Spatial autocorrelation 2: Kriging

Michael Noonan

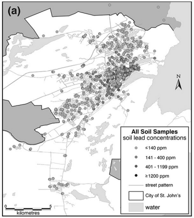
DATA 589: Spatial Statistics

Table of contents

Review

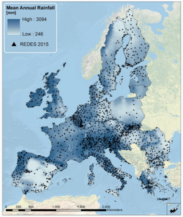
Review

Lead in St. John's



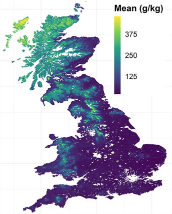
(Bell *et al.*, 2010)

Rainfall in Europe



(Ballabio *et al.*, 2017)

UK Soil Carbon



(Feeney *et al.*, 2022)

Last lecture we started covering situations where the locations of points were an arbitrary artefact of the sampling process and not the variable of interest.

We also covered how autocorrelation contains a lot of information about these spatial processes, but that working with them requires a special set of tools.

We covered some of these tools (e.g., bubble plots, Moran's I), but identified semi-variograms as being particularly useful, objective, and as having a long, proven history.

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

I also told you that the shape of a dataset's empirical semi-variogram can provide clues on how to best model the autocorrelation in the data.

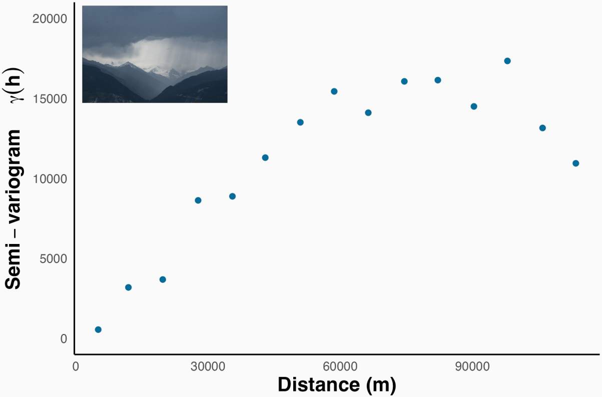
Today we will focus on how to fit models to semi-variograms, how to use these models to make predictions, and the implications for study design.

Modelling Correlation Structures

Reading variograms

Semi-variogram have a number of key features that we should be looking for (sill? range? nugget? shape?).

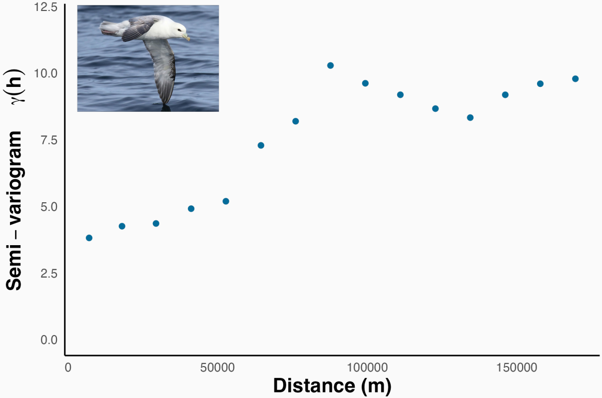
Rainfall in Switzerland



Source: gstat package

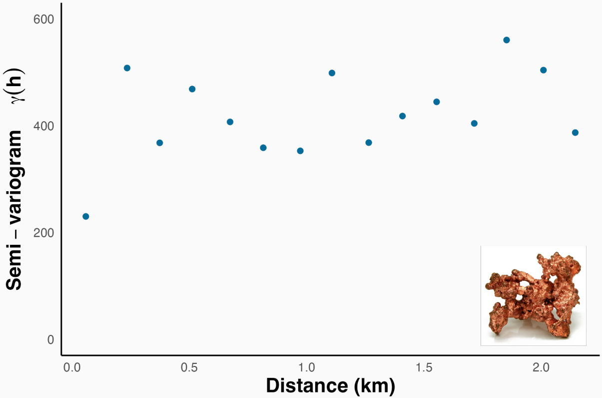
Reading variograms cont.

Fulmaris glacialis densities



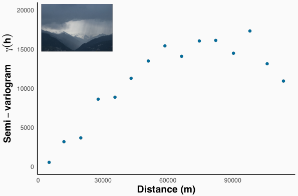
Source: gstat package

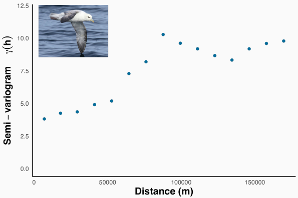
Soil copper concentrations

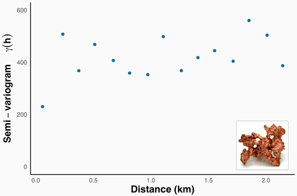


Source: gstat package

Reading variograms cont.







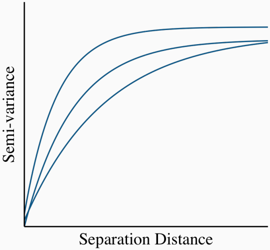
Semi-variograms have a number of key features that we should be looking for (i.e., sill, range, nugget, shape).

So what?

Usefully, the different spatial correlation models all have differently shaped theoretical variograms.

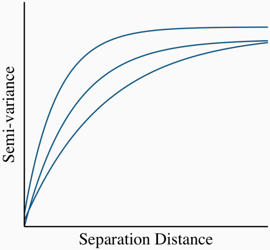
Corr. models and their variograms

Exponential $\Phi = 1 - e^{-\frac{D}{\rho}}$

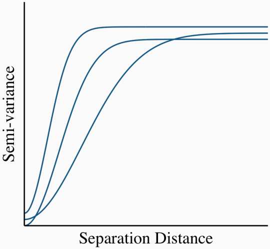


Corr. models and their variograms

Exponential $\Phi = 1 - e^{-\frac{D}{\rho}}$

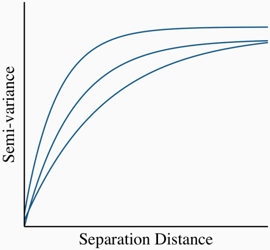


Gaussian $\Phi = 1 - e^{-(\frac{D}{\rho})^2}$

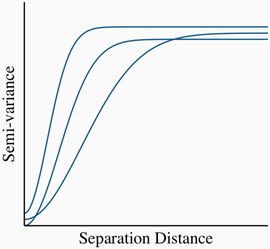


Corr. models and their variograms

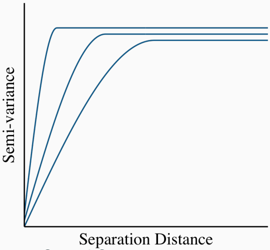
Exponential $\Phi = 1 - e^{-\frac{D}{\rho}}$



Gaussian $\Phi = 1 - e^{-(\frac{D}{\rho})^2}$

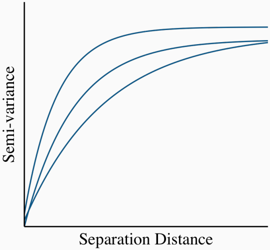


Spherical $\Phi = 1(1 - 1.5(\frac{d}{\rho}) + 0.5(\frac{d}{\rho})3)I(d < \rho)$

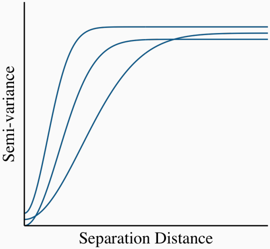


Corr. models and their variograms

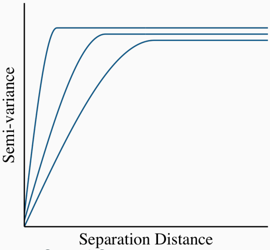
Exponential $\Phi = 1 - e^{-\frac{D}{\rho}}$



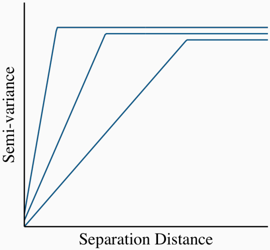
Gaussian $\Phi = 1 - e^{-(\frac{D}{\rho})^2}$

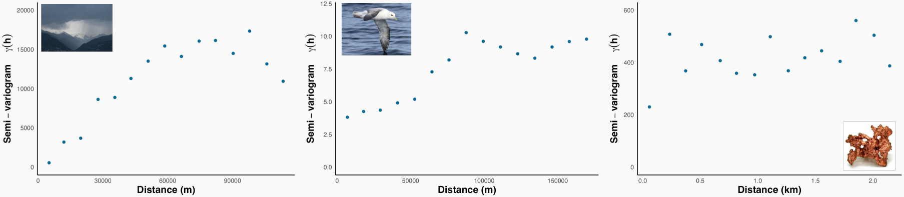


Spherical $\Phi = 1(1 - 1.5(\frac{d}{\rho}) + 0.5(\frac{d}{\rho})3)I(d < \rho)$



Linear $\Phi = 1 - (1\frac{D}{\rho})I(d < \rho)$



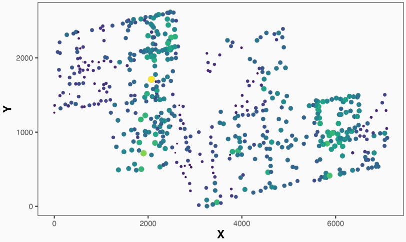


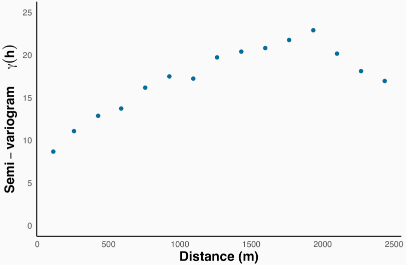
The different correlation models are fit to the semi-variogram, usually (but not necessarily) via Ordinary Least Squares, and the best fit to the data is identified.

Fitting correlation models in R

We're going to work with the dataset on forest composition in Tatarstan, Russia again.

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





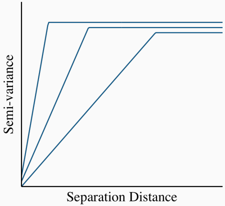
Linear spatial correlation in R

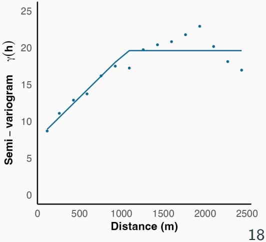
A linear spatial correlation structure can be applied via the `fit.variogram()` function with the argument `vgm("Lin")`.

```
#Data import and wrangling
data <- read.csv("Datasets/Boreality.csv")
sp::coordinates(data) <- c("x","y")

# Empirical variogram
vg <- gstat::variogram(Bor ~ 1, data = data)

#Fit linear correlation model
fit.linear <- fit.variogram(vg, vgm("Lin"))
```

```
fit.linear

  model     psill    range
1   Nug  7.649557    0.000
2   Lin 11.954061 1066.725
```

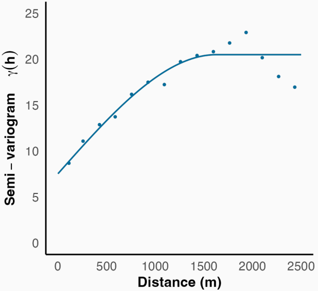
Linear $\Phi = 1 - (1\frac{D}{\rho})I(d < \rho)$



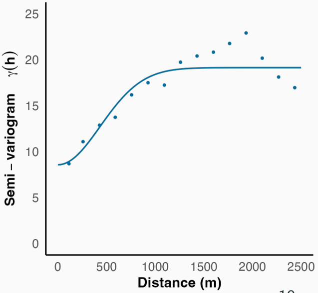


Other spatial correlations in R


```
#Fit spherical correlation model
fit.linear <- fit.variogram(vg, vgm("Sph"))

fit.spherical

  model     psill    range
1   Nug  7.539269    0.000
2   Sph 12.948714 1627.959
```

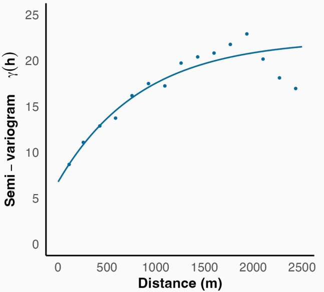



```
#Fit Gaussian correlation model
fit.Gaussian <- fit.variogram(vg, vgm("Gau"))

fit.Gaussian

  model     psill   range
1   Nug  8.566378   0.000
2   Gau 10.563320 611.773
```

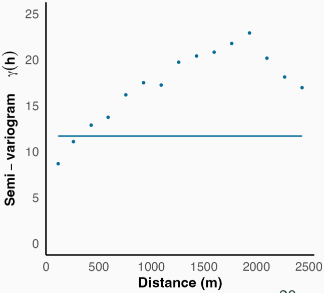


Other spatial correlations in R cont.


```
#Fit exponential correlation model
fit.exponential <- fit.variogram(vg, vgm("Exp"))

fit.exponential

  model     psill    range
1   Nug  6.783153   0.0000
2   Exp 15.545091 850.8605
```




```
#Fit nugget only model
fit.nugget <- fit.variogram(vg, vgm("Nug"))

fit.nugget

  model    psill range
1   Nug 11.68799     0
```

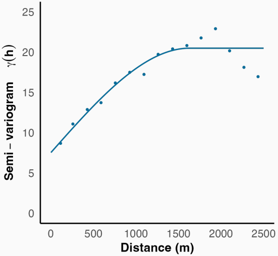


Selecting the best structure

We just fit 5 different autocorrelation models, but how do we know which is the best fit to the data?

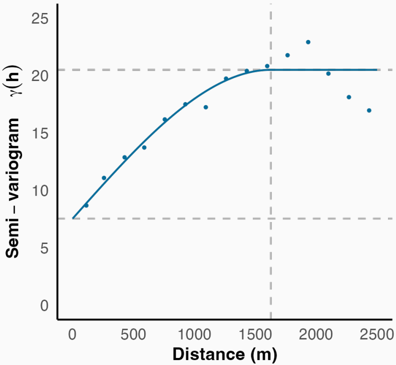
```
# Extract sum of squared errors
results <- data.frame(model = c("spherical", "linear", "Gaussian",
                                "exponential", "nugget"),
                      SSErr = c(attr(fit.spherical, "SSErr"),
                                attr(fit.linear, "SSErr"),
                                attr(fit.Gaussian, "SSErr"),
                                attr(fit.exponential, "SSErr"),
                                attr(fit.nugget, "SSErr")))

#Ordered by lowest to highest SSErr
results <- results[order(results$SSErr),]
```

```
        model      SSErr
1   spherical 0.06565859
4 exponential 0.06942162
2      linear 0.10490507
3    Gaussian 0.14084834
5      nugget 2.95142242
```



What does the selected model tell us about our data?

```
fit.spherical

  model     psill    range
1   Nug  7.539269    0.000
2   Sph 12.948714 1627.959
```



Correlations persist for ∼1.6 km.

When $h \to \infty$ $\gamma_s(h) = var(Z(\mathbf{s}))$

```
fit.spherical$psill[2] + fit.spherical$
     psill[1]
[1] 20.48798
```

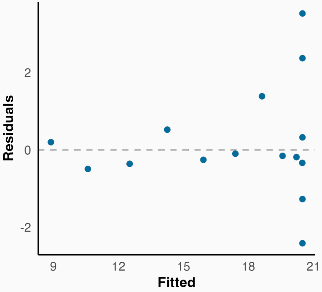
```
var(data$Bor)
[1] 17.76566
```

Variance is slightly different (non-stationarity? small-sample-size-bias? model-misspecification?).

Selected model cont.

Residuals can be manually calculated.

```
#Get fitted values
fitted <- variogramLine(fit.spherical,
                        maxdist=max(vg$dist),
                        dist_vector=vg$dist)

#Calculate residuals
residuals <- fitted$gamma - vg$gamma

#Visualise the residuals
plot(residuals ~ fitted$gamma)
```

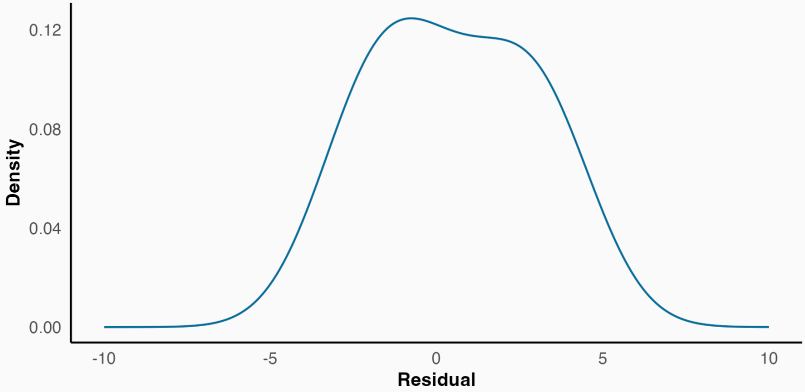


How do these look? What should they look like?

Selected model cont.

Density plot of residuals around the sill can be informative.

```
SILL <- residuals[which(fitted$gamma == max(fitted$gamma))]

plot(density(SILL))
```



Overview of correlation structures

We fit several spatial autocorrelation models:

Type	Description	gstat **Function**
Nugget	0	`vgm("Nug")`
Linear	$\Phi = 1 - (1\frac{D}{\rho})I(d < \rho)$	`vgm("Lin")`
Spherical	$\Phi = 1(1 - 1.5(\frac{d}{\rho}) + 0.5(\frac{d}{\rho})3)I(d < \rho)$	`vgm("Sph")`
Gaussian	$\Phi = 1 - e^{-(\frac{D}{\rho})^2}$	`vgm("Gaus")`
Exponential	$\Phi = 1 - e^{-\frac{D}{\rho}}$	`vgm("Exp")`

The model structures can be difficult to interpret, but their variograms have very recognizable features. Familiarising yourself with them will help you quickly narrow down what structure to use.

...but there are a lot of different candidate models to choose from.

```
gstat::vgm()

   short                                          long
1    Nug                                 Nug (nugget)
2    Exp                            Exp (exponential)
3    Sph                              Sph (spherical)
4    Gau                               Gau (gaussian)
5    Exc        Exclass (Exponential class/stable)
6    Mat                                 Mat (Matern)
7    Ste Mat (Matern, M. Steins parameterization)
8    Cir                               Cir (circular)
9    Lin                                 Lin (linear)
10   Bes                                 Bes (bessel)
11   Pen                        Pen (pentaspherical)
12   Per                               Per (periodic)
13   Wav                                   Wav (wave)
14   Hol                                   Hol (hole)
15   Log                            Log (logarithmic)
16   Pow                                  Pow (power)
17   Spl                                 Spl (spline)
18   Leg                               Leg (Legendre)
19   Err                      Err (Measurement error)
20   Int                              Int (Intercept)
```

Predicting from Spatial Autocorrelation Models

The form of the correlation model and parameter values are valuable in-and-of-themselves, but fitting these models is usually an intermediate step.

Typically, the goal of modelling these data is to predict to unsampled areas and map out the response variable.

There are many tools for interpolating spatial data, but we will focus on one of them: Kriging (based Danie Krige's MSc thesis).

There are also many forms of Kriging (ordinary, simple, universal, Bayesian, etc...), but we will focus (mostly) on ordinary Kriging.

Ordinary Kriging

In ordinary Kriging, $\hat{Z}(x_0)$ is assumed to be random variable located at an unobserved location x_0, with a constant, unknown mean (Matheron, 1963).

$\hat{Z}(x_0)$ is estimated from a linear combination of the observed values z_i and weights w_i:

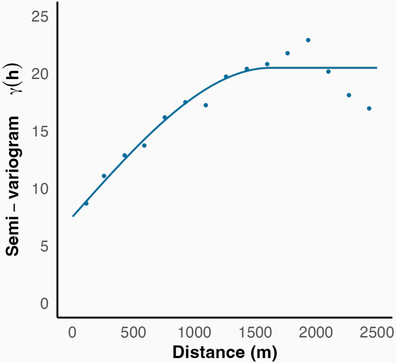
$$\hat{Z}(x_0) = \begin{bmatrix} w_1 & w_2 & \cdots & w_N \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_N \end{bmatrix} = \sum_{i=1}^{N} w_i(x_0) Z(x_i)$$

The weights are critical, and intended to reflect the proximity of samples to the estimation location x_0.

Ordinary Kriging cont.

THE UNIVERSITY OF BRITISH COLUMBIA
Okanagan Campus

We're trying to predict $\hat{Z}(x_0)$ using the known values $Z(x_i)$, and their spatial dependences.



The fitted semi-variogram model describes the spatial dependence of the samples.

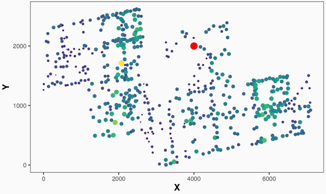
We can use this to calculate the covariance matrix (diagonal $= \sigma^2$ $=$ sill, off-diagonals $= \hat{\gamma}(h)$)

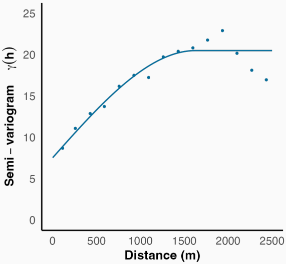
and from that the weights (with $\sum_{i=1}^{N} w_i = 1$).

Ordinary Kriging in R

We're trying to predict $\hat{Z}(x_0)$ using $Z(x_i)$, and $\hat{\gamma}(h)$.

```
# Location to predict at
x_0 <- data.frame(x = 4000,
                  y = 2000)
sp::coordinates(x_0) <- c("x","y")


# Kriged estimate
gstat::krige(Bor ~ 1,
             data,
             model=fit.spherical,
             newdata = x_0)


[using ordinary kriging]
   coordinates var1.pred var1.var
1 (4000, 2000)  12.31154 21.43853
```

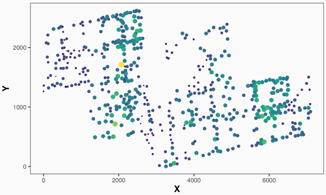


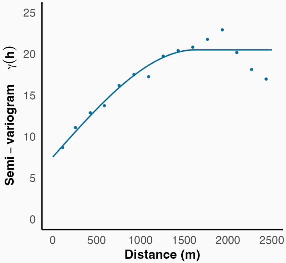


Ordinary Kriging in R cont.

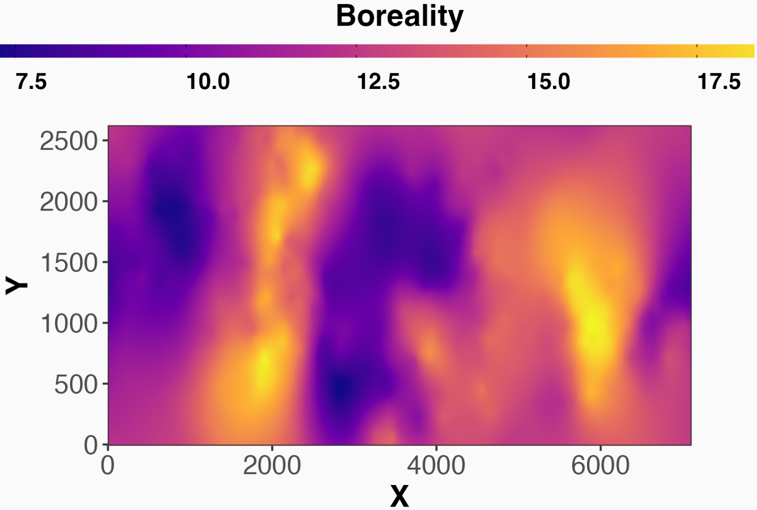
Usually we want to predict over a large spatial area.

```
# Grid over the sampled area
grid <- makegrid(data, n=200000)
names(grid) <- c("x", "y")
sp::coordinates(grid) <- c("x","y")

boreality.kriged <- krige(Bor ~ 1,
                          data,
                          newdata = grid,
                          model=fit.spherical)

head(boreality.kriged)

    coordinates var1.pred var1.var
1 (-2.89, 4.93)  12.21230 21.33647
2  (6.71, 4.93)  12.20009 21.32484
3 (16.31, 4.93)  12.18724 21.31255
4 (25.91, 4.93)  12.17376 21.29957
5 (35.51, 4.93)  12.15965 21.28588
6 (45.11, 4.93)  12.14492 21.27143
```

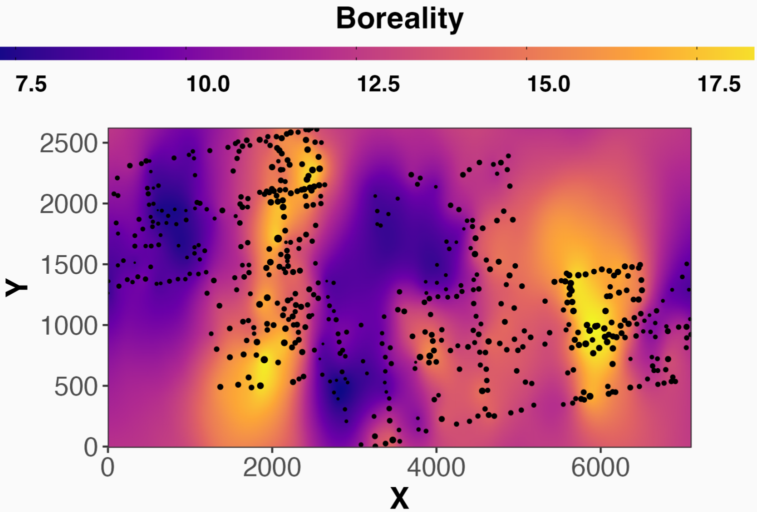




Kriged boreality map

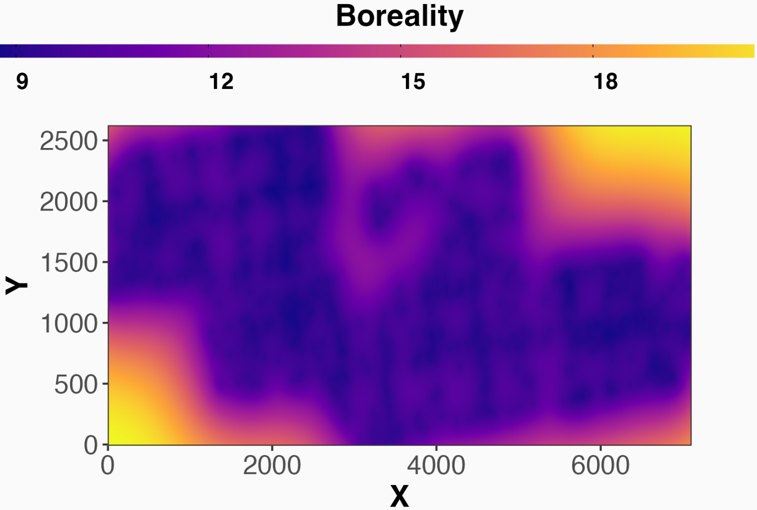


Kriged boreality map



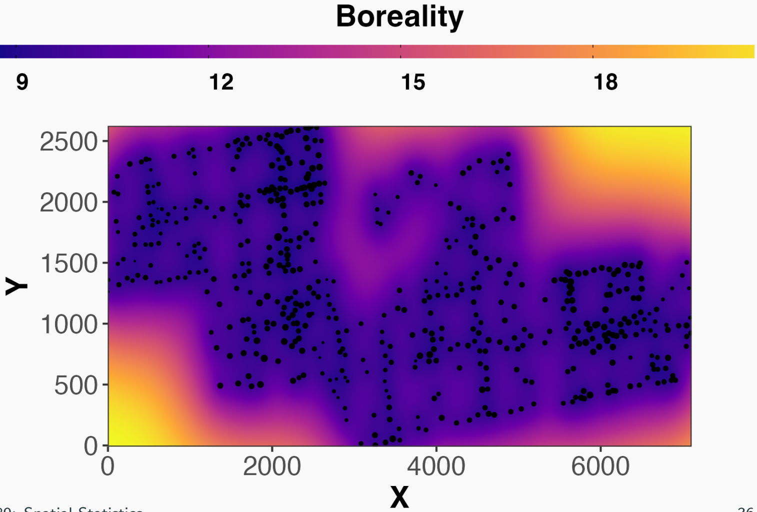
Kriged boreality variance map

We also get an estimate of the variance at x_0. Do these patterns make sense?

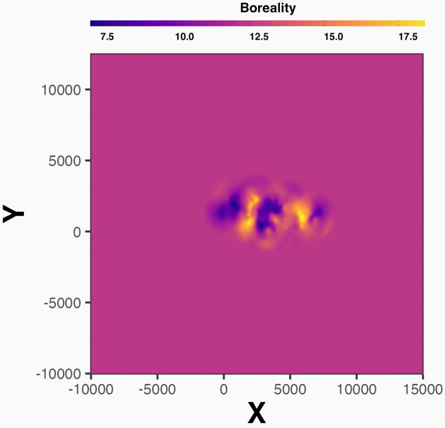


Kriged boreality variance map

Variance is lowest where we have data ($\sigma^2_{x_0}$ = nugget), and increases the further away from the samples we move.



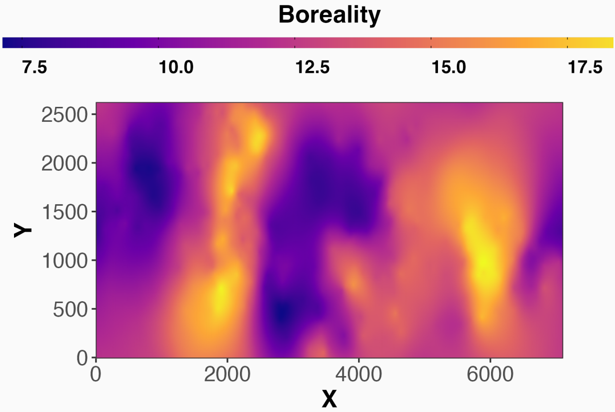
Kriging is a spatial interpolation method, so what happens if we try to extrapolate?



How far out can we reasonably predict?

Technical considerations

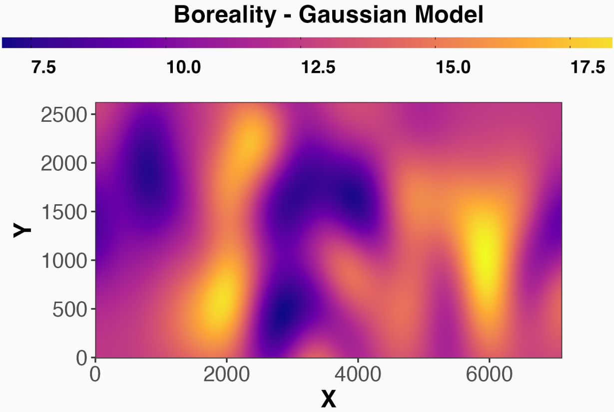
The Kriging weights (and therefore the predictions) are very sensitive to the fitted semi-variogram.



It's important to ensure the model is correctly specified.

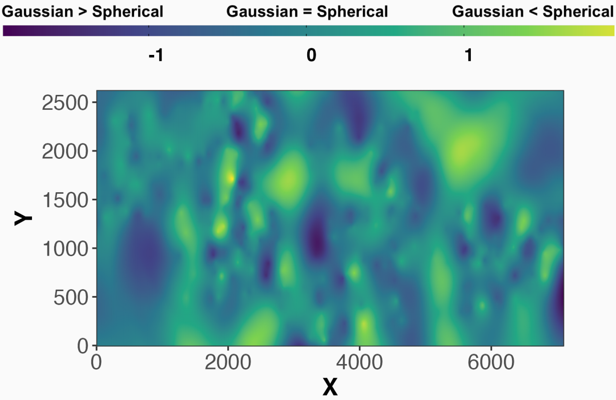
Technical considerations

The Kriging weights (and therefore the predictions) are very sensitive to the fitted semi-variogram.



It's important to ensure the model is correctly specified.

The Kriging weights (and therefore the predictions) are very sensitive to the fitted semi-variogram.



It's important to ensure the model is correctly specified.

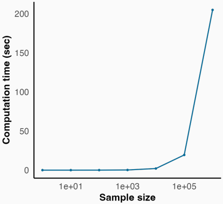
Technical considerations cont.

Est. the weights requires a matrix inversion (doesn't scale well).

```
# predict at 100 locations
grid100 <- makegrid(data, n=100)
sp::coordinates(grid1) <- c("x1","x2")

system.time(
  krige(Bor ~ 1,
        data,
        newdata = grid100,
        model=fit.spherical))

[using ordinary kriging]
   user  system elapsed
  0.049   0.001   0.050
```

```
# predict at 10000 locations
grid10000 <- makegrid(data, n=10000)
sp::coordinates(grid10000) <- c("x1","x2")

system.time(
  krige(Bor ~ 1,
        data,
        newdata = grid10000,
        model=fit.spherical))

[using ordinary kriging]
   user  system elapsed
  2.066   0.024   2.096
```

4,000 times longer!

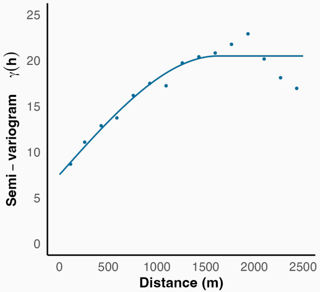


Considerations for Sampling Designs

Spatially Autocorrelated Data

Experimental designs that do not consider spatial autocorrelation risk being over/under-sampled.

Corrections exist to deal with issues of statistical bias, but they can't inject more information into a dataset when none exists.



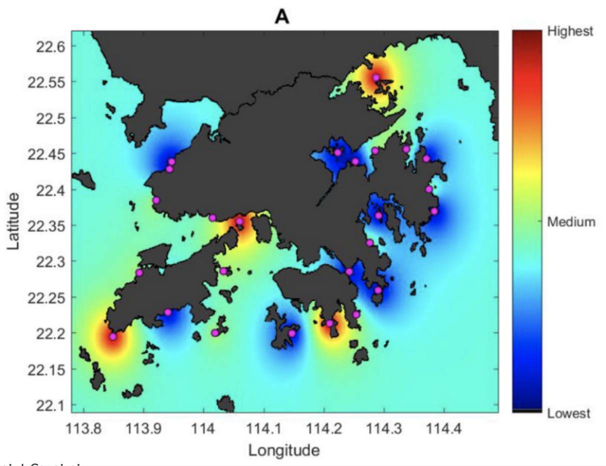
Good study design should consider spatial autocorrelation *a priori*.

If you had to collect more data for the boreality study how far apart would you sample? $\lesssim 1600m$ to see the autocorrelation, $\gtrsim 1600m$ for IID data or for the mean/sill.

Applied Kriging Analysis

Mapping plastic pollution

Coleby & Grist (2019) used a semi-variograms and kriging to map the distribution of marine plastic pollution in Hong Kong.



Take home messages

Fitting semi-variograms to spatial data can leverage the information contained in the autocorrelation structure and tell us a lot about the processes.

Kriging is a valuable tool for interpolating from spatially referenced data, but is not without limitations.

Kriging leverages information contained in the autocorrelation structure, but what about information contained in covariates?

Next lecture we will cover Kriging with covariates.

References

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.

Coleby, A.M. & Grist, E.P. (2019). Prioritized area mapping for multiple stakeholders through geospatial modelling: A focus on marine plastics pollution in hong kong. *Ocean & Coastal Management*, 171, 131–141.

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.

Matheron, G. (1963). Principles of geostatistics. *Economic geology*, 58, 1246–1266.