Spatial Statistics 9 (2014) 66–92

Contents lists available at ScienceDirect

Spatial Statistics

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

Effective sample size of spatial process models

Ronny Vallejos *, Felipe Osorio

Departamento de Matemática, Universidad Técnica Federico Santa María, Valparaíso, Chile

ARTICLE INFO

Article history:
Received 26 August 2013
Accepted 14 March 2014
Available online 21 March 2014

Keywords:
Spatial process
Effective sample size
Elliptically contoured distributions
REML estimator

ABSTRACT

This paper focuses on the reduction of sample sizes due to the effect of autocorrelation for the most common models used in spatial statistics. This work is an extension of a simple illustration highlighted in several books for an autoregressive-type correlation structure. The paper briefly reviews existing proposals to quantify the effective sample size and proposes a new definition that is a function of the correlation structure, sample size, and dimension of the space where the coordinates are defined. It describes the properties of and explicit expression for the effective sample size for processes with patterned correlation matrices, including elliptical contoured distributions. The estimation of the effective sample size is achieved using restricted maximum likelihood. Additionally, the paper describes the monotonicity of the effective sample size when two random points are uniformly distributed on the unit sphere and includes several Monte Carlo simulations to explore monotonic features of the effective sample size and to compare its behavior with respect to other proposals. Finally, this paper analyzes two real datasets, and the discussion includes topics that should be addressed in further research.

1. Introduction

Spatial analysis has developed considerably in recent decades. Particularly, problems such as determining sample sizes and how and where to sample have been studied in many different contexts. In spatial statistics, it is well known that as the spatial autocorrelation latent in georeferenced data

* Correspondence to: Departamento de Matemática, Universidad Técnica Federico Santa María, Avenida España 1680, Casilla 110-V, Valparaíso, Chile. Tel.: +56 32 2654964.
E-mail address: ronny.vallejos@usm.cl (R. Vallejos).

http://dx.doi.org/10.1016/j.spasta.2014.03.003
2211-6753/

increases, the amount of duplicated information in these data also increases. This property has many implications for the subsequent analysis of spatial data. A similar problem has been discussed by Box (1954a,b) for approximating the distribution of a quadratic form in the normal random vector by a chi-square distribution that has the same first two moments. Clifford et al. (1989) used this approach to suggest effective degrees of freedom in a modified t-test designed to assess the association between two spatial processes; a detailed discussion can be found in Dutilleul (1993). An extension, in the context of multiple correlation, involving one spatial process and several others is described in Duttilleul et al. (2008). A study on the effective number of spatial degrees of freedom of a time-varying field was conducted by Bretherton et al. (1999). The effective sample size of two glaciers determined by analyzing the spatial correlation between point mass balance measurements is discussed in Cogley (1999). The relative information content of the mean for various hydrologic series is addressed in Matalas and Langbein (1962). The method required to determine the number of independent observations in an autocorrelated time series is noted by Bayley and Hammersley (1946).

The effect of spatial correlation on statistical inference, more specifically the problem of how many uncorrelated samples provide the same precision as correlated observations, is mentioned and illustrated in classical spatial statistics books such as Cressie (1993), Haining (1990), and Shabenberger and Gotway (2005). Griffith (2005) developed a new method to determine the effective sample size for normally distributed georeferenced data for a single mean and also provided extensions for multiple sample means. Griffith's proposal is based on a regression model for which the expected value of the estimated variance of the response variable is calculated. Using this expression and the variance inflation factor, an effective sample size formula is obtained as a function of the covariance structure of the model. Other model-based alternatives and extensions to two processes are also provided. Griffith (2008) later used this method with soil samples from Syracuse, NY. Another approach based on the integral range for which the estimation process is philosophically different (i.e., it is not based on the likelihood) from that mentioned above is described in Lantuejoul (1991).

This paper addresses the following problem: if we have n data points located on a general grid in an r-dimensional space, what is the effective sample size (ESS) associated with these points? If the observations are independent and if a regional mean is being estimated, then, given a suitable definition, the answer is n. Intuitively, when perfect positive spatial autocorrelation prevails, ESS should be equal to 1; with dependence, less than n. Getis and Ord (2000) studied this type of reduction of information in the context of multiple testing of local indices of spatial autocorrelation. Note that the general approach to addressing this question does not depend on the data values; however, it does depend on the spatial locations of the points in the range of the spatial process and on the spatial dimension. We suggest a definition of the spatial effective sample size based on an alternative way of calculating the reduction of information due to the existing spatial association in the data. Our definition can be explored analytically given certain assumptions. We explore certain patterned correlation matrices that commonly arise in spatial statistics, study the effective sample size for a single normal process and extended it to CAR and SAR processes. Additionally, we consider a single mean process with errors that have an elliptically contoured distribution. We present theoretical results and examples to illustrate the features of our proposed method.

Estimation of the effective sample size is addressed via the restricted maximum likelihood estimator for the normal and elliptical cases. We discuss both how sampling design affects effective sample size and how to select a sample after sample size has been reduced to account for autocorrelation, highlighting that effective sample size is intimately related to the way in which data are collected. Additionally, we carried out numerical experiments and Monte Carlo simulations to measure the performance of the estimators of effective sample size and determined the effect of the selected variogram model. We use two examples with real data to illustrate the practical scope of our proposal. The first dataset consists of georeferenced samples from a contaminated area in Utah, USA, and the second dataset consists of forest variables based on a study of Pinus radiata plantations in southern Chile. In both cases, we calculate the effective sample size to explore the reduction of the sample size.

This paper is organized as follows. In Section 2, a motivation is given in the context of a time series previously discussed by Cressie (1993) and an extension to a spatial process is considered to illustrate the effect of dimension. Section 3 develops the effective sample size for normal, CAR, SAR, and elliptical