Spatial Statistics 22 (2017) 68–88

Contents lists available at ScienceDirect

Spatial Statistics

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

Spatial subsemble estimator for large geostatistical data

Márcia H. Barbian [a,*], Renato M. Assunção [b]

[a] Departamento de Estatística, Universidade Federal do Rio Grande do Sul - Porto Alegre, Brazil
[b] Departamento de Ciência da Computação, Universidade Federal de Minas Gerais - Belo Horizonte, Brazil

ARTICLE INFO

Article history:
Received 8 October 2016
Accepted 13 August 2017
Available online 24 August 2017

Keywords:
Large spatial data
Subsampling
U-statistics
Parallel computing
Gaussian random field
Maximum likelihood estimation

ABSTRACT

We introduce the concept of spatial subsemble, a subset ensemble estimation method useful in the analysis of large spatial random field datasets. The full dataset is sampled to give small spatially structured subsets of observations whose parameters are easily estimated; these are combined using a weighting scheme based on their cross-validation prediction ability. We show that our estimator is consistent. More importantly, we compare the spatial subsemble with competing alternatives and show that our proposed procedure is both accurate and much faster than its competitor. We illustrate the use of our method using several examples from large datasets.

1. Introduction

The computer revolution is still going on after decades. Presently, one of its marked aspects is the generation of massive amounts of data. We need to face datasets characterized by the presence of the 4 V's of big data: large volume, velocity, variety, and veracity. Mechanisms generating these data are found wherever remote sensors, satellites, and mobile devices are used, including the emerging internet of things. In particular, the size of spatial datasets has increased dramatically in the recent past with the growth of global satellite imaging, climate monitoring, and the remote recording of meteorological and air quality measurements. NOAA (the National Oceanic & Atmospheric Administration, part of the US Department of Commerce) has a website where terabytes of data are generated to represent the Earth's climate, atmospheric and meteorological states.

[*] Corresponding author.
E-mail address: helena.barbian@ufrgs.br (M.H. Barbian).

http://dx.doi.org/10.1016/j.spasta.2017.08.004
2211-6753/© 2017 Elsevier B.V. All rights reserved.

The main consequence of this growth for statisticians working in spatial statistics is the need to deal with numerical and computational difficulties brought about by the massive amount of data to be analyzed, since many methods typically crash or take too long to be useful. Consider, for example, the exact computation of the likelihood assuming a Gaussian geostatistical model, with n stations located in an irregular way on a map, generally requires $O(n^3)$ numerical operations and $O(n^2)$ memory space (Stein, 2008). These numbers scale up quickly. When $n = 1000$, most spatial statistics software packages solve the problem easily, but when $n = 50\,000$, the problem becomes a severe challenge for most machines and softwares.

Several statisticians are actively looking for improved methods to analyze large spatial datasets, both for cases where the covariance function is stationary, and where it is non-stationary. In the case of stationary processes, there are two main lines of approach. The first adopts a Bayesian viewpoint with work concentrated into two main categories: using a latent process with reduced dimension (Banerjee et al., 2008; Finley et al., 2009), and using a Markov random field to approximate the Gaussian field (Lindgren et al., 2011; Rue and Tjelmeland, 2002). The second approach has more variations. One is to taper the covariance function by setting the covariance between distant stations equal to zero (Kaufman et al., 2008; Furrer et al., 2006). Another possibility is to truncate the spectral representation to zero (Fuentes, 2007). Stein et al. (2004) and Vecchia (1988) used a composite likelihood function while Sun and Stein (2016) work with the score function, with its inverse covariance matrix approximated by a sparse matrix, and Castrillón-Candás et al. (2016) use multilevel set of contrasts. All these methods ignore some aspect of the full model in order to reduce the numerical complexity. Most of them differ by selecting different aspects to achieve a simplified likelihood function.

The non-stationary large spatial datasets case has a smaller volume of published research. Recently, results have been published by Datta et al. (2016), Katzfuss (2016), Konomi et al. (2014), Katzfuss (2013) and Sang et al. (2011).

Analysis of non-spatial big data problems are to be addressed by statisticians using subsampling techniques (Kleiner et al., 2014; Sapp et al., 2014) and divide and conquer techniques (Guha et al., 2012; Chen and Xie, 2014), which can significantly reduce the dimension of the problem, hence they can alleviate the computational demand; a review of these techniques can be found in Schifano et al. (2016) and Bühlmann et al. (2016). In the context of spatial statistics, Liang et al. (2013) developed a method for large geostatistical datasets using a resampling-based, in which small subsamples are sequentially selected and model parameter estimates are updated within the framework of stochastic approximation of Robbins and Monro (1951) and Andrieu et al. (2005).

Liang et al. (2013) provide a fast and consistent estimator which, hence it is appropriate for large datasets. However, their method has drawbacks: it has a sequential structure which prevents it from being parallelized, and it is necessary to check the stochastic convergence of the algorithm.

In this paper, we propose a new method that is simple to apply, computationally fast and requires little memory space. It allows for the calculation of confidence intervals and it has good theoretical properties for both the infill asymptotic approach as well as for the increasing domain asymptotic approach. The method is based on subsampling small spatially structured subsets of observations. In each subsample, we fit the parameters with the preferred method and combine them using a validation subset. The two main advantages of our method are: first, its simplicity, making it very easy to use; and second, its speed, since it can be parallelized. To show these advantages, we compare our spatial subsemble algorithm with resampling-based stochastic approximation (RSA) by Liang et al. (2013), and MLE (Maximum Likelihood Estimation), showing that the proposed method is accurate and substantially faster than the other methods. We illustrate our method using a large NOAA dataset of precipitation over the United States.

2. The spatial subsemble estimator

2.1. The geostatistical Gaussian model

Suppose the vector $\mathcal{Y} \equiv (Y(s_1), Y(s_2), \ldots, Y(s_n))^T$ are observed values of a random process $\{Y(s) : s \in D \subset \mathbb{R}^2\}$, where the spatial index s varies continuously across the region D. Let the