[AM-08-038] Pattern Recognition and Matching

Abstract

People recognize and characterize patterns to understand the world. Spatial data exhibit distinctive characteristics that render most aspatial recognition and matching methods unsuitable or inefficient. In past decades, a plethora of methods have been developed for spatial pattern recognition and matching to account for these spatial characteristics. This entry first focuses on the methods of spatial pattern recognition, including an overview of the basic concepts and common types. Methods for spatial pattern matching are then introduced. An example scenario of the distribution of tree species in the Arbuckle Mountains of south-central Oklahoma illustrates covered concepts. The entry concludes with brief remarks on continuing challenges and future directions in spatial pattern recognition and matching in the Big Data and artificial intelligence era.

Keywords: data mining, machine learning, pattern matching, pattern recognition, spatial analysis, spatiotemporal data mining

Author & citation

Cao, G. and Buttenfield, B. P. (2022). Pattern Recognition and Matching. The Geographic Information Science & Technology Body of Knowledge (2nd Quarter 2022 Edition). John P. Wilson (Ed.). DOI: 10.22224/gistbok/2022.2.10.

This Topic is also available in the following editions:

DiBiase, D., DeMers, M., Johnson, A., Kemp, K., Luck, A. T., Plewe, B., and Wentz, E. (2006). Pattern recognition. The Geographic Information Science & Technology Body of Knowledge. Washington, DC: Association of American Geographers.

Explanation

1. Definitions
2. Introduction
3. Spatial Pattern Recognition
4. Spatial Pattern Matching
5. An Illustrated Example
6. New Challenges and Future Directions

1. Definitions

Spatial pattern: the arrangement and the relationship among the spatial observations or spatial objects.

Spatial pattern recognition: automatic identification of spatial patterns or structures evident or implied in spatial data.

Spatial pattern matching: automatic identification of places with similar spatial patterns.

2. Introduction

If one picks up a map or looks around surrounding environments, one may notice the ubiquitous existence of spatial patterns. Landscape features don't occur randomly. Their arrangements depend upon landscape characteristics and may in turn affect nearby features. For example, lodgepole pine tends to grow well in acid soils and full sun; and as the pines grow, they create shade for certain types of ground cover and animal habitats. Patterns provide important clues about underlying processes that lead to the patterns and humans rely on examining the patterns to understand the processes. In the example above, underlying processes might include vegetative decomposition that leads to soil acidification, hydrological flow constrained by terrain elevation and slope, solar insolation, and habitat emergence or change. Human visual perception is very efficient in recognizing salient patterns in low dimensional (e.g., 2D and 3D) spaces, but often falls short for complex or latent patterns in high dimensional spaces (e.g., 3D spaces over time, or results of some statistical analyses). With the advances of computer science and technology, automatic pattern recognition and analysis have become important topics in scientific and engineering disciplines. In GIScience, spatial pattern analysis has emerged as important topics, as a rich body of theories, methods and tools have been developed in recent years.

Spatial pattern recognition and pattern matching are two common problems in spatial pattern analysis and modeling. Recognition involves detecting, identifying and characterizing how the values of spatial variables are distributed across a landscape. Because spatial pattern can relate to several underlying processes and multiple variables may be involved, pattern recognition might also address statistical associations and covariations among those variables. For example, recognizing patterns of cholera in a global south village could establish associations with population density, water sources from nearby streams, presence of contaminated food sources from crops or livestock, and proximity to periodic markets or to mobile health clinics. A variety of geospatial analytic tasks might be undertaken, for example the segmentation or classification of satellite imagery, hydrologic modeling of flow direction and accumulation in the watershed, and statistical delineation of hot-spot areas. Spatial pattern matching works from the characterization described above to locate and identify similar spatial patterns in other places.

Yuan (2001) provides an excellent example of pattern recognition and matching to understand severe storm events. Beginning with more than 8,000 layers of hourly radar images per year in Oklahoma, she established individual storm tracks based on location, theme and timestamps. Working with associated storm variables (storm rotation, barometric pressure and eyewall definition), she linked storm objects with the radar images. The pattern recognition allowed her to build storm sequences and define individual storm events. Once characterization was complete in her database, she undertook pattern matching using spatiotemporal reasoning and queries to match archived storms with newly acquired data to establish if a newer storm's characteristics were similar to earlier storms in terms of storm track, intensity, duration, etc. This example shows how recognition and matching operate in conjunction, and also points to the need for working with very large, even massive data sources. A third highlight of this example is that that both tasks (recognition and matching) can be extended into space-time settings to monitor changes among the patterns over time. Compared with aspatial data, geospatial data are spatially

indexed, and such information (e.g., location, proximity, scale, and spatial configuration) plays an explicit and critical role in pattern recognition and matching. Spatial data exhibit distinctive spatial characteristics, including scale sensitivity, spatial dependence, and spatial heterogeneity (Goodchild, 2004). These properties render many analytic methods originally developed for aspatial data unsuitable for spatial data.

For example, sensitivity to scale means that pattern recognition must anticipate the resolution at which evidence of a pattern is expected. Tobler's (1970) well- known First Law of Geography describes spatial dependence, the similarity or autocorrelation of spatial observations with nearby observations. It advises that conventional statistical methods developed under the assumption of independent observations might not work for spatial data. The less well-known Second Law of Geography describes spatial heterogeneity and the non-stationary nature of spatial data and processes (Anselin 1989), which highlights the importance of incorporating local conditions in searching for spatial patterns (Fotheringham et al. 2003, Fotheringham et al. 2017, Gelfand et al. 2003). It calls for increasingly large data samples to be collected to account for local variations not well- represented by regional or global summaries. A Third Law of Geography was proposed recently (Zhu et al., 2018) stating that similar geographic characteristics and configurations can be expected to hold similar attributes and to reflect similar spatial processes. For example, similar tree species can be expected to occur in similar soil types, climate regimes, and topography. Its intent is to extend the prediction of spatial patterns into broader geographic contexts including geographic configurations and spatial covariates, further prioritizing the need for ancillary data.

This entry will focus on the main concepts and overview the overall workflow and main categories of methods in spatial pattern recognition and matching. The remainder of the entry will introduce methods for spatial pattern recognition and then for spatial matching. Examples will be given to illustrate the concepts for both tasks. The entry will conclude with a brief discussion about ongoing challenges and future directions in the current Big Data and artificial intelligence era.

3. Spatial Pattern Recognition

Figure 1 shows the stages involved in a typical spatial pattern recognition workflow. Raw observations often need a data transformation before feeding into pattern recognition methods since the spatial patterns may not be prominent in the raw observations. Transformation projects the observations into a space so that latent patterns might become more readily distinguishable. A wide range of transformations can be applied, from logarithmic or polynomial transformations to more complex strategies (e.g., principal component analysis). The transformed raw observations are often referred to as features in machine learning literature. The stages of feature generation and selection, also known as feature engineering, can dramatically affect the recognition performance.

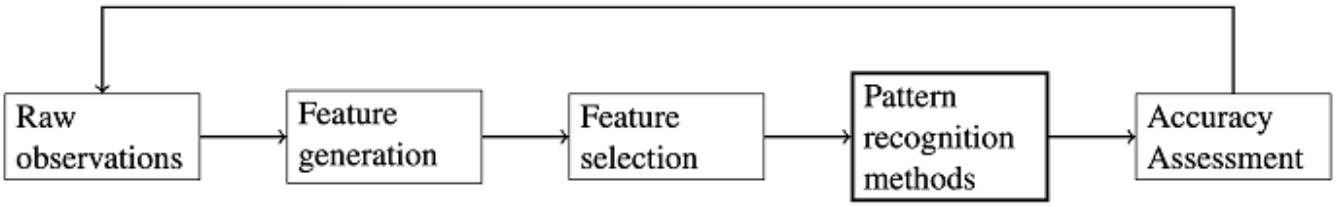

Figure 1. A typical workflow of spatial pattern recognition. Source: authors.

Common methods for spatial pattern recognition largely fall into two categories depending on the availability of training data or expert knowledge about the patterns of study. As described above, supervised methods learn data patterns and trends from the training data. Most supervised methods can be thought of as mathematical equations (rules, or knowledge) controlled by a set of parameters, and the process of learning seeks an optimal set of parameters that can best fit equations to the training data or experts' knowledge. Parameters can be optimized by minimizing the discrepancy between transformed features and estimations (e.g., error metrics) or maximizing agreement between them (e.g., maximum likelihood). One common problem with supervised learning is overfitting, which arises when a model corresponds too closely to a particular set of training data and cannot be generalized reliably to other observations. Hastie et al. (2009) describe a set of techniques to mitigate overfitting.

Compared with supervised methods, unsupervised methods do not require training. Without assuming an underlying structure or a set of groups, unsupervised methods unravel the underlying similarities of observations and arrange them into clusters. Similarity measures of spatial objects play an important role in such methods. Unsupervised methods are often used for exploratory spatial analysis to identify clusters or patterns that are worth further investigations. In common practice, unsupervised and supervised methods tend to complement each other: the former serve the purpose of exploring and identifying spatial patterns and the latter explain the patterns by incorporating auxiliary variables, relying on patterns elicited during training, or domain knowledge.

Spatial data can be characterized by geometry (points, lines, polygons) or by data models (e.g.,vector, raster or network). Additionally, dynamic information (time series or streaming video) has special properties that may complicate the recognition process, as for example in recognizing an object moving through a cluster of static objects (such as a fox running across a meadow into a grove of trees). Each type of spatial data has unique statistical characteristics that often require the development of specific pattern recognition methods. A multitude of frameworks and approaches have been developed for spatial pattern recognition and matching to address the distinctive spatial characteristics of geospatial data.

Table 1 shows the common types of spatial data types and related pattern recognition methods. Geostatistical data here refers to spatial measurements taken in sampled locations or spatial units; examples of such data include measurements of meteorological variables collected from a set of monitoring stations or a list of land cover types observed at sample locations. In Table 1, unsupervised methods are distinguished from supervised methods in terms of assumptions about the underlying structure in the data. Baxter (2006: 671) states "In unsupervised learning problems, the [goal] is often to identify previously unknown structure in the data. In supervised learning, the 'structure' (e.g., classes or

groups in the data) is assumed to be known at the outset and this 'knowledge' is used in the statistical analysis." Supervised knowledge is established by classifying a training data set that is sampled from the data in question, from expert knowledge, or by accessing a previously classified data bank whose contents are assumed to be similar to the data to be tested. Classes or groups discovered during training are then applied to the test data, thus predicting the classes or groups to be formed. Training data sets are normally very large and contain many times more items than the test data set: a training sample could involve up to 80% of the original sample, leaving only 20% for testing (Bishop, 2006).

Table 1. Typical unsupervised and supervised methods for pattern recognition related problems with common geospatial data types.

Data Types	Example Problem	Example Unsupervised Methods	Example Supervised Methods
Point data	Characterize and model the spatial distribution of point patterns (e.g., occurrences of crimes)	Kernel density estimation (Silverman, 1998)	Process-based Bayesian models (Diggle, 2014); MaxEnt (Phillips et al., 2006)
Areal data	Spatial clustering and modeling of areal maps (e.g., county maps of household income)	Local indicator of spatial association (Anselin 1995), hierarchical or K-means clustering (de Smithet al., 2018), self-organization maps (Hagenauer and Helbich, 2013)	Geographically weighted regression (Fotheringham et al. 2003), Bayesian spatial methods (Banerjee et al., 2004, Gelfand et al. 2003)
Geostatistical data	Characterize the spatial variability of measurements (e.g., air quality measurements from monitoring stations) and spatial interpolation	Triangulation, inverse distance weighted methods, spatial splines, variogram/covariogram (de Smith et al., 2018)	Kriging (Chiles and Delfiner, 1999; Cressie and Wikle 2011), Bayesian spatial methods (Banerjee et al., 2004)
Spatial grids or imagery	Satellite image segmentation (group similar neighboring pixels into groups) or classification (assign each group of pixels a label)	K-means-based methods, hierarchical clustering (de Smith et al., 2018)	Support vector machine, random forest (Hastie et al.,2009); deep learning methods (Ma et al., 2019)
Spatial videos or time series of imagery	Change detection using time series of satellite imagery or activity classification	Harmonic analysis (Zhu and Woodcock 2014), deep learning unsupervised methods (Ma et al., 2019)	Markov-based models (Liu and Cai, 2012), deep learning supervised methods (Ma et al., 2019)

The performance of supervised pattern recognition methods can be assessed by measuring the discrepancies between estimations and validations. In practice, validation data can be difficult or expensive to obtain. One common strategy is to randomly separate the available data items into subsets with one reserved for validation and the rest for training pattern recognition methods. The estimations from the trained methods are then compared with the validation subset for performance metrics. The procedure is called cross-validation when the process of subsetting and comparison is repeated a sufficient number of times to ensure that each data item serves in both training and validation subsets. Different performance metrics can be used to measure the discrepancy (or agreement) between the two. For continuous measurements (e.g., in the case of spatial interpolation), the performance can be assessed by error statistics such as root mean squared errors or mean

absolute errors. For categorical types of data (e.g., in the case of satellite imagery classification), the discrepancy can be summarized with a table, often referred to as a confusion matrix, indicating the number of data items that are misclassified, or with accuracy metrics (e.g., commission errors and omission errors). Performance metrics are also used for diagnosis and parameter tuning of pattern recognition methods or for performance comparison among different methods.

When ground truth is available, the performance of unsupervised methods can be evaluated by measuring the agreements between the result clusters and ground truth labels. In the case that ground truth is not available, two approaches are commonly used. The first is based on the tendency of the data to form clusters. Null hypothesis testing is often used: a null hypothesis of randomness is assumed to be true and the clustering tendency can then be formulated as the degree to which the data does not support this assumption (e.g., tested with a p-value). One issue with this approach is the difficulty in determining the statistical distribution of the assumed randomness, in which case statistical techniques such as Monte Carlo simulation and bootstrapping can approximate the distribution (e.g., Anselin, 1995). The second approach is based on the geometric parameters of the result clusters, such as the cohesion (or closeness) of the objects within clusters and the level of separation between clusters. Many evaluation metrics are available based on the premise that desirable clusters should have more cohesion and higher separation values (Palacio-Niño and Berzal, 2019).

4. Spatial Pattern Matching

Spatial pattern matching is an important topic for spatial analysis, spatial database evaluation and data mining. Spatial pattern recognition methods often operate hand-in-hand with pattern matching. Once patterns have been identified, spatial pattern matching can find other objects in a data archive that share similar patterns. Spatial pattern matching compares a given spatial pattern or patterns with test patterns. The central task is to develop effective similarity measures that can account for spatial characteristics. Putting it more intuitively, given a set of spatial objects or maps, spatial pattern matching quantifies or ranks the similarities between them.

As with spatial pattern recognition, the diverse types of geospatial data sources can warrant different matching objectives. For example, the comparison of boundaries or trajectories among spatial objects (e.g., water bodies or soil polygons) may require more emphasis on shape or geometry, while the comparison among metric grids (e.g., elevation or temperature) may be more concerned with local magnitudes or focal differences. A rich set of matching methods and tools have been developed to account for such heterogeneity, and these can be categorized into statistics-based methods or distance-based methods.

For example, patterns can be described by summary statistics, or binned as histograms, or modeled by probability distribution functions (PDFs). Pattern similarity can hence be derived by quantifying the differences between histograms or PDFs. Specific statistics have been developed to describe and compare spatial patterns. For point pattern data, for example, statistics tools such as K-function or kernel density estimation can be applied. Similarly for geostatistical data, variograms or covariograms describe changes in spatial measurements over distance. Datasets with similarly shaped variograms probably share

similar variation trends as the distance between measurements increase.

Distance-based metrics have been adopted to measure the differences among spatial objects while accounting for spatial characteristics. For example, the similarity among travel trajectories can be measured either by Fréchet distance, developed specifically to measure the similarity between geometrical curves, or by edit distances, originally developed to measure similarity between strings of words (Yuan and Raubal, 2014). Another metric is Earth Mover Distance (EMD) (Rubner et al., 2000) that frames the differences between two images as a transportation problem that accounts for spatial configuration. To understand EMD, consider two satellite images. Pixels with higher reflectance values can be considered metaphorically as suppliers of goods and corresponding pixels with lower values as consumers. For each pair of supplier and consumer pixels, the costs needed to move one unit of goods corresponds to the distance (i.e., difference in reflectance) between them. The EMD establishes the "cost" of equalizing reflectance between the two images.. Additional information if available can also be leveraged into similarity measures. In satellite imagery analysis, for example, reflectance values of multiple spectral bands can be considered for more reliable measures when comparing two images.

5. An Illustrated Example

An example of categorizing tree species can demonstrate the concepts discussed so far. Figure 2(a) shows a set of locations of trees of different types (elm, black oak, post oak and others) in the Arbuckle Mountains of south-central Oklahoma. Data are taken from the Public Land Survey, which is often used to study the landscapes in the pre- and early-European settled Midwestern and Western United States (Cao et al., 2014). The scenario problem is to characterize the spatial distribution of the tree species and to estimate the probability of tree species occurring at unsampled locations. This scenario provides an example of a geostatistical problem in Table 1 with species represented as categorical attributes.

As mentioned earlier, spatial phenomena tend to demonstrate spatial dependence, which if investigated carefully can provide important insights for spatial prediction. Variograms or covariograms can help explore and quantify the structures of spatial autocorrelation. Figure 2(b) shows covariograms for each tree type with the horizontal axis indicating the distances between occurrences and the vertical axis showing the covariance (similarity) between them. One can clearly see that all tree types share the same decreasing trend of similarity as distance increases, with the rate of decrease fastest for post oak and slowest for elm. Beyond roughly 800 meters, the similarity patterns for black oak and elm are quite similar, while post oak distribution seems to approach the pattern shown by other tree types.

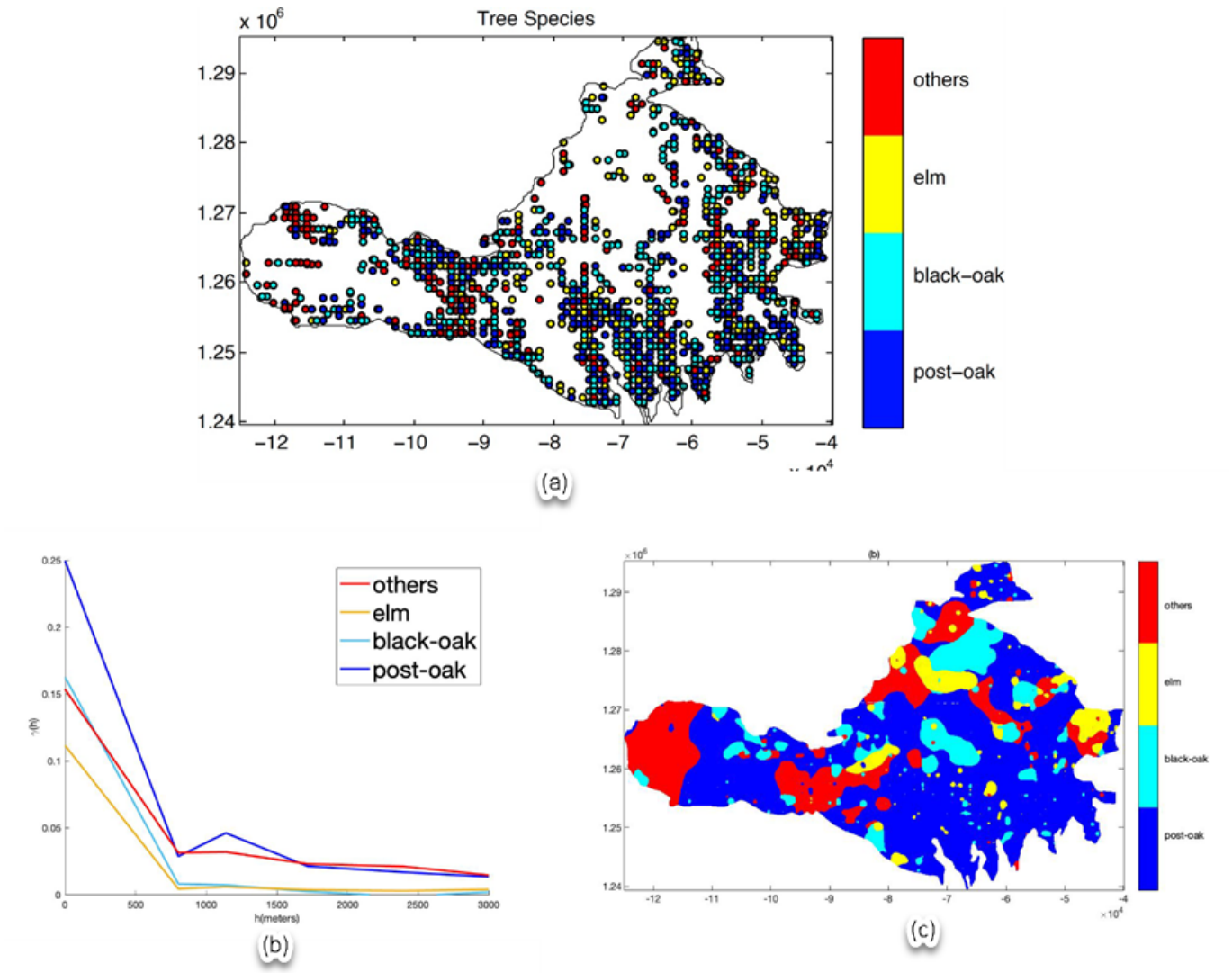

Figure 2 (a) Survey locations of the three most abundant tree species: post oak, black oak, and elm; (b) Covariogram of the sampled tree species; (c) Result prediction of a recent supervised method (Cao et al. 2014). Source: authors.

Either unsupervised or supervised methods can estimate the probability of tree species occurring at unsampled locations. An unsupervised method is inverse distance weighted interpolation, which computes the probability at any unsampled location as a weighted combination of neighboring locations. The method is unsupervised because no training process is needed. The weight of each neighbor is simply based on its reciprocal distance to estimated locations.

Typical supervised spatial methods include kriging methods (Chiles and Delfiner, 1999) and Bayesian methods (Banerjee et al., 2004). With a training process, these methods can learn the weights to best fit a training data set. Ancillary environmental conditions such as soil types, surface rock types and elevation that are closely related to tree species can also be incorporated. Figure 2(c) shows the result map of a recently developed kriging method (Cao et al., 2014) that integrates spatial covariates as well as spatial autocorrelation of tree species. Cross-validation was used for performance assessment, and the method achieved 10% higher successful classification rates (75.6% over 65.7%) than methods that ignore

spatial autocorrelation.

6. New Challenges and Future Directions

In recent years, the landscape of geospatial science and technology has shifted dramatically, marked by the advent of the Big Data revolution, advances in computing resources, and emergence of machine intelligence. This shift raises new challenges and provides new opportunities for spatial pattern recognition and matching. On one hand, geospatial data with fine spatiotemporal scales become increasingly available, which makes it possible to examine natural and social environments in unprecedented detail. However, geospatial data often demonstrates a large amount of heterogeneity, which tends to fragment environmental patterns. The data heterogeneity, along with inconsistent data scales, changing data boundaries, uneven data quality and a variety of data types, complicating pattern recognition and matching. This highlights the need for methods that can effectively reconcile highly variable data types, properties and heterogeneity to achieve more realistic spatial pattern analysis and modeling. 305On the other hand, a spatial pattern can become very complex when it involves an increasing number of spatiotemporal locations or objects. The recognition and modeling of complex spatial patterns is one of the most fundamental and yet challenging tasks in GIScience. Traditional spatial analysis methods often fall short when dealing with complex spatial patterns. The advances of machine learning methods, particularly the recent breakthrough in deep learning and artificial intelligence, dramatically improves the state-of-art in pattern recognition applications.With a deep neural network with multiple levels of processing layers, deep learning-based methods have been shown to excel at discovering intricate patterns from high-dimensional data.GIScientists are on the frontiers of the development and application of new deep learning methods for spatial pattern analysis and modeling. Collaboration and interdisciplinary efforts are leading to the emergence of the new area of GeoAI (Janowicz et al., 2019). Most recent developments however are limited to specific types of data (e.g., remote sensing imagery or spatial grids). More efforts are necessary to exploit the full power of the deep learning paradigm or spatiotemporal pattern recognition.

References

Anselin, L. (1989). What is Special About Spatial Data? Alternative Perspectives on Spatial Data Analysis (89-4). UC Santa Barbara: National Center for Geographic Information and Analysis.

Anselin, L. (1995). Local Indicators of Spatial Association -- LISA. Geographical Analysis, 27(2):93-115.

Baxter, M. J. (2006). A Review of Supervised and Unsupervised Pattern Recognition in Archaeometry. Archaeometry, 48(4): 671–694.

Cao, G., Yoo, E.-H., and Wang, S. (2014). A statistical framework of data fusion for spatial prediction of categorical variables. Stochastic Environmental Research and Risk Assessment, 28(7):1785–1799.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2003). Geographically Weighted

Regression: The Analysis of Spatially Varying Relationships. John Wiley & Sons.

Fotheringham, A. S., Yang, W., & Kang, W. (2017). Multiscale geographically weighted regression (MGWR). Annals of the American Association of Geographers, 107(6), 1247-1265.

Gelfand, A. E., Kim, H. J., Sirmans, C. F., & Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. Journal of the American Statistical Association, 98(462), 387-396.

Goodchild, M. F. (2004). The Validity and Usefulness of Laws in Geographic Information Science and Geography. Annals of the Association of American Geographers, 94(2):300–303.

Hagenauer, J. & Helbich, M. (2013). Hierarchical self-organizing maps for clustering spatiotemporal data. International Journal of Geographical Information Science, 27 (10),2026-2.

Janowicz, K., Gao, S., McKenzie, G., Hu, Y., & Bhaduri, B. (2020). GeoAI: Spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. International Journal of Geographical Information Science, 0(0), 1-13.

Liu, D. and Cai, S. (2012). A spatial-temporal modeling approach to reconstructing land-cover change trajectories from multi-temporal satellite imagery. Annals of the Association of American Geographers,102(6):1329–1347.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., and Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. ISPRS Journal of Photogrammetry and Remote Sensing, 152, 166-177.

Palacio-Niño, J.-O., and Berzal, F. (2019). Evaluation Metrics for Unsupervised Learning Algorithms. arXiv.

Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The Earth Mover's Distance as a Metric for Image Retrieval. International Journal of Computer Vision, 40(2):99–121.

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. Economic Geography 46: 234–240.

Yuan, M. (2001). Representing Complex Geographic Phenomena in GIS. Cartography and Geographic Information Science 28(2): 83-96.

Yuan, Y. and Raubal, M. (2014). Measuring similarity of mobile phone user trajectories: A Spatio-temporal Edit Distance method. International Journal of Geographical Information Science, 28(3):496–520.

Zhu, A. X., Lu, G., Liu, J., Qin, C. J., and Zhou, C. (2018). Spatial prediction based on the Third Law of Geography. Annals of GIS, 24(4):225–240.

Zhu, Z. and Woodcock, C. E. (2014). Continuous change detection and classification of land cover using all available Landsat data. Remote Sensing of Environment, 144:152–171.