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

Changing the TIGER's stripes: detecting road network change under positional uncertainty

Ashton Shortridge[1] *and Miaoying Shi*[2]

[1] Department of Geography, Michigan State University, USA 48824
ashton@msu.edu
[2] Department of Forestry, Michigan State University, USA 48824
miaoyingshi23@gmail.com

Abstract

Many sets of linear features (e.g., streets in a city) can change over time, and the identification of these changes is an important geoprocessing challenge. This challenge is exacerbated by the often low accuracy of historic network datasets: positional discrepancies between features at different times may reflect actual change, or they may simply be the product of error. While there is a substantial body of work on modeling positional uncertainty in linear features, these contributions do not appear to have been widely integrated with feature change detection algorithms. In the present work, we address this research gap by developing several uncertainty modeling approaches for use in detecting significant network change. The following paragraphs review relevant approaches to characterizing positional uncertainty and then present a typology of network data mismatch. We then describe a geostatistical approach to model this uncertainty, and contrast it with conventional epsilon band technique to identify significant change. The approach is demonstrated in a case study with US Census TIGER line data

Keywords: positional error, vector uncertainty, kriging, epsilon bands

1. Introduction

An enduring GIScience research thread has considered the problem of modeling positional uncertainty in linear phenomena such as road networks. Perkal (1966) developed the epsilon band concept, which uses a buffer with radius equal to epsilon around a linear feature, to better understand length-scale relationships; Chrisman (1982) adapted this to characterize positional uncertainty in lines. Goodchild & Hunter (1997) modified it to assess differences in linear features using a range of distances to identify buffer sizes corresponding to 90 or 95% of the overlap between features. Tveite (1999) developed the buffer approach further by developing a set of overlay metrics to assess both average displacement and mismatch/missing data. At about the same time alternatives to buffer-based metrics were being explored. Kiiveri (1997) characterized this uncertainty using smoothly varying spatial displacements to propagate the impact of this uncertainty to area estimates. Church et al. (1999) and Zhang & Kirby (2000) employed geostatistical approaches to more formally characterize the spatial structure often evident in positional error for features. More recent geostatistical work has focused on characterizing uncertainty

at locations between vertices (de Bruin, 2008). We identify several circumstances that may give rise to geometric discrepancies between two datasets characterizing a linear network at distinct times *t1* and *t2*:

1. Addition of features between *t1* and *t2*
2. Removal of features between *t1* and *t2*
3. Change in feature position between *t1* and *t2*
4. Positional error in feature location in *t1*
5. Positional error in feature location in *t2*
6. Semantic differences: datasets employ different feature definitions
7. Generalization differences: datasets contain positional information at different levels of scale or precision

For our purposes these circumstances can be conflated: items 1-3 are of particular interest for change detection. Items 4-5 need to be accounted for in order to identify 1-3. Item 6 will not be further explored in this paper, but will result in the effective 'removal' of features from one of the datasets. Item 7 can be thought of as a source of positional error in the coarser resolution dataset.

2. Modeling Positional Uncertainty

Your The approaches covered here require some source of reference data, or "ground-truth", to establish the parameters of the models. One source of these may be a subset of street intersections within the area of interest. Intersections are attractive choices, as they are relatively easy to identify and measure, either on the ground or from map data or imagery. A potential limitation is that the nature of positional error at street intersections may be different from error at locations on roads distant from these intersections. In any event, positional error in one or more datasets of spatial networks can be measured at these common intersections; the challenge then is to develop a model of error at unmeasured positions in the network. We employ two approaches to model positional error: the first is an epsilon band approach, while the second uses a geostatistical model. These approaches use distinct models of positional uncertainty, and therefore have quite different parameterization requirements and evaluate somewhat different aspects of positional uncertainty.

In the epsilon band approach, distances between intersections in reference and primary data are calculated. This distribution of distances is sorted and any desired percentile threshold can be calculated, for example, 0.95 or 0.80. This distance threshold is then used to construct buffers around the network features.

The geostatistical approach is more complex and uses sequential Gaussian simulation and roughly follows the methods adopted by Church et al. (1998). Positional error at the known locations is decomposed into x and y components; each are evaluated separately. Variograms are modeled for error in x and error in y. Gaussian models are used, as positional error is assumed to vary nearly continuously. Network vector features are densified to enable perturbation of intermediate positions along road segments. Sequential Gaussian simulation is employed to generate multiple realizations of error in x and y at the road vertices. These error realizations are then added to the vertex coordinates to produce street network realizations.

Realizations can be used in a variety of ways to understand the magnitude of positional uncertainty and to identify significant differences between street networks at different times. In this short paper, we develop epsilon bands around them to compare with the first, non-geostatistical approach. Feature-specific buffer distances are calculated; this evidently novel, local approach enables us to adjust the width of the epsilon band so that each is tailored to uncertainty in the errors in different portions of the region.

3. Example Implementation

The city of Las Vegas has been among the most rapidly growing metropolitan areas in the United States for decades, with substantial extension and change in its street network. Furthermore, digital street data for Las Vegas are available from the US Census from 1990 to 2010; while national street data production in the early 1990's was a substantial technical achievement, the positional accuracy of this data is not high. The combination of rapid change and low accuracy makes this setting an attractive case study for this research. Data for southwestern Las Vegas centered on the suburb of Spring Valley was selected for evaluation. US Census TIGER line data for 1992 and 2010 were downloaded from the Census website.

Data were projected to UTM zone 11N, datum NAD83, using OGR library tools, and clipped to a 6,000 x 5,000 meter subset. The 1992 and 2000 datasets were then imported to R for further analysis. Processing steps are:

1. Find errors in x and y for the selected 1992 street intersections
2. Develop covariance models for those errors
3. Add vertices to the 1992 street intersection data (densification)
4. Use sequential Gaussian simulation to construct error realizations for the 1992 street vertices
5. Calculate street network realizations
6. Construct epsilon bands using global and local error characteristics

In this work we assume that any positional error is in the 1992 data, and that error in the north-south direction is independent from error in the east-west direction. Measured differences at a number of street intersections are used to develop separate covariance models of error in easting and error in northing throughout the dataset. Custom code written by the authors in R was used to find common street intersections in the 1992 and 2000 datasets using both geometric and attribute matching methods. A sequential filtering process using spatial proximity and fuzzy matching of street names was used to find similar intersections. A total of 535 matches were identified.

For the basic epsilon band approach, buffer distances were calculated to encompass desired proportions of the 2010 intersections within buffers around the 1992 road network; for illustrative purposes threshold proportions of 0.80 and 0.95 were calculated.

For the geostatistical approach, positional difference in x and y was identified and stored separately. Visual inspection of positional error indicated a complex and

spatially structured pattern. Variography of each error component was conducted. Gaussian models were fit in each direction with small nuggets, based on a suggestion by Zhang & Kirby (2000). As the 1992 data had comparatively few vertices along streets, additional vertices were added via linear interpolation. Conditional sequential Gaussian simulation was then conducted to generate error realizations in x and y at both original and added vertices in the 1992 data. These errors were then subtracted from the 1992 data vertex coordinates to generate multiple perturbed street realizations.

The final step was to develop both constant-width and segment-based epsilon bands around the 1992 streets data. For constant-width bands, buffer radius r was identified such that 95% of all perturbed vertices fell within the buffer around all streets in the 1992 dataset. A novel aspect of the current paper is the development of segment-specific epsilon bands. Separate bands were calculated for each road segment, so that epsilon bands varied in width from segment to segment. The 2010 road dataset was then intersected with these buffers to identify those street segments falling outside of the bands. Such segments are deemed to be significantly different from any roadway existing in the 1992 dataset.

4. Results and Discussion

Positional differences between the 1992 and 2000 datasets were substantial at the 535 common intersections. Error in x ranged from -146 to 149 meters; the error distribution was bell-shaped with a mean of -25.5 meters; the inner quartile range was -46-5.7 meters, while the standard deviation was 36 meters. Error in y was also bell-shaped, with a range of -121 to 127 meters and an inner quartile range of -26.5 to 14 meters. Mean error was -6.5 meters with a standard deviation of 32 meters.

The basic epsilon band approach was straightforward to implement. For the 0.80 band, a distance of 65.3 meters was sufficient, while 99.27 meters was necessary for the 0.95 band. Figure 1 shows the 0.80 band for a subset of the study area; those portions of 2010 roads falling outside the buffer would be deemed significantly different from the 1992 roads and indicative of network change using this method.

Variography revealed clear spatial structure, with nearby errors being similar. Gaussian models were fitted to each error dimension. This model was used to generate 50 error realizations in x and y at the vertex locations throughout the 1992 roads dataset, conditioned on the 535 known error values. These were added to the coordinates to produce road realizations. An example is portrayed in Figure 2.

Figure 1: 0.80 epsilon band (63.3 m) for part of the study area. 2010 roads are dashed

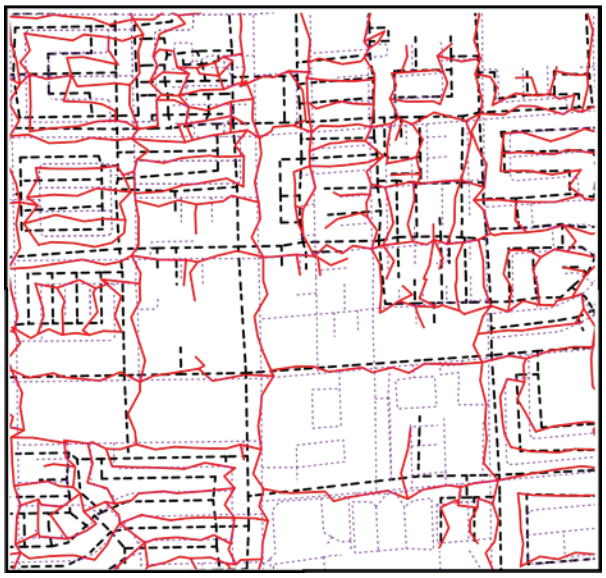

Figure 2: Streets realization for a subset of the study area.

For comparison purposes, feature-specific epsilon bands encompassing the smallest 0.80 errors across all realizations were calculated for each 1992 road segment (n=2,033). Buffer sizes ranged from 23 to 148 meters, with a mean of 59 meters. Figure 3 shows a subset of this data. Band widths vary substantially across the subset, and the results can be usefully contrasted with those in Figure 1.

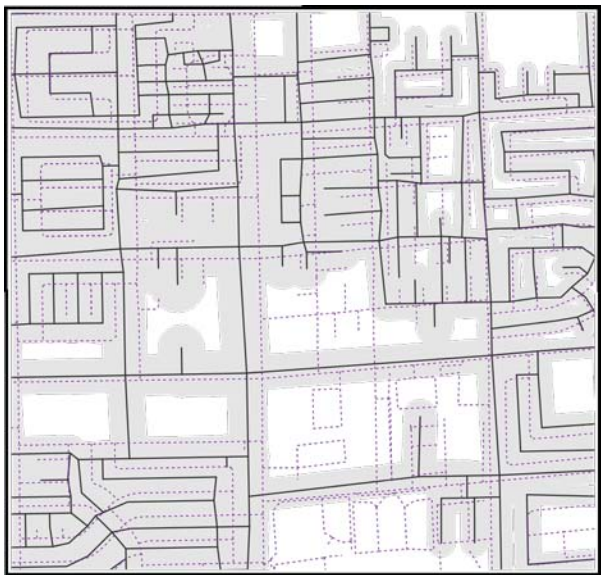

Figure 3: Feature-specific epsilon bands around streets.

This short paper presented two approaches for characterizing positional uncertainty in network data, with the goal of detecting real network change over time. The first developed epsilon bands using conventional techniques, while the second used realizations derived from geostatistical simulation to produce local, feature-specific bands. This case study served as an important proof of concept:

- ability to distinguish positional error from actual change
- an approach to derive feature-specific epsilon bands using geostatistics
- a comparison of conventional and feature-specific epsilon bands

References

Chrisman, N. R. (1982) Theory of cartographic error and its measurement in digital data bases. *Proceedings AUTO-CARTO 5*, Crystal City, VA, 22-28 August, 159-168.

Church, R., Curtin, K., Fohl, P., Funk, C., Goodchild, M., Kyriakidis, P., and Noronha, V. (1998) Positional distortion in geographic data sets. *Proceedings ACSM Annual Conference Technical Papers*, Baltimore, MD, March, 377-387.

de Bruin, S. (2008) Modelling positional uncertainty of line features by accounting for stochastic deviations from straight line segments. *Transactions in GIS* 12(2): 165-177.

Goodchild, M. F. and Hunter, G. J. (1997) A simple positional accuracy measure for linear features. *Int. J. Geographical Information Science* 11(3): 299-306.

Kiiveri, H. T. (1997) Assessing, representing, and transmitting positional uncertainty in maps. *Int. J. Geographical Information Science* 11(1): 33-52.

Perkal, J. (1965) *On the length of empirical curves*. University of Michigan Dept of Geography Discussion Paper #10, Ann Arbor, MI. 35 p.

Tveite, H. and Langaas, S. (1999) An accuracy assessment method for geographical line data sets based on buffering. *Int. J. Geographical Information Science* 13(1): 27-47.

Zhang, J. and Kirby, R. P. (2000) A geostatistical approach to modelling positional errors in vector data. *Transactions in GIS* 4(2): 145-159.