Geo-spatial method for global health applications

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

Lancaster University, Lancaster, UK

Florence, 8-11 July 2019

Pre-requisites

- Good knowledge of generalized linear models
- Basic knowledge of R
- Notions of probability calculus (e.g. conditional distribution and expectation).
- Basic mastering of mathematical equations

Learning outcomes

You should be able:

- to understand the limitations of generalized linear models;
- to test for the presence of spatial correlation using variogram-based techniques;
- to formulate a suitable geostatistical model for data-analysis;
- to understand and correctly interpret the results from a geostatistical analysis;
- to fit generalized linear geostatistical models and carry out spatial prediction using `PrevMap` in R.

Overview

- Part I. Questioning the assumptions of generalized linear regression.
- Part II. The linear geostatistical model.
- Part III: The binomial geostatistical model.

Epidemiological data

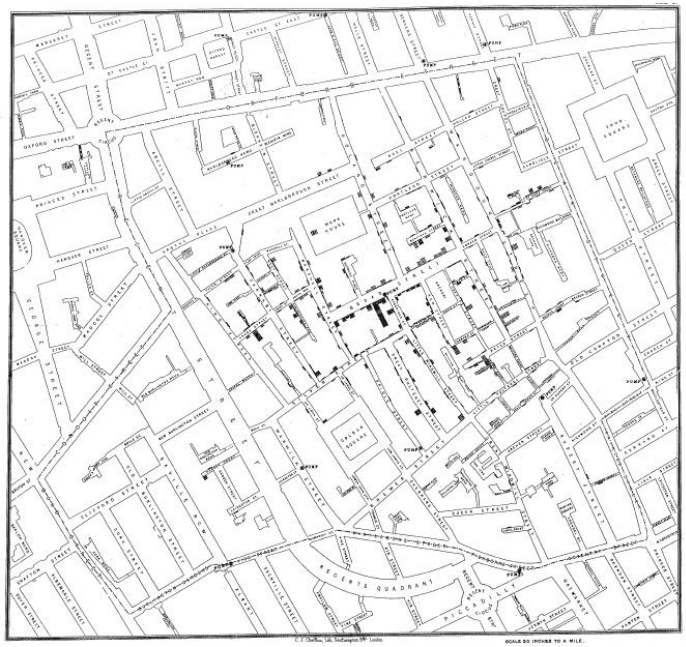
- incidence: number of new cases per unit time per unit population
- prevalence: number of existing cases per unit population
- risk: probability that a person will contract the disease (per unit time or per life-time)

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



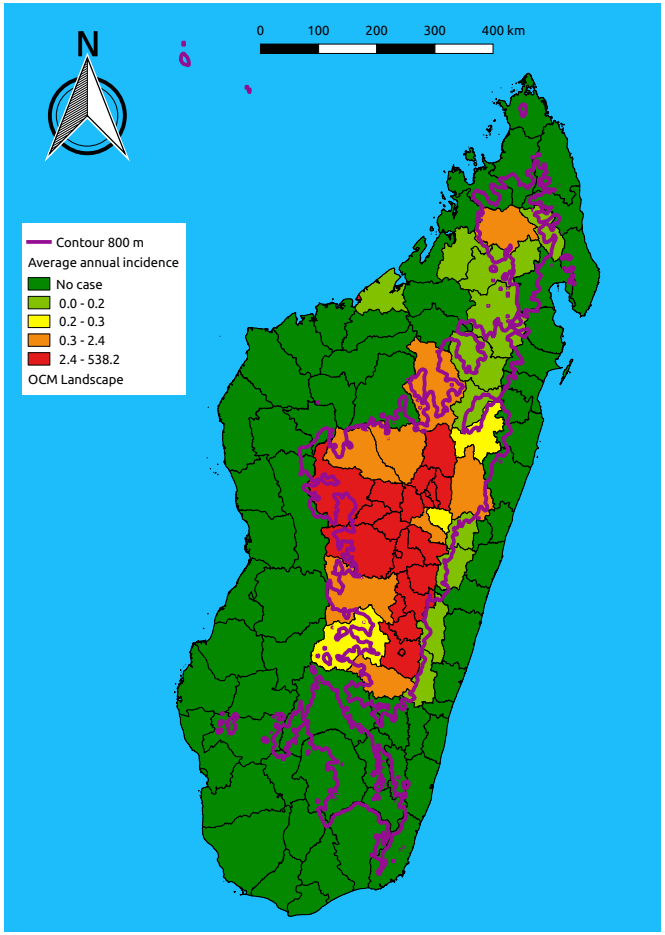
The physician John Snow famously removed the handle of the Broad Street water-pump, having concluded (correctly) that infected water was the source of the disease contrary to conventional wisdom at the time.

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

Study-designs

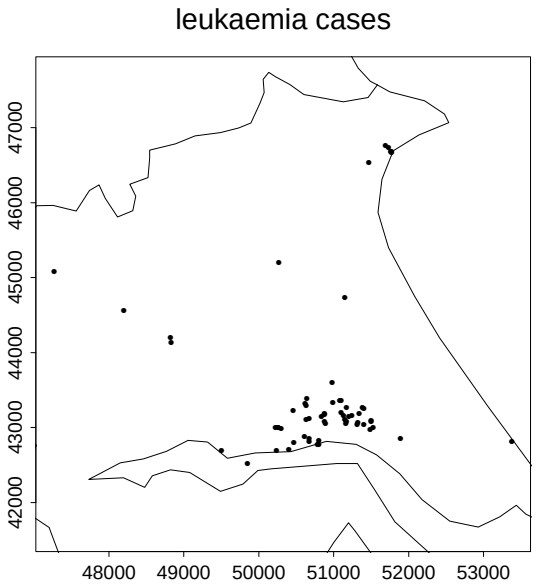
- Registry
  - case-counts in sub-regions to partition study-region (numerators)
  - population size in each sub-region (denominators)
  - collateral information from national census (covariates)
- Case-control
  - cases: all known cases within study region
  - controls: probability sample of non-cases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

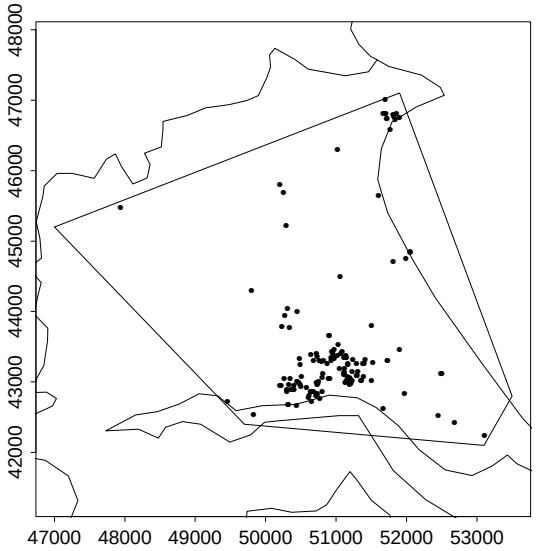
Registry example. Plague in Madagascar



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

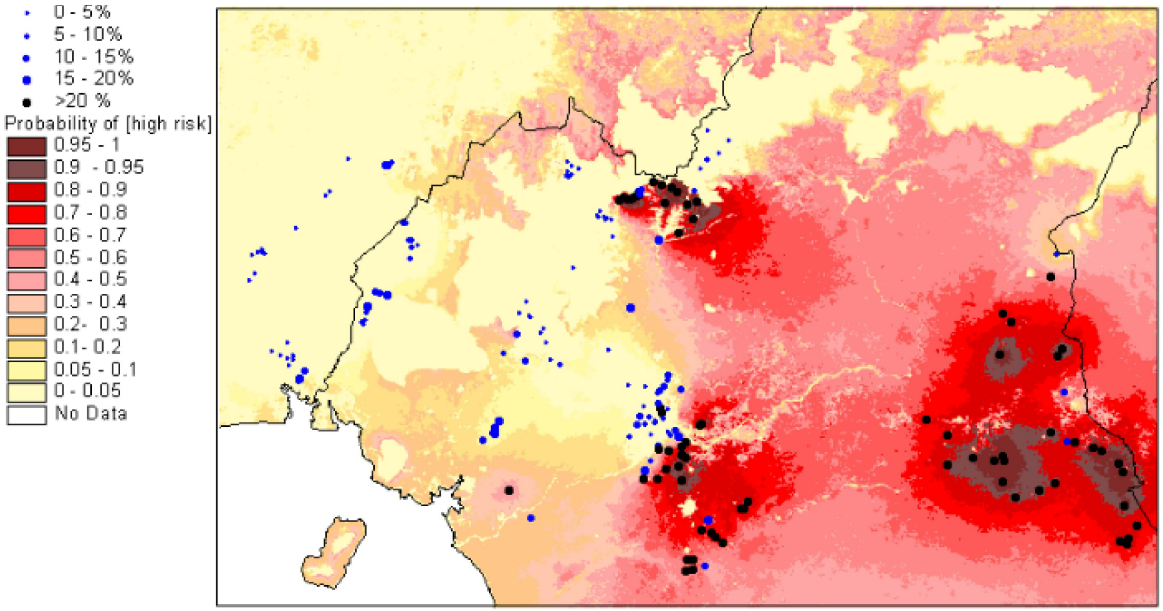


Figure 6: PCM for [high risk] in Cameroon based on LLRMr with ground truth data.

Data are empirical prevalences in surveyed villages

Map shows predictive probabilities of exceeding 20% prevalence threshold

Diggle et al (2007)

What is the public health 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)

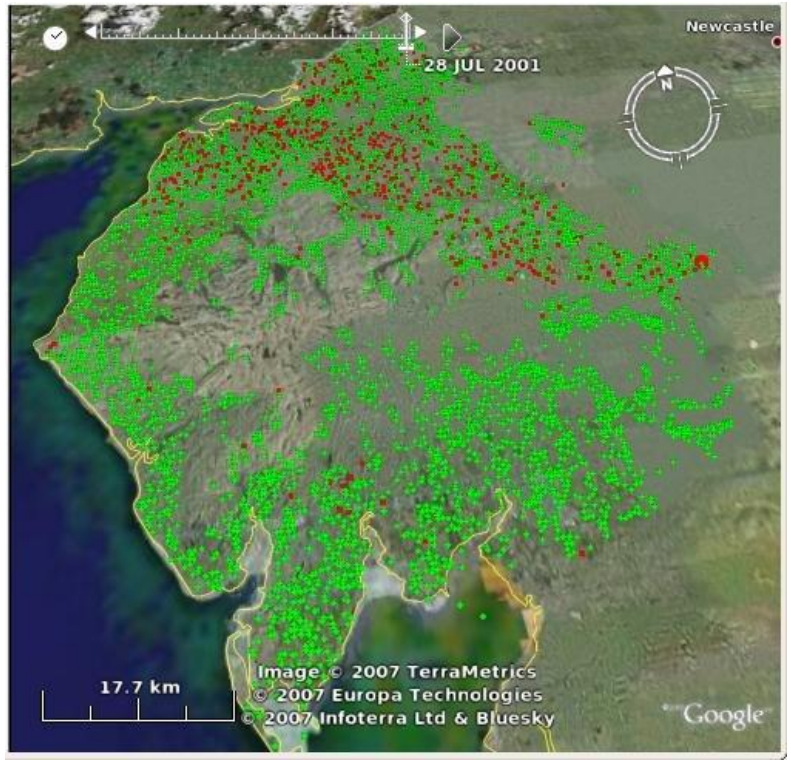
Diggle, Rowlingson and Su (2005)

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

What are the similarities and differences between the two phenomena?

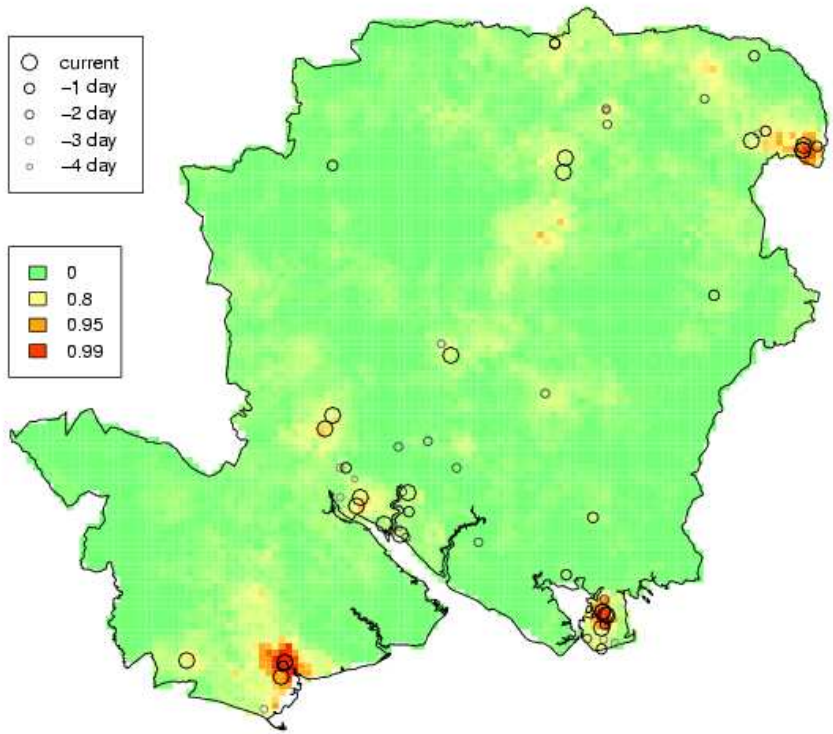
Mechanistic modelling: the 2001 UK FMD epidemic (Diggle, 2006)

- 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 $\mathcal{H}_t$

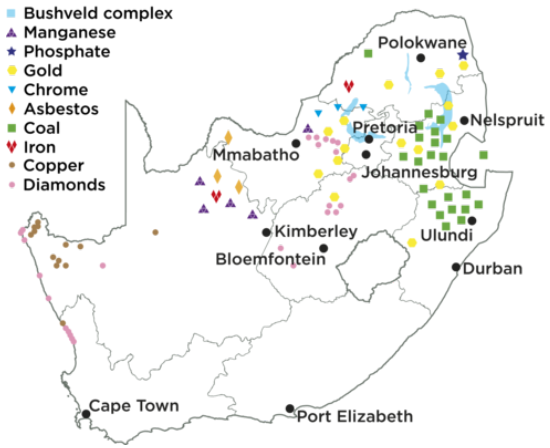


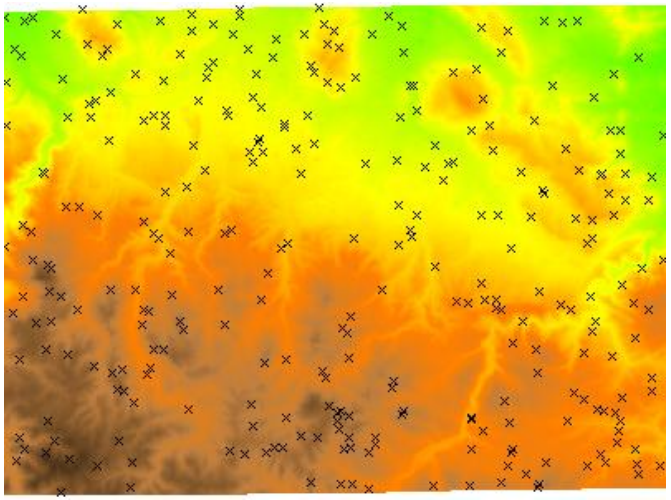
Empirical modelling: The AEGISS project (Diggle, Rowlingson and Su, 2005)

- 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 $\rho(u, v)$

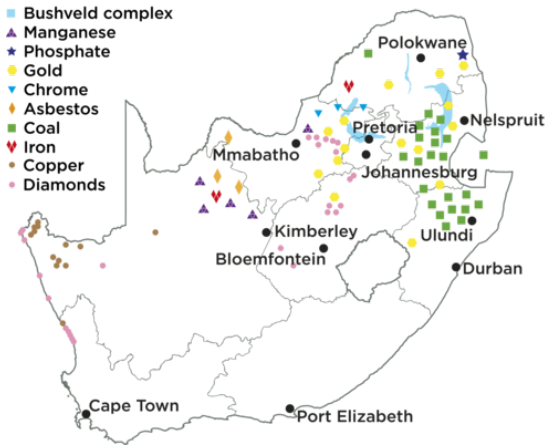


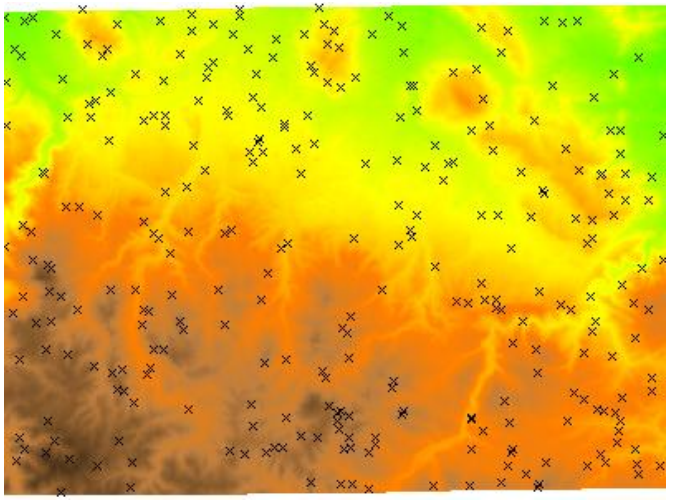
Geostatistics





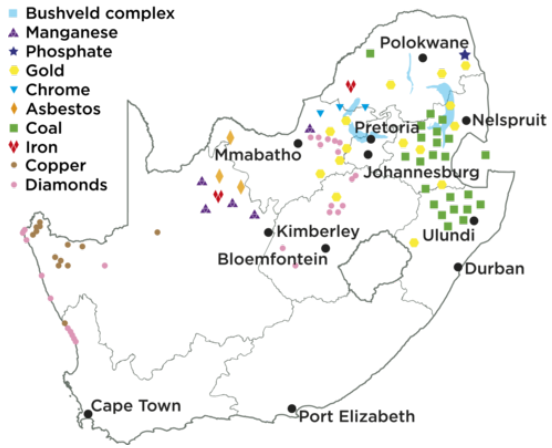
Geostatistics

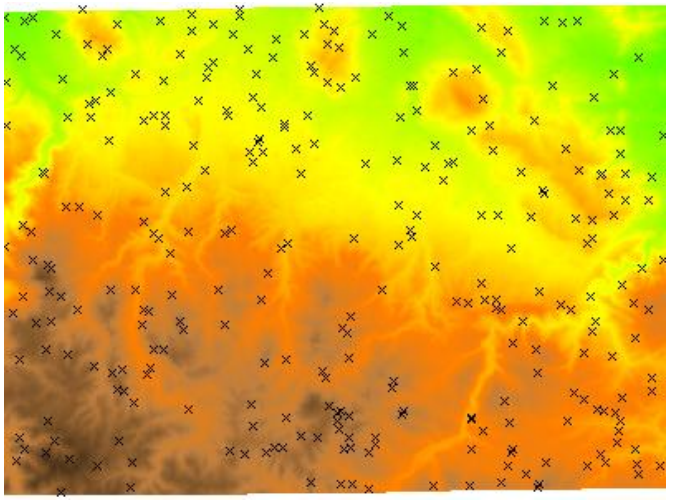




- Data: $\{(y_i, x_i), i = 1, \ldots, n\}$, $x_i \in A \subset \mathbb{R}^2$.

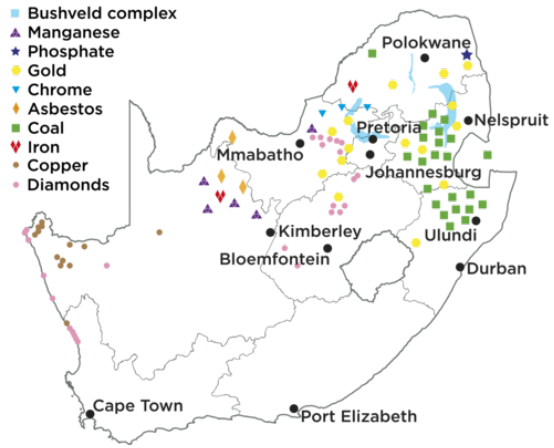
Geostatistics

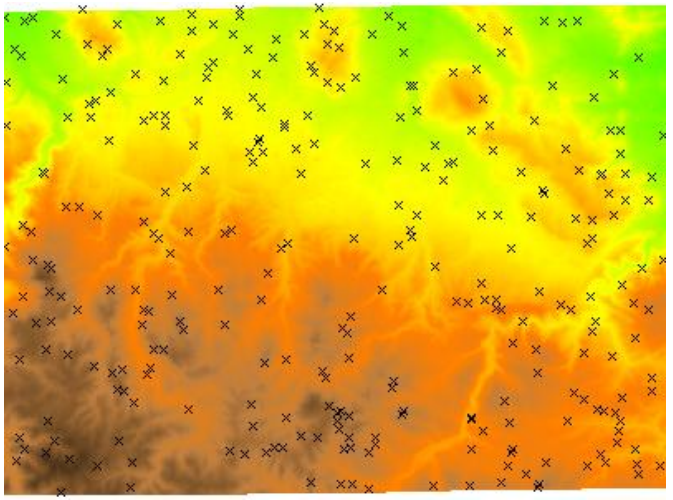




- Data: $\{(y_i, x_i), i = 1, \ldots, n\}$, $x_i \in A \subset \mathbb{R}^2$.
- Model: $Y_i = S(x_i) + Z_i$.

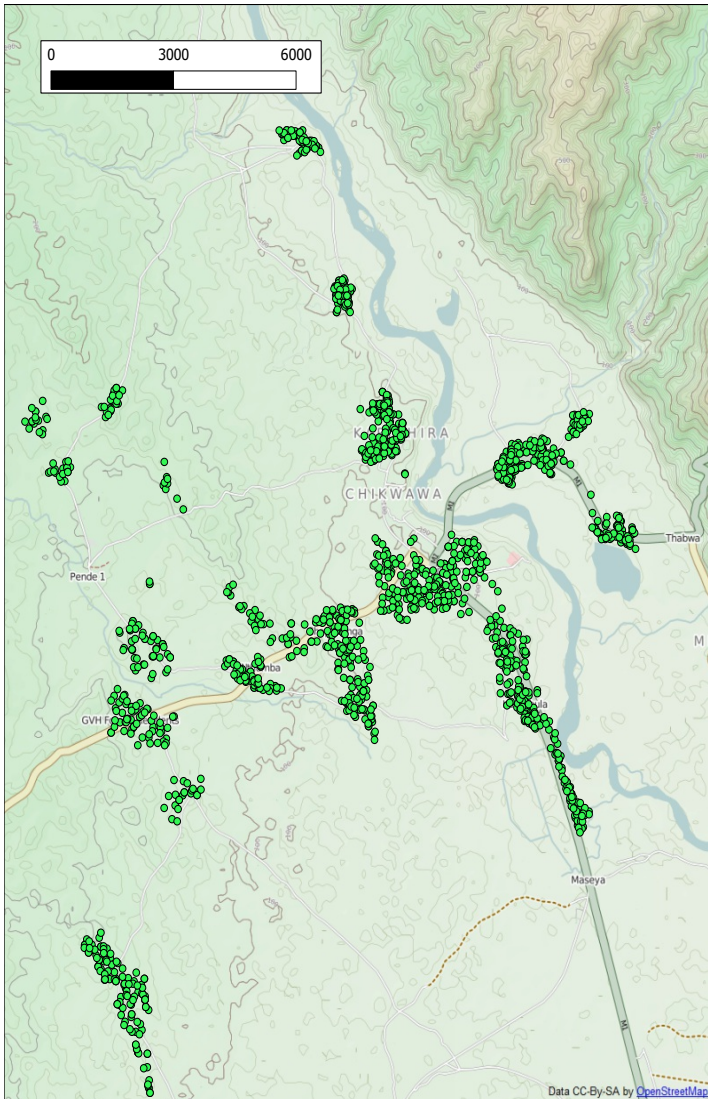
Geostatistics



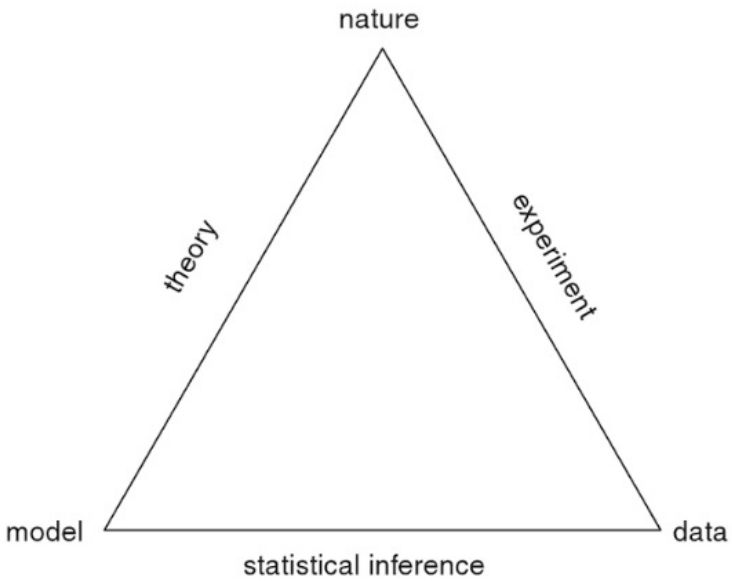


- Data: $\{(y_i, x_i), i = 1, \ldots, n\}$, $x_i \in A \subset \mathbb{R}^2$.
- Model: $Y_i = S(x_i) + Z_i$.
- Objective: $\int_A S(x)\, dx$ (yielding of a mining operation).

Model-based geostatistics

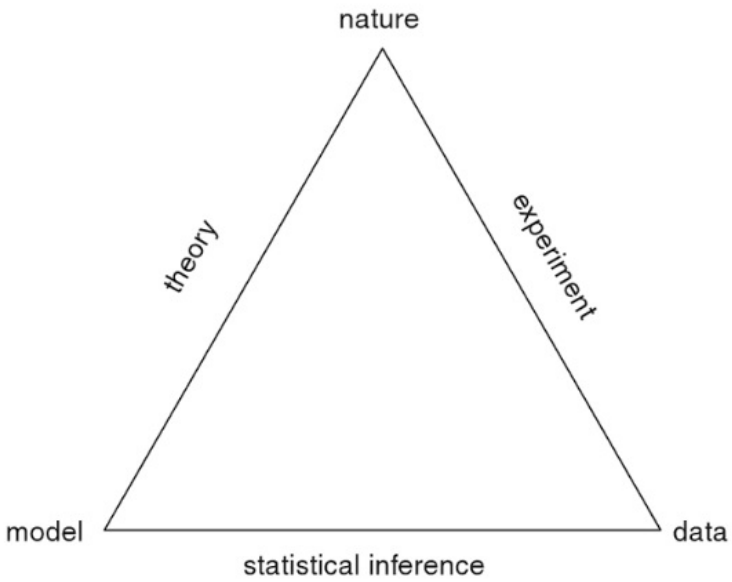


Science and statistics



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

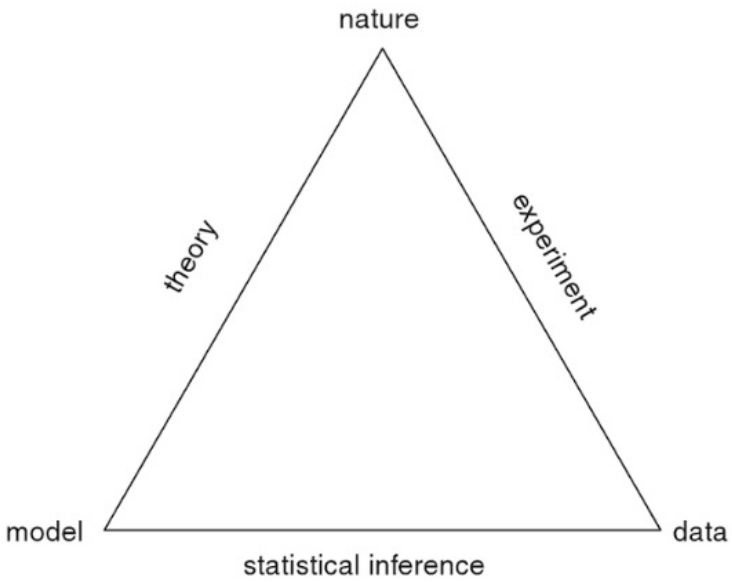
Science and statistics



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"

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

Part I: Questioning of the assumptions of generalized linear regression

A close look at generalized linear regression

Assumptions:

A close look at generalized linear regression

Assumptions:

1. $Y_i \sim f(\cdot)$ belongs to the exponential family;

A close look at generalized linear regression

Assumptions:

1. $Y_i \sim f(\cdot)$ belongs to the exponential family;
2. $E[Y_i] = m_i \mu_i$ and $Var[Y_i] = m_i V(\mu_i)$;

A close look at generalized linear regression

Assumptions:

1. $Y_i \sim f(\cdot)$ belongs to the exponential family;
2. $E[Y_i] = m_i \mu_i$ and $Var[Y_i] = m_i V(\mu_i)$;
3. $g(\mu_i) = \eta_i = d_i^\top \beta$;

A close look at generalized linear regression

Assumptions:

1. $Y_i \sim f(\cdot)$ belongs to the exponential family;
2. $E[Y_i] = m_i \mu_i$ and $Var[Y_i] = m_i V(\mu_i)$;
3. $g(\mu_i) = \eta_i = d_i^\top \beta$;
4. Y_i are mutually independent for $i = 1, \ldots, n$.

A close look at generalized linear regression

Assumptions:

1. $Y_i \sim f(\cdot)$ belongs to the exponential family;
2. $E[Y_i] = m_i \mu_i$ and $Var[Y_i] = m_i V(\mu_i)$;
3. $g(\mu_i) = \eta_i = d_i^\top \beta$;
4. Y_i are mutually independent for $i = 1, \ldots, n$.

Anything missing?

A close look at generalized linear regression

Assumptions:

1. $Y_i \sim f(\cdot)$ belongs to the exponential family;
2. $E[Y_i] = m_i \mu_i$ and $Var[Y_i] = m_i V(\mu_i)$;
3. $g(\mu_i) = \eta_i = d_i^\top \beta$;
4. Y_i are mutually independent for $i = 1, \ldots, n$.

Anything missing?

- $S = d^\top \beta$ (process of nature)

A close look at generalized linear regression

Assumptions:

1. $Y_i \sim f(\cdot)$ belongs to the exponential family;
2. $E[Y_i] = m_i \mu_i$ and $Var[Y_i] = m_i V(\mu_i)$;
3. $g(\mu_i) = \eta_i = d_i^\top \beta$;
4. Y_i are mutually independent for $i = 1, \ldots, n$.

Anything missing?

- $S = d^\top \beta$ (process of nature)
- Our model for the data: $[S][Y|S]$.

A close look at generalized linear regression

Assumptions:

1. $Y_i \sim f(\cdot)$ belongs to the exponential family;
2. $E[Y_i] = m_i \mu_i$ and $Var[Y_i] = m_i V(\mu_i)$;
3. $g(\mu_i) = \eta_i = d_i^\top \beta$;
4. Y_i are mutually independent for $i = 1, \ldots, n$.

Anything missing?

- $S = d^\top \beta$ (process of nature)
- Our model for the data: $[S][Y|S]$.
- Under the assumptions of classical GLMs, we can ignore $[S]$.

Over-dispersion

Over-dispersion

- **Definition:** the data show a greater variability than that implied by a classical GLM.

Over-dispersion

- **Definition:** the data show a greater variability than that implied by a classical GLM.

What causes over dispersion?

Over-dispersion

- **Definition:** the data show a greater variability than that implied by a classical GLM.

What causes over dispersion?

- **Example:** $Y = \sum_{i=1}^{n} Y_i$ such that $Y_i \sim \text{Bernoulli}(p)$ and $Cor(Y_i, Y_j) = \rho > 0$ $(i \neq j)$. Show that $Var(Y) > np(1-p)$.

Over-dispersion

- **Definition:** the data show a greater variability than that implied by a classical GLM.

What causes over dispersion?

- **Example:** $Y = \sum_{i=1}^{n} Y_i$ such that $Y_i \sim$ Bernoulli(p) and $Cor(Y_i, Y_j) = \rho > 0$ $(i \neq j)$. Show that $Var(Y) > np(1-p)$.

How to account for over-dispersion?

Over-dispersion

- **Definition:** the data show a greater variability than that implied by a classical GLM.

What causes over dispersion?

- **Example:** $Y = \sum_{i=1}^{n} Y_i$ such that $Y_i \sim$ Bernoulli(p) and $Cor(Y_i, Y_j) = \rho > 0$ $(i \neq j)$. Show that $Var(Y) > np(1-p)$.

How to account for over-dispersion?

1. *Marginal models.* e.g. quasi-models, $E[Y_i] = m_i \mu_i$ and $V[Y_i] = \phi m_i V(\mu_i)$ where ϕ is the over-dispersion parameter.

Over-dispersion

- **Definition:** the data show a greater variability than that implied by a classical GLM.

What causes over dispersion?

- **Example:** $Y = \sum_{i=1}^{n} Y_i$ such that $Y_i \sim$ Bernoulli(p) and $Cor(Y_i, Y_j) = \rho > 0$ $(i \neq j)$. Show that $Var(Y) > np(1-p)$.

How to account for over-dispersion?

1. *Marginal models.* e.g. quasi-models, $E[Y_i] = m_i \mu_i$ and $V[Y_i] = \phi m_i V(\mu_i)$ where ϕ is the over-dispersion parameter.
2. *Random effects models.* $S = d^\top \beta + Z$, where Z is a stochastic process.

A class of generalized linear mixed models

Assumptions:

A class of generalized linear mixed models

Assumptions:

1. Z_i are i.i.d. random variables;

A class of generalized linear mixed models

Assumptions:

1. Z_i are i.i.d. random variables;
2. $Y_i|Z_i \sim f(\cdot)$ belongs to the exponential family;

A class of generalized linear mixed models

Assumptions:

1. Z_i are i.i.d. random variables;
2. $Y_i|Z_i \sim f(\cdot)$ belongs to the exponential family;
3. $E[Y_i|Z_i] = m_i\mu_i$ and $Var[Y_i|Z_i] = m_i V(\mu_i)$;

A class of generalized linear mixed models

Assumptions:

1. Z_i are i.i.d. random variables;
2. $Y_i|Z_i \sim f(\cdot)$ belongs to the exponential family;
3. $E[Y_i|Z_i] = m_i\mu_i$ and $Var[Y_i|Z_i] = m_i V(\mu_i)$;
4. $g(\mu_i) = \eta_i = d_i^\top \beta + Z_i$;

A class of generalized linear mixed models

Assumptions:

1. Z_i are i.i.d. random variables;
2. $Y_i|Z_i \sim f(\cdot)$ belongs to the exponential family;
3. $E[Y_i|Z_i] = m_i\mu_i$ and $Var[Y_i|Z_i] = m_i V(\mu_i)$;
4. $g(\mu_i) = \eta_i = d_i^\top \beta + Z_i$;
5. $Y_i|Z_i$ are mutually independent for $i = 1, \ldots, n$.

A class of generalized linear mixed models

Assumptions:

1. Z_i are i.i.d. random variables;
2. $Y_i|Z_i \sim f(\cdot)$ belongs to the exponential family;
3. $E[Y_i|Z_i] = m_i\mu_i$ and $Var[Y_i|Z_i] = m_i V(\mu_i)$;
4. $g(\mu_i) = \eta_i = d_i^\top \beta + Z_i$;
5. $Y_i|Z_i$ are mutually independent for $i = 1, \ldots, n$.

Are the Y_i mutually independent?

- **Examples:** 1) $Y_i|Z_i \sim \text{Poisson}(e^{d_i^\top \beta + Z_i})$ and $Z_i \sim N(-\tau^2/2, \tau^2)$ i.i.d.; $E[Y_i] =$ ..... and $Var[Y_i] =$ ..... (Hint: Use the law of total expectation and variance)
  2) $Y_i|Z_i \sim \text{Poisson}(e^{d_i^\top \beta + Z_i})$, $e^{Z_i} \sim \text{Gamma}(k, k)$ i.i.d.; show that Y_i is a Negative Binomial distribution.

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:**

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

Testing for spatial correlation

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

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

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

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

The theoretical variogram

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

The theoretical variogram

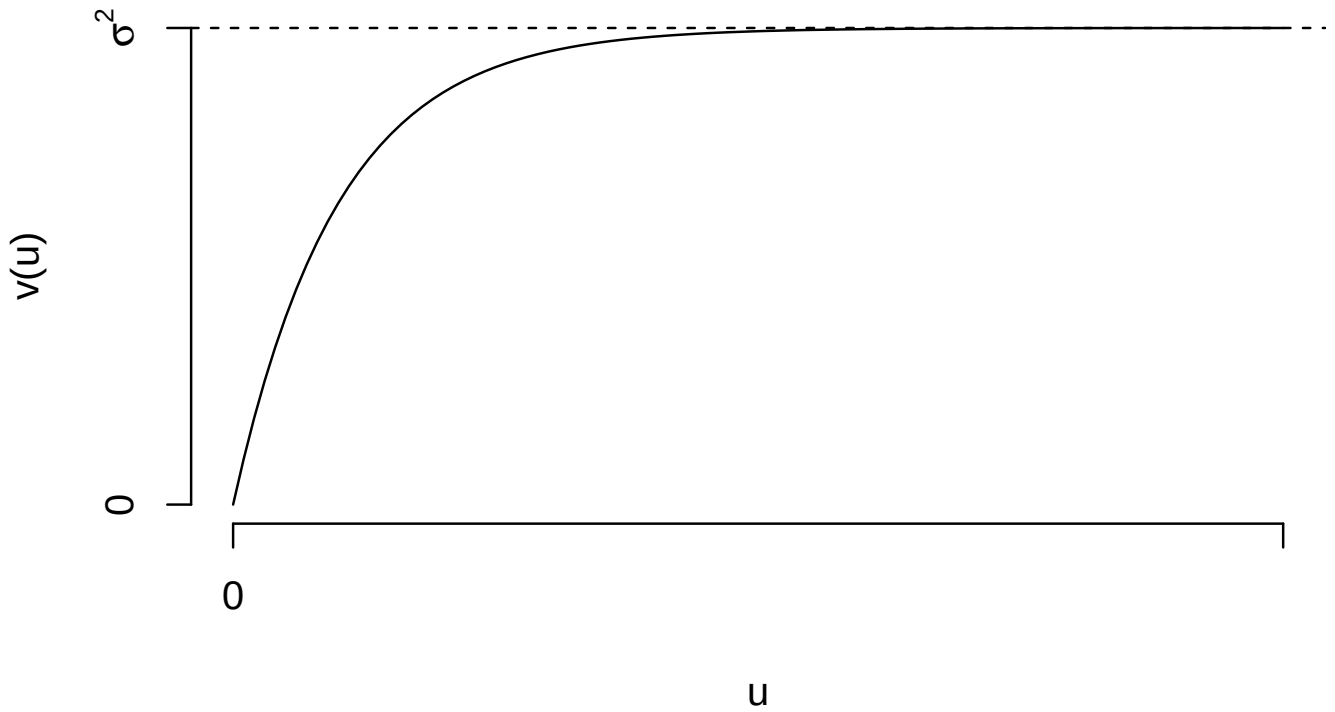
- Let $E[Y(x)] = 0$ and $Var[Y(x)] = \sigma^2$, stationary and isotropic for all x.

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

The theoretical variogram

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

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



The empirical variogram

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

The empirical variogram

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

The empirical variogram

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

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

- Residuals: $Z(x_1), \ldots, Z(x_n)$.
- 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?
  1. Permute the locations $x_1, \ldots, x_n$ while holding fix the rest of the data.

The empirical variogram

- Residuals: $Z(x_1), \ldots, Z(x_n)$.
- 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?
  1. Permute the locations $x_1, \ldots, x_n$ while holding fix the rest of the data.
  2. Compute $\hat{v}(u)$ for the permuted data.

The empirical variogram

- Residuals: $Z(x_1), \ldots, Z(x_n)$.
- 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?
  1. Permute the locations $x_1, \ldots, x_n$ 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

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

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

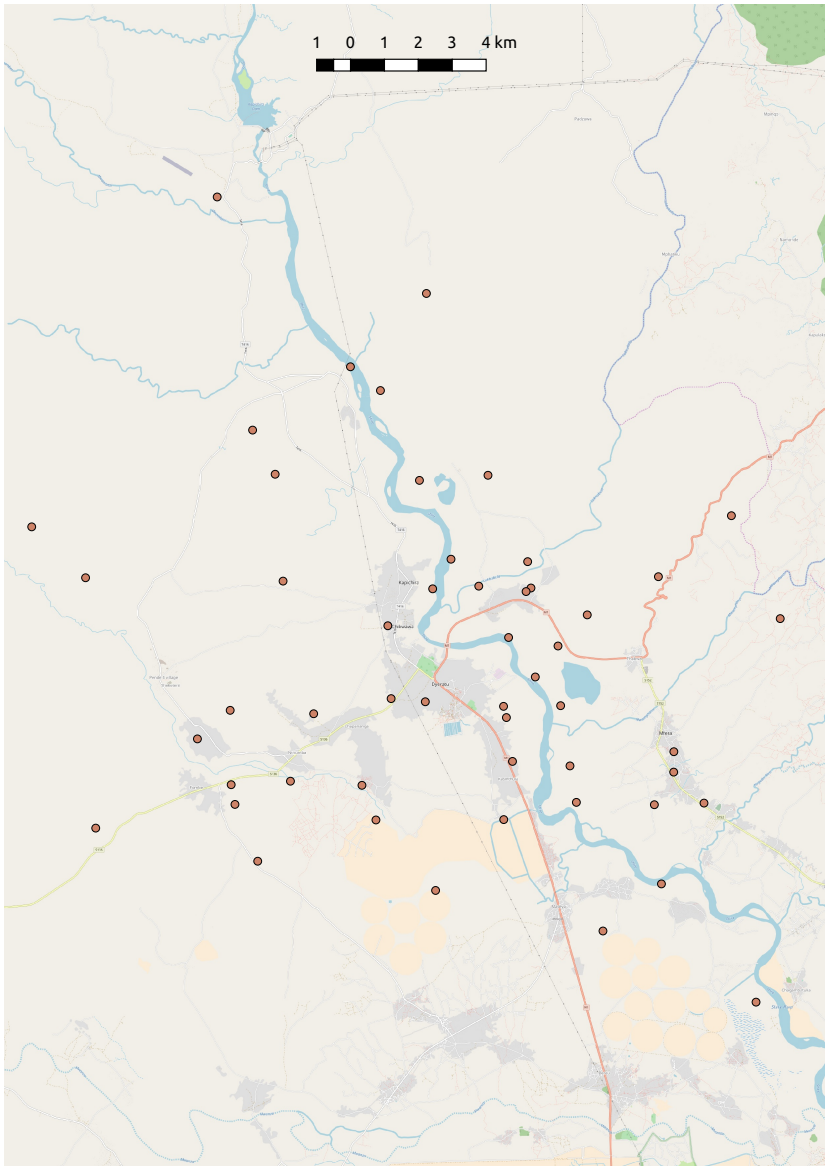
- How do we know if $\hat{v}(u)$ shows evidence of spatial correlation?
  1. Permute the locations $x_1, \ldots, x_n$ while holding fix the rest of the data.
  2. Compute $\hat{v}(u)$ for the permuted data.
  3. Repeat 1 and 2 a large enough number of times (e.g. 1,000).
  4. Compute the 95% confidence intervals (CIs) at each binned distance u.
- If the empirical variogram falls within the 95% CIs.

Example: Anaemia mapping in Chikwawa, Malawi

- `Pf`: RDT test for *Plasmodium falciparum* (1=positive, 0=negative).
- `Hb`: haemoglobin concentration (g/dl)
- `ITN`: did you sleep under an ITN last night? (1=Yes, 0=No)
- `IRS`: was there IRS in the house during the last 12 months? (1=Yes, 0=No)
- `age.months`: age in months.
- `MotherEdu`: maternal education (none, primary, secondary and above)
- `quintile`: socio-economic-status (1=poor to 5=rich)
- `febrile`: febrile on examination (1=Yes, 0=No)
- `Season`: indicator of the season (dry, rainy).
- `web_y`: y-coordinated of the household (Web Mercator projection)
- `web_x`: x-coordinated of the household (Web Mercator projection)
- `sex`: gender of the individual
- `ID`: ID of the household

Part II: The linear geostatistical model

Where we observe matters



Geostatistical modelling of Hb

Geostatistical modelling of Hb

- Y_{ij} =Hb concentration for the j-th individual at household j.

Geostatistical modelling of Hb

- Y_{ij} =Hb concentration for the j-th individual at household j.
- d_{ij} = sex

Geostatistical modelling of Hb

- Y_{ij} =Hb concentration for the j-th individual at household j.
- d_{ij} = sex

$$Y_{ij} = \alpha + \beta d_{ij} + S(x_i) + Z_{ij}$$

Geostatistical modelling of Hb

- Y_{ij} =Hb concentration for the j-th individual at household j.
- d_{ij} = sex

$$Y_{ij} = \alpha + \beta d_{ij} + S(x_i) + Z_{ij}$$

Geostatistical modelling of Hb

- Y_{ij} =Hb concentration for the j-th individual at household j.
- d_{ij} = sex

$$Y_{ij} = \alpha + \beta d_{ij} + S(x_i) + Z_{ij}$$

- How do we model $S(x)$?

Geostatistical modelling of Hb

- Y_{ij} =Hb concentration for the j-th individual at household j.
- d_{ij} = sex

$$Y_{ij} = \alpha + \beta d_{ij} + S(x_i) + Z_{ij}$$

- How do we model $S(x)$?
- We assume that $S(x)$ is a stationary and isotropic Gaussian process, i.e.

$$Cov(S(x), S(x')) = \sigma^2 \rho(u), \quad u = \|x - x'\|.$$

Geostatistical modelling of Hb

- Y_{ij} =Hb concentration for the j-th individual at household j.
- d_{ij} = sex

$$Y_{ij} = \alpha + \beta d_{ij} + S(x_i) + Z_{ij}$$

- How do we model $S(x)$?
- We assume that $S(x)$ is a stationary and isotropic Gaussian process, i.e.

$$Cov(S(x), S(x')) = \sigma^2 \rho(u), \quad u = \|x - x'\|.$$

- How do we choose $\rho(\cdot)$?

Geostatistical modelling of Hb

- Y_{ij} =Hb concentration for the j-th individual at household j.
- d_{ij} = sex

$$Y_{ij} = \alpha + \beta d_{ij} + S(x_i) + Z_{ij}$$

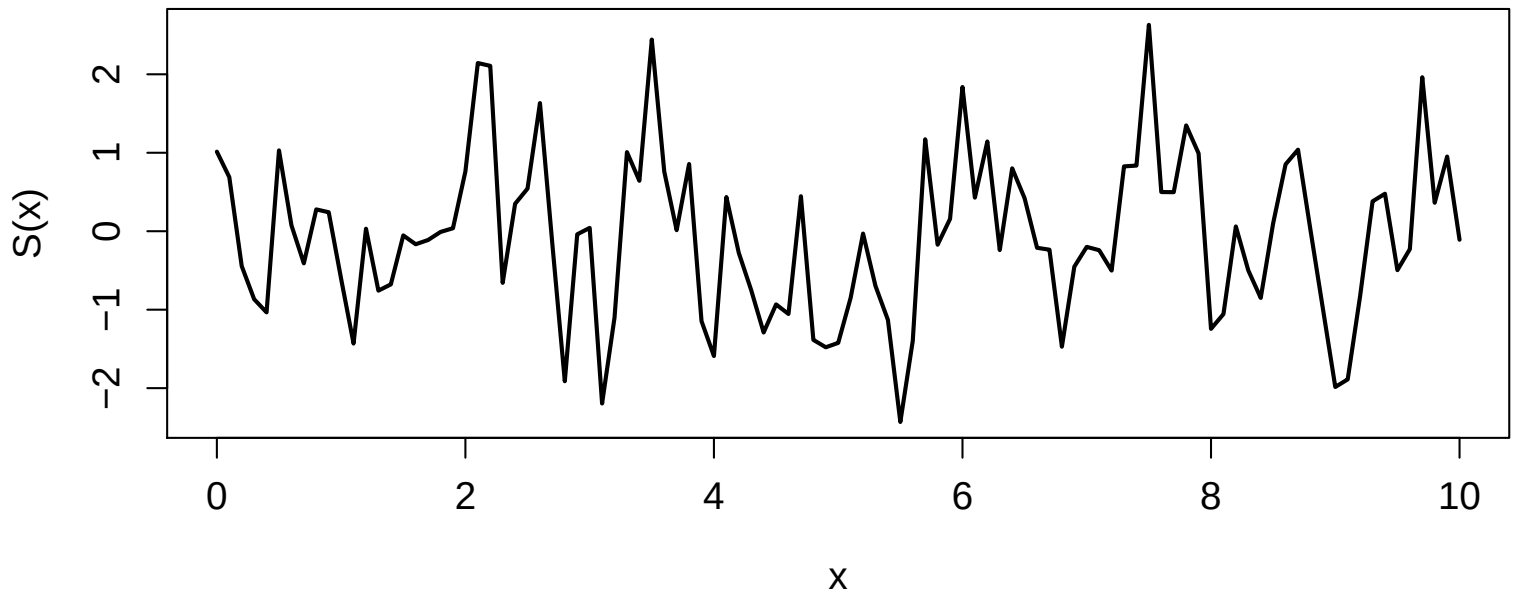
- How do we model $S(x)$?
- We assume that $S(x)$ is a stationary and isotropic Gaussian process, i.e.

$$Cov(S(x), S(x')) = \sigma^2 \rho(u), \quad u = \|x - x'\|.$$

- How do we choose $\rho(\cdot)$?
- Example: $\rho(u) = \exp\{-u/\phi\}$

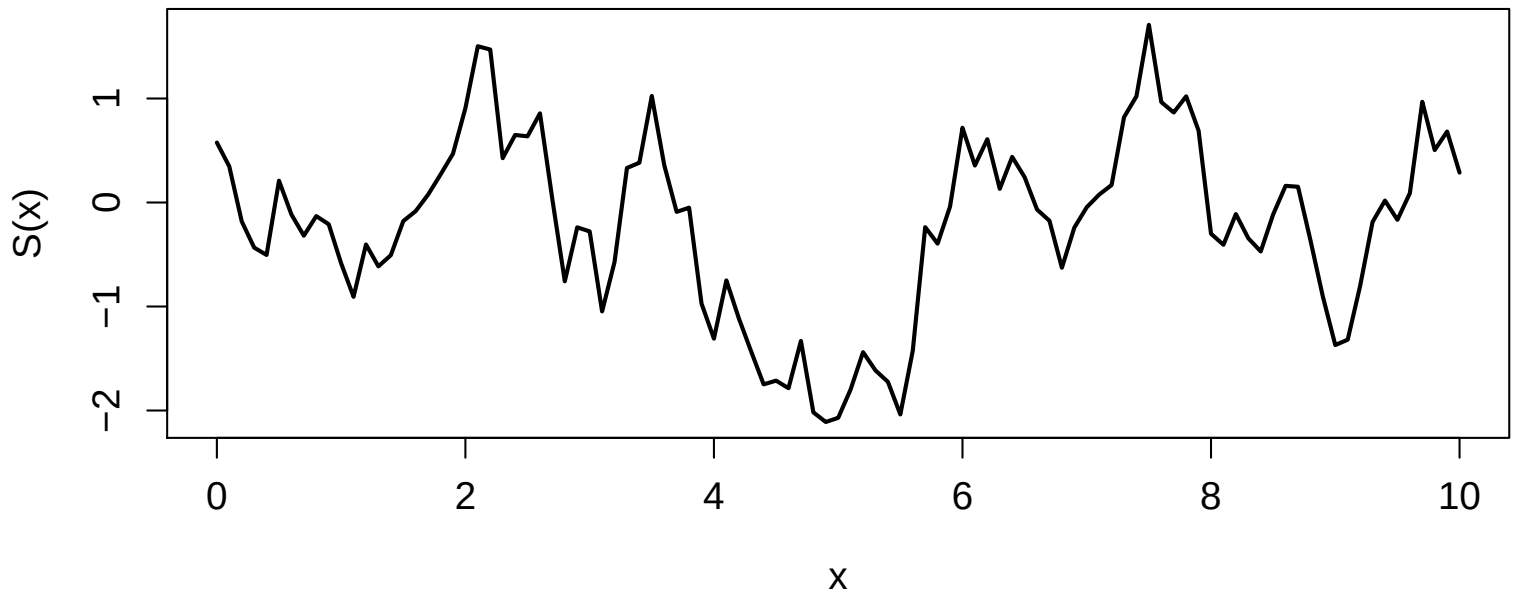
The exponential correlation function

$$\phi = 0.1$$



The exponential correlation function

$$\phi = 1$$



The likelihood function

The likelihood function

- $Z_{ij} \sim N(0, \omega^2)$, i.i.d. for all (i, j).

The likelihood function

- $Z_{ij} \sim N(0, \omega^2)$, i.i.d. for all (i, j).
- $Y_{ij}|S(x_i) \sim N(\alpha + \beta S(x_i), \omega^2)$.

The likelihood function

- $Z_{ij} \sim N(0, \omega^2)$, i.i.d. for all (i, j).
- $Y_{ij}|S(x_i) \sim N(\alpha + \beta S(x_i), \omega^2)$.
- What is the (unconditional) distribution of Y_{ij}?

The likelihood function

- $Z_{ij} \sim N(0, \omega^2)$, i.i.d. for all (i, j).
- $Y_{ij}|S(x_i) \sim N(\alpha + \beta S(x_i), \omega^2)$.
- What is the (unconditional) distribution of Y_{ij}?
- $\theta = (\alpha, \beta, \sigma^2, \phi)$

The likelihood function

- $Z_{ij} \sim N(0, \omega^2)$, i.i.d. for all (i, j).
- $Y_{ij}|S(x_i) \sim N(\alpha + \beta S(x_i), \omega^2)$.
- What is the (unconditional) distribution of Y_{ij}?
- $\theta = (\alpha, \beta, \sigma^2, \phi)$

$$\begin{aligned} L(\theta) &= [Y_{ij}, i = 1, \ldots, n, j = 1, \ldots, m_i] \\ &= MVN(\mu, \Sigma) \end{aligned}$$

The likelihood function

- $Z_{ij} \sim N(0, \omega^2)$, i.i.d. for all (i, j).
- $Y_{ij}|S(x_i) \sim N(\alpha + \beta S(x_i), \omega^2)$.
- What is the (unconditional) distribution of Y_{ij}?
- $\theta = (\alpha, \beta, \sigma^2, \phi)$

$$\begin{aligned} L(\theta) &= [Y_{ij}, i = 1, \ldots, n, j = 1, \ldots, m_i] \\ &= MVN(\mu, \Sigma) \end{aligned}$$

The likelihood function

- $Z_{ij} \sim N(0, \omega^2)$, i.i.d. for all (i, j).
- $Y_{ij}|S(x_i) \sim N(\alpha + \beta S(x_i), \omega^2)$.
- What is the (unconditional) distribution of Y_{ij}?
- $\theta = (\alpha, \beta, \sigma^2, \phi)$

$$\begin{aligned} L(\theta) &= [Y_{ij}, i = 1, \ldots, n, j = 1, \ldots, m_i] \\ &= MVN(\mu, \Sigma) \end{aligned}$$

- To estimate θ, we maximize the likelihood.

The likelihood function

- $Z_{ij} \sim N(0, \omega^2)$, i.i.d. for all (i, j).
- $Y_{ij}|S(x_i) \sim N(\alpha + \beta S(x_i), \omega^2)$.
- What is the (unconditional) distribution of Y_{ij}?
- $\theta = (\alpha, \beta, \sigma^2, \phi)$

$$\begin{aligned} L(\theta) &= [Y_{ij}, i = 1, \ldots, n, j = 1, \ldots, m_i] \\ &= MVN(\mu, \Sigma) \end{aligned}$$

- To estimate θ, we maximize the likelihood.
- $\hat{\theta}$ denotes the maximum likelihood estimate of θ.

Validation of $\rho(\cdot)$

Validation of $\rho(\cdot)$

1. Obtain θ.

Validation of $\rho(\cdot)$

1. Obtain θ.
2. Simulate a large enough number of data-sets (e.g. 1,000) from the fitted model.

Validation of $\rho(\cdot)$

1. Obtain θ.
2. Simulate a large enough number of data-sets (e.g. 1,000) from the fitted model.
3. Fit a non-spatial mixed model (i.e. $S(x_i)$ is replaced by unstructured random effects U_i).

Validation of $\rho(\cdot)$

1. Obtain θ.
2. Simulate a large enough number of data-sets (e.g. 1,000) from the fitted model.
3. Fit a non-spatial mixed model (i.e. $S(x_i)$ is replaced by unstructured random effects U_i).
4. Compute the variogram using the estimated random effects associated with a location (i.e. $\hat{U}_i$).

Validation of $\rho(\cdot)$

1. Obtain θ.
2. Simulate a large enough number of data-sets (e.g. 1,000) from the fitted model.
3. Fit a non-spatial mixed model (i.e. $S(x_i)$ is replaced by unstructured random effects U_i).
4. Compute the variogram using the estimated random effects associated with a location (i.e. $\hat{U}_i$).
5. Repeat 3 to 4, for all the simulated data-sets.

Validation of $\rho(\cdot)$

1. Obtain θ.
2. Simulate a large enough number of data-sets (e.g. 1,000) from the fitted model.
3. Fit a non-spatial mixed model (i.e. $S(x_i)$ is replaced by unstructured random effects U_i).
4. Compute the variogram using the estimated random effects associated with a location (i.e. $\hat{U}_i$).
5. Repeat 3 to 4, for all the simulated data-sets.
6. Summarize the results by computing the 95% confidence intervals for each distance u.

Validation of $\rho(\cdot)$

1. Obtain θ.
2. Simulate a large enough number of data-sets (e.g. 1,000) from the fitted model.
3. Fit a non-spatial mixed model (i.e. $S(x_i)$ is replaced by unstructured random effects U_i).
4. Compute the variogram using the estimated random effects associated with a location (i.e. $\hat{U}_i$).
5. Repeat 3 to 4, for all the simulated data-sets.
6. Summarize the results by computing the 95% confidence intervals for each distance u.

The generated 95% bandwidth from the last step indicates the band of variation for the variogram under the assumed model.

Spatial prediction

Spatial prediction

- $T(x)$ is a prediction target at location x.

Spatial prediction

- $T(x)$ is a prediction target at location x.
- Example:
  1. $T(x) = \alpha + S(x)$ (Hb of a female child)
  2. $T(x) = \alpha + S(x) + Z$ (Hb of this female child)

Spatial prediction

- $T(x)$ is a prediction target at location x.
- Example:
  1. $T(x) = \alpha + S(x)$ (Hb of a female child)
  2. $T(x) = \alpha + S(x) + Z$ (Hb of this female child)
- **Problem:** Prediction of $T(x)$ at an unobserved location x.

Spatial prediction

- $T(x)$ is a prediction target at location x.
- Example:
  1. $T(x) = \alpha + S(x)$ (Hb of a female child)
  2. $T(x) = \alpha + S(x) + Z$ (Hb of this female child)
- **Problem:** Prediction of $T(x)$ at an unobserved location x.
- **Solution:** The predictive distribution of $T(x)$, i.e. $[T(x)|\text{data}]$.

Spatial prediction

- $T(x)$ is a prediction target at location x.
- Example:
  1. $T(x) = \alpha + S(x)$ (Hb of a female child)
  2. $T(x) = \alpha + S(x) + Z$ (Hb of this female child)
- **Problem:** Prediction of $T(x)$ at an unobserved location x.
- **Solution:** The predictive distribution of $T(x)$, i.e. $[T(x)|\text{data}]$.
- Consider a linear model and let $T(x) = S(x)$. The kriging predictor for $T(x)$ is

$$\hat{T}(x) = \arg\max_{T} E[(T(x) - T)^2|\text{data}] = E[T(x)|\text{data}].$$

Spatial prediction

- $T(x)$ is a prediction target at location x.
- Example:
  1. $T(x) = \alpha + S(x)$ (Hb of a female child)
  2. $T(x) = \alpha + S(x) + Z$ (Hb of this female child)
- **Problem:** Prediction of $T(x)$ at an unobserved location x.
- **Solution:** The predictive distribution of $T(x)$, i.e. $[T(x)|\text{data}]$.
- Consider a linear model and let $T(x) = S(x)$. The kriging predictor for $T(x)$ is

$$\hat{T}(x) = \arg\max_{T} E[(T(x) - T)^2|\text{data}] = E[T(x)|\text{data}].$$

- $T(x) = f\{S(x)\}$, where $f\{\cdot\}$ is non-linear (e.g. $f(a) = \exp\{a\}$).

Spatial prediction

- $T(x)$ is a prediction target at location x.
- Example:
  1. $T(x) = \alpha + S(x)$ (Hb of a female child)
  2. $T(x) = \alpha + S(x) + Z$ (Hb of this female child)
- **Problem:** Prediction of $T(x)$ at an unobserved location x.
- **Solution:** The predictive distribution of $T(x)$, i.e. $[T(x)|\text{data}]$.
- Consider a linear model and let $T(x) = S(x)$. The kriging predictor for $T(x)$ is

$$\hat{T}(x) = \arg\max_{T} E[(T(x) - T)^2|\text{data}] = E[T(x)|\text{data}].$$

- $T(x) = f\{S(x)\}$, where $f\{\cdot\}$ is non-linear (e.g. $f(a) = \exp\{a\}$).

$$\hat{T}(x) = E[f\{S(x)\}|\text{data}] \approx \frac{1}{B}\sum_{j=1}^{B} f\{S_{(j)}(x)\}$$

Part III: The binomial geostatistical model

Gostatistical modelling of prevalence data

Gostatistical modelling of prevalence data

- x_i = location of a village/community/household

Gostatistical modelling of prevalence data

- x_i = location of a village/community/household
- y_{ij} = test outcome (1=positive, 0=negative) for j-th individual at x_i

Gostatistical modelling of prevalence data

- x_i = location of a village/community/household
- y_{ij} = test outcome (1=positive, 0=negative) for j-th individual at x_i
- n_i = total number of sampled individuals

Gostatistical modelling of prevalence data

- x_i = location of a village/community/household
- y_{ij} = test outcome (1=positive, 0=negative) for j-th individual at x_i
- n_i = total number of sampled individuals
- $d(x_i)$ = vector of spatially-referenced variables

Gostatistical modelling of prevalence data

- x_i = location of a village/community/household
- y_{ij} = test outcome (1=positive, 0=negative) for j-th individual at x_i
- n_i = total number of sampled individuals
- $d(x_i)$ = vector of spatially-referenced variables
- e_{ij} = individual characteristics

Assumptions:

1. $S(x)$ is a stationary and isotropic Gaussian process;
2. Z_i are i.i.d. random variables;
3. $Y_{ij}|S(x_i), Z_i$ are mutually independent Bernoulli$(n_i, p_j(x_i))$ such that

$$\log\left\{\frac{p_j(x_i)}{1-p_j(x_i)}\right\} = \alpha + d(x_i)^\top\beta + e_{ij}^\top\gamma + S(x_i) + Z_i$$

The likelihood function

- The likelihood function is obtained by integrating out all random effects from the model.

The likelihood function

- The likelihood function is obtained by integrating out all random effects from the model.

$$
\begin{aligned}
L(\theta) &= P(Y_i = y_i; \theta) \\
&= \int \int [S(x_i), Z_i, Y_i; \theta] \, dS(x_i) \, dZ_i \\
&= \int \int [S(x_i)][Z_i][Y_i | S(x_i), Z_i] \, dS(x_i) \, dZ_i
\end{aligned}
$$

The likelihood function

- The likelihood function is obtained by integrating out all random effects from the model.

$$\begin{aligned} L(\theta) &= P(Y_i = y_i; \theta) \\ &= \int \int [S(x_i), Z_i, Y_i; \theta] \, dS(x_i) \, dZ_i \\ &= \int \int [S(x_i)][Z_i][Y_i | S(x_i), Z_i] \, dS(x_i) \, dZ_i \end{aligned}$$

- The above integral cannot be solved in closed form.

Monte Carlo maximum likelihood (1)

Monte Carlo maximum likelihood (1)

- Let $W = S(x) + Z$.

Monte Carlo maximum likelihood (1)

- Let $W = S(x) + Z$.

$$L(\theta) = \int [W, Y; \theta] \, dW$$

Monte Carlo maximum likelihood (1)

- Let $W = S(x) + Z$.

$$\begin{aligned} L(\theta) &= \int [W, Y; \theta] \, dW \\ &= \int \frac{[W, Y; \theta]}{[W, Y; \theta_0]} [W, Y; \theta_0] \, dW \end{aligned}$$

Monte Carlo maximum likelihood (1)

- Let $W = S(x) + Z$.

$$
\begin{aligned}
L(\theta) &= \int [W, Y; \theta] \, dW \\
&= \int \frac{[W, Y; \theta]}{[W, Y; \theta_0]} [W, Y; \theta_0] \, dW \\
&= \int \frac{[W, Y; \theta]}{[W, Y; \theta_0]} [Y; \theta_0][W|Y; \theta_0] \, dW
\end{aligned}
$$

Monte Carlo maximum likelihood (1)

- Let $W = S(x) + Z$.

$$
\begin{aligned}
L(\theta) &= \int [W, Y; \theta]\, dW \\
&= \int \frac{[W, Y; \theta]}{[W, Y; \theta_0]} [W, Y; \theta_0]\, dW \\
&= \int \frac{[W, Y; \theta]}{[W, Y; \theta_0]} [Y; \theta_0][W|Y; \theta_0]\, dW \\
&\propto \int \frac{[W, Y|W; \theta]}{[W, Y|W; \theta_0]} [W|Y; \theta_0]\, dW
\end{aligned}
$$

Monte Carlo maximum likelihood (1)

- Let $W = S(x) + Z$.

$$
\begin{aligned}
L(\theta) &= \int [W, Y; \theta]\, dW \\
&= \int \frac{[W, Y; \theta]}{[W, Y; \theta_0]} [W, Y; \theta_0]\, dW \\
&= \int \frac{[W, Y; \theta]}{[W, Y; \theta_0]} [Y; \theta_0][W|Y; \theta_0]\, dW \\
&\propto \int \frac{[W, Y|W; \theta]}{[W, Y|W; \theta_0]} [W|Y; \theta_0]\, dW \\
&= E_{[W|Y;\theta_0]} \left\{ \frac{[W, Y; \theta]}{[W, Y; \theta_0]} \right\} \approx \frac{1}{B} \sum_{j=1}^{B} \frac{[w_{(j)}, Y; \theta]}{[w_{(j)}, Y; \theta_0]} = L_B(\theta)
\end{aligned}
$$

Monte Carlo maximum likelihood (2)

1. Initialize θ_0.
2. Simulate B samples $w_{(j)}$ from $[W|Y;\theta_0]$.
3. Maximize $L_B(\theta)$ with respect to θ to obtain $\hat{\theta}_B$.
4. Set $\theta_0 = \hat{\theta}_B$ and repeat 1 to 3.

Monte Carlo maximum likelihood (2)

1. Initialize θ_0.
2. Simulate B samples $w_{(j)}$ from $[W|Y;\theta_0]$.
3. Maximize $L_B(\theta)$ with respect to θ to obtain $\hat{\theta}_B$.
4. Set $\theta_0 = \hat{\theta}_B$ and repeat 1 to 3.

Exercise: Consider the binary outcome

$$Y_{ij} = \begin{cases} 1 & \text{if Hb} < 12 \text{ g/dl} \\ 0 & \text{otherwise} \end{cases}.$$

1. Formulate a geostatistical model for the data.
2. Fit the model using Monte Carlo maximum likelihood.

Spatial prediction

Spatial prediction

- $T(x)$ is a prediction target at location x.

Spatial prediction

- $T(x)$ is a prediction target at location x.
- Example:
  1. $T(x) = \exp\{\alpha + S(x)\}/(1 + \exp\{\alpha + S(x)\})$ (prevalence)
  2. $T(x) = \exp\{\alpha + S(x)\}$ (odds)

Spatial prediction

- $T(x)$ is a prediction target at location x.
- Example:
  1. $T(x) = \exp\{\alpha + S(x)\}/(1 + \exp\{\alpha + S(x)\})$ (prevalence)
  2. $T(x) = \exp\{\alpha + S(x)\}$ (odds)
- **Problem:** Prediction of $T(x)$ at an unobserved location x.

Spatial prediction

- $T(x)$ is a prediction target at location x.
- Example:
  1. $T(x) = \exp\{\alpha + S(x)\}/(1 + \exp\{\alpha + S(x)\})$ (prevalence)
  2. $T(x) = \exp\{\alpha + S(x)\}$ (odds)
- **Problem:** Prediction of $T(x)$ at an unobserved location x.
- **Solution:** The predictive distribution of $T(x)$, i.e. $[T(x)|\text{data}]$.

Spatial prediction

- $T(x)$ is a prediction target at location x.
- Example:
  1. $T(x) = \exp\{\alpha + S(x)\}/(1 + \exp\{\alpha + S(x)\})$ (prevalence)
  2. $T(x) = \exp\{\alpha + S(x)\}$ (odds)
- **Problem:** Prediction of $T(x)$ at an unobserved location x.
- **Solution:** The predictive distribution of $T(x)$, i.e. $[T(x)|\text{data}]$. Let $S = (S(x_1), \ldots, S(x_n))$ (spatial random effects at observed locations) and $y = (y_1, \ldots, y_n)$ (the data)

$$[S, T(x)|y] = [S|y][T(x)|S, y] = [S|y][T(x)|S]$$

- **Exercise:** Predict the prevalence and odds of $Hb < 12$.