Spatial Statistics 21 (2017) 319–335

Contents lists available at ScienceDirect

Spatial Statistics

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

Gaussian spatial linear models with repetitions: An application to soybean productivity

F. De Bastiani [a,*], M. Galea [b], A.H.M.A. Cysneiros [a], M.A. Uribe-Opazo [c]

[a] *Universidade Federal de Pernambuco, Avenida Professor Moraes Rego 1235, Recife 50740-540, Pernambuco, Brazil*
[b] *Pontificia Universidad Católica de Chile, Avenida Vicuña Mackenna 4860, Macul, Santiago 782-0436, Chile*
[c] *Universidade Estadual do Oeste do Paraná, Rua Universitária 2069, Cascavel 85819-110, Paraná, Brazil*

ARTICLE INFO

Article history:
Received 12 June 2016
Accepted 29 July 2017
Available online 9 August 2017

Keywords:
Diagnostics
Geostatistics
Maximum likelihood

ABSTRACT

Soybean yield in the south region of Brazil was analyzed by using Gaussian spatial linear models to account for the spatial dependence between the observations. Some special cases of Gaussian spatio-temporal linear models were also considered. The data set was collected over five years and we regarded four explanatory variables. Each year was considered to be a realization of the process. We checked the goodness of fit and since we noted the presence of an atypical year with lower productivity than in the other years, we conducted influence diagnostics analysis. Moreover, we considered two appropriate perturbation schemes in the response variable and case weight perturbation. The results showed that the methods were effective for detecting influential observations.

1. Introduction

Global soybean demand is high because it is a source of protein for the human diet, animal nutrition, cosmetics, biofuel feedstock, and oil seed. It is also used as a nutritional supplement for cattle, pigs, and poultry, and it can be found in the form of grain in nature for the composition of industrial feed or meal.

In this study, we analyze soybean productivity (also called soybean yield) data and four chemical contents from an agricultural area in south Brazil. To consider the dependence between observations,

* Corresponding author.
E-mail addresses: debastiani@de.ufpe.br (F. De Bastiani), mgalea@mat.uc.cl (M. Galea), audrey@de.ufpe.br (A.H.M.A. Cysneiros), miguel.opazo@unioeste.br (M.A. Uribe-Opazo).

http://dx.doi.org/10.1016/j.spasta.2017.07.013

the analysis is conducted by using geostatistical techniques. In traditional geostatistics, data are collected at known sites in space, from a process that has a value at every site in a certain domain. The data are then modeled as the sum of a constant or varying trend and a spatially correlated residual. Given a model for the trend, and under some stationarity assumptions, geostatistical modeling thus involves the estimation of the spatial correlation (Pebesma, 2004). Geostatistical spatio-temporal modeling is an empirical approach where the model is specified and its parameters estimated from observed data. Spatial–temporal models arise when data are collected across time as well as space.

The data set used in this study was collected over five years from 1998 to 2002 and the observations were taken from different geographical sites (experimental units), where each variable was observed five times, once per year.

According to Masuda and Goldsmith (2009), global soybean production increased by 4.6% annually from 1961 to 2007 and reached an average annual production of 217.6 million tons in 2005–2007. However, in the data set considered herein, we noted the presence of one year with lower yield than in the other years, which motivated us to develop a local influence methodology for this type of model based on the approach presented in Cook (1986).

To assess the effect of small perturbations in the model (or data) on the parameter estimates, Cook (1986) proposed the local influence method. Zhu et al. (2007) constructed influence measures by assessing the local influence of perturbations to a statistical model. Waternaux et al. (1989) suggested several practical procedures for detecting outliers for the repeated measurements model based on the global influence approach. Lesaffre and Verbeke (1998) extended the local influence methodology to normal linear mixed models both in a repeated measurement context and under the case weight perturbation scheme. When there is no repetition, Uribe-Opazo et al. (2012) used diagnostic techniques to assess the sensitivity of the maximum likelihood (ML) estimators, covariance functions and linear predictor to small perturbations in the data and/or in the Gaussian spatial linear model (GSLM) assumptions.

In this study, we extend the work presented by Uribe-Opazo et al. (2012) to the case of independent repetitions and for some special cases of spatio-temporal geostatistical models. We propose these models to analyze soybean yield and present influence diagnostic studies. The remainder of the paper is organized as follows. Section 2 presents a previous analysis for the soybean yield data set. Section 3 describes the GSLM with repetitions, the ML estimators and an explicit expression for the Fisher information matrix. In Section 4 we extend the GSLM with repetitions to take account possible temporal correlated structure. Section 5 reviews local influence concepts for both GSLM with repetitions and a special case of spatio-temporal model. Furthermore, we discuss the selection of an appropriate perturbation scheme based on the methodology proposed by Zhu et al. (2007). Section 6 contains the analysis results. Finally, Section 7 concludes. Calculations are presented in Appendices A and B.

2. The data set

The soybean yield data set was collected in a grid of 7.20 × 7.20 m in an experimental area of 1.33 ha at the Eloy Gomes Research Center at *Cooperativa Central Agropecuária de Desenvolvimento Tecnológico e Econômico Ltda* (COODETEC), in Cascavel city in Paraná State, Brazil, with Oxisol soil. We collected soybean yield data (Prod) in t ha^{-1} and four chemical contents during April in 1998, 1999, 2000, 2001 and 2002 with 253 observations each. The explanatory variables were: phosphorus (P), potassium (K), calcium (Ca), and magnesium (Mg) measured in cmolc dm^{-3}. The observations were taken at the same site for each repetition.

Table 1 presents a descriptive analysis of the response variable, namely soybean yield. Although Prod 1999 (Y_2) has the minimum value, the smallest mean yield value is for Prod 2002 (Y_5). This year has the largest variance coefficient (var. coef.) as well as the smallest values, as shown in Figs. 1 and 2. Fig. 1(a), (b), (c), (d), and (e) show plots for each repetition in the same scale, revealing that the yield values have been decreasing over time.

Fig. 2 shows the usual boxplots for yield, where only the observations taken in the harvest year 2002 do not have outliers.