

POTENTIAL MAPPING FROM GEOCHEMICAL SURVEYS USING A COX PROCESS

Raimon Tolosana-Delgado ¹, K. Gerald van den Boogaart ^{1,2} & Helmut Schaeben ³

¹ *Helmholtz Institute Freiberg for Resource Technology, Halsbruecker Str. 34, D-09599 Freiberg (Germany), r.tolosana@hzdr.de*

² *Institute for Stochastics, Technical University Bergakademie Freiberg, Pruefer Str. 9, D-09599 Freiberg (Germany), boogaart@hzdr.de*

³ *Department of Geophysics and Geoinformatics, Technical University Bergakademie Freiberg, Gustav-Zeuner Str. 12, D-09599 Freiberg (Germany), Helmut.Schaeben@geophysik.tu-freiberg.de*

Abstract. Punctual occurrence phenomena are often modeled as Poisson point processes, sometimes with an inhomogeneous, unknown intensity, that is desired to be estimated from some covariables. Within this setting, we study the case in which the covariables form a regionalized geochemical composition of stream sediments, and the known punctual occurrences are existing mineral deposits, not collocated with the available explanatory data. This is modeled by assuming a two-layer stochastic process, where the observed Poisson log-intensity is taken as a balanced log-linear function of the geochemical composition of stream sediments, which coefficients must be estimated. Estimation is possible through a pseudolikelihood device based on generalized loglinear models, though the result intensity function appears to be valid only up to arbitrary scaling and addition of constants.

Keywords. Word 1, word 2, ...

1 Introduction

Potential mapping is a loose term used to describe any technique that aims at detecting location in space where a certain phenomenon is likely to occur, by using some covariables available throughout in space and a model linking these variables with the phenomenon occurrence. Poisson point processes are, in principle, the canonical choice to model occurrences of dimensionless phenomena in space/time, thus it seems natural to approach potential mapping from the perspective of logistic regression, i.e. where the “potential” to be mapped is the intensity of an inhomogeneous Poisson process on space. This (log)intensity may then be linked through a linear model with all available covariables. These coefficients can be estimated using several state-of-the-art approaches, and here we choose the fast maximum-likelihood estimation of Baddeley and Turner (2000) as a benchmark.

The problem of these direct applications of a logistic regression approach is that the spatial uncertainty of the covariables is ignored, e.g. the likelihood to be maximized does

not take into account the fact that some of the covariables might be interpolated (e.g. with kriging) or be spatially non-representative (i.e., related to the concept of extension variance). This paper aims at presenting a method of potential mapping within the framework of inhomogeneous Poisson processes where the covariables are not available everywhere and might have important uncertainties derived from their spatial interpolation. In particular, our motivation is to map the occurrence of mineral deposits using an independent data base of geochemical compositions of stream sediments, thus the proper way of interpolating and describing the spatial uncertainty of compositional data will also be discussed.

In our example, the covariables correspond to a stream sediment geochemical survey, while the occurrences are some historical mines and mine tilings from the Grazer Palaeozoikum (Austria), but the same setting can be used for any environmental compositional survey (rock, soil, water, sediment or moss geochemistry, but also hyperspectral remote sensing) which is desired to be used as a proxy for a punctual phenomena (trees, non-contagious diseases, pollution events, etc.).

Briefly, the solution proposed is based on a Cox process, i.e. a two-layer random process. In the deep layer, the log-ratio transformed composition is considered a multivariate Gaussian random function: this is characterized as usual by estimating its mean value and a variographic structure. In the second layer, the Poisson log-intensity is linked to the composition through a log-ratio linear function, of unknown coefficients.

2 Model

Let \mathcal{D} be the geographical domain of interest, and $\mathcal{X} = \{x_1, x_2, \dots, x_Q \in \mathcal{D}\}$ the known occurrence locations of the phenomenon under study. At a set of independent locations $x_{Q+1}, x_{Q+2}, \dots, x_{Q+N} \in \mathcal{D}$ one has also available a composition $\mathbf{z}_n = \mathbf{z}(x_n)$, $n = Q + 1, \dots, Q + N$, i.e. a vector of D positive variables informing of the relative importance of D components. Regardless of their spatial dependence, compositions add to a constant κ , or they can be considered so, without loss of generality, by defining the last variable as a filler, $z_D = \kappa - z_1 - z_2 - \dots - z_{D-1}$; or by reclosing the composition to constant sum, $\mathbf{z}' = [\kappa / (\mathbf{1}^t \cdot \mathbf{z})] \mathbf{z}$, if the filler is meaningless or deemed a sampling artifact.

In the example here presented, \mathcal{D} is the Grazer Palaeozoikum (a region north of the city of Graz, Austria, where Palaeozoic rocks can be found at the surface), \mathcal{X} represents the locations of historical mining sites in this area, $x_{Q+1}, x_{Q+2}, \dots, x_{Q+N} \in \mathcal{D}$ are the locations of samples of stream sediments, and each $\mathbf{z}_n = \mathbf{z}(x_n)$ is the geochemical composition of a sediment sample. This data set and the deposit locations were presented by Weber and Davis (1990).

Let us assume \mathcal{X} observation of an inhomogeneous Poisson point process with intensity field $\lambda(x)$ and without interaction (i.e. occurrences are independent from each other, given the intensity field). Thus, the likelihood of this observed pattern is (Baddeley and Turner,

2000)

$$L(\mathcal{X}|\lambda(x)) \propto \prod_{q=1}^Q \lambda(x_q) \exp \left[- \int_{\mathcal{D}} \lambda(x') dx' \right]. \quad (1)$$

In the context of generalized linear models, one would then assume the log-intensity to be a linear function of the available covariables. However, given the compositional nature of our covariables (Aitchison, 1986; Pawlowsky-Glahn and Egozcue, 2002; van den Boogaart and Tolosana-Delgado, 2013), we must assume that to be an affine logcontrast,

$$\log \lambda(x) = \beta_0 + \sum_{i=1}^D \beta_i \log z_i(x) = \beta_0 + \boldsymbol{\beta}^t \cdot \text{clr}(\mathbf{z}(x)), \quad \mathbf{1}^t \cdot \boldsymbol{\beta} = 0, \quad (2)$$

with $\boldsymbol{\beta} = [\beta_1, \beta_2, \dots, \beta_D]$ and $\text{clr}(\mathbf{z}) = \log(\mathbf{z}) - (\mathbf{1}^t \cdot \log(\mathbf{z})/D)\mathbf{1}$ the centered logratio transformation (clr, Aitchison, 1986). Note that the clr-transformed composition has one redundant component, which might introduce singularity problems in some estimation procedures. For this reason, it is common in the compositional literature to reparametrize both $\boldsymbol{\beta}$ and the clr-scores into vectors of $D - 1$ orthonormal coefficients with a standard Gramm-Schmidt procedure, which defines the so-called isometric logratio transformation (ilr, Egozcue et al., 2003). In this case, we may equivalently write

$$\log \lambda(x) = \beta_0 + \text{ilr}^t(\mathbf{z}(x)) \cdot \boldsymbol{\beta}^*. \quad (3)$$

Such logratio transformations (clr or ilr) have the role to capture the relative and closed character of the scale of compositions, and they are all related through one-to-one linear transformations. Expressions to change between ilr and clr (and in particular linking $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}$) can be found in van den Boogaart and Tolosana-Delgado (2013) or Tolosana-Delgado and van den Boogaart (2014).

Moreover, $\mathbf{z}(x_n)$, thus $\lambda(x)$, is not known everywhere in \mathcal{D} , but assumed a regionalized composition (Pawlowsky-Glahn and Olea, 2004), i.e. a composition-valued random function. As usual in the context of Gaussian processes, we further assume that the (isometric) logratio-transformed composition at any set of N locations form a multivariate normal vector, $\boldsymbol{\zeta} \sim \mathcal{N}^{N(D-1)}(\boldsymbol{\theta})$, where the vector of parameters $\boldsymbol{\theta}$ contains the necessary descriptors of the mean vector and the variographic structure of the random function. For the sake of simplicity, we assume in this contribution that these parameters are known, i.e. a constant mean and some sill and range parameters for a set of direct and cross-variograms.

With these assumptions, we must integrate the uncertainty of the Gaussian process into the Poisson point process, which gives a likelihood for the observed point pattern of

$$\begin{aligned} L(\mathcal{X}|\boldsymbol{\theta}, \beta_0, \boldsymbol{\beta}^*) &\propto \int \prod_{q=1}^Q e^{\beta_0 + \text{ilr}^t(\mathbf{z}(x_q)) \cdot \boldsymbol{\beta}^*} \exp \left[- \int_{\mathcal{D}} e^{\beta_0 + \text{ilr}^t(\mathbf{z}(x')) \cdot \boldsymbol{\beta}^*} dx' \right] d\Phi(\boldsymbol{\theta}) \\ &\propto \int L(\mathcal{X}|\lambda(x|\beta_0, \boldsymbol{\beta}^*)) d\Phi(\boldsymbol{\theta}) \end{aligned}$$

The goal is finally to obtain $(\beta_0, \boldsymbol{\beta})$ that maximizes this likelihood (if a maximum-likelihood estimator is sought), or to find ways to integrate this likelihood with some prior information on the distribution of $(\beta_0, \boldsymbol{\beta})$ (if a Bayesian estimation is desired). The coefficients of $\boldsymbol{\beta}$ will finally inform on which variables favor the occurrence of the studied phenomenon.

3 Estimation procedure

Baddeley and Turner (2000) present methods of estimation of $(\beta_0, \boldsymbol{\beta})$ that maximize an approximation to Eq. (1) based on quadratures of the integral expression. These make use of standard software for solving weighted generalized linear models (glm). The idea is to complement the point process with dummy points. This can be done, e.g. by splitting the domain in tiles, and taking a dummy point at the center of each tile. Each point is then given a weight w_i equal to its tile area divided by the number of points (real occurrence or dummy) on that tile. Note that the sum of all weights is equal to the area of the domain \mathcal{D} . A weighted Poisson glm with log-link is then fitted with response 0 for the dummy points and $1/w_i$ for the observed occurrences, and input the logratio-transformed composition \mathbf{Z} , which must be interpolated first.

This method “as is” does not take into account the uncertainty derived from the interpolation procedure. However, it can be adapted to the present case by a pseudo Monte Carlo approach. The idea is to obtain a *simulation* of $\mathbf{z}^{(s)}(x_n)$ at all dummy points and observed occurrences, and apply the preceding algorithm using them to obtain estimates $(\hat{\beta}_0^{(s)}, \hat{\boldsymbol{\beta}}^{(s)})$. If we accept that geostatistical simulation offers us a proper description of the uncertainty of $\mathbf{Z}(x)$, and that all simulations are equally probable versions of the “truth”, then the estimates $(\hat{\beta}_0^{(s)}, \hat{\boldsymbol{\beta}}^{(s)})$ obtained for each simulation should be equally probable alternative values of the true parameters.

4 Application

4.1 A simulation example

To show the capabilities of the estimation procedures suggested, a simulation example is presented. A 4-component vector random field $\mathbf{Y}(x) = [Y_1, Y_2, Y_3, Y_4](x)$ on a square domain $\mathcal{D} = (-10, 10) \times (-10, 10)u^2$ is considered, with a variographic structure formed by two spherical structures of range 1u and 5u: the short-range structure is given a unitary sill, while the long-range structure is given sills of 5, 3, 0.5 and 0.5 for the 4 variables respectively. Isotropy and zero correlation between the variables are assumed. Note that in this example the compositional nature of the data is not accounted for, i.e. we assume that $\mathbf{Y}(x) = \text{ilr}(\mathbf{Z}(x))$ of a hypothetical composition $\mathbf{Z}(x)$.

One simulation of this vector-valued RF is obtained at a grid of $0.1 \times 0.1 \text{u}^2$, and stored as “truth” $\mathbf{Y}(x)$. Using this truth and the parameter values $\beta_0 = -4$ and $\boldsymbol{\beta} = [1, 0, 0, 0]$, the truth is transformed into an intensity $\lambda(x)$ by Eq. (2), and an inhomogeneous Poisson process is simulated. For that goal, first the integral $\bar{\lambda}$ of $\lambda(x)$ over \mathcal{D} is computed, and a number of events is drawn from a Poisson distribution with rate $\bar{\lambda}$. The locations \mathcal{X} of these events are drawn from a density on \mathcal{D} obtained normalizing $\lambda(x)$. Finally, a sample of 1000 locations is randomly picked up, and the 4 variables are taken as conditioning data $\{\mathbf{y}(x_1), \mathbf{y}(x_2), \dots, \mathbf{y}(x_{1000})\}$.

Once the simulation of the Poisson process is available, we have proceeded to estimate the link between the Poisson log-intensity and the 4-component vector random field. The pseudo Monte Carlo approach proposed here, based on Baddeley and Turner (2000) method, was applied to two sets of 300 and 3000 simulations, conditioned to $\{\mathbf{y}(x_1), \mathbf{y}(x_2), \dots, \mathbf{y}(x_{1000})\}$. The resulting estimates for the case of 300 samples are shown in Figure 1. These are kernel density estimates of the resulting estimated coefficients for the parameters β_0 and $\boldsymbol{\beta}$ compared with their true values. These comparisons show that, when the true value is different from zero, the pseudo-likelihood approach of Baddeley and Turner (2000) presents significant bias. Even though this is a bad property, it still shows that the method can potentially detect which variables *do not have* an influence on the log-intensity. Thus, the absolute values of the resulting estimated log-intensity should not be given much credibility, but their ranking can be accepted. In a given sense, the result is a potential map, just showing where “high” and “low” potential occur.

4.2 The Grazer Palaeozoikum data set

Geochemistry of stream sediments from the Grazer Palaeozoikum (Weber and Davis, 1990) is available at $N = 601$ locations, where 34 elements were measured. For this contribution, we consider $D = 10$ components, related to the major rock-forming minerals: Al, Ca, Fe, K, Mg, Mn, Na, P, Ti and the filler variable (which is mostly Si, not explicitly available in this data set). No missing values are reported. The possibility to model the mineral potential with this data was already explored by Tolosana-Delgado and van den Boogaart (2014) with a spatially-aware Fisher discriminant rule and with the pseudolikelihood device (i.e. maximizing Eq. 1) of Baddeley and Turner (2000).

Following Tolosana-Delgado and van den Boogaart (2014), the variographic structure of regionalized compositional data sets is best characterized in terms of variation-variograms,

$$t_{ij}(h) = \text{var} \left[\log \frac{z_i(x+h)}{z_j(x+h)} - \log \frac{z_i(x)}{z_j(x)} \right].$$

A linear model of coregionalization (LMC) can be equally fitted to such structural functions, albeit using symmetric sill matrices with zero diagonals. Figure 2 shows the empir-

Table 1: Sill matrices of the variation-variogram model fitted to the geochemical data set: lower triangle of the nugget matrix (\mathbf{T}_0), upper triangle of the short-range structure (\mathbf{T}_1) and lower triangle of the long-range structure (\mathbf{T}_2). Low values of these matrices indicate that the two components tend to change proportionally at that particular scale.

	Al	Ca	Fe	K	Mg	Mn	Na	P	Ti	rest
Al	0.000									
Ca	0.568	0.000								
Fe	0.049	0.414	0.000							
K	0.042	0.636	0.075	0.000						
Mg	0.373	0.326	0.204	0.419	0.000					
Mn	0.180	0.444	0.107	0.139	0.252	0.000				
Na	0.079	0.395	0.110	0.093	0.326	0.243	0.000			
P	0.106	0.293	0.042	0.113	0.212	0.127	0.081	0.000		
Ti	0.082	0.550	0.068	0.114	0.286	0.206	0.155	0.094	0.000	
rest	0.100	0.383	0.076	0.110	0.267	0.103	0.110	0.070	0.115	0.000
\mathbf{T}_0										
\mathbf{T}_1										
	Al	Ca	Fe	K	Mg	Mn	Na	P	Ti	rest
Al	0.000	2.290	0.265	0.216	0.729	0.311	0.635	0.166	0.444	0.157
Ca	0.413	0.000	3.033	1.280	1.346	3.072	3.973	2.107	3.500	1.527
Fe	0.004	0.339	0.000	0.703	0.987	0.022	0.248	0.129	0.063	0.435
K	0.028	0.473	0.044	0.000	0.367	0.766	1.205	0.343	0.958	0.051
Mg	0.242	0.029	0.183	0.316	0.000	0.944	1.531	0.528	1.195	0.351
Mn	0.085	0.523	0.076	0.207	0.308	0.000	0.365	0.135	0.085	0.495
Na	0.100	0.199	0.064	0.204	0.079	0.080	0.000	0.490	0.223	0.841
P	0.033	0.237	0.022	0.042	0.127	0.171	0.086	0.000	0.235	0.163
Ti	0.045	0.481	0.042	0.135	0.278	0.021	0.088	0.113	0.000	0.639
rest	0.043	0.308	0.025	0.134	0.150	0.033	0.021	0.066	0.025	0.000
\mathbf{T}_2										

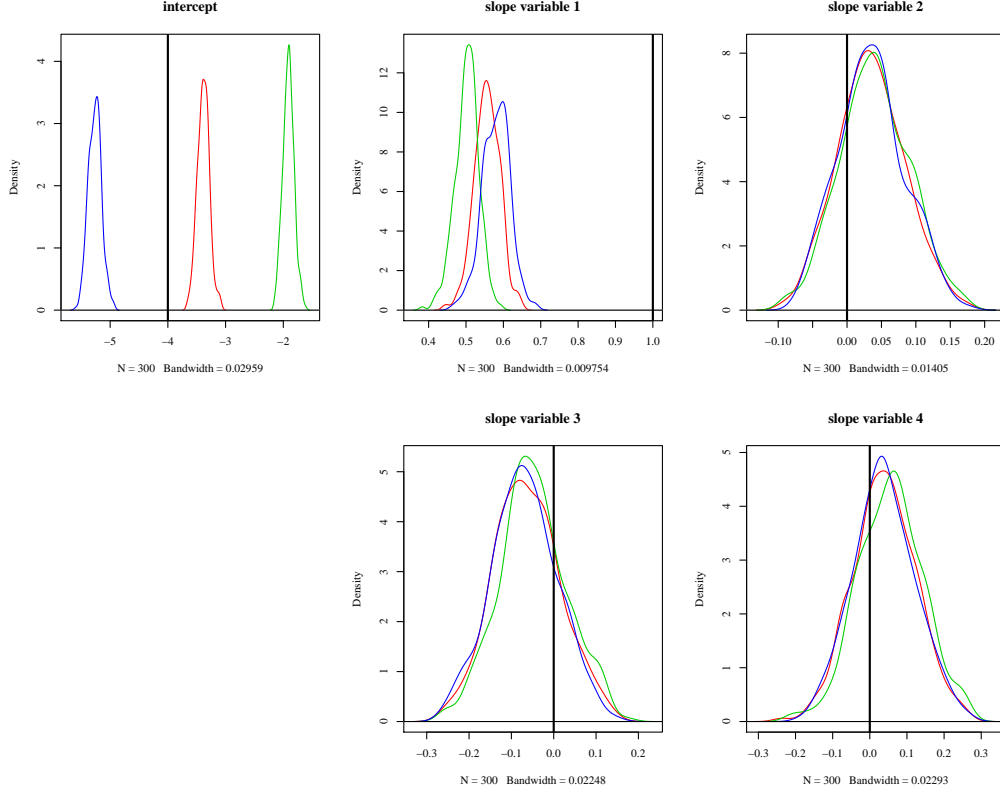


Figure 1: Monte Carlo kernel density estimates of the coefficients of the log-intensity function for the simulated case study, using 20×20 tiles (blue), 40×40 tiles (red) and 100×100 tiles (green). Equivalent results were obtained running 300 and 3000 simulations.

ical variation-variograms together with the fitted model,

$$t_{ij}(h|\boldsymbol{\theta}) = \mathbf{T}_0\delta(h) + \mathbf{T}_1\text{Sph}\left(\frac{h}{5}\right) + \mathbf{T}_2\text{Sph}\left(\frac{h}{25}\right),$$

where $\delta(h)$ represents the nugget effect, and $\text{Sph}(h/a)$ a spherical unitary variogram with range a in km. The sill matrices are reported in Table 1 in compact form.

Using the Monte Carlo/pseudo-likelihood approach illustrated before, 300 simulations of the 10-variable random field were obtained, conditioned to the $N = 601$ available data. For each simulation, Baddeley and Turner (2000) method was applied. Kernel densities of the resulting 300 estimates of each coefficient are displayed in Figure 3, and their means are reported in Table 2, showing that Fe seems to have a positive influence on the likelihood of deposit occurrence, while Mn and Al a negative one. This is a strong contrast with the results presented by Tolosana-Delgado and van den Boogaart (2014, high positive influence of Fe, Al and the rest, and high negative influence of Mn and K), which suggests that the uncertainty due to interpolation can have dramatical effects.

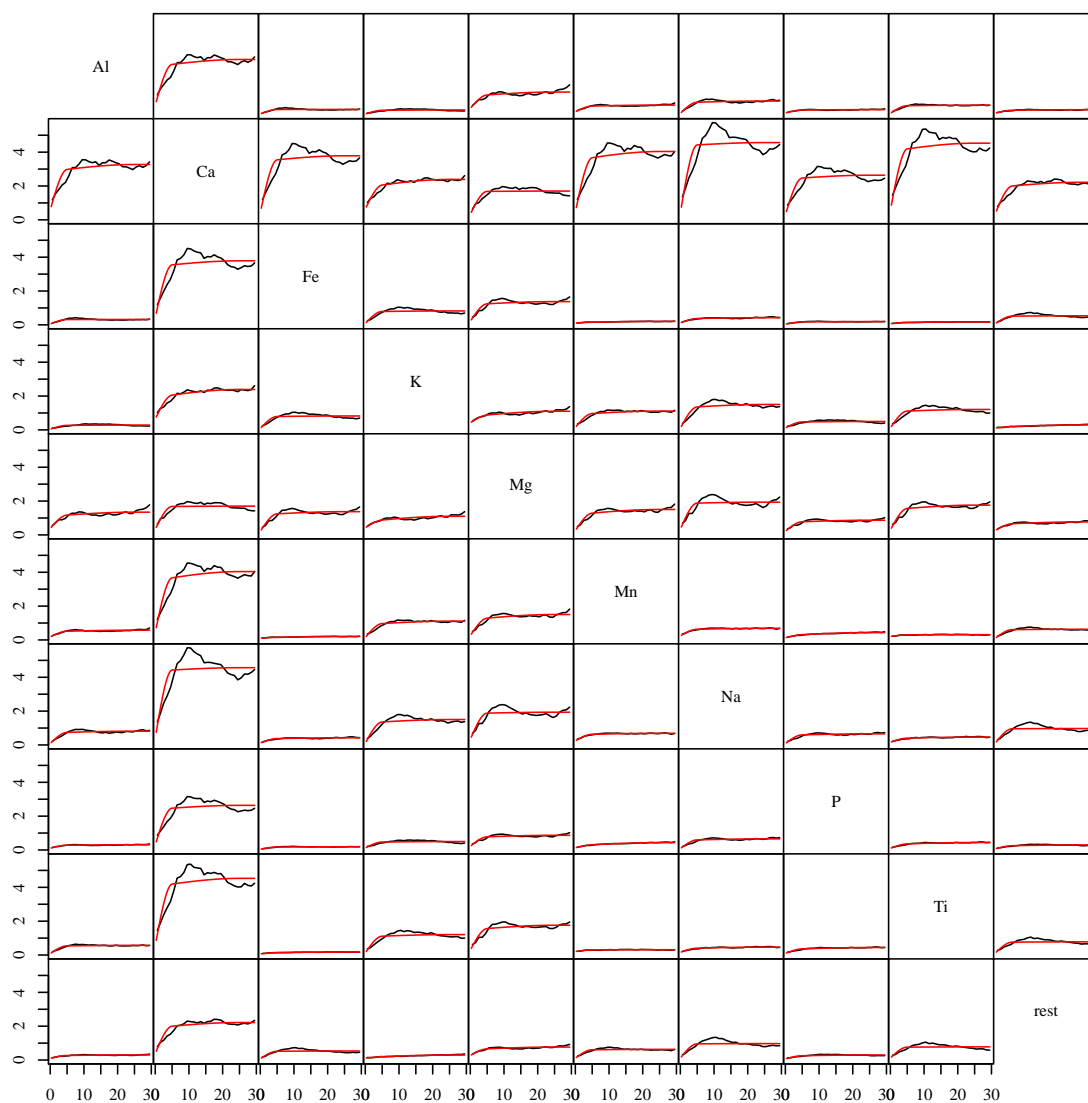


Figure 2: Variation-variograms of the geochemical dataset: empirical variograms (black line) and fitted LMC (red curve)

However, all curves in Figure 3 show that a zero coefficient is a reasonable assumption for each variable separately. To check whether all of them can be zero at the same time, the Mahalanobis distance between the set of simulated coefficients and the vector of means is compared with the Mahalanobis distance between zero and the vector of means (Fig. 4): as a sort of p-value, 27% of the simulated coefficients show a distance to the mean vector larger than the distance of zero, suggesting that all slope coefficients can be considered to be zero simultaneously, thus that no strong relation between spatial covariables and the Poisson process can be detected.

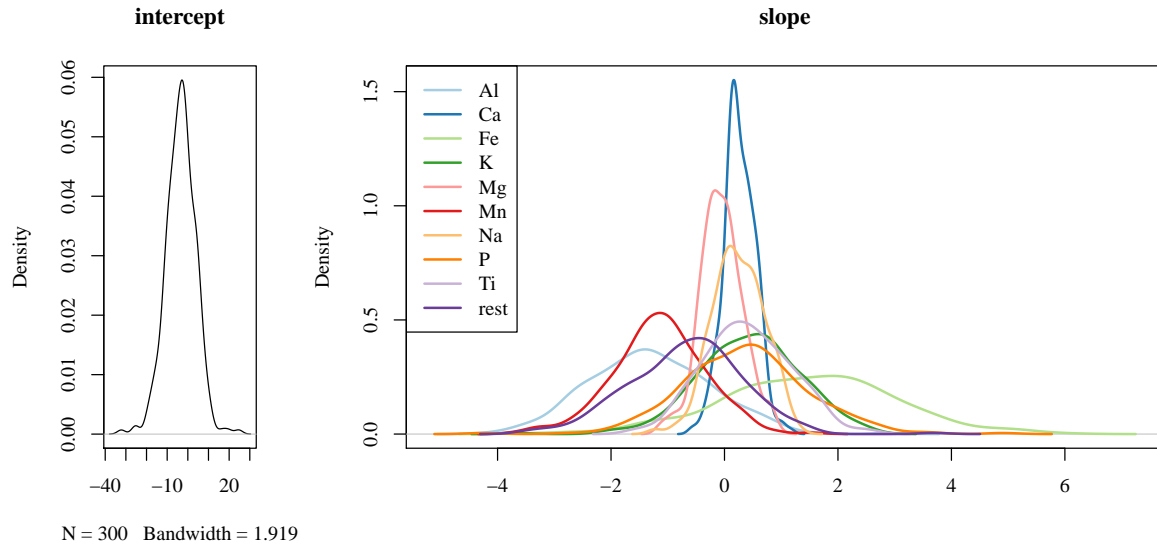


Figure 3: Kernel density estimates of the intercept and clr-slope coefficients for the Grazer Palaeozoikum data set. Note that the result suggest that most of the variables might have a null influence.

Table 2: Average estimates of the clr-slope coefficients for the Grazer Palaeozoikum data.

Al	Ca	Fe	K	Mg	Mn	Na	P	Ti	rest
-1.304	0.266	1.483	0.435	-0.055	-1.152	0.240	0.368	0.399	-0.679

References

- [1] Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall Ltd., London (UK)

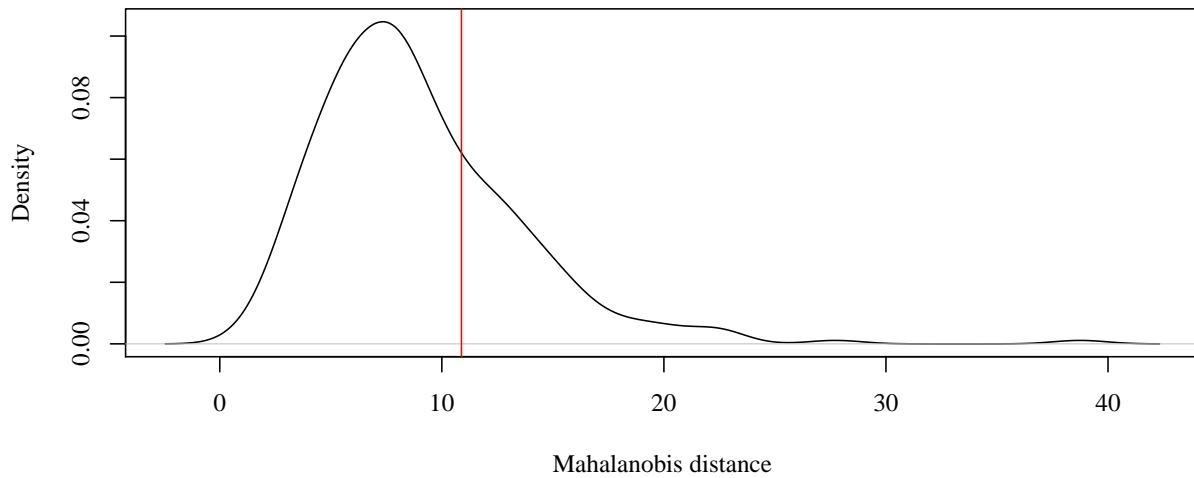


Figure 4: Kernel density estimate of the variability of the Mahalanobis distance between the estimated coefficients and their vector of means, compared with the Mahalanobis distance between that vector of means and the zero vector.

- [2] Baddeley, A. and R. Turner (2000). Practical Maximum Pseudolikelihood for Spatial Point Patterns (with Discussion). *Australian and New Zealand journal of statistics* 42, 283–322.
- [3] Egozcue, J. J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geosciences* 35, 279–300.
- [4] Pawlowsky-Glahn, V. and J. J. Egozcue (2002). BLU estimators and compositional data. *Mathematical Geosciences* 34, 259–274.
- [5] Pawlowsky-Glahn, V. and R. A. Olea (2004). *Geostatistical Analysis of Compositional Data*. Oxford University Press, Oxford.
- [6] Tolosana-Delgado, R. and K. G. van den Boogaart (2014) *Journal of Geochemical Exploration* (in press).
- [7] Tolosana-Delgado, R. and K. G. van den Boogaart (2014). Joint consistent mapping of high-dimensional geochemical surveys. *Mathematical Geosciences* 45, 983–1004.
- [8] van den Boogaart, G. and R. Tolosana-Delgado (2013). *Analysing compositional data with R*. Springer, Heidelberg
- [9] Weber, L. and J. Davis (1990). Multivariate statistical analysis of stream-sediment geochemistry in the Grazer Paläozoikum, Austria. *Mineralium Deposita* 25, 213–220.