Spatial Cluster Detection for Longitudinal Outcomes using Administrative Regions

Andrea J. Cook [1]
Biostatistics Unit, Group Health Research Institute
Seattle, WA 98101, USA
Department of Biostatistics, University of Washington
Seattle, WA 98195, USA

Diane R. Gold
The Channing Laboratory, Department of Medicine,
Brigham and Women's Hospital and Harvard Medical School
Boston, MA 02115, USA
Department of Environmental Health, Harvard School of Public Health
Boston, MA 02115, USA

Yi Li
Department of Biostatistics, Harvard University and Dana Farber Cancer Institute
Boston, MA 02115, USA

Abstract

This manuscript proposes a new spatial cluster detection method for longitudinal outcomes that detects neighborhoods and regions with elevated rates of disease while controlling for individual level confounders. The proposed method, CumResPerm, utilizes cumulative geographic residuals through a permutation test to detect potential clusters which are are defined as sets of administrative regions, such as a town, or group of administrative regions. Previous cluster detection methods are not able to incorporate individual level data including covariate adjustment, while still being able to define potential clusters using informative neighborhood or town boundaries. Often it is of interest to detect such spatial clusters because individuals residing in a town may have similar environmental exposures or socioeconomic backgrounds due to administrative reasons, such as zoning laws. Therefore these boundaries can be very informative and more relevant than arbitrary clusters such as the standard circle or square. Application of the CumResPerm method will be illustrated by the Home Allergens and Asthma prospective cohort study analyzing the relationship between area or neighborhood residence and repeated measured outcome, occurrence of wheeze in the last 6 months, while taking into account mobile locations.

Key words: Asthma; Cluster Detection; Cumulative Residuals; Repeated Measures; Wheeze

[1]Corresponding author: e-mail: cook.aj@ghc.org telephone: (206)287-4257 fax: (206)287-2871

1 Introduction

Prospective cohort studies are an increasingly common study design to address important public health questions. Typically the study population is followed over time and at certain time-points resurveyed, or revisited, which formulates a longitudinal dataset. The Home Allergens and Asthma study is an example of 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.

To be able to detect elevated rates of disease for longitudinal outcomes there was a need to develop a new spatial cluster detection method, which we propose in this manuscript named Cumulative Geographical Residuals Test (CumResPerm). There are methods available for cross-sectional 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), but there is only one method available for longitudinal outcomes (Cook, Gold, and Li, 2009).

Further, the methods available for individual-level data use arbitrary shapes such as circles or squares as potential clusters instead of informative administrative regions. The proposed CumResPerm method defines potential cluster by using informative spatial regions,

such as towns or groups of towns. Typically the environmental exposures within a town or neighborhood may be more similar due to zoning laws or other similar administrative reasons. Towns or neighborhoods that are near each other may also have similar types of environmental exposures. By incorporating this information results from our analyses may be more informative in pinpointing why the area has elevated rates of disease.

Another nuance of the CumResPerm method is how it readily incorporates information of individuals that move during the study. It is a recurring problem of how to incorporate moving locations of individuals especially when being followed over time. 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. Therefore it would be essential to include in the spatial cluster analysis all individuals who have been followed even if they move outside of the study area. The proposed method can readily incorporate moving study participants and therefore reduce missingness in the analysis. For the Home Allergens and Asthma study we will analyze the data using two different spatial locations: (1) location at birth and (2) location at age of repeated measure.

The outline of this manuscript begins by presenting in Section 2 the Cumulative Geographic Residual Permutation Test (CumResPerm) to longitudinal data. It then conducts simulations to check 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 Cumulative Geographic Residual Permutation Test (CumResPerm)

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 ($i = 1, ..., n$), at occasion k ($k = 1, ..., K_i$), Y_{ik}, is binary with a $p \times 1$

vector of covariates, $\mathbf{X_{ik}}$, and in region R_{ik}. Under the assumption that disease status is independent of geographic location (i.e. no spatial cluster), 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,

$$g(\mu_{ik}) = logit(\mu_{ik}) = \boldsymbol{\beta}\mathbf{X_{ik}}, \tag{2.1}$$

and $\boldsymbol{\beta}$ is a $1 \times p$ vector of regression parameters. For our analysis we take a generalized estimating equation (GEE) approach with an independent working covariance structure and use the robust sandwich variance estimator to handle the correlation between observations (Liang and Zeger, 1986). Then we estimate $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, by solving the score function for 0 yielding the residuals, $\hat{e}_{ik} = Y_{ik} - g^{-1}(\hat{\mu}_{ik})$. In the next section a permutation method is proposed that utilizes the residuals and the assumption of independence between residuals and geographic location to detect disease clusters.

2.1 Cumulative Geographic Residual Permutation Test (CumResPerm)

For the following spatial cluster detection method the possible clusters are defined as predefined regions such as cities, census tracks, or neighborhoods. Previous cluster detection methodology utilizing Cumulative Geographic Residuals allowed possible clusters to be arbitrary regions (Cook et al, 2007, 2009). However, it may be advantageous to use previously defined regions, or neighborhoods, particularly if these defined regions are not arbitrary, but may pinpoint underlying covariates that may be related to the outcome, such as low SES neighborhoods.

Therefore the study area is broken into L predefined non-overlapping regions or neighborhoods $\mathbf{R} = (R_1, \ldots, R_L)$. Each study participant, i, at time k, is therefore located at r_{ik} where $r_{ik} \in \mathbf{R}$. A potential spatial cluster, $C_m \subset \mathbf{R}$, is defined to be a subset of L potential regions. We constrain to allow only potential spatial clusters that form a single continuous

mass (e.g. Travel from one region to the next without crossing into non-cluster regions). Further, it may be of interest to restrict potential spatial clusters to encompass a limited number of neighborhoods/regions by restricting the number of regions in C_m to be $N_R \leq L$. This is not necessary for the method to perform well, but often the question of interest might be does a town or a small region have elevated rates of an outcome. This is similar to the restriction used in standard spatial cluster detection software to restrict clusters to be at most 50% of the study population or the size of the cluster to be at most 5 miles.

Under the null hypothesis, H_0, spatial location is independent of outcome given covariates $\mathbf{X}_i$, and assuming that the link function is correctly specified, it can be assumed that the residuals, $\hat{\mathbf{e}}_i$, are independent of location, $\mathbf{r}_i$, and therefore are independent with mean 0 and bounded variances. By these assumptions, $\{\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_n\}$ are first-moment exchangeable random vectors since unconditional first moment of $\{\hat{\mathbf{e}}_{d_1}, \ldots, \hat{\mathbf{e}}_{d_n}\}$ for any permutation $d_1, \ldots, d_n$ of the integers $1, \ldots, n$ is the same, that is, $E(\hat{e}_1, \ldots, \hat{e}_n) = E(\hat{e}_{d_1}, \ldots, \hat{e}_{d_n})$ for any $x_1, \ldots, x_n$. However, due to estimating the unknown parameters the residuals are not globally exchangeable since $Var(\hat{e}_1, \ldots, \hat{e}_n)$ is not necessarily equal to $Var(\hat{e}_{d_1}, \ldots, \hat{e}_{d_n})$ for any $x_1, \ldots, x_n$. It has been shown in simulation for the case of residuals from generalized linear models and proportional hazards models using martingale residuals that under the less stringent assumption of first-moment exchangeability that a permutation test based on residuals is valid (Jacqmin-Gadda et al, 1997; Cook and Li, 2008).

We first define the following test statistic for each potential cluster, C_m, as,

$$W_{loc}(C_m) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K_i} I(r_{ik} \in C_m) \hat{e}_{ik}. \tag{2.2}$$

To form the permutation test we fix the observed locations $(\mathbf{r}_1, \ldots, \mathbf{r}_n)$ and randomly permute the order of the observed residuals $(\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_n)$ by permuting $d_1, \ldots, d_n$ to create a permuted dataset $((\mathbf{r}_1, \hat{\mathbf{e}}_{d_1}), \ldots, (\mathbf{r}_n, \hat{\mathbf{e}}_{d_n}))$, where $\mathbf{r}_i = (r_{i1}, \ldots, r_{iK_i})^T$ and $\hat{\mathbf{e}}_{d_i} = (\hat{e}_{d_{i1}}, \ldots, \hat{e}_{d_{iK_i}})^T$. This is done a large number of times to create N_{sim} datasets and for each permuted dataset the

following test statistic for each cluster, C_m, can be calculated,

$$\tilde{W}_{loc}(C_m) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K_i} I(r_{ik} \in C_m)\hat{e}_{d_{ik}}.$$

To test for the existence of any clusters in the dataset, an empirical p-value can be calculated as,

$$\text{P-value} = \frac{\sum_{s=1}^{Nsim} I\left[S_{loc} \leq \tilde{S}_{loc,s}\right]}{N_{sim}},$$

where $S_{loc} = \sup_m W_{loc}(C_m)$ and $\tilde{S}_{loc,s} = \sup_m \tilde{W}_{loc}(C_m, s)$ for the s permuted dataset. Further for each potential cluster, C_m, the empirical p-value can be calculated as,

$$\text{P-value} = \frac{\sum_{s=1}^{Nsim} I\left[W_{loc}(C_m) \leq \tilde{S}_{loc,s}\right]}{N_{sim}}.$$

However, there is one key problem with the structure of this permutation test in the situation where individuals are lost to follow-up or have moved outside of the study area. This can be thought of as a standard missing data issue observed in longitudinal studies and there are numerous approaches to handle such issues. We have proposed for our application to handle lost to follow-up by using an address carry forward assumption which uses a participants last known location as the missing location for future follow-up time points. Of course, this could create problems since an individual may be loss to follow-up because they moved from the area. For a secondary analysis we have also assumed that the missing locations were locations outside of the study area. There are numerous other approaches to handle such missing location data, such as imputation, but since the focus of the paper is on the new statistical method we have chosen to show the two simple and more extreme cases as examples.

To handle participants that moved outside the study area we create a new location region, R_{L+1}, that is not included in the set of regions to form potential clusters, but that individual

location and outcome is still included in the dataset when permuting.

2.2 Cumulative Time-Dependent Geographic Cluster Detection

The previous section presented a global test statistic utilizing all of the longitudinal 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 than 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 $t = t_1, t_2, \ldots, t_K$,

$$W_{loc,t}(C_m, t) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K_i} I(r_{ik} \in C_m, t_{ik} = t)\hat{e}_{ik},$$

where t_{ik}, r_{ik}, and $\hat{e}_{ik}$ are the corresponding time, region, and residual for individual i at repeated measure k. The same permutation test can be formulated as in Section 2.1 for the repeated measures dataset except now conclusions can be made for each time point, t. The benefit of using the longitudinal 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.

There is a slight multiple comparison problem due to the fact that we are separately calculating K, the number of repeated measures, 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 a Bonferroni correction is known to be very conservative and the objective of the analysis is for exploratory purposes and not to make definitive conclusions. When applying this method, and interpreting it's results, one can take into account the exploratory nature of the analysis and instead run sensitivity analyses to justify the conclusion of existence of a spatial cluster at a given time point if it consistently is shown across sensitivity analyses instead of making a conservation type one error reduction. This method is applied

on the Home Allergens and Asthma study in Section 4.

3 Simulation Study

[Table 1 about here.]

We conducted simulations calculating the type I error and power for the global cumulative geographic residual test. First we will analyze the results for checking type I error. Simulations were conducted by generating 1000 test studies where the region of each individual was randomly assigned in one of the 16 possible regions displayed in Figure 1. For our simulations we chose to treat region locations of individuals as fixed over time and the maximum number of regions to form a cluster to be 3, just to reduce computational complexity. We simulate a longitudinal dataset with exchangeable correlation structure and overall probability of having the outcome to be approximately 0.2. Details of this simulation is presented in the Appendix for longitudinal outcomes. For a single measure outcome the simulated dataset was derived from a bernoulli with $P(Y = 1) = 0.2$.

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 shown 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.

[Figure 1 about here.]

For the power calculations we simulated the longitudinal study population as described for the type I error. To create a single clusters we first considered an 8x8 unit-less area and divided the area into 16 equally sized squares of size 2x2 as depicted in Figure 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 $narea_i = \sum_{k=1}^{K} Y_{ik}$ where K is the number of repeated measures and $Iarea_i$ as a random sample from a bernoulli distribution with probability $(0.4 \times \frac{narea_i}{K})$. If $Iarea_i = 1$ then individual i is randomly assigned to be in grids 6 or 10 and if $Iarea_i = 0$ then randomly assigned to be any of any of the 16 grids. Power calculations are displayed in Table 1 for simulated datasets of size 1000. 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.

Overall, the proposed cumulative geographic residual test statistic for repeated measures holds the type I error rate and has relatively high power of finding a single cluster area. The power does increase as expected given more individuals with repeated measures, but does not increase given more repeated measures in the study. This is due to our design for power testing in which we are keeping the overall probability of having a case averaged over all repeated measures as 0.2 and holding the probability of being assigned to a cluster to be at most 0.4 for a given individual. Therefore, repeated measures with fewer time points for a given individual has a higher likelihood to be in a cluster, compared to repeated measured data with more time points. What is important to note is that power stays relatively high overall for a range of numbers of repeated measures and power increases when the number of individuals studied increases.

We did not show results for the proposed time-dependent cumulative geographic residual presented in Section 2.2. The results would be similar to the repeated measured analyses except may have more power, for the same number of outcome observations, since outcomes would now not be correlated. In the next section, all of the proposed methods will be applied to the longitudinal outcome wheeze.

4 Home Allergens and Asthma Study Analysis

We now apply the proposed methods 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 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 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. We prespecified the maximum neighborhood size to be 5 cities (N_R=5), but also ran sensititivity analyses for N_R equal to 1, 2, 3, and 4.

The study area is a diverse population with a range of socioeconomic 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, 2005). These areas also correspond to lower median family income areas indicating a relationship between disparity and allergic reaction.

[Figure 2 about here.]

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, Gold, and Li, 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 to the previous outcomes, particularly hayfever (Strunk, Ford, and Taggart, 2002). 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, Gold, and Li (2007), since the area has higher IgE levels (Litonjua et al, 2005) and lower median family income (Figure 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 socioeconomic and minority areas and therefore over time the cluster 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 dataset since the complete dataset results did not change results substantially, except p-values were higher since we have fewer subjects in the complete dataset.

[Table 2 about here.]

First we ran a GEE model without considering spatial clusters to assess change in percent wheeze by age. Due 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 independent correlation structure with robust standard errors from the sandwich estimator. Table 2 summarizes the results for two analyses from running the GEE model for the full dataset and complete follow-up subset. Note that estimates, and corresponding 95% confidence intervals, do not change significantly depending on the full dataset versus complete dataset indicating missingness may be 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, or if current exposure is the important predictor. To answer this question we ran two analyses (1) keeping location constant as birth location or (2) location as the current location at evaluation. When incorporating moving we must make assumptions about location of individuals that were lost to follow-up when performing the permutation test. We ran the method assuming two assumptions (1) the participant stayed at the last address that we have recorded or (2) the participant moved outside the study area.

[Figure 3 about here.]

The first spatial analysis conducted a global spatial clustering analysis using the CumResPerm method. There was marginally significant spatial clustering when assuming birth location (p-value = 0.10) or moving location (p-values = 0.06 (previous), 0.06 (outside)). This analyses assesses if one region has on average, across age groups, higher than expected rates of wheeze compared to other regions. We would not expect one area to be consistently a wheeze cluster across age groups, because the probability of wheeze is dependent on age and the relationship between predictors that exacerbate wheeze may change based on age. Therefore it is still important to assess which age groups are influencing the overall positive result.

We then conducted age-dependent spatial clustering of wheeze at each age point. Figure 3 displays the results of the CumResPerm test assuming that the participants location was constant at birth location while Figure 4 displays the results using the participants current location. The only significant cluster for both analysis was found at 18 months of age in Boston and the southern suburbs. This was a similar location that was found for the censored outcome asthma over all ages up to age four in previous analysis in this population (Cook, Gold, and Li, 2007). Therefore there is evidence of significant spatial clustering in Boston and the lower suburbs at 18 months of age, but the spatial cluster of wheeze did not persist into later ages. This result was not dependent on assuming birth location or move location which could be due to very few study participants move in the first 18 months of life (N=34

(7%)).

In sensitivity analyses using smaller maximum cluster sizes of N_R equal to 1, 2, 3, and 4 we found similar results with clustering at 18 months of age except for the statistically significant clusters tended to be smaller, but always included Boston. For the maximum cluster size set to one there was a statistically significant elevated rate of wheeze at 6 months of age in Boston alone. Therefore, our results indicate that the strongest spatial cluster occurred at 18 months of age in Boston and it's surrounding neighborhoods, but there was some indication that, in Boston alone, there may have been elevated rates of wheeze at 6 months of age.

[Figure 4 about here.]

5 Discussion

In this paper we have proposed the cumulative geographic residual permutation test (CumResPerm) for detecting spatial clustering of longitudinal outcomes. By utilizing the cumulative geographic residual methodology, we detected significant clustering of wheeze in urban Boston for age 18 months. Further research is being conducted to look into which exposures in urban Boston may be influencing this disease clustering, such as air pollution, mold, or rodent allergens.

We also performed type I error and power calculations for the CumResPerm test. Type I error was held at the appropriate α-level under the null of no clustering. By increasing number of individuals power was shown to substantially increase. There is still a need to run more simulations to explore the effect of prevalence of outcome and its change over time, number of repeated measures, and number of individuals on the power of the CumResPerm test.

The importance of using the time-dependent cumulative geographic residual method was presented as being able to pinpoint the location and time of significant clustering. However, there is a limitation for using this method in which it reduces the information to a stratified

analysis of time points. Therefore it does not utilize previous spatial locations to indicate exposures that may still influence the current outcome. This in a way reduces the longitudinal information into a cross-sectional analysis. Further research should be conducted into spatial clustering detection procedures that not only uses current location, but also incorporates previous time point information on the individual.

6 Acknowledgements

This work was supported by National Institutes of Health [RO1 AI/EHS 35786 to DRG; RO1 CA95747 to YL and AJC].

References

Brugge, D., Vallarion, J., Ascolillo, L., Osgood, N.D., Steinbach, S., and Spengler, J. Comparison of Multiple Environmental Factors for Asthmatic Children in Public Housing. *Indoor Air* 2003; **13**: 18–27.

Celedon, J., Litonjua, A., Weiss, S., and Gold, D. 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* 1999; **104**: 495–500.

Cook, A.J., Gold, D.R. and Li, Y. Spatial Cluster Detection for Censored Outcome Data. *Biometrics* 2007; **63**: 801–818.

Cook, A.J. and Li, Y. Rejoinder to: Spatial Cluster Detection for Censored Outcome Data. *Biometrics* 2008; **64(4)**: 1289–1292.

Cook, A.J., Gold, D.R. and Li, Y. Spatial cluster detection for repeatedly measured outcomes while accounting for residential history. *Biometrical* 2009; **51**: 540–549.

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

Finkelstein, J., Fuhlbrigge, A., Lozano, P., Grant, E., Shulruff, R., Arduino, K., and Weiss, K. Parent-reported Environmental Exposures and Environmental Control Measures for Children with Asthma. *Arch Pediatr Adolesc Med* 2002; **156**: 258–264.

Huang, L., Kulldorff, M., and Gregorio, D. A spatial scan statistic for Survival Data. *Biometrics* 2007; **63**: 109-118.

Jacqmin-Gadda, H., Commenges, D., Nejjari, C., and Dartigues, J.F. Test of geographical correlation with adjustment for explanatory variables: Application to dyspnoea in the elderly. *Statistics in Medicine* 1997; **16**: 1283-1297.

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

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

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

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. Variation in total and specific IgE: Effects of ethnicity and socioeconomic status. *J Allergy Clin Immunol* 2005; **115**: 751–757.

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

Strunk, R.C., Ford, J.G., and Taggart, V. Reducing disparities in asthma care: priorities for research-National Heart, Lung, and Blood Institute workshop report. *J Allergy Clin Immunol* 2002; **109**: 229–237.

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

Tango, T. and Takahashi, K. A flexibly shaped spatial scan statistic for detecting clusters. *Int J Health Geographics* 2005; **4**: 11.

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

Appendix

Simulated Longitudinal 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 .2,

Three Repeated Measures:

Generate n Multivariate Normal Outcome, $\mathbf{Z_i} \sim N_3\left((-0.1, 0, 0.1)^T, \mathbf{V}\right)$, 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, $\mathbf{Z}_i = N_4\left((-0.1, -0.05, 0.05, 0.1)^T, \mathbf{V}\right)$, 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, $\mathbf{Z}_i = N_5\left((-0.1, -0.05, 0, 0.05, 0.1)^T, \mathbf{V}\right)$, 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.

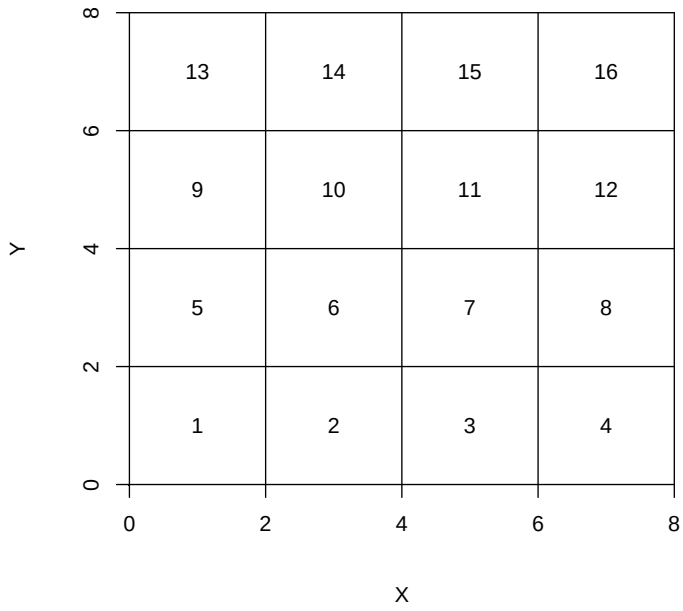

Figure 1: Study area used for the simulation study.

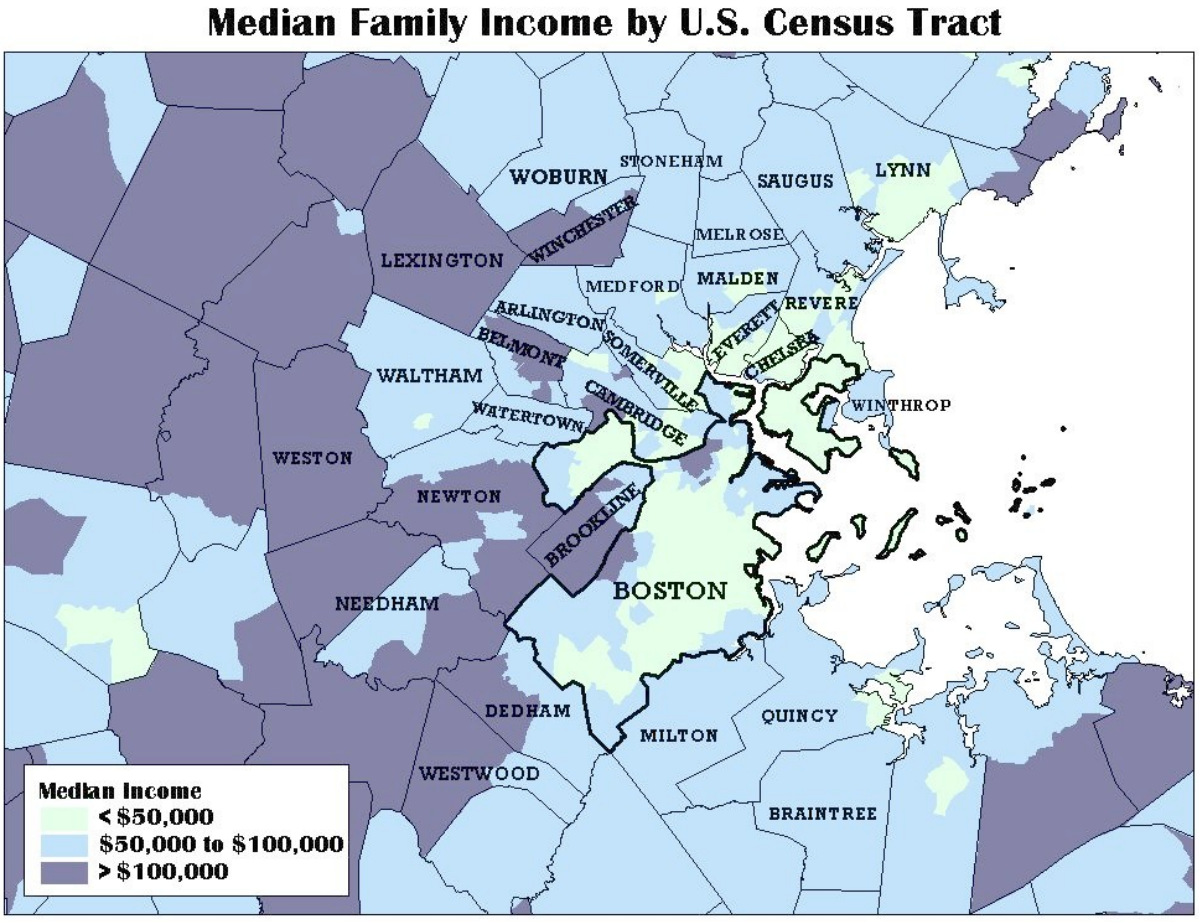

Figure 2: Indicated areas of low, medium, and high median family income by U.S. census tract in the study population area.

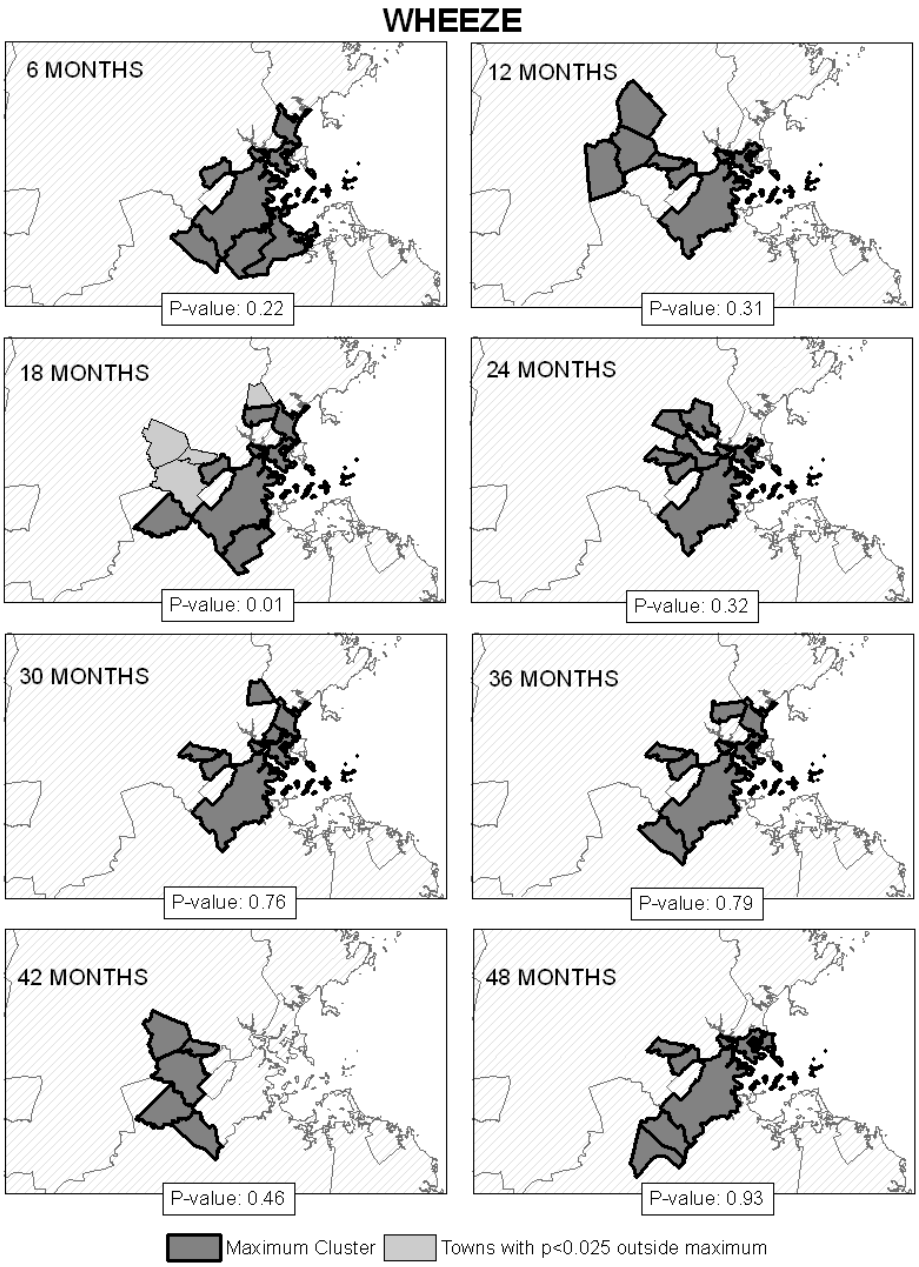

Figure 3: Indicated areas of cluster location of wheeze in the last 6 months for all time points using location at birth. For time-points without P¡0.025 the maximum cluster is the area with the most extreme test statistic, but for those time points with P¡0.025 a maximum cluster is shown along with all towns that were included in at least one cluster that had a P¡0.025.

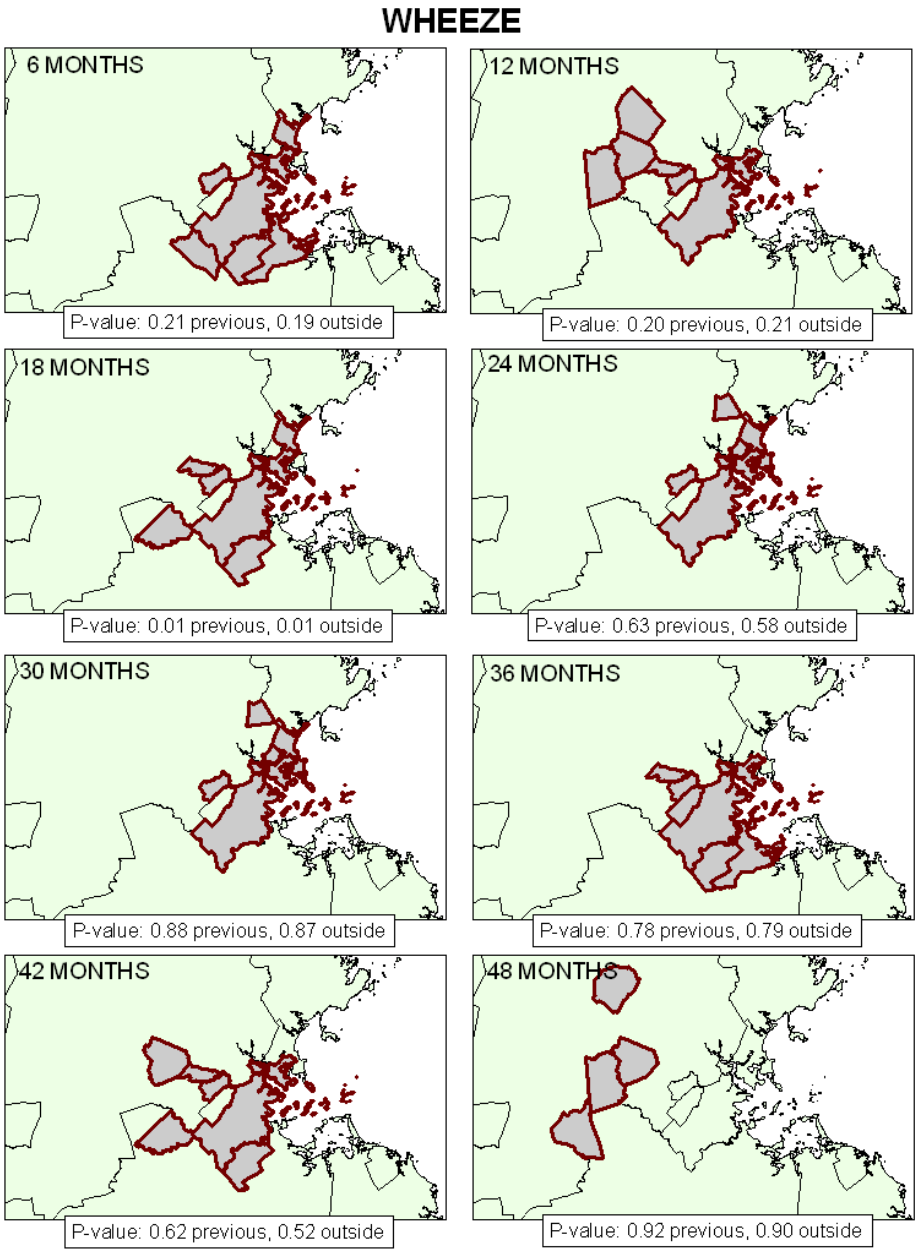

Figure 4: Indicated areas of cluster location of wheeze in the last 6 months for all time points using current location. Assumes that individuals that were loss to follow-up either stayed at there previous address or moved outside the area.

Table 1: Type I error and power of the Cumulative Geographic Residual Test for different sample sizes and number of repeated measures.

		Number of Time Points			
	N	1	3	4	5
Type I Error	100	0.039	0.041	0.059	0.047
	300	0.062	0.035	0.048	0.053
	500	0.060	0.050	0.051	0.049
Power	100	0.762	0.399	0.337	0.310
	300	0.998	0.891	0.801	0.759
	500	1.000	0.986	0.959	0.941

Table 2: 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)