ELSEVIER

Socio-Economic Planning Sciences 42 (2008) 129–149

SOCIO-ECONOMIC PLANNING SCIENCES

www.elsevier.com/locate/seps

Zip codes and spatial analysis: Problems and prospects

Tony H. Grubesic*

Department of Geography, Indiana University, Student Building 120, 701 E. Kirkwood Avenue, Bloomington, IN 47405-7100, USA

Available online 12 January 2007

Abstract

The use of zip codes for spatial, demographic, and socio-economic analysis is growing. As of August 2005, 193 articles were indexed by "zip code" in the Social Sciences Citation Index, while 386 were indexed in PubMed. All of these articles were published since 1989. While the treatment of zip codes as units of analysis varies widely in epidemiology, marketing, geography, and the socio-economic planning sciences, there are a number of common "errors" that could be avoided if analysts retained a better understanding of zip code characteristics. The purpose of this paper is to outline the problems and prospects of utilizing zip codes for spatial analysis. Issues associated with spatial contiguity, data aggregation, and boundary definitions are addressed. Results suggest that, although zip codes are not the most robust spatial units of analysis available, they retain a modest degree of utility for specialized applications. Recommendations for future research regarding zip codes and their use in socio-economic applications are offered.

Keywords: Spatial analysis; Zip code; GIS

1. Introduction

There is little doubt that the US zip code is one of the quirkier "geographies" in the world. How is it possible that a simple numeric code consisting of five numbers, can achieve near iconic status on television (Beverly Hills, 90210), be featured every month in National Geographic Magazine's column "ZipUSA" and manage to retain functionality for the United States Postal Service (USPS)? More interesting, perhaps, is the variety of ways in which zip codes do actually function.

First and foremost, zip codes are best known for being spatially based, hierarchical codes that allow the USPS to route and deliver mail (details of this process will be discussed in the next section). In addition to their functionality for the USPS, zip codes have also become the favored spatial unit of analysis for marketers—particularly those that utilize geodemographic systems [1]. Specifically, over the past 30 years, the geodemographics industry, which consists of companies like CACI, Claritas and others, has created a numerical taxonomy of zip codes. The basic premise of this taxonomy is a simple one; each zip code and its respective classification (e.g. Shotguns and Pickups) represent like-minded consumers, of similar demographic and socioeconomic status.[1] More importantly, this information, when combined with credit reports and buying patterns, creates a tapestry of small, well-defined regions of like-lifestyled households [1].

*Fax: +1 812 855 1661.

E-mail address: tgrubesi@indiana.edu.

[1]According to Claritas [2], Shotguns and Pickups is a legitimate moniker. "[I]t scores near the top of all lifestyles for owning hunting rifles and pickup trucks. These Americans tend to be young, working-class couples with large families—more than half have two or more

0038-0121/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.seps.2006.09.001

Further, because zip codes have such a strong geographical context and can be tied to a specific place, their use in the socioeconomic planning sciences, epidemiology and retailing are increasing. In part, this increased popularity can be attributed to the relative ease of collecting information at the zip code level. For example, it is a common occurrence for one to be asked, "what is your zip code?" when attempting to purchase items at the grocery store or a big-box electronics store (e.g. circuit city). This seemingly innocuous question is actually quite important. In many cases, retailers are trying to identify spatial extent of their store's trade area. The resulting zip code information is then mapped to create the primary, secondary and tertiary trade areas for the store or to track the geographic market share of an outlet. In addition, demographic, socioeconomic and geodemographic data are frequently appended to the zip code data and used to measure store performance in defined market segments [3].

A similar process can be used to determine the primary market area for downtown business districts [4,5], banks [6] or automobile dealers [7]. As mentioned previously, the use of zip codes is also exceedingly popular in the fields of epidemiology and public health. For example, zip codes are used to measure accessibility to health care [8,9], track unmarried teen births [10], establish regional/local variations in radon levels [11], and to map cancer [12].

However, they are unlike the postal codes utilized in Canada and the UK. These codes employ an extremely detailed palette of local delivery units, while the spatial characteristics of five-digit US zip codes, particularly the Census-defined zip code tabulation areas (ZCTAs), are quite varied [13].[2] More to the point, the most alluring, albeit misleading, aspects of using zip codes and ZCTAs for spatial analysis is the notion that they represent relatively "fine" levels of geographic resolution [12].[3] Although this is certainly true for urbanized areas, where ZCTAs tend to be more compact, it is generally *not* true for exurban or rural areas. For example, the average ZCTA in the State of Wyoming covers 552 square miles (1430 km^2). In a more heavily urbanized state, such as New Jersey, the average ZCTA covers 12.8 square miles (33 km^2). Clearly, this is a significant difference—one that must be considered when conducting any type of comparative spatial or statistical analysis.

In addition to the belief that ZCTAs provide better geographic resolution, an apparent benefit of using five-digit ZCTAs for spatial analysis is that they offer an improvement over alternative administrative units, such as counties [9,14]. Again, while this may be true in certain regions, or for a specific application, it is less likely that this assumption is valid in sparsely populated rural locations, where ZCTAs can be extremely large. More importantly, zip codes are not regulated by any population thresholds, and do not conform to any political boundaries (e.g. counties).[4] Certainly, a portion of the zip code's rising popularity as a unit for spatial analysis can be attributed to a "herd mentality". In essence, the widespread adoption of these units for spatial analysis is acceptable simply because everybody else uses them too [9].

A final point worth mentioning involves the use and misuse of zip codes as spatial units in both the public and private sectors in the US. One of the clearest examples in the private sector is the practice of territorial ratemaking, where a measure of risk is assigned to individual geographic units (i.e., zip codes) in an effort to set the premiums for property and causality insurance policies (e.g., car insurance). Unfortunately, because zip codes were never designed to help develop insurance rating territories, such ratemaking practices are prone to error. This exposes insurance companies to additional risk, while consumer premiums are often miscalculated, or inaccurate. In fact, the fate of zip codes and their use territorial ratemaking is currently at issue in California, where policy makers suspect that drivers residing in more rural areas are actually subsidizing the rate premiums for drivers located in more urbanized areas [15]. Further, groups representing Latinos, African–Americans, and urban senior citizens argue that zip code-based insurance rates unfairly discriminate against such segments.

(*footnote continued*)
kids—living in small homes and manufactured housing. Nearly a third of such residents live in mobile homes, more than anywhere else in the nation."

[2]Local delivery units in Canada may consist of a block face (one side of a city street between consecutive intersections), a community mailbox, an apartment building, or a mail delivery route. In many ways, this is similar to the zip + 4 system in the United States.

[3]Key differences between zip codes and ZCTAs will be discussed in the next section.

[4]Census geographies like blocks, block groups, and tracts are hierarchical and nested. That is, tracts are composed of block groups—which are composed of blocks.

Clearly, the use of zip codes and their associated data can offer insights into a variety of spatial, economic, political and health-related processes. However, there are several notable drawbacks in their use for spatial analysis and associated applications. The purpose of this paper is to provide some much-needed background information on zip codes and zip code tabulation areas from a geographic perspective—highlighting their spatial inconsistencies. In addition, several of the more problematic aspects of utilizing zip codes for spatial analysis are outlined, examining issues of spatial contiguity, data aggregation, and boundary definitions.

Although the focus of this paper is relatively US centric, important differences between domestic zip codes and other postal codes/systems (e.g., Canada) are highlighted when appropriate/useful. Moreover, while problems inherent to using zip codes and ZCTAs for spatial and statistical analyses may be unique in the US, discussions of spatial contiguity, data aggregation, boundary definitions, and the modifiable areal unit problem (MAUP) are equally valid when considering alternative areal units both in the US and abroad.

The remainder of this paper is organized as follows. Section 2 provides a brief history of the zip code in the US along with a functional review of its spatial and temporal characteristics. In Section 3, a series of spatial statistical approaches is employed to identify key sensitivities of zip codes and ZCTAs when used as the primary areal unit for analysis. The final section discusses several problems associated with the use of zip codes in the private sector, particularly the insurance industry. Directions for future research and concluding remarks are then provided.

2. A Succinct history of the zip code

The original concept for partitioning addresses into geographic zones for the delivery of mail in the United States was motivated by several factors, including World War II. As thousands of postal employees left to serve in the military, the USPS needed a way to offset the loss in manpower. The initial solution was to create a zoning address system for 124 of the largest post offices in the country [14]. In this relatively simple system, one or two numbers and their respective state identified specific zones, and were used to direct the delivery of mail and increase the overall efficiency of the USPS. For example, the number 10 represented a zone in Kansas City, Missouri. Therefore, an address might read:

John Doe
123 Main St.
Kansas City 10, Missouri

Although this system provided a relatively effective short-term solution during WWII, the postwar growth in the US economy and business related/direct mailings required a more intricate system.[5] Fig. 1 highlights both the overall growth in mail volume and the simultaneous decline in post offices in the US between 1930 and 2002.

By the year 1962, the USPS began to make plans to increase operating efficiencies through the implementation of the zip (zone improvement plan) code system. Further, it was realized that the development of the Interstate Highway system, along with additional commercial airports in the US would establish new focal points for transportation—some of which could be implemented into the realigned and revised USPS system [14]. To expedite these needed changes, an interim strategy for sorting and delivery of mail was developed. Known as the Metro System, the USPS developed a ring of transportation nodes on the outside of the 85 largest US cities. This approach helped deflect mail from heavily traveled and congested central business districts [14].[6] Once these centers were located, the USPS proceeded to allocate codes to both the centers and the post offices they served. By mid-year 1963, a five-digit zone improvement plan code (i.e., zip code) had been assigned to every US household and was used in conjunction with the Metro System nodes.

[5]By 1963, 80% of all mail in the United States was business related.

[6]Eventually, the Metro System was expanded to include 552 sectional centers, each serving between 40 and 150 surrounding post offices in a designated geographic area [14].

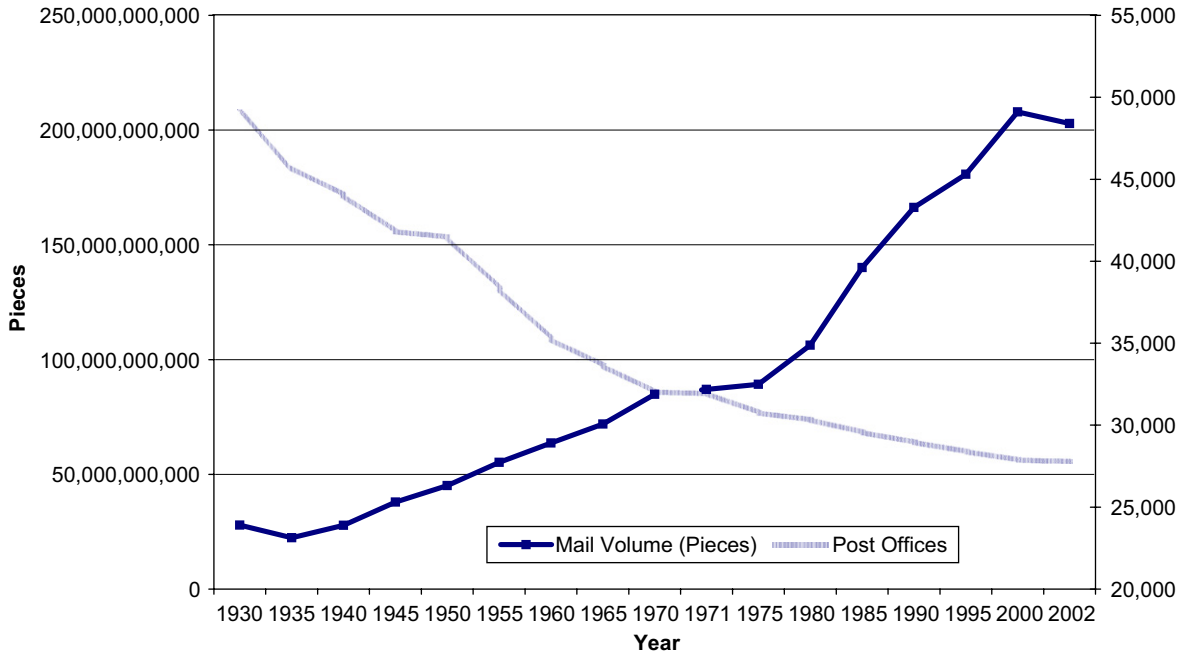

Fig. 1. Growth of the USPS: 1930–2002.

2.1. A functional review

The actual five-digit numeric codes (e.g., 43085) developed in the early 1960s are hierarchical in nature. The first digit corresponds to a multi-state region in the US For example, "7" includes the states of Arkansas (AR), Louisiana (LA), Oklahoma (OK), Texas (TX), while "1" includes the states of Delaware (DE), New York (NY), Pennsylvania (PA) (Fig. 2).

The next two digits refer to one of the 455 remaining Sectional Center Facilities in the US, all of which serve as processing and distribution nodes for USPS packages and mail. The final two digits refer to a specific post office or delivery area within a specific zone. A more recent addition is the zip + 4 code that refers to a five-digit ZIP code plus a four-digit add-on number where the latter identifies a geographic segment within the five-digit delivery zone, such as a city block, office building, individual high-volume receiver of mail, or any other distinct mail unit [16,17]. For example, Fig. 3 displays zip + 4 code locations for a portion of the 43081 zip code near Columbus, OH for the year 2002.

Given the spatial resolution of the zip + 4 codes, it is tempting to conclude that they are similar to the postal delivery units implemented in Great Britain, but this is really not the case. Although zip + 4 codes appear as points, they are not equivalent to Great Britain's Ordnance Survey Address Points. The latter are unique reference points for locating residential, business, and public postal addresses. Specifically, the system consists of 26 million postal addresses, geocoded with an internal quality control marker defining levels of accuracy. For example, the highest quality address point in this system (*positional quality* = 3), is accurate to 0.1 m [18]. Address points thus correspond to the exact physical location of individual households and businesses.

Clearly, this level of geographic resolution is much "finer" than the US zip + 4 delivery points. While zip + 4 codes can correspond to the specific location of a high-volume receiver of mail (e.g., an individual business), zip + 4 points can also represent a much larger geographic area, such as a city block. Interestingly, this level of spatial representation (viz., city blocks) is quite similar to the postal delineations utilized in the Canadian system. For example, Figs. 4a displays the forward sorting areas (FSA) for the municipality of Winnipeg, Manitoba. In Canada, FSAs are roughly equivalent to the five-digit US zip code—with similar variations in size and extent. Fig. 4b illustrates the letter carrier walk (LCW) for the R2W forward sorting area in

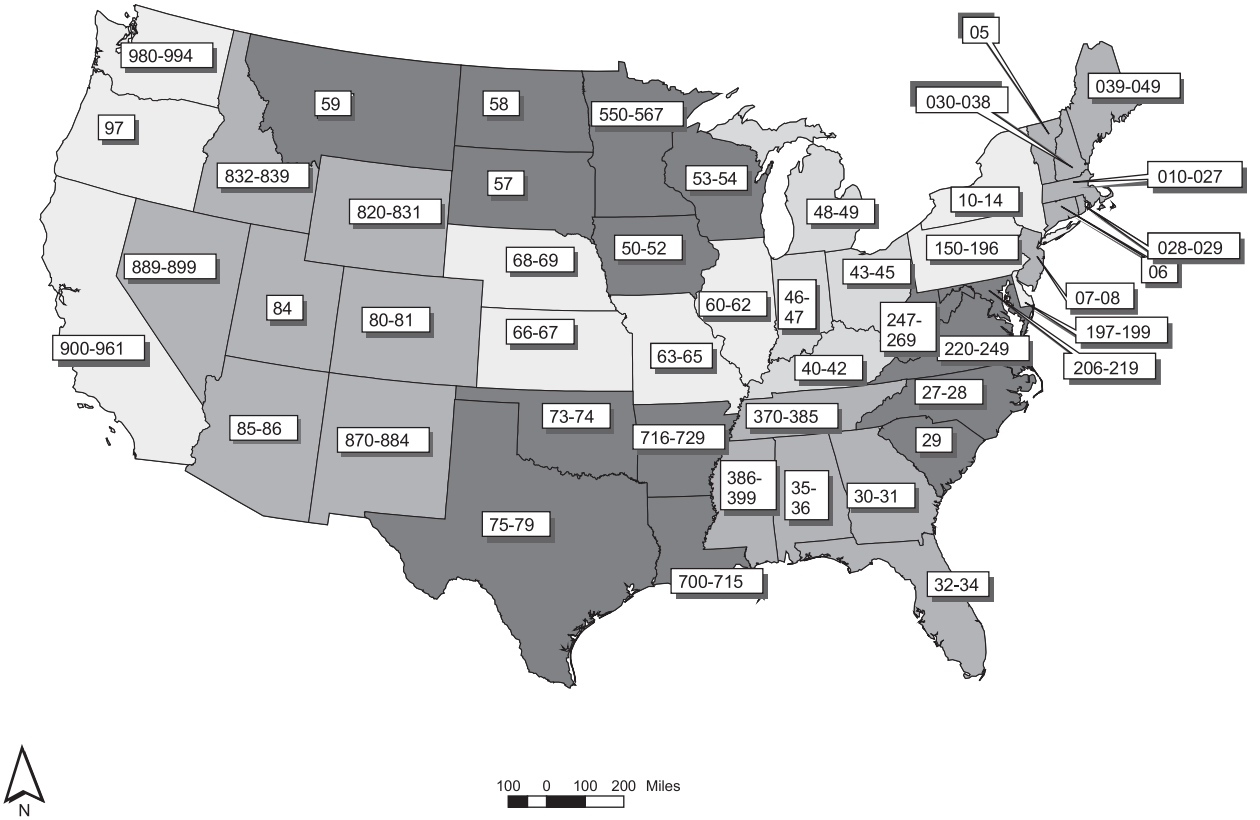

Fig. 2. US zip code zones.

Winnipeg. In effect, the LCWs represent relatively small delivery zones assigned to a specific mail carrier, and typically consist of several blocks within an FSA.

Given this abbreviated comparison of the US zip code system to those implemented in Great Britain and Canada, it would be easy to overlook one of the most important and widely misunderstood aspects of five-digit zip codes in the United States. Five digit zip codes *do not* correspond to discretely bounded geographic areas (i.e., polygons). In reality, zip codes are *linear* features associated with specific roads and addresses. Therefore, while each street in the US is assigned a specific zip code(s), areas without streets or structures corresponding to a USPS-recognized address (e.g., houses or businesses), are not formally assigned a five-digit zip code.[7] For example, the USPS does not actually assign zip codes to all of the Mojave Desert in Southern California because there are vast portions of this region that are unpopulated. This is a subtle, but important fact when considering the cartographic and spatial manifestations of zip codes.

The confusion manifests when street segments and their assigned zip codes are generalized into zones for representational convenience. For example, Fig. 5a illustrates the 43013 zip code, which is located near Columbus, OH, as a polygon for the year 2002. The boundary for this zip code is clearly demarcated by a series of streets. As noted previously, this is to be expected because all streets and their associated addresses in the US are assigned zip codes. However, while the cartographic convenience of representing 43013 as a polygon is appealing, it misrepresents the linear nature of the zip code as displayed in Fig. 5b. Thus, note there are numerous street segments within the polygon that do not belong (four are highlighted), in that they are assigned to both 43031 and 43011 rather than 43013.

Such misrepresentation is not uncommon for zip code databases as they are neither contiguous nor discrete. For analysis purposes, a basic query routine in a geographic information system (GIS) was used, wherein all

[7]To be more specific, each side of a street segment is assigned a five-digit zip code.

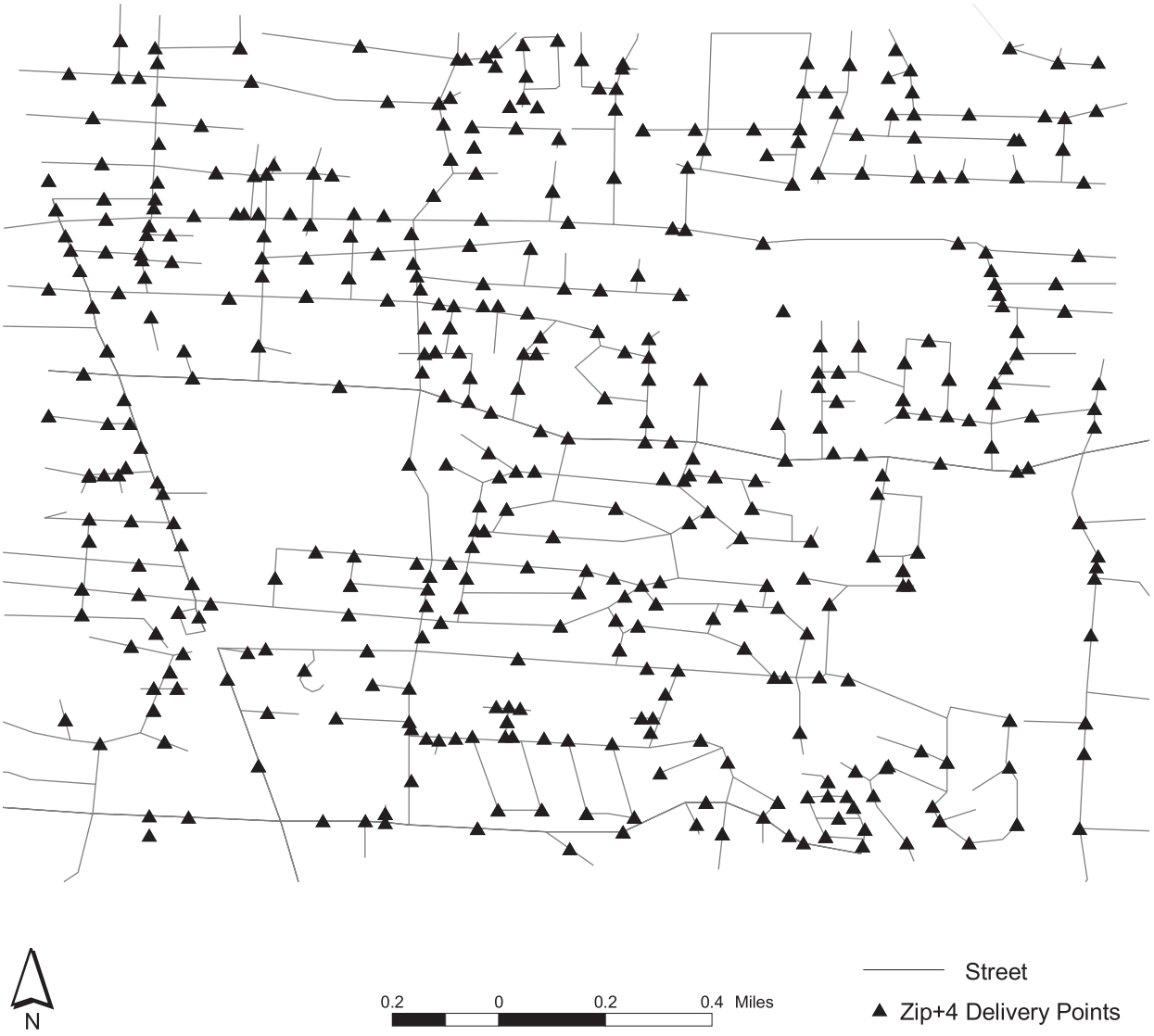

Fig. 3. Zip + 4 delivery points in 43081.

streets segments assigned to the 43013 zip code were selected. These segments were then converted into points and a convex hull (minimum bounding polygon) was generated as an enclosure. This alternative representation of the 43013 zip code is dramatically different from the original boundary, being considerably over-bounded to the north and south, while under-bounded to the east. The question of interest here is: Why?

2.2. Zip code zone interpolation

As noted throughout Section 2.1, zip codes *do not* correspond to discrete polygons. Rather, they are linear features corresponding to mailing addresses and streets serviced by the USPS. However, in many cases, five-digit zip codes can resemble areas because they are composed of clustered street ranges, many of which are contiguous. Nevertheless, as illustrated in Fig. 5, there are substantial differences between the linear "footprint" of a zip code (as defined by its associated street segments), and the zip code polygons frequently used for spatial analysis and mapping applications.

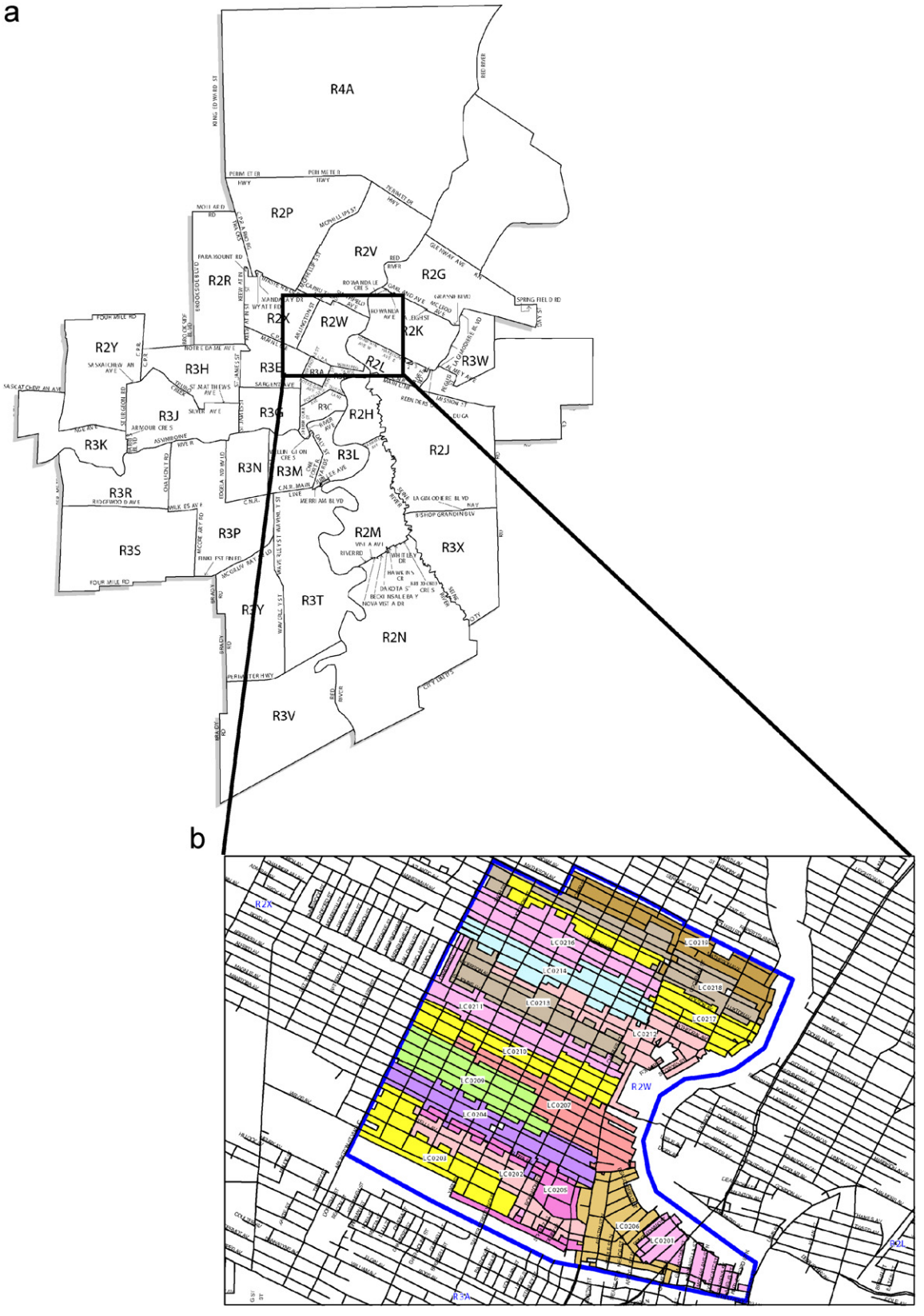

Fig. 4. . Forward sorting areas and letter carrier walks for Winnipeg, Manitoba. (a) Forward sorting areas (FSA). (b) Letter carrier walks (LCW).

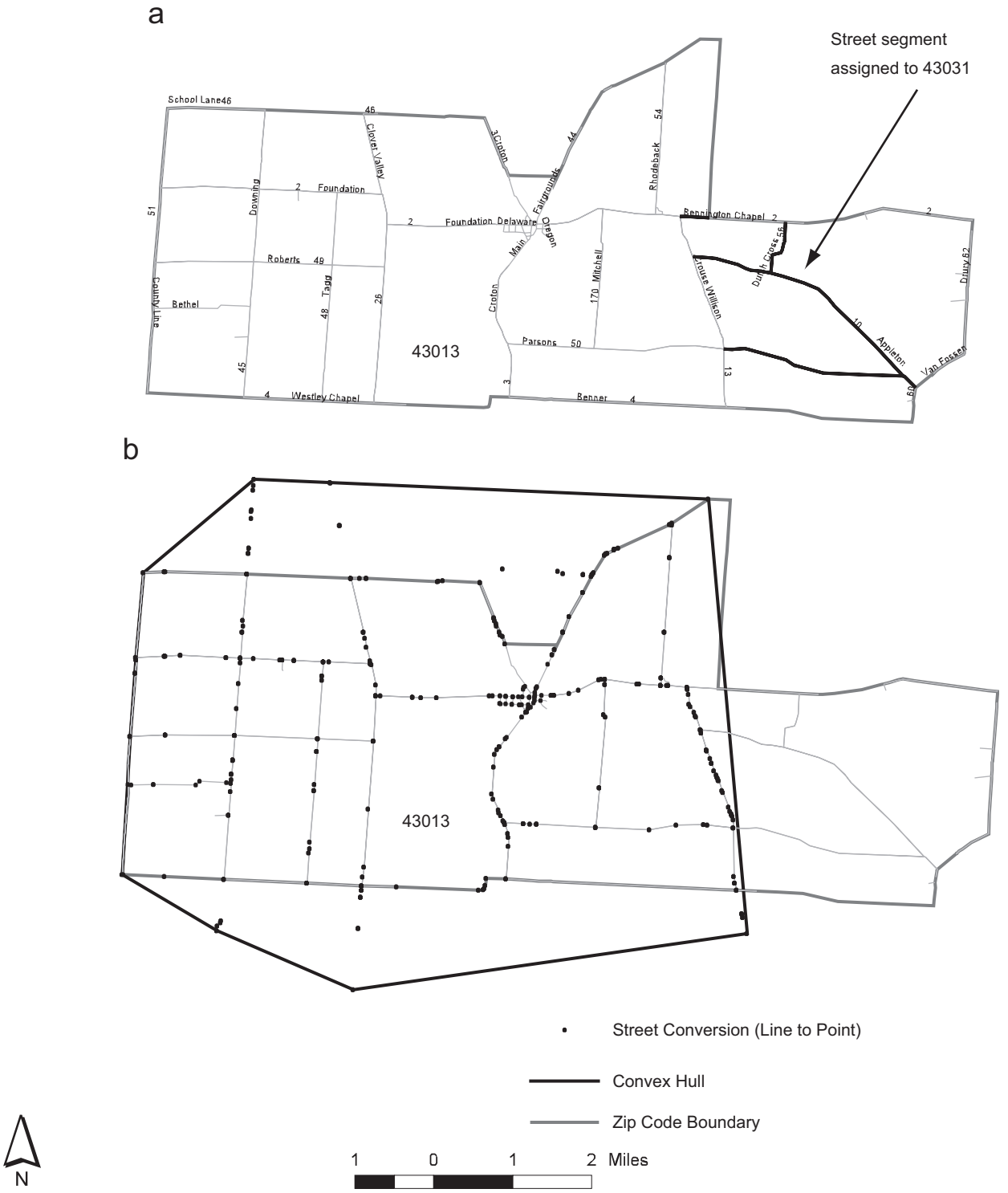

Fig. 5. Zip code boundary problems.

Part of this spatial mismatch problem can be attributed to the interpolation process used to generate the boundaries. Unfortunately, information regarding this process is difficult to find. We thus made several attempts to gather information on the methodologies used by TeleAtlas to generate zip code boundary files,

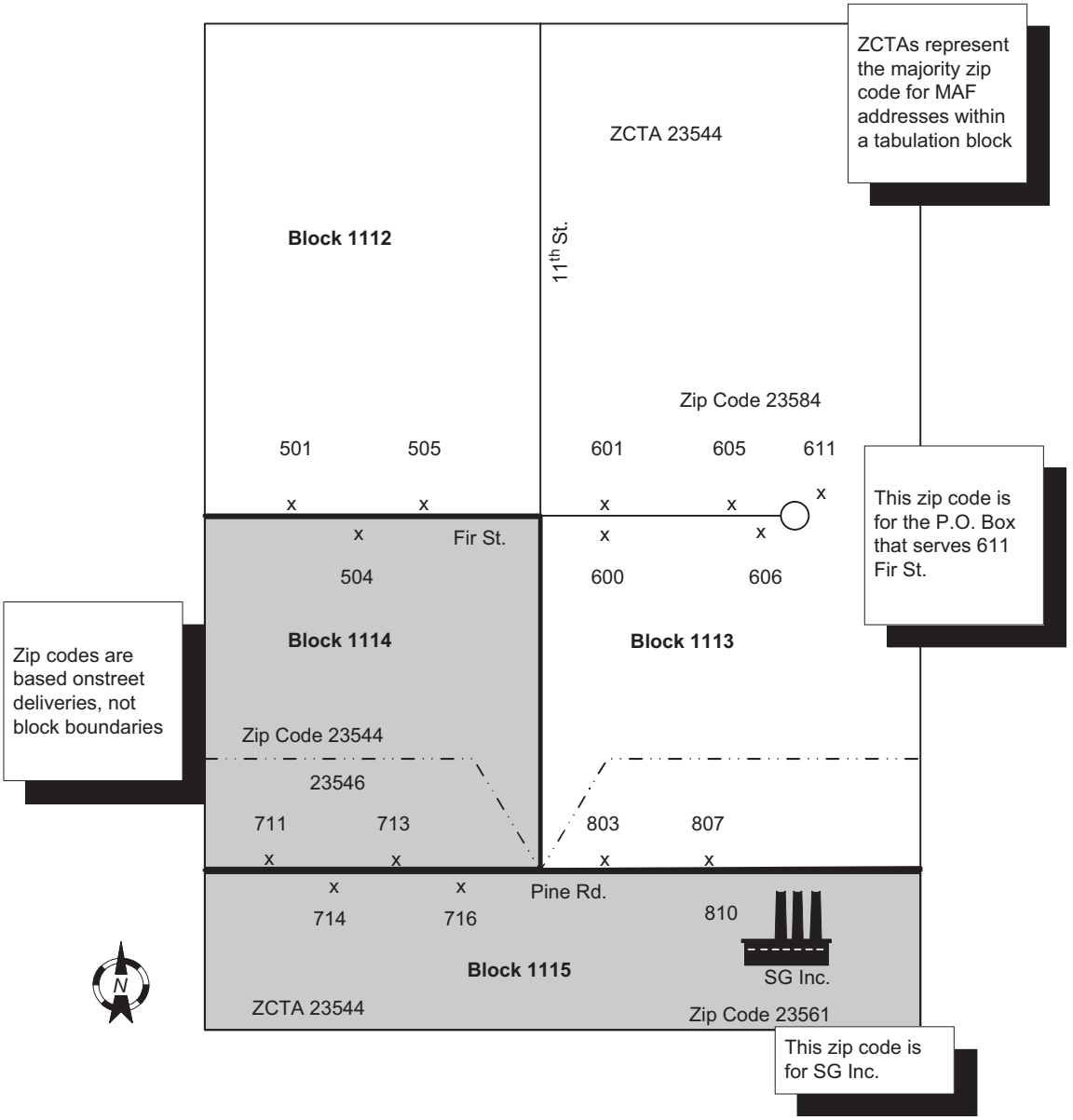

Fig. 6. Example of a ZCTA neighborhood.

but, regrettably, they were unwilling to openly share relevant "proprietary" files.[8] However, there are clues available in associated metadata files provided by second-party data distributors such as Environmental Systems Research Institute (ESRI) and Caliper, which frequently repackage TeleAtlas data in their desktop GIS products.

In brief, the methodology employed by TeleAtlas involves a spatial matching process between USPS and TeleAtlas Dynamap street data. Using USPS mail-stop information for each street segment (e.g., residential and business addresses), and other non-street features, zip code boundaries are manually created, i.e., digitized [19]. In areas where the ability to match street segments with mail-stop information is low, Geographic Data

[8]TeleAtlas, which recently acquired GDT, is one of the largest creators/distributors of five-digit zip code boundary files in the U.S market.

Technology (GDT)/TeleAtlas technicians make telephone inquiries to post offices in the area to determine the predominant zip code [19]. In addition, USPS Zip + 4 state directories are used to differentiate delivery zones and the corresponding zip code boundaries for areas that might not have a clear-cut group of street segments comprising a zip code.

In a strange twist, the US Census Bureau [20] notes that USPS zip codes often follow rear property lines rather than street centerlines. Ironically, the Dynamap streets database provided by TeleAtlas utilizes the latter as the representative entity for street segments. It is therefore not surprising that the TeleAtlas metadata we acquired clearly notes that the zip code delivery zones are "approximated".

2.3. Census zip code tabulation areas

The US Census Bureau [20] provides a more transparent, albeit complicated, interpolation process for five-digit zip code boundaries in their technical documentation for ZCTAs. As noted earlier, these are new statistical entities created by the Census Bureau to represent the generalized boundaries of USPS zip code service areas [20]. Census blocks supply the basic framework for these boundary files.[9]

Utilizing Census block boundaries and USPS zip code data, the Census Bureau compares zip code boundaries against the Census 2000 Master Address File (MAF). If zip code data are available for a block, the appropriate five-digit code is assigned. If such data are not available, the assignment process spatially extends ZCTA coverage from an adjacent area to cover the block in question. As a result, ZCTAs have the following basic characteristics [20]:

(1) They are linked to Census blocks, and every tabulation block has a single ZCTA code.
(2) They cover all tabulation blocks for the 50 states and Puerto Rico.
(3) They may consist of two or more *discontiguous areas*.
(4) A ZCTA code represents a five-digit USPS zip code where possible.
(5) In large undeveloped areas where there are no MAF addresses with five-digit zip codes, the ZCTA code assigned is based on the three-digit zip code.

It is also important to note that the ZCTAs are assigned to *whole* tabulation blocks based on the majority zip code for MAF addresses. As a result, they don't always conform to the established linear features of a zip code. Fig. 6 illustrates a fictitious example of the differences between zip codes and ZCTAs for a selected neighborhood [20].[10]

There are several aspects of this figure worth noting. First, the USPS assigns zip code 23456 to addresses on both sides of Pine Rd, but assigns 23544 to streets north of Pine, including 11th. As noted previously, because the USPS uses rear property lines rather than street centerlines for delineating zip codes, the actual zip code boundary is parallel, but offset from Pine Rd.

Addresses with dedicated P.O. Box zip codes are frequently located within zip codes used for other types of delivery. For example, zip 23584 is a delivery point/P.O. Box serving 611 Fir St., which is located in zip 23544. There are *no* spatial boundaries associated with these delivery points, nor is there a ZCTA. More importantly, the Census does not associate any demographic or socioeconomic information with these zip codes.

Similarly, delivery points that are assigned a unique zip code, such as SG Inc. (23561), can be associated with a significant geographic area (e.g., a factory campus); yet, at the same time, there is no ZCTA associated with such areas.

Finally, because ZCTAs follow Census block boundaries, blocks are assigned only one ZCTA code. Therefore, any block that contains addresses associated with more than one zip code will *not* represent the zip codes appropriately. Further, any statistical tabulation based on the ZCTAs will not necessarily reflect the true

[9]Census blocks are areas bounded on all sides by visible features, such as streets, roads, streams, and railroad tracks, and by invisible extensions of streets and roads. Generally, census blocks are small in area, for example, a block bounded by city streets. However, census blocks in remote areas may be large and irregular and contain many square miles [20].

[10]Although there are numerous examples of these errors in the real-world, the fictitious example encompasses all possible errors in a single, easy-to-read map.

characteristics of actual zip codes and their associated street segments/residences. This is particularly true if zip codes are not spatially contiguous.

This section has sought to identify and discuss many of the challenges associated with defining the geographic extent of zip codes. Most importantly, it was demonstrated that zip codes are not areas, but, rather, linear features associated with a combination of street segments and rear property lines. It was thus noted that, although the representational convenience of defining zip codes by polygons can be alluring, there are a number of potential problems associated with data aggregation, spatial contiguity, and any statistical analysis based on these boundaries.

In this context, the most pressing concern for researchers in the health and socio-economic planning sciences is the misapplication or misinterpretation of data associated with zip codes and their associated polygons. The next section provides several empirical examples highlighting the potential problems with using zip codes for spatial statistical analysis.

3. Data and methods

For comparative purposes, two boundary files were obtained for analysis. The first is a set of zip code polygons for the state of Ohio provided by GDT for the first quarter of 2000.[11] In addition to delineating the spatial extent of each five-digit zip code in Ohio, the GDT codes also provided a small set of associated demographic variables for each polygon (e.g., population and households).

The second boundary file is a set of ZCTAs for the state of Ohio provided by the US Census for the year 2000.[12] These two files do not have any demographic data associated with them; however, because they are based on Census blocks, population data for 2000 are easily aggregated to each ZCTA.

3.1. Boundary file comparison

For all intents and purposes, one would assume that the two boundary files would match relatively well. Although each was generated by a different agency, and with slightly different methods, the zip code data supplied by the USPS *should* provide a common foundation for both. Unfortunately, there are, in fact, a number of clear differences between the two boundary files. First, while the GDT file contains 1220 five-digit zip codes (with no duplicates), the ZCTA file has 1469. A portion of this difference can be attributed to the fact that many ZCTAs include separate zip codes for sparsely populated areas of Ohio. As noted in Section 2.2, there are large undeveloped areas where no MAF addresses exist with five-digit zip codes. In these cases, the Census assigns ZCTAs based on the three-digit zip code (e.g. 430HH). For the state of Ohio, there are 126 cases where a three digit code is used, primarily consisting of lakes, rivers and other waterways.

It is important to note that polygons denoted with a three-digit code are not necessarily unique (i.e., there can be more than one instance of 430HH); however, each entry does exist as a separate entity in the geographic base file. Similarly, additional disparities in zip code counts can be attributed to the presence of multiple polygons for the same five-digit zip code. For example, there are 145 separate instances where a single five-digit zip code is counted more than once in the boundary file because it corresponds to more than one polygon.

This circumstance surely complicates any attempt to make a comparative analysis between studies using incongruous boundary files. It can also have a dramatic impact on the results of aggregating data between smaller sub-units (e.g., blocks) and the zip code boundary polygons. For example, as noted previously, census blocks can vary in size and population. In the state of Ohio for the year 2000, the maximum population for a block was 5256, the minimum was 0, while the average was 41, and the standard deviation 87. While misaggregating data might not have a dramatic impact on the analysis of one or two zip codes, the

[11]As noted previously, GDT is now TeleAtlas. However, at the time these zip code polygons were constructed, the merger between these two companies had not occurred.

[12]ZCTA Boundary files were generated on January 1, 2000. The first quarter zip data from GDT are thus from a time frame directly comparable to that of the Census data.

accumulation of errors for an entire state can be substantial. Not only would the initial statistics be incorrect, but the errors would accumulate and propagate through any subsequent analysis of the data [21,22].

Clearly, these types of spatial mismatches in geographic boundary files have the potential to dramatically impact spatial and statistical analyses. This is particularly important when evaluating comparative statistical work in epidemiology, business and the socio-economic planning sciences—where the need to track spatial and temporal changes in disease rates, market share, or the diffusion of new technologies relies on high-quality geographic base files and some level of spatial consistency.

3.2. Contiguity

One approach for exploring and highlighting the spatial anomalies between zip code boundary files is *contiguity analysis*. Spatial contiguity is one of the most basic characteristics in the geographic landscape. Broadly defined, it refers to the ability to walk from any point in a polygon to any other internal point without leaving it [23,24]. Where two or more polygons are considered, spatial contiguity is the property of sharing a common boundary or vertex. Contiguity analysis provides a unique window into the spatial structure of geographic base files. For example, when comparing the GDT and ZCTA files, it would be extremely informative if one could quantify the differences in connectivity between individual zip codes, or determine the global distribution of connectivity; here, for the state of Ohio. Such information is particularly important when conducting spatial–statistical analysis—because even the slightest variation in spatial structure or representation can impact the results.

Fig. 7 displays three common measures of contiguity: Rook, Bishop and Queen.[13] Rook contiguity is measured by determining a common boundary. In Fig. 7, the central square shares a common boundary with four adjacent squares (top, bottom, left and right). In this instance, one could clearly walk from the central square to any neighboring square without leaving the matrix.

Bishop contiguity is measured somewhat differently. Instead of using common boundaries, Bishop contiguity uses common vertices. That said, Fig. 7 illustrates that it is still possible to walk from the central square to any neighboring square (i.e. diagonals) without leaving the matrix. The most serious concern with using Rook and Bishop contiguity is the potential to miss important spatial relationships.

The Queen's measure of contiguity makes up for this by incorporating both Rook and Bishop relationships into a single measure. In this instance, it is possible to travel from the central square to any square that shares a common boundary or vertex.

These types of contiguity measures can be used to analyze the connectivity of any polygonal features when tiling a geographic area. This is particularly helpful in the analysis of zip codes because of their lack of standardization (geographically or otherwise). In addition, the process of contiguity analysis can help identify the spatial anomalies of a connectivity distribution.

In order to operationalize these measures of zip code contiguity for spatial statistical analysis, one must quantify a spatial weights matrix, $\boldsymbol{W}$ [25]. Elements of $\boldsymbol{W}$ are specified as:

$$w_{ij} = \frac{c_{ij}}{\sum_{j=1}^{n} c_{ij}}, \tag{1}$$

where $c_{ij} = 1$ if i and j share a common boundary or vertex; 0 otherwise. In this context, first order properties include only those vertices and boundaries that are contiguous to the observation (e.g., zip code) in question. Second order properties extend the spatial neighborhood to vertices and boundaries that are also in contact with the first order neighborhood.

One way to summarize the results of the contiguity analysis is through a histogram. In fact, Anselin [25] notes that connectivity histograms are extremely important tools for detecting "strange" features of the connectivity distribution.

Fig. 8 displays the results of both the first- and the second-order Queen's contiguity analyses for the GDT and ZCTA zip code boundaries for the state of Ohio. There are several interesting features worth noting. First,

[13] Distance metrics can also be used to determine contiguity. For a more thorough review, see [26].

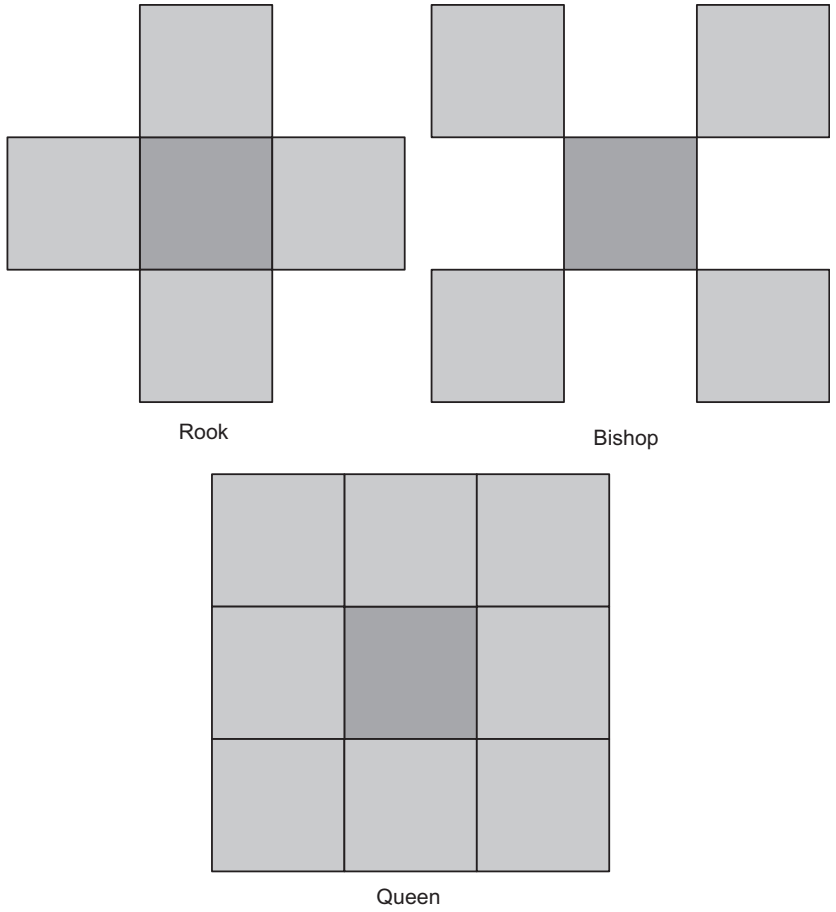

Fig. 7. Types of contiguity.

if contiguity analysis is simply viewed as an exploratory diagnostic, the differences between the GDT and the ZCTA files are obvious. Fig. 8a highlights the marked disparities in connectivity, most notably, that a small percentage of ZCTAs is connected to 18, 19, 20, 21 and, even, 28 other zip codes, whereas the maximum value for GDT is 17.

This somewhat odd result is explained by a simple reexamination of the ZCTA boundary file. The polygon with connectivity between 28 other zip codes is actually based on the three-digit assignment, 435HH. Geographically, this corresponds to the Maumee River in Northwest Ohio. Many of the other features sharing an inordinate number of neighbors are also water features assigned the three-digit zip code. This problem is exacerbated when second-order contiguity is calculated (Fig. 8b). In this instance, there is a small percentage of ZCTAs that are connected to 37, 38 or, as many as 47, other polygons. These results are indicative of dramatic differences between zip code boundary files. However, despite such issues of contiguity mismatch between the zip code boundary files, and the potential for problematic aggregations (outlined in the previous section), it is unclear how the results of a spatial statistical analysis might be impacted.

3.3. Statistical sensitivities

As noted in previous sections, slight variations in the spatial structure of geographic base files have the potential to impact the results of spatial statistical analyses. In order to determine the statistical implications of mismatched zip code boundary files, an exploratory analysis of population was conducted using local

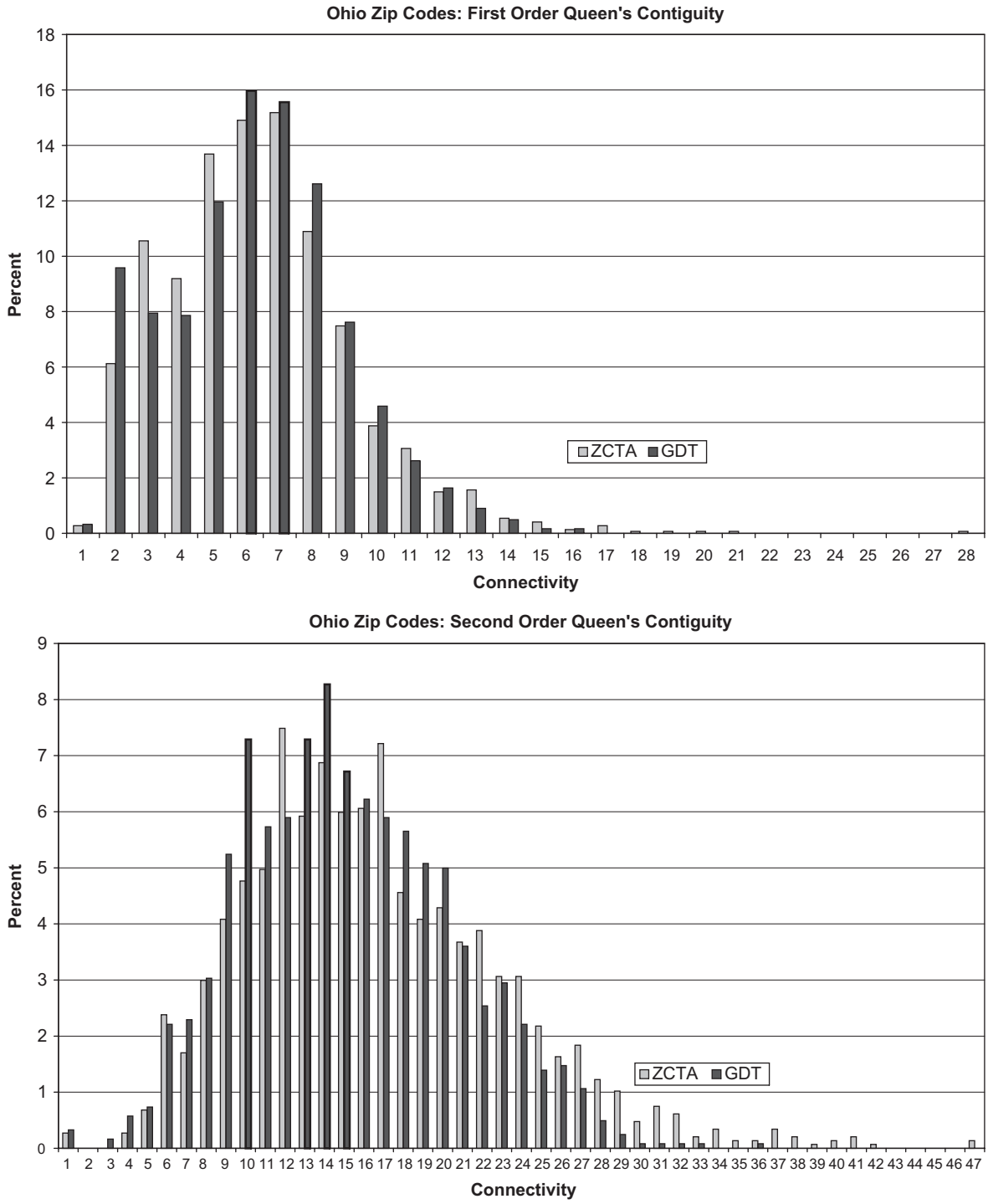

Fig. 8. Contiguity analysis of Ohio zip code boundary files.

measures of spatial autocorrelation [25]. The Local Moran's I [25] is defined as:

$$I_i = z_i \sum_j w_{ij} z_j, \tag{2}$$

where x_i and x_j are observations for locations i and j (with mean μ), $z_i = (x_i - \mu)$, $z_j = (x_j - \mu)$, and w_{ij} the spatial weights matrix with values of 0 or 1.

Recall that the GDT zip codes were provided with several demographic variables, including population for the year 2000. This was not the case for the ZCTAs, however, so population data from Census blocks for the year 2000 were aggregated to each ZCTA. As noted previously, when the GDT and ZCTA files were compared, aggregate population counts for the State of Ohio should have been nearly identical—and, in fact, they were, with a total difference of 0.0027%.

Results of the Local Moran's I based on zip code population and first order Queen's contiguity are displayed in Fig. 9. Each polygon is assigned one of five values:

- *High–high*: zip codes displaying high levels of population that are surrounded by other zip codes with similar values (i.e a hot spot).
- *Low–low*: zip codes displaying low levels of population that are surrounded by other zip codes with similar values (i.e. a cool spot).
- *Low–high*: zip codes displaying low levels of population that are surrounded by zip codes displaying relatively high values.
- *High–low*: zip codes displaying high values of population that are surrounded by zip codes displaying relatively low values.
- *Not significant*: zip codes that were not significant at the 0.05 confidence interval.

Fig. 9a displays results for the ZCTA geographic base file. Perhaps the most troubling aspect of this map is the lack of high-high values in the urban cores of Cleveland, Columbus, Cincinnati and Dayton—the most populous cities and zip codes in Ohio. Similarly, there should be a number of low–low values in the more rural, southeastern portion of the state. These problems are particularly evident when compared to Fig. 9b, which highlights results of the local Moran's I for GDT zip codes. Several identifiable clusters of high and low population (i.e., hot and cool spots, respectively) are visible. The question thus arises: Why does such a dramatic difference exist between these two files—particularly when total population counts for both are nearly identical?

The major problem with the ZCTA file involved use of a poor zip code partitioning system in creating the geographic base file. As noted in Section 3.1, assignment of a three-digit zip code to unpopulated water features, combined with the presence of multiple polygons for the same five-digit zip codes largely disrupted the spatial contiguity of the ZCTA file. This was particularly troubling when demographic or socioeconomic variables were used for testing. For example, water features created observational/spatial "gaps" in demographic data, while multiple polygons for a single zip code reduced population figures by 50% or more, depending on the number of divisions.

Hints that some of these problems existed in the ZCTA layer were highlighted earlier in the boundary file and contiguity analyses, and displayed in Fig. 8. To reiterate, Fig. 9a represents a massive level of error for both aggregation and contiguity. There is thus no real semblance of a statistical pattern that reflects the actual distribution of Ohio's population. In particular, there is almost no corroboration of spatial patterns between the ZCTA and the GDT clusters (Fig. 9b). To make matters worse, the ZCTA boundary file used for the current analysis is the unaltered, unchanged version that is available, and widely disseminated, on the US Census web site.[14]

To further explore problematic aspects of the ZCTA partitioning system, several small adjustments were made to the geographic base file. First, all zip codes that were assigned a three-digit ID by the US Census were removed (i.e., water feature). Second, all five-digit zip codes that consist of multiple polygons were dissolved on a common attribute, ID. Third, block populations were re-aggregated to this newly adjusted ZCTA boundary file. And, finally, a new Local Moran's I was calculated to determine population clusters for all of Ohio.

The results of this process are displayed in Fig. 9c. Not surprisingly, the spatial statistical results are remarkably similar to those generated using the GDT file (Fig. 9b). While several slight differences in the population cluster pattern are evident, the removal of extraneous polygon features in the ZCTA boundary file

[14] http://www.census.gov/geo/www/cob/z52000.html.

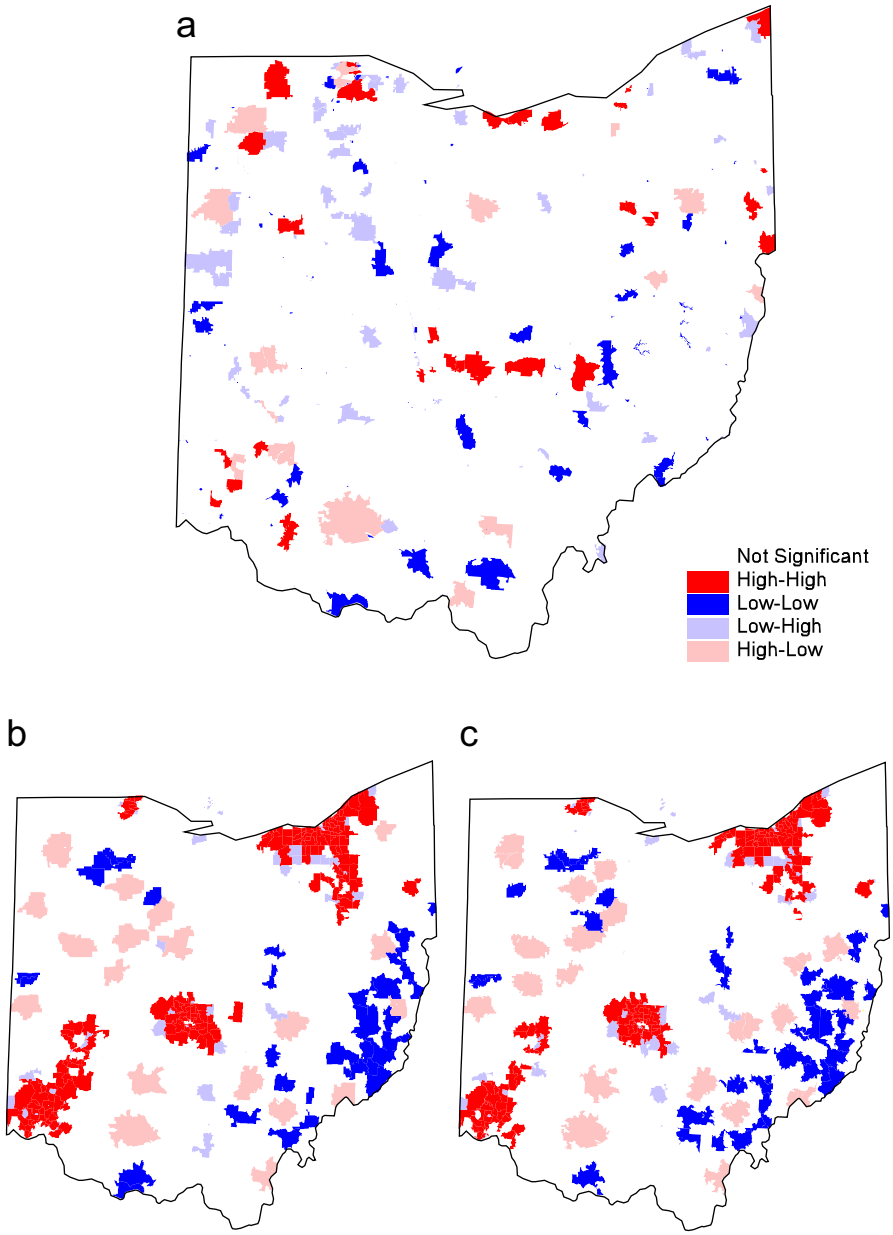

Fig. 9. Local spatial statistical sensitivities. (a) Unadjusted ZCTA population clusters. (b) GDT population clusters. (c) Adjusted ZCTA population clusters.

helped generate a more reasonable geographic base file where spatial contiguity and aggregation issues were less of a problem. This process allowed the true spatial statistical characteristics of Ohio's population distribution to emerge. Moreover, Fig. 10 displays the new, updated contiguity analysis of the adjusted ZCTAs when compared to the GDT data. Needless to say, the results suggest a significantly stronger match between the two geographic base files.

4. Discussion

The results highlighted in Section 3 suggest several significant problems in using zip codes for spatial analysis. First and foremost, five-digit zip codes are not representative of a discrete, bounded region. Instead,

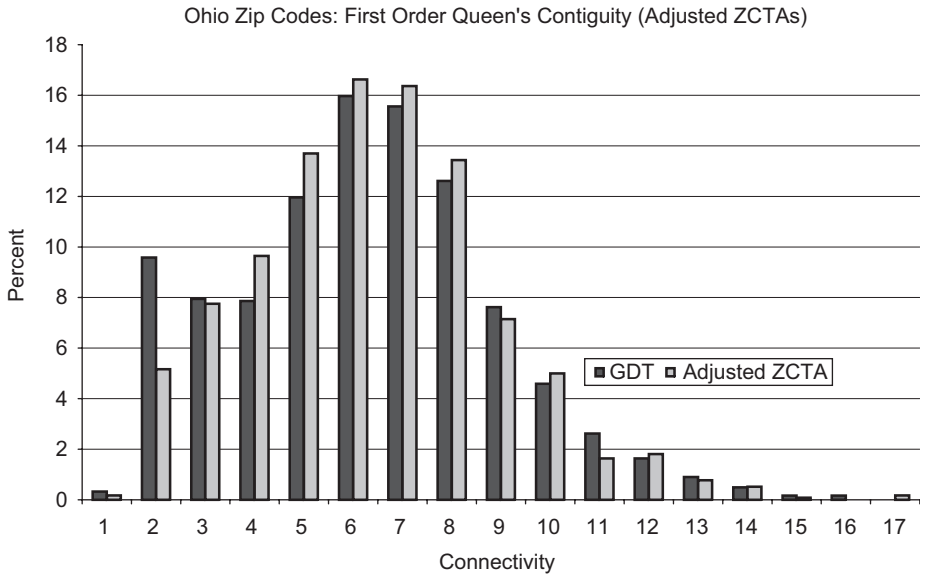

Fig. 10. GDT and adjusted ZCTA contiguity histogram.

they correspond to clustered street ranges, many of which are non-contiguous. Although the cartographic representation of zip codes as polygons is appealing for spatial analysis, particularly in the health and socioeconomic planning sciences, the boundaries are approximations—at best. At their worst, such polygons can represent lakes, rivers, unpopulated areas and other features that are relatively meaningless for social and economic analysis.

Second, there are a number of very subtle, yet insidious problems that can be linked to the interpolation process used for zip code boundary creation. Most notable are issues associated with data aggregation, including the MAUP. Broadly defined, MAUP refers to the use of arbitrary aggregation units with respect to the phenomena under investigation [27–30].

Further to this, the size and shape of the aggregation units will affect the statistical outcomes [27–30]. The results highlighted in Section 3 suggest that MAUP is a significant issue when using five-digit zip code polygons for analysis, where, in many cases, it can be attributed to variations in geographic base file construction and their impact on aggregation. For example, before the ZCTA boundary file was adjusted by dissolving zip code polygons based on a common attribute (ID), population data from Census blocks were aggregated to each of the geographic entities that represented a portion of the zip code. In other words, if zip code 43206 was split into three pieces, not only did it appear in the geographic base file (GBF) three times, population data from the blocks were aggregated to their corresponding spatial entity in the GIS. Therefore, if 43206 had a total population of 10,000—one of the GBF entries might be assigned a value of 5000, and the two others values of 2500 each. This process, in and of itself, is not problematic; however, it misrepresents the aggregate population of 43206 as a single unit. Further, because such bias was not universal in the ZCTA boundary file, population data at the zip code level would not be congruent.

While a third major issue with the use of zip codes for spatial analysis involves spatial contiguity, it is rooted within the data aggregation issues described above. As noted previously, one of the primary problems with the ZCTA boundary file is the presence of multiple spatial entities for a single zip code. In the current study, this multiplicity was somewhat artificial in nature because of the presence of water features that were assigned their own three-digit zip code (e.g., 432HH). In many cases, these areas created an artificial boundary between an otherwise contiguous zip code polygon.

In addition to the aggregation problems this creates, the simple removal of water features is not sufficient to correct the problem—particularly where spatial contiguity is concerned. Although such removal is necessary, it does not address the presence of multiple polygons associated with the same zip code. Many instances thus arose where zip codes were allowed to be their own neighbors when spatial weights matrices were constructed.

As illustrated in Fig. 10, this had a profound impact on tests for local indicators of spatial association—masking many of the known patterns of population clusters in Ohio. Fortunately, a relatively simple "dissolve" routine in a GIS can help correct this problem. As noted previously, the dissolve function removes the artificial boundaries in ZCTA polygons generated by the Census. The resulting geographies represent each zip code and its associated population with a single polygon, rather than multiple polygons.

4.1. Implications for business and industry

So, where does all this leave zip code boundary files and their utility for spatial analysis? There is little doubt that zip codes will continue to be used extensively. Data collection efforts at the zip code level are relatively easy to perform, and have become widely accepted in both the public and private sectors. Further, now that the Census Bureau has implemented zip code tabulation areas, these data are more accessible than ever. However, the results of this paper suggest that there are a number of significant problems associated with zip code boundary generation and cartographic representation. Not surprisingly, the problems identified within zip code boundaries have a direct impact on both business and industry.

For example, with respect to the insurance industry, California is currently considering massive, industry-wide changes in the way that rates are set for personal insurance lines [15]. In particular, the use of zip codes for territorial ratemaking is under scrutiny. This reassessment of property and causality lines (e.g., automobile, homeowners) has roots in a somewhat controversial statewide proposition, 103, which passed in 1988. It requires insurers to discontinue the practice of basing auto insurance rates primarily upon where one lives. Instead, rates are to be set as a function of three mandatory factors: (1) driving record, (2) number of miles driven, and (3) years of driving experience [31].

A major problem in California is that the policies set forth by Proposition 103 were never really enforced. As a result, analysts suspect that insurance premiums are largely based on zip code locations [15]. This has created serious financial consequences for thousands of drivers. In one example, data collected by the Consumers Union (1998) suggested that two young males, with identical driving records and experience, were paying dramatically different premiums [32]. In San Luis Obispo, the premium for the first driver was $1706, while in Los Angeles, the premium for the second driver was $7844 [32]. This is a difference of nearly 360%—based simply on location.

Further exacerbating this issue in California is the notion that zip code-based ratemaking discriminates against minorities and those with lower socioeconomic standing. For example, a recent insurance industry study showed that rates are likely to rise in 52 of 58 California counties in 2006 [33]. While the geographic coverage of this increase is relatively unbiased (in terms of the number and location of counties selected), the actual rate increases are not evenly distributed.

For example, premium increases in Santa Clara County are projected to be six percent or less, while increases in Imperial County, which is largely low-income, rural and Hispanic are projected to exceed 35% [15]. More importantly, data on insurance premiums that were tabulated by the Consumers Union [33] at the zip code level confirm such biases for three of the largest insurance companies that write policies in California. The biases are particularly pronounced when comparing zip codes of different demographic compositions (see Fig. 11).

Locational biases in premiums have actually motivated residents in Palo Alto, CA to petition the USPS for zip code realignment [34]. The 94303 zip code is shared by the primarily white and wealthy community of Palo Alto and its Hispanic, economically challenged neighbor, East Palo Alto. Not surprisingly, the request was denied. According to the USPS, "it is far too costly to adjust postal delivery boundaries for reasons not related to efficient process and delivery of mail" [34].

While issues of zip code boundaries and territorial ratemaking in the insurance industry are troubling, there are problems in other sectors as well. For example, the telecommunications industry is also dealing with problems associated with the use of zip codes for data collection and public policy evaluation. The Federal Communications Commission (FCC) tracks broadband competition and deployment in the United States using provider data aggregated to the zip code. In theory, this process was implemented to help ensure the equitable rollout of broadband services. The FCC's major concern is that without regulation, telecommunication companies would only provide broadband services to affluent households (where the

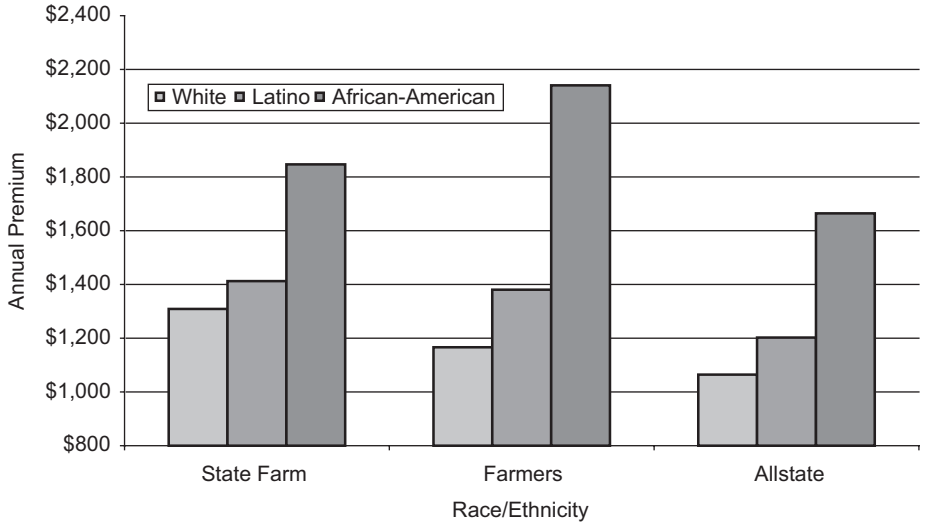

Fig. 11. 2002 premium differentials and zip code demographic composition: female, good driver, licensed 22 years, no accidents or violations, standard coverage. *White = Non-hispanic whites > 50% of total population (by zip code). Latino = Hispanics > 50% of total population (by zip code). African–American = Blacks > 50% of total population (by zip code). State Farm's 2002 Class Plan No. 02-13401. Famers' 2002 Class Plan No. 02-11204. Allstate's 2002 Class Plan No. 02-37682. *Source*: Consumers Union of the United States, Inc. (November 2005).

return on investment for infrastructure upgrades would be high), and ignore portions of their service area with higher levels of minority, or economically challenged residents [36].

The major problem with the FCCs data collection effort is the use of zip codes for aggregation. Instead of providing specifics on the exact neighborhoods where broadband is available, facilities-based telecommunication providers are simply required to report the zip codes in which they operate at least 250 terrestrial broadband lines or wireless broadband channels. Each resulting database consists of a list of zip codes that have at least one broadband service subscriber [36,37]. Unfortunately, where the FCC is concerned, these data are treated as a binary indicator of service—either the zip code is served, or it is not.

Clearly, the presence of a high-speed broadband provider in a zip code does not guarantee ubiquitous access. In many cases, the technical limitations of broadband platforms such as fixed wireless or digital subscriber lines curtail the range of service [39]. As a result, only portions of a zip code listed as having broadband access may actually receive service. This is compounded by the fact that zip codes are extremely variable in size and extent. Thus, while urban zip codes are geographically compact, rural zip codes can consist of hundreds of square miles. Such limitations, and the overall lack of spatial resolution presented by zip code geographies, severely limit the policy evaluation process and the ability to accurately track telecommunication industry investment and services.

4.2. Conclusions and implications for future

Clearly, the use of zip codes in business and industry can be problematic. For example, our research has shown that zip codes are not appropriate building blocks for territories, nor are they always appropriate for evaluating the geographic distribution of services, regardless of sector, given constant realignments by the USPS, and the fact that they do not represent discretely bounded spatial units.

In the near term, this suggests several important research frontiers concerning the zip code and its associated use in both the public and private sectors. First, there is a need to identify or develop a more appropriate geographic unit for territory definition and data collection. Alternative geographic units, such as the Census block group (BG), can be an appealing option. Block groups contain detailed demographic and socioeconomic data, while the Census Bureau attempts to maintain the optimal size of 1500 people within each BG. Further, these units more stable geographically, even when the Census Bureau opts to split a BG

between decennial surveys. The split block group will always reflect its previous geographic definition. This is something that will almost *never* occur with zip codes.

A second research frontier dealing with zip codes is the development of methods to analyze longitudinal data collected at the zip code level in an unbiased manner. As noted previously, this is a particularly challenging issue because of the dynamic spatial and temporal nature of zip codes. With every change made by the USPS, existing zip codes deviate from the previous alignment—introducing potential error into comparative, longitudinal and spatial analyses. Clearly, the need for methods to address such issues is particularly acute in those public and private sectors where zip codes serve as the primary unit for data collection [3,8,9,12,35–39].

Simply put, analysts should retain a deeper understanding of how zip code polygons are generated, and their potential uses for spatial analysis. More importantly, because zip codes are continuously updated and changed by the USPS, longitudinal studies and other comparative work is extremely susceptible to many of the errors highlighted in this paper. Obviously, zip codes are not the only choice of areal units for statistical and spatial analysis. There are many alternative options, including Census block groups. In many respects, these units provide a more stable foundation for analysis because the US Census utilizes both a nested geographic partitioning method and population constraints for each unit.

That said, information collected at the zip code level is frequently unique, non-Census related, and critical to evaluating socio-economic or epidemiological trends. An increased level of awareness concerning the spatial limitations of zip codes and their use for statistical analysis is thus seen as crucial to leverage the true value of these distinctive data.

References

[1] Phillips DJ, Curry MR. Privacy and the Phenetic Urge: Changing Spatiality of Local Practice; 2005. URL: 〈http://www.geog.ucla.edu/~curry/Phillips-Curry-GDP.pdf〉.
[2] Claritas. PRIZM NE; 2005. URL: 〈http://www.claritas.com/claritas/Default.jsp?ci=3&si=4&pn=prizmne#51〉.
[3] Kumar V, Karande K. The effect of retail store environment on retailer performance. Journal of Business Research 2000;49:167–81.
[4] Kures M, Pinkovitz B, Ryan B. Downtown and Business District Market Analysis; 2005. URL: 〈http://www.uwex.edu/ces/cced/dma/5.html〉.
[5] New Richmond Chamber of Commerce [NRCC]. Conclusions and Recommendations; 2004. URL: 〈http://www.newrichmond-chamber.com/report_section_9.pdf〉.
[6] Kaynak E, Harcar TD. American consumers' attitudes towards commercial banks: a comparison of local and national bank customers by use of geodemographic segmentation. International Journal of Bank Marketing 2005;23(1):73–89.
[7] Washington FS. Does switch to census tracts make sense? Ward's Dealer Business; 2003. URL: 〈http://wdb.wardsauto.com/ar/auto_switch_census_tracts_2/〉.
[8] Fortney J, Rost K, Warren J. Comparing alternative methods of measuring geographic access to health services. Health Services and Outcomes Research Methodology 2000;1(2):173–84.
[9] Luo W, Wang F, Douglass C. Temporal changes of access to primary health care in Illinois (1990–2000) and policy implications. Journal of Medical Systems 2004;28(3):287–99.
[10] Blake BJ, Bentov L. Geographical mapping of unmarried teen births and selected sociodemographic variables. Public Health Nursing 2001;18(1):33–9.
[11] Steck DJ, Baynes SA, Noack AP. Regional and local variation of indoor radon and radon source potentials. Environment International 1996;22(1):S729–37.
[12] Johnson GD. Small area mapping of prostate cancer incidence in New York State (USA) using fully Bayesian hierarchical modelling. International Journal of Health Geographics 2004;3(29).
[13] Krieger N, Waterman P, Chen JT, Soobader M-J, Subramanian SV, Carson R. 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 2002;92(7):1100–2.
[14] United States Postal Service [USPS]. The United States Postal Service: An American History 1775–2002; 2003. URL: 〈http://www.usps.com/cpim/ftp/pubs/pub100/〉.
[15] Kuruvila MC. Zip code used to set car insurance rates debated. The Mercury News. February 24th, 2006. URL: 〈http://www.mercurynews.com/mld/mercurynews/news/local/13956813.htm〉.
[16] Curry MR. Toward a geography of a world without maps: Lessons from Ptolemy and postal codes. Annals of the Association of American Geographers 2005;95(3):680–91.
[17] Carrier Routes. Information on Postal Carrier Routes, Saturation Mailing and Carrier Route Maps; 2005. URL: 〈http://www.carrierroutes.com〉.

[18] Ordinance Survey [OS]. Address Point User Guide; 2005. URL: 〈http://www.ordnancesurvey.co.uk/oswebsite/products/addresspoint/pdf/apuserguide.pdf〉.
[19] Geographic Data Technology [GDT]. Ohio Zip Code Areas; 2000. URL: 〈http://www.co.warren.oh.us/warrengis/metadata/ohzip.htm〉.
[20] US Census Bureau. Census 2000 ZCTAs: Zip Code Tabulation Areas Technical Documentation; 2005. URL: 〈http://www.census.gov/geo/ZCTA/zcta_tech_doc.pdf〉.
[21] Alonso W. Predicting best with imperfect data. Journal of the American Institute of Planners 1968;43:248–55.
[22] Simon SD, LeSage JP. Assessing the accuracy of ANOVA calculations in statistical software. Computational Statistics and Data Analysis 1989;8:325–32.
[23] Cova TJ, Church RL. Contiguity Constraints for Single-Region Site Search Problems. 2000; 32(4): 306–329.
[24] Wu X, Murray AT. Assessing landscape contiguity in reserve design. Systems Analysis in Forest Resources. In: Bevers M, Barrett TM, editors. Proceedings RMRS-P-000. Ogden, UT: US Department of Agriculture, Forest Service, Rocky Mountain Research Station; 2004.
[25] Anselin L. Local Indicators of Spatial Association—LISA. Geographical Analysis 1995;27(2):93–115.
[26] Wu X. Quantification and optimization of spatial contiguity in land use planning. Doctoral Dissertation. Department of Geography, The Ohio State University; 2005.
[27] Gehlke C, Biehl K. Certain effects of grouping upon the size of the correlation coefficient in census tract material. Journal of American Statistical Association 1934;29:169–70.
[28] Openshaw S, Taylor P. A million or so correlation coefficients. In: Wrigley N, editor. Statistical methods in the spatial sciences. London: Pion; 1979. p. 127–44.
[29] Openshaw S. The modifiable areal unit problem. Concepts and techniques in modern geography, vol. 38. Norwich, UK: Geo Books; 1983.
[30] Unwin DJ. GIS, spatial analysis and spatial statistics. Progress in Human Geography 1996;20(4):441–540.
[31] California Department of Insurance. Proposition 103 Fact Sheet; 2006. URL: 〈http://www.insurance.ca.gov/0200-industry/0500-legal-info/0500-gen-legal-info/prop-103-fact-sheet.cfm〉.
[32] Consumers Union. Cities of Los Angeles, Oakland and San Francisco Join Civil Rights and Consumer Groups in Lawsuits Over Zip Code-Based Auto Insurance Rates; 1998. URL: 〈http://www.consumersunion.org/finance/zipwc398.htm〉.
[33] Consumers Union. California insurers charge as much as $974 per year more to good drivers living in predominately Black, Latino zip codes; 2005. URL: 〈http://www.consumersunion.org/pub/core_financial_services/002991.html〉.
[34] Noguchi S. Drivers trying to escape higher insurance rates give up on effort to shed E. Palo Alto's Zip code. The Mercury News 2006 URL: 〈http://www.mercurynews.com/mld/mercurynews/news/local/13876384.htm〉.
[35] Botts H, Kucera JL. Personal lines territories: It's time to take another look. URL: 〈http://www.pinnacleactuaries.com/pages/publications/files/PersonalLinesTerritories.pdf〉.
[36] Grubesic TH. The geodemographic correlates of broadband access and availability in the United States: a longitudinal analysis. Telematics and Informatics 2004;21(4):335–58.
[37] Flamm K, Chaudhuri A. An analysis of the determinants of broadband access. Telecommunications Policy Research Conference, Washington, DC; September 2005.
[38] Priger JE. The supply side of side of the digital divide: Is there equal availability in the broadband Internet access market? Economic Inquiry 2004;41(2):346–63.
[39] Grubesic TH, Horner MW. Deconstructing the divide: extending broadband services to the periphery. Environment and Planning B 2006;33(5):685–704.

Tony H. Grubesic is currently an assistant professor of Geography at Indiana University and an adjunct research fellow in the Center for Urban and Regional Analysis (CURA) at The Ohio State University. His research and teaching interests are in geographic information science, technological hazards, telecommunication policy, and regional development. Dr. Grubesic has published on a range of technical and application oriented topics in journals such as the *Annals of the Association of American Geographers*, *Environment and Planning B*, *Growth and Change*, *Social Science and Medicine*, *Papers in Regional Science*, *Annals of Regional Science*, *Transportation*, *Telematics and Informatics*, *Journal of Quantitative Criminology*, *International Journal of Industrial Engineering*, and *Telecommunications Policy*. He obtained a B.A. in Political Science from Willamette University, a B.S. in Geography from the University of Wisconsin-Whitewater, a M.A. in Geography from the University of Akron, and a Ph.D. in Geography from the Ohio State University.