

# Learning Geographical Manifolds: A Kernel Trick for Geographical Machine Learning\*

Levi John Wolf<sup>1,2</sup> and Elijah Knaap<sup>2</sup>

<sup>1</sup>School of Geographical Sciences, University of Bristol, levi.john.wolf@bristol.ac.uk

<sup>2</sup>Center for Geospatial Sciences, University of California, Riverside

May 15, 2019

## Abstract

Dimension reduction is one of the oldest concerns in geographical analysis. Despite significant, longstanding attention in geographical problems, recent advances in non-linear techniques for dimension reduction, called *manifold learning*, have not been adopted in classic data-intensive geographical problems. More generally, machine learning methods for geographical problems often focus more on applying standard machine learning algorithms to geographic data, rather than applying true “spatially-correlated learning,” in the words of Kohonen. As such, we suggest a general way to incentivize *geographical learning* in machine learning algorithms, and link it to many past methods that introduced geography into statistical techniques. We develop a specific instance of this by specifying two geographical variants of Isomap, a non-linear dimension reduction, or “manifold learning,” technique. We also provide a method for assessing what is added by incorporating geography and estimate the manifold’s intrinsic geographic scale. To illustrate the concepts and provide interpretable results, we conducting a dimension reduction on geographical and high-dimensional structure of social and economic data on Brooklyn, New York. Overall, this paper’s main endeavor—defining and explaining a way to “geographize” many machine learning methods—yields interesting and novel results for manifold learning the estimation of intrinsic geographical scale in unsupervised learning.

## 1 Introduction: Manifold Learning in Geography

In the current era of Big Data, cloud computing, and adversarial artificial intelligence, quantitative research in nearly every field is turning to computational methods and machine learning techniques to help model and interpret information in its respective domain. This is equally true of geography, where machine learning models are applied ever more to multivariate problem domains contextualized by geographic space. Neighborhood analysis is a subfield of geography where this trend has been particularly visible (Arribas-Bel 2014; O’Brien et al. 2015; Delmelle 2016; Knaap 2017; Poorthuis and Zook

---

\*This material is based on work supported by the National Science Foundation under Grant #1762160

2017; Knaap et al. 2019). Apart from these new methods, however, neighborhoods have both a rich theoretical tradition and a long history of quantitative analysis, much of which exists outside the realm of spatial effects or other geographic considerations (Shevky and Bell 1955; Raudenbush and Sampson 1999; Sampson et al. 2002; Sampson 2012). For decades, much of the neighborhood research remained cordoned within disciplinary silos while geographers, sociologists, and computer scientists each applied the tools of their own trade to distill insights about neighborhoods and the people who reside within them. Today there is considerably greater exchange among disciplines than in the past, but the exchange is asymmetrical; sociologists and geographers are increasingly adopting the tools of computer science (and/or *data science*, depending on the perspective of the reader) and applying them to spatial problem domains without regard for the special properties inherent in spatial data (Anselin 2002; LeSage 2014).

While this trend toward greater adoption of machine learning methods among domain scientists is, in general a positive one, we argue that in the spatial sciences this methodological exchange has been wanting. Most applications of ML algorithms in neighborhood analysis are naive in the geographical sense. A widespread practice in the literature is the application of ML methods to reduce multivariate neighborhood data into a smaller, more tractable and interpretable dataset. Such dimensionality reduction techniques have been common in the urban social sciences for decades, with sociologists gravitating toward factor analysis in the form of "ecometrics" (Raudenbush and Sampson 1999; O'Brien et al. 2015) and geographers favoring cluster analysis often termed "geodemographics" (Singleton, Pavlis, et al. 2016; Knaap et al. 2019) but rarely has either of these traditions attempted to incorporate spatial information into the modeling frameworks<sup>1</sup>. Thus while the longevity of the neighborhood analysis literature provides strong evidence for the value of classic and emerging aspatial machine learning methods, the fact that they lack consideration of geography's first law suggests there is a wide berth for improvement (Tobler 1970).

As urban scholars continue to adopt machine learning techniques, we therefore argue there is a clear need to build a stable bridge between these methods and the traditions of neighborhood research, from the perspectives of both social theory and geographical analysis. Here, we address a portion of that gap focused on geographical analysis. Our goals in this paper are threefold; first, we introduce a new, general method for incorporating geographic information into a large family of ML methods. Second, we introduce a set of heuristics for understanding the trade-off between geographic information and multivariate attribute information that we introduce in our framework and demonstrate these trade-offs in an analysis of social and economic data on Kings County (Brooklyn), New York. Third, we stimulate discussion regarding the appropriate specification of spatial effects in non-linear models of urban phenomena, through which we hope to foster increased interdisciplinary exchange and coalesce efforts among quantitative researchers in disparate fields.

---

<sup>1</sup>of course, there are important exceptions to this statement, most notably the field of geographic regionalization. For a recent review see Knaap et al. 2019

## 1.1 Past Explorations & Prior Concerns For Geographic Dimension Reduction

Data-driven geography (Miller and Goodchild 2015) has a long and storied history in geographical analysis (Arribas-Bel and Reades 2018). While recent dialogues about “big” data have focused mainly on the problematic representativeness and meaning of large new datasets derived from “incidental” or “accidental” sensing (Goodchild 2007; Kitchin 2013; Graham and Shelton 2013; Goodchild 2013; Arribas-Bel 2014; Poorthuis and Zook 2017), the long-standing focus in geography of “big” data precedes these datasets. Indeed, a critical avenue of geographical research, that of geodemographics (R. Harris et al. 2005; Singleton and Spielman 2014), is intrinsically a method of dealing with big geographical data. This realm of quantitative geography inherits from the “social area analysis” tradition in sociology (Park and Burgess 1925; Shevky and Bell 1955; Palm and Caruso 1972; Raudenbush and Sampson 1999).<sup>2</sup> In this realm of analysis, large multidimensional census data is reduced into a few main factors model of “latent” characteristics. These latent characteristics are then analyzed, mapped, or otherwise explored to understand the “latent” spatial structure in demographic data. While these practices have routinely been scrutinized for their geographical instability and their occasionally unintuitive summaries of high-dimensional structure (Bell and Greer 1962; Berry 1971; Johnston 1971; Giggs and Mather 1975; Voas and Williamson 2001), it remains true that dimensionality reduction for complex, non-linear relationships in geographical data is an old problem indeed.

The problem arises elsewhere by different names in different areas of geographical science. Since Openshaw (1994), a geographical interest in dimensionality reduction using the Self-Organizing Map (Kohonen 1990) has sought to reduce high-dimensional structures in geographical processes and proven extensively useful. In an exploratory mode of analysis, the Self-Organizing Map and its variants (Bação et al. 2004; Henriques, Bação, et al. 2012; Henriques, Lobo, et al. 2012; Xu et al. 2017; Clark et al. 2017) have been consistently used in the analysis of complex, non-linear demographic relationships (Skupin and Fabrikant 2003; Agarwal and Skupin 2008; Pearson and Cooper 2012; Arribas-Bel, Nijkamp, et al. 2011; Delmelle et al. 2013; Spielman and Logan 2013; Psyllidis et al. 2018). Another consistent interest is the use of Self-Organizing map for data-driven map reprojection (Skupin 2003; Henriques, Bação, et al. 2009; Skupin and Esperbé 2011), which exploits the Self-Organizing Map’s distinctive properties in order to build new or better map projections and transformations. This is because a distinctive property of a Self-Organizing Map is that its transformations are “topology-preserving,” meaning that the fundamental neighbor relationships between observations are preserved by the SOM during learning. Data can be warped, its visual or apparent structural properties adjusted. But, Self-Organizing Maps are trained explicitly using “spatially-correlated learning” (Kohonen 1990, p. 1467), so the position of a point in the lower-dimensional embedding of the data incorporates the learning of surrounding points as well. These maps then allow for the exploration of complex multi-dimensional relationships in a two-dimensional grid amenable to exploratory visualization (Lobo et al. 2012; Spielman and Logan 2013) or clustering (Skupin and Fabrikant 2003; Psyllidis et al. 2018).

---

<sup>2</sup>A recent comprehensive review is provided by Knaap et al. (2019), see also Abbott (1997).

## 2 Understanding Manifold Learning

Building on this, manifold learning is a domain of non-linear dimension reduction research. Manifold learners condense complex non-linear multivariate relationships through to a lower-dimensional “embedding” of the data. The “manifold” is a surface defined in the high-dimensional space by multiple variables; it may be folded, twisted, or otherwise non-linear. These geometric properties reflect the intrinsic non-linearity of multi-dimensional geographically- and internally-dependent data. As a subtype of dimension reduction, manifold learning methods focus more on preserving the properties of this *surface* through to low dimension, rather than finding an optimal dimension reduction of the dataset when treated as individual points. This has profound implications for exploratory data analysis in geography, where non-linearities in both data and geographic structure abound.

This manifold perspective has increased in popularity dramatically in the past twenty years. Indeed, Skupin and Fabrikant (2003)’s contrast of the “discrete object” conceptualization of classical dimension reduction methods and “field conceptualization” self-organizing maps presages the contemporary tension between “local” and “global” structure underlying the manifold and general dimension reduction split:

The rationale behind these so-called manifold learners is that they attempt to retain the local structure of the data, that is, the small pairwise distances between the datapoints, in the latent space. (van der Maaten 2009, p.384)<sup>3</sup>

By this notion of “local” structure, van der Maaten (2009) means that observations that are extremely similar in high dimension should remain near one another in low dimension. Here, “nearness” might be measured by the raw distances between observations, by the set of “most similar” observations, or by defining a topological “neighborhood” as Kohonen (1990) does. Further, we demonstrate what this means, too, in the following sections.

Thus, manifold learning algorithms are a type of dimensionality reduction method. They focus on ensuring that observations that are similar in high dimension remain similar in low dimension, rather than attempting to find the optimal low-dimensional representation that preserves the (dis)similarities between observations. Most manifold learning algorithms can be encouraged to use “spatially-correlated learning” using a variety of heuristics through a method known as “kernel averaging.” We discuss two of these heuristics in the following sections for Isomap, a manifold learning technique proposed by Tenenbaum et al. (2000).

### 2.1 How Isomap Learns Manifolds

While Skupin and Fabrikant (2003)’s perspective on representation in manifold learning methods was indeed prescient, Tenenbaum et al. (2000)’s Isomap was proposed shortly before its publication. Isomap extends and improves upon classical metric Multidimensional Scaling (Mead 1992), resolving many

---

<sup>3</sup>See also page 2 of McInnes et al. (2018), that “[d]imension reduction algorithms tend to fall into two categories; those that seek to preserve the distance structure within the data and those that favor the preservation of local distances over global distance.”



of Skupin and Fabrikant (2003)’s concerns, yet has not received significant attention in dimensionality reduction for geographical analysis.<sup>4</sup> Thus, to explain how we incorporate geography into manifold learning using Isomap, we first explain how Multidimensional Scaling operates. Then, we explain how classical Isomap improves significantly upon Multidimensional Scaling, and note why Isomap is such a powerful extension of Multidimensional Scaling. Finally, we discuss how Isomap can be generalized further to respect the geographical structure of high-dimensional manifolds while becoming Isomap in the limit.

### 2.1.1 Understanding Multidimensional Scaling

Multidimensional Scaling (MDS) is one of the longest-standing methods in dimension reduction (Torgerson 1952; Mead 1992; Borg and Groenen 2005). MDS has many variants, but its most basic form reduces a matrix  $\mathbf{X}$  recording  $P$  features at  $N$  sites into a new “embedding” matrix  $\mathbf{Z}$ , with  $N$  observations over  $P_z$  features, with  $P_z < P$ . To do this, many MDS algorithms minimize *stress*,  $s$ , a function of the distances between pairs of points in the data  $\mathbf{X}$  and those in the lower-dimensional embedding,  $\mathbf{Z}$ :

$$\min s = \sqrt{\sum_{i < j}^N \frac{[d(\mathbf{x}_i, \mathbf{x}_j) - d(\mathbf{z}_i, \mathbf{z}_j)]^2}{d(\mathbf{x}_i, \mathbf{x}_j)^2}} \quad (1)$$

where  $d$  is a distance metric. In addition, *non-metric* MDS generalizes this procedure to allow for non-metric  $d$ -functions that satisfy weaker conditions (Kruskal 1964b; Kruskal 1964a). When stress is minimized, observations that are dissimilar (or, far away from one another in high dimension) in  $\mathbf{X}$  remain dissimilar in the lower-dimensional embedding,  $\mathbf{Z}$ . When looking for a two-dimensional embedding for  $\mathbf{Z}$ , this means that observations that are dissimilar in  $\mathbf{X}$  are far apart in a scatterplot of  $\mathbf{Z}$ . Thus, MDS is often explained as the best-fit distance-preserving projection of the original data into low dimension.<sup>5</sup>

## 2.2 Understanding Isomap

In the original proposal of Isomap, Tenenbaum et al. (2000) discuss dissatisfactions with MDS specifically (and linear dimension reduction methods more generally). One incisive critique from Tenenbaum et al. (2000) is that standard Euclidean distances between observations ( $d(\mathbf{x}_i, \mathbf{x}_j)$ ) may not be a good measure of the true dissimilarity of  $i$  and  $j$  in high dimension. For an explanation, consult Figure 1, reprised from Tenenbaum et al. (2000). In this example, a “global” view of the data may indicate that two points are nearby. However, the two observations exist on very different parts of a non-linear high-dimensional manifold. Using standard Euclidean distance in this case would “shortcut” the true non-linear relationship.

<sup>4</sup>This is despite Jenkins and Matari (2004)’s initial exploration into adding spatio-temporal constraints to Isomap.

<sup>5</sup>Indeed, technical documentation for `scikit-learn` (Pedregosa et al. 2011) goes as far to suggest that MDS on a spherical dataset provides results analogous to an Azimuthal Equidistant Map Projection.

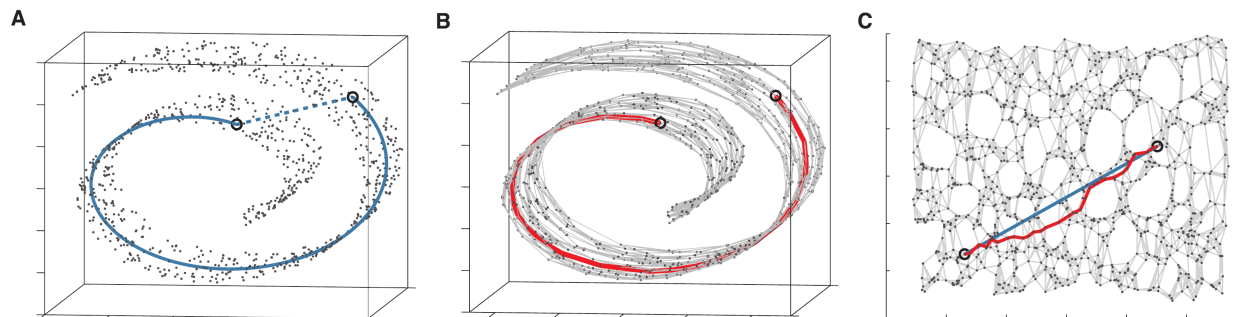


Figure 1: Tenenbaum et al. (2000) illustrate Euclidean “shortcuts” in high dimensions. In Panel A, the blue distance is drawn along the surface of the data in high-dimension. The dashed blue line connecting the two highlighted points is the Euclidean distance between the two points. The Euclidean distance involves a jump that “shortcuts” the manifold altogether, and would suggest the two points are much closer than they would otherwise be if the manifold were known. Panel B shows the distance along the manifold as modeled using “short hops” between similar points, which respects the manifold’s curvature. Finally, Panel C compares the true manifold distance (solid blue) and the “short hop” (red) path length along the manifold after building the Isomap embedding.

To resolve this, Tenenbaum et al. (2000) use local thinking to provide a better model of dissimilarity. The dissimilarity of observations in high dimension might be modeled better by “adding up a sequence of ‘short hops’ between neighboring points.” (p. 2320). In Figure 1, this “sequence of short hops” is the red path defined in Panels B and C. This problem of finding a path along a non-linear manifold is the *manifold shortest path problem*, and involves discovering the shortest path along an unknown high-dimensional surface. The use of these path lengths is central to the Isomap method of manifold learning and dimensionality reduction, and will play a critical role in introducing geography into the method.

For standard Isomap, the “short hops” are first used to build a matrix of the distances between each observation and  $k$  of its most similar peers ( $C_0$ ). Then, this is used to construct a matrix of the shortest path between all pairs of observations ( $C$ ) and MDS applied. In practice, “short hops” are moves along the  $k$ -nearest neighbor graph for  $X$ . This means that, for Isomap,  $d(x_i, x_j)$  is the length of the shortest path connecting  $i$  and  $j$  in a  $k$ -nearest neighbor graph.<sup>6</sup> For observation  $x_i$ , its  $k$ -nearest neighbors are those  $k$  observations most similar to  $i$  itself, and thus are the “closest”  $k$  observations in high dimension according to  $d$ ; the  $k$ -nearest neighbor graph is the union of all of these  $k$ -nearest connections. For clarity, let us call this  $k$ -nearest neighbor graph over  $X$  the  $k$ -similarity graph, to avoid confusing the notions of geographical and feature similarity.<sup>7</sup>

Once the distance of every “short hop” is recorded in  $C_0$ , an all-pairs shortest path algorithm (Floyd 1962) is used to build the full cost matrix  $C$  for standard MDS (Torgerson 1952). Like in the case of

<sup>6</sup>The actual dissimilarity function,  $d$ , may be any statistical distance, but common metrics include Levenshtein or Minkowski distances.

<sup>7</sup>The  $k$ -similarity graph can be symmetrized by *intersection*, such that observations  $i$  and  $j$  are mutually  $k$ -similar if and only if  $x_i$  is  $k$ -similar to  $x_j$  and  $x_j$  is  $k$ -similar to  $x_i$ . It can also be symmetrized by *union*, or left un-symmetrized. If symmetrized, solutions can be found using metric Multidimensional Scaling. Otherwise, solutions are only admissible using non-metric Multidimensional Scaling. Practically, it is simplest to use union symmetrization, as this lets jumps be made forwards and backwards, which retains the original concept of a “short hop” from Tenenbaum et al. (2000).

classic Multidimensional Scaling, points that are near one another in high dimension will still be near one another in a scatterplot of  $\mathbf{Z}$ , so long as there is a short path connecting them along the manifold. This ensures that distances in high dimension follow non-linear structures, since the higher-dimensional relationships “may be highly folded, twisted, or curved.” Tenenbaum et al. (2000, p. 2321) As such, Isomap avoids the short-cutting shown in Figure 1, and instead provides a useful model of non-linear dissimilarity using local thinking. It uses the  $k$ -similarity graph as a “greedy” approach to model the manifold shortest path. It is greedy in that it prefers paths with short link lengths, not the shortest total distance, but it still solves for the shortest total path length along the  $k$ -similarity graph. In practice, this means that the  $k$  parameter must be sufficiently large to contain paths that are close to the optimal manifold shortest path length, but small enough to avoid taking “shortcuts” through the manifold.

### 3 A General Strategy for Building Geographical Machine Learners

Geographical information provides another way to improve manifold learning, as it both expresses and constrains variation in data. Models of complex geographical relationships have tended towards the analysis of the interdependent and spatially-contingent structure of geographic processes, recognizing that everything is related to everything else (Tobler 1970). This geographical autocorrelation can affect the amount of information inherent in a single variable, serving as an intrinsic dimension-reducing factor (Griffith 2005; Griffith 2013). Further, many spatial relationships are inherently *local*: they may vary depending on the site (Fotheringham 1997). Finally, “exogenous” effects themselves can be interact with spatial structure across a variety of geographical process models in a multitude of ways, such as direct & indirect effects in simultaneous autoregressive models (LeSage and Pace 2009, §2.7.1) or nonstationarity in multilevel, (Duncan et al. 1998; Dong and R. Harris 2015), spatial interaction (LeSage and Llano 2013), or random effects models (Murakami and Griffith 2015). Thus, geography is full of studies on the nature of non-linear, intrinsically lower-dimensional dependent data.

In many geographical modeling applications, the relationship between each observation in a geographical process is measured by an affinity or a distance measure relating all pairs of sites (LeSage 2014). Conceptually, each “instance” of a spatial relationship is modelled by this affinity or distance, and the structural effect of this linkage is estimated. In geographically-weighted regression, the “instance” is a site at which a local model is estimated, and the learned effects ( $\beta$ ) change depending on the bandwidth parameter and kernel function ( $\mathcal{K}(\mathbf{C}|\theta)$ ). In a mixed regressive-autoregressive spatial model (a “spatial lag” model (Anselin 1988)), the “instance” is a single site, and the learned effects ( $\beta$ ) change depending on the autoregressive parameter and the spatial filter,  $((\mathbf{I} - \rho\mathbf{W})^{-1})$ . In both cases, the “spatializing” part (the kernel or the spatial filter, respectively) is an entire  $N \times N$  matrix modeling the interaction between sites.<sup>8</sup>

Many machine learning problems are directly based on kernels, too (Hofmann et al. 2008). The so-called “kernel trick” is a long-standing and influential realization used in models of stochastic processes

---

<sup>8</sup>In geographical problems, both kernels and distances are common. Schölkopf (2001) provides a now well-used method to convert distances into kernels, so the distinction does not pose a challenge here.

(for perspective, see Newton 2002, p. 364). Kernel methods in machine learning use direct pairwise relationships between observations to learn functional structure. Thus, just like many geographical problems, effects may not be stationary, are explicitly developed from links between datapoints, and may be highly non-linear. Indeed, nearly any linear or non-linear dimension reduction method can be solved using a non-linear Kernel Principal Components Analysis over a suitable kernel, including both MDS (Williams 2001) and Isomap (Ghodsi 2006). Thus, any geographical manifold learning is also implicitly informed by extensive past work on the spectral properties of geographical processes (Tobler 1969; Boots 1985; Griffith 1996; Griffith 2008). Beyond spectral concerns, the generality of kernel methods for instance-based learning suggests that we can *combine spatial and feature kernels* together in a single kernel for geographical machine learning.

Many existing applications of machine learning in geography involve kernel problems, but have not directly applied kernel mixing methods. At its simplest, mixing kernels involves computing a separate *feature* kernel and *geographical* kernel. For  $i = 1, 2, \dots, N$  observations of  $P$  covariates collected in a design matrix  $\mathbf{X}$ , let the site  $i$  have the  $P$ -length feature vector  $\mathbf{x}_i$  and the geographical position  $\mathbf{s}_i$ . Then, kernel mixing for geographical problems uses the fact that kernels can be combined using averages (Belanche and Tosi 2013). To combine a spatial kernel  $g$  with a feature kernel  $f$ , we use the following geometric mean:

$$\mathbf{K}_{ij} = \sqrt{(f(\mathbf{x}_i, \mathbf{x}_j) * g(\mathbf{s}_i, \mathbf{s}_j))} \quad (2)$$

Geometric kernel averages are desirable for a few reasons. First, for problems where strict constraints are required (e.g. Yuan et al. 2015), this ensures that zero-affinity elements propagate to the final kernel; linear means do not do this. Second, standard geometric means are simpler than linear weighted means because they do not require additional parameters. Third, geometric means retain the interpretation of *weighting* feature similarity by geographical similarity, rather than providing a linear trade-off between feature and geographical similarity.

This “weighting” of feature affinity by geographic structure is conceptually similar to many so-called “local” methods in geography (Fotheringham 1997) that discount or ignore distant structural relationships (Wolf, Oshan, et al. 2018). For example, geographically-weighted regression (Fotheringham et al. 2002) uses geographical weights directly in the computation of regression parameter covariance matrices. This realization has further profound implications for dimensionality reduction, classification, and unsupervised learning applications (Brunsdon et al. 2007; Mason and Jacobson 2007; P. Harris et al. 2011), where path-breaking work has also long recognized the relevance of “neighborhood model[s] for each location” (Gahegan 2000, p. 133). We use this realization of kernel mixing to power a generalization of Isomap below, but any machine learning problem for geographical data can be made to use explicit “geographical learning” using kernel mixing.

### 3.1 Geographing Isomap

For manifold learning, it is difficult to strike a balance between learning “local” structure versus the “global” structure of the data. As discussed above, Isomap focuses on “local” structures in data through

using  $k$ -similarity graph paths to model non-linear manifold distances. For geographical data, proximity provides an additional, distinct measure of non-linear “local” structure. Observations that have similar traits *and* are near one another geographically are close in both a geographical and feature-space sense. Thus, the way Isomap models manifold shortest paths using  $k$ -similarity graphs can be improved by relying on the intrinsic dimension reduction caused by geographical dependence.

Inspired by Church (2018)’s heuristic to simplify a family of facility location problems, we propose a kernel-mixing heuristic like that in Equation 2. This encourages “short hops” along the high-dimensional data surface that make geographical sense. While we can make specifications to forcibly constrain the learning to be local (e.g. Jenkins and Maturi 2004; Yuan et al. 2015), it is more useful to express this as a smooth continuum of possible transformations that exist between the classical method and a geographically-informed technique. With this, a “Geographical Isomap” algorithm is available:

### 3.1.1 Geographical Isomap

1. For all pairs of observations  $(i, j)$ ,
  - (a) compute the geographical distance between sites  $s_i$  and  $s_j$  using a distance metric  $d_s$  and store this in  $\mathbf{D}_s$ .
  - (b) compute the feature dissimilarity between  $x_i$  and  $x_j$  and using a dissimilarity metric  $d_x$  and store this in  $\mathbf{D}_x$ .
2. Combine the information about geographical distance in  $\mathbf{D}_s$  with information about feature dissimilarity in  $\mathbf{D}_x$  to get  $\mathbf{D}$ , the hybrid distance matrix:

$$\mathbf{D} = \mathbf{D}_x * \mathbf{D}_s \tag{3}$$

3. Construct a  $k$ -similarity graph using  $\mathbf{D}$ . Retain the cost of each “short hop” in  $\mathbf{C}_0$ , the first-order path cost matrix.
4. Compute the all-pairs shortest path distances for first-order path cost matrix  $\mathbf{C}_0$ , and call this matrix of all-pairs path lengths  $\mathbf{C}$ . Each path cost in this matrix is the length of a “hybrid manifold path” on  $\mathbf{D}$ .
5. Use an appropriate MDS variant to obtain an embedding in low dimension with minimal deformation of distances in  $\mathbf{C}$ .

### 3.2 Specifying Two Geographical Isomap Approaches

Together, these steps provide a very flexible, general statement for geographical manifold learning. Since  $d_s$  and  $d_x$  can be specified in a flexible manner, the method can be tuned to provide an appropriate spatial and feature metric for many different problem types. Flexibility is desirable because

geographical problems are complex and geographic representations are different depending on context. This level of flexibility, however, can also lead to significant challenges when conducting dimension reduction, as Wattenberg et al. (2016) illustrate for another flexible manifold learning method. Further, because steps 1 & 2 vary only in their inclusion of geographic information, specifying  $d_s$  well will ensure that Isomap is a special case of geographical Isomap. Just as the solution to the problem defined by Church (2018) becomes feasible and optimal to a standard warehouse location problem at a sufficient distance threshold or a geographically-weighted method approaches an aspatial technique as the bandwidth increases (Fotheringham et al. 2002), there are many specifications for  $d_s$  that have this limiting property. Thus, we provide a few recommended specifications for geographical and feature similarity, and illustrate how these specifications relate to the standard Isomap procedure.

### 3.2.1 Using the Tobler Heuristic: Locally-Constrained Isomap

The first specification is directly inspired by Church (2018)’s use of distance banding methods to reduce a problem size. Recall that Isomap models the manifold distances between observations using shortest paths on the  $k$ -similarity graph. This matrix,  $C_0$ , records the distances of the “short hops” between similar observations. Then, the all-pairs shortest paths on  $C_0$  are computed and their costs are stored in  $C$ . MDS is used on  $C$  to obtain the low-dimension coordinates. This approach is “greedy” in the sense that it suggests that manifold shortest paths are well-modelled by paths restricted to the  $k$ -shortest inbound or outbound links at each node, which is not true of shortest paths in general.

Indeed, we suggest that, when near things are more related to one another than distant things, a “useful” manifold path will tend to be both short in feature dissimilarity and geographical distance. Analogous to Church (2018), we will use  $d_s$  to impose a *geographical* threshold distance  $\bar{d}$  while building  $C_0$ . This threshold ensures that “short hops” are geographically-short, while still connecting similar observations. This is an approach that explicitly leverages spatial dependence: nearby observations (within  $\bar{d}$  of one another) are likely “more related” to one another than an equally-similar but geographically-distant observation. Since distance becomes less useful to distinguish observations in high dimension (an implication of the Curse of Dimensionality), we use geography to focus on more relevant connections that are both similar and geographically nearby. Functionally, this places a geographical restriction on edges that can enter the  $k$ -similarity graph. An edge connecting  $i$  and  $j$  in the aspatial  $k$ -similarity graph will be pruned if  $d(\mathbf{s}_i, \mathbf{s}_j) > \bar{d}$ , and the most similar other observation  $k$  will be chosen such that  $d(\mathbf{s}_i, \mathbf{s}_j) \leq \bar{d}$ . The resulting graph is the “ $\bar{d}$ -reduced  $k$ -similarity graph,” where the  $k$  most similar observations are constrained to be selected from *nearby* observations, those within  $\bar{d}$ .

Stated in terms of Eq. 3, let  $g$  be a binary distance band function, where observations are considered “linked” if they are within a threshold distance  $\bar{d}$  of one another:

$$d(\mathbf{s}_i, \mathbf{s}_j) = \begin{cases} 1, & \|\mathbf{s}_i - \mathbf{s}_j\| < \bar{d} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where  $\|\mathbf{s}_i - \mathbf{s}_j\|$  is the Euclidean distance between sites  $i$  and  $j$ . This distance threshold is often used

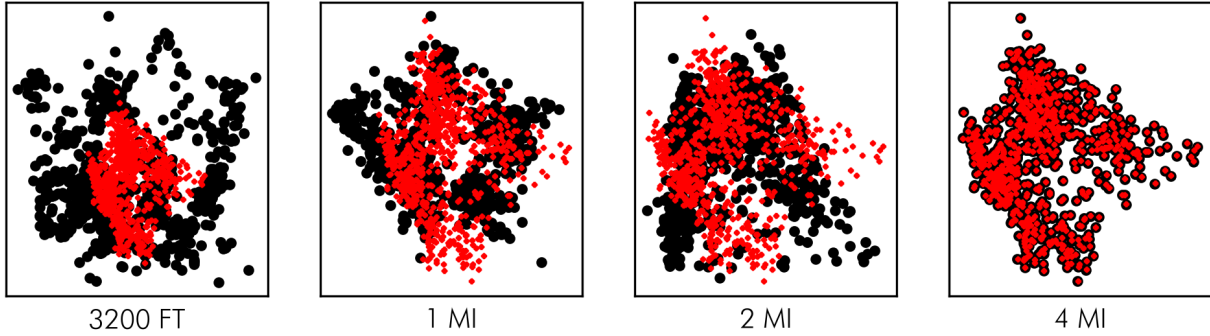


Figure 2: *Locally-Constrained Isomap becomes Isomap as  $\bar{d}$  gets large.*

to model absolute accessibility in geographical problems, since any observation reachable by moving  $\bar{d}$  away from a starting point is considered equally accessible.<sup>9</sup>

In this specification, the path cost matrix  $\mathbf{C}$  approaches that used in standard aspatial Isomap as  $\bar{d}$  becomes large. Because of this, the two techniques yield the same embedding in that case, as illustrated in Figure 2.<sup>10</sup> Intuitively, this is because the “short hops” of Tenenbaum et al. (2000) are forced to become *proximate short hops*; observations can only be considered “similar” if their geographical locations are closer than  $\bar{d}$  to one another. If the aspatial  $k$ -similarity path from  $i$  to  $j$  is composed of moves that are always less than  $\bar{d}$  in length, then that path remains in the “hybrid” geographical-feature distance path in either the geographical or aspatial variant. Since  $\bar{d}$  constrains only a few paths at a time, there is a smooth deformation of the embedding towards Isomap as  $\bar{d}$  increases.<sup>11</sup>

However, many paths in the aspatial  $k$ -similarity graph will *not* be geographically short. This is shown in Figure 3, where the  $k$ -similarity graph for links under  $\bar{d}$  is shown. Each link in this graph is part of the “hybrid” graph, built from  $k$ -similar geographically-short paths. The  $k$ -similarity path, geographical shortest path, and hybrid shortest paths are shown in red, blue, and purple, respectively. If a link is lighter-colored, the link is longer than  $\bar{d}$  (“too distant”) or is not in the  $k$ -similarity graph for links shorter than  $\bar{d} = 3200$  ft (“too dissimilar”). If a path has no lighter-colored links, the path is both feasible and optimal over the  $k$ -similar,  $\bar{d}$ -reduced graph. Only the hybrid path is both feasible and optimal.

A comparison of these different kinds of paths (using a pair of different observations) is provided in Figure 4. There, the area in blue shows US Census tracts in Brooklyn that are within 3200 feet of the originating tract in Crown Heights, Brooklyn (filled in white). The target tract in Prospect Heights is colored in dark grey. Note that the target tract is not within 3200 feet of the source tract. On the right, the most direct geographical path is shown. This is constructed by building the shortest path from “hops” that are less than 3200 feet in length. This results in a path that is very nearly the straight-line path between observations. On the left, the *manifold path* is shown. This manifold path moves through

<sup>9</sup>Euclidean distance can be replaced by travel time or another geographic distance and the problem remains the same.

<sup>10</sup>The two converge precisely at the point when  $\bar{d}$  is the longest geographical distance between an aspatially  $k$ -similar pair.

<sup>11</sup>The striking smoothness and stability of this transformation is best shown in the online supplementary material, where a video of Figure 4 is shown as  $\bar{d}$  increases to the convergence point. This highlights the fact that this process is inherently a *continuous* transformation of geographical space by data.

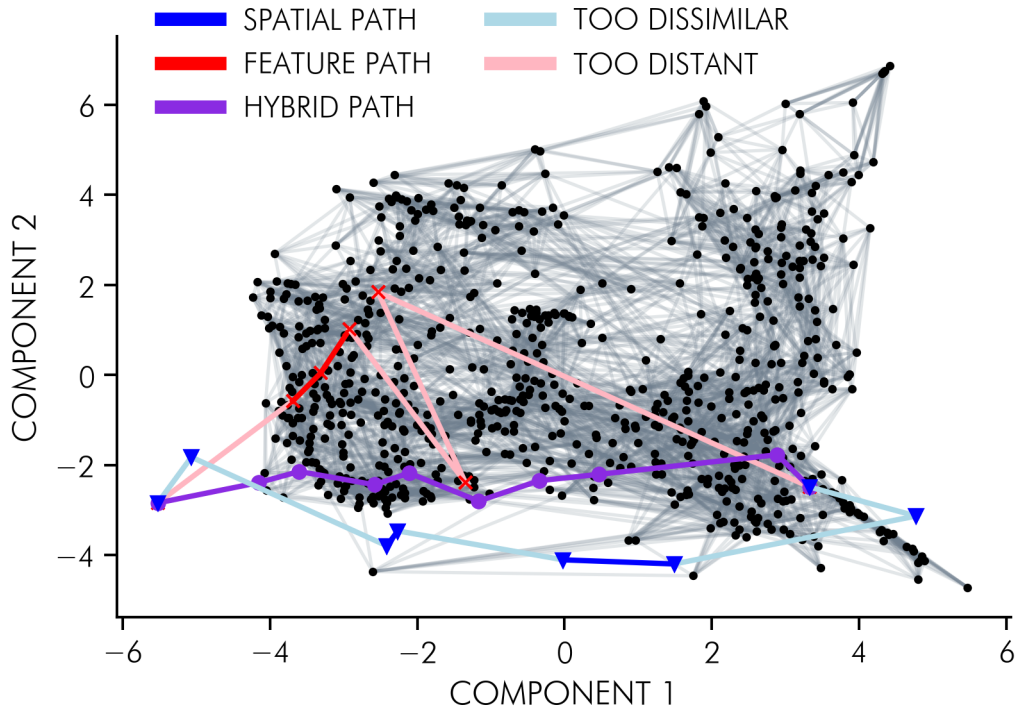


Figure 3: The “hybrid”  $k$ -similarity graph.

four intermediary tracts (shown with red triangles) before arriving at the destination tract. Each of these links is a “short hop” along a manifold in the terms of Tenenbaum et al. (2000). A link between tracts, then, means that the two tracts are  $k$ -similar, and the path overall is the shortest path found over the aspatial  $k$ -similarity graph that connects the Crown Heights origin and Prospect Heights destination tracts. These “short hops” can be of any geographical length. Indeed, two of the links are well over 3200 feet. The Isomap algorithm considers this manifold path as the proper model of the dissimilarity.

A geographical variant, however, would prefer the middle path, which avoids making the long jumps required to optimize similarity. In the center, the geographically-weighted manifold path is shown. Here, the path is constructed according to the steps discussed in Section 3.1.1. The shortest path shown for the hybrid path is from the  $k$ -similarity graph among tracts within 3200 feet of one another. The difference in total length between the manifold path and the hybrid path reflect a kind of “optimality gap,” or loss in similarity arising from more dissimilar but geographically-proximate hops. We return to this (and other) ideas of loss and gain for using geographical information later in Section 3.3.

### 3.2.2 Using the GW-Heuristic: Geographically-Weighted Isomap

As discussed in Section 3, the idea of “localizing” a process using geographical weighting and kernel functions has seen wide adoption in supervised learning, classification, and dimension reduction. Here, we can use the same GW-heuristic to model a preference for shorter paths over longer ones *without* enforcing a hard constraint. Further, we can pick still pick the  $d_s$  so that the algorithm still tends to



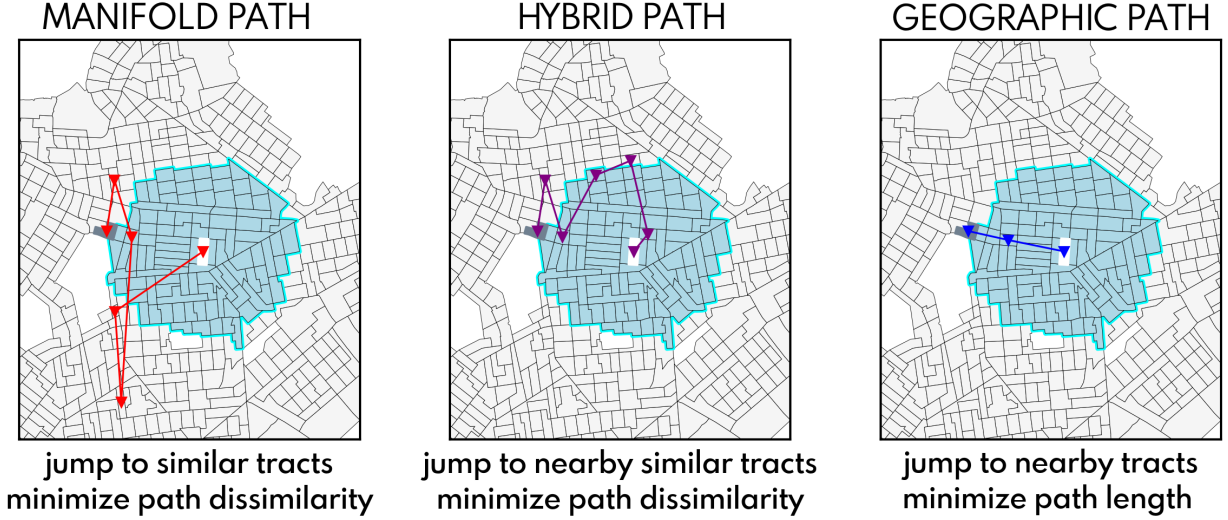


Figure 4: Visualizing the geography of three different path types.

Isomap as a limiting case. For a geographical distance function  $d_s$  with parameter  $\theta$  and Euclidean distance for  $d_X$ , a sufficient condition for convergence to Isomap when  $\theta$  approaches a special value  $\bar{\theta}$  is that the final distance matrix  $\mathbf{D}$  remains a positive scalar multiple of the original feature distance matrix:

$$\mathbf{D}_X * \mathbf{D}_s \approx c \mathbf{D}_X \quad c > 0 \quad (5)$$

One such function is the exponential distance-weighting function:

$$d(\mathbf{s}_i, \mathbf{s}_j | \alpha) = \|\mathbf{s}_i - \mathbf{s}_j\|^\alpha \quad \alpha > 0 \quad (6)$$

Using this function to model geographical proximity,  $\mathbf{C}_0$  becomes a hybrid cost matrix, containing both feature-space information on dissimilarity and geographical information about proximity:

$$\mathbf{D}_{ij} = d(\mathbf{x}_i, \mathbf{x}_j) \|\mathbf{s}_i - \mathbf{s}_j\|^\alpha \quad (7)$$

As  $\alpha \rightarrow 0$ , the algorithm again becomes Isomap: all geographical distances collapse to 1, the  $k$ -similarity graph approaches that used in classical Isomap, and geography is ignored.<sup>12</sup>

The behavior of the embedding with respect to the scaling parameters  $\alpha$  is shown in Figure 5. As  $\alpha$  increases, geographical proximity comes to dominate in importance, forcing geographically remote observations to spread further out. The effect attenuates when  $\alpha \rightarrow \infty$ , reflecting the fact that the variation in only the largest distance dominates at that scaling factor. For  $0 < \alpha < 1$ , the geographical “cost” of making distant hops decreases until  $\alpha = 0$ , where  $k$ -similar lengths of any geographical distance are considered acceptable and the Isomap embedding is recovered.

<sup>12</sup>While there may be some special  $\bar{\alpha} > 0$  exists such that the geographically-weighted  $k$ -similarity graph is the same as Isomap’s  $k$ -similarity graph, those value(s) of  $\bar{\alpha}$  depend on the geographical and feature-space structure in the data in each case.

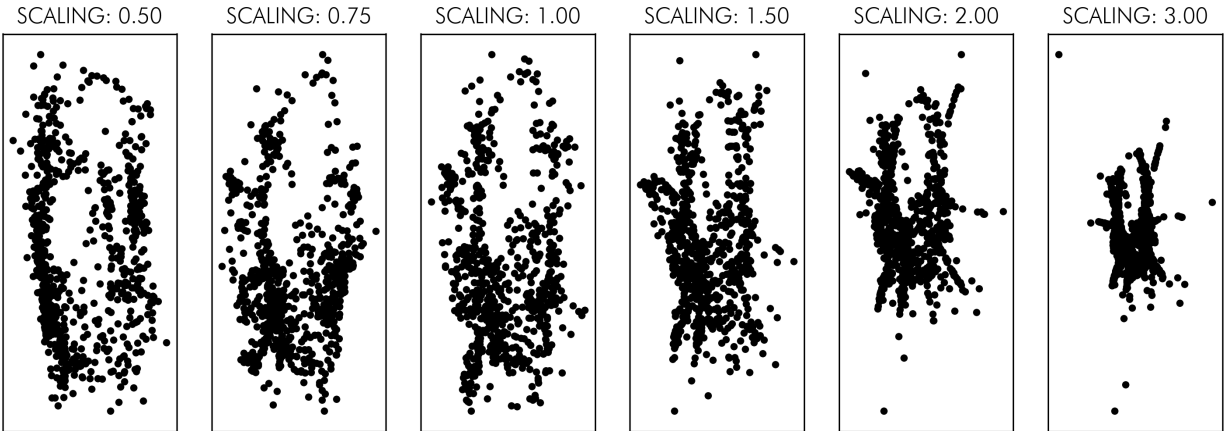


Figure 5: Distances become magnified as the scaling parameter,  $\alpha$ , increases

### 3.3 Assessing Geographically-Learned Manifolds

The “quality” of a low-dimensional embedding is difficult to characterize, as noted routinely by scholars in machine learning. While subjective visualization methods have been used extensively (Ultsch and Siemon 1990; Konig 2000; Spielman and Thill 2008; Henriques, Bação, et al. 2012; Mokbel et al. 2013; Gracia et al. 2014; Moosavi 2017), a large amount of recent work has aimed to define new measures of manifold quality itself. Many methods focus on the extent to which observations that are near one another in high dimension are preserved in low dimension (Bauer and Pawelzik 1992; Villmann et al. 1994; L. Chen and Buja 2009; Goldberg and Ritov 2009; Lee and Verleysen 2009; Lee and Verleysen 2010; Meng et al. 2011) although recent work has explored other information, such as the preservation of “landmark” properties (Zhang et al. 2012), topological structures like holes (Paul and Chalup 2017), or relative orientation (D. Chen et al. 2019). The development of both visual and statistical analytics for the examination of manifold quality and structure indicates that having multiple different ways to think about and measure manifold quality is useful.

Since geographical work on manifold learning has typically followed Openshaw (1994)’s adoption of Kohonen (1990), the “U-matrix” is often used (Ultsch and Siemon 1990), but this visualization strategy is not as useful for other manifold embedding methods. Thus, we mainly visualize the manifold embedding using paired map-plot views that link a variable’s spatial distribution to its distribution in the embedding. We also suggest using popular notions of “neighborhoods” to landmark locations in the embedding space.<sup>13</sup> However, to assess the Isomap variants’ differences, new diagnostics are needed to formalize what is “gained” using geographical information. Toward that end, we propose three new relative quality measures.

<sup>13</sup>Using neighborhoods provided by [zillow.com](http://zillow.com), an online housing market reconnaissance platform.

### 3.3.1 Relative Stress

Any method that involves MDS will have a measure of stress (Equation 1). This quantity is minimized by the MDS method. We use a relative error formula to develop *relative stress*, or the change in stress from aspatial Isomap caused by including geographic information:

$$s_r = \frac{s_v - s}{s_v} \quad (8)$$

where  $s_v$  is the stress of the standard Isomap procedure. When  $s_r$  is positive, aspatial Isomap has higher solution stress than when geography is incorporated. When  $s_r$  is negative, Isomap has lower stress than a geographical variant. As suggested by the convergence of geographical Isomap to the aspatial Isomap in Figure 2, relative stress approaches zero when solution stress becomes similar.

### 3.3.2 Relative Efficiency

We suggest a measure of *relative efficiency*, which measures the median percentage change in path lengths. Tenenbaum et al. (2000) suggests that short total paths and long individual segments may indicate more “short-cutting,” where a path ignores non-linear structure in high dimension. For geographical Isomap, paths may either increase or decrease in length. Path lengths can get longer because the path becomes more circuitous along the manifold, as it must now include both geographically-near and feature-similar observations. Decreases in path length occur because of the “greedy” nature of Isomap’s  $k$ -similarity graph solution to the manifold shortest path problem: the  $k$ -similarity graph may not necessarily contain the optimal set of “short hops” to connect two observations, and a shorter total path could be built from slightly longer hops that still travel along the manifold. However, for the geographically-weighted  $\mathbf{C}$ , we cannot compare path lengths directly: they involve distances in entirely different units due to the kernel mixing. Therefore, to ensure comparability, we instead recover the “raw” feature dissimilarity of paths used by the geographical variant.

To do this, we rely on the so-called “predecessor matrix” used in all-pairs shortest path algorithms that construct  $\mathbf{C}$ . This predecessor matrix can be used to rebuild edge sequences for paths whose costs are contained in  $\mathbf{C}$  (Ahuja 2017). Then, the predecessors are used to compute total path lengths using costs from  $\mathbf{D}_X$ , instead of  $\mathbf{C}_0$ . This matrix,  $\mathbf{C}_X$  is identical to  $\mathbf{C}$  for Isomap, but may be distinct from  $\mathbf{C}$  for geographically-mixed kernels. Thus, using  $\mathbf{C}_X$ , we can define *median relative efficiency* as the median percentage change in feature dissimilarity when geographical information is introduced:

$$e = \text{median}_{ij} \left( \frac{[\mathbf{C}_v]_{ij} - [\mathbf{C}_X]_{ij}}{[\mathbf{C}_v]_{ij}} \right) \quad i, j = 1, 2, \dots, N \quad (9)$$

where  $\mathbf{C}_v$  is the cost matrix for standard Isomap.<sup>14</sup> Efficiency approaches zero when two embeddings use the same manifold paths, but can be either positive or negative.

<sup>14</sup>The median is used to account for outlying distances. These can skew the mean change significantly, such as when the GW-Isomap scaling parameter increases beyond 1.

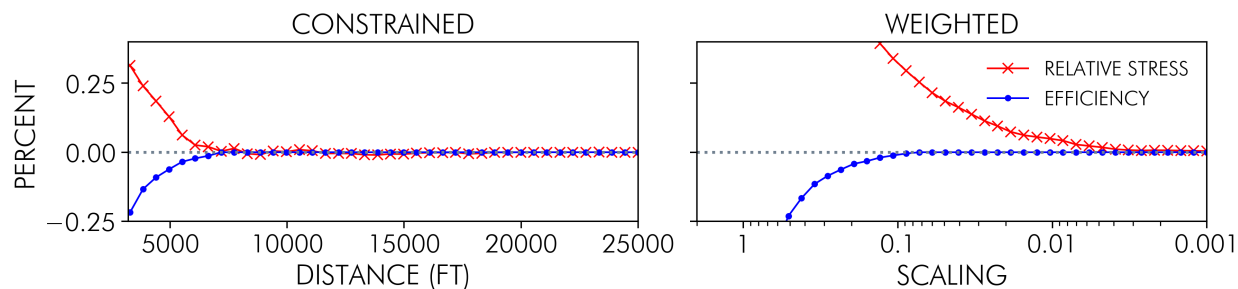


Figure 6: *Stress & Efficiency as a function of the spatial parameters in two spatial Isomap variants.*

### 3.3.3 Gain

Since both efficiency and gain are measured in percentages relative to Isomap, they are directly comparable. In most cases, efficiency is negative; the introduction of geographical information will tend to cause the typical path to increase in length. When relative stress is positive, the introduction of geographical information reduces solution stress, indicating improved solution fit. When relative stress is negative, geographical information increases the stress of the embedding and worsens solution fit. The *gain* from introducing geography to the solution can then be thought of as any reduction in stress over the loss of efficiency:

$$g = s_r + e \quad (10)$$

So, in the common case,  $g$  measures an “extra” improvement in stress over relative loss of efficiency. When Isomap has lower stress and shorter paths, gain is negative. But, if geographical information lowers solution stress, gain is positive.

## 4 Interpreting Geographical Dimension Reductions

In the sections that follow, we interpret the results of the manifold learning first by examining stress and efficiency, then by exploring visually how an embedding stretches or compresses each variable in a dataset, and closes by using “neighborhoods” as landmarks to examine demographic and geographic cohesion in the resulting embedding. Eight variates encompassing racial, ethnic, and economic information in the 2017 5-Year American Community Survey were used. These were compressed into two dimensions using the two geographical Isomap variants presented in Section 3.1.<sup>15</sup>

### 4.1 Stress, Efficiency, and Gain

First, the smoothness of change and limiting patterns presented in Figure 2 apply regardless of whether the embedding is built using geographically-weighted or locally-constrained variants. This means that small changes in the geographical parameter tend to make small changes in the embedding. As such,

<sup>15</sup>In addition to Wolf, Germuska, et al. (2016) and Pedregosa et al. (2011), all software used in this analysis is made available in free software in the Python library for spatial analysis (Rey and Anselin 2007).

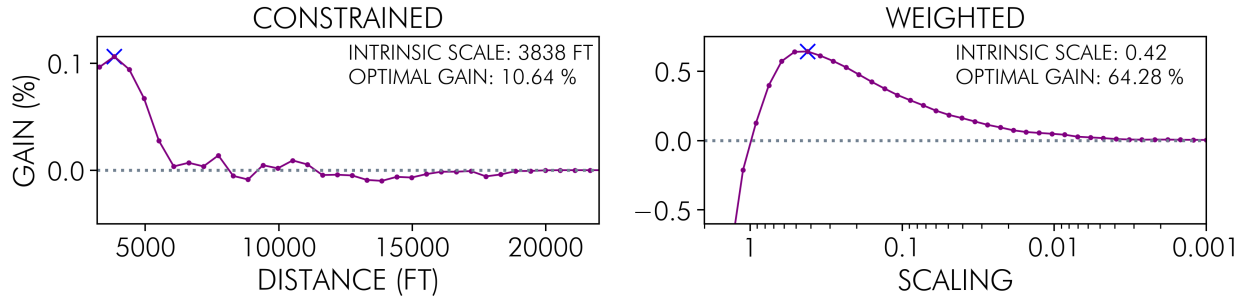


Figure 7: *Gain, the net reduction in stress relative to the increase in path lengths, as a function of the spatial parameters in two spatial Isomap variants.*

the resulting low-dimensional embedding transitions smoothly from an emphasis on geographical information to an emphasis on feature information. This is shown explicitly in the video provided in the supplementary material, which reveals both the limiting pattern behavior and the smoothness of change across geographical dissimilarity parameter values.

We demonstrate these same patterns in Figure 6, albeit in a less dynamic fashion. This figure shows the change in relative stress and efficiency measures discussed in Section 3.3 as a function of the geographical dissimilarity parameter for the two geographical Isomap variants. Notably, relative stress and efficiency values change smoothly with respect to the distance threshold or scaling parameters. For either geographical variant, this suggests what the supplementary video demonstrates directly: the kernel averaging discussed above is well-behaved since statistics about the embedding change smoothly as the geographical parameter changes. Further, the limiting properties are confirmed, as well; relative stress and efficiency both approach zero as the geographical parameter approaches its limiting value.

Since both relative stress and efficiency go to zero as each geographical Isomap variant converges to the aspatial Isomap solution, gain goes to zero, too. But, since relative stress and efficiency may go to zero at different rates, the maxima for gain can be used as a point indicating the best (linear) trade-off between reduced stress and longer paths. The geographical parameter at this maxima also provides an estimate of the “intrinsic scale” of the process, a geographical analogue to the “intrinsic dimensionality” of a process (Trunk 1968; Verveer and Duin 1995) studied elsewhere in supervised geographical learning (Wolf, Oshan, et al. 2018; Murakami, Lu, et al. 2019).

In both cases, efficiency loss approaches zero faster than the relative improvement in stress. This means that, for some range of values, geographical variants will have lower stress and an acceptable increase in the average manifold path length. We see this explicitly in Figure 7, which shows the net gain for both variants as a function of the distance parameter. Again, this function changes smoothly and can be optimized to reflect a balance between stress reduction and path distance increase.

The geographical kernel parameter at this maximum gain is an estimate for the “intrinsic scale” of the data in both problems: for the locally-constrained Isomap method, a local constraint of 3838 feet has the highest gain, meaning that it has the largest gap between the percentage improvement in stress and percentage increase in path lengths, relative to aspatial Isomap. Practically, this means that

you can find paths in the data with links no longer than 3838 feet, but that still generally avoid “short-cuts” through the manifold. In the geographically-weighted variant, the maximum gain is reached when  $\alpha = .42$ , suggesting that something close to the square-root of the map distances is the optimal geographical kernel. Likewise, this means that using the weighted distances  $d(\mathbf{x}_i, \mathbf{x}_j) \|\mathbf{s}_i - \mathbf{s}_j\|^{.42}$  provides a geographically-weighted  $k$ -similarity graph with paths that still are slightly longer in terms of manifold distance, but provide a low-dimensional embedding with considerably lower stress.

## 4.2 Geographical and Similarity Structures in Low Dimension

The relationship between the low-dimensional embedding and the geographical distribution of each variable is shown in Figures 8 & 9 for the two spatial methods at their optimal gain values. Here, the columns represent the version of Isomap used to conduct dimensionality reduction. Each row corresponds to one of the eight variables taken from the 2017 American Community Survey used in dimensionality reduction. The colors are linked, so that brighter tracts/points have higher values of that variate and darker colors have lower values. Overall, the embedding structures are quite different, but exhibit similar macro-level properties.

The constrained embeddings strongly comport with racial/ethnic axes. Shown in Figure 9, these variables tend to exist strongly into specific regions of an embedding. Each racial/ethnic group is strongly separated in the embeddings, but only the aspatial Isomap merges together the geographically-disparate mainly-Hispanic or mainly-white tracts into a single area of the embedding. The geographical methods, while compressing demographically-homogenous observations together, do not force geographically-distant but demographically-homogenous areas together.

Another way of examining what areas shrink and grow together is by visualizing the geographical distribution of the component outputs after dimension reduction, shown in Figure 10. There, we see that the patterns of the two components from *any* variant of Isomap are strongly geographically patterned, expressing the geographical structure inherent in the data and relatively low “intrinsic” dimensionality. However, we see that the gradients in each map are different, suggesting that the different hybridizations of geographical proximity and feature similarity cause different structures to be persist from high-dimension to the lower-dimensional embedding. For instance, component 1 in the locally-constrained embedding has higher scores in the north west than the geographically-weighted embedding; the weighted embedding also exhibits a nonlinearity in geographical structure, with a higher (relative) component 1 score in the hispanic areas in north Brooklyn, appearing more similar to the aspatial solution than the locally-constrained solution. Further, the locally-constrained version Altogether, the geographically-weighted version reduces much more stress, but regardless: geographical Isomaps are quite distinct from each other and from Isomap in both embedding- and geographical-space.

## 4.3 Interpreting Geographical Landmarks

Whereas the selection and interpretation of “landmark” structures in manifold learning is a known problem (Silva et al. 2006; Zhang et al. 2012), geographical “landmarks” are actually both readily available

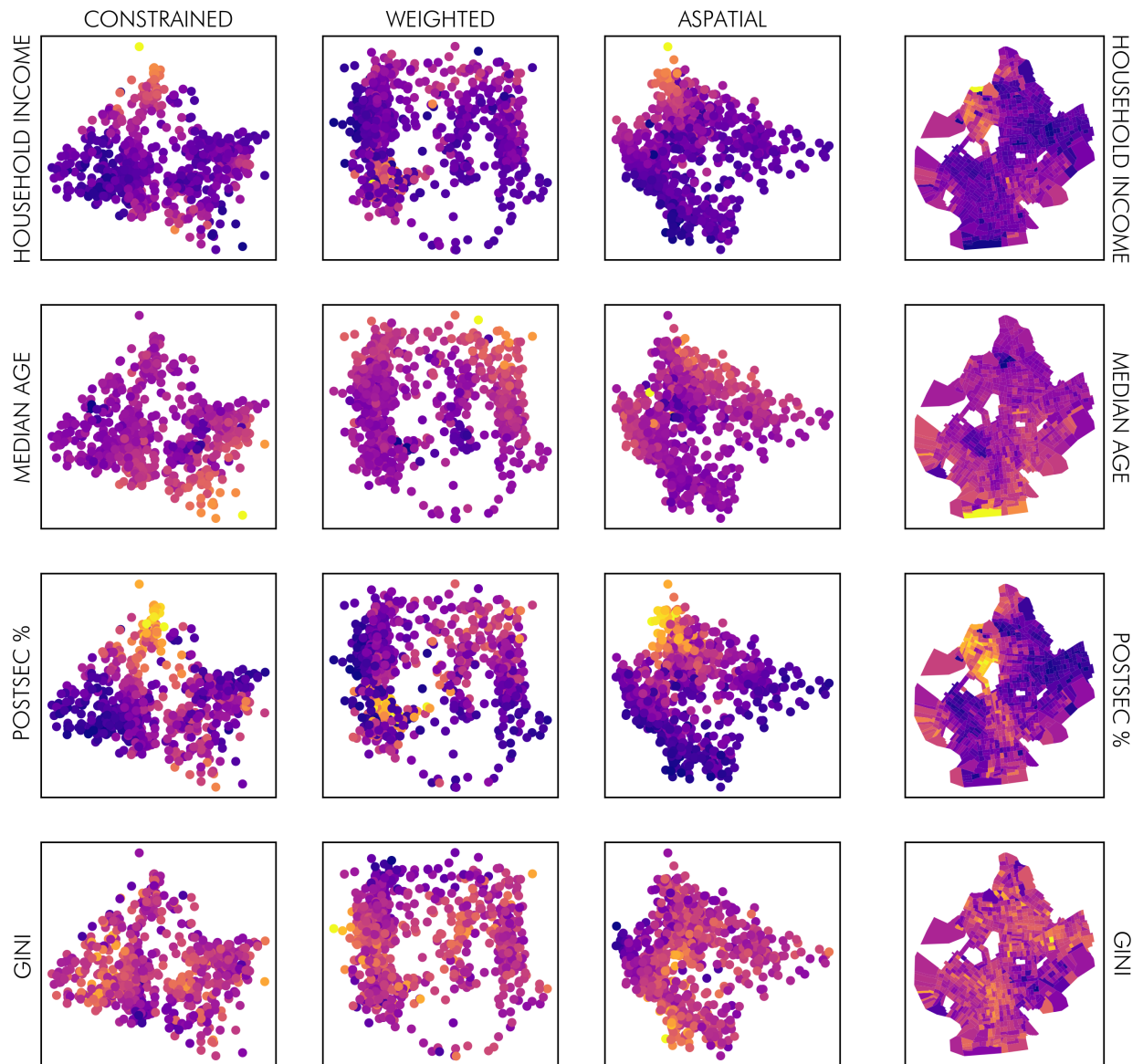


Figure 8: Embeddings at the maximum gain for geographical Isomap variants and standard Isomap, with the first four variates.



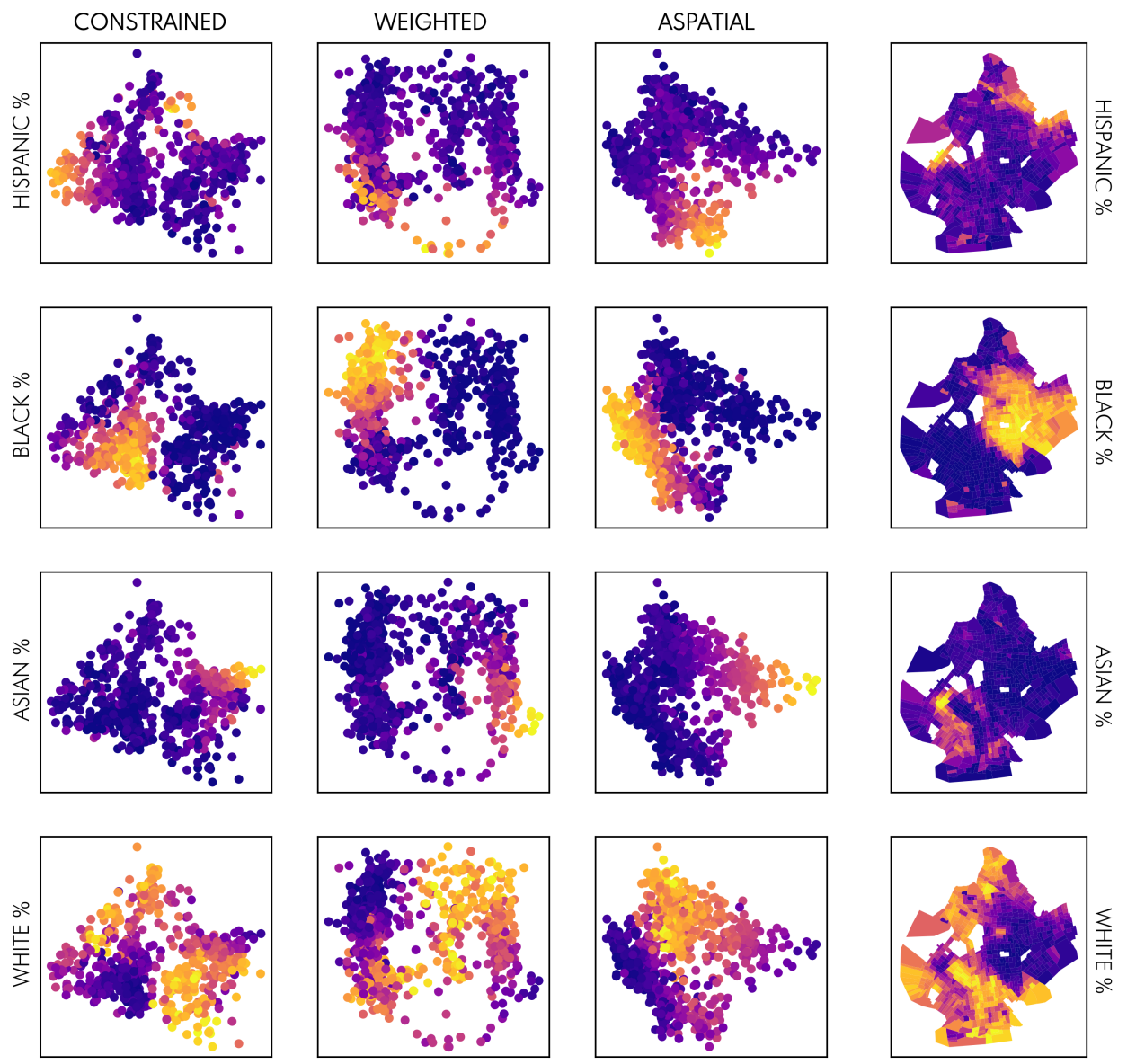


Figure 9: Embeddings at the maximum gain for geographical Isomap variants and standard Isomap, with the second four variates.



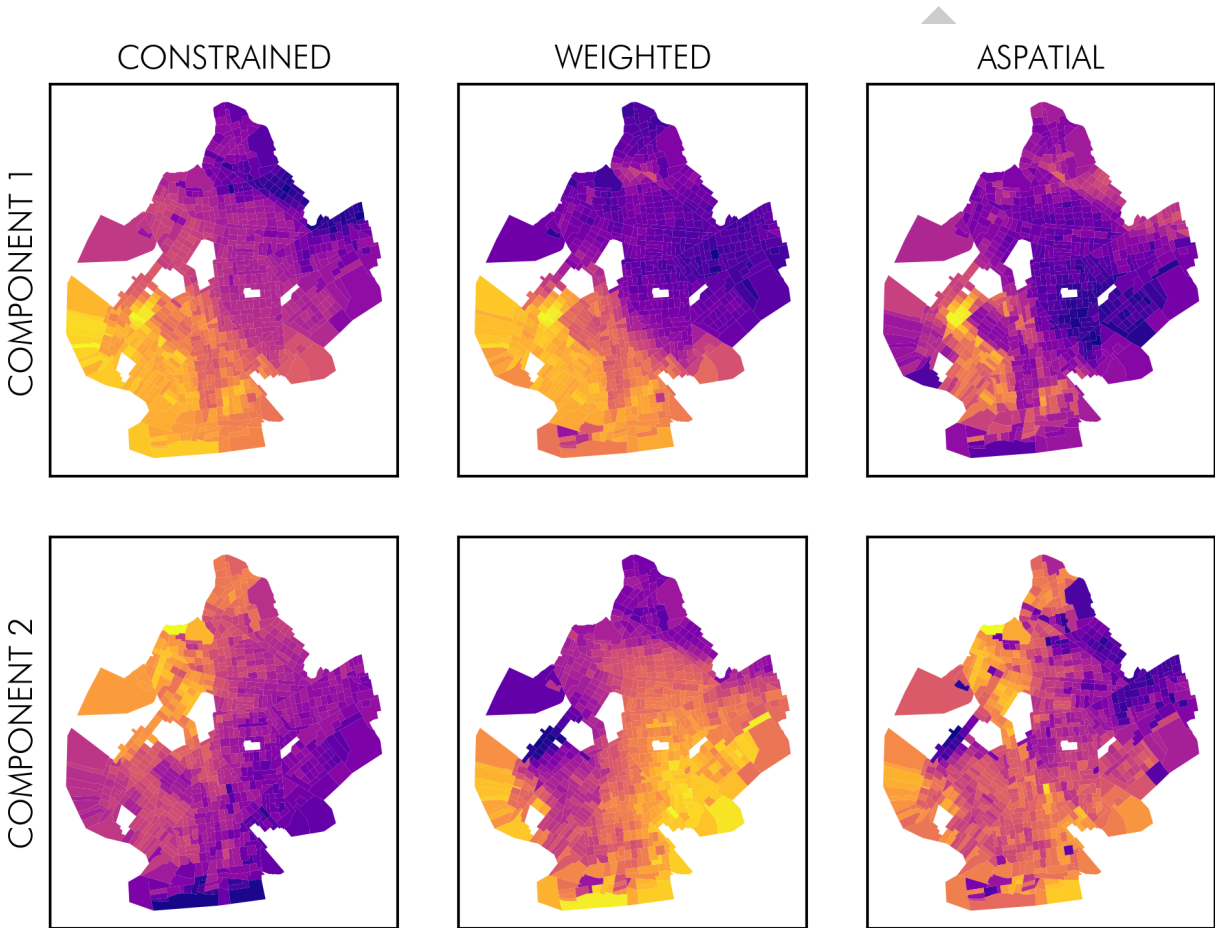


Figure 10: *Maps of the component outputs from dimension reduction.*

and simple to define for manifold learning. A “landmark” in this context is a known structure intended to be preserved during dimension reduction. For geographical problems, there are a variety of meaningful geographies used in planning, property markets, and popular language that can be used to orient ourselves in the manifold. One such way is to use neighborhoods as landmarks, linking the dimension reduction to a known region of the map. Eight neighborhoods in Brooklyn are shown in Figure 11 with varying levels of geographical remoteness and demographic distinctiveness and coherence. In that figure, we have included areas that are mainly-hispanic (Bushwick) and Asian (Sunset Park), African American (East Flatbush), and White (Greenpoint) to provide a sense of orientation in the major axes of the dimension reduction. This visualization strategy also suggests how cohesive individual neighborhoods are in terms of the learned manifold, and provides a vocabulary we can use to link regions on the map and the embedding.

For instance, Sunset Park contains both significant Asian and white populations, given its geographically-compact Chinatown. In the aspatial Isomap solution, this forces the neighborhood to distribute across the bottom-right edge of the embedding. Likewise, the neighborhood exhibits its unique demographic profile in the geographically-weighted Isomap and locally-constrained Isomap variants, too. This suggests that the neighborhood’s demographic structure and its geographical status on the western edge of Brooklyn drive its location in any of the dimension reductions. However, East Flatbush is centrally-located in Brooklyn and is predominantly African American. This demographic cohesion results in it remaining coherent in the resulting embedding; however, its geographical centrality shifts it further from the edge. A remote, largely white neighborhood like Greenpoint experiences improved coherence in the geographical embeddings, but larger diverse and relatively remote neighborhoods like Williamsburg, Bushwick, or Bedford-Stuyvesant are again made much more coherent in the spatial dimension reductions than in the aspatial variant. Overall, this suggests that geographical centrality/distance and feature coherence are clearly treated differently by the dimension reduction methods; neighborhoods that are demographically-coherent appear to be preserved after dimension reduction, but some communities are made more coherent in the dimension reduction when considering their geographical structure and relationship to other nearby areas.

## 5 Conclusion

Dimension reduction is a long-standing interest in geography, yet significant advances in computer science and machine learning have not been strongly-integrated with contemporary research. New directions in *manifold learning*, or non-linear dimension reduction, stand to significantly improve geographical analysis. Geographical problems are often characterized by high-dimensional non-linear relationships and geographically non-stationary behavior. Further, there is often no “outcome” to be modeled directly in exploratory spatial data analyses, only data to be simplified or dimensions to be identified as “useful” for further explanatory or predictive models. Thus, manifold learning is one highly-relevant frontier of dimension reduction research, and provides many entry points for geographical researchers.

As suggested by Kohonen (1990), geographical machine learning methods must deeply integrate



Figure 11: Maps of the component outputs from dimension reduction.

geographic structure into learning in order to benefit. Kernel averaging, the combination of multiple kernels, can be adapted to geographical problems in prediction, classification, and dimension reduction. As we suggest, kernel averaging provides a powerful, general purpose technique to incorporate spatial information into kernel-based machine learning problems. In making this argument, we develop a spatial variant of Tenenbaum et al. (2000) and demonstrate how geographical machine learning may provide different insights and different results when applied to unsupervised learning of high-dimensional geographic data. Through kernel averaging, we can encourage machine learning methods explicitly to learn geographically.

The preservation of geographical structure in dimension reduction is a critical property for geographical community detection, neighborhood analysis, and spatial modeling. For example, an aspatial Isomap procedure will treat observations from distinct (but attribute-similar) communities as equivalent. However, as we show here, geographical Isomap methods ensure that geographical structure enters into the embedding solution. It does so without dominating the solution, however, forcing the solution to be purely geographical as is done in explicit data-driven regionalization methods. Learning becomes “geographically-encouraged,” but not explicitly constrained or forced to respect pre-conceived geographical structure. Developments in “spatially-encouraged machine learning” will have profound consequences for geodemographics, a domain heavily involved in dimension reduction in cases where a compromise between geographical regularity, temporal stability, and feature homogeneity can be difficult to strike (Voas and Williamson 2001; Singleton and Spielman 2014; Singleton, Pavlis, et al. 2016), as well as other urban data science topics. Clustering the results of a hybrid geographical-manifold dimension reduction, where nearby observations are *both* similar and feature-near, may yet blend the dichotomy between spatial constraint and feature optimality (Knaap et al. 2019).

Beyond the implications of these methods for geodemographics, the usefulness of kernel averaging is heightened by the ability to define explicit tradeoffs between “local” and “global” ends of a geographical machine learning algorithm’s parameter space. This allows for the identification of an “optimal scale” in an unsupervised learning context; this is the scale at which the geographical solution has the largest improvement in the objective for dimension reduction, accepting that it may entail a loss of efficiency other aspects of the algorithm. This gain in fit relative to loss in efficiency is well-behaved, consistent from problem to problem, and provides a way to estimate geographical scale in complex high-dimensional data without having an explicit outcome variate or process model. Thus, this method may be of significant use in contexts where scale itself is under examination. Since many modern dimension reduction methods can be re-stated as kernel Principal Components Analyses (Ghodsí 2006), kernel mixing can immediately be applied to other methods in the same way suggested here.

Thus, mixing geographical and feature kernels provides a useful and extremely flexible method for incorporating geography into unsupervised machine learning problems. This work is part of a larger agenda to provide geographic machine learning methods that leverage geography explicitly, rather than simply using machine learning in service of geographical analysis. Geographic kernel mixing technique stands to be of extreme benefit simply from the generality of kernel learning problems in machine learning. Thus, through a new geographical Isomap, we show that kernel mixing is an effective way to

induce geographical learning in a classical machine learning technique and, in the process, opening a variety of new questions in the development of geographical machine learning methods. Future work on explicitly-geographical machine learning in geographic data science (Singleton and Arribas-Bel 2019) should continue to expand on kernel mixing as a way of “geographing” machine learning methods.

## References

- Abbott, A. (1997). Of Time and Space: The Contemporary Relevance of the Chicago School. *Social Forces*, 75(4), 1149–1182. doi:10.2307/2580667
- Agarwal, P. & Skupin, A. (2008). *Self-Organising Maps: Applications in Geographic Information Science*. John Wiley & Sons.
- Ahuja, R. K. (2017). *Network Flows: Theory, Algorithms, and Applications* (1st). Pearson Education.
- Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer.
- Anselin, L. (2002). Under the hood issues in the specification and interpretation of spatial regression models. *Agricultural economics*, 27(3), 247–267.
- Arribas-Bel, D. (2014). Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography*, 49, 45–53. doi:10.1016/j.apgeog.2013.09.012
- Arribas-Bel, D., Nijkamp, P., & Scholten, H. (2011). Multidimensional urban sprawl in Europe: A self-organizing map approach. *Computers, Environment and Urban Systems*, 35(4), 263–275. doi:10.1016/j.compenvurbsys.2010.10.002
- Arribas-Bel, D. & Reades, J. (2018). Geography and computers: Past, present, and future. *Geography Compass*, 12(10), e12403. doi:10.1111/gec3.12403
- Baço, F., Lobo, V., & Painho, M. (2004). Geo-Self-Organizing Map (Geo-SOM) for Building and Exploring Homogeneous Regions. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, . . . H. J. Miller (Eds.), *Geographic Information Science* (Vol. 3234, pp. 22–37). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-540-30231-5\_2
- Bauer, H.-. & Pawelzik, K. R. (1992). Quantifying the neighborhood preservation of self-organizing feature maps. *IEEE Transactions on Neural Networks*, 3(4), 570–579. doi:10.1109/72.143371
- Belanche, L. A. & Tosi, A. (2013). Averaging of kernel functions. *Neurocomputing. Advances in Artificial Neural Networks, Machine Learning, and Computational Intelligence*, 112, 19–25. doi:10.1016/j.neucom.2012.11.044
- Bell, W. & Greer, S. (1962). Social Area Analysis and Its Critics. *The Pacific Sociological Review*, 5(1), 3–9. doi:10.2307/1388270
- Berry, B. J. L. (1971). Introduction: The Logic and Limitations of Comparative Factorial Ecology. *Economic Geography*, 47(sup1), 209–219. doi:10.2307/143204
- Boots, B. N. (1985). Size Effects in the Spatial Patterning of Nonprincipal Eigenvectors of Planar Networks. *Geographical Analysis*, 17(1), 74–81. doi:10.1111/j.1538-4632.1985.tb00827.x
- Borg, I. & Groenen, P. J. F. (2005). *Modern Multidimensional Scaling: Theory and Applications* (2nd ed.). Springer Series in Statistics. New York: Springer-Verlag.

- Brunsdon, C., Fotheringham, S., & Charlton, M. (2007). Geographically weighted discriminant analysis. *Geographical Analysis*, 39(4), 376–396.
- Chen, D., Lv, J., Yin, J., Zhang, H., & Li, X. (2019). Angle-based embedding quality assessment method for manifold learning. *Neural Computing and Applications*, 31(3), 839–849. doi:10.1007/s00521-017-3113-6
- Chen, L. & Buja, A. (2009). Local Multidimensional Scaling for Nonlinear Dimension Reduction, Graph Drawing, and Proximity Analysis. *Journal of the American Statistical Association*, 104(485), 209–219. doi:10.1198/jasa.2009.0111
- Church, R. L. (2018). Tobler's Law and Spatial Optimization: Why Bakersfield? *International Regional Science Review*, 41(3), 287–310.
- Clark, S., Sisson, S. A., & Sharma, A. (2017). Nonlinear manifold representation in natural systems: The SOMersault. *Environmental Modelling & Software*, 89, 61–76. doi:10.1016/j.envsoft.2016.11.028
- Delmelle, E. (2016). Mapping the DNA of Urban Neighborhoods: Clustering Longitudinal Sequences of Neighborhood Socioeconomic Change. *Annals of the American Association of Geographers*, 106(1), 36–56. doi:10.1080/00045608.2015.1096188
- Delmelle, E., Thill, J.-C., Furuseth, O., & Ludden, T. (2013). Trajectories of Multidimensional Neighbourhood Quality of Life Change. *Urban Studies*, 50(5), 923–941. doi:10.1177/0042098012458003
- Dong, G. & Harris, R. (2015). Spatial autoregressive models for geographically hierarchical data structures. *Geographical Analysis*, 47(2), 173–191.
- Duncan, C., Jones, K., & Moon, G. (1998). Context, composition and heterogeneity: Using multilevel models in health research. *Social Science & Medicine*, 46(1), 97–117. doi:10.1016/S0277-9536(97)00148-2
- Floyd, R. W. (1962). Algorithm 97: Shortest Path. *Communications of the ACM*, 5(6), 345–. doi:10.1145/367766.368168
- Fotheringham, A. S. (1997). Trends in quantitative methods I: Stressing the local. *Progress in Human Geography*, 21(1), 88–96. doi:10.1191/030913297676693207
- Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley.
- Gahegan, M. (2000). On the Application of Inductive Machine Learning Tools to Geographical Analysis. *Geographical Analysis*, 32(2), 113–139. doi:10.1111/j.1538-4632.2000.tb00420.x
- Ghodsi, A. (2006). *Dimensionality Reduction: A Short Tutorial* (tech. rep. No. 37). University of Waterloo. Ontario, Canada.
- Giggs, J. A. & Mather, P. M. (1975). Factorial Ecology and Factor Invariance: An Investigation. *Economic Geography*, 51(4), 366–382. doi:10.2307/142921
- Goldberg, Y. & Ritov, Y. (2009). Local procrustes for manifold embedding: A measure of embedding quality and embedding algorithms. *Machine Learning*, 77(1), 1–25. doi:10.1007/s10994-009-5107-9
- Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, 69(4), 211–221. doi:10.1007/s10708-007-9111-y

- Goodchild, M. F. (2013). The quality of big (geo)data. *Dialogues in Human Geography*, 3(3), 280–284. doi:10.1177/2043820613513392
- Gracia, A., González, S., Robles, V., & Menasalvas, E. (2014). A methodology to compare Dimensionality Reduction algorithms in terms of loss of quality. *Information Sciences*, 270, 1–27. doi:10.1016/j.ins.2014.02.068
- Graham, M. & Shelton, T. (2013). Geography and the future of big data, big data and the future of geography. *Dialogues in Human Geography*, 3(3), 255–261. doi:10.1177/2043820613513121
- Griffith, D. A. (1996). Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying geo-referenced data. *Canadian Geographer / Le Géographe canadien*, 40(4), 351–367. doi:10.1111/j.1541-0064.1996.tb00462.x
- Griffith, D. A. (2005). Effective Geographic Sample Size in the Presence of Spatial Autocorrelation. *Annals of the Association of American Geographers*, 95(4), 740–760. doi:10.1111/j.1467-8306.2005.00484.x
- Griffith, D. A. (2008). Spatial-filtering-based contributions to a critique of geographically weighted regression (GWR). *Environment and Planning A*, 40(11), 2751–2769.
- Griffith, D. A. (2013). Establishing Qualitative Geographic Sample Size in the Presence of Spatial Autocorrelation. *Annals of the Association of American Geographers*, 103(5), 1107–1122. doi:10.1080/00045608.2013.776884
- Harris, P., Brunsdon, C., & Charlton, M. (2011). Geographically weighted principal components analysis. *International Journal of Geographical Information Science*, 25(10), 1717–1736. doi:10.1080/13658816.2011.554838
- Harris, R., Sleight, P., & Webber, R. (2005). *Geodemographics, GIS and neighbourhood targeting*. John Wiley and Sons.
- Henriques, R., Bação, F., & Lobo, V. (2009). Carto-SOM: Cartogram creation using self-organizing maps. *International Journal of Geographical Information Science*, 23(4), 483–511. doi:10.1080/13658810801958885
- Henriques, R., Bação, F., & Lobo, V. (2012). Exploratory geospatial data analysis using the GeoSOM suite. *Computers, Environment and Urban Systems*, 36(3), 218–232. doi:10.1016/j.compenvurbsys.2011.11.003
- Henriques, R., Lobo, V., & Bação, F. (2012). Spatial Clustering Using Hierarchical SOM. *Applications of Self-Organizing Maps*. doi:10.5772/51159
- Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. *The Annals of Statistics*, 36(3), 1171–1220. doi:10.1214/009053607000000677
- Jenkins, O. C. & Matari, M. J. (2004). A spatio-temporal extension to Isomap nonlinear dimension reduction. In *Twenty-first international conference on Machine learning - ICML '04* (p. 56). Banff, Alberta, Canada: ACM Press. doi:10.1145/1015330.1015357
- Johnston, R. J. (1971). Some Limitations of Factorial Ecologies and Social Area Analysis. *Economic Geography*, 47(sup1), 314–323. doi:10.2307/143213

- Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography*, 3(3), 262–267. doi:10.1177/2043820613513388
- Knaap, E. (2017). The Cartography of Opportunity: Spatial Data Science for Equitable Urban Policy. *Housing Policy Debate*, 27(6), 913–940. doi:10.1080/10511482.2017.1331930
- Knaap, E., Wolf, L. J., & Rey, S. (2019). The Dynamics of Urban Neighborhoods: A Survey of Approaches for Modeling Socio-Spatial Structure. *Geography Compass*, Under Review.
- Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9), 1464–1480.
- Konig, A. (2000). Interactive visualization and analysis of hierarchical neural projections for data mining. *IEEE Transactions on Neural Networks*, 11(3), 615–624. doi:10.1109/72.846733
- Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1), 1–27. doi:10.1007/BF02289565
- Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29(2), 115–129. doi:10.1007/BF02289694
- Lee, J. A. & Verleysen, M. (2009). Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing. Advances in Machine Learning and Computational Intelligence*, 72(7), 1431–1443. doi:10.1016/j.neucom.2008.12.017
- Lee, J. A. & Verleysen, M. (2010). Scale-independent quality criteria for dimensionality reduction. *Pattern Recognition Letters*, 31(14), 2248–2257. doi:10.1016/j.patrec.2010.04.013
- LeSage, J. P. (2014). What Regional Scientists need to know about Spatial Econometrics. *The Review of Regional Studies*, 44(1), 13–32.
- LeSage, J. P. & Llano, C. (2013). A spatial interaction model with spatially structured origin and destination effects. *Journal of Geographical Systems*, 15(3), 265–289.
- LeSage, J. P. & Pace, R. K. (2009). *Introduction to Spatial Econometrics*. Boca Raton, FL: CRC Press.
- Lobo, V., Henriques, R., & Bação, F. (2012). Exploratory geospatial data analysis using the GeoSOM suite. *Computers, Environment and Urban Systems*, 36(3). doi:10.1016/j.compenvurbsys.2011.11.003
- Mason, G. A. & Jacobson, R. D. (2007). Fuzzy Geographically Weighted Clustering. In *Proceedings of the 9th international conference on geocomputation* (p. 7). Maynooth, Ireland.
- McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [cs, stat]*. arXiv: 1802.03426 [cs, stat]
- Mead, A. (1992). Review of the Development of Multidimensional Scaling Methods. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 41(1), 27–39. doi:10.2307/2348634
- Meng, D., Leung, Y., & Xu, Z. (2011). A new quality assessment criterion for nonlinear dimensionality reduction. *Neurocomputing*, 74(6), 941–948. doi:10.1016/j.neucom.2010.10.011
- Miller, H. J. & Goodchild, M. F. (2015). Data-driven geography. *GeoJournal*, 80(4), 449–461. doi:10.1007/s10708-014-9602-6
- Mokbel, B., Lueks, W., Gisbrecht, A., & Hammer, B. (2013). Visualizing the quality of dimensionality reduction. *Neurocomputing. Advances in Artificial Neural Networks, Machine Learning, and Computational Intelligence*, 112, 109–123. doi:10.1016/j.neucom.2012.11.046



- Moosavi, V. (2017). Contextual mapping: Visualization of high-dimensional spatial patterns in a single geo-map. *Computers, Environment and Urban Systems*, 61, 1–12. doi:10.1016/j.compenvurbsys.2016.08.005
- Murakami, D. & Griffith, D. A. (2015). Random effects specifications in eigenvector spatial filtering: A simulation study. *Journal of Geographical Systems*, 17(4), 311–331. doi:10.1007/s10109-015-0213-7
- Murakami, D., Lu, B., Harris, P., Brunsdon, C., Charlton, M., Nakaya, T., & Griffith, D. A. (2019). The Importance of Scale in Spatially Varying Coefficient Modeling. *Annals of the American Association of Geographers*, in press.
- Newton, H. J. (2002). A Conversation with Emanuel Parzen. *Statistical Science*, 17(3), 357–378. doi:10.1214/ss/1042727944
- O'Brien, D. T., Sampson, R. J., & Winship, C. (2015). Econometrics in the Age of Big Data: Measuring and Assessing “Broken Windows” Using Large-scale Administrative Records. *Sociological Methodology*, 45(1), 101–147. doi:10.1177/0081175015576601
- Openshaw, S. (1994). Neuroclassification of Spatial Data. In B. C. Hewitson & R. G. Crane (Eds.), *Neural Nets: Applications in Geography* (pp. 53–70). The GeoJournal Library. Dordrecht: Springer Netherlands. doi:10.1007/978-94-011-1122-5\_3
- Palm, R. & Caruso, D. (1972). Factor Labelling in Factorial Ecology. *Annals of the Association of American Geographers*, 62(1), 122–133. doi:10.1111/j.1467-8306.1972.tb00848.x
- Park, R. E. & Burgess, E. W. (1925). *The City: Suggestions for investigation of human behavior in the urban environment*. Chicago, IL: University of Chicago Press.
- Paul, R. & Chalup, S. K. (2017). A study on validating non-linear dimensionality reduction using persistent homology. *Pattern Recognition Letters*, 100, 160–166. doi:10.1016/j.patrec.2017.09.032
- Pearson, P. T. & Cooper, C. I. (2012). Using Self Organizing Maps to Analyze Demographics and Swing State Voting in the 2008 U.S. Presidential Election. In N. Mana, F. Schwenker, & E. Trentin (Eds.), *Artificial Neural Networks in Pattern Recognition* (pp. 201–212). Lecture Notes in Computer Science. Springer Berlin Heidelberg.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Poorthuis, A. & Zook, M. (2017). Making Big Data Small: Strategies to Expand Urban and Geographical Research Using Social Media. *Journal of Urban Technology*, 24(4), 115–135. doi:10.1080/10630732.2017.1335153
- Psyllidis, A., Yang, J., & Bozzon, A. (2018). Regionalization of Social Interactions and Points-of-Interest Location Prediction With Geosocial Data. *IEEE Access*, 6, 34334–34353. doi:10.1109/ACCESS.2018.2850062
- Raudenbush, S. & Sampson, R. (1999). Econometrics: Toward a Science of Assessing Ecological Settings, with Application to the Systematic Social Observation of Neighborhoods. *Sociological Methodology*, 29(1), 1–41. doi:10.1111/0081-1750.00059

- Rey, S. J. & Anselin, L. (2007). PySAL: A Python library of spatial analytical methods. *The Review of Regional Studies*, 37(1), 5–27.
- Sampson, R. (2012). *Great American city: Chicago and the enduring neighborhood effect*. University of Chicago Press.
- Sampson, R., Morenoff, J., & Gannon-Rowley, T. (2002). Assessing “neighborhood effects”: Social processes and new directions in research. *Annual review of sociology*, 443–478.
- Schölkopf, B. (2001). The Kernel Trick for Distances. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in Neural Information Processing Systems 13* (pp. 301–307). MIT Press.
- Shevky, E. & Bell, W. (1955). *Social area analysis; theory, illustrative application and computational procedures*. Social Area Analysis; Theory, Illustrative Application and Computational Procedures. Stanford University Press.
- Silva, J., Marques, J., & Lemos, J. (2006). Selecting Landmark Points for Sparse Manifold Learning. In Y. Weiss, B. Schölkopf, & J. C. Platt (Eds.), *Advances in Neural Information Processing Systems 18* (pp. 1241–1248). MIT Press.
- Singleton, A. & Arribas-Bel, D. (2019). Geographic Data Science. *Geographical Analysis*, in press. doi:10.1111/gean.12194
- Singleton, A., Pavlis, M., & Longley, P. A. (2016). The stability of geodemographic cluster assignments over an intercensal period. *Journal of Geographical Systems*, 18(2), 97–123. doi:10.1007/s10109-016-0226-x
- Singleton, A. & Spielman, S. E. (2014). The Past, Present, and Future of Geodemographic Research in the United States and United Kingdom. *Prof. Geogr.* 66(4), 558–567.
- Skupin, A. (2003). A Novel Map Projection Using an Artificial Neural Network. In *21st International Cartographic Conference* (pp. 1165–1172).
- Skupin, A. & Esperbé, A. (2011). An alternative map of the United States based on an n-dimensional model of geographic space. *Journal of Visual Languages & Computing*, 22(4), 290–304. doi:10.1016/j.jvlc.2011.03.004
- Skupin, A. & Fabrikant, S. I. (2003). Spatialization Methods: A Cartographic Research Agenda for Non-geographic Information Visualization. *Cartography and Geographic Information Science*, 30(2), 99–119. doi:10.1559/152304003100011081
- Spielman, S. E. & Logan, J. R. (2013). Using high-resolution population data to identify neighborhoods and establish their boundaries. *Annals of the Association of American Geographers*, 103(1), 67–84.
- Spielman, S. E. & Thill, J.-C. (2008). Social area analysis, data mining, and GIS. *Computers, Environment and Urban Systems*, 32(2), 110–122. doi:10.1016/j.compenvurbsys.2007.11.004
- Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500), 2319–2323. doi:10.1126/science.290.5500.2319
- Tobler, W. R. (1969). The spectrum of US 40. *Papers in Regional Science*, 23(1), 45–53.
- Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46, 234–240.

- Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*, 17(4), 401–419. doi:10.1007/BF02288916
- Trunk, G. (1968). Statistical estimation of the intrinsic dimensionality of data collections. *Information and Control*, 12(5), 508–525. doi:10.1016/S0019-9958(68)90591-3
- Ultsch, A. & Siemon, H. (1990). Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis. In *Proceedings of the International Neural Network Conference*. Paris: Kluwer.
- van der Maaten, L. (2009). Learning a Parametric Embedding by Preserving Local Structure. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics* (Vol. 5, pp. 384–391). Clearwater Beach, Florida, USA: Journal of Machine Learning Research.
- Verveer, P. J. & Duin, R. P. W. (1995). An evaluation of intrinsic dimensionality estimators. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(1), 81–86. doi:10.1109/34.368147
- Villmann, T., Der, R., & Martinetz, T. (1994). A new quantitative measure of topology preservation in Kohonen's feature maps. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)* (Vol. 2, 645–648 vol.2). doi:10.1109/ICNN.1994.374251
- Voas, D. & Williamson, P. (2001). The diversity of diversity: A critique of geodemographic classification. *Area*, 33(1), 63–76. doi:10.1111/1475-4762.00009
- Wattenberg, M., Viégas, F., & Johnson, I. (2016). How to Use t-SNE Effectively. *Distill*, 1(10), e2. doi:10.23915/distill.00002
- Williams, C. K. I. (2001). On a Connection between Kernel PCA and Metric Multidimensional Scaling. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in Neural Information Processing Systems 13* (pp. 675–681). MIT Press.
- Wolf, L. J., Germuska, J., Eisenbarth, G., & Folch, D. (2016). *Cenpy: A Package for Automatic Spatio-Temporal Data Discovery and Retrieval*.
- Wolf, L. J., Oshan, T. M., & Fotheringham, A. S. (2018). Single and Multiscale Models of Process Spatial Heterogeneity. *Geographical Analysis*, 50(3), 223–246. doi:10.1111/gean.12147
- Xu, X.-S., Liang, X.-L., Yang, G.-Q., Wang, X.-L., Guo, S., & Shi, Y. (2017). SOMH: A self-organizing map based topology preserving hashing method. *Neurocomputing*. Good Practices in Multimedia Modeling, 236, 56–64. doi:10.1016/j.neucom.2016.08.102
- Yuan, S., Tan, P.-N., Cheruvilil, K. S., Collins, S. M., & Sorzano, P. A. (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* (pp. 1–10). IEEE.
- Zhang, P., Ren, Y., & Zhang, B. (2012). A new embedding quality assessment method for manifold learning. *Neurocomputing*, 97, 251–266. doi:10.1016/j.neucom.2012.05.013