Statistical Methods for Spatial Data Analysis

AARUSHI SETHI[1]
[1]National Institute of Standards and Technology, Boulder, CO 80305 USA (e-mail: author@boulder.nist.gov)

ABSTRACT Spatial data analysis has gained a lot of popularity over the last few decades with more devices and techniques being invented to collect spatial data. This data is of the utmost importance as it is being used to study and forecast upcoming trends and patterns. This paper discusses a few of the statistical techniques that are used to study spatial data.

KEYWORDS Spatial data; statistics; autocorrelation.

I. INTRODUCTION

With the fast changing world it is imperative that the topological, geometric, geographic and other spatial properties and features are studied thoroughly in order to develop an adept understanding of past, present and forthcoming trends. Spatial data analysis is the manipulation and study of data that is manifested in the spare. Spatial data encompasses data such as that related to location, area, distance, interaction etc. An example of spatial data analysis is the development of a smart transportation grid using sensors. In this grid spatial data like distance between vehicles, the length and area of a vehicle, its interaction with other vehicles and other parameters are organized, analyzed and modeled upon to develop and smart and safe transportation grid or environment Spatial data analysis is a broad term and one needs to streamline the area under study and the approach being used to utilize the data. Broadly speaking, spatial data analysis can be bifurcated into two categories - data driven [1], [2] and model driven [3] . As the name suggests, for the data driven approach, data drives the research. In this type of analytical research there are no preconceived theoretical hypotheses, rather the first step involves the collection of spatial data. Once the data is collected and analyzed, attempts are made to derive patterns, correlations, associations and provide structure to it. The second type, that is the model driven approach, deals with data analysis in just the opposite manner. First a model assumption or hypothesis is proposed and then confronted with relevant data. This hypothesis may be rejected or accepted based on how the data responds when related experiments are performed. In this paper, we have discussed the first approach - data driven approach and some statistical methods that are used to perform analysis and deduction on the data. Statistics for spatial data was earlier used to organize data into comprehensible patterns. However, with the advent of machine learning and AI, it has been developed into a domain of its own. The spatial data collected is not simply used to generate trends, but also used to develop regression models and time series forecasting among other developments using spatial data. In this paper, we have discussed the statistical methods that are used to analyze spatial data and trends.

II. LITERATURE REVIEW

Spatial Data Analysis (SDA) is an important topic of interest; due to this, researchers are working to develop new techniques and theories for Spatial data analysis. In this section, we will analyze some of the recent developments in the field of spatial data analysis. Also, there are many cutting-edge technologies that are useful in spatial data analysis such as big data analysis [4], [5], soft computing [5], [6] and cloud computing [7]–[9]. Author in [10] presents the SDA technique to understand carbon emission at Chian. Author in [11] presetnes a study of malnutrition in children using SDA. Author in [12] presents the spatiotemporal evolution pattern of urban resilience in the Yangtze River Delta urban agglomeration based on TOPSIS-PSO-ELM. Authors in [13], [14] analysis different deep learning techniques . Author in [15] presents a compressive sensing of medical images with confidentially homomorphic aggregations. Author in [16] presents Bayesian negative binomial regression with SDA. Author in [17] presents spatio-temporal evolution and factor explanatory power analysis of urban resilience in the Yangtze River Economic Belt. Author in [18] uses SDAto conduct larval survey of the dengue-endemic area in Samarinda. Author in [19] use SDA and estimation of spatial econometric models.Author in [20] used the location Gini coefficient and exploratory spatial data analysis to study the spatial agglomeration characteristics of grain yield. Authors in [21] analysis attack detection technique in VANET. Authors in [22] conducted exploratory spatial data analysis, with countries as the unit of analysis. Authors in [23] gives the details about cloud computing security. Authors in [24] monitors spatiotemporal characteristics of land-use carbon emissions and their driving mechanisms in the Yellow River

Delta.

III. TYPES OF DATA IN STATISTICS

Before we move on to the statistical methods and models, it is important to understand the nature of the data that we are concerned with when we use a certain statistical method. Consider a spatial process being denoted in a dimension D. The attributes are denoted by Z which will be observed in a plane P with dimension D. D is the domain for the attributes Z. The continuity and discontinuity of D and Z are separate entities which may or may not depend on each other. The types of data that concern statistical analysis are described below.

A. GEOSTATISTICAL DATA

For geostatistical data [25], [26], the domain D for a given problem is continuous. This means that Z(s) can be observed at any point $s \in D$ and infinite points can be observed between two given points s_i and s_j . This does not however define the nature of the attributes Z being observed in domain D. The continuity or discontinuity of Z depends on the nature of Z and not the nature of domain D. Since the nature of the domain is continuous, it cannot be sampled for an infinite time. Here sampling and mapping techniques come into play which help in division of continuous domain into smaller chunks or regions.

B. LATTICE DATA

Lattice data [27] is a type of data which is measured in discrete yet already defined regions or areas. There is no particular restriction on the number of data samples, however they must be predefined. These regions are more commonly known as sites. Generally, we can use the notations used for geostatistical data for lattice data as well; however since the discrete samples used for defining lattice data span a region or area of the coordinate plane P, notation such as Z(A) is also common. In order to define spatial correlations between samples, characteristics which are common to all the sites are used. For example, in order to find distance between two areas in a defined map, s1 and s2 their respective centroids can be used to calculate the data. Again, in this type of data division the nature of the domain is fixed and non-stochastic and the type of the attributes does not depend on the domain but the nature of the data being collected.

C. POINT PATTERNS

Lattice data and geostatistical data have cursory differences. However, point pattern data differs from the rest of the two types at a fundamental level. For lattice and geostatistical data, the domain D(s) is fixed and unchanging and the attributes Z(s) are of more concern. Point pattern data [28] on the other hand deals with a changing domain, which is generally the primary focus of the experiments. Consider an experiment with the domain D. For an event a

$$I(s) = 1 for a = True$$
$$I(s) = 0 for a = False$$

A new domain D is formed which is a subset of D. D depends on a series of events and is not predetermined. In this type of data analysis the domain is generally the concern however it is not to say that the attributes are not important. The analysis of such experiments heavily rely on condition based models where the occurrence of a subset of the domain further affects the data.

IV. STATISTICAL MODELS

As discussed above there are two types of approaches for spatial data analysis i.e. data driven and model driven approach. This paper deals with the data driven approach and the methods used to conduct it. Data driven approach mostly consists of Exploratory Spatial Data Analysis or ESDA [29]. Exploratory Spatial Data Analysis engrids the non-spatial tools, global and local spatial autocorrelation and spatial heterogeneity. Some of these tools and methods have been discussed below.

A. NON-SPATIAL TOOLS

1) Choropleth Map

A choropleth map [30] is nothing but a visual representation of the distribution of the variables at hand. Many regard this as the first step of ESDA as it lays down an initial baseline division of the data. A choropleth map representation does not convey the spatial relation and effects on the variables. Other tools like weighted matrices need to be used to establish the correlations. Choropleth maps divide an event space, generally a geographical area, into smaller regions or divisions to provide data distribution per region. These smaller divisions may be regular or irregular. The range of variables in a particular division is generally thematically represented using colour codes or variations.

2) Visualizing outliers and extreme values

Outliers are points in a dataset that have values that stray too far from the rest of the data, or in other words, outliers are the extreme values in a dataset. Visualizing and analyzing outlier points is absolutely essential to any statistical analysis. Firstly, knowing the extreme values of a dataset helps in determining the range of the data. Then, many statistical models do not respond well to outliers, especially the distance based models. In such a case removing the outliers is necessary. Lastly, visualizing outliers can inform about some anomalous or misinformed data which needs to be re-evaluated or discarded. The most appreciated method to visualize outliers is visualizing a box plot [31], [32]. In a box plot, an outlier is a value above or below a given multiple of the difference between the first and third quartile. For a box plot, a distribution that is adjusted to standard deviation is observed and plotted. The values that go beyond the range

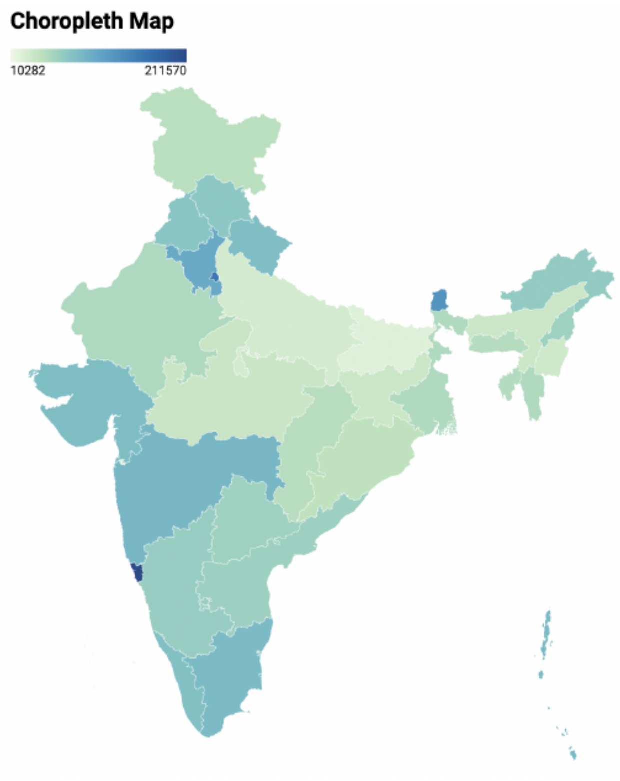

FIGURE 1: Choropleth Map for income statistics of India in 2021

of $[m - 2\sigma, m + 2\sigma]$ are considered outliers. The extreme values can also be observed by plotting a histogram for the distribution.

B. SPATIAL WEIGHT MATRIX

A spatial weight matrix [33], [34] is a matrix whose entities represent the relationship between spatial features of a given data. A spatial weight matrix converts the structured spatial data into numerical data. To establish a criteria for interaction of spatial features, it is important to know the nature of the features. On the basis of the nature of the features the method of conceptualization should be given preference. Generally, there are two fundamental branches of creating weights to define the relationship between features - binary or variable weighting. As the name suggests, the cells in the binary weight matrix consist of 0s and 1s. Some of the binary weight strategies [35] are fixed distance, K nearest neighbors, Delaunay Triangulation, contiguity, or space-time window. Variable weighting on the other hand computes weights based on the importance of one spatial feature with respect to another. Inverse distance or zone of indifference strategies are some of the methods used to calculate weighted matrices.

C. GLOBAL SPATIAL AUTOCORRELATION

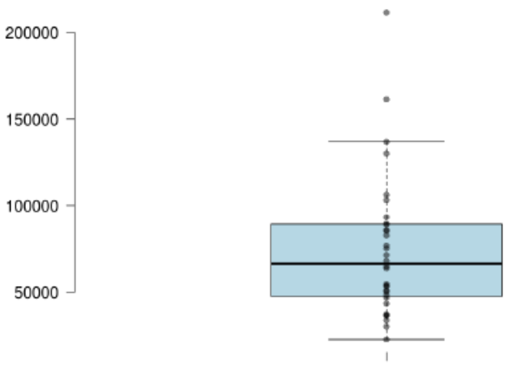

FIGURE 2: Box Plot for income statistics of India in 2021

1) Join Count Statistics

The simplest form of spatial autocorrelation is join count statistics [36], [37] because it uses binary labels to categorize quantitative values. At a nominal or ordinal level presence or absence of a characteristic or a range of features determines whether the join corresponds to a similar or different value from the neighboring values to determine the spatial relation. The quantitative notation for this method is rendered to a binary variable namely- presence or absence or thematically speaking, black or white. The spatial dependencies of data can be expressed in terms of the types of distribution i.e. grouped, random or dispersed with these types representing positive, neutral and negative dependency respectively.

2) Global Moran's I and Geary's C

The most reliable and widely used statistics to measure global autocorrelation are Global Moran's I and Geary's C [38]. In contrast to binary join count statistics, they measure the linear correlation between a variable at one point in a division in the coordinate system or domain and to the weighted average of other divisions. The tests for I and C statistics are based on Z statistics [39] and related distributions. Between I and C statistics, Moran's I statistics attest more to global spatial relations while C statistics are more relevant for local spatial relations. Moran's I can be described as

$$I = \frac{N}{S_o}\frac{\sum_i \sum_i w_{ij}(x_i - \overline{x})(x_j - \overline{x})}{\sum_i (x_i - \overline{x})^2}$$

Geary's C is given by

$$C = \frac{N-1}{S}\frac{\sum_i \sum_i w_{ij}(x_i - \overline{x})^2}{\sum_i (x - \overline{x})^2}$$

where N is the number of observations, w_{ij} is the degree of connection between the spatial units i and j, and x_i is the variable of interest in region i and $S_o = \sum_i \sum_i w_{ij}$

D. LOCAL SPATIAL AUTOCORRELATION

1) Moran Scatterplot

Moran scatterplot [40] is an important statistical tool for local spatial autocorrelation as it enables us to observe the relation and similarity of an observation to its local neighbor. The horizontal or X axis of a Moran's scatterplot represents the observation values. The Y axis or vertical axis represents the weighted average or the quantitative values in relation to the local spatial correlation to the observational values of the X axis. A Moran's scatterplot is divided into four quadrants about the origin. The upper right and lower left quadrants signify a positive correlation between values whereas the upper left and lower right quadrants indicate a negative correlation.

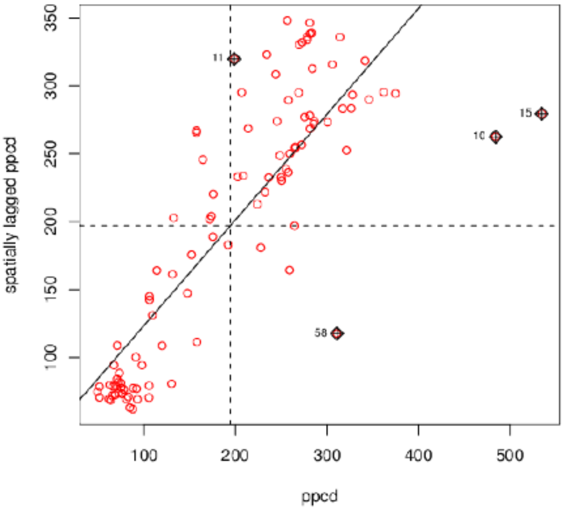

FIGURE 3: Moran's scatterplot [41]

E. OTHER INDICATORS

Major statistical methods and tools have been described above. Apart from these aforementioned approaches other notable spatial data analysis methods are - Getis–Ord statistics [42], LISA statistics, Bonferroni pseudo-significance level etc. Getis–Ord statistics or G statistics are used to calculate how high and low the local autocorrelation exists for the value of interest and its neighboring values. This is done using z-statistic and p-statistics. LISA statistics are used to determine the circle of clustering of similar values around a given observational data.

V. FURTHER DEVELOPMENTS AND CONCLUSION

Having empirical and statistical methods to study spatial data and its relation with other observational values have significantly eased and advanced the understanding of various geo-spatial phenomena. However, a lot of work still needs to be put into organizing the methodologies and statistics used between multiple disciplines. Many softwares [43] that calculate these matrices have been introduced e.g. SAGE for ESDA, DynESDA, ArcGIS Geostatistical Analyst etc. These softwares ease the calculation of the statistical analysis models however on a fundamental level there is still need for more interdisciplinary methods that not only take in account the geographical attributes, but also other socio-economic parameters like population, economics, politics etc.

REFERENCES

[1] P. Sprent, Data driven statistical methods. Routledge, 2019.

[2] E. Bartocci, L. Bortolussi, and G. Sanguinetti, "Data-driven statistical learning of temporal logic properties," in International conference on formal modeling and analysis of timed systems. Springer, 2014, pp. 23–37.

[3] J.-H. Yoo, M. S. Nixon, and C. J. Harris, "Model-driven statistical analysis of human gait motion," in Proceedings. International Conference on Image Processing, vol. 1. IEEE, 2002, pp. I–I.

[4] B. B. Gupta and et al, "Soft computing techniques for big data and cloud computing," Soft Computing, vol. 24, no. 8, pp. 5483–5484, 2020.

[5] B. B. Gupta, D. P. Agrawal, S. Yamaguchi, and M. Sheng, "Advances in applying soft computing techniques for big data and cloud computing," pp. 7679–7683, 2018.

[6] M. H. Bhatti and et al., "Soft computing-based eeg classification by optimal feature selection and neural networks," IEEE Transactions on Industrial Informatics, vol. 15, no. 10, pp. 5747–5754, 2019.

[7] K. Bhushan and et al., "Security challenges in cloud computing: state-of-art," International Journal of Big Data Intelligence, vol. 4, no. 2, pp. 81–107, 2017.

[8] C. L. Stergiou and et al., "Secure machine learning scenario from big data in cloud computing via internet of things network," in Handbook of computer networks and cyber security. Springer, 2020, pp. 525–554.

[9] P. Negi and et al., "Enhanced cbf packet filtering method to detect ddos attack in cloud computing environment," arXiv preprint arXiv:1304.7073, 2013.

[10] K. Zhou, J. Yang, T. Yang, and T. Ding, "Spatial and temporal evolution characteristics and spillover effects of china's regional carbon emissions," Journal of Environmental Management, vol. 325, 2023.

[11] B. González Cusi and O. Solano Dávila, "Study of malnutrition in children under five in peru using the exploratory spatial data analysis," Smart Innovation, Systems and Technologies, vol. 207 SIST, pp. 322–333, 2023.

[12] X. Chenhong and Z. Guofang, "The spatiotemporal evolution pattern of urban resilience in the yangtze river delta urban agglomeration based on topsis-pso-elm," Sustainable Cities and Society, vol. 87, 2022.

[13] M. Hammad and et al., "Myocardial infarction detection based on deep neural network on imbalanced data," Multimedia Systems, vol. 28, no. 4, pp. 1373–1385, 2022.

[14] J. Peng and et al., "A biometric cryptosystem scheme based on random projection and neural network," Soft Computing, vol. 25, no. 11, pp. 7657–7670, 2021.

[15] L. Wang and et al., "Compressive sensing of medical images with confidentially homomorphic aggregations," IEEE Internet of Things Journal, vol. 6, no. 2, pp. 1402–1409, 2018.

[16] F. Mutiso, J. Pearce, S. Benjamin-Neelon, N. Mueller, H. Li, and B. Neelon, "Bayesian negative binomial regression with spatially varying dispersion: Modeling covid-19 incidence in georgia," Spatial Statistics, vol. 52, 2022.

[17] C. Ye, M. Hu, L. Lu, Q. Dong, and M. Gu, "Spatio-temporal evolution and factor explanatory power analysis of urban resilience in the yangtze river economic belt," Geography and Sustainability, vol. 3, no. 4, pp. 299–311, 2022.

[18] M. Ridha and S. Sulasmi, "Larval survey of the dengue-endemic area in samarinda: guide to determine risk containers," International Journal of Public Health Science, vol. 11, no. 4, pp. 1176–1183, 2022.

[19] J. Martori, R. Lagonigro, and R. Iglesias-Pascual, "Social status and air quality in barcelona: A socio-ecological approach," Sustainable Cities and Society, vol. 87, 2022.

[20] H. He, R. Ding, and X. Tian, "Spatiotemporal characteristics and influencing factors of grain yield at the county level in shandong province, china," Scientific Reports, vol. 12, no. 1, 2022.

[21] A. Gaurav and et al., "Ddos attack detection in vehicular ad-hoc network (vanet) for 5g networks," in Security and Privacy Preserving for IoT and 5G Networks. Springer, 2022, pp. 263–278.

[22] B. Olakunde, J. Pharr, D. Adeyinka, L.-C. Chien, R. Benfield, and F. Sy, "Spatial variations in family planning demand to limit childbearing and the demand satisfied with modern methods in sub-saharan africa," Reproductive Health, vol. 19, no. 1, 2022.

[23] S. Gupta and et al., “Hunting for dom-based xss vulnerabilities in mobile cloud-based online social network,” Future Generation Computer Systems, vol. 79, pp. 319–336, 2018.

[24] Y. Yang and H. Li, “Monitoring spatiotemporal characteristics of land-use carbon emissions and their driving mechanisms in the yellow river delta: A grid-scale analysis,” Environmental Research, vol. 214, 2022.

[25] P. J. Ribeiro Jr and P. J. Diggle, “Analysis of geostatistical data,” The geoR package, version, pp. 1–6, 2006.

[26] D. Allard and G. Guillot, “Clustering geostatistical data,” in Proceedings of the sixth geostatistical conference, 2000.

[27] S. R. Sain and N. Cressie, “A spatial model for multivariate lattice data,” Journal of Econometrics, vol. 140, no. 1, pp. 226–259, 2007.

[28] R. M. Haralick, H. Joo, C.-N. Lee, X. Zhuang, V. G. Vaidya, and M. B. Kim, “Pose estimation from corresponding point data,” IEEE Transactions on Systems, Man, and Cybernetics, vol. 19, no. 6, pp. 1426–1446, 1989.

[29] R. Haining, S. Wise, and J. Ma, “Exploratory spatial data analysis,” Journal of the Royal Statistical Society: Series D (The Statistician), vol. 47, no. 3, pp. 457–469, 1998.

[30] G. Andrienko, N. Andrienko, and A. Savinov, “Choropleth maps: classification revisited,” in Proceedings ica, 2001, pp. 1209–1219.

[31] K. Potter, H. Hagen, A. Kerren, and P. Dannenmann, “Methods for presenting statistical information: The box plot.” in VLUDS, 2006, pp. 97–106.

[32] Y. Benjamini, “Opening the box of a boxplot,” The American Statistician, vol. 42, no. 4, pp. 257–262, 1988.

[33] X. Qu and L.-f. Lee, “Estimating a spatial autoregressive model with an endogenous spatial weight matrix,” Journal of Econometrics, vol. 184, no. 2, pp. 209–232, 2015.

[34] C. Lam and P. C. Souza, “Estimation and selection of spatial weight matrix in a spatial lag model,” Journal of Business & Economic Statistics, vol. 38, no. 3, pp. 693–710, 2020.

[35] J. Too, A. R. Abdullah, and N. Mohd Saad, “A new co-evolution binary particle swarm optimization with multiple inertia weight strategy for feature selection,” in Informatics, vol. 6, no. 2. MDPI, 2019, p. 21.

[36] L. Anselin and X. Li, “Operational local join count statistics for cluster detection,” Journal of Geographical Systems, vol. 21, no. 2, pp. 189–210, 2019.

[37] S. Virdee, W. C. Tan, J. C. Hogg, J. Bourbeau, C. J. Hague, J. A. Leipsic, and M. Kirby, “Spatial dependence of ct emphysema in chronic obstructive pulmonary disease quantified by using join-count statistics,” Radiology, vol. 301, no. 3, pp. 702–709, 2021.

[38] J. Ping, C. Green, R. Zartman, and K. Bronson, “Exploring spatial dependence of cotton yield using global and local autocorrelation statistics,” Field Crops Research, vol. 89, no. 2-3, pp. 219–236, 2004.

[39] B. A. Kitchenham, L. M. Pickard, S. G. MacDonell, and M. J. Shepperd, “What accuracy statistics really measure,” IEE Proceedings-Software, vol. 148, no. 3, pp. 81–85, 2001.

[40] L. Anselin, “The moran scatterplot as an esda tool to assess local instability in spatial association,” in Spatial analytical perspectives on GIS. Routledge, 2019, pp. 111–126.

[41] A. Lenzi and G. Millo, “Regional heterogeneity and spatial spillovers in the italian insurance market,” WP1/05, Assicurazaioni Generali, Trieste, Italy, 2005.

[42] P. Songchitruksa and X. Zeng, “Getis–ord spatial statistics to identify hot spots by using incident management data,” Transportation research record, vol. 2165, no. 1, pp. 42–51, 2010.

[43] L. Anselin, I. Syabri, and Y. Kho, “Geoda: an introduction to spatial data analysis,” in Handbook of applied spatial analysis. Springer, 2010, pp. 73–89.