"Spatial prediction of soil properties : the Bayesian Maximum Entropy approach"

D'Or, Dimitri

Abstract
Soil properties play important roles in a lot of environmental issues like diffuse pollution, erosion hazards or precision agriculture. With the developments of soil process models and geographical information systems, the need for accurate knowledge about soil properties becomes more acute. However, while the sources of information become each year more numerous and diversified, they rarely provide us with data at the same time having the required level of spatial and attribute accuracy. An important challenge thus consists in combining those data sources at best so as to meet the high accuracy requirements. The Bayesian Maximum Entropy (BME) approach appears as a potential candidate for achieving this task: it is especially designed for managing simultaneously data of various nature and quality ("hard" and "soft" data, continuous or categorical). It relies on a two-steps procedure involving an objective way for obtaining a prior distribution in accordance with the general knowledge...

Document type : Thèse (Dissertation)

Référence bibliographique
D'Or, Dimitri. Spatial prediction of soil properties : the Bayesian Maximum Entropy approach. Prom. : Bogaert, Patrick
Spatial prediction of soil properties, 
the Bayesian Maximum Entropy approach

Dimitri D’Or

Mai 2003

Thèse présentée en vue de l’obtention 
du grade de docteur en Sciences Agronomiques 
et Ingénierie Biologique

Promoteur : P. Bogaert
Conclusion and perspectives
The generalizing power of BME

In the last 50 years, the joint upcoming of geostatistics and increase of computational power allowed to make substantial progress in the analysis of spatially correlated variables. The congenial idea of the kriging estimator is now declined in a multitude of variants, each adapted to handle a small set of particular situations. The profusion of techniques (or sometimes should we say, tricks) does not facilitate the work of practitioners seeking the best solution for their problem. Neither it does present a global view of the situation, allowing to search for more general solutions.

In fact, kriging techniques suffer from some strong limitations: (i) the kriging predictor is only the best among the set of linear predictors, (ii) it is only the best in an absolute sense, if the variables follow a Gaussian distribution, and (iii) kriging is not able to incorporate on a sound way soft or categorical information.

In such situations, a salutary attitude often consists in trying to find a more general framework in which kriging could be embedded. Doing this would allow to relax the strong distributional assumptions of kriging and thus allow to tackle efficiently a wider range of problematics. In his paper on model-based geostatistics, Diggle et al. (1998), proposes the use of generalized linear models (GLM) in order to incorporate categorical information or non normally distributed prediction errors. However, soft information is still excluded of this prediction process.

Another attempt to bring a broader view to the general (spatial) prediction paradigm was proposed by Christakos in the early Nineties (Christakos, 1990), combining Maximum Entropy and Bayesian procedures in order to produce a very general formulation for the spatial predictor. This approach benefits from substantial advantages over classic geostatistical methods:

- The Maximum Entropy step first ensures that all the available general information is taken into account, while no spurious details are added. Compared to other Bayesian methods, it provides us with an objective tool for determining the most appropriate prior with respect to the information at hand. This step results in a joint general pdf $f_G(\cdot)$ that meets the constraints given by the general knowledge base $K_G$. The constraints may, e.g., be formulated as statistical moments (mean, covariance function) or physical laws;

- The operational Bayesian conditioning principle proposes a very flexible way for updating the prior distribution with the specific knowledge $K_S$ at hand. At the opposite to the standard Bayesianism which is based on a single probability model $f_S(\cdot)$, operational Bayesianism allows to dissociate the choice of $f_G(\cdot)$ from the choice
The generalizing power of BME

of an operator $\Xi_S$ adapted to the available $K_S$ (see Eq. 1.41). This results in most cases in highly non-linear predictors which can be described as Best Predictors (among the whole set of linear and non-linear predictors);

- The preceding feature opens possibilities for incorporating on a sound way a **wide variety of soft information** under the form of intervals, pdf's, models or charts. Soft information available at the prediction location is treated accordingly. Contrary to classic geostatistical methods, BME does not require any transformation or coding of these types of information;

- The **final estimate** of the BME approach is a **complete pdf** at each prediction location, thus allowing to dissociate the decision from the prediction step. From the set of posterior pdf's, various types of maps can easily be build by computing indicators that are adapted to the final objective of the map. Examples of such indicators are the mean, the median or the mode of the distribution, its variance, probabilities to exceed a threshold or confidence intervals. Moreover, these pdf's are not the result of a discretization, as is obtained, e.g., from IK;

- The generality of the BME approach is especially clear as a standard technique as kriging is only a particular case of the BME approach, obtained when $K_G$ is reduced to the moments of the two first orders and when only hard data are involved;

- The epistemic-based reasoning scheme of BME allows for an **easy transposition to the categorical field**, without any transformation of the data to a continuous domain. Again, the Maximum Entropy and operational Bayesian conditionalization principle produce a prior pdf that accounts for all the available $K_G$ (often presented in the form of bivariate probabilities) and allows the incorporation of a wide variety of soft data.

All these features emphasize the flexibility of BME on several levels: knowledge processing rules (with $f_G(\cdot)$ and $\Xi_S$), variables (continuous or categorical), data quality (hard and soft), and final estimates (easily derived from an entire pdf). This makes BME a very powerful method able to tackle a wide variety of problematics using the same reasoning scheme. As a proof of this, applications of BME were recently achieved, e.g., in atmospheric quality studies (Christakos et al., 2001) and in health studies (Christakos and Kolovos, 1999; Christakos and Serre, 2000).
BME and soil science

The main objectives of this research were (i) to revisit the basic BME concepts and equations, (ii) to contribute to the development of BME towards categorical variables and mixture of categorical and continuous variables, and (iii) to illustrate the power and flexibility of BME through a series of applications. In this section, we will summarise the major contributions of this research to the achievement of these objectives.

Our main contribution was to apply the BME approach for the first time in the field of soil science. After the mining engineers, soil scientists were among the very first to show an interest for the geostatistical approach. The study of the spatial distribution of soils and their properties using geostatistics has been the object of research works since the early Eighties (Burgess and Webster, 1980). The expectation is great in this area of research for new methodologies that are able to incorporate soft information and categorical variables as attested, e.g., by the works of Diggle et al. (1998); McBratney and Odeh (1997); Bierkens and Burrough (1993a); De Gruijter et al. (1997).

Simulated studies

In a first simulated experiment, we examined the prediction of sand, silt and clay fractions as these properties are often of crucial importance in many environmental applications such as the assessment of sensitivity to pollutants or the evaluation of cultural properties. We combined a limited hard data set with the exhaustive soft information provided by a soil map to produce sand, silt and clay maps. Several geostatistical techniques (SK, SKMI and BME) were compared. BME was shown to produce the best estimates, as revealed by the mean error close to zero (no bias), the reduced root mean squared error (increased accuracy), and the reduced spatial correlation of the residuals (better use of the available information). Additionally, it is worth mentioning that the prediction was not only better on a global scale as shown by these indicators, but also locally more relevant, as could be checked on the texture maps.

A second topic of major concern for soil scientists is the making of reliable soil maps. Focusing on the production of soil texture maps, we (D’Or et al., 2001, results not shown in this document) showed that BME offers an adequate methodology. Indeed, as usual in soil surveys, a reduced number of soil profile descriptions yields accurate information about the soil properties, while numerous auger borings provide less reliable information under the form, e.g., of intervals of values. The sand, silt and clay contents were estimated on a grid and then classified according to the Belgian texture triangle in order to obtain a soil texture map. Using the simulated soil map as
reference, we could establish that BME produces more accurate soil texture maps than SK.

Real scale applications

A second enhancement we achieved was to propose real scale applications. Indeed, till 1999, due to the inexistence of an adapted set of informatical routines to perform the computations, most of the illustrations of the method were made on small simulated experiments. But once the step of objective comparison was achieved, there was a need for applying BME to real scale problems.

This lead us to perform the estimation of the soil texture fractions at a regional scale using simultaneously punctual measurements (hard data) and a soil map (soft data). The results of this experiment reinforce the conclusions drawn on the simulated case studies: BME produces less biased and more accurate estimates. Moreover, the corresponding maps appeared to have more realistic spatial patterns: spatial variation is observed within a soil map unit, while OK is only able to provide a constant value within each mapping unit.

In addition, we also examined a case where no hard data were available and could conclude that BME proposes a sound way for predicting the spatial variation of a variable within a soil mapping unit, even if only soft information is at hand. Again, BME was able to provide realistic within mapping unit spatial variation, while other methods failed in achieving this goal.

Compositional data analysis

In parallel, since our central theme for the applications was the prediction of the soil texture fractions, we were confronted to the problem of compositional data. Indeed, as they are linked to each other by a sum-to-one constraint, the sand, silt and clay fractions would better be estimated together. Several methods specifically dedicated to the analysis of compositional data in a spatial context were reviewed: separate kriging, basis method, additive log-ratio transform, compositional kriging and quadratic optimization programming. Unfortunately, none of them is able to incorporate soft information. Moreover, the accuracy of the predictions could not be proved to be different from a method to another, even if some of them were questionable on a theoretical level.

In the specific context of soil texture fractions, the situation is made even worse by the complex definition of the texture classes on the texture triangle. In order to ensure that a predicted triplet of value is in agreement
with the texture class provided at the same location, we proposed a variant to the BME approach (BME/MC) allowing to take into account the complex shape of the texture classes. This method indirectly takes into account the constraint between the variable by making use of a Monte Carlo algorithm for computing the conditional expectations with respect to the texture class. We have shown that BME/MC provides estimates of triplets which are summing very close to one. Moreover, the classification error rate is drastically reduced, the remaining errors being only due to small deviations at the border between two adjacent texture classes. Up to now, BME/MC thus appears to be the only method able to deal simultaneously with soft data and compositional variables, even if the sum-to-one constraint is only indirectly taken into account.

Categorical variables analysis

In the quest for generalizing the BME approach, we contributed to the development of a theoretical framework (called BME/CAT) adapted to the analysis of categorical variables. This framework, initially proposed by Bogaert (2002), benefits from the same advantages as the BME approach for continuous variables (flexibility, sound theoretical foundations, etc.), but relies on the manipulation of multidimensional contingency tables. The set up of this technique necessitated to make links with classic statistical approaches as the log-linear models which allowed the use of well-known techniques as the iterative scaling algorithm to implement the Maximum Entropy step.

On a theoretical level, BME/CAT proposes solutions to some recurrent problems of IK. Indeed, BME/CAT estimates never suffer from order relation problems. Moreover, BME automatically ensures that the posterior conditional distributions are valid (with \( p_{ik}(x_0) \in [0, 1] \) \( \forall i \), and \( \sum_{k=1}^{n_k} p_k(x_0) = 1 \)). As in the continuous case, BME/CAT produces non-linear predictors, thus enlarging the range of candidates beyond the particular category of linear estimators classically used in geostatistics. Finally, BME/CAT relies on a completely non-parametric Probability Model (PM) to characterize the spatial structure, allowing to avoid the often questionable fitting of a LMC.

A first application of BME/CAT was presented at the 4th European Conference for Environmental Applications (GeoEnv IV) under the title "Combining categorical information with the Bayesian Maximum Entropy approach" (D’Or and Bogaert, 2003a). In this dissertation, we expose a more detailed case study dealing with the prediction of water table classes in the Ooypolder area (Netherlands).

Our first issue consisted in comparing the performance of BME/CAT and IK. Even if the accuracy of the results was not drastically changed in terms
of percentage of correctly classified locations, we showed that the mapping of the maximum probability category is more realistic with BME/CAT, yielding more indented contours and identifying small inclusions. In addition, the uncertainty associated with the selection of the class with maximum probability was shown to be significantly reduced when using BME/CAT. Finally, the conditional distributions obtained from BME/CAT appear to be more consistent than those yielded by IK. Indeed, with IK, only the categories represented in the neighborhood of the prediction location are attributed a non-null probability. This is obviously absurd since there is no valuable reason that a category has no chance to be observed when it is not represented in the surroundings of the prediction location. With BME/CAT, even the most improbable categories according to the local configuration of the data receive a small probability of occurrence.

Secondly, we examined the combination of various information sources with BME/CAT. In the Ooypolder area, we have at hand a soil map and a finite sample set of accurate data. The soil map is considered as less reliable since it had to undergo classification and aggregation steps which introduced some discrepancies between the mapping unit and the effective water table class observed at a given location. Nevertheless, the uncertainty of the soil map information can be characterized using the joint probability table between the soil map and the sample data locations. On the basis of a small selection of typical situations, we showed that BME/CAT is able to process without any loss of information the whole set of available information. BME/CAT succeeds in establishing logical rules for treating contradictory information and produces results that are corroborating intuitive feeling. Data of various quality and contradictory information sources are managed in a sound and transparent way.

Choice of an estimate

In a geostatistical soil mapping context, we proposed to separate the decision from the estimation step. This would allow to produce maps that are more goal-oriented and thus fit better the requirements of the stake-holder. In addition, a same set of estimated pdf's could be at the origin of several maps, thus sparing considerable processing time and man-power. In connection to this, BME appears to be particularly suited for achieving the estimation step as it automatically provides as estimate an entire pdf at each prediction location.

For categorical variables, the choice of the maximum probability category was questioned. Although this choice is justified by the use of several objective functions (minimum probability of error, minimum Bayes risk and minimum Kullback-Leibler distance), we tested its robustness to changes in
the definition of the loss function used to compute the minimum Bayes risk. It resulted from that experiment that BME and IK were robust to these changes as the selected category was modified at only a marginal number of locations. Moreover, those changes resulted only in half the total number of changes in a better classification. Finally, we showed once more that IK behaved on an inconsistent way, selecting sometimes a category for which the conditional probability of occurrence was equal to zero.

**Perspectives**

Our research opens real perspectives for the mapping of soil properties. Substantial progress are conceivable on the theoretical as well as on the practical levels.

**Information coding**

Most countries still do not dispose of exhaustive soil maps. The prohibitive costs in man power and sampling analyses often have stopped national soil surveys which are now replaced by smaller studies following the needs. However, soil maps remain useful sources of information in a wide variety of fields ranging from agriculture and forestry to facilities building or water tables protection. It is worth noting that geostatistical methods, contrary to what soil surveyors do, are still not able to adequately incorporate important but complex features as the topographical position, the existence of periodical soil patterns (catena's), the distance from a river or the presence of barriers. Hence, there is a crucial need in developing methodologies to build logical rules that could translate such informative features into mathematically valuable expressions.

**The soil map to come**

New ways for conveying soil related information will also be necessary in this beginning of millennium. As stated by Zhu et al. (2001), conventional soil surveys are not adapted to provide the detailed (high-resolution) soil information required by some environmental modeling and crop management modeling. They actually suffer from generalization of soils in the spatial as well as in the parameter domain (Zhu et al., 2001). Indeed, traditional soil maps are often designed to be used in a given range of closely related applications. For example, the Belgian soil map was principally built as a tool for classifying the rural areas according to their agricultural potentialities. As the range of applications wherein soil maps are involved never ceased to grow in the last years, users are rather demanding multipurpose soil data
bases from which they can extract the relevant information according to the objective of their study. The manual soil map production also slows down the efficiency of the necessary updating of soil surveys. All these reason speak for the emergence of new methods for digital mapping. Due to its indisputable flexibility, BME could be a valuable candidate for such a task: as we have shown, a wide variety of data types (continuous and categorical, hard and soft) can be used simultaneously. Moreover, newly collected data can easily be incorporated: it only requires to run again the routines while taking these data into account. Finally, as it produces an entire pdf as estimate, BME appears to be suited to the design of multipurpose maps, since the user then can easily select the adequate statistics according to the objectives of the map.

Compositional data

While BME/MC proved its efficiency in taking into account complex relations between the components of a compositional set, it still does not ensure that the components are summing up to one. In the future, it would be interesting to incorporate more explicitly this constraint into the prediction process. On the same manner OK is a limiting case of BME when only hard data are available, compositional kriging should be a limiting case of BME under constraints in the same conditions.

Towards a unified framework for continuous and categorical variables

After having contributed to the development of BME towards categorical variables, the next step in the direction of a global method was to build a unified framework wherein continuous and categorical variables could be jointly estimated. Just as generalized linear models appear to be a global concept in which are embedded methods as diverse as standard linear models, logit and probit models, it is interesting to write the basic expressions of BME in a unified form.

The advantage of such a method is, e.g., to offer the possibility of estimating the joint continuous/categorical pdf at each prediction location. This pdf is obviously more informative than the set of continuous and categorical marginal pdf’s as it provides explicitly a link between both kinds of variables.

The solution we propose offers a framework that is very similar to the one used for continuous variables. It allows to incorporate easily advanced forms of general knowledge in the form of bivariate categorical probabilities and conditional continuous expectations and covariances. It is also able to
use the same variety of soft information as both the categorical and the
continuous cases separately.

Moreover, the simple categorical and continuous cases are easily derived
from this general framework, thus emphasizing its generalization power and
its flexibility.

Links with other (geo)statistical methods like, e.g., stratified kriging,
regression kriging or GLM’s, should also be established in order to show
that methods previously considered as distinct are in fact only particular
cases of this unified BME framework.

Finally, the effects of the way the information is coded should also be
investigated. We have seen, for example, that interval- or pdf-type soft
information could easily be derived from a categorical variable (e.g., interval
for the sand content based on the knowledge of the texture class). It would
be interesting to compare BME/MIX and other BME algorithms on the
basis of their performance for retrieving the maximum of information from
the data set.

As a final word, this work opens new perspectives for the spatial analysis
of soil data. Hopefully, it will give birth to important enhancements in
matter of digital soil mapping. In this field, the integration of multiple data
sources is of crucial interest to increase the accuracy and the reliability of
the soil maps. In a near future, important political decisions will have to be
taken in relation with the present and forthcoming environmental challenges.
Those decisions need to be based on reliable information. Providing this
information is certainly one part of the task of the XXIth Century soil
scientist.