

# Spatio-temporal statistics (MATH4341)

## Michaelmas term

‘Spatial statistics’

Georgios P. Karagiannis

georgios.karagiannis@durham.ac.uk

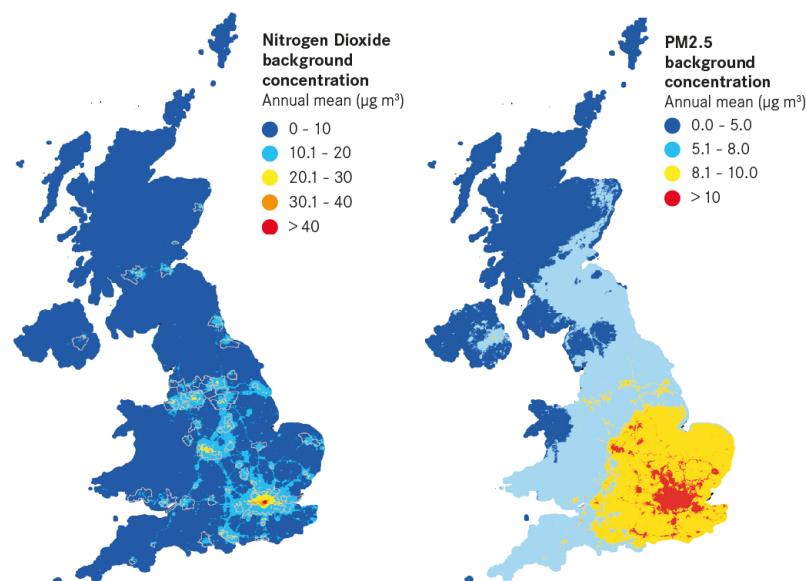
Department of Mathematical Sciences (Office MCS3088)  
Durham University  
Stockton Road Durham DH1 3LE UK

2024/10/17 at 18:33:03

Concepts

An introduction to spatial statistics:

- Regionalised statistical concepts
- Point referenced data analysis
- Aerial unit data analysis
- Point pattern data analysis
- Implementation in R



## Lecture notes

1. Lecture notes part 1: Types of spatial data and modelling
2. Lecture notes part 2: Point referenced data modeling / Geostatistics
3. Lecture notes part 3: Aerial unit data / spatial data on lattices
4. Lecture notes part 4: Spatial point pattern modeling

## Reading list

These lecture Handouts have been derived based on the above reading list.

### Main textbooks (methods, implementation, and theory):

- Cressie, N. (2015). Statistics for spatial data. John Wiley & Sons.
- Kent, J. T., & Mardia, K. V. (2022). Spatial analysis (Vol. 72). John Wiley & Sons.
- Gaetan, C., & Guyon, X. (2010). Spatial statistics and modeling (Vol. 90). New York: Springer.
- Moller, J., & Waagepetersen, R. P. (2003). Statistical inference and simulation for spatial point processes. CRC press.

### Main textbooks (software):

- Bivand, R. S., Pebesma, E. J., Gómez-Rubio, V., & Pebesma, E. J. (2008). Applied spatial data analysis with R (Vol. 747248717, pp. 237-268). New York: Springer.
- Moraga, P. (2023). Spatial statistics for data science: theory and practice with R. CRC Press.

### Supplementary textbooks (methods, implementation, and theory):

- Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). Hierarchical modeling and analysis for spatial data. CRC press.
- Ripley, B. D. (2005). Spatial statistics. John Wiley & Sons.
- Schabenberger, O., & Gotway, C. A. (2005). Statistical methods for spatial data analysis. CRC press
- Wackernagel, H. (2003). Multivariate geostatistics: an introduction with applications. Springer Science & Business Media.
- Diggle, P. J. (2013). Statistical analysis of spatial and spatio-temporal point patterns. CRC press.
- van Lieshout, M. N. M. (2019). Theory of spatial statistics: a concise introduction. CRC Press.

# Lecture notes part 1: Types of spatial data

Lecturer & author: Georgios P. Karagiannis

georgios.karagiannis@durham.ac.uk

**Aim.** To introduce the types of spatial statistical data. To get a general idea about spatial statistics modeling.

## Reading list & references:

- Cressie, N. (2015). Statistics for spatial data. John Wiley & Sons.  
– Chapter 1: pp 1- 28

## 1. MOTIVATIONS

*Note 1.* Researchers in diverse areas such as geology, soil science, image processing, epidemiology, crop science, ecology, forestry, astronomy, atmospheric science, climatology, ecology, environmental health, and real estate marketing are increasingly faced with the task of analyzing data that are geographically referenced, and often presented as maps.

*Note 2.* In several problems, the data have a space (and time) label associated with them; this gives the motivation for the development and analysis of (not necessarily statistical) models that indicate whether there is dependence between measurements at different locations.

*Note 3.* In an epidemiological investigation, for instance, one might wish to analyze lung, breast, colorectal, and cervical cancer rates by county and year in a particular state, with smoking, mammography, and other important screening and staging information also available at some level. Public health professionals who collect such data are charged not only with surveillance, but also statistical inference tasks, such as modeling of trends and correlation structures, estimation of underlying model parameters, hypothesis testing (or comparison of competing models), and prediction of observations at unobserved locations (and times).

*Note 4.* Spatial statistics is the branch of statistics that focuses on the analysis and modeling of data with inherent spatial relationships, by accounting for spatial dependencies and patterns to derive meaningful insights and make informed decisions.

## Shall I ignore spatial dependence? –No!

*Note 5.* Galton's problem (1888) arises in spatial statistics and cross-cultural research when observations are not statistically independent due to external dependencies like borrowing or common descent. For example, if two neighboring cultures share similar traits, it might

be due to cultural borrowing rather than independent development. This autocorrelation can lead to misleading conclusions if not properly accounted.

*Note 6.* From your experimental design lectures, recall R. A. Fisher's principles of randomization, blocking and replication to neutralize (not remove) spatial dependence. In his agricultural studies, he noticed that "After choosing the area we usually have no guidance beyond the widely verified fact that patches in close proximity are commonly more alike, as judged by the yield of crops, than those which are further apart." To avoid the "confounding" of treatment effect Fisher properly introduced randomization, namely the controlled introduction of uncertainty.

*Note 7.* The First Law of Geography, according to Waldo Tobler, is "*everything is related to everything else, but near things are more related than distant things.*" Perhaps, we can paraphrase it by using stats terms to "nearby attribute values are more statistically dependent than distant attribute values".

**Example 8.** <sup>1</sup>Consider a random sample  $\{Z_i \in \mathbb{R}; i = 1, \dots, n\}$  jointly following a Normal distribution with unknown  $E(Z_i) = \mu$  and known  $\text{Var}(Z_i) = \sigma^2$ .

**Iid assumption:** Assumption that the observables  $\{Z_i\}$  are independent. This is equivalent to iid model  $Z_i = \mu + \epsilon_i$  with  $\epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$ .

**AR assumption:** Assumption that the observables  $\{Z_i\}$  are positively autocorrelated as  $\text{Cov}(Z_i, Z_j) = \sigma^2 \rho^{|i-j|}$  with known  $\rho \in (0, 1)$ . This is equivalent to AR model  $Z_i = \rho Z_{i-1} + \epsilon_i$  with  $\epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2(1 - \rho^2))$ .

Consider the case that the observables  $\{Z_i\}$  are positively autocorrelated. Assume You fail to realize the presence of positive correlation in the data and You use the i.i.d model instead.

(1) Effect on parametric estimation of  $\mu$

The BLUE of  $\mu$  is

$$(1.1) \quad \hat{Z} = \bar{Z} = \frac{1}{n} \sum_{i=1}^n Z_i$$

---

<sup>1</sup>No need to worry how the equations are produced.

If the observables were iid then  $\text{Var}(\hat{Z}|\text{iid obs}) = \frac{\sigma^2}{n}$ . Since the observables are positively autocorrelated

$$(1.2) \quad \begin{aligned} \text{Var}(\hat{Z}|\text{AR obs}) &= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \text{Cov}(Z_i, Z_j) \\ &= \frac{\sigma^2}{n} \overbrace{\left( 1 + 2 \frac{\rho}{1-\rho} \left( 1 - \frac{1}{n} \right) - \frac{2}{n} \left( \frac{\rho}{1-\rho} \right)^2 \frac{1-\rho^{n-1}}{2} \right)}^{\tau_{n,\rho}=} \end{aligned}$$

If  $n = 10$ , and  $\rho = 0.26$ , the  $1 - \alpha = 95\%$  confidence interval for  $\mu$  is

$$\left\{ \bar{Z} \pm q_{\frac{\alpha}{2}} \sqrt{\text{Var}(\hat{Z}|\text{assump.})} \right\} \Rightarrow \begin{cases} \bar{Z} \pm \underbrace{1.96}_{q_{0.95}/2} \frac{\sigma}{\sqrt{10}} & \text{in i.i.d assum.} \\ \bar{Z} \pm \underbrace{2.485}_{=q_{0.95}/\sqrt{10,0.26}} \frac{\sigma}{\sqrt{10}} & \text{in AR assum.} \end{cases}$$

as  $\sqrt{\tau_{10,0.26}} = 1.608$ . Hence, if we fail to realize the presence of positive autocorrelation in the observables and wrongly the i.i.d model, the resulted confidence interval would be too narrow since the actual coverage probability is  $\Phi(2.485) - \Phi(-2.485) = 87.5\%$  instead of 95%.

By re-writing the variance (1.2) as  $\text{Var}(\hat{Z}) = \frac{\sigma^2}{n'}$  with

$$n' = n \left/ \left( 1 + 2 \frac{\rho}{1-\rho} \left( 1 - \frac{1}{n} \right) - \frac{2}{n} \left( \frac{\rho}{1-\rho} \right)^2 \frac{1-\rho^{n-1}}{2} \right) \right.,$$

we intuitively understand that the effect of the positive spatial correlation is that the equivalent independent observations in the above dataset with size  $n$  is  $n'$  (i.e.  $n' \approx n \frac{1-\rho}{1+\rho}$ , for large  $n$ ). If  $n = 10$ , and  $\rho = 0.26$ , then  $n' = 6.2$ .

## (2) Effect on predictive estimation of $Z_{n+1}$

Under the i.i.d. model, the predictor minimizing the MSPE for the next outcome  $Z_{n+1}$  is  $\hat{Z}_{n+1}^{\text{iid}} = \bar{Z}$  and has

$$\text{MSPE}(\hat{Z}_{n+1}^{\text{iid}}|\text{dep obs}) = \sigma^2 \left( 1 + 2 \frac{\rho}{1-\rho} \left( \rho^n - \frac{1}{n} \right) - 2 \left( \frac{\rho}{1-\rho} \right)^2 \frac{1-\rho^{n-1}}{2} \right)$$

Under the AR model the predictor minimizing the mse of the next outcome  $Z_{n+1}$  is

$$\hat{Z}_{n+1}^{\text{AR}} = \rho Z_n + (1-\rho) \frac{Z_1 + (1-\rho) \sum_{i=2}^{n-1} Z_i + Z_n}{n - (n-2)\rho}$$

with MSPE

$$\text{MSPE} \left( \hat{Z}_{n+1}^{\text{AR}} | \text{dep obs} \right) = \sigma^2 \left( 1 - \rho^2 \frac{(1 + \rho)(1 - \rho)^2}{n - (n - 2)\rho} \right)$$

Note that the relative efficiency of this mistake is not too bad, e.g. for  $n = 10$  and  $\rho = 0.26$

$$\text{RE} = \frac{\text{MSPE} \left( \hat{Z}_{10+1}^{\text{AR}} | \text{dep obs} \right)}{\text{MSPE} \left( \hat{Z}_{10+1}^{\text{iid}} | \text{ind obs} \right)} = \frac{1.01952}{1.1} \approx 1.$$

However, if I consider large datasize  $n \rightarrow \infty$  the asymptotic relative efficiency is

$$\text{ARE} = \lim_{n \rightarrow \infty} \frac{\text{MSPE} \left( \hat{Z}_{n+1}^{\text{AR}} | \text{dep obs} \right)}{\text{MSPE} \left( \hat{Z}_{n+1}^{\text{iid}} | \text{ind obs} \right)} = \dots = 1 - \rho^2 \approx \begin{cases} 93\% & , \rho = 0.26 \\ 75\% & , \rho = 0.5 \end{cases}$$

### Spatial data and spatial process.

*Note 9.* In spatial statistics, the basic components are data  $\{Z_1, \dots, Z_n\}$  observed at spatial locations  $\{s_1, \dots, s_n\}$  correspondingly. Classically, the locations are 2D,  $s \in S \subset \mathbb{R}^2$ , however it can be  $S \subset \mathbb{R}^1$  (such as in chromatography applications), or  $S \subset \mathbb{R}^3$  (such as in earth science, 3D imaging, etc) depending on the application.

*Note 10.* Even more exotically, the spatial domain  $S$  does not necessarily need to be a Euclidean space (which our course will focus) such as  $S \subset \mathbb{R}^d$ ,  $d = 1, 2, 3, \dots$  but any topological space, eg sphere (recall Earth is round...).

*Note 11.* The locations  $s_i \in S$  can be considered either (i.) fixed and hence used for training or (ii.) uncertain/random and hence a quantity for inference. Yet,  $\{s_i\}$  can be arranged irregularly in the space or regularly in a grid. Data  $Z_i = Z(s_i)$  are random vectors.

*Note 12.* Let  $s \in \mathbb{R}^d$  be a generic data location, and suppose the datum  $Z(s)$  at spatial location  $s$  is an uncertain and hence random vector. Considering  $s$  to vary over index set  $S \subset \mathbb{R}^d$  imposes a spatial random field (or multivariate random process)

$$\{Z(s); s \in S\}$$

which can be modeled as a random field / random function / stochastic process (to be defined later.).

*Note 13.* In spatial problems, spatial data  $\{Z_{s_i}\}_{i=1}^n$  at locations  $\{s_i\}_{i=1}^n$  are assumed to be realizations of a random field (or a stochastic processes)

$$(1.3) \quad \{Z(s); s \in S\},$$

indexed by a spatial set  $S \subset \mathbb{R}^d$ .

## 2. PRINCIPAL SPATIAL STATISTICS AREAS

*Note 14.* We can characterize the spatial statistical problems according to the type of measurement, their specified (assumed) stochastic generating mechanism, and the choice of the spatial locations. In principle, each of them is associated to different motivations, statistical/scientific problems, statistical tools, however, modern applications/problems may involve characteristics from any combination of them.

*Note 15.* Here, we will study three of spatial statistical areas corresponding to point referenced data, aerial unit data, and point patterns.

### 2.1. Point referenced data / Geostatistics.

*Note 16.* Climate or environmental data are often presented in the form of a map, for example the maximum temperatures on a given day in a country, the concentrations of some pollutant in a city or the mineral content in soil. In mathematical terms, such maps can be described as realizations of a random function (random field/stochastic process); that is, an ensemble of random quantities indexed by points in a region of interest. The aim is usually interpolation, and the associated statistical inference.

*Note 17.* Such data were first analyzed in geological sciences. Hence, for historical reasons, this area of spatial statistics is often called Geostatistics and the point referenced data are also called geocoded or geostatistical data.

*Note 18.* Mathematically speaking, the spatial domain  $S$  is a continuous fixed subset of  $\mathbb{R}^d$  that contains a  $d$ -dimensional rectangle of positive volume. The datum  $Z(s)$  is a random vector (outcome) at specific location  $s \in S$  which can vary continuously over domain  $S$ . In practice, the actual data are observations  $\{Z_i\}_{i=1}^n$  at  $n$  (finite number) fixed locations  $\{s_i\}_{i=1}^n \subset S$  such as  $Z_i = Z(s_i)$ . The locations  $\{s_i\}$  are fixed and can be arranged irregularly in the space or regularly as a grid.

*Note 19.* Geostatistics aims to answer questions about modeling, identification and separation of small and large scale variations, prediction at unobserved locations and reconstruction of the spatial process  $Z(\cdot)$  across the whole space  $S$ .

**Example 20.** (Meuse river data set) Due to pollution of the Meuse river over many years, considerable amounts of heavy metals have accumulated in the overbank sediments of the embanked floodplains of their lower river reaches. The spatial variability of metal pollution of floodplain soils, which is controlled primarily by deposition of contaminated overbank sediments during flood events is under consideration. The process governing heavy metal distribution seems that polluted sediment is carried by the river, and mostly deposited close to the river bank, and areas with low elevation.

The Meuse river R dataset `meuse{sp}` contains locations and topsoil heavy metal concentrations (such as zinc, lead, copper, and cadmium), 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). Figure 2.1 shows 155 measurements of topsoil concentrations of the above heavy metals cadmium, copper, lead, and zinc collected in a flood plain of the river Meuse. Here, the locations  $\{s_i\}$  are fixed and arranged irregularly as chosen by experimental design. The quantities of interest (QoI)  $\{(Z_i^{\text{zinc}}, Z_i^{\text{lead}}, Z_i^{\text{cooper}}, Z_i^{\text{carmium}})\}$  are the concentrations of zinc, lead, copper, and cadmium, at these locations. Interest lie in prediction of QoI at unobserved locations (i.e. interpolation), quantification of the joint distribution of the QoI (evaluate the distribution of a random function  $Z(s) = (Z^{\text{zinc}}(s), Z^{\text{lead}}(s), Z^{\text{cooper}}(s), Z^{\text{carmium}}(s))$ , for all  $s \in S$ ), and how each of QoI depends each other along the flood plain of the river Meuse is of interest (Figure 2.1f). If You ignore spatial dependency and implement standard multivariate statistical techniques (e.g. Fig 2.1e) to analyze the dependency of the QoI's you may obtain misleading results due to confounding space.

**Example 21.** (Coal ash dataset in Pennsylvania) Figure 2.2 shows 208 coal ash core measurements/samples collected on a regular grid of points in the Robena Mine in Greene County, Pennsylvania. The percentage of coal ash at the sampled locations is denoted by the colorbar. The sampled locations  $\{s_i\}$  are fixed, and regularly spaced in a grid. As  $s$  are coordinates, they vary continuously over the spatial domain which is Robena Mine. The quantity of interest is the percentage of ash coal at these locations  $\{Z(s_i)\}$ . A mining engineer could be interested in predicting the ash distributions and the washability characteristics of coal along a seam in advance of mining. A spatial statistician would be able to produce a statistical model to predict ash concentrations between sampled points as well as quantify related uncertainties. Once a reasonable model that accounts for both the global trends and the local dependencies in the data is found and validated, the mining engineer could proceed to try and fill in the gaps, in other words, to estimate the percentage of coal ash at missing grid points based on the sampled percentages.

## 2.2. Aerial unit data / spatial data on lattices.

*Note 22.* Sometimes observations are collected over areal units such as pixels, census districts, or tomographic bins. In such cases, the random field models  $\{Z(s); s \in S\}$  have a discrete index set  $S$ . The aims are usually, noise removal from an image and smoothing rather than interpolation.

*Note 23.* Mathematically speaking, the index set  $S$  of the data  $\{Z(s)\}$  is a fixed (not random) and finite collection of points (locations)  $s \in S$ . The locations  $s \in S$  can be irregular or arranged in a regular grid. Often, there is a natural adjacency relation or neighborhood

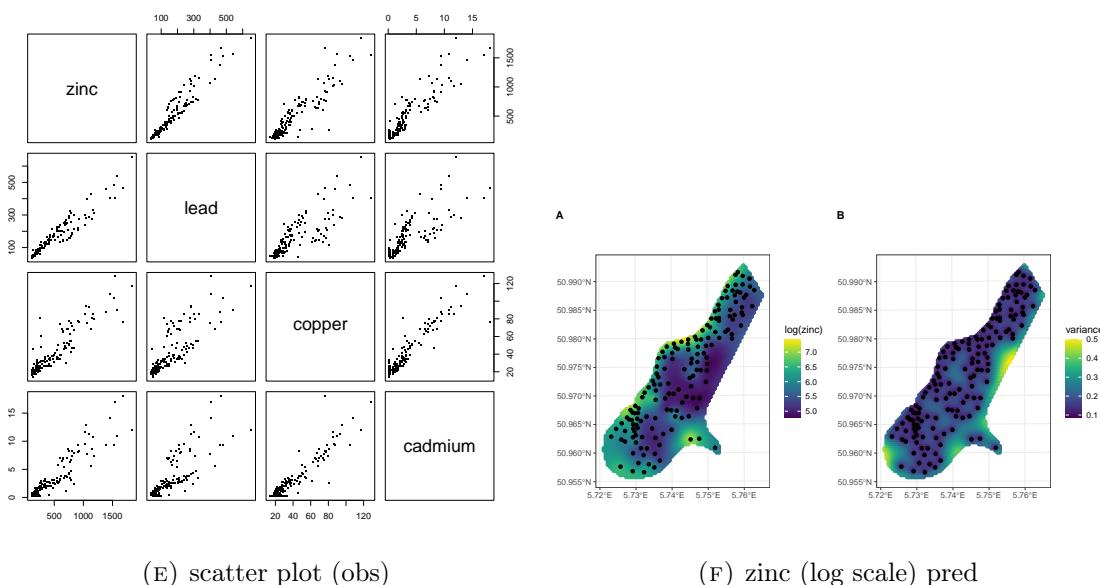


FIGURE 2.1. Map of the meuse dataset



FIGURE 2.2. (Coal ash data set) Percentage of coal ash at 208 locations.

structure. Often, datum  $Z(s)$  is a random vector at location  $s \in S$  and it represents an integral or average of the quantity of interest over some region represented by  $s \in S$ .

**Example 24.** In a UK epidemiological study,  $S$  may be the centroids of the UK counties, and  $Z(s)$  may represent the average value of a characteristic in county  $s$ . In image processing,  $S$  may be a grid of pixels (locations are fixed and regular). In statistical physics,  $S$  may be a collection of atoms and genuinely finite (locations are fixed and regular).

**Example 25.** (Image restoration data) Figure 2.3a shows an (observed) image from a gray-scale photo-micrograph of the micro-structure of the Ferrite-Pearlite steel obtained by PNNL's project supported by DoE. The lighter part is ferrite while the darker part is pearlite. We focus our analysis on the first quarter fragment of size  $240 \times 320$  pixels (red frame). This image is contaminated by noise due to the instrument errors. Interest lies in removing the noise (denoising) and recovering the real image. Figure 2.3b shows the restored image after appropriate statistical processing. Here the locations are pixels arranged in a fixed regular grid (hence discrete and not continuous). The each observation  $Z(s)$  is the color of a pixel  $s$ ; here it is scalar as the observed pixels are in tones of grey, however it could be a 3D if the pixels were colored.

**Example 26.** (North Carolina SIDS data set) Figure 2.4a shows the total number of deaths from Sudden Infant Death Syndrome (SIDS) in 1974 for each of the 100 counties in North Carolina. Figure 2.4b shows the corresponding live births in each county and same period. This is the R data set `nc{spdep}`. The centroids of the counties do not lie on a regular grid. The sizes and shapes of the counties vary and can be quite irregular. The recorded counts are not tied to a precise location but tallied up county-wise. This kind of accumulation over administrative units is usual for privacy-sensitive data in, for instance, the crime or public health domains. A public health official could be interested in spatial patterns; e.g., whether

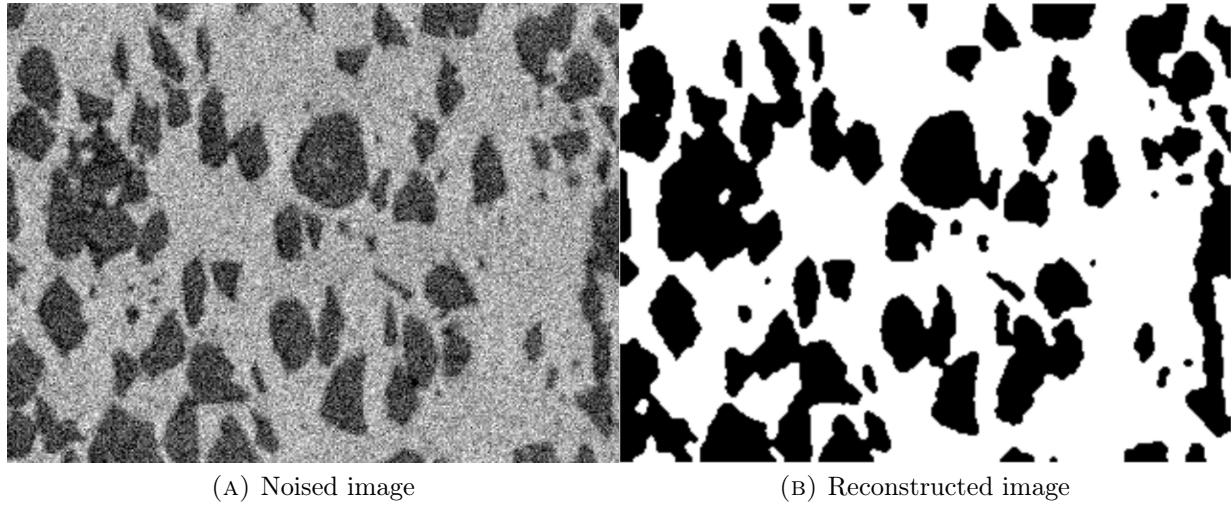


FIGURE 2.3. Ferrite-Pearlite steel image (Image restoration)

or not there are clusters of counties with a high incidence of SIDS, or areas where the SIDS counts are higher than what would be expected based on the number of live births in the area. Perhaps, we can eyeball the figures and see that there is a higher SIDS rate in the north-east areas compared to the north-west with similar birth numbers. A statistician can develop a statistical model providing inference about such questions.

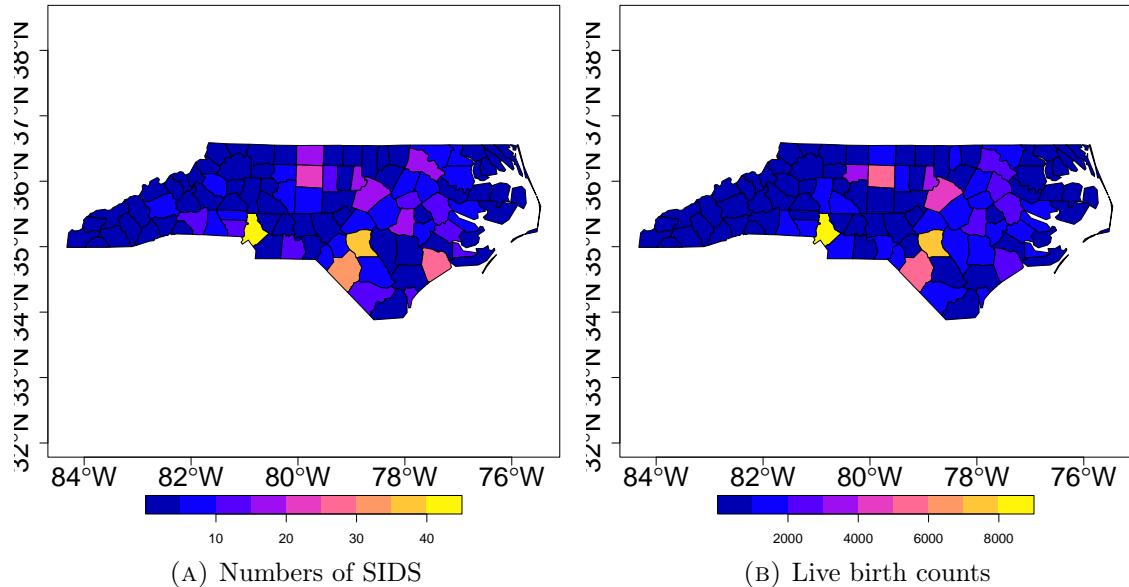


FIGURE 2.4. North Carolina SIDS data (Aerial unit data)

**Example 27.** (Columbus Columbus OH data set) Figures 2.5a, 2.5b, and 2.5c show the Property crime (number per thousand households) in 49 districts in Columbus in 1980, as  
Page 9      Created on 2024/10/17 at 18:33:08      by Georgios Karagiannis

well as the average value of the house (in 1,000 USD), and the average household income (in 1,000 USD). This is the R dataset `columbus{spdep}`. The recorded counts are not tied to a precise household but tallied up county-wise (aerial unit) for privacy reasons. The locations  $\{s_i\}$  are these areal unites; we observe they are not in a grid, and they do not have the same size or shape. Interest may lie to find whether high rates of crime are clustered in a particular areas, and if yes, perhaps what is the association of it with the value of the houses in the area.



FIGURE 2.5. Columbus Columbus OH spatial analysis dataset

### 2.3. Spatial point patterns.

*Note 28.* Sometimes the locations at which events occur are random. Typical examples include locations of trees in forests, outbreaks of forest fires, or epicentres of earthquakes. Such random patterns of locations are said to form a point pattern.

*Note 29.* Rigorously, the spatial domain  $S$  is a random set of points; specifically a point random field, in  $\mathbb{R}^d$  at which some events happened.

*Note 30.* In the simplest case, no covariate for  $Z$  is specified, and hence  $Z(s)$  represents only the occurrences of an even at location  $s$ , one could think of the data taking scalar values  $Z(s) = 1$  or  $Z(s) = 0$  when the event has occurred or not for all  $s \in S$ . We will refer to it as a spatial point random field

*Note 31.* In the most general case,  $Z(s)$  is a random vector at location  $s \in S$  (eg other covariates are associated to the location  $s$ ); these covariates are called marked variables. We will refer to it as a Marked spatial point random field.

*Note 32.* Questions in the spatial point pattern problems are mainly whether the pattern of locations is exhibiting complete spatial randomness, clustering (aggregation), or regularity (repulsiveness). In the marked spatial point random field where additional covariates are measured, we could possibly investigate the factors/variables associated to this behavior as well. A statistical approach to address such questions is needed as different observers may disagree on the amount of clustering or randomness. Usually patters from a completely random field may appear to be wrongly clustered when just eyeballed by an individual.

**Example 33.** (Tropical rain forest trees in Barro/Colorado) Figure 2.6 shows the positions (dots) of 3605 Beilschmiedia trees in a  $1000 \times 5000$  meter rectangular stand in a tropical rain forest at Barro Colorado Island, Panama. All spatial coordinates are in the Cartesian coordinate system and in meters. Dataset is available from the R package `bei{spatstat}`. The scientific question may be if the trees are distributed over the area in a uniform way, they form clusters, or they are arrange in a specific pattern. Here, the locations of the dots/trees are not fixed but random/uncertain and of course they are matter of inference. This is a point random field as each location is associated to an occurrence only and not any other covariate. The statistician's task is to design models able to test and quantify heterogeneity/homogeneity.

**Example 34.** The domain scientist is interested in knowing whether the spatial locations are completely spatially random, clustered, or regularly distributed.

- (1) (Japanese pine trees) The Japanese pine trees dataset in R `japanesepines{spatstat}` represents locations of 65 saplings of Japanese black pine (*Pinus thunbergii*) in a  $5.7 \times 5.7$  square meter sampling region in a natural forest.
- (2) (California Redwoods) The data represent the locations of 62 seedlings and saplings of California Giant Redwood (*Sequoiadendron giganteum*) recorded in a square sampling region.

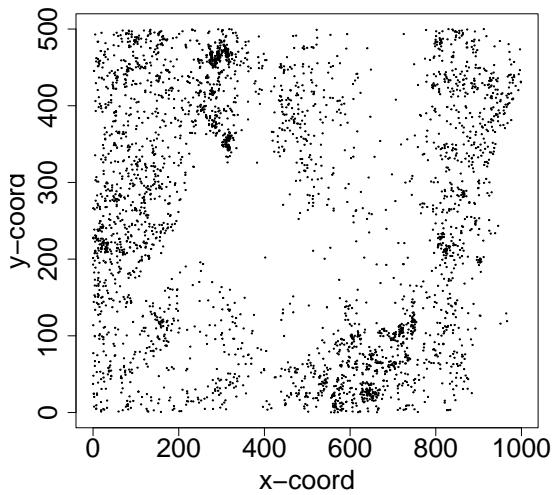


FIGURE 2.6. Locations of tropical rain forest trees in Barro/Colorado (Spatial point pattern data)

- (3) (Biological Cells) The data record the locations of the centres of 42 biological cells observed under optical microscopy in a histological section. The microscope field-of-view has been rescaled to the unit square.

Figure 2.7 presents the aforesaid spatial point patterns. It is difficult to eyeball, perhaps Figure 2.7a of the Japanese pine trees seems neither clustered nor regularly distributed but rather completely randomly distributed; Figure 2.7b of the California redwoods shows a clustered pattern; and Figure 2.7c shows a regular pattern.

**Example 35.** (Longleaf Pines Point Pattern) Figure 2.8 shows locations (as Cartesian coordinates) and relative diameters at breast height in dbh (as the size of the dot) of all longleaf pine trees in the 24ha region of the Wade Tract, an old-growth forest in Thomas County, Georgia in 1979. Dataset is available from R package `bei{spatstat}`. Longleaf pine is a fire-adapted species of trees. The domain scientist is interested in knowing whether the spatial locations are spatially random, or clustered, if large (small) trees cluster and how do large and small trees interact. A statistician can design models able to quantify such notions and provide inference. Here, the locations are random (not fixed) and in fact an object of inference. The diameter at breast height recorded along with the tree's location is the marked variable, and hence, the whole random field is a marked point random field.



FIGURE 2.7. Spatial Point Patterns

### 3. UNCERTAINTY QUANTIFICATION AND MODELING

*Note 36.* In spatial problems, uncertainty is expressed probabilistically through a spatial random field (or a stochastic process), which can be written most generally as

$$(3.1) \quad \{Y(s); s \in \mathcal{S}\},$$

To be  
defined  
rigorously  
later



FIGURE 2.8. Longleaf Pines Point Pattern (Spatial point data)

Here  $Y(s)$  is the random attribute value at location  $s$ ,  $\mathcal{S} \subset \mathbb{R}^d$  is a subset of  $\mathbb{R}^d$  ( $d = 1, 2, 3$ ), contained in  $\mathcal{S}$  is a possibly random fixed or random set  $S$  that indexes those parts of  $\mathcal{S}$  relevant to the scientific study.

### **Spatial process model.**

*Note 37.* The scientific uncertainty (i.e. the (known) uncertainty about the scientific problem) is expressed via the spatial process model. E.g., uncertainty about the real picture in Fig. 2.3a.

*Note 38.* This spatial random field can be a: geostatistical random field, lattice random field, or point random field depending on the principal spatial statistical area (Section 2) the application is associated with.

*Note 39.* The joint probability model defined by the random  $\{Y(s); s \in S\}$  is

$$(3.2) \quad \text{pr}(Y, S) = \text{pr}(Y|S) \text{pr}(S)$$

*Note 40.* The specification of  $\text{pr}(S)$  represents the three principal spatial statistical areas. E.g., for spatial data on lattices or point referenced data problems where the locations are fixed and not uncertain, we can consider  $\text{pr}(Y, S) = \text{pr}(Y|S)$  with  $\text{pr}(S) = 1_{\{S\}}(S)$  and hence ignore  $S$  and  $\text{pr}(S)$  from the notation.

### **Data model.**

*Note 41.* The measurement uncertainty is quantified via the data model. E.g. the “noisy image” in Fig. 2.3a.

*Note 42.* The data model is specified to be the conditional distribution of the data  $Z$  given the spatial random field  $Y$  and the  $S$ , namely

$$(3.3) \quad \text{pr}(Z|Y, S)$$

*Note 43.* If the data are assumed to be conditionally independent, such as  $Z(s) \perp Z(s') | Y, S$  then

$$(3.4) \quad \text{pr}(Z|Y, S) = \prod_{i=1}^n \text{pr}(Z(s_i) | Y, S)$$

*Note 44.* The spatial statistical dependence of in  $Z$ , articulated by the First Law of Geography, follows by

$$\text{pr}(Z|S) = \int \text{pr}(Z|Y, S) \text{pr}(Y|S) dY$$

### The hierarchical statistical model.

*Note 45.* To sum up the (known) uncertainty in spatial the statistics problem is expressed via the so called Hierarchical spatial model

$$(3.5) \quad \begin{cases} Z|Y, S & \text{data model} \\ Y, S & \text{spatial process model} \end{cases}$$

with

$$\text{pr}(Z, Y, S) = \text{pr}(Z|Y, S) \text{pr}(Y|S) \text{pr}(S)$$

Our interest it to learn  $\text{pr}(Z^*|Z)$  or  $\text{pr}(Y^*|Z)$  at any unseen locations  $s^*$ .

### The Empirical (Bayes) hierarchical model.

*Note 46.* Often the decomposition (3.5) is parametrized with respect to unknown parameters  $\theta \in \Theta$  we wish to learn given the observables; this is often called the Empirical hierarchical model i.e.

$$(3.6) \quad \begin{cases} Z|Y, S, \theta & \text{data model} \\ Y, S|\theta & \text{spatial process model} \end{cases}$$

with

$$\text{pr}(Z, Y, S|\theta) = \text{pr}(Z|Y, S, \theta) \text{pr}(Y|S, \theta) \text{pr}(S|\theta)$$

Our additional additional interest may be to learn the unknown  $\theta$  via MLE

$$\hat{\theta} = \arg \min_{\theta} (\text{pr}(Z|\theta))$$

### The Bayesian hierarchical model.

*Note 47.* In Bayesian statistics, the hierarchical model in (3.5) is completed by the  $\theta \sim \text{pr}(\cdot)$  adding a third layer leading to the Bayesian hierarchical model

$$(3.7) \quad \begin{cases} Z|Y, S, \theta & \text{data model} \\ Y, S|\theta & \text{spatial process model} \\ \theta & \text{hyper-parameter prior model} \end{cases}$$

with

$$\text{pr}(Z, Y, S, \theta) = \text{pr}(Z|Y, S, \theta) \text{pr}(Y|S, \theta) \text{pr}(S|\theta) \text{pr}(\theta)$$

Our additional additional interest may be to learn the posterior of  $\theta \text{pr}(\theta|Z)$ .

**Example 48.** (A naive example: NOAA weather data) As dataset we consider a fraction from daily data originated from the US National Oceanic and Atmospheric Administration (NOAA) National Climatic Data Center and can be obtained from the IRI/LDEO Climate Data Library at Columbia University<sup>2</sup>. The dataset is available from the R package in GitHub “andrewzm/STRbook”. In this example we focus on maximum temperature (Tmax) in degrees Fahrenheit ( $^{\circ}\text{F}$ ) at 138 weather stations in the central USA (between  $32^{\circ}\text{N}$ - $46^{\circ}\text{N}$  and  $80^{\circ}\text{W}$ - $100^{\circ}\text{W}$ ), recorded in 1st of May 1993. The data are not complete, in the sense that there are missing measurements at various stations and at various time points, and the stations themselves are obviously not located everywhere in the central USA. (Figure Figure 3.1)

This data set contains  $n = 138$  observations  $\{(Z_i, s_i)\}_{i=1}^n$  where the  $i$ -th observation contains the maximum temperature (Tmax) in degrees Fahrenheit ( $^{\circ}\text{F}$ )  $Z_i$  at location specified by coordinates  $s_i = (s_{1,i}, s_{2,i})^{\top}$  that is the latitude degrees north of the Equator  $s_{2,i}$ , and longitude degrees west of Greenwich  $s_{1,i}$ . See Figure 3.1.

This is definitely a geostatistics problem. Here, we will present a naive way to model it by reflecting what we discussed earlier.

**Data model:** One may consider that the observations  $\{Z_i\}$  at each location are the result of observing the real (hence unknown) maximum temperature  $Y_i$  but contaminated by “additive random noise” with unknown scale  $\sigma > 0$  due to instrumental error, i.e.

$$Z_i = Y_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad i = 1, \dots, n \implies Z_i|Y_i, \sigma^2 \stackrel{\text{ind}}{\sim} N(Y_i, \sigma^2), \quad i = 1, \dots, n$$

**Spatial process model:** One may consider that the real maximum temperature  $Y(s)$  (over the spatial domain  $s \in S$ ) is a function where at each finite set of locations  $\{s_i\}_{i=1}^n$  follows a Normal distribution with a mean  $\mu$  with  $[\mu]_i = \mu(s_i)$  parameterized

That's a  
snapshot  
for what  
follows.

---

<sup>2</sup><http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCDC/.DAILY/.FSOD/>



FIGURE 3.1. NOAA weather data (1 May 1993)

as

$$\mu(s) = \beta_0 + \beta_1 s_1 + \beta_2 s_2 + \beta_{12} s_1 s_2, \text{ at a location } s = (s_1, s_2)^\top$$

with unknown parameter  $\beta$ , and covariance matrix  $[C]_{i,j} = c(s_i, s_j)$  parameterized with covariance function

$$c(s, s') = \sigma^2 \exp(-\phi \|s - s'\|)$$

to impose that nearer locations cause stronger dependences in the model. Here  $\beta$ ,  $\phi$ , and  $\sigma^2$  are unknown parameters.

**Hierarchical model:** To sum up, we have build the hierarchical model

$$(3.8) \quad \begin{cases} Z|Y, \sigma^2 \sim N_n(Y, I\sigma^2), & \text{data model} \\ Y|\sigma^2, \beta, \phi \sim N_n(S\beta, C), & \text{spatial process model} \end{cases}$$

Figure 3.2 shows the hierarchical spatial model (3.8) for different values of  $\theta = (\sigma^2, \beta, \phi)$ ; the surface corresponds to the spatial process  $\{Y(s); s \in \mathbb{R}^2\}$  and is presented at three different instances each of them with different values for  $(\beta, \phi)$ , while the dots correspond to the observations  $\{(Z(s_i), s_i)\}_{i=1}^n$  and their deviation from the spatial process is controlled by  $\sigma^2$ . Note that marginally

$$Z|\sigma^2, \beta, \phi \sim N_n(S\beta, \sigma^2(I + C)).$$

**Bayesian hierarchical model:** If we work on the fully Bayesian framework, we can complete the model with priors on  $\theta = (\sigma^2, \beta, \phi)$  for instance  $\sigma^2 \sim IG(\kappa_\sigma, \lambda_\sigma)$ ,  $\phi \sim$



FIGURE 3.2. Examples representing the hierarchical spatial model (3.8) for different values of  $\theta = (\sigma^2, \beta, \phi)$

$\text{IG}(\kappa_\phi, \lambda_\phi)$ , and  $\beta \sim N_4(b, Iv)$ , with some known hyper-parameters  $\kappa_\sigma, \lambda_\sigma, \kappa_\phi, \lambda_\phi, b, v$ . To sum up, we have build the Bayesian model

$$\left\{ \begin{array}{ll} Z|Y, \sigma^2 \sim & N_n(Y, I\sigma^2), \text{ data model} \\ Y|\sigma^2, \beta, \phi \sim & N_n(S\beta, C), \text{ spatial process model} \\ \beta \sim & N_4(b, Iv), \text{ hyper-parameter prior model} \\ \sigma^2 \sim & \text{IG}(\kappa_\sigma, \lambda_\sigma), \text{ hyper-parameter prior model} \\ \phi \sim & \text{IG}(\kappa_\phi, \lambda_\phi), \text{ hyper-parameter prior model} \end{array} \right.$$

#### 4. SPATIO-TEMPORAL STATISTICS

*Note 49.* Spatio-temporal data arise when information is both spatially and temporally referenced. Any of the aforesaid spatial statistics cases can be extended and become temporal. Such methods will be discussed in the Epiphany term. This is where we aim at the end of the module.

**Example 50.** We extend the dataset in the Spatial statistics Example 48 by considering time.

The dataset is available from the R package in GitHub “andrewzm/STRbook”. These daily data originated from the US National Oceanic and Atmospheric Administration (NOAA) National Climatic Data Center and can be obtained from the IRI/LDEO Climate Data Library at Columbia University. The data set we consider consists the daily maximum temperature (Tmax) in degrees Fahrenheit ( $^{\circ}\text{F}$ ) at weather stations in the central USA (between  $32^{\circ}\text{N}$ – $46^{\circ}\text{N}$  and  $80^{\circ}\text{W}$ – $100^{\circ}\text{W}$ ), recorded between May-July in year 1993. These data are considered to be discrete and regular in time (daily) and geostatistical and irregular in space. However, the data are not complete, in that there are missing measurements at various stations and at various time points, and the stations themselves are obviously not located everywhere in the central USA. See Figure 4.1.

Interest lies on not only how the maximum temperature is only distributed over the spatial domain but also how it evolves during the time.

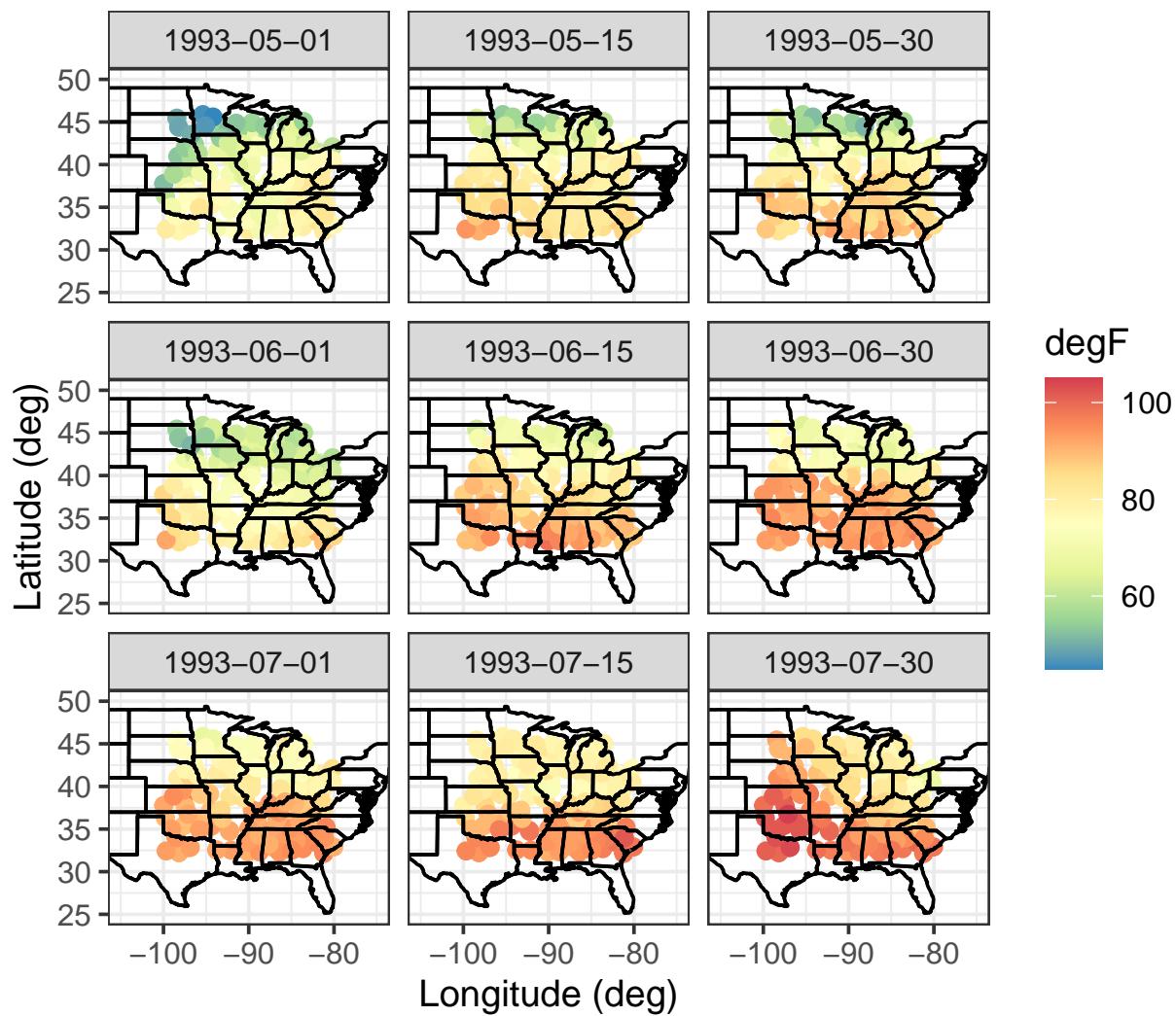


FIGURE 4.1. NOAA daily weather data

## Lecture notes part 2: Point referenced data modeling / Geostatistics

Lecturer &amp; author: Georgios P. Karagiannis

georgios.karagiannis@durham.ac.uk

**Aim.** To introduce point referenced data modeling (geostatistics) with particular focus on concepts spatial variables, random fields, semi-variogram, kriging, change of support, multivariate geostatistics, for Bayesian and classical inference.

### Reading list & references:

- [1] Cressie, N. (2015; Part I). Statistics for spatial data. John Wiley & Sons.
- [2] Kent, J. T., & Mardia, K. V. (2022). Spatial analysis (Vol. 72). John Wiley & Sons.
- [3] Chiles, J. P., & Delfiner, P. (2012). Geostatistics: modeling spatial uncertainty (Vol. 713). John Wiley & Sons.
- [4] Wackernagel, H. (2003). Multivariate geostatistics: an introduction with applications. Springer Science & Business Media.
- [5] Gaetan, C., & Guyon, X. (2010; Ch 2 & 5.1). Spatial statistics and modeling (Vol. 90). New York: Springer.

## Part 1. Basic stochastic models & related concepts for model building

*Note 1.* We discuss basic stochastic models and concepts for modeling point referenced data in the Geostatistics framework.

### 1. RANDOM FIELDS (OR STOCHASTIC PROCESSES)

**Definition 2.** A random field (or stochastic process, or random function)  $Z = (Z(s); s \in \mathcal{S})$  taking values in  $\mathcal{Z} \subseteq \mathbb{R}^q$ ,  $q \geq 1$  is a family of random variables  $\{Z(s) := Z(s; \omega); s \in \mathcal{S}, \omega \in \Omega\}$  defined on the same probability space  $(\Omega, \mathfrak{F}, \text{pr})$  and taking values in  $\mathcal{Z}$ . The label  $s \in \mathcal{S}$  is called site, the set  $\mathcal{S} \subseteq \mathbb{R}^d$  is called the (spatial) set of sites at which the random field is defined, and  $\mathcal{Z}$  is called the state space of the field.

*Note 3.* Given a set of sites  $\{s_1, \dots, s_n\}$ , with  $s_i \in \mathcal{S}$  and  $n \in \mathbb{N}$ , the random vector  $(Z(s_1), \dots, Z(s_n))^\top$  has a well-defined probability distribution that is completely determined by its joint CDF

$$F_{s_1, \dots, s_n}(z_1, \dots, z_n) := \text{pr}(Z(s_1) \leq z_1, \dots, Z(s_n) \leq z_n)$$

The family of all finite-dimensional distributions (or fidi's) of  $Z$  is called the spatial distribution of the process .

*Note 4.* According to Kolmogorov Theorem 5, to define a random field model, one must specify the joint distribution of  $(Z(s_1), \dots, Z(s_n))^\top$  for all of  $n$  and all  $\{s_i \in S; i = 1, \dots, n\}$  in a consistent way.

**Proposition 5.** (*Kolmogorov consistency theorem*) Let  $pr_{s_1, \dots, s_n}$  be a probability on  $\mathbb{R}^n$  with join CDF  $F_{s_1, \dots, s_n}$  for every finite collection of points  $s_1, \dots, s_n$ . If  $F_{s_1, \dots, s_n}$  is symmetric w.r.t. any permutation  $\mathbf{p}$

$$F_{\mathbf{p}(s_1), \dots, \mathbf{p}(s_n)}(z_{\mathbf{p}(1)}, \dots, z_{\mathbf{p}(n)}) = F_{s_1, \dots, s_n}(z_1, \dots, z_n)$$

for all  $n \in \mathbb{N}$ ,  $\{s_i \in S\}$ , and  $\{z \in \mathbb{R}\}$ , and all if all permutations  $\mathbf{p}$  are consistent in the sense

$$\lim_{z_n \rightarrow \infty} F_{s_1, \dots, s_n}(z_1, \dots, z_n) = F_{s_1, \dots, s_{n-1}}(z_1, \dots, z_{n-1})$$

or all  $n \in \mathbb{N}$ ,  $\{s_i \in S\}$ , and  $\{z_i \in \mathbb{R}\}$ , then there exists a random field  $Z$  whose fidi's coincide with those in  $F$ .

**Example 6.** Let  $n \in \mathbb{N}$ , let  $\{X_i : \mathcal{S} \rightarrow \mathbb{R}; i = 1, \dots, n\}$  be a set of constant functions, and let  $\{Z_i \sim N(0, 1)\}_{i=1}^n$  be a set of independent random variables. Then

$$(1.1) \quad \tilde{Z}(s) = \sum_{i=1}^n Z_i X_i(s), \quad s \in S$$

is a well defined random field as it satisfies Theorem 5.

### 1.1. Mean and covariance functions.

**Definition 7.** The mean function  $\mu(\cdot)$  and covariance function  $c(\cdot, \cdot)$  of a random field  $(Z(s); s \in \mathcal{S})$  are defined as

$$(1.2) \quad \mu(s) = E(Z(s)), \quad \forall s \in \mathcal{S}$$

$$(1.3) \quad c(s, s') = \text{Cov}(Z(s), Z(s')) = E((Z(s) - \mu(s))(Z(s') - \mu(s'))^\top), \quad \forall s, s' \in \mathcal{S}$$

**Example 8.** For (1.1), the mean function is  $\mu(s) = E(\tilde{Z}_s) = 0$  and covariance function is

$$\begin{aligned} c(s, s') &= \text{Cov}(Z(s), Z(s')) = \text{Cov}\left(\sum_{i=1}^n Z_i X_i(s), \sum_{j=1}^n Z_j X_j(s')\right) \\ &= \sum_{i=1}^n X_i(s) \sum_{j=1}^n X_j(s') \underbrace{\text{Cov}(Z_i, Z_j)}_{1(i=j)} = \sum_{i=1}^n X_i(s) X_i(s') \end{aligned}$$

#### 1.1.1. Construction of covariance functions.

*Note 9.* What follows provides the means for checking and constructing covariance functions.

**Proposition 10.** The function  $c : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$ ,  $\mathcal{S} \subseteq \mathbb{R}^d$  is a covariance function iff  $c(\cdot, \cdot)$  is semi-positive definite; i.e.

$$\forall n \in \mathbb{N} - \{0\}, \forall (a_1, \dots, a_n) \in \mathbb{R}^n \text{ and } \forall (s_1, \dots, s_n) \in \mathcal{S}^n : \sum_{i=1}^n \sum_{j=1}^n a_i a_j c(s_i, s_j) \geq 0$$

or in other words, the Gram matrix  $(c(s_i, s_j))_{i,j=1}^n$  is non-negative definite for any  $\{s_i\}_{i=1}^n$ ,  $n \in \mathbb{N} - \{0\}$ .

**Example 11.**  $c(s, s') = 1(\{s = s'\})$  is a proper covariance function because

$$\sum_i \sum_j a_i a_j c(s_i, s_j) = \sum_i a_i^2 \geq 0, \quad \forall a$$

*Note 12.* One way to construct a c.f  $c$  is to set  $c(s, s') = \psi(s)^\top \psi(s')$ , for a given vector of basis functions  $\psi(\cdot) = (\psi_1(\cdot), \dots, \psi_n(\cdot))$ .

*Proof.* From Proposition 10, as

$$\sum_i \sum_j a_i a_j c(s_i, s_j) = (\psi a)^\top (\psi a) \geq 0, \quad \forall a \in \mathbb{R}^n$$

□

## 2. SECOND ORDER RANDOM FIELDS (OR SECOND ORDER PROCESSES)

*Note 13.* We introduce a particular class of random fields whose mean and covariance functions exist and which can be used for spatial data modeling.

**Definition 14.** Second order random field (or second order process)  $(Z(s); s \in \mathcal{S})$  is called the random field where  $E((Z(s))^2) < \infty$  for all  $s \in \mathcal{S}$ .

**Example 15.** In second order random field  $(Z(s); s \in \mathcal{S})$  the associated mean function  $\mu(\cdot)$  and covariance function  $c(\cdot, \cdot)$  exist, because  $c(s, t) = E(Z(s)Z(t)) - E(Z(s))E(Z(t))$  for  $s, t \in \mathcal{S}$ .

## 3. GAUSSIAN RANDOM FIELD (OR GAUSSIAN PROCESS)

*Note 16.* Gaussian random field (GRF) is a particular class of second order random field which is widely used in spatial data modeling due to its computational tractability.

Also

**Definition 17.**  $(Z(s); s \in \mathcal{S})$  is a Gaussian random field (GRF) or Gaussian process (GP) on  $\mathcal{S}$  if for any  $n \in \mathbb{N}$  and for any finite set  $\{s_1, \dots, s_n; s_i \in \mathcal{S}\}$ , the random vector  $(Z(s_1), \dots, Z(s_n))^\top$  follows a multivariate normal distribution.

Example Proposition

**Proposition 18.** A GP  $(Z(s); s \in \mathcal{S})$  is fully characterized by its mean function  $\mu : S \rightarrow \mathbb{R}$  with  $\mu(s) = E(Z(s))$ , and its covariance function with  $c(s, s') = Cov(Z(s), Z(s'))$ .

*Note 19.* Hence, we denote the GP as  $Z(\cdot) \sim \mathcal{GP}(\mu(\cdot), c(\cdot, \cdot))$ .

*Note 20.* When using GP for spatial modeling we just need to specify its functional parameters i.e. the mean and covariance functions.

*Note 21.* Popular forms of mean functions are polynomial expansions, such as  $\mu(s) = \sum_{j=0}^{p-1} \beta_j s^j$  for tunable unknown parameter  $\beta$ . A popular form of covariance functions (c.f.), for tunable unknown parameters  $\phi > 0$ , and  $\sigma^2 > 0$ , are

- (1) Exponential c.f.  $c(s, s') = \sigma^2 \exp(-\phi \|s - s'\|_1)$
- (2) Gaussian c.f.  $c(s, s') = \sigma^2 \exp(-\phi \|s - s'\|_2^2)$
- (3) Nugget c.f.  $c(s, s') = \sigma^2 1(s = s')$

**Example 22.** Recall your linear regression lessons where you specified the sampling distribution to be  $y_x | \beta, \sigma^2 \stackrel{\text{ind}}{\sim} N(x^\top \beta, \sigma^2)$ ,  $\forall x \in \mathbb{R}^d$ . Well that can be considered as a GP  $Z(\cdot) \sim \mathcal{GP}(\mu(\cdot), c(\cdot, \cdot))$  with  $\mu(x) = x^\top \beta$  and  $c(x, x') = \sigma^2 1(x = x')$  in (3).

**Example 23.** Figures 3.1 & 3.2 presents realizations of GRF  $Z(\cdot) \sim \mathcal{GP}(\mu(\cdot), c(\cdot, \cdot))$  with  $\mu(s) = 0$  and differently parameterized covariance functions in 1D and 2D. In 1D the code to simulate the GP is given in Algorithm 1. Note that we actually discretize it and simulate it from the fidi.

---

**Algorithm 1** R script for simulating from a GP ( $Z(s); s \in \mathbb{R}^1$ ) with  $\mu(s) = 0$  and  $c(s, t) = \sigma^2 \exp(-\phi \|s - t\|_2^2)$

---

```
# set the GP parameterized mean and covariance function
mu_fun <- function(s) { return (0) }
cov_fun_gauss <- function(s,t,sig2,phi) {
    return ( sig2*exp(-phi*norm(c(s-t),type="2")**2) )
}
# discretize the problem in n = 100 spatial points
n <- 100
s_vec <- seq(from = 0, to = 5, length = n)
mu_vec <- matrix(nrow = n, ncol = 1)
Cov_mat <- matrix(nrow = n, ncol = n)
# compute the associated mean vector and covariance matrix of the n=100 dimensional
Normal r.v.
sig2_val <- 1.0 ;
phi_val <- 5
for (i in 1:n) {
    mu_vec[i] <- mu_fun(s_vec[i])
    for (j in 1:n) {
        Cov_mat[i,j] <- cov_fun_gauss(s_vec[i],s_vec[j],sig2_val,phi_val)
    }
}
# simulate from the associated distribution
z_vec <- mu_vec + t(chol(Cov_mat))%*%rnorm(n, mean=0, sd=1)
# plot the path (R produces a line plot)
plot(s_vec, z_vec, type="l")
abline(h=0,col="red")
```

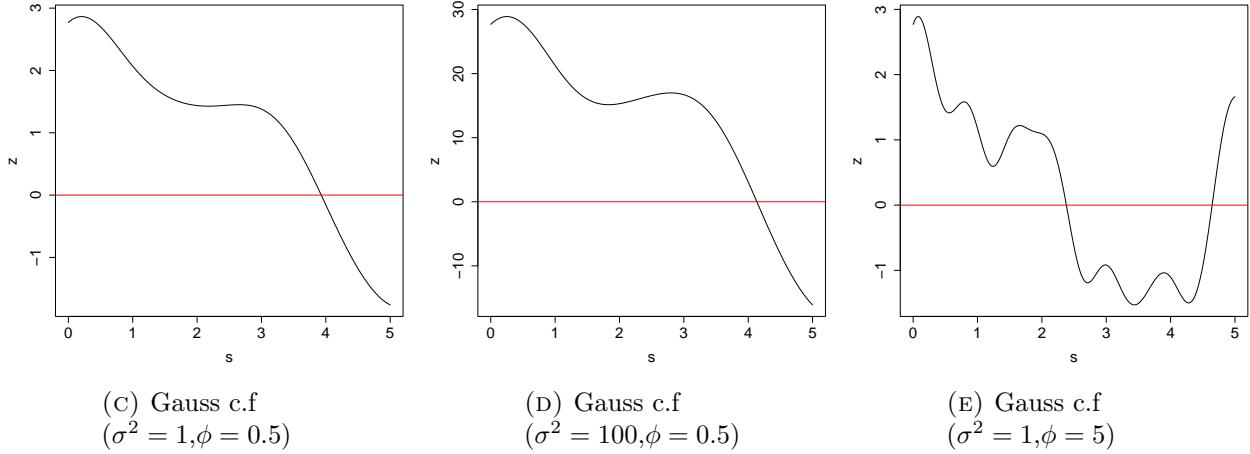
---

Nugget c.f. is the usual noise where the height of ups and downs are random and controlled by  $\sigma^2$  (Figures 3.1a & 3.1b ; Figures 3.2a & 3.2b). In Gaussian c.f. the height of ups and downs are random and controlled by  $\sigma^2$  (Fig.3.1c & 3.1d ; Figures 3.2c & 3.2d), and the spatial dependence / frequency of the ups and downs is controlled by  $\beta$  (Figures 3.1d & 3.1e ; Figures 3.2d & 3.2e). Realizations with different c.f. have different behavior (Figures 3.1a, 3.1d & 3.1e ; Figures 3.2a, 3.2d & 3.2e)



(A) Nugget c.f.  
( $\sigma^2 = 1$ )

(B) Nugget c.f.  
( $\sigma^2 = 100$ )



(C) Gauss c.f.  
( $\sigma^2 = 1, \phi = 0.5$ )

(D) Gauss c.f.  
( $\sigma^2 = 100, \phi = 0.5$ )

(E) Gauss c.f.  
( $\sigma^2 = 1, \phi = 5$ )

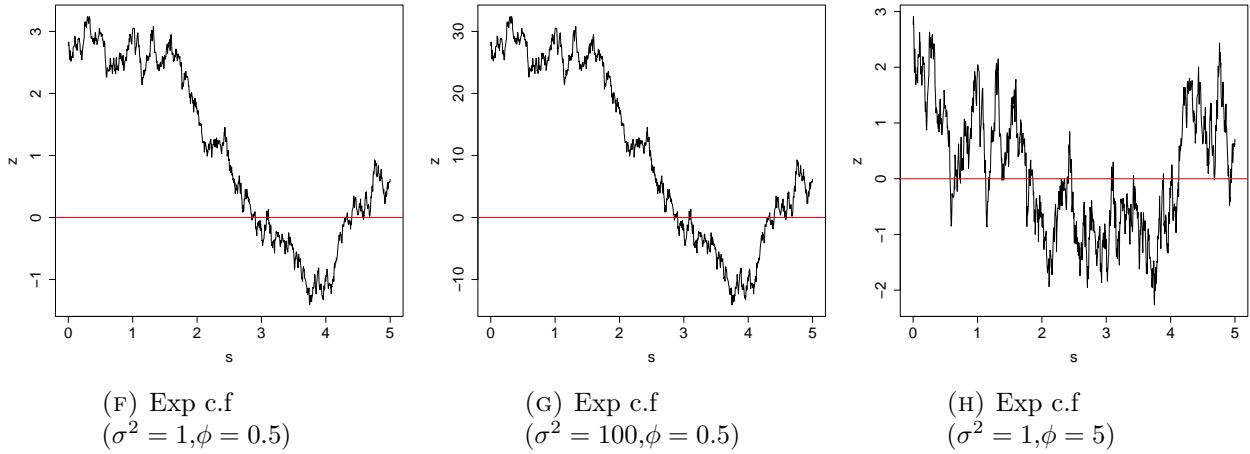


FIGURE 3.1. Realizations of GRF  $Z(\cdot) \sim \mathcal{GP}(\mu(\cdot), c(\cdot, \cdot))$  when  $s \in [0, 5]$  (using same seed)

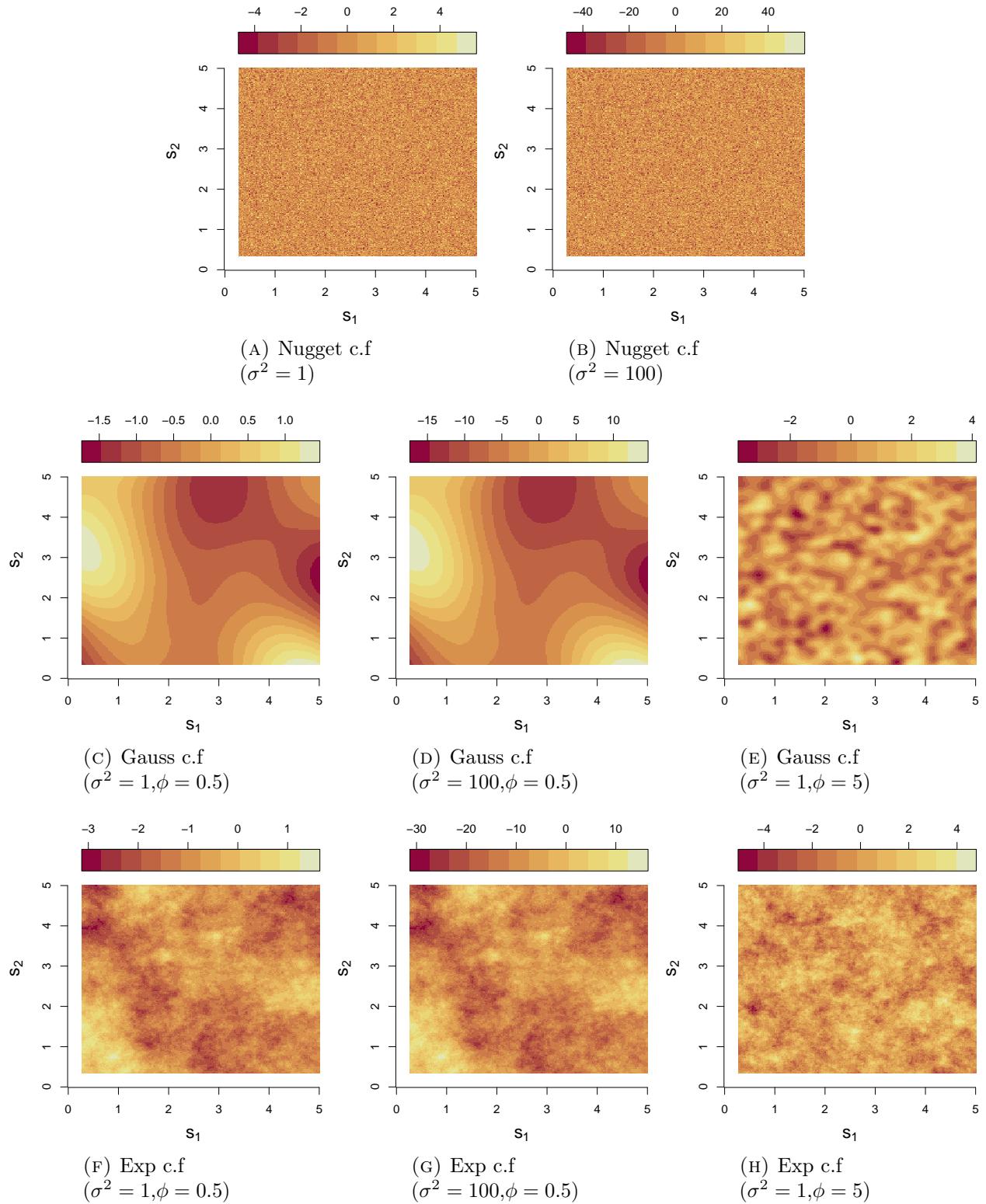


FIGURE 3.2. Realizations of GRF  $Z(\cdot) \sim \mathcal{GP}(\mu(\cdot), c(\cdot, \cdot))$  when  $s \in [0, 5]^2$  (using same seed)

#### 4. STRONG STATIONARITY

*Note 24.* We introduce a specific behavior of random field to build our models.

*Notation 25.* Formally, we define the separation (or lag) set as  $\mathcal{H} = \{h \in \mathbb{R}^d : s \in \mathcal{S}, s + h \in \mathcal{S}\}$  where  $\mathcal{S} \subseteq \mathbb{R}^d$  is the spatial domain for  $d = 1, 2, 3, \dots$ . However, we will consider cases where  $\mathcal{S} = \mathbb{R}^d$  and  $\mathcal{H} = \mathbb{R}^d$  for  $d = 1, 2, 3, \dots$  in Euclidean spaces.

**Definition 26.** A random field  $(Z(s); s \in \mathcal{S})$  is strongly stationary on  $\mathcal{S}$  if for all finite sets consisting of  $s_1, \dots, s_n \in \mathcal{S}$ ,  $n \in \mathbb{N}$ , for all  $k_1, \dots, k_n \in \mathbb{R}$ , and for all  $h \in \mathcal{H}$

$$\text{pr}(Z(s_1 + h) \leq k_1, \dots, Z(s_n + h) \leq k_n) = \text{pr}(Z(s_1) \leq k_1, \dots, Z(s_n) \leq k_n)$$

*Note 27.* Yuh... strong stationary may represent a behavior being too “restrictive” to be used for spatial data modeling as it is able to represent only limiting number of spatial dependencies.

#### 5. WEAK STATIONARITY (OR SECOND ORDER STATIONARITY)

*Note 28.* We introduce another weaker behavior of random field able to represent a larger class of spatial dependencies.

*Note 29.* Instead of working with the “restrictive” strong stationarity, we could just properly specify the behavior of the first two moments only; notice that Definition 26 implies that, given  $E((Z(s))^2) < \infty$ , it is  $E(Z(s)) = E(Z(s + h)) = \text{const} \dots$  and  $\text{Cov}(Z(s), Z(s')) = \text{Cov}(Z(s + h), Z(s' + h)) \stackrel{h=-s'}{=} \text{Cov}(Z(s - s'), Z(0)) = \text{funct of lag} \dots$

**Definition 30.** A random field  $(Z(s); s \in \mathcal{S})$  is called stationary random field (s.r.f.) (or weakly stationary or second order stationary) if it has constant mean and translation invariant covariance; i.e. for all  $s, s' \in \mathbb{R}^d$ ,

- (1)  $E((Z(s))^2) < \infty$  (finite)
- (2)  $E(Z(s)) = \mu$  (constant)
- (3)  $\text{Cov}(Z(s), Z(s')) = c(s' - s)$  for some even function  $c : \mathcal{S} \rightarrow \mathbb{R}$  (lag dependency)

**Definition 31.** Stationary (or weakly or second order stationary) covariance function is called the c.f. of a stationary random field.

#### 6. COVARIOGRAM

*Note 32.* We introduce the covariogram function able to express many aspects of the behavior of a (weakly) stationary random field and hence be used as statistical descriptive tool.

**Definition 33.** The covariogram function of a weakly stationary random field  $(Z(s); s \in \mathcal{S})$  is defined by  $c : \mathcal{H} \rightarrow \mathbb{R}$  with

$$c(h) = \text{Cov}(Z(s), Z(s+h)), \forall s \in \mathcal{S}, \forall h \in \mathcal{H}.$$

**Example 34.** For the Gaussian c.f.  $c(s, t) = \sigma^2 \exp(-\phi \|s - t\|_2^2)$  in (Ex. 20(2)), we may denote just

$$(6.1) \quad c(h) = c(s, s+h) = \sigma^2 \exp(-\phi \|h\|_2^2)$$

Observe that, in Figures 3.1 & 3.2, the smaller the  $\phi$ , the smoother the realization (aka slower changes). One way to justify this observation is to think that smaller values of  $\phi$  essentially bring the points closer by re-scaling spatial lags  $h$  in the c.f.

**Proposition 35.** If  $c : \mathcal{H} \rightarrow \mathbb{R}$  is the covariogram of a weakly stationary random field  $(Z(s); s \in \mathcal{S})$  then:

- (1)  $c(h) = c(-h)$  for all  $h \in \mathcal{H}$
- (2)  $|c(h)| \leq c(0) = \text{Var}(Z(s))$  for all  $h \in \mathcal{H}$
- (3)  $c(\cdot)$  is semi-positive definite; i.e. for all  $n \in \mathbb{N} - \{0\}$ ,  $(a_1, \dots, a_n) \in \mathbb{R}^n$ , and  $(s_1, \dots, s_n) \subseteq \mathcal{S}^n$

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j c(s_i - s_j) \geq 0$$

*Note 36.* Given there is some knowledge of the characteristic functions of a suitable distribution, the following spectral representation theorem helps in the specification of a suitable covariogram.

**Theorem 37.** (Bochner's theorem) Let  $c : \mathbb{R}^d \rightarrow \mathbb{R}$  be a continuous even real-valued function for  $d \geq 1$ . Then  $c(\cdot)$  is positive semidefinite (hence a covariogram of a stationary random field) if and only if it can be represented as

$$c(h) = \int_{\mathbb{R}^d} \exp(i\omega^\top h) dF(\omega)$$

where  $F$  is a symmetric positive finite measure on  $\mathbb{R}^d$  called spectral measure.

*Note 38.* In our course, we focus on cases where  $F$  has a density  $f(\cdot)$  i.e.  $dF(\omega) = f(\omega) d\omega$ .  $f(\cdot)$  is called spectral density of  $c(\cdot)$ , it is

$$c(h) = \int_{\mathbb{R}^d} \exp(i\omega^\top h) f(\omega) d\omega,$$

and it dies as  $\lim_{|h| \rightarrow \infty} c(h) = 0$

**Theorem 39.** If  $c(\cdot)$  is integrable, the spectral density  $f(\cdot)$  can be computed by inverse Fast Fourier transformation

$$f(\omega) = \left(\frac{1}{2\pi}\right)^d \int_{\mathbb{R}^d} \exp(-i\omega^\top h) c(h) dh$$

**Example 40.** Consider the Gaussian c.f.  $c(h) = \sigma^2 \exp(-\phi \|h\|_2^2)$  for  $\sigma^2, \phi > 0$  and  $h \in \mathbb{R}^d$ . Then, by using Theorem 37, the spectral density is

$$\begin{aligned} f(\omega) &= \left(\frac{1}{2\pi}\right)^d \int_{\mathbb{R}^d} \exp(-i\omega^\top h) \sigma^2 \exp(-\phi \|h\|_2^2) dh \\ &= \sigma^2 \left(\frac{1}{2\pi}\right)^d \prod_{j=1}^d \int_{\mathbb{R}} \exp(-i\omega_j h_j - \phi h_j^2) dh_j \\ &= \sigma^2 \left(\frac{1}{2\pi}\right)^d \prod_{j=1}^d \int_{\mathbb{R}} \exp(-\phi(h_j - (-i\omega/(2\phi)))^2) \exp(-\omega_j^2/(4\phi)) dh_j \\ &= \sigma^2 \left(\frac{1}{4\pi\phi}\right)^{d/2} \exp(-\|\omega\|_2^2/(4\phi)) \end{aligned}$$

i.e. it has a Gaussian form.

**Definition 41.** Let  $(Z(s) : s \in \mathcal{S})$  be a weakly stationary random field with covariogram function  $c : \mathcal{H} \rightarrow \mathbb{R}$  and  $c(h) = \text{Cov}(Z(s), Z(s+h))$ . The correlogram function  $\rho : \mathcal{H} \rightarrow [-1, 1]$  is defined as

$$\rho(h) = \frac{c(h)}{c(0)}.$$

## 7. INTRINSIC STATIONARITY (OF ORDER ZERO)

*Note 42.* The class of (weakly) stationary random fields may not be sufficiently general enough to model a large number of important applications. E.g., in certain applications, it has been noticed that the “underline process” we wish to model presents increments whose variance

$$\text{Var}(Z(s+h) - Z(s)) = \text{Var}(Z(s+h)) + \text{Var}(Z(s)) - 2\text{Cov}(Z(s+h), Z(s))$$

increases indefinitely with  $|h|$ ; this “process” cannot be modeled within the class of (weakly) stationary random fields whose increments have bounded variance  $\text{Var}(Z(s+h) - Z(s)) = 2(c(0) - c(h)) < 2c(0)$ . Intrinsic stationary is a weaker assumption extending the class of models we can use.

**Definition 43.** A random field  $(Z(s) : s \in \mathcal{S})$  is called intrinsic random field (i.r.f.) (or intrinsic stationary r.f.) if, for all  $h \in \mathcal{H}$ ,

- (1)  $E(Z(s+h) - Z(s))^2 < \infty$
- (2)  $E(Z(s+h) - Z(s)) = \mu(h)$  for some function  $\mu : \mathcal{H} \rightarrow \mathbb{R}$  (lag dependent)
- (3)  $\text{Var}(Z(s+h) - Z(s)) = 2\gamma(h)$  for some function  $\gamma : \mathcal{H} \rightarrow \mathbb{R}$  (lag dependent)

**Example 44.** The random field with covariance function

$$c(s, t) = \frac{1}{2} \left( \|s\|^{2H} + \|t\|^{2H} - \|t-s\|^{2H} \right), \quad H \in (0, 1)$$

is not stationary r.f. because

$$c(s, s+h) = \frac{1}{2} \left( \|s\|^{2H} + \|s+h\|^{2H} - \|h\|^{2H} \right)$$

for  $h \in \mathcal{H}$  but it intrinsic r.f. because

$$\frac{1}{2} \text{Var}(Z(s+h) - Z(s)) = \frac{1}{2} (\text{Var}(Z(s)) + \text{Var}(Z(s+h)) - 2\text{Cov}(Z(s), Z(s+h))) = \frac{1}{2} \|h\|^{2H}$$

**Example 45.** For an i.r.f.  $(Z(s) : s \in \mathcal{S})$  with  $\mu(h) = 0$ , it can be shown that

$$(7.1) \quad \text{Cov}(Z(t) - Z(s), Z(v) - Z(u)) = \gamma(t-u) + \gamma(s-v) - \gamma(s-u) - \gamma(t-u)$$

by taking expectations from

$$2(a-b)(c-e) = (a-e)^2 + (a-b)^2 - (b-c)^2 - (a-c)^2$$

*Note 46.* The price to be paid for i.r.f. offering a larger class of models by setting the assumptions on the increments only, is involve an indeterminacy regarding the actual r.f.  $Z(s)$ ; E.g. i.r.f.  $(Z(s) : s \in \mathcal{S})$  and  $(Z(s) + U : s \in \mathcal{S})$  where  $U$  a single variable leave (2) and (3) in Def 43 unchanged. When this causes problems, usual trick are: (a) “registration” (Example 47), i.e. consider an additional non-used specific site  $s_0 \in \mathcal{S}$  (called origin) at which a value is known  $Z(s_0) = z_0$  and try to work out (b) impose restrictions int eh increments.

**Example 47.** To specify the moments of an i.r.f.  $Z(s)$ . Consider an origin  $s_0 \in \mathcal{S}$  with known  $Z(s_0) = z_0$ , and specify the “registered” r.f.  $\tilde{Z}(s) = z_0 + (Z(s) - Z(s_0))$ . Then  $E(\tilde{Z}(s)) = z_0 + \mu(s - s_0)$  and  $\text{Cov}(\tilde{Z}(s), \tilde{Z}(t))$  computed from (7.1).

**Example 48.** Only the covariance of allowed linear combinations can be represented w.r.t.  $\gamma(\cdot)$ . I.e.

$$\text{Cov} \left( \sum_{i=1}^n a_i Z(s_i), \sum_{j=1}^m b_j Z(s_j) \right) = \sum_{i=1}^n a_i \sum_{j=1}^m b_j \text{Cov}(Z(s_i), Z(s_j))$$

assuming  $Z$  is i.r.f. hence covariogram may not be defined, we consider origin  $s_0 \in \mathcal{S}$  with  $Z(s_0)$  and we restrict to the sum-to-zero linear combinations. Hence, by (7.1)

$$\begin{aligned}\text{Cov} \left( \sum_{i=1}^n a_i Z(s_i), \sum_{j=1}^n b_j Z(s_j) \right) &= \sum_{i=1}^n a_i \sum_{j=1}^n b_j \text{Cov}(Z(s_i) - Z(s_0), Z(s_j) - Z(s_0)) \\ &= \sum_{i=1}^n a_i \sum_{j=1}^n b_j (\gamma(s_i - s_0) + \gamma(s_j - s_0) - \gamma(s_j - s_i)) \\ &= - \sum_{i=1}^n \sum_{j=1}^n a_i b_j \gamma(s_j - s_i)\end{aligned}$$

## 8. INCREMENTAL MEAN FUNCTION

**Definition 49.** Incremental mean function (or drift) of the intrinsic random field  $(Z(s) : s \in \mathcal{S})$  is defined as  $\mu : \mathcal{H} \rightarrow \mathbb{R}$  with  $\mu(h) = E(Z(s+h) - Z(s))$ .

**Example 50.** Let  $(Z(s) : s \in \mathcal{S})$  be an intrinsic random field, the incremental drift is linear

$$(8.1) \quad \mu(h) = h^\top \beta$$

for some  $\beta \in \mathbb{R}^d$ . Indeed, it is

$$\begin{aligned}\mu(h+h') &= E(Z(s+h+h') - Z(s)) = E(Z(s+h) - Z(s)) + E(Z(s+h+h') - Z(s+h)) \\ &= \mu(h) + \mu(h'), \quad \forall h, h'.\end{aligned}$$

Since,  $\mu(\cdot)$  is continuous and  $\mu(0) = 0$ , than  $\mu(h)$  is linear wrt  $h$ .

## 9. SEMIVARIOGRAM

*Note 51.* A very informative tool about the behavior of the intrinsic random field is the semivariogram function defined below.

**Definition 52.** The semivariogram of an intrinsic random field  $(Z(s) : s \in \mathcal{S})$  is defined as  $\gamma : \mathcal{H} \rightarrow \mathbb{R}$ , with

$$\gamma(h) = \frac{1}{2} \text{Var}(Z(s+h) - Z(s))$$

**Definition 53.** Variogram of an intrinsic random field  $(Z(s) : s \in \mathcal{S})$  is called the quantity  $2\gamma(h)$ .

*Note 54.* A stationary random field with covariogram  $c(\cdot)$  and mean  $\mu$  is intrinsic stationary as well with semivariogram

$$(9.1) \quad \gamma(h) = c(0) - c(h),$$

and constant incremental mean  $\mu(h) = \mu$ .

**Example 55.** For the Gaussian covariance function (Ex. 34) the semivariogram is

$$\gamma(h) = c(0) - c(h) = \sigma^2 (1 - \exp(-\beta \|h\|_2^2))$$

**Proposition 56.** *Properties of semivariogram.* Let  $(Z(s) : s \in \mathcal{S})$  be an intrinsic random field, then

- (1) It is  $\gamma(h) = \gamma(-h)$ ,  $\gamma(h) \geq 0$ , and  $\gamma(0) = 0$
- (2) Semivariogram is conditionally negative definite (c.n.d.): if for all  $n \in \mathbb{N}$ ,  $(a_1, \dots, a_n) \subseteq \mathbb{R}^n$  s.t.  $\sum_{i=1}^n a_i = 0$ , and for all  $(s_1, \dots, s_n) \subseteq S^n$ , it is

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j \gamma(s_i - s_j) \leq 0$$

## 10. BEHAVIOR OF SEMIVARIOGRAM OF INTRINSIC RANDOM FIELDS

*Note 57.* The semivariogram  $\gamma(h)$  is very informative when plotted against the lag  $h$ . Below we discuss some of the characteristics of it, using Figure 10.1

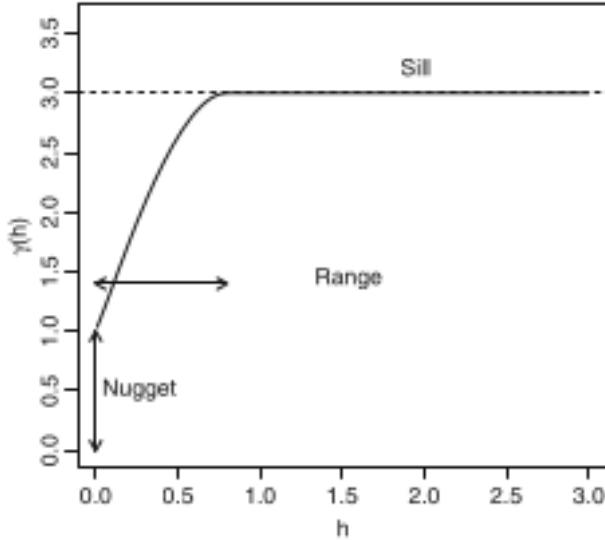


FIGURE 10.1. Semi Variogram's characteristics

*Note 58.* A semivariogram tends to be an increasing function of the lag  $\|h\|$ . Recall that for weakly stationary random fields with c.f.  $c(\cdot)$ , it is  $\gamma(h) = c(0) - c(h)$  where common logic suggests that  $c(h)$  is decreases with  $\|h\|$ .

*Note 59.* If  $\gamma(h)$  is a positive constant for all non-zero lags  $h \neq 0$ , then  $Z(s_1)$  and  $Z(s_2)$  are uncorrelated regardless of how close  $s_1$  and  $s_2$  are. Then  $Z(\cdot)$  is called white noise.

*Note 60.* Conversely, a non zero slope of the variogram indicates some structure.

Nugget Effect.

*Note 61.* Nugget effect is the semivariogram limiting value

$$\sigma_{\varepsilon}^2 = \lim_{\|h\| \rightarrow 0} \gamma(h)$$

when  $\sigma_{\varepsilon}^2 \neq 0$ .

*Note 62.* When used for modeling, nugget effect  $\sigma_{\varepsilon}^2 \neq 0$  may express (1) measurement errors (e.g., if we collect repeated measurements at the same location  $s$ ) or (2) some microscale variation causing discontinuity in the origin that cannot be detected from the data i.e. the spatial gaps because we collect a finite set of measurements at spatial locations. Ideally, a more detailed decomposition  $\sigma_{\varepsilon}^2 = \sigma_{MS}^2 + \sigma_{ME}^2$  can be considered where  $\sigma_{MS}^2$  refers to the microscale and  $\sigma_{ME}^2$  refers to the measurement error. However this may lead to non-identifiability, without any obvious tweak to address it.

Sill.

**Definition 63.** Sill is the semivariogram limiting value  $\lim_{\|h\| \rightarrow \infty} \gamma(h)$ .

*Note 64.* For intrinsic processes, the sill may be infinite or finite. For weakly random field, the sill is always finite.

Partial sill .

**Definition 65.** Partial sill is  $\lim_{\|h\| \rightarrow \infty} \gamma(h) - \lim_{\|h\| \rightarrow 0} \gamma(h)$  which takes into account the nugget.

Range .

*Note 66.* Range is the distance at which the semivariogram reaches the Sill. It can be infinite or finite.

Other.

*Note 67.* An abrupt change in slope indicates the passage to a different structuration of the values in space. This is often modeled via decomposition of processes with different semivariograms. E.g., let independent random fields  $Y(\cdot)$  and  $X(\cdot)$  with different semivariograms  $\gamma_Y$  and  $\gamma_X$ , then random field  $Z(\cdot)$  with  $Z(s) = Y(s) + X(s)$  has semivariogram  $\gamma_Z(h) = \gamma_Y(h) + \gamma_X(h)$  which may present such a behavior.