[AM-03-011] Spatial Statistics

Abstract

Spatial statistics is dedicated to describing and modeling georeferenced data through the application of statistical theories and methods. Unlike conventional statistical approaches, which often assume independence among observations, spatial statistical techniques allow to account for locational aspects observations in addition to their attributes. Modeling georeferenced data with conventional non-spatial statistical approaches can lead to bias and unreliable results. This article first discusses measurements of spatial arrangements including mean center and standard distance deviation. It then reviews statistical methods for the types of spatial data—point data, geostatistical data, and areal data. Following this, it examines Bayesian spatial models, which offer a flexible framework for incorporating spatial dependence. Finally, the article concludes with a discussion of ongoing challenges in spatial statistics, including potential limitations of area-unit based observations, computational limitations, and issues related to data uncertainty.

Keywords: Bayesian, MAUP, spatial statistics

Author & citation

Chun, Y. (2025). Spatial Statistics. The Geographic Information Science & Technology Body of Knowledge (2025 Edition), John P. Wilson (ed). DOI: 10.22224/gistbok/2025.1.13

Explanation

1. Introduction
2. Measurements for Spatial Arrangements
3. Spatial Statistical Data Analysis
4. Bayesian Spatial Modeling
5. Challenges

1. Introduction

Spatial statistics focuses on describing and modeling georeferenced data by applying statistical theories and methods. Unlike conventional statistical approaches, which assume independence among observations, spatial statistics explicitly account for spatial correlations. This reflects the broader principle in spatial science that observations closer in space are more likely to interact or exhibit similarities compared to those farther apart. This phenomenon aligns with Tobler's First Law of Geography (1970), which states, "Everything is related to everything else, but near things are more related than distant things." Spatial statistics provides the theories, concepts, and tools necessary to analyze georeferenced data and develop models that account for spatial aspects. Such spatial characteristics include trends, heterogeneity, and correlation patterns among observations. While spatial trends describe systematic patterns across space, spatial heterogeneity refers to the uneven distribution or concentration of a phenomenon. The correlation observed in spatial datasets is termed spatial autocorrelation, as it involves the relationship between values arranged in a spatial structure. Spatial autocorrelation is conceptually similar to temporal

autocorrelation, where correlations appear in time series datasets due to sequential temporal observations.

Using conventional statistical methods on georeferenced data can lead to biased or unreliable results because such models do not account for spatial characteristics such as spatial heterogeneity. For example, spatial autocorrelation violates the independence assumption, a foundational principle for methods like maximum likelihood estimation. In contrast, spatial statistical techniques incorporate model specifications that account for spatial autocorrelation. These models often extend conventional specifications by introducing a parameter that quantifies the degree of spatial autocorrelation, ensuring more accurate and reliable analyses.

2. Measurements for Spatial Arrangements

Spatial arrangements reveal the distributional characteristics of a phenomenon in space with summarized values. Centrographic measures, commonly used to describe central tendency and dispersion in two-dimensional Cartesian space, extend from univariate statistics such as the mean and standard deviation. These measures include the mean center and standard distance deviation. In addition, the centroid of a polygon can be used as a single representative point for the polygon with a series of vertices.

The mean center, representing the average location of observations, is calculated as the mean of the x and y coordinates:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

where □ denotes the number of observations. Figure 1 illustrates an example using the locations of 195 California Giant Redwood trees, as reported by Strauss (1975). The red dot represents the mean center of these trees at (0.5075, 0.4635).

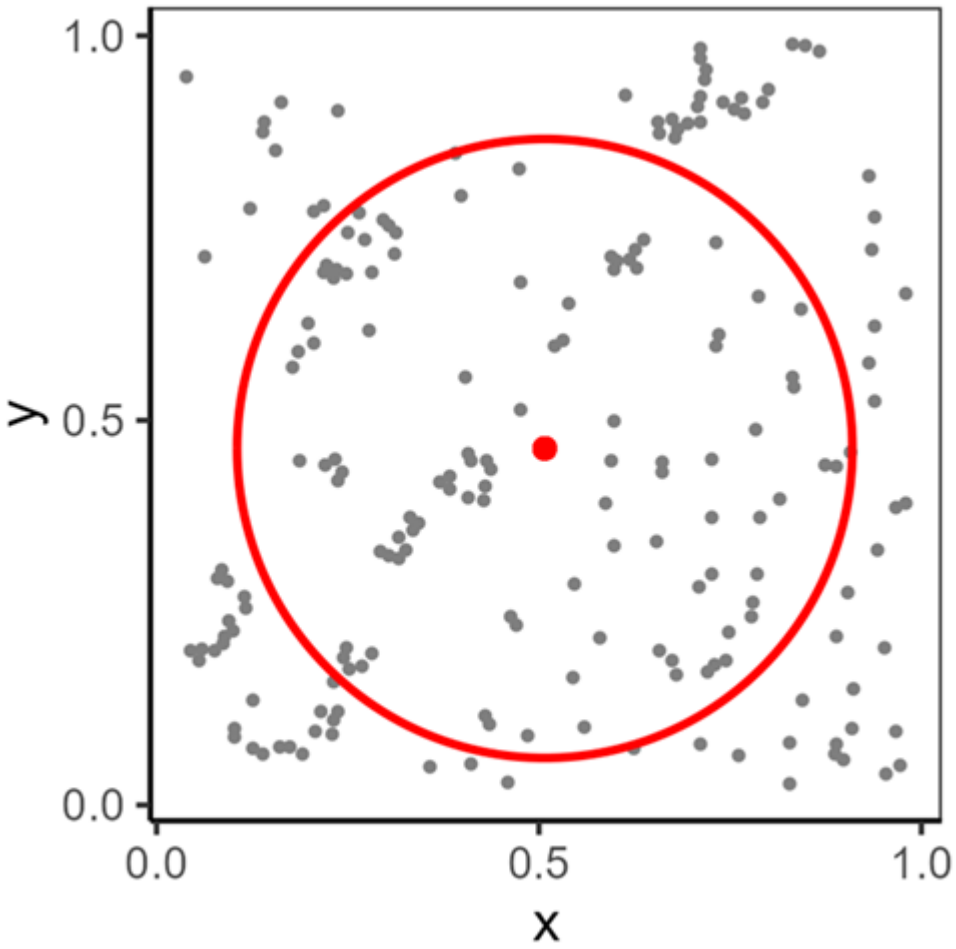

Figure 1. The mean center and the standard deviation distance among 195 California Giant Redwood tree data in Strauss (1975). Source: author.

The standard distance deviation (SDD) measures the dispersion of observations around the mean center, extending the concept of standard deviation from univariate statistics. It is given by:

$$\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}+\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n}}$$

A smaller SSD value indicates a tighter clustering of observations around the mean center, whereas a larger value suggests greater dispersion. The red circle in Figure 1 represents the standard distance deviation (0.4023) around the mean center. These centrographic measures can also be extended by incorporating unequal weights for observations, allowing for more nuanced spatial analyses.

A polygon centroid is the geometric center of the polygon and is often used as a representative point for spatial analysis. For a triangle, the centroid's x and y coordinates are computed as the mean of its vertices' coordinates. However, for a general polygon, the centroid is determined using the following formulas

$$x_c = \frac{1}{6A}\sum_{i=1}^{n}\left(x_i + x_{i+1}\right)\left(x_i y_{i+1} - x_{i+1} y_i\right),$$

$$y_c = \frac{1}{6A}\sum_{i=1}^{n}\left(y_i + y_{i+1}\right)\left(x_i y_{i+1} - x_{i+1} y_i\right)$$

where A represents the area of the polygon.

3. Spatial Statistical Data Analysis

Spatial statistical methods aim to analyze spatial data considering their characteristics that originate from their locations as well as attributes. They have different goals and model specifications based on the types of spatial data, which are generally categorized as point data, geostatistical data, and areal data (Schabenberger and Gotway, 2017). This section briefly outlines the unique characteristics of these data types and highlights commonly used statistical methods tailored to each.

3.1 Spatial Point Pattern Analysis

Spatial point pattern analysis examines the spatial distribution of point events, assessing whether they occur in a clustered, random, or regular arrangement in space. Figure 2 illustrates examples of spatial point patterns. Figure 2a depicts the locations of the 195 California Giant Redwood trees, which exhibit a clustered pattern. Figure 2b shows 100 points randomly generated with their (x, y) coordinates drawn from a uniform distribution between 0 and 1, representing a complete spatial randomness (CSR) process. Figure 2c presents 100 points arranged systematically with 0.1 spacing and minor random noise, demonstrating a regular spatial point pattern.

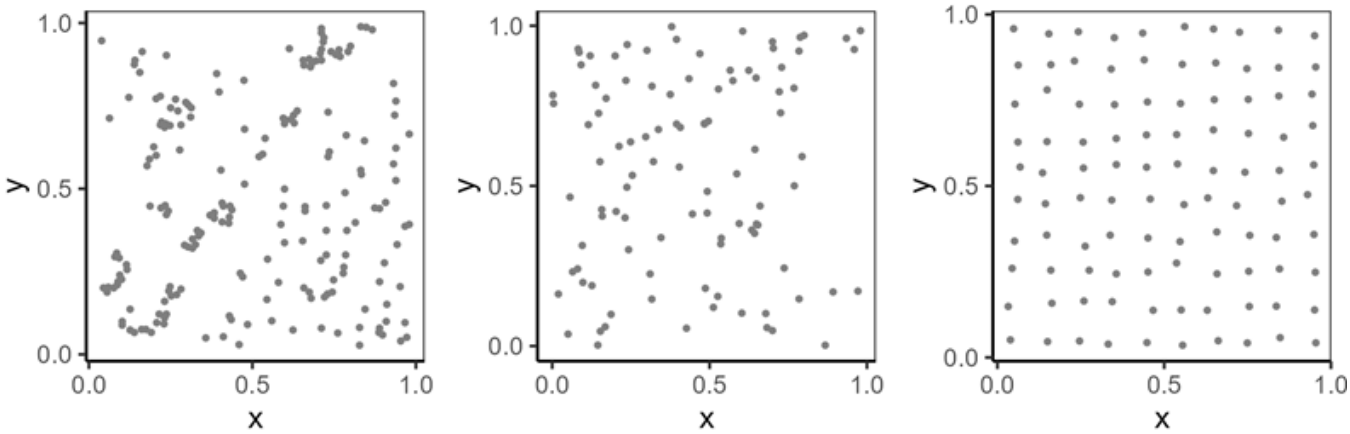

Figure 2a, 2b, and 2c (left to right). Examples of three spatial point patterns. Source: author.

Spatial point pattern analysis tests whether an observed point pattern significantly differs

from a random spatial pattern, such as the complete spatial randomness (CSR). Quadrat analysis is a method used to determine whether a spatial point pattern deviates from CSR. In this approach, a study area is divided into quadrats, and the number of points in each quadrat is counted. The pattern is then evaluated using the variance-to-mean ratio (VMR), calculated by dividing the variance of the counts by the mean. Under CSR, the VMR is expected to be 1; a VMR greater than 1 indicates a clustered pattern, whereas a VMR less than 1 suggests a dispersed pattern. These differences can be statistically tested using the chi-square distribution. Figure 3 illustrates quadrats arranged in a 5-by-5 regular pattern for the California Giant Redwood trees data. The analysis indicates that the point pattern of the trees is clustered (p-value < 0.0000). However, it is important to note that quadrat analysis can be sensitive to the choice of quadrat configuration.

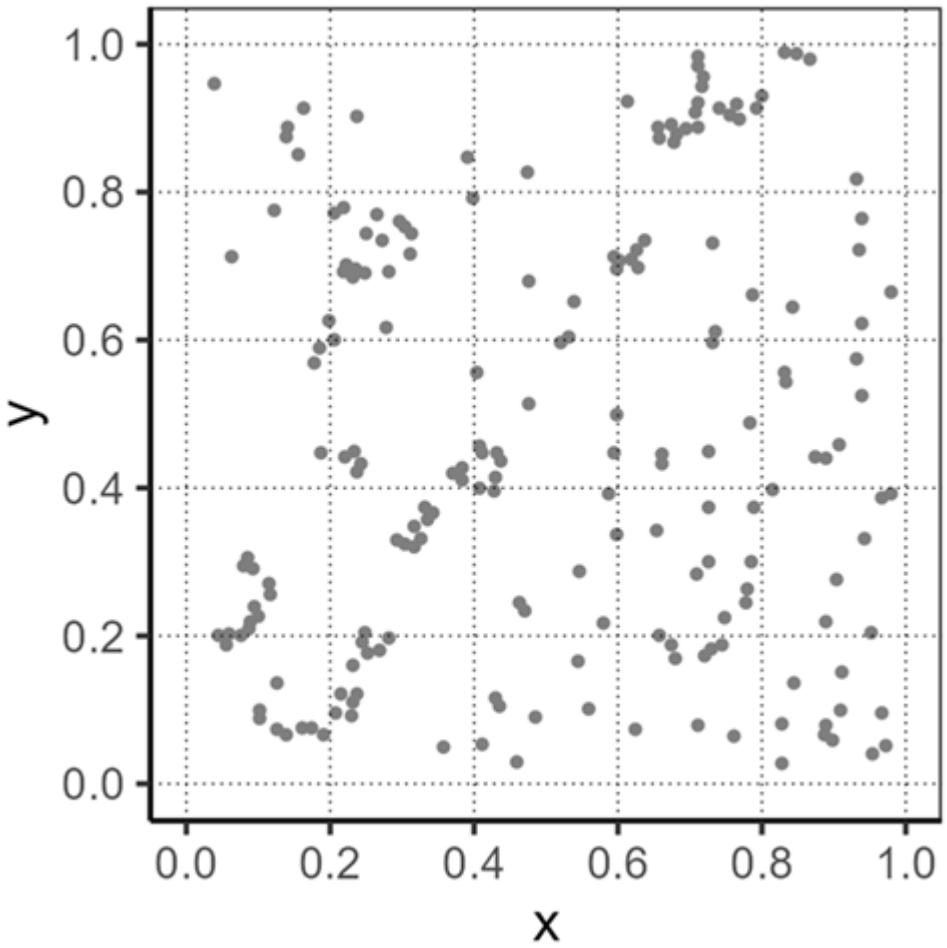

Figure 3. An illustration of quadrat analysis. Source: author.

Additional methods for spatial points analysis have been developed for spatial point analysis that rely on distance-based metrics, such as nearest-neighbor distances and pairwise distances. Statistical inferences are drawn by comparing observed distances with those expected under CSR. Detailed descriptions of these methods can be found in Yuan et al. (2020) or textbooks such as Baddeley et al. (2015). It is important to note that spatial pattern analysis focuses on spatial point locations rather than correlations in associated attributes.

3.2 Geostatistical Data Analysis

Geostatistical data represents phenomena that vary continuously across space. Geostatistical modeling predicts continuous surfaces from observed sample values, leveraging a correlation structure defined as a function of distance. Correlation is quantified using attribute values at sample locations alongside pairwise distances. Typically, correlation is strongest at shorter distances, decreases as distance increases, and stabilizes beyond a threshold distance. This spatial structure is often represented using variance or semi-variance, which complements correlation in understanding the data's spatial characteristics.

Geostatistical data analysis generates an estimation surface by predicting values at unobserved locations using spatially neighboring values and their correlation structure. Figure 4 shows the predicted surface derived from rainfall data at 112 weather stations in Puerto Rico using ordinary Kriging - a widely used geostatistical method. For more detailed information on geostatistical methods, readers are encouraged to consult additional resouraces, such as Goovaerts (2019).

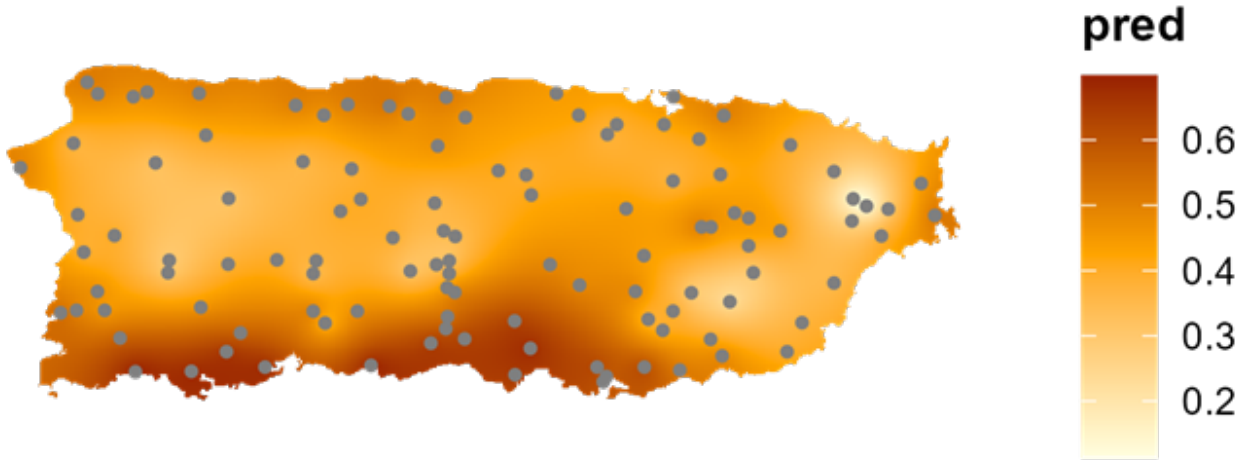

Figure 4. A prediction surface using ordinary kriging for rainfall in Puerto Rico. Source: author.

3.3 Areal Data Analysis

Area data, also referred to as lattice data, focuses on analyzing attributes collected from discrete, non-overlapping spatial units, such as administrative or census units. This type of analysis typically involves regression models, where a target variable is treated as the dependent variable. The analysis often begins with ESDA to detect the presence of spatial autocorrelation, which can suggest spatial models instead of conventional non-spatial

models. Common tools in this phase include global spatial autocorrelation measures, such as Moran's I, and visual examination of spatial patterns, such as mapping the variable of interest (Bivand, 2009). Figure 5 illustrates a map of farm densities for municipalities in Puerto Rico in 2007, with the data transformed using the Box-Cox method. Visual inspection of the map suggests positive spatial autocorrelation, with clusters of high and low values. The global Moran's I test further confirms this positive autocorrelation, yielding a z-score of 4.7358 and a p-value of less than 0.0001, suggesting significant spatial dependence in the farm density variable.

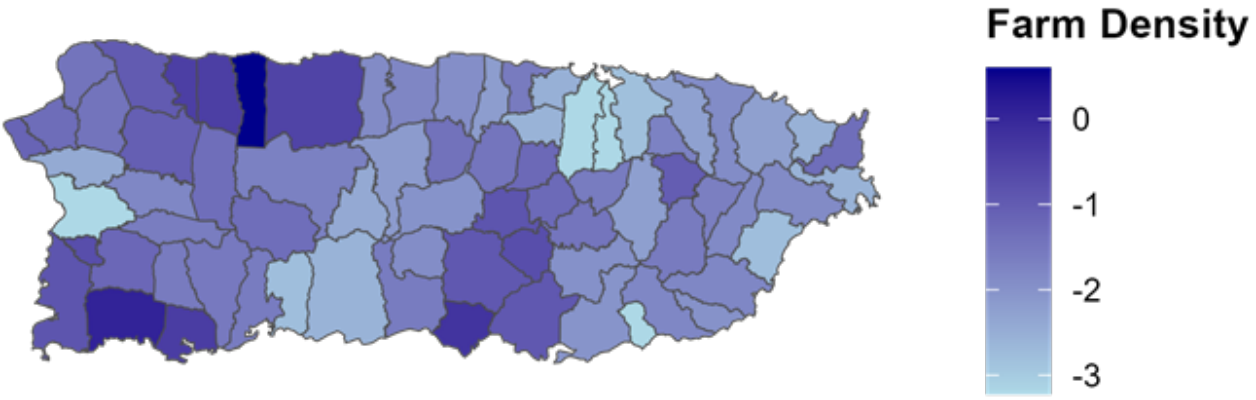

Figure 5. Box-Cox transformed farm densities for the 78 municipalities of Puerto Rico in 2007. Source: author.

Spatial regression models are designed to account for spatial autocorrelation within their structure. For instance, a spatial autoregressive model (SAR), that is also called spatial lag model, modifies a standard linear regression by adding a spatial lag term, **WY**, which are essentially average values of neighbors for each observation, to the right-hand side of the equation. Here, **W** represents a spatial weights matrix, which quantifies spatial proximity, and **Y** is the dependent variable. The SAR model is typically expressed as:

$$\mathbf{Y} = \rho\mathbf{WY} + \mathbf{X}\beta + \varepsilon$$

where ρ is the spatial autocorrelation parameter, **X** represents independent variables, **β** are the corresponding coefficients, and **ε** denotes the errors. When ρ=0 , the model reduces to a conventional linear regression model. Various other spatial regression models incorporate spatially lagged variables, such as **WY** and **WX**, along with their coefficients. These models extend conventional regression analysis by recognizing spatial dependencies between observations, ensuring more accurate and contextually relevant results. Similarly, various spatial regression models are specified with spatially lagged variables i.e., **WY** and **WX**, along with associated parameters. Further detailed discussions of spatial autoregressive models can be found in sources like Hoffman and Kedron (2023) and LeSage and Pace (2009).

Geographically weighted regression (GWR) is a spatial regression technique used to explore

local relationships between a dependent variable and one or more independent variables by allowing the regression coefficients to vary across space. The underlying idea is that the relationship between two variables may differ across areal units, which cannot be adequately captured by a single global parameter. Instead, GWR estimates a set of localized parameters based on nearby observations. A GWR model can be written as:

$$y_i = \beta_0 (u_i, v_i) + \sum_j \beta_j (u_i, v_i) x_{ij} + \varepsilon_i$$

where ui,vi denotes the geographic location corresponding to observation (i), β0 is the intercept, βj are coefficients for the independent variables, and εi denotes the random error. These localized coefficients capture spatial variations in the relationships, and their estimation employs a weighted scheme where nearby observations are given greater influence based on their proximity to the estimation point. For a more detailed description of GWR and its recent developments, please see Sachdeva and Fotheringham (2020).

4. Bayesian Spatial Modeling

Bayesian statistics provides a robust framework for modeling spatial data. In Bayesian statistics, parameters are treated as random variables, and inferences are drawn from the posterior distribution. The posterior distribution is a combination of prior knowledge, expressed through a prior distribution, and the likelihood, which represents the probability of observing the data given the parameters. This process allows for updating beliefs about parameters as new data is observed. However, calculating the posterior distribution analytically can be complex, particularly for models with multiple parameters. To overcome this challenge, Markov chain Monte Carlo (MCMC) methods are commonly employed. MCMC methods, such as Gibbs sampling and Metropolis-Hastings algorithms, allow parameters to be estimated by drawing from the conditional distributions of the parameters, which is essential for handling complex, high-dimensional models (Gelman et al., 2013).

Bayesian modeling offers significant advantages over traditional frequentist statistics when it comes to modeling spatial data. One key benefit is its flexibility in accommodating complex spatial dependency structures. In a frequentist framework, constructing a likelihood function to capture such intricate spatial dependencies can be nearly impossible. The simulation-based estimation methods offer a robust tool for estimating Bayesian models, particularly when dealing with complex spatial dependency structures. Additionally, Bayesian approaches are highly flexible and can readily accommodate hierarchical structures, which are commonly encountered in spatial data, including multi-level or nested spatial relationships. This adaptability is a significant advantage over traditional statistical methods, which often struggle to model such complex structures. By incorporating such complex dependencies, Bayesian models improve the accuracy and reliability of spatial data analysis. Detail descriptions for various Bayesian spatial models can be found in Banerjee et al. (2003), which include Bayesian kriging, spatial autoregressive models, conditional autoregressive models, generalized linear spatial models, spatially varying coefficient models, and spatio-temporal models

5. Challenges

Spatial statistics are widely applied across various research domains that involve georeferenced data, including fields such as geography, spatial econometrics, regional science, epidemiology, criminology, ecology, and environmental science. With the continuous advancements in computational capabilities, data accessibility, and analytical techniques, spatial statistical modeling continues to evolve. Despite these advancements, challenges remain in the field (e.g., Gelfand, 2020).

One significant challenge in spatial statistics is the modifiable areal unit problem (MAUP). MAUP refers to the phenomenon where statistical analysis results can vary depending on the choice of areal units used to tabulate observations (Wong, 2020). In other words, different spatial scales—such as census tracts versus census block groups—or zoning schemes can considerably influence the outcomes of spatial analyses. Although MAUP has been recognized for decades, it remains a persistent issue. A similar problem exists for temporal data, known as the Modifiable Temporal Unit Problem (MTUP), where statistical results can differ markedly depending on the temporal units used for observation (e.g., day, week, or month). Moreover, the effects of both MAUP and MTUP are even more pronounced in space-time modeling.

Recent studies highlight potential challenges in using geographic unit-based observations for individual-level analyses. These challenges are encapsulated in the Uncertain Geographic Context Problem (UGCoP) (Kwan, 2012) and the Neighborhood Effects Averaging Problem (NEAP) (Kwan, 2018). The UGCoP emphasizes that the arbitrary delineation of areal units may result in area-based attributes that do not accurately represent the intended contextual factors or align with individual-level data. In other words, reliance on commonly used geographic units can obscure the identification of a “true causally relevant” geographic context, potentially leading to inaccurate effect estimates and misleading analytical conclusions. The NEAP suggests that individuals exhibit diverse daily movement patterns, leading to considerable variation in their environmental exposures. As a result, neighborhood-level attributes—particularly those based solely on residential locations—fail to capture the heterogeneous exposures experienced by individuals. Consequently, analyses relying on such variables are limited in their ability to account for this variability and may instead produce oversimplified, averaged patterns.

Another significant challenge in spatial statistics arises from the increasing volume of spatial data, which has expanded with advancements in technologies like remote sensing and database management systems. For instance, remotely sensed images can contain millions of data points (i.e., pixels), making the modeling of such large datasets computationally intensive. Moreover, spatio-temporal data introduces an additional complexity by incorporating a temporal dimension, which can create more intricate space-time dependencies that complicate data modeling (Cressie and Wikle 2011). The sheer volume of data, especially Big Data, can also affect the validity of statistical inferences. As the number of observations increases, sample variances tend to decrease, leading to statistically significant parameter estimates even when the actual estimates are near zero. This phenomenon suggests that traditional statistical inferences may not hold the same for Big Data, highlighting the need for further investigation into the impact of large datasets on model accuracy and interpretability.

Uncertainty and bias in spatial data represent another significant challenge in spatial

statistics. Uncertainty can arise in various aspects, such as the attributes and locations of spatial data, model specifications, and the data collection process, including spatial sampling (Griffith et al., 2015). Uncertainty in both attributes and locations directly affects how accurately a phenomenon is represented and can distort the spatial relationships between observations. Additionally, incorrect model specifications, such as using linear regression for non-normally distributed data, can lead to biased or unreliable statistical inferences. Spatial sampling techniques, when properly designed, can help to better reflect the spatial distribution of a target phenomenon and minimize potential bias. However, newly emerging data sources, such as social media, may introduce new biases, as the users of these platforms do not necessarily represent the broader population. This introduces the risk of ecological fallacy, where inferences about the population at large may be misleading due to the non-representative nature of the sample. Thus, while such data sources can offer valuable insights, they must be used cautiously, with a careful understanding of their limitations and potential biases.

References

Baddeley, A., Rubak, E., & Turner, R. (2016). Spatial point patterns: methodology and applications with R. Boca Raton; London; New York: CRC Press, Taylor & Francis Group.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2003). Hierarchical Modeling and Analysis for Spatial Data, 1st Edition. Chapman and Hall/CRC.

Bivand, R.S. (2010). Exploratory Spatial Data Analysis. In: Fischer, M., Getis, A. (eds) Handbook of Applied Spatial Analysis. Springer, Berlin, Heidelberg.

Cressie, N. & Wikle, C.K. (2011). Statistics for Spatio-Temporal Data. Hoboken, New Jersey: John Wiley & Sons.

Gelfand, A. E. (2020). Statistical challenges in spatial analysis of plant ecology data. Spatial Statistics, 37, 100418.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and and Rubin, D. B. (2013). Bayesian Data Analysis, 3rd edition. Chapman and Hall/CRC.

Goovaerts, P. (2019). Kriging Interpolation. The Geographic Information Science & Technology Body of Knowledge (4th Quarter 2019 Edition), John P. Wilson (ed.).

Griffith, D. A., Wong, D. W., and Chun, Y. (2015). Uncertainty-related research issues in spatial analysis. In Shi, J, Wu, B., and Stein, A. (eds.), Uncertainty Modelling and Quality Control for Spatial Data. Taylor & Francis Group/CRC Press, 3-11.

Hoffman, T. D. and Kedron, P. (2023). Spatial Autoregressive Models. The Geographic Information Science & Technology Body of Knowledge (2nd Quarter 2023 Edition). John P. Wilson (Ed.).

Kwan, M. P. (2012). The Uncertain Geographic Context Problem. Annals of the Association of American Geographers, 102(5), 958-969.

Kwan, M.-P. (2018). The Neighborhood Effect Averaging Problem (NEAP): An Elusive Confounder of the Neighborhood Effect. International Journal of Environmental Research and Public Health, 15(9), 1841.

LeSage, J. P. and Pace, R. K. (2009). Introduction to Spatial Econometrics. Chapman and Hall/CRC Press.

Sachdeva, M. and Fotheringham, A. S. (2020). The Geographically Weighted Regression Framework. The Geographic Information Science & Technology Body of Knowledge (4th Quarter 2020 Edition), John P. Wilson (ed.).

Schabenberger, O. and Gotway, C. A. (2017). Statistical Methods for Spatial Data Analysis. Chapman and Hall/CRC.

Strauss, D. J. (1975). A model for clustering. Biometrika, 62, 467-475.

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

Wong, D. (2020). Aggregation effects in geo-referenced data. In Arlinghaus, S. (Ed.), Practical Handbook of Spatial Statistics (pp. 83-106). CRC Press.

Wu, A.-M., and Kemp, K. K. (2019). Global Measures of Spatial Association. The Geographic Information Science & Technology Body of Knowledge (1st Quarter 2019 Edition), John P. Wilson (Ed.).

Yuan, Y., Qiang, Y., Bin Asad, K., and Chow, T. E. (2020). Point Pattern Analysis. The Geographic Information Science & Technology Body of Knowledge (1st Quarter 2020 Edition), John P. Wilson (ed.).