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. 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 (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,

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

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.

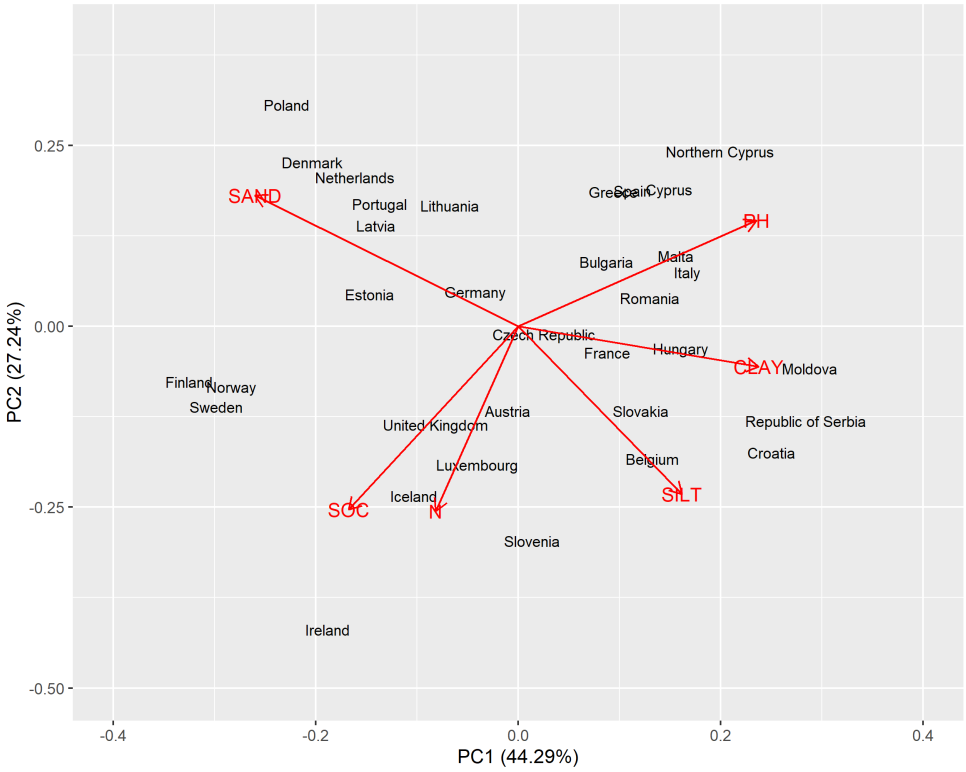

Figure A.1: A Gabriel biplot showing an optimal two-dimensional display of the Land Use and Coverage Area frame Survey data aggregated by country. For each country the median of the particular soil property was obtained.

Appendix B. Modelling results of the random forest model

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).

D	PCOLQ	CC09AVG	PDRYQ	AVGTWQ	LSTN12STD	SOC	CC12AVG	PWETQ	PWARQ	P12TOT	CCYRAVG	EVIYRAVG
	0.13	0.034	-0.035	-0.34	-0.077	-0.17	-0.63	-0.5	-0.51	-0.56	-0.45	-0.48
	1	0.68	0.6	-0.72	0.088	0.11	-0.32	-0.16	-0.43	-0.39	0.089	0.33
		1	0.76	-0.51	0.47	0.43	0.13	0.12	-0.0025	0.027	0.6	0.56
			1	-0.33	0.28	0.29	0.074	-0.036	-0.076	-0.07	0.5	0.42
				1	0.039	-0.13	0.38	0.22	0.42	0.38	0.03	-0.11
					1	0.27	0.32	0.26	0.27	0.27	0.46	0.35
						1	0.42	0.41	0.38	0.4	0.54	0.43
							1	0.84	0.89	0.91	0.83	0.56
								1	0.94	0.94	0.66	0.54
									1	0.98	0.66	0.44
										1	0.67	0.47
											1	0.69
												1

Appendix C. Biplot with Clorpt covariates

The biplot with the covariates based on the climate organisms relief parent material time (clorpt) model is presented in Figure C.2. 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).

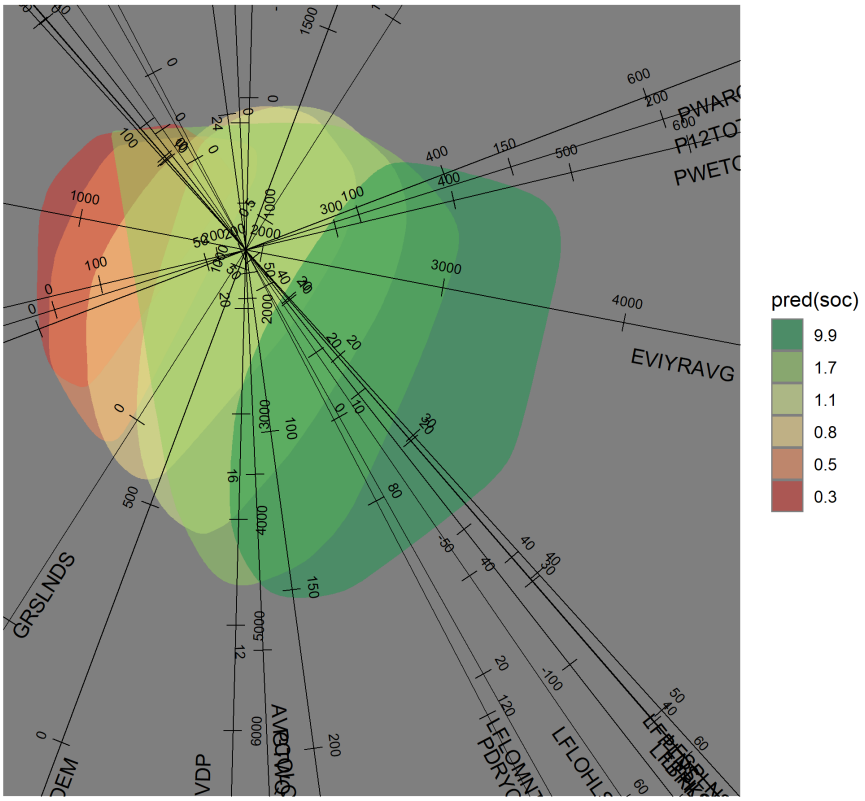

Figure C.2: Biplot with covariates that represent Jenny's clorpt model. The biplot had a quality of 0.456.

In Figures C.3 and C.4 we present biplots for the two highlighted regions based on a subset of covariates that represented Jenny's clorpt model. These included climate factors such as AVGTWQ, PWETQ, and PDRYQ. We also selected covariates that represented the organisms factor like EVIYRAVG, as well as relief factors such as DEM and VDP (Poggio et al., 2021). We did not include covariates that represented parent material and time.

The biplot for the Kahalari region is shown in Figure C.3, while the biplot for the highlighted region in KwaZulu-Natal is shown in Figure C.4. The quality of the first biplot was 0.834, and the quality of the second biplot was 0.726. The predictivity of the axes are shown in the captions of the two figures. The most predictive axis of the Kalahari biplot and the KwaZulu-Natal biplot was that of DEM with a predictivity measure of 0.934 and 0.955, respectively. The least predictive axis of the Kalahari biplot was that of AVGTWQ (0.614) and for the KwaZulu-Natal biplot it was VDP (0.488).

Figure C.3 showed mostly larger predictions for SOC in the upper section of the biplot. The axes of VDP and PDRYQ were mainly responsible for the top-to-bottom variation of the SOC predictions, indicating that valley depth and precipitation in the driest quarter greatly influenced SOC predictions. Specifically, when VDP was larger than 2 000 and when PDRYQ was more than 6.5, most predictions were less than 0.4%. In Figure C.4, VDP and PWETQ were mainly responsible for the high-to-low variation (top-to-bottom) of SOC predictions. When VDP was less than $-1\,000$ and PWETQ more than 420, SOC predictions higher than 2.3% can be expected. It was interesting to note that relief factors such as VDP contributed more notably to the explanation of SOC predictions at the regional scale as opposed to the global scale.

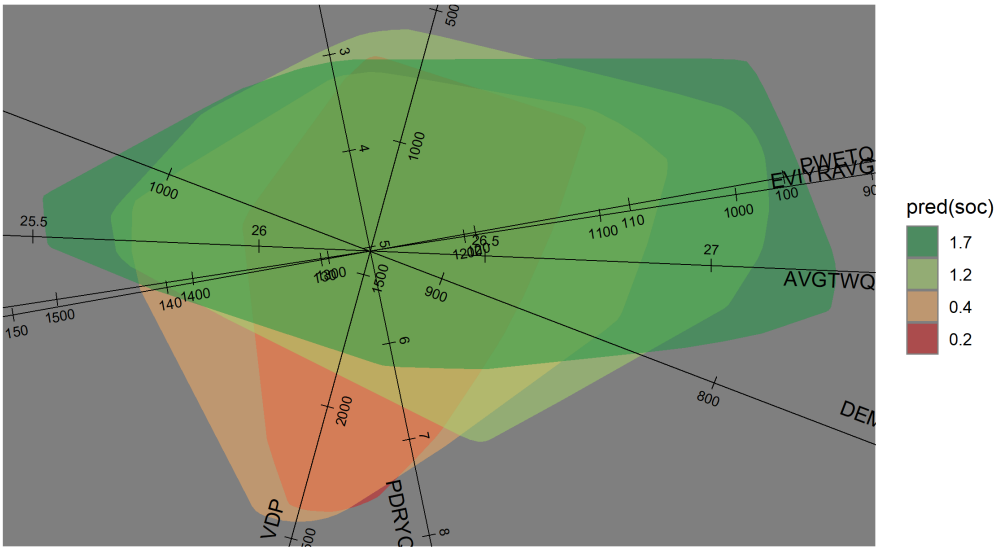

Figure C.3: principal component analysis (PCA) biplot showing the relationship between covariates and the predictions made by the RF model for the Kalahari region. Predictions for SOC are presented in %. The quality of the biplot is 0.834. The respective predicticity measures of the covariates were (sorted from highest to lowest): 0.934 for DEM, 0.926 for PWETQ, 0.904 for VDP, 0.901 for PDRYQ, 0.726 for EVIYRAVG, and 0.614 for AVGTWQ.

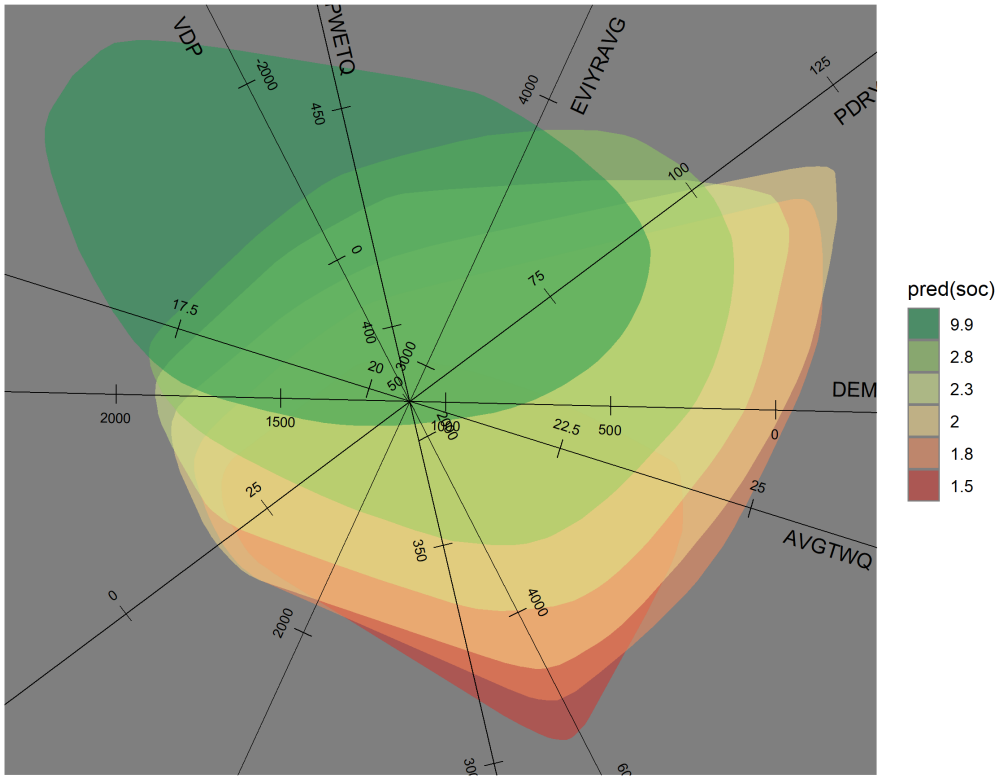

Figure C.4: PCA biplot showing the relationship between covariates and the predictions made by the RF model for the highlighted region in KwaZulu-Natal. Predictions for SOC are presented in %. The quality of the biplot is 0.726. The respective predictivity measures of the covariates were (sorted from highest to lowest): 0.955 for DEM, 0.883 for AVGTWQ, 0.825 for PDRYQ, 0.610 for PWETQ, 0.589 for EVIYRAVG, and 0.488 for VDP.

Appendix D. Important covariates based on Shapley values

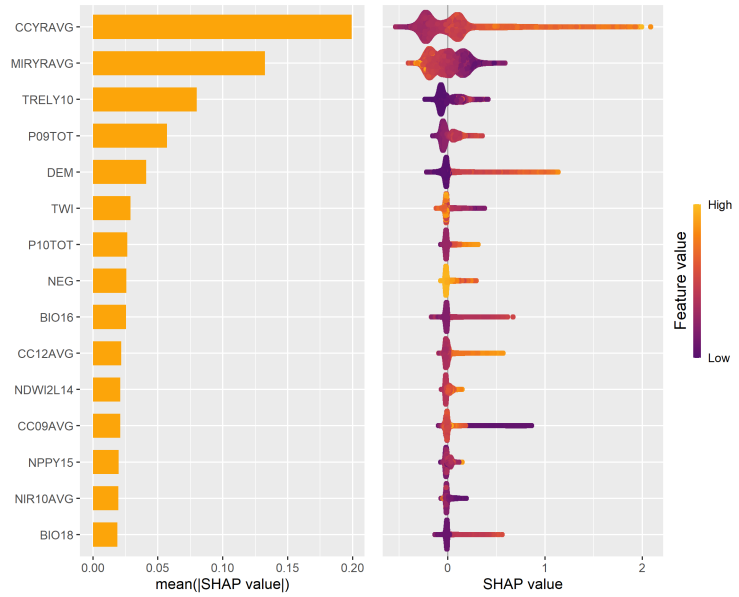

Figure D.5: Most important covariates in the RF model fitted on the South African SOC data based on the average absolute Shapley values. For details concerning the covariates we refer the reader to Poggio et al. (2021).

Appendix E. Other XML results for the highlighted regions

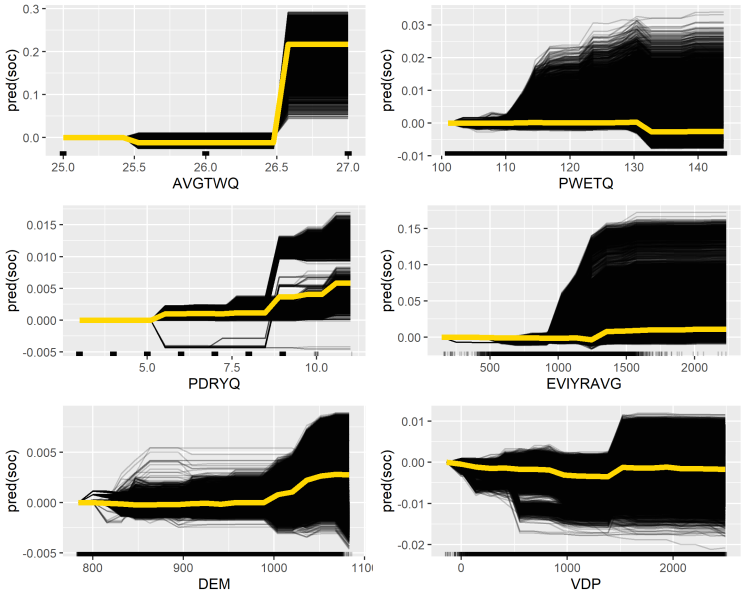

Figure E.6: ICE and PD plots for the Kalahari region.

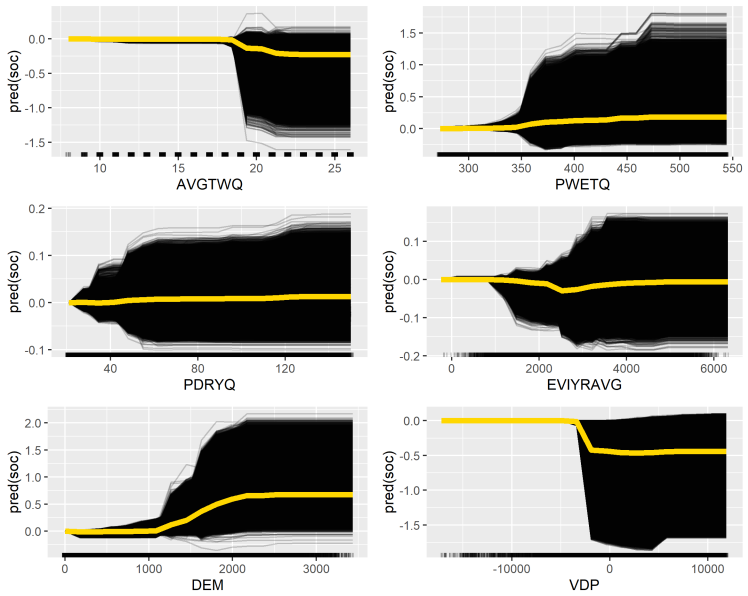

Figure E.7: ICE and PD plots for the region in Kwazulu-Natal.

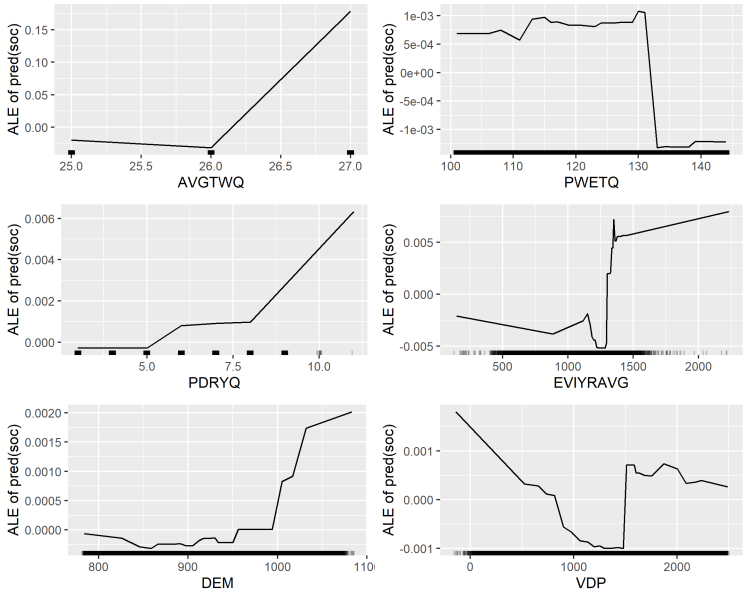

Figure E.8: ALE plots for the Kalahari region.

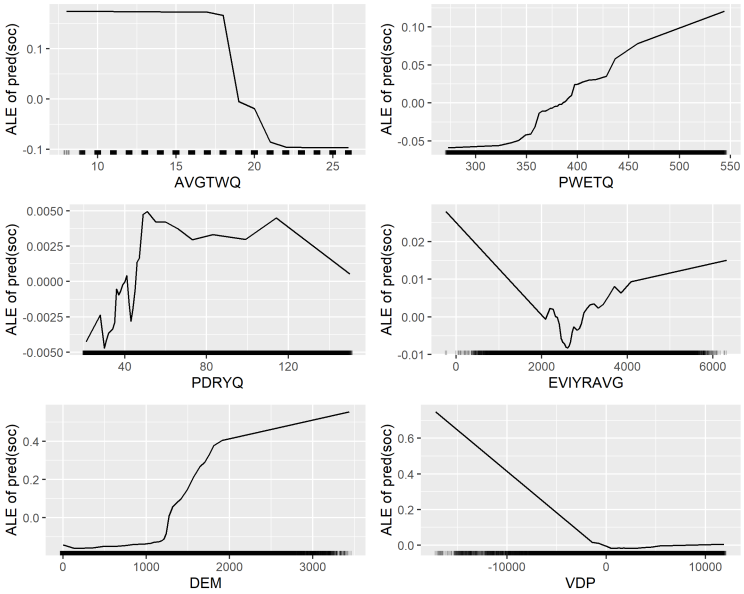

Figure E.9: ALE plots for the region in Kwazulu-Natal.

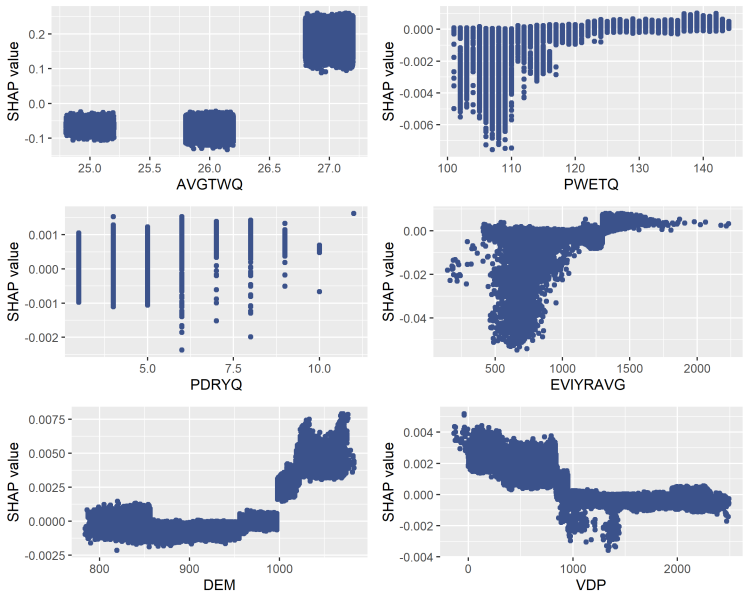

Figure E.10: Partial dependence plots with Shapley values for the Kalahari region.

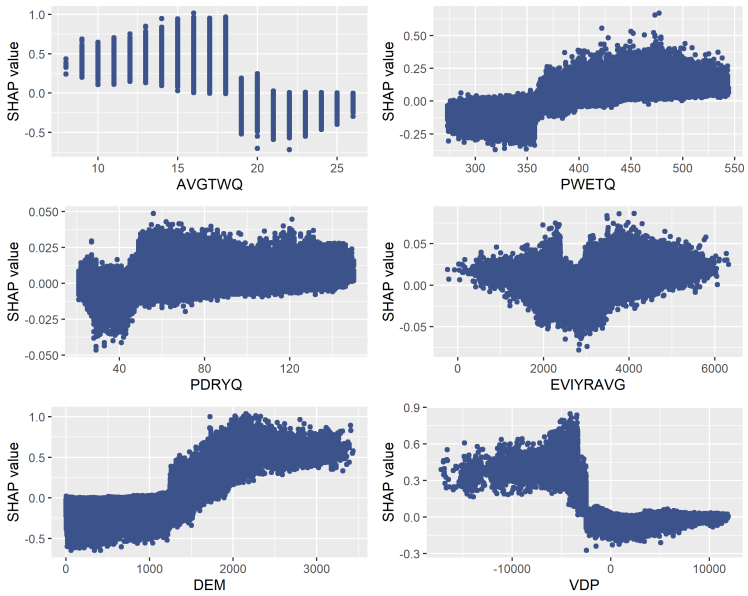

Figure E.11: Partial dependence plots with Shapley values for the region in Kwazulu-Natal.

Appendix F. Biplot R code

Link to github.

References

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.