Proceedings of the 6th AGILE
April 24th-26th, 2003 – Lyon, France

CONSISTENCY CHECKS FOR DATA QUALITY ASSESSMENT

Anders Olsson

Luleå University of Technology, Div of Geographical Information Technology

SE-971 87 Luleå, Sweden; Ph. +46 920 491466; Email: anders.olsson@sb.luth.se

1. INTRODUCTION

The Swedish National Road database (NVDB) has been produced by Swedish National Road Administration during the last four years. It contains all public, and some private, roads in Sweden, as well on the countryside as in the cities. However the completeness of the database varies considerably. This means that for instance that the database is less suitable for navigation purposes in the cities as compared to countryside.

Although high ambition in the data production process, important data quality measures such as completeness are missing. This implies that the usability of the database is difficult, if not impossible, to estimate.

To estimate the completeness of the database automatic procedures are preferred, combined with sampling strategies. Since consistency constraint can be used for error detection, the use of such constraints may be used for quality assessment.

Consistency is a term the specify the difference between as instanciation of objects and the consistency constraints as defined by the class definition. An class model is a model that describes a given application domain by a set of class definition. This mode usually includes a set of constraints such as topological, referential and domain consistency constraints. The class is populated through instanciation, where objects are created. However, in this process it is possible that some consistency constraints are violated.

Topological constraints describes spatial constraint between features that are based on closeness Example of such inconsistencies can be a polygon that is not closed, a polygon without any reference point or whit two reference points, two lines that intersect whiteout node and lines that undershoot or overshoot.

Domain consistency concerns the attribute of an object. Examples of such constraints are that values must be in a certain range, a datum must be in a certain format and according to a valid calendar, or that the values belongs to a set of predefined values (code list).

Referential constraints describes constrains on non-spatial relation between objects. An example of such constraint may be that a real estate must have a owner.

Consistency rules are mainly applied for detection of errors. Topological consistency rules for spatial connectivity are also important in spatial modeling and analysis, for instance in analysis of street networks.

Most datasets that are delivered for use in different information systems have been checked with respect to consistency and corrected by the producer. But, due to the advent of web services, where different datasets can be merged on the fly, the consistency among several datasets has to be considered. In such a case, mainly topological inconsistencies will appear, but other types of consistency rules may also be formulated. But topological inconsistencies are important, since they may cause analysis to fail, due to improper spatial connectivity.

One common problem for database producers is to estimate the data quality of their product. Especially completeness is considered to be a major problem. But since inconsistencies can be used for detecting errors, at least to some extent, the question is if inconsistencies in merged datasets can be used for this type of quality assessment.

The purpose of this paper is to study, to what degree inconsistencies among different datasets can be used for quality assessment. As a pilot, two different road datasets have been used. The goal is a method, which can estimate the completeness of the Swedish National Road Database (NVDB) within the major cities.

The theory used in this paper for investigate relationships between objects are based on a the dimensionally extended nine intersection model [2] that is a extension from the 9 intersection model [3] and the 4 Intersection Model [4]

2. CHECKING CONSISTENCY

Topological consistency may be described by using point-set topology. Point-set topology is describing the spatial relation between sets using the intersection between boundaries and interiors of two sets [1]. Definitions for interior and boundary are based on concept of closed and open set.

The *interior*, denoted by A°, of a set A is the union of all open set that are contained in A. The *closure*, denoted by $\bar{A}$, is the intersection of all closed set that contain A. The *exterior* with respect to embedding space $\Re^n$, denoted by A^-, is the set of all points of $\Re^n$ not contained in A. The *boundary*, denoted ∂A, is the intersection of the closure of A and the closure of A^-.

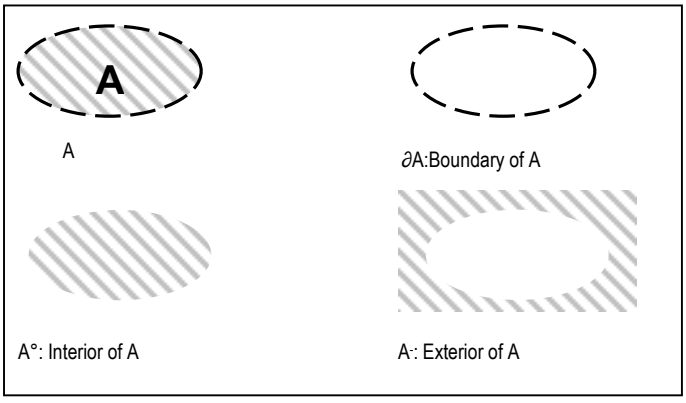


Fig. 1 Boundary, Interior and Exterior of a spatial region without hole.

To define the binary topological relation between two spatial regions, A and B, the intersections of A's interior (A°), boundary (∂A) and exterior (A^-) with the interior (B°), boundary (∂B) and exterior (B^-) of B is considered [3]. These intersections can assume value empty or non-empty. The 9 Intersection introduced by [3] is a way of denoting spatial relationships (1)

$$R(A,B) = \begin{pmatrix} A^\circ \cap B^\circ & \partial A \cap B^\circ & A^\circ \cap B^- \\ A^\circ \cap \partial B & \partial A \cap \partial B & \partial A \cap B^- \\ A^\circ \cap B^- & \partial A \cap B^- & A^- \cap B^- \end{pmatrix} \quad (1)$$

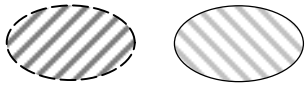
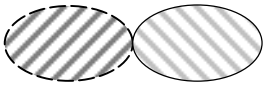
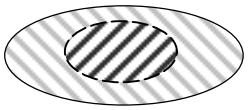
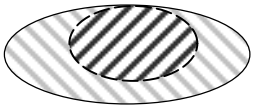
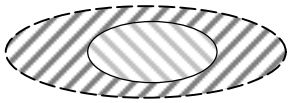
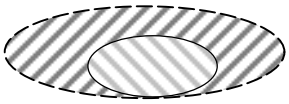
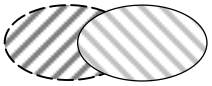
Using the 4 Intersection Model to describe topological relationship eight possible topological relations can be established for two spatial regions without hole embedded in a 2-D space [4]. They are *disjoint, meet, equal, inside, coveredBy, contains, covers, overlap*. When using the 9 Intersection Model the same eight topological relations can be established [3]. These eighth relations provides a complete coverage and the relation between two objects can be categorize exactly as one of the relations

Further work has extended the 9 Intersection Model to also include the dimension (*dim*) of the intersection between the *interior, boundary* and *exterior* of geometries [5]. *dim* returns the dimension of a point set. In the definition (2) *S* is a general point-set, which may be several disconnected part.

$$\dim(S) = \begin{cases} - & \text{if S} = \phi \\ 0 & \text{if S contains at least a point and no lines or areas} \\ 1 & \text{if S contains at least a line and no areas} \\ 2 & \text{if S contains at least an area} \end{cases} \quad (2)$$

The dimensionally extended nine intersections model (DE-9IM) has the form (3):

$$M(A,B) = \begin{pmatrix} \dim(A^\circ \cap B^\circ) & \dim(\partial A \cap B^\circ) & \dim(A^\circ \cap B^-) \\ \dim(A^\circ \cap \partial B) & \dim(\partial A \cap \partial B) & \dim(\partial A \cap B^-) \\ \dim(A^\circ \cap B^-) & \dim(\partial A \cap B^-) & \dim(A^- \cap B^-) \end{pmatrix} \quad (3)$$

Relationship (AB)		$\begin{pmatrix} \dim(A^\circ \cap B^\circ) & \dim(\partial A \cap B^\circ) & \dim(A^- \cap B^\circ) \\ \dim(A^\circ \cap \partial B) & \dim(\partial A \cap \partial B) & \dim(\partial A \cap B^-) \\ \dim(A^\circ \cap B^-) & \dim(\partial A \cap B^-) & \dim(A^- \cap B^-) \end{pmatrix}$
Disjoint		$\begin{pmatrix} -1 & -1 & 2 \\ -1 & -1 & 1 \\ 2 & 1 & 2 \end{pmatrix}$
Meet		$\begin{pmatrix} -1 & -1 & 2 \\ -1 & 1 & 1 \\ 2 & 1 & 2 \end{pmatrix}$
Equal		$\begin{pmatrix} 2 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 2 \end{pmatrix}$
Inside		$\begin{pmatrix} 2 & -1 & -1 \\ 1 & -1 & -1 \\ 2 & 1 & 2 \end{pmatrix}$
CoveredBy		$\begin{pmatrix} 2 & -1 & -1 \\ 1 & 1 & -1 \\ 2 & 1 & 2 \end{pmatrix}$
Contains		$\begin{pmatrix} 2 & 1 & 2 \\ -1 & -1 & 1 \\ -1 & -1 & 2 \end{pmatrix}$
Covers		$\begin{pmatrix} 2 & 1 & 2 \\ -1 & 1 & 2 \\ -1 & -1 & 2 \end{pmatrix}$
Overlap		$\begin{pmatrix} 2 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 2 \end{pmatrix}$

∩ Intersection, dim(x) resturn the dimensuion of the intersection, x {-1, 0, 1,2}.

Fig. 2 Possible relationship between regions without holeand there DE-91M. from Egenhoefer et. al. and Celementini and Di Felicie.

A constraint consists of two parts namely a condition and a list of object's that must fulfill the condition. If two objects violate a constraint, there is a possible inconsistency. Depending on the situation, different action may than be performed.

A constraint also has a specification that describes the cardinality of the relationship, the cardinality can be either 1) Forbidden, 2) At least *n* times, 3) At most *n* times or 4) Exactly *n* times. The specification *forbidden* is the interesting and usable one [6].

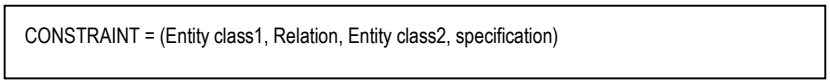


Fig. 3 The definition of a topological constraint.

3. CONSISTENCY BETWEEN LARGE SCALE BASE MAP AND NATIONAL ROAD DATABASE

3.1 National Road Database

The Swedish government has issued directives to create a nationwide road database. In the database the roads, state, municipal and private, shall be represented with specified attributes, speed limit, road number, road width, bearing capacity, etc.[7] The project is a collaboration between Swedish National Road Administration (SNRA), Central office of the National Land Survey, the Swedish Associations for Local Authorities and the forestry industry. The database is known by its Swedish abbreviation NVDB (National Road Database). A central aspect of NVDB is that data should be up-to-date and of high and documented quality.

3.2 Large Scale Base Map

Large Scale Base Maps are produced by the Municipal authorities. They are produced over densely populated communities and usual in scale 1:400-1:500. The roads in these datasets usually are represented by their outer edge.

3.3 Experiment

The approach of this study is to establish relationships between the two dataset, in order to detect the degree of completeness of NVDB. The constraints are based on the polygons formed by the different road segment (fig. 4).

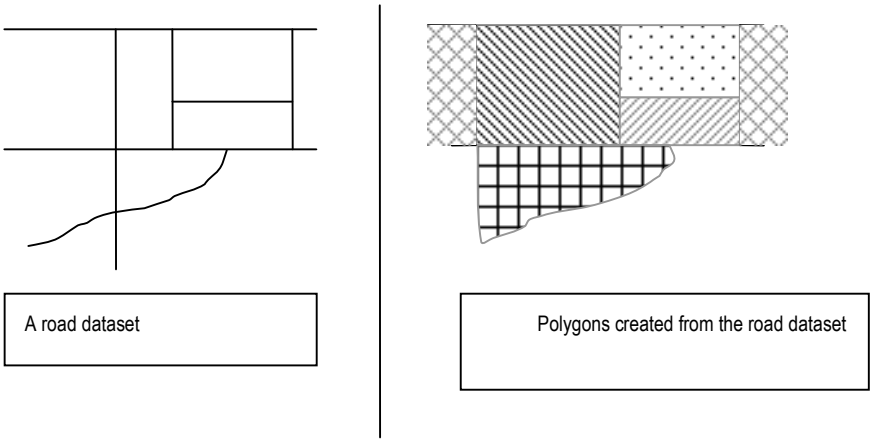


Fig. 4 One line (road) dataset and olyguon dataset genrated fron the NVBD Dataset

If we don't have any geometrical distortion, polygons created from Large Scale Base Map will be smaller than the polygons created from NVDB (fig. 5). This relation can be expressed as the *contains* constraint.

If the NVDB dataset is complete, relation operator *contains* between an NVDB polygon and Large Scale Base Map will have 1:1 relationship, i.e. only one polygon from Large Scale Base Map can be in each polygon from NVDB. If a NVDB polygon contains two Large Scale Base Map, street segments are the NVDB dataset is missing

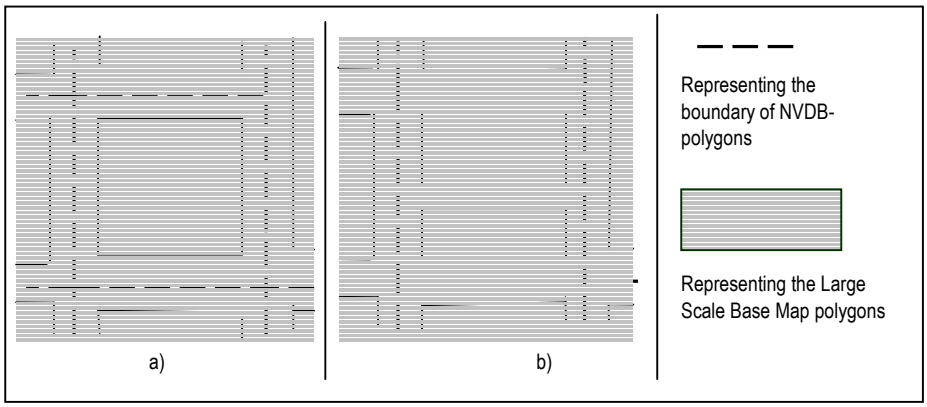


Fig. 5 a) NVDB and Large Scale Base Map Polygons , b) a NVDB polygons contains two Large Scale Base Map ploygons.

If there are geometric error for NVDB a Large Scale polygons may overlap several NVDB polygon (fig.6).

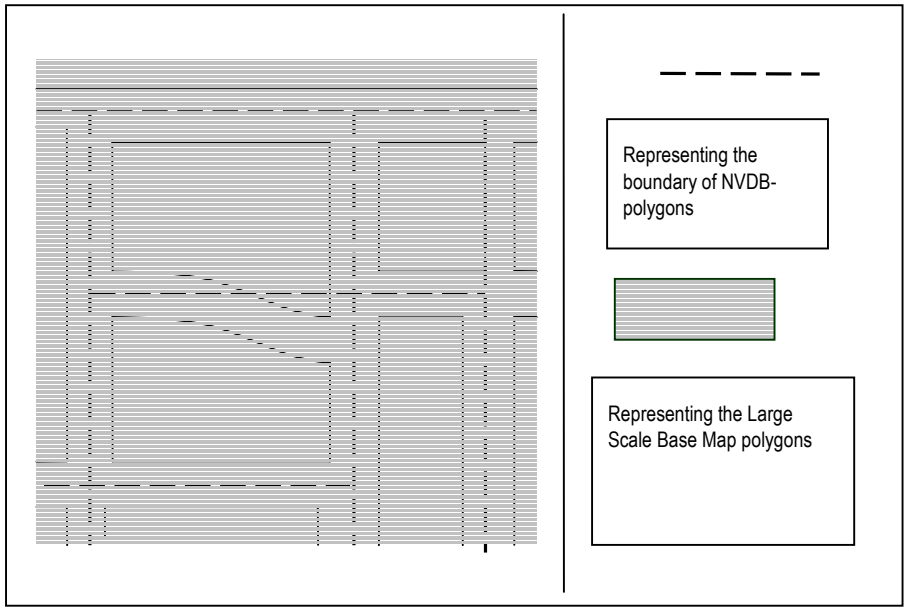


Fig. 6 Large Scale Base Maps Polygon Is overlapping two NVDB Polygon and a Polygon covers a Large Scale Polygon.

As already mentioned, constraints are used for checking inconsistency. In this experiment we want to find those polygons which interior-interior intersection has dimension 2, (dim (A° ∩ B°) =2) i.e. the intersection is a polygon. The interior-interior intersection in

non-empty in six topological relation (*equal, inside, coveredBy, contains, covers overlap*). Therefore six different constraints are formulated, (fig. 7)shows the constraint used.

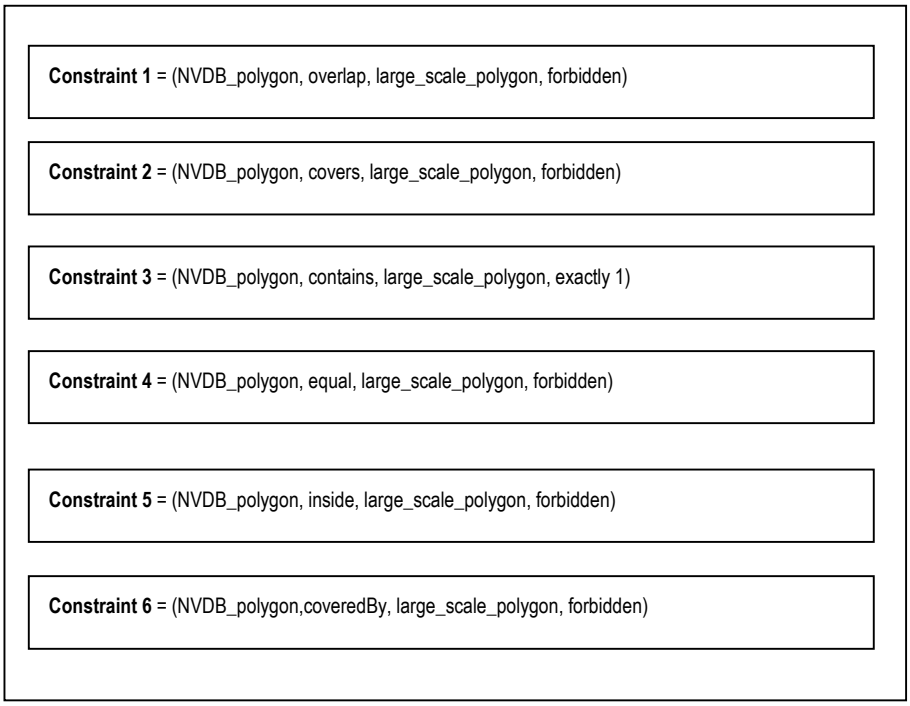


Fig. 7 Constraints used in experiment.

The first step is the procedure is to eliminate the effect of geometric errors in the NVDB dataset. This is achieved by using the overlap operators (constraint 1). If a Large Scale Polygon is overlapping two or several NVDB polygons (fig. 6), the Large Scale polygon is split and the one with smallest area is removed. Apart from the area calculating, this is completely a topological operation, so no geometric modifications needs to be done. Constraint number 2 (the covers operator) also detect geometric errors. If a NVDB polygon covers a Large Scale Polygon (fig. 6), the Large Scale Polygon is split. It can the be used to detect missing street in the NVDB dataset.

To investigate if there are any missing streets in NVDB constraint number 3 is used. If that are violated this indicates that a situation shown in (fig. 5 b)) exist, and there could be an inconsistence in the dataset.

The dataset are so Constraint 3-6 are violated this indicates a big geometric error and this demands a further investigation.

4. RESULT

No result yet obtained, but will be presented at the conference.

5. BIBLIOGRAPHICAL REFERENCES

5.1 Books

[1] Alexandroff. P., *Elementary Concepts of Topology*, Dover Publications Inc., New York, NY, 1961.

5.2 Periodicals, Journals, Website

[2] Clementini E & Di Felice P., «A Comparison of Methods for Representing Topological Relationships», *Information Systems, vol 3*, 149—178, 1995

[3] Egenhofer M. J. & Herring J., «Categorising Binary Topological Relationships Between Region, Lines and Points in Geographic databases», Technical report, Department of Surveying Engineering, University of Maine, Orono, ME, 1991.

[4] Egenhofer M. J. & Franzosa R.D., «Point-Set topological spatial relations», *International journal of Geographic Information System, vol 5,* number 2, 161-174, 1991.

[5] E. Clementini, E. Di Felice, P. and van Oosterom, P., «A Small Set of Formal Topological Relationships for End-User Interaction», *Advances in Spatial Databases - Third International Symposium, SSD'93* ; *Lecture Notes in Computer Science LNCS 692*, 277-295, Springer-Verlag, Singapore, 1993.

[6] Ubeda T. & Egenhofer M. J., «Topological Error Correcting in GIS, Advances in Spatial Databases», *Fifth International Symposium on Large Spatial Databases, SSD `97* ; *Lecture Notes in Computer Science Vol. 1262*, 283-297, Springer-Verlag, 1997.

[7] Swedish National Road Administration. http://www.vv.se/nvdb/en/index.asp, 2002-11-08.