

BAYESIAN DATA FUSION APPLIED TO SOIL DRAINAGE CLASSES SPATIAL MAPPING

Sarah Gengler & Patrick Bogaert

*Earth and Life Institute, Environmental Sciences. Université catholique de Louvain,
Croix du Sud 2/L7.05.16, B-1348 Louvain-la-Neuve, Belgium*

Abstract. Soil drainage classes spatial mapping is of great interest since drainage has direct effects on crop productivity and hydrological modelling. However, the prediction of this categorical variable often requires a laborious and expensive sampling over large areas. There is thus a need for a methodology able to combine several sources of information to improve the prediction. First developed to predict continuous variables, Bayesian Maximum Entropy (BME) has become a complete framework in the context of space-time prediction since it has been extended to predict categorical variables and mixed random fields. This method proposes solutions to combine several sources of data whatever the nature of the information. However, the various attempts that were made for adapting the BME methodology to categorical variables and mixed random fields faced some limitations, as a high computational burden. The main objective of this paper is to overcome this limitation by generalizing the Bayesian Data Fusion (BDF) theoretical framework to categorical variables, which is somehow a simplification of the BME method through the convenient conditional independence hypothesis. The BDF methodology for categorical variables is first described and then applied to a practical case study : the estimation of soil drainage classes using a soil map and point observations in the sandy area of Flanders around the city of Mechelen (Belgium). The BDF approach is compared to BME along with more classical approaches, as Indicator CoKriging (ICK) and logistic regression. Estimators are compared using various indicators, namely the Percentage of Correctly Classified locations (PCC) and the Average Highest Probability (AHP). Although BDF methodology for categorical variables is somehow a simplification of BME approach, both methods lead to very close results and have strong advantages compared to ICK and logistic regression.

Keywords. Drainage classes, Bayesian data fusion, categorical data, spatial estimation.

1 Introduction

Soil drainage is an important soil property since it has direct effects on plant growth, water flow and solute transport in soils [12]. Soil drainage classes spatial mapping is thus of great interest but it often becomes laborious and expensive using classical soil mapping methods because of the intensive sampling it requires over large areas [13]. As an increasing diversity of information sources is observed in environmental sciences, it is of great

interest to be able to account for all of them at the same time and to deal with possibly contradictory information as well. In a spatial prediction context, this idea is not new. Bayesian Maximum Entropy brought solutions to the problem of predicting categorical variables by combining several sources of information. This method has already been applied to spatially predict soil drainage classes consistently and inexpensively [8]. First developed to predict continuous variables [5], BME has become a complete framework in the context of space-time prediction since it has been extended to predict categorical variables and mixed random fields. However, the various attempts that were made for adapting the BME methodology to categorical variables and mixed random fields faced several limitations, including high computational burden [7, 15]. Bayesian Data Fusion (BDF), which also combines data fusion concept with geostatistical methods, overcomes this limitation but has been developed for continuous variables only [2]. The main objective of this paper is thus to generalize the BDF theoretical framework to categorical variables as categorical data are found in a wide variety of important applications in environmental sciences [3]. An implementation of the BDF methodology for categorical variables is proposed to estimate soil drainage classes from a real data set in the sandy area of Flanders around the city of Mechelen (Belgium). Two sources of information are available : point observations of drainage classes that can be considered as error-free (hard data) and a pre-existing soil map which is spatially exhaustive but has limited accuracy (soft data) [7]. The BDF results are compared to BME along with more classical approaches (Indicator CoKriging, logistic regression) using various indicators, namely the Percentage of Correctly Classified locations (PCC) and the Average Highest Probability (AHP).

2 Bayesian Data fusion for categorical variables

BDF aims at combining and reconciling multiple information sources relative to a same variable of interest with the goal of increasing the quality of the prediction. This method has been widely studied in environmental sciences but essentially for the spatial prediction of continuous variables [9]. An extension is proposed hereafter for categorical variables. For the sake of brevity, only the most relevant theoretical results are presented. A complete description of the theory for continuous variables can be found in Bogaert and Fusbender (2007) [2].

Let us define a categorical random variable \mathbf{Z} on a continuous spatial domain D and $\mathbf{x} = \{x_0, \dots, x_n\}$ a location vector where a secondary variable $\mathbf{y}' = \{y_0, \dots, y_n\}$ is sampled. We assume that the observable Y_i 's are linked to the Z_i 's through an error-like model, so that $\mathbf{Y} = \mathbf{g}(\mathbf{Z}) + \mathbf{E}$, where $\mathbf{g}(\cdot)$ are monotonic functionals and $\mathbf{E}' = (E_1, \dots, E_n)$ is a random errors vector which is independent from \mathbf{Z} . Let us define $p(\mathbf{z}) = p(z_0, \dots, z_n)$ as the joint probability function. What is sought for is

$$p(z_0|\mathbf{y}) = \sum_{z_1} \cdots \sum_{z_n} p(\mathbf{z}|\mathbf{y}) \quad (1)$$

From Baye's theorem we know that

$$p(\mathbf{z}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{z})p(\mathbf{z})}{\sum_{z_0} \cdots \sum_{z_n} p(\mathbf{y}|\mathbf{z})p(\mathbf{z})} \propto p(\mathbf{y}|\mathbf{z})p(\mathbf{z}) \quad (2)$$

With the same reasoning than Bogaert and Fasbender (2007) [2] and assuming $E_0 \perp E_1 \perp \dots \perp E_n$, that corresponds to the convenient conditional independence hypothesis $(Y_0 \perp \dots \perp Y_n)|\mathbf{z}$, we obtain

$$p(\mathbf{y}|\mathbf{z}) = \prod_{i=0}^n p(y_i|z_i) \quad (3)$$

with

$$p(y_i|z_i) = \frac{p(z_i|y_i)p(y_i)}{p(z_i)} \propto \frac{p(z_i|y_i)}{p(z_i)} \quad (4)$$

Plugging these results in (1), we obtain

$$p(z_0|\mathbf{y}) \propto \frac{p(z_0|y_0)}{p(z_0)} \sum_{z_1} \cdots \sum_{z_n} \left[p(\mathbf{z}) \prod_{i=1}^n \frac{p(z_i|y_i)}{p(z_i)} \right] \quad (5)$$

If needed, equation (5) can be slightly generalized in the case of m_i collocated information at each $x_0, x_1 \dots x_n$ with

$$p(z_0|\mathbf{y}) \propto \prod_{j_0=1}^{m_0} \frac{p(z_0|y_0)}{p(z_0)} \sum_{z_1} \cdots \sum_{z_n} \left[p(\mathbf{z}) \prod_{i=1}^n \prod_{j_i=1}^{m_i} \frac{p(z_i|y_{i,j_i})}{p(z_i)} \right] \quad (6)$$

Let us consider $Z = (Z_0, Z_s, Z_u)$ as the values of drainage classes, where Z_0 is the unknown drainage class at prediction location, Z_s refers to locations $\mathbf{x}_s = (x_1, \dots, x_n)$ where both Z_i 's (hard data) and Y_i 's (soft data) are known and Z_u to locations $\mathbf{x}_u = (x_n, \dots, x_{n+m})$ where only soft data are sampled. Classical probability calculus lead to

$$p(z_0|\mathbf{z}_s, \mathbf{y}) \propto \prod_{i=0}^n \frac{p(z_i|y_i)}{p(z_i)} \sum_{z_{n+1}} \cdots \sum_{z_{n+m}} \left[p(\mathbf{z}) \prod_{j=n+1}^{n+m} \frac{p(z_j|y_j)}{p(z_j)} \right] \quad (7)$$

As both Z_i 's and Y_i 's are sampled at location $\mathbf{x}_s = (x_1, \dots, x_n)$, $\prod_{i=1}^n \frac{p(z_i|y_i)}{p(z_i)}$ behaves as a constant. We finally obtain

$$p(z_0|\mathbf{z}_s, \mathbf{y}_u) \propto \frac{p(z_0|y_0)}{p(z_0)} \sum_{z_{n+1}} \cdots \sum_{z_{n+m}} \left[p(\mathbf{z}) \prod_{j=n+1}^{n+m} \frac{p(z_j|y_j)}{p(z_j)} \right] \quad (8)$$

In the context of soil drainage classes around Mechelen, two information sources are available : point observations that can be assumed as error-free (Z_i 's) and an exhaustive pre-existing soil map with limited accuracy (Y_i 's). As we are here in a peculiar case where the soil map is spatially exhaustive, the most valuable soft information when predicting drainage class at location x_0 is y_0 . Thus, simplifying equation (8) by accounting only for y_0 and the Z_i 's sampled at $\mathbf{x}_s = (x_1, \dots, x_n)$, leads to

$$p(z_0|\mathbf{z}_s, y_0) \propto \frac{p(z_0|y_0)}{p(z_0)} p(z_0, \mathbf{z}_s) \quad (9)$$

In order to predict a conditional probability function $p(z_0|\mathbf{z}_s, \mathbf{y})$, the BME methodology estimates first the complete joint probability density function $p(z_0, \mathbf{z}_s, \mathbf{y})$ by relying on a maximum entropy estimation procedure. For each prediction point, the entropy has to be maximized on a contingency table. Computational burden will thus increase exponentially with the number of data locations in the neighbourhood and the number of secondary variables taken into account [8]. The BDF methodology alleviates the need of inference for the joint probability density function $p(z_0, \mathbf{z}_s, \mathbf{y})$, since it relies on the convenient conditional independence hypothesis (see equation (3)).

3 Case study

3.1 Dataset

The study zone is an area of 30 by 30 km^2 located in the sandy area of Flanders around the city of Mechelen (Belgium) (Figure 1). The main type of soil is Spodosol. There are more clayey soils classified as Fluvents in the alluvial plains of Grote Nete, Dijle and Zenne. Topography is flat with a mean elevation of 15 m a.s.l [7].

Two sources of information are available for the prediction of soil drainage classes : (i) the Aardewerk database [14], from which we use 347 point descriptions of soil profile that can be considered as error-free (hard data) (Figure 1 a) and (ii) a pre-existing soil map which is spatially exhaustive but has limited accuracy (soft data) (Figure 1 b).

Drainage classes are obtained by grouping the original nine drainage classes into three classes : c_1 ="excessive to good drainage", c_2 ="good to moderately bad drainage" and c_3 ="moderately bad to very bad drainage" [7]. As the digitized map is of limited accuracy, conflicting information appears when comparing it to the Aardewerk database. As it can be seen from table 1, the two sources of information may give contradictory information about the drainage class.

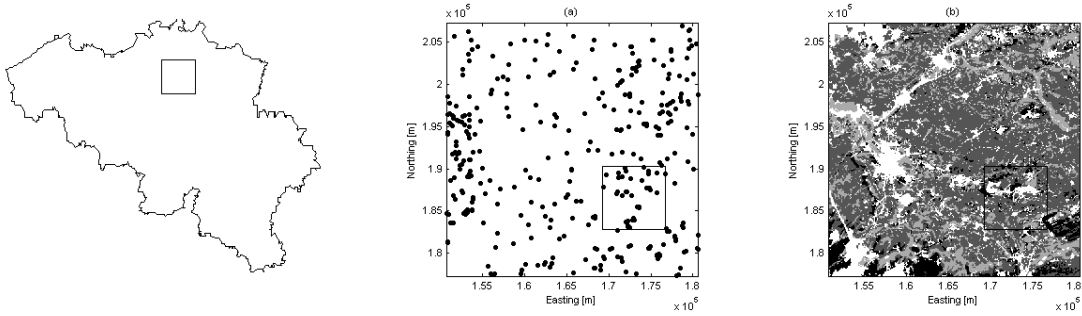


Figure 1: Study area. (a) Sampling locations for the Aardewerk database. (b) Three-classes digital drainage map, with drainage ranging from good (black areas) to bad (light gray areas). White areas are built zones. The superimposed square is the area of interest for merging the Aardewerk database and the digital drainage map [7].

Table 1: $\hat{P}(Z=c_i | Y=c_j)$ in %

	j=1	j=2	j=3
i=1	73.7	7.2	2.9
i=2	24.6	71.4	22.5
i=3	1.7	21.4	74.6

However, it is worthwhile to take into account both sources of information for the prediction since the Aardewerk database is error-free but scarcely sampled while the soil map is somehow inaccurate but spatially exhaustive.

3.2 Results

The mapping area for merging Aardewerk database and the soil map is an area of 7.5 by 7.5 km^2 around Tremelo (superimposed square in Figure 1) [7]. BDF methodology has been applied for obtaining the probabilities of the three soil drainage classes $p(z_0|\mathbf{z}_s, y_0)$ at the 10201 nodes of a square grid. For the sake of comparison, three other methods are used as well : BME, ICK and logistic regression.

BME and BDF methods lead to very close results, as it can be observed from the maps of the maximum probability drainage classes in Figures 2 c and d. Around locations where the Aardewerk samples (hard data) are available, the pre-existing soil map (soft data) is updated from the hard data, while the soil map remains unchanged at locations where there is no hard data at hand.

The map obtained with logistic regression (Figure 2 f) is very smooth and the one

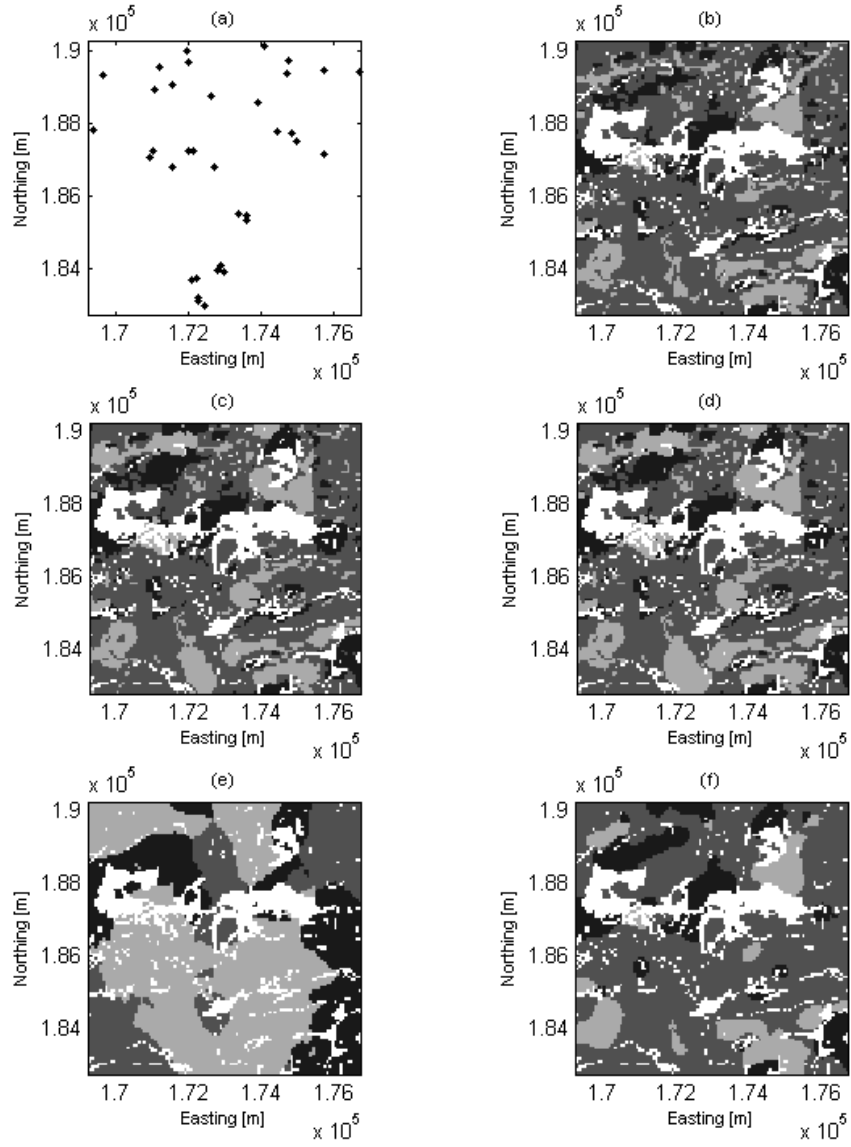


Figure 2: Maps of the maximum probability drainage classes, with (a) the Aardewerk sampled location, (b) digital drainage map, (c) BME map, (d) BDF map, (e) Indicator CoKriging map, (f) Logistic regression map.

obtained with ICK (Figure 2 e) shows even less details. The superiority of the BME and BDF methods over ICK and logistic regression is clear by the light of these results. Compared to the ICK approach which is restricted to process hard data and do not take into account any soft information [5, 10], BDF and BME methods process all information whatever their intrinsic quality (hard or soft). In this case study, ICK neglects the pre-existing soil map information and take into account the few hard data at hand, thus leading to this very smooth map. If one wants to use this source of information in a kriging technique, a soft ICK can be applied [11]. However, due to the exactitude property of kriging techniques, this would return the soft information as the prediction, since the soil map is spatially exhaustive. Regarding the logistic regression approach, it does not account for the spatial structure of the data. It thus results in a significant loss of information in this case of soil drainage classes prediction.

3.2.1 Global comparison criteria

In order to assess the ability of the estimators to correctly predict observed drainage soil classes, the Percentage of Correctly Classified locations (PCC) is computed [8] through a leave-one-out procedure [9, 4]. It corresponds to the observed average map purity (AMPo) as presented by Bierkens and Burrough (1993) [1]. Two additional indicators are calculated along with the PCC : the percentage of misclassification from a drainage class to a close one (i.e. the percentage of misclassification involving the second drainage class c_2) and the percentage of misclassification from an extreme drainage class to the other (i.e. the percentage of misclassification between c_1 and c_3).

The Average Highest Probability (AHP) evaluates the ability of the methods to discriminate unambiguously one category from the others and thus predict soil drainage classes with maximum probability. These comparison criteria are presented in table 2.

Table 2: Global comparison criteria for BME, BDF, logistic regression and ICK methods

	% correctly classified	% misclassified close class	% misclassified extreme class	Average Highest Probability
BME	0.72	0.27	0.01	0.76
BDF	0.72	0.27	0.01	0.76
Log. Reg.	0.59	0.38	0.03	0.60
ICK	0.58	0.37	0.05	0.79

It is worth noting that ICK gives positive probability values only to categories that are observed in the neighbourhood and that ICK can lead to inconsistencies (as giving a probability greater than one). For our study, more than 7 % of the maximum probabilities are greater than 1. Due to these inconsistencies, the AHP for ICK will not be taken into

account in the discussion. Note also that more than 43 % of the minimum probabilities calculated with ICK approach are negative.

Calculated indicators (Table 2) indicate similar accuracies and uncertainties for BME and BDF approaches and they confirm superiority of both methods over logistic regression and ICK.

3.2.2 Local uncertainty

In order to visualize how BME and BDF perform at a local scale, the ϕ criterion is computed for both approaches.

$$\phi = 1 - \hat{P}(C_{i,max}) \quad (10)$$

where ϕ is equal to zero when the most probable category $C_{i,max}$ has a probability equal to one (minimum uncertainty) [10].

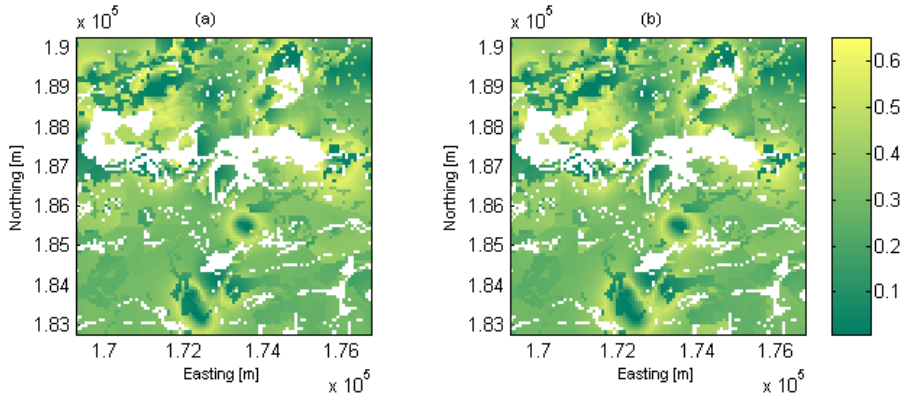


Figure 3: Maps of the ϕ criterion, for (a) the BME approach and (b) the BDF approach.

From figure 3, it appears that local uncertainty follows the same pattern with the BME and the BDF methodologies, with high uncertainty in transition zone and low uncertainty around hard data. These maps confirm the fact that using BDF instead of BME does not result in a significant decrease of the estimated probability associated with the most probable category.

4 Discussion and conclusions

BME has become a complete framework in the context of space-time prediction since it can be applied to predict continuous variables, categorical variables and mixed random fields. In this paper, BDF framework has been extended to categorical variable. By the

light of the results, BDF for categorical variable seems to be a good compromise as the prediction remains accurate with lower computation time than BME. This method has proven to be nearly as accurate as BME approach for updating the old drainage class map with recently collected samples and shows similar uncertainties. Processing is faster due to the convenient conditional independence hypothesis which alleviates the need of inferring for the joint probability density function in equation (3) [2]. BDF is thus easier to implement and can account for a larger number of information sources at a reduced computational burden.

The superiority of the BME and BDF methods over ICK and logistic regression is clear by the light of the final results. Indicator kriging techniques suffer from well-known inconsistencies (see e.g. Cressie (1991) [6], Goovaerts (1997) [10], Christakos (2000) [5]), while the logistic regression approach does not take into account the spatial structure of the data and thus lead to a significant loss of information.

As a summary, although BDF methodology for categorical variables is somehow a simplification of BME approach (conditional independence hypothesis), both methods lead to very similar results and benefit from strong advantages over ICK and logistic regression. With respect to computational burden, BDF has clearly the edge over BME. This simple application in the prediction of drainage classes makes BDF a promising method for other applications where accounting for several secondary information sources is of primary interest.

References

- [1] M. F. Bierkens and P. A. Burrough, (1993). *The indicator approach to categorical soil data. 2. application to mapping and land-use suitability*. Journal of Soil Science, 44(2):369–381.
- [2] P. Bogaert and D. Fasbender, (2007). *Bayesian data fusion in a spatial prediction context: a general formulation*. Sto, 21(6):695–709.
- [3] A. K. Bregt, J. J. Stoorvogel, J. Bouma, and A. Stein, (1992). *Mapping ordinal data in soil survey: A costa rican example*. Soil Science Society of America Journal, 56(2):525 – 531.
- [4] J. P. Chiles and P. Delfiner. *Geostatistics : Modeling Spatial Uncertainty*. John Wiley, New York, (1999).
- [5] G. Christakos. *Modern Spatiotemporal Geostatistics*. Oxford University Press, New York, (2000).
- [6] N. Cressie. *Statistics for spatial data*. Wiley series in probability and mathematical statistics, J. Wiley and Sons, New York, (1991).

- [7] D. D'Or and P. Bogaert, (2004). *Combining categorical information with the bayesian maximum entropy approach*. geoENV IV - Geostatistics for Environmental Applications, pages 295–306.
- [8] D. D'Or and P. Bogaert, (2004). *Spatial prediction of categorical variables with the bayesian maximum entropy approach: the ooypolder case study*. European Journal of Soil Sciences, 55(4):763–775.
- [9] D. Fasbender, L. Peeters, P. Bogaert, and A. Dassargues, (2008). *Bayesian data fusion applied to water table spatial mapping*. Water Resources Research, 44(12):w12422.
- [10] P. Goovaerts. *Geostatistics for Natural Ressources Evaluation*. Oxford University Press, New York, (1997).
- [11] A. G. Journel, (1986). *Constrained interpolation and qualitative information - the soft kriging approach*. Mathematical Geology, 18:269–286.
- [12] A. Kravchenko, G. Bollero, R. Omonode, and D. Bullock, (2002). *Quantitative mapping of soil drainage classes using topographical data and soil electrical conductivity*. Soil Science Society of America Journal, 66:235–243.
- [13] J. Liu, E. Pattey, M. C. Nolin, J. R. Miller, and O. Ka, (2008). *Mapping within-field soil drainage using remote sensing, dem and apparent soil electrical conductivity*. Geoderma, 143(3-4):261–272.
- [14] J. V. Orshoven, J. Maes, H. Vereecken, J. Feyen, and R. Dudal, (1988). *A structured database of belgian soil profile data*. Pedologie, 38:191–206.
- [15] M. A. Wibrin, P. Bogaert, and D. Fasbender, (2006). *Combining categorical and continuous spatial information within the bayesian maximum entropy paradigm*. Sto, 20(6):381–467.