Spatially-Encouraged Spectral Clustering: A Critical Revision of Spatially-Constrained Spectral Clustering

Levi John Wolf*

April 4, 2018

Abstract

Spatially-constrained clustering is a central concern in geographic data science. It finds applications in epidemiology, sociological neighborhood study, criminology, operations research, statistics, and econometrics, to name a few. One novel method developed by Yuan et al. (2015) provides a powerful new technique that ostensibly allows contiguity and attribute similarity to be balanced parametrically. In this paper, I investigate a free parameter omitted in their discussion. I find this free parameter has a significant impact on the solution structure; in addition to the single parameter considered by Yuan et al. (2015), this parameter also affects the balance of contiguity and attribute similarity in identified clusters. This is because what the technique uses are two separate kernels — one for spatial similarity and one for attribute similarity. Both kernels affect solution quality. Exploiting this realization, I create a generalization that leverages this behavior and apply it in two empirical examples with data of differing spatial support. This generalization, called "spatially-encouraged spectral clustering," embeds the fact that spatial and attribute information can be combined in a variety of ways, and parameters must be available for both sets of information in order to use them effectively in the quintessential geographic data science problem of cluster detection.

1 Introduction

Spectral clustering is a thoroughly-used technique in machine learning to analyze the latent spatial structure in data (Ng, Jordan, and Weiss 2002; Von Luxburg 2007). As a clustering technique, it has often been applied to graph-embedded data (White and Smyth 2005), and recently been applied to questions of geodemographic analysis of segregation and sorting (Chodrow 2017). The spectral

*School of Geographical Sciences, University of Bristol, levi.john.wolf@bristol.ac.uk, in preparation for GIS Research UK Conference, 20 April, 2018 at the University of Leicester

analysis of geographic data is not a new concept in quantitative spatial science (Tobler 1966). Indeed, the properties of the spectra of models and methods in spatial econometrics and statistics are well known (Griffith 2000; Griffith 2013). While spectral methods often find their way into supervised learning of geographic processes, the application of spectral analysis to unsupervised learning in geography is less common.

This is because spectral clustering methods have not deeply integrated geographic information into clusterings, like those required for spatially-constrained clustering (Duque, Ramos, and Surinach 2007). The problem of identifying geographically-meaningful clusters is ubiquitous in the field, and has significant common applications in epidemiology (Turnbull et al. 1990; Besag and Newell 1991; Kulldorff and Nagarwalla 1995; Neill et al. 2005; Rogerson and Yamada 2009) and econometrics (Czamanski and Ablas 1979; Rey and Mattheis 2000; Arbia, Espa, and Quah 2008). Further, the exploratory analysis of spatial outliers and "hotspots" is a common technique in exploratory spatial data analysis (Anselin 1995; Getis and Ord 1996), and is a common mode of analysis for initial geographic interrogation. The determination of spatially meaningful communities also plays a large role in spatial sociology (Galster 2001; Drukker et al. 2003; Spielman and Logan 2013), and the visualization of the spatiality inherent in demographic data is a robust subfield of geography, known as geodemographics (Harris, Sleight, and Webber 2005; Harris, Johnston, and Burgess 2007; Singleton and Longley 2009; Singleton and Spielman 2014).

While each of these domains focuses on slightly different methods of identifying geographic areas with consistent semantic meaning (or visualizing the spatiality of semantically-meaningful areas in data), measures of how "well-separated" clusters are from one another are common in the literature. Recent methods, such as local information scoring (Chodrow 2017) follow in a long line of post-hoc measures of attribute separation between identified clusters (Rousseeuw 1987). In this way, most measures of attribute similarity or spatial contiguity are either the byproduct of cluster fitting or are computed about solutions after their fit. But, by themselves, these scores provide no method to balance spatial separation and attribute homogeneity before the solution is generated. In contrast, the *spatially-constrained spectral clustering* method of Yuan et al. (2015) parameterizes the balance between contiguity and cluster cohesion before computing clusters. Critically, the balance parameter can be varied to generate different solutions with differing trade-offs between spatial separation and attribute homogeneity. This means that, instead of simply characterizing the spatial and social separation in a clustering solution after the fact, it can be controlled from the outset.

Yuan et al. (2015) provide their method in the context of land use classification over remotely sensed polygons. To provide an extension of this method for generalized spatial social science, I will review the basic theory of spectral clustering, discuss Yuan et al. (2015)'s core improvement, and then discuss two related representational choices that are required to generalize the method. First, I discuss the impact of an omitted free parameter: the attribute kernel bandwidth, τ^2. Whereas Yuan et al. (2015) only consider a spatial discrete bandwidth parameter, δ, both δ and the omitted τ^2 balance to control the solution's characteristics. This runs counter to the suggested interpretation obtained from the analogy offered to aspatial constrained spectral clustering.

The examination of the free parameter and re-interpretation of δ do not invalidate Yuan et al. (2015)'s conclusions, nor the novelty of their method. Rather, these realizations are required to generalize the method, improving the method's flexibility for a given data and adaptability for various types of data. I call this generalization *spatially-encouraged spectral clustering*. I close by demonstrating this generalization for two datasets with different spatial support and thus different conceptually-appropriate spatial representations.

2 Fundamentals of Spectral Clustering

Spectral clustering works by finding clusters in a lower-dimensional embedding of high-dimensional attribute data (Ng, Jordan, and Weiss 2002). Thus, at a high level, spectral clustering has a dimension reduction step where relevant eigenvectors are extracted from a summary matrix of the data, and then a cluster discovery step, where clusters are detected from within the lower-dimensional embedding. The data summary matrix, called the *affinity matrix*, encodes the pairwise similarities between N observations over P covariates contained in the source data matrix $\mathbf{X}$. The pairwise affinity scores are a distance metric (usually standardized between $[0, 1]$) that relates how similar observation i is to observation j; these are collected into $\mathbf{A}$, the $N \times N$ affinity matrix. Typically, attribute similarity is a kernel function, such as a negative exponential kernel, such that the resulting affinity matrix is positive semidefinite and symmetric. With an affinity matrix, spectral clustering operates on the implied

Laplacian matrix of this data, which is defined as:

$$\mathbf{L} = \mathbf{D} - \mathbf{A} \tag{1}$$

where $\mathbf{D}$ is a diagonal matrix (the *degree*) matrix, such that $\mathbf{D}_{ii} = \sum_j^N \mathbf{A}_{ij}$. Then, the set of K eigenvectors corresponding to the largest K eigenvalues extracted from $\mathbf{L}$ provide a lower-dimensional embedding over which a simpler method, such as K-means, can be used to cluster more efficiently in lower dimension (Von Luxburg 2007).

What Yuan et al. (2015) note is that, given the structure of the implied Laplacian $\mathbf{L}$, the structure of $\mathbf{A}$ can be adjusted to provide a mixture of a contiguity constraint and attribute affinity. Indeed, as long as $\mathbf{A}$ remains positive semidefinite and symmetric (and the corresponding $\mathbf{D}$ adapts to its structure), $\mathbf{A}$ may be arbitrarily adjusted. They focus on the common concern of determining contiguous clusters, with various algorithms designed explicitly for this purpose (Duque, Ramos, and Surinach 2007). To develop a spatially-constrained spectral clustering algorithm, Yuan et al. (2015) adapts constrained spectral clustering (Wang and Davidson 2010). The affinity matrix $\mathbf{A}$ is partitioned into two component matrices, $\mathbf{A}_f$, the attribute affinity matrix and $\mathbf{A}_s$, a spatial affinity matrix. For their problem, Yuan et al. (2015) parameterize the spatial affinity matrix using, δ, which they suggest reflects the extent to which spatial constraints are enforced. Thus, the form for $\mathbf{L}_s$, the spatially-informed Laplacian used for clustering, is similar to the standard aspatial spectral clustering Laplacian:

$$\mathbf{L}_s = \mathbf{D}_s - \mathbf{A}_f \circ \mathbf{A}_s(\delta) \tag{2}$$

where $\mathbf{D}_s$ is again constrained to be equivalent to the diagonalized row sums of $\mathbf{A}_f \circ \mathbf{A}_s$, and $\circ$ denotes the Hadamard (elementwise) product of the two $N \times N$ affinities. Again, by examining the top K eigenvectors of $\mathbf{L}_s$ with K-means clustering, a hybrid spatial-attribute regionalization can be constructed with the balance between spatial contiguity and attribute fit governed by δ.

2.1 The model of constraint

While the approach in Yuan et al. (2015) is novel, there are a few remaining questions about δ. Explicitly, due to the spatial nature of the problem, δ changes subtly in meaning from the typical constrained spectral clustering interpretation. In typical constrained spectral clustering, δ is a convex combination weight used to merge the affinity matrix $\mathbf{A}$ and constraint matrix $\mathbf{C}$ together:

$$(1-\delta)\mathbf{A} + \delta\mathbf{C} \tag{3}$$

where δ is constrained between 0 and 1. Thus, when δ approaches one, $\mathbf{C}$ becomes the sole considered factor, and when δ approaches zero, $\mathbf{A}$ becomes dominant.

In contrast, Yuan et al. (2015)'s more complex specification provides a δ with different behavior. In fact, their δ is a discrete bandwidth parameter, not a combination weight. They suggest a conventional geographic interpretation of contiguity as neighbors along a first-order adjacency graph representation. They call this a *linear spatial kernel*. Going further, they then state a generalization of this kernel that they call the *exponential spatial kernel*. At a given finite order of contiguity, η, the authors provide the exponential kernel as:

$$\mathbf{A}_s(\eta) = \sum_{k}^{\eta} \frac{\mathbf{A_0}^k}{k!} \tag{4}$$

where $\mathbf{A}_0$ is the first-order contiguity adjacency matrix. In fact, they note that this is a "truncated exponential kernel," since the series stops at η.[1] This implies η in this problem is an integer-valued parameter denoting a specific order of contiguity at which observations are considered neighbors. To mimic the $[0,1]$ support from standard constrained spectral clustering, they let $\delta = \eta / \max\{\eta\}$, where $\max\{\eta\}$ is the diameter of the graph, and thus the order of the longest attainable path. Further, Yuan et al. (2015) consider a *binarized truncated exponential kernel*, effectively a k-th order neighborhood, where all nonzero elements of $\mathbf{A}_s(\eta)$ are set to 1, and is zero elsewhere.

3 Interrogating δ

This redefined δ parameter is different from the aspatial δ parameter. A superficial reason indicating their difference is immediately apparent: δ is integral like η (albeit rescaled), and so is not continuous like a standard convex combination weight from Expression 3. More critically, δ is not the sole governing factor controlling the mixture of attribute affinity and spatial constraint as it would be in typical

[1] This may be familiar to those who work with simultaneous autoregressive spatial econometric models, as it is the decay of the matrix-exponential spatial autoregressive specification of (LeSage and Pace 2007) where the decay parameter is fixed to 1.

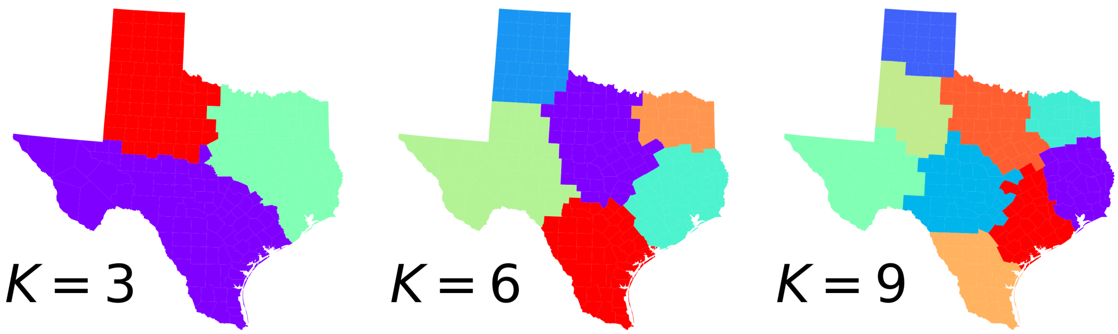

Figure 1: Clusters in the graph spectrum of rook contiguity in Texan counties.

constrained spectral clustering. Changes in the attribute kernel, $\mathbf{A}_f$, may result in dramatically different levels of contiguity in obtained solutions. Sensitivity to attribute bandwidth is a known concern in spectral clustering (Von Luxburg 2007), and the introduction of a second spatial bandwidth potentially makes their interaction even more volatile.

To resolve this, it may make more sense to understand δ only as a spatial bandwidth and not a combination weight. Further, this puts it on par as an equal to the attribute kernel bandwidth, τ^2. In this understanding, many more spatial kernels are available, whose functional form or bandwidths may differ. This attribute-spatial affinity partitioning can also be used more generally for more generic spatial data supports. In this general form, the spatial bandwidth parameter should not be understood as a contiguity relaxation parameter; this property is specific to Yuan et al. (2015)'s kernel and the specific experimental design. In general, δ does solely control contiguity. Recognizing the full generality of δ in the spatial kernel and its relationship to the attribute kernel provides a novel clustering method in its own right, adaptable for various types of clustering problems where *both* the strength of spatial cohesion (only sometimes viewed as contiguity) and attribute cohesion may be parameterized.

3.1 The Two Kernels

Yuan et al. (2015) suggest that δ, their spatial kernel parameter, behaves similarly to the constraint mixture parameter in standard constrained spectral clustering. If this were the case, δ would be the sole parameter governing the strength with which the contiguity constraint holds, as in Expression 3. Under examination, though, the affinity kernel bandwidth also affects solution contiguity.

To identify the free parameter, let us first consider the linear contiguity kernel suggested by Yuan et al. (2015), the first-order adjacency matrix $\mathbf{A}_0$. Without any attribute data, we could obtain the latent spatial clusters from the eigenspace of the connectivity graph for various K, as is shown for Texas counties in Figure 1. We see that, by Rook contiguity, the discovered clusters are indeed contiguous. Further, we observe that this contiguity kernel is also what occurs when $\eta = 0$ in the truncated exponential kernel shown in Equation 4.

The attribute affinity matrix $\mathbf{A}_f$ is also specified using a kernel. Yuan et al. (2015) follow a conventional choice in both spatial analysis and machine learning with a Gaussian kernel. This models the affinity between two observations, $\rho(\mathbf{X}_i, \mathbf{X}_j)$, as:

$$\rho(\mathbf{X}_i, \mathbf{X}_j) = \exp\left\{-\tau^2 ||\mathbf{X}_i - \mathbf{X}_j||^2\right\} \tag{5}$$

where $||.||$ denotes the euclidean distance between the observations' P-length covariate vectors.

While their simulation design focuses on the behavior of δ, Yuan et al. (2015) do not report the value of τ^2 used in the analysis. This is important, since τ^2 is a free parameter, and can be adjusted to provide different affinity structures, depending on the data, just like the spatial parameter. Their input data is the principal components derived from many mean-centered and unit-deviation standardized covariates. These principal components themselves may not necessarily have unit variances, so it should not necessarily be the case that $\tau = 1$ is an clear empirical choice. Indeed, τ is not intended to stand in as the empirical variance of $\mathbf{X}$ generally, as $\mathbf{X}$ may be $N \times P$ with different variances for each covariate, but τ^2 is scalar and used for all P. Despite the fact that Yuan et al. (2015) do not discuss the calibration of this free parameter, I will show that its calibration may significantly affect the spatial structure of obtained solutions, even when δ is fixed. Further, depending on the nature of the attribute affinity kernel, δ can have a dramatically different impact on contiguity.

For this, we need to introduce data. Let us consider a single covariate: change in the two-party vote for president from 2012 to 2016 in Texan counties. This is mapped in Figure 2. In this case, there are a few areas where groups of counties tended to swing together in the same direction. This means their vote share *intensified* together towards one party, even if that party did not win the county. What is clear from the map of swing in Texas counties is that urban areas tended to swing Democrat, meaning the Democrats increased their vote share in urban counties (even if they didn't constitute a majority

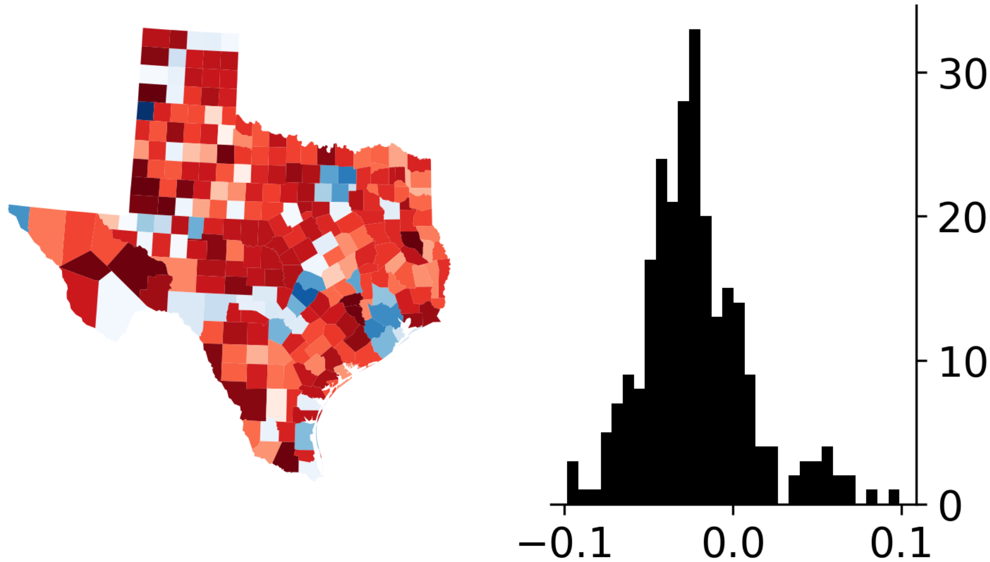

Figure 2: Change in two-party vote share in Texas, 2012 to 2016. On left is the raw distribution in a histogram, on right is the corresponding spatial distribution over counties.

there), whereas Republicans tended to consolidate support in the more rural counties in western and southern Texas which, in fact, they tended to win. This minor instantiation of partisan polarization in the electorate is well-studied in political science and electoral geography, and does not come without controversy itself.

We can visualize this for the real data on partisan swing. The distribution of affinities for the Gaussian kernel at a given τ^2 value is shown in Figure 3. The maps in that figure are produced for $K = 9$, but K is illustrative here, and is independent of the affinities shown below each map; any K could be chosen for a given τ^2 value and a similar result visualized. Further, we should reinforce that no adjustment to the spatial kernel has been made: in Yuan et al. (2015)'s terms, the linear spatial kernel (i.e. standard adjacency matrix, $\mathbf{A}_0$) is used throughout.

The graphic starts with a value of τ^2 that results in nearly all affinities tightly clustered around 1, and then proceeds to values of τ^2 that evenly distribute affinities along $(0, 1)$ to a final value of τ^2 where a plurality of observations have zero attribute affinity. Considering that δ is fixed, change in the underlying affinity distribution does change the identified clusters. Further, this change is larger for some clusters than others. Since δ is fixed, this reflects the increasing use of attribute information. When $\tau^2 = 1$, affinity scores are essentially all 1, and the cluster solution effectively ignores attribute information and recovers the $K = 9$ solution from Figure 1 But, increasing τ^2 while $\delta = 0$ does not break contiguity: all

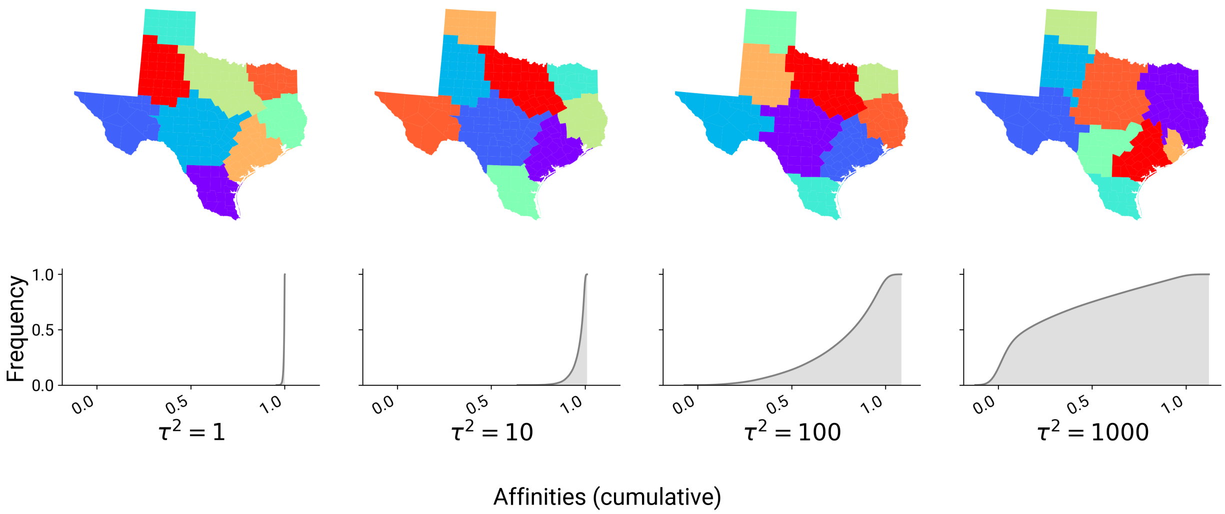

Figure 3: Spectral clusters and affinities for the presidential swing in Texas counties. Moving right, τ^2 increases. In all cases, the "linear spatial kernel," is used, so $\delta = 0$ always. The top row reflects the ultimate solution to the clustering problem for $K = 9$ clusters, and the bottom row reflects the cumulative distribution of attribute affinities contained in $\mathbf{A}_f$ for the specified τ^2.

identified clusters are internally connected.

3.2 The underlying model of clustering

Given that the clustering solution is often sensitive to τ^2, two things seem reasonable. First, it is plausible that some solutions at a fixed δ could be more relaxed than others by changing τ^2 alone. Second, it is plausible that, for two different τ^2 values (τ_a^2, τ_b^2), marginal changes in δ when $\tau^2 = \tau_a^2$ may be more impactful than when $\tau^2 = \tau_b^2$. To investigate this, a solution grid is shown in Figure 4 over values of τ^2 and δ. I focus on the binarized contiguity kernel here; each row uses the $\mathbf{A}_\eta$ connectivity matrix, connecting observations with maximum path order η, since the exponential kernel behaves substantively similarly in this case.

Both reasonable implications occur in this example. For some levels of $\eta > 0$, clusters become spatially disconnected as τ^2 increases but η is fixed. Reading across the rows, increases in τ^2 for fixed η tends to result in decreased contiguity at any order of $\eta \geq 1$. But, declining contiguity is more pronounced when η is large than when it is small. Reading down the columns, increases in η for a fixed τ^2 have nearly no impact on solution contiguity when $\tau^2 = 1$ and all attribute affinities are nearly 1. Thus, neither τ^2 nor η (alternatively, δ) is exclusively in control of the balance between contiguity and

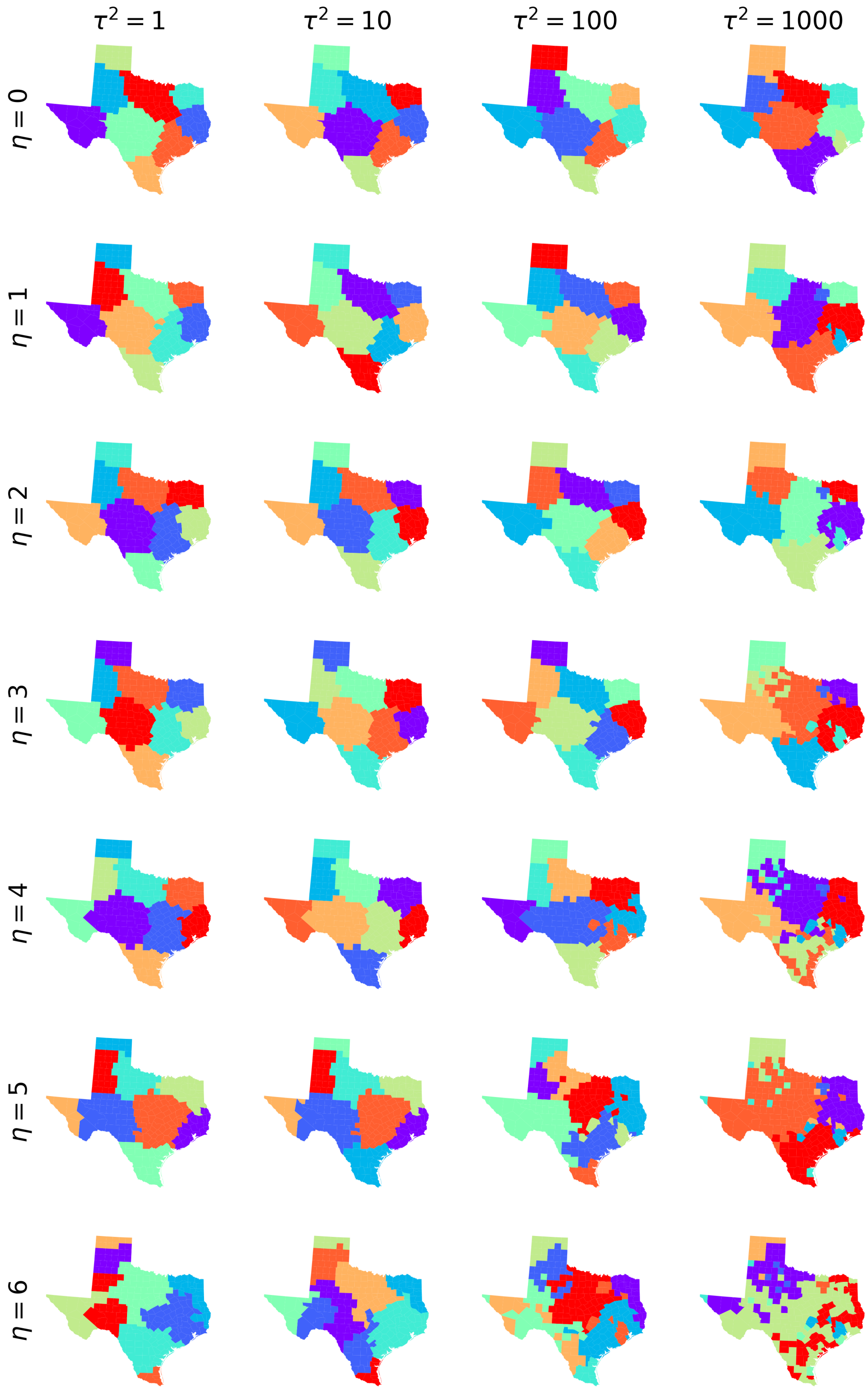

Figure 4: Clusters of counties identified in the presidential swing data, varying δ and τ^2.

attribute cohesion: they must be balanced against one another depending on the data itself.

So, why do Yuan et al. (2015) observe that, as δ increases, contiguity constraints become less important in the problem? To build intuition, note that spectral clustering solutions are an approximate solution to population-balanced minimum cut problems (Von Luxburg 2007). Thus, the use of flatter affinity distributions (larger τ^2) means that the cost of a constraint being violated is more heterogeneous, and depends more strongly on attribute affinity. In contrast, when attribute affinity scores are nearly the same, the cost of violating the spatial constraints is nearly constant and independent of attribute information.

For example, assume that the analyst sets $\tau^2 = 1$, the default in some popular data science computational software packages, such as `scikit-learn` (Pedregosa et al. 2011). For the data considered here, this results in an affinity matrix which is primarily dense, and mostly values near 1. The product of the nearly-uniformly 1 matrix and the sparse first-order spatial contiguity matrix essentially recovers the contiguity matrix itself:

$$\lim_{\mathbf{A}_f \to \mathbf{1}} \mathbf{A}_0 \circ \mathbf{A}_f = \mathbf{A}_0 \tag{6}$$

Thus, if the affinity kernel is too narrow, the spectral clustering solution will essentially ignore attribute data as seen above. As we increase the spatial bandwidth, η, $\mathbf{A}_s$ becomes significantly more dense. This means more and more of the attribute affinity matrix, $\mathbf{A}_f$, is recovered. When δ goes to one (or η goes to the graph diameter), contiguity constraints are "relaxed," but this relaxation may not be relevant to the actual obtained solutions if the attribute kernel does not enforce meaningful differences in observation affinity. Only when τ^2 begins to embody meaningfully-distinct variation over the attributes does the change in δ become relevant for the contiguity of clusters. Thus, both the form and the width of attribute and spatial kernels are necessary to fully-specify a spatially-constrained spectral clustering. Since both are relevant, both bandwidth parameters must be considered.

4 Generalizing the problem

Considering both kernels together, Yuan et al. (2015)'s work can be extended to much more general spatial clustering problems over multivariate spatial data. Indeed, for a given spatial structure matrix $\mathbf{A}_s$, the degree to which space influences the eventual clustering is governed by the distribution of values within the resulting matrix product, $\mathbf{A}_f \circ \mathbf{A}_s$. Therefore, let $\mathbf{A}_s$ now reflect *any* symmetric spatial structure matrix with nonnegative eigenvalues. Let the free parameters used in defining the spatial affinity matrix be η. Further, let the attribute kernel function, $\rho(\mathbf{X}_i, \mathbf{X}_j)$, take an arbitrary vector of free parameters, θ. Together, the same structure can be used to enable *spatially-encouraged spectral clustering*:

$$\mathbf{L} = \mathbf{D} - \mathbf{A}_f(\theta) \circ \mathbf{A}_s(\eta) \tag{7}$$

where $\mathbf{D}$ is again defined as the appropriate degree matrix. Computing a lower-dimensional clustering (like K-means) on the top-K eigenvectors of $\mathbf{L}$ provides an analogous solution to that in Yuan et al. (2015). Further, the relative importance of attributes versus contiguity is governed directly by the normal parameterization of $\mathbf{A}_f(\theta)$ in the attribute affinity matrix and $\mathbf{A}_s(\eta)$ in the spatial affinity matrix. With this, it becomes clear that Yuan et al. (2015)'s discrete contiguity kernel is one possible specification of $\mathbf{A}_s(\eta)$ applicable for clustering on spatial lattice data.

In this generalization, θ and η are stated explicitly to show that they both matter for the eventual clustering solution. As in the previous demonstration, if $\rho(\mathbf{X}_i, \mathbf{X}_j) \approx c \;\forall\; i, j$ for a constant c, geography dominates, and the obtained solutions are no different from those obtained when no data is used at all.[2] As $\mathbf{A}_f(\theta)$ widens in distribution, the clustering solutions become more distinct from the spatial-only clustering. If η changes (or the functional form of $\mathbf{A}_s(\eta)$ changes), then the spatiality of the problem will change as well.

4.1 Understanding the Parameter Tradeoffs

The practical concern for choosing θ and η, then, is the structure of the affinity distribution; if the gap between the minimum affinity and zero is large, contiguity will hold more strongly than if the gap is small, as seen in Figure 3. Both free parameters are required, however, and both simultaneously determine the resulting structure of the solution. An example of this optimization space is shown in Figure 5 for the Texas counties example. There, four maps at varying τ^2 values are shown. Underneath, three scores are shown. First, the variance ratio (Calinski-Harabasz pseudo-F (Calinski and Harabasz 1974)) score is shown, which relates the within-cluster sums of squares to the between cluster sums of squares. Second, the silhouette score is shown (Rousseeuw 1987), which relates the distances

[2] If $\mathbf{A}_s$ encodes nonlocal geographic relationships, the resulting spectral clusters will be defined with respect to that notion of geographic structure.

between observations, their source clusters, and their next-best fit clusters. In both cases, larger values are interpreted as showing clusters are more distinct or coherent, indicating that the variability within clusters is smaller than the variability between clusters. Third, the fraction of observations that are on the boundary of a cluster, the ones that touch an observation with a different label than their own, are plotted. This indicates the extent to which the map is spatially fragmented. Finally, the median attribtue affinity score is plotted on bottom.

As τ^2 increases from zero (where all observations have an attribute-affinity of 1) to being large relative to the distance metric used in the kernel (here, around 1000), clusters occasionally become more spatially fragmented according to their boundary fractions, but have larger variance ratio scores on average. In general, the change in the boundary fraction is not uniform over τ^2, however. Further, most statistics appear to be largely driven by step changes; at a critical τ^2 value, the clustering solution changes radically. The silhouette and boundary fractions, on the other hand, do not change monotonically. In fact, they are unstable over some ranges of τ^2, rapidly changing between two similar values.

In addition, the relationship between attribute coherence and spatial fragmentation is not linear. For some of the increasingly-coherent clustering solutions, the boundary fraction is smaller, such as around the step change near $\tau^2 = 1300$. Thus, small changes in τ^2 may result in significant changes in solution, and the tradeoff between attribute homogeneity and spatial regularity is neither linear nor even zero-sum: the step near $\tau^2 = 1500$ results in less spatial fragmentation and more attribute coherence, but the step change around $\tau^2 = 800$ has has effectively no impact on spatial regularity while improving the variance fraction.

If one were to use this strategy in a formal optimization routine for a combined spatial-attribute objective, solving the clustering optimally would be difficult due to the volatility in any chosen cluster scoring function. Thus, any use of this heuristic should explore τ^2 values in the neighborhood of the solution to determine the stability of the resulting solution. Since the value of τ^2 is data-dependent, it appears sufficient to start a search with τ^2 that place mass over the whole $[0, 1]$ range. Thus, a computationally-cheap exploratory method to examine the balance between contiguity and attribute affinity is to pick τ^2 that provides a reasonably low median affinity, after filtering by $\mathbf{A}_s$. Determining this initial tuning requires no evaluations of the actual eigenvectors of the affinity matrix since no clustering is computed to initialize. In fact, it can be simplified to only compute $\mathbf{A}_f$ for neighbors in a sparse $\mathbf{A}_s(\eta)$.

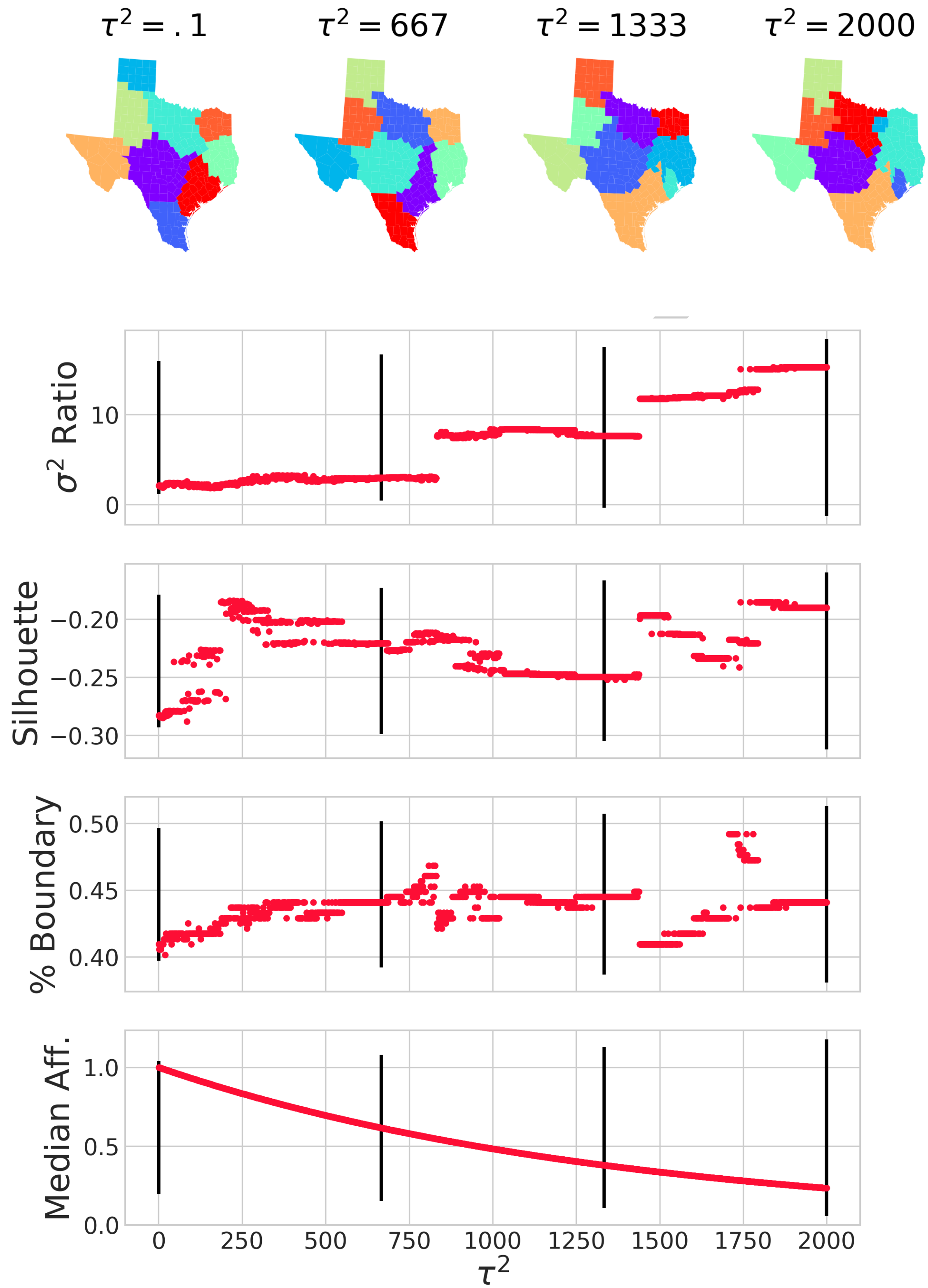

Figure 5: Clusters in electoral swing in Texas Counties as a function of τ^2, the Gaussian kernel parameter.

4.2 Induced geographic regularity for clustering on non-lattice data

Provided with this generalization over arbitrary attribute and spatial kernels, any spatial kernel function can be used. A more arbitrary form for the spatial kernel provides an arbitrary method to mix spatial proximity and attribute affinity together in a single clustering problem. Thus, I move to find price clusters in the prices of Airbnb listings in Brooklyn. These are point-referenced data, so the contiguity forms considered by Yuan et al. (2015) cannot be used. Airbnb, a popular on-demand accommodation service, offers an alternative to typical hotel accommodation. Notably, its spatial market penetration is much more widespread in Brooklyn than the equivalent presence of hostels and hotels. Thus, given enough listings, clusters in those listings' prices may identify spatially-meaningful communities where prices are similar. This is in contrast to other types of density-interpolating cluster searches useful for spatial data (like HDBScan (McInnes, Healy, and Astels 2017)) which simply identify clusters of spatial colocation and other more traditional attribute clustering methods that do not consider space.

Airbnb listings often contain a wealth of information about the potential accommodation, including various amenities made available to the renter and how often the property attracts reviews. Rental data exists as point-referenced data, with each listing having one (and only one) spatial coordinate. Here, only the price information for Airbnbs in Brooklyn scraped on October 21st, 2017 will be used. Some listings overlap, since rentals can cover single rooms or apartments within the same building. No listings are duplicated, meaning that previous price information of the same listing is excluded from the analysis. Thus, this constitutes a single, cross-sectional analysis of recent Airbnb pricing structures in the fall of 2017 in Brooklyn, NY. An example of the spatial pattern of the listings, reflecting their relative density, position, and price, is shown in Figure 6.

Identifying price clusters require that we balance price homogeneity and spatial coherence, since price clusters should ostensibly present a contiguous "conceptual area" where listers price their properties to compete with one another. By examining the spatial clumping of price information, we aim to identify competition communities, where individuals tend to price their Airbnbs to compete with other, spatially-proximate Airbnbs. Practically speaking, this means we must balance spatial coherence and attribute homogeneity, so spatial-spectral clustering is required. However, the contiguity kernel is not applicable here, since contiguity is not defined for point-referenced data.

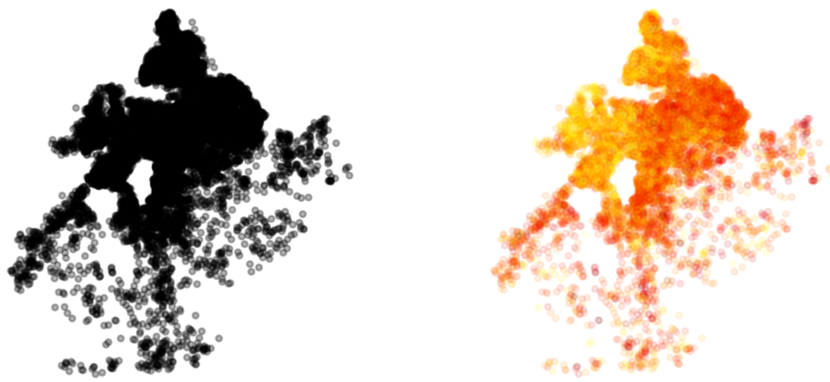

Figure 6: Prices of Airbnb rentals in Brooklyn, NY. Presence/absence is shown on the left, and their prices are shown on the right

Instead, I suggest using an adaptive Gaussian weighting for this:

$$A_s(\delta_i)_{i,j} = \exp\left\{-\delta_i^2 ||s_i - s_j||\right\} \tag{8}$$

where s_i is the spatial position of site i, and $||s_i - s_j||$ is the spatial distance between two sites. As shown in Figure 6, the spatial density of Airbnb listings decreases dramatically as one leaves the north-central Brooklyn core. Thus, using a fixed bandwidth over the entire map would not properly represent areas where listings are more spatially diffuse. An adaptive Gaussian kernel accounts for this, letting the bandwidth for each site i be determined as a function of the distance from that point and its furthest neighbor, for the p-nearest points to i. Thus, in the analysis below, I'll examine the clustering solutions as a function of the size of this nearest neighbor set.

Because the adaptive kernel weight is not symmetric, it cannot be used alone to generate the combined spatial-attribute affinity matrix. To enforce symmetry, the average of the adaptive kernel and its transpose can be used:

$$A_s = \frac{A_s(\delta_i) + A_s(\delta_i)^T}{2} \tag{9}$$

This means that the weight assigned for any dyad of points is an average of their directed weights. This still adapts the bandwidth for sparser areas, but ensures symmetry in the kernel required for spectral analysis. Further, this admits a similar type of binarized analogue: the symmetrized K-nearest neighbor weights calibrated with δ neighbors. In this case, observations are connected when at most one is a member of the other's δ-nearest neighbor clique. Practically speaking, I use δ such that the resulting spatial graph is connected in either adaptive Gaussian or binary KNN graphs to prevent clusters that are only due to spatial sparsity from resulting in the clustering analysis.

Cluster solutions varying the attribute kernel parameter and number of nearest neighbors used to construct the spatial kernel are shown in Figure 7 for the binary symmetric KNN kernel, and in Figure 8 for the adaptive Gaussian kernel. The nearest neighbors used to construct the spatial kernel changes over columns, and the attribute bandwidth changes over rows. Moving down within a column, the attribute kernel becomes sparser, meaning affinities shift from being clustered tightly around 1 to being spread over $[0, 1]$. Moving right over rows, the spatial bandwidth increases, meaning more distant observations are considered connected. In this case, increasing δ actually is associated with *more regular* spatial clusters, for either spatial kernel type. This is the reverse of the contiguity kernel case, where increasing the width of the spatial kernel causes nonlocal clustering solutions to be selected. Increasing the sparsity of the attribute kernel while holding the spatial kernel constant results in the identification of smaller clusters (when the spatial kernel is very small), and significantly affects the cluster assignments when the spatial kernel is large. Further, incredibly wide attribute kernels (when used alongside incredibly narrow spatial kernels) result in extreme imbalances in cluster size, so that some clusters when $\tau^2 = 10$ and $\eta = 6$ are effectively composed of a listings within a single building. While the separating lines are not necessarily convex (nor, indeed, even straight boundaries at a reasonable resolution), the boundaries between clusters are, to a large degree, abrupt. No boundaries are detected where stippling of different label assignment occurs, so these clusters constitute intelligible price clusters in their own right.

We can also see the impact of this varying of the spatial and attribute tuning parameters in this problem. By examining the surface of cluster quality scores, we can see how the change in spatial and attribute parameters trades off in solution quality. I show the surfaces of two common statistics, the map average silhouette score (Rousseeuw 1987) and a variance ratio “pseudo-F” statistic from (Calinski and Harabasz 1974). The silhouette provides a measure of the average separation between clusters in terms of the distance to the centers of other clusters. It varies between -1 and 1, with zero indicating that observations are about as far apart from their own observations as they are from their next best fit cluster on average. The pseudo-F relates the variability between clusters to the variability

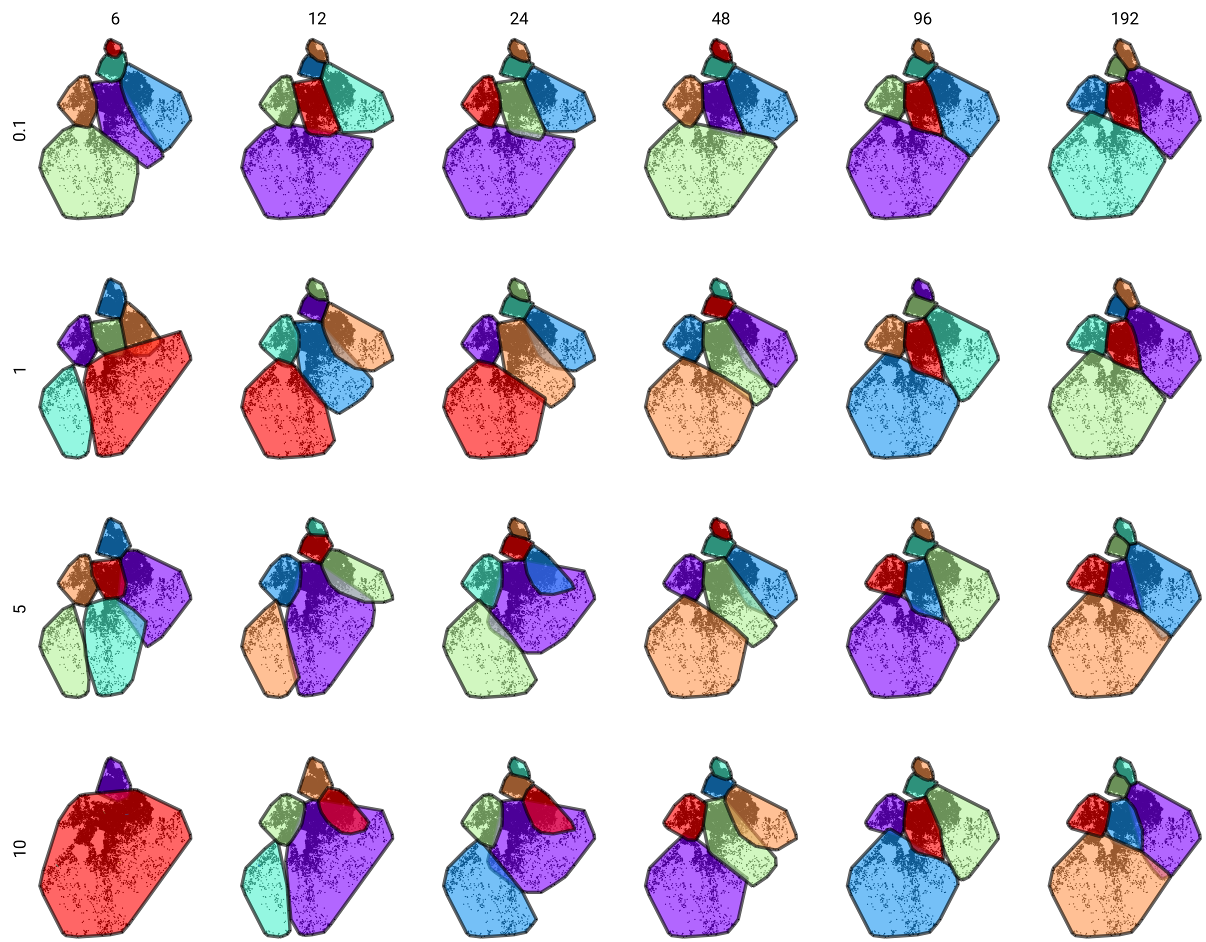

Figure 7: Clusters in Airbnb prices varying as the spatial kernel increases from six to 192 across the rows, and as the attribute kernel widens from $.01$ to 10 down the columns. Critically, the shape and size of both kernels affects the identified solutions. Here, the symmetric binary K-nearest neighbors kernel is used.

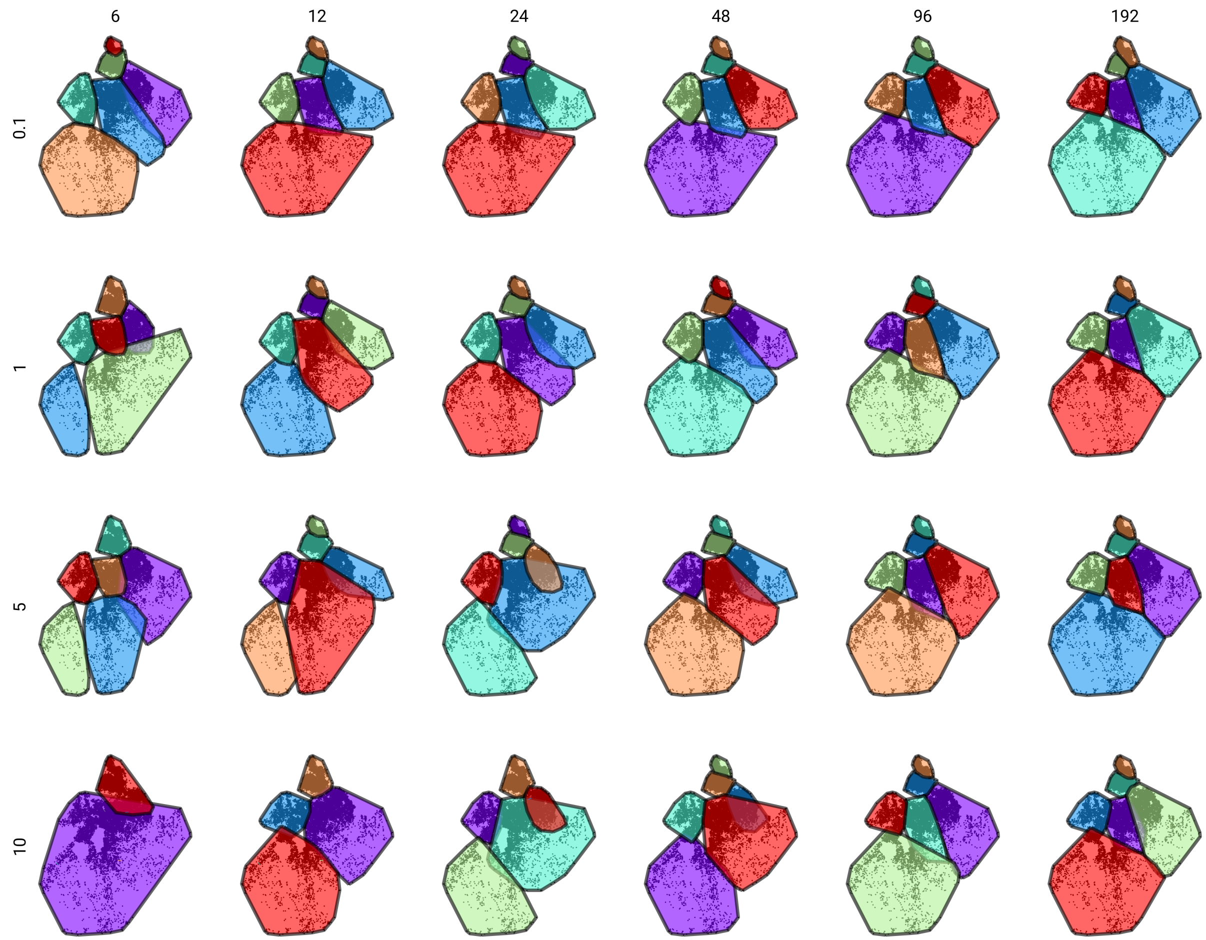

Figure 8: Clusters in Airbnb prices varying as the spatial kernel increases from six to 192 across the rows, and as the attribute kernel widens from $.01$ to 10 down the columns. Critically, the shape and size of both kernels affects the identified solutions. Here, the symmetric adaptive Gaussian kernel is used.

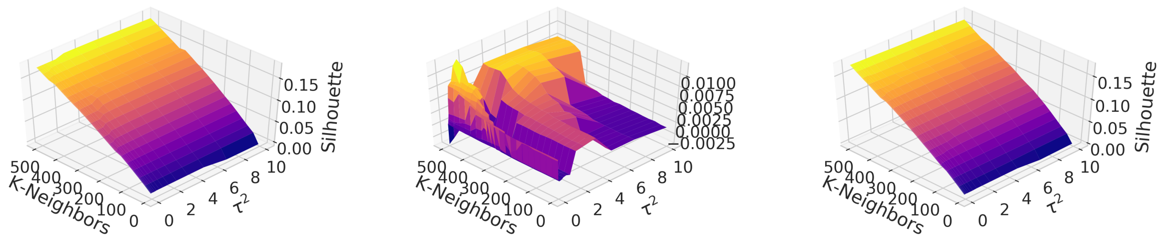

Figure 9: The surfaces of the map average silhouette score, varying the spatial and attribute kernel width. On left is the results from the adaptive Gaussian kernel $\mathbf{A}_s$. On right, the results when using symmetric KNN for $\mathbf{A}_s$. In the middle is the difference between the two curves, Gaussian kernel surface minus the KNN surface.

within clusters, with larger values indicating that variability in cluster means is larger than variability within clusters alone. Thus, in both cases, larger values indicate greater cluster separation.

In Figure 9, the relationship between the silhouette score, spatial, and attribute kernel width are plotted. On left is the kernel surface, on the right is the KNN surface, and in the middle is the difference, kernel minus the KNN surface. The attribute kernel parameter, τ^2, is measured at 20 control points distributed logarithmically between 0 and 10.[3] The spatial parameter, in this case the δ-neighbors chosen to construct the adaptive kernel (or binary adaptive kernel) are shown on another axis, and have been distributed linearly over $[6, 500]$, where 6 is the smallest required number of neighbors to keep the graph connected. What is apparent in this comparison is that the map average solution surface for this objective is relatively insensitive to the attribute kernel, but cluster homogeneity tends to improve as the spatial bandwidth increases. When the two surfaces are viewed end-on-end (collapsing the K and τ^2 dimensions), the two surfaces are not exactly coplanar, but are quite close. In general, the Gaussian adaptive kernel has slightly map average silhouette scores, indicating it provides clusters with slightly better separation. Regardless, since the attribute kernel behaves differently here from the previous section, this strongly reinforces the fact that the sensitivity of this spatial-sectral clustering to the attribute kernel parameterization may vary from dataset to dataset, and from spatial support to spatial support.

Further, the relationship between spatial and attribute kernels may also vary from objective to objective. Figure 10 is a plot of the surfaces for the pseudo-F statistic over varying spatial and attribute kernel widths. Here again, the left plot shows the surface of the pseudo-F with the Gaussian kernel, varying

[3] Plotting τ^2 on a log-scale would then result in these surface plots having a regular grid. They remain plotted on the linear scale for τ^2 to keep the intuition in terms of raw parameter values.

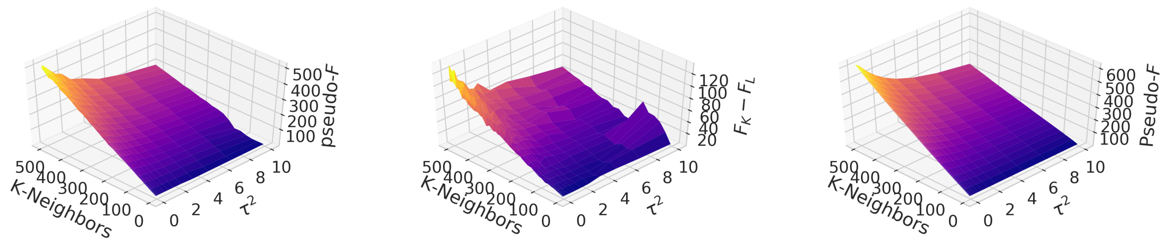

Figure 10: The surfaces of the Calinski-Harabasz pseudo-F, varying the spatial and attribute kernel width. On left is the results from the adaptive Gaussian kernel $\mathbf{A}_s$. On right, the results when using symmetric KNN for $\mathbf{A}_s$. In the middle is the difference between the two curves, KNN surface minus the Gaussian kernel surface. This is the reverse difference from Figure 9 for legibility reasons.

the spatial and attribute kernel widths, and the right plot shows the KNN binary kernel. However, here, the difference term has been flipped for orientation visibility, with the F_K standing for the F statistic for the KNN binary kernel and F_L for the Gaussian kernel. In this case, the plot demonstrates that, when the attribute kernel is narrow, widening the spatial kernel by increasing the number of neighbors tends to increase cluster homogeneity much more rapidly than when the attribute kernel is wide. Further, the KNN model in this case tends to have uniformly better variance ratios, and the difference between the two surfaces is much less noisy than that in the silhouette curve. Regardless, as shown by these two surfaces, the impact of changing the spatial kernel width on a given score is not consistent between scores, nor is it consistent for different attribute kernel widths. So, attention must be paid to both; one cannot ignore the specification of one kernel while changing the other.

5 Discussion & Conclusion

In total, spectral clustering methods hold significant promise for spatially-constrained regionalization and exploratory spatial cluster analysis. Already recognized as a computationally hard problem (Duque, Ramos, and Surinach 2007), spectral methods reduce the implementation difficulty in discovering approximately (or exactly) contiguous clusters. However, these methods are not cost-free, since the determination of eigenvectors for large sparse matrices is compute-intensive and sometimes inaccurate. Thus, efficient and accurate eigenvector computation methods constrain the broader applicability of this technique. But, this method can exploit the same computational improvements made generally for machine learning implementations of spectral methods, as this technique reduces fundamentally to spectralbibtex clustering on well-designed spatial representations of the data.

This method can be used for exploratory data regionalization, as well as data-dependent regionalization. The choice of the functional form for $\mathbf{A}_s$ and $\mathbf{A}_f$ is arbitrary, as is the choice of values for their relevant free parameters. Ideally, these should be data-appropriate and theory-driven, reflecting the unique structure of data and the desired outcomes of analysis (e.g. Chodrow 2017). An optimal fit routine could be used, where θ is chosen to maximize a convex combination of generalized contiguity (Wu and Murray 2008) and within-class homogeneity. If this were done, the resulting convex combination parameter *would* serve as a direct analogue to δ in traditional constrained spectral clustering. However, the volatility in three types of clustering score functions considered suggest that this objective may be exceptionally-poorly behaved. Alternatively, kernels may be set to search for solutions within a parameter grid in an exploratory analysis, allowing clusters to become more or less diffuse over space or homogeneous in attributes. This method was demonstrated and its impact on common cluster scoring metrics shown to be non-linear for some scores and some combinations of spatial kernels and attribute kernel widths.

The definition of spatial kernels in spatially-encouraged spectral clustering can be adapted to model the spatial relationships of any type of data support, and attribute kernels should reflect the attribute distances relevant for the data at hand. The spatial affinity structure might be highly local first-order contiguity, it may reflect neighborliness within a given distance cutoff, or it might itself be a continuous Matérn covariance function or spatial kernel. The attribute affinity structure may be a straightforward attribute kernel (e.g. Gaussian, bisquare, or triangular kernels) or may reflect various compositional affinities using more complicated divergence measures (Chodrow 2017). Regardless, when combined, typical spectral clustering methods can be applied to and clusters of varying spatial cohesion may be obtained.

This paper identifies an opportunity to extend and expand (Yuan et al. 2015) and takes it. By examining an omitted parameter in their discussion, I suggest a technique for *spatially-encouraged spectral clustering* that is significantly more general. Further, this more general method offers the end user more control over the structure of the solution, since both spatial and attribute kernels are able to be tuned. While the surfacing of more tuning parameters introduces significantly more complexity into the analysis, it also ensures that the resulting solutions can be fine-tuned and made precisely applicable for the data at hand. Further, ignoring free parameters may result in unexpected behavior, such as changes in the structure of a spatial kernel having different impacts depending on the structure of the attribute kernel. Thus, the technique developed here makes it explicit that both spatial and attribute kernel parameterizations are important, and being aware of their impact can help to tune cluster solution. In sum, this paper's refinement of Yuan et al. (2015) provides a novel, more generally-applicable technique for cluster analysis in geographic data science that can be used to model how space and attribute compete for relevance in the determination of spatial clusters.

References

Anselin, Luc. 1995. “Local indicators of spatial association-LISA”. *Geographical Analysis* 27 (2): 93–115.

Arbia, Giuseppe, Giuseppe Espa, and Danny Quah. 2008. “A class of spatial econometric methods in the empirical analysis of clusters of firms in the space”. *Empirical Economics* 34:81–103.

Besag, J., and J. Newell. 1991. “The dection of clusters in rare diseases”. *Journal of the Royal Statistical Society A* 154 (1): 143–155.

Calinski, Tadeusz, and Jerzy Harabasz. 1974. “A dendrite method for cluster analysis”. *Communications in Statistics-theory and Methods* 3 (1): 1–27.

Chodrow, Philip S. 2017. “Structure and information in spatial segregation”. *Proceedings of the National Academy of Sciences*: 11591–11596.

Czamanski, S, and L Ablas. 1979. “Indentification of industrial clusters and complexes: a comparison of methods and findings”. *Urban Stud.* 16:61–80.

Drukker, Marjan, et al. 2003. “Children’s health-related quality of life, neighbourhood socio-economic deprivation and social capital. A contextual analysis”. *Social Science & Medicine* 57, no. 5 (): 825–841.

Duque, J.C., R. Ramos, and J. Surinach. 2007. “Supervised Regionalization Methods: A Survey”. *International Regional Science Review* 30 (3): 195.

Galster, G. 2001. “On the nature of neighbourhood”. *Urban studies* 38 (12): 2111.

Getis, A., and J. K. Ord. 1996. “Local spatial statistics: an overview”. *Spatial Analysis: Modelling in a GIS Environment* 374.

Griffith, D. A. 2000. “Eigenfunction properties and approximations of selected incidence matrices employed in spatial anlaysis”. *LInear Algebra and its Applications* 321:95–112.

Griffith, Daniel A. 2013. *Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization*. Springer Science & Business Media.

Harris, Richard, Ron Johnston, and Simon Burgess. 2007. "Neighborhoods, Ethnicity and School Choice: Developing a Statistical Framework for Geodemographic Analysis". *Population Research and Policy Review* 26 (5): 553–579. ISSN: 1573-7829. doi:10.1007/s11113-007-9042-9. http://dx.doi.org/10.1007/s11113-007-9042-9.

Harris, Richard, Peter Sleight, and Richard Webber. 2005. *Geodemographics, GIS and neighbourhood targeting*. Vol. 7. John Wiley / Sons.

Kulldorff, M., and N. Nagarwalla. 1995. "Spatial disease clusters: detection and inferenc". *Statistics in Medicine* 14:799–810.

LeSage, James P, and R Kelley Pace. 2007. "A matrix exponential spatial specification". *Journal of Econometrics* 140 (1): 190–214.

McInnes, Leland, John Healy, and Steve Astels. 2017. "hdbscan: Hierarchical density based clustering". *The Journal of Open Source Software* 2 (11): 205.

Neill, D.B., et al. 2005. "Detecting significant multidimensional spatial clusters". *Advances in Neural Information Processing Systems* 17:969–976.

Ng, Andrew Y, Michael I Jordan, and Yair Weiss. 2002. "On spectral clustering: Analysis and an algorithm". In *Advances in neural information processing systems*, 849–856.

Pedregosa, F., et al. 2011. "Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research* 12:2825–2830.

Rey, Sergio J., and Daniel J. Mattheis. 2000. *Identifying Regional Industrial Clusters in California: Volume I Conceptual Design*. Technical Report. California Employment Development Department. Sacramento.

Rogerson, P, and I Yamada. 2009. *Statistical detection and surveillance of geographic clusters*. Chapman & Hall/CRC.

Rousseeuw, Peter J. 1987. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". *Journal of computational and applied mathematics* 20:53–65.

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". *Pap. Reg. Sci.* 88, no. 3 (): 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”. *Prof. Geogr.* 66 (4): 558–567.

Spielman, Seth E, and John R Logan. 2013. “Using high-resolution population data to identify neighborhoods and establish their boundaries”. *Annals of the Association of American Geographers* 103 (1): 67–84.

Tobler, W R. 1966. “Spectral analysis of spatial series”. In *Proceedings of the Fourth Annual Conference on Urban Planning Information Systems and Programs*, 179–186. University of California at Berkeley.

Turnbull, B. W., et al. 1990. “Monitoring clusters of disease: application to Leukemia incidence in upstate New York”. *American Journal of Epidemiology* 132 (1): 136–143.

Von Luxburg, Ulrike. 2007. “A tutorial on spectral clustering”. *Statistics and computing* 17 (4): 395–416.

Wang, Xiang, and Ian Davidson. 2010. “Flexible constrained spectral clustering”. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 563–572. ACM.

White, Scott, and Padhraic Smyth. 2005. “A spectral clustering approach to finding communities in graphs”. In *Proceedings of the 2005 SIAM international conference on data mining*, 274–285. SIAM.

Wu, X., and A. T. Murray. 2008. “A new approach to quantifying spatial contiguity using graph theory and spatial interaction”. *International Journal of Geographical Information Science* 22, no. 4 (): 387–407. ISSN: 1365-8816, 1362-3087, visited on 09/04/2013. doi:10.1080/13658810701405615. http://www.tandfonline.com/doi/abs/10.1080/13658810701405615.

Yuan, Shuai, et al. 2015. “Constrained spectral clustering for regionalization: Exploring the trade-off between spatial contiguity and landscape homogeneity”. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, 1–10. IEEE.