A Workflow Spatial Scan Statistic

Luiz Duczmal[1] and David L. Buckeridge[2]

[1]Universidade Federal de Minas Gerais, Brazil
[2]McGill University, Department of Epidemiology, Biostatistics and Occupational Health, Montreal, Canada

ABSTRACT

We propose a modification of the spatial scan statistic that takes account of workflow, which is the movement of individuals between home and work. The objective is to detect clusters of disease in situations where exposure occurs in the workplace, but only home address is available for analysis. In these situations, application of the usual spatial scan statistic does not account for possible differences between home and work address, thereby reducing the power of detection. We describe an extension to the usual spatial scan statistic that uses workflow data to search for disease clusters resulting from workplace exposure. We also present results from simulations that demonstrate the increased power of the workflow scan statistic over the usual scan statistic for detecting clusters arising from exposures in the workplace.

Keywords: Public Health Practice, aberration, cluster detection, spatial scan statistics, evaluation performance testing.

1. Introduction

There are many scenarios where exposure may occur in the workplace. These include exposure to infectious agents or toxins, either accidentally or intentionally. In most of these situations, especially those where the delay between exposure and symptom onset is short, timely detection of symptoms arising from the exposure can facilitate public health intervention to reduce morbidity and mortality.

Surveillance directed towards timely detection of outbreaks increasingly relies on data collected for other purposes (Mandl et al. 2004). Addresses, when available in these data, are usually home addresses, and work addresses are not routinely available. Many surveillance systems that analyze healthcare utilization data rely on the spatial scan statistic (Kulldorff, 1997,1999; Glaz et al. 2001), which uses the geographic location of cases to search for disease clusters (Odoi et al. 2004; Heffernan et al. 2004; Andrade et al. 2004). If home address is used for analysis of cases when exposure occurred in the

[1] Corresponding author: duczmal@est.ufmg.br, Universidade Federal de Minas Gerais, Statistics Dept., Campus Pampulha, Belo Horizonte, MG 31270-901 Brazil. Phone:55-31-3499-5900 fax:55-31-3499-5924

workplace, then the cases may appear falsely geographically dispersed, and this could reduce the power of detection.

In this paper, we describe a modification to the spatial scan statistic that takes account of workflow, which is the movement of individuals between home and work. The objective is to detect clusters of disease in situations where exposure occurs in the workplace, but only home address is available for analysis. The usual spatial scan statistic is reviewed briefly in section 2, before we present our proposed extension to account for workflow in section 3. We then evaluate our proposed statistic in section 4, describing our approach to creating workflow matrices, presenting our method for simulating disease outbreaks in the workplace, and then giving the results of our study to evaluate the power of the usual and workflow spatial scan statistics. Section 5 presents our conclusions.

2. Kulldorff´s Spatial Scan Statistic

Consider a study region with m cells (small geographic areas, such as ZIP codes or counties), with total population N and C total cases. Define the zone z as the collection of cells whose centroids are within a circle of given center and radius. Under the null hypothesis we assume that there is no cluster in the map, and that the number of cases in each cell is Poisson distributed according to its population. Let L_0 be the likelihood under the null hypothesis, and let $L(z)$ be the likelihood under the alternative hypothesis that there is a cluster in the zone z. The zone with the maximum likelihood is defined as *the most likely cluster*. We assume under the null hypothesis that there is no cluster in the map and the number of cases in each cell is Poisson distributed, with expected values proportional to its population size. If $\mu(z)$ is the expected number of cases inside the zone z under the null hypothesis, and $c(z)$ is the observed number of cases inside z, it can be shown that

$$U(z) = \frac{L(z)}{L_0} = \left(\frac{c(z)}{\mu(z)}\right)^{c(z)} \left(\frac{C-c(z)}{C-\mu(z)}\right)^{C-c(z)}$$

when $c(z) > \mu(z)$, and 1 otherwise. The test statistic is $\max_z U(z)$. This likelihood ratio, maximized over all the zones, identifies the zone that constitutes the most likely cluster. See Kulldorff (1997) for details.

3. The Workflow Spatial Scan Statistic

Suppose that the study region is partitioned into m cells $Z_1, Z_2, \ldots, Z_m$. Each cell may be a county or a ZIP code, for example. Let $L(k,i)$ be the proportion of the population living in cell Z_k that works at cell Z_i, $k,i = 1,\ldots,m$, and let

$$L(k,0) = 1 - \sum_{i=1}^{m} L(k,i), \quad k = 1,\ldots,m$$

be the proportion of the population living in cell Z_k that works outside the study region. For each $i = 1,\ldots,m$ sort the cells $Z_1,\ldots,Z_m$ by increasing distance from $Z_i = Z_{i_1}$ as

$Z_{i_1}, Z_{i_2}, Z_{i_3}, \ldots, Z_{i_m}$.

For each $r = 1, \ldots, r_{\max}$, $k = 1, \ldots, m$ define

$$A(k,i,r) = \sum_{t=1}^{r} L(k,i_t) .$$

The value $A(k,i,r)$ may be interpreted as the proportion of the population living in cell Z_k that works at the r nearest cells from Z_i (including itself).

For each $r = 1, \ldots, r_{\max}$ and for each $i = 1, \ldots, m$ sort the cells $Z_1, \ldots, Z_m$ as

$Z(1,i,r), Z(2,i,r), \ldots, Z(m,i,r)$

such that

$A(1,i,r) \geq A(2,i,r) \geq \ldots \geq A(m,i,r),$

then build the zones

$Y(1,i,r) = Z(1,i,r),$

$Y(2,i,r) = Z(1,i,r) \cup Z(2,i,r),$

…

$Y(m,i,r) = Z(1,i,r) \cup Z(2,i,r) \cup \ldots \cup Z(m,i,r).$

This ordering is done with the objective of collecting first the cells with the highest proportion of employees working at the r nearest cells from Z_i. For simplicity, consider this proportion as "signal", and the remaining proportion as "noise". This criterion gathers the higher working proportion cells first, trying to build a cluster with high signal-to-noise as early as possible in the process. We delay the addition of the "noisier" cells, even if they have more workers from the r nearest cells from Z_i, because they would diminish the "signal" compared to the added "noise". We note that the zones *Y(k,i,r)* may be possibly non-connected, as we should expect. Observe that for each pair i, r we build a specific sequence

$Y(1,i,r), Y(2,i,r), \ldots, Y(m,i,r),$

defined by the inequalities

$A(1,i,r) \geq A(2,i,r) \geq \cdots \geq A(m,i,r).$

For each pair i, r, and for $t = 1, \ldots, m$, define

$c(t,i,r)$ as the number of observed cases in cell $Z(t,i,r)$,

and

$p(t,i,r)$ as the population in cell $Z(t,i,r)$.

Being $N =$ total population and $C =$ total number of cases in the study region, define

$$u(k,i,r) = \sum_{t=1}^{k} A(t,i,r)\, c(t,i,r)$$

and

$$e(k,i,r) = \frac{C}{N}\sum_{t=1}^{k} A(t,i,r)\,p(t,i,r).$$

The numbers $u(k,i,r)$ and $e(k,i,r)$ have the following interpretation:

$u(k,i,r)$ = observed cases in the zone $Y(k,i,r)$ due to contamination from workers at the r nearest neighbors of cell Z_i,

$e(k,i,r)$ = expected cases in the zone $Y(k,i,r)$ due to contamination from the r nearest neighbors of cell Z_i under the null hypothesis.

Define the *workflow (spatial) scan statistic* as

$$W(Y(k,i,r)) = \left(\frac{u(k,i,r)}{e(k,i,r)}\right)^{u(k,i,r)} \left(\frac{C-u(k,i,r)}{C-e(k,i,r)}\right)^{C-u(k,i,r)}$$

if $u(k,i,r) > e(k,i,r)$, and one otherwise.

In the example shown at Figure 1, with $r=4$ and $i=i_1$, the closest neighbors Z_{i_2}, Z_{i_3} and Z_{i_4} of cell Z_{i_1} are displayed to the right. We will show, as an example, the construction of the zone $Y(3,i_1,4)$, consisting of the union of the cells $Z(1,i,4)$, $Z(2,i,4)$ and $Z(3,i,4)$. These last three cells are displayed to the left. We note that Figure 1 is just a schematic diagram, and it may be that some cells to the left and to the right are the same. The four shadings corresponding to the four cells Z_{i_1}, Z_{i_2}, Z_{i_3} and Z_{i_4}. The shaded slices within the cells $Z(1,i,4)$, $Z(2,i,4)$ and $Z(3,i,4)$ to the left indicate the proportions $A(1,i,4)$, $A(2,i,4)$ and $A(3,i,4)$ respectively. This ordering is induced by the inequality $A(1,i,4) \geq A(2,i,4) \geq A(3,i,4)$, for $i=i_1$ and $r=4$. The term $u(k,i,r)/e(k,i,r)$ in the workflow scan statistic indicates the relative incidence of observed cases within the shaded area. The other term, $(C-u(k,i,r))/(C-e(k,i,r))$, indicates the relative incidence of observed cases in the rest of the study region. This term is associated with the non-shaded area of the cells $Z(1,i,4)$, $Z(2,i,4)$ and $Z(3,i,4)$, plus the remaining cells in the study region other than $Z(1,i,4)$, $Z(2,i,4)$ and $Z(3,i,4)$. As noted before, the ordering of the inequality above depends on *r* and *i*. When $r=2$ or $r=3$ for instance, the respective orderings are different, as could be checked from figure 1.

For each i, at most $mr_{\max}$ evaluations of the statistic $W(Y(k,i,r))$ are performed. Repeating this process for $i=1,\ldots,m$, we end up with a total of $m^2 r_{\max}$ evaluations. We find the maximum value of the workflow scan statistic W for some zone $Y(K,I,R)$, called the most likely cluster for the observed cases; it has the highest likelihood of harboring the cases due to contamination from the workplace consisting of the nearest neighbors of Z_I.

Similarly to what is done with the usual spatial scan statistic, the maximum value of W is compared against the corresponding workflow scan values obtained through thousands of Monte Carlo simulations. To do this, we use randomly Poisson distributed cases according to the null hypothesis, that there is no cluster in the map, and the number of cases in each cell has expected values proportional to its population size. The p-value

of the observed cluster is then the proportion of likelihood ratio values for simulated runs that are higher than the observed value (Dwass, 1957).
§The effect of the workflow scan statistic is to "pull back" the scattered workers that were contaminated in the workplace. The $L(k,i)$ factors act in order to account only for the relative importance of cases that may be due to the workplace to living place flow. When $L(k,i)=1$ for $k=i$, and 0 otherwise, we go back to the usual spatial scan test. We implemented the workflow scan statistics in the C programming language and the software is available from the corresponding author.

4. Applying and Evaluating the Workflow Scan Statistic

In this section, we examine the power of the workflow scan statistic to detect simulated disease clusters and compare the performance of the workflow scan to that of the usual spatial scan statistic. We first describe the study region for our evaluation and the derivation of workflow matrices for two geographical zonings, counties and ZIP codes. We then describe our approach to generating simulated outbreaks, and we finish with the results of our evaluation.

4.1 The Study Region and Derivation of Workflow Matrices

We use the region surrounding Norfolk, Virginia as the setting to evaluate the workflow scan statistic (Figure 3). We define the region to include the 35 counties and 158 ZIP codes surrounding the Norfolk area. This covers approximately 160 by 200 kilometers and includes approximately 1.8 million individuals. We use data from the 2000 US census to estimate workflow between counties and between ZIP codes. The census data describe for each county, the number of employed individuals that travel to another destination county for employment (US Census Bureau, 2003a). We calculate the workflow matrix at the county level as:

$$\mathrm{L}(\mathrm{County}_k, \mathrm{County}_i) = \frac{\#(\text{employed residents of } \mathrm{County}_k \text{ who work in } \mathrm{County}_i)}{\#(\text{employed residents of } \mathrm{County}_k)}$$

The Census Bureau does not report workflow by ZIP code. Many healthcare data are available with home ZIP code as the only geographic identifier however, and health data are therefore often analyzed by ZIP code. We define the following approach to estimating workflow between ZIP codes:

$$\mathrm{L}(\mathrm{ZIP}_k, \mathrm{ZIP}_i) = \mathrm{P}(\mathrm{County}_k \mid \mathrm{ZIP}_k) \times \mathrm{L}(\mathrm{County}_k, \mathrm{County}_i) \times \mathrm{P}(\mathrm{ZIP}_i \mid \mathrm{County}_i)$$

We set the first term in this equation to 1 if the centroid of ZIP_k falls within County_k, and 0 otherwise. The middle term is taken directly from census county-level workflow data as described previously. We calculate the final term using census data that describe the number of employees in each ZIP code (US Census Bureau, 2003b):

$$\mathrm{P}(\mathrm{ZIP}_i \mid \mathrm{County}_i) = \#(\text{employees in } \mathrm{ZIP}_i) / S(\mathrm{i})$$

where $S(\mathrm{i})$ is the sum of #(employees in ZIP codes which have centroids within County_i).

Deriving ZIP workflow from county workflow is an example of the change of support problem, and solutions to this problem invariably require assumptions (Gelfand et al.

2001). Here, we assume that the distribution of outflows from a ZIP to a destination county is the same as the distribution of the outflows for the county containing the centroid of the ZIP. We also assume that the distribution of inflows to ZIPs within a county follows the distribution of employees across ZIPs within the county. Our overall approach to modeling the flow of employed individuals between locations considers the aggregate flow of employed individuals on a region-wide basis and has the advantage of being relatively simple to implement. A spatial interaction or gravity model is another approach to modeling flows that, while requiring more effort to implement, may prove more flexible and offer additional insight into workflow patterns (Haynes and Fotheringham, 1984).

4.2 Generation of Simulated Outbreaks

In the simplest situation, under the alternative hypothesis of the usual spatial scan statistic, cases are distributed randomly according to a Poisson model using a single circular cluster. More elaborate instances of alternative hypotheses include irregular and/or multiple disconnected clusters (Kulldorff et al. 2003). We establish a relative risk greater than one and identical in each cell within the cluster, and equal to one for every cell outside the cluster. The relative risks were defined such that if the exact location of the real cluster was known in advance, the power to detect it should be 0.999.

In the simplest scenario of the alternative hypothesis, using the workflow scan statistic we build test clusters for the alternative hypothesis, which are not circular clusters by home address, but rather are circular by work location. Non circular test clusters will also be considered. We first define a zone w_0 consisting of r_0 cells, and define $b(i) = 1$ if $Z_i \in w_0$, and 0 otherwise.

If cell Z_k has population p_k, then there are

$$n = \sum_{k=1}^{m} \sum_{i=1}^{m} p_k L(k,i) b(i)$$

employees during working hours in cluster w_0, that will be exposed to contamination with relative risk RR_{w_0}. The remaining workers will be exposed with relative risk 1. Then, we scatter the cases using the workflow matrix: during working hours each cell Z_k should have relative risk

$$RR(k) = \sum_{i=1}^{m} L(k,i) t(i),$$

where

$t(i) = RR_{w_0}$, if $b(i) = 1$,

and 1 otherwise. The simulated cases are Poisson distributed according to these relative risks. This process is repeated thousands of times for the Monte Carlo power evaluation. The net flow, i.e., the number of workers entering minus the number of workers leaving zone z during working hours is:

$$flow(z) = \sum_{Z_k \notin z} \sum_{Z_i \in z} p_k L(k,i) - \sum_{Z_k \in z} \sum_{Z_i \notin z} p_k L(k,i).$$

A zone z will be called an *inflow* zone when $flow(z) > 0$; otherwise, it will be called an *outflow* zone. Of course we are primarily interested in studying inflow zones. Within an inflow zone, exposure during working hours produces geographic dispersion that is partially recovered by the workflow scan statistic.

In addition to examining the performance of our statistic in detecting simulated circular clusters with equal risk throughout the cluster, we also evaluated the ability of the workflow scan statistic to detect a simulated inhalational anthrax outbreak superimposed onto real outpatient visit records. The anthrax spores, and resulting cases, were dispersed from a point release according to a Gaussian distribution with and without wind (Figure 4) and the baseline cases were the average weekly number of visits to outpatient clinics for respiratory conditions. Baseline data were drawn from the records of the TRICARE Health Maintenance Organization and were categorized by patient residential ZIP code. Buckeridge describes the baseline data and the dispersion and infection models in detail elsewhere (Buckeridge, 2005). For both the wind and no-wind scenarios, there were 5,210 baseline cases. In the no-wind scenario there were 100 additional cases due to inhalational anthrax exposure and in the wind scenario there were 150 additional outbreak cases superimposed onto the baseline. We used a higher number of cases in the wind scenario to make the maximum relative risks similar in both scenarios. The Monte Carlo procedure was repeated 10,000 times under the null and alternative hypothesis.

4.3 Evaluation Results

Using the framework described in section 4.2, we conducted power evaluations for both the workflow scan and the usual scan statistics at the County and ZIP code levels using wholly simulated data and simulated clusters superimposed onto real baseline data. The county-level example uses a study region with 35 counties in the neighborhood of Norfolk, VA. The ZIP-level examples use the same study region partitioned in 158 ZIP codes. The workflow matrices are not displayed here, and may be found at http://www.est.ufmg.br/~duczmal. County-level results are displayed in Table 1 for selected zones, with the cells shown in the first column. In both cases the level of significance is $\alpha = 0.05$. We performed 100,000 Monte Carlo simulations of both statistics under the null hypothesis for every line of Table 1, and 10,000 simulations under the alternative hypothesis for each table entry. Cells 28 and 29 isolated perform significantly better under the workflow scan – their flows are positive, respectively +4,413 and +9,511. When cells 28 and 29 are taken together (flow +13,924), the difference between the different algorithms are not so high, due to the fact that there is a high flow exchange between them. When cell 31 is added to the group formed by cells 28 and 29 (at positive flow +10,565), we observe higher power in both algorithms. The following three lines of Table 1 show a similar situation, when grouping successively the cells 30, 26 and 32. The isolated cells 13 and 25 did not perform well under the Workflow Scan, because they are outflow cells (with negative flows respectively -899, -8,052), as displayed in the last two lines of Table 1.

The power difference between the workflow and the usual scan statistics can be considerable when there are multiple population concentrations with less symmetrical distribution. That scenario is more likely to happen when the cells are smaller, e.g. for

ZIP code areas. Figure 3 displays 158 ZIP codes for the Norfolk neighborhood area. The selected ZIP codes represented by the letters A-L were used in the power simulations that relied on wholly simulated data. The results displayed in Table 2 are from the sample power evaluations using wholly simulated data, and these results were typical of the study region. The use of smaller subdivisions allowed a sharper delineation of the workflow, and as a consequence, relatively improved power results using the workflow scan statistic, compared to the usual scan statistic. The gain in power was at least 0.16 in most situations. The clusters ADEKL and DEK in Table 2 perform poorly using both scans, but even in these cases the workflow scan attained higher power. The last five clusters in Table 2 are examples of multiple clusters, i.e., two or more outbreaks occur simultaneously in separated locations in the map. Even in these cases the workflow algorithm performs significantly better than the usual scan.

The results from the power evaluations using simulated outbreaks superimposed onto real baseline data were similar to the results obtained using wholly simulated data. The power of the workflow scan statistic to detect the simulated inhalational anthrax outbreak in the no-wind scenario was 0.798 and the corresponding power for usual scan statistic in the same scenario was 0.729. In the scenario with wind, the difference between the power of the two statistics was greater, with a power of 0.780 for the workflow scan and a power of 0.681 for the usual scan statistic.

5. Conclusion

We have derived an extension to the spatial scan statistic that accounts for the mobility of individuals between home address and workplace. An analyst can use the workflow scan statistic to search for disease clusters due to workplace exposure when health records contain only residential address. Simulation studies demonstrate that in most scenarios, the workflow scan statistic has greater power than the usual scan statistic for detecting disease outbreaks due to workplace exposures. The workflow scan statistic is particularly useful when clusters are not circularly symmetrical, and thus more easily recognized by the workflow scan than by the usual spatial scan algorithm. In these scenarios the workflow scan has greater power than the usual scan when the exposure occurs at an inflow zone. The opposite situation happens with an outflow zone. The gain in power of the workflow scan over the usual scan is enhanced when events are grouped into smaller zones, such as ZIP codes, as opposed to larger zones, such as counties. This allows for a sharper delineation of workplaces and more precise workflow matrices, increasing further the power of detection of the workflow scan statistic. Workflow data are not routinely available by ZIP code, however, and the error introduced when estimating flows between ZIP codes may offset some of the benefit gained from using smaller zones.

There are some practical issues that a surveillance analyst should consider before using the workflow scan statistic. One issue is that available surveillance data will likely include records for non-employed individuals as well as workers. A simple solution to this problem is to apply the workflow scan statistic only to records for individuals that are likely to be employed, for example records for adults between 18 and 65 years of age. A more general solution is to develop a mobility model for non-employed individuals. Such a model could draw on census data describing transportation (USDOT 2004) and

activities (USDOL 2005) of non-employed individuals, but these data are less specific than workflow, and the resulting model would require careful validation. Another issue to consider is that individuals travel to work at different times of the day. Time-specific workflow data are available at the national and regional level (USDOT 2005), so one could also extend our approach to address this issue.

When interpreting the results of the workflow scan statistic, an analyst should also be aware of the ecological nature of the test. In the workflow scan, the relationship between disease outcomes and workflows are not known at the individual level. In other words, there is no guarantee that excess cases of disease among residents of a small area are, or are not, occurring among those residents employed in a specific small area. This may pose a problem if there is a confounding variable that is associated with both the disease and work location. For example, employees of a particular company may attend a picnic away from their worksite on the weekend with their families. Exposure to a disease agent at the picnic may lead to a disease outbreak in the following week among employees and family members. The workflow scan may identify that outbreak as attributable to exposure at the worksite. While in the strict sense the inference that the disease cases are caused by an exposure at the worksite is incorrect in this example, knowledge that the disease is somehow related to the work location may still facilitate public health investigation of the outbreak. Nevertheless, an analyst should be aware of the potential for ecological bias when inferring causality.

Further refinements of the workflow scan statistic may include clusters that are not circular by work location, but instead irregularly shaped ones (Kulldorff et al. 2005, Duczmal et al.2004, Patil et al. 2004). Space-time clusters extensions to the proposed method are also straightforward (Kulldorff 2001).

Acknowledgments

The authors thank the editor and the reviewers for their insightful comments, and also Martin Kulldorff for critically reading an early version of this manuscript.

6. References

1. Mandl KD, Overhage JM, Wagner MM, Lober WB, Sebastiani P, Mostashari F, Pavlin JA, Gesteland PH, Treadwell T, Koski E, Hutwagner L, Buckeridge DL, Aller RD, Grannis S. Implementing Syndromic Surveillance: A Practical Guide Informe bytheEarlyExperience, *J. Am. Med. Inform. Assoc*, 2004;11:2,141-150.
2. Kulldorff M. A spatial scan statistic. *Communications in Statistics: Theory and Methods*, 1997; 26**:**1481-1496.
3. Kulldorff M. Spatial scan statistics: Models, calculations and applications. In *Scan Statistics and Applications*, Glaz and Balakrishnan (eds.). Boston: Birkhauser, 1999; 303-322.
4. Glaz J, Naus JI, Wallenstein S. *Scan Statistics*. Springer Verlag: New York, 2001.
5. Odoi A, Martin SW, Michel P, Middleton D, Holt J, Wilson J. Investigation of clusters of giardiasis using GIS and a spatial scan statistic. *Int. J. of Health Geographics* 2004, 3:11

6. Heffernan R, Mostashari F, Das D, Karpati A, Kulldorff M, Weiss D. Syndromic surveillance in public health practice, New York City. *Emerging Infectious Diseases* 2004, 10:858
7. Andrade ALSS, Silva SA, Martelli CMT, Oliveira RM, Morais Neto OL, Siqueira Júnior JB, Melo LK, Di Fábio JL. Population-based surveillance of pediatric pneumonia: use of spatial analysis in an urban area of Central Brazil. *Cadernos de Saúde Pública,* 2004 20:2, 411-421
8. Dwass, M. Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, 1957; 28:181-187
9. US Census Bureau, *County-To-County Worker Flow Files*, March 06, 2003, U.S. Census Bureau, Population Division.
10. US Census Bureau, *ZIP Business Patterns*, March 06, 2003, U.S. Census Bureau, Population Division.
11. Gelfand, AE, Zhu L, Carlin BP, On the change of support problem for spatio-temporal data, *Biostatistics*, 2001;2:1,31-45.
12. Haynes, K.E. and Fotheringham, A.S. (1984) Gravity and Spatial Interaction Models. Sage. London.
13. Kulldorff M, Tango T, Park PJ. Power comparisons for disease clustering tests. *Computational Statistics and Data Analysis*, 2003; 42:665-684.
14. Kulldorff M, Huang L, Pickle L, Duczmal L. An elliptic spatial scan statistic. S*ubmitted*, 2005.
15. Duczmal L, Assunção R. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics and Data Analysis*, 2004; 45:269-286.
16. Patil, G.P., Taillie, C. Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics,* 11, 183-197, 2004
17. Kulldorff M. Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society A*, 2001, 164 Part 1: 61-72.
18. US Department of Transportation (2004). 2001 National Household Travel Survey - Public use dataset. Available at http://nhts.ornl.gov/2001.
19. US Department of Labor (2005). American Time Use Survey - 2003 Data Files. Available at http://www.bls.gov/tus/datafiles_2003.htm.
20. US Department of Transportation (2005). Census Transportation Planning Package 2000 - Part 3, Journey-to-Work Flow Tables. Available at http://www.fhwa.dot.gov/ctpp/dataprod.htm.
21. Buckeridge DL, 2005. A Method for Evaluating Outbreak Detection in Public Health Surveillance Systems that use Administrative Data. Ph.D. Biomedical Informatics, Stanford University, Stanford, 199pp.

	Workflow scan statistic						Usual scan statistic					
$r_{\max}$ zone cells	1	2	3	4	5	average plateau	1	2	3	4	5	average plateau
28	.73	**.86**	**.84**	**.85**	**.86**	.86	.69	**.78**	**.80**	**.80**	**.80**	.80
29	.78	**.85**	**.81**	**.85**	**.85**	.85	.71	**.78**	**.81**	**.81**	**.81**	.81
28, 29	.92	**.98**	**.97**	**.98**	**.98**	.98	.84	**.95**	**.97**	**.97**	**.97**	.97
29, 31	.78	**.85**	**.82**	**.86**	**.86**	.86	.70	**.78**	**.81**	**.82**	**.81**	.81
28, 31	.74	**.86**	**.85**	**.86**	**.87**	.86	.69	**.79**	**.81**	**.82**	**.81**	.81
28, 29, 31	.93	**.99**	**.98**	**.98**	**.98**	.98	.84	.95	**.97**	**.97**	**.97**	.97
30	.64	.60	**.73**	**.72**	**.73**	.73	.54	.55	**.64**	**.65**	**.64**	.64
30, 26	.57	.59	**.80**	**.81**	**.83**	.82	.46	.53	**.76**	**.76**	**.76**	.76
30, 26, 32	.57	.73	**.87**	**.89**	**.90**	.90	.48	.66	**.86**	**.85**	**.86**	.86
13	.61	**.70**	**.71**	**.71**	**.70**	.71	.72	.74	**.78**	**.77**	**.78**	.78
25	.15	.20	.21	**.25**	**.24**	.24	.36	.36	**.43**	**.43**	**.43**	.43

Table 1. Power evaluation for some sample zones in the Norfolk neighborhood 35 counties map of Figure 2. The table shows the power for the workflow scan statistic, compared to the usual scan statistic, with significance level $\alpha = 0.05$. In both methods the alternative hypotheses were generated for the cells and the cases were dispersed using the workflow matrix according to the scheme described in section 4, and running 10,000 Monte Carlo simulations for each table entry. For each sample zone, the computed power reached a plateau when using $r_{\max}$ values above a certain threshold, indicated by bold types. For each zone their average plateau values were displayed in the respective columns, and used for the power comparison between both statistics. Except for the last two samples, which are outflow zones, there is a significant power gain using the workflow scan statistic.

Zone ZIP Codes cells	Workflow Scan	Usual Scan
A, B, C	.63	.45
A, B, C, F, G, J	.79	.63
A, B, C, G, J	.79	.62
A, C, D, E, G, H, I	.91	.78
A, D, E, G, K, L	.76	.60
D, G	.64	.48
D, G, J	.64	.48
G, I	.65	.52
D, E, G, K, L	.70	.54
A, D, E, K, L	.53	.39
D, E, K	.51	.36
B, D	.57	.40
B, C, D, J	.73	.56
B, C, D, E, J	.77	.61
B, C, D, E	.77	.60
B, C, D, E, L	.71	.54

Table 2. Power evaluation for the workflow and the usual scan statistics using sample zones in the 158 ZIP codes map of Norfolk neighborhood area in Figure 3. Significance level $\alpha = 0.05$ and $r_{\max} = 8$ were used in each of the 10,000 Monte Carlo simulations. The use of smaller subdivisions cells allowed for a sharper delineation of the workflow. Consequently, the workflow scan statistic power gain as compared to the usual scan statistic was enhanced.

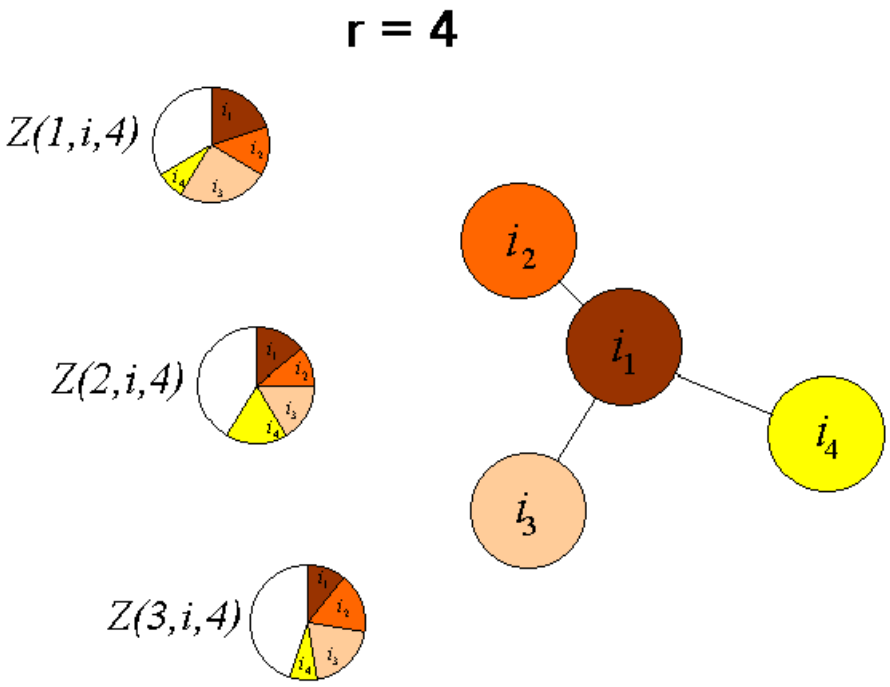

Figure 1. This example shows the construction of the zone $Y(3,i_1,4)$, $i = i_1$. The closest neighbors Z_{i_2}, Z_{i_3} and Z_{i_4} of ZIP code Z_{i_1} are displayed to the right. The shaded slices of the ZIP codes $Z(1,i,4)$, $Z(2,i,4)$ and $Z(3,i,4)$ to the left represent the respective proportions $A(1,i,4)$, $A(2,i,4)$ and $A(3,i,4)$ of the population of the components of the zone $Y(3,i_1,4)$ that work at the 4 nearest neighbors of Z_{i_1}, with the four different shadings corresponding to each cell.

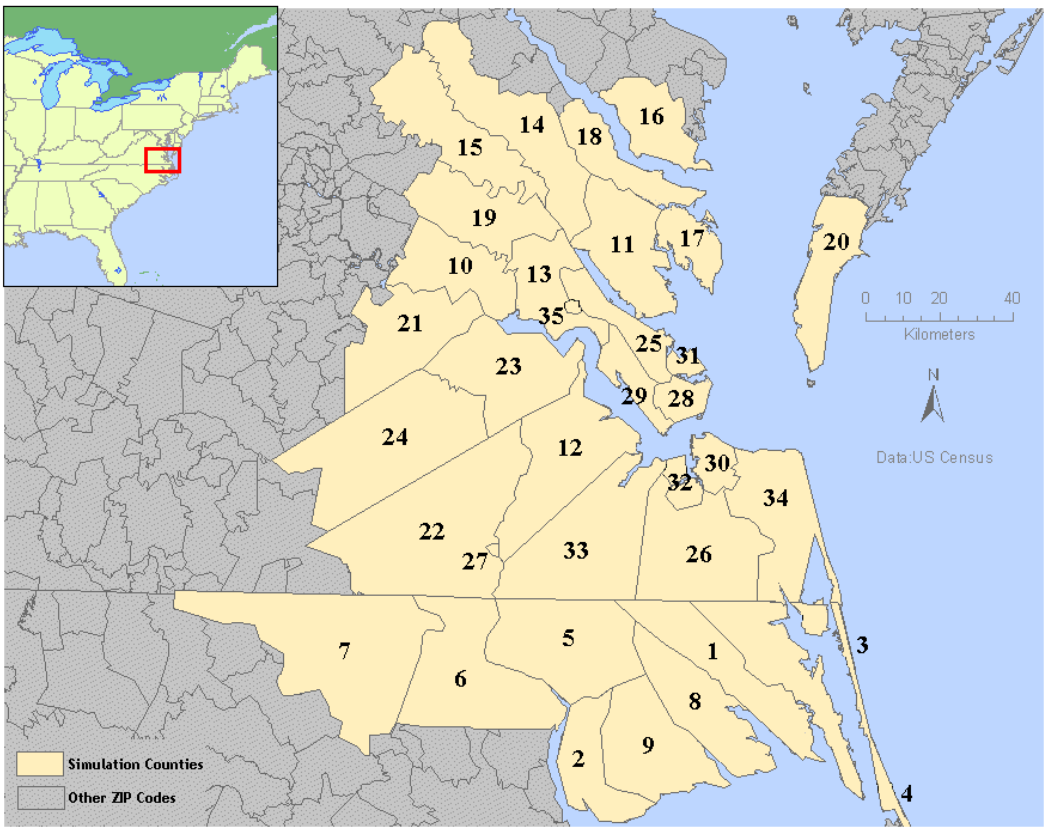

Figure 2. Study region with 35 counties in the neighborhood of Norfolk (cell 32), Newport News (cell 29) and Hampton (cell 28), VA.

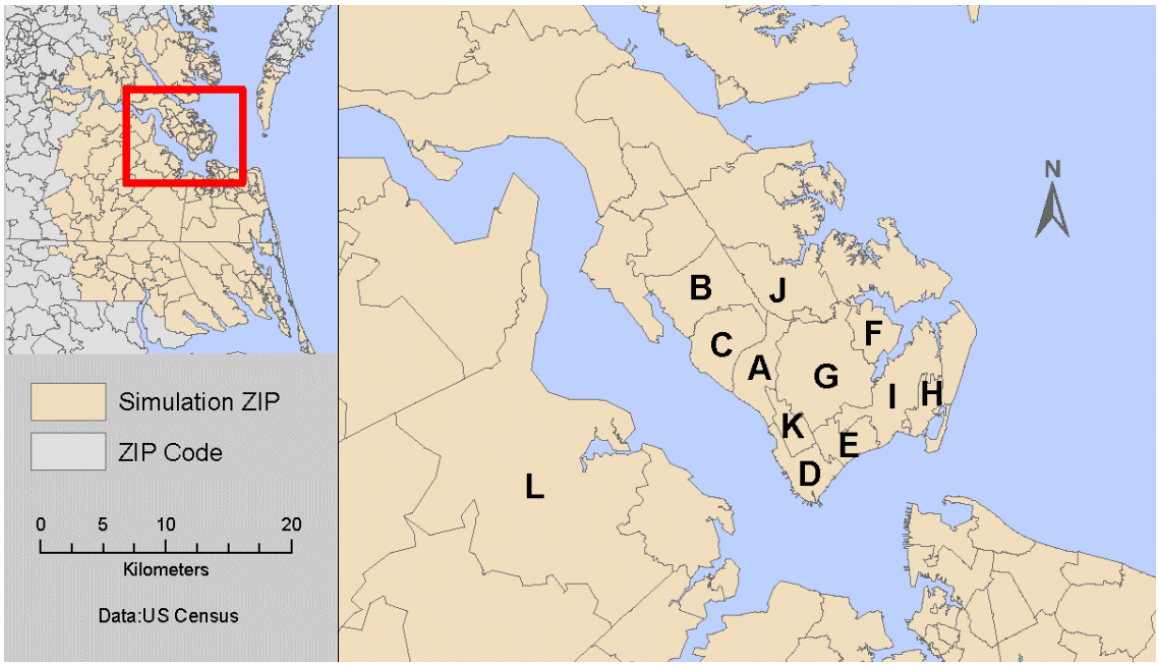

Figure 3. Map of the same study region depicted in figure 2, now divided into 158 ZIP codes. The following ZIP codes, in the neighborhood of Newport News and Hampton cities (inset), with the letters in parenthesis indicating their locations in the map, were used as examples in the simulations: 23601(A), 23602(B), 23606(C), 23607(D), 23661(E), 23665(F), 23666(G), 23668(H), 23669(I), 23693(J), 23605(K) and 23430(L).

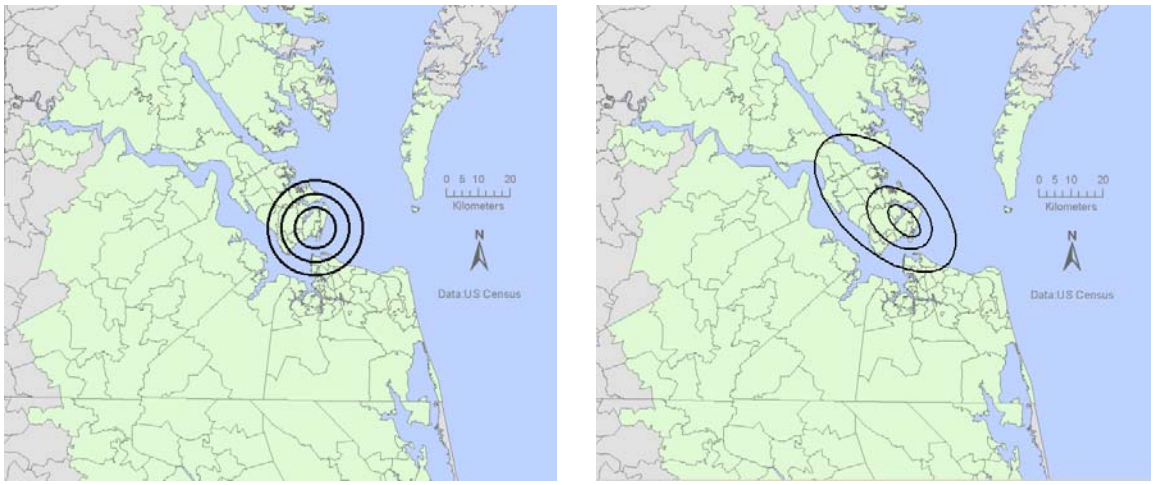

Figure 4. Simulated anthrax releases centered at ZIP 23668, without wind (left) and with wind (right). The solid lines indicate 50%, 25% and 10% intensity, respectively, compared with the center of the point release.