Geography, Spatial Data Analysis, and Geostatistics: An Overview

Robert P. Haining,1 Ruth Kerry,2 Margaret A. Oliver3

1Department of Geography, University of Cambridge, Cambridge, UK, 2Department of Geography, Brigham Young University, Provo, UT, and CRSSA, Rutgers University, New Brunswick, NJ, 3Department of Soil Science, University of Reading, Reading, UK

Geostatistics is a distinctive methodology within the field of spatial statistics. In the past, it has been linked to particular problems (e.g., spatial interpolation by kriging) and types of spatial data (attributes defined on continuous space). It has been used more by physical than human geographers because of the nature of their types of data. The approach taken by geostatisticians has several features that distinguish it from the methods typically used by human geographers for analyzing spatial variation associated with regional data, and we discuss these. Geostatisticians attach much importance to estimating and modeling the variogram to explore and analyze spatial variation because of the insight it provides. This article identifies the benefits of geostatistics, reviews its uses, and examines some of the recent developments that make it valuable for the analysis of data on areal supports across a wide range of problems.

Introduction

As an introduction to this special issue, the purpose of this article is to provide an overview of the core concepts and techniques of geostatistics, together with a short literature review of its application in the environmental sciences and in geography. Geostatistics has a long history of application in the environmental sciences where data are on a point or small regular area support, but it is now being applied to regional data where data are on an areal support that might be large and regular or irregular. We describe the new tools associated with the latter type of data and contrast them with techniques of spatial data analysis with which geographers, especially human geographers, are familiar. These techniques have descended more or less directly from work that began in the 1960s by Dacey (1968) and Cliff and Ord (1969), also the subject of a recent special issue of Geographical Analysis (2009, issue 4). Geostatistics, by contrast, has a different lineage and uses a different set of tools and techniques.

Correspondence: Ruth Kerry, Department of Geography, 690 SWKT, Brigham Young University, Provo, UT 84602
e-mail: ruth_kerry@byu.edu

The use of the term *spatial analysis* in geography can be traced back to the 1950s (see, e.g., Berry and Marble 1968). It includes several distinctive elements (Haining 2003, pp. 4–5), but the statistical analysis of spatial data is the focus here, referred to by statisticians as *spatial statistics* (Ripley 1981) or *statistics for spatial data* (Cressie 1993). Geographers often refer to these as methods for *spatial data analysis* (Haining 1993), and many of these models and techniques figure prominently in geographic information science (Goodchild and Haining 2004) and spatial econometrics (Anselin 1988).

The roots of spatial statistics can be traced back to the early part of the twentieth century to analyses of agricultural field trial data by statisticians. Geostatistics is a component of spatial statistics, although its evolution has been led principally by applied scientists and mathematicians rather than by classically trained statisticians. This historical context may explain why little cross-fertilization occurred with other branches of spatial statistics until quite recently (Cressie 1993; Diggle and Ribeiro 2007) and why geostatistics is distinctive.

Any methodology for analyzing spatial data needs to recognize that such data have the *fundamental* property of spatial dependence or spatial autocorrelation. For many attributes, values recorded at locations close together in space are correlated (autocorrelated); as the separating distance increases, autocorrelation weakens. The autocorrelation structure in a region may be complex, with several scales of variation nested within, or superimposed on, one another, varying with direction (anisotropic) and between subareas (spatially heterogeneous). Quantifying spatial dependence matters, whether the purpose of an analysis is to interpolate, to fit a regression model, or to test a hypothesis (Haining 2003, pp.33–36, 40–41). Different branches of spatial statistics model spatial dependence in different ways.

Geographical data acquire other properties as a consequence of the chosen representation of geographic space. The areal units into which a study region may be partitioned for reporting attribute values often vary in size and shape (e.g., census output areas). If the population denominator for rates (e.g., mortality rates) varies, the standard errors of such statistics are not constant across a map. Therefore, values obtained from irregularly sized areas may not be directly comparable, making map interpretation potentially problematic. Data for areas with small populations suffer from the “small number problem” (Haining 2003, pp.196–99). Methods must be able to deal not only with spatial dependence but also with the properties acquired as a consequence of a chosen representation.

This article elucidates the distinctiveness of geostatistics and how it differs from other branches of spatial statistics that are concerned with the same general problem of analyzing spatial variation and illustrates the relevance of geostatistics to geographers. Physical geographers have been somewhat receptive to geostatistics, partly because their problems and data are similar to those of other earth scientists, who were among the early practitioners of geostatistics. Nevertheless, geostatistical methods generally have been used only for basic interpolation. Human geographers, by contrast, generally have taken little interest in geostatistics for spatial data.
analysis because often attribute values are not defined everywhere in a region, and
data values are defined for irregular spatial units (e.g., census areas).

To verify some of these assertions, we searched the bibliographic databases
Geobase and Information Sciences Institute Web of Knowledge (ISI).2 These
searches produced 4377 and 1596 hits, respectively. Table 1 shows the journals
that publish most articles on geostatistics (also see Zhou et al. 2007). Physical geo-
graphers publish in these journals. However, of the institutions with authors who
have published more than three articles on geostatistics articles identified by the
Geobase search, only 9 out of 61 were geography departments, and only 50 of over
4000 articles found with this search included authors from geography departments.
Table 1 also shows specifically which geography journals2 have published the most
articles on geostatistics; none is devoted solely to human geography. Figures 1a and
1b show that the number of articles has increased over time, both in general and in
geography; but for the latter, the increase seems to have leveled out since 2000.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Number of geostatistics articles identified by Geobase and ISI searches in the top 10 journals that publish most geostatistics articles in general and in geography</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of articles in top 10 journals publishing most geostatistics articles:</td>
<td></td>
</tr>
<tr>
<td>In general</td>
<td>In geography</td>
</tr>
<tr>
<td>Journal name</td>
<td>Geobase</td>
</tr>
<tr>
<td>Mathematical Geology</td>
<td>310</td>
</tr>
<tr>
<td>Geoderma</td>
<td>161</td>
</tr>
<tr>
<td>Water Resources Research</td>
<td>143</td>
</tr>
<tr>
<td>Computers and Geosciences</td>
<td>138</td>
</tr>
<tr>
<td>Journal of Hydrology</td>
<td>105</td>
</tr>
<tr>
<td>Soil Science Society of American Journal Environmentetrics</td>
<td>78</td>
</tr>
<tr>
<td>International Journal of Remote Sensing Stochastic Environmental Research and Risk Assessment</td>
<td>66</td>
</tr>
<tr>
<td>The Environmental Monitoring and Assessment</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Between 1990 and 2008, only 3.4% of the Geobase results and 2.1% of the ISI results identified geostatistics articles that were published in geography journals.

Geostatistics: Its core concepts, techniques, and relationship with other approaches to spatial data analysis

An historical perspective on geostatistics

The use of the term geostatistics stems from Matheron’s development of a comprehensive theory for the prediction of properties in geographical space. Matheron’s (1963) theoretical framework for geostatistics was developed from D. G. Krige’s empirical ideas for improving predictions of the amount of gold in rock (Krige 1951) by using neighboring samples. Matheron (1963) uses the term kriging for the method of optimal prediction or estimation in geographical space—a spatial best linear unbiased predictor (BLUP). Matheron’s fundamental contribution is to define the covariance or variogram of a random field based on a probabilistic or stochastic approach to the analysis of spatial variation, which recognizes its complexity and treats as random that aspect of the variation that appears to be random. Matheron’s approach led to the formulation of models of spatial variation that provide weights for the BLUP (see Bilodeau, Meyer, and Schmitt 2005 for more detail on Matheron’s contribution to geostatistics).

Matheron’s ideas had been anticipated in earlier work. The article by Mercer and Hall (1911), as well as the appendix to it by Student (1911), anticipates some of the fundamental features of modern geostatistics: support, spatial dependence, correlation range, and the nugget effect. Kolmogorov (1941) devises the “structure function” to represent both spatial correlation, which we now recognize as the variogram, and his method of interpolation now known as kriging (Ripley 1981, pp. 44–50). Matérn (1960) theoretically derives some of the familiar “permissible”
functions to describe spatial covariance from random point processes. These are equivalent to Jowett’s (1955) “serial variation function,” which now is known as the variogram.

The geostatistical approach to describing spatial dependence: the covariance and variogram

Most spatial properties vary in such a complex way that variation cannot be defined deterministically, and thus the basis of geostatistics is to treat the variable of interest as a random variable. Traditionally, a variable is defined on a continuous surface such that at each point, \( s \), in space, a range of values exists for an attribute, \( Z(s) \), and the one observed, \( z(s) \), is drawn at random from a probability distribution. The set of random variables, \( \{ Z(s) : s \in D \} \), where \( D \) is a subset of two-dimensional space (\( R^2 \)), is a random field, and the actual values of \( Z \) observed are for just one of a potentially infinite number of realizations of it (the “superpopulation” view). Geostatistics is based on regionalized variable theory (RVT), which provides a sound model of how properties vary in space. It recognizes gradual change across \( R^2 \), locally erratic and structured components of variation, and uncertainty.

To describe the variation of an underlying random field and to estimate the mean and variance of an attribute, the spatial autocovariance (or covariance) is estimated to describe quantitatively the relation between pairs of points a given distance apart. This covariance is given by

\[
C(s_i, s_j) = E\{[Z(s_i) - \mu(s_i)][Z(s_j) - \mu(s_j)]\}
\]

where \( \mu(s_i) \) and \( \mu(s_j) \) are the means of \( Z \) at \( s_i \) and \( s_j \), and \( E[.\] \) is the expected value. Because only one realization of \( Z \) exists at each point, these means are unknown.

To proceed, geostatistics invokes assumptions of stationarity, which means that certain properties of a random field are assumed to be the same everywhere. We assume that the mean, \( \mu = E[Z(s)] \), is constant for all \( s \), and hence \( \mu(s_i) \) and \( \mu(s_j) \) can be replaced by \( \mu \), which can be estimated by repetitive sampling. When \( s_i \) and \( s_j \) coincide, equation (1) defines the variance or the a priori variance of a field, \( \sigma^2 = E\{[Z(s) - \mu]^2\} \) which is assumed to be finite and, as for the mean, the same everywhere. When \( s_i \) and \( s_j \) do not coincide, their covariance depends on their separation and not on their absolute positions, a property that applies to any pair of points separated by lag \( h \) (a vector in both distance and direction). Therefore, given two points \( s_i \) and \( s_j \) separated by lag \( h \),

\[
C(s_i, s_j) = E\{[Z(s_i) - \mu][Z(s_j) - \mu]\}
= E\{[Z(s)]{Z(s + h)} - \mu^2\}
= C(h)
\]

which is also constant for any given \( h \). This constancy of the first and second moments of a random field constitutes second-order or weak stationarity. Equation (2) indicates that the covariance is a function of the spatial lag and describes quantitatively the dependence between values of \( Z \) with changing separation or lag.
distance. Often the autocovariance is converted to the dimensionless autocorrelation by

$$\rho(h) = \frac{C(h)}{C(0)}$$

where $C(0) = \sigma^2$ is the covariance at lag 0.

The mean often appears to change across a region, and the variance will then appear to increase indefinitely as the extent of an area increases. The covariance cannot be defined because no value for $\mu$ exists to insert into equation (2). This situation is a departure from weak stationarity. Matheron’s (1965) solution is the weaker intrinsic hypothesis of geostatistics. Although the general mean might not be constant, it would be constant for small distances, and so the expected differences would be zero:

$$E[Z(s) - Z(s + h)] = 0$$

and the expected squared differences for those lags define their variances:

$$E[(Z(s) - Z(s + h))^2] = \text{var}[Z(s) - Z(s + h)] = 2\gamma(h)$$

(3)

The quantity $\gamma(h)$ is known as the semivariance at lag $h$, or the variance per point when points are considered in pairs. As for the covariance, the semivariance depends only on the lag and not on the absolute positions of the data points. As a function of $h$, $\gamma(h)$ is the semivariogram or, more usually, the variogram. The variogram can be applied when the assumptions of second-order stationarity do not hold or when uncertainty exists about whether they do or do not. This result makes the variogram a valuable tool, and accordingly it has become the cornerstone of geostatistics. If the field $\{Z(s) : s \in D\}$ is second-order stationary, the semivariance and covariance are equivalent.

The usual method of computing the empirical semivariances from data, $\{z(s_1), z(s_2), \ldots, z(s_n)\}$, at sample points $s_1, s_2, \ldots, s_n$, is Matheron’s (1965) method of moments (MoM) estimator:

$$\hat{\gamma}(h) = \frac{1}{2m(h)} \sum_{i=1}^{m(h)} (z(s_i) - z(s_i + h))^2$$

(4)

where $z(s_i)$ and $z(s_i + h)$ are the actual values of $Z$ at locations $s_i$ and $s_i + h$, which are separated by the lag $h$. The sum is over $m(h)$, which is the number of paired comparisons separated by $h$. By changing $h$, an ordered set of semivariances is obtained; these semivariances constitute the experimental or sample variogram. Other estimators of the variogram exist, most notably the residual maximum likelihood (REML) estimator (Pardo-Igúzquiza 1998a, b). Diggle and Ribeiro (2007) also suggest a model-based approach to geostatistics where expert knowledge, in addition to properties of the data, plays a role in determining an appropriate variogram model.

Assuming sufficient data from which to compute a reliable empirical variogram (see Webster and Oliver 1992, pp. 178, 190–91), a best-fitting model is selected
from what are known as authorized (or permissible) functions (see Webster and Oliver 2007, pp. 82–95). Several fitting procedures exist from which to choose (Cressie 1993, pp. 90–104).

The variogram can take on a variety of forms (Fig. 2), and the reader is referred to Webster and Oliver (2007, pp. 56–60) for a discussion of the most common forms and their interpretations and for explanations of commonly used terms that describe important features of the variogram, such as its “sill,” “range,” and “nugget” variance. Introductions to basic geostatistical methodology and terms are also given by Armstrong (1998), Christensen (2001), Goovaerts (1997), and Isaaks and Srivastava (1989). The variogram is a valuable exploratory data tool, regardless of whether an analyst wishes to use other geostatistical tools. For example, if data result in a variogram that appears as a horizontal set of points (i.e., pure nugget)
interpolation would merely give the local observed mean. In such cases, all that is essentially being done is the application of a local smoothing function. The variogram shape may also indicate whether the variation has a spatially random component (nugget, \( c_0 \); Fig. 2a–d) and whether more than one scale of variation is present, which requires a nested model that can be used for factorial kriging (Fig. 2d). If a variogram has no upper bound or sill (Fig. 2c), disjunctive kriging and empirical-BLUP are precluded. Anisotropy in variation can be explored by computing a variogram in different directions. The variogram map (semivariances plotted against separation distance in the x and y directions [Webster and Oliver 2007, p. 75]) can indicate the directions of greatest and least variation.

**Kriging**

Kriging is a generic term for a range of BLUP least-squares methods of spatial interpolation in geostatistics. It can be linear or nonlinear, although the former is more common. Kriging provides not only predictions but also the kriging errors or kriging variances at each prediction location. The original formulation of linear kriging, now known as ordinary kriging (Journel and Huijbregts 1978), is the most robust and most used method. Ordinary kriging assumes that the mean is unknown but constant and that the random field is locally stationary. In linear kriging, the estimate at any location \( s_0 \), \( \hat{Z}(s_0) \), is a weighted linear combination of the data:

\[
\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i z(s_i)
\]

The weights, \( \lambda_i \), are chosen to minimize \( E[(Z(s_0) - \hat{Z}(s_0))^2] \), or the kriging variance, and to ensure that the estimates are unbiased, the weights are constrained to sum to 1 (Webster and Oliver 2007, pp. 155–59). Kriging uses the spatial information described by a variogram function together with the data to predict optimally. The weights depend on the variogram and the spatial positions of the data, \( \{ z(s_i) \} \), in relation to one another and to the target point or block \( s_0 \). Observations that are nearest to a prediction point, \( s_0 \), have the largest weights, but clusters of adjacent observations that are highly correlated are individually down-weighted. Kriging is essentially a local predictor that, depending upon the aims of the prediction, can be applied to point (punctual kriging) or block supports of various sizes (block kriging) and shapes (area-to-area [ATA] kriging) even if the sample information is for points. This theory also provides the basis for designing optimal sampling schemes for kriging using a variogram (Webster and Oliver 2007, pp. 186–88).

Since its original formulation, kriging has been elaborated to tackle increasingly complex problems. Disjunctive (Matheron 1973) and indicator (Journel 1982) kriging are nonlinear forms that give probabilities that attribute values are above or below a given threshold. Both techniques have been widely used as part of the risk assessment of contaminated sites based on thresholds that define serious contamination (Brus et al. 2002; Gaus et al. 2003). Oliver, Webster, and McGrath (1996) also discuss the merits of disjunctive kriging for environmental management. Kerry
and Oliver (2007) use indicator kriging to map soil structure from categorical field observations, and Lloyd and Atkinson (2001) use it to assess the uncertainty of digital elevation model (DEM) estimates. Lark and Ferguson (2004) use both indicator and disjunctive kriging in an agricultural context to map the probability of serious nutrient deficiency. Matheron (1969) originally introduced universal kriging to deal with data with a strong deterministic component (i.e., trend), but the state-of-the-art is empirical BLUP (Stein 1999). The latter is based on a variogram estimated by REML (Lark, Cullis, and Welham 2006). For situations in which some prior knowledge about a drift or trend exists, Omre (1987) introduces Bayesian kriging.

When two or more variables are spatially correlated or co-regionalized, and one is more expensive to obtain than the other, predictions of the less densely sampled variable can be improved by ordinary cokriging (CK) (Matheron 1965). When the cross-variogram has been computed and modeled, the strength of co-regionalization and the potential benefit of CK for the dependent variable can be assessed. Several other methods in geostatistics incorporate secondary information to improve the accuracy of predictions of a primary variable. Examples include regression kriging, simple kriging with locally varying means (SKlm), kriging with an external drift (KED), and multivariate factorial kriging. These methods often have been compared with ordinary kriging (Odeh, McBratney, and Chittleborough 1994; Odeh and McBratney 2000; Rawlins et al. 2009). We refer interested readers to Chilès and Delfiner (1999), Goovaerts (1997), McBratney et al. (2000), Wackernagel (1995), and articles in this special issue for more detail on these methods, which take advantage explicitly of the co-regionalization between variables. Cross-variogram analysis also can describe how the relationship between properties varies with scale in the form of structural correlations (Webster and Oliver 2007, p. 240) or of regionalized correlation coefficients (Wackernagel 1995). These multivariate geostatistical methods have interpretive as well as predictive power, especially if the primary variable is regarded as analogous to the dependent variable and if the secondary variables are treated as the independent variables.

Factorial kriging (Matheron 1982) was developed for nested variation. Long- and short-range components of variation are estimated in a single analysis and can be filtered out. Oliver, Webster, and Slocum (2000) use factorial kriging to filter different scales of variation from remotely sensed imagery and to determine the sources of variation at each scale. Goovaerts and Webster (1994) use this approach to isolate different scales of variation in topsoil copper and cobalt concentrations and to examine the correlation between them at various scales.

The kriging equations can be used with an existing variogram to design a new optimal sampling scheme for kriging (McBratney, Webster, and Burgess 1981). No data are required for this task, and it is advantageous when variables are expensive to obtain and where previous sampling has been either too intensive or insufficient. Atkinson (1991) uses geostatistics to optimize ground sampling strategies for remotely sensed investigations. The structure of spatial autocorrelation characterized
by the variogram also can be used to improve the spatial continuity of classifications (Oliver and Webster 1989). Atkinson and Lewis (2000) use this approach to classify remotely sensed imagery, and Frogbrook and Oliver (2007) use it to identify spatially coherent management zones for precision agriculture. Geostatistical methodology also has been implemented for compressing image data (Oliver, Shine, and Slocum 2005) and for identifying redundant bands in hyperspectral imagery.

Kriging tends to smooth variation in data (see Webster and Oliver 2007, p. 267–68, for further explanation). Often this outcome is required, but sometimes the uncertainty or likely variability of observed patterns needs to be established. Determining uncertainty involves geostatistical simulation by methods such as turning bands, sequential Gaussian simulation, and LU decomposition. Given a specific variogram function and histogram, data with the desired characteristics can be simulated to produce multiple equi-probable realizations with the same variogram and histogram. Simulation also can be conditioned on existing data where the characteristics of the conditioning data are taken into account. Goovaerts (2001) provides an overview of the relative merits of kriging and simulation. Often simulation is used for risk assessment; for example, Bierkens (2006) uses conditional simulation to determine uncertainties associated with groundwater pollution and the associated costs of installing a monitoring network. Webster and Oliver (1992) use turning bands simulation to create fields with a known variogram to determine how well the generating variogram is reproduced by samples of different sizes. They also determine confidence limits for experimental variograms computed from different sample sizes.

Finally, geostatistical methods have been used to investigate temporal as well as spatial autocorrelation. Applications in this context include studies by Kyriakidis and Journel (2001a, b) of sulfate deposition over Europe and by Heuvelink, Musters, and Pebesma (1996) of soil water content. Kyriakidis and Journel (1999) give an overview of these methods.

**Geostatistics for regional data**

More recently, the basic theory of geostatistics has been adapted to predict and map regional data and to model spatial variation in an attribute (a dependent variable) in terms of other variables (independent variables), as in linear regression. This represents a particularly interesting and important development in geostatistics for human geographers.

Oliver et al. (1998) develop binomial CK to analyze the risk of childhood cancer in the West Midlands of England. To estimate the variogram of risk for binomial CK, population size is taken into account to give more weight to pairs of areas with larger populations and hence more reliable rates (Oliver et al. 1998, pp. 286–87). In addition to the kriging weights $\lambda_{ij}$ summing to 1, the weighted sum of the risk covariances between the target point $s_0$ and the centroids of the data supports $s_i$ are constrained to equal the variance of the underlying risk (Oliver et al. 1998, pp. 284–85).
When the population at risk is large and the probability is small, the binomial distribution approaches the Poisson. Monestiez et al. (2005, 2006) introduce Poisson kriging for rates where the small number problem is an issue. This approximation has since been applied to health data (Goovaerts 2005; Ali et al. 2006). Again, areas with larger populations are given more weight when estimating the variogram and when solving the kriging equations. An “error variance” term, derived from the Poisson distribution (Goovaerts 2005, 2006b) and associated with the reliability of the rates based on population size, is introduced.

Articles by Gotway and Young (2002, 2005), Kyriakidis (2004), and Goovaerts (2006b, 2008) address problems associated with change of support in the case of irregularly sized and shaped areas. This situation involves deconvolution of the variogram obtained from areal data to estimate a point-support variogram. Then area-to-point (ATP) kriging is applied where the measurement support is an area, but the prediction support is a point. Goovaerts (2008) describes an approach to ATP kriging that involves discretizing each area, where the number of discretizing points for any area will depend on its size. Each discretizing point has an associated population count obtained from small area census data; therefore, population heterogeneity is taken into account in estimating the deconvoluted variogram (Goovaerts 2008, p. 109). The distance between any two areas (required for calculating the variogram) is measured as a population-weighted average of the straight-line distances between all the points that discretize the two areas (Goovaerts 2008, p. 106). Setting aside in this review the “supplementary and unverifiable hypotheses” (Journel and Huijbregts 1978, p. 231) that ATP kriging (downscaling) involves, the creation of such maps of disease rates helps to reduce the visual bias of choropleth maps, where physically large areas dominate.

Goovaerts (2006b, 2008) combines Poisson kriging with ATP kriging (Kyriakidis 2004) for the analysis of cancer rate data. ATA Poisson kriging is applied when both the measurement and prediction supports are areas (blocks). Taking population into account filters out the influence of the small number problem.

**Comparing geostatistics with other approaches to spatial data analysis**

Regional data are encountered widely in human geography, and methods have been developed that, in many respects, contrast with those of geostatistics.

The initial interest in spatial dependence among human geographers in the 1960s took the form of adapting existing tests for spatial autocorrelation developed by Moran (1950), Geary (1954), and Krishna Iyer (1949) on regular lattices to the irregular shapes of geographical units (Cliff and Ord 1973). Moran’s I and Geary’s c statistics are

\[
I = \left[ \sum_{i,j=1}^{n} \delta(i,j) (z(i) - \bar{z})(z(j) - \bar{z}) \right] / \left[ \left( \sum_{i,j=1}^{n} \delta(i,j) \right) \sum_{i=1}^{n} (z(i) - \bar{z})^2 \right]
\]

and
\[ c = \left[ (n - 1) \sum_{i,j=1}^{n} \delta(i,j)(z(i) - z(j))^2 \right] \left[ \left( 2 \sum_{i,j=1}^{n} \delta(i,j) \sum_{i,j=1}^{n} (z(i) - z(j))^2 \right) \right] \]  

where \( z(i) \) is the observation for area \( i \) \( (i = 1, \ldots, n) \) and \( \delta(i,j) \) is 1 if \( i \neq j \) and \( i \) and \( j \) are contiguous, and 0 otherwise. The numerator of equation (6) shares similarities with an estimator for autocovariance (compare it with equation (1)), whereas the numerator of equation (7) is similar to the MoM estimator for computing the semivariances (compare it with equation (4)). Cliff and Ord’s (1973) adaptation of these statistics is achieved by replacing \( \{ \delta(i,j) \} \) by the more general \( \{ w(i,j) \} \), which specifies a connectivity or weights matrix. This specification can be based on topological and/or geometric properties of the shapes or inter-area flow (or similar) data as a measure of the “strength” of the connections between areas (Haining 2003, pp. 74–87). The resulting \( n \)-by-\( n \) matrix also could be used to identify the different orders of contiguity (similar to the vector \( h \) in geostatistics; see equation (4)), but in practice many analyses consider only the first-order or closest nearest neighbors. Testing takes the form of a null hypothesis of no spatial autocorrelation against a nonspecific alternative hypothesis that spatial autocorrelation is present. This type of testing is of particular interest when examining the residuals from a least-squares regression (for which some further modifications to equations (6) and (7) and their distribution theory are needed), because model errors are required to be independent. Investigations of the power of Moran’s \( I \) and Geary’s \( c \) by simulation and evidence that the asymptotic relative efficiency of \( c \) to \( I \) is \( <1 \) have resulted in Cliff and Ord’s (1973) modified \( I \) statistic becoming the spatial autocorrelation test statistic of choice among geographers.

If spatial autocorrelation is detected in least-squares regression residuals, a common response is to respecify the regression model using a simultaneous spatial autoregressive model for the errors, where the value of the error term at location \( i \) is modeled as a function of the values of the errors at adjacent locations plus an independent (white noise) term. Geographers analyzing regional data tend to model spatial dependence using the simultaneous spatial autoregressive model (specified either with respect to the dependent variable or the errors) rather than calculating the empirical covariances and selecting a function that best fits these auto covariances (for an overview, see Haining 2003, pp. 297–304, 350–58). This approach taken by geographers is similar to that of econometricians when modeling time-series data, which has led to the adoption, by some, of the term spatial econometrics (Anselin 1988).

Regression analysis is used to identify those independent variables that best account for the variation in a dependent variable. The linear function element of a regression model, which includes the independent variables, specifies the mean of a dependent variable. It is variation in the set of significant independent variables that model (or “statistically explains”) variation in the mean of a dependent variable. It is this variation usually that is of most interest to geographers. Spatial de-
pendence in the second-order sense, which is important in geostatistics, refers only to that part of the variation in a dependent variable that remains “unexplained” by regression. This remark applies not only to fixed-effects modeling but also to random-effects modeling of spatial data because the random effects term is often specified using some form of (conditional) spatial autoregressive mode (Besag, York, and Mollie 1991).

Two further comparative remarks merit attention here. First, frequently the effects of population size are handled in regression through weighted least squares—down-weighting most those data values with the largest error variances in model fitting (e.g., those based on the smallest population denominators). In geostatistics, this down-weighting is introduced both into modeling the variogram and into the kriging weights. Second, geographers have paid particular attention to a form of spatial heterogeneity where properties vary, or appear to vary, with location, especially when observed over large regions. For example, the mean varies in different map segments (though not in the form of a trend). In addition, or alternately, the variance and covariance, or even the relationship between a dependent and a set of independent variables, might vary in different parts of a map. To quantify this heterogeneity, statistical tests for global- or map-wide spatial dependence, such as Moran’s $I$, have been complemented with local statistics, such as local indicators of spatial association that analyze spatially defined subsets of data using moving windows (Anselin 1995). Sometimes geographically weighted regression is used with local subsets of data (see Fotheringham, Brunsdon, and Charlton 2000). Some of the geostatistical methods previously discussed can be seen as equivalent developments to deal with spatial heterogeneity. First, SKlm uses known strata from a geology map, for example, to inform about a locally varying mean, and the class residuals are kriged and added to the strata means (Goovaerts 1997). Similarly, SKlm can use regression between primary and secondary variables to inform about a local mean. McBratney, Hart, and McGarry (1991) also acknowledge that the structure of the variogram can change spatially and thus have partitioned regions based on ancillary data and computed separate variograms for subregions within a study area. Walter et al. (2001) and Haas (1990) use local or moving window variograms and find improvements in estimates over area-wide variograms in situations where data display heterogeneity. Finally, when KED is employed, allowance is made for spatial variation in the linear relationships between variables or the regression coefficients (Goovaerts 1997).

The equivalent problem to spatial interpolation for regional data occurs when there are no data in some areas. Although interest may exist in trying to estimate these missing values, for either prediction or mapping purposes (see, e.g., Haining 2003, pp.154–74 for an overview of methods, including those based on the expectation-maximization algorithm, which has similarities to kriging), the focus usually is again on fitting a regression model. The concern is less with estimating the missing values per se (although estimates are generated) and more with estimating the parameters of a regression model when there are missing values in a database.
An overview of applications of geostatistics in geography

Table 2 classifies the results from the previously described Geobase search, providing one relevant reference per category. The table shows that 40% of geostatistical references found in “mainstream” geography journals refer to the use of kriging in relation to temperature, rainfall erosivity, various atmospheric and groundwater pollutants, soil properties, vegetation and species distribution patterns, and DEM creation. By contrast, kriging has been little used in human geography, an exception being De Cola’s (2002) mapping of Lyme disease.

Few articles in geography journals consider the importance of the variogram as an interpretive tool, for example, for quantifying “spheres of influence” or the proportion of variation that is spatially uncorrelated and correlated at the scale of sampling, or for identifying variation at different scales and directions as a first step in understanding processes. This oversight might indicate that few researchers in geography are as familiar as they could be with computing and modeling the variogram and with choosing an appropriate method of kriging. A possible cause of this problem is that certain software packages, such as Surfer, offer kriging as a method of interpolation with default settings and do not provide users with access to experimental variogram and modeling options. In ArcGIS’s geostatistical analyst, a variogram model is fitted to the variogram cloud, but determining visually whether a model is a sensible fit is difficult because of the inherent skewness in the distribution of squared differences (Cressie 1993, p. 41).

Table 2 shows that multivariate geostatistical methods, which incorporate secondary information into the interpolation process (e.g., SKlm, KED, and CK), have been used for estimating air temperature and bird diversity using the normalized difference vegetation index. Factorial kriging has been used to filter different scales of variation in remotely sensed imagery and plant and soil data and to explore scale-dependent correlations in health geography. Indicator kriging has rarely been used in geography, but examples include incorporating it into the classification of hyperspectral imagery and estimating pre-settlement vegetation patterns. Although indicator and disjunctive kriging are important in risk assessment, which is particularly important to geographers studying natural hazards, no applications of these two specific techniques were found. Furthermore, disjunctive kriging appears not to have been referred to at all in the geographical literature. Also, although adaptations of kriging, such as binomial CK and Poisson kriging, are ideal for examining animal count data, which are often highly skewed and involve small numbers, no examples were found in the biogeography literature.

Geostatistical change of support has received some attention in the geographical literature; it links with scale issues and the modifiable areal unit problem. In remote sensing and biogeography, geostatistics has been used to investigate the effects of change of support in sampling, but most investigations are in human geography and are relatively recent, applying techniques such as ATP kriging.
Table 2 Summary of a Geobase search for geostatistics\(^1\) articles appearing in journals with Geogra* in the title

<table>
<thead>
<tr>
<th>Geostatistical area of investigation</th>
<th>Biogeography</th>
<th>Climatology</th>
<th>Geomorphology</th>
<th>Hydrology</th>
<th>Remote sensing</th>
<th>Human geography</th>
<th>Total refs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(comparative interpolation studies)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multivariate geostatistics (KED, CK, SKlm)</td>
<td>7 refs.—Lin et al. (2008)</td>
<td>2 refs.—Hernandez Roberto (2001)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9 refs.</td>
</tr>
<tr>
<td>Factorial kriging</td>
<td>2 refs.—Rodgers and Oliver (2007)</td>
<td></td>
<td>1 ref.—Warr, Oliver, and White (2002)</td>
<td>1 ref.—Goovaerts, Jacquez, and Greiling (2005)</td>
<td>4 refs.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Binomial/Poisson kriging</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7 refs.—Goovaerts (2005)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^1\) Kriging is a statistical method for spatial prediction that is widely used in geostatistics.
<table>
<thead>
<tr>
<th>Geostatistical area of investigation</th>
<th>Biogeography</th>
<th>Climatology</th>
<th>Geomorphology</th>
<th>Hydrology</th>
<th>Remote sensing</th>
<th>Human geography</th>
<th>Total refs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Change of support issues/AtoP, AtoA kriging</td>
<td>1 ref.—Bellehumeur and Legendre (1997)</td>
<td></td>
<td></td>
<td></td>
<td>1 ref.—Mason, O’Conaill, and McKendrick (1994)</td>
<td>3 refs.—Goovaerts (2006b)</td>
<td>5 refs.</td>
</tr>
<tr>
<td>Sampling scheme design</td>
<td>1 ref.—Lin et al. (2008)</td>
<td>1 ref.—Finley et al. (2007)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2 refs.</td>
</tr>
<tr>
<td>Data compression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1 ref.—Atkinson, Curran, and Webster (1990)</td>
<td></td>
<td>1 ref.</td>
</tr>
<tr>
<td>Spatial weighting of classification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1 ref.—Goovaerts (2002a)</td>
<td></td>
<td>1 ref.</td>
</tr>
</tbody>
</table>
Geographers study both spatial and temporal variation, and spatio-temporal
variograms and space-time kriging can be used for predicting phenomena in these
domains. However, only two applications of space-time geostatistics were found:
one for determining the representativeness of air temperature records and one for
assessing the spatio-temporal variation of groundwater salt content.

Only two studies were identified where geostatistics has been applied to sam-
pling issues: one used geostatistics to determine the location of pollutants following
a bioterrorist attack, and the other to establish sampling schemes for mapping bird
diversity. Although the storage and classification of remotely sensed image data are
important in geography, only one article was found that uses geostatistics for data
compression and only one that uses spatial weighting by the variogram to improve
classification contiguity.

Most applications of geostatistical simulation are in geology or geomorphology
for risk assessment and uncertainty, but they are also used in remote sensing and
health geography. For the latter, geostatistical simulation is used to assess the un-
certainty associated with rare disease clusters.

Several theoretical articles were found that introduce new geostatistical con-
cepts or techniques or that provide reviews. These are the second most abundant
type of geostatistics article; however, the authors are not geographers but rather
prolific publishers of geostatistical applications in the applied sciences.

Concluding remarks
Geostatistical approaches to the analysis of spatial data have been underexploited
in geographical research, particularly in human geography. Two main reasons ap-
pear to account for this. First, traditional geostatistical techniques are applicable to
attributes with a continuous spatial index (Ripley 1981; Cressie 1993, pp. 8–9),
making these techniques less relevant to many problems in human geography,
where data frequently relate to areas. Second, geostatistics is often perceived as
being of use only for spatial interpolation or kriging, including mapping and sample
design when obtaining primary data. Other reasons sometimes cited for the un-
deruse of geostatistical approaches include lack of time allotted to teaching geo-
statistics in spatial data analysis courses in geography departments compared with
other forms of spatial analysis and lack of instruction about the appropriate use of
available software.

Geostatistics embraces a broad range of tools and modeling techniques that
can be applied to many spatial problems, including prediction, determination of
the scale of spatial variation, design of sampling for primary data collection,
smoothing of noisy maps, region identification, multivariate analysis, probability
mapping, and change of support. It has a research literature that includes many
disciplines. Most important, however, geostatistics provides the spatial analyst with
a statistically rigorous model of how properties vary in space (RVT) that recognizes
the different components of variation (from locally erratic to spatially structured
components) and furnishes permissible models for those components of variation, as well as procedures for fitting them to real data.

Today, geostatistical methods can be applied to regional data, although their use with this type of data does present challenges. Proximity measures are limited to the distance between centroids, which cannot be defined uniquely when spatial units have nonzero areas. Data based on rates and proportions are aggregated, and the denominators often vary among spatial units reflecting differences in population size. Some variation in data can be a consequence of these underlying differences in sample support (Krivoruchko, Gotway, and Zhigimont 2003), and these issues need to be taken into account when dealing with regional data.

More user-friendly software is now available. A list of free and commercial software for geostatistical analysis is given at http://www.ai-geostats.org/index.php?id=107, although it is by no means exhaustive, failing to include packages such as GenStat (developed at Rothamsted Research), S-Plus, and Terraseer’s Space Time Information System, which currently has a beta package that can accommodate different types of geographic data and has capabilities for ATA, ATP, and Poisson kriging. In addition, many computer programs are available from articles published in Computers & Geosciences, as well as authors making their code freely available on their personal Web sites. Deutsch and Journel (1998) provide a comprehensive set of Fortran programs for many geostatistical techniques in Geostatistical Software Library and User’s Guide, and most of these programs have been adapted for a Windows environment in the freely available Stanford Geostatistical Modeling Software (Remy, Boucher, and Wu 2009). However, the former has no way of modeling a variogram once it has been computed without using other software, and the latter employs visual fitting. Gstat is a freely available command line software package that is favored by many and includes the capability for variogram model fitting both numerically and visually (Pebesma and Wesselting 1998). With new developments in methodology and software, exciting times lie ahead for geographers exploring the benefits of geostatistics for solving geographic problems.

Notes

1 This maps into what Tobler (1970) describes as “The First Law of Geography”: “everything is related to everything else, but near things are more related than distant things.” Banerjee, Carlin, and Gelfand (2004, p. 39) refer to “The First Law of Geostatistics,” which describes spatial structure in similar but more formal terms. This observation has statistical roots that go back to the early twentieth century (see Student 1911).

2 The strategy used to search for geostatistics articles was as follows: Year range: 1990–June 2008; Document type: all; (Geobase) Subject/Title/Abstract includes: krig* OR variogram OR geostat*; (ISI) Title includes: krig* OR variogram OR geostat*. Geostatistics articles within geography were identified as those that contained the three preceding keywords in the subject/title/abstract and also were found in a journal with geogra* in the title. We recognize that the search terms are not exhaustive and will not identify all geostatistics
articles in the academic literature in general or within geography, but some basic patterns may be observed that we think are informative. The terms used limit the number of articles identified, but few, if any, nongeostatistics ones would be identified with these terms. Simulation was not included because it would cause potential confusion. For more information arising from the bibliometric search and for more references obtained from the literature review in this article, interested readers should visit http://www.geog.cam.ac.uk/people/haining/ or contact the corresponding author.

3 See note 2. We do not claim to have identified all areas of application. However, for more details about references, go to http://www.geog.cam.ac.uk/people/haining/ or contact the corresponding author.

References


Mediterranean Sea from Sparse Count Data and Heterogeneous Observation Efforts.”


