APTS 2013/14: de-classifying spatial statistics

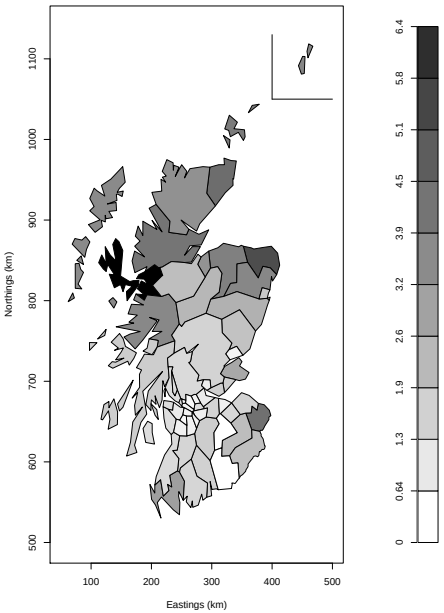
Peter J Diggle

Lancaster University and University of Liverpool

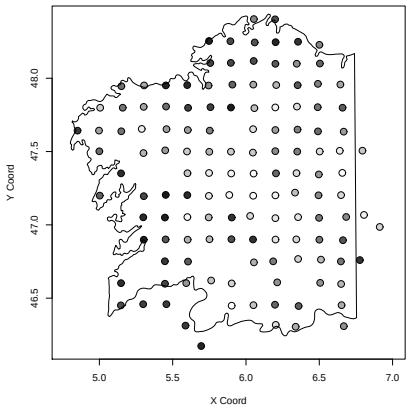
UNIVERSITY OF
LIVERPOOL

INSTITUTE OF INFECTION
AND GLOBAL HEALTH

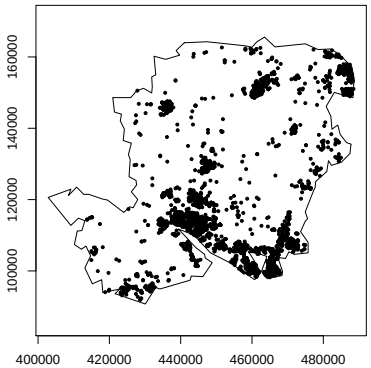
Spatial statistics according to Cressie



Lattice data

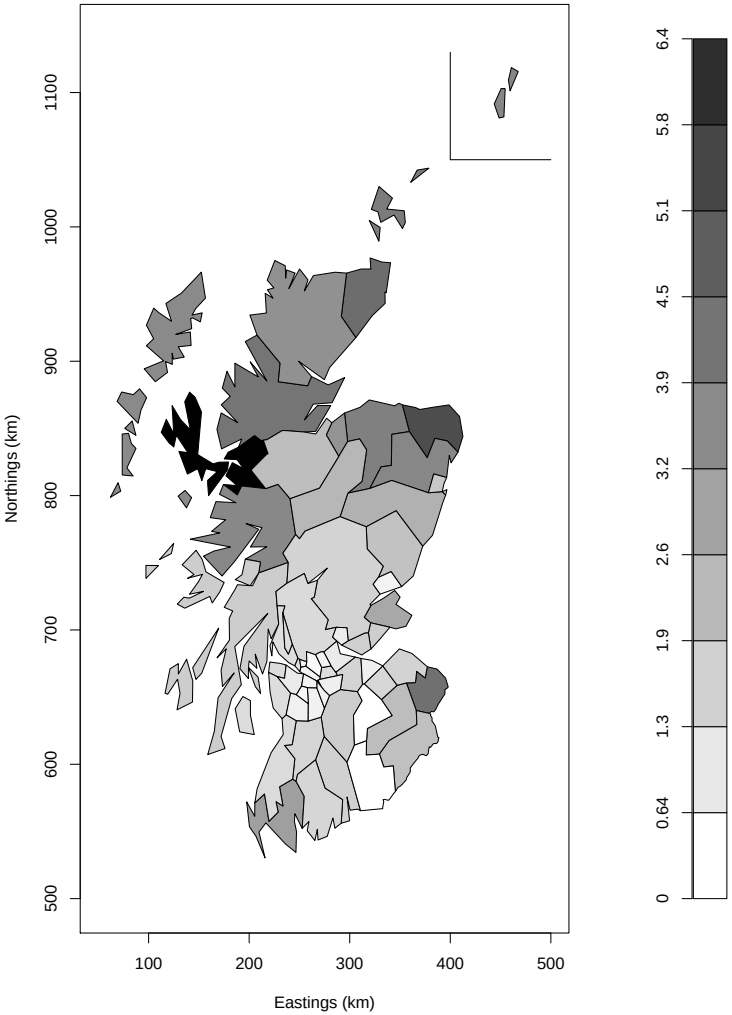


Geostatistics



Point processes

Lattice data

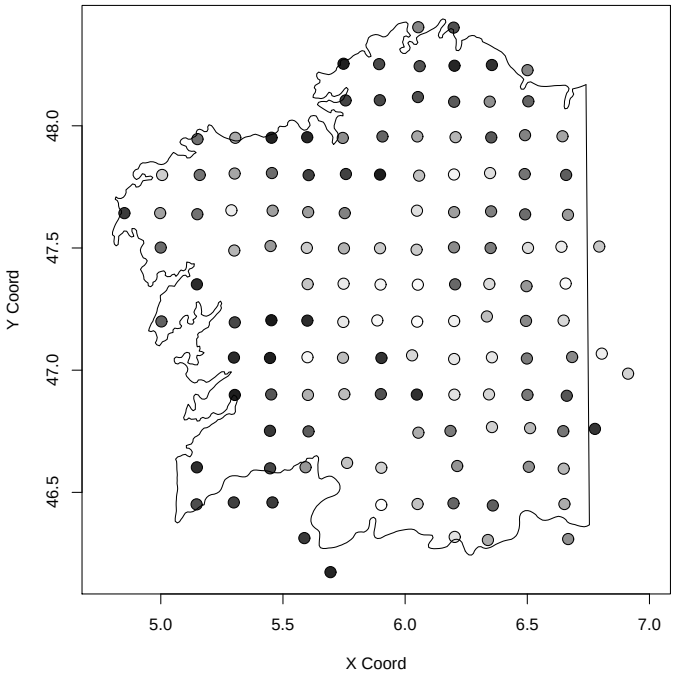


Data: outcomes $Y_i : i = 1, ..., n$

Model: Markov random field:
$[Y_i | \{Y_j : j \neq i\}] : i = 1, ..., n$

"...or indeed, for any multivariate distribution at all"
Hawkes, in discussion of Besag (1974)

Geostatistics

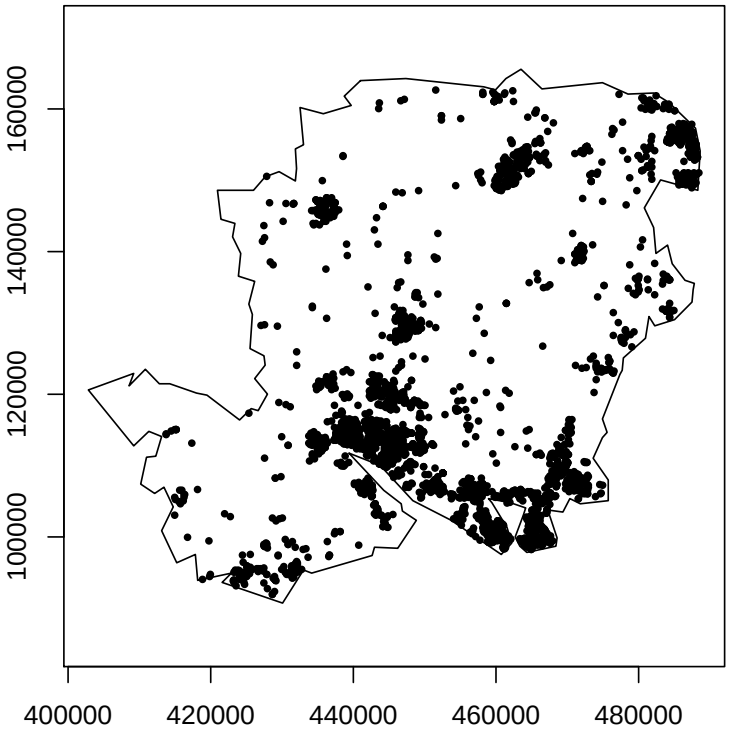


Data: outcome and locations $(Y_i, x_i) : i = 1, ..., n$ (variable n)

Model: spatially continuous stochastic process, $Y(x) : x \in \mathbb{R}^2$

Presumed scientific interest in $Y(x)$ at non-sampled locations

Point process



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

Model: countably infinite set of points, $x_i \in \mathbb{R}^2 : i = 1, 2, ...$

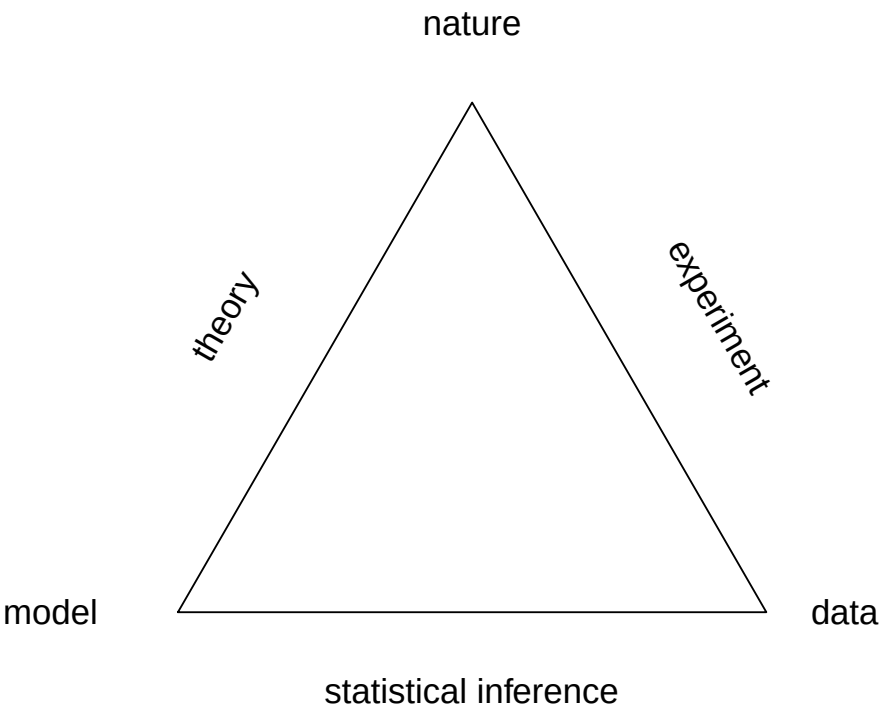
Locations are of scientific interest in themselves.

Spatial statistics according to Cressie

A classification of

- **processes?**
- **models?**
- **methods?**
- **data-formats?**

Analyse problems, not data



A statistical model is:

- **a device to answer a question**
- **a bridge between scientific theory and empirical evidence**
- **a framework to enable principled inference in the presence of uncertainty**


Programa Doenças Infecciosas
e Desigualdade Social:
Comunidade de Pau da Lima
Legenda
Escala 1: 700
AMPLI
FIOCRUZ/MS
CORNELL
ABRASCO
AMOSESM
AMRSR

Leptospirosis cohort study: Pau da Lima

- subjects i at locations x_i, blood-samples taken at times $t_{ij} \approx 0, 6, 12, 18, 24$ months
- sero-conversion defined as change from zero to positive, or at least four-fold increase in concentration
- data consist of:
  - $Y_{ij} = 0/1 : j = 1, 2, 3, 4$ (seroconversion no/yes)
  - $r_i(t)$ known and hypothesised risk-factors

Longitudinal data, binary outcome $\Rightarrow$ GEE? GLMM?

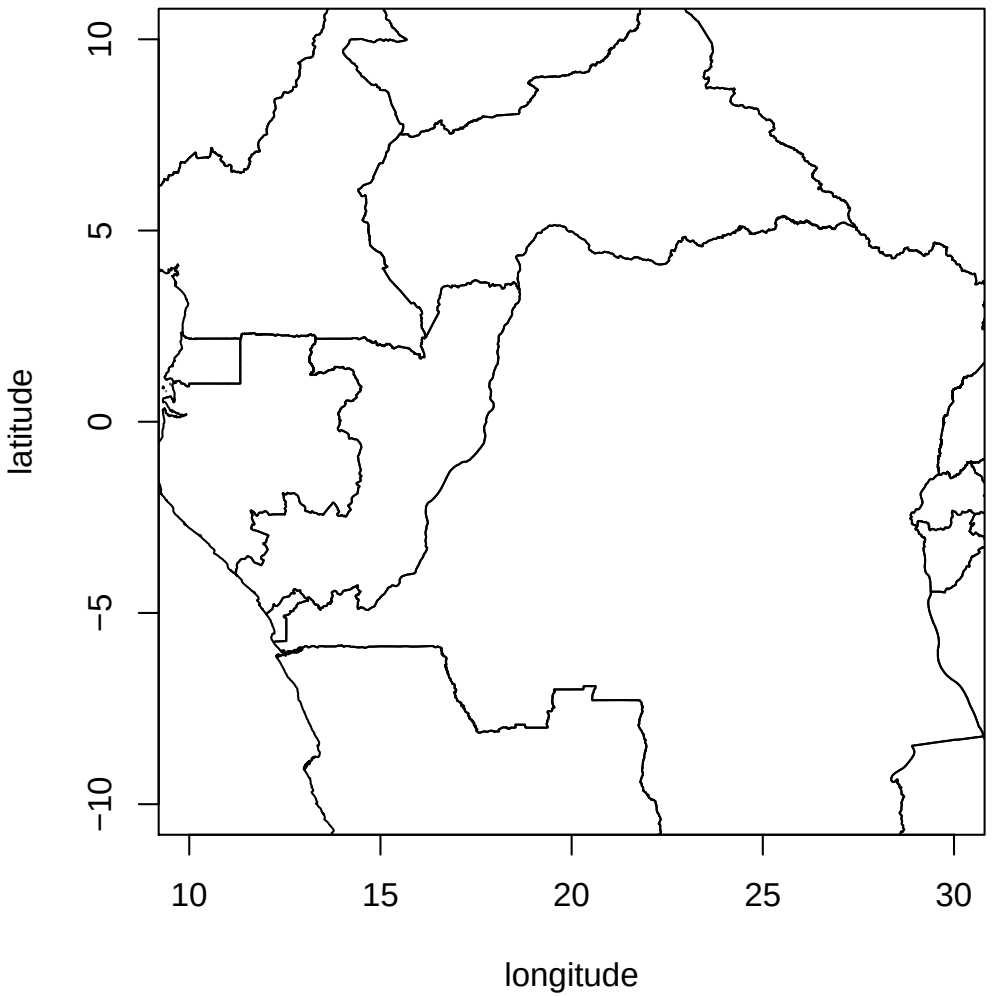
Leptospirosis cohort study: model formulation

- $Y_{it} = 1 \Leftrightarrow$ at least one infection event
- model infection events as person-specific, inhomogeneous Cox processes,

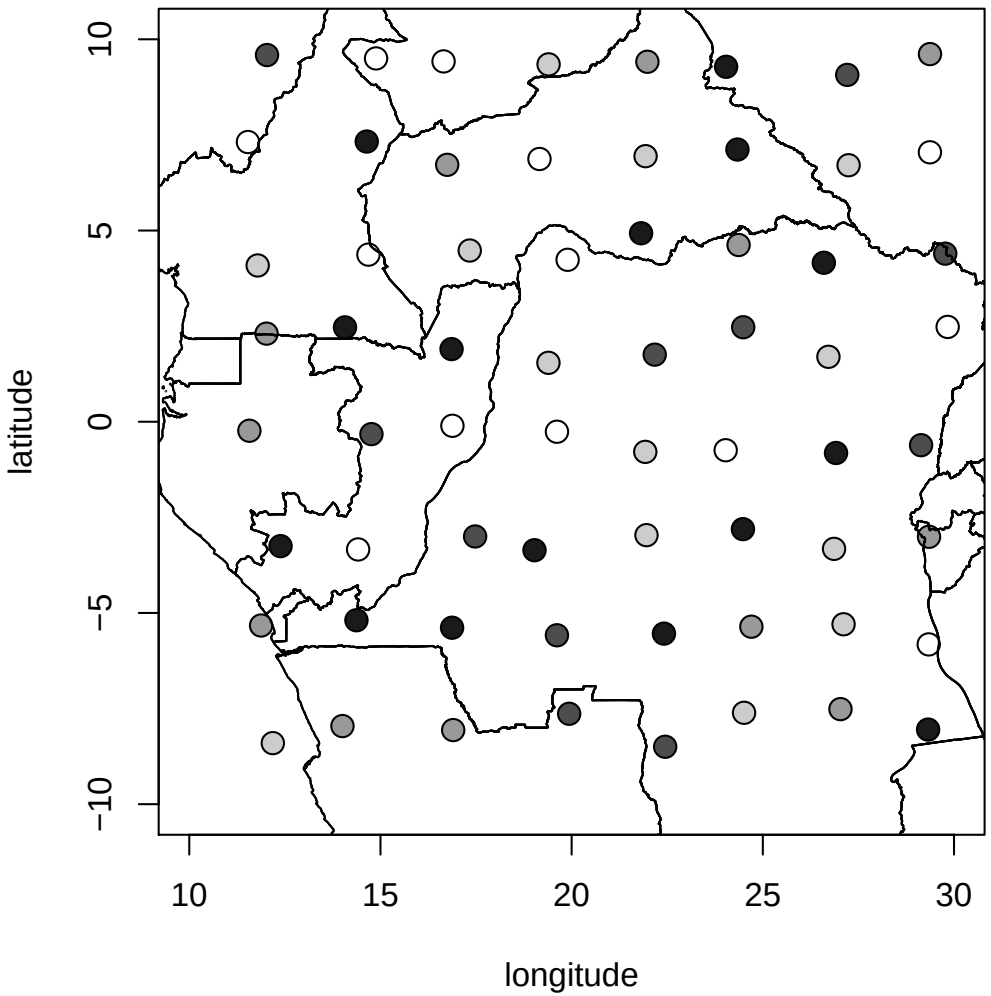
$$\Lambda_i(t) = \exp\{r_i(t)'\beta + U_i + S(x_i)\}$$

- $P(Y_{it} = 1|\Lambda_i(\cdot)\} = 1 - \exp\left\{-\int_{t_{i,j-1}}^{t_{ij}} \Lambda_i(u)du\right\}$

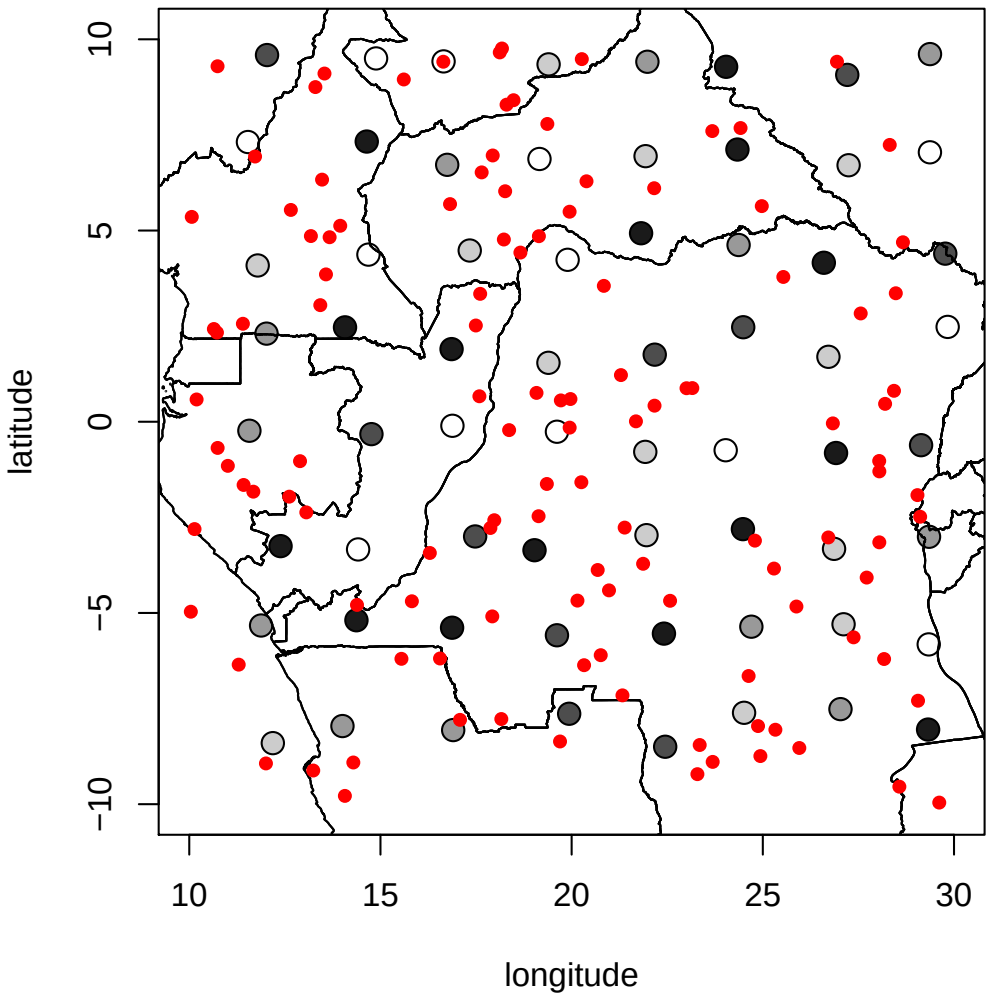
Data-synthesis



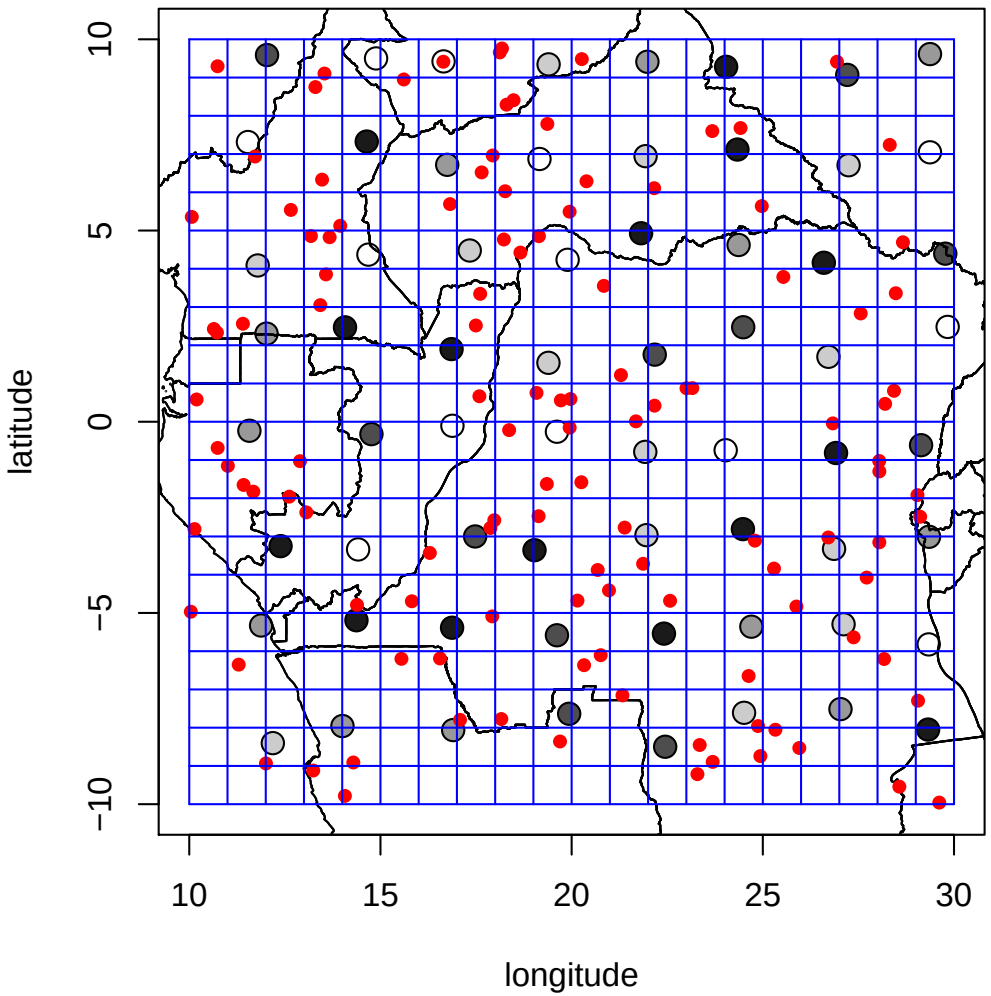
Data-synthesis



Data-synthesis



Data-synthesis



Data-synthesis

$\mathbf{S} =$ state of nature
$\mathbf{Y} =$ all relevant data
$\mathbf{T} =$ $\mathcal{F}(\mathbf{S}) =$ target for prediction

Model: $[\mathbf{S}, \mathbf{Y}] = [\mathbf{S}][\mathbf{Y}|\mathbf{S}]$

Predictive inference: $[\mathbf{S}, \mathbf{Y}] \Rightarrow [\mathbf{T}|\mathbf{Y}]$

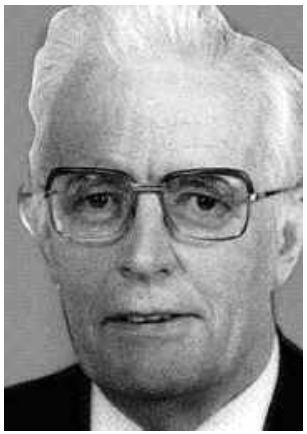
Statistics and scientific method

"Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise."

John Tukey (1915–2000)

"...the importance of making contact with the best research workers in other subjects and aiming over a period to establish genuine involvement and collaboration in their activities."

Sir David Cox (b 1924)



And I leave you with...

- **the role of modelling**

  "We buy information with assumptions"

  Coombs (1964)

- **choice of model/method should relate to scientific purpose.**

  "Analyse problems, not data"

  PJD

References

Bartlett, M.S. (1964). The spectral analysis of two-dimensional point processes. *Biometrika*, 51, 299–311.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with Discussion). *Journal of the Royal Statistical Society* B **36**, 192–225.

Besag, J., Milne, R and Zachary, S. (1982). Point process limits of lattice processes. *Journal of Applied Probability*, **19**, 210–216.

Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with Discussion). *Annals of the Institute of Statistical Mathematics*, **43**, 1–59. Coombs, C.H. (1964). *A Theory of Data*. New York : Wiley

Cox, D.R. (1955). Some statistical methods related with series of events (with Discussion). *Journal of the Royal Statistical Society*, B **17**, 129–57.

Crainiceanu, C.,Diggle, P.J. and Rowlingson, B.S. (2008) Bivariate modelling and prediction of spatial variation in Loa loa prevalence in tropical Africa (with Discussion). *Journal of the American Statistical Association*, **103**, 21–47.

Diggle, P.J. (2006). Spatio-temporal point processes, partial likelihood, foot-and-mouth. *Statistical Methods in Medical Research*, **15**, 325–336.
Diggle, P.J., Eglen, S.J. and Troy, J.B. (2006). Modelling the bivariate spatial distribution of amacrine cells. In *Case Studies in Spatial Point Processes*, ed A Baddeley, P. Gregori, J. Mateu, R. Stoica and D. Stoyan, 215–233. New York: Springer.
Diggle, P.J. , Farewell, D. and Henderson, R. (2007). Longitudinal data with dropout: objectives, assumptions and a proposal (with Discussion). *Applied Statistics*, **56**, 499–550.
Diggle, P.J., Heagerty, P., Liang, K-Y and Zeger, S.L. (2002). *Analysis of Longitudinal Data (second edition)*. Oxford : Oxford University Press.
Diggle, P.J., Kaimi, I. and Abellana, R. (2010). Partial likelihood analysis of spatio-temporal point process data. *Biometrics*, **66**, 347–354.
Diggle, P.J., Menezes, R. and Su, T.-L. (2010). Geostatistical analysis under preferential sampling (with Discussion).*Applied Statistics*, **59**, 191–232.
Diggle, P.J., Moyeed, R.A. and Tawn, J.A. (1998). Model-based geostatistics (with Discussion), *Applied Statistics*, **47**, 299–350.

Diggle, P.J. and Ribeiro, P.J. (2007). *Model-based Geostatistics*. New York: Springer

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

Eglin, S.J. and Wong, J.C.T. (2008). Spatial constraints underlying the retinal mosaics of two types of horizontal cells in cat and macaque. *Visual Neuroscience*, **25**, 209-213.

Fanshawe, T.F. and Diggle, P.J. (2011). Bivariate geostatistical modelling: a review and an application to spatial variation in radon concentrations. *Environmental and Ecological Statistics* (to appear)

Keeling, M.J., Woolhouse, M.E.J., Shaw, D.J., Matthews, L., Chase-Topping, M., Haydon, D.T., Cornell, S.J., Kappey, J., Wilesmith, J. and Grenfell, B.T. (2001). Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science*, **294**, 813–7.

Krige, D.G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, **52**, 119–39.

Liang, K-Y and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.
Little, R.J.A. (1995). Modelling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, **90**, 1112–21.
Matérn, B. (1960). *Spatial variation*. Technical Report, Statens Skogsforsningsinstitut, Stockholm (reprinted, 1986, Berlin : Springer-Verlag).
Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, **58**, 1246–66.
Mercer, W.B. and Hall, A.D. (1911). The experimental error of field trials. *Journal of Agricultural Science*, **4**, 107–132.
Møller, J. and Waagepetersen, R.P. (2004) *Statistical Inference and Simulation for Spatial Point Processes*. London: Chapman and Hall/CRC.
Ripley, B.D. (1977). Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society* B **39**, 172–212.
Ripley, B.D. (1979). AS137 Simulating spatial patterns. *Applied Statistics*, **28**, 109–12.

Ripley, B.D. (1981). *Spatial Statistics*. New York : Wiley.
Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 581–92.
Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. London: Chapman and Hall.
Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with Discussion). *Journal of the Royal Statistical Society* B **71**, 319–392
Stein, M. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer
Watson, G.S. (1972). Trend surface analysis and spatial correlation. *Geological Society of America Special Paper*, **146**, 39–46.
Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics*, **8**, 158–183.
Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley

References from preliminary material

Time series and longitudinal data analysis

Chatfield, C. (2003). *The Analysis of Time Series: an Introduction (6th edition)*. London: Chapman and Hall.

Diggle, P.J. (1990). *Time Series: a Biostatistical Introduction*.

Diggle, P.J., Heagerty, P., Liang, K.Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data (second edition)*. Oxford: Oxford University Press.

Durbin, J. and Koopman, S.J. (2001). *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.

Fitzmaurice, G.M., Laird, N.M. and Ware, J.H. (2004). *Applied Longitudinal Analysis*. New Jersey: Wiley.

References from preliminary material

Spatial statistics

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with Discussion). *Journal of the Royal Statistical Society* B **36**, 192–225.

Cressie, N.A.C. (1991). *Statistics for Spatial Data*. New York : Wiley.

Diggle, P.J. (2003). *Statistical Analysis of Spatial Point Patterns (second edition)*. London : Edward Arnold.

Diggle, P.J. and Ribeiro, P.J. (2007). *Model-based geostatistics*. New York: Springer. Whittaker, J.C. (1990). *Graphical Models in Applied Multivariate Statistics*. Chichester : Wiley.

Matérn, B. (1986). *Spatial Variation (second edition)*. Berlin : Springer-Verlag.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications.* London: CRC Press.