Spatial Statistics 15 (2016) 70–84

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta

Adaptive geostatistical design and analysis for prevalence surveys

Michael G. Chipeta [a,b,c,*], Dianne J. Terlouw [a,c,d], Kamija S. Phiri [a], Peter J. Diggle [b]

[a] College of Medicine, University of Malawi, Blantyre, Malawi
[b] Lancaster Medical School, Lancaster University, Lancaster, UK
[c] Malawi-Liverpool Wellcome Trust Clinical Research Programme, Blantyre, Malawi
[d] Liverpool School of Tropical Medicine, Liverpool, UK

A R T I C L E I N F O

Article history:
Received 26 August 2015
Accepted 24 December 2015
Available online 4 January 2016

Keywords:
Adaptive sampling strategies
Spatial statistics
Geostatistics
Malaria
Prevalence mapping

A B S T R A C T

Non-adaptive geostatistical designs (NAGDs) offer standard ways of collecting and analysing geostatistical data in which sampling locations are fixed in advance of any data collection. In contrast, adaptive geostatistical designs (AGDs) allow collection of geostatistical data over time to depend on information obtained from previous information to optimise data collection towards the analysis objective. AGDs are becoming more important in spatial mapping, particularly in poor resource settings where uniformly precise mapping may be unrealistically costly and the priority is often to identify critical areas where interventions can have the most health impact. Two constructions are: *singleton* and *batch* adaptive sampling. In singleton sampling, locations x_i are chosen sequentially and at each stage, x_{k+1} depends on data obtained at locations $x_1, \ldots, x_k$. In batch sampling, locations are chosen in batches of size $b > 1$, allowing each new batch, $\{x_{(k+1)}, \ldots, x_{(k+b)}\}$, to depend on data obtained at locations $x_1, \ldots, x_{kb}$. In most settings, batch sampling is more realistic than singleton sampling. We propose specific batch AGDs and assess their efficiency relative to their singleton adaptive and non-adaptive counterparts using simulations. We then show how we are applying these findings to inform an AGD of a rolling Malaria Indicator Survey, part of a large-scale, five-year malaria transmission reduction project in Malawi.

* Corresponding author at: Lancaster Medical School, Lancaster University, Lancaster, UK.
E-mail address: m.chipeta@lancaster.ac.uk (M.G. Chipeta).

http://dx.doi.org/10.1016/j.spasta.2015.12.004
2211-6753/

1. Introduction

Geostatistics has its origins in the South African mining industry (Krige, 1951), and was subsequently developed by Georges Matheron and colleagues into a self-contained methodology for solving prediction problems arising principally in mineral exploration; Chilès and Delfiner (2012) is a recent book-length account. Within the general statistics research community, the term geostatistics more generally refers to the branch of spatial statistics that is concerned with investigating an unobserved spatial phenomenon $S = \{S(x) : x \in D \subset \mathbb{R}^2\}$, where D is a geographical region of interest, using data in the form of measurements y_i at locations $x_i \in D$. Typically, each y_i can be regarded as a noisy version of $S(x_i)$. We write $\mathcal{X} = \{x_1, \dots, x_n\}$ and call $\mathcal{X}$ the *sampling design*.

Geostatistical analysis can address either or both of two broad objectives: *estimation* of the parameters that define a stochastic model for the unobserved process S and the observed data $\{(y_i, x_i) : i = 1, \dots, n\}$; *prediction* of the unobserved realisation of $S(x)$ throughout D, or particular characteristics of this realisation, for example its average value.

A key consideration for geostatistical design is that sampling designs that are efficient for parameter estimation are generally inefficient for prediction, and vice versa—see for example, Diggle and Ribeiro (2007); Müller (2007). Since parameter values are usually unknown in practice, design for prediction therefore involves a compromise. Furthermore, the diversity of potential predictive targets requires design strategies to be context-specific. Another important distinction is between *non-adaptive* sampling designs that must be completely specified prior to data-collection, and *adaptive* designs, for which data are collected over a period of time and later sampling locations can depend on data collected from earlier locations.

In this paper we formulate, and evaluate through simulation studies, a class of adaptive design strategies that address two compromises: between efficient parameter estimation and efficient prediction; and between theoretical advantages and practical constraints. The motivation for our work is the mapping of spatial variation in malaria prevalence in rural communities through a series of "rolling malaria indicator surveys", henceforth rMIS (Roca-Feltrer et al., 2012). rMIS is a malaria transmission monitoring and evaluation tool conducted on a monthly basis. Adaptive design is especially relevant here because resource constraints make it difficult to achieve uniformly precise predictions throughout the region of interest, hence as data accrue over the study-region D it becomes appropriate to focus progressively on sub-regions of D where precise prediction is needed to inform public health action, for example to prioritise sub-regions for early intervention.

In Section 2 we review the existing literature on adaptive geostatistical design and set out the methodological framework within which we will specify and evaluate adaptive design strategies. Section 3 describes our proposed class of adaptive designs for efficient prediction. Section 4 gives the results of a simulation study in which we compare the predictive efficiency of our proposed design strategy with simpler, non-adaptive strategies. Section 5 is an application to the design of an ongoing prevalence mapping exercise around the perimeter of the Majete wildlife reserve, Chikwawa District, Southern Malawi through an rMIS that will be conducted monthly over a two-year period. Section 6 is a concluding discussion.

2. Methodological framework

2.1. Geostatistical models for prevalence data

The standard geostatistical model for prevalence data can be formulated in a hierarchical form as follows (Diggle et al., 1998). For $i = 1, \dots, n$, let Y_i be the number of positive outcomes out of n_i individuals tested at location x_i in a region of interest $D \subset \mathbb{R}^2$, and $d(x_i) \in \mathbb{R}^p$ a vector of associated covariates. The model assumes that $Y_i \sim \text{Binomial}(n_i, p(x_i))$ where $p(x)$ is the prevalence of disease at a location x. The model further assumes that

$$\log[p(x)/\{1 - p(x)\}] = d(x)'\beta + S(x) \quad (1)$$

where $S(x)$ is a stationary Gaussian process with zero mean, variance σ^2 and correlation function $\rho(u) = \text{Corr}\{(S(x), S(x'))\}$, where u is the distance between x and x'.