Available online at www.sciencedirect.com
ScienceDirect
Computers, Environment and Urban Systems 32 (2008) 81–93
Computers, Environment and Urban Systems
www.elsevier.com/locate/compenvurbsys

Downscaling spatial structure for the analysis of epidemiological data

T.C. Matisziw [a,*], T.H. Grubesic [b], H. Wei [a,c]

[a] Center for Urban and Regional Analysis, The Ohio State University, 1036 Derby Hall, 154 N. Oval Mall, Columbus, OH 43210, USA
[b] Department of Geography, Indiana University, 701 E. Kirkwood Ave, Student Building 120, Bloomington, IN 47405, USA
[c] Department of Geography, The Ohio State University, 1036 Derby Hall, 154 N. Oval Mall, Columbus, OH 43210, USA

Received 31 January 2007; received in revised form 13 June 2007; accepted 20 June 2007

Abstract

Many widely accessible forms of socio-economic and public health data have been subjected to some form of spatial aggregation. Polygonal units, such as US Census tabulation areas, political boundaries, transportation analysis zones and ZIP codes, are often used to represent the spatial extent of these data. While convenient, this type of spatial aggregation, not only ignores the underlying variability of the data of interest, but also the spatial relationships between observations. Although privacy concerns are certainly an important issue in this context, spatial data aggregation can negatively influence one's ability to accurately surveil disease and identify locations (or clusters of locations) with increased disease incidence. The purpose of this paper is to explore the use of supplementary spatial data, particularly street networks, for downscaling the spatial structure of polygonal units. Spatial statistical tests are then used to identify clusters of disease incidence for the downscaled study area. Results suggest that the downscaled spatial structure of polygonal units provides valuable information on the spatial distribution of disease and can facilitate a more detailed spatial analysis of epidemiological data. Prostate cancer data aggregated to ZIP codes for a region of New York State are used to illustrate this concept.

Keywords: ZIP codes; Network analysis; Spatial dependence; Spatial gradation

1. Introduction

In recent years, the spatial analysis of disease has been greatly enhanced through the use of geographic information systems (GIS) (Croner, Sperling, & Broome, 1996; Jacquez, 2004; Pickle, 2002). From exploring the spatial characteristics of West Nile virus (Cooke, Grala, & Wallis, 2006), to identifying cancer clusters (Gregorio, DeChello, Samociuk, & Kuldorff, 2005), the use of spatial data and GIS is becoming fundamental to public health research. Because of this growing popularity, the misuse of spatial data, particularly for spatial statistical applications, is also growing. In many, this can be attributed to an incomplete understanding of the spatial data used for analysis (Anselin, 2006; Boscoe, Ward, & Reynolds, 2004; Grubesic & Matisziw, 2006; Jacquez, 2004). In others, automated processes such as geocoding often mask the spatial imperfections of their output, particularly as it relates to spatial precision and accuracy (Krieger, Waterman, Lemieux, Zierler, & Hogan, 2001; Rushton et al., 2006).

Another area of emerging interest related to GIS and epidemiological data is the spatial representation of data. While the modifiable areal unit problem (MAUP) is well understood in the realm of spatial analysis, epidemiology and public health (Clark & Avery, 1976; Gregorio et al., 2005; Griffith, Wong, & Whitfield, 2003; Openshaw, 1984), developing methods for dealing with multiple spatial scales and aggregations remains an area of increasing research (Louie & Kolaczyk, 2006; Manley, Flowerdew, & Steel, 2006; Rushton & Lolonis, 1996).

As noted by Louie and Kolaczyk (2006), data in spatial epidemiology usually occur in one of two forms: (1) the locations at which cases of the disease occurred, or, (2)

* Corresponding author. Tel.: +1 614 292 8232; fax: +1 614 292 6213.
E-mail addresses: matisziw.1@osu.edu (T.C. Matisziw), tgrubesi@indiana.edu (T.H. Grubesic), wei97@osu.edu (H. Wei).

0198-9715/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compenvurbsys.2007.06.002

aggregate counts of cases that occurred within geographical subregions of some overall region of study. Where the latter instance is concerned, the aggregation of individual-level data is often necessary to ensure the confidentiality of information. In this context, the disclosure of patient locations and their associated conditions is a major breach of medical privacy protocols (Brownstein, Cassa, & Mandl, 2006; Curtis, Mills, & Leitner, 2006). Aggregation of individual-level data is also related to how the data were originally collected and/or reported. For example, US Census 2000 information is not publicly available for each household. Instead, data are distributed in summary form for Census blocks, block groups, tracts and other Census tabulation units (Census, 2000a). Whether driven by privacy concerns or data collection limitations, this type of aggregation is problematic for data analysis since inferences regarding individuals become more difficult to articulate (Anselin & Tam Cho, 2002; Holt, Steel, Tranmer, & Wrigley, 1996; Robinson, 1950).

Aggregation is also an issue for spatial data analysis given that spatial relationships between individuals may not be appropriately represented by the relationships between spatial groupings or study subregions (Miller & Wentz, 2003). Many schemes for representing spatial dependence are based on geographic association of spatial units (see Aldstadt & Getis, 2006; Getis & Aldstadt, 2004; Griffith, 1996). Often, these assumed spatial relationships are imposed on the data in lieu of better information on the spatial processes at work. Therefore, any misrepresentation of spatial dependence between the observed data can obviously impact inferences about spatial processes (Getis & Aldstadt, 2004). Furthermore, attributing all areas within a region with a single measure of spatial association ignores any geographic gradation of spatial relationships. This type of spatial misrepresentation is particularly problematic for disease incidence. Simply put, disease incidence is not continuously distributed in space; rather, cases are limited to the locations of individuals.

This is an important point for several reasons. An obvious consequence of this type of classification error is that the computations of spatial measures which rely on the topologies of geographic subregions often yield a uniform distribution for the measure of interest. In effect, all spatial gradation within the polygons is ignored. Clearly, the spatial relationships between variables in two adjacent areas may not necessarily be uniform throughout each region. Areas closer to a boundary are more proximate (and perhaps more important) to adjacent areas than areas central (or less proximate) to adjacent subregions. This is somewhat akin to the fuzzy boundary transition/gradation problem whereby the spatial effect of adjacency does not pertain to the entire areal unit (Burrough & Frank, 1996; Leung, 1987; Plewe, 1997).

Given the reality that many socio-economic and epidemiological analyses are often limited to data collection and analysis at aggregate geographies, this paper seeks to address a number of issues regarding the impact of data aggregation on exploratory spatial analysis measures that depend on defining some description of spatial dependence. Specifically, given the gradation of socio-economic and epidemiological data within administrative units, how does one best represent geographic space for detecting significant spatial patterns of disease? Further, how can data aggregation and the constant neighbourhood function often associated with vector data influence the spatial statistical analysis of public health data? Can these issues be resolved? Finally, how do relatively unstructured administrative units, such as the ZIP code, impact the analysis of epidemiological data? Considering that data analysis at the ZIP code level is rapidly gaining visibility and popularity in the research community (Banerjee & Carlin, 2004; Eisen, Lane, Fritz, & Eisen, 2006; Gordon et al., 2003), this paper offers an assessment of ZIP code topologies as a natural starting point for addressing these questions.

2. Background

As discussed earlier, although socio-economic and epidemiological data are frequently collected at the individual level, they are often subject to spatial aggregation prior to being made available for research. Thus, while research premised on aggregate geographies is often unavoidable, questions remain regarding the insights that can be gained through exploratory analysis. One important issue relates to the use of aggregate data in ecological and aggregate data studies. Studies addressing these issues are generally oriented toward assessing the efficacy of inference based on area-level data or area-level data supplemented with individual-level data. The main problem of interest in this regard is how to address the bias inherent in making inference on individuals in the absence of complete individual-level data (Holt et al., 1996; Jackson, Best, & Richardson, 2006; Wakefield & Salway, 2001).

Along with obscuring individual-level variability, aggregation can also mask important information on how the variable of interest is spatially distributed. There are several basic concerns in this domain. The first concern relates to the spatial distribution of a data attribute within an areal unit. For example, when lacking supplementary information on the spatial location of individuals, it is often assumed that individuals are uniformly distributed throughout the tabulation area. Clearly, this is not always an accurate assumption. A second concern is whether or not spatial relationships between aggregate areas are representative of spatial relationships between observations in the original data. Given the assumption of a uniform spatial distribution, spatial processes are typically modelled on the perceived relationships between the tabulation areas; irrespective of the underlying geographic distribution of the measured phenomenon. Again, this has the potential to build error into the modelling framework. However, provided information on the spatial distribution of an attribute within an area of analysis, one might have more flexibility in modelling spatial dependency.

Given that we know there is a spatial component to the distribution of data within a tabulation area, many approaches have been proposed to account for this variation. In particular, areal interpolation methods can be applied to downscale spatial units of analysis by transferring geographic attributes from one spatial basis, to another, at a different spatial resolution (Goodchild, Anselin, & Deichmann, 1993). There are numerous approaches available in this regard such as areal weighting (Goodchild & Lam, 1980), pycnophylactic smoothing (Tobler, 1979), dasymetric mapping (Langford, 2006; Langford & Higgs, 2006), and geostatistical modelling (Gelfand, Zhu, & Carlin, 2001; Kyriakidis & Yoo, 2005). See Hawley and Mollering (2005) for further review and comparison of associated methods.

Particularly relevant to the goals of this paper are dasymetric approaches for spatial interpolation. While lack of spatial resolution is a problematic outcome of the aggregation process, data on the *internal* spatial structure of polygonal units are often available. This internal structure may consist of known areas of human habitation that can be represented by household locations, addresses, or transportation networks. Incorporating this information into spatial analysis of aggregate data is the basic premise of dasymetric mapping. Dasymetric mapping seeks to use such spatial information to redistribute an aggregate measure across the associated areal unit to produce a more realistic spatial rendering of the attribute. Ancillary information such as road data, which can be viewed as indicative of the distribution of human populations, has proven especially useful in this context (Mrozinski & Cromley, 1999; Reibel & Bufalino, 2005; Xie, 1996). Similarly, incorporation of ancillary geographic information can enable more defensible models of spatial dependence and association to be postulated given that the added structure permits assessment of where the spatial phenomena may actually be located. This is particularly important for determining topological relationships, such as spatial adjacency, for a study area.

As an example, consider Fig. 1a which illustrates three US ZIP code areas in Syracuse, New York (13209, 13088 and 13204). If one was to construct a simple, spatial adjacency matrix for this small study area, each of these polygons share borders and could be considered neighbours. More importantly, if epidemiological data are aggregated to these ZIP code areas, one would be forced to assume that the distribution of disease incidence is spatially uniform within the polygons. Now, consider Fig. 1b, which contains supplemental spatial information on water features within these ZIP code areas. In this instance, the spatial relationships between these polygons are somewhat muddled. Lake Onondaga is uninhabitable by humans. As a result, there is a significant spatial gap in ZIP code areas and this may lead to the inaccurate attribution of disease incidence. Finally, Fig. 1c displays the spatial correspondence between transportation infrastructure and the presence of major uninhabited features such as lakes. Where spatial relationships are concerned, particularly adjacency, the association between 13209, 13088 and 13204 is certainly less obvious. In fact, there is only one road that actually connects 13209 to 13088. This is a much different spatial relationship when compared to a simple polygonal representation of ZIP code boundaries (Fig. 1a).

Obviously, these types of results are not limited to *uninhabitable* features such as water areas, deserts or swamps. In many cases, the land within an administrative unit such as a county, ZIP code or Census tract is simply *uninhabited* – a subtle but important difference. As mentioned earlier, these types of representation errors for geographic space open a wide variety of questions pertaining to the spatial analysis of socio-economic and epidemiological data.

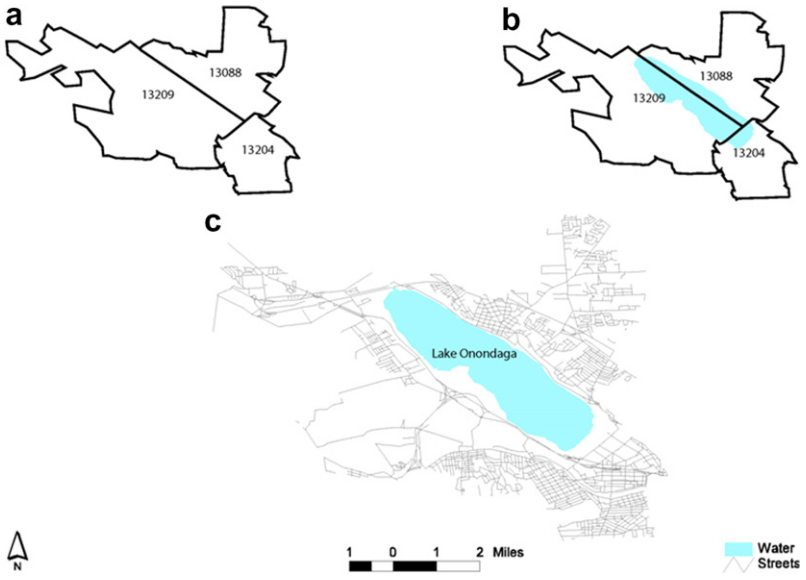

Fig. 1. Elements of spatial decomposition and spatial adjacency using supplemental spatial information: Year 2000 ZIP code areas in Syracuse, NY.

3. Representation of ZIP code data

The ZIP (Zone Improvement Plan) code refers to ranges of postal addresses along road networks which are used to facilitate mail delivery by the United States Postal Service (USPS). Functionally, ZIP codes present researchers with a geographically interesting alternative to other spatial units of analysis. First, ZIP code data is relatively easy to collect. For example, in retailing, the collection of ZIP code information is typically done at the point of sale, where customers are asked to supply their home ZIP code before a transaction takes place. A wide variety of epidemiological data is also made available at the ZIP code level. For example, the New York State Department of Health (NYSDOH) provides cancer incidence data by ZIP code (NYSDOH, 2006). From a pragmatic perspective, the process of associating individuals with ZIP codes is much less involved than associating them with actual street addresses, given that geocoding can be plagued with errors (Krieger et al., 2001; Rushton et al., 2006), requires specialized software, and involves significant computational effort. Second, many ZIP code areas are often smaller in spatial extent than counties and other US Census defined data collection units, particularly in urban areas (Grubesic, in press). As a result, it is suggested that ZIP codes offer a finer spatial resolution than other areal units for analysis (Johnson, 2004; Luo & Wang, 2003). Finally, and most importantly, since ZIP codes are based on addresses and associated network topologies, they are strongly representative of the location of human activities.

While ZIP codes may be convenient units for analysis, recent research has highlighted many potential pitfalls in the analysis of ZIP code data to which researchers should be mindful (Grubesic, in press; Grubesic & Matisziw, 2006; Krieger et al., 2002). For instance, ZIP codes do not conform to other spatial units commonly used for data collection. While Census tabulation units such as blocks, block groups and tracts are hierarchical and nested, ZIP codes are more dynamic given the needs of the USPS. Where data aggregation and representation are concerned, this makes the spatial composition of ZIP codes both highly irregular and transient (Grubesic & Matisziw, 2006). Further, although ZIP codes are linear features which are defined by postal address ranges and streets, they are often modelled as geographical zones, areas and subregions in the research community (Cooke et al., 2006; Luo & Wang, 2003). These representations have been produced and popularized by private companies such as Geographic Data Technology (now Tele Atlas) and more recently by the US Census Bureau's ZIP code tabulation area (ZCTA) (Census, 2000b). In this context, the accurate generalization of these address ranges into polygonal areas for the purpose of data collection and analysis is difficult when one considers the topological structure of ZIP code ranges. Simply put, there is not always a complete correspondence between delivery networks and derived ZIP code areas (Grubesic & Matisziw, 2006).

Given the myriad of considerations that can impact analyses based on ZIP codes, they do offer some unique prospects for spatial analysis. As discussed earlier, ZIP code address ranges defined by the USPS provide insight as to *where* human activity is actually located. In other words, network-based geographies, such as streets and their associated addresses, correspond to the locations of individuals, families, and businesses in geographic space. These are the geographic relationships that are of interest when exploring or modelling spatial phenomenon, particularly disease incidence, which is intimately tied to the daily lives of people and their local environment. This also suggests that tracking ZIP code features by address ranges and their associated network-based geographies (i.e., streets), instead of polygons, might present a viable alternative representation of geographic space for epidemiological analysis. This type of spatial decomposition is especially promising for statistical analysis which incorporates measures of spatial dependence for detecting underlying patterns – such as clusters of disease incidence. Finally, disaggregate topologies can permit a more detailed picture of spatial gradation of spatial analysis measures, particularly when it comes to adjacency metrics. Next, we will assess the potential for the use of network attributed ZIP code data in spatial analysis.

4. Implications of ZIP code representation on spatial analysis

To better explore and understand the benefits of using network-based representations of ZIP codes versus their commonly used area-based counterparts, patterns of prostate cancer incidence is analyzed. The following sections describe the data and methods involved in this study.

4.1. Data

Data on male prostate cancer incidence for the period 1999–2003 by ZIP code was obtained for seven counties in North-West New York State (NYSDOH, 2006). These counties are depicted in Fig. 2. To ensure the confidentiality of individuals involved, a small subset of ZIP codes (including those representing postal boxes) and their associated cancer incidence data were merged for public disclosure (NYSDOH, 2006). TIGER-based street networks, obtained from the Caliper Corporation (2006) and attributed with year 2000 ZIP code information, represent the geographic extent of each ZIP code. Road segments (arcs) in the study area without ZIP code information (approximately 28.6% of roads) are not further considered in this analysis.[1] Since ZIP codes relate to address ranges, the ZIP codes attributed to the left and right side of a road

[1] In most cases, this represents a combination of interstate highway segments and relatively rural segments in unpopulated areas, such as state parks. In other cases, ZIP code information for a street segment was simply not available.

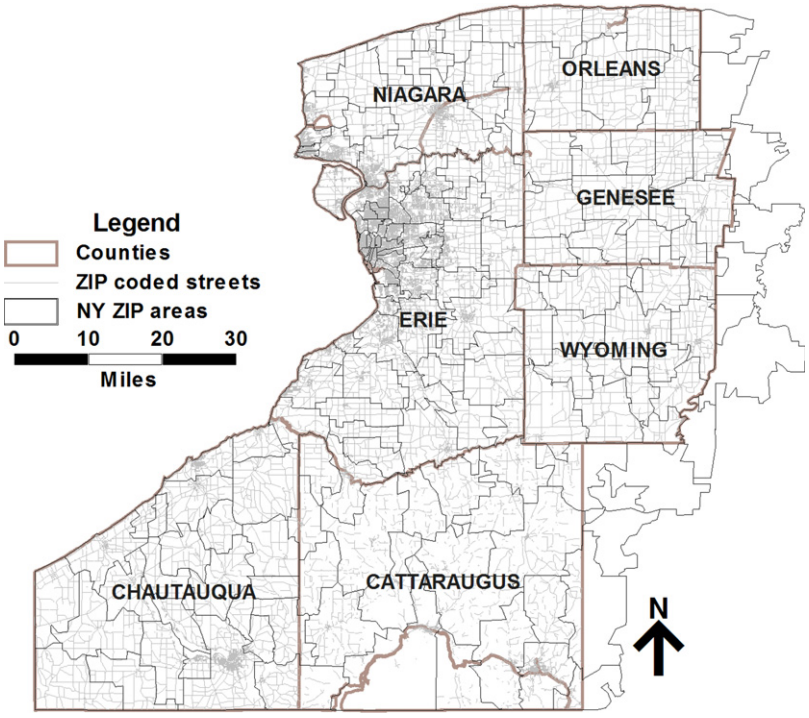

Fig. 2. Study area in North-West New York State.

may differ. To ensure that all ZIP coded arcs are included in the following analysis, arcs were added when necessary, such that both ZIP codes (left and right sides of the road) are represented. Given these modifications, a total of 68,442 network arcs are utilized for analysis in the study area (shown in Fig. 2). ZIP code boundaries for the study area (year 2000), distributed by Geographic Data Technology (GDT), were also obtained for comparative purposes.[2] Once the aggregation scheme incorporated by NYSDOH (2006) for cancer data was incorporated, a total of 171 ZIP code polygons are used for analysis (Fig. 2). Finally, for the purpose of spatial analysis, cancer incidence data were attributed to road arcs and ZIP code areas – based on the *aggregated* structure of the data. This process ensures both consistency and uniformity in the treatment of the observed cancer data for the study area. Population data for each ZIP code area and associated arcs was then obtained though aggregating Census 2000 block data on male population to the ZIP code level to calculate prostate cancer rates.

4.2. Methodology

From a topological perspective, ZIP coded network arcs and ZIP code areas are quite different – even though these two geographies are meant to represent the exact same phenomena (USPS delivery locations). For example, while ZIP code areas can have a sizable spatial extent, ZIP coded arcs are sparser and typically cover very little geographic area (in total). These differences can result in a diversity of spatial neighbourhood structures given the ZIP representation used. In this paper, three approaches for representing spatial relationships between ZIP codes are considered. These approaches involve assessing: (1) adjacency between ZIP code areas, (2) adjacency between network-based ZIP code objects, and (3) adjacency between individual components (arcs) of ZIP coded network objects.

For the purposes of this study, neighbourhoods for ZIP code areas are based on Queen's adjacency and indicate that each ZIP code area has, on average, 7.0 neighbours.[3] Where ZIP coded arcs are concerned, rather than using traditional spatial contiguity measures, adjacency is based on their proximity (i.e., distance) to other arcs. This is necessary since individually, the vast majority of arcs share relatively few contiguous neighbours.[4] Further, because many postal addresses are discontinuous (see Grubesic & Matisziw, 2006), the use of a distance threshold provides a safely conservative estimation of neighbourhood structure that can be modified for examining spatial statistical sensitivities. In the following discussion, it is useful to differentiate between two types of arcs: (1) *Native* arcs refer to neighbours of a particular road segment i that belong to the same ZIP code as i, (2) *Non-native* arcs are those neighbours of i belonging to different ZIP codes.

[2] GDT was recently acquired by Tele Atlas.

[3] Queen's adjacency defines "neighbours" as those polygons which share an arc or vertex.

[4] Other neighbourhood structures might be considered, such as the network associations employed by Black (1992).

In the first type of arc-based analysis, we treat all network arcs sharing the same ZIP code as a single network object. In this case, two ZIP code objects (m and k) are considered adjacent if they are within a specified distance of one another. Given a threshold distance R and a network object m, we identify the set of j non-native arcs (N_m), such that $d_{mj} \leqslant R$ (where d_{mj} = shortest distance between network object m and arc j). For example, Fig. 3a depicts a spatial neighbourhood associated with a single ZIP coded network object. Adjacency between objects m and k (w_{mk}) can be weighted by the number (cardinality) of neighbouring arcs found in k, $|N_m^k|$. The weights matrix can then be row standardized; in this case, $w_{mk} = \left(\frac{|N_m^k|}{|N_m|}\right)$.

The second approach, considers each arc as a representative component of a ZIP code. In this instance, the spatial neighbourhood of each arc consists of the ZIP codes asso-

Fig. 3. One mile neighbourhood for a single: (a) ZIP coded network object and (b) ZIP coded arc.

ciated with all non-native arcs falling within some prescribed distance of the arc. Thus, given a threshold distance R and an arc i, we identify the set of j non-native arcs (N_i), such that $d_{ij} \leqslant R$ (where d_{ij} = shortest distance between arcs i and j). Fig. 3b illustrates a one mile neighbourhood associated with a single ZIP coded arc. Adjacency between each ZIP coded arc and non-native ZIP code k (w_{ik}) is weighted by the number of neighbouring non-native arcs found in k, $|N_i^k|$. Again, the weights matrix can be row standardized; in this instance, $w_{ik} = \left(\frac{|N_i^k|}{|N_i|}\right)$.

Local Moran's I, a commonly employed measure of spatial autocorrelation is used to illustrate the differences between spatial analyses utilizing ZIP code areas and ZIP coded network topologies (see Anselin, 1995). The local Moran's I index can be used to assess the similarity of the value of one spatial unit with the values associated with its defined neighbourhood. Both the global and local versions of Moran's I require the specification of a weights matrix representing the spatial relationships between the objects of interest. To highlight the differences between using each of the three ZIP code representations detailed above, the resulting neighbourhood relationships for each topology are input into a spatial analysis software package, GeoDa, for further examination (Anselin, 2003). Here, two distance-based neighbourhoods (one and two miles) are specified to construct the necessary spatial weights matrices for each network-based topology. These neighbourhoods are identified using spatial query functions of ArcView, a commercial GIS software package. Global and local Moran's I are then computed for each ZIP code topology and neighbourhood specification. A permutation approach (involving 9,999 random permutations) is used to model the distributional aspects of the local Moran's I (Anselin, 1995). Given the resulting calculations, the three ZIP code representations and neighbourhood delineations are then compared based on the resulting patterns and characteristics of autocorrelation detected.

5. Results and discussion

Exploratory spatial analysis is performed using each of the three ZIP code-based topologies given the aggregation scheme employed by NYSDOH (2006). In each case, global and local Moran's I are computed, with the local Moran values classified according to deviation of cancer rate from the mean and the sign of the local I values. As such, a high–high classification refers to a positive deviation from the mean and a positive I value while a low–low classification indicates a negative deviation from the mean and a positive I value. Accordingly, high–low is used to delineate ZIP objects that exhibit positive deviations from the mean and have negative I values while low–high reflects negative deviation from the mean. In all instances, global I is positive and significant at the 0.001 level and corresponds to low-to-moderate spatial autocorrelation (specifically, spatial clustering of similar values).

5.1. Polygon-based topology

For the area-based ZIP code topology, the resulting local Moran's I classifications for prostate cancer rates are shown in Fig. 4 at the 0.001 significance level (global

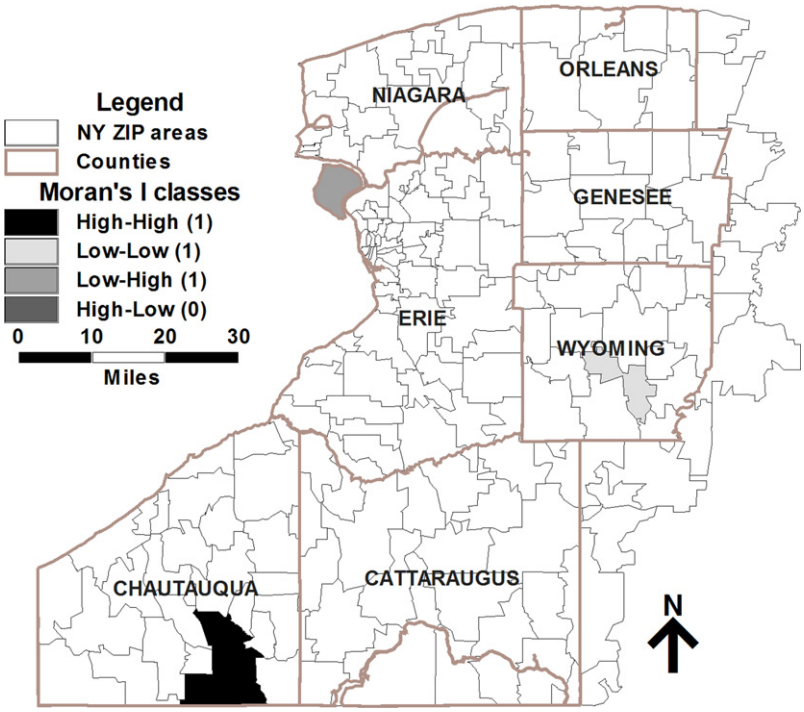

Fig. 4. Significant ($\alpha = 0.001$) clustering of ZIP code area-based prostate cancer rates.

$I = 0.1833$). In this instance, only three ZIP areas are significant (one area's significance is attributable to absence of neighbours). The pattern in this scenario indicates significant clustering of high levels of correlation in Chautauqua County and significant clustering of low values in Wyoming County. However, for the most part there are few other areas of significant spatial correlation present within the given study area. Especially interesting is the absence of any significant patterns in the more densely populated Buffalo metropolitan areas of Niagara and Erie counties. As discussed earlier, such area-based representations of spatial relationships assume that the entire contents of each area is equally influenced by the defined measure of spatial adjacency. Thus, the results here assume that spatial correlation is significant across each of the highlighted areas.

5.2. Network object-based topology

Use of the second ZIP topology entails treating all arcs attributed with the same ZIP code as a single network-based object. As discussed earlier, adjacency between ZIP code objects is weighed by the number of non-native arcs (for each non-native ZIP code) within each object's neighbourhood. Fig. 5 illustrates the resulting patterns of significant spatial correlation for prostate cancer rates given a one mile neighbourhood for each ZIP coded network object and Tables 1a and 1b summarize the local I values given specification of both one and two mile neighbourhoods. Of interest is the fact that more ZIP codes (approximately 12% for the one mile neighbourhood) are now involved in significant spatial correlation. This is a direct result of weighting adjacencies by the amount of non-native transportation infrastructure (e.g., arcs) within each ZIP code's neighbourhood. From an interpretive standpoint, the large cluster of high–high correlations in southern Chautauqua and western Cattaraugus counties is a marked deviation from the results obtained using the polygon topology (Fig. 5). That said, it is also important to note the lack of complete correspondence between derived ZIP code boundaries and associated network defined geographies. For instance, in Chautauqua County, some of the

Table 1a
Local Moran's I summary for network objects specifying one mile neighbourhood (Global $I = 0.3715$)

Moran's I Class	% Obs.	Minimum	Maximum	Mean
High–high	6.1	0.0374	9.0730	4.0210
High–low	1.2	−0.4041	−0.0288	−0.2164
Low–high	1.2	−1.5576	−0.2495	−0.9036
Low–low	3.7	0.0898	1.0879	0.4872
Not significant	87.7	−0.6188	1.6002	0.1375
Total	100.0	−1.5576	9.0730	0.3715

Table 1b
Local Moran's I summary for network objects specifying two mile neighbourhood (Global $I = 0.3546$)

Moran's I Class	% Obs.	Minimum	Maximum	Mean
High–high	6.1	0.0412	9.0847	4.0483
High–low	1.8	−0.7609	−0.0296	−0.3899
Low–high	1.2	−1.3336	−0.2477	−0.7907
Low–low	3.7	0.0925	0.5406	0.3305
Not significant	87.1	−0.5770	1.5937	0.1273
Total	100.0	−1.3336	9.0847	0.3546

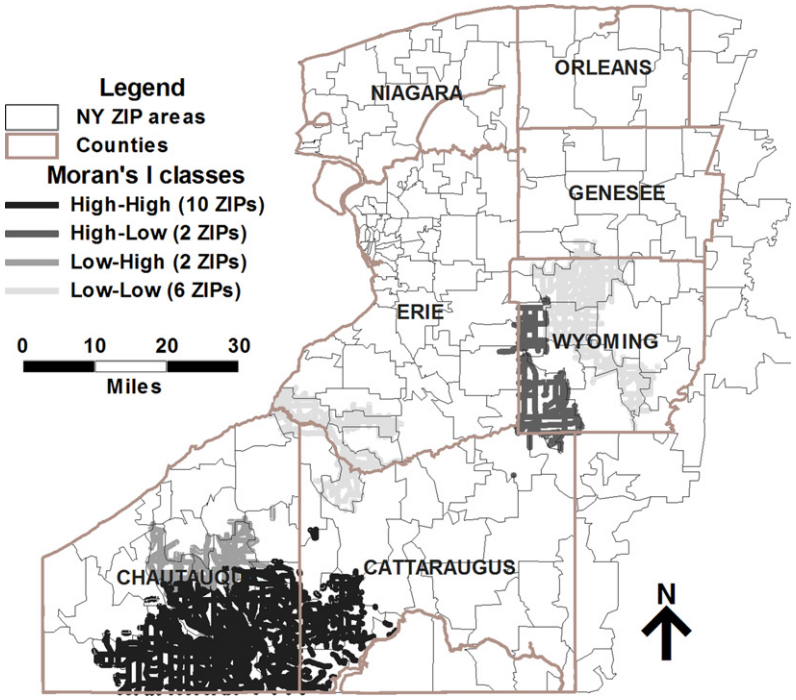

Fig. 5. Significant clustering of network object-based prostate cancer rates (one mile neighbourhood).

significant arcs clearly overlap with ZIP code areas (polygons) that do not exhibit significant correlation (both to the east and west of the initial high–high ZIP code polygon). While this certainly provides a more interesting snapshot of prostate cancer rates for the region, issues of spatial uniformity (similar to that found in Fig. 4) remain.

5.3. Individual arc-based topology

Finally, local spatial correlation is assessed provided one and two mile neighbourhoods (i.e., distance threshold) for each arc. Given a one mile neighbourhood, 18.7% of the network arcs in the study area are found to have significant local Moran's I values at the 0.001 level (Table 2a). Fig. 6a depicts those arcs exhibiting significant spatial autocorrelation in this scenario. Specification of a two mile neighbourhood resulted in 39.6% of network arcs having significant local Moran's I values at the 0.001 level (Table 2b). Perhaps most interesting here are the dramatic spatial differences in the resulting clusters when the size of the spatial neighbourhood is altered. When comparing Fig. 6a and b it becomes apparent that as the neighbourhood for each arc grows, becoming influenced by a greater number of other ZIP codes (as weighted by associated arcs), the number of significant arcs also increases. Overall, the number of ZIP code areas that contain significant arcs is greater than in the case of the area and network object-based analysis. This is reasonable given that more spatial objects are involved, enabling more variability in the classifications.

Further, although the use of both area and network object-based adjacency measures used in the generation of Figs. 4 and 5 yield some interesting results, the assumption of uniformity between polygons and network objects is not always realistic. Fig. 6a shows Low–low clusters have an obvious dominance in this study region, especially in the Buffalo area of Erie County. Interestingly, this portion of the study region was not identified as significant in the analysis of ZIP polygons or network objects.

A high–high cluster is still present in south-eastern Chautauqua County (Jamestown, NY), but the number of ZIP code areas having arcs classified as high–high was noticeably smaller. This is likely due to a sparser transportation network and the larger spatial extent of the ZIP codes found in this region, and hence the lower number of neighbours for each arc. A similar instance of Low–low clustering can be observed in Wyoming County, where given the defined distance threshold, significant spatial autocorrelation is limited to the border regions of each ZIP code polygon. This is a particularly notable result and suggests that some portions of a ZIP code may not necessarily be influenced by all other surrounding ZIP codes given the defined neighbourhood. Hence, as discussed in Section 1, the relationship between polygons is not always as clear-cut as it might appear because some areas of a polygon may not have a strong spatial connection (e.g., via transportation arcs) to other polygons. In these instances, modelling the neighbourhoods of ZIP codes based on component arcs, as done here, permits a gradient of spatial relationship to be incorporated into models of spatial analysis.

For instance, the two mile neighbourhood results depicted in Fig. 6b show an increased number of significant arcs in the Buffalo metro area. Further, a more established cluster of high–low spatial correlation emerges in this area. A majority of the arcs in these metropolitan ZIP codes are involved in significant correlations given their relatively small spatial extent and the density of the arcs involved. It is also important to note that the central regions of many ZIP codes are not involved in significant autocorrelation. This is because neighbourhoods of arcs in these areas are composed entirely of a native ZIP code arcs, hence there are no associations with non-native ZIP codes. Further, Fig. 6b displays a growth of significant high–high arcs in south-eastern Chautauqua County, while new high–high clusters are found in Erie County. Also, there is a general trend of increasing negative spatial correlation that can be observed throughout the region.

Table 2a
Local Moran's I summary for arc-based analysis specifying one mile neighbourhood (average number of neighbours = 74; Global $I = 0.2231$)

Moran's I Class	% Obs.	Minimum	Maximum	Mean
High–high	1.4	0.1256	83.9324	9.9672
High–low	2.4	−1.6779	−0.0645	−0.3528
Low–high	3.9	−1.6779	−0.0075	−0.1101
Low–low	11.0	0.0037	1.6908	0.3271
Not significant	81.3	−1.6779	4.4240	0.0705
Total	100.0	−1.6779	83.9324	0.2231

5.4. Subtleties in observed spatial correlation

As highlighted earlier, not all areas of a ZIP code polygon are necessarily involved in significant clustering of arcs. In many cases, ZIP code polygons are in fact composed of several different classes of spatial correlation. This behaviour is illustrated in Fig. 7 for the Buffalo metropolitan area. In this case, there are many instances where significant ZIP coded arcs cover only a portion of ZIP code polygons. See for instance, the 14226, 14224, and 14086 ZIP code areas. Additionally, ZIP code polygons need not be composed of arcs bearing a single classification. For example, Fig. 7 depicts several ZIP areas, notably 14221 and 14150, having statistically significant road segments of several correlation classifications. This behaviour is indicative of a transition or gradient of spatial autocorrelation within a ZIP code area. In the case of ZIP code 14221, there is a spatial gradient moving from high–high correlation in the ZIP's southwest portion to high–low correlation in the ZIP's northern regions. In another instance, ZIP code 14150 shows a gradient from low–low correlation in its south-eastern corner to low–high correlation in its northern sector. Again, the absence of significant arcs in

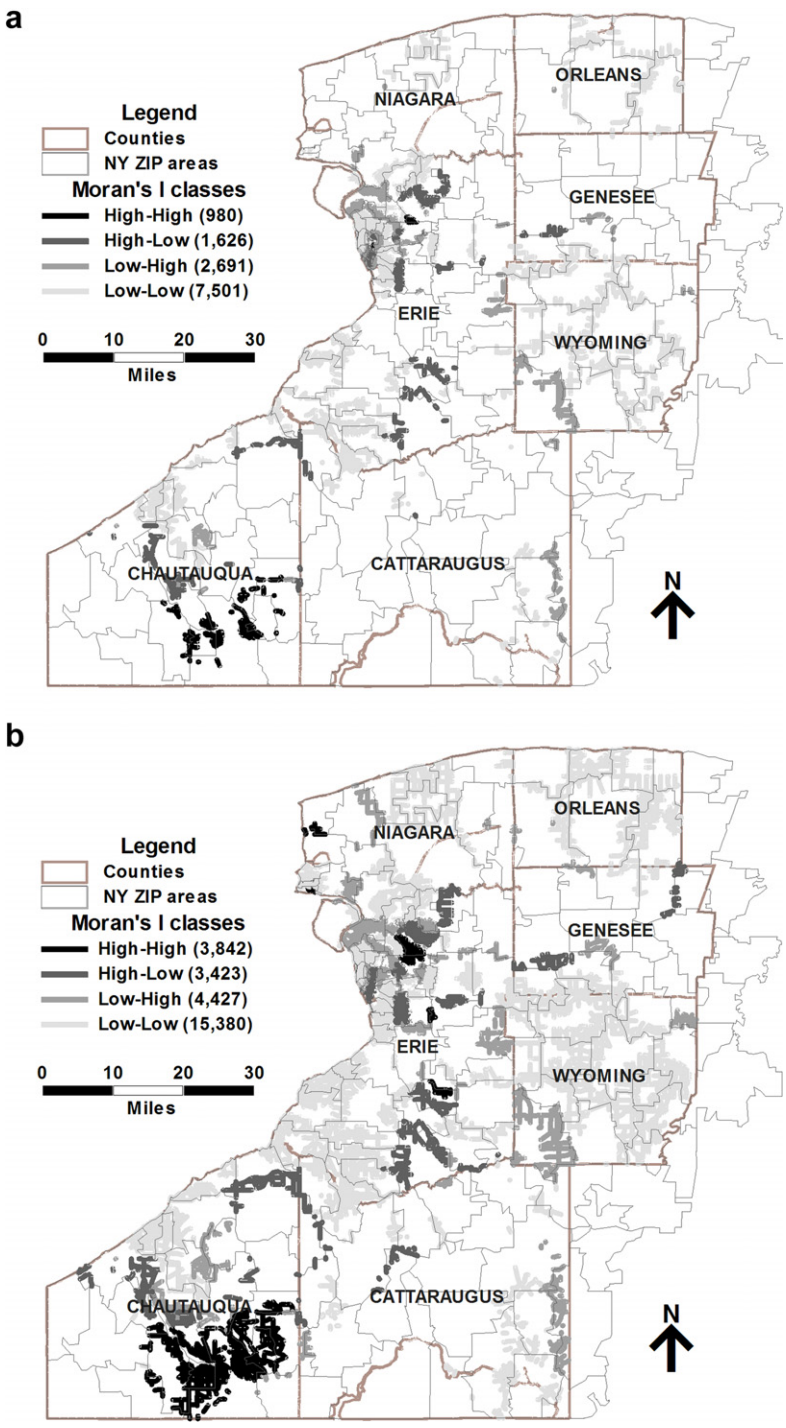

Fig. 6. Significant clustering of arc-based prostate cancer rates: (a) one mile neighbourhood and (b) two mile neighbourhood.

areas central to ZIP codes can, in many cases, be due to arcs lacking non-native ZIP code neighbours.

6. Conclusions

Traditionally, spatial analysis of aggregated data has focused on the use of the spatial unit of aggregation in deriving models of spatial dependence. However, representing all areas within a single spatial unit as uniformly possessing a single spatial relationship ignores any internal spatial structure of the phenomena of interest. This is particularly of concern given that many types of socio-economic data are not continuously sampled in space, but rather are collected from discrete locations. The motivation for this paper is therefore to experiment with downscaling the spatial structure of spatial units of analysis using available data suggestive of more realistic descriptions of spatial dependence. Such downscaling of space can be used to pro-

Table 2b
Local Moran's I summary for arc-based analysis specifying two mile neighbourhood (average number of non-native neighbours = 318; Global $I = 0.3106$)

Moran's I Class	% Obs.	Minimum	Maximum	Mean
High–high	5.6	0.0401	93.3429	4.2529
High–low	5.0	−1.6779	−0.0379	−0.2586
Low–high	6.5	−1.6779	−0.0011	−0.0944
Low–low	22.5	0.0018	1.6908	0.2250
Not significant	60.4	−1.4790	4.2155	0.0667
Total	100.0	−1.6779	93.3429	0.3106

duce a more refined spatial representation of the area of interest giving socio-economic analysts, epidemiologists and public health workers more detail as to possible underlying spatial relationships at work. This is especially important given that mitigation and management of disease can be costly and thus, a focus on geographies smaller than the available tabulation area is essential.

In this paper, US ZIP code data are used to illustrate several approaches to this problem. Unlike US Census data, ZIP code data are not geographically nested and spatial analyses are typically limited to using available area representations of ZIP codes. Here, a transportation network is used to incorporate a better spatial representation of the geographic relationships that may be of concern when assessing patterns related to prostate cancer. Representation of ZIP codes as an arc-based topology is a natural approach given that ZIP codes are actually defined for road networks serving population, not polygons. Hence, use of ZIP coded arcs permits the spatial relationships between portions of ZIP codes (as represented by associated arcs) to be modelled in greater detail. Three different representations of ZIP code data are postulated in this study: (1) polygon, (2) network object, and (3) arc-based. For each representation, we detail how descriptions of spatial neighbourhoods can be constructed for use in many types of analyses. The local Moran's I statistic, a measure of spatial association that requires spatial neighbourhoods to be defined in advance for all units of analysis, is used to illustrate the impact of geographic downscaling on spatial statistics. The resulting maps show how spatial correlation is affected by downscaling ZIP code areas based on the underlying network topology and highlight the insight that can be gained over traditional analysis – particularly with respect to identifying possible pockets of significant spatial correlation. It is shown that downscaling in this manner can help reveal patterns of spatial correlation that may otherwise be obscured. In this study, all network arcs having ZIP code information were included for analysis. However, some arcs may not (at least in part) represent populated areas. Given this, future research might work to further decompose such structures to represent only populated areas and to attribute arcs with more refined population data (e.g., dasymetric approaches). Such work would permit population variability within ZIP codes to be incorporated in support of spatial analysis.

Acknowledgement

Funding for portions of this research was provided by The Center for Urban and Regional Analysis (CURA) at The Ohio State University.

References

Aldstadt, J., & Getis, A. (2006). Using AMOEBA to create a spatial weights matrix and identify spatial clusters. *Geographical Analysis, 38*(4), 327–343.

Anselin, L. (1995). Local indicators of spatial association – LISA. *Geographical Analysis, 27*(2), 93–115.

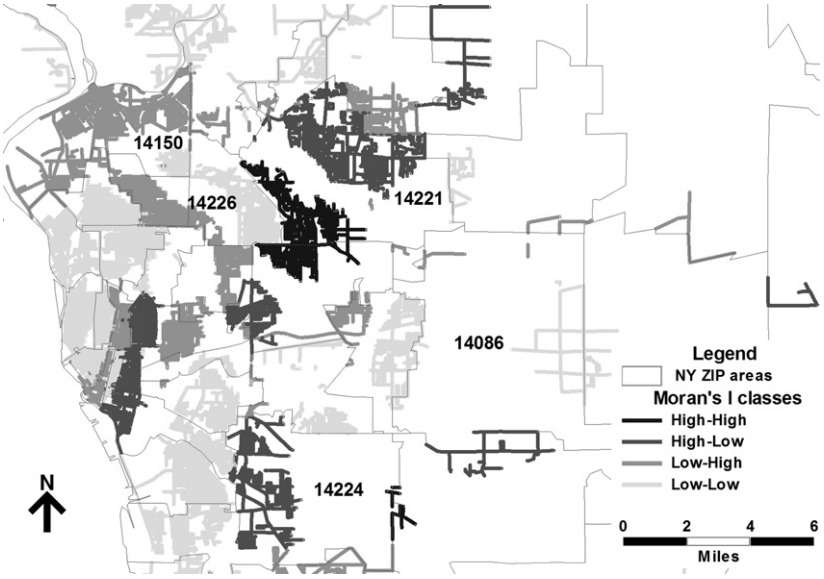

Fig. 7. Significant clustering in the Buffalo, NY Area (two mile neighbourhood).

Anselin, L. (2003). GeoDa 0.9 user's guide. Spatial Analysis Laboratory, Department of Agricultural and Consumer Economics, University of Illinois. https://www.geoda.uiuc.edu/.

Anselin, L. (2006). How (not) to lie with spatial statistics. *American Journal of Preventative Medicine, 30*(28), S3–S6.

Anselin, L., & Tam Cho, W. K. (2002). Spatial effects and ecological inference. *Political Analysis, 10*(3), 276–297.

Banerjee, S., & Carlin, B. P. (2004). Parametric spatial cure rate models for interval-censored time-to-relapse data. *Biometrics, 60*(1), 268–275.

Black, W. (1992). Network autocorrelation in transport network and flow systems. *Geographical Analysis, 24*, 207–222.

Boscoe, F. P., Ward, M. H., & Reynolds, P. (2004). Current practices in spatial analysis of cancer data: Data characteristics and data sources for geographic studies of cancer. *International Journal of Health Geographics, 3*(28).

Brownstein, J. S., Cassa, C. A., & Mandl, K. D. (2006). No place to hide – reverse identification of patients from published maps. *New England Journal of Medicine, 355*(16), 1741–1742.

Burrough, P., & Frank, A. U. (1996). *Geographic objects with indeterminate boundaries*. London: Taylor and Francis.

Caliper Corporation (2006). URL: http://www.caliper.com.

Census Bureau [United States] (2000a). Census 2000 gateway. URL: http://www.census.gov/main/www/cen2000.html.

Census Bureau [United States] (2000b). Census 2000 ZCTAs: ZIP code tabulation areas technical documentation. URL: www.census.gov/geo/ZCTA/zcta_tech_doc.pdf.

Clark, W. A. V., & Avery, K. (1976). The effects of data aggregation in statistical analysis. *Geographical Analysis, 8*, 428–438.

Cooke, W. H., Grala, K., & Wallis, R. C. (2006). Avian GIS models to signal human risk for west nile virus in Mississippi. *International Journal of Health Geographics, 5*(36).

Croner, C. M., Sperling, J., & Broome, F. R. (1996). Geographic Information Systems (GIS): New perspectives in understanding human health and environmental relationships. *Statistics in Medicine, 15*(17–18), 1961–1977.

Curtis, A., Mills, J. W., & Leitner, M. (2006). Keeping an eye on privacy issues with geospatial data. *Nature, 441*(7090), 150.

Eisen, R. J., Lane, R. S., Fritz, C. L., & Eisen, L. (2006). Spatial patterns of lyme disease risk in California based on disease incidence data and modeling of vector-tick exposure. *American Journal of Tropical Medicine and Hygiene, 75*(4), 669–676.

Gelfand, A. E., Zhu, L., & Carlin, B. P. (2001). On the change of support problem for spatio-temporal data. *Biostatistics, 2*(1), 31–45.

Getis, A., & Aldstadt, J. (2004). Constructing the spatial weights matrix using a local statistic. *Geographical Analysis, 36*(2), 90–104.

Goodchild, M. F., Anselin, L., & Deichmann, U. (1993). A framework for the areal interpolation of socioeconomic data. *Environment and Planning A, 25*, 383–397.

Goodchild, M. F., & Lam, N. (1980). Areal interpolation: A variant of the traditional spatial problem. *Geo-Processing, 1*, 297–312.

Gordon, T. E. J., Leeth, E. A., Nowinski, C. J., MacGregor, S. N., Kambich, M., & Silver, R. K. (2003). Geographic and temporal analysis of folate-sensitive fetal malformations. *Journal of the Society for Gynecologic Investigation, 10*(5), 298–301.

Gregorio, D. I., DeChello, L. M., Samociuk, H., & Kuldorff, M. (2005). Lumping or splitting: Seeking the preferred areal unit for health geography studies. *International Journal of Health Geographics, 4*, 6.

Griffith, D. A. (1996). Some guidelines for specifying the geographic weights matrix contained in spatial statistical models. In S. L. Arlinghaus (Ed.), *Practical handbook of spatial statistics*. Boca Raton: CRC.

Griffith, D. A., Wong, D. W. S., & Whitfield, T. (2003). Exploring relationships between the global and regional measures of spatial autocorrelation. *Journal of Regional Science, 43*(4), 683–710.

Grubesic, T. H. (in press). ZIP codes and spatial analysis: Problems and prospects. *Socio-Economic Planning Sciences*, in press, doi:10.1016/j.seps.2006.09.001.

Grubesic, T. H., & Matisziw, T. C. (2006). On the use of ZIP codes and ZIP code tabulation areas (ZCTAs) for the spatial analysis of epidemiological data. *International Journal of Health Geographics, 5*(58).

Hawley, K., & Mollering, H. (2005). A comparative analysis of areal interpolation methods. *Cartography and Geographic Information Science, 32*(4), 411–423.

Holt, D., Steel, D. G., Tranmer, M., & Wrigley, N. (1996). Aggregation and ecological effects in geographically based data. *Geographical Analysis, 28*(3), 244–261.

Jackson, C., Best, N., & Richardson, S. (2006). Improving ecological inference using individual-level data. *Statistics in Medicine, 25*, 2136–2159.

Jacquez, G. M. (2004). Current practices in the spatial analysis of cancer: Flies in the ointment. *International Journal of Health Geographics, 3*(22).

Johnson, G. D. (2004). Small area mapping of prostate cancer incidence in New York state (USA) using fully bayesian hierarchical modeling. *International Journal of Health Geographics, 3*(29).

Krieger, N., Waterman, P., Chen, J. T., Soobader, M.-J., Subramanian, S. V., & Carson, R. (2002). ZIP code caveat: Bias due to spatiotemporal mismatches between ZIP codes and US census-defined geographic areas – the public health disparities geocoding project. *American Journal of Public Health, 92*(7), 1100–1102.

Krieger, N., Waterman, P., Lemieux, K., Zierler, S., & Hogan, J. W. (2001). On the wrong side of the tracts? Evaluating accuracy of geocoding for public health research. *American Journal of Public Health, 91*, 1114–1116.

Kyriakidis, P. C., & Yoo, E. H. (2005). Geostatistical prediction and simulation of point values from areal data. *Geographical Analysis, 37*(2), 124–151.

Langford, M. (2006). Obtaining population estimates in non-census reporting zones: An evaluation of the 3-class dasymetric method. *Computers, Environment, and Urban Systems, 30*, 161–180.

Langford, M., & Higgs, G. (2006). Measuring potential access to primary healthcare services: The influence of alternative spatial representations of population. *The Professional Geographer, 58*(3), 294–306.

Leung, Y. (1987). On the imprecision of boundaries. *Geographical Analysis, 19*, 125–151.

Louie, M. M., & Kolaczyk, E. D. (2006). A multiscale method for disease mapping in spatial epidemiology. *Statistics in Medicine, 25*(8), 1287–1308.

Luo, W., & Wang, F. (2003). Measures of spatial accessibility to healthcare in a GIS environment: Synthesis and a case study in Chicago region. *Environment and Planning B, 30*(6), 865–884.

Manley, D., Flowerdew, R., & Steel, D. (2006). Scales, levels and processes: Studying spatial patterns of British census variables. *Computers, Environment and Urban Systems, 30*, 143–160.

Miller, H. J., & Wentz, E. A. (2003). Representation and spatial analysis in geographic information systems. *Annals of the Association of American Geographers, 93*(3), 574–594.

Mrozinski, R. D., & Cromley, R. G. (1999). Singly – and doubly – constrained methods of areal interpolation for vector-based GIS. *Transactions in GIS, 3*(3), 285–301.

New York State Department of Health [NYSDOH] (2006). New York state cancer registry. Available from: http://www.health.state.ny.us/statistics/cancer/registry/nyscr.html.

Openshaw, S. (1984). The modifiable areal unit problem. Geo Books, Norwich, United Kingdom.

Pickle, L. W. (2002). Spatial analysis of disease. In C. Beam (Ed.), *Biostatistical applications in cancer research*. Boston: Klewer Academic Publishers.

Plewe, B. (1997). A Representation-oriented taxonomy of gradation. In S. C. Hirtle & A. U. Frank (Eds.), *Spatial information theory: A theoretical basis for GIS*. *Lecture Notes in Computer Science* (1329, pp. 121–135).

Reibel, M., & Bufalino, M. E. (2005). Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems. *Environment and Planning A, 37*, 127–139.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*(3), 351–357.

Rushton, G., Armstrong, M. P., Gittler, J., Greene, B. R., Pavlik, C. E., West, M. M., et al. (2006). Geocoding in cancer research. *American Journal of Preventive Medicine, 30*(2S).

Rushton, G., & Lolonis, P. (1996). Exploratory spatial analysis of birth defect rates in an urban population. *Statistics in Medicine, 15*(7/9), 717–726.

Tobler, W. R. (1979). Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association, 74*(367), 519–530.

Wakefield, J., & Salway, R. (2001). A statistical framework for ecological and aggregate studies. *Journal of the Royal Statistical Society A, 164*, 119–137.

Xie, Y. (1996). The overlaid network algorithms for areal interpolation problem. *Computers, Environment, and Urban Systems, 19*(4), 287–306.