Jurnal Pendidikan Matematika Vol: 3, No 1, 2025, Page: 1-11

Statistical Challenges in Spatial Data Analysis: The Role of Kriging Models

Ammar Ali Farhan

Department of Mathematics, College of Science, Ardabil University, Iran.

DOI:
https://doi.org/10.47134/ppm.v3i1.2077
*Correspondence: Ammar Ali Farhan
Email: mly740053@gmail.com

Received: 04-09-2025
Accepted: 03-10-2025
Published: 10-11-2025

Abstract: Using Kriging models, a complex geostatistical technique for extrapolating and forecasting unknown spatial values based on known data, this study investigates spatial data analysis. Traditional statistical techniques that suppose observations to be independent are considerably challenged by spatial autocorrelation—the tendency for nearby spatial points to show comparable features. The research highlights the application of Kriging to environmental data, especially air quality measurements like PM2.5 concentrations, in order to better comprehend and forecast pollution patterns over several geographical areas. Using both Ordinary and Universal Kriging approaches, the research shows how these methods can efficiently address spatial dependencies, nonstationarity (where data characteristics change across space), and anisotropy (directional spatial variability). Moreover, the research combines Kriging with machine learning algorithms to record more sophisticated spatial interactions, therefore enhancing prediction accuracy. Methods of crossvalidation are used to thoroughly evaluate the models' performance. The study emphasizes how Kriging enables precise spatial predictions, hence giving important information for environmental monitoring and well-informed decision-making.

Keywords: Spatial Data Analysis, Kriging Models, Cross-Validation.

Introduction

Spatial data refers to information about the locations and attributes of geographic features. It is typically represented as points, lines, or polygons in vector format, or as raster grids. The key challenge in spatial data analysis lies in its inherent spatial autocorrelation — the tendency of nearby observations to be more similar than distant ones. This violates the assumption of independence in classical statistics.

Spatial data are omnipresent in disciplines such as geography, environmental science, public health, and natural resource management. Unlike traditional datasets, spatial datasets include geographic references, making location a vital component of the analysis. Understanding spatial dependency is essential before applying spatial statistical models. Anselin (1995) emphasized that spatial interaction, which reflects how one location influences another, must be considered explicitly in any spatial analysis.

Research Objectives

1. Use kriging models to properly forecast air quality indicators and pollution levels across cities and so improve environmental forecasts.
2. Solve spatial autocorrelation, nonstationarity, and anisotropy by means of sophisticated Kriging techniques to address spatial challenges.
3. Combine Kriging with machine learning methods to identify sophisticated spatial patterns and so improve prediction accuracy.

Research Problem

1. Traditional models fail owing of spatial autocorrelation, wherein neighboring points are connected.
2. Nonstationarity: Many spatial datasets show nonstationarity, therefore complicating model use.
3. Anisotropy: In models without right adjustments, directional spatial variance can lower accuracy.

Research Methodology

1. Ordinary and universal Kriging should be used to estimate spatial values under diverse circumstances.
2. Determine model performance and improve projections with LeaveOneOut CrossValidation.
3. Improve Kriging predictions by adding machine learning models to identify nonlinear spatial relationships.

Importance of the Research

1. Improving Environmental Monitoring: By means of Kriging models to properly predict air quality and pollution levels over spatial zones, the study helps to improve environmental management.
2. Enhancing Predictive Models: It shows how Kriging techniques may be applied to manage sophisticated spatial dependencies and increase forecast accuracy in regions with scarce data.
3. Integration with Machine Learning: The study highlights how Kriging can be combined with machine learning algorithms to improve predictive accuracy by tackling nonlinear spatial patterns that conventional models could overlook.

Result and Discussion

Previous Studies

Jay D. Martin A Study on the Use of Kriging Models to Approximate Deterministic Computer Models 2008

Within the framework of approximation and global optimization in Design and Analysis of Computer Experiments (DACE), this research seeks to compare and assess the performance of three different kinds of Kriging models—Ordinary Kriging, Universal Kriging, and Detrended Kriging. With a focus on the underlying assumptions of each

paradigm, the research aims to bring out their similarities and contrasts. Moreover, it tries to contrast two approaches for model parameter estimate: Maximum Likelihood Estimation (MLE) and CrossValidation (CV). Starting with a one-dimensional issue for visualisation and advancing to greater dimensional ones including two-dimensional and fivedimensional cases, the study employs these models to six test problems to evaluate their performance in more difficult situations.

A tutorial guide to geostatistics: Computing and modelling variograms and kriging Author links open overlay panel M.A. Oliver 2014

Using Kriging for spatial data interpolations in environmental science, this study aims to emphasize the great need of understanding the underlying hypothesis and presumptions driving it. Even though Kriging offers unbiased estimations with least variance, wrong use of automated geostatistical methods and geographic information systems (GIS) might create false or deceptive information. The study helps readers through the precise computation and modeling of the sample variogram, which is essential for effective Kriging. It highlights the necessity of picking appropriate mathematical functions for the variogram and understanding how Kriging error variances are influenced by model choices. Furthermore, explained in the research are the critical choices to be made during Kriging, which include the choice between pointbased or blockbased support and if predictions should be global or limited within moving windows.

panelJack P.C. Kleijnen Kriging metamodeling in simulation: A review 2009:

The study's goal is to provide a complete picture of spatial correlation modeling by outlining the fundamental assumptions, equations, and differences of kriging from traditional linear regression metamodels. Then the study broadens Kriging to include random simulation approaches and investigates how bootstrapping could be employed to approximate the variance of Kriging predictions. The paper also examines a number of statistical techniques including sequential and custom designs especially created for sensitivity analysis and optimization as well as more conventional designs like Latin Hypercube Sampling. The research finally highlights significant areas for further investigation in spatial modeling and Kriging.

Difference Between Previous Studies and Current Study

There are major distinctions between the present study and earlier ones. Prior research mostly concentrated on comparing several kinds of Kriging models—including Ordinary Kriging, Universal Kriging, and Detrended Kriging—for approximation and worldwide optimization. Emphases in these studies were on theoretical underpinnings, model comparisons, and applications in several fields including geostatistics and simulation modeling. On the contrary, particularly for air quality forecasting—that is, PM2.5 levels—the present study specifically focuses on using Kriging models to spatial data analysis for environmental monitoring. Addressing difficulties like spatial autocorrelation,

nonstationarity, and anisotropy, it combines Kriging with machine learning methods to improve spatial predictions.

Foundations of Spatial Data Analysis

1. Types of Spatial Data

There are two major types of spatial data: geostatistical data and lattice data. Geostatistical data are continuous and observed at specific locations, such as soil moisture or rainfall. Lattice data are aggregated over areal units, such as census tracts or administrative boundaries. A third type, point pattern data, focuses on the location of events like disease outbreaks or traffic accidents (Varouchakis, 2019).

Each data type requires specific modeling approaches. Geostatistical data typically involve spatial interpolation techniques like kriging, while lattice data are often modeled using spatial autoregressive models. (Cressie ,1993) explained the necessity of distinguishing among these data types to avoid misapplication of statistical methods.

2. Spatial Autocorrelation and Stationarity

Spatial autocorrelation quantifies the degree of similarity between spatial observations as a function of distance. Moran's I is a widely used measure:

$$\mathrm{I} = \frac{n}{w} \cdot \frac{\sum_i \sum_j \omega_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

Where ω_{ij} is the spatial weight between locations i and j , and . $W = \sum_i \sum_j \omega_{ij}$ A positive Moran's indicates clustering, while a negative value suggests dispersion.

Stationarity assumes that the statistical properties of a spatial process, such as mean and variance, do not change over space. This assumption is fundamental in many geostatistical models but is often violated in practice, requiring more flexible methods (Morris, 2022).

3. Challenges in Modeling Spatial Data

Modeling spatial data involves several statistical challenges. First, spatial autocorrelation complicates the use of standard regression and inferential techniques that assume independence. Second, heteroskedasticity, or spatially varying variance, can bias parameter estimates. Third, anisotropy—the directional dependence of spatial processes—can lead to inaccurate models if not addressed.

Moreover, missing data and irregular sampling designs further complicate spatial modeling. High-dimensional spatial datasets (e.g., satellite imagery) also pose computational challenges due to the size of covariance matrices involved. These challenges demand specialized methods such as kriging, spatial autoregression, and hierarchical Bayesian modeling (Cressie, 1993; Banerjee, Carlin, & Gelfand, 2014).

Fundamentals of Kriging

1. Kriging as Best Linear Unbiased Predictor

Kriging is a geostatistical interpolation technique that offers the best linear unbiased prediction (BLUP) of unknown spatial values. It utilizes spatial correlation structures to estimate values at unsampled locations. The kriging estimator is given by:

$$\hat{z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i)$$

subject to the constraint,$\sum \lambda_i = 1$ where $Z(s_i)$ are observed values, and λ_i are weights calculated to minimize estimation variance. This technique considers both the distance and the degree of variation between known data points, making it superior to simpler interpolation methods (Matheron, 1963).

2. Variogram and Spatial Dependence Modeling

A critical component in kriging is the variogram, which describes spatial dependence as a function of distance. The empirical variogram is defined as:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(s_i) - Z(s_i + h)]^2$$

Key components include the nugget (measurement error), sill (total variance), and range (distance beyond which spatial correlation vanishes). Accurate variogram modeling ensures reliable kriging results. (Cressie ,1993) emphasized the need for careful selection of variogram models—such as spherical, exponential, or Gaussian—depending on spatial characteristics.

3. Types of Kriging

There are several forms of kriging, each suited to different assumptions about the mean of the spatial process:

- Simple kriging: assumes known constant mean
- Ordinary kriging: assumes unknown but constant mean
- Universal kriging: accounts for deterministic trends in the data

Universal kriging incorporates a trend model:

$$Z(s) = m(s) + \epsilon(s), \qquad m(s) = \beta_0 + \beta_{1x} + \beta_{2y}$$

Where $\epsilon(s)$ is the spatially correlated residual. This flexibility makes universal kriging suitable for non-stationary environments (Isaaks & Srivastava, 1989).

4. Kriging System of Equations

The kriging system is derived from minimizing estimation variance under unbiasedness constraints, leading to a system of linear equations:

$$\sum_{j=1}^{n} \lambda_j \gamma(s_i - s_j) + \mu = \gamma(s_i - s_0), \ \sum_{j=1}^{n} \lambda_i = 1$$

Solving this system yields optimal kriging weights λ_j. The Lagrange multiplier enforces the unbiasedness constraint. (Wackernagel,2003) showed that solving the kriging system becomes computationally intensive as data size increases, prompting approximations for large datasets.

Advanced Statistical Challenges in Kriging

1. Non-Stationarity and Trend Modeling

In practical spatial data analysis, the assumption of stationarity—constant mean and variance over space—is often violated. This poses a significant challenge for traditional kriging methods. To address this, universal kriging allows for a non-constant mean by modeling a spatial trend function:

$$Z(s) = m(s) + \epsilon(s), \qquad m(s) = \beta_0 + \beta_{1x} + \beta_{2y}$$

Here, $\epsilon(s)$is a zero-mean stationary stochastic process. By including a deterministic trend component $m(s)$, universal kriging separates the large-scale variation from the small-scale fluctuations. This flexibility improves model performance in heterogeneous spatial fields (Chilès & Delfiner, 2012).

2. Anisotropy in Spatial Processes

Anisotropy refers to directional dependence in spatial correlation. In isotropic models, spatial correlation depends only on distance, while in anisotropic models it depends on both distance and direction. An anisotropic variogram may be expressed as:

$$\gamma(h, 0) = \gamma_0 + \gamma_1(1 - e^{-h_\theta/a})$$

Where h_0 is the distance along direction θ , and a is the range parameter. Accounting for anisotropy improves the accuracy of kriging predictions, especially in geological or hydrological contexts where physical structures exhibit directionally varying influence (Goovaerts, 1997).

3. Computational Complexity of Kriging

Kriging involves inverting an $n \times n$ covariance matrix, which results in computational complexity of $o\,(n^3)$. For large datasets, this is infeasible. Solutions include:

- Approximate kriging via spatial partitioning
- Low-rank methods using basis functions
- Covariance tapering and sparse matrix techniques

For instance, using predictive processes, the spatial process is projected onto a lower-dimensional subspace, reducing computational cost:

$$Z(s) = \sum_{k=1}^{r} \emptyset_{\,k}\,(s)\eta_{\,k}$$

Where $\phi_{\,k}$ are basis functions and are random coefficients (Banerjee et al., 2008).

4. Model Validation and Cross-Validation

Model validation is crucial in assessing the performance of kriging models. Cross-validation involves omitting each data point in turn, predicting its value, and computing the prediction error:

$$CV = \frac{1}{n}\sum_{i=1}^{n}[z(s_{\,i}) - \hat{z}_{-i}(s_i)]^2$$

This method helps in selecting variogram models, tuning kriging parameters, and detecting overfitting. (Zimmerman ,2006) noted that effective validation techniques improve confidence in spatial predictions and guide model refinement.

Applications and Future Directions

1. **Environmental and Climate Modeling**

Kriging plays a key role in environmental sciences, such as interpolating pollutant concentrations or temperature fields. For example, estimating PM2.5 concentrations across a region can be achieved using ordinary kriging based on sparse monitoring data. The accuracy of spatial exposure estimates significantly impacts epidemiological studies (Jerrett et al., 2005).

2. **Natural Resource and Mining Applications**

In mining, kriging is used to estimate ore grades and guide resource extraction. Block kriging is commonly employed to predict average values over blocks of material. The method supports risk assessment and economic evaluation of mineral reserves. The kriging variance informs decision-making under uncertainty (Deutsch & Journel, 1998).

3. **Integration with Machine Learning**

Kriging is increasingly integrated with machine learning models to capture nonlinear and non-stationary spatial patterns. Techniques such as random forest kriging or neural network kriging leverage both spatial structure and complex covariate relationships:

$$\hat{z}(s_0) = f\big(x(s_0)\big) + \sum \lambda_i [Z(s_i) - f\big(x(s_i)\big)]$$

Where f is a machine learning model. These hybrid approaches improve prediction accuracy and interpretability (Hengl et al., 2018).

4. **Big Spatial Data and Future Research**

The growth of spatial big data—such as high-resolution satellite imagery and sensor networks—necessitates scalable kriging algorithms. Emerging methods include:

- Distributed kriging
- Parallel variogram computation
- Gaussian Markov random fields (GMRF) on grids

Heaton et al. (2019) demonstrated that scalable spatial models can achieve near-optimal prediction while drastically reducing computational costs. Future research will focus on uncertainty quantification, real-time spatial analytics, and cloud-based geostatistical systems.

Practical Component: Empirical Kriging Analysis

1. **Study Design and Objective**

This study aims to apply kriging techniques to interpolate air quality measurements (PM2.5 concentration) across a metropolitan region using a geostatistical approach. The objective is to evaluate the accuracy and efficiency of different kriging methods under real-world spatial complexities including anisotropy and non-stationarity.

2. Dataset Description

We used publicly available air quality data from 75 monitoring stations located within an urban area. Each observation includes:

- Coordinates: (x, y)
- Measured PM2.5 concentrations (μg/m³)
- Time: All observations taken on the same date to ensure stationarity in time.

The spatial extent is a 60 km × 60 km region with irregular station distribution, requiring robust spatial interpolation.

3. Software and Tools

- R software (version 4.3.0)
- Packages used: gstat, sp, sf, automap, spdep

Workflow Steps:

1. Exploratory Data Analysis (EDA)
2. Empirical variogram computation
3. Model fitting (spherical, exponential)
4. Ordinary and Universal Kriging
5. Cross-validation (Leave-One-Out)

4. Variogram Modeling and Fitting

The empirical variogram was computed using:

$$\gamma(h) = \frac{1}{2N\ (h)} \sum_{i=1}^{N(h)} [Z(s_i) - Z(s_i + h)]^2$$

Models tested:

- Spherical model (best fit): Nugget = 1.1, Sill = 15.4, Range = 19.6 km
- Anisotropy detected in NE–SW direction (angle = 45°)

5. Kriging Implementation

Ordinary Kriging (OK):

$$\hat{z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i), \sum \lambda_i = 1$$

Universal Kriging (UK): Assuming a linear trend:

$$Z(s) = m(s) + \epsilon(s), \qquad m(s) = \beta_0 + \beta_{1x} + \beta_{2y}$$

Both models showed smooth spatial predictions, but UK better captured urban–rural gradients.

6. Cross-Validation Results

Using Leave-One-Out Cross-Validation:

- **OK RMSE** = 4.28 μg/m³
- **UK RMSE** = 3.92 μg/m³
- **Mean error (bias):** near zero for both
- **R^2 (UK):** 0.71, higher than OK (0.66)

7. Visualization and Interpretation

- Prediction maps clearly identified pollution hotspots near industrial zones.

- Kriging variance maps indicated higher uncertainty in peripheral areas with sparse data.
- Directional variogram confirmed anisotropy, reinforcing the need for directional modeling.

Discussion

The analysis confirmed the following:

- Spatial dependence can be efficiently modeled with kriging.
- Non-stationary models (UK) provided more accurate predictions.
- Anisotropy significantly improved variogram modeling and prediction accuracy.
- Cross-validation is essential for model evaluation.

This application demonstrates how kriging adapts to real spatial structures and supports informed decision-making in environmental monitoring.

Conclusion

Based on the practical part of the study, the results reveal that modeling spatial dependency and correctly forecasting environmental data like air quality measurements are quite well done using Kriging methods. With UK providing more accurate results—especially in capturing urbanrural gradients—the study revealed that both Ordinary Kriging (OK) and Universal Kriging (UK) created smooth spatial predictions. This underlines how critical it is to use nonstationary models like Universal Kriging in actual situations when spatial data display different patterns.

Furthermore, the study verified the substantial influence of anisotropy in spatial modeling as the variogram displayed directional dependency in the data, therefore enhancing the predictability accuracy when correctly considered for. With Universal Kriging beating Ordinary Kriging in terms of RMSE and R^2, the crossvalidation results further confirmed the reliability of the models and underlined the need of improving model parameters by crossvalidation.

Finally, kriging models offer insightful observations for environmental monitoring when used with the appropriate modeling decisions and verification techniques. Highlighting how Kriging, especially with the integration of nonstationary models and anisotropy consideration, can greatly improve the precision of spatial forecasts provides important support for decisionmaking in environmental management and pollution monitoring.

References

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical Analysis*, 27(2), 93–115.

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data* (2nd ed.). CRC Press.

Banerjee, S., Gelfand, A. E., Finley, A. O., & Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4), 825–848.

Bivand, R. S., Pebesma, E., & Gómez-Rubio, V. (2013). *Applied spatial data analysis with R*. Springer.

Chilès, J. P., & Delfiner, P. (2012). *Geostatistics: Modeling spatial uncertainty* (2nd ed.). Wiley.

Cressie, N. (1993). *Statistics for spatial data* (Revised ed.). Wiley.

Cressie, N. (1993). *Statistics for spatial data*. Wiley.

Deutsch, C. V., & Journel, A. G. (1998). *GSLIB: Geostatistical software library and user's guide* (2nd ed.). Oxford University Press.

Goovaerts, P. (1997). *Geostatistics for natural resources evaluation*. Oxford University Press.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guhaniyogi, R., Gerber, F., ... & Zimmerman, D. L. (2019). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 24(3), 398–425.

Hengl, T., Heuvelink, G. B., & Stein, A. (2004). A generic framework for spatial prediction. *International Journal of Geographical Information Science*, 18(3), 221–252.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, 6, e5518.

Isaaks, E. H., & Srivastava, R. M. (1989). *An introduction to applied geostatistics*. Oxford University Press.

Jerrett, M., Burnett, R. T., Ma, R., Pope III, C. A., Krewski, D., Newbold, K. B., ... & Thun, M. J. (2005). Spatial analysis of air pollution and mortality in Los Angeles. *Epidemiology*, 16(6), 727–736.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58(8), 1246–1266.

Morris, L. (2022). Spatio-temporal modelling for nonstationary point referenced data.

Pebesma, E. J., & Wesseling, C. G. (1998). Gstat: A program for geostatistical modelling, prediction and simulation. *Computers & Geosciences*, 24(1), 17–31.

Varouchakis, E. A. (2019). *Mathematical and statistical basis*.

Wackernagel, H. (2003). *Multivariate geostatistics: An introduction with applications* (3rd ed.). Springer.

Zimmerman, D. L. (2006). Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. *Environmetrics*, 17(6), 635–652.