Simplicial indicator kriging

Pawlowsky-Glahn, V.[1], R. Tolosana-Delgado[1], J.J. Egozcue[2]

[1] Univ. de Girona; Dep. de Inf. y Matemática Aplicada; Campus Montilivi, P4; E-17071 Girona (Spain); vera.pawlowsky@udg.es; raimon.tolosana@udg.es
[2] Univ. Politécnica de Cataluña, Dep. de Matemática Aplicada III; Jordi Girona 1-3; E-08034 Barcelona (Spain); juan.jose.egozcue@upc.edu

Abstract: Indicator kriging (IK) is a spatial interpolation technique devised for estimating a conditional cumulative distribution function (ccdf) at an unsampled location. The result is a discrete approximation to the ccdf, and its corresponding probability density function can be viewed as a composition. This suggests a compositional approach to IK, which by construction avoids all the standard drawbacks, like estimates outside the $(0, 1)$ interval or order-relation problems.

Keywords: Aitchison geometry; log-ratio approach; multinomial distribution.

1 Introduction

Indicator kriging (IK) is a geostatistical technique used to approximate the conditional cumulative distribution function (ccdf) at each point of a grid based on the correlation structure of indicator transformed data points (Journel, 1983). The major drawback of IK is that it might yield impossible estimates, such as negative probabilities, probabilities larger than one, or a non-monotonic ccdf. Several methods have been developed to correct the order relations violation, but none of them reconsiders the underlying hypothesis of the model itself, namely that probabilities are real numbers and obey the rules of real space as a Euclidean space. Here we discuss an approach based on the fact that relative frequencies and probabilities can be viewed as compositions with sample space a D part simplex, $\mathcal{S}^D$. This allows the application of interpolation techniques devised for compositional data (Pawlowsky-Glahn and Olea, 2004) to approximate the ccdf at an unsampled location by means of an estimating function satisfying all the required constraints. An extensive presentation of the mathematical foundations and of the method can be found in Tolosana-Delgado (2005).

2 Methodology

The new procedure relies on the fact that the D-part simplex has a Euclidean space structure different from the usual one in real space (Billheimer et al., 2001; Pawlowsky-Glahn and Egozcue, 2001), and comprises the following steps:

1) Select $D-1$ thresholds: As usual in IK, for a regionalised variable $Z(x)$ with range $A = [a_0, a_D)$, a set of thresholds $\{a_1, ..., a_{D-1}\}$ is defined, which leads to a partition $A = \bigcup_{i=1}^{D} A_i$, with $A_i = [a_{i-1}, a_i)$.

2) Express each observation as a D-multinomial probability vector: At each sampling point the partition defines a categorical random vector $\mathbf{J}(x) = (j_1(x), ..., j_D(x))$, with $j_i(x) = 1$ if $Z(x) \in A_i$ and 0 otherwise. $\mathbf{J}(x)$ follows a multinomial distribution with probabilities $\mathrm{P}[j_i(x) = 1] = \mathrm{P}[Z(x) \in A_i]$. Interpreting the vector of probabilities as a composition in $\mathcal{S}^D$, one can estimate it from a single observation of $\mathbf{J}(x)$, either using Bayesian methods (Tolosana-Delgado, 2005) or a zero substitution technique (Martín-Fernández et al., 2000). If $j_k(x) = 1$, the latter leads to $p_k(x) = 1 - \alpha$ and $p_i(x) = \alpha/(D-1)$ for $i \neq k$; it corresponds to a prior model which treats all categories equal. α is interpreted as the probability of error in the observation. This approach does not take into account the order of the categories, as the underlying model is a multinomial one.

3) Represent the vectors of estimated probabilities by their coordinates with respect to an orthonormal basis in the simplex: $\mathbf{p}(x) = (p_1(x), ..., p_D(x))$ is an element of $\mathcal{S}^D$, since the parts are strictly positive and sum up to one. The Euclidean space structure of $\mathcal{S}^D$ allows to represent compositions as coordinates with respect to an orthonormal basis (Egozcue et al., 2003). One standard option (Tolosana-Delgado, 2005) is to compute $\mathbf{c}(x) = \mathbf{\Psi} \lg \mathbf{p}(x)$, using the $(D-1) \times D$ Helmert matrix

$$\mathbf{\Psi} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{D(D-1)}} & \frac{1}{\sqrt{D(D-1)}} & \frac{1}{\sqrt{D(D-1)}} & \cdots & \frac{-(D-1)}{\sqrt{D(D-1)}} \end{bmatrix}.$$

4) Use standard variography and co-kriging techniques to obtain estimates of coordinates at unsampled locations: The vector of coordinates $\mathbf{c}(x)$ has $D-1$ unbounded real components, suitable to be treated with any existing software.

5) Represent the estimated coordinates as compositions: Estimates $\hat{\mathbf{c}}(x)$ are expressed as compositions using $\hat{\mathbf{p}}(x) = \mathcal{C}(\exp[\mathbf{\Psi} \cdot \hat{\mathbf{c}}(x)])$, where $\mathcal{C}(\cdot)$ represents the closure operation, which divides all components by their total sum, thus forcing the result to sum up to one. The properties of the exponential and the closure operation guarantee that $\hat{\mathbf{p}}(x)$ will always be valid multinomial probabilities, from which the desired ccdf is obtained.

3 Properties

When working in a Euclidean space, basic linear algebra guarantees that properties satisfied by estimators in coordinates are automatically satisfied by the same estimators when represented in any other way, although one has to be aware that not only the vector space operations and the inner product change, but also the reference measure (Eaton, 1983). Thus, the fact that co-kriging estimators are BLU in real space (the space of coordinates) guarantees that the estimators expressed as compositions are BLU in the simplex with respect to the Aitchison geometry (Pawlowsky-Glahn and Egozcue, 2002) and the associated measure (Pawlowsky-Glahn, 2003). Also, assuming the random function $\mathbf{p}(x)$ to have a joint normal distribution on the simplex, simple kriging prediction of the coordinates $\mathbf{c}(x)$ and their error variance-covariance matrix give the parameters of the true distribution of $\mathbf{p}(x_0)$ conditional on the observed data set, which is a normal distribution on $\mathcal{S}^D$ (Mateu-Figueras et al., 2003). Moreover, ordinary or universal kriging represent valid approximations to this conditional distribution up to the same extent they are for a conventional Gaussian random function, and it can be shown that the estimator and the conditional distribution do not depend on the chosen basis in the simplex $\mathcal{S}^D$.

4 Discussion

The proposed technique can be applied to interpolate discrete probability density functions if a certain degree of uncertainty is accepted when estimating the probability distribution at sampled locations. The obtained predictor leads by construction to valid probabilities, which are strictly positive and summing up to one, thus this technique overcomes the main flaws of IK. It is a BLU estimator and has all desirable properties owned by co-kriging estimators, although with respect to a different geometry and a measure different from the Lebesgue one. Also, the fact that the Euclidean structure of $\mathcal{S}^D$ can be extended to a Hilbert space structure for an infinite number of parts (Egozcue et al., 2006), opens a field of further study for the continous case. The question how results compare to estimates obtained using standard IK techniques has no answer at the present moment. Each technique relies on a different geometry and a different measure, and is thus *best* in a different sense. The real question is which model relies on underlying hypothesis that make sense to our understanding of probability, and which model leads to consistent results. For us the answer is simplicial indicator kriging.

Acknowledgments: This research has been supported by the *Dirección General de Enseñanza Superior del Ministerio de Educación y Cultura* (BFM2003-05640).

References

Billheimer, D., P. Guttorp, and W. Fagan (2001). Statistical interpretation of species composition. *Journal of the American Statistical Association* **96**(456), 1205–1214.

Eaton, M. L. (1983). *Multivariate Statistics. A Vector Space Approach.* John Wiley & Sons.

Egozcue, J. J., J. L. Díaz-Barrero, and V. Pawlowsky-Glahn (2006). Hilbert space of probability density functions based on Aitchison geometry. *Acta Mathematica Sinica* **22**.

Egozcue, J. J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology* **35**(3), 279–300.

Journel, A. G. (1983). Nonparametric estimation of spatial distributions. *Mathematical Geology* **15**(3), 445–468.

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (2000). Zero replacement in compositional data sets. In: H. Kiers, J. Rasson, P. Groenen, and M. Shader (Eds.), *Data Analysis, Classification, and Related Methods*, pp. 155–160. Springer-Verlag, Berlin (D), 428 p.

Mateu-Figueras, G., V. Pawlowsky-Glahn, and C. Barceló-Vidal (2003). Distributions on the simplex. In: Thió-Henestrosa, S. and J. A. Martín-Fernández (Eds.), *Compositional Data Analysis Workshop – CoDaWork'03.* Universitat de Girona, ISBN 84-8458-111-X, http://ima.udg.es/Activitats/CoDaWork03/.

Pawlowsky-Glahn, V. (2003). Statistical modelling on coordinates. In: Thió-Henestrosa, S. and J. A. Martín-Fernández (Eds.), *Compositional Data Analysis Workshop – CoDaWork'03.* Universitat de Girona, ISBN 84-8458-111-X, http://ima.udg.es/Activitats/CoDaWork03/.

Pawlowsky-Glahn, V. and J. J. Egozcue (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment (SERRA)* **15**(5), 384–398.

Pawlowsky-Glahn, V. and J. J. Egozcue (2002). BLU estimators and compositional data. *Mathematical Geology* **34**(3), 259–274.

Pawlowsky-Glahn, V. and R. A. Olea (2004). *Geostatistical Analysis of Compositional Data.* Oxford University Press.

Tolosana-Delgado, R. (2005). *Geostatistics for constrained variables: positive data, compositions and probabilities.* Ph. D. thesis, Universitat de Girona, Girona (E). 198 p.