

Bayesian Areal Wombling for Geographical Boundary Analysis

Haolan Lu, Bradley P. Carlin

Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN

*In the analysis of spatially referenced data, interest often focuses not on prediction of the spatially indexed variable itself, but on boundary analysis, that is, the determination of boundaries on the map that separate areas of higher and lower values. Existing boundary analysis methods are sometimes generically referred to as wombling, after a foundational article by Womble (1951). When data are available at point level (e.g., exact latitude and longitude of disease cases), such boundaries are most naturally obtained by locating the points of steepest ascent or descent on the fitted spatial surface (Banerjee, Gelfand, and Sirmans 2003). In this article, we propose related methods for areal data (i.e., data which consist only of sums or averages over geopolitical regions). Such methods are valuable in determining boundaries for data sets that, perhaps due to confidentiality concerns, are available only in ecological (aggregated) format, or are only collected this way (e.g., delivery of health-care or cost information). After a brief review of existing algorithmic techniques (including that implemented in the commercial software *BoundarySeer*), we propose a fully model-based framework for areal wombling, using Bayesian hierarchical models with posterior summaries computed using Markov chain Monte Carlo methods. We explore the suitability of various existing hierarchical and spatial software packages (notably *S-plus* and *WinBUGS*) to the task, and show the approach's superiority over existing nonstochastic alternatives, both in terms of utility and average mean square error behavior. We also illustrate our methods (as well as the solution of advanced modeling issues such as simultaneous inference) using colorectal cancer late detection data collected at the county level in the state of Minnesota.*

Background

The analysis and modeling of spatially referenced data sets have occupied statisticians and geographers for decades. Typically, such data consist of variables that

The work of the second author was supported in part by NIH grant 5-R01-ES07750-07, and by NIH grant 1-R01-CA95955-01.

Correspondence: Dr Bradley P. Carlin, Division of Biostatistics, School of Public Health, University of Minnesota, Mayo Mail Code 303, Minneapolis, MN 55455-0392
e-mail: brad@biostat.umn.edu

Submitted: March 12, 2004. Revised version accepted: October 4, 2004.

have been observed at different spatial locations, and one seeks models that capture possible spatial associations among them. Depending upon the nature of spatial referencing, spatial data are classified as either *point-referenced* (often called *geo-statistical*) or *areal* (often called *lattice*). In the former, the spatial locations are points with known coordinates, such as latitude–longitude or easting–northing pairs. For the latter, locations are usually geographic regions (such as counties, census tracts, or ZIP codes) along with information on the neighborhood structure (contiguity) of these regions.

Statistical models for spatial data depend upon the nature of the spatial referencing. The role of spatial models for point-referenced data has historically been the making of predictions and spatial interpolations over the entire domain, while models for areal data have been used primarily for smoothing raw rate maps to reveal broad spatial trends. Recently, however, spatial analysts have shown a growing interest in detecting zones or *boundaries* that reveal sharp changes in the values of spatially oriented variables. For example, in contour maps of spatial surfaces, regions with highly compact contour lines indicate zones of abrupt change; significant changes in gradients are observed as one cuts across these contours. Similarly, for areal data, the boundary separating two regions with drastically different measurements or fitted values is a boundary of abrupt change.

The general problem of identifying zones of abrupt change is known as *wombling*, after a foundational article by Womble (1951) that discusses the scientific importance of this problem. Since then, wombling has become a popular technique among geneticists, demographers, linguists, ecologists, environmental scientists, and many others in analyzing spatial relationships. Notable articles in this area include Barbujani, Jacquez, and Ligi (1990), Barbujani, Oden, and Sokal (1989), Oden et al. (1993), Bocquet-Appel and Bacro (1994), Fortin (1994, 1997), and Fortin and Drapeau (1995), with Jacquez, Maruca, and Fortin (2000) offering an excellent recent review. This literature deals primarily with point-referenced data (regularly or irregularly spaced), often investigating suitable tessellations of the spatial domain and the appropriate interpolators for approximating the surface.

The field we broadly refer to as “wombling” is also known as *barrier analysis* or *edge detection* in fields such as landscape topography, systematic biology, sociology, ecology, and public health. In all of these fields, the research goal is to identify regional differences across shared boundaries, in order to identify homogenous regions or discover important “barriers.” Ultimately, the underlying influences responsible for these barriers are typically of greatest scientific interest. For instance, the genetic study by Sokal and Thompson (1998) locates barriers (areas) over which genetic flow (population movement, through changing allele frequencies) is reduced or stopped.

As with spatial statistical models, wombling approaches must depend upon whether the data are point-referenced or areal. A significant amount of both methodological and applied research on wombling with point- (or raster-) level data exists. For example, Bocquet-Appel and Jakobi (1996) used point wombling

analysis to identify the barriers to the spatial diffusion for the demographic transition in western Europe. Barbujani, Oden, and Sokal (1989) detect a zone of main discontinuity using an algorithm operating on a local spatial variance statistic.

In studies of human populations, patient confidentiality laws often restrict available data to counts or rates over geopolitical regions. As such, in this article we restrict our attention to the areal data case. Areal wombling (also known as *polygonal* wombling) is not as well-developed in the literature as point or raster wombling, but some notable articles exist. Oden et al. (1993) provide a wombling algorithm for multivariate categorical data defined on a lattice. The statistic chosen is the average proportion of category mismatches at each pair of neighboring sites, with significance relative to an independence or particular spatial null distribution judged by a randomization test. Csillag et al. (2001) developed a procedure for characterizing the strength of boundaries examined at neighborhood level. In this method, a topological or a metric distance δ defines a neighborhood of the candidate set of polygons (say, p_i). A weighted local statistic is attached to each p_i . The difference statistic calculated as the squared difference between any two sets of polygons' local statistic and its quantile measure are used as a relative measure of the distinctiveness of the boundary at the scale of neighborhood size δ . Jacquez and Greiling (2003) estimate boundaries of rapid change for colorectal, lung, and breast cancer incidence in Nassau, Suffolk, and Queens counties in New York.

Several criteria for determining what constitutes a "barrier" have been suggested in previous boundary analysis research. Most researchers select a *dissimilarity metric* (Euclidean distance, squared Euclidean distance, Manhattan distance, Steinhaus statistics, etc.) to measure the difference in response between the values at (say) adjacent polygon centroids. An absolute (dissimilarity metrics greater than C) or relative (dissimilarity metrics in the top $k\%$) threshold then determines which borders are considered actual barriers, or parts of the boundary.

The relative (top $k\%$) thresholding method for determining boundary elements (BEs) is easily criticized, because for a given threshold, a fixed number of BEs will always be found regardless of whether or not the responses separated by the boundary are statistically different. Jacquez and Maruca (1998) suggest use of both local and global statistics to determine where statistically significant BEs are, and a randomization test (with or without spatial constraints) for whether the boundaries for the entire surface are statistically unusual.

All of the aforementioned areal wombling approaches (like traditional, point-level wombling itself) appear to be algorithmic, rather than model-based. That is, they do not involve a probability distribution for the data, and therefore permit statements about the "significance" of a detected boundary only relative to predetermined, often unrealistic null distributions. To remedy this, in this article we develop a hierarchical statistical modeling framework to perform areal wombling. Considering the underlying map and its geographical boundaries as our domain, our Bayesian statistical approach permits direct estimation of the probability that two geographic regions are separated by the wombled boundary. Our models

account for spatial association and permit the borrowing of strength across different levels of the model hierarchy, accurately estimating quantities that would be inestimable using classical approaches.

The remainder of our article is organized as follows. The next section provides a brief account of existing methods of crisp areal wombling (where each geographical border is classified as either part or not part of the boundary), as well as our hierarchical Bayesian alternative. The subsequent section introduces a model-based approach to fuzzy areal wombling (where now partial membership of a particular border in the boundary is permitted), illustrating the variety of practical benefits that accrue. The penultimate section uses simulation to demonstrate the Bayesian method's improved performance over the current, more ad hoc methods. Finally, the last section discusses our findings and suggests directions for further research.

Crisp areal wombling

Review of existing methods and software

On an areal map, geopolitical borders are usually *closed*, that is, they completely partition the study area. They are completely determined from the definition of the experimental domain and are usually summarized as an adjacency matrix A , where $a_{ij} = 1$ or 0 depending upon whether regions i and j are adjacent to each other or not. Another popular representation is to treat the map as a graph with the nodes representing the regions and edges connecting neighboring nodes.

However, in areal wombling we are interested only in *difference boundaries*, which are those borders that separate two adjacent counties having dramatically different observed response values. Difference boundaries may, in fact, not fully partition the map, but only indicate regions that greatly differ with respect to the outcome variable.

As difference boundaries are determined by a response variable that we assume has a probability distribution, difference boundaries need to be *estimated*. To be precise, suppose we have regions $i = 1, \dots, N$ along with the areal adjacency matrix A , and we have observed a response Y_i (e.g., a disease count or rate) for the i th region. Existing areal wombling algorithms assign a *boundary likelihood value* (BLV) to each areal boundary using a Euclidean distance metric between neighboring observations. This distance is taken as the dissimilarity metric, calculated for each pair of adjacent regions. Thus, if i and j are neighbors, the BLV associated with the edge (i, j) is

$$D_{ij} = ||Y_i - Y_j||$$

where $|| \cdot ||$ is a distance metric. Locations with higher BLVs are more likely to be a part of a difference boundary, as the variable changes rapidly there.

The wombling literature and attendant software further distinguishes between *crisp* and *fuzzy* wombling. In the former, BLV's exceeding specified thresholds are

assigned a *boundary membership value* (BMV) of 1, so the wombled boundary is $\{(i, j) : \|Y_i - Y_j\| > c, i \text{ adjacent to } j\}$ for some $c > 0$. The resulting edges are called *boundary elements*. In the latter (fuzzy) case, BMVs can range between 0 and 1 (say, $\|Y_i - Y_j\| / \max_{ij}\{\|Y_i - Y_j\|\}$) and indicate *partial* membership in the boundary. (We note that the use of the word “fuzzy” here does not refer to the notion of fuzzy logic, but only to the fact that an edge’s membership or nonmembership in the boundary is no longer absolute; see the section on “Fuzzy areal wombling”).

Example: Minnesota colorectal cancer late detection data. As an illustration, we consider boundary analysis for a data set recording the rate of late detection of several cancers collected by the Minnesota Cancer Surveillance System (MCSS), a population-based cancer registry maintained by the Minnesota Department of Health. The MCSS collects information on geographic location and stage at detection for colorectal, prostate, lung, and breast cancers. For each county, the late detection rate is defined as the number of regional or distant case detections divided by the total cases observed, for the years 1995–1997. As higher late detection rates are indicative of possibly poorer cancer control in a county, a wombled boundary for this map might help identify barriers separating counties with different cancer control methods. Such a boundary might also motivate more careful study of the counties that separate it, in order to identify previously unknown covariates (population characteristics, dominant employment type, etc.) that explain the difference.

In this example, we consider n_i , the total number of colorectal cancers occurring in county i , and Y_i , the number of these that were detected late. To correct for the differing number of detections (i.e., the total population) in each county, we womble not on the Y scale, but on the *standardized late detection ratio* (SLDR) scale. This is the ratio of observed to expected late detections,

$$\text{SLDR}_i = \frac{Y_i}{E_i}, \quad i = 1, \dots, N$$

where the expected counts are computed via internal standardization as $E_i = n_i \bar{r}$, where $\bar{r} = \sum_i Y_i / \sum_i n_i$, the statewide late detection rate.

We imported the SLDR_i values into *BoundarySeer*, a commercial wombling software package, and computed $D_{ij} = |\text{SLDR}_i - \text{SLDR}_j|$ for all adjacent counties i and j . Fig. 1 shows the resulting crisp wombling boundaries, using both the upper 20% (thin black lines) and the upper 50% (thick white lines) for visual comparison. County borders not part of the wombled boundary are shown as thin light gray lines. The SLDR values themselves are indicated by the shading of the counties, as in an ordinary grayscale choropleth map.

Fig. 1 shows many attractive features of the traditional approach to wombling. Boundaries are easily seen, and a rough picture of the SLDRs themselves is not sacrificed. However, statistically speaking, the results leave something to be desired. First, the extremeness in the SLDRs in the figure is likely due not to huge

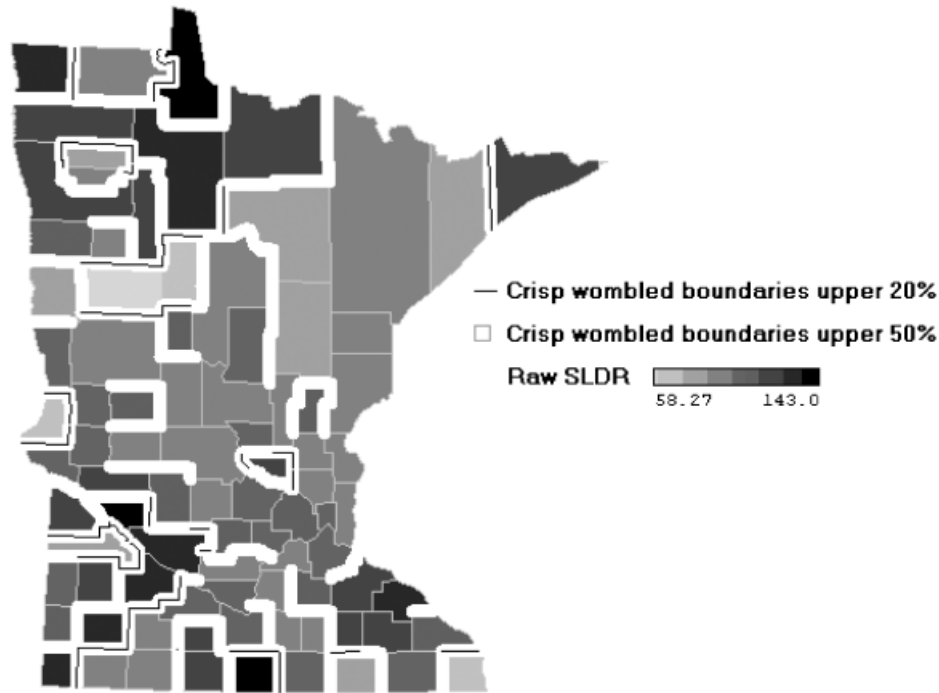


Figure 1. BoundarySeer map of colorectal cancer late detection standardized late detection ratio (SLDR) and crisp wombled (top 20% and 50%) boundaries: differences based directly on raw SLDR values.

differences in the true underlying county-level risks, but only to random variation in the observed data for thinly populated counties. Thus, as already mentioned, we still require a statistical model for the data in order to properly account for all sources of uncertainty (both spatial and nonspatial) and enable formal statistical inference. Second, we need an approach that will smooth the observed data (to reflect spatial similarity, as well as our differing degrees of confidence in the observed detection rates for urban and rural areas) while simultaneously determining the boundaries. (Note that a two-step approach applying a standard wombling algorithm to spatially smoothed rates would offer only a partial solution to this second goal, as the smoothing and boundary analysis tasks cannot properly be thought of as independent.)

Hierarchical modeling approach

We now present a statistical modeling framework to rectify the previous subsection’s omission of both variability and spatial correlation in the response variable Y . In particular, consider again data having both observed and expected counts (Y_i, E_i) for the i th county. Several authors (e.g., Hodges, Carlin, and Fan 2003) have investigated areal data models for normal (Gaussian) data, but such models provide little computational simplification here, and in any case are inappropriate for the

count data most commonly encountered in areal settings. As such, we instead employ the Poisson log-linear form

$$Y_i \sim \text{Poisson}(\mu_i) \quad \text{where } \log \mu_i = \log E_i + \mathbf{x}_i' \boldsymbol{\beta} + \phi_i. \quad (1)$$

This model allows a vector of region-specific covariates \mathbf{x}_i (if available), and a random effect vector $\boldsymbol{\phi} = (\phi_1, \dots, \phi_N)'$ that is given a conditionally autoregressive (CAR) specification (Besag 1974). A common form of this distribution (often called the *intrinsic* CAR, or IAR model) has improper joint distribution, but intuitive conditional distributions of the form

$$\phi_i | \boldsymbol{\phi}_{j \neq i} \sim N(\bar{\phi}_i, 1/(\tau m_i)) \quad (2)$$

where N denotes the normal distribution, $\bar{\phi}_i$ is the average of the $\phi_{j \neq i}$ that are adjacent to ϕ_i , and m_i is the number of these adjacencies; this distribution is usually abbreviated as CAR(τ). Finally, τ is typically set equal to some fixed value, or assigned a distribution itself (usually a relatively vague γ distribution).

Markov chain Monte Carlo (MCMC) samples $\mu_i^{(g)}$, $g = 1, \dots, G$ from the marginal posterior distribution $p(\mu_i | \mathbf{y})$ can be obtained for each i (see, e.g., Banerjee, Carlin, and Gelfand 2004, Sec. 4.3), from which corresponding samples of the (theoretical) SLDR,

$$\eta_i = \frac{\mu_i}{E_i}, \quad i = 1, \dots, N$$

are immediately obtained. We may then define the BLV for boundary (i, j) as

$$\Delta_{ij} = |\eta_i - \eta_j| \quad \text{for all } i \text{ adjacent to } j \quad (3)$$

Crisp and fuzzy wombling boundaries are then based upon the posterior distribution of the BLVs. In the crisp case, we might define ij to be part of the boundary if and only if $E(\Delta_{ij} | \mathbf{y}) > c$ for some constant $c > 0$, or if and only if $P(\Delta_{ij} \geq c | \mathbf{y}) > c^*$ for some constant $0 < c^* < 1$.

Model (1)–(2) can be easily implemented in the WinBUGS software package, freely available from www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml. WinBUGS includes an easy-to-follow manual and many worked examples; in addition, www.statslab.cam.ac.uk/~krice/winbugsthemovie.html provides a helpful online Flash tutorial. WinBUGS has no limit on N , the number of areal units. The case of *multivariate* response variables would require multivariate CAR models (see the concluding section); such models are not yet implemented in WinBUGS, but could likely be added using the new WinBUGS Development Interface (WBDev); see the WinBUGS homepage for details.

Posterior draws $\{\eta_i^{(g)}, g = 1, \dots, G\}$ and their sample means $\hat{E}(\eta_i | \mathbf{y}) = \frac{1}{G} \sum_{g=1}^G \eta_i^{(g)}$ are easily obtained for our problem by mimicking the approach used in the Scottish lip cancer example (click on “Map,” pull down to “Manual,” and then click on “Examples”), and may then be exported for future use. Bayesian wombling would naturally obtain posterior draws of the BLVs in (3) by simple transformation

as $\Delta_{ij}^{(g)} = |\eta_i^{(g)} - \eta_j^{(g)}|$, and then base the boundaries on their empirical distribution. For instance, we might estimate the posterior means as

$$\hat{E}(\Delta_{ij}|\mathbf{y}) = \frac{1}{G} \sum_{g=1}^G \Delta_{ij}^{(g)} = \frac{1}{G} \sum_{g=1}^G |\eta_i^{(g)} - \eta_j^{(g)}| \quad (4)$$

and take as our wombled boundaries the borders corresponding to the top 20% or 50% of these values. A choropleth map of the $\hat{E}(\eta_i|\mathbf{y})$ then completes a display similar to Fig. 1.

Unfortunately, BoundarySeer is not designed to permit this sort of analysis, as it does not expect repeated samples of Δ_{ij} values to be available. However, we can “trick” the program into producing what we want by exploiting a feature intended for use with multivariate data. Specifically, suppose we have K random variables Z_{ki} , $k = 1, \dots, K$ observed at each location i . BoundarySeer defines an omnibus BLV comparing *all* the variables observed at locations i and j as

$$D_{ij} = \sum_{k=1}^K |Z_{ki} - Z_{kj}| \quad (5)$$

BoundarySeer refers to this D_{ij} as *Manhattan distance*, apparently due to its similarity (when $K=2$) to distance as measured by cab drivers in New York City (city block distance). If we instead use k to index a collection of K Gibbs samples from $p(\boldsymbol{\eta}|\mathbf{y})$, then if we set $Z_{ki} = \eta_i^{(k)}$, equation (5) delivers exactly $K \times \hat{E}(\Delta_{ij}|\mathbf{y})$. As K is merely a fixed constant, the wombled boundaries will be unaffected by its inclusion. The one real limitation here might be that the program currently caps K at 2000, but this should be large enough to deliver sufficient Monte Carlo accuracy in $\hat{E}(\Delta_{ij}|\mathbf{y})$, provided the sampled $\eta_i^{(g)}$ values are not too highly autocorrelated.

Fig. 2 uses thick white lines to show the wombled boundaries arising from the top 20% of the $\hat{E}(\Delta_{ij}|\mathbf{y})$ values based on this trick using $K = G = 1000$ Gibbs samples, superimposed on the fitted colorectal cancer detection true SLDR(η_i) map. Note there are several differences between the boundaries in this map and the raw data-based boundaries in Fig. 1. As a side comment, we also experimented with basing our wombled boundaries only on a subsample of every M Gibbs samples, where M is chosen large enough so that the subsampled chain is nearly uncorrelated. In our case, this meant setting $M=5$; the resulting crisp boundaries (not shown) differ from those shown in Fig. 2 by only two segments.

Hierarchical model selection

Given our analytic goal is the detection of differences between adjacent regions (i.e., steep gradients in the surface), use of a spatial smoothing model like (2) may seem odd. It is certainly true that CAR models imply a fairly high degree of smoothing, which may indeed smooth over some apparent boundaries in the raw SLDR surface. However, the use of raw, unsmoothed measurements may identify bound-

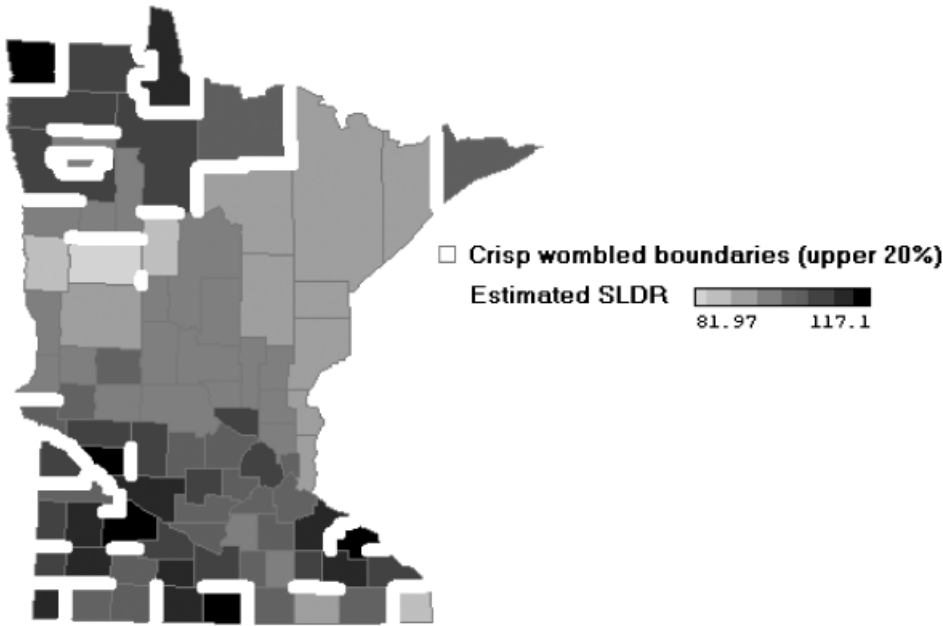


Figure 2. Fully Bayesian wombled boundaries based on $G = 1000$ Gibbs samples, Boundary Seer.

aries that are artificial (say, arising from extreme measurements in two adjacent, thinly populated regions) rather than real.

As mentioned above, the improper CAR model (2) is often referred to as the *intrinsic* CAR, or IAR. Proper versions of the CAR model, as recommended by Cressie (1993) and others, allow a smaller and, to some extent, data-determined degree of spatial smoothing. As such, in Table 1 we compare the fit of five models to our colorectal cancer detection data. Three of these have a single set of random effects ϕ_i : our original IAR, a proper CAR that induces somewhat less spatial smoothing, and a nonspatial model that assumes the ϕ_i to be independent and identically distributed (i.i.d.) Gaussian white noise random variables. The

Table 1 Goodness-of-Fit \bar{D} , Effective Model Size p_D , and Overall DIC Score for Five Competing Random Effects Models, Minnesota Colorectal Cancer Detection Data

Model	\bar{D}	p_D	DIC
Nonspatial	515.3	50.71	566.0
IAR	511.6	34.63	546.2
Proper CAR	511.6	34.58	546.2
IAR+heterogeneity	520.4	60.82	581.2
Lawson and Clark	518.6	57.83	576.4

DIC, deviance information criterion; CAR, conditionally autoregressive; IAR, intrinsic CAR.

remaining two models have more than one set of random effects. The “IAR plus heterogeneity” model, popularized by Besag et al. (1991), replaces the log relative risk in (1) with

$$\log \mu_i = \log E_i + \mathbf{x}_i' \boldsymbol{\beta} + \theta_i + \phi_i$$

where the ϕ_i are distributed CAR and the θ_i are i.i.d. normal. Finally, Lawson and Clark (2002) propose the mixture model

$$\log \mu_i = \log E_i + \mathbf{x}_i' \boldsymbol{\beta} + \theta_i + p_i \phi_i + (1 - p_i) \xi_i \quad (6)$$

where the θ_i are again i.i.d. normal, the ϕ_i are IAR, the ξ_i are also IAR but under an L_1 (median-based) distance metric, and the p_i are weights that trade off between the spatially smoother ϕ_i and the rougher ξ_i , and are such that $0 \leq p_i \leq 1$. While Lawson and Clark (2002) suggest a symmetric β prior for the p_i , we simply set $p_i = 1/2$ for all i , because even with rather informative prior distributions, our data could not identify the four sets of random effects implied by (6) well enough for our MCMC algorithm to converge.

While there is disagreement among Bayesians on how best to choose among competing hierarchical models, the usual penalized likelihood approaches (notably, Akaike Information Criterion [AIC] and Bayesian Information Criterion [BIC]) do not seem sensible here as the theory that supports them does not extend to the random effects setting. Indeed, it is not even clear how “big” any of these models are, as their effective size depends on the amount of shrinkage in the random effects toward each other, and this will be different for every data set. For instance, the ϕ_i in (2) could contribute anywhere from 0 to N effective parameters. To settle the issue, Spiegelhalter et al. (2002) recommend use of a *deviance information criterion* (DIC) consisting of the sum of a deviance score \bar{D} and an effective model size p_D . Smaller \bar{D} values correspond to models that provide better fit to the data, while smaller p_D correspond to more parsimonious models. As there will often be a trade-off here (with larger models offering better fit but less parsimony), models with smaller overall DIC scores are to be preferred.

The `car.normal`, `car.proper`, and `car.L1` distributions in WinBUGS permit us to fit the five models to the Minnesota colorectal cancer detection data. The \bar{D} , p_D , and DIC scores in Table 1 were obtained from three independent sampling chains of 10,000 iterations each, following a 1000-iteration burn-in period. The results indicate that the proper CAR and our original IAR model offer the best combination of fit and parsimony. The proper CAR gravitates toward a fit nearly as spatially strong as the IAR: it shares 93% of the IAR’s crisp (upper 20%) wombled boundaries. The nonspatial model results in a higher effective parameter count, but not one that improves model fit, and is thus unacceptable. The two multiple random effects models similarly do not deliver better fit for their even higher effective parameter burdens. As the IAR remains slightly more popular in practice than the proper CAR, in what follows we consider results only for the one-random effect IAR model.

Comparison of hierarchical and traditional areal wombling

Previously, we noted that `BoundarySeer` is designed to compute the BLV as the Manhattan (or other) distance between two region-specific estimates. Taking the posterior means as these estimates, this leads to basing our wombled boundaries on

$$\Delta_{ij}^* = |\hat{E}(\eta_i|\mathbf{y}) - \hat{E}(\eta_j|\mathbf{y})| \quad (7)$$

instead of $\hat{\Delta}_{ij} \equiv \hat{E}(\Delta_{ij}|\mathbf{y})$ as given in (4). Hierarchical modeling theory would suggest that boundaries based on $\hat{\Delta}_{ij}$ (the posterior mean of the absolute difference) should be superior to those based on Δ_{ij}^* (the absolute difference of the posterior means), as the former will properly account for uncertainty in the η_i values throughout the process, rather than averaging this uncertainty out before the BLV is computed. While not a proof of this conjecture, Fig. 3 does indicate that these two sets of boundaries can be very different: the figure shows histograms of the Δ_{ij}^* and $\hat{\Delta}_{ij}$ values for every potential county boundary segment in the Minnesota colorectal cancer detection analysis. The latter have a distribution that is far less skewed and stochastically larger.

As the absolute value function is convex, from Jensen's inequality we have

$$\hat{\Delta}_{ij} = E(|\eta_i - \eta_j||\mathbf{y}) \geq |E(\eta_i - \eta_j)|\mathbf{y}| = |E(\eta_i|\mathbf{y}) - E(\eta_j|\mathbf{y})| = \Delta_{ij}^*$$

and thus the stochastic ordering holds for every segment ij . However, Fig. 3 cannot reveal whether there is reordering among the potential BLVs that will lead to

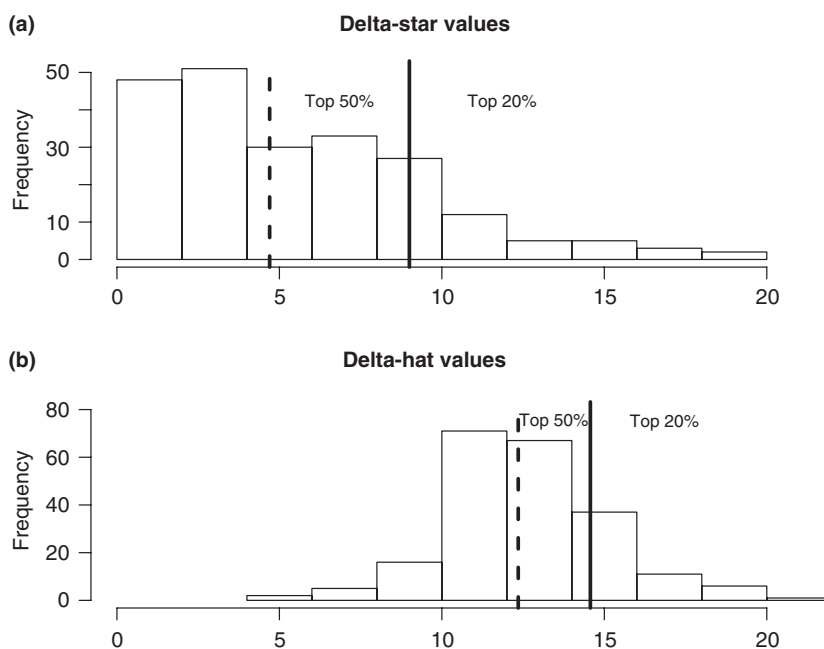


Figure 3. Histograms of (a) Δ_{ij}^* and (b) $\hat{\Delta}_{ij}$ values, Minnesota colorectal cancer detection data, with upper 50% (dashed) and 20% (solid) cutoffs as indicated.

Table 2 Comparison of the Upper 20% of the Empirical Distributions of Δ_{ij}^* and $\hat{\Delta}_{ij}$

	$\hat{\Delta}_{ij}$ not in the upper 20%	$\hat{\Delta}_{ij}$ in the upper 20%
Δ_{ij}^* not in the upper 20%	150	23
Δ_{ij}^* in the upper 20%	23	20

substantially different wombled boundaries. To check this, Table 2 compares the pairs of Δ_{ij}^* and $\hat{\Delta}_{ij}$ values to see how they match with respect to the upper 20% of their empirical distributions (the cutoffs for which are marked with solid vertical lines in Fig. 3). This table reveals that substantial differences do exist that could lead to different wombled boundaries. We defer more formal comparison of Δ_{ij}^* and $\hat{\Delta}_{ij}$ to the section on “Simulation study.”

Fuzzy areal wombling

Traditional approach

Crisp wombled boundaries are easy to comprehend, but their uncertainty is difficult to assess, as they do not indicate how close to the threshold the mean BLV is, nor how concentrated its distribution is. Maps of the BLV posterior standard errors can help somewhat, as can animated sequences of crisp boundaries (see, e.g., www.biostat.umn.edu/~haolanl/movie.gif). Still, fuzzy wombled boundaries may offer a better alternative here, as they do not resort to binary BMVs. In this subsection, we explore the traditional approach to fuzzy wombling, and again provide a hierarchical Bayesian version of the procedure.

As mentioned above, fuzzy wombling permits BMVs other than 0 or 1, so that some areas are allowed to be more important in determining the boundary than others. Fig. 4 shows how the user may draw this distinction through the choice of two BLV cutoffs, m_t and m_c . Specifically, a linear increase in BMV is used between these two cutoffs, with $\text{BMV}(\text{BLV} = m_t) = 0$ and $\text{BMV}(\text{BLV} = m_c) = 1$. Locations with BLVs below m_t are excluded from the boundary, while locations with BLVs above m_c are the “core” boundary ($\text{BMV} = 1$). The m_t and m_c values may be based on percentile choices; for example, one can choose m_c such that 20% of the BLVs belong to the core.

To illustrate, Fig. 5 gives a set of fuzzy wombling boundaries for the Minnesota colorectal cancer detection data drawn in *S-plus*. Here we have preset m_c so that the core boundary contains 20% of all $\hat{E}(\Delta_{ij}|\mathbf{y})$ values, and m_t so that the fuzzy boundary contains the remaining 80% of these values (i.e., every candidate BE is at least part of the fuzzy boundary). The boundary segments in Fig. 5 are grayscaled so that the darker the boundary is, the more likely it is above the threshold and thus the more important it is in determining the boundary.

Hierarchical modeling approach

Traditional fuzzy wombling is appealing in that it avoids a dogmatic 0–1 decision about whether or not to include a particular map segment in the boundary. The

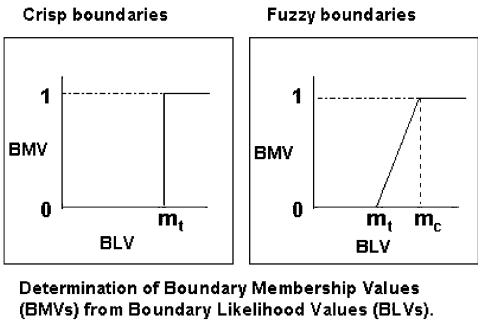


Figure 4. Illustration of the difference between crisp and fuzzy wombling.

resulting gradations of grayscale in Fig. 5 provide some idea as to how certain we are that each segment should be included. However, while a fuzzy BMV is between 0 and 1, it may *not* be interpreted as a “probability of being part of the boundary,” as no stochastic model is associated with the traditional fuzzy algorithm. Yet assigning a degree of confidence to each segment is entirely natural, and may well be the interpretation mistakenly adopted by naive viewers of such plots.

Fortunately, here again the hierarchical Bayesian approach offers a direct and convenient solution. Suppose we select a cutoff c such that, were we *certain* a

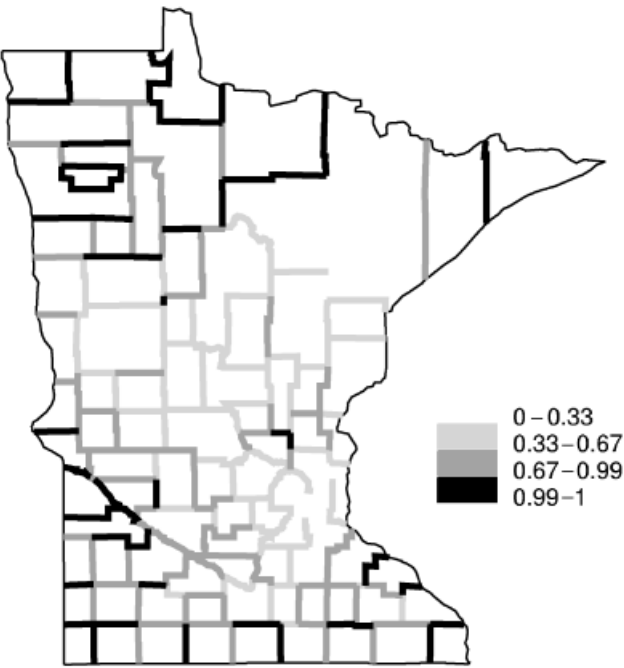


Figure 5. S-plus fuzzy wombling boundaries with grayscale boundary membership values, Minnesota colorectal cancer detection data.

particular BLV exceeded c , we would also be certain the corresponding segment was part of the boundary. As our statistical model (1)–(2) delivers the full posterior distribution of every Δ_{ij} , we can compute $P(\Delta_{ij} > c | \mathbf{y})$, and take this probability as our fuzzy BMV for segment ij .

In fact, the availability of the posterior distribution provides another benefit: a way to directly assess the *uncertainty* in our fuzzy BMVs. Our Monte Carlo estimate of $P(\Delta_{ij} > c | \mathbf{y})$ is

$$\hat{p}_{ij} \equiv \hat{P}(\Delta_{ij} > c | \mathbf{y}) = \frac{\#\Delta_{ij}^{(g)} > c}{G} \quad (8)$$

This is nothing but a binomial proportion, where its components independent, basic binomial theory implies an approximate standard error for it would be

$$\widehat{\text{se}}(\hat{p}_{ij}) = \sqrt{\frac{\hat{p}_{ij}(1 - \hat{p}_{ij})}{G}} \quad (9)$$

Of course, our Gibbs samples Δ_{ij} are *not* independent in general, as they arise from a Markov chain, but we can make them approximately so simply by subsampling as mentioned above, retaining only every M th sample. Note that this subsampling does *not* remove the *spatial* dependence among the Δ_{ij} , so repeated use of formula (9) would not be appropriate if we wanted to make a *joint* probability statement involving more than one of the Δ_{ij} at the same time; we return to this “simultaneous inference” issue below.

Fig. 6 gives posterior probability areal wombling maps for the Minnesota colorectal cancer detection data using three illustrative values of c (5, 15, and 30), a subsampling interval of $M = 5$, and $G = 2000$. The first row gives boundaries based on the \hat{p}_{ij} from equation (8). These three panels show how many and what sort of boundaries are produced by insisting that the absolute difference in fitted SLDR between adjacent regions exceed some particular cutoff (5%, 15%, or 30%) with some probability (indicated by the degree of shading). The second row gives corresponding standard error estimates from equation (9). As expected, the probability of each segment being a member of the boundary decreases as c (the threshold for being a BE) increases. The \hat{p} maps suggest little evidence of strong boundaries between counties; there are only a few county boundaries estimated to separate regions with true SLDRs that differ by more than 15%. Still, county 63 (Red Lake, a T-shaped county in the northwest part of the state) does seem “isolated” from its two neighbors; again, see a further discussion of this (simultaneous inference) issue below. The standard error plots reveal that the overall uncertainty associated with each segment tends to decrease for the more extreme c (5 and 30), as we become more certain that most segments either are or are not part of the boundary.

Fig. 7 provides histograms of the fuzzy posterior probability wombled BLVs and associated standard error estimates given in Fig. 6. The decreasing means are now

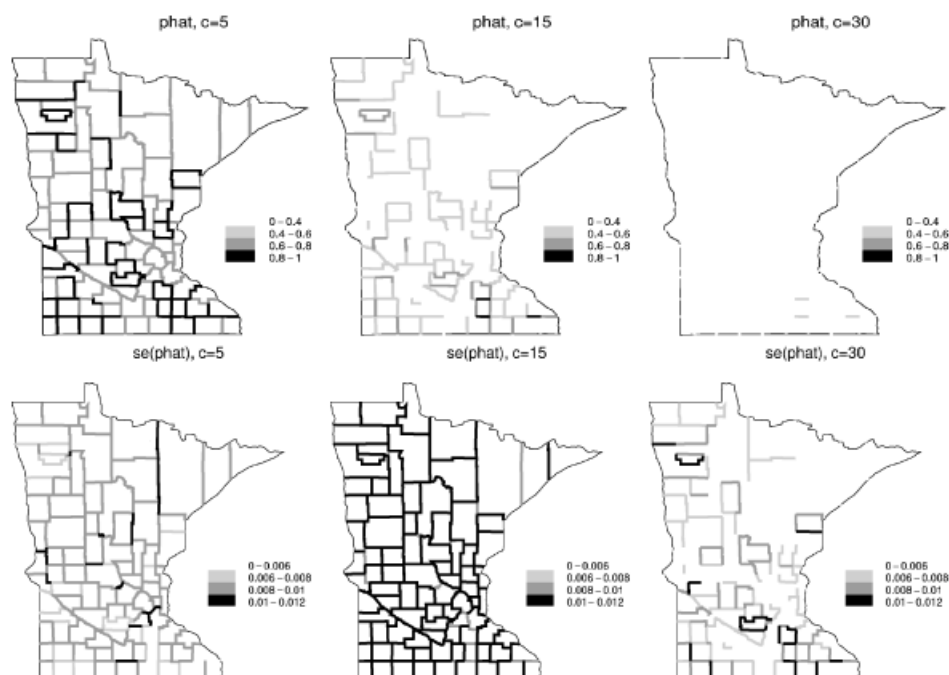


Figure 6. Posterior probability areal wombling for Minnesota colorectal cancer detection data. First row, wombled boundaries using \hat{p}_{ij} ; second row, associated estimated standard errors $\hat{se}(\hat{p}_{ij})$.

much easier to see. Note also that $\sqrt{0.25/G} \approx 0.012$ forms an upper bound on the variance estimates as the formula in (9) has this value as its maximum (obtained when $\hat{p}_{ij} = 0.5$).

An important related point concerns simultaneous inference for a particular collection of the Δ_{ij} . Consider again the case of county 63 (Red Lake), which has counties 57 and 60 as its only neighbors. We may assess whether this county truly is “isolated” from the others by evaluating

$$\tilde{p}_{63} \equiv P(\Delta_{63,57} > c \cap \Delta_{63,60} > c | \mathbf{y}).$$

As our $\{\Delta_{ij}^{(g)}\}$ samples come from the *joint* posterior of $\Delta \equiv \{\Delta_{ij}\}$, Monte Carlo estimates analogous to (8) and (9) are immediately available. The Bayesian approach thus allows simultaneous inference without reference to the troublesome frequentist notion of multiple comparisons (which would in turn require Bonferroni or similar corrections). We could easily repeat this calculation for all \tilde{p}_i , $i = 1, \dots, 87$, perhaps standardizing by the number of segments in each case for a fairer comparison. Alternatively, we might attempt to consider all pairs of adjacent segments on the map, or any predetermined subdivision of the map into pieces

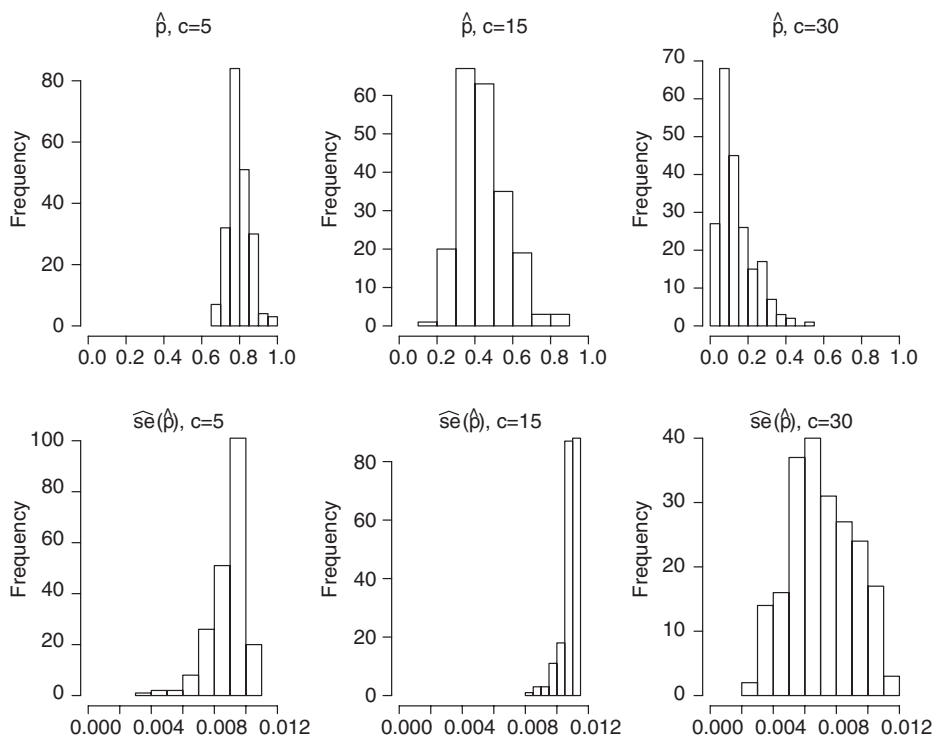


Figure 7. Histograms of posterior wombling probabilities and associated estimated standard errors, Minnesota colorectal cancer detection data. First row, wombled values for \hat{p}_{ij} ; second row, associated estimated standard errors $\widehat{se}(\hat{p}_{ij})$.

suggested by context (say, a boundary around the Twin Cities metro area, within which colorectal screening might be hypothesized to be better or more accessible).

We close this section by remarking that wombling on the spatial residuals ϕ_i instead of the fitted SLDRs η_i changes not only the scale of the c cutoff (to difference in *log*-relative risk), but the interpretation of the results as well. Specifically, we can borrow an interpretation often mentioned (but rarely actually used) in spatial epidemiology regarding spatially oriented covariates \mathbf{x}_i still missing from model (1). Because boundaries based on the ϕ_i separate regions that differ in their unmodeled spatial heterogeneity, a careful comparison of such regions identified by the wombled map should prove the most fruitful in any missing covariate search. Indeed this approach should offer a more direct solution than the usual one of searching for patterns in fitted choropleth disease maps, as it provides a measure of the differences between regions, rather than the regional levels themselves. On the other hand, if no significant boundaries exist in the wombled residual map, this is evidence of (covariate-adjusted) mapwide equity, known to public policy makers as “environmental justice” when the response variable measures an environmental exposure.

Simulation study

In this section, we conduct a small simulation study to compare the performance of Δ_{ij}^* , as given in (7), and $\hat{\Delta}_{ij}$, as given in (4). Recall the former is the absolute difference of the SLDR posterior means, while the latter is the posterior mean of the absolute difference in the SLDRs. The $\hat{\Delta}_{ij}$ are more justifiable from a theoretical point of view, but we might prefer the Δ_{ij}^* if they perform acceptably, because as we have seen they are computable in *BoundarySeer*.

Our study proceeds as follows. At iteration k , we first draw a “true” value $\phi^{(k)} \equiv \{\phi_i^{(k)}\}$ from an assumed CAR model, followed by a simulated data vector $\mathbf{Y}^{(k)} \equiv \{Y_i^{(k)}\}$ from our Poisson model (1). We then compute the posterior distribution of the $\eta_i^{(k)}$ for each i , hence the posterior summaries $\Delta_{ij}^{*(k)}$ and $\hat{\Delta}_{ij}^{(k)}$. Averaging over the $R = 216$ boundary segments in the Minnesota county map, the mean square errors (MSEs) of these two estimators relative to the “truth” are

$$\text{mse}^{*(k)} = \frac{1}{R} \sum_{i,j} (\Delta_{ij}^{*(k)} - \Delta_{ij}^{(k)})^2 \quad \text{and} \quad \widehat{\text{mse}}^{(k)} = \frac{1}{R} \sum_{i,j} (\hat{\Delta}_{ij}^{(k)} - \Delta_{ij}^{(k)})^2$$

Finally, averaging these scores over K simulated data sets $\mathbf{Y}^{(k)}$ generated in this way, we would obtain the overall scores $\overline{\text{mse}}^*$ and $\overline{\widehat{\text{mse}}}$, with smaller numbers of course indicating better performance.

Our simulations use the same Minnesota county map and adjacency structure as above, as well as the expected counts E_i from the Minnesota colorectal cancer detection data set. For convenience we set $\beta = 0$ in (1) and $\tau = 1$ in (2). However, the usual CAR prior is improper: note that if the same constant is added to every ϕ_i , ϕ_i increases by this same amount, and the probability specification in (2) is unchanged. This location invariance means it is not immediately obvious how to obtain a single draw from the CAR prior. One solution here is to again use *WinBUGS*: We simply run the same code as used for computing the posterior, but after deleting the Poisson likelihood. *WinBUGS* adds a “sum-to-zero” constraint ($\sum_{i=1}^N \phi_i = 0$) by recentering the ϕ_i after every iteration of the Gibbs sampler, resolving the impropriety. Iterating over the full conditional distributions in (2), we obtain the desired CAR draw at “convergence” (say, 100 iterations) of this preliminary sampler.

We considered MSE performance over $K = 100$ simulated data sets. Obtaining the K posterior summaries requires repeated use of the *WinBUGS* language. Such iteration is possible using within the R programming environment (www.r-project.org); see www.stat.columbia.edu/~gelman/bugsR or www.biostat.umn.edu/~brad/software/BRugs/BRugs_install.html for details. We obtained the $\text{mse}^{*(k)}$ and $\widehat{\text{mse}}^{(k)}$ values plotted versus k in Fig. 8. The fully Bayesian $\widehat{\text{mse}}^{(k)}$ values are smaller in all 100 cases. Averaging over the $K = 100$ simulated data sets we obtain $\overline{\text{mse}}^* = 307.70$ and $\overline{\widehat{\text{mse}}} = 255.66$. Thus once again, $\hat{\Delta}$ emerges as the better performer, motivating use of the fully Bayesian approach.

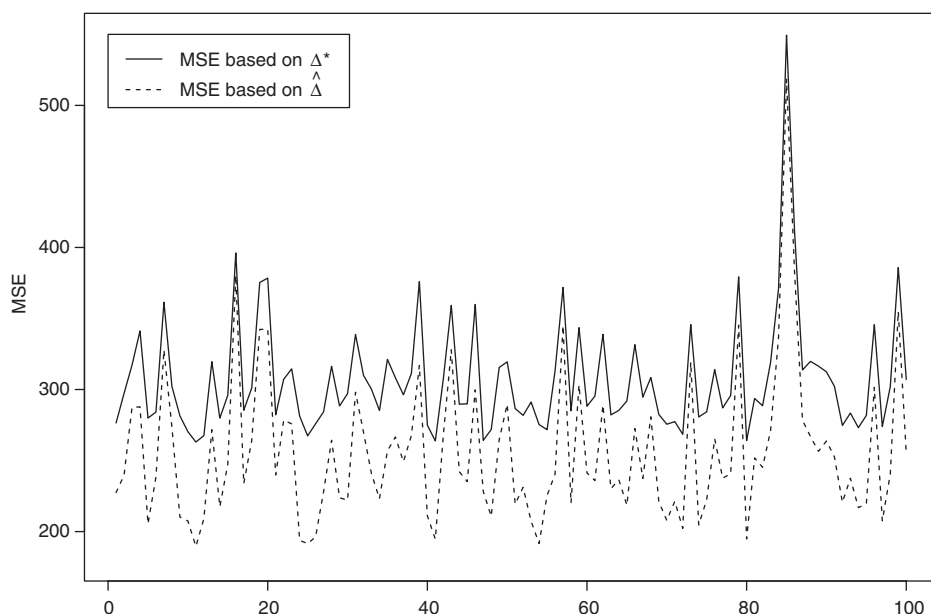


Figure 8. Mean square errors (MSEs) comparison between Δ^* and $\hat{\Delta}$ over 100 simulated data sets.

Discussion and future work

In this article, we have presented a model-based approach to areal wombling, using Bayesian hierarchical models and implemented via Markov chain Monte Carlo computing methods. The approach enables direct probability statements regarding the likelihood that a particular geographic border is a boundary, and also correctly accounts for both spatial and nonspatial uncertainty in the data.

The problem of aggregation bias has been much discussed in the geographical and spatial statistical literature, and we hasten to add that our Bayesian methods can fall victim to the same “ecological fallacy” issues that would plague any aggregated data analysis. A complete discussion of this issue is beyond the scope of this article, but we refer the reader to Banerjee, Carlin, and Gelfand (2004, p. 175) for more discussion of this problem in the context of hierarchical spatial models.

While in this article we have focused on just a few software packages (BoundarySeer, S-plus, and WinBUGS), they are by no means the only ones that might be used in our context. For example, on the mapping side, the public domain C/C++ program *shapelib* can be used to extract polygons from ArcView shape files for map regions. On the Bayes/MCMC side, our hierarchical models could be implemented in C/C++, R/S-plus, or Matlab. This last package also includes a “Spatial Econometrics Toolbox” that permits both mapping and MCMC estimation in a single software environment. In our work, we accomplish this same goal via an R/S-plus routine (see www.biostat.umn.edu/~yuecui/) to import boundaries from ArcView into WinBUGS. We prefer this approach as WinBUGS

permits not only mapping but also automatic generation of the spatial adjacency matrix.

A strong advantage of our approach is that it uses only standard algorithms and software, freely available and familiar to working spatial statisticians. However, future work looks to expanding beyond our current model class in order to allow even more flexibility. For example, suppose we seek boundaries based on observations of $K > 1$ variables over each region. That is, we now have Y_{ki} where k indexes the variable (or “working map”). We might assume these Y_{ki} were still independent Poisson random variables as in (1), but where now the $\phi_i = (\phi_{1i}, \dots, \phi_{Ki})'$ are assigned a *multivariate CAR* distribution, generalizing (2); see Gelfand and Vounatsou (2003), or Banerjee, Carlin, and Gelfand (2004, Sec. 7.4). Wombling may now proceed much as before, but with a model that captures correlation both across space and among the K variables.

Returning to the univariate setting, note that CAR model (2) can be written as

$$\phi_i | \phi_{(-i)} \sim N \left(\frac{\sum_j w_{ij} \phi_j}{\sum_j w_{ij}}, \frac{1}{\tau \sum_j w_{ij}} \right) \quad (10)$$

where the *weights* w_{ij} are equal to 1 if $i \neq j$ and regions i and j are adjacent, and 0 otherwise. Model (10) remains a valid distributional specification provided $0 \leq w_{ij} \leq 1$, offering a far richer class of possibilities for spatial smoothing. For example, we might choose the w_{ij} inversely proportional to the distance separating the centroids of regions i and j .

Even more generally, we could think of the w_{ij} as additional unknown parameters to be estimated, thus allowing the data (and perhaps other observed covariate information) to help determine the degree and nature of spatial smoothing. For example, suppose for all adjacent regions i and j , we follow an idea from statistical social network analysis (Wang and Wong 1987; Hoff, Raftery, and Handcock 2002) and model the w_{ij} as

$$w_{ij} | p_{ij} \sim \text{Bernoulli}(p_{ij}), \text{ where } \log \left(\frac{p_{ij}}{1 - p_{ij}} \right) = \mathbf{z}'_{ij} \boldsymbol{\gamma} \quad (11)$$

That is, border segment ij is a BE with probability p_{ij} , which in turn arises from a logit model that depends on a set of boundary-specific covariates \mathbf{z}_{ij} with corresponding parameter vector $\boldsymbol{\gamma}$. A wide variety of covariates might be considered here; for instance, we might set $z_{1ij} = 1$ (so that γ_1 is an intercept parameter); $z_{2ij} = d_{ij}$, the distance between the centroids of regions i and j ; $z_{3ij} = (\text{area}_i + \text{area}_j)/2$, the average area of the two regions; and $z_{4ij} = |x_i - x_j|$, the absolute difference of some regional covariate (percent urban, percent of residents who are smokers, or even the region's expected age-adjusted disease count). Such covariates may arise without error (as from a census), or with error (as from a survey asking whether respondents have undergone a colorectal cancer screening procedure within the last 2 years) that may be acknowledged by adding an errors-in-covariates term to the model (see, e.g., Xia and Carlin 1998). Reilly (2001) shows that $\boldsymbol{\gamma}$ is estimable from the data, even under a noninformative prior distribution for $\boldsymbol{\gamma}$ (say, a multi-

variate normal with a vague covariance specification). For γ_4 , say, we would expect a negative value to emerge, so that regions that are adjacent but have dissimilar covariate values would be less likely to be considered “neighbors” in the spatial model. Wombléd boundaries might then be based on the posterior distribution of the Δ_{ij} as before, or perhaps on the posterior of the w_{ij} themselves, as they represent the extent to which i and j should be thought of as similar or dissimilar. For instance, crisp boundaries could arise as those segments ij having $P(w_{ij} = 0|\mathbf{y}) > c^*$, while fuzzy boundaries could use the $P(w_{ij} = 0|\mathbf{y})$ values themselves as the BMVs.

As a final, similar but more streamlined generalization, we might delete the p_{ij} from model (11) and instead place the logit structure directly on the w_{ij} , that is, $\log\left(\frac{w_{ij}}{1-w_{ij}}\right) = \mathbf{z}'_{ij}\boldsymbol{\gamma}$ for i, j adjacent. The w_{ij} will now have distributions (induced by γ_0 and γ_1) residing on the entire interval $[0, 1]$, making their posterior means very natural fuzzy BMVs.

In summary, advanced areal wombling models of this type offer the potential to further expand on the aforementioned advantages hierarchical modeling methods enjoy over traditional approaches to boundary analysis. As spatially referenced areal data become more and more available in an ever-increasing range of application areas, the need to precisely measure and quantify the significance of boundaries will exhibit commensurate growth. Our preliminary efforts suggest a bright future for the merger of statistical methods with geographical analysis tools and software in this regard.

Acknowledgements

We are grateful to Prof. Sudipto Banerjee and Dr Sally Bushhouse for providing and permitting analysis of the data sets herein, as well as for helpful discussions that were essential to this project's completion.

References

- Banerjee, S., B. P. Carlin, and A. E. Gelfand. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, FL: Chapman & Hall/CRC Press.
- Banerjee, S., A. E. Gelfand, and C. F. Sirmans. (2003). “Directional Rates of Change Under Spatial Process Models.” *Journal of the American Statistical Association* 98, 946–54.
- Barbujani, G., G. M. Jacquez, and L. Ligi. (1990). “Diversity of Some Gene Frequencies in European and Asian Populations V. Steep Multilocus Clines.” *American Journal of Human Genetics* 47, 867–75.
- Barbujani, G., N. L. Oden, and R. R. Sokal. (1989). “Detecting Regions of Abrupt Change in Maps of Biological Variables.” *Systematic Zoology* 38, 376–89.
- Besag, J. (1974). “Spatial Interaction and the Statistical Analysis of Lattice Systems (with Discussion).” *Journal of the Royal Statistical Society, Series B* 36, 192–236.
- Besag, J., J. C. York, and A. Mollie. (1991). “Bayesian Image Restoration, With Two Applications in Spatial Statistics.” *Annals of the Institute of Statistical Mathematics* 43, 1–59.
- Bocquet-Appel, J. P., and J. N. Bacro. (1994). “Generalized Wombling.” *Systematic Biology* 43, 442–48.

- Bocquet-Appel, J.-P., and L. Jakobi. (1996). "Barriers for the Spatial Diffusion for the Demographic Transition in Europe." In *Spatial Analysis of Biodemographic Data*, 117–29, edited by J. P. Bocquet-Appel, D. Courgeau, and D. Pumain. Londres et Paris, France: Eurotext, John Libbey.
- Cressie, N. A. C. (1993). *Statistics for Spatial Data*, revised ed. New York: Wiley.
- Csillag, F., B. Boots, M.-J. Fortin, K. Lowell, and F. Potvin. (2001). "Multiscale Characterization of Boundaries and Landscape Ecological Patterns." *Geomatica* 55, 291–307.
- Fortin, M.-J. (1994). "Edge Detection Algorithms for Two-Dimensional Ecological Data." *Ecology* 75, 956–65.
- Fortin, M.-J. (1997). "Effects of Data Types on Vegetation Boundary Delineation." *Canadian Journal of Forest Research* 27, 1851–58.
- Fortin, M.-J., and P. Drapeau. (1995). "Delineation of Ecological Boundaries: Comparisons of Approaches and Significance Tests." *Oikos* 72, 323–32.
- Gelfand, A. E., and P. Vounatsou. (2003). "Proper Multivariate Conditional Autoregressive Models for Spatial Data Analysis." *Biostatistics* 4, 11–25.
- Hodges, J. S., B. P. Carlin, and Q. Fan. (2003). "On the Precision of the Conditionally Autoregressive Prior in Spatial Models." *Biometrics* 59, 317–22.
- Hoff, P. D., A. E. Raftery, and M. S. Handcock. (2002). "Latent Space Approaches to Social Network Analysis." *Journal of the American Statistical Association* 97, 1090–98.
- Jacquez, G. M., and D. A. Greiling. (2003). "Geographic Boundaries in Breast, Lung and Colorectal Cancers in Relation to Exposure to air Toxics in Long Island, New York." *International Journal of Health Geographics* 2, 4.
- Jacquez, G. M., and S. L. Maruca. (1998). "Geographic Boundary Detection." In *Proceedings of the 8th International Symposium on Spatial Data Handling*, edited by T. K. Poiker and N. Chrisman. Burnaby, BC, Canada: International Geographic Union, Geographic Information Science Study Group.
- Jacquez, G. M., S. Maruca, and M.-J. Fortin. (2000). "From Fields to Objects: A Review of Geographic Boundary Analysis." *Journal of Geographical Systems* 2, 221–41.
- Lawson, A. B., and A. Clark. (2002). "Spatial Mixture Relative Risk Models Applied to Disease Mapping." *Statistics in Medicine* 21, 359–70.
- Oden, N. L., R. R. Sokal, M.-J. Fortin, and H. Goebel. (1993). "Categorical Wombling: Detecting Regions of Significant Change in Spatially Located Categorical Variables." *Geographical Analysis* 25, 315–36.
- Reilly, C. (2001). "Modeling Adjacency in Lattice Models." *Research Report*, Division of Biostatistics, University of Minnesota.
- Sokal, R. R., and B. A. Thompson. (1998). "Spatial Genetic Structure of Human Populations in Japan." *Human Biology* 70, 1–22.
- Spiegelhalter, D. J., N. Best, B. P. Carlin, and A. van der Linde. (2002). "Bayesian Measures of Model Complexity and Fit (With Discussion)." *Journal of the Royal Statistical Society, Series B* 64, 583–639.
- Wang, Y. J., and G. Y. Wong. (1987). "Stochastic Blockmodels for Directed Graphs." *Journal of the American Statistical Association* 82, 8–19.
- Womble, W. (1951). "Differential Systematics." *Science* 114, 315–22.
- Xia, H., and B. P. Carlin. (1998). "Spatio-Temporal Models with Errors in Covariates: Mapping Ohio Lung Cancer Mortality." *Statistics in Medicine* 17, 2025–43.