Ε Λ Λ Η Ν Ι Κ Ο Μ Ε Σ Ο Γ Ε Ι Α Κ Ο

Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο

Σχολή Ηλεκτρολόγων Μηχανικών
Τμήμα Οργάνωση και Διοίκηση για Μηχανικούς

Διπλωματική Εργασία

ΕΦΑΡΜΟΓΗ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ K- MEANS ΣΤΗΝ ΟΜΑΔΟΠΟΙΗΣΗ ΓΕΩΓΡΑΦΙΚΩΝ ΠΕΡΙΟΧΩΝ

ΑΝΔΡΟΝΙΚΗ ΚΑΠΠΑ
Μ.Ο. 36

Εισηγητής
ΞΑΝΘΟΣ ΓΕΩΡΓΙΟΣ

Ηράκλειο, Φεβρουάριος 2020

“All the real knowledge which we possess depends on methods by which we distinguish the similar from the dissimilar. The greater number of natural distinctions this method comprehends the clearer becomes our idea of things. The more numerous the objects which employ our attention the more difficult it becomes to form such a method and the more necessary.” (Linnæus 1737)

«Όλες οι πραγματικές γνώσεις που κατέχουμε εξαρτώνται από τις μεθόδους που διακρίνουμε τα παρόμοια από τα ανόμοια. Όσο μεγαλύτερος είναι ο αριθμός των φυσικών διακρίσεων που κατανοεί η μέθοδος αυτή, τόσο πιο σαφής γίνεται η ιδέα μας για τα πράγματα. Όσο πιο πολλά αντικείμενα απασχολούν την προσοχή μας τόσο περισσότερο δύσκολο γίνεται να διαμορφωθεί μια τέτοια μέθοδος και πιο απαραίτητη.» (Linnæus 1737)

Copyright © Ανδρονίκη Καππά, 2019

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Η έγκριση της πτυχιακής εργασίας από το πρόγραμμα δεν υποδηλώνει απαραιτήτως και αποδοχή των απόψεων του συγγραφέα εκ μέρους του Τμήματος.

Ευχαριστίες

Ως την ελάχιστη δυνατή μνεία, με την παρούσα παράγραφο οφείλω αλλά και θέλω να ευχαριστήσω όλους όσους συνέβαλαν στην εκπόνησή της:

Τον επιβλέποντα καθηγητή μου, κύριο Ξανθό Γιώργο, για την εμπιστοσύνη που μου έδειξε εξ' αρχής, αναθέτοντάς μου το συγκεκριμένο θέμα, για την επιστημονική του καθοδήγηση, για την ανεκτίμητη υποστήριξή του, τις υποδείξεις του καθώς και το πολύ καλό κλίμα συνεργασίας που διαμόρφωσε.

Τέλος, θα ήθελα εκφράσω την ευγνωμοσύνη μου στην οικογένειά μου και ιδιαίτερα στο σύζυγό μου για όλη τη στήριξη, τη συμπαράσταση και την κατανόησή του, καθ' όλη τη διάρκεια των σπουδών μου.

Περιεχόμενα

Ευρετήριο Σχημάτων 5
Εισαγωγή 8
ΚΕΦΑΛΑΙΟ 1: Ομαδοποίηση δεδομένων 11
1.1 Γενικά περί ομαδοποίησης 11
1.2 Τύποι αλγορίθμων ομαδοποίησης 13
1.3 Αλγόριθμος ομαδοποίησης k-means 15
1.4 Πώς λειτουργεί ο αλγόριθμος k-means 16
1.5 Πλεονεκτήματα και μειονεκτήματα του αλγορίθμου k-means 18
1.6 Μέθοδοι αξιολόγησης αλγορίθμων ομαδοποίησης 19
1.6.1 Η μέθοδος Elbow 20
1.6.2 Η ανάλυση Silhouette 21
1.6.3 Άλλοι μέθοδοι 21
ΚΕΦΑΛΑΙΟ 2 : Χρήση του αλγορίθμου k-means σε γεωγραφικές περιοχές 23
2.1 Σκοπός ομαδοποίησης γεωγραφικών περιοχών 23
2.2 Γενικά συμπεράσματα 27
ΚΕΦΑΛΑΙΟ 3: Χρησιμότητα εφαρμογής του αλγορίθμου k-means σε στατιστικά δεδομένα των περιφερειών της Ελλάδας 33
3.1 Χρησιμότητα και οφέλη 33
ΚΕΦΑΛΑΙΟ 4: Εφαρμογή του αλγορίθμου k-means σε στατιστικά δεδομένα των περιφερειών της Ελλάδας 34
4.1 Πιθανά ερευνητικά ερωτήματα και προβληματισμοί 34
4.2. Στατιστική ανάλυση των δειγμάτων 36
4.3. Εφαρμογή του αλγορίθμου k-means στα στατιστικά δεδομένα των μεταβλητών 39
4.3.1 Ορισμός αριθμού συστάδων k 39
4.3.2 Επεξήγηση του αλγορίθμου k-means στο Excel 39
4.4 Αντίκτυπος των μεταβλητών στην ομαδοποίηση 41
4.4.1 Δοκιμές ομαδοποίησης με 7 μεταβλητές και 3 συστάδες. 41
4.4.2 Επηρεασμός των κεντροειδών από πολλαπλές μεταβλητές 44
4.4.3 Τελική Ομαδοποίηση με 6 μεταβλητές και 5 συστάδες. 46
4.4.3.1 Συστάδα 1 47
4.4.3.2 Συστάδα 2 47
4.4.3.3 Συστάδα 3 48

4.4.3.4 Συστάδα 4.....48
4.4.3.5 Συστάδα 5.....49
ΚΕΦΑΛΑΙΟ 5: Συμπεράσματα.....50
5.1 Συμπεράσματα της εφαρμογής του αλγορίθμου k-means.....50

Ευρετήριο Σχημάτων

Σχήμα 1.4: Λογικό διάγραμμα του αλγορίθμου k-means.....17

Σχήμα 1.6.1: Η μέθοδος Elbow για την εύρεση του κατάλληλου k.....20

Σχήμα 4.2.1: Δείγματα διασποράς μεταξύ μεταβλητών. 38

Σχήμα 4.4.1.1: Ομαδοποίηση δεδομένων με τις μεταβλητές«Κατά κεφαλή ΑΕΠ» και «Ακαθάριστη προστιθέμενη αξία ανά εργαζόμενο». 42

Σχήμα 4.4.1.2: Ομαδοποίηση δεδομένων με 7 μεταβλητές και 3 συστάδες 43

Σχήμα 4.4.1.3: Ομαδοποίηση δεδομένων με 7 μεταβλητές και 3 συστάδες 44

Περίληψη

Αυτή η εργασία παρέχει μια κατευθυντήρια γραμμή για την πραγματοποίηση ανάλυσης συστάδων σε γεωγραφικές περιοχές (νομούς) προκειμένου να εξορυχθούν δεδομένα για τις περιφέρειες στις οποίες ανήκουν.

Ο σκοπός της ερευνητικής πρότασης είναι να δείξει πως με τη χρήση ενός απλού λογισμικού είναι δυνατό να επιτευχθούν σύνθετες διεργασίες. Πιο συγκεκριμένα, στη παρούσα έρευνα, με τη χρήση του Excel και μόνο, θα επιτευχθεί ομαδοποίηση των γεωγραφικών περιοχών στην Ελλάδα, με βάση τα κοινά χαρακτηριστικά των κατοίκων των επιμέρους περιοχών τους. Τα χαρακτηριστικά που θα ληφθούν υπόψη είναι δημογραφικά, οικονομικά και κοινωνικά.

Η μεθοδολογία που θα χρησιμοποιηθεί είναι εκτεταμένη βιβλιογραφική έρευνα και εξόρυξη δεδομένων από στατιστικά στοιχεία της Στατιστικής Υπηρεσίας. Ο αλγόριθμος ομαδοποίησης που χρησιμοποιείται είναι ο k-means.

Τα βασικά συμπεράσματα της έρευνας είναι:

- Το Excel μπορεί να αποτελέσει ένα εργαλείο για τέτοιου είδους ομαδοποιήσεις, με χρήση απλών έτοιμων συναρτήσεων χωρίς την ανάγκη προγραμματισμού.
- Οι νομοί της Ελλάδος χωρίζονται σε 5 συστάδες η κάθε μία από τις οποίες έχει τα δικά της ιδιαίτερα χαρακτηριστικά.
- Υπάρχει περιθώριο συνεργασίας μεταξύ των διοικητικών περιφερειών ώστε νομοί που αντιμετωπίζουν τις ίδιες προκλήσεις ή έχουν παρόμοια δυναμική ανάπτυξης να έχουν κοινό σχέδιο διοίκησης και ανάπτυξης**.**

Abstract

This work provides a guideline for conducting cluster analysis in geographical areas in order to determine the assessment areas.

The purpose of the research proposal is to show how using simple software can accomplish complex processes. Specifically, in the present study, using Excel alone, a grouping of geographical areas in Greece will be achieved, based on the common characteristics of the inhabitants of their individual regions. The characteristics to be considered are demographic, economic, social as well as medical.

The methodology to be used is extensive bibliographic research and data extraction from Statistical Service of Greece. The clustering algorithm used is k-means. Excel can be a tool for this kind of grouping.

The main findings of the research are:

- Excel can be a tool for this kind of grouping, using existing functions.
- Greek Prefectures can be divided in 5 clusters, each with their own general characteristics, problems and areas for development .
- There is an opportunity for collaboration between greek Administrative Regions that have prefectures of the same cluster in order to join their forces when facing similar challenges**.**

Εισαγωγή

Μία πρόκληση της σημερινής εποχής, είναι η διαμόρφωση μεθόδων με τις οποίες θα καθίσταται δυνατή η εξαγωγή χρήσιμων πληροφοριών από την ολοένα και αυξανόμενη ποσότητα δεδομένων. Η ομαδοποίηση δεδομένων είναι μία ισχυρή και χρήσιμη τεχνική που χρησιμοποιείται κυρίως στη Μηχανική Μάθηση (Machine Learning) με στόχο την άντληση πληροφορίας σχετική με τη δομή των δεδομένων και απώτερο σκοπό την ανακάλυψη ενός μοντέλου πυκνότητας των δεδομένων.

Τα δύο πιο βασικά ερωτήματα που αφορούν την ομαδοποίηση δεδομένων είναι «Πόσες ομάδες υπάρχουν;» και «Ποιες είναι οι ομάδες;». Υπάρχουν πολλοί αλγόριθμοι που απαντούν στο δεύτερο ερώτημα αλλά δεν ισχύει το ίδιο και για το πρώτο ερώτημα. Ο αλγόριθμός k-means που θα χρησιμοποιηθεί στη παρούσα έρευνα, απαντά και στα δύο βασικά ερωτήματα.

Όταν πρόκειται για μικρό όγκο δεδομένων, η διαμόρφωση ομάδων είναι μια διαδικασία που ο άνθρωπος μπορεί να επιτελέσει και χωρίς τη χρήση μηχανών. Οι ομάδες που θα δημιουργηθούν στη παρούσα εργασία, θα αφορούν τον πληθυσμό της Ελλάδας και τα χαρακτηριστικά του, πράγμα που σημαίνει ότι ο αριθμός των διαστάσεων είναι μεγάλος και ο όγκος των δεδομένων ξεπερνά τις δυνατότητες που έχει ο άνθρωπος. Βέβαια, αυτό δε σημαίνει πως απαιτούνται ακριβά, δύσχρηστα και πολύπλοκα προγράμματα για να επιτευχθεί η επιθυμητή ομαδοποίηση. Αντίθετα ένα πρόγραμμα που οι περισσότεροι άνθρωποι έχουν ήδη εγκαταστημένο στον υπολογιστή τους και χρησιμοποιούν καθημερινά, μπορεί να χρησιμοποιηθεί με ικανοποιητικά αποτελέσματα.

Ο στόχος της ερευνητικής πρότασης είναι να δείξει πως με τη χρήση ενός απλού λογισμικού είναι δυνατό να επιτευχθούν σύνθετες διεργασίες. Πιο συγκεκριμένα, στη παρούσα έρευνα, με τη χρήση του excel και μόνο, θα επιτευχθεί ομαδοποίηση των γεωγραφικών περιοχών στην Ελλάδα, με βάση τα κοινά χαρακτηριστικά των κατοίκων των επιμέρους περιοχών τους. Τα χαρακτηριστικά που θα ληφθούν υπόψη είναι δημογραφικά, οικονομικά και κοινωνικά.

Ο απώτερος σκοπός της έρευνας είναι να συγκρίνει τις σημερινές 13 περιφέρειες της Ελλάδος, διαμορφώνοντας ομάδες με όμοια και ομάδες με ανόμοια χαρακτηριστικά των κατοίκων της. Η σύγκριση αυτή, θα έχει ως αποτέλεσμα αρχικά τη πιθανότητα να εμφανιστούν νομοί που ανήκουν σε μια περιφέρεια και έχουν ανόμοια χαρακτηριστικά σε σχέση με τους όμορους νομούς της περιφέρειας. Επιπρόσθετα, από τη σύγκριση θα

προκύψουν ομάδες, χρήσιμες τόσο σε διοικητικό επίπεδο αλλά και σε επιχειρησιακό (τμηματοποίηση της αγοράς)

Η αξία της εν λόγω προτεινόμενης ερευνητικής πρότασης μπορεί να εκτιμηθεί από το γεγονός ότι τα αποτελέσματά της, θα αποτελέσουν χρήσιμη πηγή πληροφορίας για τις περιφέρειες. Η ομαδοποίηση είναι μία τεχνική που βοηθάει στη λήψη αποφάσεων. Για παράδειγμα, μελετώντας, μεταβλητές ιατρικού περιεχομένου, μπορεί να επιτευχθεί μια κατηγοριοποίηση των ασθενών με όμοια συμπτώματα και να προσδιοριστούν οι αιτίες των ασθενειών ή να ληφθούν μέτρα πρόληψης και καλύτερης αντιμετώπισης. Υπάρχει ακόμη και η δυνατότητα να αναγνωριστούν οι δολιοφθορές εις βάρος κάποιας περιφέρειας, αν οι μεταβλητές που μελετώνται είναι κατάλληλες. Η τεχνική της ομαδοποίησης σχετίζεται άμεσα και με την τεχνική της βελτιστοποίησης και οι εφαρμογές της πολλαπλασιάζονται αν το ζητούμενο είναι η εύρεση ομοιογενών ομάδων, το μικρό κόστος ή το μέγιστο κέρδος.

Αρχικά ότι το excel όντως μπορεί να αποτελέσει ένα εργαλείο για τέτοιου είδους ομαδοποιήσεις. Και συμπεράσματα για ανόμοια και όμοια χαρακτηριστικά σε νομούς μιας περιφέρειας.

Η ερευνητική μέθοδος (συλλογή δεδομένων) θα ακολουθήσει τα εξής στάδια:

Το πρώτο στάδιο της συλλογής των δεδομένων είναι η επιλογή των κατάλληλων μεταβλητών βάσει των οποίων θα γίνει η ομαδοποίηση του πληθυσμού. Είναι καίριας σημασίας οι μεταβλητές που θα επιλεχθούν, να οδηγούν σε καθαρά διαφοροποιημένες ομάδες και να έχουν μεταξύ τους μικρό βαθμό συσχέτισης. Είναι κοινή πρακτική, και θα τη υιοθετήσει και η παρούσα έρευνα, για να αποφευχθεί η πιθανότητα χρήσης μεταβλητών με μεγάλο βαθμό συσχέτισης, να διεξάγεται αρχικά μια παραγοντική ανάλυση και στη συνέχεια στα αποτελέσματα αυτής να εφαρμόζεται η ομαδοποίηση. Στο ίδιο στάδιο θα πρέπει να γίνει η επιλογή μεγέθους του δείγματος η οποία είναι άρρηκτα συνδεδεμένη με την επιλογή μεταβλητών.

Το δεύτερο στάδιο, περιλαμβάνει τη μεθοδολογία ταξινόμησης που θα επιλεγεί. Στη παρούσα έρευνα, έχει ήδη αποφασιστεί η χρήση του αλγορίθμου k-means λόγω της κοινής του αποδοχής ως κατάλληλη για τέτοιου είδους δεδομένα αλλά και γιατί απαιτεί μικρή υπολογιστική δύναμη γεγονός που εξυπηρετεί το σκοπό της έρευνας. Ο αλγόριθμος αυτός ανήκει στη κατηγορία των διαιρετικών μεθόδων ταξινόμησης (Partitioning Clustering) και πιο συγκεκριμένα στην υποκατηγορία Επαναληπτικής Διαιρετικής Ταξινόμησης (Iterative Divisive Partitioning).

Το επόμενο και τελευταίο στάδιο, περιλαμβάνει την επιλογή μέτρου ομοιότητας για τη σύγκριση των μεταβλητών.

ΚΕΦΑΛΑΙΟ 1: Ομαδοποίηση δεδομένων

1.1 Γενικά περί ομαδοποίησης

Η εξόρυξη δεδομένων (data mining) αναφέρεται σε ένα σύνολο προσεγγίσεων και τεχνικών για την ανεύρεση προτύπων σε μεγάλα σύνολα δεδομένων που επιτρέπουν την εξαγωγή πολύτιμων πληροφοριών. Επίσης, στη βιβλιογραφία συναντάται ως ανακάλυψη της γνώσης σε Βάσεις Δεδομένων. Για παράδειγμα, ένας κατασκευαστής καταναλωτικών προϊόντων μπορεί να χρησιμοποιήσει την εξόρυξη δεδομένων για να κατανοήσει καλύτερα τη σχέση των πωλήσεων ενός συγκεκριμένου προϊόντος με τις στρατηγικές προώθησης, τα χαρακτηριστικά πώλησης των καταστημάτων και τα δημογραφικά στοιχεία των περιφερειών (Olkin & Sampson, 2001). Μια από τις τεχνικές εξόρυξης δεδομένων είναι η ομαδοποίηση.

Η ομαδοποίηση ή συσταδοποίηση (clustering) είναι η διαδικασία διαίρεσης ενός συνόλου από δεδομένα σε διακεκριμένες ομάδες. Η διαδικασία επιτελείται με τέτοιο τρόπο ώστε τα δεδομένα που ανήκουν σε μια ομάδα να έχουν ομοιότητες ενώ τα δεδομένα που ανήκουν σε άλλες μεταξύ τους ομάδες να έχουν διαφορές (Alsabti, Ranka and Singh, 1997). Μια **συστάδα** ή **σύμπλεγμα** (cluster) αναφέρεται ως μια συλλογή αντικειμένων (σημείων δεδομένων) συγκεντρωμένων μεταξύ τους λόγω ορισμένων ομοιοτήτων. Έτσι, τα αντικείμενα που είναι παρόμοια πρέπει να βρίσκονται στην ίδια συστάδα και τα αντικείμενα που είναι ανόμοια πρέπει να είναι σε διαφορετικές συστάδες. Έτσι, η ομαδοποίηση μπορεί να οριστεί ως η διαδικασία διαίρεσης αντικειμένων σε συστάδες έτσι ώστε αντικείμενα εντός της ίδιας συστάδας να έχουν υψηλό βαθμό ομοιότητας, ενώ αντικείμενα που ανήκουν σε διαφορετικές συστάδες να έχουν υψηλό βαθμό ανομοιότητας (Kaufman και Rousseeuw, 1990). Με άλλα λόγια, προσπαθούμε να βρούμε ομοιογενείς ομάδες ή υποομάδες μέσα στα δεδομένα έτσι ώστε τα σημεία των δεδομένων σε κάθε συστάδα να είναι όσο το δυνατόν παρόμοια, σύμφωνα με ένα μέτρο ομοιότητας, όπως είναι η απόσταση (που βασίζεται κυρίως στην Ευκλείδεια απόσταση) μεταξύ των αντικειμένων ή την απόσταση που βασίζεται σε

συσχετισμό τους. Η ομαδοποίηση ονομάζεται συχνά μη επιτηρούμενη ταξινόμηση επειδή παράγει το ίδιο αποτέλεσμα με την ταξινόμηση αλλά χωρίς προκαθορισμένες ομάδες.

Η Ευκλείδεια απόσταση των σημείων x και y ορίζεται ως εξής:

$$d = \sqrt{\sum_{j=1}^{n} (x_j - y_j)^2}$$

Οι Kaufman και Rousseeuw (1990) ορίζουν την ανάλυση συστάδων ή η διαδικασία ομαδοποίησης των δεδομένων (cluster analysis) ως την τέχνη της εύρεσης ομάδων σε αντικείμενα (δεδομένα). Η ανάλυση συστάδων εφαρμόζεται με τέτοιο τρόπο ώστε τα περισσότερο όμοια δεδομένα να εντάσσονται σε ίδιες συστάδες (ομάδες) στοιχεία (παρατηρήσεις) περισσότερα όμοια μεταξύ τους (Aldenderfer & Blashfield, 1984). Η τεχνική της ομαδοποίησης υπάγεται στην ευρύτερη κατηγορία των τεχνικών μάθησης χωρίς επίβλεψη.

Όλα τα αντικείμενα που ανήκουν στην ίδια συστάδα πρέπει να (Τζαμπαλάτης & Δασύρα, 2011):

- μοιράζονται τις ίδιες ιδιότητες ή να σχετίζονται στενά,
- έχουν μικρές αμοιβαίες αποστάσεις ή ανομοιότητες,
- έχουν “επαφές” ή “σχέσεις” με τουλάχιστον ένα άλλο αντικείμενο της ομάδας,
- ξεχωρίζουν ευκρινώς από το υπόλοιπα αντικείμενα του συνόλου δεδομένων,
- υπάρχουν συνεχείς και σχετικά υψηλής πυκνότητας σημείων περιοχές στο χώρο.

Ο τελικός στόχος της ομαδοποίησης είναι να παρέχει στους χρήστες σημαντικές πληροφορίες από τα αρχικά δεδομένα, ώστε να γίνει όσο το δυνατόν ολοκληρωμένη κατανόηση του προς επίλυση προβλήματος.

Ένα **κεντροειδές** (centroid) είναι η φανταστική ή πραγματική θέση που αντιπροσωπεύει το κέντρο του συστάδας. Ένα κεντροειδές είναι ένας φορέας που περιέχει έναν αριθμό για κάθε μεταβλητή, όπου κάθε αριθμός είναι ο μέσος όρος μιας μεταβλητής για τις παρατηρήσεις σε αυτή τη συστάδα. Το κεντροειδές μπορεί να θεωρηθεί ως ο πολυδιάστατος μέσος όρος της συστάδας. Το άθροισμα των τετραγώνων μέσα στο σύμπλεγμα είναι ένα μέτρο της μεταβλητότητας των παρατηρήσεων μέσα σε κάθε συστάδα. Γενικά, μια συστάδα που έχει μικρότερη μέση απόσταση είναι πιο συμπαγής από μια συστάδα που έχει μεγαλύτερη μέση απόσταση.

Συστάδες που έχουν υψηλότερες τιμές παρουσιάζουν μεγαλύτερη μεταβλητότητα των παρατηρήσεων. Κάθε αντικείμενο (σημείο δεδομένων) κατανέμεται σε κάθε μία από τις συστάδες, μειώνοντας το άθροισμα των τετραγώνων της απόστασής τους από το κεντροειδές της συστάδας.

1.2 Τύποι αλγορίθμων ομαδοποίησης

Υπάρχουν πολλές τεχνικές για την στατιστική ομαδοποίηση παρατηρήσεων σε ένα σύνολο δεδομένων, όπου όλες βασίζονται στην προϋπόθεση της χρήσης ενός συνόλου χαρακτηριστικών για τον ορισμό κλάσεων ή κατηγοριών παρατηρήσεων που είναι παρόμοιες, αλλά διαφέρουν μεταξύ των ομάδων. Ο τρόπος με τον οποίο ορίζεται η ομοιότητα μεταξύ των ομάδων και η ανομοιογένεια μεταξύ τους και ο τρόπος με τον οποίο λειτουργεί ο αλγόριθμος ταξινόμησης είναι ο τρόπος με τον οποίο διαφέρουν οι τεχνικές, καθώς και αυτό που καθιστά το καθένα από αυτά ιδιαίτερα κατάλληλο για συγκεκριμένα προβλήματα ή τύπους δεδομένων. Δεδομένου ότι η ομαδοποίηση είναι υποκειμενική, τα μέσα που μπορούν να χρησιμοποιηθούν για την επίτευξη αυτού του σκοπού καθώς και η μεθοδολογία τους είναι άφθονα. Κάθε μεθοδολογία ακολουθεί ένα διαφορετικό σύνολο κανόνων για τον ορισμό της «ομοιότητας» μεταξύ των σημείων δεδομένων.

Βιβλιογραφικά δεν προκύπτει μια και μόνο μέθοδος ομαδοποίησης δεδομένων αλλά πολλές διαφορετικές, όπου η ποιότητά τους κρίνεται μετά από την εφαρμογή τους, εξετάζοντας τις ομάδες που δημιουργούνται. Στην πραγματικότητα, είναι γνωστοί περισσότεροι από 100 αλγόριθμοι ομαδοποίησης αλλά λίγοι από αυτούς χρησιμοποιούνται ευρέως, οι οποίοι είναι (Omran, et al., 2007):

Αλγόριθμοι συνδεσιμότητας: Όπως υποδηλώνει το όνομα, αυτοί οι αλγόριθμοι βασίζονται στην ιδέα ότι τα σημεία δεδομένων πλησιέστερα στο χώρο δεδομένων παρουσιάζουν μεγαλύτερη ομοιότητα μεταξύ τους από τα σημεία δεδομένων που βρίσκονται πιο μακριά. Είναι πολύ εύκολο να ερμηνευτούν αλλά δεν διαθέτουν δυνατότητα κλιμάκωσης για το χειρισμό μεγάλων συνόλων δεδομένων. Παραδείγματα αυτών των μοντέλων είναι ο ιεραρχικός αλγόριθμος ομαδοποίησης και οι παραλλαγές του.

Αλγόριθμοι με κεντροειδή: Αυτοί είναι αλγόριθμοι επαναληπτικής ομαδοποίησης όπου η έννοια της ομοιότητας προέρχεται από την εγγύτητα ενός σημείου δεδομένων στο κεντροειδές των συστάδων. Σε αυτούς τους αλγορίθμους, ο αριθμός των

συστάδων που απαιτούνται στο τέλος πρέπει να αναφερθεί εκ των προτέρων, πράγμα που καθιστά σημαντικό να έχουμε προηγούμενη γνώση του συνόλου δεδομένων. Αυτά τα μοντέλα τρέχουν επαναληπτικά (αναδρομικά) για να βρουν το τοπικό βέλτιστο. Ο αλγόριθμος ομαδοποίησης k-means είναι ένας δημοφιλής αλγόριθμος που εμπίπτει σε αυτήν την κατηγορία. Μια πιο αποτελεσματική έκδοση του k-means ονομάζεται διόρθωση k-μέσων (bisecting k-means).

Αλγόριθμοι κατανομής: Αυτοί βασίζονται στην έννοια του πόσο πιθανό είναι όλα τα σημεία δεδομένων στη συστάδα να ανήκουν στην ίδια κατανομή (π.χ., κανονική, Gaussian). Αυτοί οι αλγόριθμοι συχνά υποφέρουν από υπερφόρτωση. Ένα δημοφιλές παράδειγμα είναι ο αλγόριθμος μεγιστοποίησης προσδοκιών που χρησιμοποιεί πολλαπλές κανονικές κατανομές (McLachlan and Krishnan 1997). Χρησιμοποιείται κυρίως για την εκτίμηση παραμέτρων παρουσία ορισμένων άγνωστων δεδομένων.

Αλγόριθμοι μερικής ομαδοποίησης: Οι αλγόριθμοι μερικής ομαδοποίησης χωρίζουν το σύνολο δεδομένων σε έναν συγκεκριμένο αριθμό συστάδων. Έτσι, προσπαθούν να ελαχιστοποιήσουν ορισμένα κριτήρια (π.χ. μια λειτουργία τετραγώνου σφάλματος) και συνεπώς μπορούν να αντιμετωπίζονται ως προβλήματα βελτιστοποίησης.

Αλγόριθμοι πυκνότητας: Αναζητούν το χώρο δεδομένων για περιοχές με ποικίλη πυκνότητα σημείων δεδομένων στον χώρο δεδομένων. Απομονώνουν διάφορες περιοχές διαφορετικής πυκνότητας και εκχωρούν τα σημεία δεδομένων σε αυτές τις περιοχές στην ίδια συστάδα.

Αλγόριθμοι ασαφούς ομαδοποίησης: Σε αυτούς κάθε αντικείμενο ανήκει σε κάθε συστάδα με βάρος που είναι μεταξύ 0 και 1. Με άλλα λόγια, τα σμήνη δεδομένων θεωρούνται ασαφή σύνολα. Μαθηματικά, ένα ασαφές σύνολο ορίζεται ως ένα σύνολο στο οποίο ένα αντικείμενο συσχετίζεται με οποιοδήποτε σύνολο με βάρος που κυμαίνεται μεταξύ 0 και 1. Στην ασαφή ομαδοποίηση, συνήθως θέτουμε τον πρόσθετο περιορισμό και το άθροισμα των βαρών για κάθε αντικείμενο πρέπει να είναι ίσο σε 1. Ο αλγόριθμος Fuzzy C-means (FCM) είναι η ασαφής εκδοχή του k-means και προτάθηκε από τον Bezdek (1981). Ο FCM βασίζεται σε μια ασαφή επέκταση του κριτηρίου ελάχιστου τετραγώνου σφάλματος.

Ο αλγόριθμος k-means είναι η πιο συνηθισμένη ως προς τη χρήση μέθοδος για την ομαδοποίηση δεδομένων (Nisar, Fard, and Schardl, 2010) και θα προτιμηθεί στη παρούσα έρευνα. Ο αλγόριθμος k-means είναι ένας αλγόριθμος που ομαδοποιεί αντικείμενα με βάση τα χαρακτηριστικά των k ομάδων (MacQueen, 1967). Με βάση τον k-means αλγόριθμο τα χαρακτηριστικά ενός αντικειμένου δημιουργούν ένα

διανυσματικό χώρο. Ο σκοπός του αλγορίθμου είναι να ελαχιστοποιήσει τη συνολική διακύμανση κάθε ομάδας καθώς και να έχει το μικρότερο δυνατό σφάλμα.

1.3 Αλγόριθμος ομαδοποίησης k-means

Ο αλγόριθμος ομαδοποίησης k-means είναι ένας απλός αλγόριθμος μη εποπτευόμενης μάθησης και εξόρυξης δεδομένων που αναπτύχθηκε από τον J. MacQueen το 1967 και στη συνέχεια από τους J.A. Hartigan και M.A. Wong το 1975. Μερικές φορές αναφέρεται ως αλγόριθμος Lloyd's, ιδιαίτερα στους κύκλους της επιστήμης των υπολογιστών, επειδή ο πρότυπος αλγόριθμος προτάθηκε για πρώτη φορά από τον S. Lloyd το 1957. Ο όρος k-means δόθηκε από τον J. McQueen. Χρησιμοποιείται συνήθως στην ιατρική απεικόνιση, τη βιομετρία, ομαδοποίηση αποτελεσμάτων αναζήτησης, τμηματοποίηση εικόνας, γεωγραφικές περιοχές και άλλους συναφείς τομείς. Ο αλγόριθμος ομαδοποίησης k-means επιδιώκει ένα βέλτιστο διαχωρισμό των δεδομένων ελαχιστοποιώντας το άθροισμα του κριτηρίου του τετραγωνικού σφάλματος (Sum of the Squared Error, SSE) (Hamerly and Elkan 2002).

Το άθροισμα του τετραγωνικού σφάλματος (SSE) ορίζεται ως εξής:

$$SSE = \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} ||\mathbf{m}_i - \mathbf{x}||^2$$

Όπου, m_i το κεντροειδές της i-της συστάδας με n_i αντικείμενα που ορίζεται από τη σχέση:

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in C_i} \mathbf{x}$$

Σε αυτή την προσέγγιση, τα αντικείμενα δεδομένων ταξινομούνται σε αριθμό k συστάδων στις οποίες κάθε αντικείμενο ανήκει στη συστάδα με τον πλησιέστερο μέσο όρο (ίδιο κεντροειδές). Συνήθως, οι αλγόριθμοι βγάζουν συμπεράσματα από σύνολα δεδομένων χρησιμοποιώντας μόνο φορείς εισόδου, χωρίς αναφορά σε γνωστά ή επισημασμένα αποτελέσματα. Η προσέγγιση που ακολουθεί η ομαδοποίηση k-means για να λύσει το πρόβλημα ονομάζεται προσδοκία-μεγιστοποίηση.

Ο στόχος του αλγορίθμου k-means είναι απλός: ομαδοποίηση παρόμοιων δεδομένων και συγκέντρωση των υποκείμενων μοτίβων. Για να επιτευχθεί αυτός ο στόχος, ο αλγόριθμος k-means αναζητά έναν σταθερό αριθμό k συστάδων σε ένα σύνολο

δεδομένων. Με άλλα λόγια, ο αλγόριθμος k-means προσδιορίζει τον αριθμό k των κεντροειδών και στη συνέχεια κατανέμει κάθε σημείο δεδομένων στην πλησιέστερη συστάδα. Το μέσο στον αλγόριθμο k-means αναφέρεται στο μέσο όρο των δεδομένων.

Μαθηματικά, αυτός ο αλγόριθμος στοχεύει στην ελαχιστοποίηση της εξής συνάρτησης:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

όπου,

$x_i^{(j)}$ = Αντικείμενο (σημείο δεδομένων)

c_j = Κεντροειδές της συστάδας

n = Αριθμός αντικειμένων (σημείων δεδομένων)

k = Αριθμός συστάδας

$\left\| x_i^{(j)} - c_j \right\|^2$ = Απόσταση μεταξύ σημείου δεδομένων $x_i^{(j)}$ και του κεντροειδούς της συστάδας c_j.

Ο αλγόριθμος k-means είναι μια ντετερμινιστική προσέγγιση που επιλύει ευριστικά το πρόβλημα βελτιστοποίησης της πιο πάνω συνάρτησης με την εξεύρεση τοπικού ελάχιστου.

1.4 Πώς λειτουργεί ο αλγόριθμος k-means

Ο τρόπος με τον οποίο λειτουργεί ο αλγόριθμος k-means που θα χρησιμοποιηθεί σε αυτή την εργασία είναι ο εξής:

1. Καθορισμός του αριθμού των k συστάδων.
2. Αρχικοποίηση των κεντροειδών επιλέγοντας τυχαία τα σημεία δεδομένων των k συστάδων για τα κεντροειδή.
3. Υπολογισμός του αθροίσματος της τετραγωνικής απόστασης μεταξύ των σημείων δεδομένων και όλων των κεντροειδών.
4. Τοποθέτηση κάθε σημείου δεδομένων στο πλησιέστερο κεντροειδές κάθε συστάδας.

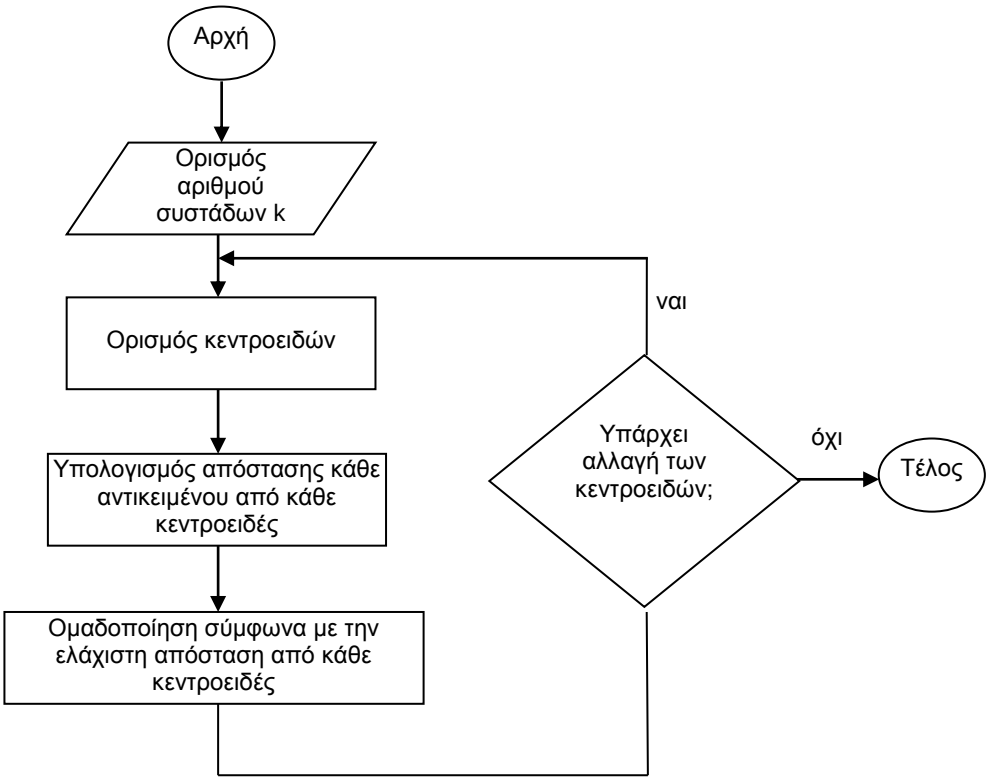
5. Υπολογισμός των κεντροειδών για όλες τις συστάδες λαμβάνοντας τον μέσο όρο όλων των σημείων δεδομένων που ανήκουν σε κάθε συστάδα.

Επανάληψη των βημάτων 3, 4 και 5 μέχρις ότου:

- τα κεντροειδή έχουν σταθεροποιηθεί - δεν υπάρχει καμία αλλαγή στις τιμές τους, επειδή η ομαδοποίηση έχει επιτευχθεί,
- ή έχει επιτευχθεί ο καθορισμένος αριθμός επαναλήψεων.

Γενικά, Ο αλγόριθμος επιλέγει αυθαίρετα k σημεία ως αρχικά κεντροειδή (c) κάθε συστάδας (k μέσα). Κάθε σημείο του συνόλου δεδομένων αντιστοιχεί στη συστάδα, με βάση την απόσταση (συνήθως την Ευκλείδεια απόσταση) μεταξύ κάθε αντικειμένου και του κεντροειδούς κάθε συστάδας. Κάθε κεντροειδές συστάδας αναπροσαρμόζεται ως ο μέσος όρος των αντικειμένων σε αυτή τη συστάδα.

Στο σχήμα 1.4 φαίνεται το λογικό διάγραμμα του αλγορίθμου k-means.



Σχήμα 1.4: *Λογικό διάγραμμα του αλγορίθμου k-means*

Δεδομένου ότι οι αλγόριθμοι ομαδοποίησης, συμπεριλαμβανομένων του k-means, χρησιμοποιούν μετρήσεις βάσει αποστάσεων για να προσδιορίσουν την ομοιότητα μεταξύ των σημείων δεδομένων, συνιστάται η τυποποίηση των δεδομένων ώστε να

έχουν μέση τιμή μηδέν και τυπική απόκλιση ένα, δεδομένου ότι σχεδόν πάντα τα χαρακτηριστικά σε οποιοδήποτε σύνολο δεδομένων έχουν διαφορετικές μονάδες μετρήσεων.

Με δεδομένη την επαναληπτική φύση του k-means και την τυχαία αρχικοποίηση των κεντροειδών στην αρχή του αλγορίθμου, διαφορετικές αρχικοποιήσεις μπορούν να οδηγήσουν σε διαφορετικές συστάδες, αφού ο αλγόριθμος k-means μπορεί να κολλήσει σε τοπικό βέλτιστο και να μην συγκλίνει σε γενικό βέλτιστο. Επομένως, συνιστάται να εκτελείται ο αλγόριθμος χρησιμοποιώντας διαφορετικές αρχικοποιήσεις των κεντροειδών και να επιλέγονται τα αποτελέσματα της εκτέλεσης που δίνουν το μικρότερο άθροισμα της Ευκλείδειας απόστασης.

Ένα από τα κύρια μειονεκτήματα της ομαδοποίησης k-means είναι το γεγονός ότι πρέπει να καθοριστεί ο αριθμός των συστάδων ως είσοδος στον αλγόριθμο. Όπως έχει σχεδιαστεί, ο αλγόριθμος δεν είναι σε θέση να καθορίσει τον κατάλληλο αριθμό συστάδων και πρέπει να οριστεί εκ των προτέρων.

Είναι συχνά μια καλή ιδέα να πειραματιζόμαστε με διαφορετικές τιμές του k για να προσδιορίσουμε την τιμή των συστάδων που ταιριάζει καλύτερα στα δεδομένα.

1.5 Πλεονεκτήματα και μειονεκτήματα του αλγορίθμου k-means

Ο αλγόριθμος k-means έχει τα ακόλουθα κύρια πλεονεκτήματα (Turi 2001): είναι πολύ εύκολο να υλοποιηθεί και η πολυπλοκότητα του χρόνου είναι O (nkl), όπου n είναι ο αριθμός των αντικειμένων, k είναι ο αριθμός των συστάδων και το l είναι αριθμός των επαναλήψεων (Anderberg 1973), καθιστώντας τον κατάλληλο για πολύ μεγάλα σύνολα δεδομένων. Τα αντικείμενα μοιράζονται σε συστάδες με αυτόματο τρόπο. Είναι αρκετά γρήγορος, τουλάχιστον σε σχέση με τις ιεραρχικές μεθόδους. Ο χρόνος εκτέλεσης του αλγορίθμου εξαρτάται γραμμικά από τα στοιχεία του προβλήματος, όπως το πλήθος των συστάδων k, το πλήθος των αντικειμένων και το πλήθος των επαναλήψεων. Για τον λόγο αυτό, είναι πιο κατάλληλος από άλλες μεθόδους για την ομαδοποίηση μεγάλων συνόλων αντικειμένων.

Ωστόσο, ο αλγόριθμος k-means έχει τα ακόλουθα μειονεκτήματα (Davies 1997): Ο στόχος του αλγορίθμου k-means είναι η ομαδοποίηση των σημείων δεδομένων σε ξεχωριστές συστάδες που δεν αλληλεπικαλύπτονται. Κάνει πολύ καλή δουλειά όταν οι συστάδες έχουν ένα είδος σφαιρικών σχημάτων (με ακτίνα ίση με την απόσταση μεταξύ του κεντροειδούς και του πλέον απομακρυσμένου σημείου δεδομένων). Δεν

αποδίδει καλά καθώς τα γεωμετρικά σχήματα των συστάδων αποκλίνουν από τα σφαιρικά σχήματα. Προσπαθεί πάντα να κατασκευάσει ένα σφαιρικό σχήμα γύρω από το κάθε κεντροειδές. Αυτό σημαίνει ότι, αν οι συστάδες έχουν περίπλοκα γεωμετρικά σχήματα ή με πολύ διαφορετικά μεγέθη, ο αλγόριθμος k-means δεν αποδίδει καλά.

Επίσης, ο αλγόριθμος k-means δεν επιτρέπει στα σημεία δεδομένων που απέχουν πολύ το ένα από το άλλο να μοιράζονται την ίδια συστάδα, παρόλο που προφανώς ανήκουν στην ίδια συστάδα.

Αν δημιουργήσουμε δεδομένα από πολυμεταβλητές κανονικές κατανομές με διαφορετικά μέσα και τυπικές αποκλίσεις θα έχουμε ομάδες δεδομένων όπου κάθε συστάδα θα έχει δημιουργηθεί από διαφορετική κανονική κατανομή πολλαπλών μεταβλητών (διαφορετική μέση τιμή και τυπική απόκλιση). Μια συστάδα θα έχει πολύ περισσότερα σημεία δεδομένων από τις άλλες.

Επίσης, ο αλγόριθμος είναι εξαρτώμενος από τα δεδομένα, είναι άπληστος αλγόριθμος που εξαρτάται από τις αρχικές συνθήκες, οι οποίες μπορεί να προκαλέσουν σύγκλιση του αλγορίθμου σε υποβέλτιστες λύσεις. Το τελικό αποτέλεσμα εξαρτάται σε σημαντικό βαθμό από την επιλογή των αρχικών κέντρων. Επιλογή διαφορετικών κέντρων μπορεί να οδηγήσει σε σημαντικά διαφορετικές συστάδες και είναι πολύ ευαίσθητος στην ύπαρξη αντικειμένων με ακραίες τιμές.

Επιπλέον, δεν μαθαίνει επίσης τον αριθμό των ομαδοποιήσεων από τα δεδομένα και απαιτεί τον προκαθορισμό τους και απαιτείται κλιμάκωση/τυποποίηση των δεδομένων κατά την εφαρμογή του αλγόριθμου k-means.

Τέλος, ο αλγόριθμος k-means ενδέχεται να ομαδοποιεί τα δεδομένα ακόμη και αν δεν μπορούν να ομαδοποιηθούν, όπως τα δεδομένα που προέρχονται από ομοιόμορφες κατανομές.

1.6 Μέθοδοι αξιολόγησης αλγορίθμων ομαδοποίησης

Σε αντίθεση με την εποπτευόμενη μάθηση, όπου έχουμε τη βασική αλήθεια για να αξιολογήσουμε την απόδοση του μοντέλου, η ανάλυση ομαδοποίησης δεν έχει μια σταθερή μετρική αξιολόγησης που μπορούμε να χρησιμοποιήσουμε για να αξιολογήσουμε το αποτέλεσμα διαφορετικών αλγορίθμων ομαδοποίησης. Επιπλέον,

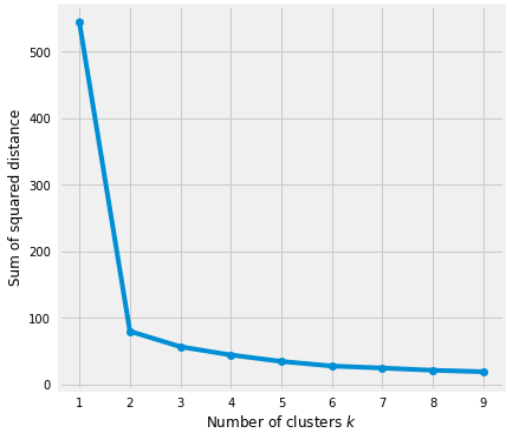
δεδομένου ότι τα k-means απαιτούν το k ως είσοδο και δεν το μαθαίνουν από τα δεδομένα, δεν υπάρχει σωστή απάντηση όσον αφορά τον αριθμό των συστάδων που πρέπει να έχουμε σε οποιοδήποτε πρόβλημα. Στη μεθοδολογία πρόβλεψης της συστάδας, μπορούμε να αξιολογήσουμε πόσο καλά εκτελούνται τα μοντέλα με βάση διαφορετικές συστάδες από τη χρήση συστάδων στη μεταγενέστερη μοντελοποίηση.

Οι βασικές μέθοδοι που υπολογίζουν το βέλτιστο αριθμό των συστάδων είναι:

- η μέθοδος Elbow και
- η ανάλυση Silhouette
- οι μέθοδοι σύμφωνα με τον κανόνα και διασταυρωμένη επικύρωση.

1.6.1 Η μέθοδος Elbow

Η μέθοδος Elbow προτάθηκε από τον Andrew (2012) μας δίνει το βέλτιστο αριθμό k συστάδων που βασίζεται στο άθροισμα της τετραγωνικής απόστασης (Sum of Squared Distance, SSE) μεταξύ των σημείων των δεδομένων και των κεντροειδών των συστάδων τους. Επιλέγουμε το k στο σημείο όπου η SSE αρχίζει να σταθεροποιείται. Το παρακάτω γράφημα δείχνει ότι το k=2 είναι μια καλή επιλογή. Μερικές φορές είναι δύσκολο να υπολογίσουμε έναν βέλτιστο αριθμό συστάδων, επειδή η καμπύλη μειώνεται μονότονα και μπορεί να μην έχει ένα προφανές σημείο όπου η καμπύλη αρχίζει να σταθεροποιείται.



Σχήμα 2.6.1: *Η μέθοδος Elbow για την εύρεση του κατάλληλου k*

1.6.2 Η ανάλυση Silhouette

Προτάθηκε από τους Kaufman και Rousseeuw (1990). Η ανάλυση Silhouette μπορεί να χρησιμοποιηθεί για τον προσδιορισμό του βαθμού διαχωρισμού μεταξύ των συστάδων. Για κάθε δείγμα απαιτούνται τα εξής:

- Υπολογισμός της μέσης απόσταση από όλα τα σημεία δεδομένων στη ίδια συστάδα (a^i).
- Υπολογισμός της μέσης απόστασης από όλα τα σημεία δεδομένων στην πλησιέστερη συστάδα (b^i).
- Υπολογισμός του συντελεστή:

$$\frac{b^i - a^i}{max(a^i, b^i)}$$

Ο συντελεστής αυτός μπορεί να λάβει τιμές στο διάστημα [-1, 1].

Αν είναι 0, τότε το δείγμα είναι πολύ κοντά στις γειτονικές συστάδες.

Αν είναι 1, τότε το δείγμα είναι πολύ μακριά από τις γειτονικές συστάδες.

Αν είναι -1, τότε το δείγμα αντιστοιχεί σε λανθασμένες συστάδες.

Επομένως, θέλουμε οι συντελεστές να είναι όσο το δυνατόν μεγαλύτεροι και κοντά στο 1 για να έχουμε καλές συστάδες.

1.6.3 Άλλοι μέθοδοι

α) Σύμφωνα με τον κανόνα

Είναι μια απλή μέθοδος. Αυτή η μέθοδος μπορεί να εφαρμοστεί σε οποιοδήποτε τύπο συνόλου δεδομένων. Η τιμή του k υπολογίζεται από την εξής σχέση:

$$k \approx \sqrt{\frac{n}{2}}$$

Όπου n είναι ο αριθμός των αντικειμένων (σημεία δεδομένων).

β) Διασταυρωμένη επικύρωση

Η διασταυρούμενη επικύρωση είναι μια άλλη προσέγγιση για την εκτίμηση του αριθμού των συσπειρώσεων που προτείνει ο Smyth (1996). Βασίζεται στη σταθερότητα των συμπλεγμάτων. Αυτή η μέθοδος χωρίζει τα δεδομένα σε δύο ή περισσότερα τμήματα. Ένα μέρος χρησιμοποιείται για την ομαδοποίηση και το άλλο μέρος χρησιμοποιείται για επικύρωση. Η ιδέα πίσω από τη σταθερότητα της συσσωμάτωσης είναι ότι ένας "καλός" αλγόριθμος τείνει να παράγει επανειλημμένα παρόμοια συσσωμάτωση στα δεδομένα που προέρχονται από την ίδια πηγή. Με άλλα λόγια, ο αλγόριθμος είναι σταθερός σε σχέση με την τυχαία εισαγωγή.

ΚΕΦΑΛΑΙΟ 2 : Χρήση του αλγορίθμου k-means σε γεωγραφικές περιοχές

2.1 Σκοπός ομαδοποίησης γεωγραφικών περιοχών

Η χωρική επιστήμη παραμένει ουσιαστική συνιστώσα της σύγχρονης ανθρώπινης γεωγραφίας. Είναι πολύ ποσοτική, αλλά οι επίσημες (γεωμετρικές) θεωρίες θέσης που βασίζονται σε μια ενιαία αιτιώδη μεταβλητή (χώρος) έχουν εγκαταλειφθεί σε μεγάλο βαθμό: η αναζήτηση χωρικής τάξης δεν προβλέπει την ανακάλυψη τακτικών δομών ούτε επιδιώκει καθολικούς νόμους χωρικής συμπεριφοράς (Johnston, 2001).

Ο Sayer (1992) επέδειξε μια σημαντική διάκριση μεταξύ εκτεταμένης και εντατικής έρευνας: η πρώτη αναζητά εμπειρικές κανονικότητες, ενώ η τελευταία διερευνά τις αιτιώδεις αλυσίδες που ευθύνονται για συγκεκριμένα αποτελέσματα. Η χωρική επιστήμη απαιτεί εκτεταμένη έρευνα, παραβλέποντας τις εμπειρικές γενικεύσεις, για να εντοπίζει σημαντικά χαρακτηριστικά και τάσεις στο σύνολο των στατιστικών δεδομένων που χαρακτηρίζουν τις σύγχρονες κοινωνίες.

Το πρόβλημα της ομαδοποίησης γεωγραφικών περιοχών γενικά είναι ένα θεμελιώδες πρόβλημα, το οποίο περιγράφει πώς συλλέγουμε τα δεδομένα για την ταξινόμηση προτύπων και την παραγωγή γνώσης. Έτσι, η εξόρυξη δεδομένων είναι το εργαλείο για την ανάλυση δεδομένων των ομαδοποιημένων γεωγραφικών περιοχών και την εξαγωγή των πληροφοριών από αυτά (ανακάλυψη γνώσεων και την εξόρυξη δεδομένων). Η αναγνώριση των γεωγραφικών ομάδων μπορεί να αποτελέσει σημαντικό στάδιο μιας γεωγραφικής ανάλυσης. Σε ένα γεωγραφικό σύνολο δεδομένων, το σύστημα ασχολείται με δεδομένα σχετικά με τη γεωγραφία και τη γεωγραφική κλίμακα των πληροφοριών μέτρησης. Μία από τις πρώτες σκέψεις είναι τι πρέπει να επιλέξουμε για τα γεωγραφικά δομικά στοιχεία των προτεινόμενων περιοχών και ποια από αυτά θα συγκρίνουμε για να καταλήξουμε σε συμπεράσματα. Με την αύξηση της διαθεσιμότητας και της πολυπλοκότητας των συστημάτων γεωγραφικών πληροφοριών

(GIS), η χρήση των γεωγραφικών περιοχών για χωροταξικά δομικά στοιχεία έχει καταστεί ευκολότερη στην εφαρμογή τους (Singleton, et al., 2016).

Στην περίπτωση της ανάλυσης χωρικών δεδομένων, υπάρχει ένα υποσύνολο μεθόδων που παρουσιάζουν ιδιαίτερο ενδιαφέρον για πολλές κοινές περιπτώσεις στην Γεωγραφική Επιστήμη Δεδομένων. Αυτές είναι οι λεγόμενες τεχνικές περιφερειοποίησης (regionalization). Οι μέθοδοι περιφερειοποίησης μπορούν να λάβουν επίσης πολλές μορφές, αλλά στο επίκεντρό τους όλες περιλαμβάνουν την στατιστική ομαδοποίηση παρατηρήσεων με τον πρόσθετο περιορισμό ότι οι παρατηρήσεις πρέπει να είναι σε γεωγραφικά γειτονικές περιοχές που βρίσκονται στην ίδια κατηγορία.

Τα χωρικά δεδομένα είναι το σύνολο των ερμηνειών μας για γεωγραφικά φαινόμενα. Σε ψηφιακή μορφή, τα δεδομένα είναι οι κύριες πληροφορίες που απαιτούνται από τα συστήματα γεωγραφικών πληροφοριών, τα εργαλεία λογισμικού που χρησιμοποιούνται για την ανάλυση χωρικών δεδομένων. Τα χωρικά δεδομένα χρησιμοποιούνται ως όρος που περιλαμβάνει σύνολα δεδομένων γενικού σκοπού, όπως ψηφιακά χαρτογραφικά δεδομένα, εικόνες τηλεπισκόπησης και περιγραφές απογραφών, καθώς και πιο εξειδικευμένα σύνολα δεδομένων όπως σεισμικά προφίλ, κατανομή λειψάνων σε ένα αρχαιολογικού χώρου ή στατιστικών μετανάστευσης (Guptill, 2001).

Στην κοινωνικο-οικονομική σφαίρα, τα σημεία, οι γραμμές και οι μονάδες της περιοχής είναι οι θεμελιώδεις οντότητες για την εκπροσώπηση των χωρικών φαινομένων. Αυτή η μορφή χωρικής αναφοράς είναι επίσης ένα χαρακτηριστικό γνώρισμα του GIS (π.χ., Γεωγραφικά Πληροφοριακά Συστήματα, Υποδομή Χωρικών Δεδομένων). Μπορούν να διακριθούν τρεις ευρείες κατηγορίες χωρικών δεδομένων (Fischer, 2015):

- δεδομένα αντικειμένων όπου τα αντικείμενα είναι είτε σημεία (μοτίβα χωρικών σημείων ή δεδομένα τοποθεσίας, δηλαδή θέσεις σημείων στα οποία έχουν συμβεί γεγονότα ενδιαφέροντος) ή περιοχές (δεδομένα περιοχής ή πλέγματος, που ορίζονται ως διακριτές μεταβολές χαρακτηριστικών στο χώρο),
- δεδομένα εδάφους (που ονομάζονται επίσης γεωστατικά ή χωρικά συνεχή δεδομένα), δηλαδή παρατηρήσεις που σχετίζονται με συνεχή μεταβολή στο διάστημα, δεδομένες τιμές σε σταθερά σημεία δειγματοληψίας και
- δεδομένα χωρικής αλληλεπίδρασης (μερικές φορές αποκαλούμενα στοιχεία σύνδεσης ή ροής) που συνίστανται σε μετρήσεις κάθε μία από τις οποίες συνδέεται με μια σύνδεση ή ζεύγος θέσεων που αντιπροσωπεύουν σημείο ή περιοχές.

Η ανάλυση χωρικών δεδομένων επικεντρώνεται στην ανίχνευση μοτίβων και στη διερεύνηση και μοντελοποίηση σχέσεων μεταξύ τέτοιων μοτίβων, προκειμένου να κατανοηθούν οι διαδικασίες που είναι υπεύθυνες για τα παρατηρούμενα πρότυπα. Με αυτόν τον τρόπο, η ανάλυση χωρικών δεδομένων τονίζει το ρόλο του χώρου ως δυνητικά σημαντικό ερμηνευτή των κοινωνικοοικονομικών συστημάτων και προσπαθεί να ενισχύσει την κατανόηση της εργασίας και της εκπροσώπησης του χώρου, των χωρικών προτύπων και των διαδικασιών (Fischer, 2015).

Σύμφωνα με τον Haining (2015), η ανάλυση και η μοντελοποίηση χωρικών δεδομένων παρουσιάζουν μια σειρά προβλημάτων. Λύσεις σε πολλές από αυτές είναι προφανείς, άλλες απαιτούν εξαιρετική προσπάθεια για την επίλυσή τους. Τα δεδομένα ασκούν δύναμη που μπορεί να οδηγήσει σε παρερμηνείες και άχρηστα αποτελέσματα. Η ποσοτική ανάλυση εξαρτάται σε μεγάλο βαθμό από την ποιότητα των δεδομένων. Τα καλά δεδομένα είναι αξιόπιστα, περιέχουν λίγα ή καθόλου λάθη και μπορούν να χρησιμοποιηθούν με σιγουριά. Δυστυχώς, σχεδόν όλα τα χωρικά δεδομένα είναι ελαττωματικά σε κάποιο βαθμό (π.χ., βλέπε χωρικά δεδομένα). Μπορούν να προκύψουν σφάλματα κατά τη μέτρηση τόσο των ιδιοτήτων των θέσεων όσο και των χαρακτηριστικών, αλλά μπορούν επίσης να συσχετιστούν με τις μηχανογραφικές διαδικασίες που είναι υπεύθυνες για την αποθήκευση, την ανάκτηση και τον χειρισμό χωρικών δεδομένων. Η λύση στο πρόβλημα της ποιότητας των δεδομένων είναι να λάβουν τα απαραίτητα μέτρα για να αποφευχθούν τα ελαττωματικά δεδομένα που καθορίζουν τα αποτελέσματα της έρευνας. Η χωρική δειγματοληψία είναι η διαδικασία της συλλογής δεδομένων σύμφωνα με ορισμένες καθορισμένες ομάδες κανόνων και στη συνέχεια η χρήση των δεδομένων για την εξαγωγή συμπερασμάτων σχετικά με τον πληθυσμό από τον οποίο έχει ληφθεί το δείγμα.

Τα μοντέλα ανθρωπο-γεωγραφίας χρησιμοποιούνται για στατιστική έρευνα, βελτιστοποίηση, ή προσομοίωση και πρόβλεψη. Η στατιστική έρευνα είναι ίσως η πιο συνηθισμένη, συχνά χρησιμοποιούμενη μέθοδος για έλεγχο υποθέσεων. Για παράδειγμα, οι γεωγράφοι έχουν χρησιμοποιήσει εξειδικευμένα μοντέλα παλινδρόμησης για τη διερεύνηση θεμάτων όπως η αποκέντρωση των πόλεων, η περιφερειακή ανισότητα εισοδήματος, τα ιστορικά πρότυπα ψηφοφορίας και οι κοινωνικοοικονομικές πτυχές της αποψίλωσης. Αυτό που διακρίνει αυτούς τους τύπους ερευνών από εκείνους άλλων κοινωνικών επιστημών είναι ότι οι χωρικές διαστάσεις αναλαμβάνουν κεντρική εστίαση.

Τα μοντέλα και οι μέθοδοι διερευνητικής χωρικής ανάλυσης δεδομένων έχουν προσελκύσει μεγάλη προσοχή (Getis, 1999).

Τα μοντέλα βελτιστοποίησης χρησιμοποιούνται για κανονιστικούς σκοπούς για τον προσδιορισμό της βέλτιστης λύσης για ένα δεδομένο πρόβλημα. Όλα τα μοντέλα βελτιστοποίησης έχουν μια ποσότητα που πρέπει να μεγιστοποιηθεί (ή να ελαχιστοποιηθεί), όπως αντικατοπτρίζεται σε μια αντικειμενική λειτουργία.

Δύο ευρείες κατηγορίες στατιστικής μοντελοποίησης στη γεωγραφία χρησιμοποιούνται. Η πρώτη προσέγγιση είναι διερευνητική, όπου εστιάζεται στην αναγνώριση του χωροταξικού προτύπου, στις δοκιμές για τη χωρική συσσωμάτωση και στις αναχωρήσεις από τυχαία δειγματοληψία. Η δεύτερη προσέγγιση στη στατιστική μοντελοποίηση είναι γνωστή ως επιβεβαιωτική ανάλυση χωρικών δεδομένων (Ray, 2015).

Η βασική ιδέα της στατιστικής ομαδοποίησης είναι η συνοπτική παρουσίαση των πληροφοριών που περιέχονται σε διάφορες μεταβλητές, δημιουργώντας ένα σχετικά μικρό αριθμό κατηγοριών. Κάθε παρατήρηση στο σύνολο δεδομένων στη συνέχεια αντιστοιχεί σε μία και μόνο σε μία κατηγορία, ανάλογα με τις τιμές της για τις μεταβλητές που είχαν αρχικά ληφθεί υπόψη στην ταξινόμηση. Αν γίνει σωστά, τότε μειώνεται η πολυπλοκότητα ενός πολυδιάστατου προβλήματος διατηρώντας παράλληλα όλες τις σημαντικές πληροφορίες που περιέχονται στο αρχικό σύνολο δεδομένων. Αυτό συμβαίνει επειδή, αφού ταξινομηθεί, ο αναλυτής πρέπει να εξετάσει μόνο σε ποια κατηγορία εμπίπτει κάθε παρατήρηση, αντί να εξετάσει τις πολλαπλές τιμές που σχετίζονται με καθεμία από τις μεταβλητές και να προσπαθήσει να καταλάβει πώς να τις συγκεντρώσει με μια συνεκτική έννοια. Όταν η ομαδοποίηση πραγματοποιείται σε παρατηρήσεις που αντιπροσωπεύουν περιοχές, η τεχνική ονομάζεται συχνά γεω-δημογραφική ανάλυση (Arribas-Bel, 2019).

Τα γεω-δημογραφικά συστήματα εκτιμούν τα πιο πιθανά χαρακτηριστικά των ανθρώπων με βάση το συγκεντρωτικό προφίλ όλων των ανθρώπων που ζουν σε μια περιοχή. Οι γεω-δημογραφικοί δείκτες είναι σύνθετα μέτρα που περιγράφουν την κοινωνικο-χωρική δομή μικρών γεωγραφικών περιοχών (Harris et al., 2005). Μια γεω-δημογραφική ταξινόμηση παρέχει ένα σύνολο κατηγορηματικών περιλήψεων των κοινωνικο-οικονομικών χαρακτηριστικών των μικρών γεωγραφικών περιοχών. Πολλές ταξινομήσεις, δημιουργούνται εξ ολοκλήρου από δεδομένα που προέρχονται από μια ενιαία δεκαετή απογραφή του πληθυσμού. Αυτές οι ταξινομήσεις συχνά επικρίνονται ότι καθίστανται λιγότερο χρήσιμες με την πάροδο του χρόνου εξαιτίας της μεταβαλλόμενης σύνθεσης μικρών γεωγραφικών περιοχών. Τα γεωδημογραφικά έργα αποδίδονται βασικά με την ανάθεση συστάδων δεδομένων υψηλής και χαμηλής τάξης (ονομαστικής κλίμακας, κλίμακας διαστήματος και κλίμακας αναλογίας) σε διακριτές κατηγορίες.

Υπάρχουν πολλές υποκειμενικότητες που είναι εγγενείς σε αυτή τη διαδικασία ιδίως, ότι ο χώρος είναι αποτελείται από ομάδες παρόμοιων περιοχών, δεν έχουν αυθαίρετα αναλυθεί και ότι υπάρχει ένας "βέλτιστος" αριθμός ομάδων που είναι πιο κατάλληλος για μια πλήρη ανάλυση των δεδομένων (Singleton, et al., 2016).

2.2 Γενικά συμπεράσματα

Ο στόχος της ομαδοποίησης είναι να προσδιοριστεί η εγγενής ομαδοποίηση σε ένα σύνολο μη επισημασμένων δεδομένων. Αλλά πώς θα αποφασίσουμε τι συνιστά μια καλή ομαδοποίηση; Μπορεί να αποδειχθεί ότι δεν υπάρχει απόλυτο κριτήριο που θα ήταν ανεξάρτητο από τον τελικό στόχο της ομαδοποίησης. Κατά συνέπεια, ο χρήστης πρέπει να παρέχει αυτό το κριτήριο, έτσι ώστε το αποτέλεσμα της ομαδοποίησης να ταιριάζει στις ανάγκες του. Ένα κριτήριο ομοιότητας είναι η απόσταση, όπου δύο ή περισσότερα αντικείμενα ανήκουν στην ίδια συστάδα αν είναι κοντά, σύμφωνα με μια δεδομένη απόσταση. Αυτό ονομάζεται ομαδοποίηση βασισμένη στην απόσταση.

Ένα άλλο είδος ομαδοποίησης είναι η εννοιολογική ομαδοποίηση, όπου δύο ή περισσότερα αντικείμενα ανήκουν στην ίδια συστάδα αν αυτή ορίζει μια έννοια κοινή για όλα αυτά τα αντικείμενα. Με άλλα λόγια, τα αντικείμενα ομαδοποιούνται σύμφωνα με την προσαρμογή τους στις περιγραφικές έννοιες, όχι σύμφωνα με τα μέτρα απλής ομοιότητας.

Με δεδομένο ένα σύνολο δεδομένων, η ομαδοποίηση ανιχνεύει συγκεκριμένο αριθμό συστάδων που είναι εσωτερικά ομοιογενείς και τα μέλη διαφορετικών συστάδων έχουν μέγιστη ανομοιογένεια. Πολλοί αλγόριθμοι έχουν σχεδιαστεί για την εκτέλεση χωρικής ομαδοποίησης σε σχέση με τις χωρικές διαστάσεις των αντικειμένων. Σε πολλές γεωγραφικές εργασίες εντοπισμού της γνώσης, προτιμάται η εφαρμογή χωρικής ανάλυσης συστάδων εξαιτίας της ικανότητάς της να εξάγει δομές απευθείας από τα δεδομένα χωρίς να χρησιμοποιεί οποιαδήποτε εκ των προτέρων γνωστές ιεραρχίες χωρικών εννοιών (Estivill-Castro και Houle 2001). Στην πραγματικότητα, ένας αλγόριθμος χωρικής ομαδοποίησης επιδιώκει έναν καθορισμένο αριθμό αντιπροσωπευτικών σημείων στον χώρο. Αυτά τα σημεία οδηγούν στα σμήνη και τα μέλη τους.

Η εξόρυξη χωρικών δεδομένων έχει γίνει ένα δημοφιλές και ισχυρό μέσο για τη σύνθετη ανάλυση τεράστιων ποσοτήτων γεωγραφικών δεδομένων. Η εξόρυξη χωρικών δεδομένων ορίζεται ως η αυτόματη διαδικασία ανεύρεσης ενδιαφέρουσας και

έμμεσης γνώσης από μεγάλες ποσότητες χωρικών δεδομένων. Ο κοινός μεγάλος όγκος γεω-χωρικών βάσεων δεδομένων έχει μετατρέψει τις πτυχές της αποδοτικότητας και της κλιμάκωσης στις κύριες ανησυχίες στο σχεδιασμό και την ανάπτυξη αλγορίθμων εξόρυξης χωρικών δεδομένων (Javadi & Shahrab, 2014).

Η ανίχνευση χωρικών συστάδων καθοδηγείται από γεωγραφικές πληροφορίες που αντιστοιχούν στη θέση των δραστηριοτήτων, απαιτώντας κατάλληλη και ουσιαστική επεξεργασία των χωρικών σχέσεων σε συνδυασμό με τα παρατηρούμενα χαρακτηριστικά της τοποθεσίας και των γεγονότων. Μέχρι σήμερα, αυτό σήμαινε τη χρήση ειδικών μέτρων και τεχνικών για τη δομή και την κατανόηση της απόστασης, των γειτόνων, της συνοχής, της ακανόνιστης γεωγραφικής μορφολογίας κ.ο.κ. Δυστυχώς, όλες οι προσεγγίσεις χωρικής ομαδοποίησης, ανεξάρτητα από τη θεωρητική τους υποστήριξη, τη στατιστική θεμελίωση ή τις μαθηματικές προδιαγραφές, έχουν περιορισμούς στην ακρίβεια, την ευαισθησία και την υπολογιστική προσπάθεια που απαιτείται για την αναγνώριση των συστάδων (Grubesic et al., 2014).

Τα γεωγραφικά δεδομένα των χωρικών συστάδων δεν είναι τυχαία, λόγω του φαινομένου της χωρικής αυτοσυσχέτισης (spatial autocorrelation), η οποία έχει αποτελέσει την αφορμή για την επέκταση των κλασικών στατιστικών μεθόδων σε ειδικές μεθόδους ανάλυσης γεωγραφικών φαινομένων. Η χωρική αυτοσυσχέτιση είναι ένας όρος ο οποίος υποδηλώνει το προφανές γεγονός ότι τα δεδομένα από τοποθεσίες που είναι κοντά στον γεωγραφικό χώρο είναι πιθανότερο να έχουν περισσότερες ομοιότητες μεταξύ τους σε σχέση με δεδομένα τα οποία προέρχονται από μακρινές τοποθεσίες. Αν η χωρική αυτοσυσχέτιση δεν ήταν ο κανόνας, δεν θα υπήρχε ενδιαφέρον για την ανάλυση των γεωγραφικών φαινομένων (Ηλιοπούλου, 2017).

Η γεω-δημογραφική ανάλυση, η οποία είναι ένα από τα πιο ενδιαφέροντα διεπιστημονικά ερευνητικά θέματα μεταξύ Γεωγραφικών Πληροφοριακών Συστημάτων και Εξόρυξης Δεδομένων, διαδραματίζει πολύ σημαντικό ρόλο στις αποφάσεις των πολιτικών, στη μετανάστευση πληθυσμών και στη διανομή υπηρεσιών. Μεταξύ ορισμένων υπολογιστικών μεθόδων που χρησιμοποιούνται για αυτό το πρόβλημα, η ομαδοποίηση σε συστάδες είναι η πιο δημοφιλής επειδή έχει πολλά πλεονεκτήματα σε σχέση με τις υπόλοιπες, όπως ο γρήγορος χρόνος επεξεργασίας, η ποιότητα των αποτελεσμάτων και ο χρησιμοποιούμενος χώρος μνήμης (Son, 2014).

Η γεω-δημογραφία είναι μια χωρικά σαφής ταξινόμηση των κοινωνικο-οικονομικών δεδομένων, τα οποία μπορούν να χρησιμοποιηθούν για να περιγράψουν και να αναλύσουν χαρακτηριστικά των πολιτών. Αφορά στον κατακερματισμό του πληθυσμού

σε ομοιογενείς ομάδες με βάση ένα φάσμα χαρακτηριστικών που επιτρέπουν τον προσδιορισμό του προφίλ των γειτονικών περιοχών για εφαρμογές σχεδιασμού εμπορικών και δημόσιων υπηρεσιών (Longley 2017). Οι γεωδημογραφικές ταξινομήσεις υποθέτουν ότι κάθε ομάδα είναι σε μεγάλο βαθμό ομοιογενής με βάση την υπόθεση ότι: "birds of a feather flock together" (πουλιά με ένα φτερό ομαδοποιούνται ως σμήνος) (Harris et al., 2005). Ως εκ τούτου, οι άνθρωποι με παρόμοια χαρακτηριστικά τείνουν να βρεθούν σε παρόμοιες τοποθεσίες.

Ωστόσο, υπάρχει ακόμη περιορισμένη κάλυψη στη βιβλιογραφία σχετικά με την ανάπτυξη γεω-δημογραφικών μεθοδολογιών που βασίζονται στο άτομο. Η ικανότητα ταξινόμησης ατόμων βάσει αποκλειστικά προσωπικών χαρακτηριστικών μπορεί να οδηγήσει σε πιο ομοιογενείς συστάδες από τα γεω-δημογραφικά συστήματα βασιζόμενα σε περιοχές.

Κατά την ομαδοποίηση γεω-δημογραφικών στατιστικών δεδομένων πρέπει να λάβουμε υπόψη τα εξής:

Σύμφωνα με τον Everitt κ.ά., (2001), προβλήματα θα προκύψουν εάν υπάρχουν διαφορετικές κλίμακες ή μεγέθη μεταξύ των μεταβλητών. Γενικά, οι μεταβλητές με μεγαλύτερες τιμές και μεγαλύτερες διακυμάνσεις θα έχουν μεγαλύτερη επίπτωση το τελικό μέτρο ομοιότητας. Μετά την ολοκλήρωση της ομαδοποίησης, είναι δυνατόν να εξεταστούν τα μέσα κάθε συστάδας για κάθε διάσταση (μεταβλητή) προκειμένου να εκτιμηθεί ο διακριτικός χαρακτήρας των ομάδων.

Σύμφωνα με τον Milligan (1996), όταν επιλέγουμε τον αριθμό των συστάδων, πρέπει να εξεταστούν τρία βασικά θέματα:

Θέμα 1: Ανάλυση της μέσης απόστασης από τα κέντρα των συστάδων για κάθε επιλογή αριθμού συστάδας. Η ιδανική λύση θα είναι ο αριθμός ομάδων που δίνει τη μικρότερη μέση απόσταση από το κέντρο της συστάδας σε όλες τις ομάδες.

Θέμα 2: Ανάλυση ομοιογένειας μεγέθους συμπλεγμάτων για κάθε επιλογή αριθμού συμπλέγματος. Θα ήταν χρήσιμο, όπου αυτό είναι δυνατό, να υπάρχουν συστάδες όσο το δυνατόν μεγαλύτερου μεγέθους όσον αφορά τον αριθμό των αντικειμένων τους.

Θέμα 3: Ο αριθμός των συνεργαζόμενων ομάδων πρέπει να είναι όσο το δυνατόν πλησιέστερος στον αντιληπτό ιδανικό. Αυτό σημαίνει ότι ο αριθμός των συστάδων πρέπει να είναι μεγέθους που είναι χρήσιμος για περαιτέρω ανάλυση.

Τέλος, οι Shashi και Hui (2008) περιγράφουν τον τρόπο δημιουργίας γεω-δημογραφικής ταξινόμησης ως εξής:

- Χρησιμοποιήστε τη μέθοδο ομαδοποίησης για να αξιολογήσετε τον αριθμό των πιθανών συστάδων που θα έπρεπε να έχει η γεω-δημογραφική ταξινόμηση.
- Αποφασίστε σχετικά με τη γεωγραφική περιοχή που πρόκειται να χρησιμοποιήσετε για την ομαδοποίηση.
- Αποφασίστε για το σύνολο των βασικών μεταβλητών για τις γεωγραφικές περιοχές που σχεδιάζετε να συγκεντρώσετε.
- Προετοιμάστε τα δεδομένα για την ομαδοποίηση (μετασχηματισμοί, τυποποίηση, έλεγχος για ακραίες τιμές).
- Αποφασίστε για τον αριθμό των συστάδων.
- Υπολογίστε και συγκρίνετε τα χαρακτηριστικά των ομάδων, γράψτε τα πορτραίτα για κάθε σύμπλεγμα και ορίστε τα ονόματα σε κάθε σύμπλεγμα.

Οι γεω-δημογραφικές ταξινομήσεις υπολογίζονται χρησιμοποιώντας τα ακόλουθα τέσσερα βήματα (Adnan, 2011):

α) Προδιαγραφές

Αυτό είναι το πρώτο βήμα για τη δημιουργία μιας ταξινόμησης γεω-δημογραφικών δεδομένων. Το βήμα αυτό αφορά επιλογή πηγών δεδομένων, επιλογή μεταβλητών από τις πηγές δεδομένων και προδιαγραφή βαρών για κάθε επιλεγμένη μεταβλητή.

β) Κανονικοποίηση

Όταν οι μεταβλητές έχουν επιλεγεί για την ταξινόμηση, το επόμενο βήμα είναι η τυποποίηση των δεδομένων εισόδου. Αυτό το βήμα μετατρέπει τις μεταβλητές σε ποσοστά ή μέτρα στην ίδια κλίμακα που επιτρέπουν τη σύγκριση. Υπάρχουν πολλοί τεχνικές ομαλοποίησης που χρησιμοποιούνται για την τυποποίηση δεδομένων. Η πιο συνηθισμένη χρήση οι τεχνικές είναι οι τιμές z, η τυποποίηση της κλίμακας και η ανάλυση βασικών συστατικών.

γ) Ομαδοποίηση

Το τρίτο βήμα για τη δημιουργία μιας γεω-δημογραφικής ταξινόμησης είναι η δημιουργία ομάδων ή ομογενείς ομάδες. Ο αλγόριθμος ομαδοποίησης k-mean χρησιμοποιείται για τη δημιουργία των γεω-δημογραφικών κλάδων με το καλύτερο επίπεδο.

δ) Οπτικοποίηση

Η ομαδοποίηση των μεταβλητών δεδομένων χωρίζει τα στοιχεία δεδομένων (στην περίπτωση της δημιουργίας γεωγραφικών ταξινομήσεων σε ομογενείς ομάδες (π.χ., περιοχές παραγωγής"). Σε κάθε περιοχή έχει εκχωρηθεί ένας αναγνωριστικός αριθμός που αντιπροσωπεύει τη δική της "Κατηγορία γεω-δημογραφίας".

Είναι σημαντικό να καταλάβουμε ποιες μεταβλητές στο σύστημα πληροφοριών οδηγούν στην ταξινόμηση. Η κατανόηση αυτών θα μας επιτρέψει να λάβουμε τεκμηριωμένες αποφάσεις όταν επιλέγουμε μεταβλητές για μεταγενέστερες διευρυμένες ταξινομήσεις. Δεν υπάρχουν ενδείξεις για ποιες μεταβλητές είναι οι σημαντικότερες στο σχηματισμό συστάδων. Μερικές θα είναι σημαντικές, δεδομένου ότι θα έχουν ισχυρό αντίκτυπο στην περίπτωσης πολυπαραγοντικού ταξινομικού χώρου. Άλλες μπορεί απλά να αναπαράγουν τις πληροφορίες (και ως εκ τούτου τη θέση) που έχουν ήδη (Debenham, 2002).

Ο Field (2000) υποστηρίζει ότι, σε έναν ιδανικό κόσμο, μια μεταβλητή θα πρέπει να έχει ένα φορτίο μεγάλου παράγοντα για έναν άξονα και χαμηλές βαθμολογίες για οποιονδήποτε άλλο παράγοντα, υποδεικνύοντας ότι η μεταβλητή σχετίζεται μόνο με έναν παράγοντα. Αυτό δεν σημαίνει ότι αυτές οι μεταβλητές είναι όλες λιγότερο σημαντικές για την ταξινόμηση, αλλά απλά ότι μπορούν να εξηγήσουν περισσότερα από μια υποκείμενη διάσταση στα δεδομένα. Ωστόσο, θα μπορούσε να υποστηριχθεί ότι αυτές οι μεταβλητές είναι που δεν περιλαμβάνονται στις βασικές συνιστώσες θα ήταν λιγότερο σημαντικές.

Μια δυσκολία με τα γεωγραφικά δεδομένα είναι ότι συχνά αποτελούν ομαδοποίηση δεδομένων τα οποία έχουν συλλεχτεί σε πιο λεπτομερές επίπεδο. Χαρακτηριστικό παράδειγμα αποτελούν οι Απογραφές Πληθυσμού στις οποίες τα δεδομένα συλλέγονται σε επίπεδο νοικοκυριού. Αυτό το πρόβλημα μπορεί να ονομαστεί πρόβλημα της κλίμακας και της ομαδοποίησης των δεδομένων σε μεγαλύτερες γεωγραφικές ενότητες, με την έννοια ότι η ανάλυση γεωγραφικών φαινομένων σε διαφορετικές κλίμακες μπορεί να επηρεάσει τα αποτελέσματα. Η οικολογική πλάνη συμβαίνει όταν αποτελέσματα τα οποία παρατηρούνται σε κάποιο επίπεδο ομαδοποίησης θεωρείται ότι ισχύουν σε ένα πιο λεπτομερές επίπεδο ανάλυσης.
Τα προβλήματα που προαναφέρθηκαν με τη χρήση κλασικών στατιστικών μεθόδων για την ανάλυση των γεωγραφικών φαινομένων έχουν σε ορισμένες περιπτώσεις αντιμετωπιστεί με την ανάπτυξη στατιστικών μεθόδων ειδικών για την ανάλυση γεωγραφικών φαινομένων, οι οποίες ενσωματώνουν σημαντικές έννοιες όπως η

απόσταση, η γειτνίαση και η εγγύτητα των γεωγραφικών οντοτήτων (O'Sullivan & Unwin, 2010).

Η χωρική δειγματοληψία εφαρμόζεται και σε μελέτη πληθυσμών για τους οποίους δεν είναι διαθέσιμο το δειγματοληπτικό πλαίσιο, αλλά βρίσκονται σε οριοθετημένη γεωγραφική περιοχή.

ΚΕΦΑΛΑΙΟ 3: Χρησιμότητα εφαρμογής του αλγορίθμου k-means σε στατιστικά δεδομένα των περιφερειών της Ελλάδας

3.1 Χρησιμότητα και οφέλη

Ο απώτερος σκοπός της έρευνας είναι να συγκρίνει τις σημερινές 13 περιφέρειες της Ελλάδος, διαμορφώνοντας ομάδες με όμοια και ομάδες με ανόμοια χαρακτηριστικά των κατοίκων της. Η σύγκριση αυτή, θα έχει ως αποτέλεσμα αρχικά τη πιθανότητα να εμφανιστούν νομοί που ανήκουν σε μια περιφέρεια και έχουν ανόμοια χαρακτηριστικά σε σχέση με τους όμορους νομούς της περιφέρειας. Επιπρόσθετα, από τη σύγκριση θα προκύψουν ομάδες, χρήσιμες τόσο σε διοικητικό επίπεδο αλλά και σε επιχειρησιακό (τμηματοποίηση της αγοράς). Το πιο σημαντικό ίσως όμως σε όλη αυτή τη προσπάθεια είναι πώς η εφαρμογή του αλγορίθμου θα γίνει με τη χρήση ενός διαδεδομένου εργαλείου, το Excel, το οποίο βρίσκεται σε κάθε οικιακό αλλά και επαγγελματικό υπολογιστή.

Η αξία της εν λόγω προτεινόμενης ερευνητικής πρότασης μπορεί να εκτιμηθεί από το γεγονός ότι τα αποτελέσματά της, θα αποτελέσουν χρήσιμη πηγή πληροφορίας για τις περιφέρειες. Η ομαδοποίηση είναι μία τεχνική που βοηθάει στη λήψη αποφάσεων. Για παράδειγμα, μελετώντας, μεταβλητές ιατρικού περιεχομένου, μπορεί να επιτευχθεί μια κατηγοριοποίηση των ασθενών με όμοια συμπτώματα και να προσδιοριστούν οι αιτίες των ασθενειών ή να ληφθούν μέτρα πρόληψης και καλύτερης αντιμετώπισης. Το ίδιο μπορεί να επιτευχθεί μελετώντας δεδομένα που αφορούν την εγκληματικότητα. Η τεχνική της ομαδοποίησης σχετίζεται άμεσα και με την τεχνική της βελτιστοποίησης και οι εφαρμογές της πολλαπλασιάζονται αν το ζητούμενο είναι η εύρεση ομοιογενών ομάδων, το μικρό κόστος ή το μέγιστο κέρδος.

ΚΕΦΑΛΑΙΟ 4: Εφαρμογή του αλγορίθμου k-means σε στατιστικά δεδομένα των περιφερειών της Ελλάδας

4.1 Πιθανά ερευνητικά ερωτήματα και προβληματισμοί

Το πρώτο ερευνητικό ερώτημα που θα πρέπει να απαντηθεί στην παρούσα εργασία, είναι «Ποιες μεταβλητές πρέπει να μελετηθούν;». Για την απάντηση σε αυτό το ερώτημα θα πρέπει να ληφθούν υπόψη οι μεταβλητές που χρησιμοποιούνται σε σχετικές έρευνες στη βιβλιογραφία αλλά και οι μεταβλητές για τις οποίες τα δεδομένα είναι διαθέσιμα και προσβάσιμα σε επίπεδο νομού. Μετά την ανάλυση, εκ των υστέρων δηλαδή, είναι πιθανό να αποδειχθεί πως κάποιες από τις μεταβλητές που επιλέχθηκαν είναι αδιάφορες, δεν προσφέρουν κάποια πληροφορία και να αποφασιστεί η αφαίρεσή τους.

Ένα δεύτερο ερευνητικό ερώτημα που θα πρέπει να απαντηθεί, είναι η επιλογή ενός μέτρου ομοιότητας για τη σύγκριση των μεταβλητών. Τα μέτρα ομοιότητας είναι δύο, οι μετρικές αποστάσεις και οι συνιστώσες συσχέτισης. Οι αποστάσεις μετράνε πόσο κοντά είναι τα κεντρικά σημεία των συστάδων που σχηματίζονται ενώ οι συσχετίσεις μετράνε συμπεριφορές. Το πόσο επιτυχημένο θα είναι το αποτέλεσμα της ομαδοποίησης δεδομένων σχετίζεται με το μέτρο ομοιότητας που θα επιλεγεί. Με βάση τη βιβλιογραφία συνηθίζεται η χρήση μετρικής απόστασης και πιο συγκεκριμένα η Ευκλείδεια απόσταση, ειδικότερα δε όταν τα δεδομένα που μελετώνται είναι συνεχή. Στην παρούσα εργασία το αν θα επιλεγεί αυτό το μέτρο ομοιότητας εξαρτάται από το πλήθος των μεταβλητών που θα μελετηθούν, γιατί σε περιπτώσεις πολλών μεταβλητών η χρήση της Ευκλείδειας απόστασης δεν ενδείκνυται (Mirkin, 2006). Σημαντικό ρόλο στην απόφαση θα διαδραματίσει και το είδος των μεταβλητών που θα ληφθούν υπόψη.

Ένα τρίτο ερευνητικό ερώτημα είναι το αν το πλήθος των μεταβλητών που θα ληφθούν υπόψη επηρεάζει την ομαδοποίηση. Αν η απάντηση σε αυτή την ερώτηση είναι αρνητική τότε θα χρησιμοποιηθούν λιγότερες μεταβλητές για οικονομία χρόνου αλλά και για διευκόλυνση στην έρευνα.

Για την ανάλυση των δεδομένων, αρχικά είναι απαραίτητη η ανάπτυξη του αλγορίθμου k-means με τη χρήση συναρτήσεων στο λογισμικό MS Office excel.
Το τελικό στάδιο της ανάλυσης δεδομένων περιλαμβάνει την αξιολόγηση και την ερμηνεία των αποτελεσμάτων της ομαδοποίησης. Η ανάλυση και η ερμηνεία αυτή θα χωριστεί σε κατηγορίες ανάλογα με τα χαρακτηριστικά των μεταβλητών που χρησιμοποιήθηκαν. Τέτοιες κατηγορίες μπορεί να είναι οικονομικής, δημογραφικής φύσεως κτλ. Η ανάλυση των ομάδων που δημιουργήθηκαν θα περιλαμβάνει την εξέταση των κέντρων βάρους των ομάδων, ώστε να εξασφαλιστεί ότι οι ομάδες που σχηματίστηκαν έχουν νόημα, δηλαδή είναι εννοιολογικά διακριτές.

Ένα πιθανό πρόβλημα που πιθανότατα να προκύψει από τη χρήση του αλγορίθμου K-means είναι η ευαισθησία του αλγορίθμου στην επιλογή των αρχικών συνθηκών (Milligan, 1980). Η προτεινόμενη λύση σύμφωνα με την βιβλιογραφία στο συγκεκριμένο πρόβλημα, είναι οι επαναλαμβανόμενες εκτελέσεις του αλγορίθμου με τυχαία σύνολα επιλεγμένων αρχικών κεντροειδών και η μετέπειτα επιλογή των ομάδων με το μικρότερο σφάλμα ομαδοποίησης.

Ένα ακόμη πιθανό πρόβλημα που μπορεί να προκύψει, είναι τα αποτελέσματα του αλγορίθμου να επηρεάζονται έντονα από περιοχές πυκνοκατοικημένες, όπως η Αθήνα. Ανάλογα με την έκταση του προβλήματος, είναι πιθανό τέτοιες περιοχές να εξαιρεθούν από την έρευνα.

Οι 13 περιφέρειες που θα μελετηθούν σε αυτή την εργασία είναι:

- Περιφέρεια Θεσσαλίας
- Περιφέρεια Νοτίου Αιγαίου
- Περιφέρεια Πελοποννήσου
- Περιφέρεια Δυτικής Ελλάδας
- Περιφέρεια Κεντρικής Μακεδονίας
- Περιφέρεια Βορείου Αιγαίου

- Περιφέρεια Ηπείρου
- Περιφέρεια Ανατολικής Μακεδονίας-Θράκης
- Περιφέρεια Αττικής
- Περιφέρεια Δυτικής Μακεδονίας
- Περιφέρεια Ιονίων Νήσων
- Περιφέρεια Στερεάς Ελλάδας
- Περιφέρεια Κρήτης

Η κάθε περιφέρεια με τη σειρά της απαρτίζεται από περιφερειακές ενότητες όπως παρουσιάζονται στο παράρτημα 1.

4.2. Στατιστική ανάλυση των δειγμάτων

Τα στατιστικά δεδομένα των μεταβλητών που θα μελετηθούν σε αυτή την εργασία πάρθηκαν από την ΕΛΣΤΑΤ και παρουσιάζονται στο παράρτημα 2. Για τη μελέτη τους θα εξετάσουμε αρχικά τα βασικά στατιστικά στοιχεία. Στον πίνακα 4.2.1 παρουσιάζονται ο μέσος όρος και η τυπική απόκλιση του δείγματος κάθε μεταβλητής και στον πίνακα 4.2.2 οι συντελεστές συσχέτισης μεταξύ των μεταβλητών.

Πίνακας 4.2.1. Μέσοι όροι και τυπικές αποκλίσεις των μεταβλητών

Μεταβλητή	Μέσος όρος	Τυπική απόκλιση	Σύνολο νομών
Κατά κεφαλή ΑΕΠ	13.166	2.862	51
Απασχόληση %	36	3	51
Ακαθάριστη προστιθέμενη αξία ανά εργαζόμενο	11.492	2.498	51
Κατανάλωση πετρελαιοειδών ανά 1.000 κατοίκους	7	6	51
Μηνιαία δαπάνη νοικοκυριού	872	302	51
Εκτάσεις και παραγωγή (στη γεωργία)	309	149	51
Εκμεταλλεύσεις και αριθμός ζώων (στην κτηνοτροφία)	52	10	51

Μια αρχική πρόβλεψη από τα δεδομένα είναι ότι οι μεταβλητές του πίνακα 4.2.1 μπορούν να χρησιμοποιηθούν για ομαδοποίηση λόγω των μεγάλων τυπικών αποκλίσεων.

Πίνακας 4.2.2. Συντελεστές συσχέτισης μεταξύ των μεταβλητών

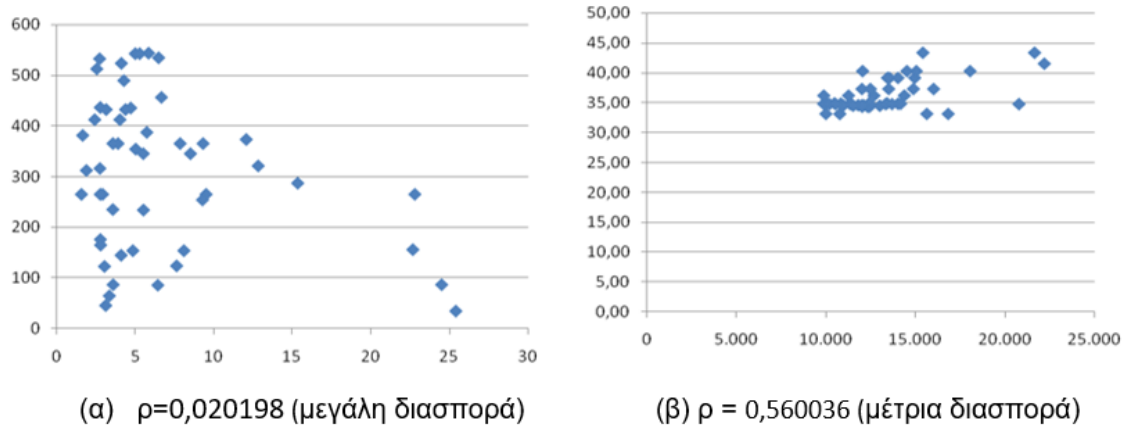
	C	D	E	F	G	H	I
C	1	0,560036	1	-0,07764	0,537162	-0,25063	-0,2936
D	0,560036	1	0,560036	0,114521	0,502485	-0,18016	-0,20171
E	1	0,560036	1	-0,07764	0,537162	-0,25063	-0,2936
F	-0,07764	0,114521	-0,07764	1	0,020198	-0,30717	-0,10543
G	0,537162	0,502485	0,537162	0,020198	1	-0,32729	-0,21644
H	-0,25063	-0,18016	-0,25063	-0,30717	-0,32729	1	0,449936
I	-0,2936	-0,20171	-0,2936	-0,10543	-0,21644	0,449936	1

Τα C,D,E,F,G,H,I αντιπροσωπεύουν τα εξής:

C: Στήλη στο Excel με τα δεδομένα της μεταβλητής «Κατά κεφαλή ΑΕΠ»
D: Στήλη στο Excel με τα δεδομένα της μεταβλητής «Απασχόληση %»
E: Στήλη στο Excel με τα δεδομένα της μεταβλητής «Ακαθάριστη προστιθέμενη αξία ανά εργαζόμενο»
F: Στήλη στο Excel με τα δεδομένα της μεταβλητής «Κατανάλωση πετρελαιοειδών ανά 1.000 κατοίκους»
G: Στήλη στο Excel με τα δεδομένα της μεταβλητής «Μηνιαία δαπάνη νοικοκυριού»
H: Στήλη στο Excel με τα δεδομένα της μεταβλητής «Εκτάσεις και παραγωγή (στη γεωργία)»
I: Στήλη στο Excel με τα δεδομένα της μεταβλητής «Εκμεταλλεύσεις και αριθμός ζώων (στην κτηνοτροφία)»

Αν εξαιρέσουμε τις «Κατά κεφαλή ΑΕΠ» και «Ακαθάριστη προστιθέμενη αξία ανά εργαζόμενο» που έχουν μια πολύ ισχυρή γραμμική σχέση μεταξύ τους, οι υπόλοιπες έχουν σχεδόν μηδενική γραμμική σχέση («Μηνιαία δαπάνη νοικοκυριού» και «Κατανάλωση πετρελαιοειδών ανά 1.000 κατοίκους») έως μια ασθενή γραμμική σχέση («Απασχόληση %» με «Κατά κεφαλή ΑΕΠ» και «Ακαθάριστη προστιθέμενη αξία ανά εργαζόμενο») που σημαίνουν μεγάλη διασπορά (Σχήμα 4.2.1 (α)) και μέτρια διασπορά (Σχήμα 4.2.1(β)) διασπορά αντίστοιχα. Εκεί που έχουμε μεγάλη διασπορά μπορούμε

να έχουμε περισσότερες ομαδοποιήσεις σε σχέση με τη μέτρια και μικρότερη διασπορά. Οι θετικές τιμές δηλώνουν ότι όταν αυξάνεται η μια μεταβλητή αυξάνεται και η άλλη, ενώ οι αρνητικές τιμές δηλώνουν ότι όταν αυξάνεται η μια μεταβλητή μειώνεται η άλλη.



(α) ρ=0,020198 (μεγάλη διασπορά) (β) ρ = 0,560036 (μέτρια διασπορά)

Σχήμα 4.2.3: Δείγματα διασποράς μεταξύ μεταβλητών.

Για την ανάλυσή τους μπορεί να βρούμε μη αναμενόμενες σχέσεις ανάμεσα στα δεδομένα καθώς και να τα συνοψίσουμε με νέους τρόπους που είναι κατανοητοί και χρήσιμοι στους χρήστες. Για παράδειγμα, μεγάλη διασπορά του πληθυσμού γύρω από τον χωρικό μέσο υποδηλώνει αρχικά στάδια της αστικοποίησης, ενώ μείωση της διασποράς σε μεταγενέστερη περίοδο μπορεί να σημαίνει τάσεις αστυφιλίας (Κουτσόπουλος, 2009). Όσο υψηλότερη η συσχέτιση μιας ανεξάρτητης μεταβλητής με την εξαρτημένη, τόσο σημαντικότερη εικάζεται καταρχήν ότι θα είναι η συνεισφορά της στο μοντέλο (Πανταζής, 2005)

Το μέτρο ομοιότητας μεταξύ των ομάδων που θα χρησιμοποιηθεί είναι η Ευκλείδεια, εφόσον εκ των προτέρων έχει αποφασιστεί η ομαδοποίηση με τον αλγόριθμο k-means.

Ο έλεγχο εγκυρότητας των ομάδων (cluster validity) θα γίνει με αποτίμηση αντικειμενικών και ποσοτικών κριτηρίων του αποτελέσματος της ομαδοποίησης. Ο αλγόριθμος k-means παρουσιάζει προβλήματα όταν τα δεδομένα περιέχουν σημεία δεδομένων που διαφέρουν σημαντικά σε απόσταση από τα άλλα (outliers) κάτι που θα διερευνηθεί στα δεδομένα μας.

4.3. Εφαρμογή του αλγορίθμου k-means στα στατιστικά δεδομένα των μεταβλητών

4.3.1 Ορισμός αριθμού συστάδων k

Αρχικά θα χρησιμοποιηθεί η μέθοδος «σύμφωνα με τον κανόνα».

$$k \approx \sqrt{\frac{n}{2}}$$

Ο υπολογισμός του k

$$k \approx \sqrt{\frac{n}{2}} \approx \sqrt{\frac{51}{2}} \approx 5$$

Όπως θα αποδειχθεί πιο κάτω αυτός ο αριθμός των συστάδων ισχύει κάτω από προϋποθέσεις (πρόβλημα της κλίμακας) διότι δημιουργούνται κεντροειδή τα οποία δεν είναι εννοιολογικά διακριτά με αποτέλεσμα κάποιοι νομοί με διαφορετικά «τρεξίματα» του αλγορίθμου πότε να ομαδοποιούνται σε μια ομάδα και πότε σε μια άλλη. Επίσης, θα αποδειχθεί ότι οι τρείς συστάδες για τις συγκεκριμένες μεταβλητές, και τις τιμές που αυτές έχουν, δημιουργούν εννοιολογικά διακριτά κετροειδή.

4.3.2 Επεξήγηση του αλγορίθμου k-means στο Excel

Ο αλγόριθμος που έχει υλοποιηθεί στο Excel είναι επαναληπτικός (iterative) και περιγράφεται στην ενότητα 1.4.

Οι περιοχές στις οποίες υπολογίζονται και επανυπολογίζονται τα κεντροειδή είναι:

Η περιοχή των δεδομένων της μεταβλητής «Κατά κεφαλή ΑΕΠ» που έχει όνομα feat1.

Η περιοχή των δεδομένων της μεταβλητής «Απασχόληση %» που έχει όνομα feat2.

Η περιοχή των δεδομένων της μεταβλητής «Ακαθάριστη προστιθέμενη αξία ανά εργαζόμενο» που έχει όνομα feat3.

Η περιοχή των δεδομένων της μεταβλητής «Κατανάλωση πετρελαιοειδών ανά 1.000 κατοίκους» που έχει όνομα feat4.

Η περιοχή των δεδομένων της μεταβλητής «Μηνιαία δαπάνη νοικοκυριού» που έχει όνομα feat5.

Η περιοχή των δεδομένων της μεταβλητής «Εκτάσεις και παραγωγή (στη γεωργία)» που έχει όνομα feat6.

Η περιοχή των δεδομένων της μεταβλητής «Εκμεταλλεύσεις και αριθμός ζώων (στην κτηνοτροφία)» που έχει όνομα feat7.

Η περιοχή των τιμών των κεντροειδών όπου αποδίδονται με τυχαίο τρόπο αρχικά και επανυπολογίζονται σε κάθε επανάληψη βρίσκεται δίπλα από τις μεταβλητές και έχει όνομα cnumb. Η αρχικοποίηση των τιμών των κεντροειδών δίνεται από την εντολή RANDBETWEEN(1;N), όπου N είναι ο αριθμός των συστάδων που θέλουμε να δημιουργηθούν.

Οι μεταβλητές που χρησιμοποιούνται είναι:
Η μεταβλητή iteration_count που μετράει τις επαναλήψεις του αλγορίθμου όσο η τιμή της μεταβλητής indicator_2 είναι TRUE.
Η μεταβλητή max_count που ορίζει το μέγιστο αριθμό των επαναλήψεων του αλγορίθμου.
Η μεταβλητή indicator_1 που παίρνει την τιμή TRUE, αν ισχύει iteration_count< max_count, αλλιώς παίρνει την τιμή FALSE.
Η μεταβλητή indicator_2 που παίρνει την τιμή της μεταβλητής indicator_1.

Οι εντολές που χρησιμοποιούνται είναι:
Η εντολή SUMIF(cnumb; X; featY) προσθέτει τα δεδομένα της περιοχής featY (Y είναι η αντίστοιχη μεταβλητή), όπου στην αντίστοιχη περιοχή των κεντροειδών cnumb υπάρχει η τιμή X=1,2,3,… του κεντροειδούς.
Η εντολή COUNTIF(cnumb;X) βρίσκει το σύνολο των κεντροειδών στην περιοχή cnumb που έχουν την τιμή ίση με X.
Η εντολή IF(indicator_1; Q_{χ}; RANDBETWEEN(1;3)) αντιγράφει τις τιμές των κεντροειδών της τελευταίας επανάληψης, που βρίσκονται στη θέση Q_{χ}, στην αντίστοιχη θέση της περιοχής cnumb, αν η τιμή της μεταβλητής indicator_1 είναι TRUE, αλλιώς εκτελεί την εντολή RANDBETWEEN(1;3) για την απόδοση των αρχικών τυχαίων τιμών των κεντροειδών.
Η διαίρεση SUMIF(cnumb; X; featY)/COUNTIF(cnumb; X) δίνει τη μέση τιμή των δεδομένων κάθε μεταβλητής (αντίστοιχο featY) που ανήκουν σε μια συστάδα, όπου το κεντροειδές έχει τιμή X.
Η εντολή SQRT((C109-C12)^2+(D109-D12)^2+(E109-E12)^2+(F109-F12)^2+(G109-G12)^2+(H109-H12)^2+(I109-I12)^2) υπολογίζει την Ευκλείδεια απόσταση από τη μέση τιμή των δεδομένων της αντίστοιχης περιοχής featY, που ανήκει στο κεντροειδές που έχει τιμή X (στο συγκεκριμένο παράδειγμα X=1), από την αντίστοιχη τιμή στην περιοχή featY. Η εντολή αυτή εφαρμόζεται σε όλες τις τιμές της περιοχής featY σε αντίστοιχα κελιά (όσα είναι τα δεδομένα).
Η εντολή MIN(N109:P109) βρίσκει τη μικρότερη τιμή (απόσταση) των δεδομένων που βρίσκονται στα κελιά από N109 ως P109

Η εντολή MATCH(MIN(N109:P109);N109:P109;0): Αποδίδει την αντίστοιχη τιμή του κεντροειδούς που ανήκει ένα αντικείμενο (συγκεκριμένη τιμή μιας μεταβλητής) και έχει τη μικρότερη απόσταση που βρίσκεται στα κελιά N109 ως P109.

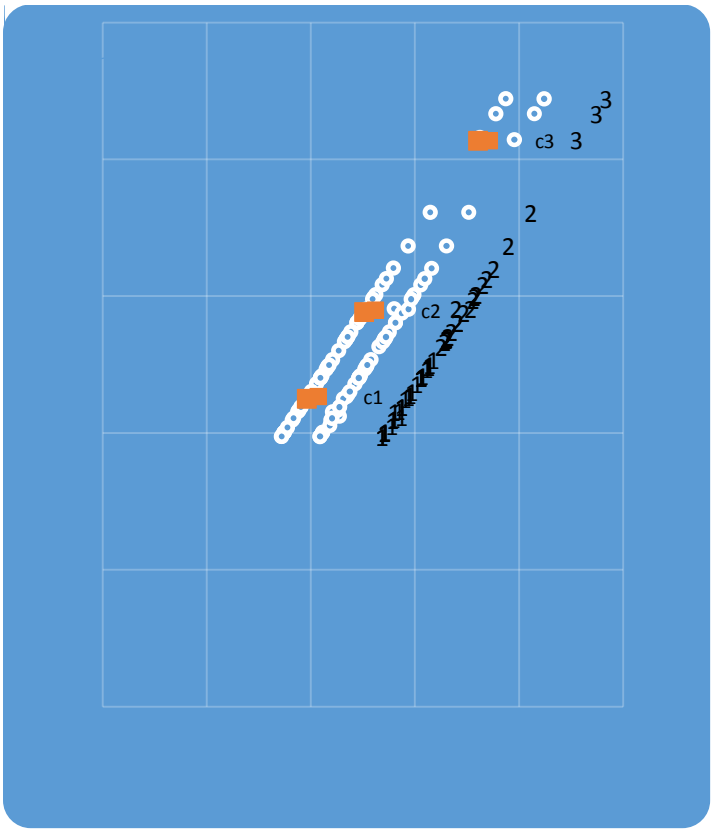
4.4 Αντίκτυπος των μεταβλητών στην ομαδοποίηση

Στην ενότητα αυτή θα μελετήσουμε κατά πόσο οι μεταβλητές επηρεάζουν την ομαδοποίηση μεταβάλλοντας τον αριθμό των μεταβλητών που λαμβάνουν μέρος στην ομαδοποίηση. Επίσης, κατά πόσο επηρεάζουν οι τιμές των μεταβλητών τον υπολογισμό των κεντροειδών. Για να αποφασίσουμε για την εγκυρότητα των συστάδων δεν είναι εύκολο μόνο με την οπτικοποίηση, αλλά πρέπει να εμπλακούν δείκτες εγκυρότητας. Στην παρούσα εργασία θα χρησιμοποιηθούν σχετικά κριτήρια εγκυρότητας των συστάδων. Δηλαδή, η εκτίμηση της δομής ομαδοποίησης με συστάδες με άλλες δομές που προκύπτουν από τον ίδιο αλγόριθμο με διαφορετικές τιμές παραμέτρων. Για τη μελέτη της ομαδοποίησης ως ανεξάρτητη μεταβλητή θα χρησιμοποιηθεί η μεταβλητή «Κατά κεφαλή ΑΕΠ». Στα πλαίσια της προεργασίας για την ανάλυση, δοκιμάσαμε πολλαπλές δομές μεταβλητών και συστάδων με σκοπό την αποσαφήνιση της σημασίας των μεταβλητών. Τα σημαντικότερα ευρήματα από τις δοκιμές παρουσιάζονται παρακάτω.

4.4.1 Δοκιμές ομαδοποίησης με 7 μεταβλητές και 3 συστάδες.

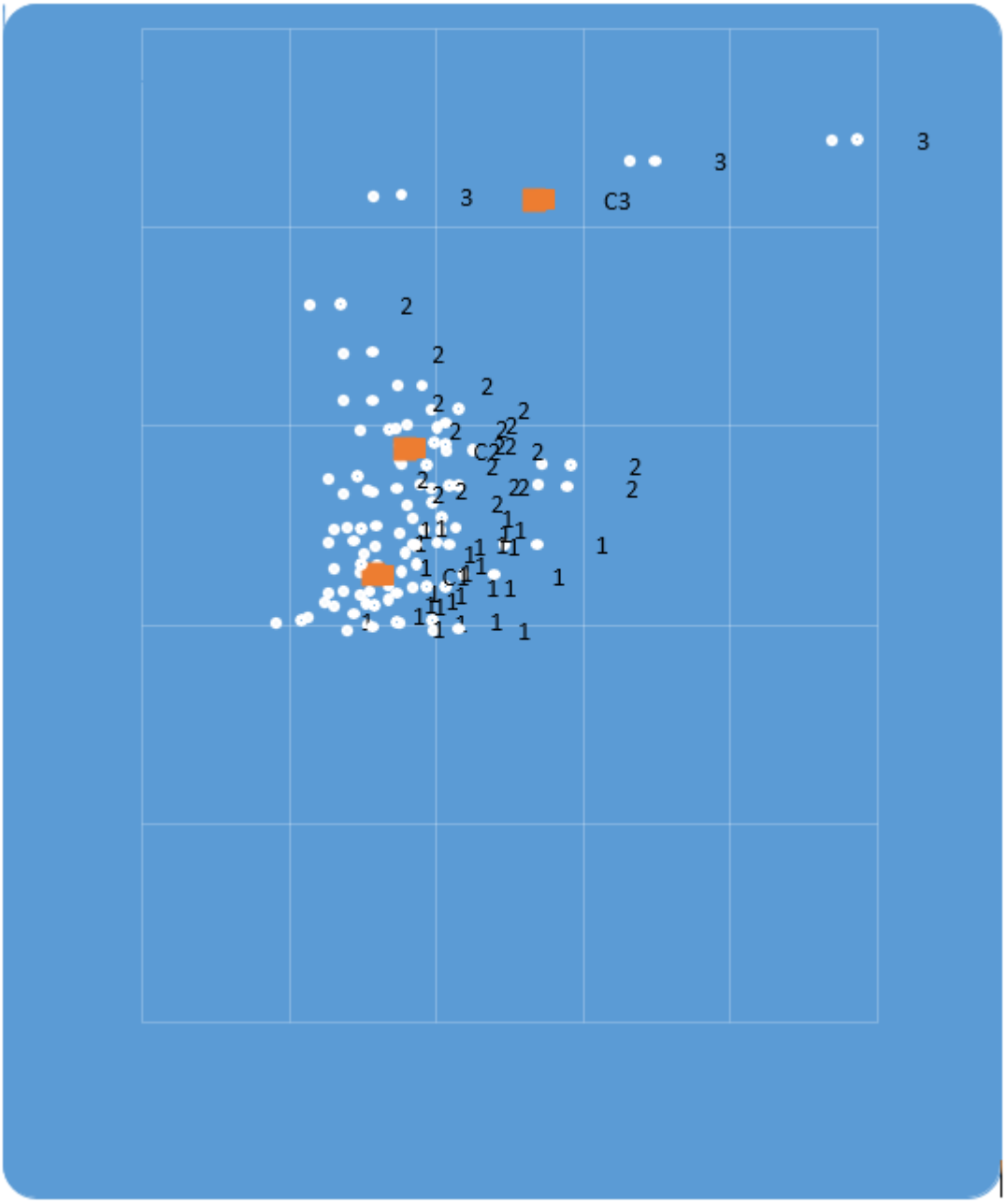
Κατά την συγκεκριμένη δοκιμή, έγινε φανερό ότι οι μεταβλητές «Κατά κεφαλή ΑΕΠ « και «Ακαθάριστη προστιθέμενη αξία ανά εργαζόμενο» έχουν γραμμική σχέση μεταξύ τους, και αποφασίστηκε να μην χρησιμοποιηθεί η δεύτερη.
Συγκεκριμένα, στο σχήμα 4.4.1.1 φαίνεται η ομαδοποίηση των μεταβλητών «Κατά κεφαλή ΑΕΠ» (άξονας Y) και «Ακαθάριστη προστιθέμενη αξία ανά εργαζόμενο» (άξονας X). Όπως αναφέρθηκε κατά την εξέταση των συντελεστών συσχέτισης μεταξύ των μεταβλητών, αυτές οι δυο μεταβλητές έχουν συντελεστή συσχέτισης ίσο με 1. Επομένως, είναι προφανές -και από το διάγραμμα- ότι έχουν γραμμική σχέση μεταξύ τους, περιέχουν δηλαδή την ίδια πληροφορία. Όσον αφορά στην ομαδοποίηση, γίνεται ομαδοποίηση, αλλά τα δυο κεντροειδή C1 και C2 απέχουν πολύ μικρή απόσταση μεταξύ τους, ενώ το κεντροειδές C3 απέχει κατά λίγο περισσότερο από τα άλλα δυο. Τα δεδομένα των συστάδων με κεντροειδή τα C1 και C2 διαχωρίζονται ελάχιστα μεταξύ τους δυσκολεύοντας κάποιον να βγάλει συμπεράσματα για αυτά. Καταλήξαμε λοιπόν

να αφαιρέσουμε από την περαιτέρω ανάλυση την «Ακαθάριστη προστιθέμενη αξία ανά εργαζόμενο».



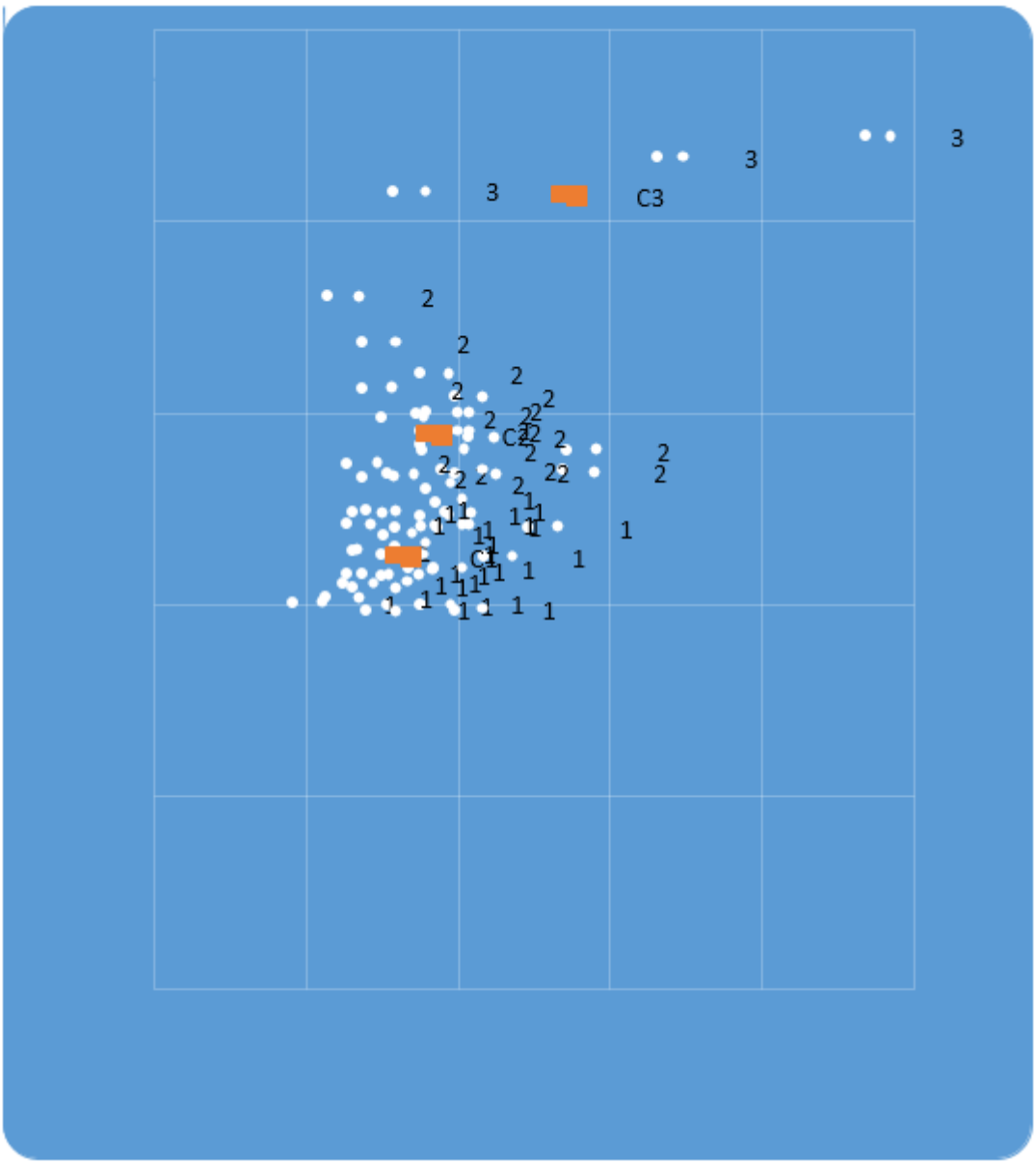
Σχήμα 4.4.1.4: Ομαδοποίηση δεδομένων με τις μεταβλητές«Κατά κεφαλή ΑΕΠ» και «Ακαθάριστη προστιθέμενη αξία ανά εργαζόμενο».

Στο σχήμα 4.4.1.2 φαίνεται η το διάγραμμα των μεταβλητών «Κατά κεφαλή ΑΕΠ» και «Μηνιαία δαπάνη νοικοκυριού». Από την ομαδοποίηση προκύπτει ότι τα δεδομένα διαχωρίζονται σε τρείς συστάδες. Τα δυο κεντροειδή C1 και C2 απέχουν μικρή απόσταση μεταξύ τους, ενώ το κεντροειδές C3 απέχει αρκετά από τα άλλα δυο. Τα δεδομένα απλώνονται ως προς τον άξονα «Μηνιαία δαπάνη νοικοκυριού» λόγω της μικρής συσχέτισης των δυο μεταβλητών δυσκολεύοντας να βγάλουμε συμπέρασμα για το κατά πόσο η μεταβολή της μεταβλητής αυτής επηρεάζει τη μεταβλητή «Κατά κεφαλή ΑΕΠ» (για παράδειγμα κάποια σημεία της συστάδας c3 βρίσκονται δεξιότερα από κάποια σημεία των συστάδων c1 και c2.



Σχήμα 4.4.5.2: *Ομαδοποίηση δεδομένων με 7 μεταβλητές και 3 συστάδες*

Στο σχήμα 4.4.1.3 φαίνεται η ομαδοποίηση των μεταβλητών «Κατά κεφαλή ΑΕΠ» (άξονας Υ) και «Μηνιαία δαπάνη νοικοκυριού» (άξονας Χ). Βλέπουμε ότι οι δύο αυτές μεταβλητές δεν έχουν γραμμική σχέση μεταξύ τους, άρα μπορούν να μείνουν στην ανάλυσή μας καθώς περιέχουν διαφορετική πληροφορία. Με αντίστοιχο τρόπο, αποφασίσαμε να κρατήσουμε και τις υπόλοιπες μεταβλητές.


3
3
3
C3
2
2
2

κανονικοποιηθούν οι μεταβλητές με την μέθοδο ελάχιστου-μέγιστου, ενώ επιλέχτηκε η κλίμακα 0-100 για λόγους καλύτερης απεικόνισης στο excel.

4.4.3 Τελική Ομαδοποίηση με 6 μεταβλητές και 5 συστάδες.

Λαμβάνοντας υπόψιν τα παραπάνω, έγινε η τελική ομαδοποίηση με τον αλγόριθμο K-means, η οποία μας έδωσε τις παρακάτω συστάδες που αναλύονται στη συνέχεια:

1	2	3	4	5
Λέσβος	Ηράκλειο	Ξάνθη	Αττικής	Λασίθι
Ικαρία	Χανιά	Ροδόπη	Χίος	Ρέθυμνο
Λήμνος	Έβρος	Δράμα	Θάσος - Καβάλα	Γρεβενά
Σάμος	Πέλλα	Ημαθία	Θεσσαλονίκη	
	Άρτα	Κιλκίς	Χαλκιδική	
	Καρδίτσα	Πιερία	Ζάκυνθος	
	Τρίκαλα	Σέρρες	Κάλυμνος Κάρπαθος Κως Ρόδος	
	Λάρισα	Κοζάνη	Άνδρος Θήρα Κέα Μήλος Μύκονος Νάξος Πάρος Σύρος Τήνος	
	Αιτωλοακαρνανία	Καστοριά		
	Αχαΐα	Φλώρινα		
	Ηλεία	Πρέβεζα		
	Εύβοια	Θεσπρωτία		
	Αρκαδία	Ιωάννινα		
	Κορινθία	Μαγνησία		
		Κέρκυρα		
		Ιθάκη - Κεφαλληνία		
		Βοιωτία		
		Ευρυτανία		
		Φθιώτιδα		
		Φωκίδα		
		Αργολίδα		
		Λακωνία		
		Μεσσηνία		
		Λευκάδα		

4.4.3.1 Συστάδα 1

Η συστάδα 1 αποτελείται από τη Λέσβο, τη Λήμνο, την Ικαρία και τη Σάμο και το κέντρο της έχει τις παρακάτω τιμές:

Κατά κεφαλή αεπ	12265,49
Απασχόληση %	35,38
Κατανάλωση πετρελαιοειδών ανα 1000 κατοίκους	24,19
Μηνιαία δαπανη νοικοκυριού	918,30
Εκτάσεις και παραγωγή (στη γεωργία)	92,64
Εκμεταλλεύσεις και αριθμός ζώων (στην κτηνοτροφία)	35,33

Παρατηρούμε ότι το ιδιαίτερο χαρακτηριστικό της συγκεκριμένης συστάδας είναι η πολύ υψηλή κατανάλωση πετρελαιοειδών, το οποίο οφείλεται ότι οι συγκεκριμένοι νομοί έχουν αναλογικά την μεγαλύτερη αναλογία προσφύγων, άρα και πληθυσμό.

Αξίζει να σημειωθεί επίσης ότι η συστάδα αυτή, ταυτίζεται με την Περιφέρεια Βορείου Αιγαίου.

4.4.3.2 Συστάδα 2

Η συστάδα 2 αποτελείται από 14 νομούς με το κέντρο της να έχει τις παρακάτω τιμές :

Κατά κεφαλή αεπ	12571,90
Απασχόληση %	35,95
Κατανάλωση πετρελαιοειδών ανα 1000 κατοίκους	4,46
Μηνιαία δαπανη νοικοκυριού	918,30
Εκτάσεις και παραγωγή (στη γεωργία)	443,20
Εκμεταλλεύσεις και αριθμός ζώων (στην κτηνοτροφία)	72,13

Το ιδιαίτερο χαρακτηριστικό της συγκεκριμένης συστάδας είναι οι μεγάλες τιμές πρωτογενούς παραγωγής (γεωργία-κτηνοτροφία). Παρατηρώντας περαιτέρω τους νομούς που την απαρτίζουν, βλέπουμε ότι περιλαμβάνει πρακτικά όλες τις μεγάλες πεδινές «περιοχές» της χώρας, καθώς και το μεγαλύτερο μέρος της Πελοποννήσου.

4.4.3.3 Συστάδα 3

Η συστάδα 3 αποτελεί την πολυπληθέστερη και χαρακτηρίζει τον «τυπικό» νομό της Ελλάδος.
Το κέντρο της συστάδας έχει τις παρακάτω τιμές:

Κατά κεφαλή αεπ	12578,57
Απασχόληση %	35,38
Κατανάλωση πετρελαιοειδών ανα 1000 κατοίκους	4,93
Μηνιαία δαπανη νοικοκυριού	723,20
Εκτάσεις και παραγωγή (στη γεωργία)	328,08
Εκμεταλλεύσεις και αριθμός ζώων (στην κτηνοτροφία)	46,73

Αποτελείται από 24 νομούς και χαρακτηρίζεται από σχετικά χαμηλό μέσο εισόδημα, απασχόληση και κατανάλωση, ενώ δεν ξεχωρίζουν ούτε στην πρωτογενή παραγωγή τους στη γεωργία και στην κτηνοτροφία. Περιλαμβάνει το μεγαλύτερο μέρος από τις περιφέρειες Δυτικής Μακεδονίας, Ανατολικής Μακεδονίας και Θράκης, Ηπείρου και Στερεάς Ελλάδας.

4.4.3.4 Συστάδα 4

Η συστάδα 4 αποτελείται από 8 νομούς και το κέντρο της έχει τις παρακάτω τιμές:

Κατά κεφαλή αεπ	16368,79
Απασχόληση %	38,37
Κατανάλωση πετρελαιοειδών ανα 1000 κατοίκους	5,15
Μηνιαία δαπανη νοικοκυριού	1223,39
Εκτάσεις και παραγωγή (στη γεωργία)	107,35
Εκμεταλλεύσεις και αριθμός ζώων (στην κτηνοτροφία)	29,62

Βλέπουμε ότι χαρακτηρίζονται από το –με διαφορά- υψηλότερο κατά κεφαλή ΑΕΠ, ενώ έχει τα πρωτεία στην Απασχόληση και στη δαπάνη νοικοκυριού. Αντίθετα έχει τις χαμηλότερες τιμές στον πρωτογενή τομέα, που μας οδηγεί στο συμπέρασμα ότι οι υπηρεσίες είναι ο βασικός τομέας της οικονομίας στη συστάδα αυτή.

Θα χαρακτηρίζαμε λοιπόν τη συστάδα αυτή ως εκείνη με τους «Πλούσιους» νομούς της Ελλάδος (Αττική, Θεσσαλονίκη, Κυκλάδες, Δωδεκάνησα κλπ).

4.4.3.5 Συστάδα 5

Η συστάδα 5 αποτελείται από 3 νομούς και το κέντρο της έχει τις παρακάτω τιμές:

Κατά κεφαλή αεπ	12791,72
Απασχόληση %	37,09
Κατανάλωση πετρελαιοειδών ανα 1000 κατοίκους	16,99
Μηνιαία δαπανη νοικοκυριού	872,34
Εκτάσεις και παραγωγή (στη γεωργία)	290,99
Εκμεταλλεύσεις και αριθμός ζώων (στην κτηνοτροφία)	70,00

Στα ιδιαίτερα χαρακτηριστικά της συγκεκριμένης συστάδας, περιλαμβάνεται ο υψηλός δείκτης της κτηνοτροφίας σε αντιδιαστολή με το χαμηλό δείκτη της γεωργίας, ενώ οι νομοί που περιλαμβάνονται σε αυτή τη συστάδα παρουσιάζουν και υψηλή κατανάλωση πετρελαιοειδών. Βλέποντας τους νομούς παρατηρούμε ότι πρόκειται για 3 από τους πιο ορεινούς νομούς της Ελλάδος.

ΚΕΦΑΛΑΙΟ 5: Συμπεράσματα

5.1 Συμπεράσματα της εφαρμογής του αλγορίθμου k-means

Το κύριο και βασικότερο συμπέρασμα της παρούσας έρευνας είναι ότι η χρήση του excel αρκεί και ικανοποιεί τις συνθήκες που απαιτεί η εφαρμογή του αλγορίθμου k-means για την ομαδοποίηση γεωγραφικών περιοχών σε επίπεδο νομών αλλά και περιφέρειας με βάση συγκεκριμένα χαρακτηριστικά. Είναι ένα πολύ σημαντικό συμπέρασμα αν λάβει κανείς υπόψη ότι το εργαλείο αυτό υπάρχει εγκατεστημένο στη πλειονότητα των υπολογιστών στην Ελλάδα και ειδικότερα στους υπολογιστές του δημόσιου τομέα. Οπότε με ένα εργαλείο που ήδη έχουν στη κατοχή τους οι περιφέρειες, μπορούν να πάρουν καλύτερες αποφάσεις σχετικές με τη διοίκηση, την οικονομία και την ανάπτυξη των περιφερειών τους.

Από την ανάλυση είδαμε ότι, σε αρκετές περιπτώσεις, η διοικητική διαίρεση της Ελλάδας σε 13 περιφέρειες, συμφωνεί με την ανάλυσή μας, όπως συμβαίνει στη συστάδα 1 και στη συστάδα 3, με βάση την περιφέρεια και συγκεκριμένες συνθήκες. Από την άλλη όμως, παρατηρούμε περιπτώσεις όπου νομοί από διαφορετικές διοικητικές περιφέρειες έχουν πολλά κοινά χαρακτηριστικά, γεωγραφικά χαρακτηριστικά (συστάδες 2 και 5), αλλά και οικονομικά στοιχεία (συστάδες 2 και 4). Φαίνεται λοιπόν ότι η Άρτα, έχει περισσότερα κοινά χαρακτηριστικά, ανάγκες και δυνατότητες με τον Έβρο από ότι με τα γειτονικά Ιωάννινα.

Συνοπτικά, η συστάδα 2 αποτελείται από νομούς με σημαντικό πρωτογενή τομέα, που γεωγραφικά περιλαμβάνουν πρακτικά όλες τις μεγάλες πεδινές εκτάσεις της Χώρας. Αντίστοιχα η συστάδα 5 περιλαμβάνονται νομοί ορεινοί, με ανεπτυγμένη κτηνοτροφία και όχι τόσο την Γεωργία. Τέλος η συστάδα 4 περιλαμβάνει τις πλουσιότερες περιοχές της Ελλάδος και βασίζονται στην παροχή υπηρεσιών για την ευμάρειά τους.

Θεωρούμε ότι μία πρόταση λοιπόν για περαιτέρω έρευνά θα ήταν η μελέτη σκοπιμότητας της δημιουργίας ενναλακτικών δομών που θα μπορούν να παίρνουν στοχευμένες διοικητικές και οικονομικές αποφάσεις σε περιοχές με όμοια

χαρακτηριστικά όπως αυτά προαναφέρθηκαν μια και οι περιοχές αυτές έχουν τις ίδιες ανάγκες.

Ακόμα θα ήταν δόκιμο να γίνουν συμπράξεις μεταξύ περιφερειών που έχουν νομούς με παρόμοια χαρακτηριστικά (π.χ. Κρήτης και Κεντρικής Μακεδονίας) για την από κοινού μελέτη των προβλημάτων και των δυνατοτήτων ανάπτυξης τους με κοινές επιτροπές και ομάδες εργασίας.

Βιβλιογραφία

1. Adnan, M., 2011. Towards real-time geodemographic information systems: design, analysis and evaluation. PhD Thesis. Department of Geography, University College London.
2. Aldenderfer M.S. & Blashfield R.K.,1984, Cluster Analysis. Sage Publications, Newbury Park, pp 88
3. Alsabti K., Ranka S. and Singh V., 1998, An efficient k-means clustering algorithm, First Workshop High Performance Data Mining, Orlando, Florida, Mar. 1998
4. Anderberg, M., 1973. Cluster analysis for applications. Academic Press, New York
5. Andrew, Ng, 2012. Clustering with the K-Means Algorithm, Machine Learning,
6. Arribas-Bel, D., 2019. Geographic Data Science (Lectures). University of Liverpool. Το άρθρο ήταν διαθέσιμο στην ιστοσελίδα: http://darribas.org/gds19/ και η πρόσβαση σε αυτό πραγματοποιήθηκε στις 17/9/2019.
7. Berman, O. & Krass, D., 2002. The generalized maximal covering location problem. Computer Operation Res 29(6). pp. 563–581.
8. Bezdek, J., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press.
9. Burns, L., Linda See, L.,Heppenstall, A. & Birkin, M., 2018. Developing an Individual-level Geodemographic Classification. Applied Spatial Analysis and Policy, Volume 11, Issue 3, pp. 417–437.
10. Dabbura, I. K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. Το άρθρο ήταν διαθέσιμο στην ιστοσελίδα: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a και η πρόσβαση σε αυτό πραγματοποιήθηκε στις 10/11/2019.
11. Davies. E., 1997. Machine Vision: Theory, Algorithms, Practicalities. Academic Press, 2nd Edition,.
12. Debenham, J. 2002. Understanding Geodemographic Classification: Creating The Building Blocks For An Extension. Working Paper. School of Geography, University of Leeds.
13. Different Types of Clustering Algorithms. Το άρθρο ήταν διαθέσιμο στις ιστοσελίδες: https://www.geeksforgeeks.org/different-types-clustering-algorithm/ και https://www.javatpoint.com/data-mining-different-types-of-clustering και η

πρόσβαση σε αυτές πραγματοποιήθηκε στις 1/11/2019 και 3/11/2019 αντίστοιχα.

14. Estivill-Castro, V. & Houle, M., 2001. Robust distance-based clustering with applications to spatial data mining. Algorithmica 30(2), pp. 216–242.
15. Fischer, M. M., 2015. Spatial Analysis in Geography. In International Encyclopedia of the Social & Behavioral Sciences (Second Edition).
16. Grubesic, T. H., Wei, R. & Murray, A.T., 2014. Spatial Clustering Overview and Comparison: Accuracy, Sensitivity, and Computational Expense. Journal Annals of the Association of American Geographers. Volume 104, Issue 6. Pages 1134-1156.
17. Guptill.S.C., 2001. Spatial Data. In International Encyclopedia of the Social & Behavioral Sciences.
18. Haining, R., 2015. Spatial Autocorrelation. In International Encyclopedia of the Social & Behavioral Sciences (Second Edition).
19. Hamerly, G. and Elkan, C., 2002. Alternatives to the K-means Algorithm that Find Better Clusterings. In Proceedings of the ACM Conference on Information and Knowledge Management (CIKM-2002), pp. 600-607.
20. Harris, R. J., Sleight, P. and Webber, R.J., 2005. Geodemographics GIS and Neighbourhood Targeting, London: Wiley.
21. Hartigan, J.A., 1975. Clustering Algorithms. John Wiley & Sons, New York.
22. Javadi M. & Shahrab J., 2014. New spatial clustering-based models for optimal urban facility location considering geographical obstacles. Journal of Industrial Engineering International.
23. Johnston, R.J., 2001). Geography. In International Encyclopedia of the Social & Behavioral Sciences.
24. Kaufman L., & Rousseeuw P., 1990. Finding Groups in Data: An Introduction to Cluster Analysis, New York: J. Wiley & Son.
25. MacQueen, J., 1967. Some Methods for Classification and Analysis of Multivariate Observations. In Proceedings Fifth Berkeley Symposium on Mathematics, Statistics and Probability, vol. 1, pp. 281-297.
26. McLachlan, G. and Krishnan, T., 1997. The EM algorithm and Extensions. John Wiley & Sons, Inc.
27. Milligan G. W., 1980, An examination of the effect of six types of error perturbation on fifteen clustering algorithms, Psychometrika, Vol 45, pp 325-342, 1980

28. Milligan, G. W., 1996. Clustering validation: Results and implications for applied analyses. In Arabie, P., Hubert, L. J. and De Soete, G. Eds., Clustering and Classification. Singapore: World Scientific.
29. Mirkin B., 2006, Different Perspectives at Clustering: The Number-of-Clusters Case, School of Computer Science Birkbeck College, University of London, IFCS 2006, pp 70.
30. Nisar M.U., Fard A. and Schardl Tao B., 2010, A work-efficient parallel breadthfirst search algorithm (or how to cope with the nondeterminism of reducers). In Parallelism in algorithms and architectures, 22nd ACM symposium, Thira, Santorini, Greece, June 13 – 15, 2010, Proceedings of the 22nd ACM symposium on, pp 303-314
31. Olkin, I. & Sampson, A.R., 2001. Multivariate Analysis: Overview. In International Encyclopedia of the Social & Behavioral Sciences.
32. Omran, M., Engelbrecht, A. & Salman, A., 2007. An overview of clustering methods. Intelligent Data Analysis. 11, pp. 583-605.
33. O' Sullivan, D., & Unwin, D.J. 2010. Geographic information analysis. New Jersey: John Wiley & Sons.
34. Rey, S. J., 2015. Mathematical Models in Geography. In International Encyclopedia of the Social & Behavioral Sciences (Second Edition).
35. Sayer, A., 1992. Method in Social Science: A Realist Approach. London: Hutchinson University Library (Second Edition).
36. Shashi, S., & Hui, X., 2008. Uk Geodemographic Classifications Review. Encyclopedia of GIS,USA:Springer. pp 347-355.
37. Singleton, A., Pavlis, M. & Longley, P. A., 2016. The stability of geodemographic cluster assignments over an intercenal period. Journal of Geographical Systems, Volume 18, Issue 2, pp. 97–123.
38. Smyth, P., 1996. Clustering using Monte Carlo cross validation. In Proc. 2nd International Conference Knowledge Discovery & Data Mining (KDD-96), Portland.
39. Son, L.H., 2014. Enhancing clustering quality of geo-demographic analysis using context fuzzy clustering type-2 and particle swarm optimization. Applied Soft Computing. Volume 22, pp. 566-584.
40. Trupti M., Kodinariya &. Prashant R., 2013. Makwana Review on determining number of Cluster in K-Means Clustering. International Journal of Advance Research in Computer Science and Management Studies. Volume 1, Issue 6, pp. 90-95

41. Turi, R.H., 2001., Clustering-Based Colour Image Segmentation, PhD Thesis. Monash University, Australia,.
42. Umawat, A. & Srivastava, S., 2015. Searching in Geographical Dataset by Using Modified k-mean Clustering Algorithm. International Bulletin of Mathematical Research Volume 2, Issue 1, pp. 236-240, ISSN: 2394-7802.
43. Υπουργείο Εσωτερικών. (2018) Διαθέσιμο στο: http://www.ypes.gr/el/Regions/programma/
[πρόσβαση: 1η Ιανουαρίου 2018]
44. Πανταζής Παναγιώτης. 2005. ΧΩΡΟΣ ΚΑΙ ΠΟΣΟΤΙΚΕΣ ΑΝΑΛΥΣΕΙΣ – ΜΕΘΟΔΟΙ ΚΑΙ ΤΕΧΝΙΚΕΣ (ΣΗΜΕΙΩΣΕΙΣ). http://library.tee.gr/digital/kma/kma_m1256/kma_m1256_pantazis.pdf
45. Πολυξένη Ηλιοπούλου. 2017. Γεωγραφική Ανάλυση. Εθνικό Μετσόβιο Πολυτεχνείο. ΣΥΝΔΕΣΜΟΣ ΕΛΛΗΝΙΚΩΝ ΑΚΑΔΗΜΑΪΚΩΝ ΒΙΒΛΙΟΘΗΚΩΝ/Heal Link.
46. Κουτσόπουλος, Κ. 2009. Πραγματεία ανάλυσης χώρου (Τόμος Α΄ Θεωρία, Τόμος Β΄ Μέθοδοι). Αθήνα: Παπασωτηρίου
47. Παππάς Β. 2000. Μεθοδολογική προσέγγιση υπολογισμού χωρικών δεικτών. Παραδείγματα από τον ελλαδικό χώρο. Στο Δεκαεπτά κείμενα για την πόλη και την ανάπτυξη. Βόλος: Τμήμα Μηχανικών Χωροταξίας και Περιφερειακής Ανάπτυξης, Πανεπιστήμιο Θεσσαλίας.
48. Τζαμπαλάτης, Ν. & Δασύρα Ζ., 2011. Συσταδοποίηση καταναλωτών ηλεκτρικής ενέργειας βάσει ομοιότητας αντιπροσωπευτικών καμπυλών. Πτυχιακή Εργασία. Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Πολυτεχνική Σχολή, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών (Τομέας Ηλεκτρικής Ενέργειας)

Παραρτήματα

Παράρτημα 1: Οι περιφερειακές ενότητες της κάθε περιφέρειας

Περιφέρεια	**Περιφερειακή Ενότητα**
1. Θεσσαλίας	1.1. Καρδίτσας
	1.2. Λάρισας
	1.3. Μαγνησίας
	1.4. Σποράδων
	1.5. Τρικάλων
2. Περιφέρεια Νοτίου Αιγαίου	2.1. Άνδρου
	2.2. Μυκόνου
	2.3. Θήρας
	2.4. Νάξου
	2.5. Καλύμνου
	2.6. Πάρου
	2.7. Καρπάθου
	2.8. Ρόδου
	2.9. Κέας-Κύθνου
	2.10. Σύρου
	2.11. Κω
	2.12. Τήνου
	2.13. Μήλου
3. Περιφέρεια Πελοποννήσου	3.1. Αργολίδος
	3.2. Αρκαδίας
	3.3. Κορινθίας
	3.4. Λακωνίας
	3.5. Μεσσηνίας
4. Περιφέρεια Δυτικής Ελλάδας	4.1. Αιτωλοακαρνανίας
	4.2. Αχαΐας
	4.3. Ηλείας
5. Περιφέρεια Κεντρικής Μακεδονίας	5.1. Ημαθίας
	5.2. Θεσσαλονίκης
	5.3. Κιλκίς
	5.4. Πέλλας
	5.5. Πιερίας
	5.6. Σερρών
	5.7. Χαλκιδικής
6. Περιφέρεια Βορείου Αιγαίου	6.1. Ικαρίας
	6.2. Λέσβου
	6.3. Λήμνου
	6.4. Σάμου
	6.5. Χίου
7. Περιφέρεια Ηπείρου	7.1. Άρτας
	7.2. Θεσπρωτίας
	7.3. Ιωαννίνων
	7.4. Πρέβεζας
8. Περιφέρεια Ανατολικής Μακεδονίας-Θράκης	8.1. Δράμας
	8.2. Έβρου
	8.3. Θάσου
	8.4. Καβάλας
	8.5. Ξάνθης
	8.6. Ροδόπης
9. Περιφέρεια Αττικής	9.1. Βορείου Τομέα
	9.2. Δυτικού Τομέα
	9.3. Κεντρικού Τομέα
	9.4. Νοτίου Τομέα
	9.5. Πειραιώς
	9.6. Νήσων

	9.7. Ανατολικής Αττικής
	9.8. Δυτικής Αττικής
10. Περιφέρεια Δυτικής Μακεδονίας	10.1. Γρεβενών
	10.2. Καστοριάς
	10.3. Κοζάνης
	10.4. Φλώρινας
11. Περιφέρεια Ιονίων Νήσων	11.1. Ζακύνθου
	11.2. Ιθάκης
	11.3. Κέρκυρας
	11.4. Κεφαλληνίας
	11.5. Λευκάδας
12. Περιφέρεια Στερεάς Ελλάδας	12.1. Βοιωτίας
	12.2. Εύβοιας
	12.3. Ευρυτανίας
	12.4. Φθιώτιδας
	12.5. Φωκίδας
13. Περιφέρεια Κρήτης	13.1. Ηρακλείου
	13.2. Λασιθίου
	13.3. Ρεθύμνης
	13.4. Χανίων

Πηγή: Υπουργείο Εσωτερικών (2018)

Παράρτημα 2

Νομοί	Κατά κεφαλή ΑΕΠ	Απασχόληση %	Ακαθαριστη προστιθεμενη αξια ανα εργαζομενο	Κατανάλωση πετρελαιοειδών ανα 1000 κατοίκους	Μηνιαία δαπανη νοικοκυριού	Εκτάσεις και παραγωγή (στη γεωργία)	Εκμεταλλεύσεις και αριθμός ζώων (στη κτηνοτροφία)
Αττική	22.204	41,46	19380,05376	3,348852773	####	65	12
Λέσβος, Λήμνος	12.325	34,29	10758,09406	24,50139705	876	87	34
Ικαρία, Σάμος	12.433	34,29	10852,23406	25,40431897	956	35	46
Χίος	12.019	34,29	10490,26833	6,436687252	####	86	25
Κάλυμνος, Κάρπαθος, Κως, Ρόδος	15.411	43,29	13451,04581	3,120467211	987	46	26
Άνδρος, Θήρα, Κέα, Μήλος, Μύκονος, Νάξος, Πάρος, Σύρος, Τήνος	21.660	43,29	18905,97177	3,590522909	####	87	45
Ηράκλειο	13.538	39,10	11816,39899	8,510911631	####	345	78
Λασίθι	14.956	39,10	13054,04205	22,79462047	865	265	65
Ρεθύμνο	13.432	39,10	11723,50979	15,33637009	987	287	73
Χανιά	14.019	39,10	12236,30692	5,511999295	879	345	68
Έβρος	12.403	34,84	10825,66276	6,666531306	654	456	89
Ξάνθη	10.015	34,84	8741,749488	9,313891513	875	365	23
Ροδόπη	10.026	34,84	8751,061405	5,843877332	456	543	34
Δράμα	10.466	34,84	9134,842992	3,569291712	654	235	62

Θάσος, Καβάλα	13.410	34,84	11704,54972	4,849581408	765	154	34
Ημαθία	10.827	34,78	9450,338491	9,506980032	687	265	61
Θεσσαλονίκη	14.023	34,78	12239,71742	7,632789345	####	124	24
Κιλκίς	11.408	34,78	9957,250652	2,920633342	654	265	46
Πέλλα	11.324	34,78	9883,717795	4,710254527	745	435	64
Πιερία	10.816	34,78	9440,164945	2,778898643	634	176	36
Σέρρες	9.863	34,78	8609,170251	1,569812563	698	265	52
Χαλκιδική	14.168	34,78	12366,25899	4,100104087	876	145	36
Γρεβενά	9.989	33,09	8719,071291	12,82593834	765	321	72
Κοζάνη	15.632	33,09	13644,08987	2,771124603	687	265	56
Καστοριά	10.765	33,09	9396,181915	5,522674225	745	234	45
Φλώρινα	16.828	33,09	14688,40941	5,026723093	687	354	28
Άρτα	10.958	34,55	9564,170328	1,88572217	834	312	72
Πρέβεζα	12.089	34,55	10552,00542	3,04072148	635	123	61
Θεσπρωτία	11.792	34,55	10292,24363	2,791996277	756	165	72
Ιωάννινα	11.999	34,55	10473,00747	3,897520247	934	365	39
Καρδίτσα	9.870	36,14	8614,964532	4,379468023	989	432	89
Τρίκαλα	11.254	36,14	9822,929417	4,999044525	####	542	79
Λάρισα	14.375	36,14	12546,8309	2,772460462	####	436	68
Μαγνησία	12.676	36,14	11063,97513	2,751085313	923	316	56
Ζάκυνθος	18.054	40,23	15758,41215	8,098987627	569	154	35
Κέρκυρα	15.048	40,23	13134,41722	5,280098017	897	542	28
Ιθάκη, Κεφαλληνία	14.520	40,23	12673,12278	9,282765341	875	254	46
Λευκάδα	12.041	40,23	10510,14362	22,67034068	923	156	26
Αιτωλοακαρνανία	11.509	34,46	10045,37619	3,145819718	794	432	64
Αχαΐα	13.004	34,46	11350,08373	2,419444974	897	412	73
Ηλεία	10.941	34,46	9549,374367	4,111396536	921	523	72
Βοιωτία	20.786	34,73	18142,60131	3,57855847	784	365	47
Εύβοια	13.322	34,73	11628,10519	2,550433292	687	512	78
Ευρυτανία	10.190	34,73	8894,249845	12,06185567	564	373	48
Φθιώτιδα	13.689	34,73	11947,88693	1,652088801	634	381	53
Φωκίδα	10.530	34,73	9191,098452	5,73437795	623	387	45
Αργολίδα	14.882	37,24	12989,74033	7,853645015	745	365	56
Αρκαδία	16.015	37,24	13978,27502	6,488152848	872	534	63
Κορινθία	13.501	37,24	11784,34064	4,273417994	945	489	53
Λακωνία	12.005	37,24	10478,49555	4,021373878	789	412	43
Μεσσηνία	12.475	37,24	10888,91807	2,730922689	698	532	38

Πηγή: ΕΛΣΤΑΤ 2016