Accounting for Source Location and Transport Direction into Geostatistical Prediction of Contaminants

HIROTAKA SAITO AND PIERRE GOOVAERTS*

Environmental and Water Resources Engineering, Department of Civil and Environmental Engineering, The University of Michigan, Ann Arbor, Michigan 48109-2125

This paper presents a variant of the well-known kriging with a trend that allows one to account for the pollutant source coordinates and information about transport process into the spatial prediction of pollutant concentration. The new technique is illustrated using lead data from a Dallas metropolitan area and cadmium data from Palmerton (PA) NPL Superfund site. Instead of modeling the local spatial trend as low-order polynomials of coordinates, it is here expressed as a function of two factors that likely control the pollution spread: the distance to the smelter and the deviation from the major wind direction. Four different combinations of these two factors are developed, and their prediction performances are evaluated for a range of wind directions using cross-validation. Comparison with two traditional algorithms (OK and KT) shows that the proposed approach leads to smaller mean square errors of prediction when the correct wind direction is determined. The best combination of trend model and wind direction is site-dependent and derived using cross-validation. Kriging of residuals from a global trend model is an alternative to the use of local trends, and both techniques are shown to outperform ordinary kriging.

Introduction

Site characterization is a critical step for the design of efficient remediation plans at contaminated sites. Geostatistics is increasingly used to model the spatial variability of contaminant concentrations and map them using generalized least-squares regression, known as kriging (1–4). Among many variants of kriging algorithms, ordinary kriging (OK) is the most straightforward and has been widely used (5). An implicit assumption of OK is the lack of spatial trend within the search window used to select data retained in estimation (assumption of local stationarity, 6). The presence of local trends can be accounted for by using kriging with a trend (KT), traditionally named and recognized as universal kriging. The trend component is typically modeled as a smoothly varying function of spatial coordinates, the coefficients of which are estimated within each search window. In most studies, low-order polynomials of the coordinates are blindly used mainly because of its computational convenience.

In some situations, the geostatistician has physical ground for choosing a particular type of analytical trend functions. For example, Séguret and Huchon (7) used a series of sine and cosine functions to model the periodic trend of geomagnetic data. Another case is a site contaminated by one or more point sources of pollution. Depending on dispersion processes, pollutant concentrations are expected to decline as a function of distance and possibly direction with respect to the source. When the location of these sources are known, they should be incorporated into environmental models to obtain better predictions and remediation decisions.

To our knowledge, the location of pollutant sources has never been incorporated directly into the kriging system. Instead a global spatial trend is usually fitted to the data, and kriging is used to interpolate the residuals, i.e., differences between measurements and trend values. For example, Mohammadi et al. (8) fitted the spatial trend in an airborne Cd-contaminated area as a function of distance to the zinc ore smelter and wind directions. More complete models have been developed to account for additional factors, such as closeness to old houses for lead paint effects (9, 10), and allow one to discriminate between site emissions and urban background contributions to pollutant concentrations. A potential limitation of global trends is that their parameters are constant over the study area; hence, they do not account for possible changes in the impact of different factors over the contaminated site.

In this paper, we propose a variant of KT algorithm that allows one to integrate information about the location of the pollutant source and transport process into the spatial mapping of contaminants. The technique is applied to a Dallas metropolitan area contaminated with lead emitted by an industrial smelter (11, 12) and a Palmerton (PA) site contaminated with cadmium emitted by two zinc smelters (13, 14). Cross-validation is used to compare, over a range of wind directions, the prediction performances of the new technique with ordinary kriging, the traditional implementation of kriging with a trend, and the kriging with a global trend model.

Theory

Consider the problem of estimating the value of an attribute z (e.g., heavy metal concentration) at an unsampled location \( \mathbf{u} \), where \( \mathbf{u} \) is a vector of spatial coordinates. The information available consists of measurements of \( z \) at \( n \) locations \( \mathbf{u}_i \), \( (z(\mathbf{u}_i), \alpha = 1, 2, \ldots, n) \) plus the coordinates of \( S \) pollutant sources, \( \mathbf{u}_s \), \( s = 1, 2, \ldots, S \). This section reviews the kriging paradigm and shows how spatial coordinates can be replaced by other factors, such as the distance to the source, into the trend modeling.

All kriging estimators are variants of the linear regression estimator defined as

\[
\hat{z}(\mathbf{u}) - m(\mathbf{u}) = \sum_{\alpha=1}^{n} \lambda_{\alpha}(\mathbf{u})[Z(\mathbf{u}_\alpha) - m(\mathbf{u}_\alpha)]
\]

(1)

where \( \lambda_{\alpha}(\mathbf{u}) \) is the weight assigned to datum \( z(\mathbf{u}_\alpha) \), interpreted as a realization of the random variable \( R \). \( Z(\mathbf{u}) \) and \( m(\mathbf{u}) \) is the trend component. In practice, only the observations closest to \( \mathbf{u} \) being estimated are retained, i.e., the \( n(\mathbf{u}) \) data within a given neighborhood or window \( W(\mathbf{u}) \) centered on \( \mathbf{u} \). The RV \( Z(\mathbf{u}) \) is usually decomposed into a residual \( R(\mathbf{u}) \) and a trend component \( m(\mathbf{u}) \):

\[
Z(\mathbf{u}) = R(\mathbf{u}) + m(\mathbf{u})
\]

(2)

with \( \text{E}[R(\mathbf{u})] = 0 \) and so \( \text{E}[Z(\mathbf{u})] = m(\mathbf{u}) \).
In traditional application of kriging, the trend is modeled as a smoothly varying function of spatial coordinates as

\[ m(\mathbf{u}) = \sum_{k=0}^{K} a_k f_k(\mathbf{u}) \quad \forall \mathbf{u} \in W(\mathbf{u}) \]  

(3)

where \( f_k(\mathbf{u}) \) are known functions, and the coefficients \( a_k \) are unknown but considered constant within each window \( W(\mathbf{u}) \).

By convention \( f_0(\mathbf{u}) = 1 \); hence, the simplest case \( K = 0 \) corresponds to the situation where the trend component is constant and unknown within \( W(\mathbf{u}) \), yielding the well-known ordinary kriging (OK) estimator. The general case \( K > 0 \) is referred to as kriging with a trend (OK).

The general case corresponds to the situation where the trend component is unknown but considered constant within each window \( W(\mathbf{u}) \).

Then the unsampled value \( z(\mathbf{u}) \) is estimated as a linear combination of neighboring observations:

\[ Z^*(\mathbf{u}) = \sum_{n=1}^{n(u)} \lambda_n(\mathbf{u})Z(\mathbf{u}_n) \]

(4)

where kriging weights are solutions of the following system of linear equations:

\[ \sum_{n=1}^{n(u)} \lambda_n(\mathbf{u})\gamma_k(\mathbf{u}_n - \mathbf{u}) = \gamma_k(\mathbf{u}_n - \mathbf{u}) \quad \alpha = 1, \ldots, n(u) \]

(5)

\[ \sum_{n=1}^{n(u)} \lambda_n(\mathbf{u}) = 1 \]

(6)

\[ \sum_{n=1}^{n(u)} \lambda_n(\mathbf{u})f_k(\mathbf{u}_n) = f_k(\mathbf{u}) \quad k = 1, \ldots, K \]

(7)

The unknown trend coefficients \( a_k \) are filtered from the estimator by imposing the following \( (K + 1) \) constraints:

\[ \sum_{n=1}^{n(u)} \lambda_n(\mathbf{u})f_k(\mathbf{u}_n) = f_k(\mathbf{u}) \quad k = 0, \ldots, K \]

(8)

In the case of a single source located at \( \mathbf{u}_s \) the trend component could be modeled as

\[ m(\mathbf{u}) = m(d, \theta) = a_0 + a_1d + a_2\Delta \theta \]

(9)

where \( \Delta \theta \leq 180^\circ \) is the azimuth discrepancy between \( \theta \) and the prevailing wind direction \( \phi \). There are other possibilities for the trend functions, and several models are later introduced and compared to the traditional trend model (eq 8). The approach is easily extended to the situation where multiple sources of pollution are present. For instance, for three sources \((S = 3)\) model (eq 11) becomes

\[ m(\mathbf{u}) = m(d, \theta) = a_0 + a_1d_1 + a_2d_2 + a_3d_3 + a_4\Delta \theta_1 + a_5\Delta \theta_2 + a_6\Delta \theta_3 \]

(10)

FIGURE 1. One-dimensional example illustrating major differences between OK and KT with a traditional linear trend, \( m(\mathbf{u}) = a_0 + a_1x \), (KT0), or a trend that accounts for the distance to the location of the source depicted by the vertical arrow, \( m(\mathbf{u}) = a_0 + a_1 \log(d) \), (KT1). Ten Cd concentrations are actual data, but the location of the source is purely fictitious.

\[ d = |\mathbf{u}_s - \mathbf{u}| = \sqrt{(x_s - x)^2 + (y_s - y)^2} \]

(11)

where \( \Delta \theta \leq 180^\circ \) is the azimuth discrepancy between \( \theta \) and the prevailing wind direction \( \phi \). There are other possibilities for the trend functions, and several models are later introduced and compared to the traditional trend model (eq 8). The approach is easily extended to the situation where multiple sources of pollution are present. For instance, for three sources \((S = 3)\) model (eq 11) becomes

\[ m(\mathbf{u}) = m(d, \theta) = a_0 + a_1d_1 + a_2d_2 + a_3d_3 + a_4\Delta \theta_1 + a_5\Delta \theta_2 + a_6\Delta \theta_3 \]

(12)

Figure 1 illustrates the application of three different kriging algorithms to estimate cadmium concentration along a transect. For the new KT, the trend was modeled as a linear function of the logarithm of the distance to a fictitious point source. All three estimated profiles pass through the data (exactitude property). Major discrepancies between algorithms occur next to the source where the proposed technique (KT1) yields larger estimates, which is intuitively expected if it is actually the main source of cadmium. Other differences are observed in extrapolation situation (right part of the transect) since the choice of the trend model becomes critical as the location being estimated gets farther away from data locations.

The troublesome part for KT is the inference of the residual semivariogram, since there is no direct way to compute it from the available \( z \) values. The most common approach is to estimate the semivariogram using only the data pairs that are unaffected or slightly affected by the trend, i.e., data pairs along the direction perpendicular to the trend direction (6). The latter can be determined from a variogram map or information related to the dispersion of contaminants from the source, e.g., wind or flow direction. The residual semivariogram is then considered isotropic, and any anisotropy in the data is accounted for in the trend model.

**Materials and Methods**

**Heavy Metal Data Sets.** The different kriging algorithms are illustrated and compared using two different data sets: 200...
soil lead concentrations measured in the Dallas metropolitan area (11, 12), and 412 soil cadmium concentrations collected in Palmerton, PA (13, 14). The Dallas metropolitan area is the site of an industrial lead smelter that is depicted by a diamond in the location map of Figure 2a. Higher concentrations are found around the smelter, which indicates that spatial coordinates of the smelter might be worth incorporating in the trend model. Following previous study (11), logarithmic transforms of lead concentrations are used in the geostatistical analysis, and their cumulative distribution is shown in Figure 2c. At the Palmerton site, metal oxide fumes emitting from two zinc smelters have caused higher concentration of zinc, cadmium, copper, and lead in soils (13). In this paper, the focus is on cadmium because of high levels of contamination found in soils and vertebrates at the site (15). Like the Dallas data, logarithmic transforms of concentrations are used to reduce the impact of extremely high values on semivariogram inference, and their cumulative distribution is shown in Figure 2d. The two zinc smelters, depicted by diamonds in the location map of Figure 2b, are referred to as the east and west plants, respectively. The Dallas lead data set was used as an example of a single source case, while the Palmerton data set illustrates the application to multiple sources (5 = 2). The exact coordinates of the smelters (chimneys) are unknown and were approximated from the information available, such as regional maps. The impact of

FIGURE 2. Location maps of logarithmic transforms of Pb concentration data at Dallas metropolitan area (panel a) and Cd concentration at Palmerton site (panel b). Smelters are depicted by diamonds. The corresponding cumulative distributions and semivariograms with the model fitted are displayed in panels c–f.
source locations onto the prediction of concentrations is discussed later.

According to Myers and Bryan (11), the average yearly wind direction at the Dallas site is S–N, which is consistent with the presence of large concentrations to the north of the smelter. The impact of the trend should then be the weakest along the E–W direction, and the residual semivariogram is computed along that direction. Figure 2e shows the E–W experimental semivariogram with the model fitted that will be used as the isotropic residual semivariogram in KT. The omnidirectional semivariogram is also computed and modeled since it is required for OK. Regardless of the major wind direction, E–W and omnidirectional semivariograms are not too different, which reflects a moderate contribution of the wind direction to the spatial pattern. Several studies (13, 15) showed the average yearly wind direction in Palmerton site is W–E, which might explain the existence of a large proportion of high concentrations on the east side of both

![Image of concentration maps](image-url)
smelters. The residual semivariogram is then computed and modeled along the direction (N–S) perpendicular to the average wind direction, see Figure 2f. For this comparative study, these semivariogram models are used for all sampling intensities and trend models so that differences in performance mainly arise from the algorithms themselves.

**Trend Model.** As mentioned above, the trend model could be a function of various factors, such as the distance to the source or the deviation from the predominant wind direction. The following four models \((S = 1)\) are used in this paper:

\[
KT_1: m_1(u) = a_0 + a_1 \log(d) \tag{13}
\]

\[
KT_2: m_2(u) = a_0 + a_1 \log(d) + a_2 \Delta \theta \tag{14}
\]

\[
KT_3: m_3(u) = a_0 + a_1 \log(d) \times \Delta \theta \tag{15}
\]

\[
KT_4: m_4(u) = a_0 + a_1 \Delta \theta + a_2 \log(d) \times \Delta \theta \tag{16}
\]

where \(d\) is the distance to the source of contamination and \(\Delta \theta\) is the deviation from the major wind direction.

The first trend model, \(m_1\), considers only the distance to the source. Since pollutant concentrations often decrease exponentially as the distance to the source increases \((8)\), the logarithm of the distance is used here. The next two models account for both the distance and the azimuth. In \(m_2\), the two factors are assumed independent so that their contributions are additive, while \(m_3\) considers possible interactions between the two factors. In other words, \(m_2\) amounts at using a smaller rate of decline of concentration along the major wind direction because the contaminant itself is transported farther away. On the other hand, in the directions perpendicular or opposite to the prevailing wind, the transport is much less important so that the concentration decreases rapidly. The last model, \(m_4\), combines \(m_2\) and \(m_3\). Multiple sources \((S > 1)\) are easily incorporated by adding the components associated to each source into the trend model as illustrated in eq.12.

**Prediction Performances.** For each data set, the mean square error (MSE) of prediction by OK and various KLT algorithms was assessed using cross-validation whereby one observation at a time is temporarily removed from the data set and re-estimated from remaining data \((6)\). Prediction errors were systematically computed over all possible major wind directions \((0^\circ, 1^\circ, ..., 359^\circ)\) to identify the direction with the best prediction performance, which was then compared
to the one reported by several authors for each site (11, 13–15).

The benefit of using global trends was investigated by fitting a global trend to the data using linear regression, followed by simple kriging of residuals. For each type of trend model (eqs 14–16), all possible wind directions were tried, and the one with the largest $R^2$ was retained as the predominant wind direction. Interpolated residuals were then added to trend estimates to obtain concentration estimates. Prediction errors (MSE) for such kriging with a global trend were computed using cross-validation and compared to other KLT scores.

To investigate the impact of the number of data on prediction performances, two different sampling intensities were considered for each data set: original sample size and a random subset of 50 observations. All techniques introduced in this paper have been implemented by modifying the FORTRAN source codes that are part of the public domain software library, GSLIB (16).

Results and Discussion

The impact of the kriging algorithm (OK, $KT_0$, and $KT_1$) and sampling intensity (200 data and 50 data) on the mapping of lead concentration (Dallas data set) is first illustrated in Figure 3. Accounting for the smelter location increases the concentration estimates to the north of the source, particularly when data are sparse (50 data). Even though the general trend is captured similarly by all kriging algorithms, OK maps display much smoother distribution of lead concentration than the others.

Figure 4 shows, for each data set and algorithm (OK and local trends), the MSE as a function of the wind direction $\phi$. Note that for $KLT_2$, any pair of directions ($\phi, \phi + 180^\circ$) yields the same results because of the linearity in both the kriging system (eq 7) and the trend model (eq 14). For the Dallas data set with 200 data (Figure 4a), best results (minimum MSE) for $KT_2$–$KT_4$ were obtained using a wind direction of $0^\circ$ (S–N), which corresponds to the direction reported in the literature (11). However, for the random subset of 50 data best results were obtained for a wind direction of $40^\circ$ for $KT_2$ and $38^\circ$ for $KT_3$ and $KT_4$ (Figure 4b). The impact of $\phi$ changes on MSE scores is also more pronounced, which confirms that the choice of wind direction can become critical when data are sparse and estimation is increasingly influenced by the trend model. It is noteworthy that the benefit of the new approach over OK is greater when data are sparse.

For the Palmerton data set, best results are almost always obtained for a wind direction $\phi = 90^\circ$ ± 5°, see Figure 4c,d), which is again in agreement with the information found in the literature (13). Once the correct wind direction is
TABLE 1. Mean Square Errors (MSE) of Prediction Obtained for a Range of Kriging Algorithms, Local and Global Trend Models, and Sampling Intensities

<table>
<thead>
<tr>
<th>data set</th>
<th>trend model</th>
<th>OK</th>
<th>KT₀</th>
<th>KT₁</th>
<th>KT₂</th>
<th>KT₃</th>
<th>KT₄</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dallas 200 data</td>
<td>global wind (φ)</td>
<td>0.7097</td>
<td>0.7036</td>
<td>0.7470</td>
<td>0.7570</td>
<td>0.7272</td>
<td>0.7584</td>
</tr>
<tr>
<td></td>
<td>local wind (φ)</td>
<td>0.7097</td>
<td>0.7219</td>
<td>0.7051</td>
<td>0.6957</td>
<td>0.6971</td>
<td>0.6959</td>
</tr>
<tr>
<td>Dallas 50 data</td>
<td>global MEE</td>
<td>0.6401</td>
<td>0.6040</td>
<td>0.6479</td>
<td>0.5671</td>
<td>0.5728</td>
<td>0.5160</td>
</tr>
<tr>
<td></td>
<td>local wind (φ)</td>
<td>0.6401</td>
<td>0.6539</td>
<td>0.6213</td>
<td>0.6012</td>
<td>0.5751</td>
<td>0.5957</td>
</tr>
<tr>
<td>Palmerton 412 data</td>
<td>global MEE</td>
<td>0.4157</td>
<td>0.4077</td>
<td>0.3941</td>
<td>0.4028</td>
<td>0.4011</td>
<td>0.4039</td>
</tr>
<tr>
<td></td>
<td>local wind (φ)</td>
<td>0.4157</td>
<td>0.4173</td>
<td>0.4175</td>
<td>0.4131</td>
<td>0.4081</td>
<td>0.4145</td>
</tr>
<tr>
<td>Palmerton 50 data</td>
<td>global MEE</td>
<td>0.6202</td>
<td>0.6723</td>
<td>0.5421</td>
<td>0.4872</td>
<td>0.6244</td>
<td>0.4862</td>
</tr>
<tr>
<td></td>
<td>local wind (φ)</td>
<td>0.6202</td>
<td>0.4564</td>
<td>0.4921</td>
<td>0.3832</td>
<td>0.5012</td>
<td>0.4059</td>
</tr>
</tbody>
</table>

*For trends KT2–KT4, results are provided for the wind direction that leads to the smallest errors.

**Literature Cited**


Received for review January 29, 2000. Revised manuscript received September 12, 2001. Accepted September 18, 2001.

ES010580F