Spatio-temporal statistics (MATH4341) Michaelmas term, 2024

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. [1]Consider a random sample $\{Z_i \in \mathbb{R}; i = 1, ..., n\}$ jointly following a Normal distribution with unknown $\mathrm{E}(Z_i) = \mu$ and known $\mathrm{Var}(Z_i) = \sigma^2$.

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

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

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 μ

The BLUE of μ is

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

[1]No need to worry how the equations are produced.

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

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

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

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

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 $\Phi\left(2.485\right) - \Phi\left(-2.485\right) = 87.5\%$ instead of 95%.

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

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

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

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

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 $n\to\infty$ the asymptotic relative efficiency is

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

Spatial data and spatial process.

Note 9. In spatial statistics, the basic components are data $\{Z_1,...,Z_n\}$ observed at spatial locations $\{s_1,...,s_n\}$ 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 $S\subset\mathbb{R}^d$, $d=1,2,3,...$ 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)

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

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

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.

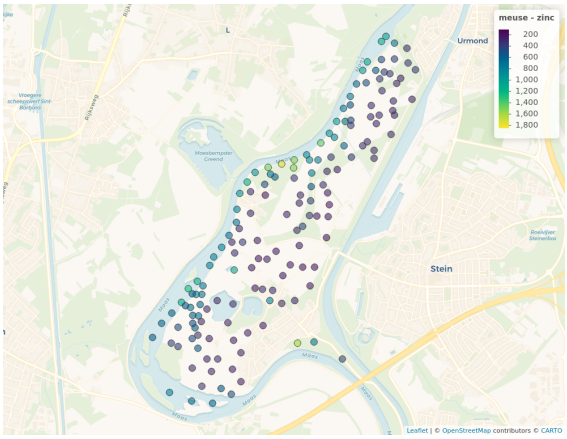
The Meuse river R dataset `meuse{sp}` contains locations and topsoil heavy metal concentrations (such as zinc, lead, copper, and cadmium), along with a number of soil and landscape variables at the observation locations, collected in a flood plain of the river Meuse, near the village of Stein (NL). Figure 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) $\left\{\left(Z_i^{\text{zinc}}, Z_i^{\text{lead}}, Z_i^{\text{cooper}}, Z_i^{\text{carmium}}\right)\right\}$ 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 $Z(s) = \left(Z^{\text{zinc}}(s), Z^{\text{lead}}(s), Z^{\text{cooper}}(s), Z^{\text{carmium}}(s)\right)$, 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 confunding 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.

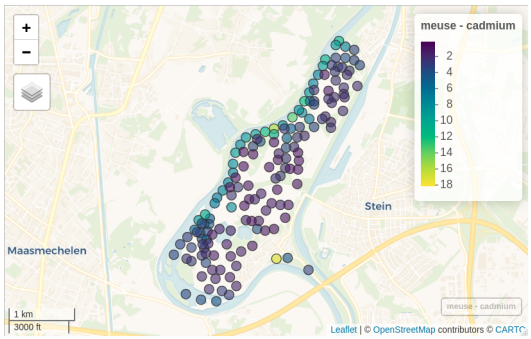
2.2. **Aerial unit data / spatial data on lattices.**

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.

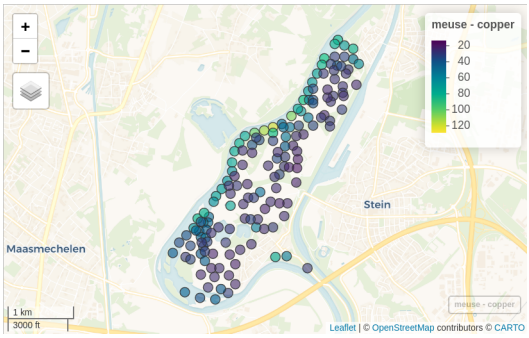
Note 23. Mathematically speaking, the index set S of the data $\{Z(s)\}$ is a fixed (not random) and finite collection of points (locations) $s \in S$. The locations $s \in S$ can be irregular or arranged in a regular grid. Often, there is a natural adjacency relation or neighborhood



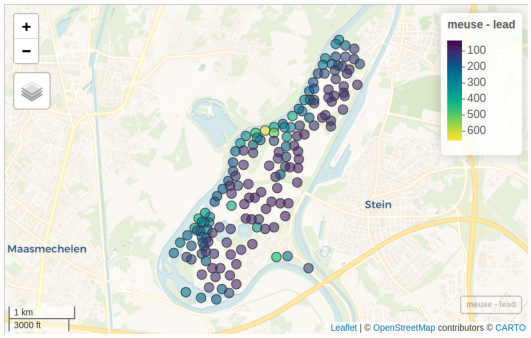
(A) zinc concentration (obs)



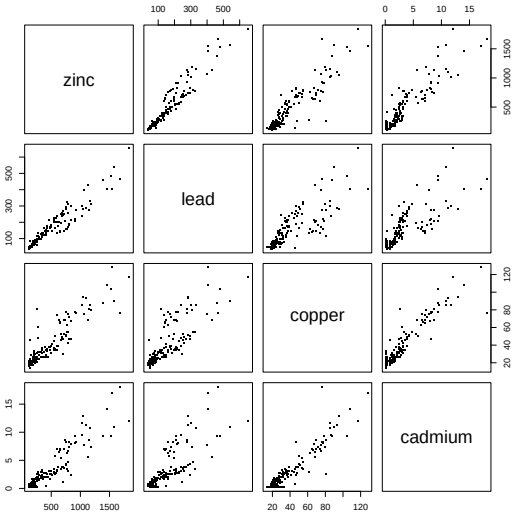
(B) cadmium concentration (obs)



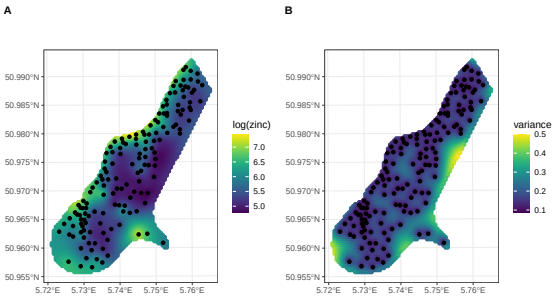
(C) copper concentration (obs)



(D) lead concentration (obs)



(E) scatter plot (obs)



(F) zinc (log scale) pred

FIGURE 2.1. Map of the meuse dataset

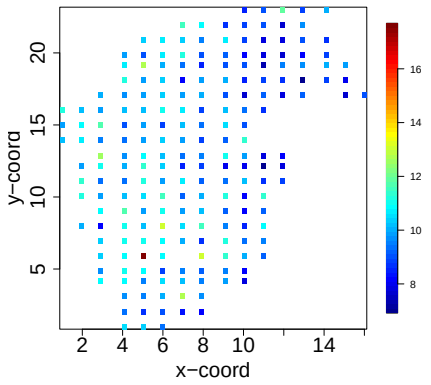


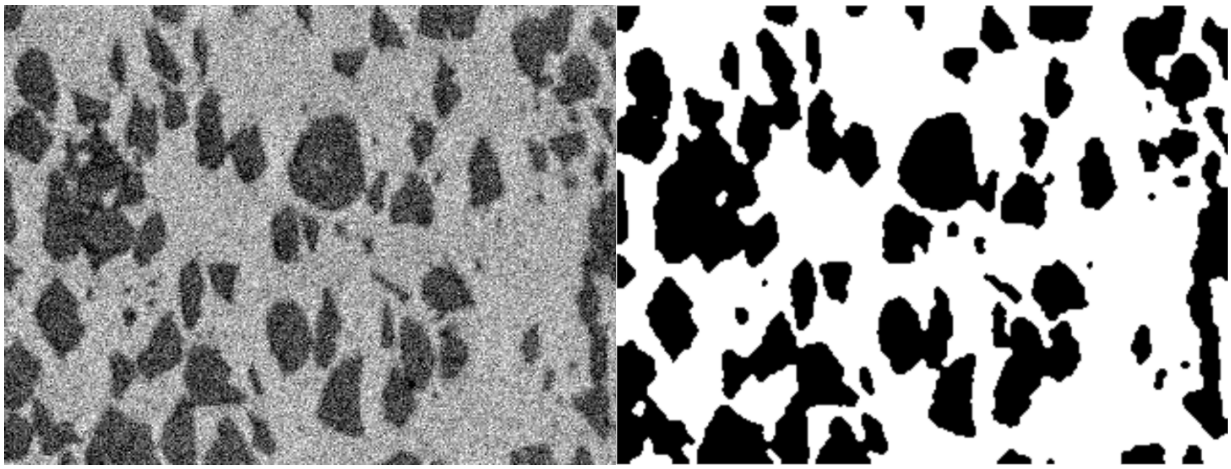
FIGURE 2.2. (Coal ash data set) Percentage of coal ash at 208 locations.

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 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. 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 240×320 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 continues). 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.

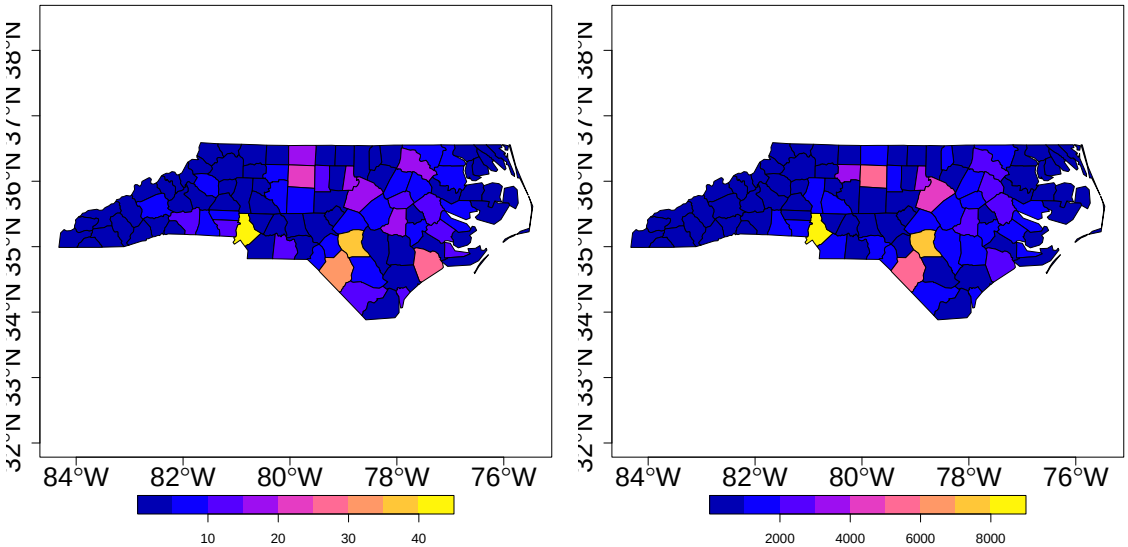
Example 26. (North Carolina SIDS data set) Figure 2.4a shows the total number of deaths from Sudden Infant Death Syndrome (SIDS) in 1974 for each of the 100 counties in North Carolina. Figure 2.4b shows the corresponding live births in each county and same period. This is the R data set `nc{spdep}`. 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



(A) Noised image (B) Reconstructed image

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.

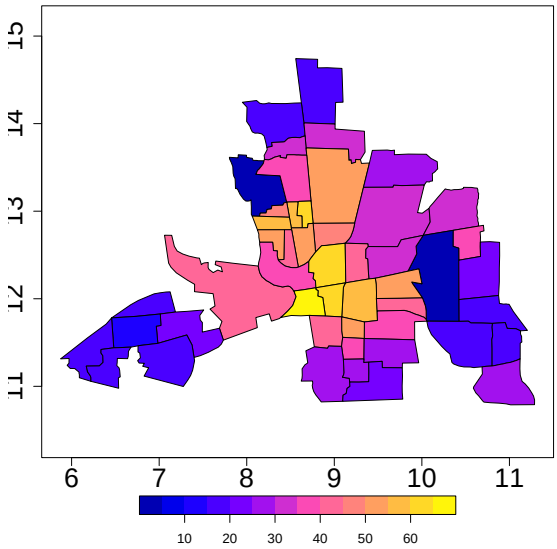


(A) Numbers of SIDS (B) Live birth counts

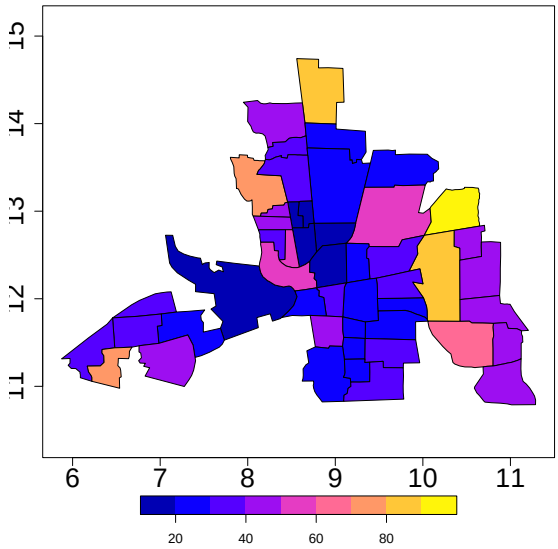
FIGURE 2.4. North Carolina SIDS data (Aerial unit data)

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

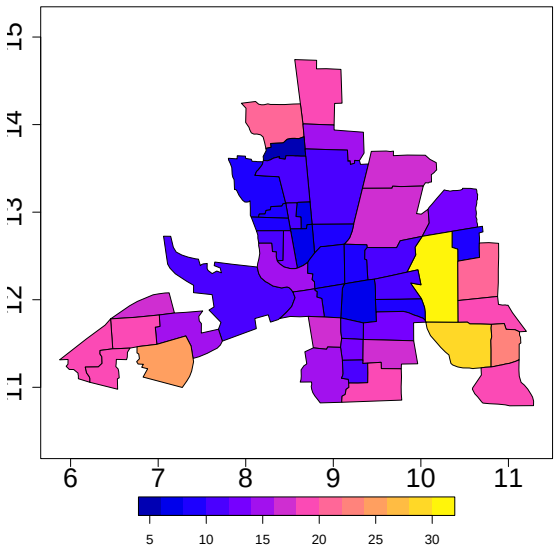
well as the average value of the house (in 1,000 USD), and the average household income (in 1,000 USD). This is the R dataset `columbus{spdep}`. The recorded counts are not tied to a precise household but tallied up county-wise (aerial unit) for privacy reasons. The locations $\{s_i\}$ are these areal 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.



(A) Property crime, i.e. residential burglaries and vehicle thefts per thousand households in the neighborhood



(B) Average housing value (in 1,000 USD)



(C) Average household income (in 1,000 USD)

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 1000×5000 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 5.7×5.7 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.

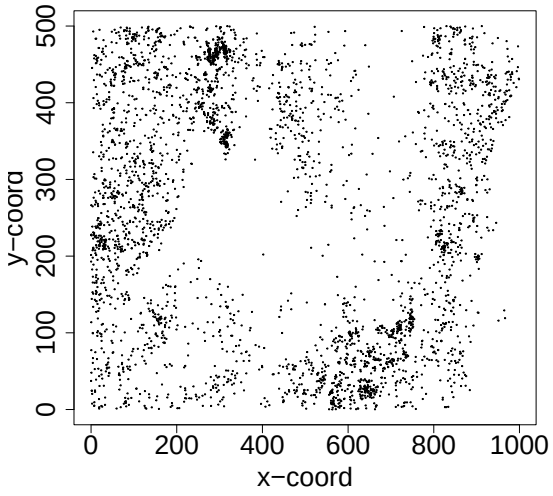
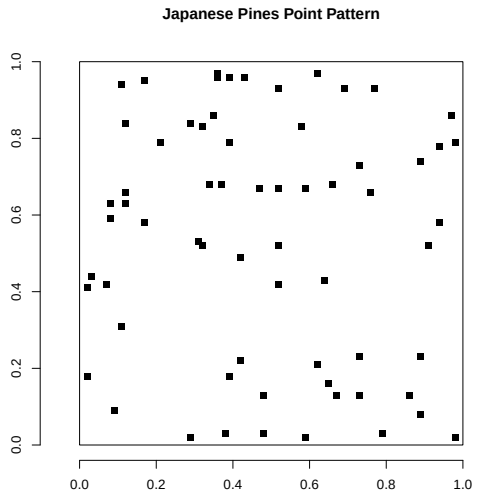


FIGURE 2.6. Locations of tropical rain forest trees in Barro/Colorado (Spatial point pattern data)

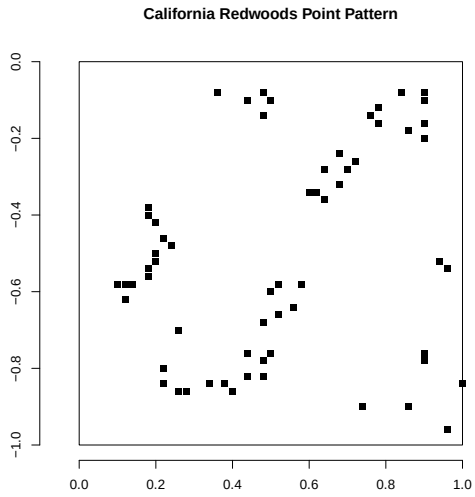
(3) (Biological Cells) The data record the locations of the centres of 42 biological cells observed under optical microscopy in a histological section. The microscope field-of-view has been rescaled to the unit square.

Figure 2.7 presents the aforesaid spatial point patters. 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 a regular pattern.

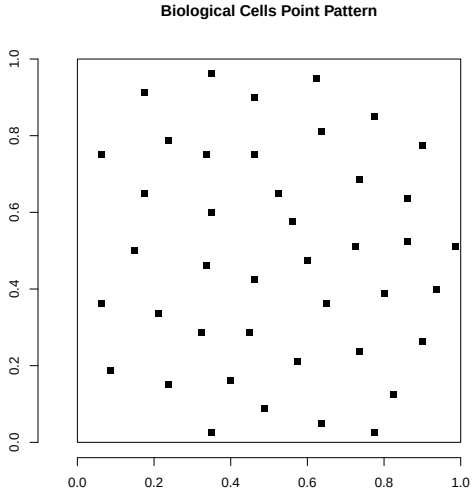
Example 35. (Longleaf Pines Point Pattern) Figure 2.8 shows locations (as Cartesian coordinates) and relative diameters at breast height in dbh (as the size of the dot) of all longleaf pine trees in the 24ha region of the Wade Tract, an old-growth forest in Thomas County, Georgia in 1979. Dataset is 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.



(A) Japanese Pines



(B) California Redwoods



(C) Biological Cells

FIGURE 2.7. Spatial Point Patterns

3. Uncertainty quantification and modeling

To be defined rigorously later

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

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

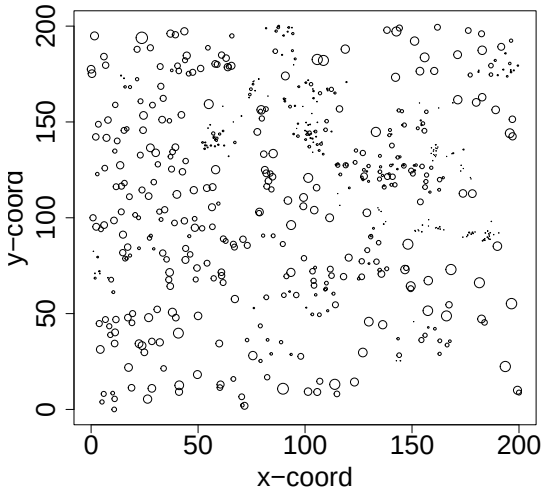


FIGURE 2.8. Longleaf Pines Point Pattern (Spatial point data)

Here $Y(s)$ 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.

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

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

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 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 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

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

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

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

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

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

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

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

with

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

Our interest it to learn $\text{pr}\left(Z^*|Z\right)$ or $\text{pr}\left(Y^*|Z\right)$ 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.

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

with

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

Our additional additional interest may be to learn the unknown θ via MLE

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

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

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

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 θ $\text{pr}(\theta|Z)$.

That's a snapsot for what follows.

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[2]. The dataset is available from the R package in GitHub "andrewzm/STRbook". In this example we focus on maximum temperature (Tmax) in degrees Fahrenheit (∘F) at 138 weather stations in the central USA (between 32∘N-46∘N and 80∘W-100∘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 (∘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.

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

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 μ with $[\mu]_i = \mu(s_i)$ parameterized

[2] http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCDC/.DAILY/.FSOD/

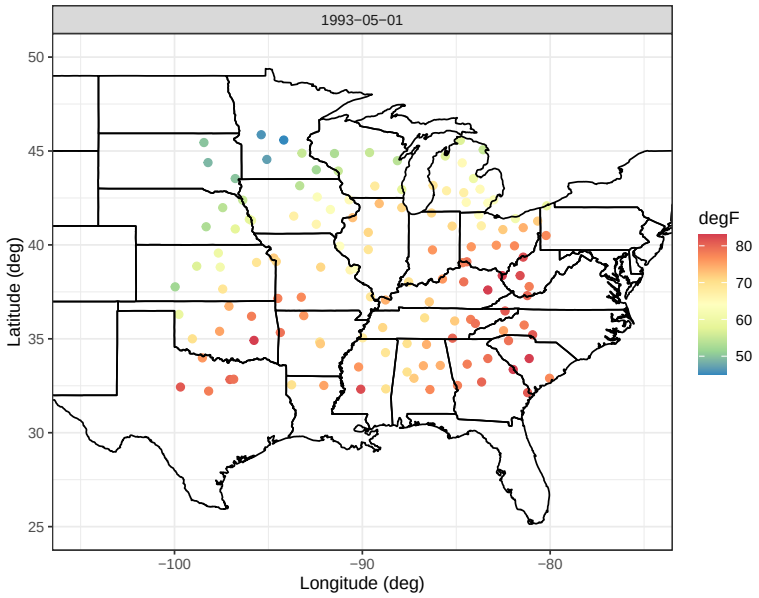


FIGURE 3.1. NOAA weather data (1 May 1993)

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 β, 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 β, ϕ, and σ^2 are unknown parameters.

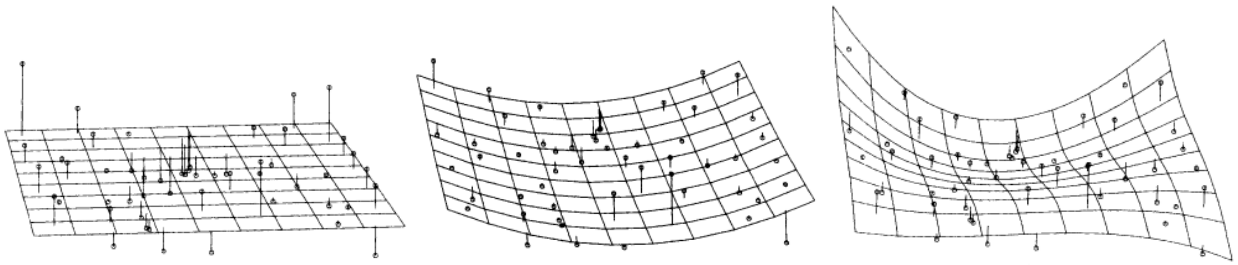
Hierarchical model: To sum up, we have build the hierarchical model

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

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 (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 . Note that marginally

$$Z|\sigma^2, \beta, \phi \sim \mathrm{N}_n\left(S\beta, \sigma^2(I + C)\right).$$

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 $\sigma^2 \sim \mathrm{IG}(\kappa_\sigma, \lambda_\sigma)$, $\phi \sim$



(A) (B) (C)

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

$\mathrm{IG}\left(\kappa_\phi, \lambda_\phi\right)$, and $\beta \sim \mathrm{N}_4\left(b, Iv\right)$, with some known hyper-parameters $\kappa_\sigma, \lambda_\sigma, \kappa_\phi, \lambda_\phi, b, v$. To sum up, we have build the Bayesian model

$$\begin{cases} Z|Y, \sigma^2 \sim & \mathrm{N}_n\left(Y, I\sigma^2\right), \text{ data model} \\ Y|\sigma^2, \beta, \phi \sim & \mathrm{N}_n\left(S\beta, C\right), \text{ spatial process model} \\ \beta \sim & \mathrm{N}_4\left(b, Iv\right), \text{ hyper-parameter prior model} \\ \sigma^2 \sim & \mathrm{IG}\left(\kappa_\sigma, \lambda_\sigma\right), \text{ hyper-parameter prior model} \\ \phi \sim & \mathrm{IG}\left(\kappa_\phi, \lambda_\phi\right), \text{ hyper-parameter prior model} \end{cases}$$

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 (∘F) at We extend the dataset in the Spatial statistics Example 48 by considering time. weather stations in the central USA (between 32∘N–46∘N and 80∘W–100∘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.

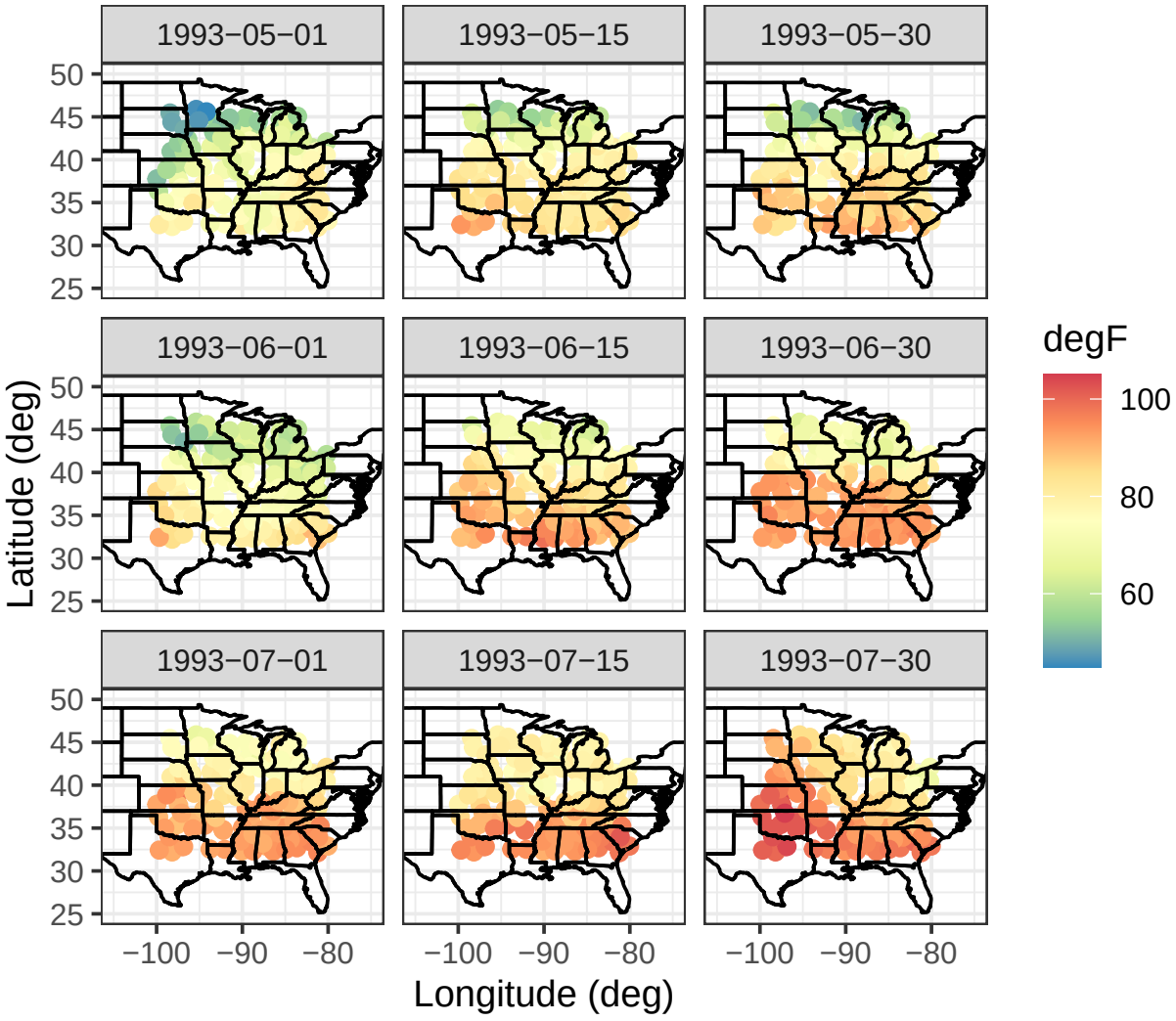


FIGURE 4.1. NOAA daily weather data