376

MODELS FOR UNCERTAINTY IN AREA-CLASS MAPS

Michael F. Goodchild
National Center for Geographic Information and
Analysis and Department of Geography
University of California, Santa Barbara
CA 93106-4060, USA.
Phone: +1 805 893 8049 Fax: +1 805 893 3146,
Email: good@geog.ucsb.edu

Abstract

Area-class maps are the cartographic equivalent of nominal fields. Although much progress has been made on defining models of uncertainty for discrete objects and for interval/ratio fields, the nominal and ordinal cases remain problematic, for many reasons. The paper outlines the characteristics of nominal fields, and summarizes them as a series of six requirements for models of uncertainty. Existing models are then reviewed in the light of the requirements, and deficiencies are identified. A model is presented that satisfies all of the requirements, and its strengths and weaknesses are explored.

Introduction

Geographic information scientists recognize two alternative approaches to the representation of geography: discrete objects, and continuous fields (Worboys, 1995). In the former case objects are well-defined features readily identified in the real world, and it is possible to model positional accuracy separately from the accuracy of feature attributes. In the latter case, the representation records the value of a variable at every location in the mapped space, $z = f(\mathbf{x})$, where $\mathbf{x}$ is a location of appropriate dimensionality (one, two, three, or four dimensions depending on the application). The variable z may be a single value measured on a nominal, ordinal, interval, ratio, or cyclic scale; a vector of such values; or a tensor.

Separation of positional and attribute accuracy is possible for continuous fields only in certain unusual circumstances, since it is generally not possible to resolve whether errors arise from the correct value being recorded at the wrong location, or the wrong value being recorded at the correct location, or intermediate conditions between these two extremes. Separation is possible only when a singularity occurs in the real world, such as a peak, pit, sharp ridge, channel, or cliff; where the singularity is sufficiently obvious in the field to be identified unambiguously; and where the representation records its location with sufficient accuracy, either explicitly or implicitly.

Mark and Csillag (1989) define an area-class map as consisting of homogeneous areas separated by boundaries, each area having a single class. Examples include maps of soil class, vegetation-cover class, climate class, and land-use class. An area-class map is clearly a cartographic representation of a nominal field, since a single value of a nominal variable is associated with every location. To emphasize the nominal nature of the variable the symbol c will be used in this paper, $c = f(\mathbf{x})$. The ordinal case will be considered separately where appropriate.

All types of geographic data are subject to uncertainty, since it is impossible to create a perfect representation of the infinitely complex real world. The literature on uncertainty is now extensive; Zhang and Goodchild (2002) provide a recent review. Several theoretical frameworks have been employed by researchers in this area, including geostatistics (Isaaks and Srivastava 1989; Goovaerts 1997) and traditional error analysis. But while useful models exist for discrete objects, and for interval/ratio fields, there remain vexing theoretical and practical questions associated with the modeling of uncertainty in nominal and ordinal fields. The purpose of this paper is to examine the state of knowledge in this area; to propose some tests of adequacy; and to review a possible solution.

The nature of uncertainty

Distortions

The process by which area-class maps are created is long and complex, and involves human interpretation at several stages. Some such maps are created from aerial photographs or satellite images by supervised or unsupervised classification, which may or may not be checked against ground truth. Others, such as soil maps, are created by taking samples at a small number of points, and generalizing from these points to a complete map using aerial photographs, ground observation, and maps of variables assumed to be correlated with soil class. Definitions of classes are commonly vague, and the rules used by human interpreters are often more detailed than the rules recorded and shared with others. As a result, the process is not replicable in a strict scientific sense. Two maps of the same theme for the same area will not agree, because of differences in:

- the level of detail of the mapping, or its inverse, the degree of generalization;
- variation among observers in their interpretations of the definitions of classes, subjective additions to those definitions, training, and knowledge of the phenomenon being mapped;
- errors in measuring instruments;
- the use of different types of classifiers, and different training sites for supervised classification; or
- differences in the sensors used to obtain imagery.

A common approach to error analysis is to assume the existence of a true value, and to regard the observed or measured value as a distortion of the true value, $z' = z + \delta z$. In other words, scientific measurements are disturbed by stochastic effects. A suitable measure of the unsigned magnitude of δz is termed the *accuracy* of the measurement, while measures of the variation of δz about its own mean are termed *precision*. Accuracy measures differences from the truth, and assumes that truth can be defined, whereas precision measures only the repeatability of measurements.

Unfortunately it is normally not possible to regard a geographic data set as an assemblage of such measurements, because the links between original measurements and the final map are typically lost during the process of interpretation, compilation, and transformation. These processes generally induce strong spatial autocorrelations in δz, such that nearby values are likely to be more similar than distant values (an instance of Tobler's First Law of Geography; Tobler 1970), and such autocorrelations must be built into successful error models. Moreover, even if the accuracy of measuring instruments is known, it is typically not possible to propagate such knowledge through the processes of map creation, and instead estimates of δz must be obtained by other means.

The same general conditions apply to area-class maps, but in this case it is obviously not possible to regard error as an additive distortion. Instead, much use has been made of transition probabilities, compiled as an *error matrix* or *confusion matrix*, which captures the probability that a true value c will be recorded as c' (Longley *et al.* 2001). The same argument regarding autocorrelation clearly applies, since the probability of misclassification at some point $\mathbf{x}$ is conditional on the probability of misclassification at another point $\mathbf{x}+\delta\mathbf{x}$ where $\delta\mathbf{x}$ is small.

Some strategy is needed for dealing with the many situations in which it is impossible to believe in a true class c. The strategy adopted here is to define c in such cases as a *consensus* class, or the modal class over a large series of repeated mappings.

Properties of area-class maps

Two alternative representations are commonly used in geographic information systems (GIS) to record the contents of area-class maps. In *raster* representations the map is divided into an array of rectangular cells, and a single class is recorded in each cell. Boundaries between classes must follow cell boundaries, and are consequently jagged. The spatial resolution of the raster is defined by the cell size.

In *vector* representations the boundaries of the area-class map are recorded explicitly, as sequences of

points connected by straight lines. Two vector options are common. In the *coverage* model (the terminology used here is that of ESRI and its GIS products) the boundaries are defined as collections of *arcs*, each arc being defined by the stretch of boundary between two junctions or *nodes*, and separating two adjacent classes. In the *shapefile* model each area is recorded separately as a complete polygon. Coverages have several advantages: each common boundary is recorded only once, as an arc, rather than twice; there is consequently no possibility of divergence between the two versions; and the map can be digitized more economically. Note that no obvious basis exists for determining spatial resolution in the vector case.

Vector representations suggest the use of uncertainty models for area-class maps that emulate those used for discrete objects. In such models, the components of the vector representation—the polygons, arcs, and nodes—would each be analyzed for uncertainty, with positions and attributes treated separately. Models would be found for the uncertainty of polygon attributes, arc positions, and node positions. But there are several obvious objections to such approaches:

- repeated mappings will differ not only in attributes and positions, but also in the numbers of polygons, arcs, and nodes—in other words, they will differ *topologically*;
- positional uncertainties in boundaries (arcs) will vary widely, depending on the clarity with which the boundary is identified in the field; and
- confusion of polygon attributes will vary widely within polygons, but an object-based model has no way of addressing within-polygon confusion.

With respect to the last point, one might assume that confusion is greater near polygon boundaries than in the center of polygons (see, for example, the *egg-yolk* model of Cohn and Gotts 1996). But Goodchild (2001) has argued to the contrary—that there are good reasons to believe that confusion is sometimes greater in the center of the polygon.

Area-class maps exist at many levels of generalization, and it is common to separate *geometric* generalization from *thematic* generalization. Both types of generalization typically occur when moving from detailed to coarse mapping: the number of polygons is reduced, and the number of classes is also reduced. Conversely, a more detailed map will contain more polygons, and will also use a more detailed classification.

An important question arises, however, when maps at different levels of generalization are compared. Attempts to automate the process of generalization often make use of *merge* rules, reducing the numbers of polygons by removing the arcs between suitably chosen neighbors. This would imply that when a classification is refined, existing boundaries remain and are augmented by new boundaries between new subclasses. But in practice a very different form of behavior is observed: as the classification is refined, new polygons emerge along the boundaries between classes, in areas where rapid change was poorly approximated at coarse scales.

Requirements for a model of area-class uncertainty

We are now in a position to identify a number of requirements for models of area-class uncertainty, based on the preceding arguments.

1. The model should address the variable confusion that exists *at every point* between the recorded class c' and the true or consensus class c.
2. When implemented as a stochastic process, repeated runs of the model should produce *realizations*, each of which has the properties of an area-class map, and variations between realizations should successfully emulate variations between repeated compilations of the same map, including topological variations.
3. The model should successfully emulate the *autocorrelations* that exist between outcomes at neighboring points.
4. The model should successfully emulate the effects that occur when area-class maps are

generalized, both geometrically and thematically.

5. Realizations generated by the model should be invariant under changes in the underlying representation. For example, if a raster is used, outcomes should not depend on cell size or geometry, since cells are an artifact of the representation and have no meaning in reality.
6. The model should preserve the properties of nominal or ordinal fields, as appropriate. In the nominal case, results should be invariant under a reordering of classes, since the order of classes has no significance in nominal data.

Models of uncertainty

In this section we review an assortment of models of uncertainty in area-class maps, and assess each one by comparison to the list of requirements above. In some cases the use of the term *model* may be a stretch, since the model may be implicit in a method of measurement, rather than explicit. After the first two, all models assume that the outcome at any point $\mathbf{x}$ is a realization under a vector of probabilities $P(\mathbf{x}) = \{p_1, p_2, \ldots p_m\}$ where m is the number of classes, and p_i denotes the probability that the observed class at the point will be i. Note that this model does not explicitly recognize a true or consensus class at $\mathbf{x}$, although this might be interpreted as the class with the highest probability. The various models described below that use this basic structure differ in how they ensure autocorrelation between neighboring outcomes.

The confusion matrix

As noted earlier, the confusion matrix describes the probabilities of misclassification, that is, the probabilities that c' will be observed given c. It is applied on both a *per-polygon* and a *per-point* basis. In the former case the assumption that c cannot vary within polygons is clearly at variance with (1). The absence of any statement about autocorrelation is at variance with (3). A per-polygon implementation cannot produce variable topology among realizations, and is thus at variance with (2). Finally, the model has no means of addressing generalization.

The epsilon band and related models

These models address positional accuracy in boundaries and arcs (Perkal, 1966). In the deterministic version some minimum distance ε is defined, such that all observed boundaries lie within this distance of all true or consensus boundaries. In a probabilistic version (Goodchild and Hunter 1997) a given percentage of the length of true or consensus boundaries is assumed to lie within this distance of observed boundaries. Since the model addresses only positional accuracy in a fixed topology, and assumes uniformity in the degree of positional accuracy, it is clearly at variance with (1), (2), and (4).

Convolution

We now move to the set of models based on vectors of probabilities, as discussed earlier. A simple way to realize this model would be to approximate the space as a raster, and to select a random outcome in each cell based on the average of probabilities in the cell. For brevity, $P(\mathbf{x})$ will be assumed to refer to this average probability over the cell.

The results of such a simple random assignment would be at variance with (3), since outcomes in neighboring cells would be statistically independent, although perfectly correlated within cells. They would also be at variance with (5), since cell size would clearly affect the outcome. Visually, realizations would exhibit an unreasonable *salt and pepper* effect.

A simple way to remove this visual effect that is often used in image processing is to apply a *convolution*, an operation in which each cell's value is replaced with a neighborhood value. For nominal data, the appropriate value is the *mode*, or the commonest class in the neighborhood. A convolution could be applied over a 3x3 neighborhood centered on each cell, or over a larger neighborhood, removing the visual salt and pepper, and inducing an autocorrelation in outcomes. The size of the neighborhood could be determined by comparison with the real world, and thus made independent of cell size, satisfying (5).

But this method fails in one crucial respect. In the convolution process rare classes are less likely to

survive, and hence the posterior proportion of cells allocated to any one class is not algined with the prior probability assigned to that class.

Sequential assignment

A number of methods exist for overcoming the objection to the convolution method, that posterior proportions do not equal prior probabilities. Goodchild, Sun, and Yang (1992) describe a method using random interval/ratio fields. In every cell a variable z is generated with a known statistical distribution, and with a known autocorrelation structure. This variable is transformed to a uniform distribution between 0 and 1, and then used to assign a class to the cell based on the prior probabilities. For example, if the probabilities are {0.2,0.3,0.5}, values of z in the range {0.0,0.2} result in Class 1, values in the range {0.2,0.5} result in Class 2, and values in the range {0.5,1.0} result in Class 3. We assume that ties are vanishingly improbable. Posterior proportions are now aligned with prior probabilities. (1) is satisfied because probabilities are assigned to each cell. Realizations show topological variations, satisfying (2). The autocorrelation in the variable z ensures that outcomes in neighboring cells will be correlated, satisfying (3). Moreover, the parameters of the autocorrelation can be made independent of cell size, satisfying (5).

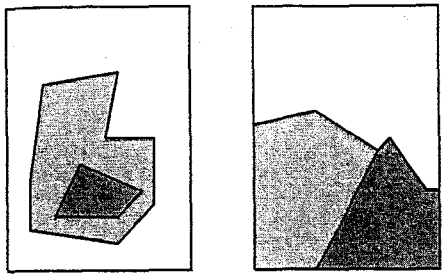

Figure 1. (Left) Adjacencies are possible only between classes that are adjacent in order; note the absence of three-valent nodes and the resemblance to a contour map. (Right) A typical area-class map dominated by three-valent nodes.

The generalization requirement can be handled in several ways. First, cell size can be increased, leading to fewer, coarser outcomes. Second, the autocorrelation structure of the random field can be coarsened by changing the values of the relevant parameters, leading to outcomes that persist over longer distances. Finally, the spatial variation in P can be coarsened using a simple convolution. However although these approaches yield realizations of different levels of generalization given the same inputs, they do not provide a method for generalizing a single realization.

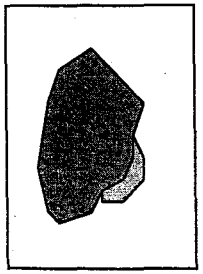

Figure 2. In a sequential assignment the second and subsequent classes assigned (light grey) tend to have different shapes from the first class (dark grey).

However, results are not invariant under reordering of classes, which is at variance with (6) for nominal data. Since the random field is autocorrelated, pairs of adjacent cells are likely to receive either the same class, or classes that are adjacent in the order of the probability vector. In the example above, a small change in z between adjacent cells will produce either no change of class, or a change to a neighboring class. Hence adjacencies between Classes 1 and 2, and Classes 2 and 3, will be much more likely than adjacencies between Classes 1 and 3. The visual effects of this problem are quite obvious: each realization shows a closer resemblance to a contour map, on which contours never meet, than to a choropleth map on which nodes are common, with each node tending to be at the junction of three arcs (to be *three-valent*). Figure 1 presents a caricature of the problem.

Indicator Kriging can also be used as a model, within the geostatistical framework (Isaaks and Srivastava, 1989; Goovaerts, 1997). The autocorrelation structure is captured in the form of the variogram or correlogram. Classes are assigned sequentially; in the first stage, cells are assigned either to Class 1 or not. A second stage then assigns some of the remaining cells to Class 2, and the process continues through m-1 stages. This process also generates results that are order-

dependent, and at variance with (6) for nominal fields. The results are rather different visually, however. Three-valent nodes are common, as illustrated in Figure 2, but the shapes of areas of Class 1 are distinct from those of other classes (acute angles at nodes and concavities are more common in subsequent classes).

Shuffling across realizations

The convolution method discussed earlier attempted to resolve the lack of autocorrelation by a simple neighborhood operation. An alternative would be to generate a large number of realizations by simple random assignment, such that the proportion of classes assigned to each cell matches the prior probabilities. But instead of a convolution, a process would then be invoked to swap the contents of cells across realizations. The proportions of classes assigned to a given cell would be unchanged, so there would be no inconsistency with prior probabilities. But swaps could be evaluated to determine whether there is an improvement in autocorrelation, relative to some defined goal. After a large number of swaps, each realization would be close to the target autocorrelation. The concept of swaps to achieve autocorrelation is discussed by Goodchild (1980).

Cell swapping methods satisfy requirements (1), (2), (3), and (6). For the results to be independent of cell size, it would be necessary to measure spatial autocorrelation in an appropriate way. This could be done, for example, by using a variogram as the standard of autocorrelation, with ground distance as the ordinate. Cell swapping methods satisfy the generalization requirement, using the same approaches outlined in the previous section. However they are similarly unsuitable for generating different levels of generalization of the same realization.

A phase-space model

In a search for suitable methods for simulating area-class maps, Goodchild and Dubuc (1987) proposed a model that can be adapted easily to the more general problem of modeling uncertainty. Generate n random fields covering the area, to provide a vector of values $Z = \{z_1,z_2,\ldots,z_n\}$ at any point. The model can be interpreted causally, if each variable is assumed to be a factor responsible for determining class at a point. For example, the soil class at a point might be determined by soil moisture content, soil organic content, soil pH, *etc.*; or vegetation class might be determined by mean annual temperature, and mean annual precipitation. An n-dimensional space defines how each combination of variables results in a class; the space is simply partitioned into zones associated with classes. Goodchild and Dubuc (1987) term this a *phase* space by analogy to physical processes. Figure 3 represents the stages of the model, which is essentially a multidimensional version of the Goodchild, Sun, and Yang (1992) model.

The role of the phase space is open to different interpretations. In the examples of soil or vegetation class mapping outlined above the phase space forms a real part of a physical process, and is presumably recoverable from analyses of vegetation or soil maps. The interval/ratio variables Z are also in principle measurable. In another context, the process resembles strongly that of image classification, and the Z values represent measured radiation in different parts of the electromagnetic spectrum. Different classifiers impose different constraints on the process: the *parallelepiped* classifier is named for the shapes of the domains in phase space assigned to each class; while the boundaries between classes take a different form with maximum-likelihood classifiers. Finally, phase space may have to be calibrated from observations, if no other evidence exists to help determine its properties, or if the variables defining it have no physical interpretation.

The calibration of phase space is clearly problematic, because of the number of unknowns. These include the number of random fields, and the sizes and shapes of the various domains. Sizes can be obtained from observations of the relative abundance of different classes. For example, if the random variables are uniformly distributed in the interval {0,1} then the proportion of outcomes of Class c is equal to the area assigned to that class in phase space. The adjacencies of classes can also be determined easily, since only classes that are adjacent in phase space can appear adjacent in geographic space, given the necessary assumption of continuity of the variables Z.

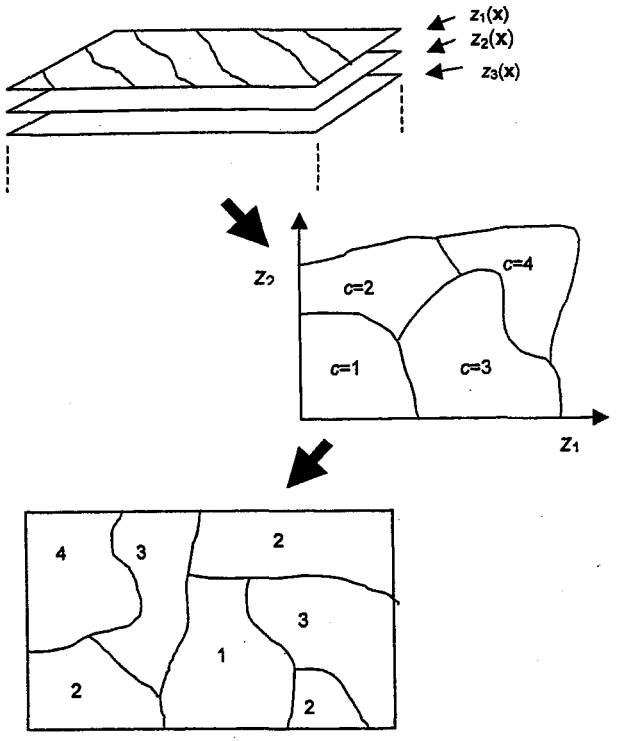

Figure 3. A phase-space model. A number of continuous fields is generated. A phase space assigns every vector of field values to a class. Finally, the interpolated fields and phase space are combined to form a nominal field.

Two versions of uncertainty are possible with this model. In the first, each realization uses independent random fields. The system has no *memory* between one realization and the next, and the proportions of outcomes at any point are the same, independent of location. This is the approach proposed by Goodchild and Dubuc (1987) for simulation of random maps. An alternative approach, however, proposes that each variable has a true value at every point, and that individual realizations are distortions of this true value using an error model of the type $z' = z + \delta z$. Outcomes vary from realization to realization, but there is strong persistence between realizations as well. In the earlier models this persistence was the result of the prior probabilities, but in this case no prior probabilities are specified; instead, they are implied by the values of Z, the magnitude and distribution of distortion, and the configuration of the phase space. Alternatively, the phase space might be treated as non-stationary, as it effectively is in the case of the one-dimensional Goodchild, Sun, and Yang (1992) model, but this raises additional issues of calibration.

The model clearly satisfies (1) for the reasons just discussed, and also (2), (3), (5), and (6). The thematic aspects of within-realization generalization are easily handled by refining the phase-space configuration. Geometric aspects might be handled by smoothing the variables Z, which will result in smoother boundaries, or more directly by applying some kind of geometric smoothing to the arcs.

Discussion

The phase-space model clearly satisfies the requirements. Although it might also be acceptable as a model of how classes arise in practice, and of how processes of vegetation dominance are actually determined on

the real landscape, it nevertheless will be very difficult to apply as a model of uncertainty, in the absence of any inferred process, because of the difficulties associated with calibration. In essence, then, the choice of model must depend to some degree on the availability of information. When estimates of P are available, shuffling across realizations will be the best choice; but when there is some reasonable basis for estimating Z, the phase-space model is clearly preferable.

Sequential assignment was faulted for its lack of invariance under reordering of classes. But this is an acceptable property when the focus is on an ordinal field, and there may be circumstances where this assumption is justified. Consider, for example, a classification of land cover into agricultural, urban, and forest. It is most likely that the urban class represents an encroachment on the other classes, and that the shape of the urban-class areas is characteristic of urban processes rather than agricultural processes. In such cases maps may well resemble Figure 2, where the shape of the urban encroachment is distinctly different from the other shapes. If classes can be ordered, such that Class i is an encroachment on Class i+1 for all i (in other words, i+1 is antecedent to i), then a method that is not invariant under class reordering is clearly appropriate. The problem with the Goodchild, Sun, and Yang (1992) method is somewhat different, and results from the use of a one-dimensional phase space.

The phase-space model emphasizes the importance of an approach that is *model-based*, in the sense that it bears some relationship to the process that actually assigns classes on the landscape. Understanding of uncertainty will always be better informed when it can be based on such knowledge, just as traditional error analysis relies on the knowledge that measuring instruments are commonly subject to Gaussian errors. Metrics of uncertainty that have no such grounding, such as the epsilon-band model, must necessarily invent a process of error creation that has no correspondence to any real process, and is consequently inherently suspect.

The various methods reviewed in this paper serve to emphasize once again the importance of spatial autocorrelation in understanding the impacts of uncertainty. Almost all GIS operations involve a *joint* analysis of conditions at multiple locations, and thus require some knowledge of the joint distribution of uncertainty. Since spatial autocorrelation is endemic, the joint distribution of properties at two locations is never a simple combination of the marginal distributions at the two locations, but instead requires some knowledge of autocorrelation. This applies whether the purpose of the analysis is as simple as the measurement of area, or as complex as many environmental models.

The issue of generalization within realizations has been addressed at several points in this paper. An interesting problem in this area is as follows. Suppose one has access to a comparatively coarse dataset, and recognizes that much more detailed data are required for a given application. One also has access to both coarse and detailed data for another area (the *reference* area). It would be useful to be able to generate realizations of fine data that are consistent with the available coarse data, subject to the constraint that the additional detail be similar to that observed in the reference area. In the interval/ratio case the problem can be tackled by co-Kriging (Kyriakidis, Shortridge, and Goodchild, 1999). For the nominal/ordinal models that rely on P, no information is available from the coarse dataset to assign values other than 0 or 1 to probabilities. Conditional co-Kriging might be used, but subject to the problems of sequential assignment noted earlier. But the coarse dataset might be used to obtain constraints on a phase space, and comparison between the coarse and fine datasets might provide an additional basis for calibration. These possibilities might form a topic for useful further research.

References

Cohn AG, Gotts NM, 1996. The 'egg-yolk' representation of regions with indeterminate boundaries. In PA Burrough, AU Frank, editors, *Geographical Objects with Indeterminate Boundaries*, pp. 171–187. New York: Taylor and Francis.

Goodchild MF, 1980. Algorithm 9: simulation of autocorrelation for aggregate data. *Environment and Planning A* 12: 1073–1081.

Goodchild MF, 2001. A geographer looks at spatial information theory. In DR Montello, editor, *Spatial Information Theory: Foundations for Geographic Information Science: Proceedings of the International Conference COSIT 2001, Morro Bay, CA, September*. Lecture Notes in Computer Science 2205, pp. 1–13. Berlin: Springer.

Goodchild MF, Dubuc O, 1987. A model of error for choropleth maps with applications to geographic information systems. *Proceedings, Auto Carto 8*, pp. 165–174. Falls Church, VA: ASPRS/ACSM.

Goodchild MF, Hunter GJ, 1997. A simple positional accuracy measure for linear features. *International Journal of Geographical Information Systems* 11(3): 299–306.

Goodchild MF, Sun G, Yang S, 1992. Development and test of an error model for categorical data. *International Journal of Geographical Information Systems* 6(2): 87–104.

Goovaerts P, 1997. *Geostatistics for Natural Resource Evaluation*. New York: Oxford.

Isaaks EH, Srivastava RM, 1989. *Applied Geostatistics*. New York: Oxford.

Kyriakidis PC, Shortridge AM, Goodchild MF, 1999. Geostatistics for conflation and accuracy assessment of digital elevation models. *International Journal of Geographical Information Science* 13(7): 677–708.

Longley PA, Goodchild MF, Maguire DJ, Rhind DW, 2001. *Geographic Information Systems and Science*. New York: Wiley.

Mark DM, Csillag F, 1989. The nature of boundaries on 'area-class' maps. *Cartographica* 26(1): 65–78.

Perkal J, 1966. *On the Length of Empirical Curves*. Discussion Paper No. 10. Ann Arbor: Michigan Inter-University Community of Mathematical Geographers.

Tobler WR, 1970. A computer movie simulating urban growth in the Detroit Region. Economic Geography 46: 234–240.

Worboys MF, 1995. *GIS: A Computing Perspective*. New York: Taylor and Francis.

Zhang J, Goodchild MF, 2002. *Uncertainty in Geographical Information*. New York: Taylor and Francis.