CHAPTER 15

Formulation and Test of a Model of Positional Distortion Fields

C. Funk, K. Curtin, M. Goodchild, D. Montello, and V. Noronha

INTRODUCTION

Hunter and Goodchild (1996) and Kiiveri (1997) have proposed a simple model of positional distortion in geographic data sets to account for the differences between true and measured locations (throughout this chapter *true* should be taken to mean *as determined from a source known to be of higher accuracy*). A point's true location (x,y) is distorted by the addition of a small vector displacement (e_x,e_y) to give the measured location of the point $(x+e_x,y+e_y)$. The displacements vary, but are assumed to be strongly spatially autocorrelated; collectively they form a vector field **e** whose components at any point are (e_x,e_y). In order to avoid folding or ripping, it is necessary that the components of **e** be constrained such that there are no cliffs in either surface, and such that their derivatives with respect to x and y, respectively, are always greater than -1.

In order to detect positional distortion it must be possible to match point locations in the data set in question to points in another data set of higher accuracy (see Church et al., 1998, for an extensive discussion of this issue). This is often difficult, particularly if there is ambiguity in the definition of mapped features, such as often exists in the natural resources area. For example, it would be very difficult to match points between two versions of a soil map, or even to agree that one was of higher accuracy than the other. Higher accuracy is often associated with a larger scale, but if the scale is different between the two data sets there would be no reason to expect pairs of features to match at all. On the other hand, location matching is much more reasonable in the case of well-defined cultural features, particularly when two sources exist at the same scale for the same area but with independent lineages. We consider such a case in this chapter.

In most industrial countries it is now possible to obtain so-called *street centerline* databases, containing representations of the approximate centerlines of the road and street network. Such databases are of immense value for emergency response, road maintenance, delivery and collection services, routing and navigation, and many other functions. Their importance is such that a significant commercial production and dissemination sector has emerged. In the United States, this industry initially added value to the public-sector TIGER database (produced for the 1980 census through a collaboration between the U.S. Bureau of the Census and the U.S. Geological Survey), and many products still have a TIGER legacy. More recently, however, companies have tended to build from scratch, beginning with aerial photography or vehicle-mounted GPS, in order to obtain higher positional accuracy. At this time, therefore, it is possible to obtain several databases covering a given area of the U.S., with identical features in varying positions. The research described in this chapter is based on an analysis of six such databases for an area of Goleta, California.

There are many practical motivations for wanting to determine the positional distortion between two databases. If one knew **e**, one could in principle correct one database to the other, removing any positional differences. Our research is driven by the problem of interoperation between street centerline databases in *in-*

telligent transportation systems (ITS). Consider the following scenario: a vehicle is being driven through an area, and information is being provided to the driver, including driving directions, from a system in the vehicle that includes a copy of Database A. The system is receiving updates on the state of the road system from a central server, maintained by the Department of Transportation, working off Database B. The position of an accident is broadcast from the server, as a coordinate location. The system in the vehicle attempts to match the location to its own street network, but the match fails because the positional difference between the two databases is sufficient to allocate the accident to the wrong street. Our analyses of the databases in the test area show that differences are frequently this large.

In this chapter we present an analysis of the differences between two of these databases, and test a model of **e**. We show how this model might be used to support error-free interoperation of the two databases using the concept of an *ITS Datum*, a proposed sparse network of control points.

Figure 15.1. Plot of a section of the two databases, superimposed on an interpolated field showing the magnitude of the distortion vector.

SAMPLING THE ERROR FIELD

Figure 15.1 shows two of the databases for a small part of the study area. The error field was sampled by determining its two components at street intersections, after matching such intersections by the names of the intersecting streets. As in all such databases, it was necessary to deal with differences in the naming and spelling of streets, and with problems caused by multiple intersections between the same pairs of streets. In some areas, particularly the lower left, the magnitudes of the positional differences clearly approach the distances between adjacent streets. The two sets of streets are shown superimposed on an interpolated field of distortion magnitude, or $(e_x^2 + e_y^2)^{1/2}$.

While this method of intersection matching works reasonably well in this case, it clearly will cause problems in rural areas where intersections are few and far between, or in cases where the names of streets are not available. In such cases it may be possible to find match points entirely from geometric information, but this seems inherently more risky and was not pursued in this study.

MODELS OF THE ERROR FIELD

Node matching provides an irregularly-spaced sample of evaluations of **e**. In order to make a complete adjustment of one database to fit the other, we need to have exhaustive rather than sample knowledge of **e**. Kiiveri (1997) argues that **e** must be everywhere smooth, and fits simple polynomial and trigonometric functions of the coordinates to both components of **e**. Church et al. (1998) argue that any of a wide range of spatial interpolation methods might also be used; they argue that they are to be preferred over Kiiveri's functions because we have no reason to suppose that **e** is a simple function of the coordinates. In this chapter we focus on the smoothness condition, and argue that it may be unwarranted for two reasons.

Consider first the processes by which such databases are made, and the likely implications of those processes for positional error. Three methods are considered here. First, a street centerline database can be constructed from existing pieces, such as digitized subdivision maps, digitized topographic maps, or existing databases. When such fragments are merged, one can expect differences of registration to show up as failure to match along the fragment edges. But edgematch problems will only appear in areas where streets or other features cross fragment edges; where features are absent such problems may be invisible, or may simply be ignored. In effect, the result is a mosaic of fragments, with distortions that persist within fragments and change sharply at fragment edges.

Second, consider a database constructed from aerial photography. Here the significant sources of positional inaccuracy relate to registration, and to the practice of

merging fragments of aerial photographs into a single mosaic. Again, we can anticipate that misregistration errors will persist over areas that depend on the same set of registration points, and change sharply at edge boundaries; and that errors at edge boundaries will be detected only when edges are crossed by features.

Finally, consider a database constructed by driving streets with in-vehicle GPS. Errors in GPS positions are known to be strongly autocorrelated through time and to change sharply when the set of observed satellites changes. While the result will clearly depend on the strategy used to traverse the street network, it seems that this method also is capable of producing sharp discontinuities in the error field. In summary, there are good reasons to believe that the positional error field of a street centerline database will show jumps and discontinuities under each of the standard methods of production.

Consider now the implications of discontinuities in the error field. A discontinuity is defined by the condition that $\mathbf{e}(\mathbf{s}+\delta\mathbf{s}) - \mathbf{e}(\mathbf{s})$ tends to some finite vector $\mathbf{a}$ as the magnitude of $\delta\mathbf{s}$ tends to 0; in other words, two points an infinitely small distance apart can have different positional error vectors. It was argued earlier that cliffs could not exist in a distortion field because any such cliff would introduce a sharp break in any linear feature that crossed it. Let $\mathbf{s}$ and $\mathbf{s}+\mathbf{ds}$ denote two points on a linear feature a small displacement $\mathbf{ds}$ apart. Then the linear feature will show a sudden horizontal offset or jump with a direction and magnitude defined by $\mathbf{a}$. If this occurred during the production of the database, the producer could adopt one of two strategies: (1) smooth out the jump by editing the feature (Figure 15.2), or (2) revise the production process. But if no feature crossed the cliff, the cliff would go undetected.

The traditional gridiron city of the nineteenth century was highly connected, with interruptions in the network only along rivers, ravines, railroads, or other features too difficult to cross. In the mid-twentieth century, urban planning practice changed to a new model, with high connectivity within subdivisions but low connectivity between them and along major arteries. Access points to subdivisions were limited, either for reasons of security (the gated community being an extreme example), or to limit access to major arteries. The practice is also convenient commercially, if stores can be positioned at or near subdivision access points.

In summary, we propose a model of the distortion field that is very different from that of Kiiveri (1997), or that implied by the interpolated surface of Figure 15.1. In our model, the plane is partitioned into a number of zones; within each zone, $\mathbf{e}$ is constant. We term this the *piecewise constant* model. The model generalizes readily to the case where $\mathbf{e}$ is a piecewise function of the coordinates, as for example where the distortion within one zone is described by an affine transformation, but we do not consider that generalization here. In the next section we implement the model using a novel clustering technique based on geostatistical principles.

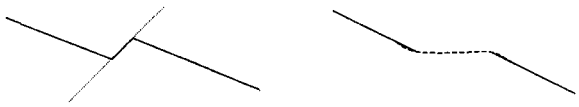

Figure 15.2. Effect of a cliff on a linear feature (left); editing with a smooth line (right).

CLUSTERING THE ERROR FIELD

"Regioning" seems to be a natural component of human thought—we speak of this place and that with confidence and regularity. Yet quantitative methods for automatically locating these regions given a set of geographic data are still insufficient to cope elegantly with many real-world problems. Geographic clustering problems are "hard," both mathematically and statistically. Posed as optimization problems, clustering problems are part of the larger class of location-allocation problems whose computational complexity rises exponentially with the number of observations. Posed within a statistical framework, spatial clustering problems immediately inherit the set of difficulties attached to most spatial datasets, chief of which is the lack of independence among nearby observations. Geostatistics is the branch of statistics which has most directly addressed the problem of spatial dependence between observations (e.g., Isaaks and Srivastava, 1989). It thus seems natural that spatial clustering should in some way benefit from the insights of geostatistics. That is the goal we pursue here.

We begin to build the bridge between geostatistics and clustering by making use of a common geostatistical tool, the empirical semivariogram. The semivariogram plots the variance of a pair of spatial observations as a function of their distance apart. The result is a summary statistic that describes the *average* variance between a given point and its neighbors, expressed solely as a function of the distances between that point and its neighbors. Since the variogram is a type of average (defined below), roughly half the data

points will have variances above and half below the variogram value. We take advantage of this relationship to define a *ratio of areal dependence* (RAD) statistic, which has an expected value of 0. When calculated for each point and plotted, this statistic is a useful exploratory spatial data analysis technique, helping the analyst to assess the number and locations of clusters, if indeed they exist at all. In addition, we show that it can be used as a type of "variance filter," allowing us to arrive at more meaningful clusters by removing points with high RAD values.

The Angular Semivariogram

The spatial dependence of values may be measured by means of the semivariogram (Cressie, 1993; Deutsch and Journel, 1992; Isaaks and Srivastava, 1989; Matheron, 1963). The semivariogram produces bucket estimates of the variance between pairs of points at varying distances or lags h. The semivariance of the set of points at a given lag, $\gamma(h)$, defined for a stationary random process (Z), is described as:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{(i,j)|h(i,j)=h} \left(Z_i - Z_j\right)^2 \qquad (1)$$

where $\gamma(h)$ is the semivariance for the set of points whose interpoint distances fall within the span of lag h, $(i,j)|h(i,j)=h$ denotes all the pairs of points whose distances place them in lag h, $N(h)$ is the total number of pairs found within this lag, and $(Z_i - Z_j)^2$ is the squared difference in the process between any two points separated by lag h.

The error field $\mathbf{e}$ is a vector field, with two components at any location, whereas the variogram is normally computed for a scalar field Z. In our cluster model, we hypothesize that $\mathbf{e}$ will be constant within each zone. For the purposes of this chapter, we assume that constancy of *direction* within each zone is an adequate test of the model, and ignore the possibility of variation in *magnitude*. But the difference between two angles cannot simply be described by subtracting one angle from the other. We now describe our proposed method for the computation of a semivariogram for angular data, based on principles of circular statistics (see Batschelet, 1981, for an excellent review of circular statistics).

Let $\mathbf{e}_i$ and $\mathbf{e}_j$ denote samples of the error field at two points i and j. We define $\mathbf{r}_{ij}$ as the mean of unit vectors in the directions of the error field at the two points; that is:

$$\mathbf{r}_{ij} = \left\{\left(e_{xi}/|\mathbf{e}_i| + e_{xj}/|\mathbf{e}_j|\right)/2, \left(e_{yi}/|\mathbf{e}_i| + e_{yj}/|\mathbf{e}_j|\right)/2\right\} \qquad (2)$$

Thus if $\mathbf{e}_i$ and $\mathbf{e}_j$ are the same they will both equal $\mathbf{r}_{ij}$, but $\mathbf{r}_{ij}$ will have no magnitude when the directions are 180 degrees apart.

We define S^2 as one minus the magnitude of $\mathbf{r}$ and substitute it in the computation of the semivariogram, as follows:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{(i,j)|h(i,j)=h} S_{ij}^2 \qquad (3)$$

Applying this equation to the distortion dataset previously discussed yields the empirical angular semivariogram shown in Figure 15.3.

This figure shows how the variation of the distortion angles increases with distance. There is a rather high inherent variability, with a minimum semivariance or nugget of about 0.15. The range is about 800 m; at this distance the semivariance reaches a sill of about 0.35.

The Ratio of Areal Dependence

We began this section by remarking that the semivariogram describes the average variance of observations at different lags. In this section we use this observation to construct a rather simple exploratory data analysis tool called the *ratio of areal dependence*, or RAD statistic. The RAD statistic has an expected value of 0.0. Positive values indicate that the local variance between a point and the set of neighbors found within the range is much higher than the expected variance for a set of points at those distances. Negative values depict the opposite. The RAD statistic is the ratio of the actual local variance at some point (V_i) divided by the expected (or modeled) variance at that point (M_i), or:

$$RAD_i = 1 - \frac{V_i}{M_i} \qquad (4)$$

Note that both V_i and M_i are measured between point i and a set of neighbors within a given range. Let N_r be the set of indices for all neighbors whose distance from the ith point is less than the range, and let n_r be the number of these points. Then the RAD statistic is a measure of actual versus expected local variance, where the actual local variance is defined as:

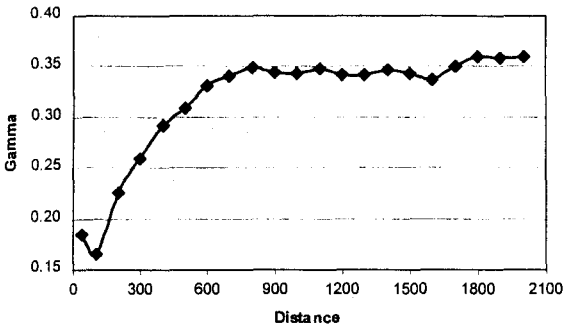

Figure 15.3. Semivariogram of angular distortion values.

$$V_i = \left(\sum_{j=1}^{N_r} S_{ij}^2 \right) / n_r \qquad (5)$$

and the expected or model variance is:

$$M_i = 2\left(\sum_{j=1}^{N_r} \gamma(d_{ij}) \right) / n_r \qquad (6)$$

A plot of the RAD values is shown in Figure 15.4. The filled circles have RAD values of greater than or equal to 0.0, indicating a higher than average spatial contiguity. The points denoted by triangles have relatively low (< 0.0) RAD values, indicating that they are relatively uncorrelated with their neighbors. This plot, and the RAD statistic in general, illustrate a fact sometimes forgotten by practitioners of spatial statistics: the variogram models average variance values, generated from ensembles of points at various lags. Individual points will tend to be more or less correlated with their neighbors than the variogram predicts. Those points which exhibit high spatial autocorrelation will have a *high* RAD, and vice versa. Thus regions with high RAD values designate natural regional patches which share similar values, and regions with high RAD values are dissimilar from their neighbors and are natural "barriers" demarcating one region from another.

Results

The angular variogram provides a statistically valid, somewhat interesting way to measure the spatial dependence of circular data. For our study region, the variogram exhibited a well-defined, highly localized

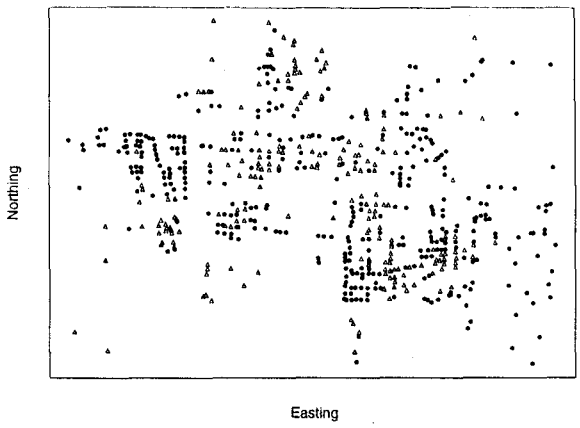

Figure 15.4. A plot of the RAD values associated with angular distortions.

spatial autocovariance structure, with points farther than 800 meters apart showing virtually no correlation. Even nearby intersections exhibited a marked degree of variance, as illustrated by the high nugget of the fitted variogram. Often, nearby streets exhibited dissimilar error vector orientation.

The RAD values plotted above depict the degree of spatial autocorrelation for the angular components of the distortion vectors. Visual inspection of Figure 15.4 indicates how successful any spatial clustering or "patchwise" modeling scheme is likely to be. Some areas, particularly the pockets of homogeneous street intersections in the upper left and lower right of the image, seem like natural "clusters," and are likely to have distortions created by a unique or similar process, which could be modeled with ease. Thus high RAD values suggest areas dominated by sources of "systematic" error. Other regions, such as the pockets of triangles in the top and bottom of the image, show what may be "random" error. Since the purpose of clustering is typically driven by a desire to model or elucidate systematic effects, we explored the effects of screening out the high RAD values in conjunction with an application of Ward's clustering algorithm. The idea is similar to that employed when removing outliers from a dataset, and the hope is that the resulting spatial partition depicts the systematic component of the variation.

As an example, Figure 15.5 shows a clustering of the distortion data. The following six fields were used to cluster the data: the x and y coordinates, the x and y components of the error vectors, and the magnitude and angle of the error vectors. Each of these variables

was standardized to have a mean of zero and variance of one. The standardized data set was then clustered via Ward's method, a hierarchical clustering technique that seeks to minimize the squared distances (or trace) between the cluster centers and their constituent points (Hartigan, 1975).

This clustering is far from ideal. As a preliminary step in a patch-wise error estimation process, the clusters depicted above would be difficult to interpret and use, since they do not clearly partition the space into mutually exclusive regions. The classes appear to overlap in geographic space, and reappear sporadically throughout the image. There are often situations in which such a pattern is legitimate and even desirable. However, for the problem at hand, we have assumed that errors are, to some degree, spatially determined—nearby points in the same clusters should share the same source of error. This is where the RAD statistic can be of assistance. We suggested earlier that this calculated variable differentiates between points exhibiting "systematic errors" (high RAD values) and "random errors" (low RAD values). Furthermore, in our discussion of the angular variogram, we noticed that the error vectors appeared to exhibit a large nugget effect, from which we inferred that about 42% of the angular variance was not explained by the values' spatial distribution. If these conclusions are valid, then it seems reasonable to exclude points with low RAD values, since presumably these values are dominated by random as opposed to systematic effects. To examine the results of such a clustering we excluded points with RAD statistics of less than zero and reran our clustering on the remaining points. The results appear in Figure 15.6.

This clustering is appealing at an intuitive level. Note that the clusters effectively partition the geographic region, and demarcate well the various patches. Compared to the clustering performed on all the points, the results are much more spatially cohesive, which is not surprising, since we rejected all the data points which differed dramatically from their neighbors. Whether this rejection of some points and retention of others improves our capacity to comprehend and capture the spatial processes under study is difficult to evaluate. Certainly the clusters defined above seem very reasonable. It is difficult to put "space" into clustering algorithms. In general, including the x and y locations, as we have done here, imparts a measure of spatial compactness to the resulting clusters. Their relative importance can be estimated by the sum of their variances compared to the sum of the variances of all the included

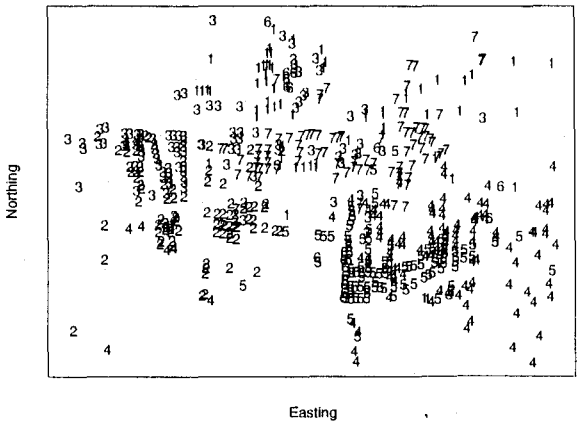

Figure 15.5. Initial clustering based on RAD values.

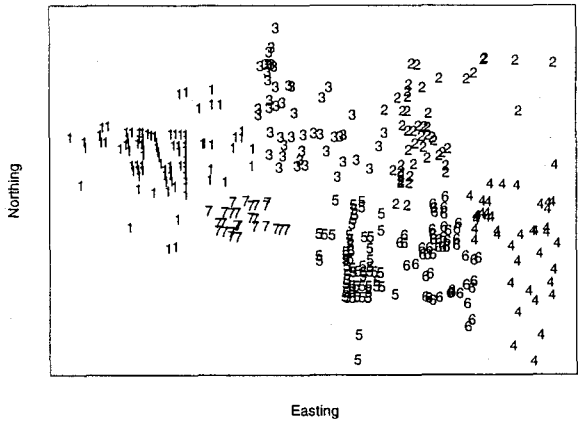

Figure 15.6. Clustering using only high RAD values.

variables. In this example, since the variables were standardized and six variables were used, x and y represent 33% of the total variation. Thus the relative weighting of space in these clusterings is about one-third. By increasing the variance of the x and y variables, we could eventually ensure spatially compact clusters. This compactness comes at a price. As the importance of the x and y variables increases, there will be increased variability within the cluster (in terms of the devaluated variables), as well as the tendency to lose coherent regions which are narrow bands. Clusters will become as round as possible. The RAD filter seems to obviate this undesirable tradeoff. In the clustering above, the clusters seem to be spatially cohesive while still allowing classes to take on complex shapes. One would hope this suggests an increased capability to follow the contour lines of the underlying "truth."

Conclusion

This is clearly a preliminary study. The angular semivariogram can be a useful geostatistical tool for examining the spatial distribution of vector valued data, but the angular component alone is insufficient to describe the errors found in spatial databases. Still, it was our hope that by defining rigorously the way one variable varied spatially over the study site, additional insight and leverage into the clustering problem could be obtained. In this sense, the RAD statistic seems a somewhat promising but still rather ad hoc tool. Simply throwing away outlying points may be a great way to bring out underlying patterns in data, even if they are not really there. Clearly, the RAD statistic must be used with caution as a "variance filter."

We have tried to explore how one might use geostatistics to bring space intelligently into the problem of clustering datapoints into homogeneous units that effectively partition geographic space. The latter two criteria are often at odds. If, however, the spatial data under consideration are in fact the result of a process that really imparts a pattern of spatial correlation, then excluding "variance outliers" may help to reveal the knowledge underneath the noise.

We have argued in this chapter that a piecewise approach to positional distortion may make more sense, in light of what is known about production processes, than an approach that assumes smooth variation. The results appear to bear this out, in a preliminary way, though more testing is clearly needed in a wider range of circumstances.

ACKNOWLEDGMENTS

The National Center for Geographic Information and Analysis is supported by the National Science Foundation. Funding for NCGIA's research on uncertainty in spatial databases comes from the National Imagery and Mapping Agency. Funding for this analysis of street centerline databases comes from Caltrans and Viggen Corporation. The assistance of companies providing data for these analyses is gratefully acknowledged.

REFERENCES

Batschelet, E. *Circular Statistics in Biology*, Academic Press, New York, 1981.

Church, R., K.M. Curtin, P. Fohl, C. Funk, M.F. Goodchild, P. Kyriakidis, and V. Noronha. Positional Distortion in Geographic Data Sets as a Barrier to Interoperation, in *Proceedings, ACSM Annual Convention, Baltimore*, 1998.

Cressie, N.A.C. *Statistics for Spatial Data*, Revised Edition. John Wiley & Sons, New York, 1993.

Deutsch, C.V. and A.G. Journel. *GSLIB: Geostatistical Software Library and User's Guide*, Oxford University Press, New York, 1992.

Hartigan, J.A. *Clustering Algorithms*, John Wiley & Sons, New York, 1975.

Hunter, G.J. and M.F. Goodchild. A New Model for Handling Vector Data Uncertainty in Geographic Information Systems. *J. Urban Reg. Inf. Syst. Assoc.*, 8(1), pp. 51–57, 1996.

Isaaks, E.H. and R.M Srivastava. *Applied Geostatistics*, Oxford University Press, New York, 1989.

Kiiveri, H.T. Assessing, Representing, and Transmitting Positional Uncertainty in Maps. *Int. J. Geogr. Inf. Sci.*, 11(1), pp. 33–52, 1997.

Matheron, G. Principles of Geostatistics. *Econ. Geol.*, 58, pp. 1246–1266, 1963.