The need of a framework to compare Geometric Conflation Algorithms

Carlos López-Vázquez and Carlos H. González(*)
Technical University of Madrid, Spain. (*) ch.gonzalez@alumnos.upm.es

ABSTRACT:

There exist a number of different algorithms devoted to harmonize the geometry of two vector datasets. We applied some of them to an urban area in Spain, using ortorectified images of higher accuracy as a ground reference. The metric of success was the US National Standard for Spatial Data Accuracy figures, based upon RMSE statistics. To simulate a realistic situation where control points might or might not be identified over the available vector dataset, we choose at random only 20 well identified points, and afterwards apply all methods. With just a few trials, the results do not show that the method A is clearly superior to method B. We discuss the methodological implications of using just one data case to discard method B in favor of method A. We propose to create a statistically sound framework to test existing and new methods by Monte Carlo simulation in order to build confidence while choosing one method as the best.

INTRODUCTION:

When two different datasets of the same region, independently developed, are compared almost always geometric differences arise. Saalfeld (1993) denoted as Conflation the process to transform one dataset A to coexist with other dataset B as a single piece. In general there might appear problems at the geometric, semantic and even topological level; this work deals with the geometric one. Following Saalfeld, the reconciliation procedure requires a) to identify homologue objects in both cartographies, and record its coordinates b) to find and apply an adequate numerical transformation to objects in A trying to fit the corresponding ones in B. Usually B is a cartography regarded as of "higher accuracy" than A, but it is not mandatory.

As part of a larger research project we should choose the best method coming from a list of suitable numerical transformations, to be applied either on-line or off-line to datasets of varying quality and lineage. This paper deals with the procedure devised to create such ordered list. The ordering criterion is related to the metric of success. For the time being, the discrepancies between A and B are measured following the National Standard for Spatial Data Accuracy (hereinafter NSSDA, 1998), which records the accuracy as an RMSE of the discrepancies. A lower value implies a better agreement.

Given a set of homologue objects (points in our case) located at irregularly spaced placed we estimate the displacement field given those isolated values. For this paper we have considered a narrow list of methods. They are: a) Ordinary kriging (Samper and Carrera, 1987) b) Inverse Distance Weighting c) GRIDDATA (Matlab® 2006) and d) GRIDFIT (D'Errico, 2006). The calculations were performed in Matlab®

DATA AND METHODS:

Data refer to the City of Gandia, Valencia, located in southeastern Spain . Covering an approximate area of 56.77 km2

Figue 1: Gandia in Spain

The approximate coordinates of the area are:

X=743043.27 m Y=4325798.07 m
X=748312.00 m Y=4317314.00 m
X=743635.16 m Y=4314269.63 m
X=738187.52 m Y=4322678.45 m

Orthophotos belong to the NATIONAL PLAN OF ORTHOPHOTOS AEREA - PNOA (*Plan Nacional de Ortofoto Aerea*) PNOA Gandia 792-1. Orthophotos correspond to the projection ED-50 (European Datum) y Huso 30 dated 1996, posterior and separate from the Numeric Base Cartographic –BCN (*Base Cartografica Numerica*).

Resolution is 0.5 m per pixel**.**

Figure 2.: PNOA 792-1

Cartography BCN , which is the document to conflate, corresponds to the BCN25 the area Gandia., scale 1:25 000

The reference system is UTM area 30 Datum ED50

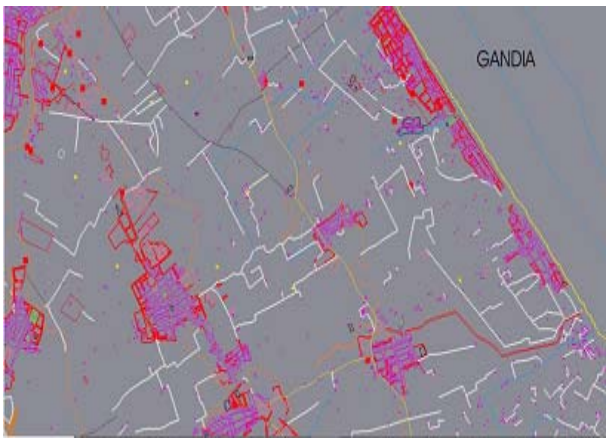

Figure 3.: BCN h10_796_1-1

We selected the set of 20 control points with a mixed criterion. 8 out of 20 were deterministically selected because they define the convex hull of the data. The remaining 12 points were selected at random, simulating a realistic situation where data can be available or not at some places. All the remaining 80 points were used to measure the resulting accuracy.

The Ordinary Kriging method requires a variogram, which has been estimated according with the method of Maximum Likelihood Cross Validation (Samper and Carrera, 1990). Only 20 control points were available for the calculations. Inverse Distance Weighting finds an estimate as a linear combination of data at the control points with weights which are the inverse of the distance squared; it has no parameters. GRIDFIT (D'Errico, 2006) is an approximation method closely related to Thin Plate Splines; we selected default parameters except for 'smoothness' set to 2. GRIDDATA is one standard interpolation routine in Matlab®, and we used it with 'v4' parameter.

RESULTS:

Just to illustrate our point, we performed five runs of the calculations. They differ in the set of control points selected. From the Table 1, it is apparent that some methods perform better than others in some cases and not in others, while some methods seems to be always losers. Select the "best" method is clearly a tricky business with just five trials, and we certainly should work harder to improve confidence.

Run#1	Run#2	Run#3	Run#4	Run#5	Methods Name
6.369	18.7753	11.6646	25.4906	5.4519	Ordinary Kriging
7.8936	26.9874	16.2904	23.1669	4.7862	GRIDFIT with smoothness =2
5.3122	20.6557	11.7092	21.5482	8.3135	IDW
18.1788	22.9777	22.7673	37.9036	9.7173	GRIDDATA with 'v4'

Table 1: Values of the NSSDA accuracy statistics for different methods. Each column has a different set of control points, taken partially at random from the ones available. In yellow the best ones in each realization.

CONCLUSIONS:

We have applied some readily available methods to the problem of conflating one vector map onto an ortophoto, taken as a reference dataset. The NSSDA statistic has been chosen as the metric of success. To gain insight in the importance of control point selection, we select partially at random only 20 control points out of the 100 available, trying to build a realistic scenario under our particular project conditions. After a few trials we realize that no clear winner arises. We feel that further calculations are required to favor one method over of the others. Assuming that most of the variability arises from the choice of control points, we might try to perform a Monte Carlo experiment and extract confidence intervals from the results. However, such approach ignores other sources of variability.

If the creation of a vector map is the result of a well defined deterministic process, but somewhat affected by random choices and options by the operator, it is fit to devise yet another formulation. We could take original map A and create at random perturbed versions of it, mimicking the variations observed between comparable maps of the same region. The perturbation model should take into consideration the large spatial correlation observed in practice (because spatial relations between close objects are usually well represented in different maps), while at the same time showing some degree of randomness. Conditional simulation based upon Ordinary Kriging and/or Cokriging (Samper and Carrera, 1987) might be adequate to model at least partly the process. However, some cartographic restrictions might not be satisfied by such simple procedures (Casado, 2006) which will require an extra effort.

The hierarchy of methods is obviously dependent upon the metric of success. We have considered the NSSDA accuracy statistic because it is a single number for the dataset; however, real datasets with similar NSSDA statistic might have both zones with good and bad agreement, while others have a more uniform pattern over the whole area. Other metrics, carefully related to the application area, should be considered in other cases.

BIBLIOGRAPHY

Beeri, B; Kanza, Y; Safra, E & Sagiv, Y, 2004 Object Fusion in Geographic Information Systems, In *Proceedings of the 30th Very Large Data Bases Conference*, Toronto, Canada, 2004, vol. 30, pp. 816-827, ISBN:0-12-088469-0

Casado, M L 2006, 'Some Basic Mathematical Constraints for the Geometric Conflation Problem', In *Proceedings of the 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, M. Caetano and M. Painho (eds), pp. 264-274

D'Errico, J. 2006 Understanding GRIDFIT, 2006. Available for download at http://www.mathworks.com/matlabcentral/fileexchange/8998 (last accessed 20090414)

Saalfeld A, 1993 Conflation: Automated Map Compilation, *Ph.D. Thesis, University of Maryland, (College Park, Maryland)*, pp. 1-133,

Samper, J. and J. Carrera, 1990 Geostadística: aplicaciones a la Hidrología Subterránea. Ed. CIMNE. Barcelona. Spain, 484 pp.

NSSDA 1998, Geospatial Positioning Accuracy Standards; Part 3: National Standard for Spatial Data Accuracy, *Federal Geographic Data Committee*, FGDC-STD-007.3, Washington, D.C. 28 pp. (http://www.fgdc.gov/standards/projects/FGDC-standards-projects/accuracy/part3/chapter3 accesed 20090414)

MATLAB 2009 Data gridding (http://www.mathworks.com/access/helpdesk/help/techdoc/ref/griddata.html) accessed 20090414