UPTEC W10 004

Examensarbete 30 hp
April 2010

Explorative Multivariate Data Analysis of the Klinthagen Limestone Quarry Data

Linus Bergfors

ABSTRACT

Explorative Multivariate data Analysis of the Klinthagen limestone quarry data
Linus Bergfors

The existing quarry planning at Klinthagen is rough, which provides an opportunity to introduce new exciting methods to improve the quarry gain and efficiency. Nordkalk AB, active at Klinthagen, wishes to start a new quarry at a nearby location. To exploit future quarries in an efficient manner and ensure production quality, multivariate statistics may help gather important information.

In this thesis the possibilities of the multivariate statistical approaches of Principal Component Analysis (PCA) and Partial Least Squares (PLS) regression were evaluated on the Klinthagen bore data. PCA data were spatially interpolated by Kriging, which also was evaluated and compared to IDW interpolation.

Principal component analysis supplied an overview of the relations between the variables, but also visualised the problems involved when linking geophysical data to geochemical data and the inaccuracy introduced by lacking data quality.

The PLS regression further emphasised the geochemical-geophysical problems, but also showed good precision when applied to strictly geochemical data.

Spatial interpolation by Kriging did not result in significantly better approximations than the less complex control interpolation by IDW.

In order to improve the information content of the data when modelled by PCA, a more discrete sampling method would be advisable. The data quality may cause trouble, though with sample technique of today it was considered to be of less consequence.

Faced with a single geophysical component to be predicted from chemical variables further geophysical data need to complement existing data to achieve satisfying PLS models.

The stratified rock composure caused trouble when spatially interpolated. Further investigations should be performed to develop more suitable interpolation techniques.

Keywords: Multivariate analysis, interpolation, PCA, principal component analysis, PLS, projection to latent structures, partial least squares, Limestone quarry, Klinthagen, Kriging.

Department of Information Technology, Uppsala University
Box 337 SE-751 05 Uppsala Sweden

REFERAT

Utforskande multivariat analys av Klinthagentäktens projekteringsdata

Linus Bergfors

Brytningsplaneringen vid kalkbrottet Klinthagen är idag mycket grov. Detta öppnar för möjligheten att utveckla nya metoder för att effektivisera och förbättra arbetet vid brottet. Nordkalk AB som bedriver brytningen i Klinthagen vill utöka sin verksamhet med ett nytt brott i samma område av Gotland. Multivariat analys av prospekteringsdata kan bidra till att samla nyttig information, som förbättrar exploateringen av framtida objekt.

Genom analys av borrhålsdata från Klinthagen utvärderades i detta examensarbete möjligheterna med de multivariata metoderna PCA (principialkomponentsanalys) och PLS regression (partiell minstakvadrat- anpassning). Data från PCA modeller interpolerades rumsligt med Krigingmetoden, vilken jämfördes med inverterade distansmetoden (IDW).

Principialkomponentanalysen förmedlade en överblick över datat. Genom detta blev problematiken då kemiska och fysikaliska data ska sammanlänkas tydlig. Samtidigt belystes även vikten av god datakvalitet.

PLS regressionen visade goda resultat då enbart kemiska data användes. Svårigheterna att koppla ihop kemiska och fysikaliska data förtydligades ytterligare under denna del av analysen.

Vid jämförelsen mellan Kriging och IDW interpolation av Klinthagendatat kunde ingen egentlig fördel tillskrivas den mer komplexa Krigingmetoden.

Metoderna PCA och PLS kan sägas fungera för geokemiska data, men för att förbättra framtida analyser bör en mer diskret datainsamlingsmetod tillämpas. Den periodvis låga datakvaliteten, förmodligen beroende på den långa insamlingsperioden orsakar även den vissa problem.

Det krävs mer än enbart geokemiska data då den fysikaliska parametern, termiskt sönderfall ska predikteras med PLS regression. Kompletterande fysikaliska data som till exempel kornstorlek kan vara lämpligt.

Eftersom berget har avsatts i lager med tvära förändringar av kalkstenstyp blir interpolationen svår. Vidare undersökningar krävs för att etablera goda interpolationsmetoder på grund av kalkstenens komplexa struktur.

Nyckelord: Multivariat analys, interpolation, PCA, principalkomponentsanalys, PLS, projektion till latenta strukturer, partiell minstakvadrat- anpassning, Kalkbrott, Klinthagen, Kriging.

Institutionen för informationsteknologi, Uppsala universitet
Box 337 SE-751 05 Uppsala

PREFACE

This master engineering thesis, earning 30 university credits, was performed at IVL Swedish Environmental Research Institute. The thesis supervisor was Erik Lindblom environmental consult at IVL. The subject was reviewed by Professor Bengt Carlsson at the Department of Information Technology, Uppsala University.

I would like to thank Kenneth Fjäder and Tomas Kjellin at Nordkalk AB, without the field-data and support supplied by Nordkalk this thesis would not have been possible. Furthermore, I am directing my gratitude towards Erik Lindblom, who enabled me to write this thesis at IVL and provided helpful guidance throughout the project.

Anders Björk deserves a special mentioning for all modelling advises and for his help with reviewing the thesis.

Linus Bergfors
Stockholm 2010

UPTEC W 10 004, ISSN 1401-5765
Printed at the Department of Earth Sciences, Geotryckeriet, Uppsala University, Uppsala, 2010.

POPULÄRVETENSKAPLIG SAMMAFATTNING

Utforskande multivariat analys av Klinthagentäktens projekteringsdata

Linus Bergfors

Nära Storugns på norra Gotland ligger kalkbrottet Klinthagen, som drivs av Nordkalk AB. Där bryts och förädlas en rad olika kalkstensprodukter. Dessa används framförallt i olika industriella processer. Den viktigaste användaren är stålindustrin, som använder kalksten i sin förädlingsprocess. Klinthagens kalksten har visat sig vara mycket väl lämpad för detta ändamål. Tyvärr börjar tillgångarna av kalksten i Klinthagen att ta slut och man räknar med att bryta den sista stenen i brottet under 2012. På grund av detta har Nordkalk ansökt om att få öppna ett nytt kalkbrott i samma område på Gotland.

För att i framtiden utnyttja kalkbrottstäkter på ett bättre och effektivare sätt finns ett behov att utveckla nya metoder för planering och kartläggning. Brytningen och planering av nyttjandet av Klinthagentäkten är oprecis och grov, vilket lämnar stora möjligheter för förbättringar.

Kalksten består av gamla korallrev och andra vattenlevande organismer som sjöliljor, och svampdjur. Klinthagens kalksten bildades för mer än 400 miljoner år sedan, då revet låg någonstans i närheten av ekvatorn. Berget har sedan genom rörelser i landmassorna förflyttats och pressats upp till den plats det är nu. Eftersom kalkstenen bildas på ett så speciellt sätt får den en lagerstruktur som beror på vilken typ av organsim som ligger till grund för det lagret.

I denna uppsats utvärderas möjligheterna med de multivariata analysmetoderna PCA (Principal Component Analysis) och PLS (Partial Least Squares) då de används på data från ett kalkbrott (Klinthagen). Av särskilt intresse för Nordkalk är om kalkens temperaturkänslighet och svavelinnehåll kan förutspås. För att vidareutveckla undersökningen genomfördes även ett försök att beräkna hur kalkstenen förändras mellan provtagningspunkterna med hjälp av en metod kallad Kriging.

PCA är en statistisk metod för att göra data som innehåller många variabler mer överskådlig. Analysen ger information om trender och avvikelser i materialet. Dessutom beskriver metoden hur de olika variablerna påverkar varandra.

PLS är en utveckling av den teknik som används i PCA men informationen om hur variablerna påverkar varandra används för att skapa samband, som kan förutspå hur en eller flera variabler kommer att bete sig.

Då Kriging används för att uppskatta hur berget förändras mellan borrhålen analyseras först hur långt bort från en punkt omgivningen påverkas av dess värde. Därefter används informationen för att, utifrån de punkter där data finns, beräkna vad som kan finnas mellan dessa punkter.

Analyserna med PCA visade att metoden fungerar bra för den här typen av material, men flera olika omständigheter försvårade analysen. Bland annat var datakvaliteten väldigt varierande och den komplicerade bergstrukturen gjorde analyserna svåra att tolka. För att förbättra framtida analyser bör provtagningsmetoden förändras något för att få ett mer lättolkat material.

Det fungerade bra att med PLS förutspå svavelinnehåll i kalkstenen, däremot gick det inte att förutspå temperaturkänslighet. För att ta fram modeller, som klarar detta måste det befintliga datamaterialet kompletteras med ytterligare information. Delar av resultaten tyder på att kornstorlek och stenens småskaliga struktur till stor del påverkar dess motståndskraft mot höga temperaturer.

Analysen av beräkningarna av bergets utseende mellan provtagningspunkterna visade att det med det använda datamaterialet inte finns någon fördel med att använda sig av Kriging. Återigen var det de komplicerade variationerna i kalkstenen som bidrog till svårigheterna. På grund av lagerstrukturen i berget kan förändringar ske mycket snabbt, vilket är svårt att förutse då provtagningarna är gjorda på en mycket grövre skala. Om Kriging ska användas måste noggranna mätningar för hur kalkens värden varierar genomföras. Frågan om det är lämpligt att använda Kriging bör behandlas noga innan försöken påbörjas.

TABLE OF CONTENTS

1 INTRODUCTION

The limestone products quarried at Klinthagen, Sweden are used in a wide variety of processes, which all demand different characteristics. To meet the demand, an industrial exploit of the resource has been developed since the start in 1987 (Karlsson, 2008). In year 2004 about 3.2 million tons of limestone products were produced at Klinthagen (Karlsson, 2008). However, since the resources are reaching its end a decrease in productivity has been inflicted at the location, and the site is expected to be fully exploited in 2012 (Nordkalk, 2010a). Nordkalk is therefore presently applying for permission to start a new quarry, Bunge, close to Klinthagen.

The Limestone at Klinthagen is a heterogenic sedimentary rock type, which originates from coral reefs and other sea-living life forms living more than 400 million years ago. This has caused the quarry, of about 180 hectares, to be constituted by lens-like reef bodies reaching up to 200 meters in length, 70 meters in width and 20 meters in depth. Every reef body is by itself composed by layers of rock originated from different life-forms, which are also mixed with clay materials and eroded reef fragments. To view the list with limestone types represented at Klinthagen, see Appendix IV. The information in this paragraph was obtained from Nordkalk (2010b).

During a visit at the quarry in the middle of November the stratified nature of the rock was noticed, which also is visible in the picture (Figure 1).

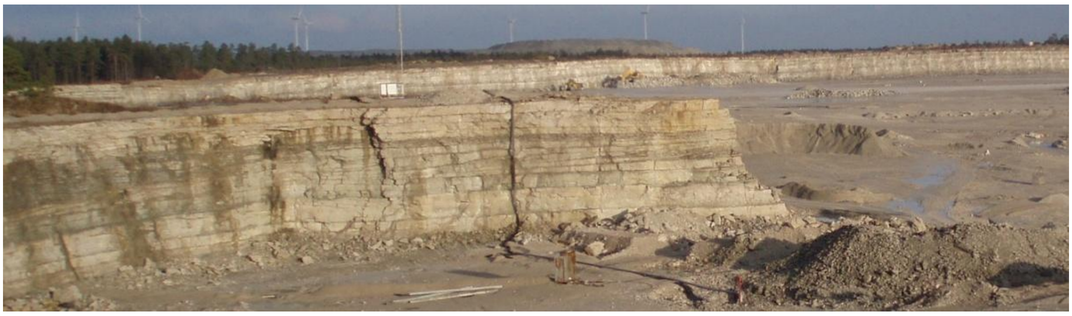


Figure 1 Picture overlooking a part of the Klinthagen quarry, notice the stratified rock composure. The rock wall is about 15 meters in height.

The Klinthagen limestone is known to be a product of high quality with low levels of contaminants, such as sulphur, and a low tendency to break at high temperatures, which is rather rare and of great value to the iron and steel industry. If the areas of high quality may be mapped and therefore more efficiently extracted it could reduce the impact on the environment at future locations and increase quarry efficiency.

Using multivariate statistics to analyse the data from the Klinthagen quarry valuable information to improve future quarry operation may be found. Multivariate statistical analysis is mainly used when large datasets containing many variables are evaluated.

The Klinthagen quarry location may be viewed in Figure 2.

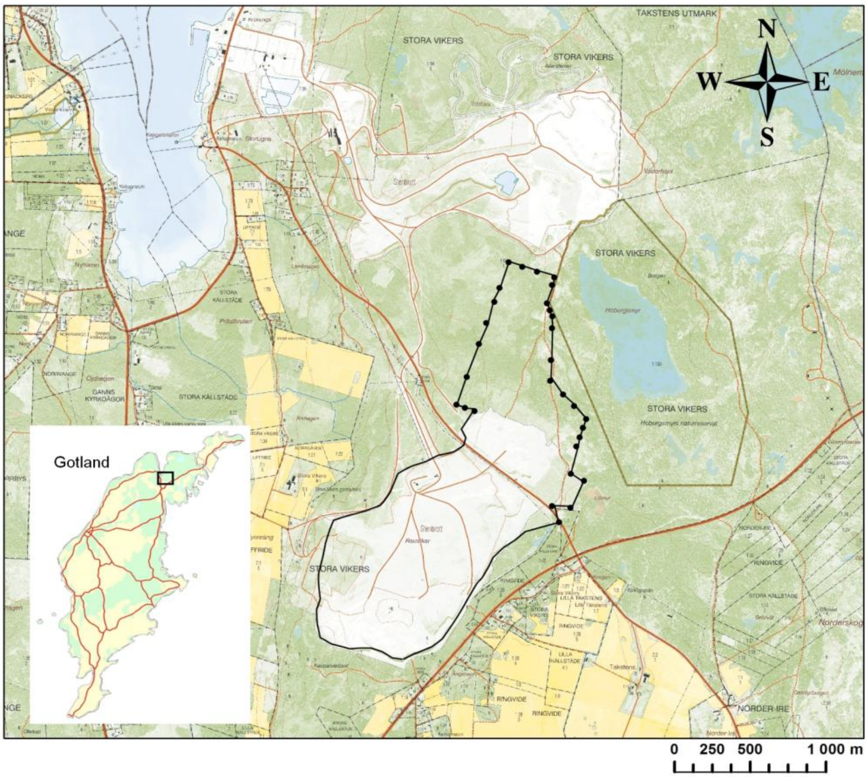


Figure 2 Klinthagen limestone quarry at Gotland. The dotted line represents the area not yet quarried (Karlsson, 2008) Printed with permission from Lantmäteriverket (I 2010/0058).

There are several studies using multivariate methods principal component analysis (PCA) and in some cases partial least square projection to latent structures (PLS) on geological data. Of special interest are those studies, which contemplate the chemical structures and spatial decomposition of bedrock. Esbensen et al. (1987) presented a study linking chemical data from overburden to geophysical data as density, magnetic susceptibility etc. They used PCA in the initial steps of the analysis, moving on with PLS to predict the geophysical characteristics from the chemical structures of the overburden, followed by Kriging interpolation of the results to obtain a map showing the spatial distribution of predicted geophysics in the project area.

Jimenez-Espinosa et al. (1993) used PCA to analyse the soil chemistry data of an area located in North-Western Spain. They let the first principal component represent six highly correlated components as a single variable. Jimenez-Espinosa then derived spatial images through Kriging analysis to visualise how this new variable was distributed in the area as to identify anomalties.

In southern Portugal a quarry used for cement production, was examined for quality by a combination of multivariate and image analysis, see Almeida et al. (2004). Different ratios of chemical components are used by the company active in the area as quality parameters; this was the starting point of the analysis. Almeida et al. used data from a set of sparse bores, creating a large block model. They divided the area into smaller blocks, which characteristics were then estimated through PCA and a simulating routine; this resulting in a set of images visualising the distribution of the variables. The images were analysed according to the “quality parameters” to identify interesting sub-areas in the quarry (Almeida et al. 2004).

1.1 PURPOSE

1. The main purpose of this thesis is to evaluate the possibilities of multivariate statistical analysis (PCA & PLS) and suggest improvements in order to enhance its applicability when applied to geochemical data.

2. Investigate the possibility of predicting thermal disintegration index or sulphur contents from geochemical data.

3. Evaluate the spatial interpolation method Kriging, when applied to PCA data from Klinthagen.

4. Present a documentation of the multivariate techniques; PCA and PLS, and the theory of spatial interpolation by Kriging.

2 MULTIVARIATE ANALYSIS

Multivariate analysis is a powerful tool used to deal with large datasets. The refined measuring techniques of today and possibilities to store large datasets often render datasets of immense magnitude. A vast amount of variables and objects in a dataset may make it impossible to distinguish trends, groups, outliers etc. Multivariate analysis consists of a variety of different methods of handling large matrix problems. In general all the methods subordinated to multivariate analysis are designed to simplify the interpretation of the data. However, depending on the objective of the analysis the methods used have to be chosen carefully.

The objectives where multivariate techniques are most commonly used are described as (Johnson, 1992):

- Data reduction or structural simplification
- Sorting and grouping data
- Variable dependency investigation
- Prediction
- Constructing and testing hypothesis

The techniques of multivariate analysis have been used in many different fields of science such as physics, chemistry, medicine and social studies but also in economics and business studies (Johnson, 1992). Most interesting for this thesis however is its prior use in mining and prospecting (Eriksson, 2001).

2.1 MULTIVARIATE NORMAL DISTRIBUTION

Most multivariate analysing techniques are based on the assumption of a multivariate normal distribution of the dataset (Johnson, 1992). In this thesis the main analysing methods (PCA & PLS) are based on projections, which are not restricted by the distribution of the data (Johnson, 1992). However, if the data is approximately normally distributed it may simplify the analysing process. To have an understanding of the data distribution prior to modelling can be valuable when making decision during the analysis; therefore it is a basic step of data analysis to determine the distribution.

In the univariate case, the samples of one variable are studied to evaluate the probability of a certain outcome if a new sample was to be taken. The probability is calculated from the normal distribution function, which forms a bell-shaped curve with maximum peak at the mean of the variable. Depending on the standard deviation of the samples the curve will be more or less stretched towards the edges. The area under the curve describes the probability of a sample to be within a certain interval. The normal distribution is given by:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-[(x-\mu)/\sigma]^2/2} \qquad -\infty < x < \infty \qquad (2.1)$$

where σ is the standard deviation and μ is the mean.

In the multivariate case the probability is described by the multivariable normal distribution function (2.2), which is analogous with the univariate function (Johnson, 1992). The function will describe a p-dimensional surface (Figure 3), where p is the number of variables included. To evaluate the probability, the volume under the surface over a region formed by intervals has to be determined (Johnson, 1992). Analogous to the univariate case the standard deviations and the covariance affect the shape of the surface greatly. When the dimension exceeds two, $p > 2$, it is hard to obtain a satisfactory graphical illustration. The multivariate normal distribution is given by:

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \cdot e^{-(\mathbf{X}-\vec{\mu})' \Sigma^{-1}(\mathbf{X}-\vec{\mu})/2} \qquad (2.2)$$

where p is the dimension, X is a matrix with p variables, Σ is the covariance matrix and $\vec{\mu}$ is a vector of expected values for each variable in X.

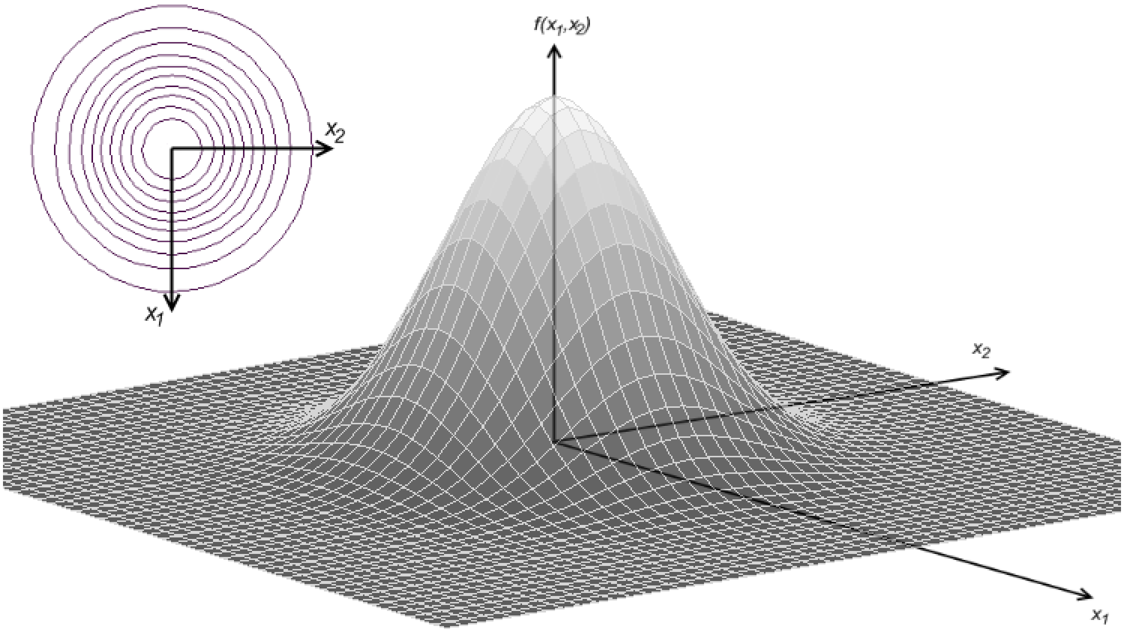


Figure 3 A graphical representation of a two dimensional normal distribution, where the variables variation is the same and no correlation occur. In the top left corner the distribution is shown as contours from above.

If it is concluded that a dataset has a multivariate normal distribution the following stands true (Johnson, 1992):

- Linear combinations of components of X are normally distributed
- All subsets of the components of X have a (multivariate) normal distribution
- Zero covariance implies that the corresponding components are independently distributed
- The conditional distributions of components are (multivariate) normal

2.2 MULTIVARIATE PROJECTION METHODS

Projection techniques deal with three aspects of the analysis: data overview, classification and discrimination and regression modelling (Eriksson, 2001). The analysing procedure often contains all of these three aspects, starting with an overview, moving on with classification and discrimination and finally approximating a model predicting one or more of the variables involved (Eriksson, 2001).

Principal component analysis (PCA) and partial least square projection to latent structure (PLS) are both multivariate projection methods.

PCA applied to the entire dataset will provide an overview of the variables and observations to be analysed. From this overview it is possible to extract information about the relations between observations, groups of observations and deviating observations. Other important information, which PCA may reveal are trends and shifting in the data. The overview also contains information on the correlation of the variables and how the variables are connected to the observations.

If the initial PCA shows distinguishable groupings in the data, this stresses the question of classifying observations. It may be necessary to perform additional PCA for each group separately in order to obtain further knowledge about groups and their characteristics (Eriksson, 2001). These new PCA-models or class-models provide the possibility to classify new observations. However, should a new observation prove not to fit any of the established classes, this becomes an interesting sample and will need to be examined more closely.

The PLS technique may make it possible to achieve a model able to predict a certain set of variables as responses to a set of new observations, which is desired. This is often the main objective of the data analysis. The model provides the opportunity to study how the observations affect the responses and how the responses correlate (Eriksson, 2001). When applying PLS to a dataset, it is important to separate causality from correlation. A causative relationship between observation and response means that a change in the observed variable causes the response to change, whereas for a correlation the change in the observed variable and the response may in fact be caused by another unknown variable and the observation and response are simply mutually affected.

2.3 PRINCIPAL COMPONENT ANALYSIS

Principal component analysis mainly tries to represent high dimensional data in a space with reduced dimensions (Jolliffe, 1986). The method could be described by a rotation of the axes as to find new variables, which represent the variability in a least square sense in the data to the highest degree (Eriksson, 2001). The new variables are called principal components and are calculated so that the first component represents the most variance and the second represent the second most variance and so on (Jolliffe, 1986).

2.3.1 Computing Principal Components

Principal components may be calculated from either the covariance matrix or the correlation matrix depending on the problem, however the manner of determining are the same. Below, the principals for calculating the components from the covariance matrix are described.

Consider a matrix X with *n* observations of *p* variables. The principal components are defined by seeking the linear combinations, which maximises the variance (Johnson, 1992). This is done by studying the sample covariance matrix, S, by respect to eigenvalues and eigenvectors (Johnson, 1992). The sample covariance matrix is calculated by:

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}'\mathbf{X} \tag{2.3}$$

where *n* is the number of observations.

The eigenvalues of the sample covariance matrix represent the variation in the direction of the corresponding eigenvector (Johnson, 1992). Since the matrices are normally large the eigenvalues could be calculated with the power method or the QL- algorithm (Jolliffe, 1986). When the eigenvalues have been calculated they may be ordered by decreasing number. The first principal component is chosen as the eigenvector corresponding to the largest eigenvalue.

$$\mathbf{y}_1 = \mathbf{e}_1'\mathbf{X} \tag{2.4}$$

where y_1 is the first principal component, e_1' is the transpose of the normalised eigenvector corresponding to the largest eigenvector and X is sample data matrix. The values of y_1 are called *scores* and represent the observations of X projected onto the new axis with the coefficients of e_1. It should be noted that the principal components are to be orthogonal and therefore there should not be any covariance between the components (Johnson, 1992). Obvious from the calculation method PCA is sensitive to scaling, hence the sample data matrix is often centered and normalised before these operations, that is to say the means are subtracted from the observations and vectors are adjusted to be of the same length (Eriksson, 2001). The normalisation is often done by dividing the vector by its standard deviation (Eriksson, 2001).

If the procedure is followed through to the last component all of the variance in X will be accounted for (Johnson, 1992). The eigenvectors will form a matrix A containing the directions of all the orthogonal principal components. The *scores* of the observations for all the components may be expressed as (Jolliffe, 1986):

$$\mathbf{Y} = \mathbf{XA} \tag{2.5}$$

Another, more direct approach for calculating the principal components is obtained through singular value decomposition (Jolliffe, 1986). This states that the sample matrix, X, can be written as

$$\mathbf{X} = \mathbf{ULA}' \tag{2.6}$$

The decomposition is based on finding eigenvalues and eigenvectors to the matrices X'X and XX' (Golub, 1965). The columns of U consist of the eigenvectors of XX' and the columns of A are, as before, the eigenvectors of X'X (Golub, 1965). Considering X being a matrix with *n* observations and *p* variables implies that the dimensions of U and A should be ($n \times n$) and ($p \times p$) respectively. The matrix L is diagonal with the singular values of X as elements. The singular values are the square roots of eigenvalues to either X'X or XX' (Golub, 1965). The elements of L are normally ordered as decreasing from the left and the dimensions are ($n \times p$). The sample matrix X is often rectangular, either more observations then variables or vice versa, therefore follows that L will be filled with zeroes to reach desired dimensions (Golub, 1965).

The column vectors of U and A are determined under the constraint of orthonormality, which implies:

$$\mathbf{U'U} = \mathbf{I} \tag{2.7}$$

$$\mathbf{A'A} = \mathbf{I} \tag{2.8}$$

Equation (2.6) together with (2.8), provides the opportunity to multiply A from the right. By this follows that

$$\mathbf{XA} = \mathbf{UL} \tag{2.9}$$

Comparing this with the result in (2.5) it is obtained that:

$$\mathbf{Y} = \mathbf{UL} \tag{2.10}$$

It is now possible to see that the singular value decomposition performed in this manner provides both the coefficients of the principal components in the matrix A and the *scores* projected to the components in the matrix UL. To link the results to the earlier discussion of principal components retrieved from the sample covariance matrix, it should be noticed that the singular values of X is in fact the square roots of the eigenvalues of the sample covariance matrix multiplied by (*n*-1) (Jolliffe, 1986).

2.3.2 Geometrics of PCA

When faced with a sample matrix X with *n* observations and *p* variables, the observations form a swarm of points in a *p*-dimensional space (Figure 4) (Eriksson, 2001).

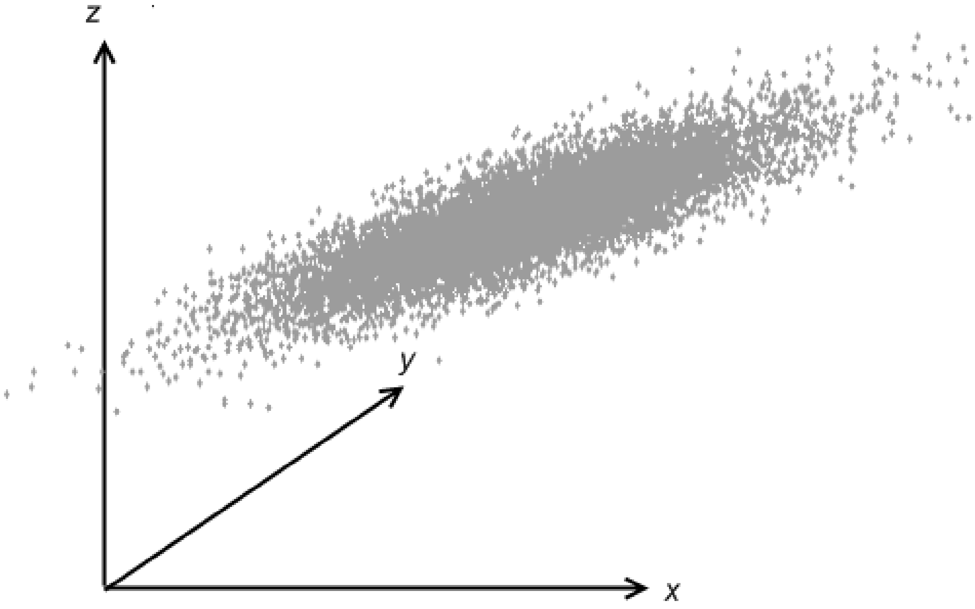


Figure 4 Point swarm in three dimensional space.

The first steps of the PCA will cover the scaling and centering of the points, which will standardise the impact of each variable variance and move the origin of the axes to the mean value. These steps will then be followed by the computing of the first principal component (PC) and the projections of the observations, the *scores*. The first PC is, as earlier explained, an axis in the direction representing the variance of the data to the highest degree (Figure 5).

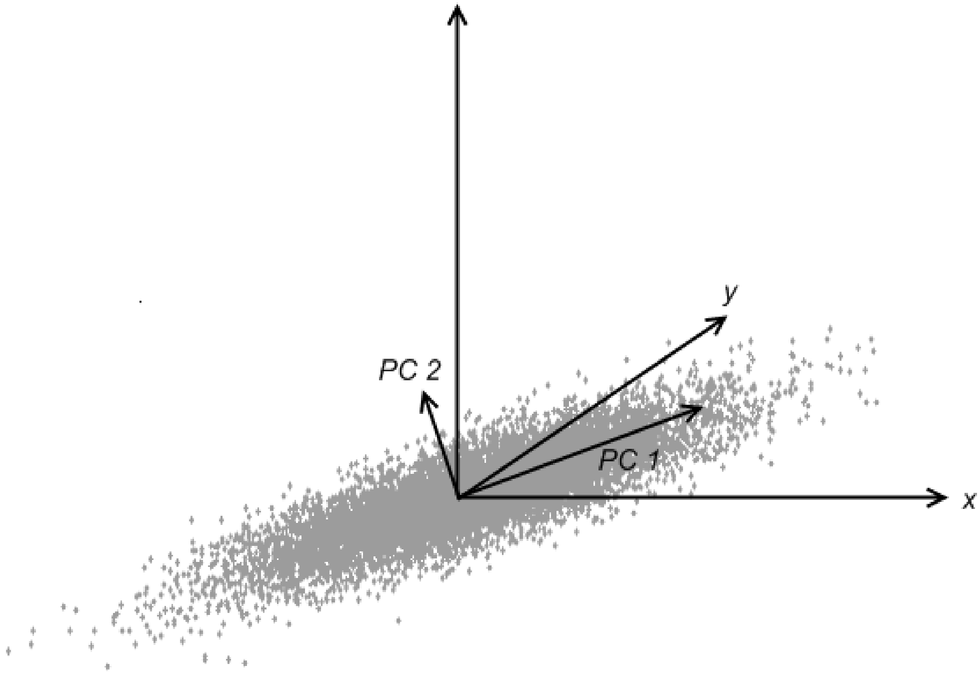


Figure 5 Centered point swarm with two principal components in three a dimensional space.

However, the *scores* of the first PC alone are often not enough to gather a sufficient understanding of the data, therefore a second PC is inserted in the data swarm, which represents the second highest degree of variance. This is sometimes continued by a third and fourth PC but the fraction of variance described decreases for each PC calculated, thus also the correlation to other variables. Normally, the *scores* are viewed in 2-dimensional plot over the first PC and any of the additional components. This can geometrically be described as inserting a plane into the point swarm, and projecting the observations onto it (Figure 6 & Figure 7).

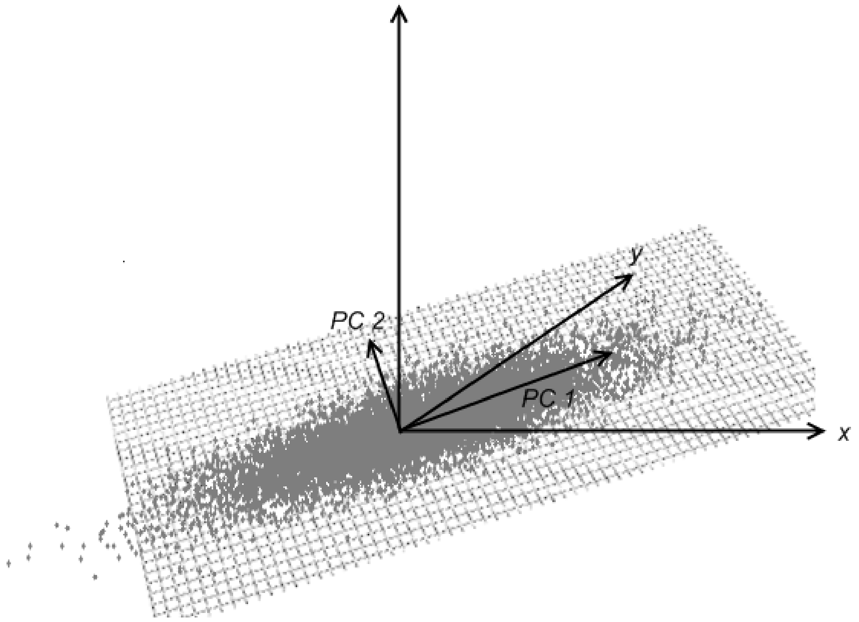


Figure 6 Plane inserted in the point swarm created by the PCs.

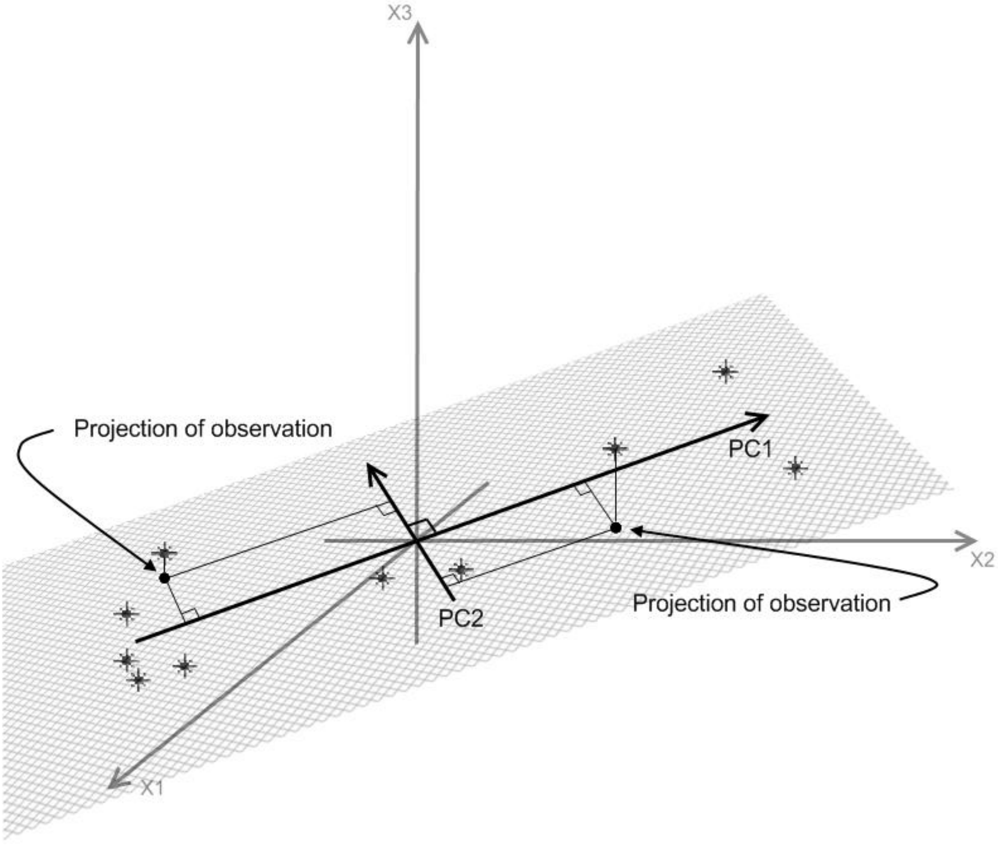


Figure 7 Projection of observation onto the plane.

The plane with projected observations is called a *score plot*, (Figure 8). The *score plot* reveals information about the observations such as groups, trends and outliers; it is often desired however, to relate the *scores* to the original variables.

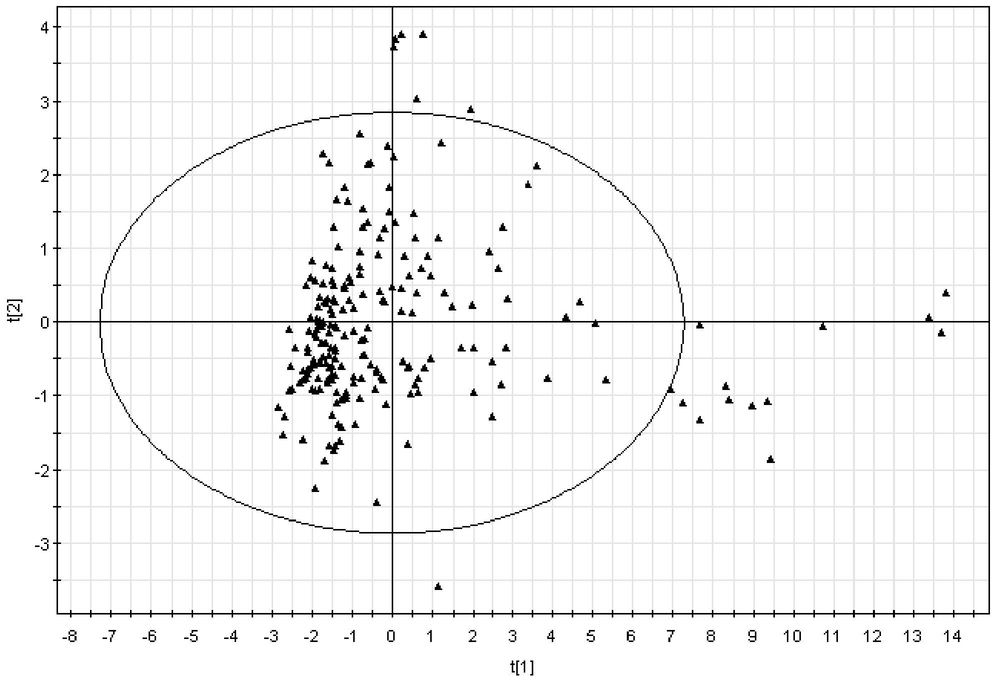


Figure 8 Score plot showing the projected observations, where the circle represents the hotelling's 95 percent confidence interval.

A way of viewing these relations is to examine how the plane is inserted into the original *p*-dimensional space, which is revealed by studying the coefficients of the principal components (Esbensen et al., 1998). The coefficients are called *loadings*, simply because they show how strongly a variable influences the PC. Geometrically the *loadings* are defined as cosine of the angle, *α*, from the variable axis to the PC (Figure 9) (Eriksson, 2001).

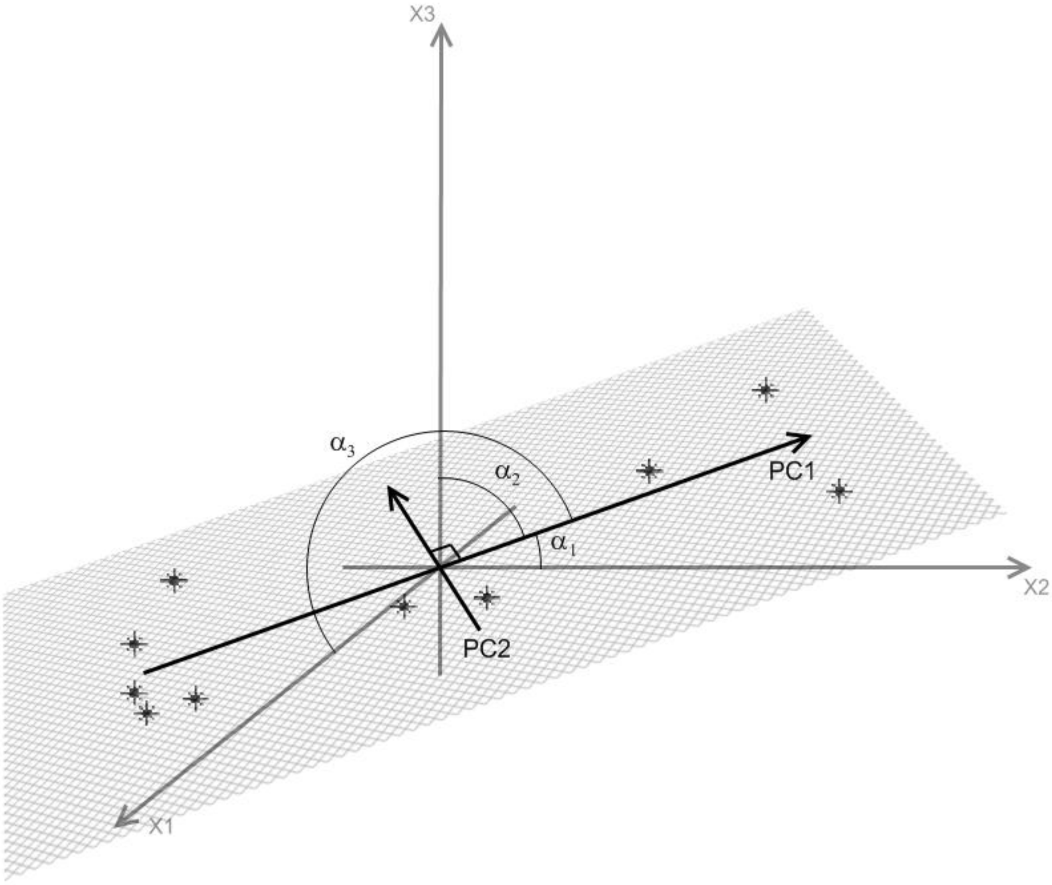


Figure 9 Angles from variables to first PC; showing how the component is inserted in a three dimesional space.

The *loadings* of two components are displayed in a *loading plot* (Figure 10), showing the variable correlation and how they affect the PCs. Variables far from the origin have a greater impact on the PCs opposed to those closer to origin. Variables close to each other may be positively correlated, and those on opposite sides of the origin may be negatively correlated (Esbensen et al., 1998).

Comparing the *score plot* with the *loading plot* is very effective since they complement each other. The directions in the plots are the same, which implies that if observations in the *score plot* are situated close to the location of a variable in the *loading plot* it is likely that these observations are affected by this variable (Eriksson, 2001).

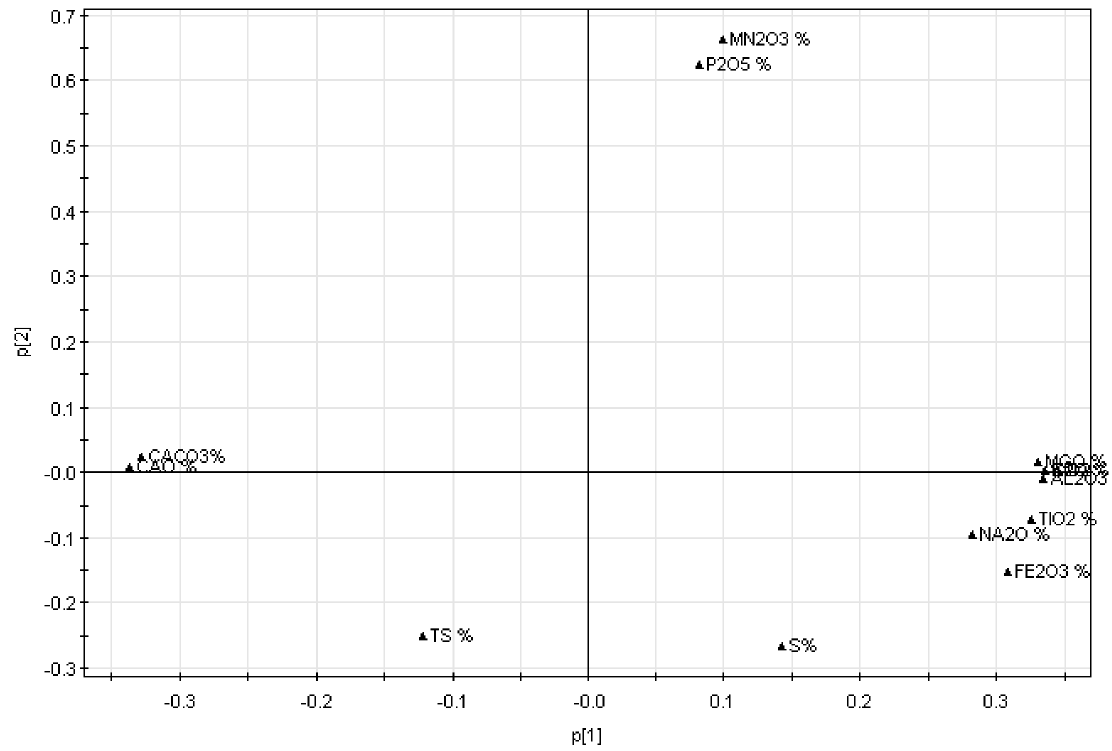


Figure 10 Loading plot showing variables relations.

2.3.3 The PC-model

A PC-model is an attempt to describe the variance of a sample data matrix X in an effective and simple way. It should be noticed though that reducing the dimensions comes with a cost of lost information. The losses are described as noise in the matrix, E. Assuming the data matrix X has been centered by subtraction of mean values and normalised; the resulting data matrix is denoted as X_{sc}. The data may then be written as:

$$\mathbf{X}_{sc} = \mathbf{E} \tag{2.11}$$

where E is the noise, which in this case accounts for the entire variance of the data. Equation (2.11) is sometimes referred to as the zero component model.

Computing the first PC will provide a vector of *scores*, t_1, and a vector of *loadings*, p_1. The model can then be described as (Esbensen et al., 1998):

$$\mathbf{X}_{sc} = \mathbf{t}_1 \cdot \mathbf{p}_1' + \mathbf{E}_1 \tag{2.12}$$

where E_1 is the new noise matrix, which in comparison to (2.11) has reduced by the variance accounted for by the first PC.

The model work continues by adding one PC after another. However, for each PC added the fraction of variance explained by the new PC decreases. The gain of adding a PC should be considered as it brings with it the cost of a more complex model. With the final number of PCs decided the model may be expressed as:

$$\mathbf{X}_{sc} = \mathbf{T} \cdot \mathbf{P}' + \mathbf{E} \tag{2.13}$$

where T is formed by the *score* vectors, t_i, and P consists of the *loading* vectors, p_i. The index, *i*, ranges from 1 to the number of PC decided upon.

Since, the PCs do not account for all the variance in the data; the matrix E represents an important and informative part of the model, which reveals how the observations and variable deviate from the model. Geometrically the noise is the distance from an observation to the plane spanned by the PCs (Figure 11).

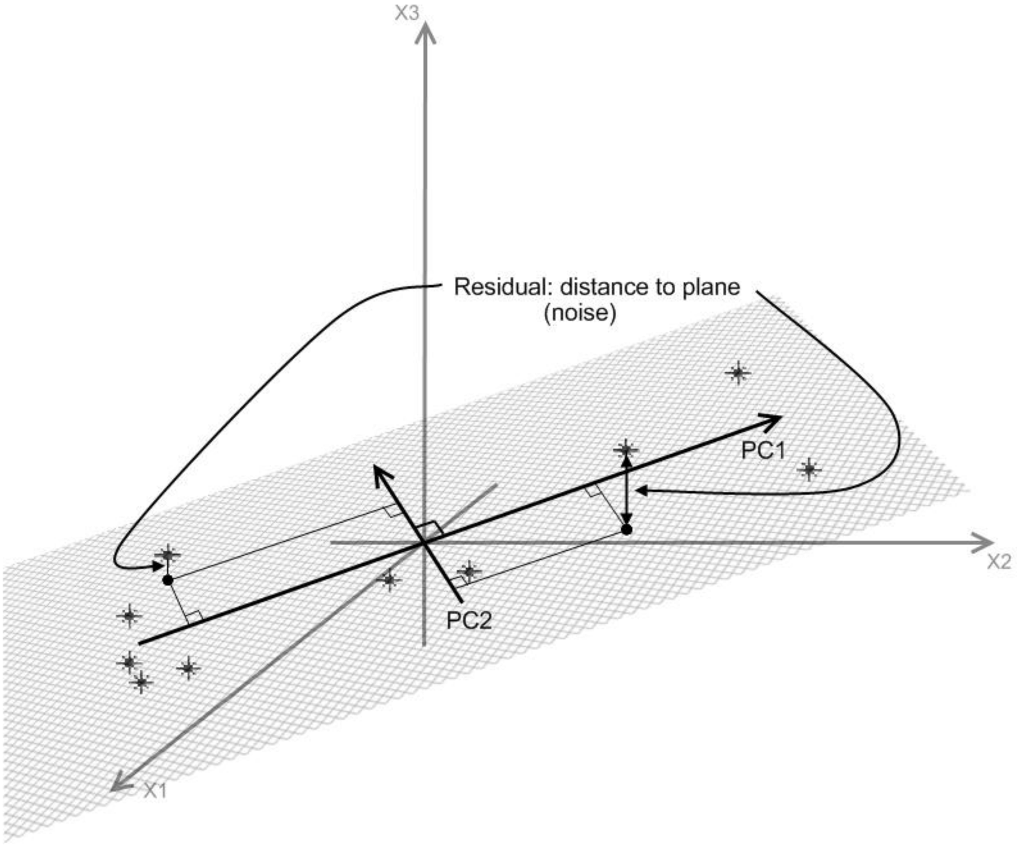


Figure 11 The noise (residuals) is the distances from the observations to the projections on the plane.

These distances are called residuals and are the content of E. By plotting how each observation deviates from the model (Figure 12), it may be possible to identify outliers not spotted in the *score plot* (Eriksson, 2001). It may also reveal if there are shifts in the data (Eriksson, 2001).

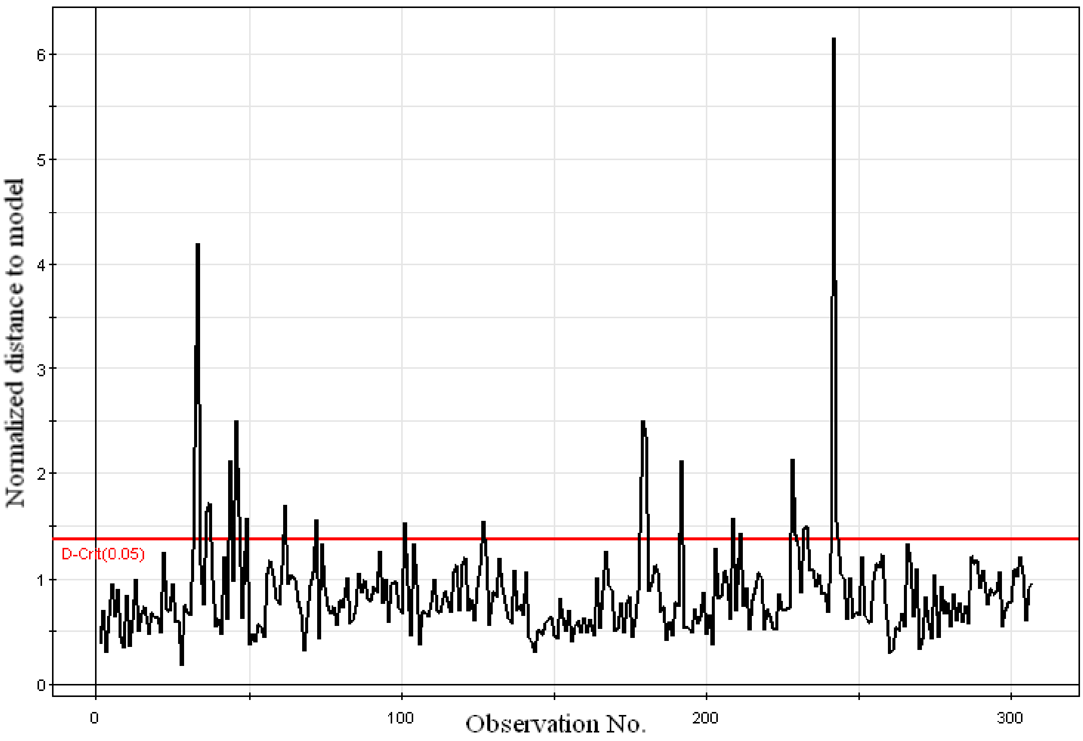


Figure 12 The observation residual plot shows the distance from the observation to the model.

Furthermore, plotting the residuals of a certain variable provides information on how well that variable is explained by the model (Eriksson, 2001). Often this is plotted in a cumulative manner (Figure 13), by adding the fraction of the residuals accounted for by a principal component to the next (Eriksson, 2001). In this way it is possible to understand which variables and to what extent they are explained by each PC.

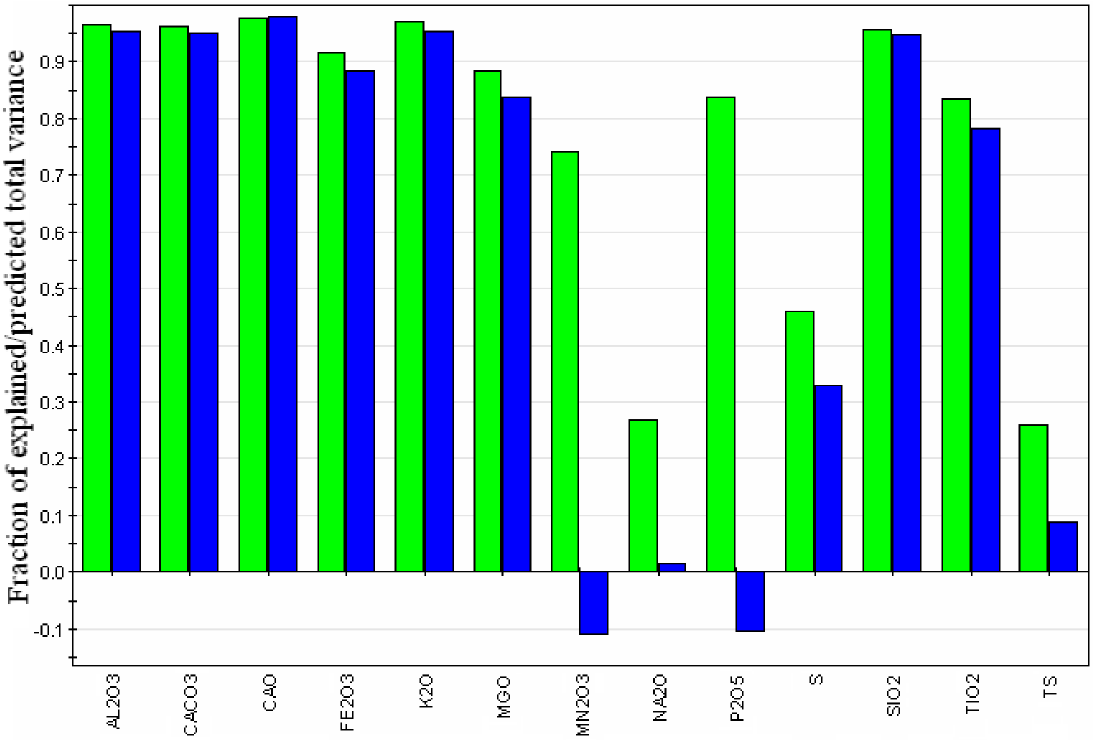


Figure 13 Variable residual plot, the left bar shows residual and the right shows how well the variable is predicted against a validation set.

The decision on number of PCs to use is a complicated matter. Examining the residual decrease provides a good guidance on how many to choose. The total residual variance is expressed by (Esbensen et al., 1998):

$$e_{tot}^2 = \sum_{i=1}^{n} e_i^2 \qquad i = 1, 2, ..., n \qquad (2.14)$$

where e^2 is the squared residuals of the observations, e_{tot}^2 is the total residual variance and n is the number of observations.

2.4 COMMON MULTIVARIATE REGRESSION METHODS

2.4.1 Multivariate Linear Regression and Principal Component Regression

In data analysis response prediction is often done by some regression method. The simplest and most commonly used is the univariate linear regression, $y = a + bx$ (Esbensen et al., 1998). In the multivariate case the corresponding technique is called MLR (multivariate linear regression), which fits a linear combinations of several variables, $x_1, x_2, ..., x_n$, to describe the response, y (Esbensen et al., 1998).

$$y = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n + f \qquad (2.15)$$

where b_i are the regression coefficients, x_i are the observed variables and f is the factors not included in the model and noise.

The coefficients b_i can be estimated from the least square approximation:

$$\hat{\mathbf{b}} = (\mathbf{X'X})^{-1}\mathbf{X'y} \qquad (2.16)$$

where the vector **y** contain the observed values of the response.

In equation (2.16) the limitation of the MLR is revealed. The least square estimate includes the inverse of the matrix (X’X), which may cause trouble if it is singular or close to being singular i.e. containing any co-linearity or dependencies among the variables, x_i (Esbensen et al., 1998).

A way around the problem of co-linearity is to use a PCR (principal component regression) (Esbensen et al., 1998). The PCR is actually a combination of PCA and MLR. The data matrix, X, is first fully decomposed to a set of principal components, which by definition are orthogonally independent. A MLR is then performed on the new dataset to predict the response, y.

The PCR comes with one great drawback: It is not certain that the chosen PCs, who represent the largest variances of the *predictors*, X, actually include the factors that control the *response* (Esbensen et al., 1998). To be forced to compute the entire set of PCs would cause the model to be more complex and advantage of reduced dimensionality would be lost. An attempt to ensure that the model describes the desired correlations from X to Y is done through PLS-regression.

2.5 PARTIAL LEAST SQUARES REGRESSION

The PLS–regression uses the variance information stored in the response and then applies this when decomposing the X matrix in PCs. This manner of performing the decomposition implies that the variance in the responses, Y, will be explained more efficiently then with a PCR in Section 2.4 (Abdi, 2003). PLS may be used on either a single variable response or a block of several response variables.

Decomposing the matrix of *predictors* with the help of the information stored in the *responses* results in a shift of directions of the PCs compared to those of a PCR (Geladi & Kowalski, 1986). PLS decomposes X into a set of *scores*, *t*, and *loadings*, *p*, and at the same time describes the *response*, Y, as a set of *scores*, *u*, and *weights*, *c*.

$$\mathbf{X} = \mathbf{TP'} + \mathbf{E} \qquad (2.17)$$

$$\mathbf{Y} = \mathbf{UC'} + \mathbf{F} \qquad (2.18)$$

where T is as before the *score* matrix of X and P the *loading* matrix. The matrices U and C contain the *scores* and *weights* from the decomposition of Y. E and F are the residuals not described by the model.

During the decomposition, the structure of Y is allowed to influence the decomposition of X by letting the Y-*scores*, *u*, be a part of the forming of the X-*scores*, *t*; this forming an inner relationship as:

$$u_k = b_k t_k \qquad k = 1, 2, ..., n \qquad (2.19)$$

where b_k is a regression coefficient and *n* is the number of observations.

It should be mentioned that equation (2.19) is a linear relationship, which is the simplest, but not necessarily the best. There are ways to account for non-linearities by replacing equation (2.19) with relations of a higher order or extending X with for instance squared or cubic terms (Björk, 2007). If the inner relationship equation (2.19) is included in the model a possibility to estimate the responses, Y, from the *scores* of X is presented as (Abdi, 2003):

$$\hat{\mathbf{Y}} = \mathbf{TBC'} + \mathbf{F} \qquad (2.20)$$

where B is a diagonal matrix with the regression coefficients on the diagonal and $\hat{\mathbf{Y}}$ represents the estimate of Y.

The prediction of Y may also be expressed as a relation directly to X; this is done by using W* instead of W, which connects back to X instead of the residuals of X. Then the estimate is written as (Eriksson, 2001):

$$\hat{\mathbf{Y}} = \mathbf{B}_{PLS}\mathbf{X} \qquad (2.21)$$

where; $\mathbf{B}_{PLS} = \mathbf{W}(\mathbf{P'W})^{-1}$

B_{PLS} is called the PLS regression coefficient matrix and is useful in interpretation of the model (Eriksson, 2001). It shows how each predicting variable is contributing to the response.

Geometrically, PLS describes a plane or hyper-plane in the space spanning X (Wold et al., 2001). However, the *scores*, *t,* of X and the *weights* of Y, *c*, indicate directions, in this plane, with highest correlation to Y (Figure 14), which performs a link between the *predictors* and the *responses* (Wold et al., 2001).

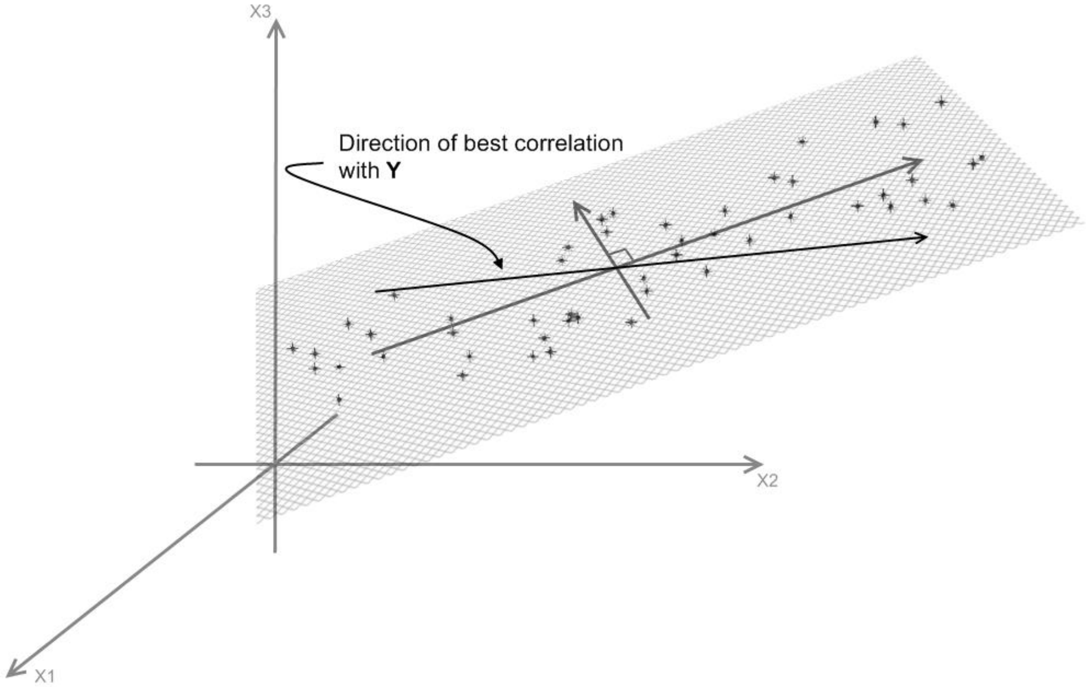


Figure 14 The line shows the direction with highest correlation to Y in the plane formed by two PLS components.

PLS is an iterative method calculating one PLS component at the time. Starting the algorithm the matrices X and Y are considered as residuals, E_0 and F_0 respectively; then for each calculated component its contribution to the residuals is subtracted (Abdi, 2003). The sizes of residual matrices are often measured by the total sum of squares, SS_E and SS_F. They serve as a measurement of how much of the residuals are explained by each component (Abdi, 2003). However, the risk of over-fitting the model is severe and cross-validation may therefore be more reliable when choosing the number of components (Wold et al., 2001). The aim is to achieve a model with the smallest possible residual matrix, consisting of as few PLS components as possible.

The algorithm of PLS is performed as (Wold et al., 2001):

1. $\mathbf{u}_{start} = \mathbf{y}_i$ starting Y-*score vector*
2. $\mathbf{w} = \mathbf{X'u} / \mathbf{u'u}$ which calculates the *weight vector* (directions) in X for the *score vector*, u.
3. w should be normalised as $\|\mathbf{w}\| = 1$
4. $\mathbf{t} = \mathbf{Xw}$ computes the corresponding *score vector*, t, of X.
5. $\mathbf{c} = \mathbf{Y't} / \mathbf{t't}$ determines the *weight vector*, c, of Y.
6. $\mathbf{u} = \mathbf{Yc} / \mathbf{c'c}$ calculates the updated *score vector*, u.

7. If $\|\mathbf{t}_{old} - \mathbf{t}_{new}\| / \|\mathbf{t}_{new}\| < \varepsilon$, where ε defines the tolerance limit of convergence. If it has not converged: return to step 2, otherwise proceed two step 8.
8. $\mathbf{p} = \mathbf{X'}\mathbf{t}/(\mathbf{t'}\mathbf{t})$ determines the *loading vector*.
9. $b = \mathbf{t'u}$ computes the regression coefficient
10. $\mathbf{X} = \mathbf{X} - \mathbf{tp'}$ subtracts the contribution of the *scores* and *loadings* from the residuals
11. $\mathbf{Y} = \mathbf{Y} - b\mathbf{tc'}$ deflates Y.

Restart to calculate the next PLS component.

The above algorithm is one of the simplest for PLS regression and is called NIPALS (Wold et al., 2001). There are other alternatives derived for different shapes of data (Wold et al., 2001). For instance the NPLS deals with matrices of more than 2 dimensions e.g. matrices of cubic form (Björk, 2007).

As with PCA, PLS is also closely related to singular value decomposition (Abdi, 2003). It can be shown from the algorithm that the *weight vector*, w, is the first right singular vector to the matrix X'T and the *weight vector*, c, is the first right singular vector (Abdi, 2003). The first *score vector*, t, may be calculated as the first eigenvector of the matrix XX'YY' and the first *score vector*, u, is the first eigenvector of YY'XX' (Abdi, 2003). This may be repeated to retrieve following *score and weight vectors* by using the deflated matrices (Wold et al., 2001).

2.5.1 Interpreting and Analysing the Model

The characteristics of the PCA model may be analysed from the scores, loadings and residuals; while the interpretation of a PLS-model is mainly done from the weights, regression coefficients and VIP (variable influence on projection) (Eriksson, 2001).

As with PCA the score plot is a tool to identify outliers and trends in the model (Eriksson, 2001). The scores from the X and Y blocks may be plotted separately to reveal the model structure of each block, but also the score from X may be plotted against the corresponding score vector in Y (Eriksson, 2001). This enables the possibility to identify non-linearities between X and Y (Figure 15), which may indicate the need for transformations of the data.

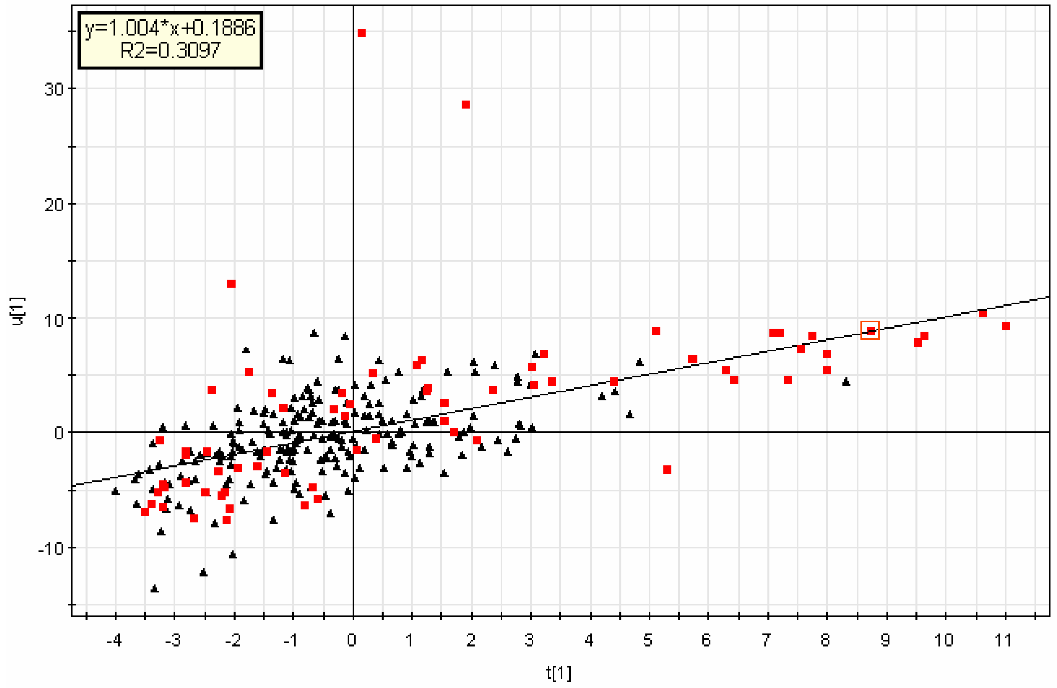


Figure 15 The plot of the scores from X and Y projection, a tool to discover non-linear relationship.

The residuals also provide information about the model. For instance the size of the residuals could be viewed as an indication on model quality (Wold et al., 2001). It is also possible to distinguish moderate outliers, who could not be identified by the score plot, and to examine how much of the variation in the variable is explained by the model (Eriksson, 2001). This is done in the same manner as with the PCA described in Section 2.3.3.

Pressing on with analysing the model; the first tool is to analyse the weights of the *predictors* and the *responses*. The weights are plotted either separately for the predictors and responses or jointly, where the latter shows how the predictor variables affect the responses, and the separate plots displays how the responses or predictors relate to each other (Eriksson, 2001).

Studying the sizes and directions of the PLS regression coefficients; it is possible to distinguish how strong impact the predicting variables has on one response variable (Eriksson, 2001). This is visualised by a bar chart showing the size and direction with the chosen confident interval (Eriksson, 2001). In accordance to this there will be one chart for each response (Figure 16).

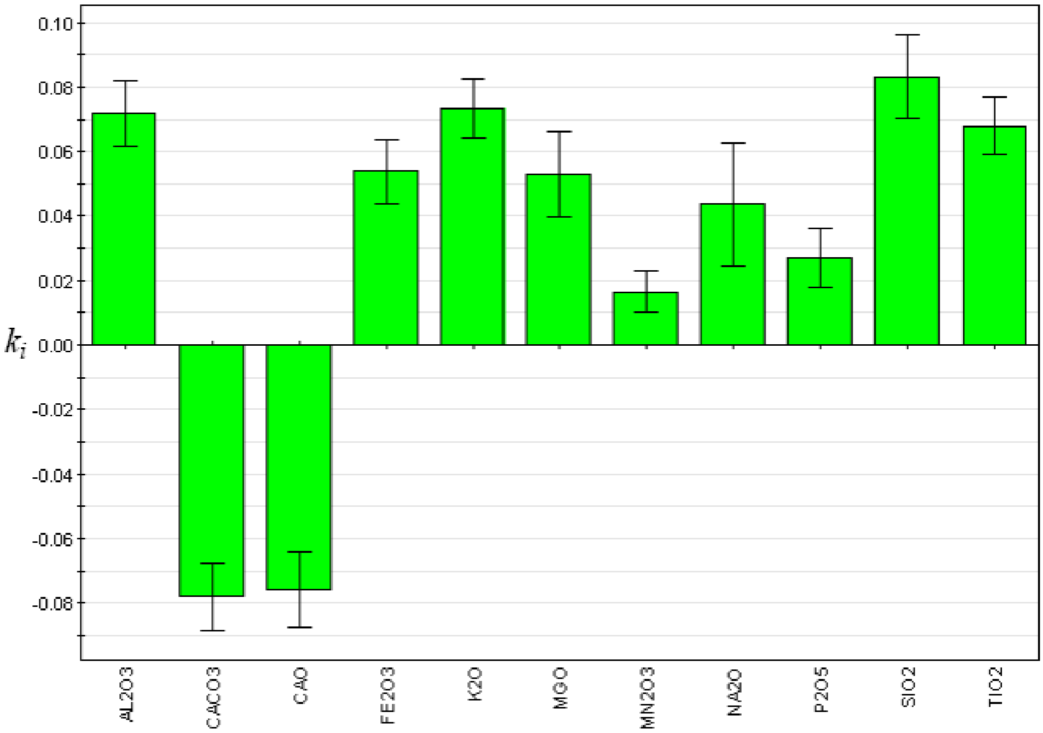


Figure 16 The PLS coefficient plot displays how the predicting variables affect the response.

Variables influence on projection (VIP) is a measurement of how great the importance of a predicting variable is to the entire model (Wold et al., 2001). It takes into account both how great the variable influences the representation of X and its importance in estimating Y (Wold et al., 2001). For each model, there is only one vector representing the VIP values, which makes it easy to overview. The VIP value is calculated as:

$$VIP_{Ak} = \sqrt{\left(\sum_{a=1}^{A}\left(w_{ak}^{2}\cdot\left(SSY_{a-1}-SSY_{a}\right)\right)\cdot\frac{K}{SSY_{0}-SSY_{A}}\right)} \qquad (2.22)$$

where *A* is the number of PLS components, *SSY* is the sum of square from the residual matrix of Y, which represents the explanation and *K* is the number of predictor variables (Eriksson, 2001).

Equation (2.23) shows that the VIP values are always positive and that the sum of all the VIPs is equal to the number of predictor variables, which further implies that variables with greater VIP values than one has the largest influence on the model. This may be visualised in a VIP bar chart (Figure 17).

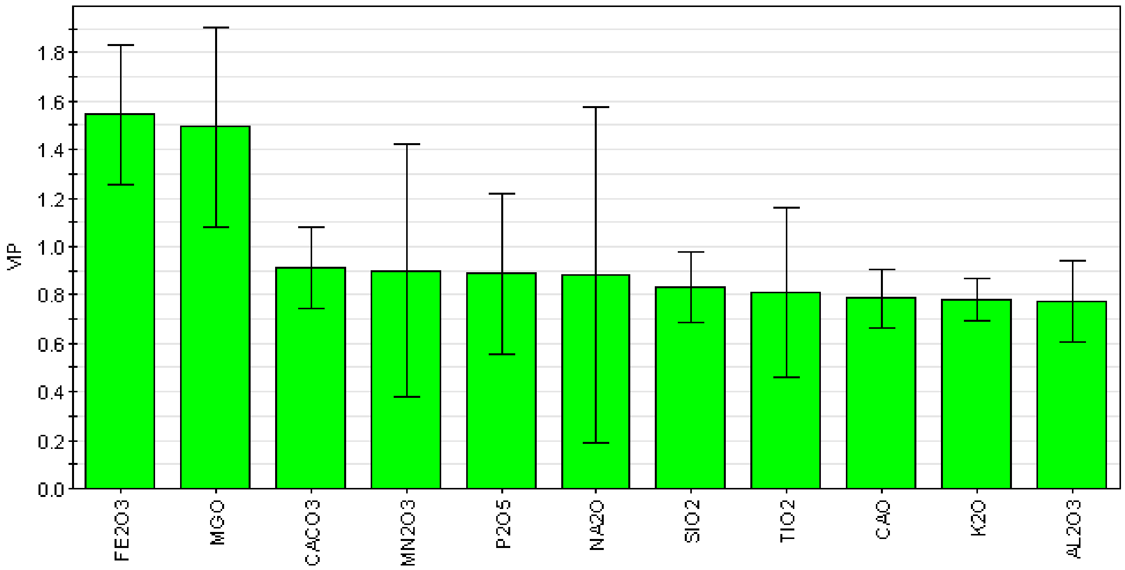


Figure 17 The variable importance plot shows how much a variable affects the model.

2.6 MULTIVARIATE SPATIAL INTERPOLATION

The two different interpolation methods considered in this thesis are described below. *Kriging*, the more complex of the two, takes into account how fast the value changes in the field. The other method, called *inverse distance weighted* (IDW) interpolation, only considers the distance between points.

2.6.1 Semivariogram

The semivariogram is a function describing the spatial variance in a field (Cressie, 1991). In other words it may be considered as an expression of the spatial dependency. Supposing the variations in the field are not completely stochastic, the variance is likely to increase with distance, i.e less dependency is visible (Figure 18). Calculating a semivariogram from spatial sample data is done by (Cressie, 1991):

$$\hat{\gamma}(h) = \frac{1}{2 \cdot |N(h)|} \sum_{N(h)} \left(Z(s_i) - Z(s_j)\right)^2 \qquad (2.23)$$

where $N(h) = \{(s_i, s_j) : s_i - s_j = h; i, j = 1, \ldots, n\}$, i.e. the number of observations within the same relative distance to each other, and Z is the observations. It is easy to see the resemblance with the estimator of variance.

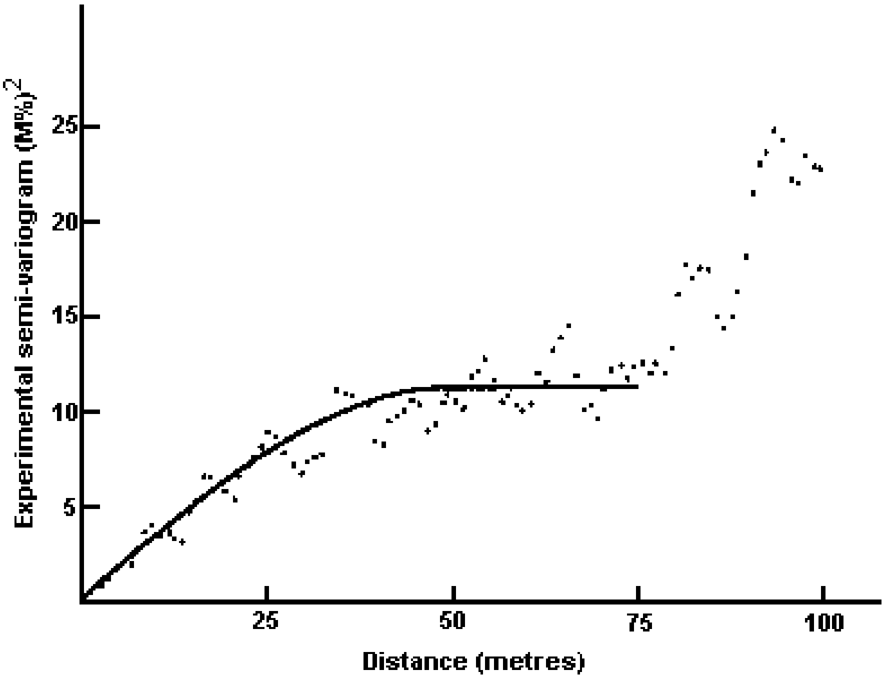


Figure 18 Showing an example of a semivariogram, with a fitted model. Notice how the variance increases with distance (Clark, 2001).

The variations in the field often depend on the direction, which is called anisotropy (Cressie, 1991). For instance the semivariogram in east-western direction may differ from north-south, thus the calculations are commonly performed separately for each direction. This may also be performed in different ways either as only observations in a certain direction, for instance east-west or as observations within a certain angle of tolerance, e.g. all observations between 45 degrees north and 45 degrees south of the east-western line may be included.

When a semivariogram has been estimated from the empirical data a mathematical model is fitted to approximate the function, There are a number of different models commonly used for this (Cressie,1991); however, only the exponential model, which was used in this thesis will be described here. The model is given by:

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ c_0 + c_1\left(1 - e^{-\|h\|/a}\right) & h \neq 0 \end{cases} \qquad (2.24)$$

where c_0 is the nugget effect most likely caused by measurement error and microscale processes (Cressie, 1991). Often the experimental semivariogram does not start at zero but has an offset called nugget. The model converges towards $c_0 + c_1$. The distance is described by h and a is called the range (Cressie, 1991).

2.6.2 Ordinary Kriging

The spatial interpolation or prediction technique evaluated in this thesis is called Ordinary Kriging. It allows the mean value to vary throughout the field, but assumes a constant mean locally in a smaller sub-field or neighbourhood (Bohling, 2005). The spatial interpolation is denoted as (Cressie, 1991):

$$\hat{z}(s^*) = \sum_{i=1}^{n} \lambda_i \cdot Z(\mathbf{s}_i) \tag{2.25}$$

where $\hat{z}(s^*)$ is the value estimate at the spatial location s^*, λ_i are the interpolation weights specific for the location s^* and $Z(s_i)$ are the observations at the known locations s_i. The weights λ are determined under the constraint of:

$$\sum_{i=1}^{n} \lambda_i = 1 \tag{2.26}$$

The basis of determining the interpolation weights is the semivariogram model, which supplies the semivariance at the locations to be estimated. The weights may be calculated by solving the system (Cressie, 1991):

$$\begin{pmatrix} \gamma(s_1, s^*) \\ \vdots \\ \gamma(s_n, s^*) \\ 1 \end{pmatrix} = \begin{pmatrix} \gamma(s_1, s_1) & \cdots & \gamma(s_1, s_n) & 1 \\ \vdots & \ddots & \vdots & 1 \\ \gamma(s_n, s_1) & \cdots & \gamma(s_n, s_n) & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \mu \end{pmatrix} \tag{2.27}$$

where γ indicates the semivariance obtained from the model in (2.24) when the distance between the two locations are inserte. μ is called the Lagrange parameter, which with last row and column of ones ensures the constraint in equation (2.26). The Lagrange parameter is also used when calculating the prediction error, which is given by (Cressie, 1991):

$$\sigma_{OK}^2 = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \mu \end{pmatrix}' \begin{pmatrix} \gamma(s_1, s^*) \\ \vdots \\ \gamma(s_n, s^*) \\ 1 \end{pmatrix} \tag{2.28}$$

where

$$\sigma_{OK}^2 = \mathrm{var}(\hat{z}(s^*) - z(s^*)) \tag{2.29}$$

A well modelled semivariance is very important. Without a thorough investigation to obtain a satisfactory semivariogram the results may not be reliable (Cressie, 1991).

2.6.3 Inverse Distance Weighting

The inverse distance weighted interpolation does not consider how the distance is connected to the relationship between observations. It focuses instead only on the distance and is given as (Shepard, 1968):

$$z(s^*) = \sum_{k=1}^{n} \frac{w_k}{\sum_{k=0}^{n} w_k} z_k \tag{2.30}$$

where

$$w_k = \frac{1}{d(s^*, s_k)^p} \tag{2.31}$$

where *d(s*,s$_k$)* is the distance between the known observation at location s_k and the unknown at location *s**. The exponent, *p*, is called a power parameter and was for this thesis set to two.

3 MATERIAL AND METHOD

The modelling could be divided into four different parts all related to the multivariate methods described in the theory chapter. However, these four parts could also be subordered into two separate areas; one dealing with the data structure, relations between variables and prediction; and the other using spatial interpolation to overview the spatial distribution.

3.1 DATA DESCRIPTION

Nordkalk uses two different sampling methods; diamond drill core and drill cuttings. The chemical data from diamond drill cores has been collected from a congregated sample of a three meter drill core. This could be viewed as the mean chemical composition of the three meter core column. From the same core, a thermal disintegration test has been performed. During the diamond core drilling procedure the sample is continuously rinsed from drill cuttings by water. This causes soft parts of the rock, such as clay, to be rinsed away, which affects the analysis results.

The drill cuttings have been gathered, while drilling to place the explosives. Also in this case the analysis has been performed as mean samples over three meters (most cases, deviations occur). A problem with these analyses were that the depths of the samples only were approximately correct, which adds a spatial uncertainty to the data. Opposed to diamond drilling there is no rinsing present with this method.

There were three different datasets. Two sets containing chemical data from drill cuttings. The third dataset originated from diamond drill cores, which were analysed with respect to both chemical components and the thermal disintegration index. The quarry has been exploited in two levels. The drill cuttings datasets are each originated from one of these levels and the diamond core drill covered the depth of both levels in a single dataset (Figure 19). In all datasets the chemical component were given as ratios in percent. The chemical components in the data were: Al_2O_3, $CaCO_3$, CaO, Fe_2O_3, K_2O, MgO, Mn_2O_3, MnO, Na_2O, P_2O_5, S, SiO_2 and TiO_2.

The physical parameter, thermal disintegration index, measures how resistant the stone is to thermal strain. In the analysis the stone is crushed into a 5 – 10 mm fraction and is slowly exposed to rising temperature. When the test is finished the disintegrated ratio of the stone is measured. A high index indicates low resistance.

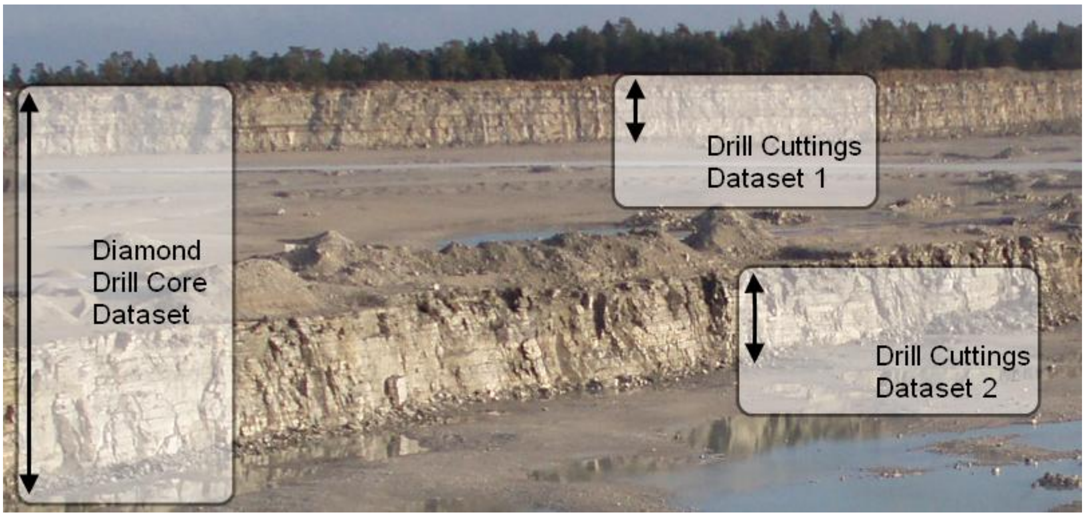


Figure 19 Illustration of how the datasets are represented in the field.

In Figure 20 the three most commonly observed problems in the series plots are encased in red circles: outliers, rounding errors and zeros. It should be emphasized that this is not a time series, but 3-dimensional spatially distributed data; hence it is probably not possible to determine trends, if present, from this plot. Outliers should also be treated with caution since it may not be obvious to see if nearby samples support the sample or not.

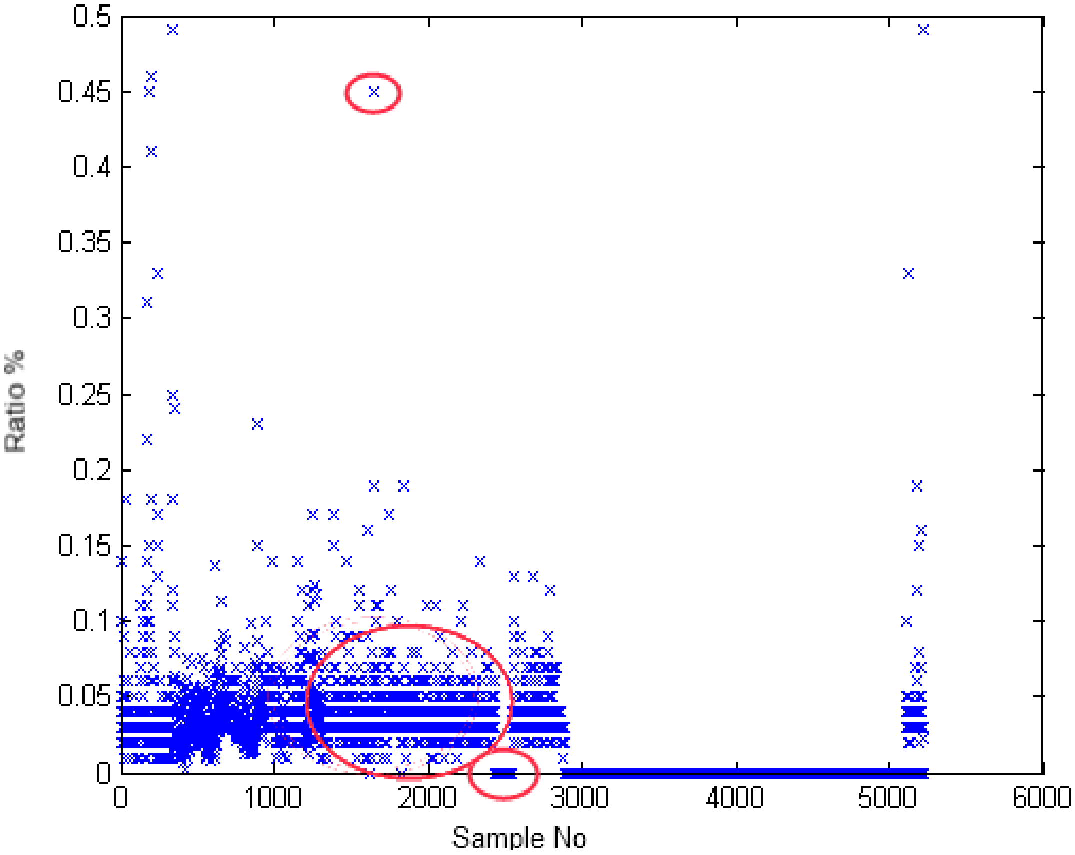


Figure 20 An example of a variable plotted separately and the three most commonly occurring problems: outliers, missing values and rounding errors.

The zeros in the data were concluded to be originated from missing analysis (Fjäder, personal communication, 2009); a problem mainly referring to variables considered to be of low interest. A few variables also contained values set to minus one. These were concluded to be referring to analysis not reaching the limit of detection (Fjäder, personal communication, 2009).

The distributions of the separate variables were determined to be either positively or negatively skew, and may not be considered as normally distributed. In Figure 21 an example of the distributions in the dataset is displayed.

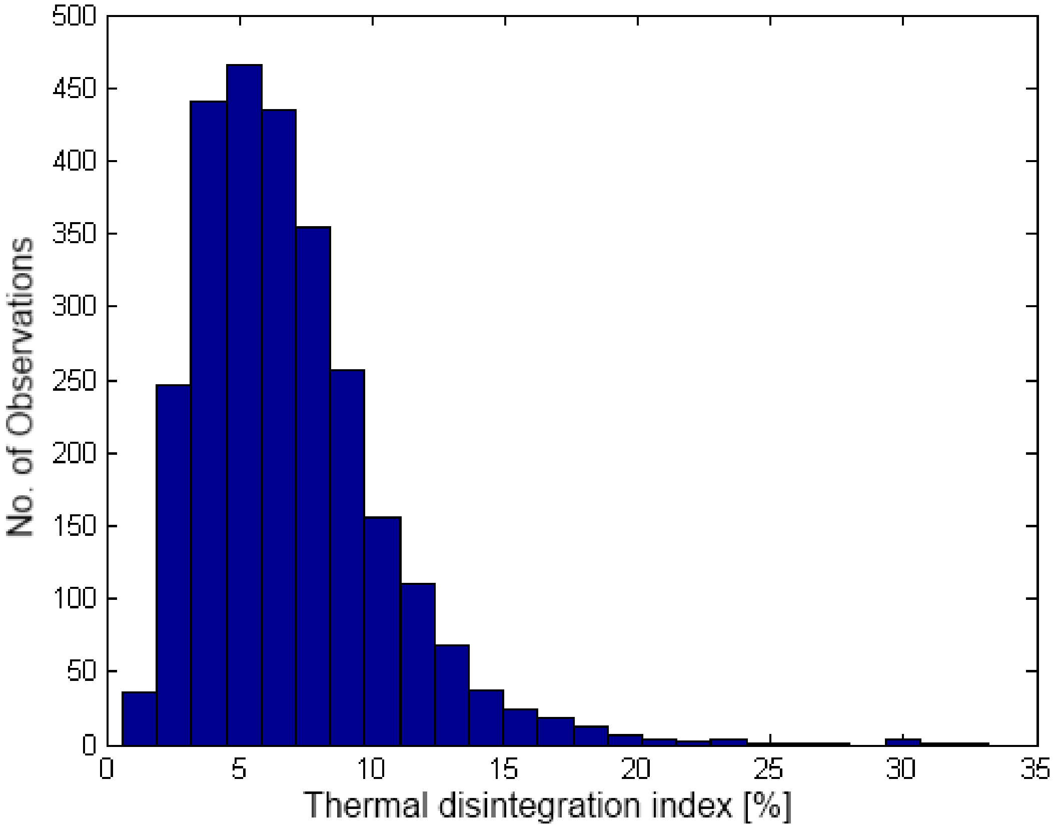


Figure 21 The positive skew distribution of the thermal disintegration index.

3.2 DATA PREPERATION

The immediate problems with the data were presented as: outliers, rounding errors, missing analysis and non-normal distribution. Since the PCA is a powerful tool to identify outliers, these were not dealt with in the data preparation, but noticed. The rounding errors were accepted as an existing problem, which had to be considered when analysing the results. To prevent the zeros and the minus ones from affecting the analysis, they were replaced by missing values.

Since the variables were to be used in PCA and PLS, which are considered to be robust towards deviations from normality, transformations to improve normality were, strictly speaking, not needed. However, logarithmic transformations were in some analysis used to shift the variables to appear more normal.

3.3 DATA STRUCTURE

This modelling part forms the foundation in the efforts to predict the thermal disintegration index and sulphur content from the geochemical data. While predicting sulphur from the drill cuttings may be considered quite trivial, thermal disintegration index on the other hand proved more complex. Perhaps not so surprising since thermal disintegration index is a geophysical property, which may not have any correlation with the geochemical composition.

3.3.1 Data Overview

The preliminary step to obtain an overview was to calculate an overall PCA model of the entire diamond core dataset. To start up the modelling the data was imported to SIMCA-P, which was used to calculate all the models based on multivariate analysis. A first PCA model was computed for diamond core dataset and several outliers were now excluded.

3.3.2 Levelled Data Overview

Since the data is distributed over three dimensions the data was sorted according to depth and modelled by PCA separately for each level. Outliers were excluded continuously during the modelling, but cautiously not to overfit the models. The layerwise models were compared to the other layers and to the overall model.

3.3.3 Transformed PCA

Since all variables were determined non-normal, logarithmic transformations with an offset were applied.

3.4 PLS PREDICTION

In order to predict the thermal disintegration index from the drill cuttings data a PLS model predicting thermal disintegration index from the diamond core data was computed. This was followed by an attempt to establish a connection between the drill cuttings data and the diamond core data, which would enable a possibility to predict thermal disintegration index directly from the drill cuttings.

Since no correlation was found between thermal disintegration index and the sulphur content, a separate model was calculated to predict sulphur. No model from drill cuttings to diamond core data was needed for sulphur since it is measured in the drill cuttings.

The PLS modelling was conducted both layerwise and for the entire diamond dataset. Sulphur and thermal disintegration index were modelled. In the dataset every fourth observation was excluded and used as validation set. A linear model was determined and then two different transformation approaches were used: first logarithmic as in section 3.3.3, and secondly by trying to determine a non-linear relation between thermal disintegration index and each variable separately, and inverting it.

3.5 CORRELATION OF DIAMOND CORE DATA AND DRILL CUTTINGS DATA

Analysis to establish a connection between drill cuttings data and diamond drill core data was conducted. Initially the measurements from the diamond samples were linked to the corresponding measurements from the drill cuttings analysis. The observations were sorted by X direction, then by Y direction and finally by depth from surface. A program connecting the observations by position and depth was written in Matlab. From the more than 3000 initial diamond observations and 5000 drill cuttings observations only 25 matching locations were found. These were imported to SIMCA-P and modelled.

3.6 SPATIAL INTERPOLATION

By representing the data from the field in just a few principal components it is possible to obtain an overview of the field characteristics, presuming that the components are chosen carefully.

An area with high spatial sample frequency was chosen as model data. The chosen area forms a rectangular field starting at point (3,000; 10,600) and ending in (3,750; 11,000) in the local spatial coordinates.

A two component PCA model was constructed from the geochemical data of field A and the score vectors were exported together with their spatial reference to MATLAB. The interpolation method is sensitive to non-normality. This was dealt with by logarithmically transforming the input data before determining the PCA. The scores, which could be considered as the output data from the PCA showed near normal distribution.

The semivariograms of the two PCs were calculated from the observations. Differences in variance occurring due to anisotropy were not considered, since the sample intervals were to large to support the analysis, i.e. isotropy was assumed.

Exponential models were fitted to the semivariograms for both PCs. These models were later used in the calculations of the weights in the Kriging interpolation (Table 1).

Table 1 Semivariance models used in Kriging interpolations

PC	Model equation	θ
1	$\gamma(h;\theta)=\begin{cases} 0, & h=0 \\ c_0+c_1\left(1-e^{-\|h\|/a}\right) & h\neq 0 \end{cases}$	c_0 =1,9 c_1=4,3 a=160
2	$\gamma(h;\theta)=\begin{cases} 0, & h=0 \\ c_0+c_1\left(1-e^{-\|h\|/a}\right) & h\neq 0 \end{cases}$	c_0 =1,07 c_1=0,32 a=200

The results from the Kriging interpolation were compared by terms of mean squared error (MSE) to an inverse distance weighted interpolation. About one fifth of the scores were excluded from the interpolations to serve as validation.

The original data had the spatial frequency of 50 m in X-direction and 100 m in Y direction, and had the total length of 850 m in X-direction and 400 m in Y-direction. When the 15 validation points had been excluded each depth layer contained 65 observations which were used in the interpolation. The interpolations were carried out with 170 points in X-direction and 80 points in Y-direction; a total of 13,600 points per layer.

As a comparison the more trivial inverse distance weighted interpolation method was used. The weights were calculated as the inverse squared distance from the evaluated point. Points within a distance of 200 meters were included in the interpolation.

4 RESULTS

4.1 DATA STRUCTURE

4.1.1 Diamond Core Data Overview

The first PCA model, which was calculated for the entire diamond core dataset, is presented by score plot (Figure 22) and loading plot (). The model was constructed by two principal components; together representing almost 76 % of the variance in data. The model showed a predictive power of about 62 %.

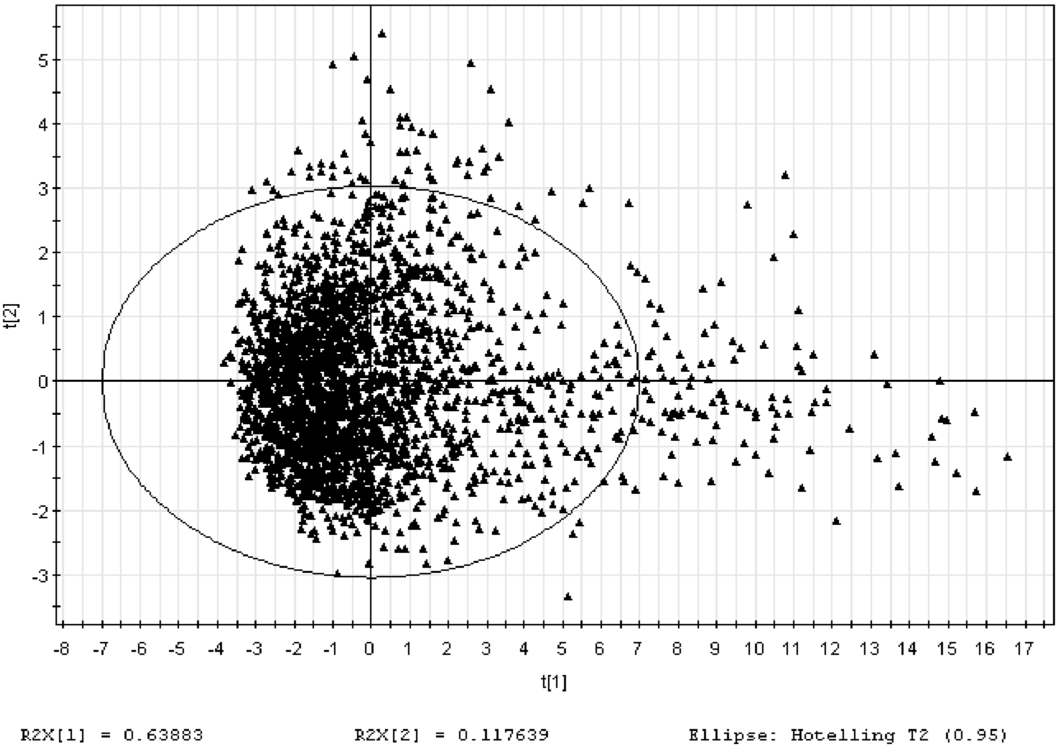


Figure 22 The score plot of the diamond core dataset.

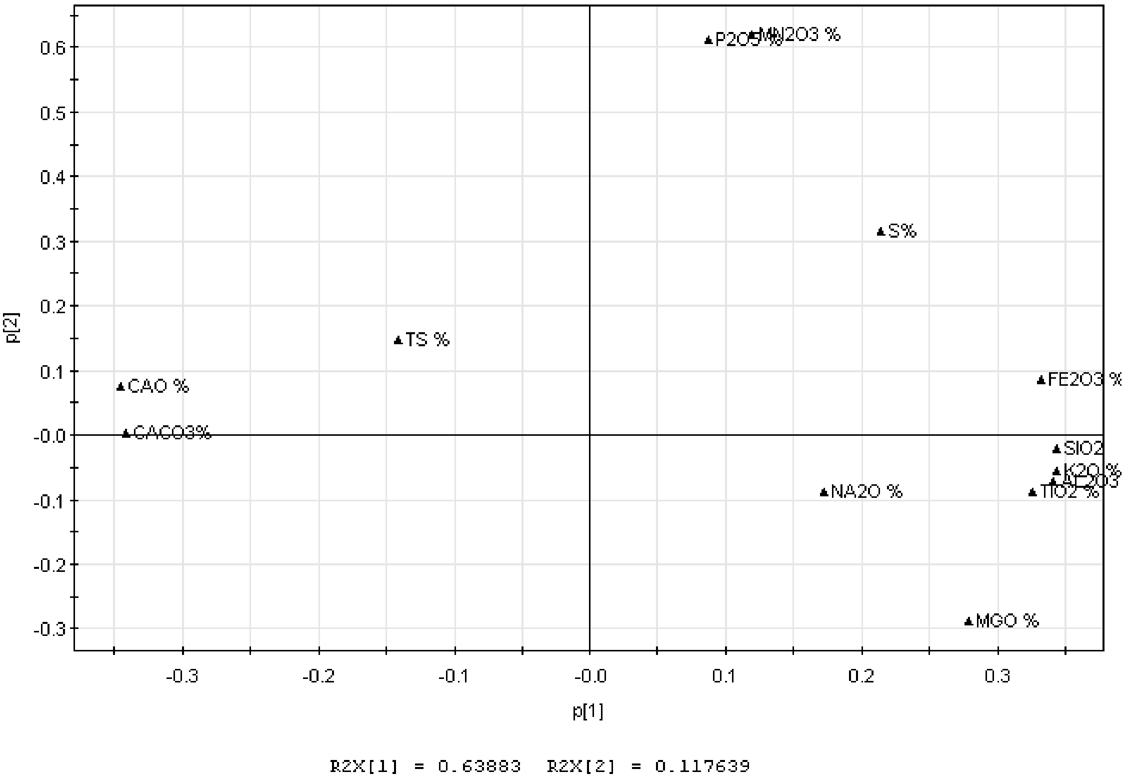


Figure 23 Loading plot of the diamond core dataset

It is possible to see that the first component is strongly influenced by the relations between CaO, $CaCO_3$ and SiO_2, K_2O, Al_2O_3, Fe_2O_3, TiO_2, which are two groups, internally correlated and negatively correlated to each other. The score plot shows many deviating observations associated with high values in the second group mentioned previously.

The dominating variables to the second PC are P_2O_5 and Mn_2O_3, which also seem closely related. Observations deviating in the second PC direction often seemed to be connected to high values in these two compounds.

Sulphur and thermal disintegration index, which are of interest to control in the production does not seem to have any close correlations to other variables. The loading plot also shows thermal disintegration index, TS, to be situated quite close to the origin. The explanatory ratio of each variable is visualised in Figure 24, only about 20 % of thermal disintegration index was explained by the model.

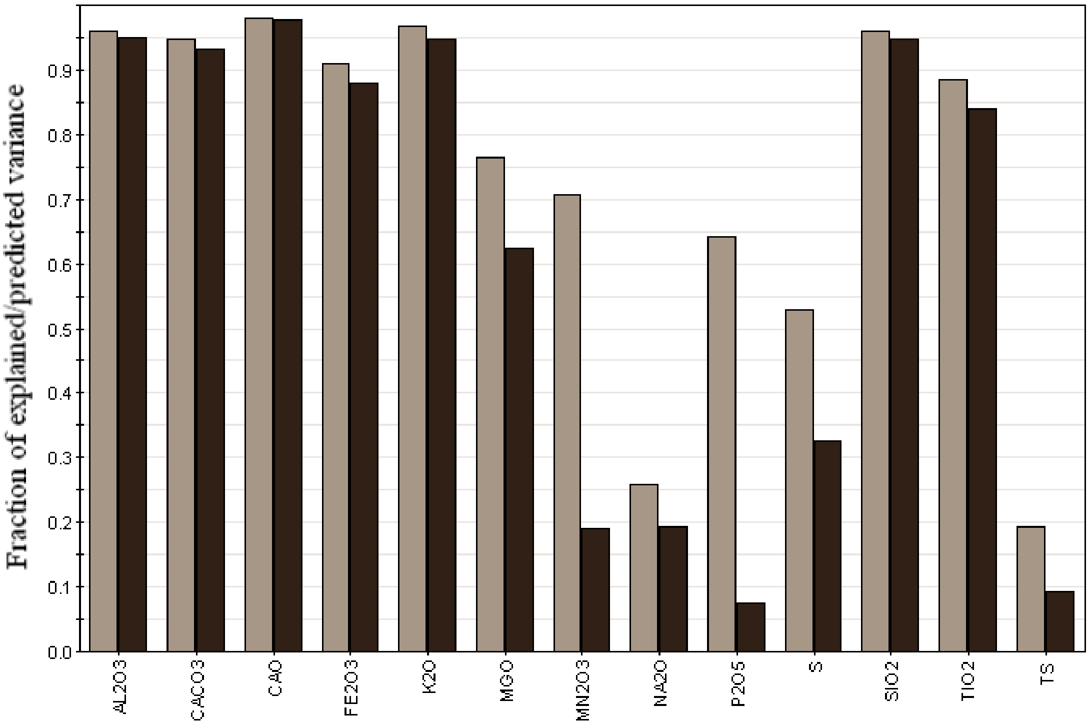


Figure 24 The fraction of explained variance for each variable calculated from the residual size and the fraction of predicted variance calculated from an external validation dataset.

Figure 24 shows that the variables closely related to the first PC were modelled and predicted well. The variables, which were described by the second component, although in some cases well fitted, seemed difficult to predict.

4.1.2 Levelled PCA of Diamond Core Data

The topmost two or three layers differed slightly from the deeper layers in structure, mainly related to the second PC. Relations between Mn_2O_3, P_2O_5 and sulphur varied in particular (Figure 25 & Figure 26). In general though, all PCAs much resembled the overview PCA of the entire set.

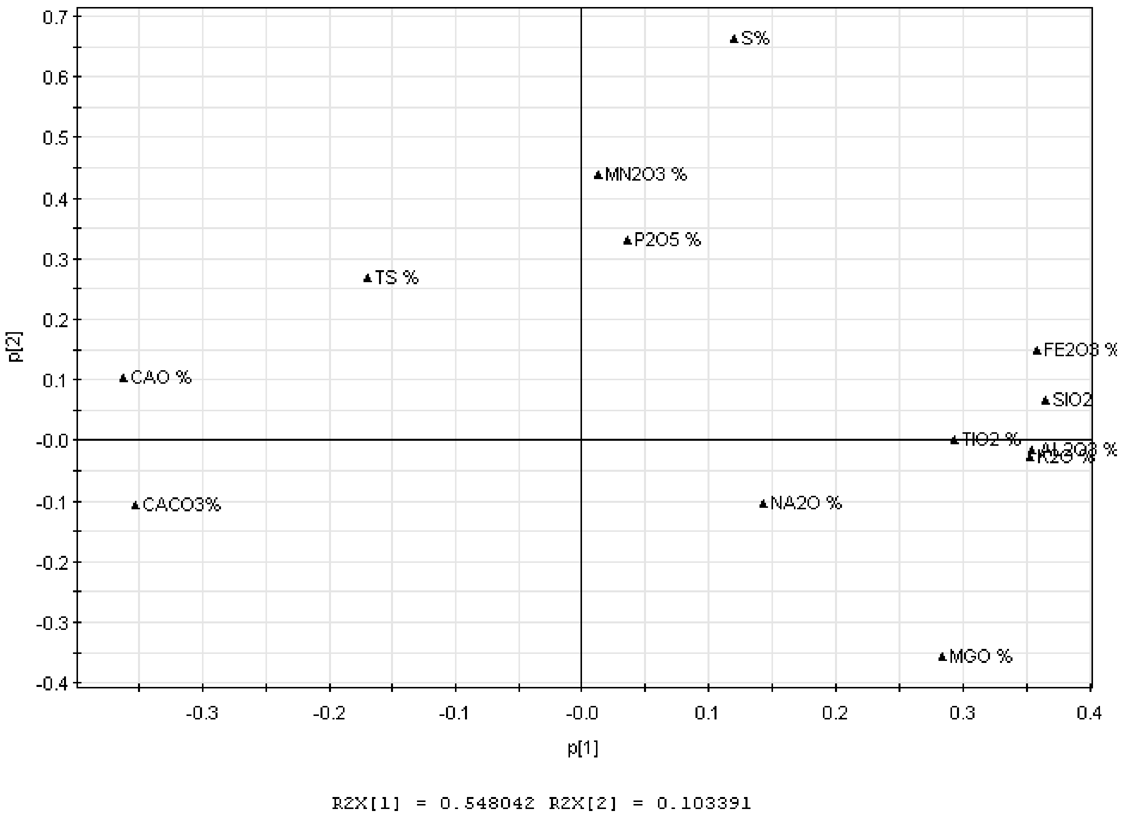


Figure 25 The loading plots of layer 1.

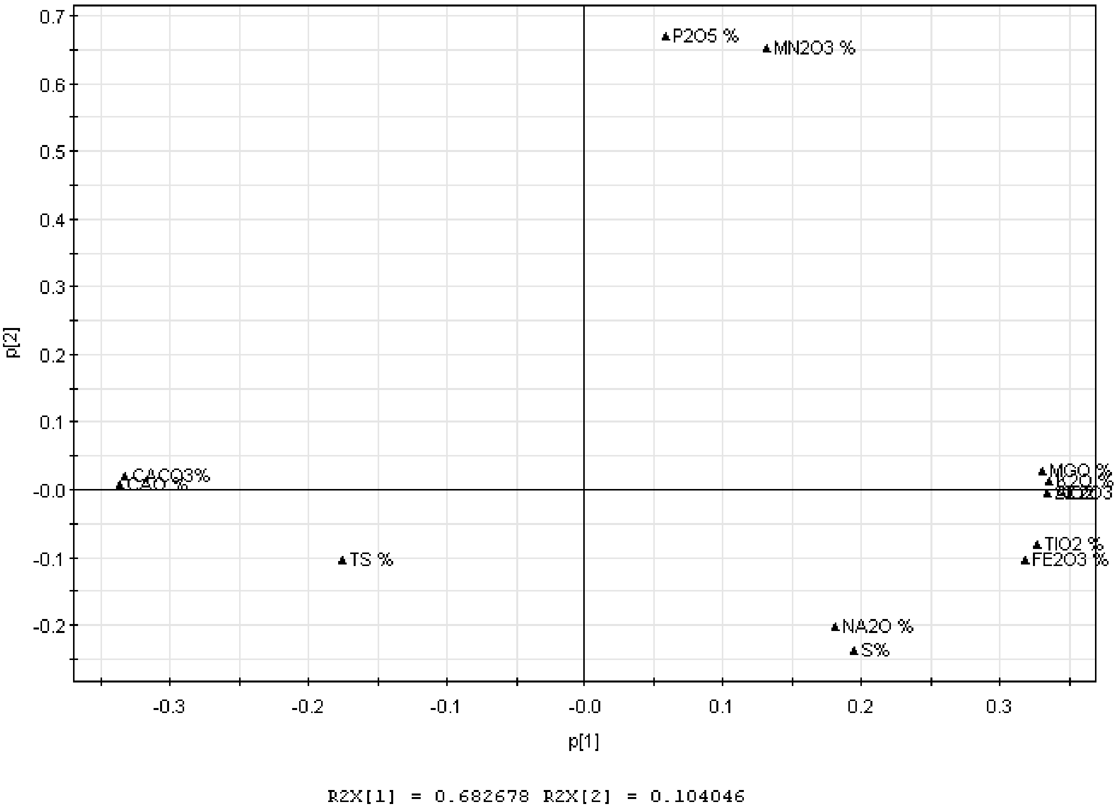


Figure 26 Loading plot of layer 8.

Both the explained variance ratio and the predictive power increased to some extent towards deeper levels (Table 2). The layers are numbered from one starting with topmost layer.

Table 2 The variance explained and validated by the models for each layer

Layer	Explained variance [%]	Predictive power [%]	Notes
1	65.1	44.5	
2	70.9	52.2	
3	70.5	51.9	
4	76.5	62.9	
5	76.3	64.3	
6	76.0	61.0	
7	79.5	64.9	
8	78.7	64.5	
9	78.8	62.7	
10	80.8	59.5	Few values
11	91.2	64.2	Few values, different structure.

The variables thermal disintegration index and sulphur, which are to be predicted in later chapters by PLS, naturally are of high interest. How well the model for each layer explains these characteristics may be viewed in Figure 27. In Figure 28 the goodness of prediction for the layer models is shown. Similar plots for all variables and layers may be viewed in Appendix I.

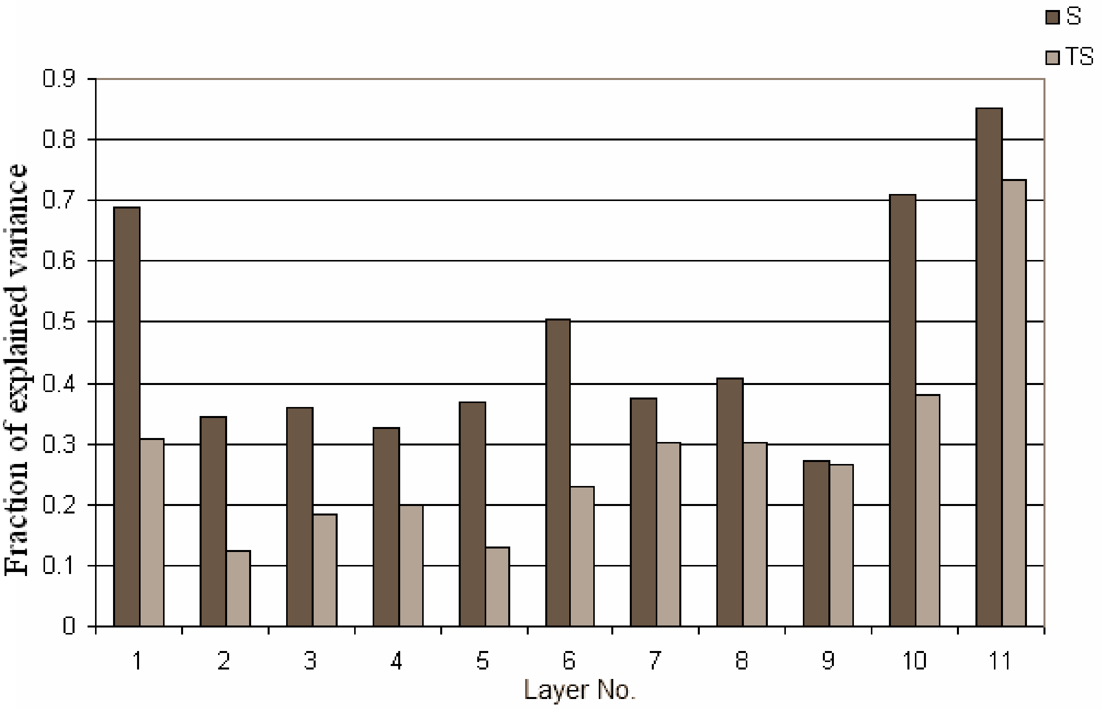


Figure 27 The explained variance fractions for sulphur and thermal disintegration index shown layer by layer.

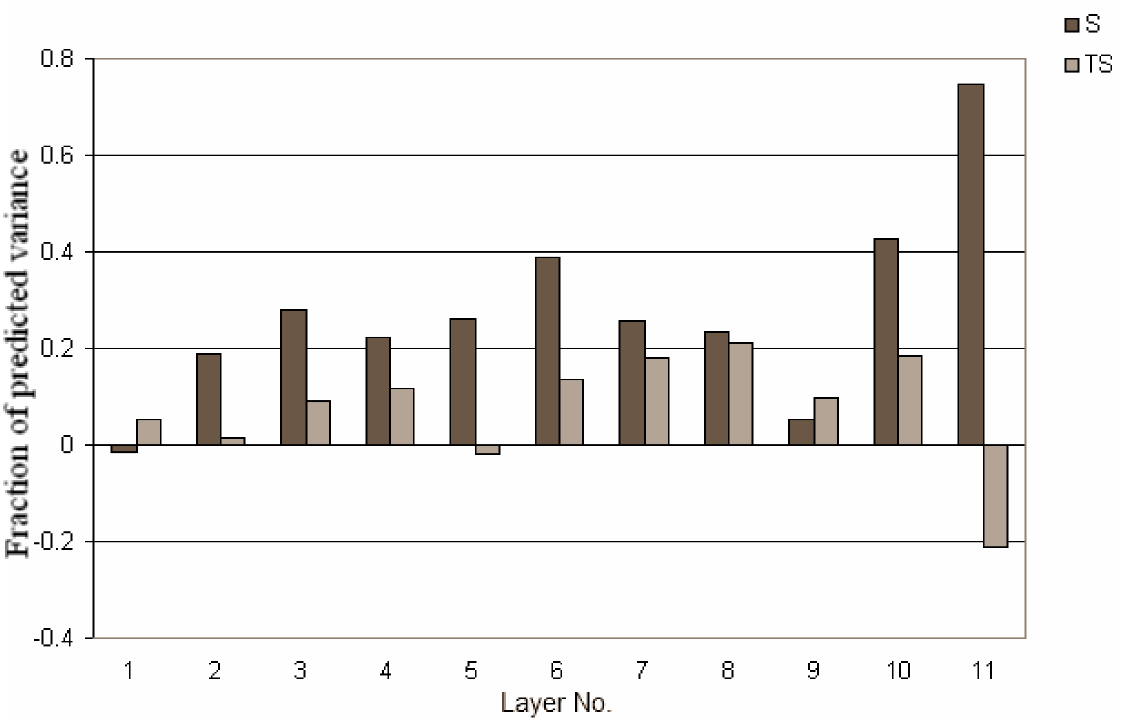


Figure 28 The predicted variance fractions for sulphur (S) and thermal disintegration index (TS) shown layer by layer.

4.1.3 Transformed PCA

No modelling improvements were revealed, with the applied transformation.

4.2 PLS PREDICTION

4.2.1 Prediction from Entire Diamond Core Dataset

A first attempt was made to create a model from the entire diamond core data, which predicted sulphur and thermal disintegration index at the same time. This model was modelled from centered and standardised data with no transformations. As mentioned earlier the data was not normally distributed, which is not necessary but desirable. The model consisted of five components describing 94 percent of the variance in the **X** matrix, while about 53 percent of the variance in the response matrix was accounted for. Sulphur was predicted quite well by the model, whereas thermal disintegration index was poorly predicted. Figure 29 shows how well the model predicts thermal disintegration index when compared to an external validation set.

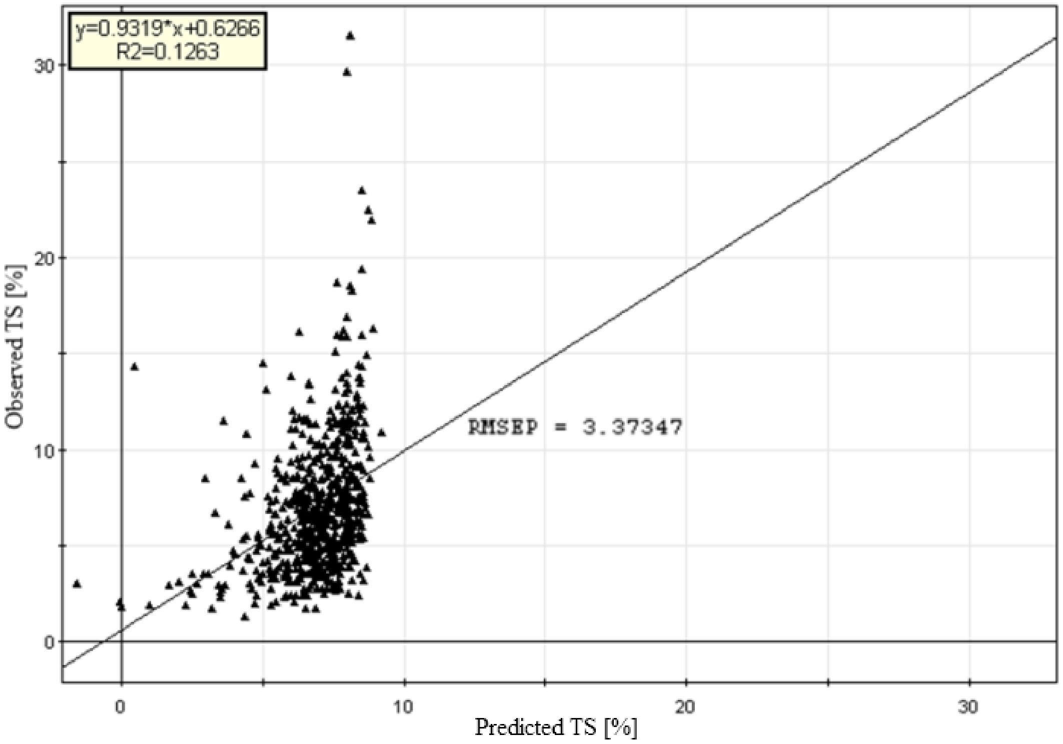


Figure 29 Thermal disintegration index (TS) predicted values plotted against observed values. The straight line represents the PLS model.

In the same manner the model predicting sulphur may be viewed in Figure 30.

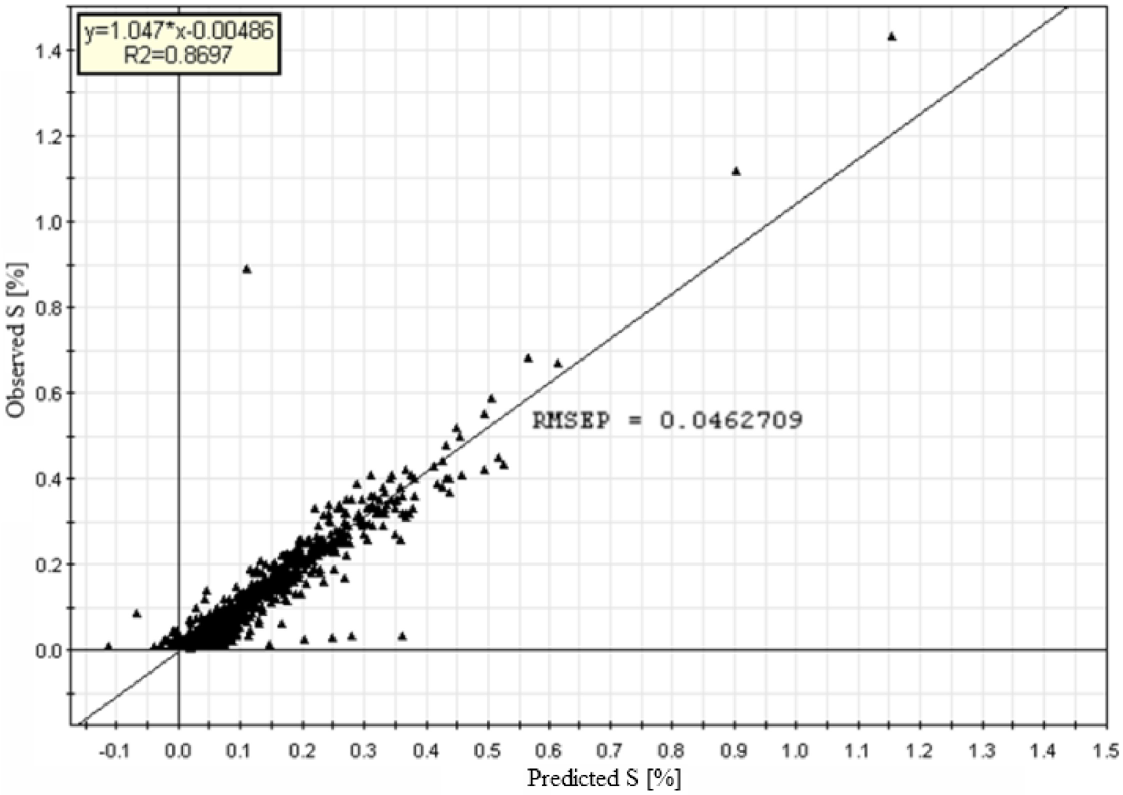


Figure 30 Predicted values for sulphur plotted against observed values. The straight line represents the PLS model.

When studying the variable importance plots and the PLS coefficients the thermal disintegration showed to be most influenced by silica, potassium and aluminium (clay minerals). High clay values pointed towards low disintegration index. Sulphur seemed strongly influenced positively by iron content.

4.2.2 Prediction from Layered Diamond Core Data

In the same manner as with the PCA further PLS models were calculated for each layer. They all showed great similarity in appearance with the models described in Figure 29. These models may be studied in detail in Appendix III. The prediction of sulphur seemed to improve when leaving the topmost layers. An example of this is visible when Figure 31 is compared to Figure 32. The prediction of thermal disintegration index on the other hand did not improve noticeably in any layer.

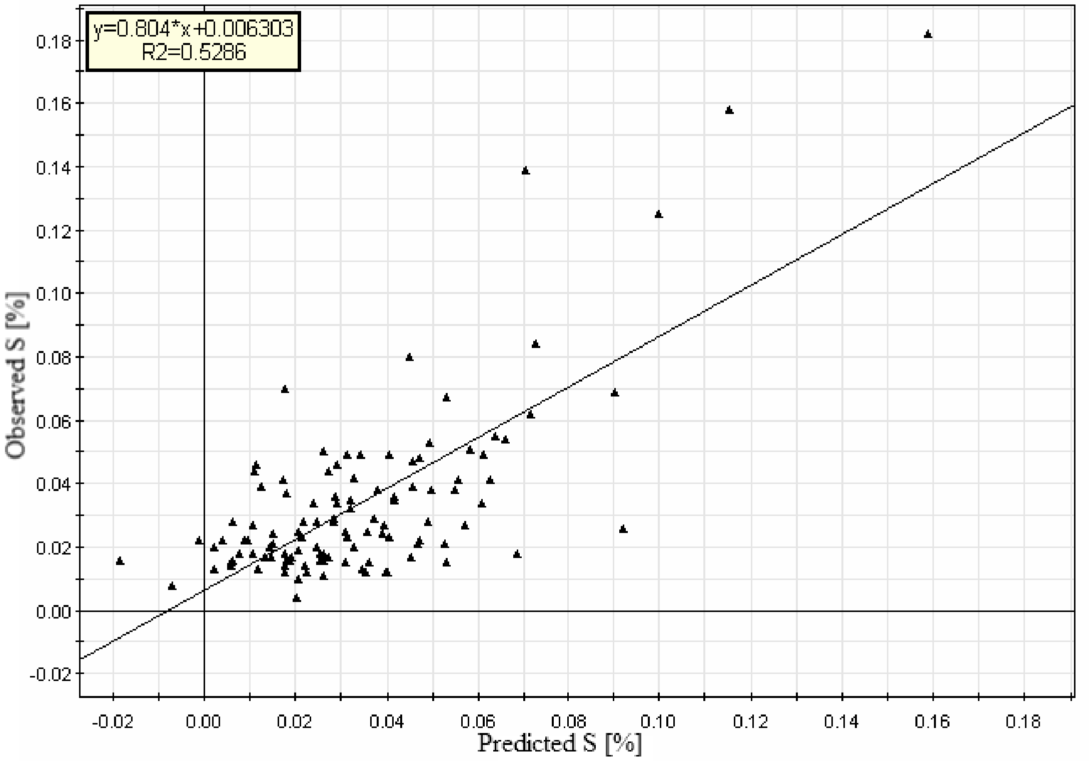


Figure 31 Predicted values of sulphur plotted against observed values for layer 1 (0-6 m).

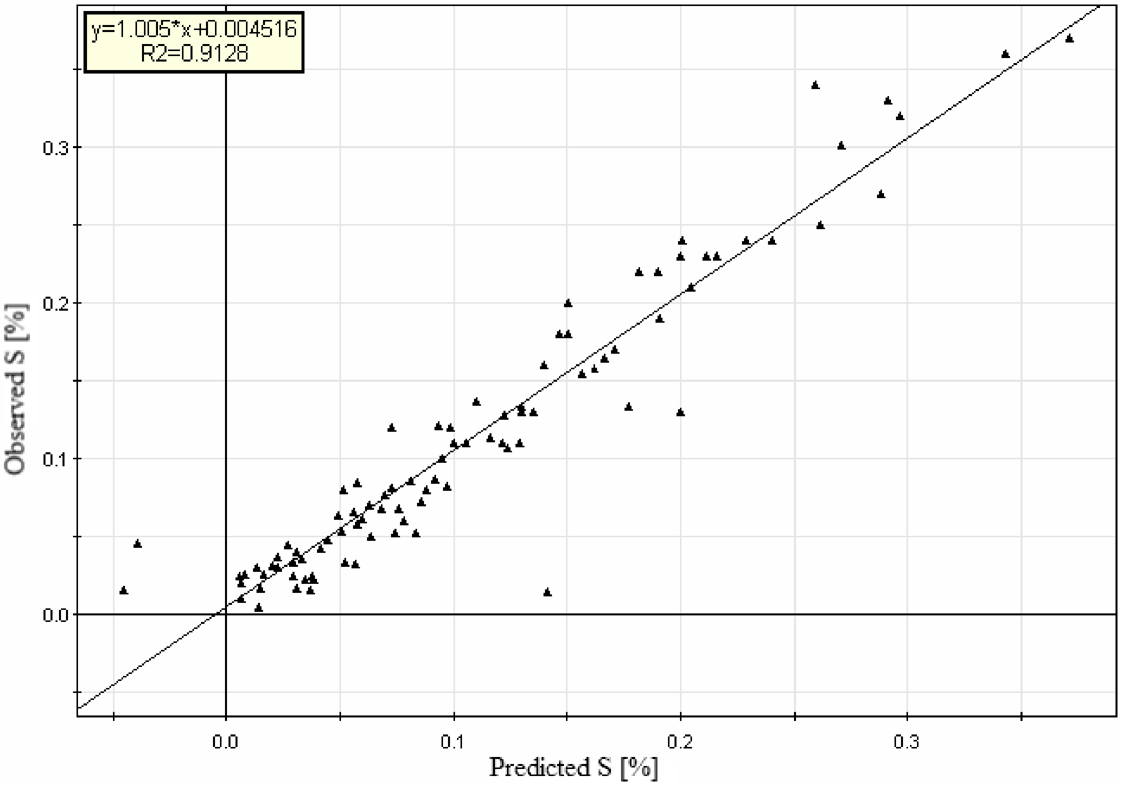


Figure 32 Predicted values of sulphur plotted against observed values for layer 4 (12-15 m).

4.2.3 Transformed Predictions

Neither of the two transformation approaches, logarithmic and linear relation to thermal disintegration index, displayed any modelling advantages. Although near normality was achieved in the first approach and improved linearity for some variables in the second, the transformations rather seemed to induce non-linearity to the models.

4.3 CORRELATION OF DIAMOND CORE DATA AND DRILL CUTTINGS DATA

The calculated PCA model consisted of four components, which explained 83 % of the variance in data. The predictive power in the model was weak, only 25 %. Figure 33 shows the loading plot of the first two components, together accounting for almost 67 percent of the variance.

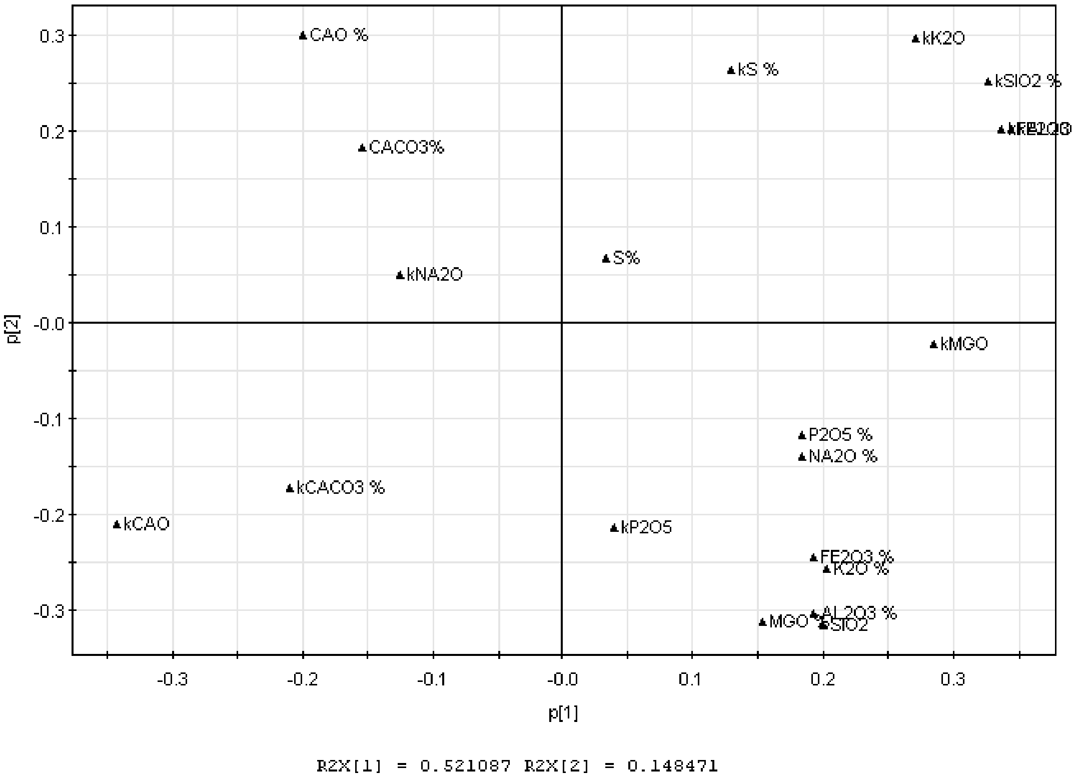


Figure 33 Loading plot of the first two components of the diamond core drill and drill cuttings dataset. Variables starting with a lower case k, indicates variables from the drill cuttings.

The observations from diamond drill core were mainly located in the second and fourth quadrant, whereas the drill cuttings observations were located in the first and third quadrant.

When examining the third and fourth component some correlation between the variables of the drill cuttings data and the diamond core data may perhaps be noticed, but since only a small part of the variance in data were described by these component it is hard to decipher what this might indicate.

The inner structures of each dataset were examined through individual PCA models fitted to the data. Both models were composed by two principal components. Figure 34 shows how the variables are related to each other in the diamond drill dataset.

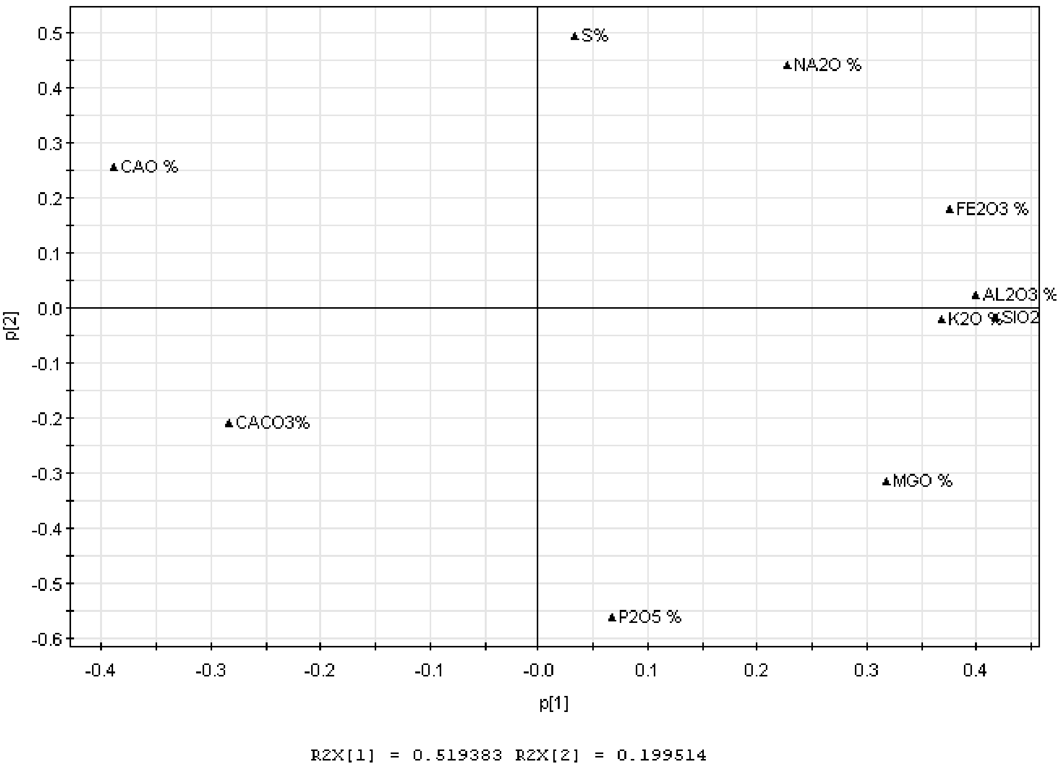


Figure 34 Loading plot of PCA model fitted to the diamond observations.

The structure of the diamond samples were compared to the structure of the drill cuttings data (Figure 35).

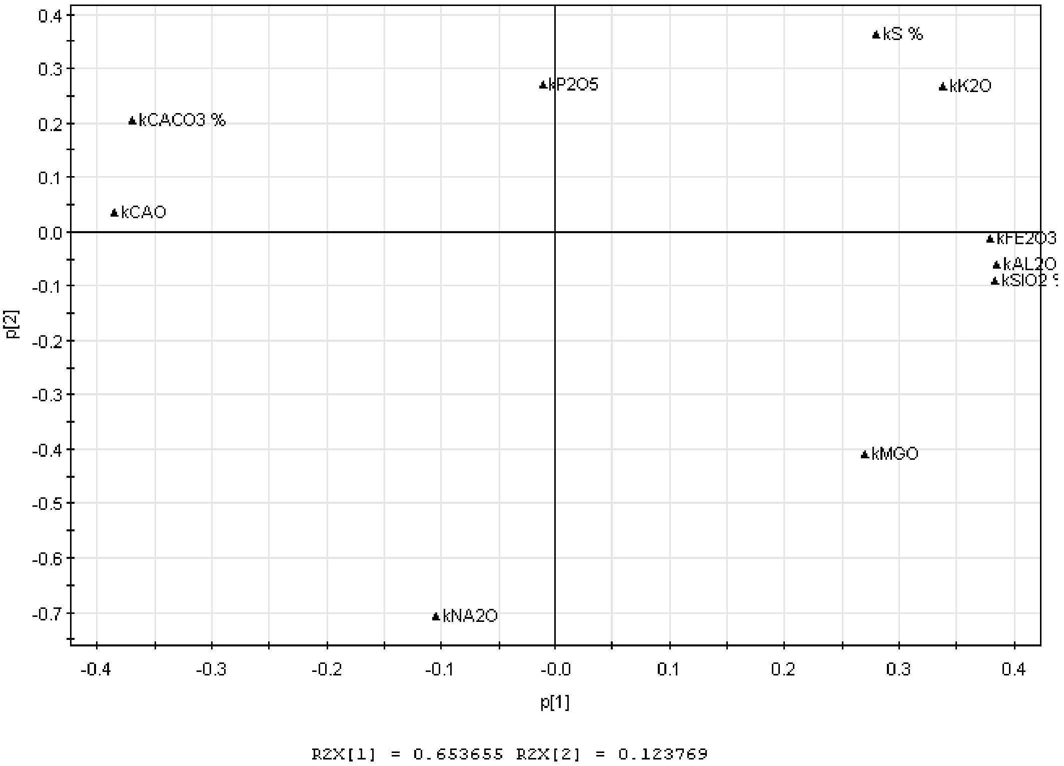


Figure 35 Loading plot of PCA model fitted to the drill cuttings observations. Observation No. 12 was excluded from the model.

The inner structure of the two sample types showed to be similar with respect to the first component. However, P_2O_5 related differently to the other components in the drill cuttings data compared to the relations in the diamond drill dataset.

It was not possible to achieve any satisfactory PLS model to predict neither of the two sample types from the other.

4.4 SPATIAL INTERPOLATION

In Figure 36 the calculated semivariogram from the first PC is shown.

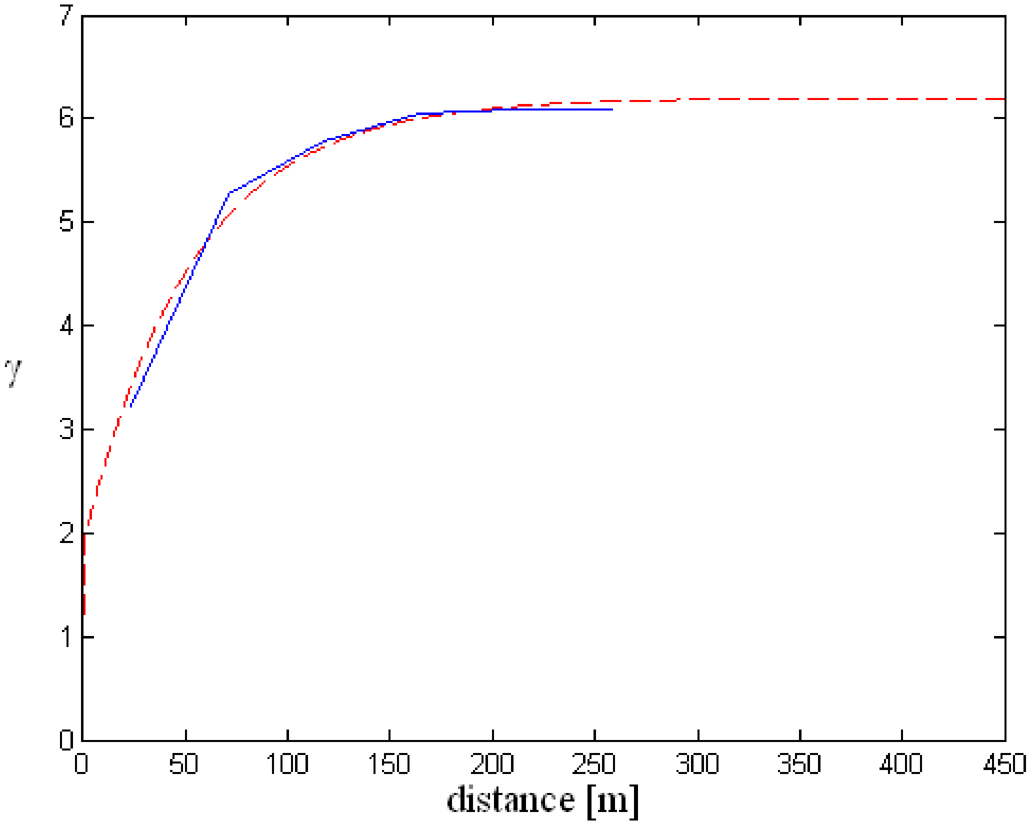


Figure 36 Semivariogram from scores in the first PC. The solid line shows the experimental variance; the dashed line is the fitted model.

Since the original samples were received as mean value of an entire 3 meter core column, a three dimensional interpolation was not possible. It allowed however to interpolate the horizontal spread. In Table 3 the mean squared errors are displayed and may be compared to the observed variance in the field.

Table 3 shows how the error varies greatly depending on the layer and that the two methods were on most occasions quite similar in error. In most cases, the error of the models were less than the observed variability in the field.

Table 3 The mean squared errors (MSE) and the observed variance for each interpolation technique, PC and layer

Layer	PC	MSE (Kriging)	MSE (idw)	Observed variance
1	1	1.80	2.26	2.00
	2	0.21	0.21	0.31
2	1	1.83	1.66	3.70
	2	0.49	0.43	0.54
3	1	3.53	3.48	5.18
	2	0.59	0.61	0.43
4	1	5.60	5.92	6.45
	2	0.30	0.22	0.78
5	1	2.51	2.39	5.46
	2	0.35	0.30	0.72
6	1	2.99	2.76	4.83
	2	0.48	0.52	0.42
7	1	6.62	6.58	9.83
	2	0.66	0.57	0.57
8	1	5.33	4.47	9.25
	2	0.66	0.56	0.65
9	1	5.40	5.14	6.58
	2	1.07	1.10	0.74

Every interpolation contains uncertainty, which could be considered as the real value variability from the interpolated value. As the distance from the observed location increases the uncertainty will also be larger. In Figure 37 it is visualized how the uncertainty or Kriging error varies over the interpolated field.

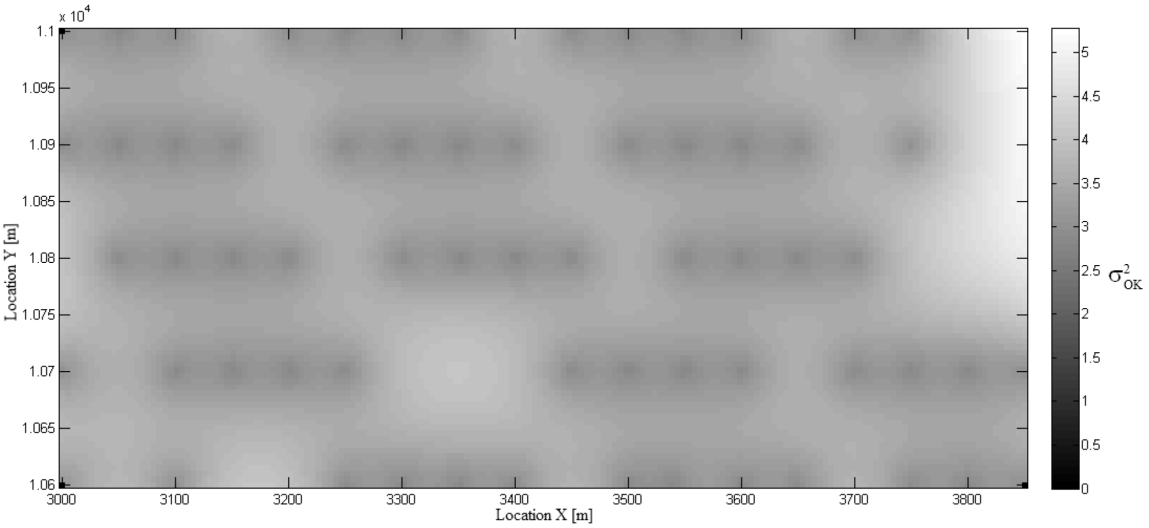


Figure 37 The Kriging error of layer 5.

It is clearly noticeable how the uncertainty is significantly smaller closer to the observed locations. Further on, from Figure 37 it is easy to see where the validation locations have been excluded and how missing observations in the upper right corner affects the interpolation variability; an area which in this case actually could be considered as being extrapolated. It is important to remember that the Kriging error calculation demand accuracy in the semivariogram to be reliable.

An example of what the interpolated field looks like is displayed in Figure 38.

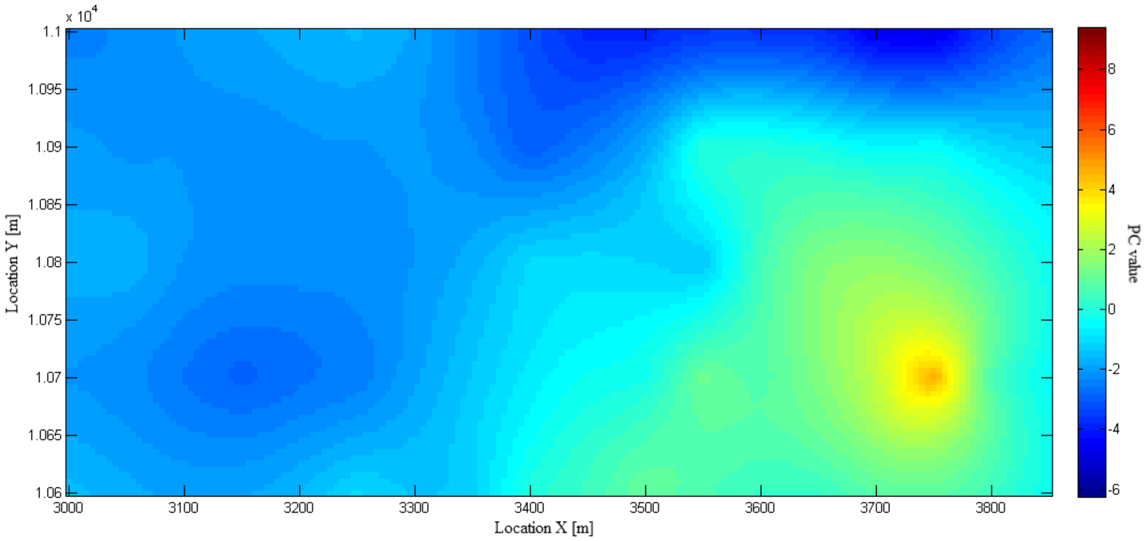


Figure 38 Kriging interpolation of layer 5 in the field. Red indicates the abundance of minerals connected to clay presence.

As to illustrate and emphasise the differences between Kriging and Inverse Distance Weighted interpolation the same layer is shown in Figure 39 with IDW interpolation. Notice the more drastic changes.

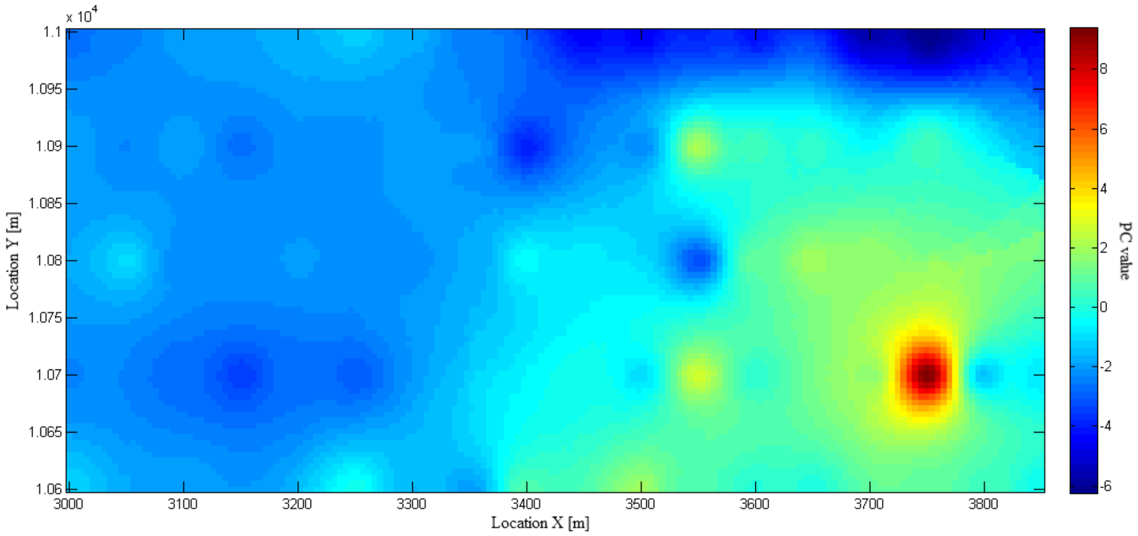


Figure 39 IDW interpolation of layer 5.

5 DISCUSSION

5.1 DATA PREPARATION

Many suspected outliers were found when examining the data one variable at the time. The main problem however, was to identify if the outlier had support from surrounding measurements. Perhaps this would have been easier to see if values were plotted as images layer by layer with accurate spatial relations. That would not include the surrounding samples in the third dimension, and therefore still exclude valuable information.

Clarifying if a whole observation or only a single variable in the observation were deviating was not possible when data was examined as series or images. The outliers were ultimately left to be dealt with in the PCA, which has the advantage of being able to identify both deviations in entire observations and in single variables. However, the spatial problem is not solved by PCA.

The organised patterns discovered in data, occurring in variables present in low quantities, were most likely originated from rounding and the limit of detection. This may be due to either the limit of detection of analysing instruments or the measurement take down by operator. With data stretching over a time period of over 30 years this may not be surprising although still important to avoid in the future, since the inaccuracies in data may inflict difficulties when modelling or analysing the data.

Obviously during periods of the quarry operation some variables were not considered of importance; resulting in large areas of missing values. The consequence is that less information is included in the analysis.

Cressie (1991) mentions that normal distribution is generally uncommon when analysing geochemical data. A possible explanation in limestone may be the composure of the analysed volume. The variability in different limestone types may cause the disturbance. Perhaps, the dominating composures are normally distributed, and smaller fractions of differently composed rocks shift the overall distribution from normality. An idea would be to consider samples originating from the same rock type separately to reduce the variability in data.

5.2 DATA STRUCTURE

The skew distributions observed during the univariate analysis were clearly visible in the PCA models. Deviating observations were mainly concentrated around the first principal component and inclining towards high positive values. This was probably related to the clay content in the sample. From the loading plots it was apparent that the clay minerals and calcium are negatively correlated, which indicates that these components share space.

Phosphorous and manganese showed to have different relations to the other components, possibly due to the low amounts of these compounds in the rock, and the difficulties with measurements close to the limit of detection. Another explanation could be the crystal structure. It is known that sulphur for example is bound inside the calcite crystal, whereas clay minerals are rather situated in the fractures. Whatever underlying reason which separates the two components, phosphorous and manganese are often present together.

Of special interest during the analysis was to establish a relation between the physical characteristic thermal disintegration index and the chemical composition. However from the PCAs thermal disintegration seemed to have little in common with other components. The variable residual plots also proved this to be the case. Probably the chemical composition of the rock has very little to do with its resistance to disintegration at high temperatures. To better explain the mechanisms involved the data needs to be complemented with other parameters e.g. grain size etc..

When the analysis was continued to include layerwise PCA models a change in structure was noticed. Mainly sulphur, phosphorous and manganese were involved, which could be explained by a number of reasons. Perhaps the most plausible explanation may be that this was caused by the stratified nature of the rock. This may also be the explanation to the somewhat varying modelling results. One layer may be much more homogene compared to another depending on the ratio of samples in the layer originated from the same rock type. Since the present observations serve as an average value over a three meter drill core or three meters worth of drill cuttings the data becomes harder to characterise and valuable information may be lost.

Another thought was that the differences may be caused by the age of the limestone or that the rock at the topmost twelve meters or so, which deviated in structure from the deeper layers, were perhaps formed in a different manner, under different conditions or in a different environment. Since limestone is originated from coral reefs it has to be considered that changes in the surroundings may affect the rock.

5.3 PLS PREDICTION

The predictive part of this thesis could be divided into two different categories, of course with connection to the two distinct responses, sulphur and thermal disintegration index. Category number one would involve finding a relationship between chemical composure and a physical characteristic. Category two involves determining a chemical component, with a slightly different behaviour, from the presence of other components.

In the results presented in section 4.2.1 and 4.2.2 the thermal disintegration index showed not to be easily predicted from the chemical data. As suspected from the earlier PCA models, the lack of correlation became a problem during the PLS modelling. In the prediction plot in Figure 29 the thermal disintegration index appeared as if it may have some non-linear relation to the chemical components. Even though several different transformations were tried, nothing seemed to improve the relation significantly. The transformations rather induced non-linear behaviour in other variables and even shifted the models further from normality.

The work with the PLS model for thermal disintegration index revealed that it may be connected to aluminium, potassium and silica, which are mainly related to clay presence. This could indicate that it is grain size or the rock structure rather than the chemical structure that affects the resistance to disintegration. Though it may seem trivial, a model to predict sulphur was also calculated and proved to be rather good. However, as with any model derived from sample data, some deviations were visible. It was interesting to see that the models noticeably improved when modelled in deeper layers, a phenomena witnessed in PCA modelling as well. The explanation could be the same as discussed earlier with stratified rock and better representation in the samples depending on the rock type distribution in the layers. Perhaps the exceptionally poor prediction result from the top layer (0-6m), shown in Figure 31, may be due to the effect of exposure to the atmosphere.

5.4 CORRELATION OF DIAMOND CORE DATA AND DRILL CUTTINGS DATA

Correlation between the diamond drill core analysis and the drill cuttings analysis was not found. It should be remembered that since only 25 overlapping sample locations existed in the data, the analysis may serve at most as an indication of the difficulties involved when evaluating the sample results.

In the model the principal components of highest variance seemed to be directed somewhere in between the variables of the drill cuttings and the diamond core samples. This shows that there are correlations between the two sample methods but also significant differences causing difficulties when trying to predict one from the other. Most likely the distinctions between the datasets were associated with the differences between sampling methods. Apparently these dissimilarities inflict noise on the system, which overshadows the correlations. It may be possible that the second component describes the distinctions between the drilling methods rather than correlation of the variables.

5.5 SPATIAL INTERPOLATION

A Kriging interpolation of the PCA scores was studied. The idea was to enable an overview of the field in one or maybe two variables only. This study should be considered as an evaluation of the method and its possibilities with the data at hand.

Since the Kriging interpolation is based on the semivariogram, this introduces the first problem. In order to get an accurate semivariogram the data must be normally distributed and it is important that the samples include the changes in variability. This exhibits the need of careful planning when deciding how to distribute the observations spatially. In the data at hand the observations were often too wide apart. Only in the X direction could any actual variation change be witnessed.

The second problem relates back once again to the semivariogram but is also closely related to the nature of the limestone. Kriging interpolation is based on how much the observations are thought to affect the interpolated point depending on their distance to it. However, since the limestone is stratified this could result in large sudden changes in structure over short distance. Should this be true the interpolation will deliver a poor result, whereas if it is not, the method is likely to work well. All in all these circumstances damages the reliability of the method.

The results in this case, which were compared to the most commonly used spatial interpolation method, IDW, showed Kriging neither to be better nor worse. In most cases the mean squared error was about the same. It should be remembered though that an enhanced semivariogram could perhaps improve the interpolation.

It was at first desired to carry out the interpolation in three dimensions, but since the data was sampled from the entire three meter column of diamond core this left no possibility for interpolation in depth. Could this have been considered as discrete data, and not as mean over three meters, the semivariogram and model could perhaps have been improved and expanded to three dimensions.

6 CONCLUSIONS

6.1 DATA OVERVIEW AND PCA

The varying data quality, due to the long time period of collection, induces problems in the analysis. This is a complication hard to address retrospectively, hence it is important to take this into account in future measurements.

Principal component analysis extracts information from the data efficiently and in this case emphasizes the importance of experimental design. In future exploring of limestone areas a more discrete manner of sample collection is advisable. The present sample method of three meter measurements should be removed in favour of point observations with as accurate spatial reference as possible. Preferably in an acknowledged coordinate system. Analysis taken at both predestined depths and in accordance to the variations of the rock type would enable more precise models and less skew distributions in the variables.

6.2 PLS PREDICTION

The predicting model was able to predict sulphur from the other components with rather good accuracy. The most important variable in the prediction was iron, which perhaps could serve as an indicator or guide when fast analysis are desired. Further analysis may be needed to develop this possibility. Improvements concerning measurement techniques mentioned earlier may enhance the results.

In future analysis of the thermal disintegration, the physical factors may not be ignored, but rather emphasised in importance if reliable models are desired. Obviously some important information was not represented in the data, which caused the models to fail, when trying to predict the thermal disintegration. Results indicating that high clay content had a positive effect on the disintegration suggest that small grain size is desirable.

6.3 CORRELATION OF DIAMOND CORE DATA AND DRILL CUTTINGS DATA

Increased sample overlapping and enhanced sampling methods are needed in order to deal with the noise complication. A more discrete sampling method would probably also improve the results.

6.4 SPATIAL INTERPOLATION

The stratified rock causes trouble for the interpolations, thus more suitable sample collection or more complex interpolation techniques, perhaps considering geophysical data as well, would be needed to reach desired reliability.

If a Kriging interpolation is to be used, the sample spatial frequency should be given the outmost attention in order to obtain an accurate semivariogram. The data used in this thesis was not distributed in a manner, which allowed a satisfactory semivariogram to be determined.

The results in the interpolation analysis showed that there is little to gain from performing the Kriging interpolation in favour of the less complex inverse distance weighted interpolation when faced with this kind of data.

7 REFERENCES

7.1 BIBLIOGRAPHY

Abdi, H., (2003) “Partial least squares (PLS) regression”, *The Sage Encyclopedia of Social Science Research Methods*, (2003), 792–795.

Almeida, J., Rocha, M., Teixeira, A., (2005). “Spatial characterization of limestone and marl quality in a quarry for cement manufacturing”, *Geostatistic Banff* (2004), 399-408.

Bohling, G., (2005). “Kriging”, unpublished, Kansas Geological Survey, http://people.ku.edu/~gbohling/cpe940/Kriging.pdf

Björk, A., (2007). “Chemometric and signal processing methods for real time monitoring and modelling using acoustic sensors, Applications in the pulp and paper industry”, Royal Institute of Technology, School of Chemical Science and Engineering, Department of Chemistry, ISSN 1654-1081.

Clark, I., (2001). “Practical Geostatistics”unpublished, Geostokos Ltd, (2001)

Cressie, N., (1991). *Statistics for Spatial Data*, John Wiley & Sons, Inc., ISBN 0-471-84336-9.

Eriksson, L., Johansson, E., Kettaneh-Wold, N. and Wold, S., (2001). *Multi- and Megavariate Data Analysis*, Umetrics AB, ISBN 91-973730-1-X.

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 (1987), 161-175.

Esbensen, K., Schönkopf, S., Midtgaard, T., Guyot, D., (1998). *Multivariate Analysis – in practice*, 3rd edition, Camo ASA, ISBN 82-993330-1-6.

Geladi, P., Kowalski, B., (1985), “Partial Least-Squares Regression: A Tutorial”, *Analytica Chimica Acta, 185,*(1985), 1-17.

Golub, G., Kahan, W., (1965), “Calculating the singular values and pseudo-inverse of a matrix”, *J. Siam Numerical Analysis,* Ser. B, Vol. 2, No. 2 (1965), 205-225.

Jimenez-Espinosa, R., Sousa, A.J., Chica-Olmo, M., (1993), “Identification of geochemical anomalies using principal component analysis and factorial Kriging analyis”, *Journal of Geochemical Exploration*, 46 (1993), 245-256.

Johnson, R., Wichern, D., (1992). *Applied Multivariate Statistical Analysis*, Prentice-Hall, Inc., ISBN 0-13-041807-2.

Jolliffe, I., (1986). *Principal Component Analysis*, Springer-Verlag, New York Inc., ISBN 0-387-96269-7.

Karlsson, J., (2008), “Efterbehandling av bergtäkter - Förslag till en hållbar efterbehandling av Klinthagentäkten”, Report at IVL, (2008).

Miller, J. & Miller J., (2000). *Statistics and Chemometrics for Analytical Chemistry*, 4th edition, Pearson Education Limited, ISBN 0-130-22888-5.

Nordkalk, (2010a), ”Verksamhet :: Nordkalk oyj”, http://www.storugns.se/default.asp?viewID=942, (2010-01-14).

Nordkalk, (2010b), “Historia :: Nordkalk oyj”, http://www.storugns.se/default.asp?viewID=947, (2010-01-14).

Shepard, D., (1968). “A two-dimensional interpolation function for irregularly-spaced data“, ACM National Conference, (1968).

Umetrics, (2005). SIMCA-P and SIMCA-P+ 11; User Guide and Tutorial, Umetrics AB.

Wold, S., Sjöström, M., Eriksson, L., (2001), ”PLS-regression: a basic tool of chemometrics”, *Chemometrics and Intelligent Laboratory Systems,* 58 (2001), 109-130.

7.2 PERSONAL COMMUNICATION

Kenneth Fjäder, Geologist, Nordkalk AB, (2009-09 – 2009-12).

APPENDIX I

SCORE AND LOADING PLOTS; DAIMOND CORE LEVELED DATA

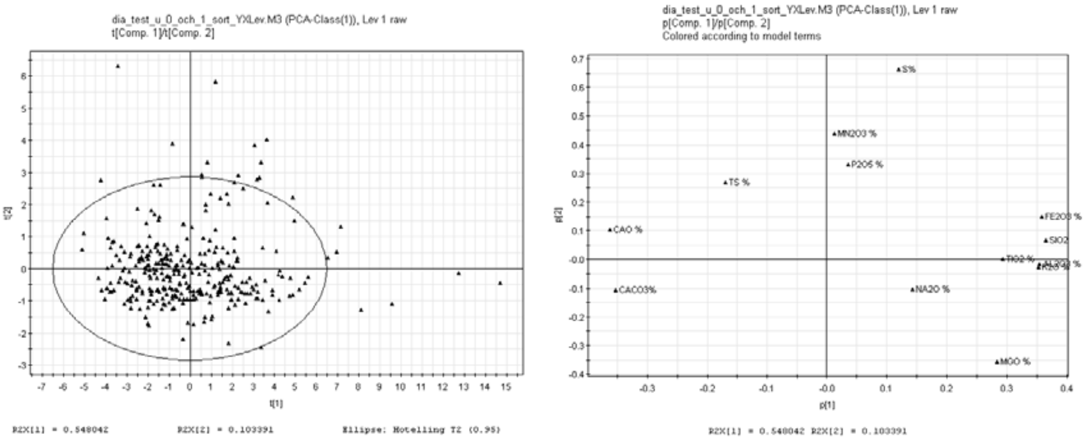


Figure 40 The Score and loading plot of the first layer, 0-6 m.

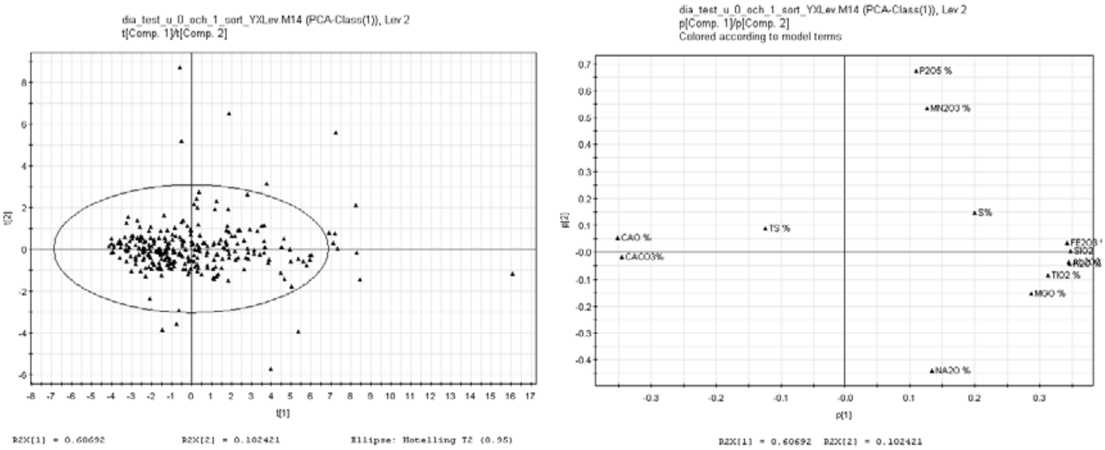


Figure 41 The Score and loading plot of the second layer, 6-9 m.

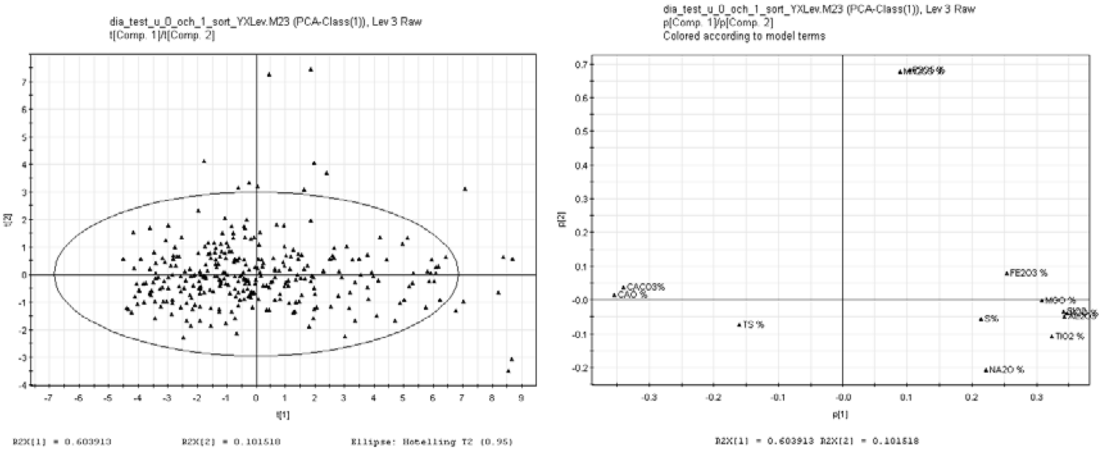


Figure 42 The Score and loading plot of the third layer, 9-12 m.

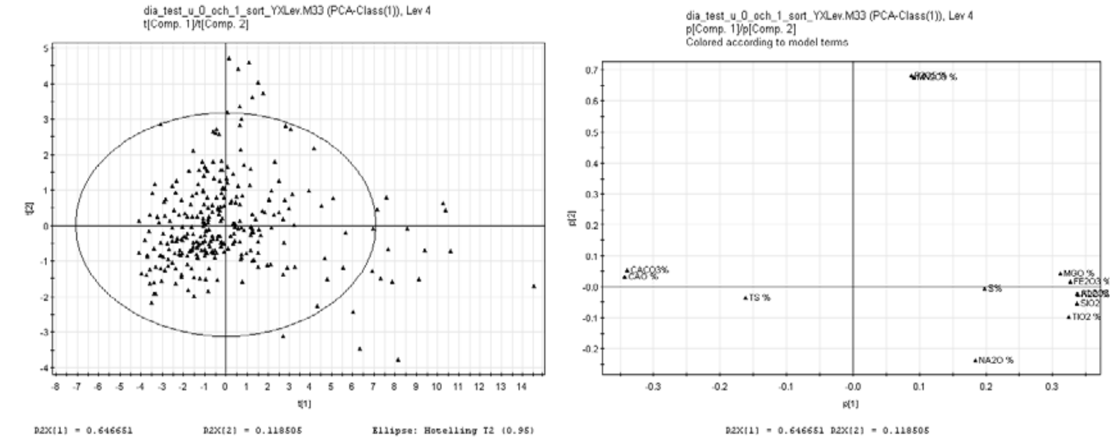


Figure 43 The Score and loading plot of the fourth layer, 12-15 m.

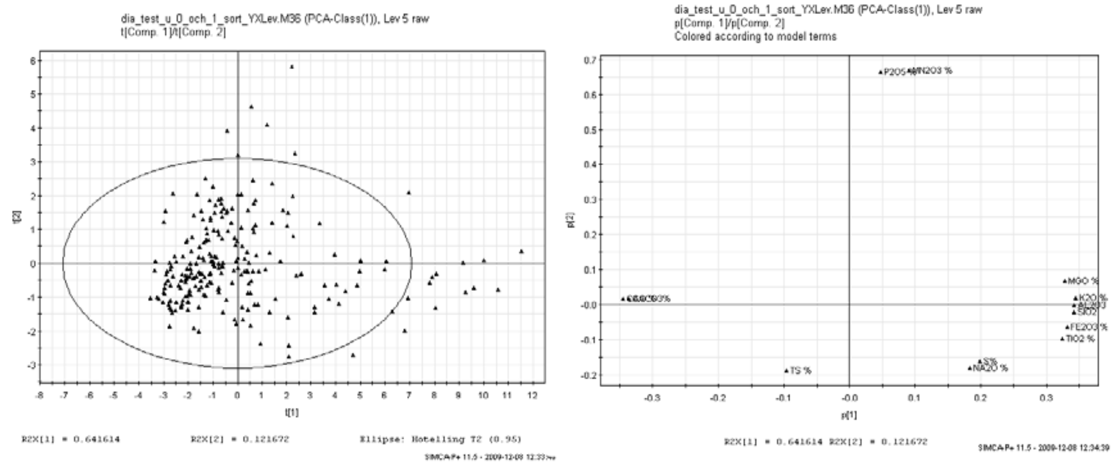


Figure 44 The Score and loading plot of the fifth layer, 15-18 m.

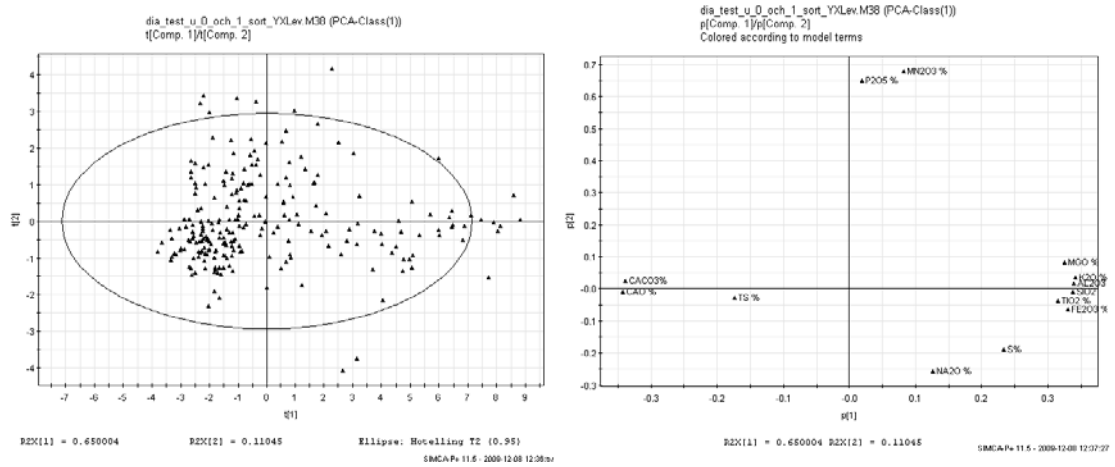


Figure 45 The Score and loading plot of the sixth layer, 18-21 m.

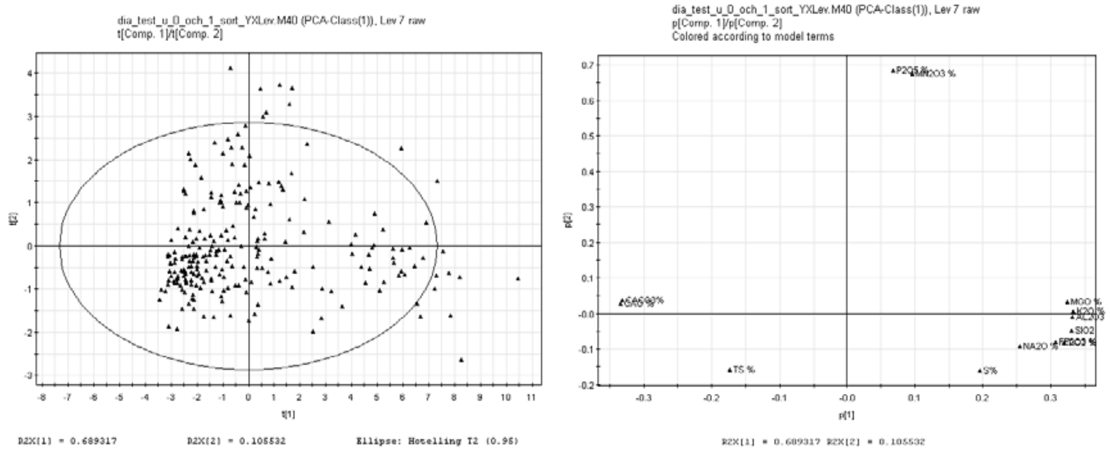


Figure 46 The Score and loading plot of the seventh layer, 21-24 m.

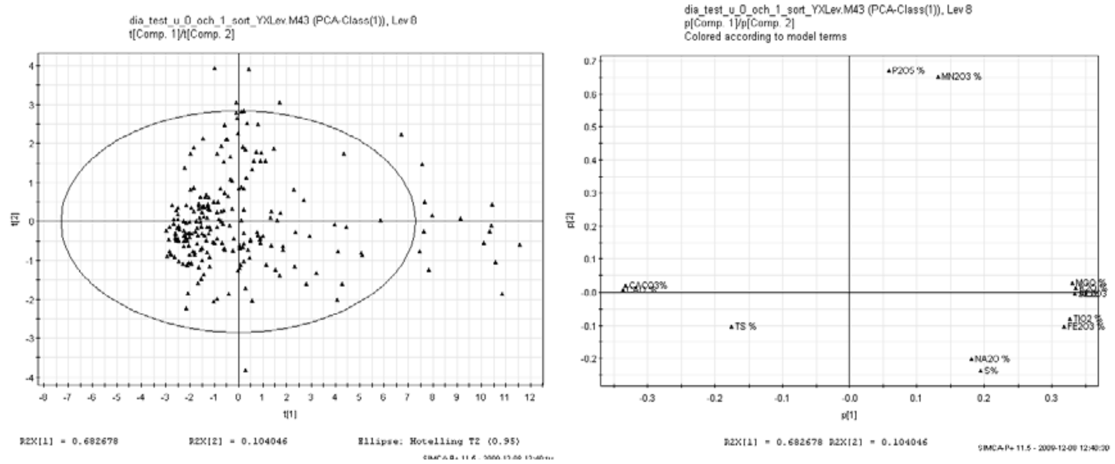


Figure 47 The Score and loading plot of the eighth layer, 24-27 m.

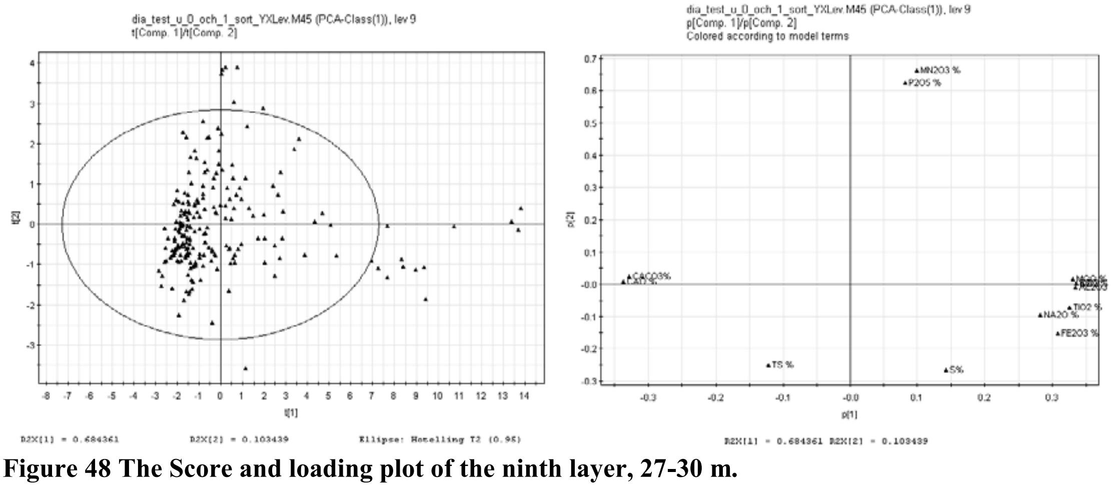


Figure 48 The Score and loading plot of the ninth layer, 27-30 m.

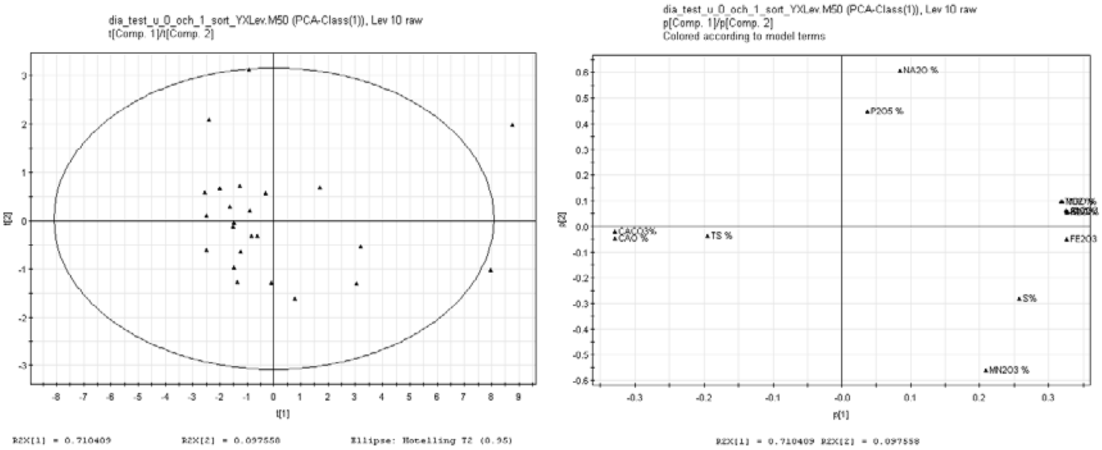


Figure 49 The Score and loading plot of the tenth layer, 30-33 m.'

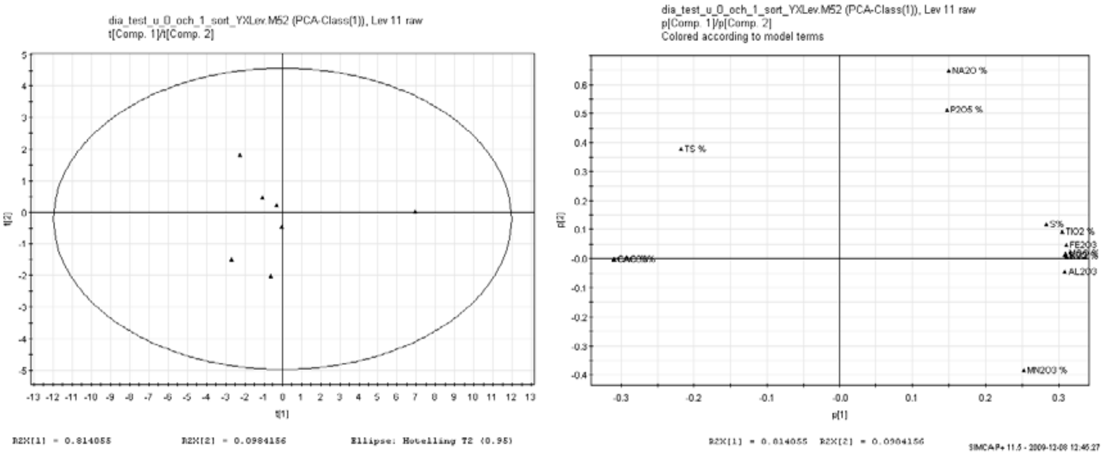
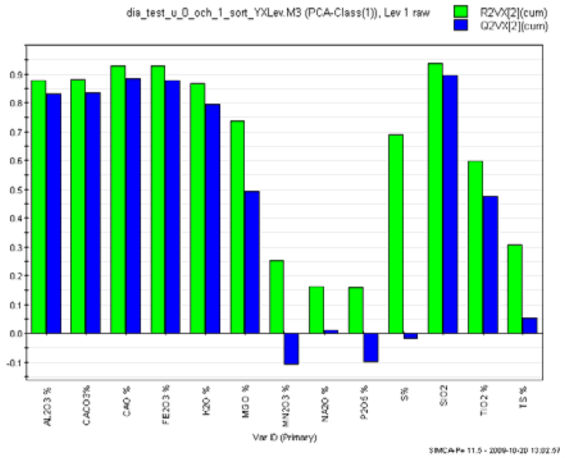


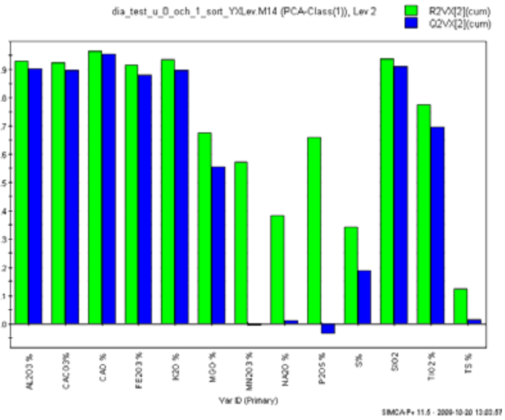
Figure 50 The Score and loading plot of the eleventh layer, 33-36 m.

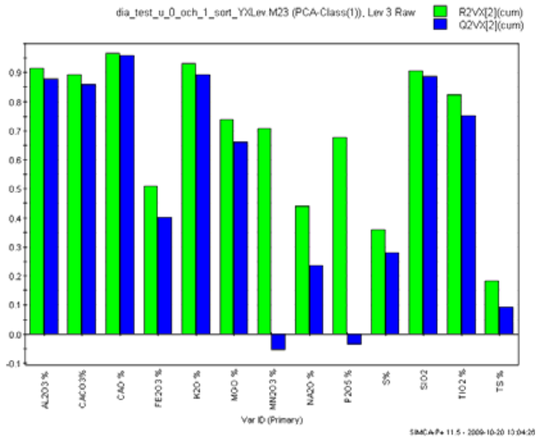
APPENDIX II

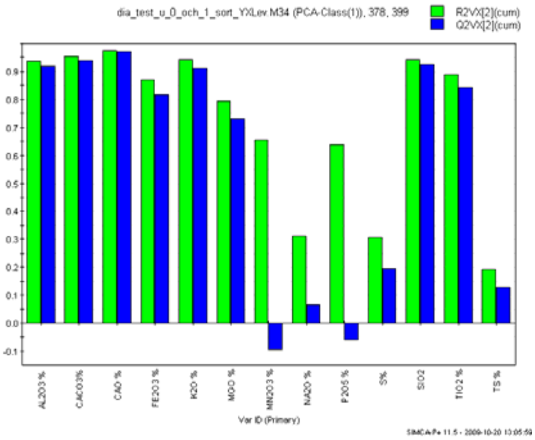
EXPLAINATORY PLOTS

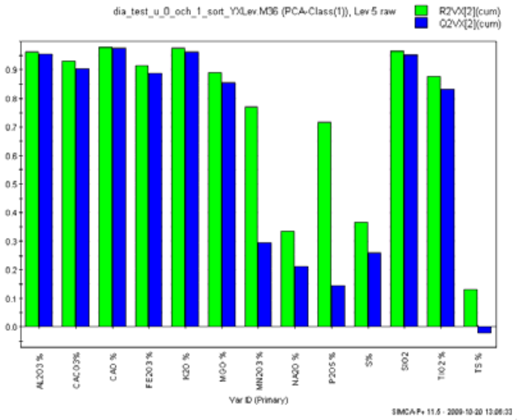
Plots for all PCA model modeled with two –PCs; starting with layer No. 1

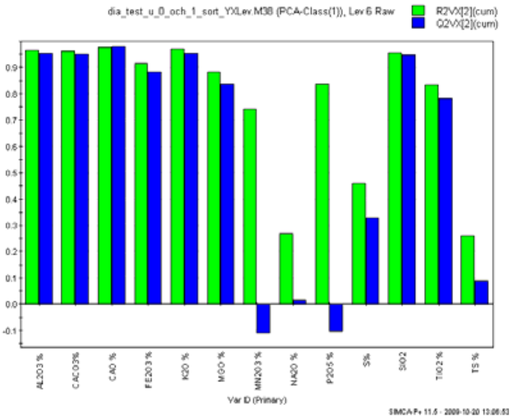


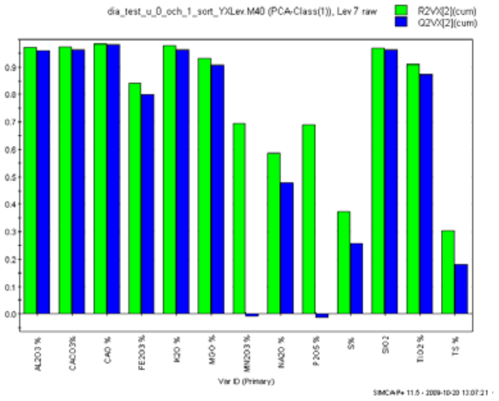


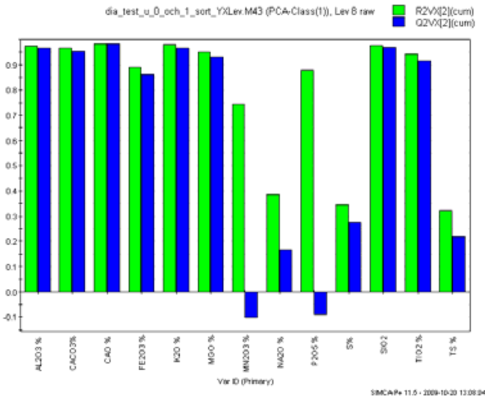


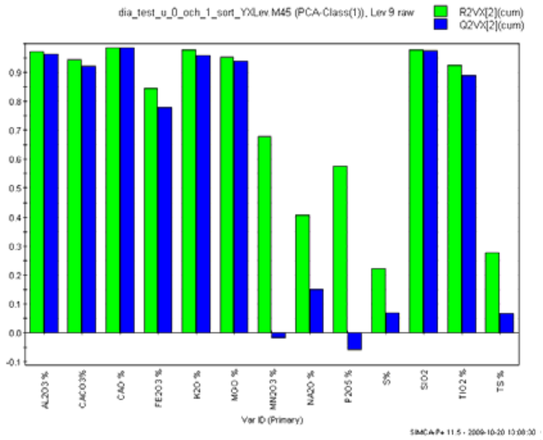


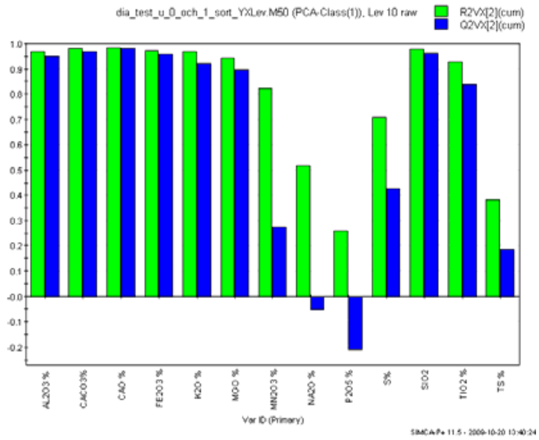


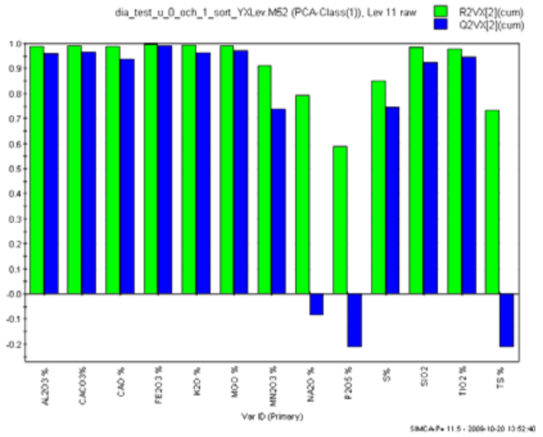



dia_test_u_0_och_1_sort_YXLev.M40 (PCA-Class(1)), Lev 7 raw
R2VX[2](cum)
Q2VX[2](cum)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
AL2O3 %
CACO3%
CAO %
FE2O3 %
K2O %
MGO %
MN2O3 %
NA2O %
P2O5 %
S%
SIO2
TIO2 %
TS %
Var ID (Primary)
SIMCA-P+ 11.5 - 2009-10-20 13:07:21


dia_test_u_0_och_1_sort_YXLev.M43 (PCA-Class(1)), Lev 8 raw
R2VX[2](cum)
Q2VX[2](cum)
-0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
AL2O3 %
CACO3%
CAO %
FE2O3 %
K2O %
MGO %
MN2O3 %
NA2O %
P2O5 %
S%
SIO2
TIO2 %
TS %
Var ID (Primary)
SIMCA-P+ 11.5 - 2009-10-20 13:08:04


dia_test_u_0_och_1_sort_YXLev.M45 (PCA-Class(1)), Lev 9 raw
R2VX[2](cum)
Q2VX[2](cum)
-0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
AL2O3 %
CACO3%
CAO %
FE2O3 %
K2O %
MGO %
MN2O3 %
NA2O %
P2O5 %
S%
SIO2
TIO2 %
TS %
Var ID (Primary)
SIMCA-P+ 11.5 - 2009-10-20 13:08:30

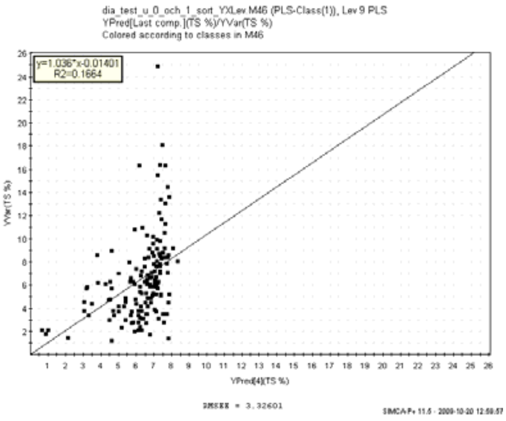

dia_test_u_0_och_1_sort_YXLev.M50 (PCA-Class(1)), Lev 10 raw
R2VX[2](cum)
Q2VX[2](cum)
-0.2
-0.1
-0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
AL2O3 %
CACO3%
CAO %
FE2O3 %
K2O %
MGO %
MN2O3 %
NA2O %
P2O5 %
S%
SIO2
TIO2 %
TS %
Var ID (Primary)
SIMCA-P+ 11.5 - 2009-10-20 13:40:24

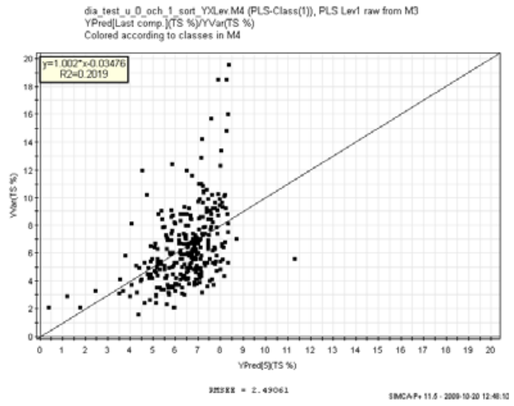

dia_test_u_0_och_1_sort_YXLev.M52 (PCA-Class(1)), Lev 11 raw
R2VX[2](cum)
Q2VX[2](cum)
-0.2
-0.1
-0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
AL2O3 %
CACO3%
CAO %
FE2O3 %
K2O %
MGO %
MN2O3 %
NA2O %
P2O5 %
S%
SIO2
TIO2 %
TS %
Var ID (Primary)
SIMCA-P+ 11.5 - 2009-10-20 13:52:40

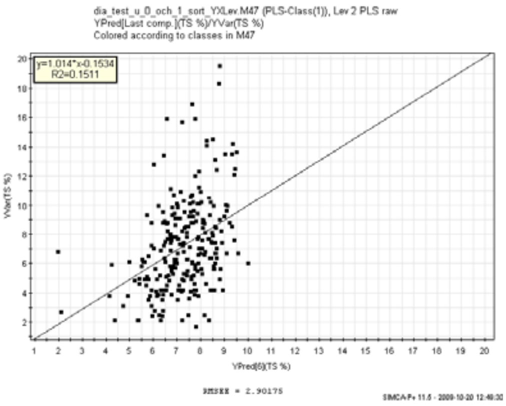
APPENDIX III

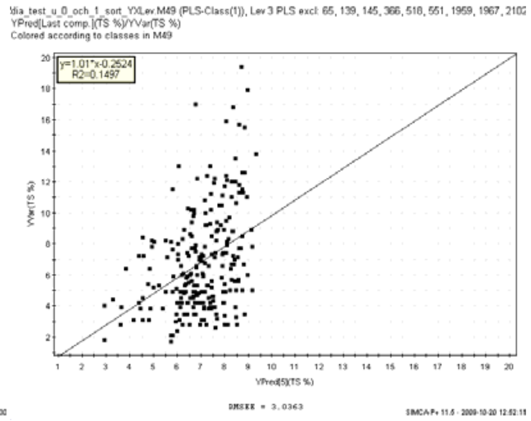
PREDICTION VS OBSERVATION PLOTS

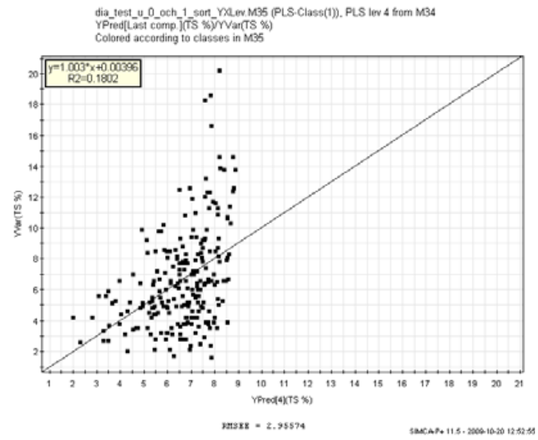
TS

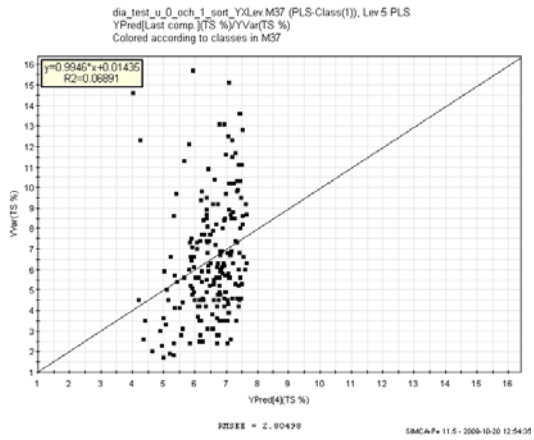


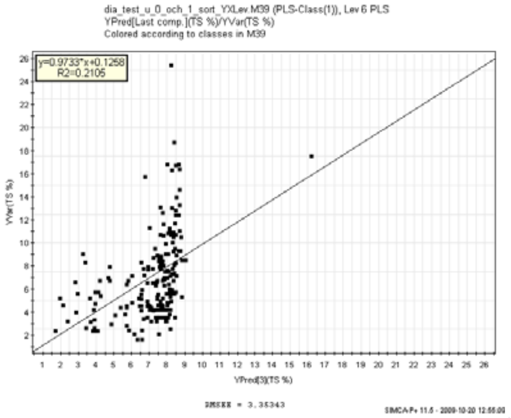


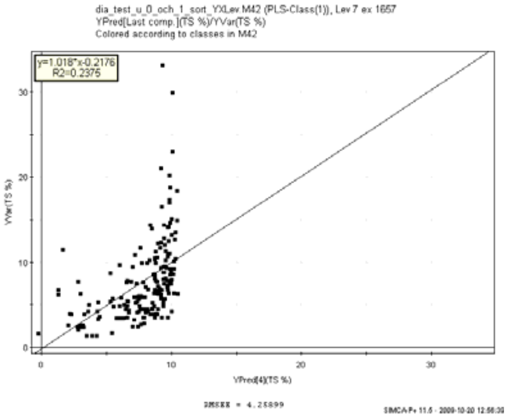


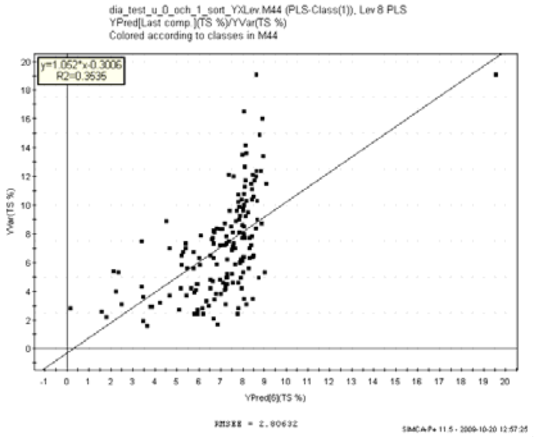


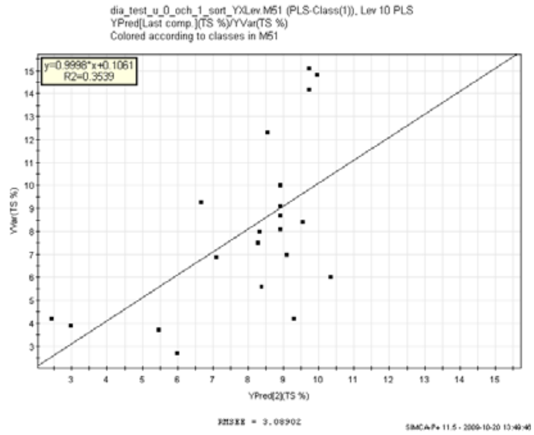


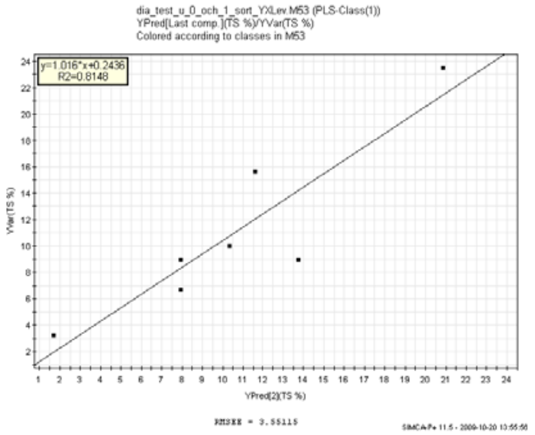



dia_test_u_0_och_1_sort_YXLev.M39 (PLS-Class(1)), Lev 6 PLS
YPred[Last comp.](TS %)/YVar(TS %)
Colored according to classes in M39
y=0.9733*x+0.1258
R2=0.2105
YVar(TS %)
2 4 6 8 10 12 14 16 18 20 22 24 26
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
YPred[3](TS %)
RMSEE = 3.35343
SIMCA-P+ 11.5 - 2009-10-20 12:55:09

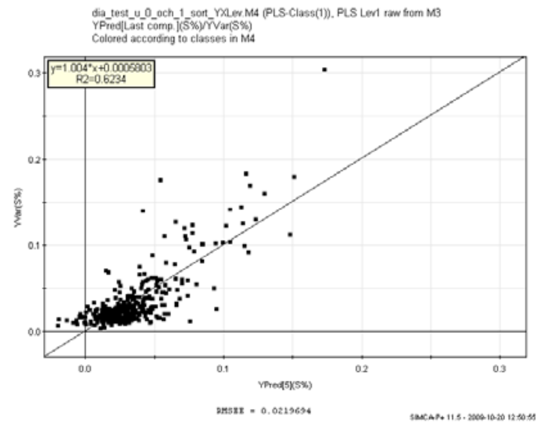

dia_test_u_0_och_1_sort_YXLev.M42 (PLS-Class(1)), Lev 7 ex 1657
YPred[Last comp.](TS %)/YVar(TS %)
Colored according to classes in M42
y=1.018*x-0.2176
R2=0.2375
YVar(TS %)
0 10 20 30
0 10 20 30
YPred[4](TS %)
RMSEE = 4.25899
SIMCA-P+ 11.5 - 2009-10-20 12:56:39

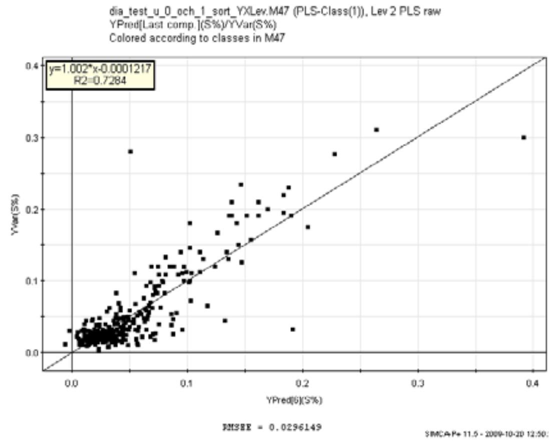

dia_test_u_0_och_1_sort_YXLev.M44 (PLS-Class(1)), Lev 8 PLS
YPred[Last comp.](TS %)/YVar(TS %)
Colored according to classes in M44
y=1.052*x-0.3006
R2=0.3535
YVar(TS %)
0 2 4 6 8 10 12 14 16 18 20
-1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
YPred[6](TS %)
RMSEE = 2.80632
SIMCA-P+ 11.5 - 2009-10-20 12:57:25

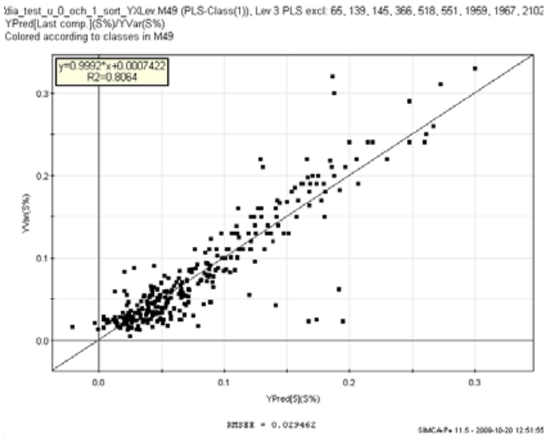

dia_test_u_0_och_1_sort_YXLev.M51 (PLS-Class(1)), Lev 10 PLS
YPred[Last comp.](TS %)/YVar(TS %)
Colored according to classes in M51
y=0.9998*x+0.1061
R2=0.3539
YVar(TS %)
3 4 5 6 7 8 9 10 11 12 13 14 15
3 4 5 6 7 8 9 10 11 12 13 14 15
YPred[2](TS %)
RMSEE = 3.08902
SIMCA-P+ 11.5 - 2009-10-20 13:49:46

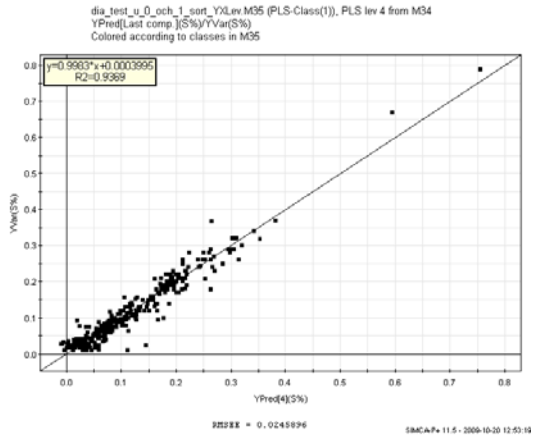

dia_test_u_0_och_1_sort_YXLev.M53 (PLS-Class(1))
YPred[Last comp.](TS %)/YVar(TS %)
Colored according to classes in M53
y=1.016*x+0.2436
R2=0.8148
YVar(TS %)
2 4 6 8 10 12 14 16 18 20 22 24
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
YPred[2](TS %)
RMSEE = 3.55115
SIMCA-P+ 11.5 - 2009-10-20 13:55:56

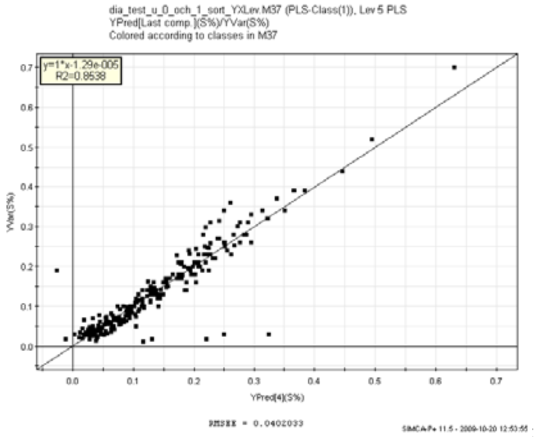
S

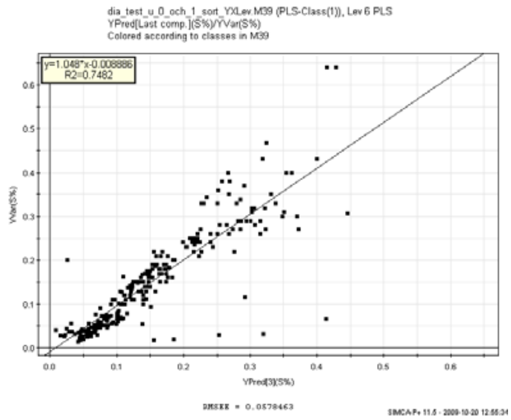

dia_test_u_0_och_1_sort_YXLev.M4 (PLS-Class(1)), PLS Lev1 raw from M3
YPred[Last comp.](S%)/YVar(S%)
Colored according to classes in M4

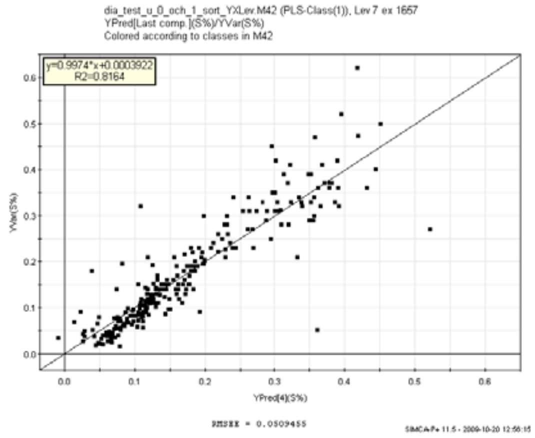

dia_test_u_0_och_1_sort_YXLev.M47 (PLS-Class(1)), Lev 2 PLS raw
YPred[Last comp.](S%)/YVar(S%)
Colored according to classes in M47

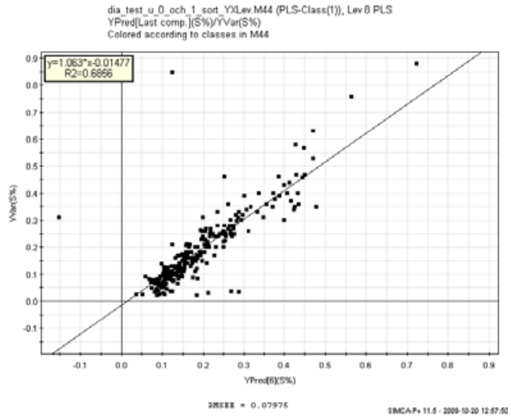

dia_test_u_0_och_1_sort_YXLev.M49 (PLS-Class(1)), Lev 3 PLS excl: 65, 139, 145, 366, 518, 551, 1959, 1967, 2102
YPred[Last comp.](S%)/YVar(S%)
Colored according to classes in M49

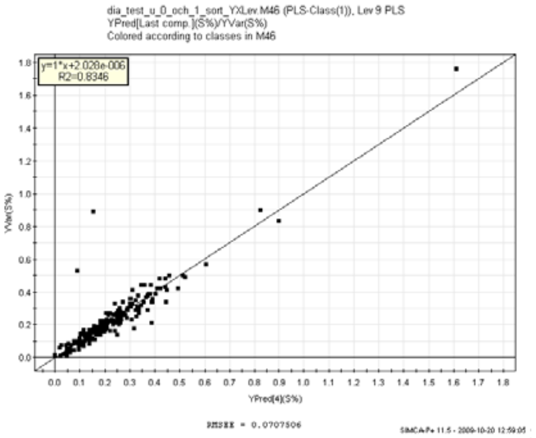

dia_test_u_0_och_1_sort_YXLev.M35 (PLS-Class(1)), PLS lev 4 from M34
YPred[Last comp.](S%)/YVar(S%)
Colored according to classes in M35

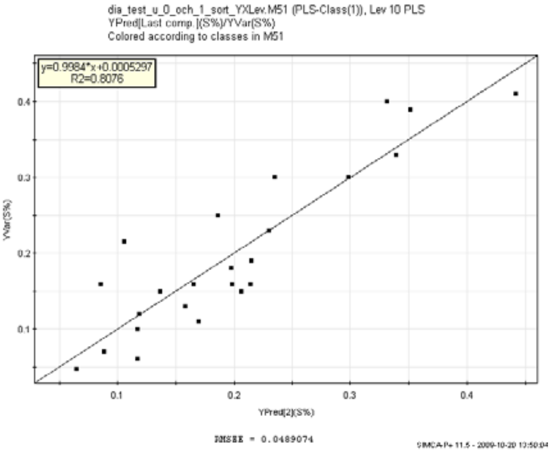

dia_test_u_0_och_1_sort_YXLev.M37 (PLS-Class(1)), Lev 5 PLS
YPred[Last comp.](S%)/YVar(S%)
Colored according to classes in M37

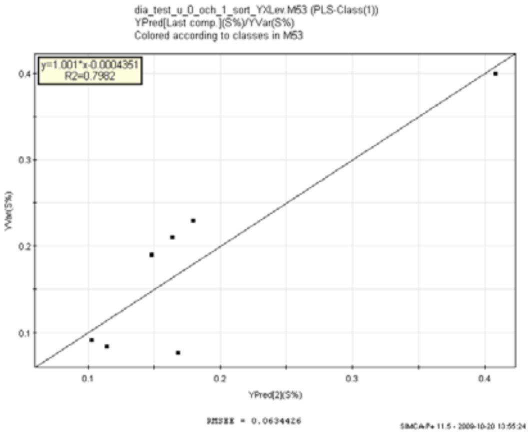

dia_test_u_0_och_1_sort_YXLev.M39 (PLS-Class(1)), Lev 6 PLS
YPred[Last comp.](S%)/YVar(S%)
Colored according to classes in M39


dia_test_u_0_och_1_sort_YXLev.M42 (PLS-Class(1)), Lev 7 ex 1657
YPred[Last comp.](S%)/YVar(S%)
Colored according to classes in M42


dia_test_u_0_och_1_sort_YXLev.M44 (PLS-Class(1)), Lev 8 PLS
YPred[Last comp.](S%)/YVar(S%)
Colored according to classes in M44


dia_test_u_0_och_1_sort_YXLev.M46 (PLS-Class(1)), Lev 9 PLS
YPred[Last comp.](S%)/YVar(S%)
Colored according to classes in M46
y=1*x+2.028e-006
R2=0.8346
YVar(S%)
YPred[4](S%)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
0.1
0.3
0.5
0.7
0.9
1.1
1.3
1.5
1.7
RMSEE = 0.0707506
SIMCA-P+ 11.5 - 2009-10-20 12:59:05


dia_test_u_0_och_1_sort_YXLev.M51 (PLS-Class(1)), Lev 10 PLS
YPred[Last comp.](S%)/YVar(S%)
Colored according to classes in M51
y=0.9984*x+0.0005297
R2=0.8076
YVar(S%)
YPred[2](S%)
0.1
0.2
0.3
0.4
RMSEE = 0.0489074
SIMCA-P+ 11.5 - 2009-10-20 13:50:04


dia_test_u_0_och_1_sort_YXLev.M53 (PLS-Class(1))
YPred[Last comp.](S%)/YVar(S%)
Colored according to classes in M53
y=1.001*x-0.0004351
R2=0.7982
YVar(S%)
YPred[2](S%)
0.1
0.2
0.3
0.4
RMSEE = 0.0634426
SIMCA-P+ 11.5 - 2009-10-20 13:55:24

APPENDIX IV

Bergartskoder:

Vi arbetar med 5 st. huvudtyper av kalksten, samt en serie variationer och övergångstyper:

S	Stromatoporoidékalksten
K	Krinoidékalksten
R	Revkalksten
Fr	Fragmentkalksten
M	Märgel

SK	Växellagrande stromatoporoidé- och krinoidékalksten
LK	Lerig krinoidékalksten
AK	Arenitisk krinoidékalksten
KFr	Krinoidékalksten med inslag av fragment
SFr	Stromatoporoidékalksten med inslag av fragment
LFr	Lerig fragmentkalksten
LR	Lerig revkalksten
MK	Mörk krinoidékalksten
LMK	Lerig mörk krinoidékalksten
LjK	Ljus krinoidékalksten
J	Jord

APPENDIX V

Base function of the Kriging interpolation

```
clear all
close all
load BA;
pos=levkrig(pos);
xlag=170;
ylag=80;
x=pos(:,1);
y=pos(:,2);
lev=11;




f=find(pos(:,4)==lev);
x1=x(f);
y1=y(f);

z1=val(f,:);
valid=round(linspace(1,length(x1),round(length(x1)*0.8)));
valid=valid';

x=x1(valid);
y=y1(valid);
z=z1(valid,:);
x1(valid)=[];
y1(valid)=[];
z1(valid,:)=[];
gamest=zeros(size(y));

for i=2:length(x)
d(i)=isequal(x(i),x(i-1));
end
f=[];
f=find(d==1);
x(f)=[];
y(f)=[];
z(f,:)=[];
%[c01,c11,a1,type1,c02,c12,a2,type2] = semivar1(x,y,z);
% [binc,sv,bin_array,svM,gamma_array,dist_array]=semivar([x
y],z(:,1));
% [binc2,sv2,bin_array2,svM2,gamma_array2,dist_array2]=semivar([x
y],z(:,2));
% plot(binc,sv,'*');figure;plot(binc2,sv2,'*');

[W1,we1,z,x,y,xest,yest,dx,dy]    = kriga([x
y],1.9,4.3,160,'exp',z,xlag,ylag);
[W2,we2]                    = kriga([x
y],1.07,0.32,200,'exp',z,xlag,ylag);

[Zest1,zest1] = kriget(W1,z(:,1),xlag,ylag);
[Zest2,zest2] = idw(x,y,z(:,1),xest,yest,xlag,ylag,200);
[Zest3,zest3] = kriget(W2,z(:,2),xlag,ylag);
[Zest4,zest4] = idw(x,y,z(:,2),xest,yest,xlag,ylag,200);
[errK1]  = krigvar(W1,we1,xlag,ylag);
[errK2]  = krigvar(W2,we2,xlag,ylag);

[valpc1]=valet(zest1,z1(:,1),x1,y1,xest,yest);
```

```
[validw1]=valet(zest2,z1(:,1),x1,y1,xest,yest);

[valpc2]=valet(zest3,z1(:,2),x1,y1,xest,yest);
[validw2]=valet(zest4,z1(:,2),x1,y1,xest,yest);

%% figure
colormap(jet);
imagesc(dx,dy,Zest1);
title('Krig PC 1')
colorbar;axis equal tight xy;
xlabel('Local X');ylabel('Local Y');

figure
colormap(jet);
imagesc(dx,dy,errK1);
title('err PC 1')
colorbar;axis equal tight xy;
xlabel('Local X');ylabel('Local Y');

figure
colormap(jet);
imagesc(dx,dy,Zest2);
title('IDW PC 1')
colorbar;axis equal tight xy;
xlabel('Local X');ylabel('Local Y');

figure
colormap(jet);
imagesc(dx,dy,Zest3);
title('Krig PC 2')
colorbar;axis equal tight xy;
xlabel('Local X');ylabel('Local Y');

figure
colormap(jet);
imagesc(dx,dy,errK2);
title('err PC 2')
colorbar;axis equal tight xy;
xlabel('Local X');ylabel('Local Y');

figure
colormap(jet);
imagesc(dx,dy,Zest4);
title('IDW PC 2')
colorbar;axis equal tight xy;
xlabel('Local X');ylabel('Local Y');

 %save(num2str(lev))
```

Calculating the weights of Kriging

```
function [W,we,z,x,y,xest,yest,dx,dy] =
kriga(pos,c0,c1,a,type,val,xlag,ylag)
x=pos(:,1);
y=pos(:,2);
z=val;%
% for i=2:length(x)
% d(i)=isequal(x(i),x(i-1));
% end
%
```

```
% f=find(d==0);
% x=x(f+1);
% y=y(f+1);
% gamest=zeros(size(y));
% z=val(f+1,:);


    switch lower(type)
        case {'sph'}
             for i=1:length(y)
                 dist=sqrt((x-x(i)).^2+(y-y(i)).^2);

                f=find(dist<=a);
                gamest(f,i)=c0 + c1.*(1.5.*(dist(f)./a)-
0.5.*(dist(f)./a).^3);
                f=[];
                f=find(dist>a);
                gamest(f,i)=c0+c1;

            end
            V=[gamest ones(length(y),1);ones(1,length(y)) 0];

            dx=linspace(3000,3850,xlag);
            dy=linspace(10600,11000,ylag);
            xest=[];
            yest=[];
            for i=1:length(dy)
                for i1=1:length(dx)
                    newxest(i1,1)=dx(i1);
                    newyest(i1,1)=dy(i);
                end
                xest=[xest; newxest];
                yest=[yest; newyest];
            end
%             [L,U]=lu(V);
            we=[];
            for i=1:length(xest)

                he=sqrt((x-xest(i)).^2+(y-yest(i)).^2);
                f=[];
                f=find(he<=a);
                he(f)= c0 + c1.*(1.5.*(he(f)./a)-0.5.*(he(f)./a).^3);

                f=[];
                f=find(he>a);
                he(f)=c0+c1;

                he=[he;1];
                we=[we he];

%                 b=L\he;
%                 W(:,i)=U\b;
%                 info=round(i/length(xest)*100);
%                 disp(num2str(info))
            end

            W=V\we;

        case {'exp'}
```

```
            for i=1:length(y)
                dist=sqrt((x-x(i)).^2+(y-y(i)).^2);

                f=find(dist>0);
                gamest(f,i)= c0 + c1.*(1-exp(-(dist(f)./a)));
                f=[];
                f=find(dist==0);
                gamest(f,i)=0;

            end
            V=[gamest ones(length(y),1);ones(1,length(y)) 0];

            dx=linspace(3000,3850,xlag);
            dy=linspace(10600,11000,ylag);
            xest=[];
            yest=[];
            for i=1:length(dy)
                for i1=1:length(dx)
                    newxest(i1,1)=dx(i1);
                    newyest(i1,1)=dy(i);
                end
                xest=[xest; newxest];
                yest=[yest; newyest];
            end

%             [L,U]=lu(V);
            we=[];
            for i=1:length(xest)
                he=sqrt((x-xest(i)).^2+(y-yest(i)).^2);

                f=find(he>0);
                he(f)= c0 + c1.*(1-exp(-(he(f)./a)));

                f=find(he==0);
                he(f)=0;

                he=[he;1];
                we=[we he];
%                 b=L\he;
%                 W(:,i)=U\b;
%                 info=i/length(xest)*100;
%                 disp(num2str(info))
            end
            W=V\we;

        case {'lin'}
             for i=1:length(y)
                dist=sqrt((x-x(i)).^2+(y-y(i)).^2);

                f=find(dist~=0);
                gamest(f,i)=c0 + c1.*dist(f);
                f=[];
                f=find(dist==0);
                gamest(f,i)=0;

            end
            V=[gamest ones(length(y),1);ones(1,length(y)) 0];

            dx=linspace(3000,3850,xlag);
            dy=linspace(10600,11000,ylag);
            xest=[];
            yest=[];
            for i=1:length(dy)
                for i1=1:length(dx)
                    newxest(i1,1)=dx(i1);
                    newyest(i1,1)=dy(i);
                end
                xest=[xest; newxest];
                yest=[yest; newyest];
            end
%             [L,U]=lu(V);
            we=[];
            for i=1:length(xest)

                he=sqrt((x-xest(i)).^2+(y-yest(i)).^2);
                f=[];
                f=find(he~=0);
                he(f)= c0 + c1.*he(f);

                f=[];
                f=find(he==a);
                he(f)=0;

                he=[he;1];
                we=[we he];

%                 b=L\he;
%                 W(:,i)=U\b;
%                 info=round(i/length(xest)*100);
%                 disp(num2str(info))
            end

            W=V\we;

        otherwise
            disp('choose either spherical (sph) or exponetial (exp)')
    end
```

Kriging interpolation

```
function [Zest,zest] = kriget(W,z,xlag,ylag)

zest=W(1:end-1,:)'*z;

for i=0:ylag-1
    for i1=0:xlag-1
        Zest(i+1,i1+1)=zest(i*xlag+1+i1);
    end
end
```

IDW interpolation

```
function [Zest1,zest]=idw(x,y,z,xest,yest,xlag,ylag,toldist)

w=[];
for i=1:length(xest)
    w=(sqrt((x-xest(i)).^2+(y-yest(i)).^2));
    f=find(w<=toldist);
    w=1./(w.^2+eps);
    zest(i)=(w(f)'*z(f))/sum(w(f));
```

```
end

for i=0:ylag-1
    for i1=0:xlag-1
        Zest1(i+1,i1+1)=zest(i*xlag+1+i1);

    end
end
zest=zest';
```

Validation code

```
function [v]=valet(zest,z1,x1,y1,xest,yest)
xes=[];
yes=[];
zes=[];
for i=1:length(x1)

    k=abs(yest-y1(i));
    kmin=min(k);
    f=find(k==kmin);

    xe=xest(f);
    ye=yest(f);
    ze=zest(f);

    k=[];
    k=abs(xe-x1(i));
    kmin=min(k);
    f=find(k==kmin);

    xes=[xes; xe(f(1))];
    yes=[yes; ye(f(1))];
    zes=[zes; ze(f(1))];
end


v=(zes-z1).^2;
v=mean(v);
```

Code for computing semivariogram with anisotropy

```
function [c01,c11,a1,type1,c02,c12,a2,type2] = semivar1(x,y,z)
%x=pos(:,1);y=pos(:,2);z=val;
% distx=zeros(length(x));disty=zeros(length(x));
dist=zeros(length(x));
close all


%% Semivariogram North
lags=linspace(60,410,8);
zmean=[];
for p=2:length(lags)
    qz11=[];qz12=[];
    for i=1:length(x)
%         distx=x-x(i);
%         disty=y-y(i);
        dist=sqrt((x-x(i)).^2+(y-y(i)).^2);
        ang = [((x-x(i))./dist) ((y-y(i))./dist)];
        newZ(:,1)=(z(:,1)-z(i,1)).^2;
        newZ(:,2)=(z(:,2)-z(i,2)).^2;

        f=[];newdist=[];newz=[];
        f=find(ang(:,2)>= sin(60/360*2*pi) & dist <= max(lags));
        newdist=dist(f);
        newz=newZ(f,:);
        f=[];
        f=find(newdist >= lags(p-1) & newdist< lags(p));
%         newdist=newdist(f);
        qz11=[qz11; newz(f,1)];
        qz12=[qz12; newz(f,2)];

    end
    zmean(p-1,1)=mean(qz11)/2;
    zmean(p-1,2)=mean(qz12)/2;

    zmean(p-1,3)=(lags(p)+lags(p-1))/2;
end
[gamest1,he,c01,c11,a1,type1]=semivarmod(1.9,4.1,100,'sph');
[gamest2,he,c02,c12,a2,type2]=semivarmod(1.07,0.32,200,'exp');
figure
subplot(321);plot(he,gamest1,'--r');
hold on
plot(zmean(:,3),zmean(:,1),'-x');
hold off
title('PCA 1 North')
axis([0 max(zmean(:,3))+10 0 max(zmean(:,1))+1])
subplot(322);plot(he,gamest2,'--r');
hold on
plot(zmean(:,3),zmean(:,2),'-x');
hold off
title('PCA 2 North')
axis([0 max(zmean(:,3))+10 0 max(zmean(:,2))+1])
%% Semivariogram for East
lags=linspace(0,300,8);
zmean=[];
for p=2:length(lags)
    qz11=[];qz12=[];
    for i=1:length(x)
%         distx=x-x(i);
%         disty=y-y(i);
        dist=sqrt((x-x(i)).^2+(y-y(i)).^2);
        ang = [((x-x(i))./dist) ((y-y(i))./dist)];
        newZ(:,1)=(z(:,1)-z(i,1)).^2;
        newZ(:,2)=(z(:,2)-z(i,2)).^2;

        f=[];newdist=[];newz=[];    lags=linspace(0,300,8);
        f=find(ang(:,1) >= cos(30/360*2*pi)  & dist <= max(lags));
        newdist=dist(f);
        newz=newZ(f,:);
        f=[];
        f=find(newdist >= lags(p-1) & newdist< lags(p));
%         newdist=newdist(f);
        qz11=[qz11; newz(f,1)];
        qz12=[qz12; newz(f,2)];

    end
    zmean(p-1,1)=mean(qz11)/2;
    zmean(p-1,2)=mean(qz12)/2;
    zmean(p-1,3)=(lags(p)+lags(p-1))/2;
end

subplot(323);plot(he,gamest1,'--r');

hold on
plot(zmean(:,3),zmean(:,1),'-x');
hold off
title('PCA 1 East')
axis([0 max(zmean(:,3))+10 0 max(zmean(:,1))+1])
subplot(324);plot(he,gamest2,'--r');

hold on
plot(zmean(:,3),zmean(:,2),'-x');
hold off
title('PCA 2 East')
axis([0 max(zmean(:,3))+10 0 max(zmean(:,2))+1])
%% Semivariogram for North-West and North-East
lags=linspace(70,480,10);
zmean=[];
for p=2:length(lags)
    qz11=[];qz12=[];qz21=[];qz22=[];
    for i=1:length(x)
%         distx=x-x(i);
%         disty=y-y(i);
        dist=sqrt((x-x(i)).^2+(y-y(i)).^2);
        ang = [((x-x(i))./dist) ((y-y(i))./dist)];
        newZ(:,1)=(z(:,1)-z(i,1)).^2;
        newZ(:,2)=(z(:,2)-z(i,2)).^2;

        f=[];newdist=[];newz=[];
        f=find(ang(:,1) <= cos(15/360*2*pi) &...
            ang(:,1) >= cos(75/360*2*pi) &...
            sign(ang(:,2))==1 & dist <= max(lags));
        newdist=dist(f);
        newz=newZ(f,:);
        f=[];
        f=find(newdist >= lags(p-1) & newdist< lags(p));
%         newdist=newdist(f);
        qz11=[qz11; newz(f,1)];
        qz12=[qz12; newz(f,2)];

        f=[];newdist=[];newz=[];
        f=find(ang(:,1) <= cos(105/360*2*pi) &...
            ang(:,1) >= cos(165/360*2*pi) &...
            sign(ang(:,2))==1 & dist <= max(lags));
        newdist=dist(f);
        newz=newZ(f,:);
         f=[];
        f=find(newdist >= lags(p-1) & newdist< lags(p));
%         newdist=newdist(f);
        qz21=[qz21; newz(f,1)];
        qz22=[qz22; newz(f,2)];

    end
    zmean(p-1,1)=mean(qz11)/2;
    zmean(p-1,2)=mean(qz12)/2;
    zmean(p-1,3)=mean(qz21)/2;
    zmean(p-1,4)=mean(qz22)/2;
    zmean(p-1,5)=(lags(p)+lags(p-1))/2;
end

subplot(325);plot(he,gamest1,'--r');
hold on
plot(zmean(:,5),[zmean(:,1) zmean(:,3)],'-x');
hold off
title('PCA 1 North-East & North-West')
legend('model','N-E','N-W')
axis([0 max(zmean(:,5))+10 0 max([zmean(:,1); zmean(:,3)])+1])

subplot(326);plot(he,gamest2,'--r');
hold on
plot(zmean(:,5),[zmean(:,2) zmean(:,4)],'-x');
hold off
title('PCA 2 North-East & North-West')
legend('model','N-E','N-W')
axis([0 max(zmean(:,5))+10 0 max([zmean(:,2); zmean(:,4)])+1])
frame=getframe;
```

Code for semivariogram with isotropy

```
% semivar_exp : Calcualte experimental variogram
%
%[hc,garr,h,gamma,hangc,head,tail]=semivar_exp(pos,val,nbin,nbinang)
%
% pos : [ndata,ndims]
% val : [ndata,ndata_types]
%
% nbin : [integer] number of bins on distance anxes
%        [array] if specified as an array, this is used.
%
% nbinang : [integer] number of arrays between 0/180 degrees
%                     (default 1)
%
% Example : load jura data
%
dwd=[mgstat_dir,filesep,'examples',filesep,'data',filesep,'jura',files
ep];
%   [p,pHeader]=read_eas([dwd,'prediction.dat']);
%   idata=6;dval=pHeader{idata};
%   pos=[p(:,1) p(:,2)];
%   val=p(:,idata);
%   figure;scatter(pos(:,1),pos(:,2),10,val(:,1),'filled');
%     colorbar;title(dval);xlabel('X');ylabel('Y');axis image;
%
% Example isotrop:
%   [hc,garr]=semivar_exp(pos,val);
%   plot(hc,garr);
%   xlabel('Distance (m)');ylabel('semivariance');title(dval)
%
% Exmple directional
%   [hc,garr,h,gamma,hangc]=semivar_exp(pos,val,20,4);
%   plot(hc,garr);
%   legend(num2str(180*hangc'./pi))
%   xlabel('Distance (m)');ylabel('semivariance');title(dval)
%
%
%
```

```
% TMH/2005-2009
%
%
function
[hc,garr,h,gamma,hangc,z_head,z_tail,dp,f]=semivar_exp(pos,val,nbin,nb
inang)


ndata=size(pos,1);
ndims=size(pos,2);

if ndims==1;
  % THIS SHOULD BE CHECKED FOR BUGS
  pos=[pos 0.*pos];
  ndims=2;
end

ndata_types=size(val,2);

% First calculate the 'distance' vector
nh=sum(1:1:(ndata-1)); % Find number of pairs of data

h=zeros(nh,1);
dp=zeros(nh,ndims);
z_head=zeros(nh,ndata_types);
z_tail=zeros(nh,ndata_types);
gamma=zeros(nh,ndata_types);
vang=zeros(nh,1);

i=0;
for i1=1:(ndata-1)
for i2=(i1+1):ndata
  i=i+1;
  if ((i/20000)==round(i/20000))
    disp(sprintf('semivar_exp : i=%d/%d',i,nh))
  end

  p1=[pos(i1,:)];
  p2=[pos(i2,:)];
  dp(i,:)=p1-p2;
  h(i)=sqrt( (p1-p2)*(p1-p2)' );
  z_head(i,:)=val(i1,:);
  z_tail(i,:)=val(i2,:);
  gamma(i,:)=0.5*(val(i1,:)-val(i2,:)).^2;
  % ANGLE
  aa=sqrt(sum(p1.^2));
  bb=sqrt(sum(p2.^2));
  ab=(p1(:)'*p2(:));

  pp=p1-p2;

  % WORKS ONLY FOR 2D
  if pp(1)==0
    vang(i)=pi/2;
  else
%     vang(i)=atan(pp(1)./pp(2));
    vang(i)=atan(pp(2)./pp(1));
  end
```

```
end
end
vang=vang+pi/2;

%%%%%%%%%%%%%%%%%%%%%%%%%%
% BIN INTO ARRAY BINS
if exist('nbin')==0
  nbin=10;
else
  if length(nbin)~=1
    h_arr=nbin;
    nbin=length(h_arr)-1;
  end
end

if exist('h_arr')==0
  h_arr=linspace(0,max(h).*.3,nbin+1);
end
hc=(h_arr(1:nbin)+h_arr(2:nbin+1))./2;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% BIN INTO ANGLE BINS

if exist('nbinang')==0
  nbinang=1;
else
  if length(nbinang)~=1
    ang_array=nbinang;
    nbinang=length(nbinang)-1
  end
end
if exist('ang_array')==0
  ang_array=linspace(0,pi,nbinang+1);
end
hangc=(ang_array(1:nbinang)+ang_array(2:nbinang+1))./2;


clear garr
for i=1:nbin
  for j=1:nbinang
  f=find(h>=h_arr(i) & h<h_arr(i+1) & vang>=ang_array(j) &
vang<ang_array(j+1));
  if (sum(gamma(f,:))==0)
    garr(i,j,:)=NaN;
  else
    garr(i,j,:)=mean(gamma(f,:));

  end
  end
end
```