Spatio-temporal statistics (MATH4341) Michaelmas term, 2023

Handout 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
- Datasets:
  `https://github.com/georgios-stats/Spatio-Temporal_Statistics_Michaelmas_2023/tree/main/Datasets/`

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 to the development and analysis of (not necessarily statistical) models that indicate when 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 a 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. 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".

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.

Spatial data and spatial process.

Note 7. In spatial statistics, the basic components are data $\{Z_{s_1}, ..., Z_{s_n}\}$ observed at locations spatial locations $\{s_1, ..., s_n\}$. 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. The locations $s_i \in S$ can be considered either (i.) fixed and hence used for training or (ii.) random and hence a quantity for inference. Yet, $\{s_i\}$ can be arranged irregularly in the space or regularly in a grid. Data $Z_{s_i} = Z(s_i)$ are random vectors.

Note 8. 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 process (or multivariate random field)

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

which can be modeled as a stochastic process (to be defined later.).

Note 9. 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 spatial process (or a multivariate random field)

$$\{Z(s) ; s \in S\}, \tag{1.1}$$

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

2. Principal spatial statistics areas

Note 10. 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/problem may involve

characteristics from a combination of them. Here, we will study three of spatial statistical areas.

2.1. **Point referenced data (Geostatistics).**

Note 11. 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 realisations from a random field, 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 12. 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 13. 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. The actual data are observations $\{Z(s_i)\}_{i=1}^n$ at n fixed locations $\{s_i\}_{i=1}^n \subset S$. The locations $\{s_i\}$ are fixed and can be arranged irregularly in the space or regularly as a grid.

Note 14. Geostatistics tries to answer questions about modeling, identification and separation of small and large scale variations, prediction at unobserved locations and reconstruction of the spatial process $Z(s)$ across the whole space S.

2.1.1. *Examples.*

Example 15. (Ground water pollution in the Central Valley of California[1]) California's Central Valley is one of the most productive agricultural regions in the world. With an increase in population, groundwater consumption is expected to increase. Agricultural irrigation heavily draws on the groundwater system. Pumping from increasingly deeper parts of the aquifer has increased the rate of downward groundwater flow, which have been linked to the release of, for example, uranium. The question therefore concerns how we can maintain groundwater quality while dealing with this increased need for it. Understanding this trade-off is key to sustainable groundwater management. Simply increasing groundwater by supply, for example through a process of recharge (e.g., flooding a field), may affect its

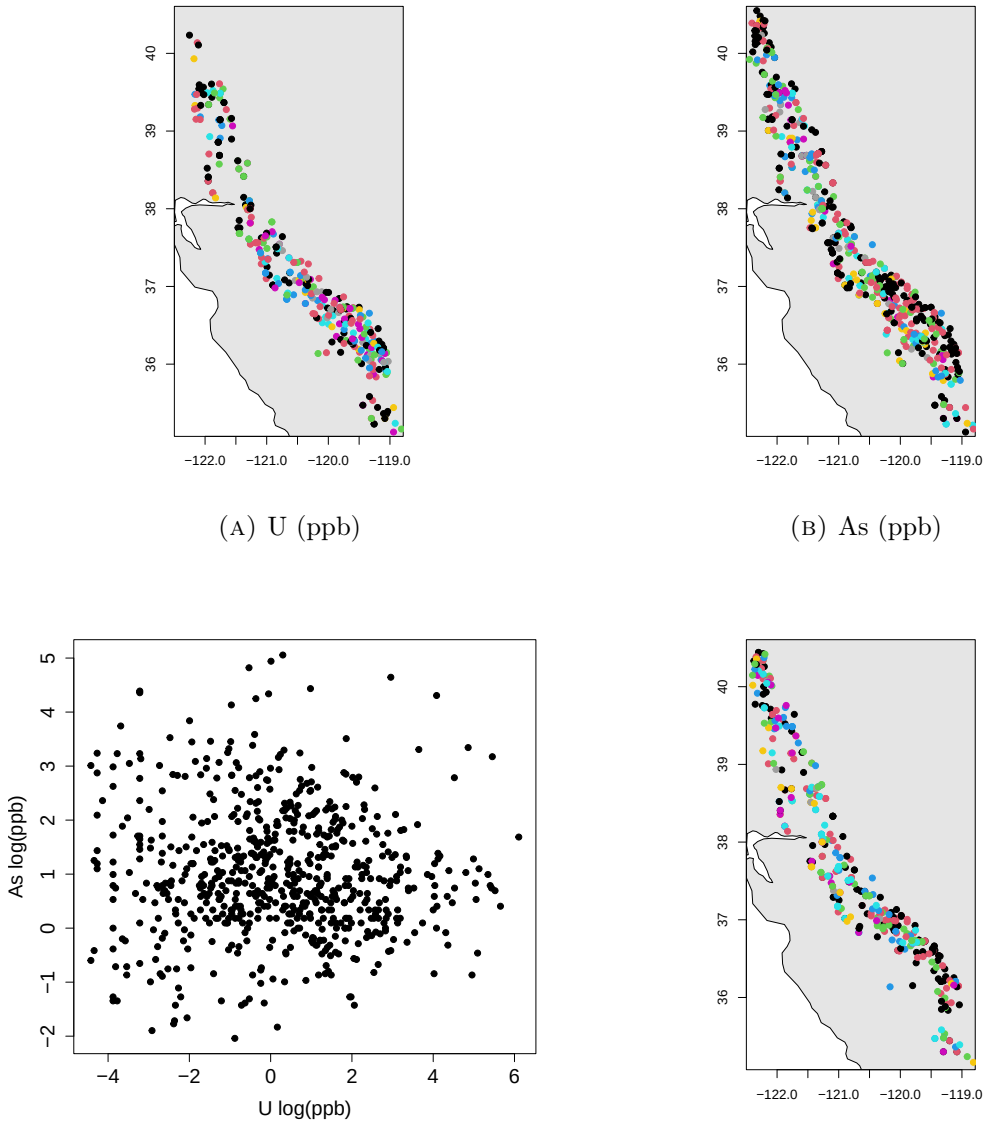
[1]Fakhreddine, S., Babbitt, C., Sherris, A., et al. (2019). Protecting Groundwater Quality in California, Management Considerations for Avoiding Naturally Occurring and Emerging Contaminants. Environmental Defense Fund. `www.edf.org/sites/default/files/documents/groundwater-contaminants-report.pdf`

quality. It may lead to an increased introduction of contaminants such as pesticides. In other words, groundwater management actions may have unintended consequences. It is important to understand how groundwater quality is affected by any management actions on, for example, seawater intrusion, land subsidence, or declining water levels. A substantial amount of geochemical analysis has been collected from existing wells. Interest lies in understanding the important processes, either natural or anthropogenic, that cause variation in these data. We may find a certain "signature" or "patterns" in the data that can determine the process of contamination in a particular area. For example, the Central Valley has what are termed "geogenic" contaminants, which means it has arsenic (As), chromium (Cr), uranium (U), which you don't want to drink, naturally occurring. A simple analysis could look for high levels of these elements, indicating possible anthropogenic contamination, although they may be naturally occurring. The point we wish to make is that a signature of a feature is more than a single high value. What we are looking for is a combination of elements and of a certain composition. Some scientific questions involve:

(1) What combination of elements are indicative of a human impact in water quality versus a natural occurrence?
(2) What caused this impact? Agriculture? Pollution?
(3) Where in the Central Valley can we find these combinations of elements, thereby informing mitigation action?

Figure 2.1c presents the scatter plot of As and U in a naive manner as it ignores spatial dependency. Figures 2.1d, 2.1b, and 2.1a show the Groundwater concentration (parts per billion [ppb]) of chromium (Cr), Arsenic (As), Uranium (U) from January 2018 to January 2019. The point coordinates are the geographical locations, and the color denotes the value of the corresponding values of Cr, As, U in ppb. The locations are fixed/known and hence part of the training observations. The locations are irregularly scattered/spaced and hence not on a regular grid of points. The spatial statistician's task may involve producing statistical models able to provide predictive inference for quantities Cr, As, U (and others) at unseen/unobserved locations. Obviously, special dependencies should be taken into account in the model, e.g. the concentration of U in two neighboring cities is expected to be more similar than two far distant cities. As seen latter this gives rise to a 'regionalized statistical analysis'. Along with the spatial dependency, and in the same model, it would be wise to take into account the dependency (e.g. correlation) between different variables, such as As and U. As seen latter this gives rise to a 'co-regionalized statistical analysis'.

Example 16. (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



(A) U (ppb)

(B) As (ppb)

(C) U vs As (log ppb)

(D) Cr (ppb)

FIGURE 2.1. Ground water pollution in the Central Valley of California

the colorbar. The sampled locations are fixed, and regularly spaced in a grid. 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. 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.

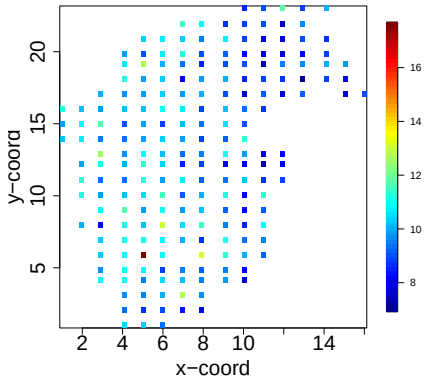


FIGURE 2.2. (Coal ash data set) Percentage of coal ash at 208 locations.

Example 17. (Air pollution in Piemonte.) Figure 2.3 presents the Average PM10 ($\mu g/m^3$) concentration during October 2005–March 2006 for the 24 monitoring stations in the Piemonte region (Northern Italy). The data (measurements) are at fixed locations at irregular grid points. PM10 is one of the most troublesome pollutants in the area. Environmental agencies need models to predict PM10 at unmonitored sites in order to assess PM10 concentration over an entire region. A geostatistician can build a model which is satisfactory in terms of goodness of fit, interpretability, parsimony, prediction capability and computational costs with purpose to build reliable PM10 concentration maps, equipped with the corresponding uncertainty measure.

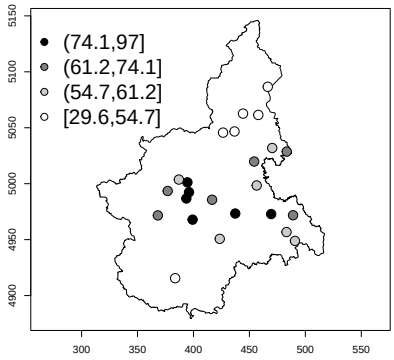


FIGURE 2.3. (Air pollution data) Average PM10 ($\mu g/m^3$) concentration during October 2005–March 2006 for the 24 monitoring stations in the Piemonte

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

Note 18. 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 19. 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 of arranged in a regular grid. Often, there is a natural adjacency relation or neighborhood 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 $s \in S$.

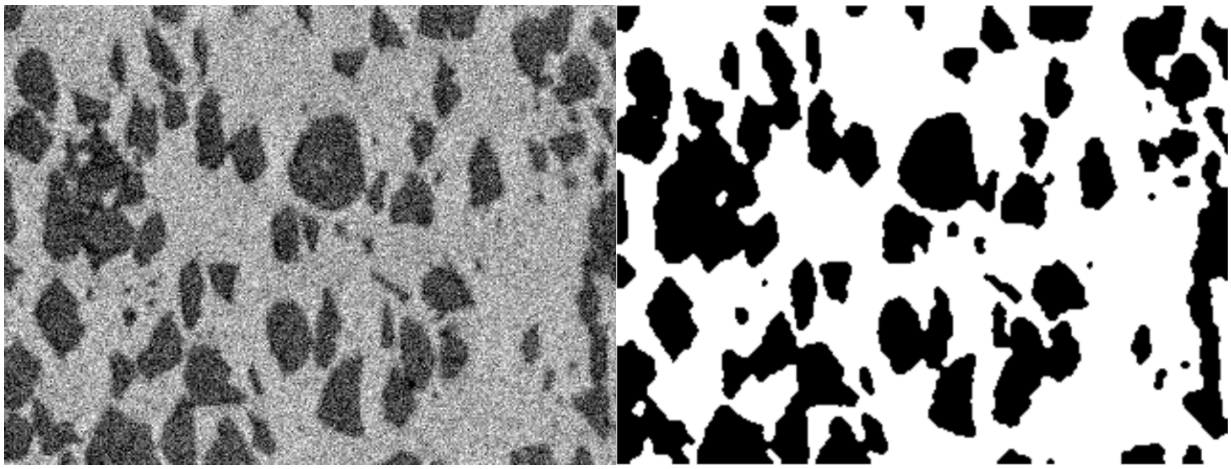
Example 20. In image processing, S may be a grid of pixels (locations are fixed and regular).

Example 21. In a UK epidemological study, S may be the centroids of the UK counties, and $Z(s)$ may represent the average value of a characteristic in county s.

Example 22. In statistical physics, S may be a collection of atoms and genuinely finite (locations are fixed and regular).

Example 23. (Image restoration data) Figure 2.4a 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 240×320 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 2.4b shows the restored image after appropriate statistical processing. Here the locations are pixels arranged in a fixed regular grid (hence discrete and not continues). The each observation $Z(s)$ is the color of a pixel s.

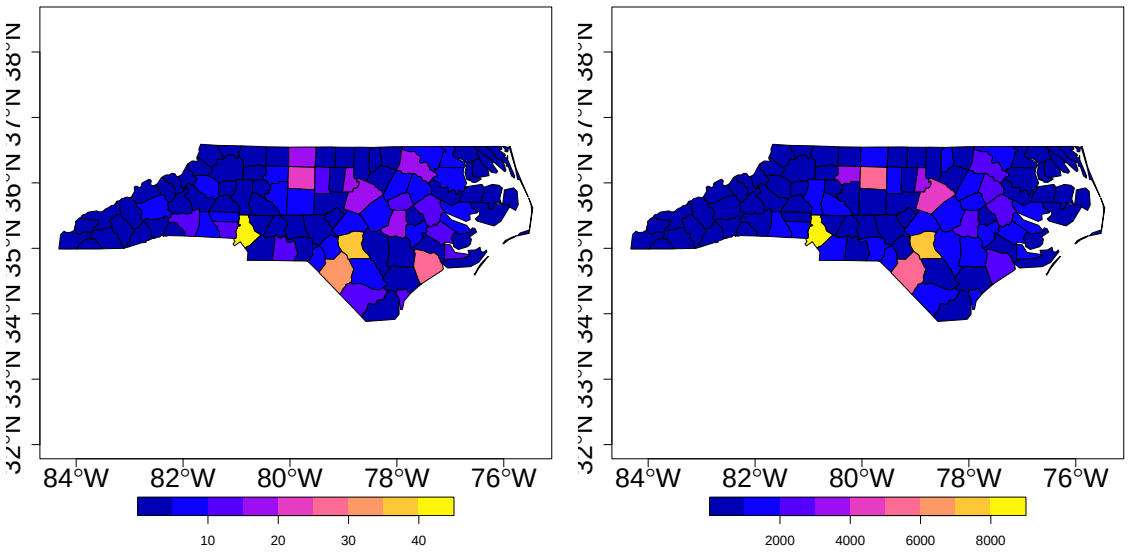
Example 24. (North Carolina SIDS data set) Figure 2.5a shows the total number of deaths from Sudden Infant Death Syndrome (SIDS) in 1974 for each of the 100 counties in North Carolina. Figure 2.5b shows the corresponding live births in each county and same period. This is the R data set `nc{spdep}`. 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 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



(A) Noised image (B) Reconstructed image

FIGURE 2.4. Ferrite-Pearlite steel image (Image restoration)

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.



(A) Numbers of SIDS (B) Live birth counts

FIGURE 2.5. North Carolina SIDS data (Aerial unit data)

2.3. **Spatial point pattern data.**

Note 25. Sometimes the locations at which events occur are random. Typical examples include outbreaks of forest fires, or the epicentres of earthquakes. Such random patterns of locations are said to form a point process.

Note 26. Rigorously, the spatial domain S is a point process, a set of points, in $\mathbb{R}^d$ at which some events happened.

Note 27. 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,. This generates a Marked spatial point process.

Note 28. In the simplest case, no covariate for Z is specified, and hence 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 $s \in S$. This generates a spatial point process

Note 29. Questions in the spatial point pattern problems are mainly whether the pattern of locations is exhibiting complete spatial randomness, clustering, or regularity. In the marked spatial point process 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 process may appear to be wrongly clustered when just eyeballed by an individual.

Example 30. (Tropical rain forest trees in Barro/Colorado) Figure 2.6 shows the positions (dots) of 3605 Beilschmiedia trees in a 1000×5000 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 process as each location is associate to an occurrence only and not any covariate. The statistician's task is to design models able to test and quantify heterogeneity/homogeneity.

Example 31. (Longleaf Pines Point Pattern) Figure 2.7 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

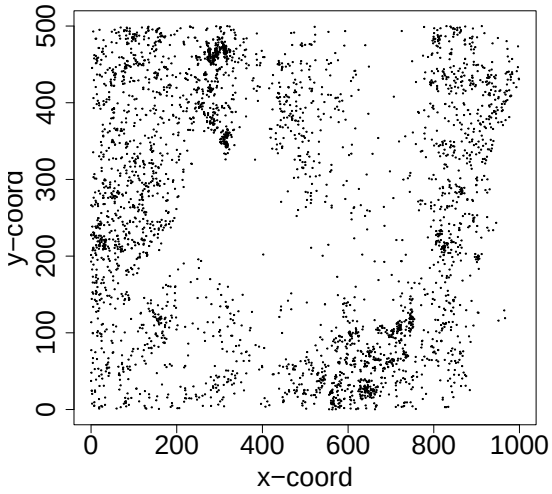


FIGURE 2.6. Locations of tropical rain forest trees in Barro/Colorado (Spatial point pattern data)

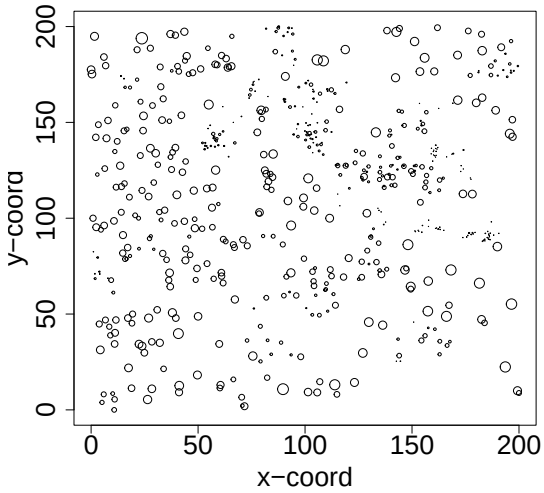


FIGURE 2.7. Longleaf Pines Point Pattern (Spatial point data)

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 process is a marked process.

Note 32. In spatial problems, uncertainty is expressed probabilistically through a spatial stochastic process (or a multivariate random field), which can be written most generally as

$$\{Y(s); s \in \mathscr{S}\}, \tag{3.1}$$

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

3.1. **Spatial process model.** 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. 2.4a.

Note 33. This spatial stochastic process can be a: geostatistical process, lattice process, or point process depending on the principal spatial statistical area (Section 2) the application is associated with.

Note 34. The joint probability model defined by the random $\{Y(s); s \in S\}$ is

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

Note 35. 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 were 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 36. The measurement uncertainty is quantified via the data model. E.g. the "snow" in Fig. 2.4a.

Note 37. The data model is specified to be the conditional distribution of the data Z given the spatial stochastic process Y and the S, namely

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

Note 38. If the data are assumed to be conditionally independent, such as $Z(s) \perp Z(s')|Y, S$ then

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

Note 39. The spatial statistical dependence of in Z, articulated by the First Law of Geography, follows by

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

The hierarchical statistical model.

Note 40. To sum up the (known) uncertainty in spatial statistics about a scientific problem is expressed via the so called Hierarchical spatial model

$$\begin{cases} Z|Y,S & \text{data model} \\ Y,S & \text{spatial process model} \end{cases} \tag{3.4}$$

with

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

The Empirical (Bayes) hierarchical model.

Note 41. Often the decomposition (3.4) is parametrized with respect to unknown parameters $\theta\in\Theta$ we wish to learn given the observables, i.e.

$$\begin{cases} Z|Y,S,\theta & \text{data model} \\ Y,S|\theta & \text{spatial process model} \end{cases} \tag{3.5}$$

with

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

...more details in the next lecture.

The Bayesian hierarchical model.

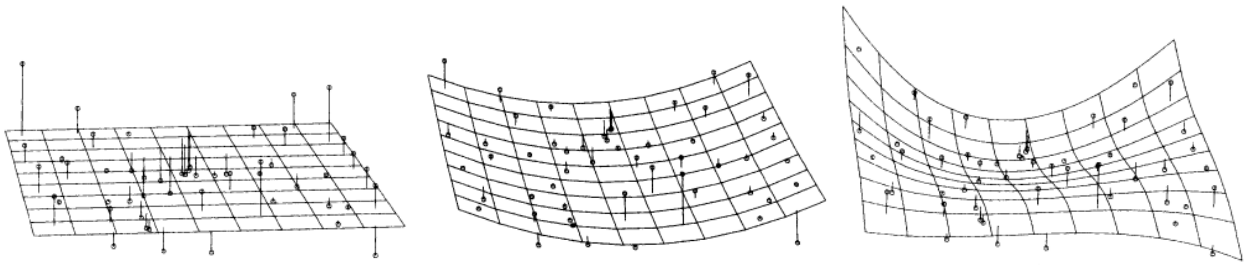
Note 42. In Bayesian statistics, the hierarchical model in (3.4) is completed by the $\theta\sim\text{pr}\left(\cdot\right)$ adding a third layer as

$$\begin{cases} Z|Y,S,\theta & \text{data model} \\ Y,S|\theta & \text{spatial process model} \\ \theta & \text{hyper-parameter prior model} \end{cases} \tag{3.6}$$

with

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

Exercise 43. (A naive example) Consider Example 15, and observations $\{(Z_i,s_i)\}_{i=1}^n$ where Z_i is the Cr measurement in ppb at the i-th location $s_i\in\mathbb{R}^2$. Perhaps one may consider that the real Cr, lets denoted as Y, may follow a Normal distribution with a mean $\mu=S\beta$ parametrized as $[\mu]_i=s_{(0),i}\beta_0+s_{(1),i}\beta_1+s_{(2),i}\beta_2+s_{(1),i}s_{(2),i}\beta_{12}+...$ at a location s (to consider spatial dependence) with some unknown parameter β, and covariance matrix parametrized as $[C]_{i,j}=c\left(s_i,s_j\right)$ with $c\left(s,s'\right)=\sigma^2\exp\left(-\phi\left\|s-s'\right\|\right)$; here β, ϕ, and σ^2 are unknown parameters. One may consider that the measurements Z at each location are the result



(A) (B) (C)

FIGURE 3.1. Examples representing the hierarchical spatial model (3.7) for different values of $\theta = (\sigma^2, \beta, \phi)$

of observing Y_i (the real Cr) but contaminated by additive random noise, as $Z_i = Y_i + \epsilon_i$, $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$. To sum up, we have build the hierarchical model

$$\begin{cases} Z|Y, \sigma^2 \sim & \mathrm{N}(Y, I\sigma^2) \\ Y|\sigma^2, \beta, \phi \sim & \mathrm{N}(S\beta, C) \end{cases} \tag{3.7}$$

Figure 3.1 shows the hierarchical spatial model (3.7) 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 (β, ϕ), 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 σ^2 . If we work on the fully Bayesian framework (!!!), we can to complete the model with priors on $\theta = (\sigma^2, \beta, \phi)$ as $\sigma^2 \sim \mathrm{IG}(\kappa_\sigma, \lambda_\sigma)$, $\phi \sim \mathrm{IG}(\kappa_\phi, \lambda_\phi)$, and $\beta \sim \mathrm{N}(b, Iv)$, with some known hyper-parameters $\kappa_\sigma, \lambda_\sigma, \kappa_\phi, \lambda_\phi, b, v$. To sum up, we have build the Bayesian model

$$\begin{cases} Z|Y, \sigma^2 \sim & \mathrm{N}(Y, I\sigma^2) \\ Y|\sigma^2, \beta, \phi \sim & \mathrm{N}(S\beta, C) \\ \beta \sim & \mathrm{N}(b, Iv) \\ \sigma^2 \sim & \mathrm{IG}(\kappa_\sigma, \lambda_\sigma) \\ \phi \sim & \mathrm{IG}(\kappa_\phi, \lambda_\phi) \end{cases}$$