Math Geosci
DOI 10.1007/s11004-017-9686-x

SPECIAL ISSUE

Environmental Monitoring and Peat Assessment Using Multivariate Analysis of Regional-Scale Geochemical Data

Jennifer M. McKinley[1] · Eric Grunsky[2] · Ute Mueller[3]

Received: 28 December 2016 / Accepted: 4 April 2017

Abstract A compositional multivariate approach was used to analyse regional-scale soil geochemical data obtained as part of the Tellus Project generated by the Geological Survey of Northern Ireland. The multi-element total concentration data presented comprise X-ray fluorescence (XRF) analyses of 6862 rural soil samples collected at 20-cm depth on a non-aligned grid at one site per 2 km^2. Censored data were imputed using published detection limits. Each soil sample site was assigned to the regional geology map, resulting in spatial data for one categorical variable and 35 continuous variables comprised of individual and amalgamated elements. This paper examines the extent to which soil geochemistry reflects the underlying geology or superficial deposits. Since the soil geochemistry is compositional, log-ratios were computed to adequately evaluate the data using multivariate statistical methods. Principal component analysis (PCA) and minimum/maximum autocorrelation factors (MAF) were used to carry out linear discriminant analysis (LDA) as a means to discover and validate processes related to the geologic assemblages coded as age bracket. Peat cover was introduced as an additional category to measure the ability to predict and monitor fragile ecosystems. Overall prediction accuracies for the age bracket categories were 68.4 % using PCA and 74.7 % using MAF. With inclusion of peat, the accuracy for LDA classification decreased to 65.0 and 69.9 %, respectively. The increase in misclassification due to the presence of peat may reflect degradation of peat-covered areas since the creation of superficial deposit classification.

✉ Jennifer M. McKinley
j.mckinley@qub.ac.uk

[1] School of Natural and Built Environment, Queen's University Belfast, Belfast BT7 1NN, UK

[2] Department of Earth and Environmental Sciences, University of Waterloo, Waterloo, ON N2L 3G1, Canada

[3] School of Science, Edith Cowan University, Joondalup, WA 6027, Australia

Published online: 03 May 2017

Keywords Compositional data analysis · Minimum/maximum autocorrelation factors (MAF) · Linear discriminant analysis (LDA) · Log-ratios · Centred log-ratio (clr) · Geochemistry

1 Introduction

A diversity of rock types observed across Northern Ireland represents three basement terranes (Fig. 1a after Mitchell 2004). The Grampian Terrane and associated rocks in the northwest have metamorphic igneous and sedimentary origins spanning the Proterozoic Era. Psammites and semi-pelites are the dominant rock type, with subordinate sandstones and conglomerates. The Midland Valley Terrane hosts Palaeozoic igneous formations and Late Devonian–Early Carboniferous sedimentary rocks. Rock types comprise red sandstones, limestone and mudstones with less common conglomerates. The Southern Uplands–Down–Longford Terrane consists of granitic igneous intrusives and Lower Palaeozoic Ordovician and Silurian marine sedimentary rocks (lithic arenites and sandstones). Palaeogene flood basalts and lava-derived sedimentary clays cover a large portion of the Midland Valley and Grampian basement rocks in the northeast of the country. The advance of ice sheets and their meltwaters over the last 100,000 years has left at least 80 % of the bedrock covered by superficial deposits, including glacial till and post-glacial alluvium and peat (Fig. 1b). Monitoring peat coverage has become important in calculating soil carbon stocks due to the relatively high carbon density of peat and organic-rich soils. This is particularly important for Ireland (and other Northern European countries), where some 16 % of the land surface is covered by peat bog. In Northern Ireland, previous work has estimated the total amount of carbon stored in vegetation to be 4.4 Mt compared with 386 Mt stored within soils such as peat (Cruickshank et al. 1998).

A number of comprehensive regional and national soil sampling programmes have been completed across the UK, including the British Geological Survey (BGS) G-BASE survey and regional soil surveys for Northern Ireland (Jordan et al. 2001). These

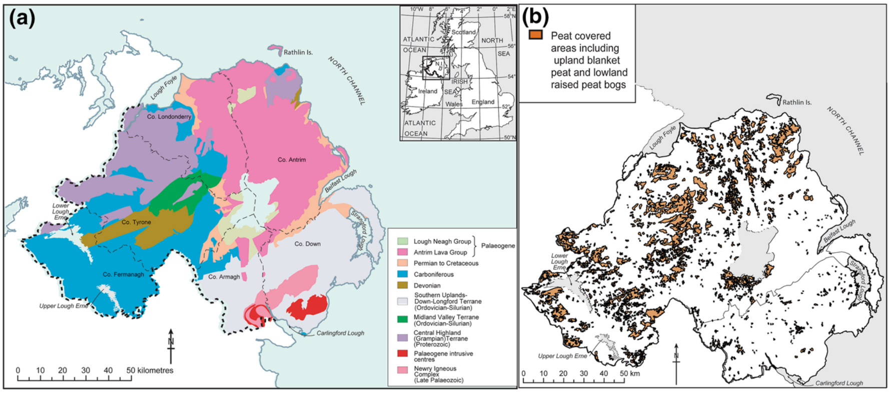

Fig. 1 **a** Simplified geology of Northern Ireland (after Mitchell 2004); **b** map showing post-glacial peat-covered areas

have been used to assess baseline element concentrations in soils and normal background concentrations of contaminants (for example Ander et al. 2013). The Northern Ireland Tellus Survey (GSNI 2007; Young and Donald 2013) included a ground-based geochemical survey in which 6862 rural soil samples were collected between 1994 and 2006. Samples were collected on a grid of one sample site every 2 km^2, with soils being collected at depths of 20 and 50 cm. The samples were analysed by X-ray fluorescence spectrometry (XRFS) for 60 elements and inorganic compounds. Tellus Survey field methods and analytical methodology are described in Smyth (2007) and Young and Donald (2013). Soils and parent geology across Northern Ireland are typical for the geological and pedological conditions across the UK (Jordan et al. 2001) as well as for several parts of Northern Europe, therefore the Tellus data set provides the basis for a comprehensive study with relevancy not only for the whole of the UK, but also Northern Europe. This paper explores the extent to which soil geochemistry can be used to classify the underlying geology and moreover differentiate superficial deposits such as peat.

2 Methods

Previous work by Grunsky et al. (2014) has demonstrated the effectiveness of applying log-ratio transforms, principal component analysis (PCA) and linear discriminant analysis (LDA) to differentiate lithologies and glacial processes. This study involves the use of both process discovery and validation methods within a compositional multivariate approach to analyse the rural Tellus soil data sampled at 20-cm depth. Many geochemical data sets contain values that are reported at less than the lower limit of detection, and these values are generally termed ‘censored’ (Grunsky 2010). For this study, published detection limits (Young and Donald 2013) were used to impute censored data; this resulted in 47 elements (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, Sc, Se, Si, Sm, Sn, Sr, Th, Ti, Tl, U, V, W, Y, Yb, Zn, Zr and total C) available for initial analysis. Elements that occupy the same sites and/or substitute for other elements in mineral structures can display collinearity when estimating statistical moments. In many cases, the elements behave in identical ways and these elements can be summed to a single variable without any loss of information or structure in the data. This is often the case for rare-earth elements, and the degree of collinearity can be observed in the loadings of the variables in a principal component biplot. An application of clr-based PCA showed that several elements were near-collinear and prompted an amalgamation of the data based on the following proxies: light rare-earth elements (LREE), La, Ce, Nd along with Th; heavy rare-earth elements (HREE), Yb and Hf; elements associated with mafic minerals (Mafic), Fe, V, Cr and Sc; alkali elements associated with feldspars (Ba_Na_K_Rb), Ba, Na, K and Rb; and lastly elements associated with a feldspar silicate framework (Al_Ga), Al and Ga. This reduced the total number of elements and amalgamated elements to 35, and these were the basis of all further analysis.

Each soil sample site was assigned to the regional geology map and then age brackets provided by Geological Survey Northern Ireland (GSNI) using the dominant

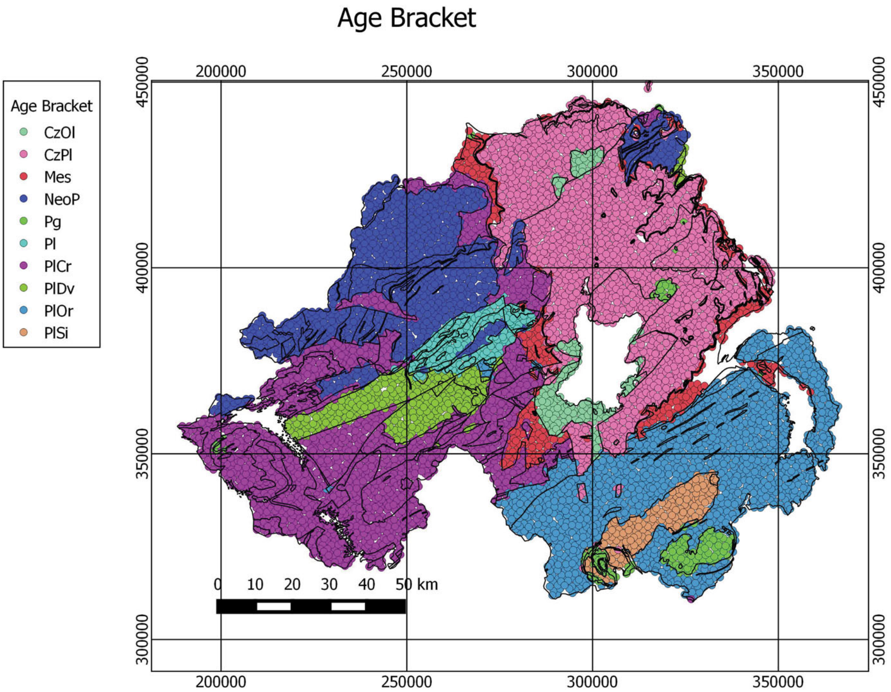

Fig. 2 Map showing samples tagged with age bracket classification. Full age bracket classes are provided in Table 1

age bracket for the map polygon (Fig. 2; Table 1). A process discovery approach (Grunsky 2010) involves use of unsupervised multivariate methods such as principal component analysis (PCA). An essential part of the process discovery phase is a suitable choice of transform to overcome the problem of closure. Therefore, to account for the requirements of compositional data analysis and avoid closure issues, the geochemical data were transformed using centred log-ratios (clr, Aitchison 1986).

Analysis of variance (ANOVA) is a commonly used method for determining the statistical distinctiveness of groups based on the observations and variables (Grunsky and Kjarsgaard 2016). The results will determine whether or not there is sufficient statistical separation of the groups to undertake a classification such as linear discriminant analysis (LDA, Venables and Ripley 2002). Garrett (1989) suggests that the ratio of observations to the total number of variables and groups should be a minimum of 3:1 and ideally greater than 10:1 (Garrett 1989). To ensure that this requirement is fulfilled, data compression was applied. Two potential compression methods are the use of principal components and minimum/maximum autocorrelation (MAF) factors. The use of principal components to represent the variables is well established. However, in this approach, spatial relationships between the variables are ignored. One way to address this failure is to use minimum/maximum autocorrelation factors instead (Mueller and Grunsky 2016). These are derived from the given data by first performing a principal component analysis and then a rotation based on the spectral

Table 1 Geological age bracket classification

Code	Number of tagged Tellus samples	Geological age bracket
CzOl	151	Cenozoic, Palaeogene, Oligocene
CzPl	1671	Cenozoic, Palaeogene, Palaeocene
Mes	330	Mesozoic, Cretaceous, Jurassic, Triassic
NeoP	978	Neoproterozoic, Dalradian Supergroup
Pg	93	Palaeogene
Pl	146	Palaeozoic, Lower–Middle Ordovician, Lower Palaeozoic
PlCr	1526	Palaeozoic, Carboniferous, Mississippian, Pennsylvanian
PlDv	296	Palaeozoic, Upper Devonian–Carboniferous, Lower Devonian, Middle Devonian, Devonian
PlOr	1289	Palaeozoic, Middle–Upper Ordovician, Ordovician–Silurian, Permian, Silurian
PlSi	163	Caledonian (Silurian–Devonian)

decomposition of a variance–covariance matrix of the PCA-transformed and scaled data at a lag (h), typically chosen to represent the sample spacing (Desbarats and Dimitrakopoulos 2000). The MAF transformation can also be seen as the solution of a generalised eigenvalue problem (Bandarian and Mueller 2008), and this is the approach taken here. For both PCA and MAF, the decision on the number of factors to use to represent the data is determined experimentally based on the behaviour of the F-values in the ANOVA.

Following the data compression step, n-fold cross-validation LDA can be used to derive posterior probabilities of class membership. Since these probabilities are compositional (van den Boogaart and Tolosana-Delgado 2006; Egozcue and Pawlowsky-Glahn 2016), an appropriate transformation is required prior to their interpolation in order to ensure that estimates are non-negative and adhere to the constant sum constraint. In this study, ordinary co-kriging was applied to the isometric log-ratio (ilr)-scores of the posterior probabilities, and this step was followed by back-transformation to interpolated probabilities.

3 Results and Discussion

Based on the geological age bracket classification, without peat, PCA showed that 72 % of the variation was determined by the first four principal components (PCs), implying ‘significant’ structure in the data. Analysis of variance showed that only 10 PCs were necessary to classify the soil geochemical data. To consider an improvement over PCA that uses the spatial relationships of the data, a classification based on MAF analysis was undertaken using the first eight dominant factors (Fig. 3a). Twenty-fold cross-validation LDA, using PCA and MAF, resulted in overall classification accuracy of 68.4 % (PCA) and 74.7 % (MAF) for the geological age bracket classification. More specifically, except for PlDv and PlOr, the use of MAF and cross-validation LDA resulted in improvements in both recall (quotient of total number of correctly

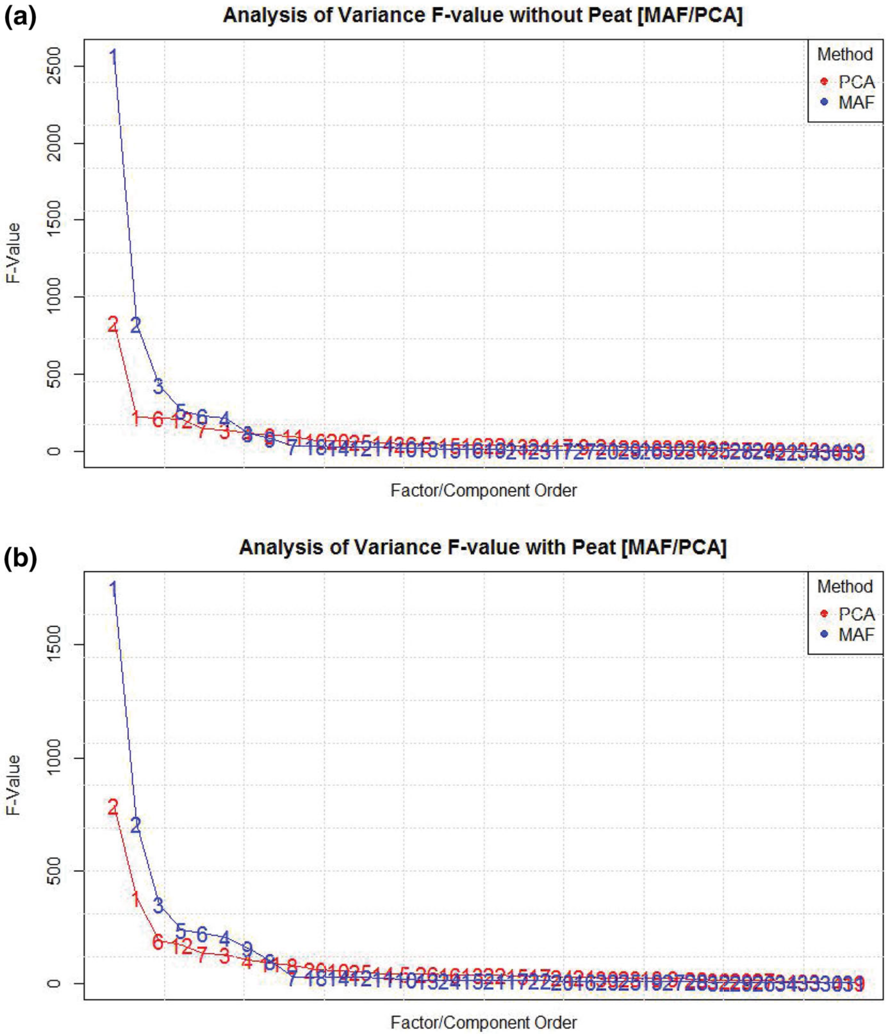

Fig. 3 ANOVA for soil geochemistry age bracket classification using PCA and MAF analysis: **a** without peat and **b** including peat in the classification

classified sites and total number of sites in the class) and precision (quotient of total number of correctly classified sites and total number of sites allocated to the class), particularly so for the classes with low number of sites (Table 2; Fig. 4).

The mapped posterior probabilities for six of the age bracket classes (CzPl, NeoP, PlCr, PlDv, PlOr and PlSi) are shown in Fig. 5. In the case of the CzPl and PlSi age brackets (Fig. 5a, f), the locations of non-zero probability are quite well constrained and the probabilities of occurrence are generally very high. The Palaeocene Antrim Lava Group, within the CzPl age bracket class, dominates the landscape of the North West of Ireland comprising the largest remnant of the North Atlantic Igneous Province (Cooper

Table 2 Recall and precision for classification results based on PCA and MAF

Code	Number of sites	PCA			MAF		
		% Classified correctly	Precision	Recall	% Classified correctly	Precision	Recall
CzOl	151	3.97	0.38	0.04	13.91	0.51	0.14
CzPl	1671	86.12	0.91	0.86	87.07	0.92	0.87
Mes	330	8.18	0.25	0.08	36.06	0.33	0.36
NeoP	978	66.36	0.61	0.66	83.64	0.71	0.84
Pg	93	69.89	0.71	0.7	73.12	0.73	0.73
Pl	146	0.00	0	0	9.59	0.3	0.1
PlCr	1526	71.76	0.56	0.72	70.38	0.7	0.7
PlDv	296	7.09	0.54	0.07	28.38	0.52	0.28
PlOr	1289	85.10	0.7	0.85	90.15	0.78	0.9
PlSi	163	90.18	0.69	0.9	90.18	0.84	0.9

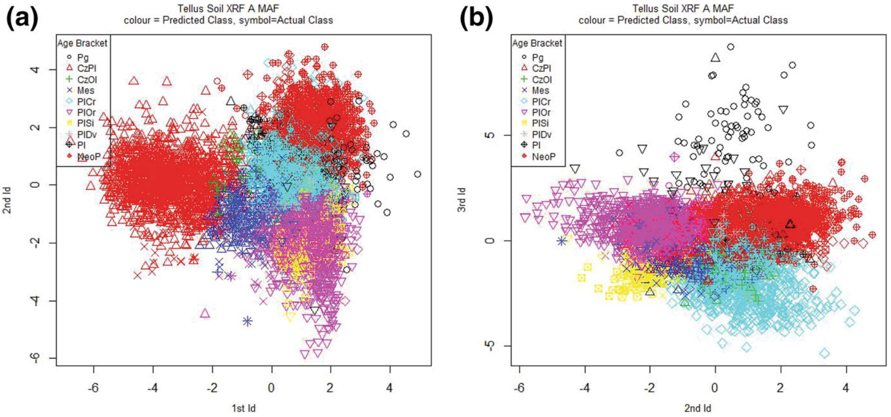

Fig. 4 Linear discriminant analysis (LDA) using MAF analysis for soil geochemistry age bracket classification (without peat) using the first eight dominant factors. *Colour* shows the predicted class, while the *symbol* shows the actual class

and Mitchell 2013). Petrographical studies demonstrate the consistent geochemical signature of the fine-grained, olivine basalt lava formations. This helps to explain the well-constrained and high probabilities of occurrence for this age bracket. Likewise the PlSi class (Fig. 5f) closely constitutes the Caledonian granodiorite igneous complex. In contrast, the PlDv class (Fig. 5d), which is also quite well constrained, has much lower probabilities of occurrence. This class, which has 296 tagged samples, represents a broad range of Palaeozoic rocks representing a fault-bounded non-marine sedimentary basin sequence of interbedded sandstones, mudstones and conglomerates. The spatial maps of co-kriging estimates of the posterior probabilities, shown for the same age

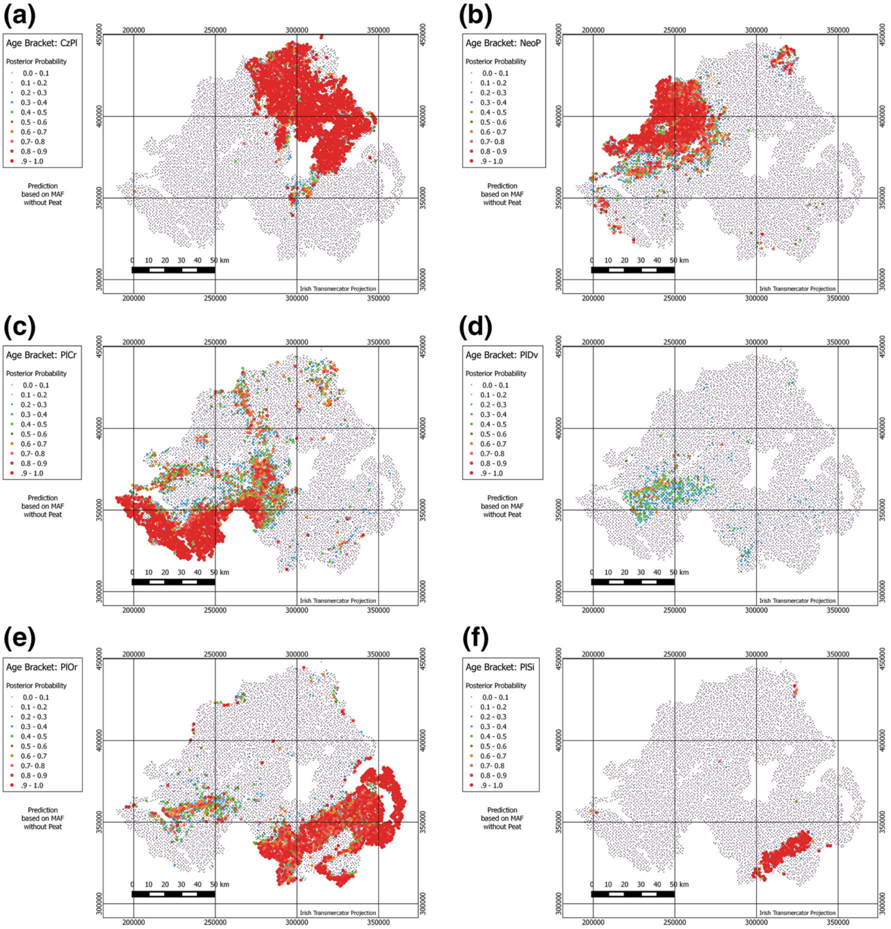

Fig. 5 Posterior probabilities for age bracket classes **a** CzPl, **b** NeoP, **c** PlCr, **d** PlDv, **e** PlOr and **f** PlSi. Geological age bracket classes provided in full in Table 1

bracket classifications (Fig. 6), consolidate the findings from the mapped posterior probabilities (Fig. 5).

3.1 Peat Assessment

Since the aim of this study is to elucidate the relationship between soil geochemistry and post-glacial deposits for environmental monitoring, the next stage of the analysis explored whether peat cover could be predicted from the classification. For this, the geological age bracket designation was adapted to include the presence of peat based on GSNI superficial deposit polygons (Fig. 1b) and linear discriminant analysis (LDA) undertaken. As was the case for the classification based on the age brackets only, 10 PCA factors and 8 MAF factors were sufficient for the classification via LDA (Fig. 3b).

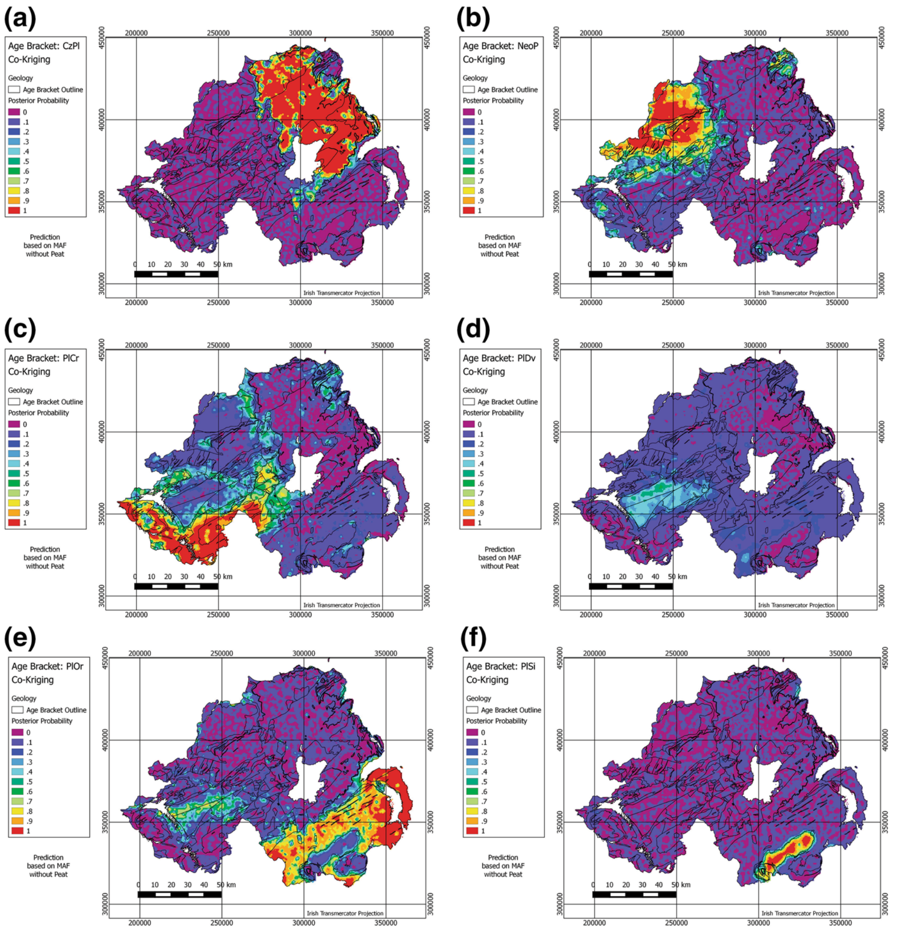

Fig. 6 Spatial maps of co-kriging estimates of the posterior probabilities for **a** CzPl, **b** NeoP, **c** PlCr, **d** PlDv, **e** PlOr and **f** PlSi. Geological age bracket classes provided in full in Table 1

Using LDA, the presence of peat was clearly differentiated from the other classifications (Fig. 7). The classification accuracies for the age bracket classes, including peat, were 65.0 % (PCA) and 69.9 % (MAF). The prediction accuracies for both data types are lower than for those where peat was not included as a class. The presence of peat results in more classification confusion, most likely because the peat composition retains geochemical characteristics of the underlying lithology and the organic composition of the peat itself. The plotted MAF posterior probabilities (Fig. 8) demonstrate a good match between the reported peat areas (Fig. 1b) and the highest probability for peat. However, there are areas of mapped peat where the estimated probabilities of occurrence of peat are low. The explanation for this misclassification of peat may be twofold: these areas may indicate degradation of peat-covered areas since the creation of the superficial deposit classification, or a further refinement in the classification of peat is required. Irish peatland is divided into blanket peatland (approximately 85 %)

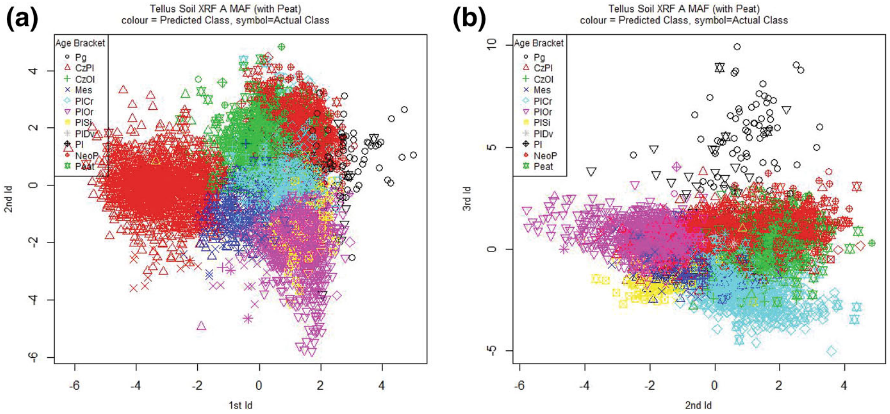

Fig. 7 Linear discriminant analysis (LDA) including the presence of peat (shown as *green symbols*) using MAF analysis. *Colour* shows the predicted class, while *symbols* show the actual class

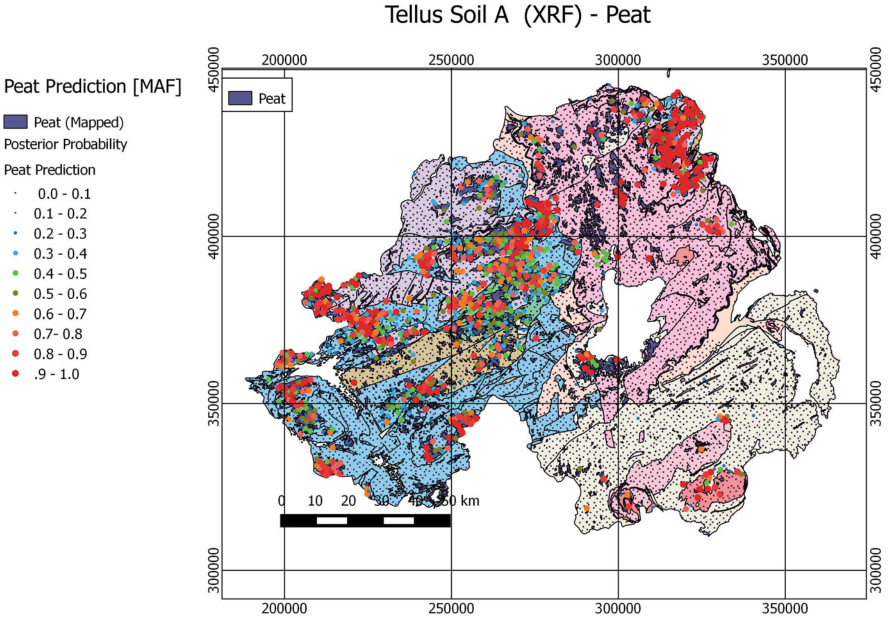

Fig. 8 Plotted MAF posterior probabilities overlain on areas of reported peat

and raised peat bogs (approximately 15 %; Tomlinson and Davidson 2000). Blanket bogs typically form on gentle slopes within upland regions [>315 m above sea level (SL); Hamilton 1982]. The distribution of blanket bogs is more spatially continuous and associated with areas of high precipitation (rainfall exceeding 1200 mm). Raised bogs develop primarily in lowland areas (<200 m above SL; Wheeler and Shaw 1995) where accumulating peat in fens becomes isolated from the groundwater supply. This process of accumulation gradually forms a dome of ombrogenous peat above the fen, giving raised bogs a distinct topography, with steep margins to the main bog expanse. Raised bogs are more limited in extent and occur as isolated features. Prediction of peat-covered areas using MAF analysis methods, which use the spatial relationships of the data, has been more successful in predicting the more extensive upland blanket bogs than lowland raised bogs.

4 Conclusions

Compositional multivariate techniques, PCA and MAF analysis methods were used to determine the influence of underlying geology on the soil geochemistry signature. The approach was explored for environmental monitoring of peat to ascertain whether peat cover could be predicted from the classification. Using LDA, the presence of peat was clearly differentiated from the other lithological classifications. Moreover, the prediction accuracy for LDA classification improved using MAF analysis. In an attempt to reduce the number of areas of misclassification of peat, further work will examine the influence of underlying lithologies on elemental concentrations in peat composition and the effect of this in classification analysis.

Acknowledgements The Tellus Project was carried out by GSNI and funded by The Department for Enterprise, Trade and Investment (DETINI) and The Rural Development Programme through the Northern Ireland Programme for Building Sustainable Prosperity.

References

Aitchison J (1986) The statistical analysis of compositional data. Chapman and Hall, New York

Ander EL, Johnson CC, Cave MR, Palumbo-Roe B, Nathanail CP, Lark RM (2013) Methodology for the determination of normal background concentrations of contaminants in English soil. Sci Total Environ 454–455:604–618. doi:10.1016/j.scitotenv.2013.03.005

Bandarian EM, Mueller UA (2008) Reformulation of direct minimum/maximum autocorrelation factors as a generalised Eigenvalue problem. In: Ortiz JM, Emery X (eds) Geostats 2008, vol 2. Universidad De Chile, Santiago, Chile, pp 1173–1178

Cooper M, Mitchell W (2013) Geology. In: Young ME, Donald AE (eds) A guide to the tellus data 2013. Geological Survey of Northern Ireland, Belfast, pp 5–20

Cruickshank MM, Tomlinson RW, Devine PM, Milne R (1998) Carbon in the vegetation and soils of Northern Ireland. Biol Environ Proc R Ir Acad 98B(1):9–21

Desbarats AJ, Dimitrakopoulos R (2000) Geostatistical simulation of regionalized pore-size distributions using min/max autocorrelation factors. Math. Geol. 23(8):919–941

Egozcue JJ, Pawlowsky-Glahn V (2016) Changing the reference measure in the simplex and its weighting effects. Austrian J Stat 45:25–44. doi:10.17713/ajs.v45i4.126

Garrett RG (1989) A robust multivariate allocation procedure with applications to geochemical data. In: Agterberg F, Bonham-Carter GF (eds) Statistical applications in earth sciences, geological survey paper 89-9: 309–318

Geological Survey Northern Ireland (GSNI) (2007) Tellus project overview. https://www.bgs.ac.uk/gsni/Tellus/index.html. Accessed 7 Mar 2017

Grunsky EC, Kjarsgaard BA (2016) Recognizing and validating structural processes in geochemical data. In: Martin-Fernandez JA, Thio-Henestrosa S (eds) Compositional data analysis, vol 187. Springer proceedings in mathematics and statistics, pp 85–116, 209 pp. doi:10.1007/978-3-319-44811-4_7

Grunsky EC, Mueller UA, Corrigan D (2014) A study of the lake sediment geochemistry of the Melville Peninsula using multivariate methods: applications for predictive geological mapping. J Geochem Explorn 141:15–41. doi:10.1016/j.gexplo.2013.07.013

Grunsky EC (2010) The interpretation of geochemical survey data. Geochem Explor Environ Anal 10(1):27–74

Hamilton AC (1982) Peatland. In: Cruickshank JG, Wilcock DN (eds) Northern Ireland environment and natural resources. QUB, Belfast, pp 185–206

Jordan C, Higgins A, Hamill K, Cruickshank JG (2001) The soil geochemical atlas of Northern Ireland. Department of Agriculture and Rural Development, NI

Mitchell W (2004) The geology of Northern Ireland: our natural foundation. Geological Survey of Northern Ireland, Belfast

Mueller UA, Grunsky EC (2016) Multivariate spatial analysis of lake sediment geochemical data. Melville Peninsula, Nunavut, Canada. Appl Geochem 75:247–262. doi:10.1016/j.apgeochem.2016.02.007

Smyth D (2007) Methods used in the Tellus geochemical mapping of Northern Ireland. British geological survey open report or/07/022

Tomlinson RW, Davidson L (2000) Estimates of carbon stores in four Northern Irish lowland raised bogs. Suo 51:169–179

van den Boogaart KG, Tolosana-Delgado R (2006) Compositional data analysis with the R. In: Buccianti A, Mateu-Figueras G, Pawlowsky-Glahn V (eds) Compositional data analysis: from theory to practice. The Geological Society, London

Venables WN, Ripley BD (2002) Modern applied statistics with S. Springer. ISBN 0-387-95457-0

Wheeler BD, Shaw SC (1995) Restoration of damaged peatlands. HMSO, London

Young ME, Donald AE (2013) A guide to the Tellus data. Geological Survey of Northern Ireland, Belfast