Spatially Weighted Log-likelihoods in a Bayesian Approach to Hazard Mapping

K. G. van den Boogaart[1], R. Tolosana-Delgado[2]

[1]Institut für Mathematik und Informatik, Ernst-Moritz-Arndt-Universität Greifswald, D-17487 Greifswald (Germany);
E-mail: boogaart@uni-greifswald.de
[2]Dept. Informàtica i Matemàtica Aplicada, Universitat de Girona, E-17071 Girona (Spain)

1. Abstract

The joint log-likelihood of a normally-distributed vector can be expressed as a linear combination of the log-likelihoods of each marginal variable in that random vector. The coefficients of this linear combination are linked to Fisher information concepts, but also to the kriging weights, when the random vector is a spatial distribution. This allows the exact estimation of the posterior or the maximum-likelihood distribution at an unsampled location conditional on the observed values and the covariance structure when the joint distribution is Gaussian, but offers also a valid approach when the marginal distributions are of any specified type.

2. Introduction

Estimation of probability density functions (pdfs) in geostatistical applications is usually done by a kriging technique, in general a predictor expressed as a linear function of the observed data, usually regarded as model-free. This approach has philosophical (estimating instead of giving the conditional distribution), technical (using average squared errors instead of distances suited for probabilities, and the linearity of the predictor) and practical (negative probabilities, or not summing up to one) limitations, especially in hazard assessment.

Bayesian estimation of pdfs in hazard assessment is contrarily almost-always done by assuming a model of distribution for the sample, which is taken as independent. Then, a prior model for the parameters of the distribution is updated by the likelihood of the sample through Bayes Theorem, and a posterior model for the parameters is obtained. Finally, this posterior easily yields the predictive distribution or quantiles. The keystone which precludes the application of such a procedure in geostatistical problems is the computation of the likelihood assuming an independent sample.

We propose to express the joint log-likelihood of a dependent sample by a weighted linear combination of log-likelihoods of each element in the data set taken independently. The weights are obtained by solving a kriging-like linear system. Assuming a joint Gaussian model, the log-likelihood is exactly reproduced, whereas for other models we obtain a first-order approximation. The weights

are nevertheless not used in a kriging estimator but to obtain directly a probability density function, conditional on the prior knowledge, the observations and the assumed marginal model for the data set.

This preliminary approach always yields valid conditional distributions, which do not present negative probabilities or order-relation violations. Furthermore, the calculated conditional pdf is an approximation and not only an estimation, as in indicator or disjunctive kriging. However, we still lack an assessment on the goodness of this (first-order) approximation, a rigorous treatment for general models, and algorithms to estimate the necessary conditional metric covariances.

3. A Closer Look to the Problem

We are considering the geostatistical prediction of probabilities at unobserved points from observed observations at different locations. This estimation of conditional probabilities is classically treated by indicator kriging or disjunctive kriging. However both approaches have the problem: they can yield negative or inconsistent probabilities. This can not be solved in a linear framework and is in our opinion mainly due to the dishonouring of the geometry of probabilities, where 0.2 is much more similar to 0.1 than 10^{-5} to 10^{-10}, and especially than 0.01 to 0 or much worse to -0.09. The whole approach of using mean squared errors for probabilities ignores that a zero probability is a very radical assertion when quantifying hazard. This problem is very similar to the discussion on the geometry in the simplex (Aitchison, 1986) since a probability can be seen as an element in a simplex of two parts.

There might be situations where such a relative reasoning is regarded as inadequate (e.g., integrated probabilities—like the probability of exceeding a small threshold over a block—, or the proportion of blocks sent to the mill) or as unimportant when only small estimation variation is involved. However when trying to estimate a *consistent* conditional probability distribution (taking care of constraints and reasonable extreme probabilities), this log-like scale is the first step towards a new approach.

The following idea in that line might be: let us transform the probability to log-odds and do kriging on them. However, trying to implement it, immediately reveals a second theoretical misconception in the indicator approach for probabilities. The idea of indicators is to interpolate a 0-1 field and not to estimate a conditional probability. And this is exactly why it works so well for ore block approach mentioned above, where we ask for an expected portion of good ore. Conditional probabilities do not have a true value independent of observation. The conditional probability changes with every new observation. This is really different from the classical interpolation problem of kriging, where the value at the unobserved location stays the same regardless what we observe. If we had the full multivariate distributional model expressed by a joint density $f(Z(x_0), Z(x_1),..., Z(x_n))$ we could calculate the conditional distribution by the total probability law:

$$f(Z(x_0) \mid Z(x_1),..., Z(x_n)) = \frac{f(Z(x_0), Z(x_1),..., Z(x_n))}{\int f(Z(x_0), Z(x_1),..., Z(x_n)) dZ(x_0)}$$

Here conditional probabilities are calculated rather then estimated. However it is practically impossible to estimate the whole joint probability from a set of observations at some locations, when the joint probability distribution is unknown. In this case, the choice of a model of distribution is unavoidable. Note that Bayesian-Maximum Entropy methods (Christakos, 1990) select the joint distribution with highest Entropy among those satisfying some constraints. Our approach will be to simply choose bivariate models and combine them in an adequate way.

4. Bayes Updates and Log Likelihoods

The general idea of disjunctive kriging is to reduce the knowledge needed to give an optimal predictor to bivariate distributions by restricting the class of allowable estimators to sums of functions depending on one of the observations only. For the calculation of conditional distributions this seems straightforward by the Bayesian theorem (Leonard and Hsu, 1999):

$$f(Z(x_0)\,|\,Z(x_1)=z_1,...,Z(x_n)=z_n)=\frac{f_0(Z(x_0))\prod_{i=1}^{n} f_i(Z(x_i)\,|\,Z(x_0))}{\int f_0(Z(x_0))\prod_{i=1}^{n} f_i(Z(x_i)\,|\,Z(x_0))dx_0}$$

Taking logs and removing the closing constant this reads in terms of loglikelihoods

$$l(Z(x_0)\,|\,Z(x_1)=z_1,...,Z(x_n)=z_n)=l_0(Z(x_0))+\sum_{i=1}^{n} l_i(Z(x_i)\,|\,Z(x_0))+c$$

showing a linear structure. However this is only valid, when the $Z(x_1),...,Z(x_n)$ are independent conditional to $Z(x_1)$. If not the sequential updating procedure becomes much more complicated with dependent observations:

$$f(Z(x_0)\,|\,Z(x_1)=z_1,...,Z(x_n)=z_n)=\frac{f_0(Z(x_0))\prod_{i=1}^{n} f_i(Z(x_i)\,|\,Z(x_0),...,Z(x_{i-1}))}{\int f_0(Z(x_0))\prod_{i=1}^{n} f_i(Z(x_i)\,|\,Z(x_0),...,Z(x_{i-1}))dx_0}$$

and needs again the whole joint distribution encoded in conditionals. Taking logs reveals the difference:

$$l(Z(x_0)\,|\,Z(x_1)=z_1,...,Z(x_n)=z_n)=l_0(Z(x_0))+\sum_{i=1}^{n} l_i(Z(x_i)\,|\,Z(x_0),...,Z(x_{i-1}))+c$$

It is still a linear combination of information in the log likelihoods, but each likelihood depend on the whole set of observations (thus there must be a removal of repeated information, e.g. if $Z(x_3)$ is independent of $Z(x_0)$ conditionally on $Z(x_2)$, then $l_i(Z(x_3)\,|\,Z(x_0),...,Z(x_2))$ is constant seen as a

function of $Z(x_0)$ and does not change anything). In kriging, a similar effect is known as screening.

Our suggestion is to replace $l_i(Z(x_i)|Z(x_0),...,Z(x_{i-1}))$ with $l_i(Z(x_i)|Z(x_0))$, but to remove the redundant information as simply as possible, using only a minimum of information on the joint probability law. We expect that a linear unbiased approach would only need information on the *covariance of likelihoods*, and would successfully remove the repeated information about the mean. Therefore, we look for a set of weights λ_{ij} such that the following expression is approximate in some sense:

$$l(Z(x_i)|Z(x_0)=z_1,...,Z(x_{i-1})=z_{i-1}) \approx \sum_{j=1}^{n} \lambda_{ij} l_j(Z(x_j)|Z(x_0)) + c_i$$

Here each λ_{ij} can be seen as a full tensor operator, as a diagonal operator or as a scalar. In this paper, we consider only the last option, thus $\lambda_{ij} \in \Re$. Whatever procedure we finally use to estimate λ_{ij}, we will end with a simpler form of an estimated or approximated version $\hat{l}(Z(x_0)|Z(x_1)=z_1,...,Z(x_n)=z_n)$ of $l(Z(x_0)|Z(x_1)=z_1,...,Z(x_n)=z_n)$, in the form of a weighted sum of univariate likelihoods.

$$\hat{l}(Z(x_0)|Z(x_1)=z_1,...,Z(x_n)=z_n) = l_0(Z(x_0)) + \sum_{i=1}^{n} \lambda_i l_i(Z(x_i)|Z(x_0)) + c$$

for some new weights λ_i.

5. Finding Weights for the Joint Loglikelihood

A first consideration to find the optimal weights is that $l(Z(x_0)|Z(x_1)=z_1,...,Z(x_n)=z_n)$ can not be predicted from the other likelihoods, since otherwise it would be possible to use that information to get a better update. In a linear framework we can only consider linear dependency, which is expressed by correlation. Thus the weights should make the $\hat{l}(Z(x_0)|Z(x_1)=z_1,...,Z(x_n)=z_n)$ pairwise uncorrelated. Since our loglikelihoods are functions and we are restricted to scalar weights, we need a scalar measure of correlation. This can be given by a metric covariance in a Hilbert space of the loglikelihoods. We propose to use $L^2(\mu)$ for some μ with the scalar product $(f,g) = \int_{\Re} f(z)g(z)d\mu(z)$, which means that the metric covariance

$\mathrm{cov}(f,g) = E[(f - E[f], g - E[g])]$ will be the average covariance over all values of z with respect to some measure μ. A good choice for μ could be the marginal distribution of z.. The important object here is the expected conditional metric covariance of the likelihoods:

$$k_{ij} \underset{def}{=} E\left[\left(l(Z(x_i)\,|\,Z(x_0)) - E[l(Z(x_i)\,|\,Z(x_0))\,|\,Z(x_0)\right], l\left(Z\left(x_j\right)\,|\,Z(x_0)\right) - E\left[l\left(Z\left(x_j\right)\,|\,Z(x_0)\right)\,|\,Z(x_0)\right]\right)\right]$$

This last definition is – for any fixed choice of the likelihood – consistent with the likelihood principle (Leonard and Hsu 1999, Robert 1996) and independent of the measure underlying the likelihood. To remove redundancy we use a simple idea now: Estimate the mean of the loglikehoods as good as possible and then multiply the means to get an as good as possible estimate of the sum, which exploits the information optimally and thus avoids double use of redundant information. An optimal weighting scheme for the mean is given by the inverse of the covariance matrix and thus from generalised least squares the weights λ would be given by $\mathbf{K}\lambda = \alpha\mathbf{1}$ for some α, where $\mathbf{1} = (11...1)^t$ denotes a vector containing ones only. Now a curious property of likehoods is that more variance means more information and thus the mean should honor things with big variances proportional to it: $\mathbf{K}\lambda \sim \alpha\mathbf{S1}$. Furthermore we should multiply with some constant afterwards, because we are not really interested in the mean, but in a sum. Looking at the special case of uncorrelated likelihoods, we see that $\alpha = 1$ would be a good choice resulting in a formula $\lambda = \mathbf{K}^{-1}\mathbf{S1}$. This can be interpreted in as following: More covariance means more *joint* information. And this redundant information means less information in total and thus down weighting. On the other hand: More direct variance means more information and this exactly counteracts to the down weighting. The $\mathbf{1}$ means taking a sum.

Clearly this motivation is extremly simplistic and more theory is needed here. However these more advanced arguments need deep considerations on Fisher information theory to quantify the virtual and the true information in likelihoods and a more developed theory of the connection of the likelihood spaces to these information concepts. These arguments are not fully developed yet and would clearly exceed the scope of this extended abstract. We leave it to a research paper to be published later. However the results of the simplistic idea work well in simple test cases as shown in the next section.

6. Comparision with Simple Kriging

In the situation of simple kriging (Cressie, 1993) with a Gaussian random field the conditional distribution can be computed directly from the conditional expectation and the conditional variance. It can be shown that the approximate solution given in section 3 gives the same result as the exact solution in this case. The calculation goes as follows:
With a multivariate normal model:

$$\begin{pmatrix} x_0 \\ \mathbf{x} \end{pmatrix} \sim N\left(0, \begin{pmatrix} 1 & \mathbf{c}^t \\ \mathbf{c} & \mathbf{C} \end{pmatrix}\right)$$

and $\mathbf{F} = (f_{ij})_{ij} \underset{def}{=} \mathbf{C} - \mathbf{c}\mathbf{c}^t$, $\mathbf{c} = (c_i)$, $\mathbf{C} = \left(c_{ij}\right)_{ij}$ the conditional distributions for the vector $\mathbf{x} = (x_i)$

and individual x_i are given by (Cressie 1993, p. 110)

$$P(\mathbf{x} \mid x_0) = N(\mathbf{c}x_0, \mathbf{F})$$

$$P(x_i \mid x_0) = N(c_i x_0,\ f_{ii} = c_{ii} - c_i c_i)$$

Therefore the individual loglikelihoods have the form:

$$l_i(x_i \mid x_0) =_a (x_i - \tfrac{1}{2} c_i x_0) f_{ii}^{-1} c_i x_0 + const$$

where the small *a* at the equal means equal up to an additive constant. The joint loglikelihood has the form

$$l(\mathbf{x}_i \mid x_0) =_a (\mathbf{x} - \tfrac{1}{2}\mathbf{c}x_0)\mathbf{F}^{-1}\mathbf{c}x_0 + const$$

Using the linear equality

$$l(\mathbf{x} \mid x_0) = \sum_{i=1}^{n} \lambda_i l_i(x_i \mid x_0) + const$$

we can see by comparison of coefficients (and using the notation) $\mathbf{B} = (\delta_{ij} f_{ii}^{-1} c_i)_{ij}$, the weights are given by the equation

$$\mathbf{B}\lambda = \mathbf{F}^{-1}\mathbf{c}$$

On the other hand the conditional metric covariation is given by

$$k_{ij} = c_i f_{ii}^{-1} f_{ij} f_{jj}^{-1} c_j$$

And thus $\mathbf{K} = \mathbf{BFB}$ and $\mathbf{S} = (\delta_{ij} c_i^2 f_{ii}^{-1})_{ij}$ we can check whether the weights are equal in both approaches by checking the equation $\mathbf{B}\lambda = \mathbf{F}^{-1}\mathbf{c}$ for $\lambda = \mathbf{C}^{-1}\mathbf{S1}$. We get: $\mathbf{BK}^{-1}\mathbf{S1} \stackrel{?}{=} \mathbf{F}^{-1}\mathbf{c}$

$\Leftrightarrow \mathbf{BB}^{-1}\mathbf{F}^{-1}\mathbf{B}^{-1}\mathbf{S1} \stackrel{?}{=} \mathbf{F}^{-1}\mathbf{c} \Leftrightarrow \mathbf{B}^{-1}\mathbf{S1} \stackrel{?}{=} \mathbf{c} \Leftrightarrow f_{ii} c_i^{-1} c_i^2 f_i^{-1} = c_i \Leftrightarrow c_i = c_i$

And thus the approximate likelihood approach yields the same result as the exact solution. The results in the likelihood approach do not change when data is transformed and thus it will yield the exact solution in any transformed Gaussian, i.e. trans-Gaussian (Cressie 1993, p.137) field too. However the whole construction did not rely on a Gaussian assumption, just like kriging does not rely on the Gaussian assumption. Clearly with non-(trans)-Gaussian fields the approximation of the conditional likelihoods will not be perfect anymore and we will lose something, as we lose optimality when kriging is based on second order arguments only and not on a Gaussian assumption.

7. Conclusions

We present a first approximation to the calculation of probability distributions at unsampled locations conditional on observed data by using linear combinations of log-likelihoods of the observations

conditional on the value which is sought. Parallelisms may be drawn between our approach and disjunctive kriging, regarding the optimal use of bivariate distributions. The approach is fully justified when assuming a Gaussian random field (or a transformation of it), but is still a valid approach for other models. However, its practical implementation in these cases still lacks methods and models to handle the spatial covariance between the likelihoods at two locations conditionals on a third one.

8. Acknowledgements

This work was elaborated during a research stage funded by the projects A/04/33586 from the German academic exchange office (DAAD) and 2004-BE-00147 from the Catalan universitary grant agency (AGAUR), and by a Student Grant of the International Association for Mathematical Geology.

9. References

Aitchison, J.,1986. The Statistical Analysis of Compositional Data, Monographs on Statistics and Applied Probability. Chapman Hall Ltd., London (UK).

Christakos, G. 1990. A bayesian/maximum entropy view to the spatial estimation problem. Math. Geol., 22(7), 763–777.

Cressie, N. A. C., 1993. Statistics for spatial data, Wiley Series in Probability and Mathematical Statistics, John Wiley and Sons

Journel, A., 1983. Nonparametric estimation of spatial distributions, Math. Geol. 15(3), 445-468.

Leonard, T., J. S. J. Hsu, 1999. Bayesian Methods, An Analysis for Statisticians and Interdisciplinary researchers, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Robert, C. P. 1996. The Bayesian Choice, a Decision Theoretic Motivation, Springer.