Title
Adjusting for sampling variability in sparse data: geostatistical approaches to disease mapping

Journal
International Journal of Health Geographics, 10(1)

ISSN
1476-072X

Authors
Hampton, Kristen H
Serre, Marc L
Gesink, Dionne C
et al.

Publication Date
2011-10-06

DOI
http://dx.doi.org/10.1186/1476-072X-10-54

Peer reviewed

INTERNATIONAL JOURNAL OF HEALTH GEOGRAPHICS

METHODOLOGY Open Access

Adjusting for sampling variability in sparse data: geostatistical approaches to disease mapping

Kristen H Hampton[1], Marc L Serre[2*], Dionne C Gesink[3], Christopher D Pilcher[4] and William C Miller[1,5]

Abstract

Background: Disease maps of crude rates from routinely collected health data indexed at a small geographical resolution pose specific statistical problems due to the sparse nature of the data. Spatial smoothers allow areas to borrow strength from neighboring regions to produce a more stable estimate of the areal value. Geostatistical smoothers are able to quantify the uncertainty in smoothed rate estimates without a high computational burden. In this paper, we introduce a uniform model extension of Bayesian Maximum Entropy (UMBME) and compare its performance to that of Poisson kriging in measures of smoothing strength and estimation accuracy as applied to simulated data and the real data example of HIV infection in North Carolina. The aim is to produce more reliable maps of disease rates in small areas to improve identification of spatial trends at the local level.

Results: In all data environments, Poisson kriging exhibited greater smoothing strength than UMBME. With the simulated data where the true latent rate of infection was known, Poisson kriging resulted in greater estimation accuracy with data that displayed low spatial autocorrelation, while UMBME provided more accurate estimators with data that displayed higher spatial autocorrelation. With the HIV data, UMBME performed slightly better than Poisson kriging in cross-validatory predictive checks, with both models performing better than the observed data model with no smoothing.

Conclusions: Smoothing methods have different advantages depending upon both internal model assumptions that affect smoothing strength and external data environments, such as spatial correlation of the observed data. Further model comparisons in different data environments are required to provide public health practitioners with guidelines needed in choosing the most appropriate smoothing method for their particular health dataset.

Keywords: disease mapping, sampling variability, spatial distribution, spatial analysis, epidemiological methods

Background

Disease maps that summarize the spatial and spatio-temporal variation in rates of disease have a wide range of applications, from hypothesis generation to public health surveillance. Identification of areas with unusually high or low rates may indicate clusters or 'hot spots' of disease that can aid decisions regarding intervention or prevention programs, allocation of health care resources, or provide context for further epidemiological studies [1-4]. At the same time, the utility of disease mapping often depends on how accurately the value being mapped estimates the spatial process of interest. Estimating disease risk requires a defined, closed population and cannot be measured directly with surveillance data; accordingly, incidence rates are often used to approximate disease risk [5]. However, calculating crude rates from routinely collected health data indexed at a small geographical resolution poses specific statistical problems due to the sparse nature of the data, especially for rare diseases [1-6]. In particular, crude rate maps of small areas will be dominated by sampling variability. Due to variation in population size among areas, a map displaying crude rates will tend to be dominated by areas with small populations since small changes to the observed number of cases will result in large changes to the rate [5-7]. Meanwhile, observed high rates based on small populations are likely to be artificially elevated due to the high variability in the estimates. In other words, error due to sampling variability introduces observational noise into the map that may obscure the

* Correspondence: marc_serre@unc.edu
[2]Department of Environmental Sciences and Engineering, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
Full list of author information is available at the end of the article

underlying spatial process of interest, and, if not adjusted for, lead to incorrect inference about the spatial pattern of interest.

One general method to improve the stability of observed rates is to increase population sizes by upscaling from finer to coarser resolution levels, such as from census tracts or ZIP codes to counties. Increasing the aggregation area, however, causes a loss in the resolution of the data, thereby masking spatial details needed to identify, analyze, and monitor health problems at the community level [5,7,8]. At the same time, analyzing health data at too fine a resolution, such as with case event or point maps, can compromise patient confidentiality and may be misleading in some circumstances because the underlying population density is not considered [8]. This paper focuses on the issue of sampling variability in disease maps at a geographical resolution corresponding to an aggregation of counts over *fixed small* geographical regions - in this case, postal codes of residence (US postal service 'ZIP codes').

Both deterministic and model-based approaches have been proposed to address the issue of sampling variability, also referred to as the 'small number problem,' when mapping health data at fine resolutions [1,5,9]. These approaches reduce the noise in rates of small areas through 'spatial smoothing,' which allows areas to borrow strength from neighboring regions to produce a more stable estimate of the value associated with each region [2-4,6,7,10,11]. Among the different methods, however, a tradeoff exists between the computational complexity of each method and the ability to account for spatial correlation and uncertainty in the smoothed results [12]. For example, methods such as disk smoothing, population-weighted averages, and empirical Bayes estimates have low computing requirements but do not account for spatial autocorrelation or quantify the amount of uncertainty in the smoothed rates [12,13]. In contrast, full Bayesian hierarchical models are able to incorporate multiple covariates in the model parameters to yield the full posterior distribution of the estimated rate, yet are computationally cumbersome due to time-consuming iterative procedures and challenging for non-statisticians to implement and interpret due to their complexity [2,12,13].

Geostatistical methods modified to account for the heteroscedasticity of health data, namely that the variance at each location varies as a function of the population size, offer a compromise between computational burden and quantification of the uncertainty in the smoothed rate estimate. Methods such as Poisson kriging are able to account for spatial correlation and yield a full posterior distribution while being computationally faster than fully Bayesian hierarchical models [12-14]. Furthermore, simulation studies of cancer mortality showed that Poisson kriging outperformed simple population-weighted averages, empirical Bayes smoothers, and the Besag-York-Mollié Bayesian hierarchical model in measures of estimation accuracy and degree of smoothing, particularly when background rate values were spatially correlated [12,13]. While spatial smoothing is needed to improve the stability of observed rates based on small populations, too much smoothing reduces the ability to identify areas of high or low rates that may indicate clusters or outbreaks of disease.

In this paper, we examine extensions of two geostatistical methods, ordinary kriging and Bayesian Maximum Entropy (BME), which have been used extensively in the analysis of fields exhibiting spatial variability and are computationally efficient [15-22]. Traditional geostatistics do not account for the heteroscedasticity of disease rates and must be modified to address both the numerator and denominator of rate data [12-14]. Poisson kriging as an extension of ordinary kriging was first introduced by Monestiez et al. [23] and applied to health data by Goovaerts [12]. Here, we introduce a uniform model extension of BME (UMBME) that, like Poisson kriging, can account for changes in variance due to population size and compare the performance of UMBME to that of Poisson kriging in measures of estimation accuracy and smoothing strength. We first examine differences in underlying model assumptions by applying each model to simulated datasets where the true spatial variation in the data is known. We then apply both models to the real data example of human immunodeficiency virus (HIV) infection with a spatial analysis of infection occurring in North Carolina HIV testing data from 1994 to 2002. Attempts to describe or monitor the distribution of HIV cases detected through testing have been limited by the variability in sampling that is inherent in the testing situation.

Methods

Model definitions

In traditional epidemiology, the incidence rate of disease is defined as the number of new cases in the at-risk population divided by the person-time, or summation of each person's observation time, over a specified period [5,24]. With rare diseases at steady state, the loss of person-time per diseased person is minimal, and the total person-time at risk may be approximated by the total size of the at-risk population over the period times the length of the period (day, month, year) [5,24]. In other words, the incidence rate corresponds to a proportion of new cases in the at-risk population divided by the time duration considered to enumerate the new cases.

Following this concept, we define in the spatial context the *latent* incidence rate as the theoretically possible rate of infection built on a hypothetical infinite

underlying population at risk. In this paper, we consider observation datasets aggregated to a single time period t, resulting in space only analyses and temporally independent maps that can be used to examine the spatial variability of the random fields. Therefore, without loss of generality, the latent incidence rate of disease in a given region i can be defined as:

$$X_i = \lim_{n_i \to \infty} \frac{Y_i}{n_i}, \qquad i = 1, \ldots I \tag{1}$$

where X_i is the latent disease rate, Y_i is the number of new cases of disease, and n_i is the size of the population at risk in area i. In other words, as the size of the population n_i reaches infinity, then the observed disease rate Y_i/n_i approaches the latent disease rate for that location. In this hypothetical situation, sampling variability is minimized because all regions being mapped are of equally large populations and small changes in the number of cases do not result in large changes to the observed rate.

In practice, however, only a finite population at risk n_i in area i may be measured, producing a finite number of cases Y_i and an observed rate $R_i = Y_i/n_i$. Note that even when n_i is an exhaustive count of the entire population at risk in area i, maps of the observed rate R_i are subject to noise due to variation in population size among regions, and areas with rates based on small populations will tend to dominate the map. The difference between the observed rate R_i and latent rate X_i is directly related to the observation that R_i can only be defined in increments of $1/n_i$, while the underlying latent rate can take any value on the real number continuum. The difference between R_i and X_i is inherently a measurement error ε_i that can be expressed as:

$$R_i = X_i + \varepsilon_i \tag{2}$$

where we refer to ε_i as the error due to sampling variability in area i. In this paper, we contrast two approaches in modeling sampling variability, Poisson kriging and the UMBME method, which use unbounded and bounded distributions, respectively, for the error ε_i.

Poisson kriging

With Poisson kriging, the observed number of cases is modeled as a random variable Y_i that follows a Poisson distribution with one parameter. This parameter is the expected number of cases defined as the product of the population size n_i by the latent rate X_i, such that:

$$Y_i \mid X_i \sim \text{Poisson}\,(n_i X_i) \tag{3}$$

As detailed by Goovaerts [12], the Poisson kriging latent rate in area i is estimated as a linear combination of K neighboring observed rates, such that:

$$\hat{X}_i^{PK} = \sum_{j=1}^{K} \lambda_{ij} R_j \tag{4}$$

where the weights λ_{ij} are derived from a Poisson kriging specific system of linear equations, in which the sampling variability at i is accounted for by a normally distributed error ε_i with variance equal to $\sigma_i^2 = m^*/n_i$, where m^* is the population-weighted mean of the observed rates [12]. We implemented the Poisson kriging adaptation of ordinary kriging in the MATLAB programming environment [25]. Detailed descriptions of Poisson kriging may be found in Goovaerts (2005) and Goovaerts and Gebreab (2008).

Uniform model extension of Bayesian Maximum Entropy (UMBME)

An underlying assumption of UMBME is that the observed population n_i is a representative sample of the population at risk in area i and measured without error. In this case, the observed number of cases Y_i given the latent rate X_i can be interpreted as the product of the population size and latent rate rounded to the nearest integer, such that:

$$Y_i \mid X_i = \text{round}\,(n_i X_i) \tag{5}$$

The discrete nature of case counts requires Y_i to be a whole number value. Therefore, when given an observed case count Y_i and population n_i, the latent rate X_i follows a uniform distribution bounded by the rounding error, such that:

$$X_i \mid R_i = U\left(R_i - \frac{0.5}{n_i}, R_i + \frac{0.5}{n_i}\right) \tag{6}$$

where R_i is the observed rate Y_i/n_i. In other words, in UMBME the measurement error ε_i is assumed to be uniformly distributed in an interval of size $1/n_i$, which expresses the fact that R_i is interpreted as an observation of the latent rate X_i in increments of $1/n_i$. To derive the UMBME latent rate estimate from observed values, the BME method treats the observed values as 'soft data' with measures of uncertainty as defined by the uniform distribution. As detailed by Christakos et al. [17], the BME method processes information known about the rate field in three main stages: (i) Structural (or prior) stage, (ii) Specificatory (or meta-prior) stage, and (iii) Integration stage. Programs to perform each stage were derived from functions found in the BMElib software package for the MATLAB programming environment [25,26]. Detailed descriptions of the BME method may be found in Christakos and Li (1998) and Christakos et al. (2002), among others. A general introduction to information processing in BME is presented in Supplementary Materials, Appendix A.

In this study, the UMBME estimate $\hat{X}_i^{UMBME}$ of the latent rate in area i is the expected value of the posterior probability density function (pdf) derived in the Integration stage. Other possible estimators include the mode or median of the posterior pdf. Furthermore, the variance of the posterior pdf, or BME variance, provides a useful measure of the estimation uncertainty.

Neighborhood selection

Poisson kriging and UMBME offer at least two advantages in neighborhood selection over their counterparts, particularly simple smoothers that weight all neighbors falling within an arbitrary fixed distance of areal centroids. First, both methods have the ability to limit the number of neighbors selected to a user-defined maximum, which can account for changes in the spatial density of centroids, and thus the size of geographical units, across the study area [12,17]. For example, an urban area may have dozens of neighbors falling within a fixed distance while a rural area may only have a handful of neighbors falling within the same distance. Without limitations on the number of selected neighbors, the urban area would have significantly more contributors to its smoothed value than the rural area. Second, Poisson kriging and UMBME are able to determine smoothing weights based upon the data's covariance model. If the neighborhood distance selected is large enough such that it includes all the neighborhood information, then increasing the distance or number of points does not add more information or change the model estimation.

Simulated data analysis

We applied both Poisson kriging and UMBME to simulated data where the true spatial variation in the latent rate field $X(\boldsymbol{s})$ is known in order to examine the effect of underlying assumptions on each model's performance in measures of estimation accuracy and smoothing strength. Simulations allow us to both model the true latent disease rate and to control variables in order to explore how changes impact the smoothed result [12,13]. For example, we considered a study area divided into square regions with the spatial location $\boldsymbol{s}$ of each region defined as the center of the corresponding square (Figure 1), which minimized the effect of spatial support and the modifiable areal unit problem on the smoothed estimates since regions were of uniform size and shape [12]. With real data, not knowing the true latent rate or relative impact of error sources makes it difficult to assess the accuracy and validity of different smoothing methods.

Since sampling variability generates the greatest distortion in studies of rare diseases, a low latent rate field $X(\boldsymbol{s})$ was simulated where the latent rate of disease in each region ranged in value from 0 to 7.5 per 1,000 persons (Figure 1A). We modeled the latent rate field by first identifying one region each of relative low, medium, and high latent rates with an associated decay rate. We then calculated the latent rates for surrounding areas based upon the distance from each central location. The resulting disease field where higher rates tend to be grouped around a central location follows the core area theory of sexually transmitted diseases such as syphilis [19,27].

The sample population size $n(\boldsymbol{s})$ was then determined using a random number generator, resulting in values ranging from 1 to 767 (Figure 1B). While general population density tends to be spatially correlated, it is rare for the populations sampled by public health investigations and surveillance systems to be perfectly representative of the general population. Instead, factors such as outreach efforts and concentrations of special populations can cause the populations sampled in neighboring areas to vary significantly. Furthermore, population density is dependent on geographic scale. Even in urban study areas, the population density of, for example, census block groups may represent a random patchwork depending on the locations of commercial and residential areas.

A fundamental difference in the model assumptions of Poisson kriging and UMBME is the probability distribution of the observed cases, $Y(\boldsymbol{s})$, given the latent rate, $X(\boldsymbol{s})$. If the observed cases, $Y(\boldsymbol{s})$, are drawn from a Poisson distribution, then Poisson kriging should outperform UMBME in estimating the latent rate. Conversely, if $Y(\boldsymbol{s})$ integer values are derived from the product of the latent rate, $X(\boldsymbol{s})$, and population size, $n(\boldsymbol{s})$, then UMBME should outperform Poisson kriging in estimation accuracy. We derived two realizations of the case count, $Y(\boldsymbol{s})$, to verify these characteristics and examine differences in spatial variability between assumptions. Given the latent rate field, $X(\boldsymbol{s})$, and sample population, $n(\boldsymbol{s})$, one realization of $Y(\boldsymbol{s})$ was sampled from a Poisson distribution (Figure 2A), while the other resulted from the product of $X(\boldsymbol{s})$ and $n(\boldsymbol{s})$ (Figure 2B). In both realizations, $Y(\boldsymbol{s})$ ranged in value from 0 to 3 cases.

For a given realization, we have the simulated (true) $X(\boldsymbol{s})$ and the observed rate, $R(\boldsymbol{s})$, from which we calculated both the Poisson kriging estimate, $\hat{X}_i^{PK}$, and UMBME estimate, $\hat{X}_i^{UMBME}$. We then calculated two statistics as measures of estimation accuracy: the mean square error (MSE) and Lin's Concordance Correlation Coefficient (LCCC) between the estimated value ($\hat{X}_i^{PK}$ or $\hat{X}_i^{UMBME}$) and the simulated true value for $X(\boldsymbol{s})$. Divergence between the estimated and true latent rate values decreases as the MSE approaches zero and LCCC

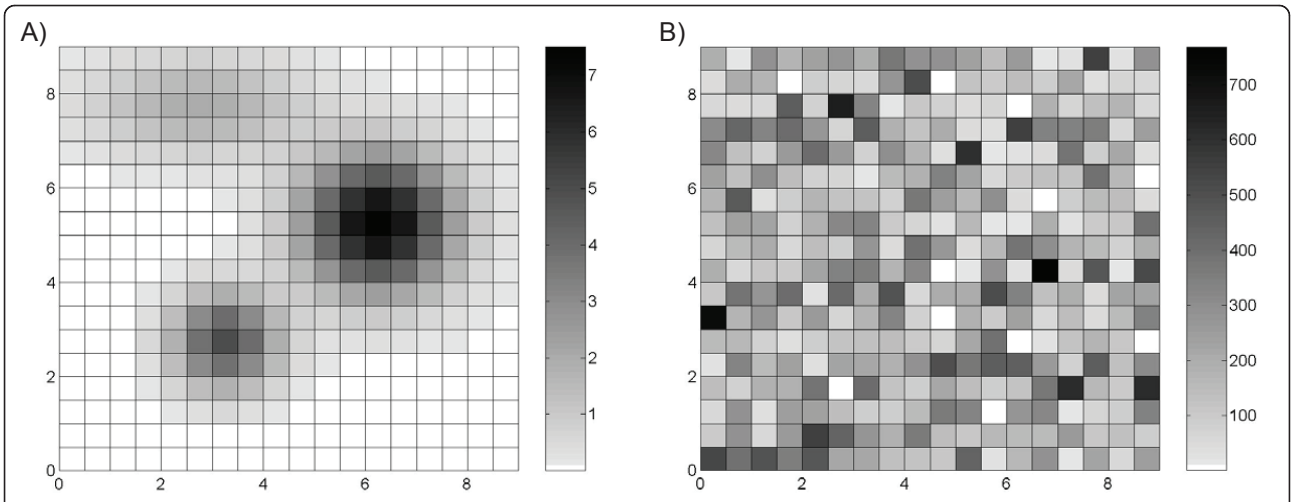

Figure 1 Maps of the simulated A) latent rate field *X(s)* and B) sample population size *n(s)*. The values of *X*(**s**) and *n*(**s**) ranged from 0 to 7.5 per 1,000 persons and from 1 to 767 persons, respectively.

approaches one [28]. Meanwhile, smoothing strength was described as the degree to which each technique modified high or low observed rates towards the global mean. More smoothing indicates greater movement towards the global mean and less spatial variability in estimated values. While smoothing strength may be qualitatively assessed by examining, for example, whether regions of high observed rates are discernable in the estimated map, we also calculated the mean absolute difference (MAD) between estimated values and the mean of observed rates. As the MAD approaches zero, the average deviation of estimated values from the global mean decreases, indicating greater smoothing of observed values.

Mapping HIV in North Carolina

To examine model performance in real world situations, we mapped HIV disease rates in North Carolina using both Poisson kriging and UMBME. Previous studies of sexually transmitted infections (STIs) demonstrate that cases tend to be concentrated in residential core areas with high rates of infection in small, definable geographic areas [19,21,27,29-31]. In North Carolina, approximately 1,700 new diagnoses of HIV disease are reported to the North Carolina public health surveillance system each year, with about 25 per cent of North Carolina's HIV disease reports consistently coming from rural, or non-metropolitan, areas since the early 1990s [32]. Regional differences in the spread of HIV highlight

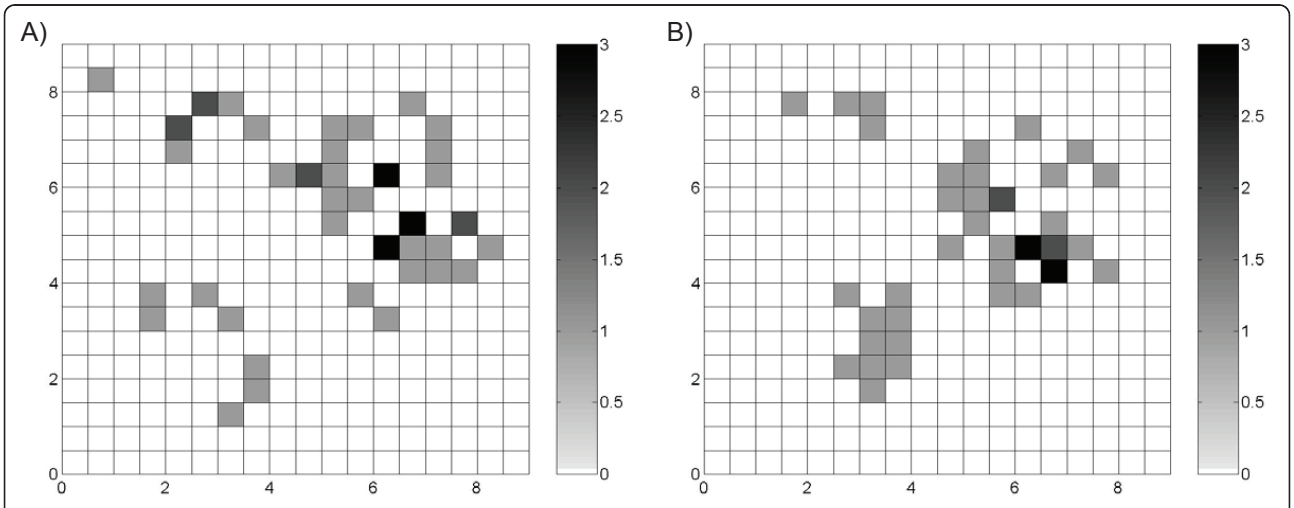

Figure 2 Maps of the simulated case count *Y(s)* under the A) Poisson assumption and B) uniform assumption. Under the Poisson assumption, the observed cases are drawn from a Poisson distribution. Under the uniform assumption, the observed cases are derived from the product of the latent rate and population size. In both realizations, the values of *Y*(**s**) ranged from 0 to 3 cases.

the necessity for accurate information on the spatial distribution of the disease in order to effectively target prevention activities and identify possible community-level risk factors contributing to the epidemic.

In this analysis, we calculated rates of new HIV disease reports in North Carolina's testing population from 1994 to 2002. Specifically, clients at risk of HIV infection presenting to publicly-funded Voluntary Counseling and Testing (VCT) sites distributed across the state of North Carolina were tested for long-term HIV infection. ZIP code of residence was recorded in a research database for each de-identified test result. Records for clients reporting previous positive HIV tests were excluded from analysis. The remaining records were geocoded to the residential-weighted centroid of the reported ZIP code of residence [33]. There were 938,889 total tests and 5,677 new HIV cases reported between 1994 and 2002, of which 96% to 97% matched to a location. Reasons for failing to geocode included having a missing, invalid, or out-of-state ZIP code. We modeled the observed HIV infection rate using Poisson kriging and UMBME and evaluated model performance.

In real world situations, the true latent rate of disease is unknown. Therefore, in order to assess model fit, we used cross-validatory predictive checks, which have been shown to be useful in determining whether divergent rates are due to poor model fit or to actual 'hot-spots' of disease that warrant further investigation [34,35]. The basic concept of cross-validation is to remove each observed rate in turn, in this case observed rates assigned to a ZIP code population centroid, then reanalyze the original data without that observation, and assess the model's ability to predict the removed area's data value. Estimation accuracy was then measured as the MSE and LCCC between the model-predicted value and the observed rate. Smoothing strength was assessed both qualitatively and as the MAD between model-predicted values and the mean of observed rates.

Results and discussion

Simulated data analysis

In applying smoothing models to the simulated rate field $R(\mathbf{s})$, we considered first the spatial moments of the latent rate spatial random field $X(\mathbf{s})$. In the absence of prior information regarding the general behavior of $X(\mathbf{s})$, we assigned no mean trend to the model and used the information provided by the rate field $R(\mathbf{s})$ directly in our calculations rather than as a residual with the mean trend removed. This enabled us to model the spatial variability of $X(\mathbf{s})$ using the covariance of $R(\mathbf{s})$. A useful feature of the covariance model is that it does not require prior distributional assumptions of the data and can be applied to both the Poisson and uniform simulations. Other estimators, such as Monestiez et al.'s population-weighted semivariogram and Yu et al.'s Poisson regression model in BME, have measurable advantages but are dependent on the Poisson assumption [23,36]. Furthermore, covariance plots provide a quantitative assessment of the correlation between pairs of points and are useful in assessing the strength and scale of the disease pattern [19]. For example, the covariance range, or distance where the curve becomes asymptotic to the x-axis, may be used to identify the neighborhood of influence around an observation point. Observations within the covariance range may influence values at the current location, whereas observations outside this range may not be influential. In addition, local disease patterns may be described by the behavior of the covariance model near the origin. Steep curves indicate rapid change over a short distance, while long-tailed curves indicate slower change and less variability over the same distance.

Poisson assumption

The experimental covariance calculated from the corresponding observed rates shows the change in the correlation between pairs of points as the distance between points increase in space. For the Poisson assumption, the experimental covariance indicated high local variability with a neighborhood of influence extending less than 2 spatial units from each point (Figure 3). The corresponding $R(\mathbf{s})$-covariance model (Figure 3A) was obtained by fitting a nugget-exponential-Gaussian nested model to the experimental $R(\mathbf{s})$-covariance values, such that

$$c_R(\rho) = c_1\delta_\rho + c_2 \exp\left(\frac{-3\rho}{a_{\rho 1}}\right) + c_3 \exp\left(\frac{-3\rho^2}{a_{\rho 2}{}^2}\right) \quad (7)$$

where ρ is the spatial lag, c_1, c_2, and c_3 are constants whose sum equals the variance of $R(\mathbf{s})$, δ_ρ is the Kronecker delta function, and $a_{\rho 1}$ and $a_{\rho 2}$ represent the spatial ranges of the exponential and Gaussian components, respectively. The nugget component of this covariance model corresponds to the initial drop of the covariance, while the exponential and Gaussian components correspond to the tail of the covariance curve. This covariance model is adequate if we are interested in mapping the observed rate field $R(\mathbf{s})$ over the geographical region of interest using observed rate data. On the other hand, if we are interested in mapping the true latent rate field $X(\mathbf{s})$, we then need to model its covariance function. We obtain the $X(\mathbf{s})$-covariance model by simply recognizing that the nugget component (initial drop of covariance) of the $R(\mathbf{s})$-covariance is due to the sampling variability error term ε defined in 2. Therefore, we obtained the $X(\mathbf{s})$-covariance model (Figure 3B) by removing the nugget component of the $R(\mathbf{s})$-covariance

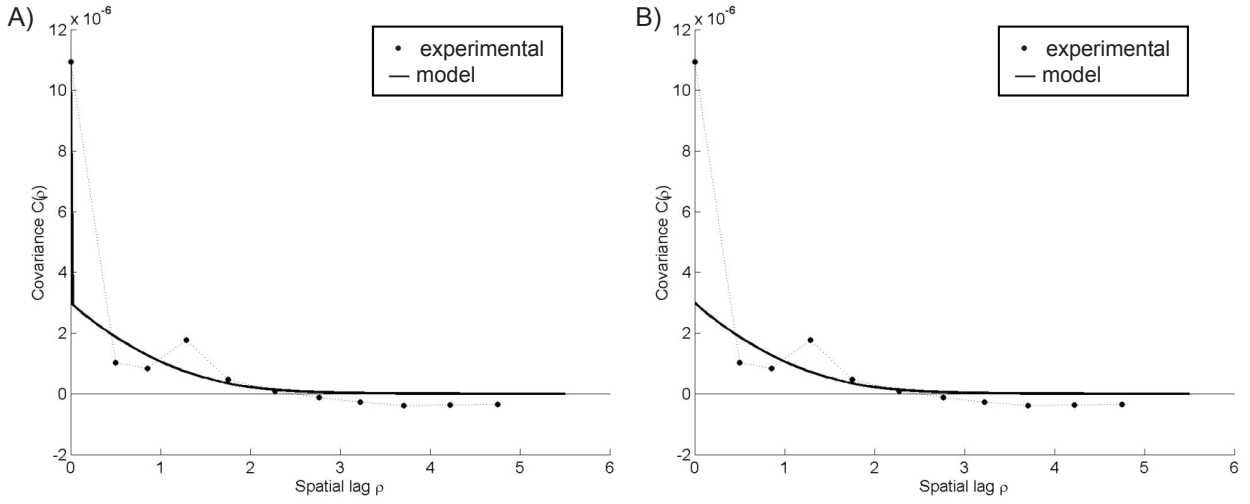

Figure 3 Plots of the experimental rate covariance with covariance models under the Poisson assumption. The nugget component of the A) observed rate field *R*(**s**)-covariance model is due to sampling variability error and removed to obtain the B) *X*(**s**)-covariance model.

model and keeping only the exponential and Gaussian components, such that

$$c_X(\rho) = c_2 \exp\left(\frac{-3\rho}{a_{\rho 1}}\right) + c_3 \exp\left(\frac{-3\rho^2}{a_{\rho 2}{}^2}\right) \quad (8)$$

where ρ is the spatial lag, c_2, and c_3 are constants, and $a_{\rho 1}$ and $a_{\rho 2}$ represent the spatial ranges of the exponential and Gaussian components, respectively.

The observed rates R_i, calculated from the cases simulated under the Poisson kriging assumption, ranged in value from 0 to 29.1 per 1,000 persons (Figure 4A). Using these observed rates and the corresponding covariance models (Figure 3), we obtained the Poisson kriging estimate $\hat{X}_i^{PK}$ (Figure 4B) and the UMBME estimate $\hat{X}_i^{UMBME}$ (Figure 4C). Spatially, $\hat{X}_i^{PK}$ and $\hat{X}_i^{UMBME}$ exhibited nearly identical spatial distributions of non-zero values. This result is due in part to the similar identification of neighbors and integration of spatial correlation between the models. However, when compared with the observed rate map, Poisson kriging exhibited greater smoothing strength than UMBME, particularly in areas of high observed rates, such as region (6,6) of the study domain. Similarly, the Poisson kriging MAD was lower than that of UMBME (Table 1), indicating greater smoothing towards the observed mean with Poisson kriging than UMBME. With regards to estimation accuracy, scatterplots of the observed rates R_i, the Poisson kriging estimates $\hat{X}_i^{PK}$, and the UMBME estimates $\hat{X}_i^{UMBME}$ versus the true latent rates X_i (Figures 4D, E, and 4F, respectively) show that the closer a point is to the 45 degree best fit line, the more accurately the observed or estimated value approximates the true $X(\boldsymbol{s})$-value. Artificially elevated rates fall above the best fit line, with the observed rate plot (Figure 4D) showing the greatest deviation from the true $X(\boldsymbol{s})$-value. Quantitatively, Poisson kriging exhibited the highest estimation accuracy in measures of both MSE and LCCC when compared with the true $X(\boldsymbol{s})$-value (Table 1).

Uniform assumption

Under the uniform assumption, the observed cases, *Y(**s**)*, were derived from the product of the latent rate field, $X(\boldsymbol{s})$, and sample population, *n(**s**)* (Figure 2B). We then obtained experimental and model covariance values for the corresponding observed rate field *R(**s**)* (Figure 5A) and latent rate field *X(**s**)* (Figure 5B). Similar to those of the Poisson assumption, the experimental covariance indicated a neighborhood of influence extending less than 2 spatial units from each point. However, the experimental covariance under the uniform assumption had a smaller nugget effect, indicating lower local variability [5]. In other words, observations that were close in space had more similar values under the uniform assumption than the Poisson assumption.

The observed rates R_i, calculated from the cases simulated under the uniform assumption, ranged in value from 0 to 7.5 per 1,000 persons (Figure 6A). Using these observed rates and the corresponding covariance models (Figure 5), we obtained the Poisson kriging estimates $\hat{X}_i^{PK}$ (Figure 6B) and the UMBME estimates $\hat{X}_i^{UMBME}$ (Figure 6C). Once again, the Poisson kriging and UMBME maps displayed nearly identical spatial distributions of non-zero values, while Poisson kriging exhibited greater smoothing strength than UMBME with

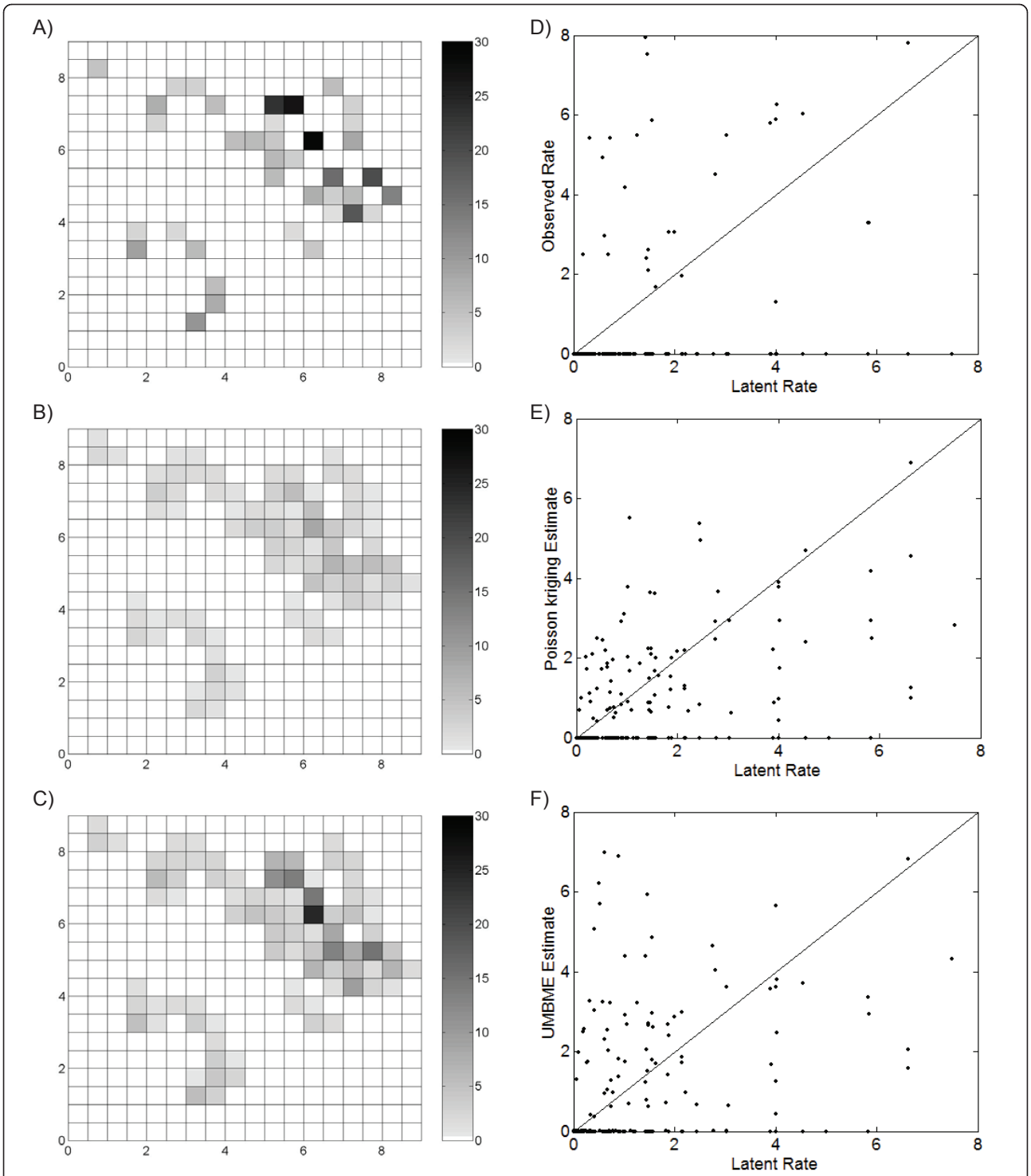

Figure 4 Maps and scatterplots of the observed, Poisson kriging estimated, and UMBME estimated rates under the Poisson assumption. Compared with the A) observed rate map, the B) Poisson kriging map displayed greater smoothing than the C) UMBME map, particularly in areas of high observed rates (cases per 1,000 persons). Meanwhile, scatterplots of the D) observed, E) Poisson kriging estimated, and F) UMBME estimated rates versus the true latent rate, $X(\mathbf{s})$, combined with MSE and LCCC calculations (Table 1), demonstrated that Poisson kriging produced the highest estimation accuracy under the Poisson assumption.

Table 1 Measures of model estimation accuracy, smoothing strength, and variance quality for the simulated Poisson assumption data, simulated uniform assumption data, and real HIV data.

Method	MSE	LCCC	MAD	*G*
Poisson Simulation				
Observed	1.00E-05	0.214		
Poisson kriging	**1.47E-06**	**0.550**	**0.938**	0.669
UMBME	4.83E-06	0.396	1.33	**0.798**
Uniform Simulation				
Observed	1.52E-06	0.618		
Poisson kriging	1.15E-06	0.555	**0.510**	0.670
UMBME	**6.43E-07**	**0.794**	0.701	**0.781**
HIV Data				
Observed	1.87E-04	0.013		
Poisson kriging	1.80E-04	0.039	**0.00297**	
UMBME	**1.79E-04**	**0.040**	0.00369	

The MSE and LCCC for the simulated data describe divergence between the model estimated and true latent rate values. For the HIV data, the MSE and LCCC values describe divergence between the model cross-validation results and observed rate values. Bolded values represent the model with the greatest estimation accuracy (MSE, LCCC), greatest smoothing strength (MAD), or best fit variance (*G*) in each dataset.

both less discernible high rate areas, such as region (6,5), and a lower MAD value (Table 1). In this case, however, the additional smoothing provided by Poisson kriging resulted in greater disparity with the true $X(\boldsymbol{s})$-value than UMBME, with a greater proportion of UMBME estimated points falling along the 45 degree best fit scatterplot line than those of Poisson kriging (Figures 6D, E, and 6F). Quantitatively, UMBME exhibited the highest estimation accuracy in measures of both MSE and LCCC (Table 1).

Mapping HIV infection in North Carolina

We applied Poisson kriging and UMBME to North Carolina HIV disease data to examine their performance in a real-world situation. We derived experimental and model covariance values for the observed rate field (Figure 7A) and corresponding $X(\boldsymbol{s})$ field (Figure 7B). The experimental covariance indicated a spatial neighborhood of influence extending less than 20 km from each location. While a nugget effect was present, without a point of comparison, it was difficult to ascertain whether the amount of spatial correlation was large or small for the given disease and geographical region. We then obtained maps of the HIV observed rate, the Poisson kriging estimates $\hat{X}_i^{PK}$, and the UMBME estimates $\hat{X}_i^{UMBME}$ (Figures 8A, B, and 8C, respectively). Similar to the simulated data analyses, Poisson kriging exhibited greater smoothing strength than UMBME with both less distinction between areas of high and low rates, such as in the northeast corner of the state, and a lower MAD value (Table 1). In cross-validation, UMBME performed only slightly better than Poisson kriging in MSE and LCCC values, with both models performing better than the observed data model with no smoothing (Table 1).

Model comparison

Under both the Poisson and uniform assumptions of the simulated data, Poisson kriging exhibited greater smoothing strength than UMBME. The two methods utilize the same values for a number of parameters that

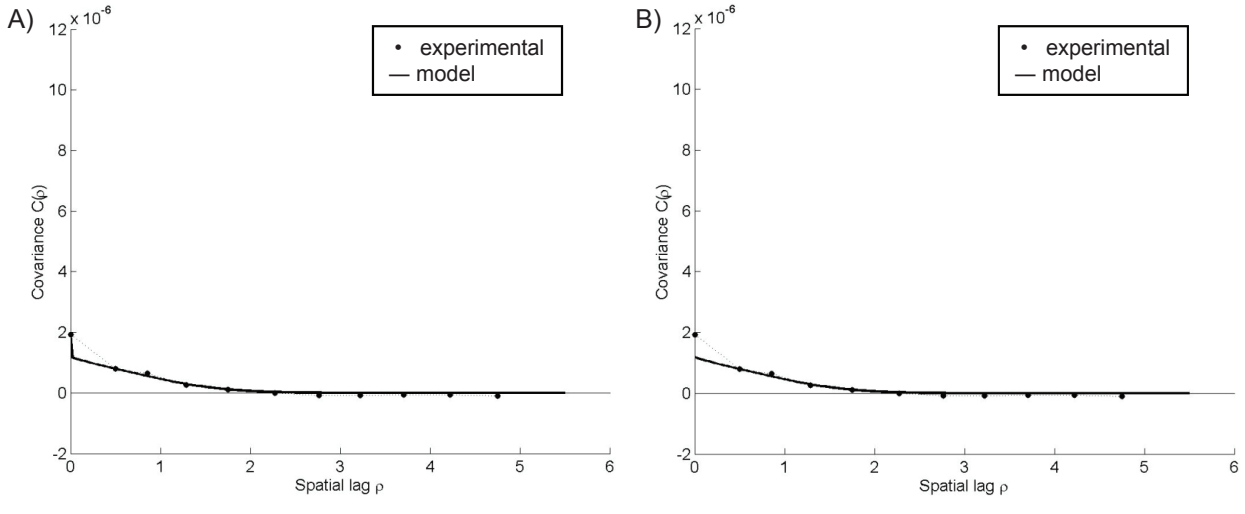

Figure 5 Plots of the experimental rate covariance with covariance models under the uniform assumption. The nugget component of the A) observed rate field $R(\mathbf{s})$-covariance model was removed to obtain the B) $X(\mathbf{s})$-covariance model.

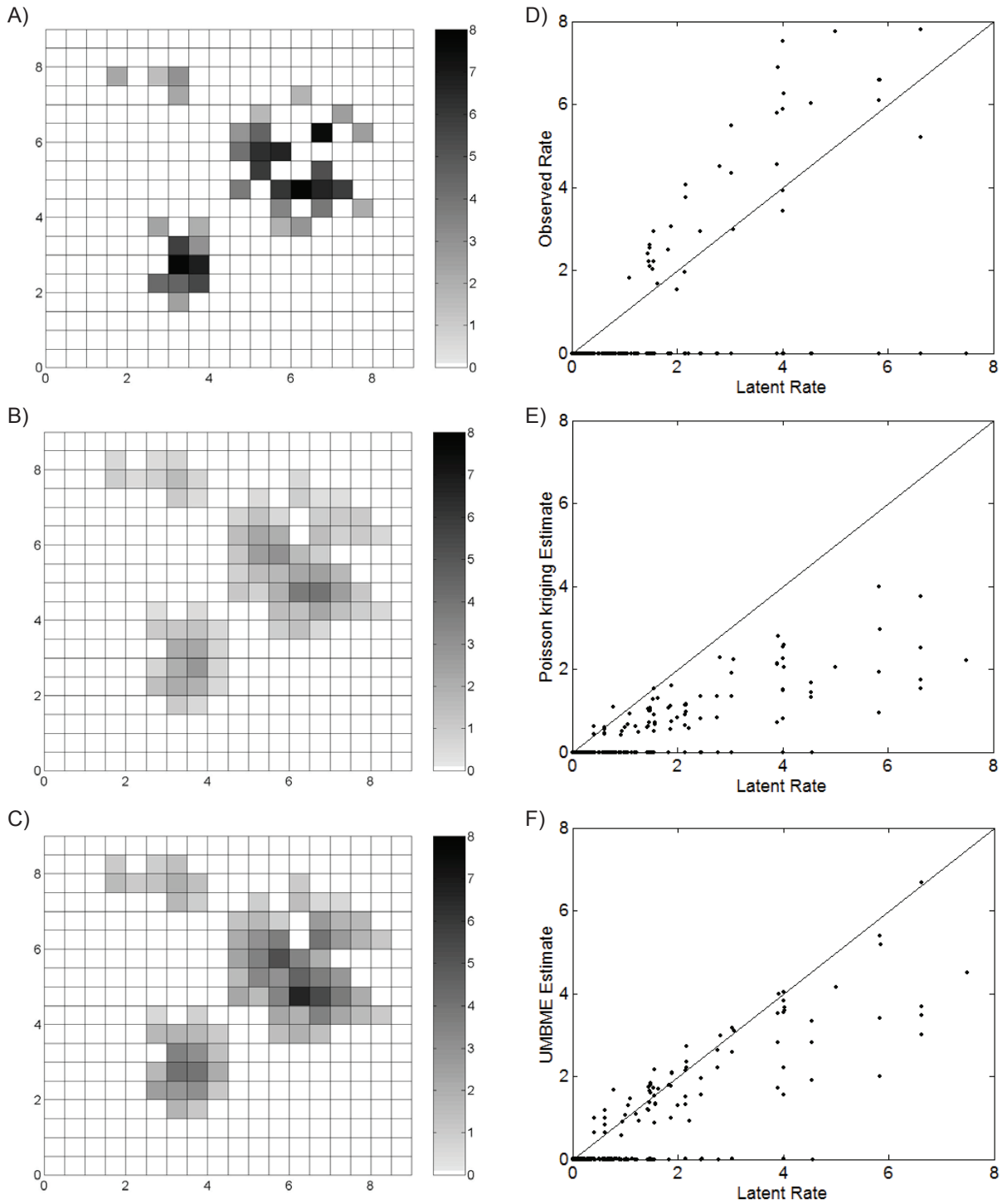

Figure 6 Maps and scatterplots of the observed, Poisson kriging estimated, and UMBME estimated rates under the uniform assumption. Compared with the A) observed rate map, the B) Poisson kriging map displayed greater smoothing than the C) UMBME map (cases per 1,000 persons). However, scatterplots of the D) observed, E) Poisson kriging estimated, and F) UMBME estimated rates versus the true latent rate, $X(\mathbf{s})$, combined with MSE and LCCC calculations (Table 1), demonstrated that UMBME produced the highest estimation accuracy under the uniform assumption.

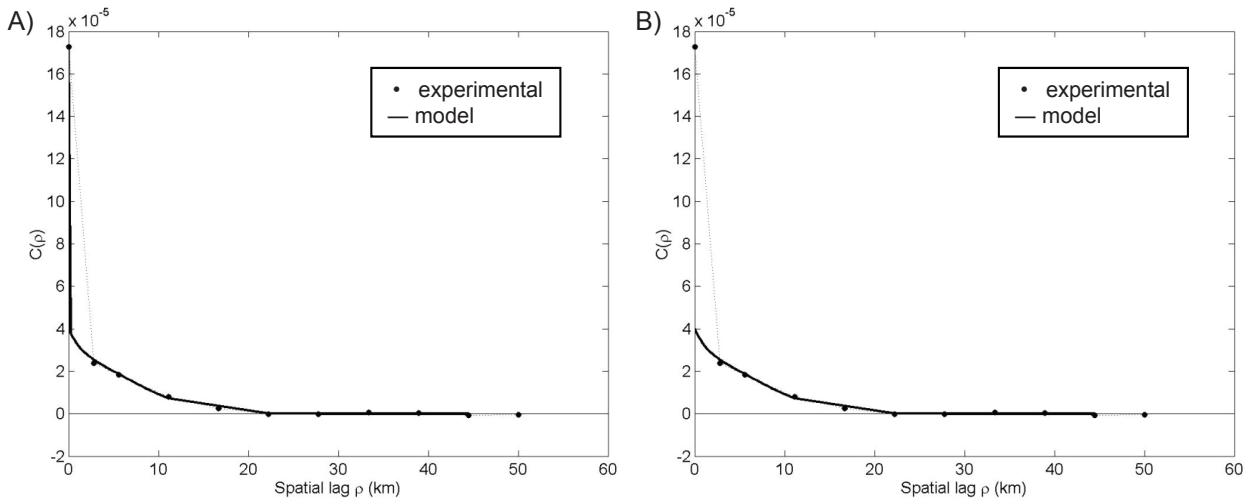

Figure 7 Plots of the experimental rate covariance with covariance models for the North Carolina HIV data. The experimental covariance indicated a spatial neighborhood of influence extending less than 20 kilometers from each location. The nugget component of the A) observed rate field $R(\mathbf{s})$-covariance model was removed to obtain the B) $X(\mathbf{s})$-covariance model.

affect smoothing strength, such as the maximum number of neighbors used in the estimation, the population size, and the covariance model [13]. The two methods differ, however, in their measurement error variances of the observed values and in their distributions of the posterior pdf. The error variance term is m^*/n_i for Poisson kriging, where m^* is the population-weighted mean of the observed rates and n_i is the population in area i, while the UMBME error term is distributed in an interval of size $1/n_i$. Furthermore, Poisson kriging yields a Gaussian posterior pdf characterized by the kriging estimate and variance, thereby allowing extreme values due to the tail of the Gaussian distribution [13]. UMBME, on the other hand, yields a non-Gaussian posterior pdf that is truncated by the interval soft data. Possible UMBME estimated values are therefore limited by the bounds of the truncated posterior pdf.

The distributional form of the posterior pdf may also contribute to differences in the estimated variance of each method. Under both the Poisson and uniform assumptions of the simulated data, the Poisson kriging and UMBME estimated variances were generally inversely proportional to the areal population size, while the magnitude of the Poisson kriging variance was consistently higher than that of UMBME. However, of greater interest is how well the estimated variance captures the true latent rate. Assuming normality of the prediction errors, we calculated each method's statistical coverage, or probability that the true latent rate in each area falls within the p-probability interval characterized by the estimated mean and variance. For example, for a p value of 0.95, the ideal model coverage probability would also equal 0.95. We then calculated Deutsch's "goodness" statistic [12,37] to examine divergence between the model-estimated and theoretical probability intervals, such that

$$G = 1 - \frac{1}{L}\sum_{l=1}^{L} w(p_l) \mid \varsigma(p_l) - p_l \mid \text{with } 0 \le G \le 1 \quad (9)$$

where L is the discretization level of the computation (i.e. L = 95 when p_l = 0.95), $\varsigma(p_l)$ is the model-estimated coverage for probability interval p_l, and $w(p_l) = 1$ if $\varsigma(p_l) > p_l$, and 2 otherwise. Double weight is given to deviations where the model-estimated coverage is less than expected [12]. The goodness of fit of the estimated variance increases as the value of G approaches one. With L = 100, UMBME outperformed Poisson kriging in G values for both the Poisson and uniform simulations (Table 1).

While a drawback of Poisson kriging's greater smoothing strength is that it becomes harder to distinguish areas of high or low estimated rates, in measures of estimation accuracy Poisson kriging performed better than UMBME under the Poisson assumption, UMBME performed better than Poisson kriging under the uniform assumption, and both models performed better than the observed rate with no smoothing. The Poisson assumption experimental covariance had a greater nugget effect than that of the uniform assumption, indicating less spatial correlation among the Poisson-derived observed rates than the uniform-derived rates, as shown in Figures 3 and 5. These results support those of previous simulation studies, which show that methods with

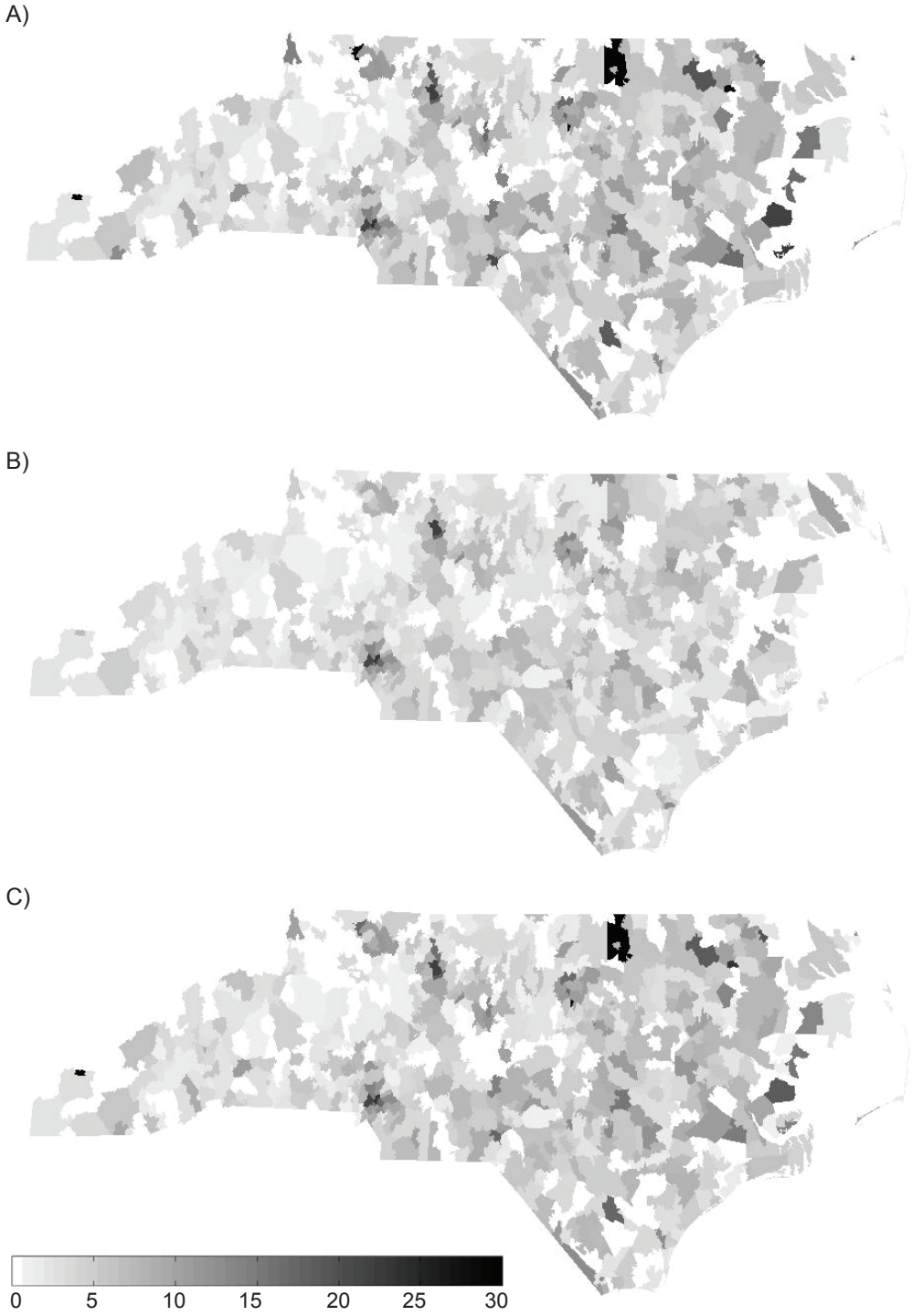

Figure 8 Maps of the North Carolina HIV A) observed, B) Poisson kriging estimated, and C) UMBME estimated rates. While the Poisson kriging and UMBME maps displayed nearly identical spatial distributions of non-zero values, Poisson kriging exhibited greater smoothing strength than UMBME, thereby providing less distinction between areas of high and low rates (cases per 1,000 tests).

greater smoothing strength are more accurate estimators when the spatial autocorrelation of the observed data is low, while methods that smooth less have greater estimation accuracy with data that is more spatially correlated [12].

In a real-world situation where the true latent rate of disease is unknown, the Poisson kriging and UMBME estimated HIV rate maps compared with the observed HIV rate map (Figure 8) show how the noise generated by artificially high rates due to sampling variability may be reduced through smoothing. Smoothing out sampling variability is particularly important in the study of HIV infection in North Carolina because it facilitates correction of the rates observed in rural areas with small testing populations by borrowing strength from rates observed in neighboring areas with larger testing populations. By exhibiting greater smoothing strength, the Poisson kriging map (Figure 8B) also eliminated much of the distinction between high and low rates needed to identify outbreaks or clusters of disease. The UMBME map (Figure 8C), on the other hand, provided a similar spatial representation of HIV that smoothed out extreme rates based on small populations while maintaining areas of high observed rates that may be useful in monitoring the HIV epidemic. For example, the number of areas with rates above 10 cases per 1,000 tests was reduced by 59% with Poisson kriging, but only 12% with UMBME as compared with the observed rates (Table 2). Both Poisson kriging and UMBME, however, produced nearly identical cross-validation MSE and LCCC values. Without knowing the true latent rate of infection, additional information, such as the distributional form of the dataset, is needed to assess model fit. Similarly, we could not calculate Deutsch's goodness statistic, *G*, for the real data to assess model uncertainty because the true latent rate of infection was unknown. However, as with the simulated data, the model-estimated variances (Figure 9A for UMBME) were generally inversely proportional to the underlying population size (Figure 9B).

Conclusions

Geostatistics modified to account for the specific nature of health data provides a rich class of methods able to incorporate and output more information about the disease field than simple smoothers while being computationally faster and easier to implement than full Bayesian hierarchical methods. This facilitates the investigation of both internal and external factors that affect model performance. As shown in this paper, both Poisson kriging and UMBME smoothing models more accurately predict the latent rate than the observed data with no smoothing. Meanwhile, Poisson kriging yielded smoother results than UMBME due to internal model assumptions.

Choosing the most appropriate smoothing method depends heavily on the characteristics of the disease being studied and the geographical space. As shown with the simulation data, accuracy of each estimation model was associated with the observed spatial correlation of the disease field. Methods that smoothed less performed better as the spatial correlation of the disease field increased. However, the observed spatial correlation, in turn, depends on the assumptions and characteristics of the latent and observed rate fields. For the latent rate, spatial characteristics of the disease, such as geographic risk factors and disease frequency, must be considered. For example, STIs tend to be concentrated in small, definable geographic areas where cases reside [19,21,27,29-31]. Therefore, maps of the residential locations of STI cases would typically exhibit greater spatial correlation than maps of the residential locations of a non-transmissible disease, such as leukemia. As a result, estimation methods with limited smoothing, such as UMBME, would be expected to produce more accurate predictions of the latent rate than methods that smooth more. However, the robustness of the observed data must also be considered. Factors that would increase or decrease the observed spatial correlation include the spatial support of the study area, such as the size and shape of geographical units, and the temporal resolution of the study. Observations measured over small time durations, such as monthly for STIs in North Carolina, tend to be spatially less correlated than observations identified over longer time periods, such as quarterly or annually.

Finally, in real world situations, the relative degree of spatial correlation in the observed data is difficult to ascertain when no point of reference exists. Further research is needed to compare the advantages and disadvantages of different smoothing models, examine model performance in different data environments, and examine model performance using different estimators. Additional research is also needed to examine whether fitting real data by a distributional form improves model performance and to identify other information about the disease field that would aid investigators in choosing the most appropriate smoothing model. Better disease map

Table 2 Number of ZIP codes with values above the 90th percentile of observed rates.

Method	No. ZIPs > 10 cases/1,000 tests	%Δ from Observed
Observed	74	–
UMBME	65	-12%
Poisson kriging	30	-59%

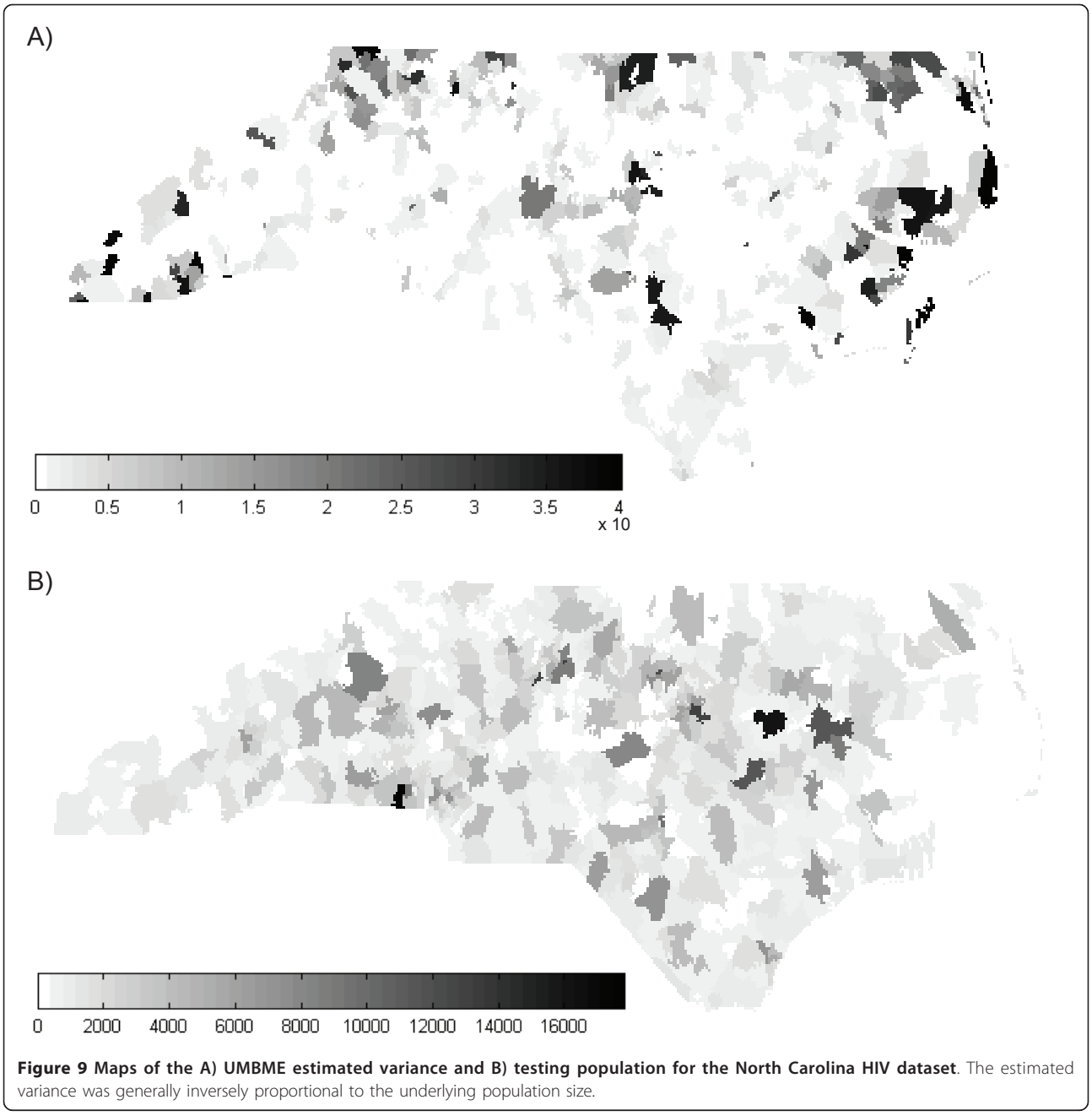

Figure 9 Maps of the A) UMBME estimated variance and B) testing population for the North Carolina HIV dataset. The estimated variance was generally inversely proportional to the underlying population size.

construction would improve the ability of public health officials to monitor spatial and temporal trends in disease rates, creating new opportunities for the definition of at-risk populations, identification of outbreaks, and allocation of resources toward areas and populations most affected by disease.

Supplementary Materials

APPENDIX A: The BME method

In the spatial application of BME, the distribution of a disease rate field is represented in terms of a Spatial Random Field (SRF), $X(\boldsymbol{s})$, where $\boldsymbol{s}$ is the spatial location. Each region i is defined in terms of a point spatial location, such as the latitude and longitude of the areal centroid, where $\boldsymbol{s_i} = (s_{1i}, s_{2i})$ and $i = 1,...I$. The spatial moments of $X(\boldsymbol{s})$ can be used to describe the behavior of the SRF. For example, the mean function

$$m_x(\boldsymbol{s}) = \overline{X(\boldsymbol{s})} \tag{A1}$$

(the overbar denotes stochastic expectation), characterizes trends and systematic structures in space, while

the covariance function

$$c_x(\rho) = \overline{[X(s) - \overline{X}(s)][X(s') - \overline{X}(s')]} \quad \text{(A2)}$$

expresses relevant correlations and dependencies between pairs of points in $X(\boldsymbol{s})$, where $\rho = |\boldsymbol{s}' - \boldsymbol{s}|$ denotes the spatial lag.

The BME approach categorizes all prior information known about the rate field into two major knowledge bases (KB): the general KB $\mathcal{G}$ and the specificatory (or site-specific) KB $\mathcal{S}$, where the total knowledge base $\mathcal{K} = \mathcal{G} \cup \mathcal{S}$. As discussed in Choi et al. (2003), the $\mathcal{G} - \text{KB}$ is considered 'general' in the sense that it characterizes global characteristics of the rate field, such as its mean trend, spatial moments, relevant epidemiologic laws or theories, and other assumptions about the behavior of the field that may apply. On the other hand, the $\mathcal{S} - \text{KB}$ refers to information that is 'specific' to each mapping situation, such as observed rate values measured at specific data points. In general, the vector of random variables $\boldsymbol{x}_{map}$ representing the rate field $X(\boldsymbol{s})$ consists of the vector of hard data random variables $\boldsymbol{x}_{hard}$ representing the latent disease rate at all locations where exact measurement values could be obtained, the vector of soft data random variables $\boldsymbol{x}_{soft}$ representing the latent rate at all locations with uncertain measured values (expressed in terms of confidence intervals or probabilistic functions), and the unknown latent rate value x_k to be estimated at some estimation point, such that $\boldsymbol{x}_{map} = (\boldsymbol{x}_{hard}, \boldsymbol{x}_{soft}, x_k)$.

Processing the information known about the rate field consists of three main stages (Gesink Law et al., 2006), as follows (Christakos and Li, 1998; Choi et al., 2003):

(i) *Structural (or prior) stage*: The general $\mathcal{G} - \text{KB}$ of the rate field is considered at all mapping points corresponding to $\boldsymbol{x}_{map}$. The structural probabilistic density function (pdf) of the random variables $\boldsymbol{x}_{map}, f_{\mathcal{G}}(\boldsymbol{\chi}_{map})$, where the vector of values $\boldsymbol{\chi}_{map}$ is a realization of the random variables $\boldsymbol{x}_{map}$, is derived by selecting the $f_{\mathcal{G}}(\boldsymbol{\chi}_{map})$ that maximizes entropy for the given general knowledge base. Using the Shannon measure of information we have

$$\begin{aligned} Info[\boldsymbol{x}_{map}] &= \log(1/f_X(\boldsymbol{x}_{map})) \\ &= -\log f_X(\boldsymbol{x}_{map}) \end{aligned}, \quad \text{(A3)}$$

and the entropy is defined in terms of the corresponding expected information, i.e.

$$\begin{aligned} H(\boldsymbol{x}_{map}) &= E[Info[\boldsymbol{x}_{map}]] \\ &= -\int_{-\infty}^{\infty} d\boldsymbol{\chi}_{map} f_G(\boldsymbol{\chi}_{map}) \log f_G(\boldsymbol{\chi}_{map}) \end{aligned} \quad \text{(A4)}$$

where $H(\boldsymbol{x}_{map})$ is the Entropy function, and E[.] is the stochastic expectation operator. Information is processed by selecting the $f_{\mathcal{G}}(\boldsymbol{\chi}_{map})$ that maximizes $H(\boldsymbol{x}_{map})$ for the given $\mathcal{G} - \text{KB}$.

(ii) *Specificatory (or meta-prior) stage*: The specificatory S-KB is identified and expressed in terms of hard and soft data, and the estimation point at which the field is to be estimated is defined. Here, we consider the specific framework in which all observed rates in the map are soft data $\boldsymbol{\chi}_{soft}$ with measures of uncertainty due to sampling variability. In processing the data, 6 may be rewritten as

$$\boldsymbol{\chi}_{soft} : \Pr\left[R(\boldsymbol{s}) - \frac{0.5}{N(\boldsymbol{s})} \le X(\boldsymbol{s}) \le R(\boldsymbol{s}) + \frac{0.5}{N(\boldsymbol{s})}\right] = 1 \quad \text{(A5)}$$

where Pr[.] is the probability operator, $R(\boldsymbol{s})$ is the observed rate, $N(\boldsymbol{s})$ is the population size, and $X(\boldsymbol{s})$ is the latent rate, and $\boldsymbol{s}$ represents each region i where these soft data are available. Furthermore, the aim is to provide a more accurate estimate of the latent rate field only at points $\boldsymbol{s}$ where soft data is available. In other words, a soft datum expressed by A5 is available for each estimation point.

(iii) *Integration stage*: The integration (posterior) pdf, $f_{\mathcal{K}}$, is derived by means of an operational Bayesian conditionalization rule that considers the total KB, $\mathcal{K} = \mathcal{G} \cup \mathcal{S}$, such that

$$f_{\mathcal{K}}(\chi_k) = A^{-1} \int_{-\infty}^{\infty} d\boldsymbol{\chi}_{soft} f_{\mathcal{S}}(\chi_k \boldsymbol{\chi}_{soft}) f_{\mathcal{G}}(\boldsymbol{\chi}_{map}) \quad \text{(A6)}$$

where $A = \int_{-\infty}^{\infty} d\chi_k \int_{-\infty}^{\infty} d\boldsymbol{\chi}_{soft} f_{\mathcal{S}}(\chi_k \boldsymbol{\chi}_{soft}) f_{\mathcal{G}}(\boldsymbol{\chi}_{map})$ is the normalization parameter, and $f_{\mathcal{S}}(\chi_k \boldsymbol{\chi}_{soft})$ is the multivariate uniform pdf defined by A5 jointly for the estimation point and the soft data points in its neighborhood.. The integration pdf provides a full stochastic assessment of the value x_k of the latent rate field at any estimation point, from which one may derive a variety of estimators. In this work, the expected value of the integration pdf was used as an appropriate estimator $\hat{X}_k$ of the latent rate, i.e.

$$\hat{X}_k = E[x_k] = \int_{-\infty}^{\infty} d\chi_k f_K(\chi_k), \quad \text{(A7)}$$

which we refer to as the BME mean estimator. Other estimators include the mode or the median of the integration pdf. Additionally, the variance of the integration pdf, or BME variance, provides a useful measure of estimation uncertainty.

List of Abbreviations

BME: Bayesian Maximum Entropy; HIV: Human Immunodeficiency Virus; LCCC: Lin's Concordance Correlation Coefficient; MAD: Mean Absolute Difference; MSE: Mean Square Error; Pdf: probability density function; STI: Sexually Transmitted Infection; UMBME: Uniform Model extension of Bayesian Maximum Entropy; VCT: Voluntary Counseling and Testing.

Acknowledgements

The authors thank Peter A. Leone, Del Williams, and the North Carolina Department of Health and Human Services, HIV/STD Prevention and Care Branch. This work was supported in part by the National Institute of Mental Health, R01-MH068686 (CDP) and by the National Institute of Allergies and Infectious Diseases, R01-AI067913 (WCM).

Author details

[1]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [2]Department of Environmental Sciences and Engineering, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [3]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada. [4]HIV/AIDS Division, San Francisco General Hospital, University of California-San Francisco, San Francisco, CA, USA. [5]Department of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.

Authors' contributions

KHH contributed to the study design, completed the analyses, and drafted the manuscript. MLS contributed to the study design, execution of the analyses, and interpretation of results. DCG, CDP, and WCM contributed to the study design and interpretation of results. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 24 May 2011 Accepted: 6 October 2011
Published: 6 October 2011

References

1. Best N, Richardson S, Thomson A: **A comparison of Bayesian spatial models for disease mapping.** *Statistical Methods in Medical Research* 2005, **14(1)**:35-59.
2. Leyland AH, Davies CA: **Empirical Bayes methods for disease mapping.** *Statistical Methods in Medical Research* 2005, **14(1)**:17-34.
3. Wakefield J, Elliott P: **Issues in the statistical analysis of small area health data.** *Statistics in Medicine* 1999, **18(17-18)**:2377-2399.
4. Wakefield JC, Best NG, Waller L: **Bayesian approaches to disease mapping.** In *Spatial Epidemiology: Methods and Applications*. Edited by: Elliott P, Wakefield J, Best N, Briggs D. Oxford: Oxford University Press; 2000:104-127.
5. Waller LA, Gotway CA: **Applied Spatial Statistics for Public Health Data.** Hoboken, NJ: John Wiley & Sons, Inc.; 2004.
6. Pascutto C, Wakefield JC, Best NG, Richardson S, Bernardinelli L, Staines A, Elliott P: **Statistical issues in the analysis of disease mapping data.** *Statistics in Medicine* 2000, **19(17-18)**:2493-2519.
7. Lawson AB: **Statistical Methods in Spatial Epidemiology.** Chichester: John Wiley & Sons Ltd.; 2001.
8. Liao H-H, Laymon P, Shull K: **Automated Process for Accessing Vital Health Information at Census Tract Level.** In *Geographic Information Systems in Public Health: Proceedings of the Third National Conference: 17-18 August 1998; San Diego*. Edited by: Williams RC, Howie MM, Lee CV, Henriques WD. Atlanta: Centers for Disease Control and Prevention; 2000:119-136, retrieved September 2005 from http://www.atsdr.cdc.gov/gis/conference98/index.html.
9. Bithell JF: **A classification of disease mapping methods.** *Statistics in Medicine* 2000, **19(17-18)**:2203-2215.
10. Clayton D, Kaldor J: **Empirical Bayes estimates of age-standardized relative risks for use in disease mapping.** *Biometrics* 1987, **43(3)**:671-681.
11. Devine OJ, Louis TA, Halloran ME: **Empirical Bayes methods for stabilizing incidence rates before mapping.** *Epidemiology* 1994, **5(6)**:622-630.
12. Goovaerts P: **Geostatistical analysis of disease data: estimation of cancer mortality risk from empirical frequencies using Poisson kriging.** *International Journal of Health Geographics* 2005, **4**:31.
13. Goovaerts P, Gebreab S: **How does Poisson kriging compare to the popular BYM model for mapping disease risks?** *International Journal of Health Geographics* 2008, **7**:6.
14. Ali M, Goovaerts P, Nazia N, Haq MZ, Yunus M, Emch M: **Application of Poisson kriging to the mapping of cholera and dysentery incidence in an endemic area of Bangladesh.** *International Journal of Health Geographics* 2006, **5**:45.
15. Christakos G, Li X: **Bayesian Maximum Entropy Analysis and Mapping: A Farewell to Kriging Estimators?** *Mathematical Geology* 1998, **30(4)**:435-462.
16. Serre ML, Christakos G: **Modern geostatistics: computational BME analysis in the light of uncertain physical knowledge - the Equus Beds study.** *Stochastic Environmental Research and Risk Assessment* 1999, **13(1-2)**:1-26.
17. Christakos G, Bogaert P, Serre ML: **Temporal GIS: Advanced Functions for Field-Based Applications.** New York: Springer-Verlag; 2002.
18. Choi K-M, Serre ML, Christakos G: **Efficient mapping of California mortality fields at different spatial scales.** *Journal of Exposure Analysis and Environmental Epidemiology* 2003, **13(2)**:120-133.
19. Law DCG, Serre ML, Christakos G, Leone PA, Miller WC: **Spatial analysis and mapping of sexually transmitted diseases to optimise intervention and prevention strategies.** *Sexually Transmitted Infections* 2004, **80(4)**:294-299.
20. Christakos G, Olea R, Serre ML, Yu H-L, Wang L: **Interdisciplinary Public Health Reasoning and Epidemic Modelling: The Case of Black Death.** New York: Springer-Verlag; 2005.
21. Gesink Law DC, Bernstein KT, Serre ML, Schumacher CM, Leone PA, Zenilman JM, Miller WC, Rompalo AM: **Modeling a syphilis outbreak through space and time using the Bayesian Maximum Entropy approach.** *Annals of Epidemiology* 2006, **16(11)**:797-804.
22. Lee S-J, Yeatts KB, Serre ML: **A Bayesian Maximum Entropy approach to address the change of support problem in the spatial analysis of childhood asthma prevalence across North Carolina.** *Spatial and Spatio-temporal Epidemiology* 2009, **1(1)**:49-60.
23. Monestiez P, Dubroca L, Bonnin E, Durbec J-P, Guinet C: **Geostatistical modelling of spatial distribution of *Balaenoptera physalus* in the Northwestern Mediterranean Sea from sparse count data and heterogeneous observation efforts.** *Ecological Modelling* 2006, **193(3-4)**:615-628.
24. Rothman KJ, Greenland S: **Measures of Disease Frequency.** In *Modern Epidemiology*.. 2 edition. Edited by: Rothman KJ, Greenland S. Philadelphia: Lippincott-Raven; 1998:29-46.
25. MathWorks, Inc: **MaTLab, the language of technical computing.** *using MATLAB version 6.1* Natick, MA: The MathWorks, Inc; 2001.
26. **The Bayesian Maximum Entropy software for Space/Time Geostatistics and temporal GIS data integration.** [http://www.unc.edu/depts/case/BMELIB/].
27. Gesink DC, Sullivan AB, Miller WC, Bernstein KT: **Sexually transmitted disease core theory: roles of person, place, and time.** *American Journal of Epidemiology* 2011, **174(1)**:81-89.
28. Lin LI: **A concordance correlation coefficient to evaluate reproducibility.** *Biometrics* 1989, **45(1)**:255-268.
29. Rothenberg RB: **The geography of gonorrhea. Empirical demonstration of core group transmission.** *American Journal of Epidemiology* 1983, **117(6)**:688-694.
30. Becker KM, Glass GE, Brathwaite W, Zenilman JM: **Geographic epidemiology of gonorrhea in Baltimore, Maryland, using a geographic information system.** *American Journal of Epidemiology* 1998, **147(7)**:709-716.
31. Zenilman JM, Ellish N, Fresia A, Glass GE: **The geography of sexual partnerships in Baltimore: applications of core theory dynamics using a geographic information system.** *Sexually Transmitted Diseases* 1999, **26(2)**:75-81.
32. North Carolina Department of Health & Human Services, Division of Public Health: **North Carolina Epidemiologic Profile for HIV/STD Prevention & Care Planning, July 2005.** Raleigh, NC; 2005.
33. Geographic Data Technology, ESRI: **U.S. ZIP Code Points.** *ESRI Data & Maps 2004* Redlands, CA: ESRI; 2004.
34. Stern HS, Cressie N: **Posterior predictive model checks for disease mapping models.** *Statistics in Medicine* 2000, **19(17-18)**:2377-2397.
35. Marshall EC, Spiegelhalter DJ: **Approximate cross-validatory predictive checks in disease mapping models.** *Statistics in Medicine* 2003, **22(10)**:1649-1660.
36. Yu H-L, Yang S-J, Yen H-J, Christakos G: **A spatio-temporal climate-based model of early dengue fever warning in southern Taiwan.** *Stochastic Environmental Research and Risk Assessment* 2011, **25(4)**:485-494.

37. Deutsch CV: **Direct assessment of local accuracy and precision.** In *Geostatistics Wollongong '96. Volume 1*. Edited by: Baafi EY, Schofield NA. Dordrecht, The Netherlands, Kluwer Academic Publishers; 1997:115-125.

doi:10.1186/1476-072X-10-54
Cite this article as: Hampton *et al.*: Adjusting for sampling variability in sparse data: geostatistical approaches to disease mapping. *International Journal of Health Geographics* 2011 **10**:54.