Addressing alternative approaches for spatial modeling of herbicide retention in soil

Giannini Kurina F.[1], *S. Hang*[2], *A. Rampoldi*[1,2], *M. Cordoba*[1,2], *E. R. Macchiavelli*[3], *M. Balzarini*[1,2]

[1] francagianninikurina@gmail.com, Consejo Nacional de Investigaciones Científicas y Técnicas de Argentina.

[2] Facultad de Ciencias Agropecuarias Universidad Nacional de Córdoba.

[3]Universidad de Puerto Rico, Mayagüez.

Abstract

Glyphosate retention coefficient (Kd) is modeled as function of soil variables, from a regional sampling, using: Ordinary and Partial Least Square regression, Random Forest, Generalized Boosted regression (GB), and Bayesian modelling with INLA; all regressions were fitted using spatial constraint on residuals. INLA produced the best fit, but GB the best spatial prediction.

Keywords: Predictive model, spatial data.

1. Introduction

Diffuse pollution derived from the use of plant protection products is a problem associated with agricultural intensification [1]. It requires deeper studies to adapt the available knowledge and generate useful technologies for decision making. The use of herbicides carries environmental risks that depends on the pesticide molecule and environmental characteristics (soil, climate, and management). Models to evaluate environmental hazards require, as inputs, variables that synthesize the interaction between herbicide and soil. An interaction that regulates pesticide behaviour in soil is retention, which is parameterized by the adsorption coefficient (Kd) [2]. It is a continuous variable that expresses the relationship between both, the amount of herbicide retained and the amount that remains in soil solution. Low values of Kd are often related to losses of leaching and runoff, while the potential losses by soil erosion are associated with high Kd values. Several soil variables may be conditioning the retention of the herbicide molecule in a site [3]. These variables are often spatially structured whereby modeling herbicide retention, through Kd coefficient, requires accounting for the underlying spatial variability. In general terms, a linear model for spatial data contains a deterministic component and a random one, which is explained by the spatial autocorrelation process and a net residual term. Models for spatial prediction can be generated from different strategies that allow fitting both, the deterministic component and the random one. In this work, approaches of different nature from which it is possible to adjust a regression model, including the spatial autocorrelation in the residual term, are addressed. We compare the results of three approaches to fit predictive regression models for spatial data: the frequentist [4] , the Bayesian [5], [6] and the Boosting-based[7].

2. Materials and Methods

A soil survey was conducted in Córdoba province, Argentina (29^o to 35^oS, and 61^o to 65^oW) that collected samples from the upper 15 cm of soil using a regular 40×40 km grid (90 sites)[3]. For each soil sample, the following variables were obtained: pH, total nitrogen, total organic carbon, Na, K, Ca, Mg, Zn, Mn, Cu, cation exchange capacity, the percentage of sand, silt and clay, water holding capacity and aluminium

and iron oxides. The glyphosate Kd was determined in lab according to the bach-equilibrium technique for the preparation of soil suspensions. The concentration of herbicide in the soil solution under equilibrium (Ceq) was quantified by high pressure liquid chromatography (HPLC) according to Marek and Koskinen (2014). The adsorbed concentration (Cad) was calculated as the difference between the initial concentration and the concentration at equilibrium in the solution. The Kd was obtained as Cad/Ceq. The Kd values were transformed to the log scale because of the skewness distribution. To help the selection of soil variables that best explain Kd variability a regression tree improved by resampling (Boosting Regression Tree) [8], was used. The R package gbm, with the functions gbm.step and gbm.simplify, was implemented to select the subset of variables that minimize deviance. To fit predictive model for Kd, we assessed the following strategies: Multiple Linear Regression (ML) with spatially correlated errors through an spatial exponential function [9]; Random Forest Regression (RF), and Generalized Boosted Regression (GB), with spatially correlated residuals according with the methodology proposed by proposed by [7]. In the implementation an exponential model was fitted to the empirical semivariogram of the residuals of both machine learning algorithm. The same procedure to account for spatiality was implemented on the residuals obtained after fitting a Partial Least Square regression model (PLS). Additionally, a Bayesian model approached with Integrated Nested Laplace Approximation (INLA) [6] was fitted on the same data. To account for spatial correlation during the Bayesian modeling the Matern function solved by spatial partial differential equation (SPDE) [10], on the spatially structured random component was used. The complete R code to fit predictive models presented here is available at https://github.com/francagiannini. For all models, the mean square error (MSE) was obtained from the differences between observed and predicted value. In the Bayesian framework, it was calculated from the difference between the observed value and the mean of the posterior distribution for a missing data. The predictive ability of all compared methods was assessed using the leave-one-out cross-validation method to produce a global error measurement. The Mean Squared Prediction Error (MSPE) was obtained averaging the differences between the observed Kd value with the predicted one at each site. Additionally, a punctual prediction error, expressed as percentage of the Kd at each site, was calculated for all methods. It was referred as Site-Specific-Error (SSE) and categorized as smaller than 20%, between 20% and 40%, and greater than 40% of the site Kd, to visually interpret the goodness of prediction. The spatial patterns of the predictions (SEE mapping) of all methods were examined for their validity.

3. Results and Discussion

Predictive modeling [11] is the process that provides a mathematical tool (model) to predict an output from a convenient selected set of data. It demands the full understanding of the undergoing data generating process, model fitting, and its validation. In this study, we compare the results of three approaches to fit predictive regression models for spatial regional data. The soil variables of greater relevance to explain the variability of Kd?s in Cordoba, as indicated by the BRT algorithm, were aluminum oxides, pH, sand percentage and clay percentage. Consequently, all predictive models were fitted using those soil properties as explanatory variables. The Bayesian model with INLA produced the best fit (MSE=3.9% of the average Kd). However, in the cross-validation process to measure predictive ability, the Bayesian model was overcome by the PLS regression with spatial correlated residuals. The lowest MSPE was 18.9% of the average Kd mean (Table 1). The relation between MSE (a measure of goodness of fit), and MSPE (a measure of predictive ability) suggest that the Bayesian modeling with INLA and the SPDE, can overfit data. Thus, the PLS regression, accounting for spatiality, was a good approach in term of global error measurements. This result is probably explained by the collinearity among input variables (being the highest correlation coefficient

equal to -0.75 between sand and clay fractions, p<0.001). The asymmetrical nature of Kd distribution and the multicollinearity among inputs made Boosted-base algorithms competitive. Such machine learning methods had been reported as more robust under these conditions [12]. Boosting-based methods, like RF and GB, have shown their superior performance in various disciplines, but they are commonly used with non-spatial data. Some uses of these methods with spatial[13], requires accounting the spatial structure through distance measurements which are incorporated as explanatory variables in the model [14]. However, as implemented here, spatiality was modeled through an autocorrelation spatial process on the residuals [7] which is easier to implement from a computational perspective.

Model	**MSE[%]**	**MSPE[%]**
ML	25.2	27.3
PLS	10.2	18.9
RF	11.9	19.9
GB	11.2	19.5
INFLA	3.9	25.4

Table 1: Goodness of fit (MSE) and average predictive ability (MSPE) of alternative regression models for glyphosate soil adsorption coefficient as a function of four soil variables. Ordinary (ML) and Partial Least Square regression (PLS), Random Forest (RF) and Generalized Boosted regression (GB); Bayesian modeling (INLA).

A deeper observation of the SSE showed that most Kd were well predicted in most of the sites (Figure 1) and high SSE where closely located (Northwest of Cordoba). In these sites, SSE was far superior to 40% and they consequently raised the global error measurement. However, it is important to highlight that the sites with high SSE (red points Figure 1) had extremely low Kd values and even with high prediction error, they are classified as low Kd sites. Thus, these SSE did not lead to misunderstanding of glyphosate retention.

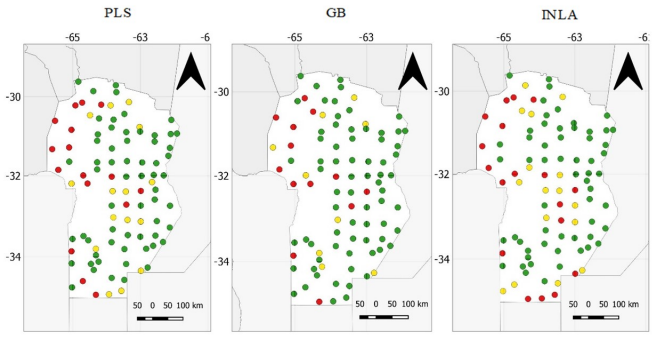

Figure 1: Site specific errors (SSE) for a model of glyphosate soil adsorption coefficient as a function of aluminum oxides, pH, clay, and sand at the site. SSE was categorized as smaller than 20% (green), between 20% and 40% (yellow), and greater than 40% (red) of the site Kd.

In the Fig. 1 we show that the GB model improved the SSE pattern in spatial predictions with respect to both PLS and INLA. GB was the procedure that presented the biggest proportion of prediction errors smaller than 20% of the site mean. Then GB regressions, with spatially correlated errors, can produce competitive results with respect to the frequentist and the Bayesian approaches. The GB advantage is that it requires much less statistical assumptions, and it is easier to automate. However, a better understanding of the process may require models that explicitly show the impact of each input variable like, the produced with INLA which best fitted the observed data. Further work on modeling the Glyphosate Kd distribution in

Córdoba may demand modeling strategies which contemplate mixture of distribution because the Kd values in the Northwest it may come from a different biological process than the others.

4. Conclusions

Visual examination of site-specific prediction errors proved to be an essential tool in assessing the spatial predictions. This study has simultaneously compared alternative methods for spatial interpolation of an environmental property (Glyphosate adsorption coefficient). Results confirm the effectiveness of Bayesian modeling with INLA to obtain good fitting, and the high predictive ability of GB regression in the context of models with several covariables and a spatial autocorrelation process underlying in the random component.

Bibliography

[1] Holland J. M. (2004). *The environmental consequences of adopting conservation tillage in Europe: reviewing the evidence*, Agric. Ecosyst. Environ.,vol.103, no.1, 1-25.

[2] Calvet R. (2005). *Les pesticides dans le sol: conséquences environnementales*. France Agri Editions.

[3] Hang S. , Negro G., Becerra A., and Rampoldi A. E. (2015). *Suelos de Córdoba: Variabilidad de las propiedades del horizonte superficial*. Córdoba, Argentina: Jorge Omar Editorial.

[4] Webster R. and Oliver M. A. (2007). *Geostatistics for environmental scientists*, vol. 1, no. 2. John Wiley & Sons.

[5] Wang X., Ryan Y. Y., and Faraway J. J. (2018). *Bayesian Regression Modeling with INLA*. Chapman and Hall/CRC.

[6] Blangiardo M. and Cameletti M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.

[7] Li J., Heap A. D., Potter A., and Daniell J. J. (2011). *Application of machine learning methods to spatial interpolation of environmental variables*, Environ. Model. Softw., vol. 26, no.12,1647-1659.

[8] Elith J., Graham C. H., Anderson R. P., Dudik M., Ferrier S., Guisan A., Hijmans R. J., Huettmann F., Leathwick J. R., and Lehmann A. (2006). *Novel methods improve prediction of species distributions from occurrence data*, Ecography (Cop.)., vol. 29, no. 2, 129-151.

[9] Pinheiro J., Bates D., DebRoy S. , and Sarkar D.(2010). *R package*, Version 3.

[10] Krainski E. T. and Lindgren F. (2013). *The R-INLA tutorial: SPDE models Warning: work in progress... Suggestions are welcome to elias@ r-inla. org*.

[11] Kuhn M. and Johnson K. (2013). *Applied predictive modeling*, vol. 26. Springer.

[12] Duffy N. and Helmbold D. (2002). *Boosting methods for regression*, Mach. Learn.,vol.47,no. 2-3, 153-200.

[13] Kanevski M., Timonin V., and Pozdnukhov A. (2009). *Machine learning for spatial environmental data: theory, applications, and software*. EPFL press.

[14] Hengl T., Nussbaum M., Wright M. N., Heuvelink G. B. M., and Gräler B. (2018). *Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables*, PeerJ, vol. 6, p. e5518.