Spatial Autocorrelation I: Variograms

Michael Noonan

DATA 589: Spatial Statistics

Table of contents

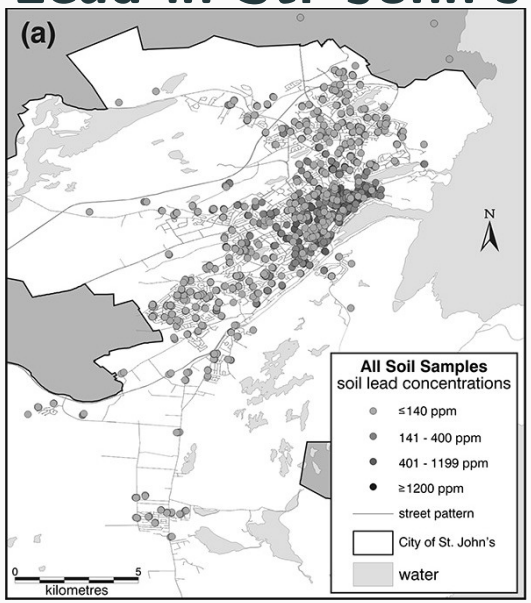
Overview

Overview

So far we have covered situations where locations were the variables of interest, and we studied these data using point processes

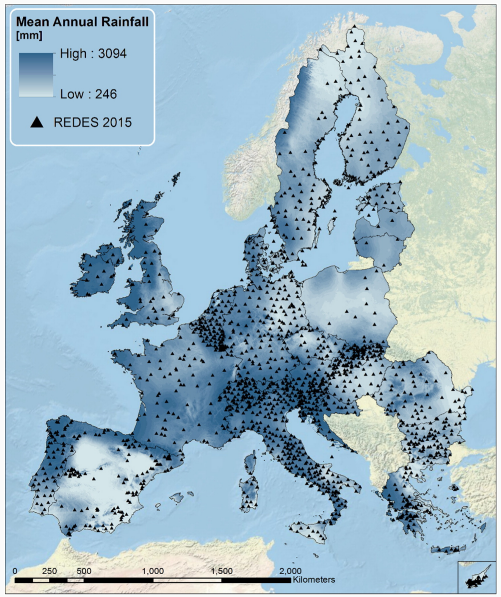
...but there are many situations where locations are arbitrary:

Lead in St. John's



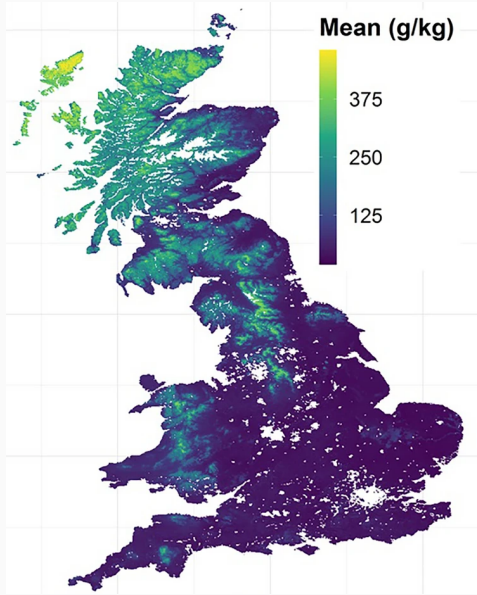
(Bell *et al.*, 2010)

Rainfall in Europe



(Ballabio *et al.*, 2017)

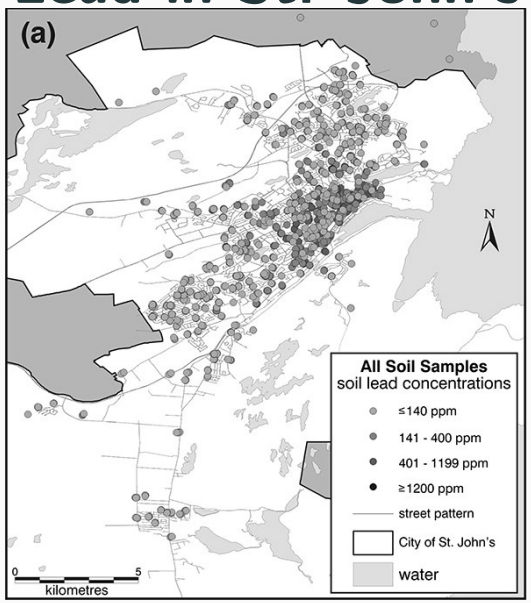
Soil Carbon in the UK



(Feeney *et al.*, 2022)

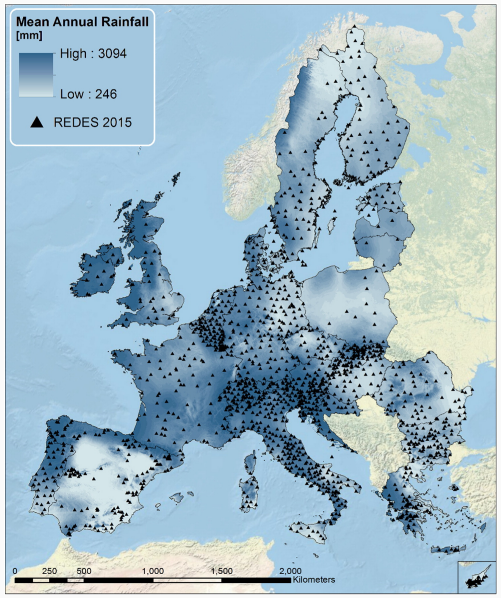
Overview

Lead in St. John's



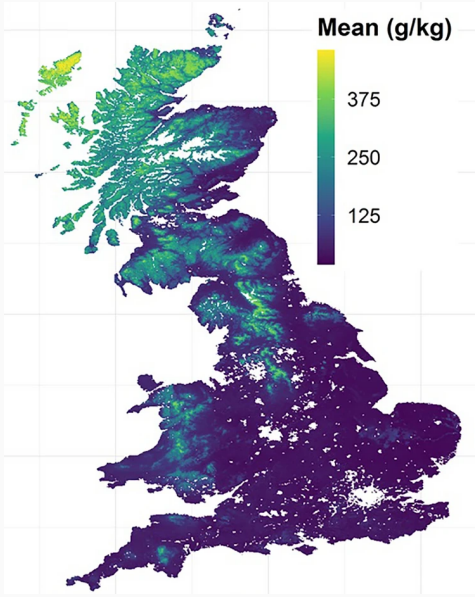
(Bell *et al.*, 2010)

Rainfall in Europe



(Ballabio *et al.*, 2017)

UK Soil Carbon



(Feeney *et al.*, 2022)

These data can't be treated as a point process because the locations of the points are (mostly) meaningless

...but they also can't be modelled using off the shelf techniques because they break an important assumption; that the data are IID.

Spatial Autocorrelation

Data are often collected by measuring quantities over space (e.g., abundances, concentrations, etc.).

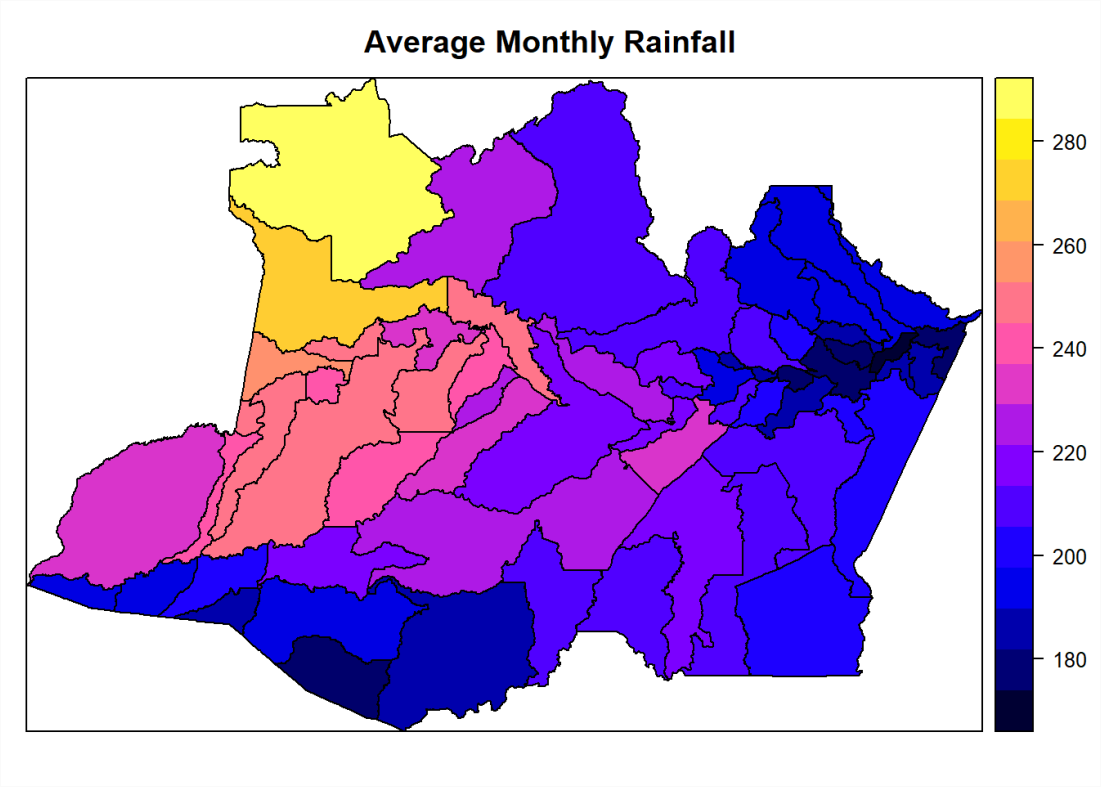
When this is the case, spatial autocorrelation can arise when the variation between the values of the datapoints is affected by their spatial distance (i.e., data that are close together in space more similar than data collected further apart).

The underlying reason for this is that many of the drivers (e.g., environmental conditions, topography, geology, etc.) act at large spatial scales.

The result is that we can not assume that spatially collected data are IID.

Spatial Autocorrelation cont.

What if we're studying the effect of rainfall on species diversity in the Amazon?



Source: Tadashi Fukami and Jes Coyle

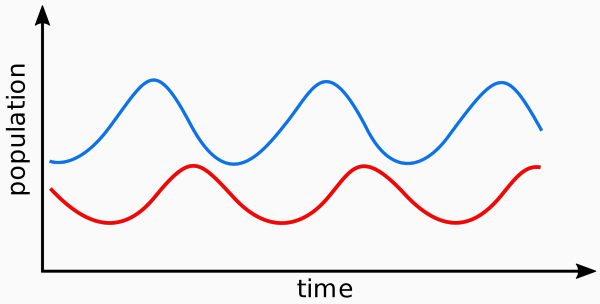
Because rainfall is correlated in space, species diversity will also be correlated in space (if the relationship exists).

Impacts of autocorrelation

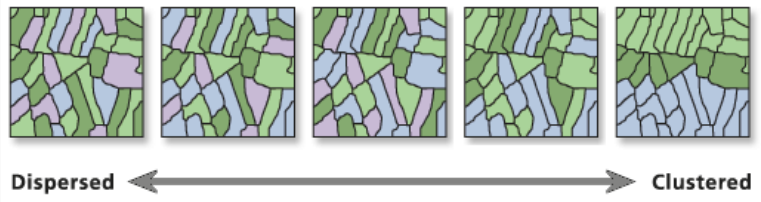
Sources of autocorrelation

Anything that causes some data points to be more similar to each other than others can result in autocorrelation.

- **Time**: Data that are close together in time are more related.



- **Space**: Data that are close together in space are more related.



We will be focusing on spatial autocorrelation, but the ideas translate to other sources of autocorrelation.

Autocorrelation is important

© Chris Sorensen

Dr. Sam Wang, Neuroscientist
—Princeton Election Consortium

"It is totally over. If Trump wins I will eat a bug."

© ABC News

Nate Silver, Statistician
—FiveThirtyEight.com

"Trump Is Just A Normal Polling Error Behind Clinton."

Autocorrelation is important

© Chris Sorensen

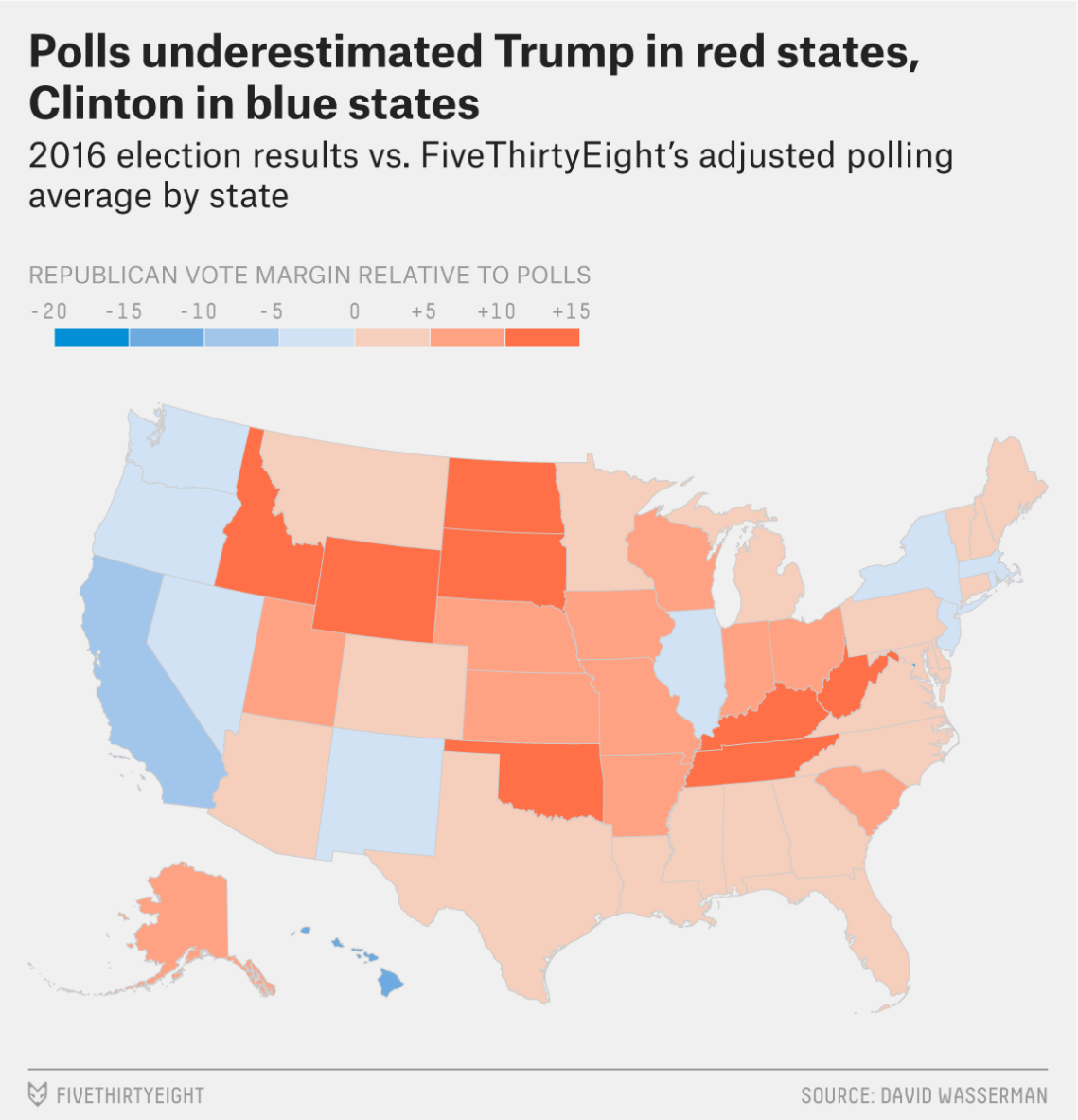
© ABC News

Nate Silver, Statistician

—FiveThirtyEight.com

"Trump Is Just A Normal Polling Error Behind Clinton."

Ignoring non-ind. $\rightarrow$ overconfidence



Are these polling errors independently distributed?

This same statistical issue that caused overly confident predictions of Clinton's 2016 victory can result in overconfidence in parameter estimates and predictions in regression models.

Autocorrelation impact

Sample size, n is the denominator when calculating SEs and CIs.

$$\text{SE} = \frac{\sigma}{\sqrt{n}} \qquad\qquad 95\%\text{CI} = \bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

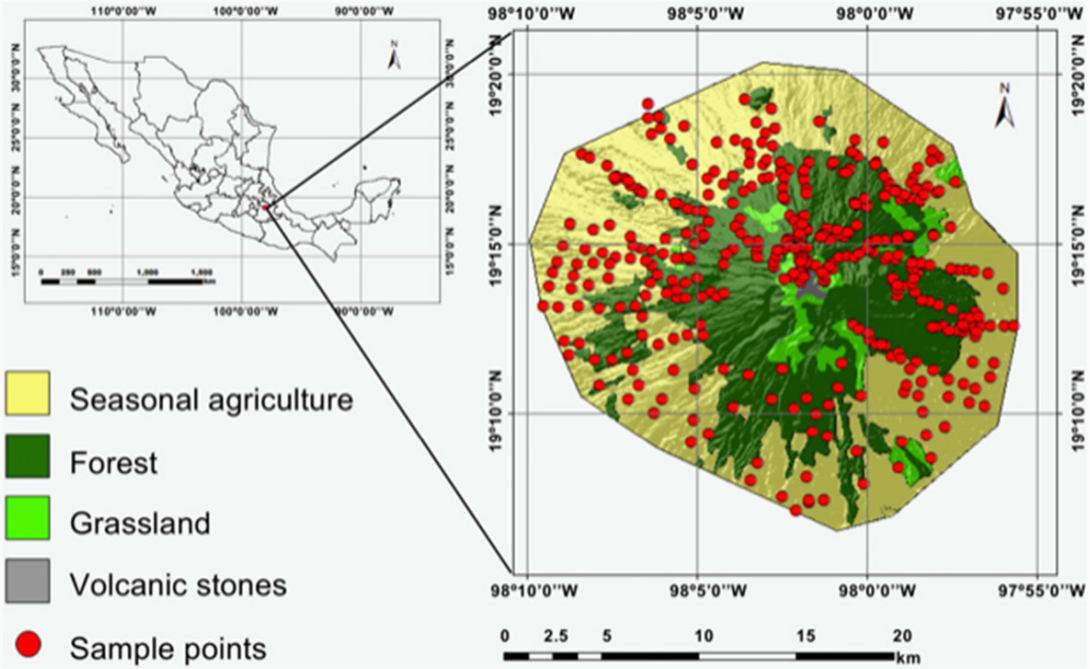
All else equal: $\uparrow n = \downarrow \text{SE}$ & $\downarrow \text{CI}$

But with autocorrelated data each new datapoint is related to a previously collected datapoint and does not bring a full independent datapoint worth of information (e.g., 90% autocorr. $\approx$ 10% new info).

When data are autocorrelated $n_{\text{effective}} < n$, meaning SEs and CIs shrink faster than they should, resulting in a false sense of confidence.

Autocorrelation contains information

Spatially referenced measurements contain valuable information (e.g., soil organic carbon on a mountain in Mexico)

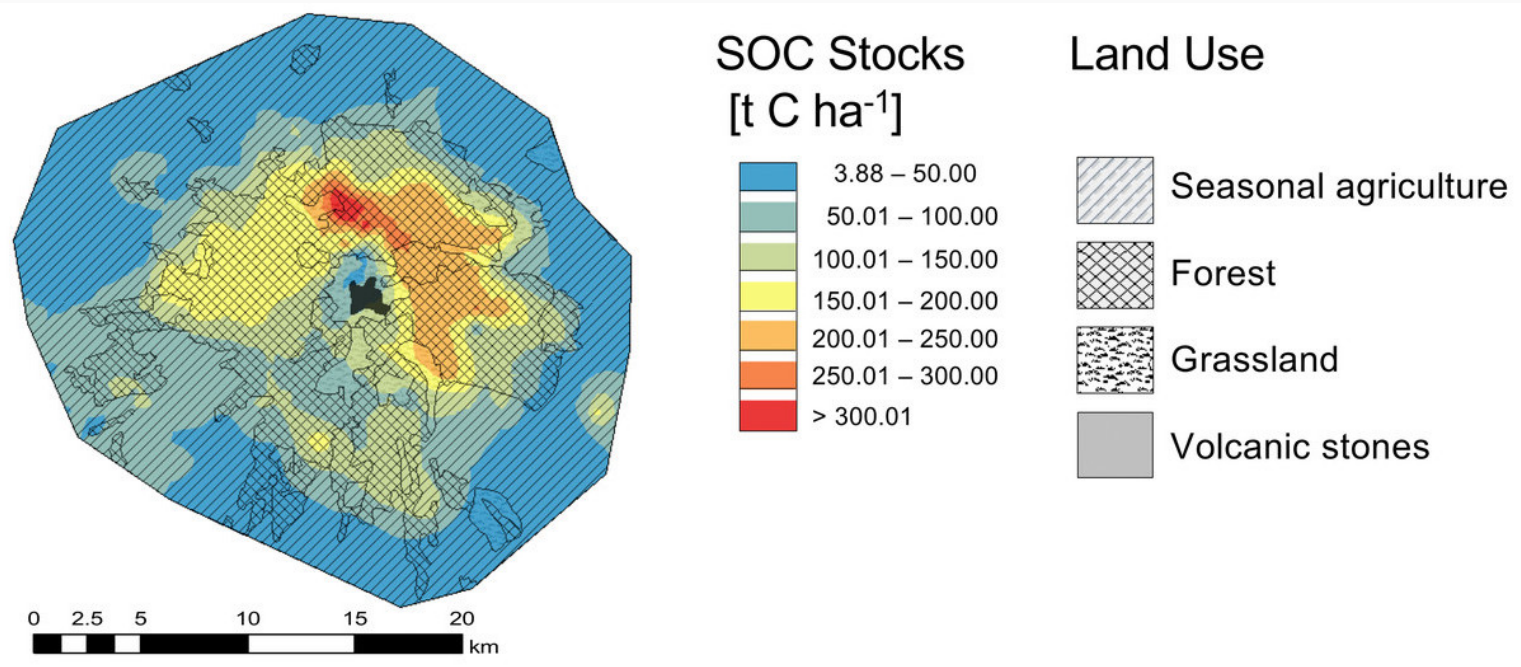


(Fusaro *et al.*, 2019)

... but what if we want to know how much carbon is at a nearby, un-sampled location?

Autocorrelation contains information

Leveraging the information contained the spatial autocorrelation between samples allows us to make predictions to unsampled areas



(Fusaro *et al.*, 2019)

Autocorrelation summary

Data that are collected across space will likely be autocorrelated, breaking the IID assumption.

Ignoring this autocorrelation results in biased, over-confident models, and (more importantly) does not leverage the full amount of information contained in the data.

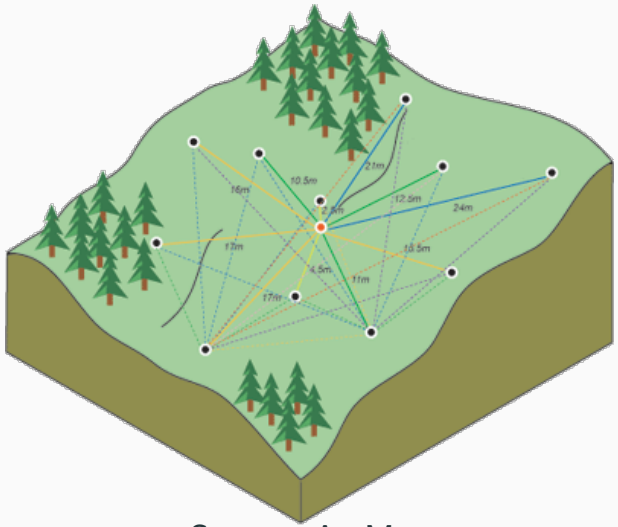
During the rest of this course we will focus on ways to visualise, detect, and work with the autocorrelation contained in spatial data.

Detecting Spatial Autocorrelation

Geostatistics

Scientists in many fields of research find themselves collecting data repeatedly over space, but the tools originally come from the field of geostatistics.

Can you think of data that are collected over space that might drive statistical innovation? Data where the ability to predict where things occur might be profitable?



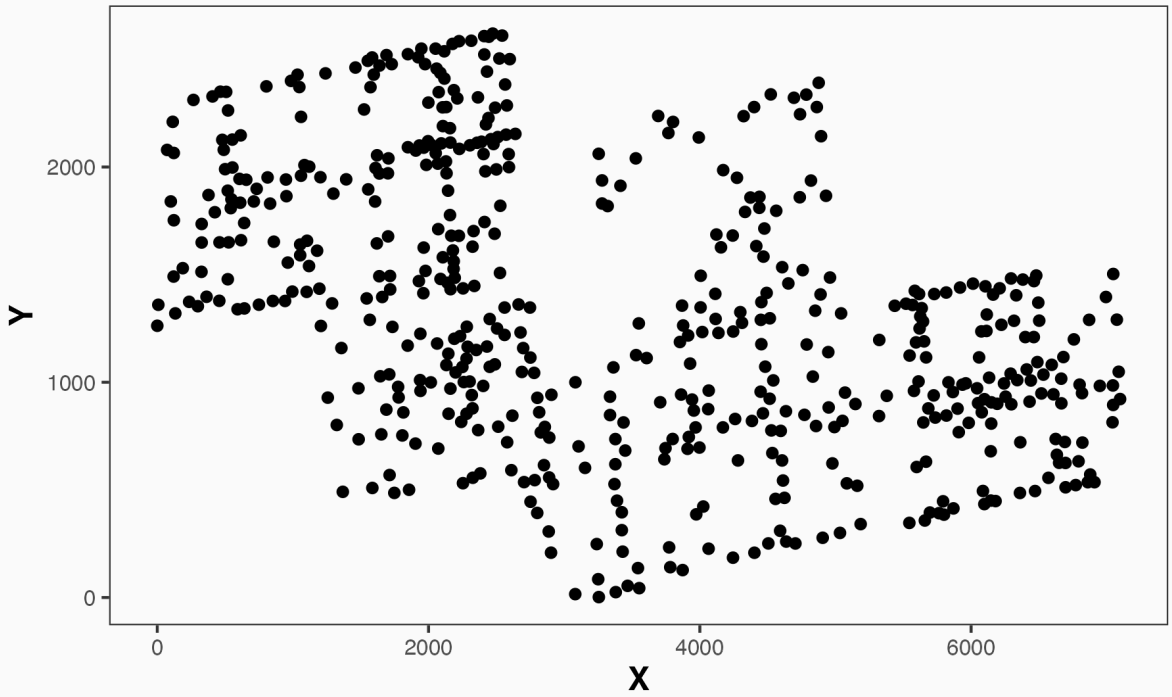
Source: ArcMap

The tools for working with spatial autocorrelation were developed for the goal of mapping mineral deposits.

The Russian Boreal Data

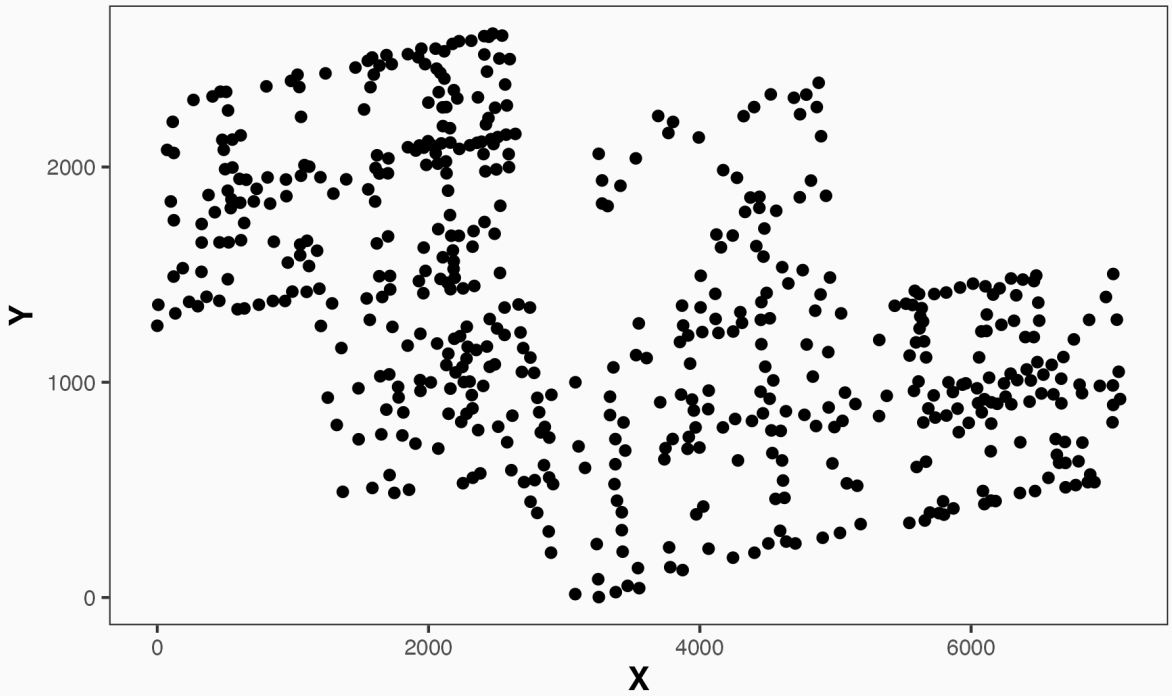
We're going to work with a dataset on forest composition in Tatarstan, Russia from Zuur *et al.* (2007).

The variable of interest is a measure of boreality (∼percent boreal species at a site).



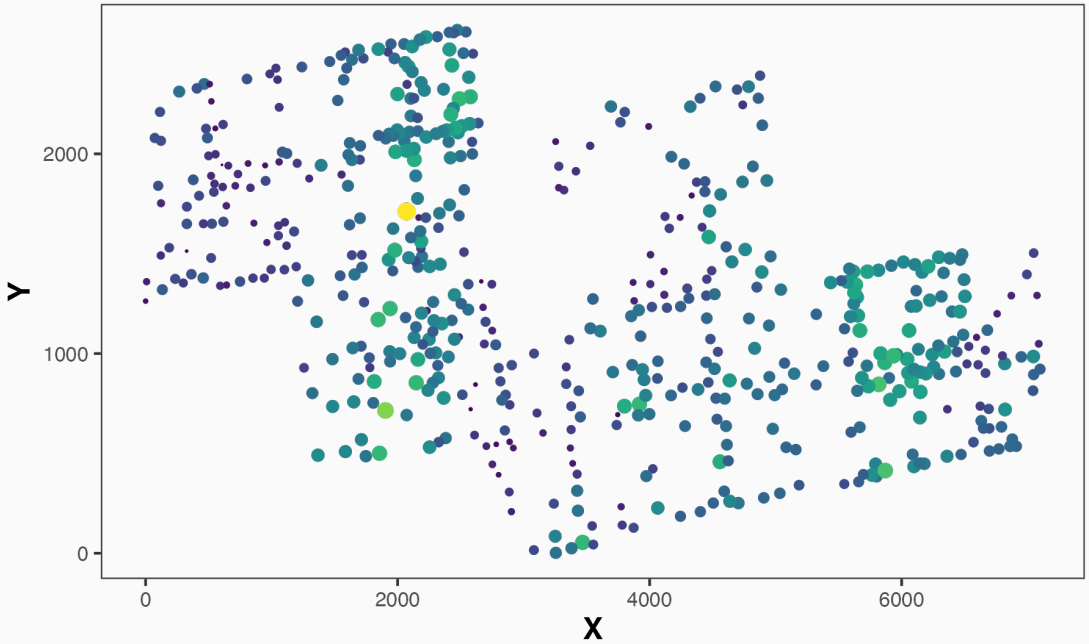
Bubble plots

Spatial autocorrelation can be difficult to see in a simple x vs. y scatterplot (not designed for this purpose).



Bubble plots

'Bubble plots' are an easy tool to quickly assess for autocorrelation.



Values are plotted in space, and sizes/colours are proportional to their values. The idea is to look for patterns.

Bubble plots are quick and easy to generate, but can be hard to read and are not particularly formal.

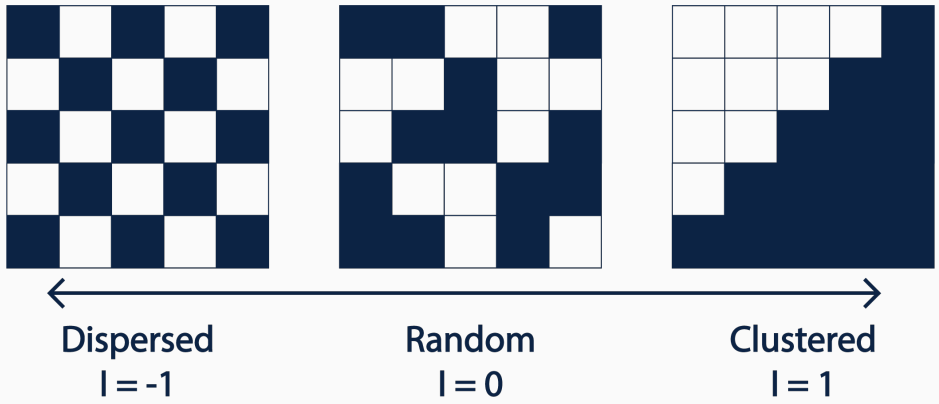
Moran's I

Moran's I is a correlation coefficient that measures the overall spatial autocorrelation of a data set (think of it as $\sim$ weighted covariance):

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

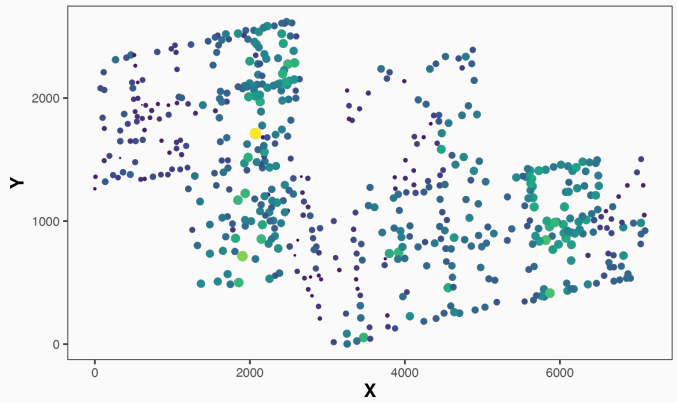
N is the number of spatial units indexed by i and j;
x is the variable of interest and $\bar{x}$ is the mean of x;
w_{ij} is a matrix of spatial weights and W is the sum of all w_{ij}.

Values of I usually range from -1 to +1.



Many R packages for calculating Moran's I

```
library(ape)
library(fields)

#Vector of spatial coordinates
coords = cbind(data$x, data$y)

#Matrix of distances for the weights
w = fields::rdist(coords)

#Calculate Moran's I
ape::Moran.I(data$Bor, w = w)

$observed
[1] -0.03019649

$expected
[1] -0.001879699

$sd
[1] 0.001368412

$p.value
[1] 3.991055e-95
```



The *p*-value tells us we have significant spatial autocorrelation.

Moran's I cont.

Moran's I can be a useful tool for identifying the presence of autocorrelation and is quite popular.

The challenge is how to act on this information (i.e., lets you know if you have autocorrelation, but doesn't help in modelling it)?

Moran's I cont.

Moran's I is also very sensitive to how you define the weights:

"The idea is to construct a matrix that accurately reflects your assumptions about the particular spatial phenomenon in question. A common approach is to give a weight of 1 if two zones are neighbors, and 0 otherwise, though the definition of 'neighbors' can vary. Another common approach might be to give a weight of 1 to k nearest neighbors, 0 otherwise. An alternative is to use a distance decay function for assigning weights. Sometimes the length of a shared edge is used for assigning different weights to neighbors. The selection of spatial weights matrix should be guided by theory about the phenomenon in question."

– Wikipedia

Semi-variograms

Semi-variograms are functions describing the degree of spatial dependence in a spatial stochastic process, $Z(\mathrm{s})$.

Semi-variance, $\gamma(h)$, is a measure of the degree of similarity between pairs of points separated by distance h, given by

$$\gamma(h) = \frac{1}{2V} \iint_V [f(M+h) - f(M)]^2 \, dV,$$

where M is a point in the spatial field V,
$f(M)$ is the value at point M (in arbitrary units);
h is the separation distance (in e.g., meters or km); and
the double integral is over 2 dimensions.

Semi-variograms properties

- $\gamma(h) \geq 0$, since it is the expectation of a square.
- Since $Z(s_1) - Z(s_1) = 0$, $\gamma(0)$ is always 0.
- If the process is stationary, in the limit where $h \to \infty$ $\gamma_s(h) = var(Z(s))$.
- If a stationary process has no spatial dependence, the semi-variogram is constant everywhere except at the origin, where it is zero.
- The semi-variogram might be discontinuous at the origin (the height of the jump at the origin is referred to as nugget).

For a discussion of these properties see Bachmaier & Backes (2011).

Empirical semi-variograms

To obtain the semi-variogram for a given $\gamma(h)$, all pairs of points at that exact distance, h, would need to be sampled. In practice, this is impossible, so the empirical semi-variogram is used instead.

For values of $Z(\mathrm{s})$ for all pairs separated by distance h, $\gamma(h)$ is estimated as:

$$\hat{\gamma}(h \pm \delta) := \frac{1}{2|N(h \pm \delta)|} \sum_{(i,j) \in N(h \pm \delta)} |z_i - z_j|^2$$

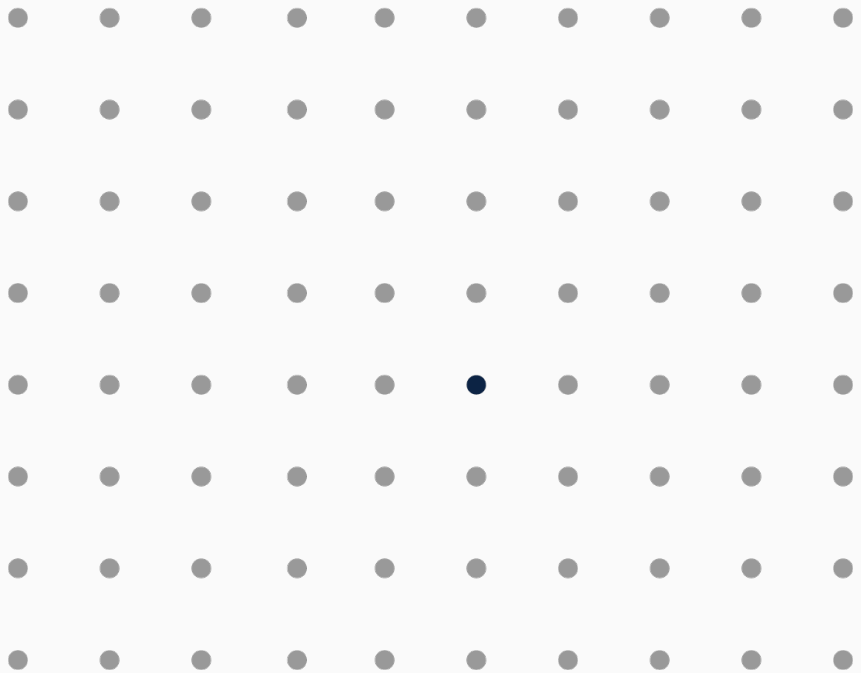
where δ is some bin width.

Empirical semi-variograms cont.

$$\hat{\gamma}(h \pm \delta) := \frac{1}{2|N(h \pm \delta)|} \sum_{(i,j) \in N(h \pm \delta)} |z_i - z_j|^2$$

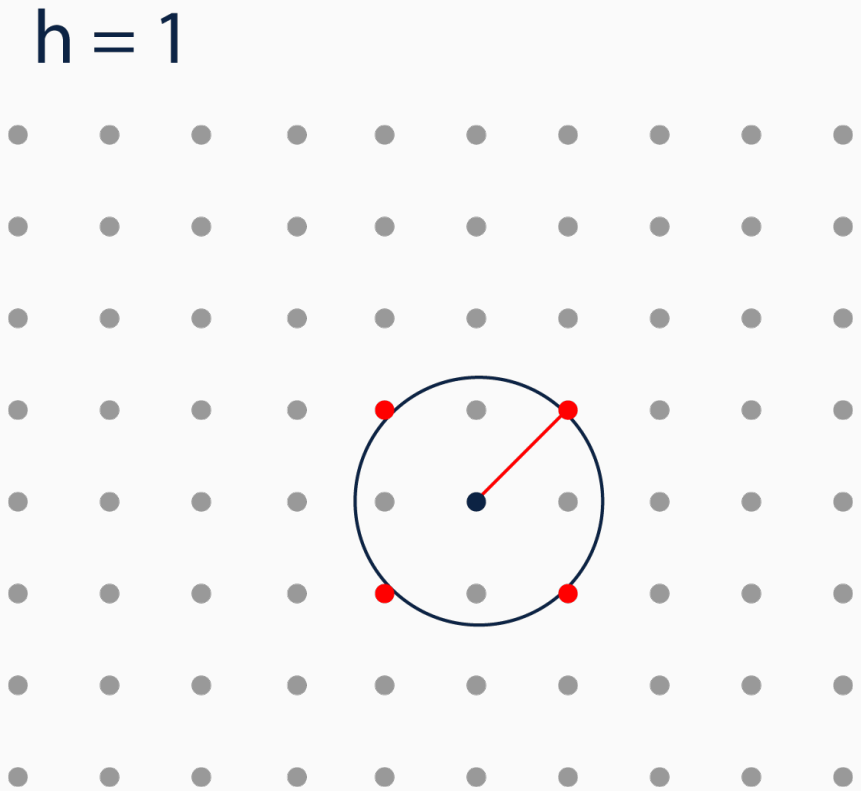
Empirical semi-variograms cont.

$$\hat{\gamma}(h \pm \delta) := \frac{1}{2|N(h \pm \delta)|} \sum_{(i,j) \in N(h \pm \delta)} |z_i - z_j|^2$$



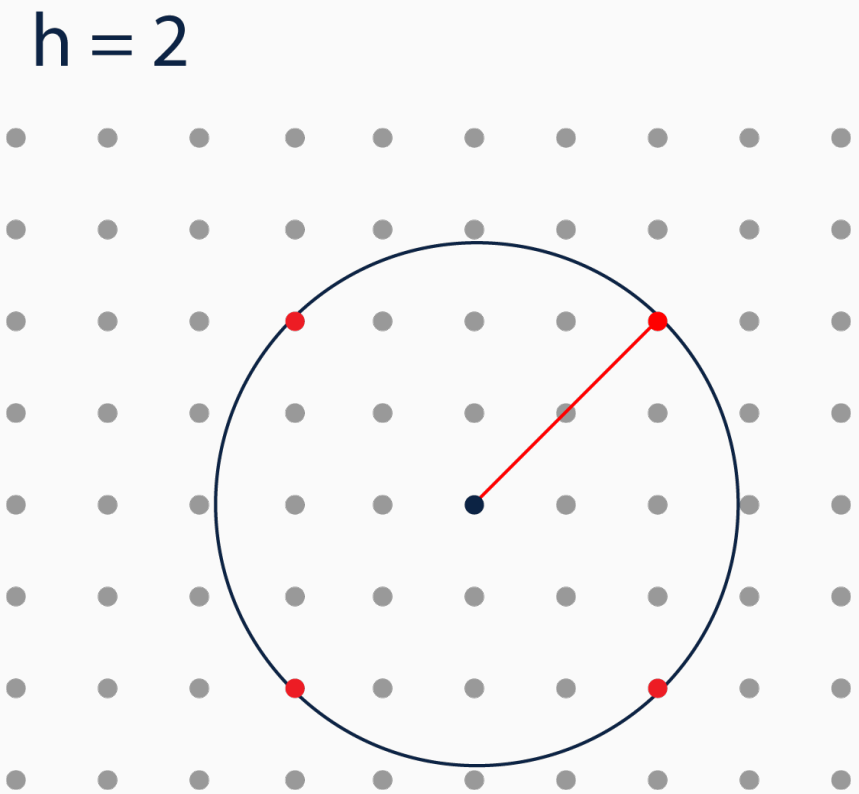
Empirical semi-variograms cont.

$$\hat{\gamma}(h \pm \delta) := \frac{1}{2|N(h \pm \delta)|} \sum_{(i,j) \in N(h \pm \delta)} |z_i - z_j|^2$$



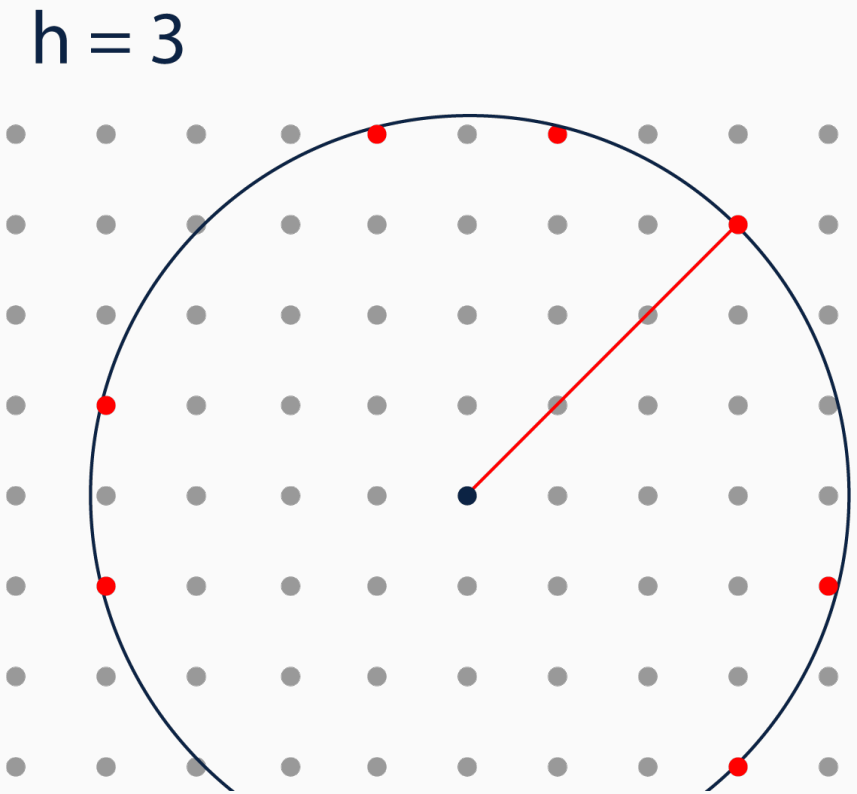
Empirical semi-variograms cont.

$$\hat{\gamma}(h \pm \delta) := \frac{1}{2|N(h \pm \delta)|} \sum_{(i,j) \in N(h \pm \delta)} |z_i - z_j|^2$$



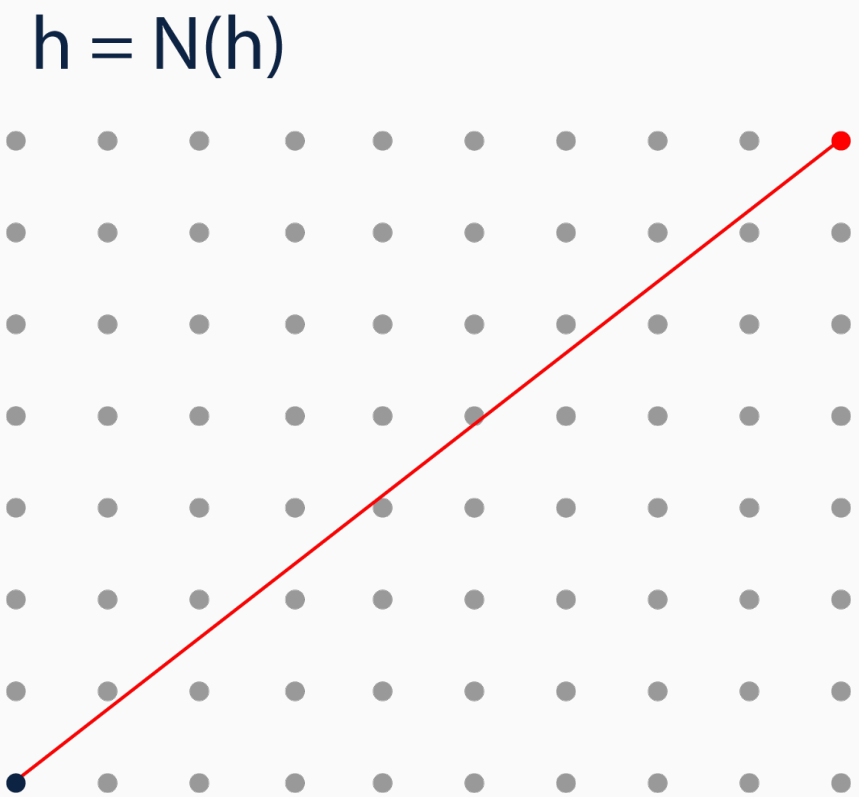
Empirical semi-variograms cont.

$$\hat{\gamma}(h \pm \delta) := \frac{1}{2|N(h \pm \delta)|} \sum_{(i,j) \in N(h \pm \delta)} |z_i - z_j|^2$$



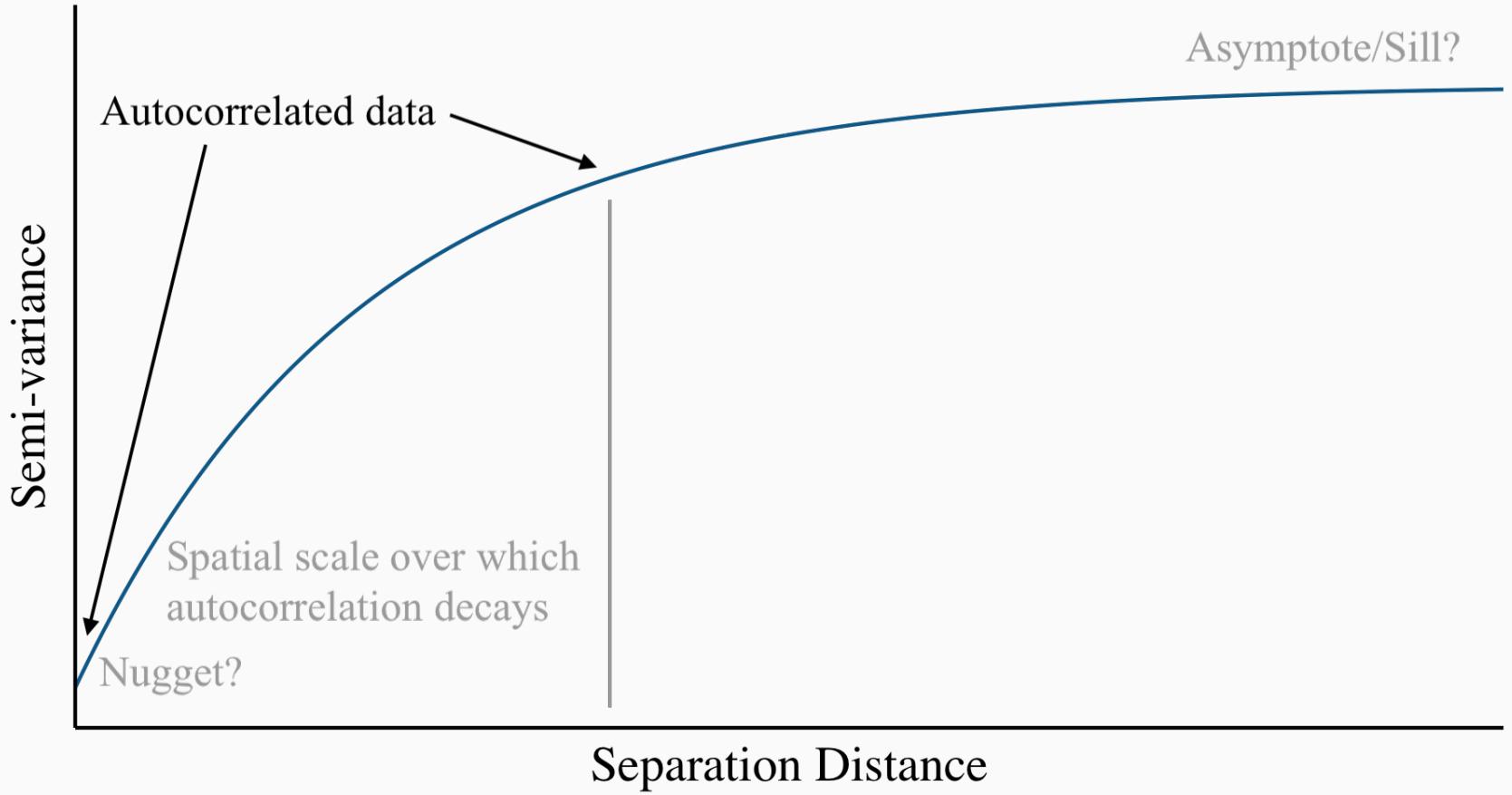
Empirical semi-variograms cont.

$$\hat{\gamma}(h \pm \delta) := \frac{1}{2|N(h \pm \delta)|} \sum_{(i,j) \in N(h \pm \delta)} |z_i - z_j|^2$$



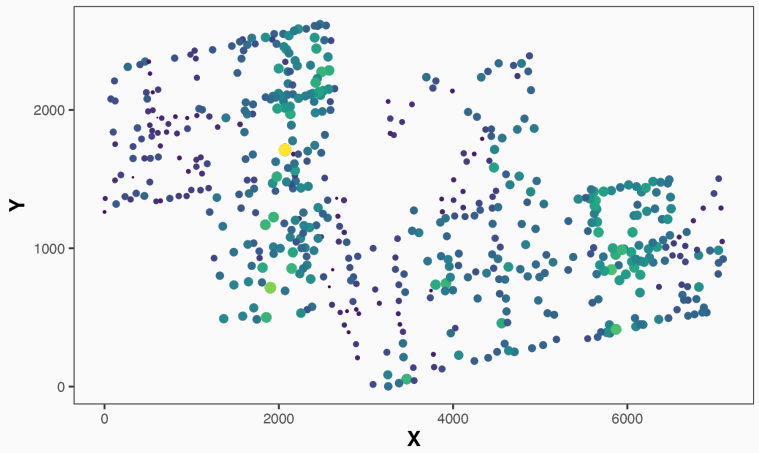
Reading a variogram

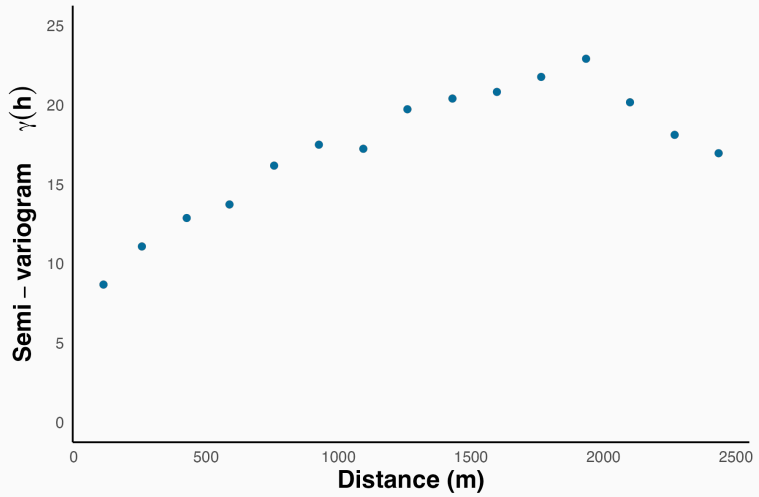
Plots of $\hat{\gamma}(h)$ vs. h are called a semi-variograms and facilitate the visual assessment of autocorrelations in spatial data.



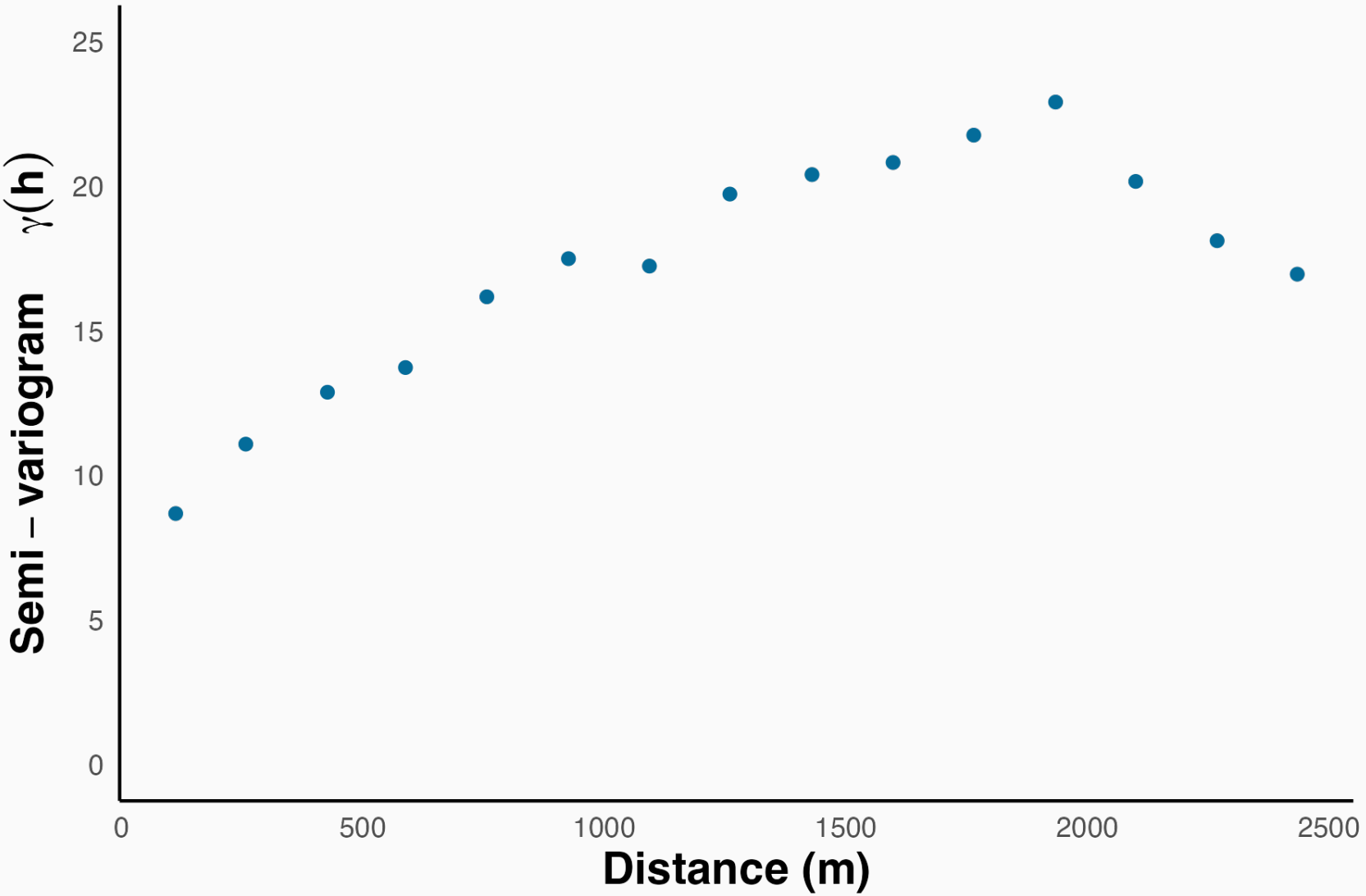
Autocorr. in the boreality data in R


```r
# Import the data
data <- read.csv("Datasets/Boreality.csv")

# Spatial data frame of boreality
DATA <- data.frame(Z_s = data$Bor,
                   x = data$x,
                   y = data$y)

# Define coordinates
sp::coordinates(DATA) <- c("x","y")

# object of class = "SpatialPointsDataFrame"

# Calculate empirical variogram
vg <- gstat::variogram(Z_s ~ 1, data = DATA)

# object of class = "gstatVariogram"

plot(vg)
```





Autocorr. in the boreality data in R



Take home messages

Autocorrelation contains a lot of information about a spatial process.

There are many methods for visualising and working with spatially correlated data (e.g., bubble plots, Moran's I).

Semi-variograms are useful tools for visualising spatial autocorrelation, are objective, and have a long, robust history.

Usefully, the shape of a dataset's empirical variogram can also provide clues on how to best model the autocorrelation in the data, which we will cover next lecture.

References

Bachmaier, M. & Backes, M. (2011). Variogram or semivariogram? variance or semivariance? allan variance or introducing a new term? *Mathematical Geosciences*, 43, 735–740.

Ballabio, C., Borrelli, P., Spinoni, J., Meusburger, K., Michaelides, S., Beguería, S., Klik, A., Petan, S., Janeček, M., Olsen, P., Aalto, J., Lakatos, M., Rymszewicz, A., Dumitrescu, A., Tadić, M.P., Diodato, N., Kostalova, J., Rousseva, S., Banasik, K., Alewell, C. & Panagos, P. (2017). Mapping monthly rainfall erosivity in europe. *Science of The Total Environment*, 579, 1298–1315.

Bell, T., Campbell, S., Liverman, D.G., Allison, D. & Sylvester, P. (2010). Environmental and potential human health legacies of non-industrial sources of lead in a canadian urban landscape – the case study of st john's, newfoundland. *International Geology Review*, 52, 771–800.

Feeney, C., Cosby, B., Robinson, D., Thomas, A., Emmett, B. & Henrys, P. (2022). Multiple soil map comparison highlights challenges for predicting topsoil organic carbon concentration at national scale. *Scientific reports*, 12, 1–13.

Fusaro, C., Sarria-Guzmán, Y., Chávez-Romero, Y.A., Luna-Guido, M., Muñoz-Arenas, L.C., Dendooven, L., Estrada-Torres, A. & Navarro-Noya, Y.E. (2019). Land use is the main driver of soil organic carbon spatial distribution in a high mountain ecosystem. *PeerJ*, 7, e7897.

Zuur, A., Ieno, E.N. & Smith, G.M. (2007). *Analyzing ecological data*. Springer.