

Web Appendix: Supplementary materials for “Biplots for understanding machine learning predictions in digital soil mapping”

Stephan van der Westhuizen^{a,b,e,*}, Gerard B. M. Heuvelink^{b,c}, Sugnet Gardner-Lubbe^{a,e}, Catherine E. Clarke^d

^a*Dept. of Statistics and Actuarial Science, Stellenbosch University, Stellenbosch, South Africa*

^b*Soil Geography and Landscape Group, Wageningen University, Wageningen, The Netherlands*

^c*ISRIC-World Soil Information, Wageningen, The Netherlands*

^d*Dept. of Soil Science, Stellenbosch University, Stellenbosch, South Africa*

^e*Centre for Multi-dimensional Data Visualisation (MuViSU), Stellenbosch University, Stellenbosch, South Africa*

Appendix A. Additional information on biplots

Appendix A.1. Gabriel biplot

An example of a Gabriel biplot is provided in Figure A.1. The biplot is constructed from the 2018 Land Use and Coverage Area frame Survey (LUCAS) which has been aggregated by country. That is, for each country we determined the median soil organic carbon (SOC), total nitrogen, pH and particle-size fractions. The points in Figure A.1 have been labelled by the names of the countries, and the variables are displayed as red arrows. Note that the vertical and horizontal sides of the graph are called scaffolding axes and are irrelevant. The length of the arrows approximates the variances of the variables such that a longer line represents a larger variance. For instance, we note from the biplot that sand has a larger variance compared to clay. The angle between the lines

*Corresponding author

Email address: stephanvdw@sun.ac.za (Stephan van der Westhuizen)

(i.e., the cosine of the angle between the lines) approximates the correlation between the variables. An angle close to 90° or 270° represents no correlation, and an angle of 0° or 180° represents a perfect correlation of 1 or -1 . In Figure A.1 we note a strong correlation between SOC and nitrogen, and a weak correlation between pH and silt. Each point can be read off from a variable axis by projecting it perpendicularly to the line. For example, if Poland, Denmark, and The Netherlands are projected onto the Sand axis we note high values for the median sand content for these countries while in comparison we note low values for Croatia and Serbia (one can extend the axis beyond the red line). The distance between two points approximates the Euclidean distance between those two points in the multivariate space. Therefore, points that are far away from each other have a large Euclidean distance, and vice versa. This property of a biplot allows the user to detect clusters. For example, in light of median sand content, we note that Finland, Norway and Sweden are close to each other, but far away from Croatia.

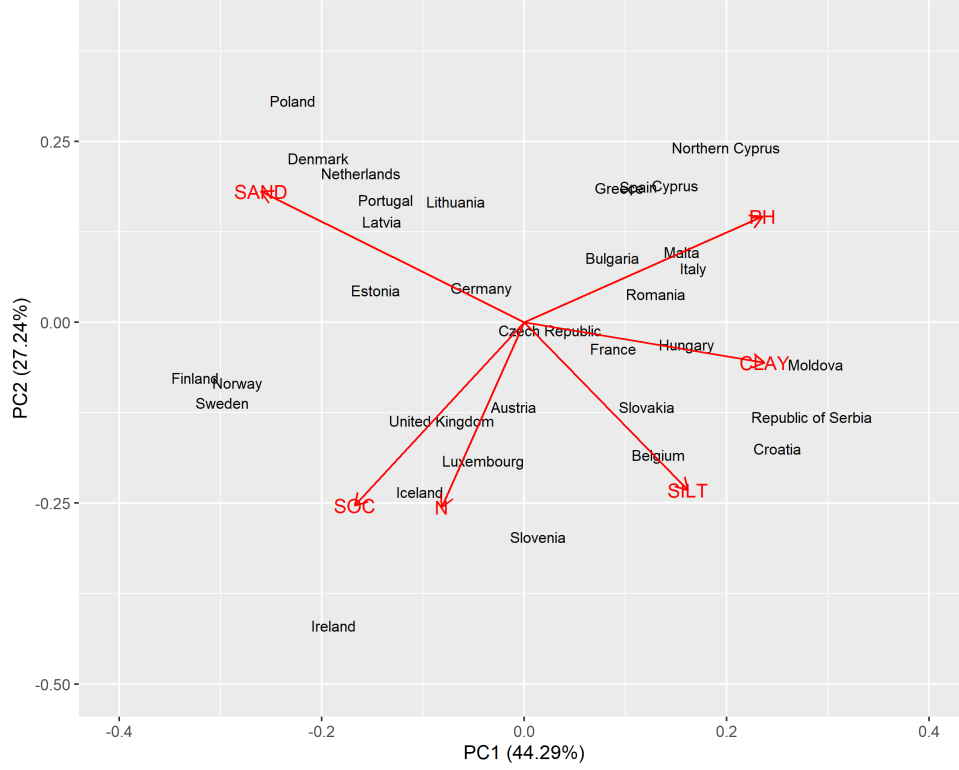


Figure A.1: A Gabriel biplot showing an optimal two-dimensional display of the Land Use and Coverage Area frame Survey data aggregated by country. For each country the median of the particular soil property was obtained.

Appendix A.2. Mathematical background for PCA biplots

Suppose that \mathbf{X} , an $(n \times p)$ standardised data matrix (i.e., columns are zero-mean and have unit variances), with n the number of observations or locations, p the number of covariates and $n \geq p$, is to be represented in an r -dimensional display, with $r < p$. principal component analysis (PCA) can be used to approximate \mathbf{X} in r dimensions, denoted by $\hat{\mathbf{X}}$, such that the least squares error between \mathbf{X} and $\hat{\mathbf{X}}$ is a minimum. PCA makes use of the singular value decomposition (Everitt

et al., 2001) of \mathbf{X}

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T, \quad (\text{A.1})$$

where \mathbf{U} is a $(n \times k)$ orthonormal matrix with $k = \min(n, p)$, and \mathbf{V} is a $(p \times p)$ orthonormal matrix with the eigenvectors of \mathbf{X} . \mathbf{D} is a $(n \times k)$ matrix with the singular values of \mathbf{X} , λ_k , as the (k, k) entries for $k = 1, 2, \dots, \min(n, p)$. Note that when $k = p$, \mathbf{D} reduces to a diagonal $(p \times p)$ matrix. Note also that the singular values of \mathbf{X} are the square roots of the eigenvalues of $\mathbf{X}^T\mathbf{X}$ (Everitt et al., 2001). The approximated data matrix is determined with

$$\hat{\mathbf{X}} = \mathbf{U}\mathbf{D}\mathbf{J}\mathbf{V}^T, \quad (\text{A.2})$$

where \mathbf{J} is a $(p \times p)$ matrix written as

$$\mathbf{J} = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

and \mathbf{I}_r is a $(r \times r)$ identity matrix. Finally, to obtain a biplot, the principal component scores, (i.e., the points in the biplot) $\mathbf{X}\mathbf{V}\mathbf{J} = \mathbf{U}\mathbf{D}\mathbf{J}$, are plotted. To add the lines, also called the principal component loadings, the columns of $\mathbf{V}\mathbf{J}$ are plotted. To obtain a two-dimensional display one would set $r = 2$ so that two principal components are plotted. Note that the first two are usually selected, but the first and third or other binary combinations of principal components can also be visualised in a two-dimensional display.

To assess the goodness-of-fit of the biplot one can obtain the “quality” of the biplot. This is calculated as the proportion of the variances of the columns in \mathbf{X} explained by $\hat{\mathbf{X}}$ (Gower et al., 2011)

$$\frac{\text{trace}(\mathbf{\Sigma}\mathbf{J})}{\text{trace}(\mathbf{\Sigma})}, \quad (\text{A.3})$$

where $\Sigma = \mathbf{D}^2$ is the variance-covariance matrix of \mathbf{XV} . Eq. (A.3) yields a value between zero and one with one indicating a perfect fit. The quality of the biplot in Figure A.1 was 0.715 which is the sum of the variances explained by the first two principal components, $0.443 + 0.272 = 0.715$.

Predictivity of a variable axis or a point in a biplot refers to how well the axis or point is approximated in the two dimensional display. The predictivity of the axes and the points of a biplot are given by (Gower et al., 2011)

$$\text{diag}(\hat{\mathbf{X}}^T \hat{\mathbf{X}})(\text{diag}(\mathbf{X}^T \mathbf{X}))^{-1}, \quad (\text{A.4})$$

and

$$\text{diag}(\hat{\mathbf{X}} \hat{\mathbf{X}}^T)(\text{diag}(\mathbf{X} \mathbf{X}^T))^{-1}, \quad (\text{A.5})$$

respectively. Eqns. (A.4) and (A.5) yield values between zero and one for each axis and point, respectively. A value close to one means that the corresponding axis or point is very accurately approximated in the biplot. The predictivity of the axes for the biplot in Figure A.1 are shown in Table A.1 while the predictivity of the points are shown in Table A.2. For example, these results illustrate that the axis for sand is very accurately represented in the biplot while the axis for N is not so accurate. In addition, the point for Finland is very accurate and hence is reliable, but the point for the Czech Republic is very inaccurately represented and care should be taken to interpret or project this point onto an axis.

Table A.1: Axes predictivity of the LUCAS biplot.

Variable	Predictivity
Sand	0.970
pH	0.758
SOC	0.747
Silt	0.654
Clay	0.646
N	0.517

Table A.2: Point predictivity of the LUCAS biplot.

Finland 0.992	Denmark 0.980	Latvia 0.978	Sweden 0.966	Italy 0.964	Poland 0.951
Spain 0.929	Netherlands 0.928	Slovenia 0.900	Greece 0.899	Lithuania 0.898	Hungary 0.886
N. Cyprus 0.878	Estonia 0.855	Portugal 0.851	Norway 0.829	Cyprus 0.828	Austria 0.803
Serbia 0.803	Slovakia 0.792	Croatia 0.790	France 0.738	Luxembourg 0.687	Malta 0.686
Ireland 0.666	Bulgaria 0.624	Germany 0.605	Belgium 0.600	UK 0.537	Moldova 0.400
Romania 0.269	Iceland 0.225	Czech Republic 0.082			

Appendix B. Modelling results

Figure B.1(a) presents a density plot illustrating the distribution of topsoil SOC in South Africa. Additionally, Figure B.1(b) displays separate density plots for the provinces of the country. The distribution of SOC across the entire nation was skewed to the right, with roughly 90% of observations falling below 2.5%. KwaZulu-Natal exhibited the highest concentration of SOC, followed by Mpumalanga, while the Northern Cape displayed the lowest concentration with approximately 90% of observations falling below 0.8%.

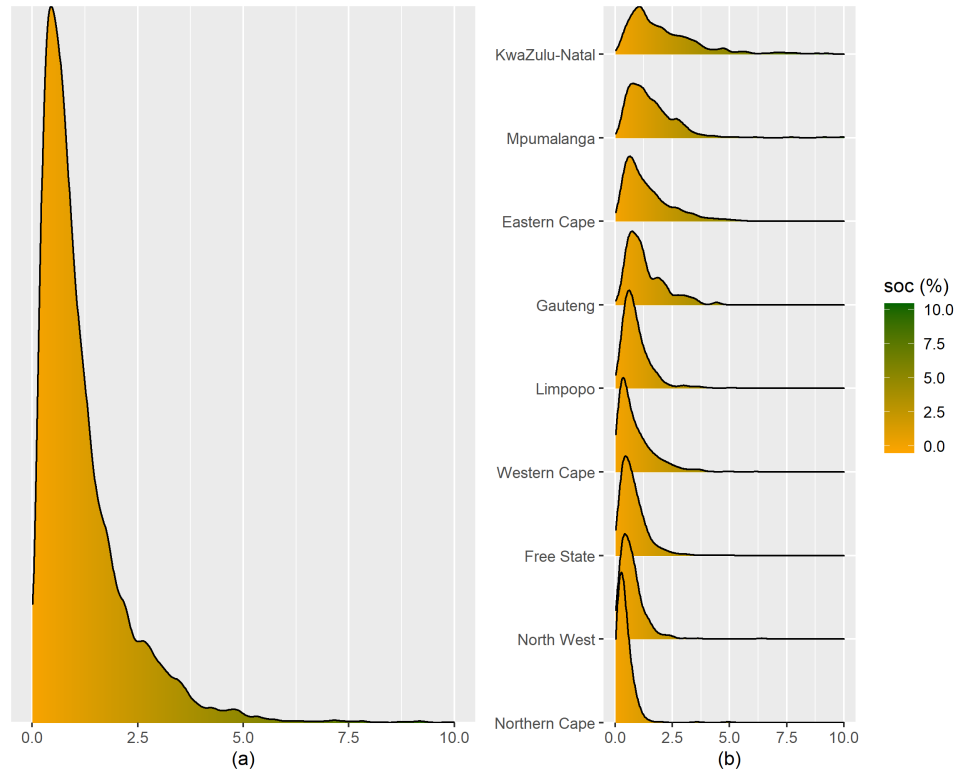


Figure B.1: (a) Density plot of topsoil SOC for South Africa. (b) SOC densities shown for each province separately.

Table B.1: Validation results of the random forest (RF) model for the outer-folds of the 10-fold nested cross-validation. The results per fold are also shown.

Outer-fold	ME	RMSE	MEC	CCC
1	0.027	0.638	0.598	0.737
2	0.056	0.693	0.541	0.707
3	0.034	0.681	0.590	0.734
4	0.011	0.832	0.547	0.678
5	0.039	0.697	0.541	0.711
6	0.023	0.697	0.567	0.723
7	0.043	0.728	0.493	0.670
8	0.048	0.717	0.532	0.691
9	0.010	0.670	0.578	0.731
10	-0.016	0.793	0.533	0.667
mean	0.027	0.715	0.552	0.705

	Covariates	Importance
1	CLM_MOD_CCYRAVG	0.808
2	MOR_ENV_DEMM	0.625
3	MOR_MRG_VDP	0.373
4	CLM_MOD_CC12AVG	0.287
5	CLM_MOD_CC09AVG	0.262
6	MOR_MRG_CRV	0.245
7	CLM_MOD_LSTD09STD	0.197
8	CLM_MOD_LSTN12STD	0.193
9	MOR_MRG_TPI	0.192
10	LUC_GFC_TRELY10	0.170
11	CLM_WCL_BIO18	0.169
12	CLM_MOD_LSTN05STD	0.168
13	VEG_MOD_EVIYRSTD	0.166
14	CLM_WCL_BIO16	0.163
15	CLM_MOD_LSTN04STD	0.153
16	CLM_MOD_LSTD04STD	0.145
17	SAT_MOD_MIRYRAVG	0.136
18	CLM_WCL_BIO08	0.135
19	MOR_MRG_NEG	0.135
20	CLM_MOD_LSTD01STD	0.130
21	CLM_WCL_P12TOT	0.127
22	CLM_MOD_LSTN06STD	0.124
23	CLM_MOD_LSTN09STD	0.111
24	MOR_MRG_TWI	0.111
25	VEG_MOD_NPPY15	0.107
26	VEG_MOD_EVIMAX	0.103
27	CLM_WCL_P09TOT	0.102
28	SAT_MOD_NIR10AVG	0.101
29	SAT_MOD_NIR05AVG	0.100
30	CLM_MOD_LSTD08STD	0.097
31	CLM_WCL_P10TOT	0.095
32	VEG_MOD_EVI07AVG	0.094
33	MOR_MRG_CRU	0.093
34	GEO_GLM_L04	0.093
35	MOR_MRG_POS	0.092
36	CLM_MOD_LSTD02STD	0.082
37	CLM_MOD_LSTD08AVG	0.082
38	SAT_MOD_NIR09AVG	0.082
39	MOR_MRG_DVM	0.082
40	CLM_MOD_LSTD12STD	0.079

Table B.2: Top 40 covariates in the RF model as indicated by the mean decrease in accuracy. For details about the covariates refer to Poggio et al. (2021).

Table B.3: Correlation matrix of SOC and 5 of the top RF predictors, calculated from the calibration data.

	CCYRAVG	CC12AVG	CC09AVG	LSTD09STD	DEM	soc
CCYRAVG	1.000	0.841	0.632	-0.423	0.072	0.531
CC12AVG		1.000	0.188	-0.610	0.329	0.434
CC09AVG			1.000	0.043	-0.444	0.415
LSTD09STD				1.000	0.046	-0.158
DEM					1.000	0.122
soc						1.000

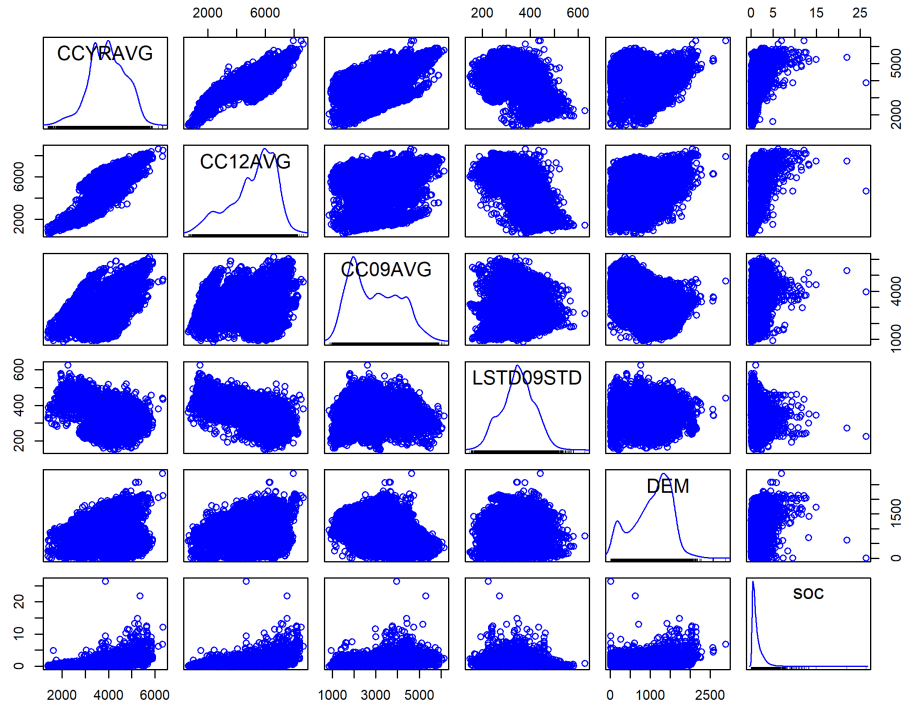


Figure B.2: Scatterplot matrix of SOC and 5 of the top RF predictors, calculated from the calibration data.

Appendix C. Important covariates based on Shapley values

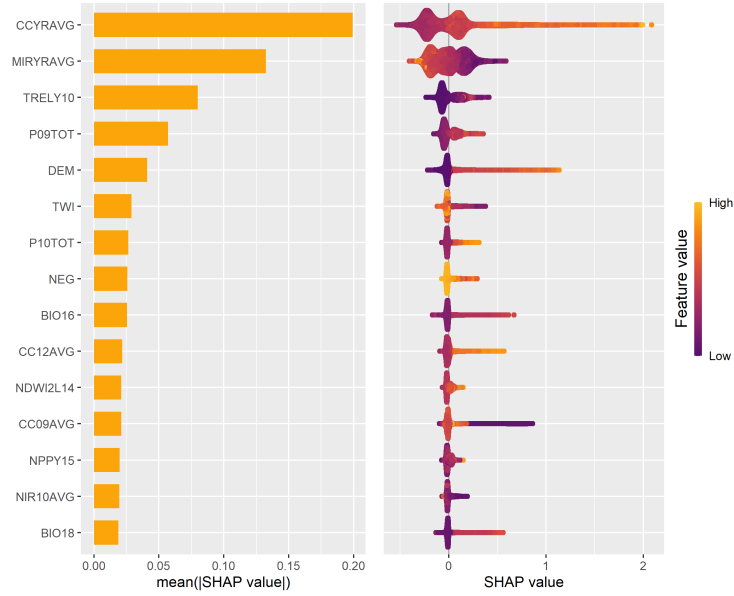


Figure C.1: Most important covariates in the RF model fitted on the South African SOC data based on the average absolute Shapley values. For details concerning the covariates we refer the reader to Poggio et al. (2021).

Appendix D. Biplot with Clorpt covariates

The biplot with the covariates based on the climate organisms relief parent material time (clorpt) model is presented in Figure D.1. The climate factors included covariates such as average annual temperature for the wettest quarter (AVGTWQ) in Celsius, total precipitation of the wettest quarter (PWETQ), driest quarter (PDRYQ), warmest quarter (PWARQ) and total precipitation of the coldest quarter (PCOLQ), and total precipitation for December (P12TOT), all of which are measured in mm. The organisms factors included the annual average EVI index (EVIYRAVG) and grasslands cover for 2010 (GRSLNDS) in percentage. The relief factors included DEM and VDP and various land forms such as foothills (LFBRKS), flat plains (LFPLNS), hills (LFHLS), low hills (LFLOHLS), low mountains (LFLOMNT) and smooth plains (LFSPLNS), all of which are measured in percentage. Parent material, represented by different rock types, was omitted due to its insignificant relationship to SOC in our data set. The quality of the biplot was 0.456, and the most predictive axes were that of the precipitation covariates ranging from 0.651 to 0.969 and that of EVIYRAVG with a predictivity measure of 0.755. The least predictive axes were that of the land forms with measures ranging between 0.033 and 0.104.

One covariate that stood out with regards to the low to high (left-to-right) variation in SOC predictions was EVIYRAVG. Specifically, when EVIYRAVG was more than 3 000, SOC predictions were mostly larger than 1.7%. Other covariates that also contributed to explaining these variations in SOC predictions were the land forms and total precipitation of the driest quarter (seen from the top-left corner to the bottom-right corner), as well as total precipitation of the wettest and warmest quarter as the total precipitation in December (mid-left to top-right). However, caution should be exercised when interpreting the axes of the land forms as these had low predictivity measures. It is also interesting to

note that relief factors such as DEM and VDP did not contribute much to the global variation in SOC predictions (seen vertically from mid-top to bottom).

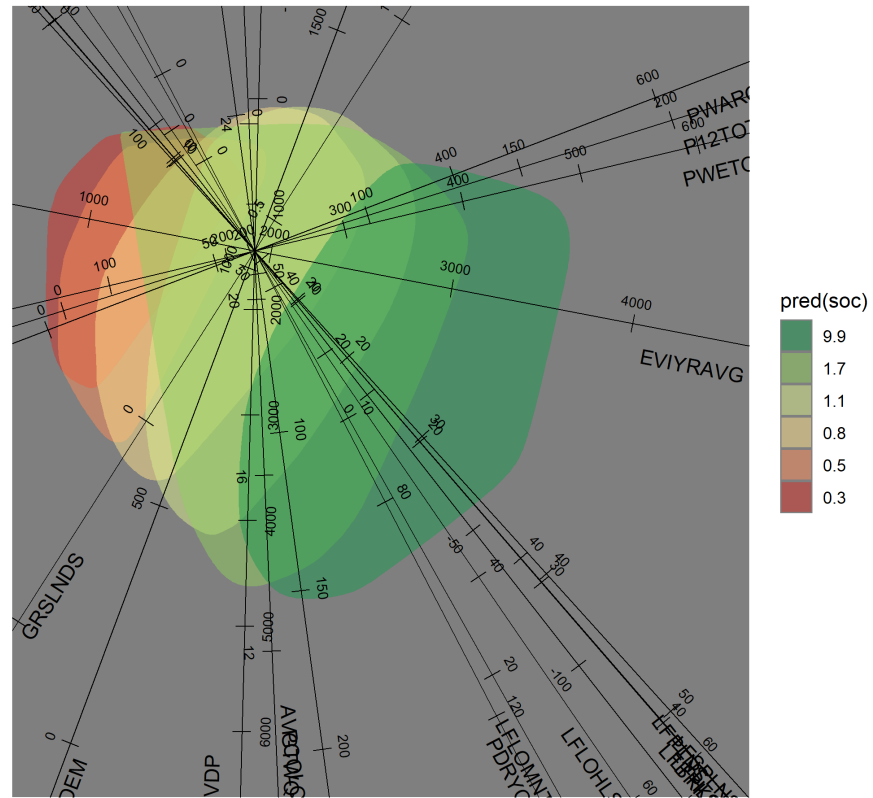


Figure C.1: Biplot with covariates that represent Jenny's clorpt model. The biplot had a quality of 0.456.

Appendix E. Biplots and other XML methods at regional scale

We selected a region in Kwazulu-Natal (see Figure E.1) to perform the analyses of explainable machine learning (XML) and that of the proposed biplot methodology. This region was selected, because of its importance in agriculture. In Figure E.2 a biplot with the colour arrangements according to the predictions is shown. In comparison to the biplot in Figure 8, relief factors such as VDP differentiate between larger and smaller predictions more so at regional scale compared to national scale.

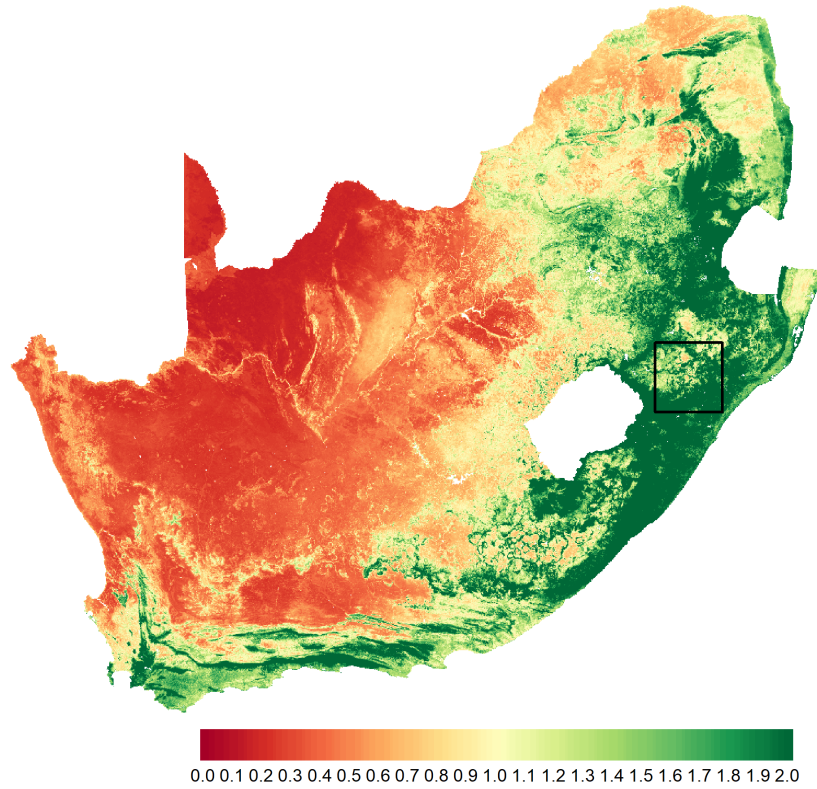


Figure E.1: Highlighted region in South Africa used for regional XML analysis.

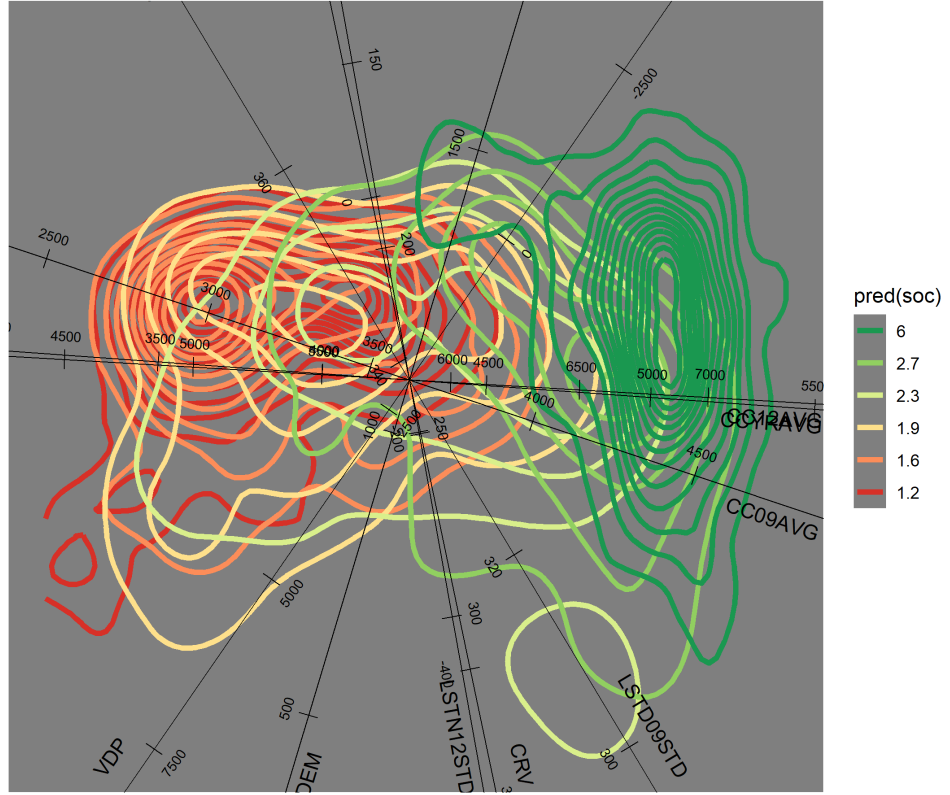
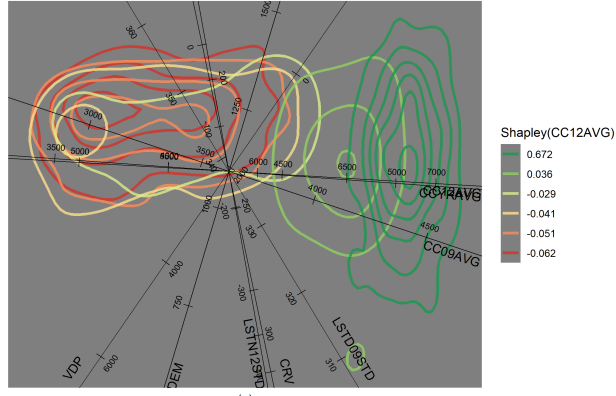
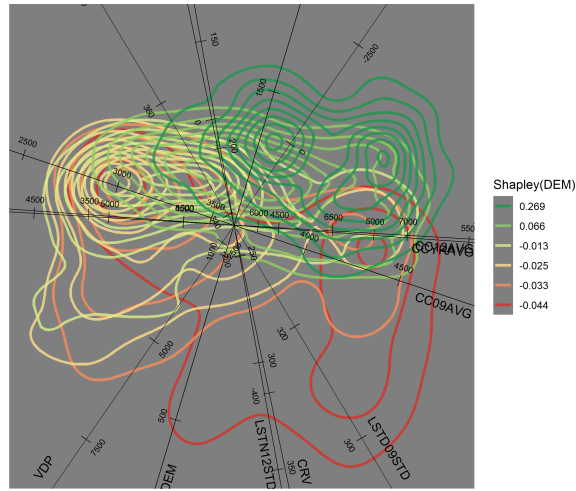


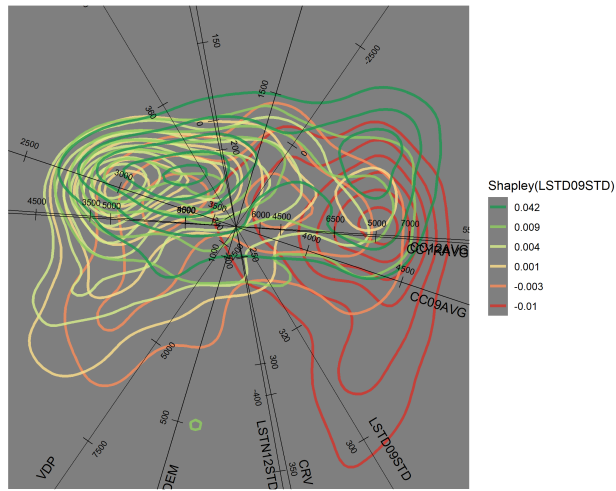
Figure E.2: PCA biplot showing the relationship between covariates and the predictions made by the RF model for the highlighted region in KwaZulu-Natal. Predictions for SOC are presented in %. The quality of the biplot is 0.653. The respective predictivity measures of the covariates were (sorted from highest to lowest): 0.961 for CCYRAVG, 0.933 for CC12AVG, 0.909 for CC09AVG, 0.752 for DEM, 0.704 for VDP, 0.429 for LSTN12STD, and 0.237 for LSTD09STD.



(a)



(b)



(c)

Figure E.3: The biplot in Figure E.2 with colour arrangements according to the Shapley values of (a) CC12AVG, (b) DEM, and (c) LSTD09STD.

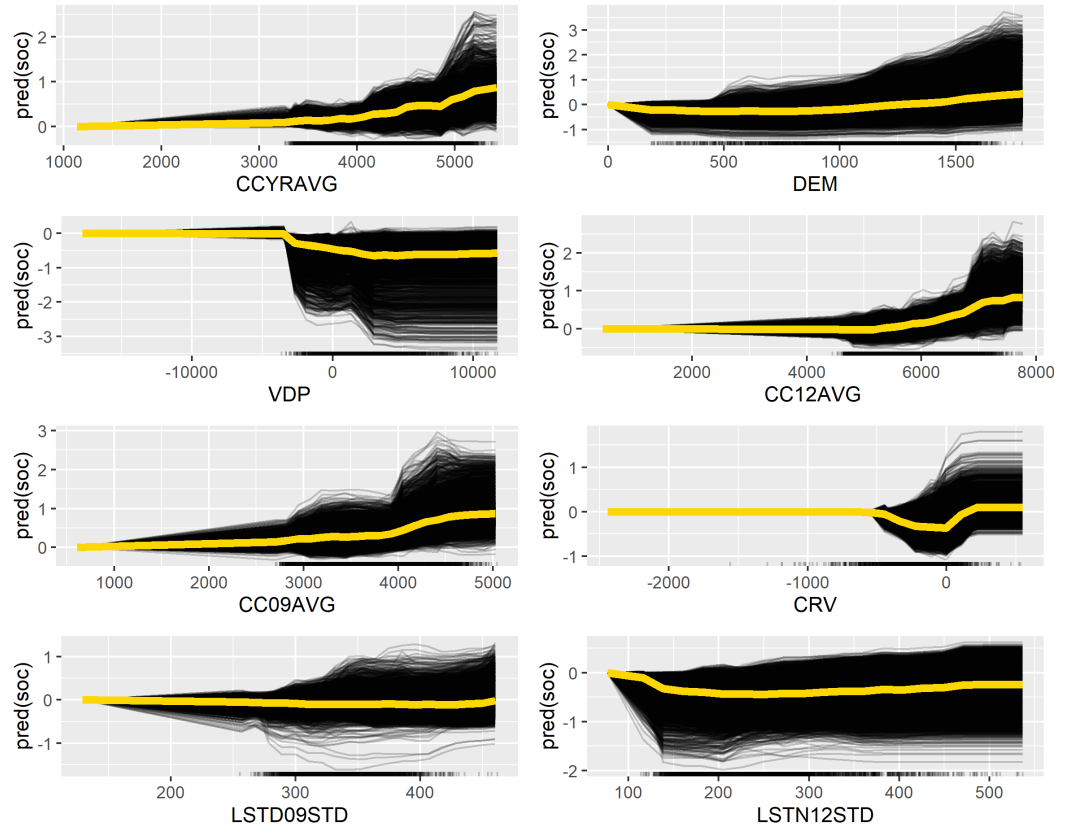


Figure E.4: Centred independent conditional expectation (ICE) curves shown in black with the corresponding partial dependence (PD) curves in yellow for the top 8 covariates according to the RF model for the region in KwaZulu-Natal. Predictions for SOC are presented in %.

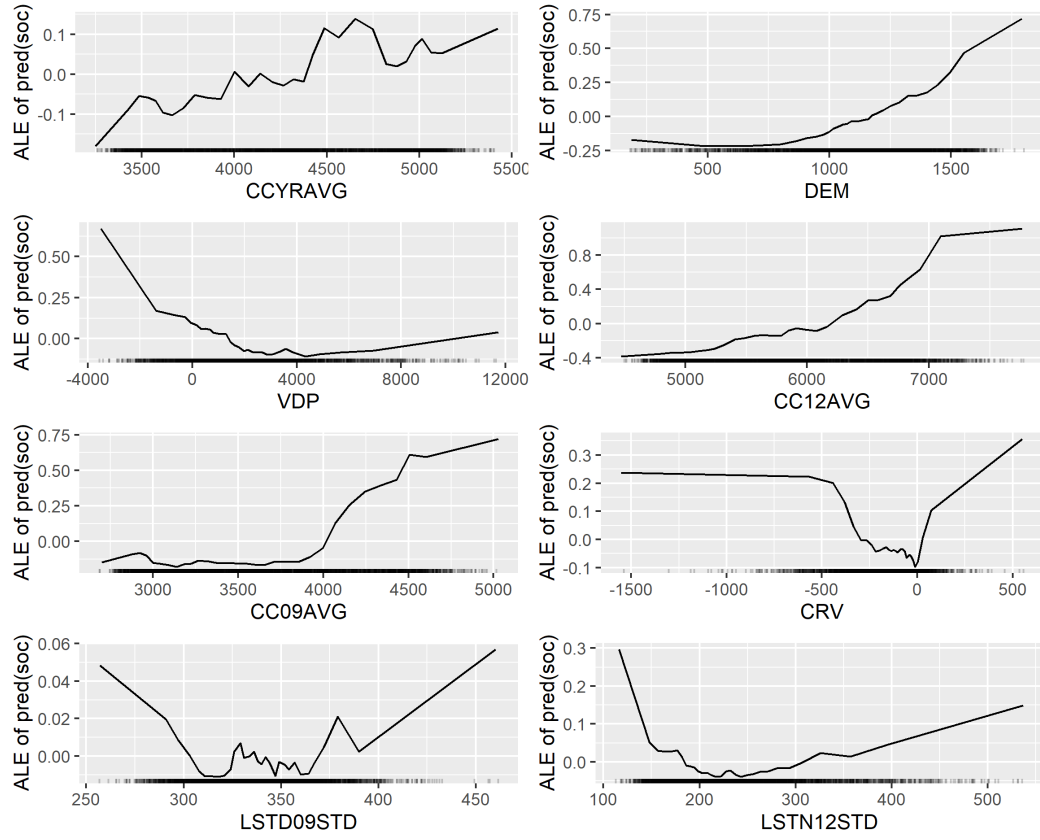


Figure E.5: Accumulated local expectation (ALE) plots the top 8 covariates in the RF model for the region in KwaZulu-Natal.

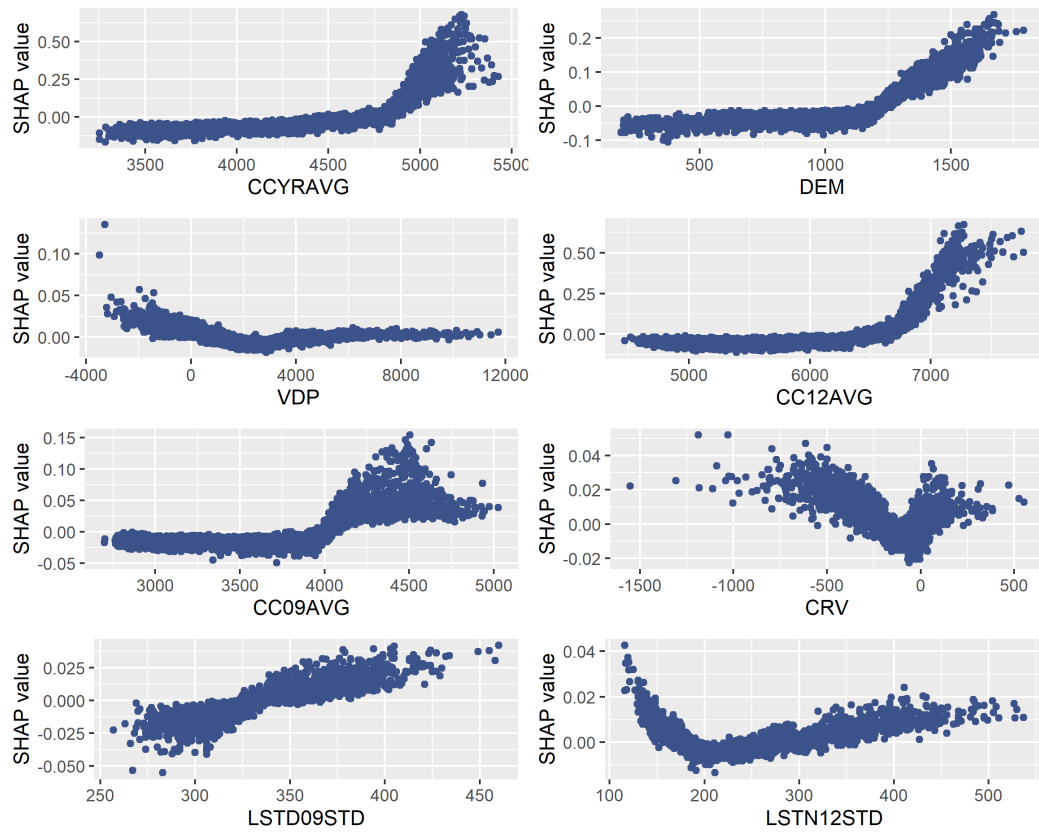


Figure E.6: Partial dependence plots with Shapley values for the top 8 covariates in the RF model for the region in KwaZulu-Natal.

Appendix F. Biplot R code

Link to [github](#)¹.

References

- Everitt, B., Dunn, G. et al. (2001). *Applied multivariate data analysis* volume 2. Wiley Online Library.
- Gower, J. C., Gardner Lubbe, S., & Le Roux, N. J. (2011). *Understanding biplots*. John Wiley & Sons.
- Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., & Rossiter, D. (2021). Soilgrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *SOIL*, 7, 217–240. doi:doi:10.5194/soil-7-217-2021.

¹<https://github.com/CSVDW/PCA-Biplots-for-machine-learning-predictions>