Published in final edited form as:

Math Geosci. 2010 July 1; 42(5): 535–554. doi:10.1007/s11004-010-9286-5.

Combining Areal and Point Data in Geostatistical Interpolation: Applications to Soil Science and Medical Geography

Pierre Goovaerts

Biomedware, Inc., 3526 W Liberty, Suite 100, Ann Arbor, MI 48103, USA

Abstract

A common issue in spatial interpolation is the combination of data measured over different spatial supports. For example, information available for mapping disease risk typically includes point data (e.g. patients' and controls' residence) and aggregated data (e.g. socio-demographic and economic attributes recorded at the census tract level). Similarly, soil measurements at discrete locations in the field are often supplemented with choropleth maps (e.g. soil or geological maps) that model the spatial distribution of soil attributes as the juxtaposition of polygons (areas) with constant values. This paper presents a general formulation of kriging that allows the combination of both point and areal data through the use of area-to-area, area-to-point, and point-to-point covariances in the kriging system. The procedure is illustrated using two data sets: (1) geological map and heavy metal concentrations recorded in the topsoil of the Swiss Jura, and (2) incidence rates of late-stage breast cancer diagnosis per census tract and location of patient residences for three counties in Michigan. In the second case, the kriging system includes an error variance term derived according to the binomial distribution to account for varying degree of reliability of incidence rates depending on the total number of cases recorded in those tracts. Except under the binomial kriging framework, area-and-point (AAP) kriging ensures the coherence of the prediction so that the average of interpolated values within each mapping unit is equal to the original areal datum. The relationships between binomial kriging, Poisson kriging, and indicator kriging are discussed under different scenarios for the population size and spatial support. Sensitivity analysis demonstrates the smaller smoothing and greater prediction accuracy of the new procedure over ordinary and traditional residual kriging based on the assumption that the local mean is constant within each mapping unit.

Keywords

Soil map; Cancer; Disaggregation; Change of support; Indicator; Binomial

1 Introduction

Since its origin, geostatistics has been routinely used to predict block averages from point data. More recently, several authors (Gotway and Young 2002, 2005, 2007; Kyriakidis 2004; Goovaerts 2008) proposed the use of kriging to predict point values from areal data, an approach referred to as area-to-point (ATP) kriging, following the terminology in Kyriakidis (2004). This approach allows mapping the variability within geographical units while ensuring the coherence of the prediction so that the sum or average of disaggregated estimates is equal to the original areal datum. However, looking at the general formulation

of kriging (Journel and Huijbregts 1978), it is clear that this approach can accommodate different spatial supports for the data, such as a mixture of point data and irregular block values.

The issue of combining data measured on different spatial supports has been the topic of much research in soil science. Indeed, the information available for mapping continuous soil attributes often includes point field data and choropleth maps (e.g. soil or geological maps) that model the spatial distribution of soil attributes as the juxtaposition of polygons (areas) with constant values. One common approach is to use soil map information to inform the local mean of the random function (Goovaerts and Journel 1995; Hengl et al. 2004; Liu et al. 2006). Variography and kriging are then conducted on the stationary residuals (that is, differences between point measurements and mapping units' means) and results are combined to obtain the final estimates. More recently, Goovaerts (2010) proposed to replace the choropleth map of local means by an isopleth map created using ATP kriging to allow for smooth transitions between units. In either case, because residual kriging proceeds in two steps, there is no guarantee that the final map of kriging estimates will honor areal data: the average of interpolated values within each area typically does not equal the areal datum.

Another field that is faced with the challenge of incorporating data measured over different spatial supports is medical geography or spatial epidemiology, which is concerned with the study of spatial patterns of disease incidence and mortality and the identification of potential causes of disease, such as environmental exposure or socio-demographic factors (Waller and Gotway 2004; Goovaerts 2007, 2009a). Maps of health outcomes, such as cancer mortality or incidence of late-stage diagnosis, are used by public health officials to identify areas of excess and to guide surveillance and control activities, including consideration of health services needs and resource allocation for screening and diagnostic testing. Quality of decision-making thus relies on an accurate quantification of risks from observed rates which can be very unreliable when computed from sparsely populated geographical units or recorded for minority populations. Data available for human health studies fall within two main categories: individual-level data (e.g. location of patients' residences) and aggregated data (e.g. incidence rates of a disease computed within administrative entities). For mapping purposes, both types of data are typically processed independently using different sets of methods. For example, kernel density estimation is used to create isopleth risk maps from individual-level data, such as the risk of late-stage cancer diagnosis computed from the location of residences of patients that were diagnosed late or early (Talbot et al. 2000; Rushton et al. 2004). On the other hand, the noise attached to unreliable rates recorded for sparsely populated areas is filtered using smoothing algorithms to create reliable choropleth maps of health outcomes (Best et al. 2005; Goovaerts 2005). Recently, Goovaerts (2009a) proposed a two-step approach to combining individual-level data (e.g. patient residences) and area-based data (e.g. rates recorded at census tract level) into the mapping of late-stage cancer incidences, with an application to breast cancer in three Michigan counties. Spatial trends in cancer incidences are first estimated from census data using area-to-point binomial kriging. This prior model is then updated using residual indicator kriging and individual-level data. However, as discussed earlier for soil data, this form of residual kriging does not guarantee that the final map of kriging estimates will honor areal data, which is the main focus of the present research.

This paper presents a general formulation of kriging that allows the combination of both point and areal data through the use of area-to-area, area-to-point, and point-to-point covariances in the kriging system. This approach capitalizes on the availability of GIS to discretize polygons of irregular shape and size and the knowledge of the point-support semivariogram model that can be inferred directly from point measurements, thereby eliminating the need for deconvolution procedure (Goovaerts 2008). A similar approach was

recently implemented within a stochastic simulation framework by Liu and Journel (2009). Methodological developments are illustrated using two data sets: the geological map and heavy metal concentrations recorded in the topsoil of the Swiss Jura (Goovaerts 1997) and breast cancer cases diagnosed over 17 years in three Michigan counties. Traditional ordinary kriging is introduced for the mapping of soil properties, while a new formulation of binomial kriging (Webster et al. 1994a; Oliver et al. 1998; Walker et al. 2008) is used for the mapping of health outcomes. The performance of the proposed approaches, relative to ordinary kriging or a traditional residual kriging with choropleth trend model, is assessed using jackknife. Performance criteria include the magnitude of prediction errors, the accuracy of the model of uncertainty, the smoothness of interpolated maps, and the ability to discriminate between early- and late-stage cancer cases.

2 Setting the Problem

The combination of point field data and areal map data in spatial interpolation will first be illustrated using a case study related to heavy metal contamination of an area of the Swiss Jura. In the spring of 1992, the Swiss Federal Institute of Technology surveyed the topsoil of a 14.5 km² region near La Chaux-de-Fonds and measured the concentrations of seven trace metals at 359 locations (Atteia et al. 1994) (see Fig. 1A). Webster et al. (1994b) investigated the effect of geological formation and land use on topsoil concentrations and found much smaller concentrations for most metals on the Argovian formation. The geological map displayed in Fig. 1C will thus act as our source of areal information. This map includes 35 polygons that belong to one of the five geological formations. In the present application, no independent calibration of the rock maps exists and, for the purpose of illustration, the mean chromium concentration within each formation was simply computed as the weighted average of all samples collected on that formation. The weight is the area of influence of each sample (that is, Thiessen polygon) in order to account for data clustering. In a situation where a soil map is available, areal data would simply be identified with concentrations recorded on representative profiles for each mapping unit (e.g. legacy soil data, see Kerry et al. 2010b).

The second case study is borrowed from the field of medical geography. Invasive breast cancer cases, diagnosed during the calendar years 1985 through 2002 in Michigan, were compiled by the Michigan Cancer Surveillance Program (MCSP) and successfully geocoded at residence at time of diagnosis. The current study focuses on cases diagnosed for white women (age 65–74 years) in 83 census tracts of three counties in south-western Michigan (see Fig. 1B). Out of the 937 women diagnosed with breast cancer during that time period, 18.46% of cases were defined as late-stage (that is, regional and distant metastatic cancer) according to the SEER General Summary Stage classification (Young et al. 2001). Areal data are incidence rates computed from 937 cases at the level of census tracts (Fig. 1D) which are frequently used in the poverty literature as proxy for neighborhoods in which residents are likely to face similar social and economic conditions (Barry and Breen 2005).

3 Methodology

3.1 Area-and-Point Kriging

Consider the problem of estimating the value of a continuous attribute z at any location $\mathbf{u}$ within a study area A . The information available consists of a set of point data collected at n discrete locations $\mathbf{u}_\alpha \{z(\mathbf{u}_\alpha); \alpha = 1, \dots, n\}$, supplemented by a set of B areal data $\{z(v_k); k = 1, \dots, B\}$ recorded for mapping units v_k of various sizes and shapes. Both point and areal data can be simultaneously incorporated into the prediction using the area-and-point (AAP) kriging estimate, defined as

$$z_{AAP}^*(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_{\alpha}(\mathbf{u}) z(\mathbf{u}_{\alpha}) + \sum_{k=n(\mathbf{u})+1}^{n(\mathbf{u})+K} \lambda_k(\mathbf{u}) z(v_k), \quad (1)$$

where $n(\mathbf{u})$ and K are the number of surrounding point and areal data, respectively. Point observations are typically selected based on their distance to the interpolation node $\mathbf{u}$ while areal data are chosen according to adjacency rules; for example, all polygons adjacent to the polygon including $\mathbf{u}$ are used in the estimation.

The kriging weights are the solution of the following ordinary kriging system

$$\begin{aligned} \sum_{j=1}^{n(\mathbf{u})+K} \lambda_j(\mathbf{u}) \bar{C}(x_i, x_j) + \mu(\mathbf{u}) &= \bar{C}(x_i, \mathbf{u}), \quad i=1, \dots, n(\mathbf{u})+K, \\ \sum_{j=1}^{n(\mathbf{u})+K} \lambda_j(\mathbf{u}) &= 1, \end{aligned} \quad (2)$$

where $\mu(\mathbf{u})$ is the Lagrange multiplier, $x_i = \mathbf{u}_i$ if $i \leq n(\mathbf{u})$ and $x_i = v_i$ otherwise. The quantity $\bar{C}(x_i, x_j)$ represents a point-to-point, point-to-block or block-to-block covariance depending on the indices i and j . Like in traditional block kriging, the block-to-point covariance $\bar{C}(v_k, \mathbf{u})$ is approximated by the average of the point support covariance $C(\mathbf{h})$ computed between the location $\mathbf{u}$ and a set of P_k points discretizing the block v_k . A similar procedure is used for the block-to-block covariance $\bar{C}(v_k, v_{k'}) = \text{Cov}\{Z(v_k), Z(v_{k'})\}$ and involves averaging $C(\mathbf{h})$ computed between any two points discretizing the blocks v_k and $v_{k'}$. A major difference between AAP kriging and the related algorithms (area-to-area and area-to-point kriging), introduced recently in the geostatistical literature, is the availability of point data here. Thus, the point support semivariogram can be inferred directly from the observations without any need for a deconvolution of the areal semivariogram (Goovaerts 2008). The prediction variance associated with the AAP kriging estimator is computed as

$$\sigma_{AAP}^2(\mathbf{u}) = C(0) - \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_{\alpha}(\mathbf{u}) C(\mathbf{u}_{\alpha}, \mathbf{u}) - \sum_{k=n(\mathbf{u})+1}^{n(\mathbf{u})+K} \lambda_k(\mathbf{u}) \bar{C}(v_k, \mathbf{u}) - \mu(\mathbf{u}). \quad (3)$$

3.2 Exactitude Property and Coherence Constraint

Because kriging is an exact interpolator, predicted values must honor both point and areal data. Prediction is typically conducted at the nodes of a regular grid for mapping purposes, hence the exactitude property entails that the averaging point estimates at the P_k nodes falling within any given entity v_k must yield the areal data $z(v_k)$

$$z(v_k) = \frac{1}{P_k} \sum_{s=1}^{P_k} z^*(\mathbf{u}_s). \quad (4)$$

Constraint (4) is satisfied if (1) the P_k interpolation nodes are also used as discretizing points for the area v_k , and (2) the same K areal data and $n(\mathbf{u})$ point data are used for the estimation of each of the point $\mathbf{u}_s$. This second condition is fulfilled by conducting the search for closest areal and point neighbors based on the distance to the centroid of the unit v_k . Beware that for small $n(\mathbf{u})$ this search strategy can create discontinuities away from the unit's

centroids (e.g. near the unit boundaries), since the interpolation might not be based on the closest observations.

3.3 Binomial Kriging

The application of AAP kriging to the medical geography case study must account for the fact that the K areal data have varying degrees of reliability. These observations are incidence rates that tend to become unstable when the denominator (that is, the number of cancer cases in this particular example) is small. On the other hand, point data can be viewed as an extreme case where the population size is one (individual-level data). The information about each cancer case, referenced geographically by its residence's spatial coordinates $\mathbf{u}_\alpha = (x_\alpha, y_\alpha)$, takes the form of an indicator of early/late-stage diagnosis

$$i(\mathbf{u}_\alpha) = \begin{cases} 1 & \text{if late-stage diagnosis,} \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

Poisson kriging (Goovaerts 2005; Kerry et al. 2010a) was recently introduced to account for the population denominator in the geostatistical processing of rate data, such as cancer mortality or crime rates. To acknowledge the fact that late-stage cancer diagnosis is not a rare event, the methodology was here modified by replacing the Poisson distribution by the binomial distribution in the derivation of the system of equations (Webster et al. 1994a; Walker et al. 2008). Under this model, the number of late-stage cases $d(v_i)$ is interpreted as a realization of a random variable $D(v_i)$ that follows a binomial distribution with two parameters: the population size $n(v_i)$ and the local risk $Y(v_i)$

$$D(v_i) | Y(v_i) \sim \text{Bi}[Y(v_i), n(v_i)], \quad i=1, \dots, N, \quad (6)$$

where $Y(v_i)$ has a mean m and a variance $\sigma_Y^2 = C_Y(0)$. Given the risk value $Y(v_i)$, the count variables $D(v_i)$ are assumed to be conditionally independent. The conditional mean and variance of the rate variable $Z(v_i)$ are defined as

$$E[Z(v_i) | Y(v_i)] = Y(v_i), \quad (7)$$

$$\text{Var}[Z(v_i) | Y(v_i)] = m(1-m)/n(v_i). \quad (8)$$

Following Monestiez (personal communication) and Webster et al. (1994a), the unconditional mean and variance are as follows

$$E[Z(v_i)] = m, \quad (9)$$

$$\text{Var}[Z(v_i)] = \frac{m(1-m) + (n(v_i) - 1)\sigma_Y^2}{n(v_i)}. \quad (10)$$

In addition, $C_Y(\mathbf{h}) = C_Z(\mathbf{h})$ when individual-level indicator data (5) are available, which is the case here since the residence of all cancer patients is known.

The area-and-point (AAP) kriging estimate is now expressed as a linear combination of point indicator data and areal incidence rates

$$z_{AAP}^*(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_{\alpha}(\mathbf{u}) i(\mathbf{u}_{\alpha}) + \sum_{k=n(\mathbf{u})+1}^{n(\mathbf{u})+K} \lambda_k(\mathbf{u}) z(v_k), \quad (11)$$

The kriging weights are the solution of the following system of linear equations (Webster et al. 1994a)

$$\begin{aligned} \sum_{j=1}^{n(\mathbf{u})+K} \lambda_j(\mathbf{u}) \left[\bar{C}_l(x_i, x_j) + \delta_{ij} \frac{a}{n(v_i)} \right] + \mu(\mathbf{u}) &= \bar{C}_l(x_i, \mathbf{u}), \quad i=1, \dots, n(\mathbf{u})+K, \\ \sum_{j=1}^{n(\mathbf{u})+K} \lambda_j(\mathbf{u}) &= 1. \end{aligned} \quad (12)$$

where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise, $a = m^*(1 - m^*) - \bar{C}_l(v_i, v_i)$, and m^* is the population-weighted mean of the N rates ($N = 83$ census tracts here). The addition of the error variance term $a/n(v_i)$ for a zero distance accounts for variability arising from population size, leading to smaller weights for less reliable incidence rates based on fewer cases. Note the following:

- (1) The error variance term a is zero for point indicator data; the number of cases $n(v_i)$ is one, hence $\bar{C}_l(v_i, v_i) = \sigma_Y^2 = m^*(1 - m^*)$ and $a = \bar{C}_l(v_i, v_i) - \bar{C}_l(v_i, v_i)$.
- (2) Under the simplifying assumption that all census tracts have the same size and shape and $\bar{C}_l(v_i, v_i) = \sigma_Y^2$ (Oliver et al. 1998), the term $\bar{C}_l(v_i, v_i) + a/n(v_i)$ actually corresponds to the unconditional variance of the rate variable (10).
- (3) If all tracts can be assimilated with points and include only one case (that is, individual-level indicator data), then the error variance term a is always zero (see item #1), making the binomial kriging system (12) identical to an indicator kriging system (Journal 1983).
- (4) If the proportion of late-stage cases m^* becomes very small (i.e. rare disease) and the number of cancer cases is much larger than 1 in every census tract, then the variance (10) becomes $\sigma_Y^2 + m/n(v_i)$, making the binomial and Poisson kriging systems identical.
- (5) The coherence constraint (4) is not satisfied by the binomial kriging estimates because of the addition of the variance error term to the diagonal of the kriging matrix. Indeed, binomial kriging aims to filter the noise attached to the areal data (that is, unstable incidence rates for census tracts with few cancer cases); hence, these areal data should not be reproduced. In other words, the exactitude property applies to the point data for which the error variance term is zero but not to the areal data.

The prediction variance associated with the binomial kriging estimator is computed as

$$\sigma_{AAP}^2(\mathbf{u}) = C(0) - \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_{\alpha}(\mathbf{u}) C_l(\mathbf{u}_{\alpha}, \mathbf{u}) - \sum_{k=n(\mathbf{u})+1}^{n(\mathbf{u})+K} \lambda_k(\mathbf{u}) \bar{C}_l(v_k, \mathbf{u}) - \mu(\mathbf{u}). \quad (13)$$

4 Case Study

4.1 Mapping Chromium Concentrations and Late-Stage Cancer Incidence Rates

Figure 2 (left column) shows the maps of chromium concentration estimated on a 25-meter spacing grid using alternative interpolation techniques. The reference approach is ordinary kriging (OK) that uses only the 369 field data (Fig. 2A). The other two maps incorporate areal data that take the form of average chromium concentration per geological mapping unit. These concentrations, ranging from 30.37 mg/kg (Argovian formation) to 39.25 mg/kg (Sequanian formation), were used either as local means in residual kriging or directly incorporated into the AAP estimator (1). In the first case, the average concentrations were subtracted from the 369 point data, and the residuals were interpolated using simple kriging and the residual semivariogram model of Fig. 3A. The final map (Fig. 2E) was obtained as the sum of the trend model and the kriged residuals. In the second case, the estimation was based on the 16 closest point data and the areal data that are first and second order neighbors of the kernel mapping unit v_k that includes the estimation grid node. To compute the area-to-area and area-to-point covariances in AAP kriging system (2), each mapping unit was discretized using the interpolation grid nodes; this ensures that the kriged estimates in Fig. 2C honor the coherence constraint, as demonstrated by the scatter plot in Fig. 4C.

As expected, the Cr residual semivariogram has a lower sill than the original semivariogram since a small part of the total variance is captured by the trend model. The residual semivariogram model has also a much shorter range, leading to bull's-eye effect around sample points in the map created by residual kriging (Fig. 2E). In contrast, the AAP map (Fig. 2C) is much smoother and clearly displays the lower concentrations expected on the Argovian formation, in particular when compared with the ordinary kriging map. Differences between the three maps are the largest in sparsely sampled areas where the choice of a trend model becomes preponderant (Goovaerts 1997). In particular, incorporating the geological information leads to smaller estimates on the section of Argovian formation where no sample was collected (dashed circle in Fig. 2C) and in a small Argovian mapping unit that must satisfy the coherence constraint despite the presence of larger concentrations recorded in the field (solid circle in Fig. 2C). The scatter plot in Fig. 4E indicates that residual kriging (RK) leads to a better reproduction of the areal data than ordinary kriging (Fig. 4A). Yet, there is no guarantee that the coherence constraint will be honored once the kriged residuals are added to the trend estimates. Such a constraint can only be imposed through the joint incorporation of field point data and areal map data by AAP kriging (Fig. 4C). Although the map of AAP kriged estimates is smoother than the OK and RK maps when all 359 observations are used, cross-validation studies in Fig. 7 will demonstrate the opposite for lower sampling densities.

A similar analysis was conducted for the health outcome data displayed in Figs. 1B–D. To account for the wide range of separation distances between cancer cases (from a few metres to 112 km), the indicator semivariogram in Fig. 3B was computed using two series of lag classes: 37 lags of 30 metres to characterize the small-scale variation of the data and 37 lags of 800 metres to look at the regional pattern. The spatial variability is clearly nested: most of the variance occurs over short distances (first range = 397 metres, see detailed view in Fig. 3C) and is superimposed on a regional structure with a range of 19.57 km. Late detection cases do not occur randomly in space, yet individual-level factors such as age or family history generate a large variability over short distances. Subtracting the census tract mean from indicator values slightly reduces the sill of the semivariogram, yet it has no impact over shorter distances (e.g. less than 1 km) which are smaller than the average dimension of the tracts.

All incidence maps were created using the 32 closest point indicator data and, for AAP kriging, the rates recorded in census tracts that share a boundary or vertex with the tract, including the interpolation node (first-order adjacency). Incorporating census-tract information through residual kriging adds more details to the map but generates discontinuities at the tract boundaries. On the other hand, accounting for adjacent areal data in AAP kriging leads to a map that has more compact spatial features, yet a larger variance, than the indicator kriging map. Scatter plots in Fig. 4 (right column) compare the average of kriged estimates within each census tract with the incidence rates that were noise-filtered by area-to-area binomial kriging (Goovaerts 2009a). Unlike for the Jura data set, ordinary kriging of late-stage diagnosis indicators leads to a better reproduction of areal data than residual kriging. Because of the greater compactness of census tracts relative to the intricate arrangement of geological formations, point data used in ordinary kriging are more likely to belong to the same areal unit for the health data set than the soil data set. In addition, the local means used in residual kriging ignore the small-number problem; differences displayed in Fig. 4F reflect the rate uncertainty.

4.2 Kriging Weights and Kriging Variances

To better understand the relative contribution of point data versus areal data to the computation of the AAP kriging estimator, three types of statistics were computed at each interpolation grid node: sum of weights of point data, weight of the kernel areal datum (which includes that grid node), and sum of weights of other areal data. General conclusions can be drawn from the maps of kriging weights displayed in Fig. 5: (1) the contribution of point data is the largest around clusters of field observations (A, B); (2) the kernel areal datum receives a larger weight in the vicinity of the area centroid (that is, the centre of geographical units) and in sparsely sampled areas, such as along the edges of the study area (C, D); (3) the other areal data (that is, the adjacent mapping units) receive much smaller weights on average and their contribution increases at the edges of the mapping units, in particular for larger kernel units when the distance to the centroid will tend to be the largest (E, F). Results for the two data sets differ, however, in terms of sign and magnitude of kriging weights.

For the soil data set, the small nugget effect and large range of the semivariogram model relative to the size of mapping units cause a strong screening effect by the kernel areal data that receive large positive weights: the average is 1 and the range is -0.33 to 2.56 . For example, dividing the range of autocorrelation by a factor of ten would lead to much smaller kernel weights ranging between -0.07 and 1.825 . On the other hand, the sums of weights for point and adjacent areal data average zero and are negative for 64 and 55% of the grid nodes, respectively. Because binomial kriging is not an exact interpolator for areal data (recall Sect. 3.3), the average of kriging weights across the study area is not one for kernel areal data and is not zero for point and adjacent areal data. In fact, the averages are positive for all three types of weight and are respectively 0.32, 0.43, and 0.25. The sums of weights for point and adjacent areal data are now negative for only 0.2 and 2% of the grid nodes, respectively.

Figure 6 shows the maps of prediction variance associated with the chromium concentration and late-stage cancer incidence maps of Fig. 2. The variance for residual kriging (Figs. 6E–F) was computed as the sum of the simple kriging variance for residuals' interpolation and the variance associated with the trend model (that is, the variance of the arithmetical average of observations within each mapping unit). For the soil data set, the prediction variance maps for OK and RK display similar spatial patterns: the variance is lower in the vicinity of point data and larger in extrapolation situation (e.g. along the edges of the study area). Again, the shorter range of the residual semivariogram model creates bull's-eye effect around sample points (Fig. 6E). When both point and areal data are used, the kriging

variance depends not only on the proximity to point data, but also the distance to the unit's centroids (Fig. 6C). Lower variances are thus observed in smaller and more compact mapping units, and in the vicinity of field samples. The AAP and ordinary kriging variances are on average similar across the study area, yet the maximum AAP variance is smaller: 112.8 versus 126.2 (mg/kg)².

The weaker spatial autocorrelation of health data, combined with the more uniform geometry of census tracts relative to soil mapping units, greatly reduces differences between OK and AAP kriging variances (Figs. 6B–D) compared to the soil data set (Figs. 6A–C). The two sets of kriging variances have the same magnitude, a smaller maximum for AAP kriging (0.138 versus 0.141), and they display a linear correlation coefficient of 0.93. The kriging variance is generally smaller in the proximity of cancer patients' residences. Like the predicted values in Fig. 2F, the map of residual kriging variance (Fig. 6F) shows strong discontinuities at the borders between census tracts. In fact, unlike the two other algorithms, the RK estimates and variances are strongly correlated. The linear correlation coefficient of 0.52 reflects the relationship between the mean and variance of indicator data (8), which are used as local trend in residual kriging.

4.3 Performance Comparison

Figures 2 through 6 provided useful information on the properties of the three types of kriging estimator: ordinary kriging, AAP kriging and residual kriging. However, since the true chromium concentration and incidence rate are unknown, one cannot conclude that one estimator outperforms the others. The prediction performances of the different interpolation techniques with respect to sampling density were investigated using the following procedure:

1. Select a random subset (prediction set) of n field data with $30 \leq n \leq 255$ for the soil data set and $100 \leq n \leq 850$ for the cancer data set. For each of the 16 sampling intensities, 100 different random subsets were selected to account for sampling fluctuations.
2. For each random subset and sampling intensity,
 - Predict the heavy metal concentration or risk of late-stage diagnosis at the $N = 369 - n$ or $N = 937 - n$ remaining locations (validation set) using each of the three algorithms and the same search strategy applied in Fig. 2, except that a minimum separation distance of 125 metres was imposed on neighboring field data to avoid overly optimistic prediction accuracies for clustered observations. A coarser discretization geography (100-m instead of 25-m grid) was used for the soil data set to reduce the computational time required by each of the 1,600 simulation runs. In each case, the experimental semivariogram was computed and modeled using weighted least-square regression. For the cancer data set, census-tract incidence rates are estimated for each simulation using the n indicator data available.
 - Compute the mean absolute error (MAE) of prediction as

$$\text{MAE} = \frac{1}{N} \sum_{\alpha=1}^N |z^*(\mathbf{u}_{\alpha}) - z(\mathbf{u}_{\alpha})|. \quad (14)$$

- Compute the mean square standardized residual (MSSR) as

$$MSSR = \frac{1}{N} \sum_{\alpha=1}^N \left[\frac{z^*(\mathbf{u}_{\alpha}) - z(\mathbf{u}_{\alpha})}{\sigma_k(\mathbf{u}_{\alpha})} \right]^2. \quad (15)$$

If the actual estimation error is equal, on average, to the error predicted by the model, the MSSR statistic should be about one (Wackernagel 1998). To penalize equally the over- and under-estimation of the prediction errors by the kriging variance, the inverse of MSSR was considered if the quantity (15) exceeded one.

- Compute the variance of the set of N heavy metal concentration estimates to assess the magnitude of the smoothing effect caused by each type of kriging algorithm.
- Compute the ability of interpolation methods to discriminate between early- and late-stage cancer diagnoses. The discriminatory power is measured by the ratio of averaged probability estimate at residences of early- and late-stage cases (Goovaerts 2009a).

Figure 7 indicates that over the 100 random subsets, the simultaneous incorporation of areal and point data using AAP kriging will most likely give the best results for all four performance criteria. The benefit is the most striking for the soil data set (Fig. 7, left column), in particular when less than 150 observations out of 359 are available. As the size of the point data set increases, the smoothing induced by ordinary kriging decreases (Fig. 7E) and both OK and AAP kriging yield similar mean absolute errors of prediction (Fig. 7A). Yet, the AAP kriging variance is clearly a more accurate indicator of the magnitude of prediction errors (Fig. 7C).

The impact of sampling density on results is the opposite for the cancer data set because the areal data (that is, late-stage cancer incidences) are computed directly from the point (individual-level) data. When these point data are few, census tract rates tend to be unreliable and their incorporation in the interpolation by AAP kriging leads to worse predictions than ordinary kriging that ignores this information. As more individual-level data become available, the probabilities estimated by AAP kriging become more accurate (Fig. 7B), leading to a better discrimination between residences of patients being diagnosed late or early (Fig. 7F). Once again, the AAP kriging variance is the best predictor of the actual magnitude of prediction errors (Fig. 7D).

5 Conclusions

The ability to combine data measured at various scales and over different spatial supports in kriging is becoming a pressing need, especially as the field of geostatistical applications now encompasses social and health sciences. Whereas the first analytical developments of kriging clearly demonstrated its flexibility to accommodate different measurement and prediction supports, geostatistical analysis of a mixture of point data and irregular blocks has rarely been implemented in practice, mainly because of its lack of application in mining. Joint advances in GIS software and computational resources now allow the application of kriging to the complex geographies found in social and health sciences. In addition, the recent development of binomial and Poisson kriging allows one to take into account both the spatial extent of the geographical unit and the size of the population under study within that unit (that is, the number of breast cancer cases) in the interpolation.

Geostatistics provides a framework to model the spatial correlation among continuous attributes measured over irregular geographic supports, allowing the mapping of the distribution of attribute values within each mapping unit while fulfilling the coherence constraint. Unlike traditional point kriging, the estimates and prediction variances computed using AAP kriging depend on both the distance to point data but also the proximity to the centroid and edges of geographical units. One advantage of area-and-point over area-to-area and area-to-point kriging is the knowledge of the point-support semivariogram model that can be inferred directly from point measurements, thereby eliminating the need for deconvolution procedures. This paper has shown how to implement AAP kriging within the binomial framework and demonstrated its relationship with indicator kriging and Poisson kriging. Binomial kriging, like Poisson kriging, is a noise-filtering algorithm; hence, by construction, areal data are not honored by kriging estimates.

Sensitivity analysis conducted for contrasted geographies (nested irregular geological formations versus semi-regular polygonal census tracts) and types of kriging (traditional versus binomial) demonstrated the overall better prediction performance of AAP kriging over ordinary kriging and residual kriging with the choropleth-map trend model. In particular, when sampling is sparse, these two case studies indicated that incorporation of areal data tends to improve the prediction accuracy while the exactitude property of areal data decreases the smoothness of interpolated surfaces. Similar results were also obtained by Kerry et al. (2010b) using areal data that were derived independently of the point data.

This paper covered the situation where areal data provide a spatially exhaustive coverage of the study area, for example, a map of soil classes or administrative units. However, the methodology is completely general and can be applied to a mixture of point data and spatially disconnected areal data. The approach can also be implemented within a stochastic simulation framework. Liu and Journel (2009) recently introduced a software package that uses either direct sequential simulation or error simulation to account for both point and block support data in stochastic simulation. Their algorithms are capable of handling different block geometries and different layouts of block data, including irregular-shaped and overlapping blocks.

Acknowledgments

This research was funded by grants R43-CA135814-01 and R44-CA132347-02 from the National Cancer Institute. The views stated in this publication are those of the author and do not necessarily represent the official views of the NCI. The author thanks two anonymous reviewers for their very pertinent comments.

References

- Atteia O, Dubois J-P, Webster R. Geostatistical analysis of soil contamination in the Swiss Jura. *Environ Pollut* 1994;86:315–327. [PubMed: 15091623]
- Barry J, Breen N. The importance of place of residence in predicting late-stage diagnosis of breast or cervical cancer. *Heal Place* 2005;11:15–29.
- Best NG, Richardson S, Thomson A. A comparison of Bayesian spatial models for disease mapping. *Stat Meth Med Res* 2005;14:35–59.
- Goovaerts, P. Geostatistics for natural resources evaluation. Oxford University Press; New York: 1997.
- Goovaerts P. Geostatistical analysis of disease data: estimation of cancer mortality risk from empirical frequencies using Poisson kriging. *Int J Heal Geogr* 2005;4:31. doi:10.1186/1476-072X-4-31.
- Goovaerts, P. Spatial uncertainty in medical geography: a geostatistical perspective. In: Shekhar, S.; Xiong, H., editors. *Encyclopedia of GIS*. Springer; Berlin: 2007. p. 1106-1112.
- Goovaerts P. Kriging and semivariogram deconvolution in presence of irregular geographical units. *Math Geosci* 2008;40:101–128.

- Goovaerts P. Combining area-based and individual-level data in the geostatistical mapping of late-stage cancer incidence. *Spat Spatio-tempor Epidemiol* 2009a;1:61–71.
- Goovaerts P. Medical geography: a promising field of application for geostatistics. *Math Geosci* 2009b;41:243–264.
- Goovaerts P. A coherent geostatistical framework for combining choropleth map and field data in the spatial interpolation of soil properties. *Eur J Soil Sci*. 2010 accepted.
- Goovaerts P, Journel AG. Integrating soil map information in modeling the spatial variation of continuous soil properties. *Eur J Soil Sci* 1995;46:397–414.
- Gotway CA, Young LJ. Combining incompatible spatial data. *J Am Stat Assoc* 2002;97(459):632–648.
- Gotway, CA.; Young, LJ. Change of support: an interdisciplinary challenge. In: Renard, Ph; Demougeot-Renard, H.; Froidevaux, R., editors. *geoENV V—Geostatistics for environmental applications*. Springer; Berlin: 2005. p. 1-13.
- Gotway CA, Young LJ. A geostatistical approach to linking geographically-aggregated data from different sources. *J Comput Graph Stat* 2007;16(1):115–135.
- Hengl T, Heuvelink GBM, Stein A. A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma* 2004;120:75–93.
- Journel AG. Nonparametric estimation of spatial distributions. *Math Geol* 1983;15(3):445–468.
- Journel, AG.; Huijbregts, CJ. *Mining geostatistics*. Academic Press; London: 1978.
- Kerry R, Goovaerts P, Haining V, Ceccato RP. Applying geostatistical analysis to crime data: car-related thefts in the Baltic States. *Geogr Anal* 2010a;42:53–75.
- Kerry, R.; Rawlins, BG.; Goovaerts, P. Area-to-point kriging of organic carbon and soil texture: an efficient use of legacy soil data from polygon maps for regional or national scale digital soil mapping. *Proceedings of 4th global workshop on digital soil mapping*; Rome, Italy. May 24–26, 2010; 2010b.
- Kyriakidis P. A geostatistical framework for area-to-point spatial interpolation. *Geogr Anal* 2004;36(2):259–289.
- Liu Y, Journel AG. A package for geostatistical integration of coarse and fine scale data. *Comput Geosci* 2009;35(3):527–547.
- Liu TL, Juang KW, Lee DY. Interpolating soil properties using kriging combined with categorical information of soil maps. *Soil Sci Soc Am J* 2006;70:1200–1209.
- Oliver MA, Webster R, Lajaunie C, Muir KR, Parkes SE, Cameron AH, Stevens MCG, Mann JR. Binomial co-kriging for estimating and mapping the risk of childhood cancer. *IMA J Math Appl Med* 1998;15(3):279–297.
- Rushton G, Peleg I, Banerjee A, Smith G, West M. Analyzing geographic patterns of disease incidence: Rates of late-stage colorectal cancer in Iowa. *J Med Syst* 2004;28(3):223–236. [PubMed: 15446614]
- Talbot TO, Kulldorff M, Forand SP, Haley VB. Evaluation of spatial filters to create smoothed maps of health data. *Stat Med* 2000;19:2399–2408. [PubMed: 10960861]
- Wackernagel, H. *Multivariate geostatistics*. 2nd completely revised edition. Springer; Berlin: 1998.
- Walker, E.; Monestiez, P.; Renard, D.; Bez, N. Kriging of the latent probability of a binomial variable: application to fish statistics. In: Ortiz, J.; Emery, X., editors. *Geostatistics 2008*. Santiago. GECAMIN; 2008. p. 981-990.
- Waller, LA.; Gotway, CA. *Applied spatial statistics for public health data*. Wiley; New Jersey: 2004.
- Webster R, Oliver MA, Muir KR, Mann JR. Kriging the local risk of a rare disease from a register of diagnoses. *Geogr Anal* 1994a;26:168–185.
- Webster R, Atteia O, Dubois J-P. Co-regionalization of trace metals in the soil in the Swiss Jura. *Eur J Soil Sci* 1994b;45:205–18.
- Young, JL., Jr; Roffers, SD.; Ries, LAG.; Fritz, AG.; Hurlbut, AA., editors. *SEER summary staging manual—2000: codes and coding instructions*. National Cancer Institute. National Institutes of Health; Bethesda, MD: 2001. Pub # 01-4969

Fig. 1.

Information available for mapping topsoil heavy metal concentration and late-stage breast cancer incidence. **(A)** Soil field measurements collected at 359 point locations. **(B)** Location of 937 patient residences (data shuffled for confidentiality reasons). **(C)** Choropleth map of the main geological formations. **(D)** Choropleth map of late-stage breast-cancer incidence rate (age group 64–75 years) in three Michigan counties, by census tract, 1985–2002

Fig. 2.

Maps of chromium concentration and late-stage breast-cancer incidence rate created by alternative interpolation techniques. (A, B) Ordinary kriging. (C, D) Kriging that combines both point and areal data. (E, F) Residual kriging with a choropleth trend model. The same color scale is used for each series of three maps

Fig. 3. Semivariogram models used for geostatistical interpolation. **(A)** Models fitted to semivariograms of chromium concentration before and after subtraction of trend estimates (i.e. residuals). Models fitted to indicator semivariograms of health outcomes before and after subtraction of trend estimates: entire model **(B)** and detailed view of first few lags **(C)**

Fig. 4. Scatter plots of areal chromium concentrations and ATA kriged incidence rates versus averages of kriging estimates within each mapping unit (i.e. geological formation or census tract). Only area-and-point kriging ensures the reproduction of areal data (**C, D**), while other interpolation techniques do not honor the coherence constraint

Fig. 5.

Maps of weights assigned to different types of data in the interpolation of chromium concentration and late-stage cancer incidences using area-and-point kriging in Figs. 3C–D. (A, B) Point data (sum for 16 Cr observations or 32 late-stage indicators). (C, D) Kernel areal datum. (E, F) Neighboring areal data (second-order adjacency for soil data and first-order adjacency for cancer data). *Open circles* denote the location of point data while *open triangles* in (E) and (F) correspond to geographical centroids of areal data. The same color scale is used for all six maps

Fig. 6. Maps of prediction variance associated with the maps of Fig. 2. (A, B) Ordinary kriging. (C, D) Kriging that combines both point and areal data. (E, F) Residual kriging with a choropleth trend model. The same color scale is used for each series of three maps

Fig. 7.

Impact of the sample size on the proportion of simulations for which an interpolation technique outperforms the other two. Cross-validation statistics include: the mean absolute error of prediction, the mean square standardized residual, the variance of the set of kriged estimates (smoothness), and the discriminatory power between early- and late-stage cases