Weighted Normal Spatial Scan Statistic for Heterogeneous Population Data

Lan Huang [*], Ram C. Tiwari [†], Zhaohui Zou [‡], Martin Kulldorff [§], Eric J. Feuer [¶]

October 24, 2008

[*]Lan Huang is Mathematical Statistician, contractor, Statistical Research and Applications Branch, Division of Cancer Control and Population Sciences, National Cancer Institute, Rockville, MD 20852 (email: huangla@mail.nih.gov).

[†]Ram Tiwari is Associate director in the office of Biostatistics, CDER, FDA, Silver Spring, MD, USA. Most of this author's work was done while the author was at the National Cancer Institute, NIH. The views expressed in this article do not necessarily represent those of the US Food and Drug Administration, and of the National Cancer Institute. (email: ram.tiwari@fda.hhs.gov).

[‡]Zhaohui Zou is Statistical Programmer, Information Management Services, Inc., Silver Spring, MD 20904 (email: zouj@imsweb.com).

[§]Martin Kulldorff is Associate Professor, Harvard Medical School and Harvard Pilgrim Health Care, Boston, MA 02215 (email: martin_kulldorff@hms.Harvard.edu).

[¶]Eric J. Feuer is Chief, Statistical Research and Applications Branch, Division of Cancer Control and Population Sciences, National Cancer Institute (email: feuerr@mail.nih.gov).

Abstract

In geographical spatial epidemiology and disease surveillance, all the existing spatial scan methods for cluster detection using continuous data are designed for evaluating clusters of individuals and analyzing individual-level data. Motivated by growing demands to study the spatial heterogeneity of continuous measures in population data, such as mortality rates, survival rates, average body mass indexes and pollution at state, county, and census tract levels, we propose a weighted normal scan statistic for investigating the clusters of the cells (geographic units such as counties) with unusual high/low continuous regional measures, where the weights reflect the uncertainty of the regional measures or sample size (number of observed cases) in the cells. Power, precision, the effect of the weights, and the sensitivity of the proposed test statistic to data from various distributions are investigated through intensive simulation. The method is applied to 1988-2002 stage I and II lung cancer survival data in Los Angeles County in order to search for clusters of geographic units with high/low survival rates in a short-term/long-term survival after diagnosis, and to 1999-2003 breast cancer age-adjusted mortality rate data in the US collected by the Surveillance, Epidemiology and End Results (SEER) program in order to evaluate the clustering pattern of counties with high mortality rate. The proposed method will be included in the new release of SaTScan software (www.satscan.org).

Keywords: weighted normal spatial scan statistic, cluster detection, geographic variation, continuous regional data, lung cancer survival, breast cancer mortality

1 Introduction

With the wide availability of geographical information systems, there is an increasing interest in the geographical aspects of disease, which has inspired the development of statistical methods for spatial epidemiology. Most of the time, the interest is in the geographical distribution of individuals with different disease status. For example, we may want to know the geographical distribution of prostate cancer survival. An area with shorter survival could, for example, reflect a higher proportion of a more aggressive form of the disease due to personal genetic or localized environmental factors, or it could reflect a local hospital with substandard care. There exist a large number of spatial cluster detection methods for analyzing such individual data, depending on the nature of the data, such as dichotomous variables for prevalence of cases and controls, Poisson variables for incidence or mortality (Duczmal and Assunção 2004; Kulldorff, Huang, Pickle and Duczmal 2006; Patil and Taillie 2004; Kulldorff 1997; Turnbull, Iwano, Burnett, Howe and Clark 1990), ordinal variables for cancer stage or histology (Jung, Kulldorff, and Klassen 2007), and continuous variables for length of survival (Huang, Kulldorff, and Gregorio 2006) or for birth weight or BMI (www.satscan.org). There are also other types of analysis, such as disease mapping (Lawson, Browne, and Widal Rodeiro 2003; Richardson, Thomson, Best, and Elliot 2004; Besag, York, and Mollie 1991; Knorr-Held and Rass 2000), cluster evaluation around a prespecified point source (Diggle's D (Diggle 1990), Lawson-Waller's Score test (Waller, Turnbull, Clark and Nasca 1992 and Lawson 1993), Bithell's linear rank score test (Bithell 1995), Isotonic regression (Stone 1988), and evaluations of global clustering throughout the map (Moran's I (1950), Tango's statistics (1995 and 2000), Besag Newell's R (1991), Oden's Ipop (1995)).

Sometimes the interest is not at the level of individuals, but in the geographical distribution of a continuous variable at some aggregate level. For example, each hospital has an average length of survival for their prostate cancer patients, and we may want to know if hospitals with shorter survival times are located close to other hospitals with short survival times. If we have

near perfect measures about the average survival from each hospital, with negligible error, then we can use the same statistical methods as for individuals with a continuous outcome variable. If the sample size is the same for each hospital, we could also use the established methods for continuous data. The problem occurs when we have different sample sizes for different hospitals, and we then have to take this variable uncertainly by hospital into account when doing the analysis for cluster detection. That is the focus of this paper.

There are several other types of application of this kind. For example, rather than studying the prevalence of smoking in individuals, we may be interested in the geographical distribution of low and high prevalence of smoking in schools. As another example, we may view each county as having a certain risk of breast cancer, and we want to know if there are any clusters of neighboring counties with high breast cancer risk, which is a very different question compared to whether there are any clusters of individuals with breast cancer, even though the counties included in the two types of clusters could be the same. As a third example, consider air pollution measurements from a number of distinct sites. If we have only one measurement from each site, those are individual measurements and can be analyzed using existing methods for continuous data, treating each site as an individual. If we have a number of measurements from each site, we need to take the variances into account when looking for geographical differences of the average pollution level in the sites rather than in the individual measurements.

While earlier methods for individual data have considered observations from both categorical and continuous variables, at the average level we are typically interested in continuous data. We develop a new spatial scan statistic, such as the weighted normal scan statistic, that accommodates continuous data with varying regional reliability/uncertainty within each geographic unit (such as county). Specifically, we develop the statistic for cluster detection on a regional measure for location z using a weight δ_z (the inverse of the uncertainty) associated with location z, to adjust for the uncertainty of the observed w_z value. The cluster detected is then the collection of geographic units (cells) with high/low regional measures that directly reflects the behavior of

the cells, instead of the individuals inside cells.

The difference between the methods developed in this paper and previous methods for continuous data is not whether data are available at one level or the other, but in the scientific question being asked. Is the interest in the geographical distribution of individual observations or in the geographical distribution of some unknown underlying variables at an aggregate level, which can only be estimated through a sample of individual observations? In fact, the exact same data can be used for both types of analyses, with very different results due to very different questions being asked from the data. This problem arises in population-based data, where, due to confidentiality concerns, researchers can not have access to individual-level information and have to rely on the estimates based on regional level data. In this case, the hypothesis that they can investigate is limited to the behavior of the regions. Note that in the analysis of individual risk, both individual case and control data, and aggregated Poisson count data can be used. However, aggregated survival time or BMI or pollution measures do not allow the analysis of individual behavior but only the regional behavior. The distinction is that counts can be aggregated but often still motivate studies of individual attributes, while the others represent average values.

The weight δ_z, for each z, in the construction of the scan statistic (more details are in Section 2) is assumed to be a known measure proportional to the inverse of the uncertainty in each z. For example, in pollution data, we may record different values of a pollution measure from several locations and different time points within one geographic unit (cell) and report the averaged (perhaps through geostatistical methods) value by cell as the final reporting measure w_z. We also record the variance of the reporting measure as the uncertainty of the cell. From a complex survey, we may obtain the regional BMI values and associated variances from some weighting procedure. In case of survival analysis, we may obtain the k-year (e.g., $k = 1, 3, 5, 10$) survival rates, the w_z's, using methods such as the LifeTable method or the Kaplan-Meier method, and the variances of the rates for each z, using Greenwood's method (Armitage and Berry 1994). In

mortality rate analysis, we can obtain the direct and indirect age-adjusted mortality rates and associated variances in each z from SEER*Stat (http://seer.cancer.gov/seerstat/). In the above cases, the related weights δ_z could be taken to be the inverse of the variances of the w_z's. In practice, we may not always have a large sample size in each z to obtain a reliable estimate of the variance in each cell. For example, if there is one individual measure recorded in a geographic unit z, the variance is not estimable. Also, if there are no deaths observed in a particular period in z, the variance of the survival rate for that period is zero. In such situations, we may use population size or sample size (i.e., the number of cases observed) as a proxy for the inverse of variance.

The rest of the paper is organized as follow. The weighted normal scan statistic is developed in section 2.1. The performance of the proposed scan statistic on varying data is evaluated using simulation studies, and discussed in section 3. The application of the method on the survival rates of stage I and II lung cancer in Los Angeles County and the breast cancer age-adjusted mortality rates in US are given in section 4. The paper ends with a discussion in section 5.

2 Weighted Normal Scan Statistic

We construct the weighted normal scan statistic based on the likelihood ratio test theory, with the likelihood incorporating the weights and location information. We maximize the likelihood over the search zones (Z's), where each Z is an arbitrary zone including a collection of cells (z) in the whole study region G.

2.1 Weighted Normal Likelihood and MLEs

First, we assume that the weight δ_z, associated with w_z, is an inverse function of the regional uncertainty measure of w_z in z. For example, if w_z represents the average of individual BMI values in z (county) from a complex survey, the weight δ_z can be taken as the inverse of the estimated variance provided by those different data sources. For a particular zone Z, we assume

that $w_z|\delta_z \sim N(\mu_Z, \frac{\sigma_G^2}{\delta_z})$, when $z \in Z$ and $w_z|\delta_z \sim N(\mu_{Z^c}, \frac{\sigma_G^2}{\delta_z})$, when $z \in Z^c (= G - Z)$. Here, μ_Z and μ_{Z^c} are the means of the measurements in and out Z respectively, $\sigma_G^2/\delta_z (= \sigma_{w_z}^2)$ is the variance of w_z ($z \in G$) after adjusting the local weight δ_z. Thus, given δ_z, the w_z's are assumed to be independent and normally distributed with the same mean but different variances in a particular Z. The likelihood function for an arbitrary zone Z in G is

$$L(\mu_Z, \mu_{Z^c}, \sigma_G^2) \propto \prod_{z \in Z} \frac{\sqrt{\delta_z}}{\sigma_G} \exp(-\frac{\delta_z}{2\sigma_G^2}(w_z - \mu_Z)^2) \prod_{z \notin Z} \frac{\sqrt{\delta_z}}{\sigma_G} \exp(-\frac{\delta_z}{2\sigma_G^2}(w_z - \mu_{Z^c})^2), \quad (1)$$

and its logarithm is

$$ln(L(\mu_Z, \mu_{Z^c}, \sigma_G^2)) \propto -\frac{1}{2} \sum_{z \in G} [ln(\frac{\sigma_G^2}{\delta_z})] - \sum_{z \in Z} [\frac{\delta_z}{2\sigma_G^2}(w_z - \mu_Z)^2] - \sum_{z \in Z^c} [\frac{\delta_z}{2\sigma_G^2}(w_z - \mu_{Z^c})^2].$$

The maximum likelihood estimates (MLEs) for μ_Z, μ_{Z^c} and σ_G^2 are $\hat{\mu}_Z = \frac{\sum_{z \in Z}(\delta_z w_z)}{\sum_{z \in Z} \delta_z}$, $\hat{\mu}_{Z^c} = \frac{\sum_{z \in Z^c}(\delta_z w_z)}{\sum_{z \in Z^c} \delta_z}$, and $\hat{\sigma}_G^2 = \frac{\sum_{z \in Z} \delta_z (w_z - \hat{\mu}_Z)^2 + \sum_{z \in Z^c} \delta_z (w_z - \hat{\mu}_{Z^c})^2}{n_G}$, respectively. Thus, the variance for w_z inside/outside Z, given the weight δ_z, is estimated as

$$\hat{\sigma}_{w_z}^2 = \frac{\hat{\sigma}_G^2}{\delta_z} = \frac{\sum_{z' \in Z} \delta_{z'} (w_{z'} - \hat{\mu}_Z)^2 + \sum_{z' \in Z^c} \delta_{z'} (w_{z'} - \hat{\mu}_{Z^c})^2}{n_G \delta_z},$$

where n_G is the total number of z's in G, and z' denotes a distinct geographical cell index from z. So for a given Z, the loglikelihood function is maximized at

$$\begin{aligned} Ln(L(\hat{\mu}_Z, \hat{\mu}_{Z^c}, \hat{\sigma}_G^2)) &\propto -\sum_{z \in G} ln(\frac{\sum_{z' \in Z} \delta_{z'} (w_{z'} - \hat{\mu}_Z)^2 + \sum_{z' \in Z^c} \delta_{z'} (w_{z'} - \hat{\mu}_{Z^c})^2}{n_G \delta_z}) - n_G \\ &\propto -\sum_{z \in G} ln(\sum_{z' \in Z} \delta_{z'} (w_{z'} - \frac{\sum_{z' \in Z} w_{z'} \delta_{z'}}{\sum_{z' \in Z} \delta_{z'}})^2 + \sum_{z' \in Z^c} \delta_{z'} (w_{z'} - \frac{\sum_{z' \in Z^c} w_{z'} \delta_{z'}}{\sum_{z' \in Z^c} \delta_{z'}})^2) + \sum_{z \in G} ln(n_G \delta_z) - n_G \end{aligned} \quad (2)$$

Maximizing the loglikelihood $ln(L(\hat{\mu}_Z, \hat{\mu}_{Z^c}, \hat{\sigma}_G^2))$ is equivalent to maximizing the expression in equation (2). Because the summation $\sum_{z \in G}$ is independent of Z, maximizing (2) is equivalent to maximizing

$$-ln(\sum_{z \in Z} \delta_z (w_z - \sum_{z \in Z} \frac{\delta_z}{\sum_{z \in Z} \delta_z} w_z)^2 + \sum_{z \in Z^c} \delta_z (w_z - \sum_{z \in Z^c} \frac{\delta_z}{\sum_{z \in Z^c} \delta_z} w_z)^2). \quad (3)$$

In equation (3), the log function is a monotone function. Therefore, maximizing (3) is equivalent to maximizing

$$-(\sum_{z\in Z}\delta_z w_z^2 - \frac{(\sum_{z\in Z}\delta_z w_z)^2}{\sum_{z\in Z}\delta_z}) - (\sum_{z\in Z^c}\delta_z w_z^2 - \frac{(\sum_{z\in Z^c}\delta_z w_z)^2}{\sum_{z\in Z^c}\delta_z}) \tag{4}$$

$$= -\sum_{z\in G}\delta_z w_z^2 + \frac{(\sum_{z\in Z}\delta_z w_z)^2}{\sum_{z\in Z}\delta_z} + \frac{(\sum_{z\in Z^c}\delta_z w_z)^2}{\sum_{z\in Z^c}\delta_z} \tag{5}$$

The first term in equation (5) is independent of Z. Therefore, maximizing $\frac{(\sum_{z\in Z}\delta_z w_z)^2}{\sum_{z\in Z}\delta_z} + \frac{(\sum_{z\in Z^c}\delta_z w_z)^2}{\sum_{z\in Z^c}\delta_z}$ over the Z's is essentially maximizing the loglikelihood $ln(L(\hat{\mu}_Z, \hat{\mu}_{Z^c}, \hat{\sigma}_G^2))$ and the likelihood $L(\hat{\mu}_Z, \hat{\mu}_{Z^c}, \hat{\sigma}_G^2)$. This simple expression in equation (5) without the first term will be used to construct the likelihood based spatial scan statistic in section 2.2. Note that $\hat{\mu}_Z$ is unbiased since

$$E(\hat{\mu}_Z) = \frac{1}{\sum_{z\in Z}\delta_z}\sum_{z\in Z}\Big[\delta_z E(w_z)\Big] = \frac{1}{\sum_{z\in Z}\delta_z}\mu_Z\sum_{z\in Z}(\delta_z) = \mu_Z, \tag{6}$$

and, similarly, $\hat{\mu}_{Z^c}$ is also unbiased. However, since

$$E(\hat{\sigma}_G^2) = E\frac{1}{n_G}[\sum_{z\in Z}\delta_z(w_z-\hat{\mu}_Z)^2 + \sum_{z\notin Z}\delta_z(w_z-\hat{\mu}_{Z^c})^2] = \frac{n_G-1}{n_G}\sigma_G^2, \tag{7}$$

$\hat{\sigma}_G^2$ is asymptotically unbiased when n_G is large. When $\mu_Z = \mu_{Z^c} = \mu_G$, the MLEs become $\hat{\mu}_G = \frac{\sum_{z\in G}(\delta_z w_z)}{\sum_{z\in G}\delta_z}$, and $\hat{\sigma}_G^2 = \frac{\sum_{z\in G}\delta_z(w_z-\hat{\mu}_G)^2}{n_G}$. One can increase the n_G by subdividing the cells into smaller cells but with different δ_z (for example, subdivide county into census tracts), or simply enlarging the whole study region G (for example, if the G is the state of Washington with county as cell, we can add more counties from Oregon into the study region). The asymptotic property in equation (7) is still valid for both cases with large n_G, but the σ_G^2 on the right side of the equation will be different from that in the original data when we change the data in either way.

In this section, we obtained closed forms of the MLEs for the parameter estimates. It was also shown that the maximum likelihood over different search zones Z's can be expressed in a simple closed form. These results help us to derive the spatial scan statistic, based on the likelihood ratio, which is analytically tractable and computationally feasible.

2.2 Spatial Scan Statistic and Hypothesis Test

Let Z be an arbitrary zone in G. The zones could have shapes of circles, ellipses, and irregular shapes with varying geographic size. We use zones with circular shapes with varying radius in our analysis and any z whose centroid is in a circular area belongs to a particular Z. We maximize the likelihood or log likelihood over the zone Z's in the whole study region G under both the null and alternative hypotheses, which is the same irrespective of the shapes used for the zones Z. The null hypothesis is that the mean of $w_z, z \in Z$ is homogeneous in G so that $\mu_Z = \mu_{Z^c} = \mu_G$ for any $Z \subset G$.

Define $\theta_Z = (\mu_Z, \sigma_G^2)$ and $\theta_{Z^c} = (\mu_{Z^c}, \sigma_G^2)$, where $Z \subset G$. The likelihood for a given zone Z is written as $L(\theta_Z, \theta_{Z^c}) = \prod_{z \in Z} L(\theta_Z) \prod_{z \in Z^c} L(\theta_{Z^c})$. Given Z, the estimated maximum likelihood is $L(\hat{\theta}_Z, \hat{\theta}_{Z^c})$, where the MLEs $\hat{\theta}_Z$ and $\hat{\theta}_{Z^c}$ are computed using the method described in section 2. The likelihood ratio test statistic for null $H_0 : \mu_Z = \mu_{Z^c} = \mu_G$ vs. alternative $H_a : \mu_Z \neq \mu_{Z^c}$, for any zone Z, is written as

$$\lambda = \frac{\max_{Z, \mu_Z \neq \mu_{Z^c}, \sigma_G^2} L(\theta_Z, \theta_{Z^c})}{\max_{Z, \mu_Z = \mu_{Z^c}, \sigma_G^2} L(\theta_Z, \theta_{Z^c})} = \frac{\max_{Z, \mu_Z \neq \mu_{Z^c}, \sigma_G^2} L(\theta_Z, \theta_{Z^c})}{L_0} = \frac{L(\hat{Z})}{L_0}, \tag{8}$$

where L_0 depends on $\hat{\mu}_G$ and $\hat{\sigma}_G^2$ and is independent of the search zone Z. Therefore, the λ is maximized when the likelihood $L(\theta_Z, \theta_{Z^c})$ in the numerator of equation (8) is maximized. And maximizing $L(\theta_Z, \theta_{Z^c})$ under the alternative hypothesis is equivalent to maximizing the equation (5) over all search zones Z's.

For the alternative $H_a : \mu_Z > \mu_{Z^c}$, the test statistic becomes

$$\lambda = \frac{\max_{Z, \mu_Z > \mu_{Z^c}, \sigma_G^2} L(\theta_Z, \theta_{Z^c}) I(\hat{\mu}_Z > \hat{\mu}_{Z^c})}{\max_{Z, \mu_Z = \mu_{Z^c}, \sigma_G^2} L(\theta_Z, \theta_{Z^c})}, \tag{9}$$

which identifies the clusters with unusual means higher than those outside the clusters. Similarly, to test $H_a : \mu_Z < \mu_{Z^c}$, we use the indicator function $I(\hat{\mu}_Z < \hat{\mu}_{Z^c})$ in the numerator of λ to identify the clusters of measures with unusually small mean values.

Constraints can be set on the maximum and minimum size of Z. We normally restrict the maximum size of Z to be less than 50% of the total number of z's or 50% of total population in

G and the minimum size of Z to be at least two cells. The constraints on both the maximum and minimum sizes depend on the purpose of the study or the features of the data.

As there is no closed form for the distribution of λ, a Monte Carlo hypothesis testing procedure is employed, wherein M datasets are generated by random permutation (the pairs $(w_z, \delta_z), z \in G$, are randomly moved among the existing geographical locations). Note that moving $(w_z, \delta_z), z \in G$ together keeps the maximum likelihood estimate of the overall mean $\hat{\mu}_G = \frac{\sum_{z \in G} \delta_z w_z}{\sum_{z \in G} \delta_z}$ invariant with respect to the permutations. The M datasets are called the null datasets without spatial clustering of cells with high/low means because we randomly move the pairs (w_z, δ_z) around. The $M + 1$ values of the statistic λ's are computed for the observed and permuted datasets. At the α-level, H_0 is rejected if the *rank* of the λ obtained from the observed data is among the $\alpha(M + 1)$ largest λ's, and the p-value is $1 - \frac{rank}{M+1}$ (Dwass 1957). Note that we used random permutation to generate the null datasets instead of simulation, because we do not know what are the true common mean and variance for the null datasets. If we use a set of pre-estimated means and variances ($\hat{\mu}_G$ and $\hat{\sigma}^2_G$) from the whole observed data to simulate null datasets from a normal distribution, our alternative hypothesis would be $\mu_Z \neq \hat{\mu}_G$ instead of $\mu_Z \neq \mu_{Z^c}$. Furthermore, since the permutation procedure does not require a distribution assumption in order to generate the data under the null hypothesis, the procedure is more robust to data with varying continuous distributions. In this way, the permutation procedure maintains the correct α level error if the observations do not come from a normal distribution. Therefore, for evaluating the spatial homogeneity of the means in G, it is more appropriate to use random permutation procedure.

Note that the random permutation testing procedure does not have power to detect a cluster consisting of a single geographic unit z because the permutation procedure randomly moves the locations of the measures, but does not change the values of the measures. After the random relocation, the spatial clustering patterns among cell measures are removed, but unusually high/low values in cells remain the same in the permuted datasets. This is reasonable because we

are looking for clusters of cells, instead of clusters of individual cases within a single geographic unit or across cells.

The zone Z that maximizes the likelihood under the alternative hypothesis in (8) is the most likely cluster, that is, the cluster that is least likely to be due to chance. We denote the value of λ associated with this particular Z by $\lambda^{(1)}$. If the p-value associated with $\lambda^{(1)}$ is less than α, Z is significantly different from the areas outside $Z(Z^c)$.

We can also find the zones that do not maximize the likelihood under the alternative hypothesis in (8), but provide the 2nd maximum likelihood, the 3rd maximum, etc., by allowing the denominator to remain the same in (8), and the numerator being the 2nd maximum, the 3rd maximum, and so on. Note that the zones must be mutually exclusive. We define the corresponding values of statistic λ as $\lambda^{(2)}, \lambda^{(3)}$, and so on. These are called secondary statistics. Comparing the $\lambda^{(s)}, s > 1$, from the observed data with the M $\lambda^{(1)}$'s from the permuted datasets, if the rank of the $\lambda^{(s)}, s > 1$, is still among the $\alpha(M + 1)$ largest λ's, we reject the H_0 even without the presence of the most likely cluster, and we claim that those associated zones are statistically significant secondary clusters.

3 Power and Precision of Cluster Detection

A simulation study is conducted to understand the performance of the proposed weighted normal scan methods, described in section 2, on continuous data in terms of power and precision of cluster detection. We only evaluate the detection of a single cluster with higher mean value compared with those outside. The alternative hypothesis is $H_a : \mu_Z > \mu_{Z^c}$ for at least one Z, and therefore, the computation of the statistic includes $I(\hat{\mu}_Z > \hat{\mu}_{Z^c})$. We fixed the number of permutations for each simulated dataset to be $M = 999$.

We simulate continuous grid data in a 10×10 matrix of cells, which represents 100 geographic units (z) in the whole region. We do not simulate data in a real map with cells as counties or census tracts because of computational time restrictions. The true cluster Z^* is defined to be the

circular area with center at (row=3, column=6) and radius of length 2. Any cell, whose center is in the circular area, belongs to the true cluster. Note that the selection of the cluster location is random. However, if the cluster center is on the border of the study region, the power may be a little bit lower because of the edge effect. There are a total of 13 cells included in the true cluster area out of the 100 cells. To evaluate the change in power of the proposed statistic λ as a function of mean difference inside/outside the selected cluster region, without loss of generality, we simulate w_z from N(0,1) outside the selected cluster, and from $N(0 + c\sqrt{2}, 1)$ inside the cluster, with $c = 0.5, 1, 1.5, 2, 3$ (Table 1, Case 1). Note that c can be interpreted as the number of standard error units of the difference between the means of the w_z's in and outside Z. Also, to understand the effect of weights, we simulate data from N(0,1) outside the true cluster, and from $N(0+1.5\sqrt{2}, 1)$ inside the true cluster, and let the weight $\delta_z(= \eta) = 1, 2, 4, 8, 10, 100, 1000$. Table 1 gives the simulation results for cases with different mean differences inside/outside the true cluster and varying weights.

Robustness of the developed methods to varying distributions is also of interest. Therefore, we simulate data not only from the normal distributions, but also from the double exponential (DoubleE), logistic, uniform, lognormal and Poisson distributions as shown in Table 2. The means and variances for DoubleE, logistic and uniform distributions are taken to be the same as those for the normal distributions. For the lognormal distribution, since it can not have a zero mean, we let the mean outside the cluster be 2 instead of 0. The mean difference, $c\sqrt{2}$, and the variances in lognormal distributions are set to be the same as those in normal distributions. For Poisson distributions, since the mean equals the variance (> 0), we simulate data from $Poisson(1)$ outside of the true cluster, and from $Poisson(1 + c\sqrt{2})$ inside the true cluster. The variability of the Poisson data is then bigger than the data from other distributions.

The criteria for evaluating the performance of the weighted normal scan method in the simulation are power of rejecting the null hypothesis of $\mu_Z = \mu_{Z^c}$, for any zone $Z \in G$ when the data are generated from varying alternatives ($\mu_Z > \mu_{Z^c}$); the proportion of detected cluster

in the true cluster (r_T); and the proportion of the true cluster in the detected cluster (r_D). We calculate the three measures through simulation as follows:

$$power = \frac{\text{number of simulations with p-value } <0.05}{L},$$

$$r_T = \frac{1}{L}\sum_{l=1}^{L} \frac{\text{number of common cells in true and detected clusters in } l^{th} \text{ simulation}}{\text{number of cells in the true cluster}},$$

and

$$r_D = \frac{1}{L}\sum_{l=1}^{L} \frac{\text{number of common cells in true and detected clusters in } l^{th} \text{ simulation}}{\text{number of cells in the detected cluster in } l^{th} \text{ simulation}},$$

where $L(= 1000)$ is the total number of alternative datasets simulated in each scenario.

The last two measures, r_T and r_D, are for evaluating the precision of the cluster location detection. The r_T evaluates sensitivity and r_D evaluates positive predictive value (PPV). The values of r_T and r_D are between 0 and 1, and larger values of r_T and r_D together imply higher precision of detecting the right location of the true cluster. However, a large value of r_T and a very small r_D together does not mean good precision, and vice versa.

Effect of mean difference

As shown in Table 1 (Case 1), as the mean difference ($c\sqrt{2}$) increased, the power and precision measures (r_T and r_D) increased. The power reaches close to 100% when the mean difference is $1.5\sqrt{2}$ or bigger. When the power reaches 100%, both r_T and r_D are close to 1.

Effect of weights

The effect of the varying weights is evaluated by considering the weight values and the location of the weights. Large variation in weights introduces large regional variability for w_z because $\sigma^2_{w_z} = \sigma^2_G/\delta_z$. We evaluate the effect of weights by increasing the values of weights from 1 to 1000, and also by allowing the high weights to be assigned to one cell (6,3), or several cells including 3 or 5 cells inside true cluster, or in the z's in a bigger area (inside/outside the true cluster including 13 cells).

Overall, when the variation of weights ($\delta_z = \eta$) increased, the power and precision measures declined (Case 2a, 2b, 3, and 4). We also note that when the cells in the true cluster area

with very high values of w_z have low variability or large sample size, we still have the power to detect them when weights are included in the formulation of the statistic. But when the cluster of unusual values has large variability or small sample size, the power of detecting this cluster becomes very low when weights are considered. This is actually the purpose of adding weights into the scan statistic λ; that is, we want to avoid detecting a cluster that is unreliable with very small sample size (observed cases). The precision is generally good when the power is above 70% (both sensitivity and PPV are high).

Effect of variance of w_z in and out the true cluster Z^*

For Case 5 in Table 1, we notice the decreasing patterns in power and precision measures when η increases. Therefore, a high variance in $w_z, z \in G$ is associated with lower power of detecting the cluster and lower precision. The r_D (PPV) becomes bigger than r_T (sensitivity) as the variance in the cluster increased, which implies that the detected cluster size is decreasing. However, most of the detected cluster is still in the true cluster region. Even for $\eta = 1000$, the r_D is not very low (0.67).

Sensitivity to varying distributions

From Table 2 and Figure 1 (A, B, and C), the weighted normal scan statistic given in equation (9) is very robust to varying distributions of the continuous measure w_z. For the homogeneous weight case ($\delta = 1$ in all cells), the graphs of power vs. c shows that only the data from Lognormal and Poisson distributions are different with the others. It is not surprising that Poisson data have the lowest power, because the Poisson data are counts with larger variance. The precision measures, r_T and r_D, show similar patterns as that of the power, and with bigger difference in r_T (sensitivity) and less difference in r_D (PPV) across the distributions.

When we fixed the mean difference to be $1.5\sqrt{2}$ and allowed the weight in the center of the cluster to vary from 1 to 1000, we noticed that the power for the data from all the distributions decreased slightly with the Poisson distribution having the greatest difference (Figure 1D, 1E and 1F).

4 Applications

4.1 Spatial Clustering of Short-Term and Long-Term Lung Cancer Survival Rates in Los Angeles County

The exponential-based scan statistic (Huang et al. 2006) and semiparametric scan statistic (Cook, Gold and Li 2007) searches for the clusters of individuals with high or low mean survival time during the study period and compare the likelihood function that is based on the survival data for all follow-up years (i.e., the entire survival curve). The spatial pattern in survival rates may also vary for different survival times (e.g., 1 year, 3 years, 5 years) because of the progression of the disease and the impact of screening and treatment during different parts of the survival (short term and long term); for example, new treatments may impact early survival, but the survival advantage may not persist over time. In this case, the interest is not to compare the whole survival curve and the overall average survival time, but to compare the geographic patterns of the survival rate at a point on the survival curve. The new weighted normal scan method is an ideal tool for this kind of study in searching for clusters of cells (not individuals) with unusual survival rate in short-term follow-up and long-term follow-up, separately. With more flexibility, we can estimate the survival rates using any of the existing methods such as nonparametric Kaplan-Meier or LifeTable methods (Klein and Moseschberger 1997), a semiparametric model (Cox 1972), a parametric model including exponential model (Klein et al. 1997), or a cure model (Yu, Tiwari, Cronin and Feuer 2004).

We first use the Kaplan-Meier (KM) method to estimate the k-year survival rates and associated variances (k=1, 3, and 5) by medical service study areas (MSSAs) (California Office of Statewide Health Planning and Development, 2005) and by census tracts on the survival data with individual survival information in Los Angeles (LA) County for stage I and II lung cancer cases diagnosed in 1988-2002 with follow-up to 2002. We use the same survival data described by Huang, Pickle, Stinchcomb and Feuer (2007), where they analyzed the spatial variation of mean survival time for individuals by an exponential model based scan model. The data include

a total of 9,242 stage I and II lung cancer cases diagnosed in LA County, with 60% of the cases censored by the end of follow-up. There are 2054 tracts and 100 MSSAs in LA County. The MSSAs are aggregations of tracts based on the supplies of medical service. We use the two kinds of cells to illustrate the performance of the methods on a limited number of large cells (MSSA), versus a large number of smaller cells (tracts). The advantage of working with smaller size areas is that we can more precisely specify the location of a cluster. Also note that one of the major differences between the MSSA and tract analyses is that the cluster detected in the MSSA analysis is the collection of the MSSA areas with similar rates and reflects the clustering behavior of the MSSA areas; however, in the tract analysis, the clustering pattern of tracts is evaluated. Even though they may provide similar cluster locations with high/low survival rates and both have geographic units (cells) as the study subjects, the meaning of the cluster is not exactly the same.

If there are no deaths observed beyond 5 years due to small sample size in some MSSAs and tracts, we assume that the survival probability remains the same after the time of the last death observed in a particular MSSA or tract (last point carry-over). The sample size, here, is the number of cases diagnosed with stage I and II lung cancer during the study period in a specific cell (MSSA or tract). Because of this last point carry-over approach, the survival rate may be overestimated in a cell with small sample size and large proportion of censoring. Obviously, with a total of 9,242 cases, sample size is not an issue for most of the 100 MSSAs, but we do have very small sample sizes for tracts and the existence of overestimation. As discussed in the earlier sections, we can use the weight to reduce the effect of the areas with unreliable overestimated survival rates. The tracts without any cases observed are treated as empty areas. Thus, there are 1885 tracts used in our analysis after the exclusion of 6 empty tracts. This will still result in good estimates on the survival rates in the areas covered by the search window with aggregation of tracts. The advantage of doing the analysis at a very small level like the tract-level is that the statistical technique employed will aggregate the tracts rather than having pre-formed cells

such as the MSSAs.

As described above, we estimate 1-year, 3-year and 5-year survival rates by MSSA and by tract in LA County. We then apply the weighted normal models on w_z, the estimated k-year survival rates, to study the spatial variation of the rates. At the tract-level, the variances of the k-year survival rates in z's are not available for 981 tracts when k is 1, for 839 tracts when k is 3, and for 1052 tracts when k is 5. We lose a lot of locations if we use weights as inverse of the variances in tracts for studying the spatial clustering pattern of the k-year survival rates. However, the sample size is available for all the 1885 tracts, and thus we use the sample size at tract-level as a substitute for the inverse of variance as the weight. We use both the sample size and inverse of the variance of the rates in each z as the weight in the MSSA analysis. The minimum cluster size is selected to be two MSSAs and 10 tracts, respectively, and the maximum is 25 percent of the total locations (either MSSA or tracts). Note that as mentioned earlier, both the minimum and maximum size can be changed according to the research interest.

The detected clusters are shown in Figure 2. The cell is MSSA in Figures 2A-2B, 2D-2E, and 2G-2H, and tract in Figures 2C, 2F, and 2I. The weight is inverse of variance in the 1st column, and sample size in the 2nd and the 3rd columns. We present the clustering pattern for 1-yr survival in Figures 2A-2C, for 3-yr survival in Figures 2D-2F, and for 5-yr survival in Figures 2G-2I. The summary information of the clusters is shown in Table 3a and 3b. The clusters detected in the 1st column of maps in Figure 2 are similar to the ones in the 2nd column, which shows that using sample size (observed cases) as the weight provides similar results to using the inverse of the variance as the weight in this situation. We calculate the k-year survival rate inside/outside the clusters using the KM method directly on the individual survival time data. The 1-year survival rate is 0.85 for the high survival rate cluster, 0.76 for the low survival rate cluster, and 0.81 for the cells outside these clusters in Figure 2C. The 3-year survival rate is about 0.68 for the high survival rate cluster, 0.51-0.54 for the low survival rate cluster, and 0.6 for cells outside the detected clusters in Figures 2D-2F. The 5-year survival rate is about

0.58-0.60 for high survival rate cluster, 0.43 for low survival rate cluster and 0.51 outside the detected clusters in Figures 2G-2I.

At tract-level, we detect significant high and low rate clusters of 1-year survival; however, there is no cluster of high/low 1-year survival rate detected at MSSA level. This implies that large cells (like MSSAs) with bigger sample size may provide more reliable estimates of the survival rate by the cell, but we lose some power in detecting significant spatial variation because a part of the variation is smoothed out inside the cell. Usually, when using spatial scan method searches through the cells, having more cells in the whole study area provides more flexibility in the location of the clusters and yields in better power.

One concern is that the estimates of the survival rates at tract level might be too variable because many tracts have only one or two cases diagnosed and no deaths observed in the study period. Using sample size as the weight, the new spatial scan method gave more weight to the w_z's in those areas with more reliable estimates (usually areas with large sample size) and compared the weighted average of the estimated survival rates inside/outside Z ($\hat{\mu}_Z = \frac{\sum_{z\in Z}\delta_z w_z}{\sum_{z\in Z}\delta_z}$). If one is worried about the high percentage of censored cases, one can use the observed number of deaths instead of the number of diagnosed cases as weights. The weighting process adjusts for differences in reliability and allows better comparisons. As shown in Table 3a, the estimates using the weighted average are always bigger than those from direct KM methods (overestimation) for tract analysis. The differences between the weighted means and the KM estimates for tract analysis are close to those for MSSA analysis in the 3-year survival rate analysis, and slightly bigger for the 5-year survival rate analysis. This suggests that we can work with the cells with small sample size in cluster detection studies. This message is also supported by observing the similar spatial patterns with high or low survival rates in the maps with both MSSA as cell and tract as cell.

The spatial pattern of the survival rates in this study is consistent with the pattern discovered in Huang et al., 2007. In Table 3b, we summarize the information on race and sex from the

SEER data we used for the survival analysis. The information on SES (social economic status) are from the 1990 census because we are more interested in the socioeconomic condition that the patients experienced both before and after diagnosis (1988-2002). There is no good summary statistic for smoking at tract level in US population data, so smoking status is not included in this analysis even though smoking is a well known risk factor for survival of patients with lung cancer. As shown in Table 3b, high survival rate clusters are located around the areas with high social economic status (high median income, more people with advanced education, higher percent of home owners, more people with health insurance, less people living below poverty level). They are in the western and northwestern portion of LA County. Low rate clusters are in the south LA County, around the low SES areas with more blacks and more males.

Our analysis also reveals that there is less spatial variation for 1-year survival rates. The difference between the survival rates between high rate clusters and low rate clusters is smaller for 1-year survival compared with 3-year and 5-year survival cases in the tract level analysis. Spatial clusters of high or low rates are only found for 3-year and 5-year survival rates in the MSSA analysis. The p-values for the clusters in 3-year and 5-year survival rates are much lower (< 0.001) than the ones (0.016 and 0.008) for clusters of 1-year survival rate in the tract analysis. We observe the location change of clusters of high rates (3-year to 5-year) from west to northwest in Figure 2 (2E to 2H and 2F to 2I), which implies that some regions in south western LA County may have better short and middle-term survival, but lose advantage in long-term survival (5-year survival). The locations of the low survival rate cluster remain almost the same in all the survival maps. The numbers in Table 3b for the cluster of high survival rates at tract-level reveal that the income, percent of home owner, percent of people with health insurance increase, and the percent of poverty, and blacks decrease when the high rate cluster changes from 2C to 2F, and to 2I. However, percent of advanced education and gender do not have consistent patterns when the high rate cluster changes the locations. The areas in the 3-year high rate cluster (in 2F), but outside the 5-year high rate cluster (in 2I), indicated in parenthesis, have

much lower median income compared with the values in the 5-year high rate cluster (40K vs. 51K), much lower percent of home owners (43% vs. 58%), much lower percent of people with health insurance (25% vs. 30%), more poor people (11% vs. 8%), and much more blacks (25% vs. 2%) as shown in Table 3b. While not a formal statistical test, these comparisons suggest that income, insurance, and race may play a bigger role on the long term advantage in the stage I and II lung cancer survival in the related areas.

4.2 Geographic Variation of Breast Cancer Mortality in US

Breast cancer is one of the most common cancers in the United States, and currently the second leading cause of cancer death in females (American Cancer Society 2007). The mortality rate for breast cancer has been declining since the middle of the 1990s (report to nation-2005 update). Beside the temporal trend and variation, the spatial heterogeneity of the breast cancer mortality is also of interest to epidemiologist and policy makers (Canto, Anderson and Brawley 2001 1998; Goodwin, Freeman, Mahnken, Freeman and Nattinger 2002; Jacquez and Greiling 2003; Sheehan et al., 2004). Here, we will apply the weighted normal scan method on the breast cancer mortality rate data to evaluate the geographic variation of rates in the US in terms of clustering of counties.

4.2.1 Applying the Weighted Normal Model on Rates

Note that the mortality rates are usually adjusted for age to avoid the confounding of age effect. There are two kinds of age-adjusted rates, namely direct and indirect (Curtin and Klein 1995). We briefly describe the calculation of the two measures and related variance in the following.

The direct age-adjusted cancer rates (DAR) can be written as

$$DAR_z = \sum_{j=1}^{J} \gamma_j \frac{d_{zj}}{n_{zj}},$$

where d_{zj} and n_{zj} are the number of cancer deaths and the mid-year population for the age-group j and the location z (e.g., tract, county or state), respectively, and the $\gamma_j (= \frac{p_{sj}}{\sum_j p_{sj}})$ are the

normalized proportion of mid-year population for the age-group j in the standard population (p_{sj}), so that $\sum_j \gamma_j = 1$. Let $d_j = \sum_{z \in G} d_{zj}$ and $n_j = \sum_{z \in G} n_{zj}$. Also, let $\gamma_{zj} = \frac{\gamma_j}{n_{zj}}$. Assuming that d_{zj} are Poisson counts, the variance for the direct rate in z is, $\text{var}(DAR)_z = \sum_{j=1}^{J} \gamma_{zj}^2 d_{zj}$ (Kim, Fay, Feuer and Midthune 2000).

The indirect age-adjusted rate (IAR) is defined as

$$IAR_z = (\frac{d_z}{\sum_j c_{sj} p_{zj}}) \times (\frac{d_s}{p_s}) = (\frac{c_s}{\sum_j c_{sj} p_{zj}}) d_z, \tag{10}$$

with $c_s = \frac{d_s}{p_s}$, where d_z is the number of deaths in z^{th} county; d_s, p_s and $c_{sj} = \frac{d_{sj}}{p_{sj}}$ are the total number of deaths, the total population, and the mortality rate for j^{th} age group in a standard population respectively; and p_{zj} is the population in j^{th} age group and z^{th} county. Then,

$$\text{var}(IAR)_z = (\frac{c_s}{\sum_j c_{sj} p_{zj}})^2 \text{var}(d_z),$$

where d_z follows Poisson distribution, c_s and c_{sj} are assumed to be known from a standard population. Thus, the estimate for the variance of IAR in z is $(\frac{c_s}{\sum_j c_{sj} p_{zj}})^2 d_z$.

As discussed in Pickle and White (1995), under some conditions, DAR and IAR are equivalent or similar; but the conditions are not always satisfied in practice. Indirect rates are usually used when age-specific numbers of deaths are not available in each cell or when the number of deaths is small (data is sparse). However, indirect rates may not be comparable across geographic areas when the age and area effects have interaction. Therefore, it is good to show the patterns using both measures if they are available.

In our analysis, we used the 5-year mortality data to reduce the instability of local age-specific rates in DAR. The mortality DAR and IAR (1999-2003) are computed using the mortality case data and population data from the SEER program. The age-adjusted mortality rate represents the number of deaths per 100,000 people controlling for age. The study area G is the United States excluding Alaska and Hawaii. The standard population is the 2000 US female population excluding those in Alaska and Hawaii. The cell units are counties in this analysis and there are a

total of 3108 counties in G. Ages are grouped into 19 categories $(0, 1-4, 5-9, 10-14, \cdots, 85+)$. We treat DAR_z (or IAR_z) as w_z, with $\delta_z = \frac{1}{\text{var}(DAR)_z}$ (or $\delta_z = \frac{1}{\text{var}(IAR)_z}$), $z \in G$, we then use the weighted normal scan statistic based on equation (8) on the data, searching for clusters of areas with high age-adjusted mortality rates. By introducing the weight into the scan statistic, the method becomes less sensitive to the unreliable estimates from low population and locations with high uncertainty when the data is sparse or has large variation. Since there are more than 3000 counties in the whole country and we are more interested in big clustering patterns instead of outliers (tiny clusters), we use 10 counties as the minimum search window size and 50% of total population as the maximum search window sizes for the mortality rate analysis.

The clustering pattern of counties with high DAR and IAR are the same in this analysis, so we use Figure 3A to present the pattern. The detected cluster is located in the northeastern USA with p-value as 0.001. The average value of DAR is 27.05 inside the cluster and 24.32 outside. For IAR, the average rate is 29.89 in the cluster and 26.71 outside. The average weight over the counties inside/outside the cluster is 0.13/0.07 for IAR and 0.16/0.08 for DAR. Note that the DAR and IAR maps may not always be the same and all the average values mentioned in this part are weighted averages with weights as the inverse of the corresponding variances within each cell.

4.2.2 Comparison of Weighted Normal Model and Poisson Model on Evaluating Breast Cancer Mortality

The Poisson model based spatial scan statistic (Kulldorff 1997) is a spatial statistic that has been widely used for detecting clusters of high mortality counts or rates. It uses the indirect standardization technique to obtain the expected cases under the null hypothesis of homogeneous relative risk in each strata (www.satscan.org). The key difference between the Poisson model and the weighted normal model is that the Poisson model approach works on count data with population information and evaluates clusters of individual cases, however, the weighted normal model approach works on regional continuous data with varying regional uncertainty and

evaluates clusters of regions (such as counties). In the Poisson model approach, the datasets under the null hypothesis of homogeneous individual disease risk are simulated by randomly relocating the location of the individual cases among population at risk, based on a multinomial distribution with the restriction of homogeneous relative risk, fixed total cases and the population in each cell (e.g. county in this study). However, in the weighted normal approach, the datasets under the null hypothesis of homogeneous age-adjusted rates for counties are generated by randomly permuting the observed values of county rates and their associated weights among locations using a distribution-free permutation procedure, in order to see if high values tend to occur near other high values. So, for studying the spatial clustering of individual cases, we move those individual cases and assign them to different locations in order to reduce the spatial patterns of individual cases for the null dataset generation. For studying the spatial clustering of cells that may include many cases, we only move those cells in order to reduce the geographic clustering pattern among cells without touching the cases inside each cell.

We apply the Poisson model based SaTScan method on the same data for the breast cancer mortality rate study, but the response becomes mortality cases. The maximum window size is 50% of the total population. The result is shown in Figure 3B. The cluster No. 1 in Figure 3B (average IAR=29.4, p-value=0.001) has large proportion of common areas with the cluster in Figure 3A (average IAR=29.9). One more cluster (cluster No. 2 in Figure 3B) is detected in the South Mississippi River area in Figure 3B with p-value as 0.001. The relative risk of dying from breast cancer is 1.1 in cluster No.1 and 1.16 in cluster No. 2. The average IAR is 29.4 in No. 1, 31.4 in No. 2, and 26.4 outside. In terms of the rate, cluster No. 2 has higher values inside compared with outside, but the average weight inside cluster No. 2 is 0.045, only half of that outside (0.082). So the IAR values in cluster No. 2 have more uncertainty even though the values of IAR are higher. The average weight in cluster No. 1 is 0.091(Figure 3B) and is similar to that of outside, namely, 0.082. The cluster No. 2 in Figure 3B is not detected as a cluster of counties, but a collection of individuals with high chance of dying from breast cancer.

We summarize some information regarding possibly related factors inside/outside the clusters (Table 4). The hospital number is from 2000 Area Resource File (http://sodapop.pop.psu.edu/data-collections/arf). The other factors in this table are from 1994-2003 Behavioral Risk Factor Surveillance System (BRFSS). It is interesting to see that the cluster area in Figure 3A has less poverty and more women ages 50-64 who had a mammogram, but also has less oncology hospitals and more smokers than the country average. This may indicate the numbers of oncology hospitals and smokers are the key factors associated with the high breast cancer mortality in this detected cluster area. Cluster No. 1 in 3B is very large and actually covers areas including most of the cluster areas in 3A. For the two clusters in 3B, we notice that they are both in higher poverty areas with fewer hospitals, and fewer people with advanced education. There are more smokers in cluster No.1(3B), but less in cluster No. 2, compared with areas outside the clusters.

5 Conclusion and Discussion

The proposed weighted normal scan methods take into account the uncertainty of continuous measures within cells (e.g., blocks, tracts, counties). The standard normal scan method (www.satscan.org) originally developed for individual data can also be treated as a special case of the weighted normal scan method with homogeneous weight ($\delta_z = 1, z \in G$). This new tool can be widely used for any spatial clustering study based on continuous measures (symmetric or not) because the hypothesis testing is based on a permutation test, which is very robust to the varying distributions of the data. The applications presented in this article are selected examples of the use of the proposed new methods on different types of continuous data. However, the methods are not restricted to analyze only the mortality rates and survival rates. If the continuous data are highly skewed, a transformation is recommended.

Researchers should be careful when choosing the available scan methods for a clustering study. First, the choice of the method depends on the purpose of the study and the hypothesis

that the researchers want to evaluate. For example, a patient with cancer may like to know if he/she has high risk of cancer mortality or not (the risk can be affected by both personal and environmental factors). In this case, the study subject should be the individual. A government health care planner is interested in the collection of counties with higher cancer mortality rates on average after adjusting for county uncertainty or county population, in order to decide how to allocate resources more efficiently to counties. In this case, the study subject should be the counties.

Second, the selection of the method also depends on the available data structure. In the case that there is only population data released at cell level (county or state), the proper choice for the researchers is to study the spatial pattern of the cells and apply the weighted normal model on the regional data with uncertainty within each cell.

Note that we use permutation rather than simulation in our appraoch. For count data, when simulating the cases, we only assume the same relative risk of disease, or equal chance of being an event in a Poisson model or a Bernoulli model. This assumption is very general. But simulating continuous data requires more assumptions. It is a very restrictive set-up, since we have to define the exact true distribution (with a specific mean and variance if it's normal) under the null hypothesis. Therefore, simulation would make the weighted normal method (developed for continuous data analysis) very data/application dependent, which is not preferred.

In this article, we have illustrated the use of sample size as a substitute for the variance within one cell when the variance is not available. In this situation, it is also possible to assume that the weights δ_z follow a random distribution (e.g., Gamma distribution) that allows some variability around the point estimate of the mean weight. The distribution of the weights (δ_z) could be assumed to be the same for all z in G (e.g., homogeneously distributed in G) or inhomogeneously distributed.

With a circular window, we sometimes include low rate areas in the detected cluster of high rates because of the restriction of the shape. With more flexible search window (Patil et al.

2004, Tango and Takahashi 2005, Kulldorff et al. 2006), we could have better precision in the location of the true clusters. The current weighted normal scan model is only designed for purely spatial analysis. It can be easily extended to a space-time scan statistic when time series data is available (Kulldorff, Athas, Feuer, Miller and Key 1998).

The approach proposed in this paper has similar goals of local indicators of spatial association (LISAs) (Anselin 1995). The main purpose of LISAs is to provide a local measure of similarity between each region's associated value (a count or a rate) and those of nearby regions. A map of the regional LISA values can provide insight into the location of regions with comparatively high or low local association with neighboring values. One of the most popular LISAs is a local version of Moran's I, which also adjusts for trend and spatial heterogeneities in regional variances. However, there are many limitations in the use of LISAs for cluster detection and in the applications of LISAs to public health data (Waller and Gotway 2004). The proposed method is more proper for detecting unusual clusters of regions with similar behavior in terms of rates or other continuous values beyond rates (e.g. survival times).

The method developed here provides a new tool to study the spatial heterogeneity and geographic clustering pattern of cells (such as counties, tracts, hospitals, and schools) instead of individual persons in a broader range of population data. This new model will be included in a future release of SaTScan software (www.satscan.org).

Acknowledgments

The authors wish to thank the editor, associate editor, and the referees for their valuable comments and suggestions. Kulldorff gratefully acknowledge the support of National Institute of Child Health and Human Development grant R01 HD048852.

References

American Cancer Society (ACS). (2007), *Cancer Facts & Figures 2007,*

online linkage: http://www.cancer.org/downloads/STT/CAFF2007PWSecured.pdf.

Anselin L. (1995), "Local indicators of spatial association: LISA," *Geographical Analysis,* 27(2), 93-116.

Armitage P. and Berry G. (1994), *Statistical methods in medical research 3rd edition,* Oxford: Blackwell scientific publications.

Besag J and Newell J. (1991), "The detection of clusters in rare diseases," *Journal of the Royal Statistical Society Series A,* 154, 327-333.

Besag J., York J., and Mollie A. (1991), "Bayesian image restoration with two applications in spatial statistics," *Annals of the Institute of Statistical Mathematics,* 43, 1-59.

Bithell J.F. (1995), "The choice of test for detecting raised disease risk near a point source," *Statistics in Medicine,* 14(21), 2309-22.

California office of statewide health planning and development. (2005), *Medical service study areas,* online linkage: http://gforge.casil.ucdavis.edu/docman/view.php/60/102/MSSA2000.zip.

Canto M.T., Anderson W.F., and Brawley O. (2001), "Geographic variation in breast cancer mortality for white and black women: 1986-1995," *CA cancer J Clin,* 51, 367.

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

Cox D.R. (1972), "Regression models and life tables (with discussion)," *Journal of the Royal Statistical Society Series B,* 34, 87-220.

Curtin L.R., and Klein R.J. (1995), "Direct standardization," *Healthy people 2000 statistical notes,* Centers for disease control and prevention, national center for health statistics.

Diggle P. (1990), "A point process modeling approach to raised incidence of a rare phenomenon

in the vicinity of a prespecified point," *Journal of the Royal Statistical Society Series A,* 153, 349-362.

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.

Dwass M. (1957), "Modified randomization tests for nonparametric hypothesis," *Annals of mathematical statistics,* 28, 181-187.

Goodwin J.S., Freeman J.L., Mahnken J.D., Freeman D.H. and Nattinger A.B. (2002), "Geographic variations in breast cancer survival among older women: Implications for quality of breast cancer care," *J. Geontol. Med Sci.,* 57, M401-M406.

Huang L., Kulldorff M., and Gregorio D. (2006), "A Spatial Scan Statistic for Survival Data," *Biometrics,* 63, 109-118.

Huang L., Pickle L.W., Stinchcomb D., and Feuer E.J. (2007), "Detection of Spatial Clusters: Application to Cancer Survival as a Continuous Outcome," *Epidemiology,* 18, 73-87.

Jacquez G.M. and Greiling D.A. (2003), "Geographic boundaries in breast, lung and colorectal cancers in relation to exposure to air toxics in Long Island, New York," *International Journal of Health Geographics,* 2, 4.

Jung I., Kulldorff M., and Klassen A. (2007), "A spatial scan statistic for ordinal data," *Statistics in Medicine,* 26(7), 1594-1607.

Kim H-J., Fay M.P., Feuer E.J., and Midthune D.N. (2000), "Permutation tests for joinpoint regression with applications to cancer rates," *Statistics in Medicine,* 19, 335-351.

Klein J.P. and Moseschberger M.L. (1997), *Survival analysis techniques for censored and truncated data,* New York: Springer-verlag New York, Inc.

Knorr-Held L. and Rasser G. (2000), "Bayesian detection of clusters and discontinuities in

disease maps," *Biometrics,* 56(1), 13-21.

Kulldorff M. (1997), "A spatial scan statistic," *Communications in Statistics: Theory and Methods,* 26, 1481-1496.

Kulldorff M., Athas W., Feuer E., Miller B., and Key C. (1998), "Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, New Mexico," *American Journal of Public Health,* 88(9), 1377-1380.

Kulldorff M., Huang L., Pickle L., and Duczmal L. (2006), "An elliptic spatial scan statistic," *Statistics in Medicine,* 25(22), 3929-43.

Lawson A.B. (1993), "On the analysis of mortality events associated with a prespecified fixed point," *Journal of Royal Statistical Society Series A,* 156, 363-377.

Lawson A.B., Browne W.J., and Widal Rodeiro C.L. (2003), *Disease mapping with WinBUGS and MLwiN,* London: Wiley.

Moran PAP. (1950), "Notes on continuous stochastic phenomena," *Biometrika,* 37, 17-23.

Oden N. (1995), "Adjusting Moran's I for population density," *Statistics in Medicine,* 14, 17-26.

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

Pickle L.W. and White A.A. (1995), "Effects of the choice of age-adjustment method on maps of death rates," *Statistics in Medicine,* 14, 615-627.

Richardson S, Thomson A, Best N, and Elliot P. (2004), "Interpreting posterior relative risk estimates in disease-mapping studies," *Environmental Health Perspectives*, 112, 1016-1025.

Sheehan T.J., Dechello L.M., Kulldorff M., Gregorio D.I., Gershman S., and Mroszcayk M. (2004), "The geographic distribution of breast cancer incidence in Massachusetts 1988 to 1997,

adjusted for covariates," *International Journal of Health Geographics,* 3, 17.

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.

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

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

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

Waller L.A. and Gotway C.A. (2004), *Applied spatial statistics for public health data,* New Jersey: A John Wiley and Sons, Inc.

Waller L.A., Turnbull B.W., Clark L.C., and Nasca P. (1992), "Chronic disease surveillance and testing of clustering of disease and exposure: Application to leukemia incidence and TCE-contaminated dumpsites in upstate New York," *Environmetrics,* 3, 281-300.

Yu B., Tiwari R.C., Cronin K.A., and Feuer E.J. (2004), "Cure fraction estimation from the mixture models for grouped survival data," *Statistics in Medicine,* 23, 1733-1747.

Table 1: Power, r_T (sensitivity), r_D (positive predictive value) of the weighted normal scan method under varying cases. The true cluster for all cases is centered at cell (6,3) with radius 2 that includes 13 cells. The minimum window size is 2 cells and maximum search window size is 50% of the total cells.

Case	Distribution		Weight δ	c	η	Power (%)	r_T	r_D
	in cluster	outside						
1	N($c\sqrt{2}$,1)	N(0,1)	1 everywhere	0.5	1	25	0.60	0.50
				1.0	1	88	0.92	0.89
				1.5	1	100	0.99	0.99
				2.0	1	100	1.00	1.00
				3.0	1	100	1.00	1.00
2a	N($c\sqrt{2}$,1)	N(0,1)	η at center of cluster (6,3);	1.5	2	100	0.99	0.99
			1 elsewhere	1.5	4	100	0.99	0.99
				1.5	8	99	0.99	0.98
				1.5	10	98	0.98	0.98
				1.5	100	97	0.92	0.96
				1.5	1000	96	0.91	0.95
2b			η at (6,3),(6,2),(7,3), 1 elsewhere	1.5	1000	63	0.53	0.80
			η in circle centered at (6,3),	1.5	1000	48	0.37	0.68
			with radius 1; 1 elsewhere					
3	N($c\sqrt{2}$,1)	N(0,1)	η in true cluster; 1 outside	1.5	2	100	0.99	0.99
				1.5	4	100	0.98	0.99
				1.5	8	99	0.97	0.96
				1.5	10	99	0.97	0.95
				1.5	100	34	0.65	0.56
				1.5	1000	12	0.43	0.40
4	N($c\sqrt{2}$,1)	N(0,1)	1 inside true cluster; η outside	1.5	2	99	0.99	0.98
				1.5	4	71	0.94	0.90
				1.5	8	12	0.68	0.57
				1.5	10	8	0.56	0.44
				1.5	100	5	0.21	0.14
				1.5	1000	5	0.19	0.13
5	N($c\sqrt{2}, \eta$)	N(0,1)	1 everywhere	1.5	2	99	0.95	0.98
				1.5	4	93	0.84	0.96
				1.5	8	81	0.70	0.93
				1.5	10	76	0.66	0.90
				1.5	100	44	0.41	0.75
				1.5	1000	34	0.35	0.67

Table 2: Performance of the weighted normal scan model on data with varying distributions. The $\eta = 1$ implies that we have homogeneous weight 1 in G. The $\eta > 1$ in this table implies that the weight $\delta_z(= \eta)$ is higher at one cell (6,5) in the true cluster, and 1 elsewhere. The data have $D(\mu, \sigma^2) = D(0, 1)$ inside the true cluster and $D(\mu, \sigma^2) = D(0 + c\sqrt{2}, 1)$ inside. D is selected to be Normal, Double exponential (DoubleE), Logistic, Uniform. For Lognormal data, $(\mu, \sigma^2) = (2, 1)$ outside and $(\mu, \sigma^2) = (2 + c\sqrt{2}, 1)$ inside the true cluster. For Poisson data, $(\mu, \sigma^2) = (1, 1)$ outside and $(\mu, \sigma^2) = (1 + c\sqrt{2}, 1 + c\sqrt{2})$ inside the true cluster.

c	η	Power (%)					
		Normal	DoubleE	Logistic	Uniform	Lognormal	Poisson
0.5	1	25	23	25	30	13	19
1.0	1	88	86	86	90	71	63
1.5	1	100	100	100	100	99	95
2.0	1	100	100	100	100	100	99
3.0	1	100	100	100	100	100	100
1.5	2	100	100	100	100	99	95
1.5	4	100	100	100	100	99	93
1.5	8	100	100	100	100	99	91
1.5	10	100	100	100	100	99	90
1.5	100	98	98	99	99	97	88
1.5	1000	98	98	98	99	97	86

c	η	r_T						r_D					
		Normal	DoubleE	Logistic	Uniform	Lognormal	Poisson	Normal	DoubleE	Logistic	Uniform	Lognormal	Poisson
0.5	1	0.60	0.58	0.59	0.54	0.46	0.45	0.50	0.51	0.49	0.50	0.43	0.52
1.0	1	0.92	0.92	0.92	0.92	0.89	0.75	0.89	0.90	0.88	0.89	0.85	0.84
1.5	1	0.99	0.99	0.99	1.00	0.99	0.89	0.99	0.99	0.99	0.99	0.98	0.96
2.0	1	1.00	1.00	1.00	1.00	1.00	0.96	1.00	1.00	1.00	1.00	1.00	0.99
3.0	1	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00
1.5	2	0.99	0.99	0.99	0.99	0.99	0.89	0.99	0.99	0.99	1.00	0.98	0.96
1.5	4	0.99	0.99	0.99	0.99	0.99	0.88	0.99	0.99	0.99	0.99	0.97	0.96
1.5	8	0.99	0.98	0.99	0.98	0.98	0.87	0.98	0.98	0.98	0.99	0.96	0.94
1.5	10	0.98	0.98	0.98	0.98	0.97	0.87	0.97	0.97	0.97	0.97	0.95	0.94
1.5	100	0.93	0.93	0.92	0.92	0.93	0.83	0.81	0.84	0.82	0.84	0.84	0.80
1.5	1000	0.91	0.92	0.91	0.92	0.92	0.82	0.77	0.71	0.74	0.76	0.76	0.70

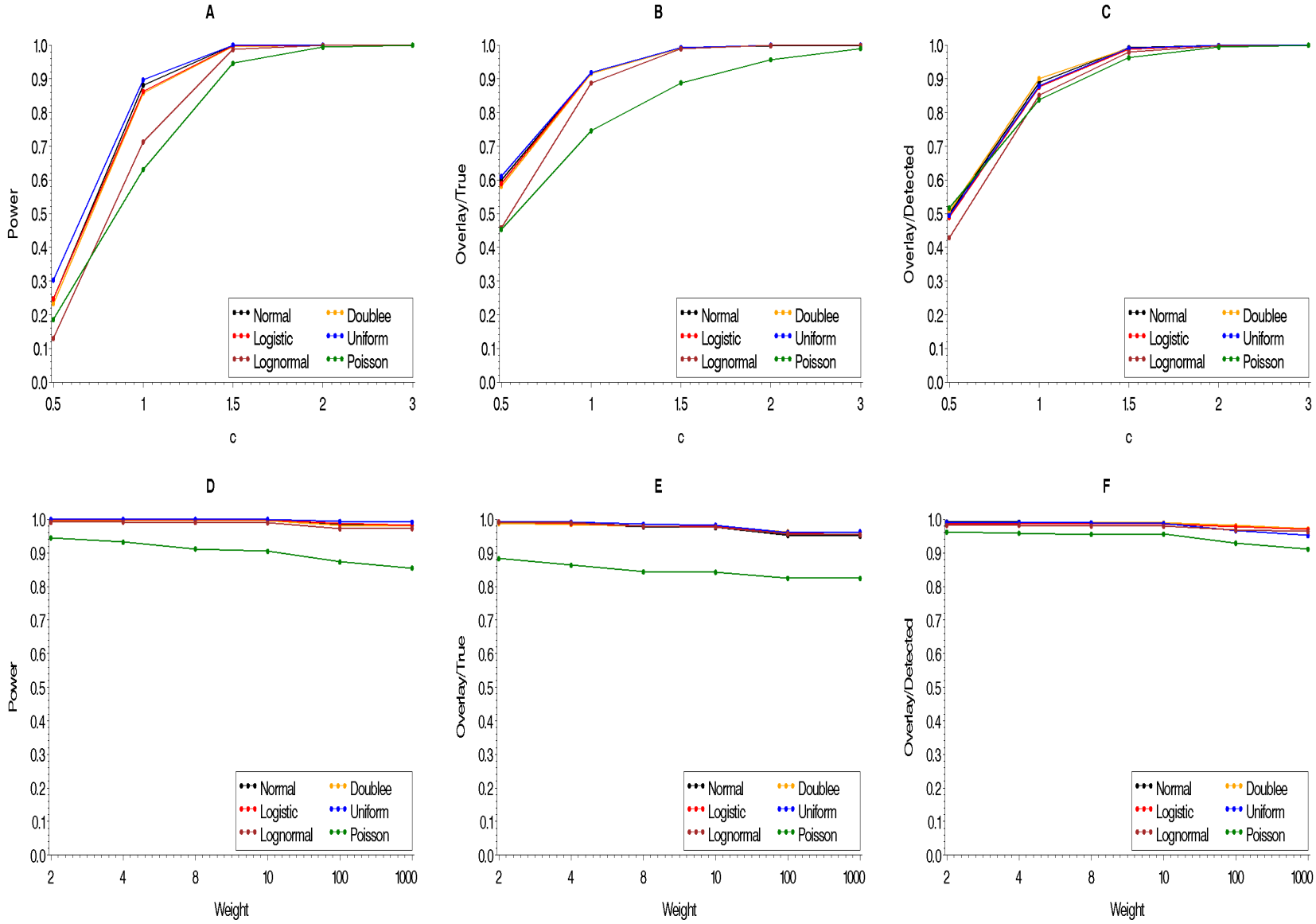

Figure 1: Power, r_T (sensitivity) and r_D(positive predictive value) for data with varying distributions. A (power), B(r_T) and C(r_D) are the results for data with varying c but same η(=1), and D(power), E(r_T) and F(r_D) are the results for data with the same c(=1.5) but varying η as shown in Table 1.

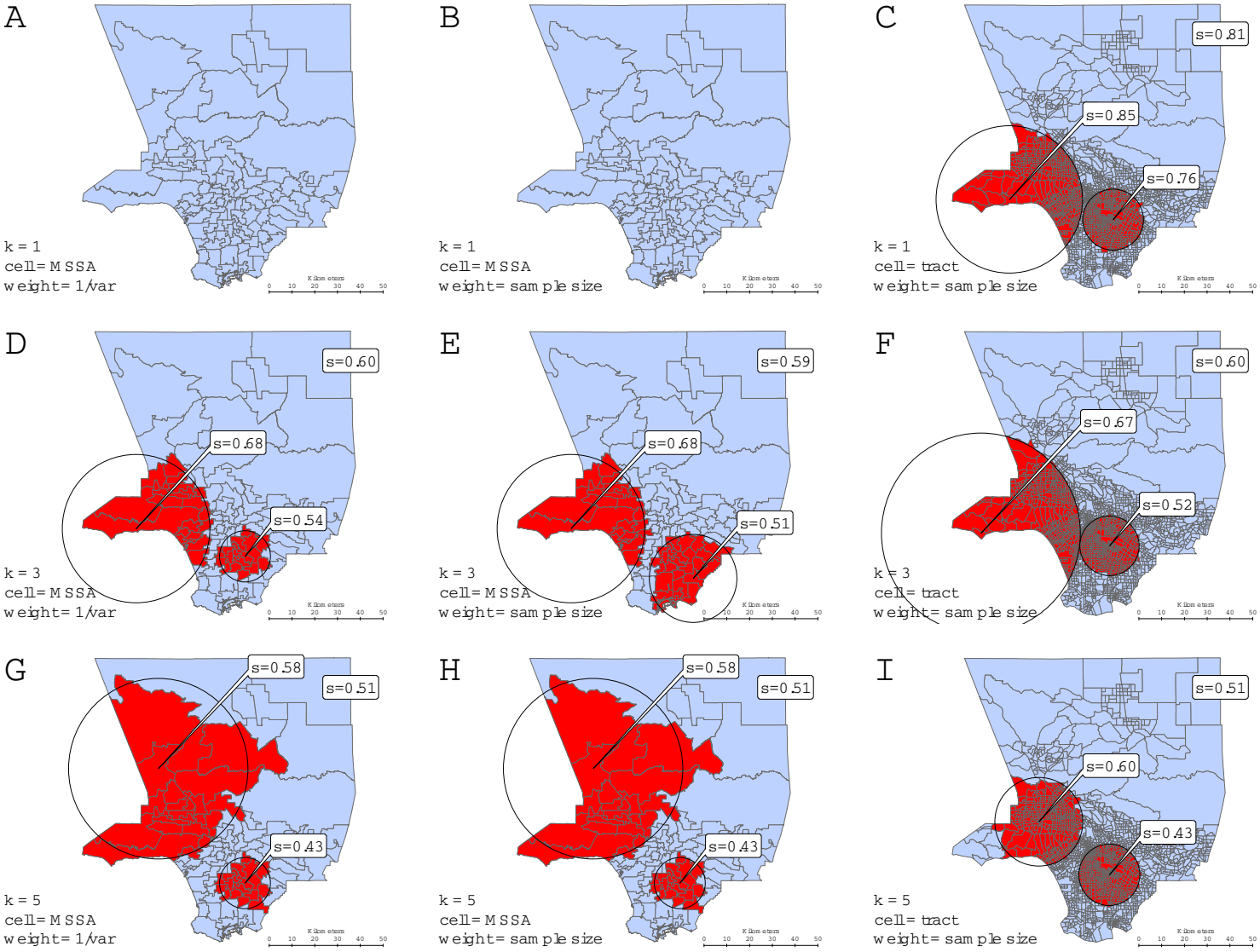

Figure 2: Clusters of k-year survival rate (k=1,3,5) detected in LA by weighted normal scan method for patients diagnosed with stage I and II lung cancer from 1988-2002. The s represents k-year survival rate. The rate numbers in the maps are computed by KM method inside/outside clusters. Other information on the clusters is included in Table 3.

Table 3a: The k-year survival rate (k=1,3,5) for patients with stage I and II lung cancer inside/outside clusters detected by the weighted normal scan method on data with cells as Medical Service Study areas (MSSAs) or tracts. The clusters are shown in Figure 2. The weighted mean is the weighted average of cell survival rates inside/outside clusters with weight as either 1/variance or sample size. The p-value is the p-value of the particular cluster when testing if the mean of k-year survival rate in the cluster is higher/lower than that outside using the scan method. The KM estimates are the k-year survival rate estimated directly from Kaplan Meier method on individual survival times in the related areas. Diff is weighted mean minus KM estimate.

map	cell	weight	k	cluster of high k-year survival rate				cluster of low k-year survival rate				outside clusters		
				weighted mean(%)	p-value	KM est (%)	diff (%)	weighted mean (%)	p-value	KM est (%)	diff (%)	weighted mean (%)	KM est (%)	diff (%)
2A	MSSA	1/variance	1						NA					
2B	MSSA	sample size	1						NA					
2C	Tract	sample size	1	87.26	0.016	84.56	2.7	79.08	0.008	76.09	3.0	83.76	81.08	2.7
2D	MSSA	1/variance	3	70.48	< 0.001	67.52	3.0	56.99	< 0.001	54.26	2.7	62.84	59.72	3.1
2E	MSSA	sample size	3	70.05	< 0.001	67.52	2.5	53.56	< 0.001	50.80	2.8	61.81	59.35	2.5
2F	Tract	sample size	3	69.78	< 0.001	67.44	2.3	55.33	< 0.001	51.94	3.4	63.16	59.70	3.5
2G	MSSA	1/variance	5	60.06	< 0.001	58.33	1.7	42.37	0.002	42.99	-0.6	52.35	50.61	1.7
2H	MSSA	sample size	5	59.87	< 0.001	58.33	1.5	42.22	0.001	42.99	-0.8	52.18	50.61	1.6
2I	Tract	sample size	5	62.19	< 0.001	59.81	2.4	46.91	< 0.001	42.67	4.2	54.23	51.00	3.2

Table 3b: Summary of the factors inside/outside clusters with cell as tract. Advanced education is the percent of adult with at least four year college education. Poverty is the percent of persons living below poverty level. The numbers in the parentheses are the average values in the areas covered by the 3-year high rate cluster, but not covered by the 5-year high rate cluster.

map	risk factors	k	average values inside/outside clusters		
			cluster of high rate	cluster of low rate	outside clusters
2C		1	47.48	26.43	37.91
2F	Median income ($1,000)	3	48.32 (40.47)	25.65	37.88
2I		5	50.51	26.31	38.14
2C		1	34.38	8.54	20.41
2F	Advanced education (%)	3	34.69 (33.48)	8.75	20.37
2I		5	33.36	8.99	22.02
2C		1	51.85	46.51	52.11
2F	Home owner (%)	3	52.84 (43.06)	44.73	52.18
2I		5	57.63	45.90	51.71
2C		1	28.97	17.26	22.78
2F	Health insurance (%)	3	29.15 (25.38)	17.15	22.80
2I		5	30.02	17.56	23.06
2C		1	9.24	22.84	14.19
2F	Poverty (%)	3	8.90 (10.70)	23.65	14.10
2I		5	8.30	22.98	13.86
2C		1	8.95	30.75	7.77
2F	Black (%)	3	8.44 (25.32)	36.42	6.16
2I		5	2.17	35.71	8.63
2C		1	46.59	59.29	51.54
2F	Male (%)	3	46.00 (45.14)	59.69	51.72
2I		5	47.11	59.22	50.63

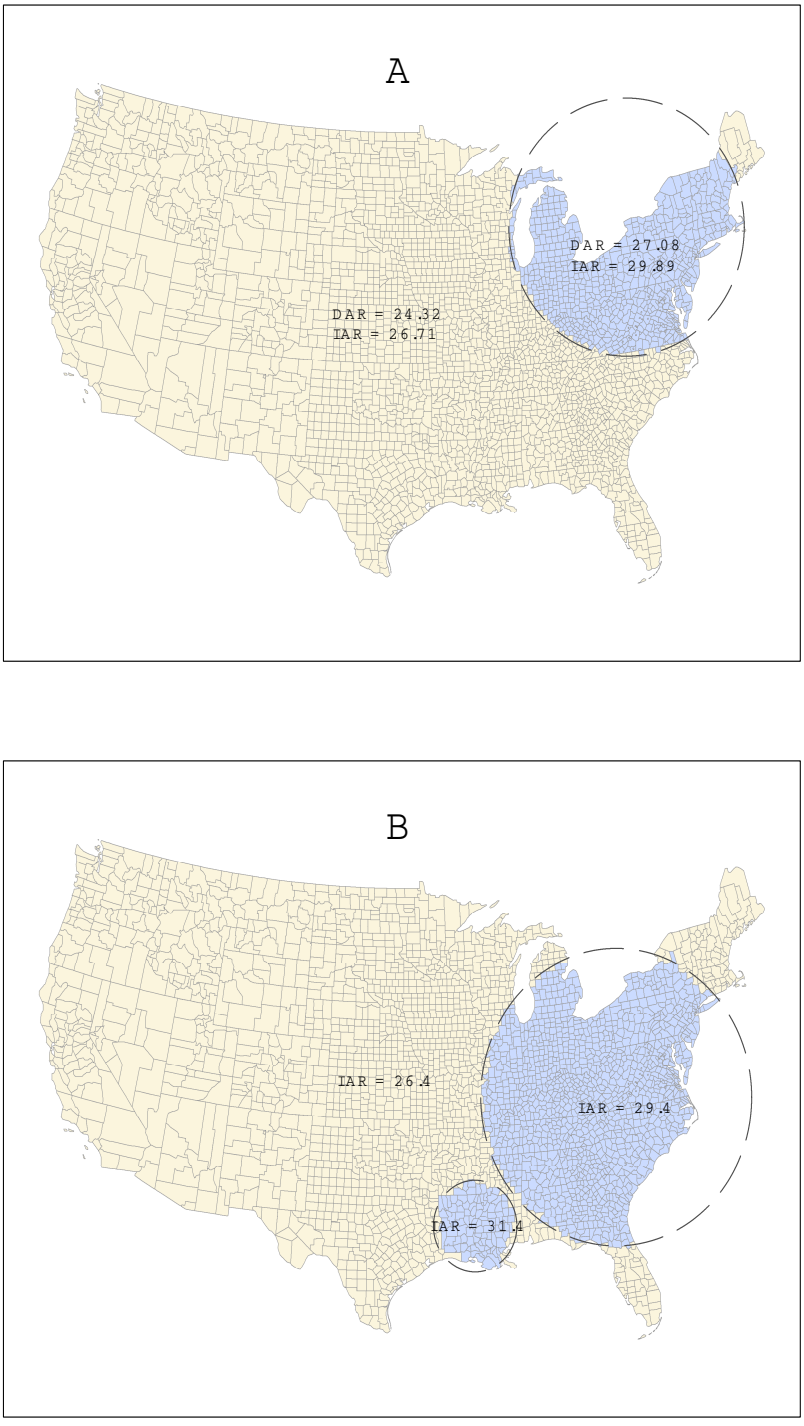

Figure 3: A cluster of high breast cancer mortality rate (1999-2003) after adjusting regional uncertainty (3A) from the weighted normal scan method and high mortality counts adjusting population (3B) from the Poisson model based SaTScan method. DAR is direct age-adjusted mortality rate, and IAR is indirect age-adjusted mortality rate.

Table 4: Average values of the factors over the counties inside/outside the clusters shown in Figure 3. The first three factors are from 2000 census. All numbers are percentages.

Factors	Figure 3A cluster		Figure 3B clusters		
	out	in	out	in No. 1	in No. 2
% of blacks	9.37	7.89	3.76	13.83	36.88
% of adult with 4+ year of college education	15.70	17.98	17.07	15.30	13.48
% of people with income below poverty level	15.16	11.88	13.96	14.32	23.66
2000 county oncology hospitals per 1000 population	1.50	1.28	1.81	1.02	0.95
1999-2003 % women ages 50-64 who had a mammogram in past 2 years	75.66	80.92	75.38	79.31	69.70
1994-1998 % women ages 50-64 who had a mammogram in past 2 years	68.33	73.86	69.04	71.21	58.68
1999-2003 % adults without health insurance	16.41	13.08	16.04	14.56	23.65
1994-1998 % adults without health insurance	15.12	13.13	14.79	14.03	20.80
1999-2003 % of females ages 18+ who ever smoked cigarettes	40.68	45.45	41.09	43.02	36.51
1994-1998 % of females ages 18+ who ever smoked cigarettes	38.30	43.71	38.88	40.71	34.92