ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΑΓΡΟΝΟΜΩΝ ΤΟΠΟΓΡΑΦΩΝ ΜΗΧΑΝΙΚΩΝ
ΤΟΜΕΑΣ ΤΟΠΟΓΡΑΦΙΑΣ

ΟΜΑΔΟΠΟΙΗΣΗ ΑΡΙΘΜΗΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

ΒΥΡΩΝΑΣ ΝΑΚΟΣ

ΑΘΗΝΑ 2006

Περιεχόμενα

Ομαδοποίηση αριθμητικών δεδομένων

1. Εισαγωγή

Οι διάφορες μέθοδοι ομαδοποίησης δεδομένων είναι τρόποι ταξινόμησης ενός συνόλου αριθμητικών δεδομένων σε διαδοχικές ομάδες δεδομένων, οι οποίες περιέχουν αριθμό παρατηρήσεων όσο το δυνατόν περισσότερο «γεωγραφικά» ισοδύναμο. Ο αριθμός των ομάδων εξαρτάται στενά από το όριο διαφοροποίησης της οπτικής μεταβλητής των συμβόλων που υλοποιούν την οπτικοποίηση των ομάδων. Στους χωροπληθείς χάρτες η χρησιμοποιούμενη οπτική μεταβλητή είναι η ένταση. Το ανθρώπινο μάτι λόγω της οπτικής αντίληψης μπορεί να διαβάζει αποτελεσματικά από το χάρτη πέντε έως το πολύ οκτώ διαφορετικούς τόνους του γκρι ή εντάσεις μιας απόχρωσης ανάλογα με το μέσο απόδοσης (π.χ. εκτυπωτή ψεκασμού μελάνης ή μονάδα εκτύπωσης τετραχρωμίας).

Σε ορισμένες περιπτώσεις τα όρια των ομάδων των δεδομένων πρέπει να συμπίπτουν με τιμές των φαινομένων που προέρχονται από εξωγενείς παράγοντες της διαδικασίας της απεικόνισης (κρίσιμες τιμές). Οι παράγοντες αυτοί εξαρτώνται από το χαρακτήρα του απεικονιζόμενου γεωγραφικού φαινομένου (π.χ. το εισόδημα που χαρακτηρίζει το επίπεδο πτώχευσης). Γενικότερα, το πρόβλημα της ομαδοποίησης εστιάζεται στην ταξινόμηση ενός συνόλου δεδομένων σε διακριτές ομάδες ανάλογα με την κατανομή που εμφανίζουν οι τιμές των δεδομένων. Κατά τη διαδικασία της ομαδοποίησης πρέπει να ταξινομούνται τα δεδομένα με τρόπο που οι τιμές τους να παρουσιάζουν ομοιογένεια μέσα στις ομάδες και σημαντικές διαφορές μεταξύ των ομάδων. Οι διάφορες μέθοδοι ομαδοποίησης δεδομένων κατατάσσονται σε τρεις μεγάλες κατηγορίες:

- Μέθοδοι σταθερών τιμές ή ίσων διαστημάτων,
- Μέθοδοι συστηματικά άνισων διαστημάτων,
- Μέθοδοι ακανόνιστων ή μεταβαλλόμενων διαστημάτων.

Παρά το γεγονός ότι έχει αναπτυχθεί μεγάλος αριθμός μεθόδων ομαδοποίησης δεδομένων στη χαρτογραφία, δεν είναι εύκολο να χαρακτηριστεί

μία από αυτές ως η «καλύτερη». Αντίθετα, κάθε φορά χρειάζεται να γίνει ανάλυση του συνόλου των δεδομένων με σκοπό να προσδιοριστεί η «βέλτιστη» μέθοδος για την ομαδοποίησή του συγκεκριμένου συνόλου δεδομένων.

2. Μέθοδοι σταθερών τιμών ή ίσων διαστημάτων

Στην κατηγορία αυτή περιλαμβάνονται τέσσερεις μέθοδοι για την ομαδοποίηση αριθμητικών δεδομένων.

2.1 Μέθοδος ίσων αριθμητικών διαστημάτων

Με τη μέθοδο αυτή, το διάστημα μεταξύ της μικρότερης και μεγαλύτερης τιμής των αριθμητικών δεδομένων υποδιαιρείται σε τμήματα ίσου εύρους ανάλογα με τον αριθμό των ομάδων. Είναι η πιο απλή μέθοδος ομαδοποίησης αριθμητικών δεδομένων τόσο στο στάδιο της εφαρμογής αλλά και στο στάδιο της ερμηνείας. Η μέθοδος αυτή είναι κατάλληλη για δεδομένα που εμφανίζουν ομοιόμορφη κατανομή. Τα όρια των διαστημάτων προσδιορίζονται από αριθμητικές προόδους με βάση τη σχέση:

$$a_n = a_1 + (n-1)\,\omega,$$

όπου a: τα όρια των διαστημάτων, n: ο αριθμός των ομάδων και ω: το εύρος των διαστημάτων.

2.2 Μέθοδος παραμέτρων κανονικής κατανομής

Η μέθοδος αυτή βασίζεται στις παραμέτρους της κανονικής κατανομής. Η μέση τιμή και η τυπική απόκλιση των αριθμητικών δεδομένων χρησιμοποιείται για τον ορισμό των ορίων των διαστημάτων της ομαδοποίησης. Η μέθοδος των παραμέτρων κανονικής κατανομής είναι κατάλληλη για δεδομένα που εμφανίζουν κανονική κατανομή.

2.3 Μέθοδος κανονικής τμηματοποίησης (quantiles)

Με τη μέθοδο αυτή το σύνολο των παρατηρήσεων των αριθμητικών δεδομένων ταξινομείται κατά αύξουσα ή φθίνουσα σειρά και στη συνέχεια υποδιαιρείται σε τμήματα με ίσο αριθμό παρατηρήσεων το καθένα. Η μέθοδος της κανονικής

τμηματοποίησης είναι κατάλληλη για δεδομένα που αναφέρονται σε ισοδύναμες επιφάνειες ή για δεδομένα που διαφοροποιούνται ως προς την κλίμακα τάξης. Η μέθοδος είναι ακατάλληλη για αριθμητικά δεδομένα που αναφέρονται σε επιφάνειες των οποίων το εμβαδόν διαφοροποιείται σημαντικά.

2.4 Μέθοδος ίσων διαστημάτων εμβαδού

Η μέθοδος των ίσων διαστημάτων εμβαδού είναι παραλλαγή της μεθόδου των ίσων αριθμητικών διαστημάτων με χρήση του εμβαδού των επιφανειών στις οποίες αναφέρονται τα δεδομένα ως βάρος για την ταξινόμηση. Με τη μέθοδο αυτή ταξινομούνται τα αριθμητικά δεδομένα κατά αύξουσα ή φθίνουσα σειρά και προσδιορίζονται τα όρια των διαστημάτων με τρόπο που να περιλαμβάνουν ισοδύναμες ως προς το εμβαδόν παρατηρήσεις. Ο προσδιορισμός των ορίων των διαστημάτων μπορεί να γίνει με τη βοήθεια του αθροιστικού διαγράμματος του εμβαδού.

3. Μέθοδοι συστηματικά άνισων διαστημάτων

Οι μέθοδοι των συστηματικά άνισων διαστημάτων βασίζονται στην εφαρμογή αναγωγικών σχέσεων ακολουθιών ή γεωμετρικών προόδων για τον προσδιορισμό των ορίων τους. Η εφαρμογή τους μπορεί να γίνει με τους ακόλουθους συνδυασμούς:

- Αύξουσες με σταθερό ρυθμό
- Αύξουσες με αυξανόμενο ρυθμό
- Αύξουσες με φθίνοντα ρυθμό
- Φθίνουσες με σταθερό ρυθμό
- Φθίνουσες με αυξανόμενο ρυθμό
- Φθίνουσες με φθίνοντα ρυθμό

Στις περιπτώσεις που τα αριθμητικά δεδομένα εμφανίζουν σημαντικές διαφοροποιήσεις τα όρια των διαστημάτων προσδιορίζονται από αναγωγικές σχέσεις ακολουθιών:

$$a_n = a_{n-1} + (n-1)\,\omega,$$

όπου a_i: τα όρια των διαστημάτων, n: ο αριθμός των ομάδων και ω: ο συντελεστής μεταβολής του εύρους των διαστημάτων.

Στις περιπτώσεις που τα αριθμητικά δεδομένα εμφανίζουν πολύ μεγάλες διαφοροποιήσεις τα όρια των διαστημάτων προσδιορίζονται από γεωμετρικές προόδους με βάση τη σχέση:

$b_n = b_1 \, \omega^{n-1}$

όπου b_i: τα όρια των διαστημάτων, n: ο αριθμός των ομάδων και ω: ο συντελεστής μεταβολής του εύρους των διαστημάτων.

4. Μέθοδοι ακανόνιστων ή μεταβαλλόμενων διαστημάτων

Οι μέθοδοι των ακανόνιστων ή μεταβαλλόμενων διαστημάτων διακρίνονται σε δύο κατηγορίες. Στην πρώτη κατηγορία, τα όρια των διαστημάτων προσδιορίζονται με τη βοήθεια γραφικών τεχνικών ενώ στη δεύτερη με επαναληπτικές τεχνικές.

4.1 Γραφικές τεχνικές ομαδοποίησης

Οι γραφικές τεχνικές στηρίζονται στην αξιοποίηση τεσσάρων διαγραμμάτων στατιστικής προέλευσης, με τη βοήθεια των οποίων προσδιορίζονται τα όρια των διαστημάτων. Τα χρησιμοποιούμενα διαγράμματα είναι:

- Το ιστόγραμμα:
  Στο διάγραμμα αυτό οι παρατηρήσεις απεικονίζονται στον άξονα: x και οι τιμές τους στον άξονα: y.
- Η κλινογραφική καμπύλη:
  Στην κλινογραφική καμπύλη απεικονίζεται το σύνολο των παρατηρήσεων στον άξονα: x και οι αθροιστικές τιμές τους στον άξονα: y. Κρίσιμα σημεία της κλινογραφικής καμπύλης είναι εκείνα στα οποία εμφανίζονται απότομες αλλαγές της κλίσης της. Τα κρίσιμα αυτά σημεία λαμβάνονται υπόψη για τον προσδιορισμό των ορίων των διαστημάτων.
- Το διάγραμμα συχνοτήτων:
  Στο διάγραμμα συχνοτήτων στον άξονα: x απεικονίζονται ομαδοποιημένες οι τιμές των αριθμητικών δεδομένων ενώ στον άξονα: y οι συχνότητες που εμφανίζουν οι αντίστοιχες ομαδοποιημένες τιμές. Το διάγραμμα αυτό αναπαριστά με παραστατικό τρόπο τα αριθμητικά χαρακτηριστικά της κατανομής των δεδομένων.
- Το διάγραμμα των αθροιστικών συχνοτήτων:

Στο αθροιστικό διάγραμμα των συχνοτήτων στον άξονα: x απεικονίζονται ομαδοποιημένες οι τιμές των αριθμητικών δεδομένων ενώ στον άξονα: y οι αθροιστικές συχνότητες που εμφανίζουν οι ομαδοποιημένες τιμές. Κρίσιμα σημεία του διαγράμματος αθροιστικών συχνοτήτων αποτελούν οι «οροφές» και τα «πατώματα» των χαρακτηριστικών μοτίβων που παρουσιάζει η καμπύλη, δεδομένου ότι εκφράζουν τις συσσωρεύσεις των τιμών των δεδομένων.

4.2 Επαναληπτικές τεχνικές ομαδοποίησης

Οι επαναληπτικές τεχνικές βασίζονται σε στατιστική επεξεργασία των αριθμητικών δεδομένων. Πιο συγκεκριμένα, ο προσδιορισμός των ορίων των διαστημάτων της ομαδοποίησης προκύπτει με την εφαρμογή ενός στατιστικού κριτηρίου. Συνήθως χρησιμοποιούνται δύο κριτήρια:

- Το κριτήριο της βέλτιστης προσαρμογής της μεταβλητότητας.
  (Goodness of the variance fit – GVF).
  Με το κριτήριο της βέλτιστης προσαρμογής της μεταβλητότητας προσδιορίζονται τα όρια των διαστημάτων της ομαδοποίησης με ελαχιστοποίηση των τετραγώνων των αποκλίσεων από τη μέση τιμή του κάθε διαστήματος.
- Το κριτήριο της βέλτιστης προσαρμογής της απόλυτης απόκλισης.
  (Goodness of absolute deviation fit - GADF)
  Με το κριτήριο της βέλτιστης προσαρμογής της απόλυτης απόκλισης προσδιορίζονται τα όρια των διαστημάτων της ομαδοποίησης με μεγιστοποίηση της βέλτιστης προσαρμογής των απολύτων αποκλίσεων από τη μέση τιμή του κάθε διαστήματος.

Το μεγαλύτερο πλεονέκτημα των επαναληπτικών τεχνικών αποτελεί η μεγιστοποίηση της ομοιογένειας που επιτυγχάνεται στις παρατηρήσεις που κατατάσσονται σε κάδε διάστημα της ομαδοποίησης και η σημαντική διαφοροποίηση των παρατηρήσεων που κατατάσσονται σε διαφορετικά διαστήματα της ομαδοποίησης, όπως καθορίζονται από το στατιστικό κριτήριο που έχει εφαρμοστεί.

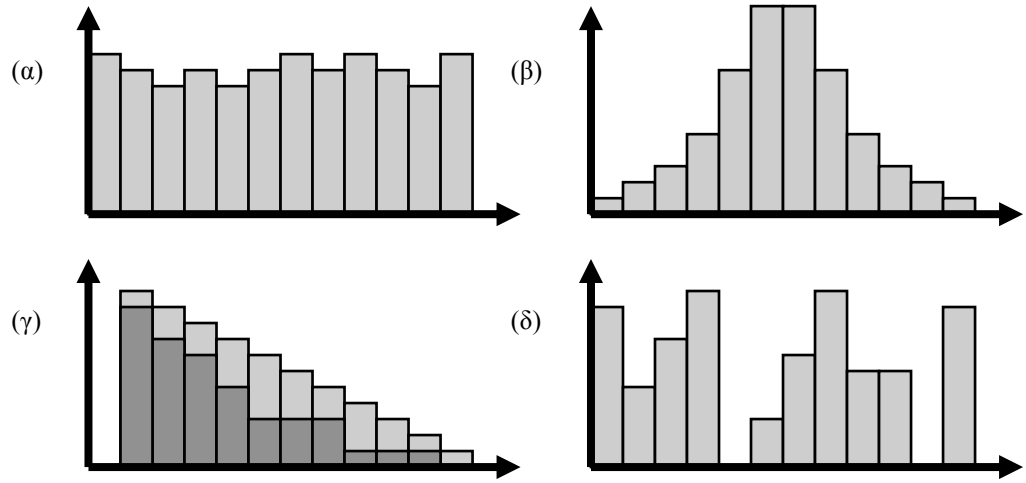
5. Μεθοδολογία ομαδοποίησης δεδομένων

Μια μεθοδολογία ομαδοποίησης αριθμητικών δεδομένων πρέπει να στοχεύει στον προσδιορισμό της καταλληλότερης μεθόδου για της ομαδοποίηση ενός

συγκεκριμένου συνόλου δεδομένων. Ο προσδιορισμός των ορίων των διαστημάτων ομαδοποίησης πρέπει να λαμβάνει υπόψη τους ακόλουθους παράγοντες:

- Τις κρίσιμες τιμές
- Τα όρια των διαστημάτων να έχουν στρογγυλεμένες τιμές
- Να εφαρμόζονται εναλλακτικές μέθοδοι
- Σε περιπτώσεις παραγωγής σειράς θεματικών χαρτών να εφαρμόζεται ενιαία τεχνική ομαδοποίησης για όλους τους χάρτες
- Να αξιοποιούνται τα διαγράμματα (κυρίως το ιστόγραμμα και το διάγραμμα κατανομής συχνοτήτων)

Κάθε διαδικασία ομαδοποίησης προϋποθέτει την ύπαρξη ενός συγκεκριμένου συνόλου αριθμητικών δεδομένων και τον επιθυμητό αριθμό διαστημάτων για την ταξινόμησή τους. Η επιλογή της καταλληλότερης μεθόδου ομαδοποίησης κυρίως έχει σχέση με την κατανομή που εμφανίζουν οι τιμές των αριθμητικών δεδομένων. Για το λόγο αυτό η κατασκευή του διαγράμματος των συχνοτήτων είναι πολύ χρήσιμη. Με την ερμηνεία της μορφής του διαγράμματος συχνοτήτων του εκάστοτε συνόλου αριθμητικών δεδομένων μπορεί να επιλεγεί η καταλληλότερη μέθοδος ή ένας κατάλληλος συνδυασμός μεθόδων. Οι πλέον χαρακτηριστικές μορφές ενός διαγράμματος συχνοτήτων (Σχήμα 1) είναι οι ακόλουθες:



Σχήμα 1. Τα τέσσερα πρότυπα κατανομών του διαγράμματος των συχνοτήτων. (α) ομοιόμορφη, (β) κανονική, (γ) στρεβλή (γραμμική/μη-γραμμική) και (δ) ακανόνιστη

- Ομοιόμορφη
- Κανονική
- Στρεβλή (γραμμική/μη-γραμμική)
- Ακανόνιστη

Σύμφωνα με την ερμηνεία του διαγράμματος συχνοτήτων μπορούν να εφαρμοστούν οι ακόλουθοι μέθοδοι ομαδοποίησης:

1. Το διάγραμμα συχνοτήτων αναπαριστά ομοιόμορφη κατανομή. Οι καταλληλότερες μέθοδοι ανήκουν στην κατηγορία των σταθερών τιμών ή ίσων διαστημάτων.
2. Το διάγραμμα συχνοτήτων αναπαριστά κανονική κατανομή. Η καταλληλότερη μέθοδος είναι η μέθοδος των παραμέτρων της κανονικής κατανομής.
3. Το διάγραμμα συχνοτήτων αναπαριστά γραμμική στρεβλή κατανομή. Ο καταλληλότερος τρόπος για τον προσδιορισμό των ορίων των διαστημάτων είναι η εφαρμογή αριθμητικών προόδων. Η μέθοδος αυτή ανήκει στην κατηγορία των συστηματικά άνισων διαστημάτων.
4. Το διάγραμμα συχνοτήτων αναπαριστά μη-γραμμική στρεβλή κατανομή. Ο καταλληλότερος τρόπος για τον προσδιορισμό των ορίων των διαστημάτων είναι η εφαρμογή γεωμετρικών προόδων. Η μέθοδος ανήκει στην κατηγορία των συστηματικά άνισων διαστημάτων.
5. Το διάγραμμα συχνοτήτων αναπαριστά ακανόνιστη κατανομή. Στις περιπτώσεις αυτές είναι χρήσιμη και η ερμηνεία του ιστογράμματος των δεδομένων. Η προσοχή κατά την ερμηνεία πρέπει να εστιαστεί στον πιθανό εντοπισμό φυσικών διακοπών των τιμών των δεδομένων. Εάν εμφανίζονται φυσικές διακοπές στο ιστόγραμμα των δεδομένων αυτές πρέπει να λαμβάνονται υπόψη στον προσδιορισμό των ορίων των διαστημάτων της ομαδοποίησης. Στη συνέχεια, εξετάζεται αν κάποια από τις προαναφερθείσες μεθόδους ή συνδυασμός τους μπορεί να εφαρμοστεί συνολικά ή τμηματικά. Τέλος, στις περιπτώσεις που δεν είναι δυνατό να εφαρμοστεί καμία από τις παραπάνω μεθόδους τότε τα όρια των διαστημάτων της ομαδοποίησης προσδιορίζονται με τη μέθοδο της κανονικής τμηματοποίησης.

6. Βιβλιογραφία

Dent, D.B. (1990). *Cartography. Thematic Map Design*. (2nd ed.). Wm C. Brown Pub., Dubuque, pp. 433.

Jenks, G.F. and M.R Coulson (1963). "Class intervals for statistical maps*"*. *International Yearbook of Cartography*, **3**, pp. 119-134.

Jenks, G.F. (1967). "The data model concept in statistical mapping". *International Yearbook of Cartography*, **7**, pp. 186-190.

Robinson, H.A., R.D. Sale, J.L. Morisson and Ph.C. Muehrcke (1985*). Elements of Cartography*. (5th ed.). John Willey & Sons, New York, pp. 544.

Yue-Hong Chou (????). *Exploring spatial analysis in Geographic Information Systems*. ISBN 1-56690-119-7.