Modelling uncertainty using geostatistics, a case study in Ecuador

Chicaiza Elena[*1] **and Buenaño Xavier**[1]

[1]Technical University of Madrid (UPM), Spain

*Corresponding author: gachisbar@gmail.com

Abstract

Spatial data includes positional and thematic components. In this study, the uncertainty analysis has been focused in thematic aspect. A total of 200 soil samples with metal determinations were analyzed. Graphical and numerical tools were employed to summarize and detect critical metals. Copper was identified as the cation of major concern in the study area. Kriging and Co-located Co-Kriging (with Zinc as secondary variable) were employed to carry out predictions. The leave-one-out cross validation method was employed to assess the variogram model used in estimation process, specifically mean square deviation ratio (MSDR) statistic was considered. Finally, unconditional simulations (500 realizations) were computed in order to get mean value and uncertainty of the random field of variable analyzed. The mean of variances obtained from simulations is 1.6 times higher than the mean of kriging variances.

Keywords

Uncertainty, simulation, validation.

I INTRODUCTION

According Griffith et al. (2015), spatial data comprise two components: attribute and location. In this work, we focused the analysis in attribute component.

Currently, the statement that descriptions of spatial phenomena are subject to uncertainty is now generally accepted as explained by Chiles and Delfiner (2009). In this context, Chiles and Delfiner (2009) argues Geostatistics can be defined as "the application of probabilistic methods to regionalized variables".

This study uses R Development Core Team (2013) open-source statistical software. In order to carry out the geostatistical analysis, two specific packages or libraries developed by Ribeiro Jr. and Diggle (2001) (geoR) and Pebesma (2004) (gstat) were employed.

The study area is located in the southeastern of Ecuador, in Zamora province. The area is inside a mining concession where informal mining activities are carried out. The accumulation of heavy metals in soils can bring human health problems, specially when these are exposed directly, for example in this kind of mining activities.

In this study, metals in 200 soil samples were analyzed. A multivariate data exploration analysis was carried out with some metals (Cu, Zn, Cd, and Pb). Kabacoff (2011) recommends the use of relatively new graphical statistical analysis tool named corrgram (see Figure 1) that summarize the relationships between variables very well.

The Figure 1 shows that these variables have a positive correlation (hatching in 45 degrees) with exception of Cd-Pb relation (hatching in -45 degrees); the color saturation in the figure repre-

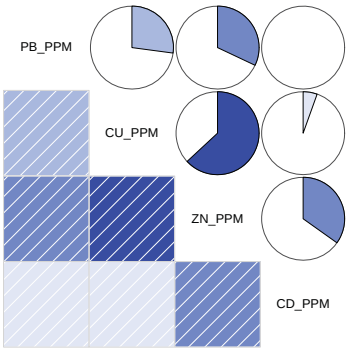

Figure 1: A corrgram of heavy metals soil samples.

sents magnitude of correlation between variables. Notice the significant correlation between Cu and Zn.

II CALCULATIONS

Copper has been identified as an element of major concern. For this reason, the Cu variable was chosen for the spatial estimation by two methods: a) Ordinary Kriging and b) Co-located Co-Kriging. It's interesting to estimate this variable with a covariate, in this case the Zn metal, in order to define the uncertainty threshold between the two methods employed. The Figure 2 shows variograms of Cu and Zn and the cross-variogram.

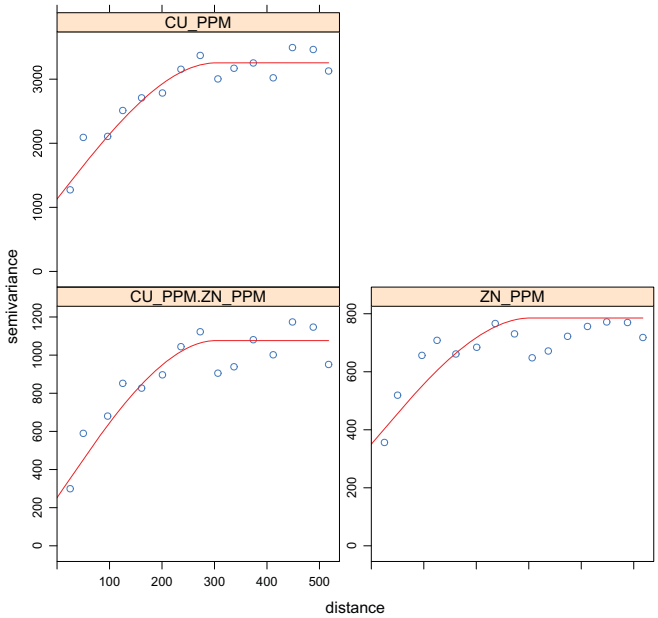

Figure 2: Direct variograms of Cu, Zn, cross-variogram and fitted functions.

A initial exploration data analysis (EDA) of Cu variable was carried out in order to identify outliers and a possible presence of trend that could limit the application of geostatistical theory.

Through directional clouds plotting, spatial trend was discarded. Directional variograms (0, 45, 90, 135 degrees) confirm this assumption.

An important aspect to notice is that exists a nugget effect in the variogram model of both elements analyzed (Cu and Zn). In this case, the nugget effect is originated in the presence of micro-structure.

The same approach was carried out to analyze Zn variable, and the results were similar.

Some variogram fitting techniques can be employed, but according Oliver and Webster (2014) residual maximum likelihood (REML) method is the current best practice. This approach was carried out in the present study.

According Goovaerts (2001), the choice of an approach for uncertainty modeling should be guided by the answer to these questions: What type of uncertainty model is being sought?, What is the support for the uncertainty assessment?, Why do we model uncertainty?.

In this context, it is important to carry out a validation procedure. There are some ways to validate the model decisions, one of them is the cross validation technique.

Bivand et al. (2008) denotes that cross validation splits the data set into two sets: a modeling set and a validation set. When the number of splits is equal to the number of observations, the procedure is called leave-one-out cross validation.

For the cross-validation, leave-one-out technique was used. There are some statistics to analyze the cross-validation. The most interesting is the mean squared deviation ratio (MSDR), which is the mean of the squared errors divided by the corresponding kriging variances according Oliver and Webster (2014). The formula is shown in Equation 1.

$$MSDR = \frac{1}{N}\sum_{i=1}^{N}\frac{\{z(x_i) - \hat{Z}(x_i)\}^2}{\hat{\sigma}_K^2(x_i)} \tag{1}$$

In these equation $z(x_i)$ is the ith datum at x_i, $\hat{Z}(x_i)$ is the kriged prediction there, and $\hat{\sigma}_K^2(x_i)$ is the kriging variance.

Also mean error (ME) and root mean squared error (RMSE) were computed. The results are shown in Table 1.

	ME	**RMSE**	**MSDR**
Ordinary Kriging	0.104	45.318	1.017
Co-Kriging	-0.085	40.688	1.099

Table 1: Uncertainty results with two geostatistical methods.

Another important tool to define uncertainty is simulation. Bivand et al. (2008) defines geostatistical simulation as the simulation of possible realisations of a random field, given the specifications for that random field (e.g. mean structure, residual variogram, intrinsic stationarity) and possibly observation data.

In this context, the drawing of envelopes is an interesting technique. Envelopes were computed assuming a (transformed) Gaussian random field model. According Ribeiro Jr. and Diggle (2001), simulated values are generated at the data locations, given the model parameters obtained by REML. Empirical variogram is computed for each simulation using the same lags as for the original variogram of the data. The envelopes are computed by taking, at each lag, the

maximum and minimum values of the variograms for simulated data. In this case, 99 simulations were carried out. Figure 3 shows the envelopes based in this criteria. The figure shows the range in that we would expect to model the variogram of Copper variable.

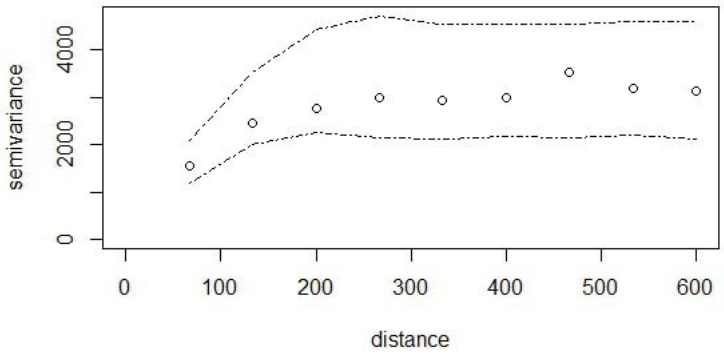

Figure 3: Envelopes of simulations.

Also, unconditional simulations that ignore observations and only reproduce means and prescribed variability, have been carried out. A mean value of 71, corresponding to the mean of prediction map was considering to carried out the simulations.

The results of four (4) first realizations are shown in Figure 4.

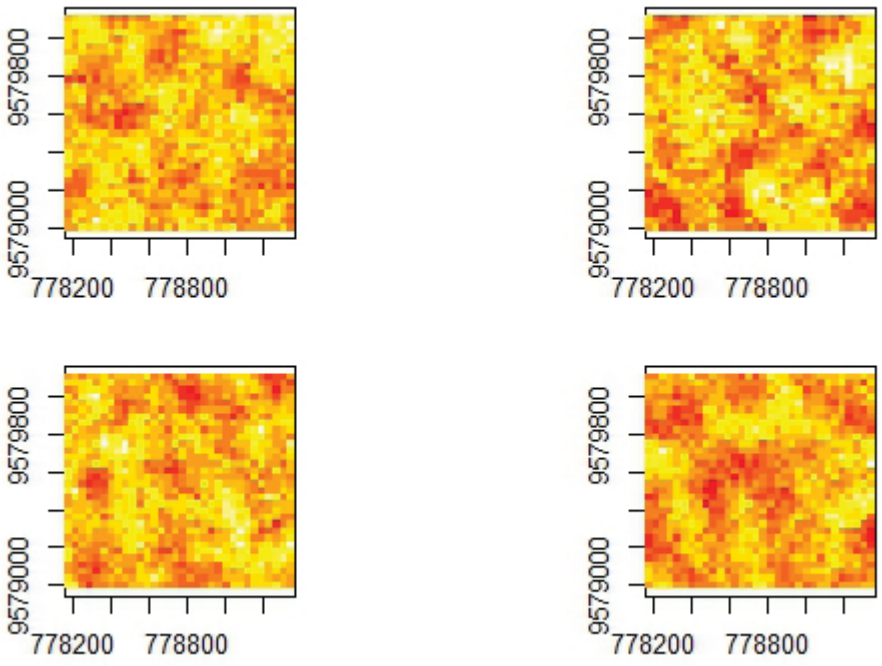

Figure 4: Unconditional simulation of 4 realizations with REML variogram parameters

A total of five hundred (500) realizations were carried out. The mean of variances obtained in these was computed; this mean is 1.6 times higher than the mean of kriging variances. This is a reasonable value, considering that observations were not included in simulations. Nevertheless

the high values of kriging variances and simulations's mean variances could be explained for nugget effect inclusion in the models. The mean of random field generated by simulations was 70.38 and the mean of kriging predictions was 70.96. The proximity of these values reflects that the random field was modeled correctly.

III RESULTS

ME of two methods is almost equal. This situation demonstrates that kriging (in general) is an unbiased estimator. RMSE of ordinary kriging is higher than co-kriging due to the auxiliary variable provides more information to the model prediction.

On the other hand, MSDR of Ordinary Kriging is slightly lower than MSDR of co-kriging, probably due to the characteristics of this statist. Lloyd (2010) in his study concerned with mapping precipitation amount, found that OK outperformed CK (with elevation as the secondary variable) and this was due, at least in part, to the weak global relationship between the two variables.

The map of kriging prediction is shown in Figure 5. The gradient color represents the Copper variable estimation and the contour lines represent variance prediction. The highest variances are located in the corner of southwestern sector.This zone exhibits a lower sampling density in relation with the rest of study area.

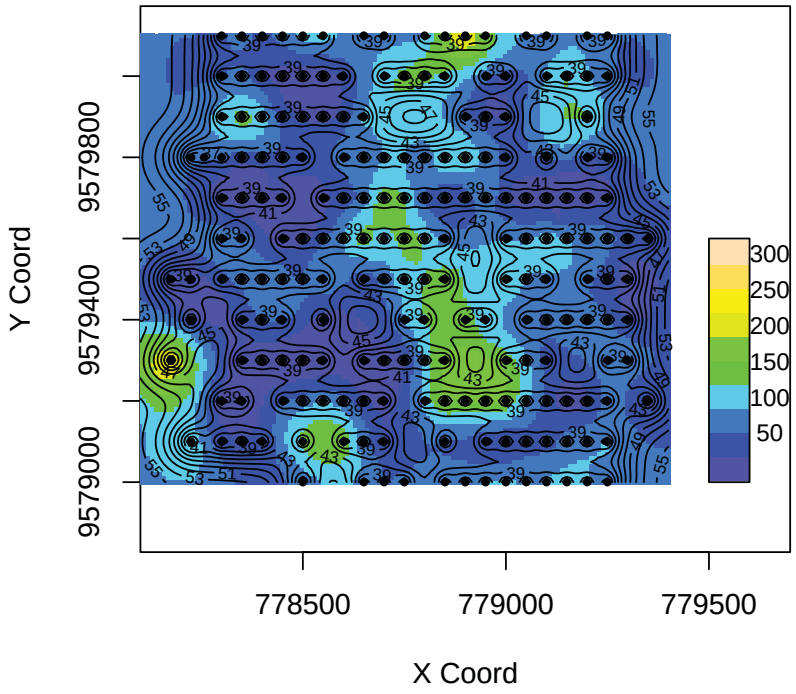

Figure 5: Map of kriging prediction and variance.

References

Bivand R. S., Pebesma E. J., Gomez-Rubio V. (2008). *Applied spatial data analysis with R. Springer*. Springer.

Chiles J.-P., Delfiner P. (2009). *Geostatistics: modeling spatial uncertainty*, Volume 497. John Wiley & Sons.

Goovaerts P. (2001). Geostatistical modelling of uncertainty in soil science. *Geoderma 103*(1), 3–26.

Griffith D., Wong D., Chun Y. (2015). Uncertainty-related research issues in spatial analysis. *Uncertainty Modelling and Quality Control for Spatial Data*, 3.

Kabacoff R. (2011). *R in Action*. Manning Publications Co.

Lloyd C. D. (2010). *geoENV VII – Geostatistics for Environmental Applications*, Chapter Multivariate Interpolation of Monthly Precipitation Amount in the United Kingdom, pp. 27–39. Dordrecht: Springer Netherlands.

Oliver M., Webster R. (2014). A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *Catena 113*, 56–69.

Pebesma E. J. (2004). Multivariable geostatistics in s: the gstat package. *Computers & Geosciences 30*, 683–691.

R Development Core Team (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Ribeiro Jr. P., Diggle P. (2001). geoR: a package for geostatistical analysis. *R-NEWS 1*(2), 15–18.