# Spatial Statistics Notes

Brian Cervantes Alvarez

2025-02-04

## Abstract

Spatial statistics is a crucial field that focuses on analyzing spatially referenced data, incorporating geographic and locational information to model relationships and dependencies. These notes cover foundational concepts such as spatial data types, coordinate reference systems, and spatial dependence, along with advanced topics including geostatistical models, spatial regression, and simultaneous autoregressive (SAR) models. Key topics include the multivariate normal distribution, variograms, kriging techniques, and methods for handling spatially structured big data, such as Nearest-Neighbor Gaussian Processes (NNGP). The notes also delve into areal data analysis, spatial autocorrelation measures like Moran's I and Geary's C, and model-based approaches for predicting spatial outcomes. A strong emphasis is placed on practical implementation, using R packages such as `sp`, `sf`, `gstat`, `spmodel`, and `spNNGP` for real-world spatial data analysis. Applications span environmental monitoring, urban planning, epidemiology, and more, demonstrating the critical role of spatial statistics in decision-making and scientific research.

# Table of contents

# Chapter 1

# Introduction to Spatial Statistics

## 1.1 Preliminaries

### 1.1.1 Tobler's First Law of Geography

"Everything is related to everything else, but near things are more related than distant things." — Tobler (1970)

**Why does this matter?** Imagine you're studying temperature changes across a region. If two locations are close, their temperatures are likely more similar than two locations far apart. This principle is fundamental in spatial analysis, affecting how we model relationships between data points.

### 1.1.2 Types of Spatial Data

1. **Geostatistical Data**: Continuous spatial variables, e.g., soil pollution measurements.
2. **Areal (Lattice) Data**: Aggregated over regions, e.g., census data by county.
3. **Point Pattern Data**: Locations of events, e.g., earthquake occurrences.

**Example:** Consider air quality measurements across a city. If measured at scattered sensor locations, it's geostatistical data. If summarized at the neighborhood level, it's areal data. If recording only pollution source locations, it's point pattern data.

## 1.2 Spatial Coordinate Systems

A coordinate system provides a mathematical way to specify locations. Important terms:

- **Datum**: Defines the origin and scale of a coordinate system.
- **Geodetic Datum**: Relates the coordinate system to Earth's shape.
- **Coordinate Reference System (CRS)**: Specifies how a dataset's coordinates relate to the real world.

**Analogy:** Think of CRS like different map apps—Google Maps, Apple Maps, and OpenStreetMap all display the same Earth, but their coordinate systems may differ slightly.

**Example:** The Universal Transverse Mercator (UTM) system divides the world into zones, making local measurements more accurate.

# Chapter 2

# Multivariate Normal Distributions

## 2.1 Bivariate Normal Distribution

A two-variable generalization of the normal distribution, described by:

- **Means** $(\mu_1, \mu_2)$: Expected values.
- **Variances** $(\sigma_1^2, \sigma_2^2)$: Spread of each variable.
- **Correlation** $(\rho)$: Strength of their relationship.

**Why does this matter?** Many spatial models assume normality, making this distribution crucial for prediction and inference.

**Example:** If you measure temperature and humidity across a city, they likely follow a bivariate normal distribution, where hotter days tend to be more humid.

## 2.2 Variance-Covariance Matrices

For multiple variables, we use a variance-covariance matrix:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \text{cov}(Y_1, Y_2) \\ \text{cov}(Y_1, Y_2) & \sigma_2^2 \end{bmatrix}$$

**Properties:**

- Always symmetric ($\text{cov}(Y_1, Y_2) = \text{cov}(Y_2, Y_1)$).
- Positive definite (ensures meaningful variability).

**Analogy:** Think of covariance like how two stocks move together—if one rises when the other does, their covariance is positive.

## 2.3 Multivariate Normal Distribution

A generalization for multiple correlated variables, defined by a mean vector and covariance matrix:

$$f(y) = (2\pi)^{-n/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\right)$$

**Example:** If you analyze rainfall, temperature, and humidity together, they likely follow a multivariate normal distribution, helping in climate modeling and weather prediction.

# Chapter 3

# More on Variograms; Spatial Prediction With Covariates

## 3.1 Practical Range and Matérn Variogram Model

The **practical range** is the distance at which covariance drops to 5% of the partial sill, varying by model type.

**Why does this matter?** It helps determine the spatial scale of correlation, guiding interpolation decisions.

**Example:** In environmental science, it informs sensor placement for air pollution monitoring.

### 3.1.1 Matérn Variogram Model

The Matérn model introduces flexibility via the **smoothness parameter ( )**, controlling how covariance decays.

**Analogy:** Think of like road conditions—higher values make the transition from smooth to rough more gradual.

## 3.2 Universal Kriging

**Universal kriging** extends ordinary kriging by incorporating covariates like elevation.

**Why does this matter?** It allows for better predictions when spatial trends exist.

**Example:** Predicting soil contamination based on both spatial location and land elevation.

# Chapter 4

# Regression With Geostatistical Data

## 4.1 Prediction vs. Estimation

- **Prediction**: Forecasts values at unsampled locations.
- **Estimation**: Determines underlying relationships between variables.

**Example:** Predicting temperature at an unmeasured site versus estimating how temperature depends on elevation.

## 4.2 Spatial Regression

Spatial regression extends traditional models by accounting for spatial dependence.

### 4.2.1 Key Model Differences:

1. **Ordinary Regression** assumes independence.
2. **Spatial Regression** incorporates a spatial covariance structure.

**Example:** Modeling heavy metal contamination while accounting for proximity effects.

### 4.2.2 Practical Implementation

- `lm()` for independent models.
- `splm()` for spatial models, estimating covariance parameters.

**Why does this matter?** Spatial regression reduces bias and improves inference in geographically structured data.

# Chapter 5

# Simultaneous Autoregressive (SAR) Models

## 5.1 Row-Standardized Weights

SAR models rely on spatial weight matrices to define relationships between spatial units.

**Example:** Consider housing prices in a city—neighboring properties often influence each other. SAR models capture this dependency through a structured matrix.

**Types of Contiguity:**

- **Queen contiguity:** A single shared boundary point defines a neighborhood.
- **Binary weights:** Assigns 1 if two areas share a boundary and 0 otherwise.
- **Row-standardized weights:** Adjusts for different numbers of neighbors to ensure interpretability.

**Analogy:** Think of binary weights like direct friendships, whereas row-standardized weights are like adjusting for popularity—someone with many friends has each connection weighted less.

## 5.2 Simultaneous Autoregressive (SAR) Models

SAR models incorporate spatial dependence by modifying regression equations:

$$Z = X\beta + B(Z - X\beta) + d$$

where:

- $B = \phi W$ describes spatial relationships,
- $W$ is the spatial weights matrix,
- $d$ is the error term.

### 5.2.1 Variance-Covariance Structure

The SAR model leads to:

$$\Sigma_{SAR} = \sigma^2_{de}[(I - \phi W)(I - \phi W)^T]^{-1}$$

This accounts for spatial correlation in the dependent variable.

**Why does this matter?** Ordinary regression assumes independence, but ignoring spatial dependence can lead to biased estimates. SAR models correct this.

## 5.3 Implementing SAR Models in R

Using the `spmodel` package:

```
library(spmodel)
boston_sar <- spautor(MEDV ~ 1, data = boston_sf, spcov_type = "sar", row_s
 ↪  = FALSE)
summary(boston_sar)
```

**Interpretation**

- **Intercept:** Represents the average property value.
- **SAR spatial covariance parameters:** Indicate spatial dependence strength.

## 5.4 Row-Standardization and Its Effect

Row-standardization ensures each row sums to 1, making comparisons fair across regions:

```
boston_sar <- spautor(MEDV ~ 1, data = boston_sf, spcov_type = "sar", row_s
 ↪  = TRUE)
summary(boston_sar)
```

**Example:** The impact of neighbors is scaled relative to the number of connections, preventing dominant effects from highly connected regions.

**Analogy:** Think of row-standardization like adjusting for social media influence—someone with 1,000 friends has less influence per friend compared to someone with just 10.

**Why does this matter?** Row-standardization helps maintain stability in the model, particularly for regions with varying numbers of neighbors.

# Chapter 6

# Additional Topics

## 6.1 Spatial Point Processes

Spatial point processes model the random occurrence of events in space. They're crucial when the data consist of discrete event locations—think tracking the spread of disease outbreaks or modeling the locations of trees in a forest.

### 6.1.1 Poisson Point Process

The **homogeneous Poisson point process** assumes that events occur completely at random with a constant intensity $\lambda$ over the study region. The number of events in a region of area $A$ follows a Poisson distribution:

$$P(N(A) = k) = \frac{(\lambda A)^k e^{-\lambda A}}{k!}$$

**Example:** If you're tracking meteorite impacts over a vast area, you might assume they occur randomly following a Poisson process.

#### 6.1.1.1 R Implementation

```r
library(spatstat)

# Define a rectangular window for the study area

win <- owin(xrange = c(0, 100), yrange = c(0, 100))

# Simulate a homogeneous Poisson process with intensity 0.05 events per unit

 ↪  area

pp_poisson <- rpoispp(lambda = 0.05, win = win)

plot(pp_poisson, main = "Homogeneous Poisson Point Process")
```

### 6.1.2 Cluster and Inhibition Processes

- **Cluster Processes:** Events tend to occur in groups. For instance, disease cases might cluster in certain neighborhoods due to environmental factors.
- **Inhibition Processes:** Events tend to repel each other (e.g., trees that compete for resources).

**Analogy:** Imagine a party—people might cluster around a snack table (cluster process) or spread out on the dance floor to give everyone room (inhibition process).

## 6.2 Bayesian Spatial Models

Bayesian approaches allow you to incorporate prior knowledge into spatial models, which is especially useful in data-scarce or highly uncertain environments.

### 6.2.1 Hierarchical Bayesian Spatial Models

In a hierarchical setup, spatial data are modeled on multiple levels. For example, the observed data might depend on latent spatial random effects that capture unmeasured spatial variation:

$$Y(s) = X(s)\beta + w(s) + \epsilon(s)$$

where: - $w(s)$ is a latent spatial process (often modeled as a Gaussian Process), - $\epsilon(s)$ is independent error.

**Why does this matter?** By explicitly modeling spatial random effects, you capture dependencies that traditional models might miss, leading to improved uncertainty quantification.

#### 6.2.1.1 R Implementation with spBayes

```r
library(spBayes)
# Assume we have spatial data with coordinates 'coords' and response 'Y'
# Define a spatial decay parameter and variance components
starting <- list("phi" = 3, "sigma.sq" = 1, "tau.sq" = 0.1)
tuning   <- list("phi" = 0.1, "sigma.sq" = 0.1, "tau.sq" = 0.05)
priors   <- list("phi.Unif" = c(0.1, 10), "sigma.sq.IG" = c(2, 1),
 ↪  "tau.sq.IG" = c(2, 0.1))
# Fit a simple spatial regression model
bayes_model <- spLM(Y ~ X1 + X2,
                    data = your_data,
                    coords = your_data[, c("x", "y")],
                    starting = starting,
                    tuning = tuning,
                    priors = priors,
                    cov.model = "exponential",
                    n.samples = 1000)
summary(bayes_model)
```

**Analogy:** Think of Bayesian models as using “wisdom from the past” (priors) to help guide predictions when new data are sparse or noisy.

## 6.3 Nearest-Neighbor Gaussian Processes (NNGP)

For large spatial datasets, traditional Gaussian Process models can be computationally prohibitive because of the inversion of large covariance matrices. **NNGP** approximates the full Gaussian Process by assuming that each location is conditionally independent given its nearest neighbors.

### 6.3.1 Key Benefits

- **Scalability:** Efficiently handles datasets with thousands (or more) spatial locations.
- **Flexibility:** Retains much of the interpretability of full Gaussian Processes while reducing computational overhead.

**Example:** In urban planning, when analyzing city-wide sensor data (e.g., air quality), NNGP allows you to process massive datasets quickly without sacrificing much accuracy.

#### 6.3.1.1 R Implementation with spNNGP

```r
library(spNNGP)

# Assume we have data 'your_data' with coordinates and response 'Y'

coords <- as.matrix(your_data[, c("x", "y")])

# Define priors and starting values for the model parameters

starting <- list("phi" = 3, "sigma.sq" = 1, "tau.sq" = 0.1)

priors   <- list("phi.Unif" = c(0.1, 10), "sigma.sq.IG" = c(2, 1),
 ↪  "tau.sq.IG" = c(2, 0.1))

# Fit the NNGP model using 10 nearest neighbors

nngp_model <- spNNGP(Y ~ X1 + X2,

                     data = your_data,

                     coords = coords,

                     starting = starting,

                     priors = priors,

                     n.neighbors = 10,

                     cov.model = "exponential",

                     n.samples = 1000,

                     n.threads = 2)

summary(nngp_model)
```

**Analogy:** Think of NNGP like a squad of Gen Z influencers—each location only "listens" to its closest peers rather than the entire network, making the model both trendy and computationally sleek.