Transactions in GIS, 2006, 10(4): 615–632

Research Article

The Role of External Variables and GIS Databases in Geostatistical Analysis

Edzer J Pebesma
Department of Physical Geography
Utrecht University

Abstract
Although many geostatistical studies only study a measured attribute in relation to its spatial coordinates, this paper argues that other layers in the GIS database can be of additional use for spatial prediction purposes. They may enter the prediction equations as predictors in a regression model, or as correlated measurements. In an example we will show how this is done for predicting PCB138, a sediment pollution variable, over the North Sea floor. Issues of exploratory data analysis, required sample size, sample configuration, local versus global neighbourhoods, non-stationarity, non-linear transformations, change of support and conditional simulation will be discussed in the light of this example.

1 Introduction

Whereas traditional statistics is usually concerned with questions like *how much?*, or *how* are x and y related? the main focus of spatial statistics is the *where* question: *where* do certain features occur, *where* is the pollution largest, *where* do certain relations change. In his book on spatial statistics, Cressie (1993) distinguishes three types of spatial data: point pattern data, lattice data and geostatistical data. Point pattern data record locations of incidences of a certain kind. Examples include the locations of trees in a forest or the exact locations of accidents or crimes in some area over a given period of time. Lattice data accrue when the data are collected for larger regions such as administrative areas that have nothing specific in common with the features collected; in this case the exact feature locations are lost. Geostatistical analysis is concerned with the analysis of geostatistical data, and the typical question is what we would have measured had we measured elsewhere, or even everywhere. Given the data, geostatistical

Address for correspondence: Edzer J. Pebesma, Department of Physical Geography, Geosciences Faculty, Utrecht University, P.O. Box 80.115, 3508 TC Utrecht, The Netherlands. E-mail: e.pebesma@geo.uu.nl

methods are used to make probabilistic statements about the non-measured quantities of interest. Excellent overviews of the origin and history of geostatistics are given by Ripley (1981) and Cressie (1993).

Geostatistical data can be continuous variables such as topographic altitude, discrete variables such as land use or soil type, or even densities of point pattern data such as number of diamonds in a rock sample or density of sea birds over a fixed area strip transect count. In the latter cases, the underlying process is a point pattern, but the measurement technique integrates this over a certain measurement volume, intentionally or necessarily leading to geostatistical data. The measurement area or volume (or even time), also called the measurement *support*, may be much smaller than the support for which estimates are required. The reasons for the need of estimates integrated over larger areas can be diverse:

- In a mining context, core samples from a bore hole are much smaller than the smallest unit that can be mined.
- In an environmental context, policy makers may be interested in aggregated, regional estimates from certain smaller or larger regions.
- From a statistical context, estimates for small or moderately large sub-regions can often be estimated with a much higher accuracy than values for regions the size of the measurements.

A large branch of geostatistics has been developed to deal with measurements and spatial estimates having a different support.

Many books have been written on geostatistics. Good introductory texts include the very accessible Isaaks and Srivastava (1989) and the much more comprehensive Chilès and Delfiner (1999). General overviews from a GIS perspective include Burrough and McDonnell (1998) and Heuvelink (1998). Other books are directed towards certain applications, such as mining (Journel and Huijbregts 1983), soil science (Goovaerts 1997), or hydrologeology (Kitanidis 1997). More mathematical statistical books are Ripley (1981), Christensen (1991), Cressie (1993), and Wackernagel (1998).

Although many texts on geostatistical analysis emphasize the theory of spatial prediction, this paper emphasizes the application of geostatistics. We will show how it is done by analyzing a multi-temporal data set on sea floor surface sediment pollution in the Dutch part of the North Sea (Laane et al. 1999), which was provided by the Dutch National Institute for Coastal and Marine Management (RIKZ). Data, software and a script with all the analysis steps are freely available through links on the author's website (see http://www.geog.uu.nl/~pebesma/ga/ for additional details). The data analysis will start with an exploratory data analysis, but first we will introduce the general linear geostatistical model.

2 The Linear Geostatistical Model

In linear geostatistics, we assume that the variability of the n observations $Z(s)$, with s denoting spatial location, is the sum of a trend $m(s)$ and a residual $e(s)$:

$$Z(s_i) = m(s_i) + e(s_i), \qquad i = 1, \ldots, n \tag{1}$$

The trend $m(s)$ is either an (unknown) constant m, or a deterministic, linear function of $p + 1$ unknown constants β_i and known covariates, $f_i(s)$:

$$m(s_i) = \beta_0 + \beta_1 f_1(s_i) + \ldots + \beta_p f_p(s_i) \quad (2)$$

where $f_0(s_i) \equiv 1$. Note that covariates may be both continuous variables ("regressors"), or categorical variables, in which case they are transformed into a series of (0/1 encoded) dummy variables.

The residual is a zero mean random variable with a stationary covariance, i.e. a covariance that depends on separation vectors $s_i - s_j$ only:

$$Cov(e(s_i), e(s_j)) = C(s_i - s_j) \quad (3)$$

which, in case of isotropy (direction independence, opposite anisotropy) further reduces to $C(s_i - s_j) = C(|\, s_i - s_j\,|) = C(h)$ with h the separation *distance*.

Instead of using covariances, geostatistics looks at semivariances $\gamma(h) = 0.5E(Z(s) - Z(s+h))^2$. In most cases, semivariances are related to covariances by $\gamma(h) = C(0) - C(h)$ (Cressie 1993). Semivariances are estimated from data point pairs by averaging over the N_i point pairs with spatial separation distance in the interval $\tilde{h}_k = [h_k, h_{k+1}]$:

$$\hat{\gamma}(\tilde{h}_k) = \frac{1}{2N_k} \sum_{j=1}^{N_k} (Z(s_j) - Z(s_j + h))^2, \quad k = 0, \ldots, q \quad h \in \tilde{h}_k,\ h_0 = 0 \quad (4)$$

for several consecutive distance intervals, usually ranging up to one third of the area size.

For predicting $Z(s_0)$ at a location s_0 where no measurements of Z are available (also called *kriging*, Table 1), we need to know $f(s_0)$. Therefore, the $f(s)$ need to be complete coverages in the GIS database. Variables that are always present are the spatial $\{x, y\}$ coordinates of s_0, but these may not carry very useful information with regard to explaining variability in Z. Better variables that may be "cheap" to obtain are distances to source locations when the variable Z is subject to dispersion processes (e.g. point source pollution, or plant or animal abundances).

The challenge of geostatistical analysis is to apply this mode to our data such that: (1) all information available is used optimally; (2) the agreement of model and data is

Table 1 Prediction and prediction standard error equations for the linear geostatistical model (equation 1)

Let $(1, f_1(s_i), \ldots, f_p(s_i))$ be the i-th row in the $n \times (p+1)$ matrix F with predictors or covariates, and let $(1 + f_1(s_0), \ldots, f_p(s_0))$ be the $1 \times (p+1)$ vector $f(s_0)$. Then, given the covariance matrix V of $e(s)$, the best linear unbiased predictor (or universal kriging predictor) of $Z(s_0)$ is:

$$\hat{Z}(s_0) = f(s_0)\hat{\beta} + v' V^{-1}(Z(s) - F\hat{\beta})$$

with $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)' = (F'V^{-1}F)^{-1}F'V^{-1}Z(s)$ is the generalized least squares estimate of the trend coefficients, where F' denotes the transpose of F, and with $v = (Cov(Z(s_0), Z(s_1), \ldots, Cov(Z(s_0), Z(s_n)))'$ where $Cov(\cdot,\cdot)$ denotes covariance, defined as $Cov(Z(s_1), Z(s_2)) = E[(Z(s_1) - E(Z(s_1)))(Z(s_2) - E(Z(s_2)))]$; where $E(\cdot)$ denotes expectation. The corresponding prediction error variance is:

$$\sigma^2(s_0) = \sigma_0^2 - v'V^{-1}v + (f(s_0) - v'V^{-1}F)(F'V^{-1}F)^{-1}(f(s_0) - v'V^{-1}F)'$$

where $\sigma_0^2 = Var(Z(s_0))$. When trend coefficients are known, β is substituted for $\hat{\beta}$ in the first equation, and the third term of the right hand side of the second equation disappears.

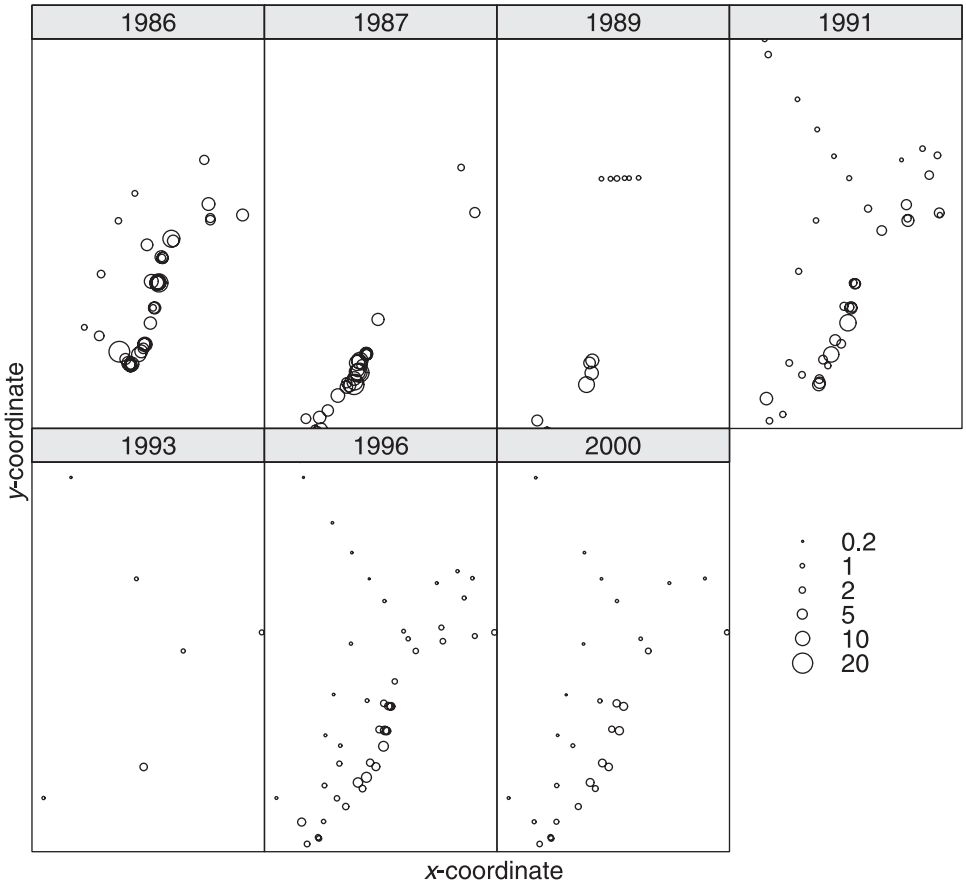

Figure 1 Maps with PCB138 measurements (μg/kg dry matter) for each monitoring year. The unsampled white area in the south-east corner of the maps approximates the outer coast line of the Netherlands. In the area shown, the *x*-coordinates range from 464,000 m to 739,000 m, and the *y*-coordinates range from 5,696,500 m to 6,131,500 m, projection UTM31

satisfactory; and (3) the model yields adequate predictions. Obviously, success depends on the quality of both measurements $Z(s)$ and covariates $f(s)$, and the variability present in the data.

3 Exploratory Data Analysis

The sediment data set is collected by the RIKZ during a monitoring program, aimed at describing spatial and temporal variability in sea floor sediment. Variables that were measured comprise heavy metals and organic contaminants, comprising a number of polychlorinated biphenyls (PCBs). The measurements for one such PCB, PCB138 are shown in Figure 1. Table 2 shows summary statistics for PCB138 concentration, for

Table 2 PCB138 (μg/kg dry matter) summary statistics; years marked with an * are the regular monitoring years, other years result from additional sampling programs

year	1986*	1987	1989	1991*	1993	1996*	2000*	all
mean	7.29	8.39	4.08	3.70	1.03	1.58	1.27	4.20
median	6.90	7.50	2.65	3.05	0.775	1.40	0.90	2.85
max	21.1	19.7	12.3	13.1	2.7	4.9	3.3	21.1
min	1.60	2.10	1.00	0.70	0.25	0.20	0.20	0.2
n	45	29	14	42	6	49	31	216

each year with measurements. The main monitoring program aims at measuring every five years, and this was done in 1986, 1991, 1996 and 2000. For some years in between, samples of smaller size were collected. The samples were collected using box core samplers, and only the fraction smaller than 63 μ was analyzed for contaminants.

The spatial pattern of PCB138 measurements, shown for each year in Figure 1, reveals that PCB138 concentration tends to decrease when moving from the coast. The summary statistics of Table 2 indicate that the PCB138 concentration decreases over time. However, this tendency may be partly attributed to the increase of the fraction of off-shore sampling points over time (Figure 1): the temporal variability of the spatial sampling scheme does not make such an analysis trivial.

How concentrations depend on water depth is shown in Figure 2 where PCB138 was graphed on a log-scale to make the relationship shown closer to linear. This relation does not come as a surprise: most of the polluted sediment originates from the rivers (Rhine, Meuse, Schelde) that enter the North Sea, and get transported further by the North Sea along-coast flow in a north-eastern direction.

In the exploratory stage we look for data irregularities, possible outliers, suitable data transforms, and explore relations between measured variables and other variables (such as sea water depth) that are available and may help predict the measured variable over the spatial domain. In our case, water depth is not *the* variable that causes PCB138 to have certain values; behind this variability a complete transport process with dynamic sources, convection and dispersion, and complex sea water flow patterns is hidden. In absence of knowledge of this process, water depth does seem to be a good proxy to much of this process, and it explains a fair proportion of the variability. Figure 2 does not give evidence to remove outliers.

The simple approach after fitting the trend (Figure 2) would be to predict log-PCB138 at unobserved locations as a function of year and depth:

$$\hat{Z}(s_0) = \hat{\alpha}_{year} + \hat{\beta}_1 f_1(s_0) \qquad (5)$$

with $\hat{\alpha}_{year}$ the year-dependent intercept and $f_1(s_0)$ the depth at location s_0. This regression model explains 77% of the variability in the data. It is assumed that the slope β_1 is constant over time; year-dependent slopes did not improve the linear regression model significantly. The maps that result from applying Equation (5) spatially are depth maps with a modified legend. The data however may well carry more information than only this trend, and one way to find out is considering the residuals spatial correlation.

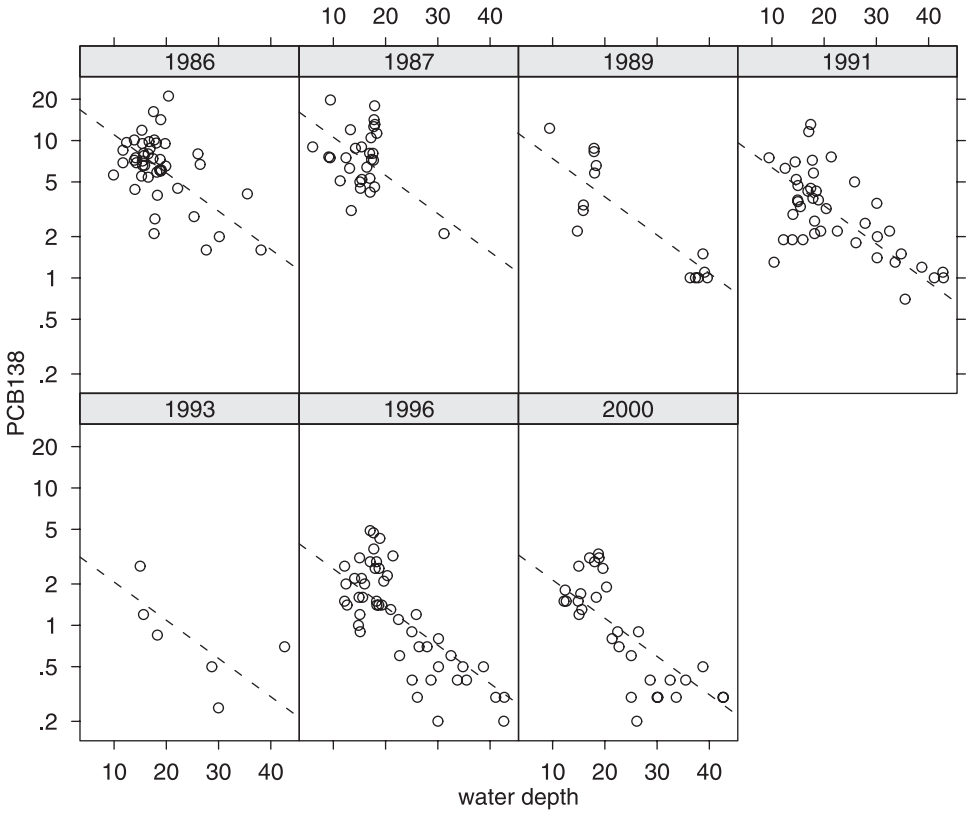

Figure 2 PCB138 concentrations (µg/kg dry matter) as a function of water depth (m); the dashed lines have a constant slope and a year-dependent intercept; R^2 for the regression model (equation 5) is 0.77

4 Variography

The modeling of spatial correlation lies at the heart of geostatistics. Usually, spatial correlation is modeled by calculating sample variograms (equation 4) and fitting parametric models to them. Another approach is fitting of variogram parameters direct to the (quadratic form of the) data by maximum likelihood or restricted maximum likelihood (Kitanidis 1983, Stein 1999). Here we will explore and model sample variograms, calculated from estimated residuals obtained by ordinary least squares. Although estimated residuals are not equal to the "true" (but unknown) residuals, they are suitable for modeling residual spatial correlation (Kitanidis 1993).

Figure 3 shows the sample variograms for log-PCB138 for each year. In this figure many of the variograms show an erratic structure (1987, 1989, 1991, 1993, and 2000), others do slightly less so. Erratic variograms may occur for various reasons, including small sample sizes, very skew data distributions, outliers contaminating the sample data, sample locations that are highly clustered, non-stationary situations or combinations of the above. We will consider all these issues.

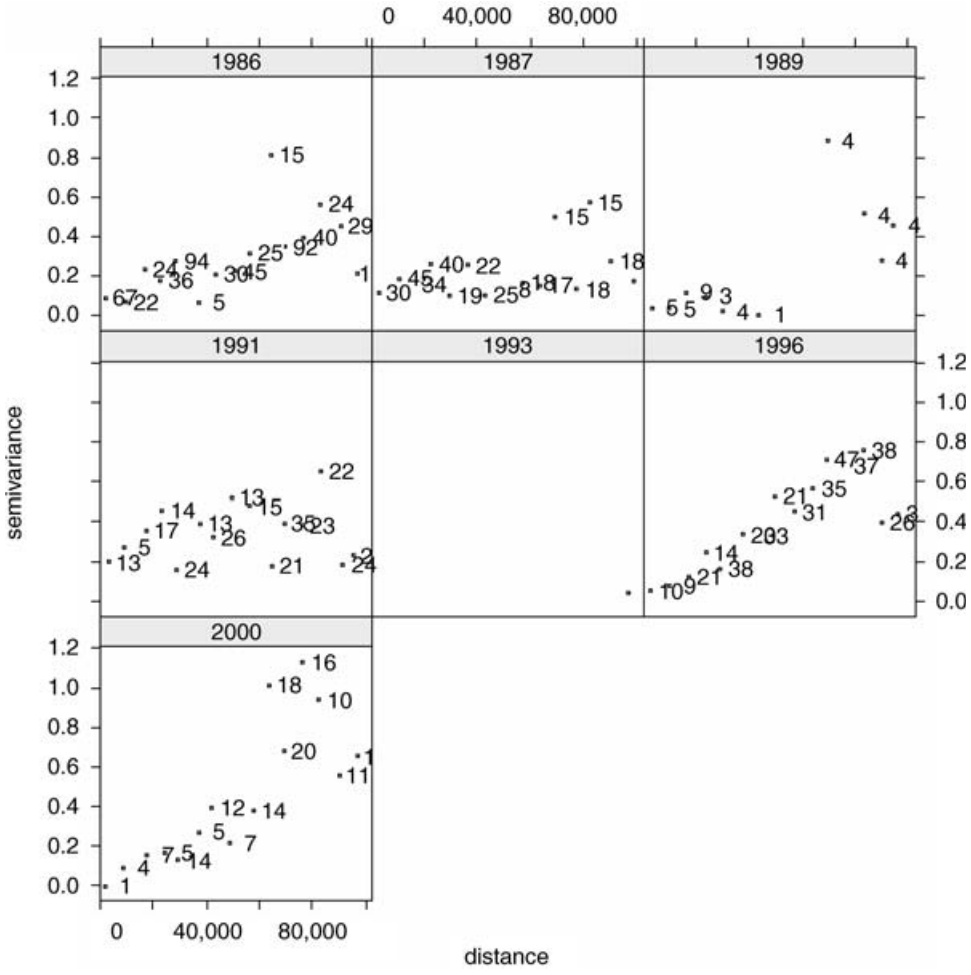

Figure 3 Sample variograms per year for long-PCB138; numbers denote the number of pair points N_i (over which a variogram was averaged)

4.1 Sample Size

Given n observations, we can form $n(n-1)/2$ point pairs, so from 45 observations we can form 990 pairs, from 100 we can form 4,950. This seems a lot, but any two point pairs that share a common point are strongly correlated. As becomes clear from comparing sample sizes in Table 2 with sample variograms in Figure 3, larger samples give usually better (less erratic) sample variograms. It is impossible to say how large a sample should be ("at least 100!"), because distribution and spatial pattern play a role. In Figure 3, the 1986 and 1996 variograms do suggest that a sample size of 50 may suggest the spatial correlation reasonably well.

The point pairs formed are to be divided over distance classes $\tilde{h}_k$ in equation (4). The key feature of a variogram is its behaviors near the origin, when $h \to 0$. It is tempting to make the distance classes $\tilde{h}_k$ (equation 4) narrow, especially near zero, but

this also decreases the number N_i, which in turn makes the variogram more erratic. A trade-off has to be found between regularity and detail at small distances. We will show this in the following example.

When sample sizes are small, there is little hope of ever getting good estimates of spatial correlation. For the sediment data set, however, we do have several years of data available, and we could try to combine this information to get a better idea of the spatial correlation in the residuals. One approach would be simply merge all residuals, and calculate their sample variogram. This is shown in the first panel of Figure 4 ("ignoring year"). This approach assumes a temporal persistence of the actual residual spatial *pattern*, which is rather unlikely for phenomena in a dynamic environment. Another approach would be to average ("pool") sample semivariances over different years to a single "pooled" sample variogram. This latter approach only assumes persistence of the *nature* of the residual spatial variability, which is a much weaker assumption that seems in accordance with Figure 3. The pooled variogram includes only point pairs that are measured in the same year. The second panel in Figure 4 shows this sample variogram. Although the first semivariance estimate now has 177 points pairs (compared to 829 point pairs for the non-pooled semivariance estimate) it does reveal a much stronger spatial correlation, confirming our expectations about temporal dynamics in the residual pattern. When we zoom in at the short distances, the first semivariance estimate with 177 point pairs, $\hat{\gamma}(4,178) = 0.13$, splits and reveals in the third panel of Figure 4 even more spatial correlation: with 36 point pairs the semivariance decreases to $\hat{\gamma}(590) = 0.08$. The final panel shows a fitted exponential model. This model was fitted to the last sample variogram using weighted least squares fit with weights proportional to N_i, while the nugget effect (the value where the model reaches $h = 0$) was fixed at 0.08.

Data sets too small for reliable modeling of spatial correlations occur frequently. Other general strategies that deal with obtaining spatial correlation for small data sets are:

- When variability within several sub-regions is considered, variograms from different sub-regions may be pooled, in which case it may be useful to normalize variances before pooling, thus assuming equal spatial autocorrelation instead of semivariances.
- Variograms can be obtained from larger, similar areas (or time periods for which many more measurements are available) as the class from which a variogram is borrowed, possibly after normalizing variances (Pebesma and De Kwaadsteniet 1997).

One of the problematic issues of applying kriging (Table 1) is that it assumes that all covariances (or semivariances) are known, and not subject to uncertainty. A more complete approach would be to take uncertainty with respect to variogram model coefficients and/or parameterizations into account, e.g. as in the Baysian approach of Diggle et al. (1998).

4.2 Skewed Distribution and Outliers

A nice example of the influence of a single observation on the sample variogram is present in the Cadmium observations of the sea floor surface sediment data set, as shown in Figure 5. A pooled within-year sample variogram for (untransformed) cadmium observations is shown in Figure 5a. It shows that there is spatial correlation, as semivariances increase with distance, but the variogram is very erratic. To investigate

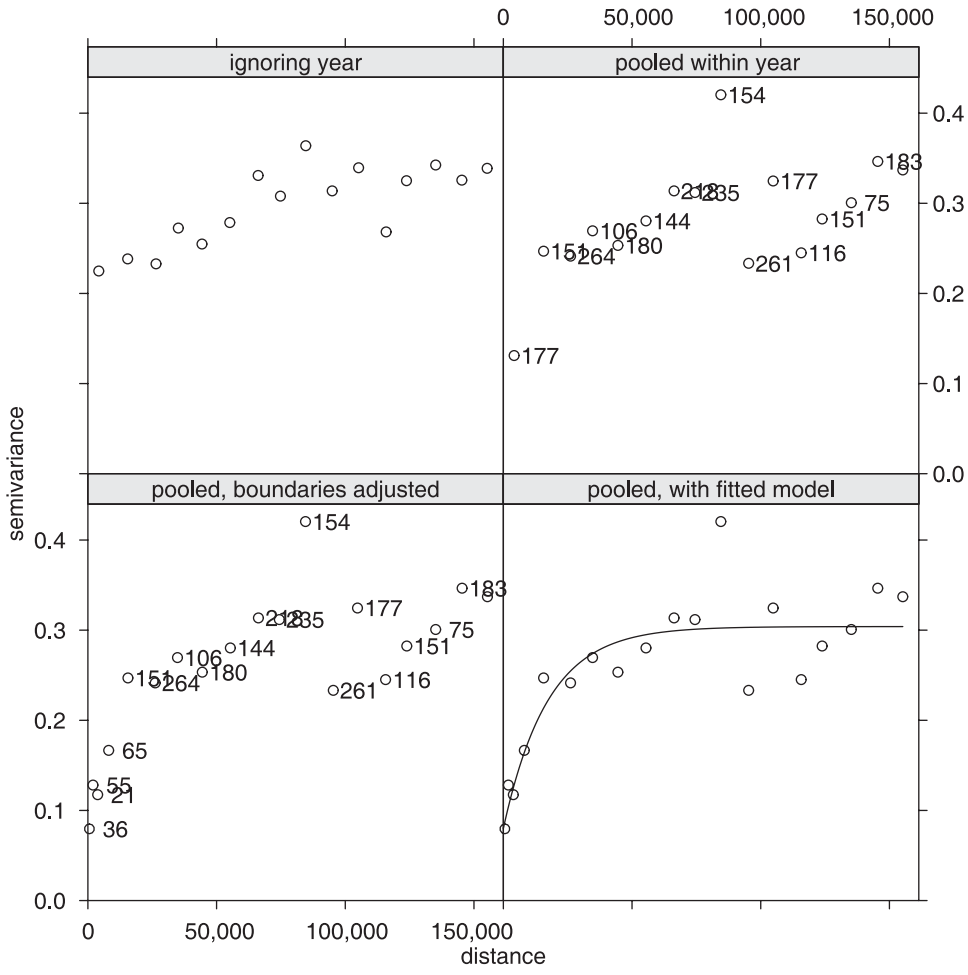

Figure 4 Sample variograms for log-PCB138 residuals. Upper left: ignoring years, simply merging residuals; upper right: including only point pairs from the same year; lower left: as upper right, zooming in at small distances; lower right: as lower left, but with fitted model. Numbers reflects the number of point pairs, N_i. The fitted model is $\gamma(h) = 0.08\delta(h) + 0.224(1 - \exp(-h/17{,}247))$ with $\delta(h) = 0$ if $h = 0$ and $\delta(h) = 1$ if $h > 0$

whether there are single observations that cause this behavior, we plotted the sample variogram cloud (Figure 5b), which shows for each point pair $0.5(Z(s_i) - Z(s_j))^2$ plotted against separation distance $h = | s_i - s_j |$. Note that the plot in Figure 5b shows exactly the same information present in Figure 5a, but omits the averaging over distance classes $\bar{h}_k$. In the plot reproduced in Figure 5b we can identify individual point pairs, or in this case digitize the whole area with semivariances above 20, as this group seems anomalous. The plot in Figure 5c shows the spatial locations of these selected pairs, and the star shape indicates that they all share a single observation: the maximum value that seems to be outlying in the normal probability plot of Figure 5d. The strong variability of semivariance estimates in Figure 5a seems to be caused by the fact that high estimates

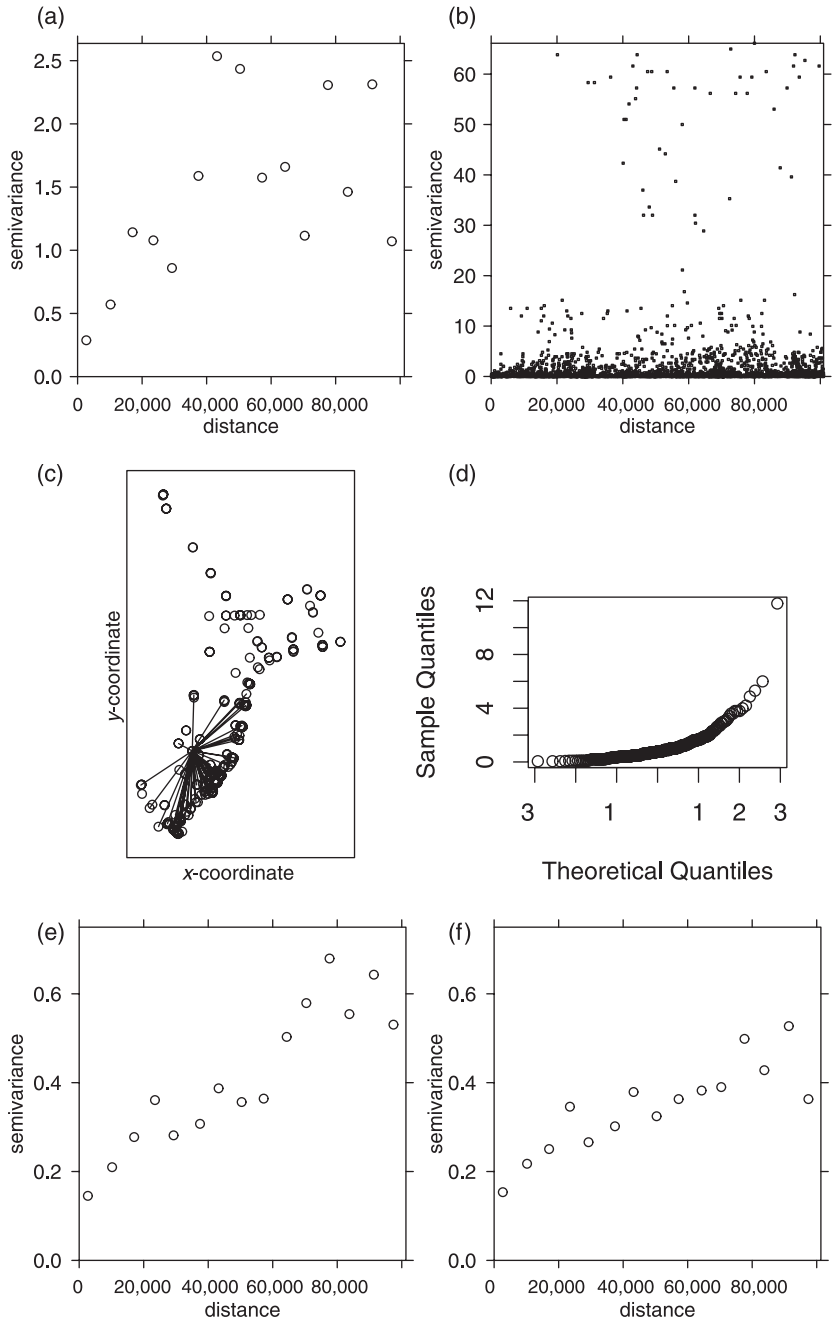

Figure 5 Cadmium in the sea floor surface sediment data set: (a) cadmium variogram (pooled, within-year); (b) sample variogram cloud for (a); (c) circles: data locations, lines: point pairs that have a semivariance in (b) larger than 20; (d) normal probability plot for cadmium; (e) pooled within-year sample variogram for log cadmium; (f) pooled within-year sample variogram for log cadmium after trend removal (including only a linear-depth effect and a year-dependent intercept)

include many point pairs with this extreme, and the low estimates do less so: a single observation dominates the sample variogram in a disturbing way. On the log-scale, the influence of this outlying point is much less and the sample variogram of Figure 5e for log cadmium, and Figure 5f for log cadmium after a linear effect for depth was subtracted, reveal the spatial correlation much better. Comparison of Figures 5e and f further show that removal of a slow, gradual trend (as depth is) does influence the long-distance residual variability much more than the short-distance behavior.

A special case of the distribution problem occurs with binary, e.g. 0/1 encoded, variables: if the proportion of ones is very close to zero (or one), then even very large samples may prove difficult for modeling the spatial correlation. Continuous variables that have very few large outliers behave the same as linearly scaled 0/1 data with very few non-zero values. Continuous variables can always be transformed into 0/1 variables, e.g. indicating whether concentrations lie above a certain threshold. Applying such as transformation before spatial analysis may, however, remove valuable information.

Whether an extreme is to be considered an outlier, and thereby a candidate for removal before further analysis, remains a difficult issue. In the above example of cadmium concentrations, a normal probability plot of log cadmium (not shown) did indicate the largest value to be not outlying but perfectly conforming to a log-normal distribution. In such a case, working on the log-scale seems the obvious approach.

4.3 Sample Configuration

Irrespective the spatial structure in the variable studied, the success of inferring this structure from sample data is determined by the spatial configuration of the sample locations. If the samples are spread over a regular grid, then no information is available about distances shorter than the grid spacing. If the aim is interpolation between grid nodes, then the absence of this information may be disastrous, because the correlation at short distances is the dominant factor when interpolating near observation locations. For reliable modeling of variograms, the availability of observations taken at short distances from each other (or even replicates taken at the same location) is of utmost importance.

Variograms will reveal whether measurements at small distance are present in the data set. However, they will not reveal *where* they occur. If the sample variogram estimate at the shortest distance has 36 point pairs (Figure 4, lower left panel) then we still do not know whether these are obtained from nine points in the cluster ($9 \times 8/2 = 36$) or from 72 points pairwise ($72/2 = 36$) close together but otherwise well spread over the study area. The nine points cluster is of course not to be preferred, both from the vague notion of being not geographically "representative" (although the cluster may have been located at random!) as well as from the statistical side: 72 points carry much more information than nine do.

Seldom is the estimation and modeling of variograms a goal in itself, usually we want to estimate the measured value over the study area as a whole. Given knowledge of the variogram, regular sampling is preferred for interpolation. If secondary data (covariates, like depth in our sediment data set) are available, and relations with covariates need to be explored, then a spread of the data over the range of the covariates is also advisable. Without knowledge of the variogram, in addition a considerable effort should be made to ensure that the short-distance behavior of the variogram can be inferred from the data.

4.4 Non-stationary Behavior

The assumption underlying most of linear geostatistics, called the *intrinsic hypothesis*,

$$E(Z(s)) = m(s), \;\; E(Z(s) - Z(s+h))^2 = 2\gamma(h)$$

basically assumes that (1) the trend is modeled correctly, and (2) residual variability and spatial correlation is independent of spatial *location*. This is a model, and as with all models, rejecting this model is only a matter of collecting sufficient evidence (i.e. data) against it. Rejecting the model is therefore in itself of little value. Only if reality differs clearly and substantially from the model, refinement is justified. Refinements may include modifications of the trend model, choosing a non-linear transformation of $Z(s)$ (e.g. log-transformation, see Figure 5), splitting the area in a number of sub-domains and modeling trends and/or variograms for each of the strata separately, or modeling direction-dependent (anisotropic) variograms.

5 Spatial Prediction and Conditional Simulation

Given observed data and the variogram, the prediction equations in Table 1 can be applied when residual covariances $C(h)$ are replaced by residual generalized covariances $C - \gamma(h)$, irrespective of the choice of a constant C (Kitanidis 1993).

For log-PCB138, results are shown in Figure 6. Using the trend model shown in Figure 2 and the residual variogram of Figure 4, we could predict log-PCB138 for each year, for each location. The maps show a pattern similar to the spatial pattern for sea floor depth, and in areas where no data are nearby the predictions heavily reply on the correctness of the assumption that the linear relationships shown in Figure 2 can be extrapolated to non-sampled locations. In areas where data are nearby (Figure 1), the maps deviate from this pattern. The series of maps also show the gradual decrease in PCB138 levels over the years, and the persistence of the dependence on depth. As can be seen by comparing the data locations (Figure 1) with the maps of predicted value (Figure 6), large areas where we predict have none or very sparse data. The extent to which we can believe the predictions is linked to the extent we may believe that the fitted trend in Figure 2 is a good model for the whole area. The predictions are however subject to prediction errors, the variance of which is quantified in the second of the two equations shown in Table 1.

The prediction (or kriging) standard error $\sigma(s_0)$ is a measure of the quality of the prediction. Showing prediction standard error maps is of little value, as their main use is in relation to the predicted value. Approximate 95% prediction intervals can be formed by back-transforming $[\hat{Z}(s_0) - 2\sigma(s_0), \hat{Z}(s_0) + 2\sigma(s_0)]$, and Pebesma and De Kwaadsteniet (1997) give two simple ways of presenting maps with prediction intervals that avoid the display of maps with only a single side of a prediction interval.

Another approach to visualize prediction errors is by conditional simulation. This technique generates Monte Carlo realizations of the complete field $Z(s)$, and each realization: (1) "follows" the trend; (2) has data values at data locations; and (3) has the variability and spatial correlation structure of $Z(s)$. Point (3) may seem obvious, but it should be noted that maps with predicted (i.e. expected $\hat{Z}(s)$) values are by definition much smoother than reality, $Z(s)$. Figure 7 shows eight conditional simulations for 5 km × 5 km block median values of PCB138 for 1991. The spatial variability *within* a

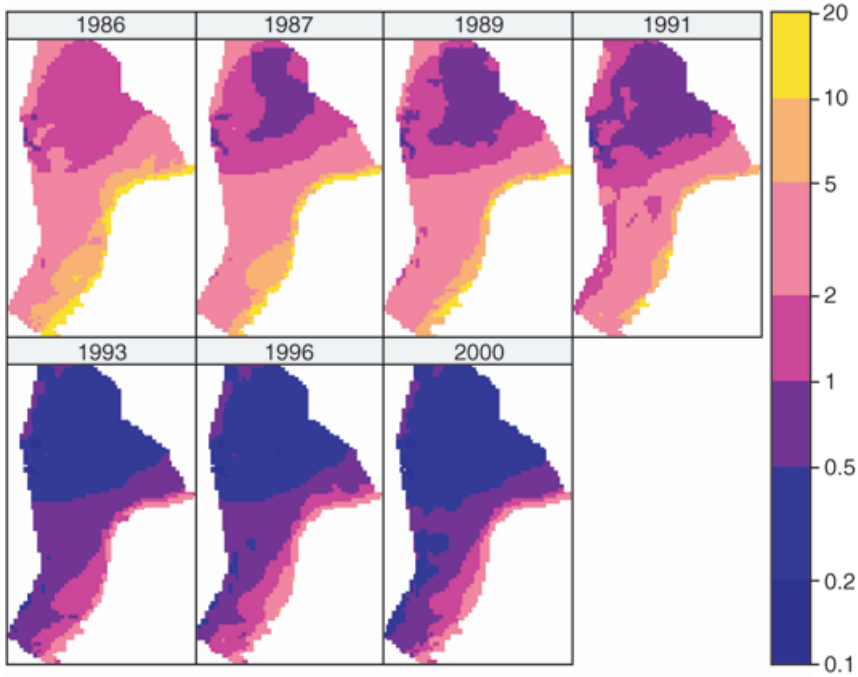

Figure 6 PCB138 predictions: predicted block median values from 5 km × 5 km square blocks; trend model according to Figure 2, residual spatial correlation according to Figure 4. This figure appears in colour in the electronic version of this article and in the plate section at the back of the printed journal

single realization has the same spatial variability (i.e. it has the same variogram) as the real block median values. Differences *between* realizations reflect our uncertainty, inherent in predicting the values of Figure 6. This uncertainty results from the fact that we only have a very limited sample of observed concentrations (Figure 1), and try to predict values for the whole area. The simulations shown here only address residual variability, and incorrectly assume that the trend coefficients are known.

Suppose that the 10 μg/kg contour is an important criterion for this variable. Looking at Figure 6, for 1991 only a few locations close to the coast (SE border) exceed this level. Based on the conditional simulations (Figure 7), one could argue that many more locations may exceed this level, and also that these locations may be much more offshore than Figure 6 suggested.

When we simulate a large number of realizations, their mean will equal the universal kriging prediction, and their standard deviation will equal the universal kriging prediction standard error, so we do not need simulations for that (an often overlooked triviality!). However, for approximating the block *averages* instead of the block median in Figure 6, we could use many point support simulations on the log-scale of a fine grid, take their exponent, and calculate block averages for each of the realizations, and calculate their mean over all realizations. Although it should give the same results, this approach avoids the rather cumbersome log-normal kriging equations given, for example, in Journel and

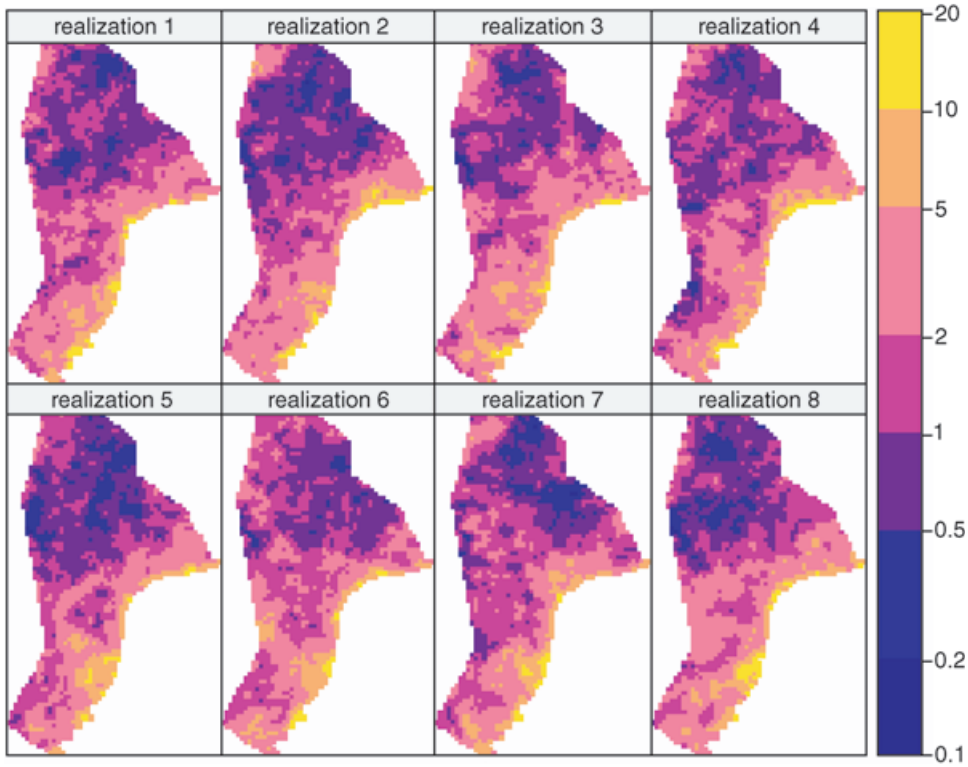

Figure 7 Eight distinct PCB138 conditional simulation realizations for 1991; simulated block median value for 5 km x 5 km square blocks, trend model according to Figure 2, residual spatial correlation according to Figure 4. This figure appears in colour in the electronic version of this article and in the plate section at the back of the printed journal

Huijbrechts (1983). Other, non-linear measures can only be obtained by simulations; an example is a probabilistic estimate (e.g. an interval estimate) of the areal fraction for which concentrations are above (or below) a certain threshold (Pebesma and Heuvelink 1999), or more complex functions of the variable, e.g. generated by a process model that simulates transport, flow, dispersion or diffusion, etc.

5.1 Change of Support

As the predictions for 5 km $\times$ 5 km block average values underlying Figure 6 were calculated on the log-scale, we cannot obtain block *average* values on the original scale by simply taking the exponent. The values obtained by taking the exponent (which was done here) can, however, be interpreted as block median values, i.e. the mid value of all point support values within each block, when, on the log-scale, point support values inside a block are symmetrically distributed. In other cases, it can be interpreted as the block geometric mean (Pebesma and De Kwaadsteniet 1997).

5.2 Kriging in Local Neighborhoods

One of the features of the general linear geostatistical model (Equation 1) is that the mean m or regression coefficients β_i are spatially constant. A weaker version of this assumption is that these coefficients are only constant in a local neighborhood around the prediction location. This neighborhood should be sufficiently large to estimate the coefficients accurately. This weaker assumption, underlying *local* kriging, may suit the data better than the global assumption.

Another (and, historically the original) reason for working with local neighborhoods is that when data are abundant (say, $n >> 1{,}000$), it may became very cumbersome or even impossible to compute V^{-1} in the two equations reproduced in Table 1. In such cases, restricting attention to a local neighborhood can be many orders of magnitude faster, provided that an efficient neighborhood search algorithm is used, such as PR-bucket quadtree-based search index (Hjaltason and Samet 1995, Pebesma 2004). When neighborhoods are fairly large, e.g. when they contain at least 50 points, and when spatial correlation is strong, and when the trend model only contains a mean value (intercept), then using global or local neighborhoods both result in practically identical prediction and prediction standard error maps.

5.3 Dealing with Auxiliary Information: Multivariate Prediction

Treating GIS coverages as $f(s)$ in the regression model of the trend is an effective way of using auxiliary information to predict the primary variable, but it does assume that the coverage is *complete*. Suppose that our sea floor surface depth data were not given as a GIS layer covering the study area, but rather as a set of points. In that case, interpolating depth and treating it as a known coverage would not be correct, as depth at interpolated locations was only predicted, not measured, and prediction errors may be large.

The alternative is to treat the secondary, measured information as a random field (instead of a given, fixed covariate $f(s)$) and to model spatial cross-correlation between the primary and secondary variable, either by using the classic cross-variogram:

$$\gamma_{12}(h) = 0.5E(Z_1(s) - Z(s+h))(Z_2(s) - Z_2(s+h)) \tag{6}$$

which can be used when Z_1 and Z_2 have sufficient observation locations in common, or by using the pseudo cross-variogram:

$$\gamma_{12}(h) = 0.5Var(Z_1(s) - Z_2(s+h))^2 \tag{7}$$

Given models for the primary and secondary variables variograms and their cross-variogram, cokriging (Cressie 1993, Wackernagel 1998) is used to predict each of the variables and cosimulation (Gómez-Hernández and Journel 1993) can be used for their simultaneous simulation. Even if the secondary variable is available as a complete coverage, Rivoirard (2002) showed that under certain circumstances it may be beneficial to treat it as a random variable, i.e. in a cokriging setting.

5.4 Temporal Change

In addition to prediction variances cokriging also yields the prediction error covariance $Cov(\hat{Z}_1(s_0), \hat{Z}_2(s_0))$, a quantity that is needed for example to estimate the standard error of the predicted differences $\hat{Z}_1(s_0) - \hat{Z}_2(s_0)$. This difference estimates temporal changes

when $Z_1(s)$ and $Z_2(s)$ are measurements at two moments in time. In this approach, we assume that time yields multiple, correlated realizations of the measured variable.

Another approach for modeling spatial-temporal variability is to consider the variable measured as coming from a space-time random function $Z(s, t)$ with a single variogram defined over continuous space-time. The difficulty of this latter approach is inferring this variogram, because we need to compare distances between points separated in both space and time. An advantage of this approach is that it allows predictions of $Z(s, t)$, at any location/moment combination (s_0, t_0).

Kyriakidis and Journel (1999) provide an overview of different geostatistical approaches to space-time modeling. Pebesma and Duin (2005) provide a space-time analysis of the PCB-138 data used here.

5.5 Other Non-Linear Approaches

As with any linear statistical model, the linear geostatistical model works best for data with a Gaussian (residual) distribution, and in practice, it also works well for data with a distribution that is not too far from Gaussian. In the example on PCB138 and cadmium we showed how log-transformation made the variables well suited to this model.

Other non-linear approaches include approaches where the variable is transformed to Gaussian (the "multi-Gaussian" approach (Goovaerts 1997), or the disjunctive kriging approach (Rivoirad 1984)), or approaches where a variable is transformed to binary values depending on whether its value is below (1) or above (0) a certain threshold. The latter is indicated with the indicator approach (Goovaerts 1997).

The indicator approach seems a natural start when the dependent variable is binary, or categorical. Gotway and Stroup (1997) give an extension of this approach, building upon generalized linear models (McCullagh and Nelder 1989) for modeling the trend, and mean-dependent covariances.

6 Conclusions

In this paper, we have shown for a sample data set that a successful analysis can be obtained by modeling the trend in the data using coverages available in the GIS database, and by modeling the spatial correlation in the residual by ways of variogram analysis. Much of the geostatistical analysis of today concentrates only on the latter, ignoring possibly informative external variables. This is partly driven by the options provided by geostatistical software. For example, the Geostatistical Analyst extension for ArcGIS 8.3 does provide universal kriging where the trend is driven by the spatial coordinates, but not where the trend is driven by external variables, as in the example shown here (sometimes referred to as "external drift kriging"). In this author's experience, spatial coordinate regression and variogram analysis seldom inform us about the physical (causal) relationship regarding the variable studied and seldom help increase the understanding of its variability. Relevant external variables do so, are usually present, and should be used wherever possible to drive the geostatistical predictions.

Data, Software, and Acknowledgements

The sea floor surface sediment data used in this paper are available from the author's web site. The software used throughout this paper is the R system (Ihaka and Gentleman 1988), which is an open source implementation of the S language (Becker et al. 1988). Within R, we used the gstat package for R or S-PLUS (Pebesma and Wesseling 1998; Pebesma 2003, 2005), which is also in open source form available from http://www.gstat.org/. This package extends the model interface of S (Chambers and Hastie 1992) to multivariable geostatistical models. The model interface takes care of automatic translation of categorical variables into the necessary dummy variables and allows a simple definition of interactions or nested effects for example. The sea floor surface sediment data set and financial support for the development of the gstat S package were gratefully obtained from the Dutch National Institute for Coastal and Marine Management (RIKZ). Richard Duin (RIKZ; http://www.rikz.nl) played a stimulating role in the work presented here.

References

Becker R A, Chambers J M, and Wilks A R 1988 *The New S Language*. London, Chapman and Hall

Burrough P A and McDonnell R A 1998 *Principles of Geographical Information Systems*. New York, Oxford University Press

Chambers J M and Hastie T J 1992 *Statistical Models in S*. London, Chapman and Hall

Chilès J and Delfiner P 1999 *Geostatistics: Modeling Spatial Uncertainty*. New York, John Wiley and Sons

Christensen R 1991 *Linear Models for Multivariable, Time Series and Spatial Data*. New York, Springer-Verlag

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

Diggle P J, Tawn R A, and Moyeed R A 1998 Model-based geostatistics. *Applied Statistics* 47: 299–350

Gomez-Hernandez J J and Journel A G 1993 Joint sequential simulation of multiGaussian fields. In Soares A (ed) *Geostatistics Troia '92*. Dordrecht, Kluwer: 85–94

Goovaerts P 1997 *Geostatistics for Natural Resources Evaluation*. In New York, Oxford University Press

Gotway C A and Stroup W W 1997 A generalized linear model approach to spatial data analysis and prediction. *Journal of Agricultural, Biological and Environmental Statistics* 2: 157–78

Hjaltason G and Samet H 1995 Ranking in spatial databases. In Egenhofer M J and Herring J R (eds) *Advances in Spatial Databases*. Berlin, Springer-Verlag Lecture Notes in Computer Science No. 951: 83–95

Heuvelink G B M 1998 *Error Propagation in Environmental Modeling with GIS*. London, Taylor and Francis

Ihaka R and Gentleman R 1996 R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 5: 299–314

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

Journel A G and Huijbregts C J 1983 *Mining Geostatistics*. Cambridge, Oxford University Press

Kitanidis P K 1983 Statistical estimation of polynomial generalized covariance functions and hydrologic applications. *Water Resources Research* 19: 909–21

Kitanidis P K 1993 Generalized covariance functions in estimation. *Mathematical Geology* 25: 525–40

Kitanidis P K 1997 *Introduction to Geostatistics: Applications in Hydrogeology*. Cambridge, Cambridge University Press

Kyriakidis P C and Journel A G 1999 Geostatistical space-time models: A review. *Mathematical Geology* 31: 651–84

Laane R W P M, Sonneveldt H L A, Van der Weyden A J, Loch J P G, and Groeneveld G 1999 Trends in the spatial and temporal distribution of metals (Cd, Cu, Zn and Pb) and organic compounds (PCBs and PAHs) in Dutch coastal zone sediments from 1981 to 1996: A model case study for Cd and PCBs. *Journal of Sea Research* 41: 1–17

McCullagh P and Nelder J A 1989 *Generalized Linear Models*. London, Chapman and Hall

Pebesma E J 2003 Gstat: Multivariable geostatistics for S. In *Proceedings of the Third International Workshop on Distributed Statistical Computing (DSC 2003)*, Vienna, Austria

Pebesma E J 2004 Multivariable geostatistics in S: The gstat package. *Computers and Geosciences* 30: 683–91

Pebesma E J and De Kwaadsteniet J W 1997 Mapping groundwater quality in the Netherlands. *Journal of Hydrology* 200: 364–86

Pebesma E J and Heuvelink G B M 1999 Latin hypercube sampling of Gaussian random fields. *Technometrics* 41: 303–12

Pebesma E J and Wesselling C G 1998 Gstat: A program for geostatistical modeling, prediction and simulation. *Computers and Geosciences* 24: 17–31

Pebesma E J and Duin R N M 2005 Spatio-temporal mapping of sea floor sediment pollution in the North Sea. In Renard Ph and Froidevaux R (eds) *Proceedings of the Fifth European Conference on Geostatistics for Environmental Applications (GeoENV 2004)*. New York, Springer: 367–68

Ripley B D 1981 *Spatial Statistics*. New York, John Wiley and Sons

Rivoirard J 1984 *Introduction to Disjunctive Kriging and Non-linear Geostatistics*. Oxford, Oxford University Press

Rivoirard J 2002 On the structural link between variables in kriging with external drift. *Mathematical Geology* 34: 797–808

Stein M L 1999 *Interpolation of Spatial Data: Some Theory for Kriging*. New York, Springer-Verlag

Wackernagel H 1998 *Multivariate Geostatistics: An Introduction with Application* (Second edition). Berlin, Springer-Verlag