Statistica Sinica Preprint No: SS-13-215wR2	
Title	A spatial scan statistic for compound poisson data, using the negative binomial distribution and accounting for population stratification
Manuscript ID	SS-13-215wR2
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.2013.215w
Complete List of Authors	Hsing-Ming Chang and Rhonda J. Rosychuk
Corresponding Author	Hsing-Ming Chang
E-mail	hsing-ming.chang@ubc.ca

Statistica Sinica: Preprint
doi:10.5705/ss.2013.215w

A SPATIAL SCAN STATISTIC FOR COMPOUND POISSON DATA, USING THE NEGATIVE BINOMIAL DISTRIBUTION AND ACCOUNTING FOR POPULATION STRATIFICATION

Hsing-Ming Chang[1] and Rhonda J. Rosychuk[2]

[1] *University of British Columbia,* [2] *University of Alberta*

Abstract: Since the interest in studying spatial relations in plant populations was raised in the 1950s, much effort has been devoted to the development of methods for spatial data analysis. One such development focused on techniques for detecting spatial clusters of cases and events in the biological sciences and epidemiology during the late 1980s and the following decade. More recently, research has examined detecting clusters of correlated count data associated with health conditions of individuals. Such a method allows researchers to examine spatial relationships of disease-related events rather than just incidents or prevalent cases. We introduce a spatial scan test that identifies clusters of events in a study region. Because an individual case may have multiple (repeated) events, we base the test on a special compound Poisson model. Based on this special class (a compound Poisson representation of the negative binomial distribution), advantages in computation over the general compound Poisson model that relies on a recursive formula are realized. We illustrate our method for cluster detection on emergency department visits, where individuals may make multiple asthma-related visits. We also demonstrate the spatial scan test adjusted by key population characteristics such as sex or age.

Key words and phrases: cluster detection, spatial scan, compound Poisson, negative binomial, stratification, surveillance.

1. Introduction

Spatial cluster detection (SCD) methods that provide tools to find proximities where certain events occur significantly more (or less) often than expected have become popular in the surveillance of diseases (e.g. Besag and Newell (1991), Kulldorff and Nagarwalla (1995), Nhu Le, Petkau and Rosychuk (1996), Jung, Kulldorff and Klassen (2007)). The spatial pattern of disease spread or other

Statistica Sinica: Preprint
doi:10.5705/ss.2013.215w

health outcomes is often of interest to health authorities; they collect substantial health data that can lead to important timely information when analysed by appropriate statistical methods. SCD methods project an objective and statistically sound approach for surveillance tools that monitor health across an administrative or geographic region.

Several authors have advanced SCD methods (see Marshall (1991), Lawson et al. (1999), Kulldorff et al. (2003a) and Kulldorff (2006) for reviews). Test may be classified as general or focused (Besag and Newell (1991)) and may detect clusters or the tendency to cluster. General tests are designed to detect clusters within the overall pattern of disease in a complete region and the cases of disease are assumed to occur at random: each individual in the population has an equal chance of developing the disease. For these tests, no specific alternative distribution for the cases is hypothesized.

Our health data applications typically have diverse population sizes and we focus our review on a few key general tests. Kulldorff and Nagarwalla (1995) introduced the spatial scan statistic (a maximum likelihood ratio statistic) that assesses if the individuals in a zone (circle) have greater disease risk than those outside the zone. The method has become popular in applications of spatial analysis and a standard for comparing with new research methods. Some recent developments in and adaptations to spatial scan methods include: Neill et al. (2005) propose an expectation-based Poisson method for larger outbreak sizes and is a variant to the test in Kulldorff (1997); Kulldorff, Fang and Walsh (2003b) propose a tree-based scan statistic for the purpose of data mining; Chan and Walther (2013) discuss the comparison of the optimality of test power of the scan statistic and the average likelihood ratio (ARL) statistic; a modified ARL statistic is proposed and its performance is studied in a more technical fashion; Wang and Yue (2013) propose a two-stage algorithm in which a binomial approximation is used in the second stage for spatial cluster detection; and Rosychuk and Chang (2013) provide a spatial scan to detect geographic areas with high numbers of disease-related events using a compound Poisson model.

Methods of Besag and Newell (1991) and Tango (1995) identify areas with a tendency to cluster. Besag and Newell (1991) combine regions with nearest neighbors and compare the number of neighbors that must be combined to contain

Statistica Sinica: Preprint
doi:10.5705/ss.2013.215w

a pre-specified number of cases. A chi-square statistic based on the discrepancy between observed and expected relative frequencies and a closeness measure is proposed by Tango (1995). Stone (1988) proposes a general test for elevation of disease risk around a point source and Morton-Jones, Diggle and Elliott (1999) extend Stone's test with covariate adjustment. Bailey (2001) discusses the general classes of problems in geographical epidemiology and reviews key statistical methods and the software to implement the methods available at that time. The paper is a good source for the history and background of spatial data modeling, and provides extensive references to this area of research. The SCD methods described are all based on detecting excess cases of disease and more recent developments and extensions have included the detection of excess events related to a particular condition or disease (e.g. Rosychuk, Huston and Prasad (2006) and Rosychuk and Stuber (2010)). These excess events tests are based on a strategy similar to that in Besag and Newell (1991).

Very few SCD methods are appropriate for diverse population sizes consider that stratification of the population by characteristics may influence the geographic pattern of disease. For example, Cuzick and Edwards (1990) demonstrate the nearest neighbours test with covariate adjustment, Kulldorff (1997) provides the spatial scan test in the Poisson model adjusted by race, and Besag and Newell (1991) and Rosychuk, Huston and Prasad (2006) offer tests that adjust for covariates through stratification. We consider a spatial scan for compound Poisson data to detect geographic areas with excess events (such as emergency department (ED) visits or physician visits), adjusting for key population characteristics (e.g. sex). Our method uses a negative binomial distribution (in contrast to the recursions as in Rosychuk and Chang (2013)) for disease-related events as the primary unit of analysis for SCD rather than analyzing data of individuals in a case/non-case fashion. Such a model enables us to detect geographical clusters of events when individuals in a population may have multiple correlated events related to a disease or condition. In Section 2, we introduce the model and test statistic. In Section 3, we extend the approach to stratify by important population characteristics. We describe our administrative health data on asthma and present case studies to illustrate our methodology, stratified by sex and age group, in Section 4. Some concluding remarks and future research ideas are given

Statistica Sinica: Preprint
doi:10.5705/ss.2013.215w

in Section 5.

2. Methodology

We assume that administrative health data can be collected from I non-overlapping geographic subregions and that each subregion has a health centre as centroid (e.g., geographic or population based). A zone Z, defined by a circular spatial scan window of radius r and center at the coordinate of a centroid, consists of only and all individuals in those subregions whose centroids lie inside the circle (Kulldorff and Nagarwalla (1995)).

For a two-dimensional scan test, we choose an upper bound r_i^*, $i = 1, \cdots, I$, on the radius of the circular scan window such that the population size of any zone defined by the window centered at centroid i does not exceed β percent of the total population in the study region. The choice of the upper bound on r_i should be made prior to analysis (Kulldorff and Nagarwalla (1995)). For each i, all test zones can be generated by combining nearest subregions with subregion i by varying the radius of the defining circle from 0 to r_i^* and from centroid to centroid. Each zone coincides with a single subregion when $r_i = 0$ for all i. Zones with such a definition have irregular geographic boundaries that depend on the size and shape of those subregions whose centroids lie inside the spatial scan window.

Figure 2.1 shows a section of the Alberta sub-regional health authorities with the coordinates of the centroids handpicked (non-official) for illustrative purpose. Circular scan windows of various radii are centered at the centroids with region ID 60, 53, and 36. Starting from each centroid, a new test zone is formed when a new neighboring centroid is enclosed by the scan window. The value of r_i^* varies for each centroid i depending on the population size of cell i and of its nearby neighbors and the chosen β.

2.1 Notation

Let Z be a collection of distinct subregions (administrative areas) in a geographic region R. For each predetermined $\mathrm{Z} \subset \mathrm{R}$, let $\mathrm{Z}' \subset \mathrm{R}$ denote the compliment of Z. Let there be a total of I non-overlapping and contiguous subregions S_i (such as districts of land space), $i = 1, \cdots, I$, in R (such as a state or a province).

Let the variable C_{ik}, with observed value c_{ik}, be the number of individuals observed with k events in subregion S_i ($k \in \mathbb{N}$, $i = 1, \ldots, I$). Let $C_i = \sum_{k \in \mathbb{N}} C_{ik}$,

Statistica Sinica: Preprint
doi:10.5705/ss.2013.215w

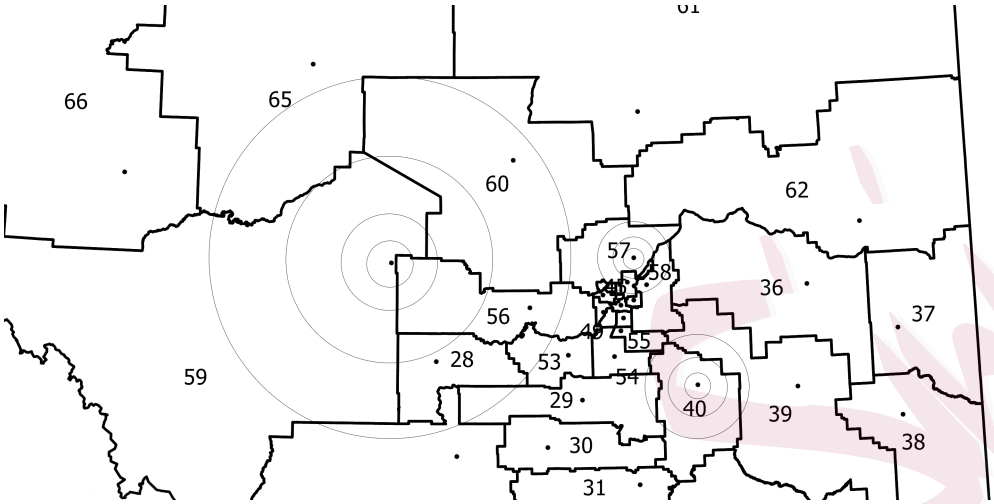

Figure 2.1: A section of Alberta subregional health authorities.

with observed value c_i, be the total number of individuals with at least one event in subregion S_i. Let $C_{\mathrm{Z}} = \sum_{\substack{i=1 \\ \mathrm{S}_i \in \mathrm{Z}}}^{I} C_i$, with observed value c_{Z}, be the number of individuals (cases) with at least one event in Z and, similarly for outside the test zone, $C_{\mathrm{Z}'} = \sum_{\substack{i=1 \\ \mathrm{S}_i \notin \mathrm{Z}}}^{I} C_i$ with observed value $c_{\mathrm{Z}'}$. Let $C = C_{\mathrm{Z}} + C_{\mathrm{Z}'}$ denote the total number of cases in R. We wish to detect zones in R that have significantly higher numbers of events relative their surrounding regions. The spatial scan statistic we propose is based on a likelihood ratio test in the same spirit of Kulldorff and Nagarwalla (1995).

We assume the population size of subregion S_i can be measured and is denoted by n_i for $i = 1, \ldots, I$, and that $C_i \sim \mathrm{POI}(\lambda_i n_i)$ where $\lambda_i > 0$ are standardized Poisson intensities. We consider that within a time period, say fiscal year, individuals will only have event(s) within the subregion of their residence. Of the individuals with at least one event, let $X_{i\ell}$ denote the number of event(s) generated by the ℓth individual in subregion S_i, $\ell = 1, \ldots, C_i$. The density distribution of $X_{i\ell}$ can be arbitrary depending on the context.

The total number of events from the population of subregion S_i can be written as

$$U_i = \sum_{\ell=1}^{C_i} X_{i\ell}$$

Statistica Sinica: Preprint
doi:10.5705/ss.2013.215w

where it is reasonable to assume that C_i is independent of $X_{i\ell}$ for $\ell = 1, \ldots, C_i$ and $i = 1, \ldots, I$. Thus, C_i could be used to represent, over a fixed period of time, the number of people having respiratory symptoms in subregion i and $X_{i\ell}$ the number of times the ℓth individual of subregion i visited for hospital emergency service with such symptoms. Any individual with symptoms of asthma would present to an ED because of a real or perceived need for health care. The individual would not know *a priori* the number of ED visits made over the time period and region, and would therefore not have such information available to alter the decision to seek health care.

With this formulation, U_i is a compound Poisson random variable and the distribution of U_i is rarely tractable. In general, the recursive formula (Panjer (1981))

$$\begin{aligned} \Pr(U_i = 0) &= e^{-\lambda_i n_i}, \\ \Pr(U_i = u_i) &= \frac{\lambda_i n_i}{u_i} \sum_{x=1}^{u_i} x f(x;\theta_i) \Pr(U_i = u_i - x) \quad u_i = 1, 2, \cdots \end{aligned} \tag{2.1}$$

can be used to obtain its distribution. For special classes of the compounding distribution $f(x;\theta_i)$, more efficient recursions are derived in De Pril (1986a) and Chadjiconstantinidis and Pitselis (2009). Large data sets make computations involving a recursion formula lengthy. Suitable exact distributions of U_i (due to a suitable selection of $f(x;\theta_i)$) are desirable, especially when numerical optimization is involved, to avoid recursions so that computation can be done more efficiently.

We assume in this paper that $X_{i\ell}$ is discrete and follows a Logarithmic distribution with density

$$\Pr(X_{i\ell} = x) = f(x;\theta_i) = \begin{cases} \dfrac{-\theta_i^x}{x \ \log(1-\theta_i)} & x = 1, 2, \cdots, \ 0 < \theta_i < 1 \\ 0 & \text{elsewhere} \end{cases} \tag{2.2}$$

which has mean $-\theta_i[(1-\theta_i)\log(1-\theta_i)]^{-1}$. Supposing $\lambda_i n_i = -\sigma_i n_i \log(1-\theta_i)$, the probabilities $\Pr(U_i = u_i)$ can be obtained using the negative binomial distribution (see Quenouille(1949) and Gurland (1957)),

$$\Pr(U_i = u_i) = \binom{u_i + \sigma_i n_i - 1}{u_i} \theta_i^{u_i} (1-\theta_i)^{\sigma_i n_i} \quad u_i = 0, 1, 2, \cdots \tag{2.3}$$

Statistica Sinica: Preprint
doi:10.5705/ss.2013.215w

and are denoted by $\mathrm{NB}(\sigma_i n_i, \theta_i)$ with mean $\sigma_i n_i \theta_i (1-\theta_i)^{-1}$, where $\sigma_i > 0$ and $0 < \theta_i < 1$. Given σ_i and θ_i, $\Pr(U_i = 0)$ can then be easily calculated using (2.3). The likelihood function for the negative binomial model can be written as

$$L(\sigma_1, \cdots, \sigma_I, \theta_1, \cdots, \theta_I) \propto \prod_{\substack{i=1 \\ C_i>0}}^{I} \binom{u_i + \sigma_i n_i - 1}{u_i} \theta_i^{u_i} (1-\theta_i)^{\sigma_i n_i} \tag{2.4}$$

Under the null hypothesis $\mathrm{H_0}$: $\lambda_i = \lambda$ and $\theta_i = \theta$ for all i, if $\lambda n_i = -\sigma n_i \log(1-\theta)$, the likelihood can be expressed as

$$L(\hat{\sigma}, \hat{\theta}) = \prod_{i=1}^{I} \binom{u_i + \hat{\sigma} n_i - 1}{u_i} \hat{\theta}^{u_i} (1-\hat{\theta})^{\hat{\sigma} n_i} \tag{2.5}$$

where $\hat{\sigma}$ and $\hat{\theta}$ are maximum likelihood estimates (mle's) of σ and θ. In direct joint estimation of σ and θ using (2.5), record level data on the individual event numbers is not used in estimating θ. For instance, if we observe in a subregion S_i having five cases ($C_i = 5$), then both $\{X_{i1} = 4, X_{i2} = 2, X_{i3} = 3, X_{i4} = 3, X_{i5} = 7\}$ and $\{X_{i1} = 2, X_{i2} = 2, X_{i3} = 2, X_{i4} = 2, X_{i5} = 11\}$ give $U_i = 19$ and yield the same $\hat{\sigma}$ and $\hat{\theta}$ based on (2.5) alone. We suggest that $\hat{\theta}$ be obtained first by the maximum likelihood method that maximizes

$$\log \prod_{i=1}^{I} \prod_{\ell=1}^{C_i} \frac{-\hat{\theta}^{x_{i\ell}}}{x_{i\ell} \log(1-\hat{\theta})},$$

then $\hat{\theta}$ substituted into (2.5) to obtain a mle of σ that maximizes $L_{\hat{\theta}}(\hat{\sigma})$.

Under the alternative hypothesis $\mathrm{H_a}$: $\lambda_i = \mu$, $\theta_i = p$ and $\mu n_i = -\gamma n_i \log(1-p)$ for $\mathrm{S}_i \in \mathrm{Z}$, $\lambda_i = \nu$, $\theta_i = q$ and $\nu n_i = -\delta n_i \log(1-q)$ for $\mathrm{S}_i \notin \mathrm{Z}$, $\mu \neq \nu$, $\gamma p (1-p)^{-1} > \delta q (1-q)^{-1}$, and $p \neq q$, the probabilities $\Pr(U_i = u_i)$ can be obtained by separate negative binomial distributions, $\mathrm{NB}(\gamma n_i, p)$ for $\mathrm{S}_i \in \mathrm{Z}$ and $\mathrm{NB}(\delta n_i, q)$ for $\mathrm{S}_i \notin \mathrm{Z}$, $i = 1, \ldots, I$. We can interpret the inequality $\gamma p (1-p)^{-1} > \delta q (1-q)^{-1}$ as: per fixed population size, the expected number of events incurred inside a test zone is higher than that outside the zone. Conditional on $U_i = u_i$ for $i = 1, \ldots, I$, the likelihood under the alternative becomes

$$L_{\hat{p},\hat{q}}(\mathrm{Z}, \hat{\gamma}, \hat{\delta}) = \prod_{\substack{i=1 \\ \mathrm{S}_i \in \mathrm{Z}}}^{I} \binom{u_i + \hat{\gamma} n_i - 1}{u_i} \hat{p}^{u_i} (1-\hat{p})^{\hat{\gamma} n_i} \prod_{\substack{i=1 \\ \mathrm{S}_i \notin \mathrm{Z}}}^{I} \binom{u_i + \hat{\delta} n_i - 1}{u_i} \hat{q}^{u_i} (1-\hat{q})^{\hat{\delta} n_i} \tag{2.6}$$

Statistica Sinica: Preprint
doi:10.5705/ss.2013.215w

where, conditioned on Z, $\hat{p}$ and $\hat{q}$ are mle's that maximize

$$\log \prod_{\substack{i=1 \\ S_i \in Z}}^{I} \prod_{\ell=1}^{C_i} \frac{-\hat{p}^{x_{i\ell}}}{x_{i\ell}\ \log(1-\hat{p})} \quad \text{and} \quad \log \prod_{\substack{i=1 \\ S_i \notin Z}}^{I} \prod_{\ell=1}^{C_i} \frac{-\hat{q}^{x_{i\ell}}}{x_{i\ell}\ \log(1-\hat{q})},$$

respectively, and $\hat{\gamma}$ and $\hat{\delta}$ are mle's that maximize $L_{\hat{p},\hat{q}}(Z, \gamma, \delta)$.

If $C_i = 0$ occurs in an observed data set, then S_i is usually a subregion with a small population size n_i. We consider S_i to be at a low risk of being a possible event cluster and do not include it separately as a test zone for cluster detection analysis; however, the inclusion and exclusion of S_i alone as a test zone under such criteria, later in assessing the test significance, depends on the Monte Carlo simulation and is irrelevant to the observed data set.

2.2 Likelihood Ratio Test Statistic

The likelihood ratio test statistic of our choice is

$$\eta = \frac{\max\limits_{Z \subset R} L_{\hat{p},\hat{q}}(Z, \hat{\gamma}, \hat{\delta})}{L_{\hat{\theta}}(\hat{\sigma})} \quad \text{or} \quad \eta = \max_{Z \subset R} \log \frac{L_{\hat{p},\hat{q}}(Z, \hat{\gamma}, \hat{\delta})}{L_{\hat{\theta}}(\hat{\sigma})} \tag{2.7}$$

with $\hat{\sigma}$ denoting the mle of σ under the null hypothesis. Note that, in practice, we only consider η from zones such that the condition

$$\frac{\hat{\gamma}\hat{p}}{1-\hat{p}} > \frac{\hat{\delta}\hat{q}}{1-\hat{q}} \quad \text{or} \quad \phi = \frac{\hat{\gamma}\hat{p}(1-\hat{q})}{\hat{\delta}\hat{q}(1-\hat{p})} > 1 \tag{2.8}$$

is satisfied since our interest is in finding regions of high expected number of events relative to their surrounding areas.

As with Kulldorff (1997), the exact distribution of η in analytical form is difficult to obtain. Monte Carlo simulation is to be employed to assess the significance of an observed value of η under the null hypothesis by taking the following steps. Each replicate of the data set is generated conditional on $C = c$ and $X_j = x_j$, the number of events generated by the jth case in the study region regardless of original cell reference, for $j = 1, \ldots, c$. Under the assumption that $C_i \sim \text{POI}(\lambda n_i)$ and the distribution of $X_{i\ell}$ is identical across all i and ℓ, in each data replication a new cell reference is assigned randomly to each case by relative frequencies n_i/n to ensure spatial randomness of $X_{i\ell}$.

1. Conditioning on $C = c$ and $X_j = x_j$, sample randomly a subregion ID in $\{1, 2, \ldots, 70\}$ for each x_j, $j = 1, \cdots, c$. The sampling distribution has

Statistica Sinica: Preprint
doi:10.5705/ss.2013.215w

weights n_i/n for $i = 1, \cdots, I$. Depending on the generated subregion ID for each x_j, new C_i and $X_{i\ell}$, hence U_i, are generated for $\ell = 1, \ldots, C_i$ and $i = 1, \ldots, I$.

2. Calculate the test statistic η as defined in (2.7).

3. Repeat Steps 1 and 2 for 999 trials and record the test statistic of each simulation trial.

4. Rank the 999 simulated likelihood ratio statistics and the observed statistic η from the data.

In our model, the likelihood $L_{\hat{\theta}}(\hat{\sigma})$ is not a constant under the null hypothesis over each simulation trial, because $L_{\hat{\theta}}(\hat{\sigma})$ depends on $U_i = u_i$ for $i = 1, \cdots, I$, and U_i's are not fixed in the simulation trials. Therefore, the numerator and the denominator of (2.7) need to be computed in Step 2 of each trial. The hypothesis test can be considered significant at the 1000α percent level if the value of the observed η calculated from data is among the 1000α (an integer) highest of the 1000 ranked statistics in Step 4. A significant test indicates that the collection of subregions which yields the observed η in the spatial scan test is the most likely cluster having higher expected numbers of events per fixed population size. Other zones that have nonoverlapping subregions with the most likely cluster and have high values of the test statistic under condition (2.8) should be examined for the possibility of being secondary clusters.

3. Accounting for Stratification

Suppose the population in geographic areas differs by a key characteristic B (for instance, sex, age group, or ethnicity) related to the condition under examination, and that we have such data on the entire population. Analyses can be adjusted for the population distribution on the key characteristic through stratification. If B with b categories is anticipated to have an effect on the distribution of correlated event count, we define $X_{ij\ell}$ to be the number of event generated by the ℓth individual who has category j of the characteristic of interest in subregion S_i for $\ell = 1, \ldots, C_{ij}$ and $j \in \{1, \cdots, b\}$, where $C_{ij} = \sum_{k \in \mathbb{N}} C_{ijk}$ with C_{ijk} the number of individuals in category j with k events in subregion S_i. The

Statistica Sinica: Preprint
doi:10.5705/ss.2013.215w

density distribution of $X_{ij\ell}$ is now assumed to be

$$\Pr(X_{ij\ell}=x)=f(x;\theta_{ij})=\begin{cases}\dfrac{-\theta_{ij}^{x}}{x\ \log(1-\theta_{ij})} & x=1,2,\cdots,\ 0<\theta_{ij}<1\\ 0 & \text{elsewhere}\end{cases}$$

The likelihood function in (2.4) under the influence of B is

$$L(Z,\sigma_{11},\cdots,\sigma_{Ib},\theta_{11},\cdots,\theta_{Ib})\propto\prod_{i=1}^{I}\prod_{\substack{j=1\\ C_{ij}>0}}^{b}\binom{u_{ij}+\sigma_{ij}n_{ij}-1}{u_{ij}}\theta_{ij}^{u_{ij}}(1-\theta_{ij})^{\sigma_{ij}n_{ij}},$$

where all parameters and variables have their original definition and are under the same assumptions. A parameter or variable with indices i and j now is associated with the ith geographical region and jth level of the population characteristic.

Under the null hypothesis $\mathrm{H_0}$: $\lambda_{ij}=\lambda_j$ and $\theta_{ij}=\theta_j$ for all i, if $\lambda_j n_{ij}=-\sigma_j n_{ij}\log(1-\theta_j)$, the likelihood accounting for stratification can be expressed as

$$L_{\hat\theta_1,\cdots,\hat\theta_b}(\hat\sigma_1,\cdots,\hat\sigma_b)=\prod_{i=1}^{I}\prod_{j=1}^{b}\binom{u_{ij}+\hat\sigma_j n_{ij}-1}{u_{ij}}\hat\theta_j^{u_{ij}}(1-\hat\theta_j)^{\hat\sigma_j n_{ij}}$$

where $\hat\theta_j$ are mle's of θ_j that maximize

$$\log\prod_{i=1}^{I}\prod_{\ell=1}^{C_{ij}}\frac{-\hat\theta_j^{x_{ij\ell}}}{x_{ij\ell}\ \log(1-\hat\theta_j)}$$

for each j, and $\hat\sigma_1,\cdots,\hat\sigma_b$ maximize $L_{\hat\theta_1,\cdots,\hat\theta_b}(\hat\sigma_1,\cdots,\hat\sigma_b)$.

If we assume the alternative hypothesis $\mathrm{H_a}$: $\lambda_{ij}=\mu_j$, $\theta_{ij}=p_j$ and $\mu_j n_{ij}=-\gamma_j n_{ij}\log(1-p_j)$ for $\mathrm{S}_i\in\mathrm{Z}$, $\lambda_{ij}=\nu_j$, $\theta_{ij}=q_j$ and $\nu_j n_{ij}=-\delta_j n_{ij}\log(1-q_j)$ for $\mathrm{S}_i\notin\mathrm{Z}$, $\mu_j\neq\nu_j$ and $p_j\neq q_j$ for at least one j in $\{1,\ldots,b\}$, and $\sum_{j=1}^{b}\gamma_j p_j(1-p_j)^{-1}>\sum_{j=1}^{b}\delta_j q_j(1-q_j)^{-1}$, the likelihood under $\mathrm{H_a}$ is

$$\begin{aligned}&L_{\hat p_1,\cdots,\hat p_b,\hat q_1,\cdots,\hat q_b}(\mathrm{Z},\hat\gamma_1,\cdots,\hat\gamma_b,\hat\delta_1,\cdots,\hat\delta_b)\\&=\prod_{\substack{i=1\\ \mathrm{S}_i\in\mathrm{Z}}}^{I}\prod_{j=1}^{b}\binom{u_{ij}+\hat\gamma_j n_{ij}-1}{u_{ij}}\hat p_j^{u_{ij}}(1-\hat p_j)^{\hat\gamma_j n_{ij}}\prod_{\substack{i=1\\ \mathrm{S}_i\notin\mathrm{Z}}}^{I}\prod_{j=1}^{b}\binom{u_{ij}+\hat\delta_j n_{ij}-1}{u_{ij}}\hat q_j^{u_{ij}}(1-\hat q_j)^{\hat\delta_j n_{ij}}\end{aligned}$$

where, conditioned on Z, $\hat p_j$ and $\hat q_j$ maximize the log likelihoods

$$\log\prod_{\substack{i=1\\ \mathrm{S}_i\in\mathrm{Z}}}^{I}\prod_{\ell=1}^{C_i}\frac{-\hat p_j^{x_{ij\ell}}}{x_{ij\ell}\ \log(1-\hat p_j)}\quad\text{and}\quad\log\prod_{\substack{i=1\\ \mathrm{S}_i\notin\mathrm{Z}}}^{I}\prod_{\ell=1}^{C_i}\frac{-\hat q_j^{x_{ij\ell}}}{x_{ij\ell}\ \log(1-\hat q_j)},$$

Statistica Sinica: Preprint
doi:10.5705/ss.2013.215w

respectively, for each j and $\hat{\gamma}_1, \cdots, \hat{\gamma}_b, \hat{\delta}_1, \cdots, \hat{\delta}_b$ maximize $L_{\hat{p}_1,\cdots,\hat{p}_b,\hat{q}_1,\cdots,\hat{q}_b}(\mathrm{Z}, \hat{\gamma}_1, \cdots, \hat{\gamma}_b, \hat{\delta}_1, \cdots, \hat{\delta}_b)$. We can interpret the inequality in the alternative hypothesis as: under the influence of population characteristic B, per fixed population size, the expected number of events incurred inside a test zone is higher than that outside the zone.

As the number of categories of characteristic B increases, the chance of not observing a case and event for some strata in some subregion(s) grows. For example, we may observe in subregion S_i that 5 individuals (all female) made presentations to the ED because of domestic violence. If population stratification is categorized by sex, p_{male} cannot be estimated for S_i. In such a situation, we substitute $\hat{\theta}_j$ for $\hat{p}_j$.

The likelihood ratio test statistic is now

$$\eta = \max_{\mathrm{Z}\subset\mathrm{R}} \log \frac{L_{\hat{p}_1,\cdots,\hat{p}_b,\hat{q}_1,\cdots,\hat{q}_b}(\mathrm{Z}, \hat{\gamma}_1, \cdots, \hat{\gamma}_b, \hat{\delta}_1, \cdots, \hat{\delta}_b)}{L_{\hat{\theta}_1,\cdots,\hat{\theta}_b}(\hat{\sigma}_1, \cdots, \hat{\sigma}_b)},$$

where $\hat{\sigma}_1, \cdots, \hat{\sigma}_b$ are mle's of $\sigma_1, \cdots, \sigma_b$ under the null. The procedure to assess the significance of an observed η is the same as before, although the population characteristic B should be carefully taken into account when spatially re-arranging the case data.

If the population differs by key characteristics (related to disease) across the geographic areas, then stratification by the key characteristics is usually preferred. If there are several strata variables and many test zones, then there may be substantive areas where an event is not observed. The stratified analysis may then yield unreliable parameter estimates and significance calculations.

4. Application

We illustrate our spatial scan on asthma-related ED presentations by children and youth (age $\leq$19 years) based on ICD-9-CM or ICD-10-CM codes in the western Canadian province of Alberta during six fiscal years (April 1, 1999, to March 31, 2005), see Rosychuk et al. (2010a and 2010b). Each ED presentation during the study period is considered to be an event and a case is defined as an individual who had at least one ED presentation for asthma in Alberta during the study period. All tables and figures referenced in this section can be found in the online supplementary document. We used 70 subregional health authorities (HAs) as the cells (see Figure 1). A subset of Aboriginals and Welfare recipients

Statistica Sinica: Preprint
doi:10.5705/ss.2013.215w

were selected for illustrating the methods when comparing the compound Poisson model and the negative binomial model. The full dataset was then used for analyses with sex and age stratification. We used four age groups to stratify the population: *preschool* (from birth to the age of 4); *primary school* (from 5 to 9); *preadolescence* (from 10 to 14); and *adolescence* (from 15 to 19). Five year age groups are typical for medical and epidemiological studies although other choices could be made. The analyzed data set contains individual record level information including age, sex, HA of residence, and fiscal year at the time of ED visit. The cell population sizes are reported in Table 1, and the aggregated case and event counts are tabulated in Table 2. In all the analyses, the radius size r_i, $i = 1, \dots 70$, varied from 0 to an upper limit, r_i^*, which restricted test zones to contain no more than $\beta = 7\%$ of the province's population size of a fiscal year.

4.1 Compound Poisson Model vs. Negative Binomial Model

We first examine each fiscal year separately using both the negative binomial model (2.3) and the spatial scan for events based on the compound Poisson model (Rosychuk and Chang (2013)). This examination allows us to compare the approaches, in particular the reduced computational time of the negative binomial model, before considering the effect of stratification. When $X_{i\ell}$ is discrete and follows a Logarithmic distribution, the results of event cluster detection analysis results based on the negative binomial model can be found in Table 3. The most likely cluster of each year is reported with other calculated statistics including computation time (CT) in hours. All computer programs were implemented in MATLAB (2012a), of which the command `fmincon` is used for constrained optimization to seek the maximizer of an objective loglikelihood function on a PC computer equipped with an Intel i7-2600 processor and 32 GB of memory. We also report candidate clusters with likelihood ratios in the top 50 rankings that do not intersect the most likely cluster as secondary clusters. It is common that many of the secondary candidate clusters intersect with each other. For candidate zones that have at least one common cell ID, we manually report the zone with the highest ranking likelihood ratios. Secondary clusters may be defined in different ways, and one may report the cell(s) which appear the most often among all secondary candidate clusters to the health authority for further investigation. When it is assumed that $X_{i\ell}$ is discrete and follows a zero-truncated

Statistica Sinica: Preprint
doi:10.5705/ss.2013.215w

Poisson distribution, the results of event cluster detection analysis based on the compound Poisson model can be found in Table 4.

Analyses based on the two assumptions of the compounding distribution may suggest very different primary event clusters. The commonly reported secondary candidate clusters would suggest that these subregions deserve further investigation. For instance, in the 2001/2002 fiscal year, zone $\{18, 19, 20, 21, 22, 26\}$ is detected as the most likely event cluster and $\{36, 62\}$ and $\{27\}$ are significant secondary candidate clusters assuming a zero-truncated Poisson distribution as the compounding distribution. Meanwhile, zone $\{63, 64\}$ is the most likely cluster and $\{27\}$ is a secondary candidate assuming a Logarithmic distribution as the compounding distribution. Similarly, $\{27\}$ appeared to be a secondary candidate under both model assumptions in 2002/2003. It appears that the results of the clustering detection is quite sensitive to model selection of the compounding distribution in describing the intra-person correlation. This diagnostic tool may require more subjective judgement when reporting potential event clustering communities to health authorities for further investigation.

As a comparison, instead of assuming that $X_{i\ell}$ follows a zero-truncated Poisson distribution with density

$$\Pr(X_{i\ell} = x) = f(x; \theta_i) = \begin{cases} \dfrac{\theta_i^x}{x!(e^{\theta_i} - 1)} & x = 1, 2, \cdots \quad \text{and } \theta_i > 0 \\ 0 & \text{elsewhere} \end{cases} \tag{4.1}$$

we report in Table 5 the probabilities of zero-truncated Poisson distributions corresponding to various $\hat{\theta}$ values estimated by the method of maximum likelihood under the null hypothesis based on the event per case data from years 1999/2000 to 2004/2005 (from the left to the right), and in Table 6 the probabilities of Logarithmic distributions. Clearly the zero-truncated Poisson distributions approach zero faster than the Logarithmic distributions as x increases. When comparing the distributions in the two tables to the relative frequencies of the number of events per individual for each fiscal year reported in Table 7, even without a formal goodness of fit test the distributions are closer to the ones in Table 6 in each year, and there is a good reason to believe that the Logarithmic compound model is preferable in our studied data set.

4.2 Stratification by Sex and Age

Statistica Sinica: Preprint
doi:10.5705/ss.2013.215w

To have suitable data for demonstrating event cluster detection accounting for population stratification under the negative binomial model, we focused on the entire children and youth population of age $\leq$19 when ED presentations were made. The event cluster detection results, without stratification, appear in Table 8, and the results when sex and age group are used as strata variables appear in Tables 9 and 10, respectively.

It is interesting to observe that by examining further some of the secondary significant event clusters in each fiscal year that, when not accounting for population stratification, the detected primary and secondary event clusters together are quite consistent to cover those regions in the most likely cluster detected while considering age or sex stratification. The required computing time is considerably less, in our experience, when analyses are carried out without stratification. The most likely reason is that the optimization algorithm utilized by MATLAB (`fmincon`) is quite fast, however, the number of repetitions of initiating the optimization algorithm increases significantly as more characteristics are used to stratify the population and creates more subsets of data in analysis.

5. Discussion

We have explored the use of a compound Poisson model in analyzing disease-related events. Whereas our earlier work (Rosychuk and Chang (2013)) used a zero-truncated Poisson distribution for the compounding distribution, here we took a Logarithmic compounding distribution and, subsequently, a negative binomial. Across all studies, the negative binomial had a definite advantage in computing time, especially when case and event counts are large in a study region. We have considered variability of the population by accounting for population stratification before analysis. It is advised to carry out some testing to determine if stratification is needed, and a goodness of fit test on the choice of the compounding distribution, before adopting our spatial scan test. Further work will investigate formal tests to assess the choice of compounding distribution.

A limitation of the proposed test is that it lacks the consideration of overdispersion as described in Loh and Zhu (2007) and Zhang et al. (2012). Overdispersion can be caused by correlations existing among the λ_i and it causes inflation of type I error in the classical spatial scan test analyzing cases alone, and assuming $C_i \sim \text{POI}(\lambda_i n_i)$. If spatial correlations exist, our spatial scan test can

Statistica Sinica: Preprint
doi:10.5705/ss.2013.215w

inherit the problem of inflation in type I error. A possible solution is to adopt the quasi-Poisson model as in Zhang et al. (2012), where C_i is shown to marginally follow a negative binomial distribution, in a specific setting, governed by λ_i and dispersion parameter $n_i(\phi - 1)^{-1}$ when $1 < \phi < 2$. When C_i follows a negative binomial distribution and $X_{i\ell}$ is discrete, results in Panjer (1981) may address the issue; it deserves further investigation in detail.

Tests for detecting case clusters remain the most popular for their fast computing time and performance in certain detection models. We have demonstrated the advantages of tests based on the compound Poisson model which takes into consideration the intra-person correlation of disease-related events. The goal is to provide additional approaches to current methods and offer strategies when confronted with spatial event data having a compound Poisson structure. Whether the use of an empirical compounding distribution will improve the performance of our approach of spatial scan test remains an open question. It certainly provides an alternate choice to the compound Poisson model and is worth further investigation.

Acknowledgment The authors would like to thank the co-editors and the three anonymous reviewers for their comments that helped to strengthen areas of this paper and improve its readability. Helpful questions hinted at future research directions in improving our proposed spatial scan test. The work was funded by an operating grant from the Canadian Institutes of Health Research. Rhonda J. Rosychuk is salary supported by Alberta Innovates-Health Solutions (AI-HS) as a Health Scholar. The authors thank Dr. Dandan Luo for her initial suggestions.

This study is based in part on data provided by Alberta Health. The interpretation and conclusions contained herein are those of the researchers and do not necessarily represent the views of the Government of Alberta. Neither the Government nor Alberta Health express any opinion in relation to this study.

Supplementary Document The online Supplement file contains the tables and figures referenced in Section 4.

References

Statistica Sinica: Preprint
doi:10.5705/ss.2013.215w

Bailey, T. C. (2001). Spatial statistical methods in health. *Cadernos de Saúde Pública*, **17**, 1083–1098.

Besag, J. and Newell, J. (1991). The detection of clusters in rare diseases. *Journal of the Royal Statistical Society, Series A*, **154**, 143–155.

Chadjiconstantinidis, S. and Pitselis, G. (2009) Further improved recursions for a class of compound Poisson distributions. *Insurance: Mathematics and Economics*, **44**, 278–286.

Chan, H. P. and Walther, G. (2013). Detection with the scan and the average likelihood ratio. *Statistica Sinica*, **23**, 409–428.

Cuzick, J. and Edwards, R. (1990). Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society, Series B (Methodological)*, **52**, 73–104.

De Pril, N. (1986a). Improved recursions for some compound Poisson distributions. *Insurance: Mathematics and Economics*, **5**, 129–132.

Gurland, J. (1957). Some interrelations among compound and generalized distributions. *Biometrika*, **44**, 265–268.

Jung, I., Kulldorff, M., and Klassen, A. C. (2007). A spatial scan statistic for ordinal data. *Statistics in Medicine*, **26**, 1594–1607.

Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters: Detection and inference. *Statistics in Medicine*, **14**, 799–810.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in statistics - theory and methods*, **26**, 1481–1496.

Kulldorff, M., Tango, T., and Park, P. J. (2003a). Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis*, **42**, 665–684.

Kulldorff, M., Fang, Z., and Walsh, S. J. (2003b). A tree-based scan statistic for database disease surveillance. *Biometrics*, **59**, 323–331.

Kulldorff, M. (2006). Tests of spatial randomness adjusted for an inhomogeneity. *Journal of the American Statistical Association*, **101**, 1289–1305.

Statistica Sinica: Preprint
doi:10.5705/ss.2013.215w

Lawson, A., Biggeri, A., Bohning, D., Lesaffre, E., Viel, J. F., and Bertollini, R. (1999). *Disease mapping and risk assessment for public health.* John Wiley & Sons, Chichester, UK.

Nhu Le, D., Petkau, A. J., and Rosychuk, R. (1996). Surveillance of clustering near point sources. *Statistics in Medicine*, **15**, 727–740.

Loh, J. M. and Zhu, Z. (2007). Accounting for spatial correlation in the scan statistic. *The Annals of Applied Statistics*, **1**, 560–584.

MATLAB and Statistics Toolbox Release 2012a. The MathWorks, Inc. Natick, Massachusetts, United States.

Marshall, R. J. (1991). A review of methods for the statistical analysis of spatial patterns of disease. *Journal of the Royal Statistical Society, Series A*, **154**, 421–441.

Morton-Jones, T., Diggle, P., and Elliott, P. (1999). Investigations of excess environmental risks around putative sources: Stone's test with covariate adjustment. *Statistics in Medicine*, **18**, 189–197.

Neill, D. B., Moore, A. W., Sabhnani M. R., and Daniel, K. (2005). Detection of emerging space-time clusters. *Proc. 11th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 218–227.

Panjer, H. H. (1981). Recursive evaluation of a family of compound distributions. *Astin Bulletin*, **12**, 22–26.

Quenouille, M. H. (1949). A relation between the logarithmic, Poisson, and negative binomial Series. *Biometrics*, **5**, 162–164.

Rosychuk, R. J and Chang, H-M. (2013). A spatial scan statistic for compound Poisson data. *Statistics in Medicine*, **32**, 5106–5118.

Rosychuk, R. J, Huston, C. and Prasad, N. G. N. (2006). Spatial event cluster detection using a compound Poisson distribution. *Biometrics*, **62**, 465–470.

Rosychuk, R. J. and Stuber, J. L. (2010). An exact test to detect geographic aggregations of events. *International Journal of Health Geographics*, **9**, 1–14.

Statistica Sinica: Preprint
doi:10.5705/ss.2013.215w

Rosychuk, R. J., Voaklander, D. C., Klassen, T.P., Senthilselvan, A., Marrie, T.J., and Rowe B.H. (2010a). Asthma Presentations by Children to Emergency Departments in Alberta, Canada: A Large Population-Based Study. *Pediatric Pulmonology*, **45**, 985–992.

Rosychuk, R. J., Voaklander, D. C., Klassen, T.P., Senthilselvan, A., Marrie, T.J., and Rowe B.H. (2010b). A Population-based Study of Emergency Department Presentations for Asthma in Regions of Alberta, Canada. *Canadian Journal of Emergency Medicine*, **12**, 339–346.

Stone, R. A. (1988). Investigations of excess environmental risks around putative sources: statistical problems and a proposed test. *Statistics in Medicine*, **7**, 649–660.

Tango, T. (1995). A class of tests for detecting 'general' and 'focused' clustering of rare diseases. *Statistics in Medicine*, **14**, 2323–2334.

Wang, T-C and Yue, C-S J. (2013). A binary-based approach for detecting irregularly shaped clusters. *International Journal of Health Geographics*; **12**, 25.

Zhang, T., Zhang, Z., and Lin, G. (2012). Spatial scan statistics with overdispersion. *Statistics in Medicine*, **2**, 762–774.

Irving K. Barber School of Arts and Sciences, 1177 Research Road, Kelowna, British Columbia V1V 1V7, Canada.

[1]E-mail: hsing-ming.chang@ubc.ca or hsingmin@ualberta.ca

Department of Pediatrics, Edmonton Clinic Health Academy, 11405 87 Avenue NW, Edmonton, Alberta T6G 1C9, Canada.

[2]E-mail: rhonda.rosychuk@ualberta.ca