International Journal of the Physical Sciences Vol. 5(5), pp. 476-483, May 2010
Available online at http://www.academicjournals.org/IJPS
ISSN 1992 - 1950 © 2010 Academic Journals

Full Length Research Paper

Topological error correction of GIS vector data

Süleyman Sırrı Maraş[1]*, Hakan Hadi Maraş[2], Bahadır Aktuğ[2], Erdem Emin Maraş[3] and Ferruh Yildiz[4]

[1]Vocational School of Technical Sciences, Selcuk University, Konya, Turkey.
[2]Department of Photogrammetry, General Command of Mapping, Ankara, Turkey.
[3]Geomatics Engineering, Faculty of Engineering, Ondokuz Mayıs University, Samsun, Turkey.
[4]Department of Geodesy and Photogrammetry, Selcuk University, Konya, Turkey.

Accepted 15 April, 2010

The detection and removal of the errors and inconsistency of spatial data in vectors are the main concerns of the Geographic Information Systems (GIS) since the GIS templates in vector form are obtained from the raster maps and plans. In order to increase the accuracy and the reliability of GIS analyses, a thorough and reliable error detection and removal procedure is indispensible. In this study, the possible geometric and topological errors are explained in detail; the methods of detecting such error sources are given along with methods of determination and visualization.

Key words: Geographic information systems, spatial data, topology.

INTRODUCTION

In simple sense, GIS is defined as the storing, processing and presentation of spatial data after relating them to maps (Star and Estes, 1990; Antenucci et al., 1991). Considering the fact that the whole system runs in a computer environment, it is ideal that template maps are directly produced digitally. Nowadays, production, storage and usage of maps in digital environments either in photogrammetric or classical methods are quite common (Burrough, 1997). In this respect, a layer hierarchy as well as ownership of the vector data can be assigned to individual features in the maps. On the other hand, considering that majority of maps show the ownership of real-estates, huge efforts are needed to convert them to the necessary vector equivalents; otherwise, it is inevitably necessary to reproduce all the available, which is not efficient in terms of cost and time. Consequently, the most common method of producing vector maps is the precise scanning of analog maps into raster formats and then digitizing into vector forms (Grimshaw, 1994).

While there are many alternatives in the choice of digitizing software with most of them working automatically, the human interaction is necessary to some extent during the transformation from raster to vector forms (Aronoff, 1989). Typical examples of such interaction could be given for the interpretation, determination of the ownership of the attributes with respect to the features, specifying the distinct features where color or layer information is not accessible (Chambers, 1989).

The applicability of the project directly depends on the accuracy and topological consistency of the vector data that form the basis for GIS. The existence of errors or inconsistency in the vector data could easily prevent a reliable GIS analysis (Feuchtwanger, 1989). Therefore, it is necessary to do an overall check as well as to compare them in terms of compatibility before hand in order to minimize the errors of the digitization process. In particular, user awareness about the amplitude of the expected errors is extremely important (Yomralıoğlu, 2000). The most costly step in GIS application is data collection. The ratio of the cost of data collection to total cost is estimated as between 60 and 80% (Dale and McLaughlin, 1988).

In the digitization procedure, it is often practical to use a CAD (Computer-Aided Design) software package rather than a GIS software package (Burrough, 1986; Hodgson et al., 1989; Cowen, 1990). In this respect, the AutoCAD software could work sufficiently to provide a cost advantage as well as a customization opportunity through its intrinsic LISP (List Processing) programming language. Moreover, the DCL visualization and windows

*Corresponding author. E-mail: ssm@selcuk.edu.tr.

programming module that runs interactively with LISP provides new possibilities. Furthermore, the DXF (Drawing Exchange Format) formatted data are very common and easy to exchange (Dangermond, 1990).

In this study, a new module designed in AutoLISP and DCL is introduced, which detects the possible topological errors in the vector data sets and minimizes the interaction burden (Maraş, 2005). It should also be noted that while employed method falls in the field of GIS applications, the visualization and adaptation software could not be considered simple GIS software.

SPATIAL DATA STRUCTURES

Different methods have been developed to represent information on the maps, produced by modeling of the spatial features in computers. These methods are generally based on the use of vector or raster data structures (Berry, 1993; ERIN, 1996).

Raster data structures

Information belonging to spatial objects at raster data structures is represented as matrix or cells forming grid network. Each cell can be represented by the binary number system to point out whether it is part of a graphic element or distinguishable from other features or tone and color values of the cells (Casettari, 1993). In raster data structures, when black/white color is used, only spatial information is stored, and when tone or color values are used besides the spatial information, limited attribute information can be stored (Maraş, 2005).

Vector data structures

The basic element in the vector data is the point. Points create lines and set of lines creates polygon. To represent the spatial features in vector data structure, initially coordinate pairs that make up those vectors are stored in computers. Then these data consisting of coordinate pairs will have to be observed enough for query and analysis related with the spatial side (Cuthbertson, 1993). Over time, vector data consisting of coordinate pairs are enhanced for spatial analysis. Vector data structures in accord with spatial analysis can be divided into two basic classes:

1. Non-topological data structure (non-structured vector data).
2. Topological data structure (structured vector data).

Non-topological data structure

Spatial features in non-topological spatial data structure called spaghetti data structure are represented with the three basic geometric forms (point, line, polygon) in connection with the scale (Guptill, 1987). Features that are represented by point geometric shape are zero-dimensional elements and each one is defined by one coordinate pair (x, y). Line-shaped features are one-dimensional elements defined by (x, y) coordinates series that follow each other. Polygon-shaped features are defined as two-dimensional closed shapes that are formed by lines starting and ending at the same point.

Problems that prevent the spatial data analysis due to the non-topological structured data usually obtained as a result of the digitization process are (Chung, 1995; Pequet and Marble, 1990):

a) Point features may not be at the intersection point of line features (bridge at the intersection of river and road).
b) Polygon features are not closed properly.
c) Neighborhood relations are unclear.
d) Contact points do not coincide (e.g. river does not coincide on the edge of the lake).
e) Since neighboring or same location features are represented twice, there is no full coincidence at the point and line features. Overlaps or gaps in polygon features do happen.
f) Due to exclusion of information, polygon, point or line features included in polygons are unclear.
g) Navigation is not possible since there is no direction concept in the line features.

Topological data structure

Topology examines the characteristics such as neighborhood and merges outside the coordinate system information of geometric objects (Chung, 1995). The aim of the topological knowledge in GIS is to increase spatial analysis opportunities. To represent the spatial information as well as spatial relations (neighborhood, coincidence, directions, links) of the features in topological data structures on the computer; the node elements corresponding to point, edge (arc) elements corresponding to the lines and the face elements corresponding to the polygons are used. Note that more than two line intersection points may be an exposed end of the line or a single point not connected to any side. Edge is a set of coordinate pairs starting with a node and ending with a node. Face is the largest two-dimensional space restricted by the edges and cannot be divided by an edge. According to this, point features are composed of nodes, line features with one or more edges, and the face features surrounded by one or more edges (Tomlinson, 1990).

Topology is the geometric relationship between edges, nodes and the faces they created. According to the other definition, topology is a way or method in which logical relations can be defined such as neighborhood, coincidence, inclusion, intersection, sharing, in addition to metric relationships such as the geometrically identifiable coordinate, length, area (Bank, 1997).

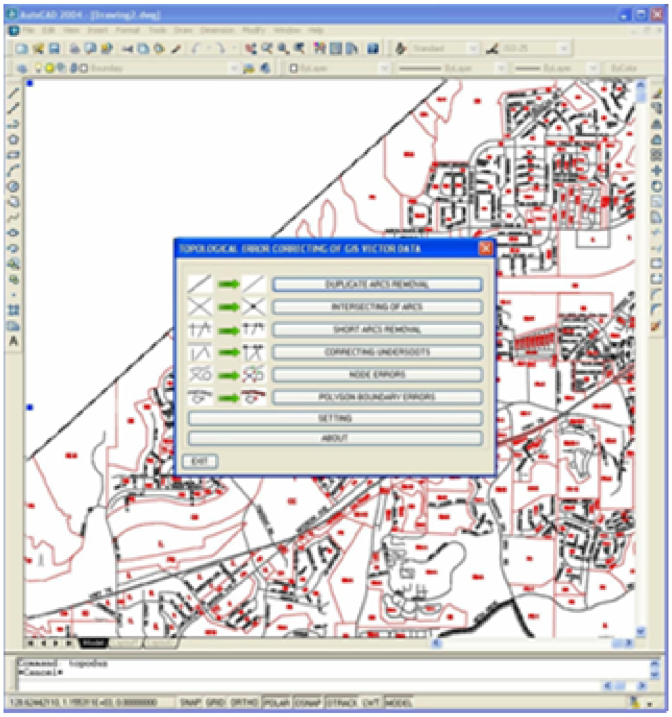
To be able to evaluate a topological database, in addition to the geometric properties the following relationships must be determined and stored:

a) Edges making up the boundaries of each polygon (polygon topology table);
b) Neighborhood relations between the polygons (edge topology table);
c) Connections at the intersection points (node topology table); and
d) Start and end points of edges (edge-coordinate data table).

In the databases that topological properties can be stored and used, information about a line feature is connected to another line feature, depending on where, and at what point line features are combined, the polygon features to the right and left line features can be easily and quickly queried (Healey, 1991). The existence of the topological data in addition to the spatial data and attribute data in database will cause the increase in volume of the database. Advantage of topological data that allows spatial analysis and queries expected from GIS offers very significant advantage compared to data storage disadvantages (Maraş, 1998).

For example, the advantages provided by digitization and storing of the polygons with common edges are:

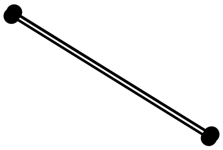
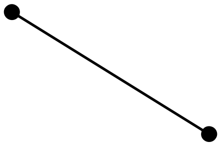
a) Time loss (digitization of the same borders twice) will be blocked.
b) There will be no consistency loss (digitization of the same borders for the second time with the same coordinates)
c) Features will be represented by fewer records because of unrepeated data.



Command: topodua
Cancel

Figure 1. Snapshot of the main menu of the interface running on AutoCAD.

Table 1. Removal of coincident arcs errors.

Topological error correction process	Before	After
Removing of duplicating arcs		

d) Spatial analysis will be easier.
e) Costs will be reduced.

TOPOLOGICAL ERRORS IN SPATIAL VECTOR DATA AND ITS VISUALIZATION AND ELIMINATION BY MEANS OF AUTOCAD VE AUTOLISP

Since spatial vector data contain many topological errors arising from the quality of the source material, data collection techniques and the human interaction, acquisition of vector data for GIS requires further processing in order to use it for some advanced spatial analysis. Before using spatial data for analysis, it must be made topologically-correct. Vector features can be made to respect spatial integrity through the application of topology rules such as ‘’polygons must not overlap’’. To make data usable or to make it clean, some quality checks should be applied. All GIS software has the functionality of performing this kind of processes. Most of the GIS software is also capable of providing the spatial integrity and cleanness during data capturing (Cowen, 1990).

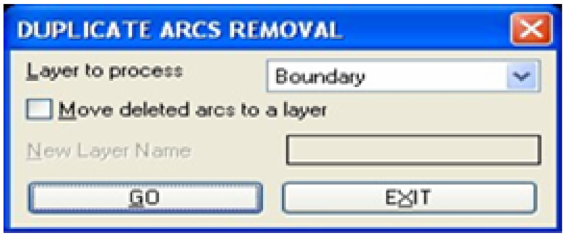
The topological errors arising from the conversion into the digital maps could be easily removed through the software written in AutoCAD LISP. The main window is shown in Figure 1. Command buttons in the dialog bring an interactive dialog to view and repair specific topological errors in the vector data.

The most common topological error types in spatial vector data:

1. Floating or short lines
2. Overlapping lines
3. Overshoots and undershoots
4. Unclosed and weird polygons

Removing duplicating arcs

According to topological integrity rules lines must not overlap (Hodgson, 1989). Duplicating lines must be removed to avoid obtaining false results from spatial analysis. For instance the removal of duplicate arc is shown in Table 1. For this operation, the relevant layer should be chosen. The duplicate arc could be optionally moved to a new layer. The software automatically checks the duplicating arcs and makes the necessary corrections via the


DUPLICATE ARCS REMOVAL

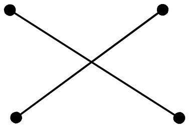
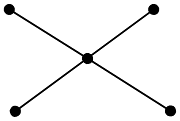
Layer to process: Boundary

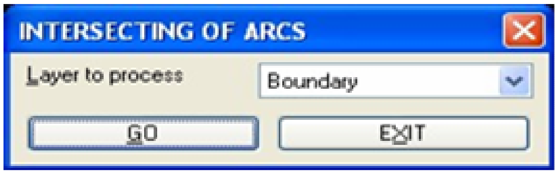
□ Move deleted arcs to a layer

New Layer Name

Figure 2. Removing of duplicating arcs dialog

Table 2. Fixing missing nodes error.

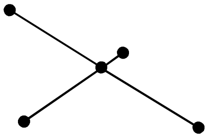
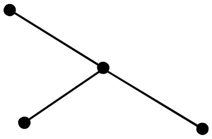
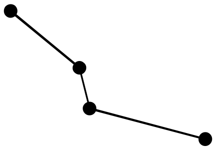
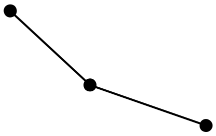
Topological error correction process	Before	After
Splitting the arcs and creating a node feature at the intersection		


INTERSECTING OF ARCS

Layer to process: Boundary

Figure 3. The dialog for intersecting arcs.

Table 3. Correcting overshoots and short lines error.

Topological error correction process	Before	After
Correcting overshoots		
Short line removal		

dialog shown in Figure 2.

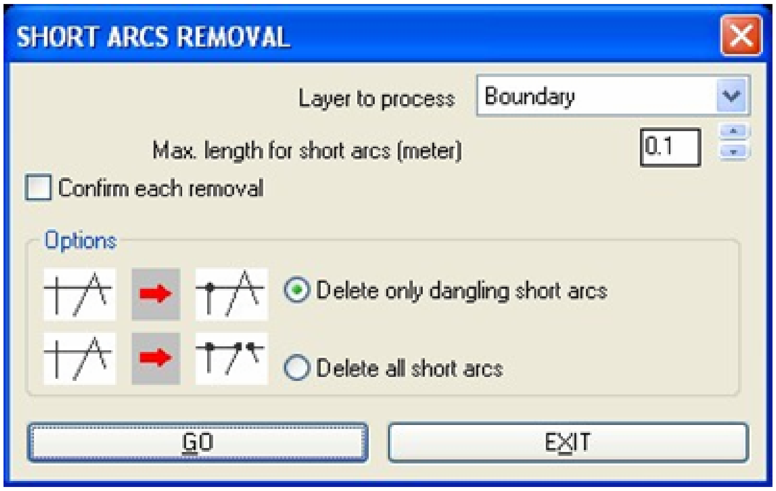
Intersecting arcs

The intersecting point of the lines and polylines should be a node feature in topology. For example, lines must connect with nodes at an intersection, especially in a network type data representing the transportation (Cowen, 1990).

If not supported by digitization software, such connections are omitted as shown in Table 2. Such an error disables the analysis that relies on the area and neighborhood. The software automatically checks the inconsistencies and creates necessary nodes at the intersection points via the dialog shown in Figure 3.

Removing nodes extends too far (overshoots) and too short (zero-length) arcs

During data digitizing or data conversion some floating or dangling short lines may occur as shown in Table 3. Such lines do not only bias the results of GIS analyses but also may completely prevent



SHORT ARCS REMOVAL

Layer to process: Boundary

Max. length for short arcs (meter): 0.1

□ Confirm each removal

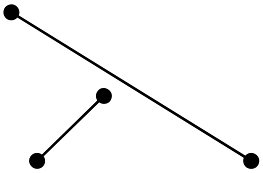
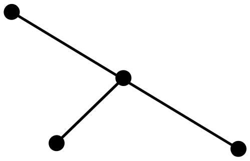
Options

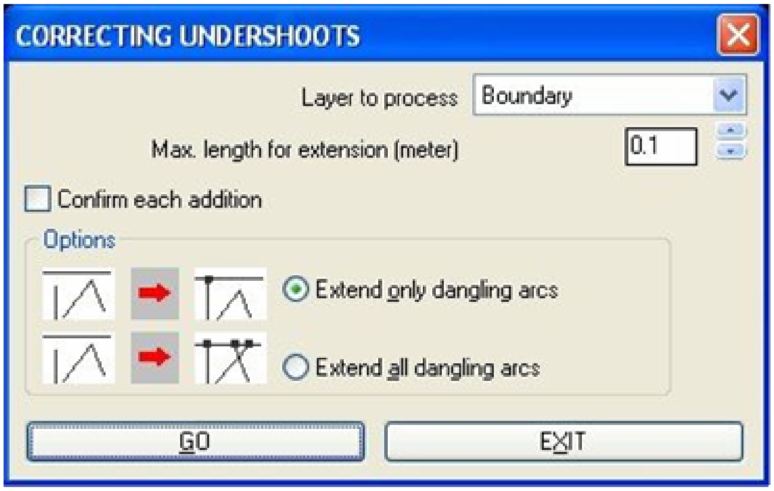
● Delete only dangling short arcs

○ Delete all short arcs

Figure 4. Short arcs removal dialog window.

Table 4. Correcting undershoot errors.

Topological error correction process	Before	After
Correcting nodes which falls short (undershoot)		



CORRECTING UNDERSHOOTS

Layer to process: Boundary

Max. length for extension (meter): 0.1

□ Confirm each addition

Options

● Extend only dangling arcs

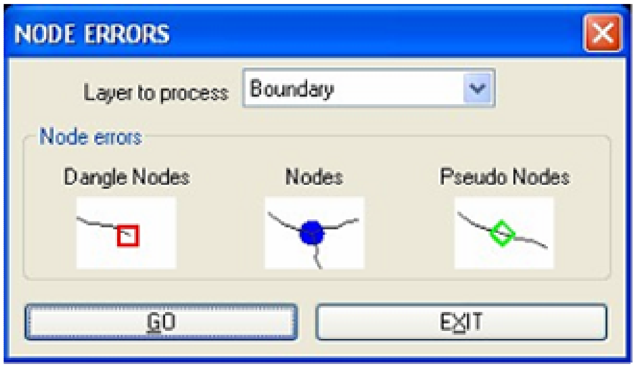
○ Extend all dangling arcs

Figure 5. Undershoots correction dialog window.

getting a reliable result (Hodgson et al., 1989). In accordance with the criteria defined by the user these line segments can be removed by means of prepared application software as tuned through the input dialog window shown in Figure 4. For allowing more control over the lines to be deleted, an optional confirmation can be requested by the user at each individual operation.

Correcting nodes falls short (undershoot)

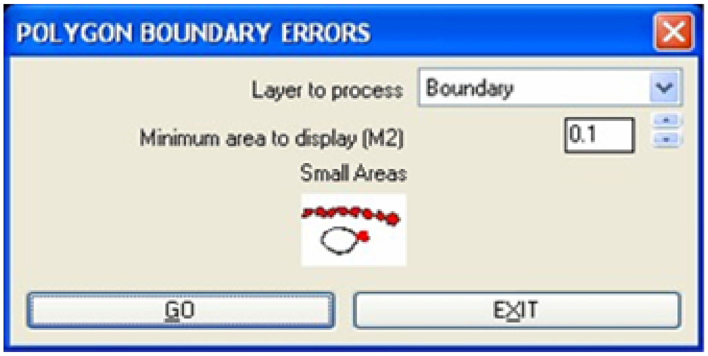
For the cases where the segments fall short, a procedure similar to overshoots can be followed (Table 4; Figure 5). In such topological errors, the end of a line falls short and is disconnected near the end of the neighboring line. To compensate such errors, a user-defined



NODE ERRORS

Layer to process: Boundary

Figure 6. Node errors display dialog window.



POLYGON BOUNDARY ERRORS

Layer to process: Boundary

Minimum area to display (M2): 0.1

Figure 7. Polygon boundary errors dialog window.

limit can be given as tolerance along with an optional confirmation that can be requested by the user at the each individual operation. The default value for the under shoots is defined as 0.1 m.

Displaying node errors

The decision about whether some inconsistencies are the true errors could be a task of the user. Such examples of inconsistencies can be given as the dangle nodes, nodes and pseudo nodes as shown in Figure 6.

Displaying tiny polygons (slivers and gaps) error

The errors of the polygons could arise from the overlapping, disconnected nodes or the digitization errors. The omission of one or more vertices during the formation of the neighboring polygons is one of the most common errors in practice (Dangermond, 1990). In such cases, either duplication (sliver) or the exclusion (gaps) of some parts of the polygons emerges. The overlapping errors in polygon feature could seriously bias the results obtained in the analyses (Figure 7 and Table 5).

RESULTS AND DISCUSSION

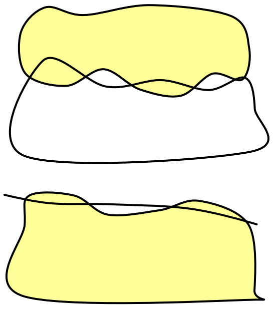
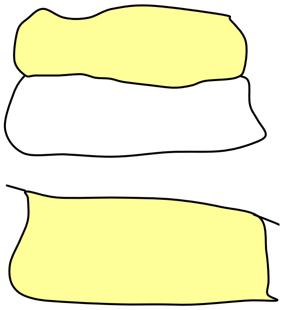
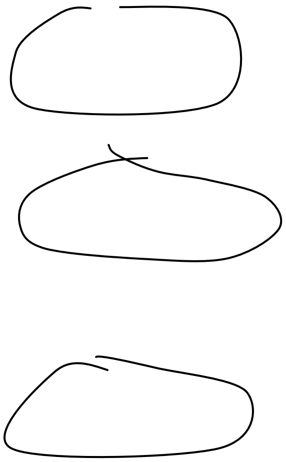
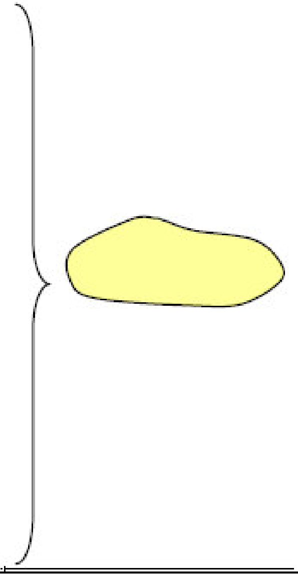
Speed, accuracy and economical criteria gain more importance as the computer technology develops. In this respect, decision support systems with intensive GIS analysis are becoming more widespread day by day. The analog maps produced by public and private service are eligible to be used as templates for GIS analysis.

Transformation of geographic data into the databases manually is often avoided due to the high possibility of adding extra errors and it is preferred instead to digitize from scratch. The data collection is the most costly and time-consuming step in GIS project. Rather than collecting data directly from the analogue sources, it is rather compiled from the existing vector data available at our institutions and bringing them to the level of prescribed standards. Therefore, it is needed to produce and develop the necessary semi-automatic data transformation software that will facilitate this process.

It is possible to produce vector data from the analogue maps through the CAD software in DXF format as long as the attributes are ignored. However, this kind of process is not only time-consuming and but also involves errors and relatively lower precision. In summary, an application that will meet the above-mentioned demands should:

1. have a user-friendly interface,
2. allow a high-degree user interaction with the modules,
3. be written in a common world-standard and flexible programming language like LISP which also allows

Table 5. Tiny polygon removal.

Topological error correction process	Before	After
Fixing tiny polygon features		
Fixing polygon closure errors		

manipulation of topological data,

4. be designed as an add-on module rather than an independent application which takes advantages of the standard operations of a commercial and matured CAD software such as zoom in, zoom out, pan, trim, extend, erase, importing and exporting various data formats.

Consequently, it was observed that the designed software with the properties outlined above proved to be successful in correcting the possible errors as well as visualization of the vector data so as to be used as GIS templates.

REFERENCES

Antenucci JC, Brown K, Croswell LP, Kevany JM, Archer H (1991). Geographic Information Systems. Van Nostrand Reinhold. New York.

Aronoff S (1989). Geographic Information Systems: A management perspective. pp. 16, WDL Publications. Ottawa. Canada.

Bank E (1997). Coğrafi Bilgi Sistemlerinde Topoloji. Sayılı Harita Dergisi. Ankara. 118: 65-74. (in Turkish).

Berry JK (1993). Cartographic Modelling: The Analytical Capabilities of GIS. in Environmental Modelling with GIS (Goodchild MF, Parks BO. Steyaert LT). pp 57-74. NewYork,. Oxford.

Burrough PA (1986). Principles of Geographical Information Systems for Land Resources Assessment. Clarendon Press. Oxford University Press. New York.

Burrough PA (1997). Geo-information needs: effect of scale. ITC Journal. 3/4: 235-236.

Casettari S (1993). Introduction to Integrated Geo-information Management. Chapman & Hall, London, pp. 24-27, 31-32,50-56.

Chambers D (1989). Overview of GIS Database Design. ARC News. Vol.11. No.2. Redlands. California.

Chung M (1995). An Object Oriented Approach for Handling Topology in Vpf Products, Proceedings Of GIS/LIS '95 Annual Conference And Exposition, Volume I, Nashville Convention Center, Tennessee, pp. 163-174.

Cowen DJ (1990). GIS versus CAD versus DBMS: what are the differences?, pp. 52-61, Introductory Readings in Geographic Information System, Taylor&Francis Ltd., Burgess Science Press, London.

Cuthbertson M (1993). A Database for Environmental Research Programmes, J. Environ. Manage. Vol. 37(4): 291- 300.

Dale P, McLaughlin J (1988). Land Information Management. Oxford University Press. Oxford.

Dangermond J (1990). A Classification of Software Components Commonly Used in Geographic Information Systems. pp. 30-51. Introductory Readings in Geographic Information System, Taylor&Francis Ltd., Burgess Science Press.London.

ERIN (1996). Glossary of Geographic Information Systems and Metadata Terms. Environmental Resources Information Network in

Australia. http://www.erin.gov.au/gis/gis_gloss.html (indexed at Aug 17, 1998).

Feuchtwanger M (1989). Geographic Logical Database Model Requirements. AUTO-CARTO 9 Proceedings. Maryland. pp. 599-609.

Grimshaw DJ (1994). Bringing Geographical Information Systems into Business. Longman, Harlow.

Guptill SC (1987). Desirable Characteristics of a Spatial Database Management System. AUTO-CARTO 8 Proceedings. Washington. pp. 278-281.

Healey RG (1991). Database Management Systems. in Geographic Information Systems Volume I: Principles. (Maquire DJ, Goodchild MF, Rhind DW). pp 119-134.Longman. London.

Hodgson ME, Barrett AL, Plews RW (1989). Cartographic Data Capture Using CAD, AUTO-CARTO 9 Proceedings, Maryland, pp. 406-415.

Maraş HH (1998). Coğrafi Veri Tabani Güncelleştirmesine Yönelik Coğrafi Bilgi Sistemi Tasarimi ve Uygulamasi, PhD Thesis, I.T.Ü. Fen Bilimleri Enstitüsü, Istanbul. (in Turkish).

Maraş SS (2005). Konumsal Vektör Verilerdeki Topolojik Hatalarinin Görselleştirilmesi ve Giderilmesi, MsC Thesis, S.Ü. Fen Bilimleri Enstitüsü, Konya. (in Turkish).

Pequet DJ, Marble DF (1990). Introductory Readings in Geographic Information System. pp. 5-6, Taylor&Francis Ltd., Burgess Science Press. London.

Star JL, Estes JE (1990). Geographic Information Systems: An Introduction, pp. 2-6, Prentice-Hall inc. Englewood Cliffs. New Jersey.

Tomlinson RF (1990). Geographic Information Systems - a new frontier. pp. 18-29. Introductory Readings in Geographic Information System, Taylor&Francis Ltd., Burgess Science Press. London.

Yomralioğlu T (2000). Coğrafi Bilgi Sistemleri Temel Kavramlar ve Uygulamalari Akademi Kitabevi, Trabzon. (in Turkish).