New Software for Fitting Indicator Covariances and for Indicator Kriging/Simulation

Clayton V. Deutsch[1], Julián M. Ortiz[2], and Chad Neufeld[1]

[1]Centre for Computational Geostatistics (CCG)
University of Alberta

[2]Department of Mining Engineering
University of Chile

Indicator based geostatistical techniques provide great flexibility to integrate soft data and account for more complex spatial features than conventional Gaussian techniques. A longstanding problem, however, has been order relation deviations partly caused by inconsistently modeled indicator variograms. The bivariate distributions summarized by indicator variograms must be valid distributions, that is, there can be no negative probabilities, the probabilities must sum to one and the corresponding marginal distributions must be reasonable. These consistency constraints are derived and quantified. Indicator variograms can be checked and the constraints used in semiautomatic fitting of indicator variograms.

Typical implementations of indicator kriging and simulation consider independent kriging of each category/threshold. Cokriging with cross covariances is avoided primarily because of awkward inference and modeling. The semiautomatic fitting permits all direct and cross variograms to be fit simultaneously. We implement full indicator cokriging as an option in indicator kriging and simulation, which brings potential benefits.

Introduction

Indicator techniques are commonly used for categorical variables such as facies and rock types. The indicator variogram for each rock type is unique, which provides flexibility to capture the architecture of different rock types. It is also straightforward to account for locally varying rock type proportions through non-stationary simple kriging. Indicator methods such as indicator kriging (IK) and sequential indicator simulation (SIS) are well suited to settings where the rock types are structured, but with no clear geological object shapes. Object-based modeling would be suitable for clear geologic shapes.

Continuous variables are also modeled with indicator techniques. Gaussian techniques are simpler, but are less flexible than indicator techniques such as IK and SIS. The spatial continuity patterns of the low, median and high values tend to be different. Indicator techniques permit the different variograms for different thresholds. Secondary data are easily coded as soft indicator data for integration with a form of cokriging or non-stationary kriging.

Multiple point statistics (MPS) permit complex spatial variation patterns to be imposed on geostatistical models; however, they call for a training image and do not use statistics such as variograms calculated directly from the available hard data. MPS are relevant when a deemed reliable training image is available and all available hard and soft data can be integrated.

Notwithstanding the availability of alternate techniques, there are cases where indicator based methods are suitable and appropriate.

This paper is concerned with the best practice of implementing indicator methods. In particular, this paper is concerned with the bivariate spatial law implicit to the modeled indicator variograms and cross variograms. The current practice of independently modeling the indicator variograms for different rock types or thresholds makes it easy to create an implausible spatial law, that is, bivariate distributions that have negative probabilities or inconsistent univariate distributions. We derive the constraints on the indicator variograms and show how they can be used to check available indicator statistics or included in semiautomatic variogram fitting

Spatial Correlation of Categorical Variables

Many principles of indicator statistics are the same with categorical variables and continuous variables; however, their definition and how they discretize a probability distribution are quite different. An indicator for a categorical variable is 1 if it is present and 0 if not, which is like a discrete probability coding. An indicator for a continuous variable is 1 if the value is less than a threshold, which is a cumulative distribution function coding.

Consider K different categories. These categories could represent rock types, facies or any other categorical variable. They are mutually exclusive – only one category can exist at a particular location. They are exhaustive – one of the categories must exist at all locations. The categorical variable is expressed as a series of K indicator variables:

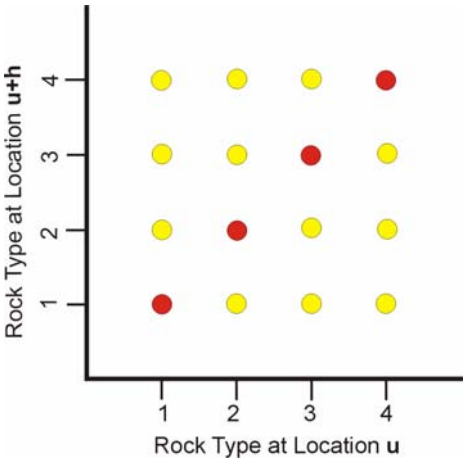
$$i(\mathbf{u};k) = \begin{cases} 1, \text{ if category } k \text{ prevails at location } \mathbf{u} \\ 0, \text{ otherwise} \end{cases}, \quad k = 1,...,K \tag{1}$$

An indicator variable is often interpreted as the probability for a category to prevail at a particular location: the probability is 1 if it does prevail and 0 if it does not. Hard local measurements are coded into 1s and 0s. Imprecise or soft measurements may be coded into continuous probabilities between 0 and 1.

The bivariate distribution for a particular $\mathbf{h}$ lag vector is fully defined by K^2 transition probabilities.

$$p(\mathbf{h};k,k') = \text{Prob}\{\mathbf{u} \in \text{k}, \mathbf{u}+\mathbf{h} \in \text{k}'\}, \quad k,k' = 1,...,K \tag{2}$$

A schematic illustration of the 16 transition probabilities (for K=4) is shown to the right. The K dark colored dots along the diagonal are when both locations are in the same category. The remaining $K(K-1)$ light colored dots represent cases where the category changes.



There are a number of constraints on the bivariate transition probabilities. The sum of all K^2 transition probabilities must be 1. These probabilities would be inferred as proportions, that is, we would scan through the available data and identify the number of pairs corresponding the lag distance: $N(\mathbf{h})$, then we would calculate $p(\mathbf{h},k,k')$ as the proportion of k/k' pairs.

If the domain is reasonably large and the lag distance is reasonably small, then the univariate probability of each category will match the naïve equal weighted proportion of the category. This provides *2K* additional constraints:

$$\sum_{j=1}^{K} p(\mathbf{h};\, j,k) = p_k \quad \text{and} \quad \sum_{j=1}^{K} p(\mathbf{h};k,\, j) = p_k\,, \qquad k = 1,\ldots,K \tag{3}$$

A strong assumption of symmetry would entail that $p(\mathbf{h};k,k') = p(\mathbf{h};k',k)$, which would provide a further $K(K-1)/2$ constraints. The semivariogram will implicitly make this assumption, but geological phenomena often exhibit particular trends such as fining upward that contradicts this assumption. The *K* indicator variograms are defined as:

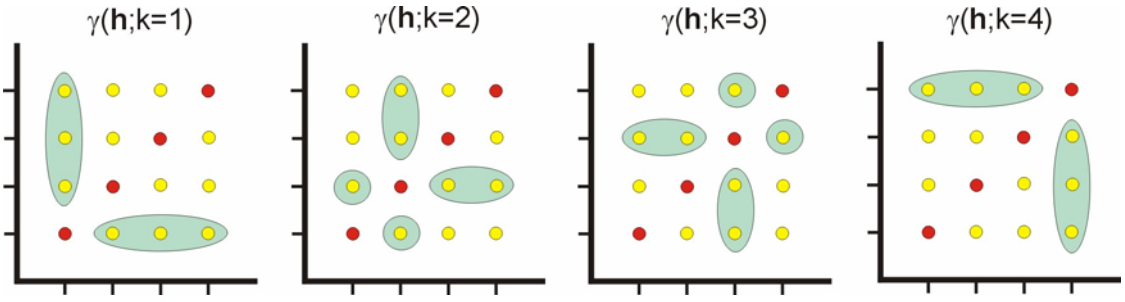
$$2\gamma(\mathbf{h};k) = E\left\{\left[I(\mathbf{u};k) - I(\mathbf{u}+\mathbf{h};k)\right]^2\right\}, \quad k = 1,\ldots,K \tag{4}$$

There are only four scenarios in the indicator variogram calculation for a particular lag h and category k:

1. Both **u** and **u+h** in the k^{th} category: $(1-1)^2 = 0$
2. Both **u** and **u+h** not in the k^{th} category: $(0-0)^2 = 0$
3. **u** in the k^{th} category and **u+h** not: $(1-0)^2 = 1$
4. **u+h** in the k^{th} category and **u** not: $(0-1)^2 = 1$

Thus, we see that there is a contribution to the indicator variogram when there is a transition from one category to another. Each indicator variogram summarizes *2(K-1)* different bivariate transition probabilities.

A schematic illustration of the transition probabilities summarized by the indicator variograms (for a *K=4* example) is shown below. The four illustrations are for the *k=1,2,3,4* indicator variograms. The bivariate points that are shaded are included in the indicator variograms.



The shaded areas are represented by the following equation:

$$2\gamma(\mathbf{h};k) = \sum_{\substack{k'=1 \\ k'\neq k}}^{K} p(\mathbf{h};k,k') + \sum_{\substack{k'=1 \\ k'\neq k}}^{K} p(\mathbf{h};k',k), \qquad k = 1,\ldots,K \tag{5}$$

Direct indicator variograms cannot quantify asymmetry; they are the average transition probabilities from category *k* to any other category and from any other category to *k*. Although indicator variograms do not quantify the direct transitions (the red dots on the diagonal), in

practice they are used to compute the direct transitions. Assuming stationarity, the row and column sums for a particular category k would equal p_k (see Equation 3), thus the row sum and the column sum plus two times the diagonal element would equal two times the univariate proportion of p_k:

$$\begin{aligned} 2\gamma(\mathbf{h};k) + 2p(\mathbf{h};k,k) &= 2p_k \\ \gamma(\mathbf{h};k) &= p_k - p(\mathbf{h};k,k), \text{ or} \qquad k = 1,\ldots,K \\ p(\mathbf{h};k,k) &= p_k - \gamma(\mathbf{h};k) \end{aligned} \tag{6}$$

Conventional indicator kriging only requires these direct terms. A full indicator cokriging would require the full matrix of transition probabilities for all lag distances. In practice, we would not consider the transition probabilities directly: we would center them under an assumption of stationarity and model them as covariances:

$$\begin{aligned} C(\mathbf{h};k,k') &= E\{I(\mathbf{u};k)\cdot I(\mathbf{u}+\mathbf{h};k')\} - p_k \cdot p_{k'}, \quad k,k' = 1,\ldots,K \\ &= p(\mathbf{h};k,k') - p_k \cdot p_{k'} \end{aligned} \tag{7}$$

Using the full matrix of K^2 covariances in indicator kriging/simulation is straightforward. The challenge is to fit them all to provide a licit model of coregionalization. This is where we must consider some form of semiautomatic fitting. In addition to imposing a licit model like the linear model of coregionalization, we would have to ensure than no transition probability ever goes negative, that is,

$$C(\mathbf{h};k,k') \geq p_k \cdot p_{k'}, \quad k,k' = 1,\ldots,K \tag{8}$$

The semiautomatic fitting will be discussed below.

Spatial Correlation of Continuous Variables

We consider continuous variables separately for simplicity. Consider K different threshold values, z_k, $k=1,\ldots,K$. These thresholds are in increasing order, that is, $z_{k'} > z_k$ for all $k' > k$. The indicator variable is defined as:

$$i(\mathbf{u};k) = \begin{cases} 1, \text{ if } z(\mathbf{u}) \leq z_k \\ 0, \text{ otherwise} \end{cases}, \quad k = 1,\ldots,K \tag{9}$$

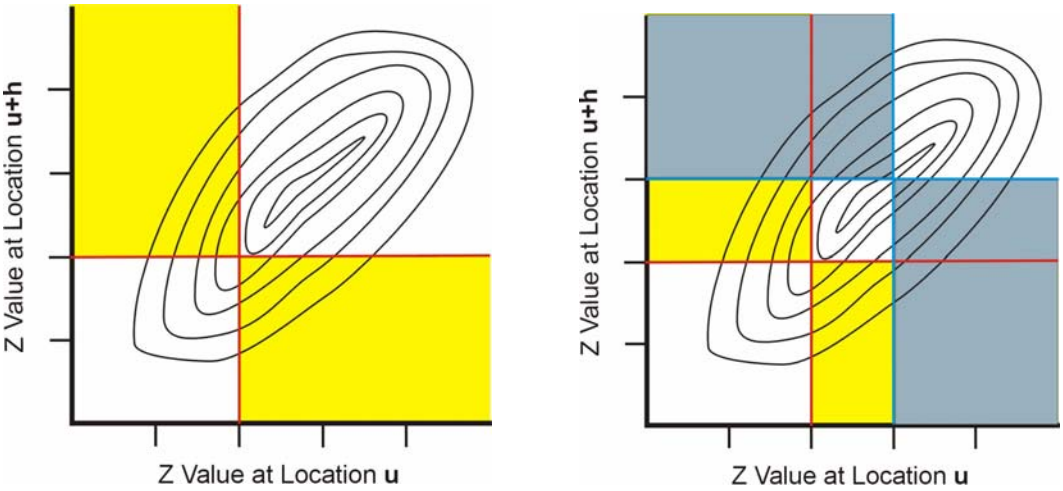
The indicator transform at a particular threshold is often interpreted as the probability for the variable to be less than or equal to that particular threshold, which is like a CDF coding of the data rather than the PDF coding used for categorical variables.

The bivariate distribution for a particular **h** lag vector is characterized by bivariate probabilities at the chosen thresholds:

$$F(\mathbf{h};z_k,z_{k'}) = \text{Prob}\{z(\mathbf{u}) \leq z_k, z(\mathbf{u}+\mathbf{h}) \leq z_{k'}\}, \quad k,k' = 1,\ldots,K \tag{10}$$

A full description of the bivariate distribution would require an infinite number of thresholds, but the main features are captured with as few as the nine deciles.

The definition of the indicator variogram for the continuous variable case remains the same as the categorical variable case (see Equation 4). Every pair of points Z(**u**) and Z(**u**+**h**) There are four possible outcomes; however, there is a contribution to the indicator variogram only when there is a transition, that is, both locations are *not* above or *not* below the cutoff. See the yellow shaded areas in the left hand side of the sketch below. As the threshold increases, the portion of the bivariate distribution that contributes to the indicator variogram changes (see the blue shaded area in the right hand side of the sketch).



The indicator variogram is linked to the bivariate CDF:

$$2\gamma(\mathbf{h};k) = F(\mathbf{h};k,K+1) + F(\mathbf{h};K+1,k) - 2F(\mathbf{h};k,k), \qquad k = 1,...,K \tag{11}$$

The first two terms on the right hand side are univariate CDF values and the last term is the bivariate CDF at the k^{th} threshold. The order relation property of probability distributions requires that all univariate and bivariate CDF values not decrease as the thresholds increase. This must be accounted for in fitting indicator variogram fitting.

As in the categorical variable case, conventional indicator kriging only requires these direct terms. A full indicator cokriging would require the full matrix of indicator covariances for all lag distances. The centered direct and cross indicator covariances would be calculated and fit under an assumption of stationarity and model them as covariances:

$$\begin{aligned} C(\mathbf{h};k,k') &= E\{I(\mathbf{u};k)\,I(\mathbf{u}+\mathbf{h};k')\} - F(k)\,F(k'), \quad k,k' = 1,...,K \\ &= F(\mathbf{h};k,k') - F(k)\,F(k') \end{aligned} \tag{12}$$

Once again, the challenge is to fit the matrix of K^2 covariances with a licit model of coregionalization. This is where we must consider some form of semiautomatic fitting. In addition to imposing a licit model of coregionalization, we must ensure than no univariate or bivariate cumulative probability ever decreases:

$$F(k) \geq F(k') \text{ for all } k > k'$$
$$\text{and}$$
$$C(\mathbf{h};k,k') + F(k)\,F(k') \geq C(\mathbf{h};l,l') + F(l)\,F(l') \tag{13}$$
$$\text{for all } k \geq l, k' \geq l', \quad k,k',l,l' = 1,...,K$$

These constraints are somewhat more complex than in the categorical variable case, but it is still easy to code in the semiautomatic fitting of indicator variograms, see below.

Semiautomatic Fitting of Indicator Covariances

The `varfit` program described in CCG report five/six and in the Computers & Geosciences journal was adapted for fitting a K^2 linear model of coregionalization of categorical or continuous indicator variables. Two new programs were written based on the program `varfit_lmc`: (1) `varfit_ind_cat` for fitting categorical indicator LMC's, and (2) `varfit_ind_cont` for fitting continuous indicator LMC's. The new versions of `varfit` require the output from `gamv2004` as input; they will not work with the `gamv` program from GSLIB (Neufeld and Deutsch, 2004). The ouput from the variogram fitting can be used directly as input to the indicator cokriging and simulation programs.

Categorical Indicator Variogram Fitting. Fitting a categorical indicator LMC was a simple extension of the existing LMC fitting program. The number of categories, the categories codes, and the proportion of each category are required as input. The input variogram types are limited to direct and cross categorical indicator variograms. No other changes were made to the input parameters.

The sills of the direct variograms are fixed based on the proportion of each facies type using the following formula:

$$\text{sill}_k = \sigma_k^2 = p_k(1 - p_k) \tag{14}$$

Neither the user nor the optimization is allowed to change the calculated sill value. The cross variogram sills are set at initial values to ensure the LMC is valid. The cross variogram sills are calculated as follows:

$$\text{sill}_{k,l} = \sigma_k^2 \cdot \sigma_l^2 \tag{15}$$

These sills can be changed by the optimization or fixed to a different value by the user.

Continuous Indicator Variogram Fitting. Continuous indicator LMC's have an additional constraint compared with categorical indicator LMC's; that is to make sure that no univariate or bivariate cumulative probability decreases, Equation (13). This check is done after the optimization procedure is complete. If an inconsistency is found, an error is reported to the screen. The number of thresholds, the threshold values, and the cumulative probability of each threshold are required as input. The input variogram types are limited to direct and cross continuous indicator variograms.

The sills of the direct variograms are fixed based on the cumulative probability for each threshold using the following formula:

$$\text{sill}_k = \sigma_k^2 = F\left(z_k\right)\left(1 - F\left(z_k\right)\right) \tag{16}$$

Neither the user nor the optimization is allowed to change the calculated sill value. The cross variogram sills are set at initial values to ensure the LMC is valid. The cross variogram sills are calculated using Equation (15). These sills can be changed by the optimization or fixed to a different value by the user.

Indicator Kriging

Local indicator data are used to estimate the distribution of uncertainty at unsampled locations. Commonly, K estimates $i^*(\mathbf{u};k)$, $k=1,\ldots,K$ are calculated by kriging with indicator data for the same category (or threshold):

$$\begin{aligned} i_{SK}^*\left(\mathbf{u};k\right) - p_k &= \sum_{\alpha=1}^{n} \lambda_\alpha^{SK}\left(\mathbf{u};k\right)\left[i\left(\mathbf{u}_\alpha;k\right) - p_k\right] \\ i_{SK}^*\left(\mathbf{u};k\right) &= \sum_{\alpha=1}^{n} \lambda_\alpha^{SK}\left(\mathbf{u};k\right) i\left(\mathbf{u}_\alpha;k\right) + \left[1 - \sum_{\alpha=1}^{n} \lambda_\alpha^{SK}\left(\mathbf{u};k\right)\right] p_k \end{aligned} \tag{14}$$

The K direct indicator covariances are required and will be reproduced by this classical simple kriging formalism. These estimates continuously between 0 and 1; they depart from the global probabilities in presence of relevant local data. Regardless of how the indicators are interpolated, they do not always satisfy the order relation requirements (1) they must be non-negative, (2) the categorical indicator estimates must sum to 1.0, and (3) the continuous indicator estimates must be non-decreasing. Standard order relation correction procedures exist (Deutsch and Journel, 1992). Order relation deviations are less when the underlying spatial model is consistent.

Sequential indicator simulation applies IK in a sequential fashion where a simulated value is drawn by Monte Carlo simulation at each location. All locations are visited sequentially with an increasing level of conditioning. A random order is followed to avoid artifacts. The simulated realization reproduces the indicator variograms according to the same simple kriging principle used in continuous variables. Order relation deviations lead to a lack of reproduction.

The `ik3d` program and the `sisim` program from GSLIB (Deutsch and Journel, 1998) provide public domain implementations of IK and SIS. No modifications are necessary.

Indicator Cokriging

Rather than just use indicators at the same category/threshold, we could use the full set of indicators in a cokriging estimator:

$$i_{FCK}^*\left(\mathbf{u};k\right) - p\left(\mathbf{u};k\right) = \sum_{\alpha=1}^{n}\sum_{l=1}^{K} \lambda_{\alpha,l}^{FCK}\left[i\left(\mathbf{u}_\alpha;l\right) - p\left(\mathbf{u};l\right)\right] \tag{15}$$

This formalism requires the full set of K^2 direct and cross covariances. The benefit of cokriging in an indicator context is questionable. The indicators at the same data location are highly correlated with each other; there is a very predictable succession of 0s and 1s. This relationship between the indicators is sometimes used to avoid full indicator cokriging. The most important reason to avoid full cokriging is the difficult fitting of a licit model of coregionalization. `varfit_ind` removes this concern.

The `ik3d_fc` and `sisim_fc` programs were derived from the GSLIB programs to implement full cokriging.

An Example

The Jura data set (Goovaerts, 1997) consists of 359 sample sites where several variables are measured. Land use and rock type are defined at each location, and concentrations of Cd, Cu, Pb, Co, Cr, Ni and Zn are also available. To illustrate the use of the new software discussed, the programs have been used to estimate the probability of prevalence of different land uses in this region of Switzerland using indicator cokriging. Four possible land uses are defined (Figure 1): forest, pasture, meadow, and tillage.

The indicator variograms and cross-variograms are calculated and modeled using the `gamv2004` and `varfit_ind_cat` programs. The parameters for variogram calculation are summarized in Table 1. The semi-automatic fitting requires an initial model that is positive definite. This is not automatically set up by the program, hence the total sills in the variograms and cross-variograms were set to provide a valid initial model. These sills were calculated based on the experimental proportions of the land uses (Table 2). The validity of the model was checked by verifying that the principal minors of the matrix of sills are positive (Table 3). The principal minors are the determinants of the k by k submatrices (k=1,…,n), whose diagonal elements lie on the principal diagonal of the covariance matrix.

The program `varfit_ind_cat` provided a positive definite model (Figure 2 and Table 4), which was used for indicator cokriging. The probability of a given land use to prevail at each estimation location was calculated. The four probabilities should sum to one at every location, however small deviations occur, which have to be corrected to provide a consistent model.

Figure 3 shows four location maps where the probability of a given land use prevailing at every location are depicted. These probabilities have been corrected to sum to one. Based on these probabilities, a classification can be performed, by assigning the most likely land use to every location, but honoring the global proportions. The classification is shown in Figure 4.

The implementation of indicator cokriging is greatly facilitated by the new software, allowing the integration of all cross-correlations between the categories.

Conclusions

Constraints for the computation of indicator direct and cross-variograms have been proposed, which, along with the programs provided can greatly facilitate the modeling of the spatial relationships between categories, through the use of the linear model of coregionalization (LMC).

Once the LMC model has been defined, implementation of indicator kriging and simulation are straightforward. The use of variograms and cross-variograms to determine the conditional cumulative distribution function through indicator cokriging allows an improved result as compared to the current practice of computing the conditional probabilities by kriging each indicator independently. This approach can be easily integrated in simulation, with the improved reproduction of relationships between categories.

Acknowledgements

The authors thank the industry sponsors of the Centre for Computational Geostatistics at the University of Alberta for financial support. Dr. Deutsch holds the Alberta Chamber of Resources Industry Chair in Mining Engineering; the supporting companies are gratefully acknowledged. Dr. Ortiz would like to thank Codelco Chile for sponsoring his research at the University of Chile. NSERC is also thanked for the industrial postgraduate scholarship that Mr. Neufeld holds; that scholarship is due to special Placer Dome support.

References

- Alabert, F.G., 1987, *Stochastic Imaging of Spatial Distributions Using Hard and Soft Information*, Master's Thesis, Stanford University, Stanford, CA.
- Deutsch, C.V. 2002, *Geostatistical Reservoir Modeling*. Oxford University Press, New York.
- Deutsch, C.V. and Journel, A.G., 1998. *GSLIB: Geostatistical Software Library: and User's Guide*. Oxford University Press, New York, 2nd Ed.
- Goovaerts, P. 1997, *Geostatistics for Natural Resources Evaluation*. Oxford University Press, New York.
- Goovaerts, P., 1994, *Comparative Performance of Indicator Algorithms for Modeling Conditional Probability Distribution Functions*, Mathematical Geology, Vol. 26, No. 3, pp 385-410.
- Gómez-Hernández, J.J. and Srivastava, R.M., 1990, *ISIM3D: An ANSI-C Three Dimensional Multiple Indicator Conditional Simulation Program*, Computers & Geosciences, Vol. 16, No. 4, pp. 395-410.
- Journel, A.G., 1893. *Nonparametric estimation of spatial distribution*. Mathematical Geology, Vol. 15, No. 3, pp. 445-468.
- Journel, A.G. and Isaaks, E.H., 1984. *Conditional indicator simulation: Application to a Saskatchewan uranium deposit.* Mathematical Geology, Vol. 16, No. 7, pp. 685-718.
- Neufeld, C.T. and Deutsch, C.V., 2004. *Developments in semiautomatic variogram fitting.* Centre for Computation Geostatistics, Report 6.

Type	Onmidirectional
Number of lags	6
Lag separation distance	0.1 km
Lag tolerance	0.05 km

Table 1: Parameters for variogram and cross-variograms calculation.

	Forest	Pasture	Meadow	Tillage
Forest	0.1116	-0.0271	-0.0820	-0.0024
Pasture	-0.0271	0.1670	-0.1360	-0.0040
Meadow	-0.0820	-0.1360	0.2301	-0.0120
Tillage	-0.0024	-0.0040	-0.0120	0.0190

Table 2: Sills of the direct (diagonal) and cross-variograms (off-diagonal) to ensure a valid initial model for `varfit_ind_cat`.

Check	Determinant
Submatrix (1:1,1:1): 0.1116	0.1116
Submatrix (1:2,1:2): 0.1116, -0.0271; -0.0271, 0.1670	0.0179
Submatrix (1:3,1:3): 0.1116, -0.0271, -0.0820; -0.0271, 0.1670, -0.1360; -0.0820, -0.1360, 0.2301	0.0003
Full Matrix (1:4,1:4): 0.1116, -0.0271, -0.0820, -0.0024; -0.0271, 0.1670, -0.1360, -0.0040; -0.0820, -0.1360, 0.2301, -0.0120; -0.0024, -0.0040, -0.0120, 0.0190	2.3×10^{-7}

Table 3: Positive definiteness check for the initial sills.

		Structure1: Spherical		Structure 2: Exponential	
Category	**Nugget**	**Sill**	**Range**	**Sill**	**Range**
Forest	0.0001	0.041	0.352	0.070	0.618
Forest-Pasture	0.000	-0.009	0.352	-0.018	0.618
Forest-Meadow	0.000	-0.039	0.352	-0.043	0.618
Forest-Tillage	0.000	0.000	0.352	-0.002	0.618
Pasture	0.001	0.086	0.352	0.080	0.618
Pasture-Meadow	0.000	-0.070	0.352	-0.066	0.618
Pasture-Tillage	0.000	0.000	0.352	-0.004	0.618
Meadow	0.0001	0.119	0.352	0.111	0.618
Meadow-Tillage	0.000	-0.008	0.352	0.001	0.618
Tillage	0.002	0.011	0.352	0.006	0.618

Table 4: Variogram and cross-variogram models.

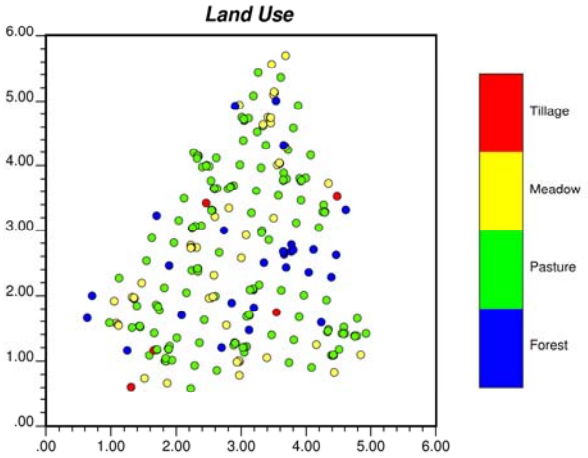


Figure 1: Location map of samples indicating the land use.

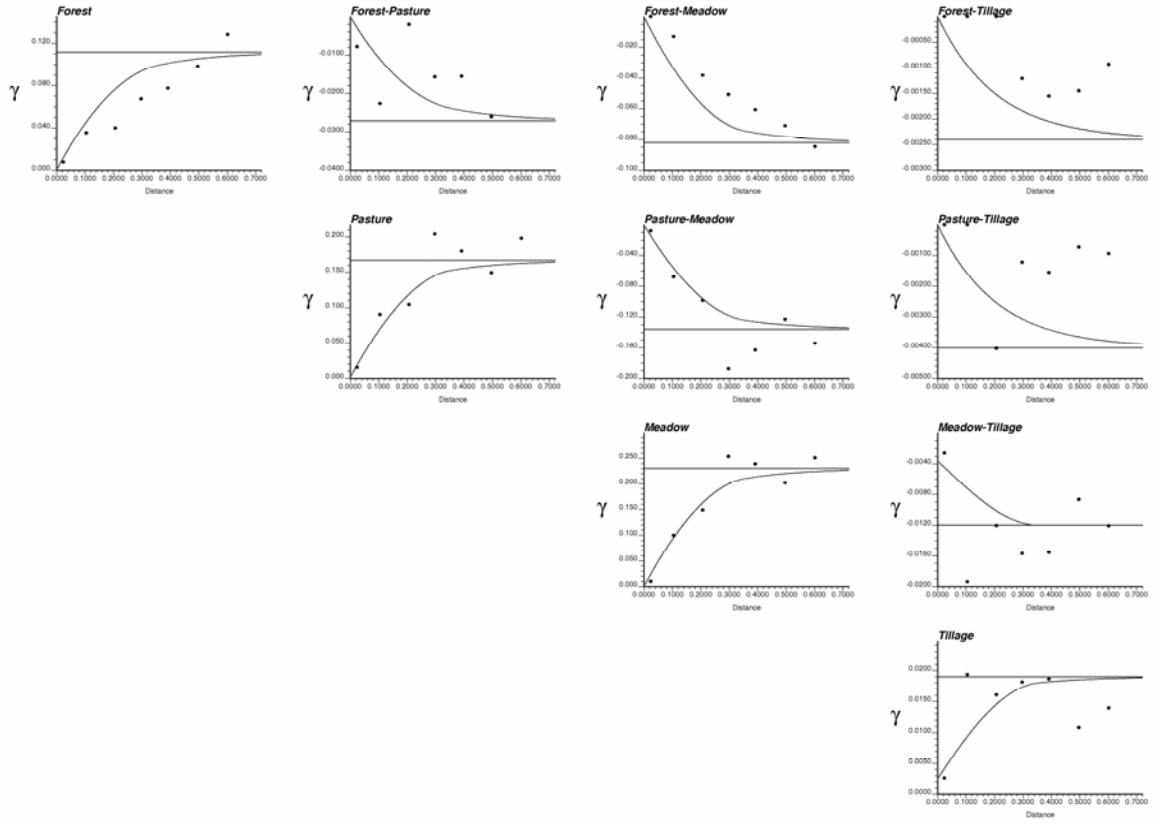


Figure 2: Fitting of indicator variograms and cross-variograms with two nested structures.

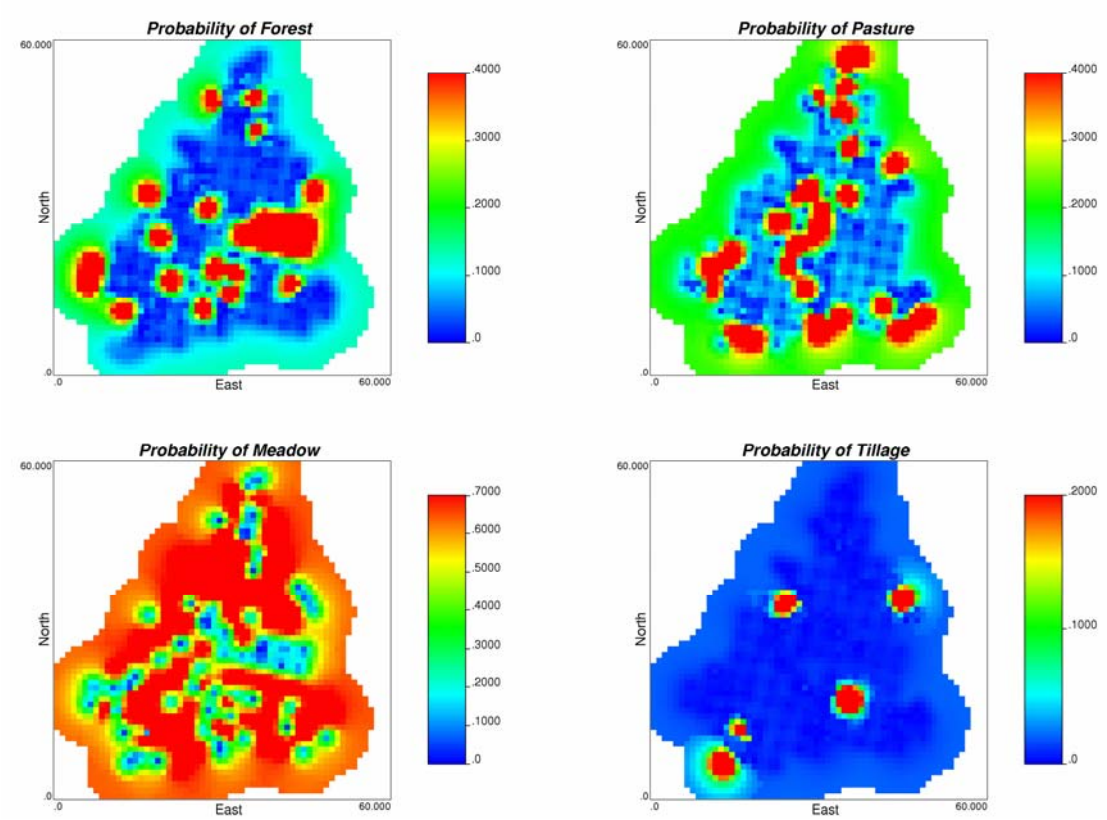


Figure 3: Location maps showing the probability of each land use to prevail at locations in the domain.

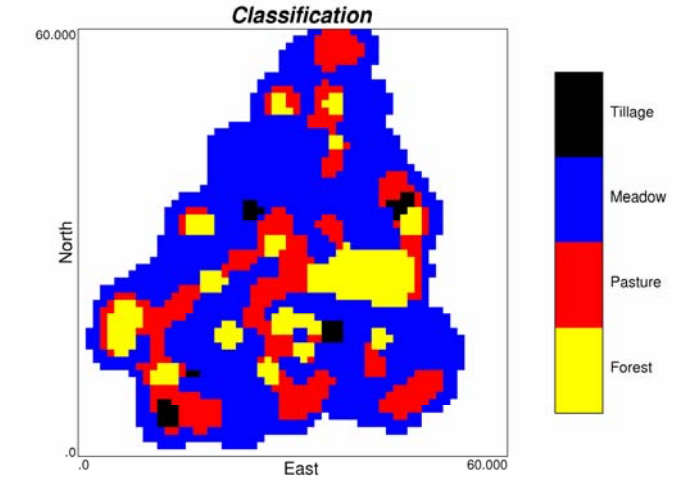


Figure 4: Classification map showing the assigned land use to each location in the domain.