Geostatistical tools for national-scale soil monitoring

Ben P. Marchant and **R. Murray Lark**
Rothamsted Research
Harpenden, UK
ben.marchant@bbsrc.ac.uk

Nicolas P.A. Saby, Claudy C. Jolivet and **Dominique Arrouays**
Infosol
INRA
Orléans, France

Abstract -**National-scale soil datasets exhibit variation over widely disparate spatial scales. This includes variation because of localised anomalous processes such as point source pollution or geological anomalies which lead to outliers within the dataset. Conventional geostatistical methods are not suited to the analysis of datasets which include outliers. We demonstrate that robust geostatistical methods can be used to predict the underlying variation of a soil property in the presence of outliers whereas copula-based models are appropriate if the outliers are to be included in the predictions.**

Keywords: Soil monitoring, robust geostatistics, copulas

I. INTRODUCTION

There is a need for high intensity soil surveys and monitoring schemes to quantify the variation of soil properties at the national-scale and to ensure that soils remain healthy and capable of supporting human activities and ecosystems. One particular concern is the contamination of soils by heavy metals which have implications for human health and the environment.

The concentrations of heavy metals in the soil vary over disparate spatial scales because of the effects of factors such as geology, relief, diffuse pollution, isolated geological anomalies and point source pollution. Thus the distributions of heavy metals are complex and often include extreme values or outliers. In some situations policy makers might wish to explore the underlying variation of heavy metals so they can broadly assess the quality of soils across the country without these assessments being distorted by isolated anomalies. Policy makers also require assessments of the risk of heavy metal concentrations exceeding regulatory thresholds. Such risk assessment must account for the isolated anomalies.

Observations of soil properties can be interpolated across a study region by geostatistical methods. However conventional geostatistical techniques are unsuitable for datasets which include outliers. The outliers lead to the underlying variation of the dataset being overestimated and the spatial extent of the anomalies is often exaggerated. Geostatistical risk assessments generally assume that the property of interest is (possibly following a transformation) a realization of a Gaussian random process. Such an assumption is not appropriate if the underlying variation of the process has been contaminated by anomalous processes. Therefore in this paper we explore techniques for the spatial analysis of national-scale datasets which include outliers.

Robust geostatistical techniques (Hawkins and Cressie, 1984) can be used to predict the underlying variation of spatial properties which include outliers. Spatial risk assessments of such properties require that the model of variation of the property permits non-Gaussian behavior. Kazianka and Pilz (in press) demonstrated that a copula framework can be used to represent very general patterns of variation of spatial properties. We apply these techniques to observations of soil cadmium (Cd) concentrations from the baseline survey of the French national soil monitoring network (RMQS: Rèseau de Mesures de la Qualitè des Sols).

II. STATISTICAL THEORY

A. Robust Geostatistical Methods

Conventional geostatistical techniques fit a spatial model or variogram to observations of the property of interest and then use this model to predict the property across the study region by kriging. The variogram describes how the expected covariance between a pair of observations varies with their separation distance. Kriging expresses a prediction of the property at an unsampled site as a weighted sum of neighboring observations of the property. The weights are determined from the fitted variogram.

If a conventional geostatistical approach is applied to a property which includes outliers then the outliers distort the variogram so it does not describe the underlying variation of the property. Therefore a number of authors have suggested robust variogram estimators which are reviewed by Lark (2000). These include the Dowd variogram estimator which we use here.

The outliers also distort the kriging procedure and lead to inflated predictions adjacent to the outliers when in reality the anomalous process only acts over very short distances. Hawkins and Cressie (1984) suggested a robust kriging procedure to account for this effect. This procedure is based upon a robust variogram of the underlying process. This variogram is used to determine which of the observations are outliers and a winsorizing procedure is used to divide each outlier into two components. One component is consistent with the underlying variation of the property. The other reflects the contribution of the anomalous process. The effects of the anomalous processes are subtracted from the

Work funded by BBSRC and "GIS Sol"

outliers such that the underlying variation can be predicted across the study region by the standard kriging procedure.

B. Copulas

Spatial risk assessments of soil properties can be based upon model-based geostatistical techniques (Diggle and Ribeiro, 2007). These methods fit statistical models to the observations by likelihood techniques and then use these models to predict at unsampled locations by a procedure equivalent to kriging. The result is a complete description of the distribution of the property at each site from which any statistical descriptor can be extracted. These models are generally based on the assumption that the observations (possibly following a transformation) can be described by a multivariate Gaussian distribution. The Gaussian distribution is preferred because the likelihood and distribution function can be easily calculated. However the Gaussian distribution is often not appropriate for real soil properties particularly when they include outliers and therefore a more general distribution that is equally amenable to analysis is required.

Bárdossy and Li (2008) found a solution to this problem from financial risk assessment literature (e.g. Embrechts, 2009). Non-Gaussian financial properties are often described in a copulas-based framework. The copula is a multivariate function which describes the dependence structure of a property independent of its marginal distribution. In theory the copula itself can be non-Gaussian but we limit ourselves to Gaussian copulas. We generalize our model by combining a Gaussian copula with a non-Gaussian marginal distribution, namely the extreme valve distribution (EVD; Fig. 2). Kazianka and Pilz (in press) demonstrate that such a model can be fitted by likelihood methods and derive a corresponding predictor

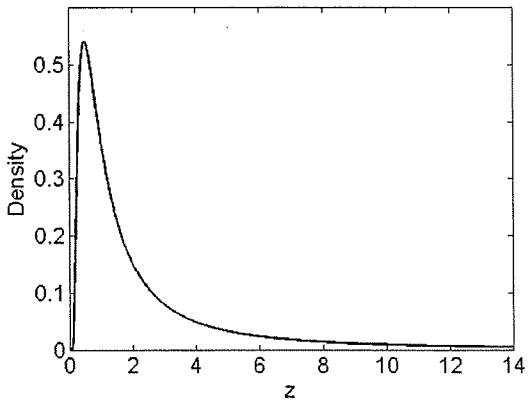

Figure 1. The Extreme value distribution.

III. METHODS

A. The French National Soil Monitoring Network

The RMQS consists of 2200 observations of 40 soil properties on a 16-km regular grid across the 550 000 km2 French metropolitan territory (Saby et al., 2006). The sites of the 1842 observations available at the time of this study are shown in Fig. 1.

The sites were selected at the centre of each 16×16-km cell. At each site, 25 individual core samples were taken of the topsoil (0-30 cm) layer, using an unaligned sampling design within a 20×20-m area. Core samples were bulked to obtain a composite sample for each site. Soil samples were air-dried and sieved to 2 mm before analysis. The total concentration of Cd was determined by inductively coupled plasma mass spectroscopy after dissolution with hydrofluoric and perchloric acids. Soil metal concentrations are known to vary with parent material so the study region was divided into 12 parent material classes based upon the soil database of Europe (King et al., 1995).

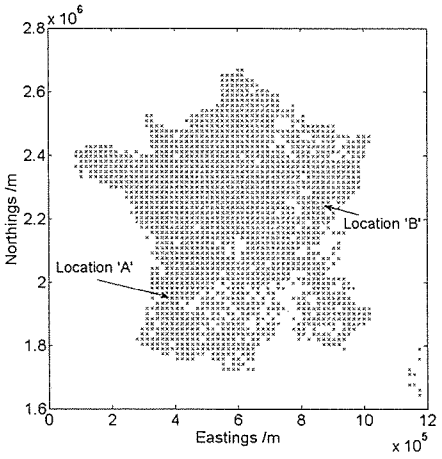

Figure 2. Sampling sites and locations featured in Figure 4.

B. Robust Geostatistical Analysis

The variation of soil Cd with parent material type was modeled by a robust regression procedure and then the robust geostatistical analysis of the residual variation was based upon the algorithm of Hawkins and Cressie (1984). A Matérn variogram was fitted to the observations using Dowd's robust estimator (Dowd, 1984). The Hawkins and Cressie (1984) winsorizing procedure was used to identify the contributions of anomalous processes. These were plotted separately to the remaining underlying variation which was predicted across the study region by ordinary kriging. The validity of the robust variogram model was tested by calculating the mean and median squared standardized prediction errors (SSPE) upon leave-one-out cross-validation of the winsorized observations. Under the assumption that the errors are normally distributed the mean value should be 1.0 and the median 0.45. Confidence limits on these values were determined by calculating the values for multiple simulated realizations of the fitted model. The model was said to be valid if both values were within the 90% confidence limit. Full details of the robust analysis can be found in Marchant et al. (2010). Copula-based models

A copula-based model with a Gaussian dependence structure but a non-Gaussian (EVD) marginal distribution was fitted to the observations by maximum likelihood. The mean of each marginal distribution was permitted to vary according to parent material type. This fitted model was used to predict the concentration of soil Cd across France. The procedure resulted in a prediction of the entire Cd distribution

at each unobserved site from which we extract the probability that the Cd concentration exceeds 0.8 mg/kg which is the regulatory limit in Switzerland.

IV. Results and Discussion

A. Robust Geostatistical Analysis

The robust procedure of Cressie and Hawkins (1984) led to a statistically valid representation of the underlying variation of Cd (mean SSPE = 1.0, median SSPE = 0.43). The underlying variation of soil Cd across France is the result of the combined effects of many different natural and anthropogenic processes (Fig. 3A). Small Cd concentrations are evident in the Landes of Gascony in the south west. This region is known to have a sandy parent material which does not bind to Cd. In contrast large Cd concentrations can be seen in the Jura Mountains in the west where large natural Cd concentrations have previously been observed (Atteia et al., 1994). The effects of diffuse pollution are evident in industrialized regions in the north of France and around Paris. Similarly the causes of Cd outliers (Fig. 3B) include both natural (e.g. geological anomalies in the south of France) and anthropogenic (e.g. outliers caused by industrial sites on the River Seine close to Paris) processes.

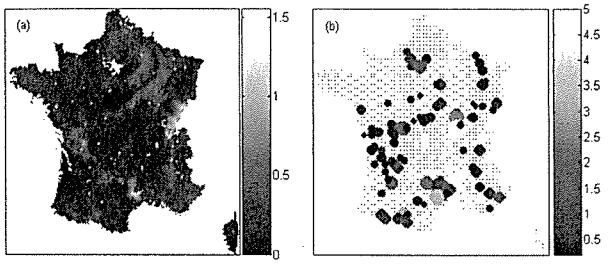

Figure 3. (a) Underlying variation of Cd across France mg/kg, (b) winsorized Cd outliers mg/kg.

B. Copula-based models

The cross-validation statistics for the copula based model fall outside the 90% confidence limit (Table 1). They are however a substantial improvement on the corresponding statistics for a Gaussian model or a Box-Cox trans-Gaussian model (Diggle and Ribeiro, 2007) which we consider to be the current state-of-the art solution to such problems. These findings are indicative of the copula-based model permitting more general behavior than existing models. However the assumptions within the model (primarily that with the exception of the mean the parameters of the marginal distribution are the same at each site) are still too restrictive to represent real soil properties at the national-scale.

The copula-based model is flexible enough to produce markedly different prediction distributions at different sites (Fig. 4). In the Landes of Gascony (Fig. 4A) the tail of the distribution decays rapidly and concentrations greater than 0.5 mg/kg are unlikely. In contrast the distribution at the site in the Jura Mountains (Fig. 4B) has a long tail and concentrations greater than 4.5 mg/kg are plausible. The plot

TABLE I. Cross-validation statistics for risk assessment

Model	Mean SSPE	Median SSPE
Gaussian	1.0	0.09
B-C Gaussian	1.0	0.27
Copula	1.0	0.34

Lower limit on 90 % confidence interval for median is 0.39

of the risk of exceeding 0.8 mg/kg at each site (Fig. 5) reflects many of the features of the underlying variation determined by robust analysis. However the copula-based model allows this quantity and other descriptors to be calculated more accurately than existing techniques.

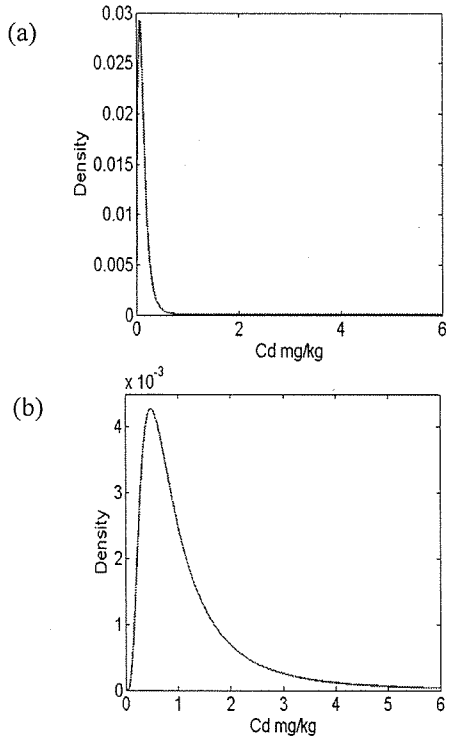

Figure 4. (a) Predicted Cd density function at Location 'A' in Fig. 1. (b) Predicted Cd density function at Location 'B' in Fig. 1.

V. Conclusions

Robust geostatistical methods are a useful tool for understanding the underlying variation of soil properties in the presence of contamination. If the contribution of anomalous processes is removed from observations of soil properties at the national-scale by robust statistical methods then a statistically valid model of the underlying variation can be fitted. However in risk assessments it is often the anomalous processes which are of primary importance. Existing model-based methods cannot adequately represent the variation of soil properties such as Cd in France because outliers within the dataset violate the assumption that the observations are a realization of a Gaussian (or transformed Gaussian) multivariate distribution. Copula-based models are more general than these existing models and produce more

accurate predictions of statistical descriptors of the property of interest.

Copula-based models are a recent innovation in spatial statistics and further work is required to explore how they can be further generalized either by using a non-Gaussian dependence structure or by permitting further variations in marginal distribution across the study region. The general nature of the copula-based model means they are applicable to other non-Gaussian soil properties including those with multi-modal distributions such as soil organic carbon content across areas which are a mixture of mineral and organic soils.

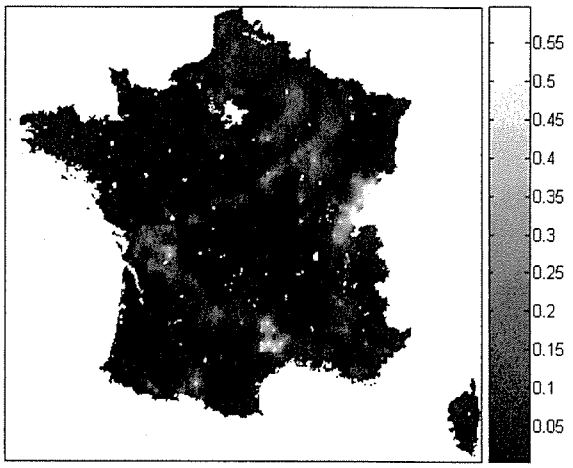

Figure 5. Probability that Cd concentration exceeds 0.8 mg/kg according to Copula model.

REFERENCES

Atteia, O., Dubois, J. and Webster, R. (1994). Geostatistical analysis of soil contamination in the Swiss Jura. *Environmental Pollution*, 86, 315-327.

Bárdossy, A. and Li, J. (2008). Geostatistical interpolation using copulas. *Water Resources Research*, 44, W07412.

Dowd, P.A. (1984). The variogram and kriging: robust and resistant estimators. In: Verly, G., David, M., Journal, A.G. and Marechal, A. (Eds) *Geostatistics for natural resources characterization* Part 1, (pp. 91-106). Dordrecht: D. Reidel.

Diggle, P.J., and Ribeiro Jr., P.J. (2007). *Model-based geostatistics*. New York: Springer.

Embrechts, P. (2009). Copulas: A Personal View. *Journal of Risk and Insurance*, 76, 639-650.

Hawkins, D.M. and Cressie, N. (1984). Robust kriging - A proposal. *Journal of the International Association for Mathematical Geology*, 16, 3-18.

King, D., Burrill, A., Daroussin, J., Le Bas, C., Tavernier, R. and Van Ranst E. (1995). The EU soil geographic database. In: King, D., Jones, R.J.A., Thomasson, A.J., (Eds). *European land information systems for agro-environmental monitoring*. EUR 16232 EN. (pp. 43-60). Luxembourg: Office for Official Publications of the European Communities.

Kazianka, H., and Pilz, J. (in press). Copula-based geostatistical modeling of continuous and discrete data including covariates. *Stochastic Environmental Research and Risk Assessment*.

Lark, R.M. (2000). A comparison of some robust estimators of the variogram for use in soil survey. *European Journal of Soil Science*, 51, 137-157.

Marchant, B.P., Saby, N.P.A., Lark, R.M., Bellamy, P., Jolivet, C.C. and Arrouays, D. (2010). Robust prediction of soil properties at the national scale: Cadmium in French soils. *European Journal of Soil Science*, 61, 144-152.

Saby, N., Arrouays, D., Boulonne, L., Jolivet, C., and Pochot, A. (2006). Geostatistical assessment of Pb in soils around Paris, France. *Science of the Total Environment*, 367, 212-221.