Geostatistical Fisher discriminant analysis

R. TOLOSANA-DELGADO[1*], J. McKINLEY[2] and K.G. van den BOOGAART[1,3]

[1]Department Modelling and Valuation – Helmholtz Zentrum Dresden Rossendorf, Helmholtz Institute Freiberg for Resource Technology, Germany, r.tolosana@hzdr.de
[2]School of Geography, Archaeology and Palaeoecology – Queen's University Belfast, UK
[3]Institute for Stochastics – Technische Universität Bergakademie Freiberg, Germany
*presenting author

Abstract

A geostatistical version of the classical Fisher rule (linear discriminant analysis) is presented. This method is applicable when a large dataset of multivariate observations is available within a domain split in several known subdomains, and it assumes that the variograms (or covariance functions) are comparable between subdomains, which only differ in the mean values of the available variables. The method consists on finding the eigen-decomposition of the matrix $\mathbf{W}^{-1}\mathbf{B}$, where $\mathbf{W}$ is the matrix of sills of all direct- and cross-variograms, and $\mathbf{B}$ is the covariance matrix of the vectors of weighted means within each subdomain, obtained by generalized least squares. The method is used to map peat blanket occurrence in Northern Ireland, with data from the Tellus survey, which requires a minimal change to the general recipe: to use compositionally-compliant variogram tools and models, and work with log-ratio transformed data.

1 Introduction

If a sufficiently large regionalized multivariate data set has been obtained across several subdomains, it is possible to attempt to discriminate the several subdomains by making use of the available regionalized variables. This can be of use, for instance, to delineate on a map geologic or environmental units from geochemical information (predictive mapping, e.g. Grunsky, Mueller and Corrigan, 2014; or mineral potential mapping, e.g. Schaeben, 2012). This contribution presents a solution to this problem based on adapting the well-known Fisher rule of discriminant analysis to the case of spatially-dependent data.

2 The method

2.1 Notation and the non-spatial Fisher rule

Let the region of interest R be split into K different, disjoint sub-domains $R = R_1 \cup R_2 \cup \cdots \cup R_K$. Let a regionalized P-component random vector $\mathbf{Z}(x)$ be intrinsic stationary defined on the whole region R, but having a different mean within each sub-domain, i.e. if the location $x \in R_i$ then $E[\mathbf{Z}(x)] = \boldsymbol{\mu}(x) = \boldsymbol{\mu}_i$, where $E[\cdot]$ denotes an expected value. That intrinsic stationarity condition implies that there exists a matrix-valued function $\boldsymbol{\Gamma}(h)$ describing the covariance matrix of spatial increments, i.e. $Cov[\mathbf{Z}(x+h) - \mathbf{Z}(x)] = \boldsymbol{\Gamma}(h)$ if both x and $x+h$ belong to the same sub-domain; or in general

$$E\left[\big(\mathbf{Z}(x+h) - \mathbf{Z}(x)\big)^t \big(\mathbf{Z}(x+h) - \mathbf{Z}(x)\big)\right] = \boldsymbol{\Gamma}(h) + \big(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\big)^t \big(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\big), \quad (1)$$

if they belong to different sub-domains R_i and R_j. Here $E[\cdot]$ denotes the P-component mean vector of of a random vector. A regionalized data set of size N will be denoted by $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N\}$, where at each of location x_i a vector of P observations $\mathbf{z}_i = \mathbf{z}(x_i) = [z_{i1}, z_{i2}, \ldots, z_{iP}]$ is available. All vectors are considered row-vectors in this contribution.

In a non-spatial context, the Fisher rule states that the best (linear) discrimination between the domains (known as data groups) occurs on the subspace defined by the eigenvectors of the non-symmetric matrix (e.g. Fahrmeir and Hammerle, 1984)

$$\mathbf{Q} = \mathbf{W}^{-1}\mathbf{B}, \tag{2}$$

where $\mathbf{W}$ is the common within-group covariance matrix, and $\mathbf{B}$ is the between groups covariance matrix. Denote the number of samples, mean (row-)vectors and (maximum likelihood) covariance matrix estimates within the i-th group are denoted respectively as N_i, $\widehat{\boldsymbol{\mu}}_i$ and $\widehat{\boldsymbol{\Sigma}}_i$. Take the global mean

$$\widehat{\boldsymbol{\mu}} = \sum_{i=1}^{K} \omega_i \widehat{\boldsymbol{\mu}}_i \tag{3}$$

i.e. a weighted mean with weights $\omega_i = N_i/N$, and the total size of the data set $N = \sum_{i=1}^{K} N_i$. Then the within-group covariance matrix is estimated by $\widehat{\mathbf{W}} = \sum_i^K \omega_i \widehat{\boldsymbol{\Sigma}}_i$ and the between-group covariance as

$$\widehat{\mathbf{B}} = \sum_{i=1}^{K} \omega_i (\widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\mu}})'(\widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\mu}}), \tag{4}$$

where $\mathbf{a}'$ denotes $\mathbf{a}$ transposed. Plugging these estimates into Eq (2) delivers an estimate $\widehat{\mathbf{Q}}$ to treat with any existing eigenvalue decomposition routine.

2.2 The (theoretical) spatial Fisher rule

The adaption of the Fisher rule to the case of a regionalized random vector is quite trivial, as the within-group covariance matrix is by definition the double of the common variogram sill

$$\mathbf{W} = \lim_{h \to \infty} Cov[\mathbf{Z}(x+h), \mathbf{Z}(x)] = 2\boldsymbol{\Gamma}(\infty)$$

(assuming that x and $x+h$ remain within the same sub-region or group, no matter which). Here $Cov[\cdot]$ denotes the $(P \times P)$- matrix of the coefficients of a P-component vector with the coefficients of another P-component vector. The theoretical between-group variance does not imply a major difficulty either, as the weighting can simply be done using the size $|R_i|$ of each region R_i, i.e. $\mathbf{B} = \sum_{i=1}^{K} \omega_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})'(\boldsymbol{\mu}_i - \boldsymbol{\mu})$, with $\omega_i = |R_i|/|R|$ and $\boldsymbol{\mu} = \sum_{i=1}^{K} \omega_i \boldsymbol{\mu}_i$. Plugging these values into Eq (2) delivers the matrix $\mathbf{Q}$ of interest.

3 The practical spatial Fisher rule

3.1 Estimating the within-region covariance

If a regionalized multivariate data set is available, one can derive estimates of the matrices $\mathbf{W}$ and $\mathbf{B}$ which take the spatial dependence between the observations of the regionalized data set into account. For this, the only requisite is to have obtained a fitted model $\boldsymbol{\Gamma}(h|\hat{\theta})$ of the theoretical variogram $\boldsymbol{\Gamma}(h)$, no matter in which way. Again, $\mathbf{W}$ is estimated by the sill of the variogram model, by definition,

$$\widehat{\mathbf{W}} = 2\boldsymbol{\Gamma}(\infty|\hat{\theta}). \tag{5}$$

3.2 Estimating the region means and the global mean

The empirical between-group variance requires obtaining estimates of the mean of each sub-region, $\widehat{\boldsymbol{\mu}}_1$to $\widehat{\boldsymbol{\mu}}_K$, as well as a sort of global mean $\widehat{\boldsymbol{\mu}}$. Each group mean can be obtained with a generalized least squares approach (Wackernagel, 2002),

$$\widehat{\boldsymbol{\mu}}_i = \left(\mathbf{F}_i{}'\mathbf{C}_i^{-1}\mathbf{F}_i\right)^{-1}\left(\mathbf{F}_i{}'\mathbf{C}_i^{-1}\mathbf{Z}_i\right), \tag{6}$$

with $\mathbf{Z}_i$ the vector of all observation in R_i, and the following $(N_i \times N_i)$- and $(N_i \times 1)$-block matrices

$$\mathbf{C}_i = \begin{bmatrix} \mathbf{C}_{11} & \cdots & \mathbf{C}_{1N_i} \\ \vdots & \ddots & \vdots \\ \mathbf{C}_{N_i1} & \cdots & \mathbf{C}_{N_iN_i} \end{bmatrix} \text{ and } \mathbf{F}_i = \begin{bmatrix} \mathbf{I}_P \\ \vdots \\ \mathbf{I}_P \end{bmatrix}$$

formed by blocks of the form $\mathbf{C}_{ij} = \boldsymbol{\Gamma}(\infty|\hat{\theta}) - \frac{1}{2}\boldsymbol{\Gamma}(x_i - x_j|\hat{\theta})$ and $\mathbf{I}_P$ the $(P \times P)$-identity matrix; that is, $\mathbf{C}_i$ has $(PN_i \times PN_i)$ elements, while $\mathbf{F}_i$ has $(PN_i \times P)$ elements and $\mathbf{Z}_i$ is a vector of (PN_i) components. Note that these matrices are exactly the same as those used in ordinary cokriging, and that Eq. (6) is actually equivalent to block cokriging within the region R_i (Wackernagel, 2002). With these insights, the global „mean" estimate $\hat{\boldsymbol{\mu}}$ (whatever it is), could be obtained by block cokriging within the whole region R, i.e with the analogous expression

$$\hat{\boldsymbol{\mu}} = \left(\mathbf{F}'\mathbf{C}^{-1}\mathbf{F}\right)^{-1}\left(\mathbf{F}'\mathbf{C}^{-1}\mathbf{Z}\right), \tag{7}$$

with $\mathbf{Z}$ the vector of all (PN) observations, $\mathbf{F}$ the $(PN \times P)$-component matrix of identity blocks, and the matrix $\mathbf{C}$ is filled with blocks $\mathbf{C}_{ij} = 2\boldsymbol{\Gamma}(\infty|\hat{\theta}) - \frac{1}{2}\boldsymbol{\Gamma}(x_i - x_j|\hat{\theta})$ as well, as if the two data points would be placed on the same domain.

3.3 Estimating the between-region covariance

A final difficulty might appear to estimate the between-region covariance **B**, namely that we do not know the weights $\omega_1, \omega_2, \ldots, \omega_K$ in the regionalized case. Our proposal is to derive them from the interpretation of the global mean as a weighted average, i.e. from Eq. (3). Unfortunately, to the authors knowledge, there is no guarantee that this linear system of P equations (one for each component of these mean vectors) on K unknowns (the weights) has a unique solution, or that these weights are positive and sum to one, as required to use them in Eq. (4). In general, these two conditions (which define an end-member problem) will only be satisfied if the global mean $\hat{\boldsymbol{\mu}}$ belongs to the interior of the convex hull defined by the K sub-region means, $\hat{\boldsymbol{\mu}}_1$to $\hat{\boldsymbol{\mu}}_K$ (Weltje, 1997; Tolosana-Delgado, von Eynatten and Karius, 2011; Konsulke et al, 2015). However, given the smoothing nature of the kriging systems (Eq. 6) and (Eq. 7) and that the estimation of $\hat{\boldsymbol{\mu}}$ involves averaging all the data averaged in each $\hat{\boldsymbol{\mu}}_i$, it appears reasonable to assume that this convex hull condition will be satisfied most of the times, at least approximately. Thus, we can assume that some positive weights $\hat{\omega}_1, \hat{\omega}_2, \ldots, \hat{\omega}_K$ summing to 1 will be available by some of the methods mentioned by the authors cited before.

Once these weights are available, it is immediate to use them with all regional and global mean estimates on Eq. (4) to derive an estimate $\widehat{\mathbf{B}}$, that can be used together with $\widehat{\mathbf{W}}$ (Eq. 5), and then plugged into Eq. (2) to obtain the matrix $\widehat{\mathbf{Q}}$ to eigendecompose.

Note that all these complications do not occur in the case that the discrimination is desired between only two groups. In this case, the global mean and the weights are not necessary because the desired eigenvectors can be obtained using the matrix $\widehat{\mathbf{B}} = (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)'(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)$ within Eq. (2).

3.4 Mapping the discriminant functions

Let us assume that the eigenvectors $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_r\}$ of $\widehat{\mathbf{Q}}$ are available. The question is now how to generate maps of the discriminant scores $\mathbf{y}_i = \mathbf{y}(x_i) = [y_{i1}, y_{i2}, \ldots, y_{ir}]$, where each score $y_{ij} = \mathbf{z}_i\boldsymbol{v}_j'$ is a scalar product (projection) of one original observation times with one eigenvector. Three options appear, ordered by increasing computational complexity and decreasing (co)kriging error (Myers, 1983):

1. To apply ordinary kriging to each score separately, namely: (1) choose one eigenvector $\boldsymbol{v}_j$; (2) project the matrix-valued variogram model onto its direction, $\gamma_j(h|\hat{\theta}) = \boldsymbol{v}_j\boldsymbol{\Gamma}(h|\hat{\theta})\boldsymbol{v}_j'$; (3) take its scores $\{y_{1j}, y_{2j}, \ldots, y_{Nj}\}$; and (4) krige them with that projected variogram $\gamma_j(h|\hat{\theta})$.
2. To apply ordinary cokriging to all r scores simmultaneously, namely: (1) order the eigenvectors in a (r, P)-matrix $\mathbf{V}$; (2) project the (P, P)-matrix-valued variogram model onto a (r, r)-matrix-valued one, $\boldsymbol{\Gamma}_\mathbf{V}(h|\hat{\theta}) = \mathbf{V}\boldsymbol{\Gamma}(h|\hat{\theta})\mathbf{V}'$; (3) take all scores $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$; and (4) cokrige them with that projected variogram $\boldsymbol{\Gamma}_\mathbf{V}(h|\hat{\theta})$.

3. To apply ordinary cokriging to the original data with the original the (P, P)-matrix-valued variogram model to predict the original variables on a location x_0, and afterwards project each vector of interpolated values $\hat{\mathbf{z}}(x_0)$ onto each eigenvector, namely $\hat{y}_j(x_0) = \hat{\mathbf{z}}(x_0)\boldsymbol{v}_j{}'$.

Note that strictly speaking, all these algorithms should either be applied within a subregion only, or else first each observation should be subtracted from its subregion mean, then (co)kriging applied and then the mean of the subregion of the interpolated point added to the result. Both approaches are useless if we want to use this Fisher rule to map the unknown boundaries of a subregion. Hence, we recommend taking the resulting maps as a sort of heuristic rule. For this reason, in our opinion the computational burden of options 2 or 3 is not paid off by their theoretical better error against option 1.

3.5 Modifications with geochemical (and other compositional) data

In case that the regionalized data set $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N\}$ is formed by D-part compositional data (probably, the most common case, as all spatially-resolved surveys will fall in this category: soil particle size, soil geochemistry, groundwater geochemistry, stream geochemistry, etc) the whole theory and practice presented before is applicable, with a simple pre-processing step: to transform each composition by any one-to-one log-ratio transformation,

$$\mathbf{z}_i^* = (\ln \mathbf{z}_i) \cdot \boldsymbol{\Psi}',$$

i.e. for a (P, D)-matrix $\boldsymbol{\Psi}$ of rank $D - 1$, and apply the whole proposed methodology to the logratio score set $\{\mathbf{z}_1^*, \mathbf{z}_2^*, \ldots, \mathbf{z}_N^*\}$. Final results obtained (i.e. the maps of the discriminant functions) will be exactly the same whichever logratio transformation is used because the key equations of this method (Eqs. 2, 5-7) are affine equivariant, as proven by Filzmoser, Hron and Templ (2012) for Eq. (2) and Tolosana-Delgado (2006) for Eqs. (5) to (7).

Variation-variogram models proposed by Tolosana-Delgado and van den Boogaart (2013) to treat regionalized compositional data can also be used. In this case, the variogram model for $\boldsymbol{\Gamma}(h|\hat{\theta})$ used through Sections 3.1 and 3.2 must be obtained first from the variation-variogram model $\mathbf{T}(h|\hat{\theta})$, by

$$\boldsymbol{\Gamma}(h|\hat{\theta}) = -\frac{1}{2}\boldsymbol{\Psi}\mathbf{T}(h|\hat{\theta})\boldsymbol{\Psi}'.$$

The rest of the methodology applies directly, with $\boldsymbol{\Gamma}(h|\hat{\theta})$ describing the matrix-valued variogram model of the logratio score data set $\{\mathbf{z}_1^*, \mathbf{z}_2^*, \ldots, \mathbf{z}_N^*\}$.

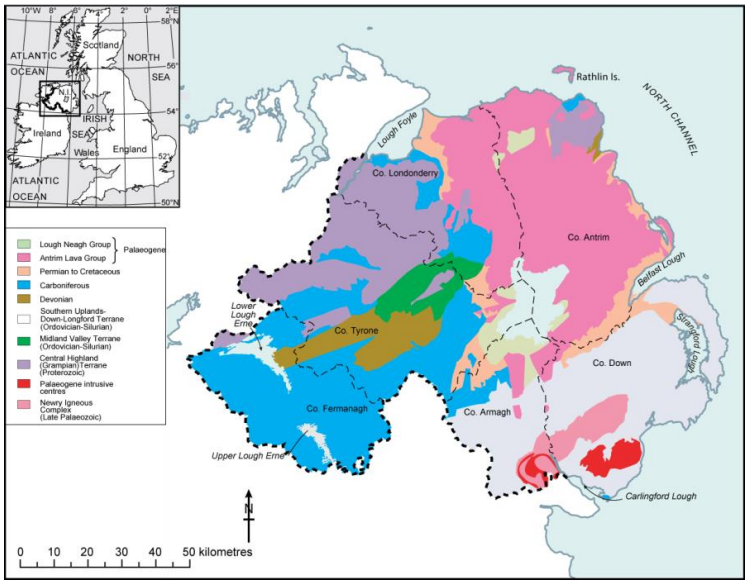

Figure 1. Geologic sketch and situation of Northern Ireland, after Mitchell (2004)

4 Application

4.1 The Tellus survey

Northern Ireland, part of the United Kingdom situated in the north east of the island of Ireland (Figure 1) covers less than 14,000 km2 in area but exhibits geology ranging from the Mesoproterozoic to Palaeogene in age (Mitchell, 2004). The bedrock can be simplified into a series of Caledonide terranes and part of a Palaeogene igneous province with distinct geological characteristics. Superficial peat cover of NI was estimated by Cruickshank et al. (1998) as 16% of the land surface, more recently peat cover is estimated as 12% of the land area of NI (Davies and Walker 2013). A large proportion of soil carbon is held within peat and organic-rich soils due to their high carbon content. In fact Malone and O'Connell (2009) cite that peat contains approximately 25% to 33% of the total soil organic carbon (SOC) worlwide (~450 GtC). This makes peat internationally significant with regard to the Kyoto Agreement and the global carbon cycle.

The Tellus project (Young and Donald, 2013), managed by the Geological Survey of Northern Ireland (GSNI), included a geochemical survey that saw the collection of nearly 30,000 soil, stream-sediment and stream-water samples across Northern Ireland between 2004 and 2006. The NI Tellus Survey soil samples were collected on a 2 km2 grid at depths of 5 – 20 cm ('A') and 35 – 50 cm ('S'). The 'A' soil sediment samples used comprises 6862 observations of 19 geochemical variables analysed by x-ray fluorescence spectrometry (XRFS). Full analytical and field methods employed by these comprehensive regional geochemical surveys can be found in Smyth (2007).

4.2 Variograms

For the following geostatistical analysis, 23 variables were considered, including the major oxides, LOI and some trace elements which had no zero issue. The variation-variograms, the set of direct variograms of all possible pairwise logratios, was computed and a model was fitted (Figure 2), with nugget effect and three isotropic spherical structures, with ranges 5km, 20km and 70km.

4.3 Spatial discrimination of blanket peat

The current algorithm available for the fitting of this spatial Fisher rule is not able to cope with the size of the Tellus data set. Hence we opted for randomly subsampling the data set taking samples of size N=1000 and fitting the algorithm each time. This was repeated 200 times, thus showing the capabilities of the model and some of its uncertainties at once. Figure 3 shows the boxplots and kernel density estimates of the resulting eigenvector (200 realizations), in centered logratio representation (Aitchison, 1986: the centered logratio considers each variable normalized by the geometric mean of all variables considered). In this representation, if the boxplot of two components overlap, then their logratio does not influence the favorability of having blanket peat. Moreover, components with high positive (or negative) values have a notable positive (or negative) influence of the blanket peat favorability. Hence, LOI, Na2O, MgO or Sn (but not any of their ratios) and clearly Se have a positive effect, and CaO, SiO2, Ni and As a negative one. Group means are not reported for lack of space.

The conditional density plot shows the contrast between the kernel estimated probabilities for each value of the discriminant score to belong to a blanket peat area or not (Figure 4 left). This shows that the probability of belonging to a peat area is never notably larger than 2/3 for very high values of the favorability score (>1), and that it drops to zero for low values of it (< -2). In between it mostly evolves as a sigmoid curve, to a first order, with a 1/2 probability around -0.8. However, a comparison of the areas of known peat covered areas with the spatial distribution of the samples colored after the resulting Fisher score (Figure 4 right) reveals two aspects that suggest our result is more valuable than a simple misclassification rate would tell. First, areas that would be identified as peat by the Fisher score are actually extensions of several surveyed peat occurrences, which could be an effect of a peat survey underestimating the real extension of peat. And second, almost all actual peat areas are found by the proposed methodology, with the exception of small raised bog areas, being the largest of these the southern coast of Lough Neagh, or potentially regions of peat degradation or extraction immediately north of Lough Neagh.

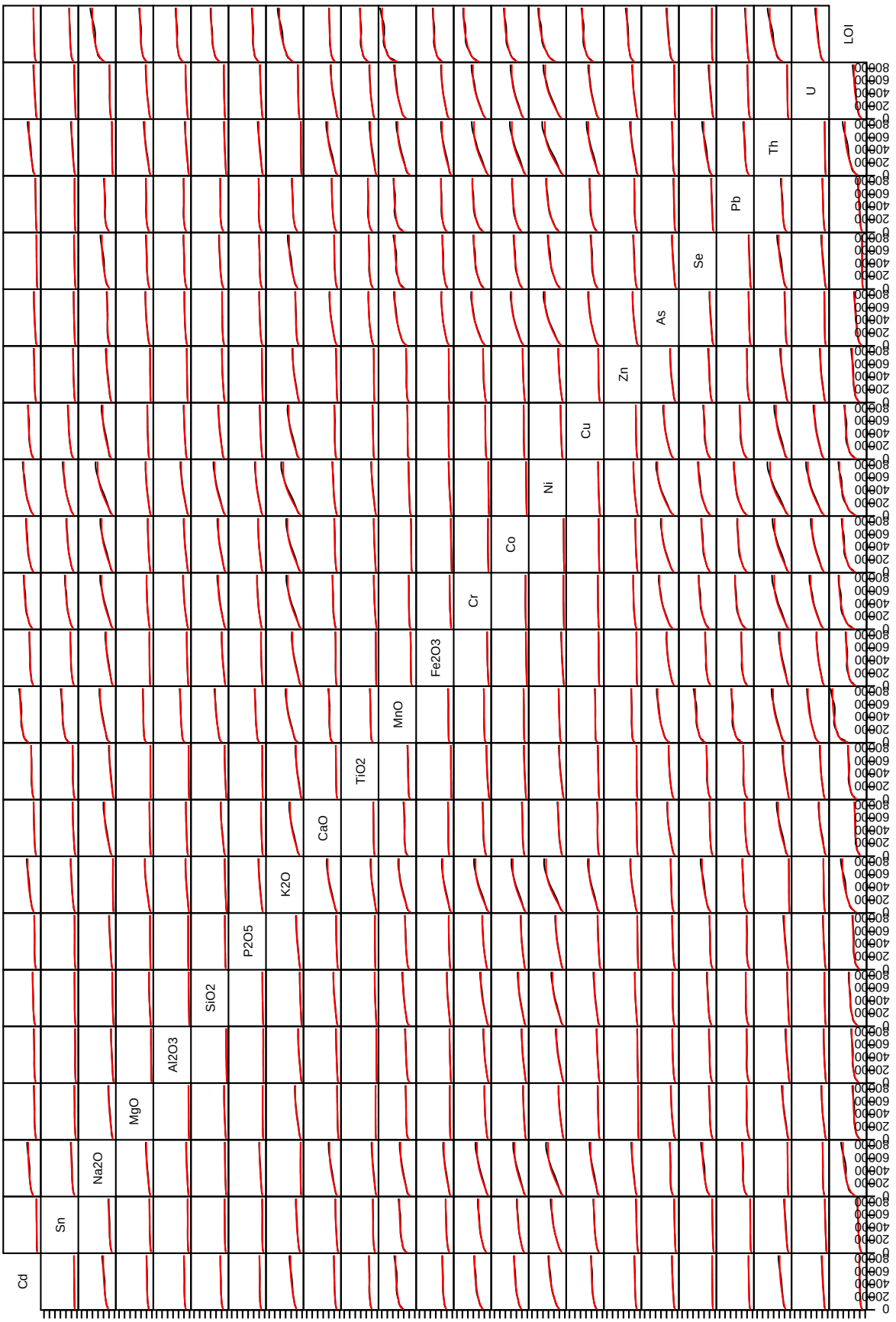

Figure 2. Variation-variogram of the 23 components considered (dimensionless vertical scale up to value 6)

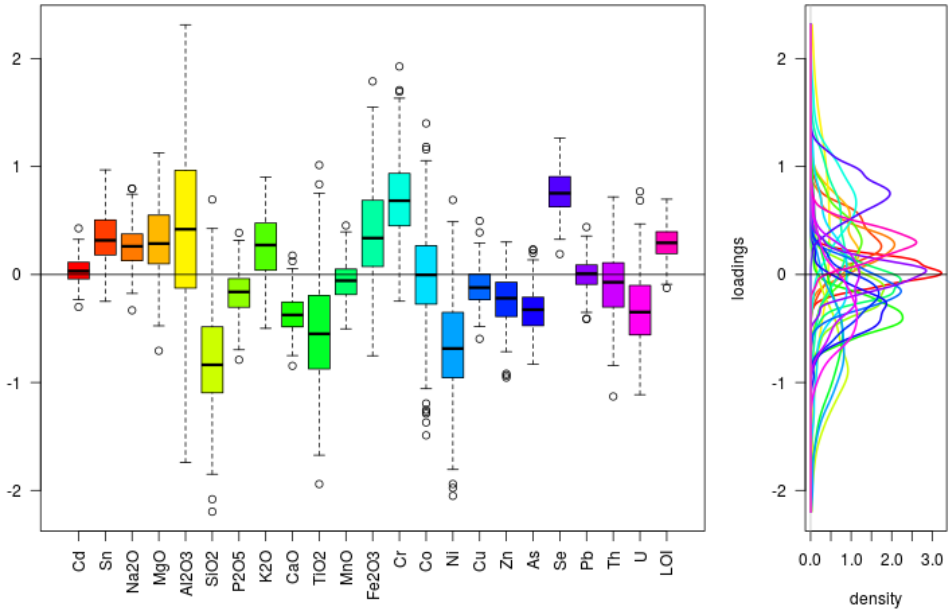

Figure 3. Resampling boxplots and kernel density estimates of the loadings of the discriminating function

5 Discussion and Conclusions

We presented a modification of the well known Fisher discrimination rule able to cope with spatial and compositional properties of explanatory data. This, applied to a case of peat classification from a soil geochemistry survey in Northern Ireland provided a good identification of peat covered areas. The variables having a clearer effect on the discrimination were LOI or SiO2 (as expected), but other less expected ones, like a positive Se effect, require further attention. Peat appears to be favored by high Cr/Ni ratios, a consistent effect with the known Cr sink effect of peat (McIlwaine et al, 2014). The spatial extent of some peat areas was overestimated, while other were underestimated. In the first case our classification may indicate a greater spatial extent of upland blanket peat areas. In the second case, our classification did not identify all areas of lowland raised peat bog, and may indicate peat covered areas which may be at risk due to degradation.

References

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.

Cruickshank, M. M., R.W. Tomlinson, P. M. Devine and R. Milne (1998) Carbon in the vegetation and soils of Northern Ireland. Biology and Environment: Proceedings of the Royal Irish Academy, 98B (1): 9–21

Davies, H. and S. Walker (2013) Strategic Planning Policy Statement (SPPS) for Northern Ireland: Strategic Environmental Assessment (SEA) Scoping Report. Leeds (UK).

Fahrmeir, L. and A. Hamerle, Eds. (1984) Multivariate Statistische Verfahren, Walter de Gruyter, Berlin (D), 796 p.

Filzmoser, P., K. Hron, and M. Templ (2012). Discriminant analysis for compositional data and robust parameter estimation. Computational Statistics and Data Analysis 27(4): 585-604

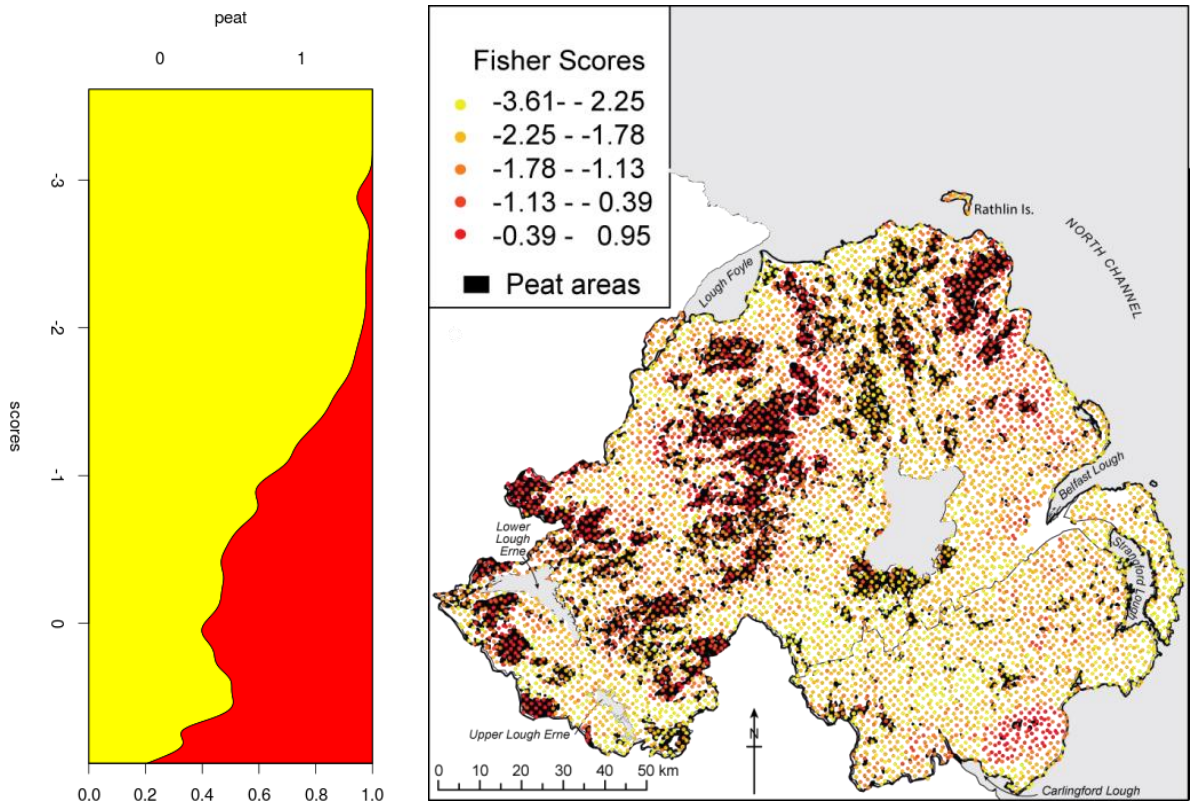

Figure 4. (left) Estimated conditional density plot, giving the odds of having Blanket peat for a certain Fisher score. (right) Map of Fisher scores for the samples compared with regions of known blanket peat

Grunsky, E.C., U.A. Mueller and D. Corrigan (2013). A study of the lake sediment geochemistry of the Melville Peninsula using multivariate methods: Applications for predictive geological mapping. Journal of Geochemical Exploration, 141(1): 15-41

Konsulke, S., S. Hopfe, R. Tolosana-Delgado, S. Matos Camacho and K.G. van den Boogaart (2015) Method for constructing a mineralogical composition from a measured sample of single components. *Proceedings of IAMG'2015, the Annual Meeting of the International Associa*tion for Mathematical Geosciences (this volume).

Malone S. and C. O'Connell (2009) *Ireland's Peatland Conservation Action Plan 2020* - halting the loss of peatland biodiversity. Irish Peatland Conservation Council, Dublin (Ireland).

McIlwaine, R, S. Cox, R. Doherty, S. Palmer, U. Ofterdinger and J. M. McKinley, (2014), Comparison of methods used to calculate typical threshold values for potentially toxic elements in soil. Environmental Geochemistry and Health, 36(5): 953-971 ., 10.1007/s10653-014-9611-x

Mitchell, W.I., ed. (2004). The Geology of Northern Ireland. Second Edi. Geological Survey of Northern Ireland, Belfast (UK).

Myers, D.E. (1983) Estimation of linear combinations and co-kriging. Mathematical Geology 15, 633–637.

Schaeben, H. (2012) Comparison of mathematical methods of potential modeling. Mathematical Geosciences 44, 101–129.

Smyth, D. (2007). Methods used in the Tellus geochemical mapping of Northern Ireland. British Geological Survey Open Report OR/07/022.

Tolosana-Delgado, R. (2006) Geostatistics for constrained variables: positive data, compositions and probabilities. Application to environmental hazard monitoring. PhD Dissertation, Universitat de Girona, Girona (Spain), available online.

Tolosana-Delgado, R., H. von Eynatten and V. Karius (2011) Constructing modal mineralogy from geochemical composition: A geometric-Bayesian approach. Computers & Geosciences 37: 677–691

Tolosana-Delgado, R. and K.G. van den Boogaart (2013). Joint consistent mapping of high-dimensional geochemical surveys. Mathematical Geosciences 45(8): 983–1004..

Tolosana-Delgado, R. and K.G. van den Boogaart (2014) Towards compositional geochemical potential mapping. Journal of Geochemical Exploration. 141(1): 42-51

Young, M.E., A.E. Donald, (2013). A Guide to the Tellus Data. Belfast: Geological Survey of Northern Ireland, Belfast (UK).

Wackernagel, H. (2002) Multivariate Geostatistics, An Introduction With Applications. 3 ed., Springer Verlag, Berlin (D). 387 p.

Weltje, J.G. (1997) End-member modeling of compositional data: numerical statistical algorithms for solving the explicit mixing problem. Mathematical Geology 29(4), 503–549.