To be presented at the 2nd International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Modeling, Fort Collins, Colorado, 21-23 May, 1996

Experimental Development of a Model of Vector Data Uncertainty

G.J. Hunter[1], B. Höck[2], M. Robey[1] and M.F. Goodchild[3]

[1] Department of Geomatics
The University of Melbourne
Parkville, Victoria, Australia

[2] Resource Monitoring Unit
Forest Research Institute
Rotorua, New Zealand

[3] National Center for Geographic Information & Analysis
University of California
Santa Barbara, CA 93106
USA

Experimental Development of a Model of Vector Data Uncertainty

G.J. Hunter[1], B. Höck[2], M. Robey[3] and M.F. Goodchild[4]

Abstract.—In a recent paper by Hunter and Goodchild (1996), a model of vector data uncertainty was proposed and its conceptual design and likely manner of implementation were discussed. The model allows for probabilistic distortion of point, line and polygon features through the creation of separate, random horizontal positional error fields in the *x* and *y* directions. These are overlaid with the vector data so as to apply coordinate shifts to all nodes and vertices to establish new versions of the original data set. By studying the variation in the family of outputs derived from the distorted input data, an assessment may be made of the uncertainty associated with the final product. This paper is a continuation of that initial work and discusses the experimental development undertaken thus far to implement the model in practice.

INTRODUCTION

In a recent paper by Hunter and Goodchild (1996), a model of vector data uncertainty was proposed and its conceptual design and likely manner of implementation and application were discussed. This paper is a continuation of that initial work and discusses the experimental development since undertaken to implement the model in practice. In the context of this research, we suggest there is a clear distinction to be made between 'error' and 'uncertainty', since the former implies that some degree of knowledge has been attained about differences between actual results or observations and the truth to which they pertain. On the other hand, 'uncertainty' conveys the fact that it is the lack of such knowledge which is responsible for hesitancy in accepting those same results or observations without caution, and the term 'error' is often used when it would be more appropriate to use 'uncertainty'.

The model that has been developed can be defined as a stochastic process capable of generating a population of distorted versions of the same reality (such as a map), with each version being a sample from the same population. The traditional Gaussian model (where the mean of the population estimates the true value and the standard deviation is a measure of variation in the observations) is one attempt at describing error, but it is global in nature and says nothing about local variations or the processes by which error may have accumulated.

The model applied here is viewed as an advance on that approach since it has the ability to show spatial variation in uncertainty, and the capability to include in its realizations, the probable effects of error propagation resulting from the

[1] *Assistant Professor, Department of Geomatics, The University of Melbourne, Victoria, Australia*

[2] *Scientist, Forest Research Institute, Rotorua, New Zealand*

[3] *PhD student, Department of Geomatics, The University of Melbourne, Victoria, Australia*

[4] *Director, National Center for Geographic Information & Analysis, University of California, Santa Barbara, CA*

combined to assess final output uncertainty. While the model requires an initial error estimate for creation of the two distortion grids, it is the resultant uncertainty arising from the use of perturbed data due to simulation which is under investigation (in conjunction with the spatial operations that are subsequently applied)—hence its label as an 'uncertainty' model.

DEVELOPMENT OF THE MODEL

Choosing the Error Grid Spacing

As discussed in Hunter and Goodchild (1996), the first step required to implement the model is to determine an appropriate error grid spacing. If it is too large, the nodes and vertices of small features in the source data will receive similar-sized shifts in *x* and *y* during perturbation and the process will not be random, giving rise to unwanted local autocorrelation between shifts (Figure 2). Conversely, if it is too small then processing time can increase dramatically as additional grid points are needlessly processed.

The authors suggest an appropriate spacing be selected from one of the following options: (1) the standard deviation of the horizontal positional error for the source data; (2) a distance equal to 0.5mm at source map scale where the data has been digitized; or (3) a threshold value smaller than the user would care to consider given the nature of the data to be processed. Thus far, the authors have found the standard deviation (as supplied by the data producer) to be a useful value to apply, since it tends to be smaller than (2) and therefore more conservative, and of similar magnitude to practical estimates of (3).

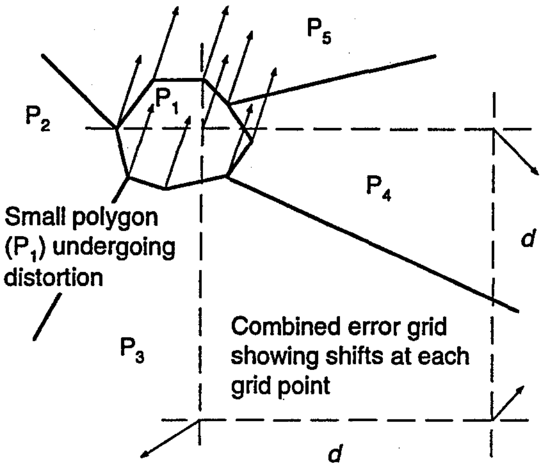
P_5
P_2
P_1
P_4
Small polygon (P_1) undergoing distortion
d
P_3
Combined error grid showing shifts at each grid point
d

Figure 2.—If the error grid spacing (d) is too large, then unwanted local autocorrelation between shifts may occur in small features.

Generating the Initial Error Grids

Using the development diagram on the following page as a guide for the remaining discussion (Figure 3), the next step is to generate the *x* and *y* error grids. The Arc/Info and Arc Grid software modules were used as the development platform and the commands referred to in the paper apply to these packages. To ensure the grids completely cover the extent of the source data, the dimensions of the grids were pre-determined by setting a window equal to the data set's dimensions (SETWINDOW command), and the cell size equivalent to the chosen grid spacing (SETCELL command). Two grids (named accordingly for the *x* and *y* directions) were created automatically using these parameters and then populated with randomly placed, normally distributed values having a mean of zero and a standard deviation defined by the data producer (NORMAL causing unwanted loops).

The solution requires filtering of any such 'offending' pairs of shifts which requires that the masking grid be extended for a distance of at least five standard deviations either side of the original data—given that with the normal distribution it is rare for a value to occur greater than this distance from the mean. This buffering of the masking grid will permit neighboring error grid shifts to be used in a subsequent filtering process.

To achieve this, the grid spacing is compared with the standard deviation of the data and the equivalent number of cells is calculated (remembering that the grid spacing will not always equal the standard deviation). For example, using a standard deviation of 20 m and a grid spacing of 30 m, the correct number of cells equals $|(100/30) + 1| = 4$ (with the extra cell and the modulus sign ensuring 'rounding up' of the answer). Using this value, a spread function is applied to the masking grid (EUCDISTANCE command). A new masking grid is created during the process and any (previously) NODATA cells affected by the operation are automatically returned to active status.

The masking grid is then overlaid with the initial error grids to provide reduced versions of the *x* and *y* grids that contain only shift values lying within the required regions. Finally, the error grids are expanded by the width of a cell on all sides (and given NODATA values) to support later processing (SELECTBOX command).

Filtering the Error Grids

To preserve topological integrity in the source data upon distortion, some means of adjustment must be introduced to control the magnitude of positional shifts between neighbouring points in the *x* and *y* error grids. As previously mentioned, if this does not occur there is a likelihood that neighbouring points in the original data set may be transposed in position—causing unwanted 'fractures'

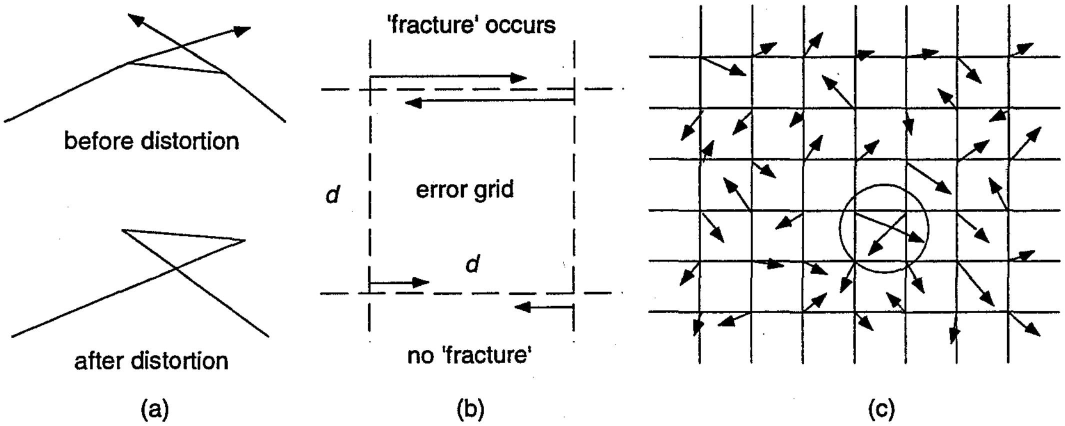

Figure 4.—In (a), uncontrolled shifts between neighbouring error grid points can cause unwanted 'fractures' or transposition of features. In (b), 'fractures' occur when the difference between neighbouring shifts is larger than their separation (*d*). In (c), a 'fracture' is circled, requiring filtering on the basis of neighbouring shifts.

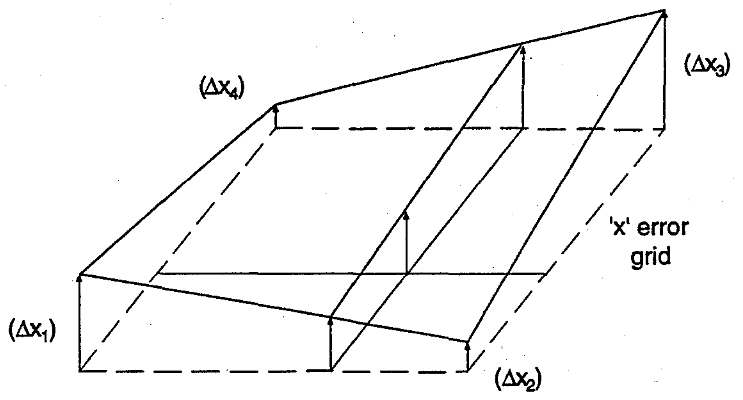

Figure 5.—The *x* and *y* shifts for a data point not coinciding with the error grid are calculated using bilinear interpolation based on the four surrounding grid values.

positions outside the original extent of the data set, this decision remains consistent with the original concept of distorting their values to form alternative, but equally probable versions of the data. At the time of writing, the experimental development of the model has not proceeded beyond this stage and the following discussion remains conceptual.

Calculation of Data Point Shifts

In the final step of the model's development, values in the error grids must be transferred to the data set being perturbed. Inevitably, it is expected that few if any nodes or vertices in the source data will coincide exactly with the error grid points, and a method is required for calculating *x* and *y* shifts based on the neighboring values in the grid. A simple bilinear interpolation procedure (Watson, 1992) is proposed in which the *x* and *y* shifts assigned to each point are calculated on the basis of the respective shifts of the four surrounding grid points (Figure 5).

An ASCII feature file containing the identifier and coordinates of each data point can be automatically derived (UNGENERATE command), and the four surrounding error shifts will be determined for each point and used to interpolate the correct values to be applied. The distorted coordinates are then written to a new file and when all records are processed, the file topology is rebuilt. Clearly, the need may arise for high performance computers to be employed for this task, particularly when many thousands of nodes and vertices may have to be perturbed, and current research into GIS and supercomputers (such as described in Armstong, 1994) is being investigated by the authors.

CONCLUSIONS

This paper has described the experimental development of an uncertainty model for vector data, which operates by taking an input data set of point, line or polygon features and then applying simulated positional error shifts in the *x* and *y* directions to calculate new coordinates for each node and vertex. In effect this produces a distorted, but equally probable, representation of the data set that can be used to create a family of alternative outputs, usually in map form.

Association (AURISA).

Barbara Höck has an MSc in operations research and has worked as a software developer and spatial analyst for a telecommunications company, a land and hydrographic survey company, and a forest research organisation. She currently manages the Resource Monitoring Unit at the New Zealand Forest Research Institute, Rotorua.

Mark Robey is a PhD student in the Department of Geomatics at The University of Melbourne. He has an MSc from the University of London in environmental policy and rural resources, and has worked as both a financial software programmer and as a consultant in agricultural economics and environmental policy for the University of London's Centre for European Agriculture Studies.

Michael Goodchild is Professor of Geography at the University of California, Santa Barbara, and Director of the National Center for Geographic Information and Analysis (NCGIA). He was editor of *Geographical Analysis* for several years and serves on the board of six other journals and book series. He is well known for his interest in quality and accuracy issues in GIS, being the editor of the well known book *Accuracy of Spatial Databases*.