Evaluating Patterns of Spatial Relations to Enhance Data Quality

David Gadish, Ph.D.
Assistant Professor of Information Systems
California State University Los Angeles
5151 State University Drive
Los Angeles, CA, 90032
Tel: 323-343-2924
Email: dgadish@calstatela.edu

Key Words - Consistency, Patterns, Spatial Relations

Abstract - Effective use of data stored in spatial databases requires methods for evaluation and enhancement of the quality of the data. Spatial data quality can be evaluated using a measure of internal validity, or consistency, of a data set. Capturing spatial data consistency is possible with a multi-step approach.

A distance measure is used to detect implicit spatial relations between neighboring objects. The next step involves identifying the types of relations between these neighboring objects using topology based constraints.

The semantic information of objects, together with topological relations are combined to discover patterns, or rules, in the data. These rules are based on the analysis of the relations between each object and each of its neighbors, as well as between each object and all of its neighbors.

Patterns of spatial relations, represented as rules are validated using available metadata, as well as trend analysis and Monte Carlo simulation techniques. These can now be used as the basis for automated detection of inconsistencies among spatial objects, where possible inconsistencies are detected when one or more rules are violated. Detected inconsistencies can then be adjusted, thus increasing the quality of spatial data sets.

1 Introduction

Geographic information systems (GIS) play an increasing role in the decision-making processes involving the management and operations of many organizations. Many GIS's are complex to build and maintain. They typically require a variety of human experts, and involve large costs to support data acquisition, system setup and data maintenance. In spite of these substantial financial and human resource costs, spatial data in these systems are often uncertain and inconsistent. Because of this, the quality of spatial data are often unsuitable for many applications, and may lead to unreliable results in analysis of these data.

Effective use of spatial data stored in these GIS's requires an evaluation and improvement of the quality of the data. The use of GIS as a decision support tool is therefore dependent on reducing the inconsistency with the development of spatial data models that incorporate inherent topological and semantic properties. This paper presents a technique for discovering rules for managing inconsistencies in vector-based spatial data of a single GIS or a combination of two or more GIS databases. A data model is proposed to facilitate discovery of patterns in spatial data. These patterns form rules, which describe topological relations between objects, based on their semantic information.

The discovered rules must be checked to ensure their validity as they are later used as the basis for automated detection of inconsistencies among spatial objects, where inconsistencies are

detected when one or more rules are violated. Detected inconsistencies can then be adjusted, thus increasing the quality of spatial data in GIS databases.

2 Related Work

In GIS, spatial relations, which exist between geographic objects, play a key role both at the query definition level and at the query processing level (Clementini, Felice, and Van Oosterom, 1993). Three areas for each representation must be understood (Mark, Frank, Egenhofer, Freundshuh, McGranghan, and White, 1989). These areas are: (1) The logical view of the spatial representations, the way in which space is organized, and the features/objects that are representable within this view; (2) The spatial operations/relations that can exist for these features within this view; (3) The rules for their combination, and manipulation (their algebra). In systems that deal with more than one representational paradigm, the mappings (morphism) between these inherent concepts must be understood, since it is the morphisms that allow questions and answers to pass between the different paradigms. Each spatial method of representation adopts a fundamental paradigm for handing spatial concepts, and for organizing the spatial objects and the relations that "exist" between them. A topologically based GIS deals with at least four of the above types, with their associated morphisms: (1) The topology structure; (2) The underlying Euclidean geometric structure used for storage; (3) The various projections, and/or elevation models. (4) The unified spatial and attribute query language.

Topological relations are those that are invariant under topological transformations, that is, they are preserved if the objects are translated, rotated, or scaled. Since topology is a purely a qualitative concept, independent of any quantitative measures, it has been difficult to find appropriate formal data models for topological relations, as well as, methods to combine topological knowledge and reason about them (Hayes-Roth, Waterman and Lenat, 1983). A theory of binary topological relations between n-dimensional spatial objects (Egenhofer, Herring, 1991) provides a complete coverage; that is, any possible pair of spatial objects can be described by exactly one of the relations defined. The formalism is based on concepts of algebraic topology and set theory. Spatial regions are modeled as point-sets and the binary topological relations are then defined in terms of the intersections of the boundaries and interiors of two point-sets. Considering empty and non-empty intersections identifies sixteen potential relations, eight of which actually exist between two point-sets embedded in a two-dimensional space. This 4-intersection model proposed by Egenhofer (1995) is a method to distinguish different topological relations between two objects in a topological space.

The 4-intersection is a subset of the 9-intersection in which topological relations are defined in terms of the interior, boundaries and exterior of point sets, so that the 9-intersection would be able to distinguish more details. For region-region configuration in R^2, the 4-intersection and the 9-intersection provide the same eight relations. However, for line-line and region-line relations, the 4-intersection distinguishes only 16 and 11 relations respectively. The major difference for line-line relations is that the 4-intersection does not suffice to establish an equivalence relation, because several different line-line configurations have the same empty/non-empty 4-intersection. Similarly for region-line relations, the 4-intersection does not distinguish between certain topologically distinct configurations that may be critical for defining natural-language spatial predicates to be used in spatial query languages. With the 9-intersection, these problems are overcome (Egenhofer and Herring, 1990) such that every different set of 9-intersections describes a different topological relation, and relations with the same specifications will be considered to be topologically equivalent. Therefore, the 9-intersection can be employed to analyze whether or not two different configurations have the same topological relation. The work presented in this paper applies the 4 and 9-intersection models to derive topological knowledge about types of relations between neighboring objects. This information is then used to detect consistency

violations in spatial data by checking the relations between objects against a set of integrity constraints.

Servigne et. al. (2000) define integrity constraints using topological relations described by the 9-intersection model, as the association of two geographical objects, a topological relation between them and a specification (Forbidden; At least n times; At most n times; Exactly n times). They suggest that the specification forbidden is the most interesting and usable one, and limit their algorithm accordingly. Topological integrity constraints defined using this specification are a mean for the end-users to describe topological situations they do not want to see in their database (via a visual interface consisting of a dialog box in which the user can choose a pair of entities, a topological relation, and a specification). These constraints are specified manually into the system, and need to be adapted to the data set processed, which requires expert knowledge. If experts are unaware of some constraints, corresponding inconsistencies will not be detected and corrected. Medeiros et. al. (1995) also propose manual specification of rules, such as "An electric cable should start and end at a pole". In contrast, this paper looks at an automated approach for constraint, or rule discovery.

3 Topology Based Database (TopDB) Model

Different spatial representations are implemented by commercial GIS's to store, analyze, extract and present spatial data. Therefore, a generic representation, that can handle data from various geographic information systems is needed. A Topology-based Database (TopDB) model is defined to facilitate quality management of spatial data. Such a model accommodates the most common aspects of existing GIS databases and explicitly stores the topological relations that are detected between pairs of objects.

Analysis of data in GIS database(s), which are not in TopDB format, requires conversion of GIS data to the proposed TopDB representation. Rule discovery as well as detection and adjustment of inconsistencies is performed on the TopDB database. Once this is complete, the adjusted data are converted back to these original GIS format(s).

One key aspect of detecting and adjusting inconsistencies in GIS is to represent the geometric relationships between objects. Topological relations are special geometric relations, which are preserved under topological transformations such as scaling, translation, and rotation. They are based on algebraic topology, which deals with symbols representing geometric configurations and their relations (Egenhofer and Herring, 1992). This representation is not tied to distance or coordinate measure. Analysis of constructed topological relations may reveal inconsistencies in spatial data, and adjusting these inconsistencies will improve the quality of such data.

Since existing commercial GIS databases come in many formats, their data content should be translated to a common representation before they are subjected to analysis. The TopDB spatial representation is defined in terms of objects to represent real world entities, layers to represent the semantic information of the objects, and topological relations to represent the relations between the entities in the real world. The use of topological relations in the context of GIS data modeling as proposed by Egenhofer (1995) is adopted since it provides an unambiguous, complete coverage; that is, any possible pair of spatial objects can be described by exactly one of the relations defined.

The TopDB is a multi-layered spatial representation, which is defined by its objects and layers.

Definition 1 – Layers and Objects

A Layer is a subset of objects in a spatial representation with common observed characteristics and functions in the real world, that is, objects with a common semantic interpretation. A Layer is denoted as L_i. The k'th object on layer L_i in the representation is denoted as O_{ik}.

For example, one layer may consist of all the objects that represent individual roads. Another layer may consist of all the objects that represent individual residential parcels of land.

Definition 2 – Multi-Layered Spatial Representation

A Multi-layered spatial representation is a set of objects O_{ik} which are abstractions of real world entities of various semantic interpretations in a two dimensional space. Each object can appear in one and only one layer L_i. Therefore $L_i \cap L_j = \varnothing$ for all $i \neq j$, i.e. if $O_{ik} \in L_i$, then $O_{ik} \notin L_j$ where $i \neq j$.

Objects are the base graphic entities of the spatial representation. Each object represents one real world entity, and therefore may reside in one layer of a spatial representation. This paper concentrates on the analysis of shape objects only, since the ability to handle them will yield ability to manage the simpler topological structures of lines and points. Formal definitions of geometric object types are based on the point-set approach, where shape objects are defined as follows:

Definition 3 - Shape Object

A shape object is a close shaped graphic element, denoted as O_{ik}. It has two important attributes, namely, measurement attributes and topology attributes. Measurement attributes consist of N vertices $(x_1,y_1) \ \ldots \ (x_n,y_n)$, where the last vertex has coordinates that are identical to those of the first vertex. Topology attributes consist of $\left(\partial(O_{ik}),(O_{ik})^o\right)$. The boundary of a shape object O_{ik} is composed of the accumulation of all points on the circumference, and is denoted as $\partial(O_{ik})$. The interior of a shape object, which consists of all points of the shape excluding those on the circumference and is denoted as $(O_{ik})^o$.

The 4-Intersection model (Egenhofer and Franzosa, 1995) classifies topological binary relations between shape objects. The classification is based on the intersection of the boundaries and interiors of object pairs. These are represented as a 4-tuple in terms of: (1) The intersection of the boundaries of the two objects, denoted as $\partial O_{ik} \cap \partial O_{jl}$; (2) The intersection of the boundary of the first object with the interior of the second object, denoted as $\partial O_{ik} \cap O_{jl}{}^o$; (3) The intersection of the interior of the first object with the boundary of the second object, denoted as $O_{ik}{}^o \cap \partial O_{jl}$. (4) The intersection of the interiors of the two objects, denoted as $O_{ik}{}^o \cap O_{jl}{}^o$.

Each of these four sets may be empty $(\varnothing)$ or non-empty $(\neg\varnothing)$. This results in a total of $2^4 = 16$ combinations. Only eight of these combinations result in valid topological relations in $\Re^2$. The remaining combinations do not produce meaningful topological relations. The different relation types between pairs of shape objects are:

(1) $(\varnothing,\varnothing,\varnothing,\varnothing) \rightarrow O_{ik} \ \ Disjoint \ \ O_{jl}$ (2) $(\varnothing,\neg\varnothing,\varnothing,\neg\varnothing) \rightarrow O_{ik} \ \ Contains \ \ O_{jl}$

(3) $(\varnothing,\neg\varnothing,\neg\varnothing,\varnothing) \rightarrow O_{ik}$ *Inside* O_{jl} (4) $(\neg\varnothing,\varnothing,\varnothing,\varnothing) \rightarrow O_{ik}$ *Meet* O_{jl}

(5) $(\neg\varnothing,\neg\varnothing,\varnothing,\varnothing) \rightarrow O_{ik}$ *Equal* O_{jl} (6) $(\neg\varnothing,\neg\varnothing,\varnothing,\neg\varnothing) \rightarrow O_{ik}$ *Covers* O_{jl}

(7) $(\neg\varnothing,\neg\varnothing,\neg\varnothing,\varnothing) \rightarrow O_{ik}$ *Covered − by* O_{jl} (8) $(\neg\varnothing,\neg\varnothing,\neg\varnothing,\neg\varnothing) \rightarrow O_{ik}$ *Overlap* O_{jl}.

The relations Disjoint, Contains, Inside, Meet, Equal, Covers, Covered-by, and Overlap provide complete coverage for relations between pairs of shape objects, and are mutually exclusive, i.e. exactly one of these topological relations holds true between any two shape-objects in $\Re^2$.

This spatial representation can accommodate data from one or multiple source GIS databases by implementing import routines, which transfer the objects from the source database(s) to the target database. Since most GIS formats currently store data in terms of objects, there is a need to detect the implicit relations between the objects once they have been converted to the TopDB format. A method to detect such relations between pairs of neighboring objects is discussed in Section 4.

4 Detection of Implicit Spatial Relations

Spatial queries can be easily solved if all geometric relations between the objects of interest are explicitly stored. However, such a scenario is unrealistic, even for relatively small data collections (Davis, 1986). It would need tremendous amounts of storage space (n^2 values for each kind of spatial relation between n objects). Since objects far apart are not likely to be related, relations between objects are of interest for those objects that are at a close proximity.

Therefore, the neighbors of each object in the database are first determined (see Section 4.1). This is followed by detecting the types of relations that exist for these neighboring objects (see Section 4.2).

4.1 Detection of Neighboring Objects

The neighboring objects are determined for objects in the database. A threshold distance y_{ij} between all objects on layers L_i and L_j is used to determine whether two objects are at a close proximity or not. If the distance between two objects is less than or equals to y_{ij}, the two objects are said to be Neighboring Objects, otherwise they are not.

Definition 4 – Neighboring Objects

Two objects O_{ik} and O_{jl} are said to be neighbors if the distance between them is less then the threshold y_{ij}, i.e. $NBR(O_{ik},O_{jl}) = \begin{cases} T & if \quad D(O_{ik},O_{jl}) \le y_{ij} \\ F & otherwise. \end{cases}$ Where $D(O_{ik},O_{jl})$ is the Euclidean distance between objects O_{ik} and O_{jl}.

Each pair of layers L_i and L_j in a spatial database is assigned a corresponding threshold y_{ij}. The y_{ij} may be set to the same or different values as other layer pairs. This will depend on what the corresponding objects represent in the real world. The value y_{ij} for each pair of layers L_i and L_j is defined as follows: (1) Consider a representative sample of objects on L_i; (2) Determine the average length of the diagonal of the bounding rectangle (which is the smallest

rectangle that contains an object) of these objects. Denote it $\partial(L_i)$; (3) The minimal threshold for objects on layer L_i is set to the average length of the diagonal.

This procedure is performed for each data layer. Then, the threshold y_{ij} is the smaller of the two values, that is $y_{ij} = \min(\partial(L_i), \partial(L_j))$.

This idea is illustrated with the following examples. Consider a layer L_1 with objects representing trees, where $\partial(L_1) = 3$ meters, and a layer L_2 with objects representing properties where $\partial(L_2) = 27$ meters. Then two trees are considered neighbors if they are at most $y_{11} = \min(\partial(L_1), \partial(L_1)) = \min(3,3) = 3$ meters apart; Two properties are considered neighbors if they are at most $y_{11} = \min(\partial(L_2), \partial(L_2)) = \min(27,27) = 27$ meters apart; and a tree and a property are considered neighbors if they are at most $y_{11} = \min(\partial(L_1), \partial(L_2)) = \min(3,27) = 3$ meters apart.

4.2 Relation Type Detection Among Neighboring Objects

The relation between each pair of neighboring objects is determined.

Definition 5 - Observed Relations

Let a_{ikjl} denote the observed relation between object O_{ik} on layer L_i and O_{jl} on layer L_j. The observed relation can take one of eight relation type (denoted by q) values $a_{ikjl} \in \{q \mid q = 1,2,...8\}$, where $q = 1$ if the relation between O_{ik} and O_{jl} is Equal, etc.

Therefore two neighboring objects O_{ik} and O_{jl} are observed to have relation $a_{ikjl} \in \{1,2,...,8\}$.

The content of TopDB databases can now be analyzed to discover interesting relationships in the data.

5 Rule Discovery

Automated rule discovery in spatial databases is achieved by developing statistical techniques to quantify relations among objects in spatial data using the TopDB data model presented in the previous section.

Rules, which consider topological relations alone, are based only on the shape of the object. This is not sufficient for detecting inconsistencies in spatial data, as this does not translate to comprehensive descriptions of real world situations. Semantic information is defined using the meaning of geographic objects, that is, the meaning of real world entities described by the objects. This information is combined with topological relations to discover rules.

Considering the semantics of objects, it is possible to define which topological relations are consistent and which are not. For example, a building inside a residential property is a permitted relation, while a building inside a road is not. In either case, the topology of the scenario is a polygon inside another polygon. Therefore the semantics of each object is required to determine which relationship is allowed and which is not.

Rules governing relations between objects can be broadly classified based on the semantics of the real world entities the objects represent. Since each layer's objects correspond to real world entities with different semantic meaning, this discussion continues in terms of layers.

Rules governing relations between objects are therefore determined based on the layers of corresponding objects.

Two types of rules are defined. Rules based on the analysis of the relation between each object and each of their neighbors are discussed in Section 5.1. Additional rules based on the analysis of the relations between each object and all of its neighboring objects are presented in Section 5.2.

Another possible source for rules that govern relations between objects is the knowledge that exists about the data in documentation or in human experts. This information is therefore said to be part of the metadata of the GIS. Rules derived from the metadata may be used to validate the discovered rules. In the absence of metadata-based rules, discovered rules can be validated using trend analysis techniques and monte carlo simulation techniques.

5.1 Rule Discovery Based on Pairs of Neighboring Objects

The aim is to determine rules such as "A tree is Inside a property" or "A road Meets a property". This section presents the process by which such rules are discovered.

The observed relation $a_{ik}j_l$ of type q between objects O_{ik} and O_{jl} is a nominal level measurement. No assumptions of ordering or distance are made. For such information measured at the nominal scale, central tendency is evaluated by the mode – the class of highest frequency in the distribution.

Let f_{ij}^{q} denote the number of object pairs O_{ik} and O_{jl} on layers L_i and L_j that have an observed relation $a_{ik}j_l$ of type q in the database.

Observe all relations $a_{ik}j_l$ between pairs of neighboring objects for layers L_i and L_j. The relation type q that has the largest frequency f_{ij}^{q} in this set of observations is called the Mode Relation Type and is denoted as $\overline{a}_{ij}$. It represents the central tendency such that if a relation between two objects O_{ik} on layer L_i and O_{jl} on layer L_j were randomly selected, it is most likely that the observed relation $a_{ik}j_l$ between the two objects would have the same relation type as the mode relation type $\overline{a}_{ij}$.

Let $\underset{q}{Max}\left(f_{ij}^{q}\right)$ be the frequency of the mode relation type $\overline{a}_{ij}$ between objects that exist on layers L_i and L_j. Let n_{ij} be the total population of all relation types between all neighboring objects on layers L_i and L_j. The dispersion of the nominal scale measurement $\overline{a}_{ij}$ is evaluated by the variation ratio v_{ij} where $v_{ij} = 1 - \left(Max\left(f_{ij}^{q}\right) \Big/ n_{ij}\right)$.

A smaller value of the variation ratio v_{ij} indicates a more concentrated case, and therefore the mode is a better indicator of the trend. For example, if the dominant relation type is the relation disjoint, and it occurs 80% of the time, then $v_{ij} = 0.2$. If this dominant relation type occurs 40% of the time, $v_{ij} = 0.6$. In the first case the mode is a better indicator of the trend.

Variation ratio v_{ij} information corresponding to each mode relation type $\overline{a}_{ij}$ for layers L_i and L_j is stored in a relational table.

Binary rules between pairs of objects on layers L_i and L_j that have relation type q are written as r_{ij}^{q}, $q \in \{1,2,...8\}$, that is $r_{ij}^{q} \in \{T,U\}$.

Definition 6 - P-Rules.

A rule based on pairs of neighboring objects (P-Rule) is created indicating that a relation may exist for object O_{ik} and object O_{jl} if there exists a small variation ratio $v_{ij} \leq t$ for one of the eight types of relations. This is stated as: $r_{ij}^{\overline{q}} = T$. A P-Rule indicating that the existence of a relation is not known for relation type q is stated as: $r_{ij}^{q} = U$.

These P-Rules are stored in groups of eight based on the levels of the objects they govern in a relational table. The particular rule stored in this table indicates that objects on layer L_i can only be Disjoint with objects on layer L_j.

The threshold t is derived using the monte-carlo simulation and comparing results to metadata based rules.

5.2 Rule Discovery Based on Multiple Neighboring Objects

In many situations there may not be one dominant relation type, since the dispersion of $\overline{a}_{ij}$, namely v_{ij}, is greater than the optimal threshold t. Therefore the discovered P-Rules are stated as $r_{ij}^{q} = U$. In such cases, additional information about those relations which occur frequently in the data, and therefore likely to form rules, and those which do not, must be established.

The aim is to determine rules such as "a Building is Inside a Property 1 time" or "A property Meets a road between 0 and 2 times". Such rules are established as follows: (1) Determine the number of times each object O_{ik} is related to it's neighbors on a layer L_j in a relation type q; (2) Determine the cumulative number of times all objects on layer L_i are related to other objects on layer L_j in a relation type q (3) Determine the average number of times an object on layer L_i is related to an object on layer L_j in a relation type q; (4) Determine the standard deviation for the mean number of times an object on layer L_i is related to an object on layer L_j in a relation type q; (5) Determine the lower and upper bound for the number of times an object on layer L_i is likely to be related (95% of the time) to an object on layer L_j in a relation type q.

The following is a detailed description of this process:

Calculate of the number of times the object participates in each $a_{ik}j_l$ relation with all its neighbors on each of the layers. Let b_{ikj}^{q} be the number of times an object O_{ik} is related to all its neighboring objects that are on layer L_j in a relation of a relation type q. Store the following information for each object: The object, its layer and the layer of the second object, as well as the relation type and frequency are stored in a relational table.

Calculate the cumulative number of times all objects on layer L_i are related to other objects on layer L_j in a relation type q. The frequency j_{ij}^{q} is defined as the cumulative number of

times the objects O_{ik}, $k = 1..K_{ij}^q$ are related to all neighbor objects O_{jl} in a relation of type q. That is $j_{ij}^q = \sum_{k=1}^{K_{ij}^q} b_{ikj}^q$.

Calculate the average number of times an object on layer L_i is related to an object on layer L_j in a relation type q. The average Frequency m_{ij}^q is defined as the average number of times each of the objects O_{ik}, $k = 1..K_{ij}^q$ are related to all neighbor objects O_{jl} on layer L_j in a relation of type q. That is $m_{ij}^q = \frac{j_{ij}^q}{K_{ij}^q}$, where K_{ij}^q is the total number of objects on L_i that are related to at least one object on L_j in a relation of type q.

The standard deviation for the average number of times an object on layer L_i is related to an object on layer L_j in a relation type q. The standard deviation S_{ij}^q is calculated as follows:

$$S_{ij}^q = \sqrt{\frac{\sum_{k=1}^{K^q} \left(b_{ikj}^q - m_{ij}^q \right)^2}{n-1}}$$

The larger the standard deviation S_{ij}^q, the greater is the dispersion, while a smaller standard deviation indicates more concentrated distribution of the number of times a particular type of relation q occurs in relations between objects on layers L_i and L_j.

The lower and upper bound for the number of times an object on layer L_i is likely to be related (95% of the time) to an object on layer L_j in a relation type q. According to the empirical rule, for normal distribution of our data, which is assumed here, the range $m_{ij}^q \pm 2S_{ij}^q$ will contain approximately 95% of the data.

A rule based on multiple neighboring objects (M-Rule) is created for object O_{ik} as follows:

An object O_{ik} should be related to one or more objects on L_j within the range of $m_{ij}^q \pm 2S_{ij}^q$. The lower bound is $m_{ij}^q - 2S_{ij}^q$ rounded down to the next integer value. The upper bound is $m_{ij}^q + 2S_{ij}^q$ rounded up to the next integer value. This rounding up and down is needed since the number of objects in the various situations encountered is an integer.

If an object O_{ik} is found related to one or more objects on L_j outside the range of $m_{ij}^q \pm 2S_{ij}^q$ an inconsistency, which includes object O_{ik} is said to exist. These M-Rules are placed in a relational table.

That is, M-Rules are defined in terms of two layers, a relation type, and upper and lower bounds for the number of objects on the second layer that can be in the specified relation type with the object on the first layer.

Since automated rule discovery is proposed, the validity of these rules must be established. The following section presents some possible techniques for validation of these discovered rules.

5.3 Rule Validation

The discovered P-Rules and M-Rules play important roles in detection and adjustment of inconsistencies. It is important to check their validity, therefore, three methods for the validation of rules are proposed. Discovered rules can be validated based on a comparison with metadata-based rules if such rules are readily available; they can be validated with trend analysis; or they can be validated with Monte-Carlo simulation. These methods can be used individually or in combination. The three methods are briefly discussed in the following paragraphs:

5.3.1 Validation Against Metadata-Based Rules

Rules may be manually derived from the metadata of the GIS being analyzed. These metadata-based rules are created by reviewing documentation pertaining to the content and structure of the elements of the various data layers in the GIS or conducting interviews to solicit expert knowledge. The discovered rules can be presented to users via a simple graphic user interface that shows the users each rule and asks them to confirm or reject the discussed rules based on the metadata. This approach is possible when improvement of the consistency of a spatial database in a production environment is needed. It can however be costly, time consuming, and prone to human error.

In situations where improvement of consistency of one or more data sets in a query environment is required, for transmittal to clients over the Internet or by other means, this approach proves impractical. The following two sections present automated approaches to verify the validity of the discovered rules.

5.3.2 Validation Using Trend Analysis

With this approach to rule validation, rules are discovered for the complete region; the region is then subdivided into four quadrants and the rule discovery process is repeated for each of the quadrants of the region. If any discovered rule is 'valid' and applicable throughout the region, the same rule should be discovered in each of the quadrants as well. Such rules will be called stationary rules. Rules that are found to be different in one or more quadrants in the region are said to be non-stationary in these quadrants. Alternate variations of validation using trend analysis are also possible.

5.3.3 Validation Using Monte Carlo Simulation

The discovered rules can be further validated by application of a monte carlo simulation. The idea of this method is to run the rule discovery N times, each resulting in a set of rules; then compute and store sample statistics from the N output sets of rules. The accuracy of the monte carlo method is inversely related to the square root of the number of runs N. This means that to double the confidence in the rules, four times as many runs are needed. Each run, the following parameters are slightly altered: (1) The minimum acceptable variation ratio for creation of P-Rules. Potential rules where the variation ratio is greater than this number will be eliminated; (2) The minimum number of objects for formation of P-Rules. P-Rules must not be formed based on a relatively small number of observed relations. These may be actual inconsistencies rather then rules, and will be processed during the Inconsistency Detection process; (3) The minimum number of objects for M-Rules. M-Rules must not be formed based on a relatively small number of observed relations. These may be actual inconsistencies rather then rules, and will be processed during the Inconsistency Detection process.

6 Conclusion

The TopDB model was defined in terms of objects along with their semantic information to represent real world entities, as well as topological relations to represent the interaction between the entities in the real world.

Statistical techniques were developed to quantify relations among objects in spatial data. The proposed rules are based on the analysis of the relation between each object and each of its neighbors, as well as between each object and all of its neighboring objects. The semantic information of objects, together with topological relations, are combined to discover rules. The rules therefore, capture the essence of the corresponding laws that govern the entities in the real world.

It is important to make sure that rules are valid since they play a key role in the proposed inconsistency detection process. Rules derived from the metadata may be used to validate the discovered rules in a production environment. In a query environment, discovered rules can be validated using trend analysis techniques, and monte carlo simulation techniques.

The rules discovery techniques discussed in this paper serve as the first component of a spatial data quality management system. The discovered rules serve as input to processes which detect inconsistencies between pairs of objects as well as among multiple objects in spatial data. These inconsistencies are then adjusted as part of the spatial quality management system which is described in a pending publication. The techniques for the discovery of rules in spatial data can also be utilized in the field of spatial data mining.

The statistics used to discover rules can be used as a means for describing the structure of spatial data. There are a large number of spatial databases with vast amounts of data, which would benefit from discovery of meaningful and useful knowledge. Such knowledge will be vastly compounded if multiple databases are analyzed simultaneously for the same region.

REFERENCES

Egenhofer, M.J., Franzosa, R.D. 1995. On Equivalence of Topological Relations, 1995, *International Journal of Geographical Information Systems*, Vol. 9, No. 2, p133-152.

Clementini, E., Felice, P.D., Van Oosterom, P., 1993, *A Small Set of Formal Topological Relations Suitable for End-User Interaction*, Advances in Spatial Databases, 3rd International Symposium, SSD'93, Singapore, Lecture Notes in Computer Science, Vol. 692 (Springer-Verlag, New York) p277-295.

Abler, R., 1987, *The national science foundation national center for geographic information and analysis*, International Journal of Geographic Information Systems, 1(4): p303-326.

Egenhofer, M.J., Franzosa, R.D., 1991, *Point-set topological spatial relations*, International Journal of Geographic Information Systems, 5(2): p161-174.

Egenhofer, M.J., Herring, J.R., 1992, *Categorizing binary topological relations between regions, lines, and points in geographic databases*, Technical report, Department of Survey Engineering, University of Maine, Orono, ME.

Mark, D.M., Egenhofer, M.J., 1992, *An evaluation of the 9-intersection for region line relations*, In Proceedings of GIS/LIS '93, San Jose, CA, p513-521.

Hadzilacos, T., Tryfona, N., in Frank, A., Campari, I., and Formentini, U., 1992, *A Model for Expressing Topological Integrity Constraints in Geographic Databases*, Proceedings of the International Conference on Theories and Models of Spatio Temporal Reasoning in Geographic Space. Pisa, Italy. Lecture notes in Computer Science, Vol. 639 (Springer Verlag, New York), p252-268.

Hayes-Roth, F., Waterman, D., and Lenat, D., 1983, *Building Expert Systems*. Addison-Wesley Publishing Company, Reading, MA.

Mark, D., Frank, A., Egenhofer, M., Freundshuh, S., McGranghan, M., White, R.M. editors, 1989, *Languages of Spatial Relations: Report on the Specialist Meeting for NCGIA Research Initiative* 2. Technical Report 89-2, national Center for Geographic Information and Analysis.

Egenhofer, M.J., Herring, J.R., 1991, *A Mathematical Framework for the Definition of Topological Relations*, NCGIA technical paper 91-97

Smith, R., Park, K.K., 1991, *An Algebraic Approach to Spatial Reasoning*, NCGIA technical paper 91-97.

Davis, E., 1986, *Representing and Acquiring Geographic Knowledge*, Morgan Kaufmann Publishers inc. Los Altos, CA.

Egenhofer, M., Herring, J., 1990, *A Mathematical framework for the definition of Topological Relations*, Proceedings of Fourth International Symposium on Spatial Data Handling, 803-813. Zurich, Switzerland.

Veregin, H., 1991, *Data Quality Measurement and Assessment*, NCGIA Core Curriculum in Geographic Information Science.

Redman, T.C., 1992, *Data Quality*, Bantam.

Aronoff, S., 1995, *Geographic Information Systems: A Management Perspective*, WDL Publications.

Servigne, S., Ubeda, T., Zuricelli, A., Laurini, R., 2000, *A Methodology for Spatial Consistency Improvement of Geographic Databases*, GeoInformatica 4:1, p.7-34.

Medeiros, C.B., Cilia, M., 1995, *Maintenance of Binary Topological Constraints Through Active Databases*. Proceedings of 3rd ACM Workshop on GIS.