Journal of Statistical Planning and Inference 146 (2014) 70–81

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Sampling of a continuous flow in the plane based on two-dimensional variogram models

Thaung Lwin [a,*], David Albrecht [b]

[a] CSIRO Mathematics, Informatics and Statistics, Private Bag 10, Rosebank MDC, Clayton, VIC 3169, Australia
[b] Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia

ARTICLE INFO

Article history:
Received 4 January 2013
Received in revised form
14 August 2013
Accepted 19 August 2013
Available online 14 October 2013

Keywords:
Two-dimensional variograms
Empirical modelling of variograms
Plane sampling

ABSTRACT

Theory of sampling from one-dimensional continuous material flows is well established in developments following the original path breaking work by P. Gy, based on an extensively employed tool named variogram and its modelling. These developments could be regarded as a continuous analogue of the discrete time series approach in one dimensional sampling due to W. G. Cochran, whose work was further unified with the variogram approach by G.H. Jowett. Extension of Cochran's discrete time series approach to plane sampling, from equally spaced grid locations in two dimensions, was carried out, independently, by M. H. Quenouille and R. Das. Gy's development is an off-shoot of a closely related geostatistical theory, pioneered by G. Matheron, for collecting and analysing data in three dimensional space. Geostatistical approach generally employed a variogram based on a Euclidean distance to provide a three dimensional modelling of spatial variation structures. Most of the successful variogram modelling approach is essentially based on geometrically isotropic one-dimensional variogram models, instead of variogram models based on generalised distance functions or distance functions in separate dimensions. Quenouille explicitly proposed a geometrically anisotropic two-dimensional autocorrelation function which could be readily reparameterised to obtain a two-dimensional elliptical variogram. Both plane sampling approach and geostatistical approach provide techniques of assessing the uncertainty of the sample average as an estimate of the overall population mean, by the respective techniques of two-dimensional discrete sampling and spatial variogram modelling. The present paper (i) outlines the connection between the precision formula based on plane sampling approach and that based on traditional geostatistical approach, via a unified framework in which the two methods can be compared on an equal footing, with or without equal spacing along the *X*-direction and/or the *Y*-direction (ii) proposes an elliptical empirical variogram model, as a two-dimensional exponential type variogram, obtained as a re-parameterised version of Quenouille's elliptical auto-correlation function and (iii) provides a computationally robust algorithm for fitting a two-dimensional elliptical variogram model to an observed variogram in two dimensions by an approximate likelihood method, in addition to a demonstration of the developed methodology to a data set, available in the published literature.

1. Introduction

An original formulation of the problem of survey sampling in two dimensions arose in the form of sampling from a finite population of units arranged in a grid on a planar region. The problem was formalised and treated, almost simultaneously,

* Corresponding author. Tel.: +61 3 9545 8016; fax: +61 3 9545 8080.
E-mail address: t.lwin@bigpond.net.au (T. Lwin).

0378-3758/$ - see front matter
http://dx.doi.org/10.1016/j.jspi.2013.08.014

by Quenouille (1949) and Das (1950) who both extended the techniques of one-dimensional discrete time series approach in finite sampling theory, due to Cochran (1946), to the case of two dimensions, using two-dimensional analogues of autocovariance and autocorrelation functions. Cochran's work was also exploited by Jowett (1952) who introduced variogram as an alternative tool to autocorrelation function and demonstrated an analysis of the variation structure of a time series in finite sampling context. Meanwhile, a parallel development, due to Gy (1982), appeared in particulate sampling theory, based on the variogram modelling and analysis technique. The basic frame work of Gy, however, was the sampling of a continuous flow in one dimension and thus can be regarded as a continuous analogue of the developments in the finite population sampling theory.

1.1. The problem of generating and analysing data from plane sampling

The observed data is assumed to be generated at the set of coordinates:

$$\{(x_i, y_j) : i = 1, \ldots, m_1, j = 1, \ldots, m_2\}, \tag{1.1}$$

selected from a planar region according to a predetermined sampling scheme. Let $M_1 = m_1 k_1$ and $M_2 = m_2 k_2$. A set of coordinate pairs is selected from a finite population of $N = M_1 M_2$ units identified by the two dimensional co-ordinates:

$$\{(x_i, y_j) : i = 1, \ldots, M_1, j = 1, \ldots, M_2\}, \tag{1.2}$$

according to a defined random sampling scheme, explained in detail in Section 2. Note that the same symbols are used to describe a typical coordinate pair of the sample as well as the population with additional explanation whenever necessary. At the point with location (x_i, y_j), a sampling unit is selected and an observation is made on an associated characteristic variable such as the quality content of the particular sampling unit, so that a two-dimensional array of data is available as observations in a set:

$$\mathbf{Q} = \{q(x_i, y_j); i = 1, \ldots, M_1; j = 1, \ldots, M_2\} \tag{1.3}$$

The set $\mathbf{Q}$ of two-dimensional data sequence of values can be regarded in two ways: (i) as a sample of a discrete set of $N = M_1 M_2$ finite population units in a plane identified by the two dimensional co-ordinates $\{(x_i, y_j) : i = 1, \ldots, M_1, j = 1, \ldots, M_2\}$ and the finite population itself is then assumed to be a sample of a super population, with elements $q(x_i, y_j)$, whose variation structure is described by a continuous process of a characteristic random function $q(x, y)$ defined in two dimensions; (ii) as a digital sampling of a continuous process in a two dimensional space:$\{q(x, y); -\infty < x < \infty, -\infty < y < \infty\}$. In either case the problem is to estimate a characteristic parameter, such as the overall (grand) mean or the overall (total) variance, over a certain specified subregion of the two-dimensional space, of the underlying assumed continuous process using a subset of the set of data $\mathbf{Q}$. In addition it is also required to assess the uncertainty of the estimate of the overall mean. In the specific modelling of the data, it is assumed that the grid points in the population are equally spaced; thus we have$(x_i, y_j) = (i, j)$.

1.1.1. A motivating example

A well known spatial data set is the data set provided by Mercer and Hall (1911) on wheat yields (given in pounds) on a (20×25) grid of points for a planar region. Each of the 20 rows run in the east–west direction and each of the 25 columns run in the north–south direction. Fig. 1 contains a contour plot of the data and the complete data are listed in Cressie (1993).

To evaluate the uncertainty of the average of a subset of the observations, selected in a random scheme from the above wheat yield data set, it is required to model the variation structure of the above data. A commonly employed method of data analysis carried out for this purpose is to construct the observed variogram, in two-dimensions, defined in Section 1.2 and inspect its features to aid the selection of an appropriate empirical model.

1.2. The observed variogram

Let $\mathbf{S} = \{(x_i, y_j) : i = 1, \ldots, m_1, j = 1, \ldots, m_2\}$ be a sample of $m_1 m_2$ points from the planar region, and let $q_{ij} = q(x_i, y_j)$. Then the variogram cloud is defined as:

$$C(\mathbf{S}) = \left\{(u, v, z) : u = |x_{i,j} - x_{s,t}|, \quad v = |y_{i,j} - y_{s,t}|, \quad z = \frac{1}{2}(q_{i,j} - q_{s,t})^2\right\} \tag{1.4}$$

and the observed variogram $\hat{V}(u, v)$ is defined as follows:

$$\hat{V}(u, v) = \frac{1}{|Z(u, v)|} \sum_{z \in Z(u,v)} z \tag{1.5}$$

where

$$Z(u, v) = \{z : (u, v, z) \in C(\mathbf{S})\} \tag{1.6}$$

and $|Z(u, v)|$ is the number of values of z corresponding to (u, v)in the cloud (1.4).

The values of $\hat{V}(u, v)$were calculated for all possible combinations of $(u, v)$for the wheat yield data and the resulting observed variogram is depicted in Fig. 2.