Spatial prediction using geostatistical models with an auxiliary variable

Shuxian LI [1,3] , Anne Gégout Petit [2] & Lucia Guérin Dubrana [1,3]

[1] *Université de Bordeaux ISVV UMR 1065 INRA*
sli@bordeaux.inra.fr
[2] *Institut Elie Cartan, Université de Lorraine, Nancy, France*
anne.gegout-petit@univ-lorraine.fr
[3]*Bordeaux Sciences Agro, Gradignan, France*
lucia.guerin@agro-bordeaux.fr

Résumé. Dans cette étude, nous utilisons des modèles géostatistiques pour la prédiction spatiale d'une variable d'intérêt observée en peu de points, à l'aide d'une variable auxiliaire observée en un nombre très élevé de points, structurés en ligne, ce qui est souvent le cas dans le domaine de l'agriculture. Etant donné que les deux variables ne sont pas mesurées aux même points, la procédure de prédiction nécessite une étape d'interpolation spatiale de la variable auxiliaire et une étape de régression spatiale de la variable d'intérêt. Lors de la première étape, les données étant hétérogènes sur la surface étudiée, une analyse locale a été utilisée. Pour traiter la structure en ligne de la variable auxiliaire, nous avons eu recours à un algorithme "one-way Median Polish" qui extrait "l'effet colonne". Enfin nous avons effectué un krigeage ordinaire sur les résidus obtenus. Pour la deuxième étape, nous avons prédit la variable d'intérêt en utilisant modèles de regression et un modèle géostatistique tel que le krigeage universel avec dérive externe. Nous discutons l'intérêt de ces modèles et choisissons celui qui donne la meilleure prédiction en comparant leurs performances à l'aide de validations croisées.

Mots-clés. Modèle geostatistique, krigeage, regression spatiale

Abstract. In this study, we use geostatistical models for the spatial prediction of a target variable observed at only a few location points. To compensate, we introduce an auxiliary variable observed at a very great number of points, distributed in lines, which is common in the agricultural domain. As the two variables were not measured at exactly the same points, a prediction procedure was needed. This was composed of two steps: spatial interpolation of the auxiliary variable followed by spatial regression of the target variable on the auxiliary variable. At the first step, local analysis was performed because the auxiliary variable data were heterogeneous in the area under study. A one-way Median Polish algorithm was then used to extract the "column effect" from the array data, so that ordinary kriging could be executed on the residuals. We then predicted the target variable at the second step, using regression models and a specific geostatistical model called universal kriging with external drift. We discussed the interest of these models and chose the one that gave the better prediction by comparing their performances using cross-validation.

Keywords. Geostatistical model, kriging, spatial regression

1 Introduction

In various fields, many scientists need to predict spatial distributed variables based on their different samples. However, when the target variable is sparsely measured in a particular area, the predicting can be very difficult, not only because there are insufficient samples, but also because of the strong spatial heterogeneity. Thus, we must resort to more abundant and accessible ancillary information to predict a sparse target variable. Specifically, we want to predict a target variable whose measurement is sparse, Z_1, at required locations, using a densely measured auxiliary variable Z_2.

In this paper, we use geostatistical methods such as kriging and statistical regression models. The auxiliary variable we study has a special spatial structure, with its samples being well aligned. By contrast, the target variable samples were measured between the lines.

So these two variables are not measured at the same locations. Consequently, the prediction process is composed of two steps. Initially, it is necessary to interpolate the auxiliary variable at the measured and required prediction locations of the target variable. A one-way Median Polish method is used here to adapt to the array structure of the Z_2 observations given above (Cressie, 1993). The second step is to predict the target variable at the required unsampled locations, using auxiliary variable values. The usual regression models and universal kriging model were compared in this step.

2 Framework

Supposing that we have target variable Z_1 and the auxillary variable Z_2 valued in a spatial field $\mathcal{D}$, their spatial observations can be represented by $\mathbb{Z}_1 = (Z_1(s_1), ..., Z_1(s_m))'$, $\mathbb{Z}_2 = (Z_2(s_{m+1}), ..., Z_2(s_{m+n}))'$, where $\{s_1, s_2, ..., x_{m+n}\} \in \mathcal{D}$ correspond to their observed locations respectively, and $m << n$. They can be considered as partial samplings of the realisation of random processes $Z_1(s), Z_2(s), s \in \mathcal{D}$. We want to predict the Z_1 at unsampled locations $\{s_{1'}, s_{2'}, ..., s_{l'}\}$ using observations of Z_2 .

3 Methods

3.1 Kriging models

First, ordinary kriging is used to predict Z_2 and then universal kriging with external drift, is used to predict Z_1 by Z_2.
These two methods are presented here:
For the *ordinary kriging*, the predictor of variable Z for an unsampled location s_0 is the linear combination of n neighbourhood sampling points,

$$Z^*(s_0) = \sum_{\alpha=1}^{n} \omega_\alpha Z(s_\alpha) \tag{1}$$

To guarantee the consistency of the estimators, the weights have to be constrained to sum up to one. The weights are calculated using an estimated variogram by minimizing the estimation variance under the constraint.
Kriging with external drift (KED) is a particular formulation of universal kriging. It allows the ancillary information to be used to account for the spatial variation of the target variable Z_1 local mean. The auxiliary variable Z_2 is chosen for its strong correlation with the target variable. The auxiliary variable should be measured or estimated at every location of the target variable and every estimation point. The linear link between Z_1 and Z_2 is incorporated in Equation (1) which gives:

$$\begin{aligned} E[Z_1^*(x_0)] &= \sum_{\alpha=1}^{n} \omega_\alpha E[Z_1(x_\alpha)] = a_0 + b_1 \sum_{\alpha=1}^{n} \omega_\alpha Z_2(x_\alpha). \\ E[Z_1^*(x_0)] &= E(Z(x_0)) = a_0 + b_1 Z_2(x_0) \end{aligned} \tag{2}$$

This implies that the weights should be consistent with an exact interpolation of $s(x)$

$$Z_2(x_0) = \sum_{\alpha=1}^{n} \omega_\alpha Z_2(x_\alpha). \tag{3}$$

The objective function ϕ to be minimized in this kriging system consists of the estimation variance σ_E^2 and of two constraints.

$$\phi = \sigma_E^2 - \mu_0(\sum_{\omega_\alpha=1}^{n} -1) - \mu_1(\sum_{\alpha=1}^{n} \omega_\alpha Z_2(x_\alpha) - Z_2(x_0))$$

The supplementary universality condition, concerning one or several external drift variables measured exhaustively in the spatial domain is incorporated into the kriging system. (Goovarest, 1997)

3.2 Median Polish

To descibe the irregular gridded spatial data (two-way array) in which the grid spacings do not have to be equal in either the horizonal direction or the vertical direction, Cressie (1993) speaks of a mean structure obtained by additive decomposition of the row and column effect:

$$u(s_i) = a + r_k + c_l, \quad s_i = (x_l, y_k) \tag{4}$$

In order to avoid bias and the influence of the extreme values, a specific approach called Median Polish, has been proposed to estimate the additive effects given above using Median theory (Cressie, 1993). Median Polish proceeds by repeated extraction of the row and column medians until convergence, with respect to a stopping criterion chosen by the investigator. It gives new estimators of a, r_k, c_l, which we write as $\tilde{a}, \tilde{r_k}, \tilde{c}_l$, So the original spatial data can be expressed as :

$$Z(s_i) = \tilde{a} + \tilde{r_k} + \tilde{c}_l + R(s_i) \tag{5}$$

where $R(s_i)$ is the Median-Polish residual which is detrended to allow ordinary kriging to be carried out:

$$\tilde{R}(s_0) = \sum_{i=1}^{n} \lambda_i R(s_i) \tag{6}$$

For s=(x,y)' in the region bounded by lines joining the four nodes, $(x_l, y_k)'$, $(x_{l+1}, y_k)'$, $(x_l, y_{k+1})'$, $(x_{l+1}, y_{k+1})'$, where $x_l < x_{l+1}$ and $y_k < y_{k+1}$ define the planar interpolant:

$$\tilde{u}(s) \equiv \tilde{a} + \tilde{r_k} + (\frac{y - y_k}{y_{k+1} - y_k})(\tilde{r}_{k+1} - \tilde{r_k}) + \tilde{c}_l + (\frac{x - x_k}{x_{k+1} - x_k})(\tilde{c}_{k+1} - \tilde{c}_l) \tag{7}$$

Thus the Median-Polish kriging predictor is :

$$\tilde{Z}(s_0) \equiv \tilde{u}(s_0) + \tilde{R}(s_0) \tag{8}$$

4 Prediction procedure

In accordance with the prediction plan mentioned in the introduction, we are now going to explain the prediction procedure in detail.

4.1 Predict auxiliary variable at target variable sampled and required un-sampled locations

In this step, we have to interpolate auxiliary variable Z_2 at target variable sampled locations $\{s_1, s_2, ..., s_m\}$ and at required un-sampled locations $\{s_{1'}, s_{2'}, ..., s_{l'}\}$ using observations $(z_2(s_{m+1}), ..., z_2(s_{m+n}))$.

Unlike other interpolation models, kriging models have the advantage that they integrate representation of the average spatial variability by estimating the variogram, and give us a best linear unbiased estimator. So here we have chosen a kriging model to execute this step. In order to adapt to heterogeneous spatial structures which often leads to the non-stationarity of increments and to improve the assessment of Z_2, here, a local kriging with local estimated variogram is applicable, because the number of predicting locations m is sufficiently small (Walter, 2001).
So for each predict location $s_i, i \in 1, ..., m$, we choose a neighbourhood D_i, and only the spatial units $Z_2(s_{ij}), s_{ij} \in D_i$ have been used to estimate $Z_2(s_i)$.
Local analysis can partly reduce non-stationarity on the increments, but it still leaves a non-stationary part, due to the column effect from the array structure.

4.2 One-way Median Polish for array data

We only consider the global effect and column effect to our data due to the array structure of the auxiliary variable observations. So the observations of Z_2 can be decomposed as:

$$Z_2(s_i) = a + c_l + \epsilon(s_i), \quad s_i = (x_l, \text{whatever } y) \in \mathcal{D}. \tag{9}$$

A simplified one-way Median Polish algorithm gives us the effect estimators $\tilde{a}, \tilde{c_l}$. A variogram analysis and an ordinary kriging can be performed on the residuals $\epsilon(s)$. Then, we use (7) and (8) to estimate the auxiliary variable value at required locations $s_{i'}, i \in 1, ..., m$.

4.3 Spatial estimates of target variable by auxiliary variable

Finally, as we have obtained the estimated values of Z_2 at all the required locations, it only remains to estimate Z_1. There are mainly two ways of estimating the target value. One is to consider only the relationship between target and auxillary variable Z_1, Z_2 and to model it using statistical models such as GLM (Generalized linear model) or SLR(Simple linear regression). The other way is to apply universal kriging with external drift.
The difference between these two types of models is that the geostatistical model take the spatial structure of the target variable's observations into account. Since the measurements are scarce, the geostatistical model might be risky because the variogram estimation could be too imprecise.
The methods discussed in this paper have been applied to selected data in an agricultural domain. The results of this study will be given in an oral presentation.

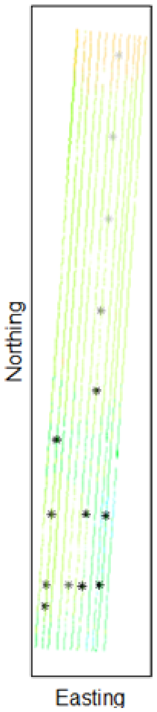

Figure 1: An example of described sampling structure, with m (the number of Z_1 measurements)=14 and n (the number of Z_2 measurements)=50000.

Bibliograpy

[1] Cressie, N. (1993). *Statistics for spatial data, revised edition*, John Wiley & Sons, New York.
[2] Goovarest, P. (1997), *Geostatistics for Natural Resources Evaluation*, Oxford University Press.
[3] Walter, C., McBratney, A. B., Douaoui, A., & Minasny, B. (2001), Spatial prediction of topsoil salinity in the Chelif Valley, Algeria, using local ordinary kriging with local variograms versus whole-area variogram *Soil Research* 39(2), 259-272.