



# Linking Disease Outcomes to Environmental Risks: The Effects of Changing Spatial Scale

Chetan Tiwari, David Sterling, and Leslie Allsopp

## Introduction

Geographic information systems (GIS) enable the assessment of environmental risks and their potential impacts on human health outcomes by providing a mechanism to overlay maps of exposure and disease outcomes. While the ability to overlay various geographic datasets is a common function provided by most GIS software packages, the ability to derive meaningful associations between these layers is limited by data quality, issues pertaining to the accuracy of exposure assessment, and problems of representing the intensity of disease outcomes over space and time. In a recent review of the role of geographic information science (GISc) in the analysis of health and place, Mennis and Yoo (2018) identify major challenges and opportunities including problems associated with scale of analysis in health research. In this context, they argue that most GIS-based health research in this area has focused on problems of sparse or missing data and emphasizes the need for more research to understand the implications

of resolution and spatial and temporal sampling frameworks on assessments of individual-level environmental exposures (Mennis and Yoo 2018). The problem of deriving associations between layers of geographic data with inconsistent scales is of particular concern in the context of big data where personalized health information through electronic health records and individual measures of exposure via low-cost sensors are becoming more common. Limitations on the use of such data due to privacy concerns or inconsistent quality often lead to the production and dissemination of datasets aggregated to different levels of spatial resolution. While GIS software can be used to combine such layers of geographic data, it is critical to note that inconsistent scales and misaligned boundaries resulting from the use of disparate spatial units will likely result in incorrect and/or misleading conclusions.

The problems of changing spatial scales and misaligned geographic boundaries are well documented in discussions of spatial uncertainty. Spatial uncertainty is broadly defined as the problem of identifying and quantifying error in the geographic location of objects. Such error may result in biased interpretations of the true relationships between the location of objects in space and surrounding contextual or environmental factors. There are two issues associated with spatial uncertainty – the change of support problem (CoSP) and the uncertain geographic context problem (UGCoP). CoSP is concerned with the problem of drawing inferences about observations at a spatial scale that is different from the scale at which those observations have occurred. Kwan (2012) defines the uncertain geographic context problem (UGCoP) as the problem of identifying the effects of spatial displacement between the geographic definitions of contextual units and the “true causally relevant” context. The problem presented by aggregation may be considered as a subset of the CoSP and is similar to the well-known modifiable areal unit problem (MAUP) which states the patterns observed on

---

C. Tiwari (✉)  
Departments of Geosciences & Computer Science, Georgia State University, Atlanta, GA, USA

Center for Disaster Informatics & Computational Epidemiology, Georgia State University, Atlanta, GA, USA  
e-mail: [ctiwari@gsu.edu](mailto:ctiwari@gsu.edu)

D. Sterling  
Department of Biostatistics and Epidemiology, School of Public Health, University of North Texas Health Science Center, Fort Worth, TX, USA

School of Population and Public Health, University of British Columbia, Vancouver, BC, Canada

L. Allsopp  
Department of Biostatistics and Epidemiology, School of Public Health, University of North Texas Health Science Center, Fort Worth, TX, USA  
e-mail: [Leslie.allsopp@unthsc.edu](mailto:Leslie.allsopp@unthsc.edu)

a map; therefore, any inferences drawn are likely to change if the shape and scale of the map unit change. In many population-based studies, the UGCoP arises from the need to use area-based measures of statistical observations, such as census tracts, to reconstruct a fundamentally vague definition of space, such as the risk of disease that arises from some kind of exposure to environmental risk. See Kwan (2018), Fisher et al. (2018), and Griffith (2018) for an overview of spatial analytic approaches to identify, measure, and address this problem.

The utility and pitfalls of GIS for exposure assessment are well described in Nuckols et al. (2004). They describe several studies that use GIS in various ways to estimate exposures to a variety of environmental risks. Some examples of studies cited in their work include an assessment of associations between residential proximity to landfill sites and adverse birth outcomes (Elliott et al. 2001), examination of possible neurobehavioral effects of exposure to trichloroethylene using a simulation model (MODFLOW) (Reif et al. 2003), and a population-based study to evaluate lung cancer outcomes to urban air pollution using a combination of dispersion modeling and geostatistical techniques (Bellander et al. 2001; Nyberg et al. 2000). Other studies that use GIS to produce fine-scale assessments of environmental risk include an assessment of the health risks posed by urban heat islands detected using remote sensing methods including airborne or satellite platforms (Tomlinson et al. 2011), an overview of techniques to produce fine-scale estimates of fine particulate matter using remotely sensed data and geostatistical approaches (Al-Hamdan et al. 2009, 2014), estimation of spatio-temporal variations in hot weather conditions of Hong Kong using statistical techniques (Shi et al. 2019), and mapping of exposures to particulate matter using remotely sensed data and geostatistical modeling techniques (Leelasakultum and Kim Oanh 2017). In all cases, assessments of the risk of environmental exposure are estimated for various levels of spatial resolution that may not necessarily conform to the scale at which disease data are commonly made available.

Under current practice, makers of disease maps select census or other administrative units for which both disease and demographic data are available. GIS software are then used to compute and visualize disease *rates* among the populations contained within those administrative units. The choice of administrative unit influences the resolution and statistical reliability of observed disease rates. Patterns of disease rates displayed on maps produced using such spatial units represent a tradeoff between spatial resolution and statistical reliability. Maps with high degrees of spatial resolution generally exhibit poor statistical reliability as the population support – numbers of persons at risk – used in calculating each rate is often small. As more sources of georeferenced health and demographic data become available, so does the opportunity to control the numbers of people at risk

and the geographic size of the areas mapped. In geographical circles, the spatial resolution of a map refers to the size or area used to measure the spatial variation of a disease rate. If the areas mapped were of equal size, the map would be said to have the same geographic resolution across the map. Since most maps use administrative areas as the spatial units to map, the common spatial resolution of a map is the average size of the administrative areas used. A second meaning of spatial resolution is when the minimum size mapped is the smallest size for which the common geography between disease data and demographic data realizes a fixed level of statistical reliability. The goal of the actual spatial resolution achieved by the map is not, therefore, a common spatial size, but, instead, a minimum sized spatial unit at any location on the map that realizes the statistical reliability desired by the mapmaker. A third meaning of spatial resolution has arisen more recently in the era of digital maps when the smallest spatial unit on the map is a pixel. If the map is constructed so that pixel values change according to some function of relative location, then the earth size corresponding to one pixel is the geographic resolution of the map in question. In this chapter, we advocate for a disease mapping approach that focuses on a deliberate choice of geographic resolution and statistical reliability. To this end, we demonstrate how the two characteristics can be controlled using a simulated dataset on disease outcomes that are influenced by four randomly selected locations of environmental exposure.

---

### Relevance of Disease Mapping for Assessing Public Health Impacts

Disease mapping refers to the process of constructing a map that shows the spatial distribution of disease within a specific geographic region. Disease maps improve public health decision-making by providing a mechanism to identify geographic areas that are in most need of interventions or resources (Bertollini and Martuzzi 1999; Moore and Carpenter 1999; Ricketts 2003). They can help answer such as the following questions: What populations are at risk? Where they are located? What are the underlying conditions in those areas? The common spatial context enables researchers and public health practitioners to link various geographic layers of data to explore associations between a multitude of complex processes that include various combinations of social, cultural, and environmental determinants. In 1850, John Snow created the first disease map of cholera distribution in London and initially showed the importance of cartographic representation of disease in serving public health (Koch 2004). Snow's point map shown in Fig. 1 describes the spatial patterns of cholera deaths and its geographical association with other features on the landscape, including the broad street pump, which was subsequently identified as serving the

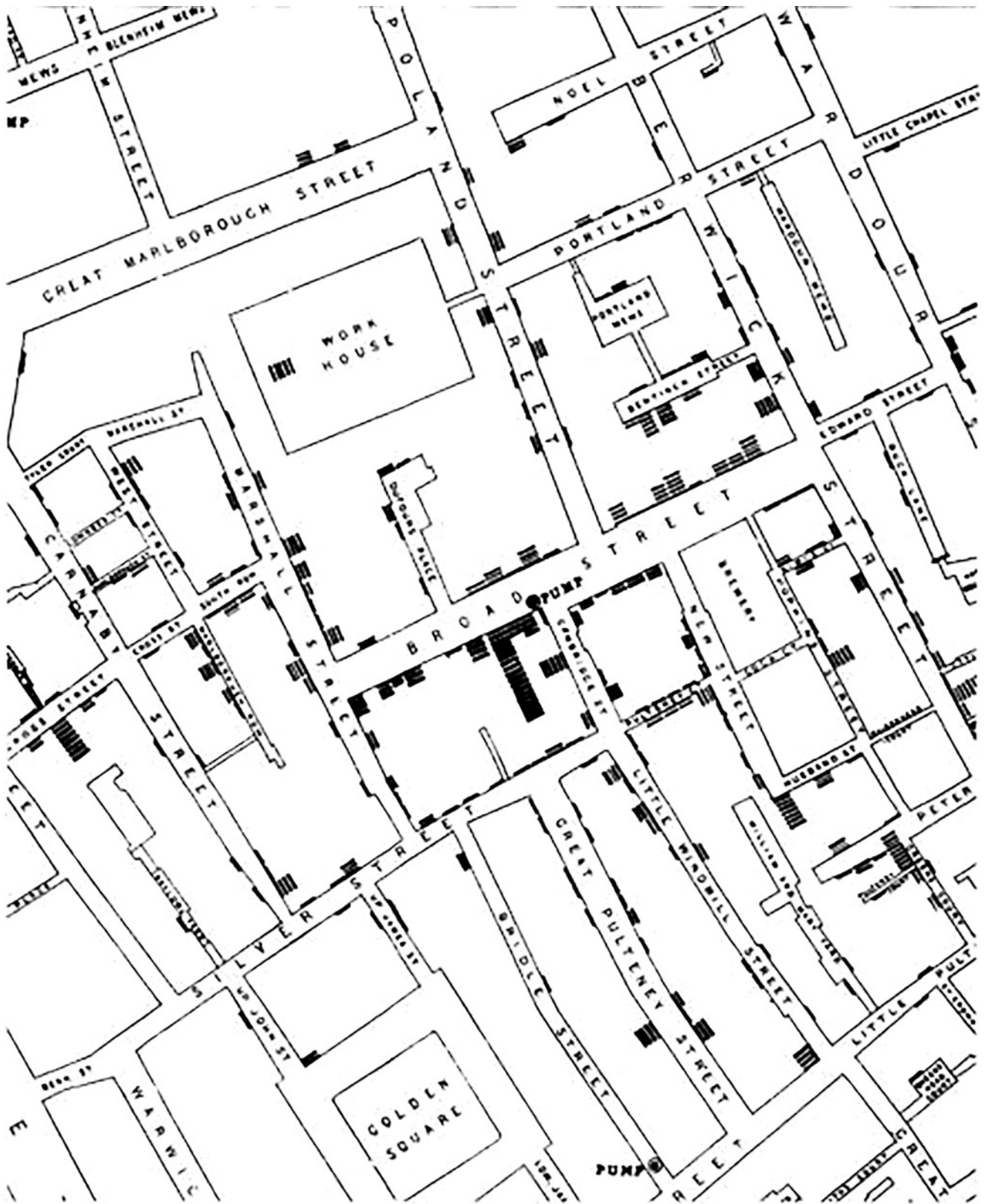


Fig. 1 John Snow's map of cholera death in London (McLeod 2000)

population with sewage tainted water (McLeod 2000; Shiode et al. 2015). John Snow's map was among the first studies to utilize disease maps for understanding public health issues.

In the modern context, disease maps are commonly used to identify spatial relationships between disease outcomes, risk factors in the environment, and population characteristics (Croner et al. 1996; English et al. 1999; Gatrell et al. 2003; Glass et al. 1995; Goodchild et al. 1992). The rise of computerized mapping software and easy access to aggregated health data have enabled the production and delivery of maps via interactive websites or as applications on mobile devices like cell phones. The Centers for Disease Control and Prevention (CDC) publishes mortality and other environmental datasets for the United States via a web-based portal called CDC WONDER. Users of this website are also able to create online maps of various health outcomes. The purpose of this web portal is twofold – (1) it enables researchers and practitioners to create their own maps to aid research and/or public health intervention and (2) it allows the public to produce maps for their own interest. Other examples of publicly available mapping portals for health data include AIDSvu, National Cancer Institute's GIS Portal, among others. While emerging GIS technology has led to the democratization of mapping, thus enabling better public participation in understanding the social and environmental determinants of health, it may also lead to misleading or biased perceptions when maps are not interpreted or used correctly. The use of choropleth maps as the default map type for representing disease burdens is of particular concern for various reasons discussed in the section titled "Methods: Linking Maps of Disease Outcomes to Environmental Risks" (see Fig. 5 for an example). Maps that represent unreliable information may lead to biased and/or incorrect perceptions about the complex relationships between environmental risks, disease burdens, and population characteristics. In addition to careful selection of map type, it is imperative that mapmakers communicate information about the intended purpose of the map, the process by which it was produced, and other information considered important for interpreting the observed patterns.

## Data

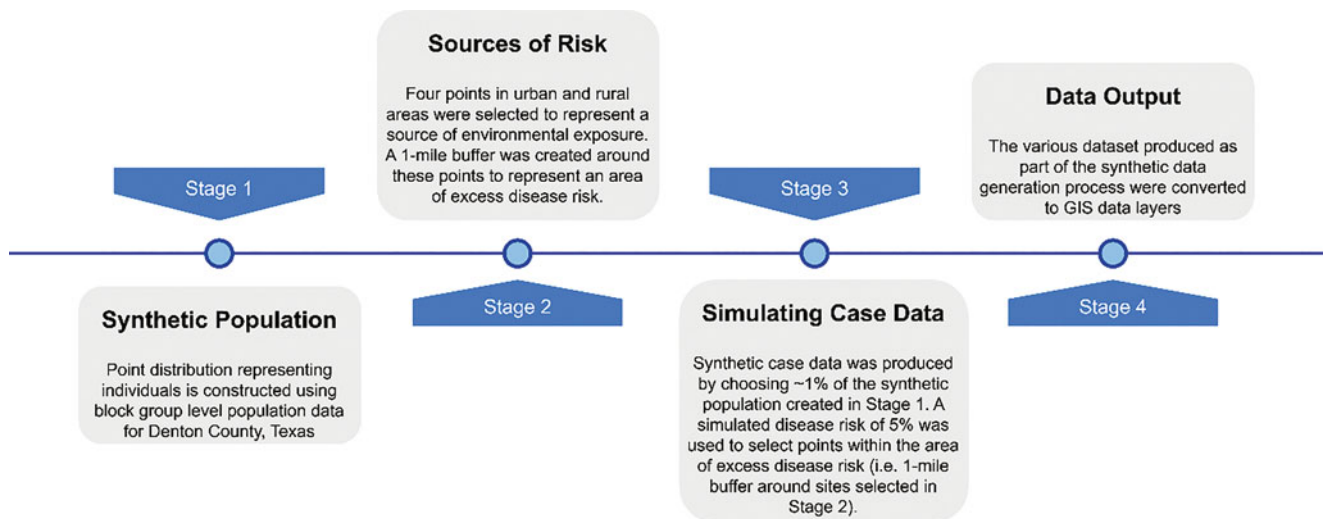
The synthetic data generation process consists of four stages as described in Fig. 2. In stage 1, block-group-level population data for Denton County in Texas was used to create a point distribution representing individuals. Population values were divided by 10 for computational efficiency. The resulting dataset consists of 66,092 points where each point represents an individual. Note that the spatial distribution of these points is proportional to the block-group population distribution in Denton County. In stage 2, four sites of simulated environmental risk were selected in Denton County.

These sites were selected to cover urban and rural contexts. We assumed a 1-mile radius of "exposure" around each of these four sites. We will refer to these buffers as "high-risk" areas. In stage 3, case data were created using two levels of simulated disease risk: (1) a 1% risk of disease among the population overall and (2) a 5% risk of disease among populations within high-risk areas. Finally, in stage 4, the simulated datasets were converted into GIS layers for use in subsequent analysis. The final synthetic dataset consisted of 89 simulated cases and 1508 individuals in the high-risk area (rate = 0.0097) compared to 625 simulated cases and 63,870 individuals overall (rate = 0.059). Block-group-level population data were obtained from the US Census. Alteryx software was used to create the synthetic datasets.

## Methods: Linking Maps of Disease Outcomes to Environmental Risks

A dot density map is the simplest way to represent disease patterns over space. Such maps are typically produced by randomly placing dots or other point symbols within the spatial extent of each geographic unit such that the total number of dots within that unit is equal or proportional to the observed number of disease cases. When producing such maps, the mapmaker chooses a numerical value that each dot represents on the map; for example, the mapmaker may decide that one dot represents five disease cases. Areas containing many dots indicate high concentrations of disease cases, whereas areas with fewer dots represent lower concentrations. As illustrated in Fig. 3, the dot value along with dot size can result in maps with vastly different presentations of disease concentration and spread. Larger spatial units such as the census tracts located in the northern and western parts of the county are more likely to be distorted as the dots are not placed in accordance with population density – instead they are randomly disbursed across the entire spatial extent of each tract. Further, such maps do not take population into consideration and are generally inadequate for measuring the intensity of a disease within a population.

Choropleth maps are a commonly used alternative (Diggle 2000). They are constructed by grouping areas (typically representing administrative units) into categories and are assigned a color based on the value of the variable being mapped. Choropleth maps are commonly used for many reasons – they are easy to produce and interpret; they rely on existing spatial units that typically represent administrative boundaries for which other demographic and secondary data are collected; and the process of aggregating data to some administrative unit often addresses privacy and confidentiality concerns. The process of constructing a choropleth maps typically requires the following three major decisions:



**Fig. 2** Synthetic data generation process

### 1. Choice of Map Unit

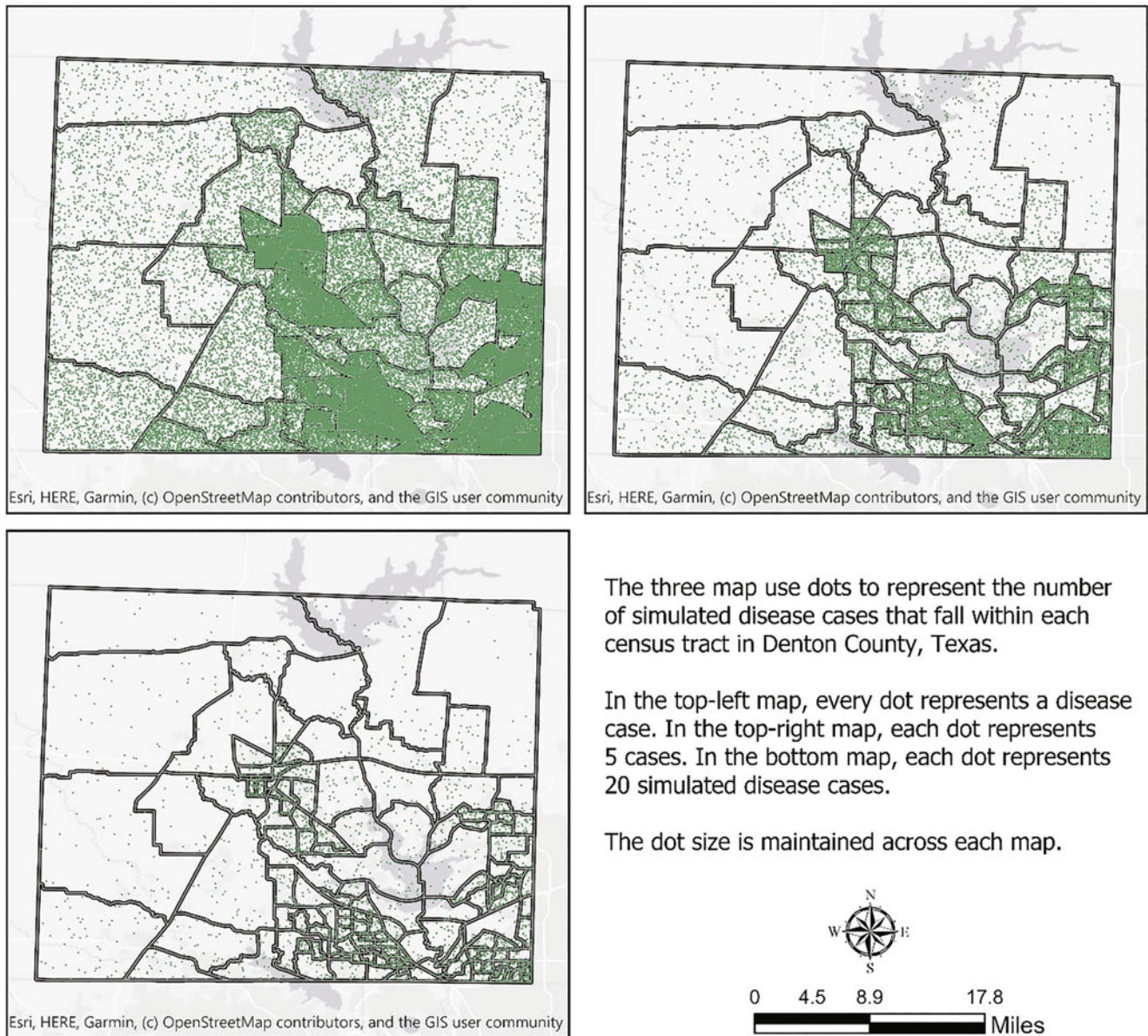
Choropleth maps are based on an existing system of boundaries which form the basic spatial units using which the map is produced. These units represent the geography at which data are collected and/or made available for mapping. In the United States, census entities such as tracts, block groups, or zipcode tabulation areas (ZCTAs) form the basis of many choropleth maps. The choice of map unit influences the patterns that are observed on the map and present several challenges that are discussed in detail below. Figure 4 shows various census entities that are commonly used in the United States. Census tracts are represented by the dark black borders. Census block groups represent finer spatial units and are represented by the yellow lines. Note that census block groups are perfectly contained within census tracts. Zipcode tabulation areas or ZCTAs are a census unit that approximates area representations of zipcode service areas that are created by the US Postal Service for the purposes of mail delivery. Zipcodes are a dynamic entity that do not conform to traditional census statistical data units such as block groups or tracts. This presents a problem **wherein** demographic and/or socioeconomic data collected by the census cannot be linked to zipcodes, which are commonly used descriptions of residential addresses. Although ZCTAs provide a mechanism to link census data to residential zipcodes, it is important to note that they are approximations of zipcodes. The error between the “true zipcode boundary” and ZCTAs is not consistent over space with some areas presenting a greater magnitude of misclassification compared to others. On a related note, one must also be careful when comparing choropleth maps constructed from different spatial units as the underlying geography supporting the statistic being visualized may be different across maps.

### 2. Choice of Classification Method

The process of classification takes a large number of observations and groups them into categories or classes. Creating maps from fewer, well-defined classes makes them easier to read and understand when compared to a map produced from raw data values. The mapmaker typically specifies the number of classes and classification method. Generally, a map must not have more than seven classes. Although more classes result in less data generalization, they may clutter the map with too much detail, thus rendering it ineffective. Commonly employed classification methods include equal intervals, quantiles, and natural breaks. The equal interval method divides the data into equal-sized classes (Fig. 5a). It works best when data values are spread across the entire range. This method must not be applied on a skewed dataset as it may result in a washed-out map where one color (class) dominates. The quantile method places an equal number of observations within each class (Fig. 5b). This method generally results in attractive maps as every color (class) has approximately equal representation. A drawback of this method is that it may result in classes that have varying numerical ranges. The natural breaks method examines the data to identify natural groupings of data that aim to group similar values while maximizing difference between classes (Fig. 5c). The Jenks Natural Breaks algorithm (Jenks 1963) is used in most common GIS software.

### 3. Choice of Color and Map Context

To produce an effective map, the mapmaker must think about the aesthetic qualities of the final map. Considerations include choice of map colors, inclusion of map elements such as a north arrow and scalebar, use of data layers to provide context, labeling styles, among others. Qualitative data are represented using differences



**Fig. 3** Using dot density maps to display disease outcomes

in hue, while quantitative data that contain a progression of low to high values are represented by varying the levels of saturation or lightness of a particular color. Brewer et al. (1997) provide several guidelines on selecting color schemes for mortality maps. The [colorbrewer2.org](http://colorbrewer2.org) website is an excellent resource for mapmakers looking for recommendations of color schemes based on the type of data and map use (Brewer 2003; Harrower and Brewer 2003). Most GIS software allow mapmakers to selectively include various map elements such as north arrows, neatlines, and scalebars. GIS software including ArcGIS and QGIS typically include various options for each map element, thus allowing for high levels of customization

in the production of the final map. Secondary data layers such as road networks, satellite imagery, or topographic maps can be used to provide background or contextual information that can aid the map reader. Examples of how such data can be used in disease maps can be found in Beyer et al. (2012).

While choropleth maps are easy to produce and interpret, they also present several problems, particularly for portraying rates of disease in a population. Such maps are subject to the modifiable areal unit problem (MAUP) which states that any change in the scale (level of aggregation) or shape of map units (such as administrative boundaries) will result in

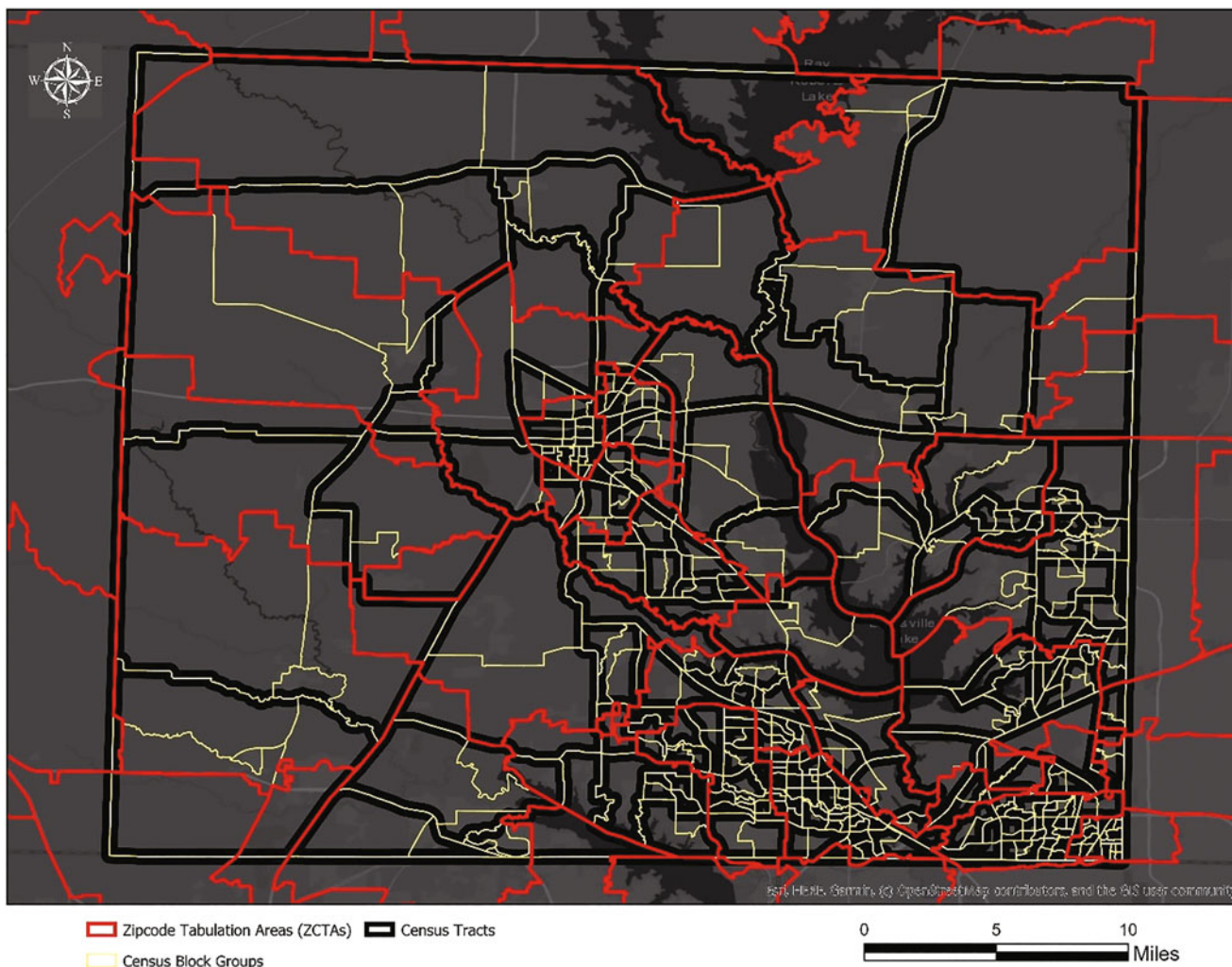


Fig. 4 Commonly used census boundaries in the United States

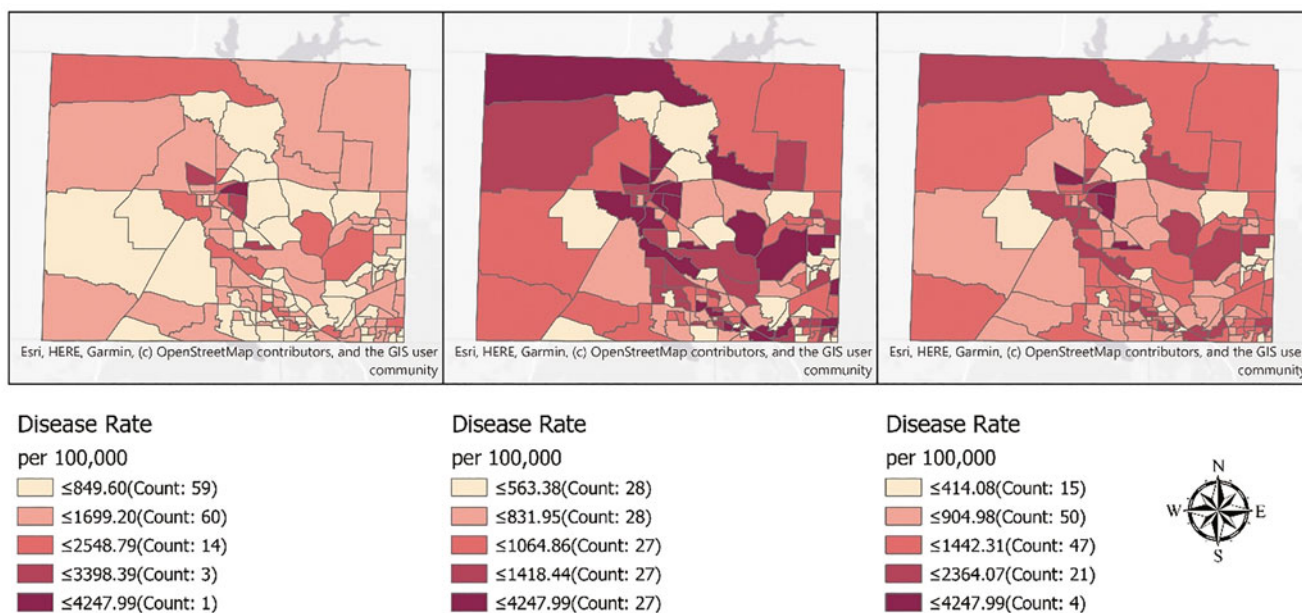


Fig. 5 Commonly used map classification methods

changing map patterns. Simply put, any change in the shape or size of the unit being mapped will result in maps with different spatial patterns of disease burdens. Cressie (1993) showed that administrative boundaries tend to change based on socioeconomic, demographic, and environmental criteria for which health event data are collected and can influence the observed rates and patterns of disease distribution. Bell et al. (2006) added that health data are aggregated on predefined spatial scales and any change in boundary does not represent true numerical information about the region. In other words, the aggregation of data into arbitrary administrative units can lead to loss of information about how diseases are distributed within those units themselves. Further, choropleth maps of disease rates are subject to statistical variability due to small numbers problem. In other words, areas with sparse population counts are likely to yield estimates of disease rates that are highly unstable and may dramatically change with the addition or deletion of a few cases.

Methods to address the small numbers problem aim to increase the population basis of support by aggregating data over space and/or time to create collections of larger, contiguous spatial units known as *spatial supports* (Beyer et al. 2012; Hansen 1991; Mungiole et al. 1999; Rushton et al. 2000). Other methods rely on the use of geostatistical modeling approaches (Berke 2005; Goovaerts 2005; Goovaerts 2006) or other types of statistical techniques (Clayton and Kaldor 1987; Devine et al. 1994; Lawson et al. 2000; Marshall 1991; Mollie and Richardson 1991). A third category of disease maps represents disease risk as a continuous function over geographical space. Kernel density estimation methods are commonly used to produce such maps (Talbot et al. 2000; Tiwari and Rushton 2005). Maps produced using these methods use a kernel or spatial filter characterized by a particular shape, size, and density function (Carlos et al. 2010; Shi 2010) to compute the intensity of a disease along a set of sampling locations overlaid across the study area. Disease rates are computed at each sample location by dividing the number of cases that fall within a kernel placed at that point by the population contained within it. The size of the kernel is determined using one of two strategies: (1) a fixed size is used at each sample point, thus ensuring consistent spatial support but variable population support and (2) kernel sizes expand or contract to meet a minimum population threshold, thereby ensuring consistent population support but variable spatial support (Talbot et al. 2000; Tiwari and Rushton 2005; Tiwari 2013). Variable sized kernels or adaptive spatial filters are preferred over fixed-size filters as they address problems of undersmoothing or oversmoothing. Undersmoothing results when the kernel size is not large enough and continues to compute disease rates using sparse population counts. This may occur in rural areas where population densities tend

to be low. Oversmoothing occurs when the kernel size is larger than what would be needed to compute a stable disease rate. Oversmoothing occurs in densely populated urban areas and results in loss of resolution on a map. Variable sized kernels contract and expand in size such that each kernel contains some minimum, user-defined population threshold. Resulting maps provide consistent levels of statistical reliability across all areas and high levels of geographic detail in areas where such detail is expected (e.g., urban contexts). In the work discussed in this chapter, we used the Web-Based Disease Mapping and Analysis Program (WebDMAP) to produce such maps. Following are the three major steps involved.

### 1. Create Data Files

WebDMAP requires three data files to compute disease rates using the kernel density estimation method. The grid file provides point locations on which kernels or spatial filters will be constructed. The other two files provide the locations of disease cases and populations, respectively. If individual-level data are available, each location represents an individual. Alternatively, each location can also represent aggregated counts of case/population data for some spatial unit such as a census block group or ZCTA. Location data must be provided in unprojected coordinates (i.e. latitude and longitude). Simulated disease and population data used in this chapter can be downloaded from <http://webdmap.com/kdedata>.

### 2. Define Minimum Population Threshold

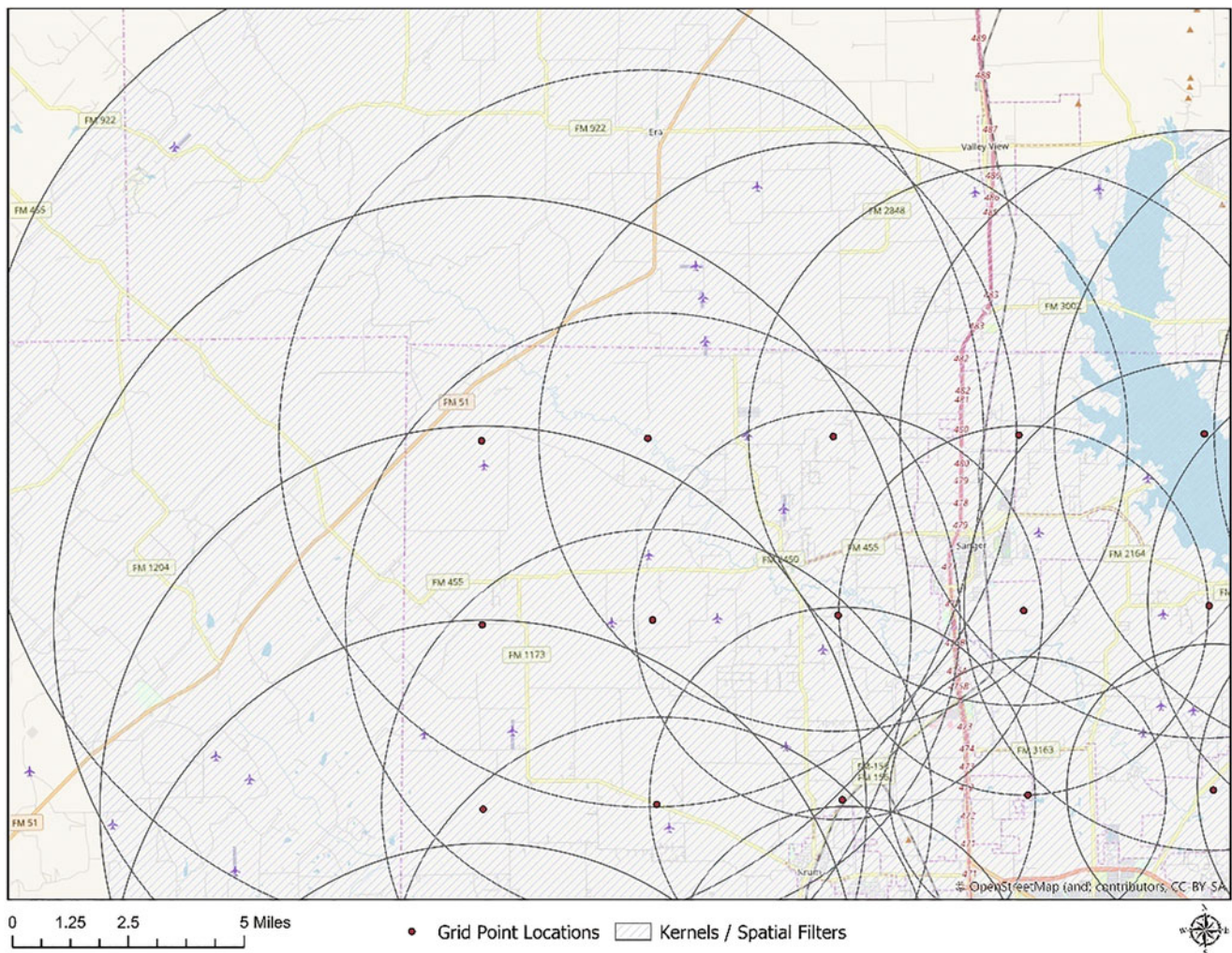
Recall that the size of the kernel/spatial filter that is placed at each grid point is determined by some user-defined minimum population size value. Note that the size of the spatial filters is determined by this user-specified parameter. Large population thresholds in areas with sparse populations will result in the largest filter sizes. Conversely, small population thresholds in areas with dense populations will result in the smallest filter sizes. In the work discussed in this chapter, we used a population threshold of 1000 individuals. The study area, Denton County, comprises dense urban areas (central and south-eastern portions) as well as sparsely populated rural areas (northwestern portions). Correspondingly, we see a combination of small and large filter sizes across the study region (Fig. 6).

### 3. Compute Rates and Produce Maps

The algorithm for computing rates using this method is described below:

- (a). Compute distance strings for the case and population data. Distance strings are a kind of data structure that were originally designed for efficiently storing information about travel costs between nodes





**Fig. 6** Grid points and spatial filters

and were used frequently in location allocation algorithms (Densham and Rushton 1992; Hillsman 1984; Sorensen and Church 1995). The basic idea behind a distance string is that it stores information about travel costs (i.e., distance) between a base node (i.e., every grid point) and all other nodes (i.e., cases and population locations) in an increasing order of distance. The procedure is implemented in a PostgreSQL database, thus enabling the calculation of distance strings for large datasets. To improve computational efficiency, distance strings are truncated at a user-defined cutoff value. By doing so, we assume that spatial filters will never be larger than a certain size and therefore terminate distance string calculations at the user-defined cutoff value. In this analysis, the distance string cutoff was set to 100 miles – i.e., we assumed that spatial filters will never exceed a 100-mile radius.

To further improve computational efficiency, spatial indexes were created for all data tables, thereby resulting in substantially faster database query processing times. See Nguyen (2009) for details on how spatial indexes work within the PostgreSQL/PostGIS relational database. Distance strings are computed for the case and population data.

- (b). For each grid point, use the population distance strings table to identify the distance associated with the user-defined population threshold value. This is implemented using database functions that query the population distance strings table to identify the distance value that corresponds with the row where the cumulative population weight exceeds the user-defined population threshold. This is the size of the spatial filter. For example, in Fig. 7, if the user-defined population threshold is set to 200, the algorithm will

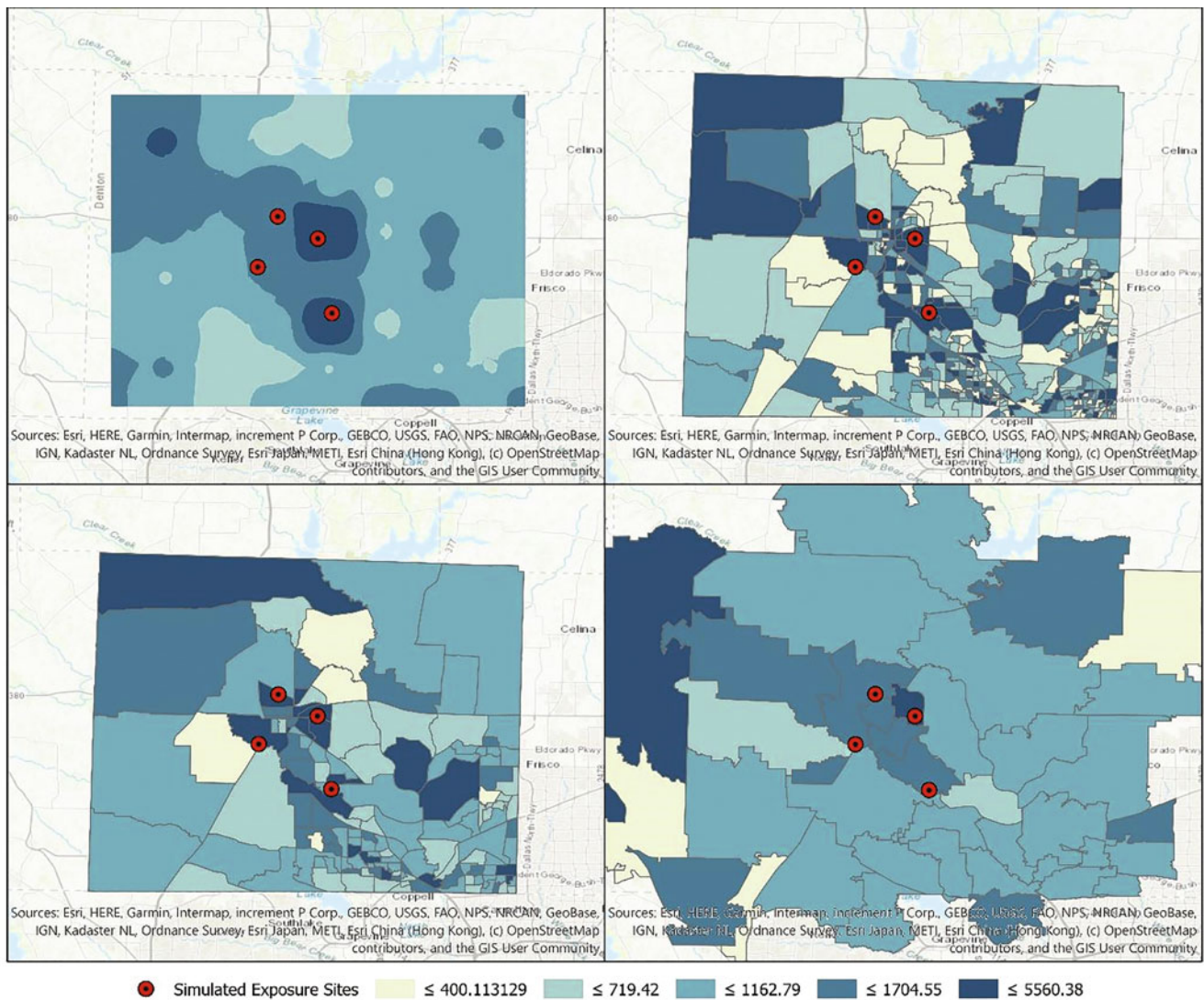
gridpoint	populationlocation	distance	population	cumulativepopulation
1	a	1.1 miles	30	30
1	b	1.4 miles	100	130
1	e	1.9 miles	75	205
1	g	3.4 miles	20	225
...		...		
2	t	4.1 miles	10	10
2	f	4.8 miles	5	15
2	b	7.4 miles	15	30
2	a	12.3 miles	25	55
...		...		

**Fig. 7** Population distance strings example

gridpoint	caselocation	distance	cases	cumulativecases
1	a	1.1 miles	5	5
1	b	1.4 miles	2	7
1	e	1.9 miles	7	14
1	g	3.4 miles	3	17
...		...		
2	t	4.1 miles	4	4
2	f	4.8 miles	1	5
2	b	7.4 miles	5	10
2	a	12.3 miles	8	18
...		...		

**Fig. 8** Case distance strings example

- select the record highlighted in orange to define the size of the spatial filter (i.e., 1.9 miles). Note that the actual population contained within this spatial filter is 205. This may occur when aggregated data are used. This process is repeated for every grid point.
- (c). For each grid point, query the case distance strings table and note the cumulative weight value that is associated with the distance noted in step b above. This is the number of cases that fall within the spatial filter size at that grid point. For example, if the distance value at grid point 1 is 1.9 miles (step b), then the number of disease cases that are contained within the spatial filter constructed at that grid point is 14.
  - (d). Compute a rate at every grid point by dividing the cumulative number of cases (step c) (Fig. 8) by the cumulative population (step b) (Fig. 7).
  - (e). Repeat steps b through d to compute a rate for all the grid points.
  - (f). A continuous surface map can be created from the grid points using the inverse distance weighted (IDW) interpolation method in any standard GIS software. The IDW method with 8 neighbors and a power of at least 2 is recommended to avoid any “double” smoothing that may occur in addition to what has already been performed by the spatially adaptive filter method.



**Fig. 9** Disease maps produced by us

The maps shown in Fig. 9 were constructed using kernel density estimation (Fig. 9a) and as choropleth maps (Fig. 9b–d). The four red dots indicate the sites of some environmental exposure. These sites were randomly selected to include urban and rural contexts. As described earlier, areas within a 1-mile buffer around each of the four points represent an area of elevated disease risk (five times the overall rate). To construct the map in Fig. 9a, a grid of points, placed 4 miles apart, was overlaid on top of the study area (Denton County, Texas). At each point of this grid, variable sized kernels were constructed such that each kernel or spatial filter contained exactly 1000 individuals. If aggregated data were used instead of individual-level point data, each kernel would contain a collection of spatial units with a minimum population size of 1000. Kernel size (radius) ranged from a

minimum of 0.72 miles to a maximum of 10.834 miles. The average kernel size was 4.163 miles. The number of cases falling within each kernel were assigned to each grid point. These ranged from a minimum of 4 cases to a maximum of 33 with an average of 10.736 cases. Rates were computed at each grid point by dividing the case count by population. Disease rates computed at each grid point ranged from a minimum value of 400 per 100,000 to a maximum value of 3300 per 100,000. The average rate was 1073 cases per 100,000 population. Rate values at each grid point were converted into a continuous surface of disease risk using the inverse distance weighted (IDW) interpolation technique. Disease rates were computed using the Web-Based Disease Analysis and Mapping Program (Web-DMAP). Final map output was created using ArcGIS Pro.

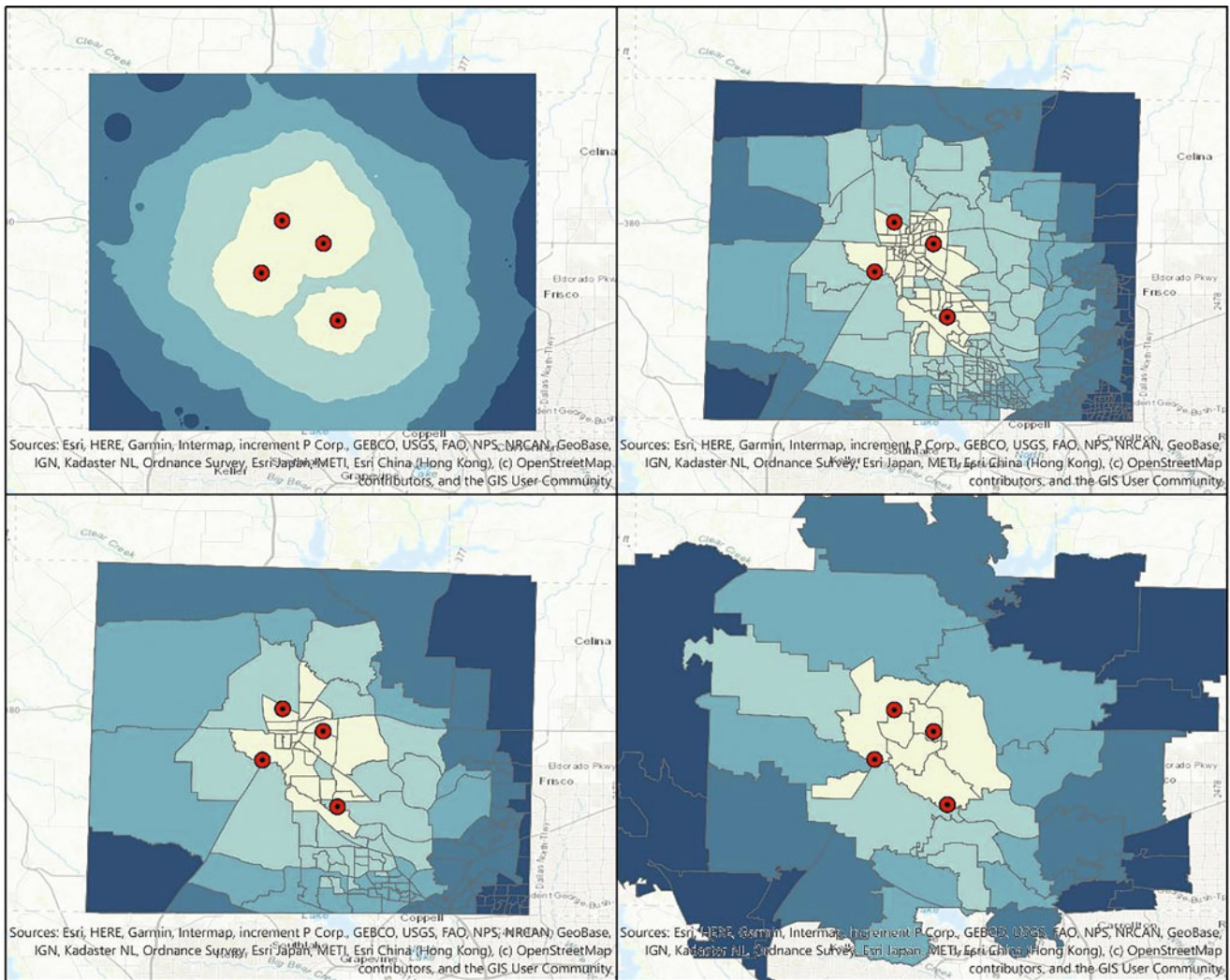
The maps in Fig. 9b–d were created by aggregating case and population counts to three sets of administrative boundaries with varying levels of spatial resolution. The maps used census block groups, census tracts, and zip code tabulation areas (ZCTAs), respectively. For each spatial unit (i.e., each block group, tract, or ZCTA), a rate was calculated by dividing the total number of cases within that unit by its population. For all four maps, rates were classified into five groups using the quantile classification method.

The spatial patterns of disease rates in each of the four maps present slight variations when compared to each other. Note that the underlying data used in each map are identical. Differences in observed patterns are a result of the different levels of aggregation and the method used to construct the map. Among the four maps, Fig. 9b presents the most geographic detail. However, due to the relatively small size of census block groups, they also contain the most variability in populations ranging from a minimum of 6 persons within a block group to a maximum of 646. Due to the variable population sizes, block groups also portrayed the most variability in disease rates with an average rate and standard deviation of 1045.67 and 939.26 per 100,000 population, respectively. The highest rates of disease were observed around the four exposure sites along with pockets in the northwest and eastern sections of the county. In contrast, the map in Fig. 9d contains the least amount of geographic detail. The use of large ZCTA boundaries tends to wash away any fine-scale variations in disease rates. Only one of the four exposure sites is located in an area of highest disease risk. Pockets of high rates are observed in the northwestern parts of the county. While the level of geographic detail presented in this map is low, it also contains the most stable estimates of disease rates. The average rate in ZCTAs in Denton County was found to be 1066.53 cases per 100,000 population with a standard deviation of 574.17. The map in Fig. 9c represents a balance between the maps presented in 9b and 9d. It uses census tracts, which are slightly larger in size (and population) compared to block groups and considerably smaller in size compared to ZCTAs. Areas surrounding the four exposure sites are classified as areas of highest disease risk in addition to pockets of high rates in the northwestern and eastern parts of the county. The average disease rate is estimated at 1040.8 per 100,000 population along with a standard deviation of 642.97. Finally, the map in Fig. 9a identifies areas of highest disease risk surrounding two of the four exposure sites. Unlike a choropleth map, this map does not use discrete spatial units to represent rates across Denton County. Instead, a continuous surface of disease risk is used to identify areas of highest and lowest rates. As discussed earlier in this chapter, the average rate is found to be 1073 cases per 100,000 population with a standard deviation of 462.99. Among the four maps, the one produced using kernel density estimation presents the best balance

between resolution and reliability. The ability to control the population basis of support ensures that consistent sample sizes are used in the calculation of every disease rate across the map. This map presents a desirable tradeoff between geographic resolution and reliability – maintaining high levels of geographic detail in urban areas while preserving high levels of statistical reliability in rural areas. While the map produced using census block groups presents high levels of visual detail, such maps must be used with caution due to the problem of unstable rates caused by small population counts. Conversely, a map that uses coarse spatial units such as ZCTAs not only maintains statistical reliability but also leads to severe loss in geographic detail across the entire map.

The choice of disease map and/or the spatial resolution at which disease data are available influence the ability to detect associations between disease burdens and environmental exposures. The maps in Fig. 10a–d represent the spatial patterns of population exposure to four point locations of simulated environmental risk. These four locations are denoted by red dots in Figs. 9a–d and 10a–d. Exposure is measured as the Euclidean distance between each individual, represented as a point in the synthetic dataset, and the closest point location representing a site of environmental risk. Map 10a was produced by interpolating distances computed for each individual point in Denton County. Lighter colors represent closer distances compared to darker colors. As expected, areas close to the four red dots on the map show lower distance values. The maps in Fig. 10b–d represent the average exposure distance for populations aggregated to census block groups, tracts, and ZCTAs, respectively. As expected, areas within close proximity of the four red dots portray lower exposure distances. Note that the block-group-level map shows better geographic resolution when compared to the other maps. This is generally a desirable property in maps of environmental exposure when compared to disease maps, where high levels of geographic detail typically represent poor statistical reliability. However, it is critical to note that valid map comparisons can only be made when the underlying geographic or spatial basis of support is consistent across all maps that are being compared. Inconsistent spatial supports can result from differences in resolution, scale, or boundary definitions. For example, one cannot directly compare a disease map constructed using the KDE method with a block-group-level map of environmental exposure. Dasymetric mapping (REF) or other geostatistical modeling techniques including interpolation (REF PYCNO) may be used to reconcile maps that do not have consistent spatial supports.

The scatter plots in Fig. 11a–d show the directionality and strength of the relationships between exposures and disease outcomes. Figs. 11a, b represent the relationship between disease and exposure data measured at fine geographic scales (individual- and block-group levels), whereas plots in Fig.

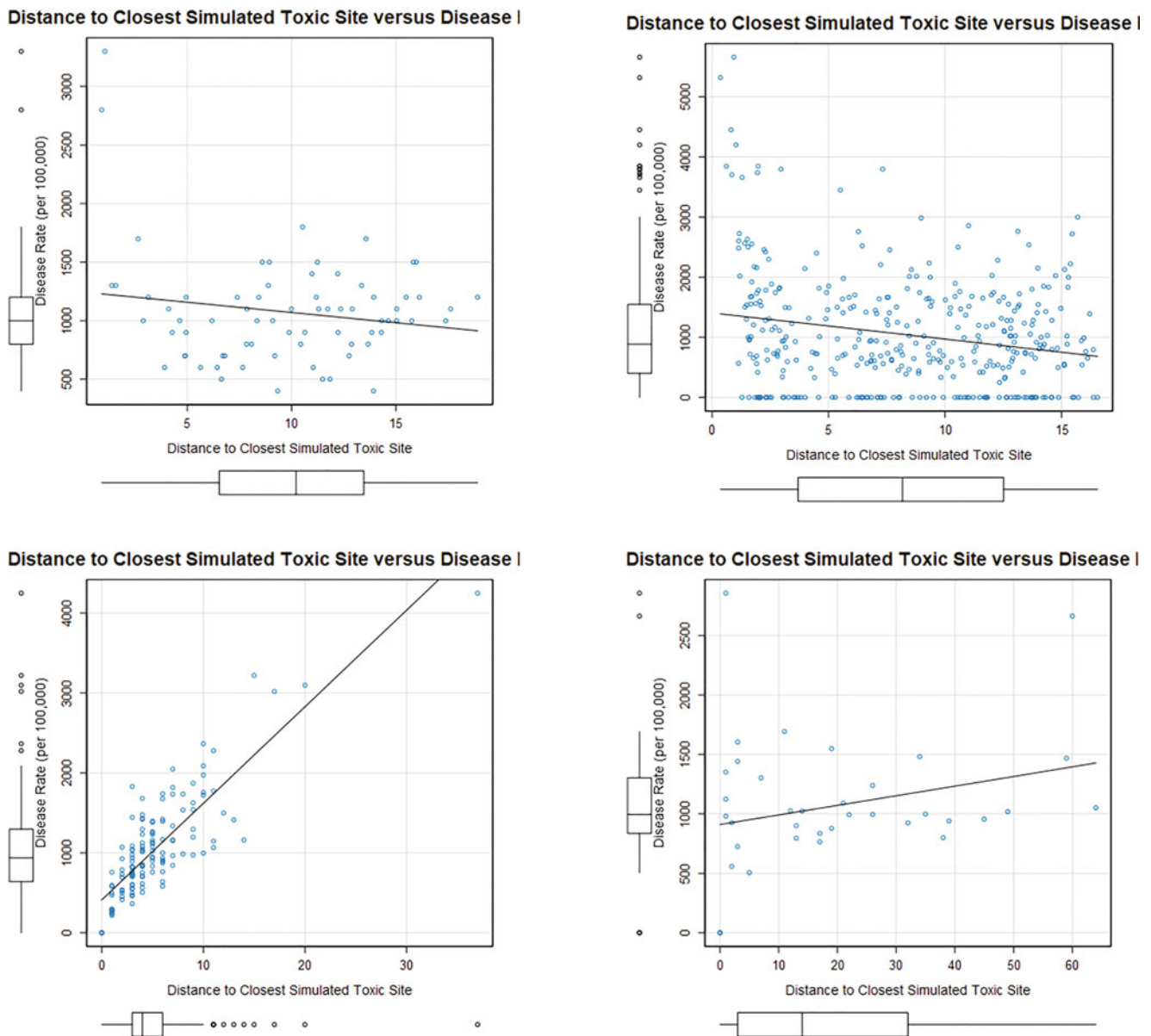


**Fig. 10** Exposure maps

11c, d represent this relationship at coarser geographic scales (tract and ZCTA levels, respectively). It is interesting to note that scatter plots produced using data at finer geographic scales correctly describe a negative relationship between distance and disease intensity. Note that the synthetic dataset was produced using a five times greater risk of disease in individuals within 1 mile of a simulated site of environmental exposure. Conversely, the scatter plots produced using coarser data represent an inverse relationship, suggesting that disease risk increases as one moves away from these sites of environmental concern. The synthetic data produced do not support this conclusion. It is also important to note that, although the relationship between distance and disease rate is correctly represented in Fig. 11b, the variability in disease rates as indicated by the boxplot beside the y-axis is likely to bias the strength of the relationships between distance and disease rates.

## Conclusions

The type of mapping method used to produce maps of disease outcomes or environmental risks as well as their parameters influences the observed patterns of disease distribution and consequently our interpretations of associated risk factors. It is important to remember that a map merely represents one abstraction of complex underlying processes that control how diseases and environmental risks manifest themselves across space and time. The construction of an “honest map” requires full disclosure of the methods used, scale of analysis, quality of data, and other parameters used in the final construction of the map. The objective of this chapter is not to identify the “best” mapping method but to demonstrate that each method comes with advantages and disadvantages and, importantly, have an impact on the patterns and relationships that are observed.



**Fig. 11** Relationships between exposures and disease outcomes

## References

- Al-Hamdan, M.Z., W.L. Crosson, S.A. Economou, M.G. Estes Jr., S.M. Estes, S.N. Hemmings, et al. 2014. Environmental public health applications using remotely sensed data. *Geocarto International* 29 (1): 85–98.
- Al-Hamdan, M.Z., W.L. Crosson, A.S. Limaye, D.L. Rickman, D.A. Quattrochi, M.G. Estes Jr., et al. 2009. Methods for characterizing fine particulate matter using ground observations and remotely sensed data: Potential use for environmental public health surveillance. *Journal of the Air & Waste Management Association* 59 (7): 865–881.
- Bell, B.S., R.E. Hoskins, L.W. Pickle, and D. Wartenberg. 2006. Current practices in spatial analysis of cancer data: Mapping health statistics to inform policymakers and the public. *International Journal of Health Geographics* 5 (1): 49.
- Bellander, T., N. Berglind, P. Gustavsson, T. Jonson, F. Nyberg, G. Pershagen, and L. Järup. 2001. Using geographic information systems to assess individual historical exposure to air pollution from traffic and house heating in Stockholm. *Environmental Health Perspectives* 109 (6): 633–639.
- Berke, O. 2005. Exploratory spatial relative risk mapping. *Preventive Veterinary Medicine* 71 (3–4): 173–182.
- Bertollini, R., and M. Martuzzi. 1999. Disease mapping and public health decision-making: Report of a WHO meeting. *American Journal of Public Health* 89 (5): 780.
- Beyer, K.M., C. Tiwari, and G. Rushton. 2012. Five essential properties of disease maps. *Annals of the Association of American Geographers* 102 (5): 1067–1075.
- Brewer, C.A. 2003. A transition in improving maps: The ColorBrewer example. *Cartography and Geographic Information Science* 30 (2): 159–162.

- Brewer, C.A., A.M. MacEachren, L.W. Pickle, and D. Herrmann. 1997. Mapping mortality: Evaluating color schemes for choropleth maps. *Annals of the Association of American Geographers* 87 (3): 411–438.
- Carlos, H.A., X. Shi, J. Sargent, S. Tanski, and E.M. Berke. 2010. Density estimation and adaptive bandwidths: A primer for public health practitioners. *International Journal of Health Geographics* 9 (1): 39.
- Clayton, D., & Kaldor, J. 1987. Empirical bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*: 671–681.
- Cressie, N.A. 1993. *Statistics for spatial data*. New York: John Wiley and Sons. Inc.
- Croner, C.M., J. Sperling, and F.R. Broome. 1996. Geographic information systems (GIS): New perspectives in understanding human health and environmental relationships. *Statistics in Medicine* 15 (18): 1961–1977.
- Densham, P.J., and G. Rushton. 1992. Strategies for solving large location-allocation problems by heuristic methods. *Environment and Planning A* 24 (2): 289–304.
- Devine, O.J., Louis, T.A., & Halloran, M.E. 1994. Empirical bayes methods for stabilizing incidence rates before mapping. *Epidemiology*: 622–630.
- Diggle, P.J. 2000. Overview of statistical methods for disease mapping and its relationship to cluster detection. *Spatial Epidemiology: Methods and Applications* 87: 103.
- Elliott, P., D. Briggs, S. Morris, C. de Hoogh, C. Hurt, T.K. Jensen, et al. 2001. Risk of adverse birth outcomes in populations living near landfill sites. *BMJ* 323 (7309): 363–368.
- English, P., R. Neutra, R. Scalf, M. Sullivan, L. Waller, and L. Zhu. 1999. Examining associations between childhood asthma and traffic flow using a geographic information system. *Environmental Health Perspectives* 107 (9): 761–767.
- Fisher, R., T. Walshe, P. Bessell-Browne, and R. Jones. 2018. Accounting for environmental uncertainty in the management of dredging impacts using probabilistic dose–response relationships and thresholds. *Journal of Applied Ecology* 55 (1): 415–425.
- Gatrell, A.C., J.C. Harman, B.J. Francis, C. Thomas, S.M. Morris, and M. McIlmurray. 2003. Place of death: Analysis of cancer deaths in part of north West England. *Journal of Public Health* 25 (1): 53–58.
- Glass, G.E., B.S. Schwartz, J.M. Morgan III, D.T. Johnson, P.M. Noy, and E. Israel. 1995. Environmental risk factors for Lyme disease identified with geographic information systems. *American Journal of Public Health* 85 (7): 944–948.
- Goodchild, M., R. Haining, and S. Wise. 1992. Integrating GIS and spatial data analysis: Problems and possibilities. *International Journal of Geographical Information Systems* 6 (5): 407–423.
- Goovaerts, P. 2005. Geostatistical analysis of disease data: Estimation of cancer mortality risk from empirical frequencies using poisson kriging. *International Journal of Health Geographics* 4 (1): 31.
- . 2006. Geostatistical analysis of disease data: Accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point poisson kriging. *International Journal of Health Geographics* 5 (1): 52.
- Griffith, D.A. 2018. Uncertainty and context in geography and GIScience: Reflections on spatial autocorrelation, spatial sampling, and health data. *Annals of the American Association of Geographers* 108 (6): 1499–1505.
- Hansen, K.M. 1991. Head-banging: Robust smoothing in the plane. *IEEE Transactions on Geoscience and Remote Sensing* 29 (3): 369–378.
- Harrower, M., and C.A. Brewer. 2003. ColorBrewer.org: An online tool for selecting colour schemes for maps. *The Cartographic Journal* 40 (1): 27–37.
- Hillsman, E.L. 1984. The p-median structure as a unified linear model for location—Allocation analysis. *Environment and Planning A* 16 (3): 305–318.
- Jenks, G.F. 1963. Class intervals for statistical maps. *International Yearbook Cartography* 3: 119–134.
- Koch, T. 2004. The map as intent: variations on the theme of John Snow. *Cartographica: The International Journal for Geographic Information and Geovisualization* 39(4): 1–14.
- Kwan, M. 2012. The uncertain geographic context problem. *Annals of the Association of American Geographers* 102 (5): 958–968.
- . 2018. The limits of the neighborhood effect: Contextual uncertainties in geographic, environmental health, and social science research. *Annals of the American Association of Geographers* 108 (6): 1482–1490.
- Lawson, A.B., A.B. Biggeri, D. Böhning, E. Lesaffre, J.F. Viel, A. Clark, et al. 2000. Disease mapping models: An empirical evaluation. Disease mapping collaborative group. *Statistics in Medicine* 19 (17): 2217–2241.
- Leelasakultum, K., and N.T. Kim Oanh. 2017. Mapping exposure to particulate pollution during severe haze episode using improved MODIS AOT-PM10 regression model with synoptic meteorology classification. *GeoHealth* 1 (4): 165–179.
- Marshall, R.J. 1991. Mapping disease and mortality rates using empirical bayes estimators. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 40 (2): 283–294.
- McLeod, K.S. 2000. Our sense of snow: The myth of john snow in medical geography. *Social Science & Medicine* 50 (7–8): 923–935.
- Mennis, J., and E.E. Yoo. 2018. Geographic information science and the analysis of place and health. *Transactions in GIS* 22 (3): 842–854.
- Mollie, A., and S. Richardson. 1991. Empirical bayes estimates of cancer mortality rates using spatial models. *Statistics in Medicine* 10 (1): 95–112.
- Moore, D.A., and T.E. Carpenter. 1999. Spatial analytical methods and geographic information systems: Use in health research and epidemiology. *Epidemiologic Reviews* 21 (2): 143–161.
- Mungiole, M., L.W. Pickle, and K.H. Simonson. 1999. Application of a weighted head-banging algorithm to mortality data maps. *Statistics in Medicine* 18 (23): 3201–3209.
- Nguyen, T.T. 2009. Indexing PostGIS databases and spatial query performance evaluations. *International Journal of Geoinformatics* 5 (3): 1.
- Nuckols, J.R., M.H. Ward, and L. Jarup. 2004. Using geographic information systems for exposure assessment in environmental epidemiology studies. *Environmental Health Perspectives* 112 (9): 1007–1015.
- Nyberg, F., Gustavsson, P., Järup, L., Bellander, T., Berglind, N., Jakobsson, R., & Pershagen, G. 2000. Urban air pollution and lung cancer in Stockholm. *Epidemiology*: 487–495.
- Reif, J.S., J.B. Burch, J.R. Nuckols, L. Metzger, D. Ellington, and W.K. Anger. 2003. Neurobehavioral effects of exposure to trichloroethylene through a municipal water supply. *Environmental Research* 93 (3): 248–258.
- Ricketts, T.C. 2003. Geographic information systems and public health. *Annual Review of Public Health* 24 (1): 1–6.
- Rushton, G., G. Elmes, and R. McMaster. 2000. Considerations for improving geographic information system research in public health. *Urisa-Washington DC* 12 (2): 31–50.
- Shi, X. 2010. Selection of bandwidth type and adjustment side in kernel density estimation over inhomogeneous backgrounds. *International Journal of Geographical Information Science* 24 (5): 643–660.
- Shi, Y., C. Ren, M. Cai, K.K. Lau, T. Lee, and W. Wong. 2019. Assessing spatial variability of extreme hot weather conditions in Hong Kong: A land use regression approach. *Environmental Research* 171: 403–415.

- Shiode, N., S. Shiode, E. Rod-Thatcher, S. Rana, and P. Vinten-Johansen. 2015. The mortality rates and the space-time patterns of john snow's cholera epidemic map. *International Journal of Health Geographics* 14 (1): 21.
- Sorensen, P.A., & Church, R.L. 1995. A comparison of strategies for data storage reduction in location-allocation problems (95-4).
- Talbot, T.O., M. Kulldorff, S.P. Forand, and V.B. Haley. 2000. Evaluation of spatial filters to create smoothed maps of health data. *Statistics in Medicine* 19 (17-18): 2399-2408.
- Tiwari, C. 2013. Methods for creating smoothed maps of disease burdens. In *Geographic health data: Fundamental techniques for analysis*, ed. F. Boscoe. Wallingford: CABI.
- Tiwari, C., & Rushton, G. 2005. Using spatially adaptive filters to map late stage colorectal cancer incidence in Iowa. *Developments in spatial data handling* (pp. 665-676). Springer, Berlin, Heidelberg.
- Tomlinson, C.J., L. Chapman, J.E. Thornes, and C. Baker. 2011. Remote sensing land surface temperature for meteorology and climatology: A review. *Meteorological Applications* 18 (3): 296-306.