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

Peter J Diggle and Emanuele Giorgi
(CHICAS, Lancaster Medical School, Lancaster University, UK)

September 25, 2015

Pre-requisites

This five-day course is aimed at biostatisticans and numerate public health researchers. Pre-requisites for the course are:

1. an understanding of probability theory including independent and dependent random variables, correlation, conditional probability and standard distributions (binomial, Normal);
2. an understanding of basic statistical theory and method including hypothesis testing, parameter estimation, confidence intervals and regression modelling;
3. experience of using statistical packages to analyse data;
4. familiarity with basic epidemiological concepts including populations and samples, prevalence, incidence and risk.

Previous experience of using the `R` computing environment would be advantageous, but is not essential. The first day of the course will include an introduction to R

We will be using `R` in all of the computing lab sessions. `It is essential that you bring to the course your own computer with R installed`. Instructions for how to do this on a Windows or Mac computer are available on the course web-site:
`www.lancs.ac.uk/staff/diggle/Malawi2015`.

Structure

The course will be delivered using a combination of *lectures*, *demos* (interactive software demonstrations interspersed with short computing exercises) and *labs*. For the lab classes, students will be given access to pre-prepared scripts that are intended to reinforce the material taught in the lectures and software demos, together with more open-ended optional assignments that would require modification and extension of the pre-prepared scripts.

Timetable

Day	Session	Time	Topic
1	1	09.00 - 10.30	Introduction to R (demo)
		10.30 - 11.00	Break
	2	11.00 - 12.30	Introduction to R (demo, continued)
		12.30 - 13.30	LUNCH
	3	13.30 - 15.00	Overview: spatial epidemiology and model-based geostatistics (lecture)
		15.00 - 15.30	Break
	4	15.30 - 17.00	Linear and logistic regression models (lecture)
2	1	09.00 - 10.30	Fitting generalized linear models (lab)
		10.30 - 11.00	Break
	2	11.00 - 12.30	Map-making in R (demo)
		12.30 - 13.30	LUNCH
	3	13.30 - 15.00	Map-making in R (demo, continued)
		15.00 - 15.30	Break
	4	15.30 - 17.00	Exploratory analysis of geostatistical data (lecture)
3	1	09.00 - 10.30	Exploratory analysis of geostatistical data (lab)
		10.30 - 11.00	Break
	2	11.00 - 12.30	Linear geostatistical models (lecture)
		12.30 - 13.30	LUNCH
	3	13.30 - 15.00	Fitting linear geostatistical models (lab)
		15.00 - 15.30	Break
	4	15.30 - 17.00	Geostatistical design (lecture)
4	1	09.00 - 10.30	Geostatistical prediction (lab)
		10.30 - 11.00	Break
	2	11.00 - 12.30	Binomial geostatistical models (lecture)
		12.30 - 13.30	LUNCH
	3	13.30 - 15.00	Binomial geostatistical models (lab)
		15.00 - 15.30	Break
	4	15.30 - 17.00	Prevalence mapping (lecture)
5	1	09.00 - 10.30	Prevalence mapping (lab)
		10.30 - 11.00	Break
	2	11.00 - 12.30	Q and A, additional topics
		12.30	CLOSE

Outline syllabus TO BE CHANGED

1. **Overview: spatial epidemiology and model-based geostatistics**
   Epidemiological study-designs: case-control, survey, registry. Regression modelling: linear and logistic models. Spatial variation: deterministic and stochastic variation. Definition of geostatistical data. Objectives of geostatistical analysis.

2. **Introduction to R**
   Calculations: scalars, vectors and matrices. Reading from data-files. Graphical methods. Simple statistical summaries. Looping. Writing R functions.

3. **Modelling and mapping with R**
   Linear and logistic regression models. Handling spatial structures, map-making.

4. **Exploratory analysis of geostatistical data**
   Basic structure of a geostatistical data-object. Transformations: log and empirical logit. Plotting geostatistical data. Numerical and graphical summaries. Exploring spatial correlation structure: the sample variogram, trend-removal, residual variogram.

5. **Linear geostatistical models**
   Spatial correlation: first law of geography, Matérn class of correlation functions, interpreting the correlation parameters - range and smoothness. Measurement error: the nugget effect and its dual interpretation. Relationship between variance, correlation and theoretical variogram. Parameter estimation: curve-fitting to the residual variogram, maximum likelihood for joint estimation of regression parameters and variogram parameters.

6. **Geostatistical prediction**
   Understanding statistical prediction: a meteorological time series. Spatial prediction: interpolation or smoothing, model-based geostatistics (kriging), predicting properties of a partially observed spatial surface.

7. **Geostatistical design**
   Non-adaptive and adaptive design. Spatially random, regular and intermediate designs. Design with practical constraints: the Majete study.

8. **Binomial geostatistical models**
   Model formulation: binomial generalized linear model with spatially correlated random effect. Parameter estimation by maximum likelihood. Plug-in prediction. Bayesian inference: parameter estimation and spatial prediction as a single process.

9. **Prevalence mapping**
   Practical application of the binomial geostatistical model. Mapping prevalence: exceedance probability maps. Extension: adding a nugget effect.

10. **Q and A, additional topics**
    An opportunity to ask about anything that was (or was not) covered in the course

Recommended background reading

The following books and articles contain far more material than will be covered in the course. They are suggested primarily for future reference, but students would benefit from reading the indicated chapters beforehand if possible. Additional references will be given during the course.

1. Brunsdon, C. and Comber, L. (2015). *An Introduction to R for Spatial Analysis and Mapping*. Sage: London.

   This includes detailed information on how to use `R` for handling spatially structured data-files and for presenting both data and the results of a spatial statistical analysis as high-quality maps. You may find it useful to read *Chapters XXX* before the start of the course.

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

   This is probably the most accessible of the many books on spatial statistical methods that are now available. You may find it useful to read *Chapters 1 and 2* before the start of the course.

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

   This a technically more challenging book than Waller and Gotway (2004), but includes examples of `R` code implementing some of the methods we will discuss in the course. You may find it useful to read *Chapters 1 and 2* before the start of the course.

4. Giorgi, E. and Diggle, P.J. (2015). PrevMap: an R Package for Prevalence Mapping. `Journal of Statistical Software` (submitted).

   This paper describes, and gives worked examples of, the `R` package `PrevMap` that we will be using in the computing lab sessions. It is available on the course web-site, `www.lancs.ac.uk/staff/diggle/Malawi2015`.

5. Ribeiro, P.J. and Digglg, P.J. (2001). GeoR: a package for geostatistical analysis. *R News*, **1/2**, 15–18.

   This short note describes the `R` package `geoR` that we will be using in the computing lab sessions. It is available on the course web-site, `www.lancs.ac.uk/staff/diggle/Malawi2015`.