Geostatistics

Random Fields

Nicolas Desassis
freely adapted from Thomas Romary's course

nicolas.desassis@minesparis.psl.eu
Mines Paris
PSL University
Centre for geosciences and geoengineering

19 November 2024

Introduction

- A regionalized phenomenon is a phenomenon that presents a structure in the space and or the time
- It is characterized by a mathematical function, called regionalized variable, defined in every point of a domain of study $\mathcal{X}$:

$$z(x), x \in \mathcal{X} \subset \mathbb{R}^d$$

Example: Phenomenon: mountain, variable: altitude

Introduction

- To describe and understand a phenomenon, a model is needed
- The deterministic approach is based on a phenomenological modeling
  - requires perfect knowledge of the mathematical/physical laws that govern the phenomenon and the initial/limit conditions
  - not likely to honor the data
- In general, the phenomenon under study is too complex

  $\Longrightarrow$ Probabilistic approach

The probabilistic approach in geostatistics

Idea

Build a model integrating both knowledge and uncertainties about the phenomenenon

The probabilistic approach in geostatistics

Idea

Build a model integrating both knowledge and uncertainties about the phenomenenon

$\Longrightarrow$ A "classical" statistical model?

The probabilistic approach in geostatistics

Idea

Build a model integrating both knowledge and uncertainties about the phenomenenon

$\Longrightarrow$ A "classical" statistical model?

Problem

The observations of the regionalized variable cannot be considered as independent and identically distributed in a regionalized phenomenon

- The observations are not always measured in the same condition (support, captors,...)
- Intuitively, in close locations, the regionalized variable should take similar values

Solution: Probabilistic geostatistics

The regionalized variable will be considered to be a realization of a random field (or random function)

Formalization

Fundamental hypothesis

$\exists \omega \in \Omega, \forall x \in \mathcal{X}$

$$z(x) = Z(\omega, x)$$

Consequence

- The values taken by the regionalized variable on $\mathcal{X}$ can be seen as drawn from an infinite family of random variables, called the random field (RF) or random function.
- The random field model will explain the structure of the data
- In virtue of the unicity of the realization of the regionalized variable, the RF model should be ergodic in the sense where all its structural parameters can be inferred from a unique realization

Characteristisation of a random field

Spatial distribution

A Random Field is characterized by its spatial distribution, that is the knowledge, $\forall n \in \mathbb{N}^\star, \forall x_1, \ldots, x_n \in \mathcal{X}$, of the functions:

$$(z_1, \ldots, z_n) \in \mathbb{R}^n \mapsto \mathbb{P}(Z(x_1) \leq z_1, \ldots, Z(x_n) \leq z_n)$$

Characteristisation of a random field

Spatial distribution

A Random Field is characterized by its spatial distribution, that is the knowledge, $\forall n \in \mathbb{N}^\star, \forall x_1, \dots, x_n \in \mathcal{X}$, of the functions:

$$(z_1, \dots, z_n) \in \mathbb{R}^n \mapsto \mathbb{P}(Z(x_1) \leq z_1, \dots, Z(x_n) \leq z_n)$$

- In practice, only a few models exist such that the spatial distribution is completely known (e.g. Gaussian)
- We rely on several simplifying assumptions (e.g. existence of finite moments, stationarity)

Notion of stationarity

The inference is made difficult because

- We only observe a unique realization of the phenomenon (the world we live in)
- We only know its value on a finite set of locations

Notion of stationarity

The inference is made difficult because

- We only observe a unique realization of the phenomenon (the world we live in)
- We only know its value on a finite set of locations

Hence the notion of stationarity: the phenomenon is sufficiently homogeneous in the space such that we can replace a repetition of the RF by a repetition in space

Notion of stationarity

The inference is made difficult because

- We only observe a unique realization of the phenomenon (the world we live in)
- We only know its value on a finite set of locations

Hence the notion of stationarity: the phenomenon is sufficiently homogeneous in the space such that we can replace a repetition of the RF by a repetition in space

Definition: strict stationarity

A RF is strict stationary when its spatial distribution is translation invariant: $\forall n \in \mathbb{N}^\star, \forall x_1, \ldots, x_n \in \mathcal{X}, (z_1, \ldots, Z_n) \in \mathbb{R}^n$ and $\forall h \in \mathbb{R}$

$$\mathbb{P}(Z(x_1) \leq z_1, \ldots, Z(x_n) \leq z_n) = \mathbb{P}(Z(x_1 + h) \leq z_1, \ldots, Z(x_n + h) \leq z_n)$$

NB: we also talk about strong stationarity

Notion of stationarity

The inference is made difficult because

- We only observe a unique realization of the phenomenon (the world we live in)
- We only know its value on a finite set of locations

Hence the notion of stationarity: the phenomenon is sufficiently homogeneous in the space such that we can replace a repetition of the RF by a repetition in space

Definition: strict stationarity

A RF is strict stationary when its spatial distribution is translation invariant: $\forall n \in \mathbb{N}^\star, \forall x_1, \ldots, x_n \in \mathcal{X}, (z_1, \ldots, Z_n) \in \mathbb{R}^n$ and $\forall h \in \mathbb{R}$

$$\mathbb{P}(Z(x_1) \leq z_1, \ldots, Z(x_n) \leq z_n) = \mathbb{P}(Z(x_1 + h) \leq z_1, \ldots, Z(x_n + h) \leq z_n)$$

NB: we also talk about strong stationarity

Can we work with a weaker assumption?

Covariance and variogram

- Expectation

$$m(x) = E[Z(x)]$$

Covariance and variogram

- Expectation

$$m(x) = E[Z(x)]$$

- Covariance

$$C(x, y) = \mathrm{Cov}(Z(x), Z(y)) = \mathbb{E}[(Z(x) - m(x))(Z(y) - m(y))]$$

Covariance and variogram

- Expectation

$$m(x) = E[Z(x)]$$

- Covariance

$$C(x, y) = \mathrm{Cov}(Z(x), Z(y)) = \mathbb{E}[(Z(x) - m(x))(Z(y) - m(y))]$$

- Property

$$\mathbb{V}\left(\sum_{i=1}^{n} \lambda_i Z(x_i)\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j C(x_i, x_j)$$

Covariance and variogram

- Expectation

$$m(x) = E[Z(x)]$$

- Covariance

$$C(x, y) = \mathrm{Cov}(Z(x), Z(y)) = \mathbb{E}[(Z(x) - m(x))(Z(y) - m(y))]$$

- Property

$$\mathbb{V}\left(\sum_{i=1}^{n} \lambda_i Z(x_i)\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j C(x_i, x_j)$$

- Variogram

$$\begin{aligned} \gamma(x, y) &= \frac{1}{2}\mathbb{V}(Z(x) - Z(y)) \\ &= \frac{C(x, x) + C(y, y)}{2} - C(x, y) \\ &= \frac{\mathbb{V}(Z(x)) + \mathbb{V}(Z(y))}{2} - C(x, y) \\ C(x, y) &= \frac{\mathbb{V}(Z(x)) + \mathbb{V}(Z(y))}{2} - \gamma(x, y) \end{aligned}$$

Covariance and variogram

- Expectation

$$m(x) = E[Z(x)]$$

- Covariance

$$C(x, y) = \mathrm{Cov}(Z(x), Z(y)) = \mathbb{E}[(Z(x) - m(x))(Z(y) - m(y))]$$

- Property

$$\mathbb{V}\left(\sum_{i=1}^{n} \lambda_i Z(x_i)\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j C(x_i, x_j)$$

- Variogram

$$\begin{aligned}\gamma(x, y) &= \frac{1}{2}\mathbb{V}(Z(x) - Z(y)) \\ &= \frac{C(x, x) + C(y, y)}{2} - C(x, y) \\ &= \frac{\mathbb{V}(Z(x)) + \mathbb{V}(Z(y))}{2} - C(x, y)\end{aligned}$$

$$C(x, y) = \frac{\mathbb{V}(Z(x)) + \mathbb{V}(Z(y))}{2} - \gamma(x, y)$$

Remark: Variogram always exists when the covariance exists. But a variogram is not sufficient to define a covariance function.

Requirements

We need to find a function C such as

- $\forall n \in \mathbb{N}^\star$
- $\forall x_1, \ldots, x_n \in \mathbb{R}^d$
- $\forall \lambda_1, \ldots, \lambda_n \in \mathbb{R}$

$$\mathbb{V}\left(\sum_{i=1}^{n} \lambda_i Z(x_i)\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j C(x_i, x_j) \geq 0$$

In other words, any covariance matrix $(C(x_i, x_j))_{i,j}$ built with C has to be (semi) positive-definite.

How to choose C?

Second-order stationarity

A random field Z over $\mathcal{D} \subset \mathbb{R}^d$ is second-order stationary if for all x, y in $\mathcal{D}$, the following properties hold:

- $\mathbb{E}[Z(x)] = m < \infty$
- $\mathbb{V}(Z(x)) < \infty$
- $C(x, y) = f(x - y)$ (only depends on $x - y$)

How to choose C?

Second-order stationarity

A random field Z over $\mathcal{D} \subset \mathbb{R}^d$ is second-order stationary if for all x, y in $\mathcal{D}$, the following properties hold:

- $\mathbb{E}[Z(x)] = m < \infty$
- $\mathbb{V}(Z(x)) < \infty$
- $C(x, y) = f(x - y)$ (only depends on $x - y$)

Consequence $\mathbb{V}(Z(x)) = f(0)$ is constant

How to choose C?

Second-order stationarity

A random field Z over $\mathcal{D} \subset \mathbb{R}^d$ is second-order stationary if for all x, y in $\mathcal{D}$, the following properties hold:

- $\mathbb{E}[Z(x)] = m < \infty$
- $\mathbb{V}(Z(x)) < \infty$
- $C(x, y) = f(x - y)$ (only depends on $x - y$)

Consequence $\mathbb{V}(Z(x)) = f(0)$ is constant

Requirement: f is a positive-definite function (or a kernel) on $\mathbb{R}^d$ if

- $\forall n \in \mathbb{N}^\star$
- $\forall x_1, \ldots, x_n \in \mathbb{R}^d$
- $\forall \lambda_1, \ldots, \lambda_n \in \mathbb{R}$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j f(x_i - x_j) \geq 0$$

How to choose C?

Second-order stationarity

A random field Z over $\mathcal{D} \subset \mathbb{R}^d$ is second-order stationary if for all x, y in $\mathcal{D}$, the following properties hold:

- $\mathbb{E}[Z(x)] = m < \infty$
- $\mathbb{V}(Z(x)) < \infty$
- $C(x, y) = f(x - y)$ (only depends on $x - y$)

Consequence $\mathbb{V}(Z(x)) = f(0)$ is constant

Requirement: f is a positive-definite function (or a kernel) on $\mathbb{R}^d$ if

- $\forall n \in \mathbb{N}^\star$
- $\forall x_1, \dots, x_n \in \mathbb{R}^d$
- $\forall \lambda_1, \dots, \lambda_n \in \mathbb{R}$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j f(x_i - x_j) \geq 0$$

In other words, any covariance matrix built by applying f to a "distance" matrix of $\mathbb{R}^d$ has to be (semi) positive-definite.

How to find a positive-definite function?

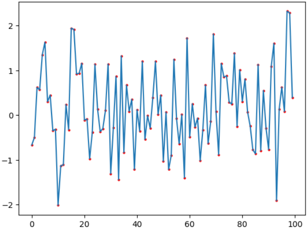
- Constructive approach: build a stationary random field and compute its covariance function
- Use a characterisation theorem (e.g. Bochner's theorem)

Example on $\mathbb{N}$

Consider a sequence of independant random variables $\{X(i), i \in \mathbb{N}\}$ with $\mathbb{E}[X(i)] = 0$ and $\mathbb{V}(X(i)) = 1$.

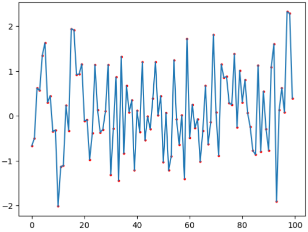
Example on $\mathbb{N}$

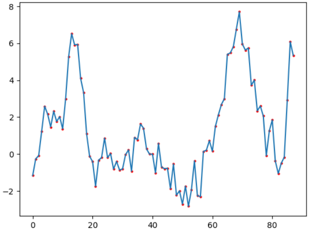
Consider a sequence of independant random variables $\{X(i), i \in \mathbb{N}\}$ with $\mathbb{E}[X(i)] = 0$ and $\mathbb{V}(X(i)) = 1$.



Example on $\mathbb{N}$

Consider a sequence of independant random variables $\{X(i), i \in \mathbb{N}\}$ with $\mathbb{E}[X(i)] = 0$ and $\mathbb{V}(X(i)) = 1$.

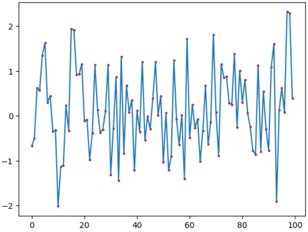


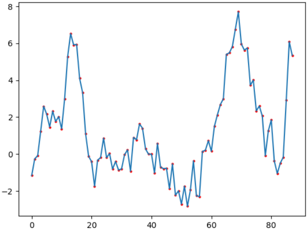


$$\text{Define } Z(i) = \sum_{j=i-p}^{i+p} X(j)$$

Example on $\mathbb{N}$

Consider a sequence of independant random variables $\{X(i), i \in \mathbb{N}\}$ with $\mathbb{E}[X(i)] = 0$ and $\mathbb{V}(X(i)) = 1$.





$$\text{Define } Z(i) = \sum_{j=i-p}^{i+p} X(j)$$

For all $(i, j) \in \mathbb{N}^2$, compute:

- $\mathbb{E}[X(i)]$ and $\mathbb{E}[Z(i)]$ for all $i \in \mathbb{N}$
- $\mathrm{Cov}(X(i), X(j))$ and $\mathrm{Cov}(Z(i), Z(j))$

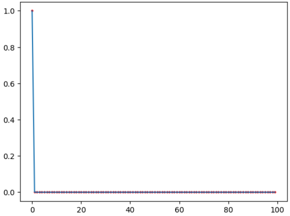
Solution

$\mathbb{E}[X(i)] = 0$ and $\mathbb{E}[Z(i)] = 0$ for all $i \in \mathbb{N}$

Solution

$\mathbb{E}[X(i)] = 0$ and $\mathbb{E}[Z(i)] = 0$ for all $i \in \mathbb{N}$

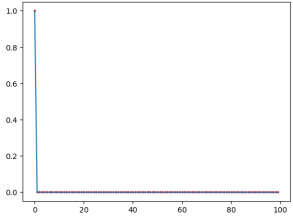
$$C_X(i,j) = \mathrm{Cov}(X(i), X(j)) = \delta_{i,j}$$
$$= \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$



Solution

$\mathbb{E}[X(i)] = 0$ and $\mathbb{E}[Z(i)] = 0$ for all $i \in \mathbb{N}$

$$C_X(i,j) = \mathrm{Cov}(X(i), X(j)) = \delta_{i,j}$$
$$= \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

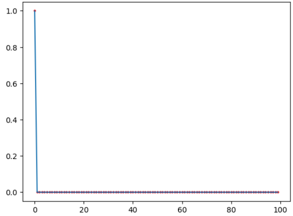


$$C_Z(i,j) = \mathrm{Cov}(Z(i), Z(j)) = \sum_{k=i-p}^{i+p} \sum_{l=j-p}^{j+p} \mathrm{Cov}(X(k), X(l))$$

Solution

$\mathbb{E}[X(i)] = 0$ and $\mathbb{E}[Z(i)] = 0$ for all $i \in \mathbb{N}$

$$C_X(i,j) = \mathrm{Cov}(X(i), X(j)) = \delta_{i,j}$$
$$= \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

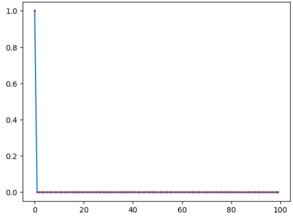


$$C_Z(i,j) = \mathrm{Cov}(Z(i), Z(j)) = \sum_{k=i-p}^{i+p} \sum_{l=j-p}^{j+p} \mathrm{Cov}(X(k), X(l))$$
$$= \sum_{k=i-p}^{i+p} \sum_{l=j-p}^{j+p} \delta_{k,l}$$

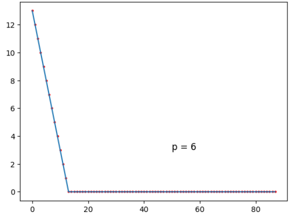
Solution

$\mathbb{E}[X(i)] = 0$ and $\mathbb{E}[Z(i)] = 0$ for all $i \in \mathbb{N}$

$$C_X(i,j) = \mathrm{Cov}(X(i), X(j)) = \delta_{i,j}$$
$$= \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$



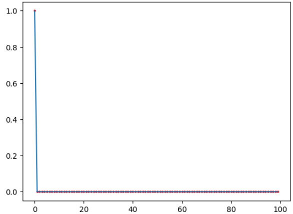
$$C_Z(i,j) = \mathrm{Cov}(Z(i), Z(j)) = \sum_{k=i-p}^{i+p} \sum_{l=j-p}^{j+p} \mathrm{Cov}(X(k), X(l))$$
$$= \sum_{k=i-p}^{i+p} \sum_{l=j-p}^{j+p} \delta_{k,l}$$
$$= \begin{cases} 2p+1-|i-j| & \text{if } |i-j| \leq 2p+1 \\ 0 & \text{otherwise} \end{cases}$$



Solution

$\mathbb{E}[X(i)] = 0$ and $\mathbb{E}[Z(i)] = 0$ for all $i \in \mathbb{N}$

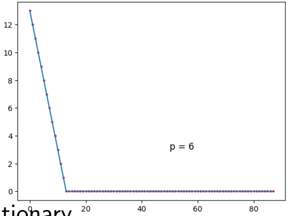
$$C_X(i,j) = \mathrm{Cov}(X(i), X(j)) = \delta_{i,j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$



$$C_Z(i,j) = \mathrm{Cov}(Z(i), Z(j)) = \sum_{k=i-p}^{i+p} \sum_{l=j-p}^{j+p} \mathrm{Cov}(X(k), X(l))$$

$$= \sum_{k=i-p}^{i+p} \sum_{l=j-p}^{j+p} \delta_{k,l}$$

$$= \begin{cases} 2p+1-|i-j| & \text{if } |i-j| \leq 2p+1 \\ 0 & \text{otherwise} \end{cases}$$



The random fields X and Z are second-order stationary

Generalization on $\mathbb{R}$

- Let $W(.)$ a standardized white noise on $\mathbb{R}$, i.e an orthogonal random measure: for all $(u,v) \in \mathbb{R}^2$
    - $\mathbb{E}[W(du)] = 0$
    - $\mathbb{E}[W(du)W(dv)] = \delta_u(v)du$ with $\delta_u(.)$ is a Dirac measure at u.

Generalization on $\mathbb{R}$

- Let $W(.)$ a standardized white noise on $\mathbb{R}$, i.e an orthogonal random measure: for all $(u, v) \in \mathbb{R}^2$
  - $\mathbb{E}[W(du)] = 0$
  - $\mathbb{E}[W(du)W(dv)] = \delta_u(v)du$ with $\delta_u(.)$ is a Dirac measure at u.
- For a function $g \in L^2(\mathbb{R})$, let consider for

$$Z(x) = \int g(x-u)W(du)$$

Generalization on $\mathbb{R}$

- Let $W(.)$ a standardized white noise on $\mathbb{R}$, i.e an orthogonal random measure: for all $(u, v) \in \mathbb{R}^2$
  - $\mathbb{E}[W(du)] = 0$
  - $\mathbb{E}[W(du)W(dv)] = \delta_u(v)du$ with $\delta_u(.)$ is a Dirac measure at u.
- For a function $g \in L^2(\mathbb{R})$, let consider for

$$Z(x) = \int g(x-u)W(du)$$

- Z is a second-order stationary random field with

$$\mathbb{E}[Z(x)] = 0$$

Generalization on $\mathbb{R}$

- Let $W(.)$ a standardized white noise on $\mathbb{R}$, i.e an orthogonal random measure: for all $(u, v) \in \mathbb{R}^2$
  - $\mathbb{E}[W(du)] = 0$
  - $\mathbb{E}[W(du)W(dv)] = \delta_u(v)du$ with $\delta_u(.)$ is a Dirac measure at u.
- For a function $g \in L^2(\mathbb{R})$, let consider for

$$Z(x) = \int g(x-u)W(du)$$

- Z is a second-order stationary random field with

$$\mathbb{E}[Z(x)] = 0$$

$$\mathrm{Cov}(Z(x), Z(y)) = \int g(x-u)g(y-u)du$$

Generalization on $\mathbb{R}$

- Let $W(.)$ a standardized white noise on $\mathbb{R}$, i.e an orthogonal random measure: for all $(u, v) \in \mathbb{R}^2$
  - $\mathbb{E}[W(du)] = 0$
  - $\mathbb{E}[W(du)W(dv)] = \delta_u(v)du$ with $\delta_u(.)$ is a Dirac measure at u.
- For a function $g \in L^2(\mathbb{R})$, let consider for

$$Z(x) = \int g(x-u)W(du)$$

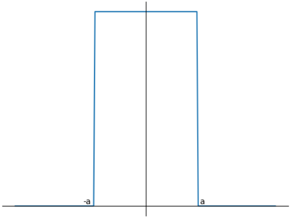
- Z is a second-order stationary random field with

$$\mathbb{E}[Z(x)] = 0$$

$$\begin{aligned}\mathrm{Cov}(Z(x), Z(y)) &= \int g(x-u)g(y-u)du \\ &= \int g(h+v)g(v)dv \text{ with } h = x - y\end{aligned}$$

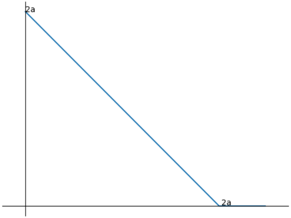
Example

- $g(x) = \mathbf{1}_{[-a,a]}(x)$



- Z has covariance function

$$f(h) = \int g(h+v)g(v)dv = (2a - |h|)\mathbf{1}_{[0,2a]}(h)$$



How to find a positive-definite function?

Bochner's theorem

A function f is positive-definite if and only if it is the Fourier transform of a non-negative measure μ on $\mathbb{R}^d$.

$$\forall x \in \mathbb{R}^d, f(x) = \int_{\mathbb{R}^d} e^{i\langle x,u\rangle} d\mu(u)$$

Proof of Bochner's Theorem (Sufficient Condition)

Let μ be a finite positive measure on $\mathbb{R}^d$. We suppose that f is the Fourier transform of μ:

$$f(x) = \int_{\mathbb{R}^d} e^{-i\langle x,u\rangle}\, d\mu(u)$$

For a finite set of points $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ and a set of real coefficients $\lambda_1, \lambda_2, \ldots, \lambda_n$, we consider the quadratic form:

$$\begin{aligned}
Q &= \sum_{j=1}^{n}\sum_{k=1}^{n} \lambda_j \lambda_k f(x_j - x_k) \\
&= \sum_{j=1}^{n}\sum_{k=1}^{n} \lambda_j \lambda_k \int_{\mathbb{R}^d} e^{-i\langle x_j - x_k, u\rangle}\, d\mu(u) \\
&= \int_{\mathbb{R}^d} \sum_{j=1}^{n}\sum_{k=1}^{n} \lambda_j \lambda_k e^{-i\langle x_j - x_k, u\rangle}\, d\mu(u)
\end{aligned}$$

Proof of Bochner's Theorem (Sufficient Condition)

$$\sum_{j=1}^{n}\sum_{k=1}^{n}\lambda_j\lambda_k e^{-i\langle x_j - x_k, u\rangle} = \left(\sum_{j=1}^{n}\lambda_j e^{-i\langle x_j, u\rangle}\right)\left(\sum_{j=1}^{n}\lambda_j e^{i\langle x_j, u\rangle}\right)$$

Therefore:

$$Q = \int_{\mathbb{R}^d}\left|\sum_{j=1}^{n}\lambda_j e^{-i\langle x_j, u\rangle}\right|^2 d\mu(u) \geq 0$$

We have shown that f is positive definite.

Properties

- If C is a positive-definite function, then λC is positive-definite for all $\lambda \geq 0$
- If C_1 and C_2 are positive-definite functions, then $C_1 + C_2$ is positive-definite
- If C_1 and C_2 are positive-definite functions, then $C_1 \times C_2$ is positive-definite
- If C is a positive-definite function, then $h \to C(Th)$ is positive-definite for all matrix $T \in \mathbb{R}^{d \times d}$

Properties

- If C is a positive-definite function, then λC is positive-definite for all $\lambda \geq 0$
- If C_1 and C_2 are positive-definite functions, then $C_1 + C_2$ is positive-definite
- If C_1 and C_2 are positive-definite functions, then $C_1 \times C_2$ is positive-definite
- If C is a positive-definite function, then $h \to C(Th)$ is positive-definite for all matrix $T \in \mathbb{R}^{d \times d}$

Consequence: we can work with a catalog of valid covariance functions and build new ones by combining them.

Properties

- If C is a positive-definite function, then λC is positive-definite for all $\lambda \geq 0$
- If C_1 and C_2 are positive-definite functions, then $C_1 + C_2$ is positive-definite
- If C_1 and C_2 are positive-definite functions, then $C_1 \times C_2$ is positive-definite
- If C is a positive-definite function, then $h \to C(Th)$ is positive-definite for all matrix $T \in \mathbb{R}^{d \times d}$

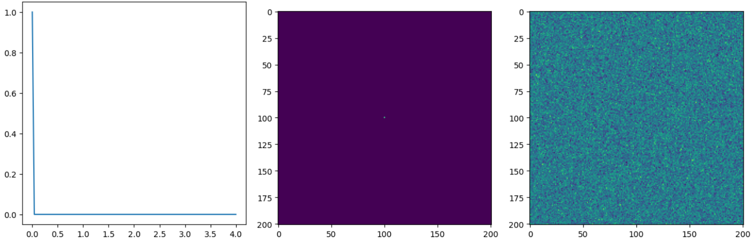
Consequence: we can work with a catalog of valid covariance functions and build new ones by combining them.

Other properties:

- For all $x \in \mathbb{R}^d$, we have $\mathbb{V}(Z(x)) = C(0)$
- $C(h) = C(-h)$ for all $h \in \mathbb{R}^d$
- $C(h) \leq C(0)$ for all $h \in \mathbb{R}^d$

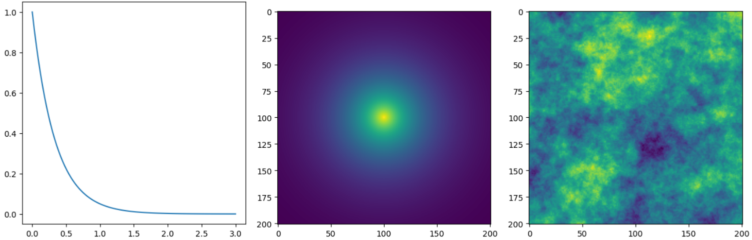
Nugget

$$C(h) = \delta_0(h)$$



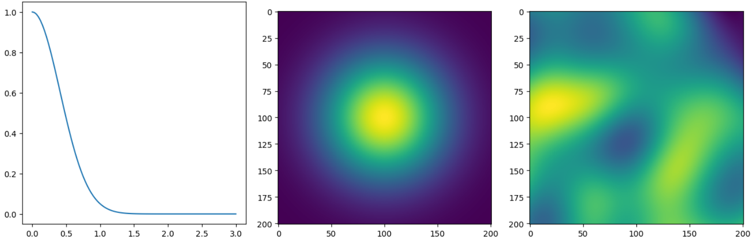
Exponential covariance function

$$C(h) = e^{-\frac{|h|}{a}}$$



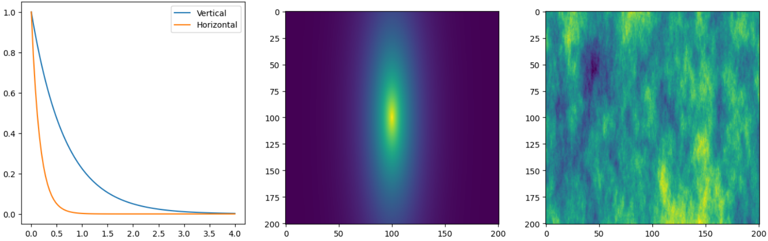
Gaussian covariance function

$$C(h) = e^{-\frac{|h|^2}{a^2}}$$



Anisotropy

$$C(h) = e^{-\sqrt{\frac{h_x^2}{a_x^2} + \frac{h_y^2}{a_y^2}}}$$



Variogram and Covariance

If C is a (stationary) covariance function, then we can define the variogram γ

$$\gamma(h) = \frac{1}{2}\mathbb{V}(Z(x+h) - Z(x))$$

Variogram and Covariance

If C is a (stationary) covariance function, then we can define the variogram γ

$$\gamma(h) = \frac{1}{2}\mathbb{V}(Z(x+h) - Z(x))$$

- $\gamma(h) = C(0) - C(h)$
- Conversely, if the variogram is bounded, we have

$$C(h) = \gamma(\infty) - \gamma(h), \text{ where } \gamma(\infty) = \lim_{|h|\to\infty} \gamma(h)$$

Variogram and Covariance

If C is a (stationary) covariance function, then we can define the variogram γ

$$\gamma(h) = \frac{1}{2}\mathbb{V}(Z(x+h) - Z(x))$$

- $\gamma(h) = C(0) - C(h)$
- Conversely, if the variogram is bounded, we have

$$C(h) = \gamma(\infty) - \gamma(h), \text{ where } \gamma(\infty) = \lim_{|h|\to\infty} \gamma(h)$$

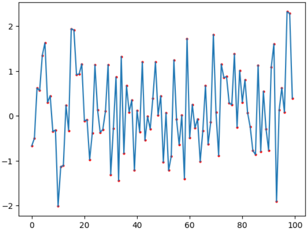
Can we define a weaker stationarity assumption to work with unbounded variograms?

Example on $\mathbb{N}$

Consider a sequence of independent random variables $\{X(i), i \in \mathbb{N}\}$ with $\mathbb{E}[X(i)] = 0$ and $\mathbb{V}(X(i)) = 1$.

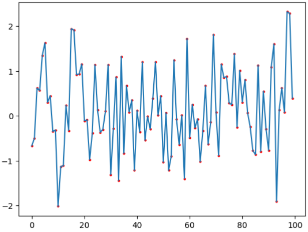
Example on $\mathbb{N}$

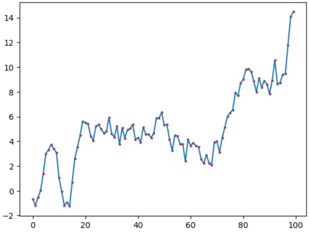
Consider a sequence of independent random variables $\{X(i), i \in \mathbb{N}\}$ with $\mathbb{E}[X(i)] = 0$ and $\mathbb{V}(X(i)) = 1$.



Example on $\mathbb{N}$

Consider a sequence of independent random variables $\{X(i), i \in \mathbb{N}\}$ with $\mathbb{E}[X(i)] = 0$ and $\mathbb{V}(X(i)) = 1$.

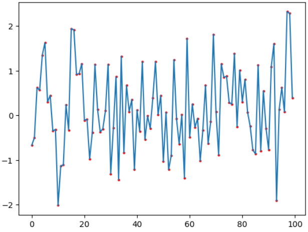


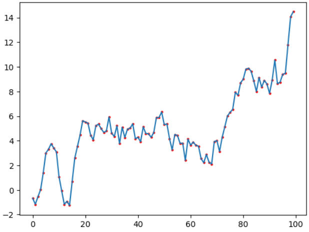


Define $Z(i) = \sum_{j=0}^{i} X(j)$

Example on $\mathbb{N}$

Consider a sequence of independent random variables $\{X(i), i \in \mathbb{N}\}$ with $\mathbb{E}[X(i)] = 0$ and $\mathbb{V}(X(i)) = 1$.





Define $Z(i) = \sum_{j=0}^{i} X(j)$

For all $(i, j) \in \mathbb{N}^2$, compute:

- $\mathbb{E}[Z(i)]$ for all $i \in \mathbb{N}$
- $\mathbb{V}(Z(i))$
- $\gamma(i, j) = \frac{1}{2}\mathbb{V}(Z(i) - Z(j))$

Solution

$\mathbb{E}[Z(i)] = 0$ for all $i \in \mathbb{N}$

Solution

$\mathbb{E}[Z(i)] = 0$ for all $i \in \mathbb{N}$

$$\mathbb{V}(Z(i)) = \mathrm{Cov}(Z(i), Z(i)) = \sum_{k=0}^{i} \sum_{l=0}^{i} \mathrm{Cov}(X(k), X(l))$$

Solution

$\mathbb{E}[Z(i)] = 0$ for all $i \in \mathbb{N}$

$$\begin{aligned}
\mathbb{V}(Z(i)) = \mathrm{Cov}(Z(i), Z(i)) &= \sum_{k=0}^{i} \sum_{l=0}^{i} \mathrm{Cov}(X(k), X(l)) \\
&= \sum_{k=0}^{i} \sum_{l=0}^{i} \delta_{k,l} = (i+1)
\end{aligned}$$

Solution

$\mathbb{E}[Z(i)] = 0$ for all $i \in \mathbb{N}$

$$\begin{aligned}\mathbb{V}(Z(i)) = \mathrm{Cov}(Z(i), Z(i)) &= \sum_{k=0}^{i}\sum_{l=0}^{i} \mathrm{Cov}(X(k), X(l)) \\ &= \sum_{k=0}^{i}\sum_{l=0}^{i} \delta_{k,l} = (i+1)\end{aligned}$$

$$\gamma(i,j) = \frac{1}{2}\mathbb{V}(Z(i) - Z(j))$$

Solution

$\mathbb{E}[Z(i)] = 0$ for all $i \in \mathbb{N}$

$$\begin{aligned}\mathbb{V}(Z(i)) = \mathrm{Cov}(Z(i), Z(i)) &= \sum_{k=0}^{i} \sum_{l=0}^{i} \mathrm{Cov}(X(k), X(l)) \\ &= \sum_{k=0}^{i} \sum_{l=0}^{i} \delta_{k,l} = (i+1)\end{aligned}$$

$$\gamma(i,j) = \frac{1}{2}\mathbb{V}(Z(i) - Z(j)) = \frac{1}{2}\mathbb{V}\left(\sum_{k=\min(i,j)}^{\max(i,j)} X(k)\right)$$

Solution

$\mathbb{E}[Z(i)] = 0$ for all $i \in \mathbb{N}$

$$\mathbb{V}(Z(i)) = \mathrm{Cov}(Z(i), Z(i)) = \sum_{k=0}^{i} \sum_{l=0}^{i} \mathrm{Cov}(X(k), X(l))$$

$$= \sum_{k=0}^{i} \sum_{l=0}^{i} \delta_{k,l} = (i+1)$$

$$\gamma(i,j) = \frac{1}{2} \mathbb{V}(Z(i) - Z(j)) = \frac{1}{2} \mathbb{V} \left(\sum_{k=\min(i,j)}^{\max(i,j)} X(k) \right)$$

$$= \frac{1}{2} \sum_{k=\min(i,j)}^{\max(i,j)} \mathbb{V}(X(k))$$

Solution

$\mathbb{E}[Z(i)] = 0$ for all $i \in \mathbb{N}$

$$\mathbb{V}(Z(i)) = \mathrm{Cov}(Z(i), Z(i)) = \sum_{k=0}^{i} \sum_{l=0}^{i} \mathrm{Cov}(X(k), X(l))$$

$$= \sum_{k=0}^{i} \sum_{l=0}^{i} \delta_{k,l} = (i+1)$$

$$\gamma(i,j) = \frac{1}{2}\mathbb{V}(Z(i) - Z(j)) = \frac{1}{2}\mathbb{V}\left(\sum_{k=\min(i,j)}^{\max(i,j)} X(k)\right)$$

$$= \frac{1}{2} \sum_{k=\min(i,j)}^{\max(i,j)} \mathbb{V}(X(k))$$

$$= \frac{|i-j|}{2}$$

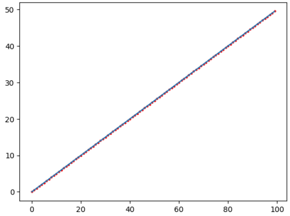
Solution

$\mathbb{E}[Z(i)] = 0$ for all $i \in \mathbb{N}$

$$\mathbb{V}(Z(i)) = \mathrm{Cov}(Z(i), Z(i)) = \sum_{k=0}^{i} \sum_{l=0}^{i} \mathrm{Cov}(X(k), X(l))$$
$$= \sum_{k=0}^{i} \sum_{l=0}^{i} \delta_{k,l} = (i+1)$$

$$\gamma(i,j) = \frac{1}{2}\mathbb{V}(Z(i) - Z(j)) = \frac{1}{2}\mathbb{V}\left(\sum_{k=\min(i,j)}^{\max(i,j)} X(k)\right)$$
$$= \frac{1}{2} \sum_{k=\min(i,j)}^{\max(i,j)} \mathbb{V}(X(k))$$
$$= \frac{|i-j|}{2}$$

only depends on the distance



Relaxing the second-order stationarity assumption

- The second-order stationarity assumption is somtimes too strong (e.g. unbounded variograms)
- It is often relaxed to a weaker assumption: intrinsic stationarity

Relaxing the second-order stationarity assumption

- The second-order stationarity assumption is somtimes too strong (e.g. unbounded variograms)
- It is often relaxed to a weaker assumption: intrinsic stationarity

Intrinsic stationarity

A random field Z over $\mathcal{D} \subset \mathbb{R}^d$ is intrinsically stationary if the increments are stationary: $\forall x, y \in \mathcal{D}$

- $\mathbb{E}[Z(y) - Z(x)] = 0$
- $\gamma(x, y) = \frac{1}{2}\mathbb{V}(Z(y) - Z(x)) = \gamma(x - y)$ is a function of $x - y$

Working with intrinsic stationarity

- An authorized linear combination (ALC) is a linear combination

$$\sum_{i=1}^{n} \lambda_i Z(x_i)$$

  for which we are allowed to compute the expectations and the variances
- We can show that the ALC of an intrinsic model are the combinations for which

$$\sum_{i=1}^{n} \lambda_i = 0$$

Variance of an authorized linear combination

$$\mathbb{V}\left(\sum_{i=1}^{n}\lambda_i Z(x_i)\right) = -\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(x_i - x_j)$$

Variance of an authorized linear combination

$$\mathbb{V}\left(\sum_{i=1}^{n}\lambda_i Z(x_i)\right) = -\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(x_i - x_j)$$

Proof:

- Let set $Y(x) = Z(x) - Z(x_0)$ for a point x_0.

Variance of an authorized linear combination

$$\mathbb{V}\left(\sum_{i=1}^{n}\lambda_i Z(x_i)\right) = -\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(x_i - x_j)$$

Proof:

- Let set $Y(x) = Z(x) - Z(x_0)$ for a point x_0.
- The variogram of Y is γ

Variance of an authorized linear combination

$$\mathbb{V}\left(\sum_{i=1}^{n}\lambda_i Z(x_i)\right) = -\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(x_i - x_j)$$

Proof:

- Let set $Y(x) = Z(x) - Z(x_0)$ for a point x_0.
- The variogram of Y is γ
- The covariances $C_{xy} = \mathrm{Cov}(Y(x), Y(y))$ are well defined and we know that

$$C_{xy} = \frac{C_{xx} + C_{yy}}{2} - \gamma(x - y)$$

Variance of an authorized linear combination

$$\mathbb{V}\left(\sum_{i=1}^{n} \lambda_i Z(x_i)\right) = -\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma(x_i - x_j)$$

Proof:

- Let set $Y(x) = Z(x) - Z(x_0)$ for a point x_0.
- The variogram of Y is γ
- The covariances $C_{xy} = \mathrm{Cov}(Y(x), Y(y))$ are well defined and we know that
$$C_{xy} = \frac{C_{xx} + C_{yy}}{2} - \gamma(x - y)$$
- Now we can compute the variance of the ALC

$$\mathbb{V}\left(\sum_{i=1}^{n} \lambda_i Z(x_i)\right) = \mathbb{V}\left(\sum_{i=1}^{n} \lambda_i Y(x_i)\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j C_{x_i,x_j}$$

Variance of an authorized linear combination

$$\mathbb{V}\left(\sum_{i=1}^{n}\lambda_i Z(x_i)\right) = -\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(x_i - x_j)$$

Proof:

- Let set $Y(x) = Z(x) - Z(x_0)$ for a point x_0.
- The variogram of Y is γ
- The covariances $C_{xy} = \mathrm{Cov}(Y(x), Y(y))$ are well defined and we know that
$$C_{xy} = \frac{C_{xx} + C_{yy}}{2} - \gamma(x - y)$$
- Now we can compute the variance of the ALC

$$\mathbb{V}\left(\sum_{i=1}^{n}\lambda_i Z(x_i)\right) = \mathbb{V}\left(\sum_{i=1}^{n}\lambda_i Y(x_i)\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j C_{x_i,x_j}$$
$$= -\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(x_i - x_j)$$

Variogram functions

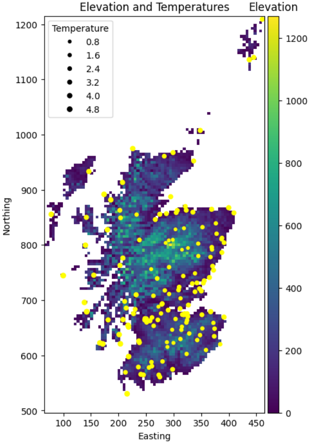
A function γ is a valid variogram function on $\mathbb{R}^d$ if it is conditionally negative definite:

- For any $n \in \mathbb{N}^\star$
- For any $x_1, \ldots, x_n \in \mathbb{R}^d$
- For any linear combination such as $\sum_{i=1}^n \lambda_i = 0$

$$-\sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(x_i - x_j) \geq 0$$

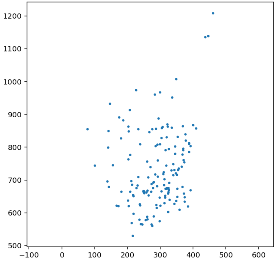
Data-sets

Temperatures in January + DTM (Scotland)



Data-sets

Temperatures in January + DTM (Scotland)



Variogram cloud

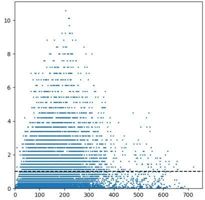
- The data is modeled as *samples of a regionalized variable* z*, i.e. as evaluations at locations* $x_1, .., x_n$ *of a variable* z *defined across a spatial domain:*
$$\{z_i = z(x_i) : i = 1, ..., n\}.$$
- *The variogram cloud is the set of pair of points defined as*

$$\left\{ \left(\|x_i - x_j\|, \frac{\left(z(x_i) - z(x_j)\right)^2}{2} \quad \textit{where} \quad 1 \leq i \leq j \leq n \right\}$$

Variogram cloud

- The data is modeled as *samples of a regionalized variable* z, *i.e. as evaluations at locations* $x_1, .., x_n$ *of a variable* z *defined across a spatial domain:*
$$\{z_i = z(x_i) : i = 1, ..., n\}.$$
- *The variogram cloud is the set of pair of points defined as*

$$\left\{ \left(\|x_i - x_j\|, \frac{(z(x_i) - z(x_j))^2}{2} \quad \textit{where} \quad 1 \leq i \leq j \leq n \right\} \right.$$



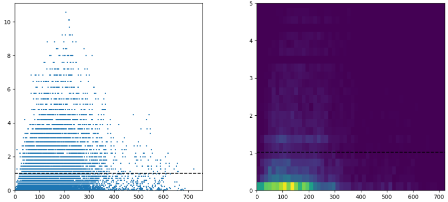
Variogram cloud

- The data is modeled as *samples of a regionalized variable* z, *i.e. as evaluations at locations* $x_1, .., x_n$ *of a variable* z *defined across a spatial domain:*

$$\{z_i = z(x_i) : i = 1, ..., n\}.$$

- *The variogram cloud is the set of pair of points defined as*

$$\left\{ \left(\|x_i - x_j\|, \frac{(z(x_i) - z(x_j))^2}{2} \quad \textit{where} \quad 1 \leq i \leq j \leq n \right\} \right.$$



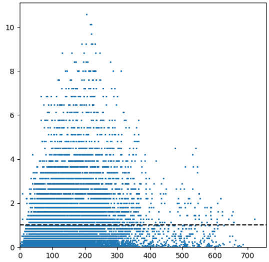
Empirical variogram

- The data is modeled as *samples of a regionalized variable* z*, i.e. as evaluations at locations* $x_1, .., x_n$ *of a variable* z *defined across a spatial domain:*

$$\{z_i = z(x_i) : i = 1, ..., n\}.$$

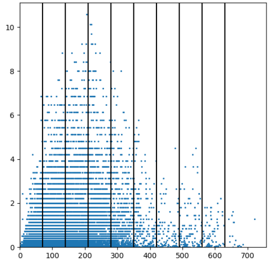
- *The variogram cloud is the set of pair of points defined as*

$$\left\{\left(\|x_i - x_j\|, \frac{(z(x_i) - z(x_j))^2}{2}\right) \quad \textit{where} \quad 1 \leq i \leq j \leq n\right\}$$



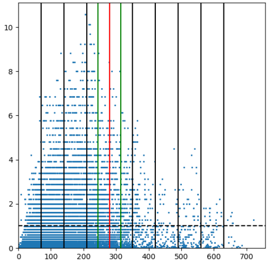
Empirical variogram

- Divide the distance span into classes (lag) I_k



Empirical variogram

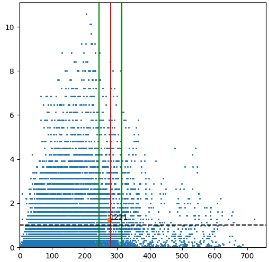
- Divide the distance span into classes (lag) I_k
- Choose a tolerance (% of the class range)



Empirical variogram

- Divide the distance span into classes (lag) I_k
- Choose a tolerance (% of the class range)
- Compute the averages inside the classes

$$N_k(h) = \sharp\{1 \leq i \leq j \leq n; ||x_i - x_j|| \in I_k\}$$

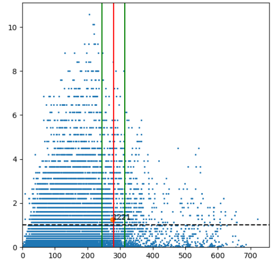


Empirical variogram

- Divide the distance span into classes (lag) I_k
- Choose a tolerance (% of the class range)
- Compute the averages inside the classes

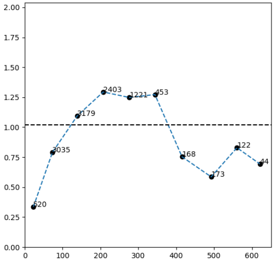
$$N_k(h) = \sharp\{1 \leq i \leq j \leq n; ||x_i - x_j|| \in I_k\}$$

$$h_k = \frac{1}{N_k(h)} \sum_{||x_i - x_j|| \in I_k} ||x_i - x_j|| \qquad \hat{\gamma}_k = \frac{1}{N_k(h)} \sum_{||x_i - x_j|| \in I_k}^{N_k(h)} \frac{(z(x_i) - z(x_j))^2}{2}$$



Empirical variogram

- Divide the distance span into classes (lag) I_k
- Choose a tolerance (% of the class range)
- Compute the averages inside the classes
- Repeat for each interal I_k



Anisotropy and directional variograms

- (Omni-directional) variograms implicitely assume that the squared differences

$$(z(x_i) - z(x_j))^2$$

only depends on the distances

$$||x_i - x_j||$$

Anisotropy and directional variograms

- (Omni-directional) variograms implicitely assume that the squared differences

$$(z(x_i) - z(x_j))^2$$

  only depends on the distances

$$||x_i - x_j||$$

- This is often not realistic since spatial variations can be different for different directions of the space (anisotropy)

Anisotropy and directional variograms

- (Omni-directional) variograms implicitely assume that the squared differences

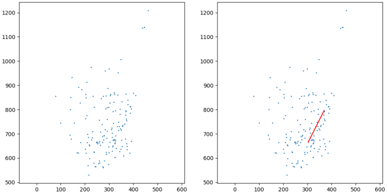
$$(z(x_i) - z(x_j))^2$$

only depends on the distances

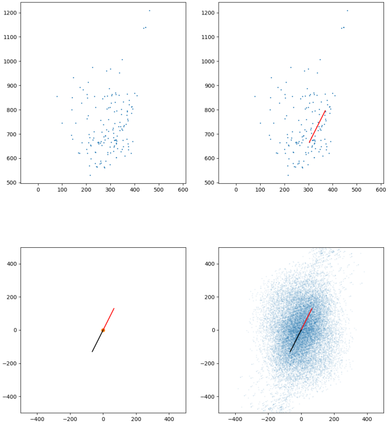
$$||x_i - x_j||$$

- This is often not realistic since spatial variations can be different for different directions of the space (anisotropy)
- Idea: use directional variograms

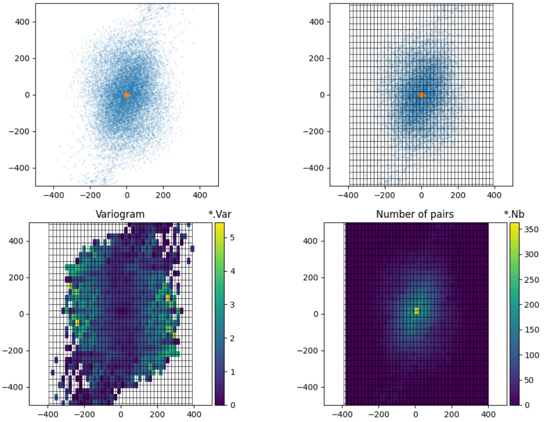
2D variogram cloud



2D variogram cloud

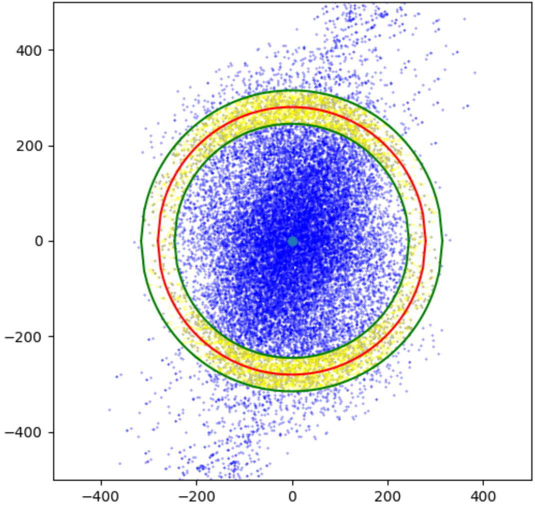


2D variogram cloud



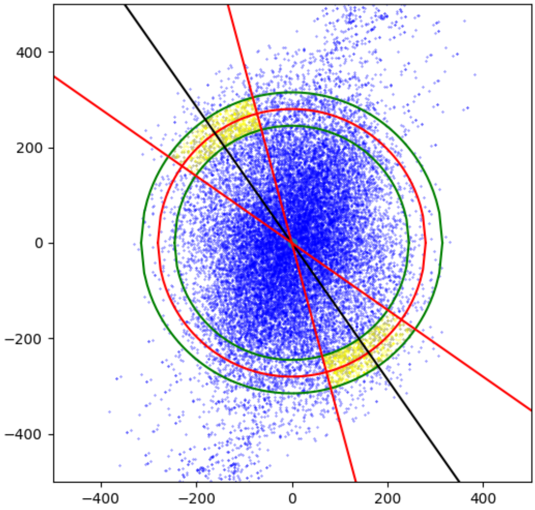
Directional variograms

Omni



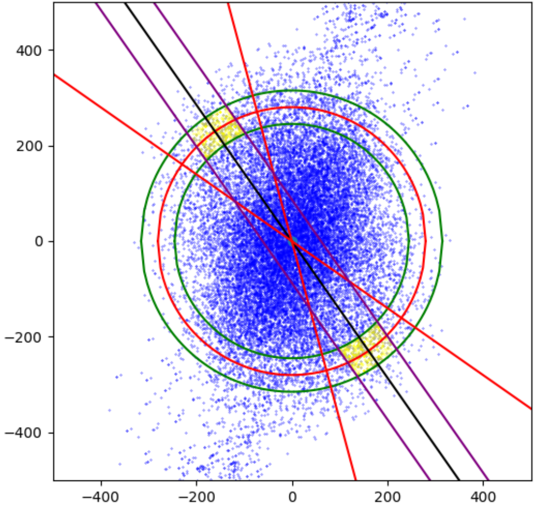
Directional variograms

Particular direction with angular tolerance



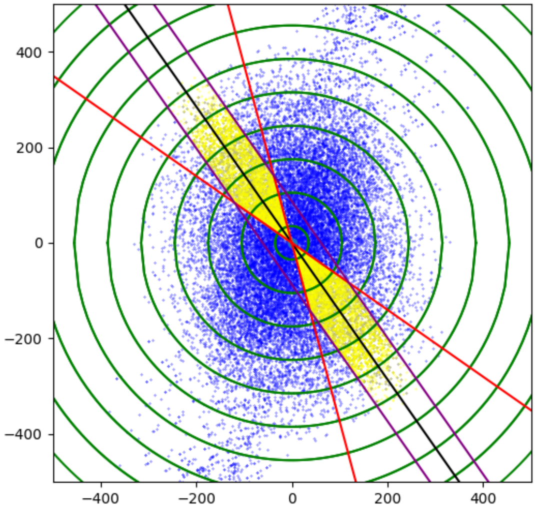
Directional variograms

Cylinder



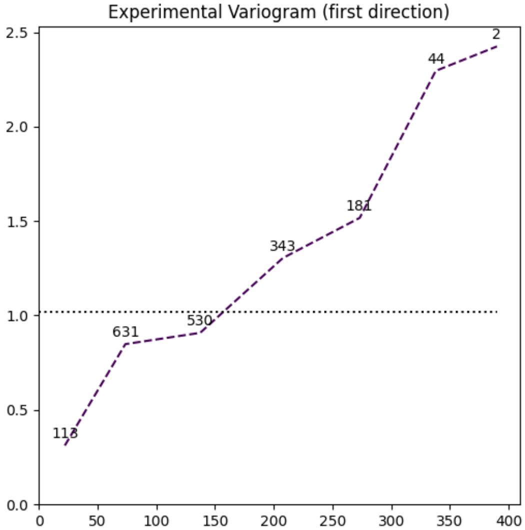
Directional variograms

Compute for all intervals I_k



Directional variograms

Result



Directional variograms

2 directions