Geostatistics

Point-referenced data models

Data of Person + Time + Space

Audrey Yeo & Sascha Stutz
Master of Science UZH in Biostatistics

game plan

Motivation !

Assumptions

Elements

Fun Part : Prediction and Application

Serious Part : Quiz

Motivation

Geography
link between natural processes and spatial structures
assumes that two points that are closer are more similar to each other

Public Health
mapping of disease

Economic policy and program
allocation of the right resources at the right time and space

Motivation

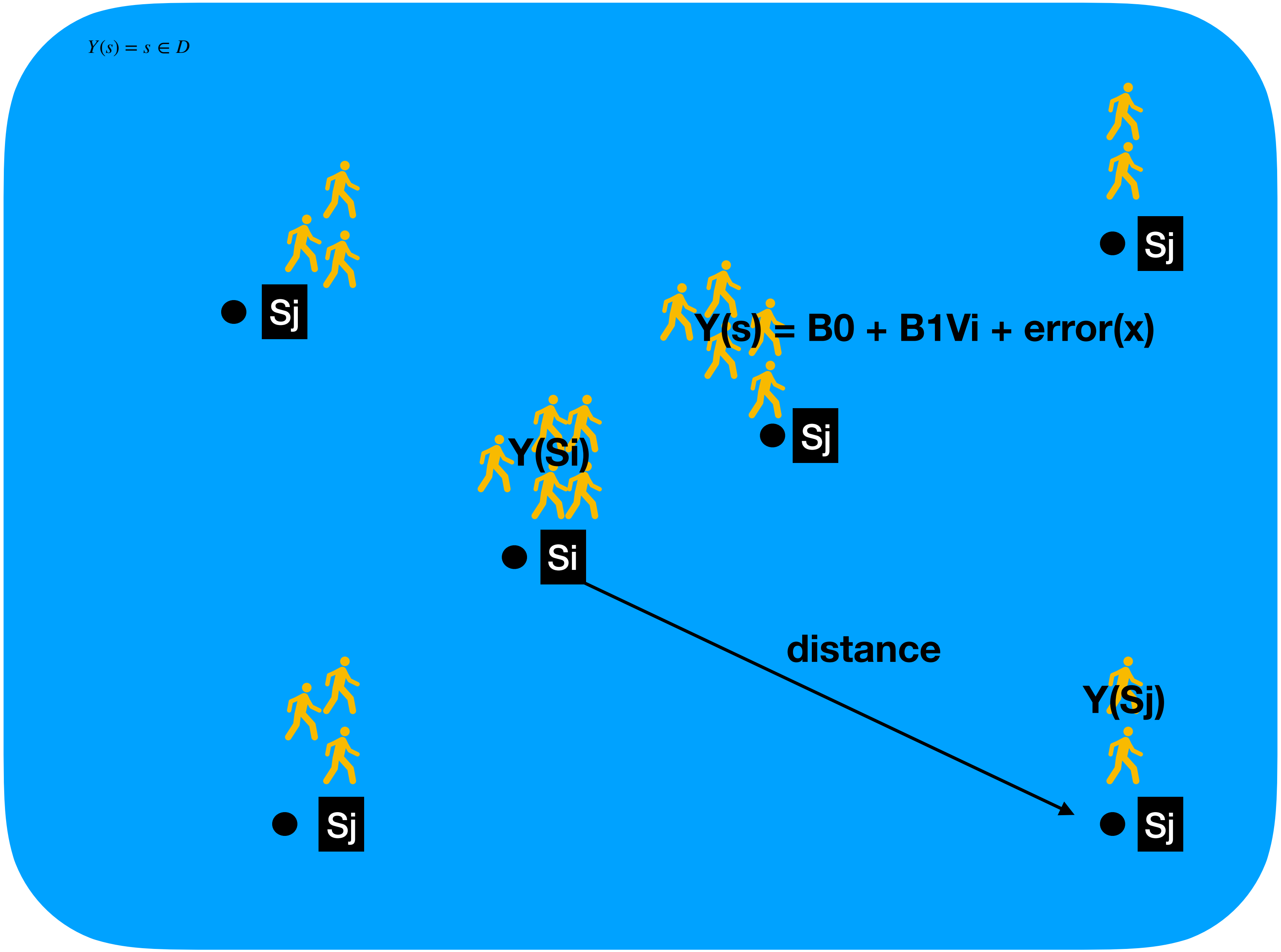
Epidemiology of esophageal cancer: Orient to Occident. Effects of chronology, geography and ethnicity (Honga et al, 2009)

Simple model: An application example

Incidence of Oesophageal Cancer	
Asia	0.4
Europe	4.2 - 7.0
Americas	3.2

Adapted from Honga et al's (2009) study

Space we're interested in which contains points at Si, Sj.
Think of the S as addresses and Y is the interested result
e.g. "prevalence of malaria" at point Si, denoted by Y(Si).



Geostatistics Assumptions

two points are more similar the closer they are

the nature of two points exist in a random process with dependence

Stationarity

- **Strictly stationarity**
  $(Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)) \sim N()$
  $(Y(\mathbf{s}_1+\mathbf{h}), \ldots, Y(\mathbf{s}_n+\mathbf{h})) \sim N()$

- **Weak stationarity (mean stationarity)**
  $\mu(\mathbf{s}) \equiv \mu$
  $\mathrm{Cov}(Y(\mathbf{s}), Y(\mathbf{s}+\mathbf{h})) = C(\mathbf{h})$

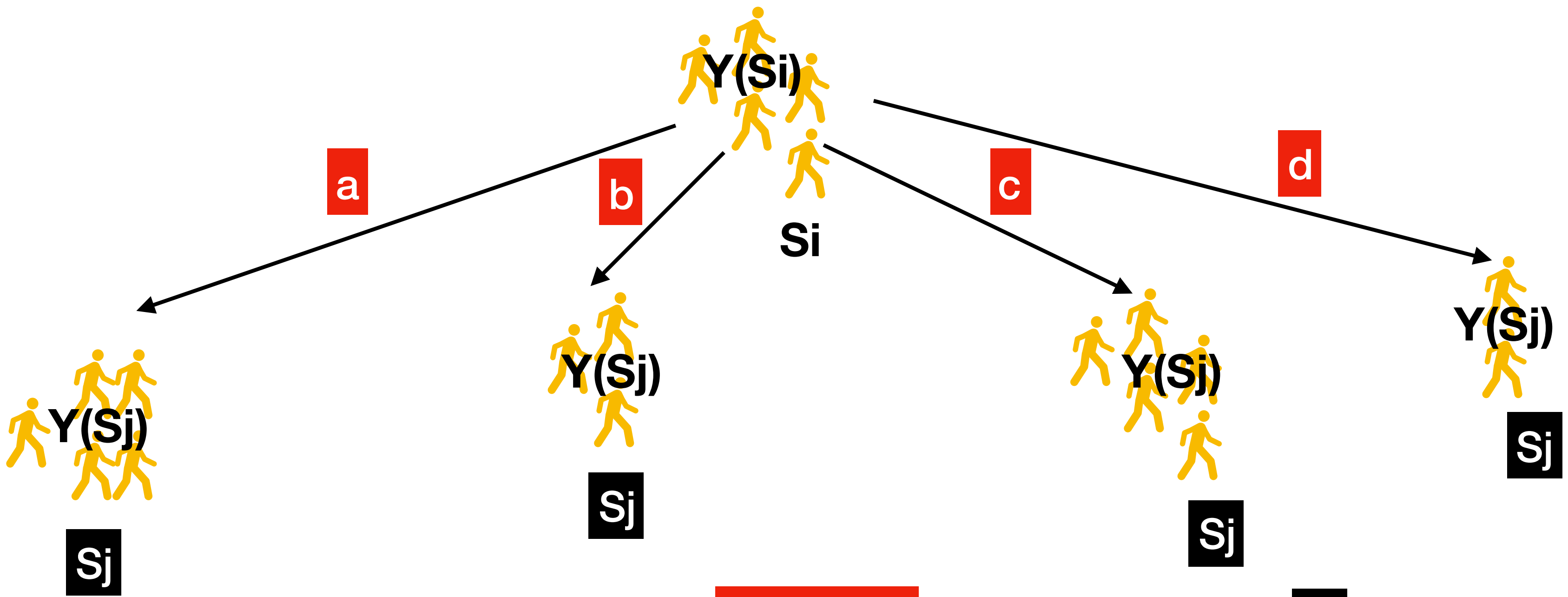
- **how do we measure this dissimilarity ?**

Semi-variogram

definition - a functional relationship between variance of the nature of two points (e.g. incidence rates) on the y-axis and distance (between two points squared) on the x-axis.

Semi-variogram

How do we do this?

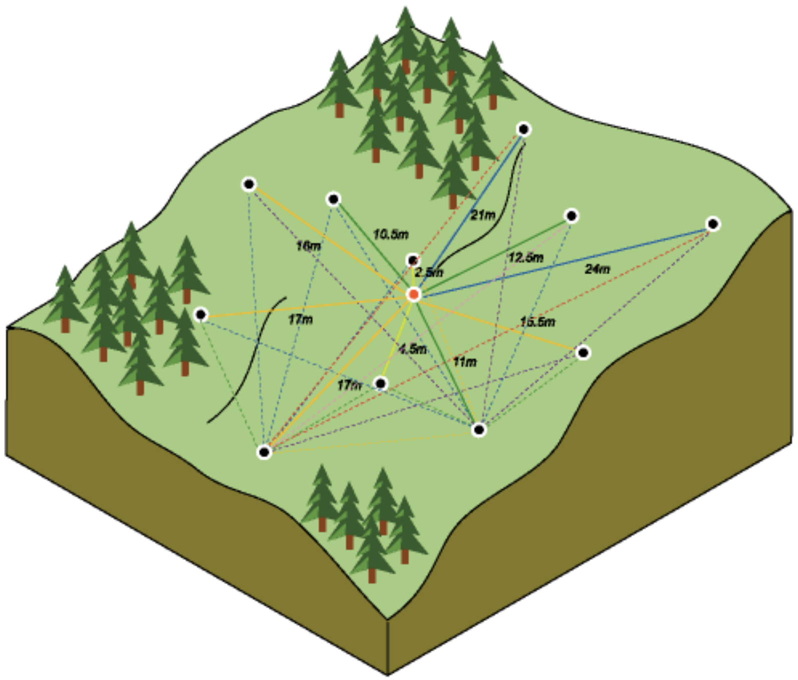


1. Expectation (squared distance between Si and Sj)
2. Variance of Y (between Si and Sj)
3. Plot Expectation on x-axis and Variance on y-axis

$$Y(S_i,S_j) = \tfrac{1}{2}\,\mathrm{var}(Y(S_i) - Y(S_j))$$

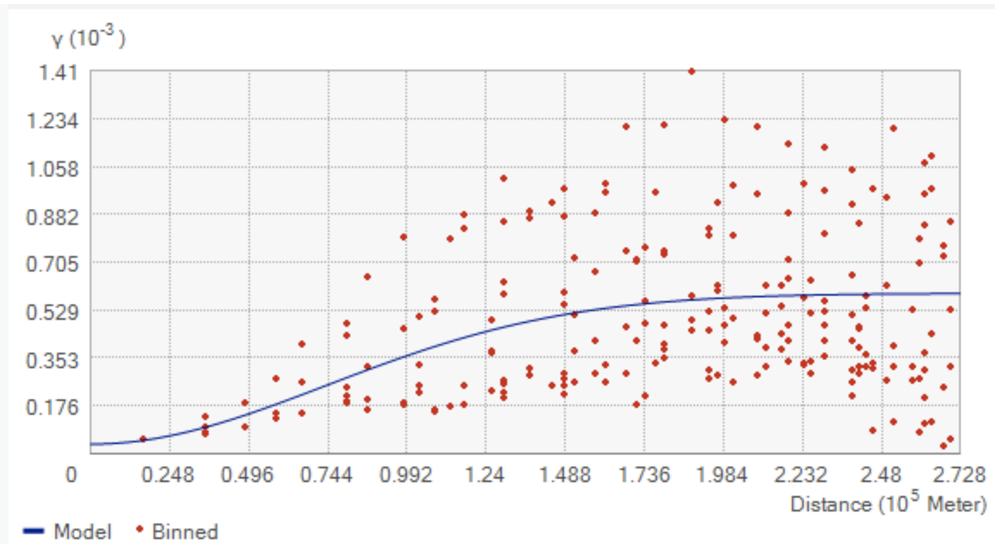
A more realistic model

Semi-variogram



Point prevalence = B0 + B1Vi + S(x)

S(x) = **spatial component**

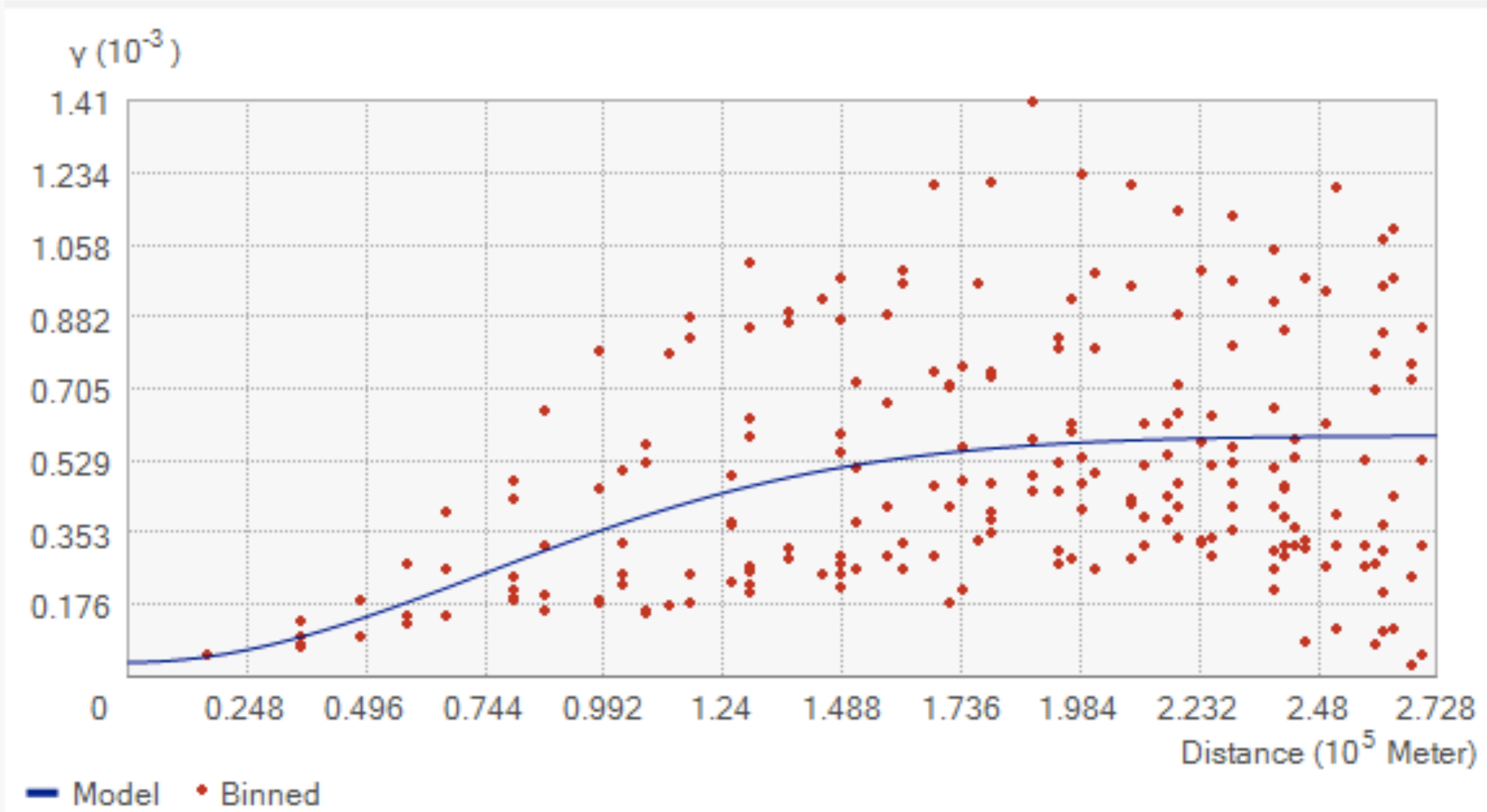


Source: ArGISPro (website) : http://pro.arcgis.com/en/pro-app/

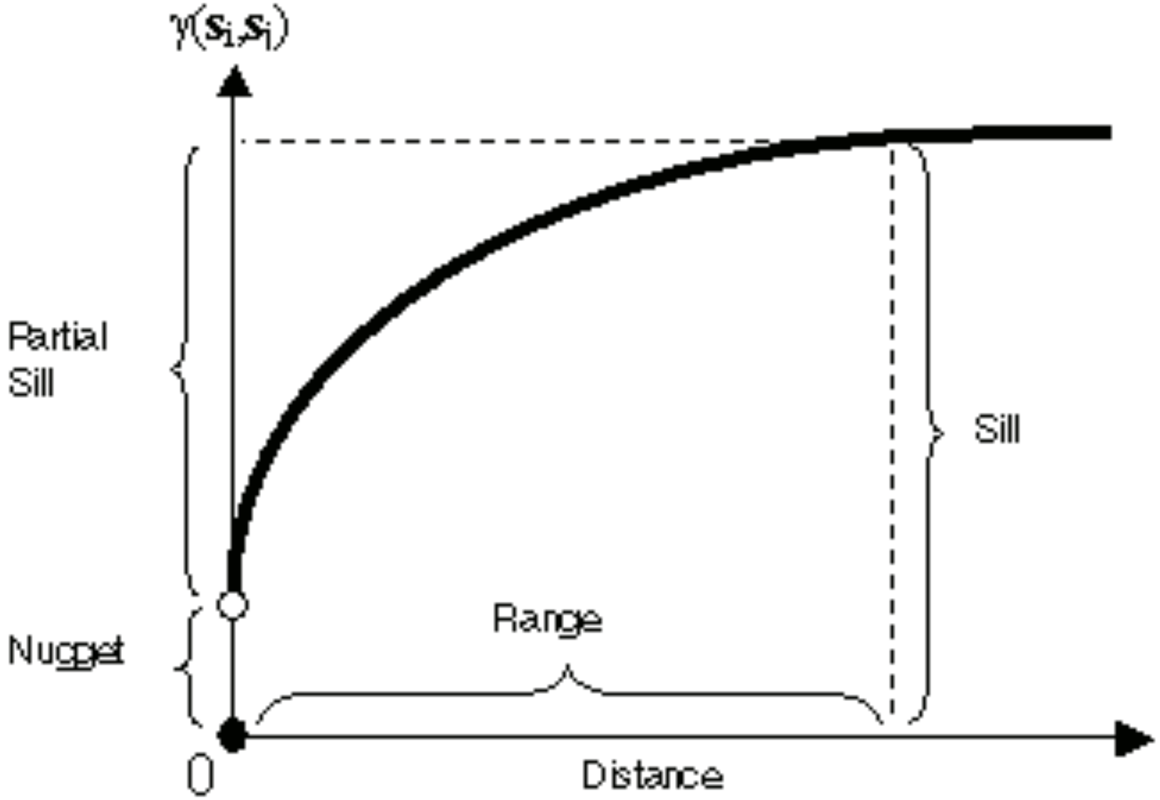
Red points = average distances squared

Blue line = line of best fit

Semi-variogram



Source: ArGISPro (website) : http://pro.arcgis.com/en/pro-app/



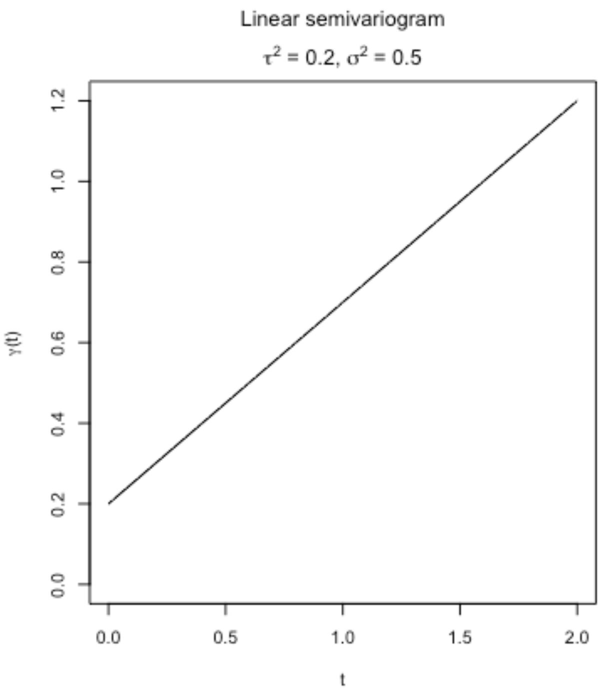
Source: ArGISPro (website) : http://pro.arcgis.com/en/pro-app/

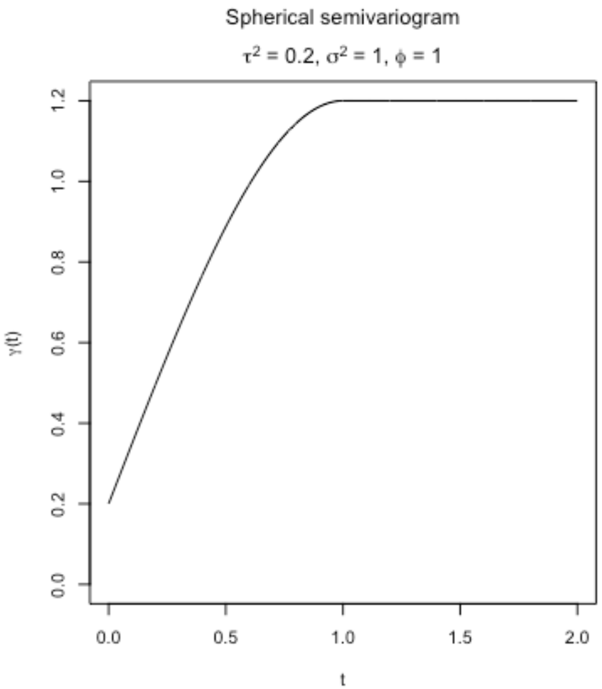
Semi-variogram

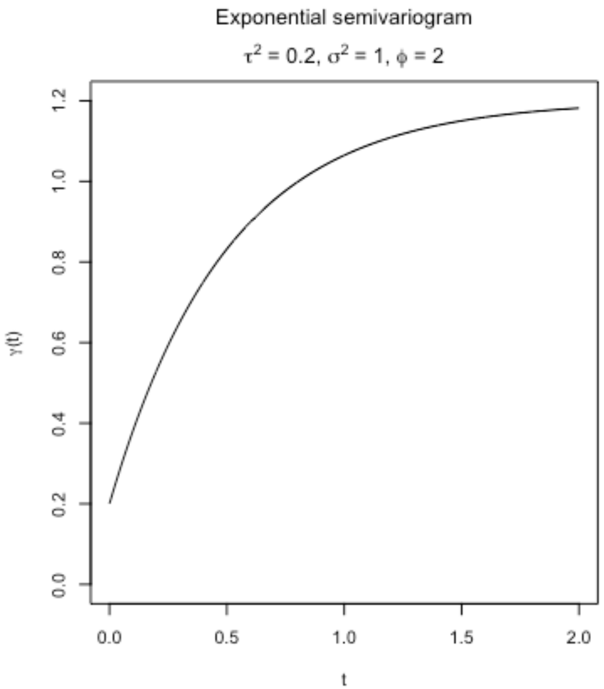
Sill
Nugget
Range

$$\gamma(si,sj) = \tfrac{1}{2}\, var(Y(si) - Y(sj))$$

Isotropy







$$\gamma(t) = \begin{cases} \tau^2 + \sigma^2 t & \text{if } t > 0,\ \tau^2 > 0,\ \sigma^2 > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\gamma(t) = \begin{cases} \tau^2 + \sigma^2 & \text{if } t \geq 1/\phi, \\ \tau^2 + \sigma^2 \{ \frac{3\phi t}{2} - \frac{1}{2}(\phi t)^3 \} & \text{if } 0 < t \leq 1/\phi, \\ 0 & \text{otherwise} \end{cases}$$

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

Recipe to choose the best one

1. Create distance bins
2. Calculate estimated semivariogram points using

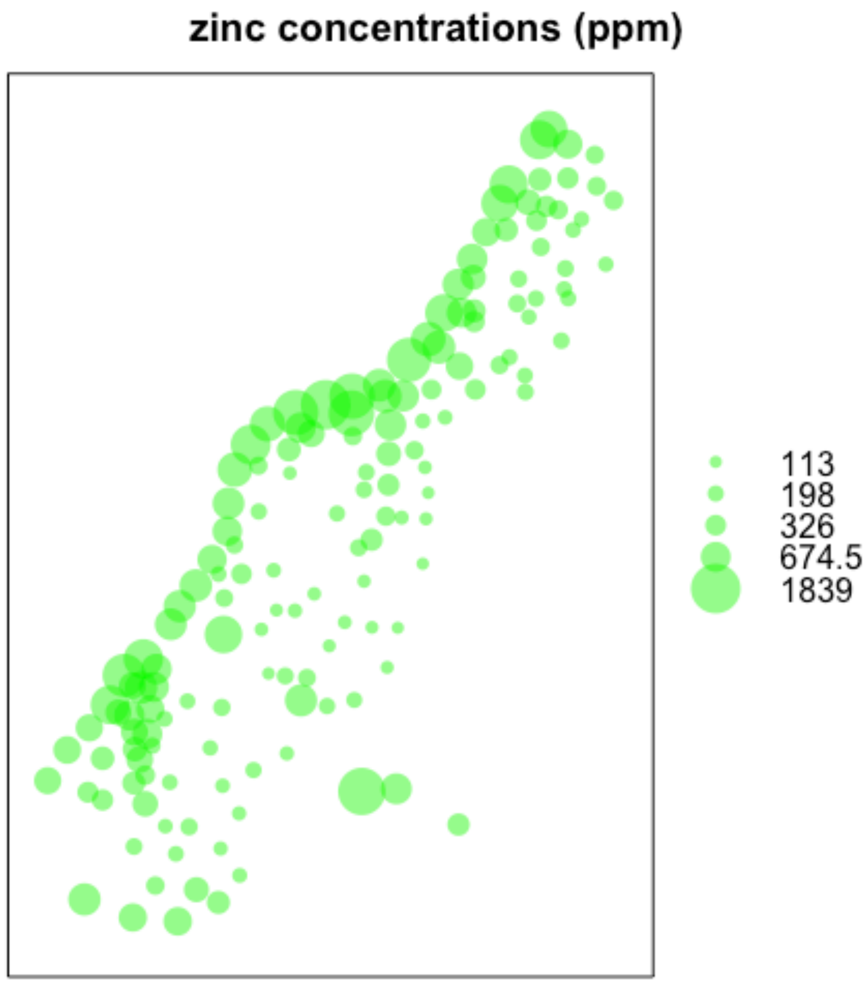
$$\hat{\gamma}(t) = \frac{1}{2N(t)} \sum_{(\mathbf{s}_i,\mathbf{s}_j)\in N(t)} \left[Y(\mathbf{s}_i) - Y(\mathbf{s}_j)\right]^2$$

$$N(t_k) = \left\{(\mathbf{s}_i, \mathbf{s}_j) : \|\mathbf{s}_i - \mathbf{s}_j\| \in I_k\right\}, k = 1, \ldots, K.$$

3. Fit the best parametric isotropic model
4. Estimate the model parameters

Demonstration in R

Output



Package : gstat

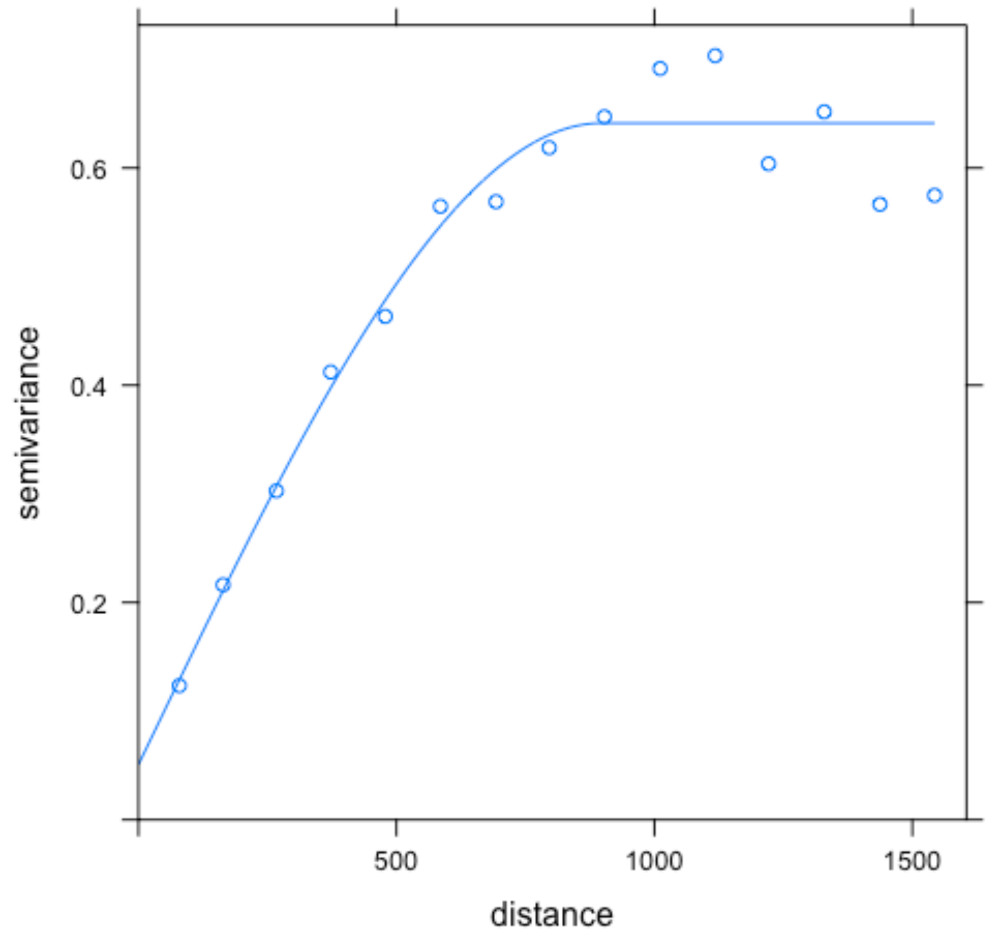
```
library(sp)
data(meuse)
head(meuse)
coordinates(meuse) = ~x+y

coordinates(meuse)[1:5,]

bubble(meuse,
"zinc",col=c("#00ff0088", "#00ff0088"),
main = "zinc concentrations (ppm)")
```

Demonstration in R

Package : gstat

Output

lzn.vgm = variogram(log(zinc)~1, meuse)
lzn.vgm

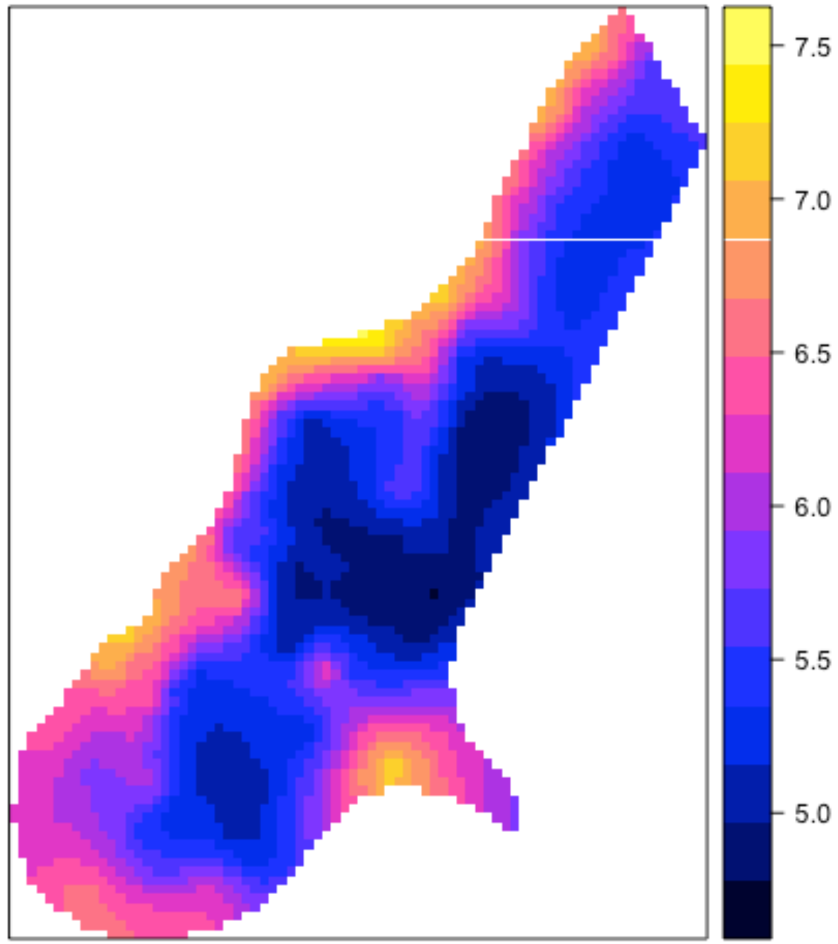


Package : gstat

Demonstration in R

Output

```
lzn.fit = fit.variogram(lzn.vgm, model = vgm
lzn.fit
plot(lzn.vgm, lzn.fit)

lzn.kriged = krige(log(zinc)~1, meuse, meus
spplot(lzn.kriged["var1.pred"])
```



Concluding example

Spatial analysis and mapping of malaria risk in Malawi using point-referenced prevalence of infection data (Kazembe et al, 2006)

Complex models: An application example

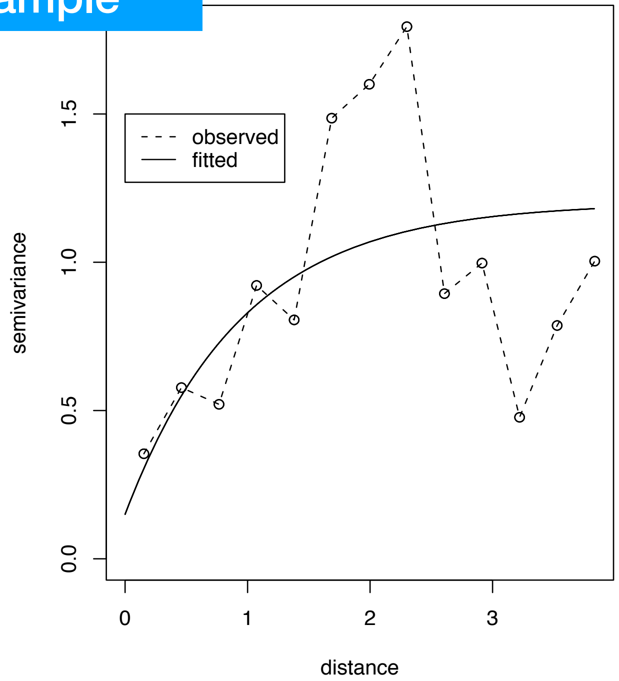


Figure 2
Empirical and fitted variogram of the logit transformed prevalence rate of infection. Separation distance is given in degrees latitude. Note: at equator one degree is approximately 120 km.

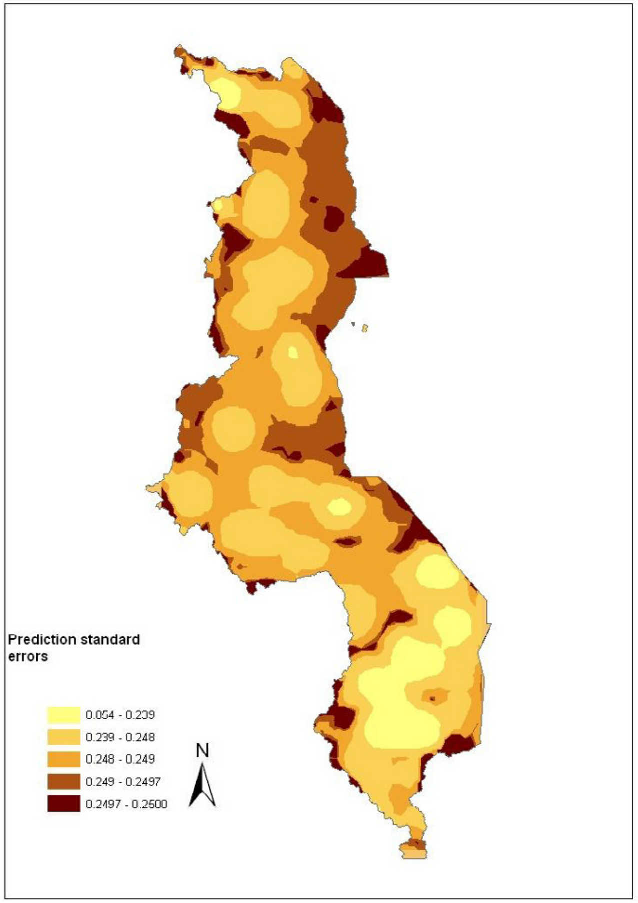


Figure 5
Map showing the prediction standard errors which are useful to quantify map precision. Cartographic visualization was carried out in ArcGIS.

QUIZ HINTS
or
Conclusion

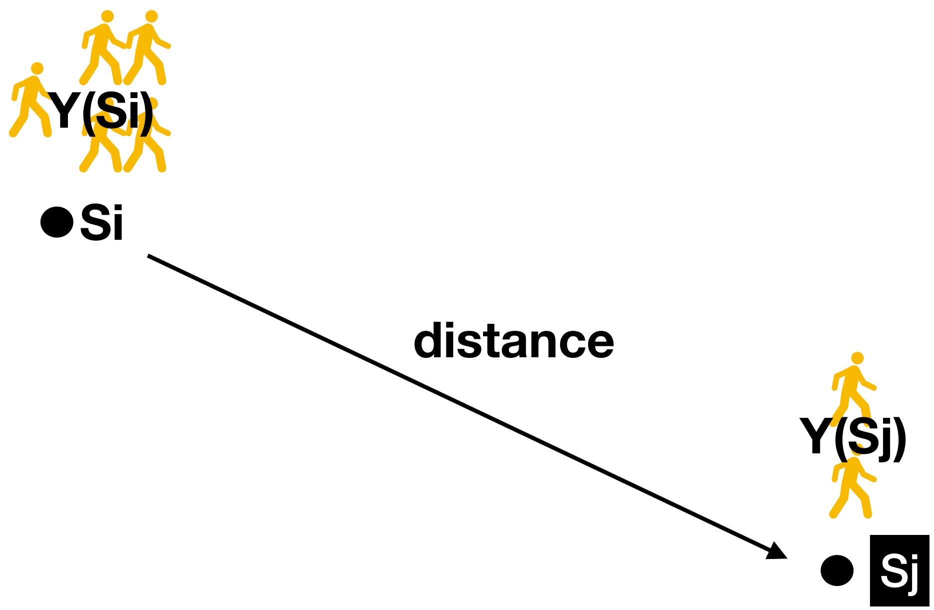
Two points closer together have a higher dependance

Stationarity

Isotropy

Semivariogram

Prediction



References

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data,* Chapman & Hall. 2, 6, 7, 9, 13, 14, 15, 16, 17, 18

Moore D.A., Carpenter, T. (1999,). Spatial analytical methods and geographic information systems: Use in health research and epidemiology. Epidemiologic Reviews, 21, 143{161. 1, 2, 4, 39

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). Hierarchical Modeling and Analysis for Spatial Data. Chapman & Hall. 2, 6, 7, 9, 13, 14, 15, 16, 17, 18

Bivand, R. S., Pebesma, E. J., and G´omez-Rubio, V. (2008). Applied Spatial Data Analysis with R. Springer New-York. 14

Esri (2017). Modeling a semivariogram. http://desktop.arcgis.com/de/arcmap/latest/ extensions/geostatistical-analyst/modeling-a-semivariogram.htm. Accessed 2018- 04-01. 8

Furrer, R. and Sain, S. R. (2010). spam: A sparse matrix R package with emphasis on mcmc methods for gaussian markov random fields. Journal of Statistical Software, 36, 1–25. 14, 16

Kjaerul, T. M., Ersboll, A. K., Gislason, G., and Schipperijn, J. (2016). Geographical clustering of incident acute myocardial infarction in denmark: A spatial analysis approach. Spatial and Spatio-temporal Epidemiology, 19, 46 – 59. 26

Kriege, D. (1951). A statistical approach to some basic mine valuation problems on the wit- watersrand. Journal of the Chemical, Metallurgical and Mining Society of South Africa, 52, 119–139. 9

Moore D.A., Carpenter, T. (1999,). Spatial analytical methods and geographic information systems: Use in health research and epidemiology. Epidemiologic Reviews, 21, 143–161. 1, 2, 4

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Computers \& Geosciences, 30: 683-691.