Spatial Cluster Detection for Repeatedly Measured Outcomes while Accounting for Residential History

Andrea J. Cook[1,2,*], **Diane R. Gold**[3,4] and **Yi Li**[5]

[1] Group Health Research Institute, Seattle, WA 98101, USA
[2] Department of Biostatistics, University of Washington, Seattle, WA 98105, USA
[3] The Channing Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA
[4] Department of Environmental Health, Harvard School of Public Health, Boston, MA 02115, USA
[5] Department of Biostatistics, Harvard School of Public Health and the Dana Farber Cancer Institute, Boston, MA 02115, USA

Received 15 December 2008, revised 20 June 2008, accepted 24 June 2008

Spatial cluster detection has become an important methodology in quantifying the effect of hazardous exposures. Previous methods have focused on cross-sectional outcomes that are binary or continuous. There are virtually no spatial cluster detection methods proposed for longitudinal outcomes. This paper proposes a new spatial cluster detection method for repeated outcomes using cumulative geographic residuals. A major advantage of this method is its ability to readily incorporate information on study participants relocation, which most cluster detection statistics cannot. Application of these methods will be illustrated by the Home Allergens and Asthma prospective cohort study analyzing the relationship between environmental exposures and repeated measured outcome, occurrence of wheeze in the last 6 months, while taking into account mobile locations.

Key words: Asthma; Cumulative residuals; Repeated measured; Spatial cluster detection; Wheeze.

1 Introduction

The prevalence of allergic diseases in children have greatly increased in the last few decades (Akinbami, Centers for Disease Control, and Prevention National Center for Health Statistics, 2006). What influences the onset of allergic diseases such as asthma and wheeze has become an increasingly important public health question. The Home Allergens and Asthma Study is an ongoing prospective cohort study investigating environmental and socioeconomic (SES) risk factors leading to early childhood respiratory diseases, such as asthma and wheezing, in the Boston, MA metropolitan area (Celedon *et al.*, 1999). Cross-sectional and longitudinal studies have tied home allergen levels (*e.g.* from cockroach and mouse), mold in the home, lower SES, and other individual or family-based measures of exposures to increased incidence or prevalence of wheeze, asthma, and allergic rhinitis (Brugge *et al.*, 2003; Finkelstein *et al.*, 2002). Fewer studies have focused on the larger area, or neighborhood, in which the individual resides as a source of environmental exposures that may influence the risk of allergic diseases (Litonjua *et al.*, 2005).

The immune development of an individual depends upon a complex interaction of factors related to genetics and environmental exposures that may derive from the larger neighborhood as well as the individual home. These exposures may have differing effects according to the age within which

* Correspondence author: e-mail: cook.aj@ghc.org, Phone: +1-206-287-4257, Fax: +1-206-287-2871

they occur and it is likely that an individual's immune development is influenced by their entire exposure history. Owing to this complexity, it is of substantial interest to detect spatial regions that have significantly higher odds of disease dependent on the age of the child. High-risk areas may indicate potential hazardous environmental sources (*e.g.* bus depots, poor housing, neighborhood waste sites, neighborhood violence).

To make conclusions about these questions of interest in the Home Allergens and Asthma Study, and other similar studies, there is a need to develop spatial cluster detection methods that handle longitudinal outcomes. Currently, numerous spatial cluster detection methods are available for the analysis of individual level data. For example, there are methods for binary outcomes assessing areas with elevated prevalence of disease and count outcomes evaluating excess rates of incidence or mortality (Kulldorff *et al.*, 2006; Tango and Takahashi, 2005; Duczmal and Assunção, 2004; Patil and Taillie, 2004; Tango, 2000; Kulldorff, 1997; Turnball *et al.*, 1990). There are even several methods for censored continuous outcomes exploring potential spatial clusters for detection of time to early event (Cook, Gold, and Li, 2007; Huang, Kulldorff, and Gregorio, 2007). However, there are no methods available for longitudinal outcomes.

The Home Allergens and Asthma Study has information about occurrence of wheeze in the last six months measured every six months from birth to age four. Previous analyses evaluated potential spatial clusters with three failure time outcomes: time to doctor diagnosed asthma or censoring, time to allergic rhinitis/hayfever or censoring, and time to eczema or censoring (Cook *et al.*, 2007). For the two outcomes, asthma and allergic rhinitis/hayfever, a significant cluster was found, but in very different neighborhoods. Wheeze is a time-varying symptom. Factors influencing wheeze and its resolution or persistence vary with age. Thus, the influence of a single geographical residence may vary with age.

A further innovation of the proposed method is the ability to incorporate study participant relocation during the study. The Home Allergens and Asthma Study has still surveyed and conducted home visits of study participants who have moved, even outside of the predefined study area. It is crucial to include this information for analysis to reduce missingness in the analysis and potential bias. For the Home Allergens and Asthma Study we will analyze the data using the following three different spatial locations: (i) location at birth, (ii) location at age of repeated measure, and (iii) weighted cumulative location history at age of repeated measure.

The outline of this manuscript begins by presenting in Section 2 a new method for spatial cluster detection for repeated measured data. We then conduct a simulation study to assess type I error and power for numerous situations in Section 3. In Section 4 the results from the analysis of the Home Allergens and Asthma Study with outcome repeated wheeze is presented. We conclude with a general discussion in Section 5.

2 Using Cumulative Residuals to Detect Clusters

We propose a new method for spatial cluster detection of repeated measured outcomes using cumulative geographic residuals, which are correlated, and generalized estimating equations (GEE). Previous cluster detection methods using cumulative geographic residuals have been developed for failure time outcomes (Cook *et al.*, 2007).

2.1 Theory of cumulative residuals for repeated measured data

We exemplify the development of our test statistic in the framework of a binary repeated outcome, though the formulation may be easily generalized to any continuous/discrete data with proper link functions (*e.g.* Poisson data with a log link function). Suppose the outcome for individual $i\,(i = 1, \ldots, n)$, at occasion $k\,(k = 1, \ldots, K_i)$, Y_{ik}, is binary with a $p \times 1$ vector of covariates, $\mathbf{X}_{ik}$, and geographic coordinate (r_{ik}, t_{ik}). Under the assumption that disease

status is independent of geographic location (*i.e.* no spatial clusters), the marginal expectation of Y_{ik} given covariates, $\mathbf{X}_{ik}$, is $E(Y_{ik}|\mathbf{X}_{ik}) = \mu_{ik}$, where μ_{ik} is linked to $\mathbf{X}_{ik}$ through a logit link function,

$$g(\mu_{ik}) = \text{logit}(\mu_{ik}) = \boldsymbol{\beta}\mathbf{X}_{ik}, \tag{1}$$

and $\boldsymbol{\beta}$ is a $1 \times p$ vector of regression parameters. The corresponding marginal variance, dependent on μ_{ik}, is $\text{Var}(Y_{ik}) = \mu_{ik}(1 - \mu_{ik})$.

Then define $\mathbf{R}_i(\boldsymbol{\alpha})$ as the 'working' correlation matrix for the $(K_i \times 1)$ response vector for individual i, $\mathbf{Y}_i = (Y_{i1}, ..., Y_{iK_i})^{\mathrm{T}}$. The matrix $\mathbf{R}_i$ may depend on unknown parameters $\boldsymbol{\alpha}$ that will need to be estimated. Define $\mathbf{A}_i = \text{diag}\{\text{Var}(Y_{i1}), \ldots, \text{Var}(Y_{iK_i})\}$. Therefore, to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ utilizing GEE theory one solves the following GEE:

$$\sum_{i=1}^{n} \mathbf{D}_i^{\mathrm{T}} \mathbf{V}_i^{-1} \boldsymbol{\varepsilon}_i = \mathbf{0}, \tag{2}$$

where $\boldsymbol{\varepsilon}_i = \mathbf{Y}_i - \boldsymbol{\mu}_i$, for $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iK_i})^{\mathrm{T}}$, $\mathbf{V}_i = \mathbf{A}_i^{1/2} \mathbf{R}_i(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2}$, and $\mathbf{D}_i = \{\partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}_j; j = 1, \ldots, p\}$.

Under mild regularity conditions and under the null hypothesis of no disease clusters, $\hat{\boldsymbol{\beta}}$ has been shown to be consistent and asymptotically normal with covariance matrix

$$\left(\sum_{i=1}^{n} \hat{\mathbf{D}}_i^{\mathrm{T}} \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i \right)^{-1} \sum_{i=1}^{n} \hat{\mathbf{D}}_i^{\mathrm{T}} \hat{\mathbf{V}}_i^{-1} \mathbf{e}_i \mathbf{e}_i^{\mathrm{T}} \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i \left(\sum_{i=1}^{n} \hat{\mathbf{D}}_i^{\mathrm{T}} \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i \right)^{-1}$$

even if $\mathbf{R}_i(\boldsymbol{\alpha})$ is misspecified, where $\mathbf{e}_i = \mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i$, $\hat{\mu}_{ik} = g^{-1}(\hat{\boldsymbol{\beta}}\mathbf{X}_{ik})$ and $\hat{\mathbf{D}}_i$ and $\hat{\mathbf{V}}_i$ are obtained by replacing unknown parameters in $\mathbf{D}_i$ and $\mathbf{V}_i$ with their sample estimators by solving (2) (Liang and Zeger, 1986).

These asymptotic results have formed the basis for checking whether the link function is correctly specified for a particular component of the covariate vector, such as, $\mathbf{X}_j$, or several components of the covariate vector, $\mathbf{X}_{q\times 1}$ with $1 \le q \le p$ (Su and Wei, 1991; Stute, 1997; Lin, Wei, and Ying., 2002). The crux of this method lies in detecting whether there are significant patterns in the residuals, e_{ik}, related to the particular covariates of interest. Patternless residuals often correspond to 'correct' model specifications (Lin *et al.*, 2002).

However, in our particular setting for spatial cluster detection we study patterns of residuals from a different perspective: instead of viewing the patterns dependent on covariates, we study whether such patterns vary by geographic locations. Presented patterns across regions may indicate excessive, or exiguous, numbers of cases within those areas. In the next section, we propose a use of cumulative residuals for cluster detection.

2.2 Cumulative geographic residuals

We begin by defining our cumulative geographic residuals, $W_{\text{loc}}(x_1, x_2|b)$, as a stochastic process indexed by (x_1, x_2) for a fixed radius b, which takes the form,

$$W_{\text{loc}}(x_1, x_2|b) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{I}(\mathbf{r}_i, \mathbf{s}_i|b, x_1, x_2)^{\mathrm{T}} \mathbf{e}_i \tag{3}$$

where $\mathbf{I}(\mathbf{r}_i, \mathbf{s}_i|b, x_1, x_2)$ is a $K_i \times 1$ vector with each row corresponding to the indicator variable, $I(x_1 - b \le r_{ik} < x_1 + b, x_2 - b < s_{ik} < x_2 + b)$, with (r_{ik}, s_{ik}) denoting the geographic location of subject i at repeated measured location $k\,(k = 1, \ldots, K_i)$, $\mathbf{e}_i = \mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i$ is a $K_i \times 1$ vector of residuals for subject i, and $2b$ is the edge length of the potential square cluster. To study the asymptotic

behavior of $W_{\text{loc}}(\cdot,\cdot|b)$ we define the stochastic process,

$$\hat{W}_{\text{loc}}(x_1,x_2|b)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\mathbf{I}(\mathbf{r}_i,\mathbf{s}_i|b,x_1,x_2)^{\text{T}}\mathbf{e}_i.+\nu^{\text{T}}(x_1,x_2,b|\hat{\boldsymbol{\beta}})\left(\sum_{j=1}^{n}\hat{\mathbf{D}}_j^{\text{T}}\hat{\mathbf{V}}_j^{-1}\hat{\mathbf{D}}_j\right)^{-1}\hat{\mathbf{D}}_i^{\text{T}}\hat{\mathbf{V}}_i^{-1}\mathbf{e}_i\right\}G_i,$$

where

$$\nu(x_1,x_2,b|\hat{\boldsymbol{\beta}})=-\sum_{i=1}^{n}\sum_{k=1}^{K_i}\left[I[(r_{ik}-x_1)^2+(s_{ik}-x_2)^2\le b^2]\mathbf{X}_{ik}\frac{\exp(\boldsymbol{\beta}\mathbf{X}_{ik})}{[1+\exp(\boldsymbol{\beta}\mathbf{X}_{ik})]^2}\right],$$

and $(G_1,\ldots,G_n)$ are independent variables from a unspecified distribution with mean 0 and variance 1 that are independent of $(Y_{ik},\mathbf{X}_{ik},r_{ik},s_{ik})$. The choice of specific distributional forms of G_i will be discussed in Section 3. Assuming that the logit link function, Equation (1), is correctly specified (*e.g.* proper adjustment has been made for known covariates) and under the null hypothesis that the geographic location is independent of outcome, it can be shown that the conditional distribution, $\hat{W}_{\text{loc}}(\cdot,\cdot|b)$ given the data $(\mathbf{Y}_i,\mathbf{X}_i,\mathbf{r}_i,\mathbf{s}_i)$, has the same limit distribution as the unconditional distribution of $W_{\text{loc}}(\cdot,\cdot|b)$. Details of the proof are provided in Appendix A. In summary it is a special case of the covariate-dependent cumulative residuals method discussed by Lin *et al.*, 2002, and follows due to the independence between residuals, $\mathbf{e}_i$, and geographic location, $(\mathbf{r}_i,\mathbf{s}_i)$, under the null. Further, by the continuous mapping theorem, $S_{\text{loc},b}=\sup_{x_1,x_2}W_{\text{loc}}(x_1,x_2|b)$ and $\hat{S}_{\text{loc},b}=\sup_{x_1,x_2}\hat{W}_{\text{loc}}(x_1,x_2|b)$ have the same limiting distribution.

Therefore, to approximate the null distribution of $W_{\text{loc}}(\cdot,\cdot|b)$, one can simulate N realized paths of $\hat{W}_{\text{loc}}(\cdot,\cdot|b)$, *e.g.* $(\hat{W}_{1,\text{loc}}(\cdot,\cdot|b),\ldots,\hat{W}_{N,\text{loc}}(\cdot,\cdot|b))$, by repeatedly simulating $(G_1,\ldots,G_n)$, while fixing the data $(\mathbf{Y}_i,\mathbf{X}_i,\mathbf{r}_i,\mathbf{t}_i)(i=1,\ldots,n)$ at their observed values.

To test for global spatial clusters of half edge length, b, would be to compute the probability of how extreme $S_{\text{loc},b}$ is under the null distribution that the residuals are not dependent on location. Formally testing this hypothesis would require calculating,

$$\hat{S}_{j,\text{loc},b}=\sup_{x_1,x_2}\hat{W}_{j,\text{loc}}(x_1,x_2|b),$$

for $j=1,\ldots,N$ simulated $\hat{W}_{\text{loc}}(x_1,x_2|b)$ and estimating the probability, under the null hypothesis that a simulated $\hat{S}_{\text{loc},b}$ is equal to or more extreme than the observed $S_{\text{loc},b}$, by the p-value $P=\sum_{j=1}^{N}I[S_{\text{loc},b}\le\hat{S}_{j,\text{loc},b}]/N$.

However, for spatial cluster detection it is particularly important to be able to range the values of b to allow the data to depict maximum cluster size. To extend the method to incorporate a finite range of edge lengths, define $\mathbf{b}=(b_1,\ldots,b_L)$ as a finite vector of potential b of length L. Since for a fixed b_l, $\hat{S}_{\text{loc},b_l}$, conditional on the data, converges weakly to the same limiting distribution as S_{loc,b_l}, Skorokhod's representation theorem implies that

$$\hat{S}_{\text{loc}}=\sup(\hat{S}_{\text{loc},b_1},\ldots,\hat{S}_{\text{loc},b_L})$$

converges weakly to the same limiting distribution as

$$S_{\text{loc}}=\sup(S_{\text{loc},b_1},\ldots,S_{\text{loc},b_L}).$$

To test for global clusters using a finite range of half edge length, $\mathbf{b}$, would be to compute the p-value $P=\sum_{j=1}^{N}I[S_{\text{loc}}\le\hat{S}_{j,\text{loc}}]/N$.

This hypothesis test can still easily be inverted to form confidence bands around the stochastic process $W_{\text{loc}}(x_1,x_2|b)$ to define values of (x_1,x_2) and b which have significantly higher cumulative residuals than expected under the null hypothesis of geographic independence. Explicitly, we can form the confidence band $\{(x_1,x_2,b):W_{\text{loc}}(x_1,x_2|b)\ge\hat{S}_{(0.95N)}\}$ where $\hat{S}_{(0.95N)}$ is the 95-th percentile of all $\hat{S}_{j,\text{loc}}$.

By using cumulative geographic residuals, one would be able to locate significant clusters with corresponding edge length, $2b$. Another advantage is the fact that location is not treated as

fixed for an individual, but can change at each repeated time point. Therefore this method incorporates moving, which previous spatial cluster detection methods do not. So far the incorporation of moving has not taken into account moving history, but only current location of the individual. The handling of moving can be made even more flexible by incorporating a weighting structure on an individual location to handle moving history. Specifically, define the test statistic as

$$W_{\mathrm{loc},h}(x_1, x_2|b) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{H}(\mathbf{r}_i^h, \mathbf{s}_i^h, \mathbf{w}_i^h|b, x_1, x_2)^{\mathrm{T}} \mathbf{e}_i, \tag{4}$$

where $\mathbf{H}(\mathbf{r}_i^h, \mathbf{s}_i^h, \mathbf{w}_i^h|b, x_1, x_2)$ is a $K_i \times 1$ vector with each row corresponding to a weight defined as $H_{ik}(\mathbf{r}_i^h, \mathbf{s}_i^h, \mathbf{w}_i^h|b, x_1, x_2) = \sum_{m=1}^{M_i} I(x_1 - b \le r_{im}^h < x_1 + b, x_2 - b \le s_{im}^h < x_2 + b) w_{ikm} \in [0, 1]$, $(\mathbf{r}_i^h, \mathbf{s}_i^h)$ is a vector of all address locations, M_i, in which individual i has resided $(r_{im}^h, s_{im}^h)(m = 1, \ldots, M_i)$, and $w_{ikm} \in [0, 1]$ is the fixed weight assumed for address m of individual i at repeated measure k with the condition that $\sum_{m=1}^{M_i} w_{ikm} = 1$ for all i and k. For example one could define the weights, w_{ikm}, as the proportion of time individual i resided at address m at time t_k. $W_{\mathrm{loc}}(\cdot, \cdot|b)$ is a special case of $W_{\mathrm{loc},h}(\cdot, \cdot|b)$ if one defined weights as 1 if individual i's current residence at time t_k is address m and 0 otherwise or some similar 0 or 1 weighting structure. Distribution of the test statistic, $W_{\mathrm{loc},h}(\cdot, \cdot|b)$, under the null hypothesis would follow the same lines as $W_{\mathrm{loc}}(\cdot, \cdot|b)$, except $\mathbf{I}(\mathbf{r}_i, \mathbf{s}_i|b, x_1, x_2)$ would be replaced by $\mathbf{H}(\mathbf{r}_i^h, \mathbf{s}_i^h, \mathbf{w}_i^h|b, x_1, x_2)$. Under the null hypothesis, the residuals and the weighted location vector are still independent similar to how the indicator location vector and residuals are independent.

The proposed spatial cluster detection test using cumulative residuals will be able to find exact locations, and size, of significant clusters and simultaneously give a p-value for the global hypothesis test of existence of geographic clusters. It can flexibly handle moving locations by applying different weighting structures by not treating location as fixed and does not require model specification of the spatial surface. Section 3 will study the properties of this approach to check the type I error and power. This approach will be applied to the Home Allergens and Asthma study in Section 4, looking at the outcome repeated wheeze.

However, the above spatial cluster detection method cannot pinpoint the times at which the significant clusters occurred. Also, clusters may occur at different locations at different time points. The previous method would only be valid if one assumes that increased risk of disease from a location is constant over age/time. This scenario is not true for most public health outcomes and, in particular, for the outcome wheeze were protective/hazardous predictors in early age can become hazardous/protective in later ages. Therefore, in the next section we will present an alternative method that can detect the time and location of the clusters.

2.3 Cumulative time-dependent geographic cluster detection

The previous section presented a global test statistic utilizing all of the repeated measured data to detect significant geographic clusters that occur throughout a study. However, often a cluster of outcomes may occur only during a certain time point of a study. For example, in the Home Allergens and Asthma Study one may hypothesize that important early in life geographic exposures are different then later in life exposures and therefore locations of significant clusters may change. To handle this important issue we present the following test statistics for each repeated time point $t = t_1, t_2, \ldots, t_K$,

$$W_{\mathrm{loc},h,t}(x_1, x_2|b) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{H}(\mathbf{r}_i^h, \mathbf{s}_i^h, \mathbf{w}_i^h, \mathbf{t}_i|b, x_1, x_2, t)^{\mathrm{T}} \mathbf{e}_i,$$

where $\mathbf{H}(\mathbf{r}_i^h, \mathbf{s}_i^h, \mathbf{w}_i^h, \mathbf{t}_i | b, x_1, x_2, t)$ is a $K_i \times 1$ vector with each row corresponding to a weight defined as $H_{ik}^t(\mathbf{r}_i^h, \mathbf{s}_i^h, \mathbf{w}_i^h, \mathbf{t}_i | b, x_1, x_2, t) = [\sum_{m=1}^{M_i} I(x_1 - b \le r_{im}^h < x_1 + b, x_2 - b \le s_{im}^h < x_2 + b) w_{ikm}] I(t_{ik} = t) \in [0, 1]$. Therefore, $W_{\mathrm{loc},h,t}$ is only summing over residuals of repeated measures that occurred at time point t. Then we define the following time-dependent stochastic process, $\hat{W}_{\mathrm{loc},h,t}(\cdot, \cdot | b)$, as

$$\hat{W}_{\mathrm{loc},h,t}(x_1, x_2 | b) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \mathbf{H}(\mathbf{r}_i^h, \mathbf{s}_i^h, \mathbf{w}_i^h, \mathbf{t}_i | b, x_1, x_2, t)^{\mathrm{T}} \mathbf{e}_i \cdot + \nu^{\mathrm{T}}(x_1, x_2, b | \hat{\boldsymbol{\beta}}) \left(\sum_{j=1}^{n} \hat{\mathbf{D}}_j^{\mathrm{T}} \hat{\mathbf{V}}_j^{-1} \hat{\mathbf{D}}_j \right)^{-1} \hat{\mathbf{D}}_i^{\mathrm{T}} \hat{\mathbf{V}}_i^{-1} \mathbf{e}_i \right\} G_i,$$

where $\nu(x_1, x_2, b | \boldsymbol{\beta})$, $\mathbf{D}$, and $\mathbf{V}$ are defined as in Section 2.2, but $I[x_1 - b \le r_{ik} < x_1 + b, x_2 - b \le s_{ik} < x_2 + b]$ is replaced by $H_{ik}^t(\mathbf{r}_i^h, \mathbf{s}_i^h, \mathbf{w}_i^h, \mathbf{t}_i | b, x_1, x_2, t)$, and $(G_1, \ldots, G_n)$ are independent mean 0 and variance 1 random variables. The asymptotic equivalency of $W_{\mathrm{loc},h,t}(\cdot, \cdot | b)$ and $\hat{W}_{\mathrm{loc},h,t}(\cdot, \cdot | b)$ under the null of geographic independence and correct model specification follows as did for the case for cumulative geographic residuals without time dependence. The benefit of using the repeated measured analysis instead of a logistic model for each time point is the reduction of variance for estimating the relationships of covariates, $\mathbf{X}_{ik}$, to outcome Y_{ik}, by using all of the repeated measured information.

To make conclusions for each time point, t, one would approximate the null distribution of $W_{\mathrm{loc},h,t}(\cdot, \cdot | b_l)$ $(l = 1, \ldots, L)$ by simulating N realizations of $\hat{W}_{\mathrm{loc},h,t}(\cdot, \cdot | b_l)$ for a finite range of half edge lengths, $\mathbf{b}$, and then taking the suprema over $(x_1, x_2, \mathbf{b})$ for the observed and simulated distributions as discussed in Section 2.2. Corresponding p-values and $(1 - \alpha)$ confidence bands can be formed for each $W_{\mathrm{loc},h,t}(\cdot, \cdot | b)$. Therefore, we would find significant cluster locations for all repeated time points t. There is a slight multiple comparison problem due to the fact that we are separately calculating K, the number of repeated measured, hypothesis tests. To be conservative one may want to use Bonferroni critical values, α/K, instead of α. We chose not to do this for our analysis since it would be overly conservative and the objective of the analysis is for exploratory purposes and not to make definitive conclusions. This method is applied on the Home Allergens and Asthma Study in Section 4.

3 Simulation Study

We conducted simulations calculating the type I error and power for the global cumulative geographic residual test. First, we will analyze the results for assessing type I error. Simulations were conducted by generating 1000 test studies where location of an individual was randomly assigned uniformly over an 8×8 grid. For our simulations we chose to treat locations of individuals as fixed over time and a finite range for b of 0.5 to 4 sequenced by 0.5, just to reduce computational complexity. We simulated a repeated measured data set with exchangeable correlation structure and overall probability of having the outcome to be approximately 0.2. The details of this simulation are presented in Appendix B.

By choosing this simulation setup the outcomes for the same individual are correlated and there is an effect of time. When running the simulation we assumed a profile analysis for the mean structure on time and an exchangeable correlation structure. The results for the type I error calculations are given in Table 1. We defined Type I error as the proportion of simulations that detect a significant ($\alpha = 0.05$) cluster. The type I error converges to the α-level of 0.05 when the number of individuals and repeated measures increase. However, it is very low when there are only 100 individuals in the study.

For the power calculations we simulated the repeated measured study population as described for the type I error. To create a single cluster we first considered an 8×8 unit-less area and divided the area into 16 equally sized squares of size 2×2 as depicted in Fig. 1. To create the cluster in consecutive grid areas 6 and 10, we gave a higher probability for individuals with more cases to be within the cluster area. First, define $S_{Yi} = \sum_{k=1}^{K} Y_{ik}$ where K is the number of repeated measures

Table 1 Type I error and power calculations of Cumulative Geographic Residual Test for different sample sizes and number of repeated measured.

		Number of time points		
	N	3	4	5
Type I error	100	0.034	0.023	0.041
	300	0.026	0.039	0.049
	500	0.024	0.040	0.051
Power	100	0.189	0.343	0.527
	300	0.675	0.871	0.977
	500	0.914	0.984	0.999
Sensitivity of highest cluster	100	0.055	0.123	0.222
	300	0.187	0.294	0.394
	500	0.246	0.340	0.409
Sensitivity of all significant clusters	100	0.045	0.098	0.169
	300	0.128	0.182	0.213
	500	0.140	0.167	0.189
Specificity of highest cluster	100	0.927	0.921	0.922
	300	0.969	0.979	0.988
	500	0.990	0.994	0.996
Specificity of all significant clusters	100	0.928	0.924	0.929
	300	0.972	0.985	0.996
	500	0.993	0.998	0.999

Type I error $= (1/1000)\sum_{j=1}^{1000} I(p\text{value}_j \leq 0.05)$. Power $= (1/1000)\sum_{j=1}^{1000} I(\{\text{CR}_j \cap \text{Grids 6 or 10}\} \neq \emptyset)$. Sensitivity $= (1/1000)\sum_{j=1}^{1000} (1/\sum_{i=1}^{N} I((x_i, y_i) \in CR_j))\sum_{i=1}^{N} I((x_i, y_i) \in \text{Grids 6 or 10}, (x_i, y_i) \in \text{CR}_j))$. Specificity $= (1/1000)\sum_{j=1}^{1000} (1/\sum_{i=1}^{N} I((x_i, y_i) \notin \text{CR}_j))\sum_{i=1}^{N} I((x_i, y_i) \notin \text{Grids 6 or 10}, (x_i, y_i) \notin \text{CR}_j))$. CR_j is the region which was detected as a spatial cluster and is $\emptyset$ if no spatial clusters were detected.

and A_i as a random sample from a Bernoulli distribution with probability $(0.15S_{Yi})$. If $A_i = 1$ then (x_i, y_i) is randomly drawn from a uniform distribution within grids 6 and 10 and if $A_i = 0$ then (x_i, y_i) is randomly drawn from a uniform distribution from the entire 8×8 study area. We defined power as the proportion of simulations that detect at least one significant ($\alpha = 0.05$) cluster and which at least one of the significant clusters detected overlaps with grids 6 or 10. We define sensitivity for a given simulation as the proportion of individuals included in a significant cluster that reside in grids 6 or 10 out of all individuals included in the significant cluster. Sensitivity is 0 if no clusters were significant. Overall sensitivity is the mean sensitivity for all simulations. We define specificity as the mean proportion of individuals not included in the significant cluster that reside outside of grids 6 or 10 out of all individuals not included in the significant cluster across all simulations. For both sensitivity and specificity, we did a calculation for the highest significant cluster and then for all spatial clusters detected with a p-value less than 0.05. Power calculations are displayed in Table 1 for simulated data sets of size 1000.

Overall, the proposed cumulative geographic residual test statistic for repeated measures holds the type I error rate and has relatively high power of finding clusters. The power, sensitivity, and specificity increase as expected given more individuals and repeated measures in the study. The sensitivity is relatively low indicating that the spatial cluster, or clusters, detected tend to be larger

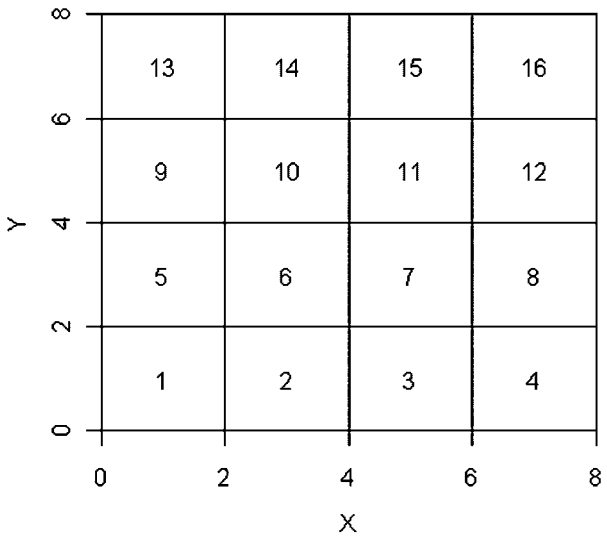

Figure 1 Grid system of study area for power simulation data sets.

than the actual cluster. The specificity is very high indicating whether the detected cluster does not include a given area, that area with high probability is not actually a cluster.

The next set of simulations evaluated the effect of covariate adjustment on cluster detection. In particular we are evaluating the assumption that the proposed spatial cluster detection method is valid even when there is dependence between the covariate and spatial location. We assumed 4 repeated measures and 300 observations for this simulation study. We first simulated the outcome data following the procedure described for type I error. For simplicity, we will assume only one covariate and that the covariate stays constant over time ($X_{ik} = X_i$). To create dependence between the covariate and outcome we simulate $X_i \sim \text{Bernoulli}(0.10 + \nu S_{Yi})$ where ν is a parameter depicting the degree of dependence between covariate and outcome. Then to create dependence between location and both the covariate and outcome we simulated $A_i \sim \text{Bernoulli}(cS_{Yi} + \gamma X_i)$ where γ and c are parameters depicting the degree of dependence between covariate and cluster and outcome and cluster, respectively. Given A_i we simulated (x_i, y_i) as described for the power calculation. We ran simulations varying γ to be 0, 0.1, and 0.2, ν to be sequenced from 0.05 to 0.20 by 0.05, and c to be 0 (no spatial cluster) and 0.15.

Table 2 displays the results from the simulation study evaluating the influence of covariate adjustment. When there was no actual spatial cluster ($c = 0$) that existed independent of the relationships between the covariate, X_i, and outcome, Y_{ij}, and the covariate and spatial location, A_i, the type I error was held at less than 0.05 when the cumulative geographic residual method was adjusted for X_i. However, when not adjusting for X_i as the dependence between X_i and Y_{ij} increases (ν increases) and X_i and A_i increases (γ increases) the power increases as would be expected. For the simulation when there is a spatial cluster ($c = 0.15$) when not adjusting for X_i the power increases as both γ and ν increase, but when adjusting for X_i the power does not remain constant, but decreases as the dependence between X_i and Y_{ij} increases (ν increases) and dependence between X_i and A_i increases (γ increases). A potential reason why the power decreases when adjusting for X_i, instead of remaining relatively constant, may be due to Y_{ij} being a binary outcome and therefore when simulating Y_{ij} dependent on X_i affects both the mean and variance of $Y_{ij}|X_i$. In a simulation study not shown for continuous outcome data, when there is no direct mean and variance relationship, the power remained constant when adjusting for the covariate. Therefore, this observation may not be due to the cluster detection method, but due to the nature of the outcome.

Table 2 Evaluation of the effect of covariate adjustment on the properties of the cumulative geographic residual test for different degrees of dependence between outcome and covariate and between covariate and location.

	ν	No cluster ($c = 0$)		Cluster ($c = 0.15$)	
		Unadjusted power	Adjusted power	Unadjusted power	Adjusted power
Independence ($\gamma = 0$)	0.05	0.034	0.040	0.885	0.867
	0.10	0.023	0.025	0.884	0.840
	0.15	0.038	0.035	0.872	0.797
	0.20	0.031	0.025	0.871	0.741
Moderate dependence ($\gamma = 0.10$)	0.05	0.037	0.038	0.902	0.862
	0.10	0.037	0.033	0.914	0.819
	0.15	0.044	0.030	0.937	0.784
	0.20	0.051	0.020	0.936	0.706
Strong dependence ($\gamma = 0.20$)	0.05	0.045	0.035	0.921	0.870
	0.10	0.062	0.033	0.945	0.817
	0.15	0.079	0.026	0.978	0.793
	0.20	0.102	0.032	0.983	0.699

$\text{Power} = \frac{1}{1000}\sum_{j=1}^{1000} I(p\text{value}_j \leq 0.05)$

.

Overall in this simulation section we have shown that the proposed cumulative geographic residual method holds the appropriate type I error with and without adjusting for covariates and power follows expected patterns as it increases with increased sample size, increased number of repeated measures, and increased dependence between outcome and spatial location. In the next section, the proposed cumulative geographic residual method will be applied to the repeated measured outcome wheeze.

4 Home Allergens and Asthma Study Analysis

We now apply the proposed method to the Home Allergens and Asthma prospective cohort study with the longitudinal outcome wheeze in the last 6 months. The study was designed to investigate potential environmental exposures and their relationship to childhood asthma and other respiratory outcomes. A total of 499 study participants were enrolled in the study after being born at Brigham and Women's hospital in Boston, MA USA between September 1994 to June 1996. Details of the study design have been previously published by Celedon *et al.* (1999). Of those 499 study participants, only 478 were used for this analysis due to the inability to geocode the missing participants' birth addresses. The investigators for this analysis were interested in areas of significant disease clusters for a range of outcomes. For this analysis we will study the clusters of the outcome repeated wheeze in the first four years of life. Therefore, the repeated measures will be observed at ages 6, 12, 18, 24, 30, 36, 42, and 48 months.

The study area is a diverse population with a range of SES levels. Figure 2 displays the median family income level in the study population. Previous analysis on the mothers of the infants screened for the study had found elevated IgE levels, an indicator of allergic response, in southern Boston, Chelsea, and Revere areas (Litonjua *et al.*, 2007). These areas also correspond to lower median family areas indicating a relationship between disparity and allergic reaction.

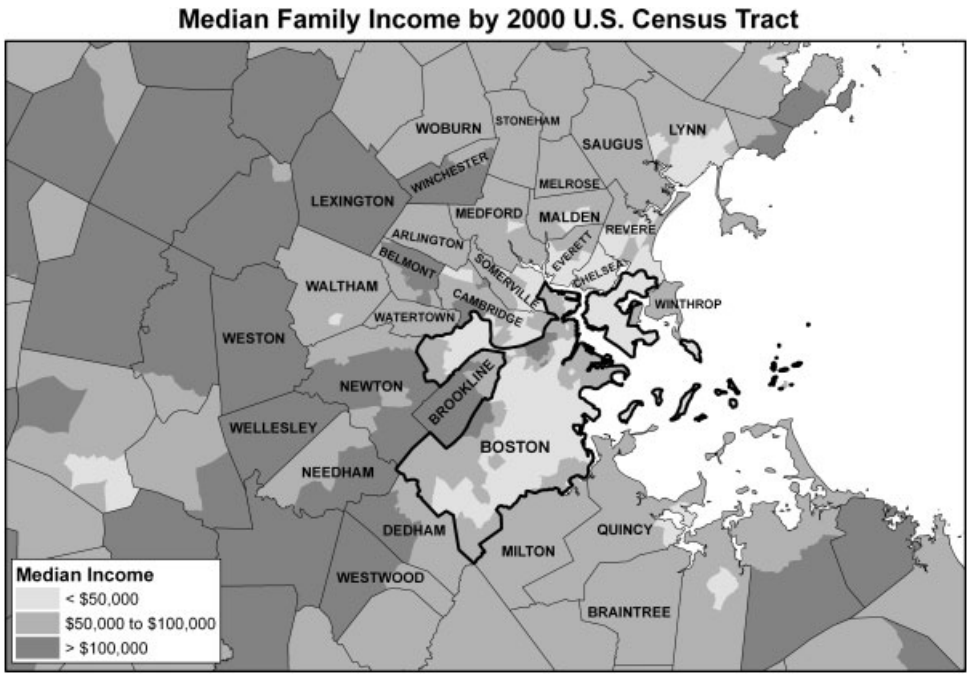

Figure 2 Indicated areas of low, medium, and high median family income by US census tract in the study population area.

A spatial cluster detection analysis on the children up to age four in this study using censored outcomes asthma, allergic rhinitis/hayfever, and eczema found a significant cluster of the censored outcome asthma in southern Boston, Chelsea, Revere, and their neighboring towns, but for the censored outcome allergic rhinitis/hayfever the significant cluster resided in the western, more affluent, towns (Cook *et al*., 2007). It is of interest to display significant disease clusters for the outcome wheeze since it may be less vulnerable to underdiagnosis in lower SES areas compared with the previous outcomes, particularly hayfever (Strunk, Ford, and Taggart, 2000). We hypothesize that a cluster will exist in the southern less affluent Boston area early in life, similar to the asthma cluster found in Cook *et al*. (2007), since the area has higher IgE levels (Litonjua *et al*., 2005) and lower median family income (Fig. 2) and location of the cluster will change over time. One reason for the change in location over time could be due to the differential drop-out within lower SES and minority areas and therefore over time the cluster may move to more affluent areas. To infer whether the cluster location movement over time is due to exposure change, or loss to follow-up, we ran the analysis using all of the data (full data) and then checked for comparable results using only the observations of study participants with complete follow-up up to age four (complete follow-up). Note that we will present results only for the full data set since the complete data set results did not change results substantially, except *p*-values were higher since we have fewer subjects in the complete data set (Table 3).

First, we ran a GEE model without considering spatial clusters to assess change in percent wheeze by age. Owing to the exploratory nature of all analyses in this manuscript, we did not adjust for other predictors except age. We used a profile mean model on age and an unstructured correlation structure with robust standard errors from the sandwich estimator. Table 3 summarizes the results for two analyses that ran the GEE model for the full data set and the complete follow-up subset. Note that estimates, and corresponding 95% confidence intervals, do not change significantly depending on the full data set *versus* complete data set indicating missingness may be

Table 3 Estimated probability wheeze per time period for all study participants and subset with complete follow-up.

AGE	Full data		Complete follow-up	
	N	π (95% CI)	*N*	π (95% CI)
6 Mos	494	0.22 (0.18, 0.26)	414	0.21 (0.18, 0.25)
12 Mos	494	0.27 (0.23, 0.31)	414	0.26 (0.22, 0.30)
18 Mos	486	0.20 (0.17, 0.24)	414	0.19 (0.16, 0.23)
24 Mos	487	0.21 (0.17, 0.24)	414	0.21 (0.17, 0.25)
30 Mos	471	0.12 (0.10, 0.16)	414	0.12 (0.09, 0.16)
36 Mos	462	0.10 (0.08, 0.13)	414	0.10 (0.07, 0.13)
42 Mos	455	0.12 (0.09, 0.15)	414	0.11 (0.08, 0.15)
48 Mos	460	0.11 (0.09, 0.15)	414	0.12 (0.09, 0.16)

Table 4 Summary of results from the cumulative geographic residual test by time point and location definition indicating probability of wheeze within and outside cluster at each time point.

	Birth				Current				Weighted			
	$N^{a)}$	p_{in}	p_{out}	S_{loc} (*p*-value)	$N^{a)}$	p_{in}	p_{out}	S_{loc} (*p*-value)	$N^{a)}$	p_{in}	p_{out}	S_{loc} (*p*-value)
6 Mos	147	0.32	0.17	3.9 (0.04)	147	0.32	0.17	3.9 (0.04)	147	0.32	0.17	3.9 (0.04)
12 Mos	46	0.57	0.25	3.0 (0.13)	55	0.51	0.25	2.9 (0.23)	53	0.52	0.25	2.9 (0.24)
18 Mos	104	0.31	0.18	2.9 (0.38)	198	0.27	0.17	3.2 (0.16)	100.4	0.31	0.18	2.9 (0.34)
24 Mos	123	0.28	0.18	1.8 (0.56)	111	0.30	0.18	2.7 (0.41)	159.1	0.27	0.18	2.7 (0.40)
30 Mos	102	0.22	0.10	3.7 (0.28)	88	0.23	0.10	3.6 (0.26)	95.7	0.22	0.10	3.7 (0.16)
36 Mos	72	0.18	0.09	2.9 (0.75)	157	0.14	0.08	3.1 (0.53)	109.4	0.15	0.09	2.5 (0.78)
42 Mos	64	0.23	0.10	3.2 (0.44)	67	0.22	0.10	3.1 (0.45)	54.9	0.26	0.10	3.4 (0.35)
48 Mos	111	0.16	0.10	2.2 (0.94)	16	0.38	0.11	1.8 (0.90)	14.7	0.38	0.11	1.7 (0.97)

a) *N* denotes the total number, or weighted number, of addresses that reside in the potential spatial cluster.

missing completely at random as assumed by the GEE. Overall, there is a definite change in probability of wheeze over time indicated by a significant drop in wheeze rates after age 30 months.

It is of interest to assess if the environmental exposure is influenced by earlier months, *i.e.* the birth location, current location, or complete location history is the important predictor. To answer this question we ran three analyses (i) keeping location constant as birth location, (ii) location as the current location at evaluation, and (iii) a weighted history of location with weights determined as length of time resided in a particular location.

Results for the three analyses are reported in Table 4 and Figs. 3–5. The only significant (α-level = 0.05) spatial cluster detected was at 6 months in the urban coastal Boston area. This is in a similar area in which the censored outcome time to asthma and time to eczema found significant clusters. The spatial clusters moved over time, starting in the urban coastal Boston area, and slowly moving toward the southern, more suburban, study area. The movement of clusters is not statistically significant since there are no significant clusters found for any time points after 6 months. However, this could be due to power issues since the prevalence of wheeze decreases over time.

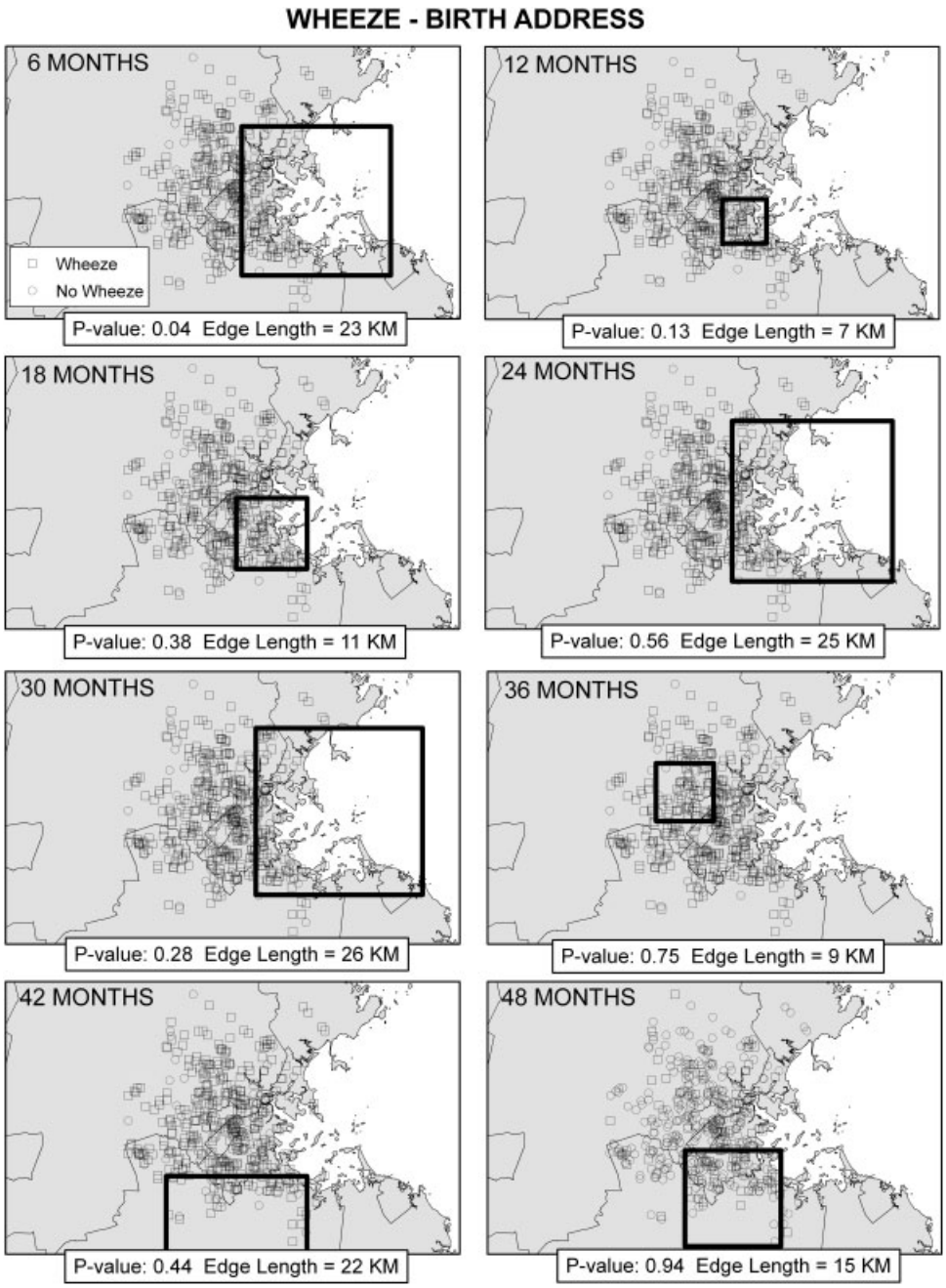

Figure 3 Spatial cluster detection every 6 months using birth place address.

It is interesting to note that the performance of the cumulative spatial cluster detection is dependent on the size of cluster detected and strength of cluster detected. The 6 months cluster comprises 31%(171/478) of the study population and was in an area with 32% prevalence of wheeze in the first 6 months compared with 17% prevalence in the rest of the study population (Table 3). At 12 months, the cluster detected,

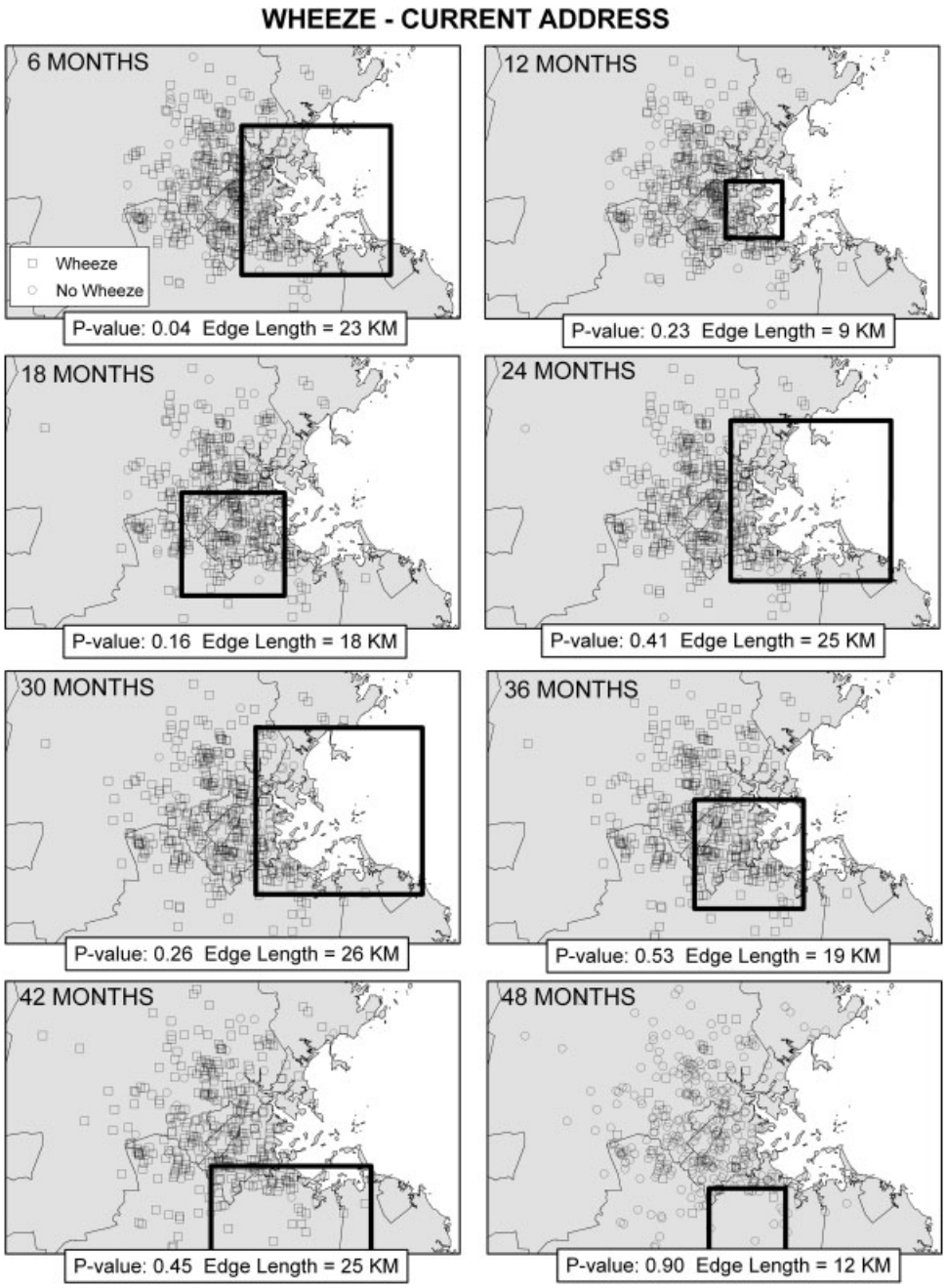

Figure 4 Spatial cluster detection every 6 months using current address.

assuming birth address, comprises of only 10% of the study population, but had a much larger difference in prevalence of wheeze of 57% compared with 25% even though the *p*-value was 0.13. There is a trade off between prevalence, effect size, and size of cluster that needs to be further explored to assess the performance of this method, but this case example indicates that the cumulative geographic residual method has more power to detect clusters of larger size.

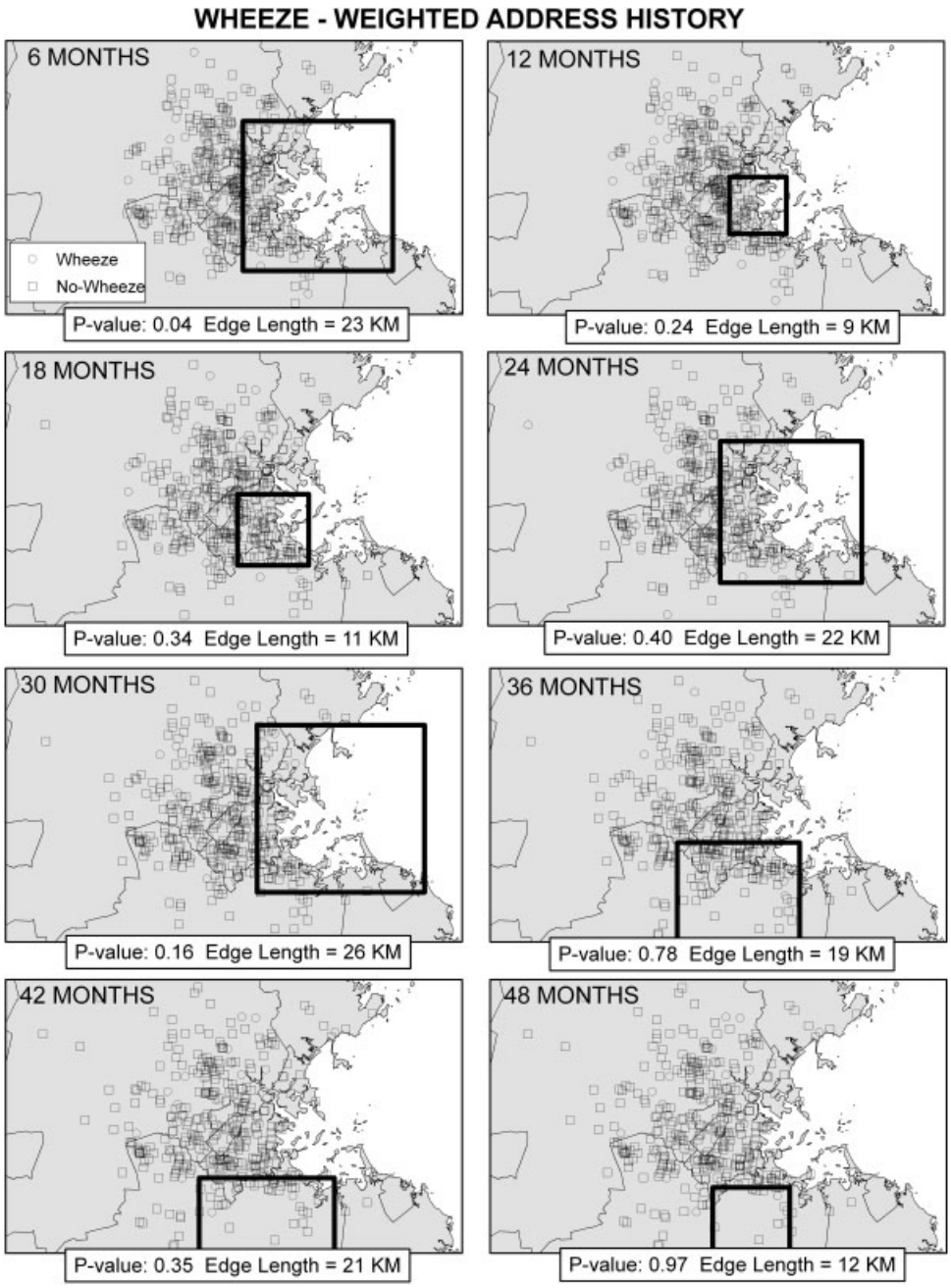

Figure 5 Spatial cluster detection every 6 months using cumulative address history.

5 Discussion

In this manuscript we have proposed a new spatial cluster detection method for repeated measured outcomes utilizing cumulative geographic residuals. Applying the new method, we detected a

significant cluster of wheeze in urban Boston for age 6 months. Further research is being conducted to look into which exposures in urban Boston may be influencing this disease cluster, such as air pollution.

We also performed type I error and power calculations for the cumulative geographic residual method. Type I error was held at the appropriate α-level under the null of no spatial clusters. By increasing the number of individuals, and repeated measures, power was shown to increase substantially. Therefore, the method is performing as expected and is valid for spatial cluster detection of repeated measured outcomes. Future work exploring the properties of the cumulative residual method, particularly how power is effected by cluster size, effect size, and overall incidence rates would be of interest to assess the performance of the method.

The importance of using the time-dependent cumulative geographic residual method was presented as being able to pinpoint the location and time of significant clusters. This method can be used to explore hypotheses and assess changes of outcomes and exposures over time which virtually no previous spatial cluster detection methods have directly been able to assess.

Supplementary Material

Software for the method can be found at `http://faculty.washington.edu/acook` developed in the `R` statistical package (R Development Core Team, 2009).

Acknowledgements Diane R. Gold was supported by NIH RO1 AI/EHS 35786 and Yi Li was supported by NIH RO1 CA95747.

Conflict of Interests Statement
The authors have declared no conflict of interest.

Appendix A

A.1 Asymptotic distribution of $W_{\text{loc}}(x_1, x_2|b)$ and $\hat{W}_{\text{loc}}(x_1, x_2|b)$, given the observed data and independence between $\mathbf{Y}_i|\mathbf{X}_i$ and r_i, s_i, for the cumulative geographic residual

Throughout this proof we assume that $g(.)$ (1) is the correct link function between $\mathbf{Y}_i$ and $\mathbf{X}_i$. We also assume that $\mathbf{Y}_i|\mathbf{X}_i$ and location, (s_i, r_i), are independent. This may be violated when $\mathbf{X}_i$ and (s_i, r_i) are dependent. We further assume that $\mathbf{X}_i, r_i$, and s_i are bounded.

Consider the following one-term Taylor series expansion of $W_{\text{loc}}(x_1, x_2|b)$ at $\boldsymbol{\beta}$:

$$\begin{aligned} W_{\text{loc}}(x_1, x_2|b) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{I}(\mathbf{r}_i, \mathbf{s}_i|b, x_1, x_2)^{\mathrm{T}} \mathbf{e}_i \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{I}(\mathbf{r}_i, \mathbf{s}_i|b, x_1, x_2)^{\mathrm{T}} \boldsymbol{\varepsilon}_i \\ &\quad + \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}(\mathbf{r}_i, \mathbf{s}_i|b, x_1, x_2)^{\mathrm{T}} \mathbf{D}_i \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + o_p(1). \end{aligned} \tag{5}$$

where $\boldsymbol{\varepsilon}_i = \mathbf{Y}_i - g^{-1}(\mathbf{X}_i\boldsymbol{\beta})$, $\mathbf{e}_i = g^{-1}(\mathbf{X}_i\hat{\boldsymbol{\beta}})$, $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, and $\mathbf{D}_i$ are defined as in Section 2.1.

It was shown in Section 2.1 that given the the conditional mean of Y_i, $E(\mathbf{Y}_i|\mathbf{X}_i)$ is correctly linked to $\mathbf{X}_i$ through $g(.)$, the $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges as $n \to \infty$ to a zero-mean Gaussian distribution with covariance matrix,

$$\left(\sum_{i=1}^{n}\mathbf{D}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\mathbf{D}_i\right)^{-1}\sum_{i=1}^{n}\mathbf{D}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\mathbf{D}_i\left(\sum_{i=1}^{n}\mathbf{D}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\mathbf{D}_i\right)^{-1}.$$

This implies that $W_{\mathrm{loc}}(x_1, x_2|b)$ is asymptotically equivalent to

$$\begin{aligned}\tilde{W}_{\mathrm{loc}}(x_1, x_2|b) =& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{I}(\mathbf{r}_i, \mathbf{s}_i|b, x_1, x_2)^{\mathrm{T}}\boldsymbol{\varepsilon}_i \\ &+ \frac{1}{n}\nu^{\mathrm{T}}(x_1, x_2, b|\boldsymbol{\beta})\left(\frac{1}{n}\sum_{j=1}^{n}\mathbf{D}_j^{\mathrm{T}}\mathbf{V}_j^{-1}\mathbf{D}_j\right)^{-1} \\ &\times \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{D}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\boldsymbol{\varepsilon}_i \\ =& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\mathbf{I}(\mathbf{r}_i, \mathbf{s}_i|b, x_1, x_2)^{\mathrm{T}} + \frac{1}{n}\nu^{\mathrm{T}}(x_1, x_2, b|\boldsymbol{\beta})(\mathbf{B})^{-1}\mathbf{D}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\right]\boldsymbol{\varepsilon}_i.\end{aligned}$$

where $\mathbf{B} = n^{-1}\sum_{j=1}^{n}\mathbf{D}_j^{\mathrm{T}}\mathbf{V}_j^{-1}\mathbf{D}_j$.

We will first establish the tightness of $\tilde{W}_{\mathrm{loc}}(x_1, x_2|b)$, which implies the tightness of $W_{\mathrm{loc}}(x_1, x_2 b|\boldsymbol{\beta})$. By the law of large numbers $\mathbf{B}$ converges to a constant matrix as $n \to \infty$. By the uniform law of large numbers $n^{-1}\nu(x_1, x_2, b|\boldsymbol{\beta})$ converges as $n \to \infty$ uniformly in x_1, x_2, and b, to a nonrandom function. Therefore,

$$\frac{1}{\sqrt{n}}\left(\frac{1}{n}\nu^{\mathrm{T}}(x_1, x_2, b|\boldsymbol{\beta})\right)(\mathbf{B})^{-1}\sum_{i=1}^{n}\mathbf{D}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\boldsymbol{\varepsilon}_i$$

converges as $n \to \infty$ to Gaussian process and therefore is tight. Since $\sum_{i=1}^{n}\mathbf{I}(\mathbf{r}_i, \mathbf{s}_i|b, x_1, x_2)^{\mathrm{T}}\boldsymbol{\varepsilon}_i$ is the sum of monotonic step functions and therefore manageable (Pollard, 1998) and by the functional central limit theorem it is tight. Hence, the entire process $\tilde{W}_{\mathrm{loc}}(x_1, x_2|b)$ is tight yielding $W_{\mathrm{loc}}(x_1, x_2|b)$ is tight.

For fixed (x_1, x_2), $\tilde{W}_{\mathrm{loc}}(x_1, x_2|b)$ is a sum of n independent and identically distributed zero-mean random vectors since $E(\boldsymbol{\varepsilon}_i) = 0$. By the multivariate central limit theorem, the finite-dimensional distributions of $\tilde{W}_{\mathrm{loc}}(x_1, x_2|b)$ are asymptotically zero-mean normal, implying the same for $W_{\mathrm{loc}}(x_1, x_2|b)$. This fact, together with thetightness of $W_{\mathrm{loc}}(x_1, x_2|b)$, implies that $W_{\mathrm{loc}}(x_1, x_2|b)$ converges as $n \to \infty$ weakly to a zero-mean Gaussian process with the covariance function between (x_{1a}, x_{2a}) and (x_{1b}, x_{2b}) being

$$\begin{aligned}&\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\{[\mathbf{I}(\mathbf{r}_i, \mathbf{s}_i|b, x_{1a}, x_{2a})^{\mathrm{T}} + \frac{1}{n}\nu^{\mathrm{T}}(x_{1a}, x_{2a}, b|\boldsymbol{\beta})(\mathbf{B})^{-1}\mathbf{D}_i^{\mathrm{T}}\mathbf{V}_i^{-1}]\boldsymbol{\varepsilon}_i\} \\ &\times\{[\mathbf{I}(\mathbf{r}_i, \mathbf{s}_i|b, x_{1b}, x_{2b})^{\mathrm{T}} + \frac{1}{n}\nu^{\mathrm{T}}(x_{1b}, x_{2b}, b|\boldsymbol{\beta})(\mathbf{B})^{-1}\mathbf{D}_i^{\mathrm{T}}\mathbf{V}_i^{-1}]\boldsymbol{\varepsilon}_i\}\end{aligned}$$

Next we will establish the weak convergence distribution of $\hat{W}_{\mathrm{loc}}(x_1, x_2|b)$. Conditional on the data $(Y_{ik}, \mathbf{X}_{ik}, r_{ik}, t_{ik})(i = 1, \ldots, n; k = 1, \ldots, K_i)$, the only random components in $\hat{W}_{\mathrm{loc}}(x_1, x_2|b)$ are $(G_1, \ldots, G_n)$. Thus, it follows from the multivariate central limit theorem that, conditional on the data, the finite-dimensional distributions of $\hat{W}_{\mathrm{loc}}(x_1, x_2|b)$ are asymptotically zero-mean normal as $n \to \infty$. Since $\hat{W}_{\mathrm{loc}}(x_1, x_2|b)$ consists of monotone functions in (x_1, x_2), which are manageable, the functional central limit theorem implies that $\hat{W}_{\mathrm{loc}}(x_1, x_2|b)$ is tight.

The conditional covariance function of $\hat{W}_{\text{loc}}(x_1, x_2|b)$ at $((x_{1a}, x_{2a}), (x_{1b}, x_{2b}))$ is,

$$\frac{1}{n}\sum_{i=1}^{n}\{\mathbf{I}(\mathbf{r}_i, \mathbf{s}_i|b, x_{1a}, x_{2a})^{\mathrm{T}}\mathbf{e}_i + \nu^{\mathrm{T}}(x_{1a}, x_{2a}, b|\hat{\boldsymbol{\beta}})(\hat{\mathbf{B}})^{-1}\hat{\mathbf{D}}_i^{\mathrm{T}}\hat{\mathbf{V}}_i^{-1}\mathbf{e}_i\}$$
$$\times\{\mathbf{I}(\mathbf{r}_i, \mathbf{s}_i|b, x_{1b}, x_{2b})^{\mathrm{T}}\mathbf{e}_i + \nu^{\mathrm{T}}(x_{1b}, x_{2b}, b|\hat{\boldsymbol{\beta}})(\hat{\mathbf{B}})^{-1}\hat{\mathbf{D}}_i^{\mathrm{T}}\hat{\mathbf{V}}_i^{-1}\mathbf{e}_i\}$$

which converges as $n \to \infty$ to the same deterministic limit as the covariance function of $W_{\text{loc}}(x_1, x_2|b)$. Therefore, $W_{\text{loc}}(x_1, x_2|b)$ and $\hat{W}_{\text{loc}}(x_1, x_2|b)$ converge to the same limiting zero-mean Gaussian process as $n \to \infty$.

Appendix B

B.1 Simulated repeated measures data

To simulate the repeated measured data under an exchangeable correlation structure we conducted the following simulation design for the different number of repeated observations holding the overall probability of being a case to be approximately 0.2.

Three repeated measures: Generate n multivariate normal outcome, $\mathbf{Z_i} \sim N_3((-0.1, 0, 0.1)^{\mathrm{T}}, \mathbf{V})$, where $\mathbf{V}$ is a 3×3 matrix with diagonal elements 1 and off diagonal elements $\rho = 0.2$. Then define binary repeated measures outcome, $Y_{ij} = I(Z_{ij} \geq 0.85)$ to use for analyses.

Four repeated measures: Generate n multivariate normal outcome, $\mathbf{Z}_i = N_4((-0.1, -0.05, 0.05, 0.1)^{\mathrm{T}}, \mathbf{V})$, where $\mathbf{V}$ is a 4×4 matrix with diagonal elements 1 and off diagonal elements $\rho = 0.2$. Then define binary repeated measures outcome, $Y_{ij} = I(Z_{ij} \geq 0.85)$ to use for analyses.

Five repeated measures: Generate n multivariate normal outcome, $\mathbf{Z}_i = N_5((-0.1, -0.05, 0, 0.05, 0.1)^{\mathrm{T}}, \mathbf{V})$, where $\mathbf{V}$ is a 5×5 matrix with diagonal elements 1 and off diagonal elements $\rho = 0.2$. Then define binary repeated measures outcome, $Y_{ij} = I(Z_{ij} \geq 0.845)$ to use for analyses.

References

Akinbami, L., Centers for Disease Control and Prevention National Center for Health Statistics (2006). The state of childhood asthma, 1980–2005. *Advance Data* **381**, 1–24.

Brugge, D., Vallarion, J., Ascolillo, L., Osgood, N. D., Steinbach, S. and Spengler, J. (2003). Comparison of multiple environmental factors for asthmatic children in public housing. *Indoor Air* **13**, 18–27.

Celedon, J., Litonjua, A., Weiss, S. and Gold, D. (1999). Day care attendence in the first year of life and illnesses of the upper and lower respiratory tract in children with a familial history of atopy. *Pediatrics* **104**, 495–500.

Cook, A. J., Gold, D. R. and Li, Y. (2007). Spatial cluster detection for censored outcome data. *Biometrics* **63**, 540–549.

Duczmal, L. and Assunção, R. (2004). A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics and Data Analysis* **45**, 269–286.

Finkelstein, J., Fuhlbrigge, A., Lozano, P., Grant, E., Shulruff, R., Arduino, K. and Weiss, K. (2002). Parent-reported environmental exposures and environmental control measures for children with asthma. *Archives of Pediatrics and Adolescent Medicine* **156**, 258–264.

Huang, L., Kulldorff, M. and Gregorio, D. (2007). A spatial scan statistic for survival data. *Biometrics* **63**, 109–118.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics* **26**, 1481–1496.

Kulldorff, M., Huang, L., Pickle, L. and Duczmal L. (2006). An elliptic spatial scan statistic. *Statistics in Medicine* **25**, 3929–3943.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Lin, D. Y., Wei, L. J. and Ying, Z. (2002). Model-checking techniques based on cumulative residuals. *Biometrics* **58**, 1–12.

Litonjua, A. A., Celedón, J. C., Hausmann, B. A., Nikolov, M., Sredl, D., Ryan, L., Platts-Mills, T. A. E., Weiss, S. T. and Gold, D. R. (2005). Variation in total and specific IgE: effects of ethnicity and socioeconomic status. *The Journal of Allergy and Clinical Immunology* **115**, 751–757.

Patil, G. P. and Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics* **11**, 183–197.

Pollard, D. (1998). *Empirical Processes: Theory and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics 2. Institute of Mathematical Sciences, Hayward, CA.

R Development Core Team (2009). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0, `http://www.R-project.org`.

Strunk, R. C., Ford, J. G. and Taggart, V. (2002). Reducing disparities in asthma care: priorities for research – National Heart, Lung, and Blood Institute workshop report. *The Journal of Allergy and Clinical Immunology* **109**, 229–237.

Stute, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics* **25**, 613–641.

Su, J. Q. and Wei, L. J. (1991). A lack-of-fit test for the mean function in a generalized linear model. *Journal of the American Statistical Association* **86**, 420–426.

Tango, T. (2000). A test for spatial disease clustering adjusted for multiple testing. *Statistics in Medicine* **19**, 191–204.

Tango, T. and Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics* **4**, 11.

Turnball, G. W., Iwano, E. J., Burnett, W. S., Howe, H. L. and Clark, L. C. (1990). Monitoring for clusters of disease: Application to leukemia incidence in upstate New York. *American Journal of Epidemiology* **132**, 136–143.

