Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 Keller Hall
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 18-013

An Introduction to Spatial Data Mining

Jamal Golmohammadi, Yiqun Xie, Jayant Gupta, Yan Li, Jiannan Cai, Samantha Detor, Abigail Roh, Shashi Shekhar

August 8, 2018

An Introduction to Spatial Data Mining

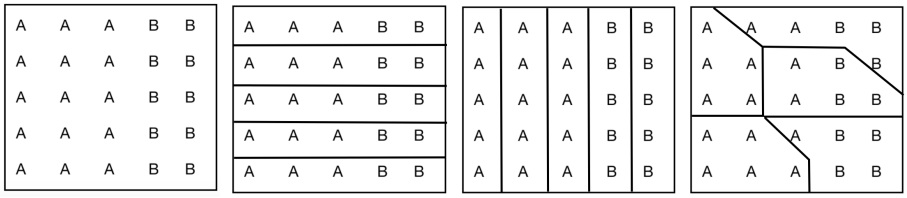
Jamal Golmohammadi, Yiqun Xie, Jayant Gupta, Yan Li,
Jiannan Cai, Samantha Detor, Abigail Roh, Shashi Shekhar
Computer Science and Engineering
University of Minnesota, Twin Cities
{golmo002, xiexx347, gupta423, lixx4266, cai00084, shekhar}@umn.edu
{detosa, rohab}@student.breckschool.org

Summary Abstract:The goal of spatial data mining is to discover potentially useful, interesting, and non-trivial patterns from spatial datasets. Spatial data mining is important for societal applications in public health, public safety, agriculture, environmental science, climate etc. For example,in epidemiology, spatial data mining helps to find areas with a high concentrations of disease incidents to manage disease outbreaks. Computerized methods are needed to discover spatial patterns since the volume and velocity of spatial data exceeds the number of human experts available to analyze it.In addition, spatial data has unique characteristics like spatial autocorrelation and spatial heterogeneity which violate the i.i.d (Independent and Identically Distributed data samples) assumption of traditional statistics and data mining methods. So, using traditional methods may miss patterns or may yield spurious patterns which are costly (e.g. ,stigmatization) in spatial applications. Also, there are other intrinsic challenges such as MAUP (Modifiable Areal Unit Problem) as illustrated by a current court case debating gerrymandering in elections. Spatial data mining considers the unique characteristics, and challenges of spatial data and domain knowledge of the target application to discover more accurate and interesting patterns.In this article, we discuss tools and computational methods of spatial data mining, focusing on the primary spatial pattern families: hotspot detection, colocation detection, spatial prediction and spatial outlier detection. Hotspot detection methods use domain information to model accurately more active and high density areas. Colocation detection methods find objects whose instances are in proximity of each other in a location. Spatial prediction approaches explicitly model neighborhood relationship of locations to predict target variables from input features. The goal of spatial outlier detection methods is to find data that are different from their neighbors.

KeyWords: Spatial Data Mining, Spatial Statistics, Spatial Patterns, Hotspot Detection, Collocation Detection, Teleconnection Discovery, Spatial Prediction, Spatial Outlier Detection, Spatial Autocorrelation, MAUP.

1 Definitions

- **Spatial data:** Any data that includes location information such as street address, (longitude, latitude), etc.



(a) Data (b) 5A - 0B (c) 3A - 2B (d) 2A - 3B

Figure 1: Example of Gerrymandering. (a)Base data; (b)Horizontal partitioning, A takes all seats, 5A - 0B; (c)Vertical partitioning, 3A - 2B; (d)Partitioning helping minority B get majority of seats, 2A - 3B.

- **i.i.d assumption:** Classical methods assume that data samples are assumed to be independent of each other and identically distributed.
- **Spatial autocorrelation**: Spatial Autocorrelation measures dependency among neighborhood points in spatial data. Dependency among neighborhood of spatial data rejects the independence assumption in spatial data.
- **Spatial heterogeneity**: Spatial heterogeneity refers to variation in events, features and relationships across a region. It violates identical distribution assumption.
- **Spatial statistics:** Spatial statistics is a generalization of traditional statistics for spatial data. For example, it models spatial dependency and heterogeneity.
- **Spatial data mining:** Spatial data mining is a generalization of traditional data mining, a field exploring trade-offs between computational scalability and statistical rigor, for spatial data. For example, it models spatial dependency and heterogeneity.
- **Hotspots** : Hotspot areas are the ones that are more active and have unusually high density.
- **Spatial outlier**: Data points that are different from their spatial neighbors.

2 Introduction

There is an explosive growth in location-aware data such as GPS tracks of mobile phones and remotely sensed satellite imagery. Availability of the data provides opportunities to discover new interesting, non-trivial spatial patterns from the data that are useful. For example, by having access to data of disease incidents, we can find areas with high concentration of the disease which may help manage disease outbreak. As shown in Table 1, spatial patterns are used in many organizations to decide reasonable actions and policies. For example in ecology and environmental management, scientists classify remote sensing images to maps of land-cover classes. In public safety, discovery of crime hotspots

Domain	Spatial data mining application
Public safety	Discovery of hotspot patterns from crime event maps
Epidemiology	Detection of disease outbreak.
Business	Allocating market stores to maximize profit
Neuroscience	Discovering patterns of human brain activity from neuroimages
Climate science	Finding positive or negative correlation between temperatures of distant places

Table 1: Example of application domains of spatial data mining

events may help to assign police resources efficiently. Also, in climate science, finding effects of distant locations on temperature of a location can lead to more accurate estimation of temperature.

Spatial data is different from traditional data because it has location information such as longitude, latitude, elevation etc. Computation on such data can reveal implicit spatial relationship between objects or features in the dataset. For example, location information for two objects can be used to compute the distance between them. The spatial relationships between objects is a vital and rich source of information which at the very least can enhance feature selection for improving the performance of traditional methods. As an example, considering distance between objects or features in traditional methods of data mining can improve performance of the methods.

Traditional data mining and machine learning may miss patterns or may yield spurious patterns which have high cost (e.g., stigmatization). This is because of nature of spatial data which violates assumptions that are common for methods of traditional data mining and machine learning. One of the assumptions is called i.i.d assumptions which presumes data samples are independent of each other and are from identical distribution. Because of spatial autocorrelation and heterogeneity of spatial data the i.i.d assumption usually is not valid in spatial domains. The sensitivity of statistical methods to space partitioning and non-stationarity of spatial data along time are other important characteristics and challenges of spatial data [9].

Spatial data mining is the process of discovering non-trivial, interesting and previously unknown, but potentially useful patterns from large spatial and spatio-temporal databases [14, 16, 18, 5]. It considers the unique characteristics,challenges and domain knowledge of spatial data. The first step of the spatial data mining process is often pre-processing of the data to correct noise, error, and missing data and explanatory space-time analysis to underlying spatial or spatiotemporal distribution. In the second step, a relevant spatial data mining lgorithm is applied to the prepro-cessed data to produce an output pattern. Well known patterns are hotspot areas, colocations, predictive models, and spatial outliers. In the final step, after post-processing of the output, domain scientists understand and explain the post-processed output to find novel insights. Some times refining the data mining algorithm is needed based on the results of the last step.

The aim of this article is to explain spatial data mining, its differences with traditional data mining, and its spatial pattern families. We do not discuss spatial statistics and related mathematics in detail. Also, spatial data mining is widely applied to many disciplines(e.g., remote sensing, geography) and domains(e.g., public health, landscape architecture, urban studies.) which is

Category	Example
Input data	Implicit relationship
Statistical foundation	Spatial autocorrelation, spatial heterogeneity
Output	Hotspot, colocation, spatial outlier

Table 2: What is special about mining spatial data?

beyond the scope of this article. A brief overview of special elements of spatial data mining is presented in Table 2. First row of Table 2 mentions implicit relationship of spatial data which is input of spatial data mining algorithms. Second row of the table shows two examples of statistical foundation of spatial data mining that are spatial autocorrelation and heterogeneity. In the last row of table 2 examples of output of spatial data mining methods are presented.

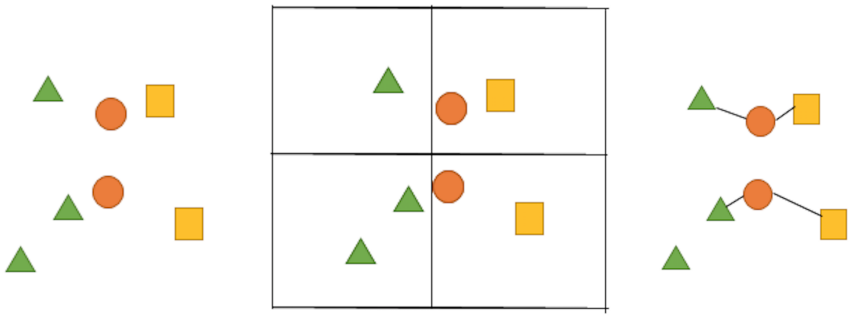
The article is organized as follow. First, We will discuss spatial statistics in section 3. Then we will explain each pattern family, its related applications, and statistical methods in section 4. In section 5, future research and trends are presented. Learning objectives and instructional assessment questions are in sections 6 and 7. Also, we provide more resources for further reading about the topic in section 8.

3 Spatial Statistics

It is important to notice that spatial statistics [2, 4] is different from traditional statistics because of the unique properties of space and time. One of the common assumptions in traditional statistics is the i.i.d. assumption. The i.i.d assumption is the foundation of major of data science and machine learning methods and statistics theorems. It is the basis for well-known methods such as maximum likelihood estimation and central limit theorem. Dependency of spatial data is a well known fact which is considered as first law of geography, "Everything is related to everything else, but nearby things are more related than distant things". First law of geometry indicates spatial autocorrelation between spatial data and violates the i.i.d assumption.

In addition to its unique characteristics, spatial data mining has other critical challenges. Result of traditional methods of statistics are sensitive to space partitioning. Formally, this is called the Modifiable Areal Unit Problem (MAUP). It is also referred as the multi-scale effect. As an example, the results of an analysis can be different when aggregation of data is in different levels of states, county or family level.Gerrymandering of election districts is an example of a MAUP. In this case a particular party or group tries to take political advantage by redrawing boundaries of districts. Consider the example in Figure **??** with 15 A and 10 B to be partitioned into 5 congenial districts. Figures 1b and 1c shows that horizontal districts give all 5 seats to A, while vertical districts give party A a slim 3-2 majority. Figure 1d shows another distract which allows party B to get majority of 3 seats, despite losing the overall popular vote 10 to 15.

In Figure 2a , there are three types of points, squares (□), circles (◯) and triangles (△). Each point type has two instances. From figure 2a, it is clear that each circle point has a square and triangle point neighboring it. For calculating



(a) (b) (c)

Figure 2: Example of spatial statistics. (a)Distribution of different points; (b)Region partitioning; (c)Neighborhood relationship based on neighborhood graph

Pairs	**Pearson's Correlation**	**Support**	**Ripley's Cross K**	**Participation Index**
◯, △	-0.9	0	0.33	0.67
◯, □	1	0.5	0.5	1

Table 3: Pearson's correlation coefficient and participation index for two event pairs

the spatial correlation between the different points we should partition the space. Figure 2b shows a partitioning of the study area. The count of each point type in each partition can be a feature of the point type for its spatial distribution. Let's order the partitions as: top-left, bottom-left, top-right, and bottom-right. Thus, triangles have features [1,2,0,0], circles have features [0, 0, 1, 1] and squares have features [0,0,1,1].

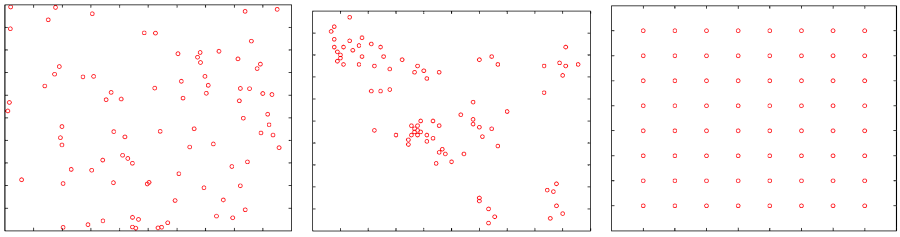
Table 1. shows Pearson's correlations between circles and triangles and circles and squares. The results contradict the observation in Figure 2a, because the correlation between triangle and circle is negative but correlation between square and circle is positive. The partitioning in Figure 2b causes the spatial relationship between circles and triangles to be lost. By contrast, Figure 2c shows that a participation index (Table 3) is able to accurately capture the adjacency. As shown in this example, using traditional space partitioning may cause the loss of spatial information. So, choosing a proper spatial model is critically important in spatial data mining.

Spatial statistics [17] methods can be categorized based on the type of spatial data they are used on: Geostatistics for point referenced data, lattice statistics for areal data, spatial point process for spatial point patterns, and graph-based spatial network statistics for network data, as shown in table 2.

In a spatial point process, we are concerned with the locations of points, especially their distributions [12]. Different statistical assumptions can be used to generate the location of a set of points. A homogeneous Poisson distribution includes complete spatial randomness (CSR) which is often used as a null hypothesis. The points in CSR are identically and independently distributed which follow Poisson distribution(see Figure 3a). There are two other assump-

Spatial Model	Spatial Statistics
Points	Spatial Point Process, Geostatistics
Field Model	Lattice statistics
Graph	Spatial network statistics

Table 4: Taxonomy of spatial statistics



(a) (b) (c)

Figure 3: Example of different statistical assumptions for generating the location of a set of points. (a)Complete spatial randomness; (b)Clustered; (c)De-clustered [3].

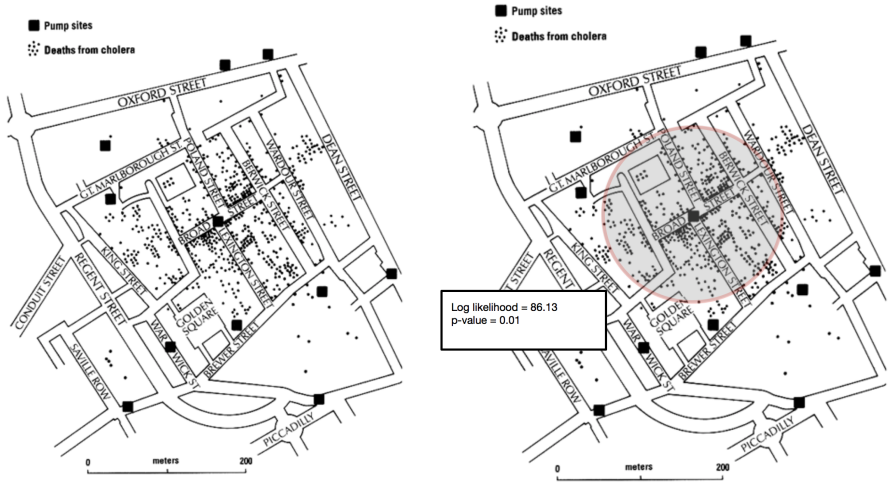
tions for generating the location of a set of points, clustered (Figure 3b) and de-clustered(Figure 3c). A cluster assumption presumes that points can be clustered to different regions while a de-cluster assumption presumes that there is not any clustering between points. Spatial statistics such as Ripley's K function, i.e., the average number of points within a certain distance of a given point over the total average intensity, can be used to compare a point pattern against CSR.

4 Spatial Pattern Families

This section describes four spatial pattern families [15]and the methods to detect them. The described pattern families are : hotspot detection, collocation detection , spatial prediction and spatial outlier detection.

4.1 Hotspot Detection

Given a set of geospatial points which are related to an activity in a domain, hotspots are the areas that are more active and have more density compared to other areas. A famous historical example of hotspot detection is the story of Dr.Snow, who collected and plotted the location of cholera infection on a map of London. He found that the highest incidence of disease was in proximity of Broad street water pump (see Figure 4a). From this he was able to hypothesized that cholera is spread by water. Today, hotspot patterns [13] can be found in domains such as epidemiology, criminology, transportation safety, ecology, environmental science, business, medicine an urban planning. It is important to note that the notion of a hotspot is domain specific and hotspot detection



(a) (b)

Figure 4: Analysis of water pump sites and deaths from cholera in London in 1854. (a) pump sites and deaths; (b)output of spatial statistical test.

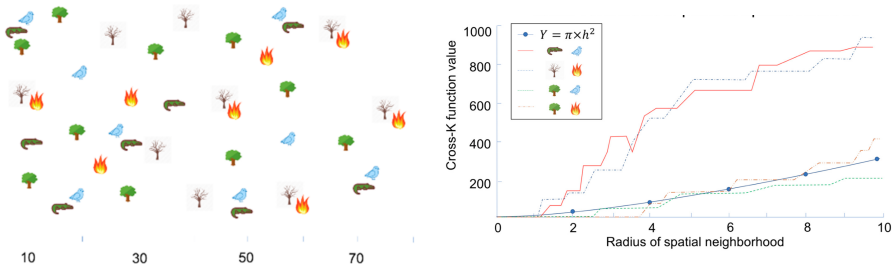
methods should consider domain knowledge to model hotspot areas correctly and effectively. For example, disease hotspots are typically modeled as circular areas. Circles are the best options in this case because epidemiologists base their models of disease spread on diffusion theory

SatScan [10] is a method to find statistically significant hotspots from spatial data. SatScan executes hypothesis testing for candidate hotspots discovered by cylinder form scanning of the space. The null hypothesis is based on CSR. In other words, it declares randomness of activity points in each cylinder. The alternative hypothesis claims that activities inside the cylander are more dense compared to outside. A Loglikelihood ratio is calculated for each candidate cylander and significance value (i.e. p-value) is the evaluation for the candidate with the highest likelihood ratio(see figure 4b).

4.2 Colocation Detection

Spatial colocation patterns [11] represent subsets of features whose their instances are located near one another. Such patterns are important in applications related to ecology, environmental science, public safety and climate science. Symbiotic relationships, for example between the Nile Crocodile and Egyptian Plover, exhibit colocation patterns. Other biological dependencies also have colocation patterns like different types of blackberry canes. Figure 5a shows the spatial distribution of instances of five features, namely, plover, crocodile, green trees, dry trees, and wild fire, in a sample dataset input into a collocation detection algorithm. In another example, identifying colocations from crime datasets can help police departments to understand more about crime patterns and hidden organization behind them.

Ripley's K function (section 3) evaluates clustering assumption on a point distribution by measuring degree of clustering. It is based on an average of



(a) (b)

Figure 5: Example of detection of colocation pattern. (a) Sample data; (b)Cross-K function of pairs of the four features [18].

points whose distance is smaller than a predefined threshold from any points. The null hypothesis of Ripley's K is also based on CSR. The cross-K function extends Ripley's K function to cases when there are multiple features. It is a spatial statistical method to detect collocation patterns between features of point events. The cross-K function between two features i and j with predefined distance h, is the ratio of the expected number of feature type i with smaller distance than h from a random feature j, to the density of feature type j (number per unit area).

Figure 5b shows the cross-K function results for the input represented in Figure 5a. As can be seen, crocodile and plover have high cross-K values which means they are more likely to locate near each other. The low value between green tree and wild fire means that these two are usually located far from each other.

Participation index is an upper bound of the cross-K function, it is a popular measure of colocation due to its computational properties [6]. The index uses a participation ratio, which is another measure for colocation detection. The participation ratio of feature f_1 in a colocation pattern CP, $pr(CP, f_1)$ is the portion of feature f_1 engaging in the pattern CP. Participation index is defined as $pi(CP) = min_{f_i \in CP} pr(CP, f_i)$ that is the minimum participation ration of all features engaging in the collocation pattern. Table 3 shows the participation index values for the collocation pattern in Figure 2a. One pattern is $(\bigcirc, \triangle)$. The $pr((\bigcirc, \triangle), \bigcirc)$ is 1 because all circles are participating in colocation pattern $(\bigcirc, \triangle)$. Also, two triangles are engaging in colocation pattern $(\bigcirc, \triangle)$ which means $pr((\bigcirc, \triangle), \triangle) = \frac{2}{3} \approx 0.67$. So, $pi(\bigcirc, \triangle) = 0.67$, which is the minimum value of the participation ratio of engaged features in the colocation pattern.

4.3 Spatial Prediction

Spatial prediction models have two kind of variables for the data items, explanatory variables (also called explanatory attributes or features) and a dependent variable (also called a target variable). Also, training samples of data is provided for spatial prediction model. The goal of a spatial prediction problem is to predict the value of dependent variables from explanatory variables by using training samples of data and the neighborhood relationships among the locations in data. When the dependent variable is discrete, the problem is called

spatial classification. When dependent variables are continuous, the problem is a spatial regression. Spatial prediction is widely used in climate science and environmental science to predict land cover types using remote sensing imagery .

Autocorrelation between samples of data and heterogeneity of data are two challenges of spatial prediction problems that violate i.i.d assumption for spatial data. [7]. For example, in Figure 6b, a decision tree is used to classify wet land and dry land using spectral features from a satellite image shown in Figure 6a . Compared to the ground truth in Figure 6c, the output of the decision tree contains a large amount of salt-and-pepper error due to inadequate consideration of spatial autocorrelation and heterogeneity [7].

Spatial properties are explicitly considered in spatial regression to handle spatial autocorrelation and heterogeneity in data. The Spatial Auto-Regressive (SAR) model belongs to the family of spatial regression models and it uses the spatial relationship between explanatory features to predict target variables. A neighborhood relationship is necessary for modeling spatial relationship of explanatory features and it is usually an additional input of SAR. The SAR model is defined as follows:

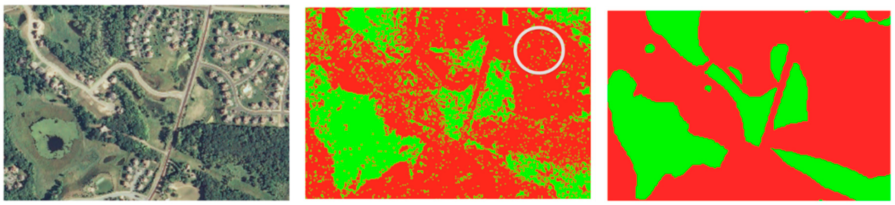
$$y = \rho W y + X\beta + \epsilon \tag{1}$$

W is the adjacency matrix in formula 1, and Wy models effect of neighborhood in addition to the effects of selected features X on the target variable y. ρ and β are the parameters that should be learned in formula 1. Notice that linear regression, which follows the i.i.d assumption, is a special case of the SAR model when ρ is zero.So, the SAR model is a more general model than linear regression model.

Geographically Weighted Regression (GWR) is another established method for modeling spatial autocorrelation. GWR does not do a regression on all data samples. It uses a search window centered at the current location and selects only a section of samples that are in the search window. Samples that are closer to the current location in the search window will get more weights.

4.4 Spatial Outlier Detection

Outliers may be gloaba or spatial. Global outliers are data samples which are inconsistent with the rest of data samples, or which deviate so much from other data that we think they were generated by a different mechanism (or distribution). Briefly, for finding global outliers we consider the whole data set for the comparison. In contrast, spatial outliers differ from other data only in their neighborhood [1, 8]. So, they do not need to be inconsistent with all data samples. For example, a new house surrounded by older houses in a developed city can be considered as a spatial outlier, but it may not be a global outlier based on the overall age of houses in the city. As another example, Figure 7 shows the 1992 United States president election results (grey vs. black) for all states. Indiana is the spatial outlier in this example. Spatial outlier detection is vital for applications that need to find unusual or suspicious activity or objects compared to their neighborhoods. Such applications include anomalous traffic monitoring in transportation engineering, fraud detection and prediction in credit card transactions and suspicious object and behavior detection in criminology.



(a) Input (b) Decision tree output (c) Output

Figure 6: Spatial classification problem. (a) input high-resolution aerial imagery; (b) decision tree prediction with salt-and-pepper errors highlighted in white circle; (c) map of ground truth: red is dry land, green is wetland

.

It is important to notice that spatial information of neighborhood points of data is applied in most of modern prediction methods. For example, CNN which is the state of art method in some of application, uses a spatial filter to consider the spatial information. Alos, spatial decision tree and Bayesian spatial binary classification are two other examples of prediction methods that incorporate the spatial information.

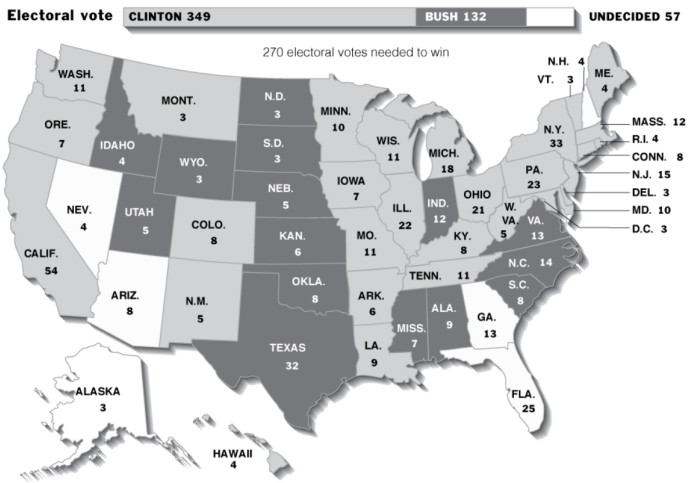


Figure 7: 1992 United States president election results (grey vs. black). Indiana is a spatial outlier. (Source: New York Times)

There are two categories of statistical tests for detection of spatial outliers, graphical tests and quantitative tests. Graphical tests detect outliers by considering visualized patterns from data. Variogram clouds and Moran scatterplots belong to Graphical tests. Quantitative tests calculate the difference between non-spatial attributes of inspected points and their spatial neighbors. When the difference is larger than a predefined threshold, an outlier is detected. Neighborhood spatial statistics and scatterplots belong to quantitative tests.

5 Future Research

Most research in spatial data mining assumes space is Euclidean and isometric(i.e., has the same statistical properties along different directions) and neighborhood are symmetric. However, in many applications, network space is the basic space which usually is not isotropic nor symmetric.For example, road networks and river networks can be modeled by network space more effectively. Considering network structure is one of the challenges of using network space and it provides more accurate insight.

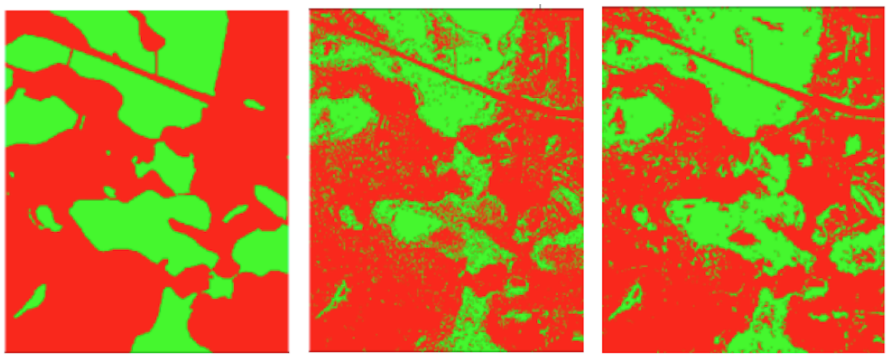
In addition to space dimension, time dimension is another important aspect of spatial data. Useful information and patterns can often be discovered by adding a time dimension to spatial data mining models. Detection of the time point that some phenomenon changed is an important problem which is called change detection. For example, change detection helps to detect when climate change happened in an area so that appropriate protective action in the area can be taken. In teleconnection discovery problem we have collection of spatial time series of different location. Teleconnection discovery aims to find pairs of positively or negatively correlated points of time series in great distance. Teleconnection discovery is used in climate science to predict more correctly temperatures of different places in the world. Adding a time dimension to spatial data mining will likely new and more complex statistical, mathematical and computational models.

Domain knowledge provides a rich source of information to enhance data-driven spatial models. Simulation models usually integrate physical rules and related domain knowledge into the data mining models to gain new and useful insights [9]. Simulation models are usually complicated from a computational perspective. So, new data science methods are needed that implement fast approximate solutions of simulation models are necessary.

As mentioned before, the cost of spurious patterns in societal applications is high(e.g. crime pattern analysis, alarm of disease epidemics). That's why, proposing methods that are statistically robust is vitally important for spatial data mining.

6 Learning Objectives

1. Explain i.i.d assumption and illustrate why it is not valid for spatial data.
2. Describe the following two key concepts in spatial statistics:
    - spatial autocorrelation
    - spatial heterogeneity



(a) (b) (c)

Figure 8

3. Define MAUP and explain gerrymandering as an example of MAUP
4. List three areas of spatial statistics and briefly explain them.
5. Name five spatial patterns and illustrate them.

7 Instructional Assessment Questions

1. Which statement(s) violate the independence assumption?
   (a) "Mohamed Lee" is a rare name, even though "Mohamed" is the most frequent first name and "Lee" is the most frequent last name.
   (b) Near things are more related than distant things.
   (c) Nearby video frames often show common people and objects.
   (d) All of the above
2. Which of three images in Figure 8 exhibits highest spatial autocorrelation?
   (a) image 8a
   (b) image 8b
   (c) image 8c
3. Which statement(s) violate the identical distribution assumption underlying traditional methods?
   (a) Cancer cell heterogeneity makes treatment of cancer difficult.
   (b) No two places on the Earth are exactly alike.
   (c) All politics is local.
   (d) All of the above
4. Which of the following are properties of spatial data?
   (a) Autocorrelation

(b) Heterogeneity
(c) Implicit relationships (e.g., neighbor)
(d) All of the above

5. Which of the following is not attributed to spatial auto-correlation?
   (a) Nearby cities have similar climate
   (b) Neighboring areas tend to plant similar farm crops
   (c) Near things are more related than distant things
   (d) Spatial data mining results are less reliable near the edges of a study area
6. Which of the following is correct about gerrymandering:
   - It is about redrawing boundaries of districts.
   - It can help a party or group to take a political advantage
   - It can change the result of an election in a way which contradicts the popular votes.
   - All of the above
7. Categorize following into hotspots, spatial outlier, colocation, location prediction:
   (a) Which countries are very different from their neighbors?
   (b) Which highway-segments have abnormally high accident rates ?
   (c) Where will a hurricane that's brewing over the ocean make landfall?
   (d) Which retail-store-types often co-locate in shopping malls?
8. Which does not illustrate a spatial-hotspot pattern family?
   (a) Roads with an unusually high rate of traffic accidents
   (b) Areas with an unusually high concentration of museums
   (c) Cities with unusually high numbers of students enrolled in this MOOC
   (d) A neighborhood with an unusually high rate of an infectious disease (or crime)
9. Which does not illustrate colocation?
   (a) A loud sound temporally follows a bright flash of lightning.
   (b) Nuclear power plants are usually located near water
   (c) Egyptian plover birds live close to Nile crocodiles
   (d) College campuses often have bookstores nearby.
10. Which of the following is false about spatial outliers?
    (a) An oasis (isolated area of vegetation) is a spatial outlier area in a desert
    (b) They may detect discontinuities and abrupt changes
    (c) They are significantly different from their spatial neighbors
    (d) They are significantly different from the population as a whole.

8 Additional Resources

- What is special about spatial data mining:http://www-users.cs.umn.edu/~shekhar/talk/2018/sdm_5_9_2018_small.pdf
- A sequence of 8 short presentations: https://www.youtube.com/playlist?list=PLN5UPhO05nn8WE4ZbzUwUhzq_p2XChK6r
- Encyclopedia of GIS [19]: It has many articles about each of mentioned topics in this paper. The book is available in thousands of institutions subscribing springer. Also, many of them are available on google books. Some of related topics in the books are listed in the following:
  1. Change detection.
  2. Colocation pattern
  3. Colocation mining
  4. Crime mapping
  5. Data mining
  6. Evolving spatial patterns
  7. Facility location problem
  8. Geostatistics
  9. Hotspot
  10. Hotspot detection and prioritization
  11. MAUP
  12. Outliter detection, spatial
  13. Partitioning
  14. Remote sensing
  15. Spatial anomaly detection
  16. Spatial big data
  17. Spatial data mining
  18. Spatial decision tree
  19. Spatial network
  20. Spatial prediction
  21. Spatial statistical analysis
- ArcGIS software: https://www.arcgis.com/

9 Acknowledgements

This article is supported by National Science Foundation under Grant No.1541876, 1029711, 1737633, IIS-1320580 , IIS-0940818, and IIS-1218168 , the USDOD under Grants No.HM1582-08-1-0017 and HM0210-13-1-0005, the Advanced Research ProjectsAgency-Energy (ARPA-E), U.S. Department of Energy under Award No.DE-AR0000795, the NIH under Grant No. UL1 TR002494, KL2TR002492,

and TL1 TR002493, the USDA under Grant No.2017-51181-27222, and the OVPR Infrastructure Investment Initiative, Minnesota Supercomputing Institute (MSI), and Provost's Grand Challenges Exploratory Research and International Enhancements Grants at the University of Minnesota. The views and opinions of authors expressed herein do not necessarily state or reflect those ofthe United States Government or any agency thereof. Also, we appreciate Kim Koffolt's helpful comments and feedbacks for enhancing readability of the paper.

References

[1] Charu C Aggarwal. "Outlier analysis". In: *Data mining*. Springer. 2015, pp. 237–263.

[2] Noel Cressie. *Statistics for spatial data*. John Wiley & Sons, 2015.

[3] Emre Eftelioglu et al. "Geospatial Data Science: A Transdisciplinary Approach". In: *Geospatial Data Science Techniques and Applications*. CRC Press, 2017, pp. 17–56.

[4] Alan E Gelfand et al. *Handbook of spatial statistics*. CRC press, 2010.

[5] Jiawei Han and Harvey J Miller. *Geographic data mining and knowledge discovery*. CRC Press, 2009.

[6] Yan Huang, Shashi Shekhar, and Hui Xiong. "Discovering colocation patterns from spatial data sets: a general approach". In: *IEEE Transactions on Knowledge and Data Engineering* 16.12 (2004), pp. 1472–1485.

[7] Zhe Jiang et al. "Focal-test-based spatial decision tree learning". In: *IEEE Transactions on Knowledge and Data Engineering* 27.6 (2015), pp. 1547–1559.

[8] James M Kang et al. "Discovering flow anomalies: a SWEET approach". In: *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE. 2008, pp. 851–856.

[9] Anuj Karpatne et al. "Theory-guided data science: A new paradigm for scientific discovery from data". In: *IEEE Transactions on Knowledge and Data Engineering* 29.10 (2017), pp. 2318–2331.

[10] Martin Kulldorff. *SaTScanTM user guide, www.satscan.org*.

[11] Pradeep Mohan et al. "Cascading spatio-temporal pattern discovery". In: *IEEE Transactions on Knowledge and Data Engineering* 24.11 (2012), pp. 1977–1992.

[12] Jesper Møller and Rasmus P Waagepetersen. "Modern statistics for spatial point processes". In: *Scandinavian Journal of Statistics* 34.4 (2007), pp. 643–684.

[13] Dev Oliver et al. "A k-main routes approach to spatial network activity summarization". In: *IEEE transactions on knowledge and data engineering* 26.6 (2014), pp. 1464–1478.

[14] Shashi Shekhar, Steven K Feiner, and Walid G Aref. "Spatial computing". In: *Communications of the ACM* 59.1 (2015), pp. 72–81.

[15] Shashi Shekhar et al. "Identifying patterns in spatial information: A survey of methods". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1.3 (2011), pp. 193–214.

[16] Shashi Shekhar et al. "Spatiotemporal data mining: a computational perspective". In: *ISPRS International Journal of Geo-Information* 4.4 (2015), pp. 2306–2338.

[17] Lance A Waller and Carol A Gotway. *Applied spatial statistics for public health data*. Vol. 368. John Wiley & Sons, 2004.

[18] Yiqun Xie et al. "Transdisciplinary Foundations of Geospatial Data Science". In: *ISPRS International Journal of Geo-Information* 6.12 (2017), p. 395.

[19] Hui Xiong, Xun Zhou, and Shashi Shekhar. *Encyclopedia of GIS*. Springer, 2017.