Published in final edited form as:

Spat Spatiotemporal Epidemiol. 2009 October 1; 1(1): 61–71. doi:10.1016/j.sste.2009.07.001.

COMBINING AREA-BASED AND INDIVIDUAL-LEVEL DATA IN THE GEOSTATISTICAL MAPPING OF LATE-STAGE CANCER INCIDENCE

Pierre Goovaerts[Chief Scientist]

BioMedware, Inc., 516 North State Street, Ann Arbor, MI 48104, USA, Tel: (734) 913-1098, Fax: (734) 913-2201

Pierre Goovaerts: goovaerts@biomedware.com

Abstract

This paper presents a geostatistical approach to incorporate 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 incidence, with an application to breast cancer in three Michigan counties. Spatial trends in cancer incidence are first estimated from census data using area-to-point binomial kriging. This prior model is then updated using indicator kriging and individual-level data. Simulation studies demonstrate the benefits of this two-step approach over methods (kernel density estimation and indicator kriging) that process only residence data.

Keywords

Breast cancer; Census tract; Binomial kriging; indicator semivariogram; Michigan; kernel density estimator

Introduction

Cancer is a major public health problem in the United States and is currently the second leading cause of death. For cancer control activities and resource allocation, it is important to be able to compare incidence and survival rates, risk behaviors, screening patterns, diagnosis stage, and treatment methods across geographical and political boundaries and at as fine a spatial scale as possible (Pickle et al., 2006). Data available for such studies fall within two main categories: individual-level data (e.g. age, marital status, and residence of cancer cases and controls) and aggregated data (e.g. cancer rates recorded at county or ZIP code level, neighborhood socioeconomic status). For mapping purposes, both types of data are typically processed independently using different sets of methods.

Although individual humans represent the basic unit of spatial analysis, the majority of maps of disease occurrence depict data in discrete or areal form (Bithell, 2000). Much attention has been devoted to the treatment of such data, leading to the development of a wide variety of statistical smoothing algorithms to filter local small-scale variations from choropleth maps (Lawson, 2000; Talbot et al., 2000). The main limitation of all these methods,

© 2009 Elsevier Ltd. All rights reserved.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

including the increasingly popular full Bayesian modeling approach (Best et al., 1999, 2005), is that the size and shape of geographical units (e.g. counties) are generally ignored in the analysis. Therefore no change of support (e.g. disaggregation) can be conducted and issues associated with the display of rates aggregated within political or administrative boundaries, or the integration of health data and covariates measured over different geographies, remain. Building on earlier work in the areas of change of support (Tobler, 1979; Gotway and Young, 2002; Kyriakidis, 2004) and kriging of count data (Monestiez et al., 2006), Goovaerts (2006) recently introduced a geostatistical approach 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). This method produces disease maps in which rates vary continuously across the study area (isopleth maps), thereby reducing the visual bias that is typically associated with the interpretation of choropleth maps. Similar efforts were undertaken by Gotway and Young (2007) for mapping the number of low birth weight (LBW) babies at the Census tract level, accounting for county-level LBW data and covariates measured over different spatial supports, such as a fine grid of ground-level particulate matter concentrations or tract population.

Despite the methodological advancements in the treatment of areal data, the degree of details in the isopleth risk maps will always be limited by the initial resolution of the choropleth map. Whenever possible, it is thus beneficial to avoid the tedious, arbitrary and inherently information-wasteful aggregation step and to process directly the point-based data (Bithell, 2000). In addition to the greater accuracy in the location of health outcomes, the analysis of geocoded data can often capitalize on detailed information on residential history and a large number of potential risk factors. In contrast to the well-developed methods for mapping aggregated epidemiologic data, the spatial mapping of point data has received much less attention (Webster et al., 2006). Mapping of individual point-based data presents challenges besides confidentiality and misalignment issues. First, simply mapping the location of health events or cases, as with the famous maps of John Snow (Mc Leod, 2000), is misleading unless the population density is spatially uniform. Even when the population at risk can be enumerated and geocoded fairly easily (e.g. newborns or patients diagnosed with cancer), mapping the location of both cases and non-cases does not provide quantitative estimates of rates; estimation and mapping of the spatial risk function requires the computation of the ratio of the case density to the population density over a regular grid of points.

The most straightforward approach is to use 'kernel density estimation', whereby the number of cases and the total number of individuals at risk are summed within sliding windows and their ratio defines the rate assigned to the center (i.e. grid node) of that window. The operation is repeated for each grid node, allowing the creation of isopleth maps of, for example, late-stage cancer rates (ratio of number of late-stage cancer cases to total number of people diagnosed with that cancer). Many variants of the same principle (e.g. Talbot et al., 2000; Rushton et al., 2004; Tiwari and Rushton, 2004; Vieira et al., 2005) have been proposed since its initial application in epidemiology, and a few of them have been implemented in the Disease Mapping and Analysis Program (DMAP, Rushton and Lolonis, 1996). The density estimation approach falls into the category of non-parametric or empirical methods, according to Bithell's classification. Empirical approaches are more straightforward and involve fewer assumptions than parametric or model-based methods that provide an analytical description of the risk surface. Generalized additive models (GAM) are the most popular type of model-based approach for mapping point-based epidemiologic data (Kelsall and Diggle, 1995, 1998; French and Wand, 2004; Webster et al., 2006). The GAM approach provides a unified statistical framework to combine smoothing with the ability to analyze binary outcome data and adjust for covariates. Recently, Kammann and Wand

(2003) showed how geostatistical kriging can be incorporated into GAM, with representation as a generalized linear mixed model. Both are promising but relatively untried methods in spatial epidemiology (Vieira et al, 2005), and the GAM method is substantially more computer intensive than empirical methods (Kelsall and Diggle, 1998).

The geostatistical analysis of geocoded data can be viewed as a particular case of the analysis of areal data where the measurement support is a point and the population size is one. Since the data represent the probability (0 or 1) that an individual is a case (e.g. late-stage cancer, birth defect), non-parametric geostatistics (Journel, 1983) developed for the analysis of categorical data seems well suited for that task. Indicator semivariograms allow one to quantify the spatial connectivity of the data, while probability maps are created by the application of kriging to indicator values. Indicator kriging has been used in many different fields, including ecology; see Rossi et al. (1992) for a review and Yabsley et al. (2005) for its use in the mapping of county serostatus for a tick-borne pathogen. Like traditional kriging methods, indicator kriging allows the incorporation of secondary information (Goovaerts and Journel, 1995) and takes into account the pattern of spatial dependence (e.g. anisotropy, range of autocorrelation) in the computation of the weights assigned to neighboring information, which should be an advantage over kernel density estimation methods.

This paper presents a geostatistical approach to identify the main scales of variation in geocoded data and to incorporate areal and individual-level data into the mapping of late-stage cancer incidence. Methodological developments are illustrated using breast cancer cases diagnosed over 17 years in three Michigan counties. Simulation studies are conducted to assess prediction performance of the geostatistical approach over common spatially adaptive filters for different sampling intensities and risk spatial patterns.

Data and Methods

Invasive breast cancer cases, diagnosed during the calendar years 1985 through 2002 in Michigan, will be used to illustrate and test the geostatistical mapping approach. Approximately 92% of these records, which were compiled by the Michigan Cancer Surveillance Program (MCSP), were successfully geocoded at residence at time of diagnosis. The current study focuses on cases diagnosed for white women in 83 census tracts of three counties in South-western Michigan: Berrien, Cass and Van Buren; see Figs. 1A and 1B (data are aggregated for confidentiality reasons). Out of the 2,118 women diagnosed with breast cancer during that time period, 18% of cases were defined as late-stage (i.e. regional and distant metastatic cancer) according to the SEER General Summary Stage classification (Young et al., 2001).

Indicator semivariograms

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/stage diagnosis:

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

The spatial pattern of these indicator data can be characterized using the experimental semivariogram computed as:

$$\hat{\gamma}_I(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} [i(\mathbf{u}_\alpha) - i(\mathbf{u}_\alpha + \mathbf{h})]^2 \quad (2)$$

where $N(\mathbf{h})$ is the number of pairs of cases within a given class of distance and direction, known as spatial lag and denoted $\mathbf{h}$. The spatial increment $[i(\mathbf{u}_\alpha) - i(\mathbf{u}_\alpha + \mathbf{h})]^2$ is non-zero only if cases at $\mathbf{u}_\alpha$ and $\mathbf{u}_\alpha + \mathbf{h}$ are diagnosed at different stages. The indicator variogram $2\hat{\gamma}_I(\mathbf{h})$ thus measures how often the stage of diagnosis of two cases a vector $\mathbf{h}$ apart is different. In other words, it quantifies the transition frequency between early and late-stage diagnosis as a function of $\mathbf{h}$. In presence of spatial clusters of early or late-stage diagnosis, the variogram value is expected to increase with the lag $\mathbf{h}$ and reaches a plateau at a distance, called range, which corresponds to the average size of these clusters. If these clusters are non-circular, different ranges will be observed along different directions, a situation referred to as spatial anisotropy. The study of indicator semivariograms can thus provide important information about the nature and scale of the process responsible for the spatial distribution of cancer stages at diagnosis.

Spatial contouring of risk

To create isopleth or contour maps of late-stage cancer incidence, one needs to estimate the spatial risk function I at any location $\mathbf{u}$. Kernel density estimation methods simply compute that risk as a linear combination of $n(\mathbf{u})$ surrounding observations falling within a window $W(\mathbf{u})$:

$$i^*(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha i(\mathbf{u}_\alpha) \quad \text{with} \quad \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha = 1 \quad (3)$$

Many variants of the same principle have been proposed since its initial application in epidemiology (Bithell, 1990):

1. the size of the window or spatial filter can remain constant over the study area (e.g. Rushton et al., 2004) or the spatial filter size is locally adjusted (spatially adaptive filters) to ensure the capture of a constant population size $n(\mathbf{u})$ (e.g. Talbot et al., 2000; Tiwari and Rushton, 2004).
2. the spatial filter is applied to individual point data (James et al., 2004; Tiwari and Rushton, 2004) or data aggregated into small census areas, such as ZIP codes (Talbot et al., 2000).
3. the spatial filter is centered directly on the nodes $\mathbf{u}$ of the mapping grid or the map is obtained by a two-step approach whereby the risk is estimated first at the location of each case $\mathbf{u}_\alpha$ using spatially adaptive filters, then interpolated at all grid nodes using kriging (Vieira et al., 2002).
4. all the observations within the spatial filter are weighted the same, i.e. $\lambda_\alpha = 1/n(\mathbf{u})$, (Talbot et al., 2000) or they are weighted according to their proximity to the center of the filter (Rushton, 2003; Han et al., 2005).
5. confounding factors, such as age, occupation or smoking status, are incorporated in the analysis (Rushton et al., 2004) or the crude unadjusted rates (or odds ratios) are mapped (Tiwari and Rushton, 2004; Vieira et al., 2005).

Whereas the computation of empirical estimates of type (3) is very straightforward, critical information regarding the pattern of spatial dependence of observations, the data geometry

(i.e. isolated cases in rural areas versus clustered cases in urban settings) and the spatial support of the observation (individual point data versus cases aggregated into small census areas) is ignored in the process. An alternative is to use geostatistical interpolation whereby the weights in equation (3) are the solution of the following system of $(n(\mathbf{u})+1)$ linear equations, known as ordinary indicator kriging system:

$$\begin{aligned} \sum_{\beta=1}^{n(\mathbf{u})} \lambda_{\beta} \gamma_l(\mathbf{u}_{\alpha} - \mathbf{u}_{\beta}) - \mu &= \gamma_l(\mathbf{u}_{\alpha} - \mathbf{u}) & \alpha=1, \dots, n(\mathbf{u}) \\ \sum_{\beta=1}^{n(\mathbf{u})} \lambda_{\beta} &= 1 \end{aligned} \quad (4)$$

where μ is a Lagrange multiplier accounting for the unit sum constraint on the weights. The kriging weights λ_{α} depend on the data spatial configuration (i.e. redundant clustered observations receive less weight than isolated observations), the proximity of the data to the node $\mathbf{u}$ (i.e. closest observations receive more weight than those further away), as well as the spatial connectivity of indicator data modelled from the experimental semivariogram (equation 2). Since kriging is a non-convex interpolator, estimated probabilities can fall outside the range $[0,1]$ and the faulty probabilities should be reset to the nearest bound, 0 or 1 (Goovaerts, 1997).

Binomial kriging

Because of the computational cost associated with solving large kriging systems and the fact that the closest data tend to screen the influence of those farther away, the number $n(\mathbf{u})$ of neighboring observations in Equation (3) rarely exceeds 100. For high density of cases, the information used for risk estimation is thus restricted to very small distances and does not capitalize on regional patterns captured through the semivariogram analysis. In addition, addresses can often fail to geocode, leaving missing or incomplete data where coarser surrogates, such as ZIP code, replace precise coordinates. One key idea in this paper is to combine individual-level data with incidence rates computed at the level of census tracts 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). More precisely, incidence rates are used to derive a local mean or prior probability $m_l(\mathbf{u})$, which is then updated into a posterior probability $i^*(\mathbf{u})$ using the following residual kriging (RK) estimate:

$$i_{RK}^*(\mathbf{u}) = m_l^*(\mathbf{u}) + \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_{\alpha} [i(\mathbf{u}_{\alpha}) - m_l^*(\mathbf{u}_{\alpha})] \quad \text{with} \quad m_l^*(\mathbf{u}) = \sum_{i=1}^K \lambda_i z(v_i) \quad (5)$$

The term $z(v_i)$ denotes the late-stage incidence rate for the i th census tract, which is computed as $d(v_i)/n(v_i)$, where $d(v_i)$ is the number of late-stage cases and $n(v_i)$ is the total number of cases. The first set of kriging weights λ_{α} are solutions of a system of type (4), except that the semivariogram model is inferred from the residual indicator data $r(\mathbf{u}_{\alpha}) = i(\mathbf{u}_{\alpha}) - m_l^*(\mathbf{u}_{\alpha})$.

Relatively to the direct use of individual-level data (equation 3), the computation of the weights λ_i assigned to each census tract must account for two important features: 1) the K rates have varying degree of reliability depending on the total number of cases $n(v_i)$ recorded in those tracts, and 2) the prediction support $\mathbf{u}$ is a point while the measurement support v_i is an area. These issues of small number problem and change of support were

recently addressed in the geostatistical literature by the introduction of Area-to-Point (ATP) Poisson kriging for creating isopleth maps of cancer mortality risk from choropleth rate maps (Goovaerts, 2006). To acknowledge the fact that late-stage cancer diagnosis is not a rare event, the methodology was modified by replacing the Poisson distribution by the Binomial distribution in the derivation of the system of equations (Webster et al., 1994; 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 $Y(v_i)$:

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

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:

$$\begin{aligned} E[Z(v_i) \mid Y(v_i)] &= Y(v_i) \\ \text{Var}[Z(v_i) \mid Y(v_i)] &= m(1 - m)/n(v_i) \end{aligned}$$

Following Monestiez (pers. com.), the unconditional mean and variance are as follows:

$$E[Z(v_i)] = m \quad (7)$$

$$\text{Var}[Z(v_i)] = \frac{m(1 - m) + (n(v_i) - 1)\sigma_Y^2}{n(v_i)} \quad (8)$$

In addition, $C_Y(\mathbf{h}) = C_I(\mathbf{h})$ when individual-level indicator data (Equation (1), $n(v_i) = 1$) are available, which is the case here since the residence of all cancer patients is known.

The kriging weights λ_i for the estimation of the local mean $m_i^*(\mathbf{u})$, which is identified with the risk value $Y^*(\mathbf{u})$, are computed by solving the following system of linear equations:

$$\begin{aligned} \sum_{j=1}^K \lambda_j \left[\bar{C}_I(v_i, v_j) + \delta_{ij} \frac{a}{n(v_i)} \right] + \mu &= \bar{C}_I(v_i, \mathbf{u}) \quad i=1, \dots, K \\ \sum_{j=1}^K \lambda_j &= 1. \end{aligned} \quad (9)$$

where $\delta_{ij} = 1$ if $i=j$ and 0 otherwise, $a = m^*(1 - m^*) - \bar{C}_I(v_i, v_i)$, 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. under the simplifying assumption that all census tracts have the same size and shape and $\bar{C}_I(v_i, v_i) = \sigma_Y^2$ (Oliver et al., 1998), the term $\bar{C}_I(v_i, v_i) + a/n(v_i)$ actually corresponds to the unconditional variance of the rate variable (Equation (8)),

2. if all tracts can be assimilated with points and include only one case (i.e. individual-level indicator data), then $\bar{C}_I(v_i, v_i) = \sigma_y^2 = m^*(1 - m^*)$ and the error variance term a is zero, making the binomial kriging system (9) identical to the indicator kriging system (4),
3. 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 (8) becomes $\sigma_y^2 + m/n(v_i)$, making the binomial and Poisson kriging systems identical.

Besides the population size, the kriging system accounts for the spatial correlation among geographical units through the area-to-area covariance terms $\bar{C}_I(v_i, v_j)$ that are numerically approximated by averaging the indicator covariance $C_I(\mathbf{h}) = C_I(0) - \gamma_I(\mathbf{h})$ computed between any two locations discretizing the areas v_i and v_j :

$$\bar{C}_I(v_i, v_j) = \frac{1}{P_i P_j} \sum_{s=1}^{P_i} \sum_{s'=1}^{P_j} C_I(\mathbf{u}_s - \mathbf{u}_{s'}) \quad (10)$$

where P_i and P_j are the number of points used to discretize the two areas v_i and v_j , respectively. To account for the wide range of census tract sizes, a flexible discretizing grid was used in this study: 50 points were distributed uniformly within each polygon according to a stratified random design.

A similar approach can be used to estimate incidence rate for a census tract v_α . The so-called Area-to-Area (ATA) binomial kriging involves solving system (9) where the area-to-point covariance $\bar{C}_I(v_i, \mathbf{u})$ is replaced by the area-to-area covariance $\bar{C}_I(v_i, v_\alpha)$. Note that by construction the ATA kriged incidence rate for each census tract is equal to the average of all ATP kriged rates within that tract (Goovaerts, 2006).

Simulation studies

An objective assessment of the prediction performance of the various mapping techniques requires the availability of the underlying risk maps, which are unknown in practice. Simulation provides a way to generate multiple realizations of the spatial distribution of cancer cases that can be analyzed using alternative approaches. Predicted risks can then be compared to the risk maps used in the simulation. The spatial distribution of early and late-stage cancer cases across the three counties was simulated using a two-step approach. Risk of late-stage diagnosis is first simulated at all 2,118 case residences using each of the three following scenarios:

1. smooth risk map obtained by application of a kernel smoother of type (3) to the original 2,118 data. At each location $\mathbf{u}_0$, all data $i(\mathbf{u}_\alpha)$ within a radius of 15 km were averaged using the following weighting scheme $\lambda_\alpha = \exp\{-(h_{\alpha 0})^2/a\}$, where $h_{\alpha 0}$ is the distance between the two locations $\mathbf{u}_0$ and $\mathbf{u}_\alpha$, and the range a is set to 30 km.
2. patchy risk map obtained by using the same kernel smoother as in scenario 1, except that λ_α is multiplied by five if the two locations $\mathbf{u}_0$ and $\mathbf{u}_\alpha$ are within the same census tract. This weighting scheme enhances differences among census tracts.
3. choropleth risk map identified with the map of census tract level incidence rates. The same three scenarios were applied to a grid with a 600 meters spacing, and the corresponding risk maps are displayed in Figure 2.

Each residence is classified as early or late-stage according to whether the simulated risk exceeds a probability p that is randomly drawn from a uniform distribution over $[0,1]$.

Results

Scales of variation of cancer stage data

The spatial connectivity of late-stage cancer cases was first explored using the indicator semivariogram (equation 2). To account for the wide range of separation distances between cancer cases (from a few meters to 112 km), two semivariograms with different lag classes were computed: 20 lags of 15 meters to characterize the small-scale variation of the data and 62 lags of 1km to look at the regional patterns. The first indicator semivariogram (Fig. 3A) indicates that late detection cases do not occur randomly in space, yet individual-level factors such as age or family history generate a large variability over very short distances (1st range = 48 meters). At a larger scale (Fig. 3B), the connectivity becomes direction-dependent: the semivariogram, hence the lack of connectivity, increases more slowly in the NE-SW direction (azimuth = 27° as measured in degrees clock-wise from the N-S axis). This spatial anisotropy reflects the gradient of decreasing late-stage cancer incidence towards the lake shore which was apparent on the incidence map of Fig. 1A.

The nested scales of variation detected on the experimental semivariogram were modelled using a nugget effect and two spherical models: an isotropic model with a range of 48.2 meters and an anisotropic model with a minor range of 30.9 km and a major range of 60.8 km. The fitting was conducted in two steps: the isotropic model was fitted first to the experimental curve of Fig. 3A, then the long-range model was fitted to the experimental values of Fig. 3B after subtraction of the short-range model. These two models are displayed on the same graph using a logarithmic scale for the distance (Fig. 3C).

To identify potential causes for the extreme variation over very short distances, pairs of cancer cases that belong to two different census tracts or two different age groups were excluded from the computation of the short-range semivariogram. Fig. 3D indicates a better spatial connectivity of late-stage cancer cases when only patients within the same age group are paired, while the contextual variable “census tract” has no impact at that short scale. This result is not surprising since the percentage of late-stage cancer cases varies greatly among age groups (Table 1). The spatial connectivity does not increase systematically for all age groups though. The largest range of autocorrelation (460m) is observed for the group 64–75 year old (Fig. 3E), whereas the semivariogram for age 64 and younger (Fig. 3F) displays almost no correlation. A possible explanation for such finding is that late-stage cancer in the younger group (<64 years old) might be controlled by individual characteristic and choices of lifestyle (genetics, diet, etc.), which may not be spatially autocorrelated, while in the older group (>64 years old) late-stage diagnosis reveals the effect of environmental factors over lifetime exposure, and these environmental factors might be spatially autocorrelated. Although these results might support the separate modeling of different age groups, all the data will be grouped for the subsequent analysis.

Mapping late-stage breast cancer incidence

Figure 4 shows maps of late-stage cancer incidence rates estimated on a 300 meters spacing grid using various algorithms. In each case, the estimation was based on the 32 closest cases and additional pieces of information, such as census-tract incidence rates. The reference method is a spatially adaptive filter that estimates the rate at each grid node as the arithmetical average (Fig. 4A) or the inverse-distance weighted average of the closest observations (Fig. 4B). Three kriging methods of increasing complexity are also used: ordinary kriging of indicator data (Fig. 4C), ordinary kriging of residual data where the local

mean is either identified with the noise-filtered census-tract incidence rate estimated by ATA binomial kriging (Fig. 4D–E) or the rate estimated at each grid node by ATP binomial kriging (Fig. 4F–G). For both types of binomial kriging, the 32 closest census tracts were included in the estimation. The average distance between a grid node and the 32 closest geocoded data is 4.3 km, which is much shorter than the ranges of the anisotropic regional structure modeled in Fig. 3C: 30.9 and 60.8 km. In comparison, the average separation distance for the 32 closest census tracts is 21.7 km. Census tract data thus allow one to capture information over a larger area.

The weighting scheme has a drastic impact on results of spatially adaptive filters. The incidence map created using equal weights is much smoother than the inverse-distance weighted results that exhibit the familiar "bull's eyes" pattern around data points. Intermediate variability is displayed by indicator kriging (IK) estimates: the map shows more compact features and a slightly more pronounced anisotropic pattern that is consistent with the better continuity inferred from the semivariogram in the direction parallel to the shore. The high risk area predicted by the spatially adaptive filter (equal weight) in the Central part and top right corner of the study area is also smaller on the kriged map. Incorporating census-tract information through residual kriging adds more details to the maps but generates discontinuities at the tract boundaries, in particular when the local mean is assumed constant throughout the census unit (Fig. 4E). Enhancing the impact of the census tracts also attenuates the anisotropy of the maps. Differences between the geostatistical models are the largest at the borders of the study area, which is expected since the impact of the trend model becomes preponderant in extrapolation situation.

Simulation studies

Most maps in Fig. 4 look like valid models for the spatial distribution of the risk of late-stage diagnosis across the three counties. Yet, since the reality is unknown, little can be concluded regarding the accuracy of those predictions. Simulation based on the processing of actual observations was used as a way to assess the prediction performance of alternative approaches within a realistic framework and without favouring one method over the others. Following the methodology described in the Data and Methods section, the three risk maps of Figure 2 were coupled with 10 random sets of probabilities, yielding a total of 30 simulated sets of 2,118 cancer cases (average frequency of late-stage diagnosis=0.18). Each set was processed using the spatially adaptive filters and kriging techniques illustrated in Fig. 4. To explore the impact of data scarcity on the results, each simulated set was randomly sampled (25% and 50% sampling rates) and the subsets of 529 and 1,059 cases were processed similarly, leading to a total of 90 simulation runs. To allow for a fair comparison, the estimation and modeling of all semivariograms, was performed automatically using only the subset of data available.

For each risk scenario, sampling intensity and interpolation technique, risk of late-stage diagnosis was estimated at the same locations as the reference risk map (total of 12,222 nodes). "Target" locations were defined as grid nodes where the risk of late-stage diagnosis exceeds a threshold set to 125% the area-wide rate. The following four statistics were then computed:

1. mean error of prediction, to detect any estimation bias.
2. mean absolute error of prediction, to quantify the accuracy of the estimation.
3. proportion of false positives committed to identify all "target" locations. This quantitative measure of detection efficiency corresponds to the relative area above the Receiver Operating Characteristics (ROC) curve (Swets, 1988; Goovaerts, 2008).

4. ability of interpolation methods to discriminate between target and non-target locations. The discriminatory power is measured by the ratio of averaged risk estimate at target and non-target locations (Goovaerts and Journel, 1995).

The first statistic (results not shown) indicates that all interpolation methods are unbiased. Regardless of the risk scenario indicator kriging and spatially adaptive filter with equal weights yield the smallest mean absolute error of prediction, see Table 2. The worst performance is obtained using inverse distance weights, confirming the lack of realism of the map in Fig. 4A.

Table 3 indicates that geostatistical algorithms systematically outperform spatially adaptive filters with respect to the detection efficiency of high risk values, as quantified by the ROC curve. Worst results are now obtained for spatially adaptive filters with equal weights which produce smoother risk maps, leading to an underestimation of high risk values. This criterion highlights the benefit of incorporating census-tract information. The gain is the largest for scenario 3 that assumes a constant risk within each census tract. As expected, the proportion of false positives decreases as more data become available.

The probability ratio is the only criterion according to which the inverse distance weighting scheme performs best. However, it is merely the result of extreme risk estimates (i.e. close to 0 or 1) produced by this estimator. Like the proportion of false positives, this criterion highlights the benefit of incorporating census-tract information over traditional indicator kriging.

5. Conclusions

The major difficulty in the analysis of health outcomes is that the patterns observed reflect the influence of a complex constellation of demographic, social, economic, cultural and environmental factors that likely change through time and space, and interact with the different types and scales of places where people live. It is thus primordial to incorporate the scale and spatial support of the data in their processing, as well as to account for the impact of population sizes on the reliability of rate estimates. Geostatistics provides a methodology to model the spatial correlation among rates measured over irregular geographic supports and to compute noise-free risk estimates over the same units or at much finer scales. In addition to the recently developed Poisson kriging, traditional non-parametric tools, such as indicator semivariograms and indicator kriging, are well suited to the analysis of geocoded data where the location of all cases and non-cases is known (i.e. registry data), which is different from the typical case/control mapping application when the controls are sampled. A two-step semivariogram modeling procedure and the use of lognormal scale for the distance axis on the semivariogram plot were here proposed to address the challenge created by the wide range of distances sampled. Because it accounts for data support, kriging allows the incorporation of information collected at various scales, such as socio-economic data at the census tract level, behavioral data at the county-level, and health data either aggregated or available at the individual-level (Gotway and Young, 2007). In addition, the proposed approach allows the incorporation of incompletely geocoded data (e.g. records for which only the census tract is known) without assigning them to an arbitrary location.

The analysis of breast cancer data in three Michigan counties illustrated the use of indicator semivariograms to identify major scales of variation in the data. Two nested scales were found: a substantial short-range variation that likely reflects the impact of individual-level factors, such as age at diagnosis or family history, on the spread of the disease and a regional anisotropic pattern corresponding to a decreasing trend in late-stage diagnosis as one gets closer to the lake shore urban communities.

Geostatistics offers a wide range of strategies to combine area-based and individual-level health data into the mapping of incidence rates. Area-to-Point kriging can first be used to disaggregate area-based data to the nodes of the interpolation grid. This information can then be incorporated as local mean in the interpolation. The use of area-based data directly into the interpolation (i.e. without disaggregation) can cause discontinuities at census tract borders. Results thus depend on the rather subjective choice of an interpolation strategy and trend model that should be guided by information on putative factors. Current research is exploring the use of secondary information, such as proximity to screening facilities and area-based measure of economic distress, to model these spatial trends, in particular in sparsely populated areas where individual-level data are lacking and rate data are less reliable.

Simulation studies demonstrated the overall better prediction performance of kriging over spatially adaptive filters, regardless the type of risk maps. As expected, less complex risk maps will always be easier to predict, in particular as more data becomes available. Incorporating census-tract data was especially useful to reduce the proportion of false positives and to improve prediction accuracy in sparsely populated areas.

Future research will investigate the contribution of positional uncertainty (i.e. geocoding errors) to the short-range variation displayed by indicator semivariograms. One should also go beyond the interpolation of incidence rates from individual-level data and conduct tests of hypothesis, such as the delineation of clusters of significantly higher late-stage incidence rates or the detection of significant disparities among ethnic groups or genders. Most studies on kernel density estimation introduced permutation procedures that relied on a random re-labeling of all observations (cases and non-cases) to test whether the observed rate was larger than the expected rate, given the null hypothesis. The expected rate was either constant across the study area (Rushton and Lolonis, 1996) or predicted from individual characteristics using logistic regression (James et al., 2004; Rushton et al., 2004). In the future, one should explore the use of geostatistical simulation to incorporate spatial autocorrelation in the randomization procedure.

Acknowledgments

This research was funded by grant R44-CA132347-01 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

- Barry J, Breen N. The importance of place of residence in predicting late-stage diagnosis of breast or cervical cancer. *Health & Place*. 2005; 11:15–29. [PubMed: 15550353]
- Best, NG.; Arnold, RA.; Thomas, A.; Waller, LA.; Conlon, EM. Bayesian models for spatially correlated disease and exposure data. In: Bernardo, JM.; Berger, JO.; Dawid, AP.; Smith, AFM., editors. *Bayesian Statistics 6*. Oxford, UK: Oxford University Press; 1999. p. 131-156.
- Best NG, Richardson S, Thomson A. A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*. 2005; 14:35–59. [PubMed: 15690999]
- Bithell JF. An application of density estimation to geographical epidemiology. *Statistics in Medicine*. 1990; 9:691–701. [PubMed: 2218172]
- Bithell JF. A classification of disease mapping methods. *Statistics in Medicine*. 2000; 19:2203–2215. [PubMed: 10960848]
- French JL, Wand MP. Generalized additive models for cancer mapping with incomplete covariates. *Biostatistics*. 2004; 5:177–191. [PubMed: 15054024]
- Goovaerts, P. *Geostatistics for Natural Resources Evaluation*. New York, NY: Oxford University Press; 1997.

- Goovaerts P. 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. *International Journal of Health Geographics*. 2006; 5:52. [PubMed: 17137504]
- Goovaerts P. Accounting for rate instability and spatial patterns in the boundary analysis of cancer mortality maps. *Environmental and Ecological Statistics*. 2008; 15(4):421–446. [PubMed: 19023455]
- Goovaerts P, Journel AG. Integrating soil map information in modelling the spatial variation of continuous soil properties. *European Journal of Soil Science*. 1995; 46:397–414.
- Gotway CA, Young LJ. Combining incompatible spatial data. *Journal of the American Statistical Association*. 2002; 97(459):632–648.
- Gotway CA, Young LJ. A geostatistical approach to linking geographically-aggregated data from different sources. *Journal of Computational and Graphical Statistics*. 2007; 16(1):115–135.
- Han D, Rogerson PA, Bonner MR, Nie J, Vena JE, Muti P, Trevisan M, Freudenheim JL. Assessing spatio-temporal variability of risk surfaces using residential history data in a case control study of breast cancer. *International Journal of Health Geographics*. 2005; 4:9. [PubMed: 15826315]
- James L, Matthews I, Nix B. Spatial contouring of risk: a tool for environmental epidemiology. *Epidemiology*. 2004; 15:287–292. [PubMed: 15097008]
- Journel AG. Nonparametric estimation of spatial distributions. *Mathematical Geology*. 1983; 15:445–468.
- Kammann EE, Wand MP. Geoadditive models. *Applied Statistics*. 2003; 52:1–18.
- Kelsall JE, Diggle PJ. Nonparametric estimation of spatial variation in relative risk. *Statistics in Medicine*. 1995; 14:2335–2342. [PubMed: 8711273]
- Kelsall JE, Diggle PJ. Spatial variation in risk of disease: a nonparametric binary regression approach. *Applied Statistics*. 1998; 47:559–573.
- Kyriakidis P. A geostatistical framework for area-to-point spatial interpolation. *Geographical Analysis*. 2004; 36(2):259–289.
- Lawson AB. Tutorial in biostatistics: Disease map reconstruction. *Statistics in Medicine*. 2000; 20:2183–2204. [PubMed: 11439429]
- McLeod KS. Our sense of Snow: the myth of John Snow in medical geography. *Social Science & Medicine*. 2000; 50:923–935. [PubMed: 10714917]
- Monestiez P, Dubroca L, Bonnin E, Durbec JP, Guinet C. Geostatistical modelling of spatial distribution of *Balenoptera physalus* in the northwestern Mediterranean Sea from sparse count data and heterogeneous observation efforts. *Ecological Modelling*. 2006; 193:615–628.
- Oliver MA, Webster R, Lajaunie C, Muir KR, Parkes SE, Cameron AH, Stevens MCG, Mann JR. Binomial cokriging for estimating and mapping the risk of childhood cancer. *IMA Journal of Mathematics Applied in Medicine and Biology*. 1998; 15:279–297. [PubMed: 9773520]
- Pickle LW, Szczur M, Lewis DR, Stinchcomb DG. The crossroads of GIS and health information: a workshop on developing a research agenda to improve cancer control. *International Journal of Health Geographics*. 2006; 5:51. [PubMed: 17118204]
- Rossi RE, Mulla DJ, Journel AG, Franz EH. Geostatistical tools for modeling and interpreting ecological spatial dependence. *Ecological Monographs*. 1992; 62:277–314.
- Rushton G. Public health, GIS, and spatial analytic tools. *Annual Review of Public Health*. 2003; 24:43–56.
- Rushton G, Lolonis P. Exploratory spatial analysis of birth defect rates in an urban population. *Statistics in Medicine*. 1996; 15:717–726. [PubMed: 9132899]
- Rushton G, Peleg I, Banerjee A, Smith G, West M. Analyzing geographic patterns of disease incidence: rates of late-stage colorectal cancer in Iowa. *Journal of Medical Systems*. 2004; 28:223–236. [PubMed: 15446614]
- Swets JA. Measuring the accuracy of diagnostic systems. *Science*. 1988; 240:1285–1293. [PubMed: 3287615]
- Talbot TO, Kulldorff M, Forand SP, Haley VB. Evaluation of spatial filters to create smoothed maps of health data. *Statistics in Medicine*. 2000; 19:2399–2408. [PubMed: 10960861]

- Tiwari, C.; Rushton, G. Using spatially adaptive filters to map late stage colorectal cancer incidence in Iowa. In: Fisher, P., editor. *Advances in Spatial data handling II*. Berlin: Springer-Verlag; 2004. p. 665-676.
- Tobler W. Smooth Pycnophylactic Interpolation for Geographical Regions. *Journal of the American Statistical Association*. 1979; 74:519–536. [PubMed: 12310706]
- Vieira V, Webster T, Aschengrau A, Ozonoff D. A method for spatial analysis of risk in a population-based case-control study. *International Journal of Hygiene and Environmental Health*. 2002; 205:115–120. [PubMed: 12018004]
- Vieira V, Webster T, Weinberg J, Aschengrau A, Ozonoff D. Spatial analysis of lung, colorectal, and breast cancer on Cape Cod: an application of generalized additive models to case-control data. *Environmental Health*. 2005; 4:11. [PubMed: 15955253]
- 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, Chile: GECAMIN Ltd; 2008. p. 981-990.
- Webster R, Oliver MA, Muir KR, Mann JR. Kriging the local risk of a rare disease from a register of diagnoses. *Geographical Analysis*. 1994; 26:168–185.
- Webster T, Vieira V, Weinberg J, Aschengrau A. Method for mapping population-based case-control studies: an application using general additive models. *International Journal of Health Geographics*. 2006; 5:26. [PubMed: 16764727]
- Yabsley MJ, Wimberly MC, Stallknecht DE, Little SE, Davidson WR. Spatial analysis of the distribution of *Ehrlichia chaffeensis*, causative agent of human monocytotropic ehrlichiosis, across a multi-state region. *American Journal of Tropical Medicine and Hygiene*. 2005; 72:840–850. [PubMed: 15964974]
- 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 Pub. # 01-4969. Bethesda, MD: 2001.

Fig. 1.
Maps of late-stage breast cancer incidence rate (A) and number of cancer cases (B) in three Michigan counties, by census tract, 1985–2002.

Fig. 2.
Reference risk maps (3 scenarios for the spatial pattern) used in the simulation studies.

Fig. 3.

Experimental indicator semivariograms with the model fitted (solid line) computed for different options: omnidirectional for short distances (A), directional for long distances (B), all distances combined (C), excluding data pairs belonging to different census tracts or age groups (D), using only cases in the age groups 0–64 (E) and 64–75 (F).

Fig. 4.

Isopleth maps of late-stage breast cancer incidence rate estimated using: spatially adaptive filters based on arithmetical (A) or inverse-distance weighted (B) averages, indicator kriging (C), kriging of residual data where the local mean is either identified with the noise-filtered census-tract incidence rate (D,E) or the rate estimated at each grid node by ATP binomial kriging (F,G). The same legend is used for all maps.

Table 1

Sample characteristics for each age group.

Age group	Number of cases	Late stage (%)
0–50	37	21.62
50–64	552	20.11
65–74	937	18.46
75–84	512	13.87
85–120	80	23.75
All	2118	18.04

Results of simulation studies for three risk scenarios and five interpolation procedures averaged over 10 realizations; mean absolute error of prediction.

Table 2

Risk scenario	Spatially adaptive filters			Kriging	
	Wgt=1/n	Wgt=1/h	IK	RK-ATA	RK-ATP
25% sampling intensity (N=529)					
I	0.068	0.124	0.068	0.079	0.078
II	0.077	0.127	0.077	0.085	0.086
III	0.066	0.127	0.066	0.076	0.077
50% sampling intensity (N=1,059)					
I	0.065	0.120	0.066	0.074	0.071
II	0.074	0.122	0.074	0.078	0.079
III	0.067	0.124	0.067	0.071	0.074
100% sampling intensity (N=2,118)					
I	0.062	0.120	0.062	0.069	0.076
II	0.070	0.121	0.069	0.071	0.072
III	0.064	0.123	0.064	0.063	0.067

Note: Kriging methods are indicator kriging (IK), and residual kriging (RK) using local means estimated using Area-to-Area (ATA) or Area-to-Point (ATP) binomial kriging. Bold numbers denote best performance for each risk scenario.

Table 3

Results of simulation studies for three risk scenarios and five interpolation procedures averaged over 10 realizations; proportion of false positives (threshold = 125% area-wide rate).

Risk scenario	Spatially adaptive filters			Kriging	
	Wgt=1/n	Wgt=1/h	IK	RK-ATA	RK-ATP
25% sampling intensity (N=529)					
I	0.440	0.383	0.387	0.377	0.371
II	0.450	0.374	0.412	0.372	0.381
III	0.460	0.415	0.419	0.359	0.383
50% sampling intensity (N=1,059)					
I	0.385	0.353	0.349	0.324	0.314
II	0.388	0.345	0.344	0.306	0.319
III	0.409	0.386	0.373	0.301	0.331
100% sampling intensity (N=2,118)					
I	0.302	0.317	0.265	0.264	0.255
II	0.317	0.314	0.276	0.253	0.264
III	0.361	0.367	0.320	0.262	0.292

Note: Kriging methods are indicator kriging (IK), and residual kriging (RK) using local means estimated using Area-to-Area (ATA) or Area-to-Point (ATP) binomial kriging. Bold numbers denote best performance for each risk scenario.

Results of simulation studies for three risk scenarios and five interpolation procedures averaged over 10 realizations; probability ratio (target locations over non-target locations).

Table 4

Risk scenario	Spatially adaptive filters		Univariate kriging		
	Wgt=1/n	Wgt=1/h	IK	RK-ATA	RK-ATP
25% sampling intensity (N=529)					
I	1.156	1.363	1.155	1.196	1.204
II	1.146	1.384	1.145	1.206	1.195
III	1.133	1.261	1.131	1.250	1.192
50% sampling intensity (N=1,059)					
I	1.239	1.438	1.237	1.280	1.293
II	1.240	1.490	1.248	1.328	1.297
III	1.204	1.328	1.202	1.349	1.271
100% sampling intensity (N=2,118)					
I	1.398	1.556	1.397	1.416	1.429
II	1.375	1.596	1.382	1.449	1.426
III	1.281	1.361	1.284	1.419	1.341

Note: Kriging methods are indicator kriging (IK), and residual kriging (RK) using local means estimated using Area-to-Area (ATA) or Area-to-Point (ATP) binomial kriging. Bold numbers denote best performance for each risk scenario.