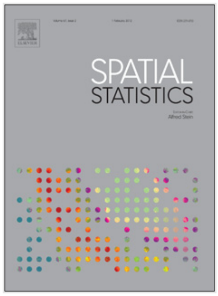
Accepted Manuscript



On the inverse geostatistical problem of inference on missing locations

Emanuele Giorgi, Peter J. Diggle

PII: S2211-6753(14)00060-8
DOI: http://dx.doi.org/10.1016/j.spasta.2014.11.002
Reference: SPASTA 78

To appear in: *Spatial Statistics*

Received date: 11 September 2014
Accepted date: 24 November 2014

On The Inverse Geostatistical Problem of Inference on Missing Locations

Emanuele Giorgi & Peter J. Diggle
Lancaster Medical School
Lancaster University

December 1, 2014

Abstract

The standard geostatistical problem is to predict the values of a spatially continuous phenomenon, $S(x)$ say, at locations x using data $(y_i, x_i) : i = 1, .., n$ where y_i is the realisation at location x_i of $S(x_i)$, or of a random variable Y_i that is stochastically related to $S(x_i)$. In this paper we address the inverse problem of predicting the locations of observed measurements y. We discuss how knowledge of the sampling mechanism can and should inform a prior specification, $\pi(x)$ say, for the joint distribution of the measurement locations $X = \{x_i : i = 1, ..., n\}$, and propose an efficient Metropolis-Hastings algorithm for drawing samples from the resulting predictive distribution of the missing elements of X. An important feature in many applied settings is that this predictive distribution is multi-modal, which severely limits the usefulness of simple summary measures such as the mean or median. We present three simulated examples to demonstrate the importance of the specification for $\pi(x)$ and show how a one-by-one approach can lead to substantially incorrect inferences in the case of multiple unknown locations. We also analyse rainfall data from Paraná State, Brazil to show how, under additional assumptions, an empirical of estimate of $\pi(x)$ can be used when no prior information on the sampling design is available.

Keywords: Geostatistics; kernel density estimation; missing locations; multi-modal distributions.

1 Introduction

Geostatistics was originally developed as a self-contained methodology for spatial prediction (e.g. Mathéron (1963)) but is now embedded as a sub-branch of spatial statistics with applications in many different disciplines. The canonical geostatistical problem is to predict the value of a spatially continuous process, $S(x)$ say, at any required location x in a region of interest $A \subset \mathbb{R}^2$, using data consisting of a set of measured values y_i at each of n locations x_i in A. A widely used geostatistical model is that the y_i are realisations of random variables $Y_i = S(x_i) + Z_i$, where Z_i are mutually independent, zero-mean Gaussian variables, and $S = \{S(x) : x \in \mathbb{R}^2\}$ is a Gaussian process (Diggle et al., 1998). Predictive inference for S is then based on the predictive distribution $[S|Y]$, where $[\cdot]$ means "the distribution of" and $Y = (Y_1, ..., Y_n)$. Conventionally, the set of measurement locations $X = (x_1, ..., x_n)$ is