Lecture 1

Geostatistical problems and spatial exploratory analysis

1. Epidemiological data; empirical and mechanistic models; statistics and scientific method
2. Geostatistical problems; examples; visualising data
3. Autocorrelation; variogram analysis

Epidemiological data

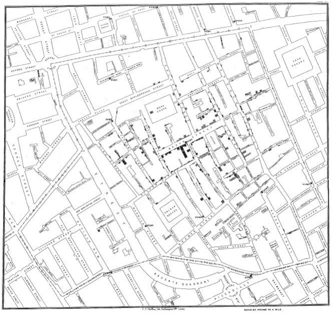
- incidence: number of new cases per unit time per unit population
- prevalence: number of existing cases per unit population
- risk: probability that a person will contract the disease (per unit time or per life-time)

Our general objective: how to understand spatial variation in disease incidence and/or prevalence and/or risk according to context

Course text

Diggle, P.J. and Giorgi, E. (2019). *Model-based Geostatistics: Methods and Applications in Global Public Health*. Boca Raton: CRC Press

In the beginning: Cholera in Victorian London, 1854



The physician John Snow famously removed the handle of the Broad Street water-pump, having concluded (correctly) that infected water was the source of the disease contrary to conventional wisdom at the time.

`https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak`

Epidemic vs endemic patterns of incidence

- Foot-and-mouth in Cumbria, UK Diggle (2006)
- Gastro-enteric disease in Hampshire, UK
  Diggle, Rowlingson and Su (2005)

Animations at: `http://www.lancaster.ac.uk/staff/diggle/`

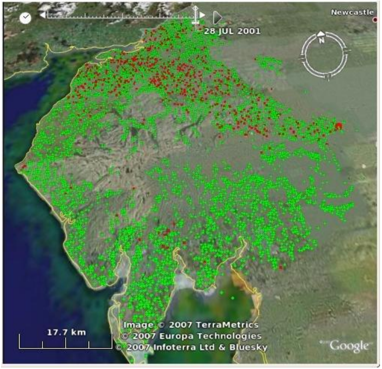
What are the similarities and differences between the two phenomena?

Diggle, P.J. (2006). Spatio-temporal point processes, partial likelihood, foot-and-mouth. *Statistical Methods in Medical Research*, **15**, 325–336.

Diggle, P., Rowlingson, B. and Su, T. (2005). Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics*, **16**, 423–34.

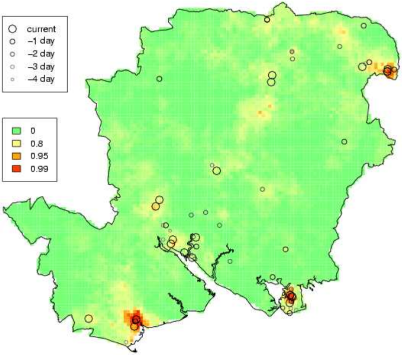
Mechanistic modelling: the 2001 UK FMD epidemic (Diggle, 2006)

- Predominantly a classic epidemic pattern of spread from an initial source
- Occasional apparently spontaneous outbreaks remote from prevalent cases
- $\lambda(x, t|\mathcal{H}_t)$ =conditional intensity, given history $\mathcal{H}_t$



Empirical modelling: The AEGISS project (Diggle, Rowlingson and Su, 2005)

- early detection of anomalies in local incidence
- data on 3374 consecutive reports of non-specific gastro-intestinal illness
- log-Gaussian Cox process, space-time correlation $\rho(u, v)$



A hierarchical modelling framework

Need to distinguish between:

- (scientific) modelling of a process whose behaviour we wish to understand;
- (statistical) modelling of data that tell us something about the process

Useful shorthand notation and a general framework

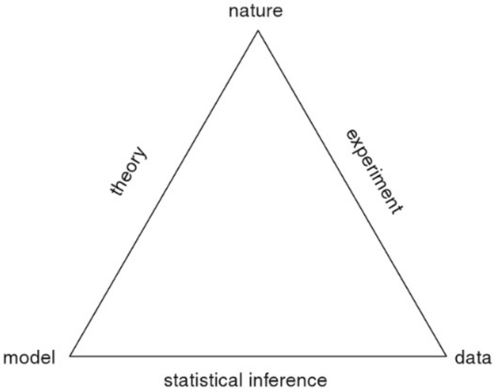
- $[\cdot]$ means the distribution of

  S : the scientific process we wish to understand
  Y : data that can help us understand the process

- hierarchical formulation:

$$[S, Y] = [S][Y|S]$$

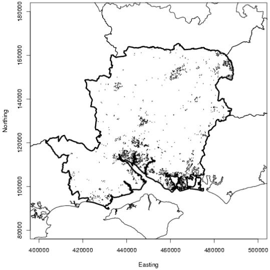
Statistics and scientific method



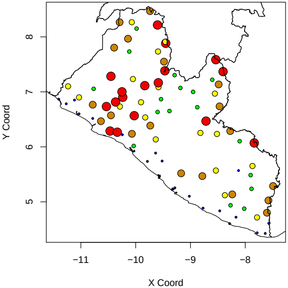
Diggle, P.J. and Chetwynd, A.G. (2011). *Statistics and Scientific Method: an Introduction for Students and Researchers*. Oxford: Oxford University Press.

Three data-sets

Cancer rates in administrative areas



Calls to NHS Direct in Hampshire, UK



Onchocerciasis prevalence surveys in Liberia

Are the three underlying processes fundamentally different?

Spatial stochastic processes

1. A stochastic process is a collection of random variables
2. A spatial stochastic process is a stochastic process in which each random variable is associated with a position in space
3. Three important types of spatial stochastic process:
   - discrete spatial variation: the random variables associate a real value with a particular, pre-specified, set of points in space, hence $\{(S_i, x_i) : i = 1, ..., n\}$
   - point processes: the random variables are the locations themselves, $\{x_i : i = 1, ..., n\}$
   - continuous spatial variation: the random variables associate a real value with every point in the space, hence $\{S(x) : x \in R^2\}$

Epidemiological study-designs

- Registry
  - case-counts in sub-regions to partition study-region (numerators)
  - population size in each sub-region (denominators)
  - collateral information from national census (covariates)
- Case-control
  - cases: all known cases within study region
  - controls: probability sample of non-cases within study-region
- **Survey (our focus)**
  - **sample of locations within study-region**
  - **collect data from each location**
  - **commonly used in developing country settings**

Geostatistics

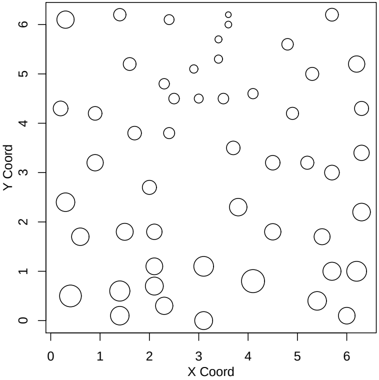
- traditionally, a self-contained methodology for spatial prediction:
  - origins in the South African mining industry
  - subsequently developed at École des Mines, Fontainebleau, France
- nowadays, that part of spatial statistics that is concerned with data obtained by spatially discrete sampling of a spatially continuous process

Model-based geostatistics: the application of general principles of statistical modelling and inference to geostatistical problems

Diggle P.J., Moyeed, R.A. and Tawn, J.A. (1998). Model-based geostatistics (with Discussion). *Applied Statistics*, **47**, 299-350.

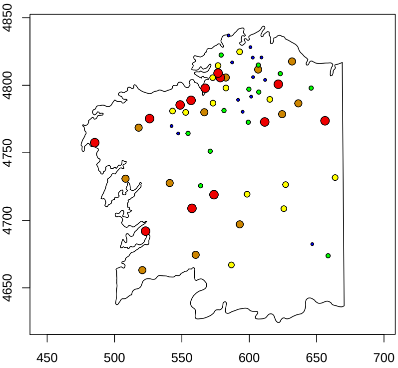
Stripped-down geostatistics

Given a set of measurements $Y_i : i = 1, ..., n$ at locations x_i in a spatial region A, presumed to be (noisy) measurements of a spatially continuous phenomenon $S(x_i)$, what can we say about the realisation of $S(x)$ throughout A?

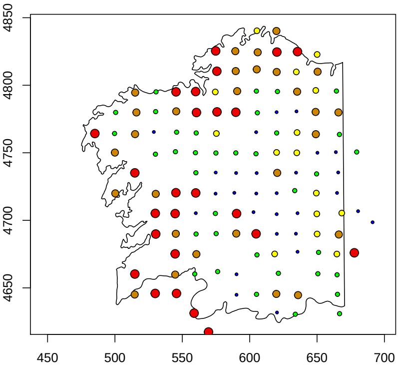


- Design: where to collect outcome data
- Estimation: how to fit a model
- Prediction: how to map the quantity of interest (prevalence)

Example 1. Environmental monitoring in Galicia, northern Spain

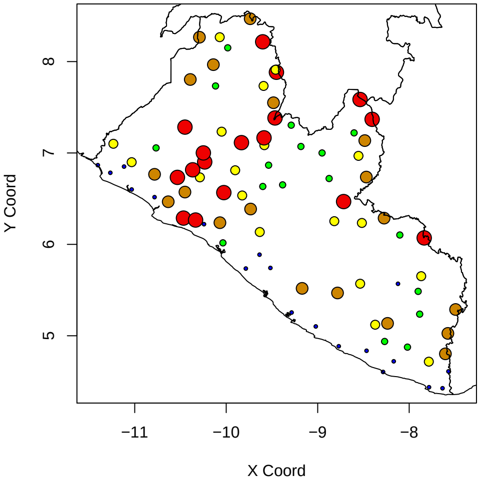


1997: why the uneven geographical coverage?



2000: a more efficient sampling design?

Example 2. Onchocerciasis in Liberia



- low prevalence near coast
- patches of high and low prevalence inland
- environmental risk-factors?

Data transformations

Modelling assumptions may be better satisfied by transforming the measurement data.

Two important transformations:

Logarithm

- converts multiplicative relationships to additive relationships

$$\log(XY) = \log(X) + \log(Y)$$

- makes skewed distributions more symmetric

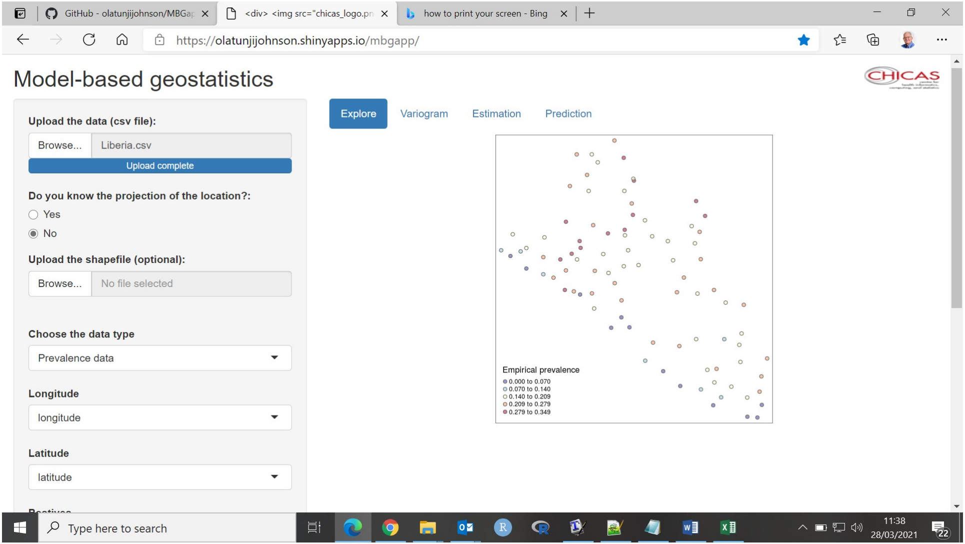
Empirical logit

- useful for exploratory analysis of prevalence data

  m_i = number tested; y_i = number positive

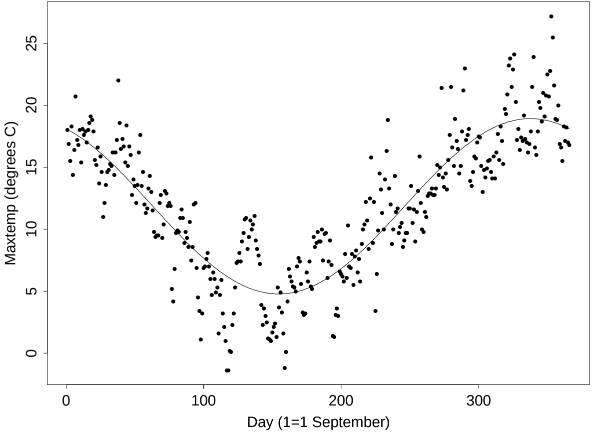
$$(m_i, y_i) \rightarrow \log\{(y_i + 0.5)/(m_i - y_i + 0.5)\}$$

Introducing the App: uploading and plotting geostatistical data



A digression into time series

- maximum daily temperatures (degrees C) at Bailrigg (Lancaster) field-station, September 1995 to August 1996
- an unusually cold Christmas 1995 was followed by a mild period in January-February



A harmonic regression model

$$
\begin{aligned}
Y(t) &= \mu + \alpha \cos(2\pi t/p + \phi) + \text{residual} \\
&= \mu + \beta_1 \cos(2\pi t/p) + \beta_2 \sin(2\pi t/p) + \text{residual}
\end{aligned}
$$

- $\mu =$ overall mean value (of time series $Y(t)$)
- $p =$ period
- $\alpha =$ amplitude
- $\phi =$ phase

Usually, the period is known, but the mean, amplitude and phase are not

Fitting the model

Use the second form of the model,

$$Y(t) = \mu + \beta_1 \cos(2\pi t/p) + \beta_2 \sin(2\pi t/p) + \text{residual}$$

Note that the following quantities are known, i.e. they can be calculated without having to estimate anything

- $x_1(t) = \cos(2\pi t/p)$
- $x_2(t) = \sin(2\pi t/p)$

Re-write the model as a linear regression model,

$$Y = \mu + \beta_1 x_1 + \beta_2 x_2$$

After fitting, amplitude and phase can be recovered using

$$\alpha = \sqrt{\beta_1^2 + \beta_2^2} \qquad \phi = \tan^{-1}(\beta_2/\beta_1)$$

Using the lm() function to fit the model

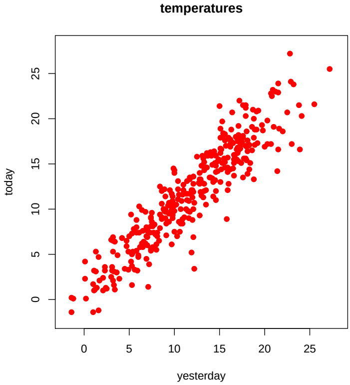
```
data<-read.csv(file.choose())
y<-data[,4];  day<-1:366
x1<-cos(2*pi*day/366); x2<-sin(2*pi*day/366)
fit<-lm(y~x1+x2)
summary(fit)
```

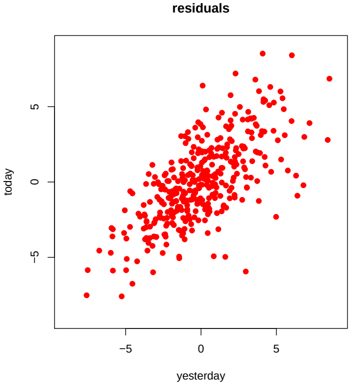
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.8467     0.1441   82.22   <2e-16 ***
x1            6.2508     0.2038   30.68   <2e-16 ***
x2           -3.3177     0.2038  -16.28   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.756 on 363 degrees of freedom
Multiple R-Squared: 0.7687,     Adjusted R-squared: 0.7674
F-statistic: 603.1 on 2 and 363 DF,  p-value: < 2.2e-16
```

Autocorrelation

- relationship between today's and yesterday's temperature?
- relationship between today's and yesterday's residual?

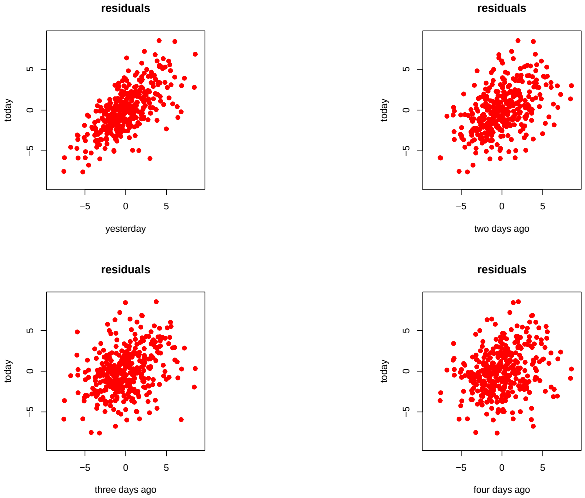




- how and why are the two relationships different?

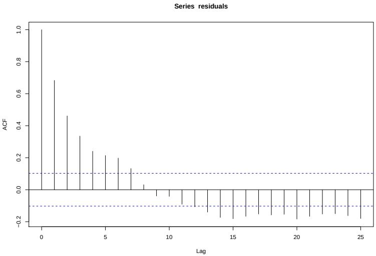
Autocorrelation (2)

How does the relationship between residuals today and k days ago change as k increases?



Autocorrelation (3)

- correlogram is a plot of correlation, r_k, between pairs of values k time-units apart



- dashed lines at $\pm 2/\sqrt{n}$ are pointwise 95% limits for uncorrelated residuals
- overall pattern is more important than indivudual values

Spatial correlation

- **First law of geography:** close things are more related than distant things.
- **Spatial correlation:**

$$Corr(Y(x_i), Y(x_j)) = f(x_i, x_j)$$

- Stationary process: $Var[Y(x)] = \sigma^2$, $f(x_i, x_j) = \rho(x_i - x_j)$.
- Stationary and isotropic process: $Var[Y(x)] = \sigma^2$, $f(x_i, x_j) = \rho(u)$, $u =$ distance between x_i and x_j

Another way of looking at correlation

Data-pairs: $(y_1, z_1), ..., (y_n, z_n)$

Means: $\bar{y} = (\sum y_i)/n \qquad \bar{z} = (\sum z_i)/n$

Variances: $v_y = \{(\sum(y_i - \bar{y})^2\}/n \qquad v_z = \{(\sum(z_i - \bar{z})^2\}/n$

Covariance: $g = \{(\sum(y_i - \bar{y})(z_i - \bar{z})\}/n$

Correlation: $r = g/\sqrt{v_y v_z}$

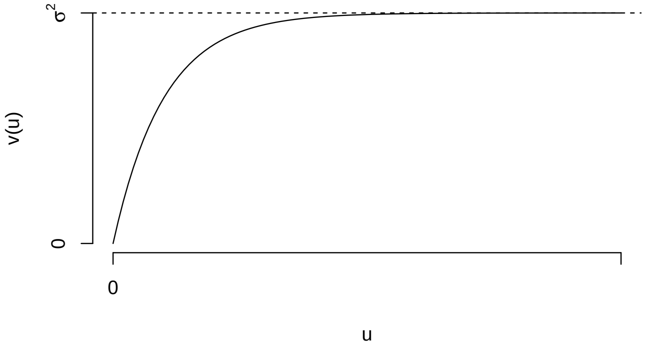
A little bit of algebra:

$$\sum\{(y_i - \bar{y}) - (z_i - \bar{z})\}^2 = \sum\{(y_i - \bar{y})^2\} + \sum\{(z_i - \bar{z})^2\} - 2\sum\{(y_i - \bar{y})(z_i - \bar{z})\}$$

The variogram

- Stationary, isotropic stochastic process $Y(x)$ with $E[Y(x)] = 0$, $Var[Y(x)] = \sigma^2$, $\mathrm{Corr}\{Y(x), Y(x'\} = \rho(u)$

$$
\begin{aligned}
V(u) &= \frac{1}{2}E[\{Y(x_i) - Y(x_j)\}^2] \\
&= \sigma^2\{1 - \rho(u)\}
\end{aligned}
$$



Estimation: the empirical variogram

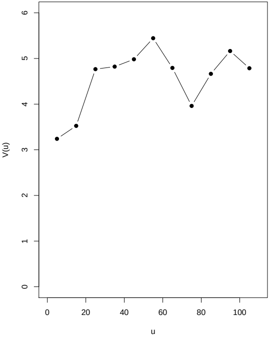
- Fit a linear regression model using available covariates
- Calculate residuals: $Z(x_1), \ldots, Z(x_n)$.
- Group pairs of locations into sets according to distance intervals u,

$$N(u) = \{(x_i, x_j) : u - 0.5h < \|x_i - x_j\| \leq u + 0.5h\}$$

- Empirical variogram:

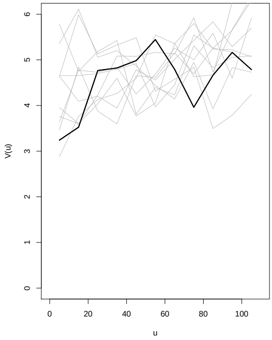
$$\hat{V}(u) = \frac{\sum \{Z(x_i) - Z(x_j)\}^2}{|N(u)|}$$

- increasing variogram corresponds to decreasing spatial correlation



Confirming existence of spatial correlation

- Randomly permute data-values across locations
- How do the resulting empirical variograms compare with the original?
- Use many more permutations for a formal test

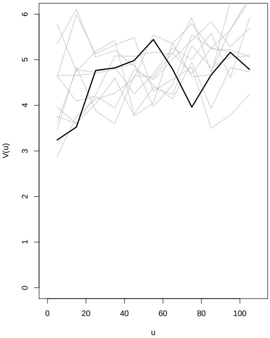


The behaviour of the variogram at small distances

$$\{Z(x) - Z(x)\}^2 = 0$$

So why might $V(u)$ not approach zero as u approaches zero?

- $V(0)$ represents the variance of any measurement error in the response
- But we can't estimate $V(u)$ at distances less than the smallest observed distance, u^* say.
- $\Rightarrow$ ambiguity between measurement error and spatial variation at distances less than u^*
- Implications for study design?



Summary: steps in the exploratory analysis of a geostatistical data-set

1. Data visualisation: map the outcome variable, also any covariates; scatterplot each covariate against the outcome variable; look for outlying data-points and find explanations for them; does transformation of the response and/or covariates lead to a more nearly linear relationship; do relationships between the outcome and covariates, whether linear or not, have face validity in context
2. Linear regression analysis: quantify the relationships between the outcome and covariates using standard regression models (but don't believe the p-values)
3. Variogram analysis: look for evidence of spatial correlation structure in the residuals from your preferred regression model

Variogram analysis using the App

- the shape of the empirical variogram depends on the set of distances that you use
- try to understand how the variogram can give you some indication of:
    - how noisy your residuals are
    - whether they are spatially correlated
    - and if so, over what approximate range of distances