Published in final edited form as:
Appl Geogr. 2012 November ; 35(1-2): 1–11. doi:10.1016/j.apgeog.2012.04.005.

Constructing Geographic Areas for Cancer Data Analysis: A Case Study on Late-stage Breast Cancer Risk in Illinois

Fahui Wang[1,*], **Diansheng Guo**[2], and **Sara McLafferty**[3]

[1]Fred B. Kniffen Professor, Department of Geography & Anthropology, Louisiana State University, Baton Rouge, LA 70803

[2]Associate Professor, Department of Geography, University of South Carolina, Columbia, SC 29208

[3]Professor, Department of Geography, University of Illinois at Urbana-Champaign, Urbana, IL 61801-3671

1. Introduction

Cancer is a rare disease. As a result, analysis of cancer data often suffers from the *small population (numbers) problem*, which can lead to unreliable rate estimates, sensitivity to missing data and other data errors, and data suppression in sparsely populated areas. Figure 1, generated from the State Cancer Profiles web site (statecancerprofiles.cancer.gov), shows age-adjusted death rates for female breast cancer in Illinois counties for 2003–2007. Rates for 37 out of 102 counties (i.e., 36.3%, mostly rural counties) are suppressed to "ensure confidentiality and stability of rate estimates" because counts were fewer than 16 cases. Cancer incidence in these counties cannot be analyzed, leaving large gaps in our understanding of geographic variation in cancer and its social and environmental determinants. For rare cancers or cancers of particular population groups, the problems of data suppression and unreliable estimates are conceivably much worse. On the other hand, using data at the county scale obscures rate variations that might exist within large urban counties such as Cook, DuPage and Kane in Illinois. Because of these problems, analyzing data by county has limited value both for the public and for researchers who are interested in cancer patterns at finer geographic scales.

The small population problem is not limited to cancer data analysis, and is present in all studies of rare events. Several strategies are commonly used to address the problem. For example, in the analysis of homicide data, criminologists have used strategies such as suppressing data for small populations, aggregating data over a longer period of time, or aggregating data to larger geographic units (Wang and O'Brien, 2005). Some analytical geographic methods have been developed to mitigate this problem. Conceptually similar to moving averages that smooth observations over a longer time interval, *spatial smoothing* computes the averages using a spatial window (Talbot et al., 2000). Spatial smoothing methods include the floating catchment area method, kernel density estimation (Wang, 2006: 36–38), empirical Bayes estimation (Clayton and Kaldor, 1987), and more recently

fwang@lsu.edu, phone: 225-578-6629, fax: 255-578-4420.

locally-weighted average (Shi, 2007) and adaptive spatial filtering (Tiwari and Rushton, 2004; Beyer and Rushton, 2009), among others. While spatial smoothing assists in revealing the overall trend of spatial patterns, the methods are ad hoc in the sense that the size of the smoothing window does not necessarily reflect knowledge of the disease characteristics or process. Another method, hierarchical Bayesian modeling (HBM), commonly used in spatial epidemiology, uses a nonparametric Bayesian approach to detect clusters of high risk and low risk with the prior model assuming constant risk within a cluster (Knorr-Held, 2000; Knorr-Held and Rasser, 2000). However, a minimum threshold population (or disease incidents count) is not incorporated in the HBM.

Another viable approach is to construct larger areas from small ones so that the base population is sufficiently large and comparable across areas. Geography has a long tradition of building regions for various purposes under the term "*regionalization*" (Cliff et al., 1975). Regionalization is to group a large number of small units into a relatively small number of regions while optimizing a given objective function and satisfying certain constraints. Traditional regionalization methods place the first priority on attribute similarity within areas, and most are implemented manually or semi-automatically. An example is the work by Haining et al. (1994) that used attribute information to first form initial regions and then applied several subjective rules and local knowledge to further adjust the regions. Advancements in Geographic Information Systems (GIS) technology have enabled researchers to develop innovative, computing-intensive methods. Among others, two earlier methods emphasize spatial proximity: Lam and Liu (1996) used space-filling curves to measure the nearness or spatial order of areal units, and grouped areas consecutively to reach a capacity constraint; and Black et al. (1996) constructed regions of approximately equal population size by beginning with an area and adding the nearest areas to form each region with the desired threshold population. Neither of these methods, however, considers within-area homogeneity of attribute. The requirement for merging only regions of similar attributes is to minimize the loss of information from aggregation. In other words, if very different regions were grouped together, much of the geographic variation, which is of primary interest to spatial analysis, would be smoothed out in the regionalization process.

There are a number of automated regionalization methods reported in the literature that account for spatial contiguity and attribute homogeneity within the derived areas: the AZP (Openshaw 1977; Openshaw and Rao 1995; Cockings and Martin 2005; Grady and Enander 2009), MaxP (Duque et al., 2007), MSSC (Mu and Wang, 2008) and REDCAP (Guo 2008; Guo and Wang 2011). For example, the AZP method starts with an initial random regionalization and then iteratively refines the solution by reassigning objects to neighboring regions to improve the objective function value, and therefore the regionalization result varies dependent upon the initial randomization state. The MSSC merges or melts adjacent and similar areas to form larger areas by following a process guided by an objective of minimizing loss of information in aggregation, but it does not guarantee that newly formed areas have population above a threshold. See Guo and Wang (2011) for more detailed reviews and comparison of existing regionalization methods.

Despite these recent advancements, our literature review indicates that no automated regionalization methods have been adapted or applied to cancer studies as a way to address the small population problem. Analysis of cancer data requires special attention to issues such as data confidentiality and privacy concerns, which require additional constraints such as a minimum threshold for region population or case numbers. This research adopted the *REDCAP model* and modified by incorporating a minimum base population (e.g., 20,000) and/or a threshold for cancer cases (e.g., 15), especially *adaptable to cancer data analysis*. Therefore, the model is termed "*REDCAPc*". The method enhances the presentation and visualization of cancer surveillance data by producing geographic areas of comparable size,

and generates reliable cancer rates in those areas. As similar areas are merged, it mitigates the spatial autocorrelation problem commonly observed in data of geographic areas and simplifies subsequent regression analysis (Mu and Wang, 2008). Spatial autocorrelation occurs when attributes of nearby areas tend to be similar (positively autocorrelated) or dissimilar (negatively autocorrelated), and thus violates the assumption of independent samples in regular regression. It is an important step towards developing frame-independent and scale-invariant methods (Kwan and Weber, 2003).

REDCAPc was developed under the support of National Cancer Institute (NCI) SEER-RRSS program, and successfully implemented in the Louisiana Tumor Registry and the Cancer Prevention Institute of California. This paper reports a case study on analyzing the late-stage breast cancer risks in Illinois to illustrate its usage and potential for cancer studies.

2. REDCAP and its modification for cancer data analysis (REDCAPc)

REDCAP refers to a family of methods, termed “regionalization with dynamically constrained agglomerative clustering and partitioning”. REDCAP extends the single-linkage (SLK), average-linkage (ALK), complete-linkage (CLK), and the Ward hierarchical clustering methods to enforce the spatial contiguity of clusters and obtain a set of regions while explicitly optimizing an overall homogeneity measure (Guo 2008). In this paper, we use the contiguity-constrained complete-linkage clustering (CLK), where the distance between two clusters is defined as the furthest pair (most dissimilar) of data points.

In essence, the goal of REDCAP is to construct a set of homogeneous regions by aggregating contiguous small areas of similar attribute values (e.g., socioeconomic structure). To achieve this goal, REDCAP constructs a cluster hierarchy based on attribute similarities among small areas and then partitions the spatially contiguous cluster tree to explicitly optimize a homogeneity measure. The homogeneity measure is the total sum of squared deviations (SSD) (Everitt 2002), as defined in Equation (1), where k is the number of regions, n_r is the number of small areas in region r, d is the number of variables considered, x_{ij} is a variable value and $\bar{x}_j$ is the regional mean for variable j. Each input data variable should be normalized and a weight can be assigned for each variable.

$$SSD = \sum_{r=1}^{k} \sum_{i=1}^{n_r} \sum_{j=1}^{d} (x_{ij} - \bar{x}_j)^2 \quad (1)$$

As shown in Figure 2, REDCAP is composed of two steps: (1) contiguity-constrained hierarchical clustering and (2) top-down tree partitioning. The color shade of each polygon represents its attribute value and similar colors represent similar values. Two polygons are considered contiguous in space if they share a segment of boundary. In the first step, as shown in Figure 2(A), REDCAP constructs a hierarchy of spatially contiguous clusters based on the attribute similarity under contiguity constraint. Two adjacent and most similar areas are grouped to form the first cluster; two adjacent and most similar clusters (based on the CLK in this study) are grouped together to form a higher-level cluster; and so on until the whole study area is one cluster. A spatially contiguous tree is generated to fully represent the cluster hierarchy (i.e., each cluster at any level is a sub-tree in the map). In the second step, as shown in Figure 2(B), REDCAP partitions the tree to generate two regions by removing the best edge (i.e., 11–15 in Figure 2(B)) that optimizes the homogeneity measure (i.e., SSD) as defined in Equation 1. In other words, the two regions are created in a way that the total within-region homogeneity is maximized. The partitioning continues until the desired number of regions is reached.

We want to emphasize that the first step (i.e., contiguity-constrained clustering) is a bottom-up process, which builds a hierarchy of spatially contiguous clusters but does not directly optimize the objective function. The second step (i.e., tree partitioning) is a top-down approach that directly optimizes the objective function. The final regions mostly likely are not the same as the top clusters suggested in the cluster hierarchy. This is why the second step is necessary, which makes the REDCAP methods different from traditional contiguity constrained hierarchical clustering. REDCAP is similar to the SKATER method (Assunç ă o et al. 2006) in terms of the two-step framework but significantly outperforms the latter in our evaluations according to criteria such as total heterogeneity, region size balance, internal variation and preservation of data distribution (Guo 2008).

For cancer data analysis, in particular for the purpose of mitigating the small population problem, REDCAP is modified, termed REDCAPc to accommodate additional constraints to be enforced such as a minimum size threshold in terms of region population and/or the number of cancer cases. Such constraints are enforced in the second step, i.e., tree partitioning. For each potential cut, if it cannot produce two regions that both satisfy the constraints, the cut will not be considered as a candidate cut. Then the best of all candidate cuts is chosen to partition a tree into two regions. If there is no candidate cut (i.e., no cut can produce regions that satisfy the constraints), then the region will not be partitioned further. If none of the current regions can be cut, the regionalization process stops. The method is deterministic. In other words, given the same criteria (definitions of attribute similarity and spatial contiguity, minimum region population and/or number of cancer cases), the method yields the same regions. The resulting regions are all large enough and have the highest homogeneity within each region.

3. Case Study on Late-Stage Breast Cancer Risks in Illinois

The variations of breast cancer mortality rates from place to place reflect both underlying differences in breast cancer prevalence and differences in diagnosis and treatment that affect the risk of death. Patients whose cancer is diagnosed early have fewer complications and substantially higher rates of survival than those whose cancer is diagnosed late. For breast cancer, access to primary care and mammography screening is critically important for early detection (Wang et al., 2008). Access is strongly influenced by financial, socio-cultural and geographic barriers (or risk factors) (Wang, 2012). In the analysis of late-stage breast cancer risks, at least three areas can benefit from the aforementioned REDCAPc of constructing suitable geographic areas:

1. reliable mapping of late-stage rates in newly-constructed areas and related exploratory spatial analysis to identify high-risk areas;
2. regression analysis of late-stage risk factors across newly-defined areas; and
3. multi-level modeling of risk factors in various configurations of “neighborhoods”.

3.1 Data preparation and variable definitions

This case study uses cancer incidence data in Illinois from the Illinois State Cancer Registry (ISCR), Illinois Department of Public Health (IDPH) in 2000. Each cancer case is geocoded to the county and zip code of residence, and includes variables such as cancer type, age group, sex, race, diagnosis stage and year. In the study period, there are 31,914 incidences of breast cancer in Illinois. The ISCR uses a classification scheme consistent with SEER summary stage to measure stage at diagnosis (Young et al., 2001). Consistent with other studies, we defined late-stage as diagnosis in stages 2 through 7 (Bradley et al., 2002). The late-stage group consists of cancers that have spread beyond the site of origin to nearby or distant tissues, organs or lymph nodes. Excluding cases with no stage information, there are

10,206 female breast cancer cases, among which 2,906 (28.5 percent) are classified as late-stage.

Given the focus of this paper on constructing geographic areas, the selection and definitions of risk factors of late-stage breast cancer is not discussed in depth. Only the most relevant literature is cited in the following discussion. This research considers the following four types of factors:

1. demographic and other attributes of *individual* cancer patients,
2. non-spatial factors (socio-demographic variables) of neighborhood level,
3. urban-rural classification assigned to each zip code area, and
4. spatial access measures to primary care physicians and to cancer screening (i.e., mammography) facilities.

Attributes of individual cancer cases from the ISCR are limited, and only age and race were available and used for this study (e.g., McLafferty and Wang, 2009). Three age groups (<40, 40–69 and ≥70 years) (Elkin et al., 2010) are coded by two dummy variables, and race (black, non-black) by one dummy variable. This set of variables is at the individual level, and the following three sets are at the level of zip code area.

Area-based nonspatial factors such as demographic and socioeconomic characteristics were extracted at the census tract level and then interpolated to the zip code level by spatial interpolation (Wang et al., 2008). Among a wide range of socio-demographic variables available from the census, 10 were selected: socioeconomic status (e.g., population in poverty, female-headed households, home ownership, and median income), environment (e.g., households with an average of more than 1 person per room, and housing units lack of basic amenities), linguistic barriers and education (e.g., non-white population, population without a high-school diploma, and households linguistically isolated), and transportation mobility (e.g., households without vehicles). Due to concerns of multicollinearity among these variables, factor analysis was used to consolidate the variables into two factors that accounted for over 70% total variance. Table 1 shows the factor loadings of the 10 variables on the two factors. The factors are labeled “socioeconomic disadvantages” and “sociocultural barriers” respectively.

A rural-urban classification code provided in the ISCR (1–9) was used to examine possible discrepancies between rural and urban areas (though not a focus of this study). Prior studies (Wang et al., 2008; McLafferty and Wang, 2009) used more categories for rural-urban continuum and highlighted the uniqueness of Chicago region. Here we adopted a binary division: (1) Chicago metro area, i.e., zip code areas coded “1” in the ISCR (in metro area with ≥ 1 million population) but excluding areas around East St. Louis, and (2) others. This simple strategy was adopted since a more detailed rural-urban breakdown would lead to many fragmented sub-areas and create a challenge to preserve these sub-areas in the process of regionalization. By doing so, the study area is basically composed of two sub-areas: “Chicago metro area” and “non-Chicago area”. A dummy variable is used to code the division. We also experimented with a 3-category scenario (areas in City of Chicago, suburban Chicago, and the rest), and the results remained largely the same and thus not reported.

Spatial access to primary care was estimated using the two-step floating catchment area method (2SFCA) (Wang, 2006: 80–82). In essence, the 2SFCA computes a numerical value that represents the ratio of the local supply of primary care physicians to the local demand (population) for primary care. Supply and demand interact within a fixed range (i.e., 30 minutes) of travel time. A high value for this spatial access measure represents better access.

Spatial access to cancer screening facility was measured as the travel time from a cancer patient (approximately by the zip code area population-weighted centroid) to the nearest mammography facility based on real-world road networks accounting for lower speeds in high-density urban areas (Wang et al., 2008).

3.2 Constructing geographic areas by REDCAPc

As discussed earlier, a major challenge for regionalization is to account for both spatial contiguity (only merging adjacent areas) and attribute homogeneity (only grouping similar areas). For this study, spatial contiguity is defined as rook contiguity. In other words, only zip code areas that share boundary line(s) (not just points) are considered contiguous. The spatial contiguity matrix is saved as a text file for subsequent clustering. The two factors, socioeconomic disadvantages and socio-cultural barriers, defined earlier by the factor analysis were used as attributes for the regionalization process. Thus, the regions are defined on the basis of both spatial contiguity and socioeconomic and sociocultural characteristics. Zip codes that have socially similar and spatially contiguous are grouped together to form regions.

A threshold number of cancer cases for the newly-defined regions is another input parameter that needs to be defined. Similar to the criterion adopted by the State Cancer Profiles, this study uses a minimum number of 15 breast cancer incidences. In other words, zip code areas with fewer than 15 cases are grouped to form a larger area that has a sufficient number of cases. In order to preserve the distinction between Chicago metro vs. non-Chicago areas in the spatial clustering process, the study area was first split to two sub-areas, and each was processed separately to construct new areas in REDCAPc. Finally the results from the two were merged together to cover the study area.

Among the 1,364 zip code areas in Illinois, 1,122 zip code areas had fewer than 15 breast cancer cases in 2000. That is to say, breast cancer rates in 82.3% zip code areas would need to be suppressed if the threshold of 15 cases is used as the criterion to ensure confidentiality and reliable rate estimates. The percentage is higher outside Chicago (984 out of 1047 or 94.0% zip code areas) than in the Chicago metro area (138 out of 317 or 43.5% zip code areas) because zip codes in the Chicago metro area tend to have larger populations. After the regionalization, a total of 341 new areas were generated with 198 new areas in the Chicago metro region and 143 outside Chicago. So there is more grouping or aggregation outside of the Chicago metro area.

Table 2 outlines the statistical distributions of total cases and late-stage cases of breast cancer, and the late-stage rates in zip code areas and newly-defined areas. Here, *late-stage cancer rate* is the ratio of number of late-stage cancer cases to the total cancer cases. Note that late-stage rates cannot be computed for the 421 zip code areas with zero cancer cases. Even among the remaining 943 zip code areas, the late-stage rates are clearly less stable (standard deviation = 0.2755) than in the areas generated by REDCAPc (standard deviation = 0.0951). Figures 3(a)–(b) show the strong contrasts in the frequency distributions of rates between the two types of areas. The distribution for zip code areas is heavily skewed to the left (with a rate of 0 for 285 out of 943 zip code areas), whereas the distribution for the new areas tends to be normal and peaks around the mean. This is an important property as many commonly used statistical test assume that variables are normally distributed.

3.3 Mapping and exploratory spatial data analysis in newly-defined areas

For the reasons discussed above, direct mapping of late-stage breast cancer rates in zip code areas displays a highly-fragmented geographic pattern with many 0 values including areas with either 0 cancer case (missing late-stage rates) or 0 late-stage cancer case (true 0 late-

stage rates), as shown in Figure 4. Figure 5 shows the variation of late-stage breast cancer rates across newly-defined areas. The elevated late-stage rates are scattered across the state with no apparent geographic patterns.

Some exploratory spatial data analysis is infeasible for zip code area data due to its fragmented pattern of late-stage breast cancer rates (zip code areas with valid rates are isolated/separated by many with missing values), but possible for the new areas. Here we use spatial autocorrelation or hot spot analysis, commonly available in commercial GIS software such as ArcGIS (http://www.esri.com/software/arcgis/index.html) or free spatial analysis packages such as GeoDa (http://geodacenter.asu.edu/software/downloads) and CrimeStat (http://www.nedlevine.com/nedlevine17.htm), for illustration. With the spatial weights defined by the polygon rook contiguity, the global Moran I for late-stage breast cancer rates in the new areas is calibrated as 0.0924, which is statistically significant at 0.01. In other words, high late-stage rates tend to cluster together; and so do low late-stage rates. In order to reveal localized cluster patterns, hot-spot analysis is conducted to obtain local Gi* indices (Getis and Ord, 1992) in the new areas. The result is shown in Figure 6. Local pockets of high late-stage rate concentrations are observed in central city of Chicago and its western and southern suburbs, as well as in several rural areas in the northern part of the state. Additional spatial exploratory analysis such as cluster analysis can also be conducted by SaTScan (http://www.satscan.org/) and other programs (Wang, 2006). The next section examines the association with various risk factors.

3.4. Regression models on risks of late-stage breast cancer

Section 3.1 discussed four types of risk factors commonly considered in analysis of late-stage breast cancer diagnosis. Various regression models can be used to examine the association of late-stage cancer with these risk factors. As explained previously, OLS regression is only applicable to the analysis of new areas where cancer rates are fairly stable and reliable. The OLS model is suitable when data of individual cancer cases are not available, and the analysis is limited to the area (neighborhood) level. In an OLS model, the dependent variable is late-stage cancer rate and independent variables are the aforementioned risk factors. Poisson regression is often used to partially account for the skewed distribution of late-stage cancer rates (Wang et al., 2008), caused by the small population problems discussed previously. In a Poisson regression model, the dependent variable is the number of late-stage cancer cases (the total number of cancer cases serves as an offset variable), and the independent variables are also limited to the area level. A multilevel logistic model examines the risk of individual cancer cases being late-stage, where the dependent variable is binary (0, 1), and independent variables include both individual- and neighborhood-level risk factors (e.g., McLafferty and Wang, 2009). Table 3 outlines three models, and the dependent and independent variables used in each. Note that all independent variables (two factor scores and two spatial accessibility measures) at the zip code level are aggregated to the new areas by the population weighted average method.

Table 4 presents the regression results: the OLS on the new areas, and the Poisson and multilevel models on both the zip code area and the new areas. The results are summarized below.

1. The three individual-level variables are all significant in the multilevel models regardless whether zip code areas or new areas are used as the neighborhood (area) level. Consistent with findings from many studies, the risk of late-stage breast cancer is higher among younger patients and lower among older patients, likely resulting from differences in frequency of primary care visits and age-related cancer screening protocols (McLafferty and Wang, 2009). The risk is higher among black cancer patients, controlling for age and area-level socioeconomic

characteristics, is consistent with finding reported in Martin and Newman (2007) among others. Some reported inconsistencies across geographic scales in racial disparities in breast cancer survival (Meliker et al., 2009).

2. The two area-level socioeconomic factors are significant with expected signs in the OLS and Poisson models. In the multilevel models, the socioeconomic disadvantages factor is no longer significant, but the sociocultural barriers factor remains significant (and the results are consistent in two neighborhood definitions). The disappearance of the socioeconomic disadvantages factor can be explained by its high correlation with the individual-level variable “black” (correlation coefficient = 0.59). In other words, the disproportionally higher presence of black patients in neighborhoods with concentrated socioeconomic disadvantages dominates the contextual effect. In contrast, sociocultural barriers remain statistically significant in the multilevel models suggesting that they may influence use of screening services and the quality and effectiveness of those services (Chu et al., 2003).

3. The urban-rural disparities do not appear to be very significant in this study (the statistical significance is 0.10 in the OLS and the two Poisson models, but not at all in both multilevel models).

4. In all models, the coefficient for travel time to the nearest mammography facility is not statistically significant, but that for spatial access to primary care is very significant. Insignificance of proximity to mammography facilities is also reported in other studies (e.g., Henry et al. 2011), but the finding here should be taken with caution since zip code centroids instead of street addresses (not available to this study) were used to approximate cancer patient locations. Most prior studies in examining the role of primary care access in cancer diagnosis stage simply used distance or travel time to physicians (e.g., Parsons and Askland, 2007; Jones et al., 2008) to measure accessibility, and did not capture the complex patients-doctors interactions as we did (also in Wang et al., 2008; McLafferty and Wang, 2009). This study indicates that living in areas with poor spatial access to primary care increases the risk of late-stage breast cancer.

4. Concluding comments

Analysis of cancer data often suffers from the small population problem, which leads to less reliable rate estimates and data suppression in sparsely populated areas. This research develops a GIS-based automated regionalization method, namely REDCAPc, that constructs larger areas that are more coherent than geopolitical areas or spatial smoothing windows in terms of socioeconomic characteristics and spatial proximity. By doing so, the study demonstrates that the cancer rates become more stable and reliable and conform to a normal distribution. This permits direct mapping, exploratory spatial data analysis, and even simple OLS regression.

Acknowledgments

The financial supports from the National Cancer Institute (NCI) under the grant 1-R21-CA114501-01 and two NCI SEER-RRSS grants (one through the Louisiana Tumor Registry and another through the Cancer Prevention Institute of California) are gratefully acknowledged. 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. We are grateful for two anonymous reviewers, whose constructive comments helped us prepare an improved and final version of the paper.

References

Assunção RM, Neves MC, Cămara G, Freitas CDC. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. International Journal of Geographical Information Science. 2006; 20:797–811.

Beyer KM, Rushton G. Mapping cancer for community engagement. Preventing Chronic Disease. 2009; 6:A03. [PubMed: 19080009]

Black, RJ.; Sharp, L.; Urquhart, JD. Analysing the spatial distribution of disease using a method of constructing geographical areas of approximately equal population size. In: Alexander, PE.; Boyle, P., editors. Methods for Investigating Localized Clustering of Disease: IARC Scientific Publications No. 135. Lyon, France: International Agency for Research on Cancer; 1996. p. 28-39.

Bradley C, Given C, Roberts C. Race, socioeconomic status and breast cancer treatment and survival. Journal of the National Cancer Institute. 2002; 94:490–496. [PubMed: 11929949]

Clayton D, Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics. 1987; 43:671–681. [PubMed: 3663823]

Cliff, A.; Haggett, P.; Ord, J.; Bassett, K.; Davis, R. Elements of Spatial Structure. Cambridge: Cambridge University Press; 1975.

Chu KC, Lamar CA, Freeman HP. Racial disparities in breast carcinoma survival rates: separating factors that affect diagnosis from factors that affect treatment. Cancer. 2003; 97(11):2853–2860. [PubMed: 12767100]

Cockings S, Martin D. Zone design for environment and health studies using pre-aggregated data. Social Science & Medicine. 2005; 60:2729–2742. [PubMed: 15820583]

Duque JC, Anselin L, Rey SJ. The max-p region problem: Regional Analysis Laboratory. 2007 (http://regionalanalysislab.org/) Working Paper 20070301 (software available at http://www.pysal.org/users/tutorials/region.html).

Elkin EB, Ishill NM, Snow JG, Panageas KS, Bach PB, Liberman L, Wang F, Schrag D. Geographic access and the use of screening mammography. Medical Care. 2010; 48:349–356. [PubMed: 20195174]

Everitt, BS. The Cambridge Dictionary of Statistics. Cambridge: University Press; 2002.

Getis A, Ord JK. The analysis of spatial association by use of distance statistics. Geographical Analysis. 1992; 24:189–206.

Grady SC, Enander H. Geographic analysis of low birthweight and infant mortality in Michigan using automated zoning methodology. International Journal of Health Geographics. 2009; 8(10)

Guo D. Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). International Journal of Geographical Information Science. 2008; 22:801–823.

Guo, D. Greedy optimization for contiguity constrained hierarchical clustering; The 4th International Workshop on Spatial and Spatiotemporal Data Mining, IEEE International Conference on Data Mining; 2009. p. 591-596.

Guo D, Wang H. Automatic Region Building for Spatial Analysis. Transactions in GIS. 2011; 15(s1): 29–45.

Haining R, Wises S, Blake M. Constructing regions for small area analysis: material deprivation and colorectal cancer. Journal of Public Health. 1994; 16:429–438.

Henry KA, Boscoe FP, Johnson CJ, Goldberg DW, Sherman R, Cockburn M. Breast cancer stage at diagnosis: is travel time important? Journal of Community Health. 2011; 36:933–942. [PubMed: 21461957]

Jones AP, Haynes R, Sauerzapf V, Crawford SM, Zhao H, Forman D. Travel times to health care and survival from cancers in Northern England. European Journal Of Cancer. 2007; 44:269–274. [PubMed: 17888651]

Knorr-Held L. Bayesian modelling of inseparable space-time variation in disease risk. Statistics in Medicine. 2000; 19:2555–2567. [PubMed: 10960871]

Knorr-Held L, Rasser G. Bayesian detection of clusters and discontinuities in disease maps. Biometrics. 56:13–21. [PubMed: 10783772]

Kwan MP, Weber J. Individual accessibility revisited: Implications for geographical analysis in the twenty-first century. Geographical Analysis. 2003; 35:341–353.

Lam NS-N, Liu K. Use of space-filling curves in generating a national rural sampling frame for HIV-AIDS research. Professional Geographer. 1996; 48:321–332.

Martin IK, Newman LA. Disparities in breast cancer. Current Problems in Cancer. 2007; 31:134–156. [PubMed: 17543945]

McLafferty S, Wang F. Rural reversal? Risk of late-stage cancer across the rural-urban continuum in Illinois. Cancer. 2009; 115:2755–2764. [PubMed: 19434667]

Meliker JR, Goovaerts P, Jacquez GM, AvRuskin GA, Copeland G. Breast and prostate cancer survival in Michigan. Cancer. 2009; 115:2212–2221. [PubMed: 19365825]

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.

Openshaw S. A geographical solution to scale and aggregation problems in regionbuilding, partitioning, and spatial modelling. Transactions of the Institute of British Geographers NS. 1977; 2:459–472.

Openshaw S, Rao L. Algorithms for reengineering 1991 census geography. Environment & Planning A. 1995; 27(3):425–446. [PubMed: 12346252]

Parsons MA, Askland KD. Cancer of the colorectum in Maine 1995–1998: Determinants of stage at diagnosis in a rural state. Journal of Rural Health. 2007; 23:25–32. [PubMed: 17300475]

Shi X, Duell E, Demidenko E, Onega T, Wilson B, Hoftiezer D. A polygon-based locally-weighted-average method for smoothing disease rates of small units. Epidemiology. 2007; 18:523–528. [PubMed: 17700240]

Talbot TO, Kulldorff M, Forand SP, Haley VB. Evaluation of spatial filters to create smoothed maps of health data. Statistics in Medicine. 2000; 19:2399–2408. [PubMed: 10960861]

Tiwari, C.; Rushton, G. Using spatially adaptive filters to map late stage colorectal cancer incidence in Iowa. In: Fisher, P., editor. Developments in Spatial Data Handling. New York: Springer-Verlag US; 2004. p. 665-676.

Wang, F. Quantitative Methods and Applications in GIS. Boca Raton, FL: CRC Press; 2006.

Wang F. Measurement, optimization and impact of healthcare accessibility: a methodological review. Annals of the Association of American Geographers. 2012

Wang F, McLafferty S, Escamilla V, Luo L. Late-stage breast cancer diagnosis and healthcare access in Illinois. Professional Geographer. 2008; 60:54–69. [PubMed: 18458760]

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 Publishing; 2005. p. 84-100.

Young, JL., Jr; Roffers, SD.; Ries, LAG.; Fritz, AG.; Hurlbut, AA. SEER Summary Staging Manual-2000: Codes and Coding Instructions. Bethesda, MD: National Cancer Institute; 2001. NIH Pub. No. 01 4969.

Highlights

- The small numbers (population) problem occurs in analysis of rare disease (including cancer) data with unstable rate estimates and data suppression in sparsely populated areas.
- This research adopts a GIS-based automated method, termed “regionalization with dynamically constrained agglomerative clustering and partitioning” for cancer analysis (REDCAPc), to construct larger areas with population or case numbers above a threshold.
- Cancer rates in these newly constructed areas have sufficiently large base population, and are thus more reliable and also conform to a normal distribution.
- This permits direct mapping, exploratory spatial data analysis, and even simple OLS regression.
- The method can be used to effectively mitigate the small numbers problem commonly encountered in analysis of public health data.

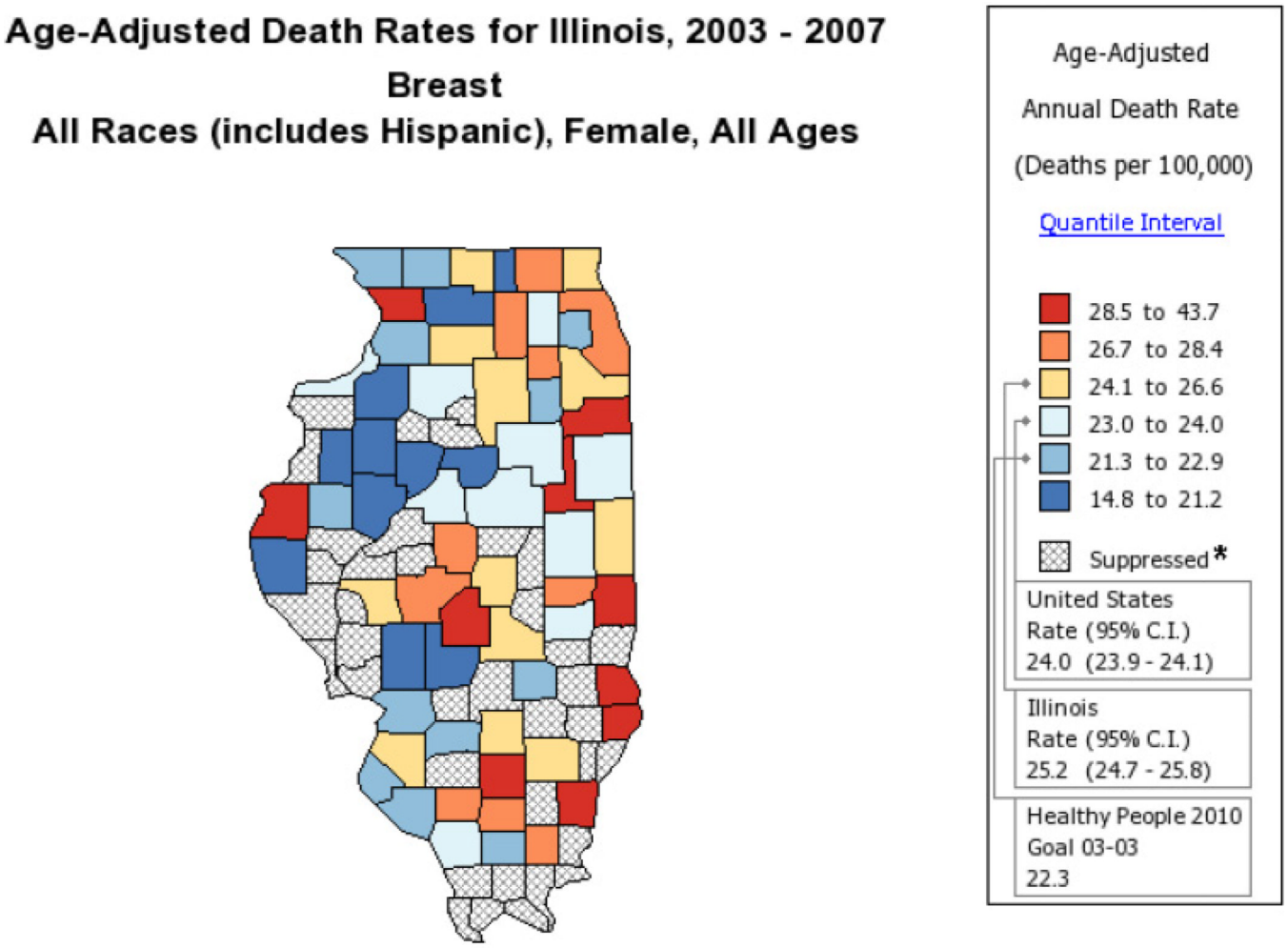


Figure 1.
Female breast cancer death rates in Illinois 2003–2007

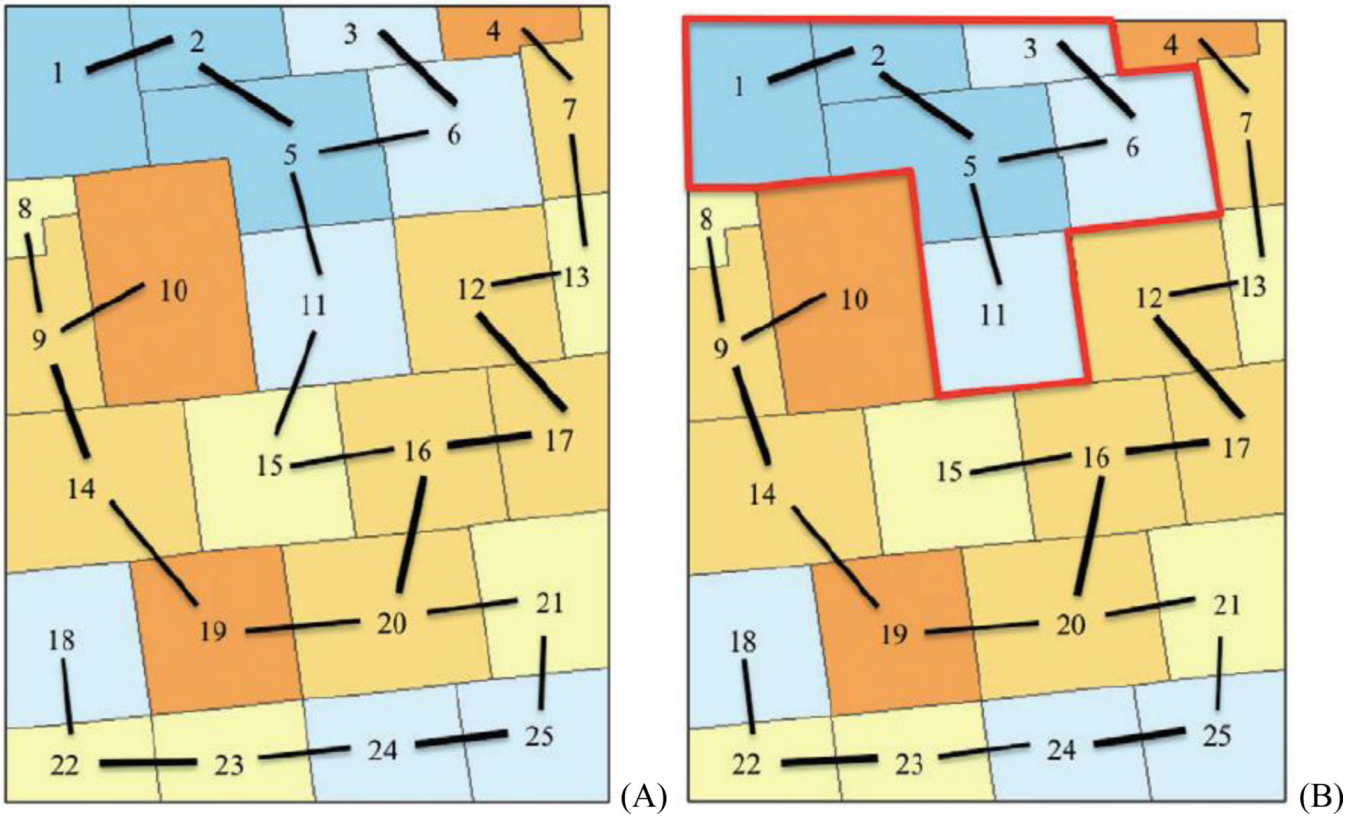
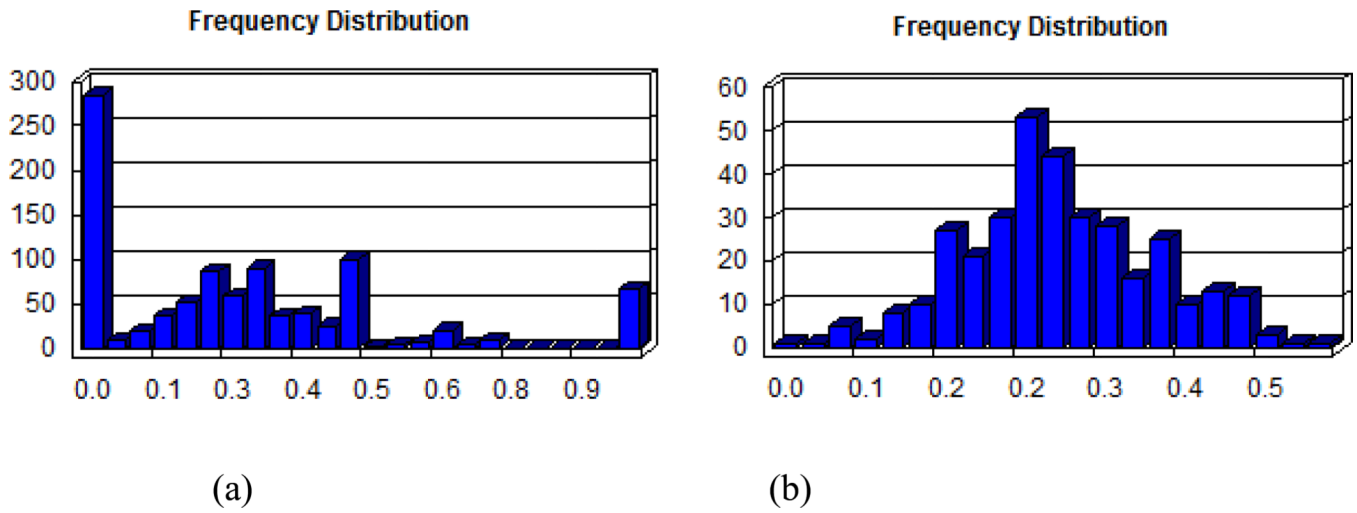


Figure 2.
An example data set illustrating REDCAP: (A) a spatially-contiguous tree is built with a contiguity constrained hierarchal clustering method; (B) partitioning the tree by removing the edge that optimizes the SSD measure.



(a) (b)

Figure 3.
Distribution of late-stage breast cancer rates in Illinois 2000: (a) 943 zip code areas and (b) 341 new areas

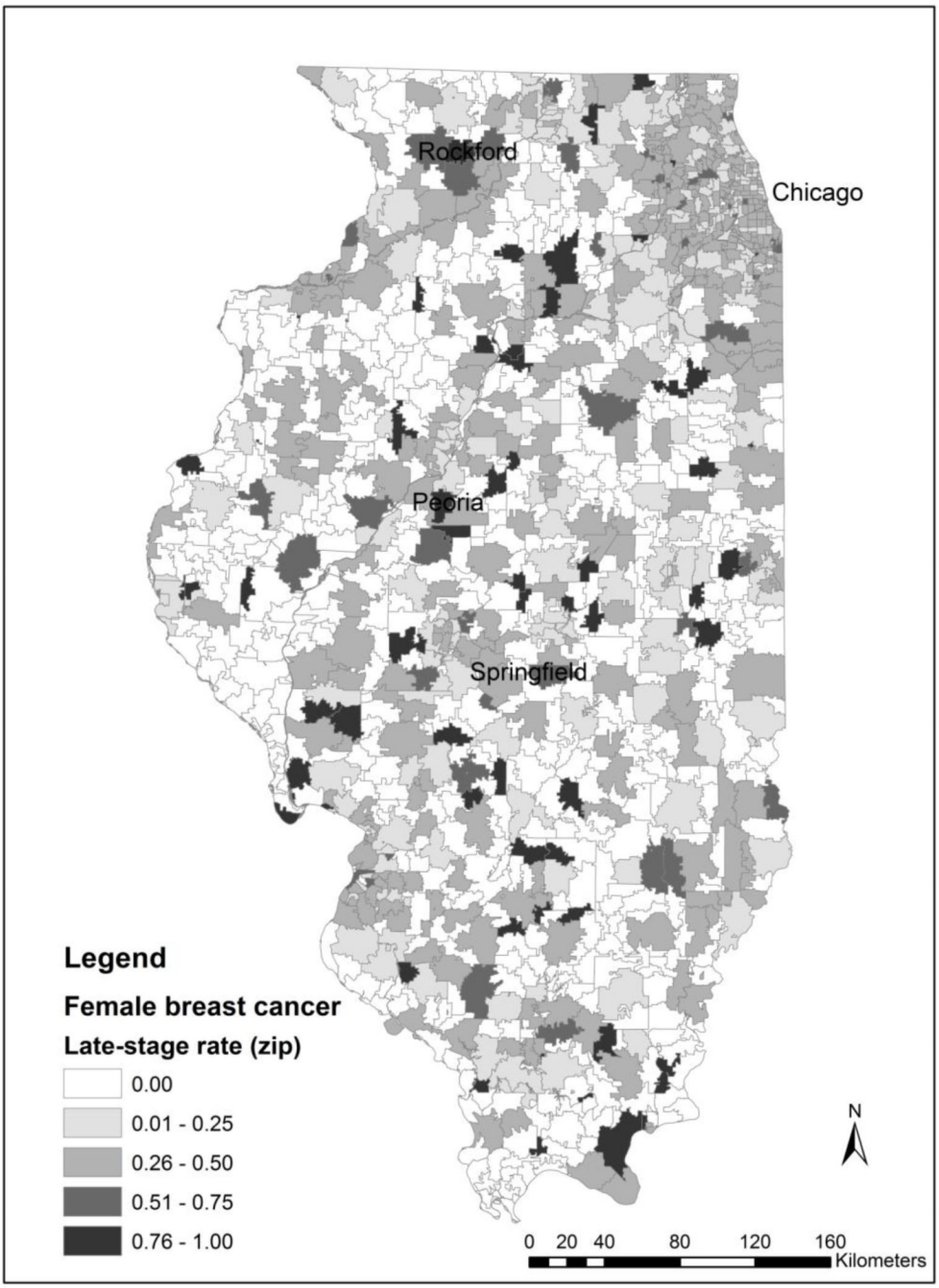


Figure 4.
Late-stage breast cancer rates in zip code areas in Illinois 2000

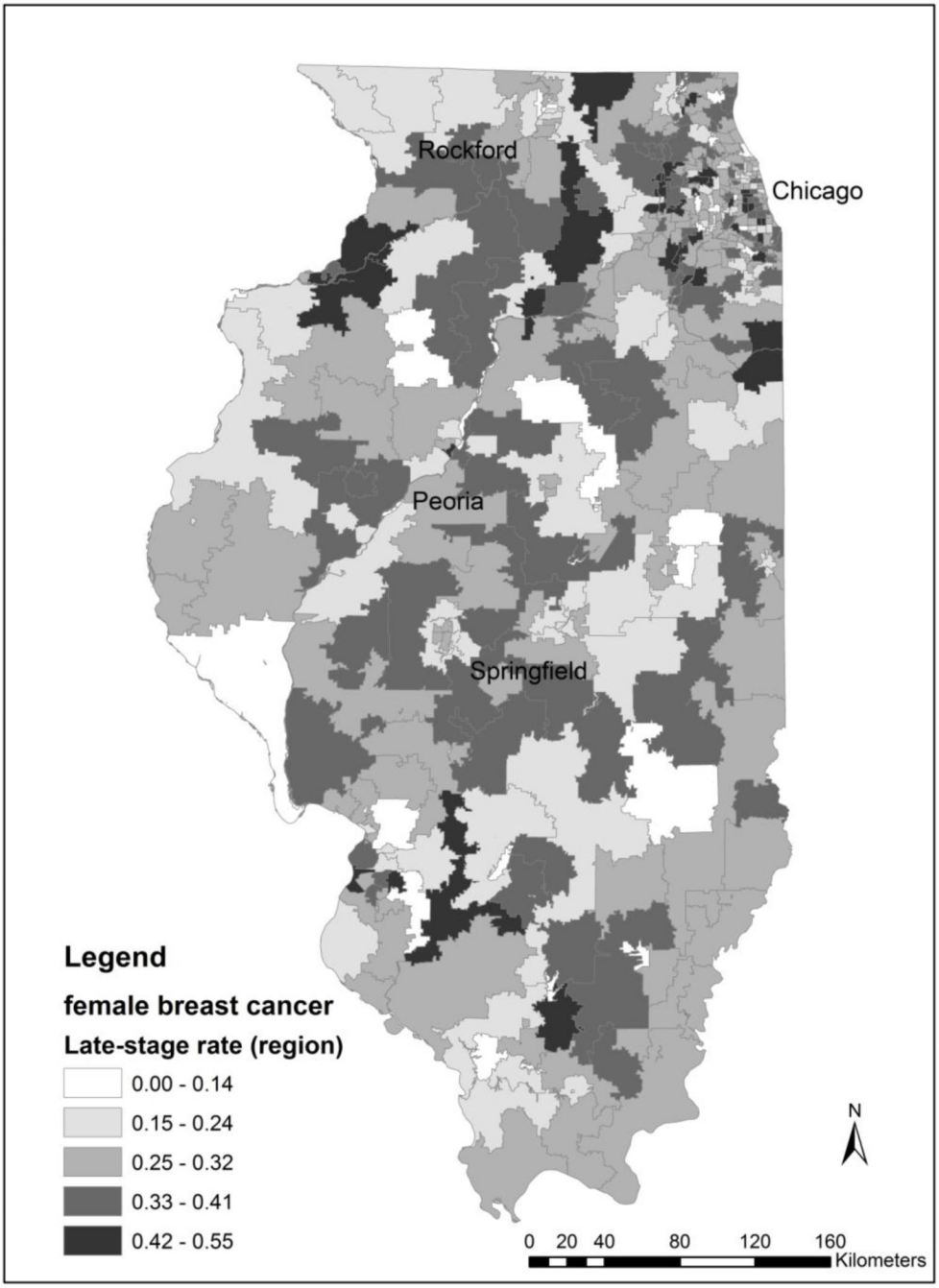


Figure 5.
Late-stage breast cancer rates in newly-defined areas in Illinois 2000

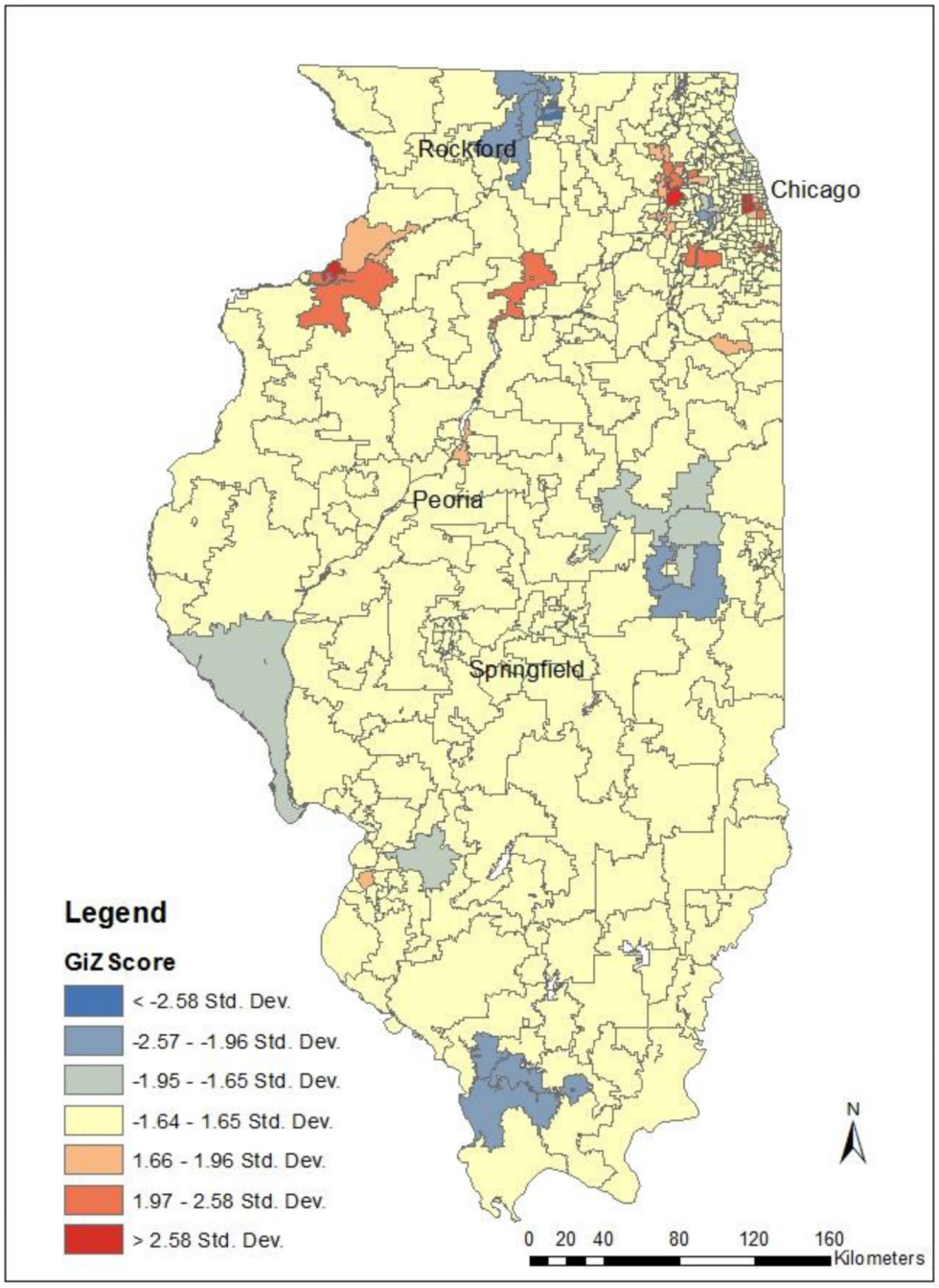


Figure 6.
Hot and cold spots of late-stage breast cancer rates in newly-defined areas in Illinois 2000

Table 1

Factor Structure of Nonspatial Factors

	Socioeconomic disadvantages factor	Sociocultural barriers factor
Non-white population (%)	0.1762	−0.0156
Female-headed households (%)	0.2477	−0.1626
Median income ($)	−0.1692	0.0515
Population in poverty (%)	0.2247	−0.0866
Homeownership (%)	−0.1503	−0.0311
Households w/o vehicles (%)	0.2077	−0.0621
Households with linguistic isolation (%)	−0.2141	0.5546
Population w/o high-school diploma (%)	0.0353	0.2476
Households with >1 person per room (%)	−0.0578	0.3889
Housing units lack of basic amenities (%)	0.0488	0.0933
Variance explained by each factor	*4.6576*	*2.3679*

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

Table 2

Descriptive statistics for female breast cancer by zip code and by REDCAP-defined areas, Illinois 2000

	Total cases	Late-stage cases	Late-stage rate
Zip code areas (n=1,364)			
Minimum	0	0	0 *
Maximum	78	22	1 *
Mean	7.4824	2.1305	0.2843 *
Standard deviation	12.4887	3.7085	0.2755 *
New areas (n=341)			
Minimum	15	0	0
Maximum	78	25	0.5517
Mean	29.9296	8.5220	0.2871
Standard deviation	13.3601	4.4705	0.0951

* Only applies to 943 zip code areas with >0 total cases (excluding 421 zip code areas with 0 total cases in 2000).

Table 3

Regression models for analyzing late-stage breast cancer risks

Variables	OLS	Poisson	Multilevel Logit
Late-stage breast cancer rate	Y		
Number of late-stage breast cancer cases		Y	
Number of all breast cancer cases		Offset	
Individual breast cancer cases (=1 for late-stage, =0 otherwise)			Y
Individual breast cancer patient socio-demographic attributes			X
Neighborhood demographic & socioeconomic factors	X	X	X
Neighborhood urban-rural classification	X	X	X
Spatial access to primary care	X	X	X
Spatial access to mammography facility	X	X	X

Note: Y indicates the dependent variable, and X indicates an independent variable in a model

Appl Geogr. Author manuscript; available in PMC 2013 November 01.

Table 4

Regression results for late-stage breast cancer risks in Illinois 2000

	OLS	Poisson		Multilevel Logit	
	New areas (n=341)	**Zip code areas (n=943, excluding ones with 0 total cases)**	**New areas (n=341)**	**10,206 cases as individual level; zip code areas as area level**	**10,206 cases as individual level; new areas as area level**
Intercept	0.3596***	−1.0646***	−1.0194***	0.3420***	0.3572***
Individual-level					
Black †				0.1103***	0.1032***
Age <40 †				0.1267***	0.1269***
Age ≥70 †				−0.0349***	−0.0352***
Neighborhood-level					
Socioeconomic disadvantages factor	0.0438***	0.1212***	0.1348***	0.0037	−0.0104
Sociocultural barriers factor	0.0474***	0.1465***	0.1542***	0.0503***	0.0522***
Urban-rural classification					
Chicago metro area	0.0212	0.0855	0.0865	0.0037	−0.0040
Spatial access					
Spatial access to primary care	−0.0295***	−0.0870***	−0.1007***	−0.0240***	−0.0278***
Travel time to nearest mammography	−0.0013	0.0006	−0.0012	−0.0001	−0.0006

† indicates a variable of individual cancer patient; others are defined at the area level;

*** significant at 0.001,

** significant at 0.01,

* significant at 0.05.