*Harvard University*
Harvard University Biostatistics Working Paper Series

*Year* 2006 *Paper* 56

# Spatial Cluster Detection for Censored Outcome Data

Andrea J. Cook* Diane Gold†
Yi Li‡

*University of Washington, acook@u.washington.edu

†Brigham & Women's Hospital and Harvard Medical School, diane.gold@channing.harvard.edu

‡Harvard University and Dana Farber Cancer Institute, yili@jimmy.harvard.edu



http://biostats.bepress.com/harvardbiostat/paper56

# Spatial Cluster Detection for Censored Outcome Data

**Andrea J. Cook*,[1,2,†] Diane Gold[‡3] and Yi Li*[4]**
[1]Department of Biostatistics, University of Washington
Seattle, WA 98105, USA
[2]Center for Health Studies, Group Health Cooperative
Seattle, WA 98101, USA
[3]The Channing Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School
Boston, MA 02115, USA
[4]Department of Biostatistics
Harvard School of Public Health and the Dana Farber Cancer Institute
Boston, MA 02115, USA

September 7, 2006

SUMMARY. While numerous methods have been proposed to test for spatial cluster detection, in particular for discrete outcome data (e.g. disease incidence), few have been available for continuous data which are subject to censoring. This paper provides an extension of the spatial scan statistic (Kulldorff, 1997) for censored outcome data and further proposes a simple spatial cluster detection method by utilizing cumulative martingale residuals within the framework of the Cox's proportional hazards models. Simulations have indicated good performance of the proposed methods, with the practical applicability illustrated by an ongoing epidemiology study, which investigates the relationship of environmental exposures with asthma, allergic rhinitis/hayfever, and eczema.

KEY WORDS: Asthma; Cluster Detection; Cumulative Residuals; Martingales; Spatial Scan Statistic.

*Supported by R01 CA 95747
† *email:* acook@u.washington.edu
‡Supported by R01 AI/EHS 35786

## 1. Introduction

The Home Allergens and Asthma study is an ongoing prospective cohort study investigating environmental and socioeconomic risk factors leading to early childhood respiratory diseases, such as asthma and allergic rhinitis (Celedon et al., 1999). Longitudinal and cross-sectional studies have linked measures of lower SES, home allergen levels (e.g., cockroach), mold in the home, and other individual or family-based measures of exposures to increased incidence or prevalence of wheeze, asthma and allergic rhinitis (Brugge et al., 2003 and Finkelstein et al., 1999). Fewer studies focus on the larger area, or neighborhood, in which the individual is situated as a source of environmental exposures that may influence the risk of allergic diseases.

An individual's immune development depends on a complex interaction of factors related to inheritance and environmental exposures that may come from the larger neighborhood/community as well as the individual home. While exposures may have differing effects according to the window within which they occur, it is likely that an individuals immune development is influenced by his/her entire exposure history up to date. Due to this complexity, it is of substantial interest to detect spatial/neighborhood regions that have significantly higher hazard rates of disease, pointing to potential hazardous environmental sources (e.g., poor housing, bus depots, neighborhood waste sites, sources of rodent infestations, neighborhood violence). Indeed, spatial cluster detection has been found as a much useful tool to fulfill these tasks. Further, the main endpoints in the Home Allergens and Asthma study, e.g. times to asthma and other respiratory outcomes, are subject to censoring due to drop out and limited time of follow-up. Hence, the data analysis calls for spatial cluster detection methods that can handle censored outcomes.

We present in this paper two general statistical approaches to quantifying spatial cluster detection for censored outcomes. The first approach, presented in Section 2, extends the

spatial scan statistic developed by Kulldorff (1997) for count and binary data and the second method, in Section 3, considers cumulative martingale residuals in the spirit of Lin et al. (1993). We conclude with a general discussion in Section 7.

## 2. Spatial Scan Statistic for Censored Outcomes

In general, the spatial scan statistic quantifies the spatial region into areas of potential clusters versus the rest of the study region and conducts a likelihood ratio test, which usually requires a full specification of the model. To allow for more flexibility, we consider using a score statistic from Cox's proportional hazards model instead of a likelihood ratio statistic to avoid specifying the baseline hazard function. We will still denote this as a spatial scan statistic since we are formulating the areas to test for disease clusters, and utilizing the permutation test to derive p-values, in the same fashion as the spatial scan statistic.

To proceed, we first form consecutive circular regions around a fine grid of points, which cover areas of ten to fifty percent of the data. Then for each $k^{th}$ defined circular region, $R_k$, an indicator covariate, $Z_{ki}$, is assigned to be 1 if the $i^{th}$ $(i = 1, \ldots, n)$ individual's geographic location $(s_i, r_i)$ is within the potential cluster area $((s_i, r_i) \in R_k)$ and 0 if outside the area. Suppose each of the study participants has a $p \times 1$ vector of covariates, $\mathbf{X_i}$, a $\delta_i$ to indicate 1 if they have the outcome and 0 otherwise and $T_i$ for time to event or censoring. Consider a Cox's proportional hazards model

$$\lambda(t|R_k, Z_{ki}, \mathbf{X_i}) = \lambda_0(t) \exp\left[\beta_{R_k} Z_{ki} + \boldsymbol{\beta} \mathbf{X_i}\right] \tag{1}$$

where $\lambda_0(.)$ is an unspecified baseline hazard function and $\boldsymbol{\beta}$ is a $1 \times p$ vector of unknown regression parameters. Hence testing if area $R_k$ has a higher hazard rate of disease than $R_k^c$ (the complement of $R_k$) is equal to testing

$$H_O : \beta_{R_k} = 0$$
$$H_A : \beta_{R_k} > 0.$$

It is thus natural to form a score test statistic (a.k.a. Log Rank Statistic) since it is relatively simple, and standard, for the Cox's proportional hazards model to formulate the score test under the null. The corresponding Log Rank test statistic is

$$LR_k = \frac{\sum\limits_{\{j:\delta_j=1\}} \left[ Z_{kj} - \frac{\sum\limits_{\{l:T_l \geq T_j\}} Z_{kl} e^{\hat{\beta}\mathbf{x_l}}}{\sum\limits_{\{l:T_l \geq T_j\}} e^{\hat{\beta}\mathbf{x_l}}} \right]}{\left[ \sum\limits_{\{j:\delta_j=1\}} \left\{ \frac{\sum\limits_{\{l:T_l \geq T_j\}} Z_{kl} e^{\hat{\beta}\mathbf{x_l}}}{\sum\limits_{\{l:T_l \geq T_j\}} e^{\hat{\beta}\mathbf{x_l}}} - \left( \frac{\sum\limits_{\{l:T_l \geq T_j\}} Z_{kl} e^{\hat{\beta}\mathbf{x_l}}}{\sum\limits_{\{l:T_l \geq T_j\}} e^{\hat{\beta}\mathbf{x_l}}} \right)^2 \right\} \right]^{1/2}}. \quad (2)$$

Large positive values of $LR_k$ signify that a region, say $R_k$, has higher hazard rates compared to the rest of the study area. For testing the existence of any spatial clusters we construct the following test statistic, $LR = \sup_k LR_k$.

As detailed by Kulldorff (1997), to determine the significance level of such a test statistic a permutation test can be formulated by fixing the locations $(\mathbf{s}, \mathbf{r})$ and randomly assigning all outcomes $(\delta, \mathbf{t})$, along with their given covariates, $\mathbf{X}$, to the fixed locations. Therefore the only component of an individuals information that is being permuted is location. This process is to be repeated a large number of times, $N$, and the test statistic, $LR$, is calculated at each simulation, $\tilde{LR}_s$. An empirical p-value is calculated by computing the frequency when the simulated data has a more extreme test statistic than the observed test statistic, P-value $= \frac{\sum_{s=1}^{N} I\left[LR \leq \tilde{LR}_s\right]}{N}$.

Finally, we note numerous ways of choosing shapes of potential clusters for testing purposes, such as a square, circle, ellipse, or an annealing algorithm that allows for any shape (Kulldorff, 1997; Duczmal and Assunção, 2004; Tango and Takahashi, 2005; Kulldorff et al., 2006). In this paper we used circular and square regions for computational readiness and did not find any significant loss of power for the scenarios considered.

## 3. Using Cumulative Martingale Residuals to Detect Clusters

The spatial scan statistic is a powerful tool, but the demanding computational burden may restrict its applicability, especially for data with large sample sizes. We consider a simple cluster detection method by using cumulative martingale residuals, which was originally proposed by Lin, Wei and Ying (1993) for model diagnostics. Indeed, as opposed to Lin, Wei and Ying (1993) 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.

### 3.1 *Cumulative Geographic Martingale Residuals for Failure Time Data*

Assume for each subject $i(i = 1, \ldots, n)$ the observed data consists of the time to event or censoring, $T_i$, the noncensoring indicator $\delta_i$, which has value 1 if the event is observed and 0 otherwise, a $p \times 1$ vector of covariates, $\mathbf{X_i}$, along with the location vector $(s_i, r_i)$. Our goal is to detect patterns in 'residuals', after controlling for covariates, $\mathbf{X_i}$, which may depend on spatial locations.

Under the null hypothesis that an individuals failure time is independent of their location, $(s_i, r_i)$, conditional on a given set of covariates, $\mathbf{X}_i$, we assume that the failure time follows a proportional hazards model,

$$\lambda(t|\mathbf{X_i}) = \lambda_0(t) \exp[\boldsymbol{\beta}\mathbf{X_i}], \tag{3}$$

where $\lambda_0(.)$ is an unspecified baseline hazard function, and $\boldsymbol{\beta}$ is a $1 \times p$ vector of unknown regression parameters. Then the partial likelihood score function for $\boldsymbol{\beta}$, conditioned on the at risk population at time $T_i$, $\sum_{l=1}^{n} V_l(T_i)$, where $V_l(t) = I(T_l \geq t)$ is the at-risk process at time $t$, is,

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i [\mathbf{X_i} - \bar{X}(\boldsymbol{\beta}, T_i)], \tag{4}$$

where,

$$\bar{X}(\boldsymbol{\beta}, t) = \frac{\sum_{l=1}^{n} [V_l(t)\mathbf{X_l} \exp(\boldsymbol{\beta}\mathbf{X_l})]}{\sum_{l=1}^{n} [V_l(t) \exp(\boldsymbol{\beta}\mathbf{X_l})]} = \frac{\sum_{l=1}^{n} [V_l(t)\mathbf{X_l} \exp(\boldsymbol{\beta}\mathbf{X_l})]}{S^{(0)}(\boldsymbol{\beta}, t)}.$$

Define the maximum partial likelihood estimator, $\hat{\boldsymbol{\beta}}$, as the solution to $U(\boldsymbol{\beta}) = 0$.

Next define the counting process, $N_i(t) = \delta_i I(T_i \leq t)$ $(i = 1, \ldots, n)$, which is the cumulative sum of events over time $t$. Thus each counting process, $N_i(t)$, has the intensity function $V_i(t)\lambda_0(t)\exp(\boldsymbol{\beta}\mathbf{X_i})$, given the assumed proportional hazards model (3). The martingale residuals are defined as,

$$\hat{M}_i(t) = N_i(t) - \int_0^t V_i(u) \exp(\hat{\boldsymbol{\beta}}\mathbf{X_i}) d\hat{\Lambda}_0(u) \ (i = 1, \ldots, n) \tag{5}$$

where

$$\hat{\Lambda}_0(t) = \int_0^t \frac{\sum_{l=1}^{n} dN_l(u)}{S^{(0)}(\hat{\boldsymbol{\beta}}, u)}.$$

These martingale residuals are similar to any other 'residual' in which it is the observed outcome, $N_i(t)$, minus the expected outcome, assuming model (3) is correctly specified.

We consider a two-dimensional moving block process over location $(x_1, x_2)$, $W_{loc}(x_1, x_2|b)$, which depends on geographic locations for a fixed block size $b$ as follows,

$$W_{loc}(x_1, x_2|b) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I\left(x_1 - b < s_i \leq x_1 + b, x_2 - b < r_i \leq x_2 + b\right) \hat{M}_i(\tau). \tag{6}$$

where $\tau$ is the pre-specified length of the study period. A spatial cluster would occur in areas with a higher intensity of an outcome which implies a larger value of $W_{loc}(x_1, x_2|b)$.

Next consider a *pseudo* moving block process in $(x_1, x_2)$, $\hat{W}_{loc}(x_1, x_2|b)$, as

$$\begin{aligned} \hat{W}_{loc}(x_1, x_2|b) &= \tfrac{1}{\sqrt{n}} \sum_{j=1}^{n} [I\left(x_1 - b < s_j \leq x_1 + b, x_2 - b < r_j \leq x_2 + b\right) \\ &- g(\hat{\boldsymbol{\beta}}, T_j, x_1, x_2, b) - \eta(x_1, x_2, b|\hat{\boldsymbol{\beta}})\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}})\{\mathbf{X_j} - \bar{X}(\hat{\boldsymbol{\beta}}, T_j)\}]\delta_j Z_j \end{aligned} \tag{7}$$

where

$$\begin{aligned} \eta(x_1, x_2, b|\hat{\boldsymbol{\beta}}) = \sum_{l=1}^{n} \ [ \ & \int_0^{\tau} V_l(t) \exp(\hat{\boldsymbol{\beta}}\mathbf{X_l}) I(x_1 - b < s_l \leq x_1 + b, x_2 - b < r_l \leq x_2 + b) \\ & \times \ \{\mathbf{X_l} - \bar{X}(\hat{\boldsymbol{\beta}}, t)\} d\hat{\Lambda}_0(t)], \end{aligned}$$

$$g(\boldsymbol{\beta}, t, x_1, x_2, b) = \frac{\sum_{k=1}^{n} V_k(t) \exp(\boldsymbol{\beta}\mathbf{X_k}) I(x_1 - b < s_k \leq x_1 + b, x_2 - b < r_k \leq x_2 + b)}{S^{(0)}(\boldsymbol{\beta}, t)}, \quad (8)$$

and $Z_j$ $(j = 1, \ldots, n)$ are independent mean 0 and variance 1 random variables that are also independent of $(T_i, \delta_i, \mathbf{X_i}, s_i, r_i)$. It follows that the asymptotic conditional distribution of the *pseudo* process $\hat{W}_{loc}(x_1, x_2|b)$ given the observed data $(T_i, \delta_i, \mathbf{X_i}, s_i, r_i)$ $(i = 1, \ldots, n)$ is equivalent to the limit distribution of $W_{loc}(x_1, x_2|b)$ assuming that geographic location, $(s_i, r_i)$, is independent of time to censoring or failure, $T_i$, after adjusting for covariates, $\mathbf{X_i}$ with the proportional hazards model (3) being correctly specified. This result can be obtained by using the independence between the martingale residuals and geographic location under the null hypothesis. Details of the proof are outlined in a Web-based Appendix 1 at http://www.tibs.org/biometrics, which is along the line of Lin, Wei and Ying (1993).

This asymptotic result immediately allows us to approximate the null distribution of $W_{loc}(x_1, x_2|b)$ with a large number, say, $N$, realizations of $\hat{W}_{loc}(x_1, x_2|b)$, $(\hat{W}_{1,loc}(x_1, x_2|b), ..., \hat{W}_{N,loc}(x_1, x_2|b))$, by repeatedly simulating independent samples of $(Z_1, ..., Z_n)$, while fixing the data $(T_i, \delta_i, \mathbf{X_i}, s_i, r_i)$ $(i = 1, \ldots, n)$ at their observed values. However, for the particular purpose of spatial cluster detection, it is important to allow the data to depict the best cluster size. Therefore, we consider a finite vector of length L of varying cluster sizes, denoted by $\mathbf{b} = (b_1, \ldots, b_L)$, where each $b_l$ denotes half of the edge length of the potential square cluster. Accordingly, we define a cluster detection test statistic to test existence of any spatial clusters,

$$S_{loc} = \sup \left[ \sup_{x_1,x_2} W_{loc}(x_1, x_2|b_1), \ldots, \sup_{x_1,x_2} W_{loc}(x_1, x_2|b_L) \right].$$

Continuous mapping theorem will show that $S_{loc}$ has the same limiting distribution as the following stochastic process, conditional on the observed data,

$$\hat{S}_{loc} = \sup \left[ \sup_{x_1,x_2} \hat{W}_{loc}(x_1, x_2|b_1), \ldots, \sup_{x_1,x_2} \hat{W}_{loc}(x_1, x_2|b_L) \right].$$

Hence, the empirical p-values can be computed as P-value $= \frac{\sum_{j=1}^{N} I\left[S_{loc} \leq \hat{S}_{j,loc}\right]}{N}$, where $\hat{S}_{j,loc}$ is the $\hat{S}_{loc}$ evaluated at the $j^{th}$ realization of $\hat{W}_{j,loc}$. In practice, to obtain the observed test statistic, $S_{loc}$, and simulated test statistics, $\hat{S}_{j,loc}$, it is necessary to create a finite grid of values over $x_1$, $x_2$, and $b$ to approximate the continuous stochastic processes.

This hypothesis test can be inverted to form confidence bands around $W_{loc}(x_1, x_2|b)$ to find the values of $(x_1, x_2, b)$ that have significantly higher hazard rate then expected assuming the null hypothesis and the proportional hazards model (3). Explicitly, $\{(x_1, x_2, b) : W_{loc}(x_1, x_2|b) \geq \hat{S}_{(.95N)}\}$, where $\hat{S}_{(.95N)}$ is the $95^{th}$ percentile of all $\hat{S}_{j,loc}$. Therefore multiple clusters can be easily detecting utilizing this proposed test statistic.

### 3.2 *Standardized Martingale Residuals (SMR)*

To reduce potential dependence between the martingale residuals and covariates, $\mathbf{X}_i$, we also consider a standardized version of test statistic:

$$W_{loc}(x_1, x_2|b) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{I(x_1 - b < s_i \leq x_1 + b, x_2 - b < r_i \leq x_2 + b)}{\sqrt{\hat{Var}\left(\hat{M}_i(\tau)\right)}} \hat{M}_i(\tau),$$

where $\hat{Var}(\hat{M}_i(\cdot))$ is an estimate of the variance for the $i^{th}$ martingale residual defined as,

$$\hat{Var}(\hat{M}_i(t)) = \int_0^t \frac{1}{n} \left[1 - \frac{\exp(\hat{\boldsymbol{\beta}}\mathbf{X}_i)}{\sum_{j=1}^{n} \exp(\hat{\boldsymbol{\beta}}\mathbf{X}_j)}\right] \exp(\hat{\boldsymbol{\beta}}\mathbf{X}_i) \sum_{k=1}^{n} V_k(s) d\hat{\Lambda}_0(s)$$

(Commenges and Rondeau, 2000). Then define another moving block process, $\hat{W}_{loc}(x_1, x_2|b)$, as

$$\begin{aligned} \hat{W}_{loc}(x_1, x_2|b) &= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} [\frac{I(x_1 - b < s_j \leq x_1 + b, x_2 - b < r_j \leq x_2 + b)}{\sqrt{Var(\hat{M}_j(\tau))}} \\ &- g(\hat{\boldsymbol{\beta}}, T_j, x_1, x_2, b) - \eta(x_1, x_2, b|\hat{\boldsymbol{\beta}}) \mathbf{I}^{-1}(\hat{\boldsymbol{\beta}}) \{\mathbf{X_j} - \bar{X}(\hat{\boldsymbol{\beta}}, T_j)\}] \delta_j Z_j \end{aligned} \tag{9}$$

where

$$\begin{aligned} \eta(x_1, x_2, b|\hat{\boldsymbol{\beta}}) = \sum_{l=1}^{n} [ & \int_0^\tau V_l(t) \exp(\hat{\boldsymbol{\beta}}\mathbf{X_l}) \frac{I(x_1 - b < s_l \leq x_1 + b, x_2 - b < r_l \leq x_2 + b)}{\sqrt{Var(\hat{M}_l(\tau))}} \\ \times & \{\mathbf{X_l} - \bar{X}(\hat{\boldsymbol{\beta}}, t)\} d\hat{\Lambda}_0(t)], \end{aligned}$$

$$g(\boldsymbol{\beta}, t, x_1, x_2, b) = \frac{\sum_{k=1}^{n} V_k(t) \exp(\boldsymbol{\beta}\mathbf{X_k}) \frac{I(x_1 - b < s_k \leq x_1 + b, x_2 - b < r_k \leq x_2 + b)}{\sqrt{Var(\hat{M}_k(\tau))}}}{S^{(0)}(\boldsymbol{\beta}, t)},$$

$$Var(\hat{M}_i(\tau)) = \int_0^\tau E\left[\left(1 - \frac{V_i(t)\exp(\boldsymbol{\beta}\mathbf{X}_i)}{S^{(0)}(\boldsymbol{\beta}, t)}\right) V_i(t)\exp(\boldsymbol{\beta}\mathbf{X}_i)\right] d\Lambda_0(t)$$

and $Z_j$ $(j = 1, \ldots, n)$ are mean 0 and variance 1 random variables that are also independent of $(T_i, \delta_i, \mathbf{X_i}, s_i, r_i)$. It follows that the asymptotic conditional distribution of $\hat{W}_{loc}(x_1, x_2|b)$ given the observed data $(T_i, \delta_i, \mathbf{X_i}, s_i, r_i)$ $(i = 1, \ldots, n)$ is equivalent to the limit distribution of $W_{loc}(x_1, x_2|b)$ assuming that geographic location, $(s_i, r_i)$, is independent of time to censoring or failure, $T_i$, after adjusting for covariates, $\mathbf{X_i}$ with the proportional hazards model (3) being correctly specified. The proof follows similarly to the unstandardized martingale residuals (UMR). In practice, we would substitute the consistent estimate of $Var(\hat{M}_i(\tau))$, $\hat{Var}(\hat{M}_i(\tau))$. Both the standardized (SMR) and unstandardized (UMR) test statistics will be applied to the Home Allergens data set in Section 6 and power calculations will be performed in Section 5.

## 4. Similarities between the Spatial Scan Statistic and Cumulative Residual Test

We further note that the proposed spatial scan statistic and cumulative geographic martingale residuals are connected under the proportional hazards model. Explicitly, the similarity occurs when the spatial scan statistic is defined by utilizing square regions instead of circles as potential clusters. Therefore, the indicator variable $Z_{ki} = I(x_{1k} - b < s_i \leq x_{1k} + b, x_{2k} - b < r_i \leq x_{2k} + b)$ where $(x_{1k}, x_{2k})$ is the centerpoint of the potential cluster area $k$ with edge length $2b$. In this case, $W_{loc}(\cdot, \cdot)$ is proportional to,

$$\sum_{i=1}^{n} Z_{ki}\hat{M}_i(\infty) = \sum_{\{j:\delta_j = 1\}} \left[ Z_{kj} - \frac{\sum_{\{l:T_l \geq T_j\}} Z_{kl} e^{\hat{\beta}\mathbf{X_l}}}{\sum_{\{l:T_l \geq T_j\}} e^{\hat{\beta}\mathbf{X_l}}} \right],$$

which is the numerator of the log rank statistic (2).

Key differences, however, do exist between these two test statistics. The spatial scan statistic is advantageous as it allows for a variety of shapes of potential clusters. However, it is computationally burdensome, requires a strong assumption of the Cox's proportional hazards model being correctly specified to have exchangeability to validate the permutation test, and is limited in its ability to define more than one significant clusters. In contrast, the cumulative geographic residual test statistics can be easily inverted to define multiple clusters. These multiple clusters can actually overlap which allows for the proper shape of the cluster to be formed. Therefore, even though the initial cluster is restricted to be a square or rectangle (in order to define a valid two-dimensional process), by overlapping significant clusters the proper cluster shape can still be detected.

## 5. Simulation Study

We conducted simulations, calculating the power and type I error for both the spatial scan statistic and cumulative geographic residual for censored outcomes. For computational efficiency for all simulations we allowed a finite range, for the radii for the spatial scan statistic and half of edge length, $b$, for the cumulative geographic residual, of 0.5 to 2 sequenced by 0.5.

We first conducted power calculations by considering an $8 \times 8$ unit-less area. A simulated data set was derived by dividing the area into 16 equally sized squares of size 2 $\times$2 as depicted in Figure 7. The study population size was 500 and each participant was randomly assigned to 1 of the 16 grids. Given the grid, the x and y coordinates for each participant were randomly assigned with a uniform distribution over the grid area.

To create a single cluster in consecutive grid areas 6 and 10 we first generated random variables $C_i$ and $F_i$ $(i = 1, \ldots, n)$ from the exponential distributions with constant hazards $\lambda_{ci}$ and $\lambda_{fi}$, respectively. If participant i, is assigned to grid 6 or 10, then $\lambda_{fi} = 1/4$ otherwise $\lambda_{fi} = 1/2$. We set $\lambda_{ci} = 1/3$ for all i. Given $F_i$ and $C_i$, define $\delta_i = I(F_i \leq C_i)$ and

$T_i = min(C_i, F_i)$ to complete the randomly generated failure time data set with a higher likelihood of failures within grids 6 and 10 and therefore a corresponding cluster.

[Figure 1 about here.]

We created 500 of the defined simulated data sets and on each data set both the spatial scan statistic and cumulative geographic residual test were performed. For the spatial scan statistic we used both circular and square regions for shapes of potential clusters. A power calculation was conducted for each test statistic by calculating the percentage of times that it found significant clusters (0.05 significance level) that overlapped either, or both, grid areas 6 and 10. The power results were 0.526 for the spatial scan statistic (0.494 using a square region) and 0.530 for the cumulative geographic residual test. Therefore, both test statistics performs equally well for a single cluster situation when the population is distributed uniformly over the study area.

The second test situation studied having two clusters in the same $8 \times 8$ unit-less grid. The first cluster was located in consecutive grid areas 6 and 10, while the second cluster was located in grid area 16. The location of each of the participants were assigned according to the same scheme as described in the first test situation. Also, using the notation from the first test situation, generate $C_i$ and $F_i$ $(i = 1, \ldots, n)$ from the following exponential distribution, if participant i is assigned to be in grids 6 or 10 then $\lambda_{fi} = 1$, if assigned to be in grid 16 $\lambda_{fi} = 1/2$, and $\lambda_{fi} = 1/3$ otherwise. We again hold $\lambda_{ci} = 1/3$ for all i. Therefore, given $F_i$ and $C_i$, define $\delta_i = I(F_i \leq C_i)$ and $T_i = min(C_i, F_i)$ to complete the randomly generated failure time data set with the highest likelihood of failures within grids 6 and 10 (Cluster 1), second highest in grid 16 (Cluster 2), and lower likelihood everywhere else.

We calculated an overall power calculation for each test. A test was deemed significant if a cluster was detected that overlapped at least 1 of the 2 cluster regions. The spatial scan statistic had an overall power of 0.936 (0.922 using a square region), which is slightly

higher than the cumulative residual test with a power of 0.922. However, when analyzing the results of which cluster each test found, almost every cluster detected overlapped the larger, Cluster 1, with a power of 0.934 for the spatial scan statistic (0.918 using a square region) and 0.002 (0.006 using square region) of the time found a cluster that overlapped cluster two. The cumulative residual found Cluster 1 with a power of 0.922, but simultaneously found Cluster 2 with a power of 0.688. So even though the spatial scan statistic has higher power to detect Cluster 1, the cumulative residual test was able to also detect Cluster 2 68.8% of the time. Many data analyses wish to find multiple clusters, which suggests cumulative geographic residual may be the preferred method without too much loss of power.

We also performed calculations to check the type I error rate for the cumulative geographic residual test and spatial scan statistic. These simulations were conducted by generating 1000 test studies where location was randomly assigned uniformly over a 10 by 10 grid and corresponding failure time outcomes were randomly assigned to each grid location. Type I error was calculated as the proportion of the 1000 simulations that found at least one cluster significant at the 0.05 significance level. The results are described in Table 1. The type I error is being held at the $\alpha$-level of 0.05 over all sample sizes and percentage of failures for both test statistics. Therefore the tests are performing as expected.

[Table 1 about here.]

The final simulation we ran compared the standardized (SMR) and unstandardized (UMR) cumulative geographic residual test to see if using the standardized martingale residual increased power. We created a binary covariate $X_i$ under two scenarios: one in which $X_i$ is related to outcome, $(T_i, \delta_i)$, but not location (independent predictor) and a scenario in which $X_i$ is dependent on outcome, $(T_i, \delta_i)$, and location (interaction). For the first scenario, where $X_i$ is an independent predictor, we conducted the same simulations as discussed without covariates for both single and multiple clusters, but altered all failure time constant

hazards, $\lambda_{fi}$, by multiplying it by $\exp(0.4 * X_i)$ where $X_i$ is a randomly generated Bernoulli covariate with mean 0.30. For the single cluster the SMR method had a power of 0.463, which was not substantially higher then the UMR method with a power of 0.457. The multiple cluster's overall power results for finding at least one of the two clusters were 0.911 for the SMR and 0.917 for the UMR results. Therefore, the standardized test statistic and the unstandardized test statistic had relatively equivalent power for cluster detection for the independent predictor scenario.

For scenario two, in which $X_i$ is an interaction, we randomly assigned all subjects to 1 of the 16 grids depicted in Figure 7. If subject $i$ is assigned to grids 6 or 10, then $X_i$ is randomly generated from a Bernoulli with mean 0.5, otherwise $X_i$ is randomly generated from a Bernoulli with mean 0.2. To obtain the censored outcomes, we followed the same simulation set up as discussed for independent predictor, where we multiply the failure time constant hazards by $\exp(0.4 * X_i)$. For the single cluster, the SMR method had a power of 0.664, which was almost equivalent to the UMR method with a power of 0.672. The multiple cluster's overall power results for finding at least one of the two clusters were 0.917 for the UMR and 0.911 for the SMR results.

Therefore, for both scenarios, the standardized and unstandardized are almost identical, but the standardized method had slightly higher power for the independent predictor, scenario one, and slightly lower power for scenario two. However, these differences were not larger then what would be expected from Monte Carlo simulation error and therefore this indicates that the standardized method does not significantly improve cluster detection power.

## 6. Home Allergens and Asthma Study Analysis

We now apply the proposed methods to the Home Allergens and Asthma prospective cohort study. 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 U.S.A. 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 with significant clusters of disease in the first four years of life for: time to asthma or censoring, time to allergic rhinitis/hayfever or censoring, and time to eczema or censoring.

[Figure 2 about here.]

There has been relatively little time to event data documenting potential risk factors for any of our outcomes and therefore we will make all a priori hypotheses based on incidence outcome results. Previous results from a study on the mothers who had been screened for the Home Allergens and Asthma study displayed higher IgE, an indicator of allergic response, in southern urban Boston, Chelsea, and Revere, all lower socioeconomic areas, and lower IgE in the western suburbs (Litonjua et al., 2005). Figure 7 displays the towns and median family income at the U.S. 2000 Census tract level for the study area. Boston, Chelsea, and Revere are displayed as having relatively low median family income compared to the rest of the study population. Further, elevated levels of IgE have been associated with a higher prevalence of asthma, eczema, and hay fever in this population (Litonjua et al., 2005).

Since asthma in children is documented to be more prevalent in minority and disadvantaged populations (Gergen et al., 1988,Schwartz et al., 1990, and Litonjua et al., 1999), we expect to find a disease cluster in southern urban Boston, Chelsea, and Revere. However, previous studies have documented atopic disorders (eczema and hay fever) as being conditions of the relatively affluent (Gergen et al., 1987 and Chen et al., 2002). This finding, for hayfever, may be the result of underreporting, or underdiagnosis, of hayfever in the disad-

vantaged population, because of various barriers to care (Strunk et al., 2002). Also, one may presume that due to these barriers to care that the time to diagnosis may also be on average longer even in the children who eventually got diagnosed. Therefore, the a priori hypothesis for the location of a hayfever/allergic rhinitis spatial cluster would be in the Western, more affluent suburbs.

There has been no documentation of underreporting of eczema in children, and given the results that elevated levels of IgE are related to eczema in the mother population, it may indicate that the cluster would occur in the southern urban Boston, Chelsea, and Revere areas. However, the a priori hypothesis assumed that it would follow a similar pattern as the hayfever/allergic rhinitis outcome. To test these a priori hypothesis we have conducted three spatial cluster detection analyses on the outcomes: (1) time to asthma or censoring, (2) time to allergic rhinitis/hayfever or censoring, and (3) time to eczema or censoring.

First displayed in Figures 7 and 7 are the results from the analysis for the outcome time to doctor diagnosed asthma or censoring. Two analyses were conducted, one without adjusting for covariates and one with adjusting for the following marginally significant predictors: parental smoking (Adjusted hazard ratio = 2.017 (p-value=0.024)), income $\geq \$50,000$ (Adjusted hazard ratio = 0.714 (p-value=0.099)), and log transformed endotoxin (Adjusted hazard ratio = 1.263 (p-value = 0.074)). For all analyses of the data, areas of potential clusters were confined to be between 1000 and 15000 meters in radius for the spatial scan statistic and between 2000 and 30000 meters in edge length for the cumulative geographic residual. Figure 7 displays that the maximum cluster, for both statistics, is located in the eastern portion of the study area including southern urban Boston, Chelsea, and Revere, but also some of the nearby surrounding towns. Note that these maximum clusters stay close to the same location even after adjusting for covariates, which include income. However, the only significant cluster, at the 0.05 level, was found using the cumulative residual method

without adjusting for other covariates. This may indicate that the predictors are sufficiently accounting for the spatial cluster assuming correct model specification. The location of the cluster is as expected given the a priori hypothesis that the cluster would be similar in location to elevated IgE levels in the mothers.

[Figure 3 about here.]

[Figure 4 about here.]

Two other analyses were conducted to study the following allergic outcomes: time to hayfever/allergic rhinitis or censoring and time to eczema or censoring. There were no marginally significant predictors for either outcome so the only analyses conducted did not adjust for covariates.

Figure 7 displays the results from these analysis using both the cumulative geographic residual and spatial scan statistic. The cumulative geographic residual found significant clusters for time to allergic rhinitis/hay fever, while the spatial scan statistic found only a marginally significant cluster for this outcome. In general the outcome hayfever/allergic rhinitis had a significant cluster in the western suburbs of Boston, while eczema did not have any significant clusters, the highest potential cluster was located in the southern study area overlapping both the hayfever and asthma clusters. The areas of disease clusters that differed most between the spatial scan statistic and the cumulative geographic residual occurred for the outcome hayfever/allergic rhinitis. The area that the spatial scan statistic found as a cluster was much smaller then the area for the cumulative geographic residual, but they do overlap. This is a common trend of the cumulative geographic residual test in which it tends to find areas of maximum clusters that are larger then the spatial scan statistic.

[Figure 5 about here.]

Even though the spatial scan statistic and the cumulative geographic residual tend to find slightly different clusters they do overlap and their Kaplan-Meier curves, displayed only for the outcome time to doctor diagnosed asthma (Figures 7), indicate that both test statistics find areas that have very different estimated survival curves of between versus within cluster. Therefore, both test statistics perform well in finding areas that have significantly higher hazard ratios.

Overall, there are significant geographic cluster for the time-to-event outcomes asthma and allergic rhinitis/hayfever observed. Asthma had observed clusters in the metropolitan Boston area, including southern Boston, Chelsea , and Revere, while hayfever/allergic rhinitis have a cluster in the Western suburbs. This indicates that asthma may be exacerbated by urban predictors. The location of the hayfever/allergic rhinitis outcomes may indicate underdiagnosis in the more disadvantaged population in southern Boston, Chelsea, and Revere, but may also be more exacerbated by suburban exposures. All conclusions are in reference to children under the age of 4.

## 7. Discussion

In this paper we have extended two techniques, the spatial scan statistic and the cumulative geographic residual test, for detecting spatial cluster with censored outcomes. By utilizing the cumulative geographic residual methodology, we detected a significant cluster of childhood doctor diagnosed asthma in the inner-city Boston area and hayfever/allergic rhinitis in the western suburbs of Boston when not adjusting for other covariates. The spatial scan statistic found only marginally significant clusters for all outcomes. It would be of interest to further investigate potential predictors that may be related to those neighborhoods in Boston, such as traffic exposure in the urban area or existence of parks/greenery in the suburbs.

We also performed power comparisons between the spatial scan statistic and cumulative

geographic residual. In all examined scenarios, these two tests seem to be comparable in terms of power. In addition, both tests statistics maintain the nominal type I error under the null hypothesis. Therefore we conclude that both methods are valid for censored outcomes.

It should be noted that there are other potential statistical methods to determine the asymptotic distribution, under the Null hypothesis of no clustering, for both test statistics. Particulary, the exact asymptotic distribution has been derived by Hashemi and Commenges (2002) and is a valid alternative. However, for the Cumulative Residual test we chose to utilize the multiplier central limit theorem framework to provide a convenient means of simulating the asymptotic stochastic process over space (Figure 7). Further, we chose to use the standard permutation approach that has been proposed for the spatial scan statistic.

Nevertheless, we do think there exist ample research topics in this relatively new area. These include extending the proposed cumulative geographic martingale residual method to utilize the linear transformation model for censored data (Fine et al., 1998), or a parametric failure time model, instead of the Cox's proportional hazards, to improve power for spatial cluster detection. Further, these test statistics could both be extended to the space-time setting to incorporate cluster detection over time in a similar fashion to the extension of the spatial scan statistic with binary and poisson outcome data to the space-time setting (Kulldorff, 2001). With such prevalence of longitudinal and survival studies, from fields such as environmental health, cancer research, community based research, and neurodegenerative disease research to just name a few, and with the rapid advancement of GIS technology, cluster detection methods for failure time outcomes will become more useful over time.

## References

Brugge, D., Vallarion, J., Ascolillo, L., Osgood, N., 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.

Chen, J., Krieger, N., Van Den Eeden, S. and Quesenberry, C. (2002). Different slopes for different folks: socioeconomic and racial/ethnic disparities in asthma and hay fever among 173,859 United States men and women. *Environmental Health Perspectives* **110**, 211–216.

Commenges, D. and Rondeau, V. (2000). Standardized martingale residuals applied to grouped left truncated observations of dementia cases. *Lifetime Data Analysis* **6**, 229–235.

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.

Fine, J., Ying, Z. and Wei, L. (1998). On the linear transformation model for censored data. *Biometrika* **85**, 980–986.

Finkelstein, J., Fuhlbrigge, A., Lozano, P., Grant, E., Shulruff, R., Arduino, K. and Weiss, K. (1999). Parent-reported environmental exposures and environmental control measures for children with asthma. *Archives of Pediatrics and Adolescent Medicine* **156**, 258–264.

Gergen, P., Mullaly, D. and Evans, R. (1988). National survey of prevalence of asthma among children in the United States, 1976 to 1980. *Pediatrics* **81**, 1–7.

Gergen, P., Turkeltaub, P. and Kovar, M. (1987). The prevalence of allergic skin test reactivity to eight common aeroallergens in the u.s. population: results from the second national health and nutrition examination survey. *Journal of Allergy and Clinical Immunology* **80**, 669–679.

Hashemi, R. and Commenges, D. (2002). Correction of p-values after multiple tests in a Cox proportional hazard model. *Lifetime Data Analysis* **8**, 335–348.

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

Kulldorff, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society A* **164**, 61–72.

Kulldorff, M., Huang, L., Pickle, L. and Duczmal, L. (2006). An elliptic spatial scan statistic. *Statistics in Medicine* Epub ahead of print.

Lin, D., Wei, L. and Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**, 557–572.

Litonjua, A., Carey, V., Weiss, S. and Gold, D. (1999). Race, socioeconomic factors, and area of residence are associated with asthma prevalence. *Pediatric Pulmonology* **28**, 394–401.

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

Schwartz, J., Gold, D., Dockery, D., Weiss, S. and Speizer, F. (1990). Predictors of asthma and persistent wheeze in a national sample of children in the United States. *American Review of Respiratory Disease* **142**, 155–62.

Strunk, R., Ford, J. and Taggart, V. (2002). Reducing disparities in asthma care: priorities for research-national heart, lung, and blood institute workshop report. *Journal of Allergy and Clinical Immunology* **109**, 229–237.

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

## Captions for Figures

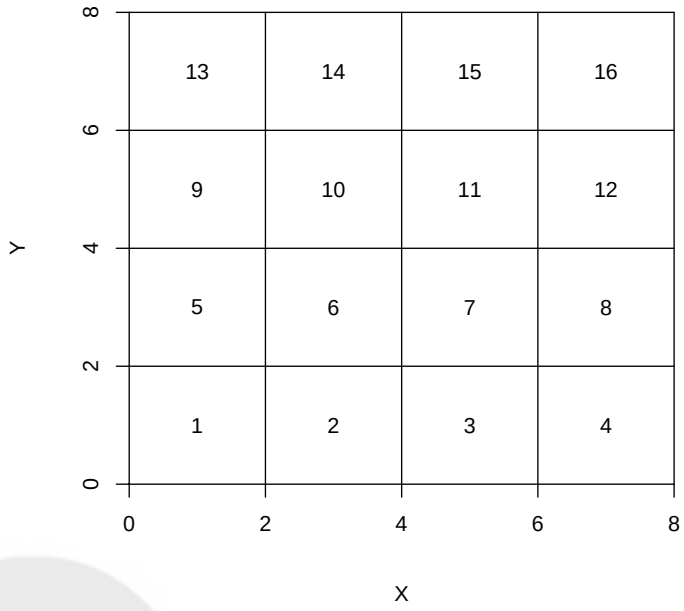
Figure 1: Grid system of study area for power simulation data sets.

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

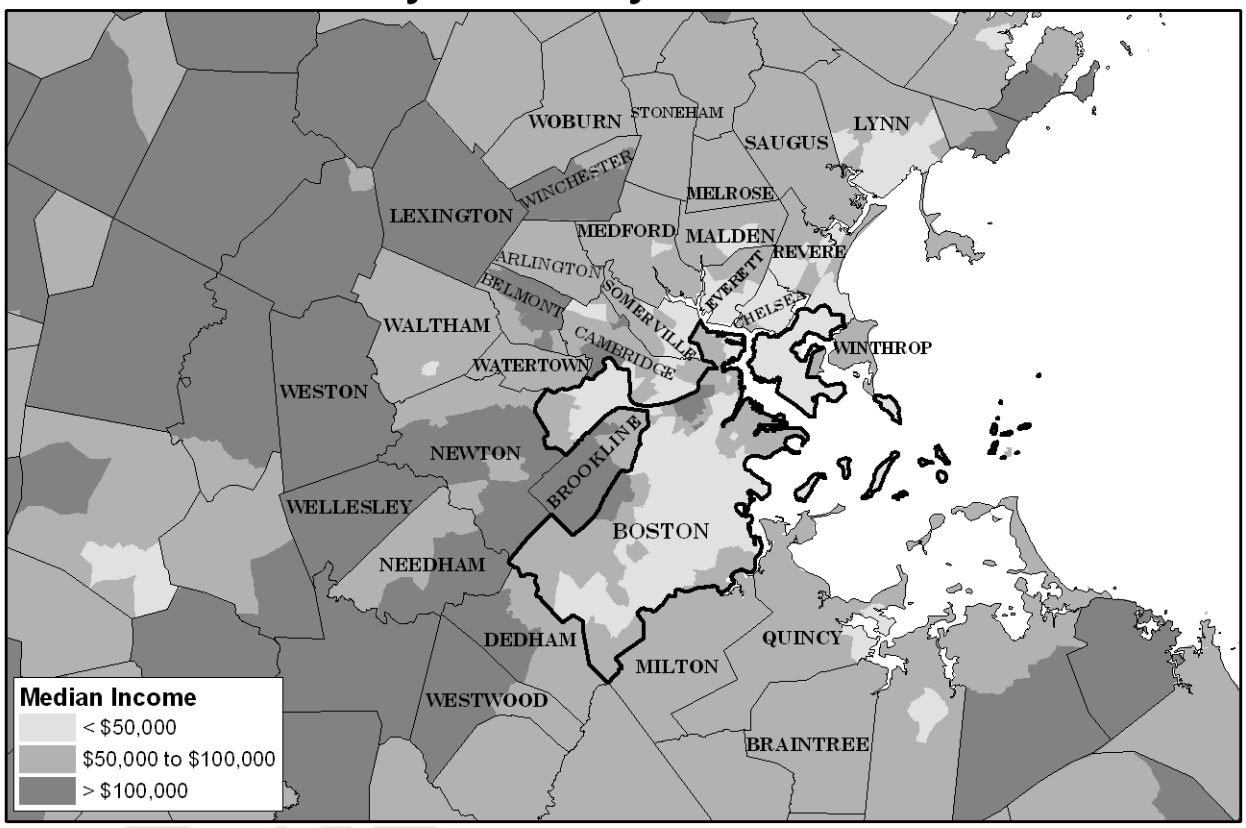
Figure 3: Indicated areas of significant cluster location of outcome time to doctor diagnosed asthma or censoring with corresponding p-values of these maximum cluster: (a) and (b) correspond to the cumulative geographic residual. (c) and (d) correspond to the spatial scan statistic. Adjusted for parental smoking, Income ($\geq \$50,000$), and log transformed endotoxin. Non-dashed lines correspond to the maximum cluster location and dashed lines corresponds to the entire area with significant (0.05 level) clusters, which potentially include multiple square areas.

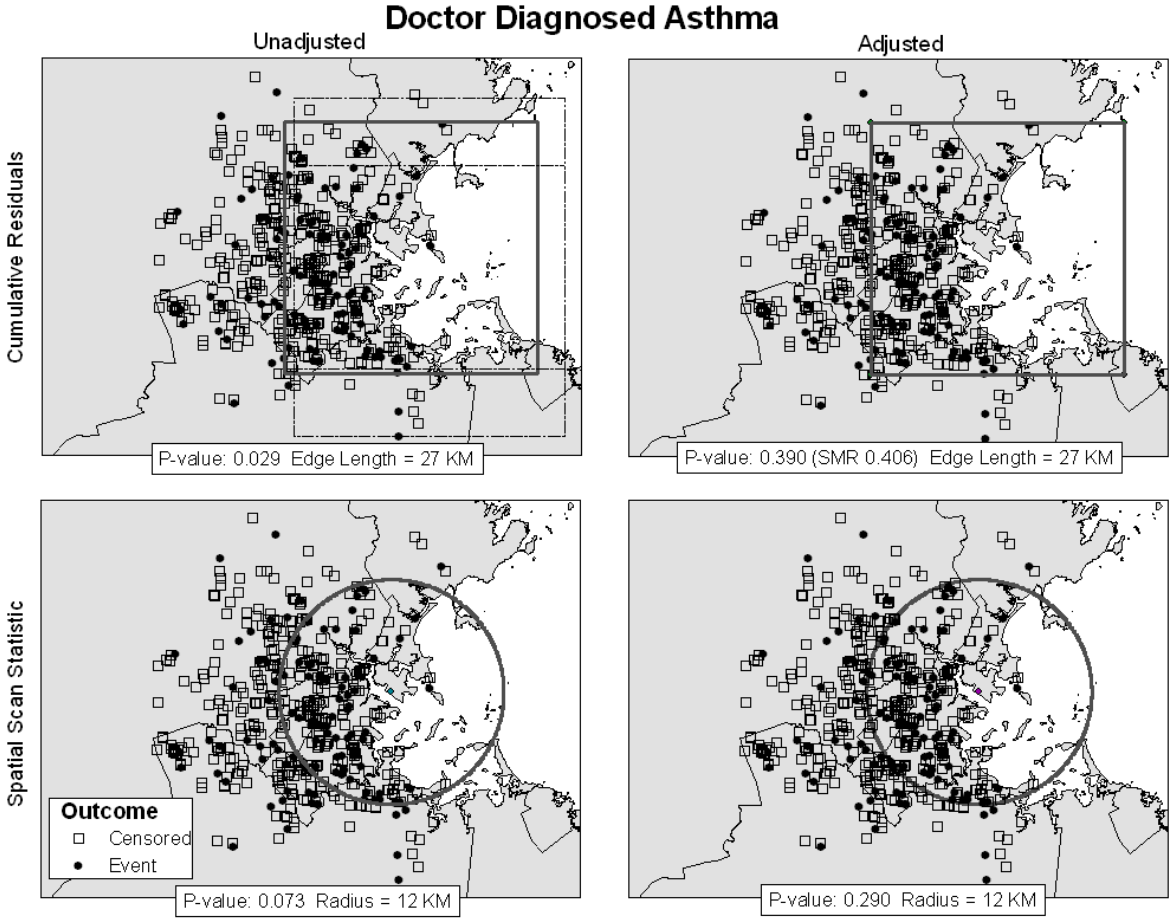
Figure 4: Kaplan Meier curves for the outcome doctor diagnosed asthma or censoring stratified by within versus outside maximum cluster: (a) and (b) correspond to the cumulative geographic residual. (c) and (d) correspond to the spatial scan statistic. Adjusted for parental smoking, Income ($\geq \$50,000$), and log transformed endotoxin.

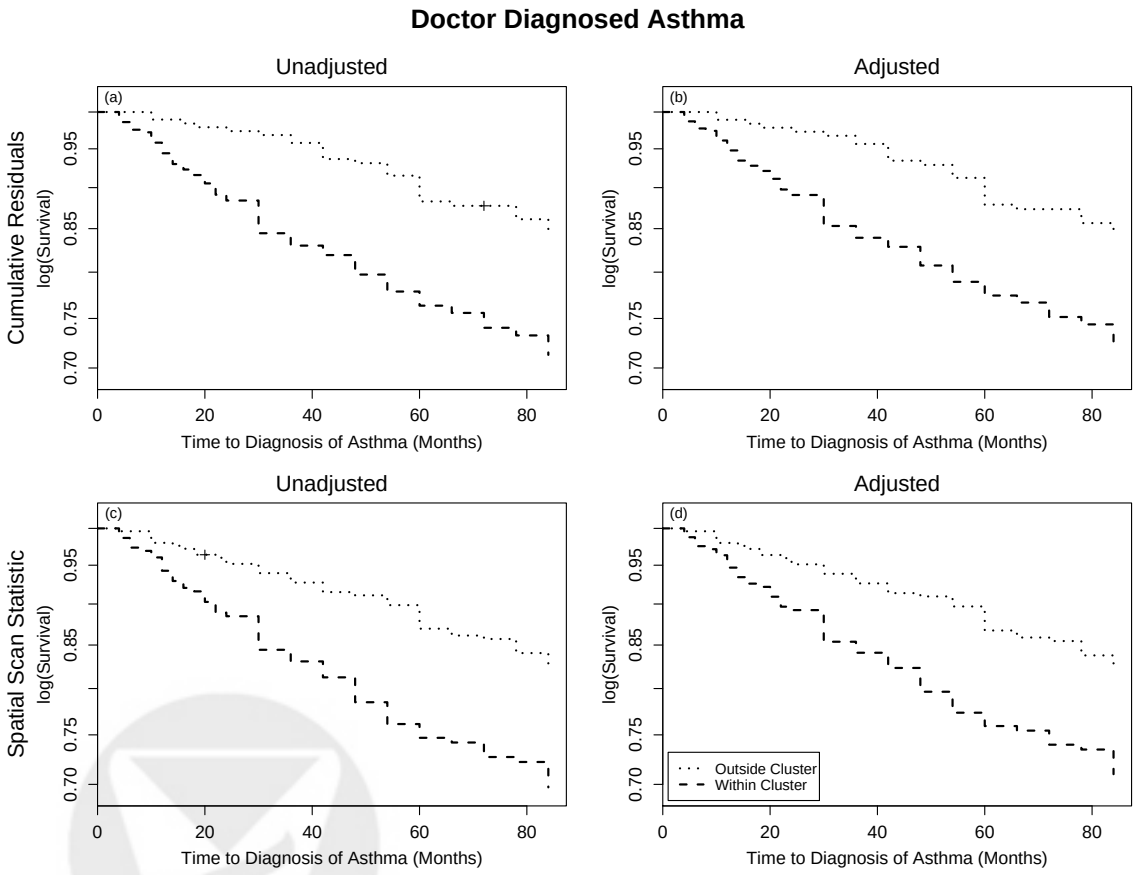
Figure 5: Indicated areas of significant cluster location of two outcomes: time to allergic rhinitis/hayfever or censoring and time to eczema or censoring. (a) and (b) correspond to the cumulative geographic residual. (c) and (d) correspond to the spatial scan statistic. Non-dashed lines correspond to the maximum cluster location and dashed lines corresponds to the entire area with significant (0.05 level) clusters, which potentially include multiple square areas.



## Median Family Income by 2000 U.S. Census Tract







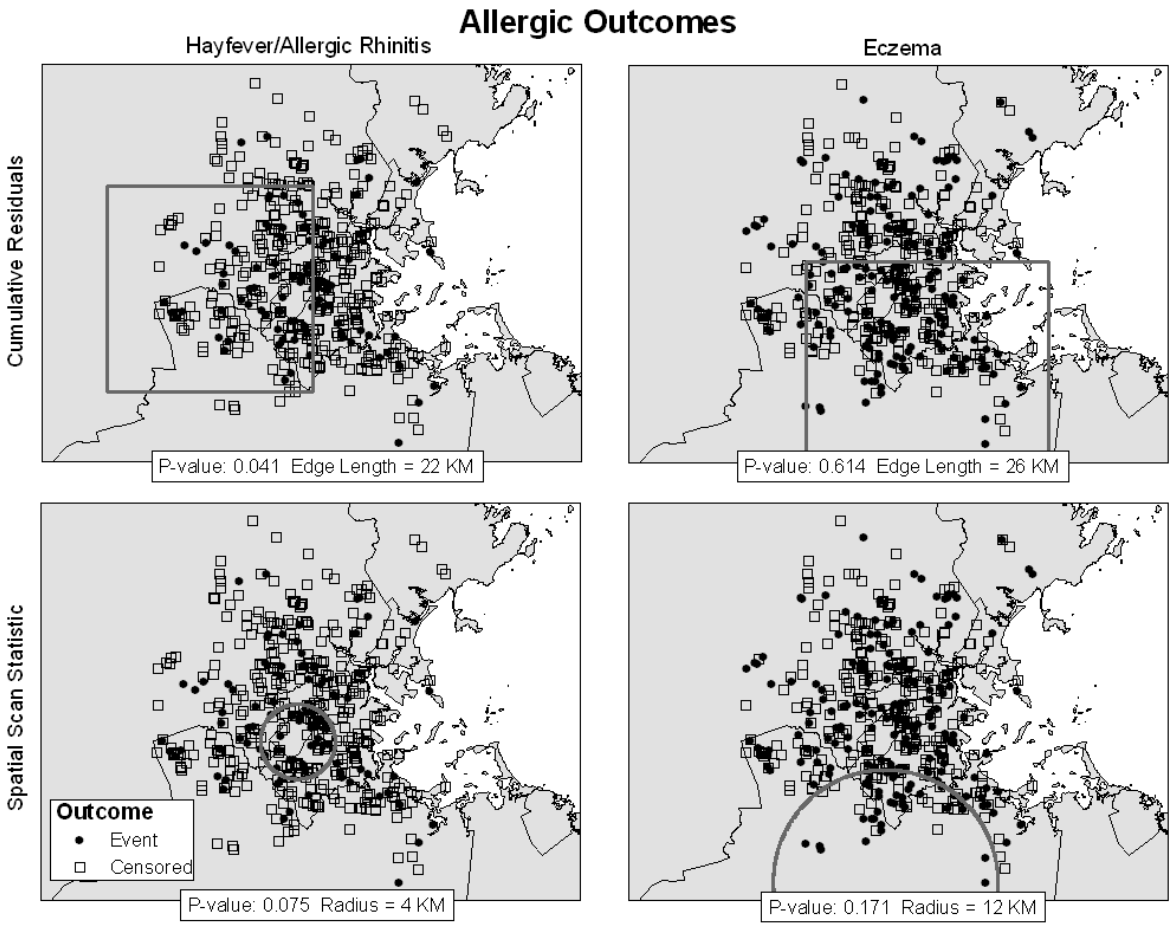


**Table 1**

*Type I error rate of Cumulative Geographic Martingale Residual Test and Spatial Scan Statistic for different sample sizes and percentage of failure events.*

| | | **Number of Observations** | | | | | |
|---|---|---|---|---|---|---|---|
| | | Cumulative Residual | | | Spatial Scan | | |
| | | 100 | 300 | 500 | 100 | 300 | 500 |
| Censoring | 80% | 0.051 | 0.058 | 0.049 | 0.055 | 0.049 | 0.062 |
| Proportion | 60% | 0.042 | 0.048 | 0.041 | 0.047 | 0.038 | 0.052 |
| | 40% | 0.041 | 0.043 | 0.049 | 0.055 | 0.052 | 0.062 |