Published in final edited form as:
Ann Assoc Am Geogr. 2014 ; 105(1): 48–66. doi:10.1080/00045608.2014.968910.

A Place-Oriented, Mixed-Level Regionalization Method for Constructing Geographic Areas in Health Data Dissemination and Analysis

Lan Mu[*], **Fahui Wang**[†], **Vivien W. Chen**[‡], and **Xiao-Cheng Wu**[‡]

[*]Department of Geography, University of Georgia

[†]Department of Geography and Anthropology, Louisiana State University

[‡]Louisiana Tumor Registry, Louisiana State University Health Sciences Center

Abstract

Similar geographic areas often have great variations in population size. In health data management and analysis, it is desirable to obtain regions of comparable population by decomposing areas of large population (to gain more spatial variability) and merging areas of small population (to mask privacy of data). Based on the Peano curve algorithm and modified scale-space clustering, this research proposes a mixed-level regionalization (MLR) method to construct geographic areas with comparable population. The method accounts for spatial connectivity and compactness, attributive homogeneity, and exogenous criteria such as minimum (and approximately equal) population or disease counts. A case study using Louisiana cancer data illustrates the MLR method and its strengths and limitations. A major benefit of the method is that most upper level geographic boundaries can be preserved to increase familiarity of constructed areas. Therefore, the MLR method is more human-oriented and place-based than computer-oriented and space-based.

Keywords

health data analysis; mixed-level regionalization (MLR); modified Peano curve algorithm (MPC); modified scale-space clustering (MSSC); place-oriented, space and place

Spatial clustering or regionalization methods are commonly used in geographic information systems (GIS) and public health for confirmatory or exploratory purposes (Cromley and McLafferty 2012b). Clustering has two different definitions and both are well accepted: partitioning, which assigns a unique cluster membership to any location in the study area, and nonpartitioning (i.e., identifying cluster centers), which does not have an inclusive requirement for all places (Neuberger and Lynch 1982; Hanson and Wieczorek 2002; Szwarcwald, Andrade, and Bastos 2002; Oliver et al. 2006; Schootman et al. 2007; Shishehbor et al. 2008; Moore et al. 2009; Nelson et al. 2009). This article focuses on regionalization, but some discussions also use the term *clustering* as a convention in the

Correspondence: Department of Geography, University of Georgia, Athens, GA 30602, mulan@uga.edu (Mu); Department of Geography and Anthropology, Louisiana State University, Baton Rouge, LA 70803, fwang@lsu.edu (Wang); Louisiana Tumor Registry, Louisiana State University Health Sciences Center, New Orleans, LA 70112, vchen@lsuhsc.edu (Chen); xwu@lsuhsc.edu (Wu).

literature. A challenge for many of these methods is not the development of algorithm, computation, or technical implementation but, rather, making sense of or interpreting the findings. Meaningful results are not just about the size and shape of clusters but the clusters' alignment with existing zonings, particularly boundaries of major geographic units. A fundamental purpose of regionalization is to group and simplify data, not to introduce further complexity by adding more boundaries that are not recognizable by administrators, public practitioners, or the general public.

"Place is security, space is freedom" (Tuan 1977, 3). Tuan's (1974, 1977, 2012) humanist geography approach has influenced generations of geographers by clarifying the relationship between place and space. Tuan illustrated the functions of boundary as bounding place to space such as an Eskimo's sense (or attachment) of trading locations and hunting space (Carpenter, Varley, and Flaherty 1959), and identified space as place with familiar landmarks and paths that are often seen as boundaries. Our regionalization method is inspired by this conceptualization of "place + space + identity + attachment" by geographers (Tuan 1974, 1977; Sack 1980, 2003; Adams, Hoelscher, and Till 2001). Yiannakoulias (2011) advocated a "placefocused" or "place-informed" approach to incorporate locally relevant factors in all aspects of human activities into forming places or regions for meaningful public health surveillance of spatial aberrations. Space is more general and abstract, and place is more attached to people and the environment. Although many regionalization methods are space-oriented, this research is designed to develop a place-oriented regionalization or clustering method that preserves major geopolitical boundaries as a key element of identity and attachment.

Boundaries are important for maintaining the familiarity and hierarchy in a map (Lloyd and Steinke 1986). Geographic, cartographic, and psychological research has shown that map readers organize and process their spatial memory hierarchically in clusters, and rely on familiar features to interpret and understand map contents (McNamara, Hardy, and Hirtle 1989; Rittschof et al. 1996; Fotheringham and Curtis 1999; Jones et al. 2004) and spatial characteristics of the environment (Hirtle and Jonides 1985). Boundary plays an interrelated role in psychological and geographical compartmentalization (Sack 2003). Boundaries and bordering are also discussed in the context of calculable space, place, security, and territory (Rose-Redwood 2012). Geographic data are provided in a hierarchical way using units of state, county, census tract, and others, and boundaries of these units serve as an essential reference to familiarity. In addition to geopolitical units, it is also important to keep other geographic boundaries, within which underlying forces and processes under study differ. For example, in F. Wang, Guo, and McLafferty (2012), a regionalization method is applied to areas of distinctive urbanicity categories separately to preserve their boundaries.

Population size usually varies substantially across areas at the same level. In public health data analysis and dissemination, it is often desirable to obtain regions of comparable population (F. Wang, Guo, and McLafferty 2012). Areas of large population need to be decomposed to gain more spatial variability, and areas of small population need to be merged to protect geoprivacy. Would keeping upper level geographic boundaries make a regionalization method more place-oriented? For example, if the data are available at the census tract level, should county boundaries be preserved as much as possible in regionalization? This research proposes a place-oriented, mixed-level regionalization (MLR) or spatial clustering method. Specifically, the conceptualization of “place = space + identity + attachment” is addressed twofold. As boundary serves as an important identifier for places, our method aims to preserve the boundaries of upper level geographic units and minimize operations at the lower level. *Attachment* is accounted for by imposing a constraint of attributive similarity on the regionalization method. By doing so, the resulting regions still look familiar or recognizable.

When working with health data, geoprivacy is a common concern that leads to aggregating individual data to area units. The overall objective of this research is to develop a regionalization method for disseminating and analyzing health data accounting for not only commonly considered spatial compactness and attributive homogeneity but also familiarity and geoprivacy. This description serves as an overarching problem statement, and detailed settings are illustrated by a case study of health (specifically cancer) data. It can certainly benefit any studies that involve the small population problem, including crime analysis (F. Wang and O’Brien 2005).

Background on Related Methods

Spatial clustering methods in general seek to balance two factors, spatial compactness and attributive homogeneity, in the derived regions.

Regionalization for Geoprivacy Protection

To protect patients’ location privacy, various geographic masking methods are developed such as affine transformation, aggregation, and random perturbations (Kwan, Casas, and Schmitz 2004; Rushton et al. 2006). Affine transformation preserves collinearity so that points on a line will still be on a line after the transformation and ratios of distances. Random perturbation, often described as *jittered*, is the disturbance of usual and regular courses (or locations). Regionalization can be viewed as an aggregation technique widely used in analysis and presentation of health data. Many clustering methods adopt two major criteria: entity connectivity and hierarchical leveling. Experimental analysis has shown that connectivity is empirically more valid than hierarchical leveling because the former reflects the “chunking” pattern or behavior in data and thus intuitively matches human perceptions of importance, whereas the latter shows no significant results as a valid basis for clustering (Moody 2003). Moody's finding reduces the requirement on human judgment to identify “nuclei” or central, major, dominant, root, or primary entities. A recent example is the work of F. Wang, Guo, and McLafferty (2012), which uses an automated regionalization method in GIS originally developed by Guo (2008) to construct geographic areas with sufficient population for cancer data analysis.

Spatial clustering or regionalization is a common approach for preparing health data for public dissemination, and its implementation strives to balance between preserving spatial variation and protecting privacy. Typical concerns include homogeneous neighborhood, delineating boundaries for convenient policymaking and administration, and preserving spatial patterns and distributions of both health and population data (Heikkila 1996; Clapp and Wang 2006; Stafford, Duke-Williams, and Shelton 2008).

Spatially adaptive filtering (Rushton 2003; Tiwari and Rushton 2005, 2010; Carlos et al. 2010; Cromley and McLafferty 2012a) is one of the most popular GIS techniques for understanding the spatial variation in data and maintaining a minimum threshold value within each filter. It also balances the spatial variation and geoprivacy of data. Conceptually, using a threshold to determine the size of the adaptive filter is quite similar to the minimum cancer count and population criteria in this study. Regionalization uses a different approach to generate a nonoverlapping partition, however, which cannot be guaranteed in a regular implementation of adaptive spatial filters.

Modified Scale-Space Clustering

Modified scale-space clustering (MSSC; Mu and Wang 2008) is developed based on scale-space theory (Witkin 1983; Koenderink 1984), an earlier algorithm (Wong 1993) and applications of the theory in remote sensing and GIS (Wong 1993; Leung, Zhang, and Xu 2000; Luo et al. 2002; Ciucu et al. 2003; M. Wang, Luo, and Zhou 2005). Using analogies of solid melting and viewing images, scale-space theory treats “scale”—corresponding to temperature in solid melting or distance in viewing images—as a parameter in describing the processes and phenomena. With the increase of scale (as temperature in the melting algorithm), a piece of metal will melt into liquid but not evenly, showing a clustering pattern; with the increase of scale (as distance in the blurring algorithm), the same image can reveal different levels of generalizations and details, or different cluster centers. The extraction of scale as a factor in modeling and analysis is in line with some work by geographers (Tobler 1989; Batty and Xie 1994; Kwan and Weber 2003; Lam 2004).

Mathematically, a Gaussian function can be used to formulate a two-dimensional (2D) scale-space (Wong 1993; Leung, Zhang, and Xu 2000). Because many GIS data are polygons with multiple attributes, Luo et al. (2002) and Mu and Wang (2008) modified the equations to capture the attributes and neighboring relationships. MSSC is an unsupervised hierarchical clustering method that considers both attributive homogeneity and spatial contiguity. Starting from small areas (e.g., census tracts), MSSC runs multiple iterations and eventually merges all tracts into one cluster, just like melting and blurring processes. The user can decide which round of clustering results to adopt. MSSC has been applied to study homicide patterns in Chicago (Mu and Wang 2008) and late-stage breast cancer distribution in Illinois (Mu and Wang 2008; Mu, Wang, and McLafferty 2010).

Peano Curve Algorithm

Peano curves are space-filling curves first introduced by Italian mathematician Giuseppe Peano (1890). A variation and more complicated form is called Hilbert curves because Hilbert (1891) visualized the spacefilling idea described in Peano curves and later referred to it as “topological monsters”(Bartholdi and Platzman 1988). Peano and Hilbert curves have been used to find all-nearest-neighbors (Chen and Chang 2011) and spatial ordering of geographic data (Guo and Gahegan 2006). Conceptually, Peano curves use algorithms to assign spatial orders to points in 2D space and map the points onto one-dimensional (1D) space. As shown in Figure 1, the spatial order of each point is calculated and labeled. Following the spatial order, a point in 2D space can be mapped onto the 1D line underneath, and the connected line in 2D space (on the right) is the Peano curve. Following the spatial orders along the 1D line, spatial clustering can be achieved by classification with many methods. Figure 1 shows an example of quartile clustering.

The algorithm to calculate the spatial order has many forms. A generic space-filling heuristic algorithm developed by Bartholdi and Platzman (1988) has been widely used in the GIS community such as in older Arc/Info commands of SPATIALORDER and COLOCATE (F. Wang and O'Brien 2005) and a more recent public domain ArcGIS Python script (F. Wang and O'Brien 2005; Mandloi 2009). We adopt this algorithm for our method, and the key function in this algorithm is presented here in Python language:

```
# Calculate and return the Peano curve coordinate for
# a pair of given x, y values. x and y are standardized locations
# in the unit square that are transformed from original
# point (x, y). k is the first k binary digits of x and y.
def Peano(x, y, k):
if (k = 0 or (x = 1 and y = 1)):
return 0.5
if x ≤ 0.5:
if y ≤ 0.5:
quad = 0
else:
quad = 1
if y ≤ 0.5:
quad = 3
else:
quad = 2
subpos = Peano(2 * abs(x−0.5), 2 * abs(y−0.5), k−1)
if (quad = 1 or quad = 3):
subpos = 1−subpos
return GetFractionalPart((quad + subpos−0.5)/4.0)
```

For spatial clustering, there are numerous methods to achieve spatial connectivity and compactness, and the space-filling curve is only one of them. Even for the space-filling curve approach, there are quite a few algorithms such as the previously mentioned Peano curve and Hilbert curve, in addition to the Sierpinski curve and several extensions or variations (Bartholdi and Goldsman 2001, 2004). Overall, they all perform satisfactorily in terms of spatial connectivity, with differences mainly in construction algorithms and applications. We decide to adopt the Peano curve algorithm for the following reasons: It is one of the earliest and most basic forms in the space-filling curve family and has been commonly used in geography, health, and other fields; the construction algorithm is simple and elegant; there are open source codes available; and it has been used in existing GIS tools.

The Modified Peano Curve Algorithm

Our place-oriented, MLR method is built on two previous clustering methods: modifying the Peano curve algorithm (Bartholdi and Platzman 1988; Mandloi 2009), termed MPC, to achieve spatial compactness, and integrating it with the MSSC (Mu and Wang 2008) to address attributive homogeneity. Therefore, it is a hybrid method. Additionally, it also considers “mixed levels” of geographic units to maximize the recognizability of resulting regions and incorporates empirical criteria such as cancer count and population count into the regionalization process.

Space-filling curves have several well-known problems such as arbitrary jumps and predefined turns. Modifications are proposed to address these problems, and therefore the method is termed the modified Peano curve algorithm. Furthermore, the MPC algorithm needs to account for previously discussed issues related to health data dissemination and analysis. For convenience of illustration, we use the case study of cancer data in Louisiana to be specific.

Setting Clustering Boundaries on the Peano Curve with Threshold Population (e.g., Lower Limits of Population and Cancer Count

Clustering follows the values of spatial order along the Peano curve with breaking points that are defined by a threshold population size. Many classification methods can be applied here. Among the four general categories (exogenous, arbitrary, idiographic, and serial), ours uses the health data release criteria to decide class breaks and thus is considered exogenous. Through iterations in programming, each cluster satisfies the three criteria:

- Ascending spatial order.
- Population ≥ 20,000.
- Cancer count > 15.

A user can change the thresholds adaptable to specific applications. Here, the example of population ≥ 20,000 is based on the Summary of the Health Insurance Probability and Accountability Act (HIPAA) Privacy Rule (U.S. Department of Health and Human Services 2003), and a minimum cancer count >15 is used by the State Cancer Profiles (National Cancer Institute 2013).

Combining Spatial Weight Matrix with Spatial Orders to Address the Jumping Problem—Disconnected Members in a Cluster

Figure 2 shows an area with forty-one units and their spatial orders are calculated and ranked with the Peano curve algorithm. The labels on the map are the ranks, and the 1D dot graph at the bottom reflects the spatial order values. Although features close in 2D space tend to have adjacent spatial order values on the 1D line such as units ranked 36 and 37, there are exceptions—for instance, there are two other spatial order values between units 11 and 14 on the 1D line. Clustering performed by simply following spatial orders would lead to disconnected cluster members such as units 11 and 12. To address this problem, we bring in the spatial weight matrix, where 1 means that two spatial objects are adjacent and 0 otherwise. When looking for the next member in a cluster after unit 11, the spatial order suggests unit 12, but the spatial weight matrix disqualifies 12 by the connectivity requirement (i.e., spatial weight between 11 and 12 is 0). The search goes on until unit 14, which is the adjacent unit with the closest spatial order.

Forcing a Cluster Membership for Unclaimed or Loose-End Units

Figure 3 shows that units 14 and 15 are left alone at the end of iterations because the aggregation of the two does not satisfy the population and cancer count criteria. They are forced to merge with cluster 1, the first cluster satisfying the Rook neighborhood criteria.

Balancing Spatial Compactness and Attributive Homogeneity in Clustering

Although the preceding three steps have modified the Peano curve algorithm to meet the requirements of spatial connectivity and compactness with additional rations such as population and cancer count, attributive homogeneity has not been counted for in this clustering procedure. We bring in previously developed MSSC to address this issue.

Attributive homogeneity has been measured using the weighted aggregation of factor scores based on factor analysis (Mu and Wang 2008; F. Wang, Guo, and McLafferty 2012), and the weights are quantified with proportions of eigenvalues, representing the captured variances by factors. The closer the aggregated scores are, the more attributively similar the two objects. The weighted aggregation score is normalized and serves as attributive order (o_{ai} in Equation 1) similar to the spatial order (o_{si} in Equation 1) concept from the Peano curve algorithm. Users can determine weighting factors for spatial (Peano) and attributive (MSSC) considerations such as

$$o_i = w_s o_{si} + w_a o_{ai} \quad (1)$$

where o_i is the integrated clustering order value for unit *i*, w_s is the weighting factor of spatial consideration, and $w_s > 0$, o_{si} is the normalized spatial order from the Peano curve algorithm, w_a is the weighting factor of attributive consideration, and o_{ai} is the normalized attributive order based on MSSC, and subject to $w_s + w_a = 1$.

In practice, the value of w_s is set to be larger than 0, and preferably larger than 50 percent, as it is the entity connectivity emphasized here (Moody 2003). The MPC method is implemented in ArcGIS 10.1 with Python and ArcPy scripts.

The Place-Oriented, Mixed-Level Regionalization Method

The aforementioned MPC method is applied to areas of multiple levels, and therefore termed *mixed-level regionalization* (or *clustering*). For illustration, we use the 2006 cancer data in Louisiana to demonstrate the general steps, summarized in a flowchart in Figure 4.

In Step 1, the method starts with surveying the data at the upper level of area units. For instance, as we are going to work with data at the census tract level, we begin with one level up (i.e., the county level) and examine whether an entire county meets the criteria of population thresholds discussed in the previous section. If so, the county itself forms a cluster; otherwise, we decide whether to further aggregate it with other counties or disaggregate it to tract-level clusters. In Louisiana, a parish is equivalent to a county in other states. There are three scenarios—disaggregation, no action, and aggregation—color-coded as red, green, and blue in all subsequent tables and figures, respectively.

1. Disaggregation: Both population and cancer count criteria overflow and the values are more than twice the tresholds, so there is a need to break a parish into subregions.
2. No action: If a parish meets both criteria with neither one twice overflown, the parish serves as a cancer data release (analysis) region with no action required. If one of the measures is more than twice the limit and the other is not, no further action is needed either.
3. Aggregation: If a parish has at least one unsatisfied criterion, it needs to be aggregated to adjacent parishes to reach the criteria. There are two types of aggregation.
   - Minimum population aggregation: Because population already exceeds 20,000, parishes with minimum population should be prioritized when merging.
   - Minimum cancer-count aggregation: Because cancer count overflows, parishes with minimum cancer count should be first considered when merging to the parish to form a cluster.

Figure 5 shows the three scenarios in a total of sixty-four parishes in Louisiana: Twenty-nine need to be divided into subregions, nineteen need no action, and sixteen parishes need to be aggregated.

Step 2 deals with the first scenario, disaggregation. Demonstrated with the same example as earlier, each of the twenty-nine parishes is divided into subregions using MPC. They are parishes with major urban centers such as New Orleans, Baton Rouge, Shreveport, Metairie, and Lafayette. Before dividing a parish into subregions, the maximum number of subregions is calculated for each parish such as

$$max_{Ni} = min\left(int\left(\frac{P_i}{P}\right), int\left(\frac{C_i}{C}\right), T_i\right) \quad (2)$$

where *i* is the the parish of interest, max_{Ni} is the maximum number of subregions in parish *i*, P_i is the population of parish *i*, *P* is the population threshold (set as 20,000 in method demonstration), C_i is the cancer count in parish *i*, *C* is the cancer count threshold (set as 15 in method demonstration), T_i is the number of tracts in parish *i*, *int* is the function to round down a number to the nearest integer, and *min* is the minimum function.

The hybrid method of MPC with a tailored Equation 1 (e.g., w_s = 90% and w_a = 10%) is applied to group tracts to meet the criteria of spatial connectivity, attributive homogeneity, and population and cancer count. All clusters derived from this step are subparish clusters made of census tracts within a parish. As discussed earlier, normalized attributive order is measured with attributive homogeneity, which is approximated with weighted aggregation of factor scores. A total of eleven socioeconomic variables are chosen to capture the sociodemographic structure of area units (tracts or counties; F. Wang, Guo, and McLafferty 2012) with a factor structure shown in Table 1.

Step 3 is for the second scenario, no action. No additional actions are needed, and thus each of these nineteen parishes forms a *single-parish cluster* (Figure 5).

Step 4 is for the group parishes in the third scenario, aggregation. The tailored MPC used in step 2 is applied again but at the level of parish instead of census tract. There are sixteen rural parishes in this category (Figure 5), each of which has a small number of census tracts ranging from two to five. All clusters in this step are multiparish ones.

Step 5, the third procedure in the section on the MPC method (forcing cluster membership), is followed to tackle isolated clusters or loose ends at different levels. Isolated tracts are merged to a nearby subparish cluster composed of tracts, and isolated parishes are merged to a nearby cluster of single-parish (from step 3) or multiparish (from step 4).

Figure 6 and Table 2 illustrate and summarize the final results. There are sixty-four parishes or 1,106 census tracts in Louisiana, and the MLR has yielded a total of 165 clusters. The underneath parish boundary map and boundary change map are provided for reference. It is observed that the MLR preserves upper level geographic units, parishes in this case, to a great degree. In terms of area or number of parishes, one third of the state has its final derived regions identical to the parishes. In terms of boundary length, only 4 percent of the parish boundaries are removed to form multiparish clusters, and 26 percent of the cluster boundaries are added for tract-level clusters. With 96 percent of parish boundaries still visible on the final clusters, the MLR generates regions of a high degree of familiarity. As shown in Figure 6 and Table 2:

- 140 tract-level clusters (within parish) are derived from twenty-nine disaggregation parishes, which have 926 tracts.
- Among the nineteen no-action parishes, sixteen remain single-parish clusters, and the other three are lost to step 5 for addressing loose-end parishes and become part of the multiparish clusters.
- Nine multiparish clusters are constructed from sixteen aggregation parishes along with the three “no-action” parishes.

Discussion

Temporal Variations and Selection of Threshold Values

Figures 7A and 7B show overall cancer incidences in Louisiana from 2002 to 2006 at census tract and parish levels, respectively. No major temporal changes are observed at either level. We use the five-year data to test the stability of the MLR method, and experiment with the threshold: population ≥ 20,000 and cancer count > 15 (Figure 8).

Although the temporal variations are negligible, the number and configurations of clusters vary in response to different threshold values of population and cancer count. To provide a general guideline to determine those values, we test multiple scenarios based on the 2006 data. The combinations of four population sizes (5,000, 10,000, 20,000, and 30,000) and six cancer counts (15, 100, 200, 300, 400, and 500) yield twenty-four scenarios. The MLR results of cluster types and numbers are shown in Table 3 and Figure 9.

Based on Figure 9A and 9B, the correlation between population and cancer count is high at both census tract and parish levels, characterized with a simple cancer/population ratio of 5 percent (i.e., about the average five-year cancer rate in the study area). Figure 9C, coupled with data in Table 3, demonstrates that the 5 percent equilibrium line divides the results into two zones. Below the line is termed the *population dominant zone*, showing little or no change in the total number of clusters and types as long as the population threshold stays the same. For example, when population threshold = 20,000, the corresponding cancer count = 100 on the equilibrium line, and any scenarios with cancer count < 100 will show quite similar results along the population = 20,000 vertical direction. Above the equilibrium line is termed the *cancer count dominant zone*, and resulting clusters will be the same or quite similar, with the same cancer count with different population thresholds. For instance, when cancer count is set to 200, the clusters stay the same for population values of 5,000, 10,000, 20,000, and 30,000. In other words, population and cancer count thresholds need to set along the equilibrium line representing the cancer rate in a study area. The rate understandably drops in a study focusing in a single year on a specific cancer type or demographic group.

Effect on Mapping

Figure 10 shows crude cancer rate maps in 2006 before (Figure 10A) and after (Figure 10B) applying the MLR method. When feasible, it is strongly advised to map age- and gender-adjusted cancer rates. Without access to cancer data by gender and age groups, the crude rates are used here for illustrating the effects of MLR method. The constraints are population ≥ 20,000 and cancer count > 15. The tract-level rate map shows a high variability of cancer rates, subject to the risk of the small population problem. The mixed-level rate map shows a much smoother pattern, yet with noticeable high-rate and low-rate clusters. Constructed regions by MLR have similar population sizes and cancer counts in clusters and are thus comparable. Population and cancer counts are also sufficiently large in the new regions and lead to more reliable cancer rate estimates, which permit meaningful mapping and subsequent exploratory spatial analysis (F. Wang, Guo, and McLafferty 2012).

Comparison Between MLR and Other Spatial Clustering Methods

Among various spatial clustering methods, *k*-mean (MacQueen 1967) and iterative self-organizing data analysis technique (ISODATA; Ball and Hall 1965) are widely used in GIS and remote sensing. Both can perform regionalization and be implemented easily in popular GIS or remote sensing software. *K*-mean clustering starts with defining *k* initial cluster centers, then assigns all objects to the closest centers and recalculates center locations. The process is iterated until center locations become unchanged. ISODATA is very similar to *k*-mean but does not require a predefined number of clusters at the beginning and thus is more flexible in cluster forming by splitting and merging (Jensen 1996). We take one of the MLR results (2006 data, with constraints of population > 20,000 and cancer count >15, 90 percent spatial + 10 percent attributive weighting) to compare with *k*-mean clustering, implemented by Grouping Analysis in ArcMap. ISODATA is mainly used for analysis of raster data in GIS and remote sensing software and therefore is not applicable to our case based on the vector data. The comparison to ISODATA is only conceptual.

Table 4 compares functionalities and outcomes of the three methods. MLR differs most in providing mixed-level results and weighing attributes and also without predefining cluster numbers and locations. Figure 11 shows the 165 clusters using either MLR or *k*-mean. Parish boundaries are provided at the bottom as a reference. The MLR best preserves the parish shapes and generates smoother clusters of more similar sizes than *k*-mean.

Conclusion

This research proposes the place-oriented MLR method based on two earlier methods: the MSSC derived from scale-space theory and the MPC based on the space-filling algorithm. The uniqueness of the MLR method lies in its ability to assign mixed-level cluster memberships. The method begins with a diagnosis of areas at the upper level (say, county) and classifies into (1) areas of large population that need to be decomposed (“disaggregation”) to gain more spatial variability, (2) areas of about the right size that require no further action, and (3) areas of small population that need to be merged (“aggregation”) to mask privacy of data. When the clustering is applied to those flagged as a disaggregation status, areas at the lower level are grouped within each upper level area (say, census tracts within each county). When the clustering is applied to those flagged as an aggregation status, areas at the upper level are grouped. It eliminates unnecessary boundary breaking in clustering and the amount of clustering operation is minimized. More important, upper level geographic unit boundaries are maximally preserved. The newly constructed regions have comparable base population, an important property in health data management and analysis. A cognitive and geo-psychological benefit of the method is that the resulting regions still look familiar or are recognizable. Therefore, the method is more place-oriented than space-oriented, as boundaries are essential for people's sense of places. It also addresses the small population problem with additional practical criteria (in this case, the minimum and approximately equal population) and generates maps of more reliable disease rates.

Some limitations of the method are inherited from its linkage to the space-filling curves with several widely known issues. Some of these are mitigated by bringing in a spatial weight matrix to remove the topological jumps and loose ends and integrating with the MSSC to account for both spatial and attributive aspects in the process. By incorporating additional criteria such as threshold population and cancer count into the clustering process, the MLR method delivers a promising result in constructing health data release regions that preserve major administrative boundaries. The concept employed here, making the best sense out of a rather arduous analytical task, echoes the thinking of human, social, and cultural geographers. The form of regions reflects the process and connection among people, space, and place. The emphasis of keeping upper unit boundaries to increase familiarity makes the clustering method more human oriented than computer oriented.

Acknowledgments

Funding: This research was supported by a Surveillance, Epidemiology, and End Results (SEER) Rapid Response Surveillance Studies (RRSS) program of the National Cancer Institute (NCI) under the grant NCI-N01-PC-54402. Points of view or opinions in this article are those of the authors and do not necessarily represent the official position or policies of NCI.

References

Adams, PC.; Hoelscher, SD.; Till, KE. Textures of place: Exploring humanist geographies. Minneapolis: University of Minnesota Press; 2001.

Ball, GH.; Hall, DJ. ISODATA, a novel method of data analysis and pattern classification. Menlo Park, CA: Stanford Research Institute; 1965.

Bartholdi JJ, Goldsman P. Continuous indexing of hierarchical subdivisions of the globe. International Journal of Geographical Information Science. 2001; 15:489–522.

Bartholdi JJ, Goldsman P. The vertex-adjacency dual of a triangulated irregular network has a Hamiltonian cycle. Operations Research Letters. 2004; 32:304–08.

Bartholdi JJ III, Platzman LK. Heuristics based on spacefilling curves for combinatorial problems in Euclidean space. Management Science. 1988; 34:291–305.

Batty M, Xie Y. From cells to cities. Environment and Planning B: Planning & Design. 1994; 21:S31–S38.

Carlos H, Shi X, Sargent J, Tanski S, Berke E. Density estimation and adaptive bandwidths: A primer for public health practitioners. International Journal of Health Geographics. 2010; 9:39. [PubMed: 20653969]

Carpenter, E.; Varley, FH.; Flaherty, RJ. Eskimo. Toronto: University of Toronto Press; 1959.

Chen HL, Chang YI. All-nearest-neighbors finding based on the Hilbert curve. Expert Systems with Applications. 2011; 38:7462–75.

Ciucu, M.; Heas, P.; Datcu, M.; Tilton, JC. Scale space exploration for mining image information content. In: Zaiane, OR.; Simoff, S.; Djeraba, C., editors. Mining multimedia and complex data. New York: Springer; 2003. p. 118-33.

Clapp JM, Wang Y. Defining neighborhood boundaries: Are census tracts obsolete? Journal of Urban Economics. 2006; 59:259–84.

Cromley, EK.; McLafferty, S. GIS and public health. New York: Guilford; 2012a. Analyzing spatial clustering of health events; p. xxiv

Cromley, EK.; McLafferty, S. GIS and public health. New York: Guilford Press; 2012b.

Fotheringham AS, Curtis A. Regularities in spatial information processing: Implications for modeling destination choice. The Professional Geographer. 1999; 51:227–39.

Guo D. Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). International Journal of Geographical Information Science. 2008; 22:801–23.

Guo D, Gahegan M. Spatial ordering and encoding for geographic data mining and visualization. Journal of Intelligent Information Systems. 2006; 27:243–66.

Hanson CE, Wieczorek WF. Alcohol mortality: A comparison of spatial clustering methods. Social Science & Medicine. 2002; 55:791–802. [PubMed: 12190271]

Heikkila EJ. Are municipalities Tieboutian clubs? Regional Science and Urban Economics. 1996; 26:203–26.

Hilbert D. Ueber die stetige Abbildung einer Line auf ein Fl€achenst€uk [On continuous mapping of a line onto a planar surface]. Mathematische Annalen. 1891; 38:459–60.

Hirtle S, Jonides J. Evidence of hierarchies in cognitive maps. Memory & Cognition. 1985; 13:208–17. [PubMed: 4046821]

Jensen, JR. Introductory digital image processing: A remote sensing perspective. Englewood Cliffs, NJ: Prentice-Hall; 1996.

Jones A, Blake C, Davies C, Scanlon E. Digital maps for learning: A review and prospects. Computers & Education. 2004; 43:91–107.

Koenderink JJ. The structure of images. Biological Cybernetics. 1984; 50:363–70. [PubMed: 6477978]

Kwan MP, Casas I, Schmitz B. Protection of geoprivacy and accuracy of spatial information: How effective are geographical masks? Cartographica: The International Journal for Geographic Information and Geovisualization. 2004; 39:15–28.

Kwan MP, Weber J. Individual accessibility revisited: Implications for geographical analysis in the twenty-first century. Geographical Analysis. 2003; 35:341–53.

Lam, NSN. Fractals and scale in environmental assessment and monitoring. In: Sheppard, ES.; McMaster, RB., editors. Scale and geographic inquiry: Nature, society, and method. Malden, MA: Blackwell; 2004. p. 23-40.

Leung Y, Zhang JS, Xu ZB. Clustering by scale-space filtering. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000; 22:1396–1410.

Lloyd R, Steinke T. The identification of regional boundaries on cognitive maps. The Professional Geographer. 1986; 38:149–59.

Luo J, Zhou C, Leung Y, Zhang Y, Huang Y. Scale-space theory based regionalization for spatial cells. ACTA Geographica Sinica. 2002; 57:167–73.

MacQueen, J. Some methods for classification and analysis of multivariate observations. In: Le Cam, LM.; Neyman, J., editors. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, CA: University of California Press; 1967. p. 281-97.

Mandloi, D. Partitioning tools. Redlands, CA: Esri; 2009.

McNamara TP, Hardy JK, Hirtle SC. Subjective hierarchies in spatial memory. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1989; 15

Moody, D. Entity connectivity vs. hierarchical levelling as a basis for data model clustering: An experimental analysis. In: Marık, V.; Retschitzegger, W.; Štepánková, O., editors. Database and expert systems applications. Berlin: Springer; 2003. p. 77-87.

Moore S, Daniel M, Gauvin L, Dub L. Not all social capital is good capital. Health & Place. 2009; 15:1071–77. [PubMed: 19482506]

Mu L, Wang F. A scale-space clustering method: Mitigating the effect of scale in the analysis of zone-based data. Annals of the Association of American Geographers. 2008; 98:85–101.

Mu, L.; Wang, F.; McLafferty, S. Analyzing spatial patterns of late-stage breast cancer in Chicago region: A modified scale-space clustering approach. In: Jiang, B.; Yao, X., editors. Geospatial analysis and modeling of urban structure and dynamics. Dordrecht, The Netherlands: Springer Science+Business Media; 2010. p. 355-72.

National Cancer Institute. [last accessed March 19, 2013] State cancer profiles. 2013. http://statecancerprofiles.cancer.gov/

Nelson MJ, Warden C, Griffiths D, Zive D, Schmidt T, Hedges JR, Daya M, Newgard CD. A geospatial analysis of persons opting out of an exception from informed consent out-of-hospital clinical trial. Resuscitation. 2009; 80:89–95. [PubMed: 19010580]

Neuberger JS, Lynch HT. Cancer of the esophagus: A model of causation. Medical Hypotheses. 1982; 9:337–42. [PubMed: 7144639]

Oliver MN, Smith E, Siadaty M, Hauck FR, Pickle LW. Spatial analysis of prostate cancer incidence and race in Virginia, 1990–1999. American Journal of Preventive Medicine. 2006; 30:S67–S76. [PubMed: 16458792]

Peano G. Sur une courbe qui remplit toute en aire plaine [On a curve which completely fills a planar surface]. Mathematische Annalen. 1890:36.

Rittschof KA, Stock WA, Kulhavy RW, Verdi MP, Johnson JT. Learning from cartograms: The effects of region familiarity. Journal of Geography. 1996; 95:50–58.

Rose-Redwood R. With numbers in place: Security, territory, and the production of calculable space. Annals of the Association of American Geographers. 2012; 102:295–319.

Rushton G. Public health, GIS, and spatial analytic tools. Annual Review of Public Health. 2003; 24:43–56.

Rushton G, Armstrong MP, Gittler J, Greene BR, Pavlik CE, West MM, Zimmerman DL. Geocoding in cancer research: A review. American Journal of Preventive Medicine. 2006; 30:S16–S24. [PubMed: 16458786]

Sack, RD. Conceptions of space in social thought: A geographic perspective. Minneapolis: University of Minnesota Press; 1980.

Sack, RD. A geographical guide to the real and the good. New York: Routledge; 2003.

Schootman M, Jeff DB, Gillanders WE, Yan Y, Jenkins B, Aft R. Geographic clustering of adequate diagnostic follow-up after abnormal screening results for breast cancer among low-income women in Missouri. Annals of Epidemiology. 2007; 17:704–12. [PubMed: 17574437]

Shishehbor MH, Gordon-Larsen P, Kiefe CI, Litaker D. Association of neighborhood socioeconomic status with physical fitness in healthy young adults: The Coronary Artery Risk Development in Young Adults (CARDIA) study. American Heart Journal. 2008; 155:699–705. [PubMed: 18371479]

Stafford M, Duke-Williams O, Shelton N. Small area inequalities in health: Are we underestimating them? Social Science & Medicine. 2008; 67:891–99. [PubMed: 18599174]

Szwarcwald CL, Andrade CL, Td, Bastos FI. Income inequality, residential poverty clustering and infant mortality: A study in Rio de Janeiro, Brazil. Social Science & Medicine. 2002; 55:2083–92. [PubMed: 12409122]

Tiwari, C.; Rushton, G. Developments in spatial data handling. Berlin: Springer; 2005. Using spatially adaptive filters to map late stage colorectal cancer incidence in Iowa; p. 665-76.

Tiwari C, Rushton G. A spatial analysis system for integrating data, methods and models on environmental risks and health outcomes. Transactions in GIS. 2010; 14:177–95.

Tobler, W. Frame independent spatial analysis. In: Goodchild, MF.; Gopal, S., editors. The accuracy of spatial databases. London: Taylor & Francis; 1989. p. 115-22.

Tuan, YF. Topophilia: A study of environmental perception, attitudes, and values Englewood Cliffs. NJ: Prentice-Hall; 1974.

Tuan, YF. Space and place: The perspective of experience. Minneapolis: University of Minnesota Press; 1977.

Tuan, YF. Humanist geography: An individual's search for meaning. Staunton, VA: George F Thompson; 2012.

U.S. Department of Health and Human Services. [last accessed 19 March 2013] Summary of the HIPAA privacy rule. 2003. http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/privacysummary.pdf

Wang F, Guo D, McLafferty S. Constructing geographic areas for cancer data analysis: A case study on late-stage breast cancer risk in Illinois. Applied Geography. 2012; 35:1–11. [PubMed: 22736875]

Wang, F.; O’Brien, V. Constructing geographic areas for analysis of homicide in small populations: Testing the herding-culture-of-honor proposition. In: Wang, F., editor. GIS and crime analysis. Hershey, PA: Idea Group; 2005. p. 84-100.

Wang, M.; Luo, JC.; Zhou, CH. Linear belts mining from spatial database with mathematical morphological operators. In: Li, X.; Wang, S.; Dong, ZY., editors. Advanced data mining and applications. New York: Springer; 2005. p. 769-76.

Witkin, AP. Scale-space filtering. In: Bundy, A., editor. The 8th International Joint Conference of Artificial Intelligence. Karlsruhe, Germany: William Kaufmann; 1983. p. 1019-22.

Wong YF. Clustering data by melting. Neural Computation. 1993; 5:89–104.

Yiannakoulias N. Spatial aberration vs. geographical substance: Representing place in public health surveillance. Health & Place. 2011; 17:1242–48. [PubMed: 21862379]

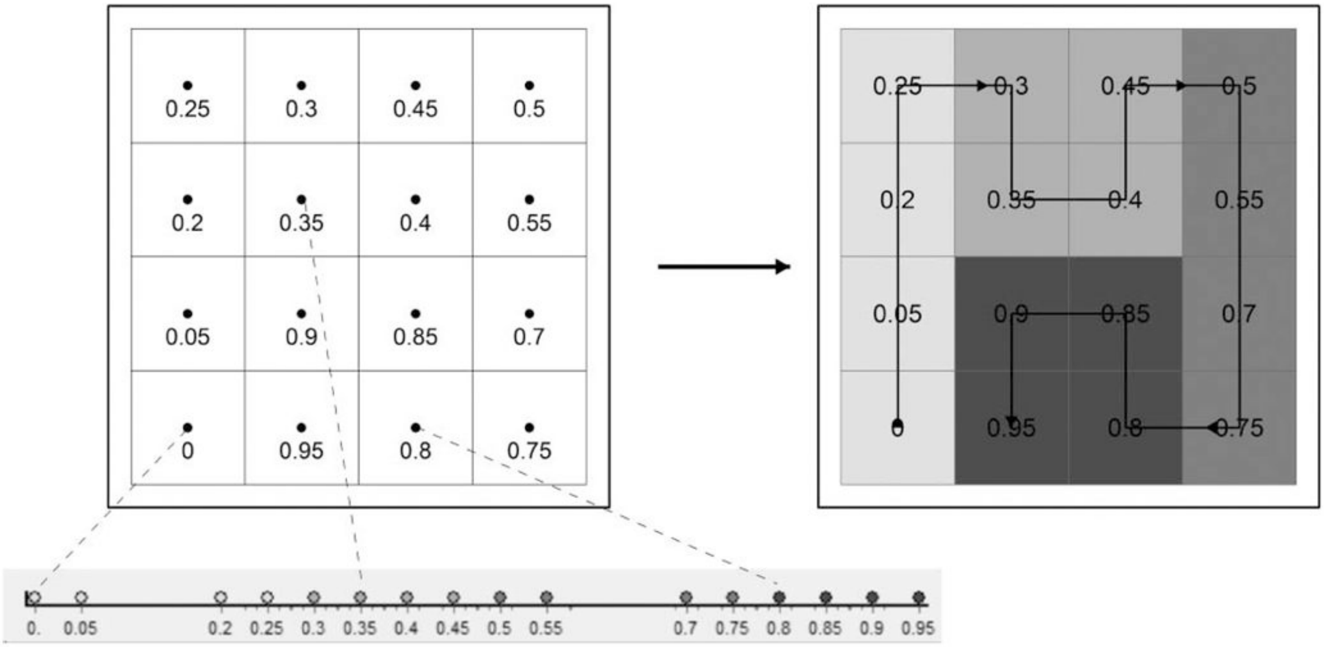
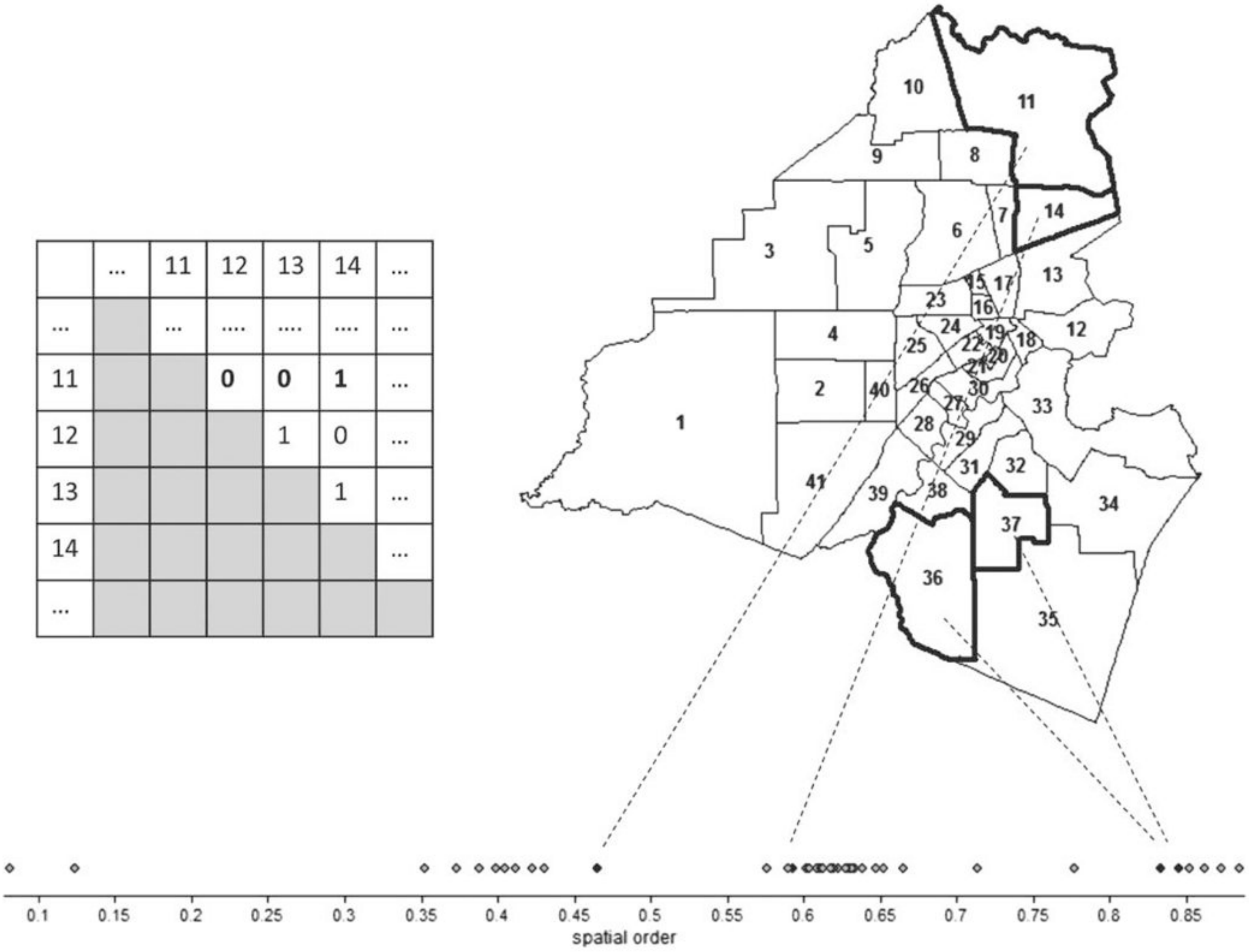


Figure 1.

Spatial order by Peano curve.



	...	11	12	13	14	...
...		...	....	....	....	...
11			**0**	**0**	**1**	...
12				1	0	...
13					1	...
14						...
...						

Figure 2.
Jumping problem in Peano curve algorithm.

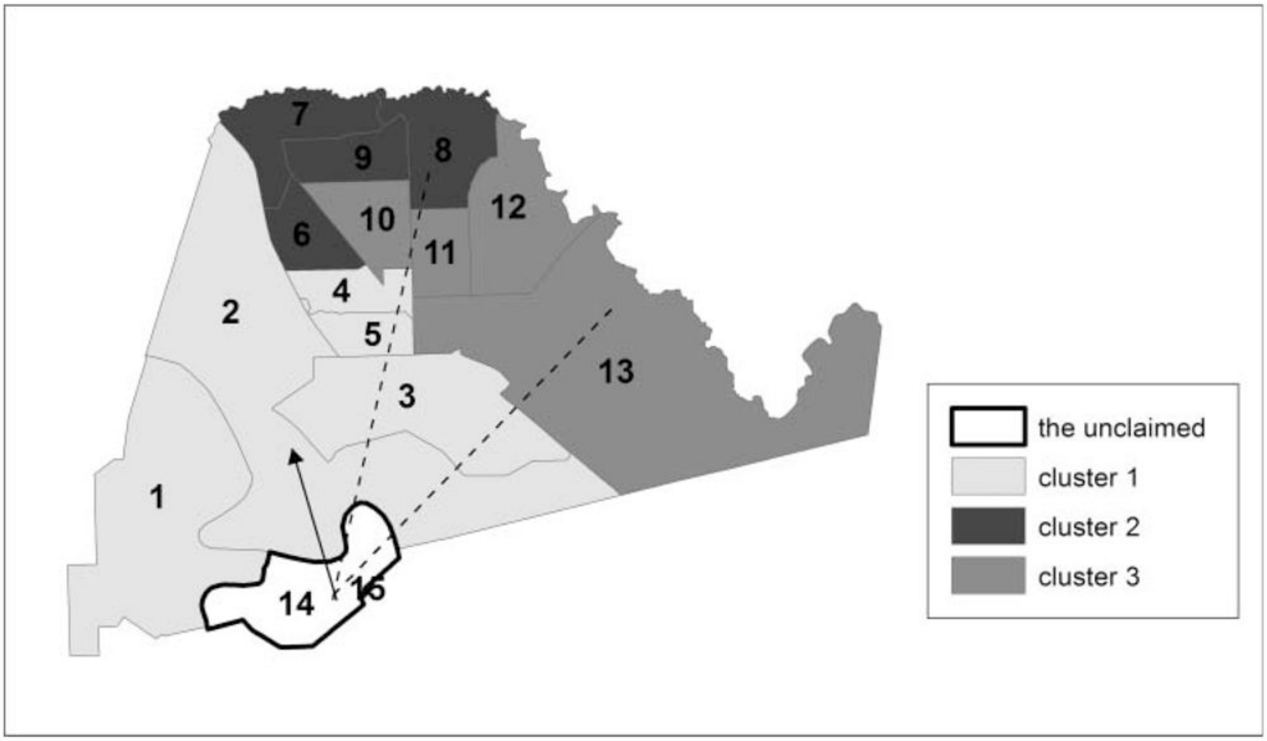


Figure 3.
Unclaimed units after the first round of clustering.

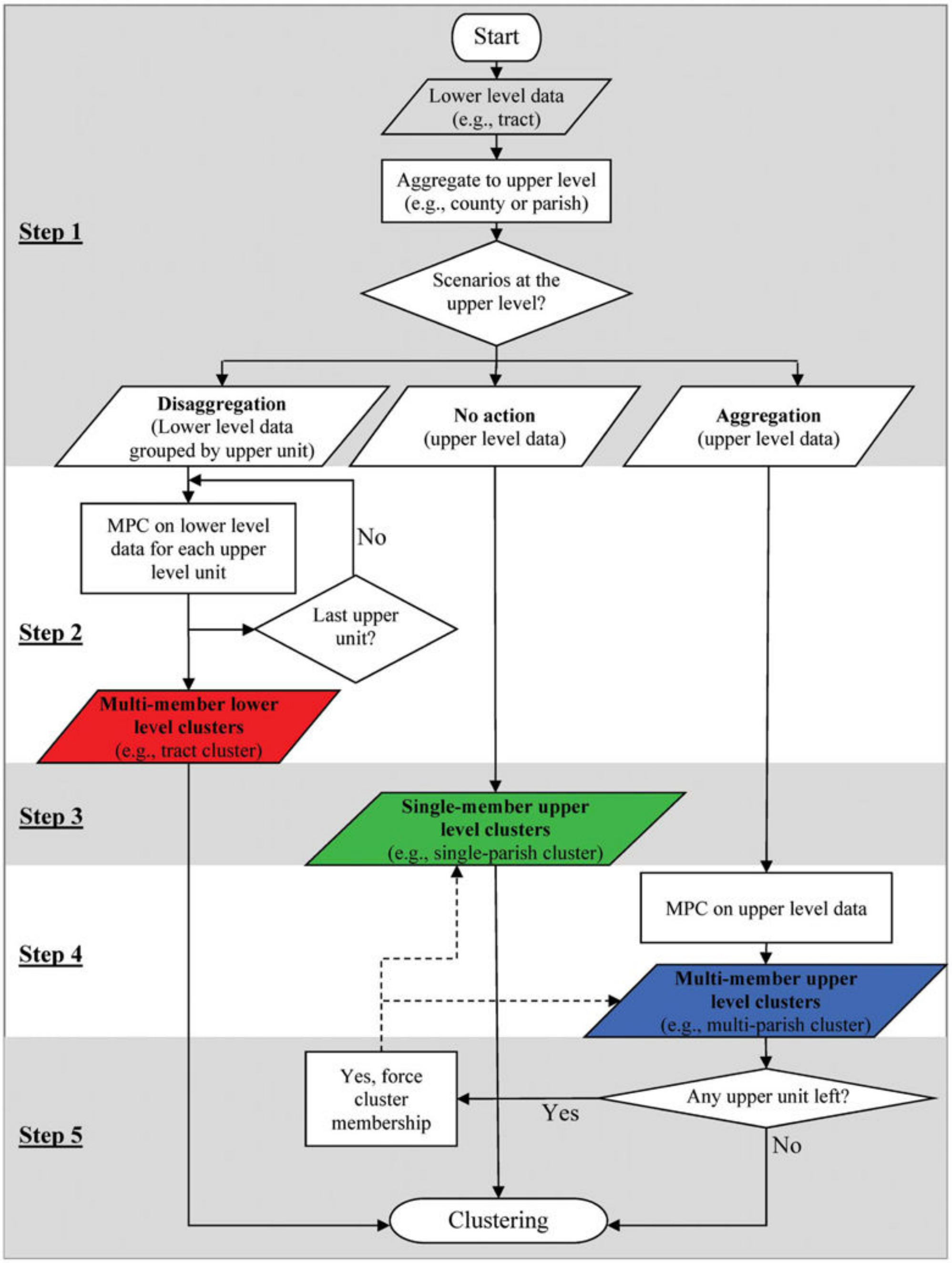
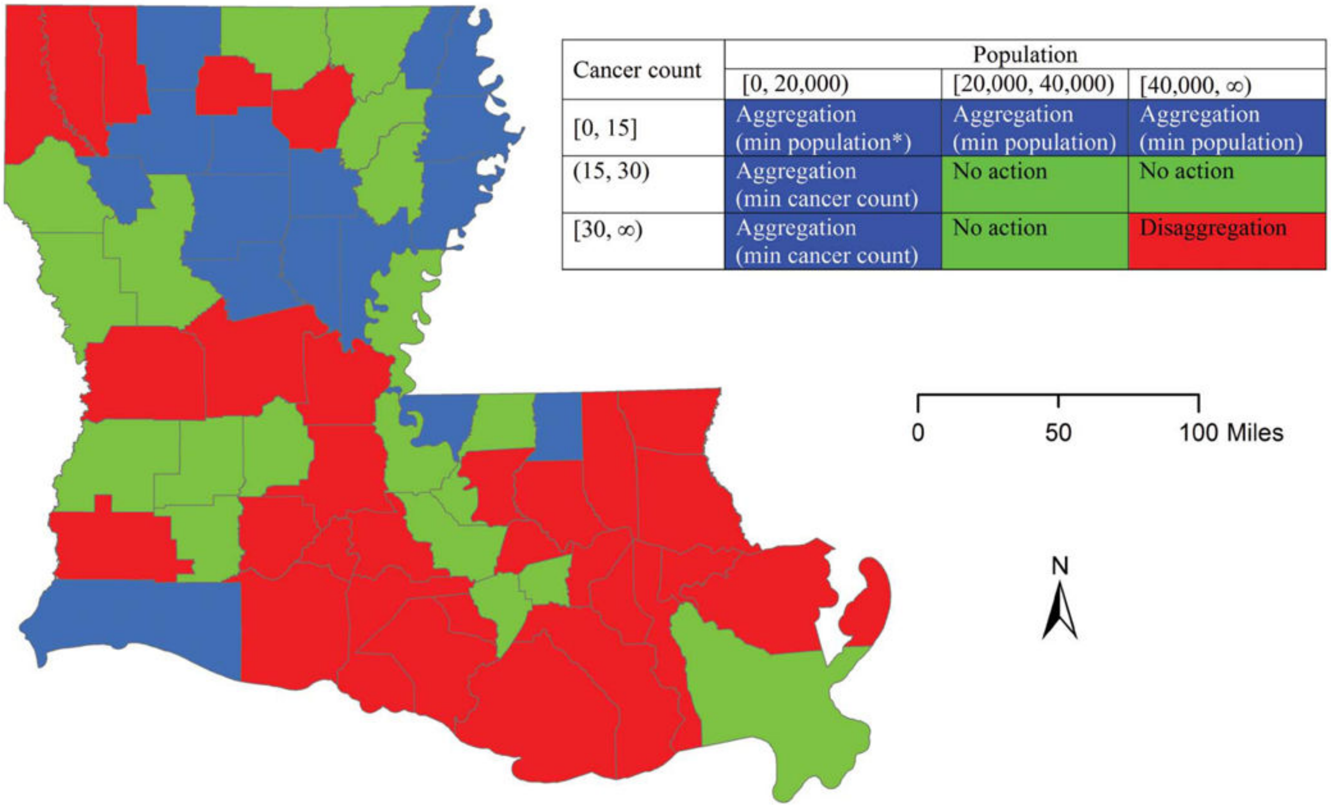

Start
Lower level data (e.g., tract)
Aggregate to upper level (e.g., county or parish)
Step 1
Scenarios at the upper level?
Disaggregation (Lower level data grouped by upper unit)
No action (upper level data)
Aggregation (upper level data)
MPC on lower level data for each upper level unit
No
Step 2
Last upper unit?
Multi-member lower level clusters (e.g., tract cluster)
Step 3
Single-member upper level clusters (e.g., single-parish cluster)
Step 4
MPC on upper level data
Multi-member upper level clusters (e.g., multi-parish cluster)
Step 5
Yes, force cluster membership
Yes
Any upper unit left?
No
Clustering

Figure 4.

Flowchart of the mixed-level regionalization (MLR) method.



Cancer count	Population		
	[0, 20,000)	[20,000, 40,000)	[40,000, ∞)
[0, 15]	Aggregation (min population*)	Aggregation (min population)	Aggregation (min population)
(15, 30)	Aggregation (min cancer count)	No action	No action
[30, ∞)	Aggregation (min cancer count)	No action	Disaggregation

Figure 5.

Three scenarios for parishes in Louisiana by mixed-level regionalization (MLR).

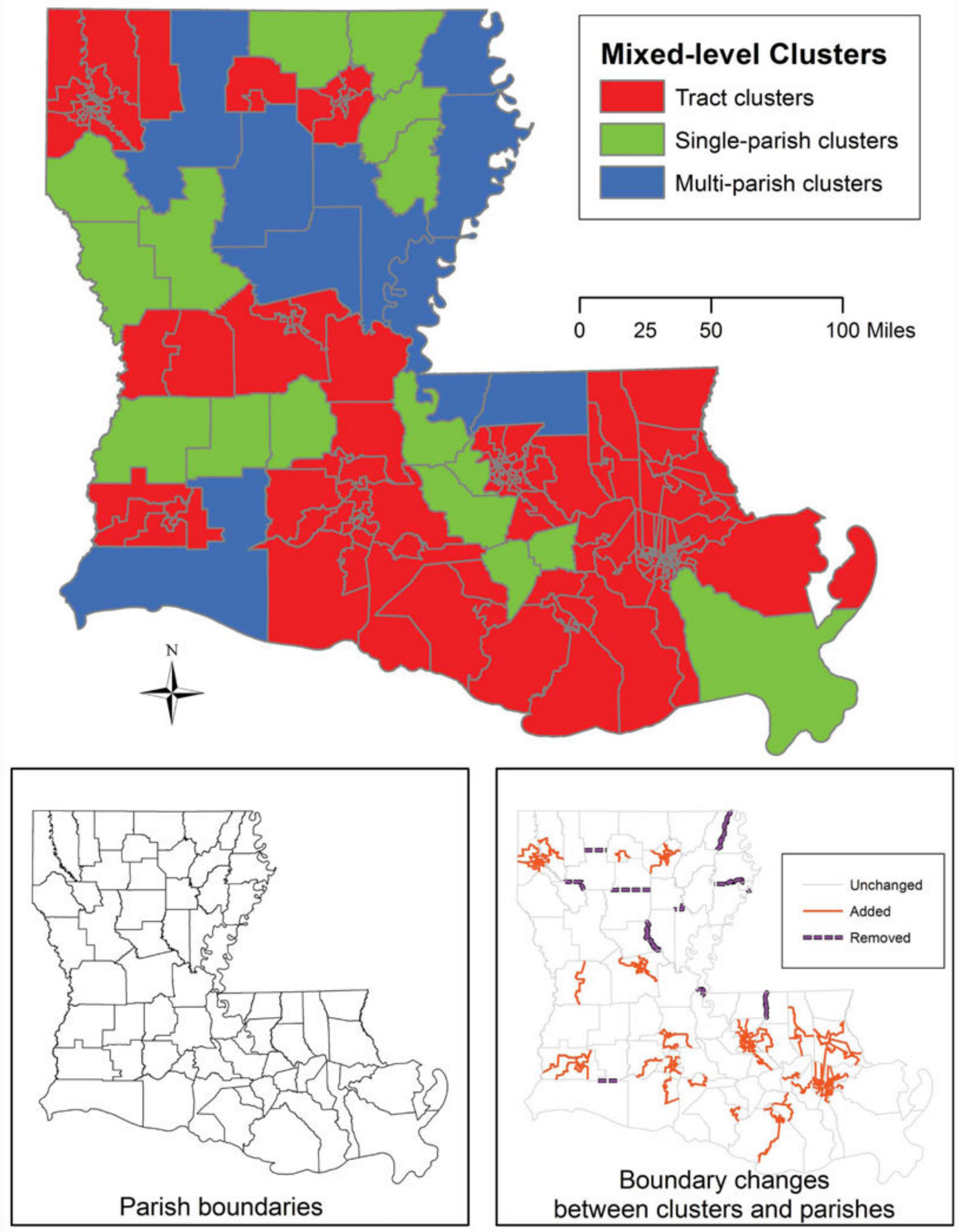


Figure 6.

Constructed regions for cancer data release by mixed-level regionalization (MLR).

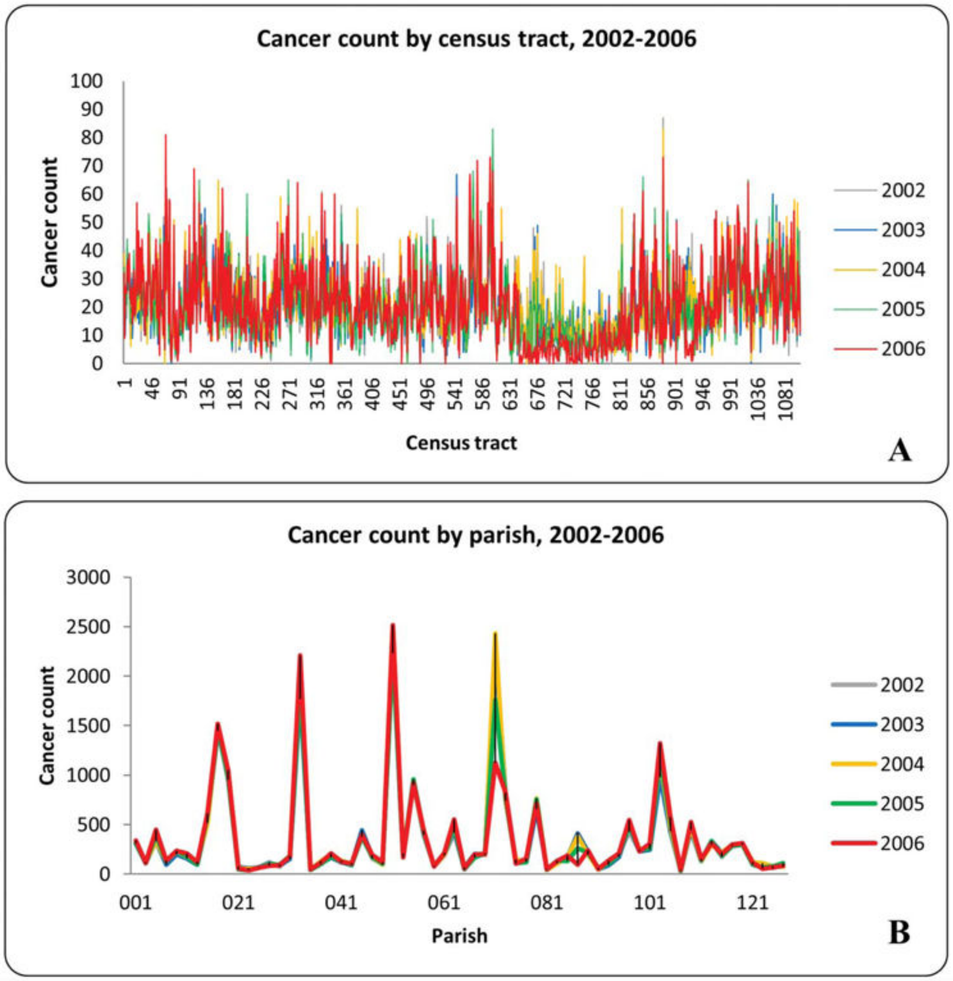


Figure 7.

Yearly variation in cancer count, 2002–2006.

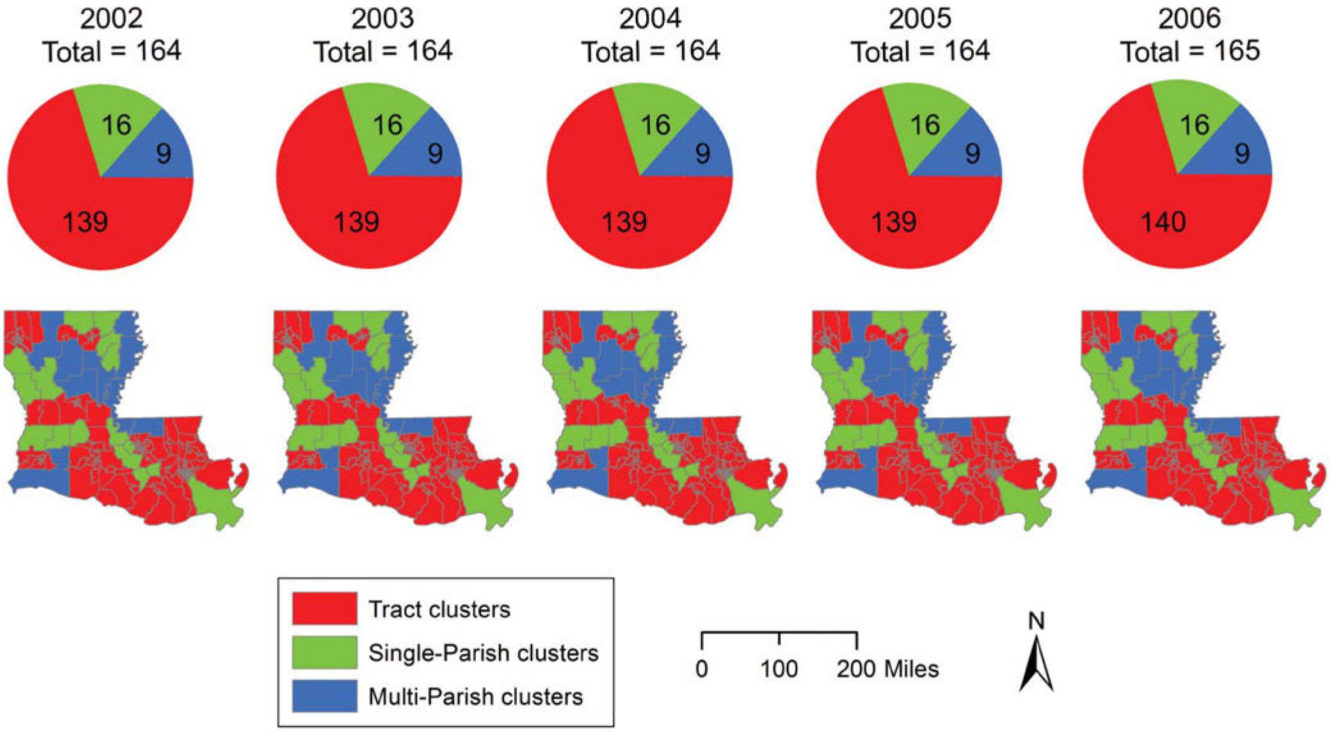


Figure 8.

Mixed-level regionalization results with population ≥ 20,000 and cancer count > 15, 2002-2006.

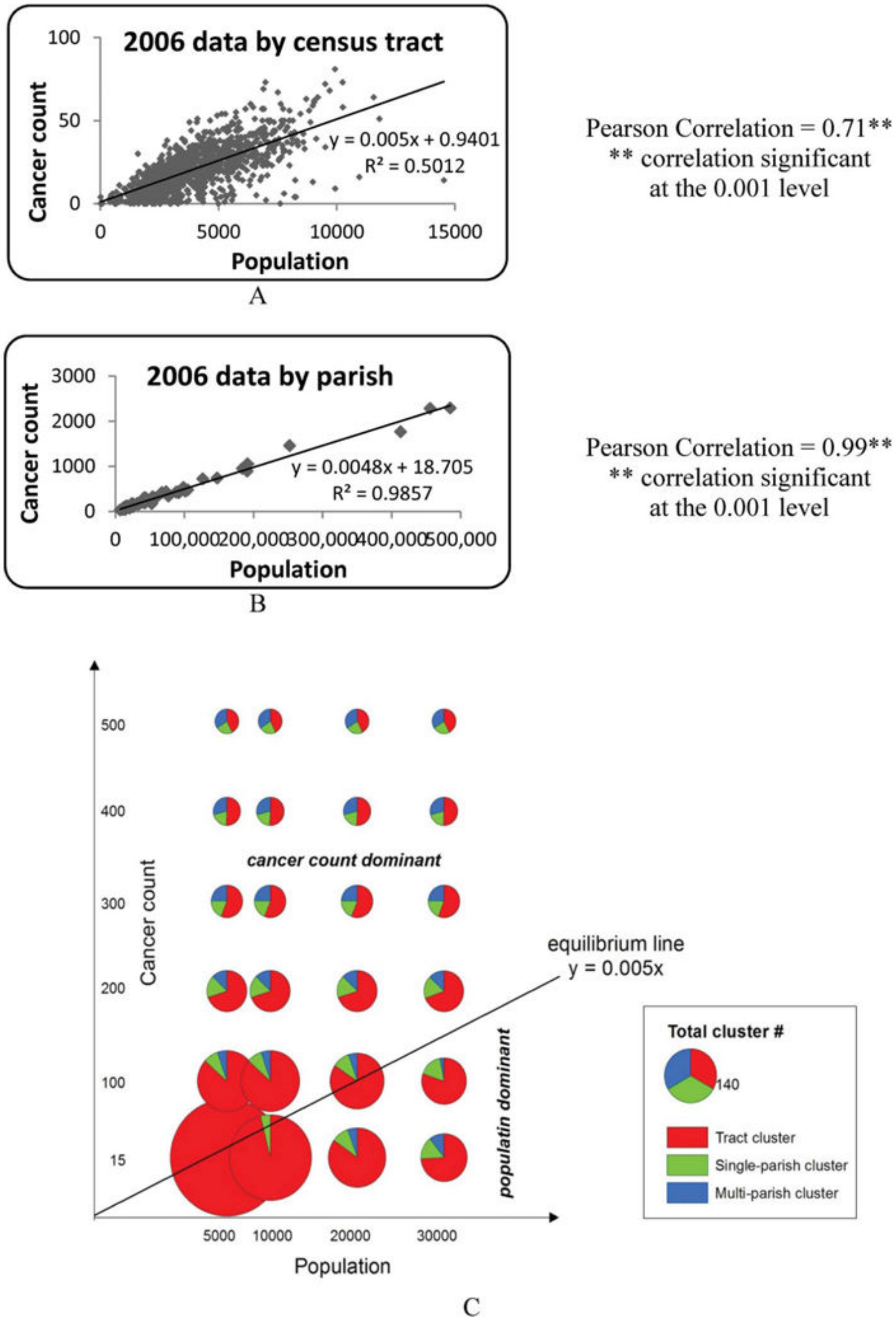


Figure 9.

Cluster types and numbers with various thresholds of population and cancer counts.

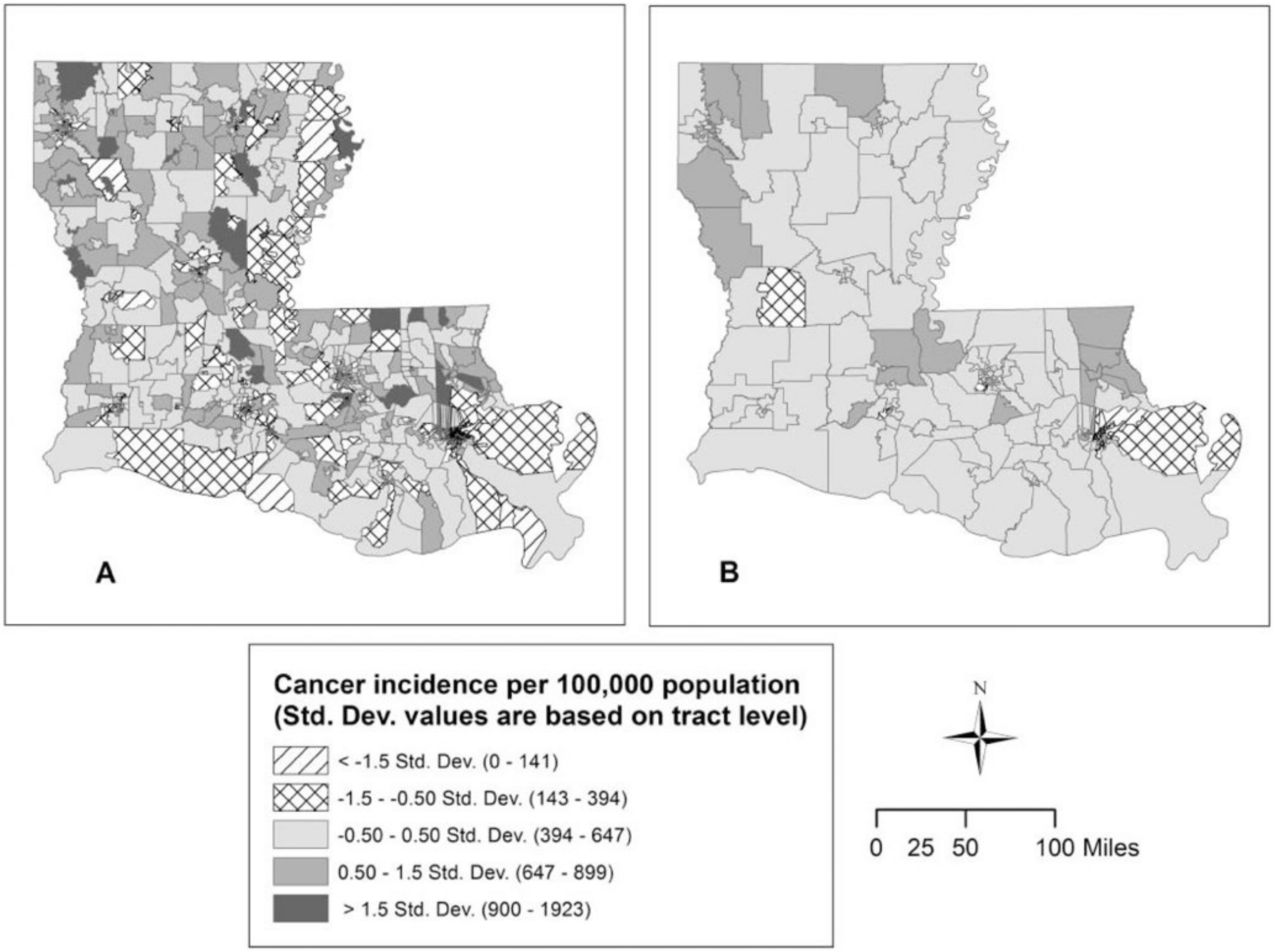

A
B
Cancer incidence per 100,000 population
(Std. Dev. values are based on tract level)
< -1.5 Std. Dev. (0 - 141)
-1.5 - -0.50 Std. Dev. (143 - 394)
-0.50 - 0.50 Std. Dev. (394 - 647)
0.50 - 1.5 Std. Dev. (647 - 899)
> 1.5 Std. Dev. (900 - 1923)
N
0 25 50 100 Miles

Figure 10.

Crude cancer rates: (A) Before the clustering (tract level), and (B) After the clustering (mixed-level).

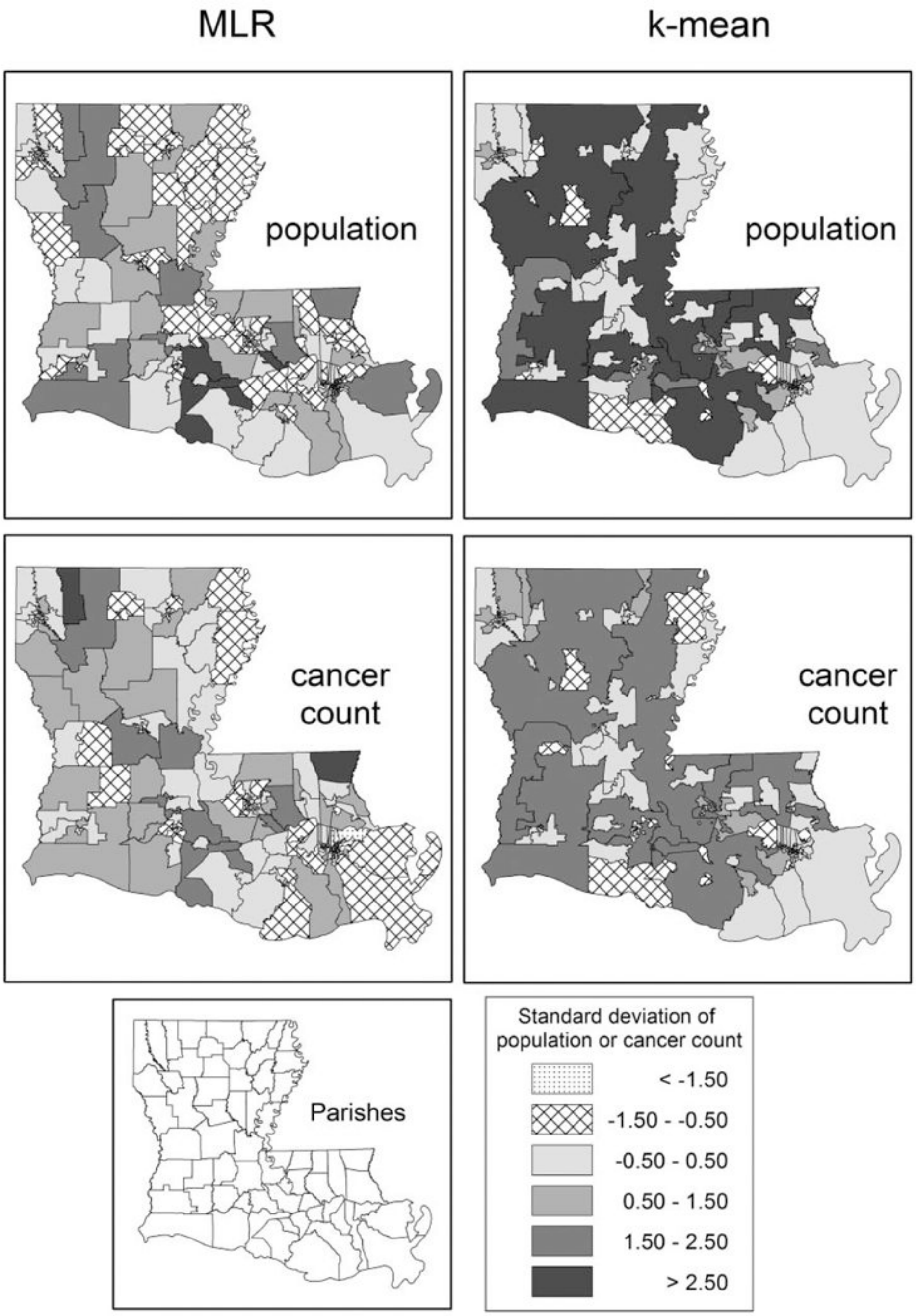


Figure 11.

Comparison between mixed-level regionalization (MLR) and *k*-mean clustering.

Table 1

Factor structure of attributive variables

Variable	Socioeconomic disadvantages factor	High health care needs factor	Language barrier factor
Nonwhite population (%)	**0.8466**	0.0251	–0.0328
Female-headed households (%)	**0.8527**	0.0529	–0.0665
Population without high school diploma (%)	**0.7467**	0.3441	0.2061
Median income ($)	**–0.8079**	0.0055	0.0639
Population in poverty (%)	**0.9339**	0.0074	–0.0091
Homeownership (%)	**–0.6711**	0.5167	0.0174
Households with > 1 person per room (%)	**0.8136**	0.1626	0.1516
Households without vehicles (%)	**0.8776**	–0.1223	–0.0576
Housing units lack of basic amenities (%)	**0.4926**	0.3056	0.0487
Demographic groups with high health care needs (%)	0.2153	**–0.7579**	–0.3578
Households with linguistic isolation (%)	–0.0196	–0.4449	**0.8847**
Eigenvalues	5.6979	1.2961	0.9918
Proportion of variance explained	0.5180	0.1178	0.0902

Note: Values shown in bold indicate the highest loading of a variable on a factor among all factors.

Table 2

Change in numbers of parishes and census tracts before and after the mixed-level regionalization (MLR)

		Original area units		MLR clusters	
		Initial survey	After step 5	Number of clusters	Type of clusters
Scenario 1:	Parish #	29	29	140	Subparish (tract cluster)
Disaggregation	Tract #	926	926		
Scenario 2:	Parish #	19	16	16	Single parish
No action	Tract #	124	108		
Scenario 3:	Parish #	16	19	9	Multiparish
Aggregation	Tract #	56	72		
Total	Parish #		64	165	
	Tract #		1106		

Table 3

Clustering results with different threshold population and cancer counts

Population	Cancer count	Tract cluster	Single-parish cluster	Multiparish cluster	Total cluster
5,000	15	534	4	0	538
5,000	100	148	14	9	171
5,000	200	61	15	11	87
5,000	300	32	11	14	57
5,000	400	23	9	13	45
5,000	500	15	8	12	35
10,000	15	298	12	1	311
10,000	100	148	14	9	171
10,000	200	61	15	11	87
10,000	300	32	11	14	57
10,000	400	23	9	13	45
10,000	500	15	8	12	35
20,000	15	140	16	9	165
20,000	100	125	15	8	148
20,000	200	61	15	11	87
20,000	300	32	11	14	57
20,000	400	23	9	13	45
20,000	500	15	8	12	35
30,000	15	83	17	12	112
30,000	100	77	16	3	106
30,000	200	59	15	11	85
30,000	300	32	11	14	57
30,000	400	23	9	13	45
30,000	500	15	8	12	35

Table 4

Comparison of mixed-level regionalization (MLR), *k*-mean, and ISODATA clustering

Method	Mixed-level results	Attributes' participation	Weights of attributes	Additional numeric constraints (population, cancer, etc.)	Multiple iterations	Predetermined number of groups	Sensitive to initial cluster
MLR	Yes	Yes	Yes	Yes	Yes	No	N/A
k-mean	No	Yes	No	No	Yes	Yes	Yes
ISODATA	No	Yes	No	Yes	Yes	Yes	Yes