Model-based Geostatistics: geospatial statistical methods for public health applications

Peter J Diggle and Emanuele Giorgi

CHICAS, Lancaster University

Overview: spatial epidemiology and model-based geostatistics

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);

Epidemic vs endemic patterns of incidence

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

  Diggle (2006)

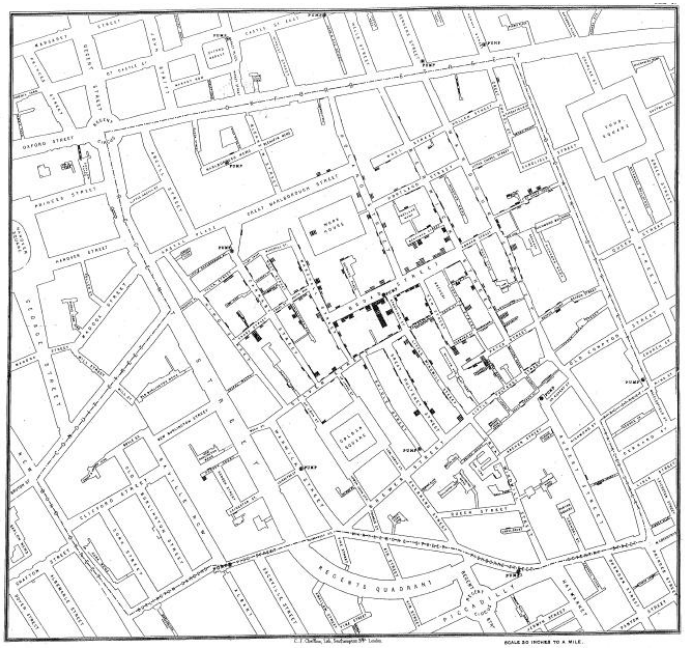
- **Gastro-enteric disease in Hampshire (AEGISS)**

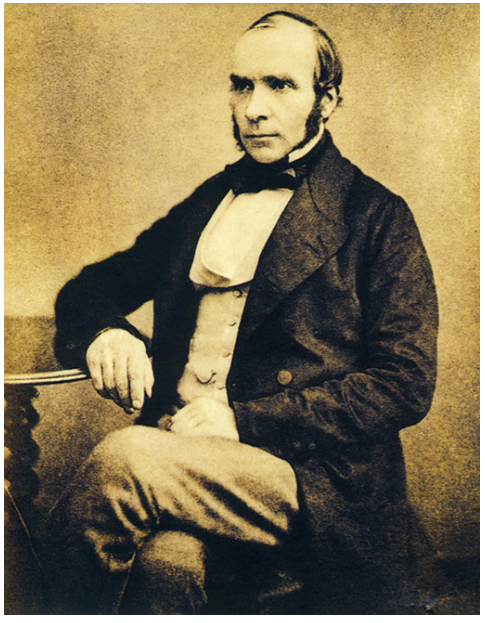
  Diggle, Rowlingson and Su (2005)

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

What are the similarities and differences between the two phenomena?

In the beginning: Cholera in Victorian London, 1854





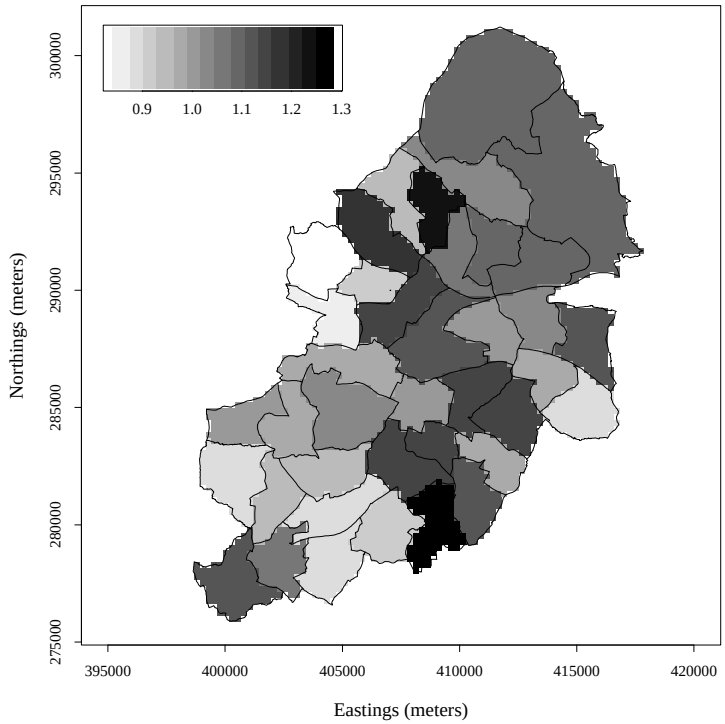
The physician John Snow famously removed the handle of the Broad Street water-pump, having concluded (correctly) that infected water was the source of the disease contrary to conventional wisdom at the time.

`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

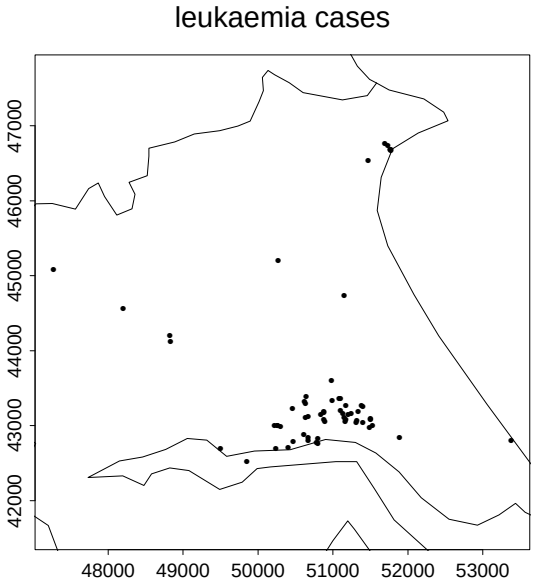
Registry example. Colorectal cancer in Birmingham

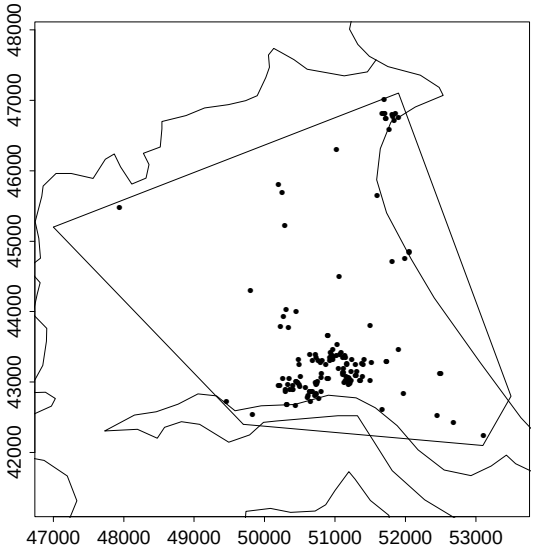


Smoothed estimates of relative risk in 36 electoral wards.

Kelsall and Wakefield (2002).

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

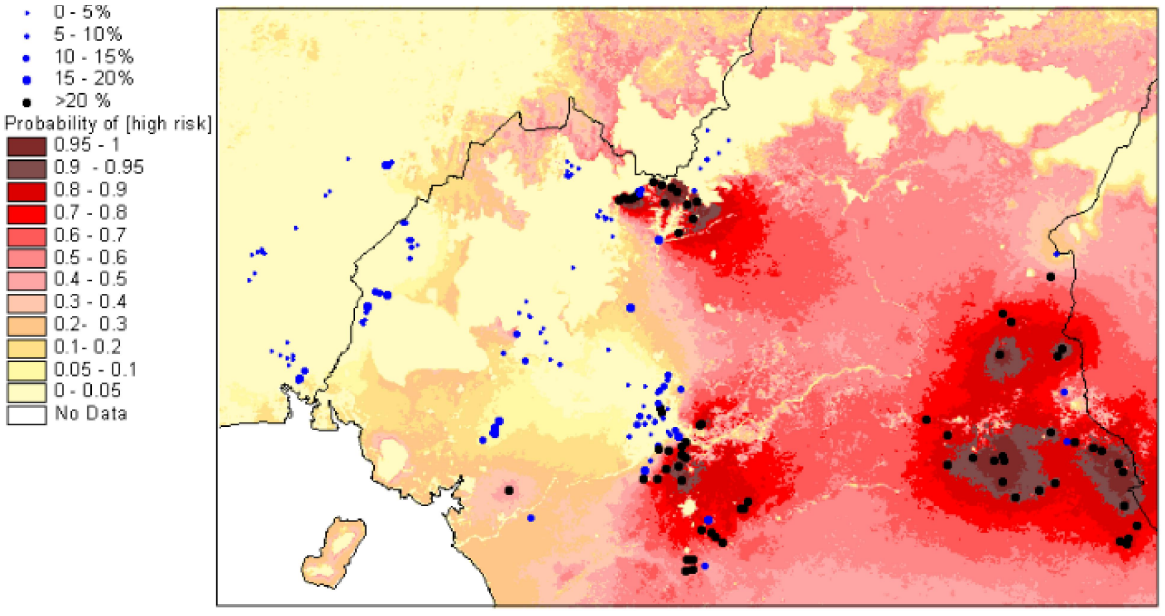


Figure 6: PCM for [high risk] in Cameroon based on ERMr with ground truth data.

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. **Colorectal cancer in Birmingham**
   - **Does the risk of contracting the disease vary spatially?**
   - **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?**

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 $\{(S_i, x_i) : i = 1, \ldots, n\}$**
   - **point processes: the random variables are the locations themselves, $\{x_i : i = 1, \ldots, n\}$**
   - **continuous spatial variation: the random variables associate a real value with every point in the space, hence $\{S(x) : x \in \mathbb{R}^2\}$**

This course covers continuous spatial variation, with a focus on its application to prevalence mapping

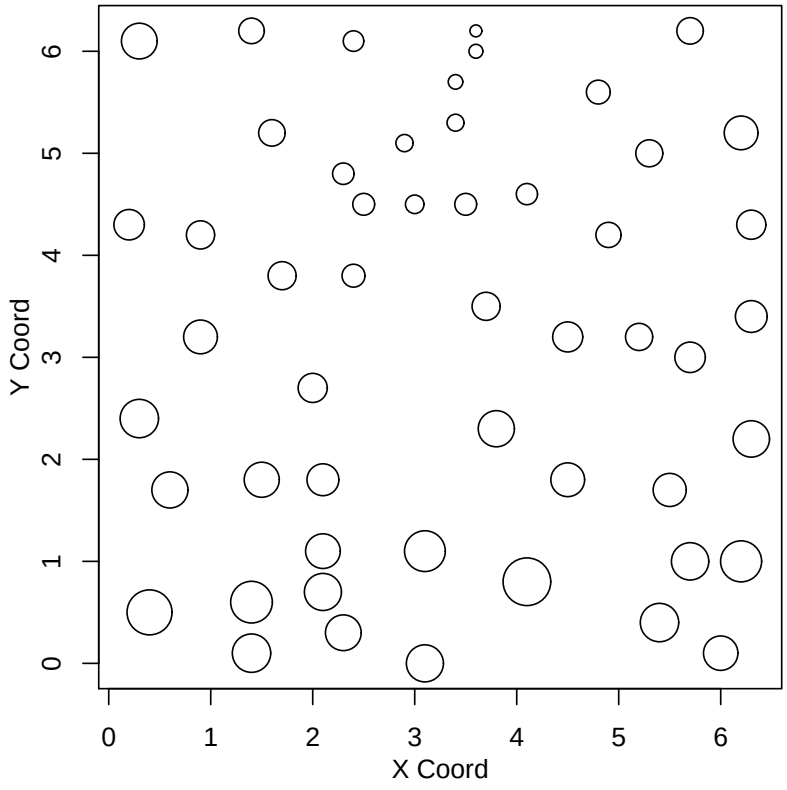
Geostatistics

- **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 which 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

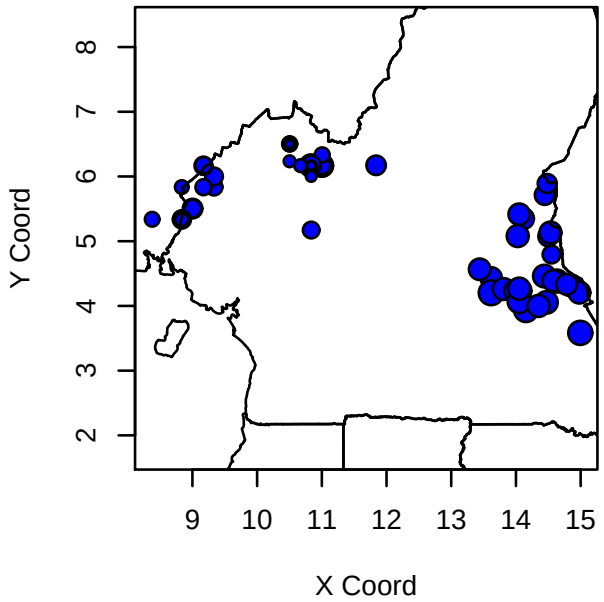
Diggle, Moyeed and Tawn (1998)

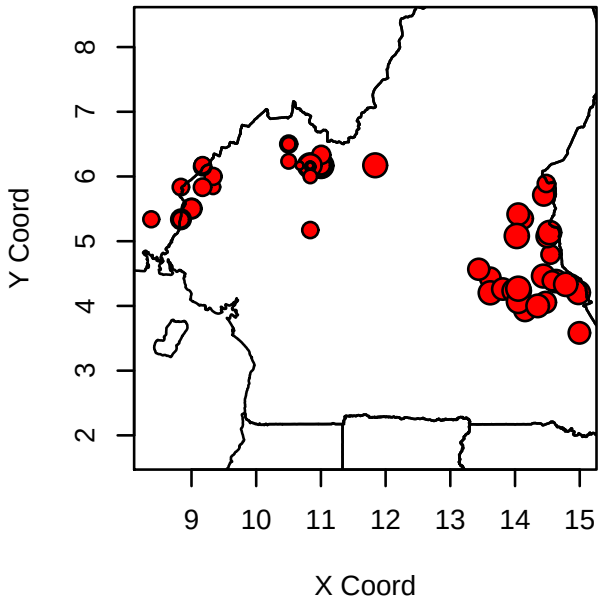
Measured surface elevations

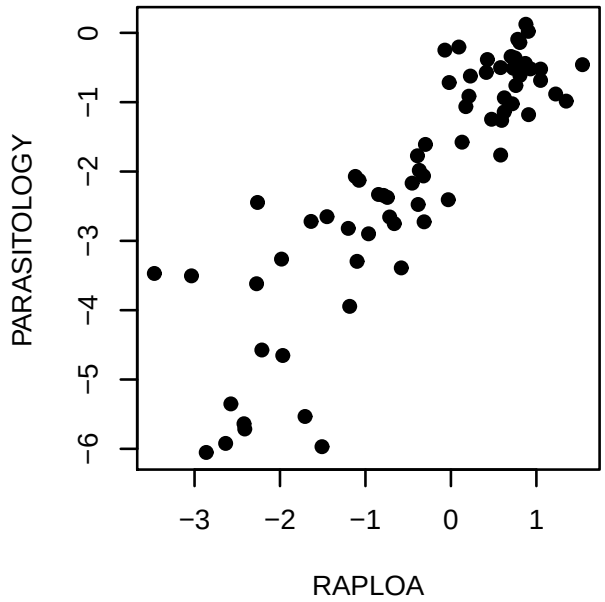


- **simplest form of geostatistical data**
- **locations and associated measurements**
- **no covariates**

Loa loa prevalence surveys



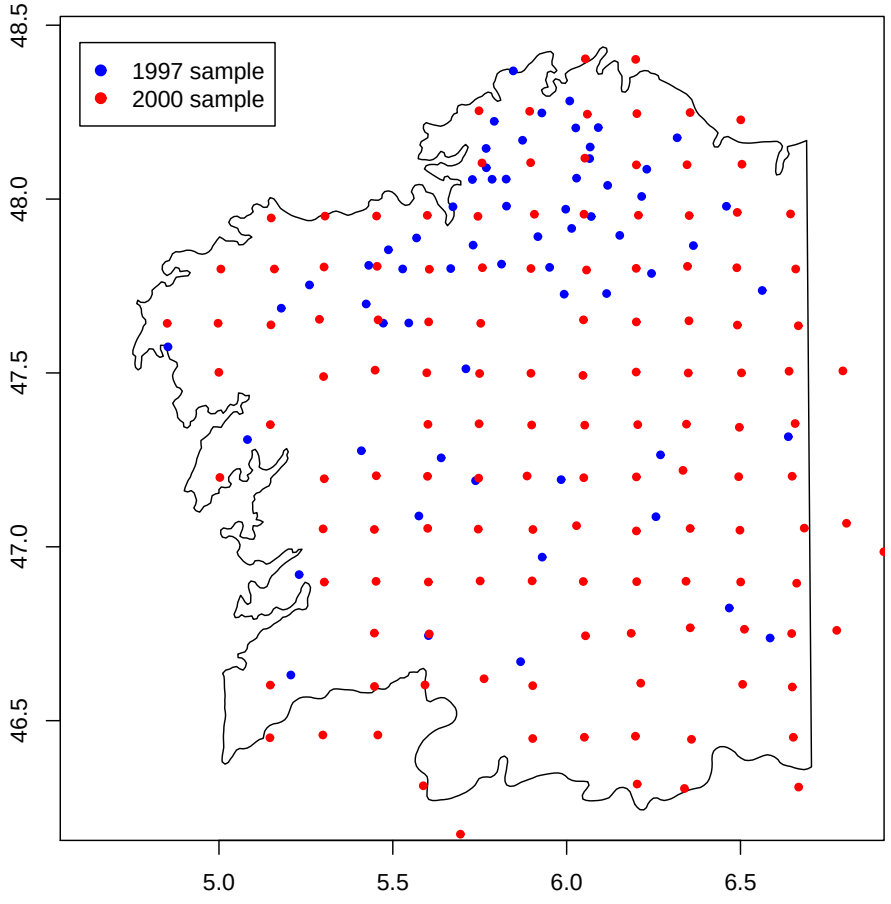




Maps show prevalence as estimated by each of two methods:

- **parasitology (blood sample, red)**
- **simple questionnaire (RAPLOA, blue)**
- **covariates include elevation and NDVI (green-ness of vegetation) at 1km resolution**

Environmental monitoring in Galicia, north-west Spain



- **spatially irregular sample in 1997**
- **potential for selection bias?**

Geostatistical problems

- **Design:** how to choose locations x_i at which to collect outcome data?
- **Estimation:** how to investigate relationship between outcome and covariates when data may be spatially correlated?
- **Prediction:** how to map (expected value of) outcome throughout the study-region?

Practical point:

- **Estimation** only requires covariate information at locations x_i
- **Prediction** requires covariate information throughout the study-region.

A non-spatial model for prevalence survey data

Design

- **Sample communities $i = 1, \ldots, n$.**
- **In community i, sample m_i individuals of whom Y_i test positive for disease of interest.**
- **Associated covariates w_i**

Model

- **ρ_i = probability that a randomly sampled individual in community i will test positive**
- **$\log\{\rho_i/(1-\rho_i)\} = \alpha + w_i'\beta$**
- **$Y_i \sim \text{Binomial}(m_i, \rho_i)$, mutually independent**

A spatial model for prevalence survey data

Design

- **Sample communities $i = 1, \ldots, n$ at locations x_i**
- **In community i, sample m_i individuals of whom Y_i test positive for disease of interest.**
- **Associated covariates $w_i = w(x_i)$**

Model

- ρ_i **= probability that a randomly sampled individual in community i will test positive**
- $\log\{\rho_i/(1-\rho_i)\} = \alpha + w(x_i)'\beta + S(x_i)$
- $Y_i \sim \mathrm{Binomial}(m_i, \rho_i)$**, conditionally independent given** $S(\cdot)$

A spatial model for prevalence survey data (continued)

Two kinds of covariates

- **$w(x_i)$ an intrinsic property of the location x_i**
- **$w(x_i)$ a property of the people who live at location x_i**

Practical implication: when mapping prevalence we need to be able to assign a value w(x) to every location in the study-region.

What is S(x)?

- **an unobserved spatially varying stochastic process**
- **a proxy for unmeasured, spatially structured covariates**

Practical implication: in any application where S(x) turns out to be important, it is worth asking what the missing covariate(s) might be.

Person or place?

Extend spatial model to

$$\log\{\rho_i/(1-\rho_i)\} = \alpha + \{w(x_i)'\beta + S(x_i)\} + \{d_i'\gamma + U_i\}$$

- **w(x) : measured properties of location x**
- **S(x) : stochastic process, proxy for unmeasured properties of x**
- **d_i : measured properties of ith community**
- **U_i : independent random variables, proxy for unmeasured properties of ith community**

References

Cuzick, J. and Edwards, R. (1990). Spatial clustering for inhomogeneous populations (with Discussion). *Journal of the Royal Statistical Society*, B **52**, 73–104.

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

Diggle, P.J. and Chetwynd, A.G. (1991). Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics*, **47**, 1155–63.

Diggle, P.J. and Giorgi, E. (2015). Model-based geostatistics for prevalence Mapping in low-resource settings (with Discussion). *Journal of the American Statistical Association* (to appear).

Diggle, P.J., Moraga, P., Rowlingson, B. and Taylor, B. (2013). Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. *Statistical Science*, **28**, 542–563.

Diggle, P.J., Moyeed, R.A. and Tawn, J.A. (1998). Model-based Geostatistics (with Discussion). *Applied Statistics* **47** 299–350.
Diggle, P., Rowlingson, B. and Su, T. (2005). Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics*, **16**, 423–34.
Diggle, P.J., Thomson, M.C., Christensen, O.F., Rowlingson, B., Obsomer, V., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Kamgno, J., Remme, H., Boussinesq, M. and Molyneux, D.H. (2007). Spatial modelling and prediction of Loa loa risk: decision making under uncertainty. *Annals of Tropical Medicine and Parasitology*, **101**, 499–509.

Elliott, P., Wakefield, J.C., Best, N.G. and Briggs, D.J. (2000). *Spatial Epidemiology*. Oxford: Oxford University Press

Gelfand, A., Diggle, P.J., Fuentes, M. and Guttorp, P. (2010). *Handbook of Spatial Statistics*. Boca Raton: CRC Press.

Giorgi, E., Sesay, S.S., Terlouw, D.J. and Diggle, P.J. (2015). Combining data from multiple spatially referenced prevalence surveys using generalized linear geostatistical models. *Journal of the Royal Statistical Society* A **178**, 445–464

Kelsall. J.E. and Wakefield, J.C. (2002). Modelling spatial variation in disease risk: a geostatistical approach. *Journal of the American Statistical Association*, **97**, 692–701.

Rothman, K.J. (1986). *Modern Epidemiology*. Boston: Little and Brown

Waller, L.and Gotway, C.A. (2004). *Applied Spatial Statistics for Public Health Data*. New York: Wiley.

Woodward, M. (1999). *Epidemiology: study design and data analysis*. Boca Raton: Chapman and Hall.

Zoure, H.G.M., Noma, M., Tekle, A.H., Amazigo, U.V., Diggle, P.J., Giorgi, E. and Remme, J.H.F. (2014). The geographic distribution of onchocerciasis in the 20 participating countries of the African Programme for Onchocerciasis Control: 2. Pre-control endemicity Levels and estimated number infected. *Parasites and Vectors*, **7**, 326