

Spatial exploratory analysis

Dr Emanuele Giorgi
Lancaster University
e.giorgi@lancaster.ac.uk

Toowoomba 21-25 October 2019

Overview

- ▶ Defining geostatistical problems.
- ▶ Spatial exploratory analysis based on the variogram.

Epidemiological data

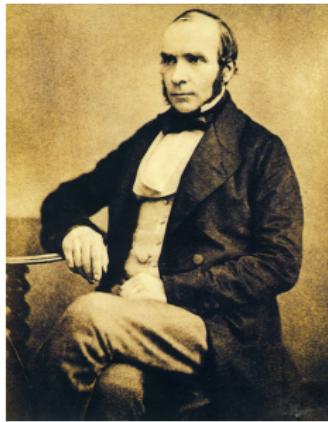
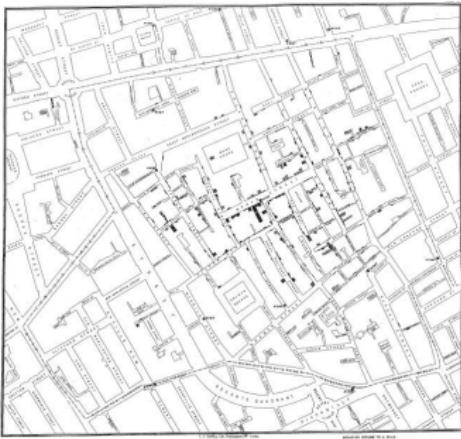
- ▶ **incidence:** number of new cases per unit time per unit population
- ▶ **prevalence:** number of existing cases per unit population
- ▶ **risk:** probability that a person will contract the disease (per unit time or per life-time)

General objective is to understand spatial variation in disease incidence and/or prevalence and/or risk according to context

Relevant books include

Elliott et al (2000); Gelfand et al (2010); Rothman (1986); Waller and Gotway (2004); Woodward (1999);

In the beginning: Cholera in Victorian London, 1854



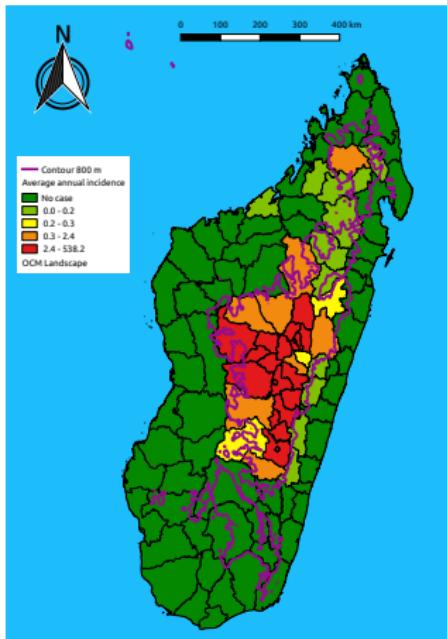
The physician [John Snow](#) famously removed the handle of the Broad Street water-pump, having concluded (correctly) that infected water was the source of the disease contrary to conventional wisdom at the time.

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

Study-designs

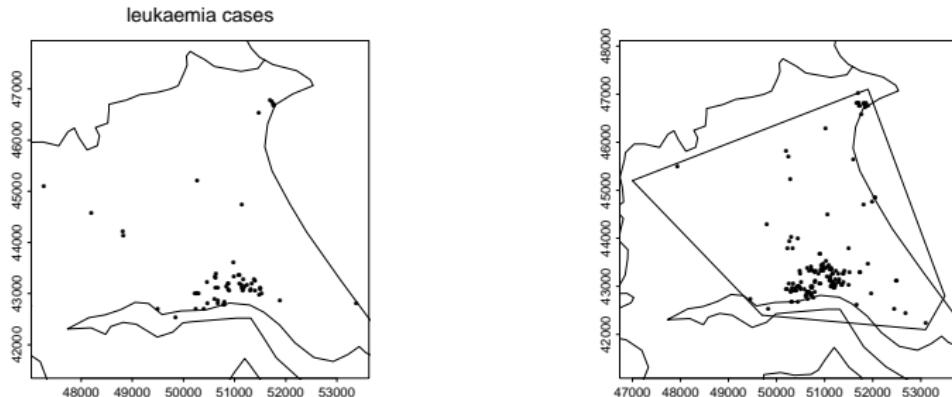
- ▶ Registry
 - ▶ case-counts in sub-regions to partition study-region (numerators)
 - ▶ population size in each sub-region (denominators)
 - ▶ collateral information from national census (covariates)
- ▶ Case-control
 - ▶ cases: all known cases within study region
 - ▶ controls: probability sample of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

Registry example. Plague in Madagascar



How much does risk in plague infection increase in areas above 800 m? Giorgi, E. et al. (2016). *Modelling of spatio-temporal variation in plague incidence in Madagascar from 1980 to 2007*. Spatial and Spatio-temporal Epidemiology. 19:125-135

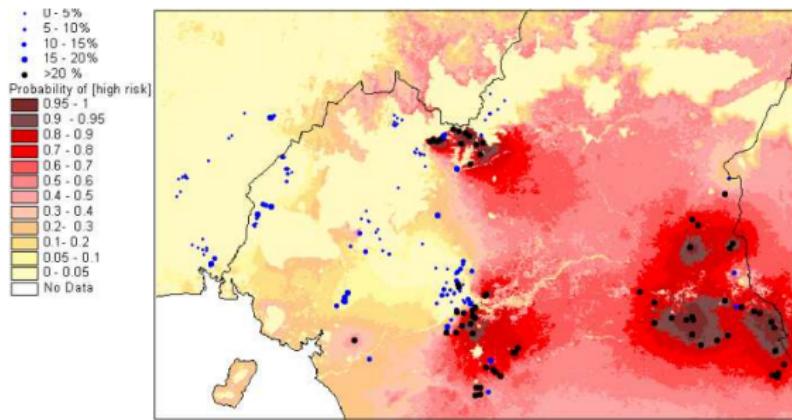
Case-control example Childhood leukaemia in Humberside



- ▶ residential locations of all known cases of childhood leukaemia in Humberside, England, over the period 1974-82;
- ▶ residential locations of a random sample of births

Cuzick and Edwards (1990); Diggle and Chetwynd (1991).

Survey example Loa loa prevalence in Cameroon



Source: M. J. Diggle et al., 2007. A spatial-temporal model for Loa loa prevalence in Cameroon.

Data are empirical prevalences in surveyed villages

Map shows predictive probabilities of exceeding 20% prevalence threshold

Diggle et al (2007)

What is the research question?

1. Plague in Madagascar

- ▶ Is elevation an important risk factor for plague infection?
- ▶ And if so, **why?**

2. Childhood leukaemia in Humberside

- ▶ Do cases show a **surprising** tendency to cluster together?

3. Loa loa in Cameroon

- ▶ What environmental characteristics affect the risk of disease?
- ▶ Can we predict where the prevalence of the disease exceeds a policy-based intervention threshold?

Epidemic vs endemic patterns of incidence

- ▶ Foot-and-mouth in Cumbria (the 2001 epidemic)

Diggle (2006)

- ▶ Gastro-enteric disease in Hampshire (AEGISS)

Diggle, Rowlingson and Su (2005)

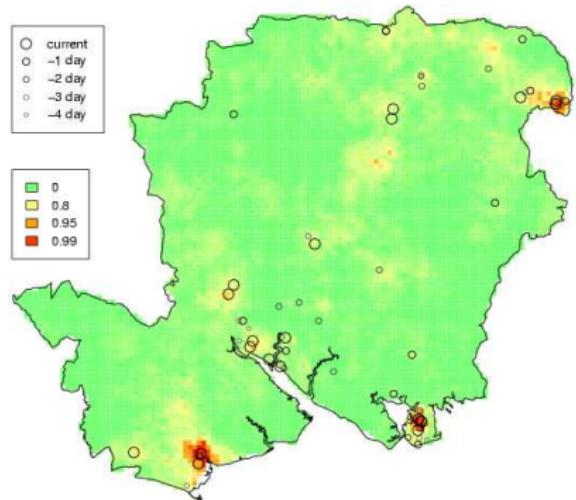
Animations at:

- ▶ <https://www.lancaster.ac.uk/staff/diggle/FMD/>
- ▶ <https://www.lancaster.ac.uk/staff/diggle/aegiss/>

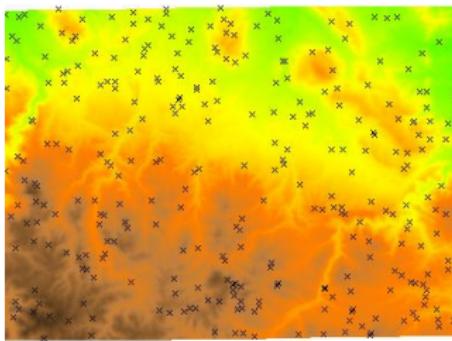
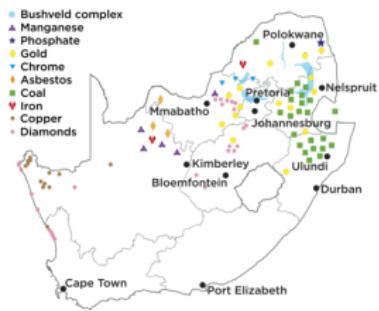
What are the similarities and differences between the two phenomena?

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

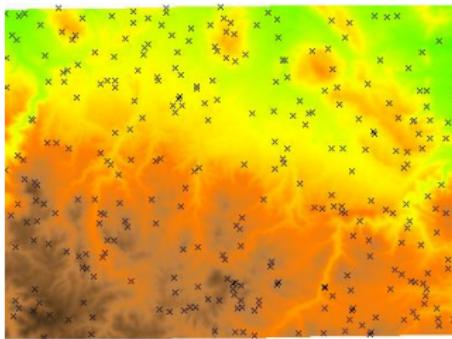
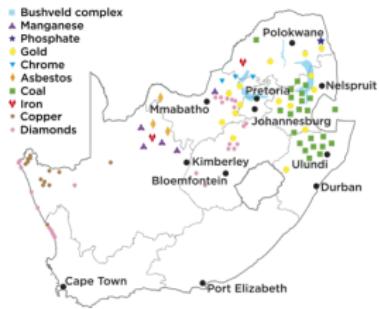
- ▶ early detection of anomalies in local incidence
- ▶ data on 3374 consecutive reports of non-specific gastro-intestinal illness
- ▶ log-Gaussian Cox process,
space-time correlation $\rho(u, v)$



Geostatistics

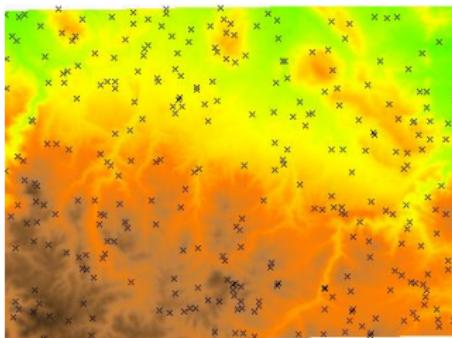
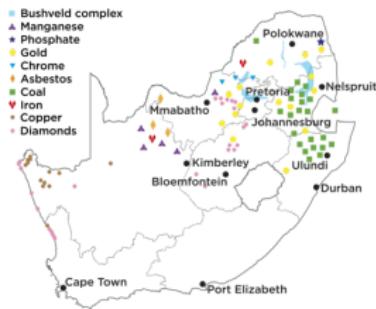


Geostatistics



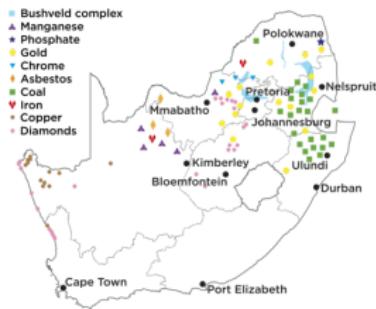
- ▶ Data: $\{(y_i, x_i), i = 1, \dots, n\}$, $x_i \in A \subset \mathbb{R}^2$.

Geostatistics



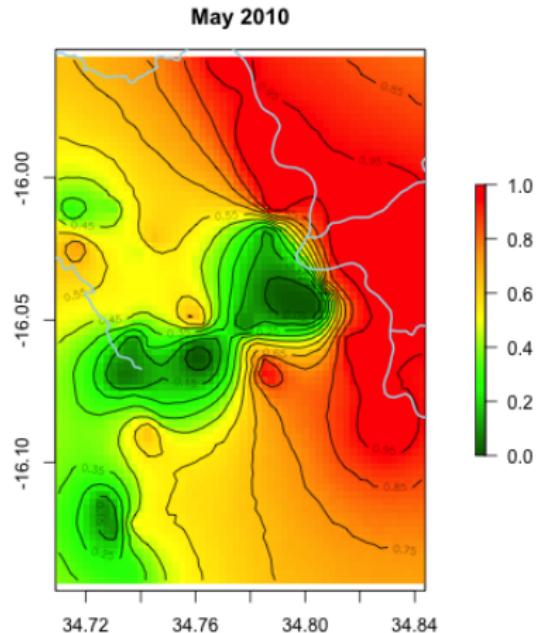
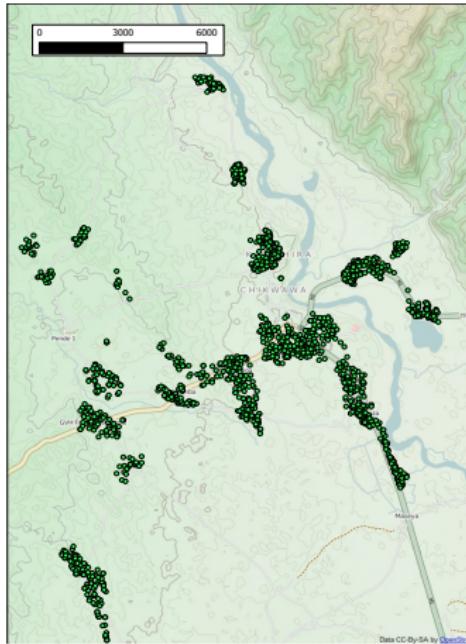
- ▶ Data: $\{(y_i, x_i), i = 1, \dots, n\}$, $x_i \in A \subset \mathbb{R}^2$.
- ▶ Model: $Y_i = S(x_i) + Z_i$.

Geostatistics



- ▶ Data: $\{(y_i, x_i), i = 1, \dots, n\}$, $x_i \in A \subset \mathbb{R}^2$.
- ▶ Model: $Y_i = S(x_i) + Z_i$.
- ▶ Objective: $\int_A S(x) dx$ (yielding of a mining operation).

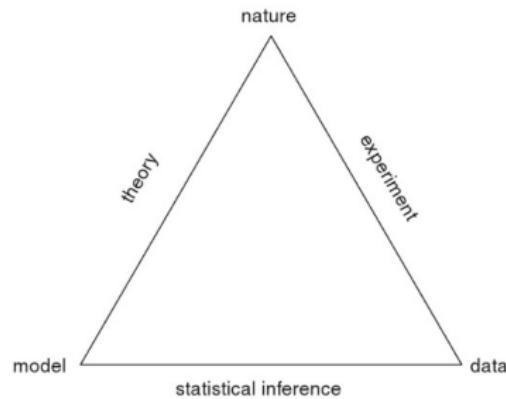
Model-based geostatistics



Animation at:

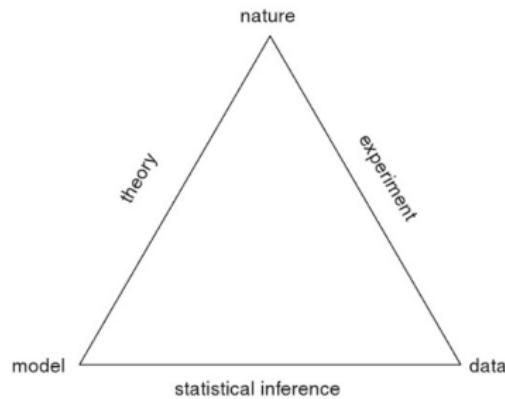
<https://www.lancaster.ac.uk/staff/giorgi/malaria/>

Science and statistics



Adapted from "Statistics and Scientific Method" (Diggle and Chatwynd, 2011).

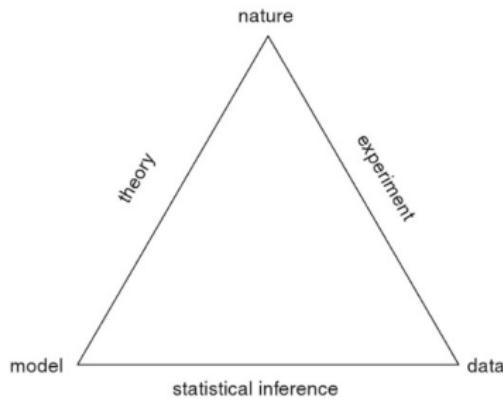
Science and statistics



Adapted from "Statistics and Scientific Method" (Diggle and Chatwynd, 2011).

- ▶ S = “process of nature”
- ▶ Y = “data”

Science and statistics



Adapted from "Statistics and Scientific Method" (Diggle and Chatwynd, 2011).

- ▶ S = “process of nature”
- ▶ Y = “data”

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

Testing for spatial correlation

Testing for spatial correlation

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

Testing for spatial correlation

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

$$\text{Corr}(Y(x_i), Y(x_j)) = f(x_i, x_j)$$

Testing for spatial correlation

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

$$\text{Corr}(Y(x_i), Y(x_j)) = f(x_i, x_j)$$

- ▶ Stationary process: $\text{Var}[Y(x)] = \sigma^2$, $f(x_i, x_j) = \rho(h)$, $h = x_i - x_j$.

Testing for spatial correlation

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

$$\text{Corr}(Y(x_i), Y(x_j)) = f(x_i, x_j)$$

- ▶ Stationary process: $\text{Var}[Y(x)] = \sigma^2$, $f(x_i, x_j) = \rho(h)$, $h = x_i - x_j$.
- ▶ Stationary and isotropic process: $\text{Var}[Y(x)] = \sigma^2$,
 $f(x_i, x_j) = \rho(u)$, $u = \|h\|$, $h = x_i - x_j$

The theoretical variogram

- ▶ Let $E[Y(x)] = 0$ and $\text{Var}[Y(x)] = \sigma^2$, stationary and isotropic for all x .

The theoretical variogram

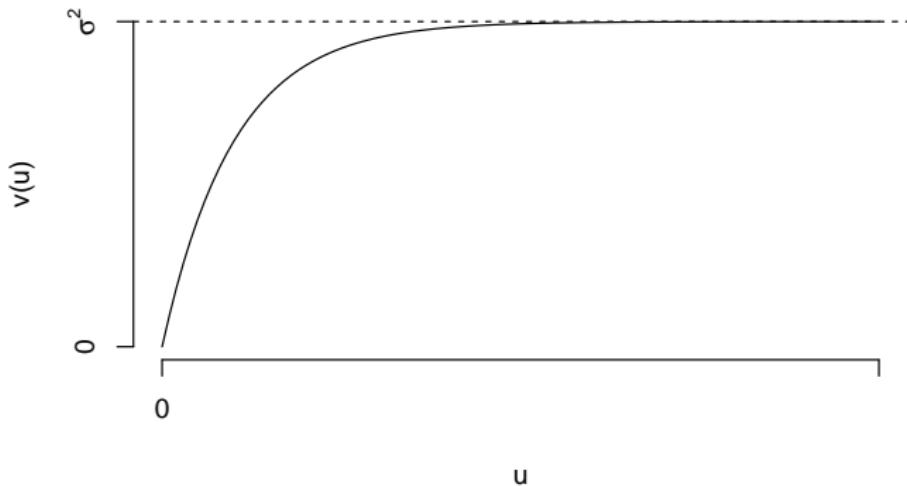
- ▶ Let $E[Y(x)] = 0$ and $\text{Var}[Y(x)] = \sigma^2$, stationary and isotropic for all x .

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

The theoretical variogram

- Let $E[Y(x)] = 0$ and $\text{Var}[Y(x)] = \sigma^2$, stationary and isotropic for all x .

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



The empirical variogram

- ▶ Residuals: $Z(x_1), \dots, Z(x_n)$.

The empirical variogram

- ▶ Residuals: $Z(x_1), \dots, Z(x_n)$.
- ▶ Set of points at distance u : $N(u) = \{(x_i, x_j) : \|x_i - x_j\| = u\}$

The empirical variogram

- ▶ Residuals: $Z(x_1), \dots, Z(x_n)$.
- ▶ Set of points at distance u : $N(u) = \{(x_i, x_j) : \|x_i - x_j\| = u\}$

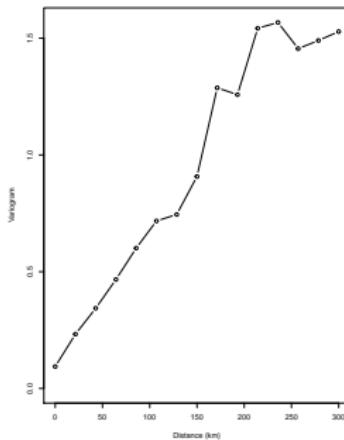
$$\hat{v}(u) = \frac{1}{|N(u)|} \sum_{(x_i, x_j) \in N(u)} \{Z(x_i) - Z(x_j)\}^2$$

The empirical variogram

- ▶ Residuals: $Z(x_1), \dots, Z(x_n)$.
- ▶ Set of points at distance u : $N(u) = \{(x_i, x_j) : \|x_i - x_j\| = u\}$

$$\hat{v}(u) = \frac{1}{|N(u)|} \sum_{(x_i, x_j) \in N(u)} \{Z(x_i) - Z(x_j)\}^2$$

- ▶ How do we know if $\hat{v}(u)$ shows evidence of spatial correlation?



The empirical variogram



An algorithm for testing spatial correlation.

The empirical variogram



An algorithm for testing spatial correlation.

1. Permute the locations x_1, \dots, x_n while holding fix the rest of the data.

The empirical variogram



An algorithm for testing spatial correlation.

1. Permute the locations x_1, \dots, x_n while holding fix the rest of the data.
2. Compute $\hat{v}(u)$ for the permuted data.

The empirical variogram



An algorithm for testing spatial correlation.

1. Permute the locations x_1, \dots, x_n while holding fix the rest of the data.
2. Compute $\hat{v}(u)$ for the permuted data.
3. Repeat 1 and 2 a large enough number of times (e.g. 1,000).

The empirical variogram



An algorithm for testing spatial correlation.

1. Permute the locations x_1, \dots, x_n while holding fix the rest of the data.
2. Compute $\hat{v}(u)$ for the permuted data.
3. Repeat 1 and 2 a large enough number of times (e.g. 1,000).
4. Compute the 95% confidence intervals (CIs) at each binned distance u .

If the empirical variogram falls within the 95% CIs.

The empirical variogram



An algorithm for testing spatial correlation.

1. Permute the locations x_1, \dots, x_n while holding fix the rest of the data.
2. Compute $\hat{v}(u)$ for the permuted data.
3. Repeat 1 and 2 a large enough number of times (e.g. 1,000).
4. Compute the 95% confidence intervals (CIs) at each binned distance u .

If the empirical variogram falls within the 95% CIs.

