2020

Biplots for compositional data derived from generalized joint diagonalization methods

Ute Mueller
Edith Cowan University

R. Tolosana Delgado

E. C. Grunsky

J. M. McKinley

10.1016/j.acags.2020.100044
Mueller, U., Delgado, R. T., Grunsky, E. C., & McKinley, J. M. (2020). Biplots for compositional data derived from generalized joint diagonalization methods. *Applied Computing and Geosciences, 8*, article 100044. https://doi.org/10.1016/j.acags.2020.100044

Contents lists available at ScienceDirect

Applied Computing and Geosciences

journal homepage: www.journals.elsevier.com/applied-computing-and-geosciences

Biplots for compositional data derived from generalized joint diagonalization methods

U. Mueller [a,*], R. Tolosana Delgado [b], E.C. Grunsky [c], J.M. McKinley [d]

[a] *Center of Ecosystems Management, School of Science, Edith Cowan University, Australia*
[b] *Helmholtz-Zentrum Dresden-Rossendorf, Helmholtz Institute Freiberg for Resources Technology, Freiberg, Germany*
[c] *Department of Earth and Environmental Sciences University of Waterloo, Waterloo, Canada*
[d] *School of Natural and Built Environment, Queen's University Belfast, Belfast, UK*

ARTICLE INFO

Keywords:
Semivariogram matrices
Spatial decorrelation
Structural analysis
Geochemical data

ABSTRACT

Biplots constructed from principal components of a compositional data set are an established means to explore its features. Principal Component Analysis (PCA) is also used to transform a set of spatial variables into spatially decorrelated factors. However, because no spatial structures are accounted for in the transformation the application of PCA is limited. In geostatistics and blind source separation a variety of different matrix diagonalization methods have been developed with the aim to provide spatially or temporally decorrelated factors. Just as PCA, many of these transformations are linear and so lend themselves to the construction of biplots. In this contribution we consider such biplots for a number of methods (MAF, UWEDGE and RJD transformations) and discuss how and if they can contribute to our understanding of relationships between the components of regionalized compositions. A comparison of the biplots with the PCA biplot commonly used in compositional data analysis for the case of data from the Northern Irish geochemical survey shows that the biplots from MAF and UWEDGE are comparable as are those from PCA and RJD. The biplots emphasize different aspects of the regionalized composition: for MAF and UWEDGE the focus is the spatial continuity, while for PCA and RJD it is variance explained. The results indicate that PCA and MAF combined provide adequate and complementary means for exploratory statistical analysis.

1. Introduction

Biplots constructed from principal components are an established means for exploring the features of a compositional data set Aitchison and Greenacre (2002). In several contributions (Mueller and Grunsky, 2016; Grunsky et al., 2017; McKinley et al., 2018) we have seen that the method of minimum maximum autocorrelation factors (MAF, Switzer and Green, 1984, Desbarats and Dimitrakopoulos, 2000, Tolosana-Delgado et al., 2019) enhances classification and improves spatial decorrelation of factors derived from regionalized compositions. It is therefore often preferred when the spatial properties of the composition cannot be ignored. However, to date biplots for MAF derived factors have not been explored, nor have biplots associated with more general joint (or simultaneous) diagonalizers based on the covariance or semivariogram function of the regionalized composition. These include uniformly weighted exhaustive diagonalization with Gauss iterations (UWEDGE, Tichavsky and Yeredor, 2009) and rotational joint diagonalization (RJD, Cardoso and Souloumiac, 1996). In what follows the inputs for biplots based on these schemes is considered. First the notion of joint diagonalization of a family of matrices is explored, along with approximate joint diagonalization. We then consider the application to multivariate spatial data and finally the application to regionalized composition and the construction of biplots. Following this exposition, an application to a subcomposition of the Tellus data set is considered.

2. Regionalized compositions and spatial decorrelation

A regionalized composition is a set $\{z(u) = [z_1(u),\ldots,z_D(u)] : z_k(u) > 0, k = 1,\ldots,D; \sum_{k=1}^{D} z_k(u) = c, u \in \mathscr{A}\}$ of compositional data defined on some study area $\mathscr{A}$, where $u \in \mathscr{A}$ denotes a location in $\mathscr{A}$ and c is an arbitrary, but fixed constant. It is common to transform compositions to log-ratios to avoid problems arising out of the fact that compositional data are closed and non-negative. The centered log-ratio (clr) transform is one of

[*] Corresponding author.
E-mail address: u.mueller@ecu.edu.au (U. Mueller).

https://doi.org/10.1016/j.acags.2020.100044
Received 12 October 2019; Received in revised form 20 September 2020; Accepted 19 November 2020
Available online 27 November 2020

the commonly used transformations to open up the simplex to real space. It is defined by $\zeta_k(u) = \ln(z_k(u)/g(u)), k = 1, \ldots, D$ where $g(u) = \left(\prod_{k=1}^{D} z_k(u)\right)^{1/D}$ denotes the geometric mean of the data at location u. The corresponding regionalized composition of clr-transformed variables will be denoted by $\{\zeta(u) = [\zeta_1(u), \ldots, \zeta_D(u)] : u \in \mathscr{A}\}$ and the corresponding variance covariance matrix by $\boldsymbol{\Sigma}_{clr}$. The image space of the clr transformation is the $(D-1)$–dimensional hyperplane H orthogonal to the vector 1_D in $\mathbb{R}^D$ all of whose entries are equal to 1. This makes $\boldsymbol{\Sigma}_{clr}$ singular, so care must be taken to deal with it appropriately.

To describe the spatial continuity of the regionalized composition, it is customary in geostatistics to use the semivariogram function, which is a matrix-valued function of the lag separation h defined by

$$\boldsymbol{\Gamma}_{clr}(h) = \frac{1}{2}E\left[(\boldsymbol{\zeta}(u) - \boldsymbol{\zeta}(u+h))^T(\boldsymbol{\zeta}(u) - \boldsymbol{\zeta}(u+h))\right] \tag{1}$$

in terms of clr transformed variables. The experimental counterparts of the regionalized composition are $\{z(u_\alpha) = [z_1(u_\alpha), \ldots, z_D(u_\alpha)] : u_\alpha \in \mathscr{A}, \alpha = 1, \ldots, n\}$ and $\{\boldsymbol{\zeta}(u_\alpha) = [\zeta_1(u_\alpha), \ldots, \zeta_D(u_\alpha)] : u_\alpha \in \mathscr{A}, \alpha = 1, \ldots, n\}$ and the experimental semivariogram $\boldsymbol{\Gamma}_{clr}(h_\ell) = [\gamma^{ij}_{clr}(h_\ell)]_{i,j=1,\ldots,D}$ at a lag value h_ℓ is defined componentwise as

$$\gamma^{ij}_{clr}(h_\ell) = \frac{1}{2N(h_\ell)}\sum_{u_\alpha - u_\beta \approx h_\ell}(\zeta_i(u_\alpha) - \zeta_i(u_\beta))\left(\zeta_j(u_\alpha) - \zeta_j(u_\beta)\right) \tag{2}$$

Thus associated with the regionalized composition there is a family $\{\boldsymbol{\Gamma}_{clr}(h_\ell), \ell = 1, \ldots, L\}$ of experimental semivariogram matrices calculated at L lag values h_ℓ chosen as appropriate to the nearest neighbor separation of the sample data.

To obtain a model of the spatial continuity of the data, a matrix-valued function is fitted to the experimental semivariograms and then used in subsequent estimation or simulation work. In the inference of an allowable model, in the case of the semivariogram function based on clr-data this model needs to be conditionally negative semidefinite (Pawlowsky-Glahn and Burger, 1992), restrictions are imposed, which make the fitting cumbersome. It is therefore often preferred to transform the data to spatially decorrelated factors and work with their family of semivariograms instead. This process of spatial decorrelation is based on a joint diagonalization of the semivariogram matrices. Since the covariance of the clr data is singular by construction, so are the semivariogram matrices. However, all spatial decorrelation methods fail under these circumstances, either because they require the inversion of one of these singular matrices, or because they end up in an underdetermined optimization problem. Several workaround strategies exist, all analogous to each other: generalized inversion, constrained optimization, change of coordinates. In the compositional literature, it is common to choose the last strategy, which is equivalent to the selection of a different log-ratio transformation (Pawlowsky-Glahn et al., 2015), arguably an isometric log-ratio transformation (ilr, Egozcue et al., 2003), which is associated with coordinates relative to an orthogonal basis of the clr image space. However, for simplicity of exposition, the choice made here is to base the mapping on the eigenvectors derived from the covariance matrix $\boldsymbol{\Sigma}_{clr}$ of the clr-data, $\boldsymbol{\Sigma}_{clr} = [\boldsymbol{U}_H, D^{-1/2}1_D]\,\boldsymbol{\Lambda}[\boldsymbol{U}_H, D^{-1/2}1_D]^T$. The matrix $\boldsymbol{U} = [\boldsymbol{U}_H, D^{-1/2}1_D]$ is orthogonal and contains the eigenvectors of $\boldsymbol{\Sigma}_{clr}$. The matrix $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_{D-1}, 0)$ is the diagonal matrix of eigenvalues arranged in descending order, with the last eigenvalue equal to 0 corresponding to the eigenvector $\mathbf{u}_D = D^{-1/2}1_D$. Specifically we put $\boldsymbol{\Gamma}_H(h_\ell) = \boldsymbol{U}_H^T\boldsymbol{\Gamma}_{clr}(h_\ell)\boldsymbol{U}_H, \ell = 1, \ldots, L$. This transformation results in a reduction of the dimension by 1, and for each $\ell = 1, \ldots, L$, the matrix $\boldsymbol{\Gamma}_H(h_\ell)$ is of size $(D-1) \times (D-1)$, and as it is a variance-covariance matrix of increments it is a positive definite matrix. This strategy is equivalent to using Moore-Penrose generalized inversion, to restricting optimization searches to the subspace orthogonal to $\mathbf{u}_D$, or to defining an ilr by means of $\boldsymbol{U}_H$.

3. Joint (approximate) diagonalization of positive definite real matrices

The general problem to be addressed is the following: Given a family of semivariogram matrices $\boldsymbol{M}_\ell = \boldsymbol{\Gamma}_H(h_\ell), \ell = 1, \ldots, L$ find a matrix $\boldsymbol{A}$ such that for all ℓ the equation $\boldsymbol{M}_\ell = \boldsymbol{A}\boldsymbol{\Lambda}_\ell\boldsymbol{A}^T$ is valid where $\boldsymbol{\Lambda}_\ell$ is a diagonal matrix. If such a matrix $\boldsymbol{A}$ exists, then the family $\{\boldsymbol{M}_\ell : \ell = 1, \ldots, L\}$ is said to be jointly diagonalizable. For real data such a matrix does not exist and so the condition of joint diagonalization is weakened to approximate diagonalization according to some suitably chosen criterion.

Moreover, the matrices are said to be jointly orthogonally diagonalizable, if the matrix $\boldsymbol{A}$ is orthogonal. In the latter case, the matrix $\boldsymbol{A}$ is unique up to permutation and scaling by a diagonal matrix, all of whose entries are ±1 (Afsari 2007). If the matrices are not jointly orthogonally diagonalizable, one can relax the requirement that the matrix $\boldsymbol{A}$ be orthogonal and instead require invertibility only. This relaxation comes at the cost that the matrix, if it exists, is only unique up to scaling and permutation.

Checking whether a family of symmetric matrices is jointly diagonalizable is straightforward: the matrices in the family need to commute pairwise (Horn and Johnson, 1985), that is the condition $\boldsymbol{M}_\ell\boldsymbol{M}_k = \boldsymbol{M}_k\boldsymbol{M}_\ell$ is satisfied for all $k, \ell = 1, \ldots, L$. In that case the matrix $\boldsymbol{A}$ is the matrix of eigenvectors of one of the matrices in the set and as such can be chosen to be orthogonal. In most applications however, pairwise commutativity is not satisfied. For example, consider the case where $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are both real positive definite matrices of the same size, but $\boldsymbol{M}_1\boldsymbol{M}_2 \neq \boldsymbol{M}_2\boldsymbol{M}_1$. In that case it can be shown that there exists an invertible matrix $\boldsymbol{C}$ such that $\boldsymbol{C}\boldsymbol{M}_1\boldsymbol{C}^T$ and $\boldsymbol{C}\boldsymbol{M}_2\boldsymbol{C}^T$ are diagonal if and only if $\boldsymbol{M}_2^{-1}\boldsymbol{M}_1$ is diagonalizable. The matrices $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are then said to be diagonalizable by congruence and this condition can be extended to larger families of matrices (Jiang and Li, 2015).

It is therefore common to distinguish between orthogonal and non-orthogonal joint diagonalization problems (OJD and NOJD). In general "experimental" settings joint diagonalization is not possible, and so approximate joint diagonalization has received a lot of attention, in particular in the context of blind source separation (for example, Cardoso and Souloumiac, 1996; Tichavsky and Yeredor, 2009).

4. PCA and MAF

The solution proposed by the PCA method consists of directly using the representation through the eigendecomposition of the variance-covariance matrix $\boldsymbol{\Sigma}_{clr}$ of the clr transformed data, $\boldsymbol{\Sigma}_{clr} = [\boldsymbol{U}_H, D^{-1/2}1_D]\boldsymbol{\Lambda}[\boldsymbol{U}_H, D^{-1/2}1_D]^T$ proposed on the preceding section. The transformation which diagonalizes $\boldsymbol{\Sigma}_{clr}$ on H is given by $\boldsymbol{\Sigma}_H = \boldsymbol{U}_H^T\boldsymbol{\Sigma}_{clr}\boldsymbol{U}_H = \boldsymbol{\Lambda}_H = \mathrm{diag}(\lambda_1, \ldots, \lambda_{D-1})$ and the eigenvalues reflect the variability represented by the corresponding factor (arranged in descending order).

For MAF, the variance-covariance matrix $\boldsymbol{\Sigma}_H$ and the semivariogram matrix $\boldsymbol{\Gamma}_H(h)$ at a chosen lag h are diagonalized jointly by congruence. Since $\boldsymbol{\Sigma}_H = \boldsymbol{\Lambda}_H$ is diagonal, matrix $\boldsymbol{A}$ is given by $\boldsymbol{A} = \boldsymbol{W}_1^T\boldsymbol{\Lambda}_H^{-1/2}$ where $\boldsymbol{W}_1$ is the orthogonal matrix which diagonalizes $\boldsymbol{\Lambda}_H^{-1/2}\boldsymbol{\Gamma}_H(h)\boldsymbol{\Lambda}_H^{-1/2}$. The matrix $\boldsymbol{A}$ is non-singular by construction and satisfies $\boldsymbol{A}\boldsymbol{\Sigma}_H\boldsymbol{A}^T = \boldsymbol{I}$ and $\boldsymbol{A}\boldsymbol{\Gamma}_H(h)\boldsymbol{A}^T = \boldsymbol{\Lambda}_1$. The eigenvalues in the matrix $\boldsymbol{\Lambda}_1$ are arranged in descending order, implying that factors will exhibit diminishing spatial continuity. In essence, MAF can thus be seen as a combination of two PCAs, the first being applied to $\boldsymbol{\Sigma}_{clr}$ and the second to $\boldsymbol{\Lambda}_H^{-1/2}\boldsymbol{\Gamma}_H(h)\boldsymbol{\Lambda}_H^{-1/2}$.

Table 1
Loading matrices for diagonalization methods and their inverses.

Method	clr-loading matrix $\boldsymbol{\Phi}$	inverse clr-loading matrix $\boldsymbol{\Psi}$
PCA	$\boldsymbol{U}_H$	$\boldsymbol{U}_H^T$
MAF	$\boldsymbol{U}_H\boldsymbol{\Lambda}_H^{-1/2}\boldsymbol{W}_1$	$\boldsymbol{W}_1^T\boldsymbol{\Lambda}_H^{1/2}\boldsymbol{U}_H^T$
RJD	$\boldsymbol{U}_H\boldsymbol{A}$	$\boldsymbol{A}^T\boldsymbol{U}_H^T$
UWEDGE	$\boldsymbol{U}_H\boldsymbol{A}$	$\boldsymbol{A}^{-1}\boldsymbol{U}_H^T$

5. Approximate joint diagonalization

Diagonalization via PCA is an OJD method, while MAF is a NOJD method. In the context of a family of semivariogram matrices that is sought to be diagonalized, the transformation derived from PCA will only diagonalize all semivariogram matrices if they commute pairwise. This condition is usually not satisfied, as the data on which calculation of the semivariogram matrices is based are noisy and there are different spatial scales that impact on the continuity of the data.

To account for phenomena of this type, the chosen blind source separation methods attempt to derive the best diagonalizer based on the entire family of matrices available. Typically some kind of fixed point iteration is used to determine the matrix $\boldsymbol{A}$ that best jointly diagonalizes the given family of symmetric matrices according to some cost criterion. In the case of OJD the cost function is set to be

$$C_1(\boldsymbol{A}) = \sum_{\ell=1}^{L} \text{trace}\left(\left(\boldsymbol{M}_\ell^{\boldsymbol{A}} - \text{diag}\left(\boldsymbol{M}_\ell^{\boldsymbol{A}}\right)\right)^T\left(\boldsymbol{M}_\ell^{\boldsymbol{A}} - \text{diag}\left(\boldsymbol{M}_\ell^{\boldsymbol{A}}\right)\right)\right) \tag{3}$$

where one seeks to minimize the sum of squares of the off-diagonal entries in $\boldsymbol{M}_l^{\boldsymbol{A}} = \boldsymbol{A}^T\boldsymbol{M}_\ell\boldsymbol{A}$. In the RJD algorithm the matrix $\boldsymbol{A}$ is constructed iteratively from Jacobi rotation matrices (Cardoso and Souloumiac 1996).

The criterion used for NOJD is not very different from it, in the case of the UWEDGE method (Tichavsky and Yeredor, 2009) a matrix $\boldsymbol{W}$ is sought that minimizes

$$C_3(\boldsymbol{W},\boldsymbol{A}) = \sum_{\ell=1}^{L} \text{trace}\left(\left(\boldsymbol{M}_\ell^{\boldsymbol{A}} - \boldsymbol{W}\boldsymbol{\Lambda}_{\ell,\boldsymbol{W}}\boldsymbol{W}^T\right)^T\left(\boldsymbol{M}_\ell^{\boldsymbol{A}} - \boldsymbol{W}\boldsymbol{\Lambda}_{\ell,\boldsymbol{W}}\boldsymbol{W}^T\right)\right) \tag{4}$$

where $\boldsymbol{\Lambda}_{\ell,\boldsymbol{W}} = \text{diag}(\boldsymbol{M}_\ell^{\boldsymbol{A}})$. The matrices $\boldsymbol{W}$ and $\boldsymbol{A}$ are called the mixing matrix and demixing matrix respectively. The naming of the matrices $\boldsymbol{W}$ and $\boldsymbol{A}$ reflects the manner in which they act on $\boldsymbol{M}_\ell$ and $\boldsymbol{\Lambda}_{\ell,\boldsymbol{W}}$ respectively: The matrix $\boldsymbol{A}$ approximately diagonalizes $\boldsymbol{M}_\ell$, ie "demixes" $\boldsymbol{M}_\ell$ while the matrix $\boldsymbol{W}$ transforms the diagonal matrix $\boldsymbol{\Lambda}_{\ell,\boldsymbol{W}}$ to a symmetric matrix, which is not necessarily diagonal (and in that sense "mixed"), but as close as possible to $\boldsymbol{M}_\ell^{\boldsymbol{A}}$.

Given the regionalized composition the biplot for any one of the methods is constructed based on the clr-transformed data. Given that the diagonalization matrix $\boldsymbol{W}$ is determined that approximately jointly diagonalizes the semivariogram matrices in the family $\{\boldsymbol{\Gamma}_H(h_\ell) : \ell = 1, \ldots, L\}$, the result needs to be recast into clr space. Indeed, if $\boldsymbol{Z}_{clr} = [z_k(u_\alpha)]_{\alpha=1,\ldots n,k=1,\ldots,D}$ denotes the $n \times D$ matrix of centered clr-scores, then the factors are given by

$$\boldsymbol{F} = \boldsymbol{Z}_{clr}\boldsymbol{U}_H\boldsymbol{A}. \tag{5}$$

The columns of $\boldsymbol{F}$ represent the scores and the rows of $\boldsymbol{\Phi} = \boldsymbol{U}_H\boldsymbol{A}$ are the loadings of the factors, i.e. the contribution of each original (clr-transformed) component onto the new factors. The original components can be recovered from the vector of scores by

$$\boldsymbol{Z}_{clr} = \boldsymbol{F}\boldsymbol{\Psi} \tag{6}$$

where $\boldsymbol{\Psi} = \boldsymbol{A}^{-1}\boldsymbol{U}_H^T$ is the pseudo-inverse of $\boldsymbol{\Phi}$. The equivalent of a scree plot can be obtained by calculating the explained variance $s_j^2 (j = 1, \ldots, D-1)$ attributable to the j^{th} factor $\boldsymbol{f}_j$. The expression in Eq. (6) may be rewritten as a sum of outer products $\boldsymbol{Z}_{clr} = \sum_{j=1}^{D-1} \boldsymbol{f}_j\psi_j$ where $\boldsymbol{f}_j$ and ψ_j denote the j^{th} column of $\boldsymbol{F}$ and the j^{th} row of $\boldsymbol{\Psi}$ respectively. Further, the total variance is given by $\text{mvar}(\boldsymbol{Z}_{clr}) = \text{trace}\ (\boldsymbol{Z}_{clr}^T\boldsymbol{Z}_{clr})$, since $\boldsymbol{Z}_{clr}$ is centered. Expanding the total variance it therefore follows that

$$\begin{aligned}\text{mvar}(\boldsymbol{Z}_{clr}) &= \text{trace}\left(\boldsymbol{Z}_{clr}^T\boldsymbol{Z}_{clr}\right)\\ &= \text{trace}(\boldsymbol{\Psi}^T\boldsymbol{F}^T\boldsymbol{F}\boldsymbol{\Psi}) = \text{trace}(\boldsymbol{\Psi}^T\text{var}(\boldsymbol{F})\boldsymbol{\Psi})\\ &= \sum\nolimits_{i=1}^{D}\sum\nolimits_{j,j'=1}^{D-1}\psi_{ji}[\text{var}(\boldsymbol{F})]_{jj'}\psi_{j'i} = \sum\nolimits_{i=1}^{D}\sum\nolimits_{j=1}^{D-1}\psi_{ji}\text{var}(\boldsymbol{f}_j)\psi_{ji}\\ &= \sum\nolimits_{j=1}^{D-1}\text{var}(\boldsymbol{f}_j)\left(\sum\nolimits_{i=1}^{D}\psi_{ji}^2\right) = \sum\nolimits_{j=1}^{D-1}\|\psi\|_j^2\text{var}(\boldsymbol{f}_j)\end{aligned}$$

since $[\text{var}(\boldsymbol{F})]_{jj'} = \text{cov}(\boldsymbol{f}_j, \boldsymbol{f}_{j'}) = 0$ for $j \neq j'$, and $[\text{var}(\boldsymbol{F})]_{jj} = \text{var}(\boldsymbol{f}_j)$, so that each

$$s_j^2 = \|\psi\|_j^2\text{var}(\boldsymbol{f}_j) \tag{7}$$

can be understood as the contribution of the j^{th} factor to the total explained variability.

To construct a (form) biplot to capture the variability of the original data from these results, we take Eq. (6) as the guiding tool, and following Graffelmann and Eeuwijk (2005), identify the analogues of the principal coordinates (to be represented as dots) as the columns of $\boldsymbol{F}$ whereas the analogues to the standard coordinates (to be represented as arrows or axes) are taken as the rows of $\boldsymbol{\Psi}$. The resulting loading matrices and their inverses expressed in clr are summarized in Table 1.

6. Data

The Tellus Survey covering the region of Northern Ireland, UK (GSNI, 2007; Young and Donald, 2013) consists of 6862 rural soil samples (X-ray

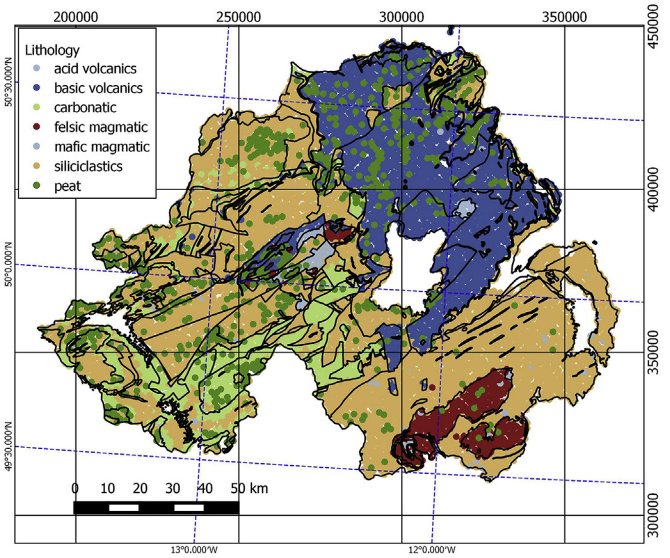

Fig. 1. Map of Northern Ireland covered by broad lithological classes, locations of peat occurrence shown in green.

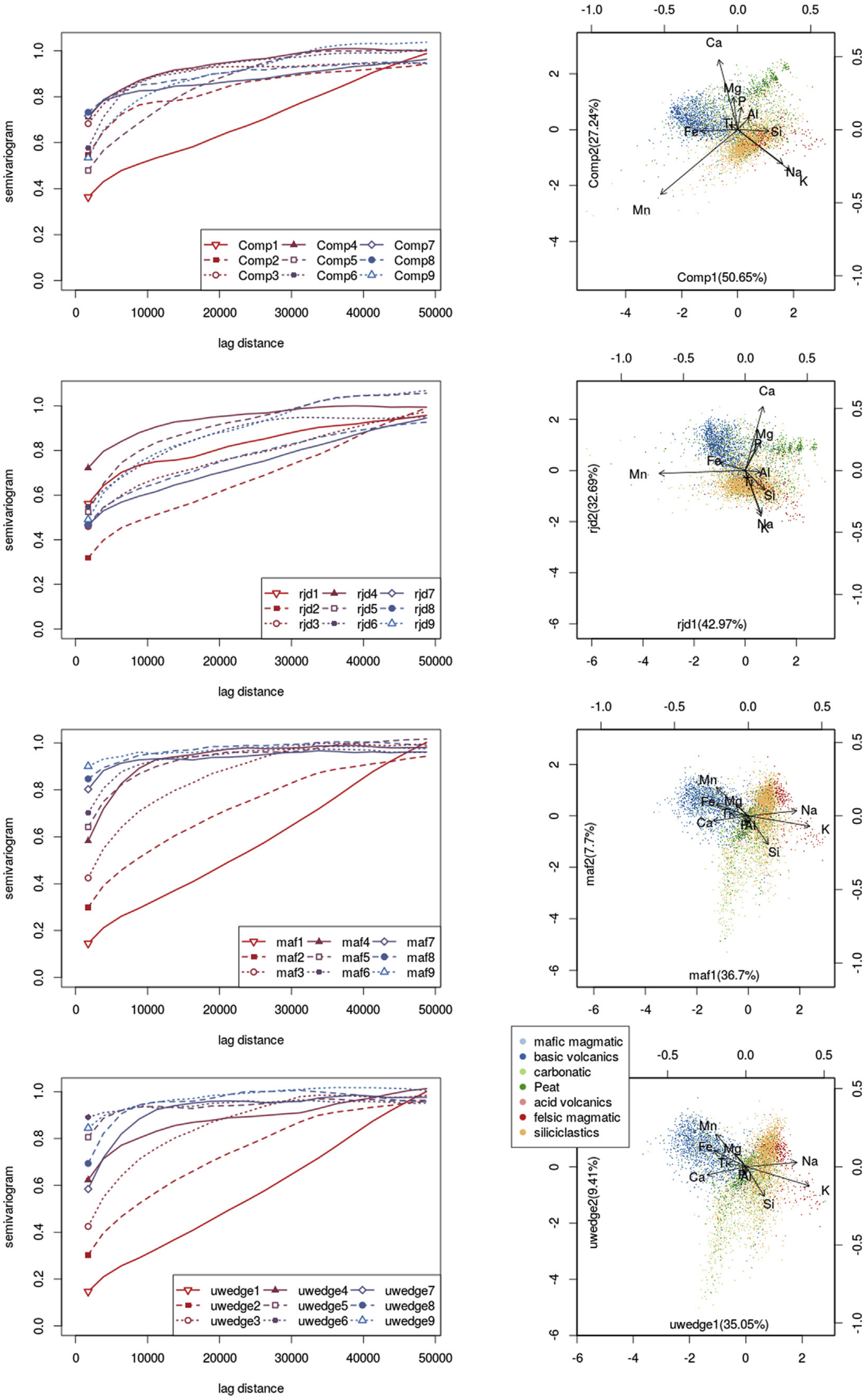

Fig. 2. Experimental semivariograms of the factors (left) and biplots of the first two factors colored by lithology (right) by method (order from top to bottom: PCA, RJD, MAF, UWEDGE, the axes labels Comp1, Comp2, rjd1, rjd2, maf1, maf 2, uwedge 1 and uwedge2 denote the first and second factor respectively).

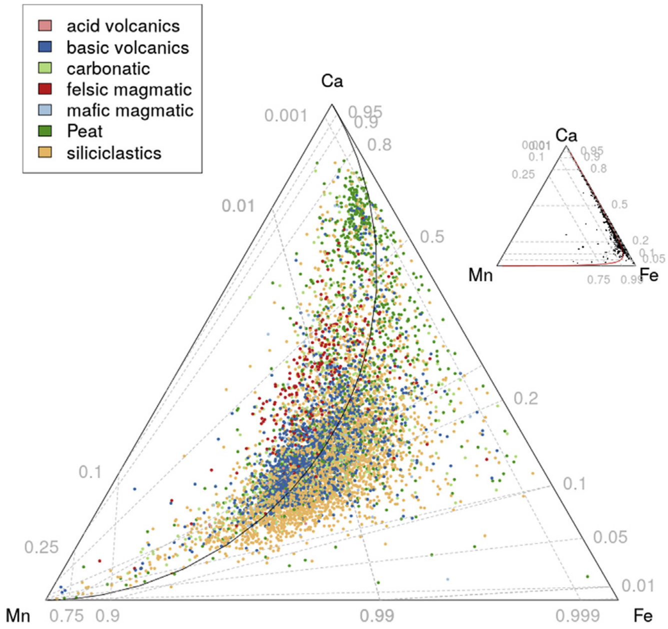

Fig. 3. Ternary diagram of the subcomposition Mn, Fe, Ca.

fluorescence (XRF) analyses). Geochemical samples presented in this study were collected at 20-cm depth, with average spatial coverage of one sample site every 2 km^2. Each soil sample site was assigned one of six broad lithological classes (acid volcanics, felsic magmatics, basic volcanics, mafic magmatics, carbonatic, silicic clastics) and in addition peat (Fig. 1) as described in Tolosana-Delgado and McKinley (2016).

At each location 50 continuous geochemical variables were available for analysis and transformed to elemental weight% prior to any work (Ag, Al, As, Ba, Bi, Br, Ca, Cd, Ce, Cl, Co, Cr, Cs, Cu, Fe, Ga, Ge, Hf, I, K, La, Mg, Mn, Mo, Na, Nb, Nd, Ni, P, Pb, Rb, S, Sb, Sc, Se, Si, Sm, Sn, Sr, Th, Ti, Tl, U, V, W, Y, Yb, Zn, Zr and Loss on Ignition (LOI)). More information on Tellus Survey field methods and analytical methodology are available in Smyth (2007) and Young and Donald (2013). Since the objective of this paper is to explore differences and similarities in the biplots obtained from the application of the various joint diagonalization approaches, the methods were applied to compute factor scores and construct biplots for the subcomposition Al, Fe, K, Mg, Mn, Na, P, Si, and Ti. We chose this 9 part subcomposition as it represents the bulk of the variability of the data. The major elements explain the variability of the compositions of the lithologies across Northern Ireland. The trace elements (omitted) account for less of the variability but can also be used to map out the major lithologies. For the sake of clarity and the minimization of redundancy, we believe the major elements are sufficient. This can be demonstrated by a simple principal component analysis of the data as documented in Tolosana-Delgado and McKinley (2016).

Experimental direct and cross variograms of the clr data were computed for 30 lags at a nominal spacing of 1 km. The MAF transform was based on an estimate of the covariance matrix and the semivariogram matrix for the first lag, for the RJD and UWEDGE methods the semivariogram matrices for all lags up to distance 20 km were used. All calculations were done in R (R Core Team, 2018), making use of package "compositions" for the log-ratio transformations, colored biplots and PCA analysis (van den Boogaart and Tolosana-Delgado, 2013), package "JADE" for RJD calculations (Miettinen et al., 2017) and package "jointDiag" for UWEDGE calculations (Gouy-Pailler, 2017).

7. Results

The experimental semivariograms of the factors and biplots of the first two components for the subcomposition of major oxides are shown in Fig. 2. For all four methods one of the experimental semivariograms shows a strong trend, for PCA, MAF and UWEDGE this is the one for the first factor. In contrast, for RJD it is factor 2. Semivariograms for PCA and RJD show greater short-scale variability than the semivariograms for the first four MAF and UWEDGE factors. For the remaining semivariograms the short scale variability of remaining MAF and UWEDGE factors is greater than that of RJD and PCA factors.

For PCA and RJD no ordering by spatial continuity results as a consequence of the transformation. In contrast for MAF there is a clear ordering of the factors by decreasing continuity as shown in the experimental semivariograms. For UWEDGE the semivariograms of the first 3 factors only show negligible differences to those from MAF, but the ordering by spatial continuity is absent for the remaining factors, however, the ordering could be restored via a permutation of the factors. The decrease in continuity is evident through a flattening of the semivariograms, with decreasing range and increasing nugget to sill ratio.

Biplots for MAF and UWEDGE are strongly similar, showing a ternary system of mafics, felsics, and siliciclastic materials. The scores corresponding to peat form a tight cluster. Comp1 for PCA is related to peat building (Tolosana-Delgado and McKinley, 2016). All four biplots show a separation of lithologies. The MAF and UWEDGE biplots are almost identical and indicate collinearity of Mn, Fe and Ca, a feature not evident in the PCA or RJD biplot. The Fe–Mn rays in the biplots are shared between the basalts and the siliciclastic materials in the PCA and RJD biplots. This association reflects the likelihood that Fe–Mn in the basalts are stoichiometrically associated with ferromagnesian mineral structures in both materials and also the likelihood that Fe–Mn oxihydroxide coatings have formed on silicate minerals in the soils as a result of groundwater circulation. In the case of MAF and UWEDGE, the Fe–Mn association is more clearly associated with the basalts, which have a significant geospatial influence and thus reflect the basalt terrain.

This is also confirmed by the ternary diagram for this subcomposition shown in Fig. 3, which also shows this one-dimensional pattern.

Scree plots in Fig. 4 suggest that the ordering by decreasing continuity

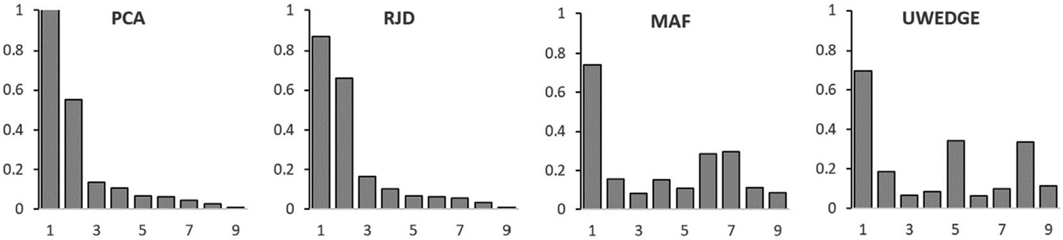

Fig. 4. Scree plots for factorization based on PCA, RJD, MAF and UWEDGE.

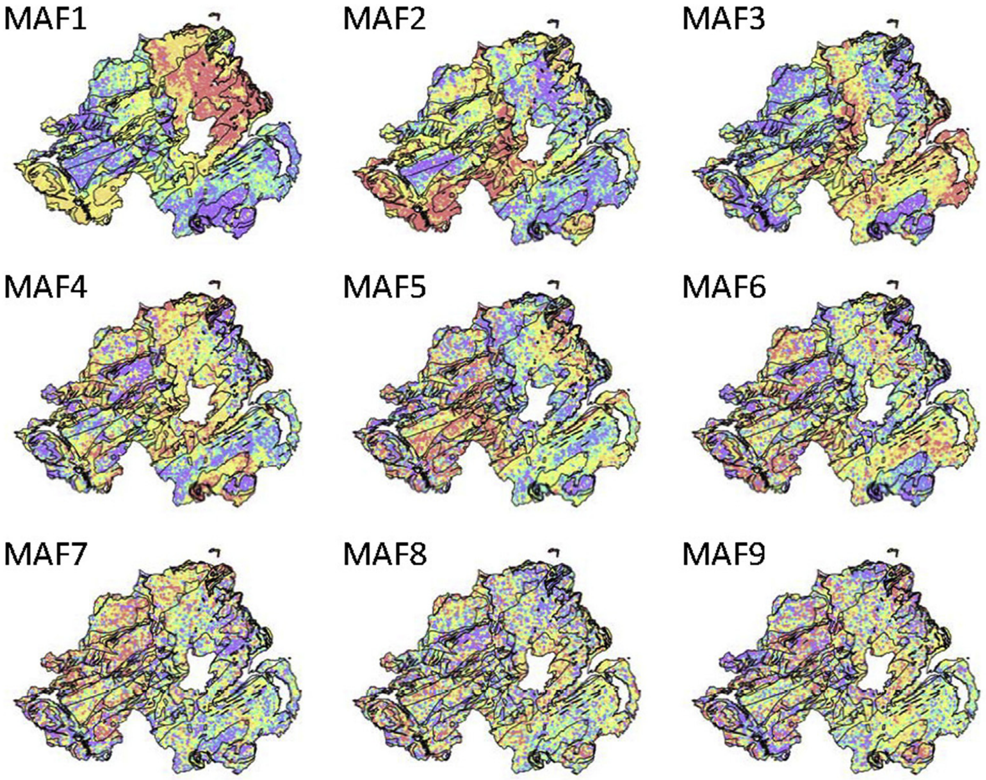

Fig. 5. Maps of MAF Factors top: 1 to 3, center: 4 to 6, bottom:7–9 showing the decrease in spatial continuity, red and blue colours indicate high and low values respectively.

apparent in the semivariograms of the factors based on MAF and UWEDGE comes at the cost of destroying the ordering by decreasing explained variance which is a feature of both PCA (by construction) and RJD. For all four methods the first factor still explains the greatest part of the variance: 50.7% in the case of PCA, 43% for RJD, 36.7% for MAF and 35.1% for RJD.

The destructurization evident in the semivariograms for the MAF and UWEDGE factors is reflected in their score maps, where decreasing scales of high and low value regions and increasing noise are evident, this is exemplified here through the maps of the MAF factors (Fig. 5).

The spatial maps of the MAF factors identify the lithologies reasonably well. The scores of the first two factors (maf1 and maf2) show a similar interpretation as a broad balance between mafic elements and felsic elements, which is related to the contrast of the Paleocene Antrim basalts in the north west with the older rocks across the remainder of Northern Ireland. The intrusive igneous granite and granodiorite are also highlighted (maf1, maf2 and maf3). Carbonates and clastics of different ages are differentiated by the fourth and fifth scores (maf4 and maf 5). The scores of the sixth, seventh and eight factors are less clear but the impact of superficial peat cover can be observed (maf3, maf6, maf7 and maf8).

Spatial maps of the scores of the first two factors for all four methods are shown in Fig. 6. They all highlight the broad balance between mafic elements and felsic elements, which is related to the contrast of the Paleocene Antrim basalts in the north west with the older rocks across the remainder of Northern Ireland. These maps support our earlier observation of greater continuity of the first two factors from MAF and UWEDGE: There is less noise evident in these maps compared to those for RJD and PCA. The greater continuity of the second factor from RJD is also evident.

8. Conclusion

The consideration of spatial aspects via MAF biplots in combination with PCA biplots provides a valuable tool for the exploratory analysis of regionalized compositions. In the case of the Tellus data the use of MAF has led to stronger grouping of samples by underlying characteristics, in the case considered here, lithology and peat. The choice of lag value for computing the covariance matrix of increments seems to have little impact. Consideration of more general approximate diagonalization methods, such as UWEDGE and RJD showed that the additional complexity introduced was not justifiable from a perspective of enhancing results: the MAF and UWEDGE biplots were typically close, leading to the conclusion that the use of MAF suffices to add a spatial perspective to the exploration. Similarly, RJD and PCA biplots were close, as were their screeplots. A combination of PCA and MAF biplots enables the exploration of the features of the data set via two different lenses: one focusing on the ordering in terms of decreasing variance explained, the other (MAF) in terms of decreasing spatial continuity, thus allowing a deeper understanding of geo-spatial aspects. The use of PCA and MAF in combination is particularly attractive, since there are closed form expressions for the MAF and PCA transformations.

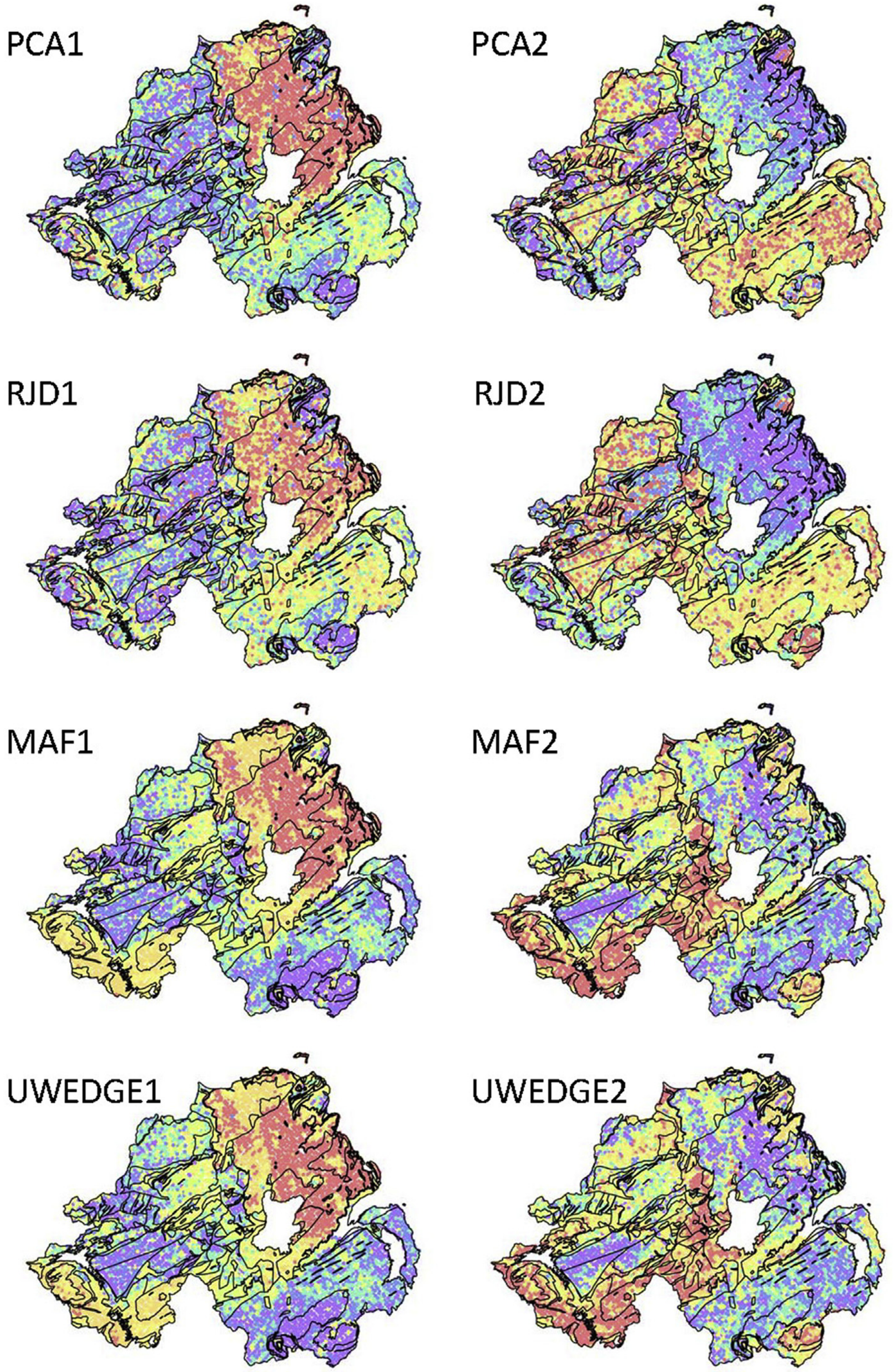

Fig. 6. Maps of factor scores for PCA, RJD, MAF and UWEDGE, left: Factor 1, right: Factor 2, red and blue colours indicate high and low values respectively.

Code availability and data

Calculation routines for the joint diagonalization methods are available in the R-package "gmGeostats" (https://cran.r-project.org/web/packages/gmGeostats/index.html) and a sample code to show the implementation is available in an electronic supplement.

The data used in this study are Regional A Soils (XRF and Aqua Regia Digest) available from http://www.bgs.ac.uk/gsni/tellus/index.html. They were further treated as described in McKinley et al. (2018). The classification by lithological class and peat is as described in Tolosana--Delgado and McKinley (2016).

Author statement

Ute Mueller: Conceptualization, Methodology, Writing – original draft, Formal analysis. Raimon Tolosana Delgado: Conceptualization, Methodology, Software, Formal analysis, Writing – review & editing. E.C. Grunsky: Interpretation of results, Writing – review & editing, Visualization. J.M. McKinley: Interpretation of results, Writing – review & editing, Visualization

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.acags.2020.100044.

References

Aitchison, J., Greenacre, M., 2002. Biplots of compositional data. Appl. Stat. 51, 375–392.

Afsari, B., 2007. What can make joint diagonalization difficult?. In: IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, vol. 3, pp. 1377–1380.

Cardoso, J.K., Souloumiac, A., 1996. Jacobi angles for simultaneous diagonalization. SIAM J. Matrix Anal. Appl. 17 (1), 161–164.

Desbarats, J.A., Dimitrakopoulos, R., 2000. Geostatistical simulation of regionalized pore-size distributions using Min/Max autocorrelation factors. Math. Geol. 32 (8), 919–942.

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

GSNI, 2007. Geological Survey Northern Ireland Tellus Project Overview. https://www.bgs.ac.uk/gsni/Tellus/index.html. (Accessed 7 March 2017).

Gouy-Pailler, C., 2017. jointDiag: Joint Approximate Diagonalization of a Set of Square Matrices. R package version 0.3. https://CRAN.R-project.org/package=jointDiag.

Graffelmann, J., Eeuwijk, F., 2005. Calibration of multivariate scatter plots for exploratory analysis of relations within and between sets of variables in genomic research. Biom. J. 47 (6), 863–879.

Grunsky, E.C., Caritat, P. de, Mueller, U.A., 2017. Using surface regolith geochemistry to map the major crustal blocks of the Australian continent. Gondwana Res. 46, 227–239.

Horn, R.A., Johnson, C.R., 1985. Matrix Analysis. Cambridge University Press, p. 561.

Jiang, R., Li, D., 2015. Simultaneous Diagonalization of Matrices and its Applications in Quadratic Programming. https://arxiv.org/abs/1507.05703v2.

McKinley, J., Grunsky, E., Mueller, U., 2018. Environmental monitoring and peat assessment using multivariate analysis of regional scale geochemical data. Math. Geosci. 50 (2), 235–246.

Miettinen, J., Nordhausen, K., Taskinen, S., 2017. Blind source separation based on Joint diagonalization in R: the packages JADE and BSSasymp. J. Stat. Software 76 (2), 1–31.

Mueller, U.A., Grunsky, E.C., 2016. Multivariate spatial analysis of lake sediment geochemical data; Melville Peninsula, Nunavut, Canada. Appl. Geochem. 75, 247–262.

Pawlowsky-Glahn, V., Burger, H., 1992. Spatial structure analysis of regionalized compositions. Math. Geol. 24 (6), 675–691.

Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2015. Modeling and Analysis of Compositional Data. Wiley, p. 272.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Smyth, D., 2007. Methods Used in the Tellus Geochemical Mapping of Northern Ireland. British Geological Survey Open Report or/07/022.

Switzer, P., Green, A.A., 1984. Min/Max Autocorrelation Factors for Multivariate Spatial Imaging. Stanford University, Palo Alto, USA, p. 14.

Tichavsky, P., Yeredor, A., 2009. Fast approximate joint diagonalization incorporating weight matrices. IEEE Trans. Signal Process. 57, 878–891.

Tolosana-Delgado, R., McKinley, J., 2016. Exploring the joint compositional variability of major components and trace elements in the Tellus soil geochemistry survey (Northern Ireland). Appl. Geochem. 75, 263–276.

Tolosana-Delgado, R., Mueller, U., van den Boogaart, K.G., 2019. Geostatistics for compositional data: an overview. Math. Geosci. 51 (4), 485–526.

van den Boogaart, K.G., Tolosana-Delgado, R., 2013. Analyzing Compositional Data with R. Springer, p. 258.

Young, M., Donald, A., 2013. A Guide to the Tellus Data. Geological Survey of Northern Ireland, Belfast, p. 233.