P-03-93

Oskarshamn site investigation

Calculation of Fracture Zone Index (FZI) for KSH01A

Lennart Lindqvist, Bergsten & Co i Värnamo AB
Hans Thunehed, GeoVista AB

October 2003

Svensk Kärnbränslehantering AB
Swedish Nuclear Fuel
and Waste Management Co
Box 5864
SE-102 40 Stockholm Sweden
Tel 08-459 84 00
+46 8 459 84 00
Fax 08-661 57 19
+46 8 661 57 19

ISSN 1651-4416
SKB P-03-93

Oskarshamn site investigation

Calculation of Fracture Zone Index (FZI) for KSH01A

Lennart Lindqvist, Bergsten & Co i Värnamo AB
Hans Thunehed, GeoVista AB

October 2003

Keywords: fracture zone index, FZI, fracture frequency, borehole geophysics, logging, multivariate analysis.

Abstract

The aim of this work was to carry out a multivariate calculation of a Fracture Zone Index, FZI, along the borehole KSH01A in the Simpevarp area. This will generalize and integrate information from geophysical logs, geological mapping and manual classification to a numerical description of the fracture properties of the rock.

The available data have been joined into a matrix with common and uniform section lengths through averaging, interpolation, resampling and manual classification in order to create comparable sections along the borehole.

A manual classification (GFZI) of the borehole in three types of classes was performed in order to define the properties that FZI is supposed to describe. These types are core of fracture zone (GFZI=2), transition zone (GFZI=1) and normal unaffected rock (GFZI=0).

Relations between objects and variables were analyzed with Principal Component Analysis (PCA) and outliers were identified and removed from the data set.

A regression model that describes the relation between the significant input variables and the manual classification, GFZI, was established with Projection to Latent Structures (PLS). FZI was then calculated for all sections along the borehole based on the PLS-model.

Contents

1 Introduction

Multivariate statistics in the form of Principal Component Analysis, PCA, and Projection to Latent Structures, PLS, is well documented analysis techniques. The methods for multivariate analysis that were developed during the 1970's are described in detail in /1/.

Multivariate analysis has become a popular tool within several sciences, including geoscience where it has been used for a long time for e.g. analysis and evaluation of chemical and petrophysical variables in exploration /2/, /3/, /4/.

Large amounts of data have been created during the last years in investigation sites managed by SKB. These data are well suited for multivariate analysis and applications are described by /5/, /6/, /7/. The methods are also used in some countries where assessments of rock volumes for localisation of a repository for spent nuclear is ongoing, e.g. /8/.

The calculation of FZI was initially described in /7/.

Analysis was performed by Bergsten & Co in Värnamo AB and GeoVista AB in accordance with the instructions and guidelines from SKB (activity plan AP PS 400-03-048 and method description MD 810.003, SKB internal controlling documents) and under supervision of Leif Stenberg, SKB.

2 Objective and scope

The aim of this work was to carry out a multivariate calculation of a Fracture Zone Index, FZI, along the borehole KSH01A in the Simpevarp area. This will generalize and integrate information from geophysical logs, geological mapping and manual classification to a numerical description of the fracture properties of the rock in a robust and objective way.

The calculation of FZI with multivariate techniques is based on measured and observed quantities along the borehole. The PLS-model used to calculate FZI can be used on data from other boreholes provided that they are from a similar geological environment.

The most important prerequisite for this analysis is the definition of what FZI is supposed to describe. This will be the quantity that the measured and observed data will try to model and predict. The borehole is therefore divided into sections of three discrete intensities of fracturing (GFZI) based on manual classification.

3 Equipment

Multivariate statistical calculations have been performed with Simca-P version 8.0 (Umetrics AB). Grapher (Golden Software) has been used for presentation of the final results.

4 Execution and results

4.1 Pre processing of data

The pre processing of input data is summarized in Table 4-1.

The upper 100 metres of KSH01A were percussion drilled. However, a cored borehole (KSH01B) was drilled close to and parallel with KSH01A. The data representing the first 100 metres in this study are hence from KSH01B and the remaining data from KSH01A.

A common section length of one metre was chosen for all variables in this work. This choice was partly based on previous experiences but also on the fact that e.g. the core-mapped fracture frequency is available in one metre sections. This choice is not critical and appears sound since significant fracture zones often have a width of several metres.

Longer sections of e.g. 5 to10 metres would probably create mixing of different zone classes and the borders between zones would be blurred.

Shorter section lengths would not create any technical problems and would even be advantageous for some variables like e.g. the sonic log which show short wave-length anomalies for fractures.

The alteration parameter was initially given in discrete sections not coinciding with the one metre sections of the data matrix. The borders between sections with different degrees of alteration were therefore rounded off to the nearest even metre and sections shorter than one half metre were removed. Almost all alteration sections were labelled “oxidation” so no distinction was made between different types of alteration. Radar reflections were initially given as coordinates of reflectors crossing the hole. This information was converted to the number of reflectors per one metre section.

The geophysical logs (magnetic susceptibility, density, natural γ-radiation, caliper, 16" and 64" normal resistivity, focussed resistivity, single point resistance and P-wave sonic) were all initially measured at 0.1 metres intervals. Average values for one metre sections were calculated. The electrical, sonic and caliper logs were also deconvolved to give weighted discrete source indications. These were summed in one metre sections.

Missing values are indicated by the number -999 in the data file.

Table 4-1. Pre processing of data for calculation of FZI.

Processed primary data	Pre processing	Resulting data file
Core-mapped fracture frequency from SICADA.		KSH01alla_var.xls (MS Excel).
Core-mapped alteration from SICADA.	Alteration: Rounding off to even metres. Omission of sections < 0.5 m.	
Geophysical logs from SICADA.	Geophysical logs: Averaging in one metre sections, resampling. Electrical, sonic and caliper logs: Deconvolution and summation.	
Radar reflections from SICADA.	Radar reflections: calculation of number of reflections per metre.	
Coded GFZI from SICADA.	Creation of common data matrix.	

4.2 Variables for analysis

A total of 18 input variables were available in the data matrix resulting from the pre-processing. Additionally there was a column describing the manual classification of fracture intensity (GFZI), a column for section identity and a column for length along the borehole.

The GFZI variable was slightly modified so that 5 sections on either side of a fracture zone were given the value 0.05 to indicate the proximity to the transition zone. These sections are called the near zone.

The list below shows the acronyms that have been used for the various variables in the text and in figures in the rest of this report.

Acronym

Id = Identity consisting of borehole number plus length in metres
e.g. 1588 means hole #1 (KSH01A) and length 588 to 589 metres
GFZI = Geological Fracture Zone Index
= 2 core of fracture zone
= 1 transition zone
= 0.05 near zone
= 0 normal rock
Ra = Radar reflex
Ff = Fracture frequency
Ms = Magnetic susceptibility (SI)
Sr = Single point resistance, SPR (Ω)
So = Sonic P-wave velocity (m/s)
N6o = 64" normal resistivity (Ωm)
N1o = 16" normal resistivity (Ωm)
Ng = Natural γ-radiation (μR/h)
F3o = Focussed resistivity (Ωm)

De = Density (kg/m^3)
Cp = Caliper (mm)
Srd = SPR deconvolved
Sod = Sonic deconvolved
N6d = 64" normal resistivity deconvolved
N1d = 16" normal resistivity deconvolved
F3d = Focres300 deconvolved
Cpd = Calip deconvolved
Al = Alteration

Output variables

NGFZI = Numerical Geological Fracture Zone Index calculated with PLS-technique
FZI = Continuous Fracture Zone Index based on NGFZI
FZI can also be assigned discrete values according to:
= 2 core of fracture zone, NGFZI > 1.5
= 1 transition zone, 0.5 < NGFZI < 1.5
= 0 unaffected rock, NGFZI < 0.5

4.3 Incomplete sections and outliers

Values are missing for all input variables in the interval 99 to 101 metres. These three sections were therefore omitted from the analysis. The section between 97 and 98 metres also contained missing values for some variables but this section was kept in the analysis. Out of total 988 sections, three were omitted and the analysis was performed on the remaining 985.

No random outliers should be allowed to influence the analysis in a serious way. Multivariate statistics was used to identify such outliers. A total of six sections were removed for this reason namely Id 1249, 1250, 1251, 1254, 1560 and 1590 and a new analysis was performed resulting in a model denoted M2.

The results from the model M2 gives a reliable and robust view of the data matrix.

37 multivariate outliers were identified to be outside the confidence limit of the PLS model where FZI is calculated.

The model becomes stable after removal of the outliers and no significant difference in the model can be observed if additional sections are removed. This makes the analysis less dependent upon which subset of the data matrix that is used to calculate the model.

The modelling with the help of PLS is also performed with cross validation, where parts of the data matrix are alternately removed to test the significance and repeatability of the model.

4.4 Multivariate analysis

Multivariate statistics is treated in detail in /1/ and summarized in Appendix 1 (See also MB 810.003, SKB internal controlling document). A brief introduction is given below.

Two main techniques are used. Principal Component Analysis (PCA) is used to describe the relations and correlations between different variables. The second technique, Projection to Latent Structures (PLS), is used to describe the relations between a number of independent variables X and a dependent variable Y. The result is similar to regression analysis with full graphical control in all steps of the measurements and variables that are related to Y. Stepwise components from the X-space are added to describe the variation in Y-space and to create the description model.

The correlation structure between the input variables is used to calculate a model. A single measured value can be assumed to consist of two parts, one that can correlate with other variables and a second that can be treated as uncorrelated noise (Figure 4-1). The benefit with multivariate statistics is that the uncorrelated part of the total value can be eliminated and only the rest remains for further analysis.

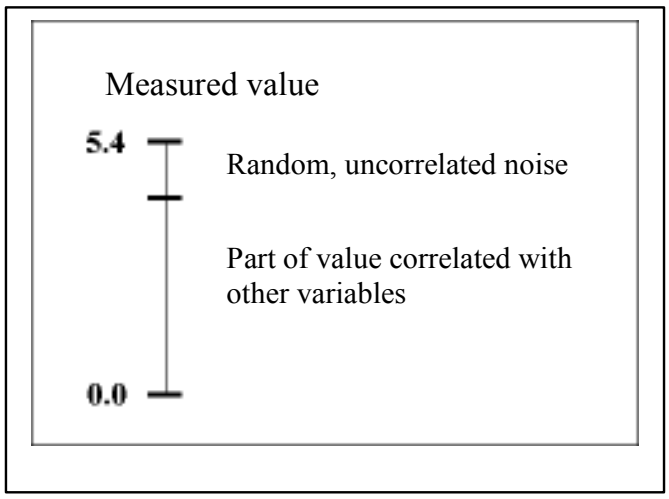


Figure 4-1. *The contents of a single measurement.*

4.5 Definition of FZI

Initially a manual classification of information from the borehole called GFZI was performed. This classification describes the intensity of fracturing in discrete levels.

The core of the fracture zone has been given the value GFZI=2 whereas the transition zone to normal rock has been given the value GFZI=1. The vicinity of the transition zone have been complemented with a near zone where GFZI=0.05. Rock out side these zones, normal unaffected rock, have been given the value GFZI=0.

The approach is that of looking for a model that describes the rock in terms of fracturing only and where other properties e.g. lithology does not interfere. The model might consist of several components with the common properties that they correlate with GFZI and that they contribute to a more robust description of FZI. Those properties

of the rock mass that are reflected in the variables but not correlate to GFZI are automatically eliminated.

4.6 General relations between variables

The general relations between data have been analyzed with PCA. No attempt has been made at this stage to make any kind of prediction of FZI. A seek of an understanding of the relations within the data set was performed.

The correlations between variables are visualised in a number of variable loading plots below. Variables that plot close to each other show correlation in these plots whereas variables on the opposite side of the plot show reverse correlation. Variables in the distal parts of the plot show strong correlation whereas variables close to the origin shows weak correlation with the other variables. The horizontal and vertical direction vectors are orthogonal and by definition uncorrelated to each other.

4.6.1 Principal Component Analysis

Principal Component transformation was performed for all variables and objects in the data matrix. The normalized distance to the centre of the model is shown in Figure 4-2. Some outliers are evident in the plot.

Sections with DmodX(PS),N >3.0, i.e. those that have a normalized distance to the centre of the model greater than 3 standard deviations, were omitted from further analysis. Six sections were removed (Id 1249, 1250, 1251, 1254, 1560, 1590) and 979 sections remained for analysis.

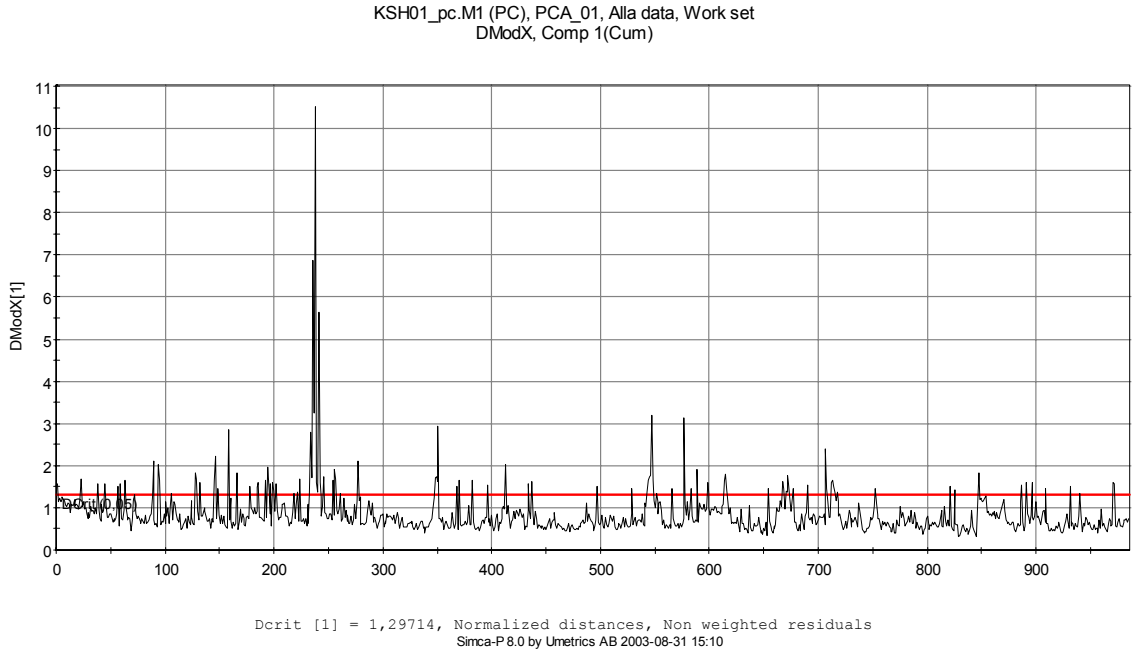


Figure 4-2. *Normalized distance (in standard deviations) of sections in the data matrix from the centre of the initial PCA-model.*

PCA for model M2

A new PCA-model was calculated for the remaining data called model M2. The normalized distance from the model centre is plotted in Figure 4-3. Some sections show distances greater than 3 in the plot but they appear to be located to the extreme ends of significant trends in the data set. They were therefore not removed.

The PC-analysis of the data matrix revealed four significant components that describes 58.7 % of the total variation in the data. The first two components show typical properties related to fracturing (first component – horizontal) and lithology (second component – vertical) (Figure 4-4 and 4-5).

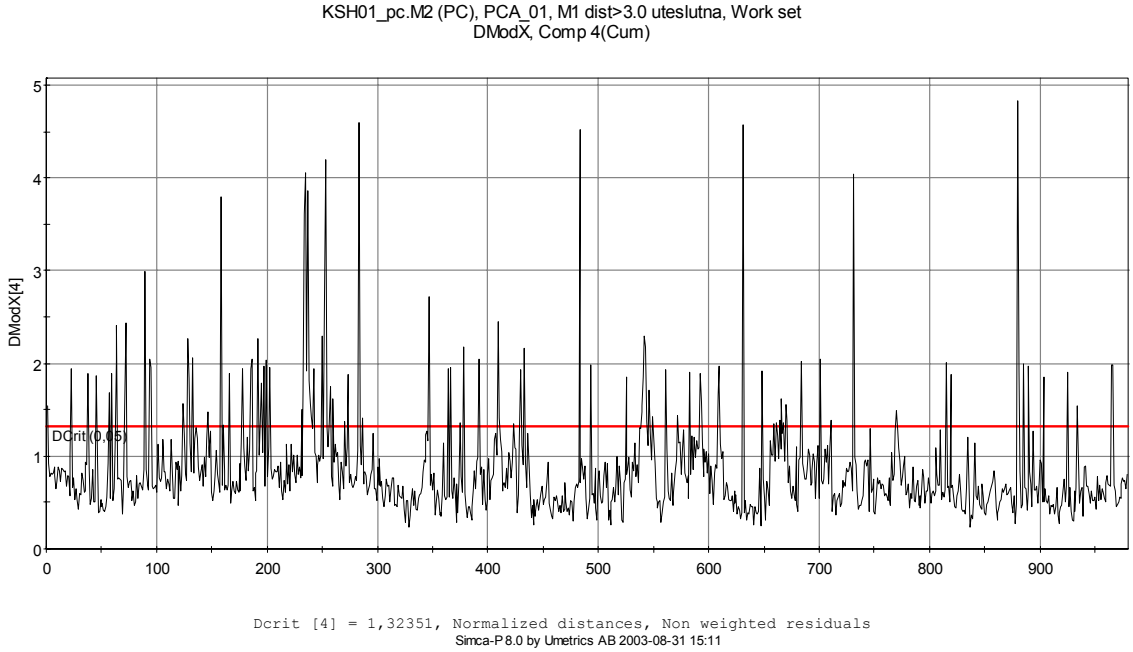


Figure 4-3. *Normalized distance (in standard deviations) of sections in the data matrix from the centre of the PCA-model M2.*

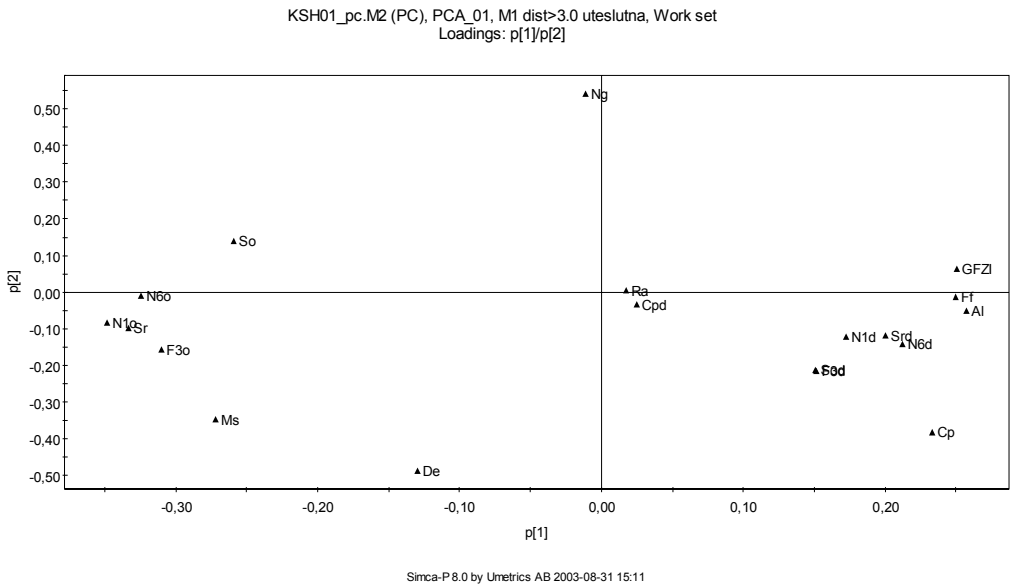


Figure 4-4. *Variable loadings for the first two PC's for model M2.*

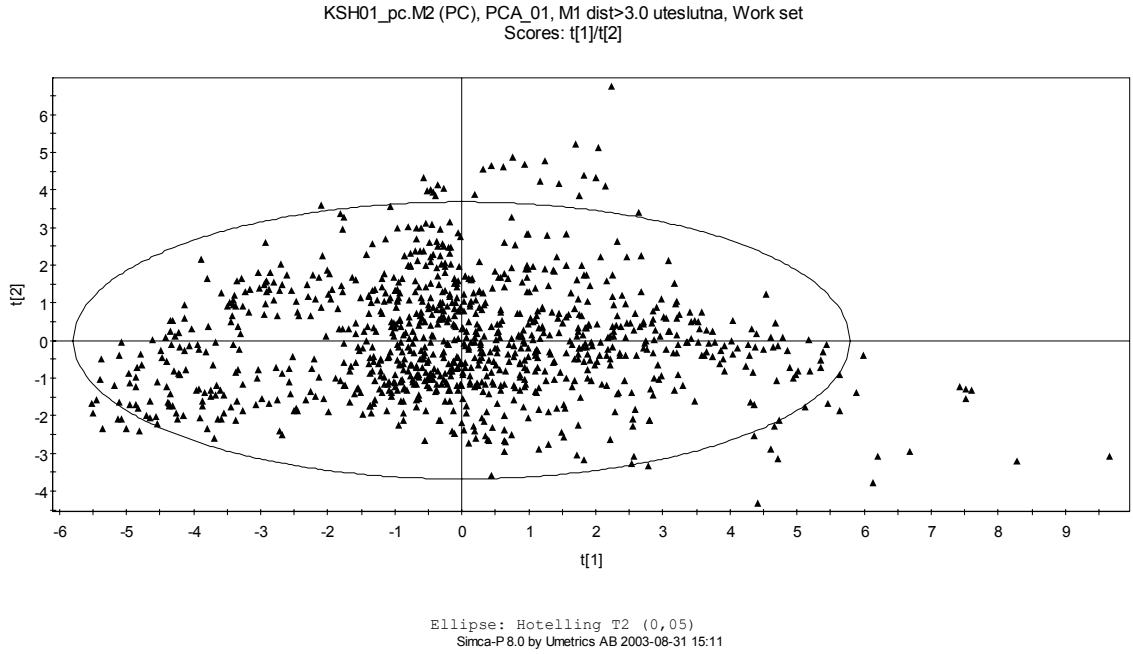


Figure 4-5. *Object scores for the first two PC's for model M2.*

Model PCA_M2: 1st PC – 29.4 % of total variation

The horizontal direction in the variable loading plot (Figure 4-4) indicates fractured rock. The variables to the right indicate fracturing with high/positive values whereas those to the left indicate fracturing with low/negative values. The interpretation is that the property of the rock that is most strongly described by the data matrix is fracturing and that it correlates with the manual classification GFZI.

Positive direction
GFZI, Ff, Al, Cp, N6d, Srd, N1d, Sod, F3d
Uncorrelated
Ng, Ra, Cpd
Negative direction
N1o, N6o, Sr, F3o, So, Ms, De

Model PCA_M2: 2nd PC – 11.9 % of total variation

The vertical direction in the variable loading plot (Figure 4-4) indicates different types of lithology. Variables in the upper part of the plot typically have high values for felsic rocks (Ng) and low values for mafic rocks. The opposite applies for variables in the lower part of the plot (Ms, De). Caliper shows correlation with density and magnetic susceptibility on the side of fractured rock. This might indicate that fractures in mafic rocks produce stronger caliper anomalies than fractures in felsic rocks. However, there is a more or less linear decrease in borehole diameter with length in the hole and at the same time felsic rocks are more abundant at depth. This might cause a correlation between caliper and rock type that is not related to the properties of the rock mass.

Positive direction
Ng
Uncorrelated
All other variables
Negative direction
Ms, De, Cp

Model PCA_M2: Pc1 & Pc2 – 41.3 % of total variation

The object score plot in Figure 4-5 (=borehole sections) shows a fairly homogeneous data set with trends towards the extreme values. An example of an object that relates to the Ng-variable is Id 1364, to increased fracturing Id 1259, to normal rock Id 1035 and to the De-variable Id 1108.

Model PCA_M2: 3rd PC – 10.0 % of total variation

The variable loadings can be seen in Figure 4-6. The variables So and De are inversely correlated with electrical logs. The object score plot (Figure 4-7) shows a funnel shape for the majority of samples and a second group of samples to the left with high N6o, Sr, N1o, F3o, Ms.

The interpretation of this component is that it at least to some extent indicates the effect of variations in the salinity of the borehole liquid that has affected the electrical logs.

Positive direction
N6o, Sr, N1o (and Sod on the side with high fracture frequency)
Uncorrelated
All other variables
Negative direction
So, De

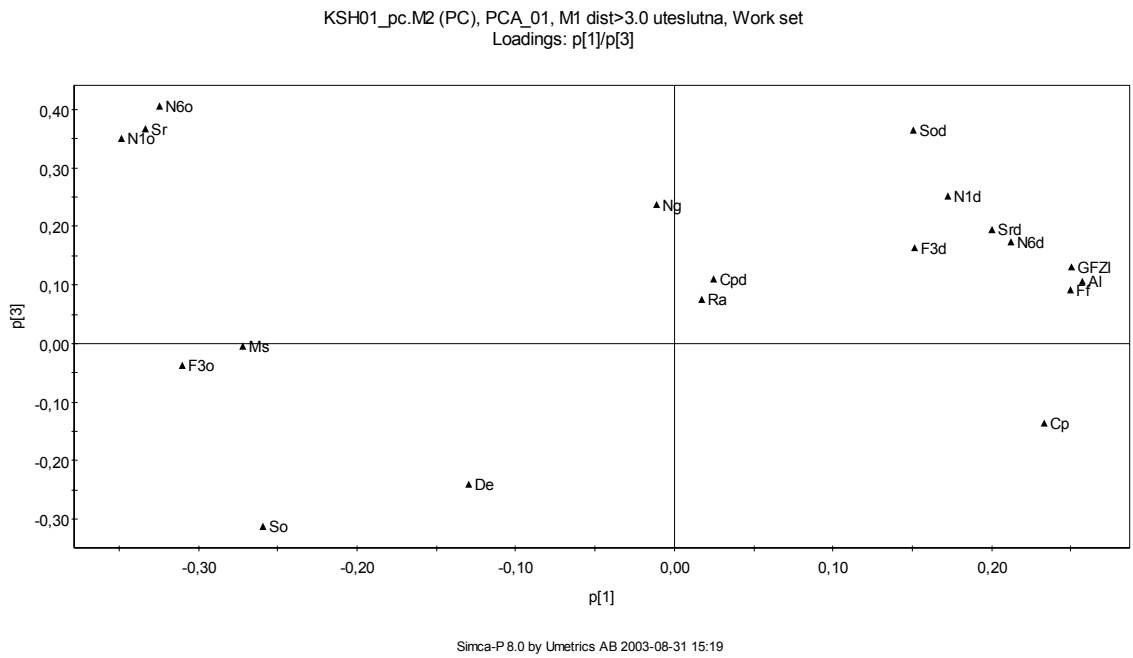


Figure 4-6. *Variable loadings for Pc1 and Pc3 for model M2.*

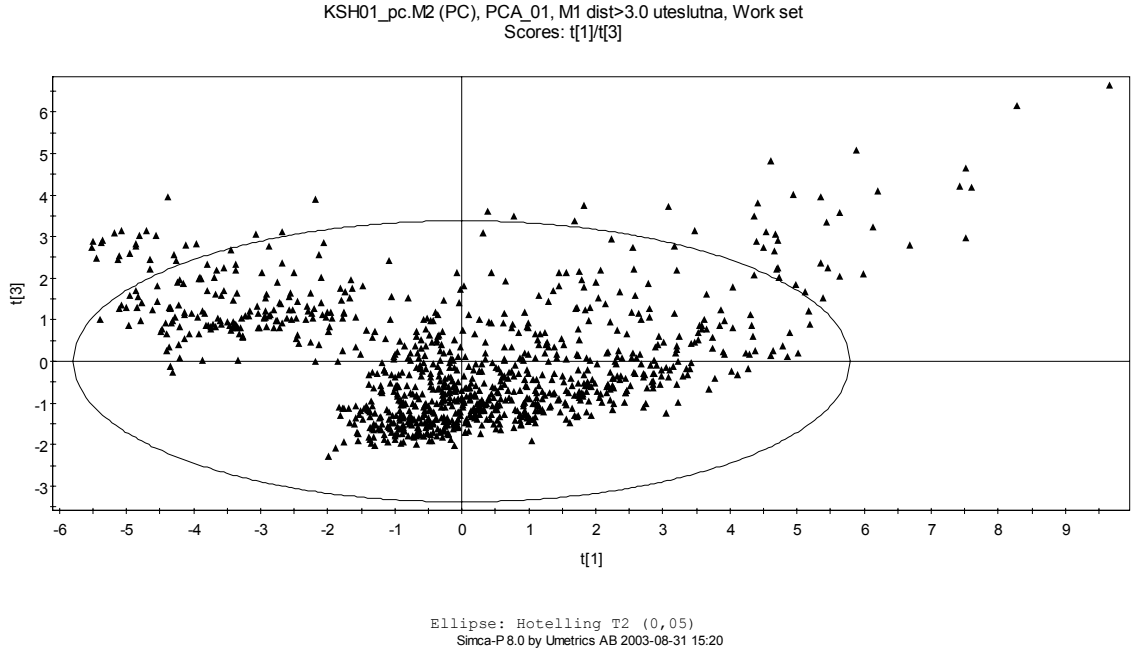


Figure 4-7. *Object scores for Pc1 and Pc3 for model M2.*

Model PCA_M2: 4th PC – 7.4 % of total variation

This PC relates to the fracturing of the rock (Figure 4-8, Figure 4-9). There is also some correlation with γ-radiation.

The component indicates fractures (negative scores) with significant anomalies in some geophysical logs with high resolution (F3d, Sod) but not in alteration and GFZI. Ng also contributes to negative scores. This indicates fractures unrelated to alteration but possibly related to felsic rocks.

Positive direction
GFZI, Al
Uncorrelated
All other variables
Negative direction
F3d, Sod and to some extent Ng

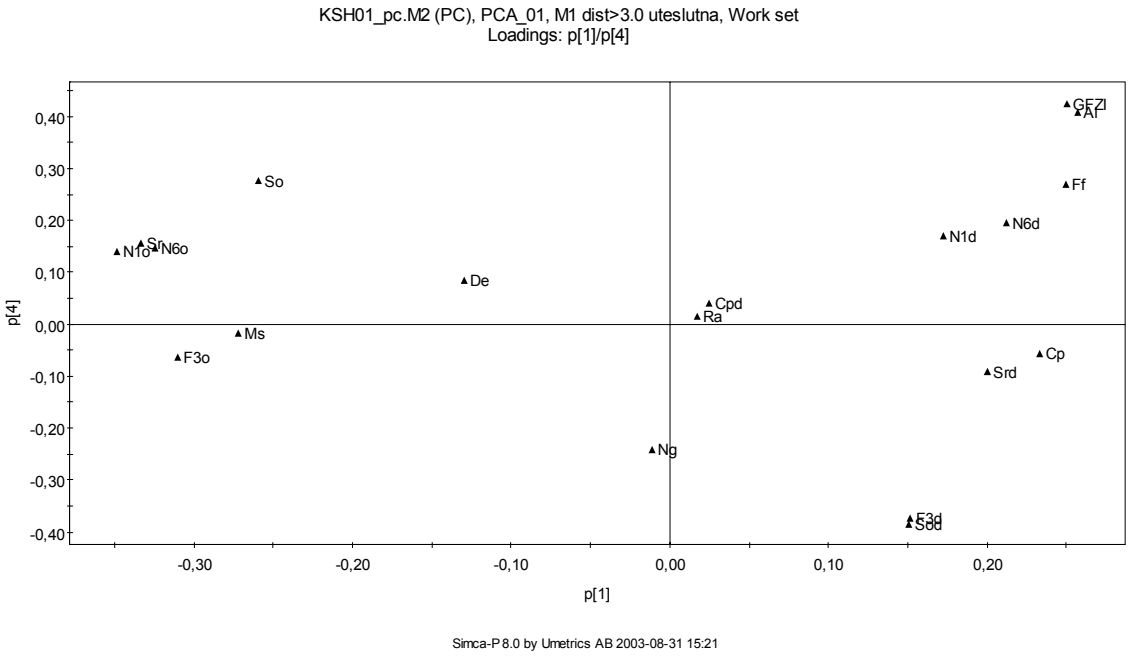


Figure 4-8. *Variable loadings for Pc1 and Pc4 for model M2.*

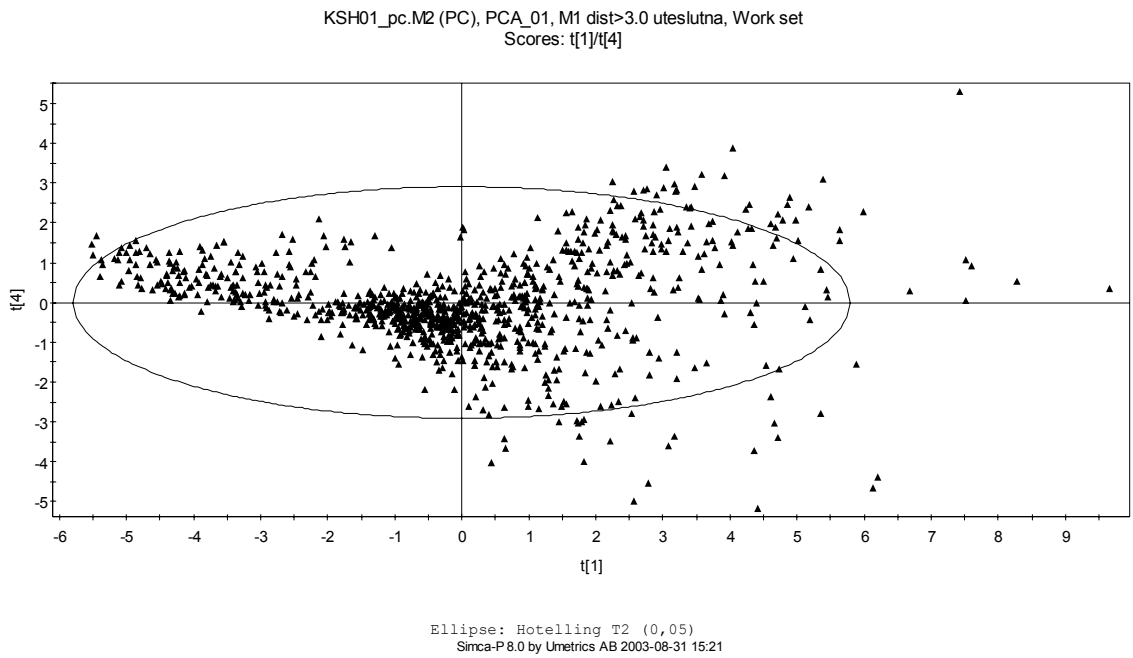


Figure 4-9. *Object scores for Pc1 and Pc4 for model M2.*

Model PCA_M2: Pc1, Pc2, Pc3 & Pc4 – 58.7 % of total variation

58.7 % of the variation in data is explained with a clear relation to fracturing and lithology.

The variables Ra and Cpd do not show any correlation with fracturing or lithology and are located close to the origin in all variable plots. These variables are independent on this general level.

For completeness, the relations between Pc2/Pc3, Pc2/Pc4 and Pc3/Pc4 are shown graphically in the Figures 4-10 to 4-15.

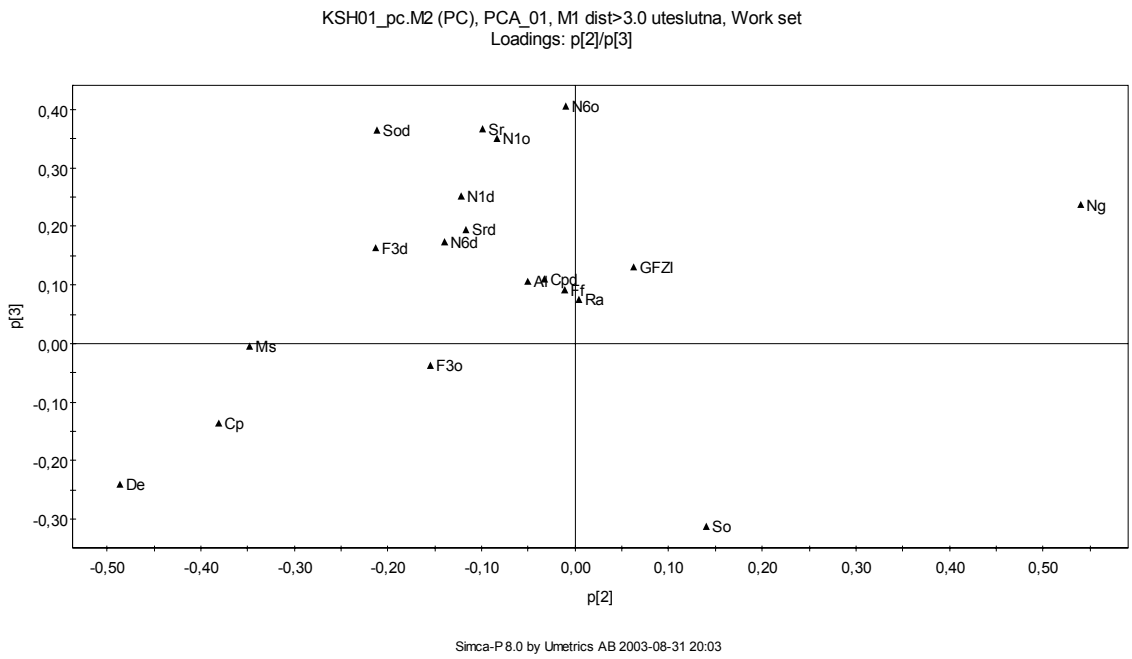


Figure 4-10. *Variable loadings for Pc2 and Pc3 for model M2.*

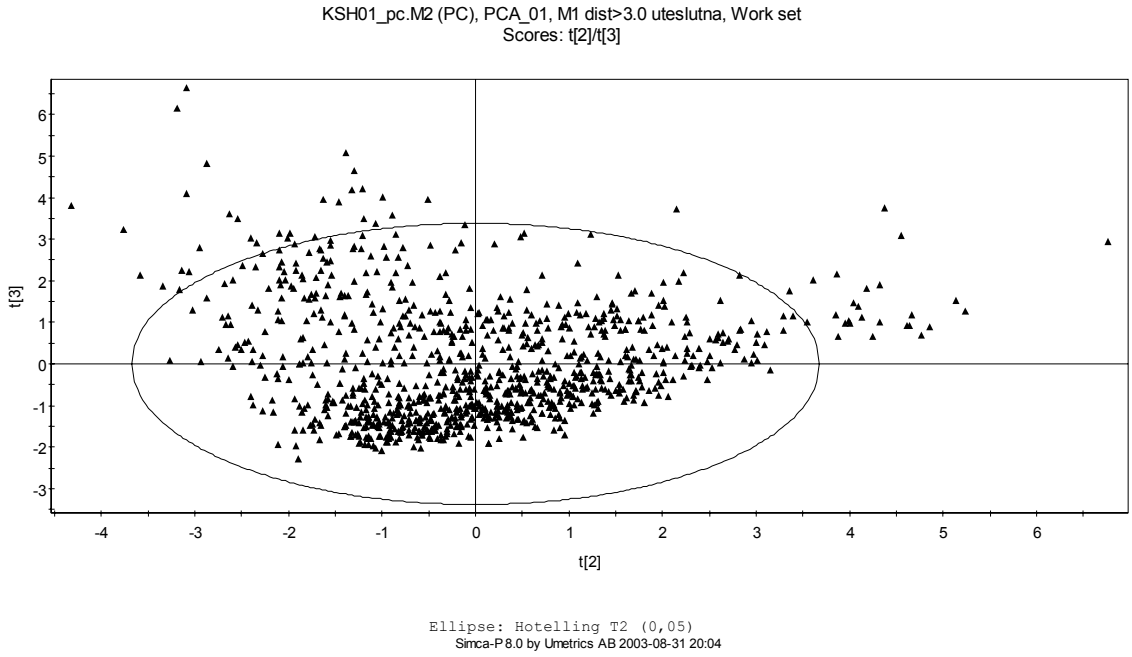


Figure 4-11. *Object scores for Pc2 and Pc3 for model M2.*

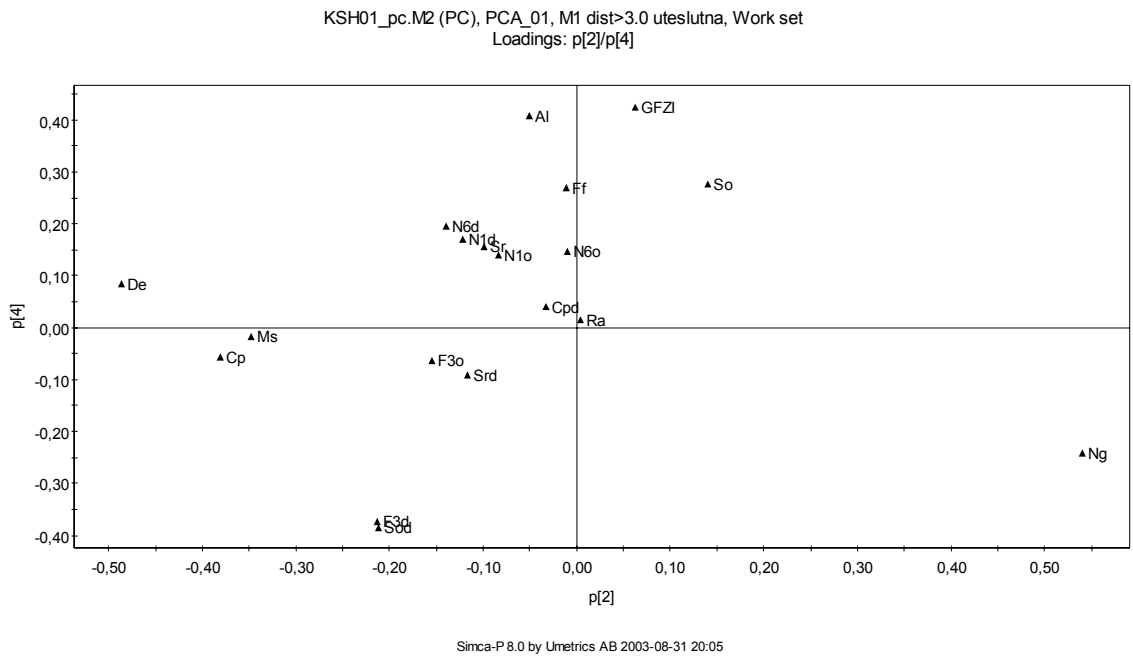


Figure 4-12. *Variable loadings for Pc2 and Pc4 for model M2.*

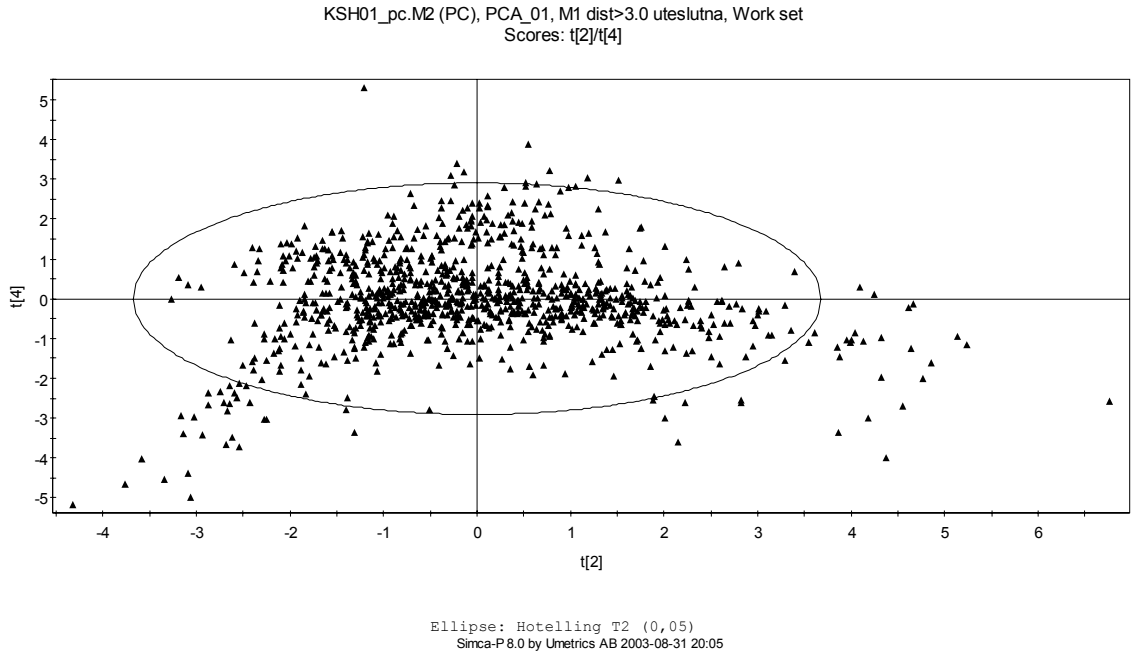


Figure 4-13. *Object scores for Pc2 and Pc4 for model M2.*

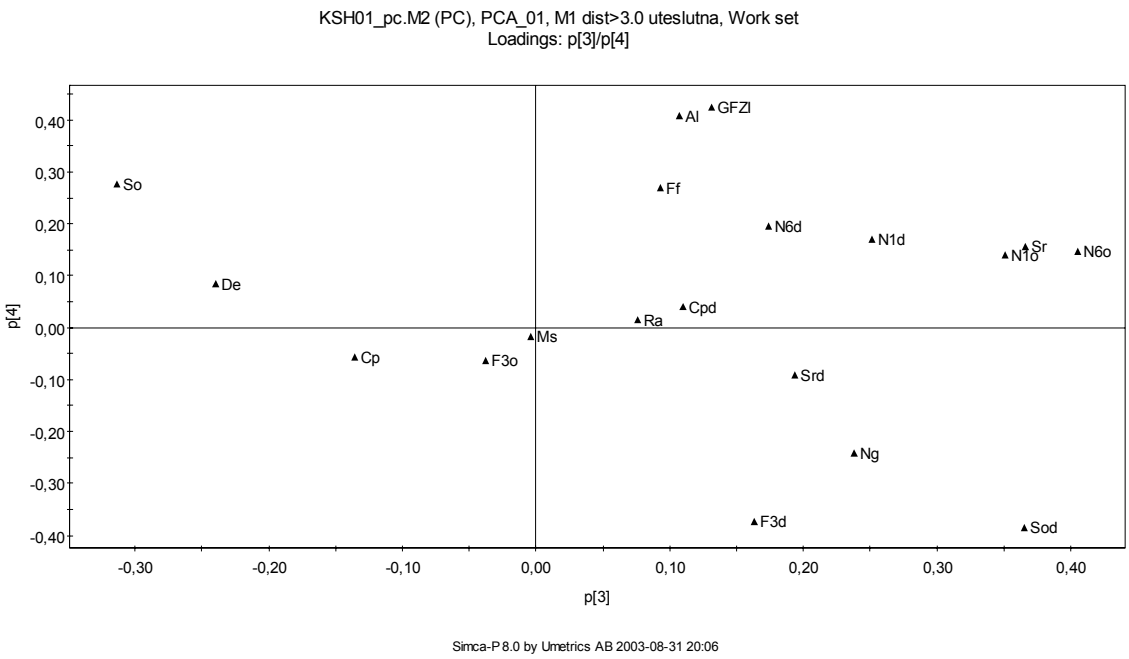


Figure 4-14. *Variable loadings for Pc3 and Pc4 for model M2.*

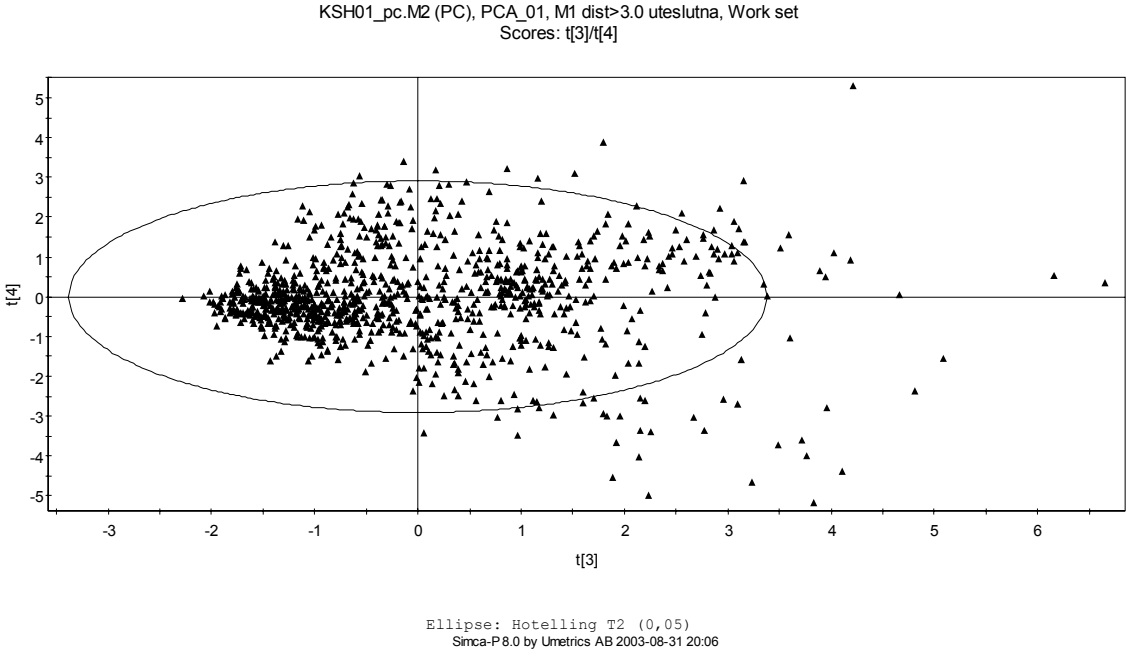


Figure 4-15. *Object scores for Pc3 and Pc4 for model M2.*

4.6.2 Summary of PC-analysis

The PC-analysis is summarized in Table 4-2.

The results from the PC-analysis generate four significant components. The position along these components for every section is denoted object score, t1, t2, t3, t4. The object scores are also plotted against length along the borehole in Figure 4-16.

Additionally, the normalized distance to the model centre has been calculated for every borehole section and is given in number of standard deviations, DModX(PS),N.

A probability value, PModX(PS), is also calculated for every section that indicates the probability for the section to be within the confidence limits of the model. If this probability is grater than 5 % we can assume that the section is within the confidence limits.

The two parameters that indicate whether a section complies with the model are defined as:

DModX(PS),N

Distance to the model in X space after n components for the observations used to fit the model. The distance is the standard deviation of the observations with scaling and centering. N stands for normalized distance.

PModX(PS)

Probability of belonging to the model in the X-space for observations used to fit the model. Observations with probability of belonging of less than 5 % are considered to be non members i.e. they are different from the normal observations used to build the model.

Table 4-2. PC-analysis of borehole data.

Processed primary data	Processing	Resulting data file
KSH01alla_var.xls	Creation of initial PCA-model Removal of outliers Creation of final PCA-model, M2 Calculation of principal component scores (t1, t2, t3, t4), distance to model [DModX(PS),N] and probability of belonging to model [PModX(PS)].	KSH01_pca_M2_01.xls (MS Excel)

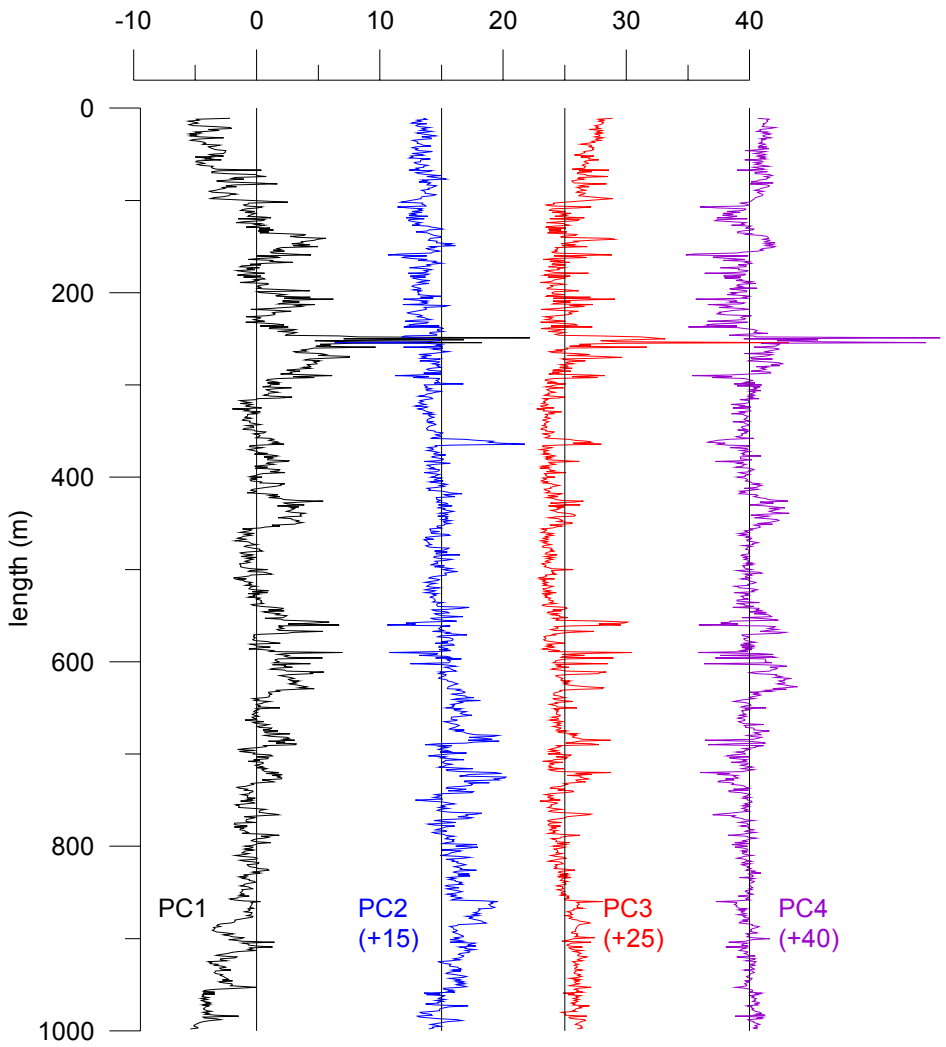


Figure 4-16. *Principal component object scores plotted versus length along the borehole. Positive scores in Pc1 indicate fracturing while negative scores indicate normal rock. Positive scores in Pc2 indicate felsic rocks whereas negative scores indicate mafic rocks. Pc3 indicates fracturing in a similar way as Pc1 but has a long wave-length trend that is reversed, probably related to liquid resistivity effects. Pc4 shows fractures unrelated to alteration and GFZI as negative scores. Constant values have been added to Pc2, Pc3 and Pc4.*

The plot in Figure 4-16 shows a long wave-length variation for both PC1 and PC3. This is most certainly due to the effect of variation in borehole fluid salinity along the borehole. The salinity affects the electrical logs. The caliper log will also introduce a trend in the PC's since the borehole diameter slightly decreases with depth. There is also a discontinuity in the caliper log where data from KSH01A and KSH01B have been merged at 100 metres length.

It might be preferable to include the fluid resistivity log in the analysis and/or perform some kind of gentle high-pass filtering of electrical and caliper logs in the pre-processing of data.

4.7 PLS-modelling of GFZI

4.7.1 Manual classification of GFZI for PLS-analysis

GFZI is a manual classification of the rock into fracture zones by taking geological and geophysical information into account. The numerical value that is assigned to each class defines the prediction achived with FZI.

A check of the coding indicates the difficulty with manual classification. A comparison between the classification of the mapped fracture frequency, Ff, for each one metre section shows great variations within each class.

GFZI		**Mapped fracture frequency,Ff**	**Number of sections**
Core of fracture zone	= 2	0 – 29 fractures/m	89
Transition zone	= 1	1 – 12 fractures /m	111
Near zone	= 0.05	0 – 9 fractures /m	60
Unaffected rock	= 0.0	0 – 19 fractures /m	725

This spread in fracture frequency might be due to the fact that the actual fracture zone borders do not coincide with the borders between one metre sections. Also, the classification is a generalization of the geology meaning that narrow fracture zones are classified as being outside any fracture zone and vice versa.

Constraints have been put on GFZI in order to make the modelling clear. Only sections that are within certain intervals of fracture frequency for each value of GFZI have been used for the modelling.

The sections that were removed from the model calculations will still be classified in the final prediction of FZI under the condition that they are within the model confidence limits.

The following constraints were put on the sections in order to include them in the model calculations:

GFZI = 2	and Ff ≥ 10	35 sections
GFZI = 1	and 3 < Ff < 10	77 sections
GFZI = 0.05	and Ff = 2 & 3	27 sections
GFZI = 0	and Ff = 0 & 1	283 sections
Total number of sections used for modelling		422 sections
Not used for modelling		563 sections

4.7.2 PLS-model for GFZI and prediction of NGFZI

The initial PLS-model is based on the selected 422 sections and all variables. Outliers are identified by analysis of DModX, the normalized distance for a section to the central part of the model (Figure 4-17).

Outliers were stepwise eliminated in order to make the dataset converge to a central homogeneous data set. 37 outliers were identified after 5 steps. This means that 385 sections remained in the four classes for the final model estimate. The distance to the model centre for the remaining sections can be seen in Figure 4-18.

The elimination of outliers had the consequence that all sections with radar reflections were removed and the variable Ra was therefore excluded from further analysis.

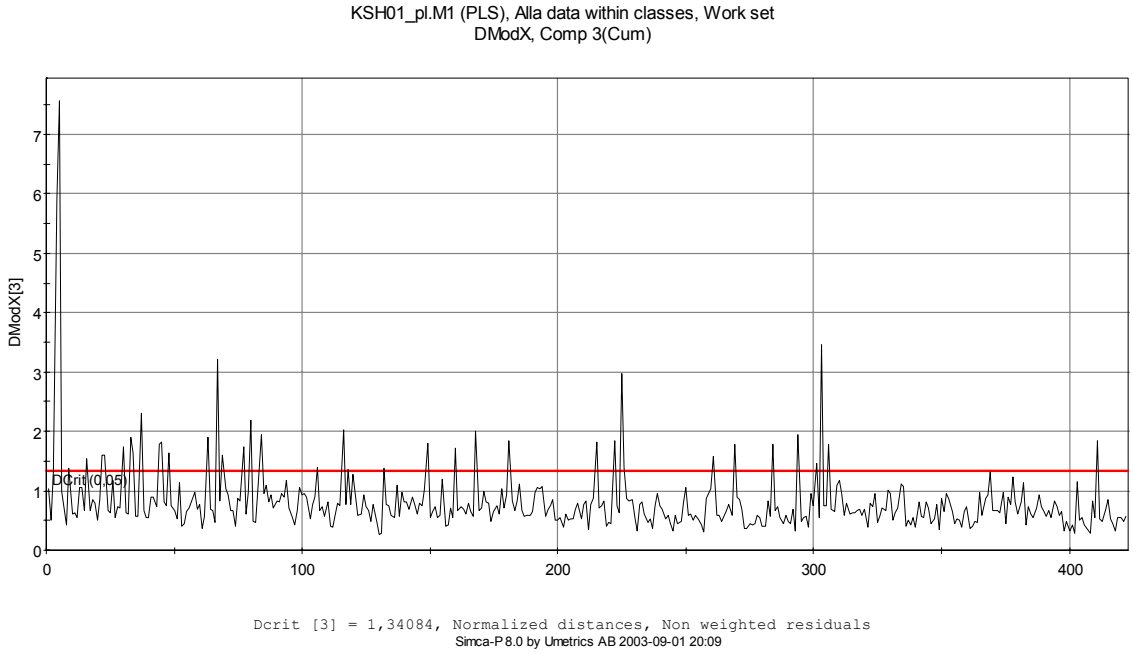


Figure 4-17. *Normalized distance to model centre for each section included in the initial PLS model estimation.*

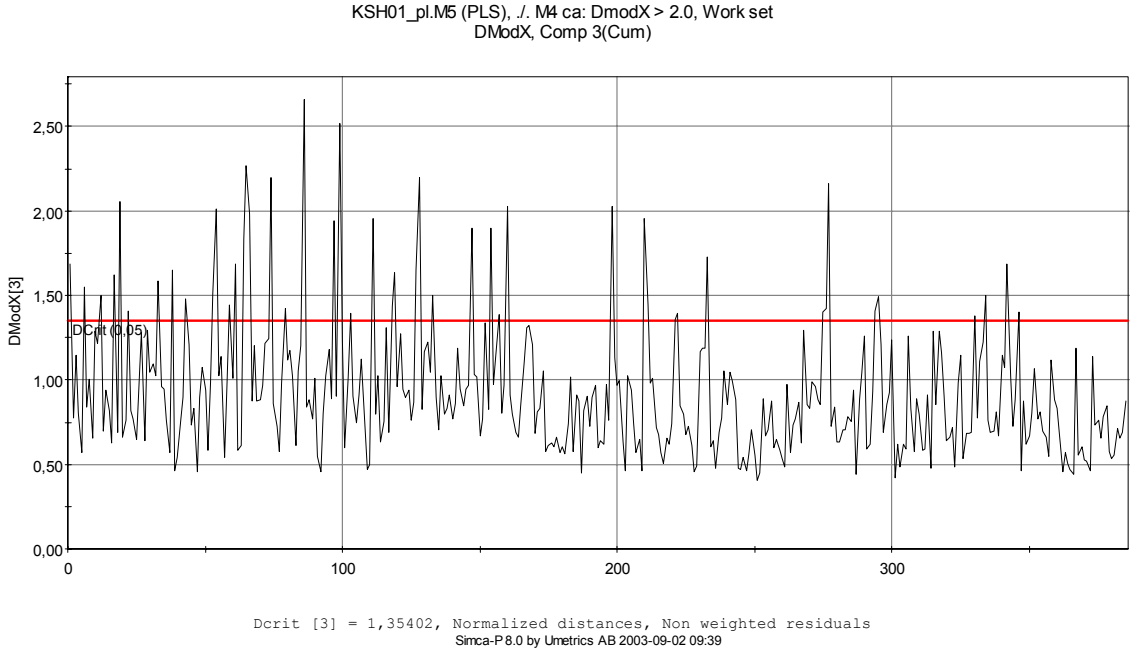


Figure 4-18. *Normalized distance to model centre after removal of outliers, model M5.*

The PLS analysis is now directed to the prediction of GFZI with the remaining variables and objects. This is done by assigning GFZI as Y-values and the other variables as X-values in a model that can be seen as a stepwise regression analysis. Two data blocks are defined in this way, X and Y. For each added PLS-component the amount of explained variation in the two blocks are given, i.e. how much of the variation in the X-block is used to explain a certain amount of variation in the Y-block.

The analysis gave three significant components after cross-validation. The model is called M5 in the discussion, tables and figures below:

PLS-M5	**X**	**Y**	**accum. Y**
Comp.1	34.7 %	50.7 %	50.7 %
Comp.2	6.2 %	32.5 %	83.2 %
Comp.3	8.7 %	3.0 %	86.2 %

With the first component 34,7 % of the variation in X is used to explain 50.7 % of the variation in Y (GFZI). For component 2, 6.2 % of the variation in X is used to explain 32.5 % of the variation in Y. Finally, with the third component 8.7 % of the variation in X is used to explain 3.0 % of the variation in Y.

A total of 49.6 % of the variation in X was hence used to explain 86.2 % of the variation in Y with the three components. The remaining variation in Y can be considered as the noise in the difference between the manual classification and a numerically continuous model.

Two types of graphs are shown below (Figures 4-19 to 4-24) where the importance of different variables in the X-block for estimation of GFZI for the three components. Figures 4-19 and 4-20 show the coefficients of the variables for the first component. Positive values for variables to the right of the origin will increase the estimated value of GFZI whereas the opposite applies for variables to the left of the origin.

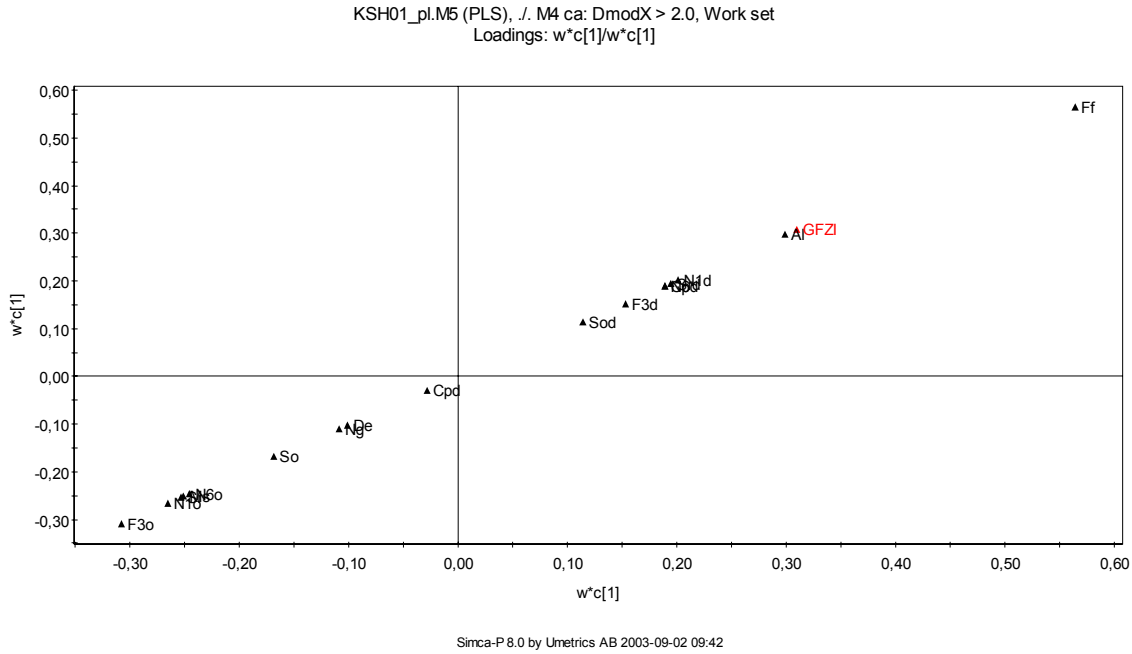


Figure 4-19. *Model PLS-M5, 1st component. The influence of variables on GFZI.*

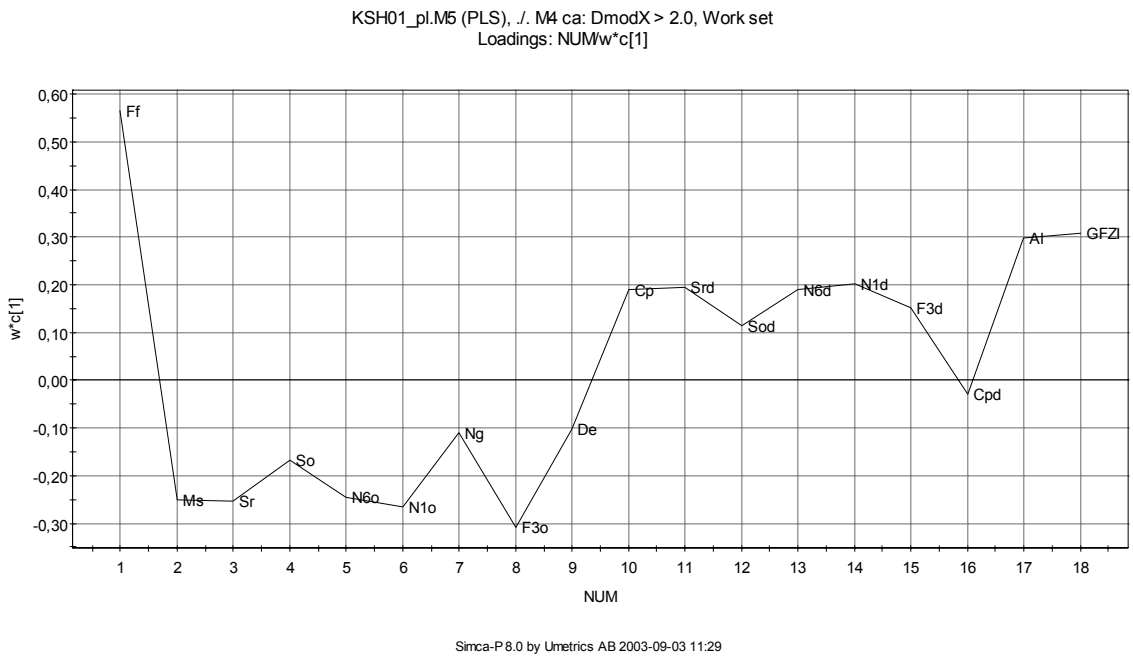


Figure 4-20. *Model PLS-M5, 1st component. The coefficients of the variables.*

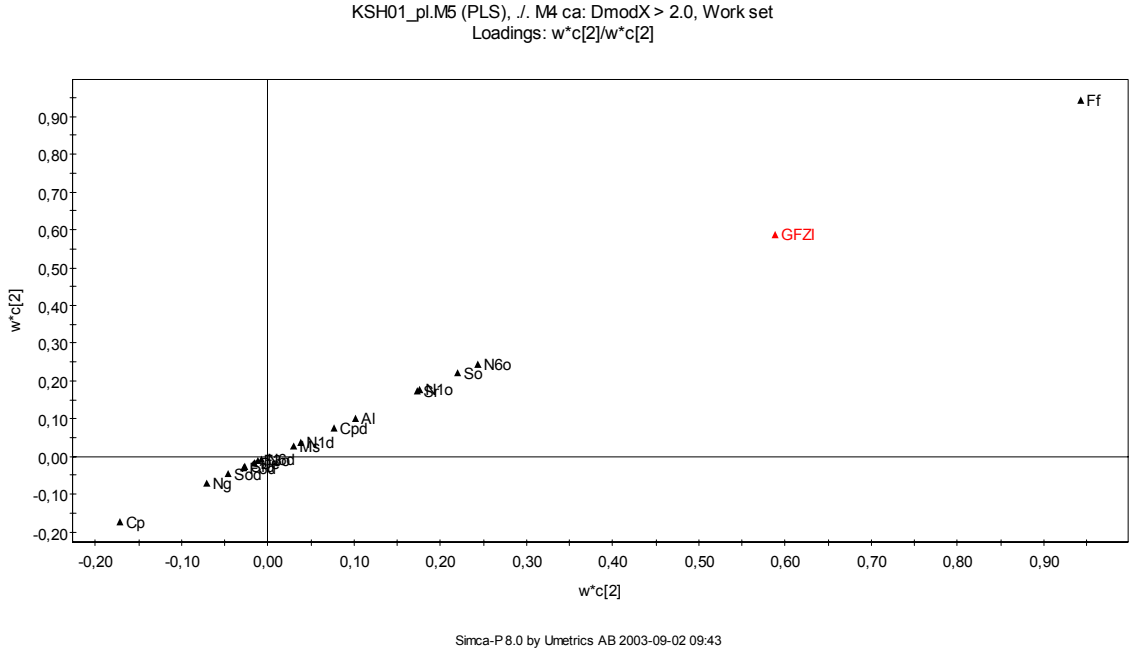


Figure 4-21. *Model PLS-M5, 2nd component. The influence of variables on GFZI.*

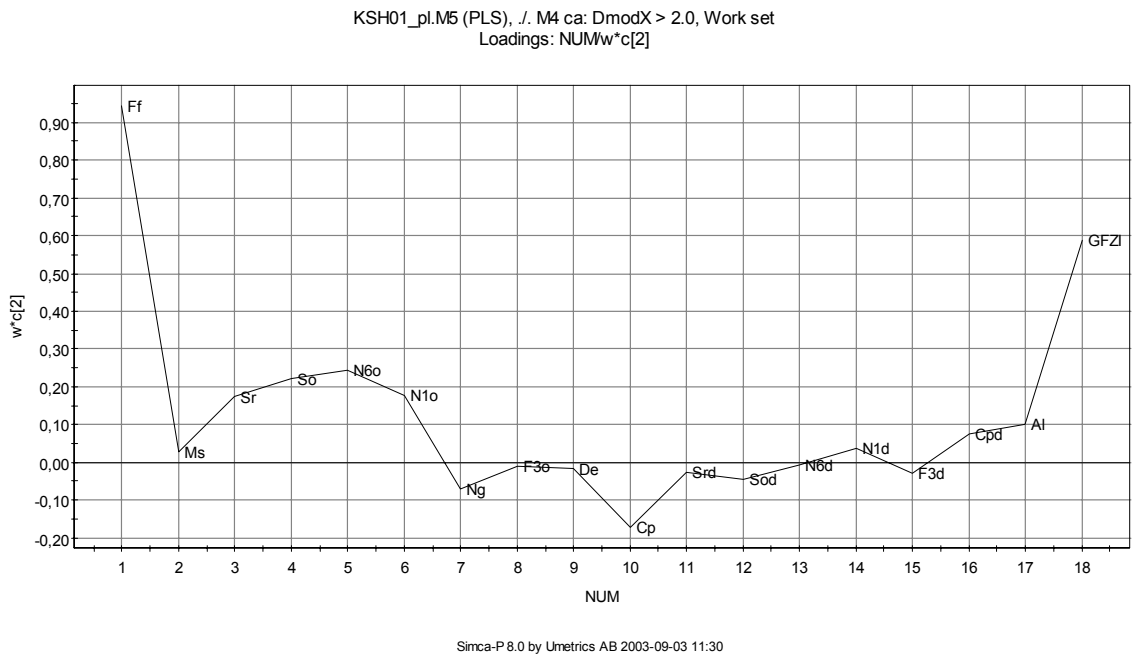


Figure 4-22. *Model PLS-M5, 2nd component. The coefficients of the variables.*

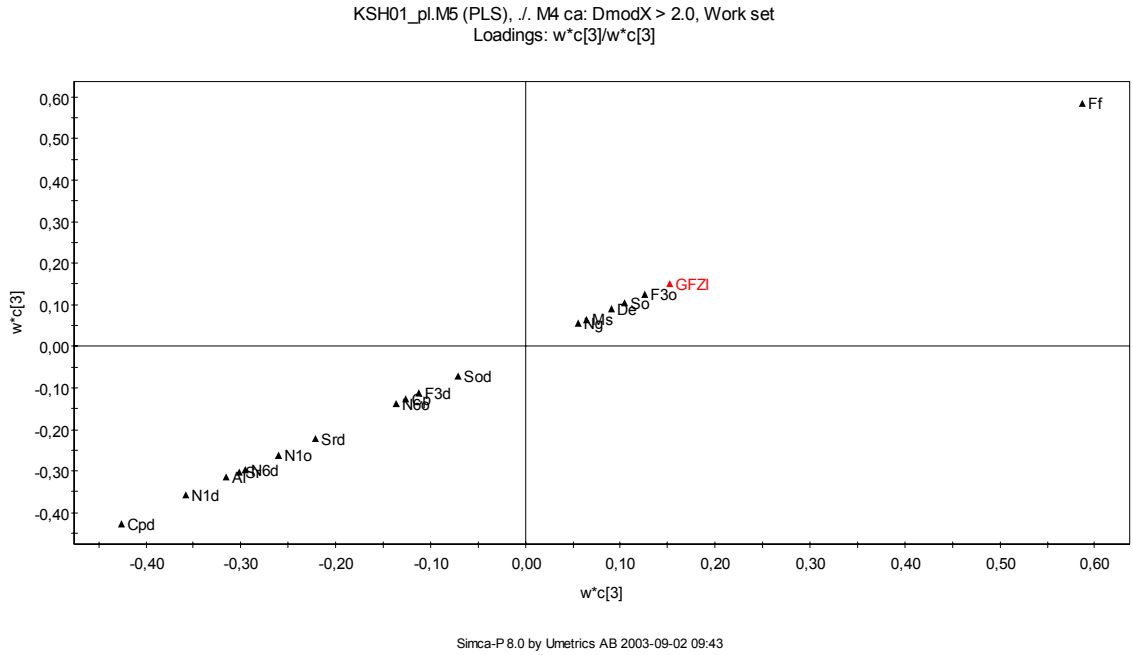


Figure 4-23. *Model PLS-M5, 3rd component. The influence of variables on GFZI.*

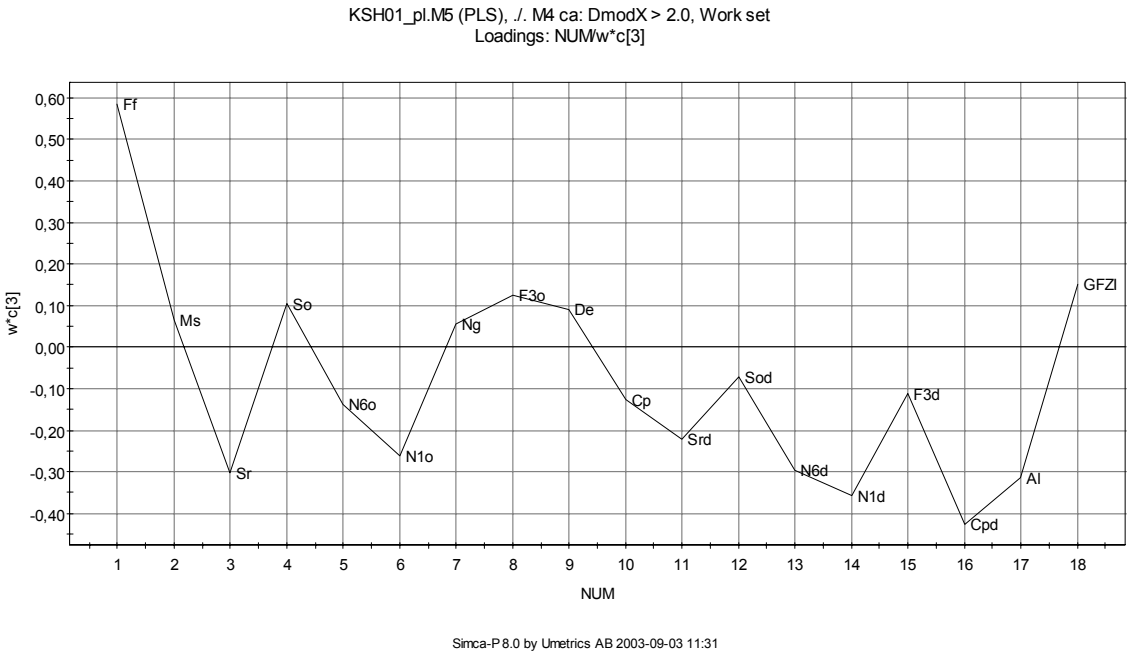


Figure 4-24. *Model PLS-M5, 3rd component. The coefficients of the variables.*

The components are interpreted in the following way:

PLS-component 1 – X: 34.7 % – Y: 50.7 %

The variables Ff and Al are most positively correlated with GFZI and the strongest negative correlation is seen for F3o. In detail, Cp, Srd, Sod, N6d, N1d, F3d are positively correlated with GFZI whereas F3o, Ms, Sr, So, N6o, N1o, Ng and De shows a negative correlation. The variable Cpd is uncorrelated with GFZI for this component.

The interpretation of this component is that it describes sections with high fracturing as well as significant alteration. These properties are also reflected to a greater or lesser extent by most geophysical logs. The negative coefficients for the electrical logs have resulted in a long wave-length anomaly in this component (Figure 4-25) due to variations in the salinity of the borehole fluid.

PLS component 2 – X: 6.2 % – Y: 32.5 %

The variable Ff is strongly correlated with GFZI . Some positive correlation is seen with Sr, So, N6o, N1o and a weak negative correlation with Cp.

The interpretation is that this component indicates high fracturing that is not affected by other variables including alteration. Apart from a long wave-length trend of the same kind as mentioned above there is a discontinuity at 100 metres length along the hole due to a shift in caliper values. Data from KSH01B have been used above 100 m length whereas data from KSH01A have been used below 100 m length.

PLS component 3 – X: 8.7 % – Y: 3.0 %

The variable Ff again shows the strongest correlation with GFZI. Deconvolved geophysical variables that are supposed to indicate fracturing, Srd, N6d, N1d, Cpd and also alteration, Al, show negative correlation with GFZI. The component is not of great magnitude but significant.

The interpretation is that negative scores in this component might indicate alteration associated with fractures but not high fracture frequency.

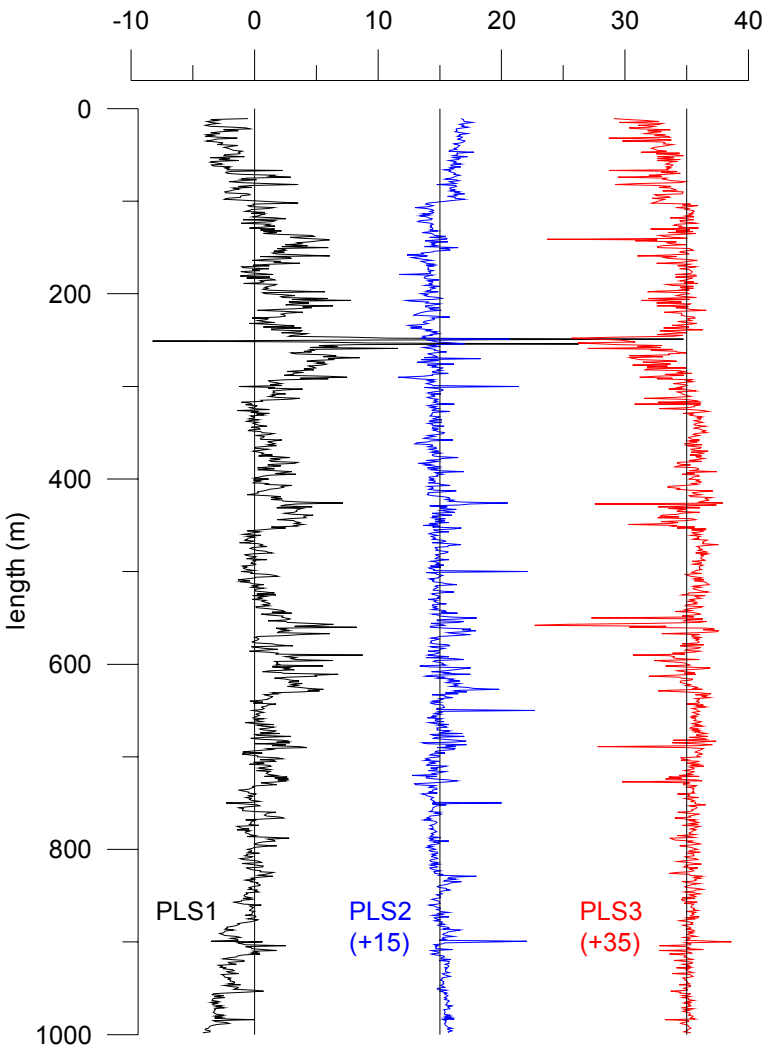


Figure 4-25. *Object scores for the three significant PLS-components of model M5 plotted against length along the hole. Constant values have been added to PLS2 and PLS3. See text for interpretation.*

4.7.3 PLS-model for GFZI without fracture frequency and alteration

An attempt was made to calculate NGFZI without using the two variables Ff and Al that relies on manual mapping of the drillcore as X-variables. The new model is called M6 below. The same data set was used as for model M5 and DModX is shown in Figure 4-26.

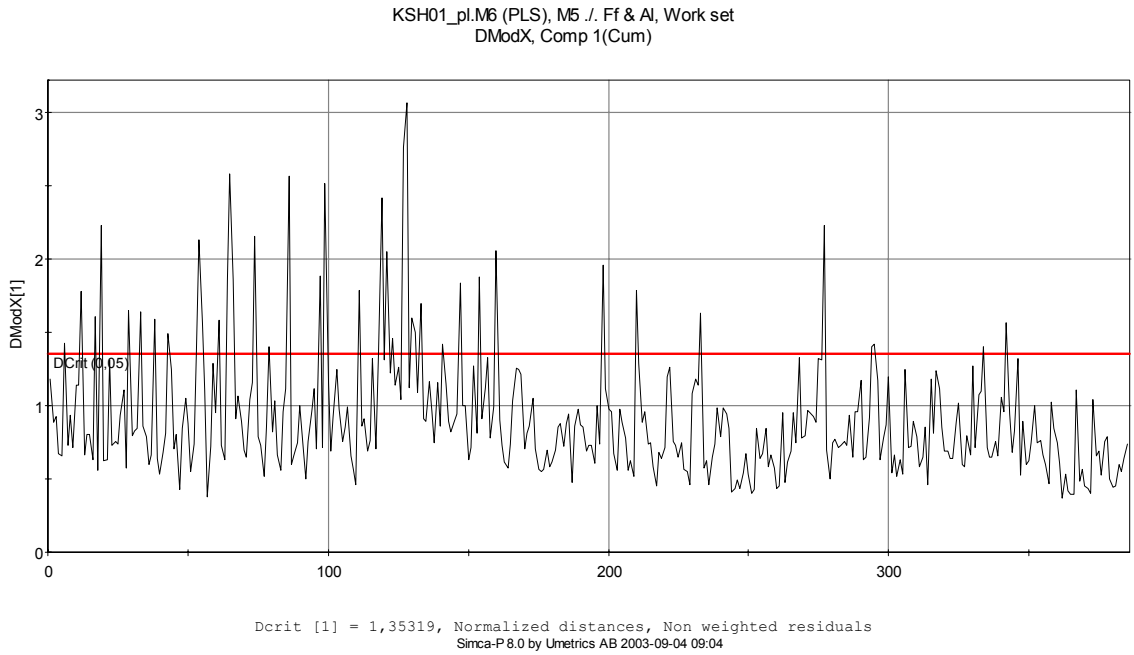


Figure 4-26. *Model PLS-M6, normalized distance to model centre.*

The analysis did only reveal one significant component after cross-validation:

PLS-M6	**X**	**Y**	**accum. Y**
Comp. 1	35.3 %	30.9 %	30.9 %

The influence of the variables shows similarity with the first component for model M5. The variation in GFZI that was explained was however 19,8 % less than for the first component of M5. The selection of sections to be included in the modelling was however partly based on their Ff-values which to some extent might explain this difference between the models. Since the second component of M5 was a “pure” Ff-component it was not possible to find any corresponding component in F6 since Ff was excluded. Figures 4-27 and 4-28 show the influence of the variables on the first component of model M6.

The first component of model M6 is given the following interpretation:

The variables show the same coefficient pattern as when Ff and Al was part of the analysis. The model shows variation in fracturing. The difference from model M5 is that manual core mapping was not used. Another difference from M5 is that there is only one significant component in the model.

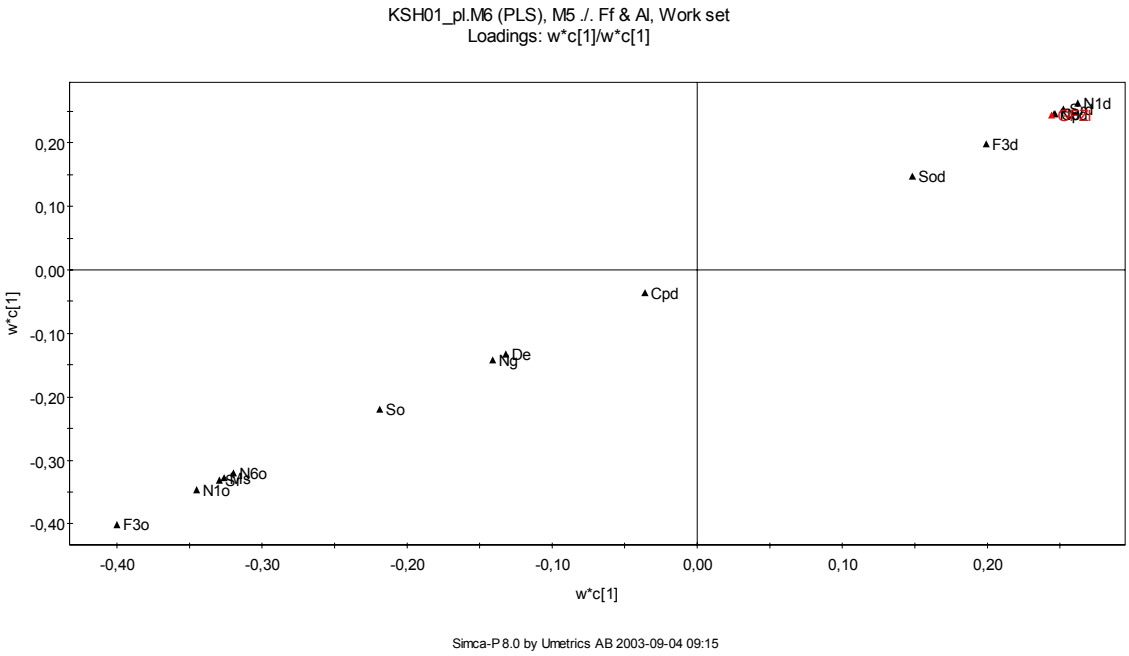


Figure 4-27. *Model PLS-M6 1st component. The influence of variables on GFZI.*

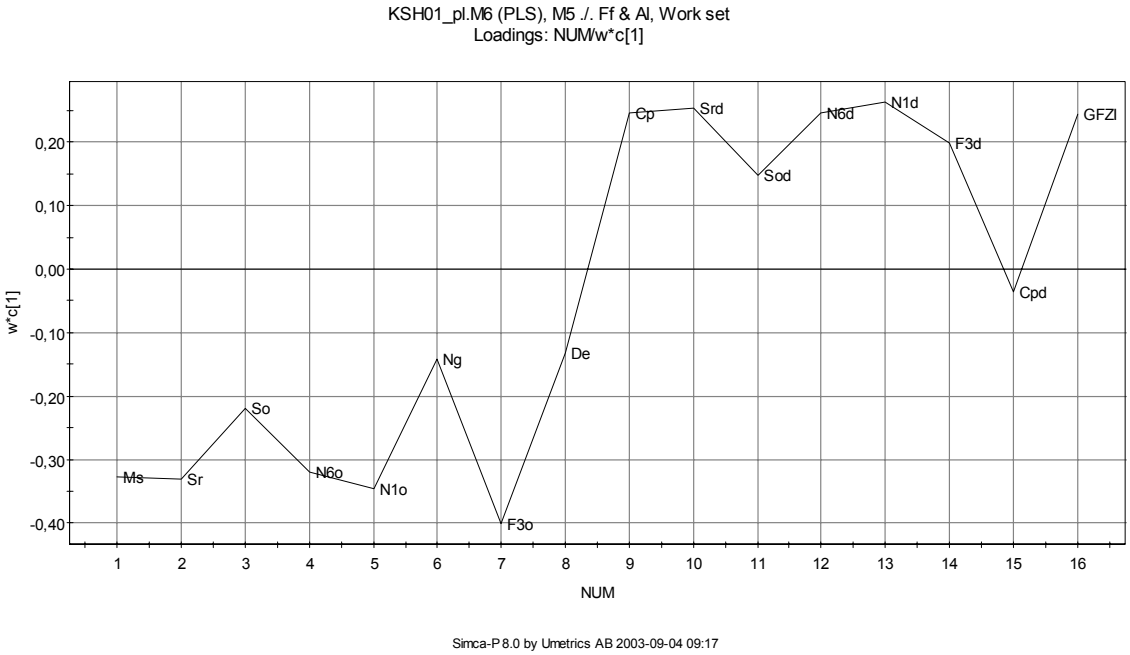


Figure 4-28. *Model PLS-M6 1st component. Coefficients of variables.*

4.7.4 Residual between observed GFZI and predicted NGFZI

A Fracture Zone Index, FZI, has been calculated with the model M5 that consists of three components. This value is here called Numerical Geological Fracture Zone Index, NGFZI. The result from the prediction of NGFZI has been analyzed in the form of a normal probability plot for the residual between the observed value of GFZI and the predicted value of NGFZI. Figure 4-29 shows two extreme values for Id 1426 and 1627 in the lower left of the graph. The remaining residual values fall along a fairly straight line, which indicates a normal distribution of the residuals.

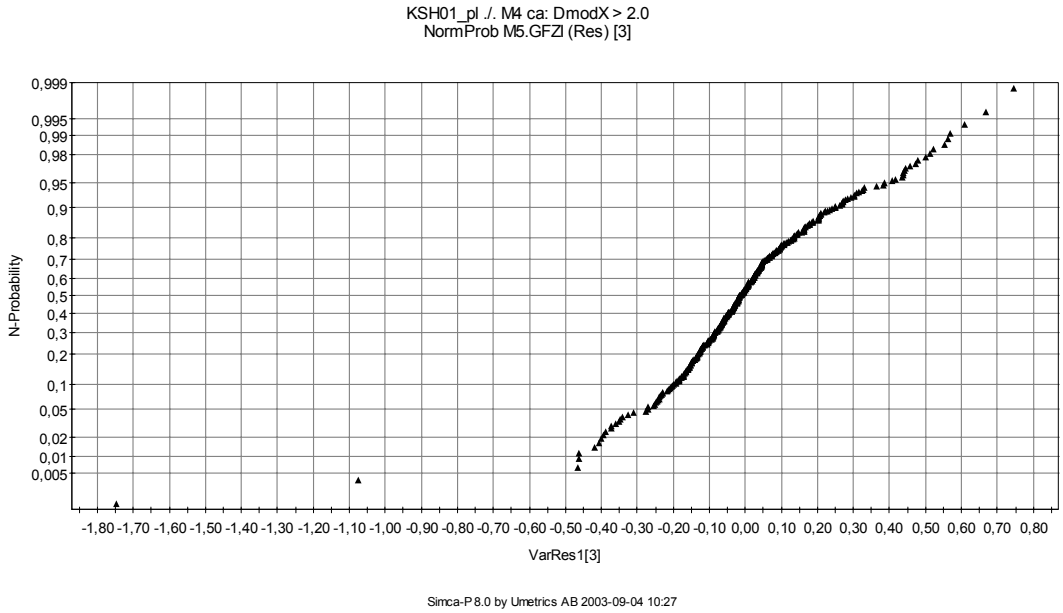


Figure 4-29. *Normal probability plot for the residual from model M5, $Y_{obs} - Y_{pred}$, ($GFZI_{obs} - NGFZI_{pred}$)*

4.7.5 Conclusions for PLS-modelling of NGFZI

The general conclusion is that only 35 % of the variation in GFZI can be explained when the variables Ff and Al were removed from the analysis. By using Ff and Al as subjective observations on the drillcore the explained variation in GFZI was increased to around 50 % for the first component. Two additional components became significant when Ff and Al were included in the analysis and the variation of GFZI that is explained by the model has increased to 86 %. The explained variation of Y of 86 % indicates a strong and robust model and the remaining 14 % of the variation in GFZI is probably noise due to generalization in the manual classification of GFZI. This noise will be filtered away by the PLS-model automatically.

This NGFZI was calculated for all sections along the borehole and describes the subdivision of the rock into classes of different fracturing with a continuous value from the core of a fracture zone to unaffected normal rock. This value is FZI after adjustments for intervals and outliers.

4.8 Calculation of FZI

NGFZI is a continuous variable varying between –0.5 and +3.5 along the borehole. Some minor adjustments have been done since the model is not aware of the limitations on the index that was predicted.

Sections that are outside the confidence limits of the model according to the variable, DModX, have been assigned a "missing value" during the analysis. In order to avoid missing predictions in the final FZI vector, these values have been replaced by the original GFZI value for the corresponding section.

The final NGFZI value has been truncated so that the minimum value is 0.0 and the maximum is 2.5. This constitutes the final continuous value of FZI. The index can be split into discrete classes according to e.g.:

Core of fracture zone	$1.5 < FZI < 2.5$
Transition zone	$0.5 < FZI < 1.5$
Unaffected rock	$0.0 < FZI < 0.5$

The PLS-modelling and FZI-calculation is summarized in Table 4-3. The result of FZI calculation is shown in Figure 4-30.

Table 4-3. PLS-analysis and FZI-calculation.

Processed primary data	Processing	Resulting data file
KSH01alla_var.xls	Creation of initial PLS-model	KSH01_pls_01.xls (MS Excel)
	Removal of outliers.	
	Creation of PLS-models M5 and M6	
	Calculation of NGFZI and DModX	
	Back substitution of GFZI-values for outliers (M5)	
	Truncation to 0<FZI<2.5 (M5)	

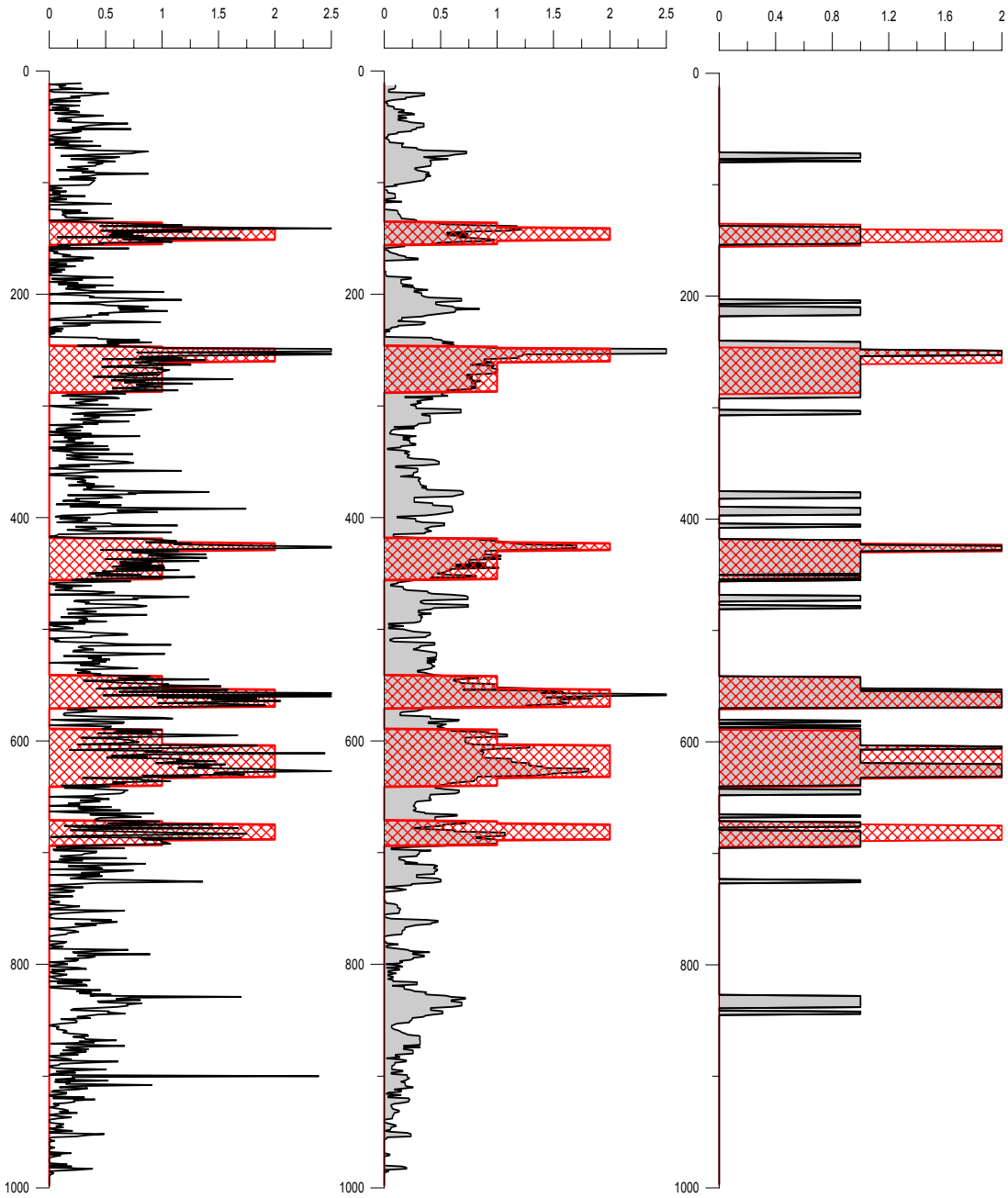


Figure 4-30. *Result of the FZI calculation. The left graph shows the predicted FZI as a black line and the original classified GFZI with red cross-hatching versus length along the borehole. The central graph shows GFZI and FZI in grey after smoothing with a 5 point median filter. The right graph shows FZI in grey as discrete classes according to the intervals: FZI<0.5 → discrete FZI=0; 0.5<FZI<1.25 → discrete FZI=1; FZI>1.25 → discrete FZI=2. Note that the border between the two highest classes was set to 1.25 in this presentation instead of 1.5 since this choice gave a slighter better agreement with GFZI.*

4.9 PC-analysis of the classes in NGFZI

NGFZI was analyzed by principal components to give some insight into the behaviour of different variables within the classes. The number of sections that were predicted in each class was:

	NGFZI	Number of sections
Core of fracture zone	1.5 – 2.5	38
Transition zone	0.5 – 1.5	273
Unaffected rock	0.0 – 0.5	674

4.9.1 PC-analysis of the class, fracture zone cores

The class for the core of fracture zone contained 38 sections. Four of these were outliers that were removed in order to make a reliable PC-analysis of the class. The model contained one significant component that explained 30.1 % of the total variation.

The plot in Figure 4-31 shows the variable loadings for Pc1/Pc2. Note that the second Pc is not significant. The first Pc shows high values for NGFZI to the left in the plot together with other parameters that are indicative for the core of a fracture zone.

It can be noted that Ff is to the right of the origin in the plot. This indicates that Ff, although important in the creation of the PLS-model, is not a strong variable in this context and even negatively correlated with NGFZI. Strong variables are: Cp and Al to the left in the plot and N1o and Sr to the right in the plot. These variables will indicate variations within the core of the fracture zone.

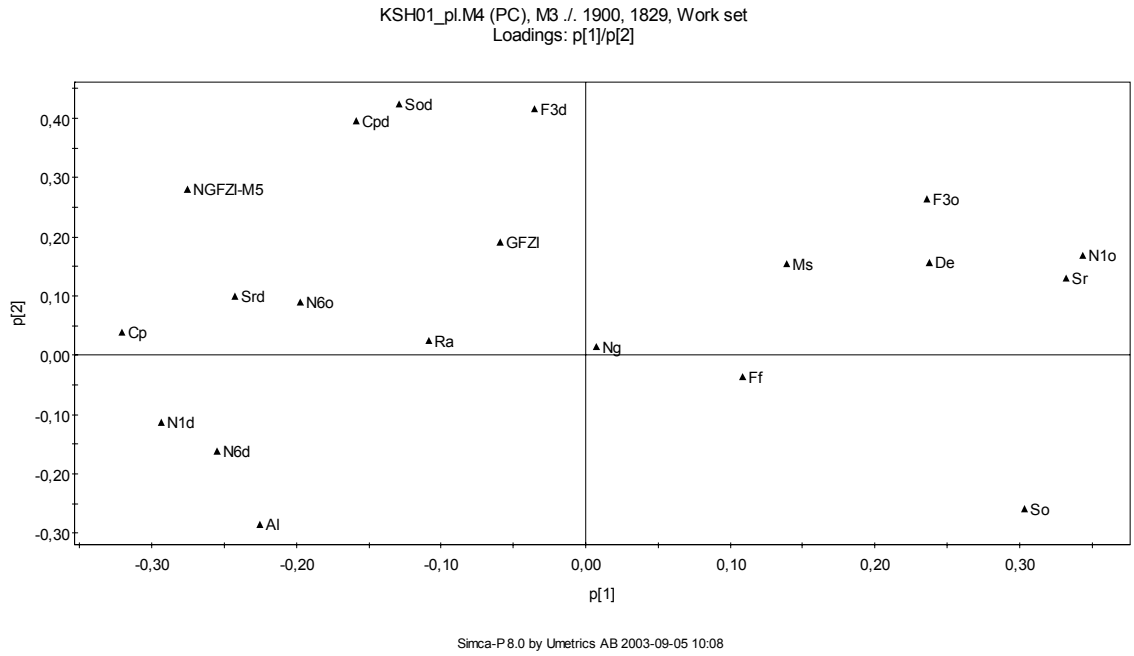


Figure 4-31. *Variable loadings for Pc1 and Pc2 for sections with NGFZI>1.5. Note that Pc2 is not significant.*

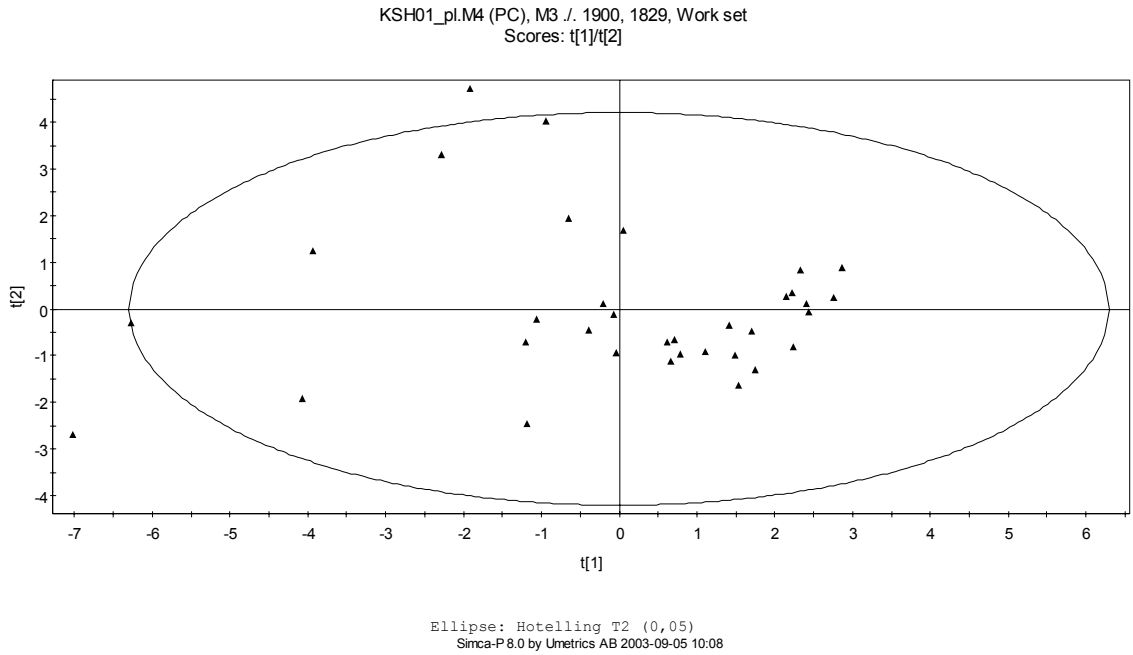


Figure 4-32. *Object scores for Pc1 and Pc2 for sections with NGFZI>1.5. Note that Pc2 is not significant.*

Since the class was based on 34 sections only, one should be somewhat cautious about the interpretation of these results and their applicability to e.g. other boreholes.

Observe that the above mentioned results are valid for the core of a fracture zone and not for sections outside the core of the fracture zone.

4.9.2 PC-analysis of the class, transition zones

The class for transition zones contained 273 sections with values for NGFZI between 0.5 and 1.5. The model showed 4 significant components describing 56.7 % of the total variation.

The results for Pc1 and Pc2 are shown in Figures 4-33 and 4-34. High values for Ff and Al will result in high values of NGFZI since they are close in the plot and thus strongly correlated. Apart from Ra and Ng all other variables will also contribute to NGFZI through either positive or negative correlation.

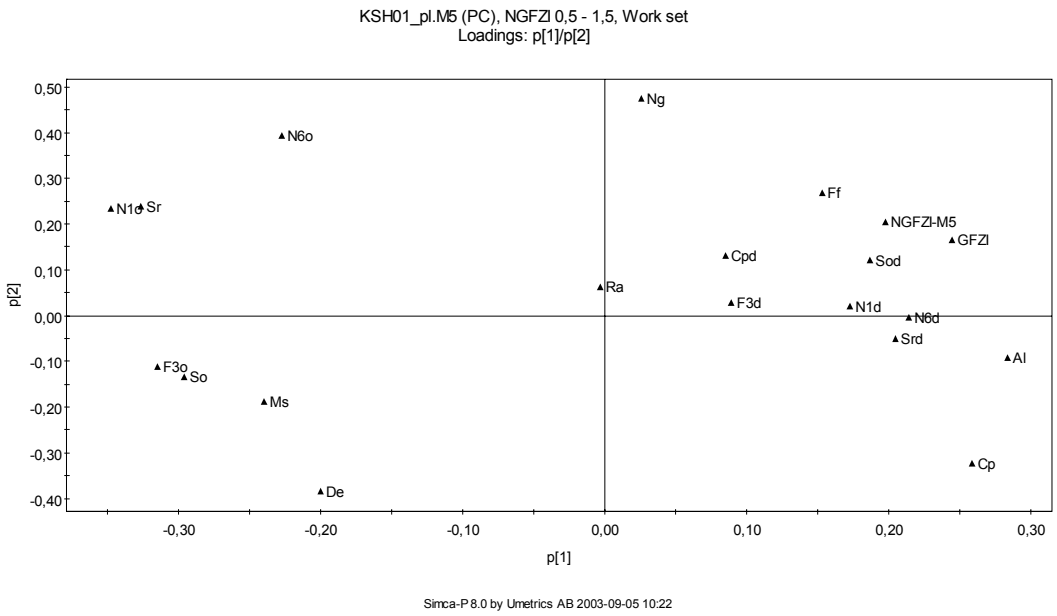


Figure 4-33. *Variable loadings for Pc1 and Pc2 for sections with 1.5>NGFZI>0.5.*

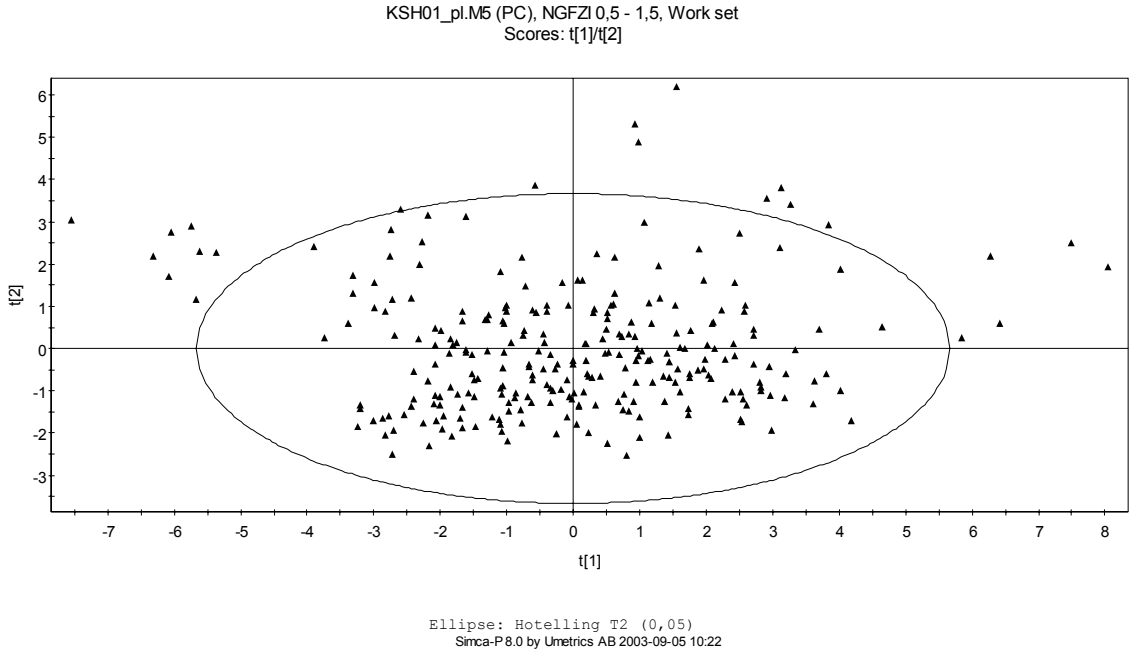


Figure 4-34. *Object scores for Pc1 and Pc2 for sections with 1.5>NGFZI>0.5.*

4.9.3 PC-analysis of the class, unaffected normal rock

The class for unaffected rock contained 674 sections with values of NGFZI less than 0.5. The model gave 6 significant components and a total of 66.5 % of the total variation was explained.

The results for Pc1 and Pc2 are shown in Figures 4-35 and 4-36. The results are similar and the interpretation is the same as for the transition zone class.

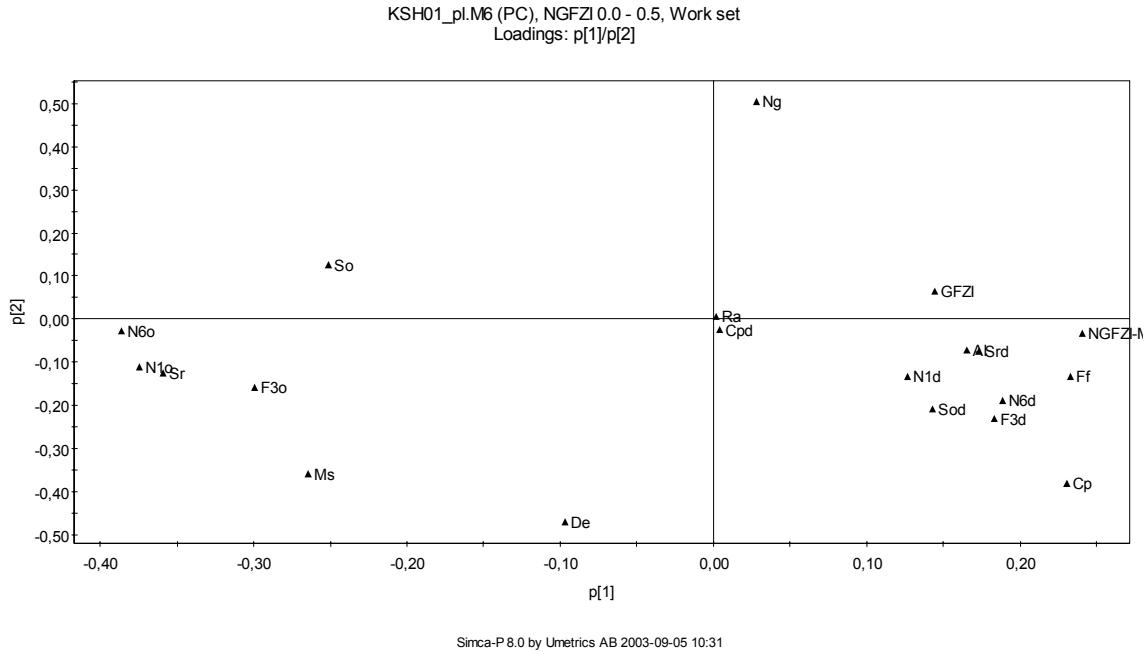


Figure 4-35. *Variable loadings for Pc1 and Pc2 for sections with 0.5>NGFZI.*

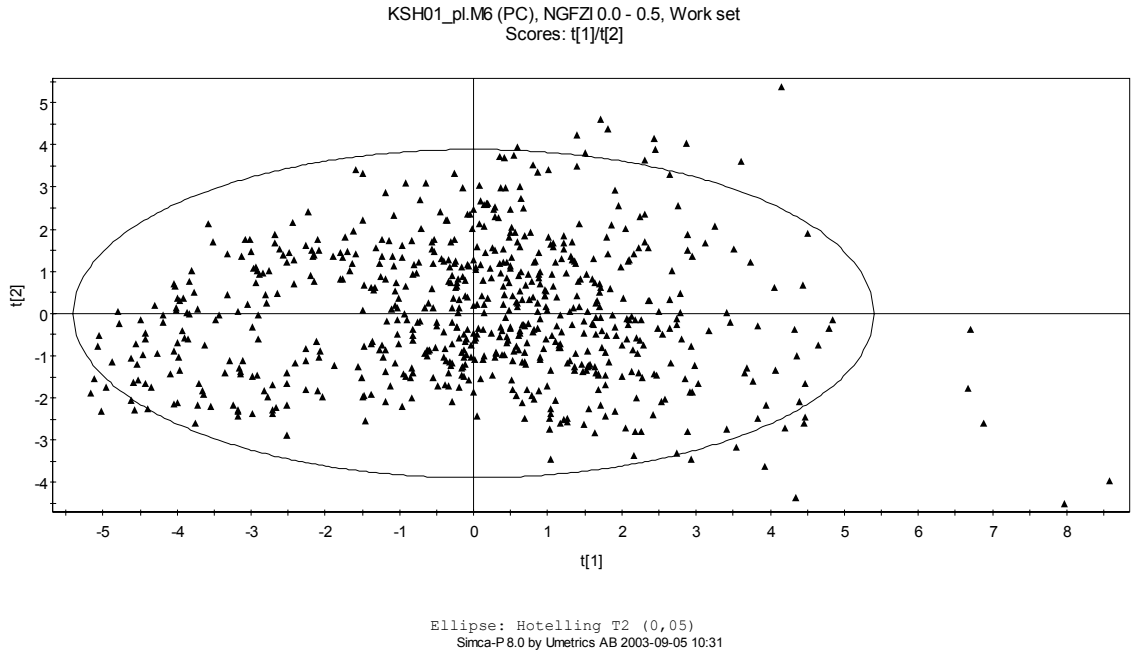


Figure 4-36. *Object scores for Pc1 and Pc2 for sections with 0.5>NGFZI.*

4.9.4 PC-analysis for subsets of all three classes

An analysis was performed on subsets of all three classes. The existing 34 sections from the core of fracture zones were used together with 34 sections from transition zones with NGFZI≈1.0 and 34 sections from unaffected normal rock with lowest values, NGFZI≈0.0.

The analysis resulted in one significant component that describes the relation between fracture zones and unaffected rock. The component is shown in Figures 4-37 and 4-38 together with the insignificant second component.

In the object scores plot in Figure 4-38, that shows the scores for every borehole section, it can be seen that the sections with NGFZI≈0 plot as a separated group whereas the other two subsets form a continuous pattern.

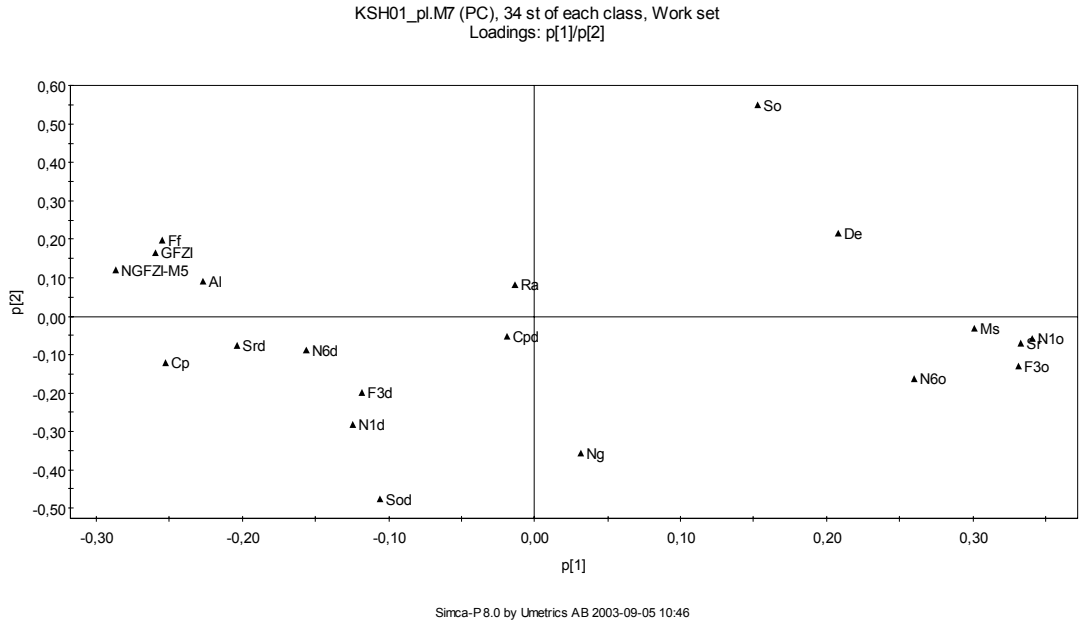


Figure 4-37. *Variable loadings for Pc1 and Pc2 for subsets of data representing the three FZI classes. Pc2 is not significant.*

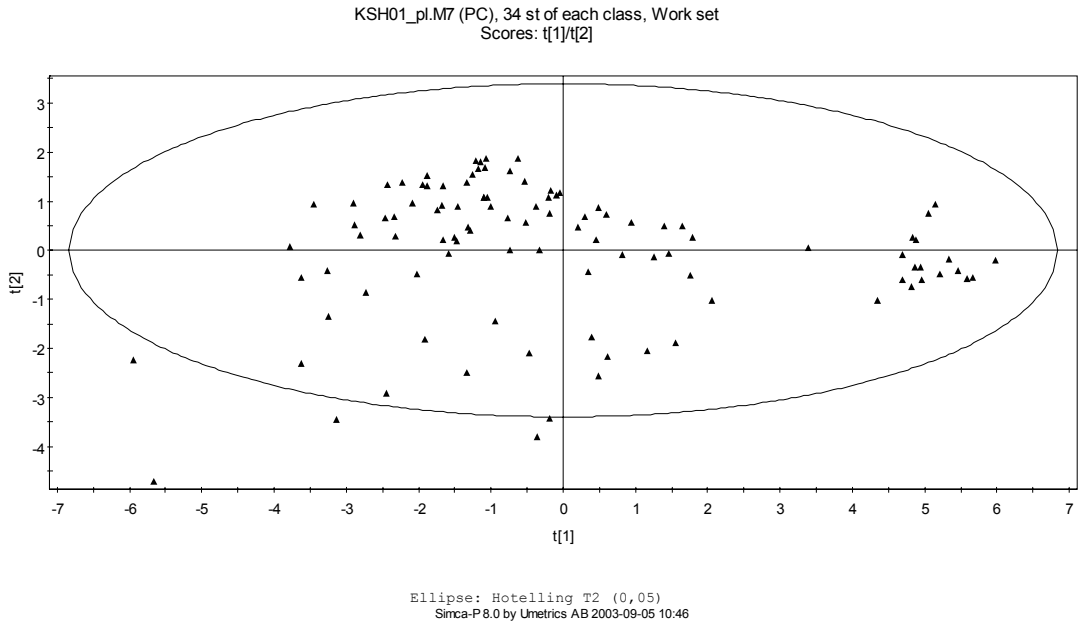


Figure 4-38. *Object scores for Pc1 and Pc2 for subsets of data representing the three FZI classes. Pc2 is not significant.*

4.9.5 Conclusion of PC-analysis of fracture zone classes

The Principal Component analysis of the FZI classes shows a similar pattern for the input variables as the original PLS-analysis that generated the NGFZI values. Inside the core of a fracture zone there is a tendency that fracture frequency is not the most indicative variable as one might believe. Some of the geophysical variables and the alteration classification are the most significant variables within the core of a fracture zone.

5 Summary and discussion

Measured and mapped variables along the borehole KSH01A have been evaluated with multivariate techniques with the purpose of calculating a Fracture Zone Index, FZI. This index should subdivide the rock into classes with information that supports interpretation of deformation zones.

By using multivariate techniques several variables can be considered simultaneously and only relevant correlated information from the variables are used for calculation of FZI. This gives a robust estimate and random and manual operator introduced noise not correlated with FZI will automatically be filtered away.

The models are based on a definition where the rock has been manually classified into three classes – core of fracture zone, transition zone and unaffected rock. Apart from this manual classification an additional constraint has been put on the sections to be included in the modelling between classes property, namely constraints in the mapped fracture frequency. The sections that constitute a class describe the properties of that class and can be regarded as the "fingerprint" of the class. Similarity to this fingerprint is then calculated in the final classification of the rock. This manual subdivision of the rock into classes is called Geological Fracture Zone Index, GFZI.

It is important that the input variables describe the entire range from the core of fracture zone to unaffected rock since this range is described by the model. The models will therefore contain variables that are not just indicative of fracture zones but also of unaffected rock. With the help of the model a Numerical Geological Fracture Zone Index, NGFZI, is calculated for all sections along the borehole, including the core of fracture zones, transition zones and the unaffected rock.

The result of FZI calculation is shown in Figure 5-1.

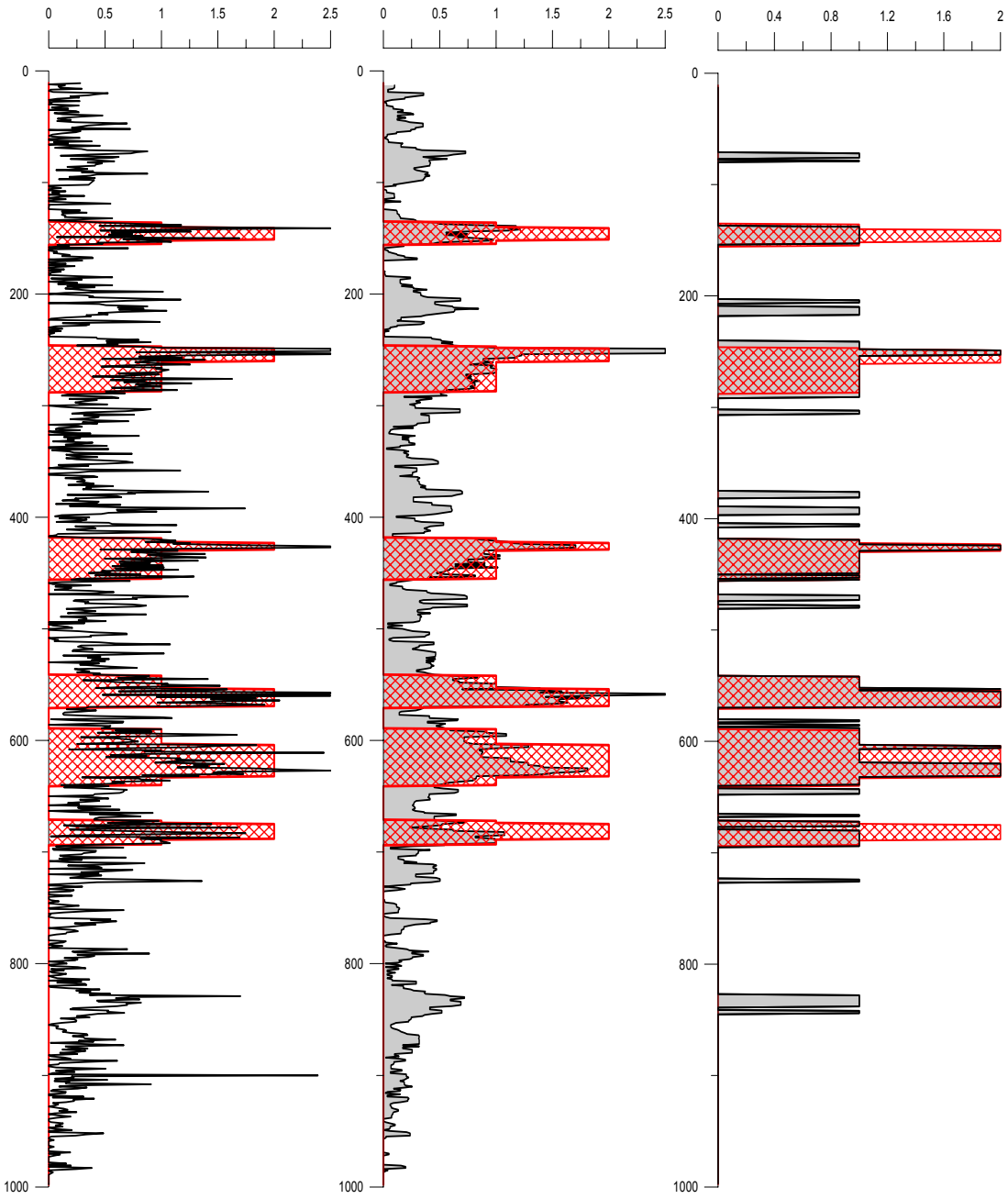


Figure 5-1. *Result of the FZI calculation. The left graph shows the predicted FZI as a black line and the original classified GFZI with red cross-hatching versus length along the borehole. The central graph shows GFZI and FZI in grey after smoothing with a 5 point median filter. The right graph shows FZI in grey as discrete classes according to the intervals: FZI<0.5 → discrete FZI=0; 0.5<FZI<1.25 → discrete FZI=1; FZI>1.25 → discrete FZI=2. Note that the border between the two highest classes was set to 1.25 in this presentation instead of 1.5 since this choice gave a slighter better agreement with GFZI.*

In order to evaluate the two variables that depends upon manual mapping of the drill core, fracture frequency and alteration, these were excluded and an analysis was performed. This had the consequence that only one PLS-component was significant and that only 35 % of the variation in GFZI could be explained. This is in contrast to the 86 % of the variation that was explained when the two variables were included.

86 % of the variation in GFZI was explained by the model and the interpretation is that the remaining 14 % mainly is noise introduced in GFZI due to generalization during the manual classification that resulted in GFZI. This noise is filtered away by the model.

The analysis of the classes for unaffected rock and transition zones shows that they mainly are correlated with an increase in fracture frequency and alteration. Inside the core of a fracture zone other variables become more important, like electrical logs, sonic and caliper.

Multivariate techniques give an objective classification of the rock as a continuous variable. Error, random and operator introduced noise in the input data is to a great extent eliminated and the resulting FZI becomes robust and repeatable. This work also shows the importance of proper pre-processing of the data. Long wave-length trends in the electrical logs and in the caliper log propagated into the multivariate components in an unwanted way, probably because they, by chance, to some extent correlated with GFZI. These logs are also represented by deconvolved data that do not show this effect. Before further work with FZI is carried out it is important that such long wave-length effects can be removed or alternatively it is checked if logs showing such trends can be omitted from the analysis.

The delivered data have been inserted in the database (SICADA) of SKB. The SICADA reference to the present activity is Field note No. 111.

References

/1/ **Wold S, Esbensen K, Geladi P, 1987.** Principal Component Analysis, Chemometrics and Intelligent Laboratory Systems, 2, p 37–52. Reprinted in Chemometrics Tutorials, 1990, 209–224, Elsevier Science Publishers.

/2/ **Lindqvist L, Lundholm I, 1985.** Effektivare prospektering med hjälp av dataanalys, Jernkontorets Annaler nr 4.

/3/ **Lindqvist L, Lundholm I, Nisca D, Esbensen K, Wold S, 1987.** Multivariate Geochemical Modelling and Integration with Petrophysical Data, Journal of Geochemical Exploration, 29, p 279–294, Elsevier Science Publishers.

/4/ **Esbensen K, Lindqvist L, Lundholm I, Nisca D, Wold S, 1987.** Multivariate Modelling of Geochemical and Geophysical Exploration Data, Chemometrics and Intelligent Laboratory Systems, 2, p 161–175, Elsevier Science Publishers.

/5/ **Andersson J E, Lindqvist L, 1988.** Prediction of Hydraulic Conductivity and Conductive Fracture Frequency by Multivariate Analysis of Data from the Klipperås Study Site, SKB Technical Report 89-11.

/6/ **Carlsten S, Lindqvist L, Olsson O, 1989.** Comparison between Radar Data and Geophysical, Geological and Hydrological Borehole Parameters by Multivariate Analysis of Data, SKB Technical Report 89-15.

/7/ **Black J, Olsson O, Gale J, Holmes D, 1990.** Site Characterisation and Validation – Stage 4 – Preliminary Assessment and Detail Predictions, Stripa Project 91-08, SKB Technical Report.

/8/ **Korkealaakso J, Vaittinen T, Pitkänen P, Front K, 1994.** Fracture Zone Analysis of Borehole Data in Three Crystalline Rock Sites in Finland – The principal Component Analysis Approach, Nuclear Waste Commission of Finish Power Companies, Report YJT-91-11.

Appendix 1

Multivariat analys

PCA-analys

Här följer en kort beskrivning av Principal Component Analysis utan ingående statistiska eller matematiska detaljer. För detaljer refereras till /Wold et al, 1987/.

Multivariat analys bearbetar en matris med data för ett antal objekt och ett antal variabler med syfte att skapa robustare analys än att bara tolka en variabel i taget.

Vanligtvis finns någon struktur i en datatabell, t ex när en variabel ökar, ökar även ett antal andra variabler och samtidigt minskar värdena för ytterligare några andra variabler. Vi kan säga att vi i tabellerna kan finna att vissa variabler är korrelerade och att vissa är omvänt korrelerade med varandra och att dessa korrelationer beskriver typiska egenskaper i tabellen som vi kan tolka och namnge.

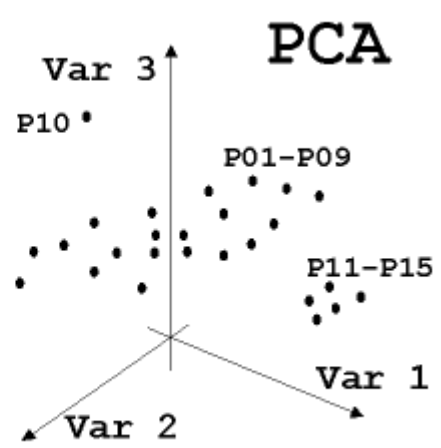
Från teori och våra erfarenheter kan vi känna igen att vissa variabler samverkar och ger oss en förståelse för hur olika egenskaper uppträder i mätvärdena. Att bara analysera en variabel i taget blir en grov och ibland allvarlig förenkling av den komplexa verkligheten där slumpmässiga variationer eller mätfel kan inverka på resultatet.

Multivariat analys ger en generell beskrivning av en datatabell i form av egenskaper som kan visas i två-dimensionella bilder. Bilderna visar extrema objekt med avvikande egenskaper samt vilka variabler och objekt som är viktiga och typiska för varje egenskap.

Ju större tabellen är desto svårare blir det att förstå innehållet och desto fler variabler tenderar att beskriva samma typ av egenskap från olika synvinklar. Genom analys av tabellerna med en multivariat metod och grafisk presentation förenklas tolkningen avsevärt. Samtidigt elimineras slumpmässiga fel när endast korrelationsstrukturen används för att beskriva en egenskap.

Nedan följer ett förenklat exempel av data och hur dessa hanteras av en multivariat analys. Figur A-1 visar en enkel tabell med en identitet och tre variabler för 15 objekt. Objekten P01–P15 kan vara våra 1-meters sektioner beskrivna med tre variabler och varje objekt får då en unik position i den tredimensionella rymden.

Obj	Var1	Var2	Var3
P01	12.4	12.9	8.2
P02	11.6	13.2	8.6
P03	12.1	14.6	8.4
P04	11.8	14.7	8.5
P05	12.5	14.9	8.2
P06	11.4	15.1	8.5
P07	12.2	15.9	8.1
P09	12.6	16.4	8.6
P10	10.2	15.9	12.8
P11	15.3	15.3	8.5
P12	15.1	15.2	8.4
P13	15.2	15.4	8.6
P14	15.1	15.3	8.5
P15	15.5	15.2	8.3

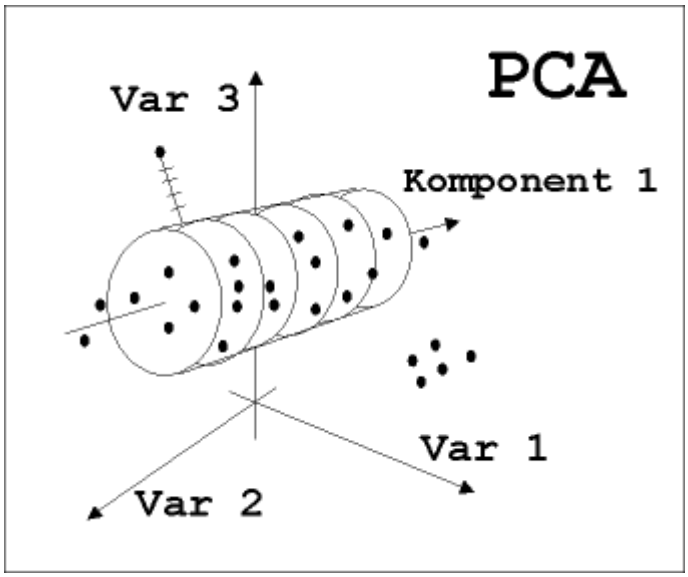


Figur A-1. *Femton objekt med tre variabler och dess position i rymden.*

Bilden visar ett extremvärde, P10 och en grupp av extremvärden, P11–P15. Dessutom finns en central grupp av objekt som visar den dominerande egenskapen i datamatrisen.

Det är dessa egenskaper vi vill kunna se grafiskt för att bättre förstå och sortera objekt och variabler i tabellen. Vi kan tänka oss att man tar den tredimensionella bilden från Figur A-1 och håller den i handen. Sedan vrider vi och vänder på bilden för att se data från olika håll.

Beroende på hur man vrider handen ser man olika trender, grupper och hur varje enskild datapunkt ligger i den tre-dimensionella rymden. På bilden i Figur A-2 har den avlånga centrala gruppen omslutits av en cylinder och tolkats som den egenskap vi vill analysera.



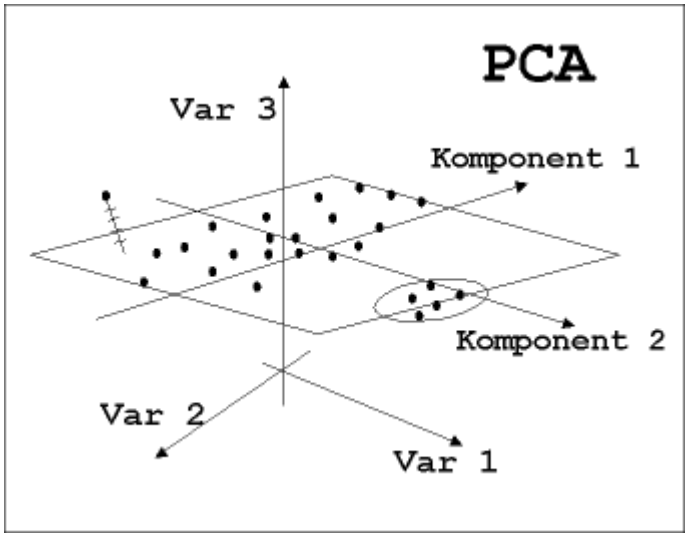
Figur A-2. *Centrala gruppen av datapunkter omsluts med en konfidensvolym.*

Idén med multivariat analys är att beräkna de riktningar som visar den största spridningen av objekten och att samtidigt projicera de ursprungliga variablerna längs dessa axlar och plan. För att göra detta använder vi vedertagna matematiska metoder.

Först beräknas den riktning som beskriver den största variationen, därefter de vinkelräta riktningarna som stegvis beskriver så stor del som möjligt av den återstående variationen tills hela matrisen har beskrivits till 100 %. Dessa riktningar kallas komponenter, egenskaper eller egenskapsriktningar och parvis kan dessa riktningar användas som projektionsplan.

Om vi anger att hela tabellen innehåller 100 % variation, kan vi också ange i procent hur stor del av variationen som varje komponent beskriver. Genom att summera dessa procenttal för de två komponenter som visas i en bild kan vi ange hur stor del av variationen som bilden visar, vilket blir en viktig del i vår analys, bedömning och förståelse av matrisen.

I Figur A-3 har de två första komponenterna lagts in i den tredimensionella rymden med tillhörande plan tillsammans med de enskilda objekten och variablerna.

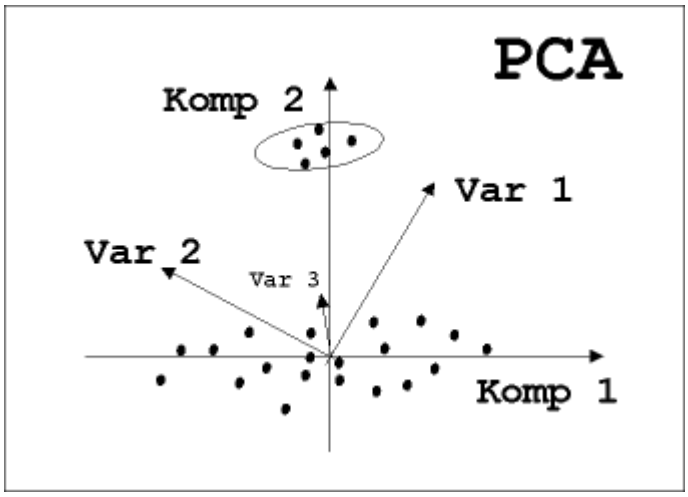


Figur A-3. *Varje punkt projiceras mot planet definierat av komponent 1 och 2.*

Vi ser även att Var 1 drar i samma riktning som komponent 2, Var 2 drar i motsatt riktning av komponent 1 och Var 3 är nästan vinkelrät mot planet och hamnar nära centrum och tillför inte någon spridning eller information till planet.

Genom att bara använda den information som finns beskriven av t ex de två första komponenterna i planet kan man eliminera residualerna och därmed eliminera dessa som icke relevant variation.

Varje tvådimensionell projektion kan sedan projiceras på en datorskärm eller skrivas ut på papper för att manuellt tolka och beskriva resultaten, Figur A-4.



Figur A-4. *Projektionsplanet kan visas som en bild på en datorskärm.*

Maximalt kan man ta fram lika många komponentriktningar som antalet ursprungliga variabler och dessa kan ses bara som ett nytt koordinatsystem vars riktningar styrs av datapunkternas spridning och påverkar inte datapunkternas läge eller inbördes avstånd.

Att använda fler än tre variabler som i detta exempel är inget problem och moderna pc-program kan hantera hundratals variabler och tusentals objekt.

Om man kan acceptera tanken att analysen endast ger en projektion av de ursprungliga objekten på ett strikt geometriskt sätt och utan att förvränga den inbördes relationen, inser man att dessa projektioner beskriver samma tabell men genom ett fönster eller projektionsplan vars riktning är baserat på flera variabler som korrelerar.

Generellt gäller att variabler eller objekt som ligger nära varandra visar stark likhet med varandra. Variabler eller objekt som ligger på vardera sidan av centrum, dvs vinkeln mellan dem är ca 180 grader genom centrum, är omvänt relaterade till varandra och således när den ena variabeln ökar så minskar den andra.

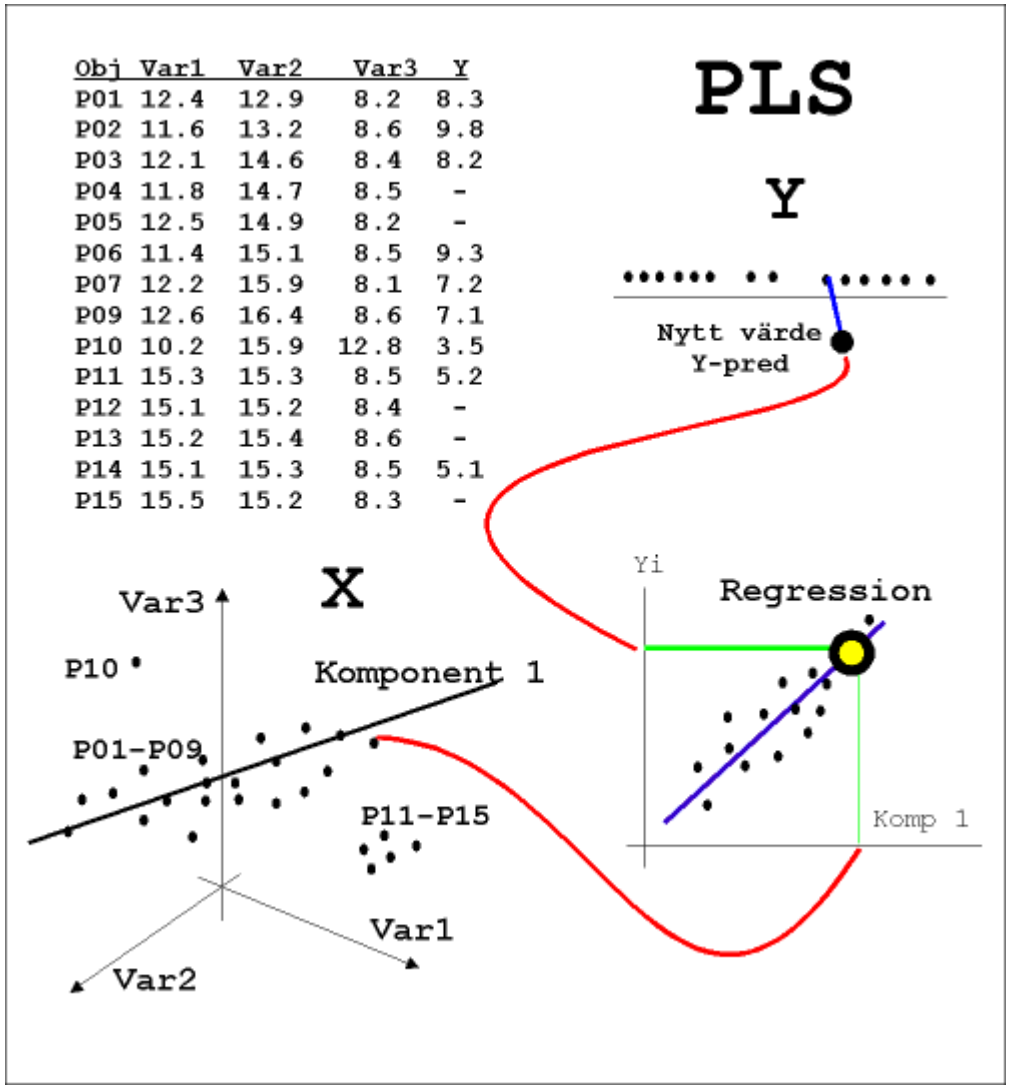
Om vinkeln mellan två objekt eller variabler genom centrum är 90 grader är variablerna oberoende och de påverkar eller samverkar inte med varandra. Om en variabel ligger nära centrum innebär detta att variabeln inte bidrar till spridningen i planet och är helt oberoende till de övriga variablerna som beskriver planet.

I en tabell ingår ofta flera variabler som beskriver likartade fenomen. Vi kan säga att ett fenomen finns beskrivet genom likvärdiga variabler. Dessa variabler hamnar nära varandra i analysen och bidrar till att beskrivningen av egenskaperna inte förändras nämnvärt om vi väljer bort en eller några av dessa likvärdiga variabler – analysen blir robust. Om vi samtidigt har eliminerat extremvärden, tillfälliga och avvikande observationer, ger detta ytterligare stöd till en objektiv och robust analys.

Genom den multivariata PCA-analysen skapas en förståelse av data samt en objektiv numerisk beskrivning av relationer mellan objekt och variabler där man har eliminerat inverkan av icke önskvärda objekt, variabler och eventuella slumpmässiga variationer.

PLS-analys

PLS-analys är en typ av regressionsanalys där man styr ett antal oberoende variabler X, att beskriva en eller flera beroende variabler Y, med syfte att kunna prediktera ett y-värde från x-värdena. För detaljer refereras till /Geladi et al, 1986/.



Figur A-5. *PLS modellering av X-variabler för prediktering av Y.*

X-variablerna från tabellen i Figur A-1 kompletteras med en Y-variabel och en första komponent beräknas med PCA-teknik för X-variablerna, Komp1. Varje objekt överförs till en ny graf där läget längs komponenten används för den horisontella axeln och värdet på Y som den vertikala axeln och en regressionslinje beräknas i den nya grafen. För varje x-värde beräknas med regressionslinjens hjälp ett nytt y-värde och metoden itererar tills y-värdet stabiliseras. Nu kan vi utläsa vilka variabler i X som visar största relationen med Y för första komponenten och hur stor del av variationen i X som används för att beskriva variationen i Y.

Om det är möjligt, skapas nästa komponent på samma sätt som för första komponenten, för den resterande variation i X, för att beskriva så stor del som möjligt av den resterande variationen i Y. Detta utförs så länge komponenterna är signifikanta och de slutliga residualerna i X och Y efter att alla signifikanta komponenter har beräknats filtreras bort från modellen.

Varje ny komponents signifikans verifieras med korsvalidering. Detta utförs genom att modelldata delas upp i sju delmängder, där en delmängd åt gången utesluts och en PLS-modell beräknas som används för prediktering av den uteslutna delmängden. Detta upprepas tills alla data element har uteslutits en gång och ett prediktionsfel beräknats för skillnaden mellan observerat och predikterat värde. En modell eller dess komponenter anses vara signifikant om prediktionsfelet är mindre än ett definierat värde.

Slutresultatet blir en regressionsformel, där man anger för varje komponent hur stor andel av variationen i X som används för att beskriva variationen i Y och bilder där man stegvis kan bedöma olika variablers och objekts inverkan på modellen tillsammans med information av typen,

Med komp. X_1	används 45 % av variationen i X för att beskriva 60 % av var. i Y
Med komp. X_2	används 15 % av variationen i X för att beskriva 20 % av var. i Y
Totalt	används 60 % av variationen i X för att beskriva 80 % av var. i Y

Regressionsformel har utseende enligt,

$$Y_{pred} = \text{constant} + AX_1 + BX_2 + CX_3 + \ldots..$$

Residual av ej beskriven variation = $Y_{obs} - Y_{pred}$

Resultaten från analysen visar automatiskt vilka X-variabler som medverkar alternativt som är omvänt korrelerade med en ökning av värdet Y, samtidigt som man ser vilka variabler som inte bidrar med någon information till beskrivning av variationen i Y.

Med den skapade relationsformeln kan man prediktera Y-värden för varje objekt, baserat på dess X-variabler även om objektet inte har något observerat y-värde.

Den slutliga modellen beskrivs även med en konfidensvolym och för varje enskilt objekt beräknas ett avstånd till modellen. Ligger ett objekt långt utanför modellen, exempelvis med ett avstånd större än 3 standardavvikelser anses detta objekt vara ett extremvärde och försiktighet skall iakttas vid användning för modellering och vid tolkningen av dess beräknade värde, Y_{pred}.

Referenser

Geladi P, Kowalski B R, 1986. Partial Least Square Regression (PLS) a tutorial, Analytical Chemistry Acta, 185, p 1–17.

Wold S, Esbensen K, Geladi P, 1987. Principal Component Analysis, Chemometrics and Intelligent Laboratory Systems, 2, p 37–52. Reprinted in Chemometrics Tutorials, 1990, 209–224, Elsevier Science Publishers.