Merging Areal and Point Data in Medical Geography and Soil Mapping

P. Goovaerts

BioMedware, 3526 W Liberty, Suite 100, Ann Arbor, MI 48103
Telephone: 001-734-913-1098 (ext. 202)
Fax: 001-734-913-2201
Email: Goovaerts@biomedware.com

1. Introduction

A common issue in spatial interpolation is the combination of data measured over different spatial supports. For example, in the field of medical geography (Goovaerts, 2009) information available for mapping disease risk typically includes point data (e.g. patients residence) and aggregated data (e.g. socio-demographic and economic data at the census track level). Similarly, soil measurements recorded at discrete locations on the ground 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 (Goovaerts, 2011). This paper presents a coherent geostatistical approach to accommodate both areal and point data in the spatial interpolation of continuous attributes. The procedure is illustrated using two datasets: 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 in Michigan for the period 1985-2002 (Figure 1).

2. Methodology

2.1 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 set of point data collected at n discrete locations $\mathbf{u}_\alpha$ $\{z(\mathbf{u}_\alpha);\ \alpha=1,\ldots,n\}$, supplemented by a set of B areal data $\{z(v_k);\ k=1,\ldots,B\}$ recorded for mapping units v_k of various size and shape. 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.

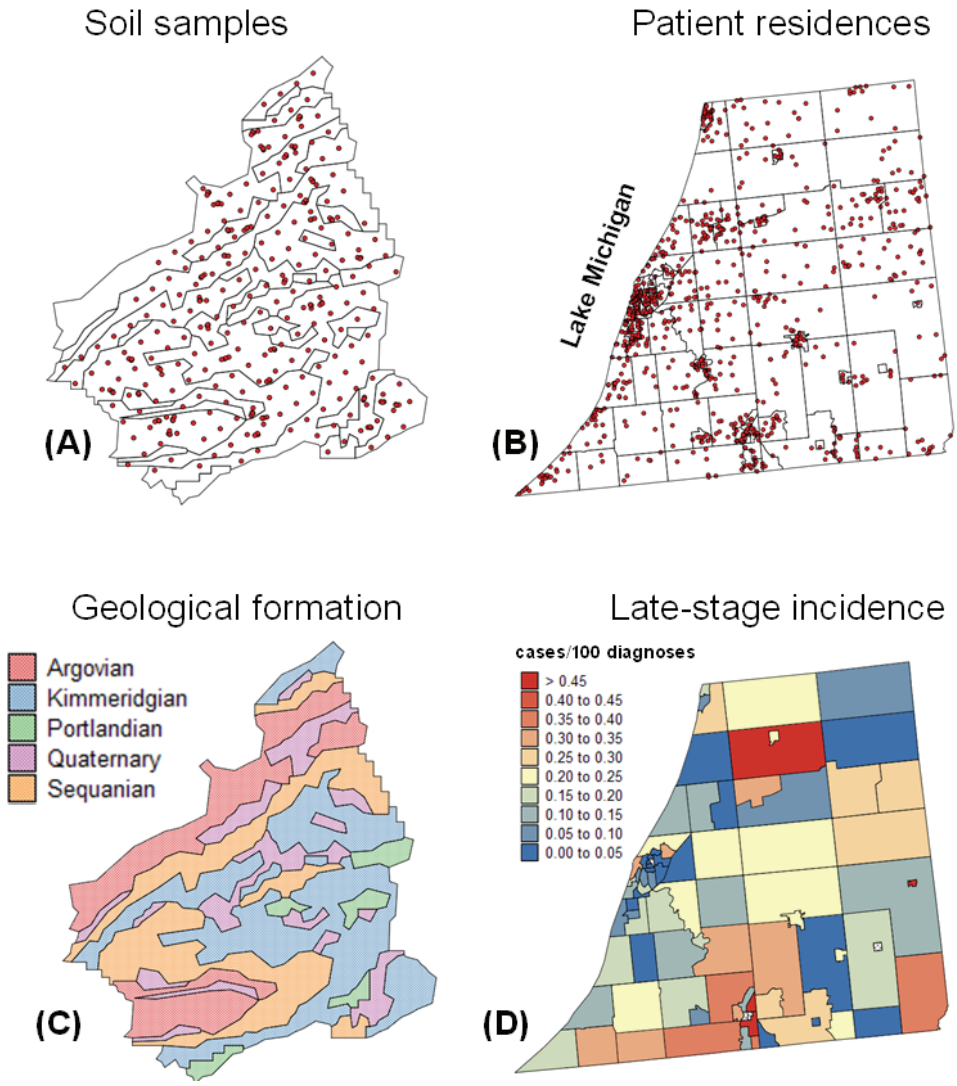

Figure 1. Information available for mapping topsoil heavy metal concentration and late-stage breast cancer incidence. (A) Soil field measurements. (C) Choropleth map of the main geological formations. (B) Location of 937 patient residences. (D) Choropleth map of late-stage breast cancer incidence rate in three Michigan counties, by census tract.

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

$$\sum_{j=1}^{n(\mathbf{u})+K} \lambda_j(\mathbf{u})\,\overline{C}(x_i,x_j) + \mu(\mathbf{u}) = \overline{C}(x_i,\mathbf{u}) \qquad \mathrm{i} = 1,\ldots,\mathrm{n}(\mathbf{u})+K$$
$$\sum_{j=1}^{n(\mathbf{u})+K} \lambda_j(\mathbf{u}) = 1. \qquad (2)$$

where $\mu(\mathbf{u})$ is the Lagrange multiplier, and $x_i=\mathbf{u}_\mathrm{i}$ if $i\leq n(\mathbf{u})$, and $x_i=v_\mathrm{i}$ otherwise. The quantity $\overline{C}(x_i,x_j)$ is 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 covariances $\overline{C}(v_k,\mathbf{u})$ are 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 covariances $\overline{C}(v_k, v_{k'}) = \text{Cov}\{Z(v_k), Z(v_{k'})\}$ 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 (Kyriakidis, 2004), 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).

2.2 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 (i.e. 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:

$$i(\mathbf{u}_\alpha) = \begin{cases} 1 \text{ if late-stage diagnosis} \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

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

$$z^*_{AAP}(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha(\mathbf{u}) i(\mathbf{u}_\alpha) + \sum_{k=n(\mathbf{u})+1}^{n(\mathbf{u})+K} \lambda_k(\mathbf{u}) z(v_k) \tag{4}$$

The kriging weights are the solution of the following system of linear equations (Webster *et al*., 1994; Goovaerts, 2010):

$$\begin{aligned} &\sum_{j=1}^{n(\mathbf{u})+K} \lambda_j(\mathbf{u}) \left[\overline{C}_I(x_i, x_j) + \delta_{ij} \frac{a}{n(v_i)} \right] + \mu(\mathbf{u}) = \overline{C}_I(x_i, \mathbf{u}) \qquad \text{i} = 1,...,\text{n}(\mathbf{u}) + K \\ &\sum_{j=1}^{n(\mathbf{u})+K} \lambda_j(\mathbf{u}) = 1. \end{aligned} \tag{5}$$

where δ_{ij}=1 if i=j and 0 otherwise, $a = m^*(1-m^*) - \overline{C}_I(v_i, v_i)$, $C_I(\mathbf{h})$ is an indicator covariance function, 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.

3. Results and Discussion

Figure 2 (left column) shows the maps of chromium concentration estimated using alternative interpolation techniques. The reference approach is ordinary kriging (OK) that uses only 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 were used either as local means in residual kriging or directly incorporated

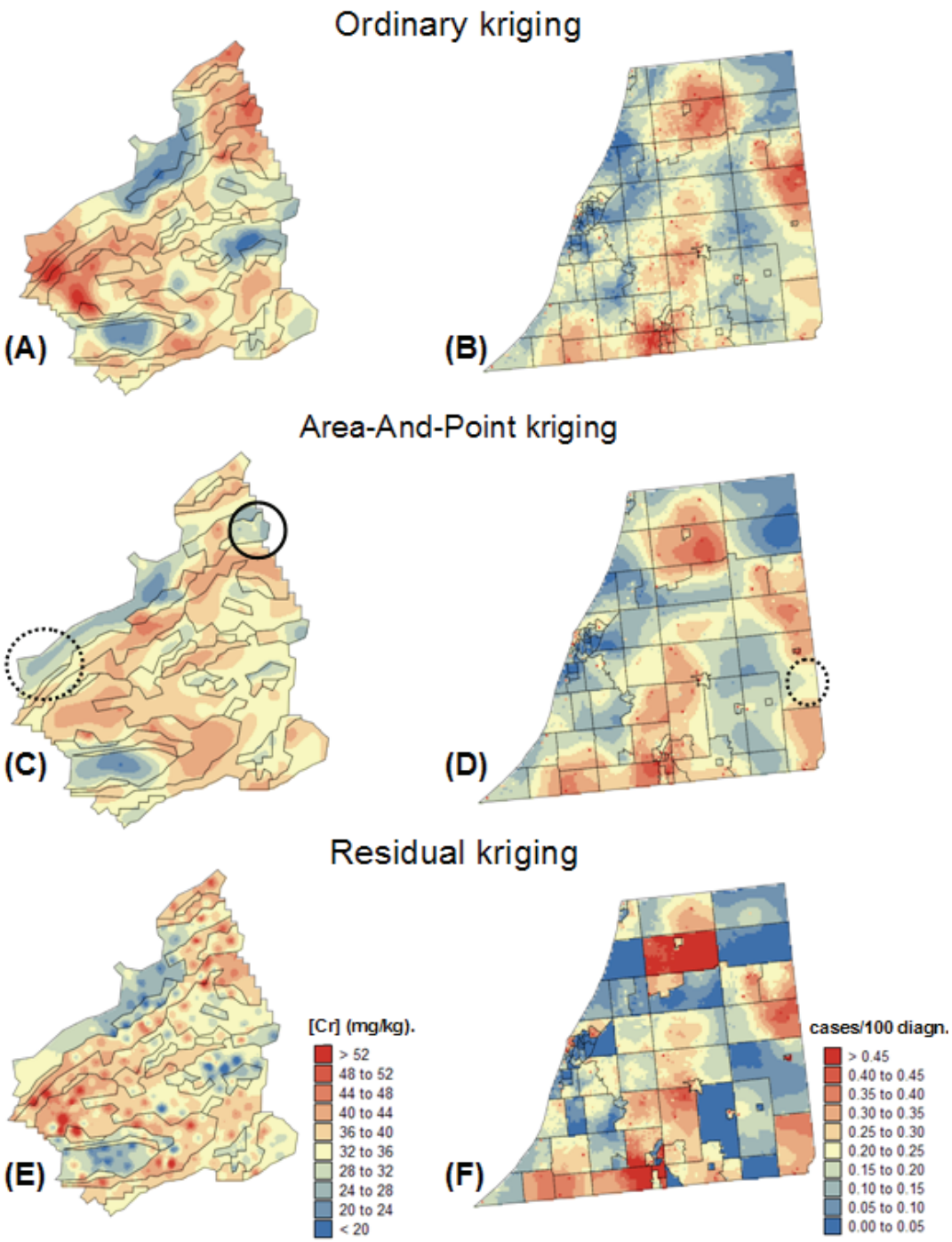

Figure 2. Maps of chromium concentration and late-stage breast cancer incidence rate created by alternative interpolation techniques. (A,B) Ordinary kriging. (C,D) Kriging that combines both areal and point data "AAP kriging". (E,F) Residual kriging with a choropleth trend model. The same color scale is used for each series of three maps.

into the Area-And-Point estimator. In the later case, the average of kriged estimates equals the mapping units' mean (coherence constraint). The residual semivariogram model has a short 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. Differences between the three maps are the largest in sparsely sampled areas where the choice of a trend model becomes preponderant. 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 sampled concentrations (solid circle in Fig. 2C).

A similar analysis was conducted for the health outcome data in Figs. 1B-D. 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 (1st 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 with more compact spatial features than the indicator kriging map.

The performance of the proposed approach, relatively to ordinary kriging or a traditional residual kriging with choropleth map trend model (e.g. constant value within each polygon), was assessed using jackknife. Performance criteria included 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. Results (Goovaerts, 2010) demonstrated the overall better prediction performance of AAP kriging over ordinary kriging and residual kriging. In particular when sampling is sparse, incorporation of areal data improves the prediction accuracy while the exactitude property of areal data decreases the smoothness of interpolated surfaces.

4. Conclusions

The ability to combine data measured at various scales and over different spatial supports in kriging is becoming a pressing need, in particular 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 (Goovaerts, 2009). 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 (i.e. number of breast cancer cases) in the interpolation.

5. Acknowledgements

This research was funded by grants R43CA150496-01 and R44CA132347-02 from the National Cancer Institute. The views stated in this publication are those of the authors and do not necessarily represent the official views of the NCI.

6. References

Kyriakidis P, 2004, A geostatistical framework for area-to-point spatial interpolation. *Geographical Analysis*, 36,259-289.

Goovaerts P, 2008, Kriging and semivariogram deconvolution in presence of irregular geographical units. *Mathematical Geosciences,* 40:101-128.

Goovaerts P, 2009, Medical geography: a promising field of application for geostatistics. *Mathematical Geosciences,* 41:243-264.

Goovaerts P, 2010, Combining areal and point data in geostatistical interpolation: Applications to soil science and medical geography. *Mathematical Geosciences*, 42:535-554.

Goovaerts P, 2011, A coherent geostatistical approach for combining choropleth map and field data in the spatial interpolation of soil properties. *European Journal of Soil Science*, in press.

Webster R, Oliver MA, Muir KR and Mann JR, 1994, Kriging the local risk of a rare disease from a register of diagnoses. *Geographical Analysis*, 26:168-185.