RESUMEN DE TESIS DOCTORAL

Similitud Semántica entre Sistemas de Objetos Geográficos Aplicada a la Generalización de Datos Geoespaciales

Semantic Similarity between Systems of Geographic Objects Applied to Generalization of Geospatial Data

Graduated: Marco Antonio Moreno Ibarra
Centro de Investigación en Computación del IPN
Av. Juan de Dios Bátiz s/n Esq. Miguel Othón de Mendizábal
C.P. 07738 México D.F.
Graduated in October 25, 2007
marcomoreno@cic.ipn.mx

Advisor: **Serguei Levachkine**
Centro de Investigación en Computación del IPN
Av. Juan de Dios Bátiz s/n Esq. Miguel Othón de Mendizábal
C.P. 07738 México D.F.
sergei@cic.ipn.mx

Abstract
The thesis presents an approach to verify the consistency of generalized geospatial data at a conceptual level. The principal stages of proposed methodology are Analysis, Synthesis, and Verification. Analysis is focused on extracting the peculiarities of spatial relations by means of quantitative measures. Synthesis is used to generate a conceptual representation (ontology) that explicitly and qualitatively represents the relations between geospatial objects, resulting in tuples called herein semantic descriptions. Verification consists of a comparison between two semantic descriptions (description of source and generalized data): we measure the semantic distance (confusion) between ontology local concepts, generating three global concepts Equal, Unequal, and Equivalent. They measure the (in) consistency of generalized data: Equal and Equivalent – their consistency, while Unequal – an inconsistency. The method does not depend on coordinates, scales, units of measure, cartographic projection, representation format, geometric primitives, and so on. The approach is applied and tested on the generalization of two topographic layers: rivers and elevation contour lines (case of study).
Keywords: semantic similarity, generalization, ontology, geographic objects

Resumen
Esta tesis presenta un método para verificar la consistencia de datos geoespaciales generalizados, utilizando únicamente una representación conceptual. Las principales etapas de la solución propuesta son Análisis, Síntesis y Verificación. El Análisis tiene como propósito extraer las particularidades que tienen los objetos geográficos usando medidas cuantitativas. La Síntesis tiene como propósito generar una representación conceptual (ontología) que cualitativa y explícitamente representa las relaciones entre los objetos geoespaciales, obteniendo tuplas llamadas descripciones semánticas. La Verificación consiste de una comparación entre dos descripciones semánticas (descripción de los datos fuente y de los datos generalizados): consiste en medir la distancia semántica (confusión) entre los conceptos locales de la ontología, generando tres conceptos: Igual, Desigual y Equivalente, los cuales miden la (in)consistencia de los datos generalizados. Igual y Equivalente representan consistencias, mientras que Desigual representa una inconsistencia. El método no depende de coordenadas, escalas, unidades de medida, proyección cartográfica, formato de representación, primitivas geométricas, entre otras. El caso de estudio es la generalización de los ríos y las curvas de nivel.
Palabras clave: similitud semántica, generalización, objetos geográficos

1 Introduction

The generalization is used to produce geographic data at coarser levels of detail, while retaining essential characteristics of underlying geographic information [Weibel, 1996]. Therefore, generalization systems should

assure the semantic consistency of generalized data. Traditionally, the problem of consistency is a numerical task based on measures that represent constraints at topological and geometrical levels [Beard, 1991]. These measures are difficult to interpret and adapt to different contexts. Thus, the present work is focused on using well studied quantitative measurements, but passing them at the conceptual operating level to facilitate the detection and interpretation of inconsistencies, which are commonly presented in generalization. To do this, we use a conceptual representation of topographic domain (ontology) those concepts (classes) are defined by numerical intervals, which in turn are the results of quantitative measurements of the geographic objects. In the case of study, the used concepts belong to two levels deep ontology fragment (ordered hierarchy). Each concept however can have more descendents at other ontological levels. In other words, a number of subclasses can be generated to conceptualize at deeper detail the numeric intervals. This allows measuring the distance between qualitative values instead of quantitative ones, that is to say, the distance between two ontology classes.

In contrast to numerical approaches the semantic distance facilitates the interpretation of the measurements and produces better results to user's satisfaction, because the concepts and their similarity can be easily understood and interpreted [Schewering and Raubal, 2005]. In addition, the detection and interpretation of inconsistencies is based only on the conceptual representation, which does not depend on coordinates, scales, units of measure, cartographic projection, representation format, geometric primitives, and so on.

The rest of the thesis is organized as follows: Section 2 describes the conceptualization of topographic domain and ontology designed for the case of study. Section 3 presents the methodology. Section 4 exposes some results for the case of study. Section 5 sketches out our conclusions and future work.

2 Conceptualization of Topographic Domain

This work is based on a conceptualization of topographic domain, which describes the main properties and relations of geographic objects (elevation contour lines and rivers). To conceptualize the domain we are use the following documents: Environmental Data Coding Specification (EDCS) [ISO 18025, 2005], WordNet [Wordnet, 2007], and documents of the National Institute of Statistics, Geography and Informatics (INEGI) [INEGI, 1995][INEGI, 1996].

Relations between rivers and elevation contour lines

Relations between rivers and elevation contour lines depend on the flow direction. In consistent data, a river can cross just once an elevation contour line; more than one crossing represents an inconsistency. From the point of view of cartographic representation, in the best case, the river should *pass by* a maximum convexity of the contour line (Fig 1a). Sometimes the river *passes by* a convexity of the elevation contour line (Fig 1b, c and d) or *passes by* its concavity (Fig. 1f, g and h); these cases represent inconsistencies. In other cases, the river passes by straight part of the elevation contour line (Fig. 1e and j).

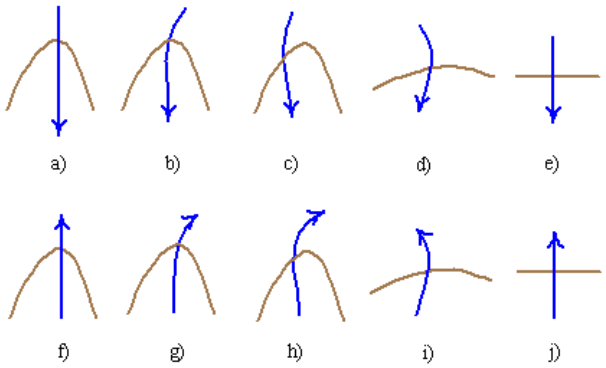

Fig. 1 Different cases of the relations between rivers and elevation contour lines

Thus, six different relations are to be considered in the following: *pass by maximum convexity*, *pass by almost maximum convexity*, *pass by convexity*, *pass by straight*, *pass by concavity*, *pass by almost maximum concavity*, and *pass by maximum concavity*.

Ontology

We define the terms of ontology to describe the relations between elevation contour lines and rivers (Fig 2). We include the relation between elevation contour lines and rivers (*pass by*) as a concept of ontology. Additionally this relation is specified in order to enrich the expressiveness of ontology by using other concepts, which describe the general term (*pass by*) at deeper level of details.

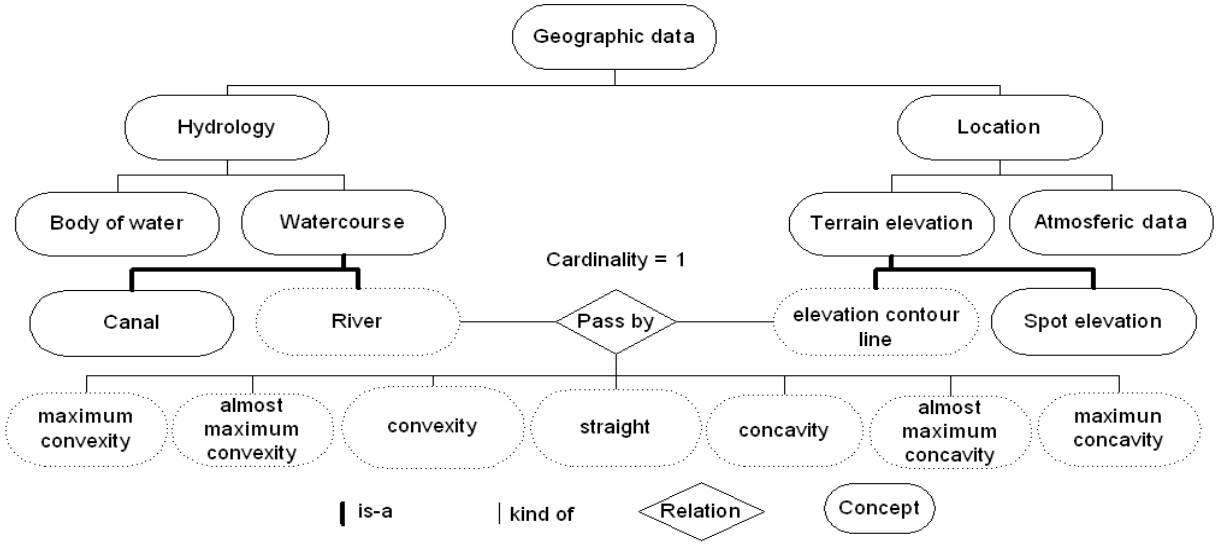

Fig. **2** Fragment of ontology to describe the relations between rivers and elevation contour lines

3 Methodology

The methodology consists of five stages: *Normalization*, *Processing*, *Analysis*, *Synthesis*, and *Verification* (Fig. 3). In normalization stage we verify the consistency of the geographic data prior to performing the following stages. In processing stage the data are automatically generalized. In analysis stage we automatically extract the relations between geographic objects. Synthesis stage is focused on generating a conceptual representation using the previous stage. In verification stage we evaluate the consistency of generalized data by using the conceptual representation and ontology.

Normalization

Normalization is used to verify the topological consistency of the source data. This stage is very important, because topological inconsistencies can affect the subsequent processing. In the case of the river networks the flow direction should be corrected, if any. Additionally, for each object an alphanumeric identifier is automatically assigned, e.g. River_1.

Analysis

Analysis stage is based on measures of geospatial data. A measurement is a computing procedure for evaluating characteristics (features, attributes, etc.) of geographical objects [Project Agent, 1999]. A measurement is a numerical value assigned to an observation that reflects the magnitude or amount of a characteristic. As result of this stage we obtain *quantitative descriptors* (*DC*).

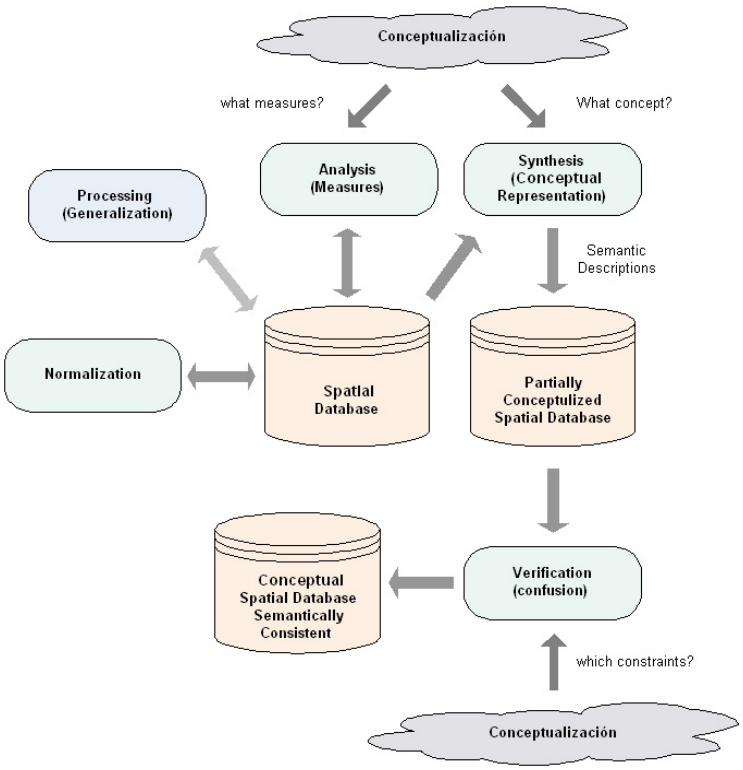

Fig. 3 Methodological framework

Quantitative descriptors of pass by relation between rivers and elevation contour lines

Herein we are focused on the measure of *pass by* relation between rivers and elevation contour lines. As result of this measure we obtain a value denominated *quantitative descriptor of pass by relation between rivers and elevations contour lines* (*DCP*). Additionally, *Number-of-Relation* (*NR*) property is calculated in this stage[1]. To better describe the relation, we follow the premise that "*topology matters, while metric refines*" [Egenhofer and Mark, 1995]. By using the *9-intersection model by* Egenhofer we can not identify all the cases of *pass by* relation (Section 2.1). Instead, we use a measure to extract the particularities of the relation, incorporating metrics for the topological relations as in [Nedas. et al., 2007]. Each measure is stored in the spatial database. DCP is computed as follows:

1. Identify the intersection of *river* and *elevation contour line* as v_p (Fig 4).
2. Define a circular area (A) with radio r and center at v_p. Two points of intersection between *elevation contour line* and A are identified as v_a and v_b and two points of intersection between the *river* and A are denominated as v_c and v_d.
3. Search for a vertex on the *elevation contour line* that form the greatest area with v_a and v_b . This vertex is denominated v_m. Compute $A_{P=} area(v_a, v_b, v_p)$[2] and $A_{M=} area(v_a, v_b, v_m)$.
4. Identify concavity or convexity. Compute $P_A = perimeter(v_a, v_b, v_p, v_c)$[3] y $P_D = perimeter(v_a, v_b, v_p, v_d)$

[1] Sequential number assigned to each relation between a river and elevation contour whenever they are intersected (following the flow of the river), starting at 1

[2] *area(a,b,c)* computes the triangular area composed of 3 points

[3] *perimeter (a,b,c,d)* computes the perimeter of figure composed of 4 points

If $P_A > P_D$, it is a, convexity then $DCP = A_P / A_M$ *(1)*
If $P_A < P_D$, it is a concavity, then $DCP = A_P / A_M$ *(-1)*
If $P_A = P_D$, it is a straight part, then *DCP =0*

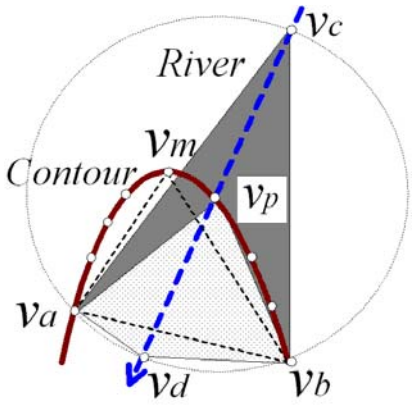

Fig. 4 Components of the river and elevation contour line to compute *DCP*

The interpretation of the values of DCP is (Fig. 5):

- *DCP* = 1 means that the river *passes by* the point of maximum convexity of elevation contour line.
- 1 > *DCP* > 0 means that the river *passes by* a convexity of the elevation contour line.
- *DCP* = 0 means that the river *passes by* a straight part of an elevation contour line (concavity and convexity do not exist).
- 0 > *DCP* > -1 means that the *river passes by* a concavity of the elevation contour line.
- *DCP* = -1 means that the river *passes by* the point of maximum concavity of an elevation contour line.

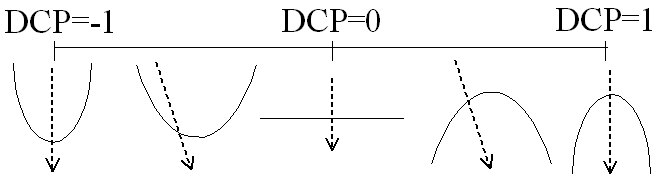

Fig. 5 Interpretation of *DCP*

Synthesis

Synthesis stage is used to generate a conceptual representation that explicitly describes the relations between geographic objects. Thus, a *Partially Conceptualized Spatial Database* (*PCDB*) is obtained. PCDB is composed of *tuples* denominated *semantic descriptions* (*SD*). SD has the form $\{O_i, \mathbf{R}, O_j\}$, where O_i, y O_j are the identifiers of geographic objects and **R** represents the relation between the objects O_i and O_j. This conceptual representation can be stored in any *relational database*. SD is generated by mapping the DCP into the conceptualization and expresses the semantics of the relation. We use different criteria to define the class of each DCP (Table 1) and a constraints in order to a DCP is belonged to certain class. For instance, suppose that after measuring the relation between the *River_1* and *Contour_8* we obtain that *DCP = 0.99*. Then, we generate a tuple of the form: *{River_1, PASS BY*

MAXIMUM CONVEXITY, Contour_8}. Additionally, *NR* property is included in the conceptual representation as an attribute in PCDB.

Table 1. Criteria to define the classes/relations between rivers and elevation contour lines

Class	Criteria
<PASS-BY-MAXIMUM-CONVEXITY>	[0.95 < *DCP* < 1]
<PASS BY ALMOST MAXIMUM CONVEXITY>	[0.85 < *DCP* < 0.94]
<PASS BY CONVEXITY>	[0.35 < *DCP* < 0.84]
<PASS BY STRAIGHT>	[-0.34 < *DCP* < 0.34]
<PASS BY CONCAVITY>	[-0.84 < *DCP* < -0.35]
<PASS BY ALMOST MAXIMUM CONCAVITY>	[-0.94 < *DCP* < -0.85]
<PASS BY MAXIMUM CONCAVIYTY>	[-1 < *DCP* < -0.95]

Processing

Processing stage consists of an automatic generalization system based on the generalization operators proposed by McMaster and Shea [McMaster and Shea, 1992] (See Fig. 6). The system is used to generalize the rivers and elevation contour lines. The user defines the parameters of generalization operators to modify the scale from 1:50,000 to 1:250,000. Here the INEGI specifications [INEGI, 1995][INEGI, 1996] are also used.

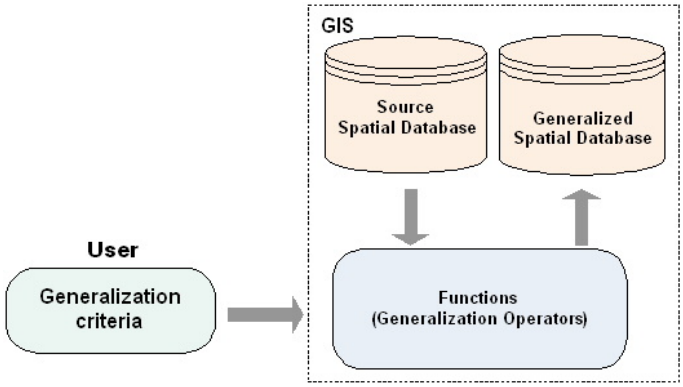

Fig. 6. Generalization System

Verification

This stage is used to verify the consistency of generalized data. We compare the *semantic descriptions of the source data* (SD_F) with the *semantic descriptions of the generalized data* (SD_G). Specifically this stage is based on the *semantic invariants*, which are relations that do not change after the generalization. The invariants depend on the consistency of the geospatial data and therefore on their semantic content (See Fig. 7).

In order to find the semantic invariants the hierarchical structure of ontology classes is used. The method to evaluate the consistency is based on *confusion* (the confusion conf(*r*, *s*) in using qualitative value *r* instead of the intended or correct value *s*). The concept of confusion allows defining the closeness to which an object fulfills a predicate as well as deriving other operations and properties among hierarchical values [Levachkine and Guzmán-Arenas, 2007].

Thus, we obtain a *semantic distance* between concepts. This distance allows us to define new concepts (*equal*, *equivalent*, *unequal*) that represent the difference between the conceptualizations and captures the change of the relations after generalization. The consistent relations (*equal, equivalent*) are stored in a *Spatial Database Semantically Consistent* (*SDSC*).

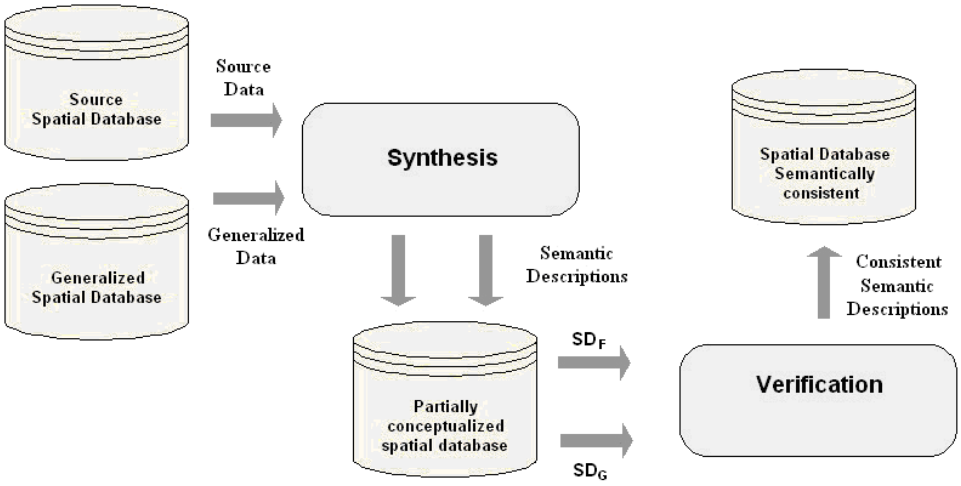

Fig. 7 Verification stage

Cases to verify the consistency

The consistency is evaluated by using confusion over a fragment of ontology. The fragment of ontology has the form of *ordered hierarchy* [12] (*pass by* relation in Fig.2). Confusion (conf (*r*, *s*)) is computed by the relative distance from the concept *r* to the concept *s* (the number of steps needed to jump from *r* to *s* in the ordering) divided by the (*cardinality* – 1) of the father (*pass by*) [Levachkine and Guzmán-Arenas, 2007].

We consider three different cases to evaluate the consistency that depend on the confusion value:

- *Equal*, conf(*r*, *s*) = 0, the concepts are equal. This means that the relation is consistent and the semantics is preserved after the generalization.
- *Unequal*, $0 < \text{conf}(r, s) \leq 1$, the concepts are unequal. This means that the relation is inconsistent and the semantics is not preserved after the generalization.
- *Equivalent*, some cases are considered to be consistent. Define a threshold (*u*). If $u < \text{conf}(r, s) < 1$, we consider that the relation is unequal. If $0 < \text{conf}(r, s) \leq u < 1$, we consider that the relation is equivalent. In a certain sense we can say that here the semantics is preserved.

By using this methodology, other errors can be identified. These errors are produced by the operators of line simplification. The identification of these errors is based on the premise "*an elevation contour line and river must cross once*".

4 Results

The system is implemented in Arc/Info 8.1.2, using Arc Macro Language (AML). The geospatial data are provided by INEGI. Some results for the case of study to verify the consistency are presented in this section. We put threshold $u = 1/6$.

Fig. 8 shows two semantic descriptions of the same relation prior (left) and after generalization (right). Applying confusion we have conf(*r*, *s*) = conf(*PASS BY MAXIMUM CONVEXITY, PASS BY MAXIMUM*

CONVEXITY) = 0; in this case the relation is *equal*. This means that the semantics is preserved in the generalized data and they are consistent.

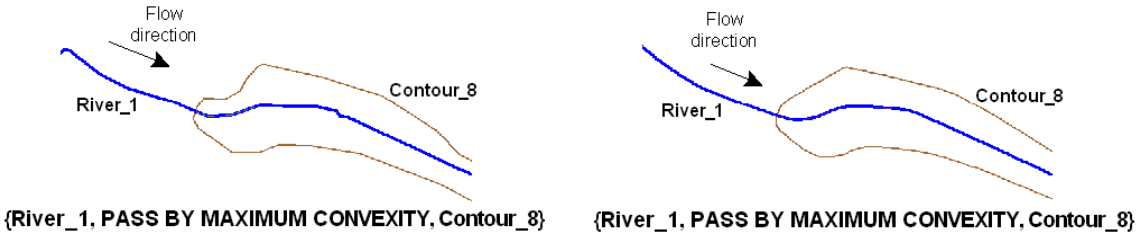

Fig. 8 Identifying a relation *Equal*

Fig. 9 depicts two semantic descriptions of the same relation prior (left) and after generalization (right). Applying confusion we have conf(*r, s*) = conf(*PASS BY MAXIMUM CONVEXITY, PASS BY ALMOST MAXIMUM CONVEXITY*) = 1/6 that is less than *u*. Thus, this relation is *equivalent* and therefore the relation in the generalized data is consistent.

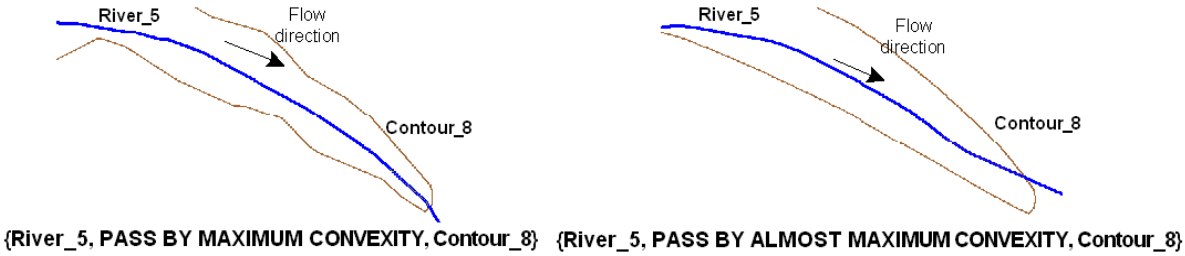

Fig. 9 Identifying a relation *Equivalent*

Fig. 8 shows two semantic descriptions of the same relation prior (left) and after generalization (right). Applying confusion we have conf(*r, s*) = conf(*PASS BY ALMOST MAXIMUM CONVEXITY, PASS BY STRAIGHT*) = 3/6 that is more than u; this means that this relation is unequal and represents an inconsistency.

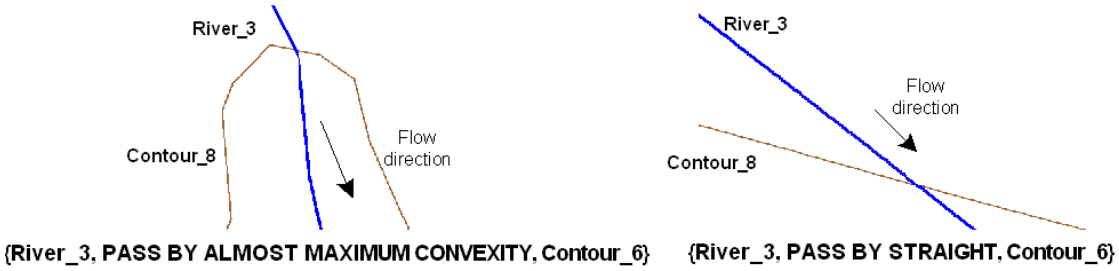

Fig. 10 Identifying a relation *Unequal*

Using the method proposed in Section 3.4 some relations are checked out in the generalized data (Fig. 9), finding that a river crosses the elevation contour line more than once and represents inconsistency. To identify these relations, we use the *NR* property: for consistent relations *NR* = 1, while for inconsistent – NR > 1.

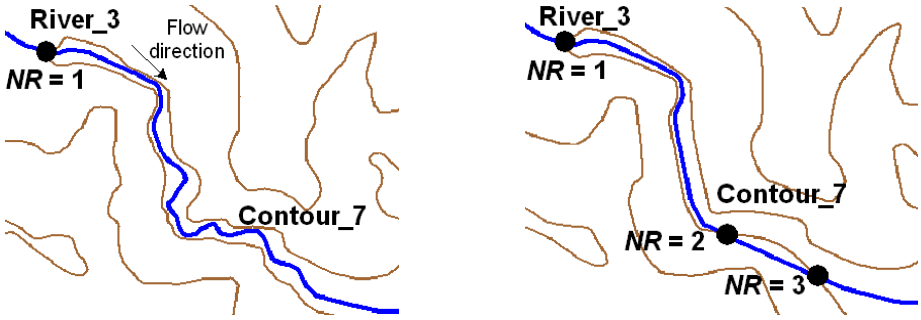

Fig. 11 Relations after the line simplification

5 Conclusions and Future Work

In this thesis we presented an approach to conceptually verify the consistency of generalized data. The method is based on a conceptual representation of spatial relations (ontology); it is generated by analyzing the quantitative metrics of topological relations between rivers and elevation contour lines. The concepts represent the interpretation of the geospatial data and the meaning of spatial relations. By using this approach, we attempt to catch the *semantic content* of the spatial data. To our knowledge, this is one of the first works based on conceptual representation to identify the inconsistencies of generalized data. Ontologies are very useful since they add a semantic component (the relations between different concepts) that usually does not consider in traditional GIS approaches. The conceptual representation does not depend on scale, cartographic projection, units of measure and format, and so on.

In the future work, more geographic information layers will be included to process the identification of the inconsistencies of generalized data as well as different kind of relations between objects from those layers will be considered.

References

1. **Weibel, R.:** A Typology of Constraints to Line Simplification. In: Kraak, M.J., Molenaar, M. (eds.) Advances in GIS Research II. 7th International Symposium on Spatial Data Handling, UK, pp. 533–546. Taylor & Francis, Abington (1996)
2. **Beard, M.K.:** Constraints on Rule Formation. In: Buttenfield, B.P., McMaster, R.B. (eds.) Map Generalization: Making Rules for Knowledge Representation, London, U.K, pp. 121–135. Longman (1991)
3. **Schwering, A., Raubal, M.:** Measuring Semantic Similarity between Geospatial Conceptual Regions. In: Rodríguez, M.A., Cruz, I., Levashkin, S., Egenhofer, M. (eds.) GeoS 2005. LNCS, vol. 3799, pp. 90–106. Springer, Heidelberg (2005)
4. **ISO 18025 (2005),** http://standards.iso.org/ittf/PubliclyAvailableStandards/
5. **Wordnet:** A Lexical Database for the English Language, http://wordnet.princeton.edu/
6. **INEGI:** Base de datos geográficos. Diccionario de datos topográficos (escala 1:50,000), Instituto Nacional de Estadística Geografía e Informática (INEGI) (1995)
7. **INEGI:** Base de datos geográficos. Diccionario de datos topográficos (escala 1:250,000), Instituto Nacional de Estadística Geografía e Informática (INEGI) (1996)
8. **Project Agent,** http://agent.ign.fr/
9. **Egenhofer, M., Mark, D.:** Naive Geography. In: Kuhn, W., Frank, A.U. (eds.) COSIT 1995. LNCS, vol. 988, pp. 1–15. Springer, Heidelberg (1995)
10. **Nedas, K., Egenhofer, M., Wilmsen, D.:** Metric Details of Topological Line-Line Relations. International Journal of Geographical Information Science 21(1), 21–48 (2007)

11. **McMaster, R.B., Shea, K.S.:** Generalization in Digital Cartography. In: Association of American Geographers (1992)
12. **Levachkine, S., Guzman-Arenas,** A.: Hierarchy as a New Data Type for Qualitative Variables. Expert Systems with Applications 32(3), 899–910 (2007)

Marco Moreno Ibarra, *actualmente es profesor investigador del Laboratorio de Procesamiento Inteligente de Información Geoespacial del Centro de Investigación en Computación del Instituto Politécnico Nacional. Ha sido autor y co-autor de más de 60 publicaciones en congresos y revistas. Es asesor de diversas asociaciones como el Instituto Federal Electoral, Compañía Minera la Parreña, Secretaría del Medio Ambiente, Recursos Naturales y Pesca, Secretaría de Medio Ambiente del D.F. Entre sus áreas de interés están el desarrollo de GIS (cartografía digital, generalización automática, métodos de representación del conocimiento espacial) e inteligencia artificial (ontologías y sistemas basados en conocimiento).*

Serguei Levachkine, *actualmente es profesor investigador del Centro de Investigación en Computación del Instituto Politécnico Nacional. Ha sido autor y co-autor de más de 150 publicaciones en congresos y revistas. Es egresado de la Universidad Estatal de Moscú (Lomonosov). Entre sus áreas de interés actual están el desarrollo de GIS (cartografía digital inteligente, métodos de representación del conocimiento espacial) e inteligencia artificial (procesamiento semántico de datos digitales, ontologías y sistemas basados en conocimiento).*