University of Glasgow

Late Night Talking

https://www.menti.com/alz6g6hw4mr7

Meet Me in the Hallway

- In the last two weeks, we looked at time series - data which vary over *time*.
- We noted that observations closer together in time were likely to be correlated, and we discussed methods for accounting for that correlation.
- Over the next two weeks, we will look at data which vary over *space*.
- This takes us into the field of spatial statistics, which focuses on developing methods to account for spatial correlation.

Introduction to Spatial Statistics

Late Night Talking

https://www.menti.com/alz6g6hw4mr7

Satellite

- Spatial data are data which have any form of geographical information attached to them.
- Spatial data are becoming increasingly frequent as a result of new technology (eg satellite weather data, GPS enabled devices including phones, CCTV systems).
- Most environmental variables of interest will vary over space to some degree, for example global temperature, air pollution, species distribution.
- We will typically use this spatial information to help us understand the relationship between our data points

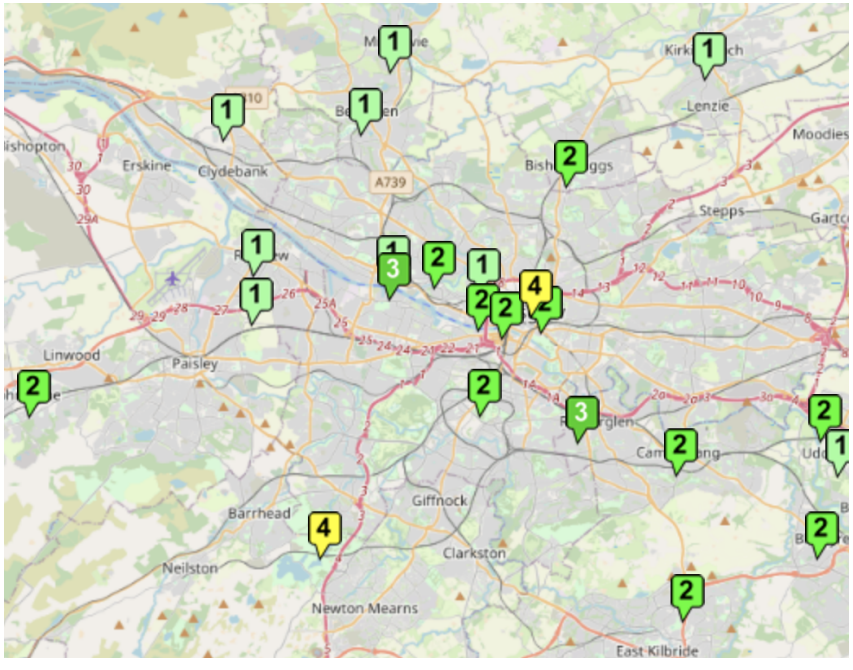
Spatial data

- Generally, each datapoint will contain a measured value in addition to information about their location (typically represented by some form of coordinate system).
- However, in some cases the locations themselves will be our data (eg the positions of trees in a forest).
- We typically assume that we have some degree of spatial autocorrelation present in our data.
- Tobler's first law of geography states that:
  "Everything is related to everything else, but near things are more related than distant things"

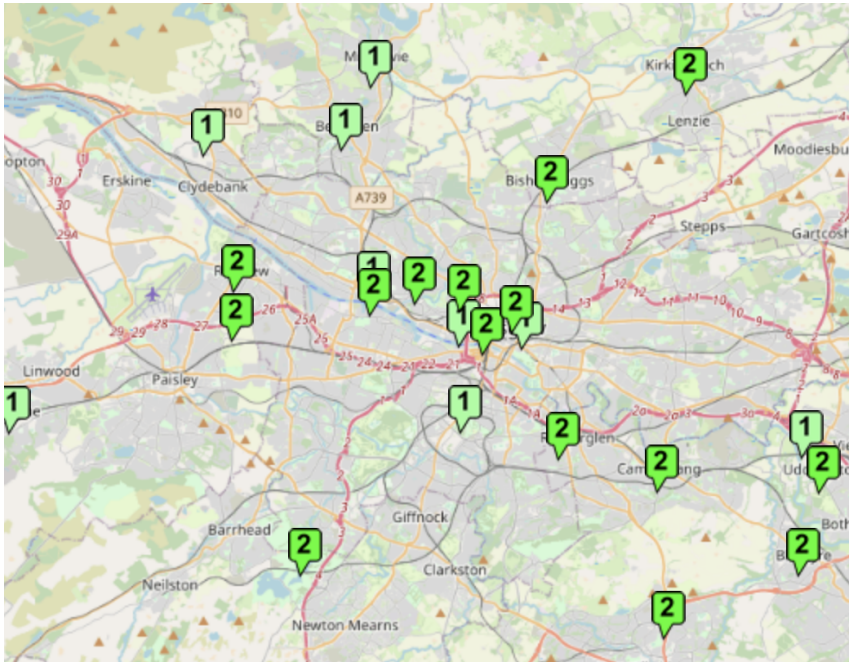
Sign of the Times

- Air pollution levels are measured at monitoring stations around Glasgow.
- 1 is "*Low*" and 10 is "*Very High*".

17th July 2022



17th January 2023



`https://www.scottishairquality.scot`

Goals of spatial analysis

- The overall goal of any piece of spatial analysis is to understand the spatial patterns in our data.
- This could involve any of the following:
  - Estimating differences in mean, variance or some other summary statistic over space.
  - Predicting the value at some unobserved location.
  - Identifying "hotspots" with a high (or low) value compared to the rest of the region.
- Typically this involves understanding and accounting for spatial autocorrelation in our data.

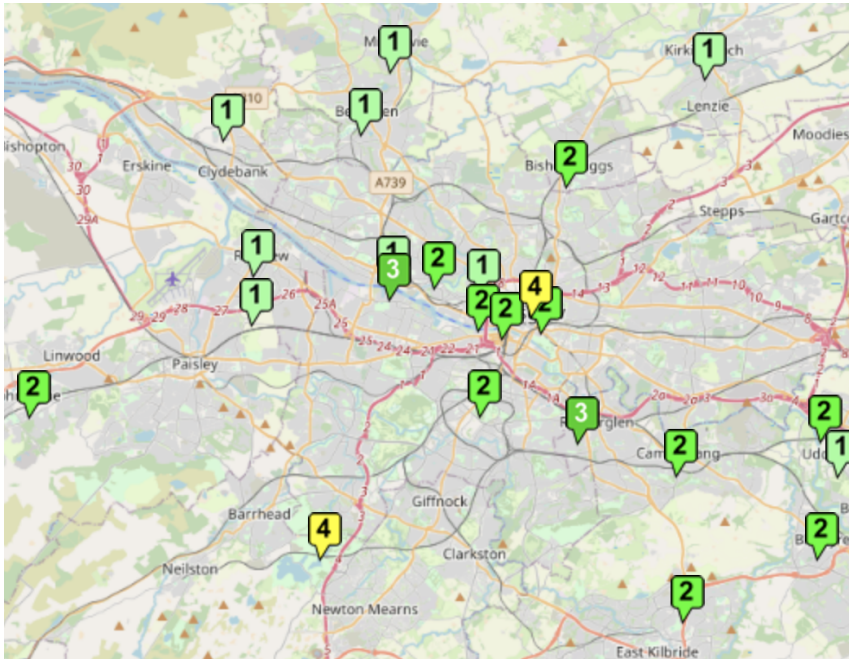
Types of spatial data

- There are three main categories of spatial data which we will look at during this course.
- These are geostatistical data, point processes and areal (or lattice) data.
- We will briefly introduce each of them and identify the main topics where they are used for environmental data.
- Note that there is a separate Spatial Statistics course which covers each of these types of data in more detail.

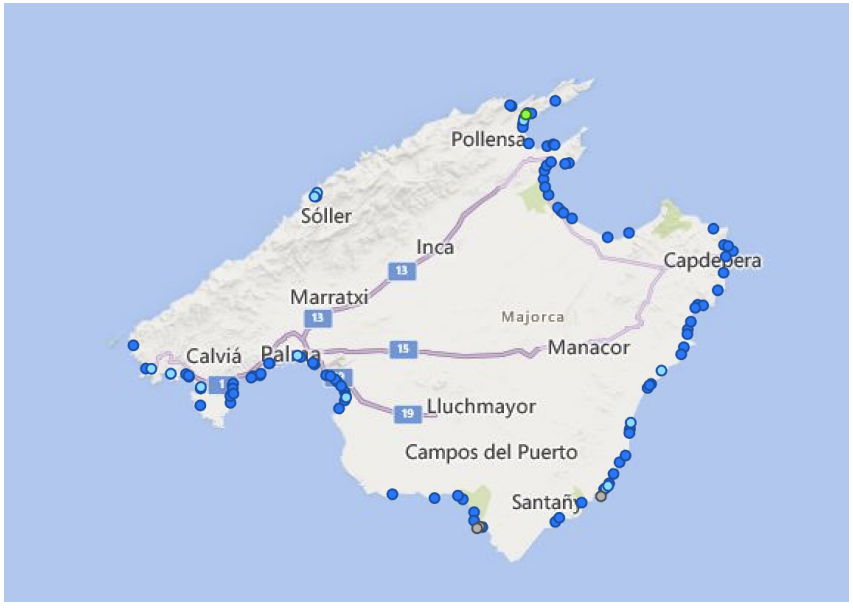
Geostatistical data

- Measurements are taken at a set of fixed locations to measure some continuous process.

Air pollution in Glasgow



Bathing water quality in Mallorca



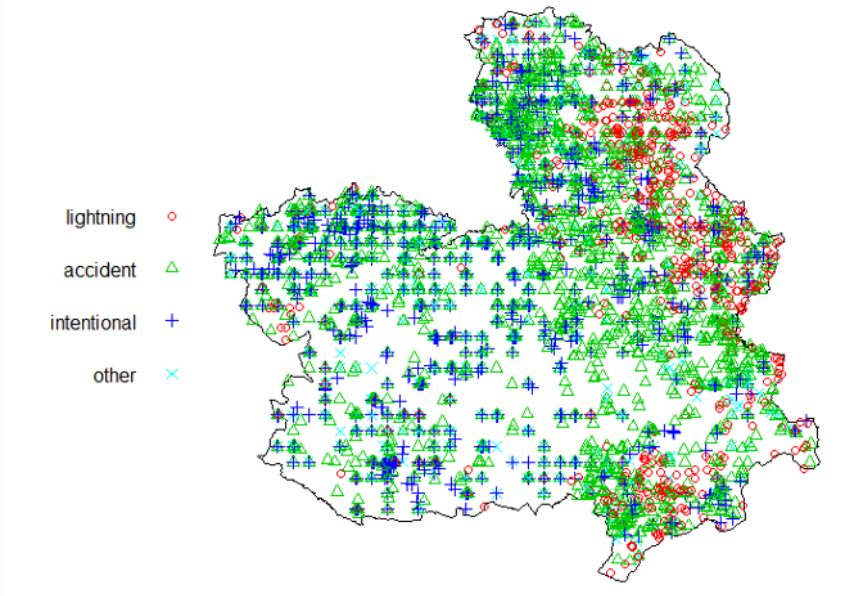
https://www.eea.europa.eu/themes/water/interactive/bathing/state-of-bathing-waters

Point process

- We measure the locations where events occur (eg trees in a forest, earthquakes) and the coordinates are our data.

Locations of flowers

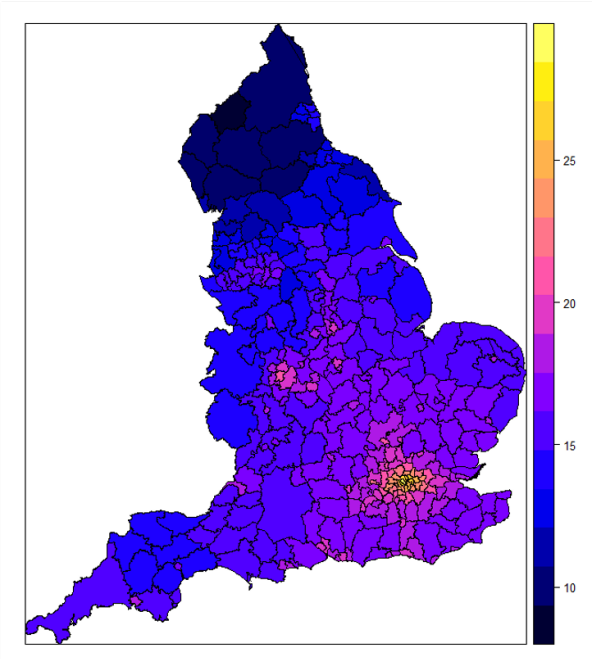
Forest fires in Castilla La Mancha, Spain



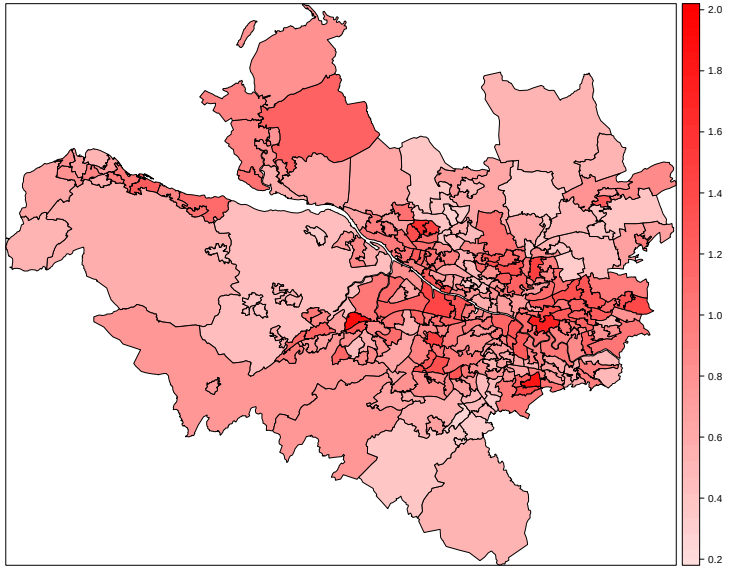
Areal

- Our data are measured as a summaries across a set of discrete, non-overlapping spatial units (eg postcode areas, council regions).

Pollution in English local authorities



Human respiratory disease in Glasgow



Introduction to Geostatistical Data

Geostatistical data

- Geostatistical data are the most common form of spatial data found in environmental setting.
- We regularly take measurements of an environmental variable of interest at a set of fixed locations.
- This could be data from samples taken across a region (eg water depth in a lake) or from monitoring stations as part of a network (eg air pollution).
- In each of these cases, our goal is to estimate the value of our variable across the entire space.

Understanding our region

- Let D be our two-dimensional region of interest.
- In principle, there are infinite locations within D, each of which can be represented by mathematical coordinates (eg latitude and longitude).
- We can identify any individual location as $\boldsymbol{s}_i = (x_i, y_i)$, where x_i and y_i are their coordinates.
- We can treat our variable of interest as a random variable, Z which can be observed at any location as $Z(\boldsymbol{s}_i)$.

Geostatistical process

- Our geostatistical process can therefore be written as:

$$\{Z(\boldsymbol{s}); \boldsymbol{s} \in D\}$$

- In practice, our data are observed at a finite number of locations, m, and can be denoted as:

$$z = \{z(\boldsymbol{s}_1), \ldots z(\boldsymbol{s}_m)\}$$

- We have observed our data at m locations, but often want to predict this process at a set of unknown locations.

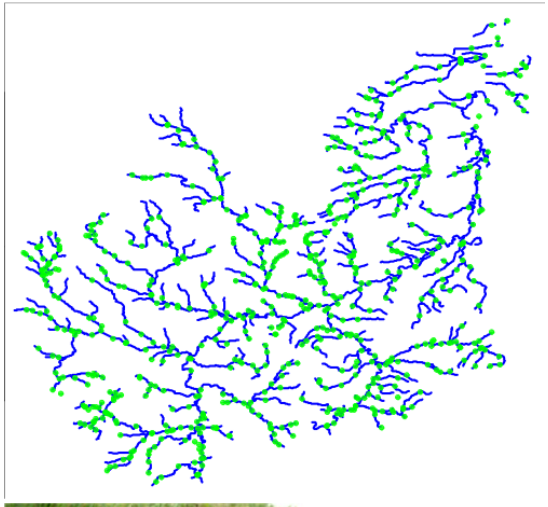
- For example, what is the value of $z(\boldsymbol{s}_0)$, where $\boldsymbol{s}_0$ is an unobserved site?

Geostatistical analysis

- There are two main steps in classical geostatistical analysis.
  1. How do I produce a statistical model for the data?
  2. How do I use my model to estimate quantities of interest?
- The first part requires us to think about how our measured datapoints relate to each other - in other words, to understand spatial autocorrelation.
- The second part requires us to use that information to predict the value at unmeasured locations, and then to produce maps or summary statistics based on this.

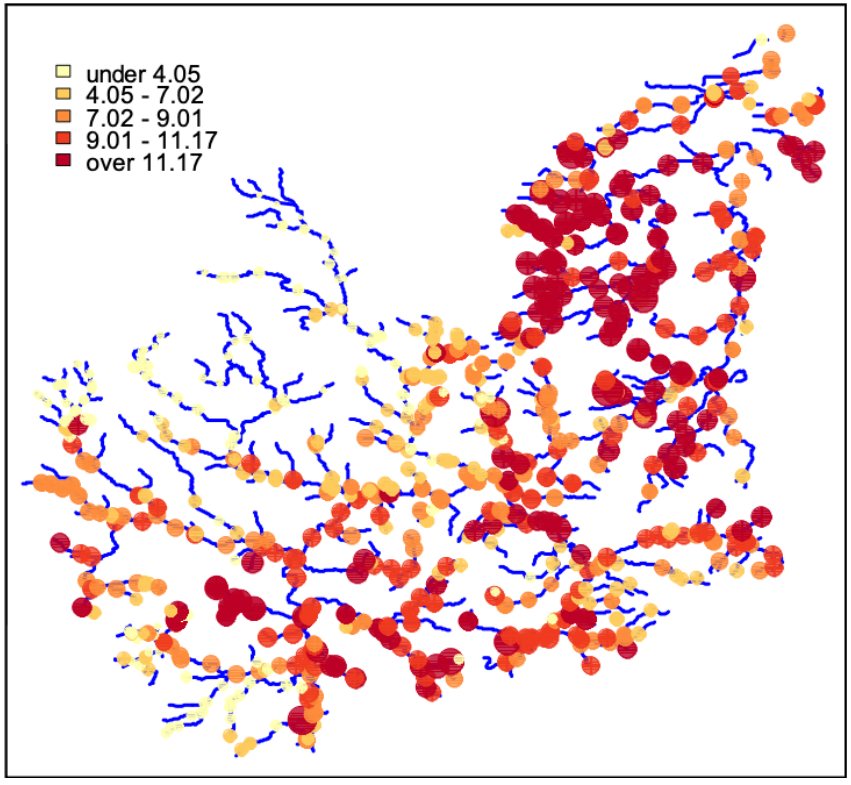
Example: River nitrogen

- We are interested in nitrate levels in the River Trent.
- A set of locations in the river network are sampled.
  - There are UK/EU directives for safe nitrate levels in water bodies such as rivers.
  - Our goal is to answer a **policy** question - which areas breach 50mg/l limit?
  - This requires us to estimate the levels at unmeasured locations.



Example: River nitrogen

- The plot below shows the average nitrate levels from 2003-2010 at each of our observed locations.
- Darker colours and larger plotting points correspond to higher levels of nitrate.

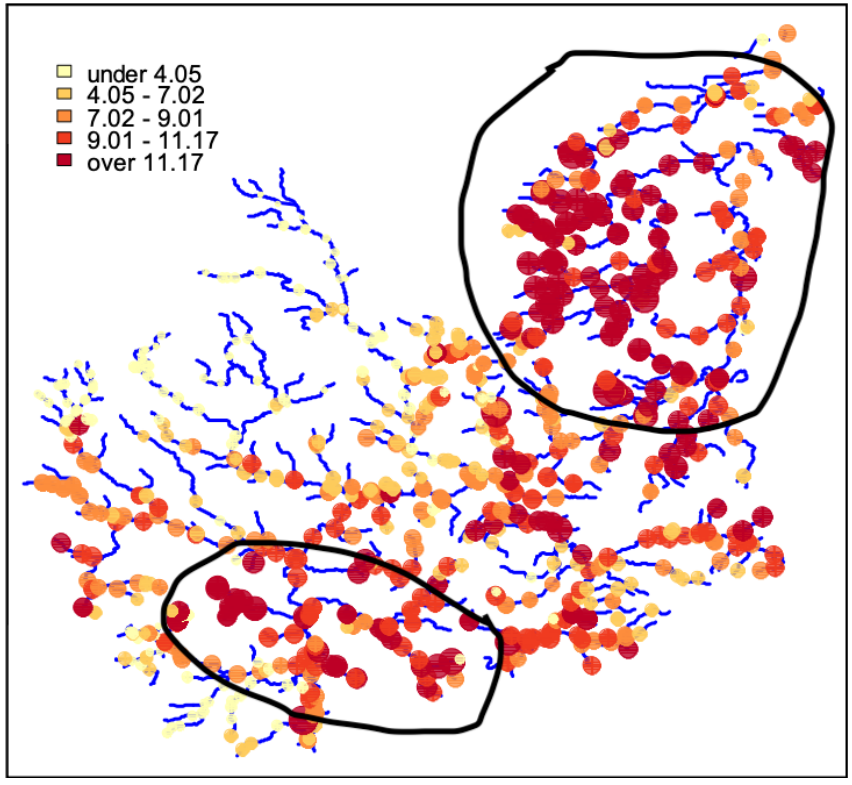


Golden

https://www.menti.com/alz6g6hw4mr7

To Be So Lonely

- Visually, it appears that the highest levels are located in the north-east and far south of the region.
- However, we require a statistical model to measure this objectively.



Modelling geostatistical data

- The key challenge in modelling geostatistical data is understanding **correlation**.
- Typically observations close together in space will be more similar than those which are further apart.
- Spatial correlation is usually driven by some unmeasured confounding variable(s) - for example, air pollution is spatially correlated because nearby areas tend to experience similar traffic levels.
- It is important that we account for these correlations in our analysis - failing to do so will lead to poor inference.

Modelling geostatistical data

- For a set of geostatistical data $\boldsymbol{z} = \{z(\boldsymbol{s}_1), \ldots, z(\boldsymbol{s}_m)\}$, we can consider the general model:

$$Z(\boldsymbol{s}_i) = \mu(\boldsymbol{s}_i) + e(\boldsymbol{s}_i)$$

- Here $\mu(\boldsymbol{s}_i)$ is a mean function which models trend and covariate effects.
- Then $e(\boldsymbol{s}_i)$ is the error process which accounts for any spatial correlation which exists after accounting for $\mu(\boldsymbol{s}_i)$
- Spatial statistics is therefore often focused on understanding the process for $e(\boldsymbol{s}_i)$.

Our key problem(s)

- We have observations at m locations

$$\boldsymbol{z} = \{z(\boldsymbol{s}_1), \ldots, z(\boldsymbol{s}_m)\}.$$

- We want to use these to obtain an estimate of $Z(\boldsymbol{s}_0)$ where $\boldsymbol{s}_0$ is an unobserved location.
- How do we model the spatial dependence between our observed sites $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_m$?
- What does this tell us about the dependence between our observed sites and our unobserved site $\boldsymbol{s}_0$?

Variograms

Modelling spatial dependence

- Spatial dependence is commonly modelled by a function known as a **variogram**.
- The variogram is similar in many ways to the autocorrelation function used in time series modelling.
- In simple terms, it is a function which measures the difference in the spatial process between a pair of locations a fixed distance apart.
- In order to define the variogram, it is important to first understand some more features of a geostatistical process.

Mean

- If we have a geostatistical process $\{Z(\boldsymbol{s}); \boldsymbol{s} \in D\}$, then its mean can be expressed as

$$\mu_z(\boldsymbol{s}) = E[Z(\boldsymbol{s})] \text{ for all } \boldsymbol{s} \in D.$$

- Our process Z is then said to be Gaussian if our random variable at the set of observed locations is multivariate normal.

- In other words, Z is Gaussian if

$$\{Z(\boldsymbol{s_1}, \ldots, Z(\boldsymbol{s_m}\} \sim \text{MVN}(\mu_z(\boldsymbol{s}), C_z(\boldsymbol{s})).$$

Variance and covariance

- Similarly, the covariance can be expressed as:

$$\begin{aligned} C_z(\boldsymbol{s},\boldsymbol{t}) &= \mathrm{Cov}\left(Z(\boldsymbol{s}), Z(\boldsymbol{t})\right) \\ &= E\left[(Z(\boldsymbol{s}) - \mu_z(\boldsymbol{s}))(Z(\boldsymbol{t}) - \mu_z(\boldsymbol{t}))\right] \end{aligned}$$

- Here, the covariance measures the strength of the linear dependence between $Z(\boldsymbol{s})$ and $Z(\boldsymbol{t})$.

- As usual, we can compute the variance of $Z(\boldsymbol{s})$ as a special case of the covariance where $\boldsymbol{s} = \boldsymbol{t}$.

Weak stationarity

- Our geostatistical process can be described as **weakly stationary** if the following criteria are met:
  1. $E[Z(\boldsymbol{s})] = \mu_z(\boldsymbol{s}) = \mu_z$ - a finite constant which does not depend on $\boldsymbol{s}$.
  2. $C_z(\boldsymbol{s}, \boldsymbol{s} + \boldsymbol{h}) = C_z(\boldsymbol{h})$ - a finite constant which can depend on $\boldsymbol{h}$ but not $\boldsymbol{s}$.
- Condition 1 states that our mean function must be constant in space, with no overall spatial trend.
- Condition 2 states that for any two locations, their covariance depends only on how far apart they are (their **spatial lag**, h), not their absolute position.

Isotropy

- A geostatistical process is said to be **isotropic** if the covariance function is *directionally invariant*.
- This means that the covariance between two points a distance h apart is the same no matter which direction you travel in.
- Mathematically, this can be denoted by

$$C_z(\boldsymbol{h}) = C_z(||\boldsymbol{h}||).$$

Daylight

https://www.menti.com/alws53ba6bzm

Limitations of this course

- In this course, we will only look at Gaussian, weakly stationary and isotropic processes.
- This means that:
  - This means our random variables across our observed locations follow a multivariate Gaussian distribution.
  - The mean of this distribution is constant over space.
  - The covariance of this distribution is only a function of the lag, not the position or direction of the points.
- Other models do exist for more complex processes, but we will not explore these.

Autocovariance

- The function describing the dependence between values of our process Z separated by different lags is known as the **autocovariance function**.
- This is similar to the autocorrelation function (ACF) used for temporal data.
- In geostatistical models, a variant of this known as a **variogram** is typically used to estimate spatial relationships.

Variogram

- The variogram measures the variance of the difference in the process at two spatial locations $\boldsymbol{s}$ and $\boldsymbol{s} + \boldsymbol{h}$.
- The variogram is defined as

$$\text{Var}[Z(\boldsymbol{s}) - Z(\boldsymbol{s} + \boldsymbol{h})] = E[(Z(\boldsymbol{s}) - Z(\boldsymbol{s} + \boldsymbol{h}))^2] = 2\gamma_z(\boldsymbol{h}).$$

- Here, $2\gamma_z(\boldsymbol{h})$ is the variogram, but in practice we use the **semi-variogram**, $\gamma_z(\boldsymbol{h})$.
- We use the semi-variogram because our points come in pairs, and the semi-variance is equivalent to the variance per point at a given lag.

Notes on (semi)-variogram

- When the variance of the difference $Z(\boldsymbol{s}) - Z(\boldsymbol{t})$ is relatively small, then $Z(\boldsymbol{s})$ and $Z(\boldsymbol{t})$ are similar (spatially correlated).
- When the variance of the difference $Z(\boldsymbol{s}) - Z(\boldsymbol{t})$ is relatively large, then $Z(\boldsymbol{s})$ and $Z(\boldsymbol{t})$ are less similar (closer to independence).
- If our process is *weakly stationary* and *isotropic* we can show that

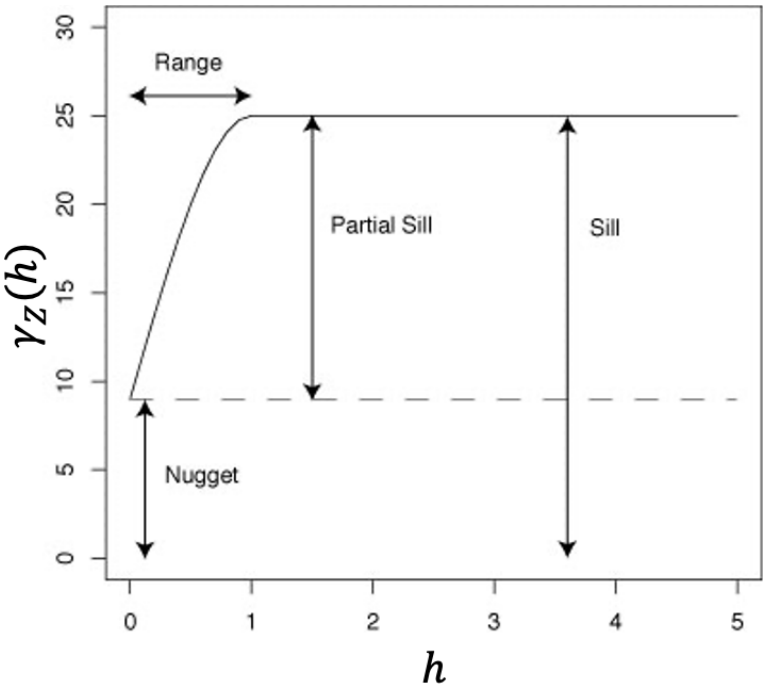
$$\gamma_z(\boldsymbol{h}) = \sigma_z^2 - C_z(\boldsymbol{h}).$$

- Therefore, if we know the covariance function, we can calculate the (semi)-variogram.

What does the variogram look like?

- The plot below shows a typical variogram, with the variance increasing as the lag increases.

- The **sill** is the maximum variance as $h \to \infty$.
- The **nugget** is the minimum variance as $h \to 0$.
- The **range** is the distance to the sill.



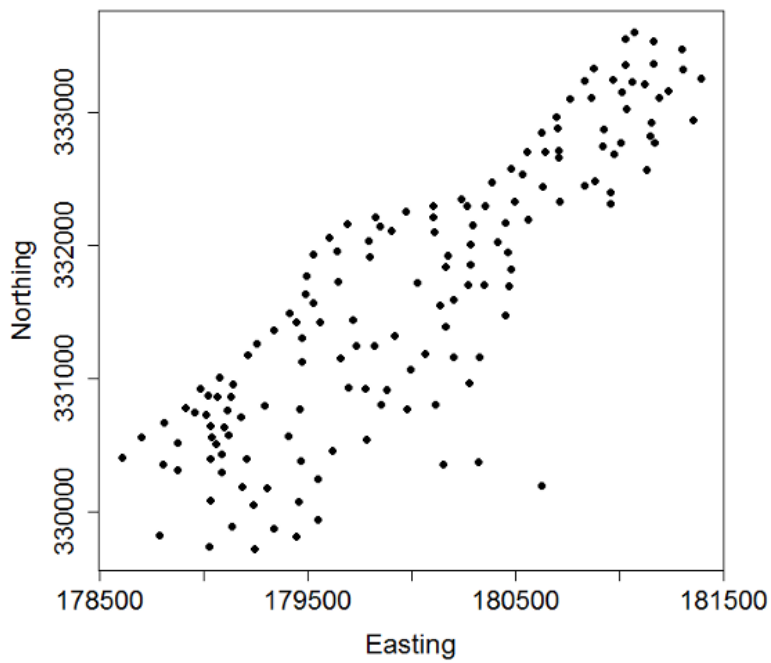
- Points further apart than the range are assumed to be uncorrelated.

Estimating the variogram

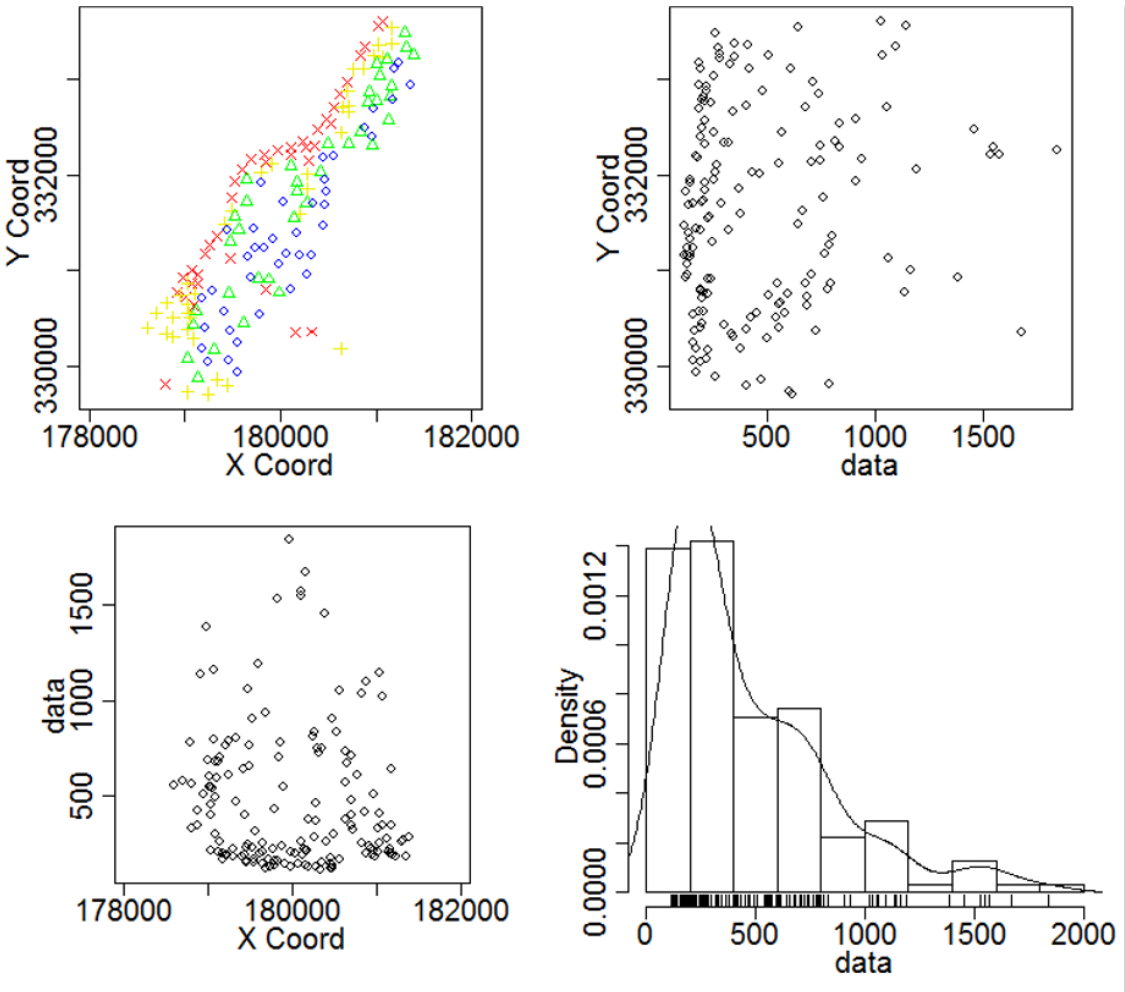
- The variogram is a function of the underlying geostatistical process Z.
- In practice, we only have access to m realisations of this process, and therefore we have to estimate the variogram.
- This is known as the *empirical variogram*.
- We obtain this by computing the semi-variance for all possible pairs of observations: $\gamma(\boldsymbol{s},\boldsymbol{t}) = 0.5(Z(\boldsymbol{s}) - Z(\boldsymbol{t}))^2$.
- We can then plot the empirical variogram by plotting these semi-variances against their corresponding lags.

Example - River Meuse

- We are interested in soil zinc levels in the flood plains of the River Meuse (in France, Belgium and the Netherlands).
- Our goal is to model the spatial correlation in the data so that we can predict zinc levels at unsampled locations.
- A set of sampling locations were selected.
- Our data consist of location coordinates and zinc levels.



Example - River Meuse

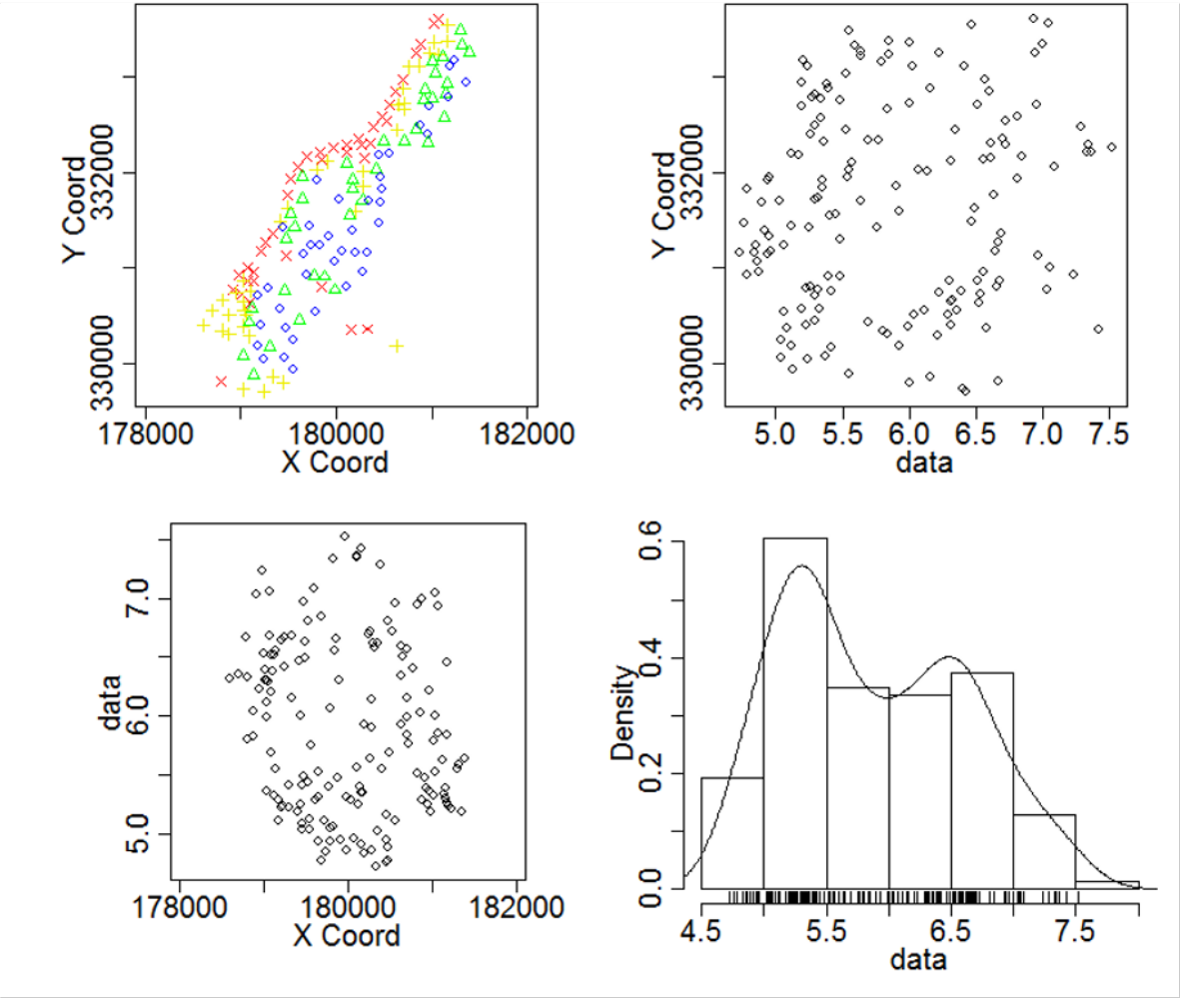


Plots: (a) map colour coded by zinc level, (b) and (c) the zinc levels plotted against the x and y coordinates and (d) a density plot of the zinc levels.

As It Was

https://www.menti.com/alws53ba6bzm

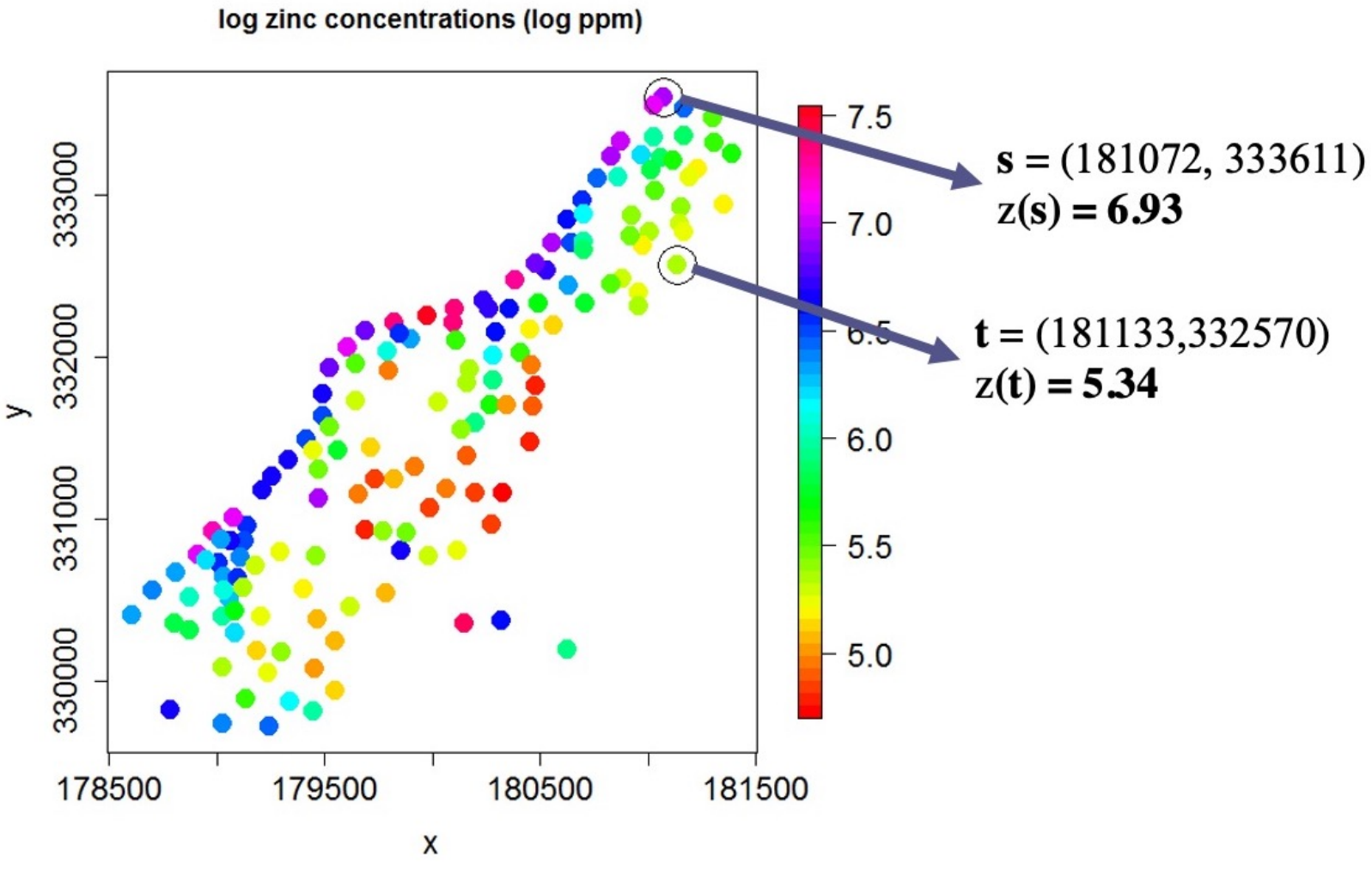
Transforming our data



Plots: (a) map colour coded by *log* zinc level, (b) and (c) the *log* zinc levels plotted against the x and y coordinates and (d) a density plot of the *log* zinc levels.

Computing the empirical variogram

- To illustrate how an empirical variogram is computed, consider the two highlighted locations below.



Computing the empirical variogram

- We can first compute the distance between the two locations using the standard Euclidean distance formula as

$$h = \sqrt{(181072 - 181133)^2 + (333611 - 332570)^2} = 1042.8$$

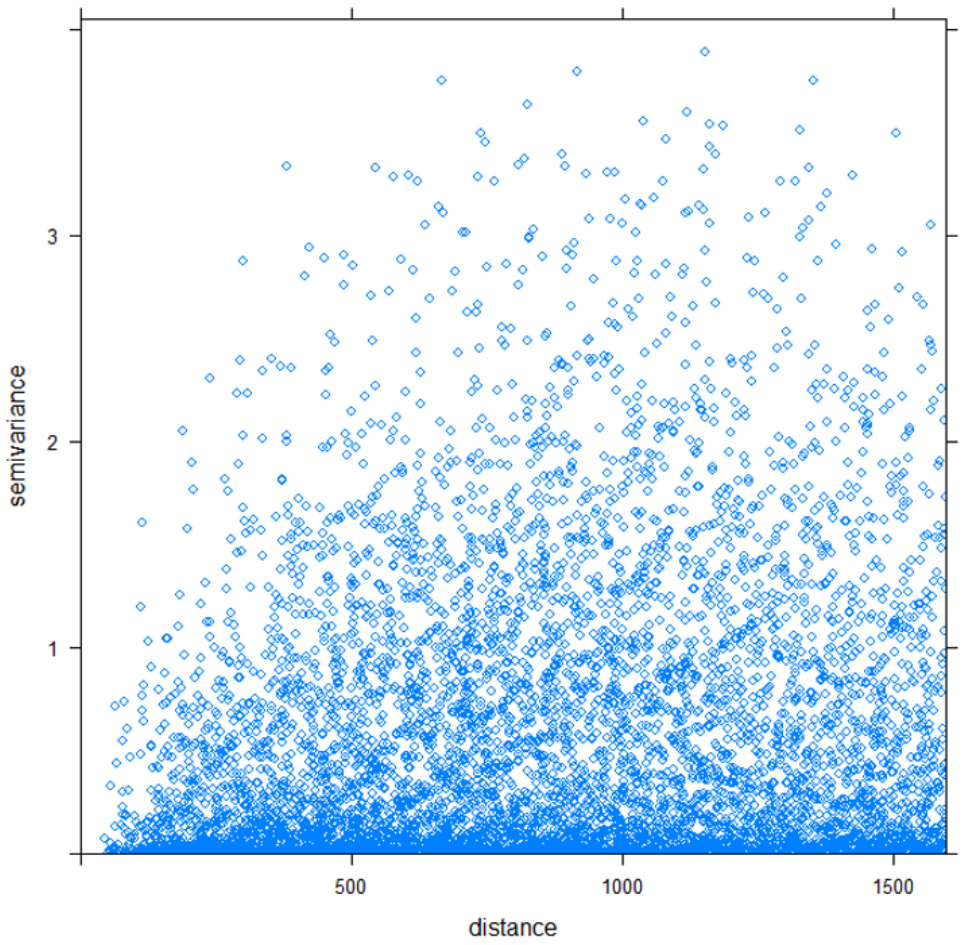
- Next, we compute the semi-variance between the points using their observed values as

$$\begin{aligned}\gamma(\boldsymbol{s}, \boldsymbol{t}) =& 0.5(Z(\boldsymbol{s}) - Z(\boldsymbol{t}))^2 \\ =& 0.5(6.93 - 5.34) = 1.27\end{aligned}$$

- We repeat this process for every possible pair of points, and plot h against $\gamma(\boldsymbol{s}, \boldsymbol{t})$ for each.

Daydreaming

- This plot shows the semi-variances for each pair of points.
- Each pair of points has a different distance, making it difficult to use this for prediction.



Computing the empirical variogram

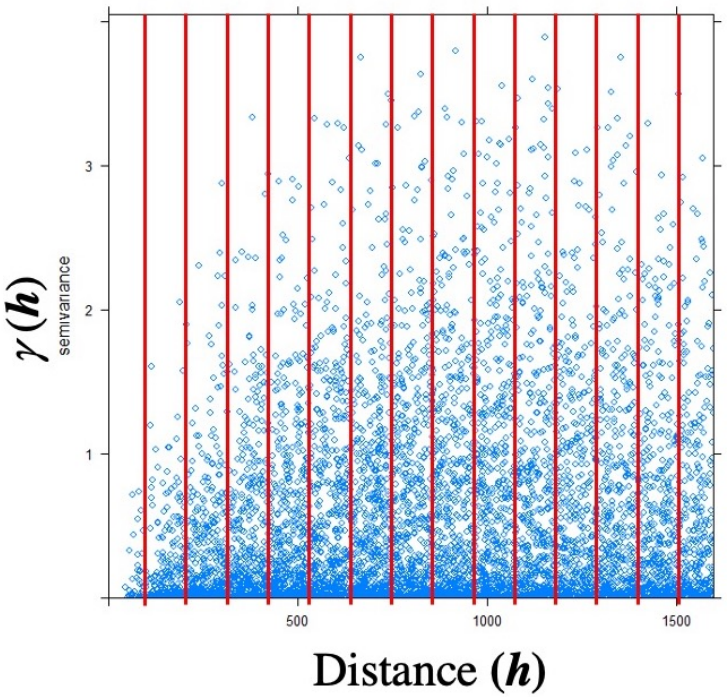
- To make the variogram easier to use and interpret, we divide the distances into a set of discrete bins, and compute the average semi-variance in each.
- We compute this binned empirical variogram as

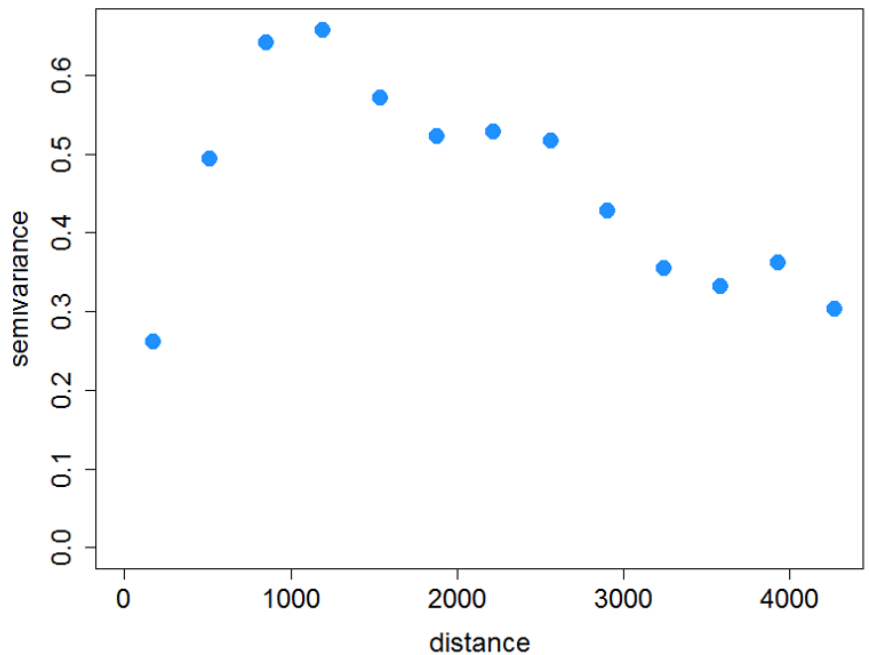
$$\gamma(\boldsymbol{h}) = \frac{1}{2N(h_k)} \sum_{(\boldsymbol{s},\boldsymbol{t})\in N(h_k)} [z(\boldsymbol{s}) - z(\boldsymbol{t})]^2$$

- Here, k is the number of bins and $N(h_k)$ is the number of points in the bin with average distance h.
- We then construct a plot of our empirical variogram and use this to estimate the covariance structure.

Fine Line

- The bins are illustrated on the left, and the empirical variogram obtained from them is shown on the right.
- We can start to think about estimating a nugget, sill and range based on this empirical variogrram.





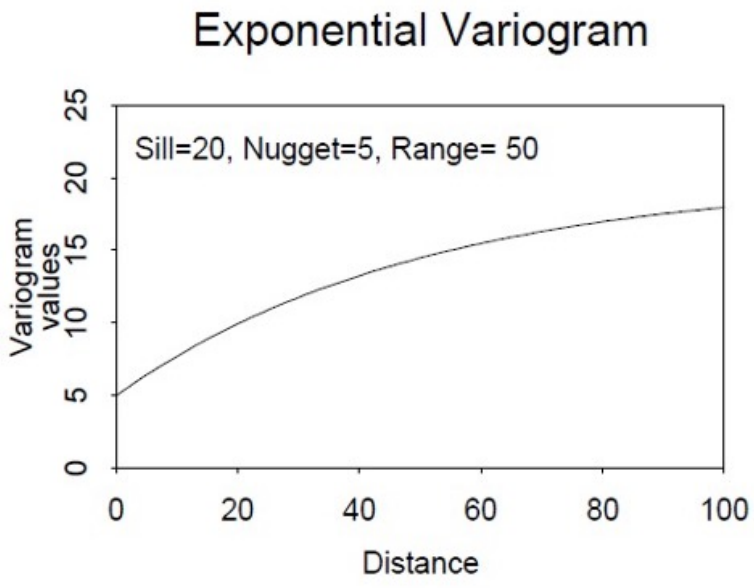
Fitting our model

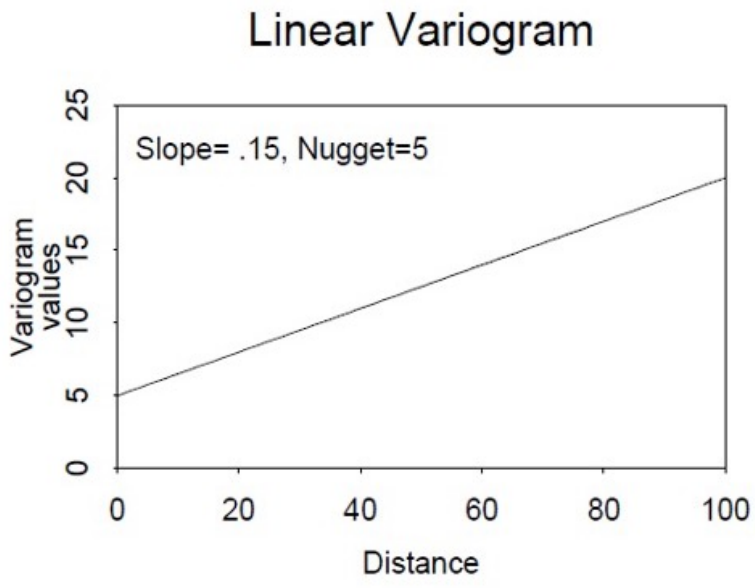
- Once we have computed an empirical variogram, we have to think about fitting a model to it.
- We only have estimates at a small number of lags, we would like a continuous model which explains the changes in dependence structure as h increases.
- Our variogram model needs to have the following properties:
  - Monotonically increasing
  - A constant maximum (sill)
  - A positive intercept (nugget)
- Several models exist which satisfy these criteria.

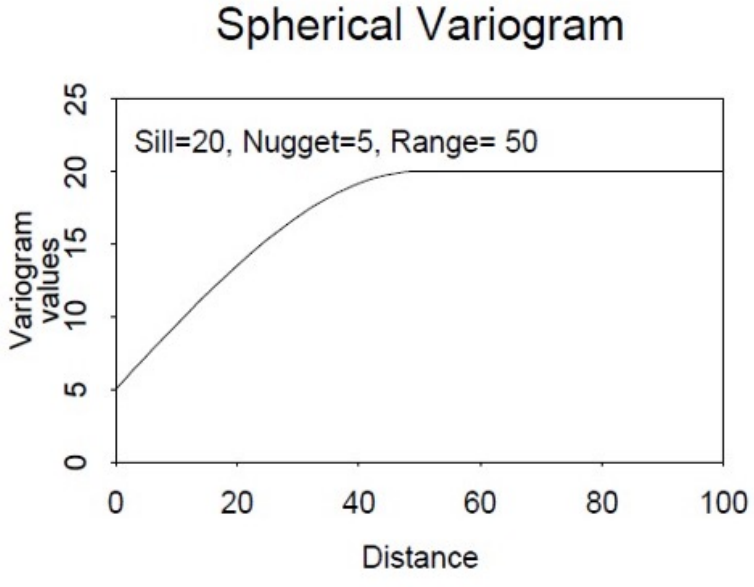
Choosing our model

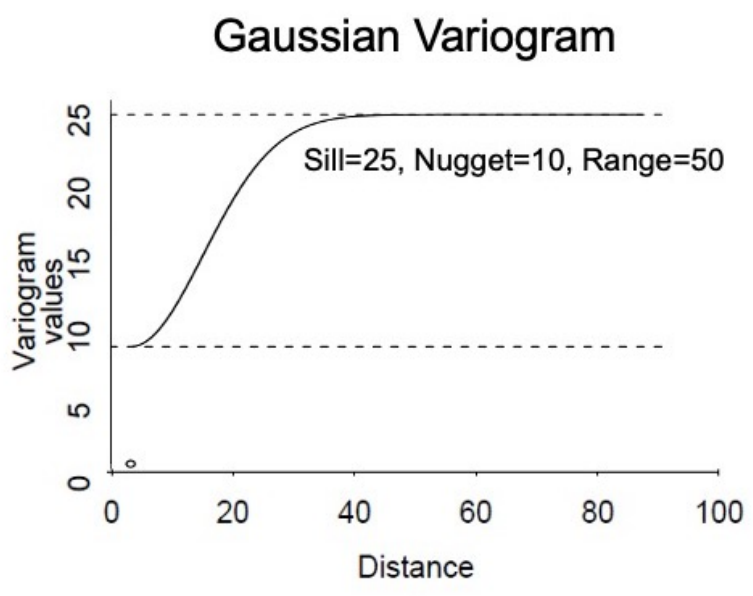
- According to Webster and Oliver (2007), choosing variogram models is one of the most controversial topics in geostatistics.
- We often choose the right model based on mathematical criteria such as least squares, maximum likelihood or AIC.
- One of the more popular approaches, proposed by Cressie (1985), is to use weighted least squares, where the weights are based on the number of observations within each ‘bin’.
- The outcome of this is that more weight is given to the lags which have been estimated with more data points, which are usually the shorter lags.

Examples of variogram models









Examples of variogram models

Random $\gamma(h) = \begin{cases} 0, & \text{if } h = 0 \\ c, & \text{otherwise} \end{cases}$

Spherical $\gamma(h) = \begin{cases} c \cdot 1.5\frac{h}{a} - 0.5\left(\frac{h}{a}\right)^3, & \text{if } h \le a \\ c, & \text{if } h \ge a \end{cases}$ *a is the range of influence.*

Exponential $\gamma(h) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-h/\phi)) & \text{if } h > 0 \\ 0, & \text{if } h = 0 \end{cases}$

Gaussian $\gamma(h) = \begin{cases} \tau^2 + \sigma^2(1 - \exp\{(-h/\phi)^2\}) & \text{if } h > 0 \\ 0, & \text{if } h = 0 \end{cases}$

Notation

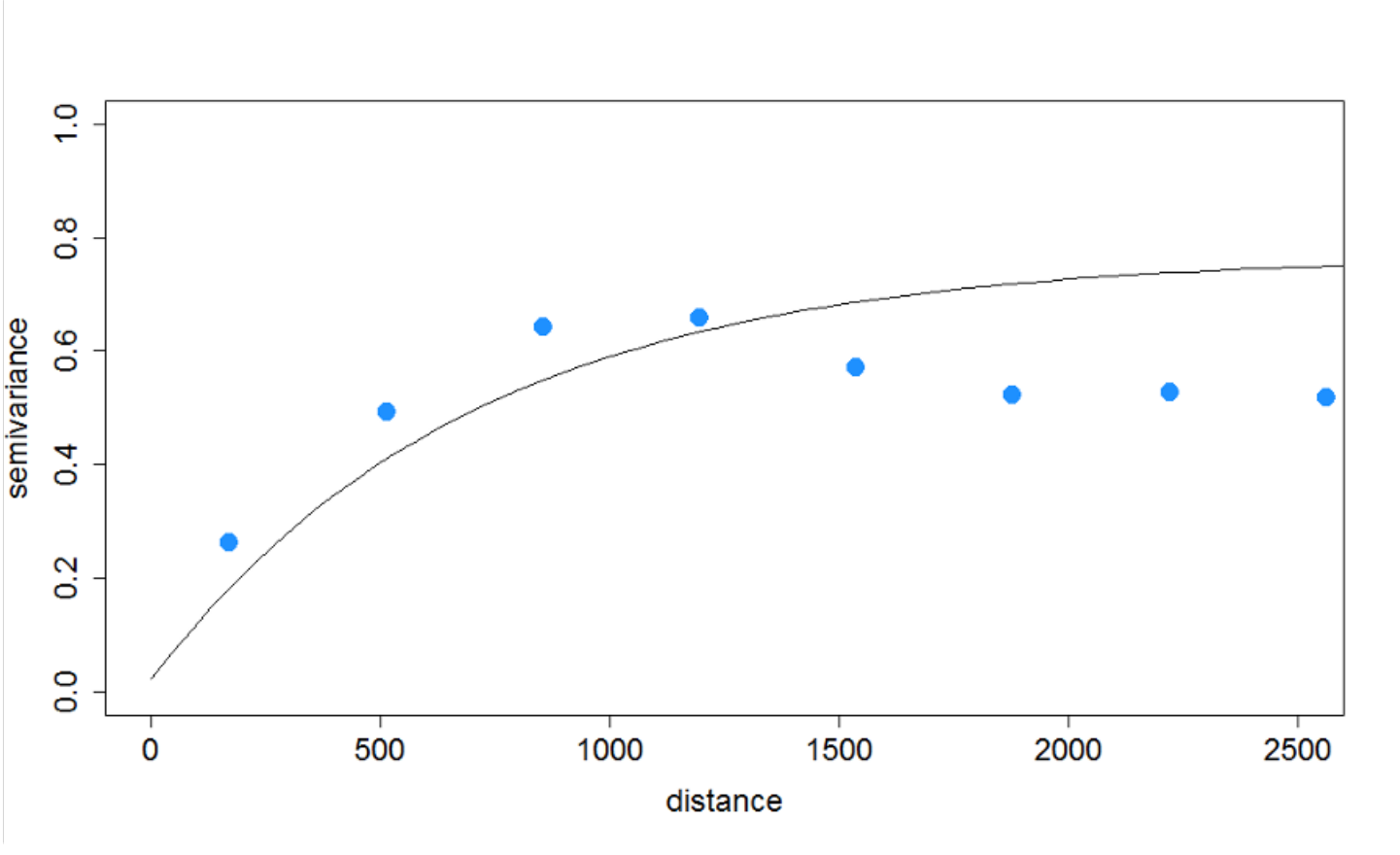
- The key features of our variogram are represented by the following model parameters
  - $\tau^2 > 0$ is the **nugget**.
  - $\sigma^2 > 0$ is the **partial sill**.
  - $\phi > 0$ is the **range parameter**.

- Note that ϕ is not the range itself, but rather a parameter which controls how quickly the covariance decays towards zero (or the variogram increases to the sill).

- A smaller value of ϕ means the covariance function decays to zero quickly, a larger value means it decays to zero more slowly.

Example - Meuse river

- We can now fit a variogram model to the empirical variogram obtained from the Meuse river example.
- The black solid line below shows an exponential variogram model (fitted using the `geoR` package).



Prediction

- Once we have estimated a variogram to account for our spatial autocorrelation, we can start to think about making predictions.
- Spatial prediction is the process of predicting the value of our variable of interest at an unobserved location $\boldsymbol{s}_0$.
- As with any statistical prediction, we use what we know about our observed data, including their values, how far our unobserved location is from them, and our variogram.
- There are many methods for spatial predictio, including regression modelling, distance weighted interpolation and an approach known as **kriging**.

Kriging

- Kriging is an approach named after its inventor D. G. Krige, who worked in the mining industry in South African in the 1950s.
- He used this approach to understand the spatial pattern of mineral resources.
- It is a relatively simple and theoretically appealing approach, and is therefore incredibly popular for geostatistical prediction.
- Kriging interpolates between previously observed locations in order to predict at new locations.

Ordinary Kriging

- Ordinary kriging is a form of kriging which assumes that the overall mean and variance of the region is constant.
- Predictions at unsampled locations are made using a weighted average of the observations.

$$z^*(\boldsymbol{s}_0) = \sum_{i=1}^{m} \lambda_i z(\boldsymbol{s}_i)$$

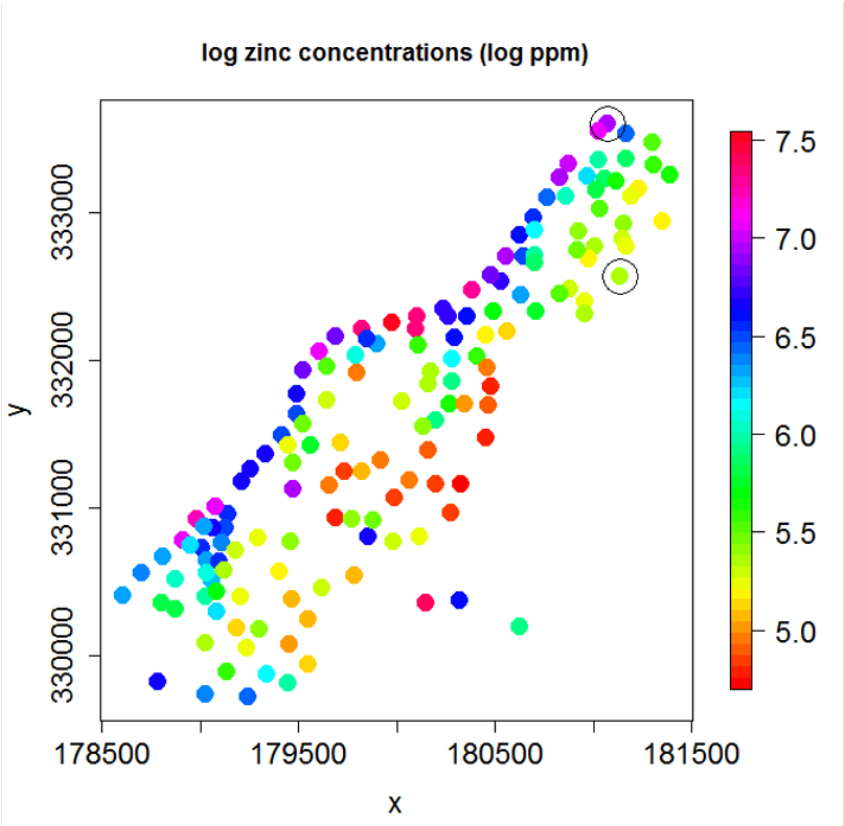
- The weights λ_i can be estimated in a number of different ways, and are commonly based on the variogram.
- The weights are therefore usually proportional to the distance between the observed and new locations.

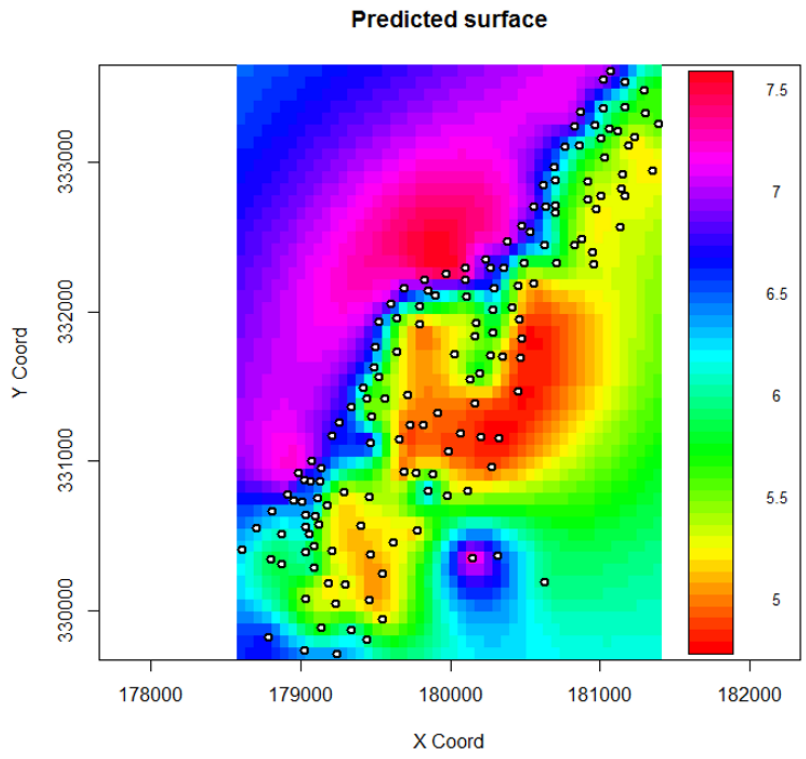
More on Kriging

- There are other types of kriging such as *universal kriging* and *block kriging* which are more appropriate for different structures of spatial data.
- Whichever method is used, we can obtain a set of predictions over a grid to allow us to plot a map of the predicted spatial surface for our variable of interest.
- Our predictions will be of better quality in areas where there are lots of observed values, and poorer quality when we have less nearby data available.
- Therefore we can also generate a map of the uncertainty across the region.

Predicted Surface

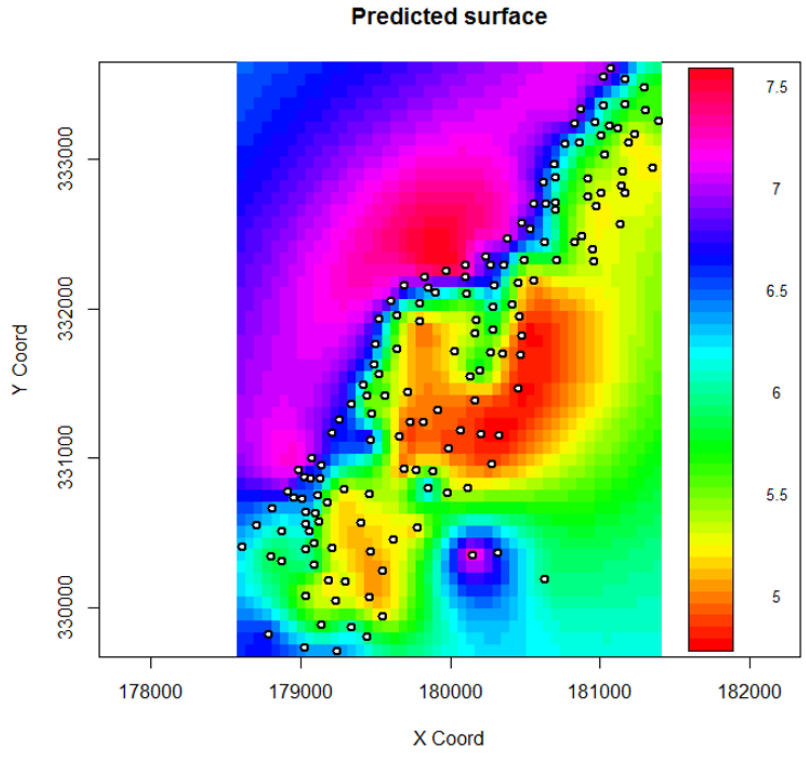
- The surface map on the right is generated by using kriging to make predictions over a fine grid.
- Notice the ‘edge effects’ - a gradual smoothing towards the mean as you move away from any actual data in the top left and bottom right.

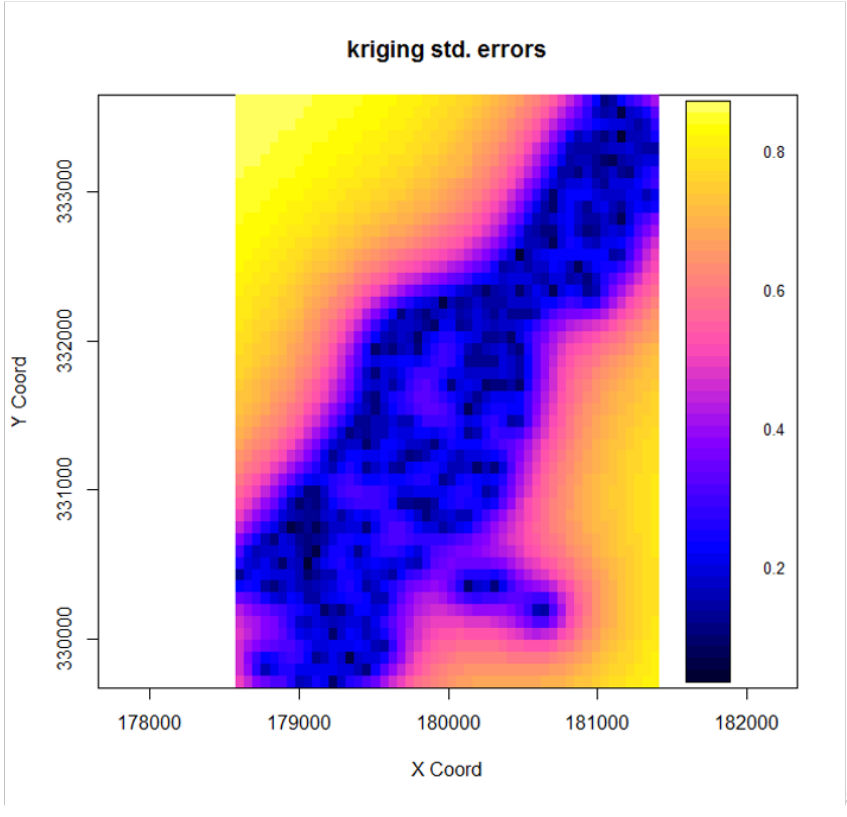




Uncertainty

- The map on the left shows the uncertainties associated with our estimated surface.
- There is lower uncertainty in the areas with lots of observed data, and higher uncertainty as we move away from the observed data.





Geostatistical modelling - summary

- When working with geostatistical data, we first have to understand what sort of spatial trend, if any, is present in our data.
- Next, we plot the empirical (semi-)variogram and choose a suitable model structure.
- We then identify the parameters (sill, nugget, range) of this chosen model structure.
- Finally, we use this model to obtain weights which can be used for interpolation via, for example, kriging, and use this to estimate the surface on a map.

Treat People With Kindness

https://www.menti.com/alws53ba6bzm

Temporary page!

LaTeX was unable to guess the total number of pages correctly
As there was some unprocessed data that should have been
added to the final page this extra page has been added to rece
it.
If you rerun the document (without altering it) this surplus pa
will go away, because LaTeX now knows how many pages to
expect for this document.