Chapter 17
Analysis of the United States Portion of the North American Soil Geochemical Landscapes Project—A Compositional Framework Approach

E. C. Grunsky, L. J. Drew and D. B. Smith

Abstract A multi-element soil geochemical survey was conducted over the conterminous United States from 2007–2010 in which 4,857 sites were sampled representing a density of 1 site per approximately 1,600 km^2. Following adjustments for censoring and dropping highly censored elements, a total of 41 elements were retained. A logcentred transform was applied to the data followed by the application of a principal component analysis. Using the 10 most dominant principal components for each layer (surface soil, A-horizon, C-horizon) the application of random forest classification analysis reveals continental-scale spatial features that reflect bedrock source variability. Classification accuracies range from near zero to greater than 74% for 17 surface lithologies that have been mapped across the conterminous United States. The differences of classification accuracy between the Surface Layer, A- and C-Horizons do not vary significantly. This approach confirms that the soil geochemistry across the conterminous United States retains the characteristics of the underlying geology regardless of the position in the soil profile.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-319-78999-6_17) contains supplementary material, which is available to authorized users.

E. C. Grunsky (✉)
Department of Earth and Environmental Sciences, University of Waterloo, Waterloo, ON, Canada
e-mail: egrunsky@gmail.com

E. C. Grunsky
China University of Geosciences, Beijing, China

L. J. Drew
United States Geological Survey, Reston, VA, USA

D. B. Smith
United States Geological Survey, Denver, CO, USA

B. S. Daya Sagar et al. (eds.), *Handbook of Mathematical Geosciences*,
https://doi.org/10.1007/978-3-319-78999-6_17

17.1 Introduction

A continental-scale soil geochemical survey was conducted over the conterminous United States from 2007 to 2010 by the U.S. Geological Survey (Smith et al. 2011, 2012, 2013, 2014). The survey collected samples at 4857 sites (Fig. 17.1), representing a density of 1 site per approximately 1600 km^2. The sampling protocol included, at each site, a sample from a depth of 0–5 cm (referred to as the surface soil for the remainder of this paper), a composite of the soil A horizon (the uppermost mineral soil), and a sample from the soil C horizon (generally the partially weathered parent material). If the top of the C horizon was at a depth greater than 1 m, a sample over a 20 cm interval was collected at a depth of approximately 1 m.

Studies on the geochemistry of two transects (east-west and north-south) across the United States and Canada, conducted as pilot studies in preparation for the continental-scale survey (Smith 2009; Smith et al. 2009) showed variability of soil geochemistry and mineralogy along both directions (Garrett 2009; Eberl and Smith 2009; Woodruff et al. 2009). As well, Drew et al. (2010) studied the two transects and demonstrated that the geochemical variability of soil is also closely associated with ecoregions (CEC 1997), which reflect continental scale features such as soil, landform, major vegetation types and climate. These studies indicate that the soil geochemistry is useful for mapping both geological and ecological domains.

Soil geochemistry, from a geological context, reflects a range of mineralogy, as a function of weathering of different parent materials, along with organic content due to biological activity. Ideally, soil geochemistry will represent underlying parent material and processes associated with the modification of those parent materials through comminution, weathering, ground water activity and biogenic processes. Grunsky et al. (2012, 2014) smf Mueller and Grunsky (2016) demonstrated that the

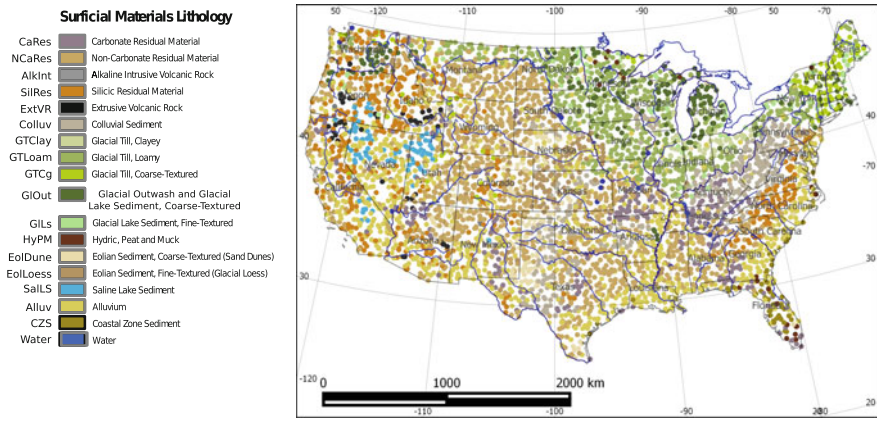

Fig. 17.1 Soil sample sites over the conterminous United States. Samples were taken at the (0–5) cm layer, the A- and C-horizons

geochemistry of lake sediment and glacial till in northern Canada can be used to predict the underlying lithologies. As part of the North American Soil Geochemistry Landscape Project (Smith et al. 2009), Grunsky et al. (2013) used soil geochemistry collected over the Maritime Provinces of Canada and the northeast United States to demonstrate that A-, B- and C-horizon soils geochemistry is useful for mapping the underlying lithologies. More recently, Grunsky et al. (2017) have shown that geochemistry of surficial soils can identify and classify underlying crustal blocks across the Australian continent, even after extended periods of weathering, transport and reworking.

The approach is based on the use of training sets of representative lithologies. Unfortunately, there are no continental-scale lithologic maps or representative training sets which can be used for predictive bedrock lithologic mapping in Canada or the United States. Sayre et al. (2009) classified the land surface of the conterminous United States according to surficial materials lithology, terrestrial ecosystems and isobioclimate. Isobioclimatic zones were subdivided into thermotypes, (temperature) and ombrotypes (moisture). It follows that soil geochemistry is a proxy for processes controlled by climatic factors. A key question that arises from this is can any of these processes be identified uniquely in the soil geochemistry and, if so, how can these processes be identified in terms of spatial continuity and distinctive chemistry? Drew et al. (2010) studied two transects across the US and demonstrated that the soil geochemistry is closely tied to zones that define the terrestrial ecosystems intersected by these transects. The objective of the current study is to address this question through the use of multivariate statistical analysis and Bayesian-based classification in conjunction with geostatistical methods that accurately describe processes in terms of distinctive geochemistry and spatial continuity.

17.2 Methods

17.2.1 Sampling and Analysis

The soil samples were analysed for geochemistry and mineralogy as described by Smith et al. (2011, 2012, 2013, 2014). The samples were air-dried and sieved to <2 mm after which the material was crushed in a ceramic mill prior to chemical analysis. Concentrations of Ag, Al, Ba, Be, Bi, Ca, Cd, Ce, Co, Cr, Cs, Cu, Fe, Ga, In, K, La, Li, Mg, Mn, Mo, Na, Nb, Ni, P, Pb, Rb, S, Sb, Sc, Sn, Sr, Te, Th, Ti, Tl, U, V, W, Y, Zn in all the soil samples (14,434) were determined using a near-total digestion using HCl-HNO_3-$HClO_4$-HF followed by inductively coupled plasma-mass spectrometry and inductively coupled plasma-atomic emission spectrometry. Mercury values were obtained using cold-vapor atomic absorption spectrometry following dissolution in a mixture of HCl and HNO_3 and Se was determined by hydride-generation atomic absorption spectrometry (HGAAS)

following dissolution in a mixture of HNO_3, HF, and $HClO_4$. Arsenic was also determined by HGAAS following fusion in a mixture of sodium peroxide and sodium hydroxide at 750 °C. Total carbon was determined by combustion. Smith et al. (2013) provides details on the analytical methods and quality control protocols. Silicon was not determined.

All A-horizon and C-horizon samples (9575) were analysed by X-ray diffraction, and the percentages of major mineral phases were calculated using a Rietveld refinement method. Splits of the <2 mm fraction were used for analysis. Complete details of the technique and quality control protocols are provided in Smith et al. (2013).

17.2.2 Data Screening and the Compositional Nature of Geochemical Data

Geochemical analyses require screening and adjustment prior to any application of statistical methods and interpretation. A generalized sequence of data screening and adjustment strategies is documented in Grunsky (2010). The data were evaluated and analysed using the R programming and statistical environment (R Core Team 2013).

Major element concentrations, reported as percentages, were converted to ppm, by multiplying the values by a factor 10,000. Summary statistics for the data are given in Smith et al. (2013). The data were screened to determine the number of values that were reported at less than the lower limit of detection. Data that are reported at less than the lower limit of detection are termed as "censored". Censored data, when used in the application of statistical procedures, can influence estimates of mean and variance and therefore a replacement value that accurately reflects an estimate of the true mean is preferred. Furthermore, geochemical data are, by definition, compositions and as such the issue of closure becomes important (Aitchison 1986). Egozcue et al. (2003) describe various transformations that assist in evaluating data that are constrained by the effect of closure. For censored geochemical data, replacement values can be determined using the several methods based on maximum likelihood estimates of replacements values (Palarea-Albaladejo et al. 2014). Elements in which >80% of the values were censored were dropped from further evaluation, which included Ag, Cs and Te.

The data were also screened for sample sites where a large number of elements were reported at less than the lower limit of detection (<LLD). In the surface soil, 8 sites were found to have more than 25 elements reported at <LLD (3 from Florida). For the A horizon, 2 sites, all from Florida, were found to have more than 25 elements reported at <LLD. For the C horizon, 3 sample sites, in Florida, were found to more that have more than 25 elements reported at <LLD. These sites were dropped from further evaluation.

Summary statistics for the elements are provided by Smith et al. (2013, 2014). The remaining 43 elements: Al, As, Ba, Be, Bi, total C, Ca, Cd, Ce, Co, Cr, Cs, Cu, Fe, Ga, Hg, In, K, La, Li, Mg, Mn, Mo, Na, Nb, Ni, P, Pb, Rb, S, Sb, Sc, Se, Sn, Sr, Th, Ti, Tl, U, V, W, Y, Zn were then evaluated for the estimate of replacement values for those results that were reported at less than the lower limit of detection. The method of nearest neighbour replacement estimates (R package: zCompositions, function **lrEM**) was used on the censored data (Palarea-Albaladejo et al. 2014). The adjusted data were then used for subsequent multivariate statistical analysis.

17.2.3 Integration of Land Surface Parameters with Soil Geochemistry

Land surface maps of the conterminous United States (Sayre et al. 2009) were used to test the effectiveness of the soil geochemistry for revealing information on surficial materials lithology, terrestrial ecosystems and isobioclimate. Isobioclimatic zones were subdivided into thermotypes, (temperature) and ombrotypes (moisture). In this study, only the surface lithologies were studied in further detail. The results of the evaluation of the soil geochemistry in the context of terrestrial ecosystems, thermotypes and ombrotypes will be provided at a later time.

The maps were obtained as raster images with a pixel resolution of 1 km and a geodetic projection of decimal degrees using the North American Datum of 1983 (NAD83). These images were re-projected to the Lambert Conformal Conic projection using the following parameters (Spheroid—GRS 1980; Central Meridian: 96° West; Standard Parallels of 32° and 44°; Latitude of Origin: 38°; False Eastings and Northings of 0 m). This projection was used throughout the study.

The Quantum Geographic Information Systems (QGIS) (QGIS Development Team 2016) was used for the integration of various data sources and the geospatial rendering of the results. Within QGIS, two procedures were used from the Geospatial Data Abstraction Library (GDAL) procedure, "**warp (reprojection)**" and "**point sampling tool**". The map images were initially re-projected to the Lambert Conformal Conic (**lcc**) projection listed above using the "**warp**" procedure. The point dataset of the geochemical sampling sites were also reprojected from latitude/longitude coordinates to the **lcc** projection. The **lcc** image of the surface lithology was then sampled at the geochemical site coordinates using the "**point sampling tool**" and the surface lithology value was integrated into the geochemical database. This methodology was carried out for the other land surface maps (terrestrial ecosystems, surface lithologies, thermotypes and ombrotypes). The values of these features were integrated into the soil geochemistry dataset for further evaluation. It should be noted that the maps produced by Sayre et al. (2009) are generalizations and expressed at a resolution of 30 m (landforms, topographic moisture), 1 km (biogeographic regions) and 15 km for the surface lithology. It is

possible that the class defined at any given point on the maps produced by Sayre does not correspond with the surface lithology, biogeographic, landform or topographic classes that were encountered during the soil survey sampling program.

For geospatial rendering purposes (interpolation), the Level 1 Ecology map of the conterminous United States was used to create a grid with a cell size of 40 km × 40 km.

Interpolation of principal component scores, posterior probabilities and measures of typicality were carried out using a geostatistical framework. The gstat package (Pebesma 2004) was used to generate and model semi-variograms with sufficient parameters to generate interpolated images through kriging. The cell size used for image interpolation was chosen as 40 km, the approximate spacing of the site sampling locations.

17.2.4 Process Discovery—Empirical Investigation of Soil Geochemistry

After screening the data for detection limit issues and missing values, the geochemical data were then subjected to an empirical investigation in which the assumptions about the data are minimal. To deal with the effect of closure, the data for 41 elements (Al As Ba Be Bi Ca Cd Ce Co Cr Cu Fe Ga Hg In K La Li Mg Mn Mo Na Nb Ni P Pb Rb S Sb Sc Se Sn Sr Th Ti Tl U V W Y Zn) were log-centred transformed after which a principal component analysis (PCA) was carried out using the methodology of Zhou et al. (1983) and Grunsky (2001). PCA was carried out on the entire set of multi-element data for the surface soil, the A and C horizons combined. PCA was also carried out on the multi-element data individually for the surface soil, A and C horizons. The rationale for this is based on enhancement of the multi-element signature for each layer rather than a principal component signature derived from the combined layers. The principal component biplots and corresponding maps of the component scores were subsequently generated for the surface soil, the A- and C-horizons independently. The biplots and interpolated maps provide insight into the orthogonal linear relationships that can reflect dominant geochemical processes that are influenced by mineral stoichiometry. The three soil layers were evaluated together in order to show any possible relationships between the two soil horizons (A and C) and the surface soil layer. To assist with insight into processes that influence the relationship of the elements and patterns of the scores of the observations, the loadings of the elements were coloured according to the classification of Goldschmidt (1937) into lithophile. siderophile or chalcophile affinity Elements associated with the atmophile affinity were not considered in this study.

17.2.5 Process Validation—Modelled Investigation of Soil Geochemistry

Using the classified information derived from the land surface maps of Sayre et al. (2009), the geochemical data were used to establish the ability to predict these classifications using a cross-validation approach in which the data are repeatedly sub-sampled as part of the classification process.

Previous studies (Grunsky et al. 2012, 2014) demonstrated that the use of multivariate statistical methods was able to classify bedrock lithologies based on lake sediment and glacial till geochemical data using discriminant analysis. The methodology employed the results of principal component analysis (described above), followed by an analysis of variance and the application of linear discriminant analysis (Venables and Ripley 2002) to determine which principal components were best at classifying and predicting the bedrock lithologies. This approach relies on having a sufficient number of degrees of freedom and homogeneity of covariance between the classes of the training sets. An alternative to linear discriminant analysis is quadratic discriminant analysis (Venables and Ripley 2002), which compensates for the classes where the condition of homogeneity of covariance cannot be met. The results of applying these methods includes measures of posterior probability in which each site is assigned a measure of probability of belonging to each of the classes and the class with the highest posterior probability is assigned to that site. Posterior probabilities are also compositions, as the sum of the probabilities for all of the classes for each site must sum to 1.0 and are, therefore, compositional in nature.

Both methods were tested for discriminating between the surface lithologies in this study. However, a comparison of results between linear discriminant and quadratic discriminant analysis showed little difference in the results and some classes had to be omitted because of an insufficient number of training sites.

To overcome some of the problems of applying classification methods in previous studies, we employed the statistical method, Random Forests (Breiman 2001) as employed by Harris and Grunsky (2015) and used as part of a remote predictive mapping strategy (Harris et al. 2008). The Random Forest method is based on the construction of classification trees (Venables and Ripley 2002, Chap. 9) in which nodes (splits in classes) are based on continuous variables from which a series of branches in the tree will correctly classify (categorical variables) all of the data. The Random Forest method "grows" many trees and each tree provides a classification. Each classification is termed a vote and a classification is assigned to the forest with the most votes. A useful description of the methodology is provided in Breiman and Cutler (2016). The function "**randomForest**", herein referred to as "RF", from the package **randomForest** (Breiman and Cutler 2016) was used for the analysis.

For each tree that is created, a training set of approximately one-third of the data is drawn, with replacement and are left out of the sample population. This is known as the out-of-bag (oob) data and is used to get a running unbiased estimate of the classification error, as trees are added to the forest. Variable importance is also

determined from the out-of-bag data. For each tree, all of the data are applied to the tree and "proximities" are determined for each pair of cases. If two cases occur at the same node, then the proximity of that pair is increased by one. When all of the trees have been estimated, the proximities are normalized by dividing by the number of trees. These proximities can be used for replacing missing data, identifying outliers and creating lower dimensional views of the data. Each tree is constructed from bootstrapping the original sample population and about one third of the data are left out from each bootstrap sample and not used in tree construction but are then classified from the tree created from the other two thirds of the sample population. An unbiased estimate of the classification error is determined from each case that is oob and did not classify correctly. Variable importance is determined by comparing oob classification results and the non-oob classification results after random permutations of each of the variables. Another measure of variable importance is determined by the Gini measure that is determined by the number of splits that are made for a given variable over all of the trees in the forest. Variables do not need to be pre-selected using techniques such as analysis of variance as the RF procedure determines which variables are the best classifiers.

Maps of the normalized votes, which are equivalent to posterior probabilities, can be created using geostatistical methods such as kriging. However, since the posterior probabilities are compositions and sum to 1.0, these values must be logratio transformed, followed by subsequent co-kriging, and then back transformed for subsequent geographic rendering (Pawlowsky-Glahn and Egozcue 2015; Mueller and Grunsky 2016). Instead, maps of the posterior probabilities for each of the classes were created by posting the sample sites with points and colours. An alternative to this would be to consider the un-normalized (raw) votes as independent and carry out kriging on these estimations. The results of these interpolations are provided in the Supplementary Annex.

17.3 Results

17.3.1 Process Discovery—Principal Component Analysis

A logcentred transform was applied to the adjusted data after which a principal component analysis was carried out. An examination of an ordered plot of eigenvalues in the form of a screeplot (Jolliffe 2002) are shown in Fig. 17.2a–d for (a) all of the data, (b) Surface Soil, (c) A horizon only and (d) C horizon only. Figure 17.2a–d display two important inflection points; at PCs 3 and 9. The first three eigenvalues define the dominant structure in the data and the next 5 display lesser but significant structure also. This is also expressed numerically in Table 17.1 where the first 10 eigenvalues are listed along with the associated cumulative contribution to the structure in the data. As shown in the screeplots of Fig. 17.2, a comparison of the first four successive eigenvalues between the C-horizon,

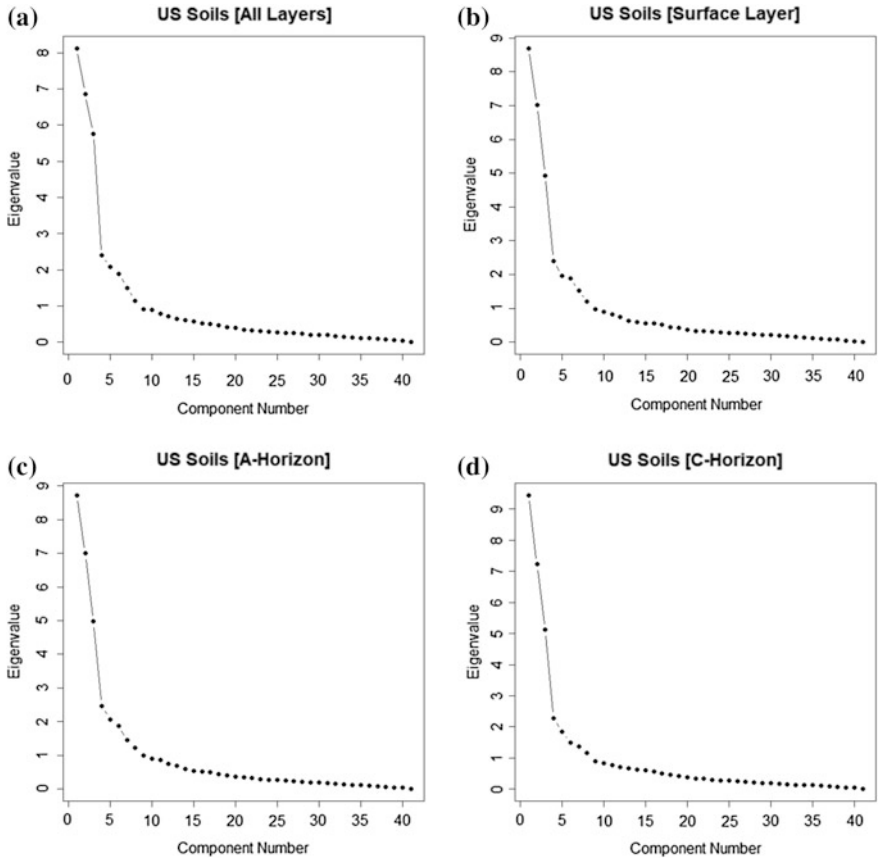

Fig. 17.2 **a**—Screeplot of eigenvalues of the soil geochemistry for the combined Surface Soil (0–5) cm layer, the A- and C- horizons, from the application of a principal component analysis to logcentred transformed data. **b**—Screeplot of eigenvalues of the soil geochemistry for the Surface Soil (0–5) cm layer from the application of a principal component analysis to logcentred transformed data for the top layer only. **c**—Screeplot of eigenvalues of the soil geochemistry for the A-horizon from the application of a principal component analysis to logcentred transformed data for the A-horizon only. **d**—Screeplot of eigenvalues of the soil geochemistry for the C-horizon from the application of a principal component analysis to logcentred transformed data for the C-horizon only

A-horizon and Surface Soil is slightly greater for the C-horizon. This implies that the linear combinations of the elements are stronger for the C-horizon than for the other two. Eigenvalues with values less than 1 and are interpreted to represent under-sampled processes or random effects (noise).

The largest eigenvalues signify that the linear combinations of the elements for these components are significant and defines “structure” in the data. This structure can be interpreted as the influence of stoichiometric control of mineralogy.

Table 17.1 Principal Component Analysis results for logcentred transformed soil geochemistry

RQPCA [clr] All layers										
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
λ	8.13	6.87	5.76	2.39	2.08	1.88	1.50	1.15	0.92	0.89
λ%	19.83	16.76	14.05	5.83	5.07	4.59	3.66	2.80	2.24	2.17
Σλ%	19.83	36.59	50.63	56.46	61.54	66.12	69.78	72.59	74.83	77.00
RQPCA [clr] surface soil										
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
λ	8.70	7.01	4.93	2.41	1.96	1.89	1.53	1.21	0.98	0.90
λ%	21.19	17.08	12.01	5.87	4.77	4.60	3.73	2.95	2.39	2.19
Σλ%	21.19	38.27	50.28	56.15	60.93	65.53	69.26	72.20	74.59	76.78
RQPCA [clr] A horizon										
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
λ	8.73	7.00	4.97	2.47	2.07	1.88	1.45	1.22	1.00	0.90
λ%	21.29	17.07	12.12	6.02	5.05	4.59	3.54	2.98	2.44	2.20
Σλ%	21.29	38.37	50.49	56.51	61.56	66.15	69.68	72.66	75.10	77.29
RQPCA [clr] C horizon										
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
λ	9.45	7.22	5.12	2.29	1.84	1.50	1.36	1.17	0.89	0.82
λ%	23.02	17.59	12.47	5.58	4.48	3.65	3.31	2.85	2.17	2.00
Σλ%	23.02	40.61	53.08	58.66	63.14	66.80	70.11	72.96	75.13	77.13

17.3.2 PCA of the Combined Surface Soil, A-Horizon, C-Horizon

Figures 17.3a, and 17.4a shows biplots (PC1-PC2 and PC2-PC3) for the principal component scores and loadings for the combined data from the surface soil, A- and C-horizons Table 17.1 shows that the first three principal components for the combined data (All Layers) account for 50.6% of the overall variation in the data.

Figure 17.3a shows the mass of data points defined by two vertices: (1) Cr-V-Ni-Co-Fe-Sc-Mn-P-Zn; (2) Hg-In-Ti-Se-Mo-As-Sb-Sn-Bi (chalcophile) and a trend of element associations: Mg-Ca-Na-Sr-Ba-K-Be-Rb-Tl that are inversely associated with the vertex defined by (2) above. The chalcophile elements are grouped along the +PC1 axis. Siderophile elements are associated with the +PC2 axis and the lithophile elements are distributed around the ±PC1/−PC2 axes and the −PC1/+PC2 axes.

Figure 17.4a shows the three sets of data (Surface Layer, A- and C-horizon) combined onto a biplot of PC2–PC3. The PC scores along the PC2 axis define a contrast between mafic (+ scores) and felsic (−scores) source material. Siderophile (Fe, Co, Ni), lithophile (Cr, V, Sc, Ti) and chalcophile elements (Cu, In) are associated along the +PC2 axis and lithophile elements (Rb, K, Tl, Ba, Th, La, Be, Ce) are concentrated along the −PC2 axis.

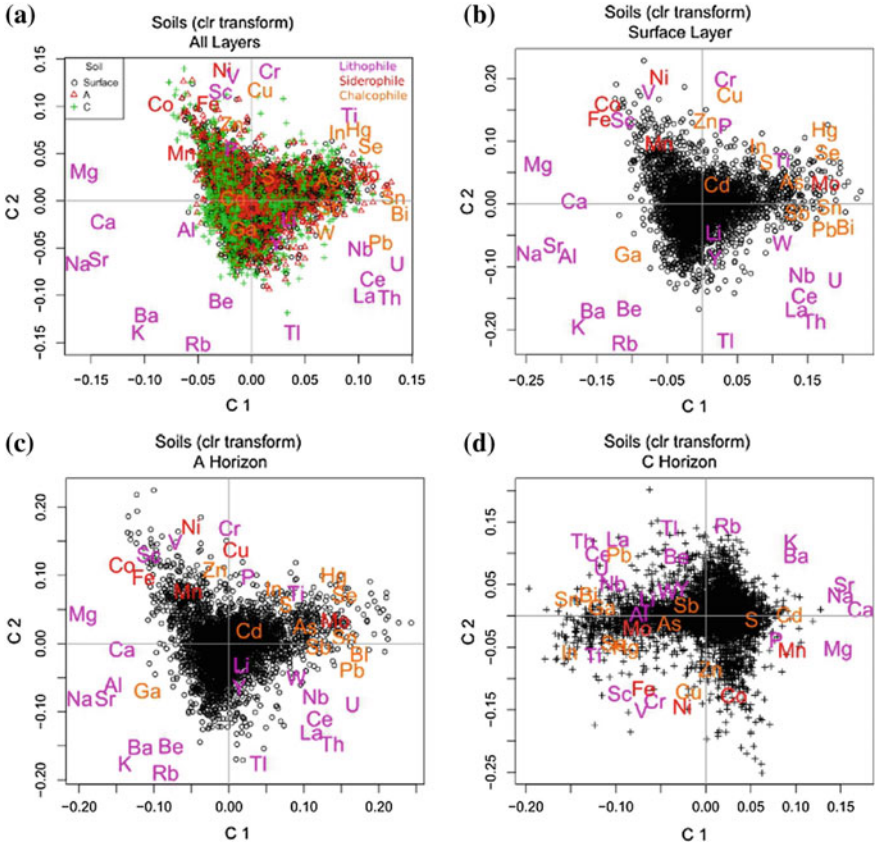

Fig. 17.3 **a**—Biplot of principal components 1 and 2 for the soil geochemistry for the combined Surface Layer, A, and C horizon soil geochemical data based on a log centred transform. The colours and symbols represent the surface soil and the soil A and C horizons. **b**—Biplot of principal components 1 and 2 for the Surface Soil geochemistry data based on a log centred transform. **c**—Biplot of principal components 1 and 2 for the A-horizon soil geochemistry data based on a log centred transform. **d**—Biplot of principal components 1 and 2 for the C-horizon soil geochemistry data based on a log centred transform

An association of chalcophile elements (Cd, S, Sb, As, Hg, Pb) occurs along the +PC3 axis with a corresponding concentration of sample sites associated with the surface layer and A-horizon, most likely representing complexing with organic rich soils. PC scores for the C-horizon are concentrated along the ±PC2 axis, which may represent a range of source material from mineral soils that are low in organic material (−PC3) to soils that are rich in organic material or derived from shales/weathered materials (+PC3).

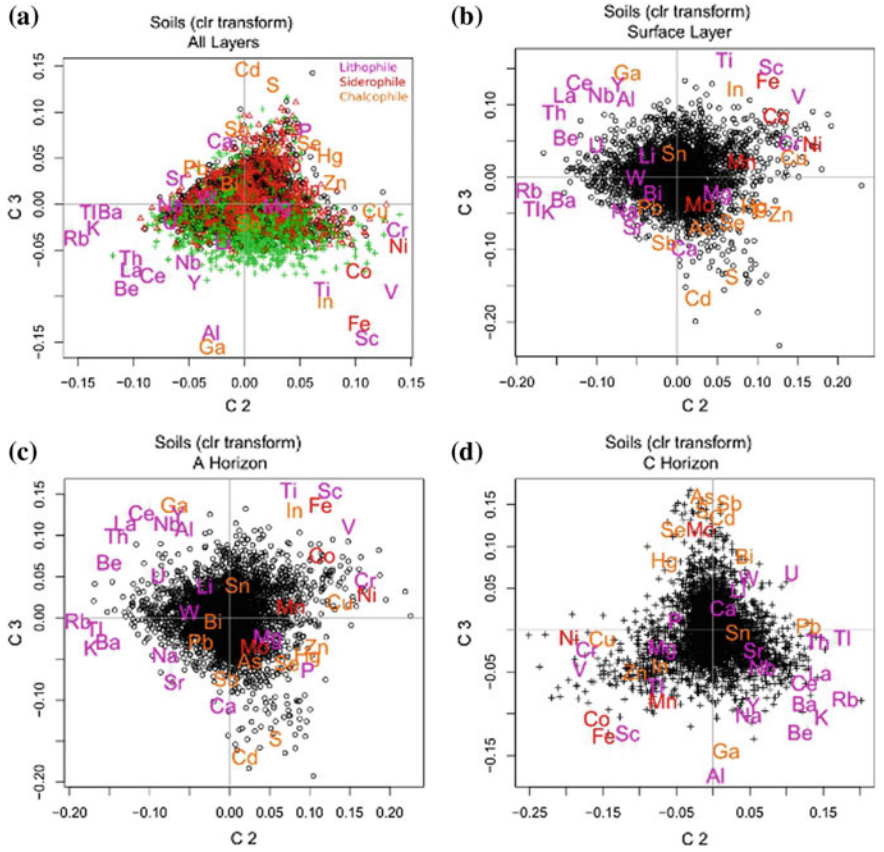

Fig. 17.4 **a**—Biplot of principal components 2 and 3 for the soil geochemistry for the combined Surface Soil, A, and C horizon soil geochemical data based on a log centred transform. The colours and symbols represent the surface soil and the soil A and C horizons as shown in Fig. 17.3a. **b**—Biplot of principal components 2 and 3 for the top layer soil geochemistry data based on a log centred transform. **c**—Biplot of principal components 2 and 3 for the A-horizon soil geochemistry data based on a log centred transform. **d**—Biplot of principal components 2 and 3 for the C-horizon soil geochemistry data based on a log centred transform

17.3.3 PCA of the Surface Soil, A-Horizon, C-Horizon

The biplots of Fig. 17.3a–c for all of the data, the surface soil data and the A-horizon data, show similar patterns in terms of the relationships of the elements with each other and the shape of the data cloud for the projection of the principal component scores onto the PC1 and PC2 axes. The biplots exhibit a range of lithophile loadings that define materials derived from mafic, feldspathic, carbonate and REE-enriched sources within the quadrants described previously. Similarly, the chalcophile element association is concentrated along the +PC1 axis for both

Fig. 17.3b, c, likely representing weathered and organic-rich material, which adsorb chalcophile elements.

The biplot of Fig. 17.3d (C-horizon) displays a different pattern in comparison with Fig. 17.3a–c. The +PC1 axis shows an association of lithophile elements (Ca-Mg-Na-Sr-P) and chalcophile elements (S-Cd), possibly representing a mix of feldspathic and/or carbonate source material. Along the PC1 axis and on the +PC2 domain, there is a contrast between (Ca-Na-Mg-S-Ba-K) and (Th-Ce-U-La-Nb-Al-Li) that may reflect a feldspathic/carbonate source environment from an environment with relative enrichment in heavy minerals.

Figure 17.4a shows a pattern and association of elements that displays a contrast of the C-horizon data with the surface soil and A-horizon data. Figure 17.4a shows a siderophile and mafic lithophile pattern of Cr-Ni-Cu-V-Co-Fe-Sc along the +PC2 axis. Along the −PC2 axis of Fig. 17.4a there is a lithophile association of Rb-K-Ti_Ba-Ce-La-Tl. The +PC3 axis in Fig. 17.4a shows a chalcophile/lithophile association of Cd-S-Sb-Ca-P-Se-Hg-As-Mo-Pb-Sr-Zn. This region of the plot is dominated by surface soil and A-horizon data although some C-horizon data are also present. A similar pattern is observed in Figs. 17.4b, c although the groups of the elements are at opposite ends of PC3 (a sign switch). In Fig. 17.4b, c, transitional between the siderophile/lithophile elements (Fe-Sc-Co-Cr-Ni) and the lithophile elements (Rb-Tl-K-Ba) is the grouping of Al-Ga-Nb-Y-Ce-La-Th-U that represents feldspars, clays and heavy minerals. As in Figs. 17.3d and 17.4d, representing the C-horizon data, shows the chalcophile enrichment trend along the +PC3 axis and a siderophile/lithophile trend along the PC2 axis. Transitional between the trend along the PC2 axis is an association of Al-Ga, likely representing feldspars and clays.

17.3.4 Mapping the Components

The first three principal components for the surface soil, the A- and the C-horizons were interpolated using the geostatistical package, gstat (Pebesma 2004). Experimental semi-variograms were generated followed by variogram model fitting with subsequent kriging. The images for the three principal components are shown in Figs. 17.5a–c, 17.6a–c and 17.7a–c.

Principal Component 1

Geospatially these patterns are observed in Figs. 17.5, 17.6 and 17.7. Figure 17.5a–c show interpolated images based on kriging of the first principal component for the surface soil, A- and C-horizons respectively. The patterns observed in Fig. 17.5a and b are consistent with the patterns observed in Fig. 17.3b and c. The +PC1 axis in Fig. 17.3b and c show relative enrichment of the previously identified chalcophile elements and relative enrichment of the mafic lithophile and siderophile elements along the −PC1 axis. In Fig. 17.5a and b, the positive scores of PC1 appear to correspond with the region in the southeast US and the negative scores of PC1

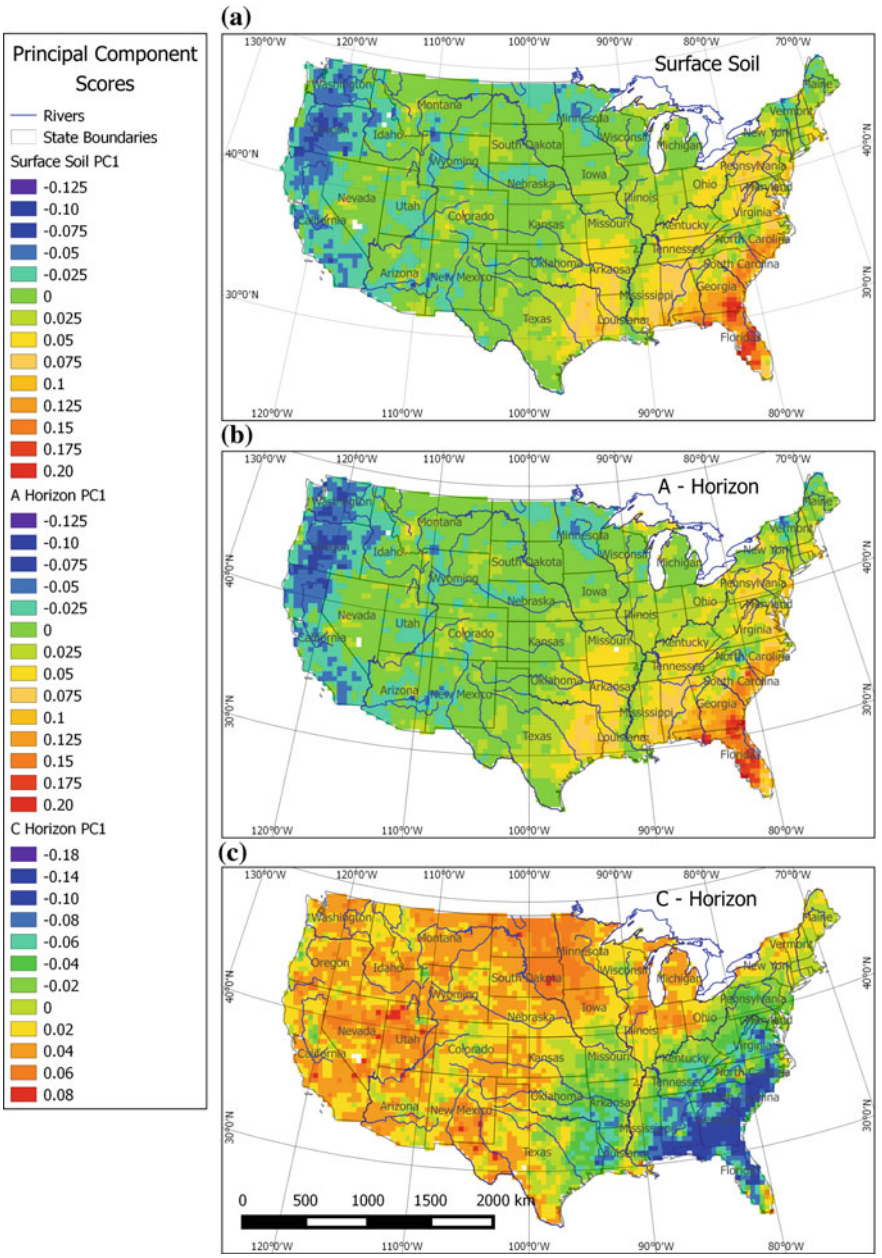

Fig. 17.5 **a–c** Map of kriged principal component 1 for the Surface Soil, A- and C-horizon data. Figures 17.4b–d provide the context for relative element enrichment/depletion associated with each of the layers

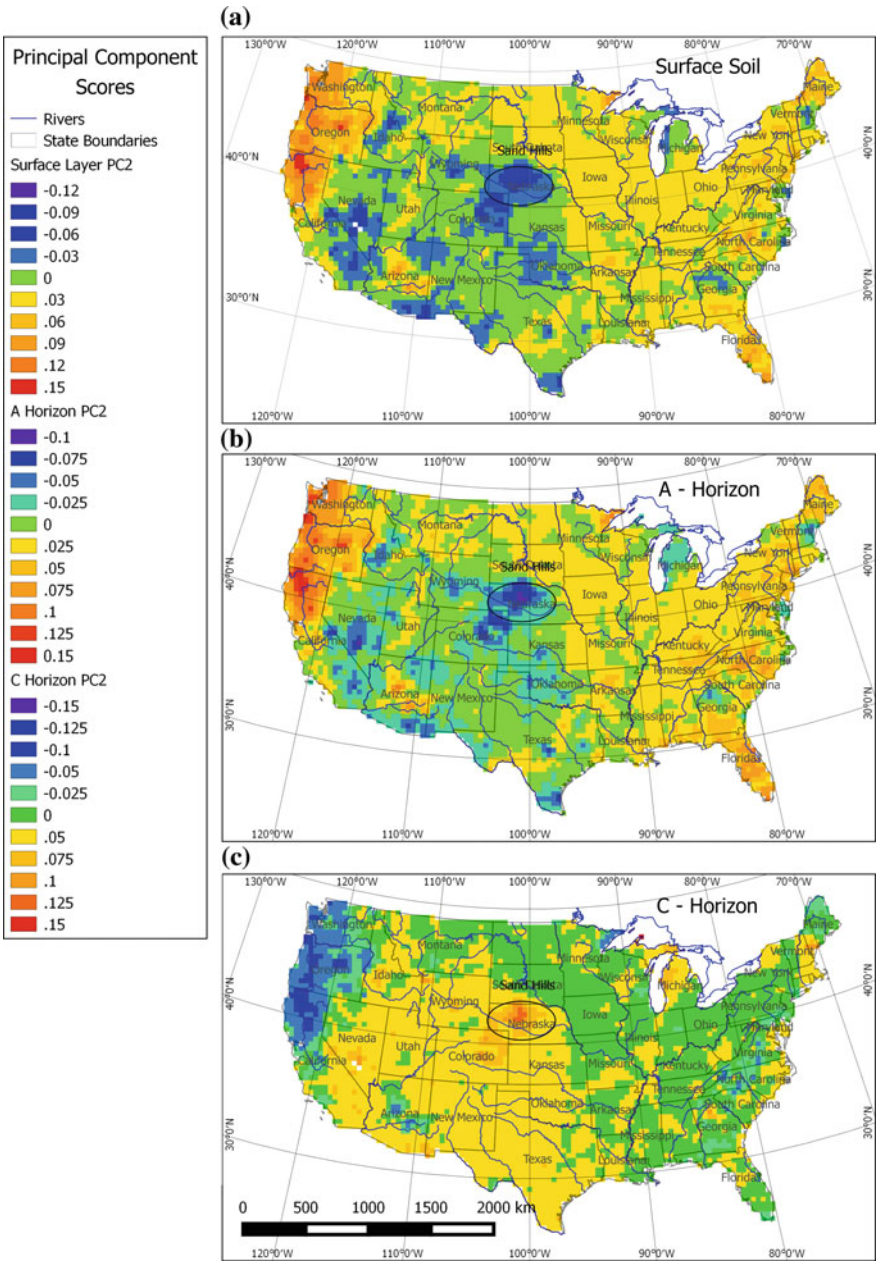

Fig. 17.6 **a–c** Map of kriged principal component 2 for the Surface Soil, A- and C-horizon data. Figures 17.4b–d provide the context for relative element enrichment/depletion associated with each of the layers

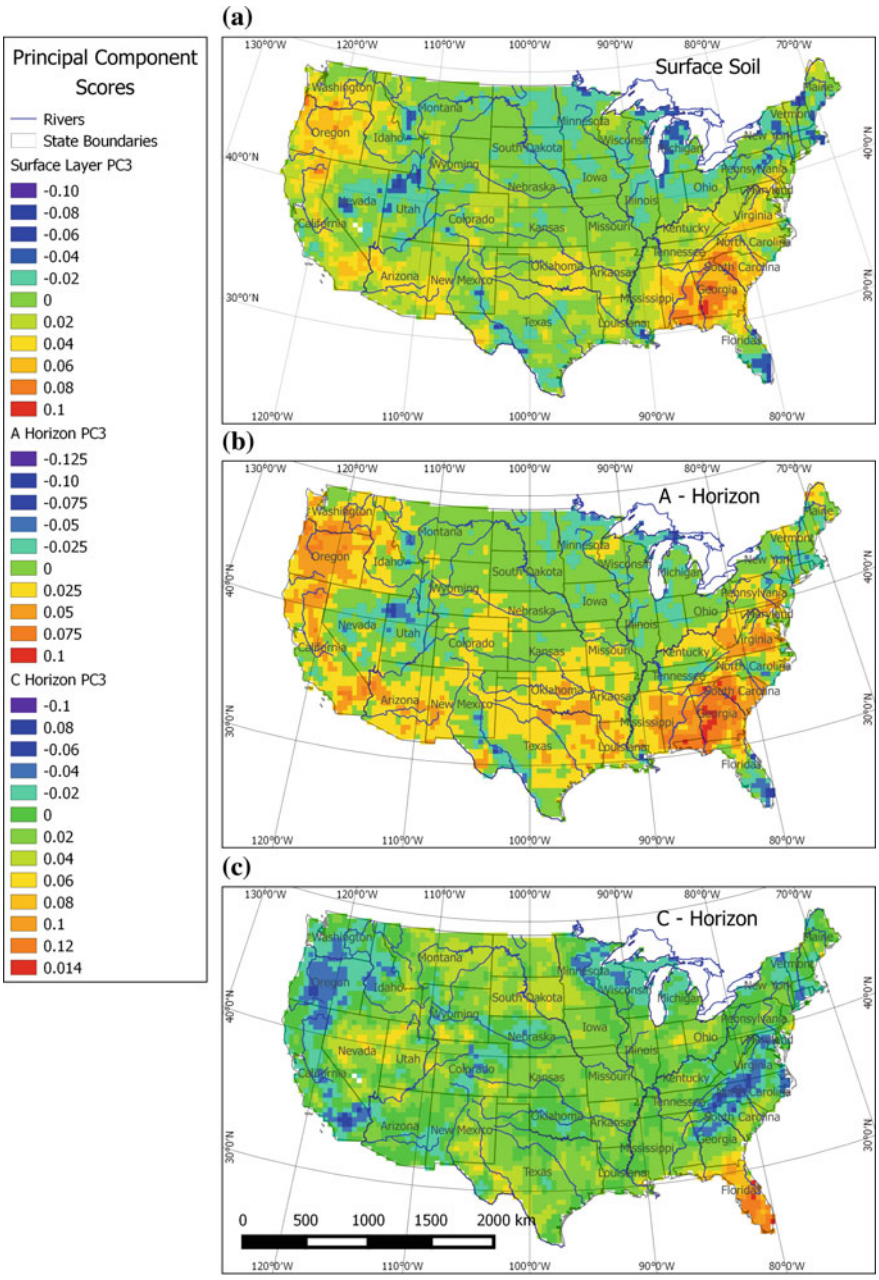

Fig. 17.7 **a–c** Map of kriged principal component 3 for the Surface Soil, A- and C-horizon data. Figures 17.4b–d provide the context for relative element enrichment/depletion associated with each of the layers

appear to occur in the northwest US and west of Lake Superior. All three figures show a pattern that coincides with the banks of the Mississippi River. Negative PC1 scores for the surface layer and A-horizon correspond to relative enrichment in Na-Sr-Al-Ca-Mg-K-Ba element associated with feldspars and/or carbonate source material.

The image of PC1 for the C-horizon data (Fig. 17.5c) shows a strong negative region in the southeast US that corresponds to the chalcophile group of elements along the negative portion of PC1 in the biplot of Fig. 17.3d. The positive portion of PC1 in Fig. 17.3d corresponds to the dominantly lithophile and siderophile groups of elements and is displayed as a large region throughout the US, with the exception of the southeast US. The same "corridor" pattern along the Mississippi River is observed in Fig. 17.5c, for the C-horizon results and represent the same relative concentration of lithophile elements observed in the surface layer and A-horizon.

Figure 17.5c shows the kriged image for the first principal component derived from the C-horizon data. In this case, the negative scores are restricted to the eastern US and reflect the chalcophile and rare earth elements indicative of detrital heavy minerals corresponding to the region of quartz enrichment accompanied with weathered and detrital materials within the erosional and weathering domain of the eastern US. Positive PC1 scores reflect a lithophile association of Ca-Na-Sr-Cd-Mg-Ba-K-Mn (Fig. 17.3d) and suggest an environment that is likely dominated by Ca-Na-K-Ba-Sr feldspars and Mg-Ca bearing ferromagnesian minerals.

An important consideration in the interpretation of the biplots is the significance of the associations of the elements. An initial interpretation of the biplots of Fig. 17.3a–d was that the associations of the chalcophile groups indicated relative enrichment of these elements (Hg-Se-As-Sb-Sn-Bi-Pb-S-In) that represent weathered materials along with the accumulation of detrital minerals within the erosional and weathering domain of the southeastern US. In fact, these elements do not reflect relative enrichment but rather relative depletion with respect to the other groups of elements, notably the siderophile and lithophile elements. Geospatially, the chalcophile association of these elements corresponds to the region of a high quartz content in the soil (Smith et al. 2014) and has been termed the "quartz dilution effect". This effect in the soil geochemistry and the subsequent multi-element associations would likely be significantly different had Si been included in the analysis. A test was carried out in which the Si content of the data was simulated as the difference from the potential total (1,000,000 ppm) from the summed content of the compositions. This simulated Si value was then included in the composition and a PCA was carried out. The first component identified the relative Si enrichment as occurring in the southeast US. The simulated value of Si was not included in this study because other elements should also be considered in a total composition, including oxygen and nitrogen.

Principal Component 2

As shown in Fig. 17.3b, c, the multi-element signature of tpc2 is nearly the same for the surface soil and A-horizon. The patterns in both figures show two trends, one

with relative enrichment in Cr-Ni-Co-Cu-V-Fe-Sc (siderophile/lithophile + Cu-Zn) and the other with relative enrichment in Hg-Se-As-Sn-Sb-Pb-Bi-In-S. (chalcophile) These two multi-element associations reflect the chemistry of mafic minerals and elements that are associated with weathering and organic complexing. This is reflected in the maps of Fig. 17.6a, b in which high PC2 values are noted in the eastern and south eastern US and the western US. The negative PC scores for the surface soil and A-horizon show relative enrichment in Rb-K-Tl-Ba-Be-Na-Sr-Al-Ga and, as shown in Fig. 17.6a, b are geospatially concentrated in the central US corresponding to the location of the Sand Hills of Nebraska, (~105° W/42° N), which is comprised of sand-sized particles of quartz and feldspar (Smith et al. 2014). There are also areas of negative PC2 scores, most likely representing feldspars associated with granitoid rocks in southern Nevada, California, Arizona, Texas, New Hampshire and Maine (Smith et al. 2014).

The map of PC2 (Fig. 17.6c) for the C-horizon data shows positive scores associated with the mafic volcanic rocks of the northwest US and corresponds to the relative enrichment of siderophile (Fe-Ni-Co), lithophile (Cr-V-Sc), chalcophile (Cu-Zn) elements as shown in Figs. 17.3d and 17.4d. The negative scores for PC2 show a similar pattern to those of the surface soil and A-horizon; relative enrichment in alkali lithophile elements (Rb-K-Ba-Be-Na-Sr) with Al-Ga representing feldspars and REE lithophile elements (U-Th-La-Ce-Ng-Tl) that represents heavy minerals and quartz (as explained previously). The geochemical expression of these minerals in PC2, which are resistant to weathering, are reflected in both horizons and the surface soil.

Principal Component 3

The positive scores for the PC3 show relative enrichment of siderophile, mafic lithophile, and light REE elements for both the surface soil and A-horizon; whereas this pattern is represented by negative scores for the C-horizon. As shown in Fig. 17.4b–d, for all three layers, there is a continual transition from relative enrichment in alkali lithophile and REE elements, including Al and Ga, representing feldspars and minerals associated with felsic domains to relative enrichment in Cr-Ni-V-Cu-C-Fe-Sc-Ti-In-Zn that represents minerals associated with mafic domains. Figures 17.7a–c show the kriged images for the third principal component. The negative scores show relative enrichment of Cd-S-Ca-Sr-Sb-P-As, which may reflect the processes of organic complexing and sulphates. Negative scores noted in Utah, Nevada, west Texas, the Mississippi delta and south Florida may have a greater component of S. Negative scores that occur in Minnesota, Michigan, Indiana and the coast of New England may reflect the presence of shales, clays and organic accumulations. The negative PC3 scores of Fig. 17.4b exhibit a bimodal pattern of relative enrichment of Fe-Sc-In-Ti and Ga-Al-Y-Nb-Ce-La. The Fe-rich pattern is associated with the mafic volcanic rocks in the northwest and southwest US and the Ga-rich pattern occurs in the eastern US and reflects the presence of feldspars in the weathering of granitoid rocks in the southern Appalachians.

As seen in Fig. 17.4c, and nearly identical to that the of surface soil, the positive scores of PC3 exhibit a bimodal pattern for the A-horizon and indicate relative

enrichment of Ti-Sc-Fe-In-V and Ga-Al-Th-La-Nb-Ce. These two groups reflect both a mafic and feldspathic/heavy mineral rich environment. Figure 17.7b shows the mafic association (Ti-Sc-Fe-In-V) in the northwest US. The positive scores in the eastern, southern, and in particular, the southeast US reflect elements associated with feldspars and heavy minerals, which reflects the concentration of minerals through the weathering process, which may be due more to gravitational effects than chemical breakdown. As in Fig. 17.7a, the negative scores of PC3 in the A-horizon demonstrate the same patterns and processes.

The C-horizon map shows two distinct geospatial patterns. The positive scores of Fig. 17.4d show relative enrichment in the chalcophile group, Sb-As-S-Mo-Se-B-Cd-Hg-U-Li-W and occur primarily in the southeast US. This pattern likely reflects both the quartz dilution effect and the presence of chalcophile elements relative to other areas throughout the US. The negative scores, which show relative enrichment of the lithophile elements Al-Ga-Na-Y-K-Be-Ba-Mn-Ti-Fe-Sc-Co, reflect a combination of mafic minerals and feldspars. These patterns are observed in the western US, Minnesota-Wisconsin, central Appalachia and the northeast US. Patterns associated with the elements that reflect mafic domains are the northwest US and Wisconsin-Minnesota. Patterns that reflect the feldspathic domains are Nebraska-Colorado, central Appalachia and the northeast US.

Evaluation of the soil geochemistry for the surface soil, the soil A horizon and the soil C horizon using a principal component approach reveals that there are continental-scale geochemical patterns that appear to be associated with the composition of the underlying soil parent material, climate, and weathering. At the scale of evaluation, details on specific lithologies are difficult to resolve, but the patterns are consistent with those mineralogical patterns delineated by Smith et al. (2014).

Process Validation Predictive Mapping of Surface Lithologies

The lithology of surficial materials by Sayre et al. (2009) is represented by 18 classes plus unknowns and listed in Table 17.2. A total of 17 classes were selected for further study. The classes “unknown” and “water” were not used as they were not considered suitable for classification.

Figure 17.8 shows a map of the sampling sites with the surface materials lithology from Sayre et al. (2009). The patterns of surface materials on the map show some similarities with the patterns observed from the first three principal components for the surface soil, A- and C-horizons. Figure 17.9 shows a biplot of the first two principal components that are coded according to the surface lithologies. The pattern of the mafic lithophile elements (Cr-Ni-Cu-V-Co-Fe-Sc) in Fig. 17.9a, b are dominated by silica-rich residual soils (SilRes), whereas the chalcophile enrichment pattern (Hg-Se-Mo-Sn-Bi-Pb-Sb-As-Ti-S-In) appears to be associated mostly with alluvium (Alluv) and coastal zone sediments (CZS). The lithophile element grouping in the negative portion of the PC2 shows a mix of several lithologies. The results of the PCA suggest that the linear combinations of elements from the PCA are related to the patterns observed in Surface Materials Lithologies of Fig. 17.8.

Table 17.2 List of surface lithologies across the conterminous United States

Mnemonic	Class description	Surface layer	A-horizon	C-horizon	Total
AlkInt	Alkaline intrusive/volcanic rocks	6	7	6	19
Alluv	Alluvium and fine-textured coastal Zone sediment	994	989	984	2967
CaRes	Carbonate residual material	265	263	260	788
Colluv	Colluvial sediment	379	379	366	1124
CZS	Coastal zone sediment, coarse-textured	44	45	43	132
EolDune	Eolian sediment, coarse-textured (Sand Dunes)	152	151	151	454
EolLoess	Eolian sediment, fine-textured (Glacial Loess)	156	155	155	466
ExtVR	Extrusive volcanic Rock	50	51	51	152
GlLs	Glacial lake sediment, fine-textured	89	85	86	260
GlOut	Glacial outwash and Glacial lake sediment, coarse-textured	221	220	221	662
GTCg	Glacial till, coarse-textured	114	111	111	336
GTClay	Glacial till, Clayey	61	61	61	183
GTLoam	Glacial till, Loamy	529	528	526	1583
HyPM	Hydrick peat muck	25	25	26	76
NCaRes	Non-carbonate residual material	1188	1174	1170	3532
SalLS	Saline lake sediment	78	82	79	239
SilRes	Silicic residual material	457	456	452	1365
Water[a]		22	21	21	64
Unknown[a]		6	6	6	18
Total		4836	4809	4775	14420

[a]Not Used

From the application of the random forest classification, the Gini Index (significance of the variables) for the surface soil, A- and C-horizons are listed in Table 17.3 and shown graphically in Fig. 17.10. The significance uses the Gini Index, which is a measure of purity based on the success of a variable in distinguishing between classes. Table 17.3 shows that generally, PC's 4, 5, 1, 2, 3 and 6 are the best variables for classification of the surface lithologies for the surface soil, A- and C-horizons. Maps of the normalized votes in point form and interpolated (kriged) maps of the raw votes are shown in the Supplementary Annex (Supplementary Figs. 1–15).

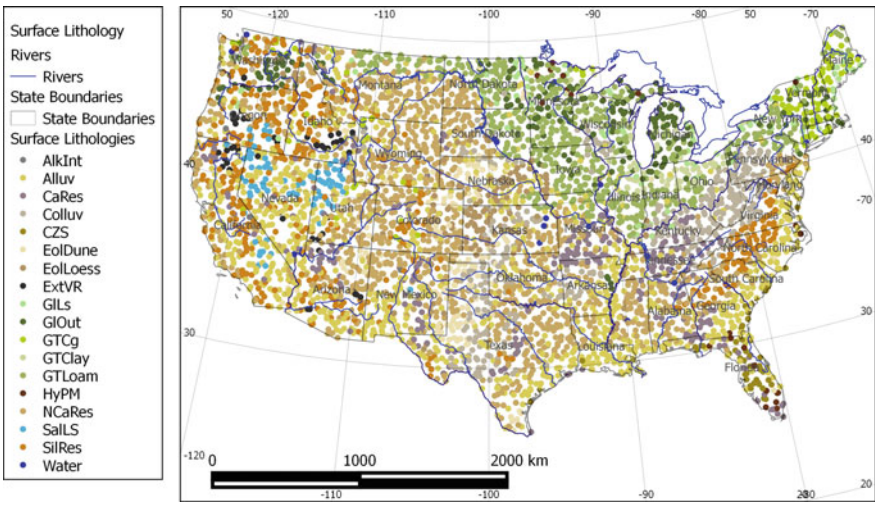

Fig. 17.8 Map of soil sample sites coded by the Surface Lithology classification. This map represents the actual classification based on the maps of Sayre et al. (2009). Colours used in this figure are the same colours used in Sayre's maps. See text for details on how the sites were selected

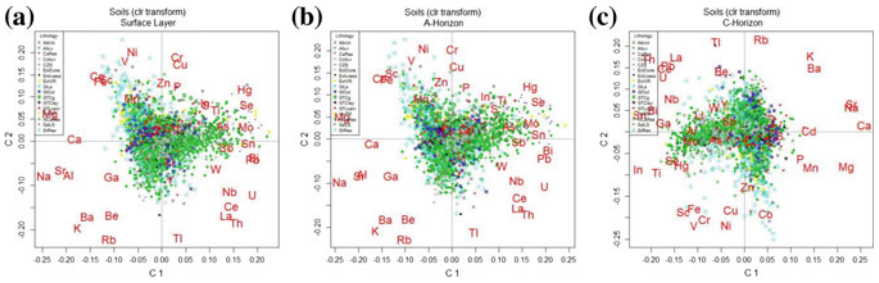

Fig. 17.9 **a–c** Principal component biplot of the surface layer (**a**), A-horizon (**b**) and C-horizon (**c**) scores that are coded and coloured according to the surface lithologies

Table 17.4 shows the accuracy of prediction for each of the surface lithologies based on the Random Forest out-of-bag classification methodology for each of the surface soil, A- and C-Horizons. The table has been ordered from the highest to the lowest prediction accuracies based on the surface soil. It is worth noting that the depth of soil has only a minor influence in the prediction accuracies, suggesting that the geochemical signature of the underlying material persists throughout the soil column. Non-carbonate residual soils (NCaRes) (~74%), loam associated with glacial till (GTLoam) (66–72%), siliceous residual soils (SilRes) (48–56%), alluvium (Alluv) (~50%) and coastal zone sediments (CZS) (45–48%) have the highest prediction accuracies, whereas the lowest accuracies are associated with hydric peat and muck (HyPM) (0%), alkalic intrusions (AlkInt) (0%), glacial lake sediments (GlLs)

Table 17.3 List of variable importance for the surface layer, A- and C-horizons as determined from Random Forest classification of the principal component results applied to the clr-transformed data. Colours reflect the most significant PCs (red) to least significant PCs (blue)

Surface Layer	Importance	A Horizon	Importance	C Horizon	Importance
PC		PC		PC	
4	198.35	4	185.34	2	165.83
5	180.88	5	172.04	4	156.36
1	163.70	1	170.09	1	154.76
2	155.81	3	150.05	3	152.11
3	152.73	2	148.14	6	131.94
9	129.17	9	127.18	16	128.21
12	109.74	6	126.50	5	115.55
6	108.61	20	119.91	11	113.04
32	106.28	29	110.74	8	109.54
23	102.15	13	102.50	7	109.05
30	100.84	12	100.25	10	106.46
8	98.87	11	98.07	14	101.91
20	98.77	8	96.86	13	96.87
40	98.19	23	94.89	12	96.57
10	96.90	10	94.48	28	95.32
21	95.83	7	93.99	9	94.97
11	93.68	16	92.35	18	93.88
15	91.76	18	92.21	17	92.57
24	91.73	19	91.46	31	92.53
13	91.49	21	89.23	34	92.34

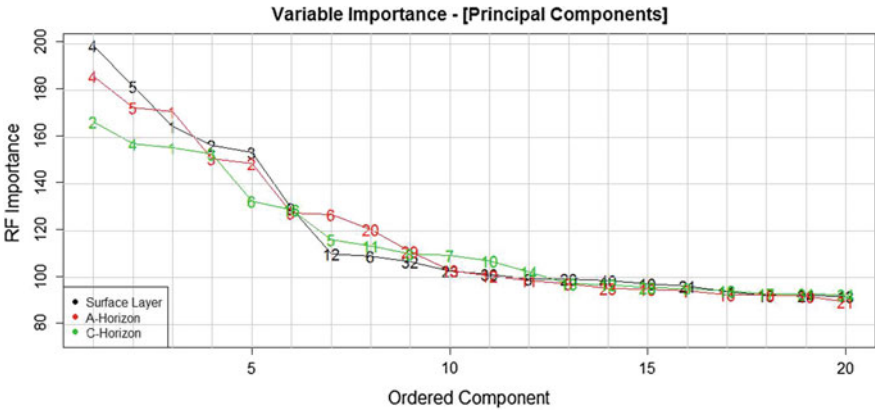

Fig. 17.10 Plot of the significance of the principal components used in the random forest classification based on the Gini Index for the Surface Layer, A- and C-horizons. See the text for a detailed explanation

Table 17.4 Measures of ordered predictive accuracy for the surface lithologies for the surface layer, the A- and C-horizons based on a Random Forest classification of the principal component results applied to the clr-transformed data

	Surface Layer	A-Horizon	C-Horizon
NCaRes	74.82	73.66	74.51
GTLoam	71.61	68.90	65.74
SilRes	52.02	47.97	55.92
Alluv	50.38	50.63	49.77
CZS	44.90	48.34	48.26
Colluv	37.14	38.20	32.73
GlOut	28.41	32.63	29.32
GTClay	27.54	37.32	42.23
GTCg	22.65	21.47	20.57
EolDune	22.25	22.40	16.47
EolLoess	21.05	26.97	30.19
CaRes	19.19	15.16	9.58
SalLS	15.22	15.69	1.25
ExtVR	1.96	5.78	0.00
GlLs	1.11	0.00	0.00
AlkInt	0.00	0.00	0.00
HyPM	0.00	0.00	0.00
Overall Accuarcy	49.92	49.37	48.61

(0–1%) and extrusive volcanics (ExtVR) (0–6%). The prediction accuracy is sensitive to the initial representation of each class in the dataset. This sensitivity is partly due to the masking and swamping effect that a large population of sites for one type of surface lithology over another (i.e. Alluvium vs. Hydric Peat and Muck).

Supplementary Tables 2, 3 and 4 provide a complete summary of the prediction accuracies for the surface soil, A- and C-horizons, respectively. The diagonal of each upper table (Tables 2a, 3a, 4a) indicates how many sample sites were classified correctly. Each row of the off-diagonal elements indicates the misclassification of the sites for each of the classes. The lower tables in Tables 2b, 3b, 4b show the classification accuracies as expressed in percentages. The overall classification accuracy is shown at the bottom of each table. Scanning the columns of Tables 2a, 3a, and 4a reveals that many classes are confused with alluvium (Alluv), siliceous residual material (SilRes), loam derived from glacial till (GTLoam) and non-carbonate residual material (NCaRes). Alluvium and non-carbonate residual material appear to overlap with almost all of the classes. The overall prediction accuracies for the surface soil, A- and C-Horizons are 50%, 49% and 49%, respectively.

The R package "**randomForests**" produces raw and normalized votes for each of the classes. Votes are a record of the number of times a site is correctly classified. As described above, normalized votes are the equivalent of a posterior probability

and are therefore compositions. Classes such as AlkInt, HyPM and other classes that have low abundance in the data create problems in the creation of co-regionalization that is required for co-kriging. Examples of the spatial distribution of the normalized and raw votes are shown below. The Supplementary Annex provides predictive maps for all of the surface lithologies, based on the normalized votes, for the surface soil, A- and C-horizons. Predictive maps for AlkInt and HyPM are not shown because the normalized votes for these two surface lithologies were very low and do not show any geospatial patterns. The prediction accuracies for the three media from Table 17.4 are: 49.9%, 49.4% and 48.6% respectively. Supplementary Tables 2, 3 and 4 provide details on the overlap of predictions for each surface lithology. In most cases, overlap is associated with non-carbonate residual soils, glacial till derived loam and alluvium. These three classes have the broadest range of compositional variation and occupy a significant amount of area across the conterminous US.

Figure 17.11 shows a map of normalized votes of Non-carbonate residual soils (NCaRes) derived from the random forest classification. Normalized votes >0.3 occur throughout the Midwest states from the Canadian border in the north to the Gulf of Mexico in the south. From Table 17.4, the overall classification accuracy is approximately 75% for the surface soil and the two soil horizons. Supplementary Tables 2, 3 and 4 show that compositional overlap occurs primarily with alluvium, which is also shown in the maps of Fig. 17.11 where a large number of sample sites show low normalized votes (~0.2–0.3). Supplementary Fig. 13a, b show the normalized and raw vote maps of the NCaRes prediction.

Figure 17.12 shows a map of normalized votes for loam derived from Glacial Till (GTLoam). The overall classification accuracy ranges from 65.7 to 71.6% over the three soil layers. Supplementary Tables 2, 3 and 4 show the overlap of the GTLoam composition is associated with non-carbonate residual material (NCaRes) and alluvium (Alluv) for the surface soil, A- and C-horizons (Supplementary Tables 2, 3, 4). The pattern of elevated normalized votes coincides with the region described by Sayre et al. (2009) that is located in the north central US and south of the Great Lakes. The pattern of elevated GTLoam follows the course of the Mississippi River, which highlights the erosional path of this material. Supplementary Figs. 12a, b show the normalized and raw vote maps of the GTLoam prediction.

Normalized votes for the prediction of alluvium (Alluv) are shown in Fig. 17.13 (Supplementary Fig. 1). The overall prediction accuracy is ~50% (Table 17.4) and compositional overlap is observed with the surface lithology non-carbonate residual soil (NCaRes) (Supplementary Tables 2, 3, 4). High predictions of alluvium are located in Nevada, western Texas and the southeast US states. The dispersed prediction of 0.2–0.3 represents the regions of compositional overlap with NCaRes, which can be seen on the map of Fig. 8. Supplementary Figs. 1a, b show the normalized and raw vote maps of the Alluv prediction and supplementary Figs. 13a, b show the normalized and raw votes of the NCaRes prediction.

Figure 17.14 shows prediction based on the normalized votes for the Eolian Dunes (EolDune) of Nebraska, southward into Texas. The patterns are the same for the surface soil, A- and C-horizon maps. The highest values of normalized votes

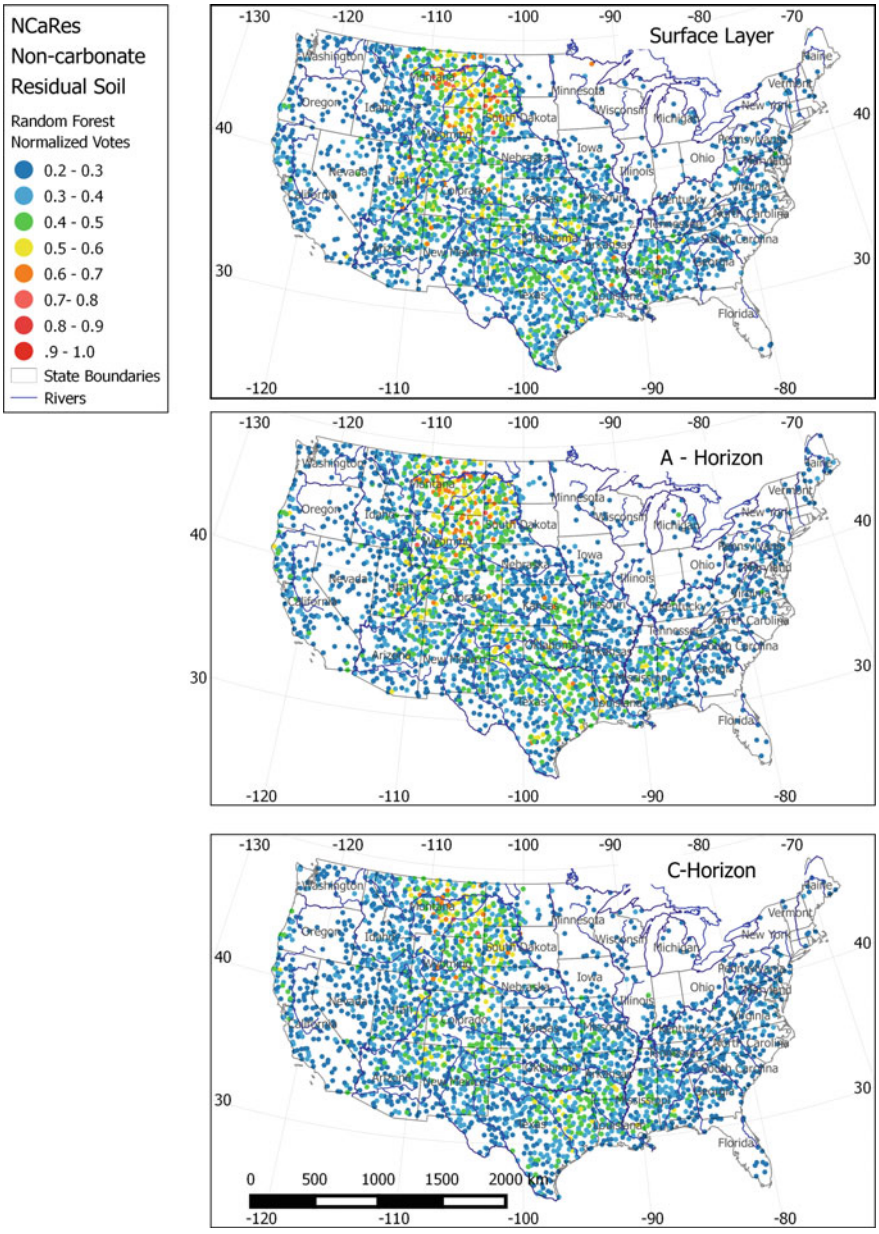

Fig. 17.11 Map of normalized votes for the surface lithology class, non-calcium residual soil (NCaRes). Sites with a normalized vote of less than 0.2 are omitted

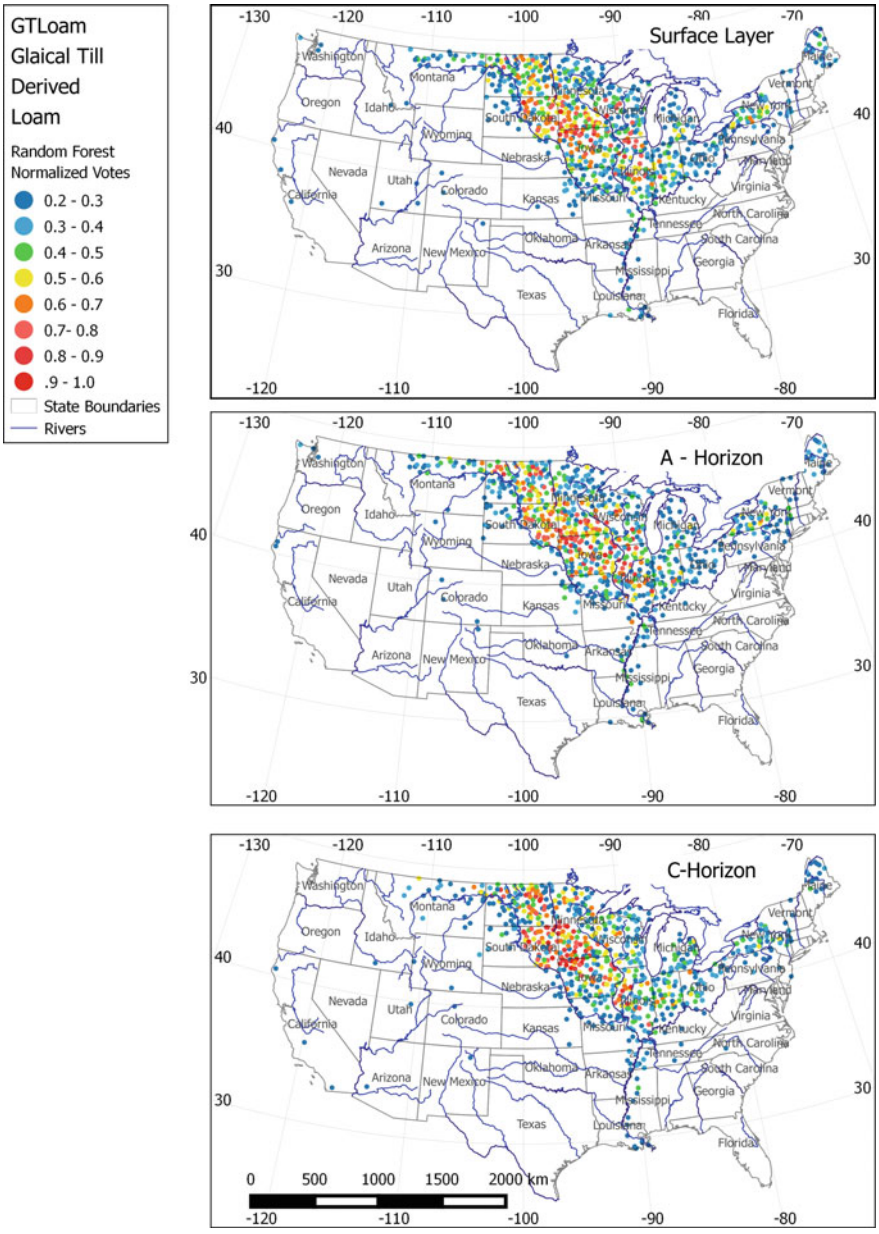

Fig. 17.12 Map of normalized votes for the surface lithology class, loam derived from glacial till (GTLoam). Sites with a normalized vote of less than 0.2 are omitted

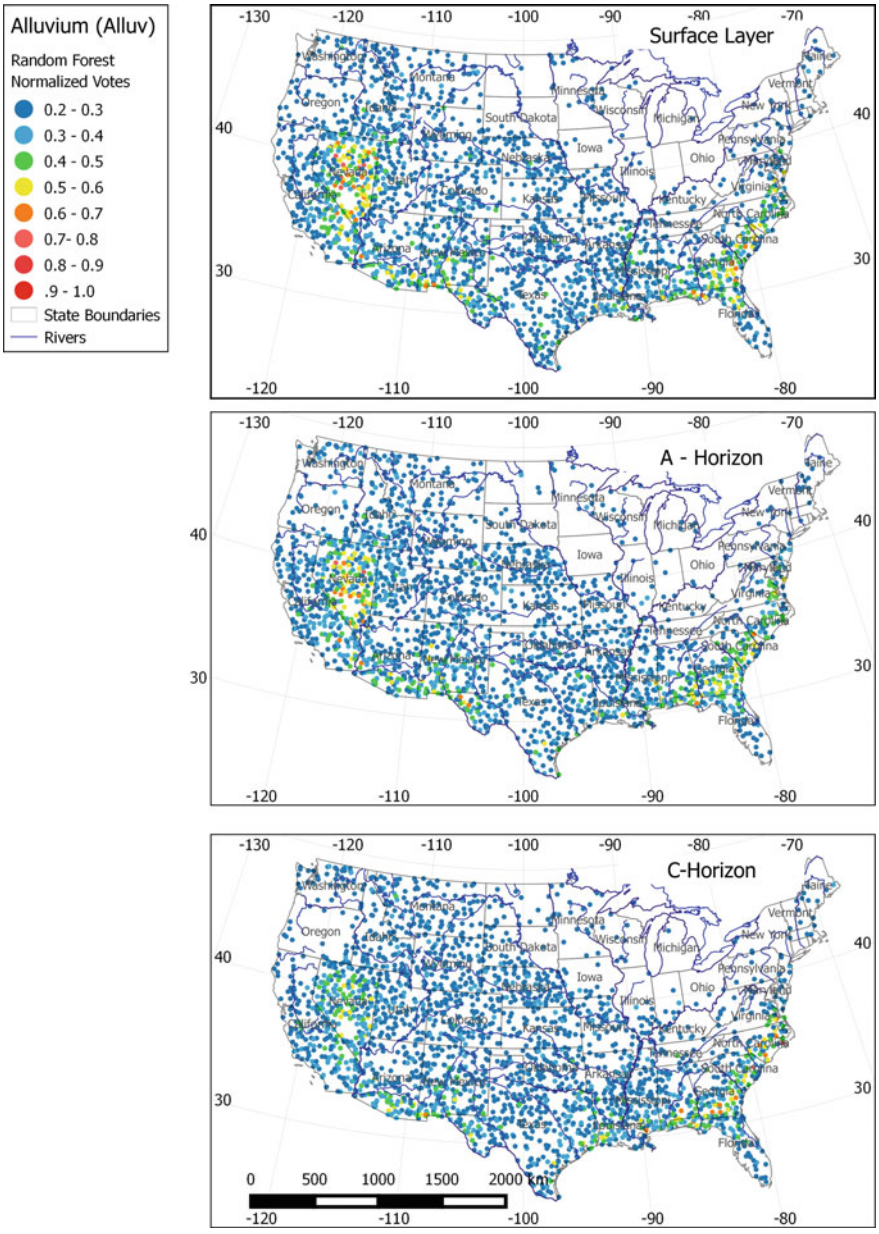

Fig. 17.13 Map of normalized votes for the surface lithology class, alluvium (Alluv). Sites with a normalized vote of less than 0.2 are omitted

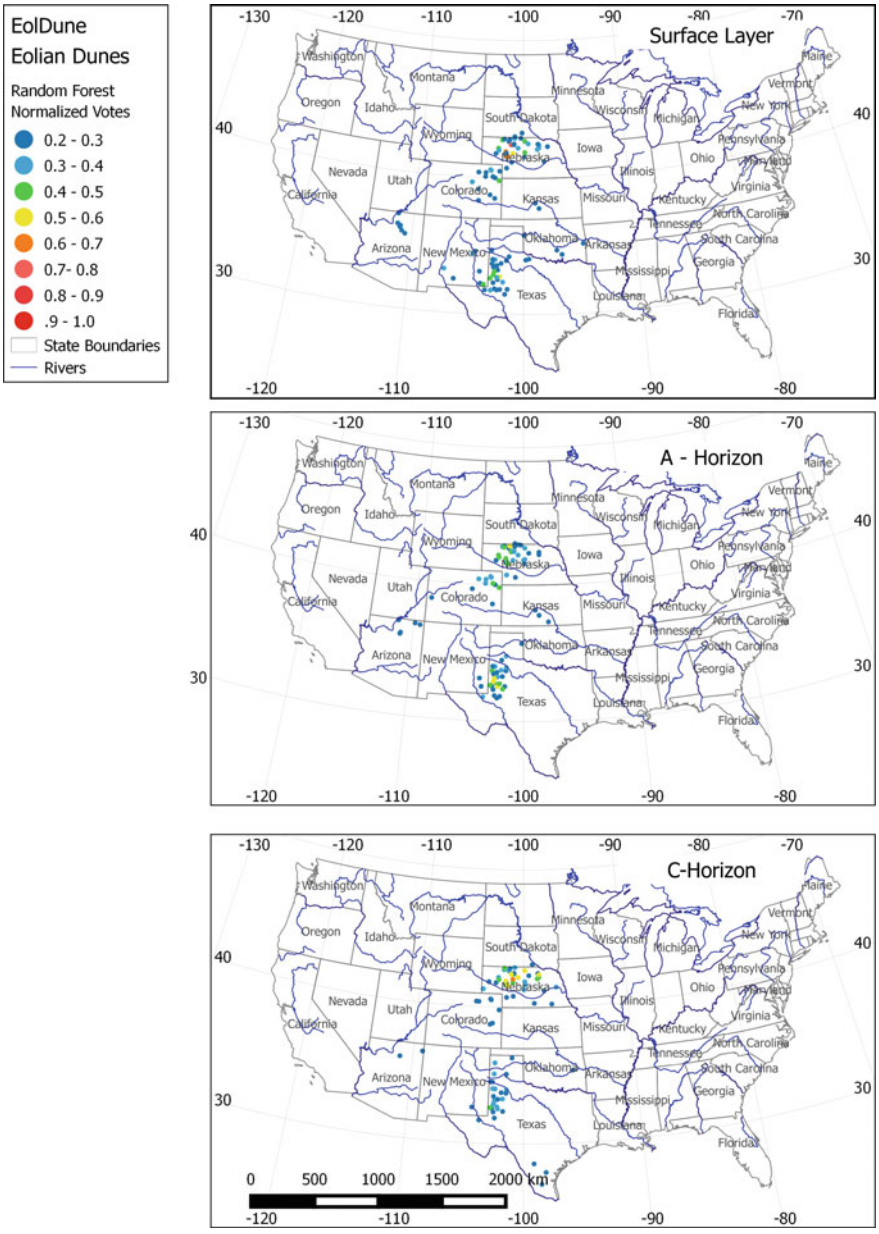

Fig. 17.14 Map of normalized votes for the surface lithology class, eolian dunes (EolDune). Sites with a normalized vote of less than 0.2 are omitted

occur in Nebraska and west-central Texas. The map of Sayre et al. (2009) shows EolDune in northern Texas and the Oklahoma Panhandle, although these two regions are not predicted in the surface soil, A- or C-Horizon results. Table 17.4 shows predictive accuracies of 22.3, 22.4 and 16.5% for the surface soil, A- and C-horizons, respectively. Supplementary Tables 2, 3 and 4 show that compositional overlap occurs with alluvium (Alluv) and non-carbonate residual soil (NCaRes). Supplementary Figs. 5a, b show the normalized and raw vote maps of the EolDune prediction.

The effects of erosion and subsequent re-deposition along the banks of the Mississippi River is observed for several of the surficial lithologies. NCaRes, CaRes and Colluv exhibit an erosional pattern along the Mississippi River, while EolLoess, GlLS, GlOut and GTLoam exhibit depositional patterns. This suggests that the recent deposition of the sediments along the banks of the Mississippi River has modified the composition of the upper layers of the soil. These classes (EolLoess, GlLS, GlOut, GTLoam—Supplementary Figs. 6a, b, 8a, b, 9a, b, 12a, b) show a distinct compositional presence down the length of Mississippi River starting from the northern Midwest states and reflecting continued transport of these materials at a continental scale.

A brief description of the maps for the surface soil, A and C-horizon data that are displayed in the Supplementary Annex are discussed in the section, Supplementary Material.

17.4 Discussion

Examination of the principal component biplots (Figs. 17.3 and 17.4) show that the multi-element patterns are very similar for the surface soil and A-horizon data. The C-horizon biplots show similar multi-element groupings, but the shape of the point patterns (Figs. 17.3d and 17.4d) are different from those of the surface soil and A-horizon (Figs. 17.3b, c and 17.4 b, c). As described previously, the element groupings for the three sampling layers are:

(1) Group 1: Tl-Rb-Be-Ba-K-Ga-Al-Sr-Na-Ca-Mg [felsic and mafic lithophile elements (silicates)]
(2) Group 2: Ni-Cr-V-Fe-Sc-Co-Cu-Zn-Mn [Ferromagnesian silicates and clays]
(3) Group3: Hg-Se-Mo-Sn-Bi-Pb-Sb-As-Ti-S-In. [Shales and organic material with adsorbed elements]

These associations are slight variants on Goldschmidt's classification of elements; lithophile (Group 1), siderophile (Group 2) and chalcophile (Group 3).

The principal component biplots, along with the maps of the dominant principal components (Figs. 17.5, 17.6 and 17.7), indicate that there is strong stoichiometric and geospatial control on the patterns that are observed. These patterns, both in the

biplots and the kriged map images, provide the justification to use the soil geochemical data to predictively map (validate) the surface lithology classification of Sayre et al. (2009). It should be noted that Sayre's map of surface lithologies does not distinguish lithologies with different mineralogies, and, hence there is considerable overlap between some of the classes defined by Sayre.

The results of the random forest classification show that for most of the surface lithology classes, the accuracy of prediction and spatial coherence of the predicted sites is variable, as shown in Table 17.4 and Figs. 17.11, 17.12, 17.13 and 17.14 and the Supplementary Tables and Figures. The surface lithologies with the lowest predictions are: Hydric Peat and Muck (HyPM), Alkalic Intrusives (AlkInt), Glacial Lake Sediments (GlLS), Extrusive Volcanic Rocks (ExtVR) and Saline Lake Sediments (SalLS). Two factors influence the classification accuracy. The first is the areal extent that a given class occupies. The compositional range of a class of small spatial extent may be swamped or masked by the compositional range of a class that is geographically adjacent to it and has a much larger areal extent. Surface lithologies such as AlkInt, HyPM ExtVr, SalLS and GlLS have limited geospatial extent and the compositions of these lithologies are similar to several other lithologies, including Alluv GTLoam and NCaRes. The second factor that influences the prediction accuracy is the common compositions of several of the surface lithology classes namely, alluvium (Alluv), non-carbonate residual soil (NCaRes), and silica-rich residual soil (SilRes). These surface lithologies are comprised of similar mineralogies and are, therefore, compositionally similar and result in compositional overlap in the statistically based prediction process.

Silicate mineralogy, including quartz, is under-represented in the data used for this study. As discussed previously, the quartz dilution effect has an influence on how the various relationships of the elements are observed, particularly in the methods that are part of the "Process Discovery" component of this study. The absence of silicon in the geochemical analysis in terms of the classifications may have some effect on the ability to distinguish between the different surface lithologies, but the exact effect is unknown at this time and further studies where Si is included and subsequently excluded in process discovery studies are warranted.

The validation of surface lithologies using soil geochemistry highlights some of the limitations on predicting distinct surface lithologies that have similar geochemical compositions but represent different processes. Despite this confusion of compositions between surface lithology classes, the predictive maps render a close representation of the maps of Sayre et al. (2009).

17.5 Concluding Remarks

The multi-element soil geochemistry over the conterminous United States contains a rich set of information that reflects the original source material and subsequent modification through weathering, mass transport, climate and biological activities.

As a result, continental-scale geochemistry may represent many processes. In this study, we have focused on the evaluation and interpretation of the multi-element soil geochemistry from the surface soil, A- and C-horizons in the context of predicting the surface lithologies.

Process discovery makes use of multivariate methods such as principal component analysis, which creates orthogonal linear combinations of the elements that often reflect processes controlled by mineral stoichiometry that comprise the parent material. This parent material may be bedrock (igneous, metamorphic, sedimentary), glacial deposits, loess or fluvial deposits. Ideally, soil geochemistry can be used to predict the composition of the underlying soil parent material. As demonstrated in this study, multivariate methods such as principal component analysis cannot decouple all of these processes. Processes such as igneous and metamorphic mineral reactions share similar mineral stoichiometry, making them indistinguishable from a geochemical perspective. Many distinct sedimentary assemblages are comprised of similar lithologies with similar mineralogy, and are thus difficult to distinguish solely on a geochemical basis.

With the exception of the surface lithology map of Sayre et al. (2009), a continental-scale map of lithology does not exist, which creates difficulty in an attempt to predictively map at large scales. However, the availability of the maps by Sayre et al. (2009) that include terrestrial ecosystems, thermoclimate, soil moisture and surface lithologies provides an opportunity to test the capacity of soil geochemistry to uniquely define these features. Although not presented here, the soil geochemistry has the ability to uniquely define terrestrial ecosystems and regional climate indicators. We intend to publish the results of using soil geochemistry to uniquely identify the terrestrial ecosystems, thermoclimatic zones and soil moisture (ombrotype) as defined by Sayre et al. (2009).

With few exceptions, there are only minor differences between the geochemical compositions of the surface soil and the A-horizon. The geochemistry of the C-horizon displays a distinct geochemical difference between the surface soil and A-horizon as it has not undergone the degree of weathering as the near-surface soils and contains less organic material.

The overall predictive accuracies for the predicting the surface lithologies for the surface soil, A- and C-horizons are 49.9%, 49.4% and 48.6%, respectively. As described above, the reasons for these low accuracies are due to the overlap of many of the lithologies with Alluvium, Non-carbonate residual soils, Siliceous soils, Eolian Dunes, Eolian Loess and materials deposited from glaciation. However, the spatial continuity of the posterior probabilities confirm the distinctiveness of these lithologies and demonstrate the effectiveness of soil geochemistry in recognizing the differences between the classes.

The geochemistry of soils represents modification of the initial parent material through weathering in response to varying precipitation and temperature, groundwater effects, meteoric water effects, biologic activity and geologic complexity.

Thus, geochemistry is a rich source of information that can be used in many ways to describe, monitor and predict processes derived from natural and anthropogenic events (Grunsky et al. 2013).

The results from the statistical evaluation of the geochemical data in the context of predicting surface lithologies across the conterminous US indicates that soil geochemistry reflects a number of physical processes. Further studies of the soil geochemistry across the US will evaluate the ability to predict terrestrial ecosystems and indicators of climate.

Acknowledgements The authors thank Karl Ellefson of the United States Geological Survey for his thoughtful and helpful review of the manuscript.

References

Aitchison J (1986) The statistical analysis of compositional data. Chapman and Hall, New York, p 416

Breiman L (2001) Randomforests. Machine. Learning 45:5–32

Breiman L, Cutler A (2016) Random forests. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#intro

Commission for Environmental Cooperation (CEC) (1997) Ecological regions of North America—toward a common perspective. Montreal: commission for Environmental Cooperation, p 71

Drew LD, Grunsky EC, Sutphin DM, Woodruff LG (2010) Multivariate analysis of the geochemistry and mineralogy of soils along two continental-scale transects in North America. Sci Total Environ 409:218–227. https://doi.org/10.1016/j.scitotenv.2010.08.004

Eberl DD, Smith DB (2009) Mineralogy of soils from two continental-scale transects across the United States and Canada and its relation to soil geochemistry and climate. Appl Geochem 24(8):1394–1404

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

Garrett RG (2009) Relative spatial soil geochemical variability along two transects across the United States and Canada. Appl Geochem 24(8):1405–1415

Goldschmidt VM (1937) The principal of distribution of chemical elements in minerals and rocks. The seventh Hugo Muller lecture, delivered before the Chemical Society on March 17th, 1937. J Chem Soc 1937, 665–673. https://doi.org/10.1039/jr9370000655

Grunsky EC (2001) A program for computing rq-mode principal components analysis for S-Plus and R. Comput Geosci 27:229–235

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

Grunsky EC, de Caritat P, Mueller UA (2017) Using surface regolith geochemistry to map the major crustal blocks of the Australian continent. Gondwana Res 46:227–239. https://doi.org/10.1016/j.gr.2017.02.011

Grunsky EC, McCurdy MW, Perhsson SJ, Peterson TD, Bonham-Carter GF (2012) Predictive geologic mapping and assessing the mineral potential in NTS 65A/B/C, Nunavut, with new regional lake sediment geochemical data. Geological Survey of Canada, Open File 7175, 1 sheet. https://doi.org/10.4095/291920

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 Explor. https://doi.org/10.1016/j.gexplo.2013.07.013

Grunsky EC, Drew LJ, Woodruff LG, Friske PWB, Sutphin DM (2013) Statistical variability of the geochemistry and mineralogy of soils in the maritime provinces of Canada and part of the northeast United States. Geochem Explor Environ Anal 13(2013):249–266. https://doi.org/10.1144/geochem2012-138

Harris JR, Schetselaar EM, Lynds T, deKemp EA (2008) Remote predictive mapping: a strategy for geological mapping of Canada's North, chapter 2. In: Harris JR (ed) Remote predictive mapping: an aid for Northern mapping. Geological Survey of Canada, Ottawa, Ontario, Canada, pp 5–27 (OpenFile5643)

Harris JR, Grunsky EC (2015) Predictive lithological mapping of Canada's North using random forest classification applied to geophysical and geochemical data. Comput Geosci 80:9–25

Jolliffe IT (2002) Principal component analysis, 2nd edn. Springer, New York, p 487

Mueller UA, Grunsky EC (2016) Multivariate spatial analysis of lake sediment geochemical data. Melville Peninsula, Nunavut, Canada, Applied Geochemistry, https://doi.org/10.1016/j.apgeochem.2016.02.007

Palarea-Albaladejo J, Martín-Fernández JA, Buccianti A (2014) Compositional methods for estimating elemental concentrations below the limit of detection in practice using R. J Geochem Explor. ISSN 0375–6742. https://doi.org/10.1016/j.gexplo.2013.09.003. Accessed 28 Sept 2013

Pawlowsky-Glahn V, Egozcue J-J (2015) Spatial analysis of compositional data: a historical review. J Geochem Explor 164:28–32. https://doi.org/10.1016/j.gexplo.2015.12.010

Pebesma EJ (2004) Multivarilabe geostatistics in S: the gstat package. Comput Geosci 30:683–691

QGIS Development Team (2016) QGIS geographic information system. Open Source Geospatial Foundation Project. http://www.qgis.org/

R Core Team (2013) R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. http://www.r-project.org

Sayre R, Comer P, Warner H, Cress J (2009) A new map of standardized terrestrial ecosystems of the conterminous United States: U.S. Geological Survey Professional Paper 1768, 17 p. http://pubs.usgs.gov/pp/1768

Smith DB (2009) Geochemical studies of North American soils: results from the pilot study phase of the North American Soil Geochemical Landscapes Project. Appl Geochem 24(8):1355–1356. https://doi.org/10.1016/j.apgeochem.2009.04.006

Smith DB, Woodruff LG, O'Leary RM, Cannon WF, Garrett RG, Kilburn JE, Goldhaber MB (2009) Pilot studies for the North American Soil Geochemical Landscapes Project—site selection, sampling protocols, analytical methods, and quality control protocols. Appl Geochem 24(8):1357–1368

Smith DB, Cannon WF, Woodruff LG (2011) A National–scale geochemical and mineralogical survey of soils of the conterminous United States. Appl Geochem 26:S250–S255

Smith DB, Cannon WF, Woodruff LG, Rivera FM, Rencz AN, Garrett RG (2012) History and progress of the North American Soil Geochemical Landscapes Project, 2001–2010. Earth Sci Front 19(3):19–32

Smith DB, Cannon WF, Woodruff LG, Solano F, Kilburn JE, Fey DL (2013) Geochemical and mineralogical data for soils of the conterminous United States. U.S. Geological Survey Data Series 801, 19 p. http://pubs.usgs.gov/ds/801/

Smith DB, Cannon WF, Woodruff LG, Solano F, Ellefsen KJ (2014) Geochemical and mineralogical maps for soils of the conterminous United States. U.S. Geological Survey Open-File Report 2014–1082, 386 p. https://pubs.usgs.gov/of/2014/1082/

Venables WN, Ripley BD (2002) Modern applied statistics with S, 4th edn. Springer, Berlin

Woodruff LG, Cannon WF, Eberl DD, Smith DB, Kilburn JE, Horton JD, Garrett RG, Klassen RA (2009) Continental-scale patterns in soil geochemistry and mineralogy: results from two transects across the United States and Canada. Appl Geochem 24(8):1369–1381

Zhou D, Chang T, Davis JC (1983) Dual extraction of R-mode and Q-mode factor solutions. Math Geol 15(5):581–606