Journal Title
XX(X):1–15
©The Author(s) 0000
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Geosilhouettes: geographical measures of cluster fit

Levi John Wolf[1], Elijah Knaap[2], Sergio Rey[2]

Abstract

Regionalization, under various guises and descriptions, is a longstanding and pervasive interest of urban studies. With an increasingly large number of studies on urban place detection in language, behavior, pricing, and demography, recent critiques of longstanding regional science perspectives on place detection have focused on the arbitrariness and non-geographical nature of measures of best fit. In this paper, we develop new explicitly-geographical measures of cluster fit. These hybrid spatial-social measures, called *geosilhouettes*, are demonstrated to capture the "core" of geographical clusters in racial data on census blocks in Brooklyn neighborhoods. These new geosilhouettes are also useful in a variety of boundary analysis and outlier detection uses. These new measures are defined, demonstrated, and new directions are suggested.

Keywords

data science, urban sociology, segregation, clustering, unsupervised learning

Introduction

Analysis of spatial community dynamics is a longstanding domain of regional science & urban geography, and a burgeoning concern for spatial data science. One common kind of geography, the "ecologically meaningful" municipal neighborhood (Drukker et al. 2003) is a prime geography used in urban data science. Neighborhoods are often analyzed for their impacts on health (Roberts 1997; O'Campo et al. 1997; Santos, Chor, and Werneck 2010; Spielman, Yoo, and Linkletter 2013) crime (Sampson, Raudenbush, and Earls 1997; Hipp and Boessen 2013) and life outcomes (Duncan, Brooks-Gunn, and Klebanov 1994). However, since these neighborhoods are often defined by government or administrative bureaucracies for convenience's sake, these neighborhood impacts measure the effect of this pre-existing geography, not the geography that might emerge latent in the data (Shelton and Poorthuis 2019).

A different and longstanding mode of analysis focuses on estimating or "bounding" the neighborhood according to some specific objective or known phenomenon under study (Isard 1956). One domain focuses on latent geographies in demographic data–the study of gemodemographics (Harris, Sleight, and Webber 2005; Singleton and Longley 2009; Singleton and Spielman 2014). Geodemographic analysis produces a demographic "typology," or collection of interpretable demographic categories, which are mapped and examined to provide a sense of the social tapestry of a (typically) urban space. In contrast to this, detecting latent ecologically-meaningful communities directly from data is growing more popular in spatial data science. While serious work "bounding" the neighborhood is not new (Galster 2001; Spielman and Logan 2013; Spielman and Folch 2015), the advent of high-quality spatio-temporal data has made this pursuit more feasible (Anselin and Williams 2016; Poorthuis 2018; Arribas-Bel and Bakens 2018; Gibbons, Nara, and Appleyard 2018; Wachsmuth and Weisler 2018). In both geodemographic and latent-neighborhood approaches, these places can be defined consistently in terms of a coherent demographic profile, containing a consistent "bundle" of attributes, behaviors, interactions, marketing, or social ties.

Latent neighborhoods may be used in a similar context as prescriptive administrative ones, but can also be used themselves as indicators of spatial social structure (Morenoff, Sampson, and Raudenbush 2001; Mikelbank 2011) or to study the perceptions or experiences of these boundaries (Hipp, Faris, and Boessen 2012; Duncan et al. 2014). Further, some modes of analysis in this latent

[1] School of Geographical Sciences, University of Bristol
[2] Center for Geospatial Sciences, University of California Riverside

Corresponding author:
Levi John Wolf, School of Geographical Sciences, University Road, Clifton, Bristol, BS8 1SS, United Kingdom
Email: levi.john.wolfbristol.ac.uk

neighborhood vein provide data-dependent geographic frames which can characterize urban dynamics, volatility, and social change (Rey et al. 2011; Duque, Anselin, and Rey 2012). Regardless, these latent spatial neighborhood analyses allow endogenously-determined areas to be identified, characterized, and used in secondary models.

For methods that analyze neighborhoods, it is often necessary to characterize how "cohesive" a detected neighborhood is. Traditional measures of cluster cohesion, or"goodness of fit," do not take into consideration the geography of the data being clustered. Although the analysis of urban "frontiers" or "boundaries" arose early in urban science (Womble 1951) and has seen consistent application in epidemiology (Jacquez 1995; Jacquez, S. Maruca, and Fortin 2000; Lu and Carlin July 2005 2007; Jacquez, Kaufmann, and Goovaerts 2008) & ecology (Fortin et al. 1996; Fitzpatrick et al. 2010), recent work on boundary analysis in demography and urban data science is sparser (Dean et al. 2018; Dong et al. 2019). In these contexts, goodness of fit statistics to measure whether some members, houses, families, or blocks are distinct from a neighborhood's general spatial-social profile; but, the way that this goodness of fit is measured or operationalized is often entirely non-geographic, and has no knowledge of spatial proximity, boundaries, or adjacency (Shelton and Poorthuis 2019). Thus, we develop a new boundary strength measure inspired by ecological & epidemiological methods, but one which is appropriate for urban data science applications like geodemographics and neighborhood-bounding.

In the following work, we explore the trade-offs involved between demographic coherence and spatial integrity in the analysis of urban structure through common geodemographic or neighborhood-bounding methods. Then, inspired by the logic of parametric statistical boundary detection (Womble 1951), we suggest a geographic innovation on Rousseeuw (1987) called the *geosilhouette*. We demonstrate the usefulness of these new measures in both an empirical-descriptive example, assessing the strength & direction of racial boundaries between Zillow neighborhoods over Brooklyn Census blocks, and in latent neighborhood/place learning, where they can be used to characterize the joint spatial-social goodness of fit. Together these new measures provide novel insight into the structure of spatial partitions and enable new analyses of the power of boundaries in quantitative human geography.

Conceptualizing "goodness" of fit

In general, geodemographic & neighborhood-bounding exercises use goodness of fit statistics to characterize the homogeneity or consistency of a given neighborhood or demographic partitioning. Further, measures of segregation are used in a similar fashion for empirical analyses of how thoroughly-mixed (or not) urban spaces are when split by race, class, or other demographic traits. These ancillary measures of neighborhood homogeneity or cluster fit are usually not leveraged directly in theory-driven analyses, but are instead a part of the barely-visible constellation of descriptive statistics used in the heuristic analysis of geographical clusters. To support the wide variety of cluster analyses, there is a similarly-wide set of goodness of fit measures.

Silhouette scores, as suggested by Rousseeuw (1987), are a useful standardized measure of how well an observation fits its cluster. The silhouette score for an observation expresses the relationship between an observation, the other observations in the same cluster, and a counterfactual "next-best-fit" cluster for that observation. In the original presentation of the silhouette score, Rousseeuw (1987) offers an intelligible conceptual motivation for this next-best fit cluster:

> [The next-best-fit cluster] is like the second-best choice for object i: if it could not be accommodated into [its current] cluster A, which cluster B would be the closest competitor? (p.55)

Thus, for demographic data, an observation's next-best-fit cluster is the cluster closest in population profile to that observation but that does not contain the observation.

The silhouette's motivating concepts are clear—"tightness" of each cluster and "separation" between clusters. Each concept has a distinct term in the formal statement of the silhouette score for observation i:

$$s(i) = \frac{\min\left\{\bar{d}_k(i)\right\} - \bar{d}_c(i)}{\max\{\min\left\{\bar{d}_k(i)\right\}, \bar{d}_c(i)\}} \tag{1}$$

where $\bar{d}_m(i)$ is the average distance from observation i to other observations j in cluster m, $j \neq i$. Here, we use c to denote a cluster that contains i and k for any cluster that does not. Taken together, this means that the *minimum* $\bar{d}_k(i)$ represents the cluster k that does not contain i, but whose observations tend to be closest to i on average. This is the "second-best choice" cluster, or the "next-best-fit" cluster (NBFC), since it does not contain i, but is the most similar alternative for i. For observation i, we denote the next best fit cluster $\tilde{k}_i$.

Silhouette values range between -1 and 1, with values close to 1 indicating i is "well-classified" into c. Conceptually, this occurs when $\min \bar{d}_k(i)$ is much larger than $\bar{d}_c(i)$, so the quotient in Eq. 1 is nearly 1. Values close to -1 indicate i is not well-classified into c, since i is much closer to members of $\tilde{k}_i$ than it is to other members of c. For a particularly poor clustering, nearly all i in cluster c may have negative silhouettes, meaning they are closer to some other cluster, $\tilde{k}_i$, than they are to their own cluster. In

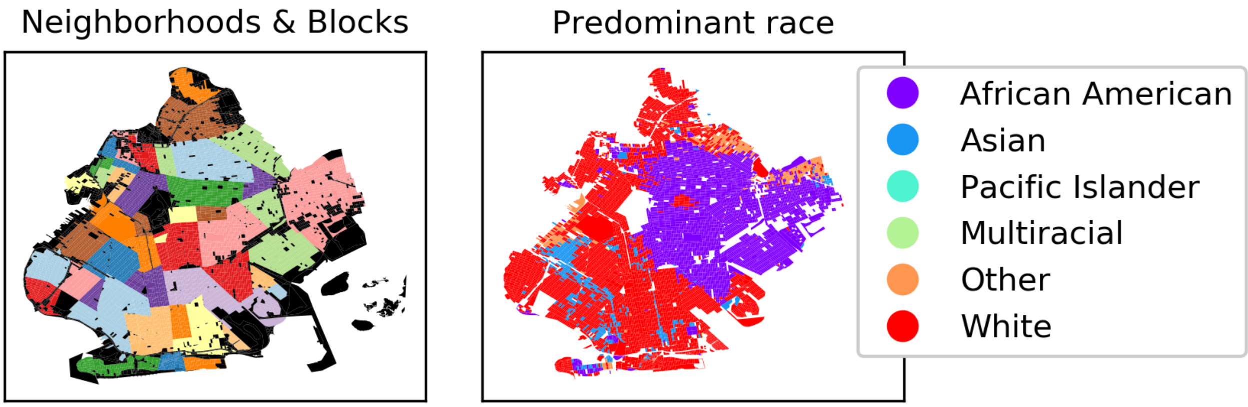

Figure 1. Zillow neighborhoods (left), with 2010 Census blocks with nonzero population under-laid for Brooklyn, NY. Tracts with low populations are shown in black on the left figure, and omitted entirely from the remainder of the analysis. The single most-predominant race in the study area is shown on left.

light of this, the median silhouette score within each group is often used to characterize the goodness of fit of that group overall, and the map median characterizes the fit of the map as a whole. In contemporary applications, silhouettes are often used to identify an appropriate number of clusters, as well as being used to identify outlying observations in clusters, or clusters with exceptionally poor fit. It is in the second sense, as a measure of the goodness of fit or outlier detection, that we extend the silhouette.

Data: Neighborhoods & Endogenous Racial Clusters

In part due to their simplicity, silhouettes have long been used to detect observations that are not well-grouped with their cluster. However, for geographic analysis, next-best-fit scores can be made more informative. As it stands, the next best fit cluster represents a group to which observation i can be most plausibly reassigned—the “second-best choice.” What “best” means is more complex in geographical analysis, though.

To examine various kinds of geographic “second-best choice“ cluster assignments, we examine self-reported race in the 2010 Census blocks across neighborhoods in Brooklyn, NY using the neighborhood boundaries provided by Zillow.[*] One view of this data is provided by Figure 1, which demonstrates the populated census blocks, neighborhood boundaries, and provides an indication of the racial composition across Brooklyn blocks. It is important to note that the neighborhoods shown on the left side of Figure 1 are different from the official government neighborhoods maintained by the New York Transit Authority (NYTA); there are 7 more of the “Zillowhoods” than the NYTA Neighborhoods and Zillowhoods are smaller on average. This is likely done to keep the overall size of communities more consistent in the NYTA definitions than in the Zillow neighborhoods, which derive from how properties are marketed. Regardless, both the NYTA and the Zillow neighborhoods serve here as exogenously determined neighborhood boundaries for our purposes, so their relative arbitrariness is not under analysis here. To contrast with these exogenous neighborhoods, we also will analyze detected clusters in the racial composition of census blocks in the 2010 US Census using an aspatial K-means approach common in geodemographics (Harris, Sleight, and Webber 2005) and a spatial-hierarchical agglomerative clustering heuristic based on Ward’s method (Ward 1963).[†]

Fragmentation in Urban Regions

Fundamentally, the idea of cluster quality in spatial cluster analysis implicates two distinct concepts: *attribute* coherence, that an observation’s characteristics are similar to its cluster; and *spatial* coherence, that the cluster itself demarcates or delineates a geographically-coherent “zone” or *region* of the overall problem frame.[‡] To varying degrees, “real” neighborhoods generally exhibit both demographic coherence and spatial coherence: they are a “bundle of spatially-based attributes associated with [a] cluster of residences” (Galster 2001, p. 2112). Both the “bundle

[*] These can be accessed at https://www.zillow.com/howto/api/neighborhood-boundaries.htm and are licensed as a Creative Commons dataset.

[†] These results, as well as all estimators developed here, depend on the efforts of `sckit-learn` (Pedregosa et al. 2011) and `pysal` (Rey and Anselin 2007). They will be made available alongside this manuscript as free and open software in `pysal`.

[‡] This is the intended meaning of “ecologically meaningful” used by Drukker et al. (2003).

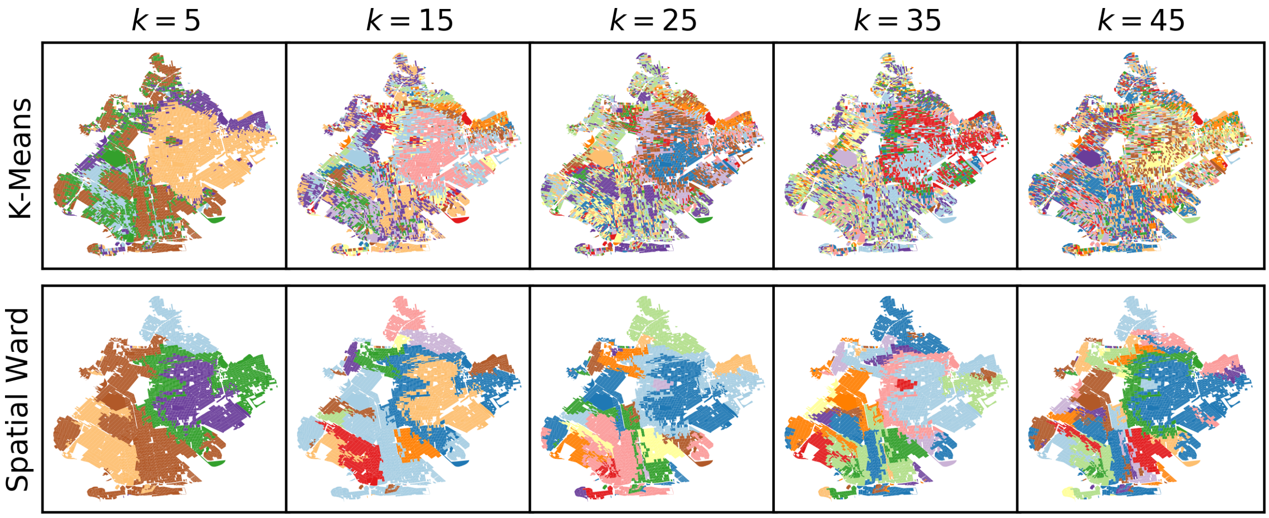

Figure 2. Demographic clusters in Brooklyn, NY for k-means and spatially-constrained Ward agglomerative clustering. Fragmentation increases dramatically as the number of clusters increases.

of attributes" and the "spatial cluster" are needed to characterize a classification's fitness in a geographical process.

However, silhouettes (like nearly all goodness of fit measures for clustering) exclusively measure attribute coherence. This is fine for non-spatial clustering applications, but is difficult to justify in geographical applications. Indeed, for the contiguous regions used in the neighborhood dynamics and neighborhood effects literature, nearly all of the "second-best choices" constructed for silhouette scores are actually *infeasible choices*: i might be nowhere near $\tilde{k}_i$ geographically. If i were to move from c to $\tilde{k}_i$, both c and $\tilde{k}_i$ would cease to be geographically coherent. Since i can not feasibly be reassigned to $\tilde{k}_i$, the counterfactual "second-best choice" considered by the silhouette score is moot.

Acknowledging this, we can leverage observations' spatial contexts (in addition to their group memberships) to extract more meaningful information about neighborhoods or spatial clusters themselves. Observations on the boundary of a spatial cluster are the only ones that could be connected to their next-best-fit spatial cluster if they were reassigned. All other *interior* observations require more than one block to be reassigned in order to be a feasible, internally-connected cluster. As clusters become less geographically coherent, the size of their interior decreases. Visually, the clustering solutions shown in Figure 2 illustrate this: as the number of clusters increases, the spatial fragmentation of clusters increases quickly.

Another view of this fragmentation is provided by Figure 3. In this composition plot, the share of all blocks that are *interior* to the cluster—those that only touch other blocks in the same cluster—is represented by the gray area. The blue fraction shows blocks that are touching their next-best-fit cluster. These are blocks where the "second-best choice" assignment is feasible, since the block could be reclassified to its second-best choice and not affect the spatial fragmentation in the map. Finally, the red area denotes the share of blocks that are on the boundary of their own cluster, but are not near any member of their next-best-fit cluster. These are the blocks where a "second-best choice" assignment would affect territorial integrity. In addition to the shares from latent/discovered neighborhoods, we show "empirical" fractions of the same quantities: census blocks in Zillow neighborhoods that touch a neighborhood that is next-most demographically similar to the block itself, or that are on the boundary of a neighborhood but do not touch a neighborhood that is next-most demographically similar. These are shown by the tickmarks on the right side of the plot.

Interpreting Figure 3, we can understand a few things. First, as is mathematically necessary, the share of interior blocks declines as the number of clusters increases. Second, despite this increasing fragmentation, the number of blocks that touch their NBFC is relatively stable as the number of clusters increases. This occurs quickly for the spatial agglomerative clusters, but both are remarkably stable at around $k = 20$. Third, we see that groups defined without spatial information (k-means) tend to be much more fragmented than either the empirical neighborhoods or the clusters discovered using the explicit spatial clustering technique. The fraction of blocks that is interior to a cluster is consistently smaller in the aspatial k-means map, and clusters are much more dramatically interspersed. The most spatially-coherent solution seen in the k-means

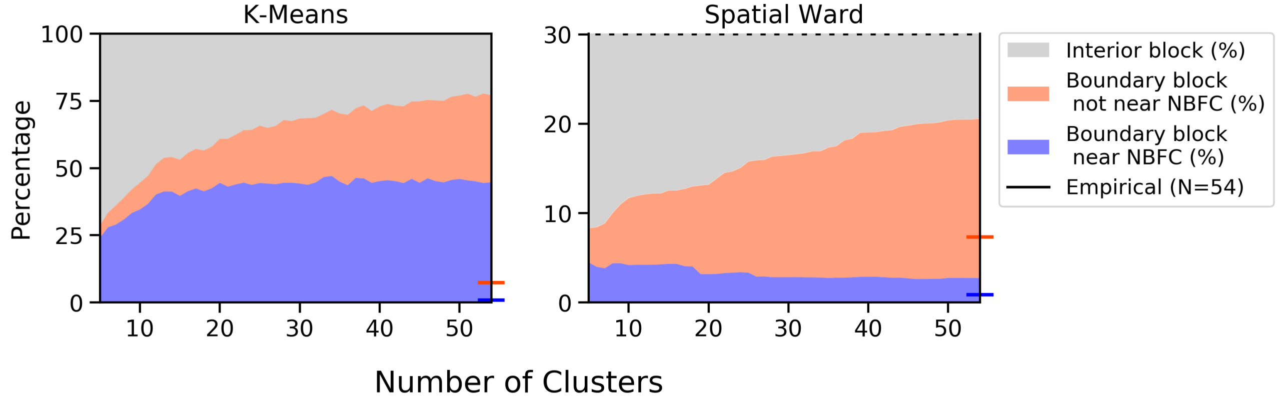

Figure 3. Breakdown of census block types based on their boundary & next-best-fit cluster relationships for clusters shown in Figures 2 & 1.

clustering solutions (that with the smallest k) is still more fragmented than the most fragmented spatial agglomerative clustering solution (that with the largest k). Finally, the empirically-observed breakdowns are quite low; given that only 7% of the 7729 blocks with non-zero population sit on the boundary between two Zillow neighborhoods, only 11% of these boundary blocks (approx $.08\%$ overall) are themselves near their next-best-fit clusters. Thus, the level of spatial cohesion in the Zillow neighborhoods is much higher than even those detected using the spatially-explicit clustering method.

Silhouette Scores are not Spatial

Focusing on the empirical case, the interplay between these NBFCs and the silhouette scores is shown in Figure 4. Note that the preponderance of silhouette scores are *negative* for real-world neighborhoods. This means that, in terms of their racial composition, census blocks are nearly *always* more similar to a different neighborhood than they are to the neighborhood in which they reside. Together, this suggests that silhouette scores will always favor "tighter" clusters in attribute space, without regard for spatial feasibility or geographical plausibility. Indeed, any realistic urban place-geography will be considered less "tight" by the silhouette score, since attribute coherence and geographic coherence are often opposing objectives. By the same logic, any spatially-informed clustering method *must also be* less "tight." Neighborhoods (empirical or embedded within the data) are often much more diverse than the maximally-homogeneous demographic partition to which the silhouette refers, so any measure of cluster fit that does not consider their inherent spatiality will demonstrate this behavior.

Indeed, neighborhood social homogeneity should not be regarded as a necessarily intrinsically-desirable normative objective when conducting place detection. Social scientists have long argued that diverse and socially integrated neighborhoods provide benefits to residents when they are able to foster meaningful social exchanges (Joseph, Chaskin, and Webber 2007; Chaskin and Joseph 2013; Talen and Koschinsky 2014; Tolsma and Meer 2018). Further, there is evidence that neighborhood diversity in the United States is increasing, carrying important benefits for residents: methods that distill neighborhoods according to maximum demographic homogeneity may be overlooking important aspects of they ways that neighborhoods are *experienced* by their residents (Logan 2013). As trends towards diversification continue, there is also recent evidence that neighborhood boundaries are perceived differently among residents from different social backgrounds (Hwang 2016), too. Together, this suggests that neighborhood definitions are tenuous, occasionally contested, and may be defined by attribute homogeneity, resident perception, or physical demarcation–and each of these definitions has unique value in different research contexts.

While silhouette scores are particularly useful for identifying spatial configurations of attribute homogeneity, (such as racial and ethnic enclaves) the point we raise here is that other definitions are important and useful for other research questions; building explicitly geographic measures of fit is necessary to improve the validity of geographical work on urban regions. Therefore, *contra* Shelton and Poorthuis (2019), it is *not* the designation of a "best fit" criterion itself that harms the construct validity of detected places; it is the inflexibility, simplicity, and arbitrariness of these criteria that makes detected regions uninteresting or unhelpful. For more interesting and helpful computational geographies, it is necessary to improve, develop, and strengthen the conceptualization and operationalization of these measures of best fit. In short, we need better geographically-aware measures of cluster fit.

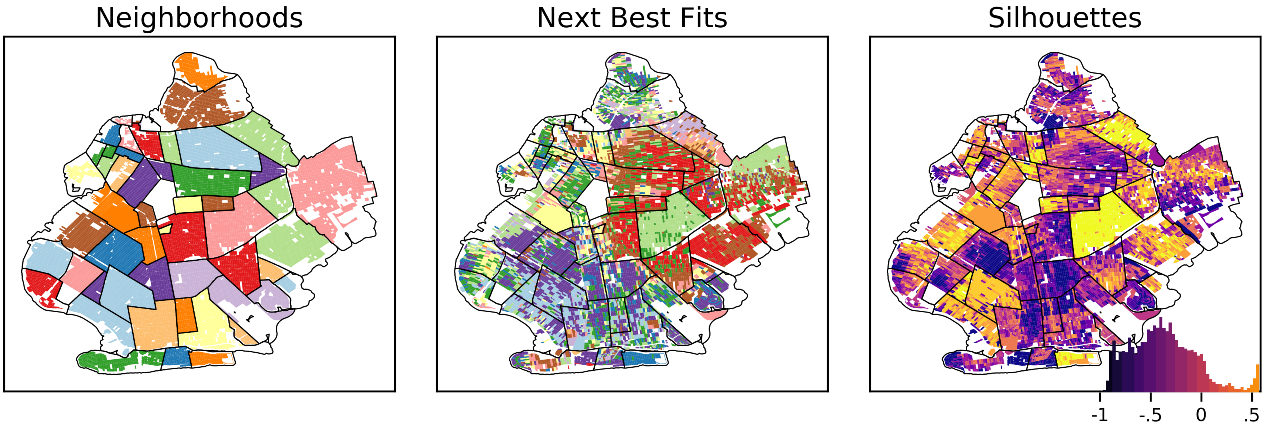

Figure 4. Sillhouettes & Next-Best-Fit Clusters (NBFCs) for Census Blocks within Zillow neighborhoods in Brooklyn, NY.

Geosilhouettes: Measures of Spatial Cluster Similarity

Thus, for many forms of empirical urban analysis, it is necessary to develop better methods to characterize the local similarity in geographical regions. Further, such a measure should bridge both the local *attribute coherence* in a way that respects or controls for *spatial coherence*. While others may examine Figure 2 and observe simple or straightforward fragmentation in shape (McGarigal et al. 2002), we instead take inspiration from literature on the statistical analysis of boundaries (Jacquez, Kaufmann, and Goovaerts 2008); both the geography and demography matter when defining "best fit" social-spatial regions. While the percentage measures used in Figure 3 are useful to describe a single map, it is unclear what the "expected" or "neutral" value of these percentages are. For some geographies and attribute distributions, it may be quite difficult to achieve even 1% NBFC proximity; the "natural" level of proximity for a given geography is unknown. Thus, these percentage-style measures are inherently map-specific and difficult to generalize, and so should only be used to characterize the relative quality of a solution over a given map.

Thus, again like silhouettes, social-spatial measures of cluster fit should be comparable between maps, have some finite range, and have a theoretically-useful zero point. Using a measure with a similar structure and meaning to the silhouette in Eq. 1 is desirable, since silhouettes have a long history in unsupervised learning, are well-understood, and are not map-specific. Below, we derive two geosilhouette specifications. One, the so-called *path* silhouette, focuses on joint attribute-spatial affinity through the use of so-called *dissimilarity paths*. The other, the *boundary* silhouette, restricts the set of each observation's next-best-fit clusters to *only* those clusters that are nearby. That is, the *boundary* silhouette constrains the next-best-fit cluster to be a feasible cluster reassignment (Duque, Church, and Middleton 2011). These two methods will be derived, discussed again in descriptive & normative/inferential applications for the analysis of race in Brooklyn Census blocks.

Path Silhouettes

One way to make the silhouette score geographically aware is to use *dissimilarity paths* rather than focusing on attributes alone. A *dissimilarity path* models the dissimilarity between two observations, i and j, as a function of the total dissimilarity between observations along the path connecting them. Underlying this model of spatial-social dissimilarity is the recognition that, in order for i and j to be included in the same geographically-contiguous cluster c, they must connected by a set of observations also in c. Thus, a "path" silhouette is a silhouette score computed using the length of the *dissimilarity path* from i to j as the the distance from i to j, rather than the dissimilarity of i and j's data directly.

For a path silhouette, first consider the $N \times N$ matrix, $\mathbf{D}$, containing every pair of distances between observations i and j. Recalling that $d_k(i)$ takes the ith row of $\mathbf{D}$ and computes the average of all j columns in cluster k, it is sufficient to modify $\mathbf{D}$ to account for spatial structure and use the same method to compute a silhouette score. to do this, we will build a complement to $\mathbf{D}$ that models the spatial relationships between observations in the problem. This can be represented by $\mathbf{W}$, an $N \times N$ spatial affinity matrix. $\mathbf{W}$ may be binary, reflecting a near/not near classification common for describing adjacency in lattice data, k-nearest neighbor proximity, or buffer/distance banding proximity. Alternatively, $\mathbf{W}$ may be continuous, formed from a spatial kernel function or a collection of spatial basis functions (Bradley, Wikle, and

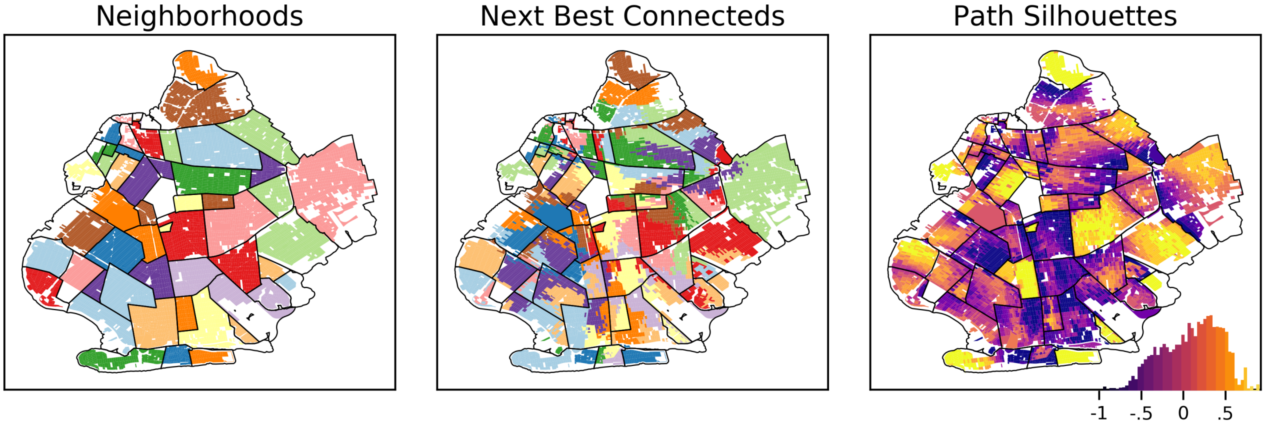

Figure 5. Neighborhoods and next-best-fit clusters using the *path dissimilarity metric*. This is the path silhouette analogue of Figure 4.

Holan 2015). Regardless, $\mathbf{W}$ should represent symmetric spatial relationships between all pairs of observations.

Bringing $\mathbf{D}$ and $\mathbf{W}$ together, a path silhouette can be constructed on $\mathbf{C}_1$, the first-order cost matrix.

$$\mathbf{C}_1 = \mathbf{D} \circ \mathbf{W} \tag{2}$$

where $\circ$ denotes element-wise (Hadamard) matrix products. The shortest path from each observation to every other observation can be computed using $\mathbf{C}_1$ as the attribute-weighted spatial adjacency matrix, or *cost matrix*, to provide a complete set of dissimilarities used in a silhouette score. Let this $N \times N$ matrix of all-pairs shortest path lengths found in $\mathbf{C}_1$ be called $\mathbf{C}$. In $\mathbf{C}$, the distance between i and j is modeled by the sum of the path lengths in attribute space connecting i and j in geographic space. When $\mathbf{W}$ is complete, then every path in $\mathbf{D}$ is shortest due to the triangle inequality, so $\mathbf{C}_1 = \mathbf{C}$. Otherwise, $\mathbf{C}_1$ expresses only the first-order path costs between observations, and $\mathbf{C}$ must be constructed using standard all-pairs shortest path algorithms (e.g. Floyd 1962). [§]

The path silhouette is then computed using the same formula as in Eq. 1, using $\mathbf{C}$ instead of $\mathbf{D}$. Since an observation's next-best-fit cluster ($\tilde{k}_i$) on $\mathbf{D}$ alone will often not be the next-best-connected cluster in $\mathbf{C}$, let us denote the next-best-connected cluster to observation i as $\overset{\circ}{k}_i$ to make this clear. The path silhouette expresses the difference between the average *path length* from i to $j \in \overset{\circ}{k}_i$ and i to other $j \in c$.[¶] When the path silhouette is close to 1, it indicates that i has short attribute-weighted paths to other $j \in c$, so it is either extremely close to $j \in c$, extremely similar to $j \in c$, or some combination thereof. Alternatively, path silhouette scores close to -1 indicate that i is much easier to connect to elements in $\overset{\circ}{k}_i$ than to other elements c; again this can be driven by spatial and/or social factors. Finally, this method can be used in clustering problems in other spatial supports as well, so long as the structure of spatial relationships can be represented in an appropriate $\mathbf{W}$.

An example of this approach to analyzing cluster quality can be seen in Figure 5. In this color ramp, the darker purple areas are those where an observation is classed as not well-fit to its cluster (since the silhouette is negative), and lighter yellow are areas where the observation is well fit. This is in the same style as Figure 4, but shows the *path silhouette* versions: the "next best fit" cluster becomes the "next best connected" cluster, and the silhouettes shown are the *path silhouette* variant.

These maps show a few things. First, the geographically remote neighborhoods in the far north, west, and south of Brooklyn exhibit strong joint spatial-social cohesion due to their joint social coherence and geographical remoteness. Second (and more critically), the empirical neighborhoods with path silhouettes closer to 1 in Figure 5 tend to remain together in the spatially-informed clusterings in Figure 6, even when they are in more central areas of the city. Since the path silhouette measures joint spatial-social similarity, it is reasonable that the spatial agglomerative clustering picks up on this. However, there is no constraint forcing this to

[§] In the case where $\mathbf{W}$ is quite sparse, this may be feasible for large N using dedicated sparse algorithms. Further, since $\mathbf{D}$ is non-negative and $\mathbf{W}$ is connected, then $\mathbf{C}$ is guaranteed to be well defined. When $\mathbf{C}_1$ has disconnected sub-graphs, no path will exist to connect them, meaning $\mathbf{C}$ will have infinite values. Thus, the connected sub-components of $\mathbf{C}_1$ should be analyzed separately if this occurs.

[¶] It is also possible to define the next-best-connected cluster solely by the single shortest cost of connection between $j \in k$ to i, rather than the shortest average cost of connection over all $j \in k$ as we suggest. However, this is also not in the spirit of the original average-of-cluster dissimilarities used by Rousseeuw (1987), and so is elided here.

occur, so this reinforces the utility of the path silhouette as an exploratory measure of the local spatial-social coherence for urban regions.

To illustrate, central Brooklyn has many neighborhoods with majority-African American populations, as shown in Figure 1. The aspatial silhouettes shown in Figure 4 show one quite clearly: East Flatbush. Recalling its atypically-high silhouette scores in Figures 4 and path silhouettes in 5, this area in the deep center of Brooklyn is spatially & socially distinctive. This distinctiveness is recognized regardless of cluster heuristic. In this area, both the silhouettes and the path silhouettes are high, showing this is an area with significant *demographic* homogeneity (a bundle of similar attributes) that also is spatially-coherent (this bundle clusters geographically). Notably, though, high path silhouettes still betray spatial-social similarity in demographically more-complex neighborhoods, such as Bushwick, along the northeast Queens-Brooklyn border. This area is not as strongly self-similar (in that it is not predominately of a single race in the Census classifications), but its profile *is* still distinct from other nearby neighborhoods. Path silhouettes pick up on this weaker form of spatial-social similarity, too. However, this does not stand out in the aspatial silhouettes, again regardless of the clustering heuristic. The "core" of this area is assigned its own cluster in the spatial Ward clustering, and shares the same high path silhouette values as the empirical neighborhoods. Thus, spatially-informed measures of cluster fit like the path silhouette can help us identify what parts of a given cluster are spatially- and socially-distinctive, while grounding empirical descriptions of existing or latent regions.

Boundary Silhouettes

While path silhouettes are a novel and potentially-useful measure of the joint social-spatial proximity of observations, it too suggests a somewhat unrealistic "second-best choice" counterfactual: when computing the "next-best-fit" cluster, the cost of moving i from c to k is modeled by the average length of paths from i to $j \in k$ that captures both spatial and social distance. But, neighborhoods, recovered or received, are usually not point-to-point paths. While we believe the joint geographically- and data-weighted path lengths is a better model of spatial reassignment costs than nothing at all, it remains only one possible model of the *actual* reassignment costs, which will be different for every heuristic and clustering objective. So, we suggest a second, more conservative measure of spatial-social proximity in clusters & regions: consider only those observations that might be reassigned without affecting other assignments. That is, focus on the cluster boundaries.

Consider that for each i, the region $\tilde{k}_i$ is identified while computing each silhouette score. In the standard silhouette, $\tilde{k}_i$ has no predetermined spatial relationship to i. Often, it is geographically distant from i. In fact, practically speaking, i's next-best-fit cluster may never plausibly contain i depending on the unique geographical structure of the clustering problem. Whereas the path silhouette considers the cost of *connecting* i and all elements in other $\tilde{k}_i$, this more conservative measure searches only for next-best-fit clusters that are near i.In this way, we are constructing the *best local alternative* cluster for i, instead of the next-best-fit cluster over the entire map.

In light of this, a boundary silhouette is defined as a restriction of the standard silhouette score. Reprising the original silhouette statement from Equation 1:

$$s(i) = \frac{\min\left\{\bar{d}_k(i)\right\} - \bar{d}_c(i)}{\max\{\min\left\{\bar{d}_k(i)\right\}, \bar{d}_c(i)\}} \qquad (3)$$

the boundary silhouette must restrict $\min \bar{d}_k(i)$ to only those where i could already be reassigned, without affecting any other j. So, to disqualify distant alternatives, for any k that is not near i (for any geographical operationalization of near), $\bar{d}_k(i)$ is set arbitrarily high. Then, our target counterfactual "second-best choice" for i—called the *best local alternative* cluster—has three properties: (A) it does not contain i, (B) it is geographically near i, and (C) it has the lowest average attribute dissimilarity to i. It is helpful to denote this as $\hat{k}_i$, since it is often the case that $\hat{k}_i \neq \tilde{k}_i$; as we discussed in the *Silhouette Scores are not Spatial* section and show in Figure 4, $\tilde{k}_i$ is often nowhere near i itself. These best local alternative clusters are quite restricted. In fact, depending on the notion of geography used to define *local* and the relative scales of the clusters and what is being clustered, there may only be one or two alternative clusters near i.[‖] It also may be true that $\hat{k}_i$ is not even a particularly good fit for i in attribute space. But, since $\hat{k}_i$ is the best cluster for which i can be reassigned without affecting other observations, it also is the best feasible second choice.

Using this idea, the boundary silhouette is the silhouette-style score between i, c, and $\hat{k}_i$, defined only for i on the boundary. To build the set of observations on the boundary, first let us use $\eta(i)$ to mean the set of all observations j that are local/nearby i. Then, the set of observations on the boundary are all i for which at least one element of $\eta(i)$ falls in a different cluster than i's cluster, c. This set of boundary observations is then:

$$\mathcal{B} = \bigcup_{i}^{N} \{i \,;\, k_j \neq c \quad \exists\, j \in \eta(i)\} \qquad (4)$$

[‖] For observations deep within the interior of a cluster, where there is *no* best local alternative, we adopt a similar convention to Rousseeuw (1987) and set their boundary silhouette to zero.

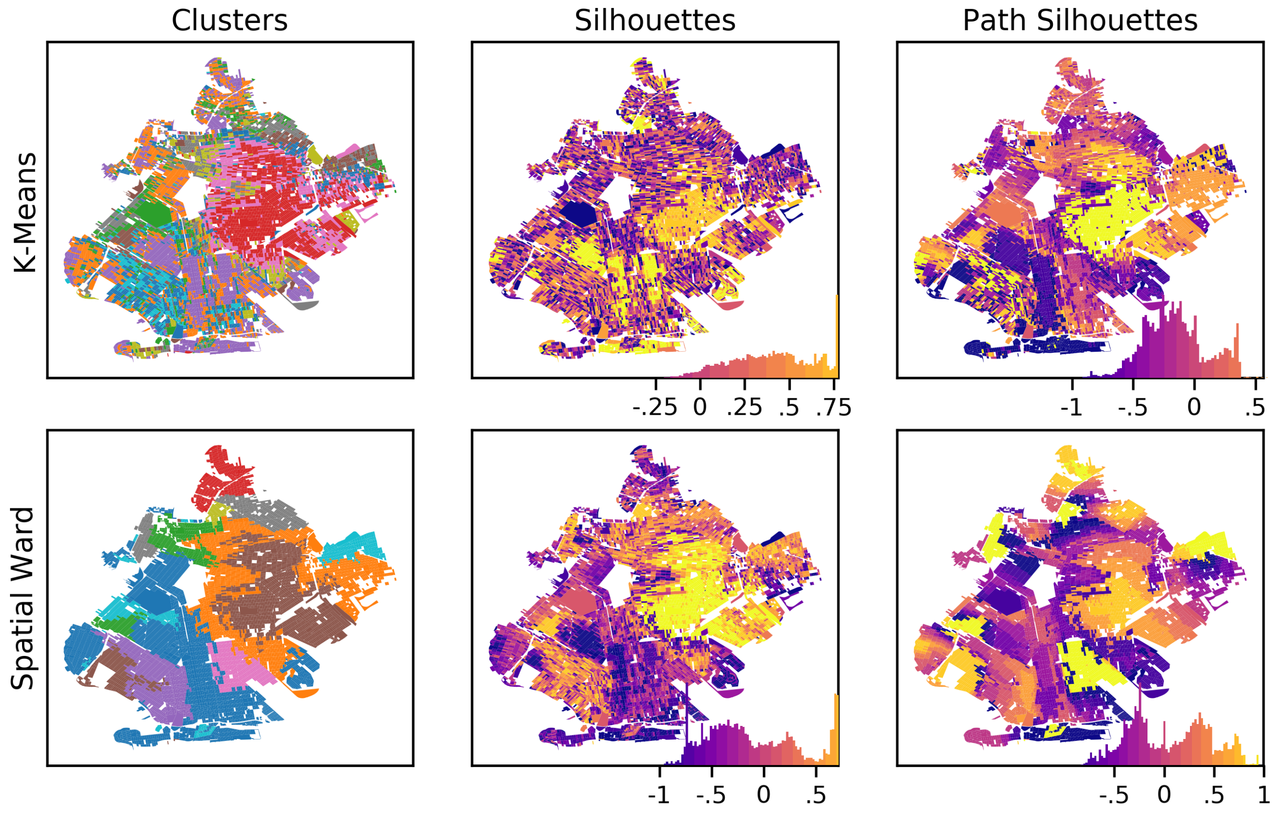

Figure 6. Assignments to 15 clusters, silhouettes, and path silhouettes for aspatial k-means and spatial Ward agglomerative clustering are shown above.

Second, the set of clusters around site i can also be define in a similar fashion:

$$\mathcal{A}_i = \{k_j; k_j \neq c \;\; j \in \eta(i)\} \tag{5}$$

Together, these definitions are sufficient to define the *boundary silhouette*. The "best local alternative," the boundary silhouette's version of the next best fit cluster, is the cluster in $\mathcal{A}_i$ that is most similar to i. With this understanding of i's best local alternative, we can state the boundary silhouette as a familiar ratio of within- and between-cluster distances:

$$s_b(i) = \frac{\min_{\mathcal{A}_i}\left\{\bar{d}_k(i)\right\} - \bar{d}_c(i)}{\max\left\{\min_{\mathcal{A}_i}\{\bar{d}_k(i)\}, \bar{d}_c(i)\right\}} \quad \forall \quad i \in \mathcal{B} \tag{6}$$

This score has the same interpretation as Rousseeuw (1987)'s silhouette discussed in Section , but measures the cost of "flipping" i over the border of c and $\hat{k}_i$. In this way, it is again a silhouette score, in that it only considers the attribute distance between i & c or i & $\hat{k}_i$, but it uses a geographical constraint common in spatial clustering: $\hat{k}_i$ should be near i.

Intuitively, when the boundary silhouette is negative, i is more similar to its best local alternative than it is to its home cluster. So, i could "cross" the boundary and improve the attribute coherence of both regions without affecting the spatial coherence of the regions significantly. At its widest, $\hat{k}_i$ can represents a very general class of "second-best choice" clusters—regions where the reassignment of i from c to $\hat{k}_i$ yields the best feasible reassignment. Further, for spatial clustering problems with stronger constraints, a stronger notion of feasibility can be adopted by imposing more restrictions on $\mathcal{A}_i$, the set of local alternatives for i.[**] Since locality or geographic coherence may only be one of many relevant constraints on feasible reassignments for i, this generality means this particular statistic is remarkably flexible.

The boundary silhouette exhibits an interesting property: it can be asymmetric for any boundary. For cluster k and cluster c, the median boundary silhouette score for observations in k bordering c may not necessarily be equal to the score for observations in c bordering k. This would imply that observations in c neighboring k may be more similar to those in k than they are to their own cluster,

[**] Common additional constraints in the literature include population minima or compactness requirements, such as in redistricting contexts.

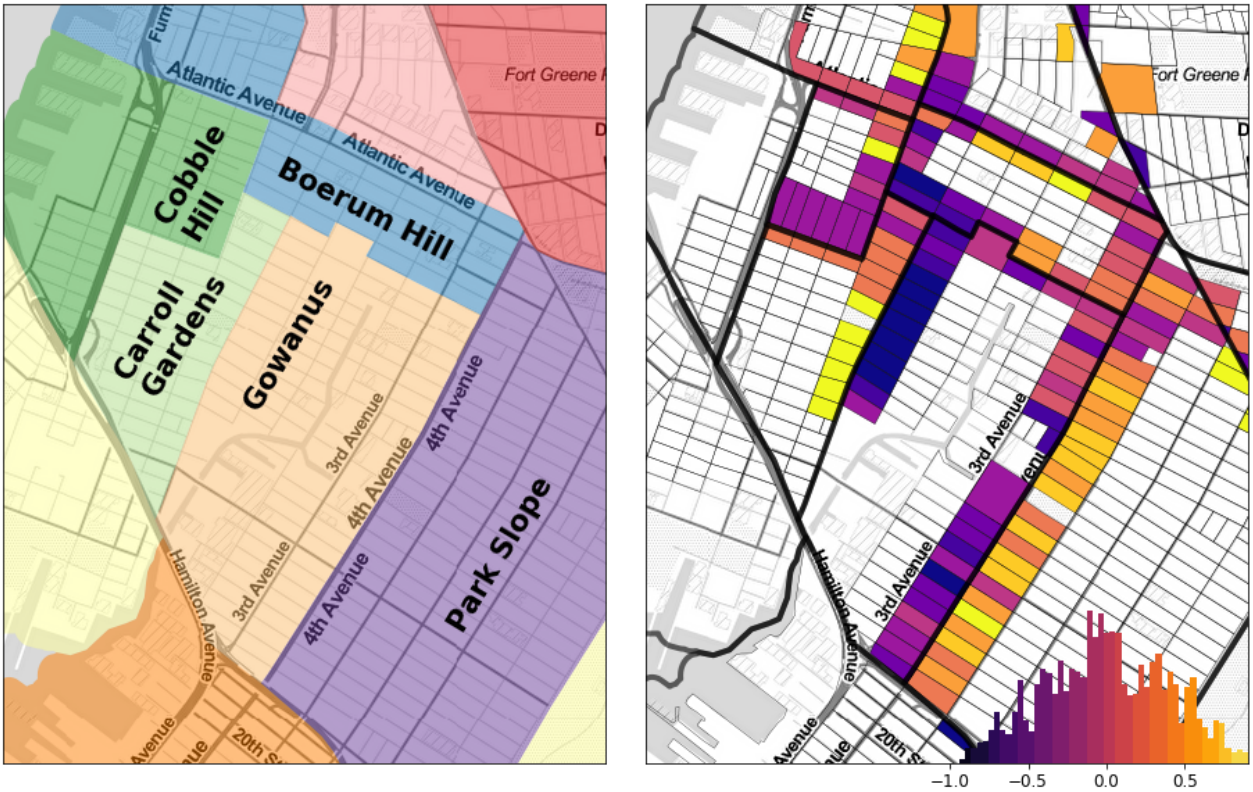

Figure 7. Detail of downtown Zillow neighborhoods in Brooklyn, with boundary silhouettes overlaid. Legends on the bottom-right of each view demonstrate the visible distributions of mapped boundary silhouettes. Basemaps are provided by Stamen Design.

neighbor focal	Boerum Hill	Cobble Hill	Carroll Gardens	Gowanus	Park Slope
Boerum Hill	0.000	-0.32	-0.358	0.274	0.122
Cobble Hill	0.627	0	-0.156	0.639	-
Carroll Gardens	0.339	0.152	0	0.710	-
Gowanus	-0.071	-0.359	-0.647	0.000	-0.168
Park Slope	0.050	-	-	0.390	0

Table 1. Median boundary silhouette values for blocks abutting each cluster in downtown Brooklyn neighborhoods. The rows record blocks in the “focal” cluster that tough the “neighbor” cluster.

but observations in k neighboring c are still closer to k. Alternatively, it may be the case that both sides of the boundary are poorly (or well) classified, indicating the clusters are poorly (or well) separated.

Practically speaking, when both sides of the boundary have a positive median boundary silhouette, it means that the parts of the clusters immediately adjacent to one another are strongly distinct. When both are negative, it suggests that the neighborhoods may be misaligned from the true underlying demographic difference in that locality. When one is positive and one is negative, the cluster on the positive side of the boundary tracts could merge with the tracts on the boundary and improve the local structure of fit without adjusting the spatial coherence of the two clusters. Thus, the boundary silhouette asses the *local* goodness of fit for a cluster by centering on the relative similarities between candidate reassignments on the cluster’s boundary.

For an example, we provide an empirical illustration of the boundary silhouettes in north-central and downtown Brooklyn in Figures 7 & 8. In addition to the figures, the median boundary silhouette values for each adjacent neighborhood pair is provided in Tables 1 & 2. On the left of each plot, the neighborhoods are labeled On the right, the boundary silhouettes are shown.

A few strongly asymmetric boundaries are apparent. Looking at the strongest asymmetry, blocks in Gowanus near Carroll gardens are more similar to Carroll Gardens than the rest of Gowanus, while the tracts in Carroll Gardens bordering Gowanus are much more similar to Carroll Gardens. Thus, the demographic profile of Carroll Gardens is a better demographic fit for those boundary tracts in Gowanus, so the similarity is directional, and the two neighborhoods may appear to change gradually in demographic composition when moving from Gowanus into Carroll Gardens. This contrasts with a socially undirected boundary, such as the one between Bedford Stuyvesant and Bushwick in figure 8. For boundaries with positive scores on both sides, social characteristics change remarkably between the boundary and its adjacent cluster. Blocks in Bushwick immediately north of Broadway Boulevard, simply could not easily be demographically passed off as a typical Bedford-Stuyvesant block.

In addition, some neighborhoods may be quite internally heterogeneous and still have positive boundary silhouettes. Plainly, a neighborhood may be an arbitrarily-bounded "bundle" of inchoate and dissimilar attributes, and yet be distinct from every other bundle nearby. Some neighborhoods may even have positive and negative boundaries of nearly equal magnitude. For instance, the boundaries for Cobble Hill are positive when abutting two neighborhoods (Boerum Hill & Gowanus) but not a third (Carroll Gardens). Indeed, an even stronger example of this is in the north-central detail shown in Figure 8 with medians in Table 2. The border area between Bedford-Stuyvesant & Williamsburg is directed towards Williamsburg, but Williamsburg overall is more heterogeneous than Bedford-Stuyvesant according to their aspatial silhouette values. Further, Williamsburg blocks on the Bushwick boundary are about equally split in their demographic similarity to Williamsburg or Bushwick. This is despite the fact that Bushwick is much more demographically cohesive than Williamsburg as a whole, measured by its median aspatial silhouette score.

Discussion

Thus, between the path and boundary silhouettes, these methods introduce spatial structure into the canon of (aspatial) methods common in spatial data science. Formally, each statistic does this using a slightly different spatial structure. Both, however, introduce a formal, direct notion of geographical proximity or distance directly into the computation of social distance used to assess the coherence of a given neighborhood or the goodness of fit for an urban cluster.

The path silhouette, by mixing together attribute similarity and spatial proximity, provides a useful mechanism to measure and assess the joint spatial-social similarity in a dataset. This strategy shows increasing promise at the methodological frontiers of urban data science (Chodrow 2017; Wolf 2019), providing a comprehensive way to introduce an explicit model of geographical similarity into the analysis of urban clusters. The pervasiveness the "cores" identified by path silhouettes to be clustered in both spatial, aspatial, and exogenously-determined boundaries suggests that this joint spatial-social similarity measure is both useful in empirical description and in unsupervised learning.

The boundary silhouette, similarly introduces spatial thinking into a classic data science measure, but does so with a different focus in mind. Instead of specifying an explicit model for joint spatial-social similarity, this measure instead aims to quantify how strongly (and in which direction) does each side of a boundary align? It provides a novel, explicitly spatial method to examine how demographic differences coincide (or fail to) in the areas where regions meet. While this is a post-hoc diagnostic (rather than a boundary *detection* method), it can easily be incorporated into the myriad heuristics that guide cluster design, too.

It is important to note that the directional structure inherent in boundary silhouettes is not simply caused by some neighborhoods being more internally cohesive than others. These boundary silhouettes are not functions of the *absolute* goodness of fit of a given observation, they indicate the relative goodness of fit comparing an observation's home cluster to its local alternatives. The aspatial silhouette also does not take into account the proximity of the next-best-fit choice; again, only 16% of blocks have their next-best-fit neighborhood as their best local alternative neighborhood. Since it is often the case that local urban structure can be quite distinct from global urban patterning (Leckie et al. 2012; Jones et al. 2015; Harris 2017), this distinction between the relative goodness of *local* fit and the global best alternative considered by the classic silhouette is novel and insightful.

Conclusion

Geosilhouttes, both path and boundary variants, are immensely useful in their own right for detecting the latent social-spatial "core" of geographical regions, identifying the strength & direction of spatial boundaries, and for understanding the local socio-geographical structure of cluster fit. There is a large variety of possible refinements available for these methods, as well as possible extensions or applications. Moving forward, a classic statistical perspective could be used to identify the formal distributional properties of silhouette statistics in conditions common in urban data science (Anselin and Rey 1991; Rey, Kang, and Wolf 2018, e.g.). Second, the strongly scale-driven reasoning embedded in the boundary silhouette could be used to generalize the analysis of boundaries between multiple levels, allowing for "local" alternatives at

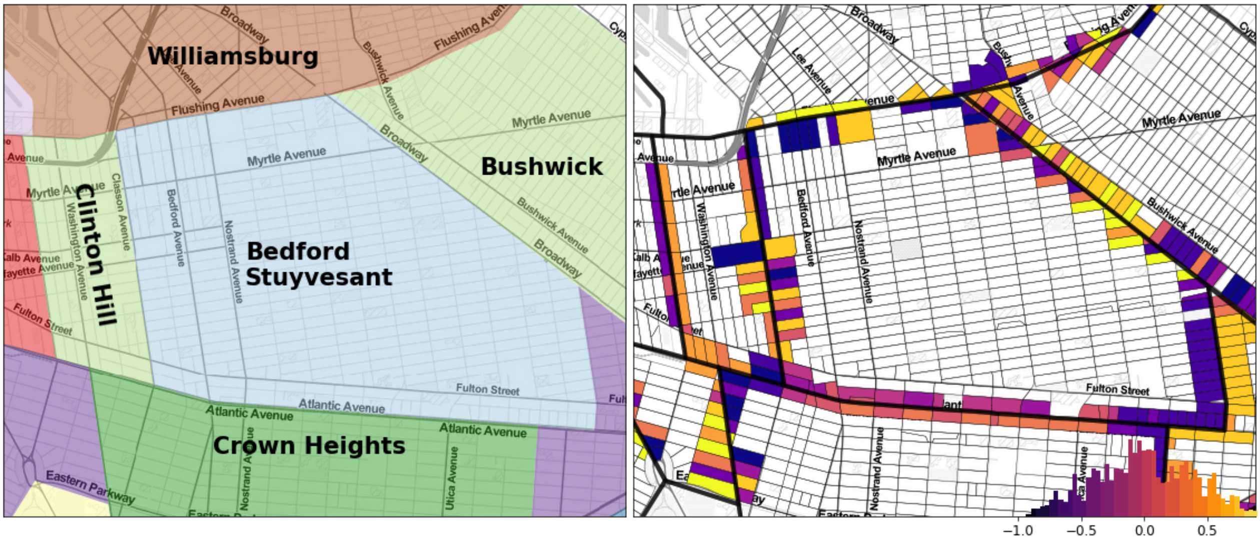

Figure 8. Detail of north-central Zillow neighborhoods in Brooklyn, with boundary silhouettes overlaid. Legends on the bottom-right of each view demonstrate the visible distributions of mapped boundary silhouettes. Basemaps are provided by Stamen Design.

neighbor	Williamsburg	Bushwick	Bedford Stuyvesant	Clinton Hill	Crown Heights
focal					
Williamsburg	0	-0.096	0.693	0.516	-
Bushwick	0.288	0	0.482	-	-
Bedford Stuyvesant	-0.478	0.198	0.000	0.006	-0.059
Clinton Hill	-0.355	-	0.358	0	0.296
Crown Heights	-	-	0.077	-0.427	0

Table 2. Median boundary silhouette values for blocks abutting each cluster in north-central Brooklyn neighborhoods. The rows record blocks in the “focal” cluster that touch the “neighbor” cluster.

a micro (i.e., primitive units such as census blocks/tracks), meso (individual clusters), or macro (citywide) scale (Harris 2017, e.g.). Third, these measures could be extended to spatiotemporal clustering, applying the conceptual logic of the “second-best choice” to alternatives in time and space, or considering the trajectories of demographic classifications using a spatio-temporal distance metric (Delmelle et al. 2013; Delmelle 2016, e.g.). Fourth, a common use case of silhouettes is for graphical heuristics to identify the “optimal” number of clusters in an aspatial context; the path silhouette should provide a similar method for geographical clustering problems, and this should be further studied in future work.

At a more conceptual level, the silhouette provides a useful formal method to introduce spatial thinking because Rousseeuw (1987) is so explicit in the operationalization of the intent of the statistic. Future work should be similarly explicit in intent. However, our choice to use silhouettes as the basic structure onto which geographical thinking can be grafted does not limit the scope of “spatializing” data science methods. Where possible, enhanced methods for spatial data science should be developed in this manner: geographical relationships or structures should be leveraged directly in the statistic or estimator, rather than entering in the *post hoc* analysis of aspatial data science on geographical processes.

In our execution of this research program, we develop two new ways of measuring the local “goodness of fit” for urban clusters. Assessing the local structure of “neighborhoods,” either detected lying latent within a dataset or exogenously determined using government or colloquially-defined boundaries, is a ubiquitous problem in urban data science. For the *path silhouette*, demographic similarity and geographical similarity are combined, providing a single measure of how cohesive neighborhoods are, both spatially and socially. For the *boundary silhouette*, local thinking is introduced into how observations’ are assessed for similarity. This provides an indication of how quickly or dramatically social characteristics change between two adjacent urban clusters, and speaks to the inherently multi-scale structure of urban geography.

Generally speaking, this effort participates in the broader project of developing new methods for urban spatial data science. Sometimes, is not enough to conceptualize fundamentally-geographical problems in aspatial structures; instead, we suggest that introducing spatial thinking directly into the way a statistic *operationalizes its core measurement* is necessary to provide new insights, as we have done. Further, it is through these better concepts and operationalizations that better, more meaningful, and more useful results on the structure of urban society will be obtained.

Funding

We are grateful to the National Science Foundation (NSF # 1733705, *Neighborhoods in Space-Time Contexts* for the generous support.

References

Anselin, L. and S. J. Rey (1991). “Properties of Tests for Spatial Dependence in Linear Regression Models”. In: *Geographical Analysis* 23, pp. 112–131.

Anselin, Luc and Sarah Williams (2016). “Digital Neighborhoods”. In: *Journal of Urbanism: International Research on Placemaking and Urban Sustainability* 9.4, pp. 305–328. ISSN: 1754-9175. DOI: 10.1080/17549175.2015.1080752.

Arribas-Bel, Daniel and Jessie Bakens (2018). “Use and validation of location-based services in urban research: An example with Dutch restaurants”. In: *Urban Studies*, p. 0042098018779554.

Bradley, Jonathan R., Christopher K. Wikle, and Scott H. Holan (2015). “Regionalization of Multiscale Spatial Processes Using a Criterion for Spatial Aggregation Error”. en. In: *arXiv:1502.01974 [stat]*. arXiv: 1502.01974 [stat].

Chaskin, Robert J. and Mark L. Joseph (2013). “Positive’ Gentrification, Social Control and the Right to the City’ in Mixed-Income Communities: Uses and Expectations of Space and Place”. In: *International Journal of Urban and Regional Research* 37.2, pp. 480–502. ISSN: 03091317. DOI: 10.1111/j.1468-2427.2012.01158.x. URL: http://doi.wiley.com/10.1111/j.1468-2427.2012.01158.x.

Chodrow, Philip S (2017). “Structure and Information in Spatial Segregation”. In: *Proceedings of the National Academy of Sciences*, pp. 11591–11596.

Dean, Nema et al. (2018). “Frontiers in Residential Segregation: Understanding Neighbourhood Boundaries and Their Impacts”. en. In: *Tijdschrift voor economische en sociale geografie* in print. ISSN: 1467-9663. DOI: 10.1111/tesg.12316.

Delmelle, Elizabeth et al. (2013). “Trajectories of Multidimensional Neighbourhood Quality of Life Change”. In: *Urban Studies* 50.5, pp. 923–941. DOI: 10.1177/0042098012458003.

Delmelle, Elizabeth C. (2016). “Mapping the DNA of Urban Neighborhoods: Clustering Longitudinal Sequences of Neighborhood Socioeconomic Change”. In: *Annals of the American Association of Geographers* 106.1, pp. 36–56. DOI: 10.1080/00045608.2015.1096188.

Dong, Guanpeng et al. (2019). “Inferring neighbourhood quality with property transaction records by using a locally adaptive spatial multi-level model”. In: *Computers, Environment and Urban Systems* 73, pp. 118–125.

Drukker, Marjan et al. (2003). “Children’s Health-Related Quality of Life, Neighbourhood Socio-Economic Deprivation and Social Capital. A Contextual Analysis”. In: *Social Science & Medicine* 57.5, pp. 825–841.

Duncan, Dustin T et al. (2014). “Examination of How Neighborhood Definition Influences Measurements of Youths’ Access to Tobacco Retailers: A Methodological Note on Spatial Misclassification”. In: *American Journal of Epidemiology* 179.3, pp. 373–381.

Duncan, Greg J, Jeanne Brooks-Gunn, and Pamela Kato Klebanov (1994). “Economic Deprivation and Early Childhood Development”. In: *Child development* 65.2, pp. 296–318.

Duque, Juan C, Luc Anselin, and Sergio J Rey (2012). “The Max-p-Regions Problem”. In: *Journal of Regional Science* 52.3, pp. 397–419.

Duque, Juan C, Richard L Church, and Richard S Middleton (2011). “The P-Regions Problem”. In: *Geogr. Anal.* 43.1, pp. 104–126.

Fitzpatrick, Matthew C. et al. (2010). “Ecological Boundary Detection Using Bayesian Areal Wombling”. In: *Ecology* 91.12, pp. 3448–3455.

Floyd, Robert W. (1962). “Algorithm 97: Shortest Path”. In: *Commun. ACM* 5.6, pp. 345–. ISSN: 0001-0782. DOI: 10.1145/367766.368168.

Fortin, Marie-Josée et al. (1996). “Quantification of the Spatial Co-Occurrences of Ecological Boundaries”. en. In: *Oikos* 77.1, p. 51. ISSN: 00301299. DOI: 10.2307/3545584.

Galster, G. (2001). “On the Nature of Neighbourhood”. In: *Urban studies* 38.12, p. 2111.

Gibbons, Joseph, Atsushi Nara, and Bruce Appleyard (2018). “Exploring the imprint of social media networks on neighborhood community through the lens of gentrification”. In: *Environment and Planning B: Urban Analytics and City Science* 45.3, pp. 470–488.

Harris, Richard (2017). “Measuring the scales of segregation: Looking at the residential separation of White British and other schoolchildren in England using a multilevel index of dissimilarity”. In: *Transactions of the Institute of British Geographers* 42.3, pp. 432–444.

Harris, Richard, Peter Sleight, and Richard Webber (2005). *Geodemographics, GIS and Neighbourhood Targeting*. Vol. 7. John Wiley and Sons.

Hipp, John R and Adam Boessen (2013). “Egohoods as Waves Washing across the City: A New Measure of “Neighborhoods””. In: *Criminology* 51.2, pp. 287–327.

Hipp, John R, Robert W Faris, and Adam Boessen (2012). “Measuring ‘Neighborhood’: Constructing Network Neighborhoods”. In: *Social networks* 34.1, pp. 128–140.

Hwang, Jackelyn (2016). “The Social Construction of a Gentrifying Neighborhood”. In: *Urban Affairs Review* 52.1, pp. 98–128. ISSN: 1078-0874. DOI: 10.1177/1078087415570643. URL: http://journals.sagepub.com/doi/10.1177/1078087415570643.

Isard, W (1956). “Regional Science, the Concept of Region, and Regional Structure”. In: *Pap. Reg. Sci.* 2.1, pp. 13–26.

Jacquez, Geoff M., Andy Kaufmann, and Pierre Goovaerts (2008). “Boundaries, Links and Clusters: A New Paradigm in Spatial Analysis?” en. In: *Environmental and Ecological Statistics* 15.4, pp. 403–419. ISSN: 1352-8505, 1573-3009. DOI: 10.1007/s10651-007-0066-4.

Jacquez, Geoffrey (1995). “The Map Comparison Problem: Tests for the Overlap of Geographic Boundaries”. In: *Statistics in Medicine* 14.21-22, pp. 2343–2361.

Jacquez, Geoffrey, S. Maruca, and M.-J. Fortin (2000). “From Fields to Objects: A Review of Geographic Boundary Analysis”. In: *Journal of Geographica* 2, pp. 221–241.

Jones, Kelvyn et al. (2015). “Ethnic Residential Segregation: A Multilevel, Multigroup, Multiscale Approach Exemplified by London in 2011”. en. In: *Demography* 52.6, pp. 1995–2019. ISSN: 0070-3370, 1533-7790. DOI: 10.1007/s13524-015-0430-1.

Joseph, Mark L., Robert J. Chaskin, and Henry S. Webber (2007). “The Theoretical Basis for Addressing Poverty Through Mixed-Income Development”. In: *Urban Affairs Review* 42.3, pp. 369–409. ISSN: 1078-0874. DOI: 10.1177/1078087406294043. URL: http://journals.sagepub.com/doi/10.1177/1078087406294043http://uar.sagepub.com/cgi/doi/10.1177/1078087406294043.

Leckie, George et al. (2012). “Multilevel modeling of social segregation”. In: *Journal of Educational and Behavioral Statistics* 37.1, pp. 3–30.

Logan, John R. (2013). “The Persistence of Segregation in the 21st Century Metropolis”. In: *City & Community* 12.2, pp. 160–168. ISSN: 15356841. DOI: 10.1111/cico.12021. arXiv: NIHMS150003. URL: http://doi.wiley.com/10.1111/cico.12021.

Lu, Haolan and Bradley P. Carlin (July 2005 2007). “Bayesian Areal Wombling for Geographical Boundary Analysis”. In: *Geographical Analysis* 37.3, pp. 265–285.

McGarigal, Kevin et al. (2002). “FRAGSTATS: Spatial Pattern Analysis Program for Categorical Maps”. In:

Mikelbank, Brian A. (2011). “Neighborhood Déjà Vu: Classification in Metropolitan Cleveland, 1970-2000”. In: *Urban Geography* 32.3, pp. 317–333. DOI: 10.2747/0272-3638.32.3.317.

Morenoff, Jeffrey D, Robert J Sampson, and Stephen W Raudenbush (2001). “Neighborhood Inequality, Collective Efficacy, and the Spatial Dynamics of Urban Violence”. In: *Criminology* 39.3, pp. 517–558.

O’Campo, P. et al. (1997). “Neighborhood Risk Factors for Low Birthweight in Baltimore”. In: *American Journal of Public Health* 87.7, pp. 1113–1119.

Pedregosa, F. et al. (2011). “Scikit-Learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Poorthuis, Ate (2018). “How to Draw a Neighborhood? The Potential of Big Data, Regionalization, and Community Detection for Understanding the Heterogeneous Nature of Urban Neighborhoods”. en. In: *Geographical Analysis* 50.2, pp. 182–203. ISSN: 1538-4632. DOI: 10.1111/gean.12143.

Rey, Sergio J. and Luc Anselin (2007). “PySAL: A Python Library of Spatial Analytical Methods”. In: *The Review of Regional Studies* 37.1, pp. 5–27.

Rey, Sergio J, Wei Kang, and Levi John Wolf (2018). “Regional inequality dynamics, stochastic dominance, and spatial dependence”. In: *Papers in Regional Science*.

Rey, Sergio J. et al. (2011). “Measuring Spatial Dynamics in Metropolitan Areas”. In: *Economic Development Quarterly* 25.1, p. 54.

Roberts, E. (1997). “Neighborhood Social Environments and the Distribution of Low Birthweight in Chicago”. In: *American Journal of Public Health* 87.5, pp. 597–603.

Rousseeuw, Peter J (1987). “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis”. In: *Journal of computational and applied mathematics* 20, pp. 53–65.

Sampson, Robert J, Stephen W Raudenbush, and Felton Earls (1997). “Neighborhoods and Violent Crime: A Multilevel Study of Collective Efficacy”. In: *Science* 277.5328, pp. 918–924.

Santos, Simone M, Dora Chor, and Guilherme Loureiro Werneck (2010). “Demarcation of Local Neighborhoods to Study Relations between Contextual Factors and

Health". In: *International journal of health geographics* 9.1, p. 1.

Shelton, Taylor and Ate Poorthuis (2019). "The Nature of Neighborhoods: using big data to rethink the geographies of Atlanta's Neighborhood Planning Unit system". In: *Annals of the American Association of Geographers* forthcoming. DOI: https://doi.org/10.31235/osf.io/vmrnf.

Singleton, Alexander D and Paul A Longley (2009). "Creating Open Source Geodemographics: Refining a National Classification of Census Output Areas for Applications in Higher Education". In: *Pap. Reg. Sci.* 88.3, pp. 643–666.

Singleton, Alexander D and Seth E Spielman (2014). "The Past, Present, and Future of Geodemographic Research in the United States and United Kingdom". In: *Prof. Geogr.* 66.4, pp. 558–567.

Spielman, Seth E and David C Folch (2015). "Reducing Uncertainty in the American Community Survey through Data-Driven Regionalization". In: *PLoS One* 10.2, e0115626.

Spielman, Seth E and John R Logan (2013). "Using High-Resolution Population Data to Identify Neighborhoods and Establish Their Boundaries". In: *Annals of the Association of American Geographers* 103.1, pp. 67–84.

Spielman, Seth E, Eun-Hye Yoo, and Crystal Linkletter (2013). "Neighborhood Contexts, Health, and Behavior: Understanding the Role of Scale and Residential Sorting". In: *Environment and Planning B: Planning and Design* 40.3, pp. 489–506. DOI: 10.1068/b38007.

Talen, Emily and Julia Koschinsky (2014). "Compact, Walkable, Diverse Neighborhoods:Assessing Effects on Residents". In: *Housing Policy Debate* 24.4, pp. 717–750. ISSN: 1051-1482. DOI: 10.1080/10511482.2014.900102. URL: http://www.tandfonline.com/doi/abs/10.1080/10511482.2014.900102.

Tolsma, J. and T.W.G. van der Meer (2018). "Trust and contact in diverse neighbourhoods: An interplay of four ethnicity effects". In: *Social Science Research* 73.April, pp. 92–106. ISSN: 0049089X. DOI: 10.1016/j.ssresearch.2018.04.003. URL: http://linkinghub.elsevier.com/retrieve/pii/S0049089X17304726.

Wachsmuth, David and Alexander Weisler (2018). "Airbnb and the rent gap: Gentrification through the sharing economy". In: *Environment and Planning A: Economy and Space* 50.6, pp. 1147–1170.

Ward, Joe H. (1963). "Hierarchical Grouping to Optimize an Objective Function". In: *Journal of the American Statistical Association* 58, pp. 236–244. DOI: 10.1080/01621459.1963.10500845.

Wolf, Levi John (2019). "Spatially-Encouraged Spectral Clustering". In: *preprint*.

Womble, W. H. (1951). "Differential Systematics". en. In: *Science* 114.2961, pp. 315–322. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.114.2961.315.