THE MAP COMPARISON PROBLEM: TESTS FOR THE
OVERLAP OF GEOGRAPHIC BOUNDARIES

GEOFFREY M. JACQUEZ
BioMedware, Inc., 516 North State Street, Ann Arbor, MI 48104-1236, U.S.A.

SUMMARY
The quality of environmental studies is often compromised by the use of statistics, such as correlation and regression for example, which presuppose a statistical model, linear or otherwise, between two variables. When investigating hypotheses about relationships among geographically distributed variables, an alternative approach is to measure the amount of boundary overlap. Boundaries are geographic zones of rapid change in the intensity of a variable, and are often of scientific interest in their own right. Examples of boundaries include ecotones, genetic hybrid zones, pollution plumes, and the front of the wave of advance of an epidemic. Boundary overlap describes zones where boundaries from two or more variables coincide, and are useful for evaluating epidemiologic hypotheses relating health to environmental exposures. This paper proposes four statistics of boundary overlap, and explores their performance using simulation models and real data describing ozone concentrations and hospital admissions for respiratory conditions. The statistics are found sensitive to different aspects of boundary overlap, and provide an additional diagnostic tool in the analysis of geographically distributed variables. Overlap statistics are expected to come into increasing use as the installed base of geographic information systems increases.

1. INTRODUCTION
Comparisons of two or more spatially distributed variables are used in the biomedical sciences both to generate and explore hypotheses describing putative relationships between the environment and human health. Data may be said to possess a spatial component whenever observations are sampled from several geographic locations. Sometimes this component is made explicit, as when a study explores spatial pattern. At other times the data's spatial nature is implicit and subsumed within the overall study design. Examples of medical inquiries involving spatially distributed variables include air pollution and respiratory illness, environmental risk factors and cancers, and agricultural and industrial exposures and cancer, to name a few. Maps arise whenever observations are made at several locations, and the problem is to determine whether maps describing the spatial distributions of two or more variables have similar pattern. Call this the map comparison problem. We may search for similarities over entire maps, or we may focus on specific aspects or sub-areas, such as zones of high intensity or of rapid change in the variables. Typically, one has several maps characterizing the study area, and the problem from an epidemiologic perspective is to determine whether health and environment maps are somehow related to one another.

1.1. Spatial fields
Define a spatial field as observations on a variable (continuous or discrete) at different geographic locations. This contrasts with the term surface, which has been used in the literature to convey the

CCC 0277–6715/95/212343–19
© 1995 by John Wiley & Sons, Ltd.
idea of a continuous spatial distribution underlying the observations. We prefer 'spatial field' because the term accurately describes the data as observations whose spatial distribution is discrete. Spatial fields may be regular or irregular. A regular spatial field arises when observations are sampled on a grid, with one observation at each grid node. Examples include satellite imagery, where each pixel constitutes a grid node, and certain agricultural trials with a complete block design. Most readers are familiar with irregular spatial fields where sample locations are scattered through an area. Irregular spatial fields are common in the biomedical arena because health and environmental variables typically are sampled at different locations. For example, health status may be evaluated at hospitals (such as, hospital admissions for respiratory conditions), while air quality is measured at air monitoring stations, and the locations of hospitals and air monitoring stations are not coincident. The lack of coincident sample locations when more than two variables are considered is a complication that arises often for irregular spatial fields, but may also arise when the grid nodes for regular spatial fields are not coincident (do not register) across variables.

1.2. Spatial autocorrelation

An important and often unavoidable issue in map comparisons is that of spatial autocorrelation. Spatial autocorrelation arises whenever the intensity (value) of a variable at one location is related to its intensity at other locations. Most data will exhibit positive spatial autocorrelation such that nearby locations tend to have similar intensities. This theme is so common that the pre-eminent geographer Waldo Tobler\textsuperscript{11} coined the first law of geography to be 'nearby things tend to be alike'. There is nothing magical about this first law; it applies because nearby places have a common history, and because environments in nearby locations are themselves similar. There are many examples of this pattern in medical geography, perhaps the foremost being the observation that patients from a given area often present similar conditions and symptoms. This occurs because socio-economic status is similar within neighbourhoods, and because individuals within a neighbourhood have common exposures and use the same services (water supplies, restaurants, schools, stores, hospitals, etc.). Environmental variables (such as air pollution) manifest Tobler's first law of geography because they are localized phenomena (for example, pollutants discharge at particular locations) whose dispersal is mediated by space–time processes (for example, spatial diffusion and transport).

A distinction is made between true, induced, spurious and interpolation autocorrelation. Causal interaction between adjacent locations gives rise to true spatial autocorrelation. Autocorrelation may be induced in a dependent variable through a causal relationship with another variable that is itself spatially patterned. When locations are statistically independent spurious autocorrelation will occasionally be observed by chance alone. Interpolation autocorrelation results when values at not-gauged locations are estimated by interpolation or extrapolation.

Measures for quantifying spatial autocorrelation include Moran's\textsuperscript{12} I, Geary's\textsuperscript{13} c, the variogram\textsuperscript{14} and the join-count statistic,\textsuperscript{15} to name a few (but see references in 15–18 for access to the literature). Computer programs for calculating spatial autocorrelation statistics include Stat\textsuperscript{19} and C2D\textsuperscript{20} (contact the author for availability). Autocorrelation is important in the context of the map comparison problem for two reasons. First, it will almost always be present, for reasons implicit in Tobler's first law of geography. Second, when present it biases traditional statistics such as correlation, regression, ANOVA and other methods that assume independent observations.\textsuperscript{15–22} In general the direction of bias is towards a false positive, so one is at risk of declaring significance when it does not actually exist (type I error).\textsuperscript{15,18} Recent research has
developed methods for correcting correlation,\textsuperscript{22–25} regression,\textsuperscript{26,27} ANOVA\textsuperscript{28} and other statistics\textsuperscript{15,21} for spatial autocorrelation.\textsuperscript{29}

1.3. Review of map comparison methods

After correction for spatial autocorrelation, regression and correlation are suited for map comparisons, and the use of these tests is detailed by authors referenced earlier.\textsuperscript{21–28} Tests for matrix association\textsuperscript{30–32} may be used when relationships among sampling locations are first expressed as distance or similarity matrices. Often the null distribution of the statistic is generated using Monte Carlo methods, but caution must be used as the permutation procedures typically assume spatial randomness (see references 33 and 34 for access to the literature on permutation tests). Some authors\textsuperscript{21,35,36} use restricted randomizations to control for spatial autocorrelation in the variables. When using Mantel’s statistic one also can account for a common geography by partialing geographic relationships out of the health-exposure correlation.\textsuperscript{37,38} This approach corrects the health-exposure correlation by partialing out association arising from the health-geography and exposure-geography correlations.

These methods require observations on all variables at the same sampling locations. When this is not the case researchers may first interpolate to obtain observations at the not-gauged locations, but this procedure introduces interpolation autocorrelation which itself may bias subsequent statistical procedures. A second alternative is to delete locations from the analysis when one or more of the variables are not observed at those locations. This can drastically reduce the final sample size and, when used with bivariate statistics such as correlation, results in different sample sizes for different elements of the correlation matrix (see references 39 and 40 for discussion on the missing location problem).

Often scientific interest may focus on geographic boundaries, which are defined as zones of rapid change in a spatial field. Boundaries are of interest for several reasons. First, boundaries represent transitional areas and are often sampled more intensively in order to characterize the variance. Second, at early stages of an inquiry one may not have sufficient knowledge to specify a formal model. In these instances boundary overlap provides a useful measure of spatial association. Finally, boundaries are of significant scientific interest in their own right because of their biomedical and biological implications. Examples include: hybrid zones in population biology,\textsuperscript{41} where gene frequency boundaries can exist between interbreeding populations; groundwater pollution in hydrology, where boundaries exist at the edges of plumes; ecotones in ecology,\textsuperscript{42} where boundaries define interfaces between vegetation communities; landscapes in conservation biology,\textsuperscript{43–45} where boundaries define the spatial extent of distinct ecosystems; and epidemic fronts in epidemiology, where a boundary is the leading edge of a space-time epidemic. There are many methods for quantifying boundaries.\textsuperscript{46} These include maximum-difference barriers\textsuperscript{47} in geography and edge detection algorithms\textsuperscript{48} in the field of image analysis.\textsuperscript{49} Gradient operators such as the Womble method\textsuperscript{50} have been used with some success in ecology\textsuperscript{51} and population genetics.\textsuperscript{52} Fortin\textsuperscript{42,53} has evaluated the use of this algorithm for quantifying vegetation and edaphic boundaries.

In general, opportunities for analysing zones of rapid change in the medical sciences are increasing as the use of geographic information systems burgeons. It is expected that the analysis of boundaries will increase our understanding of the fundamental and spatial relationships between the environment and human health. Of particular interest is the question of whether the spatial locations of boundaries in health and environmental variables co-occur. A related paper\textsuperscript{54} summarizes recent research on the topic of overlap statistics and proposes six ways of incorporating spatial structure in permutation tests. This paper proposes four statistics for boundary
overlap, and evaluates them using three spatial models. The statistics are then applied to a data set describing air pollution and human respiratory health in Ontario, Canada.

2. METHODS

2.1. Notation

The problem is to determine whether the boundaries from two different variables occupy the same locations – overlap – more than would be expected if there were no association between the two variables. Assume two variables, $G$ (describing exposure) and $H$ (health-related), with $N_G$ and $N_H$ observations, respectively ($X$ and $Y$ are used later for Cartesian co-ordinates). Let $L_G = \{l_{G1}, l_{G2}, \ldots, l_{GN_G}\}$ denote the set of locations where $G$ was observed and $L_H$ the locations where $H$ was observed. Suppose boundaries for $G$ and $H$, called $B_G$ and $B_H$, have been determined using one of the boundary detection algorithms described earlier. The boundaries are then verified to assure they are biologically and physically reasonable, so that, for example, health boundaries do not occur over water. $N_{B_G}$ is the number of locations defining boundary $B_G$, and $N_{B_H}$ is the number of locations defining boundary $B_H$. Define the distance matrix $D$ of dimension $N_{B_G} \times N_{B_H}$ whose elements, $d_{ij}$, are the distances between location $i$ in the boundary for variable $G$ and location $j$ in the boundary for variable $H$. This distance is computed as some metric on the boundary locations. In later examples we use Euclidean distance as the metric, although great circle or metrics accounting for altitude or travel distance may be used as well. Let $\min(d_{ij})$ be the smallest distance in column $j$ of this distance matrix, and $\min(d_{ij})$ be the smallest distance in row $i$.

2.2. Overlap statistics

The overlap statistics are

$$O_s = \text{card}(B_G \cap B_H)$$

$$O_G = \frac{\sum_{i=1}^{N_{B_G}} \min(d_{ij})}{N_G}$$

$$O_H = \frac{\sum_{j=1}^{N_{B_H}} \min(d_{ij})}{N_H}$$

$$O_{GH} = \frac{\sum_{i=1}^{N_{B_G}} \min(d_{ij}) + \sum_{j=1}^{N_{B_H}} \min(d_{ij})}{N_G + N_H}.$$

The operation card $(A)$ is the cardinality or count of the number of elements in set $A$. $O_s$ (equation (1)) is simply the count of the number of locations that are in both of the sets defining the boundaries for variables $G$ and $H$. It measures how frequently boundaries in the two variables overlap exactly, and was originally proposed by Barbujani and Sokal. $O_G$ (equation (2)) is the mean distance from any location in the boundaries for $G$ to the nearest location in the boundaries for $H$. $O_H$ (equation (3)) is the mean distance from any location in the boundaries for $H$ to the nearest location in the boundaries for $G$. $O_{GH}$ (equation (4)) is the average distance from a location in the $G$ or $H$ boundaries to the nearest location in the other ($G$ or $H$ as appropriate) boundary.
These statistics are sensitive to different aspects of boundary overlap, as described later in the discussion.

2.3. Null distributions and \( P \)-values

Null distributions of the overlap statistics are obtained using a five-step randomization procedure.

*Step 1: Calculate overlap statistics*

Calculate the boundaries \( B_G \) and \( B_H \) and the observed statistics \( O^*_G \), \( O^*_G \), \( O^*_H \), and \( O^*_G|H \). The asterisks indicate these are statistics calculated for the original (not randomized) data.

*Step 2: Conduct Randomization*

Recalling \( G \) is an environmental variable and \( H \) is health-related, the null hypothesis assumes \( G \) and \( H \) are independent and the epidemiologic alternative is that \( G \) influences \( H (G \Rightarrow H) \). When simple association (not causality) between \( G \) and \( H \) is hypothesized one may use *unconditional randomization* and randomize both variables. When the causal relationship \( G \Rightarrow H \) is hypothesized a *conditional randomization* (conditioned on the geographic distribution of \( H \)) may be conducted by re-allocating the observations on \( H \) among the \( N_H \) sampling locations. This is a shuffling of the vector of observations on \( H \) while holding the vector of locations (\( L_H \)) constant. The conditional randomization assumes the spatial distribution of \( G \) as a given, and the variance of the test statistic may be smaller (resulting in smaller \( p \)-values) than for unconditional randomization. Which approach is used sets the context for interpreting the results.

The randomization (conditional or unconditional) can be conducted with equal probability across the locations, corresponding to a null hypothesis of no spatial autocorrelation, or the randomization may be spatially restricted, corresponding to a more complex null hypothesis. For example, if \( H \) denotes regional counts (not rates) one may wish to restrict randomizations to reflect population size in the areas. At present we are not aware of a satisfactory means for restricting randomizations to retain a given level of spatial autocorrelation in the data, and instead use spatially random reference distributions (but see reference 56 for a discussion of restricted randomization procedures that retain spatial autocorrelation). Oden et al.\(^{57}\) have shown randomized reference distributions to result in conservative tests for the detection of geographic boundaries.

*Step 3: Boundaries for the Randomized Data*

Calculate the boundary \( B_H \) using the shuffled observations on \( H \). Then calculate a new distance matrix \( D \). The boundaries for \( G \) are unchanged.

*Step 4: Calculate Overlap Statistics for the Randomized Data*

Calculate and save the statistics \( O^*_S \), \( O^*_G \), and \( O^*_H \) and \( O^*_G|H \).

*Step 5: Construct Null Distributions*

Repeat steps 2 to 4 until the desired number of randomizations is achieved. \( P \)-values are determined by comparing the observed statistics (from step 1) to their null distributions.
2.4. Interpretation

The statistics $\bar{O}_a$, $\bar{O}_b$, $\bar{O}_c$ and $\bar{O}_{cH}$ are the means of the four overlap statistics under randomization, and $S_{O_a}$, $S_{O_b}$, $S_{O_c}$ and $S_{O_{cH}}$ are their associated standard deviations. The overlap statistics are standardized as

$$O'_a = \frac{O'_a - \bar{O}_a}{S_{O_a}}$$  \hspace{1cm} (5)

$$O'_b = \frac{O'_b - \bar{O}_b}{S_{O_b}}$$  \hspace{1cm} (6)

$$O'_c = \frac{O'_c - \bar{O}_c}{S_{O_c}}$$  \hspace{1cm} (7)

$$O'_{cH} = \frac{O'_{cH} - \bar{O}_{cH}}{S_{O_{cH}}}.$$  \hspace{1cm} (8)

Positive values of the standardized statistics indicate boundary overlap and negative values indicate overlap avoidance (hence the leading minus signs in equations (6)–(8). In our simulation studies the null distributions of the overlap statistics were symmetric and the standardized statistics had a mean near zero when there was no association between the boundaries.

2.5. Boundary Detection

As described earlier there are numerous boundary detection methods, and the overlap statistics described in this paper can be used with any of them. In this study we use the Womble\textsuperscript{58} method, which is a gradient operator increasingly used in the biomedical sciences (see references 58 and 59 for early applications, and references 42, 51–55, 59 for recent developments). The input data are regular or irregular spatial fields for two variables. If irregular the data may first be interpolated to the nodes of a grid, or, if one wishes to avoid interpolation autocorrelation, one may use a Delauney triangulation instead (see reference 42 for the Delauney approach). The tests for boundary overlap described in this paper may be used in either case. For simplicity of presentation assume regular spatial fields for both variables. The slope at each node of the spatial field is then evaluated using the bilinear function\textsuperscript{50} described by Barbujani \textit{et al.}\textsuperscript{51} which has known derivatives. Consider $X$ and $Y$ to be Cartesian co-ordinates in the geographic plane, and $Z$ to be a third dimension orthogonal to $X$, $Y$ and describing the variable under scrutiny. Let \{z$_A$, z$_B$, z$_C$, z$_D$\} be the levels of $Z$ at four locations \{A, B, C, D\}. Further assume these locations form a unit square with geographic co-ordinates A(0,0), B(1,0), C(1,1) and D(0,1). When the square is made infinitely small the derivative of the bilinear function is an estimate of the instantaneous slope. The bilinear function is

$$f(X,Y) = z_A(1-x)(1-y) + z_Bx(1-y) + z_Cxy + z_D(1-x)y.$$  \hspace{1cm} (9)

This function yields an estimate of the variable at any location $(x, y)$ within the unit square. The slope in the centre of the unit square is obtained from the magnitude of the partial derivatives as

$$m = \sqrt{(\partial f/\partial x)^2 + (\partial f/\partial y)^2}.$$  \hspace{1cm} (10)
The direction of the slope is the angle made with the X axis

\[ \theta = \arctan \left( \frac{\partial f/\partial x}{\partial f/\partial y} \right) + \Delta. \]  

(11)

This is the direction of the slope described in equation (10). Here \( \Delta = 0 \) if \( \partial f/\partial x > 0 \) and \( \Delta = 180 \) if \( \partial f/\partial y < 0 \). This produces \( \theta \) in the range 0 to 360 degrees. Leaving \( \Delta \) out of the equation results in angles in the range 0 to 180 degrees, which is desirable when opposite facing slopes are considered to have the same orientation. This situation arises when one is interested only in the absolute magnitude of the difference between locations.

Two matrices are calculated for each variable, one containing the magnitudes of the derivatives and another the directions. Together, these two matrices describe a vector field indicating magnitude and direction of slope throughout the geographic space. Barbujani et al.\textsuperscript{51,52} judge a vector to be part of a boundary when its magnitude is in the upper 5 per cent of all magnitudes, it is adjacent to other high magnitude vectors, and when the direction of the slope of adjacent vectors differ by no more than 30 degrees. This prevents areas of rapid, random variation from being classified as boundaries. The resulting boundaries are mapped by plotting their vectors as lines whose orientations are given by the direction matrix and whose lengths are given by the slope magnitude matrix, resulting in what have been described by some as ‘fuzzy caterpillars’. Alternatively, boundaries may be visualized as subgraphs which connect high magnitude vectors, as described by Oden et al.\textsuperscript{57}

When applied to regular spatial fields, the Womble algorithm (termed lattice wombling in Jacquez and Fortin\textsuperscript{56}) is subject to interpolation autocorrelation. If desired, the sensitivity of the overlap statistics to interpolation autocorrelation can be quantified by using different interpolation algorithms, or, alternatively, one may use triangulation or categorical wombling which operate directly on irregular spatial fields (see reference 56 for a description of the triangulation and categorical approaches).

3. SIMULATIONS

Simulations to evaluate the performance of the overlap statistics were conducted using spatially autocorrelated variables. Three spatial models of the relationship between the environmental (G) and health variables (H) were used: independence; small-scale boundaries, and large-scale boundaries. The spatial fields on which the data were modelled were regular and of dimension 25 x 25, for a total of 625 samples on each field. 50 samples were taken at random from both the G and H spatial fields to produce irregular spatial fields. Irregular spatial fields were used because they are representative of the irregular spatial sampling typical of environmental studies. The Womble method was used to detect boundaries in the simulated data, and the four overlap statistics G and H were applied. The spatial models are now described.

Independence

The G and H spatial fields were populated by repeatedly sampling a normal distribution of mean 50 and standard deviation 25 (Figure 1, top). These fields were then smoothed to produce positive spatial autocorrelation using a first-order moving average based on king’s moves. The smoothed fields are shown at the top of Figure 2. Fifty samples were taken at random from each of these fields, one representing the observations of G and the other the observations on H (Figure 3, top). Sampling locations differed for the two variables. This models G and H is independent, spatially autocorrelated variables without overlapping boundaries.
Figure 1. Scattergrams of the response variable ($H$) on the environmental variable ($G$) for the three models. Independence (top), small-scale boundaries (middle) and large-scale boundaries (bottom).

Small-scale overlapping boundaries

Dose–response relationships often are not linear, and dependence of the health variable on the environmental variable was modelled using

$$h = \frac{\beta}{1.0 + (\beta - \theta_0) / \theta_0 e^{-2g}} + e.$$

(12)
Figure 2. Surface plots of the simulated regular spatial fields. The environmental variable (G) is in column 1, the response variable (H) is column 2. Each row is a model: independence (top), small-scale boundaries (middle) and large-scale boundaries (bottom). Fifty samples were taken from each of these spatial fields to yield the irregular spatial fields used in the analysis (see Figure 3).
Figure 3. Surfaces interpolated from the irregular spatial fields (50 samples each). Independence (top), small-scale boundaries (middle) and large-scale boundaries (bottom); G left, H right
Parameter values of $\beta = 50$, $\alpha = 0.1$ and $g_0 = 0.05$ were used. This produced a function which increased rapidly after exceeding a 'threshold' of $g \sim 50$ (Figure 1, middle and bottom). Such a function is observed, for example, in acute responses to a noxious substance. The spatial field for $G$ was generated by sampling an $N(50, 25)$ distribution. While the normal distribution provides a starting place for evaluating the methodology, it may not be appropriate for counts of rare diseases where a log-normal distribution might instead be preferred. Sensitivity of the overlap statistics to the underlying disease process is a continuing area of research. Spatial autocorrelation was introduced by smoothing the spatial field as described earlier (Figure 2, middle left). Observed values of $H$ were obtained by using corresponding observations from the $G$ spatial field as input to equation (12) (Figure 2, middle right). The stochastic term $\varepsilon$ represents random variation in $H$, and was sampled from an $N(0,1)$ distribution. Irregular spatial fields of 50 observations each were then sampled from $G$ and $H$ (Figure 3, middle). This model assumes $H$ is non-linearly dependent on $G$, $G$ exhibits true spatial autocorrelation, and that spatial autocorrelation in $H$ is induced by the autocorrelation of $G$. Small-scale, overlapping boundaries result when $G$ induces spatial pattern in $H$.

**Large-scale overlapping boundaries**

Boundaries on a large geographic scale were incorporated in the model by sampling $G$ from an $N(50/3, 25/3)$ distribution when the row location in the grid was less than 12, otherwise $G$ was sampled from an $N(50, 25)$ distribution (Figure 2, bottom). Other parameters were identical to small-scale boundaries. This models non-linear dependence of $H$ on $G$, induced autocorrelation in $H$ and a large-scale overlapping boundary.

### 3.1. Simulation Results

Typical scattergrams for the three models are shown in Figure 1. The scattergram for the independence model appears bivariate-normal. The small boundaries model shows a rapid increase in $H$ once $G$ exceeds a threshold level. The scattergram for the large-scale boundaries model demonstrates a bimodal distribution. Smaller values were used to populate one half of the $G$ field, with higher values in the other half. This produced a well-defined boundary which appears as a 'cliff' in the surface plot (Figure 2, bottom left). This pattern might be observed, for example, at the edge of a pollution plume.

This cliff is present, although less apparent, in the surface plot for $H$ (Figure 2, lower right). The surface plots for the independence model (Figure 2, top row) are jagged, even though spatial autocorrelation in the model tends to broaden the peaks and valleys. Three peaks in the $G$ surface from the small-scale boundaries model (Figure 2, centre row) are reflected by peaks in the surface plot for the $H$ surface (Figure 2, centre).

These surfaces represent complete knowledge since the entire $25 \times 25$ spatial field was used. In the real world spatial knowledge is usually incomplete, and we often must rely on fewer samples in an irregular spatial field. Fifty observations were sampled from the complete spatial fields in Figure 1 to mimic the incomplete information typical of field studies. These irregular spatial fields were then interpolated using kriging\textsuperscript{14,62,63} to produce the surface plots in Figure 3. Notice the loss of small-scale variation caused by the reduced sample size and interpolative smoothing. None the less, many of the large-scale features are still present, in particular the cliffs in the large-scale boundary model (Figure 3, bottom) and the central peak in the small boundary model (Figure 3, centre).

Boundaries in these kriged spatial fields were then determined using the Womble procedure and the boundary elements are shown as vectors defined by slope magnitude (vector length) and direction (vector orientation) in Figure 4. Based on visual inspection it is difficult to determine
Figure 4. Boundaries calculated from the irregular spatial fields using the Wamble method. Independence (top), small-scale boundaries (middle) and large-scale boundaries (bottom); G left, H right. Wamble vectors are plotted as lines whose orientation shows the direction of maximum change in slope and whose length reflects slope magnitude.
whether the $G$ and $H$ variables have common boundaries, which arises when $G$ and $H$ boundary elements tend to occur in the same place. The independence model (Figure 4, row 1) demonstrates an area of high surface slope in the vicinity of $\{3, 6\}$ (here 3 is the location on the $X$ axis, 6 the location on the $Y$ axis), yet the surfaces are independent. The boundaries of the small boundary model (Figure 4, middle) seem unrelated, except perhaps for an area of high surface slope which appears along the western edge of the map. The large boundary model (Figure 4, bottom) shows boundaries in the centre of the maps, but they seem to be offset for the $G$ and $H$ variables. It appears that visual inspection alone cannot determine whether two variables have common boundaries.

3.2. Overlap statistics

The four overlap statistics were calculated from the Womble boundaries (Figure 4), producing three sets of overlap statistics, one each for independence, small-scale boundary and large-scale boundary models. The significance of these statistics was evaluated under randomization by holding the $G$ surface constant and permuting the 50 observations on $H$ under the null hypothesis of independence. Boundaries on the rearranged surface were then determined using Womble analysis, and the overlap statistics were calculated and recorded. This process was repeated 249 times to construct a distribution of the overlap statistics under the null hypothesis of no spatial association between the $G$ and $H$ variables.

None of the overlap statistics was significant for the independence model. Both $O_G$ and $O_{GH}$ are significant for the small-scale boundary model ($P = 0.004$ and $0.04$, respectively). $O_G$, $O_{GH}$ and $O_{GR}$ are significant ($P = 0.004$, $P = 0.04$ and $P = 0.044$, respectively) for the large-scale boundary model. Before discussing the implications of the simulation results first consider an application.

4. APPLICATION: THE RELATIONSHIP BETWEEN RESPIRATORY ILLNESS AND AIR POLLUTION

It is well known that exposure to elevated ozone levels causes acute respiratory distress leading to pulmonary oedema and that chronic exposure is associated with emphysema.\textsuperscript{54}

It is not clear, however, whether exposure in the every-day environment to zones of rapid change (boundaries) in ozone is sufficient to cause concomitant boundaries in hospital admissions for respiratory conditions (but see references 1–4 and 62–68 for studies addressing the relationship between air quality and respiratory health). We hypothesized that exposure to high environmental ozone might cause acute respiratory response in sensitive individuals, thereby resulting in increased admissions for respiratory conditions. To address this question we scanned time series of hourly observations at 26 ozone monitoring stations in Southern Ontario to identify a six-year peak in ozone which occurred during 1–7 August 1984. Inspection of maps of ozone levels during the pollution event revealed large differences in pollution levels over most of Southern Ontario defined by distinct geographic boundaries (Figure 5, top).

Admission data at 77 acute-care hospitals were obtained from the Ontario Ministry of Health. Respiratory admissions were defined based on ICD-9 primary diagnosis codes: acute bronchitis (ICD = 466); viral, bacterial and broncho-pneumonias (480, 481, 482 and 485); bronchitis, chronic bronchitis, emphysema (490, 491 and 492), and asthma (493). At each hospital a standardized daily respiratory admissions rate was calculated by subtracting the mean daily rate over the 11-week period bracketing the pollution episode and dividing by the standard deviation in
5. DISCUSSION

What do these overlap statistics tell us about whether or not two variables have common geographic boundaries? $Q_i$ is the count of the number of times locations of rapid change in
Table I. Overlap statistics and $P$-values. Asterisks indicate significance levels: *$P \leq 0.05$, **$P \leq 0.01$

<table>
<thead>
<tr>
<th>Data</th>
<th>$O_s'$</th>
<th>$O_G'$</th>
<th>$O_H'$</th>
<th>$O_{GH}'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independence</td>
<td>0.610</td>
<td>-0.212</td>
<td>0.304</td>
<td>0.085</td>
</tr>
<tr>
<td>Small boundaries</td>
<td>1.156</td>
<td>2.293**</td>
<td>0.801</td>
<td>1.724*</td>
</tr>
<tr>
<td>Large boundaries</td>
<td>3.752**</td>
<td>1.460**</td>
<td>1.403</td>
<td>1.471*</td>
</tr>
<tr>
<td>Ozone and respiratory</td>
<td>0.295</td>
<td>-0.094</td>
<td>0.236</td>
<td>0.099</td>
</tr>
</tbody>
</table>

$H$ exactly overlay locations of rapid change in $G$. $O_s$ is thus highly sensitive to the precision of overlap, that is, whether or not two boundaries occur in exactly the same place. It is not significant when the two variables are independent and spatially autocorrelated, as exemplified by the independence model. Neither is it significant when $H$ depends on $G$ but boundaries on a broad geographic scale are absent, as represented by the small-scale boundary model. However, $O_s$ is significant when geographically well-defined boundaries are present, as shown by the large-scale boundaries model. $O_s$ is expected to be significant when geographically large boundaries, caused for example by phenomena such as ecotones, hybrid zones and barriers to transport, occur in the same location for both variables. It thus should be useful in determining, for example, whether areas of rapid genetic change are geographically coincident for different species. It may also be used within a species to determine whether alleles at different loci have a common boundary, suggestive perhaps of similar selective forces. A finding of a significant $O_s$ suggests the presence of a coincident, geographically large boundary in both variables.

$O_G$ is the average distance from a location of rapid change in $G$ to the nearest location of rapid change in $H$. $O_H$ is the average distance from a place of rapid change in $H$ to the nearest location of rapid change in $G$. These two statistics will be significant when boundaries are nearer to one another than expected. When boundaries overlap exactly $O_s$ will also be significant; when boundaries are offset from one another $O_s$ may not be significant but $O_G$ and/or $O_H$ may be significant.

The statistic $O_{GH}$ is the average distance from any location in a $G$ or $H$ boundary to the nearest location in a boundary for the other variable. It is thus a general measure of simultaneous fit between boundaries of the two variables.

Notice the statistics are sensitive only to boundary overlap, and do not discern on which side of a boundary areas of high or low exposure may lie. Whether exposure and health are directly correlated is best addressed using measures of association that correct for spatial autocorrelation (see review of map comparison methods'). In practice, the overlap statistics are perhaps best used to describe spatial pattern as part of an exploratory data analysis, or EDA (see reference 19 for a description of EDA). In this capacity boundary overlap serves as an additional tool for quantifying and exploring spatial relationships.

Does the pattern of results tell us anything about causation? Causal inference from observational data is a difficult task (see references 69–72 for discussion on this topic), and quantitative conclusions based on the significance results for the overlap statistics may not be possible. However, based on our studies to date, relationships among the variables can be reflected in the overlap statistics, and further investigation may be warranted, as now described.

Notice $O_G'$ was significant for both the small- and large-scale boundary models, while $O_H'$ was not. $H$ depends on $G$ in both models, and includes a stochastic term representing natural variation. This causal relationship means almost all of the high $G$ values induce high $H$ values in nearby locations. Stochasticity in $H$, however, means that not all high $H$ values have high
$G$ values nearby. The magnitudes of $O_{G}$ and $O_{H}$ may reflect the strength and direction of the relationship between two variables. $O_{G}$ significant and $O_{H}$ not significant suggests a causal relationship such that $G \Rightarrow H$. An alternative interpretation is an unobserved common cause driving both variables; $U \Rightarrow G$ and $U \Rightarrow H$, but with a weaker relationship for $U \Rightarrow H$ such that rapid change in $H$ can occur at locations removed from areas of rapid change in $G$. $O_{H}$ significant and $O_{G}$ not significant suggests $H \Rightarrow G$, or perhaps a common cause driving both variables but with more noise in $G$, so that some of the locations of rapid change in $G$ are isolated from areas of rapid change in $H$.

Now consider the results summarized in Table 1. None of the statistics was significant for the independence model, and one correctly infers the two variables do not share common boundaries. $O_{GH}$ is significant for both small- and large-scale boundary models, and one concludes the locations of boundaries in $G$ and $H$ are near to one another. $O_{G}$ is significant for the two boundary models, while $O_{H}$ is not, indicating that while boundaries in $G$ are significantly near boundaries in $H$, some of the $H$ boundaries are not near any $G$ boundaries. This reflects the stochastic non-linear dependence of $H$ and $G$. $O_{S}$ is significant only for the large-scale boundary model, leading one to correctly infer a common, geographically large boundary. Finally, none of the overlap statistics is significant for ozone and respiratory admissions response surfaces, and one concludes they have no boundaries in common, even though visual inspection of Figure 5 seems to indicate that they do.

The methods in this paper are new and much work remains to be done. How results depend on the boundary detection algorithm remains to be determined, and the statistical power of the statistics under different space-time models needs to be evaluated. Finally, the relationship between accuracy of boundary determination and the resolution of the data needs investigation. However, these preliminary results are promising and the approach has several advantages. First, incomplete knowledge of the underlying spatial field is permitted. Second, different numbers of sampling locations for the two variables are allowed, and the variables need not be sampled at the same places. Third, the four statistics are sensitive to different aspects of overlap, as described previously, which provides insight into how the variables might be related. Fourth, the method is sensitive not only to radical, well-defined surface changes, which will result in a significant $O_{S}$, but also to small-scale covariation across surfaces, which will result in significant $O_{G}$ or $O_{H}$. Finally, the methods do not require a model of how the two variables under scrutiny are related, rather, the statistics reflect the degree to which boundaries overlap. While this may be a weaker statement than fitting a model – statistical or otherwise – to the data, it is more robust, since, no matter what the parametric relationship between the variables may actually be, the locations of rapid change in the variables are at least observable, while fitting a model often requires the estimation of parameter values which cannot be observed. The exact form of the relationship between the two variables therefore is not needed. The overlap statistics provide a tool for determining whether two variables have geographic boundaries in common.

ACKNOWLEDGEMENTS

This paper was greatly improved by the editorial comments of Lance Waller and two anonymous referees. The research was funded in part by grant R43 CA65366-01A1 from the National Cancer Institute. Its contents are solely the responsibility of the author and do not necessarily represent the official views of NCI.

REFERENCES


