8 Geographic Data Modeling

OVERVIEW

Describes the process of data modeling and the various data models that have been used in GIS.

LEARNING OBJECTIVES

By the end of this chapter students should:

- **Define what geographic data models are and discuss their importance in GIS;**
- **Understand how to undertake GIS data modeling;**
- **Outline the main geographic models used in GIS today and their strengths and weaknesses;**
- **Understand key topology concepts and why topology is useful for data validation, analysis, and editing;**
- **Read data model notation;**
- **Describe how to model the world and create a useful geographic database.**

KEY WORDS AND CONCEPTS

Topological relationships, planar enforcement, georelational model, linear referencing, dynamic segmentation, object classes, topology rules, encapsulation, inheritance, polymorphism, UML, CASE tools, database schema

OUTLINE

CHAPTER SUMMARY

8.1 Introduction

- Focuses on how geographic reality is modeled
- Differentiates between
    - *Representation* which can be considered to denote the conceptual and scientific issues
    - *Model* which is used in practical and database concepts
    - *Data model* used to distinguish it from process models

8.1.1 Data model overview

- Data model is a set of constructs for representing objects and processes in the digital environment of the computer
- Decisions about the type of data model to be adopted are vital to the success of a GIS project.

8.1.2 Levels of data model abstraction

- Four different levels of abstraction (levels of generalization or simplification) are
    - Reality is made up of real-world phenomena and includes all aspects that may or may not be perceived by individuals, or deemed relevant to a particular application.
    - The conceptual model is a human-oriented, often partially structured, model of selected objects and processes that are thought relevant to a particular problem domain.
    - The logical model is an implementation-oriented representation of reality that is often expressed in the form of diagrams and lists.
    - The physical model portrays the actual implementation in a GIS, and often comprises tables stored as files or databases

- The conceptual modeling phase begins with definition of the main types of objects to be represented in the GIS and concludes with a conceptual description of the main types of objects and relationships between them.
- The logical modeling phase leads to the creation of diagrams and lists describing the names of objects, their behavior, and the type of interaction between objects.
- The physical modeling phase involves describing the exact files or database tables used to store the data, the relationships between objects types, and the precise operations that can be performed.
- A data model provides
    - Developers with the means to represent an application domain in terms that may be translated into a design and implemented
    - Users with a description of the structure of the system, independent of specific items of data or details of the particular application

8.2 GIS data models

- The key types of geographic data models and their main areas of application are listed in Table 8.1.
- A collection of entities of the same geometric type (dimensionality) is referred to as a class or layer.
- It should also be noted that the term layer is quite widely used in GIS as a general term for a specific dataset.

8.2.1 CAD, graphical, and image GIS data models

- In a CAD system, real-world entities are represented symbolically as simple point, line, and polygon vectors.
- There are three severe problems with CAD models
    - Use local drawing coordinates
    - Individual objects do not have unique identifiers
    - Are focused on graphical representation of objects, cannot store relationships
- Computer cartography (graphical) data models store entities as points, lines and polygons, with annotation used for placenames
- Image data models store scanned aerial air photos and digital satellite images as rasters or grids

8.2.2 Raster data model

- The raster data model uses an array of cells, or pixels, to represent real-world objects

- The cells can hold any attribute values based on one of several encoding schemes including categories, and integer and floating-point numbers
- In some systems, multiple attributes can be stored
- Techniques for compressing rasters are described in Box 8.1, includes run-length encoding, block encoding and wavelet

8.2.3 Vector data model

- In the vector data model each object in the real world is first classified into a geometric type: in the 2-D case point, line, or polygon (Figure 8.7).
- Points are recoded as single coordinate pairs
- Lines as a series of ordered coordinate pairs
- Polygons as one or more line segments that close
- In some systems, curves can be defined by a mathematical function

8.2.3.1 Simple features

- Geographic entities encoded using the vector data model are usually called features
- Features of the same geometric type are stored in a geographic database as a feature class, or when speaking about physical representation the term feature table is preferred
- Simple feature datasets are sometimes called *spaghetti* because lines and polygons can overlap and there are no relationships between any of the objects.
- Simple features lack more advanced data structure characteristics, such as topology

8.2.3.2 Topological features

- Topological features are essentially simple features structured using topological rules.
- Topology is the mathematics and science of geometrical relationships.
- Topological relationships are non-metric (qualitative) properties of geographic objects that remain constant when the geographic space of objects is distorted.
- Topological structuring of layers forces all line ends that are within a user-defined distance to be snapped together so that they are given exactly the same coordinate value. A node is placed wherever the ends of lines meet or cross.
- *Data validation* topology tests include network connectivity, line intersection, overlap, duplicate lines
- Many objects share common locations and partial identities. These situations can be modeled in a GIS database as either
    - single objects with multiple geometry representations, or

    - multiple objects with separate geometry integrated for editing, analysis, and representation.
- Topologically aware tools include: manipulate common, shared polylines and nodes as single geometric objects; rubberbanding; snapping; auto-closure; tracing
- Optimized queries from topological relationships include network tracing, polygon adjacency, containment, intersection
- In a topologically structured polygon data layer each polygon is defined as a collection of polylines that in turn are made up of an ordered list of coordinates (vertices).
- Figure 8.8 shows an example of a topologically structured polygon dataset
- Storing common boundaries between adjacent polygons avoids the potential problems of gaps (slivers) or overlaps
- Downside is that drawing polygons from multiple polylines is time intensive
- *Planar enforcement* means that all the space on a map must be filled and that any point must fall in one polygon alone, that is, polygons must not overlap.
- Planar enforcement implies that the phenomenon being represented is conceptualized as a field.
- In the *georelational* model, the feature geometries and associated topological information are stored in regular computer files, whereas the associated attribute information is held in relational database management system (RDBMS) tables.
- Geometry and topology were not placed in RDBMS because until relatively recently RDBMS were unable to store and retried geographic data efficiently

8.2.3.3 Network data model

- Network topological relationships define how lines connect with each other at nodes and define rules about how flows can move through the network
- In *linear referencing systems,* the location of geographic entities are stores as distances along a network from a point of origin
- *Dynamic segmentation* is a special case of linear referencing in which data values are added dynamically to the route each time the user queries the database

8.2.3.4 TIN data model

- A TIN is a topological data structure that manages information about the nodes comprising each triangle and the neighbors of each triangle.
- Briefly explains how Delaunay triangulation is carried out
- Advantages of TINs are
    - the density of sampled points can be adjusted to reflect relief
    - they incorporate the original sample points

    - easy to calculate elevation, slope, aspect, and line-of-sight
- Limitations of TINs are
    - They are especially susceptible to extreme high and low values since there is no smoothing of the original data
    - Unable to deal with discontinuity of slope across triangle boundaries
    - Difficult to calculate optimum routes

8.2.4 Object data model

- Each geographic object is an integrated package of geometry, properties, and methods.
- Geometry is treated like any other attribute of the object
- Geographic objects of the same type are grouped together as *object classes*
- Individual objects in the class are *instances*
- Three types of *relationships* in object data models are
    - *Topological* relationships are built into the class definition such as network and polygon structures
    - *Geographic* relationships are based on geographic operators that determine the interaction between objects
    - *General* relationships define other types of relationships such as between land parcels and ownership data, light pole IDs and attributes
- *Rules* are a means of maintaining data integrity during editing tasks. They include
    - *Attribute* rules are used to define the possible attribute values that can be entered, includes both range and coded values
    - *Connectivity* rules specify the valid combinations of features
    - *Geographic* rules define what happens to the properties of objects when an editor splits or merges them

Technical Box 8.2 Object-oriented concepts in GIS

- An *object* is a self-contained package of information describing the characteristics and capabilities of an entity under study.
- An interaction between two objects is called a *relationship*.
- A collection of objects of the same type is called a *class*.
- A class can be thought of as a template for objects.
- There are three key facts of object data models
    - *Encapsulation* which describes the fact that each object packages together a description of its state and behavior

- *Inheritance* is the ability to reuse some or all of the characteristics of one object in another object.
- *Polymorphism* describes the process whereby each object has its own specific implementation for operations like draw, create, and delete.

8.3 Example of a water-facility object data model

- The goal of this section is to describe an example of a geographic object model and discuss how many of the concepts introduced earlier in this chapter are used in practice.
- Figure 8.18 shows a possible object model using Unified Modeling Language (UML) to show objects and the relationships between them
- In UML models each box is an object class and the lines define how one class reuses (inherits) part of the class above it in a hierarchy.
- Figure 8.19 shows how a computer-aided software engineering (CASE) tool is used to specify the logical model

8.4 Geographic data modeling in practice

- No step in data modeling is more important than understanding the purpose of the data modeling exercise, gained by collecting user requirements form the main users
- Once an implementation-independent logical model has been created (using CASE tools and UML, for example), this model can be turned into a system-dependent physical model
- A physical model will results in an empty *database schema* – a collection of database tables and the relationships between them

ESSAY TOPICS

1. Define what is meant by the term ‘data model’ and explain why its design is vital to the success of any GIS application.
2. Describe, with examples, five key differences between topological vector and raster geographic data models. It may be useful to consult Figure 8.3 and refer back to Chapter 3.
3. It has been suggested that the differences between the vector and raster conceptual computer models will be eliminated by technological advances. What do you think are these advances? Have the vector and raster models converged?
4. What are the deficiencies of CAD, graphical and image GIS data models in geographic data analysis?

5. In the hybrid georelational model, use is made of a standard relational data base management system. What are the advantages of this approach to both data modeler and GIS user?
6. Study Section 8.4.2 and Technical Box 8.2. Why do GI scientists favor the object oriented approach?
7. 'There is no such thing as the correct geographic data model' (page 196). Explain why.
8. What do you understand by the terms encapsulation, inheritance and polymorphism and why are they important in object oriented data modeling?
9. The Chapter lists eight different approaches to data modeling in GIS. Consider how each might be extended to create either (a) a truly three-dimensional system for use in geology or (b) a temporal GIS for use in tracking land use change over time.

MULTIPLE CHOICE QUESTIONS (MCQ)

1. Place the following terms into their correct ordering by level of abstraction involved:

Term	**Write 1 – 4**
Conceptual model	
Logical model	
Reality	
Physical model	

2. For each of the eight types of model, state whether it is best for representing a field or an object:

Geographic Data Model	**Field (F) or Object (O)?**
Computer-aided design	
Graphical, non topological	
Image	
Raster/grid	
Vector/Georelational topological	
Network	
Triangulated Irregular Network (TIN)	
Object	

3. For each of the eight models, state which applications area it is most suited to support, choosing your answers from the list of the right hand column (each application should be

selected once only):

Geographic Data Model	**Answer (a)-(h)**	**Application**
Computer-aided design		a) Surface terrain analysis
Graphical, non topological		b) Computer cartography
Image		c) Spatial analysis
Raster/grid		d) Socio-economic data analysis
Vector/Georelational topological		e) Natural hazards assessment
Network		f) Estate management
Triangulated Irregular Network (TIN)		g) River management
Object		h) Land ownership information

4. Tick the following GIS operations in which topological information is useful:
    a. Data validation
    b. Modeling integrated features
    c. Editing
    d. Spatial interpolation
    e. Query optimization

5. Give the full name of each of the acronyms listed below:

UML	
VRML	
XML	
HTML	

6. Tick which of the following statements describe a TIN model in GIS:
    a. An alternative to raster coding of surface terrain
    b. A way of coding objects in GIS
    c. A way of modeling continuous data in a vector GIS environment
    d. An image data format.

7. A data model that separates the attribute information from the feature geometries and stores it separately is called

a. Object oriented
b. Georelational
c. Topological
d. Network

8. Which of the following are not topological relationships?
    a. Containment
    b. Direction
    c. Adjacency
    d. Intersection
    e. Distance

9. A ‘class’ in data base modeling is best defined as:
    a. A grouping of similar object types
    b. A template for creating objects
    c. A set of features

CLASS AND INDIVIDUAL ACTIVITIES

1. Figure 8.3 has a SPOT satellite image together with a vector map of the same area. Make a list of the ways in which these differ, thinking about resolution, selectivity and so on.
2. Read Section 8.2, and then refer back to Figure 3.14 or an equivalent topographic map of an area known to you. Following the processes outlined in the chapter develop a data model for the information displayed on this map. An accessible guide to database design in GIS will be found at http://www.geog.ubc.ca/courses/klink/gis.notes/ncgia/u10.html#UNIT10.
3. Assume that you are working for a road maintenance agency. Your spatial objects are the roads over, say, a small city. The GIS is to support operations such as:
    a. Surface renewal as required;
    b. Avoiding clashes with other agencies that might want to dig holes in your roads; and
    c. Improvements to the road structure.
4. How would you record the *network* of roads in your database and what attributes would you attempt to collect?
5. Now assume that you are working for a bus company for the same city, and suggest a GIS to support operations such as:

Timetabling and timing

Predicting demand for the buses

6. This exercise explores the dependence of lossless compression of a raster on the nature of the data involved. Run length encoding (RLE) is an obvious way to compress a raster of values. In it we record each run of values as a pair of numbers, one giving the number itself, the other the number of repeats. Given the raster of integers:

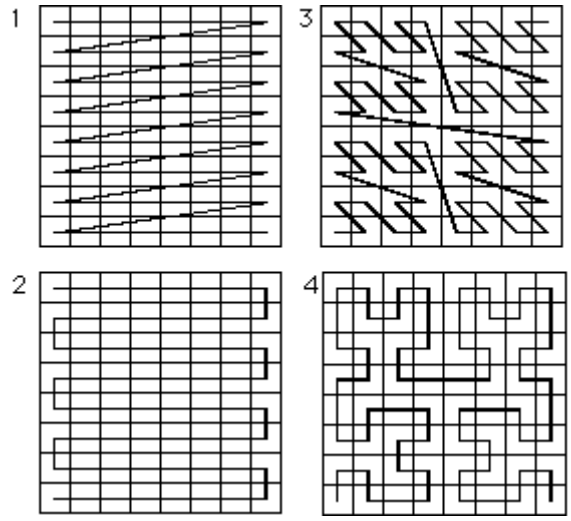
    12111234
    11112233
    11111223
    44112233
    44412223
    44442233
    44422223

7. What is the saving in store if we run length encode this?

Now, given the raster of floating-point values:

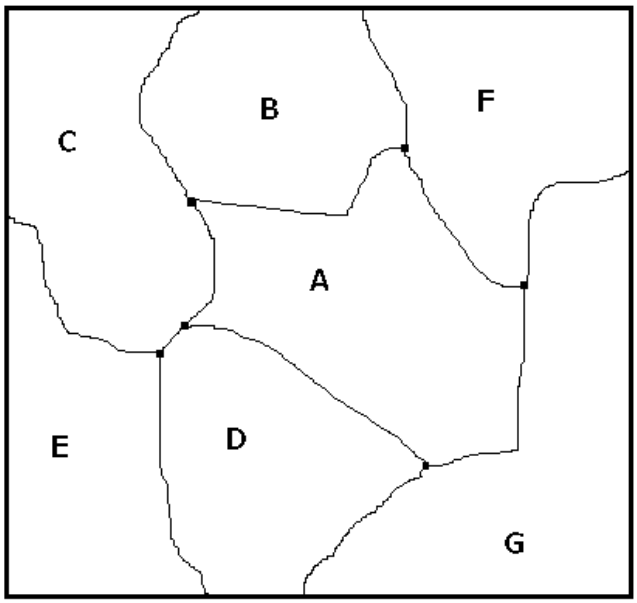
1.1 2.1 1.1 1.4 1.6 2.1 3.6 4.2
1.1 1.4 1.5 1.8 2.9 2.7 3.4 3.1
1.1 1.4 1.2 1.1 1.4 2.4 2.8 3.3
4.1 4.2 1.1 1.9 2.2 2.4 3.1 3.9
4.2 4.3 4.4 1.4 2.2 2.3 2.4 3.3
4.4 4.1 4.9 4.7 2.1 2.2 3.1 3.3
4.2 4.4 2.1 2.2 2.3 2.4 3.6 4.0

What is the saving in store if we use the same approach? To be effective RLE relies on like values being next to each other in the sequence, and in the two examples so far, this has been as in (1) below. It is possible to do better by wrapping round at the end of each row of the raster, as in (2). Repeat the exercise. Does it compress the data even further?



8. Boxes (3) and (4) show scan orders that attempt to improve things further, this time by scanning the raster in an order that, if it is present, will make maximal use of any spatial autocorrelation in the grid. Order the entries according to these schemes and repeat the exercise. Method (3) is the so-called Morton Ordering, and it has been much debated in GIS. What do you conclude from this exercise?

9. Study Figures 8.8. For the planar enforced areas A – E below number the polylines and create two tables that describe the structure. There is no need to complete the digitization of each polyline.



10. How big is mainland Australia? This exercise is taken from O'Sullivan and Unwin (2002, page 175). It can be done using a semi-automatic digitizer, but it is useful to do it by hand.

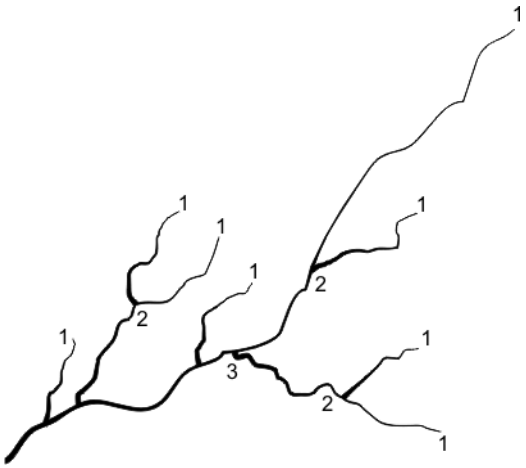
11. Trace the shoreline of Australia from a map of the continent, taking care to ensure that the source is drawn on an equal area map projection. Record the shoreline as a series of (x, y) co-ordinates. How many vertices do you need to represent the shape of Australia so that it is instantly recognizable? What is the minimum number you can get away with? How many do you think you need to ensure that you get a reasonable estimate of the total area of the continent?

    a. Use Simpson's method (see O'Sullivan and Unwin, 2010, page 191-194) to compute its area. This is easily done using any spreadsheet program. Enter your co-ordinates into the first two columns and copy these from row 2 onwards into the next two columns, displacing them 'upwards' by one row as you do so. Make a copy of the first co-ordinate pair into the last row of the copied columns. The next column can then be used to enter and calculate the trapezoid formula. The sum of this column then gives your estimate of the continent's area. You will have to scale the numbers from co-ordinate units into real distances and areas on the ground. Compare your result with the 'official' value, which is 2,974,581 km^2.
    b. O'Sullivan and Unwin thought that the minimum number of coordinate pairs needed to make the result recognizably Australia is nine. Using a 1:30 000 000 map as a source they recorded 45 co-ordinates for the shoreline and got an area of 2,964,185 km^2, which is about 0.3% too low. In fact, the closeness of this result is likely to be a happy accident. What conclusions do you draw from this exercise?

11. Illustrating structure in a tree network. This exercise is taken from O'Sullivan and Unwin (2010). From a topographic map at a scale of 1:25,000 or 1:50,000 (or nearest equivalents), trace off a drainage network represented by (usually) the 'blue lines'. Try to find a reasonably large basin with, say, about 50 or so 'sources' where streams begin.

    a. Now 'order' this network using Strahler's method. To do this, give all the fingertip tributaries the order '1'. Where two such 1^{st} order streams meet, the stream that results becomes 2^{nd} order, and should be labeled as such. Note that any 1^{st} order tributaries that join a 2^{nd} order stream will not change its order until it meets another 2^{nd} order stream, when the stream becomes '3^{rd} order'.

b. Continue until all the streams have been ordered in this way and there is just one stream with the highest order reached. This ordering scheme is :



The Strahler ordering scheme

c. Now count the number of streams in each order. In the example this count is eight 1st, three 2nd and one 3rd.

d. Plot a graph of the *logarithm of the number of streams in each order* on the vertical axis against stream order (1 , 2 , 3 , … *n*) on the horizontal axis.

Usually this results in a straight line plot, evidence of what was called the 'law of stream numbers', one of several laws due to a US engineer called Horton. This structure seems to be a property of most branching networks.

12. Representation in raster and vector codings. The basic objective of this assignment is to code a series of area polygons into vector and raster codings, and then use appropriate algorithms to calculate the total area of the object. We suggest that you proceed as follows:

a. Choose an area object such as a lake, city built over area or forest. Initially make this small enough so that considerations of Earth curvature are not significant.

b. Establish an origin at the bottom left of the plan and create an appropriate plane co-ordinate system.

c. Use this system to record the co-ordinates in an exact, vector data structure. You could do this directly into a GIS package, by digitizing off screen, but it is just as easily done using a spreadsheet.

d. Use the appropriate algorithm to compute the area of each and hence the total area.

e. Using the same origin (bottom left), create an appropriate raster grid for the entire area. Use two resolutions, a 16 by 16 grid and a 32 by 32 grid. Record

these grids using both run length encoding based on a row ordering, and compare this with the same run length encoding generated using the Morton ordering.

f. Estimate the total area from both grids.
g. Finally, write an account saying which of the two approaches is most appropriate for this problem and why.

13. As a final task, imagine that the given map is not depicting a single, small area, but an extensive, continental scale series of areas, such as the lakes of Manitoba, Canada. What implications would this have on measurement techniques and results, in particular in reference to map projection and map reference system?

At slightly more advanced level this exercise can be extended to include coding the quadtree for each of the rasters.

14. Investigating compression. Although raster compression can be achieved using run length and related encodings, use is often made of proprietary compression using, for example, discrete cosine, fractal, and wavelet approaches. Using the websites and a selection of these approaches to compress some of the files on your computer. Which seems to work most effectively for which type of data? Can the approach be used to enhance data for display? Is the method lossy or lossless? Some sites to explore are: www.lizardtech.com, www.jpeg.org, and www.pkzip.com.
15. As the chapter indicates, obtaining a clear set of user requirements is a sine qua non of data modeling. Use a role play simulation involving a data modeling team with a series of 'clients'. In this, it is helpful to develop the client needs as a set of 'use cases', each defining a specific, defined potential use of the GIS when it is established. A useful website describing the approach is at www.pols.co.uk/use-case-zone/ but there are many other sites that have similar materials.

FURTHER READING

Abel, D. and Mark D.M. (1990) A comparative analysis of some two-dimensional orderings International Journal of Geographical Information Systems, 4(1): 21-31.
A classic attempt to show how raster ordering can help compression.
Fisher, P. and D. Unwin (Eds) 2005 Re-presenting GIS, London: Wiley
A set of research papers that together provide numerous arguments for full object orientation in geographical data bases.

Gahegan, M.N. and S. Roberts (1988) An intelligent, object- oriented geographical information system, International Journal of Geographical Information Systems 2: 101-10.
An attempt to produce a fully object oriented GIS.
Goodchild, M.F., and A.W. Grandfield (1983) Optimizing raster storage: an examination of four alternatives, Proceedings, AutoCarto 6, Ottawa, 1: 400-7.
More on compression.
Peuker, T.K., and N. Chrisman (1975) Geographic Data Structures, American Cartographer 2(1): 55-69.
Peuquet, D.J. (1984) A conceptual framework and comparison of spatial data models, Cartographica 21(4): 66-113.
Two papers that explore the standard arc-node data structure.
Raper, J.F. (2000) Multidimensional GIS, London: Taylor & Francis
An advanced discussion of the need to incorporate additional dimensions, such as height/depth and time, into geographic data modeling. A case study of research at Scolt Head in England provides a powerful argument for full object orientation in data base design.
Worboys, M.F. (1992) A generic model for planar spatial objects. International Journal of Geographical Information Systems, 6(5), 353-372.
The classic paper on object types in GIS by a logician/computer scientist.
Worboys, M.F., Hearnshaw H.M. and D.J. Maguire, (1990) Object-oriented data modeling for spatial databases. International Journal of Geographical Information Systems 4: 369-383.
Another classic paper, one of the earliest calls for object orientation in GIS.
Zeiler, Michael (1999) Modeling our World: The ESRI Guide to Geodatabase Design (Redlands CA: ESRI Press).
An extremely useful survey of the approaches taken by the world-leading GIS software.

RELATED READING

Longley P.A., Goodchild M.F., Maguire D.J. and Rhind D.W. (eds) 2005 *Geographical Information Systems: Principles, Techniques, Management and Applications* (abridged edition). Hoboken, NJ: Wiley.
26. Relational databases and beyond, M R Worboys
Maguire D.J., Goodchild M.F. and Rhind D.W. (eds) 1991 *Geographical Information Systems: Principles and Applications*. Harlow, UK: Longman (text available online at www.wiley.co.uk/gis/volumes.html).
16. High-level spatial data structures for GIS, M J Egenhofer and J R Herring, pp. 227-37
19. Digital terrain modeling, R Weibel and M Heller, pp. 269-97

ONLINE RESOURCES

NCGIA Core Curriculum in GIScience, 2000 (www.ncgia.ucsb.edu/giscc)

2.3.1. Information Organization and Data Structure (051), Albert Yeung

2.3.2. Non-spatial Database Models (045), Thomas Meyer

2.4.1. Rasters (055), Michael Goodchild

2.4.2. TINs (056)

2.4.3. Quadtrees and Scan Orders (057), Michael Goodchild

2.6. Representing networks (064), Benjamin Zhan

2.9.1. Transportation Networks (183), Val Noronha

NCGIA Core Curriculum in GIS, 1990 (www.ncgia.ucsb.edu/pubs/core.html)

4. Raster GIS

11. Spatial objects and database models

12. Relationships among spatial objects

13. Vector GIS

21. Raster/vector debate

30. Storage of complex spatial objects

31. Storage of lines: chain code

35. Raster storage

36. Hierarchical data structures

37. Quadtree algorithms, spatial indexes

38. Digital elevation models

39. TIN data model

42. Temporal and 3D databases

43. Database concepts I

44. Database concepts II