Proceeding of the 10th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences
Florianopolis-SC, Brazil, July 10-13, 2012.

Vectorial analysis modeling to determine spatial uncertainty among different data production scales due to error propagation

Cárdenas A[1]., Treviño E.J[2]., Aguirre O.A[2]., Jiménez J[2]., González M.A[2]., Antonio X[3]., and Sánchez G[1] .

[1] Abraham **Cárdenas** Tristán, Guillermo **Sánchez** Díaz. Universidad Autónoma de San Luis Potosí (UASLP), Facultad de Ingeniería, Av. Dr. Manuel Nava #8, Zona Universitaria. C.P. 78290, San Luis Potosí, S.L.P. México. abraham.cardenas@uaslp.mx, guillermo.sanchez@uaslp.mx

[2] Eduardo Javier **Treviño** Garza, Oscar Alberto **Aguirre** Calderón, Javier **Jiménez** Pérez, Marco Aurelio **González** Tagle. Universidad Autónoma de Nuevo León (UANL), Facultad de Ciencias Forestales. Carretera Nacional km. 145. AP 41. C.P. 67700. Linares, Nuevo León, México. ejtrevin@gmail.com, oscar.aguirrecl@uanl.edu.mx, javier.jimenezp@uanl.edu.mx, marco.tagle@gmail.com

[3] Xanat **Antonio** Némiga. Universidad Autónoma del Estado de México (UAEM), Facultad de Geografía. Cerro de Coatepec s/n, Ciudad Universitaria, Toluca, Edo, de México. C.P. 50100. xanynemiga@rocketmail.com

Abstract

Given different processes in recent years to analyze vectorial data production at different scales as well as quality analysis of it, geometric primitive geocodes were analyzed forming the geographic objects representation in order to carry out a geometric-topological recognition from different ways of representing reality vectorially. We used the nearest neighbor rule combined with a classification, determining an iterative search algorithm to analyze polylines that form curves or geographic objects. This algorithm allows modeling certain spatial uncertainty indicators from different types of topological-geometric errors that have propagated significant differences between various scales correspondence of vectorial information.

Keywords: Spatial uncertainty, data quality, maps production, vectorial data, error propagation.

1. Introduction

Mapping production traditionally has been a long process that involves acquisition and information validation, cartographic databases development and cartographic generation at different scales. Recently, (Kumi-Boateng *et al.,* 2010), raises awareness in the policies establishment to authenticate quality of spatial data production, "is not only useful for in-house data development, but data customers and users are able to determine the validity of data by checking the sources and procedures used to create the data".

In the cartography production history, there have been an events series and elements, directed by sciences and technologies; for study, knowledge and the territories representation. One of the main problems to abstract from reality features for the territory knowledge was that there were not standardized models for abstraction process. According to (Andrew U. F., 2008), a decision processes model, from the point of view data producers, must be "sufficiently simplified, based on bounded rationality leads to an operational method to assess the fitness for use of data".

Prior to this criterion, each information producer in its own judgment, was using methodology of his choice, usually without considering metrics control and required specifications for adequate information collection. Generally, large amount research considering mapping quality from vectorial data at various scales, takes into account assessing primitive's components for any cartographic representation. Of such analysis in most cases it has been determined that if there are abnormalities in evaluating data process at different cartographic scales, these have been propagated in cartographic edition, inheriting various errors; metric, semantic, geometric and topological. This situation relates largely to that in most cases editing processes have been manual (Griffith, 2008) *"Because so much geospatial data has been manually digitized over the years, researchers have studied error associated with this digitizing process for several decades"*. The errors propagation generated in such manual editing processes, has left in most cases, an uncertain control of spatial data that show objects and territory elements. However, current vector map editing process is believed that new technologies can generate better geospatial representations data. The NCGIA initiative on "Visualizing the Quality of Spatial Information" classified the sources of data uncertainty as source errors, process errors and use errors, (Beard, M.K. *et al.,* 1991). However, it must be noted that uncertainty arises in real data in many ways, since data may contain errors or may be only partially complete (Lindley, 2006). In order to propose a model to evaluate spatial uncertainty, we proceeded to analyze spatial data sets coherence included in the corresponding databases to curves elements. To show anomalies of geometric and topological correspondences, the analysis consisted in data integrating from same sources at different scales since same sites where objects should correspond to each other. Most of the representation conflicts of vectorial information have to do with the detail level in its geometry, and for such situation, several data sets were analyzed that implicitly in its constitution have abnormalities linked to the way they are edited. Since geometric evaluation of vectors forming the curves object, and these being represented by an interconnected polylines series, expressing the terrain shape elevation, these were integrated with various technologies in order to describe its geometric shape representation over a background image in the corresponding territory. It was determined that curves representation that describes features around water bodies should be an equable set of terrain polylines to represent in a certain elevation. A curve that describes topographic features on the territories, should never enter in the surface water bodies, given that these elements constitution is unpredictable to be measured by fact that water is not a fixed component to be measurable. Therefore, to be measurable in elevation, volume and surface strictly depends on other procedures, such that determinations of hydrological and hydraulic order. Then, the level curves evidence that cross over water bodies, help us to determine inconsistencies existence in the images referencing, its constitution pixelating and vectorial data integration, and those inconsistencies, which has introduced this research idea, in the elements analysis of polylines representing curves that have been produced by different technologic mechanisms through the years. Thus, is intended to measure inconsistencies represented on the basis of vector data mismatch-image, awarded to error propagation evidence. Showing different quality aspects, we concentrated on curves near to a water body, where they fall within that body. To evaluate these inconsistencies, we used a Landsat image for the ETM + panchromatic band with 15m spatial resolution, which was integrated with different scales of vectorial data representing curves.

2. Metodology

To support consistency in the issue of representation geometric correlation between same sites information, we supported with FME workbench technology, which allowed us to integrate different vector data layers with up to three different data sets at scales 1:20 000, 1:50 000 and 1:250 000, using different spatial technology operators. By verifying the same curves correspondences at different scales, the geometric representation has to be similar, differs greatly in the logic representation of territory. Similarly, if you put a background image, it can be seen that such inconsistencies are represented in the study area. In most cases, scales 1:250 000 and 1:1 000 000 have reportedly been subjected to cartographic generalization processes of manual type, which in its due time to lack of technology and perhaps to the little given importance of the quality issue adjudicated from these scales that should represent the territory. To determine differences in the curves integration that describes water body's scales and their vectorial relationship located in specific image pixels, turn to perform various tests such as the example (Figure 1) in different water bodies of the given area.

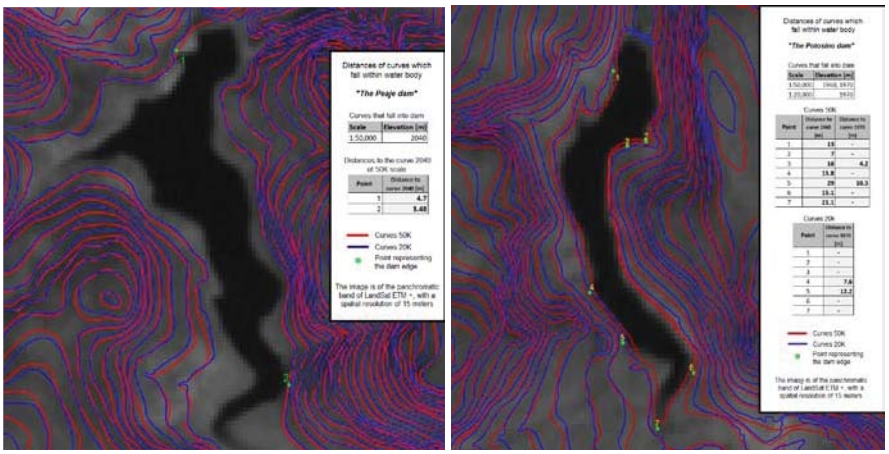

Figure 1: Integration of scales 50 and 20K on Landsat EMT

Generally, information integrations from vectorial samples at different scales analyzed show in high percentage the same problematic in the indicated correspondences. Sometimes the semantic representation evidences changes, due to data sets produced temporality, by facts that scale production 1:50 000 analysed, was edited between 1968 and 1988 being integrated at 1:20 000 scales which has been edited recently. The integration purpose of data sets mentioned above was performed to analyze the geometric relationships, topological, semantic and positional accuracy between different information scales on the same territory. For example, it should be common that contours representation at 1:50 000, that goes to each 10 meters, regarding contours at 1:20 000 which also are set to the same equidistance, these should match respectively in geometry terms and positional accuracy. However, since in the editing process the cartography at 1:50 000 emerges from a photography scale to 1:75 000 and the 1:20 000 cartography emerges from a photography scale to 1:40 000. Now, the different personnel involved which has been working on the editing process and equally in the application of different standards across

time. To validate the position correct curves and their relative pixels position in the image, we made satellite positioning, placing a point's series near water bodies, which helped us to check how accurate position curves is, in relation to the image pixels where these must correspond in position. What struck us was that while the both scales curves edited every 10 meters, these have wide correspondence differences relating to representation of the terrain they describe. In blue continuous curves are described at 50K with poor quality in editing geometry with various and frequent peaks that if we analyze at a zoom, it could be seen that edition makes no sense on respect to elevation of territory that it represents. In both situations the issue would be that the curves being represented in same elevation to a given territory, these lack logic and the problem has wanted to evaluate from a topological and geometric point of view, in order to describe a model that expresses the spatial uncertainty according to the problems described. Of recent papers that analyze images pixels and their relationship in the objects formation, and the use of algorithms and alignment patterns can cite to (Soe, W.M. *et al.,* 2001; Steele B. M., 2001; Zhang K., 2008; Xiang Zhang, 2011; Muhammad Aamir Cheema et al., 2010).

2.1 Definition of the problem.

Let U be an object universe. Each object $O_i \in U$ is described by an attribute set $R = \{x_1, x_2, ..., x_n\}$, and the objects are distributed in d-classes $\{S_1, S_2, ..., Sn\}$. Let $TM = \{O_1, O_2, ..., O_m\}$, $TM \subseteq U$ be the training set of the objects belonging to *U.* Specifically, a landsat image could be represent for a set *LI*, where $LI \subseteq U$, where each pixel *P* of *LI*, can be considered an object *O* of *U.* There are several areas of interest in *LI*, which account rivers, lakes, reservoirs, mountains, etc., which could represent by subsets of *LI.* Besides, there are obtained geometric relationships about a territory, as elevation curves in different scales, which could represents by subsets of a set *GR*, where $GR \subseteq U$. In addition, $LI \cup GR = U$. Let $S_i = \{O_{l_1}, O_{l_2}, ..., O_{l_u}\}, S_i \subseteq LI$ and $S_j = \{O_{h_1}, O_{h_2}, ..., O_{h_z}\}, S_j \subseteq GR$ the sets belonging to lakes and elevation curves of the sets *LI* and *GR,* respectively. Theoretical, these sets must comply $S_i \cap S_j = \varnothing$. However, when physical measurements were made in the territory, it was found that above property does not hold in all cases (see figures 1 and 2). Besides, is not sufficient to verify compliance with the above mentioned property, since it is desirable to know the degree of uncertainty propagated. This work proposes a way to calculate this value of the degree of uncertainty, using a classification algorithm, particularly the k-nearest neighbor (K-nn) classifier, applied over elevation curves, lakes and land surfaces.
Therefore, the problem in this work is about supervised classification, from a training set *TM,* the algorithm allows assigning to a set of objects $O_t \in U,\ O_t \notin TM$ a specific class $S_i, i = 1, ..., d.$ The following (figure 2) shows the model that helps determine the spatial uncertainty degree, according to analysis made.

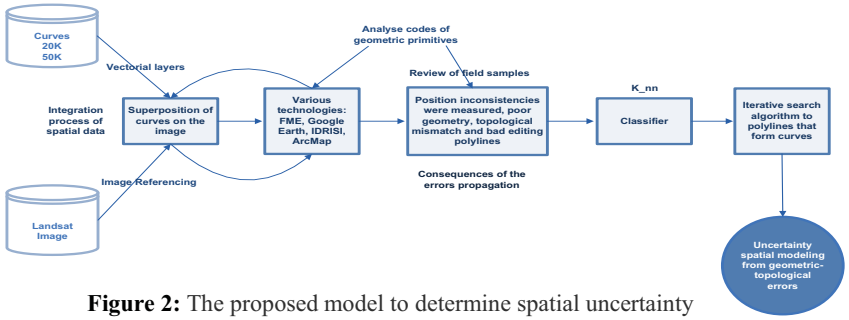

Figure 2: The proposed model to determine spatial uncertainty

3. Preliminary results

The proposed model includes the application of a classification algorithm, in order to obtain the uncertainty degree of elevation curves over lakes, analyzing the curves of geographic objects. For the particular case of satellite images shown in figures 1 and 2, it was selected the k-nn classification algorithm. In a general way, k-nn works as follow: First, the object to be classified is compared among all objects of TM, to obtain the distance between them. Second, these distances are order ascending. Finally, the first k-neighbors in the established order are only considerate for the classification. For the proposed model, the sets: $LI \subseteq TM$, $J \subseteq TM$, $LI \cap J = \varnothing$, and $GR \subseteq TS$ are considered. Where LI and J are the sets of pixels belonging to lakes and territory in the processed image, GR is the set of elevation curves belonging to the processed image and *TS* is the control or test set. Each point of a elevation curve could be represented by one pixel in the processed image, using their relative position. Then, each point of everyone elevation curves of *GR*, is processed by k-nn, obtained the k-nearest neighbors. Theoretical, all point of an elevation curve should been classified in the set *J*. However, in the practice this does not occur, because some points of elevation curves are classified in the set *LI*, which belonging to lakes in the processed image. The degree of uncertainty (*ud*) of elevation curve could be obtained using the following statement:

$$ud = 1 - a * \left(\frac{b}{n}\right)$$

where n is the number of pixels of the elevation curve; *b* is the number of pixels of the elevation curve classified erroneously into the set *LI*; and *a* is a constant of adjustment between different scales of elevation curves. Then, if the value of $ud = 1$, the elevation curve is correctly mapped. Else, if the value of *ud* is near to 1, this means that the mistake made is small. Otherwise, the mistake make is great. According to the analysis in the water body's case which have been described in the figure 1 and having applied the formula for determining spatial uncertainty, the results obtained are in (Table 1).

Table 1: Application of formula for measurement of uncertainty

Curve	Points number incidence of curve in water body (b)	Pixels Number of curve with water body incidence (n)	Scale (a)	ud
2040	2	390	50k	0.9948
1970	2	251	20k	0.9920
1970	2	251	50k	0.9920
1960	6	251	50k	0.9760

4. Conclusions

It was proposed the preliminary phase algorithm from theoretical approach; however, this algorithm has been tested in various methodological process assessments and this being specific to reviews of polylines that constitute curves and the relationship of these around water bodies. The purpose was to determine the spatial uncertainty degree in the example described, demonstrating that errors propagation in this case, can be measured in some way with certain metrics, which are found in field verifying the correct position (using positioning system) where curves should cross on terrain and its correspondence in the referenced image. The next step is to automate the proposed model, implementing the algorithm through a system, which automates measurement requirements of the spatial uncertainty.

Acknowledgments

The authors would like to thank Consejo Nacional de Ciencia y Tecnología CONACyT, which has made possible this research, we would also like to thank Instituto Nacional de Estadística y Geografía e Informatica INEGI, for their cooperation and collaboration with this project idea.

References

Kumi-Boateng B., I. Y. (2010). "Assessing the Quality of Spatial Data." European Journal of Scientific Research 43(4): 507-515.

Andrew, U. F. (2008). Data Quality - What can an Ontological Analysis Contribute? The 8th International Symposium on Spatial Accuarcy Assessment in Natural Resources and Environmental Sciences, v1: Spatial Uncertainty, Vienna, Austria, Department of Geoinformation Technical University Vienna Gusshausstrasse 27-29/E127-1 A-1040.

Griffith, D. A. (2008). Spatial Autocorrelation and Random Effects in Digitizing Error. Proceedings of the 8th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Shanghai, P. R. China, June 25-27, 2008.

Beard Kate, M., Barbara P. Buttenfield, and Sarah B. Clapham (1991). NCGIA research initiative 7: Visualization of spatial data quality. Technical Paper 91-26, Castine, Maine, October 1991.

Lindley, D. V. (2006). Understanding Uncertainty. New Jersey, 250.

Soe W. Mynt, P. G., Anthony Brazel, Susanne Grossman-Clarke, Qihao Weng (2001). "Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery." Remote Sensing of Environment RSE-07831;(ELSEVIER): 17.

Steele, B. M. (2001). "Combining Multiple Classifiers: An Application Using Spatial and Remotely Sensed Information for Land Cover Type Mapping." Remote Sensing of Environment 74(3): 545-556.

Zhang, K. Z. a. S. (2008). Estimation of Linear Vectorial Semiparametric Models by Least Squares. Proceedings of the 8th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Shanghai, P. R. China, June 25-27, 2008, pp. 16-21.

Xiang Zhang, T. A., Jantien Stoter, Menno-Jan Kraak, Martien Molenaar (2011). "Building pattern recognition in topographic data: examples on collinear and curvilinear alignments." Geoinformática Springer(DOI 10.1007/s10707-011-0146-3): 33.

Muhammad Aamir Cheema, X. L., Wei Wang, Wenjie Zhang and Jian Pei (2010). "Probabilistic Reverse Nearest Neighbor Queries on Uncertain Data." transactions on knowledge and data engineering: 14.