Geospatial Health 8(1), 2013, pp. 23-35

Impact of spatial aggregation error on the spatial scan analysis: a case study of colorectal cancer

Lan Luo

University of Illinois at Urbana-Champaign, Chicago, USA

Abstract. The paper aims to estimate the level and impact of spatial aggregation error for spatial scan statistics where disaggregated data below the zip code level are not available. Data on colorectal cancer cases in Cook county, Illinois, USA with a 5-year interval were used. An innovative procedure using SAS and Java was designed to make SaTScan auto-run. Characteristics of clusters at each reference level were compared to those at zip code level to observe differences related to spatial aggregation. The comparison reveals that spatial scan statistic at the zip code level can generate reliable clusters in areas with a large number of cases, but fail to detect clusters in areas where there are a sparse number of cases, since the spatial aggregation error is minimised in areas with sizeable numbers of cases. Without localised cancer data, zip code level data can be used effectively to identify dominant clusters. However, smaller clusters located in low-density areas may be missed.

Keywords: zip code, reference levels, colorectal cancer, spatial scan analysis, spatial aggregation error.

Introduction

The choice of geographical unit plays a very important role in analysing the uneven distribution of cancer cases and designing appropriate policies for disease control and prevention (Rushton, 1995). To protect privacy and confidentiality, cancer data obtained from surveillance systems are usually released only for predefined areal units with relatively large populations, e.g. counties or zone improvement plan (zip) codes. Because these predefined areas were not originally designed for cancer research, true patterns of cancer incidence can be distorted or obscured and thus produce misleading results. This problem has been well described as the "modifiable areal unit problem" (MAUP) (Amrhein, 1994; Openshaw and Alvandies, 1999). One of the important components of the MAUP is the spatial aggregation error, which is caused by using data at an aggregated, large-area level to generate inferences about patterns and processes at lower (small-area) levels (Hodgson et al., 1997). Although the biases brought about by the spatial aggregation error have been widely analysed (Hillsman and Rhoda, 1978; Hodgson et al., 1997; Fortney et al., 2000; Hewko et al., 2002; Luo et al., 2010), its impact on the detection of spatial clusters of cancer cases has rarely been studied.

In analysing spatial clustering of cancer, many researchers have used the spatial scan statistic to detect cluster locations (Kulldorff et al., 1997; Jemal et al., 2002; Thomas and Carlin, 2003; Gregorio et al., 2004; Pollack et al., 2006).The spatial scan statistic is a "local" spatial clustering test that identifies the locations and characteristics of statistically significant clusters of cases within a study area (Kulldorff, 1997). Studies have utilised spatial scan statistics to examine spatial disparities in cancer incidence and mortality (Kulldorff et al., 1997; Gregorio et al., 2002, 2004; Jemal et al., 2002; Roche et al., 2002). Several studies have applied spatial scan statistics to identify areas with high or low incidence rates of breast cancer (Gregorio and Samociuk, 2003) and to detect areas with an elevated proportion of late-stage breast cancer cases (Roche et al., 2002). All of these studies rely on cancer data that are spatially aggregated to units, such as towns, zip code areas, counties and census tracts. In all cases, using data at the individual level, based on precise residential locations, would likely yield different results for the SaTScan clustering test.

Specifically, some studies have compared the results of cluster tests using health data at different geographical scales in the USA. (Sheehan et al., 2000; Krieger et al., 2002; Gregorio et al., 2005). Sheehan et al. (2000) utilised the spatial scan statistic to detect significant spatial clusters of late-stage breast cancer diagnoses

Corresponding author:
Lan Luo
University of Illinois at Urbana-Champaign
55 S. Vail Avenue, Unit 1206, Arlington Heights
IL 60005, Chicago, USA
Tel. +1 217 390 7526; Fax +1 312 540 5199
E-mail: lanluo2010@gmail.com

across Massachusetts, using towns, zip code areas and census tracts. They observed that differences exist among the three geographical levels in terms of cluster sizes and the number of cases included in each cluster. However, they found that fluctuations in cluster characteristics were caused by geocoding problems, and that the fluctuations had little association with the sizes and boundaries of study units. Krieger et al. (2002) examined all-cause and cause-specific mortality rates, and all-cause and site-specific cancer incidence rates within census block groups, census tracts and zip code regions, across Massachusetts and Rhode Island. They concluded that analyses by census block group and census tract performed comparably, but results at the zip code level were contradictory. Gregorio et al. (2005) applied the spatial scan statistic to compare geographical variation in late-stage prostate and breast cancers across Connecticut, using census block groups, census tracts and towns. They reported that the local clusters identified at each scale were similar in terms of locations, populations at risk and other estimated parameters (centroid coordinates, P-values, and the ratios of observed-to-expected). Only a few differences were found in analytical results across the areal units. Schmiedel et al. (2012) analysed the clustering patterns and statistical power of different cluster detection methods using of individual and aggregated data at various levels. For the spatial scan statistic, data aggregated to small geographical areas produced the highest statistical power, and individual-level data yielded very similar results. These studies implement useful strategies for comparing different cancer spatial clusters among areal units. Particularly, Gregorio et al. (2005) summarised all the clustering parameters from the spatial scan statistic results into a straightforward table for clear comparison. They used the census block-level clusters as reference points, and compared clusters at the census tract and town levels. One metric used was the average distance between the geographical coordinates of census block-level centroids and those at town and census tract levels. Cluster comparisons were also illustrated by a nested-structure format which displays cluster locations, cluster sizes and the shared sections (overlap) among clusters on the same map.

Each study unit has pros and cons, e.g. small areal units depict local variations more clearly than larger areal units, while larger units produce more reliable and stable estimates of disease incidence or risk across a large region. Lacking a "gold standard", it is very difficult to select the optimum areal unit, and the optimum may vary from one case to another. The aforementioned studies generally observe little difference in cluster results using data at different scales ranging from census blocks to towns, indicating that the spatial aggregation error has a minimal effect on cluster detection. The approach taken in these studies is to begin with data for small areas and aggregate the data into larger areas. In this approach, there is only one outcome at each level, and the effect of the spatial aggregation error is exactly known. In many situations, however, researchers do not have access to data for small reference units, so it is important to know how much error might exist as a result of the need to work with data that are highly spatially aggregated. For example: how reliable are clusters detected based on large-area data? How likely is it that those clusters would also be detected if small-area data were analysed? Past research shows that the spatial aggregation error can be highly context-dependent (Luo et al., 2010). Specifically, it has a larger impact on cluster detection when the distribution of disease cases and at-risk population vary across the study area. Given these challenges, an important question is: How sensitive are the results of the spatial scan statistic to the choice of areal units? Therefore, approaches are needed to estimate the level and impact of potential spatial aggregation error on cluster detection results. The method used in this paper involves enumerating possible distribution patterns of cases within zip code areas using a Monte Carlo simulation approach and then examining the effects of spatial aggregation error at different geographical levels (census tract, block group and block).

This study aims to evaluate the level and impact of error caused by the aggregation of cancer data into predefined areas for situations where disaggregated data are not available. The most widely-used spatial scan test, the Bernoulli-based spatial scan statistic, is used. The impact of the spatial aggregation error on the spatial scan statistic at the zip code level is examined. Following previous research (Luo et al., 2010), a Monte Carlo simulation procedure is used to disaggregate cancer cases from the zip code level to the census tract, block group and block levels based on population demographic characteristics with a certain number of simulations. Then a Bernoulli-based spatial scan statistic method is applied to cancer cases at the zip code level and all the disaggregated sets of cancer cases at three census levels. Results of the spatial scan statistic are compared at each level to evaluate the sensitivity of results to the geographical scale of cancer data.

Materials and methods

Data

To analyse the impact of spatial data aggregation on the results of the spatial scan statistic, data on colorectal cancer (CRC) cases in Cook county, Illinois, USA were used. The health outcome analysed is the binary variable, late-stage CRC at diagnosis. CRC is classified as "late-stage" if the tumour is large and/or the disease has spread beyond the initial site when first diagnosed. People diagnosed with late-stage CRC have a higher risk of mortality and morbidity than those whose cancer is diagnosed early. Clusters of late-stage CRC were detected via SaTScan based on data at four geographical scales, from zip code to census block and results are compared.

The data were obtained from Illinois State Cancer Registry (ISCR) and include all CRC cases diagnosed in Cook county residents between 1998 and 2002. Records in the data set represent individual cancer cases, with variables including age group, sex, race, diagnosis stage, year and zip code of residence. The CRC cases were divided into early-stage (stages 0 and 1) and late-stage (stages 2 to 7) groups. Based on previous work (Luo et al., 2010), examining the influence of spatial data aggregation involved allocating CRC cases from zip codes to smaller geographical units which have stronger demographic association to reflect the age-sex race characteristics of the cancer case. A Monte Carlo simulation method was applied in this study to accomplish this; thus, the probability of cancer case's assignment from his or her residential zip code to a smaller geographical unit is proportional to the age-sex-race composition of the smaller unit's population.

To prepare the demographic link for the disaggregation, the CRC cases were divided into 12 categories representing combinations of race by age by gender. Specifically, CRC cases were aggregated into black and non-black groups; the original 5-year age groups were classified into three main groups (<50 years-old, 50-70-years old and >70 years-old), and gender was categorised as male and female. Population-level data for census areal units were derived from the Summary File 1 (SF1) data from the US Census Bureau for 2000 (US Census Bureau, 2000b, c), and categorised into the same 12 age-sex-race groups.

For comparison with the zip code level, three smaller geographical levels were selected as reference units: census tracts, census block groups and census blocks. These areal units are hierarchically structured and defined by the United States Census Bureau. Census tracts are "designated to be relatively homogeneous units with respect to population characteristics, economic status, and living conditions", and average 4,000 inhabitants in each area (US Census Bureau, 2000a). Census tracts can be subdivided into block groups and blocks, with blocks being the smallest areal units, and block groups intermediate in size between blocks and tracts. On average, 39 blocks form a block group with some small variations across the country. These three census areal units make appropriate choices because of their well-established association with demographic information, their nested structure, and their relatively stable boundaries over time.

Cook county was chosen as the study region, mainly because the spatial relationships between the four geographical levels are well-defined. Cook county is the most populated area in Illinois, and the high population density means that the three census areal units are typically smaller than zip codes. Thus, the spatial relation between census tracts and zip codes can be easily defined as "within" or "outside". In general, there is a clear hierarchical spatial relationship between the zip code level and smaller census area units. In addition, Cook county contains a large sample size of CRC cases, with 3,608 total cases and 2,353 late-staged, making it possible to detect statistically significant spatial clusters.

Disaggregation of cancer cases

Because cancer data are unobtainable at a level below the zip code scale, a Monte Carlo simulation approach developed previously (Luo et al., 2010) was applied to disaggregate cancer data from the zip code level to each smaller geographical unit. In the Monte Carlo procedure, each cancer case is randomly assigned to a census tract (or block group or block) within the zip code in which the case is located. The probability of assignment is proportional to the age-sex-race composition of the tract population. For example, a black male in the 50-70 year age group is more likely to be assigned to a tract containing a large population of that demographic group. This disaggregation process is repeated a large number of times, resulting in a large number of possible geographical disaggregations of cancer cases. For each simulated disaggregation, the SaTScan method was used to identify spatial clusters of late-stage CRC cases. Because of the intensive computation time for re-running SaTScan, the number of Monte Carlo simulations was set at 100. Consequently, at each reference level, the

spatial scan algorithm was run 100 times, each time on a separate simulated CRC dataset.

The most critical step in the disaggregation process was to define which zip code contains each census tract, so that tracts were not shared by neighbouring zip codes. In the cancer dataset, CRC patients lived in 152 out of 161 zip code areas, covering most sections of Cook county. As the smallest reference unit, census blocks are mostly completely inside of each zip code area. If a block overlapped a zip code boundary, the block was treated as within a zip code if the block centroid fell within the zip code. As a result, 64,231 blocks were assigned to the 152 zip code areas. Linking census block groups with zip codes was more complicated, because the larger size of a block group increases the chances of it overlapping multiple zip codes. Several steps were implemented to specify the spatial relation between zip codes and block groups. First, the population-weighted centroid of each block group was generated based on block-level population information; then each block group was regarded to be within a zip code if its population-weighted centroid was located inside of that zip code. Seven block groups whose population-weighted centroids were outside the study area. Two other block groups were merged with their neighbours, because their small sizes were completely within a zip code and they shared that zip code with their neighbouring block groups. This resulted in a total of 4,260 block groups. Similar strategies were implemented to assign census tracts to zip codes. Only nine of the Cook county census tracts were excluded, leaving 1,365 tracts.

Automation of SaTScan

The spatial scan statistic was utilised to analyse spatial clustering patterns of high late-stage CRC cancer cases at the level of the zip code and the three reference levels. SaTScan was chosen over other spatial clustering methods like local indicators of spatial autocorrelation (LISA) and Getis-Ord G*, because it uses a varying scanning window and an appropriate maximum likelihood test to detect clusters accurately. The specific spatial scan statistic to address the binary characteristic of late-stage diagnosis that was applied in this study was the Bernoulli-based model. In the spatial scan test, a scanning window is passed over the study area, and the number of cases computed within and outside the window. A likelihood ratio test is utilised to compare the null hypothesis of constant risk within and outside the window with the alternative hypothesis of non-equal risk. The outcomes of the maximum likelihood ratio test provide an indication of the most likely clusters. The formulation of the Bernoulli-based spatial scan statistic is provided below:

$$\lambda = \max_z \left(\frac{c_z}{n_z}\right)^{c_z} \left(1-\frac{c_z}{n_z}\right)^{n_z-c_z} \left(\frac{C-c_z}{N-n_z}\right)^{C-c_z} \left(1-\frac{C-c_z}{N-n_z}\right)^{(N-n_z)-(C-c_z)} \times I\left(\frac{c_z}{n_z} > \frac{C-c_z}{N-n_z}\right) \quad (1)$$

where c_z is defined as the total number of late-stage CRC cases and n_z the total number of CRC cases within a circular area (Z). C is the total number of late-stage CRC cases and N the total number of CRC cases in the whole study area. I denotes the indicator function (this formula only maximises the likelihood function for windows where the observed probability inside the window is larger than the one outside the window).

To implement this procedure, SaTScan uses a coordinate file to assign the location of each case (a late-stage CRC case) and each control (an early-stage CRC case). Then it generates a very large number of circular windows, whose centroids are the coordinates of cases. The radii of these circular windows vary from the smallest observed distance between a pair of cases to a user-defined threshold (Waller and Gotway, 2004). *I* set the threshold as the radius containing up to 33% of the entire population of the study area. In each circle, the likelihood ratio statistic is applied to test the null hypothesis of constant risk *versus* the alternative hypothesis that the late-stage rate within the scanning window is greater than that outside the window. The statistical significance of clusters was tested by Monte Carlo simulation with 999 replications. The Bernoulli-based spatial scan statistic normally generates a number of spatial clusters with different P-values. This study focuses on the clusters which had the smallest P-values ($P \leq 0.10$).

In comparing geographical clusters of cancer using data for different areal units, SaTScan needs to be run many times, for each randomly generated disaggregation of cancer cases at the tract, block group or block level. Each run of SaTScan involves creating unique input and destination files, a very time-consuming task using SaTScan's graphical user interface. Therefore, a need existed to automate the whole SaTScan procedure, and Abrams and Kleinman, (2007) designed the SaTScan Macro Accessory for Cartography (SMAC) package, comprising four SAS macros, to fully automate SaTScan. Nevertheless, SMAC is only available

for the Poisson-based spatial scan statistic. Additionally, the macro-syntaxes in SAS are lengthy and complicated, so that it is quite challenging for users to customize or apply the SMAC package, especially for those who are not familiar with macro-level programming in SAS. To overcome this challenge, I created scripts using SAS macro programming and Java in order to auto-run SaTScan for the analyses conducted at each geographical level. The scripts automatically generate the case, control, parameter and destination files that are required for each run of SaTScan. Details of the procedures and scripts are available from the author.

To compare SaTScan outcomes at different geographical scales, the locations, sizes and other characteristics of statistically significant spatial clusters were compared. SaTScan outcomes included the primary cluster at the zip code level, and the primary clusters from each of the 100 simulated cancer patterns at the census tract, block group and block levels. However, many of the primary clusters did not achieve statistical significance (P-value <0.1). Only the statistically significant clusters at each level were compared with the primary cluster at the zip code level. The primary clusters with statistical significance at each reference level were displayed on a map with the zip code level cluster to show the geographical similarity or difference between the results at the two levels. Additionally, the parameters of the statistically significant clusters at each reference level were compared with those at the zip code level. The geographical and statistical comparisons between zip code level and reference levels reveal the impact of spatial aggregation error on the Bernoulli-based spatial scan statistic results.

Results

Overall, more than half of CRC cases in Cook county in 1998-2002 were diagnosed at a late-stage. The late-stage percentage varied among age, gender and race groups. Generally, the ratio of late-stage to early-stage cases fell between 1.5 and 2 in each demographic category (Table 1). The most dramatic excess of late-stage CRC cases compared to early-stage was in the youngest group; however sample sizes were so small that it is difficult to generalise. The largest numbers of early-and late-stage CRC cases were observed in the elder age group for every race/gender group. Beyond these age-related differences, no gender or racial disparities in late-stage diagnosis were apparent.

Running SaTScan at the zip code level identified one primary cluster. This cluster occurred in the northwestern section of Cook county, covering the northwestern edge of Chicago city. The radius of the cluster was approximately 6 km, and it covered almost 113 km^2. The number of late-stage cases in this cluster is 288, and the relative-risk is 1.14, indicating that CRC patients living in the cluster are approximately 14% more likely to be diagnosed with late-stage CRC than those residing outside the zone. In terms of P-value (0.119), the cluster at the zip code level is not statistically significant according to standard significance levels. However, zip codes may be oversised areal units for studying the local patterns of late-stage CRC, and one might suspect that the zip code analysis will miss some significant clusters that would be detected based on small-area data. The zip code cluster is used as a benchmark for comparison: the clusters with statistically significant P-values at each reference level. This comparison suggests the validity of the zip code level cluster, and the types of clusters it might miss. These comparisons are discussed in the following sections.

At the census tract level, 14 of 100 simulations resulted in statistically significant spatial clusters containing significantly high ratios of late-to-early stage CRC cases. Table 2 displays characteristics of these clusters, including centroid coordinates, the radius and covering area of circular windows, numbers of observed late-stage cases within each cluster, the ratio

Table 1. Demographic and epidemiological summary of colorectal cancer cases in Cook county in the period 1998-2002.

Gender	Age (years)	Black			Non-black		
		Early stage (n)	Late stage (n)	Ratio*	Early stage (n)	Late stage (n)	Ratio*
Female	<50	6	18	3.00	27	82	3.04
Female	50-70	23	49	2.13	163	288	1.77
Female	>70	45	87	1.93	349	682	1.95
Male	<50	4	13	3.25	36	82	2.28
Male	50-70	39	62	1.59	212	365	1.72
Male	>70	20	50	2.50	317	563	1.78

*Late to early-stage

Table 2. Results of Bernoulli-based spatial scan statistic at zip code and census tract levels.

Cluster	Centroid coordinates	Radius (km)	Area (km^2)	Late-stage cases (n)	O/E	P-value	Relative risk	Distance (km)	Overlap area (%)*
1	39.773; -85.044	1.33	5.52	23	1.53	0.042	1.54	10.83	0.00
2	39.734; -84.990	1.22	13.40	26	1.53	0.017	1.54	4.62	4.89
3	39.809; -85.252	2.07	5.52	23	1.53	0.055	1.54	27.68	0.00
4	39.708; -85.252	0.98	3.04	21	1.53	0.090	1.54	5.27	2.47
5	39.697; -85.023	3.72	43.45	126	1.23	0.095	1.24	3.45	33.12
6	39.703; -84.955	5.27	87.38	204	1.18	0.087	1.20	0.92	76.52
7	39.798; -85.327	1.83	10.57	23	1.53	0.059	1.54	33.14	0.00
8	39.781; -85.052	1.01	3.28	35	1.49	0.012	1.50	11.98	0.00
9	39.731; -85.028	3.84	46.37	144	1.23	0.029	1.25	6.71	13.26
10	39.586; -84.808	2.01	12.68	21	1.53	0.090	1.54	18.27	0.00
11	39.712; -85.008	1.47	6.81	57	1.36	0.066	1.37	4.19	6.03
12	39.699; -84.965	6.25	122.64	278	1.15	0.090	1.17	0.22	100.00
13	39.703; -84.955	7.25	165.08	336	1.14	0.065	1.16	0.92	100.00
14	39.704; -85.002	0.99	3.10	22	1.53	0.070	1.54	3.46	2.74
Zip**	39.698; -84.963	5.99	112.88	288	1.12	0.120	1.14		

*% of zip code level cluster; **zip code level.

of observed-to-expected cases, P-values, relative risks, the distance from the zip code level centroid and the percent of zip code cluster area that overlaps with the tract cluster. This table also lists the parameters at the zip code level in the last row for comparison. Compared to the zip code cluster, the significant census tract clusters all had higher relative-risks and ratios of observed-to-expected cases. This localised clustering indicates that in a highly populated region, spatial clusters of CRC cases are more likely to be detected using data for smaller areal units than at the zip code scale. Nine census tract level clusters overlapped the zip code cluster, and the overlap percentages varied from 2.7% to 100.0%.

The census tract clusters are shown in Figs. 1 and 2 with the centroids of each cluster mapped in Fig. 2. In Fig. 1, two clusters (12 and 13 in Table 2) at the census tract level have very similar covering areas as the one at the zip code level, and the 12th cluster can almost be treated as a replica of the zip code level one except for a small curved area outside of the zip code cluster zone. The 13th cluster includes more area than the zip code one, including a crescent-shaped buffer surrounding the zip code cluster. The 6th cluster also highly overlaps the zip code cluster, covering 76.5% of its area. At the south-eastern edge of the zip code cluster, four census tract clusters are completely within the zip code cluster, and another cluster mainly falls into the zip code cluster except for a small tip outside. On the other hand, five clusters at the census tract level are completely outside the zip code level cluster: one close to the northern border of Cook county, two southeast of the zip code level cluster, and the other two locate at the southern border of the city of Chicago. However, these clusters are small and the numbers of observed cases within these clusters no larger than 35. Fig. 2 also shows the location of each tract cluster centroid in relation to the one at the zip code level. The centroid of the 12th cluster seen there is almost identical to the zip code one. The centroids of eight other clusters also closely surround the zip code centroid with distances ranging from 0.92 km to 6.71 km (Table 2). However, four clusters have centroids located more than 10 km from the zip code cluster centroid. In summary, the tract-level clusters corresponded quite well geographically to the zip code cluster in general, although the zip code cluster failed to represent some smaller, distant clusters that were detected with tract level data.

The block group level spatial scan statistic generated 18 clusters with significantly high late-to-early ratios, more than were found at either of the other two levels (Table 3). Similar to the clusters at the census tract level, these block group clusters presented higher ratios of observed-to-expected cases and larger relative risks than the one at the zip code level. Block group clusters tended to be smaller than those at the tract and zip code levels. Only one block group level cluster overlapped greatly (85.2% overlap area) with the zip code cluster. Nine other clusters overlapped with small

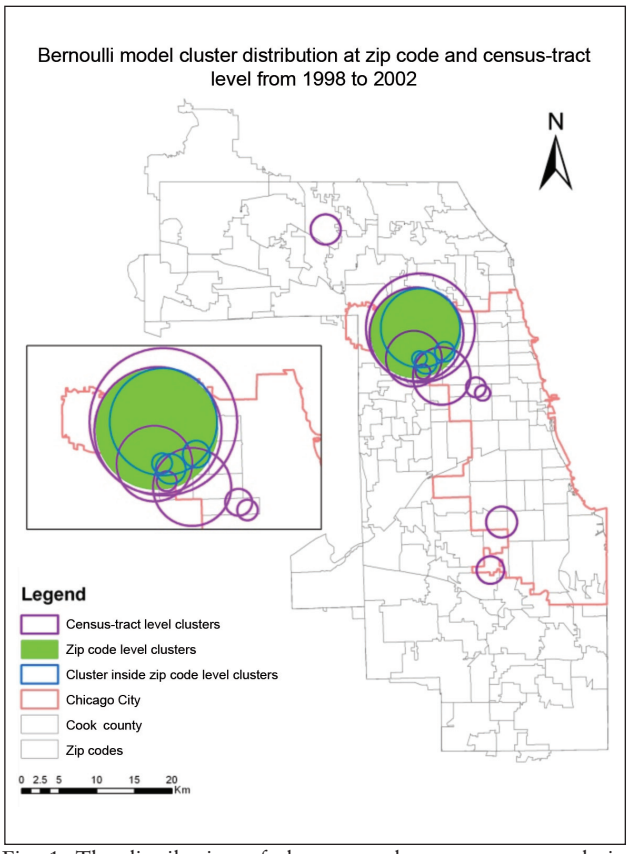

Fig. 1. The distribution of clusters at the census-tract and zip code levels in Cook county.

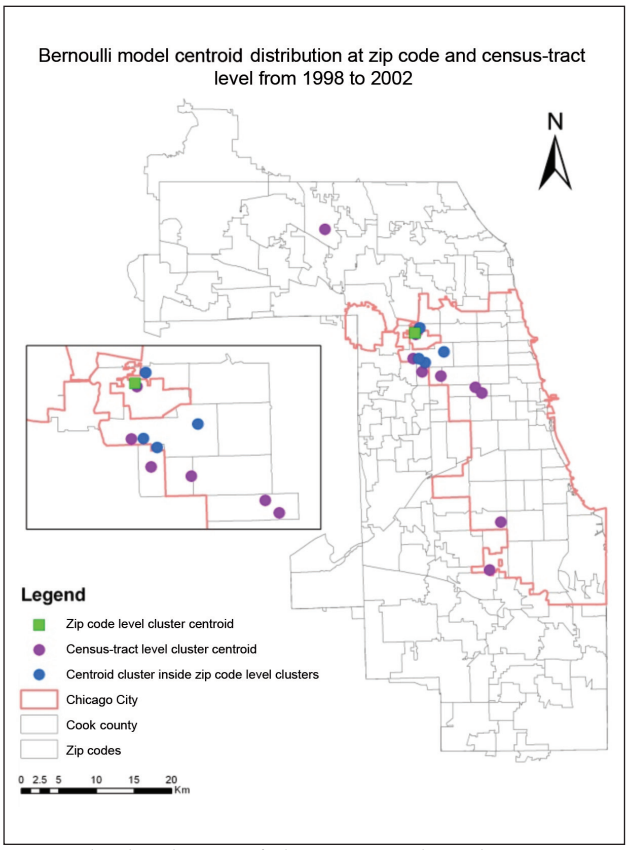

Fig. 2. The distribution of cluster centroids at the census-tract and zip code levels in Cook county.

Table 3. Results of Bernoulli-based spatial scan statistic at zip code and census block group levels.

Cluster	Centroid coordinates	Radius (km)	Area (km^2)	Late-stage cases (n)	O/E	P-value	Relative risk	Distance (km)	Overlap area (%)*
1	39.675; -84.989	2.28	16.39	25	1.53	0.044	1.54	3.39	14.52
2	39.717; -85.017	1.15	4.16	25	1.53	0.048	1.54	5.18	3.32
3	39.712; -85.008	1.50	7.05	40	1.49	0.0034	1.50	4.19	6.24
4	39.742; -85.049	2.49	19.48	47	1.41	0.069	1.42	8.83	0.00
5	39.811;-85.204	1.37	5.88	24	1.53	0.063	1.54	23.76	0.00
6	39.712; -85.017	1.62	8.19	46	1.41	0.068	1.42	4.86	6.49
7	39.713; -85.021	0.95	2.83	25	1.53	0.053	1.54	5.29	2.31
8	39.695; -84.973	5.73	103.25	244	1.18	0.041	1.20	0.98	85.24
9	39.611; -84.799	3.10	30.21	26	1.53	0.039	1.54	17.09	0.00
10	39.723; -85.025	2.03	13.00	59	1.37	0.051	1.38	6.06	5.12
11	39.729; -85.040	3.35	7.94	107	1.29	0.019	1.30	7.45	6.09
12	39.774; -85.042	1.59	3.68	27	1.53	0.024	1.54	10.80	0.00
13	39.782; -85.045	1.08	15.33	31	1.48	0.071	1.49	11.68	0.00
14	39.878; -85.291	2.21	18.06	25	1.53	0.047	1.54	34.51	0.00
15	39.723; -85.040	2.40	68.90	54	1.38	0.089	1.39	7.20	2.73
16	39.721; -85.013	1.45	6.56	32	1.48	0.054	1.49	4.98	5.21
17	39.837; -85.358	4.68	14.07	92	1.29	0.083	1.30	37.27	0.00
18	39.775; -85.042	2.12	35.23	48	1.41	0.049	1.42	10.97	0.00
Zip**	39.698; -84.963	5.99	112.89	288	1.12	0.120	1.14		

*% of zip code level cluster; **zip code level.

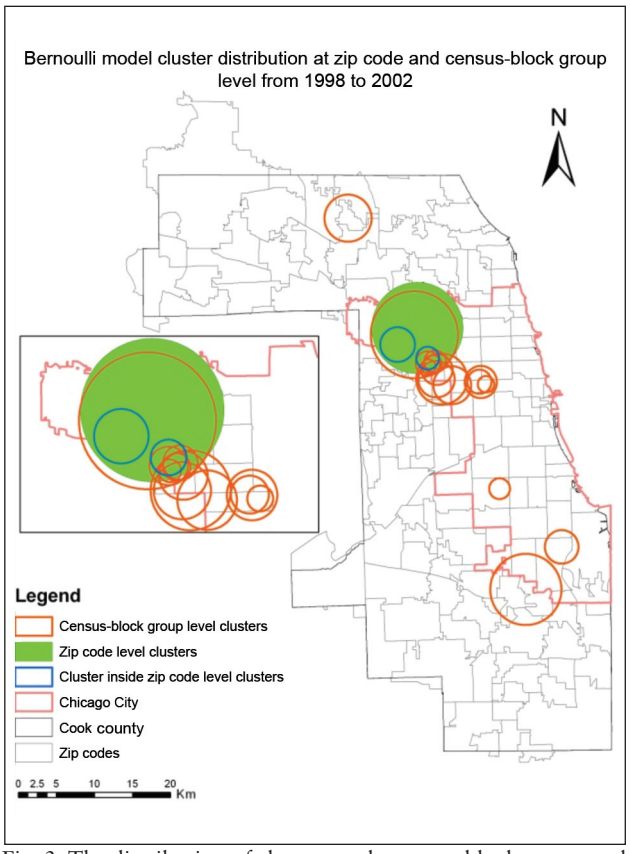

Fig. 3. The distribution of clusters at the census block-group and zip code levels in Cook county.

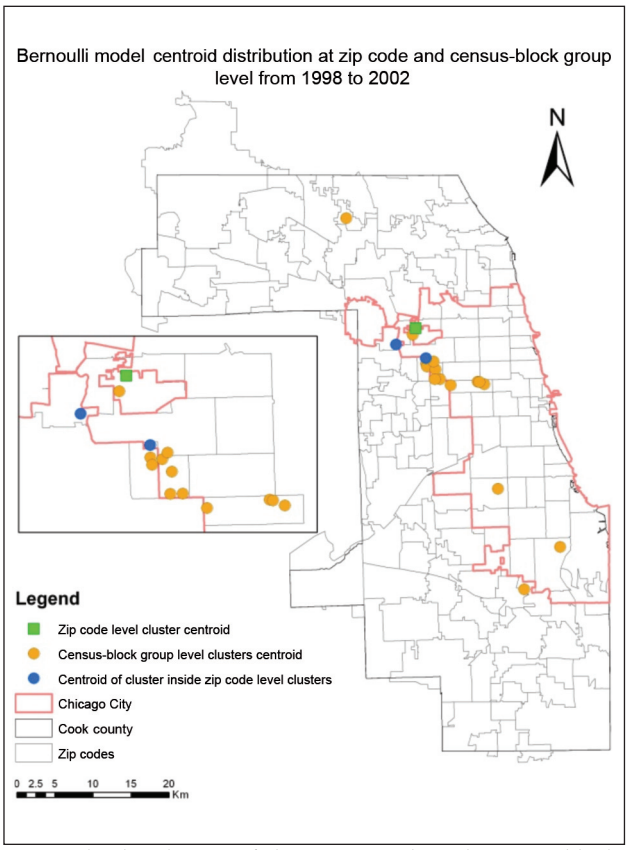

Fig. 4. The distribution of cluster centroids at the census block-group and zip code levels in Cook county.

sections (ranging from 2.3% to 14.5% of the zip code cluster area) of the zip code cluster. The number of block group level clusters that are completely outside the zip code clustering zone is eight, compared with only five at the census tract level. These "outside" clusters appeared to have larger radii and covered a larger area than the "outside" ones at the census tract level. Furthermore, the P-values at the block group level were generally smaller than those at census tract level, indicating that this smaller area level is capable of detecting more distinctive patterns of late-stage CRC clustering.

Fig. 3 describes the spatial distribution of block group level clusters. Similar to the census tract results, the block group level clusters that overlap with the zip code cluster are often located along the southern part of the zip code cluster, indicating a tendency for clusters to be focused in this area. Among the 10 clusters that overlap with the zip code cluster, only two (highlighted by blue boundary) lie completely inside, occupying 14.5% and 6.2% of the zip code cluster area, respectively. Among "outside" clusters, several appear- southeast of the zip code cluster, in locations similar to those detected with census tract data. The other "outside" clusters are also located in areas similar to clusters at the census tract level. One appear in the northern part of Cook county and two others around the southern border of the city of Chicago. Furthermore, their radii and covering areas are generally larger than those for the corresponding clusters at the census tract level. These "outside" clusters revealed that the use of data at the block group level enhanced the possibility of detecting late-stage CRC clusters outside the dominant clustering area compared to using data at the tract or zip code levels. Examining block group cluster centroid locations in Fig. 4, shows a concentration of centroids near the zip code centroid and along the south-eastern edge of the zip code cluster – a pattern similar to that observed based on tract level data. Fourteen of the 18 block group cluster centroids lie within the 11 km buffering zone of the zip code level centroid, indicating a relatively good geographical correspondence between clusters at both levels. However, four cluster centroids fall far outside, with centroid distances ranging from 17 to 37 km (Table 3). The maximum value of the block group level distances to the zip code centroid is 37 km, compared to 33 km at the census tract level.

Based on block level data, 15 clusters had significantly high late-to-early ratios. These clusters tended to be smaller in size than those at the tract or block group level (Table 4): their radii and covering areas were generally smaller than the ones at census tract and block group levels. As the smallest reference unit, blocks provided the most localised detail about the spatial clustering patterns of late-stage CRC cases. The majority of block level clusters presented high ratios of observed-to-expected late-stage cases and relative risks, indicating that more localised variation in late-stage CRC cases can be detected using data for the smallest reference unit. Numbers of observed cases in each block level cluster were generally less than those in clusters at other levels, so the block data uncover small, localised clusters of late-stage CRC. Because of the small sizes of block-level clusters, the percentages of zip code cluster area that overlapped with the block level clusters were much less than the ones for clusters at the other two scales.

Fig. 5 displays the clusters with statistically significant P-values at the block level. These clusters clearly reveal concentrations of high late-to-early ratios around the eastern and south-eastern sections of the zip code level cluster. Three clusters (highlighted by blue boundary) completely fall inside the zip code level cluster, respectively covering 15.7%, 7.4% and 5.2% of the zip code level cluster area. Five clusters at the block level are located completely outside the zip code level cluster: four southeast of the zip code level cluster and another in the southern part of Cook county. Clusters in the northern part of Cook county and around the south-western border of the city of Chicago that emerge in the tract and block group analyses do not appear in the block level analysis. The reason may be that the simulated cancer cases at the block level are more evenly distributed than those at the tract and block group levels, resulting in a less tendency towards clustering. Of course, the Monte Carlo simulation involves a random assignment procedure, in which the spatial disaggregation of cases within zip codes is only based on demographic information and otherwise spatially random. In areas with few CRC cases, disaggregation of cases to the block level may result in more dispersed geographical patterns.

In terms of distances between centroids, only the cluster located in the southern part of the city of Chicago present a relative long distance (34 km) (Fig. 6). Centroids of the three "inside" clusters are located near the zip code centroid with centroid distances of 4.1 km or less. The other clusters have distances varying from 3.3 km to 12.4 km. Compared to the clusters at census tract and block group levels, the block level clusters revealed more detailed spatial aggregations of late-stage CRC cases in areas containing large numbers of late-stage CRC cases. However, in regions with fewer late-stage CRC cases, such as the northern part of Cook county and south-eastern section of the city, the block level failed to identify clusters with significantly high ratios.

Table 4. Results of Bernoulli-based spatial scan statistic at zip code and census block levels.

Cluster	Centroid coordinates	Radius (km)	Area (km^2)	Late-stage cases (n)	O/E	P-value	Relative risk	Distance (km)	Overlap area (%)*
1	39.722; -85.021	1.65	8.51	42	1.43	0.090	1.44	5.64	4.57
2	39.785; -85.053	1.97	12.19	45	1.43	0.039	1.44	12.38	0.00
3	39.786; -85.040	1.21	4.56	33	1.53	0.003	1.54	11.82	0.00
4	39.733; -84.972	2.38	17.76	71	1.34	0.052	1.35	3.97	15.08
5	39.774; -85.046	1.67	8.77	33	1.49	0.071	1.49	11.08	0.00
6	39.724; -85.018	1.27	5.08	27	1.53	0.042	1.54	5.55	3.15
7	39.872; -85.290	1.85	10.80	24	1.53	0.085	1.54	34.11	0.00
8	39.673; -84.981	2.37	17.65	24	1.53	0.087	1.54	3.16	15.66
9	39.693; -84.954	1.63	8.30	25	1.53	0.061	1.54	0.97	7.35
10	39.721; -85.018	2.44	18.63	81	1.32	0.067	1.33	5.40	10.07
11	39.772; -85.050	1.45	6.61	27	1.53	0.040	1.54	11.16	0.00
12	39.672; -84.983	2.79	24.42	25	1.53	0.055	1.54	3.39	21.35
13	39.662; -84.978	3.28	33.88	28	1.53	0.022	1.54	4.16	23.76
14	39.668; -84.966	2.98	27.88	26	1.53	0.038	1.54	3.32	24.04
15	39.716; -85.005	1.37	5.90	31	1.53	0.007	1.54	4.10	5.22
Zip**	39.698; -84.963	5.99	112.88	288	1.12	0.12	1.14		

*% of zip code level cluster; **zip code level.

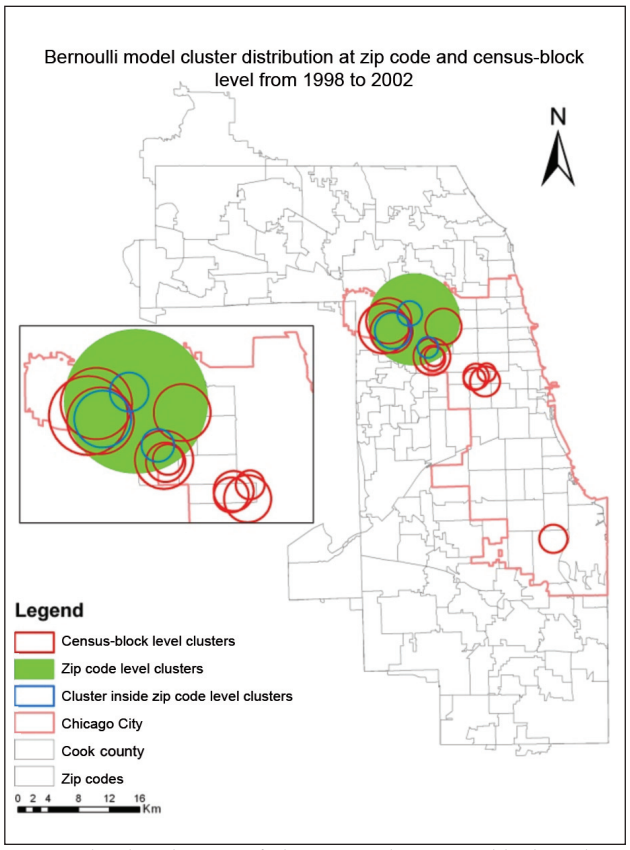

Fig. 5. The distribution of clusters at the census block and zip code levels in Cook county.

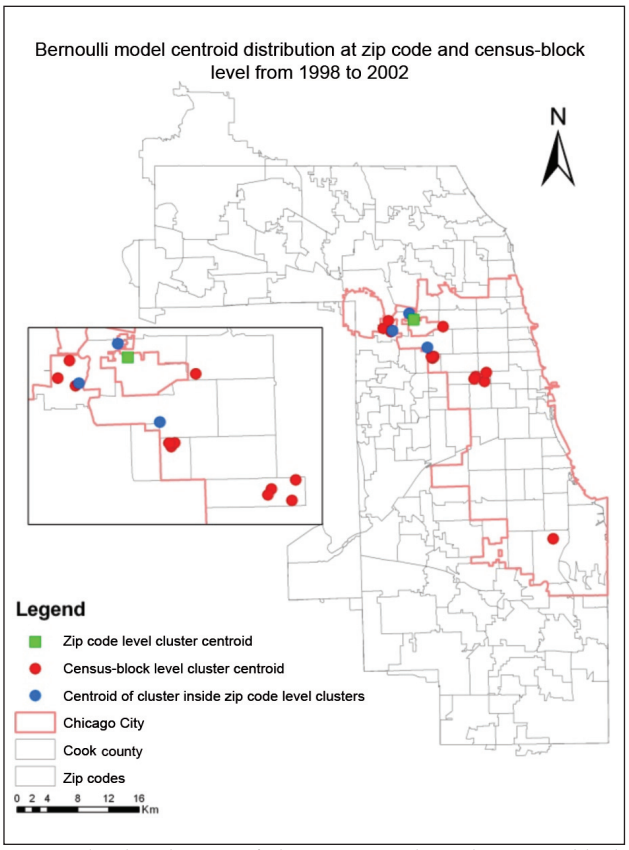

Fig. 6. The distribution of cluster centroids at the census block and zip code levels in Cook county.

Discussion

This study compared the results of the Bernoulli-based spatial scan statistic at the zip code level with the outcomes at three reference census units (census tract, census block group, and census block) to examine if reliable and accurate spatial analysis results can be generated using zip code level data. Lacking actual data on patient locations by census tract, block group and block, a Monte Carlo simulation procedure was used to disaggregate cancer cases from the zip code level to smaller census geographical units. Thus, the research focused on possible geographical patterns of CRC cases that conform to the demographic and geographical characteristics of cases at the zip code level. The number of simulated results was 100 at each reference level, and every result was tested for spatial clustering using the Bernoulli-based spatial scan statistic in SaTScan. Because the steps of importing input files and providing non-duplicated output names in each SaTScan run were tedious and time-consuming to perform manually, I designed a cost-effective procedure to automate the running of SaTScan. This procedure mainly consisted of a macro-level SAS programme to automatically generate input files and a Java programme to automate the parameter file generation. Compared with the SMAC package created by Abrams et al. (2007), my procedure is simpler, more efficient and highly adaptive to other spatial scan statistics in SaTScan, because it only comprises two small programmes and there is no need to build the major part of a parameter file.

Comparing geographical clusters with statistically significant P-values at each reference level with the zip code cluster yielded several innovative results. One important observation was that only a small number (14-18) of the simulated data patterns at each reference unit produced statistically significant clusters. Thus, the fact that the zip code level cluster had a P-value of 0.12 seems appropriate, given that 80-85% of the clusters generated based on simulated data at each reference level were not statistically significant. The spatial scan statistic at the zip code level did well at identifying a primary cluster in an area with a high density of cases. However, the spatial scan analysis at this level lost the power to detect more localised clusters. In some instances, the simulated datasets contained statistically significant clusters located in areas with smaller numbers of late-stage cases. Even in the areas with a large sample size

of cases, using zip code level data fails to detect statistically significant clusters that appeared at the census tract, block group and block levels. At these levels, clusters often were detected along an axis extending southeast of the zip code level cluster. Some of these clusters partially overlapped with the zip code cluster, while others did not.

Comparing the zip code level cluster with the clusters at the three reference levels indicates strengths and weaknesses of using the zip code level as the study unit. Specifically, the Bernoulli-based spatial scan statistic at the zip code level can detect clusters in areas with large concentrations of cases. However, even in these concentrated settings, the zip code cluster is at the global level, which means it gives general clustering information without much local detail. In other areas with fewer CRC cases, the zip code level is too large to detect "local level" clusters. Thus, spatial aggregation error may have more influence in areas where the sample size is small, compared to areas with many cases. Specifically, in areas containing fewer CRC cases, the use of zip code level data misses statistically significant clusters that are detected based on small-area data. Clusters located near the northern border of Cook county and southern border of Chicago could not be detected at the zip code level. At the block level, simulated data contained significant clusters located in the eastern, western, southeastern parts of the zip code level cluster and some surrounding areas. Although some of these clusters overlapped the zip code cluster, others were more geographically distinct. Thus, the spatial scan statistic at the zip code scale can produce reliable and stable "global-level" results; however it has difficulty in identifying clusters at a smaller and more localised level. If cancer data for small areas is not available, applying the spatial scan statistic at the zip code level can detect the dominant cluster(s) in areas where the sample size is large.

Additionally, depending on the densities of cases within different local areas, the influence of the degree of the spatial aggregation error on the spatial scan analysis may vary. The influence is typically greater in areas with a low density of cases, where the combination of low statistical power and spatial aggregation of cases makes it difficult to detect localised clusters. Although utilizing zip code level data made it possible to detect a stable and large cluster in Cook county, scan analysis at this level was less appropriate for detecting clusters in areas with a lower density of CRC cases. Thus, to detect spatial clusters using the spatial scan statistic, a trade-off strategy needs to be applied in selecting the study unit. In areas with large numbers of cases, such as metropolitan areas, work based on the smallest unit, such as census block, can reveal localised clusters in great detail. However, in areas with fewer cases, represented as suburban areas in this study, using a 'middle-size' study level which can contain enough sample size of cases without oversised concern, such as census tract or census block group, the clustering patterns can be identified better than using the smallest areal unit. In areas with sparsely populated cases, such as small towns or rural regions, using a large-unit level with strong attachment of demographic characteristics (community, town, county) rather than zip codes (that are infamously detached from demographic attributes) can collect sample sizes large enough to identify the significant patterns not possible with small- or middle-sized units. Thus, the optimum study size of the spatial scan statistic needs to be varied based on the distribution of cases in different regions across the whole study area.

To meet the growing needs of detecting health data with localised patterns and provide more accurate spatial analysis outcomes, the release level of cancer data needs to evolve to various levels rather than a traditional unified one. Explicitly, given that the results of the spatial scan statistic highly depend on the locations and density of cases in a specific area, using a uniform policy to release health data for research in different study areas is not very appropriate. In areas with a high density of cases, data can be released at a smaller areal unit without violating privacy concerns and discover the real spatial patterns. In areas including fewer cases, data can be published for areas of medium or larger areal units in order to detect significant clusters as well as conform to confidentiality regulations. A multidisciplinary effort should be taken to develop data publishing criterion that can cover the confidential issues and be utilised to reveal detailed and accurate relationship between health data and local areas. This finding is consistent with the suggestions from Yang et al. (2013), stating the necessity of development of trade-off methods for health data release to balance the privacy and spatial, analytic empowerment.

Several limitations and drawbacks need to be pointed. The distributions of CRC cases at the three reference levels were computed by simulation, and they do not represent actual CRC case locations. This study also constrained the study area to Cook county, a highly-populated and urbanised area, and the corresponding results may not be applicable to suburban or

rural areas. The edge effect may add some bias to the results of the spatial scan statistics at all levels of analysis, especially in locations along the boundary of the study area. Cook county may be too small to identify the statistical significant clusters with restricted criteria (P-value ≤0.05); thus 0.1 was chosen as the cut-off point of statistically significance to select clusters at zip code level and three reference levels for comparison. The findings may also be limited by errors in assigning tracts, block groups and block to their respective zip code areas. The observed variations of spatial aggregation error at zip code level across the study area and lack of actual data at three census reference levels can hardly quantitatively estimate the impact of the spatial aggregation on SaTScan analysis at the zip code level. Additionally, the number of simulated datasets at each reference level was constrained to 100, given the very long processing time of each run in SaTScan. The simulated data may not capture all the possible spatial distributions of cancer cases at each reference level, bringing potential bias due to the inadequate number of possibilities. With the rapid development of super computing, speeding up the application of SaTScan may become a reality in the near future. Then the number of simulated datasets can be increased to include large numbers of distribution possibilities to provide a much more unbiased analysis.

The main tasks for future research are to overcome data limitations and design more appropriate spatial relationships in the disaggregation method to deal with the problem of multiple zip code areas overlapping a single census tract/block group. Introducing zip code tabulation areas (ZCTAs) may be a good idea in terms of using their internal populations to compute the weights for assigning cancer cases from one zip code area to its shared multiple census tracts (US Census Bureau, 2001). In future research, it is also important to use an enlarged study area – a buffer zone – to deal with the edge effect and include more CRC cases to detect the statistical significant clusters with P-value ≤0.05 at the zip code level. Similar to the influence of the spatial aggregation error on zip code level statistical analysis (Luo et al., 2010), the impact of this error on spatial scan analysis has been found to vary with the number of cases across the study area. This empirical evaluation of spatial aggregation error was limited to an urban setting and may not apply to suburban-and rural-areas. More diverse study areas should be studied to obtain detailed information about the impact of the spatial aggregation error on zip code level spatial scan statistics.

Acknowledgements

Financial support from the National Cancer Institute (NCI), National Institutes of Health (NIH), under grant 1-R21-CA114501-01, is gratefully acknowledged. Points of view or opinions in this paper are those of the author, and do not necessarily represent the official position or policies of NCI. The author appreciates Dr. Sara McLafferty from University of Illinois at Urbana-Champaign, who assisted in study design, data collection, and interpretation as well as in extensive editing of the manuscript.

References

Abrams MA, Kleinman PK, 2007. A SatScan™ macro accessory for cartography (smac) package implemented with SAS software. Int J Health Geogr 6, 6.

Amrhein CG, 1994. Searching for the elusive aggregation effect: evidence from statistical simulations. Environ Plann A 27, 105-119.

Fortney J, Kathryn R, Warren J, 2000. Comparing alternative methods of measuring geographic access to health services. Hlth SORM 1, 173-184.

Gregorio DI, DeChello LM, Samociuk H, Kulldorff M, 2005. Lumping or splitting: seeking the preferred areal unit for health geography studies. Int J Health Geogr 4, 6.

Gregorio DI, Kulldorff M, Barry L, Samociuk H, 2002. Geographic differences in invasive and *in situ* breast cancer incidence according to precise geographic coordinates, Connecticut, 1991-1995. Int J Health Geogr 100, 194-198.

Gregorio DI, Kulldorff M, Sheehan TJ, Samociuk H, 2004. Geographic distribution of prostate cancer incidence in the era of PSA testing, Connecticut, 1984 to 1998. Urology 63, 78-82.

Gregorio DI, Samociuk H, 2003. Breast cancer surveillance using gridded population units, Connecticut, 1992 to 1995. Ann Epidemiol 13, 42-49.

Hewko J, Smoyer-Tomic KE, Hodgson MJ, 2002. Measuring neighborhood spatial accessibility to urban amenities: does aggregation error matter? Environ Plann A 34, 1185-1206.

Hillsman E, Rhoda R, 1978. Errors in measuring distances from populations to services centers. Ann Regional Sci 12, 74-88.

Hodgson MJ, Shmulevitz F, Körkel M, 1997. Aggregation error effects on the discrete-space p-median model: the case of Edmonton, Canada. Can Geogr 41, 415-428.

Jemal A, Kulldorff M, Devesa SS, Hayes RB, Fraumeni JF Jr, 2002. A geographic analysis of prostate cancer mortality in the United States, 1970-89. Int J Cancer 101, 168-174.

Krieger N, Chen JT, Waterman PD, Soobader MJ, Subranmanian SV, Carson R, 2002. Geocoding and monitoring of us socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter? Am J Epidemiol 156, 471-482.

Kulldorff M, 1997. A spatial scan statistic. Commun Stat Theory 26, 1481-1496.

Kulldorff M, Feuer EJ, Miller BA, Freedman LS, 1997. Breast cancer clusters in the northeast United States: a geographic analysis. Am J Epidemiol 146, 161-170.

Luo L, McLafferty S, Wang F, 2010. Analyzing spatial aggregation error in statistical models of late-stage cancer risk: a Monte Carlo simulation approach. Int J Health Geogr 9, 51.

Openshaw S, Alvandies S, 1999. Applying geocomputing to the analysis of spatial distributions. In: Geographic information systems: principles and technical issues, volume, I. 2nd edition. Longley P, Goodchild M, Maguire D, Rhind D (eds). New York: John Wiley and Sons.

Pollack LA, Gotway CA, Bates JH, Parikh-Patel A, Richards TB, Seeff LC, Hodges H, Kassim S, 2006. Use of the spatial scan statistic to identify geographic variations in late stage colorectal cancer in California (United States). Cancer Cause Control 17, 449-457.

Roche LM, Skinner R, Weinstein RB, 2002. Use of a geographic information systems to identify and characterize areas with high proportions of distant stage breast cancer. J Public Health Manag Pract 8, 26-32.

Rushton G, 1995. Methods to evaluate geographic access to health services. J Public Health Manag Pract 5, 93-100.

Schmiedel S, Blettner M, Schüz J, 2012. Statistical power of disease cluster and clustering tests for rare diseases: a simulation study of point sources. Spat Spatiotemporal Epidemiol 3, 235-242.

Sheehan TJ, Gershman ST, McDougal L, Danley RA, Mroszczyk M, Sorensen AM, Kulldorff M, 2000. Geographic surveillance of breast cancer screening by tracts, towns and zip codes. J Public Health Manag Pract 6, 48-57.

Thomas A, Carlin BP, 2003. Late detection of breast and colorectal cancer in Minnesota counties: an application of spatial smoothing and clustering. Stat Med 22, 113-127.

US Census Bureau, 2000a. American FactFinder Help-Glossary. US Census Bureau, Washington DC. Available at: http://factfinder.census.gov/home/en/epss/glossary_b.html (accessed on February 2011).

US Census Bureau, 2000b. Census Tracts and Block Numbering Areas. U.S. Census Bureau, Washington DC. Available at: http://www.census.gov/geo/www/cen_tract.html (accessed on January 2011).

US Census Bureau, 2000c. United States Census 2000 Summary File 1(SF1). US Census Bureau, Washington DC Available at: http://www.census.gov/census2000/sumfile1.html (accessed on January 2011).

US Census Bureau, 2001. Zip Code Tabulation Areas (ZCTAs™). US Census Bureau, Washington DC. Available at: http://www.census.gov/geo/ZCTA/zcta.html (accessed on March 2011).

Waller LA, Gotway CA, 2004. Applied spatial statistics for public health data. New York: John Wiley & Sons, Inc.

Yang WJ, Ma KP, Kreft H, 2013. Geographical sampling bias in a large distributional database and its effects on species richness-environment models. J Biogeogr 40, 1415-1426.