Geostatistical Analysis of Health Data: State-of-the-Art and Perspectives

P. Goovaerts

Abstract The analysis of health data and putative covariates, such as environmental, socio-economic, behavioral or demographic factors, is a promising application for geostatistics. It presents, however, several methodological challenges that arise from the fact that data are typically aggregated over irregular spatial supports and consist of a numerator and a denominator (i.e. population size). This paper presents an overview of recent developments in the field of health geostatistics, with an emphasis on three main steps in the analysis of aggregated health data: estimation of the underlying disease risk, detection of areas with significantly higher risk, and analysis of relationships with putative risk factors. The analysis is illustrated using age-adjusted cervix cancer mortality rates recorded over the 1970–1994 period for 118 counties of four states in the Western USA. Poisson kriging allows the filtering of noisy mortality rates computed from small population sizes, enhancing the correlation with two putative explanatory variables: percentage of habitants living below the federally defined poverty line, and percentage of Hispanic females. Area-to-point kriging formulation creates continuous maps of mortality risk, reducing the visual bias associated with the interpretation of choropleth maps. Stochastic simulation is used to generate realizations of cancer mortality maps, which allows one to quantify numerically how the uncertainty about the spatial distribution of health outcomes translates into uncertainty about the location of clusters of high values or the correlation with covariates. Last, geographically-weighted regression highlights the non-stationarity in the explanatory power of covariates: the higher mortality values along the coast are better explained by the two covariates than the lower risk recorded in Utah.

1 Introduction

Since its early development for the assessment of mineral deposits, geostatistics has been used in a growing number of disciplines dealing with the analysis of data

P. Goovaerts
BioMedware, Inc. 516 North State Street, Ann Arbor, MI 48104-1236, USA
e-mail: goovaerts@biomedware.com

A. Soares et al. (eds.), *geoENV VI – Geostatistics for Environmental Applications*, 3–22

distributed in space and/or time. One field that has received little attention in the geostatistical literature 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). This lack of attention contrasts with the increasing need for methods to analyze health data following the emergence of new infectious diseases (e.g. West Nile Virus, bird flu), the higher occurrence of cancer mortality associated with longer life expectancy, and the burden of a widely polluted environment on human health.

Individual humans represent the basic unit of spatial analysis in health research. However, because of the need to protect patient privacy publicly available data are often aggregated to a sufficient extent to prevent the disclosure or reconstruction of patient identity. The information available for human health studies thus takes the form of disease rates, e.g. number of deceased or infected patients per 100,000 habitants, aggregated within areas that can span a wide range of scales, such as census units, counties or states. Associations can then be investigated between these areal data and environmental, socio-economic, behavioral or demographic covariates. Figure 1 shows an example of datasets that could support a study of the impact of demographic and socio-economic factors on cervix cancer mortality. The top map shows the spatial distribution of age-adjusted mortality rates recorded over the 1970-1994 period for 118 counties of four states in the Western USA. The corresponding population at risk is displayed in the middle map, either aggregated within counties or assigned to 25 km^2 cells. The bottom maps show two putative explanatory variables: percentage of habitants living below the federally defined poverty line, and percentage of Hispanic females. Indeed, Hispanic women tend to have elevated risk of cervix cancer, while poverty reduces access to health care and to early detection through the Pap smear test in particular (Friedell et al. 1992). These socio-demographic data are available at the census block level and were assigned to the nodes of a 5 km spacing grid for the purpose of this study (same resolution as the population map).

A visual inspection of the cancer mortality map conveys the impression that rates are much higher in the centre of the study area (Nye and Lincoln Counties), as well as in one Northern California county. This result must however be interpreted with caution since the population is not uniformly distributed across the study area and rates computed from sparsely populated counties tend to be less reliable, an effect known as "small number problem" and illustrated by the top scattergram in Fig. 1. The use of administrative units to report the results (i.e. counties in this case) can also bias the interpretation: had the two counties with high rates been much smaller in size, these high values likely would have been perceived as less problematic. Last, the mismatch of spatial supports for cancer rates and explanatory variables prevents their direct use in the correlation analysis.

Unlike datasets typically analyzed by geostatisticians, the attributes of interest are here measured exhaustively. Ordinary kriging, the backbone of any geostatistical analysis, thus seems of little use. Yet, I see at least three main applications of geostatistics for the analysis of such aggregated data:

1. Filtering of the noise caused by the small number problem using a variant of kriging with non-systematic measurement errors.
2. Modeling of the uncertainty attached to the map of filtered rates using stochastic simulation, and propagation of this uncertainty through subsequent analysis, such as the detection of aggregate of counties (clusters) with significantly higher or lower rates than neighboring counties.
3. Disaggregation of county-level data to map cancer mortality at a resolution compatible with the measurement support of explanatory variables.

Goovaerts (2005a, 2006a,b) introduced a geostatistical approach to address all three issues and compared its performances to empirical and Bayesian methods which have been traditionally used in health science. The filtering method is based on Poisson kriging and semivariogram estimators developed by Monestiez et al. (2006) for mapping the relative abundance of species in the presence of spatially heterogeneous observation efforts and sparse animal sightings. Poisson kriging was combined with p-field simulation to generate multiple realizations of the spatial distribution of cancer mortality risk. A limitation of all these studies is the assumption that the size and shape of geographical units, as well as the distribution of the population within those units, are uniform, which is clearly inappropriate in the example of Fig. 1. The last issue of change of support was addressed recently in the geostatistical literature (Gotway and Young 2002, 2005; Kyriakidis 2004). In its general form kriging can accommodate different spatial supports for the data and the prediction, while ensuring the coherence of the predictions so that disaggregated estimates of count data are non-negative and their sum is equal to the original aggregated count. The coherence property needs however to be tailored to the current situation where aggregated rate data have various degree of reliability depending on the size of the population at risk (Goovaerts, 2006b).

This paper discusses how geostatistics can benefit three main steps of the analysis of aggregated health data: estimation of the underlying disease risk, detection of areas with significantly higher risk, and analysis of relationships with putative risk factors. An innovative procedure is proposed for the deconvolution of the semivariogram of aggregated rates and the disaggregation of these rates, accounting for heterogeneous population densities and the shape and size of administrative units. The different concepts are illustrated using the cervix cancer data of Fig. 1.

2 Estimating Mortality Risk from Observed Rates

For a given number N of entities v_α (e.g. counties), denote the observed mortality rates as $z(v_\alpha) = d(v_\alpha)/n(v_\alpha)$, where $d(v_\alpha)$ is the number of recorded mortality cases and $n(v_\alpha)$ is the size of the population at risk. Let us assume for now that all entities v_α have similar shapes and sizes, with a uniform population density. These entities can thus be referenced geographically by their centroids with the vector of spatial

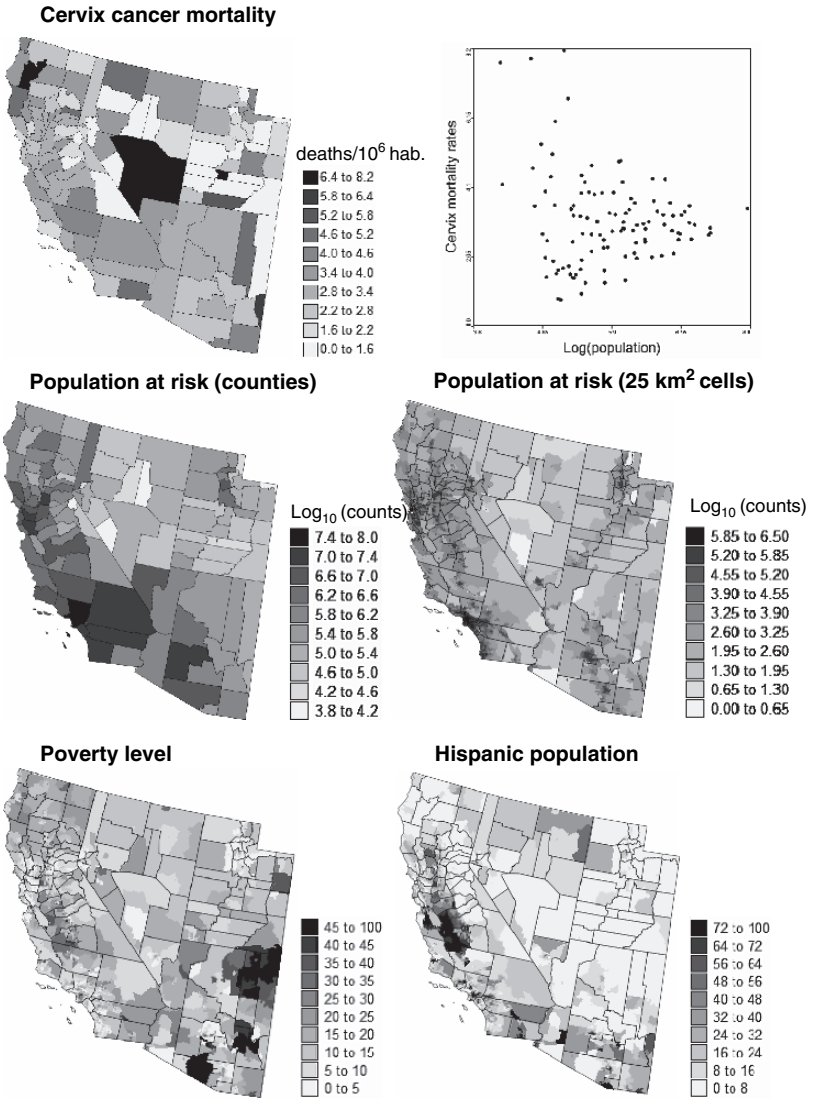


Fig. 1 Geographical distribution of cervix cancer mortality rates recorded for white females over the period 1970–1994, and the corresponding population at risk (aggregated within counties or assigned to 25 km^2 cells). Scatterplot illustrates the larger variance of rates computed from sparsely populated counties. Bottom maps show two putative risk factors: percentage of habitants living below the federally defined poverty line, and percentage of Hispanic females

coordinate's $\mathbf{u}_\alpha = (x_\alpha, y_\alpha)$. The disease count $d(\mathbf{u}_\alpha)$ is interpreted as a realization of a random variable $D(\mathbf{u}_\alpha)$ that follows a Poisson distribution with one parameter (expected number of counts) that is the product of the population size $n(\mathbf{u}_\alpha)$ by the local risk $R(\mathbf{u}_\alpha)$, see Goovaerts (2005a) for more details.

In Poisson kriging (PK), the risk over a given entity v_α is estimated as a linear combination of the kernel rate $z(\mathbf{u}_\alpha)$ and the rates observed in (K-1) neighboring entities:

$$\hat{r}_{PK}(\mathbf{u}_\alpha) = \sum_{i=1}^{K} \lambda_i(\mathbf{u}_\alpha) z(\mathbf{u}_i) \tag{1}$$

where $\lambda_i(\mathbf{u}_\alpha)$ is the weight assigned to the rate $z(\mathbf{u}_\mathrm{i})$ when estimating the risk at $\mathbf{u}_\alpha$. The K weights are the solution of the following system of linear equations:

$$\begin{aligned} &\sum_{j=1}^{K} \lambda_j(\mathbf{u}_\alpha)\left[C_R(\mathbf{u}_i - \mathbf{u}_j) + \delta_{ij}\tfrac{m*}{n(\mathbf{u}_i)}\right] + \mu(\mathbf{u}_\alpha) = C_R(\mathbf{u}_i - \mathbf{u}_\alpha) \quad \mathrm{i} = 1,\ldots,K \\ &\sum_{j=1}^{K} \lambda_j(\mathbf{u}_\alpha) = 1 \end{aligned} \tag{2}$$

where δ_{ij}=1 if $\mathbf{u}_\mathrm{i}$=$\mathbf{u}_\mathrm{j}$ and 0 otherwise, and $m*$ is the population-weighted mean of the N rates. The addition of an "error variance" term, $m*/n(\mathbf{u}_\mathrm{i})$, for a zero distance accounts for variability arising from population size, leading to smaller weights for less reliable data (i.e. measured over smaller populations). The prediction variance associated with the estimate (1) is computed using the traditional formula for the ordinary kriging variance:

$$\sigma^2_{PK}(\mathbf{u}_\alpha) = C_R(0) - \sum_{i=1}^{K} \lambda_i(\mathbf{u}_\alpha) C_R(\mathbf{u}_i - \mathbf{u}_\alpha) - \mu(\mathbf{u}_\alpha) \tag{3}$$

The computation of kriging weights and kriging variance (Equations (2) and (3)) requires knowledge of the covariance of the unknown risk, $C_R(\mathbf{h})$, or equivalently its semivariogram $\gamma_R(\mathbf{h})$=$C_R(0)$- $C_R(\mathbf{h})$. Following Monestiez et al. (2006) the semivariogram of the risk is estimated as:

$$\hat{\gamma}_R(\mathbf{h}) = \frac{1}{2\sum_{\alpha=1}^{N(\mathbf{h})} \frac{n(\mathbf{u}_\alpha)n(\mathbf{u}_\alpha+\mathbf{h})}{n(\mathbf{u}_\alpha)+n(\mathbf{u}_\alpha+\mathbf{h})}} \sum_{\alpha=1}^{N(\mathbf{h})} \left\{ \frac{n(\mathbf{u}_\alpha)n(\mathbf{u}_\alpha+\mathbf{h})}{n(\mathbf{u}_\alpha)+n(\mathbf{u}_\alpha+\mathbf{h})} [z(\mathbf{u}_\alpha) - z(\mathbf{u}_\alpha+\mathbf{h})]^2 - m* \right\} \tag{4}$$

where the different pairs $[z(\mathbf{u}_\alpha) - z(\mathbf{u}_\alpha+\mathbf{h})]$ are weighted by the corresponding population sizes to homogenize their variance.

2.1 Area-to-Area (ATA) Poisson Kriging

In the situation where the geographical entities have very different shapes and sizes, areal data can not be simply collapsed into their respective polygon centroids. Following the terminology in Kyriakidis (2004), ATA kriging refers to the case where

both the prediction and measurement supports are blocks (or areas) instead of points. The PK estimate (1) for the areal risk value $r(v_\alpha)$ thus becomes:

$$\hat{r}_{PK}(v_\alpha) = \sum_{i=1}^{K} \lambda_i(v_\alpha) z(v_i) \tag{5}$$

The Poisson kriging system (2) is now written as:

$$\begin{aligned} &\sum_{j=1}^{K} \lambda_j(v_\alpha) \left[\bar{C}_R(v_i, v_j) + \delta_{ij} \frac{m*}{n(v_i)} \right] + \mu(v_\alpha) = \bar{C}_R(v_i, v_\alpha) \quad \text{i} = 1,\ldots,K \\ &\sum_{j=1}^{K} \lambda_j(v_\alpha) = 1. \end{aligned} \tag{6}$$

The main change is that point-to-point covariance terms $C_R(\mathbf{u}_i - \mathbf{u}_j)$ are replaced by area-to-area covariances $\bar{C}_R(v_i, v_j) = \text{Cov}\{Z(v_\text{i}), Z(v_\text{j})\}$. Like in the traditional block kriging, those covariances are approximated by the average of the point support covariance $C(\mathbf{h})$ computed between any two locations discretizing the areas v_i and v_j :

$$\bar{C}_R(v_i, v_j) = \frac{1}{\sum_{s=1}^{P_i} \sum_{s'=1}^{P_j} w_{ss'}} \sum_{s=1}^{P_i} \sum_{s'=1}^{P_j} w_{ss'} C(\mathbf{u}_s, \mathbf{u}_{s'}) \tag{7}$$

where P_i and P_j are the number of points used to discretize the two areas v_i and v_j, respectively. For the example of Fig. 1 a grid with a spacing of 5 km was overlaid over the study area, yielding a total of 11 to 2,082 discretizing points per county depending on its area. The high-resolution population map in Fig. 1 clearly shows the heterogeneous distribution of population within counties. To account for spatially varying population density in the computation of the area-to-area covariance, the weights $w_{ss'}$ were identified to the product of population sizes within the 25 km^2 cells centred on the discretizing point $\mathbf{u}_\text{s}$ and $\mathbf{u}'_s$:

$$w_{ss'} = n(\mathbf{u}_s) \times n(\mathbf{u}_{s'}) \text{ with } \sum_{s=1}^{P_i} n(\mathbf{u}_s) = n(v_i) \text{ and } \sum_{s'=1}^{P_j} n(\mathbf{u}_{s'}) = n(v_j) \tag{8}$$

The kriging variance for the areal estimator is computed as:

$$\sigma^2_{PK}(v_\alpha) = \bar{C}_R(v_\alpha, v_\alpha) - \sum_{i=1}^{K} \lambda_i(v_\alpha) \bar{C}_R(v_i, v_\alpha) - \mu(v_\alpha) \tag{9}$$

where $\bar{C}_R(v_\alpha, v_\alpha)$ is the within-area covariance that depends on the form of the geographical entity v_α and decreases as its area increases. Thus, ignoring the size of the prediction support in the computation of the kriging variance (3) can lead to a systematic overestimation of the prediction variance of large blocks.

2.2 Area-to-Point (ATP) Poisson Kriging

A major limitation of choropleth maps is the common biased visual perception that larger rural and sparsely populated areas are of greater importance. A solution is to create continuous maps of mortality risk, which amounts to perform a disaggregation or area-to-point interpolation. At each discretizing point $\mathbf{u}_s$ within an entity v_α, the risk $r(\mathbf{u}_s)$ can be estimated as the following linear combination of areal data:

$$\hat{r}_{PK}(\mathbf{u}_s) = \sum_{i=1}^{K} \lambda_i(\mathbf{u}_s) z(v_i) \tag{10}$$

The Poisson kriging system is similar to system (6), except for the right-hand-side term where the area-to-area covariances $\bar{C}_R(v_i, v_\alpha)$ is replaced by the area-to-point covariance $\bar{C}_R(v_i, \mathbf{u}_s)$. The latter is approximated by a procedure similar to the one described in equation (7). A critical property of the ATP kriging estimator is its coherence, that is the aggregation of the P_α point risk estimates within any given entity v_α yields the areal risk estimate $\hat{r}_{PK}(v_\alpha)$:

$$\hat{r}_{PK}(v_\alpha) = \frac{1}{n(v_\alpha)} \sum_{s=1}^{P_\alpha} n(\mathbf{u}_s) \hat{r}_{PK}(\mathbf{u}_s) \tag{11}$$

Condition (11) differs from the constraint commonly found in the geostatistical literature (Kyriakidis, 2004) in that: 1) the observation $z(v_\alpha)$ is uncertain, hence it is the reproduction of the PK risk estimate $\hat{r}_{PK}(v_\alpha)$ that is imposed, and 2) the incorporation of the population density in the computation of the areal covariance implies that it is the population-weighted average of the point risk estimates, not their arithmetical average, that satisfies the coherence condition. The constraint (11) is satisfied if the same K areal data are used for the estimation of the P_α point risk estimates. Indeed, in this case the population-weighted average of the right-hand-side covariance terms of the K ATP kriging systems is equal to the right-hand-side covariance of the single ATA kriging system:

$$\begin{aligned} \tfrac{1}{n(v_\alpha)} \sum_{s=1}^{P_\alpha} n(\mathbf{u}_s) \bar{C}_R(v_i, \mathbf{u}_s) &= \tfrac{1}{n(v_\alpha)} \sum_{s=1}^{P_\alpha} n(\mathbf{u}_s) \left[\tfrac{1}{n(v_i)} \sum_{s'=1}^{P_i} n(\mathbf{u}_{s'}) C(\mathbf{u}_{s'}, \mathbf{u}_s) \right] \\ &= \bar{C}_R(v_i, v_\alpha) \,, \end{aligned} \tag{12}$$

per relations (7) and (8). Therefore, the following relationship exists between the two sets of ATA and ATP kriging weights:

$$\lambda_i(v_\alpha) = \frac{1}{n(v_\alpha)} \sum_{s=1}^{P_\alpha} n(\mathbf{u}_s)\lambda_i(\mathbf{u}_s) \quad i = 1, \ldots, K \tag{13}$$

which ensures the coherence of the estimation.

2.3 Deconvolution of the Semivariogram of the Risk

Both ATA and ATP kriging require knowledge of the point support covariance of the risk $C(\mathbf{h})$, or equivalently the semivariogram $\gamma(\mathbf{h})$. This function cannot be estimated directly from the observed rates, since only aggregated data are available. Derivation of a point support semivariogram from the experimental semivariogram of areal data is called "deconvolution", an operation that is frequent in mining and has been the topic of much research (Journel and Huijbregts, 1978). However, in typical mining applications all blocks (areas) have the same size and shape, which makes the deconvolution reasonably straightforward. Goovaerts (2008) proposed an iterative approach to conduct the deconvolution in presence of irregular geographical units. This innovative algorithm starts with the derivation of an initial deconvoluted model $\gamma^{(0)}(\mathbf{h})$; for example the model $\gamma_R(\mathbf{h})$ fitted to the areal data. This initial model is then regularized using the following expression:

$$\gamma_{regul}(\mathbf{h}) = \bar{\gamma}^{(0)}(v, v_h) - \bar{\gamma}_h^{(0)}(v, v) \tag{14}$$

where $\bar{\gamma}^{(0)}(v, v_h)$ is the area-to-area semivariogram value for any two counties separated by a distance h. It is approximated by the population-weighted average (7), using $\gamma^{(0)}(\mathbf{h})$ instead of $C(\mathbf{h})$. The second term, $\bar{\gamma}_h^{(0)}(v, v)$, is the within-area semivariogram value. Unlike the expression commonly found in the literature, this term varies as a function of the separation distance since smaller areas tend to be paired at shorter distances. To account for heterogeneous population density, the distance between any two counties is estimated as a population-weighted average of distances between locations discretizing the pair of counties:

$$Dist(v_i, v_j) = \frac{1}{\sum_{s=1}^{P_i}\sum_{s'=1}^{P_j} n(\mathbf{u}_s)n(\mathbf{u}_{s'})} \sum_{s=1}^{P_i}\sum_{s'=1}^{P_j} n(\mathbf{u}_s)n(\mathbf{u}_{s'}) \|\mathbf{u}_s - \mathbf{u}_{s'}\| \tag{15}$$

Note that the block-to-block distances (15) are numerically very close to the Euclidian distances computed between population-weighted centroids (Goovaerts, 2006b). The theoretically regularized model, $\gamma_{regul}(\mathbf{h})$, is compared to the model fitted to experimental values, $\gamma_R(\mathbf{h})$, and the relative difference between the two curves,

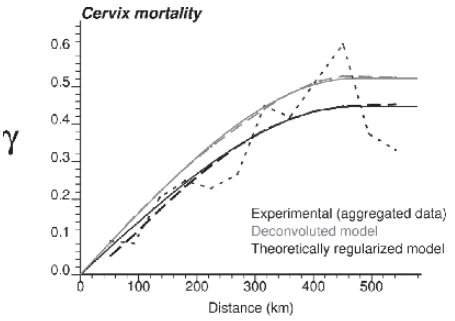
denoted D, is used as optimization criterion. A new candidate point-support semivariogram $\gamma^{(1)}(\mathbf{h})$ is derived by rescaling of the initial point-support model $\gamma^{(0)}(\mathbf{h})$, and then regularized according to expression (14). Model $\gamma^{(1)}(\mathbf{h})$ becomes the new optimum if the theoretically regularized semivariogram model $\gamma^{(1)}_{regul}(h)$ gets closer to the model fitted to areal data, that is if $D^{(1)} < D^{(0)}$. Rescaling coefficients are then updated to account for the difference between $\gamma^{(1)}_{regul}(h)$ and $\gamma_R(\mathbf{h})$, leading to a new candidate model $\gamma^{(2)}(\mathbf{h})$ for the next iteration. The procedure stops when the maximum number of allowed iterations has been tried (e.g. 35 in this paper) or the decrease in the D statistic becomes negligible from one iteration to the next. The use of lag-specific rescaling coefficients provides enough flexibility to modify the initial shape of the point-support semivariogram and makes the deconvolution insensitive to the initial solution adopted. More details and simulation studies are available in Goovaerts (2006b, 2008).

2.4 Application to the Cervix Cancer Mortality Data

Figure 2 (top graph, dark gray curve) shows the experimental and model semivariograms of cervix cancer mortality risk computed from aggregated data using estimator (4) and the distance measure (15). This model is then deconvoluted and, as expected, the resulting model (light gray curve) has a higher sill since the punctual process has a larger variance than its aggregated form. Its regularization using expression (14) yields a semivariogram model that is close to the one fitted to experimental values, which validates the consistency of the deconvolution.

The deconvoluted model was used to estimate aggregated risk values at the county level (ATA kriging) and to map the spatial distribution of risk values within counties (ATP kriging). Both maps are much smoother than the map of raw rates since the noise due to small population sizes is filtered. In particular, the high risk area formed by two central counties in Fig. 1 disappeared, which illustrates how hazardous the interpretation of the map of observed rates can be. The highest risk (4.081 deaths/100,000 habitants) is predicted for Kern County, just west of Santa Barbara County. ATP kriging map shows that the high risk is not confined to this sole county but spreads over four counties, which is important information for designing prevention strategies. By construction, aggregating the ATP kriging estimates within each county using the population density map of Fig. 1 (right medium graph) yields the ATA kriging map.

The map of ATA kriging variance essentially reflects the higher confidence in the mortality risk estimated for counties with large populations. The distribution of population can however be highly heterogeneous in large counties with contrasted urban and rural areas. This information is incorporated in the ATP kriging variance map that shows clearly the location of urban centers, such as Los Angeles, San Francisco, Salt Lake City, Las Vegas or Tucson. The variance of point risk estimates is much larger than the county-level estimates, as expected.



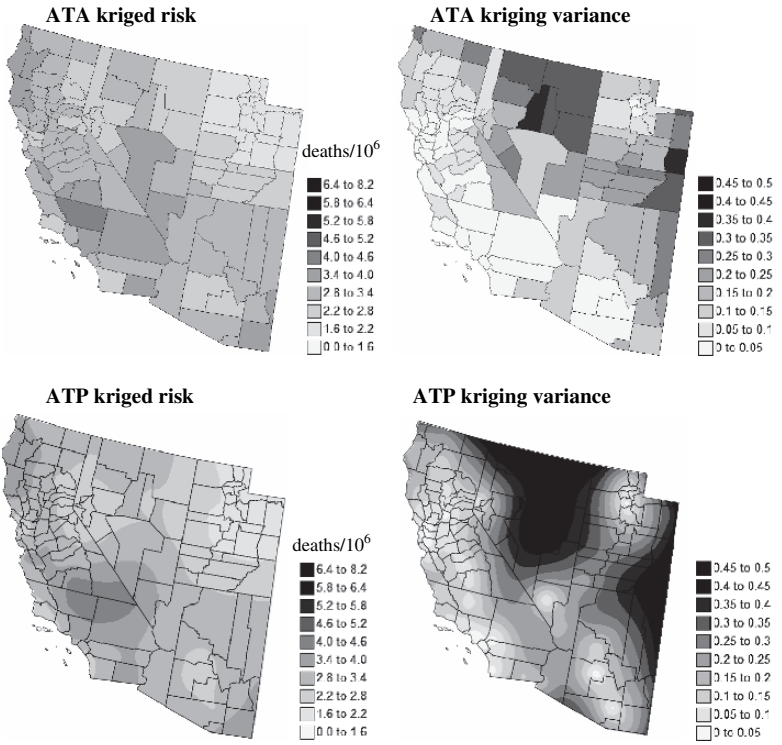


Fig. 2 Experimental semivariogram of the risk estimated from county-level rate data, and the results of its deconvolution (top curve). The regularization of the point support model yields a curve (black dashed line) that is very close to the experimental one. The model is then used to estimate the cervix cancer mortality risk (deaths/100,000 habitants) and associated prediction variance at the county level (ATA kriging) or at the nodes of a 5 km spacing grid (ATP kriging)

3 Detection of Spatial Clusters and Outliers

Mapping cancer risk is a preliminary step towards further analysis that might highlight areas where causative exposures change through geographic space, the

presence of local populations with distinct cancer incidences, or the impact of different cancer control methods.

3.1 Local Cluster Analysis (LCA)

The local Moran test aims to detect the existence of local clusters or outliers of high or low cancer risk values (Goovaerts, 2005b). For each county, the so-called LISA (Local Indicator of Spatial Autocorrelation) statistic is computed as:

$$LISA(v_\alpha) = \left[\frac{z(v_\alpha) - m}{s}\right] \times \left(\sum_{j=1}^{J(v_\alpha)} \frac{1}{J(v_\alpha)} \times \left[\frac{z(v_j) - m}{s}\right]\right) \qquad (16)$$

where $z(v_\alpha)$ is the mortality rate for the county being tested, which is referred to as the "kernel" hereafter; $z(v_\mathrm{j})$ are the rates for the $J(v_\alpha)$ neighboring counties that are here defined as units sharing a common border or vertex with the kernel v_α (1-st order queen adjacencies). All values are standardized using the mean m and standard deviation s of the set of risk estimates. Since the standardized values have zero mean, a negative value for the LISA statistic indicates a negative local auto-correlation and the presence of spatial outlier where the kernel value is much lower (higher) than the surrounding values. Cluster of low (high) values will lead to positive values of the LISA statistic.

In addition to the sign of the LISA statistic, its magnitude informs on the extent to which kernel and neighborhood values differ. To test whether this difference is significant or not, a Monte Carlo simulation is conducted, which traditionally consists of sampling randomly and without replacement the global distribution of rates (i.e. sample histogram) and computing the corresponding simulated neighborhood averages. This operation is repeated many times (e.g. $M = 999$ draws) and these simulated values are multiplied by the kernel value to produce a set of M simulated values of the LISA statistic for the entity v_α. This set represents a numerical approximation of the probability distribution of the LISA statistic at v_α, under the assumption of spatial independence. The observed statistic (Equation 16) is compared to the probability distribution, enabling the computation of the probability of not rejecting the null hypothesis of spatial independence. The so-called p-value is compared to the significance level chosen by the user and representing the probability of rejecting the null hypothesis when it is true (Type I error). Every county where the p-value is lower than the significance level is classified as a significant spatial outlier (HL: high value surrounded by low values, and LH: low value surrounded by high values) or cluster (HH: high value surrounded by high values, and LL: low value surrounded by low values). If the p-value exceeds the significance level, the county is declared non-significant (NS).

Figure 3 (left top map) shows the results of the LCA of the observed cervix cancer mortality rates. Only two counties are declared significant HL outliers, a result that must be interpreted with caution given their small population sizes. Indeed, these

two counties become non-significant when the analysis is conducted on the map of kriged risks, see Fig. 3 (right top map). Accounting for population size in the analysis reveals a cluster of low risk values in Utah, which likely reflects cultural or religious influence on sexual practices resulting in reduced transmission of human papillomavirus. Yet, the smoothing effect of kriging tends to enhances spatial autocorrelation in the risk map, with the risk of inflating artificially cluster sizes. For example, the one-county HH cluster detected in the middle of the mortality map grows to become an aggregate of seven counties on the map of kriged risks. Another weakness is that the uncertainty attached to the risk estimates (i.e. kriging variance) is ignored in the analysis.

3.2 Stochastic Simulation of Cancer Mortality Risk

Static maps of risk estimates and the associated prediction variance fail to depict the uncertainty attached to the spatial distribution of risk values and do not allow its propagation through local cluster analysis. Instead of a unique set of smooth risk estimates $\{\hat{r}_{PK}(v_\alpha),\ \alpha = 1,\ldots,N\}$, stochastic simulation aims to generate a set of L equally-probable realizations of the spatial distribution of risk values, $\{r^{(l)}(v_\alpha),\ \alpha = 1,\ldots,N;\ l = 1,\ldots,L\}$, each consistent with the spatial pattern of the risk as modeled using the function $\gamma_R(\mathbf{h})$. Goovaerts (2006a) proposed the use of p-field simulation to circumvent the problem that no risk data (i.e. only risk estimates), hence no reference histogram, is available to condition the simulation. The basic idea is to generate a realization $\{r^{(l)}(v_\alpha),\ \alpha = 1,\ldots,N\}$ through the sampling of the set of local probability distributions (ccdf) by a set of spatially correlated probability values $\{p^{(l)}(v_\alpha),\ \alpha = 1,\ldots,N\}$, known as a probability field or p-field. Assuming that the ccdf of the risk variable is Gaussian, each risk value can be simulated as:

$$r^{(l)}(v_\alpha) = \hat{r}_{PK}(v_\alpha) + \sigma_{PK}(v_\alpha) y^{(l)}(v_\alpha) \tag{17}$$

where $y^{(l)}(v_\alpha)$ is the quantile of the standard normal distribution corresponding to the cumulative probability $p^{(l)}(v_\alpha)$. $\hat{r}_{PK}(v_\alpha)$ and $\sigma_{PK}(v_\alpha)$ are the ATA kriging estimate and standard deviation, respectively. The L sets of random deviates or normal scores, $\{y^{(l)}(v_\alpha),\ \alpha = 1,\ldots N\}$, are generated using non-conditional sequential Gaussian simulation with the distance metric (15) and the semivariogram of the risk, $\gamma_R(\mathbf{h})$, rescaled to a unit sill; see Goovaerts (2006a) for a detailed description of the algorithm.

Figure 3 (medium row) shows two realizations of the spatial distribution of cervix cancer mortality risk values generated using p-field simulation. The simulated maps are more variable than the kriged risk map of Fig. 2, yet they are smoother than the map of potentially unreliable rates of Fig. 1. Differences among realizations depict the uncertainty attached to the risk map. For example, Nye County in the center of the map, which has a very high mortality rate (recall Fig. 1) but low

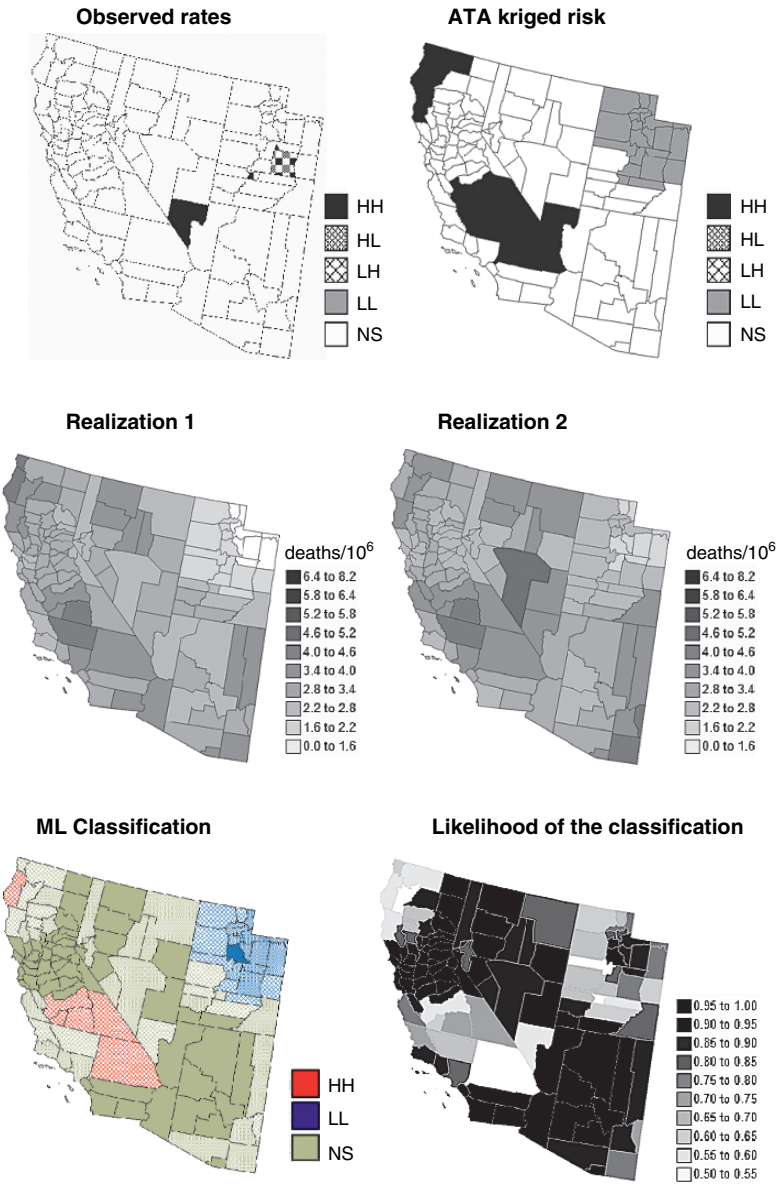


Fig. 3 Results of the local cluster analysis conducted on cervix cancer mortality rates and estimated risks (top maps); see legend description in text. Middle maps show two realizations of the spatial distribution of cervix cancer risk, while the bottom map shows the most likely (ML) classification inferred from 500 realizations. The intensity of the shading increases as the classification becomes more certain (i.e. the likelihood increases)

population, has a simulated risk that is small for realization 1 but large in the next realization. Five hundreds realizations were generated and underwent a local cluster analysis. The information provided by the set of 500 LCAs is summarized at the bottom of Fig. 3. The color code indicates the most frequent classification (maximum likelihood = ML) of each county across the 500 simulated maps. The shading reflects the probability of occurrence or likelihood of the mapped class, see Fig. 3 (right bottom graph). Solid shading corresponds to classifications with high frequencies of occurrence (i.e. likelihood > 0.9), while hatched counties denote the least reliable results (i.e. likelihood < 0.75). This coding is somewhat subjective but leads to a clear visualization of the lower reliability of the clusters of high values relatively to the cluster of low risk identified in Utah. Only one county south of Salt Lake City is declared a significant low-risk cluster with a high likelihood (0.906).

4 Correlation Analysis

Once spatial patterns, such as clusters of high risk values, have been identified on the cancer mortality map, a critical step for cancer control intervention is the analysis of relationships between these features and putative environmental, demographic, socioeconomic and behavioral factors. The major difficulty is the choice of a scale for quantifying correlations between variables that are typically measured over very different supports, e.g. counties and census blocks in this study.

4.1 Ecological Analysis

The most straightforward approach is to aggregate the finer data to the level of coarser resolution data, resulting in a common geographical scale for the correlation analysis. For example, Fig. 4 shows the county-level kriged risk and the two covariates of Fig. 1 aggregated to the same geography: percentage of habitants living below the federally defined poverty line, and percentage of Hispanic females. Both variables were logarithmically transformed, and their product defines the interaction term. Table 1 (first two rows) shows the correlation coefficient between each of the three covariates and the mortality rates before and after application of Poisson kriging. Filtering the noise due to the small number problem clearly enhances the explanatory power of the covariates: the proportion of variance explained (R^2) increases by almost one order of magnitude (6.2% to 48.8%) and all correlation coefficients become highly significant. The uncertainty attached to the risk estimates can be accounted for by weighting each estimate according to the inverse of its kriging variance, leading to slightly larger correlation coefficients and R^2 (Table 1, 3rd row).

So far the significance of the correlation coefficient is tested using the common assumption of independence of observations, which is clearly inappropriate for most

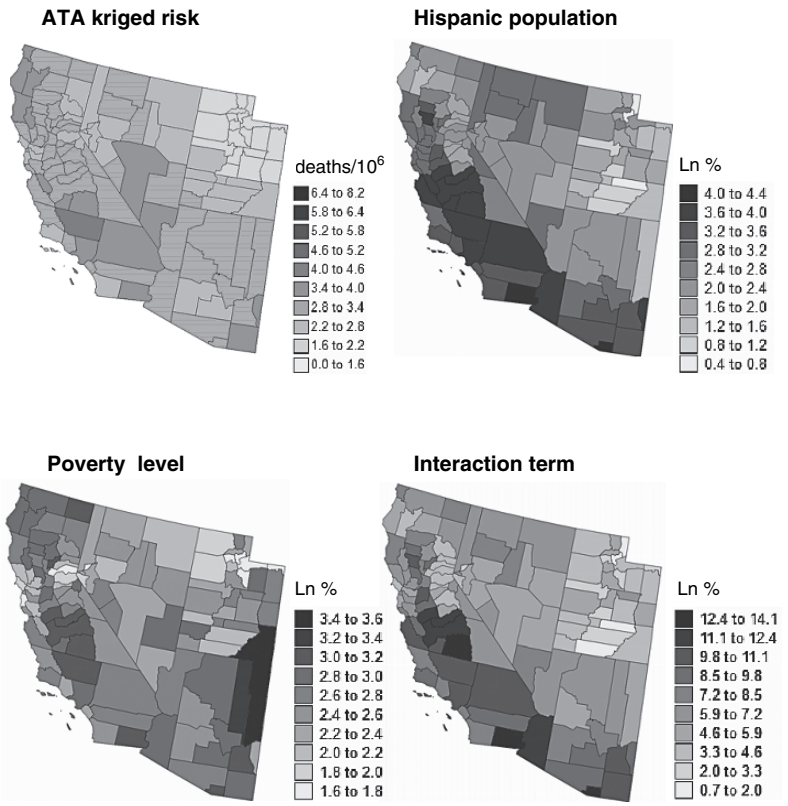


Fig. 4 Maps of cancer mortality risk estimated by Poisson kriging and the logtransformed values of three putative covariates aggregated to the county-level for conducting the ecological analysis

spatial datasets. Instead of computing the correlation between each covariate and the smoothed risk map, the correlation was quantified for each of the 500 risk maps generated by p-field simulation in Section 3.2. This propagation of uncertainty leads to a range of correlation coefficients and R^2 that can be fairly wide, see Table 1 (4th row). Next, this distribution must be compared to the one expected under the assumption of no correlation between mortality risk and each covariate. This reference distribution was obtained empirically in 2 steps. First, the maps of covariates were modified using the spatially ordered shuffling procedure proposed by Goovaerts and Jacquez (2004). The idea is to generate a standard normal random field with a given spatial covariance, e.g. the covariance of the demographic variable in this paper, using non-conditional sequential Gaussian simulation. Each simulated normal score is then substituted by the value of same rank in the distribution of proportion of Hispanic females. To maintain the correlation among covariates, all three covariate maps were modified simultaneously. The operation was repeated 100 times, yielding

Table 1 Results of the correlation analysis of cervix cancer mortality rates and kriged risks with two putative covariates, as well as their interaction. Kriging estimates are weighted according to the inverse of their kriging variance. The use of neutral models allows one to incorporate the spatial uncertainty attached to cancer risk estimates into the computation of the correlation coefficients and testing of their significance (* = significant, ** = highly significant). The last two rows show the results obtained after disaggregation

Correlation with covariates				
Regression models	Hispanic	Poverty	Interaction	R^2(%)
County-level correlation				
Rates	0.210*	0.144	0.240**	6.2
ATA kriging	0.625**	0.473**	0.690**	48.8
ATA kriging (weighted)	0.641**	0.613**	0.729**	54.1
ATA kriging (neutral model)	0.247–0.703**	0.173–0.590**	0.347–0.716**	14.4–52.0
Point-level (25 km^2 cells) correlation				
ATP kriging	0.096**	−0.036**	0.188**	9.8
ATP kriging (weighted)	0.239**	0.090**	0.321**	14.0

100 sets of covariate maps. Second, the correlation between each of the re-ordered covariate maps and each of the 500 simulated risk maps is assessed, leading to a distribution of 50,000 correlation coefficients that corresponds to an hypothesis of independence, since the covariate maps were modified independently of the risk maps. For this case study, this more realistic testing procedure does not change the conclusions drawn from the classical analysis.

Correlations computed between health outcomes and risk factors averaged over geographical entities, such as counties, are referred to as 'ecological correlations'. The unit of analysis is a group of people, as opposed to individual-based studies that relies on data collected for each cancer case. A limitation of ecological analyses is the resolution available which might be too coarse to obtain a detailed view of geographical patterns in disease mortality or incidence. The aggregation may also distort or mask the true exposure/response relationship for individuals, a phenomenon called the *ecological fallacy*. The disaggregation performed by ATP Poisson kriging eliminates the need for using averaged values, and the correlation coefficients between both risk and covariates estimated at the nodes of the 5-km spacing grid are listed in Table 1 (last rows). The correlation is much weaker than for county-level data, which might be due to the noise in the map of socio-demographic variables and/or reflects the scale-dependence of the relationship.

4.2 Geographically-Weighted Regression

The analysis in Table 1 is aspatial and makes the implicit assumption that the impact of covariates is constant across the study area. This assumption is likely

unrealistic for large areas which can display substantial geographic variation in demographic, social, economic, and environmental conditions. Several local regression techniques have been developed to account for the non-stationarity of relationships in space (Fotheringham et al., 2002). In geographically-weighted regression (GWR) the regression is performed within local windows centred on each observation or the nodes of a regular grid, and each observation is weighted according to its proximity to the centre of the window. This weighting avoids abrupt changes in the local statistics computed in adjacent windows. Local regression coefficients and associated statistics (i.e. proportion of variance explained, correlation coefficients) can then be mapped to visualize how the explanatory power of covariates changes spatially (Goovaerts, 2005c).

GWR regression was conducted using as dependent variable the mortality risk estimated by ATA and ATP kriging (20 km spacing grid). The centers of the local windows were identified to either the county population-weighted centroids or the nodes of the 5 km spacing grid. The window size was defined as the set of 50 closest observations for both county-level and point-level data. The weight assigned to each observation $\mathbf{u}_\alpha$ was computed as $C_{sph}(h_{0\alpha})/\sigma^2_{PK}(\mathbf{u}_a)$, where $C_{sph}(h_{0\alpha})$ is the value of the spherical covariance at a distance $h_{0\alpha}$ to the center $\mathbf{u}_0$ of the window, and $\sigma^2_{PK}(\mathbf{u}_a)$ is the kriging variance of the ATA or ATP kriged estimate. The range of $C_{sph}(h)$ was set to the distance between the center of the window and the most distant observation. Two statistics are displayed in Fig. 5: the proportion of variance explained within each window (left column) and the covariate with the highest significant correlation coefficient (right column).

The analysis of county-level data (Fig. 5, top maps) shows a clear SW-NE trend in the explanatory power of the local regression models: the higher mortality values along the coast are better explained by the two covariates than the lower risk recorded in Utah. In this state, none of the covariates displays significant correlation with cancer mortality. Poverty level is the best correlated covariate in Northern California while the interaction between economic and demographic variables is the most significant factor in Central California and in the South of the study area. The proportion of Hispanic females is the most significant covariate in a very small transition area between the coast where higher mortality rates and proportion of Hispanic females are observed and Utah where the same two variables have lower values. The computation of the GWR statistics over a regular grid allows one to visualize the within-county variability (Fig. 5, middle maps), yet the analysis is still based on county-level aggregates of socio-demographic variables which can be overly simplistic for some counties, recall Figure 1 (bottom maps). For example, the largest R^2 observed in the Northeast corner of the study area (Fig. 5, left bottom map) corresponds to the Eastern border of a county that display great variation for both proportion of Hispanic females and habitants below the poverty level. Differences between the GWR of county-level and point-support data are even more striking for the map of significantly correlated covariates. The pattern becomes much more complex and correlations are locally negative, see hatched areas in Figure 5 (right bottom map). These maps are mainly used for descriptive purpose and should guide further studies to interpret these local relationships.

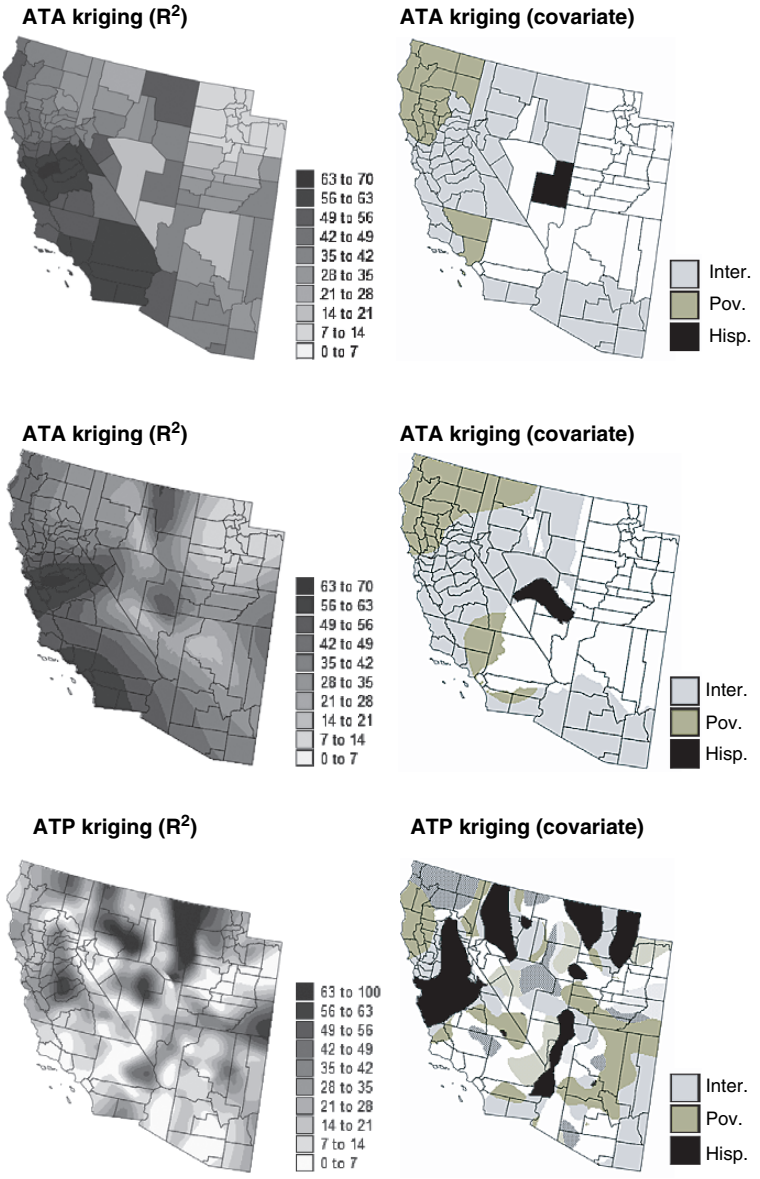


Fig. 5 Results of the geographically-weighted regression applied to the ATA and ATP kriged risk values. Left column displays the maps of the local proportion of variance explained, while the right maps show, for each county or node of the 5 km spacing grid, the covariate (Hispanic population, poverty level, and interaction) that has the highest significant correlation (hatched areas = negative correlation) with cancer mortality risk. The analysis of county-level data conducted at each node of the 5 km spacing grid is shown in the middle maps

Conclusions

The analysis of health data and putative covariates, such as environmental, socio-economic, behavioral or demographic factors, is a promising application for geostatistics. It presents, however, several methodological challenges that arise from the fact that data are typically aggregated over irregular spatial supports and consist of a numerator and a denominator (i.e. population size). Common geostatistical tools, such as semivariograms or kriging, thus cannot be blindly implemented in environmental epidemiology. This paper demonstrated how recent developments in other disciplines, such as ecology for Poisson kriging or remote sensing for area-to-point kriging, can foster the advancement of health geostatistics. Capitalizing on these results and an innovative approach for semivariogram deconvolution, this paper presented the first study where the size and shape of administrative units, as well as the population density, is incorporated into the filtering of noisy mortality rates and the mapping of the corresponding risk at a fine scale (i.e. disaggregation).

Like in other disciplines, spatial interpolation is rarely a goal per se; rather it is a step along the decision-making process. In epidemiology one main concern is to establish the rationale for targeted cancer control interventions, including consideration of health services needs, and resource allocation for screening and diagnostic testing. It is thus important to delineate areas with significantly higher mortality or incidence rates, as well as to analyze relationships between health outcomes and putative risk factors. The uncertainty attached to cancer maps needs however to be propagated through this analysis, a task that geostatisticians have been tackling for several decades using stochastic simulation. Once again the implementation of this approach in epidemiology faces specific challenge, such as the absence of measurements of the target attribute. This paper introduced the application of p-field simulation to generate realizations of cancer mortality maps, which allows one to quantify numerically how the uncertainty about the spatial distribution of health outcomes translates into uncertainty about the location of clusters of high values or the correlation with covariates. Last, this study demonstrated the limitation of a traditional aspatial regression analysis, which ignores the geographic variations in the impact of covariates.

The field of health geostatistics is still in its infancy. Its growth cannot be sustained, or at least is meaningless, if it does not involve the end-users who are the epidemiologists and GIS specialists working in health departments and cancer registries. Critical components to its success include the publication of applied studies illustrating the merits of geostatistics over current methods, training through short courses and updating of existing curriculum, as well as the development of user-friendly software. The success of mining and environmental geostatistics, as we experience it today, can be traced back to its development outside the realm of spatial statistics, through the close collaboration of mathematically minded individuals and practitioners. Health geostatistics will prove to be no different.

Acknowledgments This research was funded by grant R44-CA105819-02 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.

References

Fotheringham AS, Brunsdon C, Charlton M (2002) Geographically weighted regression: the analysis of spatially varying relationships. Wiley, Chichester

Friedell GH, Tucker TC, McManmon E, Moser M, Hernandez C, Nadel M (1992) Incidence of dysplasia and carcinoma of the uterine cervix in an Appalachian population. J Nat Cancer Inst 84:1030–1032

Goovaerts P (2005a) Geostatistical analysis of disease data: estimation of cancer mortality risk from empirical frequencies using Poisson kriging. Int J Health Geogr 4:31

Goovaerts P (2005b) Detection of spatial clusters and outliers in cancer rates using geostatistical filters and spatial neutral models. In: Renard Ph, Demougeot-Renard H, and Froidevaux R (eds) geoENV V-Geostatistics for Environmental Applications. Springer-Verlag, Berlin, Germany, pp 149–160

Goovaerts P (2005c) Analysis and detection of health disparities using Geostatistics and a space-time information system. The case of prostate cancer mortality in the United States, 1970–1994. Proceedings of GIS Planet 2005, Estoril, May 30-June 2

Goovaerts P (2006a) Geostatistical analysis of disease data: visualization and propagation of spatial uncertainty in cancer mortality risk using Poisson kriging and p-field simulation. Int J Health Geogr 5:7

Goovaerts P (2006b) Geostatistical analysis of disease data: accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point Poisson kriging. Int J Health Geogr 5:52

Goovaerts (2008) Kriging and semivariogram deconvolution in presence of irregular geographical units. Math Geology 40, in press

Goovaerts P, Jacquez GM (2004) Accounting for regional background and population size in the detection of spatial clusters and outliers using geostatistical filtering and spatial neutral models: the case of lung cancer in Long Island New York. Int J Health Geogr 3:14

Gotway CA Young LJ (2002) Combining incompatible spatial data. J Am Stat Assoc 97:632–648

Gotway CA, Young LJ (2005) Change of support: an inter-disciplinary challenge. In: Renard Ph, Demougeot-Renard H, and Froidevaux R (eds) geoENV V - Geostatistics for environmental applications. Springer-Verlag, Berlin, Germany, pp 1–13

Journel AG, Huijbregts CG (1978) Mining Geostatistics. Academic Press, London

Kyriakidis P (2004) A geostatistical framework for area-to-point spatial interpolation. Geogr Anal 36:259–289

Monestiez P, Dubroca L, Bonnin E, Durbec JP Guinet C (2006) Geostatistical modelling of spatial distribution of *Balenoptera physalus* in the Northwestern Mediterranean Sea from sparse count data and heterogeneous observation efforts. Ecol Modell 193:615–628

Waller LA, Gotway CA (2004) Applied Spatial Statistics for public health data. John Wiley and Sons, New Jersey