Spatial Statistics

January 14, 2024

Centre for Modern Beamer Themes

Spatial Stochastic Process/ Spatial Random Field**

A spatial stochastic process is a family of random variables

$$\{Z(s) : s \in D\}$$

indexed by spatial locastions $s \in D$.

D: Spatial domain (the geographical region in which observations could made)

$Z(s)$: Random variable representing the quantity that you measure at location s

Temporal Stochastic Process

A collection of random variables $\{X_t : t \in T\}$ or $\{X(t) : t \in T\}$ where T is an index set. For each $t \in T$, X_t or $X(t)$ isa random variable.

Three types of spatial data

1. Geostatistical processes

Example: Maximum temperature in Colombo District

2. Areal processes

Example: Dengue cases in each district in Sri Lanka

3. Point processes

Example: Location of dengue patients household addresses

Geostatistical processes

A geostatistical process is the stochastic process

$$\{Z(\mathbf{s}) : \mathbf{s} \in D\}$$

where D is a fixed subset of the p-dimensional space $\mathbb{R}^p$. The locations s at which data could occur vary **continuously** over D. In other words, it is possible to measure at infinitely many locations across the spatial domain D.

Areal unit process/ Lattice process

The spatial domain D is partitioned into n disjoint areal units which are denoted by

$$D = \{B_1, B_2, ..., B_n\}$$

.

The areal stochastic process is denoted by

$$Z = \{Z(B_1), Z(B_2), ..., Z(B_n)\}$$

.

Alternative formulation of areal unit processes

Let $s_1, s_2...., s_n$ be the centroids of $B_1, B_2, ..., B_n$. THen the areal stochastic process is denoted by

$$Z = \{Z(s_1), Z(s_2), ...Z(s_n)\}.$$

Point Stochastic Process

Let

$$D = \{A_1, A_2, ...A_n\}$$

, where n denotes the number of points in D. Then the stochastic process is

$$Z = \{Z(A_1), Z(A_2), ...Z(A_n)\}.$$

Goals of spatial analysis

- To find a statistical model that adequately explains the spatial dependency structure and trends, etc.
- Interpolation
- To make inferences
- To model the relationship between covariates and response

Geostatistical stochastic process

A geostatistical process is the stochastic process

$$\{Z(\mathbf{s}) : \mathbf{s} \in D\}$$

where D is a fixed subset of the p-dimensional space $\mathbb{R}^p$. The locations $\mathbf{s}$ at which data could occur vary **continuously** over D. In other words, it is possible to measure at infinitely many locations across the spatial domain D.

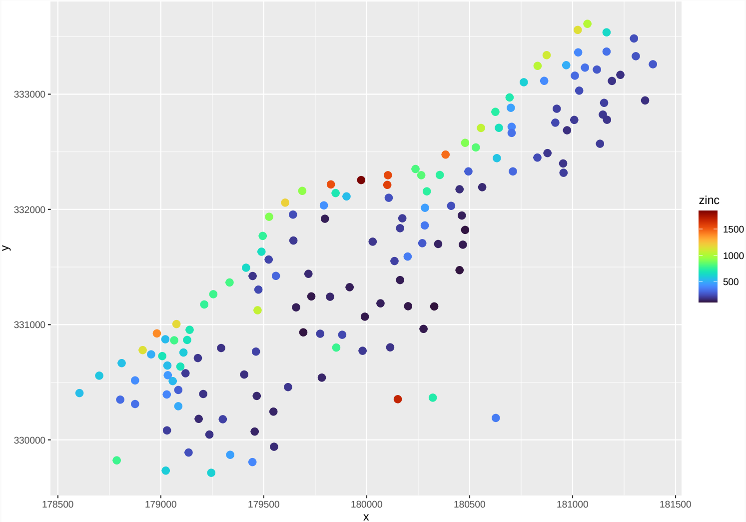
In this course, we focus on $p = 2$. That is, a location $\mathbf{s} = (s_1, s_2)$. For example, s_1 and s_2 could be longitude and latitude.

Meuse river data set

This data set gives locations and topsoil heavy metal concentrations, along with a number of soil and landscape variables at the observation locations, collected in a flood plain of the river Meuse, near the village of Stein (NL).

x	y	cadmium	copper	lead	zinc	elev	dist
181072	333611	11.7	85	299	1022	7.909	0.00135
181025	333558	8.6	81	277	1141	6.983	0.01222
181165	333537	6.5	68	199	640	7.800	0.10302

EDA



333000
332000
331000
330000
y
178500
179000
179500
180000
180500
181000
181500
x
zinc
1500
1000
500

Mean function

The mean function of $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$ is

Continuous random variable

$$\mu(s) = E[Z(\boldsymbol{s})] = \int_{-\infty}^{\infty} z f_Z z(\boldsymbol{s}) dz$$

where $f_Z z(\boldsymbol{s})$ is the probability density function of $Z(\boldsymbol{s})$.

Discrete random variable

$$\mu(s) = E[Z(\boldsymbol{s})] = \sum_{z_i \in S} z_i f_Z z(\boldsymbol{s})$$

where $f_Z z(\boldsymbol{s})$ is the probability mass function for $Z(\boldsymbol{s})$.

Autocovariance function

$$C(\mathbf{s}, \mathbf{t}) = Cov[Z(\mathbf{s}), Z(\mathbf{t})]$$

Measures the linear dependence between $Z(s)$ and $Z(t)$.

Variance function

$$V[Z(\mathbf{s})] = C(\mathbf{s}, \mathbf{s}) = \nu^2(s)$$

Theorems

1. The autocovariance function is symmetric in its arguments. That is, $C(\mathbf{s}, \mathbf{t}) = C(\mathbf{t}, \mathbf{s})$ for each $\mathbf{s}, \mathbf{t} \in D$.
2. The autocovariance function $C(\mathbf{s}, \mathbf{t})$ is a nonnegative definite function.

Autocovariance function

$$\rho(\mathbf{s}, \mathbf{t}) = Corr[Z(\mathbf{s}), Z(\mathbf{t})] = \frac{C(\mathbf{s}, \mathbf{t})}{\sqrt{C(\mathbf{s}, \mathbf{s})C(\mathbf{t}, \mathbf{t})}}$$

Properties of autocorrelation function: In class

White noise process

1. $\mu(\mathbf{s}) = \mu$ for all $\mathbf{s} \in D$
2. $$C(\mathbf{s}, \mathbf{t}) = \begin{cases} \tau^2, & \text{if } \mathbf{s} = \mathbf{t}. \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Strictly Stationary

A geostatistical process $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$ is strictly stationary if

$$f(Z(\mathbf{s}_1), ..., Z(\mathbf{s}_n)) = f(Z(\mathbf{s}_1 + h), ..., Z(\mathbf{s}_n + h))$$

for any displacement vector h and any set of n locations $\{\mathbf{s}_1, ..., \mathbf{s}_n\}$. This means, the joint distribution of a set of random variables are unaffected by spatial shifts.

Weakly stationarity

1. $E[Z(\mathbf{s})] = \mu(\mathbf{s}) = \mu$ for some finite constant μ which does not depend on $\mathbf{s}$.
2. $Cov[Z(\mathbf{s}), Z(\mathbf{s+h})] = C(\mathbf{s}, \mathbf{s+h}) = C(h)$

Here h is called the spatial lag or displacement.

Note: Strictly stationary implies it is weakly stationary, but the converse is not true in general (unless $Z(\mathbf{s})$ is a Gaussian process).

Isotropic

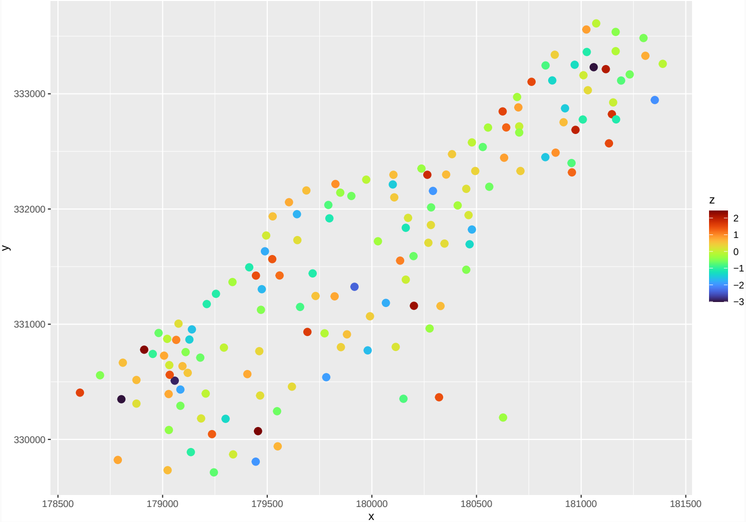
This means that the correlation between any two observations depends only on the distance between those locations and not on their relative orientation. There is no directional influence.

Spatial Continuity

Spatial continuity: Correlation between values over distance

No spatial continuity

Random values at each location



Perfect spatial continuity

Variogram

- Used to check if there is any spatial autocorrelation in the data.

Semi-variogram

$$\gamma(\mathbf{s}, \mathbf{t}) = \frac{1}{2} Var[z(\mathbf{s}) - z(\mathbf{t})]$$

Task

Show that, when the process has constant mean $\mu(s) = \mu$

$$\gamma(\mathbf{s}, \mathbf{t}) = \frac{1}{2} E[z(\mathbf{s}) - z(\mathbf{t})]^2$$

Proof: in-class

Variogram calculation

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

Important results

$$\gamma(\mathbf{h}) = \nu^2 - C(\mathbf{h})$$

Proof: In-class