Eidgenössische Technische Hochschule Zürich

Seminar für Statistik

Comparison of FAC and AAE

Karagiannis-Voules Dimitrios-Alexios
Report as part of "Statistics Lab", ETH Zürich

23rd August 2010

Abstract: This report includes the comparison of two chemical extraction methods in soil. The goal is to study if the cheaper method could efficiently replace the expensive one.

1 Problem Description

Cationexchange capacity and Exchangeable Cations are measured in soil samples by a procedure called "FAC". This procedure costs CHF 300 per sample.

There also exists another method called "AAE-Extraction" (AAE) in which extraction of the nutrient elements is done. This procedure costs CHF 50 per sample.

Consequently, there is a strong argument in favor of the AAE method.

The purpose of this project is to analyze whether it is possible to substitute-predict FAC from AAE. These predictions concern the following chemical elements: Aluminum (Al), Calcium (Ca), Kalium (K), Magnesium (Mg) and also kakeff which shows the cationexchange capacity. Kakeff represents the quantity of cations bound to soil colloids, taking into consideration the strongly pH-value-dependent charge of the organic substances. For the prediction of kakeff all the elements will be used.

2 Data

The initial data provided concerned repeated measurements (Figure 1) in soil samples by both methods. Not all the elements were measured in all soil samples. As a result, different soil samples were used to analyze each element. More sets of measurements from FAC method were included.

Apart from the two measurements there were also some other covariates in the data. These are characterized as "factors of soil formation" and concern:

- climate
- organisms

- geology
- topography

FAC and AAE may be influenced in an undetermined way by these categories.

More analytically:

Name in dataset	Explanation	Type
Al_AAE_PPM	Aluminum from AAE	Continuous
Ca_AAE_PPM	Calcium from AAE	Continuous
K_AAE_PPM	Kalium from AAE	Continuous
Mg_AAE_PPM	Magnesium from AAE	Continuous
alexcorad	Aluminum from FAC	Continuous
caexcorad	Calcium from FAC	Continuous
kexcorad	Kalium from FAC	Continuous
mgexcorad	Magnesium from FAC	Continuous
kakeffexcorad	Effective cation exchange capacity with FAC	Continuous
AktStandorttyp	Landuse	5 level factor
Ausgangsgestein	Parent material	7 level factor
Exposition	Direction	7 level factor
Mittlere Temperatur	Temperature	Continuous
Neigung	Inclination	Continuous
Niederschlagsmenge	Precipitation	Continuous
phkorr	PH	Continuous
Obergrenze	Depth	Transformed to binary
VegCode_EK1972	Vegetation	27 level factor

First of all, negative values in concentrations were coded "NA" as instructed and since they are not plausible values for concentration.

Then the data were aggregated by using the "median" in order to remain with one value per sample. At the end there were different datasets for each element. Median was used because it is a robust (not easily influenced by outliers) estimate of the mean.

The NA removal from the data resulted in having less than all the levels of the categorical variables in the data.

The depth was defined as binary in order to distinguish between topsoil and subsoil. If the upper limit (Obergrenze) of depth was larger than 0 then it was transformed to 1. Otherwise it was transformed to 0 which will be the reference category.

In two soil samples the values of Ca and Mg were corrected since they were found not plausible due to interchanged values within the data. The IDs of these samples are 18259 and 18766. The respective name of this ID in the dataset is "ProbenID".

A figure to each compound is given in section 2.2. It contains the plots for all the variables for a first visual inspection.

Dealing with non-negative data lead to using log-transformation for the two methods. Thus, there will be only sound (positive) values produced from the models.

2.1 Aggregation Explained

Aggregation example for $ProbenID = 7452$:

caexcorad (Ca FAC)	Ca_AAE_PPM (Ca AAE)
8.141663	1601
7.842792	1639
8.151968	1532
7.626368	1521
9.061520	1596
8.522981	1617
7.894321	1515

After the aggregation there is only one value left for each method, which corresponds to the median:

caexcorad (Ca FAC)	Ca_AAE_PPM (Ca AAE)
8.141663	1596

For this particular soil sample there were more values given but the purpose was to demonstrate the aggregation.

Figure 1: A soil sample is divided in subsamples. The result is multiple values per sample.

The following table was calculated from the aggregation of each compound for the two methods. It shows the averaged standard deviation (sd) of the two measurements as well as how many samples and observations per sample the calculation was based on.

The two methods give measurements with different units. The FAC gives units $cmol/kg$ and the AAE mg/kg. Thus, the standard deviations are not comparable. To make them comparable, for the AAE measurements there is only given a standard deviation of the scaled data (scaling by a factor AAE in order to have the same units as FAC). This is noted as "AAEs" in the table which stands for "AAE scaled".

Compound_Method	Averaged sd	Based on measurements x samples
Aluminum_FAC	0.211	42x1, 3x12, 2x75
Aluminum_AAEs	0.126	15x1
Calcium_FAC	0.276	42x1, 3x12, 2x92
Calcium_AAEs	0.221	15x1
Kalium_FAC	0.010	42x1, 3x12, 2x95
Kalium_AAEs	0.007	15x1
Magnesium_FAC	0.049	42x1, 3x11, 2x91
Magnesium_AAEs	0.091	15x1
kakeffcorad	0.452	74x1, 37x1, 12x1, 5x10, 4x7, 3x31, 2x362

There is not a clear lead in the accuracy of the two methods. However, for all AAE measurements there was only one soil sample measured more than once ($ProbenID = 7452$). Consequently, this is just a first impression of the AAE accuracy, and conclusions must be drawn carefully.

The calculated standard deviation of AAE cannot be trusted to do an errors-in-variables (EIV) analysis because it was based on one sample.

In EIV the variation of both methods is used in the modeling process. In this particular situation, for example, it would consider that the AAE observations/measurements are beset with some error.

2.2 Variable Plotting

This section contains the plots of the two methods against all the variables. Furthermore, there are also included the plots between all the other (i.e. not FAC, AAE) variables. For simplicity, whenever referring to log(FAC) or log(AAE), the "log()" argument will not be used. Comments will concern the plots including FAC since this is the target variable.

2.2.1 Aluminum

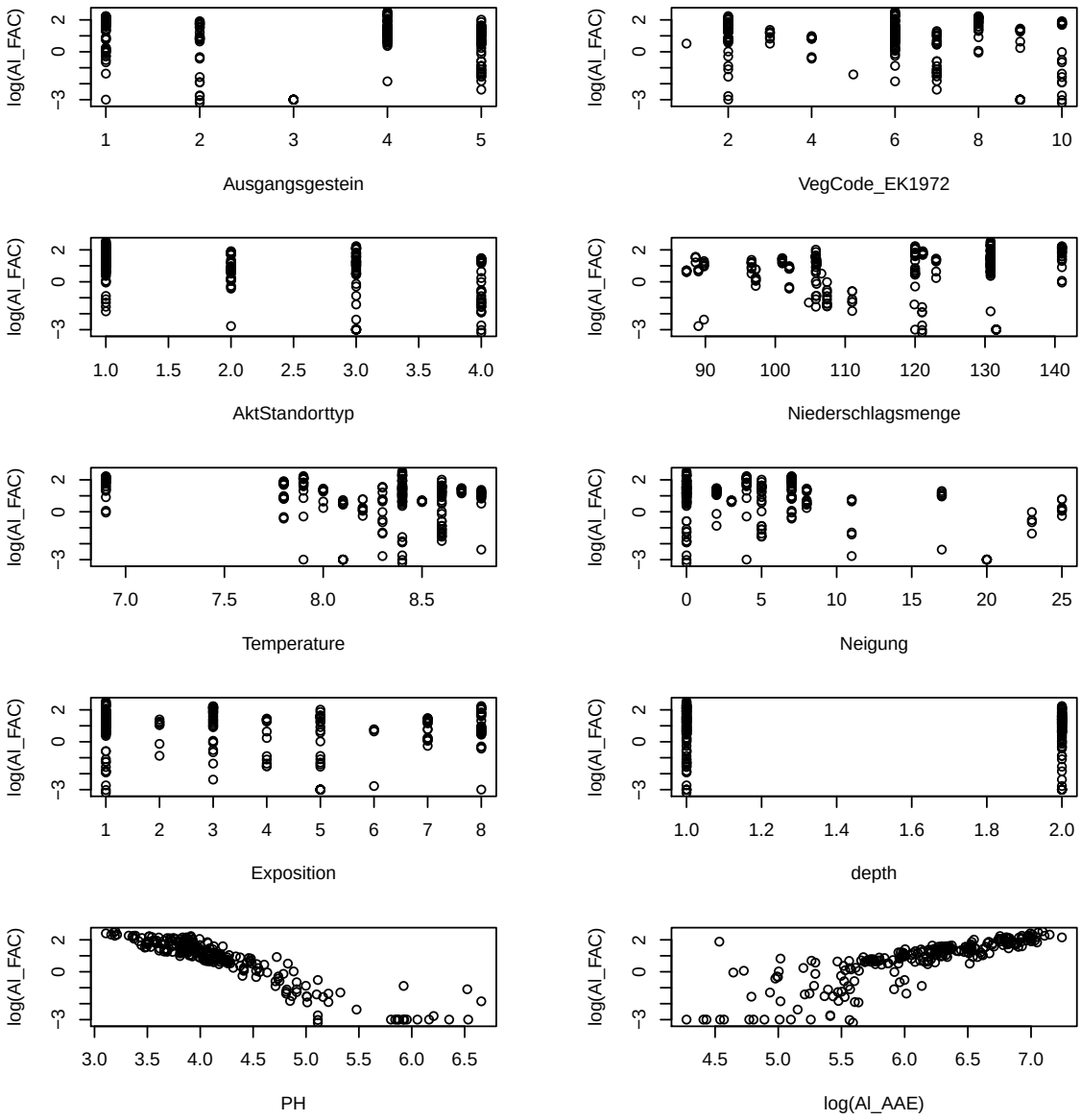

Figure 2: Scatterplot matrix of all the variables for Aluminum FAC.

The third level of Ausgangsgestein produces only low values and level 4 mostly high values of FAC. In Exposition certain levels produce only medium values whereas in the rest all values are produced. Linear dependence of FAC on PH with higher variation for larger PH. The variance shown on plotting FAC versus PH is lower than FAC versus AAE. Large variation of the two methods for small values. The previous two facts support the advice of doing an EIV analysis in the future.

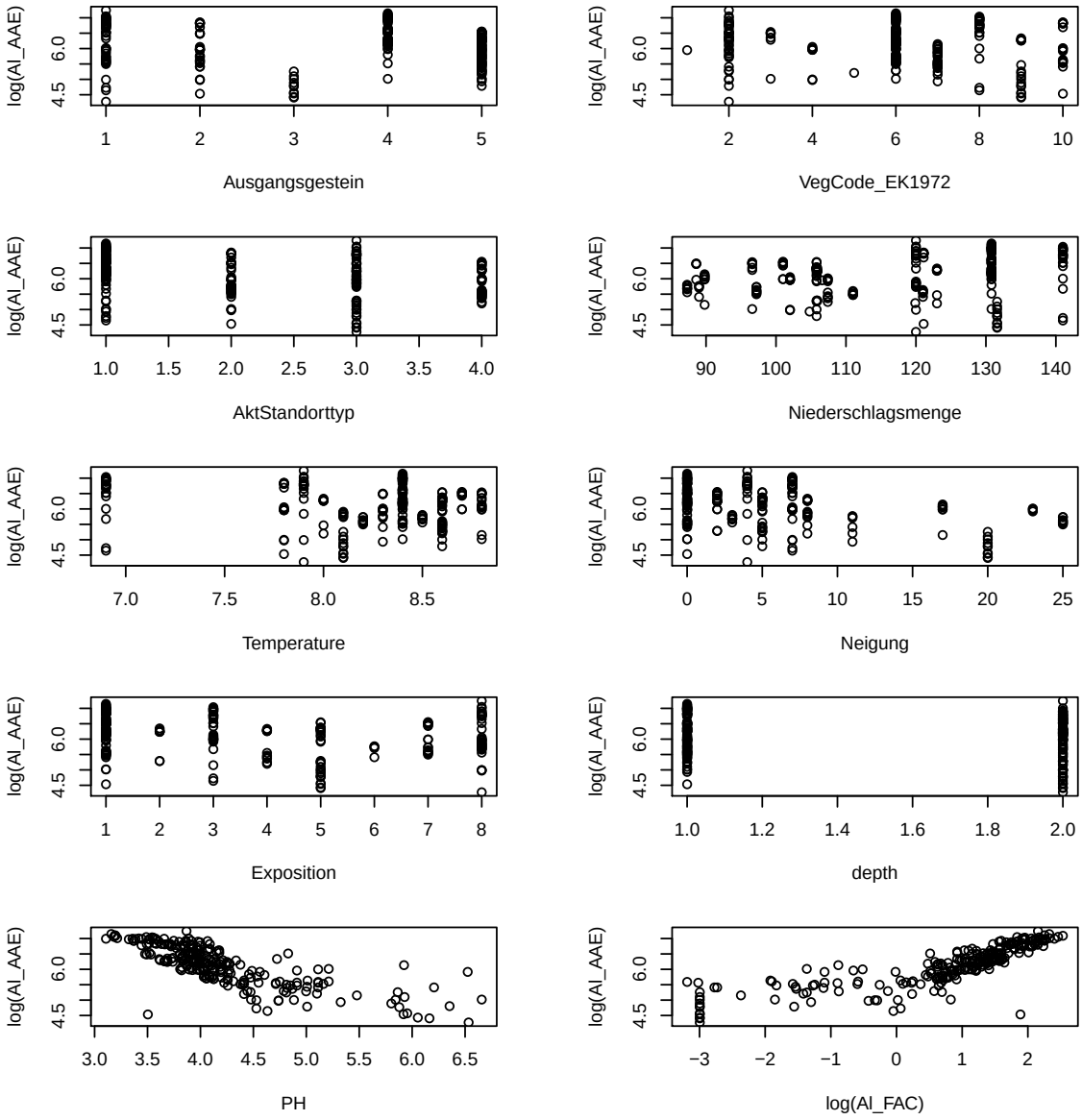

Figure 3: Scatterplot matrix of all the variables for Aluminum AAE.

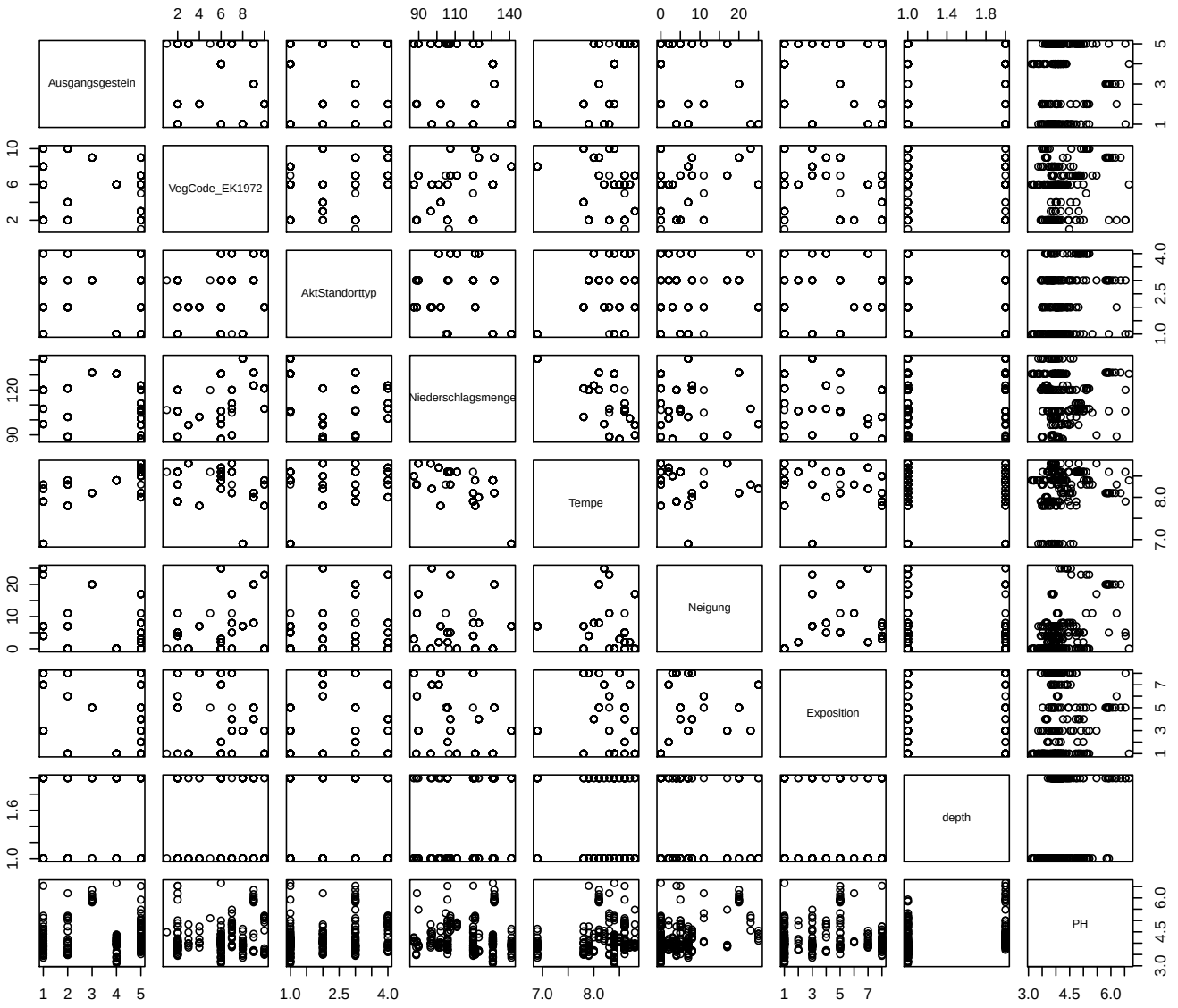

Figure 4: Scatterplot matrix of the rest of the variables for Aluminum.

2.2.2 Calcium

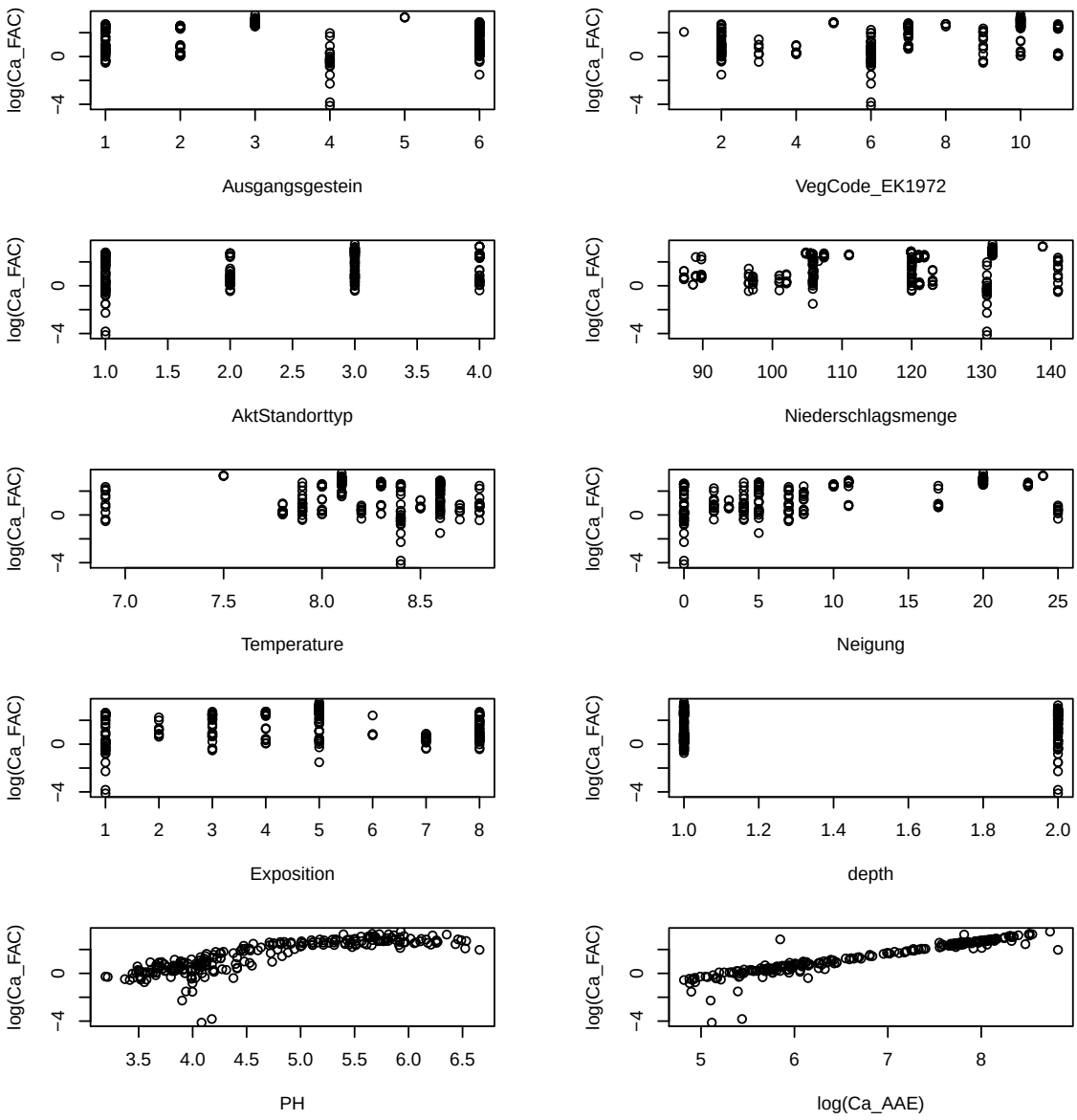

Figure 5: Scatterplot matrix of all the variables for Calcium FAC.

Ausgangsgestein has an effect in Calcium FAC too. In particular, there is one level which produces only high values and the low values produced from FAC correspond to only one level of Ausgangsgestein. There is only one level AktStandorttyp that produces values of all the possible range. For depth equal to 0 there are no low values produced. FAC has a non linear dependence with PH. So, a potential transformation of PH could be beneficial.

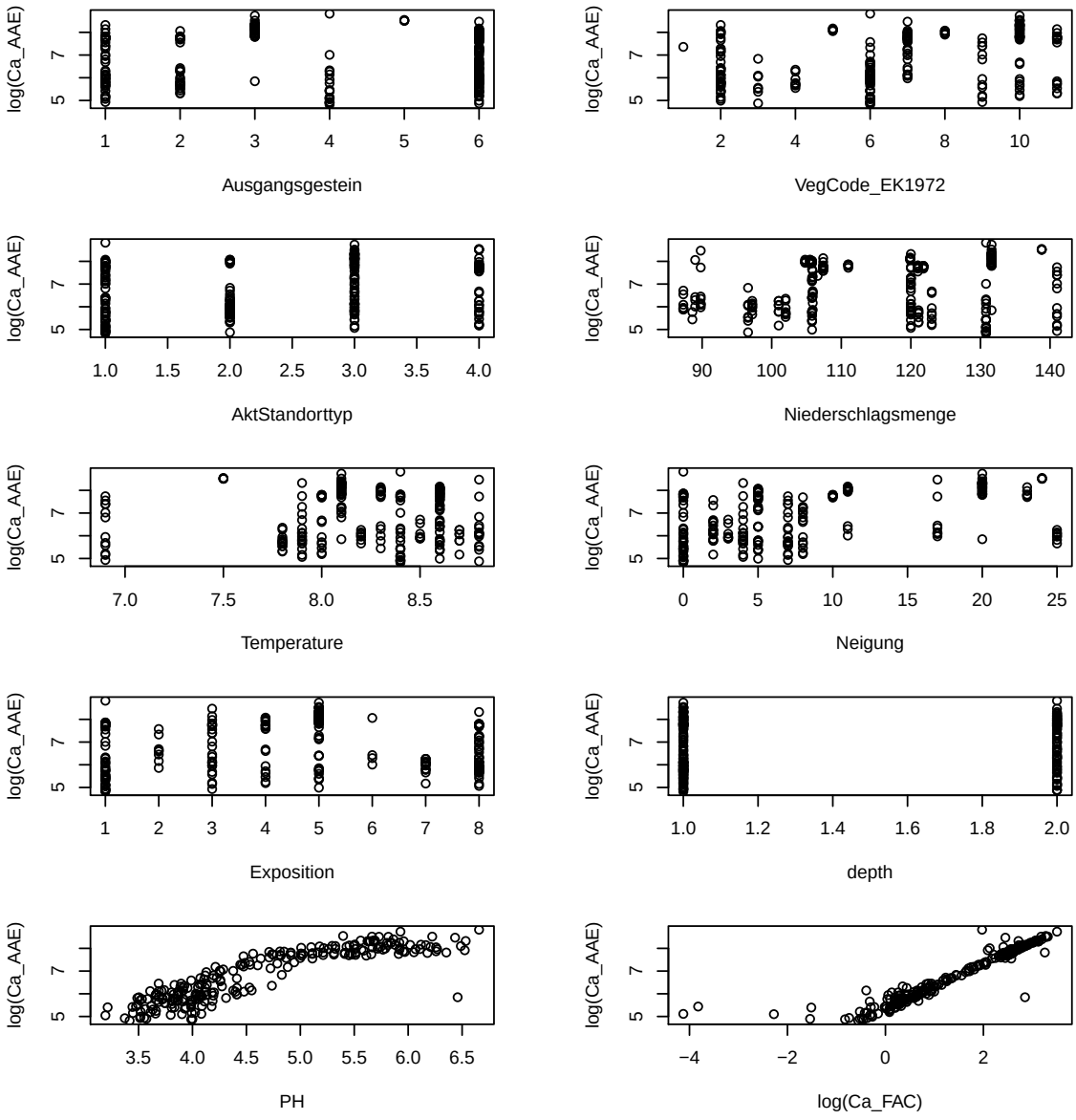

Figure 6: Scatterplot matrix of all the variables for Calcium AAE.

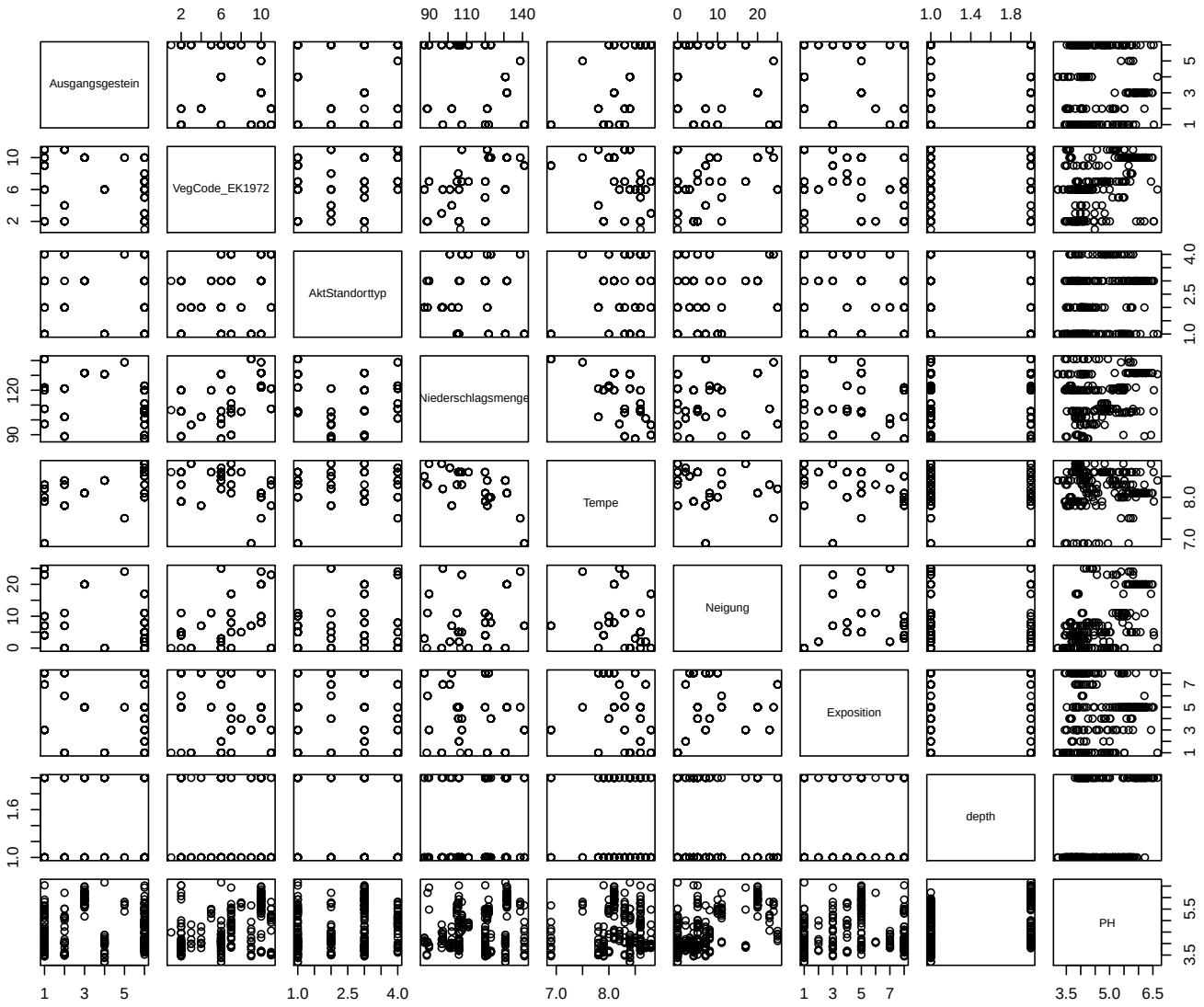

Figure 7: Scatterplot matrix of the rest of the variables for Calcium.

2.2.3 Kalium

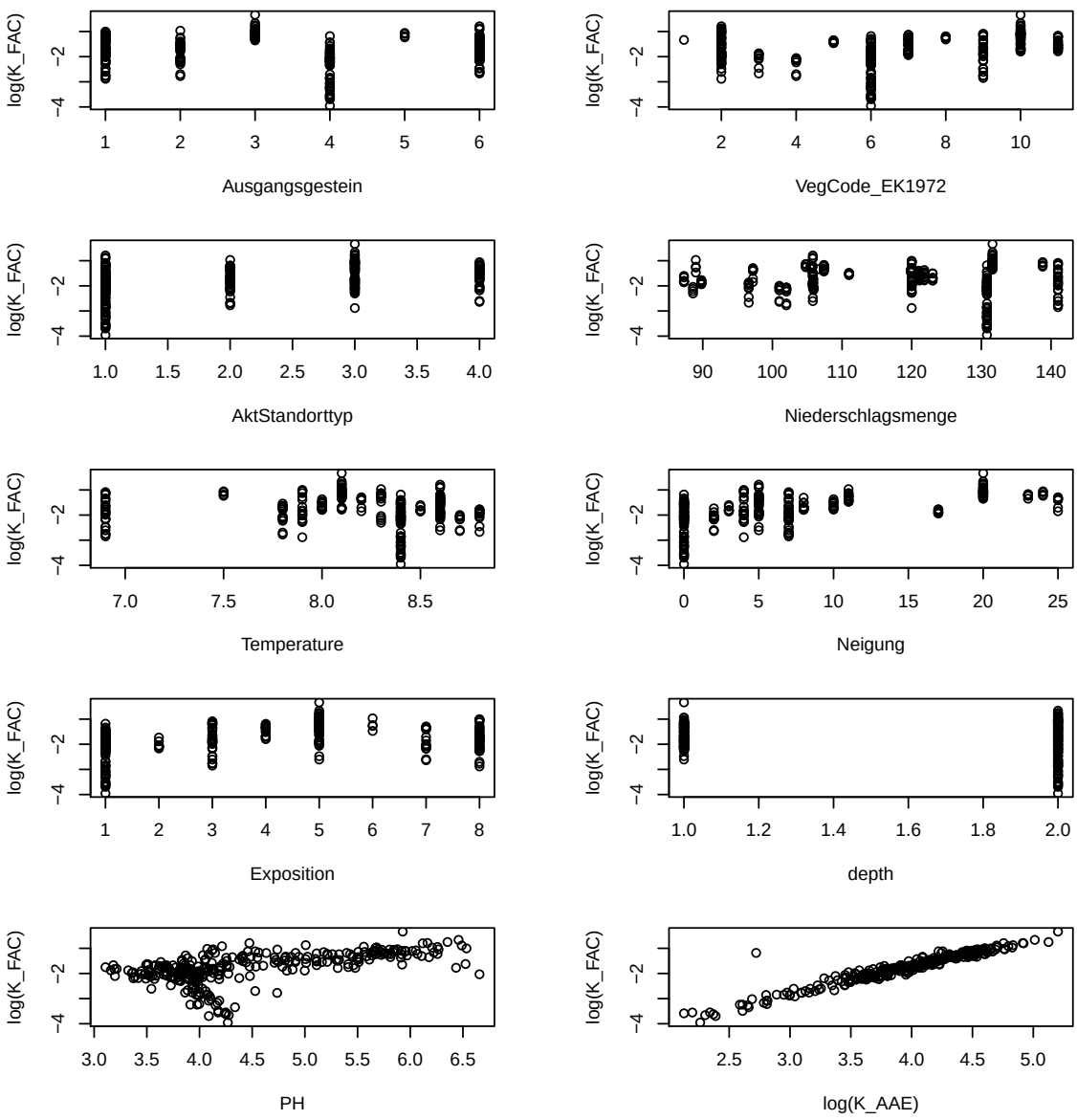

Figure 8: Scatterplot matrix of all the variables for Kalium FAC.

In Ausgangsgestein there are some levels which produce mostly "medium" (not too large and not too low) values and one certain level which produces only large values of FAC. Non linear dependence of FAC on PH. For depth equal to 0 there are no low values produced. Same agreement (no change of variation) of the two methods for all the range (data space) of values.

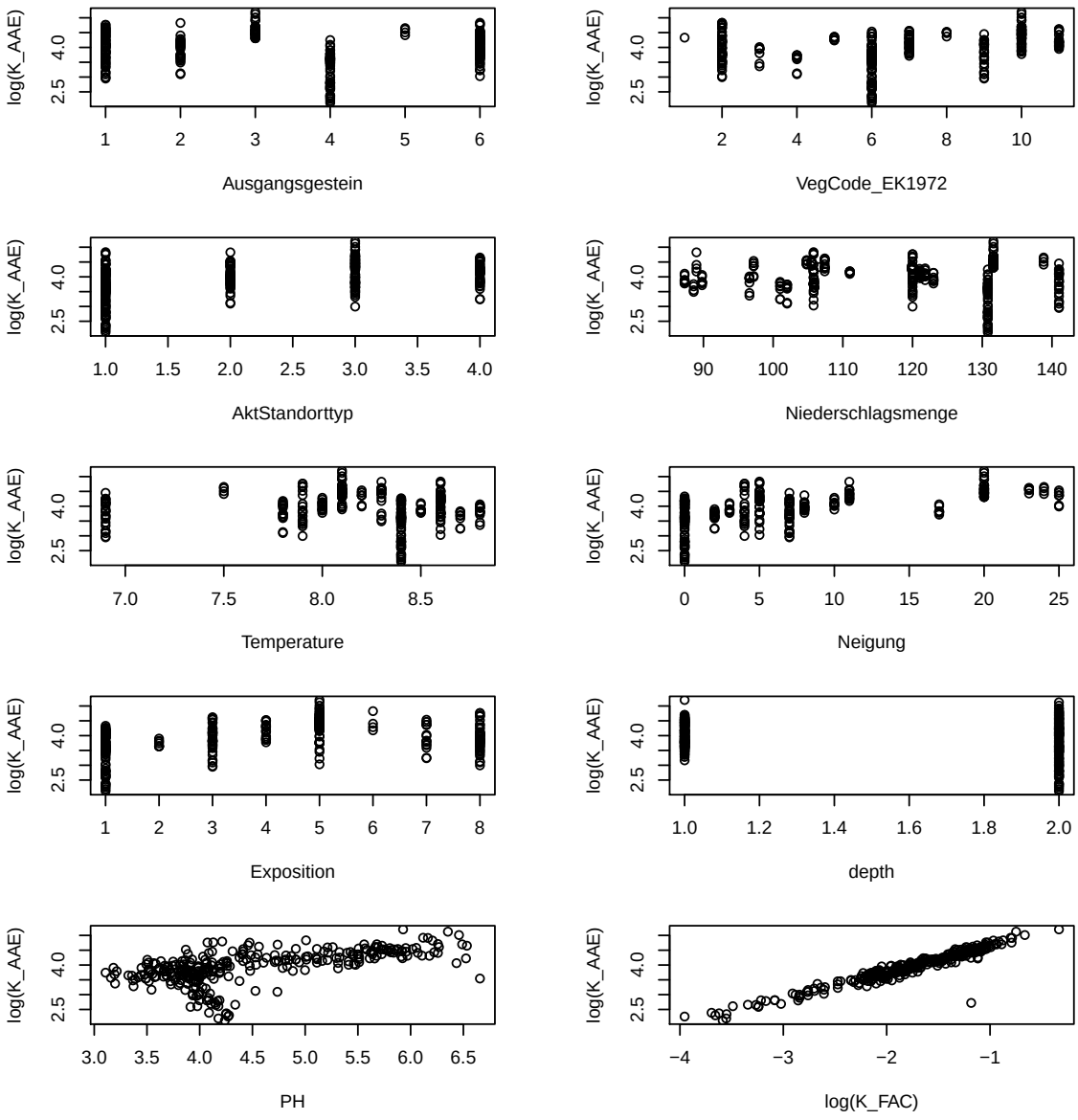

Figure 9: Scatterplot matrix of all the variables for Kalium AAE.

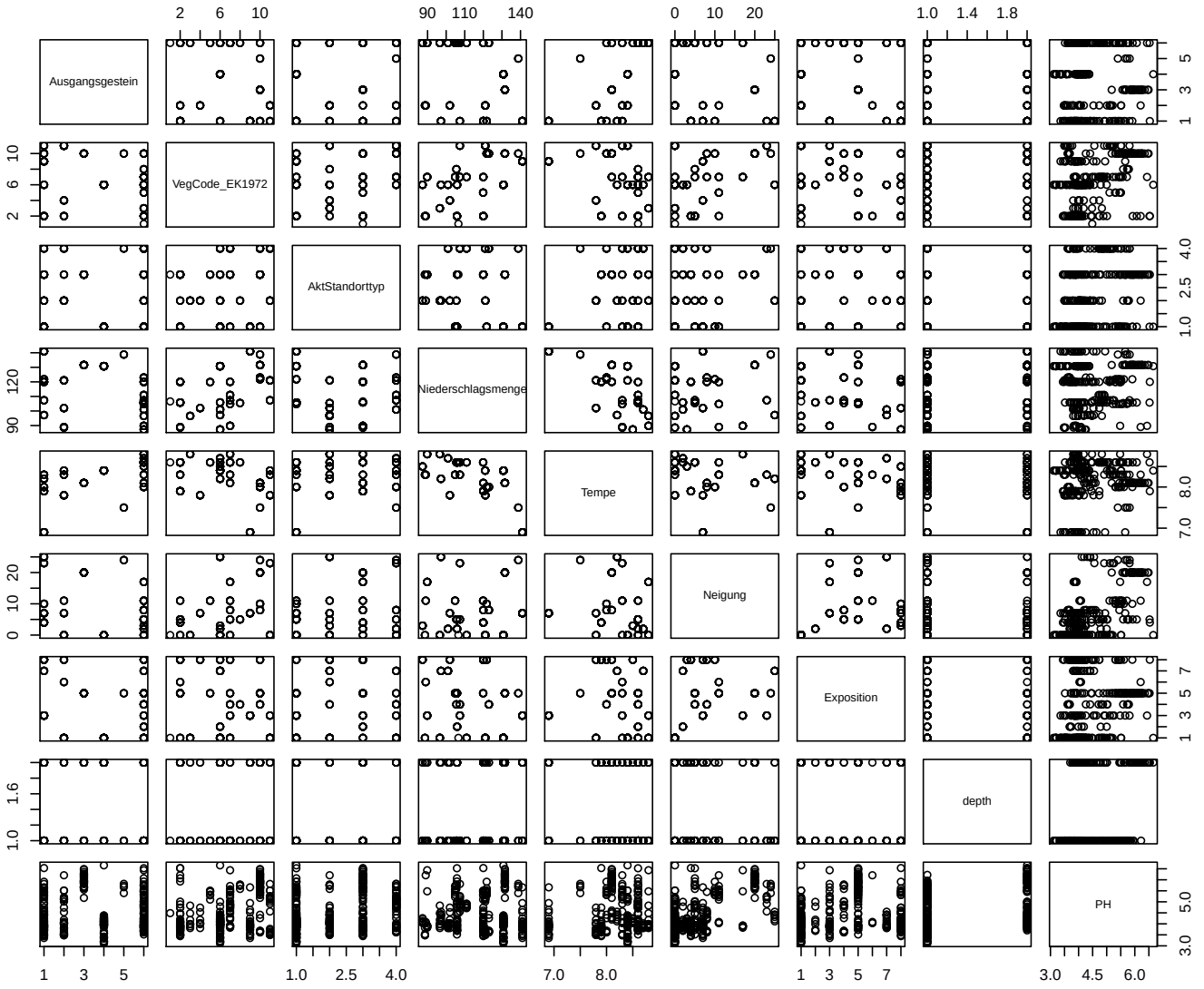

Figure 10: Scatterplot matrix of the rest of the variables for Kalium.

2.2.4 Magnesium

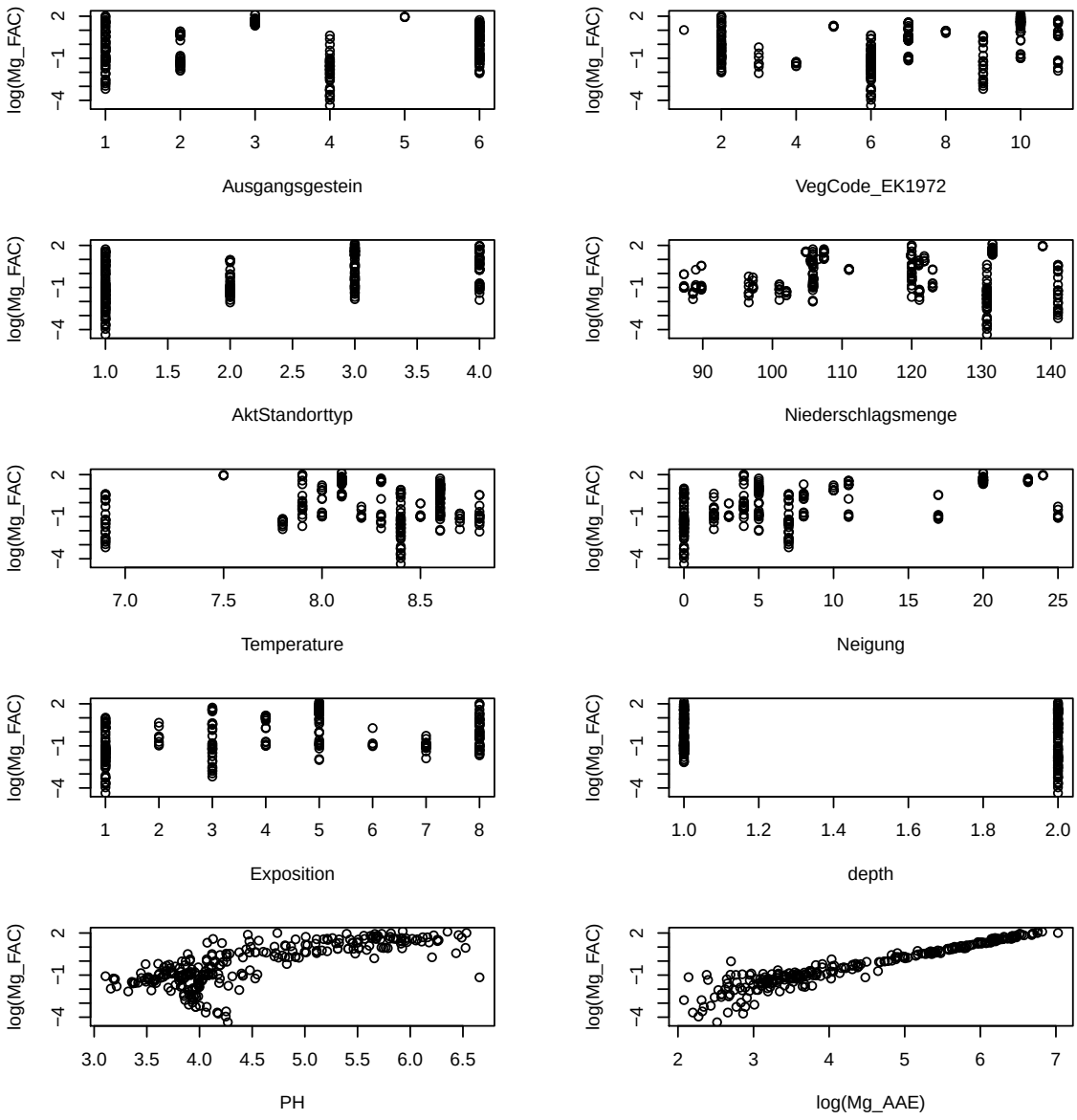

Figure 11: Scatterplot matrix of all the variables for Magnesium FAC.

Ausgangsgestein has some influence in FAC. Some levels produce only high values and there is only one level which produces the very low values. Similar influence of AktStandorttyp where only one level produces low values. For depth equal to 0 there are no low values produced. Non linear dependence of FAC on PH. For lower values there is worse agreement of the two methods than for larger values.

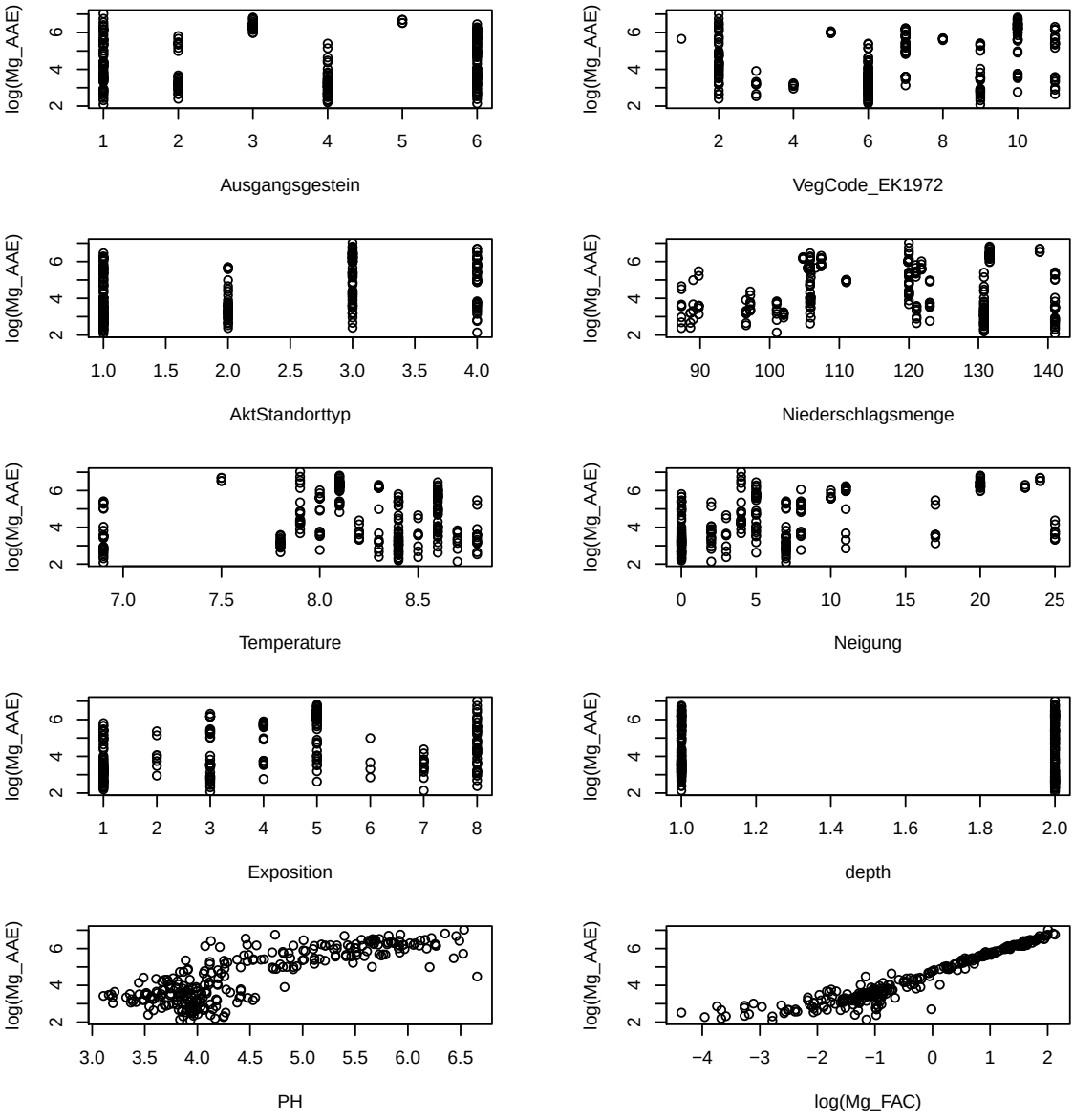

Figure 12: Scatterplot matrix of all the variables for Magnesium AAE.

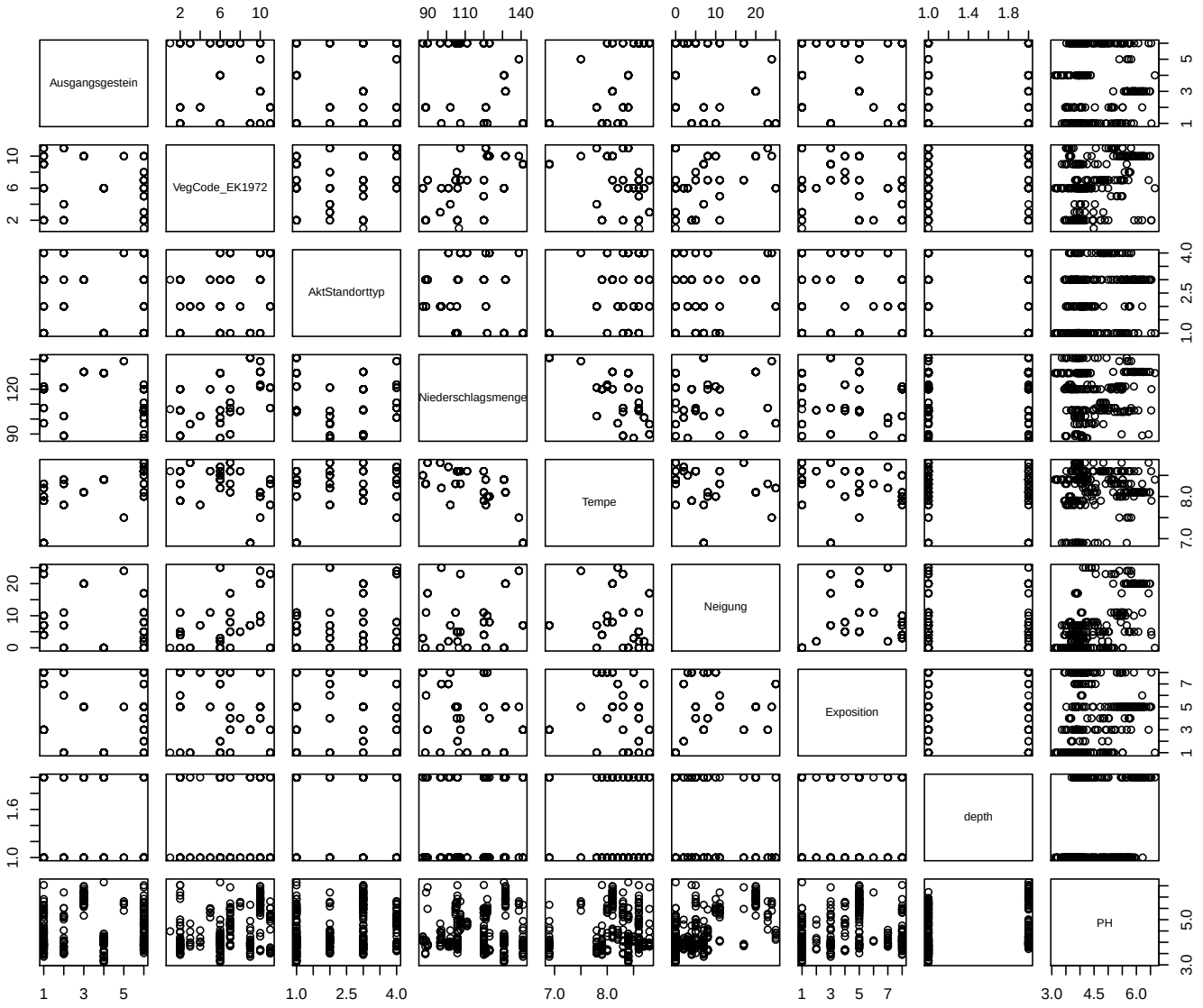

Figure 13: Scatterplot matrix of the rest of the variables for Magnesium.

2.3 Coding

The coding for the factor variables is as follows:

Ausgangsgestein:

- 2, Molasse
- 3, Riss Moraene
- 4, Rutschungen
- 5, Schotter
- 6, Solifluktion
- 7, Tormfoor
- 8, Wuerm-Moraene

AktStandorttyp:

- 1, L-ACKER
- 2, W-ANDERE
- 3, W-LAUB
- 4, W-MISCH
- 5, W-NADEL

Exposition:

- 1, ALL (no inclination, flat landscape)
- 2, E.
- 3, N.
- 4, NE
- 5, NW
- 6, S.
- 7, SE
- 8, SW
- 9, W.

VegCode_ EK1972:

- 8, 29
- 9, 6
- 10, 7(35)
- 11, 7*
- 14, 7a
- 17, 7d
- 19, 7f
- 21, 7g
- 24, 8d
- 26, 8f
- 27, 8g

As mentioned before not all the categories were in the final datasets and this is why the rest of the levels for VegCode_ EK1972 are omitted.

A possible grouping for the levels of Vegetation has been proposed. According to this 7*,7a,7d,7f,7g could be grouped and also 8d,8f,8g.

There will be given one p-value for each factor variable, which corresponds to the "whole" variable and shows its global effect. The levels colored with blue are the reference levels for all the following models (if the respective variable is included in the model).

3 Analysis

3.1 Method

The model selection was performed by :"forward selection", "backward elimination", "stepwise procedure", and looking at the possible model subsets. All the models were compared. AIC, model fit and $adj - R^2$ were some of the criteria used in the comparison.

Robust regression was used for the analysis. The reason for that is to explain better the center of the data. Estimates in robust regression are more reliable in the presence of outliers.

There are two different types of estimates for robust regression: M-estimates and MM-estimates. MM-estimates are more robust than M-estimates.

Technical (concerning the program R) details during the analysis lead in certain occasions to the usage of M-estimates.

The future usage of the models lead to finding "prediction optimal" models.

For each model there are more details in the Appendix (Section 5).

In the plots of the two methods in this section for Aluminum and Magnesium the red line corresponds to a simple linear model with only AAE as an explanatory variable. In the plots of Calcium and Kalium the red line corresponds to the reported model for the reference categories of the categorical variables.

The overall performance of each model can be visualized by the model fit where the observed values are plotted against the fitted ones from the model and "ideally" they should lie on the red line which corresponds to the identity function. These plots are in the Appendix (section 5).

The reported R^2 for each model is the adjusted R^2 (notation: $adj - R^2$) which adjusts for the number of parameters in the model. It increases from the addition of a new variable only if there is a reduction in the residual mean square which is the estimated variance of the residuals.

3.2 Aluminum

Analysis for Aluminum. Robust MM-estimation of the coefficients. There exists a large variance in both methods for Aluminum. This is an additional reason for the inclusion of more covariates in this model compared to the other 3 that concern Calcium, Kalium, Magnesium. A recommendation for the future is to measure multiple times the soil samples with both methods so that an errors-in-variables analysis will be feasible.

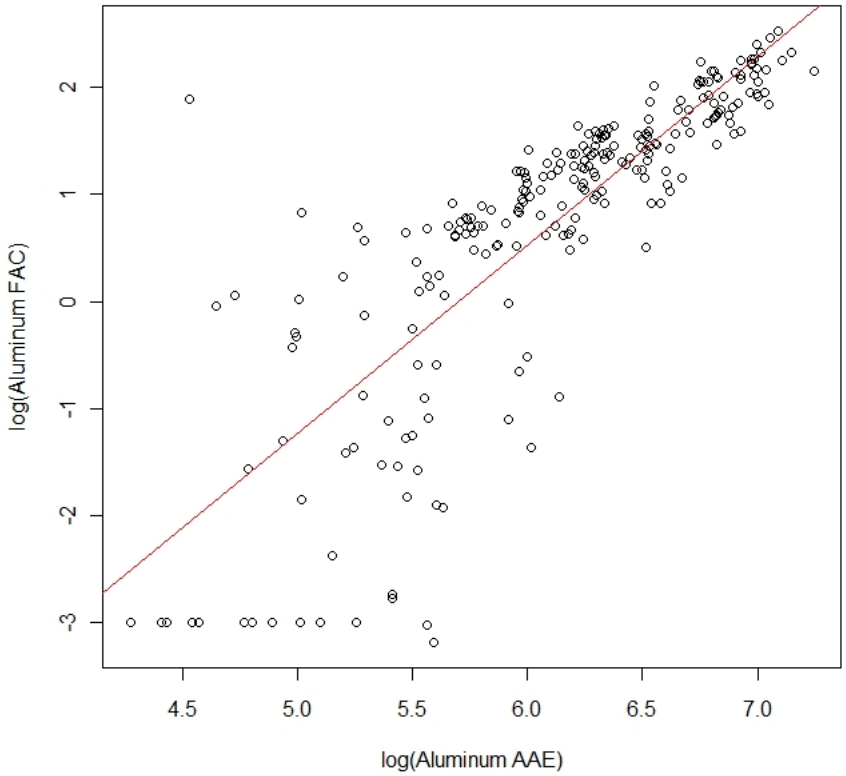

Figure 14: Plot of the two methods for Aluminum. For larger values there is better agreement.

Formula of the model for Aluminum:

$$log(Al_FAC) \quad \sim \quad PH + log(Al_AAE) + Ausgangsgestein + Tempe + Exposition + depth$$

Variable	Value	Std. Error	P-value
Intercept	5.03985	0.73162	0.00000
PH	-1.45337	0.05541	0.00000
Log Al_AAE	0.53789	0.05194	0.00000
Ausgangsgestein 3	-0.15929	0.00825	0.00000
Ausgangsgestein 4	-0.83257	0.12651	
Ausgangsgestein 5	0.09375	0.09226	
Ausgangsgestein 8	0.15210	0.07796	
Temperature	-0.21664	0.06523	0.00104
Exposition 2	0.13871	0.10469	0.00000
Exposition 3	0.09213	0.09299	
Exposition 4	-0.25317	0.09778	
Exposition 5	0.41333	0.08151	
Exposition 6	0.31056	0.13439	
Exposition 8	0.20221	0.09039	
Exposition 9	0.27824	0.07074	
depth 1	0.36197	0.03937	0.00000

As expected from figure 2, Ausgangsgestein is in the model. Exposition is also in the model and because in figure 2 there were not all the levels that showed a difference, these now have a large standard error. Depth turns out significant. This could result from the variation of the two methods which is large. Obviously PH adds essential information in the model. This was expected from figure 2 where the linear dependence on PH is visible. The higher the PH the lower the log(Al_ FAC). The effect of PH on FAC is larger than the effect of AAE.

One should remember that the models concern the log transformation and thus for the "original" model there should be a back-transformation. For instance:

$$Al_FAC \sim \exp\left(log(Al_AAE) + PH + Ausgangsgestein + Temperature + Exposition + depth\right)$$

Thus:

$$Al_FAC \sim (Al_AAE)\cdot\exp\left(PH\right)\cdot\exp\left(Ausgangsgestein\right)\cdot\exp\left(Temperature\right)\cdot\exp\left(Exposition\right)\cdot\exp\left(depth\right)$$

So, the lower (in absolute value) the estimate of a variable, the closer its exponent to 1 and the smaller its effect.

Next, is reported the corresponding $adj - R^2$ of the model as well as the one of simpler models to see the additional contribution.

Model	$adj - R^2$
Reported	0.9711
Only including log(Aluminum AAE)	0.7493
Adding only PH	0.9224

Apart from the PH and AAE the rest of the variables have a small contribution in explaining FAC.

3.3 Calcium

Analysis for Calcium. Robust M-estimation of the coefficients.

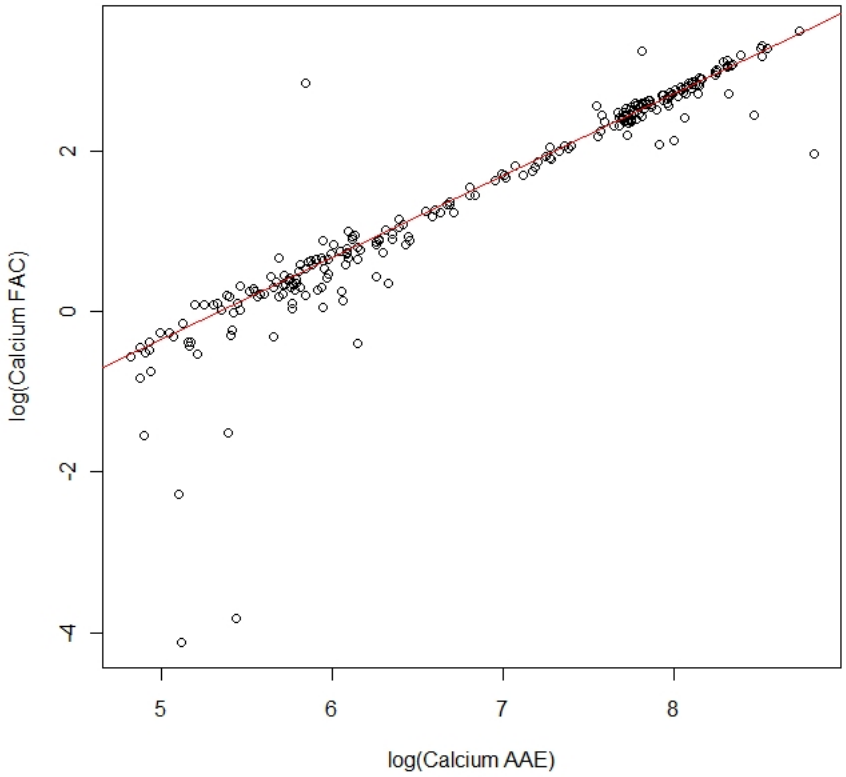

Figure 15: Plot of the two methods for Calcium. Larger values agree better.

Formula of the model for Calcium:

$$log(Ca_FAC) \quad \sim \quad log(Ca_AAE) + Ausgangsgestein + depth$$

Variable	Value	Std. Error	P-value
Intercept	-5.42399	0.058	0.00000
Log Ca_AAE	1.01619	0.008	0.00000
depth 1	-0.07499	0.0174	0.00079
Ausgangsgestein 3	-0.04421	0.0280	0.00001
Ausgangsgestein 4	0.07741	0.0291	
Ausgangsgestein 5	-0.24071	0.0323	
Ausgangsgestein 6	0.05230	0.0643	
Ausgangsgestein 8	0.00491	0.0203	

Model	$adj - R^2$
Reported	0.9831
Only including log(Calcium AAE)	0.9801

Small contribution of Ausgangsgestein and depth. Furthermore their magnitude is low (exponent close to 1). It was obvious from the plot of the two methods that they are in a good agreement and AAE would explain substantially FAC.

3.4 Kalium

Analysis for Kalium. Robust MM-estimation of the coefficients.

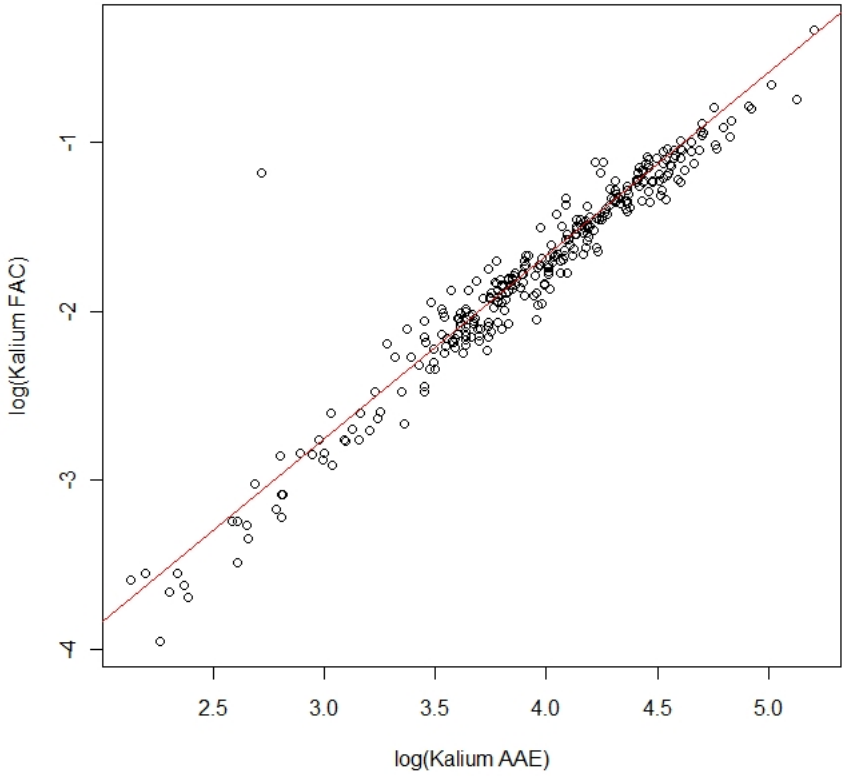

Figure 16: Plot of the two methods for Kalium.

Formula of the model for Kalium:

$$log(K_FAC) \quad \sim \quad log(K_AAE) + AktStandorttyp + depth$$

Variable	Value	Std. Error	P-value
Intercept	-6.00886	0.05623	0.00000
Log K_AAE	1.08414	0.01464	0.00000
AktStandorttyp 3	-0.08928	0.02198	0.00004
AktStandorttyp 4	-0.01539	0.01891	
AktStandorttyp 5	-0.04669	0.02306	
depth 1	-0.03584	0.01530	0.01981

The level of magnitude of depth shows that it adds little information (estimate close to 0 and thus the exponent close to 1). The same stands for AktStandorttyp in whose levels there are very small differences.

Model	$adj - R^2$
Reported	0.9735
Only including log(Kalium AAE)	0.971

Apart from AAE the rest of the covariates have a small contribution. The two methods were in a good agreement and AAE explains substantially FAC.

3.5 Magnesium

Analysis for Magnesium. Robust MM-estimation of the coefficients.

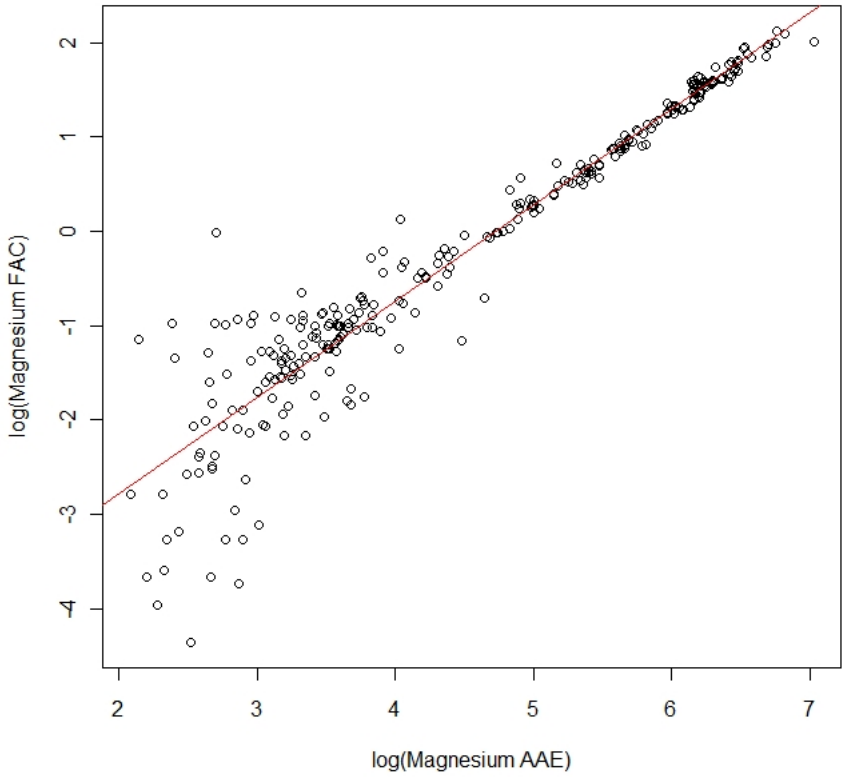

Figure 17: Plot of the two methods for Magnesium. For larger values there is better agreement.

Formula of the model for Magnesium:

$$log(Mg_FAC) \quad \sim \quad log(Mg_AAE) + Ausgangsgestein + depth + Niederschlagsmenge$$

Variable	Value	Std. Error	P-value
Intercept	-4.05396	0.17691	0.00000
Log Mg_AAE	0.98388	0.01491	0.00000
Ausgangsgestein 3	0.02966	0.05462	0.00000
Ausgangsgestein 4	0.13456	0.05298	
Ausgangsgestein 5	-0.09384	0.05581	
Ausgangsgestein 6	0.14256	0.13201	
Ausgangsgestein 8	0.00289	0.04256	
depth 1	-0.13489	0.02998	0.00001
Niederschlagsmenge	-0.00460	0.00138	0.00097

AktStandorttyp is not in the model as expected from Figure 11 and this is explained in the Appendix. Niederschlagsmenge's coefficient is not large but the values in which it ranges in the dataset are between 87.3 and 141. Depth is included in the model as expected from Figure 11 in which it was noticeable that for depth equal to 0 there were not low values but only medium and high ones. Therefore since depth 1 produces low values its estimate is negative.

Model	$adj - R^2$
Reported	0.9886
Only including log(Magnesium AAE)	0.9812

Small contribution of the rest of the covariates. The two methods were in a good agreement and AAE explains substantially FAC.

3.6 kakeff

Analysis for kakeffcorad. For kakeffcorad there is no direct AAE measurement and as a result the final model is complex. Log transformation of kakeffcorad so that the model will not produce implausible values. Furthermore, second order terms were included for log(Calcium AAE) and log(Magnesium AAE). Binary PH was selected for this analysis. An interaction between PH(binary) and log(Aluminum AAE) was added. Interaction means that two variables combined have an effect on an another variable. The reason for the transformation and the inclusion of the interaction is in the Appendix. Robust M-estimation of the coefficients.

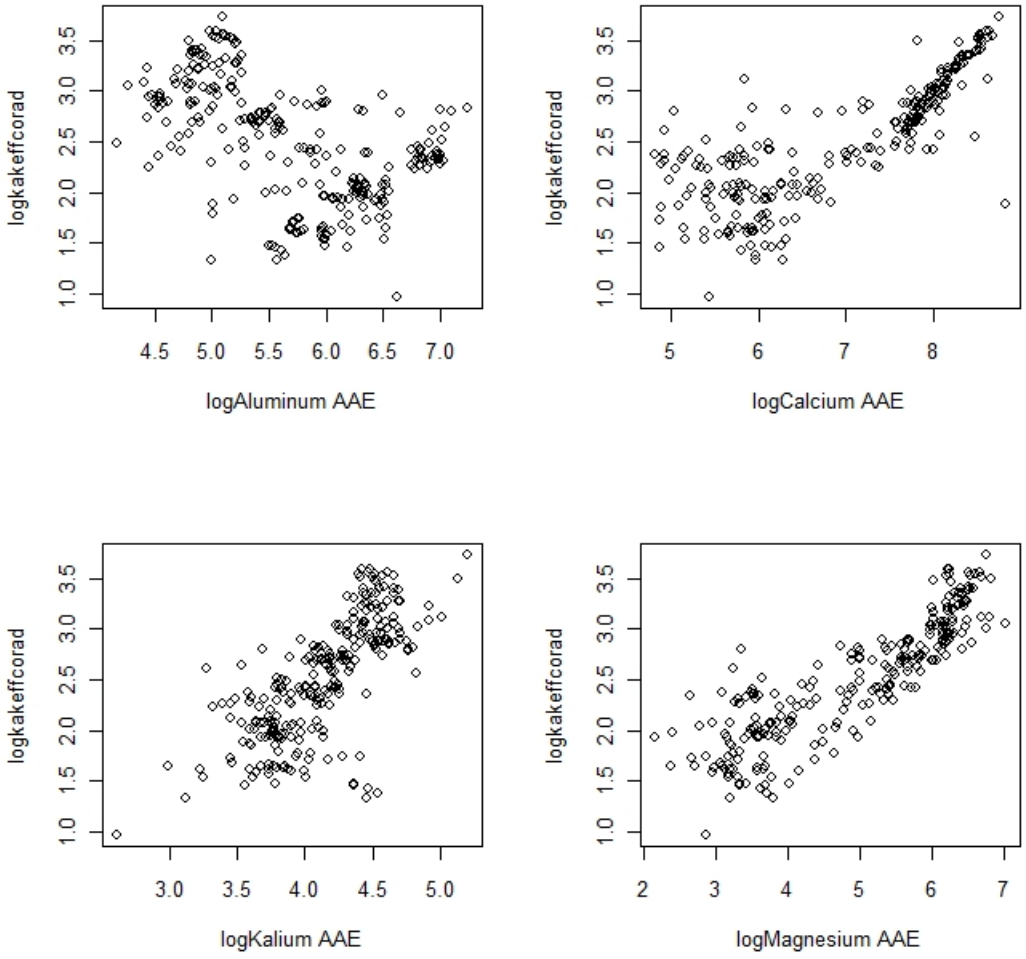

Figure 18: Plot of kakeff against all other elements. Topleft kakeff with Aluminum AAE, topright kakeff with Calcium, bottomleft kakeff with Kalium AAE, bottomright kakeff with Magnesium.

Formula of the model for kakeff:

$$\begin{aligned} log(kakeff) &\sim (log(Mg_AAE))^2 + ph + depth + Exposition + VegCode_EK1972 \\ &+ AktStandorttyp + Ausgangsgestein + Neigung + log(Al_AAE) \\ &+ log(Mg_AAE) + (log(Ca_AAE))^2 + log(Ca_AAE) + ph : log(Al_AAE) \end{aligned}$$

Variable	Value	Std. Error	P-value
Intercept	-0.68530	0.5303	0.1976
(logCaAAE)^2	0.02438	0.0097	0.0190
(logMgAAE)^2	0.07215	0.0066	0.0000
logCaAAE	-0.22191	0.1335	0.1137
logMgAAE	-0.35098	0.0625	0.0000
logAlAAE	0.54003	0.0334	0.0000
ph(>4):logAl	-0.46202	0.0377	0.0000
ph(>4)	2.75529	0.2413	0.0000
depth 1	-0.11161	0.0198	0.0000
Neigung	-0.01055	0.0047	0.0334
Ausgangsgestein 3	0.030210	0.0918	0.0047
Ausgangsgestein 4	-0.10241	0.0981	
Ausgangsgestein 5	0.20778	0.1305	
Ausgangsgestein 6	-0.22073	0.0755	
Ausgangsgestein 8	-0.07667	0.0611	
VegCode_EK1972 9	-0.00804	0.1230	0.0000
VegCode_EK1972 10	-0.23096	0.1161	
VegCode_EK1972 11	-0.27551	0.2027	
VegCode_EK1972 14	0.13997	0.1150	
VegCode_EK1972 19	-0.14894	0.1524	
VegCode_EK1972 17	0.19817	0.1212	
VegCode_EK1972 21	0.49034	0.1260	
VegCode_EK1972 24	0.15746	0.1502	
VegCode_EK1972 26	0.27519	0.1473	
VegCode_EK1972 27	-0.12136	0.1502	
AktStandorttyp 3	0.27732	0.0605	0.0000
AktStandorttyp 4	0.11365	0.0555	
AktStandorttyp 5	0.33236	0.0428	
Exposition 2	0.28796	0.1115	0.0000
Exposition 3	0.07120	0.0798	
Exposition 4	-0.32799	0.0633	
Exposition 5	0.21658	0.0795	
Exposition 6	-0.12955	0.1027	
Exposition 8	-0.11297	0.0885	
Exposition 9	-0.03971	0.0686	

Small coefficients for Exposition and the second order terms. This means that when the model is back-transformed (as shown in Section 3.2) the exponent of the coefficients will be close to 1 and thus small influence of the corresponding variables. In Ausgangsgestein only levels 5 and 6 have a considerable effect. In vegetation only levels 10 and 21 have small standard error. For AktStandorttyp all the levels have small standard error and the alteration of the coefficients shows that there is difference in the levels. The interaction term and the binary PH have an important contribution in the model. It is interesting that the larger coefficient is the one of PH ($ph > 4$).

Model	$adj - R^2$
Reported	0.9737
Only including Mg, Ca, Al, PH, and the interaction between PH and Al	0.9002

A much simpler model could explain kakeffcorad well enough. By adding all the factors and the higher order terms the $adj - R^2$ increased by only 0.07. This could be a hint for a future EIV analysis where more measurements for AAE per sample will be available. Then the resulting model could be simpler than the reported one with a comparable $adj - R^2$.

4 Usage

4.1 Prediction

A general formula for prediction interval is given. However, due to the large matrices, any model specific details shall be provided electronically.

Point prediction can be performed by the tables of the models, but even that is recommended to be done electronically.

Suppose we have measured all the predictors and we want to give a prediction interval for the response variable of a certain model.

Let X_f denote the future vector of the predictors of one observation and $\hat{Y_f}$ the corresponding predicted response. Then the prediction interval for $\hat{Y_f}$ is:

$$\hat{Y_f} \pm t(1 - \frac{a}{2}; n - p) \cdot s_f$$

where:

$$s_f = MSE \cdot (1 + X_f^T (X^T X)^{-1} X_f)$$

Explanations: n is the number of observations, p is the number of predictors in the model (including the intercept), (1-a) is the confidence level, t(b;k) is the b quantile of the t-distribution with k degrees of freedom and MSE is the estimated variance of the residuals. s_f shows the deviation of the predicted value $\hat{Y_f}$.

In all the models, log-transformation (with base e) was applied. This means that both the point prediction and the prediction intervals will have to be back-transformed.

A simple example for the model of Calcium: Let Ca_AAE=403.4288, log(Ca_AAE)=6, Ausgangsgestein 3 (Riss Moraene) and depth 1. Then the point prediction would be equal to 0.55375 and the prediction interval $[0.36346, 0.74444]$ but these predict the log(Ca_FAC). Taking the exponent of these values we would end up with point prediction 1.73977 and prediction interval $[1.43829, 2.10526]$ and these concern Ca_FAC (caexcorad). In this example n=244, p=8, a=0.05.

The prediction interval shall be interpreted as: there is 0.95 (or $1 - a$) probability that the interval contains the FAC measurement.

For more details in prediction and regression this book is recommended: "Introduction to Linear Regression Analysis" Montgomery, Peck, Vining

4.2 Conclusions

The substitution of FAC from AAE is feasible. This was obvious even from the plot of FAC versus AAE where a simple linear model could explain the majority of the data.

By using the available information there was an improvement in the fit. It is subjective to say whether this improvement was good enough or not. Nevertheless, a comparison could be made. For instance, in Aluminum there was a 23% increase in the $adj - R^2$ whereas in Calcium only 0.3% . Of course Aluminum can be seen as an exception because of the high variation of the two methods. In the rest of the models there were considerably small differences.

These small differences were expected since these two methods measure equivalent quantities.

An important remark is that both FAC and AAE are measurements and in every measurement there is an error. In Section 2.1 it was mentioned that the estimated standard errors shall not be trusted.

As a result, if precision is the target then it is recommended to have available more measurements per sample in the future. Then an EIV analysis will be possible.

It could even be more important to have reliable estimates of the errors than doing an EIV. With these estimates there could be a judgement on which method is more precise.

5 Appendix

The purpose of the appendix is to present further details and plots concerning the previous models.

For all the models the following plots will be given:

- qqplot (model residuals against theoretical ones)
- fitted values against observed ones

5.1 Aluminum

The model is in section 3.2.

In the reported model for Aluminum 11 observations received 0 weight, 226 received 1 and the rest 6 received weights between 0.55 and 0.998.

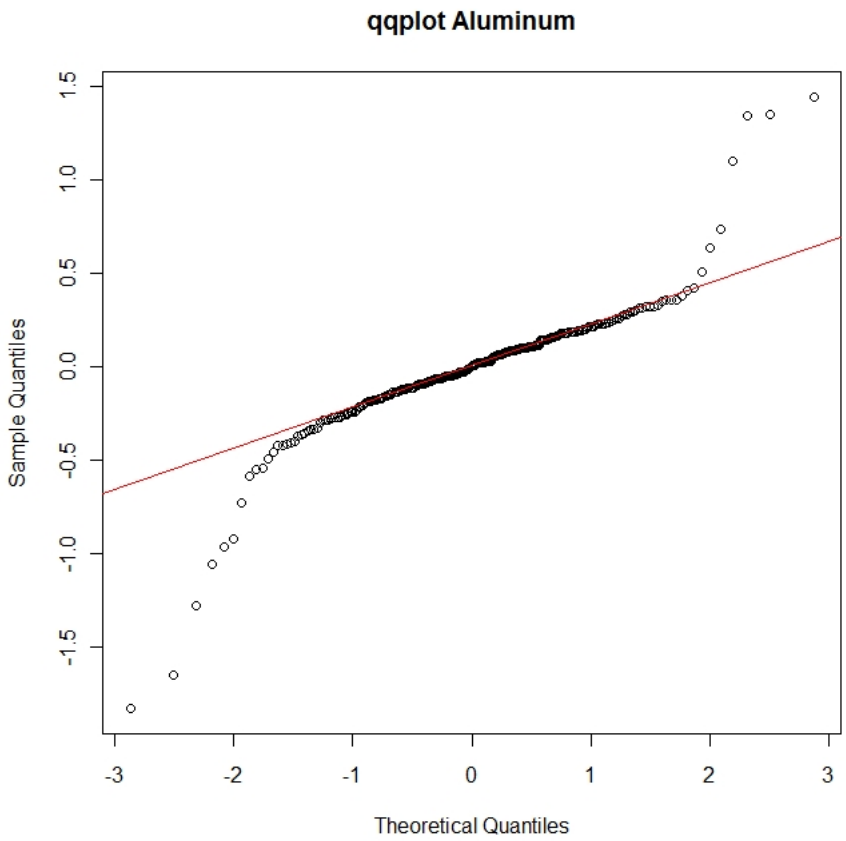

Figure 19: QQplot of the residuals of the model for Aluminum.

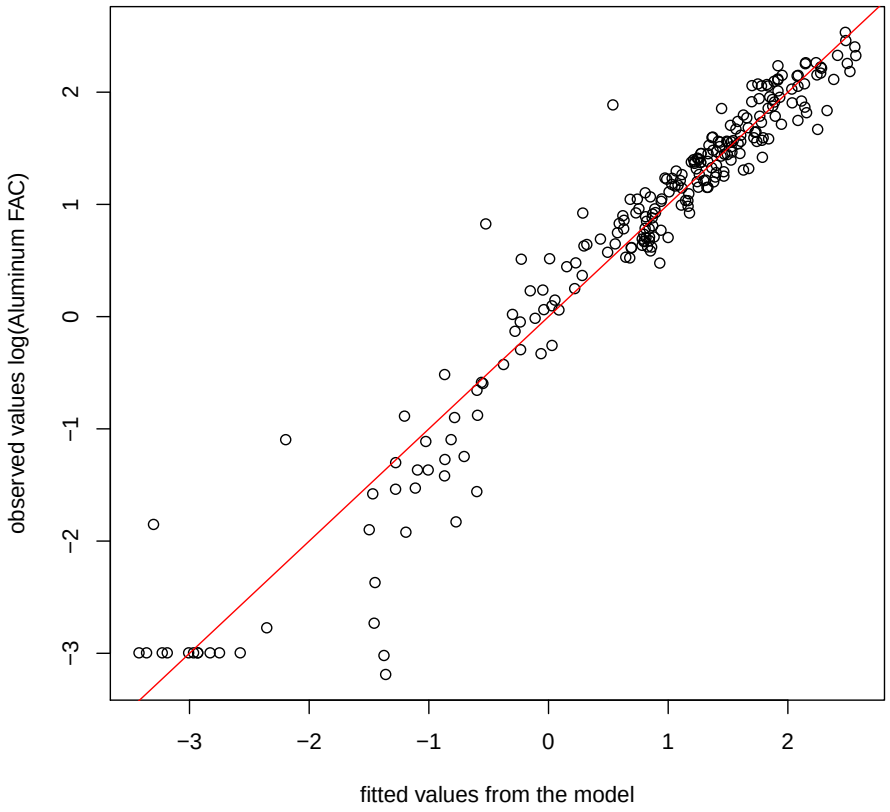

Figure 20: Model fit. As expected from figure 6 for larger values there is better fit. The model will perform better in higher values and the prediction in lower values will be more uncertain.

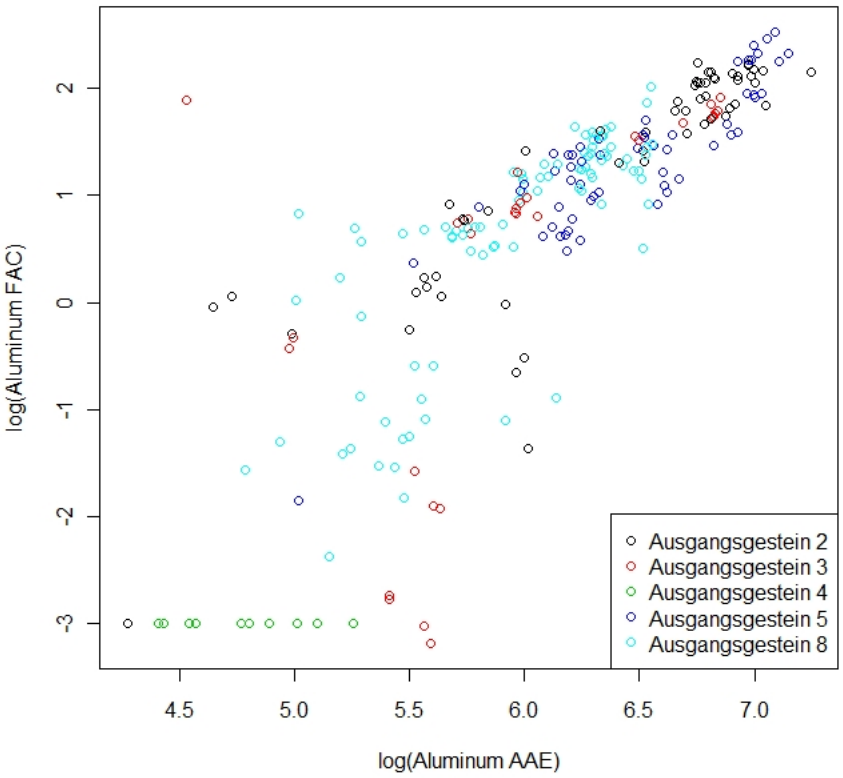

Figure 21: Plot of the two methods colored according to Ausgangsgestein. A small group of observations that behave differently corresponds to Ausgangsgestein 4.

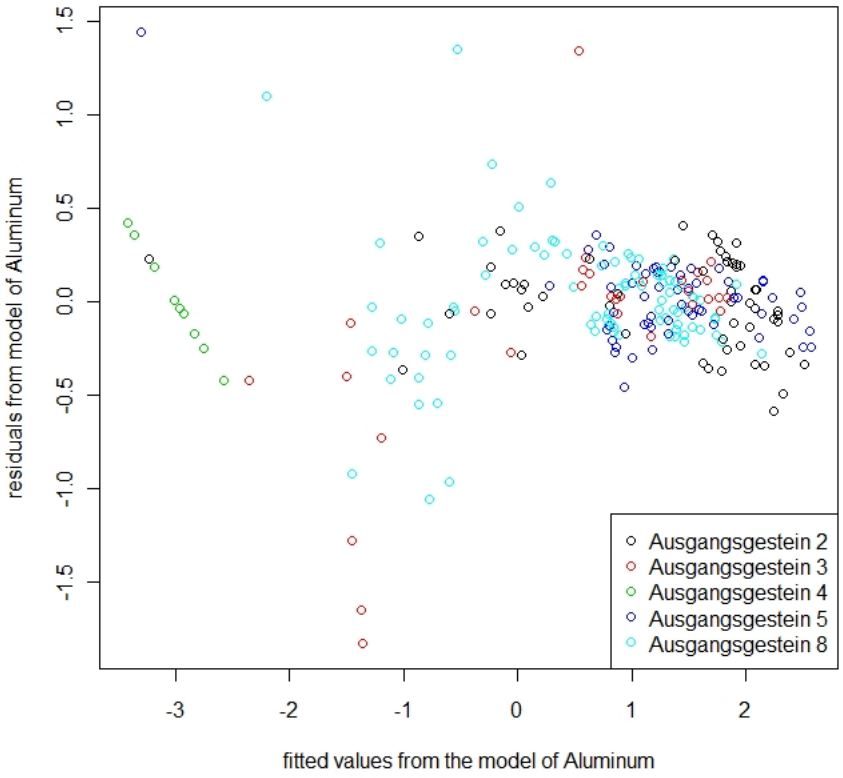

Figure 22: Residuals versus fitted values. For Ausgangsgestein 4 we get small residuals. The higher the fitted values the lower the residuals.

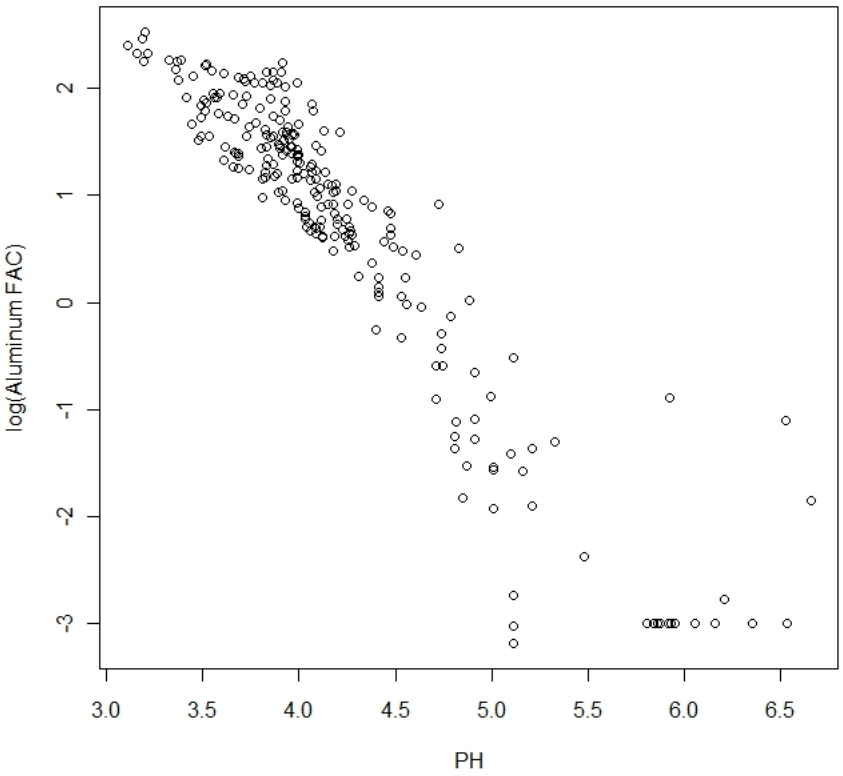

Figure 23: Plot of FAC against PH.

PH is indeed an important predictor. In a sense, PH is a better predictor than AAE for Aluminum if choosing only between these two variables. It is interesting that in the comparison of all possible models with only one explanatory variable the best model was the one with PH and not the one with log(Al_AAE)! This happened only in Aluminum where the agreement of the two methods was the worst compared to the other three compounds.

5.2 Calcium

The model is in section 3.3.

In the reported model for Calcium 176 observations received weight of 1 and the rest 66 observations received weight between 0.035 and 0.998.

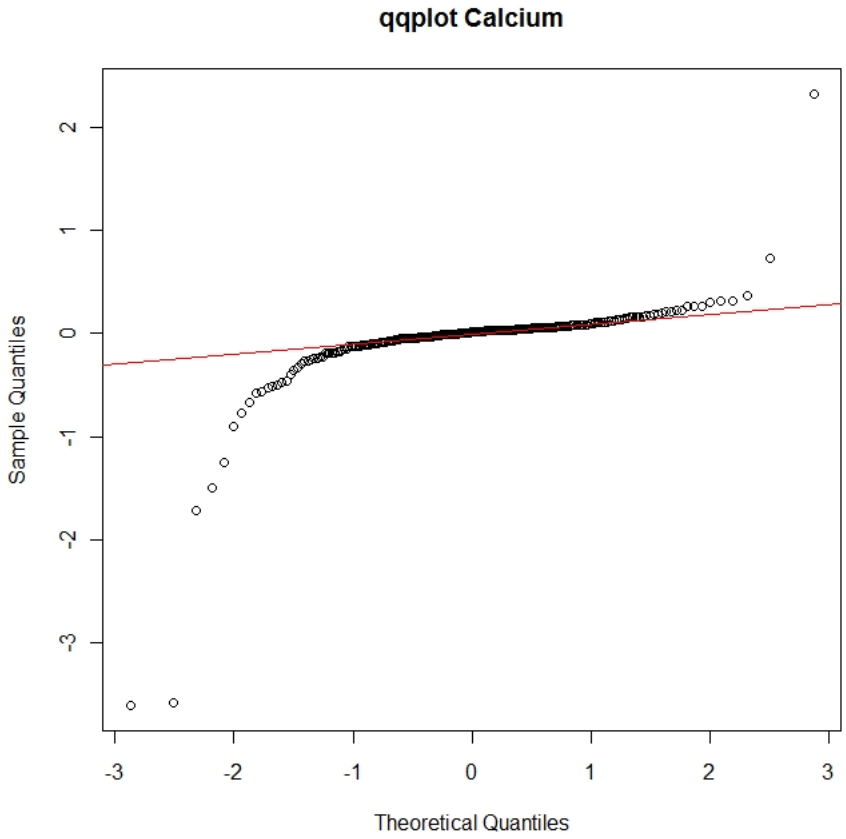

Figure 24: QQplot of the residuals of the model for Calcium.

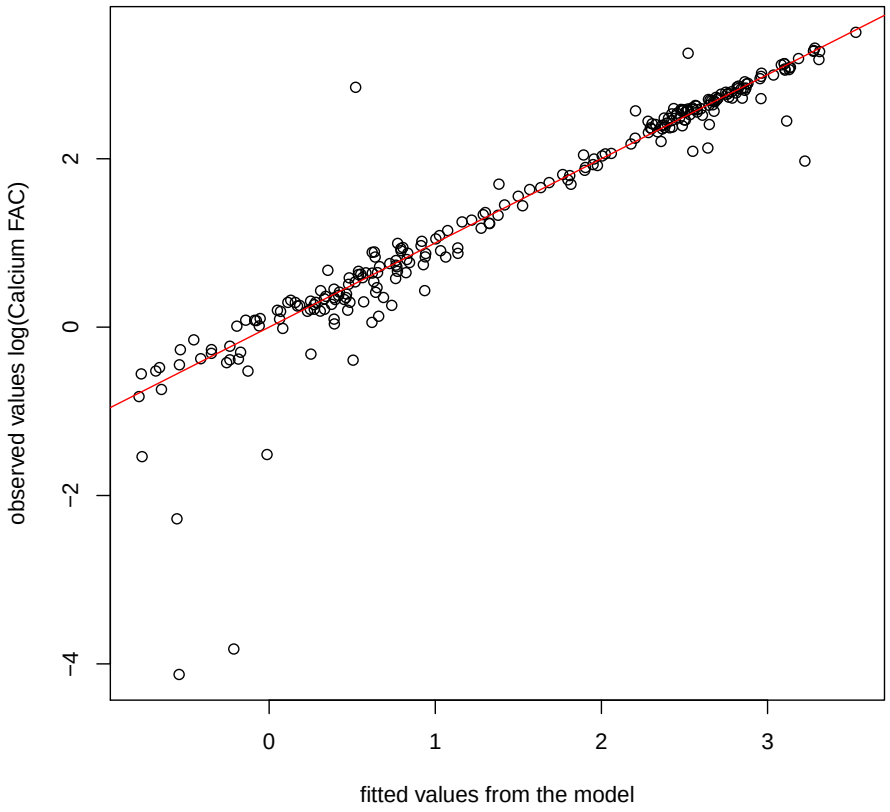

Figure 25: Model fit. As expected from figure 15 for larger values there is better fit. The prediction for lower values will be more uncertain.

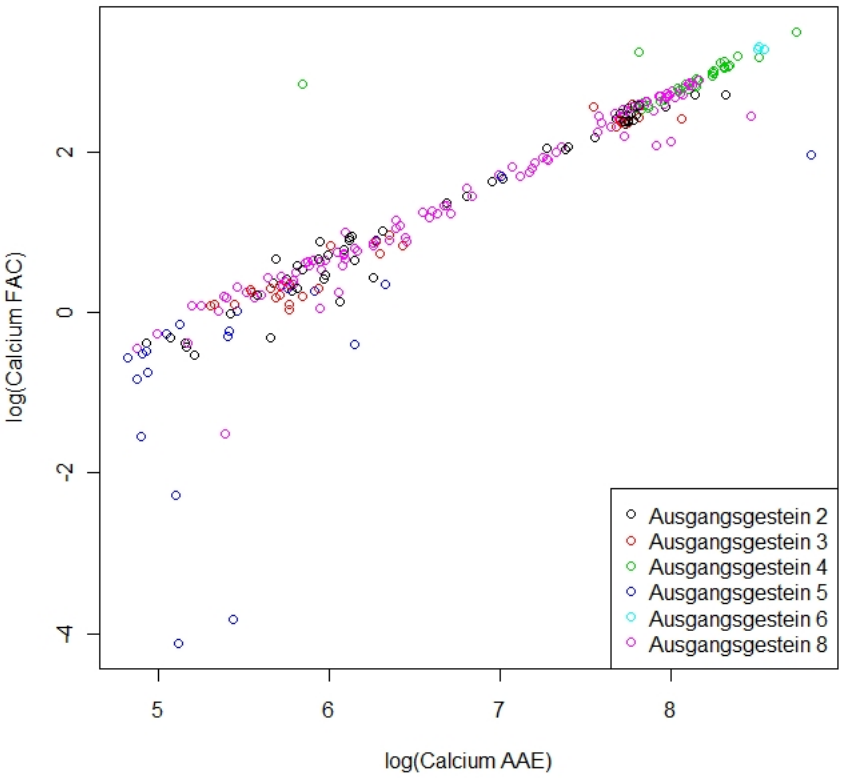

Figure 26: A group of observations with different behavior corresponds to Ausgangsgestein 5. It is also visible that some levels of Ausgangsgestein get only high or low values.

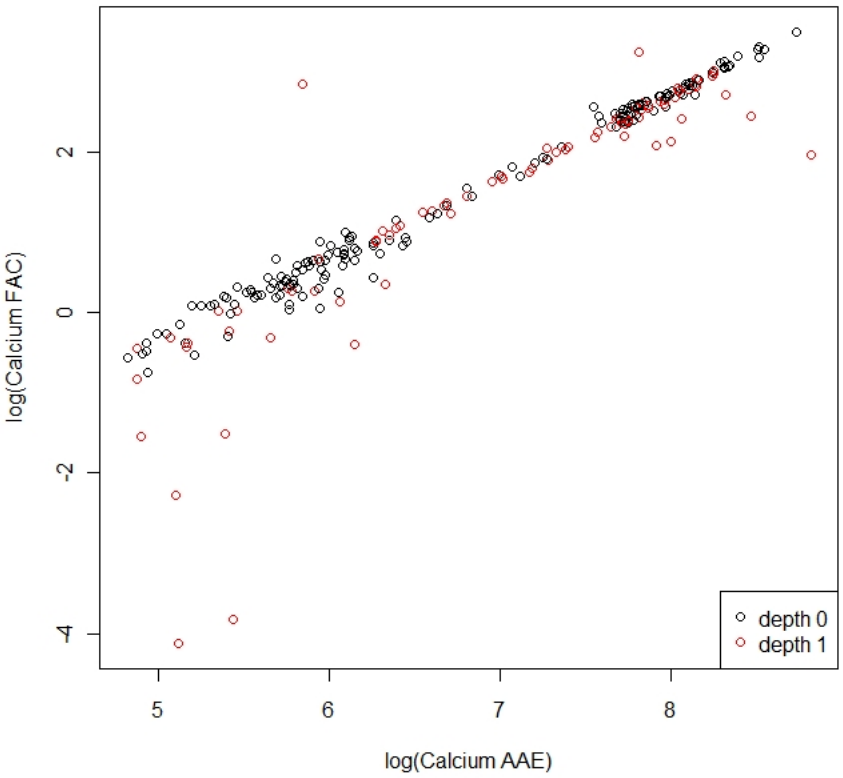

Figure 27: The observations from all the data range with different behavior come only from depth 1.

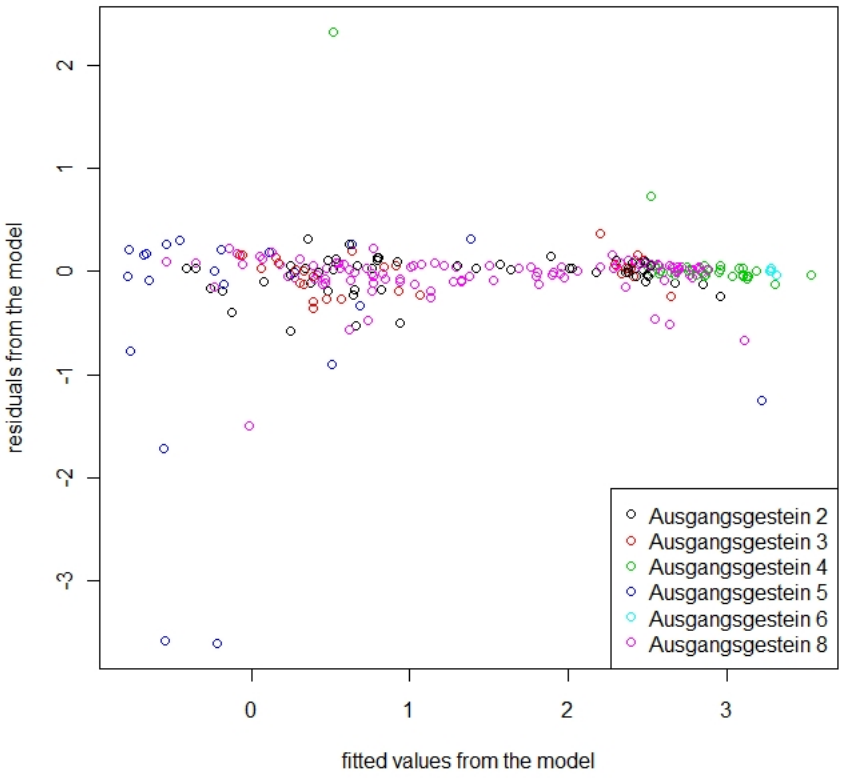

Figure 28: For Ausgangsgestein 5 the model produces larger residuals.

5.3 Kalium

The model is in section 3.4.

In the reported model for Kalium 8 observations received 0 weight, 270 observations received 1 weight and the rest 25 received weights between 0.01 and 0.99.

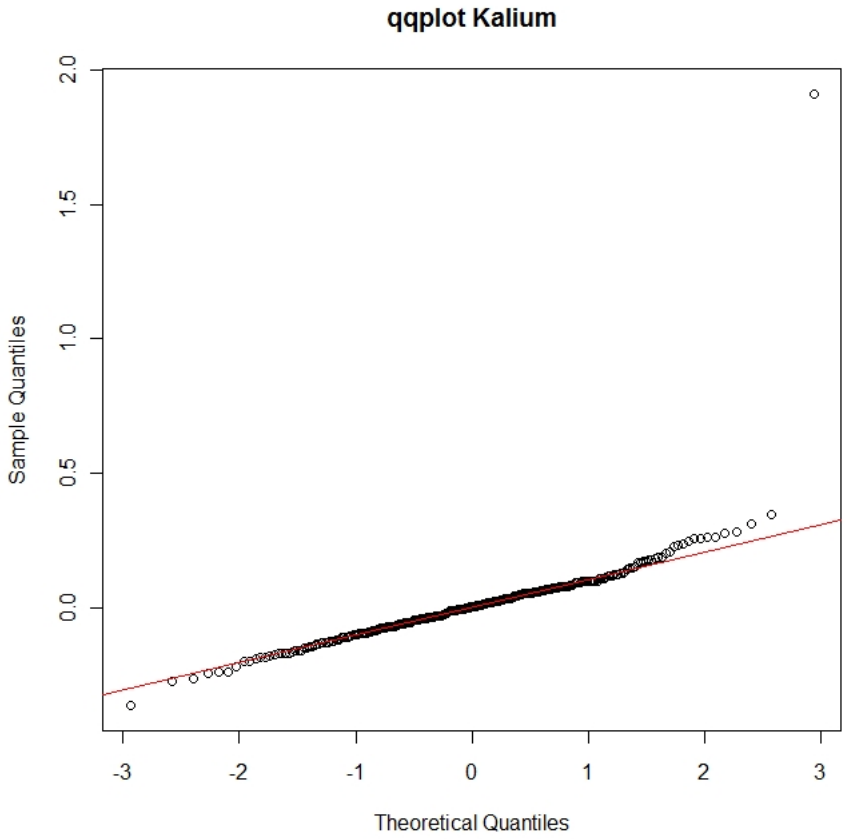

Figure 29: QQplot of the residuals of the model for Kalium.

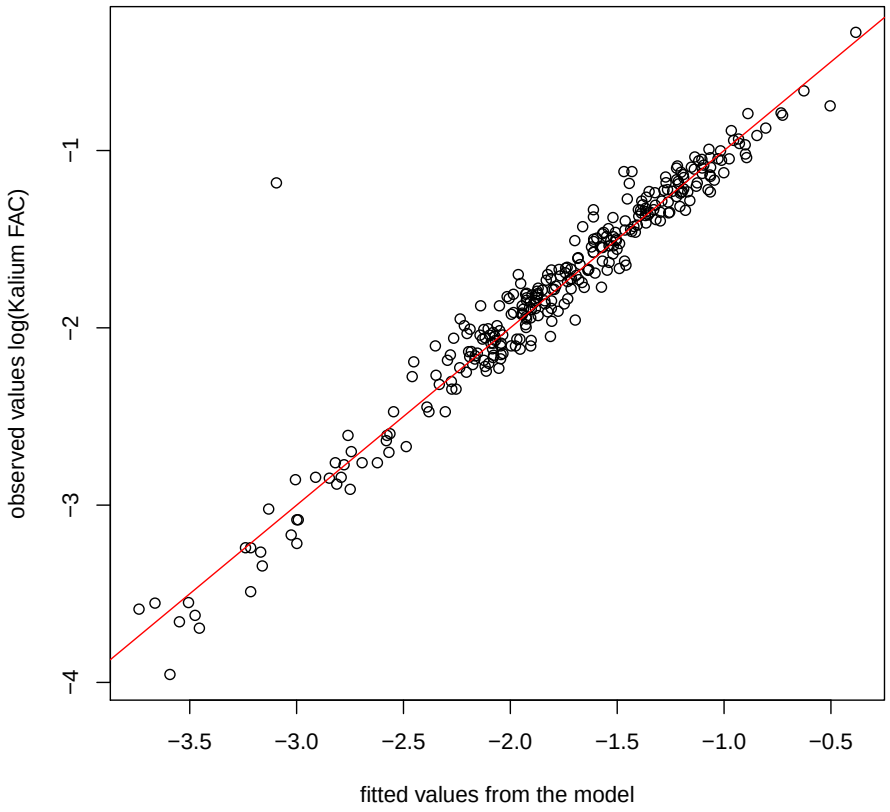

Figure 30: Model fit. For all the range of values the model performs well. The prediction is equally accurate within the data space.

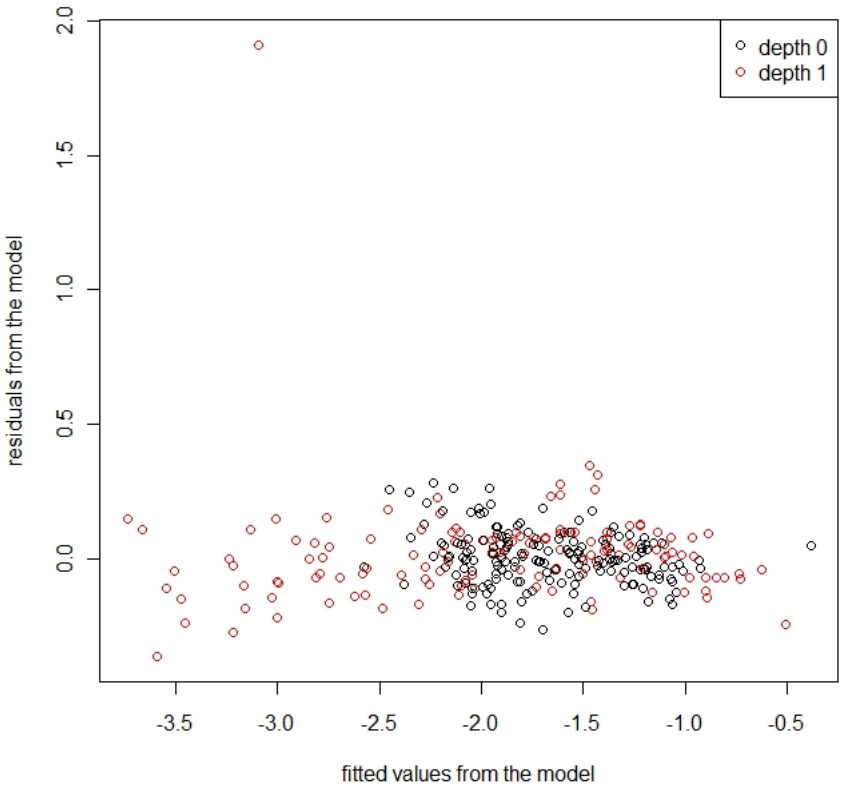

Figure 31: For depth 0 there are mostly high values. For depth 1 there are all possible values. It is a positive thing that for both levels of depth and for all the range of values there are small residuals.

5.4 Magnesium

The model is in section 3.5.

In the reported model for Magnesium 34 observations received 0 weight, 230 received 1, and the rest 22 received weights between 0.005 and 0.99.

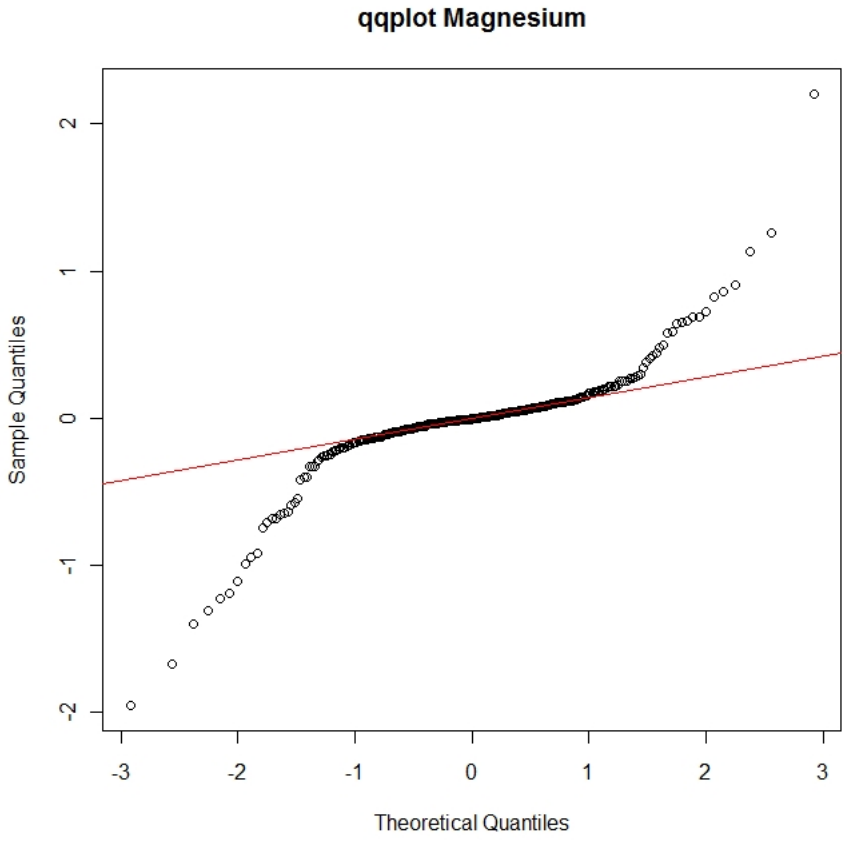

Figure 32: QQplot of the residuals of the model for Magnesium.

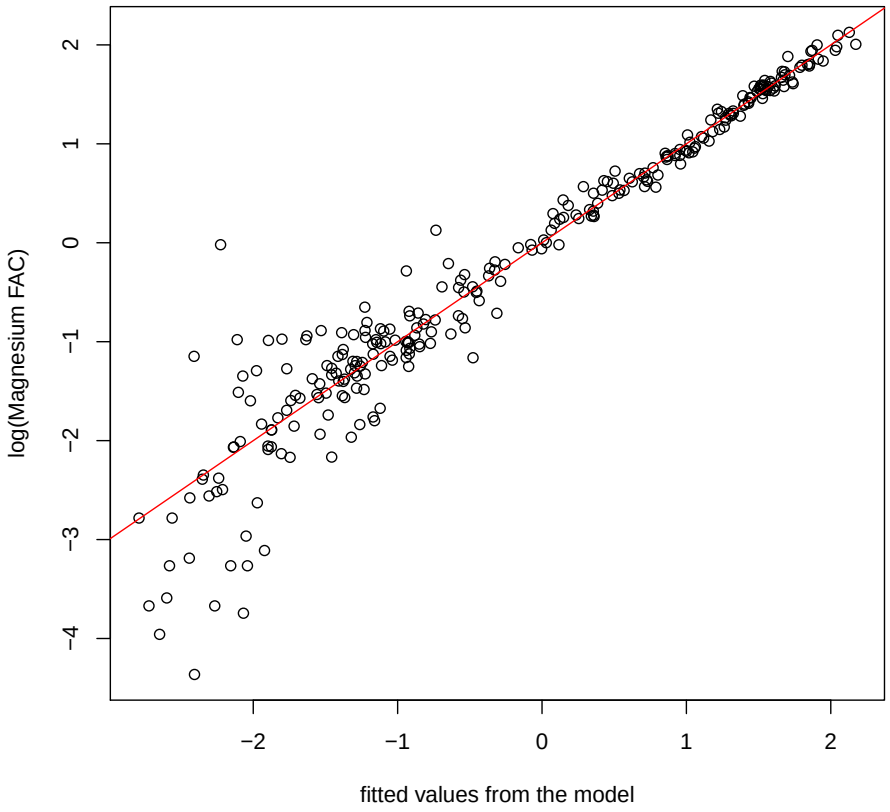

Figure 33: Model fit. As expected from figure 17 for larger values there is better fit. The prediction for lower values is more uncertain.

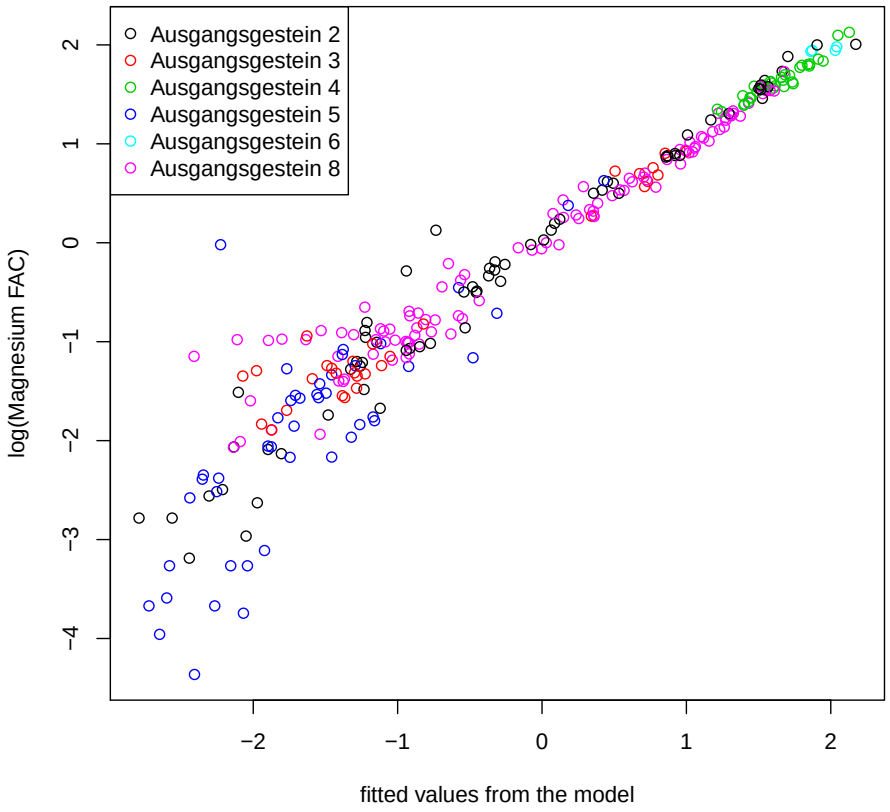

Figure 34: Model fit colored according to Ausgangsgestein. Different behavior for Ausgangsgestein 5. Including interaction between Ausgangsgestein and AAE did not improve the situation.

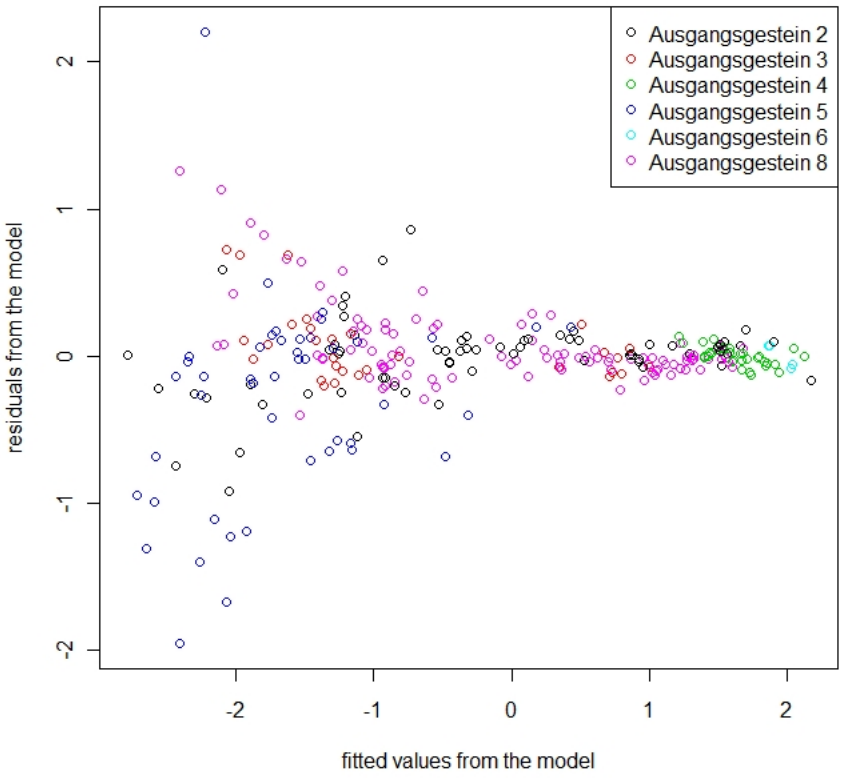

Figure 35: Large residuals for small values. The different behavior of Ausgangsgestein 5 resulted in getting large residuals for this level.

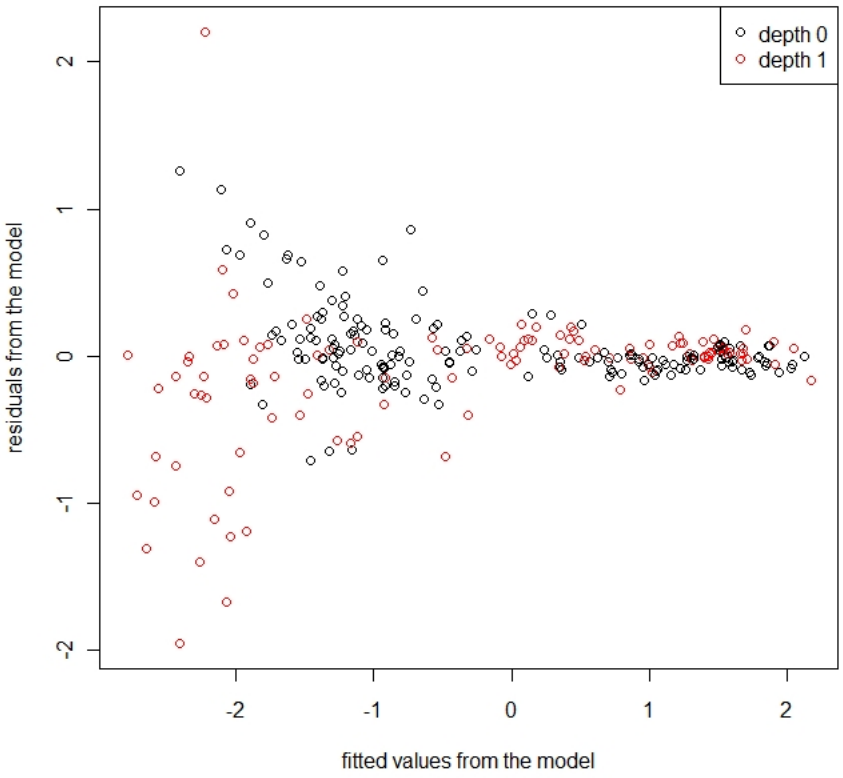

Figure 36: The observations with different behavior come mostly from depth 1. There is, though, a group of observations with depth 0 with larger residuals.

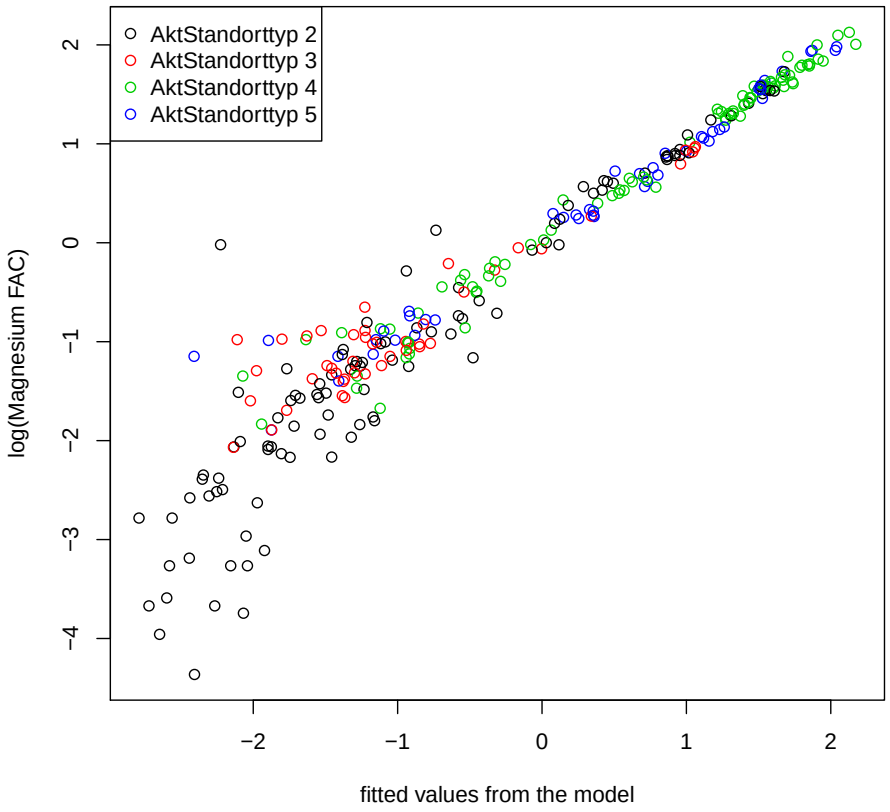

Figure 37: Model fit colored according to AktStandorttyp.

Different behavior for a group of observations of AktStandorttyp 2. This group deviates from the rest of the observations which follow a clearer line. However this group corresponds to Ausgangsgestein 5 as shown before. In fact with Ausgangsgestein 5 the group of observations with different behavior is more apparent. This is because in AktStandorttyp 2 there are also observations well explained and that do not deviate from the behavior of the rest. Although from figure 11 we would expect AktStandorttyp to be in the model it is now obvious that it is not needed and that Ausgangsgestein is more informative.

5.5 kakeff

The model is in section 3.6.

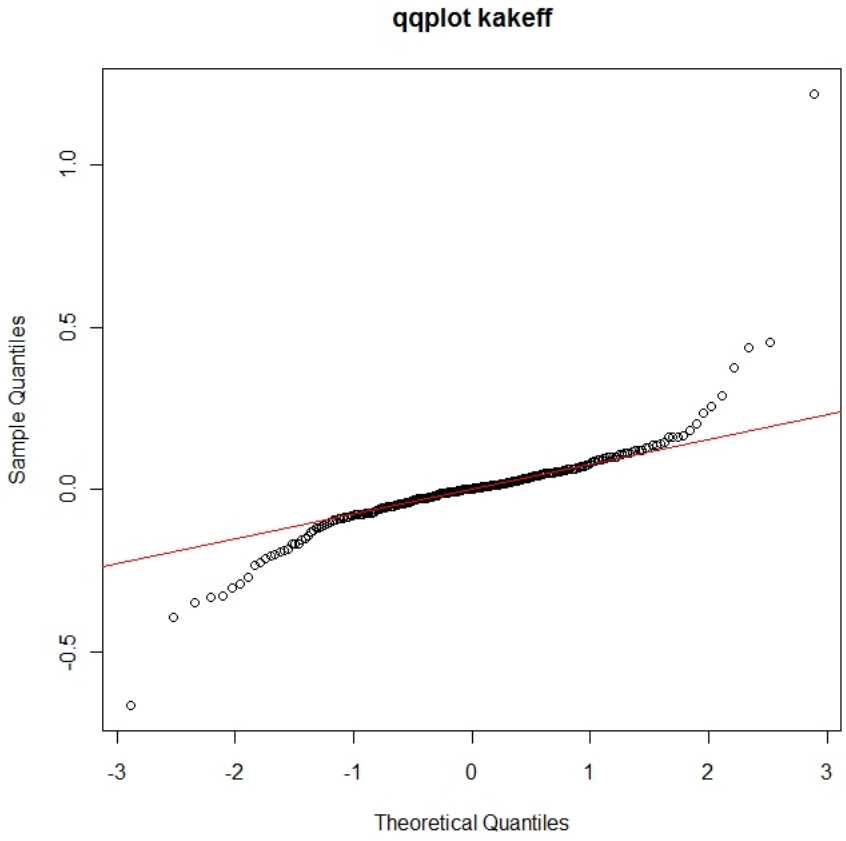

Figure 38: QQplot of the residuals of the model for kakeffcorad.

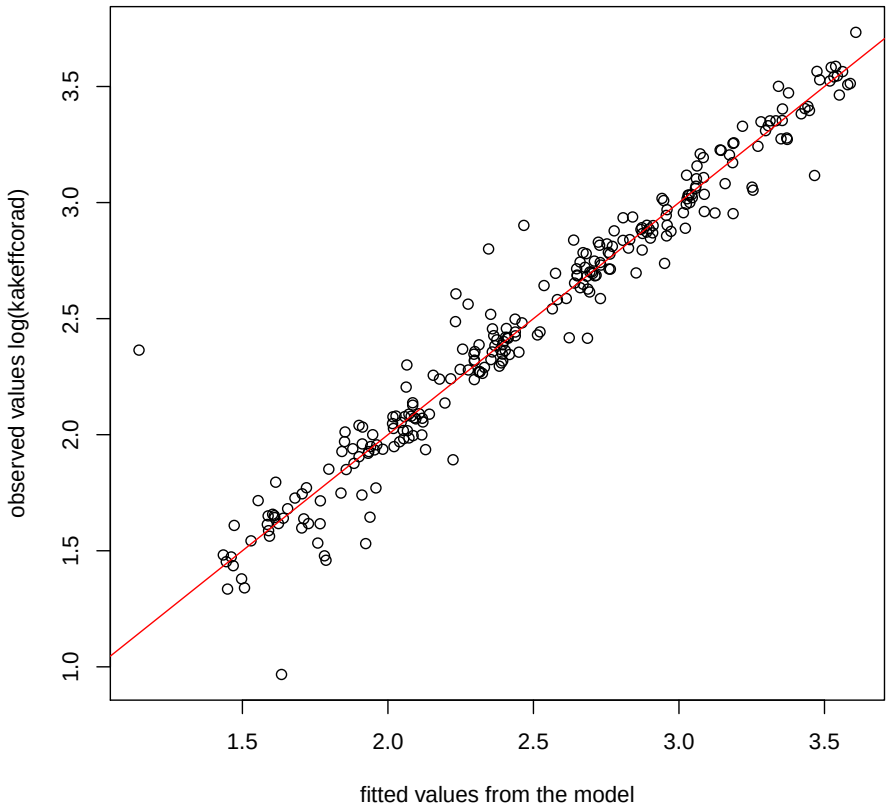

Figure 39: Model fit.

By taking into account that there is no direct AAE measurement for this quantity the fit can be characterized as very good. Another advantage is that for all (low, medium, high) concentrations there are no large residuals produced and the prediction for all the data space will be equally good and certain.

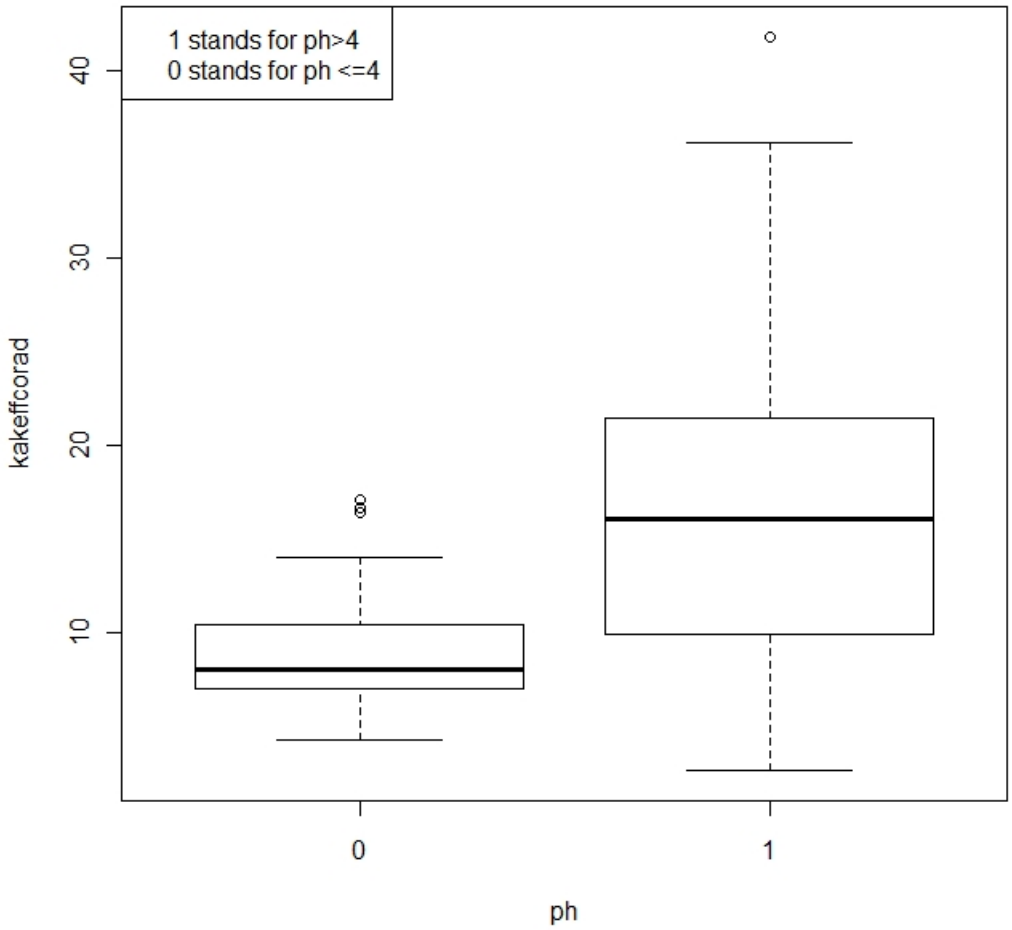

Figure 40: Boxplot of kakeffcorad against the binary PH.

Clear difference in the two groups. Models with continuous PH and binary PH where compared. The decision to include a binary PH was based in the stepwise linear relationship of kakeff on the compounds and more specifically on Aluminum.

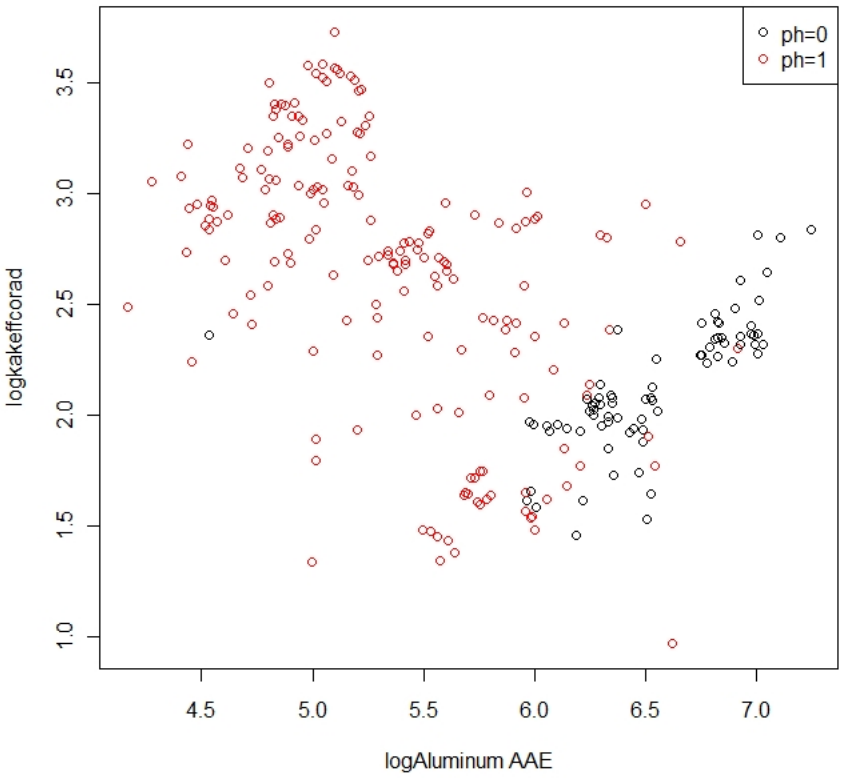

Figure 41: Plot of kakeffcorad against Al_AAE colored according to PH. This plot demonstrates the necessity of adding the interaction term.