[AM-03-019] Exploratory Spatial Data Analysis (ESDA)

Abstract

Exploratory Spatial Data Analysis (ESDA) is a crucial methodology within spatial statistics, designed to uncover and interpret spatial patterns, trends, and relationships within geographic datasets. Unlike traditional Exploratory Data Analysis (EDA), which focuses solely on data attributes without considering spatial context, ESDA integrates spatial information to explore how geographical factors influence data patterns. ESDA is a critical phase in the spatial data science pipeline that occurs after data collection and before modeling and consists of a combination of statistical techniques and visualizations to examine the data's structure, detect patterns, spot outliers, and investigate relationships between variables. This exploratory phase is essential for cleaning and preparing the data, aiding in identifying potential issues such as missing values or biases, informing the selection of appropriate models and techniques, and ensuring that subsequent steps in the research pipeline are grounded in a thorough comprehension of the data's characteristics. This chapter provides a holistic overview of ESDA and situates it within the broader spatial data science pipeline, while differentiating it from aspatial EDA, elucidating its core methodologies, and discussing its implications for understanding spatial datasets - aimed at equipping readers with a comprehensive introduction of ESDA techniques, laying the groundwork for advanced spatial data analysis.

Keywords: spatial statistics

Author & citation

Sachdeva, M. (2024). Exploratory Spatial Data Analysis. The Geographic Information Science & Technology Body of Knowledge. John P. Wilson (Ed.). DOI: 10.22224/gistbok/2024.1.28.

Explanation

1. Introduction
2. Emphasizing the Spatial in ESDA
3. Key Objectives and Methods of Exploratory Spatial Data Analysis

1. Introduction

Exploratory Spatial Data Analysis (ESDA) is a part of the spatial data science pipeline that consists of a collection of techniques to analyze spatial data and uncover underlying patterns, trends, associations, and relationships. The primary objective of using exploratory techniques within spatial analysis is to provide an initial exploration of the spatial data by identifying spatial structures and dependencies that might influence subsequent analyses. Using spatially explicit techniques within exploratory data analysis is crucial when dealing with most natural and social phenomena because these techniques account for spatial dependencies and context that aspatial methods might overlook. Techniques that account

for interdependencies, anomalies, and structure governed by space (and place) are consequential for analyzing most phenomena representing interactions among humans, nature, infrastructure, and the environment. Spatial data often exhibit patterns where nearby observations are more similar than distant ones, a concept underlying the fundamental principle of spatial dependence (Tobler, 1970), and that similar phenomena might differ across varying contexts and places, a concept underlying the fundamental tenet of spatial heterogeneity (Goodchild, 2004; Sui & Turner, 2022). Ignoring these inherently spatial properties in the data can lead to misleading conclusions, incomplete evidence for model selection for further analysis, and biased directions for framing subsequent research hypotheses, owing to the often-flawed assumption of independence of aspatial EDA techniques.

For example, if spatial dependencies are ignored in a study of city property prices, the analysis might miss that certain neighborhoods have higher or lower values due to localized factors like proximity to amenities. This could lead to incorrect assumptions about property price distribution and the selection of unsuitable models for testing research hypotheses. Incorporating spatial techniques, such as spatial autocorrelation or cluster analysis, reveals localized patterns, offering a more accurate understanding of the data.

2. Emphasizing the Spatial in ESDA

The fundamental properties that spatial phenomena and data representing such phenomena exhibit are characterized by primary and prevailing theories and principles within spatial science (Goodchild, 2022). These theories, and hence the lens they provide, add foci to the ‘special’ challenges and directions for further explorations that are specific to spatial data. Methods within ESDA are hence different from EDA methods common in other allied fields, as they are intended to unify the models and encompassing methodologies that are fundamentally spatial and operationalize hypotheses and investigations within research in quantitative human geography. The geographical principles commonly employed to emphasize the spatial in ESDA are (i) Spatial Dependence, (ii) Spatial Heterogeneity, and (iii) Spatial Scale Uncertainty.

1. Spatial Dependence: Spatial dependence is a fundamental concept in geography that explains how the value of a variable at a particular location is influenced by the value of the same variable at nearby locations (Tobler, 1970). Exploratory techniques within spatial data analysis are hence aimed toward investigating this aspect of the spatial data and, if found, help inform choices of models based on the theory of spatial dependence that assumes that the spatial dependence measured from units closer in space will be stronger in magnitude than that for more distant units.
2. Spatial Heterogeneity: The principle of Spatial heterogeneity in spatial sciences pertains to the uneven distribution and variability of spatial phenomena such as those describing human behavior patterns or natural hazard vulnerability across geographic space (Goodchild, 2022; Sui & Turner, 2022). Exploratory methods within spatial analysis are hence also explicitly aimed towards investigating and quantifying the spatial heterogeneity that spatial data exhibit. For example, local models for spatial autocorrelation, such as Anselin’s Local Indicators of Spatial Autocorrelation (LISA; Anselin, 1995), are based on the premise that patterns tend to be similar for certain ranges within space and beyond said thresholds tend to vary and exhibit heterogeneity.
3. Spatial Scale Uncertainty: Geographical data are scale-dependent, meaning they can

be collected at various resolutions, such as census units, leading to different spatial partitions. Uncertainties in results from analyses conducted at different spatial scales can give rise to uncertainties such as the Modifiable Areal Unit Problem (MAUP; Openshaw, 1983), which arises when data tabulated under different zonal systems produce inconsistent results, even with the same variables and study areas. An exploration of the spatial scale over which data might vary is hence an important step in the spatial data analysis pipeline since the selection of spatial scale in any analysis is fundamentally linked to how the research findings are interpreted, making it essential to clearly define the uncertainties that arise from using different scales (Fotheringham & Sachdeva, 2022; Sachdeva & Fotheringham, 2023). Further, just as spatial data are scale-dependent, techniques within ESDA used to analyze these data can also be categorized as either ‘local’ or ‘global’. Local statistics used to describe spatial data (such as local Moran’s I) do so in the context of localized neighborhoods or clusters within the whole dataset, while global statistics are calculated for the entire dataset at once. While we do not delve into a detailed difference between local and global statistics in this overview, we provide details on a mix of local and global ESDA techniques in the section below to guide the reader in their further explorations.

Figure 1 illustrates the role of exploratory spatial analysis within the broader research process, emphasizing its function in the early stages of theory development and model construction. It highlights how exploratory methods, informed by geographical theories, guide the selection of spatial principles that shape subsequent modeling and methodological decisions, thereby ensuring that the research remains contextually and spatially relevant.

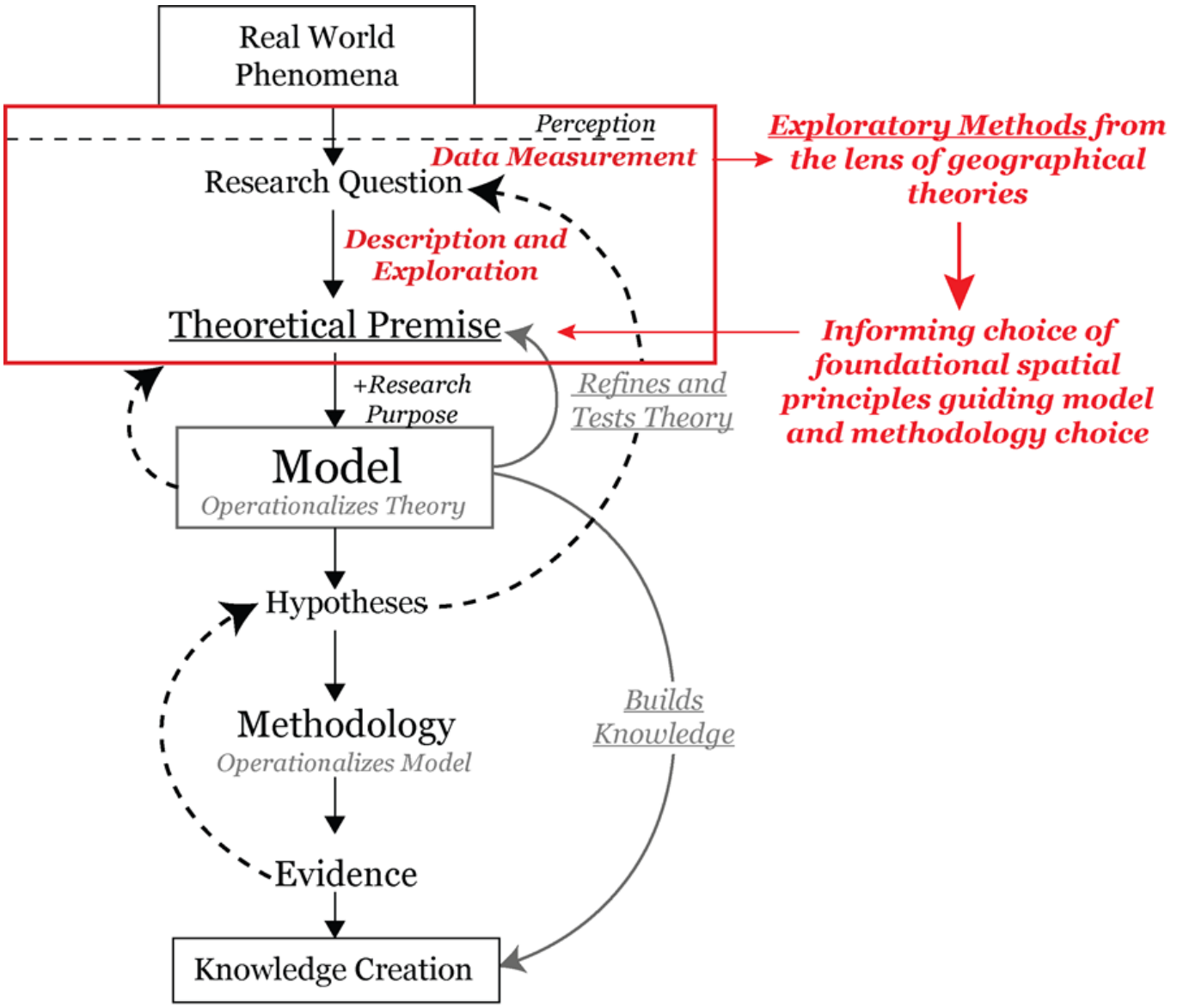

Figure 1. Situating ESDA within the broader spatial analysis research pipeline. Source: author.

3. Key Objectives and Methods of Exploratory Spatial Data Analysis

ESDA occupies a crucial role in the broader data analysis pipeline, serving as an initial step that bridges data collection and more formalized statistical modeling. ESDA techniques are applied early to help analysts uncover patterns, identify anomalies, and formulate hypotheses based on spatial relationships within the data. By visually and statistically examining spatial distributions, spatial autocorrelation, and cluster detection, techniques within ESDA provide the insights needed to guide subsequent analyses. This exploratory phase is essential for understanding the underlying structure of the data, which informs the choice of appropriate models and methods for deeper analysis. Thus, ESDA acts as a foundation for refining research questions and enhancing the robustness of later stages in the analysis pipeline.

3.1 Spatial Distribution Analysis: An important first step in the exploratory spatial analysis process involves investigating how data values are distributed across a geographical area. Spatial distribution analysis examines the geographical arrangement of phenomena, using tools such as the mean center to identify the central point of a

distribution based on spatial coordinates and spatial standard deviation that measures the spread of data points around this mean center. This stage of the analysis also includes identifying global and local outliers to understand where the spatial distribution significantly deviates from expected patterns, revealing areas of potential interest or concern.

3.2 Investigating and Quantifying Spatial Structure:

- Pattern analysis: Investigating and understanding spatial patterns helps identify the spatial structure of the data, including the clustering extent, dispersion, and spatial trends. Techniques such as kernel density estimation (KDE), K-means clustering (detailed below), and DBSCAN are employed to reveal these patterns and visualize the spatial structure of the data.
    - K-means clustering is a widely used unsupervised learning technique to partition a dataset of n points into □ distinct clusters (Lloyd, 1982). The method works by iteratively refining cluster centroids to minimize the within-cluster dissimilarity. The algorithm begins by initializing □ centroids, either randomly or through predefined criteria. Each data point is then assigned to the cluster associated with the closest centroid, where the Euclidean distance determines closeness, following which the centroids are recalculated as the mean of the points within each cluster. The process of assigning points to clusters and updating centroids is repeated until they converge/they no longer change substantially. The K-means algorithm primarily focuses on attribute similarity and does not ensure spatial contiguity in clusters. This is hence, inherently an EDA technique with no explicit spatial consideration. An approach to make it explicitly spatial adds x and y coordinates as attributes, promoting compact clusters while another method uses weighted multi-objective optimization, balancing attribute similarity with spatial proximity to create contiguous clusters (Anselin et al., 2006). K-means clustering using spatial constraints, is an effective example of how many ESDA methods have their foundations in EDA techniques.
- Spatial Autocorrelation: Key factors about the data that are investigated at this stage include whether there are clusters of high or low values and whether the spatial and covariate distribution appear random or structured. Techniques used to quantify spatial autocorrelation, such as Moran’s I (detailed below), Getis and Ord G (Getis & Ord, 1992) and Geary’s C (Geary, 1954), aid in examining these aspects of the spatial data.
    - Moran's I is a key statistical metric for evaluating spatial autocorrelation, which measures the extent to which similar values are spatially clustered (Moran, 1950). This index assesses whether a spatial pattern is clustered, dispersed, or randomly distributed by comparing the value of a variable at each location to values at neighboring locations. The Moran's I statistic is calculated as:

$$I = \frac{\{N \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} (x_i - \bar{x}) (x_j - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2 \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}} \qquad (1)$$

- Scale of Spatial Variation: Methods such as Variograms are useful in this step of the spatial analysis pipeline to quantify the distance and spatial scale over which data values might vary. Variograms quantify the degree of similarity between pairs of data points as a function of the distance separating them, providing insights into the spatial

structure of the data (Matheron, 1963). Variograms are crucial for building spatial models, such as kriging, that rely on understanding spatial autocorrelation to predict values at unsampled locations.

3.3 Spatial Relationships and Correlations: This step of the spatial data exploration involves examining how spatial variables interact with one another. For example, this step within the data exploration process might reveal how environmental factors correlate with health outcomes or how socioeconomic variables influence spatial development, to help inform subsequent analyses and hypothesis choices in the research of a phenomenon.

- Aspatial correlation statistics: Statistics such as Pearson's correlation coefficient and Correlation Ratios (Lee Rodgers & Nicewander, 1988) and graphical methods such as scatter plots can be used to investigate correlations between variables within a dataset.
- Bivariate Spatial Autocorrelation: Bivariate and multivariate counterparts of spatial autocorrelation techniques, such as bivariate Moran's I and Multivariate Geary's C (detailed below; Wartenberg, 1985) analyses, are used in this step to investigate how correlations in the covariate space (between different data values) might cluster similarly or dissimilarly with correlations in geographic space.
- The local multivariate Geary's C statistic is an extension of the univariate Geary's C, designed to assess spatial autocorrelation across multiple variables simultaneously (Wartenberg, 1985; Anselin, 2019). It measures the degree to which similar or dissimilar values of different variables are spatially clustered. The local multivariate Geary's C statistic (for a bivariate case, for example) is calculated as:

$$\sum w_{ij}\, d_{ij}^2 = \sum w_{ij}\,(z_{1,i} - z_{1,j})\text{^}2 + \sum w_{ij}\,(z_{2,i} - z_{2,j})\text{^}2 \qquad (2)$$

which represents the weighted squared distance in attribute space between the values at observation i and its geographic neighbor j, between two standardized variables, z1 and z2 (Anselin, 2019). The Geary's C statistic is hence additive in attribute space and is defined as a sum of the C statistic for all covariates in a generalized multivariate case (Anselin, 2019). The local multivariate Geary's C provides a more comprehensive understanding of spatial relationships by considering the interactions between multiple variables, making it a valuable tool for detecting spatial patterns in complex datasets.

ESDA is pivotal in guiding modeling choices and shaping research outcomes by revealing the underlying spatial patterns and relationships within data. Through techniques such as spatial autocorrelation, cluster detection, and spatial correlation analysis, ESDA helps identify areas of spatial dependence, heterogeneity, and potential anomalies. These insights are crucial for selecting appropriate spatial models, such as those that account for spatial lag or error, ensuring that the models accurately capture the underlying spatial processes. Moreover, ESDA informs the selection of relevant variables and the appropriate spatial scales for analysis, ultimately leading to more robust and reliable research outcomes sensitive to the spatial dimensions of the data.

References

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

Anselin, L. (2019). A Local Indicator of Multivariate Spatial Association: Extending Geary's c. Geographical Analysis, 51(2), 133-150.

Anselin, L., Syabri, I., and Kho, Y. (2006). GeoDa: An Introduction to Spatial Data Analysis. Geographical Analysis 38 (1), 5-22.

Fotheringham, A. S., & Sachdeva, M. (2022). Scale and local modeling: New perspectives on the modifiable areal unit problem and Simpson's paradox. Journal of Geographical Systems. 24, 475–499.

Geary, R. C. (1954). The Contiguity Ratio and Statistical Mapping. The Incorporated Statistician, 5(3), 115-146.

Getis, A., & Ord, J.K. (1992). The analysis of spatial association by use of distance statistics. Geographical Analysis, 24(3), 189-206.

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.

Goodchild, M. F. (2022). Commentary: General principles and analytical frameworks in geography and GIScience. Annals of GIS, 28(1), 85–87.

Lee Rodgers, J., & Nicewander, W. A. (1988). Thirteen Ways to Look at the Correlation Coefficient. The American Statistician, 42(1), 59–66.

Lloyd, S. (1982). Least squares quantization in PCM. IEEE Transactions on Information Theory, 28(2), 129–137.

Matheron, G. (1963). Principles of Geostatistics. Economic Geology, 58(8), 1246–66.

Moran, P. A. P. (1950). Notes on Continuous Stochastic Phenomena. Biometrika, 37: 17–23.

Openshaw, S. (1984). The Modifiable Areal Unit Problem. Concepts and Techniques in Modern Geography No. 38. Norwich, England: Geo Books, Regency House.

Rey, S., Arribas-Bel, D., & Wolf, L. J. (2023). Geographic Data Science with Python (1st Edition). Chapman and Hall/CRC.

Sachdeva, M. (2024). Models and Human Geography. In: Warf, B. (eds) The Encyclopedia of Human Geography. Springer, Cham.

Sachdeva, M., & Fotheringham, A. S. (2023). A Geographical Perspective on Simpson's Paradox. Journal of Spatial Information Science, 26, Article 26.

Sui, D., & Turner, M. (2022). General theories and principles in geography and GIScience: Moving beyond the idiographic and nomothetic dichotomy. Annals of GIS, 28(1), 1–4.

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

Wartenberg, D. (1985). Multivariate Spatial Correlation: A Method for Exploratory Geographical Analysis. Geographical Analysis, 17(4), 263–283.