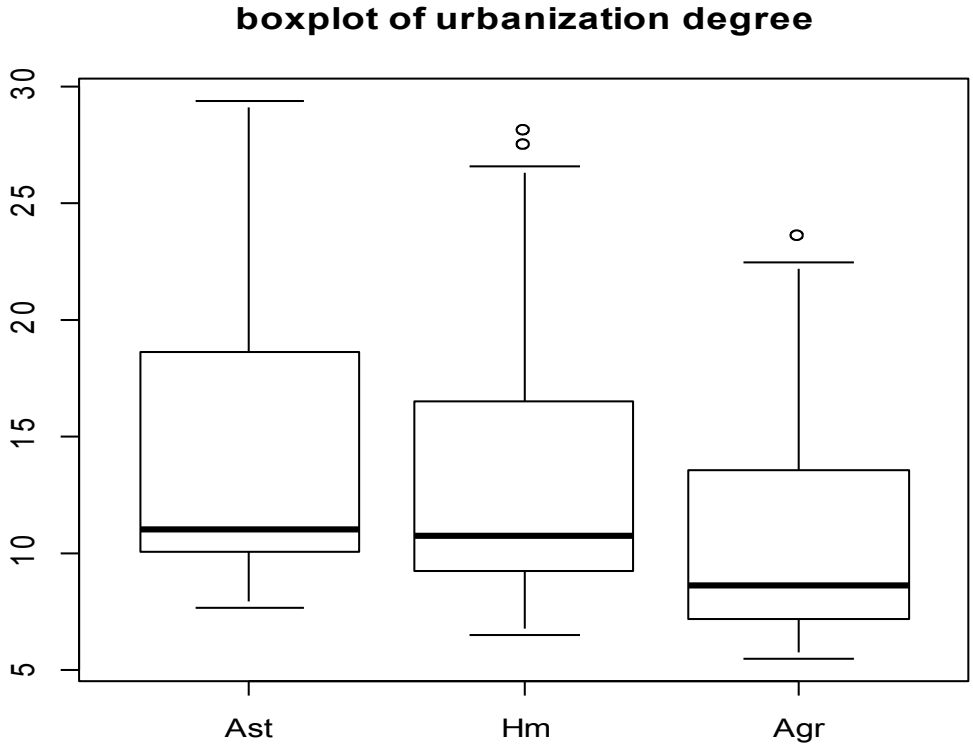
```
summary(xdata)
```

```
      Ast              Hm              Agr
 Min.   : 7.70   Min.   : 6.500   Min.   : 5.50
 1st Qu.:10.12   1st Qu.: 9.325   1st Qu.: 7.25
 Median:11.00   Median :10.750   Median : 8.60
 Mean   :14.84   Mean   :13.754   Mean   :11.14
 3rd Qu.:18.12   3rd Qu.:16.475   3rd Qu.:13.40
 Max.   :29.40   Max.   :28.100   Max.   :23.60
```

```
 sd(Ast)
[1] 7.380452
 sd(Hm)
[1] 6.613511
 sd(Agr)
[1] 5.545692
```

Κοιτώντας τα αποτελέσματα της εντολής summary φαίνεται ότι οι αστικές περιοχές παρουσιάζουν μεγαλύτερα ποσοστά κατά μέσο όρο σε σχέση με τις ημιαστικές και τις αγροτικές περιοχές. Το ενδοτεταρτημοριακό εύρος[2] και για τις τρεις περιπτώσεις δε φαίνεται να είναι πολύ διαφορετικό και κυμαίνεται στις 6 με 7 μονάδες, κάτι που σημαίνει ότι το 75% των παρατηρήσεων σε κάθε δείγμα βρίσκεται σε ίση απόσταση από τη διάμεσο του καθενός. Οι τυπικές αποκλίσεις μάλιστα φαίνεται να παίρνουν κοντινές τιμές σχετικά με την τυπική απόκλιση του δείγματος για τις αγροτικές περιοχές να είναι μικρότερη και στις αστικές περιοχές να είναι μεγαλύτερη. Η διαφοροποίηση δηλαδή των ποσοστών ανά τρίμηνο από το 2001 ως το 2014 παρουσιάζει μεγαλύτερη διαφοροποίηση στις αστικές περιοχές από ότι στις άλλες δύο. Κοιτώντας επίσης τις μέγιστες και ελάχιστες τιμές βλέπει κανείς ότι ενώ οι ελάχιστες τιμές είναι σχετικά κοντινές, με τη διαφορά τους να είναι κοντά στη μονάδα, η μέγιστες τιμές φαίνεται να έχουν μεγαλύτερη διαφοροποίηση.
Ακολουθεί και το θηκόγραμμα του Πίνακα 2.1 .

```
boxplot(xdata)
title(main="boxplot of urbanization degree")
```



Διάγραμμα 2.1: Θηκόγραμμα σχετικών συχνοτήτων επί 100 των ανέργων ανά βαθμό αστικότητας για 56 τρίμηνα από το 2001 ως το 2014

Απο το Διάγραμμα 2.1 φαίνεται ότι οι διάμεσοι δεν διαφέρουν τόσο πολύ αν και στις αστικές και στις ημιαστικές περιοχές φαίνεται υψηλότερη η τιμή της σε σχέση με αυτή στις αγροτικές περιοχές. Επίσης το ενδοτεταρτημοριακό εύρος, που είναι το μέγεθος των κουτιών στα παραπάνω θηκογράμματα φαίνεται να είναι μεγαλύτερο για τις αστικές

2 Είναι η διαφορά του τρίτου από το πρώτο τεταρτημόριο που δίνονται από την εντολή summary.

περιοχές σε σχέση με τις άλλες δύο κατηγορίες βαθμού αστικοποίησης. Τέλος το Διάγραμμα 2.1 δίνει την επιπλέον πληροφορία για την ύπαρξη ακραίων παρατηρήσεων στις κατηγορίες των ημιαστικών και αγροτικών περιοχών.

2.4 Η ανάλυση αντιστοιχιών στην R με το πακέτο "ca"

Στην περίπτωση πινάκων με τρεις γραμμές (ή στήλες), τα αντίστοιχα προφίλ μπορούν εύκολα να απεικονιστούν σε ένα χώρο δύο διαστάσεων, δηλαδή στο επίπεδο.
Το πακέτο "ca" της R είναι ένα πακέτο που εφαρμόζει την απλή ανάλυση αντιστοιχιών σε πίνακες συχνοτήτων ή σχετικών συχνοτήτων.
Εγκαθιστούμε λοιπόν το πακέτο στο παράθυρο εντολών.

```
install.packages("ca")
library(ca)
```

Στη συνέχεια αφού έχει εγκατασταθεί με επιτυχία τρέχουμε την ανάλυση αντιστοιχιών για τα δεδομένα του βαθμού αστικοποίησης και τα καταχωρούμε σε μία μεταβλητή, έστω x, με την ακόλουθη εντολή

```
x<-ca(xdata)
```

Όταν λοιπόν ζητηθεί το διάνυσμα αυτό από την R ή όταν πληκτρολογήσουμε την εντολή

```
print(x)
```

θα επιστρέψει όλα τα απαραίτητα αποτελέσματα από την εφαρμογή της μεθόδου. Το output που θα δώσει θα είναι:

```
 Principal inertias (eigenvalues):
                1           2
Value       0.002194    0.000735
Percentage  74.91%      25.09%


 Rows:
                [,1]        [,2]        [,3]        [,4]        [,5]        [,6]        [,7]
Mass       0.014340  0.013171  0.012631  0.014744  0.015149  0.012631  0.012182
ChiDist    0.073628  0.092228  0.118737  0.071932  0.084115  0.093671  0.108032
Inertia    0.000078  0.000112  0.000178  0.000076  0.000107  0.000111  0.000142
Dim. 1    -1.434085 -1.911729 -2.523369 -0.843097 -0.633078 -1.871211 -2.295211
Dim. 2     1.112235  0.815437  0.421861  2.217161  2.902744  1.219399  0.395638
                [,8]        [,9]        [,10]       [,11]       [,12]       [,13]       [,14]
Mass       0.013261 0.013935  0.011867  0.011598  0.012721 0.015508 0.013486
ChiDist    0.075552 0.068036  0.067542  0.078002  0.034073 0.101533 0.037424
Inertia    0.000076 0.000065  0.000054  0.000071  0.000015 0.000160 0.000019
```

Dim. 1	-1.140949	0.093893	-1.403109	-1.650443	-0.704708	2.072099	0.729976
Dim. 2	1.969527	2.503652	0.575316	-0.385175	0.312087	1.100399	0.561297

.
.
.
.
.
.

	[,50]	[,51]	[,52]	[,53]	[,54]	[,55]	[,56]
Mass	0.034478	0.034029	0.035647	0.036276	0.034029	0.032365	0.033444
ChiDist	0.028964	0.036553	0.009210	0.032683	0.029751	0.015869	0.008873
Inertia	0.000029	0.000045	0.000003	0.000039	0.000030	0.000008	0.000003
Dim. 1	0.089566	-0.156813	0.185770	0.659529	0.551907	-0.012401	0.178817
Dim. 2	-1.056805	-1.320445	0.111394	0.393768	-0.543170	-0.584787	0.108093

Columns:

	Ast	Hm	Agr
Mass	0.373460	0.346220	0.280320
ChiDist	0.049168	0.038764	0.073294
Inertia	0.000903	0.000520	0.001506
Dim. 1	-0.901005	-0.279609	1.545718
Dim. 2	-0.930511	1.345422	-0.422026

Από τα παραπάνω μπορεί κανείς να δει αρχικά τις ιδιόμορφες τιμές και το ποσοστό ερμηνείας της αδράνειας από την κάθε ιδιόμορφη τιμή. Έτσι στην περίπτωση του πίνακα συχνότητας για το βαθμό αστικότητας η πρώτη ιδιόμορφη τιμή είναι 0.002194 και εξηγεί το 74.91% της συνολικής αδράνειας και η δεύτερη είναι 0.000735 και εξηγεί το υπόλοιπο 25.09% της συνολικής αδράνειας.

Οι ιδιόμορφες τιμές επίσης χρησιμοποιούνται για την εύρεση των κύριων συντεταγμένων, δηλαδή τα σημεία στα οποία θα αντιπροσωπεύεται κάθε κατηγορία στη διαγραμματική απεικόνιση.

Εκτός από αυτά δίνεται από την εφαρμογή του πακέτου η μάζα, η απόσταση και η αδράνεια σε κάθε σειρά καθώς και η θέση της κάθε γραμμής. Δηλαδή για κάθε τρίμηνο από το 2001 ως το 2014 δίνονται οι συντεταγμένες της θέσης της συχνότητας της αστικότητας στους άξονες του διαγράμματος που θα δημιουργηθεί. Η απόσταση που εμφανίζει η R με την εφαρμογή του πακέτου "ca" είναι η X^2 απόσταση, η οποία δίνεται από τον τύπο

$$d = \frac{1}{2}\frac{\sum_{i=n}^{n}(x_i - y_i)^2}{x_i + y_i}$$ όπου x_i, y_i οι συχνότητες των γραμμών και των στηλών αντίστοιχα.

Στο τέλος των αποτελεσμάτων από την R εμφανίζεται επίσης ένας πίνακας που δίνει τις ίδιες ποσότητες, δηλαδή τη μάζα, την απόσταση, την αδράνεια και τη θέση στους 2 άξονες για τις 3 κατηγορίες συνολικά, τις αστικές, τις ημιαστικές και τις αγροτικές περιοχές.

Για να δει κανείς αναλυτικά τις συντεταγμένες της συχνότητας των ανέργων κάθε γραμμής χωριστά αρκεί να πληκτρολογήσει την εντολή

```
ca(xdata)$rowcoord
```

η οποία θα δώσει 2 στήλες με τις διαστάσεις 1, 2 για κάθε γραμμή.

```
           Dim1        Dim2
 [1,] -1.434084871  1.112235076
 [2,] -1.911729405  0.815436800
 [3,] -2.523368853  0.421860621
                .
                .
                .
                .
                .
                .
[53,]  0.659528692  0.393767504
[54,]  0.551907168 -0.543169961
[55,] -0.012401127 -0.584786640
[56,]  0.178816667  0.108092805
```

Επιπλέον η R μπορεί να δώσει διάφορα άλλα αποτελέσματα. Με την εντολή

```
names(x)
```

μπορεί να δει κανείς τι επιπλέον μπορεί να επιστρέψει η R. Μερικά από τα αποτελέσματα που μπορεί να επιστρέψει είναι η αδράνεια ανά στήλη και όχι μόνο ανά γραμμή, τα ονόματα των στηλών και των γραμμών, οι μάζες ανά γραμμή και ανά στήλη και άλλα.

2.5 Δημιουργία διαγραμμάτων

Για τη δημιουργία της διαγραμματικής απεικόνισης των αποτελεσμάτων της ανάλυσης αντιστοιχειών χρησιμοποιούνται οι ιδιόμορφες τιμές αλλά και τα ιδιόμορφα διανύσματα.
Κάθε ιδιόμορφη τιμή αντιστοιχεί σε έναν άξονα και όλοι οι άξονες μαζί αντιπροσωπεύουν το σύνολο των δεδομένων.
Τα ιδιόμορφα διανύσματα από την άλλη χρησιμοποιούνται για τον υπολογισμό των κύριων συντεταγμένων. Συγκεκριμένα η κύρια συντεταγμένη για τον j-άξονα της i-κατηγορίας (γραμμής) θα δίνεται από τον τύπο

$$f_{ij} = \frac{u_{ij}\gamma_j}{\sqrt{r_i}}$$

όπου

u_{ij} το ij στοιχείο του αριστερά ιδιόμορφου πίνακα U

γ_j η j ιδιόμορφη τιμή

r_i η μάζα της i γραμμής.

Σε μια γραφική αναπαράσταση της Ανάλυσης των Αντιστοιχιών, σημεία γειτονικά μεταξύ τους υποδηλώνουν και συσχετισμό ανάμεσα στις αντίστοιχες γραμμές και στήλες. Επίσης

μελετάται η ύπαρξη κάποιου είδους διάταξης μεταξύ γραμμών και μεταξύ στηλών, είτε φυσική διάταξη είτε κάποιου είδους επικάλυψης μεταξύ των διάφορων κατηγοριών των δεδομένων. Η ανάλυση αντιστοιχιών εξετάζει ακόμα αν τα ποσοστά των στηλών διαφοροποιούνται μεταξύ αυτών των γραμμών και αντιστρόφως, δηλαδή αν υπάρχει ανεξαρτησία μεταξύ στηλών-γραμμών. Γίνεται γραφική απεικόνιση του X^2 ελέγχου ανεξαρτησίας για να εξεταστεί κατά πόσο οι γραμμές και οι στήλες είναι ανεξάρτητες. Ένα ακόμα πλεονέκτημα που παρέχει η ανάλυση αντιστοιχιών είναι ότι δημιουργούνται καινούργιες μεταβλητές, οι οποίες συνοψίζουν σημαντικό μέρος της αρχικής πληροφορίας. Αυτές είναι οι νέοι άξονες που έχουν δημιουργηθεί.

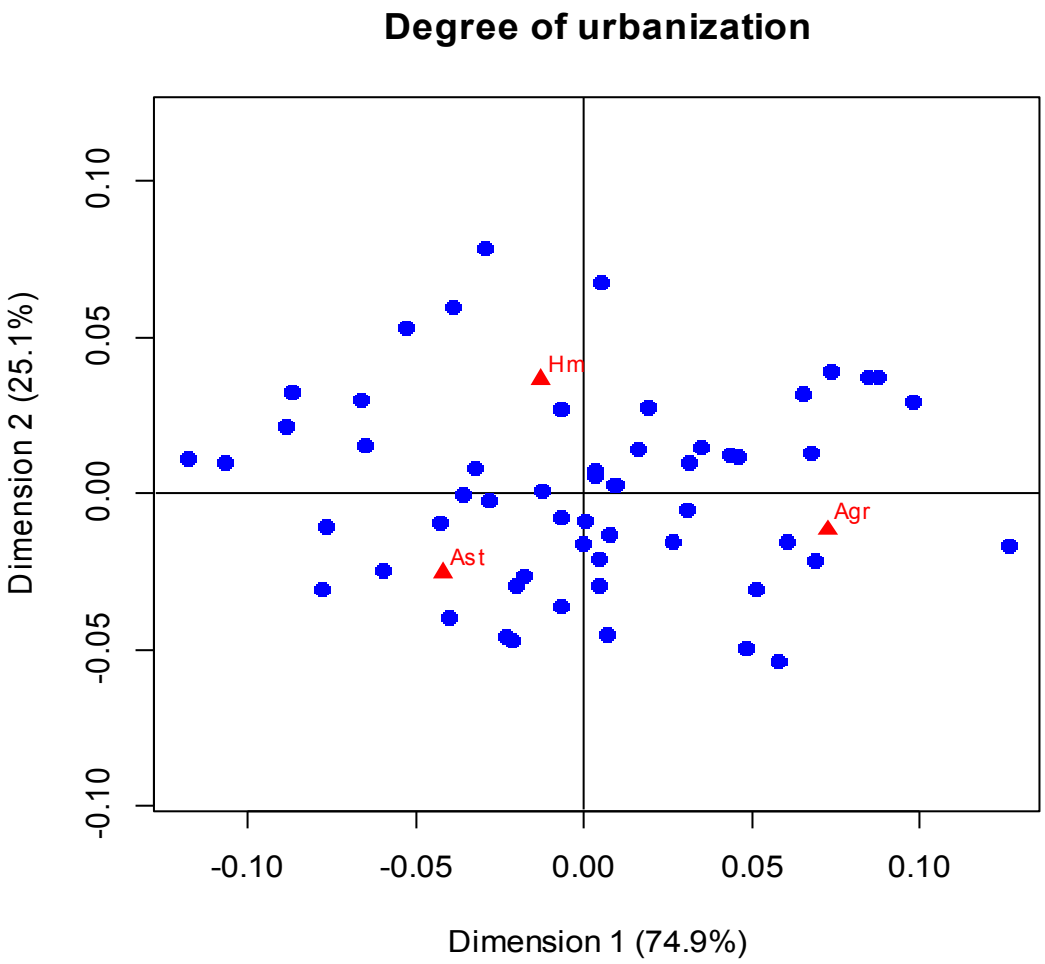
2.5.1 Κατασκευή στην R και παρουσίαση συμμετρικού διαγράμματος

Για την κατασκευή της διαγραμματικής απεικόνισης της ανάλυσης αντιστοιχειών στην R θα χρησιμοποιηθεί η εντολή plot του πακέτου ca. Αρχικά θα δημιουργήσουμε την απλή, συμμετρική απεικόνιση των σημείων όπως αυτά έχουν προκύψει από την ανάλυση αντιστοιχειών. Το γράφημα που παράγει το πακέτο περιέχει τις κατηγορίες τόσο των γραμμών όσο και των στηλών πάνω στο ίδιο διάγραμμα. Τα γραφήματα αυτού του είδους ονομάζονται συμμετρικά διγραφήματα (symetrical biplot) και παρουσιάζουν με διαφορετικό σύμβολο και χρώμα τις κατηγορίες των γραμμών και τις κατηγορίες των στηλών. Επίσης σε αυτή τη μορφή διαγράμματος οι μετρήσεις είναι κλιμακοποιημένες κατά τέτοιο τρόπο ώστε τα σημεία, είτε των γραμμών είτε των στηλών, που έχουν μεγάλη μάζα να μην επηρεάζουν σημαντικά το γράφημα.

Πληκτρολογούμε λοιπόν στο παράθυρο εντολών της R

```
plot(x)
title(main="Degree of urbanization")
```

οπότε προκύπτει το Διάγραμμα 2.2.



Διάγραμμα 2.2: Συμμετρικό διγράφημα (symetric biplot) του πακέτου "ca" για τον πίνακα σχετικών συχνοτήτων των ανέργων για 56 τρίμηνα από το 2001 ως το 2014 ανά βαθμό αστικότητας

Οι 2 ορθογώνιοι άξονες είναι τεχνητοί όπως έχουμε αναφέρει για αυτό και τα πρόσημα αλλά και οι ίδιες οι τιμές στους άξονες δεν έχουν τόσο μεγάλη σημασία. Αυτό το οποίο πρέπει να παρατηρηθεί είναι η διαφορά πρόσημων ανάμεσα σε παρατηρήσεις και όχι το ίδιο το πρόσημο.

Το Διάγραμμα 2.2 απεικονίζει με μπλε βούλες τα διάφορα τρίμηνα από το 2001 ως το 2014 στα οποία έχει καταμετρηθεί το ποσοστό της ανεργίας και με κόκκινα τρίγωνα την κατηγορία αστικότητας.
Ο πρώτος άξονας, δηλαδή ο οριζόντιος είναι και ο πιο σημαντικός καθώς εξηγεί την πλειοψηφία της συνολικής αδράνειας, το 74.9% αυτής (ο οριζόντιος άξονας του διαγράμματος). Μπορεί να δει κανείς ότι σύμφωνα με αυτό τον άξονα υπάρχει μία κατάταξη των τιμών ανάμεσα στο ποσοστό των ανέργων σε αστικές περιοχές σε ημιαστικές και σε αγροτικές περιοχές. Δηλαδή οι τρεις κατηγορίες φαίνονται διατεταγμένες ως προς τον οριζόντιο άξονα για αυτό και μπορούμε να θεωρήσουμε ότι ο άξονας αυτός που είναι μία από τις καινούριες μεταβλητές που έχουν προκύψει από την ανάλυση αντιστοιχιών δείχνει το βαθμό αστικοποίησης του ποσοστού των ανέργων. Το ποσοστό των ανέργων στις

αστικές περιοχές της χώρας είναι λίγο πιο αριστερά σαν κατηγορία από τις άλλες δύο με το ποσοστό στις ημιαστικές περιοχές να εμφανίζεται αργότερα και σαν τελευταία τιμή έχουμε το ποσοστό των ανέργων στις αγροτικές περιοχές. Αυτή η κατάταξη ταιριάζει με τις τιμές των μέσων όρων των τριών κατηγοριών.
Τα μπλε σημεία που βρίσκονται κοντύτερα είναι τα τρίμηνα που έχουν κοντινά προφίλ, δηλαδή τα τρίμηνα στα οποία παρατηρούνται παρόμοια ποσοστά ανεργίας συνολικά.
Εύκολα μπορεί να δει κανείς ότι οι αγροτικές περιοχές συγκεντρώνουν λιγότερες κουκίδες γύρω τους σε σχέση με τις άλλες δύο κατηγορίες, κάτι που οδηγεί στο συμπέρασμα ότι λιγότερα τρίμηνα έχουν μεγαλύτερη σχετική συχνότητα ανέργων στις άλλες δύο κατηγορίες και όχι στις αγροτικές περιοχές.

2.5.2 Κατασκευή στην R και παρουσίαση ασύμμετρου διαγράμματος

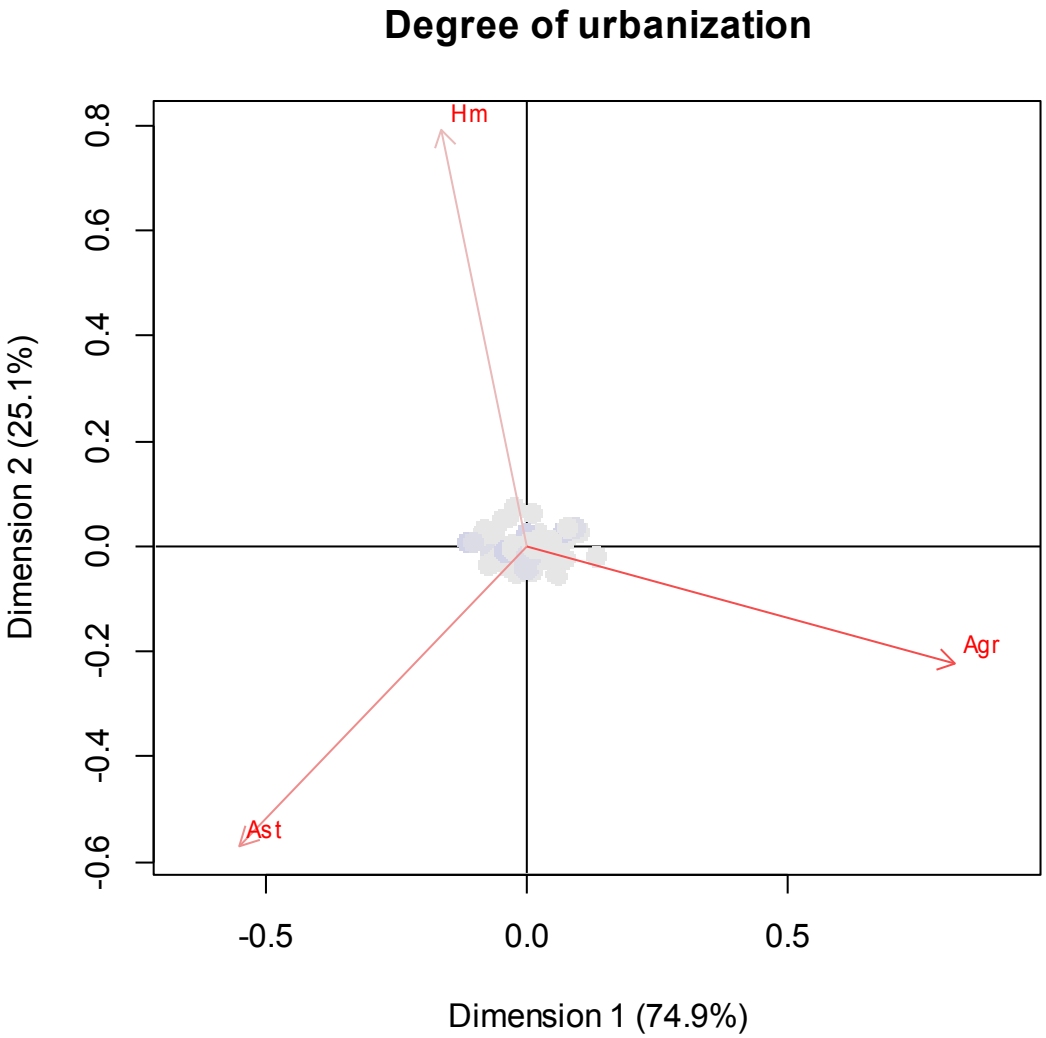
Στο ασύμμετρο διάγραμμα της ανάλυσης αντιστοιχειών και πάλι οι άξονες είναι ίδιοι με αυτούς του συμμετρικού διαγράμματος, είναι δηλαδή οι νέες μεταβλητές που εξηγούν το 74.9% και το 25.1% της αδράνειας. Σε αυτό το διάγραμμα αναπαριστάται και η μάζα κάθε τριμήνου σε μορφή σημείων. Επίσης οι στήλες του Πίνακα 2.1, δηλαδή η κατηγορία αστικότητας στην οποία έχουμε μεγαλύτερα ποσοστά ανέργων, αναπαρίστανται με τη μορφή βελών. Η ένταση του χρώματος στο διάγραμμα αφορά την έμφαση στην συνεισφορά των γραμμών, δηλαδή το πόσα τρίμηνα υπάρχουν στην συγκεκριμένη περιοχή.

Για την κατασκευή του ασύμμετρου διαγράμματος στην R θα γίνει χρήση και πάλι της εντολής plot μέσα από το πακέτο ca, χρησιμοποιώντας άλλα ορίσματα μες στην εντολή.
Συγκεκριμένα η εντολή θα είναι:

```
plot(ca(xdata), mass=TRUE, contrib="absolute", map="rowgreen",arrows=c(FALSE,TRUE))
title(main="Degree of urbanization")
```

Σε αυτή την εντολή πέρα από τον πίνακα των δεδομένων σαν όρισμα έχουν δοθεί και τα ακόλουθα.
Αρχικά με το όρισμα “mass=TRUE” δίνεται η εντολή για να εμφανιστούν οι μάζες στο διάγραμμα.
Με το “contrib=absolut” θα εμφανιστούν οι ολικές συχνότητες και όχι οι σχετικές οπότε θα έπρεπε να δοθεί η εντολή “contrib=relative” με διαφορετικό βάθος χρώματος.
Το όρισμα “map=rowgreen” είναι για το είδος του διαγράμματος. Η προεπιλογή της R είναι το “symmetric” για την κατασκευή του συμμετρικού διαγράμματος που παρουσιάστηκε παραπάνω. Η επιλογή rowgreen είναι για την κατασκευή ασύμμετρου διαγράμμτος με τις γραμμές να παίζουν τον σημαντικό ρόλο στην κατασκευή των κυρίων αξόνων και τις στήλες ως κανονικοποιημένες συντεταγμένες επί την ρίζα του αριθμού της αντίστοιχης μάζας.
Το όρισμα “arrows” είναι ένα διάνυσμα στο οποίο ορίζεται αν θα εμφανιστούν με τη μορφή σημείων ή βελών οι παρατηρήσεις των γραμμών και των στηλών. Στην πρώτη θέση του διανύσματος καθορίζεται το είδος εμφάνισης των γραμμών και στο δεύτερο το είδος

εμφάνισης των στηλών.
Η τελευταία γραμμή όπως και στα παραπάνω διαγράμματα καθορίζει τον τίτλο και τον υπότιτλο του διαγράμματος.



Διάγραμμα 2.3: Ασύμετρο διγράφημα (asymetric biplot) του πακέτου "ca" για τον πίνακα σχετικών συχνοτήτων των ανέργων για 56 τρίμηνα από το 2001 ως το 2014 ανά βαθμό αστικότητας

Στο Διάγραμμα 2.3 τα τρίμηνα συγκεντρώνονται σαν ένα νέφος κοντά στο κέντρο του διαγράμματος ενώ ο διαφορετικός βαθμός αστικοποίησης φαίνεται να εκτείνεται αρκετά μακριά. Το βελάκι για τις αστικές και ημιαστικές περιοχές φαίνεται να είναι λίγο πιο μακρύ από αυτό που αντιστοιχεί στις αγροτικές περιοχές χωρίς όμως να είναι τόσο εμφανής η διαφορά αυτή τη φορά.
Στο μη συμμετρικό διάγραμμα οι αποστάσεις ανταποκρίνονται πολύ περισσότερο στην πραγματικότητα σε σχέση με το συμμετρικό γράφημα στο οποίο έχει γίνει η απαραίτητη κλιμακοποίηση.
Μπορεί λοιπόν κανείς να συμπεράνει ότι τα ποσοστά των ανέργων ανά τρίμηνο φαίνεται να διαφέρουν αλλά είναι συγκεντρωμένα σε μία σφαίρα παρατηρήσεων και δεν είναι

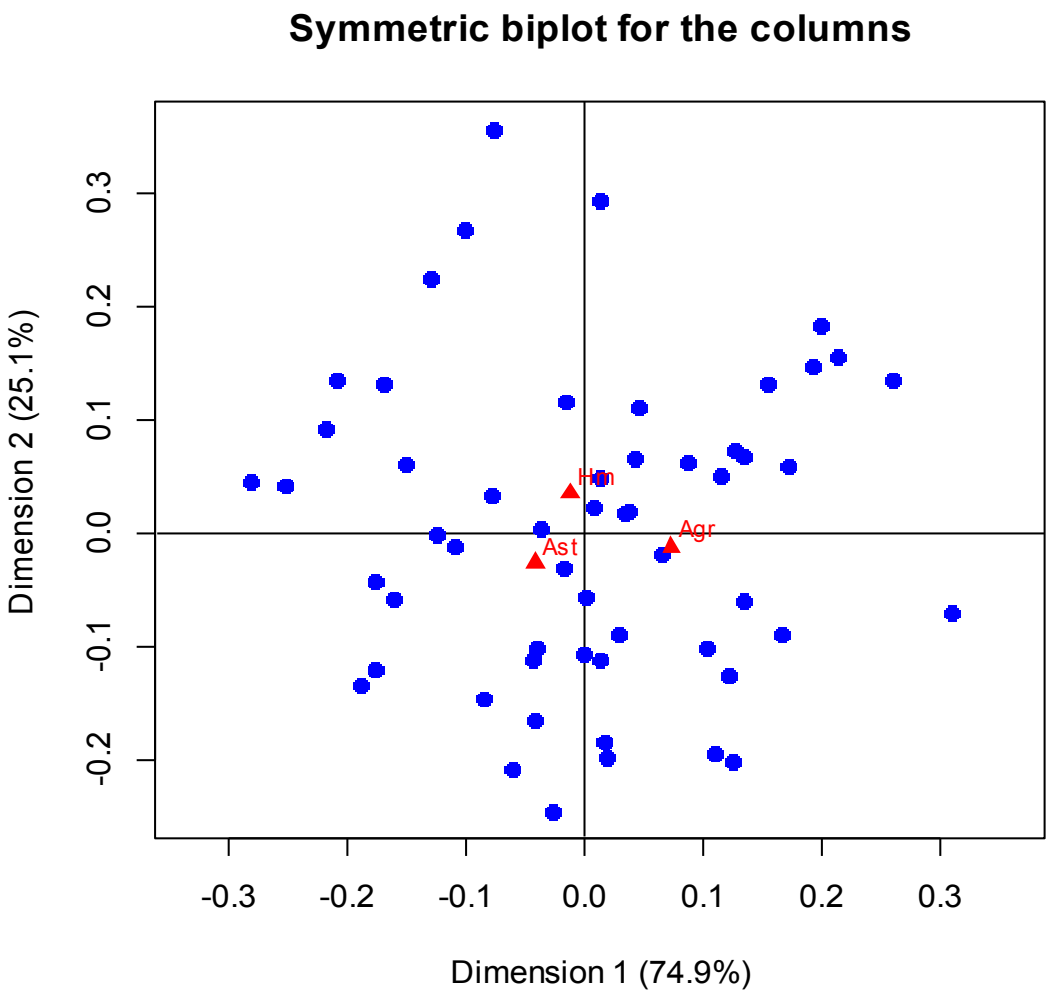
διασπαρμένα στους άξονες, παρά τις διαφορές τους στις τιμές δηλαδή δεν υπάρχουν ποσοστά που να κυμαίνονται πάνω από 50% για κανένα τρίμηνο και σε καμιά κατηγορία βαθμού αστικότητας. Στο ασύμμετρο διάγραμμα επίσης γίνεται πιο αισθητή η έννοια της απόστασης των σχετικών συχνοτήτων ανά τρίμηνο που φαίνεται πως βρίσκονται σε μία σφαίρα γύρω από το κέντρο, δηλαδή το 0 και πριν το 0.5 και το μέρος προς το οποίο τείνουν να βρεθούν δείχνει κοντά σε ποιον άξονα είναι οι παρατηρήσεις, δηλαδή σε ποιά κατηγορία βαθμού αστικοποίησης ανήκει κάθε τρίμηνο. Ενώ λοιπόν οι αποστάσεις είναι πολύ πιο ρεαλιστικές στο ασύμμετρο διγράφημα η κατηγορία βαθμού αστικοποίησης που συγκεντρώνει μεγαλύτερα ποσοστά ανέργων για τα περισσότερα τρίμηνα των ετών 2001-2014 δεν είναι εμφανής.

2.5.3 Κατασκευή στην R του συμμετρικού διαγράμματος για τις στήλες του πίνακα συχνοτήτων

Για μια ακόμη καλύτερη εικόνα των σχετικών συχνοτήτων των τριμήνων σε σχέση με τον βαθμό αστικότητας μπορεί κανείς να δώσει άλλα ορίσματα στην εντολή plot και να κατασκευάσει περαιτέρω διαγράμματα.
Η κύρια διαφορά στο είδος του διαγράμματος εξαρτάται από το όρισμα που θα δοθεί στην τιμή "map" καθώς τα άλλα ορίσματα μέσα στο plot αφορούν κυρίως αισθητικές διαφορές, όπως τα χρώματα ή υπαρξη βελών αντί για σημεία. Από τις διάφορες τιμές που μπορεί να πάρει το όρισμα "map" όμως μπορεί κανείς να δει τελείως διαφορετικά διαγράμματα καθώς έχει τη δυνατότητα να αλλάξει τους άξονες αλλά και να δημιουργήσει το αντίστοιχο γράφημα με κύριες συντεταγμένες τις γραμμές για τις στήλες.
Για παράδειγμα με την εντολή

```
plot(ca(xdata),map="colgreen",arrows=c(FALSE,FALSE))
title(main="Symmetric biplot for the columns")
```

προκύπτει το ανάλογο διάγραμμα για τον Πίνακα 2.1 μόνο που τώρα έχει υπολογιστεί η μάζα και η αδράνεια των στηλών και όχι των γραμμών και το αποτέλεσμα που προκύπτει μπορεί να δει κανείς πως είναι αρκετά διαφορετικό καθώς οι βούλες που αντιπροσωπεύουν τα τρίμηνα δεν συγκεντρώνονται στο κέντρο αλλά σκορπίζονται προς τα έξω και στο κέντρο φαίνεται να μαζεύονται οι 3 κατηγορίες του βαθμού αστικότητας. Δηλαδή η συμπεριφορά των σημείων φαίνεται να αντιστρέφεται αλλά το τελικό συμπέρασμα φαίνεται να διατηρείται, το γεγονός δηλαδή ότι οι αγροτικές περιοχές φαίνεται να βρίσκονται πιο μακριά από τις ημιαστικές και τις αστικές και συγκεντρώνουν σχετικά λιγότερα τρίμηνα γύρω τους. Μόνο που τώρα δεν είναι τόσο ξεκάθαρο και λόγω της πληθώρας των σημείων που αντιστοιχούν στα τρίμηνα δεν μπορεί να εξαχθεί κάποιο άλλο συμπέρασμα. Το Διάγραμμα 2.4 μάλιστα φαίνεται να μοιάζει αρκετά και με το συμμετρικό Διάγραμμα 2.2 που είχαμε κατασκευάσει νωρίτερα μόνο που οι αποστάσεις των κατηγοριών των στηλών ήταν μεγαλύτερες.
Το συμμετρικό διγράφημα όπως μπορεί να δει κανείς στο Διάγραμμα 2.4 είναι κάτι ενδιάμεσο σε εκείνο με κύριες συνιστώσες τις γραμμές και σε αυτό με τις στήλες.



Διάγραμμα 2.4: Συμμετρικό διγράφημα (symetric biplot) του πακέτου "ca" για τον πίνακα σχετικών συχνοτήτων 2.1 με έμφαση στις στήλες και όχι τις γραμμές

Εκτός από το πακέτο "ca" που χρησιμοποιήθηκε για τις παραπάνω γραφικές μπορεί να χρησιμοποιηθούν και άλλα πακέτα για να πραγματοποιηθεί η απλή ανάλυση αντιστοιχειών και να δημιουργηθούν διαγράμματα παρόμοια με τα παραπάνω, κάποια από τα οποία θα παρουσιαστούν παρακάτω.

2.6 Το ποσοστό των ανέργων σε τρεις κατηγορίες αστικών κέντρων

Από τις προηγούμενες παραγράφους φάνηκε ότι η πληθώρα του αριθμού των ανέργων συγκεντρώνονται κυρίως στα αστικά κέντρα της χώρας αλλά και στις ημιαστικές περιοχές,

αλλά τα βασικά αστικά κέντρα παρουσιάζουν ιδιαίτερο ενδιαφέρον λόγω του ότι συγκεντρώνουν πολύ μεγαλύτερο μέρος του πληθυσμού της χώρας για αυτό και θα παρουσιαστούν και κάποια διαγράμματα για την απεικόνιση της πορείας της ανεργίας στα βασικότερα από αυτά. Η ΕΛ.ΣΤΑΤ. στη βάση δεδομένων της περιέχει επίσης πίνακες αριθμού των ανέργων και των ποσοστών ανά το εργατικό δυναμικό στην περιφέρεια της πρωτεύουσας Αθήνας, στο πολεοδομικό συγκρότημα της Θεσσαλονίκης και σε λοιπές αστικές περιοχές για κάθε τρίμηνο από το 2001 ως το 2014.

Χρονιά και αριθμός τριμήνου	Κατηγορίες αστικών κέντρων		
	Αθήνα	Θεσσ/κη	Άλλα
2001a	10,9	11,3	14,5
2001b	10,4	10,8	13,4
2001c	10,3	10,9	13,4
2001d	10,1	11,3	14,6
2002a	9,8	12,1	14,8
2002b	9,2	11,4	13,1
2002c	9	12,7	13,1
2002d	8,8	11,8	13,7
2003a	8,6	10	14
2003b	8,5	9,7	12,8
2003c	8,8	10,3	12,4
2003d	9,1	11,1	13
2004a	8,1	11,7	13,4
2004b	8,1	11,5	12,2
2004c	9,4	12,5	12,2
2004d	9,4	12,2	12,1
2005a	9,3	12	12,2
2005b	9,2	11,6	11,1
2005c	9,6	11,5	11,5
2005d	9,6	10	11,1
2006a	8,6	9,9	11,4
2006b	8,5	9,7	10,3
2006c	8,5	9,6	9,8
2006d	9,1	9,2	10,4
2007a	8,6	9,1	10,4
2007b	7,9	9	9,5
2007c	7,9	9,4	9,5
2007d	7,2	9,8	9,6
2008a	6,8	9,1	10,2
2008b	6,1	8,9	9,2
2008c	6,5	9,3	9,1
2008d	7	9,3	10,1
2009a	7,6	11,6	11,2
2009b	8,1	11,6	10,3
2009c	9,5	13,4	10,8
2009d	10,2	12,4	11,8
2010a	10,5	13	13,6
2010b	11,4	14,4	13,4
2010c	12,8	16,2	14,1
2010d	14,1	28	16,1
2011a	14,7	20,4	17,8
2011b	15,7	22,1	28,6
2011c	18,5	22,4	20,4
2011d	21,8	25,5	22,6
2012a	22,9	27,2	24,2
2012b	23,7	28,8	26,1
2012c	26,7	29	27
2012d	27,2	30,6	28,1
2013a	28,7	31,5	29,5
2013b	28,4	31,4	29,7
2013c	28,7	31,6	29,5
2013d	28,8	31,3	29,4
2014a	28,2	30,5	29,5
2014b	27,6	30,6	27,8
2014c	27,1	29,7	26,8
2014d	26,9	28,8	28

Πίνακας 2.2: Πίνακας σχετικών συχνοτήτων επί 100 για τους ανέργους σε τρεις κατηγορίες αστικών κέντρων για 56 τρίμηνα από το 2001 ως το 2014

Η εκχώρηση της στην R γίνεται με όμοιο τρόπο όπως και στα προηγούμενα.

```
data1<-read.xls("../Desktop/diplwmatikh-domh/kef2/astika_kentra.xls",sheet=1)
```

```
A<-(data1)[,2]
B<-(data1)[,3]
C<-(data1)[,4]

xdata1<-data.frame(A,B,C)
```

Ακολουθεί μία γενική περιγραφή των στοιχείων.

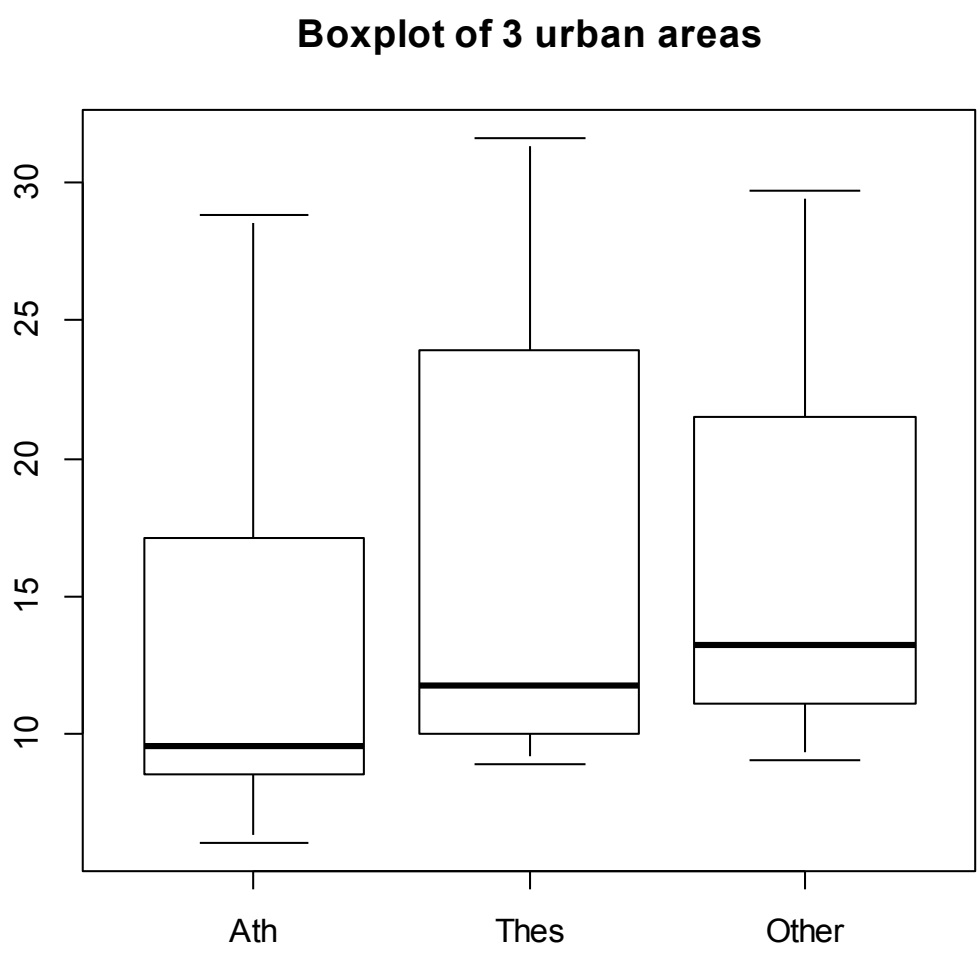
```
 summary(xdata2)
```

A	B	C
Min. : 6.100	Min. : 8.90	Min. : 9.10
1st Qu.: 8.575	1st Qu.:10.00	1st Qu.:11.10
Median : 9.550	Median :11.75	Median :13.25
Mean :13.548	Mean :16.26	Mean :16.15
3rd Qu.:16.400	3rd Qu.:23.18	3rd Qu.:20.95
Max. :28.800	Max. :31.60	Max. :29.70

```
 sd(A)
[1] 7.662684
 sd(B)
[1] 8.291366
 sd(C)
[1] 7.063556
```

και ένα θηκόγραμμα των τριών αυτών κατηγοριών αστικών κέντρων.

```
 boxplot(xdata2,names=c("Ath","Thes","Other"))
 title(main="Boxplot of 3 urban areas")
```

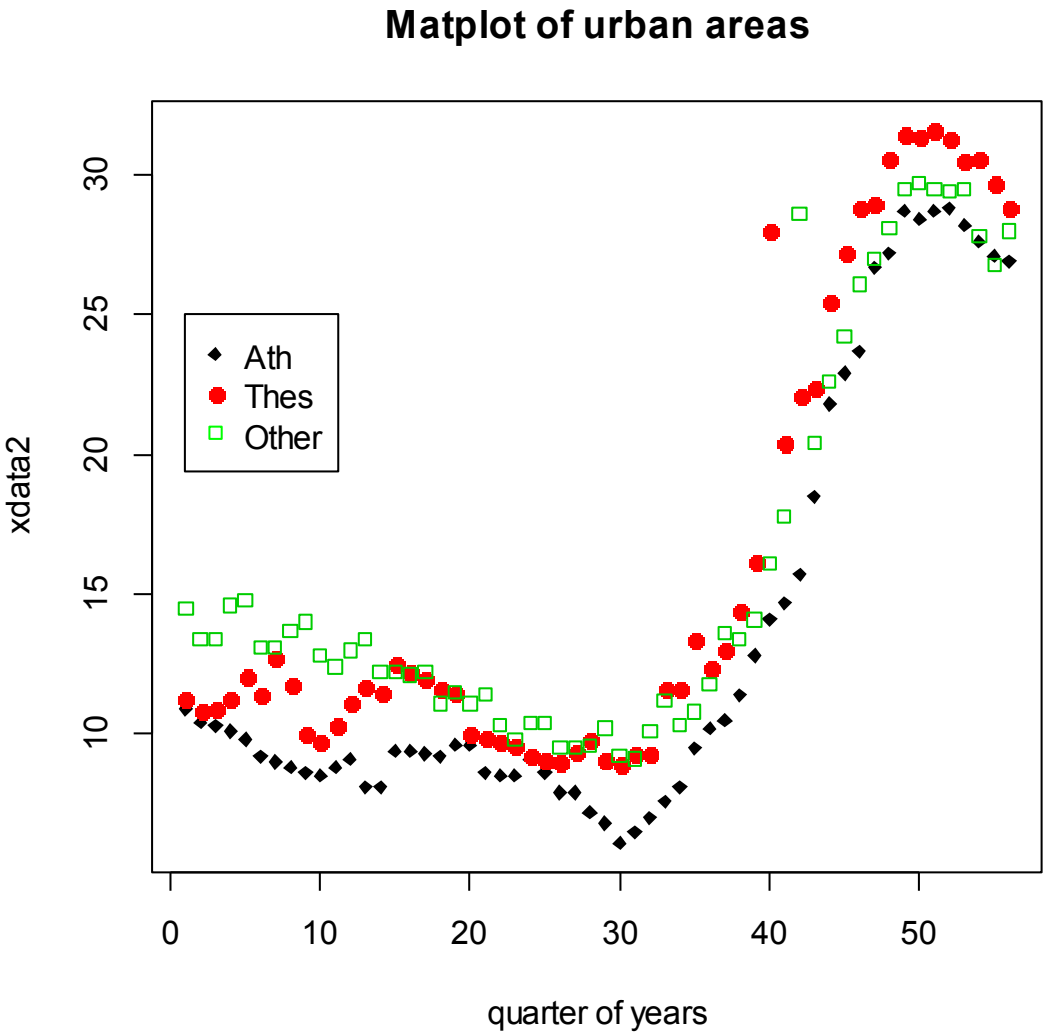


Διάγραμμα 2.5: Θηκόγραμμα της σχετικής συχνότητας επί 100 των ανέργων βασικών αστικών κέντρων για 56 τρίμηνα από το 2001 ως το 2014

Τα θηκογραφήματα στο Διάγραμμα 2.5 δείχνουν ότι το πολεοδομικό συγκρότημα της Θεσσαλονίκης παρουσιάζει κατά μέσο όρο μεγαλύτερα ποσοστά ανέργων στο σύνολο του εργατικού δυναμικού της από την περιφέρεια Αττικής, ενώ οι υπόλοιπες αστικές περιοχές φαίνεται να έχουν ακόμα μεγαλύτερο μέσο όρο ποσοστών. Το πολεοδομικό συγκρότημα της Θεσσαλονίκης μάλιστα φαίνεται να παρουσιάζει και μεγαλύτερες διαφοροποιήσεις στα ποσοστά των ανέργων. Αυτό μπορεί να οφείλεται στην συγκέντρωση των περισσότερων εταιριών και βιομηχανιών στην περιφέρεια της πρωτεύουσας της χώρας οι οποίες και απασχολούν περισσότερους εργαζομένους.

Ένα κοινό διάγραμμα διασποράς των τριών κατηγοριών για την πορεία τους για τα τρίμηνα των έτων 2001 ως 2014 θα δώσει μία καλύτερη εικόνα για τα ποσοστά της ανεργίας στις τρεις κατηγορίες αστικών κέντρων. Η δημιουργία του κοινού διαγράμματος διασπορών θα γίνει με τις ακόλουθες εντολές:

```
matplot(xdata1,pch=c(18,16,22))
legend(1,25,cex=1,col=c("black","red","green"),pch=c(18,16,22),legend=c("Ath","Thes","Other"))
title(main="Matplot of urban areas", xlab="quarter of years")
```



Διάγραμμα 2.6: Κοινό διάγραμμα διασποράς για τις τρεις κατηγορίες αστικών περιοχών μαζί για το σύνολο των 56 τριμήνων από το 2001 ως το 2014

Ο οριζόντιος άξονας παίρνει τιμές από το 1 ως το 56 που είναι ο αριθμός του κάθε τριμήνου από το δείγμα σε σειρά.
Από το Διάγραμμα 2.6 γίνεται εμφανής η διαφοροποίηση ως προς την σχετική συχνότητα των ανέργων μεταξύ των τριών κατηγοριών αστικών κέντρων. Η μαύρη γραμμή που είναι η περιφέρεια Αττικής συγκεντρώνει χαμηλότερα ποσοστά ανέργων στην πάροδο του χρόνου. Η μορφή των τριών καμπυλών παρουσιάζει ενδιαφέρον και φαίνεται να είναι σχετικά όμοια στην πορεία της για τις τρεις κατηγορίες. Παρουσιάζεται μετά το $30^{ο}$ τρίμηνο περίπου, δηλαδή περίπου το 2008 μία άυξηση του αριθμού και στις τρεις αστικές περιοχές. Στην περιφέρεια Αττικής αυτή η αύξηση είναι πιο απότομη σε σχέση με τις άλλες δύο κατηγορίες αστικών περιοχών, κάτι που επιβεβαιώνει το γεγονός ότι στην πρωτεύουσα όπου

συγκεντρώνεται μεγάλη ποσότητα του πληθυσμού (περίπου ο μισός πληθυσμός της χώρας) η αλλαγή από την οικονομική ύφεση είναι πιο εμφανής και για αυτό η συχνότητα των ανέργων είναι πολύ μεγαλύτερη. Στην πάροδο των τριμήνων όμως και πάλι στην περιφέρεια της πρωτεύουσας φαίνεται τα ποσοστά ανέργων να είναι χαμηλότερα τελικά αναλογικά και με το εργατικό δυναμικό τους ενώ το πολεοδομικό συγκρότημα της Θεσσαλονίκης φαίνεται να παρουσιάζει τα τελευταία τρίμηνα από το 2008 και μετά μεγαλύτερα ποσοστά ανέργων και από τις τρεις κατηγορίες αστικών κέντρων. Οι υπόλοιπες αστικές περιοχές φαίνεται να παρουσιάζουν μεγαλύτερα ποσοστά ανέργων τα πρώτα τρίμηνα πριν το 2005 ακόμα από τις άλλες δύο κατηγορίες ενώ στη συνέχεια φαίνεται η κατάσταση να εξομαλύνεται και τα ποσοστά ανεργίας να μην διαφέρουν τόσο πολύ.

3. Το ποσοστό των ανέργων ανά ηλικιακή ομάδα και ανά φύλο

Σε αυτό το κεφάλαιο θα γίνει μελέτη των ποσοστών των ανέργων ανάλογα με την ηλικιακή ομάδα στην οποία ανήκουν. Για την ΕΛ.ΣΤΑΤ. από την οποία προέρχονται τα δεδομένα η ηλικία από την οποία μπορεί να θεωρηθεί ένα άτομο εργαζόμενο είναι τα 15 έτη, οπότε και τελειώνει η υποχρεωτική εκπαίδευση. Τα δεδομένα που θα χρησιμοποιηθούν έχουν μετρηθεί ανά χρονιά από το 2001 ως το 2014. Η μελέτη θα γίνει με τη μέθοδο της ανάλυσης αντιστοιχιών και πάλι αλλά αντί για το πακέτο "ca" της R που χρησιμοποιήθηκε στο Κεφάλαιο 2, θα χρησιμοποιηθούν άλλα πακέτα που δίνουν τη δυνατότητα κατασκευής περισσοτέρων διαγραμμάτων. Το πακέτο "anacor" καθώς και το πακέτο "languageR" είναι πακέτα που εφαρμόζουν την ανάλυση αντιστοιχιών και θα γίνει επίσης αναφορά στο πακέτο "MASS" που όμως δεν δίνει τη δυνατότητα κατασκευής πολλών διαγραμμάτων. Επίσης θα παρουσιαστεί και ο πίνακας σχετικών συχνοτήτων του αριθμού των ανέργων ανά φύλο ώστε να εξεταστεί αν υπάρχει διάκριση στα ποσοστά ανεργίας σε αυτόν τον τομέα. Οι σχετικές συχνότητες έχουν προκύψει από τον αριθμό των εργαζομένων ανά κατηγορία στο σύνολο του εργατικού δυναμικού της εκάστοτε κατηγορίας.

3.1 Παρουσίαση των δεδομένων για την σχετική συχνότητα των ανέργων ανά φύλο και η περιγραφή τους

Ο πίνακας σχετικών συχνοτήτων στην ομάδα δεδομένων για το φύλο των ανέργων περιέχει τα ποσοστά ανέργων επί 100 από το σύνολο του εργατικού δυναμικού σε κάθε μία από τις 2 κατηγορίες φύλου σε δύο στήλες για τους άνεργους άντρες και γυναίκες με τους τίτλους "M" και "W" αντίστοιχα.

		Φύλο	
		M	**W**
χρονιά	**2001**	7,25	16,25
	2002	6,8	15,8
	2003	6,3	15,1
	2004	6,7	16,3
	2005	6,2	15,5
	2006	5,7	13,8
	2007	5,2	12,9
	2008	5,1	11,5
	2009	7	13,3
	2010	10,1	16,3
	2011	15,2	21,5
	2012	21,6	28,2
	2013	24,5	31,3
	2014	23,6	30,2

Πίνακας 3.1: Σχετικές συχνότητες των ανέργων ανά φύλο επί 100 για τα έτη 2001 ως 2014

Όπως και στις προηγούμενες περιπτώσεις η εισαγωγή των δεδομένων του Πίνακα 3.1στην R θα γίνει με την εντολή read.xls του πακέτου gdata και με την ακόλουθη εντολή.

```
sdata<-read.xls("../Desktop/diplwmatikh-domh/kef3/fulo.xls",sheet=1,header=TRUE)
```

Στη συνέχεια ορίζεται η κάθε στήλη ως διάνυσμα.

```
M<-(sdata)[,2]
W<-(sdata)[,3]
```

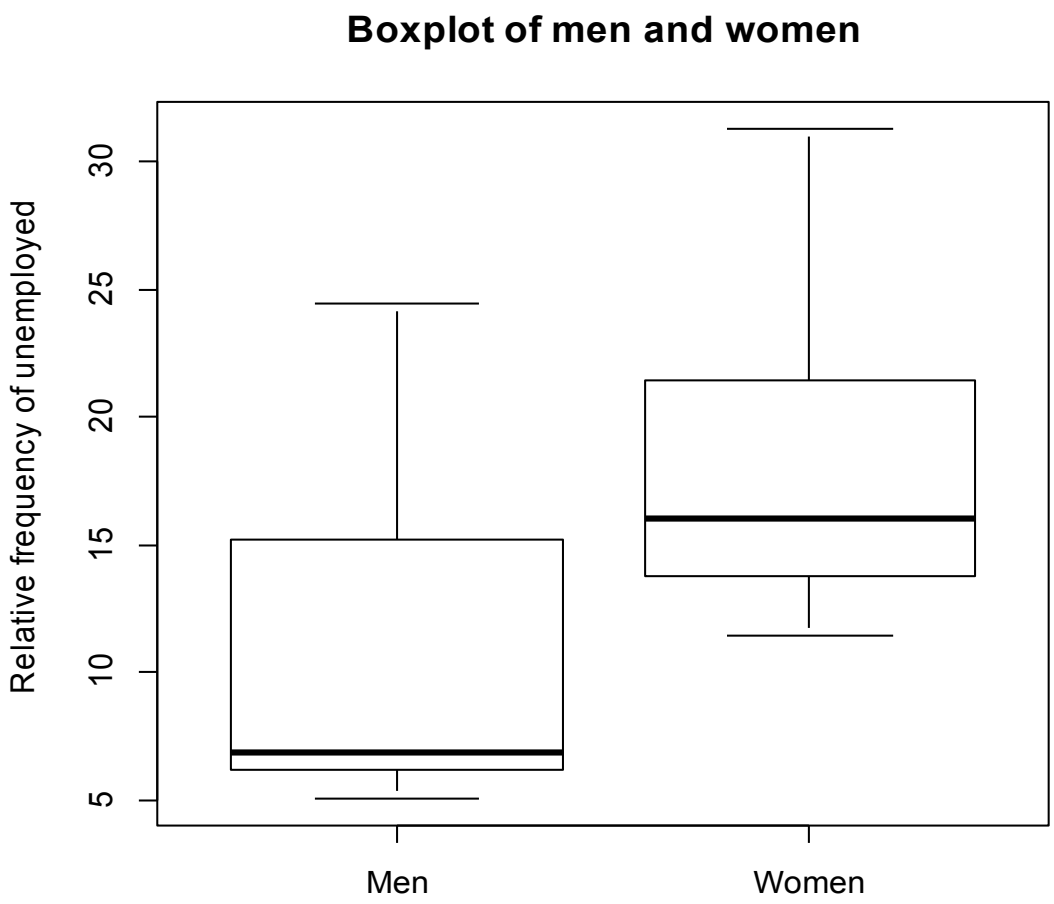
Ακολουθούν η τυπική απόκλιση και τα αποτελέσματα της εντολής summary για κάθε φύλο.

```
summary(M)
   Min. 1st Qu.  Median    Mean   3rd Qu.    Max.
  5.100   6.225   6.900   10.800   13.920   24.500
sd(M)
[1] 7.227843

summary(W)
   Min. 1st Qu.  Median    Mean   3rd Qu.    Max.
  11.50   14.13   16.02    18.42    20.20    31.30
sd(W)
[1] 6.655384
```

Όπως μπορεί να δει κανείς εύκολα από τους μέσους όρους υπάρχει διαφοροποίηση στο ποσοστό των ανέργων ανά φύλο με τις γυναίκες να έχουν πολύ μεγαλύτερο μέσο όρο. Η τυπική απόκλιση δεν παρουσιάζει σημαντικές διαφορές ανάμεσα στις δύο κατηγορίες αν και στους άντρες είναι λίγο μεγαλύτερη. Τέλος οι μέγιστες και οι ελάχιστες τιμές και στις δύο περιπτώσεις διαφέρουν αρκετά, αλλά το εύρος των τιμών είναι ανάλογο τόσο στους άντρες όσο και στις γυναίκες όπως και το ενδοτεταρτημοριακό εύρος, κάτι που προδίδει πως ανεξαρτήτως διαφοροποιήσεων των ποσοστών ανά φύλο για τα έτη 2001 ως 2014 οι διαφοροποιήσεις γύρω από τον μέσο όρο και στις δύο περιπτώσεις είναι ανάλογες, δηλαδή η πορεία των ποσοστών με την πάροδο των ετών φαίνεται να είναι όμοια για τις δύο περιπτώσεις.

Ακολουθεί το Διάγραμμα 3.1.

```
boxplot(M,W, names=c("Men","Women"))
title(main="Boxplot of men and women",ylab="Relative frequency of unemployed")
```



Διάγραμμα 3.1: Θηκόγραμμα της σχετικής συχνότητας επί 100 των ανέργων ανά φύλο

Από το Διάγραμμα 3.1 φαίνεται ότι παρά τις διαφοροποιήσεις στον αριθμό των μέσων τιμών που είχε παρατηρηθεί παραπάνω το ενδοτεταρτημοριακό εύρος των δύο κατηγοριών δε διαφέρει πολύ οπότε η πλειοψηφία των τιμών, δηλαδή το 75% των παρατηρήσεων φαίνεται να απέχει αναλογικά από τις διαμέσους τόσο για τους άνεργους άντρες όσο και για τις άνεργες γυναίκες.

Στη συνέχεια θα σχηματιστεί γραφικά ο πίνακας σχετικών συχνοτήτων με τη βοήθεια χρωμάτων. Το διάγραμμα αυτό μπορεί να το δημιουργήσει κανείς με πολλούς τρόπους ένας από τους οποίους είναι το πακέτο "cape" που μπορεί να εγκατασταθεί στην R.

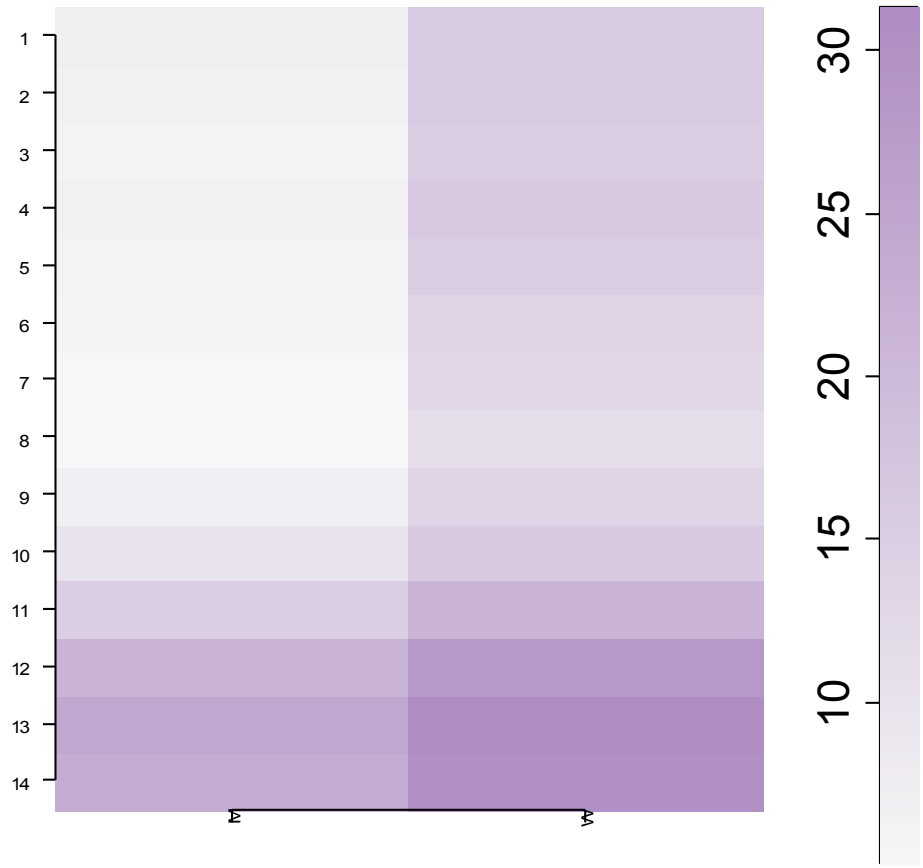
```
install.packages("cape")
library(cape)
```

Το πακέτο αυτό περιέχει την εντολή "myImagePlot" που δημιουργεί το επιθυμητό διάγραμμα και δέχεται σαν όρισμα έναν πίνακα. Πριν τη δημιουργία του διαγράμματος θα πρέπει πρώτα λοιπόν να δημιουργηθεί ο πίνακας σχετικών συχνοτήτων χωρίς τη στήλη με τις χρονιές.

```
sdata1<-data.frame(M,W)
```

Στη συνέχεια μπορούν να δοθούν οι εντολές για το σχηματισμό του Διαγράμματος 3.2.

```
myImagePlot(sdata1)
title(main="Matrix image of Men and Women")
```



Διάγραμμα 3.2: Η σχετική συχνότητα των ανέργων επί 100 ανά φύλο με χρωματική αναπαράσταση

Το Διάγραμμα 3.2 περιέχει δύο στηλές με διαφορετικές αποχρώσεις του μωβ ανάλογα με το μέγεθος της τιμής κάθε παρατήρησης που αντιστοιχούν στους άνεργους άντρες και γυναίκες. Στην κάθετη στήλη περιέχονται οι χρονιές. Ακριβώς δίπλα από το διάγραμμα υπάρχει μία στήλη που δείχνει τη διαβάθμιση των χρωμάτων. Για παράδειγμα το πολύ ανοιχτό μωβ σχεδόν άσπρο αντιστοιχεί σε τιμές κάτω από 10 μονάδες ενώ η πιο σκούρα απόχρωση αντιστοιχεί σε 30 μονάδες και πάνω. Αυτή η χρωματική κλίμακα βοηθάει στην ανάγνωση των αποτελεσμάτων.

Στο Διάγραμμα 3.2 φαίνεται ξεκάθαρα πλέον πως τα τελευταία χρόνια ο αριθμός των ανέργων είναι πολύ μεγαλύτερος σε σχέση με παλιά τόσο στους άντρες όσο και στις γυναίκες. Όπως είναι εμφανές τις χρονιές από το 2010 και πριν βλέπουμε έντονες διαφοροποιήσεις στα ποσοστά των ανέργων για τα δύο φύλα με τη δεξιά στήλη που αντιστοιχεί στις γυναίκες να εμφανίζει πιο σκουρόχρωμες περιοχές δηλαδή πιο υψηλές τιμές ανέργων σχεδόν για όλα τα έτη. Επίσης φαίνεται ότι τα τελευταία χρόνια τα ποσοστά ανέργων είναι πολύ μεγαλύτερα.

3.2 Παρουσίαση των δεδομένα για την ηλικιακή ομάδα των ανέργων

Οι πιθανές ηλικιακές ομάδες που έχουν χρησιμοποιηθεί από την ΕΛ.ΣΤΑΤ είναι 6 και ο διαχωρισμός τους έχει γίνει με βάση κοινωνικά κριτήρια κατά κύριο λόγο. Για παράδειγμα μία ηλικιακή ομάδα είναι από τα 30 ως τα 44 έτη οπότε και θεωρείτε ότι είναι το πιο παραγωγικό κομμάτι εργασίας για το άτομο και το χρονικό διάστημα στο οποίο βγάζει το περισσότερο εισόδημα. Η επόμενη ηλικιακή ομάδα είναι από τα 45 μέχρι τα 64 έτη, τα οποία θεωρούνταν η ηλικία συνταξιοδότησης για τις περισσότερες χρονιές του δείγματος. Οι κατηγορία ηλικιακής ομάδας στην οποία μπορεί να ανήκει κάποιος είναι τα

- 15-19 έτη
- 20-24 έτη
- 25-29 έτη
- 30-44 έτη
- 45-64 έτη
- 65+ έτη

Ακολουθεί ο Πίνακας 3.2 που περιέχει τις σχετικές συχνότητες των ανέργων ανά ηλικιακή ομάδα πολλαπλασιασμένες επί 100. Οι σχετικές συχνότητες έχουν προκύψει από τη διαίρεση του αριθμού των ανέργων ανά τον αριθμό του εργατικού δυναμικού στην κάθε ηλικιακή ομάδα για την εκάστοτε χρονιά.

Χρονιά	Ηλικιακές ομάδες					
	15-19	20-24	25-29	30-44	45-64	65 plus
2001	34,75	26,7	16,875	8,65	5,125	1,525
2002	33,35	25,35	16,675	8,5	4,85	1,4
2003	32,9	25,575	16,1	8,225	4,275	1,175
2004	33,9	25,175	16	9,225	5,3	1,1
2005	33,05	24,575	15,1	8,925	5,2	1,55
2006	31,75	23,825	13,95	8	4,525	1,2
2007	26,7	22	14,225	7,45	4,175	1,125
2008	26,2	21,15	13,05	6,9	3,975	0,825
2009	31,45	24,775	15,075	8,6	5,725	0,85
2010	39,25	32,125	19,725	11,725	7,8	1,35
2011	56,475	43,025	29,425	16,325	11,1	2,6
2012	65,8	53,575	37,375	23,025	16,575	4,475
2013	72,2	56,05	43,3	26	19,2	9,05
2014	61,525	51,15	40,8	25,4	19,5	10,75

Πίνακας 3.2: Η σχετική συχνότητα των ανέργων ανά ηλικιακή ομάδα επί 100 για τα έτη 2001 ως 2014

3.3 Περιγραφικά στοιχεία για τα δεδομένα για την ηλικιακή ομάδα των ανέργων από την R

Όπως και στις προηγούμενες περιπτώσεις η εισαγωγή των δεδομένων στην R θα γίνει με την εντολή read.xls του πακέτου gdata. Επειδή στο αρχείο excel ο τίτλος κάθε στήλης είναι αριθμητικές τιμές (για παράδειγμα "15-19") δεν αναγνωρίζονται κατευθείαν σαν τίτλοι των στηλών για την κάθε ηλικιακή ομάδα. Για αυτό και η εντολή χρειάζεται το όρισμα "header=TRUE" για να θεωρήσει την πρώτη γραμμή με αριθμητικές τιμές τίτλο και όχι κομμάτι των δεδομένων.

```
data<-read.xls("../Desktop/diplwmatikh-domh /kef4 /hlikiakh_omada.xls" ,sheet=1, header=TRUE)
```

Στη συνέχεια ορίζουμε κάθε στήλη ως ένα διάνυσμα με ονόματα απο το X1 ως το X6 που αντιστοιχούν από τη 2η ως την 7η στήλη αντίστοιχα για να είναι πιο εύκολη η χρήση τους αργότερα. Με X_i ορίζουμε τις στήλες με τα ποσοστά των ανέργων.

```
X1<-(data)[,2]
X2<-(data)[,3]
X3<-(data)[,4]
X4<-(data)[,5]
X5<-(data)[,6]
X6<-(data)[,7]
```

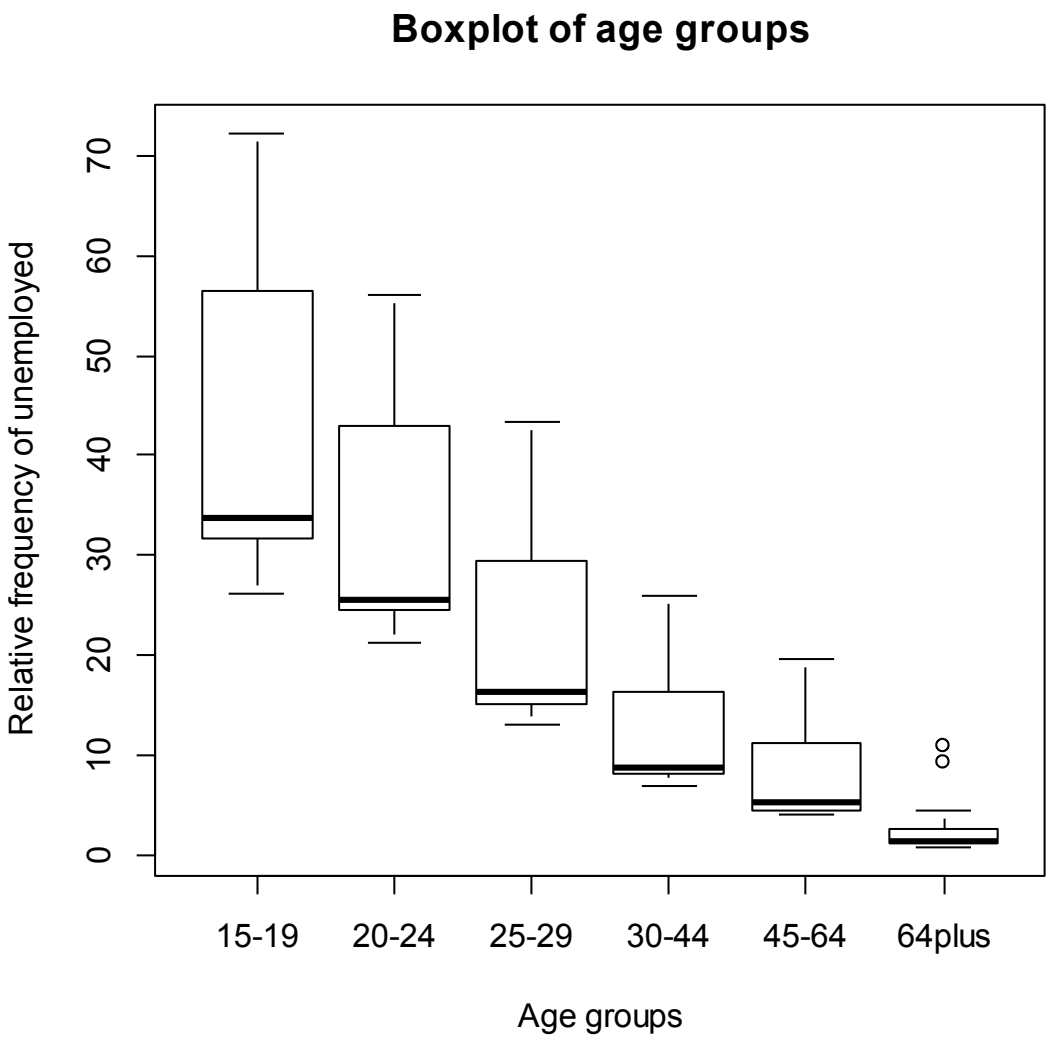
Ακολουθούν η τυπική απόκλιση και τα αποτελέσματα της εντολής summary για κάθε μία

από τις ηλικιακές ομάδες που έχουν οριστεί για την σχετική συχνότητα των ανέργων στην κάθε μία ηλικιακή κατηγορία.

```
 summary(X1)
   Min. 1st Qu.  Median    Mean   3rd Qu.    Max. 
  26.20   32.04   33.62    41.38    52.17    72.20 
 sd(X1)
[1] 15.51027
  summary(X2)
   Min. 1st Qu.  Median    Mean   3rd Qu.    Max. 
  21.15   24.62   25.46    32.50    40.30    56.05 
 sd(X2)
[1] 12.65118
  summary(X3)
   Min. 1st Qu.  Median    Mean   3rd Qu.    Max. 
  13.05   15.08    16.39    21.98   27.00    43.30 
 sd(X3)
[1] 10.84922
  summary(X4)
   Min. 1st Qu.  Median    Mean   3rd Qu.    Max. 
  6.900   8.294    8.788   12.640  15.180   26.000 
 sd(X4)
[1] 7.012747
  summary(X5)
   Min. 1st Qu.  Median    Mean   3rd Qu.    Max. 
  3.975   4.606    5.250    8.380   10.270   19.500 
 sd(X5)
[1] 5.778751
  summary(X6)
   Min. 1st Qu.  Median    Mean   3rd Qu.    Max. 
  0.825   1.138   1.375    2.784    2.338    10.750 
 sd(X6)
[1] 3.173896
```

Επίσης ακολουθεί ένα θηκόγραμμα (boxplot) για την οπτική παρουσίαση των δεδομένων.

```
boxplot(X1,X2,X3,X4,X5,X6,names=(c("15-19","20-24","25-29","30-44","45-64","64plus")))
title(main="Boxplot of age groups", xlab="Age groups",ylab="Relative frequency of unemployed")
```



Διάγραμμα 3.3: Θηκόγραμμα της σχετικής συχνότητας των ανέργων επί 100 ανά ηλικιακή ομάδα για τα έτη 2001 ως 2014

Όπως μπορεί να δει κανείς τόσο από το Διάγραμμα 3.3 αλλά και από τους μέσους όρους υπάρχει διαφορά στον αριθμό των ανέργων ανά ηλικιακή ομάδα. Για παράδειγμα ο μέσος όρος της ηλικιακής ομάδας 15-19 δεν μπορεί να θεωρηθεί ίδιος με αυτόν των ενδιάμεσων ηλικιακών ομάδων. Επίσης η μεταβλητότητα των τιμών ανά κατηγορία φαίνεται να διαφέρει αρκετά, με την μεγαλύτερη να είναι στην πρώτη ηλικιακή ομάδα και την μικρότερη στην τελευταία όπως και το ενδοτεταρτημοριακό εύρος που ακλουθεί και αυτό φθίνουσα σειρά από την πρώτη ως την τελευταία ηλικιακή ομάδα. Επίσης οι τυπικές αποκλίσεις ακολουθούν και αυτές την ίδια φθίνουσα σειρά στις τιμές τους, κάτι που σημαίνει πως η διαφοροποίηση των ποσοστών ανά έτος από τη μία ηλικιακή ομάδα στην άλλη γίνεται κατά ανάλογο τρόπο.

Ανεργία είναι η κατάσταση ενός ατόμου, που ενώ είναι ικανό, πρόθυμο και διαθέσιμο να απασχοληθεί, δεν δύναται να βρει εργασία. Η ηλικιακή ομάδα άνω των 64 περιλαμβάνει πληθώρα ατόμων που δεν εργάζονται αλλά η πλειοψηφία είναι είτε μη ικανά είτε μη πρόθυμα άτομα για εύρεση εργασίας. Επίσης στην ηλικιακή ομάδα αυτή πολλά άτομα έχουν περάσει το όριο συνταξιοδότησης και λαμβάνουν την αντίστοιχη σύνταξη που τους

αναλογεί οπότε δεν μπορούν να θεωρηθούν εργαζόμενοι πλέον. Για αυτό το λόγο και το σύνολο του εργατικού δυναμικού σε αυτή την κατηγορία είναι πολύ μικρότερο από αυτό στις άλλες ηλικιακές ομάδες ενώ ταυτόχρονα από αυτούς που ανήκουν στο εργατικό δυναμικό έχουν κάποια δουλειά που δε θέλουν να αφήσουν ακόμα ενώ έχουν την δυνατότητα. Παρατηρείται όμως η ύπαρξη κάποιων ακραίων τιμών, που κοιτώντας τον Πίνακα 3.2 μπορεί να διαπιστώσει κανείς ότι είναι οι 2 τελευταίες χρονιές από το σύνολο των δεδομένων. Τα έτη 2013 και 2014 παρατηρείται μεγάλη αύξηση του ποσοστού των ανέργων άνω των 64 ετών. Η ραγδαία αύξηση του ποσοστού αυτού μπορεί να θεωρηθεί προιόν της οικονομικής ύφεσης στην χώρα και να θεωρηθεί ότι οφείλεται στην προσπάθεια ατόμων μεγαλύτερα σε ηλικία, που υπό κανονικές συνθήκες δε θα έψαχναν για εργασία, να μπουν στη διαδικασία εύρεσης ώστε να βοηθήσουν την οικογένεια τους ή να μπορέσουν να συντηρηθούν καθώς τα έσοδα δεν επαρκούν. Ακόμα και οι ακραίες αυτές τιμές των ποσοστών όμως είναι μικρότερες από τον μέσο όρο του ποσοστού ανεργίας για την ηλικιακή ομάδα 15-19 ετών. Το ποσοστό των ανέργων σε αυτή την ηλικιακή ομάδα είναι πολύ μεγαλύτερο κατά μέσο όρο από ότι σε άλλες ηλικιακές ομάδες. Η υποχρεωτική εκπαίδευση στην Ελλάδα τελειώνει στα 15 έτη, πρόκειται για την απόκτηση πτυχίου 3 τάξεων μέσης εκπαίδευσης. Οπότε το άτομο μπορεί να μπει θεωρητικά στο σύνολο του εργατικού δυναμικού όμως η πραγματικότητα είναι ότι η πλειοψηφία των διαθέσιμων εργασιών χρειάζονται περαιτέρω μόρφωση. Επίσης οι απαιτήσεις στην αγορά εργασίας είναι πολύ υψηλότερες σε σχέση με το παρελθόν οπότε το πτυχίο της μέσης εκπαίδευσης είναι απαραίτητο για όλους σχεδόν τους τύπους εργασίας. Οπότε και αυτή η ηλικιακή ομάδα διαθέτει μεγάλο αριθμό ανέργων. Για την ηλικιακή ομάδα μεταξύ 45-64 ετών ισχύει ότι κατά κύριο λόγο έχουν εξασφαλίσει μία θέση εργασίας από το παρελθόν που έχει γίνει μόνιμη οπότε και τα ποσοστά είναι χαμηλότερα στις περισσότερες ηλικιακές ομάδες εκτός αυτής των ανέργων άνω των 64. Με μια ματιά όμως στην μέγιστη και στην ελάχιστη τιμή μπορεί να δει κανείς πολύ μεγάλη διαφοροποίηση που οδηγεί στο συμπέρασμα της αύξησης των ανέργων ατόμων σε αυτή την ηλικιακή ομάδα, στην οποία η επαναπρόσληψη ή η εύρεση νέας θέσης εργασίας είναι πολύ πιο δύσκολη. Τέλος οι υπόλοιπες ηλικιακές ομάδες έχουν ενδιάμεσα ποσοστά με χαμηλότερα στην ηλικία μεταξύ 30-44 ετών, που είναι και η πιο παραγωγική ηλικία για τον μέσο εργαζόμενο καθώς έχει τελειώσει όλη την δυνατή εκπαίδευση του και συνήθως έχει ήδη κάποια προϋπηρεσία στον τομέα του.

3.4 Η ανάλυση αντιστοιχιών με το πακέτο “anacor”

Ο τύπος διαγράμματος που προκύπτει από την ανάλυση αντιστοιχιών και μπορεί να σχεδιαστεί με το πακέτο “ca” που μελετήθηκε σε προηγούμενο κεφάλαιο είναι το διγράφημα (joint plot) που δημιουργείται με την παρουσίαση σε Ευκλείδιους άξονες τόσο των χαρακτηριστικών των γραμμών όσο και των στηλών ταυτόχρονα. Τα διαγράμματα που παρουσιάστηκαν στο προηγούμενο κεφάλαιο ήταν δύο ειδών, τα συμμετρικά και τα ασύμμετρα, των οποίων η διαφορά κατά κύριο λόγο βρισκόταν στην απόσταση των σημείων μεταξύ τους. Το πακέτο “anacor” δίνει τη δυνατότητα κατασκευής παραπάνω διαγραμμάτων που περιέχουν περισσότερη πληροφορία και διαφορετικούς τύπους

αποστάσεων και έχουν δημιουργηθεί με τη βοήθεια διαφορετικών τρόπων δημιουργίας αξόνων και κλιμακοποίησης των δεδομένων σε αυτούς.
Για την εφαρμογή της μεθόδου αντιστοιχιών θα χρησιμοποιηθεί η εντολή ‘‘anacor’’ η οποία εφαρμόζεται σε πίνακες συχνοτήτων και σχετικών συχνοτήτων και θα εφαρμοστεί στα δεδομένα για την ηλικιακή ομάδα των ανέργων που παρουσιάζουν μεγαλύτερο ενδιαφέρον. Ο πίνακας με την σχετική συχνότητα των δεδομένων όμως περιέχει και τη στήλη με τις χρονολογίες που έχει αριθμητικές τιμές και εμποδίζει την σωστή εφαρμογή της μεθόδου. Για αυτό το λόγο θα δημιουργήσουμε στην R έναν πίνακα που θα περιέχει τις υπόλοιπες στήλες του πίνακα. Αυτό θα γίνει με τις ακόλουθες εντολές:

```
xdata<-data.frame(X1,X2,X3,X4,X5,X6)
names(xdata)=c("15-19","20-24","25-29","30-44","45-64","64plus")
```

και αυτό δημιουργεί τον επιθυμητό πίνακα δεδομένων.

```
 xdata
   15-19  20-24    25-29    30-44   45-64 64plus
1  31.450 119.075 113.025 169.350  74.100  1.350
2  26.725 110.900 113.200 168.600  71.900  1.300
3  22.875 105.950 111.250 165.650  65.700  1.175
                          .
                          .
                          .
13 31.450 152.150 245.825 558.925 336.600  5.400
14 22.900 137.475 222.775 540.300 344.100  6.850
```

3.4.1 Εφαρμογή του πακέτου “anacor”

Η εγκατάσταση του πακέτου ‘‘anacor’’ και η εφαρμογή της ανάλυσης αντιστοιχειών με τη βοηθειά του για τα δεδομένα για τις διάφορες ηλικιακές ομάδες των ανέργων θα γίνει με τις ακόλουθες εντολές:

```
install.packages("anacor")
library(anacor)
res<-anacor(xdata)
res
```

Παίρνουμε τα αποτελέσματα της εφαρμογής της μεθόδου:

```
CA fit:
Sum of eigenvalues:     0.01544856
```

Total chi-square value: 26.396

Chi-Square decomposition:

	Chisq	Proportion	Cumulative	Proportion
Component 1	24.441		0.926	0.926
Component 2	1.440		0.055	0.980
Component 3	0.315		0.012	0.992
Component 4	0.156		0.006	0.998
Component 5	0.045		0.002	1.000

Τα αποτελέσματα από την εφαρμογή της μεθόδου με αυτό το πακέτο είναι λίγο διαφορετικά από εκείνα με την εφαρμογή του πακέτου "ca". Εμφανίζεται το άθροισμα των ιδιοτιμών πλέον και όχι κάθε ιδιοτιμή χωριστά. Επίσης δίνεται και πάλι η X^2 απόσταση[3] και γίνεται ανάλυση του ποσού της απόστασης αυτής που προέρχεται από κάθε στήλη.

3.4.2 Οι γραφικές παραστάσεις του πακέτου "anacor"

Ακολουθούν μερικά από τα διαγράμματα που δίνει δυνατότητα το πακέτο να σχεδιάσουμε με την περιγραφή τους. Ο τίτλος κάθε είδος διαγράμματος είναι το όρισμα που παίρνει η εντολή "plot" του πακέτου "anacor" για να δώσει το αντίστοιχο τύπου διαγράμματος.

- **Joint plot**

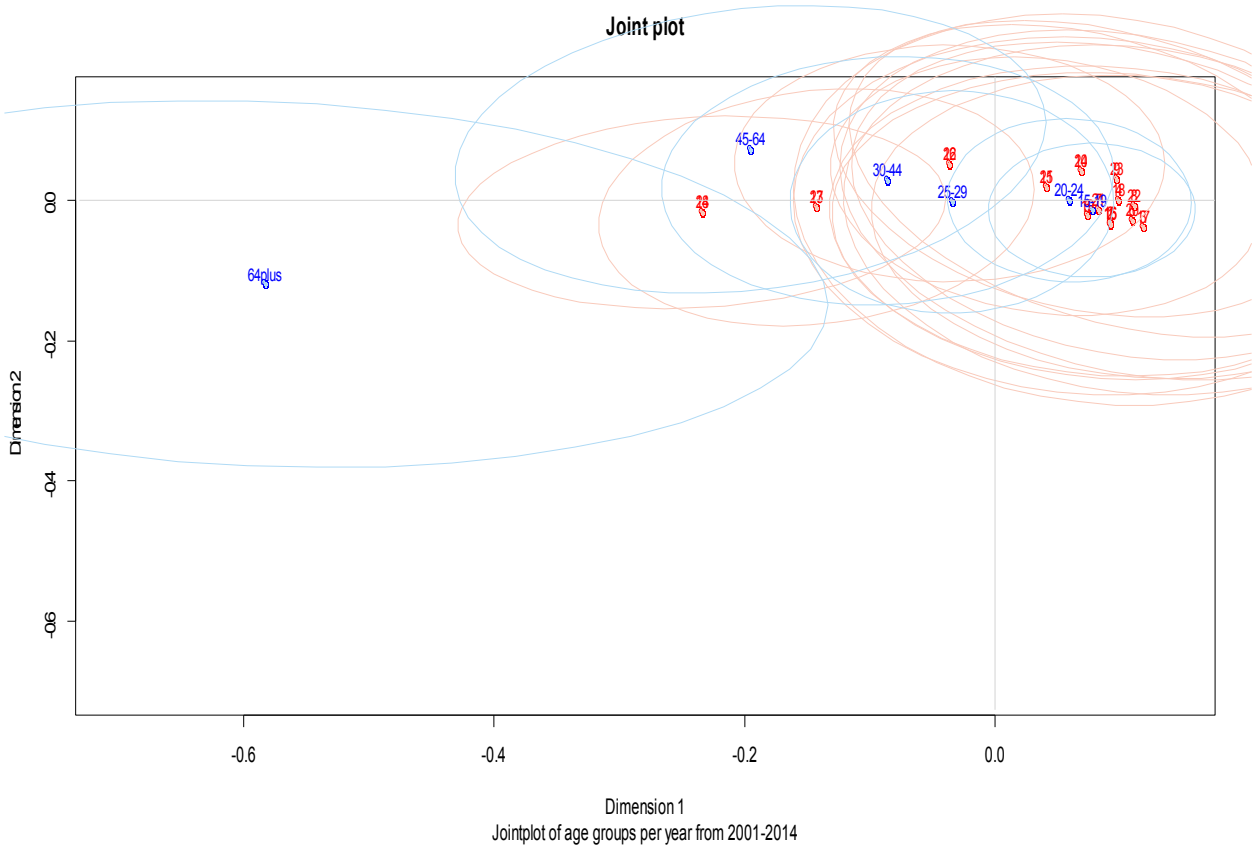
Το διάγραμμα τύπου "joint plot" είναι ένα διάγραμμα δύο διαστάσεων στο οποίο έχουν σχηματιστεί οι δύο άξονες σαν νέες μεταβλητές με τη βοήθεια των ιδιοτιμών του πίνακα συνάφειας όπως γινόταν και με το πακέτο "ca". Στο Διάγραμμα 3.4 που ακολουθεί οι γραμμές και οι στήλες τοποθετούνται στο ίδιο γράφημα. Με μπλε παρουσιάζονται οι στήλες και με κόκκινο οι γραμμές ενώ γύρω από κάθε στήλη και από κάθε γραμμή υπάρχει ένα ελλειψοειδές που περιέχει ένα 95% χωρίο εμπιστοσύνης για τις πραγματικές θέσεις των γραμμών και των στηλών επί των παραγοντικών επιπέδων από ένα πίνακα δύο κατηγορικών μεταβλητών.
Η εύρεση των χωρίων εμπιστοσύνης με αυτή τη μέθοδο βασίζεται στην εξής ιδέα. Αν έχουμε ένα σημείο A(u,v), όπου u και v είναι οι διαθέσιμες αμερόληπτες εκτιμήσεις των συντεταγμένων του σε ένα ορθογώνιο σύστημα συντεταγμένων (U×V) του $\mathbb{R}^2$ και αν θεωρήσουμε τα u και v ως τυχαίες μεταβλητές τότε ο πίνακας διασπορών – συνδιασπορών Σ περιγράφει την αβεβαιότητα της κατ' εκτίμηση θέσης του σημείου A κατά μήκων των δύο αξόνων του συστήματος συντεταγμένων. Αν οι εκτιμήσεις u και v ακολουθούν τη $N_2(\mu, \Sigma)$, με $\mu=[\mu_u, \mu_v]^T$ το διάνυσμα των αντίστοιχων μέσων τιμών, τότε μπορούμε να θεωρήσουμε ότι οι αναμενόμενες τιμές μ_u και μ_v είναι οι άγνωστες πραγματικές συντεταγμένες της θέσης του σημείου A. Αν και είναι δυνατόν να κατασκευαστούν ξεχωριστά (1-α)% διαστήματα εμπιστοσύνης για τις αναμενόμενες αυτές τιμές, ωστόσο,

3 Ο τύπος της απόστασης X^2 δόθηκε στο προηγούμενο κεφάλαιο στη σελίδα 13.

από πρακτική σκοπιά θα ήταν χρήσιμο να αναζητήσουμε μια "περιοχή εμπιστοσύνης" στο επίπεδο, ανεξάρτητα από το εν γένει αυθαίρετο σύστημα αναφοράς, τέτοια ώστε η πραγματική θέση του Α να βρίσκεται μέσα σε αυτή με προκαθορισμένη πιθανότητα P=1-α. Η ζητούμενη περιοχή θα μπορούσε να είναι μια έλλειψη τέτοια ώστε να λαμβάνεται υπόψη τόσο ο διαφορετικός βαθμός αβεβαιότητας της θέσης του σημείου κατά μήκων των αξόνων όσο και η πιθανή συσχέτιση των εκτιμητριών των συντεταγμένων. Η εξίσωση της ζητούμενης έλλειψης αυτής προκύπτει με την μετακίνηση των αξόνων των σημείων κατά τρόπο ώστε το κέντρο τους να είναι το σημείο (μ_u, μ_v) και στη συνέχεια το νέο αυτό σύστημα στρέφεται κατά θετική γωνία φ. Η γωνία φ βρίσκεται ως η απαραίτητη θετική στροφή που χρειάζεται ώστε, αν R είναι ένας πίνακας για τον οποίο ισχύει η σχέση $\vec{x}\,' = R(\vec{x} - \vec{\mu})$,ο R να διαγωνοποιεί τον πίνακα διασπορών-συνδιασπορών. Στην συνέχεια γίνεται η εφαρμογή αυτής της μεταφοράς στον πίνακα Σ και από τον τύπο $(x-\mu)^T \Sigma^{-1}(x-\mu) = \chi^2_{2,\alpha}$ όπου $x = [u, v]^T$,βρίσκουμε μετά από πράξεις την εξίσωση του χωρίου εμπιστοσύνης. Η τελευταία εξίσωση ισχύει λόγω του ότι οι εκτιμήσεις των u,v ακολουθούν την $N_2(\mu, \Sigma)$.

Η R χρησιμοποιεί αυτή τη μέθοδο και βρίσκει τα επιθυμητά χωρία. Η ερμηνεία των αξόνων βασίζεται σε κάποιους βασικούς κανόνες. Σημεία γραμμών που η αντίστοιχη ε.ε.(έλλειψη εμπιστοσύνης) δεν περιέχει την αρχή των αξόνων είναι σημαντικά και συνεισφέρουν στη συσχέτιση (εξάρτηση) των δύο μεταβλητών, όπως αυτή ερμηνεύεται από το παραγοντικό επίπεδο 1×2. Επίσης σημεία γραμμών που οι αντίστοιχες ε.ε. δεν τέμνονται έχουν διαφορετικό προφίλ ιδιαίτερα στην περίπτωση που οι αντίστοιχοι παραγοντικοί άξονες ερμηνεύουν υψηλό ποσοστό της ολικής αδράνειας. Τέλος σημεία γραμμών που οι αντίστοιχες ε.ε. έχουν σχετικά μικρή επιφάνεια έχουν και πιο σταθερή απεικόνιση, δηλαδή τα αποτελέσματα για τις κατηγορίες αυτές μπορούν να γενικευτούν και για άλλους πληθυσμούς ή για άλλες συνθήκες διεξαγωγής της έρευνας (εξωτερική εγκυρότητα) .
Ακολουθεί το Διάγραμμα 3.4 για τα ε.ε.

```
plot(res,plot.type="jointplot")
title(sub="Jointplot of age groups per year from 2001-2014")
```



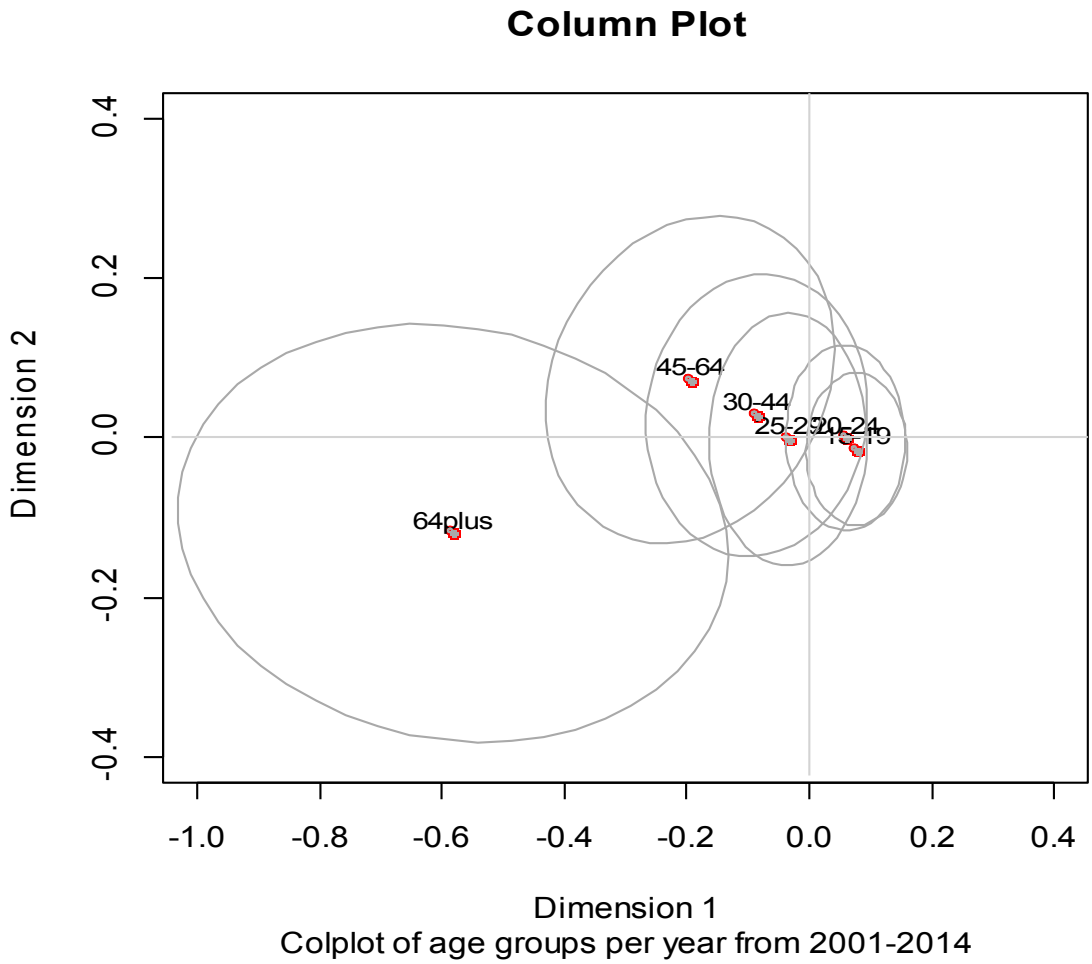
Διάγραμμα 3.4: Διάγραμμα διαστημάτων εμπιστοσύνης για το ποσοστό των ανέργων ανά ηλικιακή ομάδα του Πίνακα 3.2 που περιέχει τις σχετικές συχνότητες επί 100 των ανέργων ανά ηλικιακή ομάδα για τα έτη 2001 ως 2014

Με μία πρώτη ματιά παρατηρούμε ότι τα ελλειψοειδή δεν διαχωρίζονται εμφανώς κάτι αναμενόμενο αφού το δείγμα δεν είναι αρκετά μεγάλο για να αποφευχθούν τέτοια φαινόμενα. Οι κύκλοι με κόκκινο περιέχουν διαστήματα εμπιστοσύνης για τις γραμμές του πίνακα συχνοτήτων, δηλαδή το ποσοστό των ανέργων κάθε χρονιά σε κάθε ηλικιακή ομάδα. Με μπλε απεικονίζεται η κάθε ηλικιακή ομάδα και το διάστημα εμπιστοσύνης για το ποσό των ανέργων κάθε χρονιάς. Το Διάγραμμα 3.4 όμως δεν δίνει μία καθαρή εικόνα λόγω του γεγονότος ότι πολλές ελλείψεις συμπίπτουν για αυτό και θα είναι χρήσιμο να σχεδιαστεί το αντίστοιχο διάγραμμα χωριστά για τις γραμμές και τις στήλες που παρουσιάζουν μεγαλύτερες διαφοροποιήσεις και να σχολιαστούν καλύτερα τα αποτελέσματα. Το Διάγραμμα 3.5 που ακολουθεί στην επόμενη κατηγορία είναι το αντίστοιχο διάγραμμα κατασκευής ε.ε, για τις κατηγορίες των στηλών όπου και θα σχολιαστούν τα αποτελέσματα.

- **Row/Col plot**

Το διάγραμμα "row/col plot" είναι ένα διάγραμμα παρόμοιο με το διάγραμμα "joint plot" με την μόνη διαφορά ότι παρουσιάζει μόνο τις γραμμές (ή τις στήλες αντίστοιχα) .

```
plot(res,plot.type ="colplot",col="red",xlim=c(-1,0.4),ylim=c(-0.4,0.4))
title(sub="Colplot of age groups per year from 2001-2014")
```

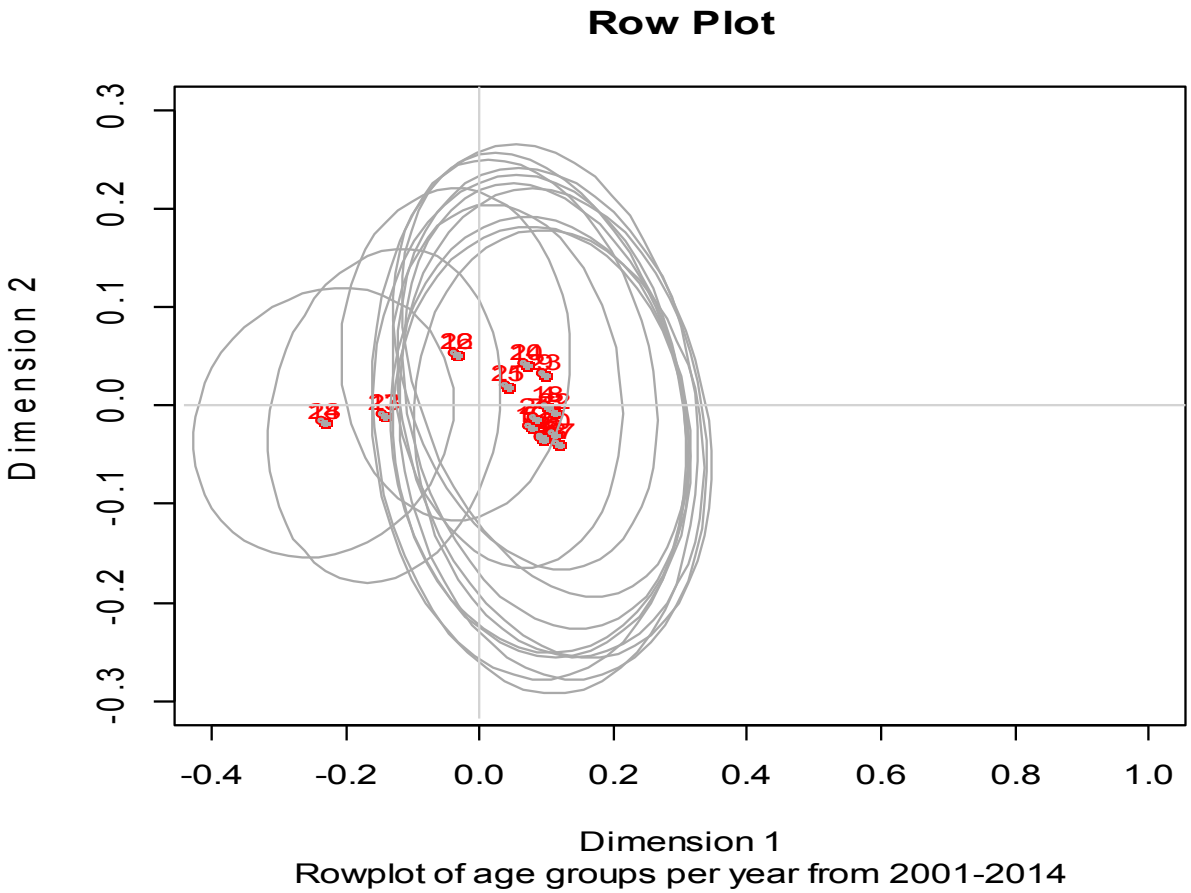


Διάγραμμα 3.5: Διάγραμμα διαστημάτων εμπιστοσύνης μόνο για τις στήλες του Πίνακα 3.2, δηλαδή για την σχετική συχνότητα των ανέργων επί 100 κάθε ηλικιακής ομάδας

Παρατηρούμε ότι τα ελλειψοειδή που δημιουργούνται για τις στήλες που αντιστοιχούν στις διάφορες ηλικιακές ομάδες δεν είναι ίδια σε μέγεθος. Όπως είχαμε δει σε προηγούμενη παράγραφο οι στήλες διαφέρουν πολύ μεταξύ τους τόσο στη μέση τιμή όσο και στη διασπορά οπότε είναι λογικό να υπάρχουν αντίστοιχα έντονες διαφοροποιήσεις και στα ε.ε. Οι ηλικιακή ομάδα ανέργων άνω των 64 φαίνεται να έχει τη λιγότερο δυνατή σταθερή απεικόνιση λόγω του μεγέθους της ελλείψεως της. Η κατηγορία αυτή των ανέργων είναι τοποθετημένη στα αριστερά του πρώτου άξονα και δεν περιέχει την αρχή των αξόνων πράγμα που σημαίνει ότι δεν είναι αρκετά σημαντική σε σχέση με άλλες για τη συσχέτιση των δύο μεταβλητών που έχει δημιουργήσει η ανάλυση αντιστοιχιών δηλαδή των δύο αξόνων, αυτό ίσως οφείλεται στις μικρές τιμές που παίρνει σε σχέση με τις άλλες ηλικιακές ομάδες. Επίσης όσον αφορά τον δεύτερο άξονα η έλλειψη για αυτή την ηλικιακή ομάδα δεν είναι ξεκάθαρα τοποθετημένη πράγμα που σημαίνει ότι η ερμηνεία της κατηγορίας αυτής ως προς την δεύτερη μεταβλητή που δημιούργησε η ανάλυση αντιστοιχιών είναι προβληματική αφού δεν μπορεί να αποδοθεί σε μία από τις δύο πλευρές που χωρίζει το

επίπεδο ο οριζόντιος άξονας, ο οποίος είχαμε δει ότι εξηγεί περίπου το 75% της σχέσης των παρατηρήσεων. Οι υπόλοιπες ηλικιακές ομάδες φαίνεται να συμμετέχουν πιο πολύ στην ερμηνεία των δύο αξόνων και να έχουν πιο σταθερή απεικόνιση. Καμία όμως δεν έχει αρκετά μικρό χωρίο ώστε να θεωρηθεί ότι μπορεί να γενικευτεί και για άλλες καταστάσεις. Δεν είναι δηλαδή αρκετά σταθερή η απεικόνιση του. Τέλος σχεδόν όλες οι ελλείψεις φαίνεται να τέμνονται ή ακόμα και να περιέχονται σε άλλες, πράγμα που σημαίνει ότι τα προφίλ των γραμμών δεν διαφέρουν σημαντικά. Δηλαδή οι μεταβολές των ποσοστών των ανέργων ανά ηλικιακή ομάδα δεν διαφέρουν με την πάροδο των ετών σε κάθε μία από τις εξεταζόμενες ηλικιακές ομάδες.
Για τις γραμμές η ερμηνεία και η κατασκευή του Διαγράμματος 3.6 γίνονται με όμοιο τρόπο.

```
plot(res,plot.type ="rowplot",col="red",xlim=c(-0.4,1),ylim=c(-0.3,0.3))
title(sub="Rowplot of age groups per year from 2001-2014")
```



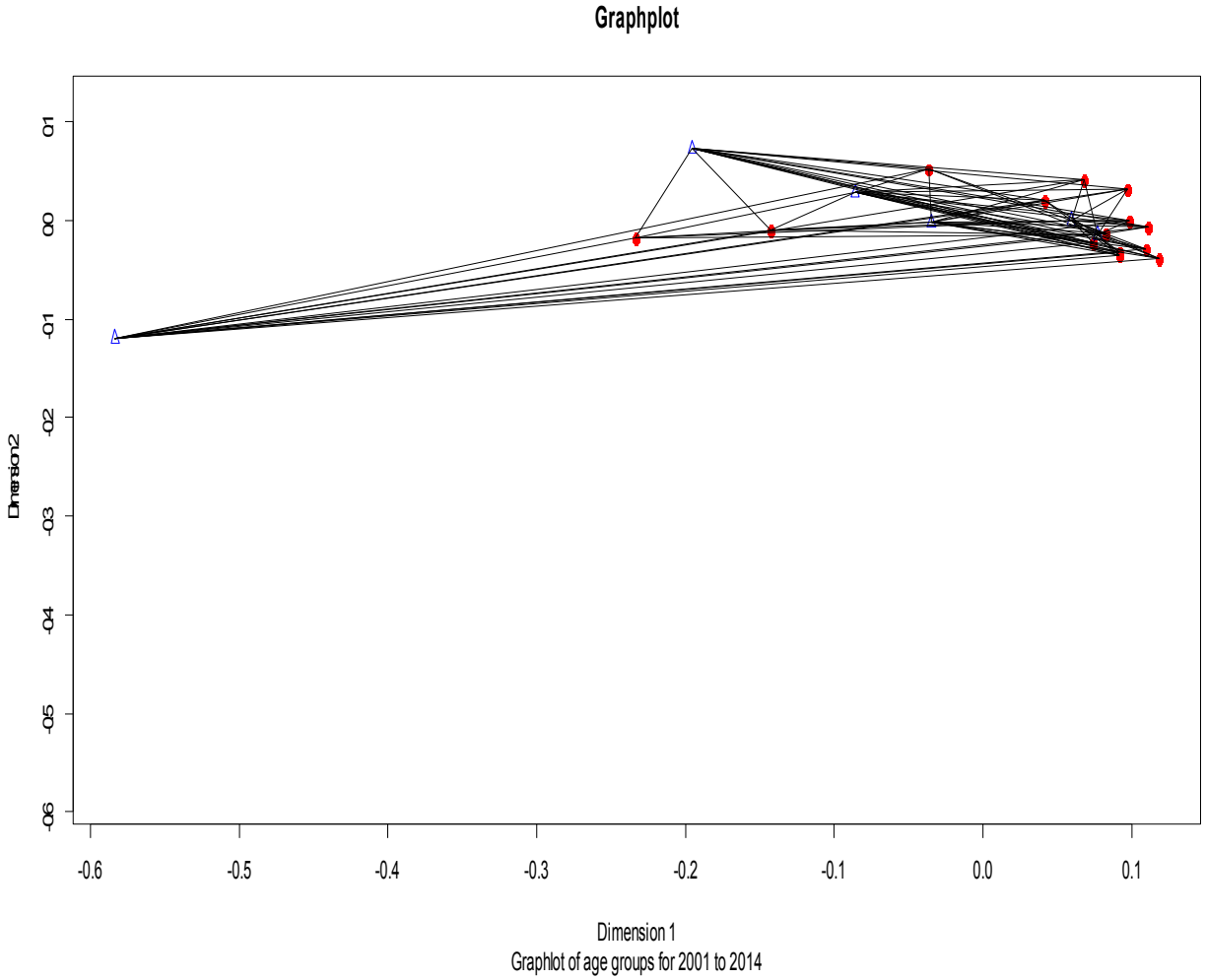
Διάγραμμα 3.6: Διάγραμμα διαστημάτων εμπιστοσύνης μόνο για τις γραμμές του Πίνακα 3.2, δηλαδή για τα έτη σχετική συχνότητα επί 100 της ηλικιακής ομάδας των ανέργων

Και για τις γραμμές τα χωρία εμπιστοσύνης φαίνεται να συμπίπτουν και μόνο μερικές χρονιές να διαφέρουν αισθητά που είναι οι τελευτείες χρονιές όπου η αύξηση των ποσοστών ανεργίας σε όλες τις ηλικιακές ομάδες είναι ιδιαίτερα αισθητή.

- **Graphplot**

Το διάγραμμα "graphplot" θυμίζει το διάγραμμα "jointplot" καθώς έχει καταστευαστεί στους ίδιους άξονες και περιέχει και αυτό τόσο τις στήλες όσο και τις γραμμές σαν σημεία με διαφορετικό σύμβολο ή/και χρώμα για τις μεν και τις δε. Η διαφορά τους είναι ότι σε αυτό το είδος διαγράμματος τα σημεία είναι όλα ενωμένα μεταξύ τους με γραμμές, το πάχος των οποίων δίχνει τη δύναμη της σύνδεσης του κάθε σημείου.

```
plot(res,plot.type="graphplot")
title(sub="Graphlot of age groups for 2001 to 2014")
```



Διάγραμμα 3.7: Διάγραμμα των στηλών και σειρών του Πίνακα 3.2 για τα ποσοστά των ανέργων επί 100 ανά ηλικιακή ομάδα τα έτη 2001 ως 2014 ενωμένων με γραμμές

Οι προεπιλογές τις R είναι να δείχνει με μπλε τρίγωνο τις στήλες και με κόκκινες βούλες τις γραμμές. Τα σημεία βρίσκονται στο ίδιο σημείο με αυτά στο από κοινού διάγραμμα διαστημάτων εμπιστοσύνης και παρατηρούμε ότι η ηλικιακή ομάδα άνω των 64 βρίσκεται πιο απόμακρα σε σχέση με τις άλλες ηλικιακές ομάδες που είναι μέσα στο πλήθος των ενωμένων γραμμών. Επίσης οι κόκκινες βούλες φαίνονται συγκεντρωμένες στο δεξί κομμάτι του Διαγράμματος 3.7 πέρα απο δύο που είναι πιο μακριά και αντιστοιχούν στα

πολύ μεγαλύτερα ποσοστά ανέργων που συγκεντρώνουν τα τελευταία έτη.

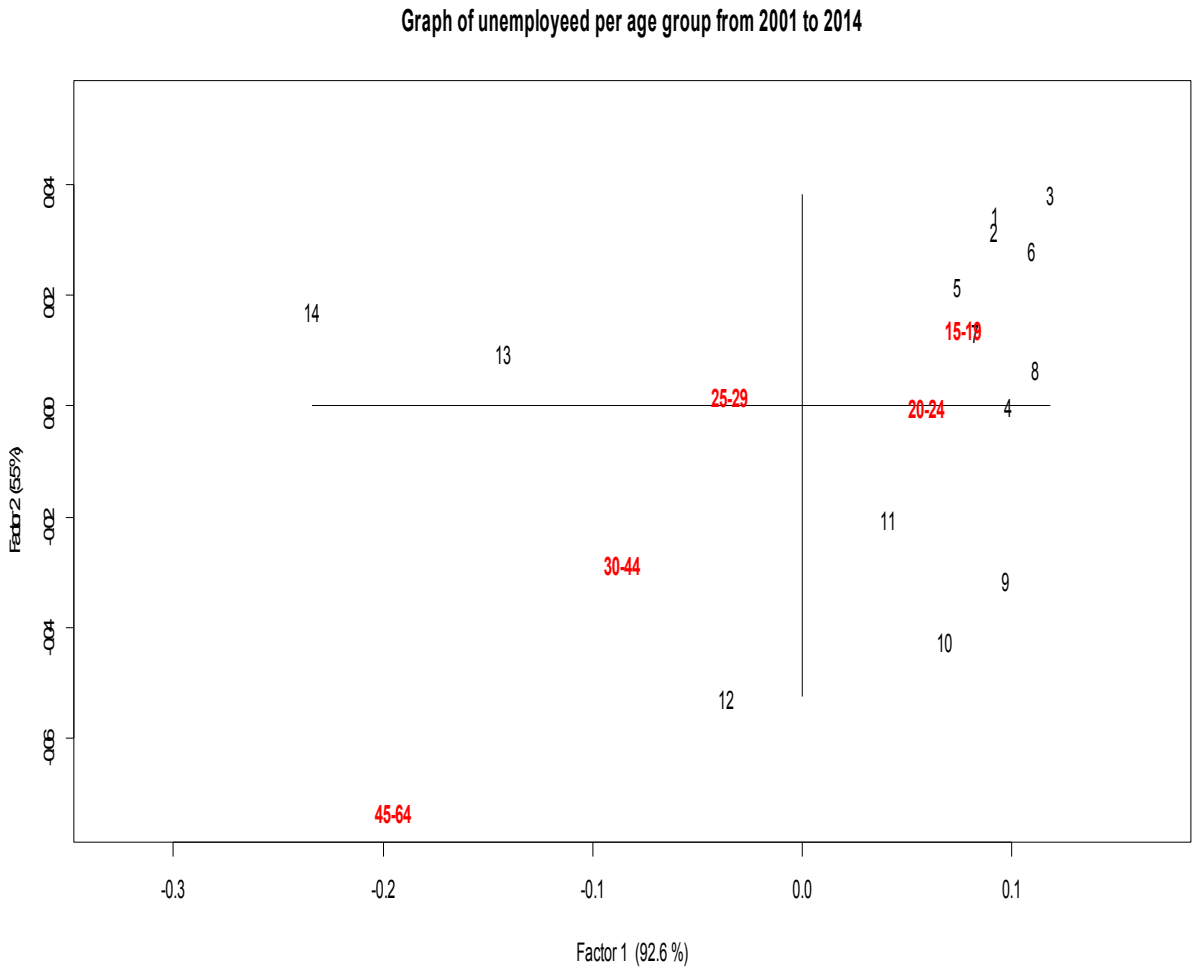
3.5 Η ανάλυση αντιστοιχιών με το πακέτο “languageR”

Το πακέτο “languageR” περιέχει μαζί με άλλες εντολές που προσφέρει και την εντολή “corres.fnc” η οποία εφαρμόζει σε έναν πίνακα συχνοτήτων ή σχετικών συχνοτήτων την ανάλυση αντιστοιχιών και δίνει τη δυνατότητα κατασκευής ενός ακόμα τύπου διαγράμματος. Γίνεται λοιπόν η εγκατάσταση του πακέτου.

```
install.packages("languageR")
library(languageR)
```

Στη συνέχεια εκτελείται η ανάλυση αντιστοιχιών για τον Πίνακα 3.2.

```
res1<-corres.fnc(xdata)

plot(res1)
title(main="Graph of unemployeed per age group from 2001 to 2014")
```



Διάγραμμα 3.9: Συμμετρικό διάγραμμα ανάλυσης αντιστοιχειών για τον πίνακα σχετικών συχνοτήτων επί 100 της ηλικιακής ομάδας των ανέργων για τα έτη 2001 ως 2014

Το Διάγραμμα 3.9 θυμίζει λίγο εκείνα που παρείχε το πακέτο "ca" στη μορφή του. Με κόκκινο βρίσκονται τοποθετημένες στους άξονες οι διάφορες ηλικιακές ομάδες και με μαύρο συμβολίζονται οι χρονιές από το 2001 ως το 2014. Τα στοιχεία "2 0" πριν από κάθε χρονολογία έχουν αφαιρεθεί καθώς είναι όμοια σε όλα τα έτη και θα καταλάμβαναν παραπάνω χώρο με αποτέλεσμα να δημιουργούνταν σύγχυση. Οι δύο άξονες έχουν δημιουργηθεί από την εφαρμογή της μεθόδου. Ο οριζόντιος άξονας φαίνεται να κατατάσει τις ηλικιακές ομάδες ανά σειρά γηραιότητας και είναι αυτός που εξηγεί την πλειοψηφία των σημείων, δηλαδή το 92,6% όπως αναγράφεται στο Διάγραμμα 3.9.
Όπως μπορεί να δει εύκολα κανείς οι περισσότερες χρονιές συγκεντρώνονται στα δεξιά του οριζόντιου άξονα και βρίσκονται κοντά στις ηλικιακές ομάδες 15-19 ετών και 20-24 ετών. Η ομάδα των ανέργων άνω των 64 δεν εμφανίζεται καν στο Διάγραμμα 3.9 καθώς βρίσκεται πολύ πιο αριστερά σε σχέση με τις άλλες ηλικιακές ομάδες. Επίσης φαίνεται ότι τα έτη μετά την αρχή της κρίσης δηλαδή μετά το 2009 βρίσκονται πιο μακριά από τις άλλες που σημαίνει ότι για όλες τις ηλικιακές ομάδες η διαφοροποίηση των ποσοστών είναι έντονη. Τα έτη 2013 και 2014 μάλιστα είναι απομονωμένα στα αριστερά του άξονα. Ο οριζόντιος άξονας αυτή τη φορά μπορεί να θεωρηθεί ότι κατατάσσει τα έτη κατά φθίνουσα σειρά για τα ποσοστά των ανέργων που συγκεντρώνουν.

3.6 Αναφορά στην ανάλυση αντιστοιχειών με άλλα πακέτα

Εκτός από τα πακέτα που παρουσιάστηκαν παραπάνω υπάρχουν και μερικά ακόμα που προσφέρουν τη δυνατότητα εφαρμογής ανάλυσης αντιστοιχιών και τη δημιουργία παρόμοιων διαγραμμάτων.
Ένα από αυτά είναι το πακέτο “amap” με την εφαρμογή της εντολής “afc( )” στον πίνακα των δεδομένων, με την οποία μπορεί να σχεδιαστεί και το διάγραμμα των αποτελεσμάτων που είναι παρόμοιο με αυτά που έχουν παρουσιαστεί μέχρι στιγμής.
Το πακέτο “ade4” επίσης παρέχει τη δυνατότητα εφαρμογής απλής ανάλυσης αντιστοιχιών με την εντολή “dudi.coa( )” .
Τέλος το ήδη εγκατεστημένο πακέτο “MASS” που περιέχει και μία σειρά από λίστες δεδομένων διαθέτει την εντολή “corresp( )” που εφαρμόζει και αυτή την ανάλυση αντιστοιχιών.

4. Το ποσοστό των ανέργων ανά επίπεδο εκπαίδευσης

Σε αυτό το κεφάλαιο θα γίνει μία μελέτη της σχέσης των ποσοστών των ανέργων σε σχέση με το επίπεδο εκπαίδευσης. Οι αριθμοί στον πίνακα των δεδομένων αποτελούν ποσοστά ανέργων επί του συνόλου του εργατικού δυναμικού σε κάθε κατηγορία. Η μελέτη θα γίνει με τη δημιουργία διαφόρων διαγραμμάτων των οποίων την υλοποίηση προσφέρει η R με τη βοήθεια διαφόρων πακέτων.

4.1 Παρουσίαση δεδομένων

Ο πίνακας δεδομένων για το μορφωτικό επίπεδο των ανέργων περιέχει δύο χαρακτηριστικά το ένα που αποτελεί τις γραμμές του πίνακα είναι η χρονιά από την οποία προέρχονται τα δεδομένα και παίρνει τιμές απο 2001 ως 2014. Οι στήλες του πίνακα αποτελούν το είδος της εκπαίδευσης των ανέργων που μπορεί να πάρει 8 διαφορετικές τιμές.
Αυτές είναι:

1. Διδακτορικό ή Μεταπτυχιακός τίτλος
2. Πτυχίο Ανωτάτων Σχολών
3. Πτυχίο Ανώτερης Τεχνολογικής Εκπαίδευσης
4. Απολυτήριο Μέσης Εκπαίδευσης
5. Απολυτήριο 3-τάξεωνΜέσης Εκπαίδευσης
6. Απολυτήριο Δημοτικού
7. Μερικές Τάξεις Δημοτικού
8. Δεν πήγε καθόλου σχολείο

Στον Πίνακα 4.1 με τις σχετικές συχνότητες των κατηγοριών τα διάφορα μορφωτικά επίπεδα συμβολίζονται με ένα κεφαλαίο γράμμα της αγγλικής αλφαβήτου με Α να είναι η πρώτη κατηγορία και Η η τελευταία με τη σειρά παρουσίασης.

		Μορφωτικό επίπεδο							
		A	B	C	D	E	F	G	H
Χρονιά	2001	8,35	7,5	13,075	13,325	13,675	7,85	4,75	7,975
	2002	7	6,575	13,8	12,7	12,375	7,525	6,45	5,275
	2003	7,675	6,675	12,95	12,075	11,15	6,95	4,8	2,725
	2004	6,4	7,75	12,8	12,075	11,35	8,925	5,525	12,2
	2005	7,2	7,475	12,425	11,35	10,8	7,875	5,975	9,95
	2006	6,475	6,275	11,575	10,325	9,65	7,075	4,175	7,3
	2007	6,475	6,375	10,575	9,25	8,45	6,95	6,2	12,4
	2008	5,475	5,6	10,025	8,425	8,65	6,175	5,65	9,825
	2009	7,5	6,725	11,1	10,7	11,15	8,15	9,6	14,5
	2010	7,875	8,75	14,975	14	14,625	11	13,075	20,3
	2011	10,525	13	20,3	19,65	19,5	16,325	24,075	28,5
	2012	13,15	16,7	27,225	26,775	27,55	24,425	27,575	38,6
	2013	15,4	18,475	30,525	29,925	32,675	26,925	38,275	40,3
	2014	13,625	19,525	27,7	29,375	30,95	25,75	41,525	36,8

Πίνακας 4.1: Πίνακας σχετικών συχνοτήτων του ποσοστού των ανέργων επί 100 ανά επίπεδο εκπαίδευσης για τα έτη 2001 ως 2014

4.2 Εισαγωγή δεδομένων στην R

Όμοια με τα προηγούμενα κεφάλαιο τα δεδομένα θα εκχωρηθούν στην R με τη βοήθεια του πακέτου “gdata”.

```
install.packages(“gdata”)
library(gdata)

data1<-read.xls("../Desktop/diplwmatikh-domh/kef3/ekpaideush.xls",sheet=1)
```

Καλώντας την μεταβλητή data1 μπορεί κανείς να διαπιστώσει ότι ο πίνακας σχετικών συχνοτήτων εκχωρήθηκε με επιτυχία στην R.

```
data1
   xronia     A     B      C      D      E     F     G     H
1    2001 8.350 7.500 13.075 13.325 13.675 7.850 4.750 7.975
2    2002 7.000 6.575 13.800 12.700 12.375 7.525 6.450 5.275
3    2003 7.675 6.675 12.950 12.075 11.150 6.950 4.800 2.725
                              .
                              .
                              .
10   2010 7.875 8.750 14.975 14.000 14.625 11.000 13.075 20.300
11   2011 10.525 13.000 20.300 19.650 19.500 16.325 24.075 28.500
```

Μετά θα οριστεί η κάθε στήλη ως διάνυσμα από τη δεύτερη ως την τελευταία. Για να γίνει αυτό θα δοθούν οι εντολές

```
A<-(data1)[,2]
B<-(data1)[,3]
C<-(data1)[,4]
D<-(data1)[,5]
E<-(data1)[,6]
F<-(data1)[,7]
G<-(data1)[,8]
H<-(data1)[,9]

year<-(data1)[,1]
```

Καλώντας ένα από αυτά τα διανύσματα θα δει κανείς ότι είναι ένα διάνυσμα που έχει τιμές την αντίστοιχη στήλη του πίνακα συχνοτήτων.

```
G
 [1]  4.750  6.450  4.800  5.525  5.975  4.175  6.200  5.650  9.600 13.075
[11] 24.075 27.575 38.275 41.525
```

4.3 Βασικά περιγραφικά στοιχεία των δεδομένων

Για να έχει κανείς μία καλύτερη εικόνα των δεδομένων χρειάζεται να δει κάποια μέτρα θέσης και κάποια μέτρα μεταβλητότητας και στη συνέχεια να δημιουργηθούν διάφορα διαγράμματα για την καλύτερη κατανόηση του Πίνακα 4.1.
Για αυτό ακολουθούν τα αποτελέσματα της εντολής summary της R για κάθε στήλη-διάνυσμα του πίνακα με τα ποσοστά των ανέργων ανά κατηγορία επιπέδου εκπαίδευσης, η οποία δίνει τον μέσο, το ελάχιστο και το μέγιστο και το πρώτο και τρίτο τεταρτημόριο που δείχνουν την παρατήρηση αριστερά από την οποία βρίσκεται το πολύ το 25% του συνόλου των παρατηρήσεων και την παρατήρηση αριστερά της οποίας βρίσκεται το πολύ το 75% του συνόλου των παρατηρήσεων αντίστοιχα.

```
summary(A)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  5.475   6.606   7.588   8.795   9.981  15.400
summary(B)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  5.600   6.600   7.488   9.814  11.940  19.520
summary(C)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.02   11.79   13.01   16.36   18.97   30.52
```

```
summary(D)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  8.425  10.860  12.390  15.710  18.240  29.920 
summary(E)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   8.45   10.89   11.86   15.90   18.28   32.67 
summary(F)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  6.175   7.188   8.012  12.280  14.990  26.920 
summary(G)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  4.175   5.556   6.325  14.120  21.320  41.520 
summary(H)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2.725   8.438  12.300  17.620  26.450  40.300 
```

Επιπλέον ακολουθεί και η τυπική απόκλιση της κάθε κατηγορίας ανέργων ανάλογα με το μορφωτικό τους επίπεδο.

```
 sd(A)
[1] 3.115457
 sd(B)
[1] 4.923235
 sd(C)
[1] 7.049418
 sd(D)
[1] 7.538445
 sd(E)
[1] 8.387501
 sd(F)
[1] 7.708891
 sd(G)
[1] 13.13607
 sd(H)
[1] 13.03165
```

Ρίχνοντας μία ματιά στους μέσους είναι εμφανές ότι το μεγαλύτερο ποσοστό ανέργων αφορά την κατηγορία των ατόμων που δεν πήγαν σχολείο ή έχουν μόνο πτυχίο μέσης εκπαίδευσης, κάτι που είναι λογικό καθώς οι απαιτήσεις της εργασίας με την πάροδο του χρόνου μεγαλώνουν και το πτυχίο γίνεται απαραίτητο για την εύρεση εργασίας. Τα μικρότερα ποσοστά ανέργων κατά μέσο όρο εμφανίζονται στις κατηγορίες ατόμων με πτυχίο και μεταπτυχιακό ή διδακτορικό αλλά και στην κατηγορία των ατόμων που απλά έχουν πάει μερικές χρονιές σχολείο. Αυτή η κατηγορία ατόμων διεκδικεί μία συγκεκριμένη μερίδα από το σύνολο των θέσεων εργασίας, όπως οι αγροτικές θέσεις και συναγωνίζεται την κατηγορία ατόμων που δεν πήγαν καθόλου σχολείο διεκδικώντας πολύ περισσότερες από αυτές τις θέσεις όπως φαίνεται από τα ποσοστά της εμφάνισης των ανέργων.
Από την τυπική απόκλιση όμως και από το ενδοτεταρτημοριακό εύρος βλέπει κανείς ότι οι

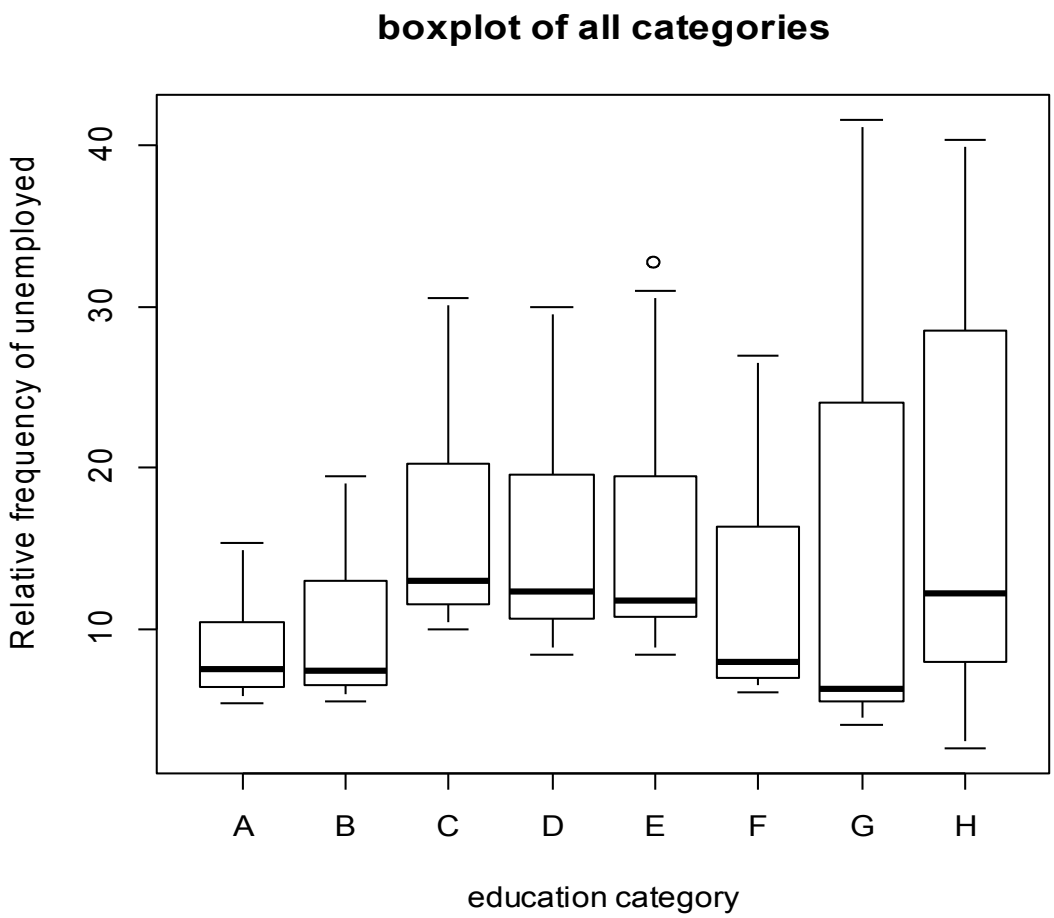
δύο τελευταίες κατηγορίες έχουν μεγάλη διαφοροποίηση ανά χρονιά καθώς έχουν μεγάλες τιμές τυπικής απόκλισης. Αυτό μπορεί να οφείλεται στο γεγονός ότι με την πάροδο του χρόνου ακόμα και τα επαγγέλματα που αφορούν τις αγροτικές εργασίες ή άλλα που θα μπορούσε να θεωρηθεί ότι δεν απαιτούν εξειδικευμένες γνώσεις και κάποιο τίτλο πτυχίου χρειάζονται κάποιες γνώσεις. Τα άτομα που δεν έχουν ολοκληρώσει το σχολείο λοιπόν μπορεί στο παρελθόν να μην αντιμετώπιζαν πρόβλημα αν στρεφόντουσαν σε αυτή την κατηγορία επαγγελμάτων. Κάτι τέτοιο όμως δεν ισχύει τα τελευταία χρόνια και μπορεί κανεις να το διαπιστώσει εύκολα κοιτώντας την πρώτη και την τελευταία τιμή του πίνακα συχνοτήτων για τις κατηγορίες G, H.

4.4 Διαγράμματα των κατηγοριών μορφωτικού επιπέδου

4.4.1 Θηκόγραμμα (boxplot)

Το θηκόγραμμα είναι ένα χρήσιμο διάγραμμα για τη σύγκριση διαφορετικών κατηγοριών για κατηγορικές μεταβλητές και έχει ήδη χρησιμοποιηθεί σε προηγούμενα κεφάλαια καθώς δίνει μία καλή εικόνα για τη σχέση κατηγοριών μεταξύ τους.
Ο κώδικας για την κατασκευή του Διαγράμματος 4.1 είναι ο ακόλουθος.

```
boxplot(data1$A, data1$B, data1$C, data1$D, data1$E, data1$F, data1$G, data1$H,
names=(c("A","B","C","D","E","F","G","H")), xlab="education category",ylab="frequency of
unemployed")
title(main="boxplot of all categories")
```



Διάγραμμα 4.1: Θηκόγραμμα για την σχετική συχνότητα των ανέργων επί 100 ανά επίπεδο εκπαίδευσης για τα έτη 2001 ως 2014

Από το Διάγραμμα 4.1 παρατηρεί κανείς ότι οι διάμεσοι σε όλες τις περιπτώσεις βρίσκονται κοντά στο κάτω άκρο του κάθε θηκογραφήματος ενώ αυτά εμφανίζουν και τιμές πολύ μεγαλύτερες, κάτι που δείχνει ότι η μεσαία παρατήρηση είναι πολύ μικρή σε σχέση με άλλες στο δείγμα. Από τον Πίνακα 4.2 μπορεί να δει κανείς πως στις περισσότερες κατηγορίες μορφωτικού επιπέδου οι 5 τελευταίες χρονιές εμφανίζουν μεγάλη αύξηση του ποσοστού ενώ οι 10 προηγούμενες έχουν πολύ χαμηλότερες τιμές. Οι κατηγορίες με το μεγαλύτερο εύρος τιμών είναι οι δύο τελευταίες όπως είχαμε δει και προηγουμένως, δηλαδή άτομα που δεν έχουν πάει καθόλου σχολείο ή έχουν πάει πολύ λίγο κάτι που μπορεί να το διαπιστώσει κανείς και από τη μεγάλη απόσταση που έχει η ελάχιστη από τη μέγιστη τιμή σε αυτές τις δύο περιπτώσεις.

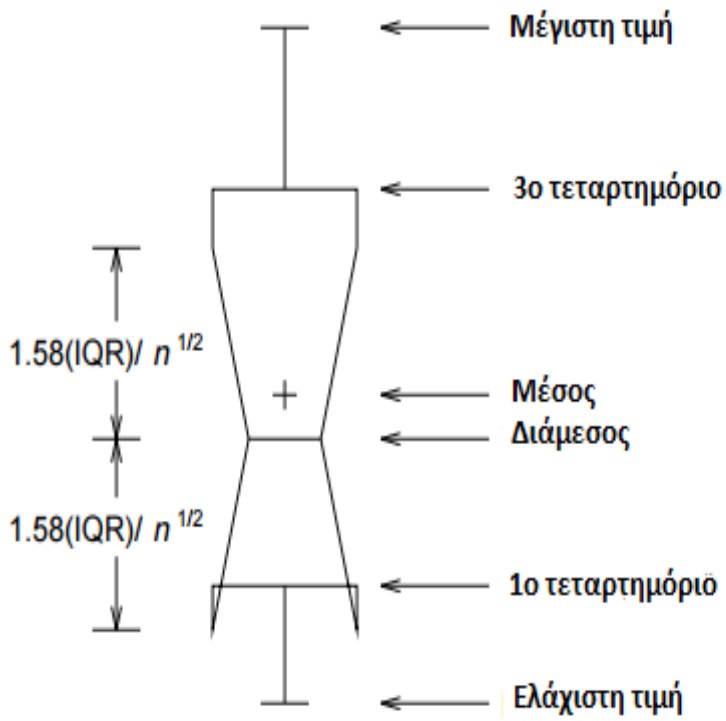
Επίσης παρατηρούμε ότι η κατηγορία Ε, δηλαδή η κατηγορία ατόμων που έχουν απολυτήριο 3 τάξεων μέσης εκπαίδευσης παρουσιάζουν και κάποια ακραία τιμή προς τα πάνω, έχουν δηλαδή κάποια χρονιά με πολύ μεγάλο ποσοστό ανέργων. Τέλος η κατηγορία με την μικρότερη δυνατή διαφοροποίηση στα ποσοστά των ανέργων που συγκεντρώνει είναι αυτή των ατόμων με διδακτορικό ή μεταπτυχιακό τίτλο. Για αυτή την κατηγορία δε φαίνεται το ποσοστό των ανέργων να διαφοροποιείται τόσο πολύ σε σχέση με άλλες κατηγορίες. Ο μέσος όρος της σχετικής συχνότητας των ανέργων όμως σε αυτή την κατηγορία μορφωτικού επιπέδου δεν είναι ο χαμηλότερος στο δείγμα όπως μπορεί να δει

κανείς από την εντολή “summary” στην Παράγραφο 4.3. Δεν είναι εύκολο λοιπόν να βγάλει κανείς συμπέρασμα από αυτό το διάγραμμα για το ποια κατηγορία συγκεντρώνει τους περισσότερους ανέργους.

4.4.2 Θηκόγραμμα με εγκοπές (boxplot with notches)

Εκτός από το απλό θηκογράφημα που παρουσιάστηκε παραπάνω η R δίνει την δυνατότητα κατασκευής ενός πιο σύνθετου αλλά παρόμοιου γραφήματος. Το θηκόγραμμα με εγκοπές όπως και το κανονικό στο κύριο μέρος του, το “κουτί” περιέχει το 50% των διατεταγμένων παρατηρήσεων και τα 2 άκρα αντιπροσωπεύουν την μέγιστη και την ελάχιστη τιμή ενώ ο μέσος όρος αντιπροσωπεύεται με ένα σταυρό και η διάμεσος με μια ευθεία γραμμή εντός του διαγράμματος. Σε αυτή την περίπτωση όμως το συνολικό ύψος του κουτιού δεν είναι το ενδοτεταρτημοριακό έυρος, δηλαδή η διαφορά του πρώτου από το τρίτο τεταρτημόριο αλλά δίνεται από τον τύπο $\frac{1.58 \cdot IQR}{\sqrt{n}}$ όπου n είναι το σύνολο των παρατηρήσεων και IQR το ενδοτεταρτημοριακό εύρος.
Αυτό το γράφημα μας δίνει ένα 95% διάστημα για να συγκρίνουμε τους μέσους και χρησιμοποιείτε για να συγκρίνουμε διαφορετικά δείγματα μεταξύ τους όπως στην περίπτωση των μορφωτικών επιπέδων. Αν δύο παράλληλα θηκογραφήματα με εγκοπές δεν επικαλύπτονται τότε μπορεί κανείς να συμπεράνει ότι οι μέσοι όροι τους είναι με 95% επίπεδο σημαντικότητας διαφορετικοί μεταξύ τους.

Μπορεί κανείς να δει τις παραπάνω εξηγήσεις στην Διάγραμμα 4.2 που ακολουθεί ώστε να αντιληφθεί ακριβώς τι σημαίνει κάθε κομμάτι του γραφήματος αυτού.



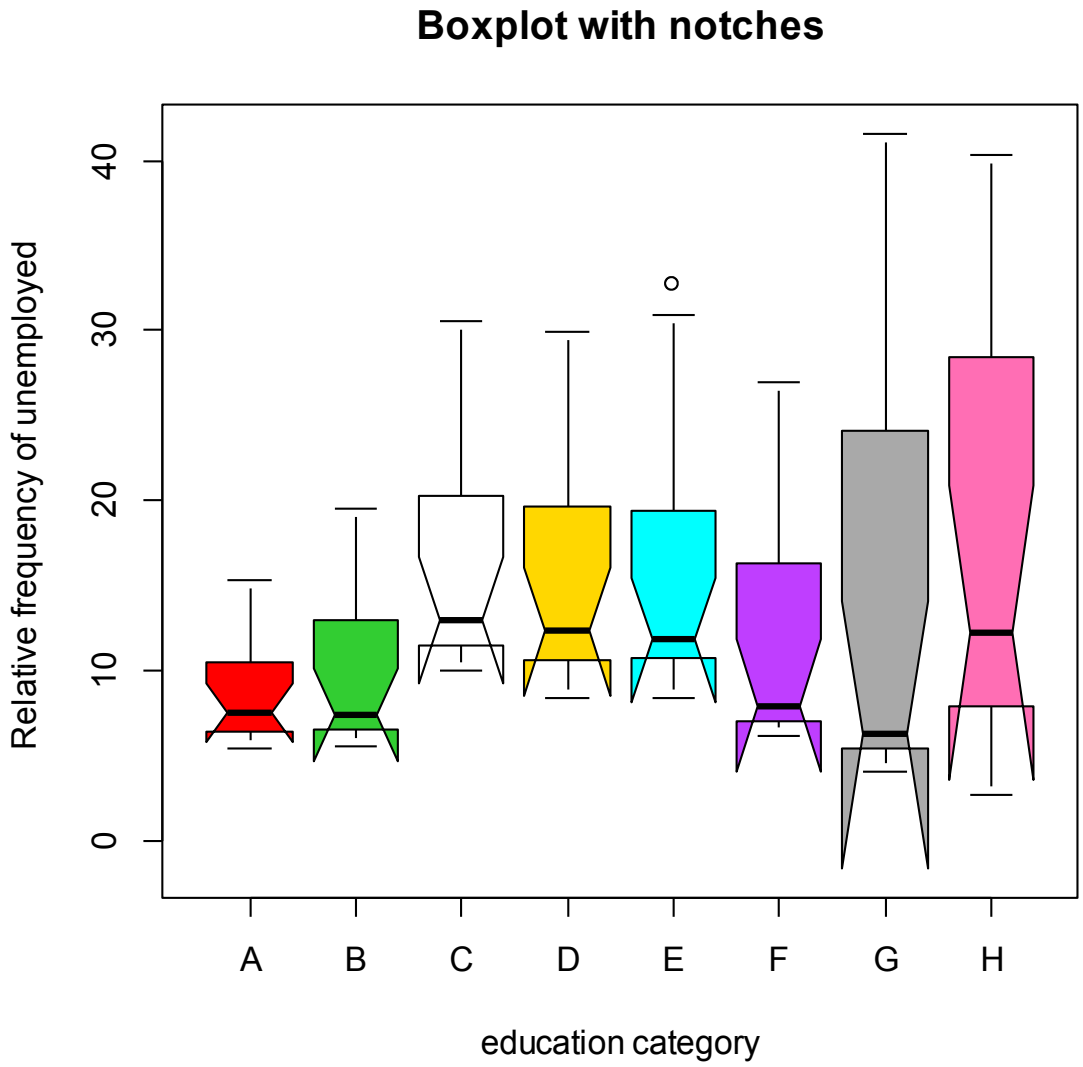
Διάγραμμα 4.2: Οπτική παρουσίαση της εξήγησης του θηκογράμματος με εγκοπές

Για την κατασκευή του Διαγράμματος 4.2 στην R για τις 8 κατηγορίες μορφωτικού επιπέδου των ανέργων δίνεται η παρακάτω εντολή

```
boxplot(A,B,C,D,E,F,G,H,notch=TRUE,col=c("red","limegreen","white","gold","cyan","darkorchid1"
,"darkgrey","hotpink1"),names=(c("A","B","C","D","E","F","G","H")),xlab="education
category",ylab=" Relative frequency of unemployed")
title(main="Boxplot with notches")
```

οπότε προκύπτει το Διάγραμμα 4.3[4].

4 Τα χρώματα που έχουν χρησιμοποιηθεί προέρχονται από την λίστα χρωμάτων της R



Διάγραμμα 4.3: Θηκόγραμμα με εγκοπές για τις διάφορες κατηγορίες επίπεδου εκπαίδευσης της σχετικής συχνότητας των ανέργων επί 100 για τα έτη 2001 ως 2014

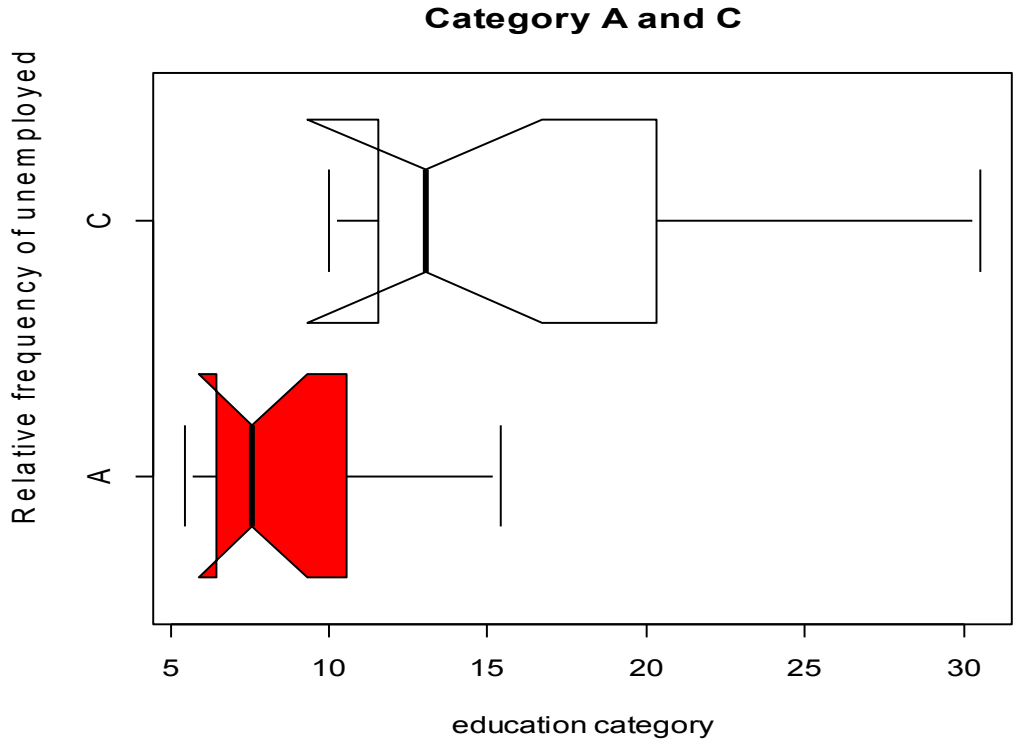
Παρατηρώντας κανείς το Διάγραμμα 4.2 μπορεί να δει πως στις περισσότερες περιπτώσεις των κατηγοριών του πίνακα των ποσοστών των ανέργων για τις κατηγορίες μορφωτικού επιπέδου το κάτω μέρος του "κουτιού" με τις εγκοπές αναδιπλώνει και δεν είναι όμοιο με το πάνω μέρος. Αυτό συμβαίνει γιατί το μέγεθος των εγκοπών είναι μεγαλύτερο από το ενδοτεταρτημοριακό εύρος. Ο τύπος που δίνει το μέγεθος των κουτιών με τις εγκοπές αυτές όπως εξηγήθηκε παραπάνω δίνεται από ένα τύπο που εξαρτάται από το μέγεθος του δείγματος. Το γεγονός ότι οι εγκοπές αναδιπλώνουν σημαίνει ότι το μέγεθος τους, δηλαδή το διάστημα εμπιστοσύνης για τους μέσους που δίνει το διάγραμμα είναι μεγαλύτερο από το ενδοτεταρτημοριακό εύρος του δείγματος. Η R το επιστρέφει σαν παρατήρηση κιόλας μετά την εντολή κατασκευής του παραπάνω διαγράμματος. Ο μόνος τρόπος για να αντιμετωπιστεί αυτό είναι να μεγαλώσουμε το μέγεθος του δείγματος οπότε θα μικρύνει το διάστημα εμπιστοσύνης και δε θα υπάρχουν οι αναδιπλώσεις αυτές. Σε περιπτώσεις όμως όπως αυτή που το μέγεθος του δείγματος είναι συγκεκριμένο δεν μπορεί κανείς να κάνει

κάτι για να αποφύγει τις αναδιπλώσεις.

Το Διάγραμμα 4.2 για τις διάφορες κατηγορίες μορφωτικού επιπέδου δίνει αποτελέσματα παρόμοια με εκείνα του κανονικού θηκογράμματος. Σε αυτό το διάγραμμα όμως μπορεί κανείς να συγκρίνει καλύτερα τους μέσους της κάθε κατηγορίας και να συμπεράνει αν η διαφορά τους είναι στατιστικά σημαντική. Όπως φαίνεται στις περισσότερες κατηγορίες υπάρχει επικάλυψη του διαστήματος εμπιστοσύνης για κάθε ένα μέσο. Για αυτό και σε επίπεδο σημαντικότητας 95% δεν μπορούμε να δεχθούμε ότι οι μέσοι όροι από κάθε κατηγορία διαφέρουν. Μπορεί όμως κανείς να συμπεράνει ότι υπάρχει κάποια διαφοροποίηση ανάμεσα στις 2 πρώτες κατηγορίες A και B με τις 3 επόμενες C, D, E καθώς δεν έχουμε παρά ελάχιστες επικαλύψεις. Ειδικά για την πρώτη με την τρίτη κατηγορία η επικάλυψη είναι αρκετά μικρή έως καθόλου.
Ακολουθεί το Διάγραμμα 4.3 για τις δύο κατηγορίες A και C ξεχωριστά σε οριζόντια μορφή για την καλύτερη οπτική παρατήρηση τους.

Όπως φαίνεται στο Διάγραμμα 4.3 για τις δύο αυτές κατηγορίες δεν υπάρχει επικάλυψη άρα μπορούμε να υποθέσουμε ότι οι μέσοι σε αυτή την περίπτωση διαφέρουν καθώς τα διαστήματα εμπιστοσύνης τους έχουν κενή τομή. Δηλαδή σε επίπεδο σημαντικότητας 95% οι κατηγορίες A και C μορφωτικού επιπέδου των ανέργων διαφέρουν.

```
boxplot(A,C,notch=TRUE,horizontal=TRUE,col=c("red","white"),names=(c("A","C")),xlab="educatio
n category",ylab="Relative frequency of unemployed")
title(main="Category A and C")
```



Διάγραμμα 4.4: Θηκογράφημα με εγκοπές για την πρώτη και τρίτη κατηγορία μορφωτικού επιπέδου της σχετικής συχνότητας των ανέργων επί 100 για τα έτη 2001 ως 2014

4.4.3 Διαγράμματα με την μέθοδο των πυρήνων για κάθε κατηγορία (density plot)

Η μέθοδος των πυρήνων είναι μία μέθοδος εκτίμησης της συνάρτησης πυκνότητας πιθανότητας μιας τυχαίας μεταβλητής. Ένας πυρήνας (kernel) είναι μία συνάρτηση, έστω K(x) η οποία έχει τις εξής ιδιότητες:

- $K(x) > 0$
- $\int_{-\infty}^{\infty} K(x)\,dx = 1$
- $\int_{-\infty}^{\infty} x \cdot K(x)\,dx = 0$
- $0 < \int_{-\infty}^{\infty} x^2 \cdot K(x)\,dx < +\infty$

Η μέθοδος των πυρήνων γενικά εκτιμά την σ.π.π. μιας τυχαίας μεταβλητής έστω Χ τοποθετώντας έναν πυρήνα (kernel) σε κάθε παρατήρηση καθιστώντας την κέντρο και στη συνέχεια "μετρώντας" πόσες παρατηρήσεις είναι κάτω από την καμπύλη του πυρήνα και γύρω από την κάθε παρατήρηση. Οι περισσότεροι πυρήνες είναι συμμετρικές κατανομές οπότε δίνουμε μεγαλύτερο βάρος στα σημεία γύρω από το κέντρο και λιγότερο στις πιο απομακρυσμένες παρατηρήσεις μέσα στον πυρήνα. Η εκτιμήτρια της σ.π.π. θα είναι το άθροισμα των βαρών που προκύπτουν. Δηλαδή δίνεται από τον τύπο

$$\hat{f}(x) = \frac{1}{n} * \sum_{i=1}^{n} \frac{1}{h} * K\left(\frac{x - x_i}{h}\right)$$

όπου h μία παράμετρος που εκφράζει το πλάτος του πυρήνα και x_i οι παρατηρήσεις από το δείγμα μας.

Στην R οι δυνατοί πυρήνες που μπορούμε να διαλέξουμε είναι οι:

- Gaussian
- Epanenchnikov
- Biweight
- Triweight
- Uniform

Το διάγραμμα της εκτιμήτριας της σ.π.π. συνήθως δεν διαφέρει ιδιαίτερα ως προς τις διάφορες επιλογές πυρήνα εκτός από τις περιπτώσεις που χρησιμοποιούμε την ομοιόμορφη (uniform) κατανομή, όπου τότε προκύπτουν γραφήματα λιγότερο λεία. Όμως και πάλι δεν χάνουμε την γενική εικόνα της κατανομής η οποία είναι ίδια με τις υπόλοιπες. Συμπερασματικά λοιπόν μπορούμε να πούμε ότι η επιλογή του πυρήνα δεν είναι τόσο σημαντική για την εύρεση της εκτιμήτριας της σ.π.π με τη μέθοδο των πυρήνων. Για αυτό και για την κατασκευή του γραφήματος θα χρησιμοποιηθεί ο κανονικός πυρήνας που είναι και η προεπιλογή της R. Σε διαφορετική περίπτωση μέσα στην εντολή για την κατασκευή του διαγράμματος θα πρέπει να δοθεί το όρισμα kernel="όνομα κατανομης".

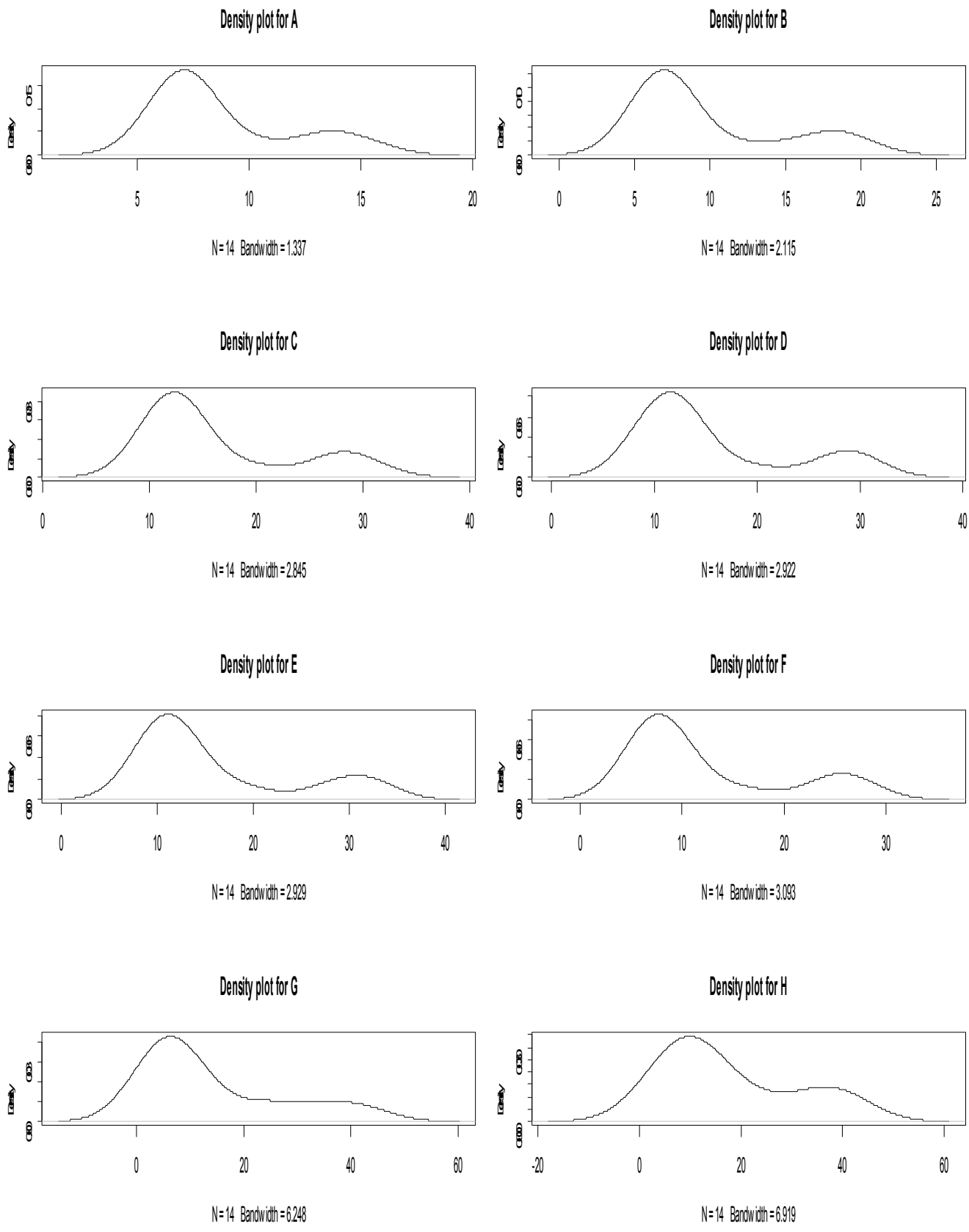
Όσον αφορά το πλάτος του πυρήνα η επιλογή του θα πρέπει να είναι προσεκτική καθώς για πολύ μεγάλο πλάτος μπορεί να έχουμε υπερβολικά λεία εκτίμηση στην οποία χάνονται οι λεπτομέρειες και να χαθεί πληροφορία για την κατανομή ενώ για πολύ μικρό πλάτος μπορεί

να υπάρξει μην είναι καθόλου λεία και να μην είναι ξεκάθαρη η εικόνα της κατανομής.
Η θεωρητική τιμή για το βέλτιστο πλάτος προκύπτει από την ελαχιστοποίηση του μέσου ολοκληρωτικού τετραγωνικού σφάλματος (MISE) της εκτιμήτριας της σ.π.π. Το μέσο ολοκληρωτικό τετραγωνικό σφάλμα δίνεται από τον τύπο

$MISE(\hat{f}(x))=\int_{-\infty}^{\infty} MSE(\hat{f}(x))dx$ με $MSE(\hat{f}(x))=Bias^2(\hat{f}(x))+Var(\hat{f}(x))$, είναι δηλαδή το ολοκλήρωμα του αθροίσματος του τετραγώνου της μεροληψίας και του τετραγώνου της διασποράς της εκτιμήτριας της συνάρτησης πυκνότητας πιθανότητας. Η θεωρητική τιμή για το βέλτιστο πλάτος του πυρήνα είναι $h_{opt}=[\frac{R(K)}{nR(f'')\sigma_{\kappa}^4}]^{-\frac{1}{5}}$, όπου $R(K)=\int_{-\infty}^{\infty} K^2(t)dt$ και $R(f'')=\int_{-\infty}^{\infty}(f''(x))^2 dx$.
Ο θεωρητικός αυτός τύπος όμως έχει το μειονέκτημα ότι πρέπει να γνωρίζουμε την άγνωστη συνάρτηση πυκνότητας πιθανότητας f.
Αν στον τύπο για το βέλτιστο πλάτος θεωρήσουμε ότι η άγνωστη σ.π.π. είναι η $N(0,\sigma^2)$ και επιλέξουμε για πυρήνα τον κανονικό τότε προκύπτει ο τύπος $h_{opt}=1.59\hat{\sigma}n^{-\frac{1}{5}}$, όπου $\hat{\sigma}$ η εκτιμήτρια της τυπικής απόκλισης με βάση τις παρατηρήσεις που διαθέτουμε.
Εναλλακτικά όταν υπάρχουν έκτροπες τιμές στα δεδομένα μας και ως μία πρώτη εκτίμηση θεωρούμε ότι η κατανομή είναι μεν συμμετρική αλλά πιθανόν με παχύτερες ουρές (π.χ. Student) χρησιμοποιούμε τον ακόλουθο τύπο για το βέλτιστο πλάτος $h_{opt}=0.9\,min(\hat{\sigma},\frac{IQR}{1.349})\cdot n^{-\frac{1}{5}}$, όπου IQR το ενδοτεταρτημοριακό εύρος των δεδομένων μας.
Ο συγκεκριμένος τύπος είναι προεπιλογή της R χρησιμοποιώντας τη συνάρτηση "density".

Για την κατασκευή των γραφημμάτων αυτών θα δοθεί στην R ο ακόλουθος κώδικας που κατασκευάζει τις προσεγγίσεις της σ.π.π. για κάθε κατηγορία μορφωτικού επιπέδου με τη μέθοδο των πυρήνων σε ένα παράθυρο.

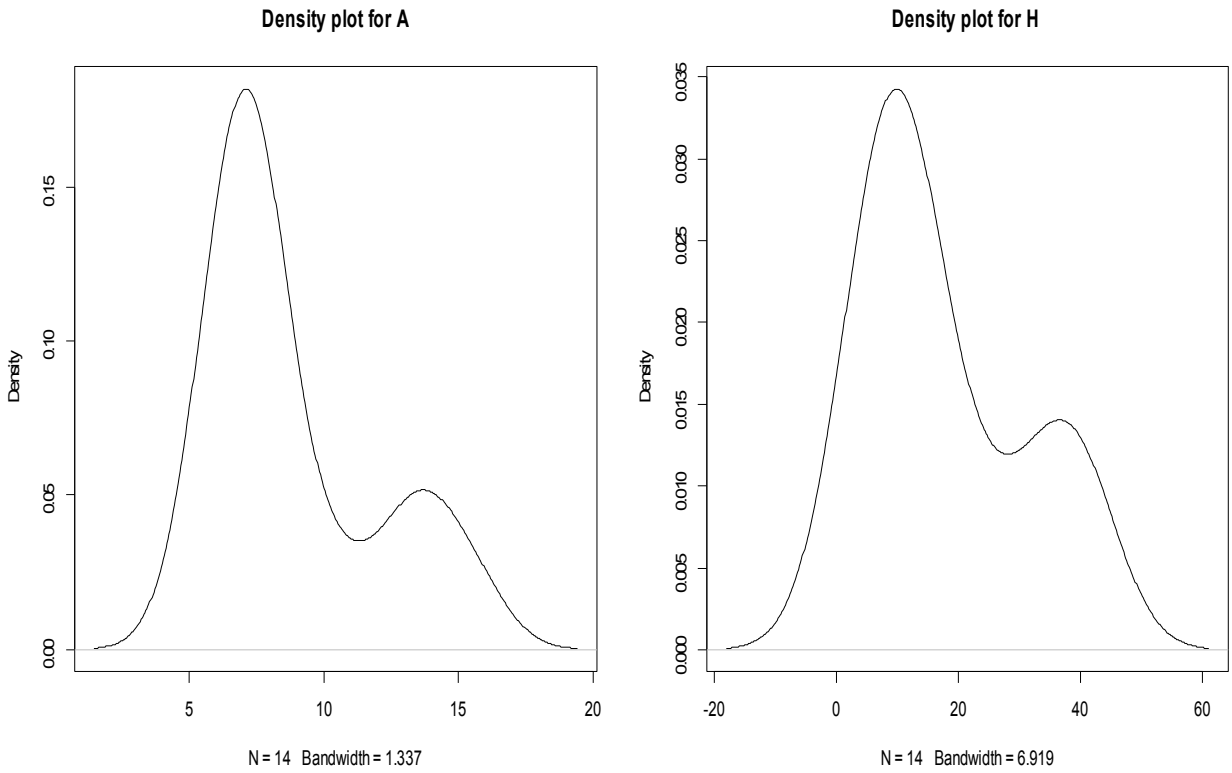
```
par(mfrow=c(4,2))
plot(density(A),main="Density plot for A")
plot(density(B),main="Density plot for B")
plot(density(C),main="Density plot for C")
plot(density(D),main="Density plot for D")
plot(density(E),main="Density plot for E")
plot(density(F),main="Density plot for F")
plot(density(G),main="Density plot for G")
plot(density(H),main="Density plot for H")
```



Διάγραμμα 4.5: Διαγράμματα προσέγγισης με τη μέθοδο των πυρήνων της σχετικής συχνότητας των ανέργων επί 100 κάθε κατηγορίας μορφωτικού επιπέδου από το 2001 ως το 2014

Τα Διαγράμματα 4.5 εκτιμούν την γραφική παράσταση της σ.π.π. του ποσοστού ανέργων ανά κατηγορία μορφωτικού επιπέδου. Παρατηρώντας τα μπορεί να δει κανείς ότι η γενική μορφή των κατανομών δε διαφέρει ιδιαίτερα πολύ ανά περίπτωση. Είναι μία κατανομή με δύο κορυφές μία μεγαλύτερη κοντά στην τιμη 10 και μία μικρότερη η οποία βρίσκεται μετά την τιμή 20. Ανάλογα όμως με την κάθε κατηγορία η δεύτερη κορυφή είναι πριν ή μετα την τιμή 30. Το ύψος των κορυφών διαφέρει επίσης και μειώνεται στις τελευταίες περιπτώσεις σε σχέση με τις πρώτες. Συγκεκριμένα για την κατηγορία Α το ύψος της μεγάλης κορυφής είναι κοντά στο 0.15 ενώ στην τελευταία κοντά στο 0.03. Επίσης το πλάτος που έχει θεωρήσει η R σε κάθε περίπτωση ιδανικό διαφέρει πολύ για κάθε κατανομή.
Στη μέθοδο των πυρήνων πολλές φορές αν έχουμε απομακρισμένες τιμές για μικρό πλάτος μπορεί να καταλήξουν να δημιουργούν μία πλασματική κορυφή στην άκρη της κατανομής. Από τα παραπάνω γραφήμματα γνωρίζουμε ότι η κατηγορία Ε παρουσίαζε μία ακραία τιμή προς τα πάνω και για αυτό παρατηρούμε ότι σε αυτή την περίπτωση η δεύτερη κορυφή είναι μετακινημένη αρκετά πιο δεξιά στον άξονα, μετά την τιμή 30 και μπορούμε να θεωρήσουμε ότι αυτό οφείλεται στην ακραία τιμή καθώς κοιτώντας το δείγμα δεν έχουμε πληθώρα παρατηρήσεων σε αυτή την τιμή. Οι δύο τελευταίες κατηγορίες είχαμε δει ότι έχουν μεγάλο εύρος τιμών στα ποσοστά των ανέργων ούτως ή άλλως οπότε δεν είναι παράλογο να βρίσκεται τόσο δεξιά η δεύτερη κορυφή. Επίσης παρατηρείται ότι η κατανομή έχει και κάποιες αρνητικές τιμές στις δύο τελευταίες περιπτώσεις, κάτι αδύνατο αφού μιλάμε για ποσοστά ανέργων. Και σε αυτή την περίπτωση η κατανομή φαίνεται να εκτίνεται προς τα αρνητικά λόγω της ύπαρξης τιμών κοντά στο 0 ενώ στις πραγματικότητα δεν υπάρχουν αρνητικές παρατηρήσεις και για αυτό η κατανομή δε θα έπρεπε να εκτείνεται προς εκείνη την κατεύθυνση. Για να αποφευχθεί αυτό αρκεί κανείς να μικρύνει το πλάτος του πυρήνα.
Ενδεικτικά θα κατασκευαστούν οι γραφικές της πρώτης κατηγορίας ξανά για να παρατηρηθεί το διαφορετικό ύψος της κορυφής με την κατηγορίας Η. Στη συνέχεια θα κατασκευαστεί το διάγραμμα για την κατηγορία Η για τέσσερα διαφορετικά πλάτη ένα από τα οποία είναι η προεπιλογή της R.
Ο κώδικας για αυτά τα διαγράμματα θα είναι

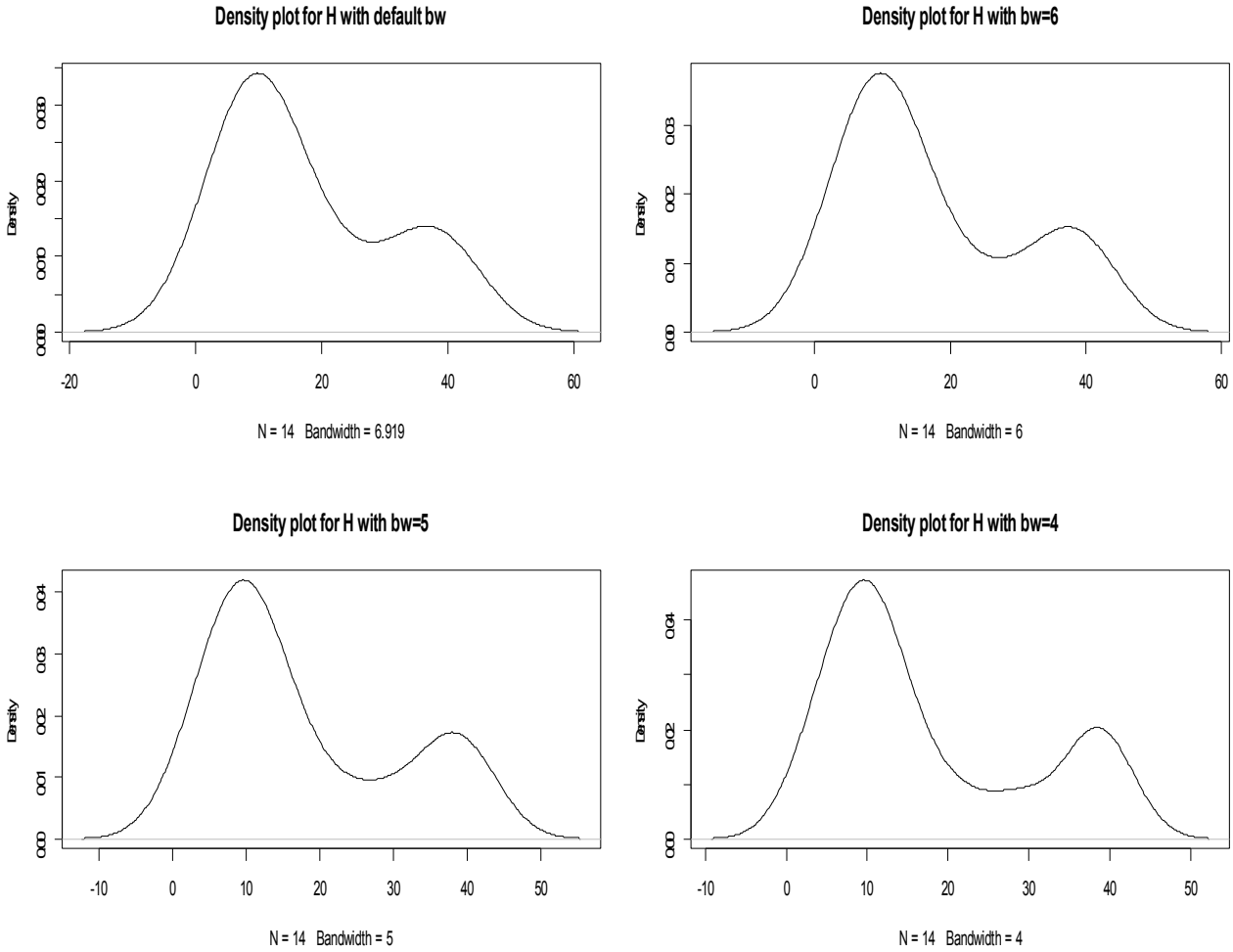
```
par(mfrow=c(1,2))
plot(density(A),main="Density plot for A")
plot(density(H),main="Density plot for H ")
```



Διάγραμμα 4.6: Διάγραμμα των προσσεγίσεων των κατανομών με τη μέθοδο των πυρήνων της σχετικής συχνότητας των ανέργων επί 100 για την κατηγορία μορφωτικού επιπέδου Α και Η για τα έτη 2001 ως 2014

Και ο κώδικας για τη δεύτερη ομάδα διαγραμμάτων θα είναι

```
par(mfrow=c(2,2))
plot(density(H),main="Density plot for H with default bw")
plot(density(H,bw=6),main="Density plot for H with bw=6")
plot(density(H,bw=5),main="Density plot for H with bw=5")
plot(density(H,bw=4),main="Density plot for H with bw=4")
```



Διάγραμμα 4.7: Το διάγραμμα με τη μέθοδο των πυρήνων για την κατηγορία Η μορφωτικού επιπέδου της σχετικής συχνότητας των ανέργων επί 100 για τα έτη 2001 ως 2014 για 4 διαφορετικά πλάτη

Πλέον μπορεί κανείς να δει από τα Διαγράμματα 4.7 την διαφορά στο ύψος της κορυφής της κατηγορίας Α και της κατηγορίας Η. Επίσης με την μείωση του πλάτους (bw) φαίνεται να μαζεύεται το εύρος της γραφικής και να συναντώνται λιγότερες αρνητικές τιμές ενώ η μορφή της δεν έχει αλλάξει. Αν μικρύνουμε όμως κι άλλο το πλάτος τότε δημιουργούνται λόγω της επαναληψιμότητας πολλές πτυχώσεις και χάνεται η γενική εικόνα του γραφήματος.

Σε περιπτώσεις όπως εδώ που παρατηρούνται έκτροπες τιμές συνήθως χρησιμοποιούμε μεταβλητό πλάτος ανάλογα με την πυκνότητα των παρατηρήσεων σε εκείνο το σημείο. Για παράδειγμα μπορεί αν έχουμε δύο απομακρυσμένες τιμές, για μικρό πλάτος (h) να κατέληγαν να δημιουργήσουν ένα "λοφάκι" στην άκρη της κατανομής δίνοντας μία παραπάνω πλασματική κορυφή. Με την εναλλαγή του h αυτό θα αντιμετωπιζόνταν.
Για αυτό υπάρχει η μέθοδος των πυρήνων με μεταβλητό παράθυρο (πλάτος). Για τα προηγούμενα δεδομένα παρατηρούνται διάφορες κορυφές που όμως δεν είναι ξεκάθαρο για όλες τις περιπτώσεις αν είναι πραγματικές. Για αυτό θα δημιουργήσουμε τα αντίστοιχα

γραφήματα για την προεπιλογή των αντίστοιχων πλατών σε κάθε περίπτωση με τη μέθοδο των πυρήνων με μεταβλητό παράθυρο.
Η μεθοδολογία της μεθόδου αυτής είναι ως εξής.

- Υπολογισμός της $\hat{f}(x)$ χρησιμοποιώντας κάποιο σταθερό h.
- Εύρεση του γεωμετρικού μέσου των διαφόρων σημείων της εκτιμήτριας της σ.π.π. $G=[\prod_{i=1}^{n}\hat{f}(x_i)]^{\frac{1}{n}}$.
- Εύρεση του $\lambda_i=\sqrt{\frac{G}{\hat{f}(x_i)}}$.
- Υπολογισμός $h(x_i)=h\cdot\lambda_i$.

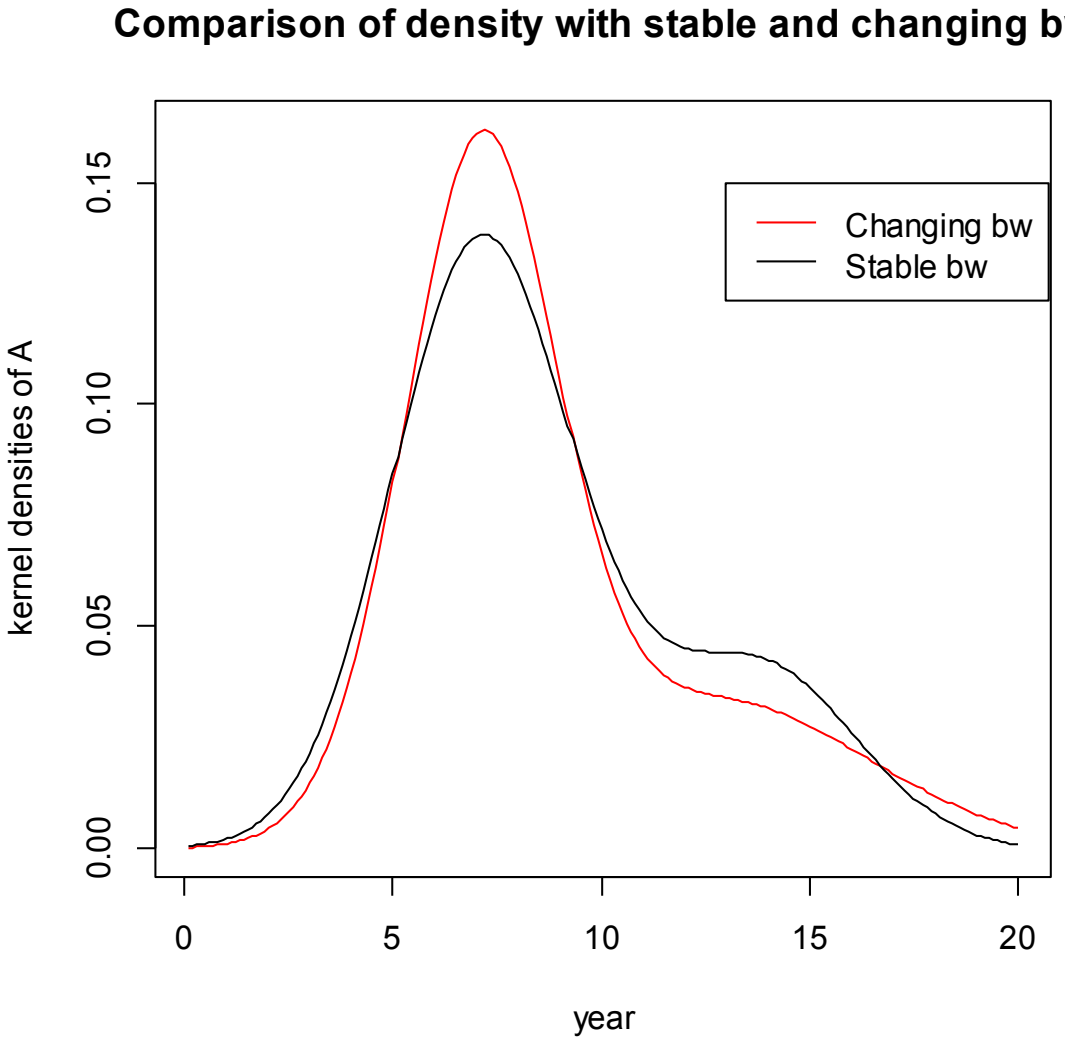
Για την πραγματοποίηση αυτής της μεθόδου θα πρέπει να δημιουργηθεί στην R μία συνάρτηση υπολογισμού της εκτιμήτριας της σ.π.π. με τη μέθοδο των πυρήνων χωρίς τη βοήθεια έτοιμων πακέτων. Θα δημιουργηθούν δύο συναρτήσεις που θα υπολογίζουν την εκτιμήτρια της σ.π.π. με σταθερό και με μεταβλητό πλάτος.

```
kerneldensity<- function(x,Y,h)
 {
y<-rep(1,length(x))
n<-length(Y)
i<-1
for (i in 1:length(x))
 {
s<-0
j<-1
for (j in 1:length(Y))
 {
s<- dnorm((x[i]-Y[j])/h,0,1)/h + s
}
y[i]<-s/n
}
y
}

kerneldensity2<- function(x,Y,h)
{
y<-rep(1,length(x))
n<-length(Y)
i<-1
for (i in 1:length(x))
{
s<-0
j<-1
for (j in 1:length(Y))
```

```
{
s<- dnorm((x[i]-Y[j])/h[j],0,1)/h[j] + s }
y[i]<-s/n
}
y
}
```

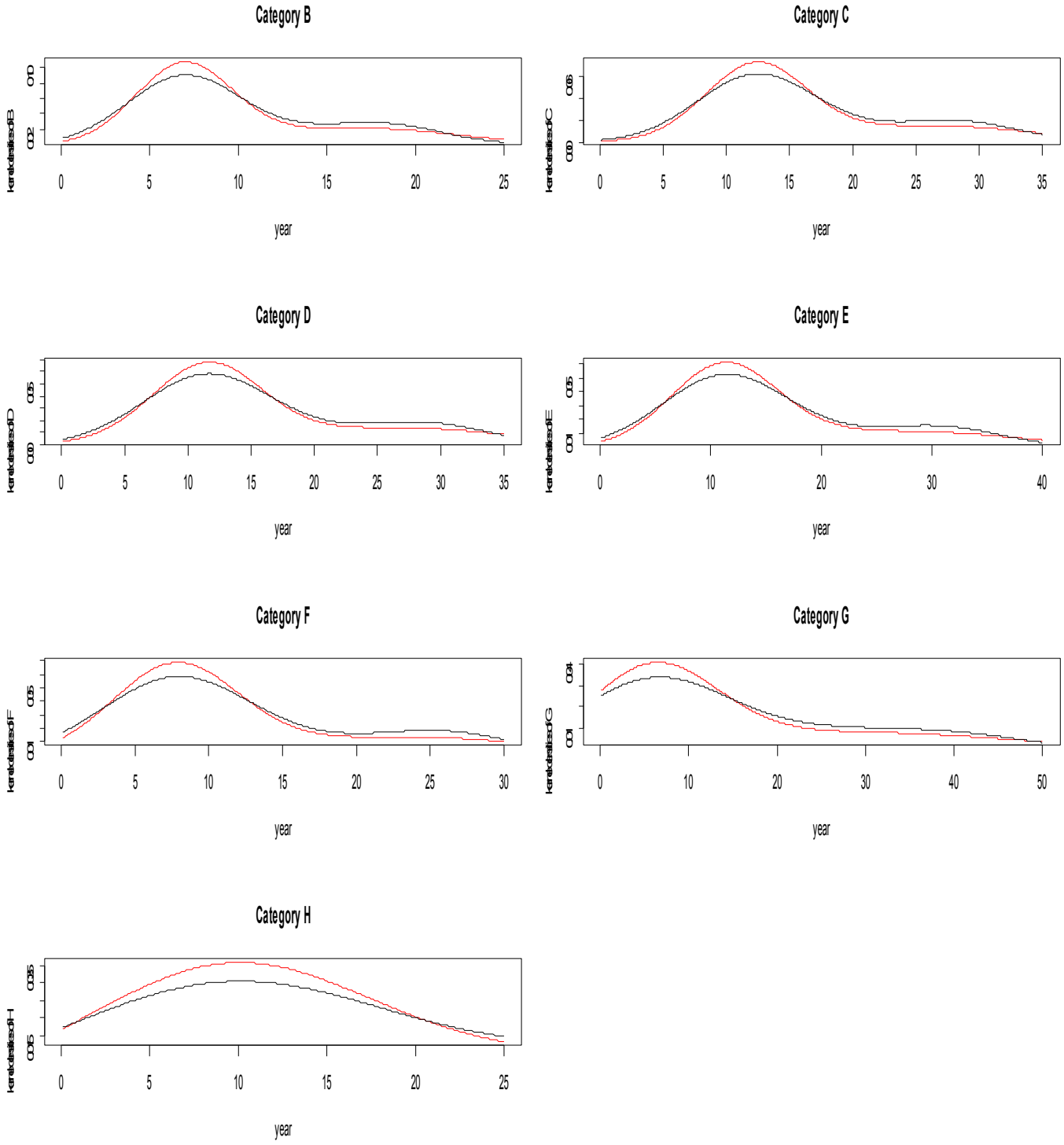
Η συνάρτηση “kerneldensity” υπολογίζει την $\hat{f}(x)$ με σταθερό πλάτος και η συνάρτηση “kerneldensity2” την $\hat{f}(x)$ για μεταβλητό πλάτος. Οι συναρτήσεις δέχονται σαν ορίσματα εκτός από το πλάτος h και τις παρατηρήσεις Y που θα είναι ένα διάνυσμα με τα δεδομένα από το 2001 ως το 2014 για την κάθε κατηγορία μορφωτικού επιπέδου καθώς και τη μεταβλητή x.
Στη συνέχεια θα γίνει η εφαρμογή της μεθόδου των πυρήνων με μεταβλητό παράθυρο για κάθε μία από τις οχτώ κατηγορίες μορφωτικού επιπέδου και θα συγκριθούν στο ίδιο διάγραμμα η εκτίμηση της σ.π.π. με σταθερό πλάτος και με μεταβλητό πλάτος για ένα σύνολο σημείων x. Η τιμή του σταθερού πλάτους είναι η βέλτιστη τιμή που προκύπτει από τον θεωρητικό τύπο για την κανονική κατανομή σε κάθε περίπτωση. Ο κώδικας για την πρώτη κατηγορία επαγγέλματος λοιπόν είναι:

```
h_opt<-1.059*sd(A)*(length(A))^(-1/5)

a<-kerneldensity(A,A,h_opt)

g<-(prod(a))^(1/length(A))

lambda<-sqrt(g/a)

h_v<-h_opt*lambda

x<-seq(0.1,20,0.1)
plot(x,kerneldensity2(x,A,h=h_v), type='l',col="red",ylab="kernel densities of A",xlab="year")
lines(x, kerneldensity(x,A,h_opt))
legend(13,0.15,col=c("red","black"),lty=c(1,1),legend=c("Changing bw","Stable bw"))
title(main="Comparison of density with stable and changing bw")
```



Διάγραμμα 4.8: Διάγραμμα σύγκρισης του εκτιμητή της σ.π.π. με σταθερό και με μεταβλητό πλάτος των ποσοστών των ανέργων επί 100 της πρώτης κατηγορίας μορφωτικού επιπέδου για τα έτη 2001-2014

Με όμοιο τρόπο θα σχηματιστούν στο ίδιο γράφημα οι δύο προσεγγίσεις των σ.π.π. για όλες τις κατηγορίες μορφωτικού επιπέδου, με κόκκινο για το μεταβλητό πλάτος και με μαύρο για το σταθερό βέλτιστο πλάτος κάθε περίπτωσης.



Διάγραμμα 4.9: Σύγκριση εκτιμήτριας σ.π.π. με τη μέθοδο των πυρήνων με σταθερό και με μεταβλητό πλάτος για τα ποσοστά των ανέργων επί 100 για τα έτη 2001 ως 2014 για τις κατηγορίες μορφωτικού επιπέδου Β ως Η

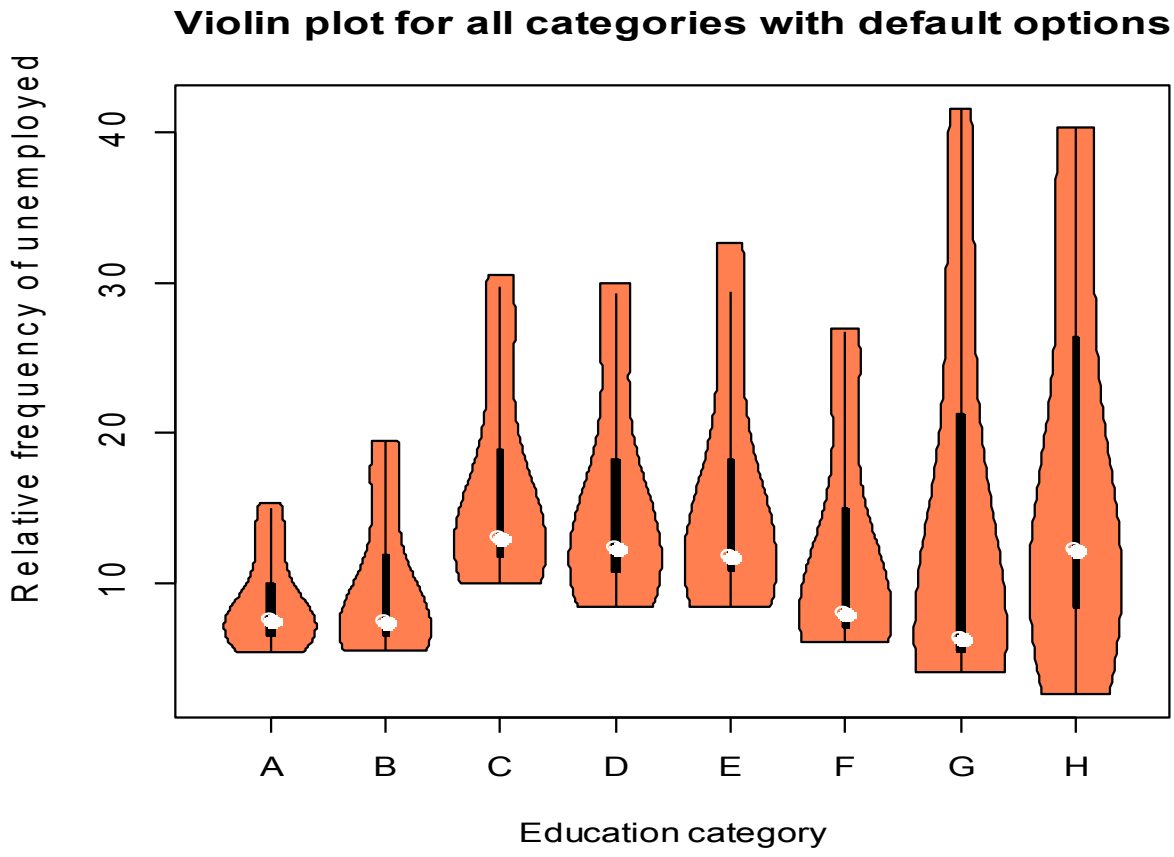
Όπως φαίνεται και στο Διάγραμμα 4.8 και στα Διαγράμματα 4.9 η μορφή της προσέγγισης της σ.π.π. δεν διαφέρει πολύ ανάμεσα σε αυτή με το σταθερό και με το μεταβλητό πλάτος αλλά στις περισσότερες κατηγορίες η δεύτερη κορυφή δεν υπάρχει πια ή είναι αρκετά πιο μικρή. Ήταν λοιπόν μία πλασματική κορυφή που είχε προκύψει από την μεγάλη απόσταση

των τιμών το 2009 με αυτές των επόμενων χρονιών.

4.4.4 Το διάγραμμα βιολί (violin plot)

Το διάγραμμα βιολί είναι ένας συνδυασμός θηκογράμματος (boxplot) και διαγράμματος προσέγγισης με τη μέθοδο των πυρήνων (kernel density plot). Η μορφή του διαγράμματος μοιάζει με εκείνη του θηκογράμματος με τη διαφορά ότι τα "κουτιά" έχουν σχεδιαστεί ώστε να είναι η προσέγγιση της συνάρτησης πυκνότητας πιθανότητας της κάθε κατηγορίας με την μέθοδο των πυρήνων στραμμένη και τοποθετημένη 2 φορές μία από τα αριστερά και μία από τα δεξιά. Το διάγραμμα αυτό μπορεί να κατασκευαστεί μέσω της εντολής vioplot( ) η οποία είναι μερος του πακέτου vioplot. Πρέπει λοιπόν να εγκατασταθεί πρώτα το πακέτο για την κατασκευή των διαγραμμάτων αυτών και στη συνέχεια να δοθεί η εντολή για την κατασκευή τους.

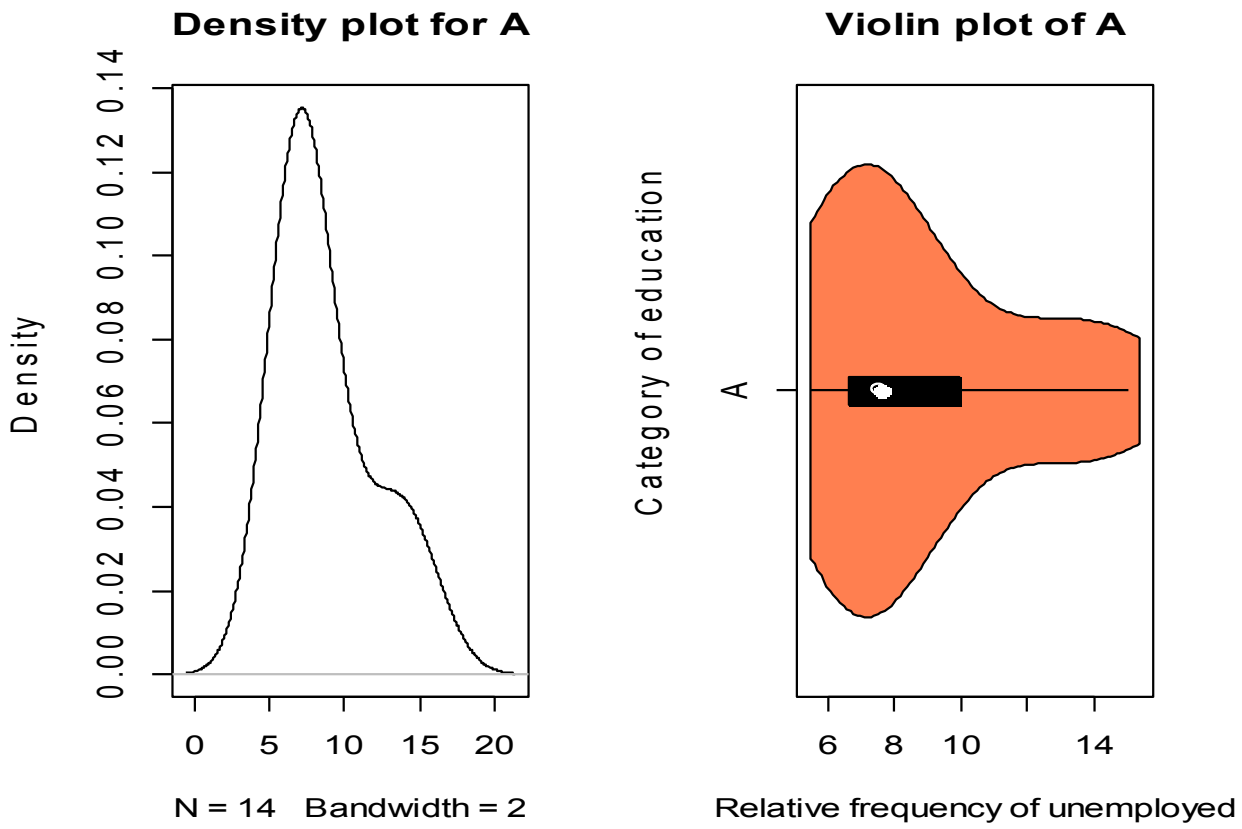
```
install.packages("vioplot")
library(vioplot)

vioplot(A,B,C,D,E,F,G,H,names=c("A","B","C","D","E","F","G","H"),col=c("coral"))
title(main="Violin plot for all categories with default options",xlab="Education
category",ylab="Relative frequency of unemployed")
```



Διάγραμμα 4.10: Violin plot για την σχετική συχνότητα των ανέργων ανά μορφωτικό επίπεδο για τα έτη 2001 ως 2014

Η διάμεσος στο Διάγραμμα 4.10 συμβολίζεται με την λευκή κουκίδα για αυτό και έγινε χρήση άλλου χρώματος και όχι του λευκού για να είναι εμφανής η θέση του ενώ το ενδοτεταρτημοριακό εύρος (IQR) είναι η πιο έντονη γραμμή στο κέντρο των γραφημάτων. Μπορεί κανείς να κατασκευάσει αυτά τα διαγράμματα και μέσω του πακέτου ggplot2 που παρέχει τη δυνατότητα κατασκευής μίας σειράς διαγραμμάτων όμως η εντολή σε εκείνο το πακέτο δεν δείχνει την διάμεσο και το ενδοτεταρτημοριακό εύρος αλλά χρειάζεται περαιτέρω ορίσματα για να γίνουν εμφανείς αυτές οι ποσότητες για αυτό και προτιμήθηκε το πακέτο vioplot. Το Διάγραμμα 4.10 κατασκευάστηκε με τις προεπιλεγμένες τιμές της R για την κατασκευή των διαγραμμάτων με τη μέθοδο των πυρήνων στο πλάι.
Από το Διάγραμμα 4.10 είναι εμφανές ότι οι δύο τελευταίες κατηγορίες μορφωτικού επιπέδου έχουν πολύ μεγαλύτερη μεταβλητότητα σε σχέση με τις άλλες ενώ η πρώτη κατηγορία φαίνεται να έχει τη μικρότερη μεταβλητότητα. Επίσης οι διάμεσοι σε όλες τις κατηγορίες φαίνεται να απέχουν αρκετά μεταξύ τους αλλά το μοτίβο της κατανομής που ακολουθούν τα ποσοστά των ανέργων σε κάθε κατηγορία μοιάζει αρκετά. Είναι επισφαλές όμως να πούμε ότι τα ποσοστά των ανέργων κατανέμονται με όμοιο τρόπο γύρω από τη διάμεσο σε κάθε κατηγορία μορφωτικού επιπέδου από αυτό το διάγραμμα καθώς δεν περιέχει πολλές λεπτομέρειες. Στην προηγούμενη παράγραφο μάλιστα που παρουσιάστηκαν οι προσεγγίσεις των σ.π.π. με τη μέθοδο των πυρήνων ξεχωριστά φάνηκε ότι ενώ οι προσεγγίσεις των σ.π.π. των κατηγριών μορφωτικού επιπέδου μοιάζουν δεν είναι ίδιες.
Η μορφή των κάθετων διαγραμμάτων πυκνότητας στο Διάγραμμα 4.10 μοιάζει με αυτή της προηγούμενης παραγράφου αλλά δεν είναι ιδιαίτερα εμφανής η δεύτερη κορυφή. Αν σχηματίσουμε το διάγραμμα για μία από τις παραπάνω κατηγορίες και το στρέψουμε μπορούμε να δούμε τις ομοιότητες με το αντίστοιχο διάγραμμα της προηγούμενης παραγράφου.

```
par(mfrow=c(1,2))
plot(density(A,bw=2),main="Density plot for A")
vioplot(A,names=c("A"),col=c("coral"),horizontal=TRUE)
title(main="Violin plot of A")
```



Διάγραμμα 4.11: Σύγκριση του διαγράμματος προσέγγισης με τη μέθοδο των πυρήνων και του διαγράματος violin plot για τις σχετικές συχνότητες των ανέργων ανά κατηγορία μορφωτικού επιπέδου για τα έτη 2001 ως 2014

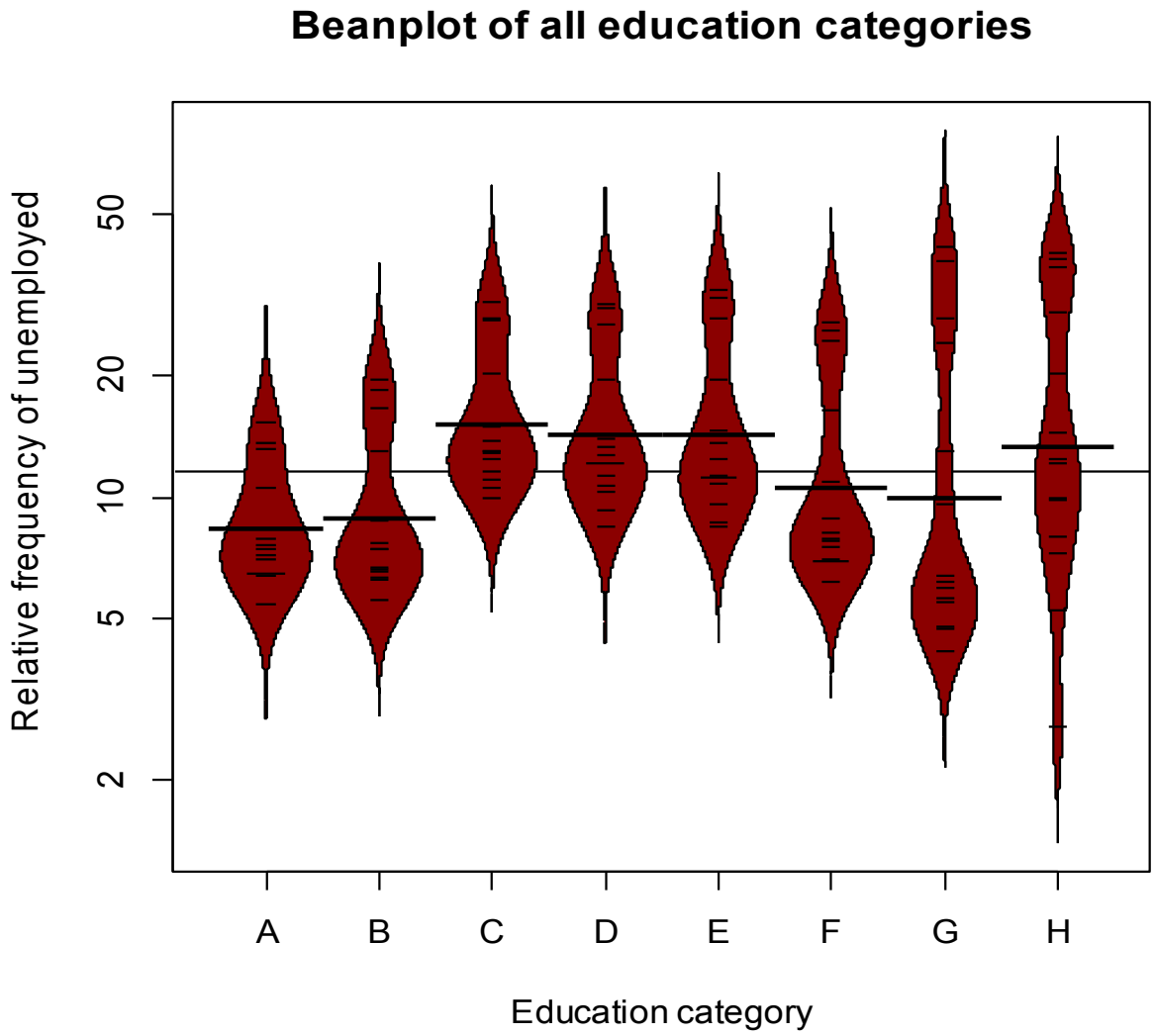
Το πλεονέκτημα αυτού του είδους διαγράμματος είναι ότι δίνει την ευκαιρία να συγκριθούν οι προσεγγίσεις των σ.π.π. με τη διάμεσο σε κάθε κατηγορία αλλά και το ενδοτεταρτημοριακό εύρος να παρουσιάζονται πάνω στη διάγραμμα. Ταυτόχρονα δίνεται η ευκαιρία να παρουσιαστούν δίπλα δίπλα τα αντίστοιχα διαγράμματα ώστε να γίνει πιο αισθητή η διαφοροποίηση ανά κατηγορία. Όμως δεν επιτρέπει την παρουσίαση τόσων λεπτομερειών όσο η μέθοδος των πυρήνων ή το θηκογράφημα ξεχωριστά.

4.4.5 Το διάγραμμα σε σχήμα φασολιού (beanplot)

Το όνομα του διαγράμματος "beanplot" έχει προκύψει από την εικόνα που παρουσιάζει αυτό το διάγραμμα που μοιάζει με φασόλι. Η γραφική της μεθόδου τον πυρήνων με τις καμπυλώσεις που κάνει συνήθως μπορεί να θεωρηθεί ότι είναι οι καμπυλώσεις που κάνει το φασόλι ενώ οι ευθείες γραμμές μοιάζουν με τα μπιζέλια μέσα σε αυτό.
Το διάγραμμα "bean plot" συνδιάζει το θηκόγραμμα με την διαγραμματική προσέγγιση με τη μέθοδο των πυρήνων (kernel density plot), όπως το violin plot συνδιάζοντας μαζί με αυτά τα δύο και ένα "rug plot", δηλαδή ένα διάγραμμα μίας διάστασης για τις παρατηρήσεις των δειγμάτων κάθε κατηγορίας. Έχει μορφή που μοιάζει με εκείνη του violin plot, δηλαδή αποτελείται από ένα κουτί του οποίου οι κάθετες πλευρές σχηματίζουν

τη μορφή της προσέγγισης της σ.π.π. με τη μέθοδο των πυρήνων με τη διαφορά ότι ενδιάμεσα υπάρχουν κάποιες λεπτές γραμμές που αναπαριστούν τις τιμές από το δείγμα. Επιπλέον υπάρχει μία πιο παχιά γραμμή που αναπαριστά τη διάμεσο. Τα διαγράμματα αυτά δεν περιέχουν πληροφορία για το ενδοτεταρτημοριακό εύρος αλλά περιέχουν όλο το εύρος των παρατηρήσεων σε αντίθεση με το violin plot.
Η εντολή για την κατασκευή του διαγράμματος αυτού περιέχεται στο πακέτο beanplot, το οποίο και θα πρέπει να εγκατασταθεί. Ο κώδικας για την κατασκευή του διαγράμματος 4.9 θα είναι:

```
install.packages("beanplot")
library(beanplot)

beanplot(A,B,C,D,E,F,G,H, names=c("A","B","C","D","E","F","G","H"),col="darkred",xlab="Education category",ylab="Relative frequency of unemployed")
title("Beanplot of all education categories")
```

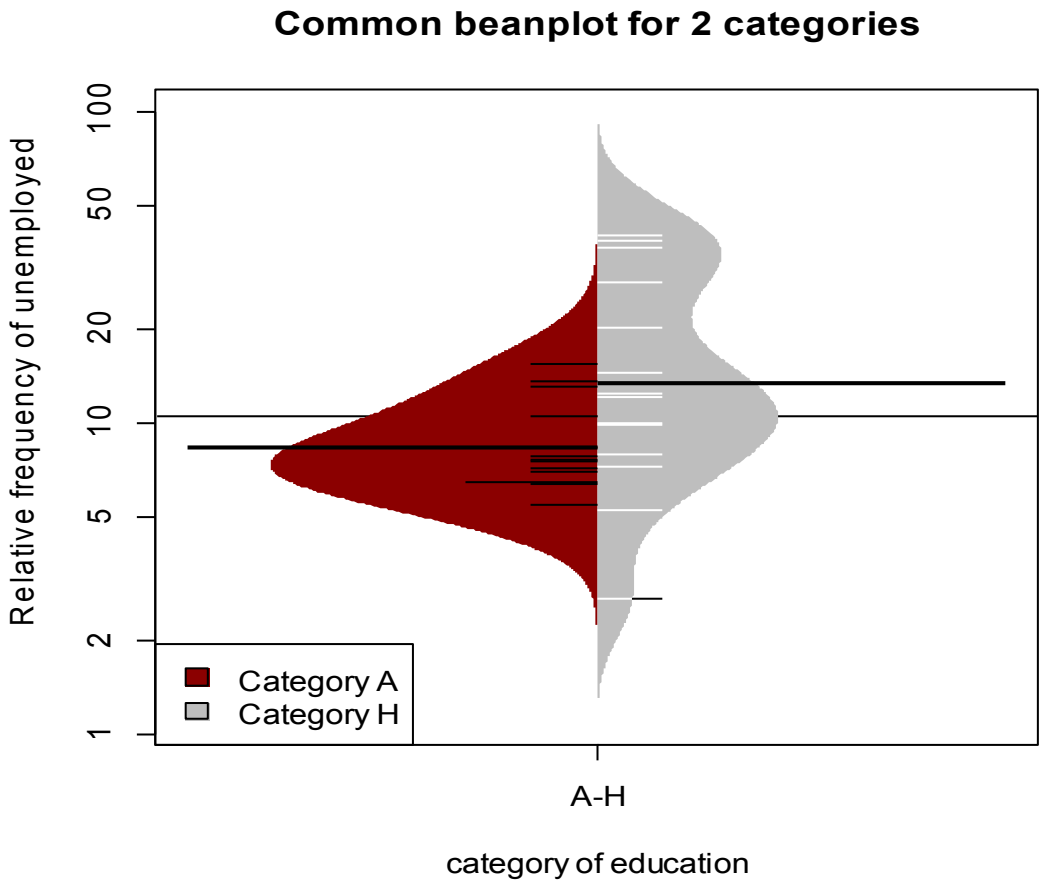


Διάγραμμα 4.12: Διάγραμμα beanplot για την σχετική συχνότητα των ανέργων επί 100 και ανά κατηγορία μορφωτικού επιπέδου για τα έτη 2001 ως 2014

Η μορφή της προσέγγισης της σ.π.π. είναι ακόμα πιο εμφανής στο Διάγραμμα 4.12 και επίσης πέρνουμε περισσότερη πληροφορία για τις τιμές των ποσοστών των ανέργων σε κάθε κατηγορία μορφωτικού επιπέδου. Μπορεί πλέον κανείς να δει ξεκάθαρα ότι υπάρχουν χρονιές που το ποσοστό των ανέργων που δε έχουν πάει καθόλου σχολείο είναι μεταξύ του 20-50% σε ποσοστό ενώ για τα άτομα που έχουν διδακτορικό ή μεταπτυχιακό τίτλο δεν συναντάται ποτέ ποσοστό μεγαλύτερο του 20%. Επιπλέον στο Διάγραμμα 4.12 φαίνεται και ο αριθμός των παρατηρήσεων που δεν το έδινε κανένα από τα παραπάνω διαγράμματα.

Στη συνέχεια θα σχηματίσουμε ένα διάγραμμα που θα είναι συνδυαστικό για τις 2 ακραίες κατηγορίες Α και Η, δηλαδή τους άνεργους κάτοχους διδακτορικού ή μεταπτυχιακού τίτλου και τους ανέργους που δεν έχουν πάει ποτέ σχολείο. Το διάγραμμα 4.13 θα περιέχει τις 2 κατηγορίες στο ίδιο "beanplot" ώστε να φανεί η διαφοροποίηση της συνάρτησης κατανομής και των διαμέσων τους.

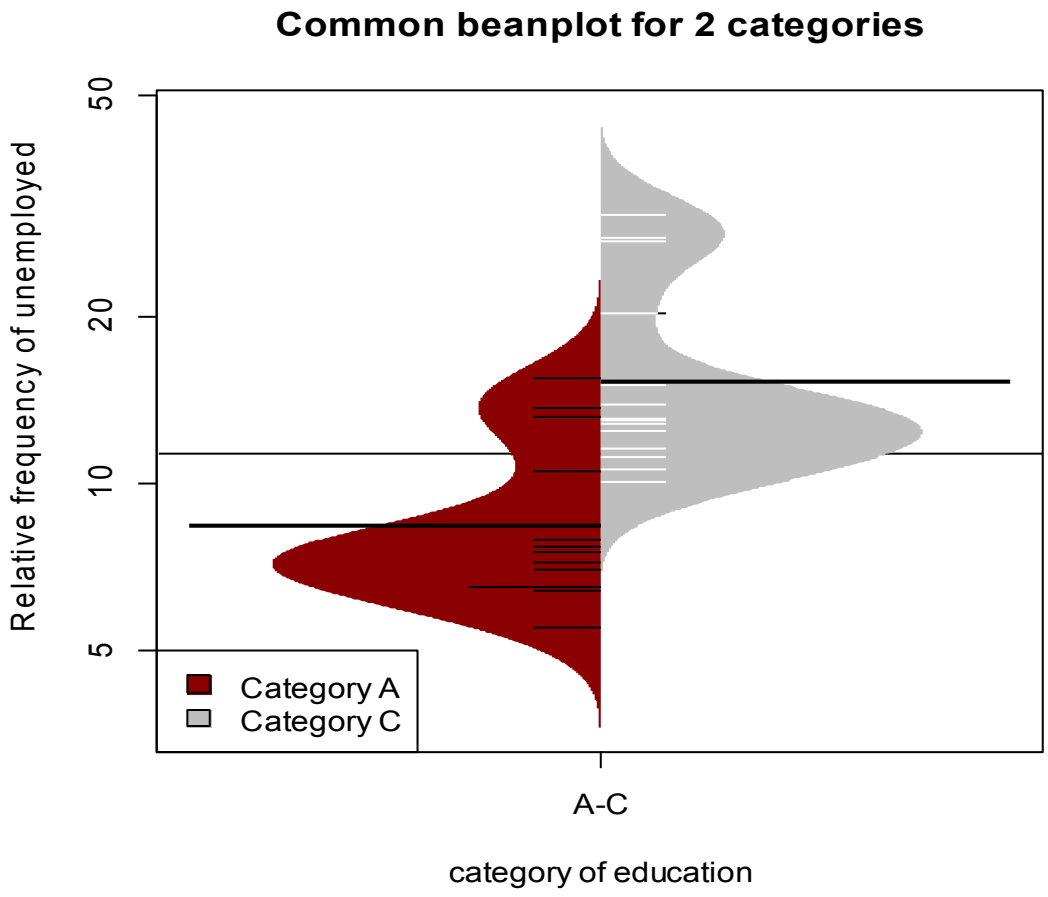
```
par(lend=1)
beanplot(A,H,side="both",border=NA,names=("A-H"),  col=list("darkred",c("grey","white")))
title(main="Common beanplot for 2 categories",xlab="category of education",ylab="Relative frequency of unemployed")
legend("bottomleft", fill = c("darkred", "grey"),legend=c("Category A","Category H"))
```



Διάγραμμα 4.13: Διάγραμμα δύο κατηγοριών μορφωτικού επιπέδου της Α και της Η στο ίδιο διάγραμμα τύπου "beanplot" για τη σχετική συχνότητα των ανέργων για τα έτη 2001 ως 2014

Στο Διάγραμμα 4.13 φαίνεται εμφανώς η διαφοροποίηση των 2 κατηγοριών. Μπορεί η διαφορά των διαμέσων να μη θεωρείται στατιστικά σημαντική αλλά το εύρος των τιμών είναι αισθητά διαφορετικό και η προσέγγιση της σ.π.π. τους είναι αρκετά διαφορετική για τα πλάτη που έχει επιλέξει η R και για τις δύο περιπτώσεις. Στην περίπτωση των ανέργων που δεν έχουν πάει σχολείο η δεύτερη κορυφή σχηματίζεται ξεκάθαρα λόγω της ύπαρξης τιμών σε εκείνες τις τιμές των ποσοστών στο 20% με 50% ενώ στην άλλη κατηγορία η δεύτερη κορυφή δεν φαίνεται πλέον σχεδόν καθόλου. Αν δοκιμάσουμε να μικρύνουμε το πλάτος δίνοντας διάφορα ορίσματα παρατηρούμε ότι με τον σχηματισμό της δεύτερης κορυφής για την κατηγορία των ανέργων Α έχουμε ταυτόχρονα σχηματισμό επιπλέον κορυφών για την κατηγορία Η.
Ακολουθεί και η σύγκριση της κατηγορίας των κατόχων διδακτορικού ή μεταπτυχιακού τίτλου με τους απόφοιτους ανώτατου εκπαιδευτικού ιδρύματος.

```
beanplot(A,C,side="both",border=NA,names=("A-C") , col=list("darkred",c("grey","white")))
title(main="Common beanplot for 2 categories",xlab="category of education",ylab="Relative frequency of unemployed")
legend("bottomleft", fill = c("darkred", "grey"),legend=c("Category A","Category C"))
```



Διάγραμμα 4.14: Διάγραμμα δύο κατηγοριών μορφωτικού επιπέδου της Α και της C στο ίδιο διάγραμμα τύπου "beanplot" για τη σχετική συχνότητα των ανέργων για τα έτη 2001 ως 2014

Στο Διάγραμμα 4.14 βλέπει κανείς ότι τα ποσοστά για τους κατόχους διδακτορικού ή μεταπτυχιακού τίτλου είναι χαμηλότερα σε σχέση με αυτά της άλλης κατηγορίας και το σύνολο των παρατηρήσεων σε κάθε περίπτωση φαίνεται να έχει παρόμοια κατανομή ενώ το εύρος των τιμών δεν έχει μεγάλη διαφορά. Απλά τα ποσοστά των ανέργων της κατηγορίας C είναι αναλογικά ψηλότερα σε σχέση με αυτά της κατηγορίας A.

5. Απασχολούμενοι ανά επαγγελματικό τομέα και ανά κλάδο οικονομικής δραστηριότητας

Σε αυτό το κεφάλαιο θα γίνει μία μελέτη για την συχνότητα των απασχολούμενων σε κάθε κλάδο οικονομικής δραστηριότητας με τη βοήθεια διαφόρων διαγραμμάτων που παρέχει η R. Έτσι θα φανεί ποιοι κλάδοι δραστηριοποιούνται ακόμα σε σχέση με τους άλλους στην Ελλάδα και τη διαφορά που υπάρχει για κάθε έναν από το 2001 ως το 2014. Στη συνέχεια θα διαπιστωθεί ποιος από τους τρεις ο πρωτογενής, ο δευτερογενής ή ο τριτογενής κλάδος οικονομίας παρουσίασε τη λιγότερη δυνατή ύφεση με την πάροδο των ετών.

5.1 Παρουσίαση των δεδομένων

5.1.1 Τα δεδομένα ανά επαγγελματικό τομέα

Σε αυτό το κεφάλαιο θα μελετηθούν δύο ειδών δεδομένα. Αρχικά θα μελετηθεί ο αριθμός των εργαζομένων που απασχολεί κάθε επαγγελματικό τομέα χωριστά. Η μελέτη θα γίνει σε 16 κατηγορίες που θεωρείται ότι συμπυκνώνουν την πληθώρα των εργαζομένων στην Ελλάδα. Οι τομείς αυτοί είναι οι εξής:

- Γεωργία/Κτηνοτροφία/Αλιεία/Δασοκομία
- Ορυχεία/Λατομεία
- Παροχή ηλεκτρικού ρεύματος/πετρελαίου/φυσικού αερίου/νερού
- Μεταποιητικές βιομηχανίες
- Κατασκευές
- Εμπόριο και επισκευή προϊόντων
- Χρηματοπιστωτικοί και ασφαλιστικοί οργανισμοί
- Διαχείρηση ακίνητης περιουσίας/Εκμισθώσεις
- Δημόσια διοίκηση/Άμυνα/Κοινωνικές ασφαλίσεις
- Υγεία και κοινωνική μέριμνα
- Ενημέρωση και επικοινωνία/Μεταφορές
- Εκπαίδευση
- Ξενοδοχεία και εστιατόρια
- Παροχή υπηρεσιών
- Επεξεργασία λυμάτων, διαχείρηση αποβλήτων και δρασηριότητες εξυγίανσης

Η τελευταία κατηγορία επαγγελματικού κλάδου δεν απασχολούσε σημαντικό αριθμό εργαζομένων τις πρώτες χρονιές του δείγματος καθώς δεν ήταν ανεπτυγμένος μέχρι πρόσφατα. Έτσι τις πρώτες χρονιές θεωρείται ότι δεν απασχολεί αριθμό εργαζομένων πριν από το 2008.

Τα δεδομένα αυτά είναι σε χιλιάδες εργαζομένους. Κάθε στήλη του πίνακα συχνοτήτων συμβολίζεται με K_i όπου i είναι ο αριθμός της κάθε στήλης. Έτσι η στήλη K_1 περιέχει τα δεδομένα για τον αριθμό των εργαζομένων που απασχολεί ο κλάδος Γεωργίας/Κτηνοτροφίας/Αλιείας/Δασοκομίας, η στήλη K_2 τα δεδομένα για τον αριθμό των εργαζομένων που απασχολεί ο κλάδος των Ορυχείων και Λατομείων και αντίστοιχα ορίζονται οι άλλες 15 στήλες.

Ο Πίνακας 5.1 παρουσιάζεται σε δύο κομμάτια λόγω της πληθώρας των στηλών που έχει.

	Κλάδος οικονομικής δραστηριότητας						
year	**K1**	**K2**	**K3**	**K4**	**K5**	**K6**	**K7**
2001	666,275	19,575	37,95	595,5	315,175	727,475	111,275
2002	652,775	19,675	37,775	590,85	326,075	739,275	105,95
2003	657,125	13,325	40,875	578,2	352,325	763,025	113,275
2004	543,475	14,9	39,025	572,025	357,05	772,05	115,875
2005	540,3	18,425	38,8	570,075	367,625	795,3	116,4
2006	530,85	18,125	41,5	570	367,3	806,375	119,3
2007	517,125	18,175	41,825	566,5	397,9	813,25	117,05
2008	513,775	16,95	34,575	545,025	397,275	840,1	121,35
2009	532,875	14,275	28,525	518,8	370,675	827,625	114,625
2010	544,2	13,275	25,95	468,275	319,625	799,3	115,775
2011	500,675	11,2	24,175	409,7	245,8	752,25	113,725
2012	480,5	11,075	26	351,425	200,9	663,7	111,1
2013	481,05	9,6	27,7	324,725	162,35	630,5	107,075
2014	479,9	11,25	27,5	316,5	151,6	625,6	93

Πίνακας 5.1a: Η συχνότητα απασχολούμενων σε χιλιάδες ανά επαγγελματικό κλάδο για τις πρώτες επτά κατηγορίες επαγγελμάτων για τα έτη 2001 ως 2014

	Κλάδος οικονομικής δραστηριότητας							
year	**K8**	**K9**	**K10**	**K11**	**K12**	**K13**	**K14**	**K15**
2001	228,275	313,75	188,025	265,45	269,8	276,925	186,55	0
2002	251,275	320,15	192,65	254,775	272,4	288,35	211,95	0
2003	253,5	329,9	189	267,15	287,5	292,3	214,925	0
2004	289,05	359,475	224,3	274,95	315,5	281	229,5	0
2005	294,5	352,65	224,05	276,3	315,925	305	227,5	0
2006	296,15	385,75	232,35	288,3	334,1	306,7	230	0
2007	303,425	387,475	244,8	269,6	325,525	319,25	240,55	0
2008	323,625	379,525	236	292,9	324,05	322,15	230,95	30,575
2009	318,35	377,05	234,425	304,55	328,75	320,95	232,475	30,475
2010	299,1	369,9	245,9	295,35	321,75	308,45	228,6	32,725
2011	296,95	354,875	238,225	271,85	304,375	295,675	206,425	26,35
2012	291,1	326,675	223,55	250	290,3	272,125	172,775	21,825
2013	262,4	325,05	212,825	249,15	274,775	259,2	163,125	22,25
2014	283,625	311,1	209,2	247,775	290,2	297,1	167,75	22,8

Πίνακας 5.1b: Η συχνότητα των απασχολούμενων σε χιλιάδες ανά επαγγελματικό κλάδο για τις οχτώ επόμενες κατηγορίες επαγγελμάτων για τα έτη 2001 ως 2014

5.1.2 Τα δεδομένα ανά οικονομικό κλάδο

Η δεύτερη κατηγορία δεδομένων είναι ο αριθμός εργαζομένων που απασχολείται από κάθε κατηγορία οικονομικού κλάδου, ανάμεσα στον πρωτογενή, δευτερογενή και τριτογενή κλάδο εργασίας. Όλες οι κατηγορίες επαγγελμάτων από τον παραπάνω πίνακα συμπτήσονται σε τρεις κατηγορίες που αντικατοπτρίζουν τους 3 βασικούς οικονομικούς κάδους μίας χώρας.
Στον πρωτογενή τομέα ανήκουν οι παραγωγικές δραστηριότητες που παρέχουν αγαθά σε φυσική κατάσταση απ΄ευθείας από τη φύση χωρίς να έχουν υποστεί καμία επεξεργασία. Ο πρωτογενής τομέας αποτελεί απαραίτητη προϋπόθεση για την ύπαρξη του δευτερογενούς και τριτογενούς τομέα.
Ο δευτερογενής τομέας περιλαμβάνει τις δραστηριότητες επεξεργασίας και μεταποίησης των πρώτων υλών που παρέχει ο πρωτογενής. Επεξεργασία είναι η μετατροπή των πρώτων υλών σε προϊόν με μικρές αλλαγές στη μορφή και τη σύστασή τους (παστεριωμένο γάλα). Μεταποίηση είναι η μετατροπή των πρώτων υλών σε προϊόν με ριζικές αλλαγές στη μορφή και τη σύστασή τους (γιαούρτι, τυρί).
Ο τριτογενής τομέας περιλαμβάνει την παροχή των υπηρεσιών. Είναι το τελευταίο στάδιο της παραγωγικής διαδικασίας και περιλαμβάνει ενέργειες που φέρνουν τα τελικά προϊόντα στον καταναλωτή. Τα τελευταία χρόνια η εισαγωγή των νέων τεχνολογιών στην παραγωγή έχει αυξήσει ραγδαία τον αριθμό των απασχολούμενων στον τομέα αυτό.
Τα δεδομένα και σε αυτή την περίπτωση είναι σε χιλιάδες άτομα αλλά περιέχουν και μία στήλη με το σύνολο των εργαζομένων στην Ελλάδα ώστε να μπορούν να υπολογιστούν και ποσοστά της απασχόλησης εργαζομένων σε κάθε κλάδο ανά το εργατικό δυναμικό.
Η κάθε στήλη έχει ονομαστεί ως A,B και C για τον πρωτογενή, δευτερογενή και τον τριτογενή τομέα αντίστοιχα ενώ η τελευταία που αφορά το σύνολο του εργατικού δυναμικού έχει τον τίτλο “Total”.

Ακολουθεί ο Πίνακας 5.2 με τις συχνότητες των απασχολούμενων ανά οικονομικό κλάδο.

year	A	B	C	Total
2001	666,275	968,275	2567,625	4202,125
2002	652,775	9,7435	26,37775	4264,95
2003	657,1	984,725	2711,325	4353,175
2004	543,475	983,075	2863,025	4389,525
2005	540,3	994,95	2908,3	4443,55
2006	530,825	996,95	2999,75	4527,5
2007	517,075	1024,375	3022,6	4564,05
2008	513,775	1024,425	3072,275	4610,5
2009	532,875	962,75	3060,4	4556
2010	544,2	859,825	2985,775	4389,75
2011	500,675	717,175	2836,475	4054,35
2012	480,5	611,25	2603,225	3694,975
2013	481,05	546,6	2485,5	3513,2
2014	479,875	529,675	2526,675	3536,3

Πίνακας 5.2: Η συχνότητα των απασχολούμενων σε χιλιάδες ανά οικονομικό κλάδο για τα έτη 2001 ως 2014

5.2 Εισαγωγή στην R και περιγραφή των δεδομένων

Όπως και στα προηγούμενα κεφάλαια έτσι και σε αυτό θα γίνει η εισαγωγή των δεδομένων στην R με όμοιο τρόπο και για τα δύο είδη δεδομένων. Τα δεδομένα για τους επαγγελματικούς τομείς θα ονομαστούν "data" και αυτά για τους κλάδους οικονομίας θα ονομαστούν "data1". Με K_i συμβολίζονται οι 15 διαφορετικοί επαγγελματικοί κλάδοι από το πρώτο σύνολο δεδομένων και με A,B,C οι στήλες για τους οικονομικούς κλάδους, ενώ με T συμβολίζεται το σύνολο του εργατικού δυναμικού για τις χρονιές 2001 ως 2014.

```
install.packages("gdata")
library(gdata)

data<-read.xls("../Desktop/diplwmatikh-domh/kefalaio5/oikon_drast.xls",sheet=1,skip=1)

data1<-read.xls("../Desktop/diplwmatikh-domh/kefalaio5/kef5.xls",sheet=1)

K1<-(data)[,2]
K2<-(data)[,3]
K3<-(data)[,4]
K4<-(data)[,5]
K5<-(data)[,6]
K6<-(data)[,7]
K7<-(data)[,8]
K8<-(data)[,9]
K9<-(data)[,10]
K10<-(data)[,11]
K11<-(data)[,12]
K12<-(data)[,13]
K13<-(data)[,14]
K14<-(data)[,15]
K15<-(data)[,16]
A<-(data1)[,2]
B<-(data1)[,3]
C<-(data1)[,4]

T<-(data1)[,5]
year<-(data)[,1]
```

Επίσης δημιουργούμε τους πίνακες δεδομένων που πρόκειται να χρειαστούν σε αυτό το κεφάλαιο.

```
xdata<-data.frame(K1,K2,K3,K4,K5,K6,K7,K8,K9,K10,K11,K12,K13,K14,K15)
xdata1<-data.frame(A,B,C)
xdata2<-data.frame(A,B,C,T)
```

Στη συνέχεια παρουσιάζεται ο μέσος και η τυπική απόκλιση για κάθε επαγγελματικό και

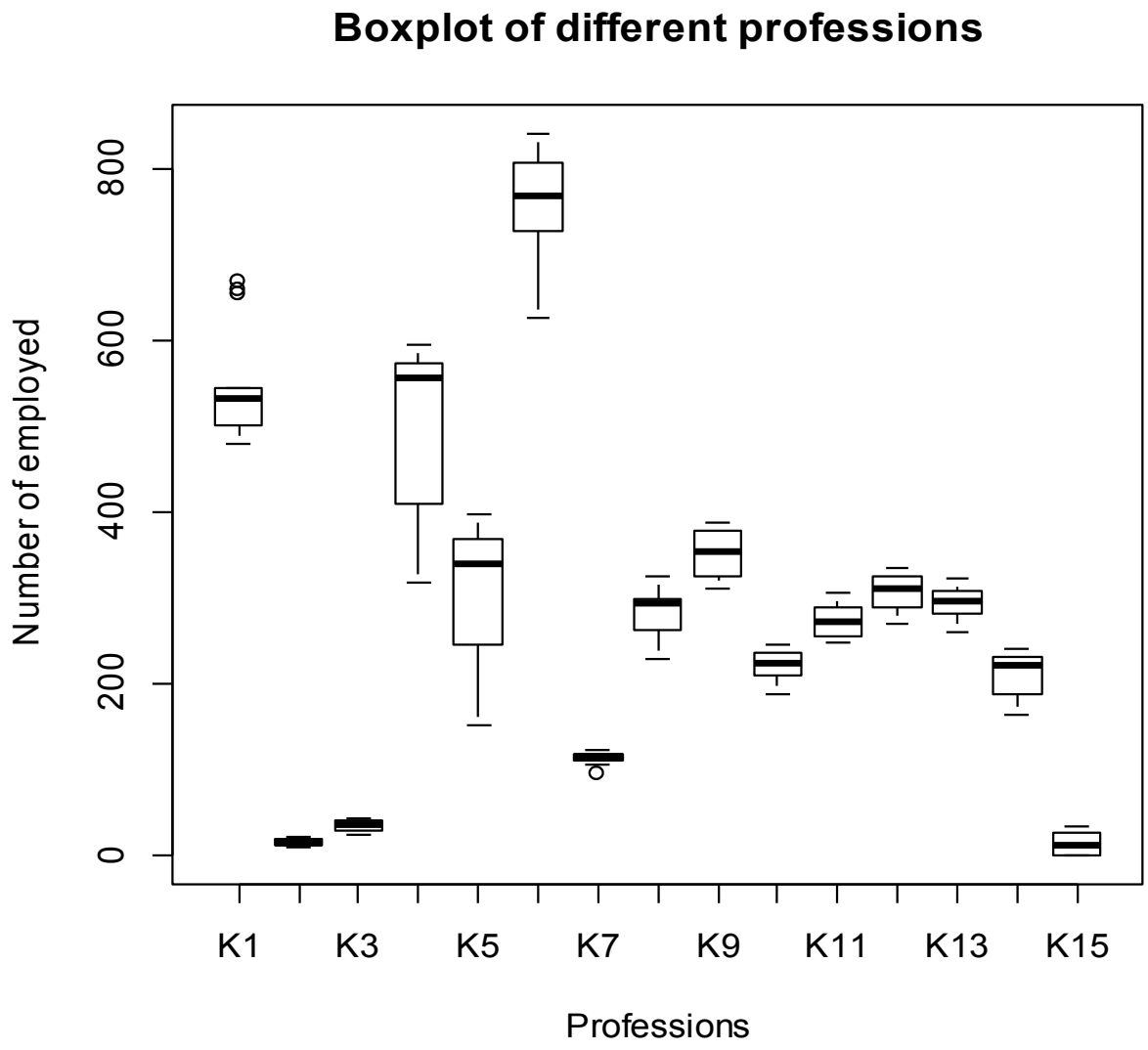
οικονομικό κλάδο.

```
mean(K1)
[1] 545.7786
sd(K1)
[1] 65.31785
mean(K2)
[1] 14.9875
sd(K2)
[1] 3.481251
mean(K3)
[1] 33.72679
sd(K3)
[1] 6.676159
mean(K4)
[1] 498.4
sd(K4)
[1] 104.1017
mean(K5)
[1] 309.4054
sd(K5)
[1] 84.54504
mean(K6)
[1] 753.9875
sd(K6)
[1] 70.20243
mean(K7)
[1] 112.5554
sd(K7)
[1] 7.04496
mean(K8)
[1] 285.0946
sd(K8)
[1] 26.90906
mean(K9)
[1] 349.5232
sd(K9)
[1] 27.85694
mean(K10)
[1] 221.0929
sd(K10)
[1] 19.93243

mean(K11)
[1] 272.0071
sd(K11)
[1] 18.17923
mean(K12)
[1] 303.925
sd(K12)
[1] 22.61069
mean(K13)
[1] 296.0839
sd(K13)
[1] 19.18317
mean(K14)
[1] 210.2196
sd(K14)
[1] 26.70045
mean(K15)
[1] 13.35714
sd(K15)
[1] 14.20252

mean(A)
[1] 545.7696
sd(A)
[1] 65.31864
mean(B)

mean(T)
[1] 4221.425
sd(T)
[1] 379.8777
```

[1] 800.9852
sd(B)
[1] 291.9159
mean(C)
[1] 2619.238
sd(C)
[1] 774.2324

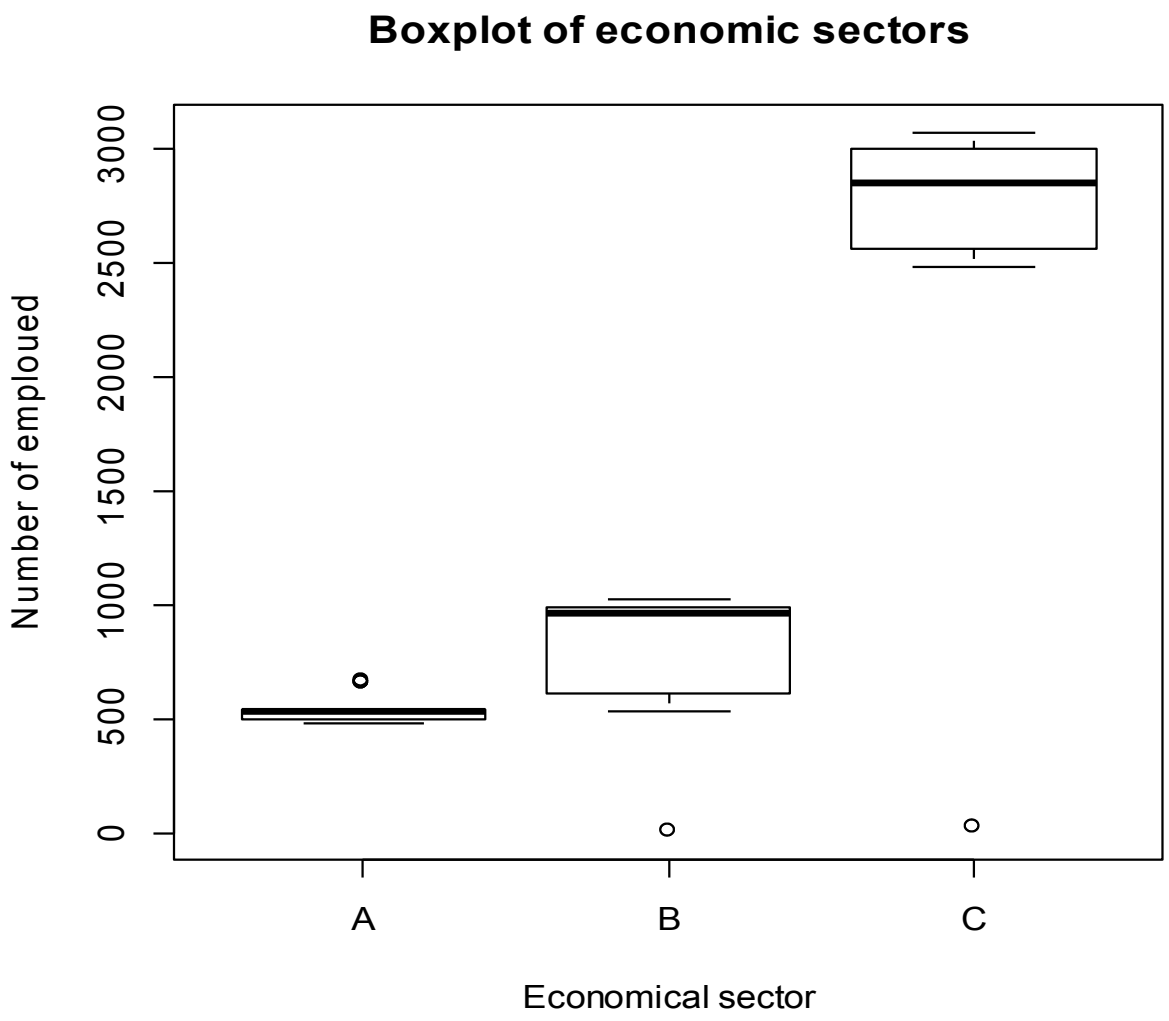
Ακολουθούν τα Διαγράμματα 5.1 και 5.2 με τον αριθμό των απασχολούμενων για κάθε κατηγορία επαγγελμάτων αλλά και για κάθε ένα από τους τρεις οικονομικούς κλάδους αντίστοιχα.

```
boxplot(xdata)
title(main="Boxplot of different professions",xlab="Professions",ylab="Number of employed")
```



Διάγραμμα 5.1: Θηκόγραμμα απασχολούμενων σε χιλιάδες για τους διάφορους επαγγελματικούς κλάδους για τα έτη 2001 ως 2014

```
boxplot(xdata1)
title(main="Boxplot of economic sectors",xlab="Economical sector", ylab="Number of emploued")
```



Διάγραμμα 5.2: Θηκόγραμμα απασχολούμενων ανά οικονομικό κλάδο σε χιλιάδες για τα έτη 2001 ως 2014

Για τους διάφορους επαγγελματικούς κλάδους το μεγαλύτερο μέσο όρο τον έχει η κατηγορία K_6 που αντιστοιχεί στο εμπόριο και την επισκεύη προϊόντων. Ενώ οι κλάδοι των Ορυχείων/Λατομείων, η παροχή ηλεκτρικού ρεύματος και νερού και οι δραστηριότητες εξυγίανσης φαίνεται να απασχολούν αρκετά μικρό αριθμό εργαζομένων σε σχέση με τους άλλους κλάδους (είναι οι κατηγορίες $K_{2,}\, K_3 \kappa\alpha\iota\, K_{15}$). Οι τυπικές αποκλίσεις σε όλους τους κλάδους φαίνονται να είναι σχετικά μικρές. Μόνο στις περιπτώσεις των κατηγοριών $K_4 \kappa\alpha\iota\, K_5$ παρατηρείται αρκετά μεγαλύτερη τιμή των τυπικών αποκλίσεων. Πρόκειται για τις μεταποιήσεις και τις κατασκευές που φαίνεται ότι με την πάροδο των ετών παρουσιάζουν αισθητή διαφορά στον αριθμό των ατόμων που απασχολούν. Τέλος μπορεί κανείς να δει από το Διάγραμμα 5.1 ότι η κατηγορία Χρηματοπιστωτικοί και Ασφαλιστικοί οργανισμοί και Γεωργία /Κτηνοτροφία/ Δασοκομία/ Αλιεία παρουσιάζουν ακραίες τιμές που σημαίνει ότι έχει υπάρξει χρονιά που παρουσιάζει μεγάλη διαφορά σε σχέση με τις άλλες στον αριθμό των απασχολούμενων σε αυτούς τους 2 κλάδους.

Όσον αφορά τους τρεις οικονομικούς κλάδους φαίνεται το μεγαλύτερο μέσο όρο απασχολούμενων να τον κατέχει ο τριτογενής κλάδος εργασίας που αφορά την παροχή υπηρεσιών. Είναι γνωστό πως είναι ο κλάδος που παρουσιάζει τη μεγαλύτερη ανάπτυξη σε σχέση με τους άλλους δύο τα τελευταία χρόνια και αυτό επαληθεύεται και για την Ελλάδα στις χρονιές 2001-2014. Φαίνεται μάλιστα να απέχει αρκετά από τους άλλους δύο στο Διάγραμμα 5.2 και η τυπική απόκλιση του έχει σχετικά μεγάλη τιμή κάτι που οφείλεται στις διαφοροποιήσεις του αριθμού απασχολούμενων λόγω της αρχικής ανάπτυξης του κλάδου αλλά και της οικονομικής ύφεσης της χώρας. Τέλος ο πρωτογενής κλάδος παρουσιάζει πολύ μικρό ενδοτεταρτημοριακό εύρος πράγμα που σημαίνει ότι οι περισσότερες παρατηρήσεις της κατηγορίας αυτής είναι συγκεντρωμένες κοντά στη διάμεσο και επίσης παρατηρείται από το Διάγραμμα 5.2 και μία ακραία παρατήρηση.

5.3 Διαγράμματα των δεδομένων ανά επαγγελματικό τομέα από το πακέτο “aplpack”

Σε αυτή την παράγραφο θα γίνει παρουσίαση και εξήγηση μιας σειράς διαγραμμάτων που η R δίνει τη δυνατότητα να κατασκευαστούν μέσω του πακέτου “aplpack” για την καλύτερη οπτική παρουσίαση και κατανόηση των δεδομένων για τους απασχολούμενους των διαφόρων επαγγελματικών τομέων. Η εγκατάσταση του πακέτου γίνεται όμοια με όλα τα προηγούμενα:

```
install.packages(“aplpack”)
library(aplpack)
```

5.3.1 Το δισδιάστατο κυτιογράφημα (bagplot)

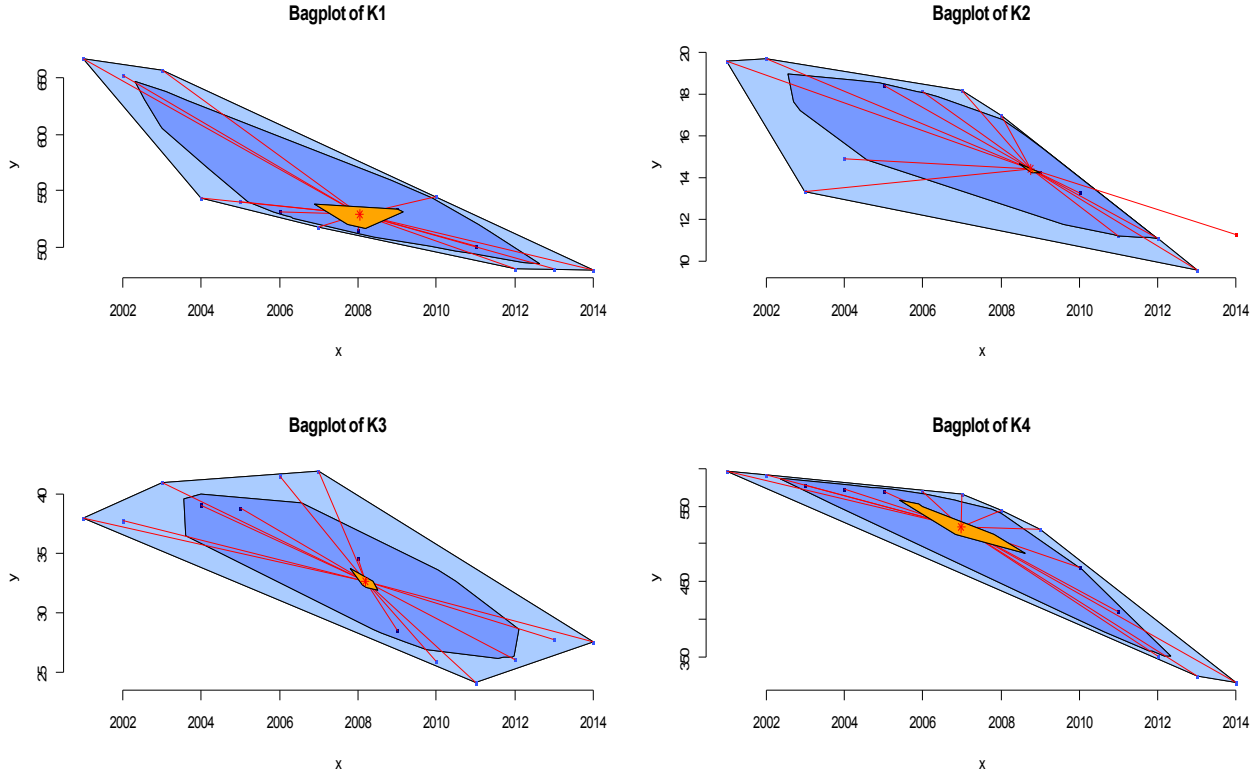
Αυτή η γραφική παράσταση είναι μία απεικόνιση δύο διαστάσεων ανάλογη του θηκογράμματος που είναι η αντίστοιχη απεικόνιση σε μία διάσταση. Ονομάζεται και “starbust plot” και προσφέρει πληροφορία για ένα σύνολο δεδομένων. Το διάγραμμα αποτελείται από τρία πολύγωνα με τα ονόματα “bag”, “fence” και “loop”. Το εσωτερικό πολύγωνο είναι αυτό με το όνομα “bag” και έχει σχηματιστεί ως εξής. Περιέχει τον μικρότερο αριθμό παρατηρήσεων που μπορεί να περιέχονται στο ημιεπίπεδο που περιέχει επίσης ένα δεδομένο σημείο. Στην προκείμενη το σημείο αυτό είναι η διάμεσος. Το εξωτερικό πολύγωνο λέγεται “fence” και βοηθάει στο σχηματισμό του γραφήματος αλλά δεν είναι μέρος του. Σχηματίζεται διογκώνοντας το εσωτερικό πολύγωνο κατά ένα συγκεκριμένο παράγοντα (συνήθως κατά 3). Όποιες παρατηρήσεις είναι εκτός του πολυγώνου αυτού θεωρούνται ακραίες τιμές. Οι παρατηρήσεις που δεν θεωρούνται ακραίες τιμές βρίσκονται μέσα στο κενό μεταξύ των δύο πολυγώνων, του μεσαίου και του εξωτερικού. Αυτός ο χώρος είναι το “loop”. Ο αστερίσκος στο κέντρο του γραφήματος συμβολίζει τη διάμεσο και οι παρατηρήσεις συνδέονται μέσω μίας λεπτής γραμμής η

καθεμία με το μικρότερο πολύγωνο που είναι ενδεικτικό της απόστασης της παρατήρησης από τη διάμεσο.
Σε αυτό το κεφάλαιο θα σχηματιστεί το διάγραμμα αυτό για κάθε έναν από τους επαγγελματικούς κλάδους σε σχέση με τη χρονιά που θα βρίσκεται στον οριζόντιο άξονα του διαγράμματος ώστε να μπορεί να συγκριθεί η θέση των σημείων σε κάθε μία περίπτωση επαγγελματικού κλάδου.
Τα διαγράμματα θα παρουσιάζονται ανά 4 στο ίδιο παράθυρο.

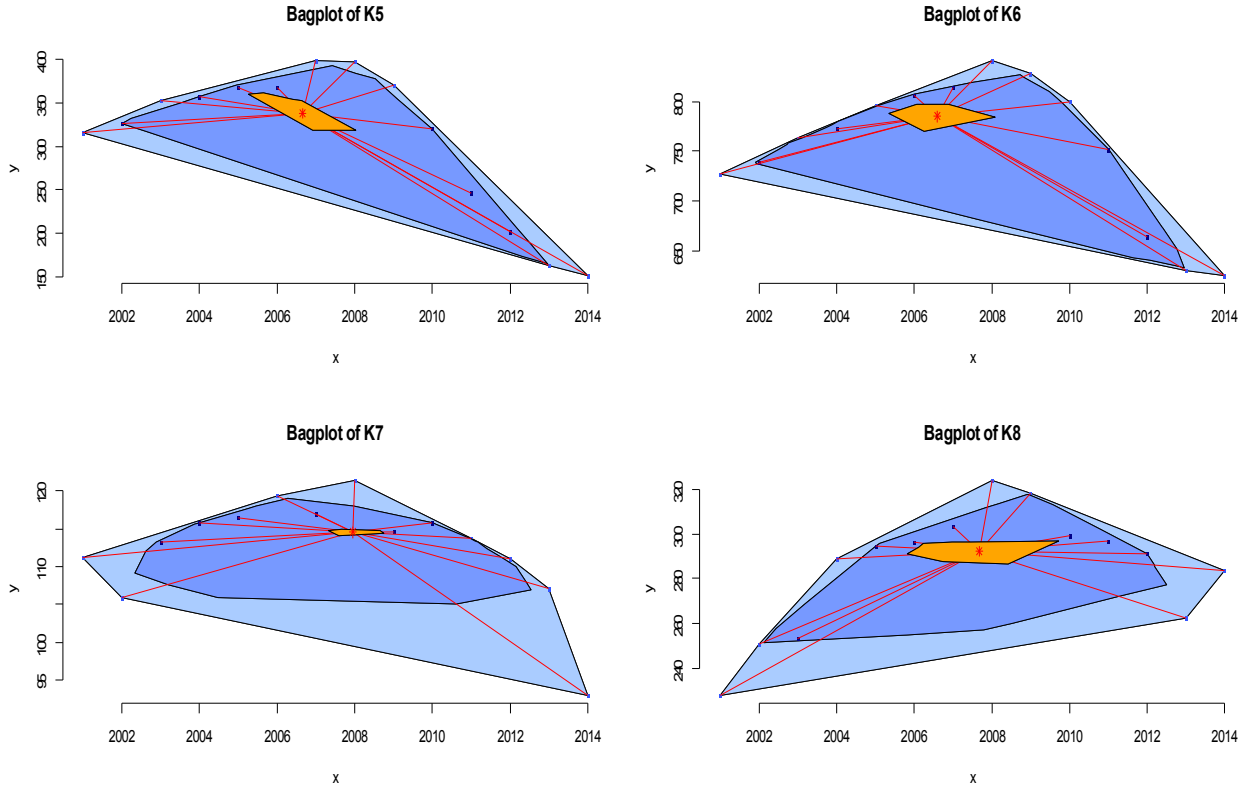
```
par(mfrow=c(2,2))
bagplot(year,K1,main="Bagplot of K1")
bagplot(year,K2,main="Bagplot of K2")
bagplot(year,K3,main="Bagplot of K3")
bagplot(year,K4,main="Bagplot of K4")
```

και όμοια ο κώδικας για κάθε τετράδα γραφημάτων θα είναι

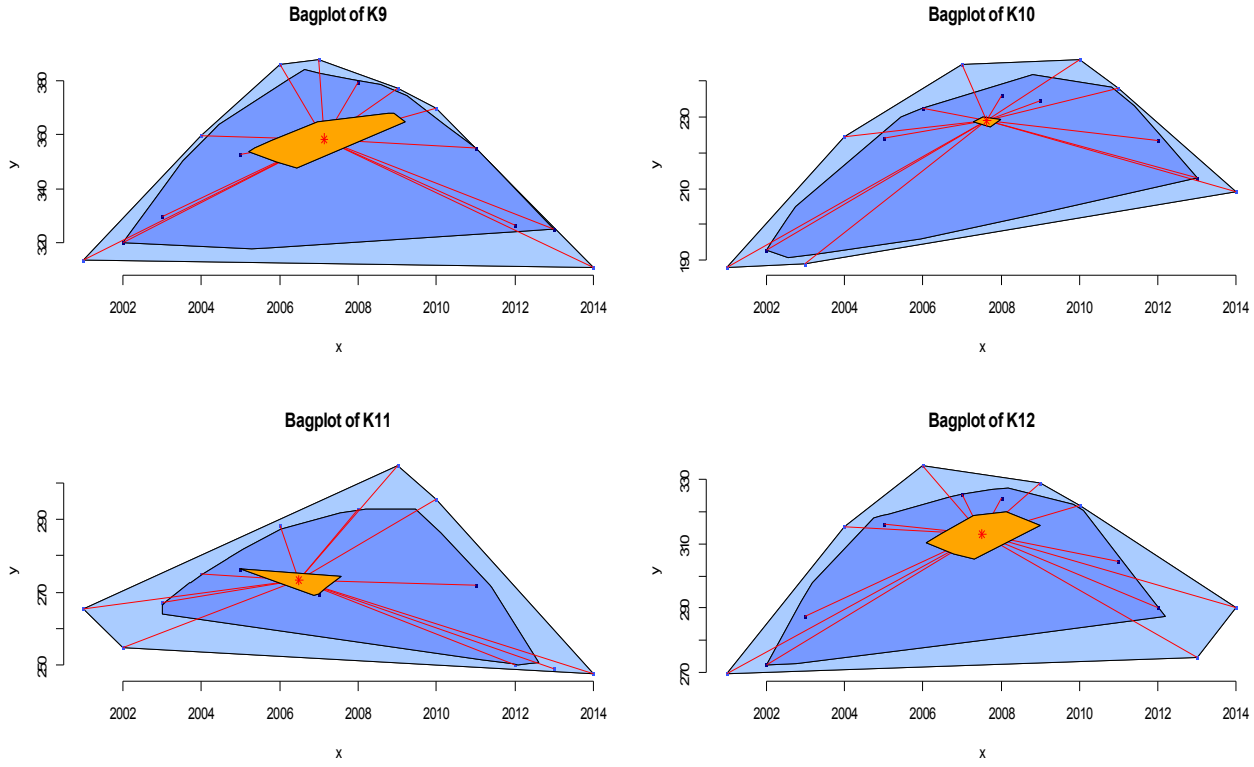
```
par(mfrow=c(2,2)
bagplot(year,Ki,main="Bagplot of Ki")
                    .
bagplot(year,K(i+4),main="Bagplot of K(i+4)")
```



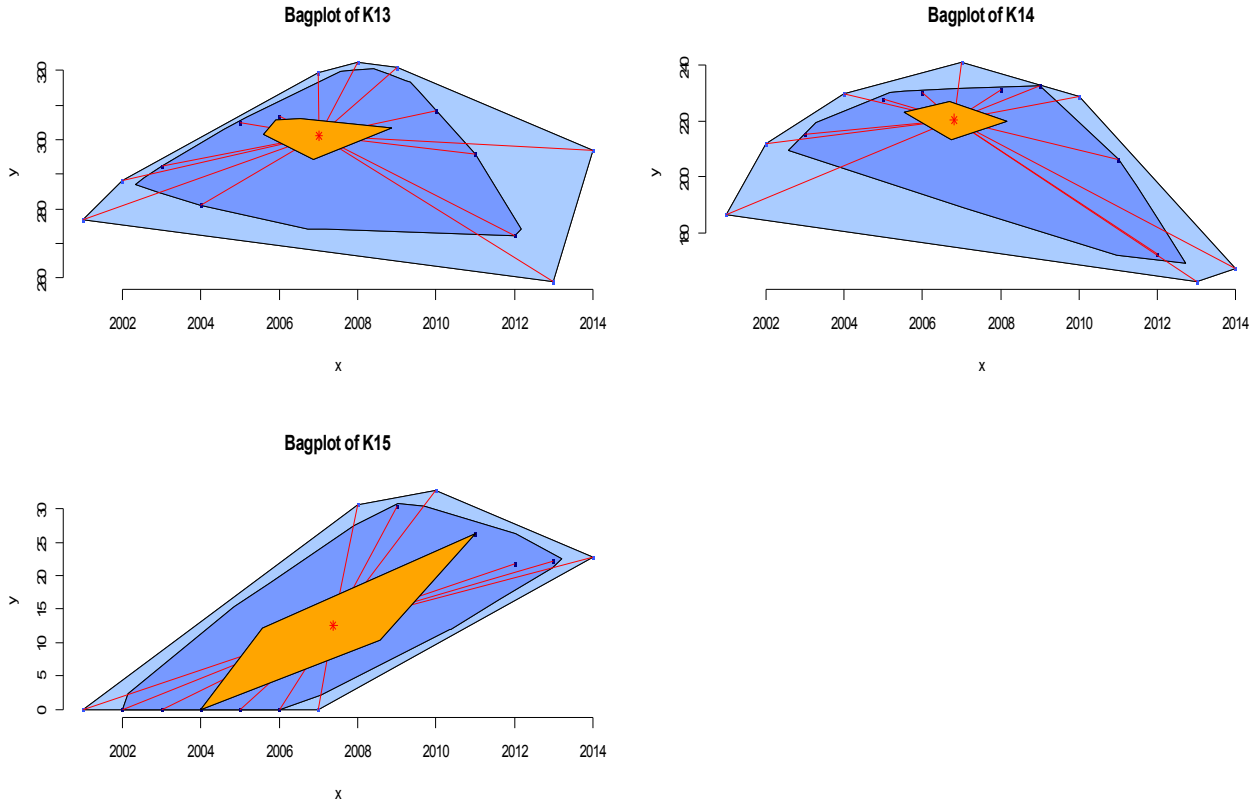
Διάγραμμα 5.3: Το δισδιάστατο κυτιογράφημα για τον αριθμό απασχολούμενων σε χιλιάδες για τις κατηγορίες K1,K2,K3 και K4 επαγγελμάτων για τα έτη 2001 ως 2014



Διάγραμμα 5.4: Δισδιάστατο κυτιογράφημα για τον αριθμό απασχολούμενων σε χιλιάδες στις κατηγορίες επαγγελμάτων K5,K6,K7 και K8 για τα έτη 2001 ως και 2014



Διάγραμμα 5.5: Δισδιάστατο κυτιογράφημα για τον αριθμό απασχολούμενων σε χιλιάδες στις κατηγορίες επαγγελμάτων K9, K10, K11 και K12 για τα έτη 2001 ως 2014



Διάγραμμα 5.6: Δισδιάστατο κυτιογράφημα για τον αριθμό απασχολούμενων σε χιλιάδες στις κατηγορίες επαγγελμάτων K13, K14 και K15 για τα λετη 2001 ως 2014

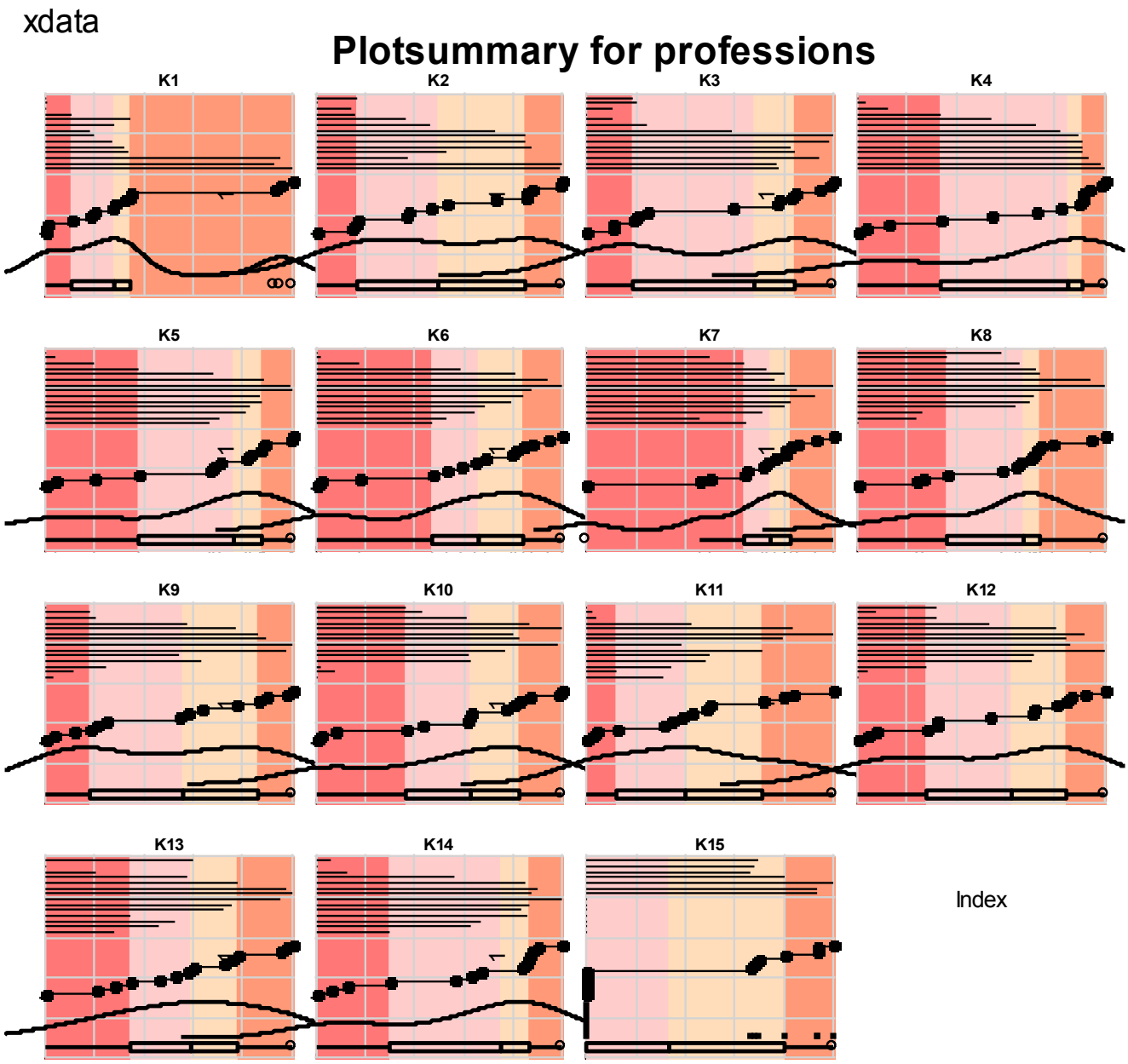
Από τα Διαγράμματα 5.3, 5.4, 5.5 και 5.6 μπορεί να παρατηρήσει κανείς ότι υπάρχουν αρκετά μεγάλες διαφορές της μορφής των δεδομένων ανά χρονιά για κάθε διαφορετικό επαγγελματικό κλάδο. Ακραίες τιμές παρατηρούνται μόνο στην κατηγορία K_2 που είναι οι απασχολούμενοι από Λατομεία και Ορυχεία ενώ οι τιμές από την πρώτη κατηγορία που είχαν θεωρηθεί ακραίες στο θηκόγραμμα έχουν συμπεριληφθεί στην περιοχή του διαγράμματος. Αυτό οφείλεται στο διαφορετικό τρόπο σχεδιασμού των δύο διαγραμμάτων και στο μέγεθος της απόστασης που έχει θεωρηθεί σημαντική σε κάθε περίπτωση. Βλέποντας τον πίνακα δεδομένων για την κατηγορία K_2 για παράδειγμα βλέπει κανείς ότι το 2014, που στο γράφημα αυτής της παραγράφου φαίνεται ακραία τιμή απασχολούνται από αυτή την κατηγορία περίπου 11.000 εργαζόμενοι που είναι περίπου οι μισοί σε σχέση με τη χρονιά 2001 οπότε το να θεωρηθεί αυτή η τιμή ακραία δε φαίνεται παράλογο. Το κεντρικό πολύγωνο στο δισδιάστατο κυτιογράφημα περιέχει το πολύ το 50% των παρατηρήσεων. Επίσης όπως ήταν αναμενόμενο παρατηρούνται διαφορετικές κλίμακες σε κάθε δισδιάστατο κυτιογράφημα. Για παράδειγμα για την τελευταία κατηγορία που οι τιμές των απασχολούμενων είναι της τάξης του 20.000 δεν μπορεί να είναι ίδια η κλίμακα με την πρώτη κατηγορία που οι απασχολούμενοι είναι της τάξης του 500.000. Εκτός αυτού ο αστερίσκος, δηλαδή η διάμεσος σε κάθε περίπτωση είναι πιο κοντά σε άλλη χρονιά. Τα σχήματα στα οποία το κίτρινο πολύγωνο είναι μεγαλύτερο περιέχουν πολύ περισσότερα σημεία κοντά στη διάμεσο, για αυτό και θεωρείται από τις προεπιλογές της R ότι βρίσκονται στο ίδιο ημιεπίπεδο με τη διάμεσο. Από την άλλη αυτά με το μικρότερο

εσωτερικό πολύγωνο όπως το K_2 θεωρείται ότι δεν περιέχουν τόσο πολλά σημεία κοντά στη διάμεσο της κατηγορίας. Τέλος το σχήμα των πολυγώνων φαίνεται να διαφέρει ανάλογα και με το πόσο διασκορπισμένα είναι τα σημεία σε κάθε κατηγορία επαγγελματικού κλάδου.
Το πακέτο αυτό επίσης παρέχει την εντολή "bagplot.pairs" για τη δημιουργία των bagplots όλων των στηλών μεταξύ τους από το σύνολο των δεδομένων, όμως αυτό θα είχε σαν αποτέλεσμα την παρουσίαση 16×16 δισδιάστατων κυτιογραφημάτων στο ίδιο παράθυρο, κάτι που δε μας επιτρέπει να κάνουμε παρατηρήσεις για τα δεδομένα.
Το διάγραμμα αυτού του είδους δεν παρέχει ξεκάθαρη πληροφορία για τις διάφορες χρονιές αλλά δείχνει την μεταβλητότητα των εργαζομένων των διαφόρων μορφωτικών επιπέδων στην πάροδο των χρόνων και τη διαφοροποίηση μεταξύ τους. Η διαφοροποίηση ανά χρονιά θα παρουσιαστεί παρακάτω αναλυτικότερα.

5.3.2 Το διάγραμμα της εντολής "plotsummary"

Το πακέτο "aplpack" παρέχει τη δυνατότητα της κατασκευής ενός διαγράμματος που δείχνει σημαντικά χαρακτηριστικά των μεταβλητών του συνόλου δεδομένων. Δημιουργεί τόσα διαγράμματα όσα η κάθε κατηγορία του συνόλου δεδομένων, για την συγκεκριμένη περίπτωση λοιπόν θα δημιουργηθούν 15 διαγράματα. Το κάθε ένα περιέχει κατά σειρά το ιστόγραμμα, την εμπειρική κατανομή, την προσέγγιση της συνάρτησης πυκνότητας πιθανότητας του κάθε επαγγελματικού κλάδου με τη μέθοδο των πυρήνων (dencity trace) και το θηκόγραμμα (boxplot). Τα 4 αυτά παρουσιάζονται σε ένα κοινό διάγραμμα με τη βοήθεια της εντολής "plotsummary" που βρίσκεται στο πακέτο αυτό.
Σαν όρισμα η εντολή δέχεται έναν πίνακα.
Για τη δημιουργία τους λοιπόν στο σύνολο δεδομένων θα πληκτρολογηθούν οι ακόλουθες εντολές.

```
plotsummary(xdata)
title(main="Plotsummary for professions")
```



Διάγραμμα 5.7: Το διάγραμμα απο την εντολή "plotsummary" για τον αριθμό απασχολουμένων σε χιλιάδες όλων των διαφορετικών επαγγελμάτων

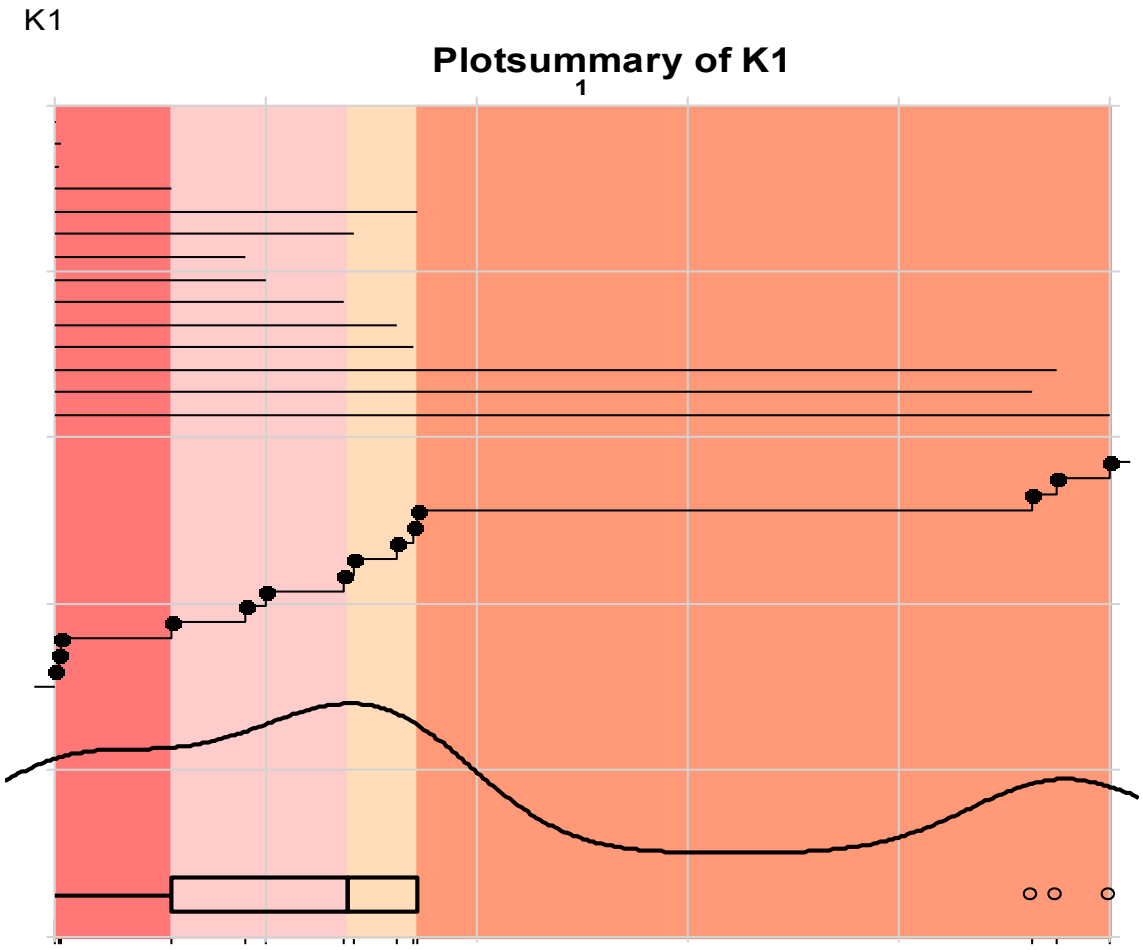
Η αλλαγή των χρωμάτων στο κάθε διάγραμμα είναι προαιρετική και δείχνει τις διαφορετικές περιοχές που δημιουργούνται από τα τεταρτημόρια για αυτό και μπορεί να δει κανείς ότι το χρώμα σε κάθε περίπτωση αλλάζει στα σημεία που δείχνει το θηκόγραμμα ως κάτω και άνω τεταρτημόριο. Για να μην υπάρχει αυτή η διαφοροποίηση στα χρώματα πρέπει κανείς να δώσει σαν όρισμα στην εντολή "plotsummary" το "mycols="no"" γιατί η R έχει σαν προεπιλογή το "mycols="RB"".

Τα θηκογράμματα είναι όμοια με αυτά που παρουσιάστηκαν σε προηγούμενη παράγραφο αυτού του κεφαλαίου σε οριζόντια όμως μορφή. Μπορεί κανείς να παρατηρήσει την αλλαγή των χρωματιστών επιπέδων στα ανάλογα σημεία. Είναι εμφανές πως οι κορυφές έχουν σχηματιστεί εκεί που υπάρχει πληθώρα σημείων κοντά. Τέλος το ιστόγραμμα έχει την αναμενόμενη μορφή σε κάθε περίπτωση και έχει σχηματιστεί με την δημιουργία γραμμών στα αντίστοιχα σημεία και έχει χρησιμοποιηθεί ελάχιστο πλάτος για την κατασκευή του.

Όπως είναι αναμενόμενο παρατηρείται διαφορά ανάμεσα σε κάθε επαγγελματικό κλάδο για το σύνολο των απασχολουμένων κάτι αναμενόμενο αφού τα δεδομένα σε κάθε περίπτωση παρουσιάζουν αρκετά διαφορετική συχνότητα και διαφοροποιούνται σε σχέση με τον αριθμό ανάπτυξης και του κάθε επαγγέλματος. Για παράδειγμα η τελευταία κατηγορία τις πρώτες χρονιές δεν απασχολεί αριθμό εργαζομένων καθώς δεν είχε αναπτυχθεί στην Ελλάδα και μόνο μετά το 2008 παρουσιάζει ανάπτυξη οπότε και απασχολεί αριθμό εργαζομένων ο οποίος όμως δεν διαφοροποιείται ιδιαίτερα αφού είναι όλες τις χρονιές γύρω στους 25.000. Αντίθετα για τη κατηγορία K_8 που είναι η διαχείρηση ακίνητης περιουσίας και οι εκμισθώσεις δεν παρατηρείται κάποιο τέτοιο φαινόμενο. Είναι ένας τιμέας που ιδιαίτερα τις τελευταίες δεκαετίες προτιμάται σχετικά για αυτό και έχει όλες τις χρονιές σχετικά μεγάλο αριθμό εργαζομένων. Φαίνεται μάλιστα από το γράφημα ότι η κορυφή του βρίσκεται προς τα δεξιά και δεν παρουσιάζει άλλη κορυφή. Αυτό σημαίνει ότι έχουμε συσσώρευση παρατηρήσεων τις τελευταίες χρονιές και δείχνει ότι με την πάροδο των ετών και παρά την οικονομική ύφεση ο αριθμός εργαζομένων έχει αυξηθεί σχετικά. Αυτό συμβαίνει και σε άλλες κατηγορίες του τριτογενούς κλάδου ανάπτυξης που πλέον προτιμάται αλλά και είναι πιο απαραίτητος σε σχέση με το παρελθόν. Αντίθετα στον κλάδο της Γεωργίας/Αλιείας/Δασοκομίας/Κτηνοτροφίας η μορφή της γραφικής είναι αρκετά διαφορετική. Στις αρχικές χρονιές έχει δημιουργηθεί μία κορυφή που δείχνει πληθώρα απασχολουμένων τις πρώτες χρονιές του δείγματος. Στη συνέχεια όμως έχουμε πτώση καθώς η παραγωγή της χώρας μειώνεται αλλά και το είδος αυτό του επαγγέλματος δεν προτιμάται. Προς το τέλος δημιουργείται και μία δεύτερη κορυφή, αρκετά μικρότερη, που οφείλεται στην στροφή κάποιων εργαζομένων ίσως προς την αγροτική ζωή μετά από περιόδους ανεργίας και στη μείωση του ρυθμού αύξησης της αστυφιλίας μετά την κρίση.

Το Διάγραμμα 5.8 που παρουσιάζεται στην συνέχεια αφορά την πρώτη κατηγορία και βοηθά στην καλύτερη κατανόηση αυτών των διαγραμμάτων αλλά και δίνει την δυνατότητα να φανούν οι λεπτομέρειες που είχαν χαθεί λόγω του περιορισμού του χώρου στο Διάγραμμα 5.7.

```
plotsummary(K1)
title(main="Plotsummary of K1")
```



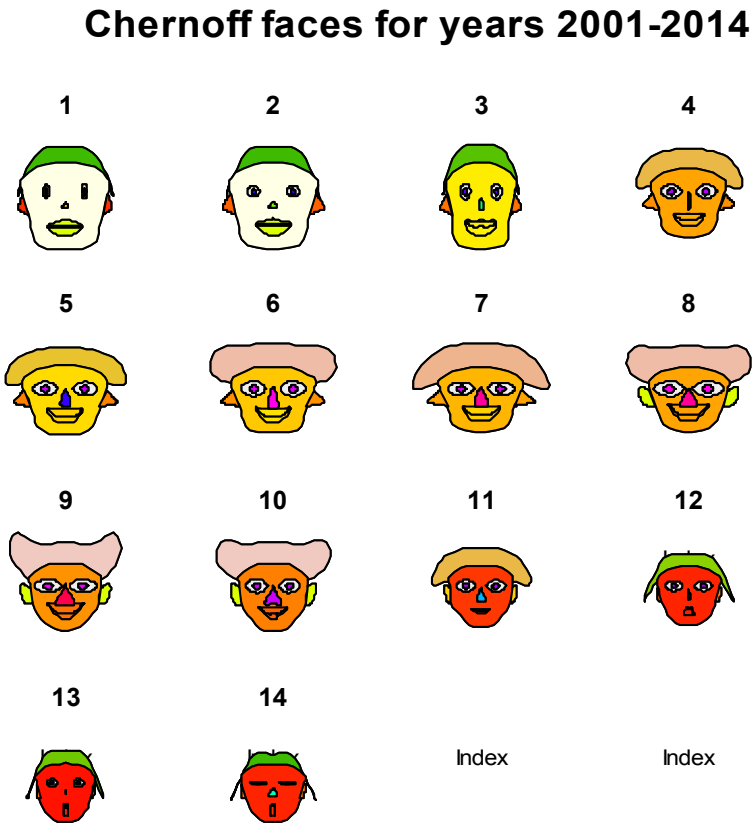
Διάγραμμα 5.8: Το διάγραμμα από την εντολή "plotsummary" για τους απασχολούμενους σε χιλιάδες της Κ1 κατηγορίας επαγγελμάτων για τα έτη 2001 ως 2014

5.3.3 Τα πρόσωπα του Chernoff

Ο σκοπός του διαγράμματος των προσώπων του Chernof είναι να παρουσιαστούν μία σειρά από διαφορετικά χαρακτηριστικά σε ένα γράφημα μέσω χαρακτηριστικών ενός προσώπου όπως το μέγεθος της μύτης και άλλα. Η αρχική ιδέα για την κατασκευή αυτών των γραφημάτων είναι ότι ο μέσος άνθρωπος μπορεί να παρατηρήσει εύκολα διαφορές στα πρόσωπα ανθρώπων στην καθημερινότητα οπότε θα μπορεί να παρατηρήσει και τις διαφορές σε πρόσωπα που αναπαριστούν μία σειρά από δεδομένα. Στην R τα γραφήματα αυτά μπορούν να κατασκευαστούν και αυτά μέσω του πακέτου "aplpack". Η εντολή για την κατασκευή των προσώπων Chernoff είναι:

```
faces(xdata)
title(main="Chernoff faces for years 2001-2014")
```

Οπότε προκύπτει το Διάγραμμα 5.9.



Διάγραμμα 5.9: Τα πρόσωπα του Chernoff για τον πίνακα δεδομένων των απασχολούμενων σε χιλιάδες σε διάφορες κατηγορίες επαγγελμάτων για τα έτη 2001 ως 2014

Εκτός από το Διάγραμμα 5.9 η R επιστρέφει και τις αντιστοιχίες κάθε χαρακτηριστικού με κάποια μεταβλητή.

effect of variables:

modified item	Var
"height of face "	"K1"
"width of face "	"K2"
"structure of face"	"K3"
"height of mouth "	"K4"
"width of mouth"	"K5"
"smiling"	"K6"
"height of eyes"	"K7"
"width of eyes "	"K8"
"height of hair"	"K9"
"width of hair "	"K10"
"style of hair"	"K11"
"height of nose "	"K12"

"width of nose " "K13"
"width of ear " "K14"
"height of ear " "K15"

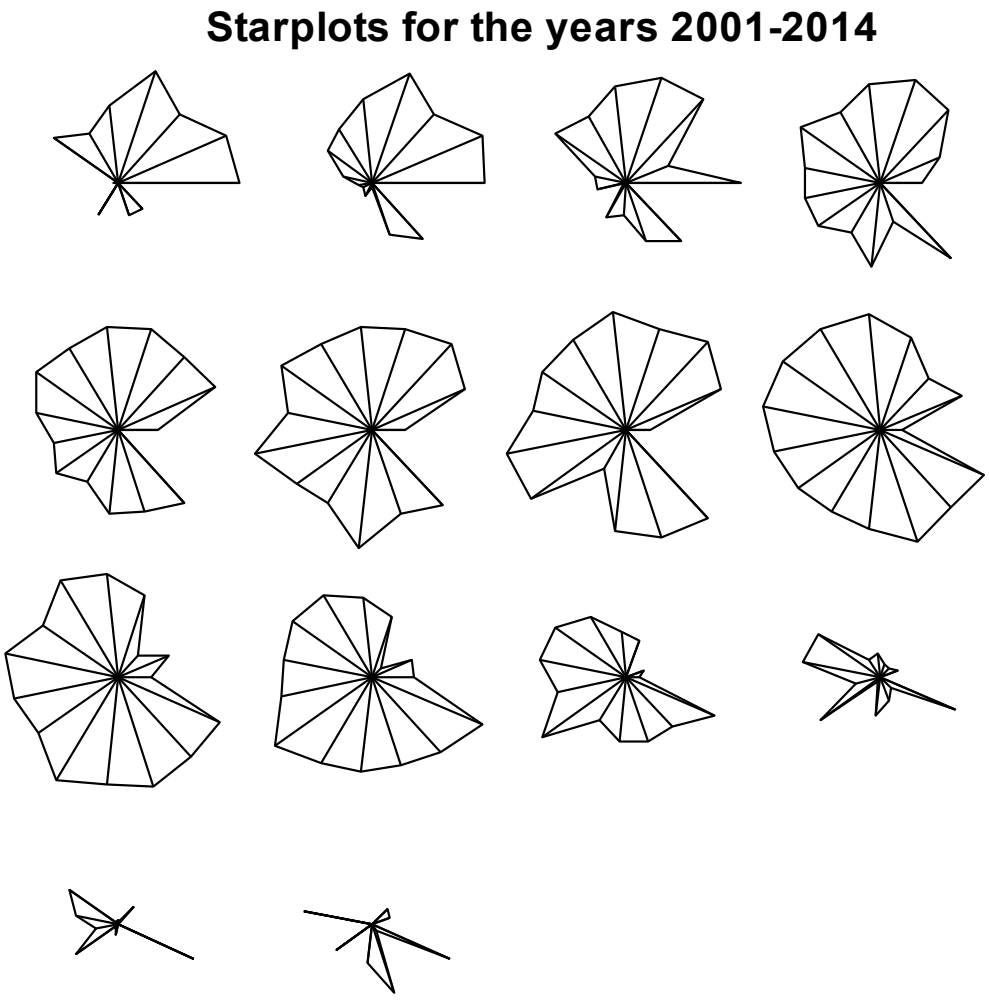
Έτσι βλέποντας κανείς το Διάγραμμα 5.9 μπορεί να δει για τις χρονιές από το 2001 ως το 2014 τον αριθμό των εργαζομένων που απασχολούνται από την κατηγορία K_1 με τη βοήθεια του ύψους του προσώπου. Είναι εμφανές πως τις 2 πρώτες χρονιές ο αριθμός των εργαζομένων είναι πολύ μεγαλύτερος από τις 2 τελευταίες. Το μήκος του προσώπου περιγράφει τον αριθμό των εργαζομένων από την κατηγορία επαγγέλματος K_2 ενώ αντίστοιχα η δομή του προσώπου αφορά τη μεταβλητή K_3 .Το ύψος και το πλάτος του στόματος στα πρόσωπα αφορά τις τιμές για τις κατηγορίες K_4 και K_5 αντίστοιχα ενώ το μέγεθος του χαμόγελου τις τιμές για τη μεταβλητή K_6 .Το ύψος και το πλάτος των ματιών καθώς και το ύψος και το πλάτος και το είδος των μαλλιών αφορούν τις κατηγορίες $K_{7,}\, K_8 \,και\, K_{9,}\, K_{10} \,και\, K_{11}$ αντίστοιχα. Το ύψος και το πλάτος της μύτης μας περιγράφουν τις τιμές για τον αριθμό απασχολούμενων των κατηγοριών K_{12} και K_{13}. Τέλος το πλάτος των αυτιών και το ύψος τους παρουσιάζει τον αριθμό εργαζομένων για τις κατηγορίες $K_{14} \,και\, K_{15}$.
Όπως μπορεί κανείς να δει τα περισσότερα χαρακτηριστικά φαίνεται να μικραίνουν όσο περνούν τα έτη εκτός ίσως από το ύψος των αυτιών στα πρόσωπα που αφορούν την τελευταία κατηγορία επαγγέλματος (Επεξεργασία λυμάτων,διαχείρηση αποβλήτων και δραστηριότητες εξυγίανσης) που στις αρχικές χρονιές απασχολούσε μηδενικό αριθμό εργαζομένων.

5.4 Το διάγραμμα-αστέρι (starplot) για τους διάφορους επαγγελματικούς τομείς

Το διάγραμμα-αστέρι είναι ένας έξυπνος τρόπος για να αναπαραστήσει κανείς παρατηρήσεις με πολλά χαρακτηριστικά. Για κάθε μια παρατήρηση ο ερευνητής κατασκευάζει ένα "αστέρι" με τόσες ακτίνες όσες είναι και οι διάφορες κατηγορίες που μελετούνται, δηλαδή το μέγεθος κάθε ακτίνας αναπαριστά την τιμή της παρατήρησης για κάποια μεταβλητή. Συνεπώς κρίνοντας από το οπτικό αποτέλεσμα μπορεί κανείς να παρατηρήσει διαφορές ανάμεσα στις παρατηρήσεις καθώς αυτές θα έχουν διαφορετικά σχήματα. Τέτοια διαγράμματα μπορούν να χρησιμοποιηθούν για μεγάλο αριθμό μεταβλητών. Από τη μορφή που παίρνει το αστέρι για κάθε παρατήρηση μπορεί κανείς να δει τις διαφορές που υπάρχουν ανάμεσα στις περιοχές. Θα πρέπει να τονιστεί πως το μήκος της ακτίνας υπολογίζεται με βάση τις τυποποιημένες τιμές. Οι τυποποιημένες τιμές είναι ένα μέτρο σχετικής θέσης των τιμών σε ένα σύνολο μετρήσεων και υπολογίζονται αφαιρώντας από κάθε παρατήρηση το μέσο όρο και στη συνέχεια διαιρώντας με την τυπική απόκλιση. Οι τυποποιημένες τιμές συμβολίζονται με το γράμμα z.

Αυτός ο τύπος διαγράμματος δε χρειάζεται κάποιο πακέτο για να κατασκευαστεί αλλά βρίσκεται στις γραφικές παραστάσεις των οποίων την κατασκευή προσφέρει η R.

```
stars(xdata)
title(main="Starplots for the years 2001-2014")
```



Διάγραμμα 5.10: Το διάγραμμα-αστέρι για τον πίνακα συχνοτήτων για τον αριθμό απασχολούμενων σε χιλιάδες σε διάφορες κατηγορίες επαγγελμάτων για τα έτη 2001 ως 2014

Το Διάγραμμα 5.10 περιέχει 14 "αστέρια" κάθε ένα από τα οποία αντιστοιχεί σε ένα έτος από το 2001 ως το 2014. Κάθε γραμμή που σχηματίζει το αστέρι αντιστοιχεί σε μία κατηγορία επαγγέλματος. Παρατηρούμε ότι από χρονιά σε χρονιά υπάρχουν αρκετά έντονες διαφοροποιήσεις. Οι τελευταίες τρεις χρονιές φαίνεται να έχουν πολύ μικρότερες τιμές σε όλους τους επαγγελματικούς κλάδους σε σχέση με τις προηγούμενες χρονιές, τόσο πολύ που η μορφή τους δε θυμίζει καθόλου τα προηγούμενα έτη. Πολλές από τις "ακτίνες" μάλιστα για τα διάφορα επαγγέλματα έχουν σχεδόν εξαληφθεί. Από το 2001 ως το 2009 φαίνεται να υπάρχει μία ανάπτυξη στον αριθμό των απασχολούμενων στις περισσότερες κατηγορίες που όμως από το επόμενο έτος φαίνεται να σταματάει ενώ ένα χρόνο μετά ο αριθμός των απασχολούμενων φαίνεται να πέφτει κατακόρυφα και να διατηρείται σε πολύ

χαμηλότερες τιμές σε σχέση με πριν για τα τελευταία έτη.

5.5 Οι καμπύλες του Andrews για τους επαγγελματικούς τομείς

Οι καμπύλες του Andrews (Andrews' curves) εισήχθηκαν από τον Andrews με σκοπό να απεικονιστούν πολυμεταβλητά δεδομένα. Ο σχηματισμός της καμπύλης γίνεται ως εξής.
Για κάθε παρατήρηση σχηματίζουμε την καμπύλη της συνάρτησης

$$f_X(t)=\frac{X_1}{\sqrt{2}}+X_2\cdot sint+X_3\cdot cost+X_4\cdot\sin(2t)+X_5\cdot\cos(2t)+.... \quad \text{με} \quad t\in(-\pi,\pi)$$

για διαφορετικές τιμές του t και στη συνέχεια φτιάχνουμε το γράφημα $(t, f(t))$ για το διάστημα $(-\pi,\pi)$.
Όπως μπορεί να δει κανείς η καμπύλη αυτή αποτελείται από ημίτονα και συνημίτονα έχει δηλαδή μια περιοδικότητα. Αυτή η περιοδικότητα όμως εξαρτάται από τις τιμές των μεταβλητών και επομένως για διαφορετικές παρατηρήσεις περιμένουμε και διαφορετικές καμπύλες. Επομένως με τη χρήση αυτών των καμπυλών ελπίζουμε πως απεικονίζοντας πολυμεταβλητές παρατηρήσεις θα μπορέσουμε να δούμε πόσο και ποιες από αυτές διαφέρουν καθώς παρατηρήσεις με διαφορετικές τιμές σε κάθε μεταβλητή θα έχουν διαφορετικά χαρακτηριστικά (π.χ. διαφορετική περιοδικότητα, μεγαλύτερη συχνότητα, κλπ). Η επιλογή της σειράς με την οποία οι μεταβλητές θα χρησιμοποιηθούν είναι σημαντική δεδομένου πως η σειρά καθορίζει τη σημαντικότητα κάθε μεταβλητής στη δημιουργία της καμπύλης, οι μεταβλητές τοποθετούνται με φθίνουσα σειρά διακύμανσης, δηλαδή η X1 είναι η μεταβλητή με τη μεγαλύτερη διακύμανση, η X2 η μεταβλητή με τη δεύτερη μεγαλύτερη διακύμανση και ούτω καθεξής.
Οι παρατηρήσεις που είναι σχετικά ίδιες μεταξύ τους θα έχουν σχετικά ίδιες καμπύλες και επομένως μπορούμε να διακρίνουμε ομάδες παρατηρήσεων. Αντίθετα παρατηρήσεις που διαφέρουν θα έχουν πολύ διαφορετικές καμπύλες. Επομένως η μέθοδος είναι ικανή να βρει ακραίες παρατηρήσεις (outliers) από ένα σύνολο παρατηρήσεων. Αν δηλαδή υπάρχει κάποια(/ες) παρατήρηση (/εις) για την οποία η καμπύλη είναι ολότελα διαφορετική αυτό σημαίνει πως η παρατήρηση αυτή είναι πολύ διαφορετική από τις άλλες και άρα είναι πιθανή έκτροπη τιμή.
Είναι πολύ σημαντικό πως οι καμπύλες Andrews έχουν την εξής ιδιότητα: η απόσταση ανάμεσα σε δύο καμπύλες είναι ανάλογη της Ευκλείδειας απόστασης ανάμεσα στις παρατηρήσεις. Επομένως οι καμπύλες Andrews αναπαριστούν τις διαφορές που υπάρχουν ανάμεσα στις παρατηρήσεις.
Εν κατακλείδι, οι καμπύλες αυτές μας προσφέρουν έναν εύκολο τρόπο να πάρουμε μια εικόνα από τα δεδομένα και κυρίως να δούμε κατά πόσο οι παρατηρήσεις μοιάζουν μεταξύ τους και αν υπάρχουν κάποιες που διαφέρουν πολύ από τις άλλες στα δεδομένα μας.

Στην R οι καμπύλες του Andrews σχηματίζονται με τη βοήθεια του πακέτου "andrews" το οποίο και εγκαθιστούμε για να χρησιμοποιήσουμε την εντολή "andrews" που θα δημιουργήσει το γράφημα.

```
install.packages("andrews")
library(andrews)
```

Η εντολή “andrews” εκτός από τον πίνακα δεδομένων και το όρισμα για τίτλο και υπότιτλο που είναι όμοιο με τις περισσότερες εντολές στην R (“main=”, “sub=”) και τα ορίσματα “type” και “clr”. Το πρώτο αφορά το είδος των καμπυλών που θα δημιουργηθεί και παίρνει 4 πιθανές τιμές

- $f_X(t)=\frac{X_1}{\sqrt{2}}+X_2\cdot sint+X_3\cdot cost+X_4\cdot\sin(2t)+X_5\cdot\cos(2t)+....$
- $f_X(t)=X_1\cdot sint+X_2\cdot cost+X_3\cdot\sin(2t)+X_4\cdot\cos(2t)+....$
- $f_X(t)=X_1\cdot cost+X_2\cdot\cos(2t)^{\frac{1}{2}}+X_3\cdot\cos(3t)^{\frac{1}{2}}+....$
- $f_X(t)=\frac{1}{\sqrt{2}}\cdot X_1+X_2\cdot(sint+cost)+X_3\cdot(sint-cost)+X_4\cdot(\sin(2t)+\cos(2t))+....$

Για το διάγραμμα αυτής της παραγράφου θα χρησιμοποιηθεί ο πρώτος τύπος από τους τέσσερις καθώς οι άλλοι θεωρούνται προσεγγίσεις που χρησιμοποιούμε για μεγαλύτερες τιμές του t.
Η R λοιπόν θα χρησιμοποιήσει τον παραπάνω τύπο τοποθετώντας στη θέση των X_i τους διάφορους επαγγελματικούς κλάδους K_i με σειρά φθίνουσας διακύμανσης.
Για να μπορέσουμε να δούμε τη σειρά κατάταξης των επαγγελματικών κλάδων με κριτήριο τη διασπορά τους μπορούμε να εκτελέσουμε τον ακόλουθο κώδικα στην R.

```
i<-1
res<-rep(0,15)
for(i in 1:15)
 {
x[i]<-var(xdata[,i])
res<-c(res,x[i])
i<-i+1
 }
order(res)
```

Ο κώδικας αυτός υπολογίζει ένα διάνυσμα το οποίο περιλαμβάνει την διακύμανση κάθε κατηγορίας επαγγελματικού κλάδου από την πρώτη στήλη ως την τελευταία του συνόλου δεδομένων και στη συνέχεια επιστρέφει ένα διάνυσμα που δείχνει τη σειρά κατάταξης κάθε θέσης. Το αποτέλεσμα είναι

```
 [1]  2  3  7 15 10 12  9 11 14  8  9  1  6  5  4
```

και μας δείχνει ότι η διακύμανση της πρώτης κατηγορίας επαγγέλματος είναι η δεύτερη μικρότερη ενώ της τρίτης κατηγορίας είναι η τρίτη μικρότερη κ.ο.κ. Η μεγαλύτερη τιμή είναι αυτή της οποίας η θέση κατατάσσεται ως 15 δηλαδή τελευταία που είναι η τέταρτη κατά σειρά. Μπορούμε να εντοπίσσουμε την κατηγορία με τη μεγαλύτερη διακύμανση και με την βοήθεια της εντολής

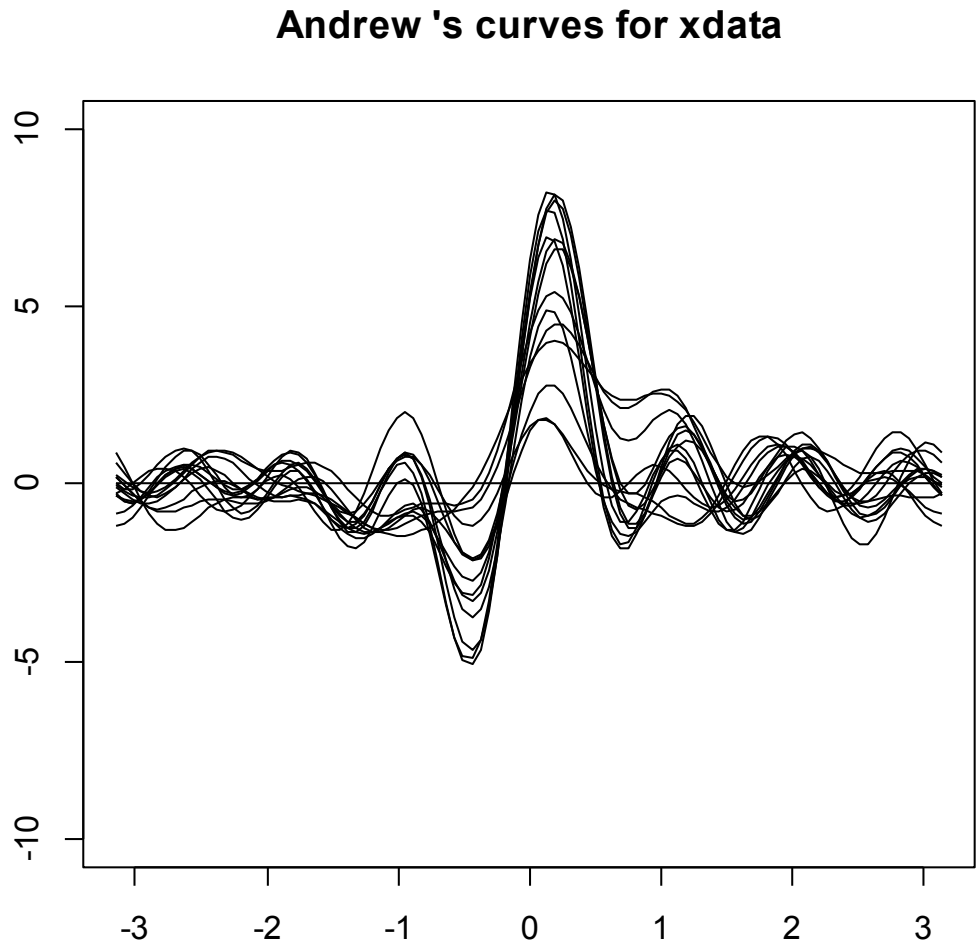
```
which( res %in% max(res))
```

που επιστρέφει κατευθείαν τον αριθμό 4, δηλαδή την τέταρτη στήλη του συνόλου δεδομένων.

Το δεύτερο όρισμα "clr" αφορά τον χρωματισμό κάποιων καμπυλών από το σύνολο των καμπυλών του γραφήματος και αφορά το χρωματισμό σημείων των καμπυλών που επηρεάζονται από κάποια συγκεκριμένη στήλη του συνόλου δεδομένων. Για παράδειγμα αν ορίσουμε την τιμή 2 που αντιστοιχεί στην δεύτερη κατηγορία επαγγέλματος και στον τύπο για τον σχηματισμό καμπύλης αυτή η κατηγορία είναι δίπλα από το ημίτονο θα χρωματιστούν τα σημεία που έχουν καμπυλώσει λόγω της ύπαρξης του ημιτόνου.

Το διάγραμμα λοιπόν θα είναι:

```
andrews(xdata,type=1,main="Andrew 's curves for xdata")
```



Διάγραμμα 5.11: Οι καμπύλες του Andrews για τον πίνακα συχνοτήτων για τον αριθμό των απασχολούμενων σε χιλιάδες ανά κατηγορία επαγγελμάτων τα έτη 2001 ως 2014

Στο Διάγραμμα 5.11 υπάρχουν 14 καμπύλες για τις 14 χρονιές στις οποίες έχει μετρηθεί ο

αριθμός των απασχολούμενων ανά επαγγελματικό κλάδο. Οι καμπυλώσεις των γραμμών έχουν σχηματιστεί από τη συνεισφορά κάθε επαγγέλματος στον τύπο για τον σχηματισμό τους. Δε φαίνεται να υπάρχει κάποια συγκεκριμένη χρονιά που να διαφοροποιείται τρομερά σε σχέση με τις άλλες τόσο που ηδιαφορά να είναι εμφανής (για παράδειγμα να έχει διαφορετική συχνότητα) αλλά ταυτόχρονα οι περισσότερες χρονιές δεν είναι ίδιες μεταξύ τους. Υπάρχουν μάλιστα τρεις ομαδοποιήσεις ανάμεσα στις διάφορες χρονιές. Μπορεί λοιπόν να υποθέσει κανείς ότι οι ομαδοποιήσεις αυτές αφορούν τις τιμές των τελευταίων ετών που είναι κοντινές (από το 2010 ως το 2014). Τα σημεία με τις έντονες καμπυλώσεις προέρχονται από τη συνεισφορά επαγγελμάτων που είχαν μεγάλες διασπορές στις τιμές των απασχολούμενων.

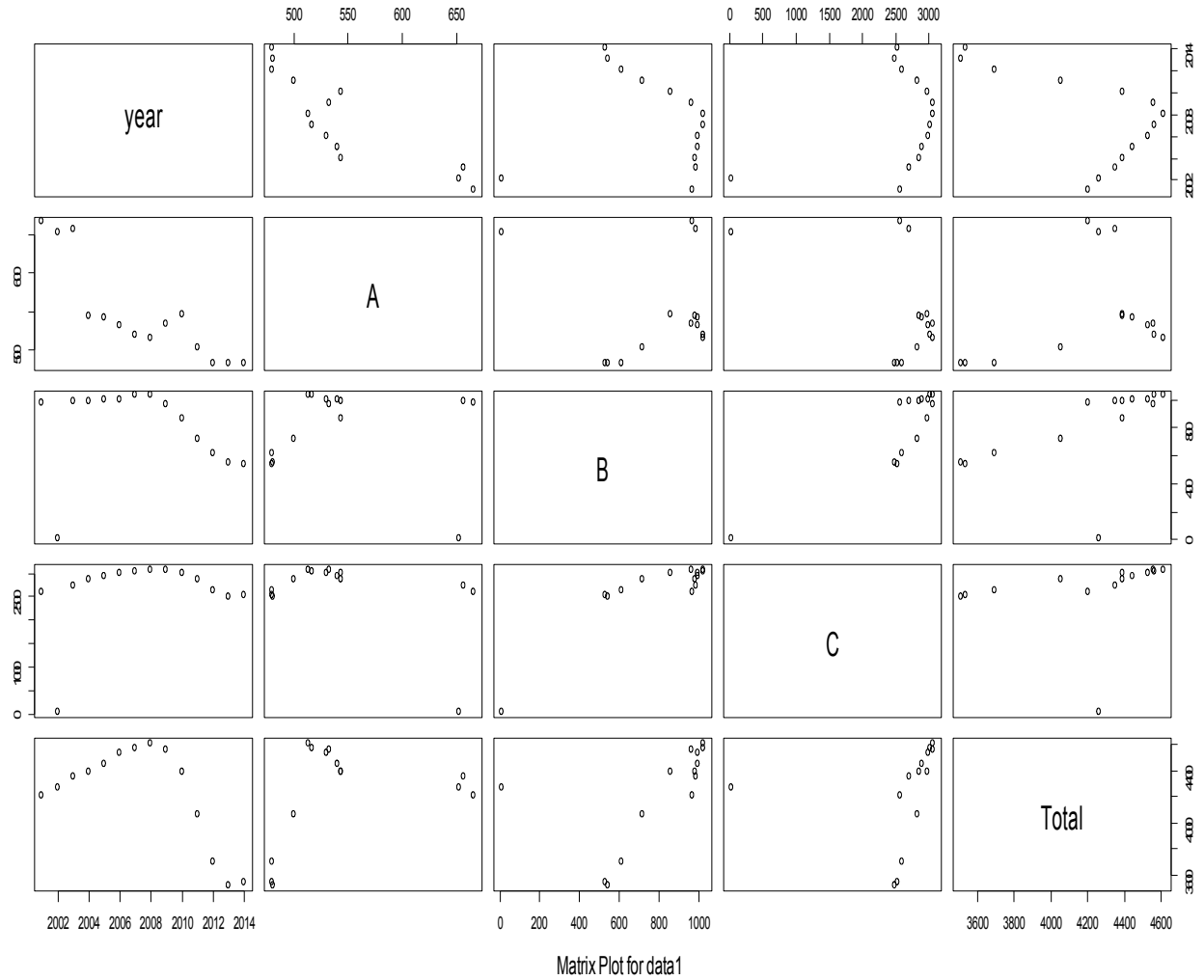
5.5 Διαγράμματα για τα δεδομένα των απασχολουμένων ανά κλάδο οικονομικής δραστηριότητας

Σε αυτή την παράγραφο θα ακολουθήσουν μία σειρά διαγραμμάτων που θα βοηθήσουν στην οπτική παρουσίαση της διαφοροποίησης του αριθμού εργαζομένων ανά οικονομικό κλάδο τις χρονιές 2001 ως 2014. Οι τρεις οικονομικοί κλάδοι που θα μελετηθούν είναι ο πρωτογενής, δευτερογενής και τριτογενής που έχουν συμβολιστεί με τα γράμματα A, B και C αντίστοιχα ενώ θα χρησιμοποιηθούν και γραφήματα για τον ολικό αριθμό των εργαζομένων στη χώρα.

5.5.1 Το διάγραμμα Matrix Plot

Το Matrix Plot δεν είναι παρά ένας οργανωμένος πίνακας από απλά διαγράμματα σημείων για ζευγάρια μεταβλητών. Το πλεονέκτημα του είναι πως μπορούμε να δούμε όλα τα δυνατά ζευγάρια ενώ επίσης επειδή οι κλίμακες είναι σταθερές μπορούμε να συγκρίνουμε ζεύγη μεταβλητών μεταξύ τους. Από ένα matrix plot μπορεί κανείς να αποκτήσει γρήγορα μια εικόνα για το ποιες μεταβλητές συσχετίζονται με ποιες άλλες. Οι ονομασίες των μεταβλητών υπάρχουν στη διαγώνιο του πίνακα. Για να βρει κανείς ποιο είναι το ζεύγος των μεταβλητών για το οποίο έχει σχηματιστεί το διάγραμμα σημείων αρκεί να βρει ποια είναι η μεταβλητή που απεικονίζεται σε κάθε γραμμή και σε κάθε στήλη του πίνακα. Το γράφημα αυτό στην R γίνεται με μία έτοιμη εντολή.

```
pairs(data1)
title(sub="Matrix Plot for data1")
```

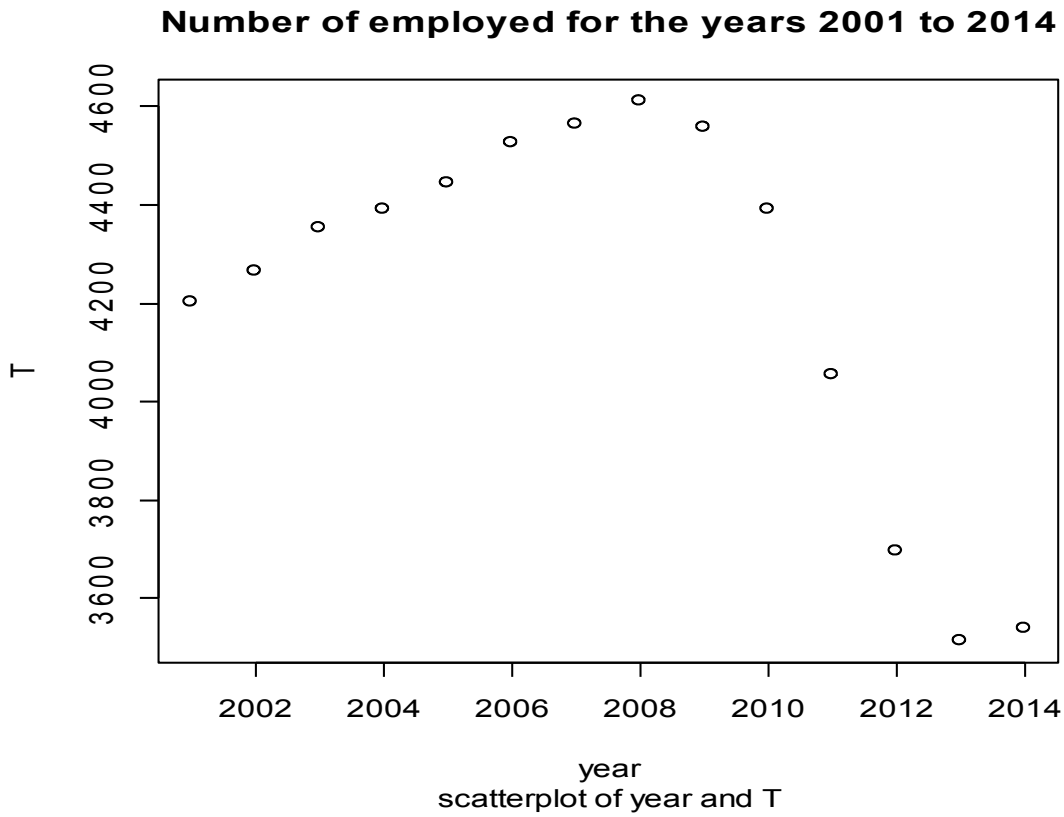


Διάγραμμα 5.12: Το διάγραμμα matrix plot για τον αριθμό των εργαζομένων σε χιλιάδες ανά οικονομικό κλάδο και του συνολικού αριθμού εργαζομένων για τα έτη 2001 ως 2014

Η πρώτη στήλη γραφικών παραστάσεων στο Διάγραμμα 5.12 είναι αυτή με την σημαντικότερη πληροφορία καθώς περιέχει τα γραφήματα κάθε οικονομικού κλάδου αλλά και του συνόλου των εργαζομένων με την πάροδο των ετών για οριζόντιο άξονα. Η πρώτη γραμμή γραφημάτων είναι τα ίδια γραφήματα με αντεστραμένους άξονες οπότε δε θα αναλυθούν περαιτέρω. Τέλος τα ενδιάμεσα γραφήματα γύρω από τη διαγώνιο παρουσιάζουν πιθανές συσχετίσεις μεταξύ του αριθμού εργαζομένων ανά οικονομικό κλάδο και του ολικού αριθμού. Τα άνω διαγώνια διαγράμματα είναι και πάλι όμοια με τα κάτω διαγώνια με αντεστραμμένους άξονες. Σε αυτά τα ενδιάμεσα διαγράμματα δεν παρατηρείται κάποια συστηματική συμπεριφορά ανά τις 3 κατηγορίες όμως φαίνεται ότι τα διαγράμματα που έχουν τον ολικό αριθμό εργαζομένων ανά κατηγορία παρουσιάζουν μία τάση γραμμικότητας για τις κατηγορίες Α,Β. Μπορεί δηλαδή να υποθέσει κανείς ότι όσο αυξάνεται ο ολικός αριθμός απασχολούμενων αυξάνεται γραμμικά και ο αριθμός απασχολούμενων στον δευτερογενή και τον τριτογενή οικονομικό κλάδο. Κάτι τέτοιο όμως δε συμβαίνει και με τον πρωτογενή κλάδο ο οποίος δε φαίνεται να ακολουθεί το ίδιο μοτίβο σε σχέση με τους άλλους δύο όσον αφορά τον συνολικό αριθμό των εργαζομένων.
Τα διαγράμματα της πρώτης στήλης περιέχουν σαν οριζόντιο άξονα τα έτη και σαν κατακόρυφους τις τρεις κατηγορίες οικονομικών κλάδων. Οι κατηγορίες Β και C

παρουσιάζουν μία σχετικά μικρή άνοδο από την αρχή ως το 2008 ενώ η πρώτη κατηγορία έχει φθίνει απότομα από το 2004 ως το 2008. Αυτό οφείλεται στην τάση των ανθρώπων προς επαγγέλματα αυτών των οικονομικών κλάδων και είναι αποτέλεσμα της αστυφιλίας αλλά και της ανάπτυξης των επαγγελμάτων που αντιστοιχούν σε αυτούς τους κλάδους στην πλειοψηφία των ανεπτυγμένων χωρών. Μετά το 2008 όμως παρατηρείται έντονη πτώση του αριθμού των εργαζομένων στις δύο τελευταίες κατηγορίες που οφείλεται στην οικονομική ύφεση η οποία ξεκινάει το 2009 ενώ ο πρωτογενής κλάδος εμφανίζει μία πολύ μικρή ανάκαμψη. Αυτό ίσως οφείλεται στην στροφή πολλών ατόμων εν μέσω κρίσης σε επαγγέλματα όπως η γεωργία μετά από αδυναμία εύρεσης εργασίας. Η διαφορά του αριθμού όμως των απασχολούμενων που στον πρωτογενή τομέα σε σχέση με πριν το 2008 δεν καλύπτει τη διαφορά στη μείωση των απασχολούμενων που έχει προκύψει από τους άλλους δύο οικονομικούς κλάδους ενώ μετά το 2010 και ο πρωτογενής τομέας εμφανίζει απότομη μείωση του αριθμού των απασχολούμενων.
Τέλος το γράφημα του συνολικού αριθμού των εργαζομένων ανά χρονιά μπορεί να δώσει μία γενική εικόνα της πορείας των διαθέσιμων θέσεων εργασίας στη χώρα από το 2001 ως το 2014. Είναι εμφανές ότι ο αριθμός απασχολούμενων ενώ παρουσίαζε μία σχετική άυξηση μειώνεται απότομα από το 2009 και μετά. Δεν έχουν μάλιστα υπάρξει στοιχεία που να αναφέρουν και ταυτόχρονη ανάλογη μείωση του πληθυσμού αυτά τα έτη οπότε μπορεί να συμπεράνει κανείς ότι έχουν μειωθεί αναλογικά οι διαθέσιμες θέσεις εργασίας.
Το τελευταίο διάγραμμα παρουσιάζεται και μόνο του παρακάτω.

```
plot(year,T)
title(main="Number of employed for the years 2001 to 2014",sub="scatterplot of year and T")
```

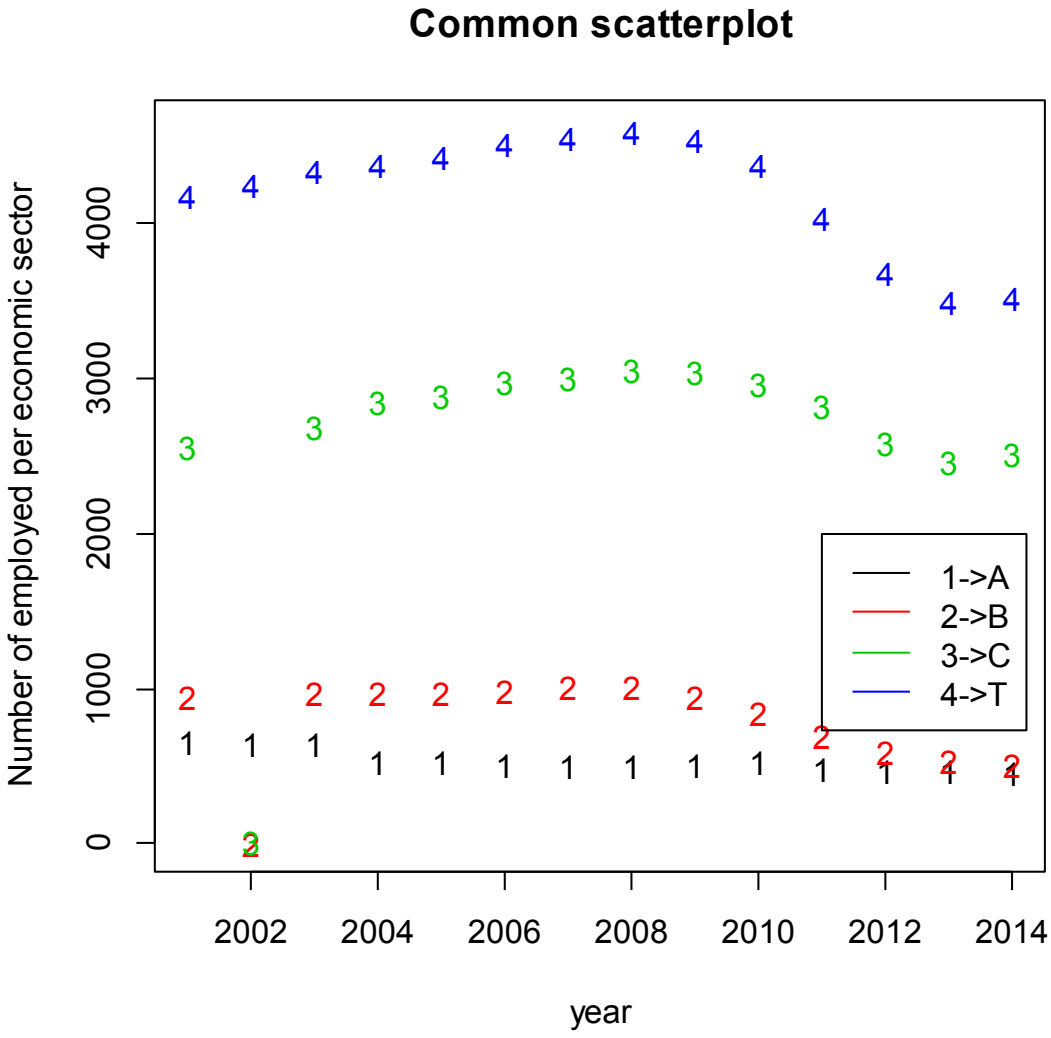


Διάγραμμα 5.13: Το διάγραμμα διασποράς του συνόλου των εργαζομένων σε χιλιάδες για τα έτη 2001 ως 2014

5.5.2 Το κοινό διάγραμμα διασπορών (matplot)

Το κοινό διάγραμμα διασπορών που μπορεί να κατασκευαστεί απευθείας στην R είναι το διάγραμμα διασποράς μίας συγκεκριμένης στήλης από το σύνολο των δεδομένων συναρτήσει όλων των υπολοίπων με τον διαχωρισμό των σημείων να γίνεται με διαφορετικούς αριθμούς και χρώματα.

```
matplot(year,xdata2,ylab="Number of employed per economic sector")
legend(2011,2000,lty=rep(1,4),col=1:2:3:4,legend=c("1->A","2->B","3->C","4->T"))
title(main="Common scatterplot")
```



Διάγραμμα 5.14: Το από κοινού διάγραμμα διασποράς για τους εργαζομένους σε χιλιάδες ανά οικονομικό κλάδο και του συνόλου των εργαζομένων για τα έτη 2001 ως 2014

Τοποθετημένα στο ίδιο διάγραμμα τα δεδομένα είναι πιο σαφή. Η γραφική παράσταση για τον αριθμό εργαζομένων ανά χρονιά στο από κοινού Διάγραμμα 5.13 είναι όμοια με εκείνη που σχηματίστηκε προηγουμένως αλλά όσον αφορά τις άλλες τρεις είναι πιο εύκολο να δει

κανείς πόσο μικρότερος είναι ο αριθμός των απασχολούμενων στις δύο πρώτες κατηγορίες σε σχέση με την τρίτη. Η πορεία ανά χρονιά του κάθε οικονομικού κλάδου σχολιάστηκε και παραπάνω και είναι ακριβώς ίδια. Τέλος στο Διάγραμμα 5.13 μπορεί κανείς να δει ξεκάθαρα ότι ο συνολικός αριθμός των εργαζομένων είναι το άθροισμα των εργαζομένων στις τρεις κατηγορίες οικονομικού κλάδου.

6. Τελική ανασκόπηση των αποτελεσμάτων

6.1 Παρουσίαση αποτελεσμάτων για κάθε μία μεταβλητή

Οι μεταβλητές που εξετάστηκαν σε αυτή την εργασία ήταν τα ποσοστά των ανέργων ανά τον βαθμός αστικότητας, το φύλο και την ηλικιακή ομάδα και το επίπεδο εκπαίδευσης καθώς και ο αριθμός των απασχολουμένων ανά επαγγελματικό τομέα και ανά κλάδο οικονομικής δραστηριότητας.

6.1.1 Ο αριθμός των ανέργων ανά βαθμό αστικότητας

Τα δεδομένα για τα ποσοστά των ανέργων ανά βαθμό αστικότητας περιείχε τρεις κατηγορίες βαθμού αστικότητας που ήταν οι άνεργοι σε αστικές, ημιαστικές και αγροτικές περιοχές. Μετά από την εφαρμογή της ανάλυσης αντιστοιχιών και μετά από διάφορα διαγράμματα το τελικό συμπέρασμα ήταν πως η πληθώρα του αριθμού των ανέργων συγκεντρώνονται στις αστικές και τις ημιαστικές περιοχές ενώ στις αγροτικές περιοχές η σχετική συχνότητα των ανέργων ήταν μικρότερη. Το συμπέρασμα αυτό ήταν αναμενόμενο δεδομένου και της αστυφιλίας που χαρακτηρίζει τη χώρα τα τελευταία χρόνια. Στη συνέχεια έγινε έρευνα για τις σχετικές συχνότητες των ανέργων σε τρία είδη αστικών περιοχών, την περιφέρεια Αττικής, το πολεοδομικό συγκρότημα Θεσσαλονίκης και όλες τις υπόλοιπες αστικές περιοχές της χώρας καθώς με αυτό τον τρόπο εξετάζεται η πορεία της ανεργίας σε σημεία ενδιαφέροντος που συγκεντρώνουν μεγάλο αριθμό πολιτών. Το κοινό διάγραμμα για αυτές τις τρεις κατηγορίες έκανε σαφές το γεγονός ότι η οικονομική ύφεση έγινε πιο αισθητή μέσω της ανεργίας στο πολεοδομικό συγκρότημα της Θεσσαλονίκης. Αυτό οφείλεται στο γεγονός ότι σαν πόλη έχει μεγάλο αριθμό κατοίκων αλλά δεν διαθέτει εργοστασιακές εγκαταστάσεις, εταιρίες και επιχειρήσεις ώστε να καλύπτουν το διαθέσιμο εργατικό δυναμικό που υπάρχει στην περιοχή.

6.1.2 Η σχετική συχνότητα των ανέργων ανά φύλο και ανά ηλικιακή ομάδα

Αρχικά όσον αφορά τα δύο φύλα ο αριθμός των ανέργων φαινόταν να διαφέρει αισθητά τις περισσότερες χρονιές. Μόνο τα τελευταία τρία έτη φαίνεται ότι τα ποσοστά ανεργίας σε κάθε φύλο δεν είναι τόσο διαφορετικά αφού τις προηγούμενες χρονιές οι γυναίκες άνεργοι φαίνεται να έχουν πολύ μεγαλύτερα ποσοστά σε σχέση με τους άντρες.Πάλι όμως η διαφοροποίηση των ποσοστών ανά φύλο είναι αισθητή.
Όσον αφορά την ηλικιακή ομάδα τα μεγαλύτερα ποσοστά ανέργων συναντώνται στις μικρότερες ηλικιακές ομάδες. Τα ποσοστά στις πολύ μικρές ηλικιακές ομάδες που το άτομο

δεν έχει αποκτήσει ακόμα πτυχίο ή κάποιο δίπλωμα ειδίκευσης αγγίζουν το 50% ενώ και οι ηλικιακή ομάδα των 30-44 ετών οπότε το άτομο θεωρείται ότι είναι στην πιο παραγωγική εργασιακά ηλικία δεν παρουσιάζουν χαμηλά ποσοστά ιδιαίτερα τα τελευταία έτη. Παρατηρείται όμως σε όλες τις ομάδες μία άνοδος στα ποσοστά μετά το 2008. Η ηλικιακή ομάδα που παρότι δεν περιέχει την πλειοψηφία των ανέργων έχει παρουσιάσει τη μεγαλύτερη αλλαγή σε σχέση με το παρελθόν είναι η ηλικιακή ομάδα 45-64 ετών. Πρόκειται για άτομα που δύσκολα επαναπροσλαμβάνονται αν έχουν χάσει την θέση εργασίας τους καθώς δεν είναι πλέον εύκολο να προσαρμοστούν σε νέες συνθήκες εργασίας και ταυτόχρονα δεν μπορούν να συνταξιοδοτηθούν ακόμα ώστε να σταματήσουν να αναζητούν θέση εργασίας. Τα μικρότερα ποσοστά ανέργων τέλος τα συγκεντρώνει η ηλικιακή ομάδα ατόμων άνω των 64 ετών. Αυτή η ηλικιακή ομάδα όμως δείχνει επίσης μία αύξηση μεγάλη αναλογικά με τα ποσοστά της τα τελευταία έτη. Ο αριθμός βέβαια φαίνεται και πάλι πολύ μικρός σε σχέση με άλλες ομάδες. Η αύξηση αυτή οφείλεται κατά πάσα πιθανότητα στην αύξηση των ατόμων σε αυτή την ηλικιακή ομάδα που μπήκαν στην διαδικασία αναζήτησης εργασίας λόγω μη κάλυψης των αναγκών τους πλέον από το υπάρχων εισόδημα.

6.1.3 Η σχετική συχνότητα των ανέργων ανά επίπεδο εκπαίδευσης

Οι άνεργοι ανά επίπεδο εκπαίδευσης παρουσιάστηκαν στο τέταρτο κεφάλαιο της εργασίας και τα δεδομένα αφορούσαν πάλι σχετική συχνότητα ανά κατηγορία επιπέδου εκπαίδευσης. Όσον αφορά τους κατόχους διδακτορικού ή μεταπτυχιακού διπλώματος όπως είναι αναμενόμενο τα ποσοστά ανεργίας είναι πολύ χαμηλότερα σε σχέση με τις άλλες κατηγορίες που οδηγεί στο συμπέρασμα ότι η υψηλή εκπαίδευση εξασφαλίζει καλύτερες πιθανότητες εύρεσης εργασίας για όλα τα έτη που μελετήθηκαν. Παρόλ'αυτά από το 2009 και μετά παρατηρήθηκε αύξηση των ποσοστών ανεργίας και σε αυτή την κατηγορία. Στη συνέχεια η κατηγορία των ατόμων με πτυχίο ανώτατου εκπαιδευτικού ιδρύματος παρουσιάζει πολύ μεγάλη διαφορά στα ποσοστά με το ποσοστό των ανέργων το 2014 να είναι διπλάσιο σε σχέση με εκείνο του 2001. Και πάλι βέβαια το ποσοστό των ανέργων δεν είναι τόσο μεγάλο σε σχέση με άλλες κατηγορίες. Ο διπλασιασμός αυτός δείχνει πως πλέον λόγω της πληθώρας των ατόμων με πτυχίο σε σχέση με το παρελθόν η κατοχή του δεν εξασφαλίζει μία θέση εργασίας. Οι τρεις επόμενες κατηγορίες περιείχαν τα άτομα με πτυχίο ανώτατου εκπαιδευτικού ιδρύματος και των ατόμων με πτυχίο λυκείου αλλά και πτυχίο τριών τάξεων ανώτερης εκπαίδευσης, δηλαδή άτομα που έχουν τελειώσει μέχρι γυμνάσιο. Αυτές οι τρεις κατηγορίες δεν παρουσιάζουν μεγάλες διαφορές μεταξύ τους αλλά έχουν αυξηθεί πολύ τα ποσοστά και στις τρεις κατηγορίες σε σχέση με παλαιότερα έτη. Τέλος οι τρεις τελευταίες κατηγορίες ατόμων με πολύ χαμηλότερη μόρφωση φαίνεται να ανταγωνίζονται μεταξύ τους με χαμηλότερα ποσοστά να συναντούνται στην κατηγορία των ατόμων που έχουν ολοκληρώσει το δημοτικό. Η διαφορά των ποσοστών σε αυτές τις κατηγορίες είναι τεράστια αφού πλέον συναντάει κανείς έως και οχταπλάσια ποσοστά το 2014 σε σχέση με το 2001. Αξιοσημείωτο είναι επίσης το γεγονός ότι τα παλαιότερα έτη η διαφορά των ποσοστών μεταξύ αυτών των τριών κατηγοριών ήταν ελάχιστη ενώ τα τελευταία χρόνια η διαφοροποίηση τους είναι ιδιαίτερα εμφανής.
Όλα αυτά οδηγούν στο συμπέρασμα ότι η κατοχή πτυχίων πλέον είναι απαραίτητη για όλες

τις ομάδες επαγγελμάτων στις οποίες θέλει να στραφεί κάποιος.

6.1.4 Ο αριθμός των ανέργων ανά οικονομικό κλάδο

Για την μεταβλητή της συχνότητας απασχολουμένων ανά επαγγελματικό τομέα εξετάστηκαν 15 επαγγελματικοί κλάδοι για τον αριθμό των εργαζομένων που απασχολούν από το 2001 ως το 2014. Στους περισσότερους από αυτούς παρατηρείται ότι υπάρχει αύξηση των απασχολούμενων από το 2001 μέχρι το 2008 ενώ από το 2009 και μετά έχουμε και πάλι πτώση του αριθμού των εργαζομένων. Υπάρχουν όμως και κατηγορίες που δεν ακολουθούν αυτό το μοτίβο. Για παράδειγμα η κατηγορία της Γεωργίας/Δασοκομίας/ Αλιείας/Κτηνοτροφίας φαίνεται να παρουσιάζει μία σταδιακή μείωση στον αριθμό εργαζομένων που απασχολεί με την πάροδο των ετών και ανακάμπτει λίγο μετά την αρχή της οικονομικής ύφεσης για να συνεχίσει την πτώση της στη συνέχεια. Ταυτόχρονα οι εργαζόμενοι σε ορυχεία και λατομεία παρουσιάζουν μία μεγάλη μείωση από το 2003 και μετά ενώ κατηγορίες όπως η δημόσια διοίκηση παρουσίασε τελικά μία μικρή πτώση με ενδιάμεσες αυξήσεις μικρής κλίμακας όμως. Ο τομέας που παρουσίασε τη μικρότερη μείωση του τελικού αριθμού εργαζομένων που απασχολεί σε σχέση με τον αριθμό εργαζομένων που απασχολούσε το 2001 είναι ο τομέας του εμπορίου που ακόμα και το 2014 συνέχισε να απασχολεί περίπου τον ίδιο αριθμό εργαζομένων. Το ίδιο συνέβη και με τους χρηματοπιστωτικούς οργανισμούς. Υπάρχει και μία κατηγορία επαγγελματικών κλάδων που παρουσίασε σχετική αύξηση του αριθμού των απασχολούμενων το 2014 εντός της οικονομικής κρίσης σε σχέση με το 2001. Πρόκειται για τους εργαζόμενους στον τομέα διαχείρισης ακίνητης περιουσίας και εκμισθώσεων. Όπως φαίνεται εντός τις κρίσης αυτή η κατηγορία έπεσε ελάχιστα σε αριθμό απασχολούμενων σε σχέση με νωρίτερα ενώ την τελευταία χρονιά παρουσίασε αύξηση, κάτι που δε συνέβη σε καμία άλλη κατηγορία επαγγελμάτων. Επίσης μεγάλη πτώση στον αριθμό εργαζομένων που απασχολεί το 2014 σε σχέση με το 2001 και τις ενδιάμεσες χρονιές παρουσιάζει ο κλάδος της παροχής υπηρεσιών. Τέλος υπάρχει και ο κλάδος της επεξεργασίας λυμάτων, διαχείρισης αποβλήτων και δραστηριοτήτων εξυγίανσης που δεν είχαν αναπτυχθεί αρκετά ως το 2008 για να απασχολεί εργαζόμενους. Η ανάγκη για την ανάπτυξη αυτού του τομέα οδήγησε στην απασχόληση περίπου 30.000 εργαζομένων και μπορεί να θεωρήσει κανείς ότι αυτό βοήθησε στην καταπολέμηση του αριθμού των ανέργων αφού υπήρξε η δημιουργία θέσεων εργασίας. Ο αριθμός των απασχολούμενων βέβαια μειώνεται και αυτός όπως σε όλες σχεδόν τις κατηγορίες επαγγελμάτων όσο η κρίση συνεχίζεται.

Όσον αφορά τους τρεις οικονομικούς κλάδους εργασίας όπως είναι αναμενόμενο ο τριτογενής κλάδος φαίνεται να απασχολεί την πλειοψηφία των εργαζομένων στη χώρα όπως συμβαίνει σχεδόν σε όλες τις ανεπτυγμένες χώρες. Ο πρωτογενής και ο δευτερογενής κλάδος σε αντίθεση απασχολούν και οι δύο μαζί λιγότερο από το μισό του συνόλου των εργαζομένων. Στον τριτογενή κλάδο, που απασχολεί την πληθώρα των εργαζομένων παρουσιάζεται μία αύξηση των απασχολούμενων ανά χρονιά ως το 2008 ενώ αργότερα παρουσιάζεται μία κατακόρυφη πτώση και μία ελάχιστη ανάκαμψη το τελευταίο έτος του συνόλου δεδομένων. Ο δευτερογενής κλάδος αν και απασχολεί πολύ μικρότερο αριθμό εργαζομένων παρουσιάζει και αυτός ένα μοτίβο που δε διαφέρει πολύ αλλά σε μικρότερη κλίμακα. Υπάρχει δηλαδή και εδώ μία αύξηση του συνόλου των εργαζομένων ως το 2008

και στη συνέχεια παρουσιάζεται μία πτώση. Η ανάκαμψη στο τέλος δεν είναι εμφανής σε αυτή την κατηγορία. Τέλος ο πρωτογενής κλάδος που εμφανίζεται να απασχολεί πολύ λιγότερους εργαζόμενους φαίνεται να έχει πιο ομαλή πορεία στην πάροδο των ετών παρουσιάζοντας μία σχετική πτώση όχι το 2008 όπως οι άλλοι δύο τομείς αλλά αρκετά νωρίτερα. Η πτώση αυτή δεν οφείλεται στην κρίση αλλά στην αστυφιλία που αναπτυσσόταν όλο και περισσότερο στη χώρα. Στη συνέχεια παρουσιάστηκε μία αύξηση του αριθμού των εργαζομένων και πάλι με την αρχή της κρίσης η οποία όμως δεν μπορεί να θεωρηθεί αρκετά μεγάλη ώστε να καλύψει τη διαφορά που έχει προκύψει από την τεράστια μείωση στον αριθμό των εργαζομένων από τους άλλους δύο οικονομικούς κλάδους. Ο ολικός αριθμός των εργαζομένων στην πάροδο των ετών λοιπόν μειώθηκε εμφανώς, κάτι που επιβεβαιώνεται και από το διάγραμμα του διανύσματος του συνόλου των εργαζομένων για τις χρονιές του δείγματος. Ταυτόχρονα δεν έχει αναφερθεί αντίστοιχη μείωση του συνόλου του εργατικού δυναμικού στον πληθυσμό οπότε η πτώση αυτή δείχνει ξεκάθαρα την αντίστοιχη αύξηση του συνόλου των ανέργων.

6.2 Τελικό συμπέρασμα

Από τα δεδομένα που χρησιμοποιήθηκαν και τα διαγράμματα που πραγματοποιήθηκαν είναι εμφανές ότι το ελληνικό μοντέλο απασχόλησης χαρακτηρίζεται από υψηλά ποσοστά ανεργίας και χαμηλό ποσοστό συμμετοχής του οικονομικά ενεργού πληθυσμού στην αγορά εργασίας. Χαρακτηριστικό είναι το γεγονός ότι την περίοδο 2000-2006 πάνω από τις μισές νέες θέσεις εργασίας δημιουργήθηκαν σε τέσσερις μόλις κλάδους παραγωγής: στις κατασκευές, στο εμπόριο, στις υπηρεσίες εστίασης και ξενοδοχείων και στην γενικότερη παροχή υπηρεσιών (αυτή η πληροφορία προέρχεται από το Κέντρο Πληροφόρησης Εργαζομένων & Ανέργων(2012)). Το γεγονός ότι η πλειονότητα των νέων θέσεων εργασίας αφορούν εργασίες χαμηλής ειδίκευσης συνεπάγεται την αδυναμία αξιοποίησης των δυνατοτήτων των εργαζομένων με αρνητικές συνέπειες τόσο στην παραγωγικότητα της εργασίας όσο και στο αίσθημα ικανοποίησής τους από την εργασία τους. Ειδικά στους νέους η κατάσταση είναι ακόμα χειρότερη, δεδομένου ότι περίπου ένας στους τέσσερις δηλώνει ότι η εργασία που κάνει δεν ανταποκρίνεται στα προσόντα του (πληροφορία από την ΕΛ.ΣΤΑΤ. από δημοσκοπική έρευνα). Σήμερα δεν είναι λίγοι οι νέοι και οι νέες που στα τριάντα τους χρόνια μένουν ακόμα στην οικογενειακή εστία ή δεν διαθέτουν ένα αυτόνομο δικό τους εισόδημα λόγω των υψηλών ποσοστών ανεργίας ειδικά για ηλικιακές ομάδες που δεν είναι φυσιολογικό να συναντούνται τόσο μεγάλα ποσοστά, όπως άτομα μεταξύ των 30 με 44 ετών. Ειδικότερα για τα έτη 2001 ως και 2014 που προέρχονται τα δεδομένα αυτής της εργασίας φαίνεται η επίπτωση της οικονομικής ύφεσης της χώρας. Τα ποσοστά από το 2008 και ύστερα φαίνεται να έχουν αυξηθεί ραγδαία και να έχουν επηρεάσει κλάδους και ομάδες ατόμων που θα μπορούσαν να θεωρηθούν ασφαλείς από την ανεργία υπό άλλες συνθήκες. Η υψηλή μόρφωση φαίνεται να είναι ακόμα ένας τρόπος βελτίωσης των πιθανοτήτων εύρεσης εργασίας αλλά ακόμα και σε αυτή την κατηγορία υπάρχουν πολλοί άνεργοι. Σύμφωνα με στοιχεία της ΕΛ.ΣΤΑΤ. μάλιστα και από το σύνολο των εργαζομένων σε αυτή την ηλικιακή ομάδα δεν έχουν εξασφαλίσει μία μόνιμη εργασία παρά ελάχιστοι. Μόνο το ένα τέταρτο του εργατικού δυναμικού φαίνεται να έχει αυτό το προσόν σήμερα και από αυτούς η πλειοψηφία είναι από παλιότερους διορισμούς πριν την ύφεση.

Βιβλιογραφία

Α) Διεθνής Βιβλιογραφία

Maindonald J. and Braun J. (2003), *Data Analysis and Graphics Using R–an Example – based Approach*. Cambridge University Press. London.

Mair P. and Jan de Leeu (2009), Simple and canonical Correspondance Analysis using the r package anacor, *Journal of Statistical Software,* **31**, **Issue 5,** 1-18.

Maravelakis P. and Bersimis S. (2009), The use of Andrews curves for detecting the out-of-control variables when a multivariate control chart signals, *Statistical Papers,* **50**, 51-65.

Nenadic O. and Greenacre M. (2007), Correspondance Analysis in R, with two and three dimensional graphics: The package ca, *Journal Of Statistical software,* **20**, **Issue 3**, 1-13.

Β) Ελληνική Βιβλιογραφία

Ελληνική Στατιστική Αρχή (2014), *Βάση στατιστικών δεδομένων/Αγορά Εργασίας/ Απασχόληση Ανεργία/Πίνακες Δεδομένων*, Αθήνα.

Ελληνικό Στατιστικό Ινστιτούτο (2009), *Λεξικό Στατιστικής Ορολογίας*, Έκδοση Ελληνικού Στατιστικού Ινστιτούτου, Αθήνα.

Καρλής Δ. (2005), *Πολυμεταβλητή Στατιστική Ανάλυση*, Εκδόσεις Σταμούλης, Αθήνα.

Κοκολάκης Γ. και Φουσκάκης Δ. (2009), *Στατιστική: Θεωρία και Εφαρμογές*, Εκδόσεις Συμεών, Αθήνα.

Μενεξές Γ., Μάρκος Α. και Παπαδημητρίου Γ. (2005), *Ελλείψεις εμπιστοσύνης στα παραγοντικά επίπεδα της ανάλυσης αντιστοιχειών*, Ελληνικό Στατιστικό Ινστιτούτο, Πρακτικά $18^{ου}$ Συνέδριου Στατιστικής.

Φουσκάκης Δ. (2013), *Ανάλυση Δεδομένων με Χρήση της R*, εκδόσεις Τσότρας, Αθήνα.

Φουσκάκης Δ. (2014), Σημειώσεις Μαθήματος *Υπολογιστικές μέθοδοι στη στατιστική*, Αθήνα.

Ψαρράκος Π. (2012), *Θέματα Ανάλυσης Πινάκων,* Εθνικό Μετσόβιο Πολυτεχνείο, Αθήνα.

Borjas G. (2003), *Τα οικονομικά της ανεργίας*, Εκδόσεις Κριτική.

Γ) Ιστοσελίδες

http://cran.r-project.org/web/packages.

http://www.statmethods.net, quick-R/accesing the power of R.

http://stats.stackexchange.com , Cross Validated, forum.

http://www.r-bloggers.com/exploratory-data-analysis-variations-of-box-plots-in-r/.

http://research.stowers−institute.org/efg/R/Color/Chart.