ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΑΓΡΟΝΟΜΩΝ & ΤΟΠΟΓΡΑΦΩΝ ΜΗΧΑΝΙΚΩΝ

ΤΟΜΕΑΣ ΤΟΠΟΓΡΑΦΙΑΣ – ΕΡΓΑΣΤΗΡΙΟ ΧΑΡΤΟΓΡΑΦΙΑΣ

Διδακτορική Διατριβή

ΑΝΑΠΤΥΞΗ ΜΕΘΟΔΩΝ ΚΑΙ ΤΕΧΝΙΚΩΝ ΑΠΟΚΤΗΣΗΣ ΓΕΩΓΡΑΦΙΚΗΣ ΓΝΩΣΗΣ

Σοφία Κονταξάκη

Αθήνα, 2010

Στα παιδιά μου Αναστασία και Κάλλια

… και στη μνήμη του γιου μου Αντρέα

ΕΥΧΑΡΙΣΤΙΕΣ

Με την ολοκλήρωση της διδακτορικής μου διατριβής, θα ήθελα να ευχαριστήσω θερμότατα όλους όσους μου συμπαραστάθηκαν και με βοήθησαν.

Αρχίζοντας, θέλω ολόψυχα να ευχαριστώ τον επιβλέποντα καθηγητή μου, κ. Μαρίνο Κάβουρα, Καθηγητή του Εθνικού Μετσόβιου Πολυτεχνείου της Σχολής Αγρονόμων Τοπογράφων Μηχανικών, ο οποίος μου έδωσε το αρχικό ερέθισμα για την ενασχόληση μου μ' ένα τόσο ενδιαφέρον και πρωτότυπο θέμα, και με βοήθησε δίνοντάς μου πολύτιμες συμβουλές καθ' όλη την εκπόνηση της παρούσας διδακτορικής διατριβής. Θέλω ακόμα να τον ευχαριστήσω για τη συμπαράσταση και την εμπιστοσύνη που μου έδειξε όχι μόνο ως καθηγητής αλλά και ως πολύτιμος φίλος σε δύσκολες για μένα στιγμές.

Ακόμη, θα ήθελα ιδιαιτέρως να ευχαριστήσω τη συνεργάτιδά μου, κ. Μαργαρίτα Κόκλα για την αμέριστη υποστήριξη και ενθάρρυνσή της σε αυτό το δύσκολο έργο. Ευχαριστώ ακόμα την κ. Ελένη Τομαή, η οποία με βοήθησε πρόθυμα στην παρουσίαση των εργασιών που εκπονήθηκαν στο πλαίσιο της παρούσας διατριβής, καθώς και τα μέλη της ερευνητικής ομάδας *OntoGEO Group*, κ. Γιώργο Πανόπουλο, κ. Θανάση Καραλόπουλο και κ. Νάνσυ Δάρρα, για την άψογη συνεργασία μας.

Τέλος, επειδή χωρίς την παρουσία, την υποστήριξη και την ανεκτικότητα κάποιων ανθρώπων, δεν θα ήταν δυνατή η υλοποίηση της παρούσας διδακτορικής διατριβής, θα ήθελα να ευχαριστήσω το σύζυγό μου, Γιάννο, για την ψυχική και επιστημονική υποστήριξή του, τα παιδιά μου, Αναστασία και Κάλλια, που με βοήθησαν με την ενθαρρυντική παρουσία τους, και τους γονείς μου, Χρήστο και Καλλιόπη, που στέκονται πάντα ακούραστα δίπλα μου.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΣΥΝΟΨΗ

Το θέμα της παρούσας διδακτορικής διατριβής είναι η ανάπτυξη μεθόδων και τεχνικών απόκτησης γεωγραφικής γνώσης. Για την επίτευξη του στόχου αυτού, ανασκοπείται η πρόοδος σε συναφείς τομείς όπως είναι η αναπαράσταση γεωγραφικής γνώσης, η επεξεργασία φυσικής γλώσσας, η δημιουργία και χρήση ελεγχόμενων γλωσσών, η οπτικοποίηση πληροφοριών και ειδικότερα η χωρικοποίηση. Ακολούθως, εισάγονται τρεις καινοτόμες μέθοδοι και τεχνικές. Η πρώτη εξ αυτών στηρίζεται σε μια νέα ελεγχόμενη γλώσσα που επιτρέπει την απόσπαση γνώσης από βάσεις γεωχωρικών δεδομένων, με τρόπο αποτελεσματικό και ταυτόχρονα οικείο προς τον άνθρωπο. Στη συνέχεια, εισάγεται μια μέθοδος παραγωγής σημασιολογικών επιγραφών για ομάδες γεωγραφικών εννοιών οι οποίες περιγράφονται από ορισμούς σε φυσική γλώσσα. Αξιοποιώντας τη γνώση που εμπεριέχεται στους ίδιους τους ορισμούς και όχι σε εξωτερικές πηγές, η μέθοδος δημιουργεί σημασιολογικές επιγραφές που μοιάζουν με ορισμούς διατυπωμένους σε φυσική γλώσσα και που συνοψίζουν το περιεχόμενο των ορισμών των γεωγραφικών εννοιών κάθε ομάδας. Ακολούθως, η διδακτορική διατριβή εξετάζει ένα πρόβλημα που άπτεται του πεδίου της οπτικοποίησης πληροφοριών και ειδικότερα της χωρικοποίησης. Το πρόβλημα αυτό έγκειται στην απεικόνιση πολυδιάστατων πληροφοριών σε χώρους περιορισμένων διαστάσεων, όπως είναι ο γεωγραφικός, κάνοντας χρήση ειδικών τεχνικών προβολής και κατάλληλων χωρικών μεταφορών. Εισάγεται νέα τεχνική που αντιμετωπίζει το ζήτημα της χωρικοποίησης πληροφοριών σε πολλά επίπεδα λεπτομέρειας, η οποία ενσωματώνεται σ’ ένα πρωτότυπο γραφικό περιβάλλον που επιτρέπει την εξόρυξη γνώσης από συλλογές πολυδιάστατων πληροφοριών μεγάλου όγκου. Στο περιβάλλον αυτό, η νέα τεχνική χωρικοποίησης παράγει τρισδιάστατες επιφάνειες κάνοντας χρήση της μεταφοράς του τοπίου πληροφοριών, για την ανάδειξη ομάδων όμοιων πληροφοριών, και της μεταφοράς της ομαλότητας του τοπίου πληροφοριών, για την αναπαράσταση πολλών επιπέδων λεπτομέρειας.

ABSTRACT

The topic of the present dissertation is the development of methods and techniques for geographic knowledge acquisition. Initially, a survey is performed to review the work in progress regarding related fields such as geographical knowledge representation, natural language processing and controlled languages, information visualization and especially spatialization. In following, three novel methods and techniques are introduced. The first involves the creation of a new controlled language, which is used to derive spatial knowledge from geospatial databases in both human familiar and effective way. The purpose is to overcome the difficulties arising from the learning of specialized and hard to manipulate languages in order to formulate search queries. The second method consists in the creation of semantic labels for clusters of geographic concepts described by natural language definitions. The aim is to create labels by taking advantage of the semantic information immanent in the definitions of geographic concepts instead of resorting to external information. The resulting labels are structured so as to epitomize the meaning of the clusters and be readily understood by users in the context of semantic-based applications. The forth section of the dissertation concerns the field of information visualization and spatialization. It tackles both the issue of granularity and the way it is handled by existing spatialization methods. A new spatialization technique is introduced and integrated into a prototyping spatialization environment for representing, exploring, and extracting knowledge from large sets of multidimensional data at different levels of granularity, by making use of kernel density estimation and spatial metaphors.

1. Εισαγωγή

1.1 Θέμα Διδακτορικής Διατριβής

Το θέμα της παρούσας διδακτορικής διατριβής είναι η *ανάπτυξη μεθόδων και τεχνικών απόκτησης γεωγραφικής γνώσης*. Πρόκειται για σύνθετο τίτλο που εμπλέκει και συνδυάζει εξίσου σύνθετους όρους. Τι εννοούμε με την έκφραση *μέθοδοι και τεχνικές* ή με την έκφραση *απόκτηση γεωγραφικής γνώσης*; Ας εξετάσουμε κάθε έκφραση ξεχωριστά, ξεκινώντας από την πρώτη.

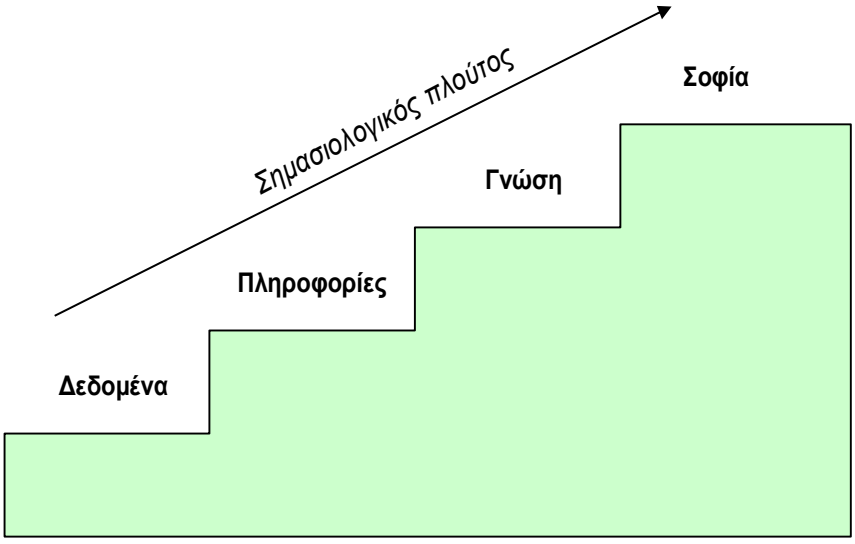
Συχνά οι όροι *μέθοδος* και *τεχνική* χρησιμοποιούνται ως συνώνυμα. Ωστόσο, προφανώς διαφέρουν. Η *μέθοδος* προέρχεται από τη σύζευξη των λέξεων *μετά* και *οδός*, που υποδεικνύει «το μεταβαίνειν προς αναζήτηση κάποιου πράγματος», στην περίπτωση της διατριβής αυτής, προς αναζήτηση της γνώσης. Η μέθοδος είναι δηλαδή ο συστηματικός και προγραμματισμένος τρόπος απόκτησης της γνώσης βάσει συγκεκριμένων κανόνων και βημάτων. Αντίθετα, η *τεχνική* αποτελεί ειδική διαδικασία που προσβλέπει στην τέλεση συγκεκριμένης εργασίας.

Το δεύτερο σκέλος του τίτλου, η έκφραση *απόκτηση γεωγραφικής γνώσης* χρήζει και αυτή διευκρίνισης. Ο όρος *απόκτηση* (acquisition) χρησιμοποιείται με την έννοια της *απόσπασης* (elicitation) γεωγραφικής γνώσης μέσα από την ανάλυση, την επεξεργασία και τη συσχέτιση πληροφοριών. Η απόκτηση εξειδικεύεται στην *εξόρυξη γνώσης* (knowledge extraction ή data mining), όταν αναφέρεται στη διαδικασία ανακάλυψης της γνώσης με τη βοήθεια ειδικών μεθόδων και τεχνικών οργάνωσης, ανάλυσης και οπτικοποίησης που αποσκοπούν στην *αναγνώριση προτύπων γνώσης* (knowledge patterns recognition) (εννοιών, σχέσεων, ομάδων, κ.ά.) μέσα από συλλογές δεδομένων, όπως π.χ. βάσεις δεδομένων, κείμενα, ιστοσελίδες, κλπ.

Ο όρος *γεωγραφική* χαρακτηρίζει τη γνώση που σχετίζεται με το γεωγραφικό χώρο δηλαδή με γεωγραφικά αντικείμενα, οντότητες και φαινόμενα καθώς και με αφηρημένα δεδομένα στα οποία αποδίδεται γεωγραφική αναφορά. Επειδή στην παρούσα διατριβή, όπως και στη γενικότερη βιβλιογραφία, εκτός του όρου *γεωγραφικός* (geographic)

χρησιμοποιούνται και οι όροι *χωρικός* (spatial) και *γεωχωρικός* (geospatial), διευκρινίζεται ότι: 1) ο όρος χωρικός θεωρείται ευρύτερος του όρου γεωγραφικός και αναφέρεται σε αντικείμενα/φαινόμενα οποιουδήποτε φυσικού χώρου, 2) η ουσιαστική διαφορά μεταξύ των όρων γεωγραφικός και χωρικός έγκειται στο ότι ο δεύτερος όρος χρησιμοποιείται περισσότερο όταν ενδιαφέρει η γεωμετρική διάσταση των αντικειμένων/φαινομένων, και 3) ο όρος γεωχωρικός, αναφέρεται σε οτιδήποτε χωρικό είναι και γεωγραφικό.

Ο όρος *γνώση* (knowledge) ορίζεται συχνά εν συγκρίσει με τα *δεδομένα* (data) και τις *πληροφορίες* (information). Τα δεδομένα αποτελούν συλλογές ακατέργαστων στοιχείων ενώ οι πληροφορίες είναι τα στοιχεία που προκύπτουν από τα δεδομένα μετά από την επεξεργασία τους. Η γνώση είναι σημασιολογικά πιο πλούσια από τις πληροφορίες. Θα έλεγε κανείς ότι πρόκειται για το αποτέλεσμα πνευματικής διαδικασίας για την κατανόηση της αντικειμενικής πραγματικότητας, η οποία βασίζεται στην εμπειρία και τη συλλογιστική (reasoning). Στη βιβλιογραφία, συναντάμε ακόμα τη *σοφία* (wisdom), ως σημασιολογικά πλουσιότερη της γνώσης (Σχήμα 1.1). Η σοφία ορίζεται περισσότερο ως πνευματική κατάσταση, η οποία βασίζεται στη συσσωρευμένη γνώση και περιλαμβάνει τη δυνατότητα λήψης αποφάσεων που αφορούν σε μελλοντικές δράσεις.



Σχήμα 1.1 Δεδομένα, πληροφορίες, γνώση, σοφία, και επίπεδο σημασιολογικού πλούτου

Πολλοί ορισμοί της γνώσης έχουν κατά καιρούς διατυπωθεί, οι οποίοι τονίζουν περισσότερο τη φιλοσοφική υπόσταση της γνώσης. Οι Davenport και Prusak [DP00] δίνουν έναν ορισμό της γνώσης που προσεγγίζει περισσότερο το πνεύμα της παρούσας διατριβής:

"Knowledge is a fluid mix of framed experience, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information."

Δηλαδή: «Η γνώση είναι ένα ρευστό μείγμα εμπειριών, αξιών, πληροφοριών σχετικών με κάποιο συγκεκριμένο πλαίσιο αναφοράς, και ενορατικότητας εμπειρογνωμόνων, η οποία παρέχει ένα πλαίσιο για την εκτίμηση και ενσωμάτωση νέων εμπειριών και πληροφοριών.»

Διευκρινίζεται τέλος ότι η γνώση μπορεί να είναι 1) *ρητή*, *σαφής* και *αντικειμενική* (explicit knowledge), οπότε και αποτελείται από οτιδήποτε είναι σαφώς καθορισμένο και δύναται να καταγραφεί, να κωδικοποιηθεί και να αρχειοθετηθεί με οποιοδήποτε τρόπο ή 2) *άρρητη*, *υποκειμενική* και *αφανής* (tacit knowledge). Σε αυτήν την περίπτωση δεν εκφράζεται άμεσα, αλλά εννοείται, όπως για παράδειγμα ένα σύνολο εμπειριών, παραστάσεων, πρακτικών κλπ. [Choo00], [BS01]. Στο πλαίσιο της παρούσας διατριβής, ενδιαφερόμαστε για την απόκτηση και εξόρυξη ρητής γνώσης.

1.2 Κύριες Ερευνητικές Κατευθύνσεις

Όπως υποδεικνύει ο τίτλος της, η παρούσα διδακτορική διατριβή έχει ως κύρια ερευνητική κατεύθυνση την απόκτηση γεωγραφικής γνώσης από συλλογές δεδομένων. Η κατεύθυνση αυτή συνδυάζεται με τη μελέτη θεμάτων που άπτονται άλλων συναφών επιστημονικών πεδίων, όπως είναι η αναπαράσταση γεωγραφικής γνώσης, η επεξεργασία φυσικής γλώσσας, η δημιουργία και χρήση ελεγχόμενων γλωσσών, η οπτικοποίηση πληροφοριών και ειδικότερα η χωρικοποίηση.

Ειδικότερα, όσον αφορά την απόκτηση γνώσης μέσα από διαδικασίες εξόρυξης γνώσης, ο Στεφανάκης [Στεφ03] διακρίνει, α) τις περιγραφικές διαδικασίες, οι οποίες ανακτούν τη γνώση που σχετίζεται με τις ιδιότητες των δεδομένων μιας συλλογής, και β) τις διαδικασίες πρόβλεψης, οι οποίες έχουν ως στόχο την πρόβλεψη της συμπεριφοράς (εξέλιξη της

κατάστασης, των τιμών τους, κλπ) δεδομένων στο χρόνο με τη βοήθεια ειδικών μοντέλων. Επιπλέον, ο ίδιος [Στεφ03] προτείνει μια πιο αναλυτική κατηγοριοποίηση των περιγραφικών διαδικασιών εξόρυξης γνώσης σε:

1) Διαδικασίες εύρεσης κανόνων συσχέτισης των κλάσεων των δεδομένων,

2) Διαδικασίες περιγραφής των κλάσεων των δεδομένων,

3) Διαδικασίες κατηγοριοποίησης των δεδομένων σε κλάσεις, κ.ά.

Από τις μεθόδους και τεχνικές που εισάγονται και περιγράφονται στην παρούσα διατριβή, η *Geo-Labeling* (τρίτο κεφάλαιο), εμπίπτει στη δεύτερη κατηγορία, δηλαδή στην κατηγορία των διαδικασιών περιγραφής των κλάσεων των δεδομένων. Πράγματι, έχοντας ως στόχο την παραγωγή σημασιολογικών επιγραφών για ομάδες γεωγραφικών εννοιών που περιγράφονται από ορισμούς σε φυσική γλώσσα, η Geo-Labeling σχηματίζει επιγραφές με τρόπον που να σχηματίζουν μια συνοπτική περιγραφή του σημασιολογικού περιεχομένου των ομάδων γεωγραφικών εννοιών.

Ακόμη, η τεχνική χωρικοποίησης που υλοποιήθηκε και εφαρμόζεται στο πρωτότυπο περιβάλλον χωρικοποίησης *GeoScape* (τέταρτο κεφάλαιο), επιχειρεί να αναδείξει ομάδες όμοιων δεδομένων σε διάφορα επίπεδα λεπτομέρειας επιλεγμένα από το χρήστη και ως εκ τούτου, εμπίπτει στην τρίτη κατηγορία διαδικασιών εξόρυξης γνώσης, δηλαδή αυτών που κατηγοριοποιούν τα δεδομένα σε κλάσεις.

Αντίθετα, η μέθοδος που περιγράφεται στο δεύτερο κεφάλαιο της παρούσας διατριβής και που προτείνει τη χρήση της ελεγχόμενης γλώσσας *Geo-Q* για την απόκτηση γνώσης, δεν εμπίπτει σε καμία από τις παραπάνω κατηγορίες διότι δεν αφορά στην εξόρυξη αλλά γενικότερα στην απόσπαση γνώσης από βάσεις γεωχωρικών δεδομένων. Πράγματι, εδώ πρέπει να γίνει διάκριση μεταξύ των απλών ερωτημάτων που υποβάλλονται προς μια βάση δεδομένων σε σχέση με τις ερωτήσεις εξόρυξης γνώσης. Στην πρώτη περίπτωση, τα ερωτήματα μεταφράζονται σε σχετικά λίγες και απλές συσχετίσεις μεταξύ των δεδομένων, π.χ. *Ποιες πόλεις βρίσκονται σε απόσταση μικρότερη από 100km από το αεροδρόμιο x*; ενώ στη δεύτερη περίπτωση, οι ερωτήσεις συνεπάγονται τη δημιουργία περισσότερων και πιο σύνθετων συσχετίσεων μεταξύ των δεδομένων, π.χ. *Επηρεάζεται η κατανομή του πληθυσμού από τη γεωμορφολογία του εδάφους;*

1.3 Διατύπωση Προβλημάτων

Το πρώτο από τα προβλήματα που προσεγγίζει η παρούσα διδακτορική διατριβή, έγκειται στη δυσχέρεια χειρισμού από πλευράς χρηστών, των εξειδικευμένων και απαιτητικών σε χρόνο εκμάθησης, γλωσσών διατύπωσης ερωτημάτων προς συστήματα που φιλοξενούν γεωχωρικά δεδομένα. Μάλιστα αρκετά συχνά απαιτείται οι χρήστες που αναζητούν απαντήσεις μέσα από τα συστήματα αυτά, να είναι έμπειροι προγραμματιστές. Έτσι, προτείνεται η χρήση ελεγχόμενης γλώσσας, συγκεκριμένα της Geo-Q, με στόχο τη διευκόλυνση του χρήστη στη διατύπωση ερωτημάτων για την απόκτηση γνώσης από συλλογές γεωχωρικών δεδομένων.

Το δεύτερο πρόβλημα αφορά στην επικοινωνία ανθρώπου – υπολογιστή και πιο συγκεκριμένα στην ανικανότητα του υπολογιστή να κατανοήσει πλήρως (με την έννοια της ικανότητας μετατροπής του λόγου σε εκτελέσιμες εντολές) τα ερωτήματα εκφρασμένα σε φυσική ανθρώπινη γλώσσα. Η επικοινωνία αυτή, δηλαδή η επικοινωνία βασισμένη στον τρόπο ομιλίας και σκέψης του ανθρώπου, ακούγεται ενδιαφέρουσα αλλά δύσκολα επιτεύξιμη. Μάλιστα, σύμφωνα με τη Rich [Rich83]: «*Οι υπολογιστές δεν θα είναι σε θέση να πράξουν όπως πράττουν καθημερινά οι άνθρωποι, εκτός κι αν αποκτήσουν την ικανότητα του λόγου*». Η αλήθεια είναι πως η πλήρης κατανόηση της φυσικής ανθρώπινης γλώσσας από τη μηχανή, εξαιτίας της πολυπλοκότητάς της (συντακτική, σημασιολογική και πραγματιστική), παραμένει αδύνατη έως τις μέρες μας. Προσεγγίζοντας το πρόβλημα αυτό, η ελεγχόμενη γλώσσα Geo-Q επιχειρεί να παράγει ερωτήματα άμεσα επεξεργάσιμα από τον υπολογιστή.

Το επόμενο πρόβλημα που τίθεται στο πλαίσιο της παρούσας διατριβής, είναι αυτό της επιλογής κατάλληλων επιγραφών για την περιγραφή των αποτελεσμάτων ομαδοποίησης γεωγραφικών σημασιολογικών πληροφοριών. Η επιλογή επιγραφών αποτελεί διαδικασία που συνήθως παραμελείται (ή και παραλείπεται) στο βωμό της ανάπτυξης όσο το δυνατό πιο αποτελεσματικών και γρήγορων αλγορίθμων ομαδοποίησης. Για την αντιμετώπιση του προβλήματος αυτού, προτείνεται μια μέθοδος παραγωγής σημασιολογικών επιγραφών, η οποία αναδεικνύει τα αποτελέσματα ομαδοποίησης, αποδίδοντας επιγραφές που συνοψίζουν το περιεχόμενο ομάδων ορισμών γεωγραφικών εννοιών.

Το παραπάνω πρόβλημα απαντάται σε αρκετές εφαρμογές. Καταρχάς αποτελεί κεντρικό ζήτημα στις εφαρμογές αυτόματης δημιουργίας οντολογιών, οι οποίες έχουν ως στόχο τη διαμόρφωση ιεραρχίας εννοιών συσχετιζόμενων μεταξύ τους με σχέσεις υπερώνυμου/υπώνυμου (IS-A), από ένα αρχικό σύνολο μη δομημένων ή ημι-δομημένων δεδομένων, όπως είναι για παράδειγμα τα κείμενα γραμμένα σε φυσική γλώσσα, τα γλωσσάρια, τα XML δεδομένα, οι βάσεων δεδομένων, κλπ. Στις εφαρμογές αυτές, είναι πολύ συχνά απαραίτητη η δημιουργία και απόδοση επιγραφών σε νέες έννοιες, οι οποίες θα αποτελέσουν υπερ-έννοιες (super-concepts) για κάποιες ομάδες σημασιολογικά όμοιων εννοιών. Παρόμοιο πρόβλημα μπορεί να προκύψει κατά τη διαδικασία ολοκλήρωσης οντολογιών, όσον αφορά την απόδοση ονομασίας και περιγραφής στις πρόσθετες έννοιες που ενδεχομένως χρειαστεί να δημιουργηθούν. Επιπροσθέτως, το πρόβλημα αυτό απαντάται σε εφαρμογές αναζήτησης σημασιολογικών πληροφοριών όταν επιχειρείται η κατηγοριοποίηση των αποτελεσμάτων σε ομάδες με σημασιολογικά συναφές περιεχόμενο.

Τέλος, στο τέταρτο κεφάλαιο της διατριβής, το περιβάλλον GeoScape σχεδιάστηκε για να αντιμετωπίσει το διττό πρόβλημα: 1) της απεικόνισης πολυδιάστατων δεδομένων σε χώρους περιορισμένων διαστάσεων, όπως είναι ο γεωγραφικός, κάνοντας χρήση ειδικών τεχνικών προβολής και μείωσης διαστάσεων, καθώς και κατάλληλων χωρικών μεταφορών και 2) της αναπαράστασης δεδομένων σε πολλά επίπεδα λεπτομέρειας βάσει μιας τεχνικής χωρικοποίησης που χρησιμοποιεί τη χωρική μεταφορά του τοπίου πληροφοριών.

1.4 Συνεισφορά και Δομή Διδακτορικής Διατριβής

Η συνεισφορά της διδακτορικής διατριβής αφορά στην ερευνητική κατεύθυνση της απόκτησης γεωγραφικής γνώσης και των συναφών σε αυτήν επιστημονικών πεδίων και, πιο συγκεκριμένα, έγκειται στην εύρεση λύσεων για την αντιμετώπιση των προβλημάτων που επισημάνθηκαν στην προηγούμενη παράγραφο.

Αναλυτικότερα, για τη διευκόλυνση του χρήστη στη διατύπωση ερωτημάτων προς τον υπολογιστή και την εξάλειψη της πολυπλοκότητας και της αμφισημίας που καθιστούν δύσκολα επεξεργάσιμη τη φυσική γλώσσα από τον υπολογιστή, εισάγεται και περιγράφεται η ελεγχόμενη γλώσσα Geo-Q, με τη βοήθεια της οποίας, μπορούν να διατυπωθούν ερωτήσεις

απόσπασης γνώσης προς βάσεις γεωχωρικών δεδομένων και μελλοντικά προς άλλες συλλογές δεδομένων. Το ζήτημα αυτό περιγράφεται στο *δεύτερο κεφάλαιο* της διατριβής.

Ακόμη, έχοντας ως στόχο την αυτοματοποίηση της διαδικασίας δημιουργίας επιγραφών στο πλαίσιο εφαρμογών που επεξεργάζονται και διαχειρίζονται σημασιολογικές πληροφορίες (π.χ. αυτόματη δημιουργία γεωχωρικής οντολογίας, ολοκλήρωση γεωχωρικών οντολογιών, κατηγοριοποίηση αποτελεσμάτων αναζήτησης πληροφοριών, κλπ), προτείνεται και περιγράφεται η μέθοδος παραγωγής σημασιολογικών επιγραφών Geo-Labeling. Η Geo-Labeling δέχεται ως δεδομένα εισόδου ομάδες ορισμών γεωγραφικών εννοιών διατυπωμένων σε φυσική γλώσσα, για τις οποίες προτείνει στη συνέχεια συνεκτικές και αντιπροσωπευτικές σημασιολογικές επιγραφές. Οι επιγραφές δομούνται με τρόπον που να μοιάζουν με σύντομους ορισμούς, οι οποίοι συνοψίζουν το περιεχόμενο ομάδων ορισμών γεωγραφικών εννοιών ενώ, ταυτόχρονα, διαφοροποιούν τις ομάδες αναμεταξύ τους. Η ανάπτυξη της Geo-Labeling πραγματοποιείται βάσει σαφώς καθορισμένων διαδοχικών βημάτων επεξεργασίας της πληροφορίας που εμπεριέχεται στους ορισμούς γεωγραφικών εννοιών. Έτσι, διευκολύνεται η μελλοντική τυποποίηση και αυτοματοποίησή της. Η Geo-Labeling και οι πιθανές εφαρμογές της εξετάζονται στο *τρίτο κεφάλαιο* της διατριβής.

Σε σχέση με το πεδίο της οπτικοποίησης πληροφοριών και ειδικότερα της χωρικοποίησης, υλοποιήθηκε το πρωτότυπο γραφικό περιβάλλον χωρικοποίησης πολυδιάστατων πληροφοριών GeoScape. Το *τέταρτο κεφάλαιο* της διατριβής αναπτύσσει το ζήτημα της χωρικοποίησης και περιγράφει λεπτομερώς το περιβάλλον GeoScape.

Τέλος, η ανακεφαλαίωση των κύριων σημείων της διατριβής και των σχετικών μελλοντικών ερευνητικών κατευθύνσεων παρουσιάζονται στο *πέμπτο κεφάλαιο*.

2. Απόκτηση Γεωγραφικής Γνώσης Βάσει Ελεγχόμενων Γλωσσών

2.1 Αναπαράσταση Γεωγραφικής Γνώσης

Για την ανάπτυξη του θέματος της απόκτησης γεωγραφικής γνώσης βάσει ελεγχόμενων γλωσσών, θεωρείται απαραίτητο να προηγηθεί μια συνοπτική περιγραφή των φορμαλισμών *αναπαράστασης γνώσης* (ΑΓ) που θα χρησιμοποιηθούν στις επόμενες παραγράφους.

Η ΑΓ αποτελεί πεδίο της Τεχνητής Νοημοσύνης το οποίο πρωτοεμφανίστηκε κατά τη δεκαετία 1980. Στη βιβλιογραφία υπάρχουν αρκετοί ορισμοί της ΑΓ. Σύμφωνα με την εγκυκλοπαίδεια του Πανεπιστημίου του Stanford[1], είναι το επιστημονικό πεδίο που *ασχολείται κυρίως με θέματα αναπαράστασης και συλλογιστικής (reasoning)*. Ένας άλλος ορισμός ο οποίος δίνεται από την εγκυκλοπαίδεια των επιστημών γνώσης του MIT [JR01] περιγράφει την ΑΓ ως την *κωδικοποίηση της γνώσης με τρόπο που να μπορεί να την επεξεργαστεί ένας Η/Υ και να εξάγει από αυτήν λογικά συμπεράσματα.*

Συνδυάζοντας τους δύο προηγούμενους ορισμούς και προσθέτοντας τη «χωρική διάσταση» που μας ενδιαφέρει, μπορούμε να ορίσουμε την *αναπαράσταση γεωγραφικής γνώσης* ως την *κωδικοποίηση της γνώσης που αφορά στο γεωγραφικό χώρο, η οποία υποστηρίζει τη διατύπωση συλλογισμών και την εξαγωγή συμπερασμάτων, επιτρέποντας παράλληλα την επεξεργασία της από Η/Υ.*

[1] http://plato.stanford.edu/entries/logic-ai, Τελευταία Προσπέλαση Ιούνιος 2010.

Οι Davis, Shrobe και Szolovitz [DSZ93], ερευνητές του MIT, επεσήμαναν πέντε διακριτές ιδιότητες της ΑΓ:

- Η ΑΓ αποτελεί **μοντέλο**, με την έννοια του αντιπροσωπευτικού υποκατάστατου μιας πτυχής της πραγματικότητας. Τα αντικείμενα, οι σχέσεις, τα φαινόμενα του πραγματικού κόσμου για τα οποία είναι φύσης αδύνατον να ενταχθούν στην ολότητά τους στο περιβάλλον ενός Η/Υ, υποκαθίστανται από συμβολικές αναπαραστάσεις που προσομοιώνουν τα χαρακτηριστικά και τη συμπεριφορά τους στο χώρο του Η/Υ.
- Η ΑΓ αποτελεί **σύνολο οντολογικών δεσμεύσεων**, δηλαδή εννοιών και σχέσεων ανά μεταξύ των εννοιών, που διέπονται από συγκεκριμένους οντολογικούς κανόνες και περιγράφουν μια πτυχή της πραγματικότητας.
- Η ΑΓ αποτελεί **τμήμα θεωρίας ευφυούς συλλογιστικής**. Η ΑΓ οφείλει να προσομοιώσει, όχι μόνο τις συμμετέχουσες έννοιες και σχέσεις σε μια πτυχή της πραγματικότητας, αλλά και τον τρόπο που αυτές συμπεριφέρονται και συναλλάσσονται. Ως εκ τούτου, η ΑΓ πρέπει να υποστηρίζει τη διατύπωση συλλογισμών και την εξαγωγή έγκυρων συμπερασμάτων.
- Η ΑΓ αποτελεί **μέσο αποτελεσματικής υλοποίησης ευφυούς συλλογιστικής** στο πλαίσιο υπολογιστικών συστημάτων, το οποίο συνεπάγεται ότι μπορεί να μετατραπεί σε μορφή επεξεργάσιμη από τα συστήματα αυτά.
- Η ΑΓ αποτελεί **μέσο έκφρασης και διευκόλυνσης της επικοινωνίας** μεταξύ των ανθρώπων που δημιουργούν την ΑΓ και των χρηστών που θα την εφαρμόσουν σε ένα συγκεκριμένο πεδίο εφαρμογής.

Τα παραπάνω αναδεικνύουν τρεις διαφορετικές αλλά αλληλένδετες όψεις της ΑΓ [KK08]: τη *σημασιολογία*, το *φορμαλισμού*, και τη *γλώσσα υλοποίησης*. Η πρώτη όψη, η σημασιολογία, καθορίζει τα αντικείμενα, τις σχέσεις και τα φαινόμενα που υπάρχουν και συνδιαλέγονται στην πτυχή της πραγματικότητας που ενδιαφέρει. Η τελευταία όψη, η γλώσσα υλοποίησης, αφορά στη δυνατότητα εφαρμογής και υλοποίησης της ΑΓ σ' ένα υπολογιστικό σύστημα. Η ενδιάμεση όψη, ο φορμαλισμός, συνιστά συνδετικό κρίκο μεταξύ

των άλλων δύο, καθορίζοντας τις παραδοχές, το συμβολισμό, την ορολογία και τους κανόνες συλλογιστικής της ΑΓ.

Ενδεικτικά, παραδείγματα φορμαλισμών ΑΓ, οι οποίοι έχουν χρησιμοποιηθεί στο γεωχωρικό τομέα, είναι οι ακόλουθοι: η θεωρία της ανάλυσης τυποποιημένων εννοιών (Formal Concept Analysis - FCA) [Will92], [GW99], οι εννοιολογικοί γράφοι (Conceptual Graphs - CG) [Sowa84], η θεωρία της ροής πληροφορίας (Information Flow - IF) [BS97], οι λογικές περιγραφής (Description Logics - DL) [Wood75], [Brac85], η θεωρία αναπαράστασης του λόγου (Discourse Representation Theory - DRT) [Kamp81], κλπ.

Η επιλογή μεταξύ των φορμαλισμών γίνεται ανάλογα με τη φύση των γνωσιακών στοιχείων που πρόκειται να αναπαρασταθούν και τους συλλογισμούς που χρειάζεται να διατυπωθούν. Για παράδειγμα, οι φορμαλισμοί FCA, CG και μερικώς ο IF, διαθέτουν εποπτικές ικανότητες που επιτρέπουν τη διαγραμματική απεικόνιση της πτυχής της πραγματικότητας που ενδιαφέρει και των εμπλεκόμενων σε αυτή εννοιών, διευκολύνοντας έτσι την άμεση αντίληψη και κατανόησή τους. Αντίθετα, οι φορμαλισμοί DL και DRT χαρακτηρίζονται από έναν μεγαλύτερο βαθμό τυποποίησης. Έτσι, οι έννοιες μοντελοποιούνται ως αντικείμενα με συγκεκριμένες ιδιότητες και οι σχέσεις αναμεταξύ τους περιγράφονται χρησιμοποιώντας λογικούς τελεστές. Αυτό ισχύει βέβαια περισσότερο για το φορμαλισμό DL, ο οποίος χρησιμοποιείται κυρίως ως τεχνική διατύπωσης λογικών προτάσεων και εξαγωγής συμπερασμάτων.

Σε ό, τι αφορά τους φορμαλισμούς FCA και CG, επιτρέπουν μια πιο ευέλικτη ΑΓ. Ειδικότερα, ο FCA παρέχει ένα μεθοδολογικό πλαίσιο για τη μοντελοποίηση, ανάλυση και ΑΓ βασισμένη στη μαθηματική θεωρία διατάξεων (Order Theory) και πιο συγκεκριμένα στη θεωρία των πλήρων δικτυωτών (Theory of Complete Lattices). Ο φορμαλισμός CG, ο οποίος θεωρείται ότι αποτελεί έναν πολύ ευέλικτο φορμαλισμό ΑΓ λόγω της δυνατότητας άμεσης μετατροπής του σε οποιαδήποτε μορφή λογικής [Sowa04b], εξετάζεται αναλυτικότερα στη συνέχεια.

2.2 Εννοιολογικοί Γράφοι (Conceptual Graphs - CG)

2.2.1. Εισαγωγή

Οι εννοιολογικοί γράφοι (Conceptual Graphs - CG) εισήχθησαν από τον John Sowa το 1984 [Sowa84] [Sowa00]. Το κύριο πλεονέκτημά τους έγκειται στο γεγονός ότι βρίσκονται πολύ κοντά στον τρόπο που ο άνθρωπος συλλογίζεται και διατυπώνει λογικές προτάσεις. Ενώ αρχικά οι CG δημιουργήθηκαν με σκοπό να διευκολύνουν το «πέρασμα» από μια μορφή λογικής σε άλλη, στη συνέχεια βρήκαν εφαρμογή και σε άλλα πεδία όπως αυτό της εξόρυξης γνώσης από κείμενα γραμμένα σε φυσική γλώσσα [Cyre97], [MGL02], [Hens04] και αυτό της αναζήτησης πληροφοριών [HOC96], [ZZLY02].

Πιο πρόσφατα, προτάθηκε η αξιοποίηση των CG στο γεωγραφικό τομέα. Για παράδειγμα, οι Karalopoulos *et al*. [KKK04] εισήγαγαν μια μεθοδολογία για την αναπαράσταση της γεωγραφικής γνώσης που εμπεριέχεται σε ορισμούς γεωγραφικών εννοιών υπό τη μορφή CG. Χρησιμοποιώντας έναν ειδικά σχεδιασμένο αλγόριθμο, πρότειναν οι ορισμοί να διατρέχονται και να αναγνωρίζονται παράλληλα τα συντακτικά τους μέρη. Τέλος, εξέφρασαν το αποτέλεσμα της συντακτικής ανάλυσης σε όρους CG.

Επιπροσθέτως, εκμεταλλευόμενοι τη δυνατότητα των CG να μετατρέπονται σε οποιαδήποτε μορφή λογικά διατυπωμένης πρότασης, οι Kavouras και Kontaxaki [KK05] εισήγαγαν την ελεγχόμενη γλώσσα Geo-Q, με σκοπό τη διατύπωση ερωτημάτων που να αναλύονται άμεσα σε όρους CG και στη συνέχεια να μετατρέπονται σε SQL εντολές. Η Geo-Q, η οποία περιγράφεται αναλυτικότερα στη συνέχεια, προτάθηκε για την απόσπαση γνώσης από βάσεις γεωχωρικών δεδομένων με τρόπο οικείο προς τον άνθρωπο και ταυτόχρονα αποτελεσματικό.

2.2.2. Ορισμός

Σύμφωνα με το διεθνές και σχετικό ISO πρότυπο [CGS02], ένας CG ορίζεται ως εξής: "*A conceptual graph g is a bipartite graph, which consists of two kinds of nodes called concepts and conceptual relations*", δηλαδή: «Ένας CG είναι ένας διμερής γράφος στον οποίο εναλλάσσονται δύο είδη κόμβων, οι έννοιες και οι εννοιολογικές σχέσεις.» Οι κόμβοι

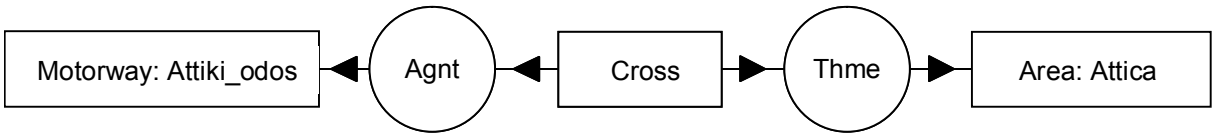
συνδέονται μεταξύ τους με συνδέσμους. Ο όρος *διμερής* αναφέρεται στο ότι οι σύνδεσμοι συνδέουν πάντα έναν κόμβο-έννοια μ' έναν κόμβο-εννοιολογική σχέση. Οι κόμβοι-έννοιες αντιστοιχούν σε οντότητες, χαρακτηριστικά ή ενέργειες. Οι κόμβοι-εννοιολογικές σχέσεις προσδιορίζουν το είδος της σχέσης μεταξύ των εννοιών.

Υπάρχουν διάφοροι συμβολισμοί για την αναπαράσταση CG. Οι συμβολισμοί αυτοί περιγράφονται ακολούθως.

2.2.3. Γραφική και Γραμμική Αναπαράσταση Εννοιολογικών Γράφων

Οι συμβολισμοί που χρησιμοποιούνται για την αναπαράσταση CG έχουν ως στόχο είτε να διευκολύνουν τη μεταφορά CG διαμέσου συστημάτων και δικτύων υπολογιστών (π.χ. Conceptual Graph Interchange Form - CGIF), είτε να μεταδώσουν πληροφορίες σε χρήστες, οπότε και υιοθετούνται πιο φιλικοί προς τον άνθρωπο συμβολισμοί. Στη δεύτερη περίπτωση, υπάρχει επίσης δυνατότητα γραφικής (Display Form - DF) ή γραμμικής (Linear Form - LF) αναπαράστασης.

Στη *γραφική αναπαράσταση* CG, οι έννοιες οπτικοποιούνται ως ορθογώνια, οι εννοιολογικές σχέσεις ως κύκλοι ή ελλείψεις και οι σύνδεσμοι ως προσανατολισμένα τόξα. Για παράδειγμα, ο CG που περιγράφει την πρόταση «*Ο αυτοκινητόδρομος 'Αττική Οδός' διασχίζει την περιοχή της Αττικής*» απεικονίζεται στο σχήμα 2.1.



Σχήμα 2.1 Παράδειγμα γραφικής αναπαράστασης CG

Σε σχέση με τη γραφική αναπαράσταση, η *γραμμική αναπαράσταση* CG είναι λιγότερο εποπτική, ειδικά για πολύπλοκους γράφους. Καταλαμβάνει όμως λιγότερο χώρο και διευκολύνει τη δημιουργία CG με το πληκτρολόγιο. Στην περίπτωση αυτή, οι έννοιες εισάγονται ως λέξεις σε αγκύλες, ενώ οι εννοιολογικές σχέσεις ως λέξεις σε παρενθέσεις. Για παράδειγμα, η γραμμική αναπαράσταση του CG που απεικονίζεται στο σχήμα 2.1 έχει ως εξής:

[Motorway: Attiki_odos] ← (Agnt) ← [Cross] → (Thme) → [Area: Attica]

2.2.4. Δομικά Στοιχεία Εννοιολογικών Γράφων

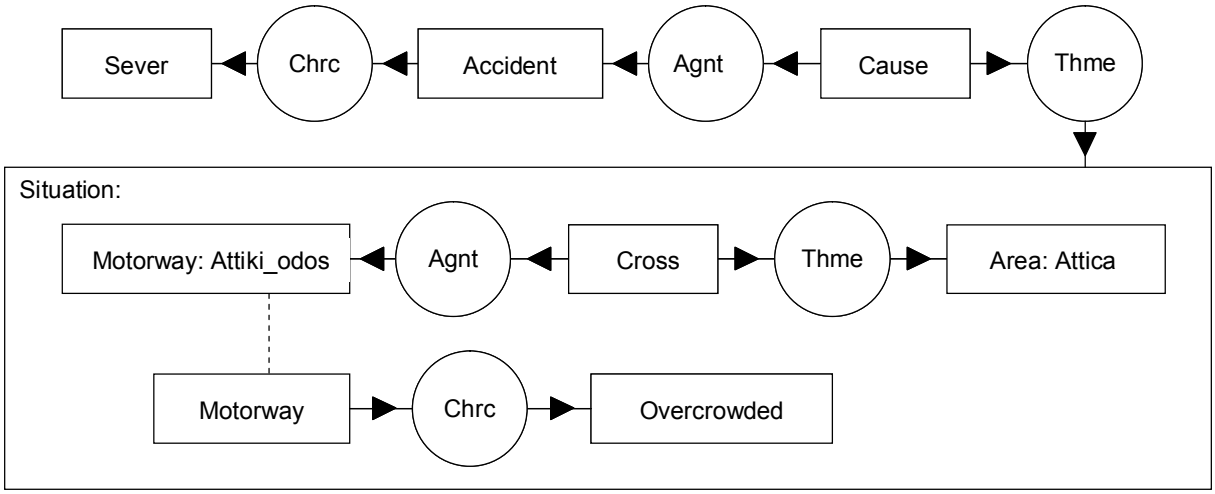
Μια έννοια δομείται από δύο μέρη: έναν *εννοιολογικό τύπο* (concept type) και μία *αναφορά* (referent). Μπορούμε να παρομοιάσουμε τον εννοιολογικό τύπο με κατηγορία εννοιών με κοινά χαρακτηριστικά και ιδιότητες. Οι τύποι εννοιών οργανώνονται σε ιεραρχία αποτελούμενη από υπο-τύπους και υπερ-τύπους και με ρίζα που ονομάζεται συμβατικά *Entity*. Ακόμη, θεωρείται ότι οι τύποι που απαρτίζουν το προ-τελευταίο ιεραρχικό επίπεδο αποτελούν όλοι υπερ-τύποι του τύπου *Absurdity* που συνθέτει συμβατικά κι από μόνος του το τελευταίο ιεραρχικό επίπεδο. Οι *περιπτώσεις* (instances) εννοιολογικών τύπων αποκαλούνται *αναφορές* (referent). Για παράδειγμα, η αναφορά της έννοιας *[Motorway: Attiki_odos]* είναι *Attiki_odos*, ενώ ο τύπος της είναι *Motorway*.

Οι αναφορές χαρακτηρίζονται από έναν *τελεστή ποσοτικοποίησης* (quantifier) και έναν *δείκτη* (designator). Ο τελεστής ποσοτικοποίησης μπορεί να είναι *υπαρξιακός* (existential) ή *ορισμένος* (defined). Ο υπαρξιακός τελεστής ποσοτικοποίησης, ο οποίος αναπαρίσταται με το σύμβολο $\exists$ ή εννοείται όταν δεν χρησιμοποιείται κανένα ειδικό σύμβολο, υποδεικνύει την ύπαρξη μιας τουλάχιστον αναφοράς. Ο ορισμένος τελεστής ποσοτικοποίησης, ο οποίος αναπαρίσταται με τη βοήθεια των συμβόλων $\forall$, *@,{*}* ή με κάποιο σύνολο αναγνωριστικών, π.χ. *{a,b,c}*, αναφέρεται στον αριθμό ή την ποσότητα των αναφορών. Για παράδειγμα, η έννοια *[Motorways: {*}]* υποδεικνύει την ύπαρξη πολλών αυτοκινητοδρόμων χωρίς όμως να προσδιορίζει τον ακριβή αριθμό τους, ενώ η έννοια *[Motorways: {*}@3]* υποδεικνύει την ύπαρξη ακριβώς τριών αυτοκινητοδρόμων. Ακόμη, η έννοια *[Motorways: $\forall$]* που χρησιμοποιεί το γνωστό μαθηματικό σύμβολο $\forall$ (για κάθε), υποδεικνύει το σύνολο των αυτοκινητοδρόμων ενώ η έννοια *[Motorways: {Attiki_odos, Ionia_odos}]* τους συγκεκριμένους αυτοκινητοδρόμους *Attiki_odos* και *Ionia_odos*. Ο δείκτης περιγράφει τις τιμές των εννοιών. Για παράδειγμα, *[Distance: 1200]*, *[Name: 'Αττική Οδός']*, *[Measure: <1200, km>]*, κλπ.

Κάθε εννοιολογική σχέση περιγράφεται από τον *τύπο* (relation type), το *σθένος* (valence) και την *υπογραφή* (signature) της. Το σθένος αναφέρεται στον αριθμό των συνδέσεων που ξεκινούν ή καταλήγουν στη σχέση. Η υπογραφή προσδιορίζει τους τύπους των εννοιών με τις

οποίες μπορεί να συνδεθεί η σχέση ενώ, όπως και στην περίπτωση των τύπων των εννοιών, οι τύποι των σχέσεων μπορούν να οργανωθούν σε ιεραρχία υπερ-τύπων και υπο-τύπων. Στο σχήμα 2.1, το σθένος της σχέσης *(Agnt)* ισούται με 2, διότι υπάρχουν συνολικά και πάντα δύο συνδέσεις που ξεκινούν ή καταλήγουν στη σχέση αυτή.

Επιπροσθέτως, σ' έναν CG, μπορούν να οριστούν *πλαίσια* (contexts). Κάθε πλαίσιο συμβολίζεται με ορθογώνιο σχήμα και αντιστοιχεί σε έννοια της οποίας η αναφορά περιγράφεται από έναν μη κενό δείκτη. Για παράδειγμα στο σχήμα 2.2, η αναφορά του πλαισίου με όνομα Situation προσδιορίζεται από το δείκτη ο οποίος αντιστοιχεί στον CG που εσωκλείεται στο ορθογώνιο. Ο συνολικός CG περιγράφει την πρόταση: «*Ένα σοβαρό ατύχημα δημιούργησε μποτιλιάρισμα στον αυτοκινητόδρομο Αττική Οδό που διασχίζει την Αττική.*»



Σχήμα 2.2 Παράδειγμα χρήσης πλαισίου σε CG

Στο παράδειγμα του σχήματος 2.2, έχει οριστεί με διακεκομμένη γραμμή μία *συναναφορά* (coreference). Μία συναναφορά συνδέει τις έννοιες εκείνες που αναφέρονται στην ίδια περίπτωση, π.χ. οι έννοιες *[Motorway: Attiki_odos]* και *[Motorway]* αναφέρονται στην ίδια περίπτωση, εν προκειμένου στην Αττική Οδό.

2.3 Ελεγχόμενες Γλώσσες (Controlled Languages)

Όπως επισημάνθηκε στην εισαγωγή, ένα από τα προβλήματα που αντιμετωπίζουν συχνά οι χρήστες συστημάτων που φιλοξενούν γεωχωρικά δεδομένα, είναι η δυσκολία στην εκμάθηση και το χειρισμό εξειδικευμένων και πολυσύνθετων γλωσσών επικοινωνίας. Μάλιστα αρκετά συχνά απαιτείται οι χρήστες να είναι έμπειροι προγραμματιστές. Όσο προφανές και αν ακούγεται, ανάλογο πρόβλημα παρουσιάζεται από την πλευρά του συστήματος, διότι ένας υπολογιστής δεν μπορεί να «κατανοήσει» στην πληρότητά του τον ανθρώπινο λόγο και να ανταποκριθεί σε οποιαδήποτε εντολή του.

Για να προσπεραστούν οι δυσκολίες αυτές, έστω και μερικώς, έχουν προταθεί πολλές προσεγγίσεις όπως (α) η ημι-αυτόματη επεξεργασία του λόγου με τη βοήθεια και παρέμβαση ενός ειδικού και (β) η δημιουργία και χρήση ελεγχόμενων γλωσσών, η οποία περιγράφεται στη συνέχεια.

Οι *ελεγχόμενες γλώσσες* αποτελούν υποσύνολα των φυσικών γλωσσών από την άποψη ότι έχουν ακριβώς την ίδια μορφή μόνο που επιβάλλεται επιπλέον ένας περιορισμός στο λεξιλόγιο και στη σύνταξη των προτάσεων. Ο περιορισμός έχει ως στόχο την εξάλειψη της πολυπλοκότητας και της αμφισημίας που χαρακτηρίζουν τις φυσικές γλώσσες. Ακόμη, λόγω της συνοχής, της ομοιομορφίας και της απλούστευσης της δομής των παραγόμενων προτάσεων, καθώς και λόγω της τυποποίησης στη διάταξη και τη μορφοποίηση των συντασσόμενων με ελεγχόμενη γλώσσα εγγράφων, επιτυγχάνονται η βελτίωση της ευχρηστίας, η δυνατότητα άμεσης μεταφοράς μεταξύ υπολογιστικών συστημάτων, η εύκολη ανάκτηση και η άμεση μετάφραση.

Οι ελεγχόμενες γλώσσες χρησιμοποιούνται σε τομείς όπου παραδοσιακά τα παραγόμενα έγγραφα είναι δυσνόητα και πολύπλοκα, όπως για παράδειγμα στον τομέα της οικονομίας και της νομικής επιστήμης. Αναφέρεται ακόμη ότι οι ελεγχόμενες γλώσσες αναπτύχθηκαν εκτός από τα αγγλικά και σε άλλες γλώσσες: γερμανικά, σουηδικά, γαλλικά, ισπανικά, κινέζικα και … ελληνικά [VMMK03].

Οι ελεγχόμενες γλώσσες εμπίπτουν σε δύο κύριες κατηγορίες: 1) εκείνων που απευθύνονται σε ανθρώπους με σκοπό τη διευκόλυνση της αναγνωσιμότητας των παραγόμενων κειμένων/εγγράφων και 2) εκείνων που στοχεύουν στη βελτιστοποίηση των

αλγορίθμων επεξεργασίας κειμένου (από πλευράς χρόνου και πολυπλοκότητας) που εκτελούνται από υπολογιστικά συστήματα.

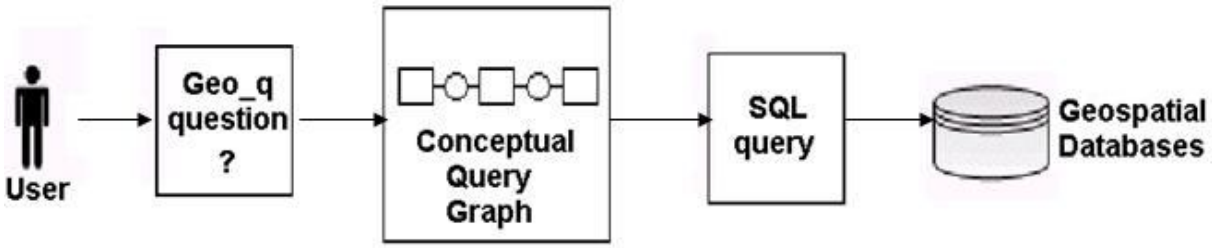
Ακόμη, ένα πολύ μεγάλο πλεονέκτημα των ελεγχόμενων γλωσσών έγκειται στο ότι μπορούν άμεσα να μετατραπούν σε οποιασδήποτε μορφής δομημένης λογικής έκφρασης όπως είναι η Πρωτοβάθμια Λογική (First-Order-Logic), οι γλώσσες προγραμματισμού, κλπ. Προς την κατεύθυνση αυτή, η ελεγχόμενη γλώσσα Attempto, η οποία δημιουργήθηκε στο Πανεπιστήμιο της Ζυρίχης και αποτελεί υποσύνολο της αγγλικής, χρησιμοποιείται για τον ορισμό των απαιτήσεων σχεδιασμού ενός συστήματος και μπορεί να μεταφραστεί σε εκτελέσιμα προγράμματα Prolog [SFS98]. Ένα άλλο παράδειγμα, η ελεγχόμενη γλώσσα Common Logic Controlled English (CLCE), δημιουργήθηκε ώστε «*οποιοσδήποτε που να μπορεί να διαβάζει αγγλικά να μπορεί να διαβάζει προτάσεις γραμμένες σε CLCE με λίγη ή και καθόλου εξάσκηση*» [Sowa04a] αλλά και για να μπορεί να μεταφραστεί άμεσα σε προτάσεις Πρωτοβάθμιας Λογικής.

2.4 Απόσπαση Γεωγραφικής Γνώσης με την Ελεγχόμενη Γλώσσα Geo-Q

Θέλοντας να αξιοποιήσουμε τις δυνατότητες των ελεγχόμενων γλωσσών στο γεωγραφικό τομέα, θέσαμε τα θεμέλια για τη δημιουργία μιας νέας ελεγχόμενης γλώσσας, της *Geo-Q*, με τη βοήθεια της οποίας, μπορούν να διατυπωθούν ερωτήσεις προς βάσεις γεωχωρικών δεδομένων, με σκοπό την απόσπαση γνώσης που σχετίζεται με τα ποιοτικά χαρακτηριστικά και τις χωρικές ιδιότητες (γεωμετρικά χαρακτηριστικά, τοπολογικές σχέσεις, κλπ) των δεδομένων.

Σε ένα πρώτο βήμα, κάθε Geo-Q ερώτηση επεξεργάζεται με τρόπο ώστε να μετατραπεί σε εννοιολογικό γράφο – ερώτηση. Σε δεύτερο βήμα, ο εννοιολογικός γράφος – ερώτηση μεταφράζεται σε κώδικα SQL, του οποίου η εκτέλεση οδηγεί στην απάντηση του αρχικού ερωτήματος.

Το σχήμα 2.3 περιγράφει τη σειριακή επεξεργασία και μετατροπή των Geo-Q ερωτήσεων που υποβάλλονται σε βάσεις γεωχωρικών δεδομένων.



Σχήμα 2.3 Σειριακή επεξεργασία και μετατροπή των Geo-Q ερωτήσεων που υποβάλλονται σε βάσεις γεωχωρικών δεδομένων [KK05b]

2.4.1. Ερωτήσεις που Υποστηρίζονται από την Ελεγχόμενη Γλώσσα Geo-Q

Στην τρέχουσα έκδοση της Geo-Q, δεν μπορούν να τεθούν μεγάλη ποικιλία ερωτήσεων προς βάσεις γεωχωρικών δεδομένων. Συγκεκριμένα, εστιάσαμε στη σύνθεση των πιο αντιπροσωπευτικών. Έτσι, υποθέσαμε αρχικά ότι οι Geo-Q ερωτήσεις διακρίνονται στα είδη Q1, Q2 και Q3. Παρακάτω παρουσιάζονται ενδεικτικά παραδείγματα:

Είδος Q1: Ερωτήσεις που παράγουν απαντήσεις τύπου *yes* ή *no*

- *Is polygon A adjacent to polygon B?*
- *Does polygon A touch polygon B?*
- *Is point X at the North of polygon A?*

Είδος Q2: Ερωτήσεις που επιστρέφουν ως απάντηση μια ή περισσότερες γεωγραφικές οντότητες

- *Which line intersects polygon A?*
- *Which polygon is closest to polygon B?*
- *Which points are at the North of polygon A?*

Είδος Q3: Ερωτήσεις που επιστρέφουν την τιμή κάποιου χαρακτηριστικού μιας γεωγραφικής οντότητας

- *What is the color of polygon A?*

- *What are the coordinates of X?*
- *What is the shape of B?*

2.4.2. Γραμματική της Ελεγχόμενης Γλώσσας Geo-Q

Για την περιγραφή της γραμματικής της Geo-Q, ορίσαμε τη λεκτική και τη συντακτική της μορφή.

a) Λεκτική Μορφή Geo-Q

Το λεξιλόγιο της γλώσσας απαρτίζεται από:

1. Λέξεις που υποστηρίζουν τη δόμηση των ερωτήσεων, όπως: *which*, *what*, *of*, *and*, *the*, *a*, *an*, *some*, *many*, *any*, κλπ.
2. Λέξεις που συμμετέχουν σε χωρικές εκφράσεις, όπως: *at the North*, *at the South*, *near*, *between*, κλπ.
3. Λέξεις που αντιστοιχούν σε ρήματα που εκφράζουν χωρικές σχέσεις και που μπορούν να χρησιμοποιηθούν στον ενικό/πληθυντικό και στην παθητική/ενεργητική φωνή, όπως: *intersects*, *intersect*, *intersected by*, κλπ.
4. Λέξεις που προέρχονται από τις ονομασίες πινάκων και πεδίων της βάσης γεωχωρικών δεδομένων.

Ακόμα, θεωρείται ότι χρησιμοποιείται το σύμβολο '*?*' στο οποίο λήγουν οι υποβληθείσες ερωτήσεις.

b) Συντακτική Μορφή Geo-Q

Το συντακτικό της Geo-Q ορίστηκε χρησιμοποιώντας κανόνες Backus-Naur[2]. Στην τρέχουσα έκδοση, υποστηρίζονται μόνο ερωτήσεις τύπου Q1, Q2 ή Q3. Έτσι, το κυρίως σώμα των ερωτήσεων περιγράφεται από τους ακόλουθους επεκτάσιμους συντακτικούς κανόνες:

<question> ::= <question_type> ?

<question_type> ::= <Q1> | <Q2> | <Q3>

Η διατύπωση ερωτήσεων τύπου Q1, Q2 και Q3 προϋποθέτει τη δυνατότητα χρησιμοποίησης εκφράσεων που περιγράφουν χωρικές σχέσεις όπως τοπολογικές, κατεύθυνσης, προσανατολισμού, εγγύτητας, κλπ. Οι χωρικές σχέσεις που συντίθενται με Geo-Q αποτελούν είτε *ρηματικές φράσεις*, οι οποίες χρησιμοποιούν το ρήμα *be* ή άλλα ρήματα, είτε *μη ρηματικές φράσεις*, οι οποίες δεν περιέχουν ρήμα στη δομή τους:

<relation_expression> ::= <verb_expr> | <geo_expr>[1]

<verb_expr> ::= <active_expr> | <passive_expr>

<active_expr> ::= <plur_verb>[2] | <sing_verb>[3]

<passive_expr> ::= <verb_be_expr> | <past_participle>[4]

<verb_be_expr> ::= <verb_be> <qualifiers> <prop_expr>

<verb_be> ::= *is* | *are*

όπου:

---

[2] International Standard ISO/IEC 14977, http://www.iso.org, τελευταία προσπέλαση Ιούνιος 2010.

1. geo_expr: μη τερματικό σύμβολο που συνθέτει εκφράσεις, όπως: *at the North*, *at the South*, *near*, *between*, κλπ.
2. plur_verb: μη τερματικό σύμβολο που συνθέτει ρηματικές φράσεις στον πληθυντικό και στην ενεργητική φωνή, όπως: *touch*, *intersect*, *overlap*, κλπ.
3. sing_verb: μη τερματικό σύμβολο που συνθέτει ρηματικές φράσεις στον ενικό και στην ενεργητική φωνή, όπως: *touches*, *intersects*, *overlaps*, κλπ.
4. past_participle: μη τερματικό σύμβολο που συνθέτει εκφράσεις οι οποίες χρησιμοποιούν ρήματα στην παθητική φωνή όπως: *touched by*, *intersected by*, *overlapped by*, etc.
5. *is* | *are*: τερματικά σύμβολα που εκφράζουν το ρήμα *be* στον ενικό ή το πληθυντικό.

Οι αναφορές σε γεωχωρικές οντότητες γίνονται με εκφράσεις που συνδυάζουν ονόματα, άρθρα και επίθετα. Τα ονόματα και τα επίθετα αντιστοιχούν στα πεδία των πινάκων της βάσης γεωχωρικών δεδομένων. Ακόμα, είναι δυνατόν να χρησιμοποιείται ο ενικός ή ο πληθυντικός όπως ακριβώς συμβαίνει και στη φυσική γλώσσα. Ενδεικτικά, παρατίθενται ακολούθως κάποιοι γραμματικοί κανόνες που περιγράφουν εκφράσεις αυτού του είδους:

<many_geoentities> ::= <geoentity> [[{, <geoentity>}] *and* <geoentity>]

<geoentity> : := [<articles>] [<qualifiers>] [*entity_type*[1]] *entity_name*[2] [<qualifiers>]

<articles> ::= (*the* | *a* | *an* | *some* | *many* | *any*)

<qualifiers> ::= <qualifier> [[{, <qualifier>}] *and* <qualifier>]

<qualifier> ::= *entity_attribute*[3]

Όπου:

1. *entity_type*: τερματικό σύμβολο που εκφράζει το όνομα γεωχωρικής οντότητας που αντιστοιχεί σε συγκεκριμένο πίνακα της βάσης γεωχωρικών δεδομένων. Για παράδειγμα: *polygon*, *line*, κλπ.

2. *entity_name*: τερματικό σύμβολο που εκφράζει το όνομα του χαρακτηριστικού γεωχωρικής οντότητας που αντιστοιχεί στο πρωτεύον κλειδί πίνακα της βάσης γεωχωρικών δεδομένων. Για παράδειγμα: *A*, *B*, *X*, etc.
3. *entity_attribute*: τερματικό σύμβολο που εκφράζει το όνομα κάποιου χαρακτηριστικού γεωχωρικής οντότητας που αντιστοιχεί σε συγκεκριμένο πεδίο ενός πίνακα της βάσης γεωχωρικών δεδομένων. Για παράδειγμα: *coordinates*, *color*, *shape*, etc.

2.4.3. Βήμα Πρώτο: Μετατροπή Geo-Q Ερωτήσεων σε Εννοιολογικούς Γράφους

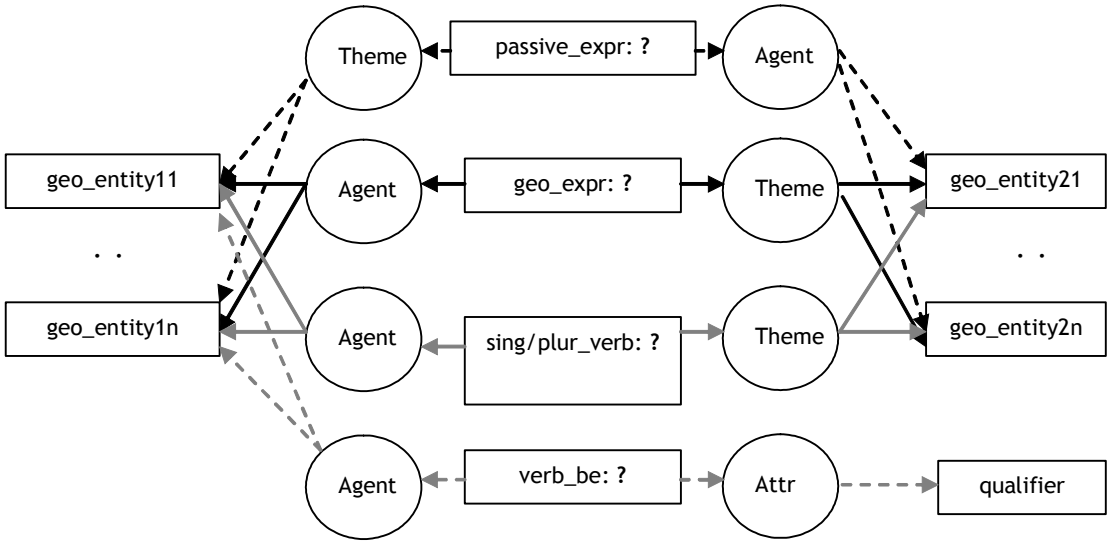
Για κάθε Geo-Q ερώτηση, δημιουργείται ένας εννοιολογικός γράφος του οποίου οι έννοιες και οι σχέσεις αντιστοιχούν στα δομικά στοιχεία της ερώτησης. Για να μπορέσει ο χρήστης να χρησιμοποιήσει όρους που αντιστοιχούν σε πίνακες ή πεδία της βάσης γεωχωρικών δεδομένων, προτείνουμε το σύστημα να υποστηρίζει την αυτόματη δημιουργία καταλόγου που να προτείνει ονόματα πινάκων και πεδίων, καθώς και συνώνυμα αυτών, και να επιτρέπει στο χρήστη να ορίσει δικές του λέξεις με κατάλληλες αντιστοιχίσεις (aliases). Όσον αφορά τα ρήματα και τις εκφράσεις που χαρακτηρίζουν τις γεωχωρικές σχέσεις μεταξύ των δεδομένων (π.χ. *touches*, *between*, κλπ), προτείνουμε το σύστημα να δημιουργεί και να ενσωματώνει τους σχετικούς πίνακες στη βάση γεωχωρικών δεδομένων, κάθε φορά που αυτό απαιτηθεί από συγκεκριμένες ερωτήσεις.

Παρακάτω περιγράφουμε πως οι ερωτήσεις όλων των τύπων μοντελοποιούνται υπό τη μορφή εννοιολογικών γράφων.

a) Ερωτήσεις Τύπου Q1

Οι ερωτήσεις τύπου Q1 εξετάζουν εάν μια σχέση ή μια ιδιότητα χαρακτηρίζει κάποια ή κάποιες γεωχωρικές οντότητες. Οι ερωτήσεις αυτές λαμβάνουν την απάντηση *yes* ή *no*. Η επεξεργασία των ερωτήσεων αυτών εξαρτάται από το αν το ρήμα που χρησιμοποιείται βρίσκεται σε ενεργητική ή παθητική φωνή ή από το αν πρόκειται για έκφραση που συνδυάζει το ρήμα *be* με γεωχωρική σχέση. Το ερωτηματικό εντοπίζεται στη «μεσαία» έννοια του

εννοιολογικού γράφου. Στο σχήμα 2.4 απεικονίζονται τα εναλλακτικά σενάρια που παράγουν ερωτήσεις τύπου Q1.



Σχήμα 2.4 Σενάρια παραγωγής Geo-Q ερωτήσεων τύπου Q1 [KK05b]

Για παράδειγμα, η Geo-Q ερώτηση *Is polygon A adjacent to polygon B?* η οποία, από την εφαρμογή των διαδοχικών συντακτικών κανόνων:

<question> ::= *Is polygon A adjacent to polygon B?*

<question> ::= <verb_be> [*entity_type*] *entity_name* <geo_rel> <prop_expr> [*entity_type*] *entity_name?*

<question> ::= <verb_be> <geoentity> <geo_expr> <geoentity>?

<question> ::= <Q1> ?

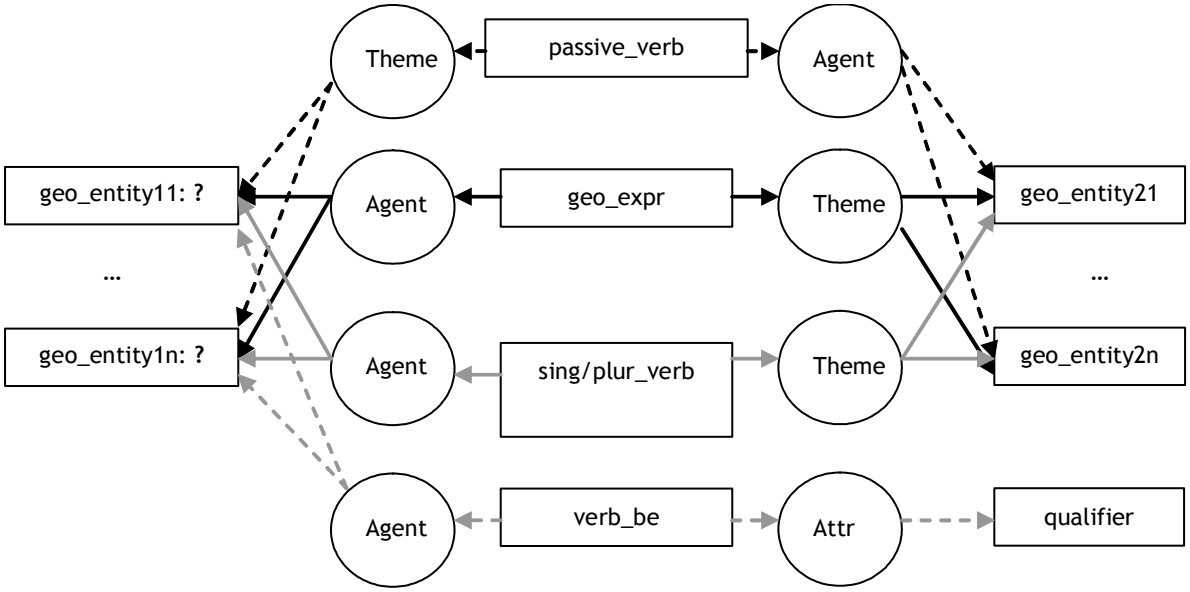
αναγνωρίζεται ως ερώτηση τύπου Q1, μετατρέπεται στον εννοιολογικό γράφο – ερώτηση:

[Polygon: A] ← (Agnt) ← [Is_adjacent_to?] → (Thme) → [Polygon: B]

b) Ερωτήσεις Τύπου Q2

Οι ερωτήσεις τύπου Q2 αναζητούν το σύνολο των οντοτήτων που συνδέονται μέσω γεωχωρικής σχέσης μ' ένα δεύτερο σύνολο οντοτήτων. Όπως και στην περίπτωση των

ερωτήσεων τύπου Q1, η επεξεργασία των ερωτήσεων τύπου Q2 εξαρτάται από το αν το ρήμα που χρησιμοποιείται βρίσκεται σε ενεργητική ή παθητική φωνή ή από το αν πρόκειται για έκφραση που συνδυάζει το ρήμα *be* με γεωχωρική σχέση. Το ερωτηματικό εντοπίζεται στις έννοιες που βρίσκονται αριστερά στον εννοιολογικό γράφο. Στο σχήμα 2.5 απεικονίζονται τα εναλλακτικά σενάρια που παράγουν ερωτήσεις τύπου Q2.



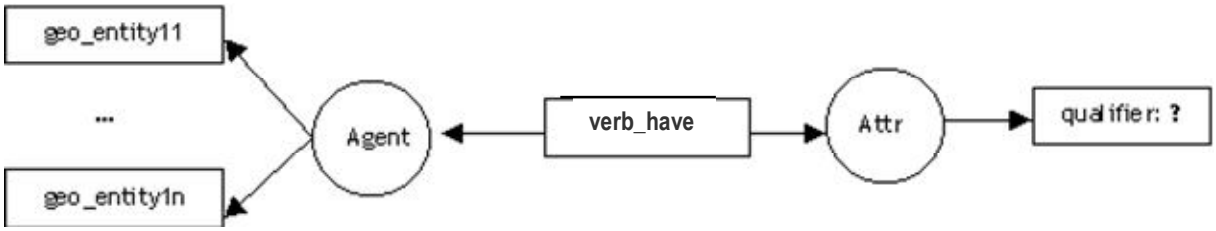
Σχήμα 2.5 Σενάρια παραγωγής Geo-Q ερωτήσεων τύπου Q2 [KK05b]

Για παράδειγμα, η Geo-Q ερώτηση *Which polygon touches A?*, με την εφαρμογή των κατάλληλων συντακτικών κανόνων που προσδιορίζουν τις ερωτήσεις τύπου Q2, οδηγεί στον εννοιολογικό γράφο – ερώτηση:

[*Polygon: {*}?*] <- (*Agnt*) <- [*Touches*] -> (*Thme*) -> [*A*]

c) Ερωτήσεις Τύπου Q3

Όσον αφορά τις ερωτήσεις τύπου Q3, αναζητείται η ιδιότητα ή το χαρακτηριστικό κάποιας ή κάποιων οντοτήτων. Το ερωτηματικό εντοπίζεται στις έννοιες που βρίσκονται δεξιά στον εννοιολογικό γράφο. Στο σχήμα 2.6 απεικονίζεται ο εννοιολογικός γράφος που παράγει ερωτήσεις τύπου Q3.



Σχήμα 2.6 Παραγωγή Geo-Q ερωτήσεων τύπου Q3 [KK05b]

Για παράδειγμα, η Geo-Q ερώτηση *What are the coordinates of X?*, με την εφαρμογή των κατάλληλων συντακτικών κανόνων που προσδιορίζουν τις ερωτήσεις τύπου Q3, οδηγεί στον εννοιολογικό γράφο – ερώτηση:

[*X*] <- (*Agnt*) <- [*Have*] -> (*Attr*) -> [*Coordinates:?*]

2.4.4. Βήμα Δεύτερο: Μετατροπή των Εννοιολογικών Γράφων των Geo-Q Ερωτήσεων σε Εντολές SQL

Το γεγονός ότι το σύστημα «κατανοεί» μια ερώτηση σημαίνει ότι είναι ικανό να την επεξεργαστεί και να τη μετατρέψει σε άλλη μορφή λογικής. Έτσι, οι εννοιολογικοί γράφοι μετατρέπονται σ' ένα δεύτερο βήμα σε κώδικα SQL.

Στη συνέχεια, θα χρησιμοποιηθεί το παρακάτω παράδειγμα ως βάση για την περιγραφή της όλης διαδικασίας. Υποθέτουμε ότι η βάση γεωχωρικών δεδομένων ορίζει κόμβους, τόξα και πολύγωνα μέσω των πινάκων 2.1, 2.2 και 2.3 και ότι διατυπώνεται η Geo-Q ερώτηση: *Which polygon touches A?*

Node	X1	X2
1	25	78
2	30	82
3	28	76
4	67	67
5	35	83

Πίνακας 2.1 Κόμβοι [KK05b]

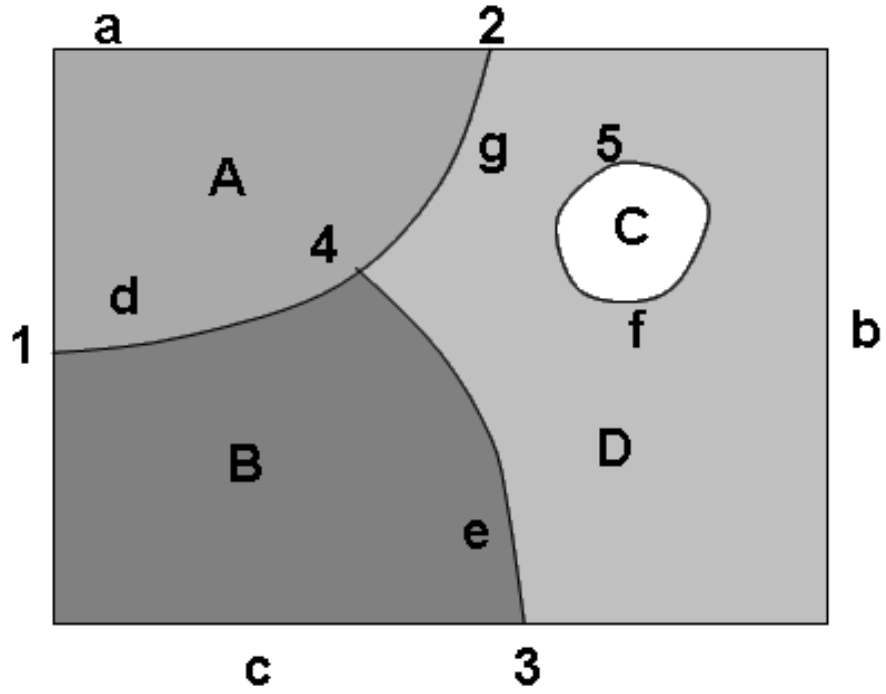
Arc	Srce_node	End_node
a	1	2
b	3	2
c	3	1
d	1	4
e	3	4
f	5	5
g	4	2

Πίνακας 2.2 Τόξα [KK05b]

Poly	Arc_num	Arc_list
A	3	a, d, g
B	3	c, d, e
C	1	f
D	4	b, e, g, f

Πίνακας 2.3 Πολύγωνα [KK05b]

Η γραφική αναπαράσταση των δεδομένων των πινάκων φαίνεται στο σχήμα 2.7.



Σχήμα 2.7 Γραφική αναπαράσταση δεδομένων παραδείγματος [KK05b]

Η μετατροπή των Geo-Q ερωτήσεων σε εντολές SQL προϋποθέτει την αναδιοργάνωση της βάσης γεωχωρικών δεδομένων με το φορμαλισμό που χρησιμοποιείται για την περιγραφή εννοιολογικών γράφων. Ο Sowa υποστηρίζει σχετικά ότι: «Η αναδιοργάνωση μιας μεγάλης βάσης δεδομένων είναι μια χρονοβόρα διαδικασία που είναι όμως καμιά φορά απαραίτητη. Αλλά για να απαντηθεί μια απλή ερώτηση, είναι συνήθως γρηγορότερο να αναδιοργανώσεις τον εννοιολογικό γράφο – ερώτηση παρά όλη τη βάση δεδομένων» [Sowa04a] . Στο παράδειγμά μας, η βάση γεωχωρικών δεδομένων είναι μικρή και θα αναδιοργανώσουμε μόνο τους πίνακες που συμμετέχουν στην ερώτηση.

Έτσι, οι πίνακες 2.1, 2.2 και 2.3 εκφράζονται, χρησιμοποιώντας το φορμαλισμό των εννοιολογικών γράφων, από τις σχέσεις 1, 2 και 3.

relation *Nodes* (**x, *y, *z*) is (1)

[*Node: ?x*] -

(*Coord*)-> [*X1: ?y*]

(*Coord*)-> [*X2: ?z*]

relation *Arcs* (**x, *y, *z*) is (2)

[*Arc: ?x*] -

(*Srce_node*)-> [*Node: ?y*]

(*End_node*)-> [*Node: ?z*]

relation *Polys* (**x, *y, *z_1, ..., *z_n*) is (3)

[*Poly: ?x*] -

(*Arc_num*)-> [*Int_num: ?y*]

(*Arc_list*) -> [*Arc: ?z_1*]

…

(*Arc_list*) -> [*Arc: ?z_n*]

} Arc_num

Επιπλέον, από το περιεχόμενο του πίνακα 2.3, συντίθεται ο πίνακας 2.4 που περιγράφει τα πολύγωνα που βρίσκονται σε επαφή.

Poly1	Poly2
A	B
A	D
B	A
B	D
C	D
D	A
D	B
D	C

Πίνακας 2.4 Πολύγωνα που βρίσκονται σε επαφή [KK05b]

Ο πίνακας 2.4, με το φορμαλισμό που χρησιμοποιείται για την περιγραφή εννοιολογικών γράφων, μετατρέπεται στην εννοιολογική σχέση *Touches*:

relation *Touches* (**x*,**y*) is [*Poly1*: *?y*] <- (*Agnt*) <- [*Touches*] -> (*Thme*) ->[*Poly2: ?x*] (4)

όπου:

type *Poly1* (**x*) is [*Poly*: *?x*] <- (*Agnt*) <- [*Touches*] (5)

type *Poly2 (*x*) is [*Poly: ?x*] <- (*Thme*) <- [*Touches*] (6)

Η ερώτηση *Which polygon touches A?* είναι ερώτηση τύπου Q2. Στο σχήμα 3.3, η έννοια *Polygon* αντιστοιχίζεται στην έννοια geo_entity11, η έννοια *Touches* στην έννοια sing/plur_verb και η έννοια *A* στην έννοια geo_entity21. Η ερώτηση μετατρέπεται στον εννοιολογικό γράφο – ερώτηση του σχήματος 2.8.

Σχήμα 2.8 Εννοιολογικός γράφος της ερώτησης *Which polygon touches A?* [KK05b]

Ο συμβολισμός {*} δείχνει ότι η έννοια Polygon χρησιμοποιείται στον πληθυντικό αριθμό. Οι έννοιες που συμμετέχουν στο γράφο είναι οι εξής: [*Polygon: {*}*], [*Touches*] και [*A*]. Οι Polygon και A μπορούν να οριστούν ως ακολούθως:

type *polygon* (**x*) is [*Poly: ?x*] (7)

type *A (*x)* is [*Poly: A*] (8)

Ο εννοιολογικός γράφος του σχήματος 3.6 μετατρέπεται σταδιακά ως εξής:

[*Polygon: {*}?*] <- (*Agnt*) <- [*Touches*] -> (*Thme*) -> [*A*] (Σχήμα 2.6)

[*Poly: {*}?*] <- (*Agnt*) <- [*Touches*] -> (*Thme*) -> [*Poly: A*] (Σχέσεις (7) και (8))

[*Poly1: {*}?*]<- (*Touches*) ->[*Poly2: A*] (9) (Σχέση (4))

Η SQL εντολή *select* εφαρμόζεται στην έννοια που περιέχει το ερωτηματικό δηλαδή στην [*Poly1: {*}?*]. Η σχέση (*Touches*) του γράφου χρησιμοποιείται ως όρισμα για το πεδίο *from* της εντολής *select* και η έννοια [Poly2:A] αντιστοιχίζεται στο πεδίο *where* που προσδιορίζει το φίλτρο επιλογής της *select*. Σαν αποτέλεσμα, παράγεται ο κώδικας SQL:

```
Select poly1
from Touches
where poly2 = 'A'          (10)
```

του οποίου η εκτέλεση οδηγεί στην απάντηση: 'B, D'.

3. Εξόρυξη Γεωγραφικής Γνώσης Βάσει Μεθόδων Επεξεργασίας Φυσικής Γλώσσας και Αναγνώρισης Προτύπων

3.1 Επεξεργασία Φυσικής Γλώσσας

3.1.1. Εισαγωγή

Η φυσική γλώσσα δεν είναι άμεσα ερμηνεύσιμη και επεξεργάσιμη από τα υπολογιστικά συστήματα λόγω της πολυπλοκότητας (complexity) και της αμφισημίας (ambiguity) που παρουσιάζει, καθώς και της εξάρτησής της από τα συμφραζόμενα (context). Το επιστημονικό πεδίο που ασχολείται με το ζήτημα αυτό ονομάζεται *Επεξεργασία Φυσικής Γλώσσας* (Natural Language Processing - NLP) και επικεντρώνεται σε προβλήματα που σχετίζονται με την αλληλεπίδραση ανθρώπου – υπολογιστή, όπως η δυνατότητα παραγωγής και κατανόησης της φυσικής γλώσσας, η αναζήτηση και εξόρυξη πληροφοριών και γνώσης, η αυτόματη μετάφραση, η αναγνώριση του προφορικού λόγου, κλπ.

Στο πλαίσιο της παρούσας διατριβής, ενδιαφερόμαστε για την αξιοποίηση του NLP με σκοπό την εξόρυξη γνώσης από πηγές σημασιολογικής πληροφορίας. Στο γεωχωρικό τομέα, πηγές σημασιολογικής πληροφορίας αποτελούν κυρίως οι γεωχωρικές οντολογίες, τα γλωσσάρια, οι θησαυροί, οι βάσεις γνώσης, κλπ. Σε αυτές τις πηγές, η φυσική γλώσσα χρησιμοποιείται για τη διατύπωση ορισμών γεωγραφικών εννοιών ή εννοιών που άπτονται γενικότερα του γεωχωρικού τομέα.

Μπορεί να θεωρηθεί ότι η σημασιολογική πληροφορία που εμπεριέχεται σε κείμενα γραμμένα σε φυσική γλώσσα, αποτελείται από *σημασιολογικά στοιχεία* (semantic elements) και ειδικότερα από *σημασιολογικές σχέσεις* (semantic relations) και *σημασιολογικές ιδιότητες* (semantic properties) [KK08]. Η αυτόματη αναγνώριση των σημασιολογικών στοιχείων μπορεί να αποδειχθεί χρήσιμη σε πολλές εφαρμογές, όπως για παράδειγμα, στην αυτόματη δημιουργία σημασιολογικών επιγραφών για ομάδες όμοιων δεδομένων [KKK10b], στην

επίτευξη διαλειτουργικότητας μεταξύ ετερογενών συστημάτων γεωγραφικών πληροφοριών, στην αναγνώριση και επίλυση των σημασιολογικών ετερογενειών μεταξύ γεωχωρικών οντολογιών [KK05a], στην αυτόματη δημιουργία γεωχωρικών οντολογιών, στην κατηγοριοποίηση των αποτελεσμάτων αναζήτησης πληροφοριών, σε εφαρμογές οπτικοποίησης πληροφοριών, κ.ά..

3.1.2. Αναγνώριση των Σημασιολογικών Στοιχείων Ορισμών Γεωγραφικών Εννοιών

Παρότι συντάσσονται σε φυσική γλώσσα, οι συλλογές ορισμών (γλωσσάρια, θησαυροί, οντολογίες, κλπ), αποτελούν σπουδαίες πηγές σημασιολογικής πληροφορίας. Από αυτές, η γνώση μπορεί να εξαχθεί με σχεδόν αυτόματο τρόπο, δεδομένου του ειδικού περιεχομένου και τρόπου σύνταξης των ορισμών.

Ειδικότερα, παρατηρείται ότι ένας ορισμός αποτελείται συνήθως από δύο διακριτά μέρη [JHR93]:

1) Το *γένος* (genus), το οποίο περιλαμβάνει τις αναφορές σε ευρύτερους όρους ή υπερώνυμα (hypernyms, broader terms), δηλαδή σε όρους περισσότερο γενικούς από την έννοια που ορίζεται.

2) Τα *διαφοροποιητικά στοιχεία* (differentiae), τα οποία αντιστοιχούν στα υπόλοιπα μέρη του ορισμού και διαφοροποιούν / διακρίνουν την οριζόμενη έννοια από τις υπόλοιπες που διαθέτουν το ίδιο γένος.

Στη μεθοδολογία που προτείνεται στο [KK08], τα γένη (genera) των ορισμών αναγνωρίζονται ως σημασιολογικά στοιχεία τύπου IS-A, ενώ τα διαφοροποιητικά στοιχεία, ως σημασιολογικά στοιχεία άλλου τύπου, π.χ. SHAPE, SIZE, ADJACENCY, κλπ. Η αναγνώριση των σημασιολογικών στοιχείων είναι μια διαδικασία που συνίσταται: (α) στη συντακτική ανάλυση των ορισμών γεωγραφικών εννοιών και (β) στην εφαρμογή κανόνων αναγνώρισης προτύπων για την εξαγωγή γνώσης.

Παρακάτω, παρατίθενται οι Πίνακες 3.1 και 3.2 οι οποίοι συγκεντρώνουν τα κύρια σημασιολογικά στοιχεία, δηλαδή τις κύριες σημασιολογικές ιδιότητες και σχέσεις, που απαντώνται σε ορισμούς γεωγραφικών εννοιών.

SEMANTIC PROPERTIES
PURPOSE
AGENT
PROPERTY-DEFINED LOCATION
COVER
PROPERTY-DEFINED TIME
POINT IN TIME
DURATION
FREQUENCY
SIZE
SHAPE

Πίνακας 3.1 Κύριες σημασιολογικές ιδιότητες γεωγραφικών εννοιών [KK08]

SEMANTIC RELATIONS
IS-A
IS-PART-OF
HAS-PART
RELATIVE POSITION
UPWARD VERTICAL RELATIVE POSITION
DOWNWARD VERTICAL RELATIVE POSITION
IN FRONT OF HORIZONTAL RELATIVE POSITION
BEHIND HORIZONTAL RELATIVE POSITION
BESIDE HORIZONTAL RELATIVE POSITION
SOURCE - DESTINATION
SEPARATION
ADJACENCY
CONNECTIVITY
OVERLAP
INTERSECTION
CONTAINMENT
EXCLUSION
SURROUNDNESS
EXTENSION
PROXIMITY

DIRECTION

Πίνακας 3.2 Κύριες σημασιολογικές σχέσεις γεωγραφικών εννοιών [KK08]

3.2 Εξόρυξη Γεωγραφικής Γνώσης με τη Μέθοδο Αυτόματης Δημιουργίας Επιγραφών Geo-Labeling

3.2.1. Εισαγωγή

Οι μέθοδοι *σημασιολογικής* ομαδοποίησης επιτυγχάνουν την εξόρυξη γνώσης δημιουργώντας ομάδες σημασιολογικά όμοιων δεδομένων και αναγνωρίζοντας κατ' αυτόν τον τρόπο κάποιες εξαρχής μη φανερές ιδιότητες και σχέσεις μεταξύ των δεδομένων ενός συνόλου.

Ενδεικτικά παραδείγματα πληροφοριών εισόδου για τις μεθόδους αυτές, αποτελούν μεταξύ άλλων τα παρακάτω:

- Κείμενα/έγγραφα σε φυσική γλώσσα
- HTML, XML, GML δεδομένα και ιστοσελίδες
- Αποτελέσματα αναζήτησης πληροφοριών από διάφορες συλλογές δεδομένων (βάσεις δεδομένων, θησαυροί, κλπ)

Στη βιβλιογραφία, οι μέθοδοι σημασιολογικής ομαδοποίησης δίνουν συνήθως περισσότερη έμφαση στη δημιουργία ομάδων, παρότι η δημιουργία επιγραφών για τις ομάδες είναι εξίσου σημαντική για την ανάδειξη των αποτελεσμάτων της ομαδοποίησης. Ακόμη, οι μέθοδοι δημιουργίας επιγραφών που έχουν ως τώρα προταθεί βασίζονται περισσότερο στον υπολογισμό της συχνότητας εμφάνισης λέξεων [MR99], [PU00], [TC06], στον προσδιορισμό επιγραφών βάσει των εννοιών που συμμετέχουν σε οντολογίες ανώτερου επιπέδου [SE04] ή στο διαδίκτυο [PR04], παρά στην ίδια τη σημασιολογία των πληροφοριών εισόδου.

Στη συνέχεια, εισάγεται η *Geo-Labeling*, μια νέα μέθοδος δημιουργίας επιγραφών για ομάδες σημασιολογικά όμοιων γεωγραφικών εννοιών που περιγράφονται από ορισμούς σε

φυσική γλώσσα [KKK10b]. Στόχος της Geo-Labeling είναι να αξιοποιήσει τη γνώση που εσωκλείεται σε κάθε ορισμό αντί να ανατρέξει αποκλειστικά σε εξωτερικές πηγές γνώσης (βάσεις γνώσης, διαδίκτυο, οντολογίες, κλπ). Οι επιγραφές δομούνται με τρόπο ώστε αφενός, να συνοψίζουν το σημασιολογικό περιεχόμενο κάθε ομάδας και αφετέρου, να γίνονται άμεσα κατανοητές από τους χρήστες των εφαρμογών που θα χρησιμοποιήσουν τη μέθοδο.

3.2.2. Σχετική Εργασία

Οι μέθοδοι δημιουργίας επιγραφών που βασίζονται στον υπολογισμό της συχνότητας εμφάνισης λέξεων, αφορούν συνήθως στην ομαδοποίηση και κατηγοριοποίηση κειμένων και εγγράφων. Ο υπολογισμός της συχνότητας γίνεται αφού πρώτα εντοπιστούν και διαγραφούν οι λεγόμενες «τερματικές» λέξεις (stop words), οι οποίες βοηθούν στη σύνταξη του κειμένου, χωρίς να προσθέτουν ουσιαστικό νόημα. Παραδείγματα τερματικών λέξεων αποτελούν οι: *the*, *or*, *and*, κλπ. Στη συνέχεια, οι επιγραφές δημιουργούνται ως λίστες που συμπεριλαμβάνουν τις πιο χρησιμοποιούμενες λέξεις. Ωστόσο, εκτός του ότι οι λέξεις αυτές δεν σχετίζονται μεταξύ τους, έχει παρατηρηθεί ότι στις επιγραφές αυτής της μορφής προστίθεται συχνά άχρηστη πληροφορία που δεν απαρτίζεται από τερματικές λέξεις. Για παράδειγμα, στην περίπτωση που η ομαδοποίηση πραγματοποιηθεί μεταξύ επιστημονικών άρθρων σχετικών με συστήματα γεωγραφικών πληροφοριών, οι λέξεις *σύστημα* ή *γεωγραφικός*, παρότι θα προστεθούν στην επιγραφή ως συχνά χρησιμοποιούμενες, δεν θα προσθέσουν παραπάνω πληροφορία σχετικά με το περιεχόμενο μιας ομάδας άρθρων, δεδομένου ότι είναι εξαρχής γνωστό στο χρήστη ότι πρόκειται για άρθρα σχετικά με συστήματα γεωγραφικών πληροφοριών.

Στο πεδίο της αναζήτησης πληροφοριών, οι Merkl και Rauber [MR99] παρουσίασαν μια διαφορετική μέθοδο αυτόματης δημιουργίας επιγραφών για ομάδες επιστημονικών άρθρων. Τα άρθρα, μοντελοποιημένα ως διανύσματα συχνοτήτων όρων, προβάλλονται σε αυτό-οργανούμενο χάρτη (Βλ. §4.4.4), ενώ στη συνέχεια αναγνωρίζονται οι ομάδες όμοιων άρθρων και δημιουργούνται επιγραφές που τοποθετούνταν στους νευρώνες, κέντρα των ομάδων. Η μέθοδος προτάθηκε για να υποστηρίζει τους χρήστες κατά την περιήγησή τους σε πολυπληθείς συλλογές άρθρων, προς αναζήτηση συγκεκριμένων πληροφοριών.

Οι Popescul και Ungar [PU00] πρότειναν δύο μεθόδους δημιουργίας επιγραφών, οι οποίες οδηγούν και οι δύο στη διαμόρφωση επιγραφών με τη μορφή συνόλου όρων. Η πρώτη μέθοδος ομαδοποιεί ιεραρχικά έγγραφα και αναπαριστά τα αποτελέσματα της ομαδοποίησης με τη βοήθεια δενδρογράμματος. Κάθε έγγραφο απεικονίζεται ως κόμβος στο δενδρόγραμμα και του αποδίδεται επιγραφή που αποτελείται από το σύνολο των πιο χρησιμοποιούμενων όρων σε αυτό. Στη συνέχεια, εφαρμόζοντας ελέγχους που οι συγγραφείς ονομάζουν «τεστ ανεξαρτησίας» (tests of independence) X2 [TD00], προσδιορίζουν, για κάθε επιγραφή, τους όρους εκείνους για τους οποίους υπάρχει πιθανότητα να διαχυθούν στους υπο-κόμβους του δενδρογράμματος. Οι όροι αυτοί «βαφτίζονται» πολύ γενικοί για να είναι ικανοποιητικά περιγραφικοί και αποκλείονται από το σύνολο των όρων που συνθέτουν τις επιγραφές των υπο-κόμβων. Η διαδικασία επαναλαμβάνεται από τη ρίζα έως τα φύλλα, μέχρις ότου απαλλαγεί το δενδρόγραμμα από τους περιττούς μη περιγραφικούς όρους. Η δεύτερη μέθοδος πραγματοποιεί αποσαφήνιση της σημασίας των λέξεων (word sense disambiguation) που χρησιμοποιούνται στα έγγραφα, εφαρμόζοντας μια διαδικασία που προτάθηκε από τον Yarowsky [Yaro92], η οποία συνδυάζει τη συχνότητα εμφάνισης των όρων με την προβλεψιμότητά (predictiveness) τους, με τρόπο που να επιλέγονται ως πιο περιγραφικοί για κάθε ομάδα εγγράφων, οι ειδικότεροι όροι και να αποκλείονται, ως μη αρκούντως περιγραφικοί, οι γενικότεροι.

Οι Stein και zuEissen [SE04] πρότειναν μια διαφορετική προσέγγιση για τη δημιουργία επιγραφών, η οποία βασίζεται στην αξιοποίηση οντολογίας. Θέλοντας να κατηγοριοποιήσουν τα αποτελέσματα αναζήτησης πληροφοριών σε ομάδες όμοιων ως προς το περιεχόμενο κειμένων, πραγματοποίησαν αντιστοιχίσεις μεταξύ των ομάδων και των πιο συναφών εννοιών μιας οντολογίας με σχετικό περιεχόμενο. Εάν μια ομάδα αντιστοιχιζόταν σε ακριβώς μια έννοια, τότε της απέδιδαν ως μονοθεματική επιγραφή το όνομα της έννοιας, διαφορετικά της απέδιδαν μια πολυθεματική (polythetic) επιγραφή αποτελούμενη από το σύνολο των συναφών εννοιών της οντολογίας.

Οι Treeratpituk και Callan [TC06] ασχολήθηκαν κι αυτοί με το πρόβλημα της δημιουργίας επιγραφών, για έγγραφα ομαδοποιημένα ιεραρχικά σε δενδρόγραμμα, ανάλογα με τη συνάφεια του περιεχομένου τους. Ισχυρίστηκαν ότι «καλές» επιγραφές είναι αυτές που βοηθούν στη διάκριση μεταξύ των αμφιθαλών (sibling) εγγράφων και αυτών που βρίσκονται

στα ανώτερα ιεραρχικά επίπεδα του δενδρογράμματος, και ως εκ τούτου, η προσέγγισή τους δεν βασίζεται στα στατιστικά στοιχεία που προκύπτουν από τη συχνότητα εμφάνισης των όρων σε κάθε έγγραφο ξεχωριστά, αλλά σε σχέση με τη συχνότητα των ίδιων όρων στα αμφιθαλή έγγραφα και στα έγγραφα ανήκοντα στα ανώτερα ιεραρχικά επίπεδα του δενδρογράμματος. Αρχικά προσδιορίζονται ως υποψήφιες επιγραφές, πενταμελείς λίστες όρων, στις οποίες αποδίδονται εν συνεχεία βαθμοί (scores) ανάλογα με την καταλληλότητά τους. Ακολούθως, οι τελικές επιγραφές επιλέγονται με την εφαρμογή ενός μοντέλου αποκοπής (cut-off model) που προσδιορίζει τον αριθμό των κατάλληλων περιγραφικών όρων.

Ως σχετική εργασία, μπορούμε επιπροσθέτως να αναφέρουμε τις εφαρμογές αυτόματης ιεράρχησης πληροφοριών (automatic information hierarchy creation). Προς αυτήν την κατεύθυνση ο Caraballo [Cara99] δημιούργησε από το περιεχόμενο ενός κειμένου, μια ιεραρχία όρων και υπερωνύμων, χρησιμοποιώντας μια συσσωρευτική ιεραρχική μέθοδο ομαδοποίησης (Βλ. §4.3.2). Αρχικά καθορίστηκαν υποψήφιες επιγραφές για τις ομάδες όμοιων όρων, με τη βοήθεια τεχνικών αναγνώρισης προτύπων που είχαν ως στόχο την εύρεση σχέσεων υπερωνύμων/υπονύμων (hypernyms/hyponyms) με βάση τους χρησιμοποιούμενους γλωσσικούς συνδέσμους. Στη συνέχεια, τα ουσιαστικά με τον μεγαλύτερο αριθμό υπονύμων στο κείμενο αποδόθηκαν ως τελικές επιγραφές στις ομάδες.

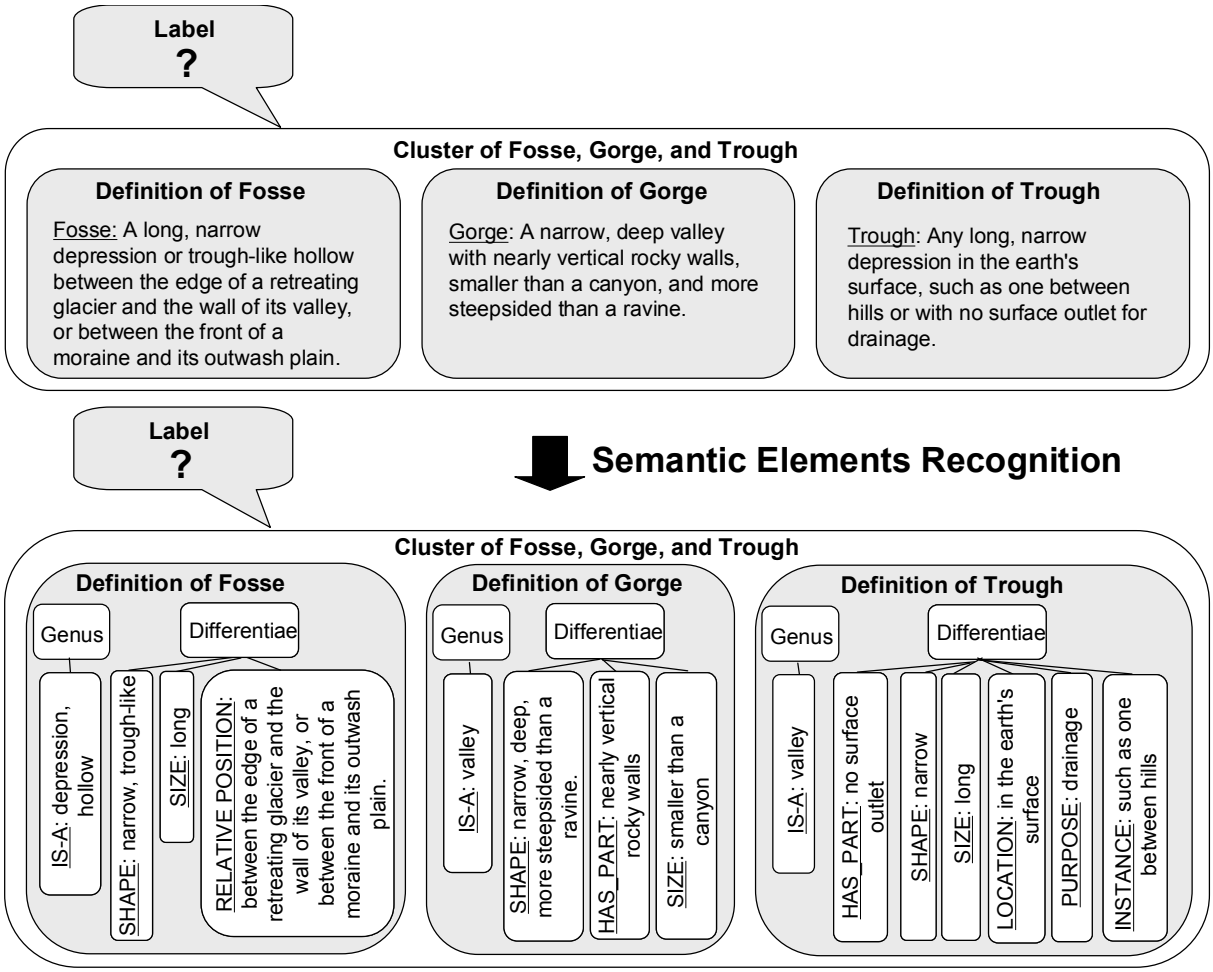
Ακόμη, με σκοπό την αυτοματοποιημένη επέκταση των γλωσσικών πηγών γνώσης, οι Pantel και Ravichandran [PR04] πρότειναν μια μέθοδο για τον αυτόματο προσδιορισμό σημασιολογικών ομάδων όρων μέσω του διαδικτύου και τη δημιουργία επιγραφών για τις ομάδες αυτές, χρησιμοποιώντας έναν αλγόριθμο ομαδοποίησης το οποίο ονόμασαν CBC (Clustering by Committee). Αρχικά, ο αλγόριθμος αυτός ανακαλύπτει μέσω του διαδικτύου σημασιολογικά όμοιους όρους και τους ομαδοποιεί. Οι όροι εξετάζονται ως προς την ύπαρξη προτύπων που αναδεικνύονται από σχέσεις υπερωνύμων/υπονύμων και στη συνέχεια, σε κάθε τέτοια σχέση, αποδίδεται βαθμολογία ανάλογη με το ποσό της πληροφορίας κοινής στα μέλη κάθε ομάδας. Το όνομα κάθε ομάδας, δηλαδή η επιγραφή της, σχηματίζεται ως λίστα με τους πέντε όρους που συγκεντρώνουν την υψηλότερη βαθμολογία. Οι όροι αυτοί θεωρούνται ως οι πιο αντιπροσωπευτικοί για την ομάδα.

3.2.3. Η Διαδικασία Εξόρυξης Γνώσης της Geo-Labeling

Για να δημιουργήσει επιγραφές, η Geo-Labeling ξεκινά με την εξόρυξη της γνώσης που εσωκλείεται στους ορισμούς των γεωγραφικών εννοιών. Η διαδικασία που ακολουθείται βασίζεται στη μεθοδολογία που παρουσιάζεται στις εργασίες [Kokl08] και [KK08], και περιγράφεται συνοπτικά στην παράγραφο §3.2. Ειδικότερα, η εξόρυξη γνώσης συνίσταται: (α) στη συντακτική ανάλυση των ορισμών των γεωγραφικών εννοιών και (β) στην εφαρμογή κανόνων αναγνώρισης προτύπων για την εξαγωγή της γνώσης υπό τη μορφή *σημασιολογικών στοιχείων* (Βλ. §3.1.1). Τα γένη (genera) των ορισμών αναγνωρίζονται ως σημασιολογικά στοιχεία τύπου IS-A ενώ τα διαφοροποιητικά στοιχεία ως σημασιολογικά στοιχεία π.χ. τύπου SHAPE, SIZE, ADJACENCY, κλπ.

Για παράδειγμα, το σχήμα 3.1 συγκεντρώνει τα σημασιολογικά στοιχεία που αναγνωρίστηκαν στους ορισμούς των γεωγραφικών εννοιών *Fosse*, *Gorge* και *Trough*, στο γλωσσάριο Glossary of Landform and Geologic Terms (USA National Resources Conservation Service[3]). Στη συνέχεια, θα χρησιμοποιηθεί αυτό το μικρό παράδειγμα για την περιγραφή των διαδοχικών βημάτων που ακολουθεί η μέθοδος.

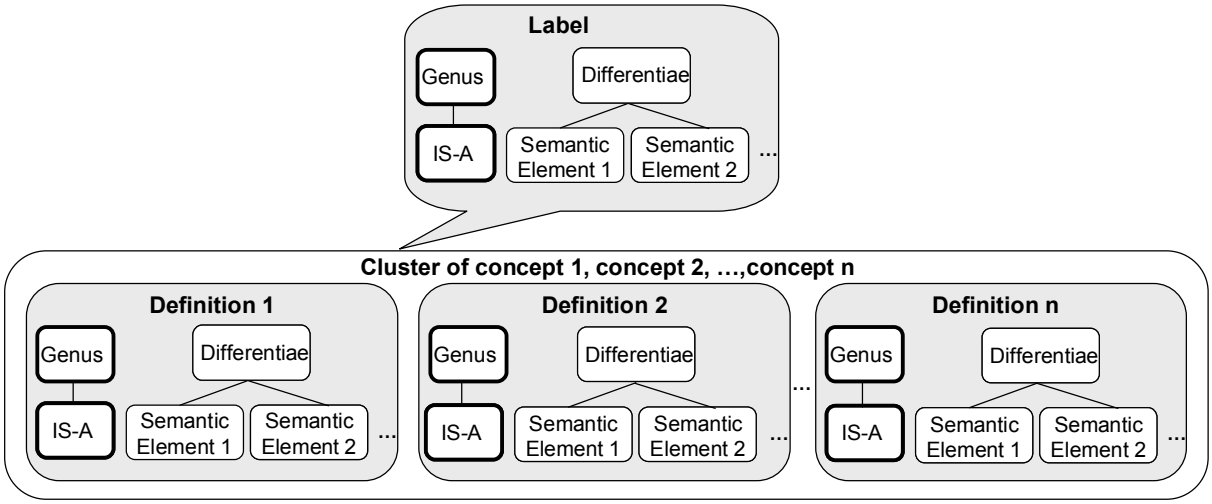
[3] Διαθέσιμο στο σύνδεσμο http://soils.usda.gov/technical/handbook/contents/part629.html, τελευταία προσπέλαση Ιούνιος 2010.



Σχήμα 3.1 Παράδειγμα αναγνώρισης σημασιολογικών στοιχείων [KKK10b]

3.2.4. Βήμα Πρώτο: Προσδιορισμός του Γένους Μιας Επιγραφής

Η Geo-Labeling σχηματίζει επιγραφές με δομή που θυμίζει ορισμούς εννοιών, δηλαδή προτάσεις διατυπωμένες σε φυσική γλώσσα, οι οποίες αποτελούνται από ένα *τμήμα – γένος* και ένα *τμήμα – διαφοροποιητικά στοιχεία*, μόνο που, αυτή τη φορά, το γένος και τα διαφοροποιητικά στοιχεία περιγράφουν συνοπτικά το σύνολο των εννοιών που εσωκλείονται σε κάθε ομάδα (Σχήμα 3.2).



Σχήμα 3.2 Δομή ορισμών γεωγραφικών εννοιών και επιγραφών [KKK10b]

Ο προσδιορισμός του γένους μιας επιγραφής προϋποθέτει τα ονόματα των ομαδοποιημένων εννοιών και τα γένη τους (δηλαδή τα αναγνωρισμένα σημασιολογικά στοιχεία τύπου IS-A) να αποσαφηνιστούν και να εμπλουτιστούν με επιπλέον πληροφορία. Η διαδικασία αυτή συνίσταται στην αντιστοίχισή τους με συνώνυμα και υπερώνυμα τα οποία παρέχονται από βάσεις γνώσης παρόμοιες με τη γνωστή βάση γνώσης Wordnet[4].

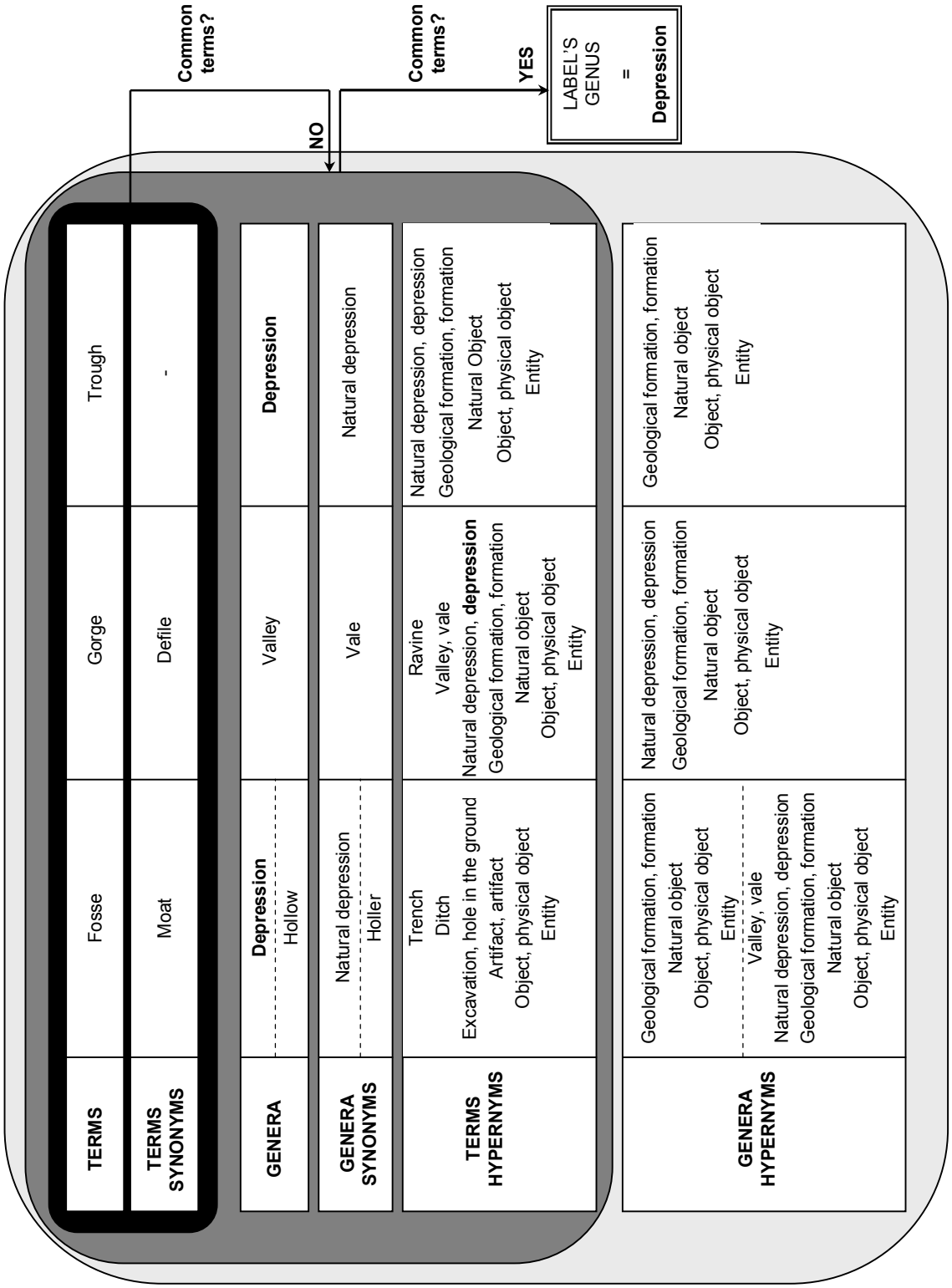
Αφού ολοκληρωθούν η αποσαφήνιση και ο εμπλουτισμός, *ο κοινός και πιο ειδικός όρος* μεταξύ των ακολούθων επιλέγεται ως το γένος της επιγραφής:

1) Τα ονόματα των εννοιών και τα συνώνυμα τους,
2) Τα γένη και τα συνώνυμα των γενών, καθώς και τα υπερώνυμα των εννοιών,
3) Τα υπερώνυμα των γενών.

[4] Διαθέσιμη στο σύνδεσμο http://wordnet.princeton.edu/, τελευταία προσπέλαση Ιούνιος 2010.

Η σειρά από 1) έως 3) ορίζει μια ταξινόμηση από τους πιο ειδικούς όρους στους πιο γενικούς. Προφανώς, *όσο περισσότερο διαφέρουν σημασιολογικά οι ομαδοποιημένες έννοιες, τόσο πιο γενικός θα είναι και ο όρος που θα επιλεγεί για να αποτελέσει το γένος της επιγραφής*.

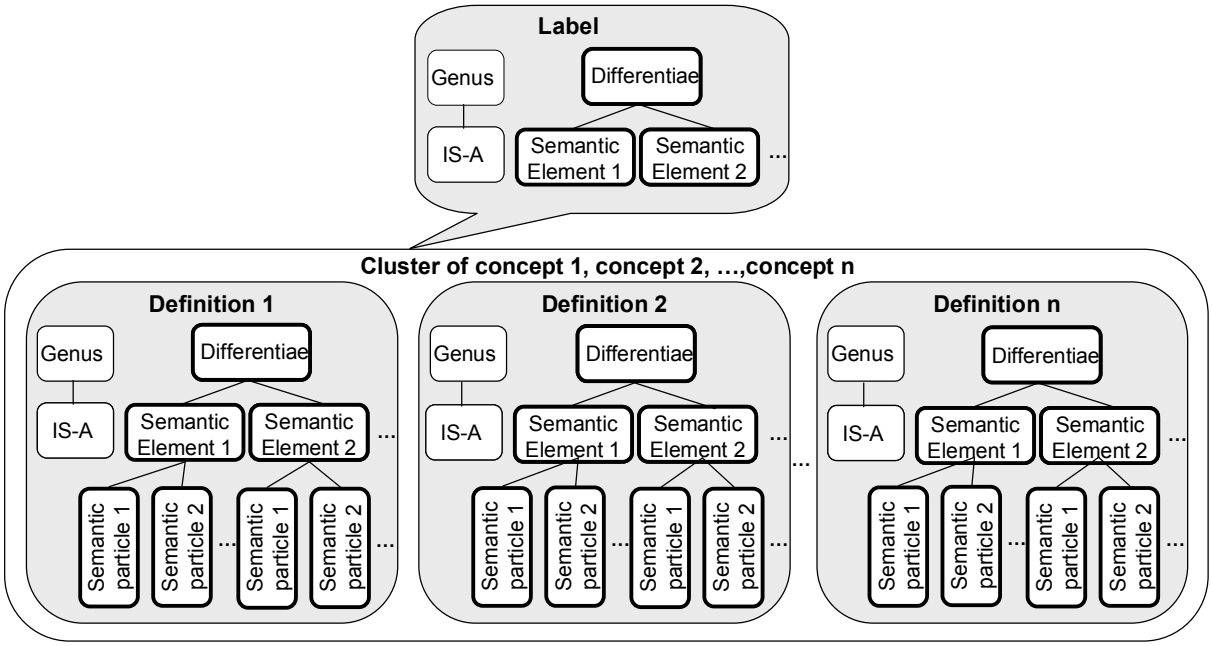
Το σχήμα 3.3 δείχνει πως επιλέγεται ο όρος *Depression* ως το γένος της επιγραφής της ομάδας αποτελούμενης από τις έννοιες *Fosse*, *Gorge* and *Trough*.



Σχήμα 3.3 Παράδειγμα προσδιορισμού του γένους μιας επιγραφής [KKK10b]

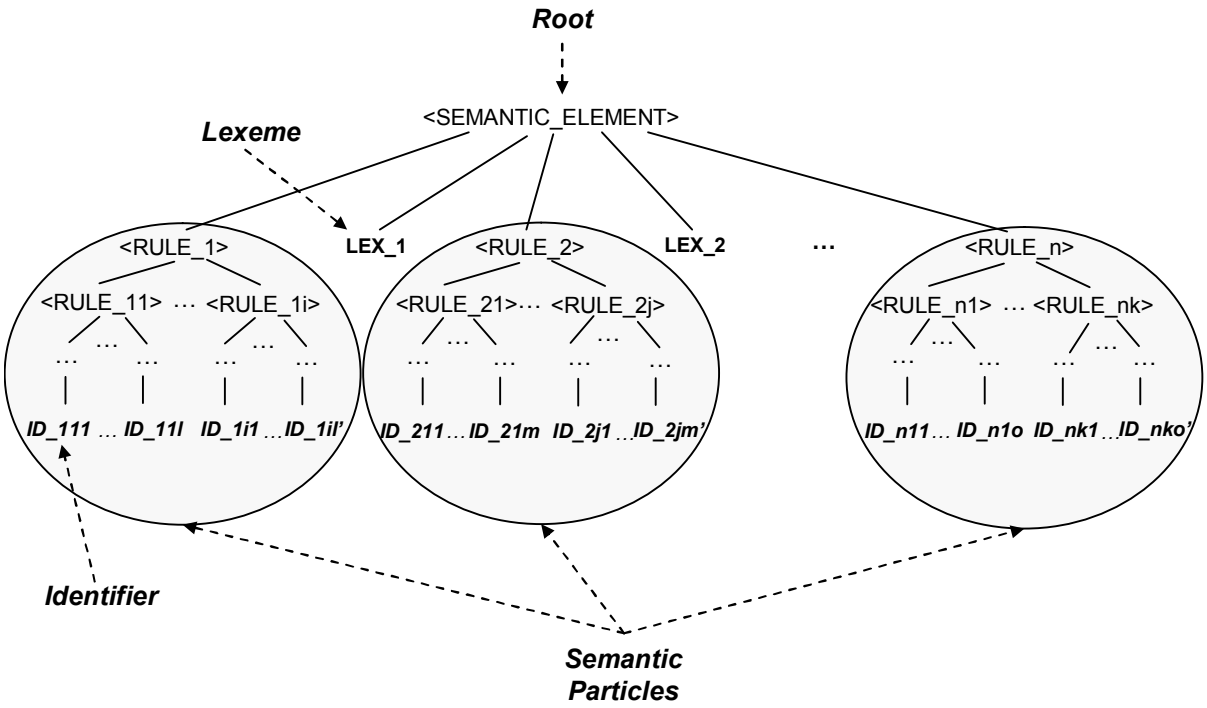
3.2.5. Βήμα Δεύτερο: Προσδιορισμός των Διαφοροποιητικών Στοιχείων μιας Επιγραφής

Ο προσδιορισμός των διαφοροποιητικών στοιχείων μιας επιγραφής απαιτεί να βρεθεί *το μεγαλύτερο τμήμα πληροφορίας που μοιράζονται τα σημασιολογικά στοιχεία των ορισμών των ομαδοποιημένων εννοιών*. Για το σκοπό αυτό, τα σημασιολογικά στοιχεία διαιρούνται σε ακόμη μικρότερες και συνάμα συγκρίσιμες ποσότητες πληροφορίας, οι οποίες καλούνται *σημασιολογικά μόρια* (semantic particles) (Σχήμα 3.4).



Σχήμα 3.4 Διαίρεση σημασιολογικών στοιχείων σε σημασιολογικά μόρια [KKK10b]

Αυτή η περεταίρω διαίρεση επιτυγχάνεται διατρέχοντας και αναλύοντας τα σημασιολογικά στοιχεία και αναγνωρίζοντας για κάθε τύπο στοιχείων συγκεκριμένα και προκαθορισμένα πρότυπα. Η διαδικασία αυτή οδηγεί στη διαμόρφωση ενός δέντρου συντακτικής ανάλυσης (parse tree) για κάθε σημασιολογικό στοιχείο. Τα δέντρα συντακτικής ανάλυσης ξεκινούν από έναν κόμβο – ρίζα, από τον οποίο «κρέμονται» τα κύρια υποδέντρα που αντιστοιχούν στα σημασιολογικά μόρια, και τελειώνουν σε φύλλα που αντιστοιχούν σε τερματικά σύμβολα (terminal tokens), τα οποία είναι είτε λεκτικές μονάδες (lexemes) (π.χ. *and*, *not*, κλπ), είτε αναγνωριστικά (identifiers) (π.χ. *narrow*, *island*, κλπ) (Σχήμα 3.5).



Σχήμα 3.5 Δέντρο συντακτικής ανάλυσης και σημασιολογικά μόρια [KKK10b]

Ως ενδεικτικό παράδειγμα, περιγράφεται ακολούθως η διαίρεση του σημασιολογικού στοιχείου <SHAPE> υπό τη μορφή κανόνων Backus-Naur[5], η οποία οδηγεί στην αναγνώριση των σημασιολογικών μορίων που απεικονίζονται στο σχήμα 3.6.

<SHAPE> ::= {*NOT*|*NO*} <SHP_PROP> [[{, {*NOT*|*NO*} <SHP_PROP>}] *AND* {*NOT*|*NO*} <SHP_PROP>]

<SHP_PROP> ::= <QUALIFIER>|<REL_SHP_PROP>

<REL_SHP_PROP> ::= [[{*MORE*|*LESS*} <QUALIFIERS> *THAN*] | [*AS* <QUALIFIERS> *AS*]] <ENTITIES>

[5] International Standard ISO/IEC 14977, διαθέσιμο στο σύνδεσμο http://www.iso.org, τελευταία προσπέλαση Ιούνιος 2010.

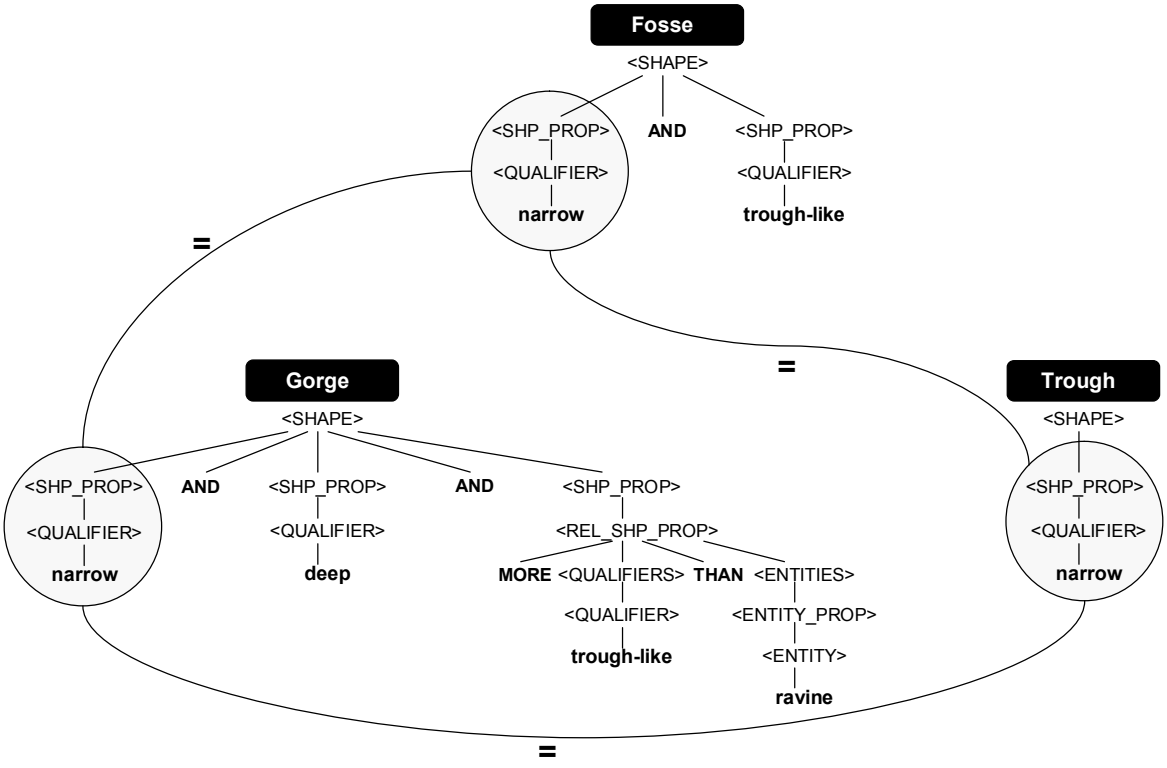
<ENTITIES> ::= <ENTITY> [[{, <ENTITY>}] (*AND* | *OR*) <ENTITY_PROP>]

<ENTITY_PROP> ::= [<QUALIFIERS>] <ENTITY> [<QUALIFIERS>]

<QUALIFIERS> ::= <QUALIFIER> [[{, <QUALIFIER>}] *AND* <QUALIFIER>]

<QUALIFIER> ::= *narrow* | *straight* | *rectangular* | *circular* | … (κλπ, επίθετα εβρισκόμενα στη βάση γνώσης)

<ENTITIES> ::= *island* | *river* | *valley* | *lagoon* | *gorge* | … (κλπ, ουσιαστικά εβρισκόμενα στη βάση γνώσης)



Σχήμα 3.6 Παράδειγμα προσδιορισμού κοινής σημασιολογικής πληροφορίας μέσω της σύγκρισης σημασιολογικών μορίων [KKK10b]

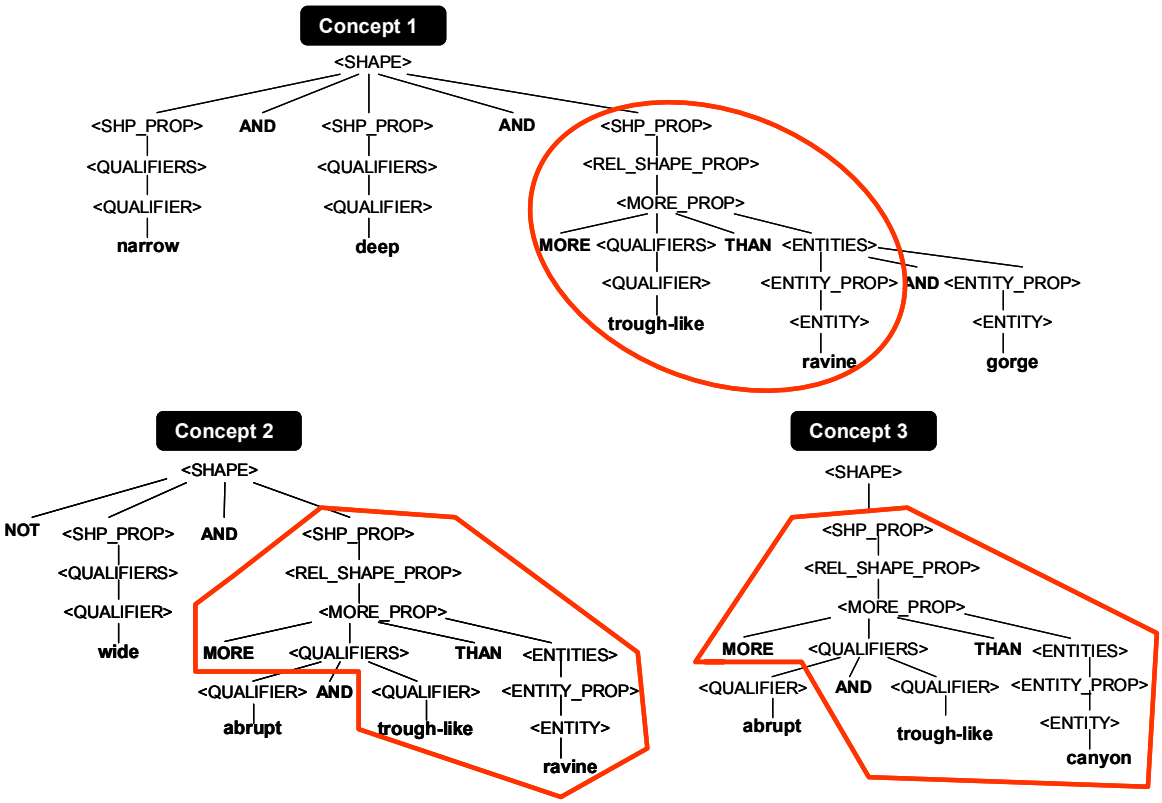
Κατά τη Geo-Labeling, τα σημασιολογικά μόρια τα οποία αντιστοιχούν σε σημασιολογικά όμοια πληροφορία, οφείλουν να διαθέτουν την ίδια συντακτική δομή, η οποία μάλιστα οφείλει να τερματίζεται σε σημασιολογικά όμοια αναγνωριστικά. Η σύγκριση των αναγνωριστικών πραγματοποιείται λαμβάνοντας υπόψη την ακολουθία των διαδοχικών μη-τερματικών συμβόλων (nonterminal tokens) από τα οποία προέρχονται.

Όσον αφορά το σημασιολογικό στοιχείο <SHAPE>, τα αναγνωριστικά είναι είτε τύπου <QUALIFIER> είτε τύπου <ENTITY>. Στην πρώτη περίπτωση, δύο αναγνωριστικά τύπου <QUALIFIER> θεωρούνται σημασιολογικά όμοια εάν συμβαίνει ένα από τα ακόλουθα: 1) αντιπροσωπεύουν το ίδιο επίθετο, 2) αποτελούν συνώνυμα επίθετα, ή 3) μοιράζονται συνώνυμα επίθετα. Εάν οι λεκτικές μονάδες *not* ή *no* προηγούνται των αναγνωριστικών, θεωρείται ότι τα αναγνωριστικά αποτελούν αντώνυμα ή μοιράζονται κοινά αντώνυμα. Στη δεύτερη περίπτωση, δύο αναγνωριστικά τύπου <ENTITY> είναι σημασιολογικά όμοια εάν συμβαίνει ένα από τα ακόλουθα: 1) αντιπροσωπεύουν την ίδια έννοια, 2) αποτελούν συνώνυμες έννοιες ή 3) μοιράζονται κάποια συνώνυμη έννοια.

Η διαίρεση του σημασιολογικού στοιχείου <SHAPE>, κοινού στις έννοιες *Fosse*, *Gorge* και *Trough*, οδηγεί στο συμπέρασμα ότι το μεγαλύτερο τμήμα πληροφορίας που μοιράζονται οι έννοιες αυτές υποδεικνύεται από τα σημασιολογικά τους μόρια που τερματίζονται στο αναγνωριστικό *narrow* (Σχήμα 3.6). Κατ' ανάλογο τρόπο, θα μπορούσε να αποδειχθεί ότι οι έννοιες αυτές δεν μοιράζονται καμία σημασιολογική πληροφορία σχετική με το κοινό σημασιολογικό τους στοιχείο <SIZE>.

Έτσι, σε αυτό το παράδειγμα μικρής έκτασης, τα διαφοροποιητικά στοιχεία της επιγραφής της ομάδας που συνθέτουν οι έννοιες *Fosse*, *Gorge* και *Trough*, συνοψίζονται στην έκφραση *narrow* και η πλήρης περιγραφή της επιγραφής γίνεται *Depression, narrow*.

Στο σχήμα 3.7 απεικονίζεται ένα παρόμοιο αλλά πολυπλοκότερο παράδειγμα, με τρεις υποθετικές έννοιες, *Concept1*, *Concept2* και *Concept3*, οι οποίες διαθέτουν όλες το σημασιολογικό στοιχείο <SHAPE>. Από τη διαίρεση του στοιχείου αυτού σε σημασιολογικά μόρια, αποδεικνύεται ότι τα διαφοροποιητικά στοιχεία της επιγραφής της ομάδας των εννοιών αυτών σχηματίζουν την έκφραση *More trough-like than ravine*. Στην περίπτωση αυτή, πρέπει ωστόσο να αποσαφηνιστεί με τη βοήθεια της βάσης γνώσης ότι η έννοια *Ravine* είναι συνώνυμη της έννοιας *Canyon*.



Σχήμα 3.7 Παράδειγμα αναγνώρισης κοινής σημασιολογικής πληροφορίας μέσω της σύγκρισης σημασιολογικών μορίων

3.2.6. Σενάρια Εφαρμογής της Geo-Labeling

a) Αυτόματη Δημιουργία Γεωχωρικής Οντολογίας

Η κεντρική ιδέα της αυτόματης δημιουργίας οντολογίας έγκειται στη δημιουργία μιας ιεραρχίας εννοιών συσχετιζόμενων μεταξύ τους με σχέσεις υπερώνυμου/υπώνυμου (IS-A), βάσει ενός αρχικού συνόλου μη δομημένων ή ημι-δομημένων δεδομένων, όπως για παράδειγμα βάσει κειμένων σε φυσική γλώσσα, γλωσσάριων, δεδομένων σε XML, βάσεων δεδομένων, κλπ. Στην τεχνική έκθεση [BN07], περιγράφονται περιληπτικά οι τρέχουσες και σημαντικότερες εξελίξεις σχετικά με το ζήτημα αυτό.

Σε αυτό το σενάριο εφαρμογής, αναδεικνύεται ο τρόπος που η Geo-Labeling δύναται να αξιοποιηθεί με σκοπό την αυτόματη δημιουργία γεωχωρικής οντολογίας. Ειδικότερα,

περιγράφεται πως, από ένα σύνολο ορισμών γεωγραφικών εννοιών διατυπωμένων σε φυσική γλώσσα, είναι δυνατόν να δημιουργηθεί μια ιεραρχία γεωγραφικών εννοιών, η οποία θα μπορούσε να αποτελέσει τη βάση για την ανάπτυξη μιας πιο ολοκληρωμένης γεωχωρικής οντολογίας.

Συγκεκριμένα, προτείνεται η δημιουργία της ιεραρχίας γεωγραφικών εννοιών να βασιστεί στις τρεις ακόλουθες αρχές:

($1^{η}$ Αρχή)

Το όνομα μιας έννοιας και τα συνώνυμα αυτού αποτελούν πιθανότατα πιο ειδικούς όρους από το γένος ή τα συνώνυμα του γένους που εμπεριέχονται στον ορισμό της. Έτσι, εάν βρεθεί ότι το όνομα μιας έννοιας X ή κάποιο από τα συνώνυμα του, ταυτίζεται με το γένος ή με κάποιο από τα συνώνυμα του γένους μιας δεύτερης έννοιας Y, τότε θεωρείται ότι το όνομα της X αποτελεί υπερώνυμο του ονόματος της Y.

($2^{η}$ Αρχή)

Εάν δύο έννοιες X και Y μοιράζονται, βάσει του ορισμού τους, το ίδιο γένος (ή κάποιο συνώνυμο του γένους τους), τότε οι X και Y θεωρούνται αμφιθαλείς έννοιες μιας γενικότερης έννοιας, έστω Z, της οποίας το όνομα ταυτίζεται με αυτό το κοινό γένος (ή το συνώνυμό του). Στην περίπτωση αυτή, θεωρείται ότι το όνομα της Z αποτελεί υπερώνυμο των ονομάτων των X και Y.

($3^{η}$ Αρχή)

Εάν τα ονόματα δύο εννοιών X και Y μοιράζονται βάσει του ορισμού τους ένα κοινό υπερώνυμο, τότε θεωρείται ότι αποτελούν ειδικότερες και αμφιθαλείς έννοιες μιας γενικότερης, έστω Z, της οποίας το όνομα ταυτίζεται με αυτό το κοινό υπερώνυμο.

Τα βήματα του αλγορίθμου που θα «χτίσουν» την ιεραρχία των γεωγραφικών εννοιών και που βασίζονται στις παραπάνω αρχές, ακολουθούν μια προσέγγιση «από-κάτω-προς-τα-πάνω» (bottom up) και περιγράφονται ως ακολούθως:

1) Από ένα αρχικό σύνολο γεωγραφικών εννοιών, **εντοπισμός** εκείνων των εννοιών που τα ονόματά τους αποτελούν υπερώνυμα ή υπώνυμα για τα ονόματα κάποιων άλλων εννοιών του συνόλου. Εφαρμόζοντας την 1η Αρχή, **δημιουργία** των αντίστοιχων IS-A σχέσεων.

2) **Αφαίρεση** από το αρχικό σύνολο εννοιών, εκείνων των οποίων το όνομα αποτελεί υπώνυμο κάποιου άλλου, και **εντοπισμός** των εννοιών που μοιράζονται το ίδιο γένος ή κάποιο από τα συνώνυμα των γενών τους. Για τις έννοιες αυτές, εφαρμόζοντας την 2η Αρχή, **δημιουργία** νέων γενικότερων εννοιών και των αντίστοιχων IS-A σχέσεων, και **απόδοση ονομάτων** σε αυτές βάσει της Geo-Labeling.

3) **Αφαίρεση** από το αρχικό σύνολο εννοιών, εκείνων των οποίων το όνομα αποτελεί (βάσει του προηγούμενου βήματος), υπώνυμο κάποιου άλλου ονόματος, και **εντοπισμός** των εννοιών που μοιράζονται την πλησιέστερη γενικότερη έννοια, χρησιμοποιώντας την ιεραρχία εννοιών που εμπεριέχεται στη βάση γνώσης. Εφαρμόζοντας την 3η αρχή, **δημιουργία** νέων γενικότερων εννοιών και των αντίστοιχων IS-A σχέσεων, και **απόδοση ονόματος** σε αυτές βάσει της Geo-Labeling.

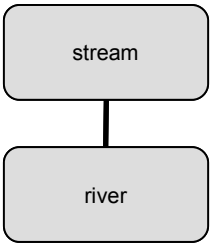
4) **Επανάληψη** του 3ου βήματος μέχρις ότου όλες οι έννοιες του αρχικού συνόλου τοποθετηθούν κάτω από γενικότερες έννοιες και δημιουργηθεί τελικά η έννοια – ρίζα της ιεραρχίας.

Για να γίνουν πιο απτά τα προηγούμενα, περιγράφεται στη συνέχεια ένα συγκεκριμένο παράδειγμα εφαρμογής. Υποθέτουμε ότι ο χρήστης ενδιαφέρεται να δημιουργήσει μια ιεραρχία γεωγραφικών εννοιών συσχετιζόμενων μεταξύ τους με σχέσεις IS-A, ακολουθώντας τον προηγούμενο αλγόριθμο και χρησιμοποιώντας τους ορισμούς που παρατίθενται στον Πίνακα 3.3. Η διαδικασία ξεκινά με την αναγνώριση των σημασιολογικών στοιχείων των ορισμών, εφαρμόζοντας τη μεθοδολογία που περιγράφεται στα [Kokl08] και [KK08].

Definitions	Semantic elements
Sea: Body covered with salt water.	IS-A: body COVER: salt water
Stream: Natural flowing body of fresh water.	IS-A: body COVER: fresh water NATURE: natural FLOW: flowing
River: Natural stream of water, normally of large volume.	IS-A: stream COVER: water NATURE: natural SIZE: large volume
Lake: Body of water surrounded by land.	IS-A: body COVER: water SURROUNDNESS: land
Marsh: Wet land with grassy vegetation.	IS-A: land COVER: grassy vegetation CHRC: wet PURPOSE: boats or irrigation
Bog: Ground with decomposing vegetation.	IS-A: ground COVER: decomposing vegetation

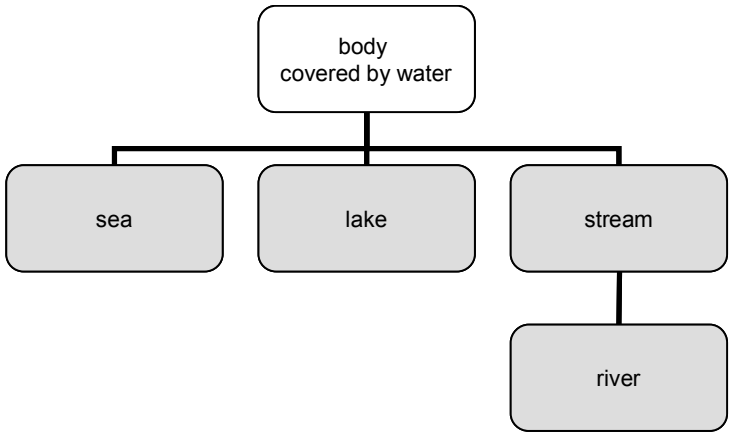
Πίνακας 3.3 Παράδειγμα αρχικού συνόλου γεωγραφικών εννοιών και των αναγνωρισμένων σημασιολογικών στοιχείων τους

Εκτελώντας το βήμα 1, ο αλγόριθμος αναγνωρίζει ότι η έννοια *Stream* έχει όνομα που ταυτίζεται με το γένος της έννοιας *River*, συνεπώς το όνομα της *Stream* αποτελεί υπερώνυμο του ονόματος της *River*. Δημιουργείται έτσι, η πρώτη IS-A σχέση της ιεραρχίας γεωγραφικών εννοιών μεταξύ των *Stream* και *River* (Σχήμα 3.8).



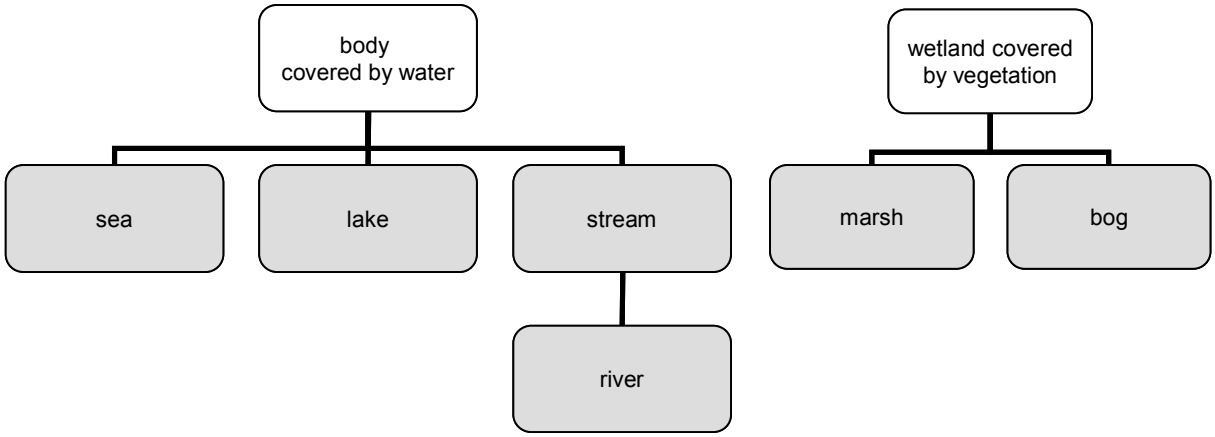
Σχήμα 3.8 Δημιουργία IS-A σχέσης μεταξύ των γεωγραφικών εννοιών *Stream* και *River*

Στη συνέχεια, μεταξύ των εννοιών του αρχικού συνόλου και έχοντας αφαιρέσει την έννοια *River*, οι *Stream* και *Lake* αναγνωρίζονται ως έννοιες που μοιράζονται το ίδιο γένος, *Body*. Εφαρμόζοντας το 2^{ο} βήμα του αλγορίθμου, δημιουργείται μια καινούργια και γενικότερη έννοια, στην οποία αποδίδεται το όνομα *Body covered by water* από τη Geo-Labeling. Η ιεραρχία παίρνει τη μορφή που φαίνεται στο σχήμα 3.9.



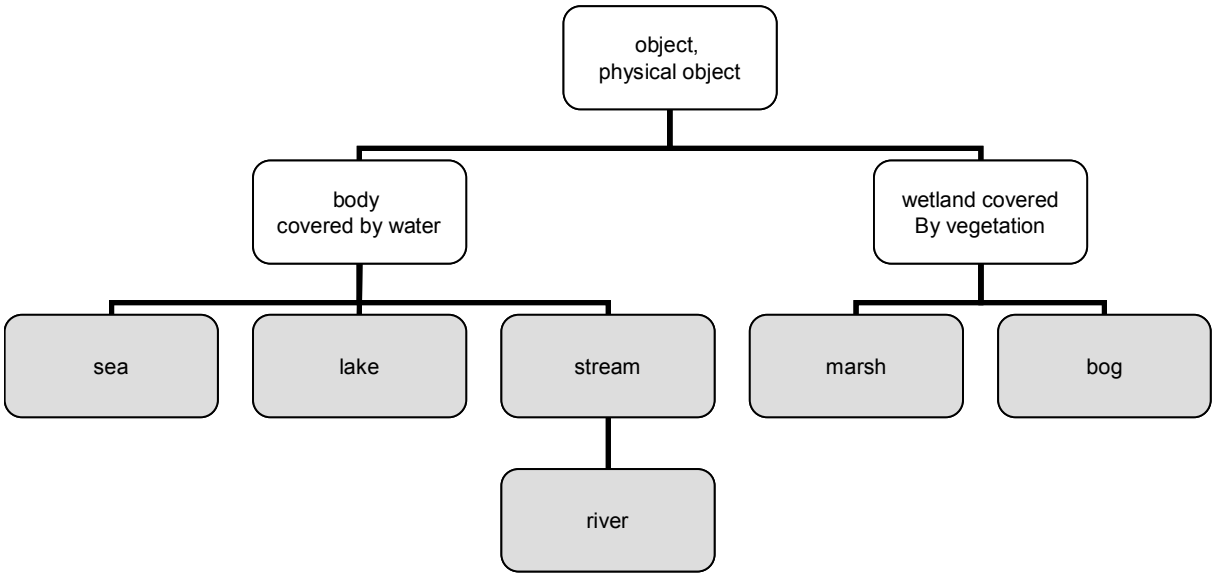
Σχήμα 3.9 Δημιουργία της έννοιας *Body covered by water*

Ακολούθως, οι *Marsh* και *Bog* αναγνωρίζονται ως έννοιες που μοιράζονται την πλησιέστερη γενικότερη έννοια Wetland (με βάση το Wordnet). Εφαρμόζοντας το 3^{ο} βήμα του αλγορίθμου, δημιουργείται μια καινούργια γενικότερη έννοια, στην οποία αποδίδεται από τη Geo-Labeling το όνομα *Wetland covered by vegetation*. Η ιεραρχία παίρνει τότε τη μορφή που φαίνεται στο σχήμα 3.10.



Σχήμα 3.10 Δημιουργία της έννοιας *Wetland covered by vegetation*

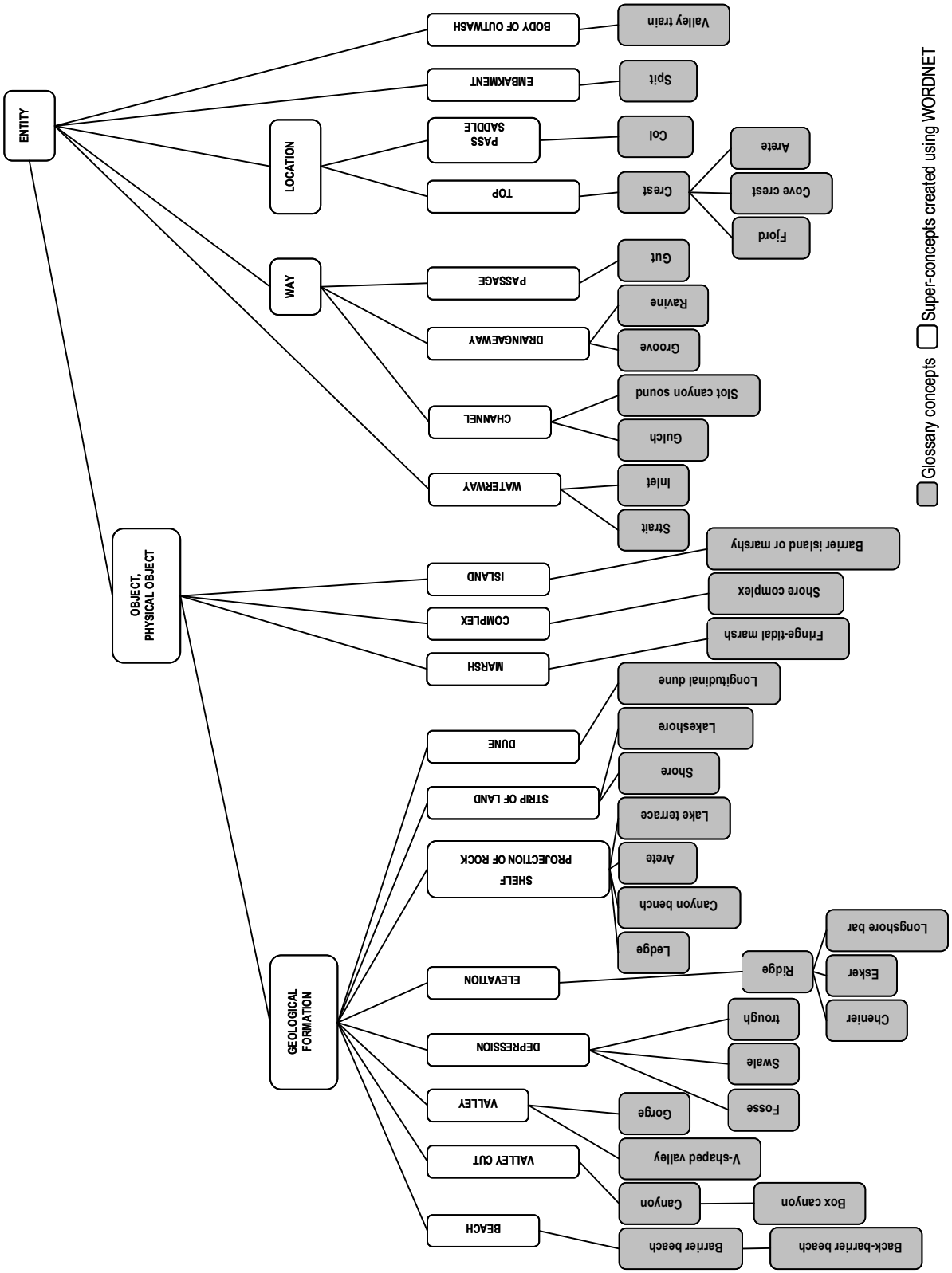
Τέλος, βάσει της βάσης γνώσης, διαπιστώνεται ότι οι έννοιες *Body covered by water* και *Wetland covered by vegetation*, με γένη *Body* και *Wetland* αντίστοιχα, μοιράζονται την πλησιέστερη γενικότερη έννοια *Object, physical object*. Εφαρμόζοντας πάλι και για τελευταία φορά το 3^{ο} βήμα του αλγορίθμου, διαμορφώνεται η ιεραρχία που απεικονίζεται στο σχήμα 3.11.



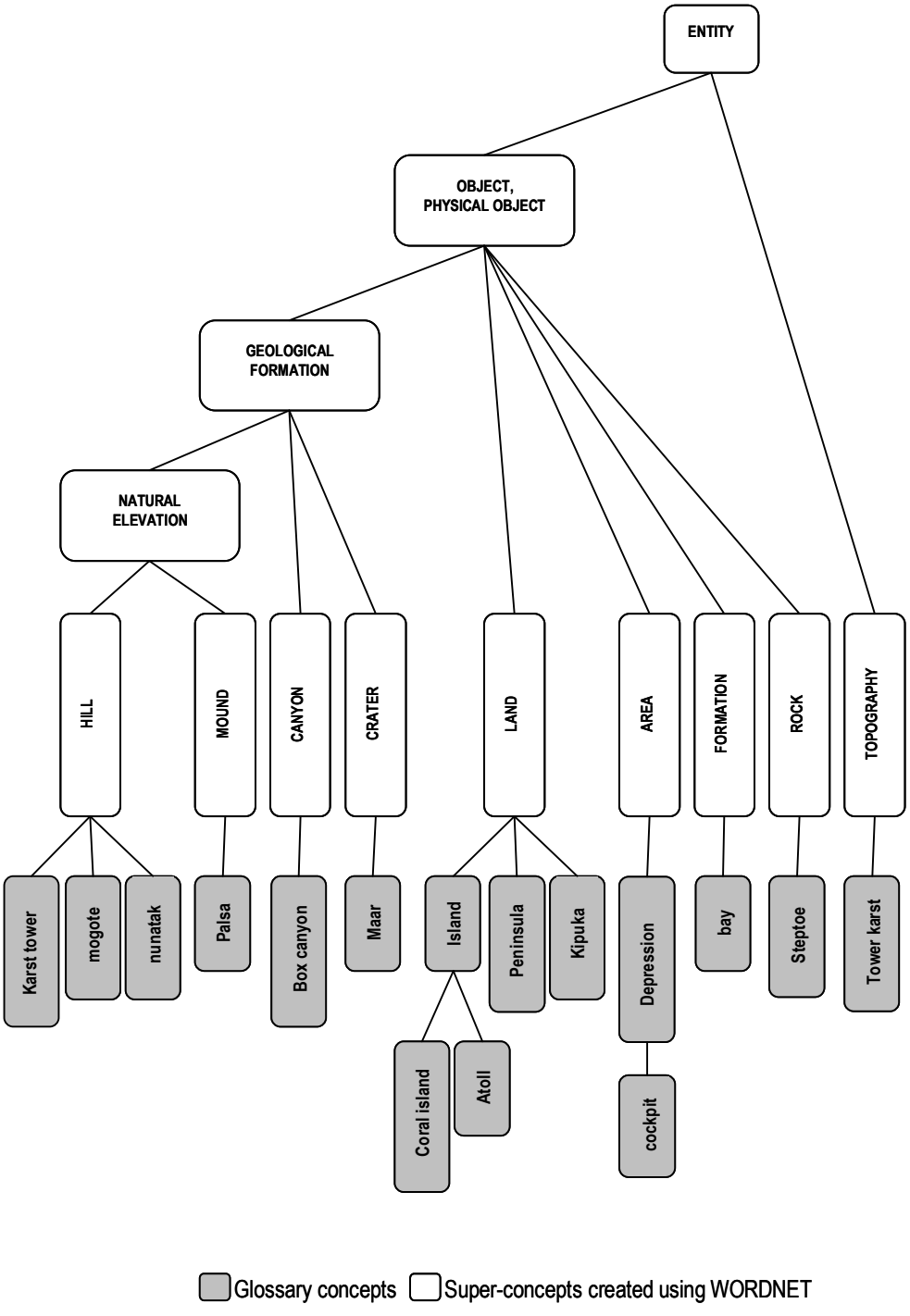
Σχήμα 3.11 Τελική μορφή ιεραρχίας γεωγραφικών εννοιών

Στη συνέχεια, παρουσιάζονται δύο παραδείγματα όπου ο αλγόριθμος εφαρμόζεται σε πραγματικά δεδομένα και συγκεκριμένα στους ορισμούς που παρέχονται από το γλωσσάριο Glossary of Landform and Geologic Terms της υπηρεσίας USA National Resources Conservation Service, στο οποίο αναφερθήκαμε σε προηγούμενη σελίδα (Βλ. §3.2.3).

Ωστόσο, για λόγους περιορισμένου χώρου ανάπτυξης των παραδειγμάτων, 1) θεωρήσαμε ως αρχικό σύνολο ορισμών για το πρώτο παράδειγμα, το υποσύνολο του γλωσσάριου που περιλαμβάνει τους ορισμούς γεωλογικών μορφωμάτων μακρόστενου σχήματος και για το δεύτερο παράδειγμα, το υποσύνολο των ορισμών των γεωλογικών μορφωμάτων που περικλείονται από κάποια γεωγραφική οντότητα και 2) εφαρμόσαμε τη Geo-Labeling, μόνο όσον αφορά το γένος των επιγραφών, παραλείποντας το τμήμα των επιγραφών που αναφέρεται στα διαφοροποιητικά στοιχεία. Τελικά, η μέθοδος οδήγησε στη διαμόρφωση της ιεραρχίας γεωγραφικών εννοιών των σημάτων 3.12. και 3.13.



Σχήμα 3.12 Πραγματικό σενάριο εφαρμογής με ορισμούς εννοιών που περιγράφουν γεωλογικά μορφώματα μακρόστενου σχήματος



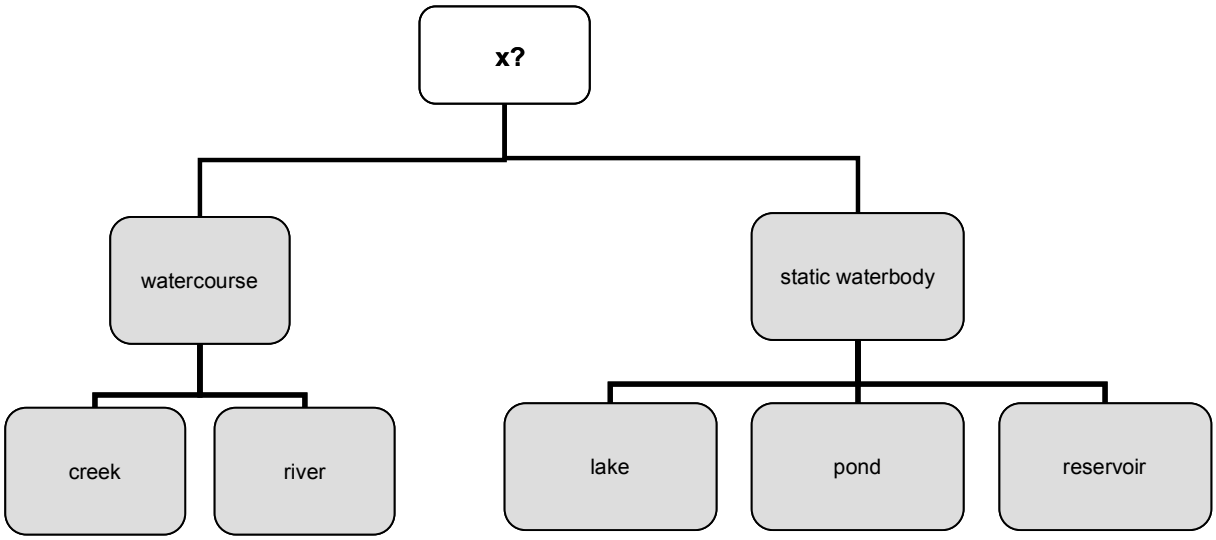
Σχήμα 3.13 Πραγματικό σενάριο εφαρμογής με ορισμούς εννοιών που περιγράφουν περικλειόμενα γεωλογικά μορφώματα

b) Ολοκλήρωση Γεωχωρικών Οντολογιών

Οι κύριες μέθοδοι ολοκλήρωσης οντολογιών είναι η ευθυγράμμιση, η μερική συμβατότητα, η ενοποίηση και η πραγματική ολοκλήρωση. *Ευθυγράμμιση* (alignment) ονομάζεται η αντιστοίχιση μεταξύ των σημασιολογικά όμοιων εννοιών που ανήκουν σε διαφορετικές οντολογίες. Η *μερική συμβατότητα* (partial compatibility) ορίζεται ως η συγχώνευση των σημασιολογικά όμοιων τμημάτων διαφορετικών οντολογιών. Τα συγχωνευθέντα τμήματα αλλάζουν μορφή προκειμένου να εναρμονιστούν με το αποτέλεσμα της ολοκλήρωσης, ενώ τα υπόλοιπα τμήματα παραμένουν ως έχουν. Η *ενοποίηση* (unification), ή *σύντηξη* (fusion), επεκτείνει τη μερική συμβατότητα σε όλη την έκταση των προς ολοκλήρωση οντολογιών. Έτσι οι οντολογίες «εξαναγκάζονται» κατά κάποιο τρόπο να αλλάξουν δομή και να συνθέσουν μία ενιαία οντολογία. Τέλος, η *πραγματική ολοκλήρωση* (true integration) οντολογιών, συνθέτει από ένα σύνολο αρχικών οντολογιών μία ενιαία οντολογία, δημιουργώντας ενίοτε επιπρόσθετες έννοιες.

Κατά τη διαδικασία πραγματικής ολοκλήρωσης, όποτε απαιτηθεί να δημιουργηθεί νέα έννοια για να αποτελέσει υπερ-έννοια (super-concept) μιας ομάδας αμφιθαλών εννοιών, η Geo-Labeling μπορεί να συμβάλλει προτείνοντας για την έννοια αυτή μια κατάλληλη περιγραφή.

Για παράδειγμα, ας θεωρήσουμε ότι από τα αποτελέσματα της πραγματικής ολοκλήρωσης δύο γεωχωρικών οντολογιών, προέκυψε η ανάγκη ένωσης δύο τμημάτων τους κάτω από μια γενική και καινούργια έννοια, έστω *X*. Η ερώτηση που τίθεται είναι «*Πως θα μπορούσε να ονομαστεί η έννοια X;*». Το σχήμα 3.14 παρουσιάζει γραφικά τη σχετική κατάσταση ώστε να αναδειχθεί με ποιον τρόπο η Geo-Labeling θα μπορούσε να προτείνει μια απάντηση στο προηγούμενο ερώτημα.



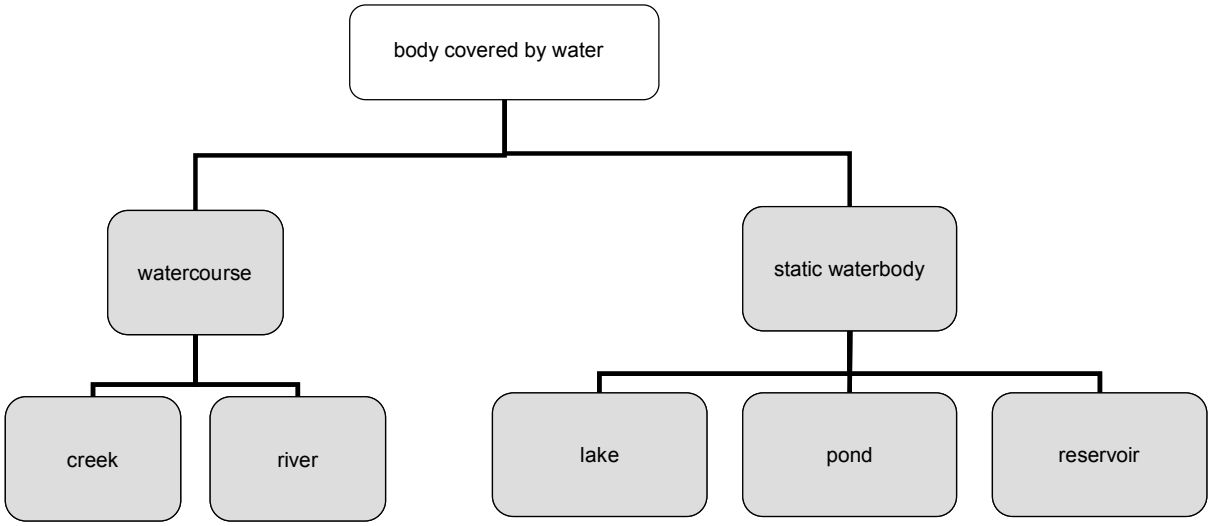
Σχήμα 3.14 Παράδειγμα ολοκλήρωσης τμημάτων γεωχωρικών οντολογιών «κάτω» από την ίδια έννοια

Η αναζήτηση κατάλληλου ονόματος για τη *X*, συνίσταται ουσιαστικά στην εύρεση επιγραφής για την ομάδα που αποτελείται από τις γενικότερες έννοιες των τμημάτων προς ένωση. Στο παράδειγμα του σχήματος 3.14, η ομάδα αυτή αποτελείται από τις έννοιες *Watercourse* και *Static waterbody*. Στους ορισμούς των εννοιών αυτών, αναγνωρίζονται τα σημασιολογικά στοιχεία τα οποία συγκεντρώνονται στον Πίνακα 3.4.

Concept	Extracted semantic elements
Watercourse	IS-A: body
	COVER: water
	NATURE: natural
	FLOWS: flows
Static waterbody	IS-A: body
	COVER: water
	NATURE: natural or artificial
	FLOWS: not flowing

Πίνακας 3.4 Τα σημασιολογικά στοιχεία των γεωγραφικών εννοιών *Watercourse* και *Static waterbody*

Οι έννοιες *Watercourse* και *Static waterbody* μοιράζονται το ίδιο γένος *Body* ενώ διαθέτουν και οι δύο το σημασιολογικό στοιχείο <COVER> με τιμή *water*. Σύμφωνα με τη διαδικασία σχηματισμού των επιγραφών που ακολουθεί η Geo-Labeling, κατάλληλο όνομα για την έννοια *X* αποτελεί το *Body covered by water*, το οποίο φαίνεται να ταιριάζει ικανοποιητικά στις έννοιες *Watercourse* και *Static waterbody* (Σχήμα 3.15).



Σχήμα 3.15 Απόδοση του ονόματος *body covered by water* στην έννοια *X*

c) Αναζήτηση Πληροφοριών

Το σενάριο αυτό σχετίζεται με ένα πρόβλημα που απαντάται στις εφαρμογές αναζήτησης πληροφοριών. Πρόκειται για το πρόβλημα της κατηγοριοποίησης των αποτελεσμάτων της αναζήτησης σε ομάδες πληροφοριών σημασιολογικά όμοιου περιεχομένου, ώστε οι χρήστες των εφαρμογών αυτών να βοηθηθούν στην αναγνώριση των πληροφοριών ενδιαφέροντος.

Για παράδειγμα, χρησιμοποιώντας το γλωσσάριο Glossary of Generic Terms[6] που παρέχεται από την Επιτροπή Committee for Geographical Names in Australia, ως πηγή ορισμών γεωγραφικών εννοιών, έστω ότι ανακτούνται όλοι οι ορισμοί που σχετίζονται με την έννοια *Sea*, εκτελώντας μια απλή αναζήτηση κειμένου βάσει της λέξης-κλειδί *sea*. Από την αναζήτηση, προκύπτει ότι 25 έννοιες περιέχουν τη λέξη sea στον ορισμό τους. Τα αναλυτικά αποτελέσματα συγκεντρώνονται στον Πίνακα 3.5.

[6] Διαθέσιμο στο σύνδεσμο http://www.icsm.gov.au/cgna/glossary_pnames.pdf, τελευταία προσπέλαση Ιούνιος 2010.

Glossary term	Definition
ARM	A narrow portion of the **sea** projecting from the mainland.
BASIN	The tract of country drained by a river and its tributaries, or which drains into a particular lake or **sea**.
BAY	An open, curving indentation made by the **sea** or a lake into a coastline
COAST, COASTLINE	The edge or margin of land next to the **sea**.
DEPRESSION	Any hollow or relatively sunken area, on land or in the **sea**.
GULF	Part of the **sea**, extending into the land; usually larger than a bay.
INLET	A basin at the lower reaches of a river, connected to the **sea** by a narrow opening and subject to tidal movements.
ISLAND	A piece of land surrounded by water, in an ocean, **sea**, lake or river.
LAGOON	An enclosed area of shallow salt or brackish water which is partly or completely separated from the **sea** by a narrow strip of land or sand banks (dunes).
LOCH	A lake or arm of the **sea**.
LOUGH	An Irish term for lake or arm of the **sea**.
POOL	A large partly enclosed arm of a **sea** or lake.
PROMONTORY	A rocky coastal headland projecting significantly into the **sea**.
PRONG	A pointed elongated arm of land protruding into the **sea**.
REACH	An arm of the **sea** or a lake extending into the land.
REEF	A ridge of rocks or coral lying near the surface of the **sea**, which may be visible at low tide, but is usually covered by water.
RIVER	A stream of fresh water which, part of the year, is larger than a brook or creek and flows by natural channel, being confined within banks, into the **sea** or a lake, or another river.
SALT MARSH	A marsh which at times is flooded by the **sea**, or an inland marsh in an arid region in which the water contains a high proportion of salt.
SANDBANK	A bank, shoal or submerged ridge of sand especially in the **sea**, or a river often exposed at low tide.
SANDBAR	A bar of sand formed in a **sea** or river by the action of the tides or currents.
SEA	One of the smaller divisions of the oceans, especially if partly enclosed by land.
SHOAL	A ridge of sand or of rocks just below the surface of the **sea** or of a river and therefore dangerous to navigation.
SOUND	A relatively long arm of the sea, forming a channel between an island and the mainland, or connecting two larger bodies of water, as a **sea** and the ocean, or two parts of the same body, but usually wider and more extensive than a strait.
TABLELAND	A plateau bounded by steep cliff-like faces which lead abruptly down to the **sea** or the adjoining lowlands.
TERRACE	A nearby level strip of land extending along the edge of a **sea**, river or lake, or on the sides of a hill or valley. It is bounded above and below by rather abrupt slopes.

Πίνακας 3.5 Αποτελέσματα αναζήτησης ορισμών γεωγραφικών εννοιών που περιέχουν τη λέξη *sea*

Στη συνέχεια, όπως ορίζει η Geo-Labeling, αναγνωρίζονται τα σημασιολογικά στοιχεία των ορισμών, γίνεται αποσαφήνιση και εμπλουτισμός (συσχέτιση με συνώνυμα και υπερώνυμα) του ονόματος και του γένους τους και τέλος, αναλύονται τα σημασιολογικά στοιχεία σε συγκρίσιμα σημασιολογικά μόρια πληροφοριών. Ακολούθως, προσδιορίζεται το μεγαλύτερο τμήμα σημασιολογικής πληροφορίας που μοιράζονται οι έννοιες αναμεταξύ τους με σκοπό την εύρεση ομάδων σημασιολογικά όμοιων εννοιών. Ως αποτέλεσμα, σχηματίζονται οι 5 διαφορετικές ομάδες εννοιών που περιγράφονται παρακάτω:

- Μια ομάδα 7 εννοιών, οι οποίες διαθέτουν όλες το σημασιολογικό στοιχείο <PART-OF> και το σημασιολογικό μόριο *sea*. Η επιγραφή που δίνεται στην ομάδα αυτή από τη Geo-Labeling είναι *Entity, part of sea* (Πίνακας 3.6).

Labels	Glossary term	Semantic elements derived from GeoNLP
ENTITY PART OF THE SEA	LOCH	IS-A: lake or arm **PART-OF: sea**
	LOUGH	IS-A: lake or arm **PART-OF: sea**
	POOL	IS-A: arm **PART-OF: sea** or lake SIZE: large SHAPE: partly enclosed
	REACH	IS-A: arm **PART-OF: sea** or lake EXTENDS_TO: extending into the land.
	SOUND	IS-A: arm **PART-OF: sea** SHAPE: forming a channel between an island and the mainland, CONNECTS: connecting two larger bodies of water, as a sea and the ocean, or two parts of the same body, but usually wider and more extensive than a strait.
	ARM	IS-A: portion **PART-OF: sea** SHAPE: narrow EXTENDS_TO: projecting from the mainland.
	GULF	**PART-OF: sea** SIZE: usually larger than a bay EXTENDS_TO: extending into the land.

Πίνακας 3.6 Ομαδοποίηση των αποτελεσμάτων της αναζήτησης γεωγραφικών εννοιών που σχετίζονται με τη θάλασσα, βάσει του σημασιολογικού στοιχείου <PART-OF>

- Μια δεύτερη ομάδα 7 εννοιών, οι οποίες διαθέτουν όλες το σημασιολογικό στοιχείο <LOCATION> και το σημασιολογικό μόριο *sea*. Η επιγραφή που δίνεται στην ομάδα αυτή από τη Geo-Labeling είναι *Object, physical object, Located in relation to sea* (Πίνακας 3.7).

Labels	Glossary term	Semantic elements derived from GeoNLP
OBJECT, PHYSICAL OBJECT LOCATED IN RELATION TO THE SEA	DEPRESSION	IS-A: area SHAPE: hollow or relatively sunken **LOCATION:** on land or **in** the **sea**.
	ISLAND	IS-A: piece COVER: land SURROUNDED_BY: surrounded by water **LOCATION: in** an ocean, **sea**, lake or river.
	SANDBANK	IS-A: bank, ridge COVER: shoal or submerged, sand, exposed at low tide. **LOCATION:** especially **in** the **sea** or a river.
	SANDBAR	IS-A: bar COVER: sand NATURE: formed by the action of the tides or currents. **LOCATION: in** a **sea** or river.
	COAST, COASTLINE	IS-A: edge, margin COVER: land **LOCATION:** next to the **sea**.
	REEF	IS-A: ridge COVER: of rock or coral, but is usually covered by water. VISIBILITY: which may be visible at low tide **LOCATION:** near the surface of the **sea**.
	SHOAL	IS-A: ridge COVER: of sand or of rocks **LOCATION:** just below the surface of the **sea** or of a river. PURPOSE: therefore dangerous to navigation.

Πίνακας 3.7 Ομαδοποίηση των αποτελεσμάτων της αναζήτησης γεωγραφικών εννοιών που σχετίζονται με τη θάλασσα, βάσει του σημασιολογικού στοιχείου <LOCATION>

- Μια τρίτη ομάδα 4 εννοιών, οι οποίες διαθέτουν όλες τα σημασιολογικά στοιχεία <CONNECTS> ή <CONNECTED_TO> και το σημασιολογικό μόριο *sea*. Η επιγραφή που δίνεται στην ομάδα αυτή από τη Geo-Labeling είναι *Entity connects/connected to sea* (Πίνακας 3.8).

Labels	Glossary term	Semantic elements derived from GeoNLP
ENTITY CONNECTS/CONNECTED TO THE SEA	BASIN	IS-A: tract PART-OF: country CONNECTED_TO: drained by a river and its tributaries, **CONNECTS:** drains into a particular lake or **sea**.
	INLET	IS-A: basin LOCATION: at the lower reaches of a river **CONNECTED_TO:** connected to the **sea** by a narrow opening MOVEMENT: subject to tidal movements.
	RIVER	IS-A: stream COVER: fresh water SIZE: part of the year, is larger than a brook or creek FLOWS: flows NATURE: by natural channel **CONNECTED_TO:** being confined within banks, into the **sea** or a lake, or another river.
	SOUND	IS-A: arm PART-OF: sea SHAPE: forming a channel between an island and the mainland, **CONNECTS:** connecting two larger bodies of water, as a **sea** and the ocean, or two parts of the same body, but usually wider and more extensive than a strait.
	TABLELAND	IS-A: plateau SURROUNDED_BY: bounded by steep cliff-like faces **CONNECTED_TO:** lead abruptly down to the **sea** or the adjoining lowlands.

Πίνακας 3.8 Ομαδοποίηση των αποτελεσμάτων της αναζήτησης γεωγραφικών εννοιών που σχετίζονται με τη θάλασσα, βάσει του σημασιολογικών στοιχείων <CONNECTS> και <CONNECTED_TO>

- Μια τέταρτη ομάδα 3 εννοιών, οι οποίες διαθέτουν όλες το σημασιολογικό στοιχείο <EXTENDS_TO> και το σημασιολογικό μόριο *sea*. Η επιγραφή που δίνεται στην ομάδα αυτή από τη Geo-Labeling είναι *Object, physical object extends to sea* (Πίνακας 3.9)

Labels	Glossary term	Semantic elements derived from GeoNLP
OBJECT, PHYSICAL OBJECT EXTENDS TO THE SEA	PROMONTORY	IS-A: headland COVER: rocky LOCATION: coastal **EXTENDS_TO:** projecting significantly into the **sea**.
	PRONG	IS-A: arm COVER: land SHAPE: pointed elongated **EXTENDS_TO:** protruding into the **sea**.
	TERRACE	IS-A: strip PART-OF: land LOCATION: nearby level **EXTENDS_TO:** extending along the edge of a **sea**, river or lake, or on the sides of a hill or valley. SURROUNDED_BY: bounded above and below by rather abrupt slopes.

Πίνακας 3.9 Ομαδοποίηση των αποτελεσμάτων της αναζήτησης γεωγραφικών εννοιών που σχετίζονται με τη θάλασσα, βάσει του σημασιολογικού στοιχείου <EXTENDS_TO>

- Μια πέμπτη ομάδα 2 εννοιών, οι οποίες διαθέτουν όλες το σημασιολογικό στοιχείο <COVER> και το σημασιολογικό μόριο *salt water*. Η επιγραφή που δίνεται στην ομάδα αυτή από τη Geo-Labeling είναι *Entity, covered by salt water* (Πίνακας 3.10)

Labels	Glossary term	Semantic elements derived from GeoNLP
ENTITY COVERED BY SALT WATER	LAGOON	IS-A: area SHAPE: enclosed COVER: of shallow **salt** or brackish **water** SEPARATED_FROM: separated from the sea by a narrow strip of land or sand banks (dunes).
	SALT MARSH	IS-A: marsh COVER: at times is flooded by the sea, the **water contains** a high proportion of **salt**. LOCATION: inland, in an arid region.

Πίνακας 3.10 Ομαδοποίηση των αποτελεσμάτων της αναζήτησης γεωγραφικών εννοιών που σχετίζονται με τη θάλασσα, βάσει του σημασιολογικού στοιχείου <COVER>

- Οι τελευταίες 2 από τις 24 έννοιες, οι οποίες παραμένουν χωρίς ομάδα, είναι οι *Sea* και *Bay*. Αυτό συμβαίνει διότι, βάσει της μεθόδου, δεν προκύπτει ότι μοιράζονται κάποιο τμήμα σημασιολογικής πληροφορίας με άλλες. Μπορεί να θεωρηθεί ότι οι έννοιες αυτές συνθέτουν από μόνες τους μια ομάδα στην οποία αποδίδεται ως επιγραφή το γένος του ορισμού τους, δεδομένου ότι το γένος αποτελεί συνήθως γενικότερη έννοια από την οριζόμενη. Στην περίπτωση που δεν ορίζεται το γένος κάποιας έννοιας, της αποδίδεται ως επιγραφή το όνομα της πλησιέστερης γενικότερης έννοιας που παρέχεται από τη βάση γνώσης. Τα αποτελέσματα απεικονίζονται στον Πίνακα 3.11.

Labels	Glossary term	Semantic elements derived from GeoNLP
BODY OF WATER, WATER	SEA	PART-OF: oceans SIZE: One of the smaller divisions SURROUNDED_BY: especially if partly enclosed by land.
INDENTATION	BAY	IS-A: indentation SHAPE: open, curving NATURE: made by the sea or a lake LOCATION: into a coastline.

Πίνακας 3.11 Απόδοση επιγραφής στις μονομελείς ομάδες που υπολείπονται

4. Εξόρυξη Γνώσης Βάσει Τεχνικών Χωρικοποίησης

4.1 Εισαγωγή

Την τελευταία εικοσαετία, η ανάγκη διαχείρισης όλο και αυξανόμενου όγκου διαθέσιμων δεδομένων σε ηλεκτρονική μορφή, καθώς και η προσπάθεια κατανόησης πολύπλοκων διαδικασιών και φαινομένων, οδήγησαν στην αναζήτηση λύσεων για τη διευκόλυνση της ανθρώπινης αντίληψης. Η οπτικοποίηση αποτελεί την ιδανικότερη των λύσεων εάν λάβουμε υπόψη μας ότι «*η οπτική πληροφορία επεξεργάζεται και αφομοιώνεται από το ανθρώπινο μυαλό πολύ πιο αποτελεσματικά από αυτήν που παρουσιάζεται σε αλφαριθμητική ή άλλη μορφή*» [Tuft83]. Με τον όρο *οπτικοποίηση*, εννοούμε την εφαρμογή τεχνικών οπτικής αναπαράστασης πληροφορίας, με στόχο την καλύτερη κατανόηση της.

Κατά τη διάρκεια της ίδιας περιόδου, στο χώρο της πληροφορικής, οι τεχνικές οπτικοποίησης αναπτύχθηκαν ραγδαίως ακολουθώντας δυο γενικές κατευθύνσεις [CMS99a]: την *επιστημονική οπτικοποίηση* (scientific visualization), που αφορά κυρίως σε δεδομένα του φυσικού κόσμου, όπως το ανθρώπινο σώμα, τη γη, τη μοριακή δομή της ύλης, κλπ, καθώς και την *οπτικοποίηση πληροφοριών* (information visualization), που εστιάζεται σε αφηρημένες έννοιες, χωρίς φυσική υπόσταση, όπως κείμενα, στατιστικά δεδομένα, κλπ.

4.1.1. Ορισμοί

Εξαιρετικά ενδιαφέρουσα είναι η περίπτωση κατά την οποία, αφηρημένα δεδομένα, απεικονίζονται, είτε στο φυσικό χώρο με τη βοήθεια χωρικής μεταφοράς (spatial metaphor), είτε σε γεωμετρικούς χώρους με τη βοήθεια χωρικών διατάξεων. Η οπτικοποίηση αυτή ονομάζεται *χωρικοποίηση* (spatialization) και έχει ως αποτέλεσμα τον ορισμό ενός πρωτότυπου χώρου, όπου οι βασικές χωρικές έννοιες αποκτούν καινούργια σημασία.

Η χωρικοποίηση ορίζει μία απεικόνιση από τον πολυδιάστατο χώρο των αφηρημένων πληροφοριών στο χώρο περιορισμένων διαστάσεων που αντιλαμβάνεται ο άνθρωπος. Δεν

πρόκειται περί απλής χωρικής συσχέτισης, όπως συμβαίνει στους θεματικούς χάρτες, αλλά περί προσομοίωσης μη χωρικού φαινομένου με χωρικό.

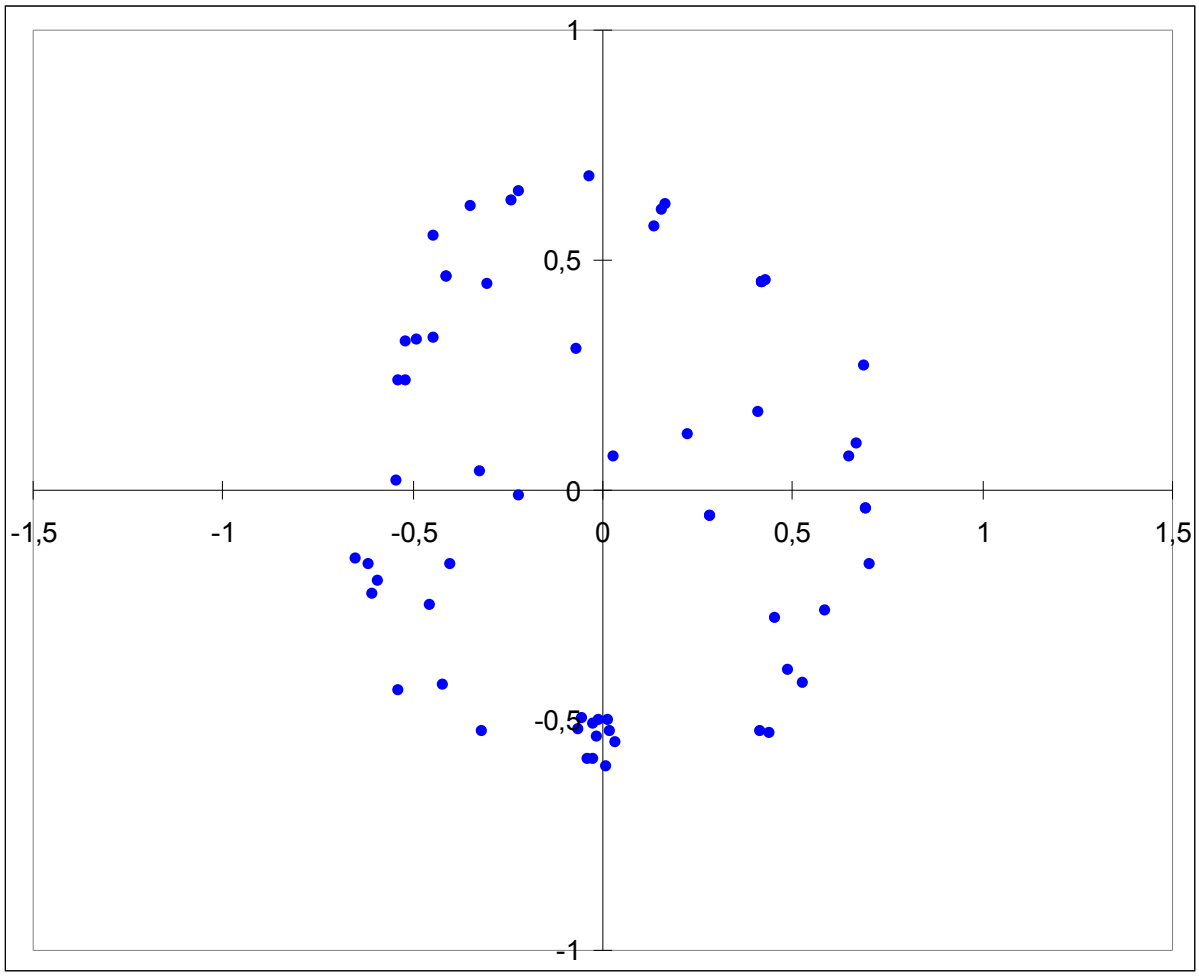
Σύμφωνα με τους Kuhn και Blumenthal [KB96], η χωρικοποίηση ορίζεται ως «*η διαδικασία κατά την οποία, αφηρημένοι χώροι πληροφορίας απεικονίζονται στο φυσικό χώρο βάσει χωρικών μεταφορών*». Η χωρική μεταφορά ορίζει πάντα μια μερική απεικόνιση από το χώρο πληροφορίας στο φυσικό χώρο. Δεν θεωρεί σε καμία περίπτωση ότι οι δυο χώροι ταυτίζονται. Για παράδειγμα, όλοι γνωρίζουμε τη χωρική μεταφορά σύμφωνα με την οποία το περιβάλλον εργασίας ενός Η/Υ παρομοιάζεται με το γραφείο (desktop), δεν συνεπάγεται όμως ότι το περιβάλλον εργασίας του Η/Υ θα αρχίζει να σκονίζεται…[KB96].

Οι Skupin και Buttenfield [SB97], γενικεύουν τον προηγούμενο ορισμό προτείνοντας ένα δεύτερο, ο οποίος δεν περιορίζει τη χωρικοποίηση σε απεικόνιση από το χώρο πληροφορίας στο φυσικό χώρο, ούτε θεωρεί αναγκαία και απαραίτητη τη χρήση χωρικής μεταφοράς. Τονίζει απλώς ότι το ζητούμενο είναι να απεικονιστεί η πολυδιάστατη πληροφορία, σε χώρο περιορισμένων διαστάσεων, που θα μπορούσε ενδεχομένως να είναι και ο φυσικός.

Οι Κάβουρας και Κόκλα [ΚΚ08] περιγράφουν τη χωρικοποίηση ως τεχνική αναπαράστασης της ομοιότητας κατά την οποία οι όμοιες πληροφορίες απεικονίζονται σ’ ένα χώρο αναπαράστασης ως αντικείμενα με παρόμοια μορφή, παρόμοιες χωρικές ιδιότητες και εγγείς τοποθεσίες.

Τέλος, ο Goodchild [Good08] ορίζει τη χωρικοποίηση ως «*οργάνωση αντικειμένων στο χώρο που σχετίζεται με τις αναμεταξύ τους ομοιότητες*».

Στο σχήμα 4.1 που ακολουθεί, απεικονίζεται το παράδειγμα της χωρικοποίησης πολυδιάστατων πληροφοριών που πραγματοποιήθηκε με την τεχνική της πολυ-ανυσματικής κλιμάκωσης. Παρατηρούμε ότι δεν έχει χρησιμοποιηθεί συγκεκριμένη χωρική μεταφορά, αλλά προβλήθηκαν και αναπαράστησαν πολυδιάστατες πληροφορίες σε χώρο δυο διαστάσεων.

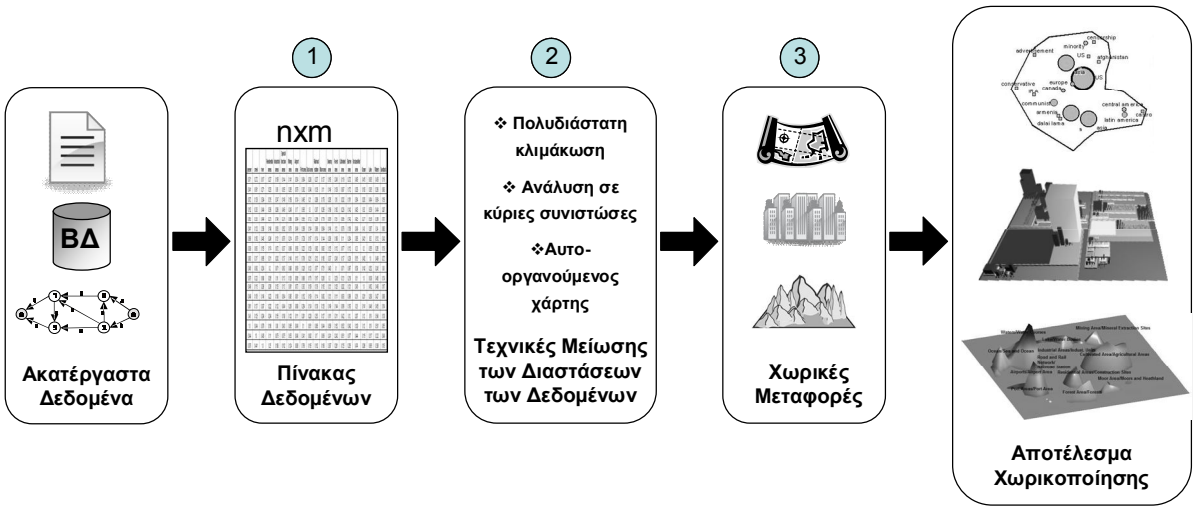


Σχήμα 4.1 Παράδειγμα χωρικοποίησης με την τεχνική της πολυ-ανυσματικής κλιμάκωσης

4.1.2. Ο Σχεδιασμός μιας Χωρικοποίησης

Δεδομένης της πολυμορφίας της δομής των αφηρημένων πληροφοριών προς χωρικοποίηση, κρίνεται απαραίτητο για τον επιτυχή σχεδιασμό μιας χωρικοποίησης, να ακολουθηθούν τα εξής βήματα (Σχήμα 4.2):

1. Προετοιμασία των πληροφοριών προς χωρικοποίηση και μετατροπή τους σε δομημένη και «επεξεργάσιμη» μορφή.
2. Επιλογή και εφαρμογή τεχνικής μείωσης των διαστάσεων των πληροφοριών.
3. Επιλογή κατάλληλων χωρικών μεταφορών για την προβολή και απεικόνιση των πληροφοριών στο χώρο αναπαράστασης.



Σχήμα 4.2 Τα βήματα για το σχεδιασμό μιας χωρικοποίησης

Το πρώτο βήμα κατά το σχεδιασμό μιας χωρικοποίησης, η *προετοιμασία* των πληροφοριών, αποτελεί απαραίτητη προϋπόθεση για τον αποτελεσματικό χειρισμό τους και τη μετέπειτα προβολή τους στο χώρο αναπαράστασης. Ειδικότερα, οι αφηρημένες πληροφορίες, οι οποίες δεν έχουν φυσική υπόσταση και συνεπώς στερούνται χωρικών χαρακτηριστικών, πρέπει να εξετάζονται και να προετοιμάζονται με ιδιαίτερη προσοχή προκειμένου να χωρικοποιηθούν.

Οι «ακατέργαστες» πληροφορίες, δηλαδή τα *δεδομένα* που πρόκειται να χωρικοποιηθούν, είναι ποικίλης μορφής. Αποτελούνται από κείμενα, στατιστικά δεδομένα, πίνακες βάσεων δεδομένων, σελίδες διαδικτύου, δεδομένα δικτύων, κλπ. Η πολυμορφία αυτή επιβάλλει τη μετατροπή των δεδομένων σε μια πιο δομημένη μορφή, με σκοπό την αλγοριθμική επεξεργασία τους. Για τον σκοπό αυτό, δηλαδή την περιγραφή των δεδομένων με δομημένο τρόπο, χρησιμοποιείται ο *πίνακας δεδομένων* [CMS99a]. Οι στήλες του πίνακα δεδομένων αντιστοιχούν στα χαρακτηριστικά των δεδομένων, ή *μεταβλητές* (variables). Από το σύνολο των τιμών μιας στήλης, προσδιορίζεται το εύρος μεταβολής της μεταβλητής. Οι γραμμές του πίνακα δεδομένων, αντιστοιχούν στους δυνατούς συνδυασμούς, ή *περιπτώσεις* (cases) δεδομένων.

Ο ορισμός πινάκα δεδομένων παρουσιάζει πολλά πλεονεκτήματα. Ένα από τα πλεονεκτήματα αυτά είναι ότι, από τις διαστάσεις του, γίνονται άμεσα αντιληπτές οι διαστάσεις των δεδομένων και το πλήθος τους. Π.χ., εάν *Π*, είναι ένας πίνακας δεδομένων

διαστάσεων *nxm*, τότε ο *Π* εσωκλείει *n* περιπτώσεις δεδομένων *m* διαστάσεων. Είναι προφανές ότι, συνήθως, το *n* είναι πολύ μεγαλύτερο του *m*. Ακόμη, από τις διαστάσεις αυτές, βγαίνουν συμπεράσματα που αφορούν στις διαστάσεις του καταλληλότερου χώρου απεικόνισης. Η τέλεια περίπτωση (από την άποψη ότι δεν χρειάζεται να μειωθούν οι διαστάσεις των δεδομένων και συνεπώς δεν χάνεται κανένα ποσοστό πληροφορίας) είναι αυτή κατά την οποία, οι διαστάσεις των δεδομένων είναι τόσες όσες και οι διαστάσεις του χώρου απεικόνισης. Δυσκολότερη αλλά συνήθης, είναι η περίπτωση όπου οι διαστάσεις των δεδομένων ξεπερνάνε τις τρεις. Έτσι, δεδομένου ότι ο χώρος αναπαράστασης μπορεί να είναι το πολύ μέχρι τριών διαστάσεων, η χωρικοποίηση δεδομένων με περισσότερες από τρεις διαστάσεις απαιτεί ορισμένα τεχνάσματα. Επιπροσθέτως, με την παρατήρηση του πίνακα δεδομένων, μπορούμε να βγάλουμε κάποια πρόχειρα συμπεράσματα για τη σημασία ορισμένων μεταβλητών. Για παράδειγμα, αντιλαμβανόμαστε ποιες μεταβλητές χαρακτηρίζουν μοναδικά τα δεδομένα, ποιες περιττεύουν και υπερφορτώνουν αδίκως το σύνολο των χαρακτηριστικών, ποιες δεν είναι ανεξάρτητες αλλά προκύπτουν από άλλες, κλπ.

Σημειώνεται τέλος, ότι ο πίνακας δεδομένων μπορεί να οριστεί και για δεδομένα που περιγράφουν τη δομή δικτύων. Στην περίπτωση αυτή, οι κόμβοι του δικτύου μοντελοποιούνται ως περιπτώσεις στον πίνακα δεδομένων ενώ χρειάζεται να συμπεριληφθεί και μια μεταβλητή που να περιγράφει τις διακλαδώσεις μεταξύ των κόμβων. Οι ιεραρχίες αποτελούν ειδική περίπτωση δεδομένων δικτύου, όπου ορίζεται ένας αρχικός κόμβος, η «ρίζα», ενώ οι διακλαδώσεις μεταξύ κόμβων δεν δημιουργούν κυκλικές διαδρομές.

Για έναν πίνακα δεδομένων *Π*, το στοιχείο *Π(i,j)* αντιστοιχεί στην τιμή του χαρακτηριστικού *j* της περίπτωσης *i* των δεδομένων. Ποια είναι όμως η φύση των στοιχείων του πίνακα δεδομένων; Είναι αριθμοί; Λέξεις; Κάτι άλλο;

Η φύση των στοιχείων του πίνακα δεδομένων καθορίζεται από το είδος των μεταβλητών. Οι μεταβλητές κατηγοριοποιούνται σε τρία βασικά είδη: το *ονομαστικό* (nominal), το *ταξινομημένο* (ordinal) και το *ποσοτικό* (quantitative) είδος.

Οι τιμές μιας μεταβλητής ονομαστικού είδους απαρτίζουν ένα μη ταξινομημένο σύνολο. Οι δυνατές πράξεις μεταξύ των στοιχείων του συνόλου αυτού είναι οι: '=' («*ίδιο με*», μη αριθμητικό ίσον) και '≠' («*διαφορετικό από*»). Για παράδειγμα, το σύνολο {'*Όσα παίρνει ο*

άνεμος', '*Η Αλίκη στο ναυτικό*', '*Καζαμπλάνκα*'} αναφέρεται στις τιμές της μεταβλητής '*Τίτλος_έργου*', ονομαστικού είδους.

Μια μεταβλητή ταξινομημένου είδους ορίζει ένα ταξινομημένο σύνολο. Οι δυνατές πράξεις μεταξύ των στοιχείων του συνόλου αυτού είναι οι πράξεις συσχέτισης: '>', '<', '=' («*ίδιο με*», μη αριθμητικό ίσον), («*διαφορετικό από*») κλπ. Για παράδειγμα, το σύνολο <*Νέος*, *Μεσήλικας*, *Ηλικιωμένος*>, που αναφέρεται στο εύρος των τιμών της μεταβλητής '*Ηλικία*', είναι ταξινομημένου είδους.

Τέλος, στο ποσοτικό είδος ανήκουν οι μεταβλητές των οποίων οι τιμές ορίζουν ένα αριθμητικό εύρος. Όλες οι πράξεις συσχέτισης και όλες οι αριθμητικές πράξεις, '+', '-', κλπ, μπορούν να εφαρμοστούν στις τιμές αυτές. Για παράδειγμα, το εύρος [*-1.000.000*, *1.000.000*] αναφέρεται στις τιμές της μεταβλητής '*Τιμή*', που είναι ποσοτικού είδους.

4.2 Χωρικές Μεταφορές

Οι χωρικές μεταφορές προσφέρουν τη δυνατότητα εξερεύνησης χώρων πληροφορίας σχεδόν διαισθητικά. Οι Benking and Judge [BJ94] απαριθμούν έξι κατηγορίες χωρικών μεταφορών:

1. Γεωμετρικές δομές (geometric forms)
2. Τεχνητές δομές (artificial forms)
3. Φυσικές δομές (natural forms)
4. Συστημικές δομές (systemic structures)
5. Συστήματα παραδοσιακών συμβόλων (traditional symbol systems)
6. Δυναμικά συστήματα (dynamic systems)

Οι ίδιοι [BJ94] υποστηρίζουν ότι τα τοπία – όπως αυτά απαντώνται στη φύση – αποτελούν μια άριστη χωρική μεταφορά για τη μοντελοποίηση και την εξερεύνηση πληροφορίας λόγω του ότι μπορούν να μοντελοποιήσουν την πληροφορία σε διάφορα ιεραρχικά επίπεδα ή επίπεδα λεπτομέρειας. Ακόμα, τα τοπία πληροφορίας καθιστούν πιο απτό και πιο οικείο το

χώρο αναπαράστασης της πληροφορίας αφού μοιάζει πολύ με τον πραγματικό γεωγραφικό χώρο [FB01].

Οι Kuhn και Blumenthal [KB96] υποστηρίζουν ότι ο χώρος απεικόνισης πρέπει πάντα να ορίζεται βάσει της ανθρώπινης εμπειρίας και φαντασίας, όπως είναι οι εικονικές πόλεις και τα τοπία πληροφορίας. Έτσι, αρκετά συχνά συναντάται η τεχνική χωρικοποίησης κειμένων που χρησιμοποιεί τη χωρική μεταφορά του ανάγλυφου τοπογραφικού χάρτη. Στην περίπτωση αυτή, πολύπλοκοι στατιστικοί αλγόριθμοι αναλύουν τα κείμενα γραμμένα στη φυσική γλώσσα, και αντιστοιχίζουν στο καθένα ένα διάνυσμα με τις πιο χρησιμοποιούμενες λέξεις, «κωδικοποιώντας» το επικρατέστερο θέμα τους. Τα επικρατέστερα θέματα των κειμένων συγκρίνονται μεταξύ τους και απεικονίζονται ως σημειακές οντότητες μεταξύ των οποίων η απόσταση εξαρτάται από το βαθμό ομοιότητάς τους. Το ανάγλυφο σχηματίζεται συναρτήσει της πυκνότητας των απεικονιζόμενων σημειακών οντοτήτων.

Συνοπτικά, η χωρικοποίηση εγγράφων με χρήση της μεταφοράς του τοπίου, πραγματοποιεί τις εξής προσομοιώσεις:

- Σημειακές οντότητες → Έγγραφα
- Υψόμετρο → Πλήθος εγγράφων με όμοιο θέμα
- Αποστάσεις μεταξύ σημειακών οντοτήτων → Σημασιολογική ομοιότητα μεταξύ θεμάτων εγγράφων
- Κορυφές ανάγλυφου → Μεγάλη πυκνότητα εγγράφων με παρόμοιο θέμα
- Βουνά και λόφοι → Ομάδες εγγράφων με θέμα που υποδεικνύεται από την επιγραφή στην κορυφή του βουνού / λόφου

4.3 Τεχνικές Χωρικής Ομαδοποίησης

Οι τεχνικές *ομαδοποίησης* έχουν ως σκοπό τη δημιουργία ομάδων δεδομένων από ένα αρχικό σύνολο δεδομένων με σκοπό την ανακάλυψη σημαντικών νοημάτων και σχέσεων μέσα από τα δεδομένα, δηλαδή την εξόρυξη γνώσης από τα δεδομένα. Όσον αφορά τις τεχνικές *χωρικής* ομαδοποίησης, θεωρώντας ότι τα δεδομένα προβάλλονται υπό τη μορφή

σημείων σ' ένα χώρο απεικόνισης, γίνεται προσπάθεια ενοποίησης των πλησιέστερων σημείων σε ομάδες, πετυχαίνοντας εμμέσως, τη δημιουργία ομάδων όμοιων δεδομένων. Χαρακτηριστικά παραδείγματα εφαρμογών που χρησιμοποιούν τεχνικές χωρικής ομαδοποίησης αποτελούν: η χωρική εξόρυξη γνώσης [KAH96], [MH01], [THL03], η χωρική ανάλυση δεδομένων [ME98], [Open98] και η εξόρυξη γνώσης από χωρικές βάσεις δεδομένων ή άλλες συλλογές χωρικών δεδομένων [EKSX96], [SEKX98], [Kola01], [EVJW06]. Πιο πρόσφατα, η χωρική ομαδοποίηση βρήκε εφαρμογή στη δημιουργία θεματικών χαρτών [JEC07].

Στις τεχνικές χωρικής ομαδοποίησης στηρίζονται πολλές εφαρμογές χωρικοποίησης, γι αυτό και κρίνεται σκόπιμο στο σημείο αυτό να γίνει μια σύντομη επισκόπηση των τεχνικών αυτών.

Καταρχάς, ως *ομάδα* ορίζουμε το σύνολο των σημείων του χώρου απεικόνισης μεταξύ των οποίων η απόσταση είναι πολύ μικρή σε σχέση με την απόσταση που έχουν από τα σημεία των υπόλοιπων ομάδων του ίδιου χώρου. Επειδή στη χωρικοποίηση η χωρική απόσταση είναι αντιστρόφως ανάλογη με την ομοιότητα, μπορούμε να διατυπώσουμε την προηγούμενη πρόταση ως εξής: *ως ομάδα ορίζουμε το σύνολο των δεδομένων μεταξύ των οποίων η ομοιότητα είναι πολύ μεγάλη σε σχέση με την ομοιότητα που έχουν με τα δεδομένα των υπόλοιπων ομάδων*.

Οι κυριότερες κατηγορίες τεχνικών χωρικής ομαδοποίησης περιγράφονται παρακάτω.

4.3.1. Διαιρετικές Τεχνικές Ομαδοποίησης (Partitioning Clustering Methods)

Οι διαιρετικές τεχνικές ομαδοποίησης επιδιώκουν να χωρίσουν ένα σύνολο σημείων του χώρου σε *k* ομάδες με τρόπο ώστε να ικανοποιείται κάποιο κριτήριο. Ενδεικτικά, η γνωστή τεχνική διαιρετικής ομαδοποίησης *k-means*, προσπαθεί να συμπεριλάβει σημεία σε ομάδες με κριτήριο η μέση τιμή των αποστάσεων των σημείων από το κέντρο βάρους της ομάδας να ελαχιστοποιείται.

Τα βήματα που ακολουθούνται κατά την εκτέλεση ενός διαιρετικού αλγορίθμου, είναι τα εξής:

1. Επιλέγονται ο αριθμός και τα κέντρα βάρους των ομάδων.
2. Υπολογίζονται οι αποστάσεις των σημείων από τα επιλεγμένα κέντρα βάρους.
3. Αποδίδεται κάθε σημείο στην πλησιέστερη ομάδα.
4. Επαναπροσδιορίζονται τα κέντρα βάρους των ομάδων, βάσει των καινούργιων μελών.
5. Συγκρίνεται η μεταβολή της θέσης των κέντρων βάρους με κάποια προκαθορισμένη τιμή κατωφλίου. Εάν βρεθεί μεγαλύτερη, βγαίνει το συμπέρασμα ότι δεν έχουν φτάσει τα κέντρα βάρους στη βέλτιστη θέση και συνεπώς ο αλγόριθμος εκτελείται πάλι από το βήμα 2. Εάν βρεθεί μικρότερη, ο αλγόριθμος σταματά.

Τα μειονεκτήματα των τεχνικών αυτών είναι αρκετά:

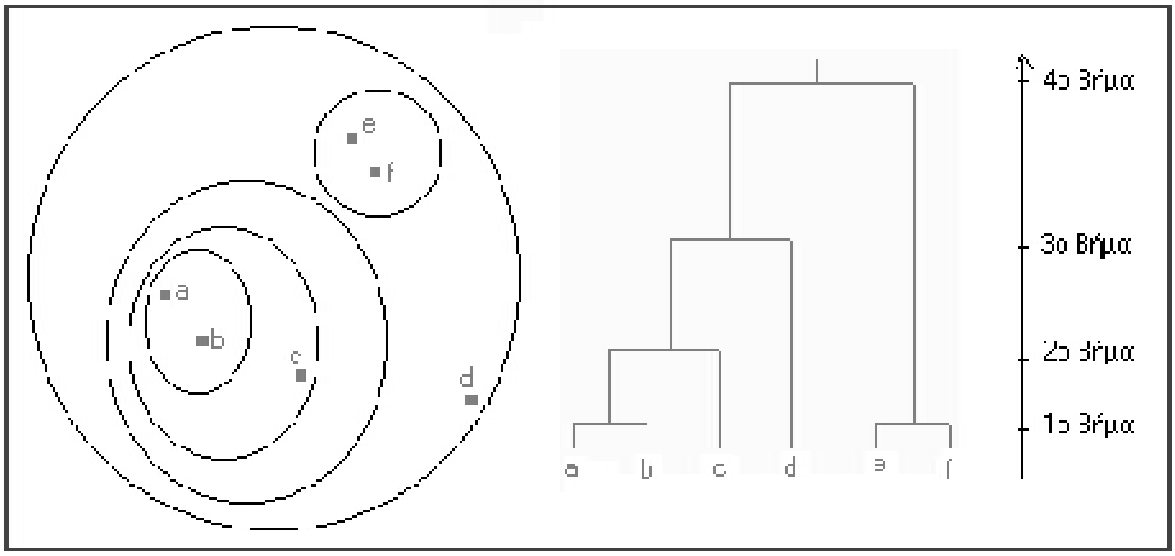
- Πρέπει να είναι εκ των προτέρων γνωστός, ο αριθμός των ομάδων.
- Το τελικό αποτέλεσμα εξαρτάται από την αρχική τοποθέτηση των κέντρων βάρους των ομάδων στο χώρο.
- Ορισμένα σημεία αποδίδονται λανθασμένα σε κάποιες ομάδες επειδή βρίσκονται πιο κοντά στο κέντρο βάρους τους, παρά στο κέντρο της κανονικής τους ομάδας. Αυτό συμβαίνει κυρίως όταν τα μεγέθη των ομάδων είναι πολύ δυσανάλογα ή όταν υπάρχουν ομάδες με κυρτό σχήμα.

4.3.2. Ιεραρχικές Τεχνικές Ομαδοποίησης (Hierarchical Clustering Methods)

Από τις πλέον γνωστές, οι ιεραρχικές τεχνικές ομαδοποίησης έκαναν την εμφάνιση τους το 1951 [FLP+51]. Από τότε, έχουν δημιουργηθεί αρκετές τεχνικές ιεραρχικής ομαδοποίησης οι οποίες χωρίζονται σε δυο βασικούς τύπους: την *συσσωρευτική* (agglomerative) και την *διαιρετική* (divisive) τεχνική, με δημοφιλέστερη την πρώτη.

Η συσσωρευτική ιεραρχική ομαδοποίηση συνίσταται στη δημιουργία μιας και μοναδικής ομάδας από ένα αρχικό πλήθος ομάδων ίσο με το πλήθος των σημείων. Στη διαιρετική ιεραρχική ομαδοποίηση, γίνεται ακριβώς το αντίθετο: στην αρχή, θεωρείται ότι υπάρχει μια και μοναδική ομάδα, η οποία περιέχει όλα τα σημεία. Σε κάθε βήμα του αλγορίθμου, η

ομάδα αυτή διαιρείται μέχρις ότου δημιουργηθούν τόσες ομάδες, όσες και τα σημεία του χώρου. Τα βήματα της συσσωρευτικής ιεραρχικής ομαδοποίησης απεικονίζονται συνήθως με τη βοήθεια μιας δενδροειδούς δομής, το *δενδρόγραμμα*, όπως φαίνεται στο σχήμα 4.3.

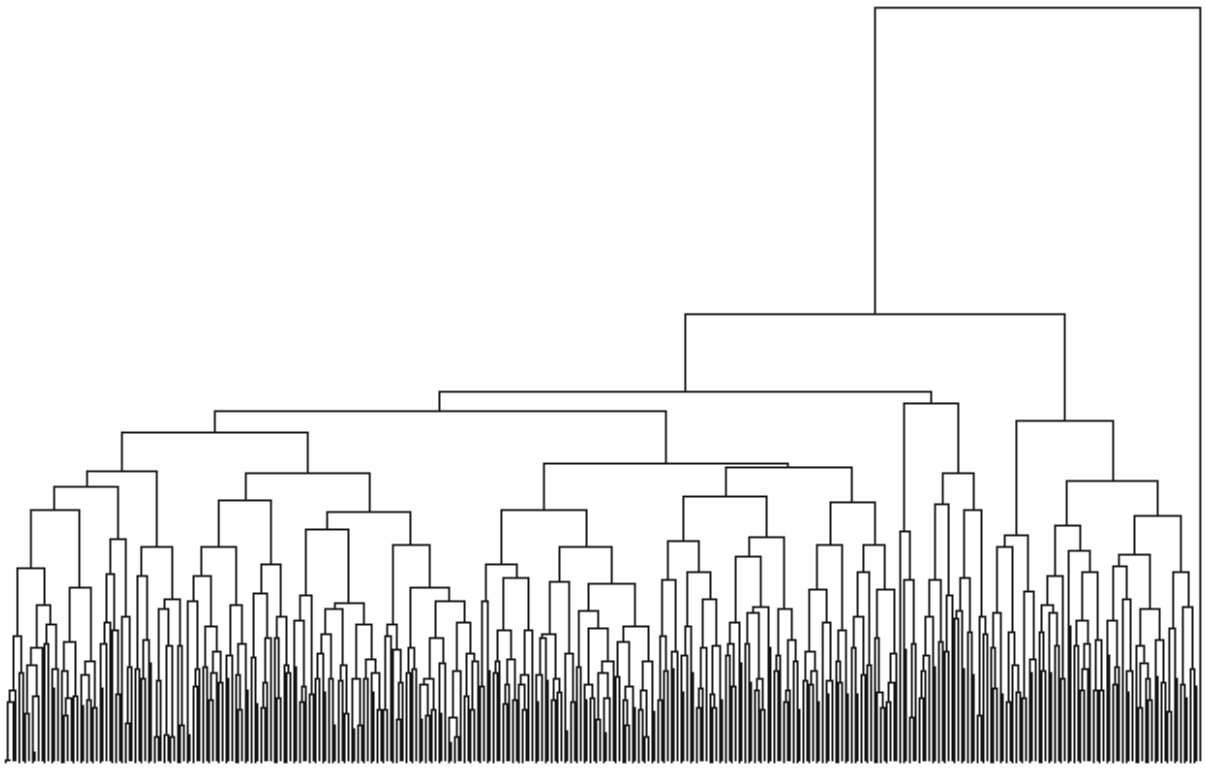


Σχήμα 4.3 Παράδειγμα σχηματισμού δενδρογράμματος

Στο σχήμα 4.3, αναδεικνύονται τα αλγοριθμικά βήματα που ακολουθούνται στην ιεραρχική ομαδοποίηση:

1. Εύρεση των δυο πλησιέστερων σημείων και δημιουργία καινούργιας ομάδας που τους περιέχει.
2. Εύρεση είτε των δυο επόμενων πλησιέστερων σημείων είτε μιας ομάδας και ενός σημείου, που απέχουν τη μικρότερη απόσταση μεταξύ τους, και δημιουργία καινούργιας ομάδας που τους περιέχει. (Ως απόσταση μεταξύ σημείου και ομάδας, μπορεί να ληφθεί υπόψη, η απόσταση του σημείου από το κέντρο βάρους της ομάδας)
3. Εφόσον υπάρχουν περισσότερες από μια ομάδα, επιστροφή στο βήμα 2.

Το δενδρόγραμμα, ενώ επιτρέπει την οπτική αναπαράσταση της εκτέλεσης των βημάτων του αλγορίθμου ομαδοποίησης, δεν δίνει ευκρινή αποτελέσματα στην περίπτωση πολύ μεγάλου πλήθους σημείων (Σχήμα 4.4).



Σχήμα 4.4 Παράδειγμα πολυπληθούς δενδρογράμματος

Μειονέκτημα της ιεραρχικής τεχνικής ομαδοποίησης θεωρείται το γεγονός ότι ένα σημείο δεν δύναται να αλλάξει ομάδα εφόσον καταχωρηθεί σε κάποια.

Η τεχνική αυτή δεν παύει όμως να αποτελεί έναν φυσικό και απλό τρόπο ομαδοποίησης σημείων. Οι επιδόσεις της στον εντοπισμό ομάδων ποικίλων σχημάτων είναι αναμφισβήτητες. Ωστόσο, το πιο σημαντικό πλεονέκτημα της μεθόδου αυτής, είναι ότι δεν απαιτείται να είναι εκ των προτέρων γνωστός ο αριθμός των ομάδων που θα σχηματιστούν. Τέλος, οι ιεραρχικές τεχνικές χρησιμοποιούνται για να βελτιώσουν σημαντικά τα αποτελέσματα των διαιρετικών τεχνικών. Για το λόγο αυτό, αρκετοί επιστήμονες θεωρούν τις τεχνικές αυτές συμπληρωματικές και όχι ανταγωνιστικές.

4.3.3. Τεχνικές Ομαδοποίησης Βάσει Πυκνότητας (Density Based Clustering Methods)

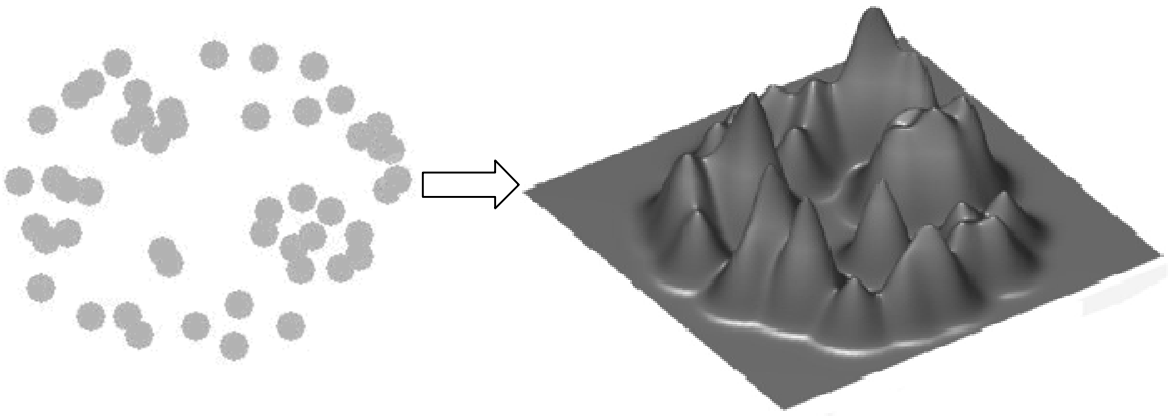
Η βασική ιδέα των τεχνικών ομαδοποίησης βάσει πυκνότητας, είναι η δημιουργία ομάδων με τρόπο που η πυκνότητα των σημείων εντός μιας ομάδας να είναι μεγαλύτερη από εκείνη

εκτός της ομάδας. Αρκετές τεχνικές τέτοιου τύπου έχουν αναπτυχθεί. Ενδεικτικά, παρουσιάζεται στη συνέχεια η τεχνική ομαδοποίησης βάσει πυκνότητας, DENCLUE.

Η τεχνική ομαδοποίησης DENCLUE, αποτελεί γενικευμένη εκδοχή πολλών τεχνικών ομαδοποίησης, όπως παραδείγματος χάριν, των διαιρετικών και των ιεραρχικών τεχνικών. Ο αλγόριθμος που εφαρμόζεται, μοντελοποιεί τη συνολική πυκνότητα των απεικονιζόμενων σημείων, ως άθροισμα συναρτήσεων επιρροής που εφαρμόζονται στα σημεία. Ακολούθως, προσδιορίζονται οι ελκυστές πυκνότητας (density-attractors) της παραγόμενης επιφάνειας, οι οποίοι συμβάλλουν στον εντοπισμό των ομάδων. Οι επιστήμονες [HK98] που εφεύραν την τεχνική ομαδοποίησης DENCLUE, υποστηρίζουν ότι χαρακτηρίζεται από τέσσερα σημαντικά πλεονεκτήματα:

- Διαθέτει ένα γερό μαθηματικό υπόβαθρο.
- Είναι κατάλληλη για τον εντοπισμό ομάδων οποιουδήποτε σχήματος.
- Δίνει καλά αποτελέσματα ακόμη και όταν εφαρμόζεται σε δεδομένα με πολύ θόρυβο.
- Η ταχύτητα εκτέλεσης του σχετικού αλγορίθμου είναι αρκετά μεγάλη.

Ένα παράδειγμα του τρόπου με τον οποίο παρουσιάζονται τα αποτελέσματα της τεχνικής αυτής, φαίνεται στο σχήμα 4.5.



Σχήμα 4.5 Παράδειγμα ομαδοποίησης βάσει πυκνότητας

Η τεχνική DENCLUE βασίζεται στην παραδοχή ότι η επιρροή που ασκεί κάθε απεικονιζόμενο σημείο στη γειτονιά του, μπορεί να μοντελοποιηθεί μαθηματικά. Η

μαθηματική συνάρτηση που χρησιμοποιείται ονομάζεται *συνάρτηση επιρροής* (impact function). Η συνάρτηση επιρροής εφαρμόζεται τοπικά σε κάθε σημείο ενώ το άθροισμα των συναρτήσεων επιρροής αντιστοιχεί στη συνάρτηση επιρροής του συνόλου των σημείων. Τα *τοπικά μέγιστα* (local maxima), ή *ελκυστές πυκνότητας* (density-attractors) συμβάλλουν στον εντοπισμό των ομάδων.

Παραδείγματα συναρτήσεων επιρροής αποτελούν η σταθερή συνάρτηση, f_{square}, και η συνάρτηση Gauss, f_{gauss}, που περιγράφονται στη συνέχεια.

Η *σταθερή συνάρτηση* επιρροής ενός σημείου *y* σ' ένα άλλο σημείο *x* του χώρου, μοντελοποιείται από τον τύπο 11:

$$f_{square}(x,y) = 1,\ d(x,y) \leq \sigma \quad \text{και} \quad f_{square}(x,y) = 0,\ d(x,y) > \sigma \qquad (11)$$

Η επιρροή ενός σημείου *y* σ' ένα άλλο σημείο *x* του χώρου, μπορεί επίσης να μοντελοποιηθεί με τη *συνάρτηση Gauss*, με τύπο 12:

$$f_{gauss}(x,y) = e^{(-d(x,y)^2/2\sigma^2)} \qquad (12)$$

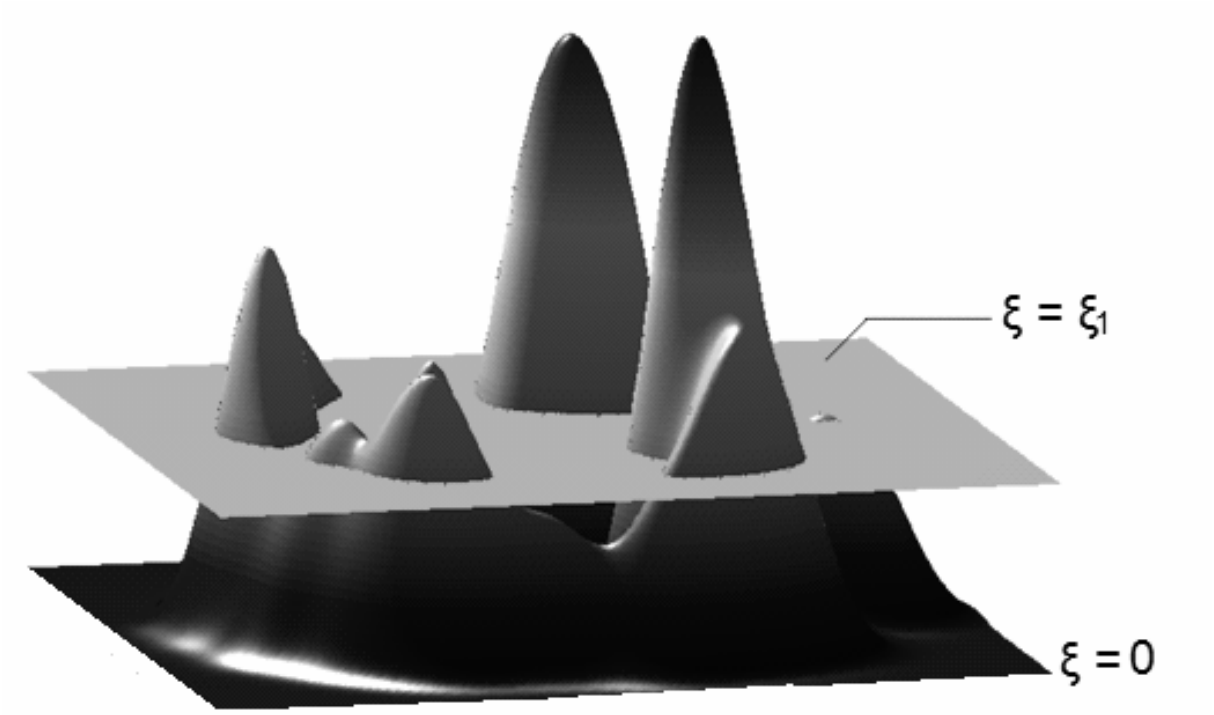
Η πυκνότητα σ' ένα σημείο *x* του χώρου, ορίζεται ως το άθροισμα των επιρροών όλων των σημείων και περιγράφεται από τον τύπο 13:

$$f_D(x) = \sum_{i=1}^{N} fx_i(x) \qquad (13)$$

Μια ομάδα μπορεί να οριστεί είτε βάσει ενός σημείου-κέντρου (βαρύκεντρου των σημείων της ομάδας), όπως στην τεχνική k-means, οπότε παίρνει το χαρακτηρισμό της *κεντρο-ορισμένης* ομάδας (center-defined cluster), είτε βάσει των κέντρων πολλών ομάδων που ενώνονται σχηματίζοντας ένα μονοπάτι, οπότε παίρνει το χαρακτηρισμό της *πολυ-κεντρο-ορισμένης* ομάδας (multi-center-defined clusters). Οι πολυ-κεντρο-ορισμένες ομάδες εντοπίζουν ομάδες οποιουδήποτε σχήματος.

Ωστόσο, οι ομάδες που σχηματίζονται μπορούν να περιγραφούν με μαθηματικό τρόπο που απαιτεί τον προσδιορισμό δυο παραμέτρων, των *α* και *ξ*. Η παράμετρος *α* ορίζει μια τιμή κατωφλίου που αντιστοιχεί στην απόσταση εκείνη όπου ένα σημείο στο χώρο παύει να έχει επιρροή. Η παράμετρος *ξ* ορίζει μια τιμή κατωφλίου που χαρακτηρίζει έναν ελκυστή

πυκνότητας σημαντικό ή ασήμαντο. Το σχήμα 4.6 απεικονίζει δύο ενδεικτικά παραδείγματα τιμών της παραμέτρου ξ.



Σχήμα 4.6 Ενδεικτικές τιμές παραμέτρου ξ

Η παράμετρος ξ ισούται με την ελάχιστη τιμή πυκνότητας που πρέπει να έχει ένας ελκυστής πυκνότητας προκειμένου να είναι σημαντικός. Εάν θέσουμε $\xi = 0$, όλοι οι ελκυστές θεωρούνται σημαντικοί. Αυτό όμως, δεν είναι πάντα επιθυμητό, ιδιαίτερα σε περιπτώσεις πολύ αραιών σημείων όπου όλα τα μεμονωμένα σημεία δύναται να αποτελέσουν ομάδες από μόνα τους. Η επιλογή της τιμής της παραμέτρου ξ εξαρτάται από την περίπτωση, ωστόσο, θα πρέπει να είναι τέτοια που να εντοπίζονται μόνο οι σημαντικοί ελκυστές πυκνότητας άρα και οι σημαντικές, ανά περίπτωση, ομάδες.

Ακόμα, δύναται να οριστεί και μια τρίτη παράμετρος, η σ, η οποία προσδιορίζει το βαθμό ομαλότητας της παραγόμενης επιφάνειας και επηρεάζει άμεσα τον αριθμό των ελκυστών πυκνότητας. Για την επιλογή της παραμέτρου σ, είναι απαραίτητη η δοκιμή διαφόρων τιμών προκειμένου να προσδιοριστεί το μεγαλύτερο εύρος τιμών της σ, $[\sigma_{max},\sigma_{min}]$, για το οποίο ο αριθμός των ελκυστών πυκνότητας παραμένει σταθερός.

Η τεχνική DENCLUE γενικεύει την τεχνική ομαδοποίησης k-means παρέχοντας τη βέλτιστη γενική διαίρεση του συνόλου των σημείων. Η τεχνική ομαδοποίησης k-means μπορεί να οριστεί σαν μια τεχνική DENCLUE, όπου:

- οι συναρτήσεις επιρροής των σημείων είναι συναρτήσεις Gauss
- οι ομάδες είναι κεντρο-ορισμένες
- η παράμετρος ξ ισούται με 0

Η ιεραρχική τεχνική ομαδοποίησης μπορεί να οριστεί σαν μια τεχνική DENCLUE, όπου:

- οι ομάδες είναι κεντρο-ορισμένες
- οι ομάδες σχηματίζουν μια ιεραρχία για διαφορετικές τιμές του σ.

4.4 Ενδεικτικά Παραδείγματα Τεχνικών Χωρικοποίησης

Στην ενότητα αυτή εξετάζεται λεπτομερώς η διαδικασία της απεικόνισης και της μείωσης των διαστάσεων των δεδομένων, στο πλαίσιο μιας χωρικοποίησης. Ξεκινώντας από την περιγραφή της προσέγγισης του Bertin, αναλύονται οι τεχνικές χωρικοποίησης της ανάλυσης σε κυρίες συνιστώσες (Principal Component Analysis - PCA), της πολυδιάστατης ή πολύ-ανυσματικής κλιμάκωσης (Multidimensional Scaling - MDS), του αυτό-οργανούμενου χάρτη (Self-Organized Map - SOM) και της δημιουργίας χώρων Benedikt.

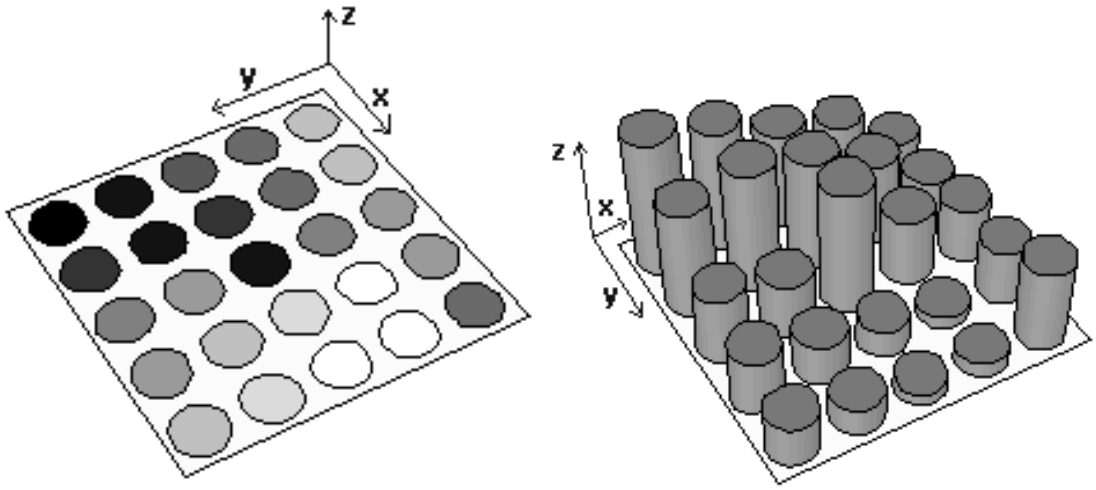
4.4.1. Η Προσέγγιση του Bertin

Ο γάλλος χαρτογράφος, Bertin, μελέτησε τη σχεδίαση γραφημάτων και την αποτελεσματικότητά τους ως προς τη μετάδοση πληροφορίας. Στα συγγράμματά του [Bert81] και [Bert83], ο Bertin ανέπτυξε μια πρωτότυπη, για την εποχή εκείνη, θεωρία, εισάγοντας την έννοια των οπτικών μεταβλητών και αναλύοντας τη σημασία των χαρτογραφικών συμβόλων. Το έργο του Bertin, αποτελεί το θεμέλιο της σημερινής έρευνας που αφορά στην οπτικοποίηση πληροφοριών.

Σύμφωνα με τον Bertin, η μετάδοση πληροφορίας μέσω γραφημάτων, στηρίζεται στην ανάδειξη σχέσεων ομοιότητας, ταξινόμησης και αναλογίας, μεταξύ των δεδομένων ενός

συνόλου. Τα γραφήματα ορίζουν ένα σύστημα σημάτων. Κάθε γράφημα αντιστοιχεί σε συγκεκριμένη υλοποίηση που παράγεται από το σύστημα αυτό. Ειδικότερα, το γράφημα μπορεί να είναι ένα διάγραμμα, ένα δίκτυο ή ένας χάρτης. Ο Bertin εξετάζει την εικόνα, ως μέσο απόδοσης των γραφημάτων, στο πλαίσιο της «θεωρίας του περί εικόνας». Ο όρος *εικόνα* υποδηλώνει έμμεσα μια στατική θεώρηση των πραγμάτων, που αδυνατεί να περιγράψει φαινόμενα εξελισσόμενα στον χρόνο. Ωστόσο, ο Bertin θεωρεί ότι τα φαινόμενα που εξελίσσονται στο χρόνο, μπορούν να περιγραφούν υπό τη μορφή ταξινομημένης συλλογής εικόνων ή καλύτερα, ακολουθίας στιγμιότυπων.

Ο Bertin τονίζει πως η κύρια δυσκολία στη δημιουργία γραφημάτων, έγκειται στο γεγονός ότι η εικόνα χαρακτηρίζεται από έναν *αξεπέραστο φραγμό τριών διαστάσεων,* ενώ καλούμαστε να οπτικοποιήσουμε δεδομένα με πάρα πολλά χαρακτηριστικά ή διαστάσεις. Με τον ισχυρισμό αυτό, επεσήμανε έμμεσα την κύρια δυσκολία στη σχεδίαση μιας χωρικοποίησης. Όπως φαίνεται στο σχήμα 4.7, κάθε στοιχείο μιας εικόνας, μπορεί να θεωρηθεί ως η συνιστώσα τριών διαστάσεων: μιας θέσης στον άξονα των x, μιας θέσης στον άξονα των y, και μιας τιμής στον άξονα των z.



Σχήμα 4.7 Χρήση των οπτικών μεταβλητών *ένταση* και *μέγεθος* στην τρίτη διάσταση

Ο Bertin προτείνει να ξεπεραστεί ο περιορισμός των τριών διαστάσεων της εικόνας με το σχεδιασμό ακολουθίας γραφημάτων και τη χρήση οπτικών μεταβλητών. Οι οπτικές μεταβλητές αντιστοιχούν στα οκτώ είδη μεταβολών που αντιλαμβάνεται το ανθρώπινο μάτι και που μπορούν να χρησιμοποιηθούν για την ανάδειξη των σχέσεων ομοιότητας,

ταξινόμησης και αναλογίας. Πρόκειται για τις δυο διαστάσεις *x*, *y* του επιπέδου, το *μέγεθος* και την *ένταση*, το *μοτίβο*, το *χρώμα*, τον *προσανατολισμό* και το *σχήμα*. Οποιαδήποτε οπτική αναπαράσταση αντικειμένων στο χώρο συσχετίζει μια τιμή, με συγκεκριμένη θέση στο επίπεδο. Οι οπτικές μεταβλητές x, y προσδιορίζουν τη θέση, ενώ όλες οι υπόλοιπες αναφέρονται στην τιμή που θα μεταβάλλεται κατά την τρίτη διάσταση.

Με εξαίρεση ορισμένες πολύ απλές περιπτώσεις, ο Bertin διευκρινίζει ότι, η υπέρθεση πολλών εικόνων, δηλαδή η ταυτόχρονη χρήση πολλών οπτικών μεταβλητών κατά την τρίτη διάσταση, έχει σαν αποτέλεσμα να χαθεί η ικανότητα άμεσης αντίληψης της πληροφορίας που μεταδίδεται από τις επιμέρους εικόνες. Μάλιστα, χρειάζεται να διαθέσουμε πολύ περισσότερο χρόνο για την κατανόηση της τελικής εικόνας, από το συνολικό χρόνο που θα διαθέταμε για την κατανόηση των επιμέρους εικόνων.

4.4.2. Ανάλυση σε Κύριες Συνιστώσες (Principal Component Analysis - PCA)

Η *ανάλυση σε κύριες συνιστώσες* (PCA) είναι καταρχάς μια τεχνική στατιστικής ανάλυσης που χρησιμοποιείται ευρέως στην ανάλυση και τη συμπίεση δεδομένων και κατά δεύτερον μια τεχνική προβολής πολυδιάστατων δεδομένων που χρησιμοποιείται ως τεχνική χωρικοποίησης. Όπως αναφέρθηκε στις προηγούμενες παραγράφους, η κύρια δυσκολία στο σχεδιασμό μιας χωρικοποίησης έγκειται στην απεικόνιση δεδομένων με περισσότερα χαρακτηριστικά (ή μεταβλητές) απ’ ότι οι διαστάσεις του χώρου προορισμού. Η PCA προσπερνά τη δυσκολία αυτή, απεικονίζοντας τα δεδομένα στο σύστημα αναφοράς που ορίζουν οι κύριες συνιστώσες των δεδομένων, περιορίζοντας έτσι το πλήθος των διαστάσεων του χώρου απεικόνισης.

Πιο συγκεκριμένα, η PCA συνίσταται στη δημιουργία ενός μικρότερου συνόλου νέων μεταβλητών, που προκύπτουν από το γραμμικό μετασχηματισμό των αρχικών μεταβλητών και όχι στην επιλογή υποσυνόλου αυτών. Η απλή επιλογή υποσυνόλου από τις αρχικές μεταβλητές θα είχε προφανώς ως συνέπεια την απώλεια σημαντικού ποσού πληροφορίας. Ενώ οι αρχικές μεταβλητές συσχετίζονται πιθανότατα μεταξύ τους και συνεπώς περιέχουν πλεονάζουσα πληροφορία, οι μετασχηματισμένες μεταβλητές είναι ασυσχέτιστες μεταξύ τους και μπορούν να αντιπροσωπεύσουν επάξια, σχεδόν χωρίς απώλεια πληροφορίας, τις αρχικές μεταβλητές.

Οι κύριες συνιστώσες υπολογίζονται βάσει του πίνακα διασποράς και των ιδιοδιανυσμάτων του. Τα ιδιοδιανύσματα αυτά προσδιορίζουν τη διεύθυνση κατά την οποία τα δεδομένα μεταβάλλονται περισσότερο. Οι προβολές των αρχικών δεδομένων στις διευθύνσεις που ορίζουν τα ιδιοδιανύσματα, σχηματίζουν τις κύριες συνιστώσες. Οι ιδιοτιμές των ιδιοδιανυσμάτων δίνουν μια προσέγγιση του ποσού της πληροφορίας που σχετίζεται με κάθε συνιστώσα. Οι συνιστώσες που αντιστοιχούν στις μεγαλύτερες ιδιοτιμές μεταδίδουν μεγαλύτερο ποσό πληροφορίας.

Στη συνέχεια, γίνεται μια σύντομη αναφορά σε έννοιες και διαδικασίες γνωστές από τη στατιστική και τη γραμμική άλγεβρα, στις οποίες στηρίζεται η PCA και οι οποίες περιγράφονται αναλυτικότερα στο [Joli86].

Έστω X, τυχαίο διάνυσμα, με $X = (x_1, \ldots, x_n)^T$ και $x_1, \ldots, x_n$, τυχαίες μεταβλητές. Η μέση τιμή του X και ο πίνακας διασποράς ορίζονται αντίστοιχα από τους τύπους 14 και 15:

$$\mu_x = E\{X\} \qquad (14) \qquad C_x = E\{(X - \mu_x)(X - \mu_x)^T\} \qquad (15)$$

Τα στοιχεία C_{ij} του πίνακα C_x, αντιπροσωπεύουν τις συνδιακυμάνσεις μεταξύ των τυχαίων μεταβλητών x_i και x_j, του τυχαίου διανύσματος X. Το στοιχείο C_{ii} είναι η μεταβλητότητα της μεταβλητής x_i. Η μεταβλητότητα μιας μεταβλητής x_i προσδιορίζει τη διασπορά της x_i ως προς τη μέση τιμή της. Εάν οι δυο μεταβλητές x_i και x_j, των δεδομένων είναι ασυσχέτιστες, οι συνδιακυμάνσεις τους C_{ji} και C_{ij}, ισούνται με 0. Ο πίνακας διασποράς είναι εξ' ορισμού συμμετρικός. Επειδή, ο πίνακας διασποράς είναι συμμετρικός, είναι δυνατός ο υπολογισμός μιας ορθογώνιας βάσης, από τις ιδιοτιμές του και τα ιδιοδιανύσματα του. Τα ιδιοδιανύσματα e_i και οι αντίστοιχες ιδιοτιμές λ_i, αποτελούν τις λύσεις της εξίσωσης 16

$$C_x e_i = \lambda_i e_i \text{ όπου } i = 1, \ldots, n. \qquad (16)$$

Υπολογίζουμε τις ιδιοτιμές, από την επίλυση της χαρακτηριστικής εξίσωσης 17:

$$|C_x - \lambda I| = 0, \qquad (17)$$

Στην εξίσωση 17, ο I είναι ο μοναδιαίος πίνακας, βαθμού ίσου με του C_x, και ο συμβολισμός |.|, αναφέρεται στην ορίζουσα πίνακα. Για λόγους απλοποίησης, υποθέτουμε ότι οι ιδιοτιμές είναι διαφορετικές ανά μεταξύ τους. Εάν το διάνυσμα δεδομένων έχει n

συστατικά, η χαρακτηριστική εξίσωση είναι n βαθμού. Αντιλαμβανόμαστε ότι όσο πιο μικρό το n, τόσο πιο εύκολα λύνεται η χαρακτηριστική εξίσωση.

Τα πολυδιάστατα δεδομένα στα οποία θα εφαρμοστεί PCA, παρομοιάζονται με πολυδιάστατα τυχαία στατιστικά φαινόμενα, ή, τυχαίες μεταβλητές. Έστω ότι επεξεργαζόμαστε ένα δείγμα m περιπτώσεων n-διάστατων δεδομένων. Κάθε περίπτωση, i, δύναται να θεωρηθεί ως η i γραμμή ενός πίνακα δεδομένων Π, m x n, ή, ως ένα n-διάστατο διάνυσμα $X_i(x_1, \ldots, x_n)$. Για το δείγμα των διανυσμάτων $X_1, X_2, \ldots, X_m$, μπορούμε να εκτιμήσουμε τη μέση τιμή και τον πίνακα διασποράς του δείγματος. Δημιουργούμε, μία ορθογώνια βάση, από τα ιδιοδιανύσματα και τις ιδιοτιμές του πίνακα διασποράς, ταξινομημένη ανά φθίνουσα σειρά ιδιοτιμών. Έτσι, το πρώτο ιδιοδιάνυσμα αντιστοιχεί στην διεύθυνση της μεγαλύτερης μεταβλητότητας δεδομένων.

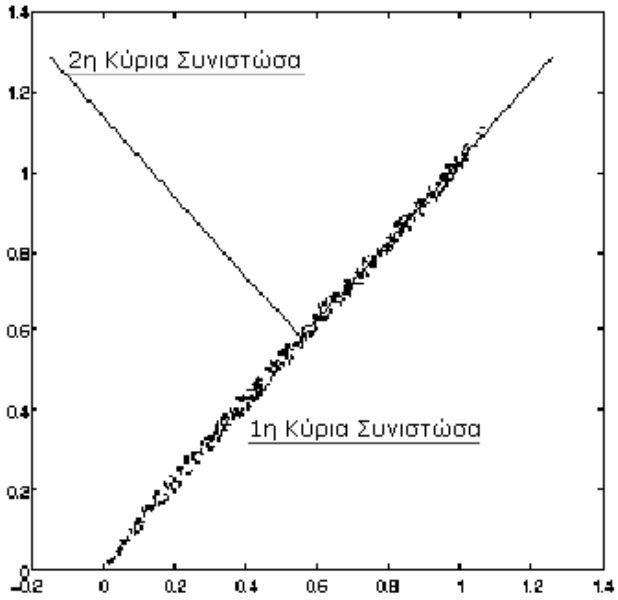
Ας υποθέσουμε ότι, για το σύνολο των δεδομένων, έχουμε υπολογίσει τη μέση τιμή και τον πίνακα διασποράς. Ορίζουμε έναν πίνακα Α, του οποίου οι γραμμές αποτελούνται από τα ιδιοδιανύσματα του πίνακα διασποράς. Μετασχηματίζοντας τα διανύσματα δεδομένων X_i σύμφωνα με το γραμμικό μετασχηματισμό 18, προκύπτουν διανύσματα Y_i, που αντιστοιχούν σε σημεία, στο ορθογώνιο σύστημα αναφοράς που ορίζουν τα ιδιοδιανύσματα.

$$Y_i = A(X_i - \mu_X) \qquad (18)$$

Τα χαρακτηριστικά των Y_i αντιστοιχούν στις συντεταγμένες τους, στην ορθογώνια βάση. Εάν λάβουμε υπόψη μας την ιδιότητα των ορθογώνιων πινάκων, $A^{-1} = A^T$, μπορούμε εύκολα να δημιουργήσουμε ξανά το αρχικό διάνυσμα δεδομένων, λύνοντας την εξίσωση μετασχηματισμού ως προς X_i ως εξής: $X_i = A^T Y_i + \mu_X$. Συνεπώς, προβάλαμε τα αρχικά διανύσματα δεδομένων στους άξονες συντεταγμένων που όριζε η ορθογώνια βάση, και τα δημιουργήσαμε ξανά με τον αντίστροφο μετασχηματισμό. Έχοντας ως στόχο τη μείωση των διαστάσεων των δεδομένων προς απεικόνιση, θα μπορούσαμε να επαναλάβουμε τη διαδικασία αυτή, λαμβάνοντας υπόψη μας μόνο k από τα ιδιοδιανύσματα, δημιουργώντας τον πίνακα A_k. Επιλέγοντας τα ιδιοδιανύσματα που αντιστοιχούν στις μεγαλύτερες ιδιοτιμές, ελαχιστοποιούμε το σφάλμα, που μεταφράζεται ως απώλεια πληροφορίας, από την παράληψη των υπόλοιπων ιδιοδιανυσμάτων. Συνεπώς, προκύπτουν δυο αντιφατικοί στόχοι: αφενός μεν, θέλουμε να απλοποιήσουμε το πρόβλημα, μειώνοντας τις διαστάσεις του χώρου

απεικόνισης, αφετέρου δε, θέλουμε να διατηρήσουμε όσο γίνεται περισσότερη πληροφορία κατά την απεικόνιση. Η PCA επιτρέπει να βρεθεί η χρυσή τομή.

Στο σχήμα 4.8 που ακολουθεί, ένα υποθετικό σύνολο δεδομένων απεικονίζεται στο δισδιάστατο χώρο που ορίζουν τα ιδιοδιανύσματα των δυο μεγαλύτερων ιδιοτιμών του πίνακα διασποράς.



Σχήμα 4.8 Παράδειγμα απεικόνισης δεδομένων στο χώρο που ορίζουν τα ιδιοδιανύσματα.

Οι κατευθύνσεις των δυο ιδιοδιανυσμάτων εμφανίζονται στο σχήμα ως δύο κάθετες ευθείες. Παρατηρούμε ότι τα μετασχηματισμένα δεδομένα κατανέμονται (μεταβάλλονται) κυρίως κατά τη διεύθυνση του πρώτου ιδιοδιανύσματος (προς τα πάνω και δεξιά), που αντιστοιχεί στη μεγαλύτερη ιδιοτιμή. Εάν τα δεδομένα είναι συγκεντρωμένα γραμμικά σε περιοχή του χώρου απεικόνισης, σημαίνει ότι η απεικόνιση επιδέχεται απλοποίηση χωρίς να συνεπάγεται παράλληλα και χάσιμο πολύτιμης πληροφορίας. Στο σχήμα 4.8, όπου το πρώτο ιδιοδιάνυσμα κατέχει σχεδόν όλη την «ενέργεια» της πληροφορίας, θα αρκούσε μια και μόνο διάσταση για την χωρική απεικόνιση των δεδομένων.

Στη βιβλιογραφία, χρησιμοποιείται ο όρος *ενέργεια* πληροφορίας για να εκφράσει το ποσό της πληροφορίας που μεταδίδεται μέσω της απεικόνισης. Συγκρίνοντας κάθε ιδιοτιμή με το

άθροισμα όλων των ιδιοτιμών, υπολογίζουμε το ποσοστό συμμετοχής της ιδιοτιμής στο συνολικό ποσό πληροφορίας. Το ποσοστό αυτό δίνεται από τον τύπο 19:

$$\%\text{Variance}_i = (\lambda_i \,.\, 100) \,/\, \sum_{k=1}^{n} \lambda_k \qquad (19)$$

Στην πράξη, έχει βρεθεί ότι, το άθροισμα των ποσοστών συμμετοχής των επιλεγμένων ιδιοτιμών, πρέπει να ξεπερνά το 75%, για να έχει νόημα η μείωση των διαστάσεων.

Κατά την ερμηνεία των αποτελεσμάτων της χωρικοποίησης με PCA, δεν εξετάζουμε τη μεταβολή των μεμονωμένων χαρακτηριστικών των δεδομένων, αλλά τη μεταβολή των γραμμικών συνδυασμών των μεταβλητών που συμμετέχουν στις κύριες συνιστώσες. Τα χαρακτηριστικά των αρχικών δεδομένων συνεισφέρουν περισσότερο ή λιγότερο στη μεταβολή των κύριων συνιστωσών, ανάλογα με το συντελεστή συνδιακύμανσης με το οποίο πολλαπλασιάζονται στον τύπο 18. Από γεωμετρικής άποψης, η PCA συνίσταται σε στροφή των αρχικών αξόνων προς τις διευθύνσεις εκείνες κατά τις οποίες τα δεδομένα κατανέμονται κατά κύριο λόγο. Ωστόσο, θεωρείται γενικά ότι η τεχνική PCA παράγει αποτελέσματα που δεν ερμηνεύονται εύκολα [Berk06].

4.4.3. Πολυδιάστατη ή Πολύ-ανυσματική Κλιμάκωση (Multidimensional Scaling - MDS)

Ο σκοπός της *πολυδιάστατης* ή *πολυ-ανυσματικής κλιμάκωσης* (MDS) έγκειται στην απεικόνιση πολυδιάστατων δεδομένων σε χώρο περιορισμένων διαστάσεων και αποτελεί δημοφιλή τεχνική χωρικοποίησης. Πιο συγκεκριμένα, τα πολυδιάστατα δεδομένα απεικονίζονται υπό τη μορφή σημείων, στα πλαίσια δισδιάστατης ή τρισδιάστατης χωρικής διάταξης. Η πληροφορία που μεταδίδεται, μέσω της οπτικοποίησης αυτής, αφορά στην ομοιότητα, κατά μιαν έννοια, των δεδομένων μεταξύ τους. Η σχετική θέση, που κατέχουν δυο δεδομένα στο χώρο της απεικόνισης, σχετίζεται άμεσα με το πόσο μοιάζουν ή διαφέρουν μεταξύ τους. Όσο μεγαλύτερη είναι η απόσταση που χωρίζει δυο σημεία στο χώρο απεικόνισης, τόσο διαφορετικά είναι τα αντίστοιχα δεδομένα στην πραγματικότητα.

Έστω ότι εξετάζουμε την περίπτωση εφαρμογής της τεχνικής MDS σε n δεδομένα διαστάσεων m. Αφού δομήσουμε τον πίνακα δεδομένων Π, n x m, δημιουργούμε έναν

πίνακα σύγκρισης Σ, n x n, στον οποίο κάθε στοιχείο Σ(i, j) αντιστοιχεί στο αποτέλεσμα της σύγκρισης κατά μια έννοια, των δεδομένων i και j. Καταχωρείται δηλαδή μια τιμή ανάλογη με το πόσο διαφέρουν τα δεδομένα μεταξύ τους,. Είναι προφανές ότι:

- τα στοιχεία Σ(i, i), για κάθε i = 1, …, n, ισούνται με 0. Κανένα δεδομένο δεν διαφέρει από τον εαυτό του.
- Ο πίνακας είναι συμμετρικός, δηλαδή Σ(i, j) = Σ(j, i), για κάθε i = 1, …, n και για κάθε j = 1, …, n

Η διαδικασία ορισμού του πίνακα σύγκρισης προϋποθέτει τον ορισμό ενός τρόπου μέτρησης της ομοιότητας δεδομένων. Υπάρχουν πολλοί δυνατοί τρόποι μέτρησης αυτής της ομοιότητας. Αναφέρουμε τη *μέτρηση τύπου απόσταση* (distance-type measure) και τη *μέτρηση τύπου ταίριασμα* (matching-type measure). Η μέτρηση τύπου απόσταση βασίζεται στον τύπο της *Ευκλείδειας απόστασης* (τύπος 20):

$$d_{ij} = \sqrt{\sum_{a=1}^{r}(x_{ia} - x_{ja})^2} \qquad (20)$$

όπου x_{ia} και x_{ja}, οι προβολές ή συντεταγμένες των σημείων στην διάσταση a (a=1,2, …, r). Υπάρχουν όμως και άλλοι ορισμοί, όπως για παράδειγμα η *απόσταση Minkowski*. Η μέτρηση της ομοιότητας βάσει του τύπου απόσταση εφαρμόζεται συνήθως σε ποσοτικά δεδομένα, αφού πρώτα εξασφαλιστεί η αναφορά τους στις ίδιες μονάδες μέτρησης και η κανονικοποίηση τους. Η κανονικοποίηση των δεδομένων εξισώνει τη μεταβλητότητα τους με τη μονάδα: Var(X)=1. Το αποτέλεσμα αυτό προκύπτει από τη διαίρεση των τιμών των χαρακτηριστικών με τις αντίστοιχες μεταβλητότητες. Με τον τρόπο αυτό εξασφαλίζεται η ισοδύναμη συμβολή όλων των χαρακτηριστικών στον υπολογισμό της ομοιότητας των δεδομένων. Ακόμα, αποκλείουν το γεγονός τα χαρακτηριστικά μεγάλης μεταβλητότητας να επισκιάσουν τα χαρακτηριστικά μικρής μεταβλητότητας και να συνεισφέρουν, τελικά, μόνα τους στη μέτρηση της ομοιότητας των δεδομένων.

Η μέτρηση τύπου ταίριασμα εφαρμόζεται σε περιπτώσεις σύγκρισης δεδομένων των οποίων το σύνολο των χαρακτηριστικών είναι ονομαστικού τύπου. Στην περίπτωση κατά την οποία τα χαρακτηριστικά είναι διαφορετικού τύπου, προσπαθούμε, εφόσον είναι δυνατόν, να μετατρέψουμε τα ονομαστικά σε ποσοτικά. Για παράδειγμα το χαρακτηριστικό *φύλο* με τιμές

Άνδρας ή *Γυναίκα*, μπορεί να μετασχηματιστεί στο ποσοτικό χαρακτηριστικό *φύλο* με τιμές *0* και *1*. Για την περιγραφή αυτού του τρόπου μέτρησης της ομοιότητας, ας εξετάσουμε το ακόλουθο παράδειγμα. Το αντικείμενο Α, με έξι χαρακτηριστικά, συγκρίνεται διαδοχικά με τα αντικείμενα Β και Γ. Έστω ότι τα Α, Β και Γ τα εξής:

Α: [a,b,c,d,e,f], Β: [a,g,h,k,e,i] και Γ: [j,b,c,d,e,f]

Συγκρίνοντας ένα προς ένα τα χαρακτηριστικά του Α προς τα αντίστοιχα των άλλων αντικειμένων, δημιουργούμε τα δυο σύνολα τιμών για την περιγραφή του αποτελέσματος της σύγκρισης.

Αποτέλεσμα σύγκρισης Α και Β:[1,0,0,0,1,0]

Αποτέλεσμα σύγκρισης Α και Γ:[0,1,1,1,1,1]

Το *1*, αντιστοιχεί σε επιτυχημένο ταίριασμα ενώ το *0*, σε αποτυχημένο ταίριασμα. Μετρώντας τώρα τον αριθμό τον επιτυχημένων ταιριασμάτων, βρίσκουμε ότι:

$S_{A/B}$ = (αρ. επιτυχημένων ταιριασμάτων / αρ. χαρακτηριστικών) = 2/6 = 0.333

$S_{A/Γ}$ = (αρ. επιτυχημένων ταιριασμάτων / αρ. χαρακτηριστικών) = 5/6 = 0.833

Παρατηρούμε ότι τα αποτελέσματα οποιασδήποτε σύγκρισης κυμαίνεται μεταξύ 0 και 1.

Μετά τον ορισμό του τρόπου μέτρησης της ομοιότητας των δεδομένων, είναι δυνατός ο ορισμός του πίνακα σύγκρισης των δεδομένων, Σ (n x n). Στη συνέχεια, ξεκινά η διαδικασία της απεικόνισης των πολυδιάστατων δεδομένων υπό τη μορφή σημείων, στο δισδιάστατο ή τρισδιάστατο χώρο. Η διαδικασία της απεικόνισης συνίσταται στην χωροθέτηση σημείων έτσι ώστε οι αποστάσεις ανά μεταξύ τους να προσεγγίσουν όσο το δυνατόν καλύτερα τις αποστάσεις που αναφέρονται στον πίνακα σύγκρισης. Αυτό συμβαίνει εφόσον ελαχιστοποιηθεί ένας παράγοντας που ονομάζεται *Stress*.

Όπως επισημαίνει ο Borgatti [Borg97], ο παράγοντας stress ορίζεται ως ο βαθμός προσέγγισης των σημείων στις τιμές του πίνακα σύγκρισης. Ο ορισμός του Kruskal [KW78] αναφέρεται στη σχέση 21:

$$\text{Stress} = \sqrt{(\Sigma\Sigma(f(x_{ij}) - d_{ij})^2 / \Sigma\Sigma d_{ij}^2)} \qquad (21)$$

όπου $f(x_{ij})$, συνάρτηση που εξαρτάται από το αν έχουμε ποσοτικά ή όχι δεδομένα. Στην καταφατική περίπτωση, $f(x_{ij}) = x_{ij}$. Με άλλα λόγια, τα ακατέργαστα δεδομένα συγκρίνονται απ' ευθείας με τις αποστάσεις στον χώρο απεικόνισης. Σε αντίθετη περίπτωση, η $f(x_{ij})$ ορίζεται ως μια συνάρτηση που ελαχιστοποιεί τον τύπο του Stress.

Η μαθηματική εξήγηση που δίνεται όταν αποτυγχάνει ο μηδενισμός του παράγοντα Stress, είναι μια και μοναδική: ανεπαρκής αριθμός διαστάσεων του χώρου απεικόνισης. Δεν μπορούν να αναπαρασταθούν τα δεδομένα σε δισδιάστατο χώρο ή σε χώρο με μικρότερο αριθμό διαστάσεων. Ωστόσο, δεν είναι απαραίτητο να πετύχουμε μηδενική τιμή του Stress για να είναι χρήσιμη η χωρικοποίηση. Μια μικρή παραμόρφωση του τελικού αποτελέσματος είναι ανεκτή. Συνήθως, ακολουθείται ο εξής κανόνας: εάν το Stress είναι μικρότερο από 0.1, έχουμε άριστη προσέγγιση και αμελητέα παραμόρφωση, ενώ αν είναι μεγαλύτερο από 0.15, η προσέγγιση είναι απαράδεχτη.

Ο αλγόριθμος απεικόνισης των σημείων περιλαμβάνει τα εξής βήματα:

1. Αποδίδουμε αρχικά στα σημεία τυχαίες θέσεις στον χώρο απεικόνισης
2. Υπολογίζουμε όλες τις ευκλείδειες αποστάσεις μεταξύ των σημείων και τις καταχωρούμε σε προσωρινό πίνακα, έστω T (n x n).
3. Συγκρίνουμε τις τιμές του πίνακα T με αυτές του πίνακα σύγκρισης Σ, υπολογίζοντας παράλληλα την τιμή του παράγοντα Stress. Όσο μικρότερη είναι η τιμή αυτή, τόσο καλύτερη προσέγγιση έχουμε πετύχει.
4. Προσαρμόζουμε τις συντεταγμένες κάθε σημείου προς την κατεύθυνση εκείνη που μειώνει το Stress.
5. Επαναλαμβάνουμε τα βήματα 2 έως 4, μέχρις ότου παρατηρήσουμε ότι δεν μειώνεται άλλο η τιμή του Stress.

Το πέρας του αλγορίθμου, ακολουθεί η ερμηνεία του αποτελέσματος της χωρικοποίησης. Καταρχήν, παρατηρούμε ότι ο προσανατολισμός των αξόνων είναι άνευ σημασίας. Ανεξάρτητα από το πώς θα στρέψουμε τους άξονες, οι σχετικές αποστάσεις των αναπαριστώμενων οντοτήτων παραμένουν αναλλοίωτες. Ο ερμηνευτής εξετάζει τη διάταξη των οντοτήτων στο χώρο. Παρατηρώντας τη γενική κατανομή των οντοτήτων στο χώρο,

προσπαθεί να διακρίνει σχηματιζόμενες ομάδες όμοιων οντοτήτων, οι οποίες πιθανότατα να μπορούν να κατηγοριοποιηθούν. Ακόμα η παρουσία ειδικών σχημάτων όπως κύκλοι, κτλ, έχουν τη δική τους ερμηνεία. Ο Kruskal και ο Wish [KW78] ασχολήθηκαν με την ερμηνεία τέτοιων αποτελεσμάτων.

Συγκρίνοντας την MDS με την PCA, θα λέγαμε ότι διαφέρουν σε πάρα πολλά σημεία. Το δυνατό σημείο της MDS είναι ότι εφαρμόζεται χρησιμοποιώντας οποιουδήποτε τύπου απόσταση ή ομοιότητα. Για το λόγο αυτό, χρησιμοποιείται ευρέως από επιστήμονες διαφόρων ειδικοτήτων: ψυχολόγους, βιολόγους, κοινωνιολόγους, φυσικούς και χαρτογράφους. Η PCA αντίθετα προϋποθέτει τον υπολογισμό του πίνακα μεταβλητότητας. Ακόμα, η PCA απαιτεί τα πολυδιάστατα δεδομένα να ακολουθούν κανονική κατανομή και να διέπονται από γραμμικές σχέσεις. Η MDS δεν έχει τέτοιο περιορισμό. Όσον αφορά τη μείωση των διαστάσεων του χώρου απεικόνισης, η PCA τείνει παράγει χώρους με περισσότερες διαστάσεις και τα αποτελέσματα της ερμηνεύονται πιο δύσκολα απ' ότι της MDS.

4.4.4. Ο Αυτό-Οργανούμενος Χάρτης (Self-Organized Map - SOM)

Ο *αυτο-οργανούμενος χάρτης* (SOM) αποτελεί μια δημοφιλή εφαρμογή της θεωρίας των νευρωνικών δικτύων. Δεν απαιτεί ούτε την ανθρώπινη παρέμβαση ούτε τη λεπτομερή γνώση των χαρακτηριστικών των δεδομένων. Θα μπορούσαμε για παράδειγμα να χρησιμοποιήσουμε το SOM για την ομαδοποίηση δεδομένων χωρίς να γνωρίζουμε εκ των προτέρων τον αριθμό των ομάδων.

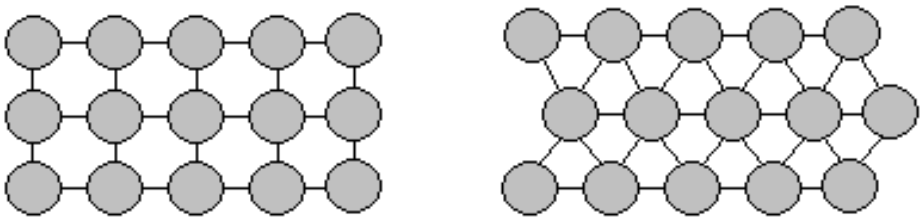
Το SOM αναπτύχθηκε από τον Teuvo Kohonen [Koho95], Καθηγητή στο Πανεπιστήμιο του Helsinki, στην αρχή της δεκαετίας του 1980 και από τότε χρησιμοποιείται σε πληθώρα εφαρμογών. Για παράδειγμα, το SOM έχει άμεση εφαρμογή στην αναζήτηση δεδομένων από βιβλιοθήκες ή από το διαδίκτυο (WEBSOM - Self-Organizing Maps for Internet Exploration, [KKL+00]). Ακόμα, ο Vesanto [Vesa00] μελέτησε την εφαρμογή του SOM στην εξόρυξη γνώσης και ο Kaski [Kask97] την εξερεύνηση πληροφοριών με χρήση SOM.

Ο Kohonen [Koho95] περιγράφει το SOM ως «*μία μη-γραμμική, ταξινομημένη, ήπια αντιστοίχιση ενός συνόλου πολυδιάστατων δεδομένων με τα στοιχεία ενός πίνακα*

περιορισμένων διαστάσεων». Ο Goodchild [Good08] θεωρεί το SOM σαν «*μια τεχνική απόδοσης χωρικής αναφοράς σε δεδομένα χωρίς χωρική υπόσταση*». Το SOM παράγει γράφους ομοιότητας που διατηρούν τις τοπολογικές σχέσεις και μπορούν να χρησιμοποιηθούν για αναγνώριση προτύπων ή ομαδοποίηση .

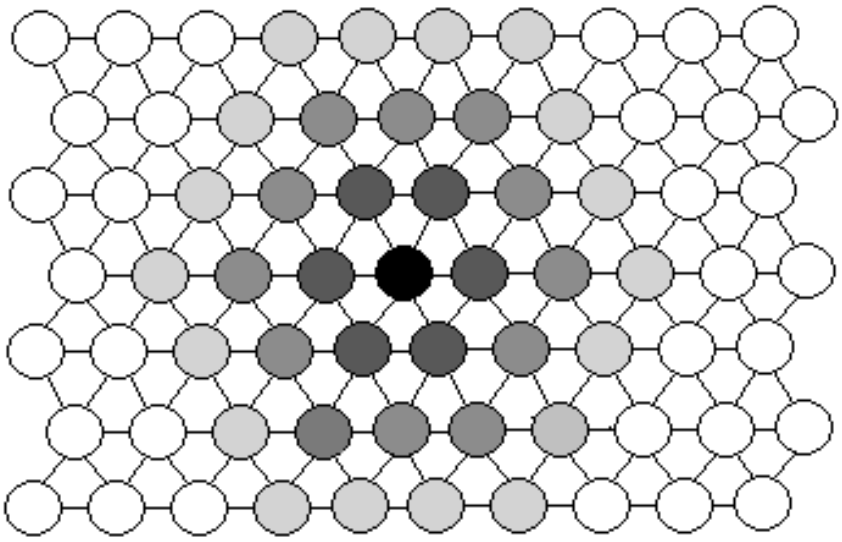
Κατά τη δημιουργία ενός SOM, λαμβάνει χώρα μια χωρικοποίηση που απεικονίζει τον πολυδιάστατο χώρο των δεδομένων, στο δισδιάστατο χώρο του αυτο-οργανούμενου χάρτη. Η απεικόνιση αυτή έχει την ιδιότητα να διατηρεί την τοπολογία των δεδομένων, ενώ, τα όμοια δεδομένα, απεικονίζονται σε γειτονικές μονάδες ή νευρώνες.

Ο Kohonen ορίζει το SOM σαν δισδιάστατο πίνακα νευρώνων: $M = \{m_1, \ldots, m_{pxq}\}$. Οι νευρώνες αναφέρονται και ως μονάδες ή στοιχεία επεξεργασίας: $m_i = [m_{i1}, \ldots, m_{in}]$, όπου $i = 1, \ldots, p$, και έχουν τις ίδιες διαστάσεις με τα διανύσματα εισόδου, δηλαδή τις γραμμές του πίνακα δεδομένων. Στο βασικό αλγόριθμο δημιουργίας SOM, ο αριθμός των νευρώνων είναι εξ' αρχής γνωστός, προσδιορίζοντας την κλίμακα και την ακρίβεια του χάρτη. Ένας νευρώνας συνδέεται με σχέσεις γειτονίας με τους πλησιέστερούς του. Ανάλογα με τη διάταξη των νευρώνων, όπως φαίνεται στο σχήμα 4.9, η τοπολογία του SOM χαρακτηρίζεται ορθογώνια ή εξαγωνική.



Σχήμα 4.9 Τοπολογίες νευρώνων

Η απόσταση μεταξύ δυο νευρώνων ορίζεται συναρτήσει της τοπολογικής σχέσης τους. Η στενή γειτονιά, N_c, του νευρώνα m_c, συμπεριλαμβάνει όλου τους νευρώνες που συνδέονται άμεσα με αυτόν. Στο σχήμα 4.10, παρουσιάζονται γειτονιές διαφόρων μεγεθών σε εξαγωνική διάταξη, όπου όσο σκουρότερη η απόχρωση του εξαγώνου, τόσο στενότερη η γειτονιά του νευρώνα με μαύρο χρώμα.



Σχήμα 4.10 Γειτονιές διαφόρων μεγεθών

Για τη δημιουργία ενός SOM, λαμβάνονται υπόψη όλα τα δεδομένα εισόδου. Εάν υπάρχουν λάθη στα δεδομένα εισόδου, πρέπει να εξαλειφθούν για να αποφευχθεί το λανθασμένο αποτέλεσμα. Η ιδιότητα αυτή είναι γνωστή ως «garbage in, garbage out». Για παράδειγμα, οι άγνωστες τιμές χαρακτηριστικών συχνά και λανθασμένα αντικαθίστανται, με μηδενικά. Η λύση για την αντιμετώπιση των άγνωστων αυτών τιμών είναι η αντικατάστασή τους με τιμές αδιάφορες («don't care» values) ή, η διαγραφή των αντίστοιχων δεδομένων εισόδου.

Πριν επεξεργαστούν από το SOM, τα δεδομένα εισόδου πρέπει να μετασχηματίζονται σε ποσοτικά και να κανονικοποιούνται. Η ποσοτικοποίηση των δεδομένων είναι αναγκαία επειδή οι συγκρίσεις μεταξύ των δεδομένων γίνονται με τη βοήθεια της Ευκλείδειας απόστασης. Για παράδειγμα, έστω ένα χαρακτηριστικό που χωρίζει τα δεδομένα στις κατηγορίες από 1 έως 10. Δεν μπορούμε να συμπεράνουμε ότι τα δεδομένα της κατηγορίας 9 είναι περισσότερο όμοια με τα δεδομένα της κατηγορίας 10 παρά με αυτά της κατηγορίας 1. Όμως, εάν μετασχηματίσουμε το συγκεκριμένο χαρακτηριστικό, ποσοτικοποιώντας την ιδιότητα των δεδομένων βάσει της οποίας γίνεται η κατηγοριοποίηση, τότε θα έχει έννοια η σύγκριση των τιμών του χαρακτηριστικού και ο ορισμός ενός μέτρου ομοιότητας.

Όπως και στην τεχνική MDS, η κανονικοποίηση των δεδομένων είναι απαραίτητη ώστε να εξασφαλιστεί η ισοδύναμη συμβολή όλων των χαρακτηριστικών στον υπολογισμό της

ομοιότητας των δεδομένων, αποκλείοντας την περίπτωση τα χαρακτηριστικά σημαντικής μεταβλητότητας να επισκιάσουν τα υπόλοιπα χαρακτηριστικά συνεισφέροντας αποκλειστικά στη μέτρηση της ομοιότητας.

Η χωρικοποίηση των δεδομένων με τη δημιουργία του SOM ξεκινάει με τη φάση της αρχικοποίησης του SOM. Ο Kohonen αναφέρει τρεις τύπους αρχικοποίησης: την τυχαία αρχικοποίηση, την αρχικοποίηση βάσει δειγμάτων από τα δεδομένα και τη γραμμική αρχικοποίηση. Στην τυχαία αρχικοποίηση, τυχαίες τιμές αποδίδονται αρχικά στους νευρώνες του SOM. Η αρχικοποίηση αυτή χρησιμοποιείται κυρίως όταν δεν γνωρίζουμε τίποτα ή λίγα σχετικά με τα δεδομένα εισόδου. Ο δεύτερος τρόπος αποδίδει αρχικά σε ορισμένους νευρώνες κάποιες από τις τιμές των δεδομένων εισόδου. Η μέθοδος αυτή έχει το πλεονέκτημα ότι στους νευρώνες αυτούς θα αποδοθούν ακριβώς οι τιμές των δεδομένων του δείγματος. Η τρίτη περίπτωση συνίσταται στην αρχικοποίηση των νευρώνων με τα ιδιοδιανύσματα εκείνα που αντιστοιχούν στις δυο μεγαλύτερες ιδιοτιμές των δεδομένων, όπως υπολογίζονται στην ανάλυση σε κύριες συνιστώσες. Η αρχικοποίηση αυτή θα επιβάλλει στο SOM να προσανατολίσει τα δεδομένα εισόδου προς τα δεδομένα εκείνα που διαθέτουν μεγαλύτερο ποσό ενέργειας πληροφορίας.

Η φάση της εκπαίδευσης ενός SOM, που έπεται της αρχικοποίησης, αποτελεί μια χρονοβόρα και επαναληπτική διαδικασία. Συνίσταται στην επιλογή διανυσμάτων εισόδου και στη μύηση των νευρώνων του SOM σ' αυτά. Σε κάθε επανάληψη, συγκρίνεται ένα από τα διανύσματα εισόδου, με όλους τους νευρώνες του SOM. Ο νευρώνας, ο περισσότερο όμοιος προς το διάνυσμα εισόδου, επιλέγεται νικητής και χαρακτηρίζεται ως η πιο ταιριαστή μονάδα (BMU - Best Matching Unit). Συνήθως, η ομοιότητα των δεδομένων υπολογίζεται βάσει της Ευκλείδειας νόρμας διανυσμάτων και της Ευκλείδειας απόστασης. Έτσι, εάν η Ευκλείδεια νόρμα ενός διανύσματος X ορίζεται ως $||X|| = \sqrt{\sum x_i^2}$, η ομοιότητα μεταξύ του διανύσματος εισόδου X, και του νευρώνα m_c, μπορεί να οριστεί ως η Ευκλείδεια απόσταση των διανυσμάτων ως: $||X - m_c|| = \min_i\{||X - m_i||\}$.

Όταν συγκριθεί το διάνυσμα εισόδου με όλους τους νευρώνες και βρεθεί η πιο ταιριαστή μονάδα, ενημερώνεται το SOM έτσι ώστε να πλησιάσει περισσότερο το διάνυσμα εισόδου. Ενημερώνονται ακόμα και οι νευρώνες στη γειτονιά της πιο ταιριαστής μονάδας.

Εάν η γειτονιά της πιο ταιριαστής μονάδας είναι μεγαλύτερη (περιλαμβάνει 16, 24, κτλ νευρώνες), τότε είναι λογικό, ο χρόνος εκπαίδευσης να αυξηθεί ανάλογα. Στην αρχή της εκπαιδευτικής φάσης, συνιστάται η χρήση μεγάλων γειτονιών και η προοδευτική μείωσή τους με την πάροδο του χρόνου. Ακόμα, είναι σημαντικό να αναφέρουμε ότι ναι μεν ο μεγάλος αριθμός νευρώνων μεγαλώνει την ακρίβεια της απεικόνισης, αλλά ο χρόνος ανεύρεσης της πιο ταιριαστής μονάδας, αυξάνεται ραγδαία.

Ο κανόνας που ισχύει κατά την ενημέρωση της μονάδας m_i του SOM, περιγράφεται από τη σχέση 22:

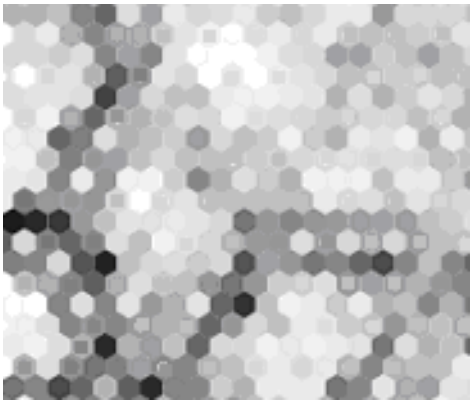
$$m_i(t + 1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)] \qquad (22)$$

Όπου t, ο χρόνος, x(t), το διάνυσμα εισόδου, m_i, ο νευρώνας i και h_{ci}, η συνάρτηση γειτονίας (neighborhood function). Το 'c' αντιστοιχεί στην πιο ταιριαστή μονάδα και το 'i' στον τρέχοντα νευρώνα. Χρησιμοποιούνται ποικίλες συναρτήσεις γειτονίας, των οποίων η τιμή φθίνει όσο μεγαλώνει η απόσταση του εξεταζόμενου νευρώνα m_i από την πιο ταιριαστή μονάδα m_c. Αναφέρουμε το παράδειγμα της Gaussian, με τύπο 23:

$$h_{ci}(t) = \alpha(t). \exp(- || r_i – r_c||^2 / 2\sigma(t)^2 \qquad (23)$$

Αναφέρουμε επίσης τη συνάρτηση γειτονίας Bubble, που είναι μια σταθερή συνάρτηση και που ενημερώνει κατά την ίδια τιμή την πιο ταιριαστή μονάδα και τους γειτονικούς νευρώνες.

Οι αποστάσεις μεταξύ νευρώνων δύναται να απεικονιστούν υπό τη μορφή αλλαγής χρώματος μεταξύ γειτονικών νευρώνων. Η αναπαράσταση ενός SOM με τον τρόπο αυτό, αναφέρεται ως *U-matrix* (Unified distance matrix) [Ults93] (Σχήμα 4.11). Το σκούρο χρώμα χρησιμοποιείται για το συμβολισμό μεγάλων αποστάσεων μεταξύ δεδομένων ενώ το ανοιχτό χρώμα, για τις κοντινές αποστάσεις. Οι περιοχές ανοιχτού χρώματος σχηματίζουν κατηγορίες, ενώ οι σκούρες περιοχές αντιστοιχούν στα όρια των κατηγοριών αυτών. Η μέθοδος U-matrix είναι πολύ χρήσιμη κατά την ομαδοποίηση δεδομένων σε ομάδες που ο αριθμός τους δεν είναι εκ των προτέρων γνωστός.



Σχήμα 4.11 U-matrix αναπαράσταση του SOM

4.4.5. Η Προσέγγιση του Benedikt

Η προσέγγιση του Benedikt [Bene91], ως τεχνική χωρικοποίησης, έχει ως αποτέλεσμα, ένα σύνολο δεδομένων, να απεικονιστεί υπό τη μορφή συνόλου αντικειμένων στον τρισδιάστατο χώρο, με συγκεκριμένη θέση και συγκεκριμένη εμφάνιση. Τα *τοπία πληροφορίας* (information landscapes) που προκύπτουν από την εφαρμογή της προσέγγισης αυτής είναι γνωστά ως *χώροι Benediktine* (Benediktine spaces). Το βασικό πλεονέκτημα της προσέγγισης αυτής, είναι ο γενικός της χαρακτήρας, που της επιτρέπει να εφαρμοστεί σε οποιουδήποτε τύπου δεδομένα.

Πιο συγκεκριμένα, σύμφωνα με την προσέγγιση του Benedikt, τα χαρακτηριστικά των δεδομένων απεικονίζονται σε *εγγενείς* (intrinsic) και *εξωγενείς* (extrinsic) διαστάσεις στο χώρο. Οι εξωγενείς διαστάσεις αντιστοιχούν σε συντεταγμένες στο χώρο απεικόνισης. Οι εγγενείς διαστάσεις αντιστοιχούν στις οπτικές μεταβλητές (σχήμα, μέγεθος, χρώμα, μοτίβο, κτλ).

Ο προσδιορισμός των εξωγενών διαστάσεων δεν πρέπει, σύμφωνα με τον Benedikt, να παραβιάζει τις αρχές του *αποκλεισμού* (Principle of Exclusion) και του *μέγιστου αποκλεισμού* (Principle of Maximum Exclusion). Η πρώτη αρχή αποκλείει δυο ξένα δεδομένα να απεικονιστούν στην ίδια θέση στο χώρο, απαγορεύοντας, στις ίδιες εξωγενείς διαστάσεις δεδομένων να είναι ίσες. Η δεύτερη αρχή, προτρέπει να οριστούν ως εξωγενείς διαστάσεις, εκείνα τα χαρακτηριστικά που ελαχιστοποιούν την πιθανότητα να παραβιαστεί η πρώτη αρχή. Ουσιαστικά, η αρχή του μέγιστου αποκλεισμού τείνει να οργανώνει τα απεικονιζόμενα

δεδομένα στο χώρο με ομοιογενή τρόπο, αποφεύγοντας το σχηματισμό συνωστισμένων ή άδειων περιοχών. Ακόμα, συνηθίζεται εκείνα τα χαρακτηριστικά που αφορούν στη θέση ή στο χρόνο να αντιστοιχίζονται άμεσα στις εξωγενείς διαστάσεις, για τη διευκόλυνση της ανθρώπινης αντίληψης.

Ο προσδιορισμός των εγγενών διαστάσεων πρέπει απαραίτητα να συνοδεύεται από πολυπληθείς επεξηγήσεις προς τον χρήστη. Πράγματι, δεν μπορεί να γίνει άμεσα αντιληπτή η λογική της αντιστοίχισης των χαρακτηριστικών στις διάφορες οπτικές μεταβλητές. Ακόμα, ενδέχεται ο κατάλογος των διαθέσιμων οπτικών μεταβλητών να εξαντληθεί, χωρίς να έχουν απεικονιστεί όλα τα χαρακτηριστικά των πολυδιάστατων δεδομένων. Υπάρχει όμως και η περίπτωση, ο χρήστης να χαθεί με τη χρήση πάρα πολλών οπτικών μεταβλητών. Ο Bertin ανέφερε μάλιστα ότι ενδείκνυται η χρήση μιας και μόνο οπτικής μεταβλητής, στην τρίτη διάσταση μιας εικόνας. Βεβαίως, ο χώρος Benediktine δεν ταυτίζεται με τη στατική εικόνα του Bertin, ωστόσο δεν μπορούμε να «βομβαρδίσουμε» το χρήστη με μια τεράστια πληθώρα οπτικών μεταβλητών. Για τη λύση του προβλήματος αυτού, ο Benedikt πρότεινε την αρχή του «*ξετυλίγματος*» (unfolding). Σύμφωνα με την αρχή του «ξετυλίγματος», μπορούν να απεικονιστούν ορισμένα χαρακτηριστικά σ' ένα δεύτερο χώρο Benediktine, ορίζοντας ένα σύνδεσμο μεταξύ του αντικειμένου και του νέου χώρου.

Ένα άλλο μειονέκτημα της προσέγγισης του Benedikt, έγκειται στο ότι ο χρήστης δεν έχει μια ολοκληρωμένη εικόνα του συνόλου των δεδομένων, αφού τα «μπροστινά» αντικείμενα κρύβουν τα «πίσω», ενώ προσανατολίζεται δύσκολα στον τρισδιάστατο καρτεσιανό χώρο. Ως απάντηση στο πρόβλημα αυτό, ο Benedikt πρότεινε τη χρήση σφαιρικών συντεταγμένων αντί των καρτεσιανών. Με τον τρόπο αυτό, τα αντικείμενα τοποθετούνται όλα στην επιφάνεια μιας σφαίρας με αποτέλεσμα ο χρήστης να έχει μια πιο εποπτική εικόνα των δεδομένων και να μπορεί πιο εύκολα να πλοηγηθεί σε αυτά.

4.5 Εξόρυξη Γνώσης από Συλλογές Πολυδιάστατων Δεδομένων Χρησιμοποιώντας το Πρωτότυπο Περιβάλλον Χωρικοποίησης GeoScape

4.5.1. Εισαγωγή

Όπως αναφέρθηκε παραπάνω, η οπτικοποίηση γνωστή ως χωρικοποίηση, επιτυγχάνει την απεικόνιση πολυδιάστατων δεδομένων σε χώρους περιορισμένων διαστάσεων, όπως είναι ο γεωγραφικός, κάνοντας χρήση ειδικών τεχνικών προβολής και μείωσης διαστάσεων, καθώς και κατάλληλων χωρικών μεταφορών. Εκτός όμως από το πρόβλημα που επιφέρει ο μεγάλος αριθμός διαστάσεων των δεδομένων, η χωρικοποίηση οφείλει να διαχειριστεί και το πρόβλημα της αναπαράστασης των δεδομένων σε πολλά *επίπεδα λεπτομέρειας* (granularity levels).

Το πρωτότυπο περιβάλλον χωρικοποίησης *GeoScape* [KKK10a] σχεδιάστηκε για να αντιμετωπίσει το διττό πρόβλημα της διαχείρισης πολλών διαστάσεων και της απεικόνισης σε πολλά επίπεδα λεπτομέρειας. Στη συνέχεια, θα εξεταστεί καταρχάς ο τρόπος που οι μέχρι τώρα προτεινόμενες εφαρμογές χωρικοποίησης πραγματοποιούν την απεικόνιση πολυδιάστατων δεδομένων σε διάφορα επίπεδα λεπτομέρειας, και στη συνέχεια θα περιγραφεί η λύση που προσφέρεται από το περιβάλλον χωρικοποίησης του GeoScape.

4.5.2. Το Υπόβαθρο της Υλοποίησης του GeoScape

Τα επίπεδα λεπτομέρειας είναι στενά συνδεδεμένα με τις αλλαγές στην κλίμακα απεικόνισης που επιφέρουν οι διαδικασίες της γενίκευσης και της ειδίκευσης. Ειδικότερα, η διαδικασία της γενίκευσης μειώνει τον αριθμό των απεικονιζόμενων αντικειμένων, με την ομαδοποίηση και την αντικατάστασή τους από αντικείμενα λιγότερα σε αριθμό και «φτωχότερα» σε λεπτομέρειες αναπαράστασης. Αντίθετα, η διαδικασία της ειδίκευσης επιφέρει τη «διαίρεση» των απεικονιζόμενων αντικειμένων σε περισσότερα και «πλουσιότερα» σε λεπτομέρειες αναπαράστασης.

Οι διαδικασίες της γενίκευσης και της ειδίκευσης μπορούν άριστα να υποστηριχθούν από κατάλληλους αλγορίθμους ιεραρχικής χωρικής ομαδοποίησης, μιας και συμβάλλουν στην αναδρομική δημιουργία μιας ιεράρχησης ομάδων βάσει κριτηρίων που σχετίζονται με τις

αποστάσεις και την ομοιότητα των απεικονιζόμενων αντικειμένων. Πράγματι, η ιεράρχηση αυτή μπορεί άριστα να προσφερθεί ως βάση για τη μοντελοποίηση των διαδοχικών συμπτύξεων και διαιρέσεων αντικειμένων ή ομάδων αντικειμένων, καθόσον καθένα από τα επίπεδά του υποδεικνύει τα αντικείμενα που πρέπει να συμμετέχουν σε συγκεκριμένη κλίμακα ή επίπεδο λεπτομέρειας μιας οπτικοποίησης.

Τα αποτελέσματα της ιεραρχικής ομαδοποίησης απεικονίζονται συνήθως με τη βοήθεια μιας δομής γνωστής ως *δενδρόγραμμα*. Το δενδρόγραμμα καθορίζει μια ιεράρχηση ομάδων που ξεκινά από το επίπεδο στο οποίο υπάρχει *μια και μοναδική ομάδα* (all-inclusive cluster) που περιλαμβάνει όλα τα αντικείμενα της οπτικοποίησης και τελειώνει στο επίπεδο όπου υπάρχουν τόσες ομάδες όσα είναι και τα αντικείμενα, δηλαδή στο επίπεδο όπου δεν υφίσταται καμία ομαδοποίηση. Ωστόσο, ένα δενδρόγραμμα μπορεί να αποκτήσει τέτοια έκταση που το τελευταίο επίπεδο (το επίπεδο των φύλλων της ιεράρχησης) να γίνει υπερβολικά «υπερφορτωμένο» [Dems06] ενώ ταυτόχρονα μια μεγάλη έκταση του χώρου αναπαράστασης να παραμείνει κενή και ανεκμετάλλευτη [CMS99b]. Ως εκ τούτου, ως αποκλειστική μέθοδος απεικόνισης δεδομένων σε διάφορα επίπεδα λεπτομέρειας, το δενδρόγραμμα παρουσιάσει αρκετά μειονεκτήματα, τα οποία αυξάνουν τόσο πολύ με τον όγκο των δεδομένων που μπορεί να προκληθεί πρόβλημα αναγνωσιμότητας.

Στο πλαίσιο μιας χωρικοποίησης, το πρόβλημα της απεικόνισης δεδομένων σε πολλά επίπεδα λεπτομέρειας αντιμετωπίζεται συνήθως με τη χρήση ιεραρχικής δομής χωρικών μεταφορών. Οι Kuhn και Blumenthal [KB96] υποστηρίζουν ότι η χρήση μιας χωρικής μεταφοράς επιφέρει αυτομάτως και τη χρήση των σχετικών *υπερ-μεταφορών* (super-metaphors) και *υπο-μεταφορών* (sub-metaphors). Για παράδειγμα, οι χωρικές μεταφορές που παρέχουν η χρήση των εννοιών του *κτηρίου* και της *χώρας* μπορεί να θεωρηθούν ότι αποτελούν αντίστοιχα υπο-μεταφορά και υπερ-μεταφορά της χωρικής μεταφοράς που παρέχει η έννοια της *πόλης*. Με τον τρόπο αυτό, μια ολόκληρη ιεραρχία μεταφορών μπορεί να χρησιμοποιηθεί για να υποστηρίξει την απεικόνιση από το χώρο των δεδομένων στο χώρο της χωρικοποίησης σε διάφορα επίπεδα λεπτομέρειας.

Στην έρευνα που παρουσιάζεται στο [DF98], η μεταφορά της πόλης εξετάζεται διεξοδικά. Υποστηρίζεται ότι παρέχει αφενός τις απαραίτητες ιεραρχικές και δυναμικές δομές για την προσαρμογή της απεικόνισης στην εκάστοτε κλίμακα ή επίπεδο λεπτομέρειας και αφετέρου

τη δυνατότητα ορισμού ξεκάθαρων χωρικών ορίων μεταξύ των απεικονιζόμενων αντικειμένων (π.χ. ένα κτήριο ή ένα δωμάτιο έχουν ξεκάθαρα όρια) και *σχέσεων περιέχων/περιεχόμενο* (relationships of containment).

Ακόμη, οι ιδιότητες της μεταφοράς της πόλης αξιοποιούνται στο [DCH+03] για την οπτικοποίηση μεγάλου όγκου πολυμεσικών δεδομένων. Η μεταφορά περιγράφεται ως πολύ εύκολα μνημονεύσιμη και τόσο πλούσια σε λεπτομέρειες που μπορεί να μεταδώσει πληροφορίες με παρά πολλούς τρόπους. Ένα άλλο παράδειγμα εφαρμογής που χρησιμοποιεί τη μεταφορά της πόλης αποτελεί η εφαρμογή CodeCity [WL08], η οποία πραγματοποιεί τρισδιάστατη χωρικοποίηση δεδομένων τα οποία αντιστοιχούν σε περιγραφές πολυσύνθετων συστημάτων λογισμικού. Στο CodeCity, οι μεταφορές της *συνοικίας* (district) και του *κτηρίου* (building) χρησιμοποιούνται ως υπο-μεταφορές της *πόλης* για την αναπαράσταση των πακέτων λογισμικού και των κλάσεων (classes) που υλοποιούνται στα πακέτα αυτά.

Άλλο ένα παράδειγμα χωρικής μεταφοράς είναι αυτό του τοπίου πληροφοριών. Τα τοπία πληροφοριών κάνουν το χώρο αναπαράστασης εύκολα αντιληπτό διότι μοιάζει με τον οικείο γεωγραφικό χώρο [FB01]. «*Τα τοπία, όπως αυτά απαντώνται στη φύση*» αποτελούν, σύμφωνα με τους Benking and Judge [BJ94], ιδανικές μεταφορές για τη μοντελοποίηση, την αναπαράσταση και την εξερεύνηση πληροφοριών σε διάφορες κλίμακες, δηλαδή σε διάφορα επίπεδα λεπτομέρειας.

Συνήθως, στη βιβλιογραφία, οι τεχνικές χωρικοποίησης που χρησιμοποιούν τη μεταφορά του τοπίου πληροφοριών επεξεργάζονται δεδομένα κειμένου (ελεύθερο κείμενο, επιστημονικά άρθρα, βιβλία, ιστοσελίδες, κλπ). Ενδεικτικά παραδείγματα είναι οι εφαρμογές SPIRE [Wise99] και VxInsight [BWD02]. Η πρώτη πραγματοποιεί τη χωρικοποίηση συλλογών εγγράφων, αλλά δεν παρέχει καμία λειτουργία αλλαγής της κλίμακας απεικόνισης ή του επιπέδου λεπτομέρειας. Η εφαρμογή VxInsight διαχειρίζεται κι αυτή δεδομένα κειμένου όπως έγγραφα, διπλώματα ευρεσιτεχνίας, ενώ παρέχει λειτουργίες εστίασης (zooming), για την πιο λεπτομερή παρατήρηση ενός μέρους του τοπίου.

Ένα άλλο παράδειγμα εφαρμογής που χρησιμοποιεί τη μεταφορά του τοπίου πληροφοριών περιγράφεται στο [ZCY08], όπου ακριβέστερα, προτείνεται η χρήση της μεταφοράς του τοπίου νήσων για την οργάνωση εικόνων που ανακτήθηκαν από το διαδίκτυο

ή από διάφορες συλλογές φωτογραφιών, σε ένα δισδιάστατο χώρο αναπαράστασης. Τα νησιά του τοπίου, τα λεγόμενα «*οπτικά νησιά*» (visual islands), αντιστοιχούν σε ομάδες όμοιων εικόνων ως προς κάποια χαρακτηριστικά όπως ο χρόνος λήψης, το χρώμα, η υφή, κλπ. Τέλος, στην έρευνα που παρουσιάζεται στο [MDP08], προτείνεται ένα νέο γραφικό μοντέλο αναπαράστασης που βασίζεται στη μεταφορά του τοπίου, το οποίο ελέγχει και διαχειρίζεται με οπτικό τρόπο τις διαδικασίες ανάπτυξης μεγάλων έργων λογισμικού.

Όσον αφορά τις τεχνικές χωρικοποίησης SOM, MDS και PCA πετυχαίνουν μεν την επιθυμητή μείωση των διαστάσεων αλλά δεν μπορούν να αναπαραστήσουν τα δεδομένα σε επίπεδα λεπτομέρειας χωρίς την υιοθέτηση άλλων τεχνικών οπτικοποίησης. Για παράδειγμα η εφαρμογή WEBSOM [HKLK97], [KKL+00], σχεδιασμένη για τη χωρικοποίηση συλλογών εγγράφων μεγάλου όγκου με τη βοήθεια ενός SOM, ενσωματώνει μια βοηθητική διαδραστική λειτουργία που δίνει τη δυνατότητα στους χρήστες να εξερευνήσουν το χώρο της απεικόνισης σε δύο επίπεδα λεπτομέρειας. Ακόμη, ο Skupin [Skup01] έδειξε πως η ανάλυση του SOM, δηλαδή ο αριθμός των νευρώνων ή μονάδων του SOM, επηρεάζει το επίπεδο λεπτομέρειας της παραγόμενης χωρικοποίησης. Απέδειξε μάλιστα ότι όσο μεγαλύτερη είναι η ανάλυση ενός SOM, τόσο πιο λεπτομερής είναι και η χωρικοποίηση που παράγεται και αντίστροφα. Επιπλέον, ο ίδιος [Skup02] εξέτασε τον τρόπο κατά τον οποίο μια συλλογή από περιλήψεις συνεδρίου μπορούν να οπτικοποιηθούν στο πλαίσιο μιας χωρικοποίησης με SOM σε διάφορα επίπεδα λεπτομέρειας κάνοντας χρήση ειδικών αλγορίθμων ιεραρχικής ομαδοποίησης.

Στην περίπτωση των τεχνικών χωρικοποίησης PCA and MDS, η διακριτή φύση των δεδομένων προς απεικόνιση διατηρείται, με αποτέλεσμα η χωρικοποίηση να οδηγεί σε κατανομές σημειακών αντικείμενων. Ωστόσο, καμία από τις δύο τεχνικές δεν μπορεί να αναπαραστήσει τα αποτελέσματα μιας χωρικοποίησης σε πολλά επίπεδα λεπτομέρειας χωρίς να γίνει χρήση και άλλων οπτικών μεταβλητών πέραν της θέσης των σημείων. Για παράδειγμα, οι Skupin and Buttenfield [SB97] πραγματοποίησαν τη χωρικοποίηση κειμένων χρησιμοποιώντας την τεχνική MDS σε συνδυασμό με τη χωρική μεταφορά του χάρτη. Κατόρθωσαν να προβάλλουν τα κείμενα σε ένα δισδιάστατο χώρο βάσει της ομοιότητας του περιεχομένου τους ενώ ταυτόχρονα τα ομαδοποίησαν ιεραρχικά σε δενδρόγραμμα. Σε κάθε αλλαγή του τρέχοντος ιεραρχικού επιπέδου στο δενδρόγραμμα, τα σημειακά αντικείμενα

συμπτύσσονταν ή διαχωρίζονταν σε επιμέρους, με ταυτόχρονη αλλαγή του χαρτογραφικού συμβόλου τους.

Συνεπώς, όπως φάνηκε από τα παραπάνω παραδείγματα, οι κατανομές σημείων που προκύπτουν από τεχνικές χωρικοποίησης, όπως η PCA ή η MDS, χρειάζονται περαιτέρω επεξεργασία για να μπορέσουν να αναπαραστήσουν πληροφορίες σε διάφορα επίπεδα λεπτομέρειας. Στην επόμενη ενότητα, εισάγεται και περιγράφεται το πρωτότυπο περιβάλλον χωρικοποίησης GeoScape το οποίο προσπαθεί να καλύψει αυτήν ακριβώς την ανάγκη.

4.5.3. Η Τεχνική Χωρικοποίησης του GeoScape

Το πρωτότυπο περιβάλλον χωρικοποίησης GeoScape έχει ως στόχο την εξόρυξη γνώσης από συλλογές πολυδιάστατων δεδομένων μεγάλου όγκου. Το GeoScape χρησιμοποιεί μια τεχνική χωρικοποίησης [KTKK10] η οποία βασίζεται στην *εκτίμηση πυκνότητας με χρήση συνάρτησης πυρήνα* (kernel density estimation) η οποία εφαρμόζεται σε δισδιάστατες κατανομές σημείων που προκύπτουν από τεχνικές μείωσης των διαστάσεων των δεδομένων. Πιο συγκεκριμένα, το GeoScape παράγει τρισδιάστατες χωρικοποιήσεις κάνοντας χρήση 1) της μεταφοράς του τοπίου πληροφοριών για την ανάδειξη ομάδων όμοιων δεδομένων και 2) της μεταφοράς της ομαλότητας του τοπίου πληροφοριών για την παραγωγή οπτικοποιήσεων σε διάφορα επίπεδα λεπτομέρειας. Από τις εφαρμογές χωρικοποίησης που χρησιμοποιούν τη μεταφορά του τοπίου, το GeoScape διαφέρει στα ακόλουθα σημεία:

- Το GeoScape δεν δέχεται κείμενα (έγγραφα, σελίδες διαδικτύου, κλπ) ως δεδομένα εισόδου αλλά επεξεργάζεται πολυδιάστατα δεδομένα με τη μοναδική προϋπόθεση να βρίσκονται οργανωμένα σε πίνακες στους οποίους οι γραμμές να αντιστοιχούν στις διακριτές περιπτώσεις δεδομένων και οι στήλες στις τιμές συγκεκριμένων χαρακτηριστικών (ή μεταβλητών) βάσει των οποίων γίνονται οι παρατηρήσεις.
- Σε αντίθεση με τα υπάρχοντα περιβάλλοντα χωρικοποίησης που χρησιμοποιούν τη μεταφορά του τοπίου πληροφοριών, το GeoScape χρησιμοποιεί επιπλέον τη μεταφορά της ομαλότητας του τοπίου για να διαμορφώσει το τοπίο σύμφωνα με το επιθυμητό επίπεδο λεπτομέρειας, πετυχαίνοντας ταυτόχρονα ιεραρχική ομαδοποίηση των δεδομένων. Έτσι, ενώ στο κατώτερο ιεραρχικό επίπεδο, το τοπίο

μοιάζει σαν φτιαγμένο από βουνά τόσο στενά που θυμίζουν κυπαρίσσια, όσο ανεβαίνει το ιεραρχικό επίπεδο της ομαδοποίησης (δηλαδή μειώνεται το επίπεδο λεπτομέρειας της χωρικοποίησης), τα βουνά συμπτύσσονται σε μεγαλύτερα και ομαλότερα ενώ η μορφή τους, στο ανώτερο ιεραρχικό επίπεδο σχηματίζει τελικά ένα και μοναδικό βουνό με πολύ απαλή κλίση, σαν «λόφος».

- Το GeoScape εφαρμόζει τη μέθοδο της εκτίμησης πυκνότητας με χρήση της τριγωνικής συνάρτησης πυρήνα, η οποία μοντελοποιεί την ομοιότητα μεταξύ των δεδομένων που απεικονίζονται στη χωρικοποίηση και η οποία θα περιγραφεί λεπτομερώς στη συνέχεια.
- Εκτός από την οπτικοποίηση των δεδομένων υπό τη μορφή τοπίου πληροφοριών, το GeoScape παρέχει παράλληλα μια άλλη μορφή οπτικοποίησης, τη λεγόμενη «όψη δενδρογράμματος» (dendrogram view), που προβάλλει ολόκληρη την ιεράρχηση των ομάδων δεδομένων και υποδεικνύει το τρέχον ιεραρχικό επίπεδο ή επίπεδο λεπτομέρειας της χωρικοποίησης.
- Το GeoScape παρέχει ακόμη τη δυνατότητα δυναμικής προσαρμογής των επιγραφών της χωρικοποίησης ώστε να διευκολυνθεί ο χρήστης στην επιλογή των πιο αντιπροσωπευτικών εξ αυτών, ανάλογα με το τι ακριβώς αναζητά ανάμεσα στα δεδομένα.

Το GeoScape πραγματοποιεί τη χωρικοποίηση των δεδομένων εισόδου σε δύο βήματα. Ξεκινά με την εφαρμογή μιας τεχνικής μείωσης των διαστάσεων των δεδομένων ώστε αυτά να μπορούν να προβληθούν ως κατανομή σημείων σε ένα δισδιάστατο χώρο αναπαράστασης. Στην τρέχουσα έκδοση του GeoScape, υποστηρίζεται για το σκοπό αυτό η κλασική τεχνική MDS αλλά έχει δημιουργηθεί υποδομή για την επέκταση του λογισμικού, ώστε να υποστηρίζει μελλοντικά οποιαδήποτε άλλη τεχνική μείωσης των διατάσεων.

Σε ένα δεύτερο βήμα, το GeoScape χρησιμοποιεί μια ειδική συνάρτηση πυρήνα που ομαδοποιεί τα σημεία της κατανομής, με δύο τρόπους:

- Πρώτον, *βάσει πυκνότητας*, όπως περιγράφηκε στην παράγραφο 4.3.3, αποβλέποντας στη μεγιστοποίηση της πυκνότητας εντός των ομάδων και στην ελαχιστοποίηση της πυκνότητας εκτός των ομάδων, δηλαδή στη μεγιστοποίηση

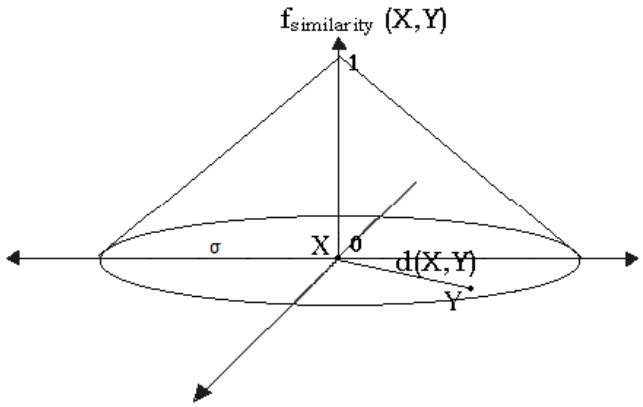
της *ομοιότητας εντός των ομάδων* (intracluster similarity) και στην ελαχιστοποίηση της *ομοιότητας μεταξύ των ομάδων* (intercluster similarity).

- Δεύτερον, με την απόδοση διαδοχικών τιμών στην ειδική παράμετρο «σ» της συνάρτησης πυρήνα, επιτυγχάνεται *ιεραρχική ομαδοποίηση* των σημείων της κατανομής. Έτσι, ξεκινώντας από την κατάσταση στην οποία όλα τα σημεία της κατανομής αντιστοιχούν σε διαφορετικές ομάδες, διαδοχικά γίνεται η ομαδοποίησή τους σε μία και μοναδική ομάδα.

Η ειδική συνάρτηση πυρήνα, $f_{similarity}$, που χρησιμοποιείται είναι τριγωνική (ονομάζεται έτσι από τη μορφή της γραφικής της παράστασης (Σχήμα 4.12)) και έχει ως τύπο τον παρακάτω:

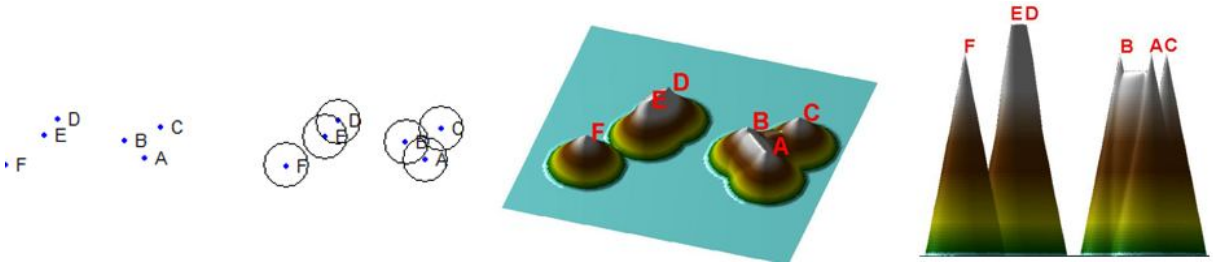
$$f_{similarity}(x,y) = 1-(1/\sigma)d(x,y), 0 \le d(x,y) < \sigma \qquad (24)$$
$$f_{similarity}(x,y) = 0, \sigma \le d(x,y)$$

Η $f_{similarity}$ μεταβάλλεται με την τιμή της ομοιότητας μεταξύ δύο σημείων στις θέσεις x και y στο δισδιάστατο χώρο αναπαράστασης και φθίνει μέχρις ότου φτάσει το κατώφλι που καθορίζεται από την παράμετρο σ. Στην ουσία το σ αντιστοιχεί στο εύρος ζώνης της συνάρτησης πυρήνα. Η απόσταση μεταξύ δύο σημείων d(x, y), παίρνει τιμές από 0 έως 1, όπου η τιμή 0 υποδηλώνει τελείως ανόμοια δεδομένα ενώ το 1 υποδηλώνει πανομοιότυπα δεδομένα. Το άθροισμα των επιμέρους συναρτήσεων πυρήνα συνθέτει τη συνολική συνάρτηση πυκνότητας, της οποίας η γραφική παράσταση παράγει την επιφάνεια του τοπίου πληροφοριών. Οι ομάδες όμοιων δεδομένων εντοπίζονται από τα τοπικά μέγιστα της επιφάνειας.



Σχήμα 4.12 Γραφική παράσταση της συνάρτησης πυρήνα $f_{similarity}$ [KTKK10]

Στο σχήμα 4.13 περιγράφεται με τη βοήθεια ενός παραδείγματος με λίγα δεδομένα εισόδου (Σχήμα 4.13(α)), ο τρόπος που σχηματίζεται το τοπίο πληροφοριών πάνω από την δισδιάστατη κατανομή σημείων. Τα όρια των συναρτήσεων πυρήνα που εφαρμόζονται σε κάθε σημείο φαίνονται στο σχήμα 4.13(β) ως κύκλοι με ακτίνα ίση προς σ. Το σχήμα 4.13(γ) δείχνει το τοπίο, το οποίο διαμορφώνεται πάνω από τα σημεία και το οποίο αντιστοιχεί στη γραφική παράσταση της συνολικής συνάρτησης πυκνότητας. Στο σχήμα 4.13(δ) φαίνεται μια διαφορετική προοπτική του τοπίου, προκειμένου να αναδειχθεί η μεταβολή του υψομέτρου σε σχέση με την πυκνότητα των σημείων. Όσο πυκνότερη είναι η κατανομή των σημείων, τόσο μεγαλύτερο είναι το υψόμετρο του ανάγλυφου.



(α) (β) (γ) (δ)

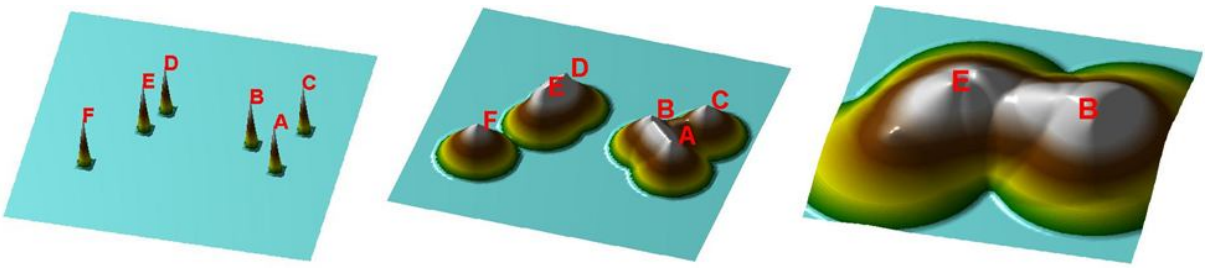
Σχήμα 4.13 Σχηματισμός τοπίου πληροφοριών [KTKK10]

Η παράμετρος σ επηρεάζει την ομαλότητα της επιφάνειας πυκνότητας και, κατ' επέκταση, τον αριθμό των απεικονιζόμενων ομάδων. Με την απόδοση διαδοχικών τιμών στο σ, πραγματοποιείται ιεραρχική ομαδοποίηση, με αποτέλεσμα το σχηματισμό ιεραρχίας εμφωλευμένων ομάδων. Ειδικότερα, τιμές του σ μικρότερες από το μισό της ελάχιστης απόστασης μεταξύ των σημείων της κατανομής, οδηγούν στην παραγωγή του χαμηλότερου

επιπέδου της ιεραρχίας ομάδων, όπου κάθε σημείο αντιστοιχεί σε ακριβώς μία ομάδα. Με τη διαδοχική αύξηση του σ, οι ομάδες συγχωνεύονται σε μεγαλύτερες, δημιουργώντας έτσι το επόμενο ιεραρχικό επίπεδο. Το υψηλότερο επίπεδο δημιουργείται όταν όλα τα σημεία ομαδοποιούνται στην ίδια και μοναδική ομάδα. Τα ιεραρχικά επίπεδα της ομαδοποίησης καθορίζουν και τα επίπεδα λεπτομέρειας της χωρικοποίησης.

Για να περιγραφεί καλύτερα η επιρροή του σ στο τοπίο πληροφοριών του σχήματος 4.13, έχουν αποδοθεί διάφορες τιμές στην παράμετρο αυτή, οι οποίες οδηγούν στο σχηματισμό τοπίων διαφορετικής ομαλότητας. Το d(E, D) είναι η μικρότερη μεταξύ των αποστάσεων των σημείων της κατανομής. Αν αποδοθούν στο σ τιμές μικρότερες από d(E, D)/2, τότε η παραγόμενη επιφάνεια θα αντιστοιχεί στο χαμηλότερο επίπεδο λεπτομέρειας (Σχήμα 4.14(α)). Στην περίπτωση αυτή, το γεγονός ότι υπάρχει μία ομάδα ανά σημείο κάνει το τοπίο να εμφανίζεται σαν ένα τοπίο με πολύ στενά βουνά, σαν «κυπαρίσσια». Καθώς το σ αυξάνεται, οι επιμέρους συναρτήσεις πυρήνα συγχωνεύονται, με αποτέλεσμα τα βουνά να ενώνονται σε μεγαλύτερα και ομαλότερα.

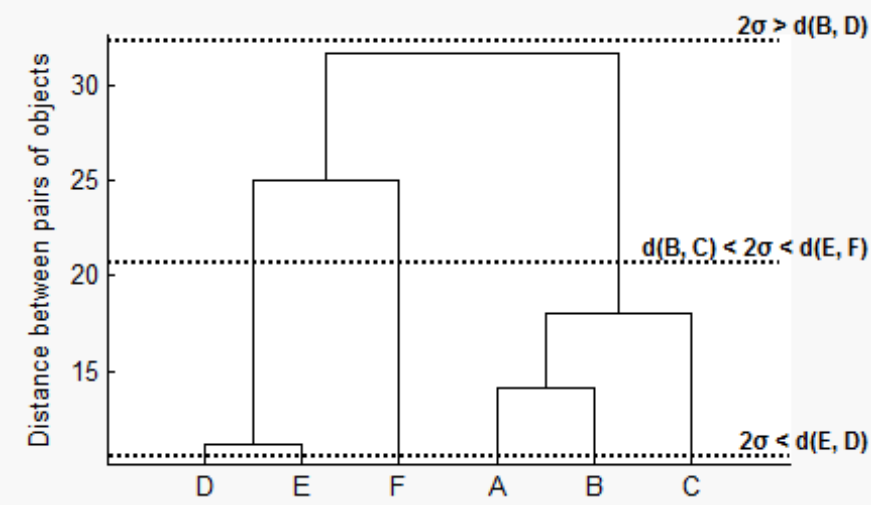
Το σχήμα 4.14(β) παρουσιάζει ένα ενδιάμεσο επίπεδο λεπτομέρειας, όπου στην παράμετρο σ έχει αποδοθεί μια τιμή μεγαλύτερη από d(B, C)/2, αλλά μικρότερη από d(E, F)/2. Τέλος, τιμές του σ ψηλότερες από d(B, D)/2 παράγουν την επιφάνεια πυκνότητας που αντιστοιχεί στο υψηλότερο επίπεδο λεπτομέρειας, όπως απεικονίζεται στο σχήμα 4.14(γ). Το τοπίο περιλαμβάνει μόνο ένα μεγάλο και ομαλό βουνό που αντιστοιχεί στην ομάδα που βρίσκεται στην κορυφή της ιεραρχίας της ομαδοποίησης. Παρατηρείται ότι όσο μεγαλύτερες είναι οι τιμές που αποδίδονται στο σ, τόσο χαμηλότερο είναι το επίπεδο λεπτομέρειας της χωρικοποίησης που προκύπτει.



(α) (β) (γ)

Σχήμα 4.14 Επιρροή της παραμέτρου σ της συνάρτησης πυρήνα $f_{similarity}$ στην ομαλότητα του τοπίου πληροφοριών [KTKK10]

Τα αποτελέσματα της χωρικοποίησης μπορούν να συγκριθούν με το δενδρόγραμμα που απεικονίζεται στο σχήμα 4.15, το οποίο προκύπτει από την ιεραρχική ομαδοποίηση των σημείων, και πιο συγκεκριμένα από την συσσωρευτική ιεραρχική ομαδοποίηση με το κριτήριο σύνδεσης του απλού συνδέσμου, των ίδιων δεδομένων. Το χαμηλότερο επίπεδο του δενδρογράμματος ορίζει 6 ομάδες και αντιστοιχεί στο τοπίο πληροφοριών που φαίνεται στο σχήμα 4.14(α). Σ' ένα ενδιάμεσο επίπεδο, όπως αυτό που απεικονίζεται στο σχήμα 4.14(β), ορίζονται 3 ομάδες δεδομένων. Το τελευταίο επίπεδο του δενδρογράμματος περιλαμβάνει μόνο μία ομάδα, όπως φαίνεται στο σχήμα 4.14(γ). Η περαιτέρω αύξηση της τιμής του σ θα είχε ως μοναδικό αποτέλεσμα να καταστήσει ομαλότερη την επιφάνεια του τοπίου χωρίς να επηρεάσει τον αριθμό των ομάδων ούτε να προσθέσει επιπλέον πληροφορία σχετικά με τα δεδομένα και, επομένως, δεν παρουσιάζει κανένα ενδιαφέρον από πλευράς χωρικοποίησης.



Σχήμα 4.15 Μεταβολή της παραμέτρου σ σε σχέση με το επίπεδο λεπτομέρειας [ΚΤΚΚ10]

Στο GeoScape, οι επιγραφές τοποθετούνται στα πιο σημαντικά (προεξέχοντα) τοπικά μέγιστα και τους αποδίδονται ονόματα σύμφωνα με μια *δυναμική στρατηγική απόδοσης επιγραφών* (dynamic labeling strategy). Ως εκ τούτου, εναπόκειται στον χρήστη να επιλέξει ποιες ή ποια από τις μεταβλητές θα είναι οι περισσότερο αντιπροσωπευτικές μεταξύ των πολλών που περιγράφουν τις παρατηρήσεις (τα δεδομένα). Στη συνέχεια, σε κάθε επιγραφή αποδίδεται ένα σύνθετο όνομα που συνοψίζει κατά κάποιον τρόπο τις τιμές των μεταβλητών των δεδομένων μιας ομάδας. Οι τύποι των μεταβλητών που υποστηρίζονται από την παρούσα έκδοση του GeoScape είναι οι εξής:

- Οι *ποσοτικές* (quantitative) μεταβλητές, οι οποίες προσθέτουν στην επιγραφή της ομάδας το συνθετικό εκείνο που αντιστοιχεί στο μέσο όρο των τιμών τους.
- Οι μεταβλητές *ταξινόμησης* (ordinal), οι οποίες προσθέτουν στην επιγραφή της ομάδας το συνθετικό εκείνο που αντιστοιχεί στην τιμή που συναντάται συχνότερα.
- Οι μεταβλητές *κειμένου* (textual ή nominal), οι οποίες προσθέτουν στην επιγραφή της ομάδας το συνθετικό εκείνο που αντιστοιχεί στην τιμή της μεταβλητής του σημείου (της παρατήρησης ή του δεδομένου) που βρίσκεται πλησιέστερα στο τοπικό μέγιστο.

Για παράδειγμα, ας υποθέσουμε ότι γίνονται κάποιες παρατηρήσεις που αφορούν σε ορισμένες γεωγραφικές περιοχές, σχετικά με τις μεταβλητές *Έκταση* (ποσοτική), *Εδαφοκάλυψη* (ταξινόμησης), και *Όνομα* (κειμένου). Σε περίπτωση που ο χρήστης επιλέξει να εμφανίσει πληροφορίες σχετικά με όλες τις μεταβλητές, η επιγραφή κάθε ομάδας θα δημιουργηθεί από τη σύνθεση του μέσου όρου της μεταβλητής *Έκταση*, της πιο συχνής τιμής της μεταβλητής *Εδαφοκάλυψη* και του ονόματος της γεωγραφικής περιοχής που βρίσκεται πλησιέστερα στο τοπικό μέγιστο που ορίζει την ομάδα. Για παράδειγμα, η ετικέτα *1000__Forest__Dadia* αποτελεί επιγραφή της ομάδας που συγκεντρώνει τις γεωγραφικές περιοχές με μέσο όρο έκτασης 1000 (π.χ. σε m^2), από τις οποίες οι περισσότερες καλύπτονται από δάσος, και μοιάζουν με την περιοχή της Δαδιάς.

Ακόμη, πρέπει να επισημανθεί ότι νέα ονόματα αποδίδονται στις επιγραφές κάθε φορά που το επίπεδο λεπτομέρειας αλλάζει, λόγω της συγχώνευσης ομάδων σε μεγαλύτερες ή λόγω της διάσπασης ομάδων σε μικρότερες. Επιπλέον, για λόγους ευκρίνειας, το μέγεθος της γραμματοσειράς των επιγραφών αυξάνεται ανάλογα με την πυκνότητα κάθε ομάδας. Τέλος, πρέπει να αναφερθεί ότι, στο χαμηλότερο επίπεδο λεπτομέρειας, δεδομένου ότι κάθε ομάδα αντιστοιχεί σε μία και μοναδική παρατήρηση ή δεδομένο, οι επιγραφές που αποδίδονται αντιστοιχούν στις ίδιες τις τιμές των μεταβλητών.

4.5.4. Το Πρωτότυπο Περιβάλλον Χωρικοποίησης GeoScape

Το GeoScape αποτελεί ένα πρωτότυπο διαδραστικό περιβάλλον που προσφέρει στους χρήστες τη δυνατότητα: (1) να εισάγουν μια συλλογή πολυδιάστατων δεδομένων υπό τη

μορφή πίνακα, (2) να περιηγηθούν σε τοπία πληροφοριών σε διάφορα επίπεδα λεπτομέρειας, (3) να μεταβάλλουν το επίπεδο λεπτομέρειας, δηλαδή την ομαλότητα του τοπίου πληροφοριών, και (4) να έχουν πρόσβαση στα λεπτομερή χαρακτηριστικά των δεδομένων εφόσον χρειαστεί. Η αλληλεπίδραση με το πρωτότυπο περιβάλλον επιτυγχάνεται μέσω *συντονισμένων προβολών* (coordinate views). Η χρησιμότητα των συντονισμένων προβολών έγκειται στην οπτικοποίηση πολλών και διαφορετικών όψεων των δεδομένων [GPQX06]. Πιο συγκεκριμένα, διατίθενται οι προβολές: «GeoScape View», «Dendrogram View» και «Details View», οι οποίες περιγράφονται ως ακολούθως:

- Η προβολή «GeoScape View» εμφανίζει τα αποτελέσματα της χωρικοποίησης, δηλαδή το τρισδιάστατο τοπίο πληροφοριών στο οποίο η εξερεύνηση γίνεται σε συγκεκριμένο και επιλεγμένο από το χρήστη επίπεδο λεπτομέρειας. Η εξερεύνηση υποστηρίζεται από λειτουργίες *εστίασης* (zooming), *μετατόπισης* (panning) και *περιστροφής* (rotating). Η λειτουργία εστίασης συνίσταται στην αλλαγή της απόστασης παρατήρησης του τοπίου και όχι στην αλλαγή του επιπέδου λεπτομέρειας. Η λειτουργία μετατόπισης μεταθέτει την εικόνα του τοπίου πάνω και κάτω, αριστερά και δεξιά. Η περιστροφή στρέφει το τοπίο προς οποιοδήποτε κατεύθυνση. Έτσι, ο χρήστης αλλάζει το πεδίο ορατότητας και εμφανίζει τα ενίοτε κρυμμένα τμήματα του τοπίου πληροφοριών.
- Η προβολή «Dendrogram View» παρέχει μια *εποπτική εικόνα ολόκληρου του δενδρογράμματος* ομαδοποίησης των δεδομένων, σημειώνοντας ταυτόχρονα το τρέχον επίπεδο λεπτομέρειας της χωρικοποίησης που εμφανίζεται στην προβολή «GeoScape View». Έτσι, ο χρήστης γνωρίζει ανά πάσα στιγμή τη θέση του στην ιεραρχία των ομάδων και μπορεί να αποφασίσει την αλλαγή του επιπέδου λεπτομέρειας, κάνοντας χρήση της παρεχόμενης γραμμής κύλισης. Με τη γραμμή κύλισης, ο χρήστης «γενικεύει» ή «ειδικεύει» το τοπίο πληροφοριών. Ο όρος *γενίκευση* περιγράφει τη λειτουργία που αυξάνει την ομαλότητα του τοπίου και, συνεπώς, κατευθύνει το χρήστη σε χαμηλότερα επίπεδα λεπτομέρειας (που αντιστοιχούν τα υψηλότερα επίπεδα του δενδρογράμματος). Αντίθετα, ο όρος *ειδίκευση* αναφέρεται στην ακριβώς αντίστροφη διαδικασία, η οποία μειώνει την ομαλότητα του τοπίου και μεταφέρει το χρήστη σε υψηλότερα επίπεδα λεπτομέρειας (που αντιστοιχούν σε χαμηλότερα επίπεδα του δενδρογράμματος).

- Η προβολή «Details View» χρησιμοποιείται από το χρήστη για να ζητήσει περαιτέρω *περιγραφικές πληροφορίες* σχετικά με κάποια ομάδα, όπως για παράδειγμα, πληροφορίες για το ποιες είναι οι εμφωλευμένες ομάδες, οι τιμές των δεδομένων που έχουν ομαδοποιηθεί, κλπ. Η προβολή αυτή ενεργοποιείται επιλέγοντας «Details View» από το μενού που εμφανίζεται με δεξί κλικ του ποντικιού πάνω σε συγκεκριμένη ομάδα (κορυφή του τοπίου).

Στη συνέχεια, περιγράφεται ένα σενάριο εφαρμογής που έχει ως στόχο την ανάδειξη της λειτουργικότητας του πρωτότυπου περιβάλλοντος λογισμικού GeoScape και των διαφορετικών προβολών που παρέχει.

Το σενάριο χρησιμοποιεί ως δεδομένα εισόδου, μια συλλογή πολυδιάστατων δεδομένων που αφορούν σε περιοχές της Ελλάδας που προστατεύονται από το Ευρωπαϊκό Δίκτυο NATURA 2000. Τα δεδομένα περιέχουν 419 παρατηρήσεις με 5 μεταβλητές: το γεωγραφικό μήκος, το γεωγραφικό μήκος πλάτος, το ελάχιστο υψόμετρο της περιοχής, το μέσο υψόμετρο και το μέγιστο υψόμετρο. Ένα απόσπασμα των δεδομένων εισόδου παρουσιάζεται στον πίνακα 4.1 ενώ το σύνολο των δεδομένων που έχουν χρησιμοποιηθεί στο παράδειγμα παρατίθεται στο παράρτημα 2. Το σενάριο ορίζει ότι η αποστολή του χρήστη συνίσταται στην εύρεση των προστατευόμενων ελληνικών περιοχών με παρόμοια γεωγραφικά χαρακτηριστικά.

Site Name	Longitude (º ' '')			Latitude (º ' '')			Alt. Mean (m)	Alt. Max (m)	Alt. Min (m)
Dadia forest, Soufli	26	10	11	41	6	52	188,52	614	15
Treis Vryses	26	0	13	41	8	17	530,23	1034	190
Fengari, Samothraki island	25	40	57	40	27	32	629,08	1600	-50
Mountains of Evros county	26	10	19	41	6	55	185,19	614	13
Delta of Evros river	26	4	31	40	45	53	1,72	66	0
Delta and west arm of Evros river	26	4	3	40	46	21	2,21	32	0
Riverside forest of northern Evros river and Arda	26	23	50	41	37	44	59,86	332	0
Cluster of Forests of southern Evros county	25	56	57	40	57	37	235,12	848	21
Mountainous region of Evros county – Valley of Derios	26	2	6	41	12	9	401,93	1064	69
Mountain Chaidou - Koula and surrounding peaks	24	48	27	41	19	27	1262,87	1820	683
Narrows of Nestos river	24	43	48	41	7	34	456,34	1281	0
Forest of Nestos river	24	42	60	41	6	34	222,16	817	0
Filiouris river	25	34	17	41	1	36	80,37	624	5
Kompsatos river (new river bed)	25	11	35	41	9	49	95,12	217	15
Maroneia cave	25	30	13	40	55	56	142,74	160	129

Πίνακας 4.1 Ενδεικτικό υποσύνολο Ελληνικών περιοχών προστατευόμενων από το Ευρωπαϊκό Δίκτυο NATURA 2000[7]

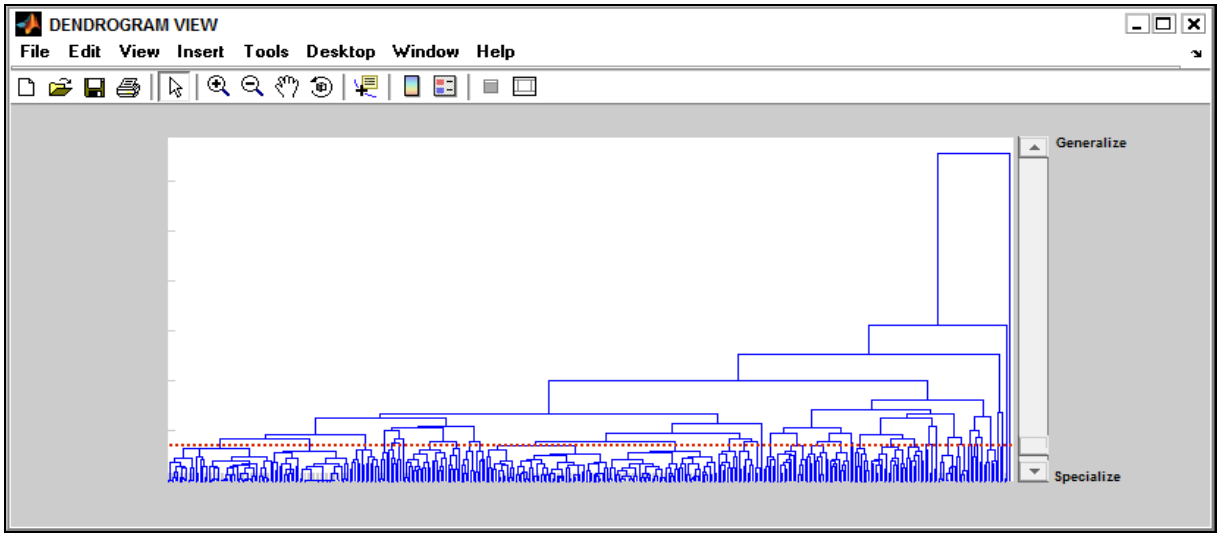
[7] Πηγή: European Environment Information and Observation Network, διαθέσιμη στο σύνδεσμο http://cdr.eionet.europa.eu/gr/eu/n2000, τελευταία προσπέλαση Οκτώβριος 2009.

Αφού εισαχθούν τα δεδομένα εισόδου, το GeoScape προχωρά στη σύγκριση των παρατηρήσεων. Για το σκοπό αυτό, χρησιμοποιείται ως μέτρο ομοιότητας για κάθε ζεύγος παρατηρήσεων, η κανονικοποιημένη Ευκλείδεια απόσταση που ορίζεται από τον τύπο 25:

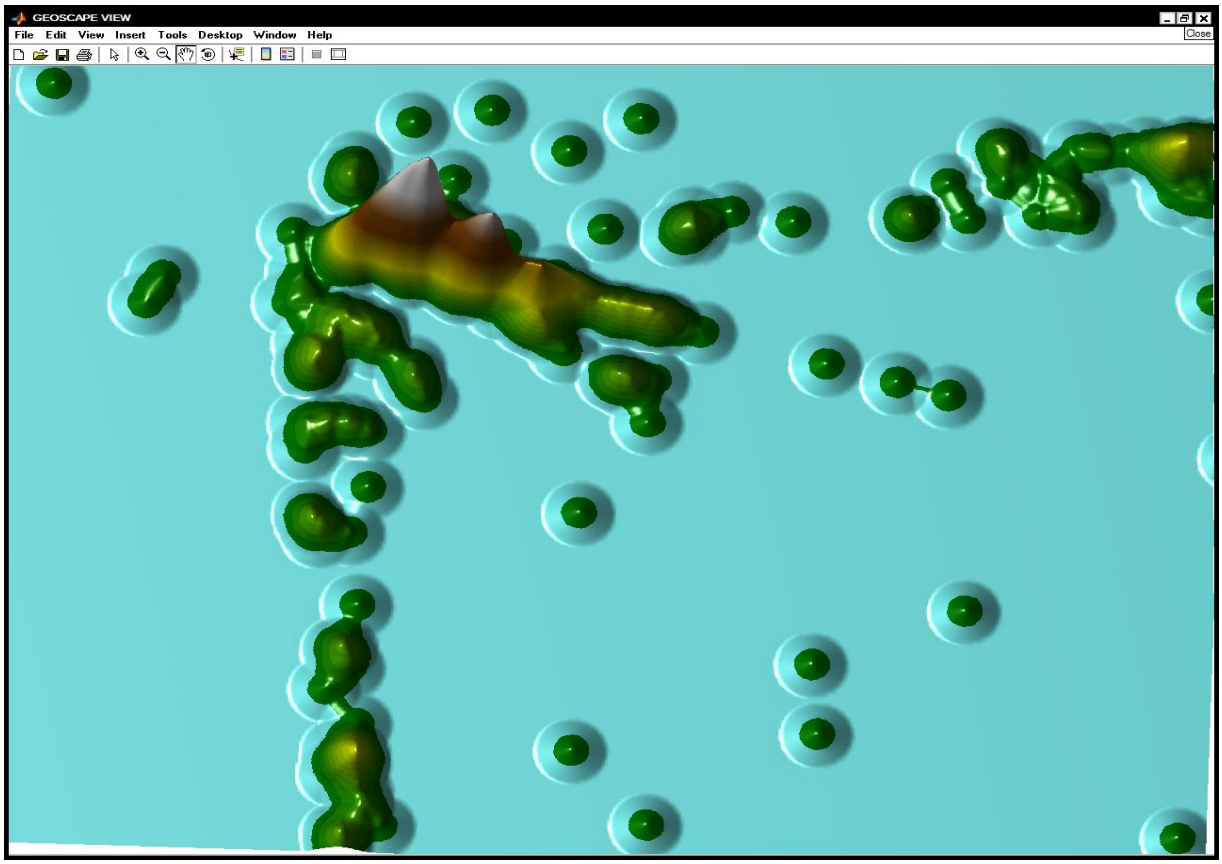
$$dist(d_i, d_j) = \sqrt{\sum_{a=1}^{n} (x_{ia} - x_{ja})^2 / s_a^2} \qquad (25)$$

όπου d_i και d_j (i, j = 1, 2, ..., 27) ζευγάρι παρατηρήσεων n διαστάσεων προς σύγκριση, x_{ia} και x_{ja}, οι τιμές των μεταβλητών των παρατηρήσεων αυτών που αναφέρονται στη διάσταση α, s_α^2, η διακύμανση των τιμών αυτών, και n = 5.

Στη συνέχεια, αφού υπολογιστούν τα μέτρα της ομοιότητας των δεδομένων, το GeoScape δημιουργεί το δενδρόγραμμα που προκύπτει βάσει της ιεραρχικής ομαδοποίησης των δεδομένων και το οποίο εμφανίζεται στην προβολή «Dendrogram View». Ταυτόχρονα, δημιουργεί το τοπίο πληροφοριών στην προβολή «GeoScape View», σε ένα προεπιλεγμένο από το σύστημα ενδιάμεσο επίπεδο λεπτομέρειας σημειωμένο με διακεκομμένη κόκκινη γραμμή στο δενδρόγραμμα του σχήματος 4.16(α). Στα σχήματα 4.16(α) και 4.16(β), παρουσιάζεται ένα στιγμιότυπο των αποτελεσμάτων.



(α)

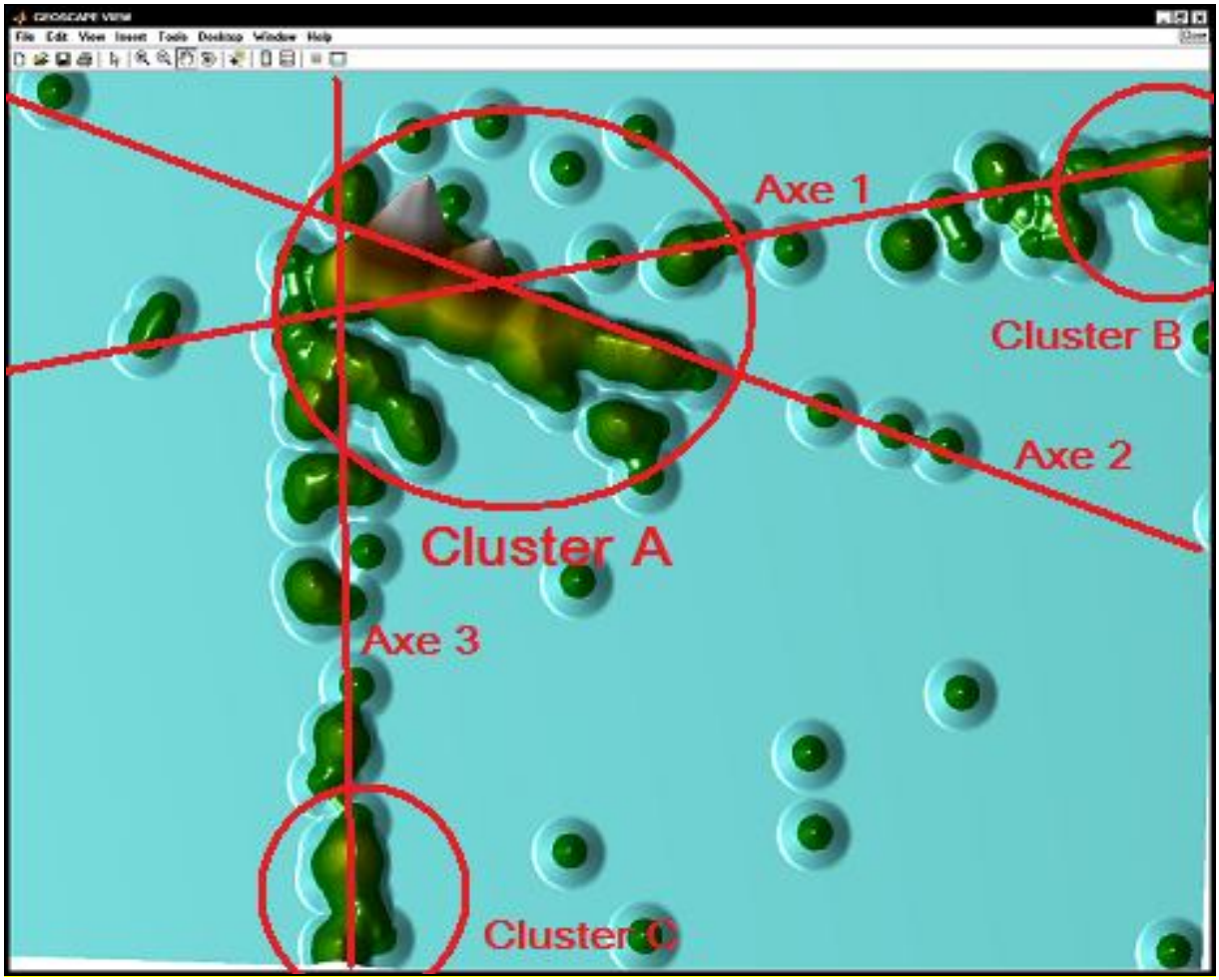


(β)

Σχήμα 4.16 Στιγμιότυπο αποτελεσμάτων που εμφανίζουν οι προβολές (α) «Dendrogram View» και (β) «GeoScape View» [KKK10a]

Σε αυτό το σημείο, πρέπει να επισημανθεί ότι δύο είδη μορφωμάτων του τοπίου πληροφοριών χρήζουν ερμηνείας από το χρήστη: 1) τα βουνά που σχηματίζονται πάνω από τις περιοχές με υψηλή πυκνότητα σημείων, δηλαδή οι ομάδες που συγκεντρώνουν τα όμοια αναμεταξύ τους δεδομένα, και 2) τα γραμμικά μοτίβα (ευθεία ή κυρτά) που εμφανίζονται ως μακρόστενα συμπλέγματα βουνών, τα οποία φανερώνουν την ύπαρξη συσχετισμού μεταξύ των τιμών των μεταβλητών κάποιων υποσύνολων δεδομένων. Σε περίπτωση που τα δεδομένα δεν συσχετίζονται μεταξύ τους, η μορφή της κατανομής των σημείων που προκύπτει μοιάζει με ακανόνιστο «σύννεφο» σημείων.

Στο Σχήμα 4.17, σημειώνονται με κόκκινο τα πιο εμφανή χαρακτηριστικά μορφώματα του τοπίου πληροφοριών. Ειδικότερα, το τοπίο αποκαλύπτει: 1) ένα ογκώδες και ψηλό ορεινό συγκρότημα, σημειωμένο μ' έναν κόκκινο κύκλο ως *Ομάδα Α* (Cluster A), 2) δύο άλλα ορεινά συγκροτήματα, επίσης σημειωμένα με κόκκινους κύκλους ως οι *Ομάδες Β* και *Γ* (Cluster B και Cluster C), και 3) τρία γραμμικά μοτίβα κατά μήκος των οποίων βρίσκονται μακρόστενα ορεινά συμπλέγματα βουνών, τα οποία σημειώνονται με άξονες κόκκινου χρώματος πάνω στο τοπίο.

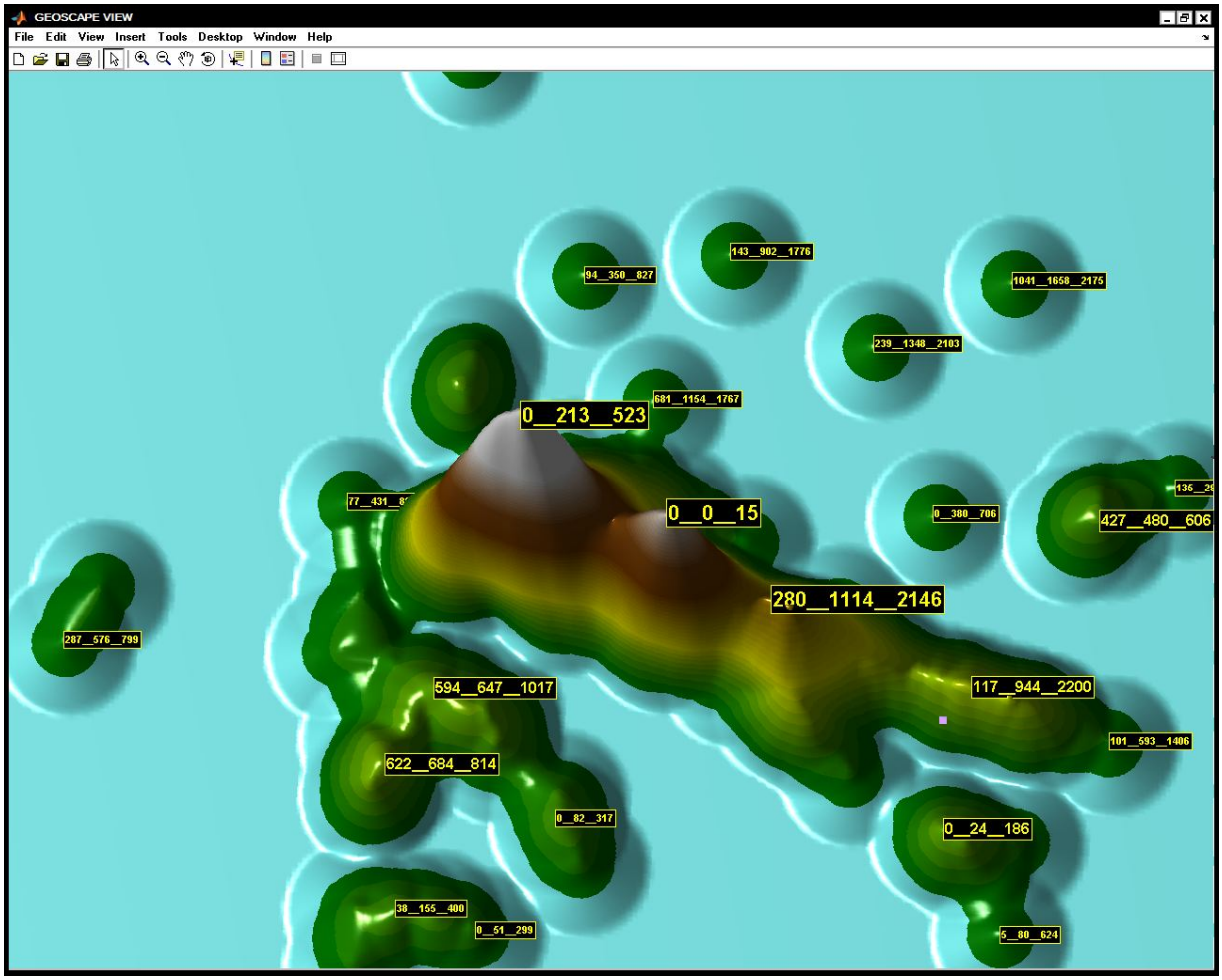


Σχήμα 4.17 Χαρακτηριστικές μορφές του τοπίου πληροφοριών του σεναρίου [KKK10a]

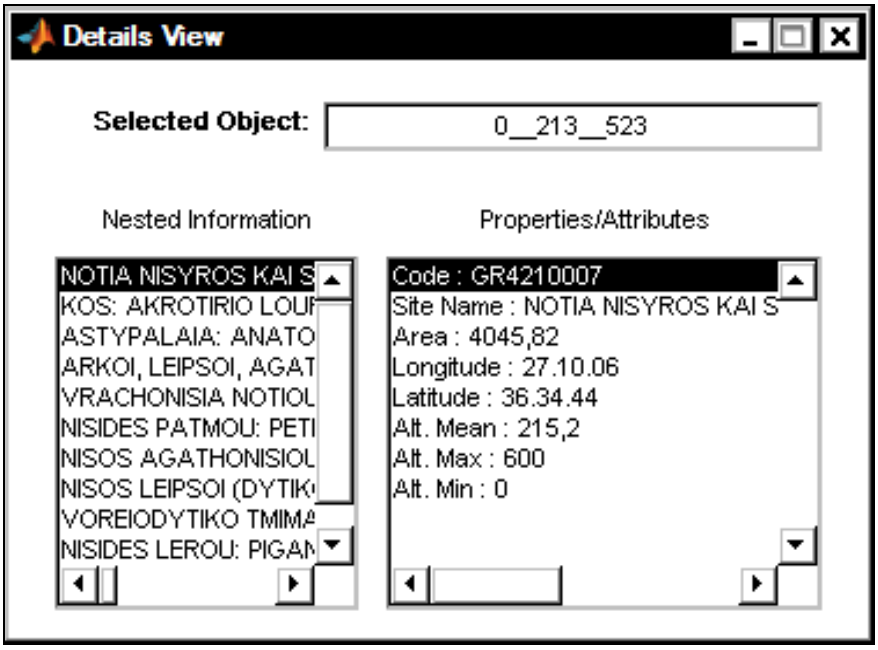
Η *ομάδα Α* φανερώνει την ύπαρξη μιας πολυπληθούς ομάδας δεδομένων με παρόμοια χαρακτηριστικά, ενώ οι *ομάδες Β* και *Γ*, φανερώνουν την ύπαρξη άλλων δύο, με μικρότερο πληθυσμό. Όπως εξηγήθηκε παραπάνω, οι άξονες (γραμμικά μοτίβα) 1, 2 και 3 δείχνουν ότι τα αντίστοιχα δεδομένα είναι τόσο συσχετισμένα που έχουν δημιουργήσει ένα συστηματικό μοτίβο, εν προκειμένου τα μακρόστενα ορεινά συγκροτήματα κατά μήκος των αξόνων.

Υποθέτοντας ότι ο χρήστης επιλέγει να μειώσει την απόσταση παρατήρησης προς την *ομάδα Α* (με zoom) και να εμφανίσει τις επιγραφές που περιέχουν τις μεταβλητές *Ελάχιστο Υψόμετρο*, *Μέσο Υψόμετρο* και *Μέγιστο Υψόμετρο* των παρατηρήσεων, η προβολή «GeoScape View» αλλάζει, όπως απεικονίζεται στο σχήμα 4.18. Από τις επιγραφές, ο χρήστης μπορεί να εξάγει το συμπέρασμα ότι η πλειοψηφία των ελληνικών περιοχών που προστατεύονται από το πρόγραμμα NATURA 2000 είναι περιοχές κοντά στην θάλασσα

επειδή έχουν μεταβλητή *Ελάχιστο Υψόμετρο* με μέσο όρο 0. Στην περίπτωση που ο χρήστης επιλέξει την κορυφή με επιγραφή *0__213__523* για να αναζητήσει μια πιο λεπτομερή περιγραφή της ομάδας, η προβολή «Details View» που απεικονίζεται στο σχήμα 4.19 ενεργοποιείται και οι περαιτέρω πληροφορίες που σχετίζονται με την επιλεγμένη ομάδα εμφανίζονται στην οθόνη. Παρομοίως, ο χρήστης μπορεί να επιλέξει να εμφανίσει πληροφορίες σχετικά με τις *ομάδες Β* και *Γ*, ή οποιεσδήποτε άλλες ομάδες (κορυφές).



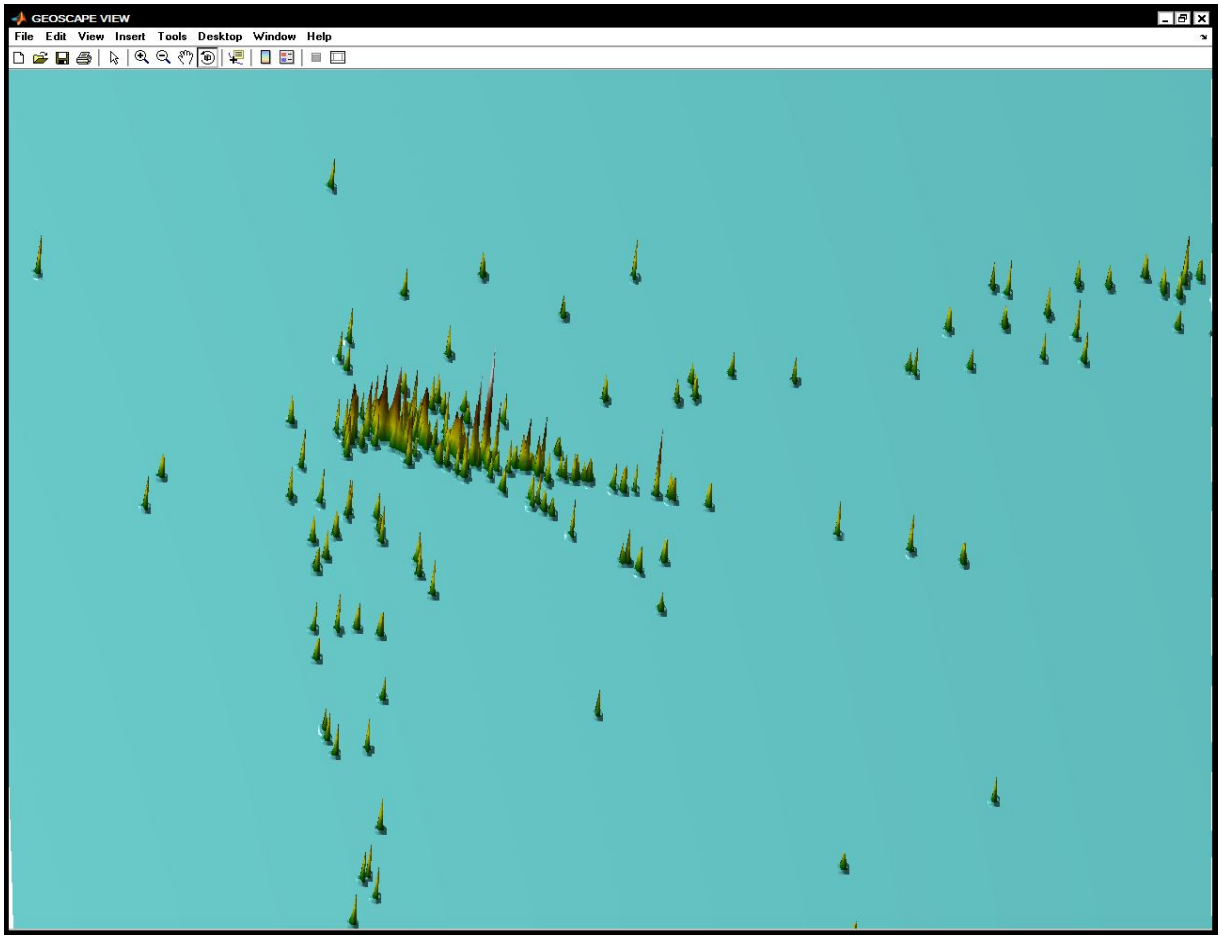
Σχήμα 4.18 Η προβολή «GeoScape View», όπως εμφανίζεται με τις επιγραφές που αντιστοιχούν στις μεταβλητές *Ελάχιστο Υψόμετρο*, *Μέσο Υψόμετρο* και *Μέγιστο Υψόμετρο* των παρατηρήσεων [KKK10a]



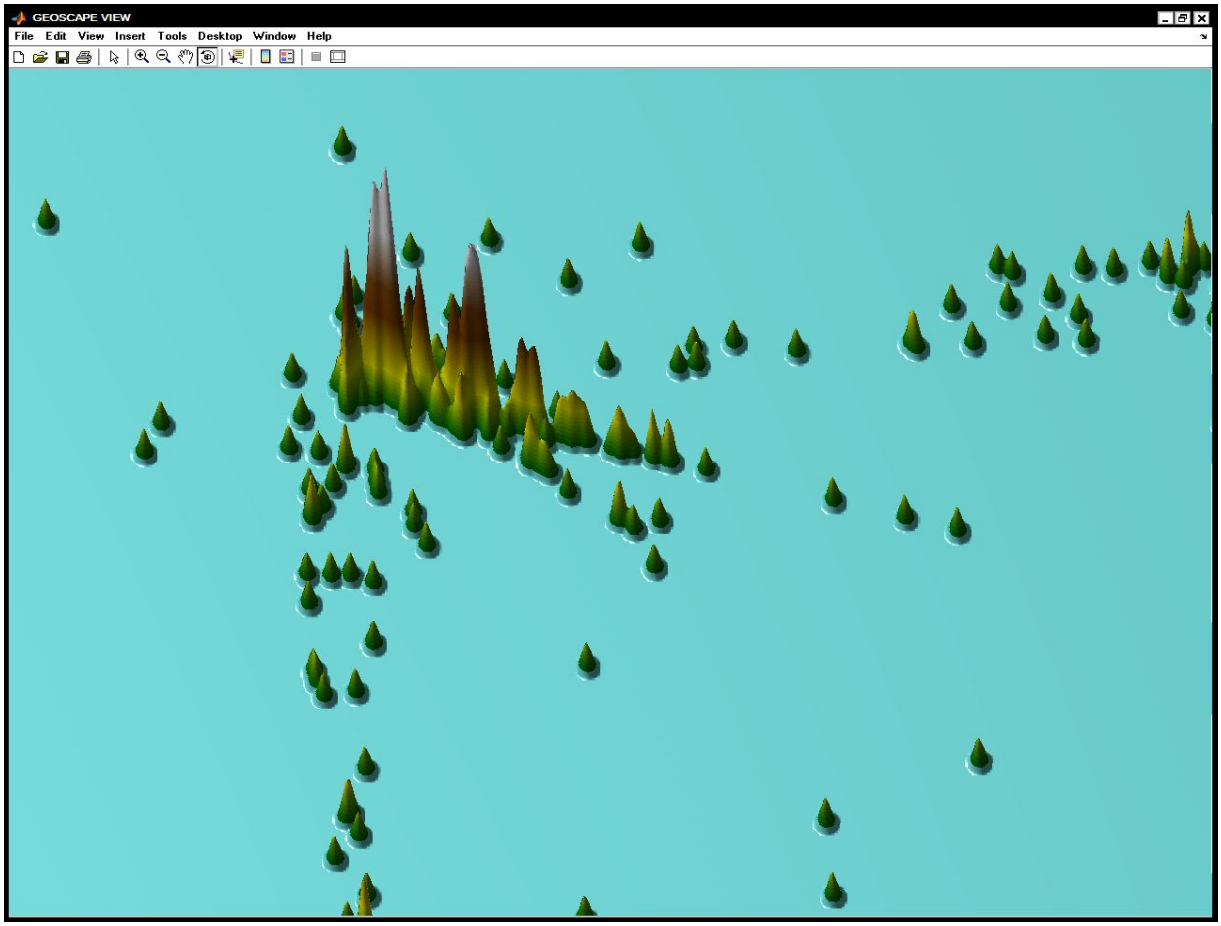
Σχήμα 4.19 Αναζήτηση λεπτομερέστερων πληροφοριών με τη βοήθεια της προβολής «Details View» [KKK10a]

Επιπροσθέτως, ο χρήστης μπορεί να επιλέξει να αλλάξει το επίπεδο λεπτομέρειας σε υψηλότερο, προκειμένου να χωρίσει τις ομάδες σε μικρότερες και να εμφανίσει μια πιο λεπτομερή εικόνα των δεδομένων. Χρησιμοποιώντας τη λειτουργία της εξειδίκευσης, μετακινώντας τη γραμμή κύλισης προς την αντίστοιχη κατεύθυνση, το επίπεδο λεπτομέρειας σταδιακά αυξάνεται. Με τον τρόπο αυτό, ο χρήστης μπορεί να βρει τις εμφωλευμένες «υπο-ομάδες» δεδομένων. Για παράδειγμα, εμφανίζοντας τις επιγραφές που περιγράφουν συνοπτικά τις γεωγραφικές συντεταγμένες των περιοχών, ο χρήστης μπορεί να βρει μέσα από την ομάδα Α με τις περιοχές που εφάπτονται της θάλασσας, και αυτές που επιπλέον είναι και γειτονικές.

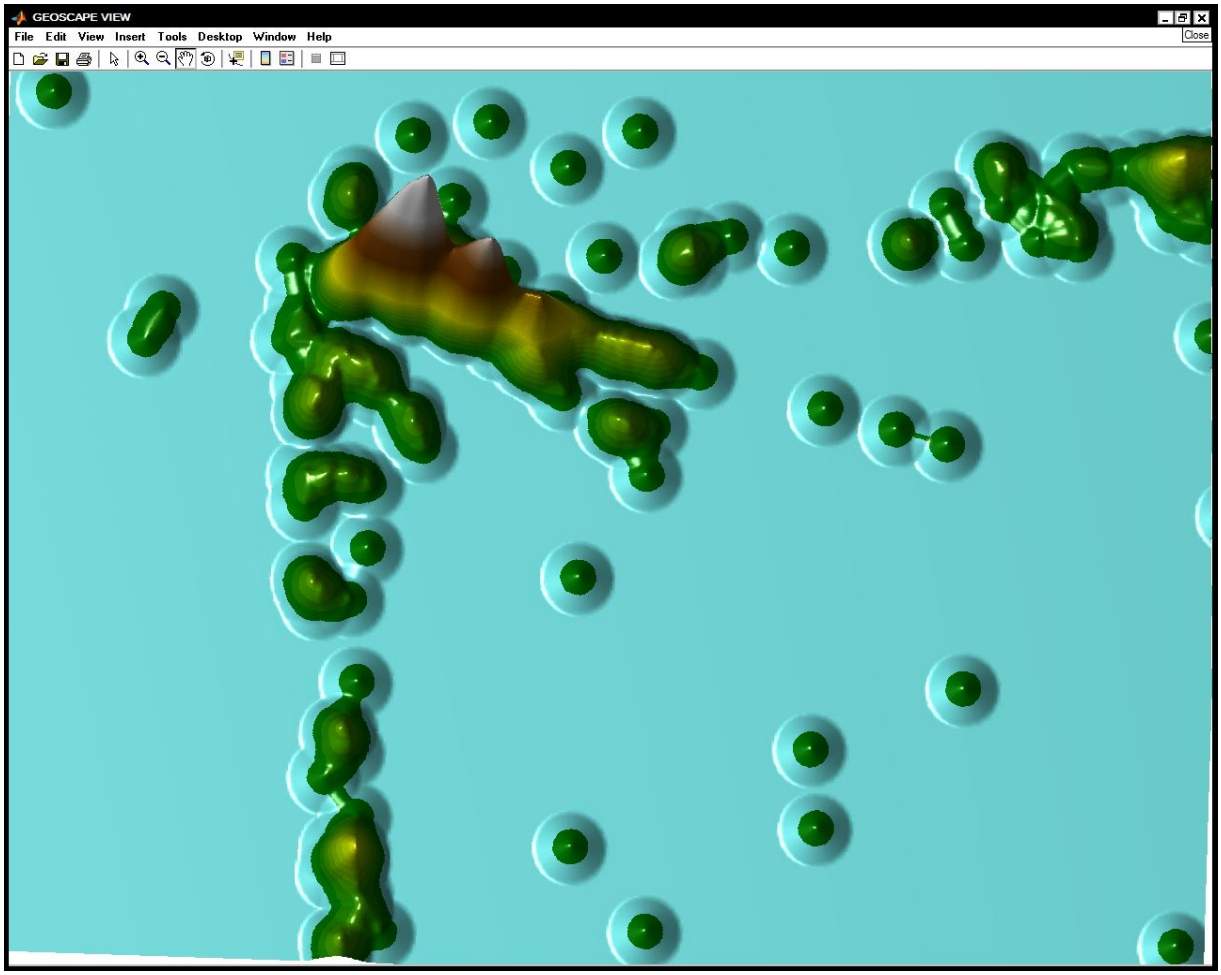
Με παρόμοιο τρόπο, χρησιμοποιώντας τη λειτουργία της γενίκευσης, μετακινώντας τη γραμμή κύλισης προς την αντίθετη κατεύθυνση, το επίπεδο λεπτομέρειας σταδιακά μειώνεται, οι ομάδες συγχωνεύονται σε μεγαλύτερες, το τοπίο ομαλοποιείται και μια λιγότερο λεπτομερής εικόνα των δεδομένων προβάλλεται. Το σχήμα 4.20 δείχνει πώς το ανάγλυφο του τοπίου πληροφοριών αλλάζει, καθώς το επίπεδο λεπτομέρειες μειώνεται.


GEOSCAPE VIEW
File Edit View Insert Tools Desktop Window Help

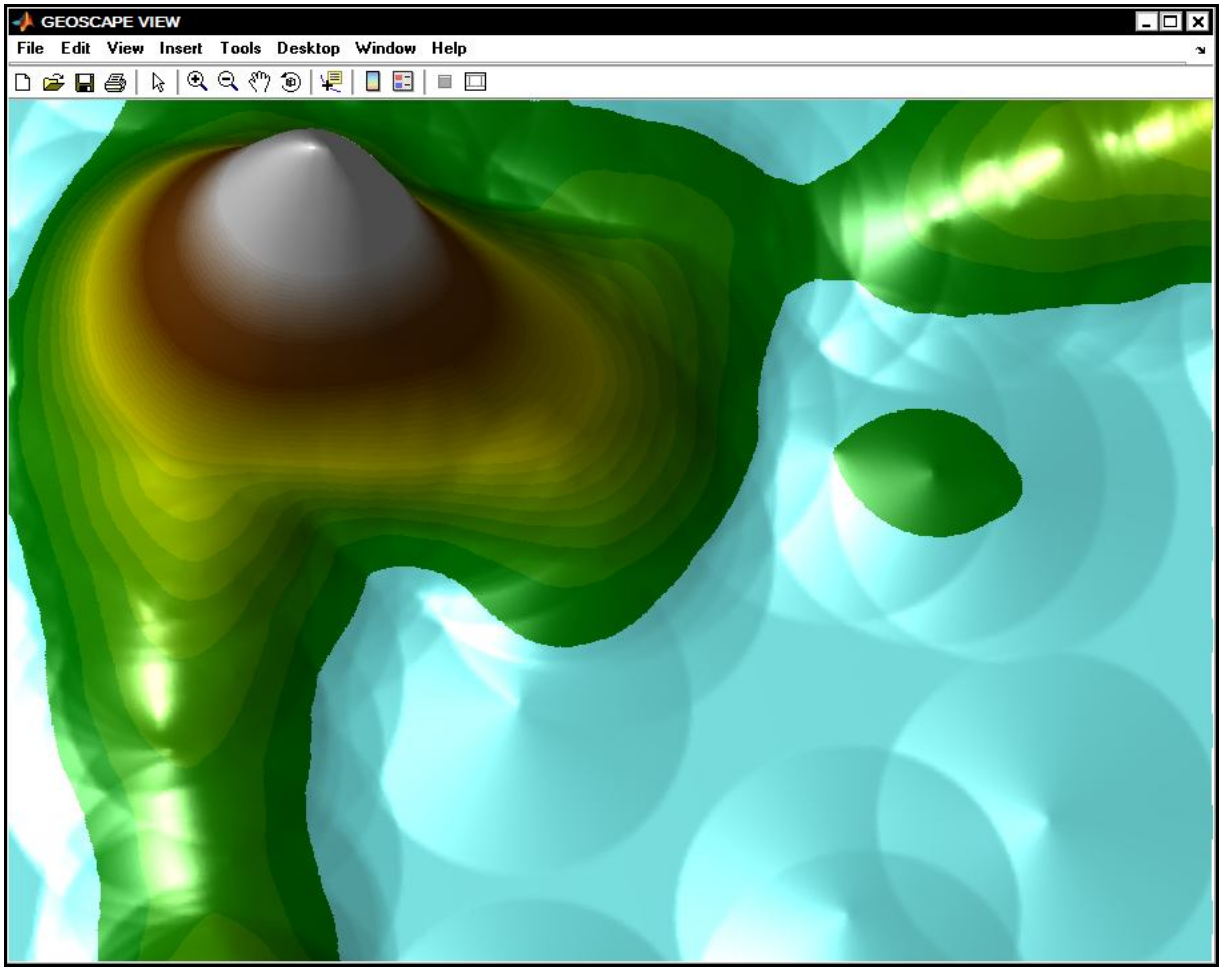
(α)


GEOSCAPE VIEW
File Edit View Insert Tools Desktop Window Help

(β)


GEOSCAPE VIEW
File Edit View Insert Tools Desktop Window Help
Close

(γ)



(δ)

Σχήμα 4.20 Διαδοχικές μεταβολές του ανάγλυφου τοπίου πληροφοριών, από ψηλό επίπεδο λεπτομέρειας (α) σε χαμηλό επίπεδο λεπτομέρειας (δ) [KKK10a]

4.5.5. Συζήτηση

Οι Lakoff και Johnson [LJ80] υποστηρίζουν ότι η χρήση μεταφορών βοηθά την ανθρώπινη αντίληψη. Ακόμα, οι Kuhn και Blumenthal [KB96] εκφράζουν την άποψη ότι οι χωρικές μεταφορές παρέχουν τις ελευθερίες και τους φυσικούς περιορισμούς που επιτρέπουν την εξερεύνηση των πληροφοριών με ένα διαισθητικό τρόπο. Ωστόσο, επειδή οι χωρικές μεταφορές προϋποθέτουν την καταβολή κάποιας επιπλέον νοητικής προσπάθειας μέχρι να επιτευχθεί εξοικείωση, είναι προτιμότερο οι τεχνικές χωρικοποίησης να εφαρμόζονται στην περίπτωση σχετικά πολύπλοκων και μεγάλου όγκου δεδομένων. Επιπροσθέτως, έχουν

αναφερθεί διαφορές στον τρόπο που ο ανθρώπινος νους επεξεργάζεται τις χωρικές μεταφορές. Σχετικά με το τοπίο πληροφοριών, οι έρευνες έχουν οδηγήσει σε μεικτά συμπεράσματα. Για παράδειγμα, η μελέτη που παρουσιάζεται στην ερευνητική εργασία [TSD09] κατέληξε στο συμπέρασμα ότι ενώ οι άνθρωποι μπορούν πιο εύκολα να αντιληφθούν διαισθητικά τη γεωμορφολογία ενός τρισδιάστατου τοπίου και να κατανοήσουν την αντιστοιχία «απόσταση-ομοιότητα», οι πληροφορίες που μεταδίδονται από μια δισδιάστατη χωρικοποίηση γίνονται πιο γρήγορα αντιληπτές.

Ακόμα, λόγω των γνωστών περιορισμών της τεχνικής MDS σχετικών με την αποτελεσματικότητά της να διαχειριστεί δεδομένα με πολύ αυξημένο αριθμό διαστάσεων [SF03], προτείνεται το περιβάλλον GeoScape να χρησιμοποιείται σε δεδομένα με πολλές, αλλά σχετικά περιορισμένες σε αριθμό διαστάσεις, όπως είναι για παράδειγμα τα δεδομένα εισόδου του σεναρίου εφαρμογής.

Το περιβάλλον GeoScape εκμεταλλεύεται την τρίτη διάσταση του χώρου για να οπτικοποιήσει πληροφορίες και να τις οργανώσει σε διάφορα επίπεδα λεπτομέρειας. Όμως, οι γνώμες διίστανται σχετικά με την αποτελεσματικότητα των τρισδιάστατων έναντι δισδιάστατων οπτικοποιήσεων. Για παράδειγμα, η έλλειψη ορατότητας προς ορισμένες κατευθύνσεις είναι ένα από τα προβληματικά ζητήματα που αναφέρονται συχνά σχετικά με τις τρισδιάστατες οπτικοποιήσεις [KKUW06]. Η προτεινόμενη τεχνική χωρικοποίησης που υλοποιείται μέσα από το περιβάλλον του GeoScape αντιμετωπίζει το πρόβλημα αυτό με την παροχή διαδραστικών λειτουργιών όπως είναι η λειτουργία της περιστροφής του τοπίου πληροφοριών, με τη βοήθεια της οποίας μπορεί να αλλάξει η γωνία παρατήρησης και να αποκαλυφθούν τα τμήματα εκείνα του τοπίου που δεν ήταν ορατά. Ακόμα, οι χρήστες έχουν τη δυνατότητα να γνωρίζουν επ' ακριβώς το επίπεδο λεπτομέρειας, δηλαδή το ιεραρχικό επίπεδο ομαδοποίησης, κατά τη διάρκεια της περιήγησής τους στο περιβάλλον της χωρικοποίησης, χάρη στις συντονισμένες προβολές που παρέχονται από το GeoScape και που οπτικοποιούν πολλές και διαφορετικές όψεις των δεδομένων. Μάλιστα ο Nöllenburg [Nöll06] υποστηρίζει ότι ο συνδυασμός πολλών διαφορετικών προβολών δεδομένων συμβάλλει στην τόνωση της οπτικής αντίληψης.

Τέλος, η χρήση χρώματος ενισχύει τα αποτελέσματα της χωρικοποίησης και διευκολύνει την έμφυτη ικανότητα του ανθρώπου να αναγνωρίζει χωρικά πρότυπα. Ως αποτέλεσμα, τα

χαρακτηριστικά του τοπίου, τα βουνά και οι λόφοι, γίνονται άμεσα αντιληπτά. Στο εγγύς μέλλον, σχεδιάζουμε να διερευνήσουμε αυτές τις παρατηρήσεις, βάσει συγκεκριμένης έρευνας ώστε να βελτιώσουμε την αποτελεσματικότητα της τεχνικής χωρικοποίησης που υλοποιήθηκε στο πλαίσιο του πρωτότυπου περιβάλλοντος GeoScape.

5. Συμπεράσματα

5.1 Σύνοψη – Μελλοντική έρευνα

Συνοψίζοντας, στην παρούσα διδακτορική διατριβή, παρουσιάστηκαν τα παρακάτω:

- Μια μέθοδος απόκτησης γνώσης από βάσεις γεωχωρικών δεδομένων με χρήση της ελεγχόμενης γλώσσας Geo-Q

Εισήχθη και περιγράφηκε συνοπτικά η ελεγχόμενη γλώσσα Geo-Q, η οποία μπορεί να χρησιμοποιηθεί για την υποβολή ερωτήσεων προς βάσεις γεωχωρικών δεδομένων με σκοπό την απόκτηση γνώσης. Η γραμματική της Geo-Q επιδέχεται εμπλουτισμό με σκοπό την υποστήριξη *μεγαλύτερης ποικιλίας* ερωτήσεων, με *πλουσιότερο λεξιλόγιο* και *πιο ευέλικτη σύνταξη*.

Ακόμη, επειδή οι Geo-Q ερωτήσεις μπορούν άμεσα να αντιστοιχηθούν με CG και τα CG μπορούν άμεσα να μετατραπούν σε οποιασδήποτε μορφής λογική, η Geo-Q θα μπορούσε μελλοντικά να ενσωματωθεί σ' ένα *σύστημα ερωταπαντήσεων* (answering system) το οποίο να προσπελαύνει *όχι μόνο* βάσεις γεωγραφικών δεδομένων αλλά και άλλες συλλογές δεδομένων (ιστοσελίδες, κείμενα σε φυσική γλώσσα, κλπ).

Τέλος, η Geo-Q θα μπορούσε στο μέλλον να αναπτυχθεί ώστε *να εξορύσσει*, και όχι μόνο να αποσπά γεωγραφική γνώση, από οποιαδήποτε συλλογή γεωχωρικών δεδομένων.

- Η μέθοδος δημιουργίας σημασιολογικών επιγραφών Geo-Labeling

Η Geo-Labeling επεξεργάζεται ορισμούς γεωγραφικών εννοιών διατυπωμένων σε φυσική γλώσσα. Προκειμένου να είναι σημασιολογικά συνεκτικές και αντιπροσωπευτικές, οι επιγραφές δομούνται με τρόπο που να μοιάζουν με σύντομους ορισμούς που συνοψίζουν το περιεχόμενο των ομάδων ορισμών γεωγραφικών εννοιών ενώ ταυτόχρονα διαφοροποιούν τις ομάδες αναμεταξύ τους.

Η μέθοδος μπορεί στο μέλλον να χρησιμοποιηθεί για τη δημιουργία επιγραφών στο πλαίσιο εφαρμογών που επεξεργάζονται και διαχειρίζονται σημασιολογική πληροφορία. Σχετικά παραδείγματα αναφέρθηκαν σε προηγούμενες παραγράφους (αυτόματη δημιουργία γεωχωρικής οντολογίας, ολοκλήρωση γεωχωρικών οντολογιών, ομαδοποίηση αποτελεσμάτων αναζήτησης πληροφοριών, κλπ).

Ακόμη, η δομημένη ανάπτυξη της μεθόδου υπό τη μορφή διαδοχικών βημάτων επεξεργασίας των ορισμών, διευκολύνει τη μελλοντική τυποποίηση και αυτοματοποίησή της.

- Το περιβάλλον χωρικοποίησης GeoScape

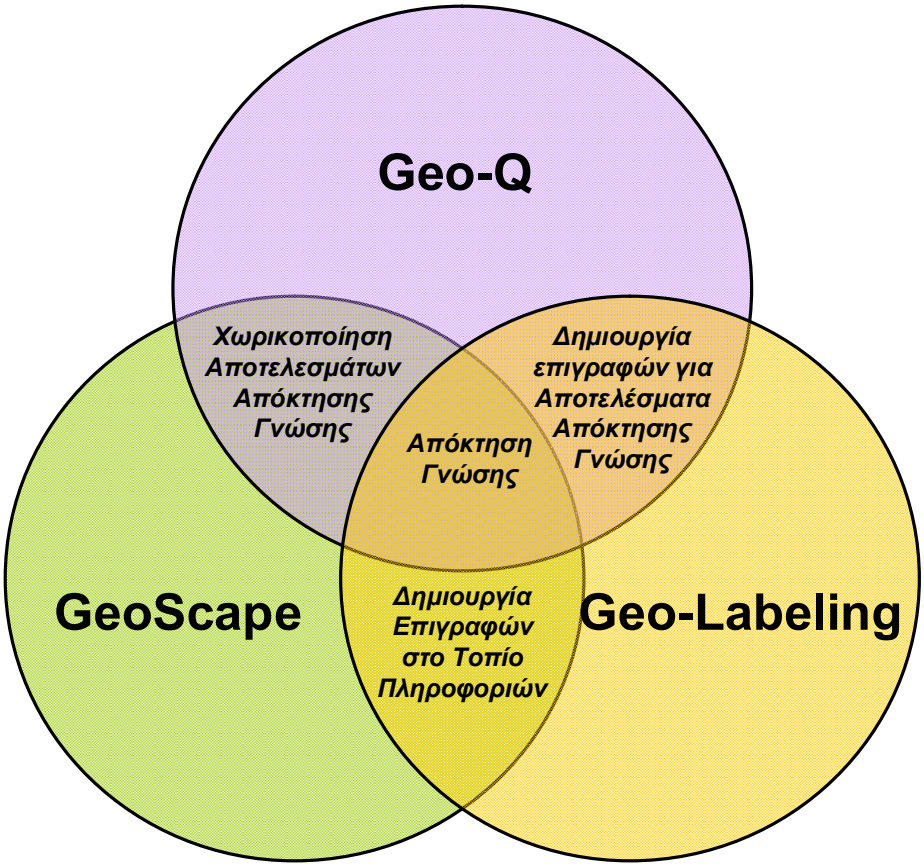
Στο τέταρτο μέρος της διατριβής, εξετάστηκε η χωρικοποίηση ως τεχνική οπτικοποίησης πληροφοριών, που οδηγεί στον ορισμό ενός πρωτότυπου χώρου απεικόνισης στον οποίο οι βασικές χωρικές έννοιες αποκτούν καινούργια σημασία. Η κλασσική χωρική απόσταση ερμηνεύεται πλέον ως το μέτρο της ομοιότητας μεταξύ των απεικονιζόμενων πληροφοριών. Η χαρτογραφική προβολή συνίσταται στη διαδικασία μείωσης των διαστάσεων των δεδομένων ενώ η κλίμακα αποκτά την έννοια του επιπέδου λεπτομέρειας.

Αποδείχθηκε ακόμη ότι η χωρικοποίηση δεν αποτελεί απλή χωρική συσχέτιση, όπως συμβαίνει στους θεματικούς χάρτες, αλλά προσομοίωση μη χωρικού φαινομένου με χωρικό. Οι χωρικές μεταφορές παρέχουν ένα «φιλικό» πλαίσιο για την εξερεύνηση και την ερμηνεία ενός εικονικού χώρου βάσει της ανθρώπινης εμπειρίας και μάλιστα, με δυνατότητα αναπαράστασης των πληροφοριών σε διάφορα επίπεδα λεπτομέρειας.

Προς αυτήν την κατεύθυνση, εισήχθη το περιβάλλον χωρικοποίησης GeoScape, για να προσφέρει στους χρήστες τη δυνατότητα: να περιηγηθούν σε τοπία πληροφοριών σε διάφορα επίπεδα λεπτομέρειας για να εξορύξουν γνώση και να μεταβάλλουν το επίπεδο λεπτομέρειας, με διαδικασίες παρόμοιες με τη χαρτογραφική γενίκευση και ειδίκευση. Η αλληλεπίδραση με το πρωτότυπο περιβάλλον επιτυγχάνεται μέσω συντονισμένων προβολών. Η χρησιμότητα των συντονισμένων προβολών έγκειται στην οπτικοποίηση πολλών και διαφορετικών όψεων των δεδομένων.

5.2 Συνδυασμός Αποτελεσμάτων

Τα παραπάνω αποτελέσματα της διδακτορικής διατριβής απεικονίζονται συνοπτικά και διαγραμματικά στο σχήμα 5.1, φανερώνοντας παράλληλα και το ενδεχόμενο μελλοντικού συνδυασμού μεταξύ των Geo-Q, Geo-Labeling και GeoScape.



Σχήμα 5.1 Αποτελέσματα διατριβής και ενδεχόμενος μελλοντικός συνδυασμός τους

Πράγματι, με το συνδυασμό των Geo-Q και Geo-Labeling, παρέχεται η δυνατότητα απόκτησης γνώσης με χρήση ελεγχόμενης γλώσσας και δημιουργίας επιγραφών για την ομαδοποίηση των αποτελεσμάτων. Με το συνδυασμό, Geo-Labeling και GeoScape, μπορούν να δημιουργηθούν επιγραφές στο τοπίο πληροφοριών για την υποστήριξη του χρήστη στην ερμηνεία του αποτελέσματος της χωρικοποίησης. Με το συνδυασμό GeoScape και Geo-Q, μπορούν τα αποτελέσματα της απόκτησης γνώσης από συλλογές γεωχωρικών δεδομένων με βάση τη Geo-Q να απεικονιστούν στο περιβάλλον χωρικοποίησης του GeoScape, για την εξόρυξη επιπλέον γνώσης. Τέλος, επισημαίνεται ότι ο συνδυασμός και των τριών

συνιστωσών, δηλαδή των Geo-Q, Geo-Labeling και GeoScape, θα μπορούσε να συνεισφέρει στην ανάπτυξη ενός *ολοκληρωμένου περιβάλλοντος απόκτησης γνώσης*.

Βιβλιογραφία

[AHS+04] Aleman-Meza B., Halaschek C., Sheth A., Arpinar I.B., and Sannapareddy G., SWETO: Large-Scale Semantic Web Test-bed. Intl. *Workshop on Ontology in Action*, Banff, Canada (2004).

[ASR+05] Arpinar B., Sheth A, Ramakrishnan C., Usery L., Azami M., and Kwan M., Geospatial Ontology Development and Semantic Analytics. Book Chapter, *Handbook of Geographic Information Science*, Eds: J. P. Wilson and A. S. Fotheringham, Blackwell Publishing, (2005).

[BA98] Bajcsy P. and Ahuja N., Location and Density-based Hierarchical Clustering Using Similarity Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(9) (1998), 1011– 1015.

[Bene91] Benedikt M., Cyberspace: Some Proposals. Cyberspace: First Steps. *MIT Press* (1991), 273–302.

[Berk06] Berkhin P., A Survey of Clustering Data Mining Techniques. In *Grouping Multidimensional Data - Recent Advances in Clustering*, Kogan J, Teboulle M, and Nicholas C (Eds), Springer Berlin Heidelberg, New York (2006), 25–71.

[Bert81] Bertin J., *Graphics and Graphic Information Processing*. De Gruyter, Translated by W. J. Berg and P. Scott (1981).

[Bert83] Bertin J., *Semiology of Graphics*, The University of Wisconsin Press (1983).

[BFR98] Bradley P.S., Fayyad U.M., and Reina C.A., *Scaling EM (Expectation-Maximization) Clustering to Large databases*. Microsoft Technical Report (1998), 98–35.

[BJ94] Benking H. and Judge A.J.N. Design Considerations for Spatial Metaphors: Reflections on the Evolution of Viewpoint Transportation

Systems. WWW document, position paper, *ACM-ECHT Conference*, *Workshop on Spatial Metaphors for Information Systems*, Edinburgh (1994), http://www.uia.org/uiadocs/spatialm.htm.

[BN07] Bedini I. and Nguyen B., *Automatic Ontology Generation: State of the Art*. University of Versailles, Technical report, December 2007, http://bivan.pagespro-orange.fr/Docs/Automatic_Ontology_Generation_State_of_Art.pdf

[Borg97] Borgatti S.P., *Multidimensional Scaling*. WWW Document, (1997), http://www.analytictech.com/borgatti/mds.htm.

[Brac85] Brachman J.R., On the epistemological status of semantic networks. In: *Readings in Knowledge Representation*, Los Altos, CA: Kaufmann, 91–215 (1985).

[Brew94] Brewer C., Color Use Guidelines for Mapping and Visualization. In *Visualization in Modern Cartography*, edited by A.M. MacEachren and D.R.F., Taylor, Elsevier Science, Tarrytown, NY (1994), 123–147.

[BS01] Bloodgood J.M. and Salisbury W.D., Understanding the influence of organizational changes strategies on information technology and knowledge management strategies. *Decision Support Systems*, 31 (2001), 55–69.

[BS97] Barwise J. and Seligman J., *Information Flow*. Cambridge University Press, Cambridge, England, 1997.

[BWD02] Boyack K.W., Wylie B.N., and Davidson G.S., Domain visualization using VxInsight for science and technology management. In *Journal of the American Society for Information Science and Technology*, 53(9) (2002), 764–774.

[Cara99] Caraballo S., Automatic construction of a hypernym-labeled noun hierarchy from text. In: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, College Park, Maryland, 120–126 (1999).

[CGS02] Committee on Information Interchange and Interpretation, *Conceptual Graph Standard*, NCITS.T2 (2002), http://users.bestweb.net/~sowa/cg/cgstand.htm.

[Choo00] Choo C.W., Working with knowledge: How information professionals help organisations manage what they know. *Library Management*, 21(8) (2000), 395–403.

[CMS99a] Card S.K., Mackinlay J.D., and Shneiderman B., *Readings in Information Visualization. Using Vision to Think*, Morgan Kaufmann, San Francisco, CA (1999).

[CMS99b] Card S., Mackinlay J., and Shneiderman B., Trees. In *Readings in Information Visualization*. Card S.K., Mackinlay J.D., and Shneiderman B. (Eds), Morgan Kaufmann Publishers, San Francisco, CA (1999), 149–151.

[Cyre97] Cyre W. R., Knowledge Extractor: A Tool for Extracting Knowledge from Text, In: *Proceedings of the Fifth Int'l. Conf. on Conceptual Structures (ICCS'97)*, Seattle, WA, 607-610, Springer-Verlag, Aug. 8, 1997.

[DCH+03] Derthick M., Christel M., Hauptmann A., Dorbin Ng, Stevens S., and Wactlar H., A Cityscape Visualization of Video Perspectives, In: *Proceedings of the National Academy of Sciences Arthur M. Sackler Colloquium on Mapping Knowledge Domains*, Irvine, CA, (2003), available at http://www.cs.cmu.edu/ sage/papers/Cityscape.pdf, Last date accessed 2.2010.

[Dems06] Demsar U., In: *Data mining of geospatial data: combining visual and automatic methods*, PhD Thesis, KTH - Royal Institute of Technology, Sweden, (2006), http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-3892, Last date accessed 10.2009.

[DF98] Dieberger A. and Frank A.U., A city metaphor for supporting navigation in complex information spaces, In: *Journal of Visual Languages and*

Computing, 9 (1998), 597–622.

[DGA+00] Dos Santos C.R., Gros P., Abel P., Loisel D., Trichaud N., and Paris J.P., Mapping Information onto 3D Virtual Worlds. In *Proceedings of IEEE International Conference on Information Visualization*, London, England, (2000).

[DLR77] Dempster A.P., Laird N.M., and Rubin D.B., Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, Series B, 39 (1977), 1–38.

[DP00] Davenport T.H. and Prusak L., *Working knowledge: How organizations manage what they know*, Boston, Harvard Business School Press (2000).

[DSZ93] Davis R., Shrobe H. and Szolovits P., What is a knowledge representation? *AI Magazine*, 14(1): 17–33 (1993).

[EBNF96] *Extended Backus-Naur form (EBNF) Syntax*. ISO/IEC 14977: standard. Draft document, University of Cambridge, Computer Laboratory (1996), http://www.cl.cam.ac.uk/~mgk25/iso-14977.pdf

[Egen02] Egenhofer J.M., Toward the Semantic Geospatial Web. In *Proceedings of the Tenth ACM International Symposium on Advances in Geographic Information Systems*, McLean, Virginia (2002).

[EKSX96] Ester M., Kriegel H., Sander J., and Xu X., A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon (1996), 226–231.

[EVJW06] Eick C., Vaezian B., Jiang D., and Wang J., Discovery of Interesting Regions in Spatial Datasets Using Supervised Clustering. In *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, Berlin, Germany, (2006), 127–138.

[Fabr01] Fabrikant S.I., Visualizing Region and Scale in Information Spaces. In *Proceedings of the 20th International Cartographic Conference*, Beijing,

China (2001), 2522–2529.

[FB01] Fabrikant S.I. and Buttenfield B.P., Formalizing Semantic Spaces for Information Access. In *Annals of the Association of American Geographers*, Blackwell Publishers, Oxford, UK, 91(2) (2001), 263–280.

[FEAC01] Fonseca F., Egenhofer M., Agouris P. and Camara G., Using Ontologies for Integrated Geographic Information Systems. *Transactions on GIS* (2001).

[FEDC02] Fonseca F.T., Egenhofer M.J., Davis C.A., and Câmara G., Semantic Granularity in Ontology-Driven Geographic Information Systems. In *Annals of Mathematics and Artificial Intelligence*, Special Issue on Spatial and Temporal Granularity, 36 (2002), 121–151.

[FLP+51] Florek K., Lukaszewicz J., Perkal J., Steinhaus H., and Zubrzycki S., Sur la liaison et la division des points d'un ensemble fini. *Colloq. Math.*, 2 (1951), 282–285.

[FSCA04] Frery A.C., Silva C.K.R., Costa E.B., and Almeida E.S., Cartographic Generalization in Virtual Reality. In *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences*, Vol. 35 (2004), 200–204.

[FWR00] Fua Y.H., Ward M.O., and Rundensteiner E.A., Navigating Hierarchies with Structure-Based Brushes. *IEEE Transactions on Visualization and Computer Graphics*, Vol.6(2), (2000), 150–159.

[GCS89] Green P.E., Carmone F.J., and Smith S.M., *Multidimensional scaling: concepts and applications*. Boston: Allyn & Bacon (1989).

[Good08] Goodchild M., Epilogue: Intelligent Systems for GIScience: Where Next? A GIScience Perspective. In *Self-organising maps, Applications in Geographic Information Science*, Agarwal P. and Skupin A. (Eds), England, John Wiley & Sons (2008), 195–198.

[GPQX06] Görg C., Pohl M., Qeli E., and Xu K., Visual Representations. In *Human-Centered Visualization Environments*, Kerren A., Ebert A., and Meyer J.

(Eds), Revised Lectures, GI-Dagstuhl Research Seminar, Dagstuhl Castle, Germany (2006), 163–230.

[Gree84] Green M., Masking by light and the sustained-transient dichotomy. *Perception and Psychophysics*, 24 (1984), 617–635.

[Gree98] Green M., Toward a perceptual science of multidimensional data visualization: Bertin and beyond. *ERGO/GERO Human Factors Science* (1998).

[GW99] Ganter B. and Wille R., *Formal Concept Analysis. Mathematical Foundations*, Berlin: Springer-Verlag (1999).

[Hens04] Hensman S., Construction of Conceptual Graph representation of text. In *Proceedings of the Student Research Workshop at HLT-NAACL*, Boston, USA (2004), 49–54.

[HG07] Hinneburg A. and Gabriel H., Denclue 2.0: Fast Clustering Based on Kernel Density Estimation. In *Proceedings of the 7th International Symposium on Intelligent Data Analysis*, Ljubljana, Slovenia (2007), 70–80.

[HK98] Hinneburg A. and Keim D., An Efficient Approach to Clustering in Large Multimedia Databases with Noise. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, (KDD98)*, New York (1998), 58–65.

[HKLK97] Honkela T., Kaski S., Lagus K., and Kohonen T., WEBSOM - self-organizing maps of document collections. In *Proceedings of WSOM_97 (Workshop on Self-Organizing Maps)*, Espoo, Finland (1997), 310–315.

[HL97] Heady R.B. and Lucas J.L., PERMAP: An Interactive Program for Making Perceptual Maps. *J Behavior Research Methods, Instruments & Computers*, 29 (1997), 450–455.

[HOC96] Huibers T., Ounis I. and Chevallet J.-P., Conceptual Graph Aboutness. In P.W. Eklund, G. Ellis en G. Mann, Eds, *Conceptual Structures: Knowledge Representation as Interlingua*, *4th International Conference*

on Conceptual Structures (ICCS'96), vol.(1115) of Lecture Notes in Artificial Intelligence, Sydney, Australia, August, Springer-Verlag, Berlin (1996), 130–144.

[JD88] Jain A. and Dubes R., *Algorithms for Clustering Data*. Prentice-Hall, Englewood Cliffs, NJ (1988).

[JEC07] Jiang D., Eick F.C., and Chen C., *On Supervised Density Estimation Techniques and their Application to Clustering*. Houston, USA, University of Houston, Department of Computer Science, Technical Report Number UH-CS-07-09 (2007).

[JHR93] Jensen K., Heidorn G., and Richardson S. (Eds.), *Natural Language Processing: The PLNLP Approach*, Kluwer Academic Publishers, USA (1993).

[Joli02] Joliffe I.T., *Principal Component Analysis*, New York: Springer-Verlag, New York, Second Edition (2002).

[JR01] Jordan M. I. and Russell S., Computational Intelligence, in Wilson, Robert A.; & Keil, Frank C. (eds.), *The MIT Encyclopedia of the Cognitive Sciences*, Cambridge, MA: MIT Press (2001).

[KAH96] Koperski K., Adhikary J., and Han J., Spatial Data Mining: Progress and Challenges Survey Paper. In *Proceedings of Workshop on Research Issues on Data Mining and Knowledge Discovery*, Montreal, Canada (1996), 55–70.

[Kamp81] Kamp H., A theory of truth and semantic representation. In: Groenendijk J.A.G., Janssen T.M.V., and Stokhof M.B.J. (Eds.), *Formal Methods in the Study of Language*, Mathematical Centre Tracts 135, Amsterdam, 277–322 (1981).

[Kask97] Kaski S., *Data Exploration Using Self-Organizing Maps*, Neural Networks Research Centre, Helsinki University of Technology (1997).

[KB96] Kuhn W. and Blumenthal B., *Spatialization: Spatial Metaphors for User Interfaces*. Reprinted Tutorial Notes from the ACM Conference on

Human Factors in Computing Systems in Vancouver, GeoInfo 8, Department of Geoinformation, Technical University of Vienna, Vienna, (1996).

[KHK99] Karypis G., Han E.H., and Kumar V., Chameleon: Hierarchical Clustering Using Dynamic Modeling. IEEE Computer, 32(8) (1999), 68–75.

[KK05a] Kokla M. and Kavouras M., Semantic information in geo-ontologies: extraction, comparison, and reconciliation. *Journal on Data Semantics III*, Lecture Notes in Computer Science, 125–142 (2005).

[KK05b] Kontaxaki S. and Kavouras M., Spatial Knowledge Extraction from Geographical Databases: An approach based on the Controlled English Query Language Geo-Q and Conceptual Graphs, In: *Proceedings of GIS Planet 95*, Estoril, Portugal, May-June, 2005.

[KK08] Kavouras M. and Kokla M., *Theories of Geographic Concepts: Ontological Approaches to Semantic Integration*. CRC Press, Taylor & Francis Group, Boca Raton, FL, USA (2008).

[KKK04] Karalopoulos A., Kokla M., and Kavouras M., Geographic Knowledge Representation Using Conceptual Graphs. In *Proceedings of the 7th AGILE Conference on Geographic Information Science*, Science, Crete, Greece (2004).

[KKK10a] Kontaxaki S., Kokla M., and Kavouras M., GeoScape, a Granularity-Depended Spatialization Tool for Visualizing Multidimensional Data Sets, *Geo-Information Science*, Springer, Vol. 4 (2010).

[KKK10b] Kontaxaki S., Kokla M., and Kavouras M., Semantic Labeling of Geo-Concept Clusters. Extended Abstract, *6th International Conference on Geographic Information Science*, Zurich, September 2010, available at http://www.giscience2010.org.

[KKL+00] Kohonen T., Kaski S., Lagus K., Salojarvi J., Paatero V., and Saarela A., Self Organization of a Massive Document Collection. In *IEEE Transactions on Neural Networks*, Special Issue on Neural Networks for

Data Mining and Knowledge Discovery, 11(3) (2000), 574–585.

[KKT03] Kavouras M., Kokla M., and Tomai E., Determination, Visualization and Interpretation of Semantic Similarity among Different Geographic Ontologies. In *Proceedings of the 6th AGILE Conference on Geographic Information Science*, Lyon, France (2003), 51–56.

[KKUW06] Kulyk O., Kosara R. Urquiza J., and Wassink I., Human-Centered Aspects. In *Human-Centered Visualization Environments*, Kerren A., Ebert A., and Meyer J. (Eds), Revised Lectures, GI-Dagstuhl Research Seminar, Dagstuhl Castle, Germany (2006), 13–75.

[Koho95] Kohonen T., *Self-Organizing Maps*, Berlin, Germany, Springer-Verlag (1995).

[Kokl08] Kokla M, 2008, GEONLP: A Tool for the Extraction of Semantic Information from Definitions. In: *Proceedings of the ISPRS 2008 Congress*, Beijing, Volume XXXVII, Commission II, 691–696.

[Kola01] Kolatch E., *Clustering Algorithms for Spatial Databases: A Survey*. Department of Computer Science, University of Maryland, College Park CMSC 725 (2001), http://citeseer.ist.psu.edu/kolatch01clustering.html.

[KPS99] Kumar H.P., Plaisant C., and Shneiderman B., Browsing Hierarchical Data with Multi-Level Dynamic Queries. *Readings in Information Visualization*, Card S.K., Mackinlay J.D., and Shneiderman B. (Eds), Morgan Kaufmann Publishers, San Francisco, CA (1999), 295–305.

[KR90] Kaufman L. and Rousseeuw P., *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, New York (1990).

[KTKK10] Kontaxaki S., Tomai E., Kokla M., and Kavouras M., Visualizing multidimensional data through granularity-dependent spatialization. In: *Proceedings of the SPIE Conference on Visualization and Data Analysis 2010*, doi: 10.1117/12.838430, SPIE Vol.7530, 75300M (2010).

[KW78] Kruskal J.B. and Wish M., *Multidimensional Scaling*. Sage (1978).

[LJ80] Lakoff G. and Johnson M., *Metaphors We Live By*, The University of Chicago Press (1980).

[MacQ67] MacQueen J., Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1 (1967), 281–297.

[MDP08] Martinez A.A., Dolado Cosin J.J., and Presedo Garcia C., A landscape metaphor for visualization of software projects. In: *Proceedings of the 4th ACM symposium on Software visualization*, Ammersee, Germany (2008), 197–198.

[ME98] Murray A.T. and Estivill-Castro V., Cluster discovery techniques for exploratory spatial data analysis. In *International Journal of Geographical Information Science*, Vol. 12(5) (1998), 431–443.

[MGL02] Montes-y-Gómez M., Gelbukh A. and López-López A., Text mining at Detail Level using Conceptual Graphs. In *Lecture Notes in Artificial Intelligence*, Springer, Vol. 2393 (2002).

[MH01] Miller H.J and Han J., *Geographic data Mining and Knowledge Discovery*. Taylor & Francis, London (2001).

[MKB80] Mardia K.V., Kent J.T., and Bibby J.M., *Multivariate Analysis (Probability and Mathematical Statistics)*, Academic Press, London (1980).

[MR99] Merkl D. and Rauber A., Automatic Labeling of Self-Organizing Maps for Information Retrieval. In *Lecture Notes in Artificial Intelligence*, 1574 (1999), 228–237, Springer Verlag.

[Nöll06] Nöllenburg M., Geographic Visualization. In *Human-Centered Visualization Environments*, Kerren A., Ebert A., and Meyer J. (Eds), Revised Lectures, GI-Dagstuhl Research Seminar, Dagstuhl Castle, Germany (2006) 257–294.

[Old02] Old L.J., Information Cartography: Using GIS for Visualizing Non-Spatial Data. In *Proceedings of ESRI International Users' Conference*, San Diego, CA, WWW document (2002), http://gis.esri.com/library/userconf/proc02/pap0239/p0239.htm.

[OP98] Ounis I., and Pasca M., Modeling, Indexing and Retrieving Images Using Conceptual Graphs. In *Proceedings of the 9th {DEXA} International Conference on Database and EXpert Systems Applications*, Vienna, Austria, Quirchmayr G. and Schweighofer E. and Bench-Capon T.J.M. (Eds) (1998), 226–239.

[Open98] Openshaw S., Building Automated Geographical Analysis and Explanation Machines. In *GeoComputation: A Primer*, Wiley Chichester, (1998), 95–115.

[PE88] Pullar D. and Egenhofer M., Toward formal definitions of topological relations among spatial objects. In *Proceedings of the Third International Symposium on Spatial Data Handling*; Sydney, Australia (1988).

[PFMA06] Pinho R., Ferreira de Oliveira M.C., Minghim R., and Andrade M.G., Voromap: A Voronoi-tool for visual exploration of multi-dimensional data. In *Proceedings of the 10th International Conference on Information Visualization* (2006), 39–44.

[PR04] Pantel P. and Ravichandran D., Automatically Labeling Semantic Classes. In: *Proceedings of the HLT-NAACL Conference*, Boston, MA 321–328 (2004).

[PU00] Popescul A. and Ungar L.H., *Automatic Labeling of Document Clusters*. WWW document, Unpublished MS, Department of Computer and Information Science, University of Pennsylvania (2000), http://www.cis.upenn.edu/~popescul/publications.html.

[RE00] Rodríguez A. and Egenhofer M. Determining Semantic Similarity Among Entity Classes from Different Ontologies. *IEEE Transactions on Knowledge and Data Engineering* (2000).

[RE03] Rodríguez A. and Egenhofer M., Determining semantic similarity among entity classes from different ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 15(2) (2003), 442–456.

[Resn99] Resnik O., Semantic Similarity in Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity and Natural Language. *Journal of Artificial Intelligence Research*, 11 (1999), 95–130.

[Rich83] Rich E., *Artificial Intelligence*. McGraw-Hill Book Co., ISBN 0-07-052261-8 (1983), 436.

[RMBB89] Rada R., Mili H., Bicknell E., and Blettner M., Development and Application of a Metric on Semantic Nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1) (1989), 17–30.

[SB05] Stein B. and Busch M., Density-Based Cluster Algorithms in Low-Dimensional and High-Dimensional Applications. In *Proceedings of the Second International Workshop on Text-Based Information Retrieval* (2005), 45–56.

[SB97] Skupin A. and Buttenfield B.P., Spatial Metaphors for Display of Information Spaces. In *Proceedings of the International Research Symposium on Computer-based Cartography*, AUTO-CARTO 13, Seattle, WA (1997), 116–125.

[SBG00] Sprenger T.C., Brunella R., and Gross M.H., H-BLOB: A Hierarchical Visual Clustering Method Using Implicit Surfaces. In *Proceedings of the IEEE Symposium on Information Visualization*, Salt Lake City, Utah (2000), 61–68.

[Schi96] Schikuta E., Grid-Clustering: a Fast Hierarchical Clustering Method for Very Large Data Sets. In *Proceedings of the 13th International Conference on Pattern Recognition*, 2 (1996), 101–105.

[SE04] Stein B. and zu Eissen S.M., Topic Identification: Framework and Application. In: Tochtermann K. and Maurer H. (Eds.), *Proceedings of the 4th International Conference on Knowledge Management* (I-KNOW

04), Graz, Austria, Journal of Universal Computer Science, 353–360 (2004).

[SE97] Schikuta E. and Erhart M., The BANG-Clustering System: Grid-Based Data Analysis. In *Proceeding of the 2nd International Symposium on Advances in Intelligent Data Analysis*, Reasoning about Data, London, UK (1997), 513–524.

[SEKX98] Sander J., Ester M., Kriegel H.P., and Xu X., Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications. *Data Mining and Knowledge Discovery*, Kluwer Academic Publishers, 2(2) (1998), 169–194.

[SF03] Skupin A. and Fabrikant S.I., Spatialization Methods: A Cartographic Research Agenda for Non-Geographic Information Visualization. In *Cartography and Geographic Information Science*, 30(2) (2003), 95–119.

[SFS98] Schwitter R., Fuchs N.E., and Schwertel U., Attempto - Controlled English (ACE) for Software Specifications. *Second International Workshop on Controlled Language Applications*, Language Technologies Institute, Carnegie Mellon University, Pittsburgh (1998).

[Shne92] Shneiderman B., Tree Visualization with Tree-Maps: A 2-D Space-Filling Approach. *ACM Transactions on Graphics*, 11(1) (1992), 92–99.

[Skup00] Skupin A., From Metaphor to Method: Cartographic Perspectives on Information Visualization. In *Proceedings of the IEEE Symposium on Information Visualization*, Salt Lake City, Utah (2000), 91–97.

[Skup01] Skupin A., Cartographic Considerations for Map-like Interfaces to Digital Libraries. *ACM+IEEE Joint Conference on Digital Libraries*, Workshop on Visual Interfaces to Digital Libraries - Its Past, Present, and Future, (2001), http://www.indiana.edu/visual01/skupin.pdf, Last date accessed 10.2009.

[Skup02] Skupin A., A Cartographic Approach to Visualizing Conference Abstracts. *IEEE Computer Graphics and Applications*, 22(1) (2002), 50–

58.

[Skup99] Skupin A., Revisiting Töpfer: Implications of the Radical Law for Scalable Spatialization. *3rd Workshop on Progress in Automated Map Generalization*, International Cartographic Association, Working Group on Map Generalization, Ottawa, Canada (1999).

[Sowa00] Sowa J., *Knowledge Representation: Logical, Philosophical and Computational Foundations*. Brooks Cole Publishing Co., ISBN 0-534-9496-7 (2000), 594.

[Sowa04a] Sowa J., *Common Logic Controlled English Specifications*. WWW document (2004), http://www.jfsowa.com/clce/specs.htm.

[Sowa04b] Sowa J., Graphics and Languages For the Flexible Modular Framework. *International Conference on Conceptual Structures (ICCS)* (2004), http://www.jfsowa.com/pubs/gal4fmf.htm.

[Sowa84] Sowa J., *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley, ISBN 05214449004 (1984), 406.

[Swer08] Swering A., Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey. *Transactions in GIS*, 12(1) (2008), 5–29.

[SZ00] Stasko J. and Zhang E., Focus + Context Display and Navigation Techniques for Enhancing Radial, Space-Filling Hierarchy Visualizations. In *Proceedings of the IEEE Symposium on Information Visualization*, Salt Lake City, Utah (2000), 57–68.

[TC06] Treeratpituk P. and Callan J., Automatically Labeling Hierarchical Clusters. In *Proceedings of the 6th National Conference on Digital Government Research*, San Diego, CA, (2006) 167–176.

[TD00] Tahmane A.C. and Dunlop D.D., *Statistics and Data Analysis*. Prentice Hall, 299−330 (2000).

[TF97] Timpf S. and Frank A.U., Using hierarchical spatial data structures for hierarchical spatial reasoning. Spatial Information Theory - A Theoretical

Basis for GIS. International Conference COSIT'97, Hirtle S.C. and Frank A.U., Berlin-Heidelberg, Springer-Verlag, In *Lecture Notes in Computer Science*, 1329 (1997), 69–83.

[THL03] Tay S.C., Hsu W., and Lim K.H., Spatial data mining: Clustering of hot spots and pattern recognition. *International Geoscience & Remote Sensing Symposium*, Toulouse, France (2003).

[Tobl70] Tobler W., A Computer Model Simulating Urban Growth in the Detroit Region. *Economic Geography*, 46(2) (1970), 234–240.

[TSD09] Tory M., Swindells C. and Dreezer R., Comparing Dot and Landscape Spatializations for Visual Memory Differences. *IEEE Trans. Vis. Comput. Graph.* **15**(6) (2009), 1033–1040.

[Tuft83] Tufte E.R., The Visual Display of Quantitative Information. *Graphics Press* (1983).

[Tver77] Tversky A., Features of Similarity. *Psychological Review*, 84(4) (1977), 327–352.

[Ults93] Ultsch A., Self-Organizing Neural Networks for Visualization and Classification. *Information and Classification - Concepts, Methods, and Applications*, Opitz O., Lausen B., and Klar R., (Eds), Berlin, Germany, Springer-Verlag (1993), 307–13.

[Vesa00] Vesanto J., *Using SOM in Data Mining, Department of Computer Science and Engineering*, Technical Report, Helsinki University of Technology (2000).

[VHAP00] Vesanto J., Himberg J., Alhoniemi E., and Parhankangas J., SOM toolbox for Matlab 5. Technical Report A57, Helsinki University of Technology, Finland (2000).

[VMMK03] Vassiliou M., Markantonatou S., Maistros Y., and Karkaletsis V., "Evaluating Specifications for Controlled Greek", *EAMT-CLAW 2003*, (2003), http://www.mt-archive.info/CLT-2003-Vassiliou.pdf

[Will92] Wille R., Concept Lattices and Conceptual Knowledge Systems. In: *Computers and Mathematics with Applications*, 23(6-9) (1992), 493–515.

[Wise99] Wise J.A., The Ecological Approach to Text Visualization. *Journal of the American Society for Information Science*, 50(13) (1999), 1224–1233.

[WL08] Wettel R. and Lanza M., CodeCity: 3D visualization of large-scale software. In *Companion of the 30th International Conference on Software Engineering*, Leipzig, Germany, 9 (2008), 921–922.

[Wood75] Woods A.W., What's in a link: foundations for semantic networks. In: Bobrow D.G. and Collins A. (Eds.), *Representation and Understanding*, ISBN 0-121-08550-3, Academic Press, New York, 35–82 (1975).

[WYM97] Wang W., Yang J., and Muntz R., STING: a Statistical Information Grid Approach to Spatial Data Mining. *In Proceedings of the 23rd Conference on Very Large Data Bases*, Athens, Greece (1997), 186–195.

[Yaro92] Yarowsky D., Word−Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. In: *Proceedings of COLING−92*, Nantes, 454−460 (1992).

[YCSZ98] Yu D., Chatterjee S., Sheikholeslami G., and Zhang A., *Efficiently Detecting Arbitrary Shaped Clusters in Very Large Datasets with High Dimensions*. SUNY, Technical Report 98–08, Buffalo, Computer Science, (1998).

[ZCY08] Zavesky E., Chang S.F., and Yang C.C., Visual Islands: Intuitive Browsing of Visual Search Results. In: *Proceedings of the 2008 International Conference on Content-Based Image and Video Retrieval*, Niagara Falls, Canada (2008), 617–626.

[ZZLY02] Jiwei Zhong, Haiping Zhu, Jianming Li and Yong Yu, Conceptual Graph Matching for Semantic Search. *ICCS* (2002), 92–196.

[ΚΚ10] Κονταξάκη Σ. και Κάβουρας Μ., Χωρικοποίηση: η Γεω-Οπτικοποίηση Αφηρημένων Πληροφοριών, 9ο Πανελλήνιο Συνέδριο Γεωγραφίας,

Αθήνα, Νοέμβριος 2010, http://www.geography2010.gr.

[Στεφ03] Στεφανάκης Ε., Βάσεις Γεωγραφικών δεδομένων και Συστήματα Γεωγραφικών Πληροφοριών. Εκδόσεις Παπασωτηρίου (2003).

Παράρτημα 1 (Κώδικας Matlab)

Παρακάτω ακολουθεί απόσπασμα του κώδικα Matlab που συντάχθηκε για την υλοποίηση της τεχνικής χωρικοποίησης που εφαρμόζει το πρωτότυπο περιβάλλον GeoScape.

%%%%%%%%%% F_SIMILARITY %%%%%%%%%

```
function zplus = similarity_impact(x0,y0)
n=120;
%n=240;
s=11;
[x,y]=meshgrid(-n:1:n);

for i=1:2*n+1
    for j=1:2*n+1
        d=sqrt((x(i,j)-x0).^2+(y(i,j)-y0).^2);
        if d <= s
            zplus(i,j)=1-(1/s)*sqrt((x(i,j)-x0).^2+(y(i,j)-y0).^2);
        else
            zplus(i,j)=0;
        end
    end
end
```

%%%%%%%% EXAMPLE DENDROGRAM %%%%%%%

```
S={'A'; 'B'; 'C'; 'D'; 'E'; 'F' };
X = [-10 30 0; 0 20 0; 10 35 0; 10 -10 0; 0 -15 0; -20 -30 0];
D = pdist(X, 'euclidean');
Y = mdscale(D, 2);
```

```
D = pdist(Y, 'euclidean');
Z = linkage(D, 'single');
dendrogram(Z, 'labels', S');
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


%%%%%%%        EXAMPLE LANDSCAPE       %%%%%%%

map=[
[   0.50196   ,   1   ,   1   ];
[   0   ,   0.50196   ,   0   ];
[   0.125   ,   0.56422   ,   0   ];
[   0.25   ,   0.62647   ,   0   ];
[   0.375   ,   0.68873   ,   0   ];
[   0.5   ,   0.75098   ,   0   ];
[   0.625   ,   0.81324   ,   0   ];
[   0.75   ,   0.87549   ,   0   ];
[   0.875   ,   0.93775   ,   0   ];
[   1   ,   1   ,   0   ];
[   0.96443   ,   0.9465 ,   0   ];
[   0.92885   ,   0.893   ,   0   ];
[   0.89328   ,   0.8395 ,   0   ];
[   0.8577 ,   0.78599   ,   0   ];
[   0.82213   ,   0.73249   ,   0   ];
[   0.78655   ,   0.67899   ,   0   ];
[   0.75098   ,   0.62549   ,   0   ];
[   0.71541   ,   0.57199   ,   0   ];
[   0.67983   ,   0.51849   ,   0   ];
[   0.64426   ,   0.46499   ,   0   ];
[   0.60868   ,   0.41148   ,   0   ];
[   0.57311   ,   0.35798   ,   0   ];
[   0.53753   ,   0.30448   ,   0   ];
```

```
[ 0.50196 , 0.25098 , 0 ];
[ 0.50562 , 0.25649 , 0.0073529 ];
[ 0.50928 , 0.26199 , 0.014706 ];
[ 0.51295 , 0.2675 , 0.022059 ];
[ 0.51661 , 0.27301 , 0.029412 ];
[ 0.52027 , 0.27852 , 0.036765 ];
[ 0.52393 , 0.28402 , 0.044118 ];
[ 0.52759 , 0.28953 , 0.051471 ];
[ 0.53126 , 0.29504 , 0.058824 ];
[ 0.56474 , 0.34539 , 0.12605 ];
[ 0.59822 , 0.39575 , 0.19328 ];
[ 0.6317 , 0.4461 , 0.2605 ];
[ 0.66518 , 0.49646 , 0.32773 ];
[ 0.69866 , 0.54681 , 0.39496 ];
[ 0.73215 , 0.59717 , 0.46218 ];
[ 0.76563 , 0.64752 , 0.52941 ];
[ 0.79911 , 0.69787 , 0.59664 ];
[ 0.83259 , 0.74823 , 0.66387 ];
[ 0.86607 , 0.79858 , 0.73109 ];
[ 0.89955 , 0.84894 , 0.79832 ];
[ 0.93304 , 0.89929 , 0.86555 ];
[ 0.96652 , 0.94965 , 0.93277 ];
[ 1 , 1 , 1 ];
[ 1 , 1 , 1 ];
[ 1 , 1 , 1 ];
[ 1 , 1 , 1 ];
[ 1 , 1 , 1 ];
[ 1 , 1 , 1 ];
[ 1 , 1 , 1 ];
[ 1 , 1 , 1 ];
[ 1 , 1 , 1 ];
[ 1 , 1 , 1 ];
```

```
[    1    ,    1    ,    1    ];
[    1    ,    1    ,    1    ];
[    1    ,    1    ,    1    ];
[    1    ,    1    ,    1    ];
[    1    ,    1    ,    1    ];
[    1    ,    1    ,    1    ];
[    1    ,    1    ,    1    ];
[    1    ,    1    ,    1    ];
[    1    ,    1    ,    1    ]
];
b = rot90(Y);
Y = b;
[m,n] = size(Y) ;
xy=zeros(101);
z=0;
for i=1:n
    z_current = similarity_impact_example(Y(1,i),Y(2,i));
    [u,v]=size(z_current);
    max_z_current=0;
    for k=1:u
        for l=1:v
            if z_current(k,l)>=max_z_current
                max_z_current =z_current(k,l);
                i_max=k; j_max=l;
            end
        end
    end
    xy(i_max,j_max) =  i;
    z = z + z_current ;
end

set(gcf, 'color', 'white');
```

```
colormap(map);
surf(z,'FaceColor','interp','EdgeColor','none');
camlight left; lighting phong;
hold on;
grid off; hold on; axis off; hold on;
xy_plus=xy;
[m,n] = size(xy_plus) ;
for i=1:n
    for j=1:m
        if xy_plus(i,j) ~= 0
            q=xy_plus(i,j);
            text(j,i, z(i,j)+0.1, S(q), 'color', [1 0
0],'FontSize',8,'HorizontalAlignment','left','FontWeight','bold');
        end
    end
end
```

Παράρτημα 2 (Υποσύνολο Δεδομένων Natura 2000)

Παρακάτω παρουσιάζονται τα πολυδιάστατα δεδομένα που επιλέχθηκαν από τη βάση γεωγραφικών δεδομένων του διαδικτυακού τόπου του Υπουργείου Περιβάλλοντος, Ενέργειας και Κλιματικής Αλλαγής[8], και στη συνέχεια συνδυάστηκαν και επεξεργάστηκαν ώστε να χρησιμοποιηθούν ως δεδομένα εισόδου στο παράδειγμα που αναπτύσσεται στην παράγραφο §4.5.4.

SITE_NAME	LON_ DEG	LON_ MIN	LON_ SEC	LAT_ DEG	LAT_ MIN	LAT_ SEC	ALT_ MEAN	ALT_ MAX	ALT_ MIN
DASOS DADIAS - SOUFLI	26	10	11	41	6	52	189	614	15
TREIS VRYSES	26	0	13	41	8	17	530	1034	190
FENGARI SAMOTHRAKIS, ANATOLIKES AKTES, VRACHONISSIDA ZOURAFA KAI THALASSIA ZONI	25	40	57	40	27	32	629	1600	-50
VOUNA EVROU	26	10	19	41	6	55	185	614	13
DELTA EVROU	26	4	31	40	45	53	2	66	0
DELTA EVROU KAI DYTIKOS VRACHIONAS	26	4	3	40	46	21	2	32	0
PARAPOTAMIO DASOS VOREIOU EVROU KAI ARDA	26	23	50	41	37	44	60	332	0
NOTIO DASIKO SYMPLEGMA EVROU	25	56	57	40	57	37	235	848	21
OREINOS EVROS - KOILADA DEREIOU	26	2	6	41	12	9	402	1064	69

[8] Διαθέσιμο από το σύνδεσμο http://www.minenv.gr/1/12/121/12103/g1210300/g12103000000.html, τελευταία προσπέλαση Ιούνιος 2010.

KOILADA ERYTHROPOTAMOU: ASVESTADES, KOUFOVOUNO, VRYSIKA	26	21	15	41	22	13	97	263	20
SAMOTHRAKI: OROS FENGARI KAI PARAKTIA ZONI	25	34	18	40	26	58	501	1598	0
OROS CHAINTOU - KOULA KAI GYRO KORYFES	24	48	27	41	19	27	1263	1820	683
STENA NESTOU	24	43	48	41	7	34	456	1281	0
AISTHITIKO DASOS NESTOU	24	42	60	41	6	34	222	817	0
POTAMOS FILIOURIS	25	34	17	41	1	36	80	624	5
POTAMOS KOMPSATOS (NEA KOITI)	25	11	35	41	9	49	95	217	15
MARONEIA - SPILAION	25	30	13	40	55	56	143	160	129
LIMNES KAI LIMNOTHALASSES TIS THRAKIS - EVRYTERI PERIOCHI KAI PARAKTIA ZONI	25	11	10	40	57	16	10	209	0
LIMNES VISTONIS, ISMARIS - LIMNOTHALASSES PORTO LAGOS, ALYKI PTELEA, XIROLIMNI, KARATZA	25	6	0	41	3	8	7	43	0
KOILADA FILIOURI	25	48	9	41	13	21	573	1225	56
KOILADA KOMPSATOU	25	10	26	41	14	19	384	1307	14
DASOS FRAKTOU	24	29	49	41	32	36	1711	1949	1366
RODOPI (SIMYDA)	24	8	38	41	29	52	1166	1530	704
PERIOCHI ELATIA, PYRAMIS KOUTRA	24	19	17	41	30	12	1430	1810	919
KORYFES OROUS FALAKRO	24	5	20	41	17	17	1382	2200	803
KENTRIKI RODOPI KAI KOILADA NESTOU	24	24	41	41	21	7	989	1940	240
OROS FALAKRO	24	5	30	41	16	10	903	2216	117
DELTA NESTOU KAI LIMNOTHALASSES KERAMOTIS KAI NISOS THASOPOULA	24	48	55	40	52	45	7	56	0
KORYFES OROUS PANGAIO	24	6	7	40	54	41	1069	1955	165
ORMOS POTAMIAS - AKR. PYRGOS EOS N. GRAMVOUSSA	24	46	13	40	42	40	20	85	0
KOLPOS PALAIOU - ORMOS ELEFTHERON	24	20	5	40	49	51	9	39	0
DELTA NESTOU KAI LIMNOTHALASSES KERAMOTIS - EVRYTERI PERIOCHI KAI PARAKTIA ZONI	24	45	36	40	55	29	8	120	0
OROS PANGAIO KAI NOTIES YPOREIES TOU	24	41	51	40	42	11	682	1940	59
THASOS (OROS YPSARIO KAI PARAKTIA ZONI) KAI NISIDES KOINYRA, XIRONISI	24	6	9	40	54	4	461	1180	0
OROS VERMIO	22	0	55	40	27	54	1287	2032	491

STENA ALIAKMONA	22	13	16	40	26	34	330	747	46
LIMNES VOLVI KAI LAGKADA - EVRYTERI PERIOCHI	23	19	38	40	40	33	71	560	27
DELTA AXIOU - LOUDIA - ALIAKMONA - EVRYTERI PERIOCHI - AXIOUPOLI	22	42	31	40	34	41	8	80	0
STENA RENTINAS - EVRYTERI PERIOCHI	23	39	15	40	39	17	193	710	0
LIMNOTHALASSA ANGELOCHORIOU	22	49	13	40	28	59	11	30	0
LIMNES KORONEIAS - VOLVIS, STENA RENTINAS KAI EVRYTERI PERIOCHI	22	10	7	41	5	30	306	1140	0
DELTA AXIOU - LOUDIA - ALIAKMONA - ALYKI KITROUS	22	42	8	40	31	3	35	353	0
LIMNOTHALASSA EPANOMIS	22	54	23	40	23	21	3	20	0
LIMNOTHALASSA EPANOMIS KAI THALASSIA PARAKTIA ZONI	22	54	22	40	23	18	3	20	0
LIMNI PIKROLIMNI	22	48	53	40	49	41	76	94	54
YDROCHARES DASOS MOURION	22	46	34	41	14	20	157	169	140
LIMNI DOIRANI	22	46	17	41	13	11	153	212	140
LIMNI PIKROLIMNI - XILOKERATEA	22	49	15	40	49	56	81	106	54
PERIOCHI ELOUS ARTZAN	22	39	25	40	59	34	65	155	20
PERIOCHI ANTHOFYTOU	22	43	21	40	50	57	54	120	34
KORYFES OROUS VORA	21	53	30	40	56	25	1324	2503	343
ORI TZENA	22	11	11	41	7	41	1171	2182	171
OROS PAIKO	22	18	14	41	1	4	939	1623	219
LIMNI AGRA	21	55	52	40	48	16	480	606	427
STENA APSALOU - MOGLENITSAS	22	7	58	40	51	18	155	400	38
LIMNI KAI FRAGMA AGRA	21	56	36	40	48	15	477	606	403
ORI TZENA KAI PINOVO	23	42	20	41	11	21	1004	2172	140
OROS VORAS	23	33	15	40	24	58	1013	2500	120
OROS PAIKO, STENA APSALOU KAI MOGLENITSAS	24	12	18	40	19	19	560	1647	34
OROS OLYMPOS	22	23	47	40	5	58	1519	2891	193
PIERIA ORI	22	13	8	40	14	38	1267	2190	162

OROS TITAROS	22	10	14	40	9	45	1342	1837	672
ALYKI KITROUS - EVRYTERI PERIOCHI	22	38	38	40	21	18	4	19	0
LIMNI KERKINI - KROUSIA - KORYFES OROUS BELES, ANGISTRO - CHAROPO	23	9	12	41	13	12	383	2005	14
EKVOLES POTAMOU STRYMONA	23	51	30	40	47	36	19	64	0
AI GIANNIS - EPTAMYLOI	23	34	52	41	5	24	54	107	34
KORYFES OROUS MENOIKION - OROS KOUSKOURAS - YPSOMA	23	47	31	41	9	47	1006	1963	227
KORYFES OROUS ORVILOS	23	36	37	41	22	5	1321	2200	757
ORI VRONTOUS - LAILIAS - EPIMIKES	23	34	56	41	16	9	1155	1767	681
TECHNITI LIMNI KERKINIS - OROS KROUSIA	23	5	37	41	10	43	244	1178	19
KOILADA TIMIOU PRODROMOU-MENOIKION	23	22	43	40	38	57	938	1958	51
OROS BELES	23	7	11	41	17	56	760	2005	14
OROS CHOLOMONTAS	23	31	28	40	26	26	673	1160	294
OROS ITAMOS - SITHONIA	23	50	8	40	8	34	267	807	0
CHERSONISOS ATHOS	24	11	48	40	16	42	329	1920	0
LIMNOTHALASSA AGIOU MAMA	23	20	37	40	14	13	9	42	0
OROS STRATONIKON - KORYFI SKAMNI	23	48	24	40	33	21	288	904	0
AKROTIRIO ELIA - AKROTIRIO KASTRO - EKVOLI RAGOULA	23	42	12	40	10	60	11	50	0
PALIOURI - AKROTIRI	23	39	28	39	58	51	7	30	0
PLATANITSI - SYKIA: AKR. RIGAS - AKR. ADOLO	23	59	58	40	2	41	6	61	0
AKROTIRIO PYRGOS - ORMOS KYPSAS - MALAMO	23	19	10	40	4	12	18	68	0
OROS CHOLOMONTAS	21	50	28	40	54	51	546	1138	102
YGROTOPOI NEAS FOKAIAS	23	19	26	40	6	34	9	42	0
CHERSONISOS SITHONIAS	23	52	37	40	5	4	263	807	0
VASILITSA	21	5	36	40	1	27	1439	2248	775
VALIA KALNTA KAI TECHNITI LIMNI AOOU	21	20	32	39	10	10	1604	2170	1002
ETHNIKOS DRYMOS PINDOU (VALIA KALNTA) - EVRYTERI PERIOCHI	21	7	16	39	54	4	1658	2175	1041

ORI ORLIAKAS KAI TSOURGIAKAS	21	15	27	39	57	14	986	1523	593
LIMNI KASTORIAS	21	17	54	40	31	22	684	814	622
KORYFES OROUS GRAMMOS	20	50	42	40	21	4	1479	2505	619
LIMNI ORESTIAS (KASTORIAS)	21	17	35	40	31	4	689	814	622
OROS VOURINOS (KORYFI ASPROVOUNI)	21	40	18	40	11	48	1226	1531	893
ORI VOREIOU VOURINOU KAI MELLIA	21	39	32	40	11	45	1107	1865	654
ETHNIKOS DRYMOS PRESPON	21	4	38	40	46	20	972	2053	579
ORI VARNOUNTA	21	12	25	40	50	43	1650	2314	900
LIMNES VEGORITIDA - PETRON	21	45	34	40	42	47	577	909	557
LIMNES CHEIMADITIDA - ZAZARI	21	33	44	40	36	13	647	1017	594
OROS VERNON - KORYFI VITSI	21	25	48	40	39	35	1396	2128	845
LIMNI PETRON	21	42	59	40	44	46	743	1192	557
LIMNES ChEIMADITIDA KAI ZAZARI	21	7	44	39	52	57	658	1067	597
PERIOCHI LIMNIS TAVROPOU	21	44	34	39	14	57	809	1791	587
AGRAFA	21	36	8	39	13	46	1454	2160	786
KATO OLYMPOS - KALLIPEFKI	22	32	11	39	55	57	780	1583	12
AISTHITIKO DASOS OSSAS	22	41	24	39	48	12	906	1965	0
KARLA - MAVROVOUNI - KEFALOVRYSO VELESTINOU - NEOCHORI	22	50	34	39	36	18	376	1053	0
AISTHITIKO DASOS KOILADAS TEMPON	22	34	0	39	52	28	171	535	9
OROS MAVROVOUNI	22	40	18	39	54	12	409	1050	0
OROS OSSA	22	41	23	39	49	6	806	1965	0
KATO OLYMPOS, OROS GODAMANI KAI KOILADA RODIAS	22	28	15	39	47	41	660	1583	12
STENA KALAMAKIOU KAI ORI ZARKOU	22	11	58	39	37	8	287	695	65
STENA KALAMAKIOU	22	14	39	39	39	40	157	412	65
PERIOCHI THESSALIKOU KAMPOU	22	25	4	39	26	59	198	725	65
PERIOCHI FARSALON	22	22	39	39	16	58	241	533	113

PERIOCHI TYRNAVOU	22	18	40	39	42	44	106	545	57
PERIOCHI ELASSONAS	22	50	22	39	36	25	257	554	160
DELTA PINEIOU	22	8	33	39	50	38	4	47	0
OROS PILIO KAI PARAKTIA THALASSIA ZONI	23	4	31	39	26	44	725	1604	0
KOURI ALMYROU - AGIOS SERAFEIM	22	44	13	39	11	43	83	96	69
SKIATHOS: KOUKOUNARIES KAI EVRYTERI THALASSIA PERIOCHI	23	24	14	39	8	52	8	32	0
ETHNIKO THALASSIO PARKO ALONNISOU - VOREION SPORADON, ANATOLIKI SKOPELOS	24	2	50	39	15	31	137	564	0
NISIA KYRA PANAGIA, PIPERI, PSATHOURA KAI GYRO NISIDES AGIOS GEORGIOS, NISOI ADELFOI, LECHOUSA, GAIDOURONISIA	24	4	13	39	20	10	95	564	0
OROS OTHRYS, VOUNA GKOURAS KAI FARANGI PALAIOKERASIAS	22	39	35	39	4	16	985	1725	157
PERIOCHI TAMIEFTIRON PROIN LIMNIS KARLAS	22	48	43	39	26	13	94	543	40
OROS PILIO	23	4	58	39	26	6	685	1622	0
ASPROPOTAMOS	21	17	16	39	38	20	1388	2095	805
KERKETIO OROS (KOZIAKAS)	21	28	27	39	34	12	944	2200	117
ANTICHASIA ORI - METEORA	21	44	9	39	43	20	594	1406	101
ANTICHASIA ORI KAI METEORA	21	33	44	40	36	20	604	1420	99
KORYFES OROUS KOZIAKA	21	34	21	39	29	53	754	1899	117
AMVRAKIKOS KOLPOS, DELTA LOUROU KAI ARACHTHOU (PETRA, MYTIKAS, EVRYTERI PERIOCHI)	20	55	24	39	1	29	25	508	0
ORI ATHAMANON (NERAIDA)	21	11	22	39	27	52	1517	2428	614
AMVRAKIKOS KOLPOS, LIMNOTHALASSA KATAFOURKO KAI KORAKONISIA	20	56	3	39	2	24	26	508	0
KOILADA ACHELOOU KAI ORI VALTOU	21	9	4	39	29	10	942	1846	270
EKVOLES (DELTA) KALAMA	20	11	34	39	34	13	35	506	0
ELOS KALODIKI	20	27	40	39	18	38	165	328	139
LIMNI LIMNOPOULA	20	27	2	39	28	44	270	600	217
STENA KALAMA	20	28	26	39	37	42	350	827	94
YGROTOPOS EKVOLON KALAMA KAI NISOS PRASOUDI	20	11	34	39	34	13	35	506	0
ELI KALODIKI, MARGARITI, KARTERI KAI LIMNI PRONTANI	20	26	14	39	19	41	168	329	139

STENA PARAKALAMOU	20	21	22	39	33	42	218	732	0
ORI PARAMYTHIAS, STENA KALAMA KAI STENA ACHERONTA	20	35	53	39	23	3	629	1644	44
ORI TSAMANTA, FILIATON, FARMAKOVOUNI, MEGALI RACHI	20	22	43	39	41	1	633	1803	85
ETHNIKOS DRYMOS VIKOU - AOOU	20	45	43	39	55	29	1290	2465	405
KORYFES OROUS SMOLIKAS	20	54	57	40	4	45	1486	2636	572
KENTRIKO TMIMA ZAGORIOU	20	51	58	39	51	34	1065	1887	589
LIMNI IOANNINON	20	53	8	39	39	35	473	679	469
PERIOCHI METSOVOU (ANILIO - KATARA)	21	12	31	39	47	13	1404	1823	824
OROS LAKMOS (PERISTERI)	21	7	38	39	39	18	1553	2286	657
OROS MITSIKELI	20	50	26	39	45	3	1215	1808	739
OROS TYMFI (GKAMILA)	20	47	24	39	57	56	1425	2477	405
OROS DOUSKON, ORAIOKASTRO, DASOS MEROPIS, KOILADA GORMOU, LIMNI DELVINAKIOU	20	29	49	39	54	41	912	2201	444
KENTRIKO ZAGORI KAI ANATOLIKO TMIMA OROUS MITSIKELI	21	44	47	39	45	43	1039	1872	480
EVRYTERI PERIOCHI POLIS IOANNINON	20	49	58	39	48	48	566	875	180
EVRYTERI PERIOCHI ATHAMANIKON OREON	20	50	32	39	36	38	1129	2424	260
EKVOLES ACHERONTA (APO GLOSSA EOS ALONAKI) KAI STENA ACHERONTA	20	30	5	39	14	1	260	1276	0
PARAKTIA THALASSIA ZONI APO PARGA EOS AKROTIRIO AGIOS THOMAS (PREVEZA), AKR. KELADIO - AG. THOMAS	20	28	30	39	12	55	5	71	0
DYTIKES KAI VOREIOANATOLIKES AKTES ZAKYNTHOU	20	43	47	37	42	27	159	463	0
KOLPOS LAGANA ZAKYNTHOU (AKR. GERAKI - KERI) KAI NISIDES MARATHONISI KAI PELOUZO	20	54	25	37	42	33	22	240	0
NISOI STROFADES	21	0	34	37	15	18	4	24	1
NISIDES STAMFANI KAI ARPYIA (STROFADES)	21	0	34	37	14	51	4	24	0
KALON OROS KEFALONIAS	20	34	34	38	20	32	598	901	181
ETHNIKOS DRYMOS AINOU	20	39	44	38	8	42	1091	1624	286
ESOTERIKO ARCHIPELAGOS IONIOU (MEGANISI, ARKOUDI, ATOKOS, VROMONAS)	20	50	13	38	34	46	20	281	0
PARAKTIA THALASSIA ZONI APO ARGOSTOLI EOS VLACHATA (KEFALONIA) KAI ORMOS MOUNTA	20	33	41	38	4	56	7	47	0

DYTIKES AKTES KEFALONIAS - STENO KEFALONIAS ITHAKIS - VOREIA ITHAKI (AKROTIRIA GERO GKOMPOS - DRAKOU PIDIMA - KENTRI - AG. IOANNIS)	20	29	20	38	20	53	24	186	0
KEFALONIA: AINOS, AGIA DYNATI KAI KALON OROS	19	32	47	39	45	51	674	1688	0
LIMNOTHALASSA ANTINIOTI (KERKYRA)	19	51	2	39	48	55	1	12	0
LIMNOTHALASSA KORISSION (KERKYRA)	19	55	5	39	26	37	36	457	0
ALYKI LEFKIMMIS (KERKYRA)	20	4	7	39	27	7	0	4	0
NISOI PAXOI KAI ANTIPAXOI	20	11	26	39	11	30	17	220	0
PARAKTIA THALASSIA ZONI APO KANONI EOS MESONGI (KERKYRA)	19	55	14	39	32	32	15	112	0
LIMNOTHALASSA KORISSION (KERKYRA) KAI NISOS LAGOUDIA	19	54	29	39	26	40	10	62	0
DIAPONTIA NISIA (OTHONOI, EREIKOUSA, MATHRAKI KAI VRACHONISIDES)	20	36	60	38	10	51	82	379	0
LIMNOTHALASSES STENON LEFKADAS (PALIONIS - AVLIMON) KAI ALYKES LEFKADAS	20	43	6	38	48	8	2	58	0
PERIOCHI CHORTATON (LEFKADA)	20	37	32	38	41	59	815	1160	257
DELTA ACHELOOU, LIMNOTHALASSA MESOLONGIOU - AITOLIKOU, EKVOLES EVINOU, NISOI ECHINADES, NISOS PETALAS	21	15	14	38	19	53	23	409	0
OROS PANAITOLIKO	21	39	23	38	42	8	1242	1921	557
OROS VARASOVA	21	35	50	38	21	40	363	903	0
LIMNES VOULKARIA KAI SALTINI	20	48	39	38	51	52	11	72	0
LIMNI AMVRAKIA	21	10	41	38	45	2	49	230	30
LIMNI OZEROS	21	13	23	38	39	16	70	222	23
LIMNES TRICHONIDA KAI LYSIMACHEIA	21	29	3	38	34	6	22	312	7
OROS ARAKYNTHOS KAI STENA KLEISOURAS	21	27	34	38	27	34	491	913	78
OROS TSEREKAS (AKARNANIKA)	20	55	3	38	44	8	496	1141	0
LIMNI LYSIMACHEIA	21	21	50	38	33	60	18	45	9
LIMNI VOULKARIA	20	50	27	38	51	35	55	432	0
DELTA ACHELOOU, LIMNOTHALASSA MESOLONGIOU - AITOLIKOU KAI EKVOLES EVINOU, NISOI ECHINADES, NISOS PETALAS, DYTIKOS ARAKYNTHOS KAI STENA KLEISOURAS	21	15	14	38	19	53	158	913	0
LIMNI AMVRAKIA	21	10	41	38	45	8	25	205	20

LIMNOTHALASSA KALOGRIAS, DASOS STROFYLIAS KAI ELOS LAMIAS, ARAXOS	21	21	48	38	5	37	5	53	0
OROS CHELMOS KAI YDATA STYGOS	22	13	31	37	57	28	1413	2337	600
FARANGI VOURAIKOU	22	10	24	38	4	59	573	1128	65
AISTHITIKO DASOS KALAVRYTON	22	5	51	38	0	46	957	1560	712
ORI BARMPAS KAI KLOKOS, FARANGI SELINOUNTA	22	0	57	38	8	36	903	1776	143
ALYKI AIGIOU	22	6	26	38	15	48	0	2	0
OROS PANACHAIKO	21	52	41	38	12	30	1224	1926	392
OROS ERYMANTHOS	21	52	15	37	58	6	1290	2208	586
SPILAIO KASTRION	22	8	30	37	56	59	767	979	660
ORI BARMPAS, KLOKOS, FARANGI SELINOUNTA	22	0	10	38	8	27	930	1771	140
YGROTOPOI KALOGRIAS-LAMIAS KAI DASOS STROFYLIAS	22	11	35	37	59	3	15	243	0
OROS ERYMANTHOS	21	50	38	37	56	44	1132	2206	357
OROS CHELMOS (AROANIA) - FARANGI VOURAIKOU KAI PERIOCHI KALAVRYTON	21	22	58	38	5	8	1228	2328	40
OROPEDIO FOLOIS	21	41	29	37	47	7	557	760	158
EKVOLES (DELTA) PINEIOU	21	14	6	37	49	8	6	18	0
OLYMPIA	21	37	38	37	38	28	74	159	30
THINES KAI PARALIAKO DASOS ZACHAROS, LIMNI KAIAFA, STROFYLIA, KAKOVATOS	21	35	20	37	31	37	27	444	0
LIMNOTHALASSA KOTYCHI, BRINIA	21	17	55	38	0	13	2	16	0
PARAKTIA THALASSIA ZONI APO AKR. KYLLINI EOS TOUMPI - KALOGRIA	21	17	18	38	1	57	45	239	0
THALASSIA PERIOCHI KOLPOU KYPARISSIAS: AKR. KATAKOLO - KYPARISSIA	21	30	8	37	34	32	1	15	0
LIMNOTHALASSA KOTYCHI - ALYKI LECHAINON	21	18	20	38	0	53	5	27	0
LIMNES YLIKI KAI PARALIMNI - SYSTIMA VOIOTIKOU KIFISOU	23	10	51	38	25	50	104	413	54
OROS PARNASSOS	24	30	33	37	56	4	1280	2446	289
OROS OCHI - KAMPOS KARYSTOU - POTAMI - AKROTIRIO KAFIREFS - PARAKTIA THALASSIA ZONI	24	30	13	38	1	57	564	1394	0
DIRFYS: DASOS STENIS - DELFI	23	50	50	38	35	49	1015	1740	432
MEGALO KAI MIKRO LIVARI - DELTA XERIA - YDROCHARES DASOS AG. NIKOLAOU - PARAKTIA THALASSIA ZONI	23	7	28	39	0	5	1	5	0

SKYROS: OROS KOCHYLAS	24	37	16	38	49	27	399	780	0
MEGALO KAI MIKRO LIVARI - DELTA XERIA	23	7	9	39	0	12	1	5	0
LIMNI DYSTOS	24	6	58	38	20	25	200	608	9
NISIDES SKYROU	24	21	28	38	50	13	61	170	1
OROS KANTILI	22	22	29	38	52	18	531	1220	0
ORI KENTRIKIS EVVOIAS, PARAKTIA ZONI KAI NISIDES	22	7	14	38	39	21	690	1738	0
OROS OCHI, PARAKTIA ZONI KAI NISIDES	22	21	51	38	24	30	485	1386	0
OROS TYMFRISTOS (VELOUCHI)	21	48	49	38	56	35	1580	2213	978
ORI AGRAFA	22	32	48	38	32	30	1312	2143	400
KOILADA KAI EKVOLES SPERCHEIOU - MALIAKOS KOLPOS	22	25	31	38	51	47	57	401	0
FARANGI GORGOPOTAMOU	22	21	33	38	49	7	711	1560	175
ETHNIKOS DRYMOS OITIS	22	17	39	38	50	17	1348	2103	239
KATO ROUS KAI EKVOLES SPERCHEIOU POTAMOU	23	25	59	38	42	21	31	262	0
OROS KALLIDROMO	22	33	10	38	46	4	834	1393	43
ETHNIKOS DRYMOS OITIS - KOILADA ASOPOU	22	20	11	38	49	34	968	2103	52
ORI VARDOUSIA	22	6	43	38	39	42	1374	2427	456
OROS GKIONA	22	17	41	38	36	19	1512	2456	558
PARALIAKI ZONI APO NAFPAKTO EOS ITEA	22	8	17	38	20	52	109	624	0
NOTIOANATOLIKOS PARNASSOS - ETHNIKOS DRYMOS PARNASSOU - DASOS TITHOREAS	22	33	32	38	32	9	1434	2427	360
KORYFES OROUS GKIONA, CHARADRA REKA, LAZOREMA KAI VATHIA LAKKA	22	16	24	38	38	1	1638	2456	558
OROS VARDOUSIA	23	50	47	38	37	19	1278	2440	440
EVRYTERI PERIOCHI GALAXEIDIOU	21	37	30	39	10	43	329	892	0
AKRONAFPLIA KAI PALAMIDI	22	48	24	37	33	42	69	219	0
ORI ARTEMISIO KAI LYRKEIO	22	27	9	37	56	33	1101	1793	580
OROS MAINALO	22	17	40	37	36	19	1179	1978	634
LIMNI TAKA	22	22	8	37	25	52	679	855	667

LIMNOTHALASSA MOUSTOU	22	45	8	37	23	0	7	78	0
MONI ELONAS KAI CHARADRA LEONIDIOU	22	49	32	37	9	15	474	1204	0
OROS PARNONAS (KAI PERIOCHI MALEVIS)	22	37	54	37	13	26	1057	1920	99
KORYFES OROUS KYLLINI (ZIRIA) KAI CHARADRA FLAMPOURITSA	22	27	29	37	57	4	1198	2364	136
LIMNI STYMFALIA	22	27	32	37	51	19	611	720	599
AKROKORINTHOS	22	51	56	37	52	50	293	566	135
OROS OLIGYRTOS	22	22	40	37	49	9	1189	1940	639
ORI GERANEIA	23	4	15	38	1	14	765	1367	295
OROS ZIREIA (KYLLINI)	22	27	19	37	36	23	1358	2365	558
ORI GIDOVOUNI, CHIONOVOUNI, GAIDOUROVOUNI, KORAKIA, KALOGEROVOUNI, KOULOCHERA KAI PERIOCHI MONEMVASIAS	22	57	30	36	54	4	562	1274	0
PERIOCHI NEAPOLIS KAI NISOS ELAFONISOS	23	0	3	36	32	41	61	521	0
EKVOLES EVROTA	22	41	21	36	49	50	12	120	0
LAGKADA TRYPIS	22	18	51	37	5	23	1015	1777	343
YGROTOPOI EKVOLON EVROTA	22	41	22	36	48	55	6	26	0
ORI ANATOLIKIS LAKONIAS	22	57	30	36	54	4	504	1274	0
NOTIA MANI	22	26	57	36	34	28	321	1203	0
FARANGI NEDONA (PETALON - CHANI)	22	9	44	37	5	24	431	889	77
NISOI SAPIENTZA KAI SCHIZA, AKROTIRIO AKRITAS	21	49	4	36	49	50	129	517	0
LIMNOTHALASSA PYLOU (DIVARI) KAI NISOS SFAKTIRIA, AGIOS DIMITRIOS	21	40	21	36	56	57	17	126	0
THINES KYPARISSIAS (NEOCHORI - KYPARISSIA)	21	40	43	37	16	6	8	88	0
OROS TAYGETOS	22	18	44	36	57	42	1056	2827	266
THALASSIA PERIOCHI STENOU METHONIS	21	43	37	36	47	35	7	68	0
LIMNOTHALASSA GIALOVAS KAI NISOS SFAKTIRIA	21	40	28	36	57	54	20	126	0
OROS TAYGETOS - LAGKADA TRYPIS	22	19	13	36	58	32	1046	2827	28
OROS PARNITHA	23	43	44	38	10	37	752	1400	234

ETHNIKO PARKO SCHINIA - MARATHONA	24	1	51	38	8	57	28	241	0
VRAVRONA - PARAKTIA THALASSIA ZONI	23	59	54	37	55	8	50	254	0
SOUNIO - NISIDA PATROKLOU KAI PARAKTIA THALASSIA ZONI	23	59	39	37	39	35	89	305	0
YMITTOS - AISTHITIKO DASOS KAISARIANIS - LIMNI VOULIAGMENIS	23	48	10	37	53	53	376	1025	0
ANTIKYTHIRA - PRASONISI KAI LAGOUVARDOS	23	16	8	35	54	60	118	360	0
NISIDES KYTHIRON: PRASONISI, DRAGONERA, ANTIDRAGONERA	23	6	13	36	13	22	11	29	0
NISIDES MYRTOOU PELAGOUS: FALKONERA, VELOPOULA, ANANES	23	27	36	36	54	57	36	200	0
NISOS ANTIKYTHIRA KAI NISIDES PRASONISI, LAGOUVARDOS, PLAKOULITHRA KAI NISIDES THYMONIES	23	18	2	35	52	5	124	360	0
KYTHIRA KAI GYRO NISIDES: PRASONISI, DRAGONERA, ANTIDRAGONERA, AVGO, KAPELLO, KOUFO KAI FIDONISI	23	3	1	36	16	15	186	456	0
PERIOCHI LEGRENON - NISIDA PATROKLOU	23	58	9	37	41	24	124	306	0
OROS YMITTOS	24	0	58	38	8	52	418	1023	64
YGROTOPOS SCHINIA	23	48	21	37	54	42	55	300	0
LIMNOS: CHORTAROLIMNI - LIMNI ALYKI KAI THALASSIA PERIOCHI	25	27	47	39	57	5	7	101	0
AGIOS EFSTRATIOS KAI PARAKTIA THALASSIA ZONI	25	1	14	39	31	25	101	200	0
LESVOS: DYTIKI CHERSONISOS - APOLITHOMENO DASOS	25	58	36	39	11	47	224	710	0
LESVOS: KOLPOS KALLONIS KAI CHERSAIA PARAKTIA ZONI	26	12	43	39	9	46	50	299	0
LESVOS: KOLPOS GERAS, ELOS NTIPI KAI OROS OLYMPOS	26	25	23	39	4	54	487	900	0
LIMNOS: LIMNES CHORTAROLIMNI KAI ALYKI, KOLPOS MOUDROU, ELOS DIAPORI KAI CHERSONISOS FAKOS	25	17	25	39	54	22	36	315	0
LESVOS: PARAKTIOI YGROTOPOI KOLPOU KALLONIS	26	29	56	39	4	55	18	180	0
NISIDES KAI VRACHONISIDES LIMNOU: NISOS SERGITSI KAI NISIDES DIAVATES, KOMPIO, KASTRIA, TIGANI, KARKALAS, PRASONISI	25	8	5	40	1	18	41	100	0
NISIDES LESVOU (SYMPLEGMA TOMARONISION, KYDONAS, AGIOS GEORGIOS, GLARONISI, KLP)	26	26	15	39	18	28	4	55	0
NOTIODYTIKI CHERSONISOS, APOLITHOMENO DASOS LESVOU	25	59	8	39	10	10	204	610	0
OROS OLYMPOS LESVOU	26	19	40	39	5	59	279	900	30
VOREIA LESVOS	26	16	5	39	20	41	305	954	0
LESVOS: KOLPOS GERAS, ELI NTIPI KAI CHARAMIDA	25	1	1	39	31	30	16	60	0

NISOS AGIOS EFSTRATIOS KAI THALASSIA ZONI	26	13	59	39	8	9	113	296	0
SAMOS: PARALIA ALYKI	27	1	8	37	42	22	25	112	0
SAMOS: OROS AMPELOS (KARVOUNIS)	26	49	12	37	45	33	747	1209	142
SAMOS: OROS KERKETEFS - MIKRO KAI MEGALO SEITANI - DASOS KASTANIAS KAI LEKKAS, AKR. KATAVASIS - LIMENAS	26	38	2	37	44	32	517	1400	0
IKARIA - FOURNOI KAI PARAKTIA ZONI	26	28	57	37	34	48	331	1000	0
NISOS IKARIA (NOTIODYTIKO TMIMA)	26	2	34	37	33	49	482	1000	0
NISOS FOURNOI KAI NISIDES THYMAINA, ALATSONISI, THYMAINAKI, STRONGYLO, PLAKA, MAKRONISI, MIKROS KAI MEGALOS ANTHROPOFAGOS, AGIOS MINAS	26	30	26	37	32	15	110	488	0
SAMOS: ALYKI PSILIS AMMOU	27	0	37	37	42	28	15	39	0
SAMOS: OROS KERKIS	26	0	54	38	7	33	426	1426	0
VOREIA CHIOS KAI NISOI OINOUSSES KAI PARAKTIA THALASSIA ZONI	26	4	28	38	31	5	388	1200	0
NISIA ANTIPSARA KAI NISIDES DASKALIO, MASTROGIORGI, PRASONISI, KATO NISI, MESIAKO, KOUTSOULIA	25	30	41	38	32	31	59	380	0
VOREIA CHIOS	26	38	45	37	44	25	484	1284	0
NISIDA VENETIKO	25	59	0	38	31	41	23	67	19
VRACHONISIDES KALOGEROI KAI THALASSIA ZONI							-100	0	-200
KASOS KAI KASONISIA - EVRYTERI THALASSIA PERIOCHI	26	55	4	35	24	24	205	600	0
KENTRIKI KARPATHOS: KALI LIMNI - LASTOS - KYRA PANAGIA KAI PARAKTIA THALASSIA ZONI	27	8	17	35	35	34	420	1200	0
VOREIA KARPATHOS KAI SARIA KAI PARAKTIA THALASSIA ZONI	27	12	3	35	48	53	217	705	0
KASTELLORIZO KAI NISIDES RO KAI STRONGYLI KAI PARAKTIA THALASSIA ZONI	29	34	59	36	8	54	42	157	1
RODOS: AKRAMYTIS, ARMENISTIS, ATTAVYROS, REMATA KAI THALASSIA ZONI (KARAVOLA-ORMOS GLYFADA)	27	50	15	36	9	40	361	1200	0
RODOS: PROFITIS ILIAS - EPTA PIGES – PETALOUDES - REMATA	28	1	42	36	18	15	240	706	51
NOTIA NISYROS KAI STRONGYLI KAI PARAKTIA THALASSIA ZONI	27	10	6	36	34	44	215	600	0
KOS: AKROTIRIO LOUROS - LIMNI PSALIDI - OROS DIKAIOS - ALYKI - PARAKTIA THALASSIA ZONI	27	15	27	36	50	53	223	800	0
ASTYPALAIA: ANATOLIKO TMIMA, GYRO NISIDES KAI OFIDOUSSA KAI THALASSIA ZONI (AKR. LANTRA - AKR. VRYSI)	26	25	39	36	35	26	83	317	0
ARKOI, LEIPSOI, AGATHONISI KAI VRACHONISIDES	26	45	24	37	17	34	36	200	0

VRACHONISIA NOTIOU AIGAIOU: VELOPOULA, FALKONERA, ANANES, CHRISTIANA, PACHEIA, FTENO, MAKRA, ASTAKIDONISIA, SYRNA - GYRO NISIA KAI THALASSIA ZONI	26	40	33	36	20	43	66	322	0
NISIDES PATMOU: PETROKARAVO, ANYDROS	26	29	34	37	24	35	30	54	0
NISOS AGATHONISIOU KAI NISIDES: PITTA, KATSAGANI, NERONISI, STRONGYLI	26	57	58	37	27	57	57	200	0
NISOS LEIPSOI (DYTIKO TMIMA) KAI NISIDES: FRAGKOS, MAKRONISI, PILAFI, KAPARI, KALAPODIA, MEGALO ASPRONISI, MAKRY ASPRONISI, KOULOURA, NOTIA ASPRA, SARAKINA, PIATO, PSOMOS, STAVRI, LIRA, ARETHOUSA, MANOLI	26	43	49	37	18	49	73	200	0
VOREIODYTIKO TMIMA ARKION KAI NISIDES: AGRELOUSA, STRONGYLI, SPALATHI, SMINERO, TSOUKA, TSOUKAKI, PSATHONISI, KALOVOLOS, MAKRONISI, AVAPTISTOS, KOMAROS	26	43	16	37	23	43	17	100	0
NISIDES LEROU: PIGANOUSA, MEGALO GLARONISI, MIKRO GLARONISI, LERIKO	26	54	20	37	6	57	20	100	0
NISIDES KALYMNOU: EPANO, NERA, SARI, TELENDOS	26	54	15	37	0	36	135	400	0
NISOI KINAROS KAI LEVITHA KAI NISIDES LIADIA, PLAKA, GLAROS, MAVRA	26	27	29	37	0	29	54	280	0
ANATOLIKO TMIMA ASTYPALAIAS KAI NISIDES KOUNOUPOI, FTENO, CHONDROPOULO, KOUTSOMYTIS, MONI, AGIA KYRIAKI, TIGANI, CHONDRI, LIGNO, FOKIONISIA, KATSAGRELI, PONTIKOUSSA, OFIDOUSSA, KTENIA	26	25	29	36	37	24	54	248	0
NISOS SYRNA KAI NISIDES MEGALOS ADELFOS, MIKROS ADELFOS, KATSIKAS, MESONISI, PLAKIDA, STEFANIA, NAVAGIO	26	40	33	36	20	51	113	322	0
NISIDES KARPATHIOU PELAGOUS: MEGALO SOFRANO, SOCHAS, MIKRO SOFRANO, AVGO, DIVOUNIA, CHAMILI, ASTAKIDONISIA	26	23	58	36	4	25	18	149	0
NISOS TILOS KAI NISIDES: ANTITILOS, PELEKOUSA, GAIDOURONISI, GIAKOUMIS, AGIOS ANDREAS, PRASOUDA, NISI	27	22	19	36	25	51	166	600	0
ANATOLIKO TMIMA NISOU SYMIS KAI NISIDES KOULOUNDROS, SESKLI, TROUMPETO, MARMARAS, KARAVALONISI, MEGALONISI, GIALESINO, OXEIA, CHONDROS, PLATY, NIMOS	27	51	12	36	33	3	150	508	0
NISOS CHALKI KAI NISIDES: KOLOFONA, PANO PRASOUDA, TRAGOUSA, STRONGYLI, AGIOS THEODOROS, MAELONISI, ALIMIA, KREVVATI, NISAKI	27	34	15	36	13	52	173	600	0
KOS: LIMNI PSALIDI - ALYKI	27	10	8	36	52	57	7	27	0
NISOS KASOS KAI SYMPLEGMA KASONISION	26	55	45	35	22	59	235	600	0
ANATOLIKI RODOS: PROFITIS ILIAS - EPTA PIGES - EKVOLI LOUTANI - KATERGO, REMA GADOURA - CHERSONISOS LINDOU - NISIDES PENTANISA KAI TETRAPOLIS, LOFOS PSALIDI	28	2	13	36	15	49	208	705	0
DYTIKI RODOS: ORI ATTAVYROS & AKRAMYTIS, TECHNITI LIMNI APOLAKKIAS KAI NISIDES GEORGIOU, STRONGYLI, CHTENIES & KARAVOLAS	27	47	26	36	8	2	420	1200	0

NOTIO AKRO RODOU, PRASONISI, YGROTOPOS LIVADI KATTAVIAS	27	46	8	35	53	2	75	206	0
NISOS NISYROS KAI NISIDES	24	54	48	37	28	24	189	680	0
ANDROS: ORMOS VITALI KAI KENTRIKOS OREINOS OGKOS	24	51	14	37	51	56	402	960	0
ANAFI: CHERSONISOS KALAMOS - ROUKOUNAS	25	49	24	36	21	4	138	459	0
SANTORINI: NEA KAI PALIA KAMENI - PROFITIS ILIAS	25	27	45	36	22	18	162	561	0
FOLEGANDROS ANATOLIKI MECHRI DYTIKI SIKINO KAI THALASSIA ZONI	24	59	8	36	36	44	154	520	0
PARAKTIA ZONI DYTIKIS MILOY	24	23	57	36	39	11	6	51	0
NISOS POLYAIGOS - KIMOLOS	24	35	37	36	46	45	128	339	0
NISOS ANTIMILOS - THALASSIA PARAKTIA ZONI	24	14	18	36	47	18	261	600	0
SIFNOS: PROFITIS ILIAS MECHRI DYTIKES AKTES KAI THALASSIA PERIOCHI	24	41	17	36	57	33	283	560	0
NOTIA SERIFOS	24	27	18	37	7	51	239	560	0
VOREIODYTIKI KYTHNOS: OROS ATHERAS - AKROTIRIO KEFALOS KAI PARAKTIA ZONI	24	24	1	37	26	15	128	320	0
ANATOLIKI KEA	24	20	20	37	35	45	257	560	0
VOREIA AMORGOS KAI KINAROS, LEVITHA, MAVRA, GLAROS KAI THALASSIA ZONI	26	1	58	36	54	34	227	820	0
MIKRES KYKLADES: IRAKLEIA, SCHOINOUSSA, KOUFONISIA, KEROS, ANTIKERIA KAI THALASSIA ZONI	25	35	23	36	55	34	86	400	0
KENTRIKI KAI NOTIA NAXOS: ZAS KAI VIGLA EOS MAVROVOUNI KAI THALASSIA ZONI (ORMOS KARADES - ORMOS MOUTSOUNAS)	25	27	39	36	56	51	332	997	0
NISOS PAROS: PETALOUDES	25	7	28	37	2	53	91	205	36
NISOI DESPOTIKO KAI STRONGYLO KAI THALASSIA ZONI	24	59	39	36	57	46	45	180	0
SYROS: OROS SYRINGAS EOS PARALIA	24	54	35	37	28	31	164	400	0
TINOS: MYRSINI - AKROTIRIO LIVADA	25	13	40	37	35	3	245	680	0
NISOS MILOS: PROFITIS ILIAS - EVRYTERI PERIOCHI	24	24	7	36	41	20	158	720	0
NISOS IRAKLEIA, NISOI MAKARES, MIKROS KAI MEGALOS AVELAS, NISIDA VENETIKO IRAKLEIAS	25	26	53	36	50	15	108	400	0
NISOI CHRISTIANA	25	12	12	36	15	9	27	168	0
ANAFI: ANATOLIKO KAI VOREIO TMIMA KAI GYRO NISIDES	25	9	35	37	37	35	127	463	0
NISOS AMORGOS (VOREIOANATOLIKO TMIMA) KAI NISIDES: PSALIDA, GRAMVOUSSA, NIKOURIA, MIKRO KAI MEGALO VIOKASTRO, KRAMVONISI, PETALIDI	26	1	21	36	54	44	320	820	0

NISIDES PAROU KAI NOTIA ANTIPAROS	25	0	17	36	57	38	34	200	0
NAXOS: ORI ANATHEMATISTRA, KORONOS, MAVROVOUNI, ZAS, VIGLATOURI	25	30	41	37	1	43	361	997	0
NISIDES MYKONOU (RINEIA, CHTAPODIA, TRAGONISI)	25	13	37	37	22	44	30	160	0
ANDROS: KENTRIKO KAI NOTIO TMIMA, GYRO NISIDES KAI PARAKTIA THALASSIA ZONI	27	10	6	36	35	24	340	993	0
SERIFOS: PARAKTIA ZONI KAI NISIDES SERIFOPOULA, PIPERI KAI VOUS	24	30	1	37	6	28	56	210	0
DYTIKI MILOS, ANTIMILOS, POLYAIGOS KAI NISIDES	24	23	55	36	39	24	149	743	0
VOREIOANATOLIKI TINOS KAI NISIDES	25	50	32	36	21	3	220	717	0
VOREIA SYROS KAI NISIDES	24	50	45	37	52	24	161	438	0
NISOS GYAROS KAI THALASSIA ZONI									
GIOUCHTAS - FARANGI AGIAS EIRINIS	25	9	4	35	13	34	470	799	90
NISOS DIA	25	12	58	35	27	13	91	255	0
DYTIKA ASTEROUSIA (APO AGIOFARANGO EOS KOKKINO PYRGO)	24	45	49	34	57	22	115	326	0
ASTEROUSIA (KOFINAS)	25	6	40	34	58	3	494	1213	0
DIKTI: OMALOS VIANNOU (SYMI - OMALOS)	25	26	3	35	4	7	973	1506	463
KROUSONAS - VROMONERO IDIS	24	55	9	35	11	16	1274	1908	516
OROS GIOUCHTAS	25	8	34	35	13	55	576	799	287
KORYFI KOUPA (DYTIKI KRITI)	25	23	2	35	4	15	735	1189	358
EKVOLI GEROPOTAMOU MESARAS	24	45	50	35	2	57	13	92	0
ASTEROUSIA ORI (KOFINAS)	25	7	4	34	58	39	411	1195	0
DIKTI: OROPEDIO LASITHIOU, KATHARO, SELENA, KRASI, SELAKANO, CHALASMENI KORYFI	25	29	30	35	9	12	1115	2146	280
NISOS CHRYSI	25	42	26	34	52	18	2	20	0
MONI KAPSA (FARANGI KAPSA KAI GYRO PERIOCHI)	26	3	33	35	2	31	380	706	0
OROS THRYPTIS KAI GYRO PERIOCHI	25	53	18	35	4	33	590	1473	0
VOREIOANATOLIKO AKRO KRITIS: DIONYSADES, ELASA KAI CHERSONISOS SIDERO (AKRA MAVRO MOURI - VAI - AKRA PLAKAS) KAI THALASSIA ZONI	26	15	6	35	14	9	75	254	0
NISOS KOUFONISI KAI PARAKTIA THALASSIA ZONI	26	8	24	34	56	8	18	79	0

VOREIOANATOLIKO AKRO KRITIS	26	14	51	35	15	29	63	213	0
LAZAROS KORYFI - MADARA DIKTIS	25	32	2	35	3	51	1351	2146	478
DIONYSADES NISOI	26	10	32	35	20	45	49	129	0
FARANGI SELINARI - VRACHASI	25	33	10	35	16	9	331	813	59
NOTIODYTIKI THRYPTI (KOUFOTO)	25	49	32	35	3	8	601	1008	327
ORI ZAKROU	26	13	24	35	3	41	308	802	0
NISOS KOUFONISI, GYRO NISIDES KAI NISIDES KAVALLOI	26	8	26	34	56	15	26	71	0
OROS KEDROS	24	36	45	35	11	20	965	1775	525
KOURTALIOTIKO FARANGI - MONI PREVELI - EVRYTERI PERIOCHI	24	28	51	35	11	40	327	902	0
PRASSANO FARANGI - PATSOS - SFAKORYAKO REMA - PARALIA RETHYMNOU KAI EKVOLI GEROPOTAMOU, AKR. LIANOS KAVOS - PERIVOLIA	24	35	28	35	15	48	399	1005	0
OROS IDI (VORIZIA, GERANOI, KALI MADARA)	24	49	25	35	13	33	1201	2452	281
SOROS - AGKATHI - KEDROS	24	35	41	35	12	56	854	1775	366
KOURTALIOTIKO FARANGI, FARANGI PREVELI	24	28	32	35	11	29	424	981	0
PRASSANO FARANGI	24	32	38	35	19	32	247	469	40
OROS PSILOREITIS (NOTIODYTIKO TMIMA)	24	44	52	35	9	56	1421	2452	413
IMERI KAI AGRIA GRAMVOUSSA - TIGANI KAI FALASARNA - PONTIKONISI, ORMOS LIVADI - VIGLIA	23	35	25	35	33	51	197	720	0
NISOS ELAFONISOS KAI PARAKTIA THALASSIA ZONI	23	31	50	35	16	12	1	26	0
CHERSONISOS RODOPOU - PARALIA MALEME	23	47	42	35	32	20	289	742	0
ELOS - TOPOLIA - SASALOS - AGIOS DIKAIOS	23	39	3	35	22	38	596	1176	89
ORMOS SOUGIAS - VARDIA - FARANGI LISSOU MECHRI ANYDROUS KAI PARAKTIA ZONI	23	45	49	35	15	13	331	765	0
LIMNI AGIAS - PLATANIAS - REMA KAI EKVOLI KERITI - KOILADA FASA	23	55	2	35	25	28	131	414	0
FARANGI THERISSOU	23	59	17	35	25	32	408	704	100
LEFKA ORI KAI PARAKTIA ZONI	24	0	26	35	17	58	1216	2448	0
DRAPANO (VOREIOANATOLIKES AKTES) - PARALIA GEORGIOUPOLIS - LIMNI KOURNA	24	17	1	35	20	11	125	508	0
FRE - TZITZIFES - NIPOS	24	8	38	35	22	1	414	779	151

ASFENDOU - KALLIKRATIS KAI PARAKTIA ZONI	24	16	42	35	12	21	653	1510	0
NISOI GAVDOS KAI GAVDOPOULA	24	4	46	34	50	49	60	362	0
ETHNIKOS DRYMOS SAMARIAS - FARANGI TRYPITIS - PSILAFI - KOUSTOGERAKO	23	54	57	35	16	13	1014	2105	0
PARALIA APO CHRYSOSKALITISSA MECHRI AKROTIRIO KRIOS	23	33	11	35	16	39	33	159	0
METERIZIA AGIOS DIKAIOS - TSOUNARA - VITSILIA LEFKON OREON	23	36	24	35	19	50	587	1176	4
CHERSONISOS GRAMVOUSSAS KAI NISIDES IMERI KAI AGRIA GRAMVOUSSA, PONTIKONISI	23	35	20	35	32	59	228	720	0
NISIDA AGIOI THEODOROI	23	55	56	35	32	16	61	152	0
FARANGI KALLIKRATIS - ARGOULIANO FARANGI - OROPEDIO MANIKA	24	15	58	35	13	19	727	1260	100
LIMNI AGIAS (CHANIA)	23	56	14	35	28	49	54	78	43
CHERSONISOS RODOPOU	23	44	22	35	39	40	273	741	0
LIMNI KOURNA KAI EKVOLI ALMYROU	24	16	37	35	19	58	27	146	0
NOTIODYTIKI GAVDOS KAI GAVDOPOULA	24	5	6	34	49	24	39	245	0