An efficient maximum entropy approach for categorical variable prediction

D. Allard[a], D. D'Or[b], R. Froidevaux[b]

[a] Biostatistique et Processus Spatiaux (BioSP), INRA, Site Agroparc, 84914 Avignon, France.

[b] Ephesia Consult, 9, rue Boissonnas, CH - 1227 Geneva, Switzerland.

Running title: Efficient maximum entropy for categorical variable prediction

Corresponding author: D. Allard, allard@avignon.inra.fr

Summary

We address the problem of the prediction of a spatial categorical variable by revisiting the maximum entropy approach. We first argue that, for predicting category probabilities, a maximum entropy approach is more natural than a least-squares approach, such as (co-) kriging of indicator functions. We then show that, knowing the categories observed at surrounding locations, the conditional probability of observing a category at a location obtained with a particular maximum entropy principle is a simple combination of sums and products of univariate and bivariate probabilities. This prediction equation can be used for categorical estimation or categorical simulation. We make connections to earlier work on prediction of categorical variables. On simulated data sets we show that our equation is a very good approximation to Bayesian Maximum Entropy (BME), while being orders of magnitude faster to compute. Our approach is then illustrated by using the celebrated Swiss Jura data set.

Introduction

We consider the problem of the spatial prediction of a categorical variable given a set of observations at surrounding locations. Such a problem is common in all branches of geosciences and particularly in soil science. In the following, prediction means either the computation of the probability of occurrence of a category at a given location, or the mapping of the most likely category (estimation), or the conditional simulation of categories (simulation). In all cases, what is needed is the conditional probability of observing a category at an unsampled location, given the observed categories at sampled locations.

An extensive body of literature has been pulished over the last two decades proposing alternative solutions to this problem. Indicator kriging (Journel, 1983) and sequential indicator simulation (Journel & Alabert, 1989) use an independent prediction for each category. This involves fitting a variogram model for each category and then estimating independently, via kriging, the probability of occurrence of each class. It has been applied in the field of soil science (water table classes) by Bierkens & Burrough (1993a,b).

Although easy to implement, this method has well known inherent problems: the probabilities of occurrence are not guaranteed to be between 0 and 1 and the sum of probabilities may not be equal to 1. A post-processing of the conditional probabilities is thus needed. From an application point of view, well-defined soil or lithological structures and asymmetries in depositional sequences are almost impossible to reproduce with sequential indicator simulation. Simulations based on transition probabilities (Carle & Fogg, 1996) amounts to a sequential co-indicator simulation, not based on a (co-) variogram model, but on transition probabilities. Although this method fares better in terms of the reproduction of lithological structures and asymmetries, some of the problems found in indicator simulation are still present.

Sequential simulation using multi-points statistics (Strebelle, 2002; Mariethoz & Renard, 2010) allows us to go beyond the two-points statistical models inherent to all variogram-based methods and provides very convincing images. The major limitation of this method is that it requires a training image as a model and such training images are very difficult to obtain for 3D applications. With the exception of multi-points statistics, all the above techniques are based on kriging which is an optimal generalized mean-squared linear predictor, i.e. a linear predictor minimizing the variance of estimation.

The Bayesian Maximum Entropy (BME) principle (Christakos, 1990) has been applied to the spatial prediction of categorical variables by Bogaert (2002), in particular for soil data sets (D'Or *et al.*, 2001; Bogaert & D'Or, 2002; & D'Or & Bogaert, 2004). The aim was to find an approximate probability distribution, denoted $\tilde{p}$, that maximizes the entropy, while respecting known univariate and bivariate probability functions. The general form of $\tilde{p}$ is a log-linear model involving main effects and interaction effects of the first order, whose parameters are estimated at each prediction point using the Iterative Proportional Fitting algorithm (Deming & Stephan, 1940). Because of CPU time and memory storage considerations, solving this estimation problem is unfortunately limited to a restricted number of categories and small numbers of neighbouring samples.

More recently a Markov Chain Random Field approach was proposed by Li (2007) with applications to the simulation of soil type spatial distribution (Li *et al.*, 2004; Li & Zhang, 2007) but the conditioning is limited to only four neighbours organized in orthogonal directions. This algorithm relies on a conditional independence between two Markov chains, but its optimality properties are not known.

To summarize briefly, the spatial prediction problem is that of selecting among all possible sets of probability distributions the one that, subject to some constraints,

minimizes some functionals such as mean squared distance (kriging paradigm) or negative entropy (maximum entropy principle). Leaving aside implementation issues that may or may not limit a particular approach, one may wonder at a more theoretical level which selection rule is 'best', and in what sense. In a very interesting paper, Csiszár (1991) addressed this problem from an axiomatic point of view. In a relatively general setting, he showed that when there are no restrictions on the values to be predicted, the only selection rule meeting some minimal requirements, such as product-consistency, distinctness, continuity and locality, is the least-square selection rule. In constrast, when the predicted values must belong to a restricted set, or when they are constrained to sum up to one (such as a probability distribution for categories), the maximum entropy selection rule is the unique regular, local, product-consistent selection rule. Although this theorem is not completely general, it provides strong evidences that the maximum entropy selection rule should be prefered for spatial prediction of categories.

There is thus a need for a prediction method which i) is based on the maximum entropy principle; ii) is simple to use, fast to compute and accurate and iii) can reproduce complex patterns, including asymmetries and depositional sequences such as catenary patterns. The maximum entropy principle is thus revisited and we establish a new result: we show that the maximum entropy solution, subject to the bivariate probabilities having the prediction point as one end-point, has a closed-form expression. The resulting conditional probability only involves sums and products of univariate and bivariate probabilities. This formulation corresponds essentially to a conditional independence assumption; it will therefore be named MCP, for Markovian-type Categorical Prediction.

MCP is easy to code and fast to compute. As a consequence, it can handle a large number of categories and as many neighbours in as many directions as desired. It

results in a genuine probability distribution which, in addition, has a very interesting 0/1 forcing property: if any bivariate probability with a neighbour is equal to 0 (or equal to 1), the resulting conditional probability will be strictly equal to 0 (or, respectively, equal to 1). In consequence, sequences of categories are easily reproduced, not only in simulations, but also in prediction when mapping the most likely category.

The outline of the paper is as follows. After a first intuitive derivation of the MCP equation, we show that it is the solution of a maximum entropy approach subject to a restricted set of constraints. We also show that the Markov Chain Random Field (Li, 2007) is a special case of our general result and we also provide some interesting connections with composite likelihoods (Nott & Ryden, 1999) used in spatial statistics. We then estimate the set of bivariate probabilities from the data and demonstrate on simulated data and on a realistic case study the performances of MCP and compare it to BME. The last section is devoted to discussions and conclusions.

Framework, notation and intial description of MCP

General framework and notations

Consider a categorical random field $Y(\mathbf{x})$ defined on a domain D of the d-dimensional Euclidean space $\mathbb{R}^d$. Here, $d = 2$, but our model could be applied to higher-dimensional spaces. We will denote $\mathbf{Y} = (Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_n))$ a n vector of $Y(\mathbf{x})$ and $\mathcal{Y}$ the set of all possible combinations of categories for $\mathbf{Y}$. In the following, P will denote the probability distribution associated to the random vector $\mathbf{Y}$. Each random variable $Y(\mathbf{x}_k)$ is a multilevel categorical variable taking values in $\mathcal{I} = \{i = 1, \ldots, I\}$. We will denote $p_i(\mathbf{x}_k) = P[Y(\mathbf{x}_k) = i]$ the univariate probability of observing category i at location $\mathbf{x}_k$. Without loss of generality, we will consider that for each category i, there exists at least one location k such that $p_i(\mathbf{x}_k) > 0$, so that no category is empty.

The spatial structure of the categorical random field $\mathbf{Y}(\mathbf{x})$ is described by means of the bivariate probabilities:

$$p_{i,j}(\mathbf{x}_k, \mathbf{x}_{k'}) = P[Y(\mathbf{x}_k) = i \text{ and } Y(\mathbf{x}_{k'}) = j], \tag{1}$$

of occurrence of categories i and j respectively at locations $\mathbf{x}_k$ and $\mathbf{x}_{k'}$. These functions called bivariate probability functions (Bogaert, 2002), fully characterize the bivariate distribution of the categorical random field. They are equal to the non-centered cross-covariance functions of $(\mathbf{1}_1(\mathbf{x}), \ldots, \mathbf{1}_I(\mathbf{x}))$, where $\mathbf{1}_i(\mathbf{x})$, $i = 1, \ldots, I$ denotes the indicator function of category i: $\mathbf{1}_i(\mathbf{x}) = 1$ if $Y(\mathbf{x}) = i$ and $\mathbf{1}_i(\mathbf{x}) = 0$ otherwise. From now on, we will make the following stationarity assumption:

Assumption 1 (Stationarity). The univariate and bivariate probability functions of the categorical random field $Y(\mathbf{x})$ are translation-invariant:

$$p_i(\mathbf{x}_k) = p_i; \quad p_{i,j}(\mathbf{x}_k, \mathbf{x}_{k'}) = p_{i,j}(\mathbf{h}_{kk'}), \quad \text{for all } \; i, j \in I, \; k, k' \in 1, \ldots, n, \tag{2}$$

where $\mathbf{h}_{kk'} = (\mathbf{x}_{k'} - \mathbf{x}_k)$.

The univariate and bivariate probability functions, being necessarily bounded, translation invariance of the bivariate probability functions is equivalent to the usual second-order stationarity in geostatistics (Chilès & Delfiner, 1999). By definition, the following relationships always hold: $\sum_{i=1}^{I} p_i = 1$; $\sum_{j=1}^{I} p_{i,j}(\mathbf{h}) = p_i$; $p_{i,j}(\mathbf{0}) = p_i$ for $i = j$ and $p_{i,j}(\mathbf{0}) = 0$ otherwise. However, in general $p_{i,j}(\mathbf{h}) \neq p_{j,i}(\mathbf{h})$ and $p_{i,j}(\mathbf{h}) \neq p_{i,j}(-\mathbf{h})$, but bivariate probabilities are antisymmetric :$p_{i,j}(\mathbf{h})$ is always equal to $p_{j,i}(-\mathbf{h})$.

A related function which will be used in this work is the conditional bivariate probability function, sometimes called transiogram:

$$P\left[Y(\mathbf{x} + \mathbf{h}) = j \mid Y(\mathbf{x}) = i\right] = p_{j|i}(\mathbf{h}) = p_{i,j}(\mathbf{h})/p_i. \tag{3}$$

Initially, we assume that bivariate probability functions are available. We will show later how they can be estimated from data.

According to some criteria which will be developed below, we want to find within this framework a 'good' approximation of the conditional probability for category i to occur at a given location $\mathbf{x}_0$, given the categories $i_1, \ldots, i_n$ observed at surrounding data locations $\mathbf{x}_1, \ldots, \mathbf{x}_n$. This probability, denoted $p_{i_0|i_1,\ldots,i_n}$ is:

$$p_{i_0|i_1,\ldots,i_n} = \frac{p_{i_0,i_1,\ldots,i_n}}{p_{i_1,\ldots,i_n}}. \tag{4}$$

A first, intuitive, derivation of MCP

We seek to approximate the conditional probability in Equation (4) using only univariate and bivariate probabilities. First, Equation (4) can be rewritten in the following way:

$$p_{i_0|i_1,\ldots,i_n} = \frac{p_{i_0,i_1,\ldots,i_n}}{p_{i_1,\ldots,i_n}} = \frac{p_{i_0,i_1,\ldots,i_n}}{\sum_{i_0=1}^{I} p_{i_0,i_1,\ldots,i_n}} = \frac{p_{i_0} p_{i_1,\ldots,i_n|i_0}}{\sum_{i_0=1}^{I} p_{i_0} p_{i_1,\ldots,i_n|i_0}}. \tag{5}$$

The conditional probability $p_{i_1,\ldots,i_n|i_0}$ is then approximated using products of bivariate conditional probabilities:

$$p^*_{i_1,\ldots,i_n|i_0} = \prod_{k=1}^{n} p_{i_k|i_0}(\mathbf{h}_{0k}). \tag{6}$$

This approximation corresponds to the conditional independence assumption that the categories are independent at $\mathbf{x}_k$ and $\mathbf{x}_{k'}$, given the category at the prediction location. The approximate conditional probability of category i at the site $\mathbf{x}_0$ then becomes

$$p^*_{i_0|i_1,\ldots,i_n} = \frac{p_{i_0} \prod_{k=1}^{n} p_{i_k|i_0}(\mathbf{h}_{0k})}{\sum_{i_0=1}^{I} p_{i_0} \prod_{k=1}^{n} p_{i_k|i_0}(\mathbf{h}_{0k})} = \frac{p_{i_0}^{1-n} \prod_{k=1}^{n} p_{i_0,i_k}(\mathbf{h}_{0k})}{\sum_{i_0=1}^{I} p_{i_0}^{1-n} \prod_{k=1}^{n} p_{i_0,i_k}(\mathbf{h}_{0k})}, \tag{7}$$

since $p_{i_k|i_0}(\mathbf{h}_{0k}) = p_{i_0,i_k}(\mathbf{h}_{0k})/p_{i_0}$.

Equation (7) is the basis of our spatial prediction method for categories. Since it is based on a conditional independence assumption, it is called Markovian-type Categorical Prediction (MCP). In the computer science, machine learning and classification

literature, Equation (7) is known as the "Naïve Bayes classifier" and Equation (6) is known as the "Naïve Bayes assumption", which performs surprisingly well in most applications (Domingos & Pazzani, 1997).

Mapping the most probable category according to Equation (7) yields a prediction map, while drawing from it by Monte-Carlo in a sequential simulation context provides simulated categories conditional on existing data. Next, we will show that Equation (7) is the solution of a maximum entropy principle, but we first make a few preliminary remarks.

(i) In the case of spatial independence, the bivariate probabilities satisfy $p_{ij}(\mathbf{h}) = p_i p_j$, or equivalently, $p_{i|j}(\mathbf{h}) = p_i$ for all $\mathbf{h}$. In this case, Equation (7) simply becomes

$$p^*_{i_0|i_1,\ldots,i_n} = \frac{p_{i_0} \prod_{k=1}^{n} p_{i_k}}{\sum_{i_0=1}^{I} p_{i_0} \prod_{k=1}^{n} p_{i_k}} = p_{i_0}.$$

(ii) It is important to note that $p_{i_0|i_1,\ldots,i_n}$ is not approximated by $\prod_k p_{i_0|i_k}(\mathbf{h}_{0k})$. If this were so it would lead, in the case of spatial independence, to the inconsistent result $p^*_{i_0|i_1,\ldots,i_n} = p^n_{i_0}$.

(iii) Being the result of additions and products of probabilities, the probabilities $p^*_{i_0|i_1,\ldots,i_n}$ are positive. Moreover, they add to 1 thus providing a genuine probability distribution.

(iv) If for one datum, the transiogram $p_{i_k|i_0}(\mathbf{h}_{0k})$ is equal to 0 (or equal to 1), the conditional probability equation (6) will be equal to 0 (or, conversely, equal to 1). Forbidden transitions $p_{i_k|i_0}(\mathbf{h}_{0k}) = 0$ or mandatory transitions $p_{i_k|i_0}(\mathbf{h}_{0k}) = 1$ will thus be respected, both in estimation and in simulation. This 0/1 forcing property is very interesting in a soil science context in which sequences of soil types such as catenary patterns must be respected.

(v) The computation of Equation (7) is very fast. It does not involve matrix inversion or a large contingency table. We can thus consider many categories and include a larger number of data locations in the neighbourhood. Moreover, we will show in the case studies that it converges rapidly as the number of neighbours increases, and therefore is also robust.

(vi) No assumptions about the isotropy or the symmetry of the bivariate probabilities are necessary; the approach is thus able to deal with highly asymmetric/anistropic patterns, such as sequential orderings or catenary patterns as will be illustrated below.

Theoretical results

Maximum entropy principle

We first recall briefly the definition of entropy and some related concepts. A complete presentation of entropy is available in Cover & Thomas (2006). The entropy of a finite probability distribution, say P, with states $\mathbf{Y} \in \mathcal{Y}$ is the quantity:

$$H(P) = -\sum_{\mathbf{Y}\in\mathcal{Y}} P(\mathbf{Y}) \ln P(\mathbf{Y}) = -E_P[\ln P(\mathbf{Y})]. \tag{8}$$

A related concept is the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951), or relative entropy, between a distribution, say Q, and P, denoted $D(P \mid\mid Q)$:

$$D(P \mid\mid Q) = \sum_{\mathbf{Y}\in\mathcal{Y}} P(\mathbf{Y}) \ln \frac{P(\mathbf{Y})}{Q(\mathbf{Y})} = E_P\left[\ln \frac{P(\mathbf{Y})}{Q(\mathbf{Y})}\right] = -H(P) - E_P[\ln Q(\mathbf{Y})], \tag{9}$$

where, by convention, $0 \ln 0/p = 0$ and $p \ln p/0 = \infty$, for any $p > 0$. Although not a distance in the mathematical sense (it is not symmetrical), the KL divergence is a measure of how 'different' two probability distributions are. It is always positive and it is equal to zero if, and only if, $P = Q$. There are strong connections between entropy and KL divergence (see Cover & Thomas, 2006). In particular, let us assume

that some quantities related to P are known, such as moments or conditional probabilities. A natural approach, very common in information theory, image processing, language processing and computer science, is to seek the distribution P^* that shares properties (moments or conditional probabilities) with P and minimizes the KL divergence $D(P^* \,||\, P)$. This is equivalent to finding the distribution P^* and maximizing its entropy $H(P^*) = \sum_{\mathbf{Y}\in\mathcal{Y}} P^*(\mathbf{Y}) \ln P^*(\mathbf{Y})$, subject to the imposed constraints.

Bayesian Maximum entropy (BME)

BME for categorical variables has been proposed by Bogaert (2002) and D'Or & Bogaert (2004). It is a two stage procedure. In the first stage, the joint probability distribution $\tilde{p}_{i_0,i_1,\ldots,i_n}$, solution of the maximum entropy principle, subject to the univariate and bivariate constraints:

$$\begin{aligned}
\sum_{i_0=1}^{I} \cdots \sum_{i_n=1}^{I} \tilde{p}_{i_0,\ldots,i_n} \mathbf{1}_{[Y_k=i]} &= p_i, \\
\sum_{i_0=1}^{I} \cdots \sum_{i_n=1}^{I} \tilde{p}_{i_0,\ldots,i_n} \mathbf{1}_{[Y_k=i;Y_{k'}=j]} &= p_{i,j}(||\mathbf{h}_{kk'}||),
\end{aligned}$$

for all $k, k' = 0, \ldots, n$ and $i, j = 1, \ldots, I$ is found. In the second stage, the conditional probability of category i_0 at $\mathbf{x}_0$ is computed as follows

$$\tilde{p}_{i_0|i_1,\ldots,i_n} = \frac{\tilde{p}_{i_0,i_1,\ldots,i_n}}{\tilde{p}_{i_1,\ldots,i_n}}.$$

Following a derivation similar to the one presented in the Appendix 1, it is possible to show that the maximum entropy joint probability distribution is of the form:

$$\tilde{p}_{i_0,\ldots,i_n} = \mu \prod_{l=0}^{n} \lambda_{i_l} \prod_{k=0}^{n-1} \prod_{k'=k+1}^{n} \nu_{i_k,i_{k'}}, \tag{10}$$

where μ, $\boldsymbol{\lambda} = (\lambda_{i_0}, \lambda_{i_1}, \ldots, \lambda_{i_n})$ and $\boldsymbol{\nu} = (\nu i_0, \nu i_1, \ldots, \nu i_n)$ are computed so as to match the given univariate and bivariate probabilities. We will call $\tilde{p}_{i_0|i_1,\ldots,i_n}$ the full-BME prediction.

For finding the values of the parameters μ, $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$, Bogaert (2002) and D'Or & Bogaert (2004) have used the Iterative Proportional Fitting algorithm (Deming & Stephan, 1940). However, since the size of the table of $\tilde{p}$ is equal to $I^{(n+1)}$, IPF suffers from two limitations: first, as it is an iterative procedure, computation time becomes very large, and second, the memory requirements to save the solution may exceed the memory of conventional computers. As a consequence, the categories and the sizes of the neighbourhood are restricted to a very limited number, which may be too strong a restriction for this approach to be used easily in practice.

A restricted BME leads to MCP

Since BME is difficult to apply because $\mu, \boldsymbol{\lambda}$ and $\boldsymbol{\nu}$ are computed so as to match the whole set of $n(n+1)$ bivariate probabilities, we restrict the problem and only consider constraints on the n bivariate probabilities between the prediction point and the data points p_{i_0,i_k}, $k = 1, \ldots, n$. We obtain the following proposition, which is the theoretical support of MCP.

Proposition The probability distribution p^*, solution of the maximum entropy principle, subject to the following univariate and bivariate constraints:

$$\begin{aligned} \sum_{i_0=1}^{I} \cdots \sum_{i_n=1}^{I} p^*_{i_0,i_1,\ldots,i_n} \mathbf{1}_{[Y_0=i]} &= p_i, \\ \sum_{i_0=1}^{I} \cdots \sum_{i_n=1}^{I} p^*_{i_0,i_1,\ldots,i_n} \mathbf{1}_{[Y_0=i;Y_k=j]} &= p_{i,j}(||\mathbf{h}_{0k}||), \end{aligned}$$

for all $k = 1, \ldots, n$ and $i, j = 1, \ldots, I$, is:

$$p^*_{i_0,i_1,\ldots,i_n} = p_{i_0}^{1-n} \prod_{k=1}^{n} p_{i_0,i_k}(\mathbf{h}_{0k}). \tag{11}$$

The proof of this proposition is given in the Appendix. Unlike a full BME, there is now a closed-form expression for p^*, allowing for fast computation. From Equation

(11), one can easily derive the conditional probability of $Y(\mathbf{x}_0)$:

$$p^*_{i_0|i_1,\ldots,i_n} = \frac{p^*_{i_0,i_1,\ldots,i_n}}{p^*_{i_1,\ldots,i_n}} = \frac{p_{i_0}^{1-n}\prod_{k=1}^{n} p_{i_0,i_k}(\mathbf{h}_{0k})}{\sum_{i_0=1}^{I} p_{i_0}^{1-n}\prod_{k=1}^{n} p_{i_0,i_k}(\mathbf{h}_{0k})}, \tag{12}$$

which is precisely our prediction Equation (7).

Returning to the full BME, for prediction purposes (estimation and/or simulation), the conditional probability $p_{i_0|i_1,\ldots,i_n}$ is our final target and the joint probability $p_{i_0,i_1,\ldots,i_n}$ is not of direct interest. Inserting the full BME solution Equation (10) into the numerator and denominator of the conditional probability Equation (4) yields:

$$\tilde{p}_{i_0|i_1,\ldots,i_n} = \frac{\mu\prod_{l=0}^{n}\lambda_{i_l}\prod_{k=0}^{n-1}\prod_{k'=k+1}^{n}\nu_{i_k,i_{k'}}}{\sum_{i_0=1}^{I}\mu\prod_{l=0}^{n}\lambda_{i_l}\prod_{k=0}^{n-1}\prod_{k'=k+1}^{n}\nu_{i_k,i_{k'}}} = \frac{\lambda_{i_0}\prod_{k=1}^{n}\nu_{i_0,i_k}}{\sum_{i_0=1}^{I}\lambda_{i_0}\prod_{k=1}^{n}\nu_{i_0,i_k}}. \tag{13}$$

This equation has the same formal structure as Equation (7). However, for the full BME it is not generally true that one can identify λ_{i_0} with p_{i_0} and ν_{i_0,i_k} with $p_{i_k|i_0}(\mathbf{h}_{0k})$. The conditional distribution obtained with the restricted set of constraints (MCP) is thus not equal to the conditional distribution obtained with the full set of constraints (full BME), even though the formal structure of the equations are similar.

MCP, (Equation (7)) is thus an approximation to the full BME solution as in D'Or & Bogaert (2004). It is however fast, easy to code and in most cases a very accurate approximation as it will be shown later. Differences between MCP and BME on the local conditional distributions will be illustrated in some particular cases in a subsequent section.

Relationship with Composite Likelihood

Elaborating on the pseudo-likeliood introduced by Besag (1974), Hjort & Omre (1994) proposed the use of products of marginal and/or conditional likelihoods, an approach called composite likelihood, for covariance estimation in geostatistics. Nott & Ryden (1999), and Heagerty & Lele (1998) have used pairwise likelihood in the

context of spatial probit regression and for inference on image models. Under usual regularity conditions, they proved that pairwise likelihood and composite likelihood estimators of the parameters converge to the true value as the number of data increases, a property called statistical consistency. They also obtained asymptotic normality of the estimators.

Our approximation Equation (7), is a type of pairwise composite likelihood, since it is a product of marginal univariate and bivariate probabilities. Clearly, the purpose here is not parameter estimation, but prediction at an unsampled location. However, the properties obtained in the context of parameter estimation provides an interesting insight into the optimality properties shared by maximum likelihood estimators and maximum entropy principles for the exponential family of distributions.

Relationship with Markov Chain Random Fields

For the estimation and simulation of spatial categorical variables, Li (2007), Li *et al.* (2004) and Li & Zhang (2007) proposed combining two orthogonal one-dimensional Markov chains, P^1 and P^2, with a conditional independence assumption between the two chains. Let us consider $\mathbf{x}_0$ as a site of the two-dimensional grid and $\mathbf{x}_1, \ldots, \mathbf{x}_4$ the four nearest neighbours in the four directions $d_1, \ldots, d_4$ with distances to $\mathbf{x}_0$ respectively equal to $h_1, \ldots, h_4$. Then, the prediction equation obtained by Li (2007; Equation (12)) is:

$$P^{LZ}_{i_0|i_1,\ldots,i_4} = \frac{P^{d_1}_{i_0|i_1}(h_1) \prod_{k=2}^{4} P^{d_k}_{i_k|i_0}(h_k)}{\sum_{i_0=1}^{I} P^{d_1}_{i_0|i_1}(h_1) \prod_{k=2}^{4} P^{d_k}_{i_k|i_0}(h_k)}, \tag{14}$$

where $P^{d_i}(\cdot)$ is obtained from transition probabilities of one of the two Markov chains, depending upon their direction. Replacing $P^{d}_{i_0|i_1}(h_1)$ by $P^{d}_{i_1|i_0}(h_1)p_{i_0}/p_{i_1}$ yields, after simplification by p_{i_1}, to an equation formally similar to the prediction Equation (7).

Our method is thus a generalization of the Markov Chain Random Field approach, allowing to consider as many neighbours in as many directions as desired.

Method

Estimation of the bivariate probability functions

Up to now, we have considered that the bivariate probability function was known. It has of course to be estimated from the data. Two very different cases must be distinguished. In geological or petroleum applications data are usually sparse and organized along lines (wells, drill cores, etc.), which does not allow the estimation of the bivariate probability functions in all directions. Geologists thus rely very often on analogues and training images. In this case, the bivariate probabilities can be estimated directly from the image for all vectors $\mathbf{h}$.

In soil science however, the estimation of the bivariate probability functions is based on some sparse and/or irregular spatial sample $\mathbf{Y}$. In general, we wish to estimate a non-isotropic, antisymmetric $I \times I$ matrix of functions $p_{ij}(\mathbf{h})$ to the data. The estimated bivariate probability functions must verify the compatibility conditions: (i) $\hat{p}_{ii}(\mathbf{0}) = p_i$ and (ii) $\sum_{j=1}^{I} \hat{p}_{ij}(\mathbf{h}) = p_i$ for all $\mathbf{h}$, where p_i is the univariate marginal proportion of category i.

Because of the potentially complex behaviour of the bivariate functions, we implement a non-parametric approach using a kernel smoothing procedure already used by D'Or & Bogaert (2004). It is briefly summarised here; more details can be found in Silverman (1986). Let us consider a fixed direction, say $\mathbf{h}$ and a positive distance, h, along $\mathbf{h}$. Let us further denote $\mathcal{P}_{\mathbf{h}}$ the set of all pairs of sample locations $\{\mathbf{x}_k, \mathbf{x}_{k'}\}$ such that the direction of the vector $\mathbf{x}_{k'} - \mathbf{x}_k$ is equal to the direction $\mathbf{h}$, up to a specified angular tolerance. Then, denoting Y_k for $Y(\mathbf{x}_k)$, the kernel estimator of $p_{ij}(\mathbf{h})$ is:

$$p_{ij}^{K}(h) = \frac{\sum_{k,k' \in \mathcal{P}_{\mathbf{h}}} \mathbf{1}_{[Y_k=i, Y_{k'}=j]} K_w(h - ||\mathbf{x}_{k'} - \mathbf{x}_k||)}{\sum_{k,k' \in \mathcal{P}_{\mathbf{h}}} K_w(h - ||\mathbf{x}_{k'} - \mathbf{x}_k||)}, \tag{15}$$

where $K_w(h) = 1/wK(h/w)$ and $K(h)$ is a kernel function, i.e. it is an even function such that $\int_{\mathbb{R}^d} K(||\mathbf{h}||)d\mathbf{h} = 1$. Here, we have used a 1-dimensional gaussian kernel

$K(h) = e^{-||\mathbf{h}||^2}/\sqrt{2\pi}$, but other kernel functions could be used. The parameter w is a bandwidth parameter. The greater the value of w, the smoother the function $p_{ij}^K(h)$. It can be chosen by the user or set automatically so as to minimize the mean-squared estimation error of $p_{ij}^K(h)$ (Silverman, 1986).

It is easy to check that Equation (15) ensures that $\sum_{i=1}^{I}\sum_{j=1}^{I} p_{ij}^K(h) = 1$. The compatibility conditions are, however, not automatically verified, and some additional computations are necessary. In order to verify condition (i), the estimation of $p_{ij}^K(\mathbf{h})$ is extended to negative values of h according to a central symmetry around $(0, p_{ij})$, where $p_{ii} = p_i$ and $p_{ij} = 0$ when $i \neq j$. Specifically, one considers the values $\{2p_i - \mathbf{1}_{[Y_k=i,Y_{k'}=j]}\}$ for the negative distances $-||\mathbf{x}_{k'} - \mathbf{x}_k||$. Because of the central symmetry, the estimated value $p_{ii}^K(h=0)$ will automatically be equal to p_i, while $p_{ij}^K(h=0) = 0$ when $i \neq j$. The result of this negative-extended kernel estimation is denoted $p_{ij}^{(0)}(h)$. Compatibility condition (ii) is then obtained with an iterative computation at each distance h. Starting with $p_{ij}^{(0)}(h)$, we iterate:

$$p_{ij}^{(r)}(h) = p_{ij}^{(r-1)}(h)\frac{p_i p_j}{\sum_{i=1}^{I} p_{ij}^{(r-1)}(h)\sum_{j=1}^{I} p_{ij}^{(r-1)}(h)}. \tag{16}$$

As $r \to \infty$, this computation will eventually converge to the estimated bivariate probability function $\hat{p}_{ij}(h)$ which will verify all compatibility conditions. In all tested situations, convergence was reached in less than 10 iterations.

The MCP approach is now illustrated by using two case studies, a synthetic one and a real one. The synthetic case aims at demonstrating the power of the approach for generating simulations showing spatial structures organized in sequence. In the real case, we demonstrate the ability of MCP for estimating categorical variables.

Synthetic case study

We first illustrate the method on a synthetic case study with three categories (Figure 1a). The reference image was the result of a 3 colour classification of sand dunes in the

Gobi desert. The 114 by 114 pixel image was sampled at 500 locations to produce the reference data set (Figure 1b). The marginal probability of occurrence vector on the reference image was [0.4274 0.2622 0.3104]. The categories may represent geological facies, soil types, land uses or any other categorical variable. In order to mimic catenary patterns, the spatial structure presented a South-West to North-East asymmetry: in this direction, repetitive transition sequence 1-2-3 was the most probable while other transitions were unlikely to appear. Both the omni-directional and the directional bivariate probability models were estimated directly from the reference image for any vector $\mathbf{h}$. The estimation and simulation cases are next developed.

As a first task, the data set was used to estimate the probability of occurrence for each category on the entire grid. Full BME and MCP methods were compared. A maximum of five data were used in the neighbourhood up to a distance of 25 pixels. The advantage of using the directional bivariate probability model instead of the omnidirectional one was also tested. The maps of the most probable categories are shown in Figure 1. Classification was compared with the reference image (Figure 1a). The correct classification rate was equal to 69.07 for BME and 67.83 for MCP using the omnidirectional model, and to 71.02 (BME) and 71.82 (MCP) using the directional model. Differences between BME and MCP in terms of classification rate were marginal.

In addition to a better classification, the impact of using a directional model resulted mostly in improved spatial organization of the categories: expected transitions were far better described. This could be quantified by computing the root mean squared error (RMSE) over all the pairs of categories of the differences in transition probabilities between any combination of method (BME, MCP) and model (omnidirectional or directional) and the reference image for the first step in the North-East direction. The values of these RMSE are given in Table 1. The use of directional models reduced the RMSE by a factor of nearly two. Using the directional model introduced the correct

category between the two others even if it was not represented in the neighbourhood. When a directional model is used, MCP scores slightly better than BME for the correct classification rate and also for describing the transitions. In terms of computation times, MCP performed the prediction in 1.3 minutes and BME in 20.8 minutes using Matlab with an AMD Athlon 64 processor, 1.80 GHz, and 512 Mb of RAM.

Table 1: Root Mean Squared Errors of the differences of transition probabilities at the first step in the North-East direction. Every combination of method and model is compared to the transition probabilities computed from the reference image.

Method	Model	RMSE 5 data	RMSE 2 data
BME	omnidirectional	0.0797	0.0275
MCP	omnidirectional	0.0817	0.0277
BME	directional	0.0489	0.0161
MCP	directional	0.0438	0.0149

Though differences in classification were small, an examination at the local conditional distributions (LCD) revealed interesting differences between MCP and BME in particular circumstances. If the data were regularly distributed around the estimation location, LCDs from MCP and BME were very similar (Figure 2a). However when the data were clustered (Figure 2b), LCDs became quite different. BME uses the complete network of bivariate probabilities between the group of data and estimation location: the factor ν_{i_0,i_k} corresponding to each data point in the right-hand group takes into account the spatial redundancy of this cluster, leading to a probability close to 1/2. MCP on the contrary uses only the relationships between the prediction location and the data, thus leading to a very large probability for the category of the right-hand group. Note, however, that with the most probable category, the two approaches lead to the same prediction. This example illustrates why differences are more visible on simulations than on predictions.

We have seen earlier the relationships between MCP and BME in terms of the maximum entropy approach. We now illustrate this in more detail. As a benchmark of efficiency, $\tilde{p}_{i_0|i_1,\ldots,i_n}$ and $p^*_{i_0|i_1,\ldots,i_n}$ were computed on a set of increasing numbers $n = 1, \ldots, 10$ of data in the neighbourhood. The BME solution with 10 neighbours, $\tilde{p}_{i_0|i_1,\ldots,i_{10}}$, was the best approximation that could be achieved in a reasonable amount of time; other solutions were thus compared to $\tilde{p}_{i_0|i_1,\ldots,i_{10}}$. The Kullback-Leibler (KL) divergence between the conditional probability distributions and $\tilde{p}_{i_0|i_1,\ldots,i_{10}}$ was then computed along with Gini index, correct classification rates and computation times, at 2000 random locations in order to limit the computation burden of the comparison. The averages of these quantities are shown in Figure 3. As expected, the BME KL divergence decreases asymptotically towards 0 as the number of data increases. More suprisingly, the MCP KL divergences decreases up to 3 data points, before increasing again. Looking in detail at the LCDs shows that the LCD comes close to $\tilde{p}_{i_0|i_1,\ldots,i_{10}}$ with very few data (2 or 3). As more data of the same category are involved, the LCD becomes more clear-cut in favour of this category (as shown in Figure 3b with the Gini index), but becomes also more different from $\tilde{p}_{i_0|i_1,\ldots,i_{10}}$. As a consequence, the KL divergence increases. The correct classification rate was very similar for both approaches, showing that very often a correct classification did not need very accurate probabilities, as long as the maximum probability category remained identical. The last plot shows that in terms of computation time MCP out-performed BME dramatically. BME is plagued with an exponential increase of computation time when the number of neighbours increases while MCP computation increases linearly with a very small slope. Finally, the transition model in the North-East direction is clearly better respected with 2 data than with 5, for MCP as well as for BME (Table 1). It appears from this experiment that MCP seems to perform better with a few data taken in account (max. 5). In this situation, KL divergence is close to its minimum, correct classification rate

is near its maximum and transitions are better reproduced.

As a second task, 50 non-conditional simulations were generated on the 114×114-node grid by both MCP and BME with the directional bivariate probability model. The same random path and seeds were used. 50 pairwise comparisons could be made between realizations. The simulation followed a classical sequential process, except that the random path was organized following a multigrid with four levels. At the first level, the grid was sub-sampled using an internode distance equal to the basic internode distance multiplied by 2^4. The nodes in this subsample were first simulated following a random path. Then, the operation was repeated for lower levels using powers from 3 to 1 to sample the grid. In this way, the random path was partially organized by spreading regularly the first simulated values. At each node of the grid, the neighbourhood consisted of the eight closest previously simulated nodes. The resulting maps for the most probable category were compared. In addition, we also compared the obtained proportions and the bivariate probability functions estimated on these images with the reference bivariate probability functions.

The three sets of realizations presented in Figure 4 show that BME and MCP are very close to each other. The North-East/South-West asymmetry was reproduced in the simulations. The maps produced by MCP had more regular boundaries between categories than those obtained with BME. This can be an advantage since more geographically coherent areas are produced.

The average proportions were very close to the original ones ([0.426 0.280 0.294]): [0.408 0.290 0.302] for MCP, and [0.435 0.280 0.285] for BME. Figure 5 depicts the mean bivariate probability functions estimated on 50 realizations for each method. The differences with the reference model and between methods are quite small indicating that the methods are able to reproduce the spatial model. Computation time is about 20 times smaller for MCP (8.5 min.) than for BME (177 min.).

Real case study

As a real case, we used the lithology data available in the well known Jura data set (Atteia *et al.*, 1994; Goovaerts, 1997). Estimation of the lithology may be important for soil scientists, e.g. for understanding the distribution of heavy metal in soils, for example. Available observations have been split into a prediction set (292 locations) and a validation set (100 locations), as represented in Figure 6a and 6b. In the original data set, five rock types were available. Following Bel *et al.* (2009), Portlandian was grouped with Quaternary into category 4, because of its very low frequency of occurrence (1.2%). Their marginal probabilities were, respectively, equal to (0.21 0.34 0.25 0.20). Two kinds of bivariate probability models were adjusted to the prediction set, omnidirectional and directional (Figure 7). The directional bivariate probability functions were computed for the eight principal directions equally distributed along the circle with a tolerance of 45. This was considered as a good compromise between maximizing the number of directions and having enough data to estimate the functions in a robust way in each direction.

For both methods, the neighbourhood was restricted to the five nearest data. Prediction results are shown at Figure 6. The correct classification rate was 64% for MCP (wether the model was directional or not) and 67% or 65% for BME, depending on wether the model was directional or not. The use of the directional model did not seem to improve the maps significantly, probably because of the absence of ordered sequences. MCP maps showed smoother inter-category boundaries and ignored some small features, but produced acceptable maps.

As a measure of uncertainty on the prediction, the Gini index of the local conditional probability $G = 1 - \sum_{k=1}^{4} p_k^{*2}$ was computed. Maps of the Gini index are shown in Figure 8. The larger dark areas on the MCP maps reveal a smaller uncertainty in

the local probability distributions. This is because MCP does not take the clustering of the data into account, thus resulting in a more clear-cut choice of the maximum probability category.

Conclusion

We have presented a new approach, MCP, for predicting a categorical variable in a spatial context. We have shown that our approach is the exact solution of a maximum entropy principle constrained to verify the bivariate probabilities having the prediction point as one end-point. Solving this conditional maximum entropy problem is a much less difficult task than solving the maximum entropy principle constrained to verify all bivariate probabilities, which is the full BME approach. In all tested situations MCP yielded very comparable classification rates with (full) BME while being orders of magnitude faster to compute. MCP is thus a fast, easy to code and accurate approximation of full BME. Moreover, MCP has an interesting 0/1 forcing property which makes it suitable for modeling soil or stratigraphic sequences, such as catenary patterns, when used with multidirectional bivariate probability functions.

Drawing from the prediction probability distribution by Monte-Carlo in a sequential simulation context provides simulated categories which can be conditionned to existing data. Simulated maps enjoy the same 0/1 forcing property than prediction maps, as illustrated on our synthetic data set.

Another advantage of the method is the flexibility of the spatial structure inference: (i) experimental bivariate probabilities functions can be estimated from a training image as well as from a point set, or even from a conceptual model drawn by the user, (ii) omnidirectional or asymetric directional functions can be computed, allowing the capture of ordered sequences, and (iii) the experimental functions are fitted using an easy-to-use kernel smoothing procedure, so avoiding parametric models which are

nearly always not sufficiently flexible to capture the subtleties of the spatial structure of a categorical variable.

Unlike kriging which is a minimum variance predictor, MCP is based on a maximum entropy principle. Contrary to kriging approaches, maximum entropy approaches do not directly provide a map of uncertainty associated with the prediction. Such a map is however easy to build using a large number of conditional simulations.

This work is the first implementation of the MCP principle for modeling spatial categorical variables. Stationarity was thus assumed and no secondary variables were considered. Future research should aim (i) at proposing a non-stationary version of MCP and (ii) accounting for secondary, discrete or continuous variables.

Appendix: proof of proposition

We want to prove the following: the probability distribution

$$p^*_{i_0,i_1,\ldots,i_n} = P^*[Y(\mathbf{x}_0) = i_0, Y(\mathbf{x}_1) = i_1, \ldots, Y(\mathbf{x}_n) = i_n] \tag{A.1}$$

maximizing its entropy $H(P^*)$ and subject to the following univariate and bivariate constraints

1. $\sum_{i_1,\ldots,i_n} p^*_{i_0,i_1,\ldots,i_n} = p^*_{i_0} = p_{i_0}$,
2. $\sum_{i_1,\ldots,i_n} p^*_{i_0,i_1,\ldots,i_n} \mathbf{1}_{[i_k=j]} = p^*_{i_0,i_k=j} = p_{i_0,j}(\mathbf{h}_{0k})$, for $k = 1, \ldots, n$ and $j = 1, \ldots, I$,

is:

$$p^*_{i_0,i_1,\ldots,i_n} = p_{i_0}^{1-n} \prod_{k=1}^{n} p_{i_0,i_k}(\mathbf{h}_{0k}) = p_{i_0} \prod_{k=1}^{n} p_{i_k|i_0}(\mathbf{h}_{0k}) \text{(A.1)}.$$

For ease of notation, we use $\sum_{i_1,\ldots,i_n}$ as a short notation for $\sum_{i_1=1}^{I} \cdots \sum_{i_n=1}^{I}$.

Proof The adequate approach is to use the Lagrange multiplier technique on the

objective function

$$\begin{aligned} J &= -\sum_{i_1,\ldots,i_n} P_{i_0,i_1,\ldots,i_n} \ln P_{i_0,i_1,\ldots,i_n} + \mu \left(\sum_{i_1,\ldots,i_n} P_{i_0,i_1,\ldots,i_n} - p_{i_0} \right) \\ &+ \sum_{k=1}^{n} \sum_{i=1}^{I} \lambda_{k,i} \left(\sum_{i_1,\ldots,i_n} P_{i_0,i_1,\ldots,i_n} \mathbf{1}_{[i_k=i]} - p_{i_0,i}(\mathbf{h}_{0k}) \right), \end{aligned}$$

where μ and the λs are Lagrange multipliers. For finding the solution of the constrained problem, we set all partial derivatives to 0. This leads to the system of equations:

$$\ln p^*_{i_0,i_1,\ldots,i_n} = -1 + \mu + \sum_{k=1}^{n} \sum_{i=1}^{I} \lambda_{k,i} \mathbf{1}_{[i_k=i]}, \tag{A.2}$$

$$\sum_{i_1,\ldots,i_n} p^*_{i_0,i_1,\ldots,i_n} = p_{i_0}, \tag{A.3}$$

$$\sum_{i_1,\ldots,i_n} p^*_{i_0,i_1,\ldots,i_n} \mathbf{1}_{[i_k=i]} = p_{i_0,i}(\mathbf{h}_{0k}), \text{ for } k = 1,\ldots,n \text{ and } i = 1,\ldots,I. \tag{A.4}$$

From Equation (A.2):

$$p^*_{i_0,i_1,\ldots,i_n} = e^{-1+\mu} \prod_{k=1}^{n} \prod_{i=1}^{I} e^{\lambda_{k,i} \mathbf{1}_{[i_k=i]}} = e^{-1+\mu} \prod_{k=1}^{n} e^{\lambda_{k,i_k}}. \tag{A.5}$$

Inserting Equation (A.5) into Equation (A.3) yields

$$e^{-1+\mu} = \frac{p_{i_0}}{\sum_{i_1,\ldots,i_n} \prod_{k=1}^{n} e^{\lambda_{k,i_k}}}, \tag{A.6}$$

and substituting Equation (A.6) into Equation (A.5) leads to

$$p^*_{i_0,i_1,\ldots,i_n} = \frac{p_{i_0} \prod_{k=1}^{n} e^{\lambda_{k,i_k}}}{\sum_{i_1,\ldots,i_n} \prod_{k=1}^{n} e^{\lambda_{k,i_k}}}. \tag{A.7}$$

Now, to find the Lagrange multipliers $\lambda_{k,i}$, let us consider the particular bivariate constraint $p^*_{i_0,i_1=j} = p_{i_0,j}(\mathbf{h}_{01})$. Then,

$$\begin{aligned} p^*_{i_0,i_1=j} &= \sum_{i_2,\ldots,i_n} p^*_{i_0,j,i_2,\ldots,i_n} \\ &= \frac{p_{i_0} \sum_{i_2,\ldots,i_n} e^{\lambda_{1,j}} \prod_{k=2}^{n} e^{\lambda_{k,i_k}}}{\sum_{i_1,\ldots,i_n} \prod_{k=1}^{n} e^{\lambda_{k,i_k}}} \\ &= \frac{p_{i_0} e^{\lambda_{1,j}} \sum_{i_2,\ldots,i_n} \prod_{k=2}^{n} e^{\lambda_{k,i_k}}}{\sum_{i=1}^{I} e^{\lambda_{1,i}} \sum_{i_2,\ldots,i_n} \prod_{k=2}^{n} e^{\lambda_{k,i_k}}} \\ &= \frac{p_{i_0} e^{\lambda_{1,j}}}{\sum_{i=1}^{I} e^{\lambda_{1,i}}}. \end{aligned}$$

In order for the constraint to be verfied, one must thus have $e^{\lambda_{1,j}} = c_1 p_{i_0,j}(\mathbf{h}_{01})/p_{i_0}$. Without loss of generality, one can set $c_1 = 1$. Considering this result to all sites $k = 1, \ldots, n$, and inserting into Equation (A.7) we finally obtain

$$p^*_{i_0,i_1,\ldots,i_n} = p_{i_0}^{1-n} \prod_{k=1}^{n} p_{i_0,i_k}(\mathbf{h}_{0k}). \tag{A.8}$$

□

As an alternative proof, a general theorem (Cover & Thomas, 2006) states that under very weak conditions, which are verified here, the maximum entropy solution is unique. Hence, verifying that the solution $e^{\lambda_{j,1}} = p_{i_0,j}(\mathbf{h}_{01})/p_{i_0}$ is correct is equivalent to solving (A.2)–(A.4).

Acknowledgements We thank Pierre Biver from Total for providing the sand dune image. We are extremely grateful to two anonymous referees and to the Guest Editor for their constructive remarks which helped us to improve greatly the quality of the manuscript.

References

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

Bel, L., Allard, D., Laurent, J.M., Cheddadi, R. & Bar-Hen, A. 2009. CART algorithm for spatial data: Application to environmental and ecological data. *Computational Statistics and Data Analysis,* **53,** 3082–3093.

Besag, J.E. 1974. Spatial interaction and the statistical analysis of lattice system (with Discussion). *Journal of the Royal Statistcal Society, B,* **36,** 192–236.

Bierkens, M.F.P. & Burrhough, P.A. 1993a. The indicator approach to categorical soil data. I. Theory. *European Journal of Soil Science,* **41,** 361–368.

Bierkens, M.F.P. & Burrhough, P.A. 1993b. The indicator approach to categorical soil data. II. Application to mapping and land use suitability analysis. *European Journal of Soil Science,* **41,** 369–381.

Bogaert, P. 2002. Spatial prediction of categorical variables: the Bayesian Maximum Entropy approach. *Stochastic Environmental Research & Risk Assessment,* **16,** 425–448.

Bogaert P. & D'Or D. 2002. Estimating soil properties from thematic soil maps : the Bayesian maximum entropy approach. *Soil Science Society of America Journal,* **66,** 1492–1500.

Carle, S.F. & Fogg, G.E. 1996. Transition probability-based indicator geostatistics. *Mathematical Geology,* **28,** 453–476.

Chilès, J.-P. & Delfiner, P. 1999. *Geostatistics: Modeling Spatial Uncertainty.* John Wiley & Sons, New-York.

Christakos, G. 1990. A Bayesian/maximum entropy view to spatial estimation problem. *Mathematical Geology,* **30,** 435–462.

Cover, T. M. & Thomas, J. A. 2006. *Elements of Information Theory; 2nd edition*, John Wiley & Sons, New-York.

Csiszár, I. 1991. Why least squares and maximum entropy ? An axiomatic approach to inference for linear inverse problems. *The Annals of Statistics,* **19,** 2032–2066.

Deming, W.E. & Stephan, F.F. 1940. On a least squares adjustement of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics,* **11,** 427–444.

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning,* **29,** 103-130.

D'Or, D. & Bogaert, P. 2004. Spatial prediction of categorical variables with the Bayesian Maximum Entropy approach: the Ooypolder case study. *European Journal of Soil Science,* **55,** 763–775.

D'Or D., Bogaert P. & Christakos G. 2001. Application of the BME approach to soil texture mapping. *Stochastic Environmental Research & Risk Assessment,* **15,** 87–100.

Goovaerts P. 1997. *Geostatistics for Natural Resources Evaluation.* Oxford University Press, New-York.

Heagerty, P.J. & Lele, S.R. 1998. A composite likelihood approach to binary spatial data. *Journal of the Amererican Statistical Association,* **93,** 1099–1111.

Hjort, N.L. & Omre, H. 1994. Topics in spatial statistics (with Discussion). *Scandanivian Journal of Statistics,* **21,** 289–357.

Journel, A.G. 1983. Non parametric estimation of spatial distributions. *Mathematical Geology,* **15,** 445–468.

Journel, A.G. & Alabert, F. 1989. Non-Gaussian data expansion in Earth Sciences. *Terra Nova,* **1,** 123–134

Kullback, S. & Leibler, R.A. 1951. On information and sufficiency. *Annals of Mathematical Statistics,* **22,** 76–86.

Li, W. 2007. Markov Chain Random Fields for Estimation of Categorical Variables. *Mathematical Geology,* **39,** 321–225.

Li, W. & Zhang, C. 2007. A Random-Path Markov Chain Algorithm for Simulating Categorical Soil Variables from Random Point Samples *Soil Science Society of America Journal,* **71,** 656–668.

Li, W., Zhang, C. Burt, J.E., Zhu, A.-X. & Feyen, J. 2004. Two-dimensional Markov Chain Simulation of Soil Type Spatial Distribution. *Soil Science Society of America Journal,* **68,** 1479–1490.

Mariethoz, G., & Renard, P. 2010. Reconstruction of incomplete data sets or images using direct sampling. *Mathematical Geosciences,* **42,** 245–268.

Nott, D. J. & Rydén, T. 1999. Pairwise likelihood methods for inference in image analysis. *Biometrika,* **86,** 661–676.

Silverman, 1986. *Density Estimation for Statistics and Data Analysis,* Chapman & Hall, London.

Strebelle, S. 2002. Conditional simulation of complex geological structures using multi-points statistics, *Mathematical Geology*, **34**, 1–22.

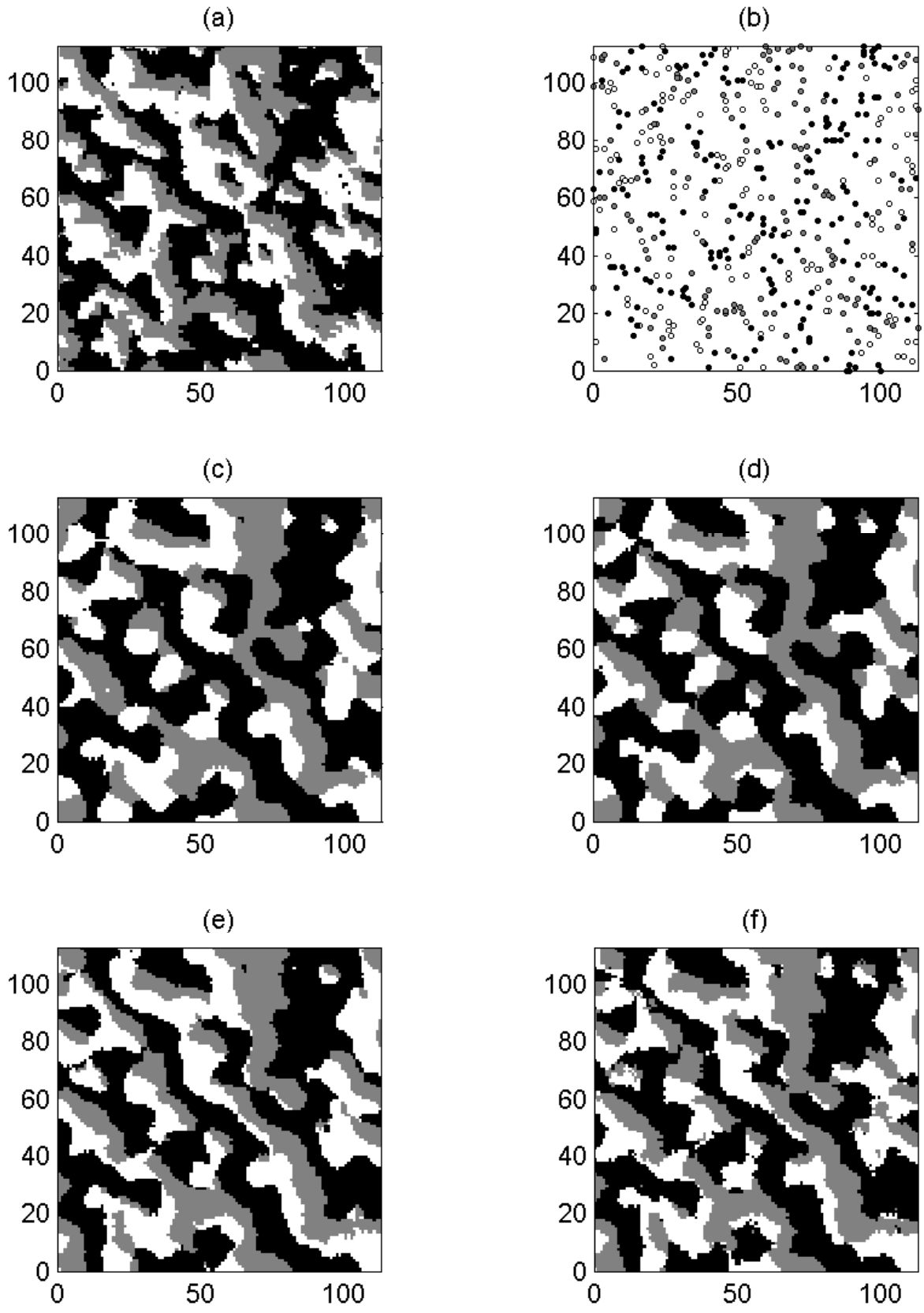
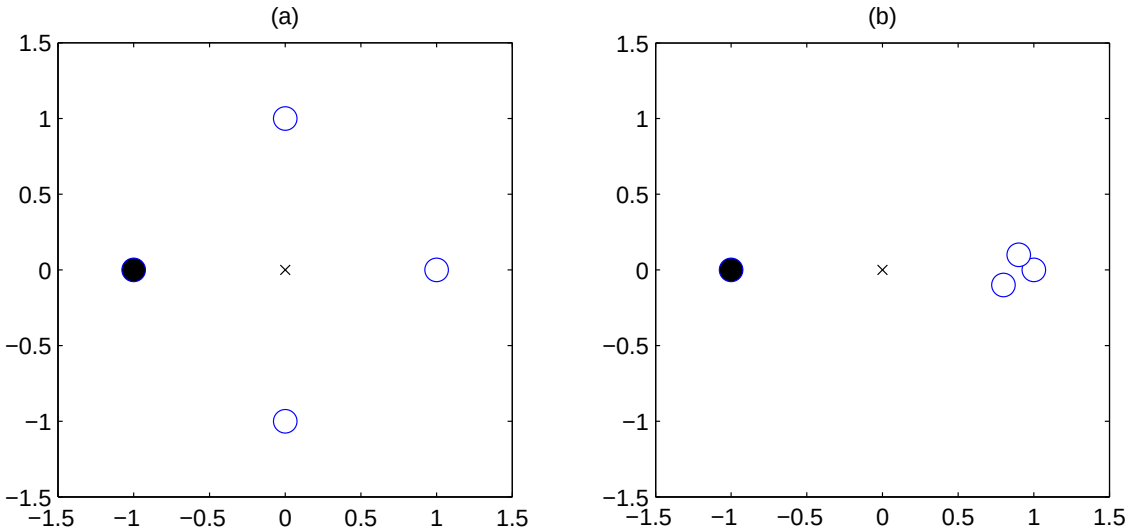


Figure 1: Reference image (a) and 500 samples data set (b). Prediction maps (most probable category) obtained with MCP ((c) and (e)), and BME ((d) and (f)). Maps (c) and (d) use the omnidirectional bivariate probability model while maps (e) and (f) are use directional one. Black: category 1, gray: category 2, and white: category 3.



	Cat. 1	Cat. 2	Cat. 3
BME	0.0014	0	0.9986
MCE	0.0012	0	0.9988

	Cat. 1	Cat. 2	Cat. 3
BME	0.4351	0.0007	0.5642
MCE	0.0033	0	0.9967

Figure 2: Taking data clustering into account. (a) a non clustered situation and (b) a clustered one. The cross symbolizes the estimation point, the black circle represents a datum of category 1 and the white circles data of category 3. Below: local conditional distributions obtained with BME and MCP using a directional bivariate probability model.

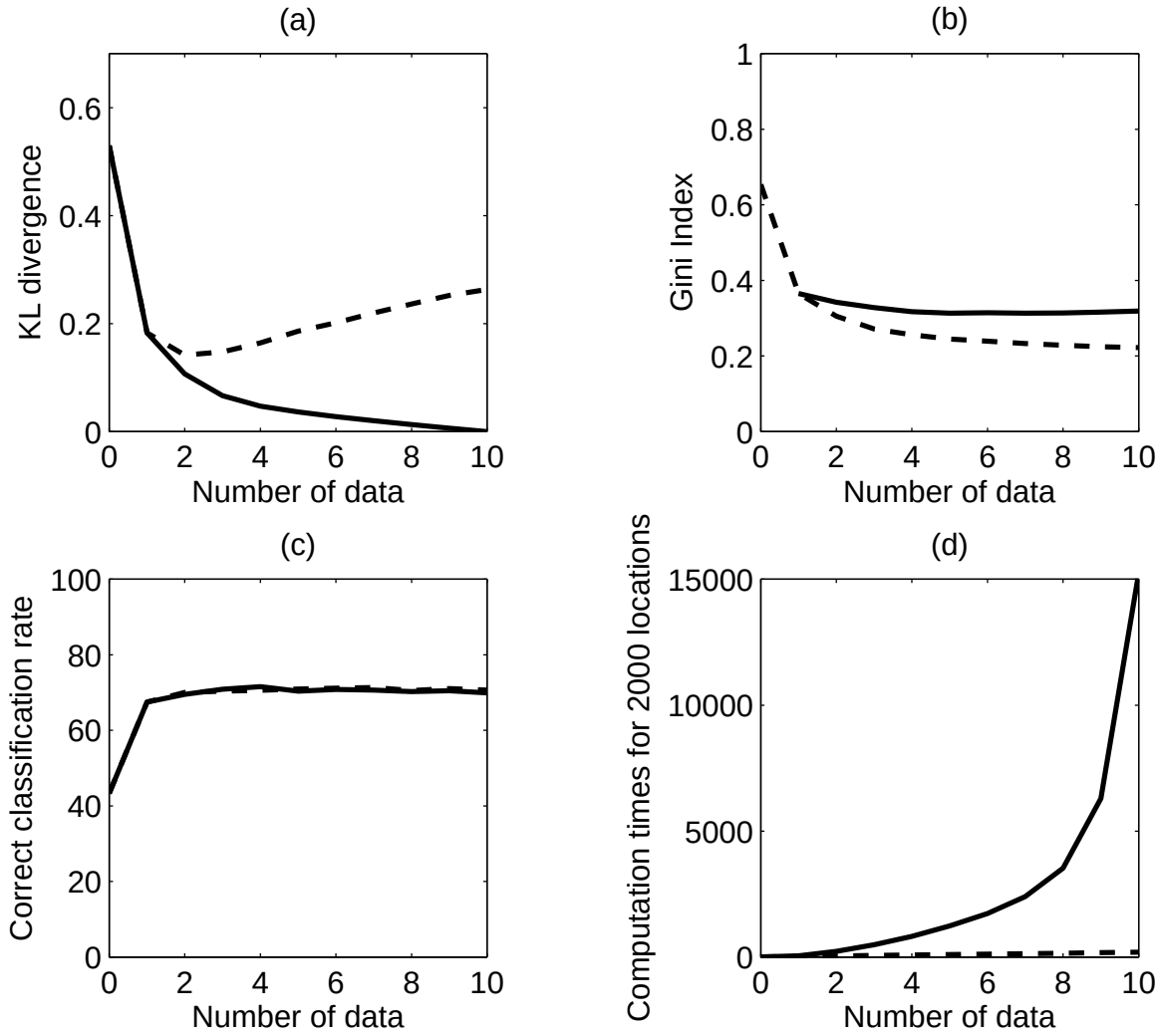


Figure 3: Effect of increasing number of data in the neighbourhood on (a) KL-distances, (b) Gini index, (c) correct classification rate and (d) computation times obtained for BME (plain lines) and MCP (dashed lines).

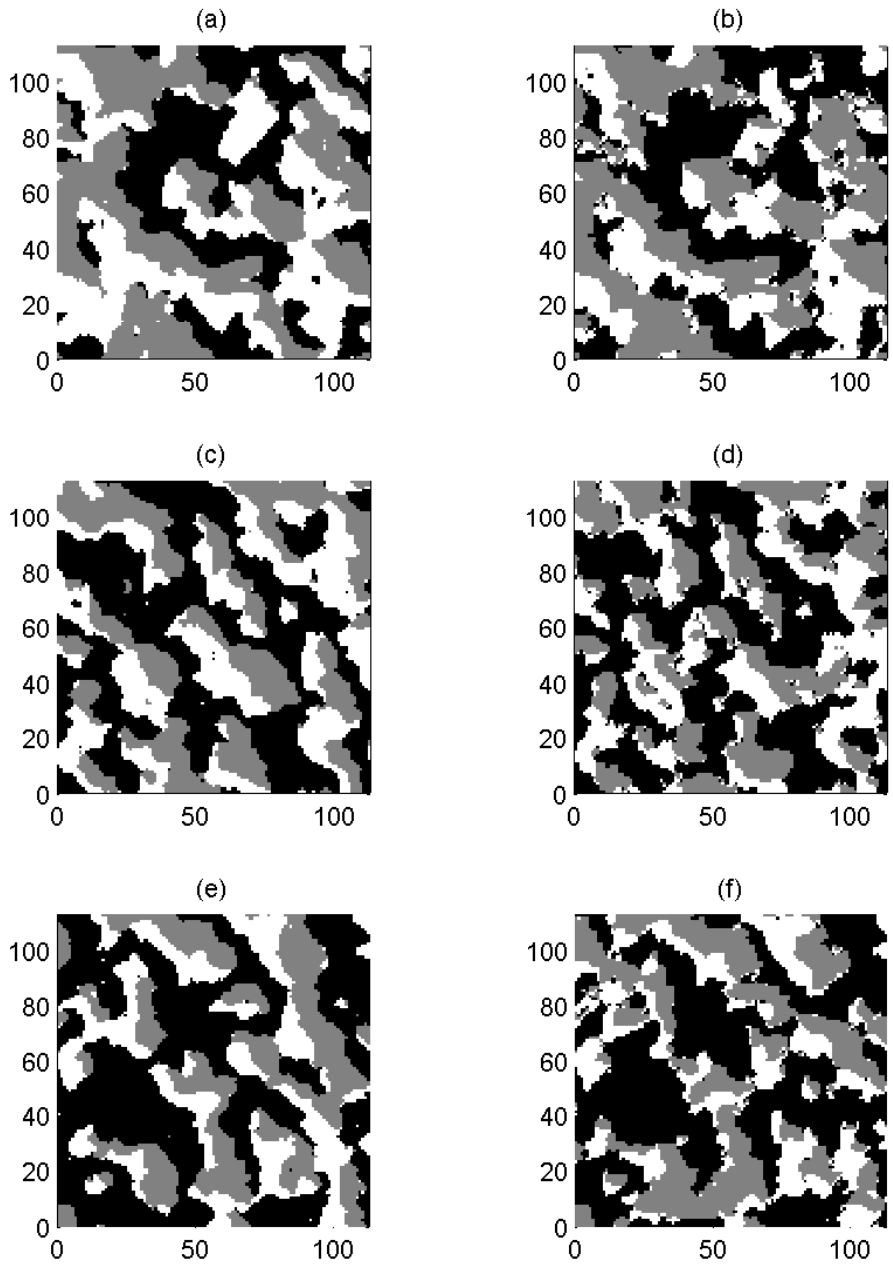


Figure 4: Comparison of realizations of non conditional simulations (maps of the most probable category) obtained with MCP (Figures (a), (c), and (e)) and BME (Figures (b), (d), and (f)). A directional bivariate probability model was used. Realizations on the same row are paired and use the same ramdom path and seed. Black: category 1, gray: category 2, and white: category 3.

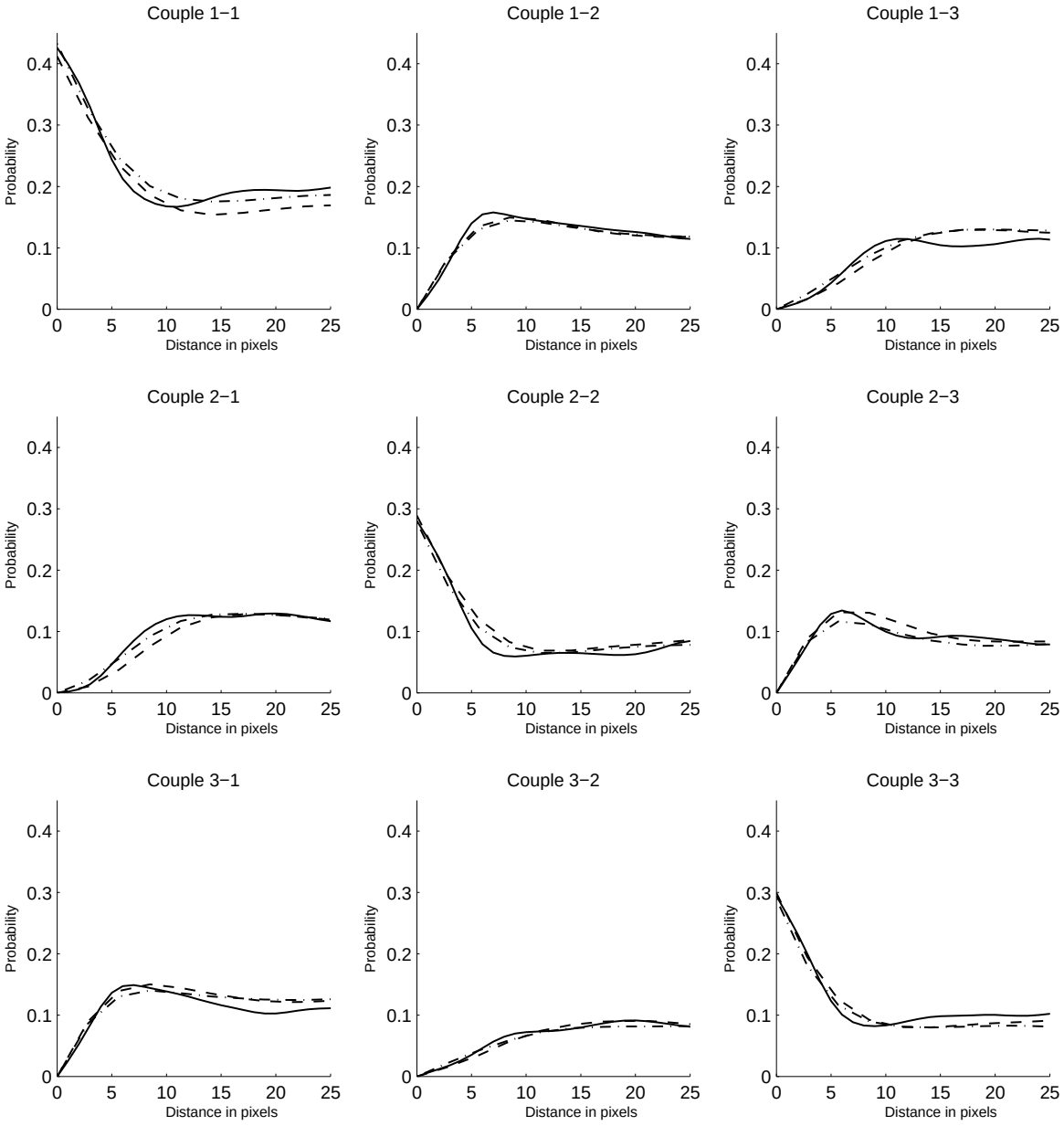


Figure 5: Bivariate probability functions for the three categories synthetic example. On each subplot, the bivariate probability functions are given for the North-East directional model, respectively for the reference model (plain lines), MCP (dashed lines) and BME (dash-dotted lines).

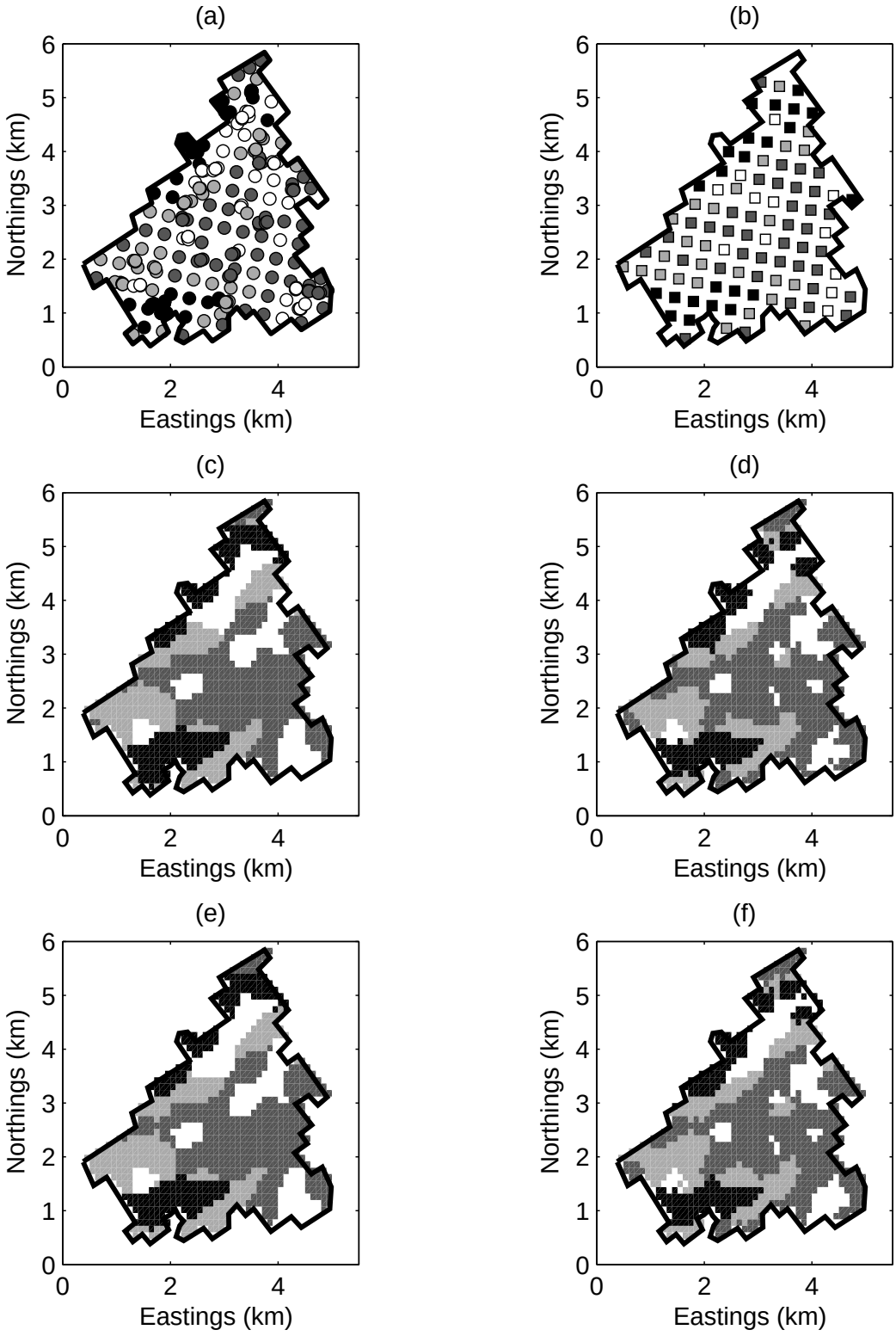


Figure 6: Jura data set prediction maps. Prediction (a) and validation (b) sets. Prediction maps (most probable category) obtained with MCP ((c) and (e)), and BME ((d) and (f)). Maps (c) and (d) use the omnidirectional bivariate probability model while maps (e) and (f) use the directional one. Categories 1 to 4 have colors from black to white.

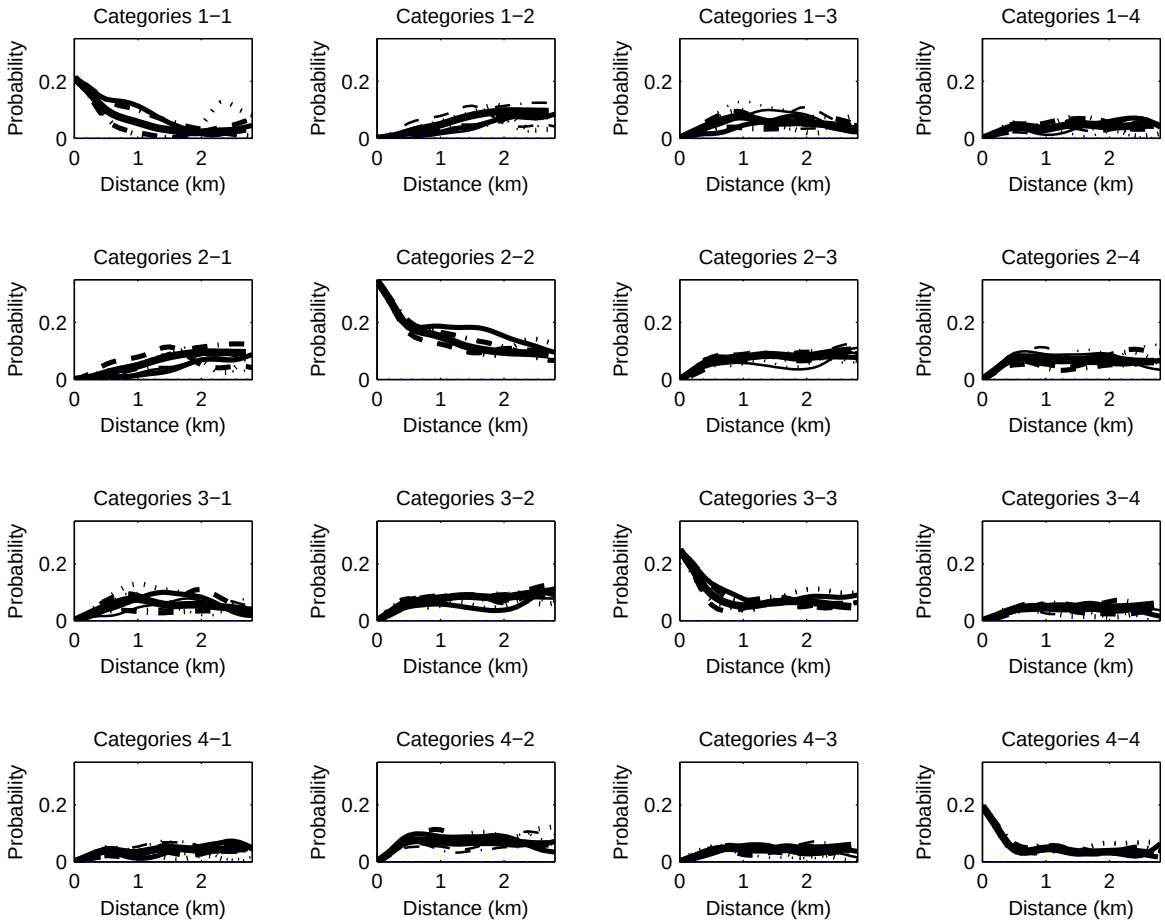


Figure 7: Jura data set bivariate probability models: omnidirectional (plain line), East-West (dashed line), South-North (dash-dotted line). On each subplot, the bivariate probability functions are given for the omnidirectional model (thick plain line) and for eight main directions: East (medium plain line), North-East (medium dash-dotted line), North (medium dashed lin), North-West (medium dotted line), West (thin plain line), South-West (thin dash-dotted line), South (thin dashed lin), South-East (thin dotted line). Note the antisymmetry of the functions.

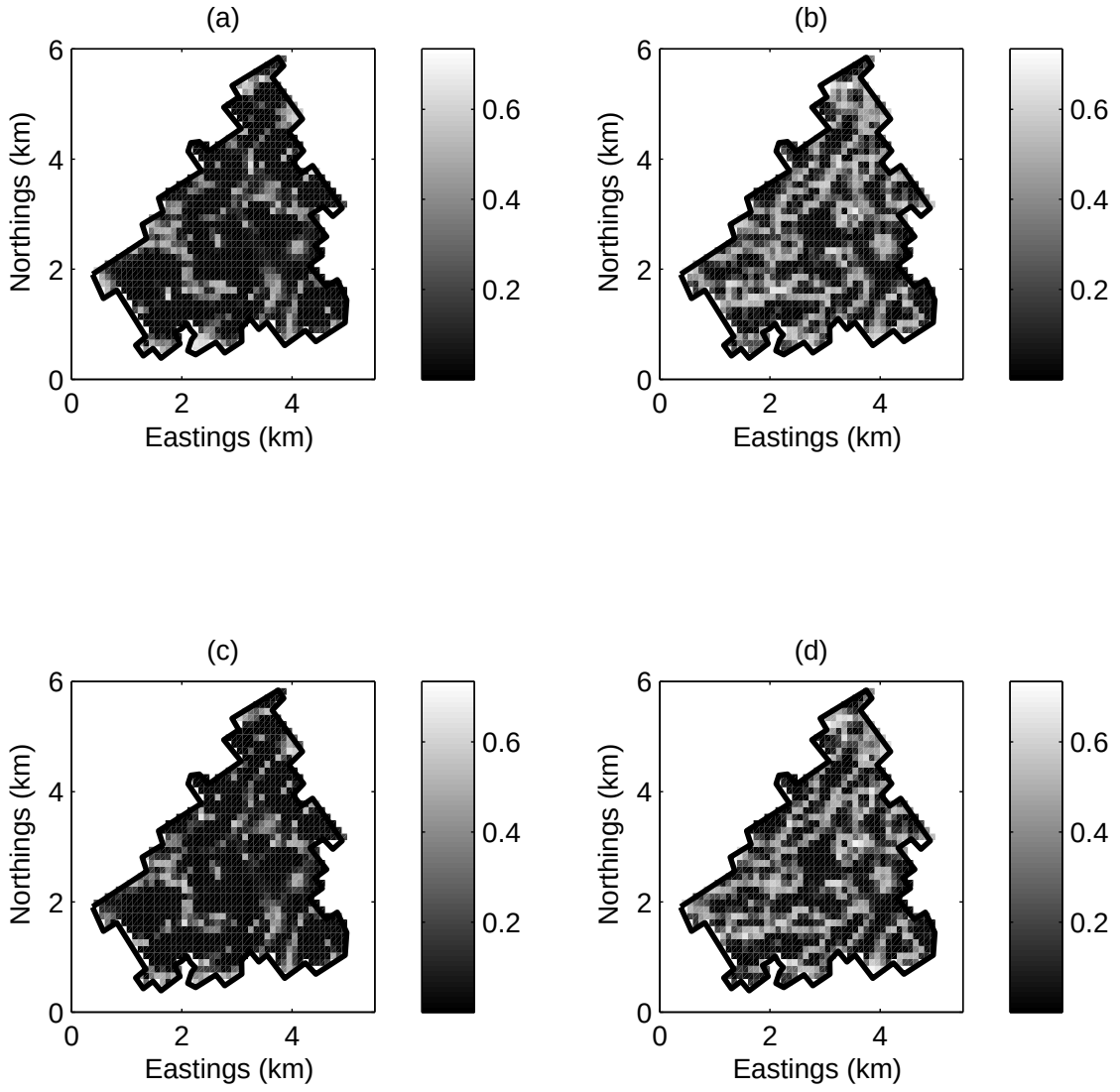


Figure 8: Jura data set uncertainty maps. Maps of the Gini index obtained with MCP ((a) and (c)), and BME ((b) and (d)). Maps (a) and (b) use the omnidirectional bivariate probability model while maps (c) and (d) use the directional one.

List of Figures

diction maps (most probable category) obtained with MCP ((c) and (e)), and BME ((d) and (f)). Maps (c) and (d) use the omnidirectional bivariate probability model while maps (e) and (f) use the directional one. Categories 1 to 4 have colors from black to white.

Figure 7. Jura data set bivariate probability models: omnidirectional (plain line), East-West (dashed line), South-North (dash-dotted line). On each subplot, the bivariate probability functions are given for the omnidirectional model (thick plain line) and for eight main directions: East (medium plain line), North-East (medium dash-dotted line), North (medium dashed lin), North-West (medium dotted line), West (thin plain line), South-West (thin dash-dotted line), South (thin dashed lin), South-East (thin dotted line). Note the antisymmetry of the functions.

Figure 8. Jura data set uncertainty maps. Maps of the Gini index obtained with MCP ((a) and (c)), and BME ((b) and (d)). Maps (a) and (b) use the omnidirectional bivariate probability model while maps (c) and (d) use the directional one.