Matematisk statistik
Stockholms universitet

Pilotstudie för prediktering av metallurgiska kvalitetsparametrar

Marcus Nygård

Examensarbete 2006:16

Postadress:

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm
Sverige

Internet:

http://www.math.su.se/matstat

Stockholms universitet
Matematisk statistik
Examensarbete 2006:16
http://www.math.su.se/matstat

Pilotstudie för prediktering av metallurgiska kvalitetsparametrar

Marcus Nygård *

Oktober 2006

Sammanfattning

Detta examensarbete redovisar en studie av huruvida metallurgiska kvalitetsparametrar kan predikteras med hjälp av processparametrar. LKAB genomför dagligen en stor mängd dyra och tidskrävande pelletstester. LKAB är för tillfället inne i en expansiv fas och bygger två nya pelletsverk. Detta innebär att antalet tester kommer att öka markant de närmaste åren. Det vore därför mycket attraktivt om det skulle vara möjligt att prediktera testresultaten med ledning av processparametrar. Den kvalitetsparameter som detta arbete analyserar är det så kallad LTB-värdet. LTB-värdet anger hur pellets bryts ner i de övre delarna av en masugn. Det visar sig att det troligen inte är möjligt att göra numeriska prediktioner av LTB-värdet med tillfredställande prediktionsfel på den analyserade datamängden. De variabler som enligt den analyserade datamängden med stor säkerhet påverkar LTB-värdet är storleken på pellets och den kemiska sammansättningen.

*E-post: marcus.nygard@gmail.com. Handledare: Rolf Sundberg.

A pilot study for prediction of metallurgical quality parameters

Abstract

The aim of this final year project (master thesis) is to conduct a study regarding the opportunity to predict metallurgical quality parameters from process parameters. LKAB daily conduct large numbers of expensive and time consuming tests on their pellets. LKAB is at present in an expansive phase and are building two new pellet works. This implies that the numbers of tests are going to increase substantially during the next couple of years. Therefore it would be very attractive to be able to predict the test results from process parameters. This project will analyze a quality parameter called the LTB. The LTB value simulates of the pellet behaves in the upper parts of a blast furnace. The work shows that it is probably not possible to make numerical predictions with acceptable prediction error on the analyzed data. The variables that most certainly are correlated with the LTB value are the size of the pellets and the chemical composition.

Förord

Detta examensarbete är utfört vid institutionen för matematisk statistik vid Stockholms universitet. Arbetet är utfört på uppdrag av LKAB under sommaren/hösten 2006. Jag vill tacka min handledare Rolf Sundberg vid Stockholms universitet för all korrespondens under arbetets gång. Slutligen vill jag tack alla på LKAB som gjort detta arbete möjligt och framförallt min handledare vid LKAB i Kiruna Henrik Thorneus.

Innehåll

1 Inledning

I detta kapitel beskrivs först bakgrunden och syftet till examensarbetet. Sedan presenteras valda analysmetoder och kapitlet avslutas med en beskrivning av examensarbetets struktur.

1.1 Bakgrund

Två nya pelletsverk byggs i malmfälten på beställning av LKAB, ett i Malmberget (MK3) och ett i Kiruna (KK4). Det innebär att antalet metallurgiska tester markant kommer att öka. Då de flesta testmetoder är mycket tidsödande är det attraktivt att försöka förutsäga resultaten med ledning av andra mätdata såsom kemisk sammansättning, fukthalt, partikelstorlek samt processdata (bl.a. temperaturer, gasflöden, bäddhöjd, tryck).

En av LKAB:s viktigaste kvalitetsparametrar är LTB-värdet (Low Temperature Breakdown). LTB-värdet anger hur järnmalmpellets uppför sig i övre delen av en masugn. Bestämningen av LTB genomförs efter en ISO-metod (ISO 13930) som används för att bestämma lågtemperaturhållfastheten (isoterm reduktion vid 500 ^{o}C) hos järnmalmspellets.

LTB är en viktig kvalitetsparameter ty skörare pellets ger upphov till mer fines (fines är små pelletsfragment). En hög andel fines kan förutom produktionsbortfall i själva masugnarna också ge upphov till miljöproblem (i form av damm) under transport och i produktionen där pellets hanteras.

1.2 Syfte

I detta examensarbete görs en inledande studie rörande prediktering av LTB-värdet i KK2 (Kiruna Kulsinterverk 2). Om det är möjligt att göra numeriska prediktioner av LTB-värdet med acceptabelt prediktionsfel kommer LKAB att kunna spara stora pengar på att kraftigt reducera antalet tester. I arbetet ingår också försök att bestämma vilka variabler som påverkar LTB-värdet mest. Om det är möjligt att reda ut hur de olika processvariablerna påverkar LTB-värdet kan processen styras för att optimera LTB-värdet. Detta skulle vara mycket attraktivt för LKAB då de kan spara stora summor på detta. Under arbetets gång upptäcks dock att det troligen inte är möjligt att göra precisa numeriska prediktioner, därför koncentreras arbetet på att försöka göra prediktioner av typen ”stor risk för lågt LTB”.

1.3 Analysmetoder

I detta problem finns en variabel som skall predikteras (LTB) och ett antal potentiellt förklarande variabler. Alla prediktorer samt responsen mäts på intervallskala vilket förenklar tolkningen av resultaten. Flera av de förklarande

variablerna kan vara kraftigt korrelerade och det medför att minsta-kvadratmetoden ej kommer att ge tillförlitliga resultat. De metoder som i huvudsak används i detta examensarbete är PCR (Principal Components Regression) och PLSR (Partial Least Squares Regression). Båda dessa metoder kan används då de förklarande variablerna ej är okorrelerade. PCR är en PCA (Principal Components Analysis) följt av en klassisk regressionsanalys på de viktigaste principalkomponenterna. Det en PCA gör är att den hittar ortogonala linjärkombinationer av de förklarande variablerna som förklarar variationen inom de förklarande variablerna. En PLS däremot försöker hitta ortogonala linjärkombinationer av de förklarande variablerna som även är relevanta för respons-variabeln.

1.4 Rapportens struktur

Rapporten är upplagd på följande vis; i kapitel 2 beskrivs de valda analysmetoderna mer ingående samt modellval för prediktion och inflytelsediagnostik, i kapitel 3 presenteras data, kapitel 4 går igenom ett antal olika analyser, i kapitel 5 sammanfattas och diskuteras de viktigaste resultaten från arbetet och i kapitel 6 återfinns slutligen en akronymlista. I appendix A finns en variabelförklaring och i appendix B respektive C återfinns alla tabeller respektive figurer.

2 Teori

Detta kapitel börjar med en genomgång av olika metoder för linjär regression. Minsta-kvadrat-metoden fungerar dock dåligt då vi ej har okorrelerade variabler. Två metoder för att komma till rätta med detta problem är PCR och PLSR och dessa beskrivs i stycke 2.2 och 2.3. Stycke 2.4 och 2.5 tar upp modellval respektive inflytelsediagnostik.

2.1 Klassisk linjär regression

Den klassiska modellen för enkel linjär regression definieras som:

$$y_i = \alpha + \beta x_i + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2) \text{ [och oberoende]} \tag{1}$$

där y_i kallas responsen, x_i kallas den förklarande variabeln, interceptet α och lutningskoefficienten β är parametrar, ϵ_i är den stokastiska komponenten och brukar kallas försöksfelet i data och $i = 1 \dots n$ där n är antal mätningar. Om vi har mer än en förklarande variabel definieras den klassiska modellen för multipel linjär regression som:

$$y_i = \alpha + x_i\beta + \epsilon_i$$

där β är en $(p \times 1)$ vektor och x_i är en $(1 \times p)$ vektor med förklarande variabler. Minsta-kvadrat-skattningen av parametrarna α och β blir enligt ekvation (2) (för mer information om klassisk linjär regression se [12]).

$$\begin{aligned} \theta &= (\alpha\, \beta)' \\ A &= \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix} \\ \hat{\theta} &= (A'A)^{-1}A'Y \end{aligned} \tag{2}$$

2.2 PCR (Principal Components Regression)

En PCA hittar oberoende linjärkombinationer av de p förklarande variablerna. De nya variablerna kan skrivas:

$$T = XW$$

där score-matrisen T är en $(n \times p)$ matris och W är en $(p \times p)$ matris med vikter. Kolumnerna i W är egenvektorer till $Cov(X)$ eller $Corr(X)$ normerade till längd 1. Följande gäller:

$$\begin{aligned} Var(T_i) &= \lambda_i \\ Cov(T_i, T_k) &= 0 \quad i \neq k \end{aligned}$$

där T_i är kolumn i i T och det är nya variabel nummer i och λ_i är egenvärdet som hör ihop med egenvektor i. Egenvärdena och egenvektorerna är sorterade så att $(\lambda_1 \geq \lambda_2 \geq \dots)$. Andelen variation som den i:te PC (Principal Component, samma sak som T_i) förklarar är $\lambda_i / \sum \lambda_k$. PCA representerar X genom att projicera X på det rum som spännas upp av en delmängd av egenvektorerna till $Cov(X)$ eller $Corr(X)$. $\hat{X}$ blir då:

$$\hat{X} = XW_{(m)}W'_{(m)}$$

där $W_{(m)}$ betecknar matrisen av de m första kolumnerna i W. Då gäller att $m \leq p$. $W_{(m)}$ är den matris som minimerar summan av de kvadratiska skattningsfelen, alltså $\sum\sum(x_{ij} - \hat{x}_{ij})^2$, bland alla matriser av storlek $(p \times m)$ ([8] sid 462, [9] sid 7). I en PCR utförs först en PCA och sedan en regression av Y på T. Y kan då skrivas enligt:

$$Y = \alpha + X\beta + \epsilon = \alpha + XWW'\beta + \epsilon = \alpha + T\beta_{ny} + \epsilon \tag{3}$$

där ϵ är enligt ekvation (1). Från ekvation (3) kan det ses att en PCR är en klassisk linjär regression med de förklarande variablerna transformerade.

2.3 PLSR (Partial Least Squares Regression)

PLS-regression reducerar också, precis som PCR, dimension på prediktorrummet men en PLSR försöker inte bara förklara variationen inom prediktorvariablerna, utan den försöker simultant förklara variationen inom responsen. Det första steget i en PLSR går ut på att hitta ortogonala linjärkombinationer av de förklaranade variablerna och sedan utföra klassisk regression av responsen på dessa. Skillnaden mot PCR är hur dessa linjärkombinationer väljs. Antag att t_1 är en linjärkombination av X, där X en $(n \times p)$ matris. Antalet observationer är n och antal förklarande variabler är p. En linjärkombination kan skrivas $t_1 = Xw_1$, där w_1 är en $(p \times 1)$ vektor med vikter. En PLS-regressionsmodell kan då skrivas som ([2] sid 18, [9] sid 11):

$$\hat{Y} = \sum_{k=1}^{m} t_k c_k \tag{4}$$

där c_k är laddningen för PLSR-faktor k på den beroende variabeln (c_k är alltså en skalär) och m är antal PLSR faktorer som ska extraheras. Den första linjärkombinationen väljs så att $t'_1 Y$ maximeras under villkoret att $w'_1 w_1 = 1$. Den andra linjärkombinationen t_2 väljs sedan så att $t'_2 Y$ maximeras samt att $t'_1 t_2 = 0$. Även denna gång ska längden på viktvektorn vara 1. Låt $d = X'Y$ och D=$X'X$. Vi ska då först maximera $(w'_1 d)^2$ med villkoret att $w'_1 w_1 = 1$. Detta ger att $w_1 \propto d$. Sedan ska $(w'_2 d)^2$ maximeras med villkoren att längden av w_2 är 1 och att $w'_2 D w_1 = 0$. Detta ger att

$w_2 \propto d - (d'Dd/d'D^2d)Dd$. Om vi fortsätter på detta sätt kan alla viktvektorerna fås fram ([3] sid 72). Vi får sedan $c_1, ..., c_m$ genom klassisk regression av Y på $t_1, ..., t_m$.

PLSR kan också beskrivas som en algoritm. Låt D_0 vara en kopia av X, matrisen med prediktorer och låt F_0 vara en kopia av Y, vektorn med responsvärden. I de flesta fall bör D_0 och F_0 standardiseras. D_0 är en $(n \times p)$ matris och F_0 är en $(n \times 1)$ vektor, där p är antal förklarande variabler och n är antal observationer. Låt sedan $k = 1$. PLSR startar med en linjärkombination, $t_k = D_{k-1}w_k$, av prediktorerna, där score-vektorn t är en $(n \times 1)$ vektor och w är en $(p \times 1)$ vektor med vikter. PLSR predikterar sedan både D_{k-1} och F_{k-1} genom regression på t:

$$\begin{aligned} \hat{D}_{k-1} &= t_k\hat{p}_k' \quad där \quad \hat{p}_k' = (t_k't_k)^{-1}t_k'D_{k-1} \\ \hat{F}_{k-1} &= t_k\hat{c}_k \quad där \quad \hat{c}_k = (t_k't_k)^{-1}t_k'F_{k-1} \end{aligned} \tag{5}$$

där p kallas laddnings-vektorn (loading på engelska) för de förklarande variablerna och c laddningen för responsen. Den linjärkombination $t_k = D_{k-1}w_k$ som väljs är den som maximerar $t_k'F_{k-1}$ (i den ursprungliga algoritmen är det definierat så men det är även möjligt att byta ut F_{k-1} mot Y). Sedan beräknas residualerna:

$$\begin{aligned} D_k &= D_{k-1} - \hat{D}_{k-1} \\ F_k &= F_{k-1} - \hat{F}_{k-1} \end{aligned}$$

Sedan sätts $k = k + 1$ och nya skattningar av D och F beräknas och så fortsätter algoritmen till dess att k är lika med antal PLSR-faktorer som ska extraheras, antag att detta är m. Vanligen väljs m så att D_{m-1} och F_{m-1} är tillräckligt små enligt något kriterium.

Antag att X^* är observerade värden på prediktorerna. X^* kan för enkelhetens skull antas vara en $(1 \times p)$ vektor, där p är antal prediktorer. X^* standardiseras sedan genom följande beräkning:

$$X^*_{stdize} = \frac{X^* - \bar{X}}{s_x}$$

där X innehåller de observationer som användes vid modellbyggandet. Följande beräkningar görs sedan (dessa gäller även PCR):

$$\begin{aligned} T &= X^*_{stdize}W \\ \hat{Y}^*_{stdize} &= TC' \\ \hat{Y}^* &= \hat{Y}_{stdize}s_Y + \bar{Y} \end{aligned}$$

där Y är de observationer på responsvariablerna som används vid modellbyggandet, W är en $(p \times m)$ matris med vikter för de oberoende variablerna

(där p är antal variabler och m antal extraherade PLSR-faktorer eller antal PC) och C är en $(1 \times m)$ vektor med responsvariabelns laddningar (vid PCR ersätts C med de vanliga parameterskattningarna).

2.4 Metodval för prediktion

Ordet prediktion kommer från latin och är en ihopslagning av orden som betyder "före" och "att säga". Från denna definition är det tydligt att prediktion på något sätt handlar om att sia om något före det har hänt. AIAA (American Institute of Aeronautics and Astronautics) definierar prediktion som; "use of a computational model to foretell the state of a physical system under conditions for which the computational model has not been validated". En prediktion är enligt denna definition en simulering via beräkningar av ett specifikt fall där information saknas. En vanligare definition av prediktion är dock att även inkludera resultatet av beräkningar på data som är observerade. I detta arbete kommer en prediktion att definieras som ett framräknat LTB-värde. En prediktion är då det LTB-värde som den antagna modellen ger till en viss observation. Denna observation kan vara observerad eller icke observerad. Skillnaden mellan det uppmätta LTB-värdet och det framräknade LTB-värdet kallas residualen om observationen i fråga har används för att bygga modellen. Annars kallas denna skillnad prediktionsfelet. Prediktion är intressant ur LKAB:s synvinkel då en prediktionsmodell med acceptabelt prediktionsfel skulle eliminera behovet att göra ett stort antal dyra och tidskrävande pelletstester.

I detta arbete kommer prediktion att vara förknippat med regressionsmodeller, då prediktion definieras som ett från en regressionsmodell framräknat LTB-värde. Regression är bara korrelation sett från en annan synvinkel och korrelationssamband behöver inte betyda att det finns orsakssamband. Det kan tydligt ses genom att korrelation alltid är symmetrisk medan orsakssamband inte behöver vara symmetriska. Det betyder alltså att prediktion inte kräver orsakssamband vilket innebär att det inte finns någon möjlighet att reda ut eventuella orsakssamband mellan variabler, om dessa överhuvudtaget finns, med regression.

Korsvalidering är ett förfarande som ofta används för att hitta en lämplig prediktionsmodell. Korsvalidering används t.ex. för att bestämma antal PLSR eller PCR-faktorer. I korsvalidering delas datamängden in i ett antal grupper. Modellen anpassas sedan till alla grupper utom en och den sista gruppen används för att bestämma prediktionsfelet. Detta förfarande upprepas sedan för alla grupper och summan av kvadraterna på prediktionsfelen är en statistika som kallas $PRESS$ (Predicted REsidual Sum of Squares). Vanligen väljs sedan den modell som uppvisar lägsta $PRESS$-värdet eller en mindre modell som uppvisar obetydligt större $PRESS$ än modellen med

lägst $PRESS$. Formlerna för $PRESS$ i detta stycke gäller under minsta-kvadrat-skattningen. En vanlig statistika som ofta redovisas i samband med modellval är det så kallade Q^2-värdet. Q^2-värdet är ett mått på modellens prediktionsförmåga och beräknas enligt:

$$Q^2 = 1 - \frac{PRESS}{\sum (y_i - \bar{y})^2} \tag{6}$$

En variant av $PRESS$ fås då en observation åt gången "lämnas utanför". Det medför att $PRESS$ kan beräknas enligt ekvation (7) eller (11).

$$PRESS = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{(1 - h_{ii})^2} \tag{7}$$

$$H = X(X'X)^{-1}X' \tag{8}$$

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{9}$$

$$\hat{y}_i = x_i\hat{\beta} \tag{10}$$

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_{i(i)})^2 \tag{11}$$

$$\hat{\beta}_{(i)} = (X'_{(i)}X_{(i)})^{-1}X'_{(i)}Y_{(i)} \tag{12}$$

$$\hat{y}_{i(i)} = x_i\hat{\beta}_{(i)} \tag{13}$$

där X är design matrisen, h_{ii} är elementet på rad i och kolumn i i H, x_i är rad i i X och (i) betyder att den i:te observationen har raderats. Dessa båda uttryck för $PRESS$ är ju ekvivalenta.

$$X'X = X'_{(i)}X_{(i)} + x'_i x_i \tag{14}$$

$$X'Y = X'_{(i)}Y_{(i)} + x'_i y_i \tag{15}$$

$$(X'X - x'_i x_i)^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}x'_i x_i (X'X)^{-1}}{1 - x_i(X'X)^{-1}x'_i} \tag{16}$$

$$h_{ii} = x_i(X'X)^{-1}x'_i \tag{17}$$

Dessa uttryck kan lätt verifieras om uttrycken skrivs ut (i ekvation (16) multiplicera $(X'X - x'_i x_i)$ på båda sidorna samt utnyttja att $x_i(X'X)^{-1}x'_i$ är en skalär). Från ekvation (14), (16) och (17) fås:

$$(X'_{(i)}X_{(i)})^{-1} = (X'X - x'_i x_i)^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}x'_i x_i (X'X)^{-1}}{1 - h_{ii}} \tag{18}$$

Från ekvation (13), (12), (18) och (15) fås:

$$\begin{aligned}\hat{y}_{i(i)} &= x_i\left((X'X)^{-1} + \frac{(X'X)^{-1}x_i'x_i(X'X)^{-1}}{1-h_{ii}}\right)(X'Y - x_i'y_i) = \\ &= x_i(X'X)^{-1}X'Y - x_i(X'X)^{-1}x_iy_i + \\ &+ \frac{x_i(X'X)^{-1}x_i'x_i(X'X)^{-1}X'Y - x_i(X'X)^{-1}x_i'x_i(X'X)^{-1}x_i'y_i}{1-h_{ii}}\end{aligned} \tag{19}$$

Från ekvation (9), (10) och (17) fås ekvation (19) till:

$$\begin{aligned}\hat{y}_{i(i)} &= \hat{y}_i - h_{ii}y_i + \frac{h_{ii}\hat{y}_i - h_{ii}^2y_i}{1-h_{ii}} = \\ &= \frac{\hat{y}_i - h_{ii}y_i}{1-h_{ii}}\end{aligned} \tag{20}$$

Om ekvation (20) sätts in i ekvation (11) fås ekvation (7). Vi kan alltså i detta fall beräkna $\hat{y}_{i(i)}$ från den ursprungliga regression då all data används.

2.5 Inflytelsediagnostik

Inflytelserika observationer är de observationer som har stort inflytande på parameterskattningarna. Då dessa ensamt kan ha stor påverkan på modellen är det viktigt att undersöka om det finns några inflytelserika observationer i datamängden. Alla gränsvärden i detta stycke kommer från [7].

En statistika som ofta brukar redovisas i samband med inflytelsediagnostik är h_{ii}. Observationer som har h_{ii} större än $2p/n$ bör undersökas närmare, där p är antal parametrar i modellen, n är antal observationer och h_{ii} är enligt ekvation (17). En annan statistika som också brukar komma på tal i samband med inflytelsediagnostik är standardiserade residualer. En typ av standardiserade residualer fås genom att varians standardisera residualerna se ekvation (22).

$$r_i = (y_i - x_i\hat{\beta}) \tag{21}$$

$$\hat{\beta} = (X'X)^{-1}X'Y$$

$$Var(Y - X\hat{\beta}) = Var((I - X(X'X)^{-1}X')Y) = \sigma^2(I-H)$$

$$r_i^* = \frac{r_i}{s\sqrt{1-h_{ii}}} \tag{22}$$

där s är den skattade standardavvikelsen, x_i är rad i i X och r_i^* är den variansstandardiserade residualen (Student Residual i SAS). En annan standardiserad residual, kallad Rstudent i SAS, beräknas enligt följande:

$$r_{i(i)}^* = \frac{r_i}{s_{(i)}\sqrt{1-h_{ii}}}$$

där (i) betyder att observation i har tagits bort. Standardavvikelsen är därmed skattad utan observation i. Observationer som har $|r^*_{i(i)}|$ större än två bör undersökas närmare. Observationer som uppfyller följande kriterium kan också tyda på att de är inflytelserika:

$$CovRatio_i = \frac{det(s^2_{(i)}(X'_{(i)}X_{(i)})^{-1})}{det(s^2(X'X)^{-1})}$$
$$|CovRatio_i - 1| \geq 3p/n$$

där *det* är determinant. Stora värden på $|Dffits|$ och $|Dfbetas|$ kan också tyda på att en observation är inflytelserik. $Dffits$ och $Dfbetas$ definieras enligt:

$$Dffits_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{s_{(i)}\sqrt{(h_{ii})}}$$
$$Dfbetas_{ik} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{s_{(i)}\sqrt{((X'X)^{-1})_{kk}}}$$

där $k = 1 \ldots p$, där p är antal parametrar i modellen. Värden på $|Dffits|$ på över $2\sqrt{p/n}$ (eller på över 2) och värden på $|Dfbetas|$ på över $2/\sqrt{n}$ (eller på över 2) kan tyda på en observation är inflytelserik.

3 Projektets data

I avsnitt 3.1 respektive 3.2 beskrivs sensordata respektive sikt- och kemiskdata. I avsnitt 3.3 förklaras hur LTB-värdet mäts. I avsnitt 3.4 beskrivs hur data ser ut när vi får det från datasystemet och hur data transformeras för att kunna användas i vidare analyser.

3.1 Sensordata

Efter hela flödeskedjan sitter sensorer som kontinuerligt registrerar processparametrar (t.ex. tryck och temperaturer). Då det skulle krävas obegränsade dator resurser för att lagra kontinuerliga datamängder måste de filtreras på något sätt. I datasystemet sparas därför bara punkter som avviker mer än en på förhand bestämd gräns från det av operatören inställda processvärdet (se figur 1). Dessa gränser är också satta så att givarnas mätosäkerhet filtreras bort. Det gör dock att en del data kan vara kraftigt missvisande (se figur 2).

Detta sätt att filtrera data är nödvändigt också därför att processförändringar måste kunna upptäckas fort och då är inte sampling vid givna tidsintervall lämpligt. Det typiska är ändå att över 100 punkter sparas per dag. Medelvärdet av de sparade punkterna kommer dock ej att vara helt korrekt men viss information ligger ändå i att veta att en större andel punkter ligger på någon sida av gränserna.

Datapunkter kan också saknas, det kan vara orsakat av t.ex. sensorfel eller produktionsstopp.

3.2 Sikt- och kemisk data

I pelletsprocessen mäts viktsandelen (i procent) av pelletsen som ligger i följande intervall; <5 mm, 5-9 mm, 9-12.5 mm, 12.5-16 mm och >16 mm. Det är sedan dessa procentandeler som blir de fem siktvariablerna. Siktvariablerna sparas sedan i datasystemet som dygnsmedelvärden.

De variabler som representerarar den kemiska sammansättning är järnhalten, fosforhalten, järnoxidhalten, kiselhalten, kalciumoxidhalten och magnesiumoxidhalten. Alla dessa halter mäts i procent. Dessa variabler sparas också som dygnsmedelvärden i datasystemet.

3.3 Hur LTB mäts

LTB-analysen genomförs på ett representativt prov från pelletsprocessen. Ett av de första stegen i LTB-analysen är att sikta det representativa provet från pelletsprocessen så att bara pellets i storleken 10-12.5 mm återstår. Om

andelen pellets är stor som har en storlek över 12.5 så kommer medelvärdet av de siktade pelletsen troligtvis att vara närmare 12.5 än 10. Det dras sedan slumpmässigt pellets från den siktade pelletsmängden (alltså bland de med storlek 10-12.5) som LTB-värdet bestäms på. Ett högt medelvärde bland de siktade pelletsen kommer således troligtvis att innebära att de pellets LTB-värdet bestäms på har en stor medelstorlek. Om pelletsprocessen tillverkar en stor andel stora pellets så medför det därför att LTB-värdet troligen kommer att bestämmas på en pelletsmängd med stor medelstorlek.

När LTB-värdet ska mätas sätts en mängd pellets in i en speciell ugn. I denna ugn utsätts pellets för förhållanden liknande de som råder i de övre delarna på en masugn. När pelletsen har körts i ugnen beräknas sedan andelen massa med storlek på över 6.3 mm, andelen massa med storlek på upp till 3.15 mm och andelen massa med storlek upp till 0.5 mm. Det är viktigt att det är en stor andel med storlek över 6.3 mm då de mindre fraktionerna kan orsaka problem. I fortsättningen kommer LTB att underförstås betyda $LTB_{+6.3}$ och därmed andelen massa med en storlek på över 6.3 mm.

En icke fullstäng LTB-analys görs efter varje skift (det ger 3 stycken per dygn). Att den är icke fullständig innebär att bara ett replikat görs per pelletsmängd samt att vissa torkningsförfaranden inte utförs. Enligt ISO beskrivningen ska minst två replikat göras beroende på utfallet av dessa. En fullständig LTB-analys tar för mycket tid och det är kritiskt att snabbt få LTB feedback så därför görs inte fullständiga LTB-analyser. En fullständig LTB-analys utförs innan båtarna skickas i väg men då det är nästan omöjligt att koppla denna LTB-analys till specifika processinställningar kan den inte användas. Att det nästan inte går att koppla båtarnas LTB-analys till specifika processinställningar beror på att cykeltiderna är svåra att uppskatta. Att bara ett replikat utförs innebär också att det inte finns någon frihetsgrad att uppskatta variationen i mätmetoden.

3.4 Datainsamling

Alla datapunkter sparas i ett datasystem där bl.a. både värdet och tidpunkten sparas. Från datasystemet kan sedan Excel ark fås där värden på olika variabler vid olika tidpunkter framgår. Då LTB-värdet endast mäts tre gånger per dygn och nästan alla övriga variabler mäts kontinuerligt så behövs någon metod för att koppla LTB-värden till processinställningarna. Bedömningen gjordes att beräknande av dygnsmedelvärden på alla variabler är det bästa sättet att göra kopplingen. Försök att koppla processinställningar till varje enskilt LTB-värde ansågs vara för tidskrävande (speciellt då andra förklarande variabler såsom den kemiska sammansättningen sparas som dygnsmedelvärden).

Ett motargument mot att använda dygnsmedelvärden är att om den första LTB-mätningen för ett dygn är låg så kommer processingenjörerna att försöka rätta till det genom att ändra i processen. Om de två nästkommande LTB-mätningarna då ger höga värden kommer det första låga värdet inte att ha så stor påverkan. Detta argument håller dock inte helt då det inte är helt känt vilka variabler som påverkar LTB. Att det görs tre mätningar per dag har mer att göra med försök att undvika uppenbara produktionsfel. Om t.ex. den första LTB-mätningen för ett dygn är låg kan det vara orsakat av en felaktigt inställd processparameter som inte mäts. Det är dessa typer av misstag som gör att tre LTB-mätningar per dygn är nödvändigt.

4 Analys

I detta kapitel kommer ett antal olika analyser att presenteras. Kapitlet börjar beskriva varför vi bytte pellets typ att analysera och fortsätter sedan att beskriva hur storleksfördelningen på pellets analyseras. I stycke 4.3 finns en analys av LTB-värdenas fördelning och resten av kapitlet, förutom stycke 4.10 och 4.11, ägnas åt att beskriva olika modelltyper. I stycke 4.10 analyseras LTB-mätmetoden och i stycke 4.11 analyseras variationskomponenterna inom respektive mellan dygn.

4.1 Byte av datamängd

Arbetet började med att försöka prediktera LTB-värdet för en pellets som kallas KPBA (Kiruna Pellets Blastfurnace Acid). Det var dock mycket svårt att hitta någon modell som kunde förklara variationen i LTB för KPBA (varken PLSR eller PCR kunde hitta någon tillfredsställande modell). Orsaken till att processparametrarna kan förklara så liten del av LTB variationen kan vara att standardavvikelsen på LTB-mätningarna bara är runt 1% och medelvärdet är runt 97%. Enligt specifikationen för LTB-mätmetoden är det fullt normalt att två upprepade LTB-mätningar skiljer sig upp till två procentenheter. Det innebär att den variation i LTB som härrör från pelletsprocessen drunknar i den variation som härrör från LTB-mätmetoden.

För att komma till rätta med problemet att LTB-värdet varierar för lite väljs en annan pelletstyp, kallad KPBO (Kiruna Pellets Blastfurnace Olivin), där LTB-värdet varierar mer. LTB-medelvärdet för KPBO datamaterialet är cirka 87% och standardavvikelsen runt 6%. En nackdel med KBPO är dock att det bara finns 44 stycken fullständiga observationer på den (för KPBA finns det ett stort antal observationer).

4.2 Siktanalys

Pellets siktas efter storlek. Det finns fem siktvariabler som beskriver storleksfördelningen på pellets. Dessa fem variabler bör summera till 1 och utgör därför vad som kallas sammansättningsdata. Exempel på värden för dessa variabler kan ses i tabell 1. I tabell 1 är $Sikt1$=andelen pellets som är <5 mm, $Sikt2$= andelen pellets som är 5-9 mm, $Sikt3$=andelen pellets som är 9-12.5 mm, $Sikt4$ andelen pellets som är 12.5-16 mm och $Sikt5$ är andelen pellets som är >16 mm.

Låt S beteckna matrisen med siktdata, där s_{ij} är observation i på $Siktj$ (givet i andelar), när alla rader utan siktobservationer rensats bort. Egenvärdena och egenvektorerna, kalla dessa $e_1, \ldots, e_5$ där e_1 hör ihop med största egenvärdet

e_2 näst största osv., till $Cov(S)$ beräknas sedan. Det medför att e_i blir:

$$e_i = \begin{pmatrix} e_{1i} \\ \vdots \\ e_{Di} \end{pmatrix} \tag{23}$$

Antag att e_i och e_j är signifikanta egenvektorer, alltså egenvektorer med stort tillhörande egenvärde. Om $S*e_i$ plottat mot $S*e_j$ uppvisar något slags krökt samband kommer en vanlig PCA att ge felaktiga resultat på grund av det krökta sambandet ([1] sid 188)(i [6] beskrivs hur en PLSR bör anpassas för sammansättningsdata). Att denna plott uppvisar krökta samband är vanligt då sammansättningsdata analyseras. Om inga krökta samband finns så försämrar transformationen i ekvation (24) ändå inte analysen. Data ska transformeras enligt:

$$\begin{aligned} s^*_{ij} &= log(s_{ij}/g(\underline{s_i})) \\ \underline{s_i} &= (s_{i1} \dots s_{iD}) \\ g(\underline{s_i}) &= (s_{i1} * \cdots * s_{iD})^{1/D} \end{aligned} \tag{24}$$

där D är antal siktvariabler. Från transformationssambandet i ekvation (24) är det tydligt att komponenter i S med värde noll inte kommer att kunna transformeras. För att komma tillrätta med det kan kolumner i S slås ihop. Om t.ex. de allra flesta nollor återfinns i kolumn fem kan kolumn fem och fyra slås ihop till en kolumn där summan av värdet i kolumn fyra och fem återfinns. Om det även efter en sammanslagning finns komponenter med värde noll kvar så kan dessa tas med i analysen genom att byta ut nollan mot det minsta observerade värdet och det är 0.1. Det är inte heller säkert att en summering över kolumnerna ger värde 1. En anledning till varför summan avviker från 1 kan vara mätfel. För att komma till rätta med det utförs:

$$\tilde{s}_{ij} = s_{ij}/(s_{i1} + \cdots + s_{iD}) \tag{25}$$

En matris transformerad enligt ekvation (24) blir dock densamma oavsett om beräkningen i ekvation (25) utförts, $\tilde{S}^* = S^*$. Antalet kolumner i S efter att problemet med nollor har åtgärdats är d, där $d \leq D$. Sedan bildas S^* enligt ekvation (24). Efter detta beräknas egenvärden och egenvektorer, kalla dessa $e_1, \dots, e_d$ där e_1 hör ihop med största egenvärdet e_2 med det näst största osv., till $Cov(S^*)$. Siktvariablerna byts då ut mot:

$$p_i = S^* * e_i \tag{26}$$

eller mot:

$$p_i = log(S) * e_i \tag{27}$$

där e_i är enligt ekvation (23) men med D utbytt mot d. Ekvation (26) och (27) är ekvivalenta.

4.2.1 Siktanalys av KPBO

Om siktvariablerna för KPBO analyseras fås figur 3, figuren visar $S * e_2$ plottat mot $S * e_1$. Från denna figur är det tydligt att dessa har något slags krökt samband. Om siktvariablerna transformeras fås figur 4. I denna figur kan inga tydliga krökta samband ses.

De två största egenvärdena till $Cov(Sny^*)$ förklarar nästan all variation, se scree plotten i figur 5. De fem siktvariablerna på rad i byts därför ut mot:

$$p1 = +0.0669 * sny^*_{i1} - 0.6489 * sny^*_{i2} - 0.1590 * sny^*_{i3} + 0.7410 * sny^*_{i4} \tag{28}$$

$$p2 = -0.8633 * sny^*_{i1} + 0.2295 * sny^*_{i2} + 0.2922 * sny^*_{i3} + 0.3416 * sny^*_{i4} \tag{29}$$

där sny_{i4} innehåller summan av $sikt4$ och $sikt5$ på rad i, detta för att $sikt5$ ofta antar värdet noll.

4.3 Fördelningen för de förklarande variablerna

Ett absolutbelopp på *skevheten* som överstiger 1.5 är ett tecken på att variabeln bör tranformeras ([5] sid 209), där *skevheten* beräknas enligt ekvation (30) om variansen beräknas genom division med antalet frihetsgrader, ej antal observationer:

$$skevheten = \frac{n}{(n-1)(n-2)} \sum z_i^3 \tag{30}$$

där z är de standardiserade observationerna och n är antal observationer. Regression bygger inte på några antaganden om normalfördelning för vare sig responsen eller de förklarande variablerna men det finns exempel där resultatet av en PCA kraftigt förbättras om den görs på variabler som transformerats för att minska $|skevheten|$ (se [5] sid 209). Om de förklarande variablerna och/eller responsen är skeva och om det verkligen råder ett linjärt samband mellan dessa, så förstörs detta linjära samband ifall vi transformerar de förklarande variablerna och/eller responsen för att få bort *skevheten*. Skevhet är därför inget tillräckligt motiv för att transformera. Stark skevhet innebär dock ofta att vissa observationer får en oönskat hög grad av inflytande på det anpassade sambandet. Ofta kan också linjära samband lättare hittas om transforamtioner för att få bort höggradig skevhet används.

Om *skevheten* beräknas på de förklarande variablerna så upptäcks att många av dem har stora absolutbelopp på *skevheten*. Den enda transformationen som verkar mothjälpa *skevheten* i dessa variabler är upphöjt till. Om dessa variabler bara upphöjs till ett nog stort tal går det att få ner *skevheten*. Det dock ytterst marginell skillnad i förklarandegrad om analysen görs med transformerade variabler, därför kommer i fortsättningen otransformerade variabler att användas.

4.4 LTB-analys av KPBO

Figur 6, 7 och 8 visar scatterplottar av LTB mot de förklarande variablerna. Från dessa scatterplotter är det svårt att urskilja några tydliga trender. Det verkar dock som att låga värden på p_1, där p_1 är enligt ekvation (28), medför låga LTB-värden.

Datamängden för KPBO består av 144 observationer det är dock bara 44 observationer som har värden för alla variabler. Till att börja med kommer bara dessa 44 observationer att användas. För att bestämma hur många PLSR-komponenter som ska användas används korsvalidering. I denna analys har data delats upp i sju grupper, det finns ingen djupare eftertanke med just sju men SAS föreslår det som ett bra nummer.

Från tabell 2 kan det ses att modellen med fem PLSR-komponenter har lägsta $PRESS$-värdet. Från denna tabell kan det också ses att modellen med fyra PLSR-komponenter inte har ett statistiskt signifikant större $PRESS$-värde än modellen med fem komponenter, därför väljs modellen med fyra komponenter. Från tabell 3 kan det ses att modellen med fyra PLSR-komponenter förklarar 88% av variationen i LTB med hjälp av 67% av variationen i de förklarande variablerna.

En PLSR-modell har formen, se avsnitt 2.3 (i både [11] och [4] finns exempel på hur PLSR kan användas):

$$X = TP' + R_x$$
$$Y = TC' + R_y$$

där T=X-score, P=X-laddning, R_x=X-residual, C=Y-laddning och R_y=Y-residual. PLSR-algoritmen hittar sedan ortogonala faktorer som maximerar kovariansen mellan varje X-score och respektive Y-score, där korrelationen ofta minskar från faktor till faktor. Då vi bara har en respons är Y-score vektorerna lika med de successiva F vektorerna, se avsnitt 2.3. Y-score1 är då lika med Y och högre Y-score är då de successiva residualerna. Vi följer här terminologin från programpaketet SAS (se [11]). De första faktorerna bör därför uppvisa en stark korrelation mellan X-score och Y-score vektorerna i en bra PLSR-modell. I figurerna 9, 10, 11 och 12 kan Y-score plottat mot X-score ses, där siffrorna i plottarna är observationsnumret. Korrelationen mellan första PLSR-komponentens X-score och Y-score är runt 0.75, för andra komponenten är korrelationen runt 0.68 och för de två sista runt 0.55.

I figur 13, 14 och 15 följer plottar av X-score mot varandra. Om dessa uppvisar krökta trender eller grupperingar kan vissa modifikationer i analysen behöva göras. Om dessa t.ex. uppvisar två tydliga grupperingar kan det vara

att föredra att analysera dessa två grupper separat. Dessa figurer uppvisar inte några tydliga grupperingar eller krökta samband (vid krökta samband kan högre ordningens termer måsta inkluderas, det gäller även vid krökta samband i Y-score mot X-score plottarna).

I figur 16 plottas residualerna mot predikterat LTB-värde och i figur 17 föjer en normalfördelningsplott av residual vektorn för responsen. Från dessa kan det ses att residualerna verkar vara tämligen väl normalfördelade men med lite tyngre svansar, samt att plotten mellan residualer och predikterat värde inte uppvisar någon tydlig struktur. Eventuell systematik i denna plott kan tolkas som en avviklese från modellen. Hypotesen att residualerna är normalfördelade kan inte heller förkastas på 5% nivån med Shapiro-Wilks normalitetstest.

I figur 18 och 19 följer plottar av avståndet från observationerna till modellen för de förklarande variablerna respektive responsen. Från dessa kan det ses att ingen observation riktigt sticker ut.

I plottarna 20 och 21 kan det ses hur stor påverkan respektive variabel har för de olika PLSR-komponenterna. I dessa plottar används X-weight (vikten på svenska, se avsnitt 2.3) men det hade gått lika bra att använda X-laddning då dessa vanligtvis är mycket lika. Från figurerna kan det ses direkt att KemFe och KemMgO har stor påverkan. En mindre subjektiv metod för att avgöra vilka variabler som har stor påverkan finns i avsnitt 4.5.

4.5 KPBO reducerad modell

För att avgöra vilka variabler som varken kan förklara variation i de förklarande variablerna eller responsen beräknas regressionskoefficienterna och respektive VIP-värde (Variable Importance for the Projection), se ekvation (31). Regressionskoefficienterna (B1 i tabell 4) avgör hur viktig varje förklarande variabel är för att prediktera responsen. VIP-värdet anger hur viktig varje förklarande variabel är när det gäller att förklara variationen i både de förklarande variablerna och responsen. Om en variabel har litet absolutbelopp av regressionskoefficienten och ett litet VIP-värde bör man överväga att ta den ur analysen. Ett litet VIP-värde är ett värde under 0.8 ([11] sid 8).

$$VIP_j = \sqrt{\frac{p * \sum_{k=1}^{m}(R^2(y;t_k) * w_{jk}^2)}{R^2(y;t_1 \dots t_m)}} \tag{31}$$

där $j = 1 \dots p$, p är antal prediktorvariabler, m är antal PLSR-faktorer, w_{jk} är de normaliserade vikterna för variabel j och PLSR-faktor k och $R^2(y;t_1 \dots t_m)$ är andelen variation i y som förklaras av PLSR-faktorerna $t_1 \dots t_m$. Följande variabler tas därför bort ur analysen: $Ovr46$, $Ovr41$,

$Ovr30$, $Ovr29$, $Ovr28$, $Ovr27$, $Ovr26$, $Ovr22$, $Ovr20$, $Ovr18$, $Ovr14$, $Ovr12$, $Ovr10$, $Ovr6$, $Ovr4$, $Ovr3$, $Ovr2$ och $Ovr1$. En PLSR med dessa faktorer borttagna blir enligt tabell 5. Från tabellen kan det ses att skillnaden mot då alla variablerna används är marginell när det handlar om antalet procent variation som modellen förklarar (jämför med tabell 3). Varken en normalfördelningsplott av residual vektorn för responsen eller plottar av Y-score mot X-score är nämnvärt förändrade. Korrelationen mellan X-score och Y-score är 0.76, 0.70, 0.57 och 0.41. Korrelationen mellan score vektorerna för den fjärde PLSR-komponenten har alltså sjunkit kraftigt medan korrelationen har ökat lite för de tre övriga.

När dessa variabler tagits bort så medför det att 46 stycken observationer som tidigare varit ofullständiga nu blivit fullständiga. Dessa observationer kommer att används till att validera modellen. I vanliga fall delas data upp i en modellanpassningsdatamängd och en valideringsdatamängd. I detta fall tillkom så många fullständiga observationer när ovan nämnda variabler eliminerades att dessa räcker för att validera modellen. I figur 22 kan skillnaden mellan LTB och predikterat LTB ses.

ISO standarden specificerar en differans på mellan 2.5 och 3 procentenheter (beroende på LTB-värdet) på två upprepade LTB-mätningar som normalt. Det medför, då bara en mätning görs, att det sanna LTB-värdet lika gärna kan ligga upp till 3 procentenheter över eller under det uppmätta värdet. LTB-mätningen inducerar därmed en mätvarians på upptill 6 procentenheter på mätningar på pelletsmängder med samma sanna LTB-värde. Om t.ex. en LTB-mätning ska göras på en pelletsmängd med 85% som sant LTB-värde är det helt normalt att resultatet av denna mätning blir 88% men resultatet hade lika gärna kunnat bli 82%. Det är alltså inte ovanligt att skillnaden mellan två mätningar på två olika pelletsmängder med samma sanna LTB-värde är upptill 6 procentenheter. Det gör att snäva prediktionsgränser omöjliggörs då upptill ± 3 procentenheters fel härrör från LTB-mätningen. I figur 23 kan residualerna och prediktionsfelen plottat mot de predikterade LTB-värdena ses. Medelvärdesbildningen över dygn minskar dock variansen inom dygn och därmed det inducerade mätfelet inomdygn, om samma mätutrustning hade använts. Ett problem är dock att vi inte vet om de olika mätningarna inom samma dygn har utförts med samma mätutrustning.

I figur 24 kan avståndet från observationerna till modellen ses för de förklarande variablerna. Det kan ses att de observationer på de förklarande variablerna som används för modellanpassningen har kortare avstånd till modellen än de observationer som används för valideringen. De observationer som användes för modellanpassningen är därför inte helt representativa för hela datamängden.

4.6 LTB-analys av KPBO utan $Ovr30$ (Kolflödet)

Om det antas att kolflödet inte påverkar LTB-värdet så ökar antalet fullständiga observationer från 44 till 90 stycken (att det saknas observationer på kolflödet beror på ombyggnationer). Dessa 90 observationer delas sedan upp i en datamängd med jämna observationsnummer (datamängd1), den innehåller 43 observationer och en mängd med udda observationsnummer (datamängd2), med 47 observationer.

Om en PLSR görs på datamängd1 eller datamängd2 upptäcks att andelen förklarad variation är lägre än i den tidigare analysen då kolflödet ej uteslutits. Den tidigare analysen visade dock att kolflödet inte skulle ha nämnvärd påverkan på LTB. Orsaken till att andelen förklarad variation sjunkit från den första modellen med 44 observationer är därför troligen mer beroende på att det är större spridning på observationerna än att kolflödet tagits ur modellen. Modellen byggd med datamängd1 har också lägre prediktionsfel än den tidigare modellen. I fortsättningen kommer alla analyser att göras utan kolflödet, detta för att få fler fullständiga observationer.

Ett problem är dock att de mest signifikanta variablerna inte är desamma oberoende av vilken datamängd som används för att bygga modellen. Om datamängd2 används för att bygga modellen så fås andra variabler som mest signifikanta (stort absolutbelopp på den standardiserade regressionskoefficienten och stort VIP-värde, se ekvation (31)) än om datamängd1 används. Oavsett vilken datamängd som används för att bygga modellen är dock inte prediktionsfelen tillfredsställande små.

Om observationerna delas upp efter observationsnummer fås att medel LTB-värdet för de första 45 observationerna är runt 84% och standardavvikelsen är runt 6.5%. För de 45 sista observationerna är medel LTB-värdet runt 91% och standardavvikelsen är runt 2.6%. För de 45 sista observationerna är variationen så liten att den variation i LTB som härrör från pelletsprocessen drunknar i den variation som härrör från mätutrustningen. Om datamängden visuellt analyseras upptäcks att de allra flesta observationer med lågt LTB-värde kommer från en avgränsad tidsperiod (juni 04). Observationer som är tidsmässigt nära varandra kan vara mer lika än man önskar av ett oberoende stickprov. Det kan därför vara så att de observationer med lågt LTB inom denna tidsperiod kan ha haft något gemensamt som inte behöver kännetecknа observationer med lågt LTB från en annan tidsperiod.

I tabell 6 kan det ses att vissa variabler har stor skillnad på totala medelvärdet och medelvärdet för juni 04. De variabler som har större skillnad än 30% på medelvärdena är; $KemFeO$, $Ovr7$, $Ovr9$, $Ovr29$, $Ovr34$, $Ovr36$, $Ovr37$, $Ovr43$, $Ovr44$ och $p1$.

4.7 $LTB2$-analys av KPBO

Meningen med $LTB2$ är att undersöka om det går att avgöra med hjälp av processparametrarna om LTB-värdet kommer att över- eller understiga 80%, därför definieras $LTB2$ som:

$$LTB2 = \begin{cases} 1 & \text{om } LTB > 80 \\ 0 & \text{om } LTB < 80 \end{cases}$$

$\widehat{LTB2}$ definerias som det av modellen predikterade $LTB2$-värdet. $\widehat{LTB2}$ avrundas på vanligt sätt, $\widehat{LTB2}$ avrundas därmed till 1 om $\widehat{LTB2}$ >0.5 annars avrundas den till 0. Om varannan observation används för att bygga modellen och de övriga för att validera modellen fås att prediktionen för tre observationer blir fel (med PLSR och två faktorer). Detta oavsett om jämna observationsnummer används för att bygga modellen och udda för att validera eller tvärt om. I båda fallen predikteras dock den observation som har lägst LTB och som inte är från juni 04 fel (de flesta observationer med lågt LTB kommer ju från juni 04, se avsnitt 4.6). Det kan tala för att vi predikterar "junivecka 04" istället för "lågt LTB".

4.8 Analys av KPBO med bara låga LTB

Meningen med $LTB2$, se avsnitt 4.7, är att avgöra om vi kommer att få ett högt eller lågt LTB-värde. Vi är vidare bara intresserade av låga LTB och speciellt om det går att göra numeriska prediktioner av dessa.

För att se om dessa observationer med lågt LTB passar in i det normala variationsmönstret för de olika förklaringsvariablerna utgår vi från standardiserade värden och söker variabler där lågt LTB hänger ihop med konsekvent och extremt utfall av en förklarande variabel. Om datamängden standardiseras och de observationer med LTB större än -1.28 elimineras så återstår nio stycken observationer, kalla denna datamängd datamängdQ. Vi valde -1.28 för att det motsvarar den punkt där den standardiserade normalfördelningen har 90% av massan till höger samt att -1.28 i standardiserat LTB motsvarar 79.9 i icke standardiserat LTB. Vi behåller i princip bara de observationer där $LTB2$=0. I figur 25, 26 och 27 följer plotter av LTB mot de förklarande variablerna för datamängdQ. De flesta obervationer i datamängdQ kommer från juni -04 (se avsnitt 4.6), endast en av dem kommer inte från den perioden. Denna observation är markerad med * i figurerna.

Kalla de förklarande variablerna i datamängdQ för z_{ij}, där i är observationsnummer och j är variabel nummer. Sedan beräknas följande matris (viktigt att notera är att datamängdQ inte längre är en standardiserad datamängd):

$$T = \begin{pmatrix} \frac{1}{9}\sum_{i=1}^{9} z_{i1}^2 & \cdots & \frac{1}{9}\sum_{i=1}^{9} z_{i1}z_{ip} \\ \vdots & \ddots & \vdots \\ \frac{1}{9}\sum_{i=1}^{9} z_{i1}z_{ip} & \cdots & \frac{1}{9}\sum_{i=1}^{9} z_{ip}^2 \end{pmatrix} \tag{32}$$

denna matris är en skattning av korrelationsmatrisen då vi vet att alla variabler är dragna ur en mängd med medelvärde 0 och varians 1, om variablerna är orelaterade till låga LTB. Om och endast om en z-variabel inte har inflytande på $LTB2$ så skattar dess diagonalelement i T motsvariande varians i z, som då har värdet 1 genom standardiseringen. Om t_{ii} är stort så förväntas variabel i ha ett samband med låga LTB-värden, där t är elementen i T. Diagonalen i T är då medelvärdet av de kvadratiska avvikelserna från medelvärdet i den ursprungliga datamängden för varje variabel. Om t_{ii} är stort så betyder det att en större andel än normalt av variabel i:s värden ligger längre från centrum i den ursprungliga fördelningen än normalt. De variabler som i figurerna 25, 26 och 27 inte har observationer spridda runt noll är ett tecken på att de har ett samband med lågt LTB. Vissa variabler har dock inte den observation som inte kommer från juni 04 på samma sida om nollan som de som kommer från juni 04 (se t.ex. $Ovr44$ i figur 26). Det kan vara ett tecken på att det inte finns något samband mellan lågt LTB-värde och onormalt variabel värde.

Sedan beräknas egenvärden och egenvektorer till T. Egenvärdena till T blir enligt figur 28 där alla egenvärden från och med egenvärde nummer tio är noll. Detta beror på att vi bara har nio stycken observationer. Sedan beräknas följande:

$$C = \begin{pmatrix} c_{11} & \dots & c_{19} \\ \vdots & \ddots & \vdots \\ c_{91} & \dots & c_{99} \end{pmatrix} = \begin{pmatrix} z_{11} & \dots & z_{1p} \\ \vdots & \ddots & \vdots \\ z_{91} & \dots & z_{9p} \end{pmatrix} * \begin{pmatrix} e_{11} & \dots & e_{19} \\ \vdots & \ddots & \vdots \\ e_{p1} & \dots & e_{p9} \end{pmatrix} \tag{33}$$

där $(e_{11} \dots e_{p1})'$ är första egenvektorn till T och p är antalet förklarande variabler. I figur 29 kan resultatet av en regressionsanalys med bara den första PC (alltså den första kolumnen i C kalla den $c1$) som förklarande variabel ses och inget intercept. Om ingen förklarande variabel har samband med lågt LTB så kommer de att ha z-värden runt 0 och detta medför att $c1$ kommer vara runt 0. Om $c1$ är 0 och alltså de förklarande variablerna inte har något med lågt LTB att göra, så vill vi att predikterat LTB ska bli ett normalt värde i den ursprungliga datamängden och detta fås då predikterat LTB=0. $PRESS$-värdet fås till 7.529, där $PRESS$ kan beräknas enligt ekvation (7) eller (11). Om Q^2 beräknas fås det till 0.864, se ekvation (6).

För att avgöra vilka variabler som har störst inflytande kan egenvektorerna undersökas. I tabell 7 kan egenvektorn som hör ihop med det största egenvärdet ses. De variabler som har störst inflytande är därmed $KemFeO$, $Ovr25$, $Ovr37$, $Ovr38$, $Ovr44$ och $p1$. I tabell 8 kan uppmätt och predikterat LTB-värde ses.

I tabell 9 kan inflytelsediagnostik för regressionsmodellen med $c1$ som förklarande variabel ses, där $c1$ är första kolumnen i C se ekvation (33). Det betyder att

$c1$ är första principalkomponenten vid en PCA utan vare sig medelvärdes- eller variansstandardisering. Det kan ses att ingen observation verkar vara speciellt inflytelserik, se avsnitt 2.5. Tabell 10 visar prediktionsintervall för denna modell, där prediktionsintervallet beräknas enligt ekvation (34). Från denna tabell framgår det att 95%:iga prediktionsintervall är runt fyra enheter breda då LTB är standardiserad och det blir runt 24 procentenheter för icke standardiserade LTB.

$$x_i\hat{\beta} \pm t_{\alpha/2}(n-1)\sqrt{(1+h_{ii})s^2} \tag{34}$$

Tabell 11 visar resultatet av en PLSR med en faktor (istället för en PCR som tidigare). Vi valde ju ut de observationer med standardiserat LTB<-1.28. Vi gör sedan en PLS-regression med dessa utvalda observationer utan att genomföra en till standardisering (i SAS specificeras detta som "no-center" och "noscale"). Om tabell 11 och figur 29 jämförs kan det ses att PLS-regressionen ger lite högre förklarandegrad (också Q^2-värdet blir något högre). Regressionskoefficienten blir 0.4790 och de förklarande variablernas vikter kan ses i tabell 12. Från denna tabell kan det ses att variablerna $KemFeO$, $Ovr25$, $Ovr37$, $Ovr38$, $Ovr44$ och $p1$ har störst inflytande. Det är alltså samma variabler som då PCR används. I figur 30 kan laddnings-vektorn för en PCR plottat mot laddnings-vektorn för en PLSR ses. Denna figur visar vilka variabler som är mest betydelsefulla för både en PCR och PLSR, med en PC respektive en faktor. I figur 31 kan X-score för en PCR plottat mot X-score för en PLSR ses. Denna figur visar vilka observationer som är mest extrema.

Om en PCR görs och två principalkomponenter, alltså kolumn ett och två C, används som förklarande variabler och utan intercept fås $PRESS$-värdet till 2.745. Om Q^2 beräknas fås det till 0.950, se ekvation (6). Ett 95%:igt prediktionsintervall blir också lite snävare om två PC används. Det bredaste intervallet har en bredd på runt 17.5 procentenheter för icke standardiserat LTB. Om en PLSR används istället så kommer predikteringen för alla observationer att komma närmare det uppmätta värdet. En PLSR ger typiskt värden som ligger runt 1 procentenhet närmare det uppmätta värdet, för icke standardiserat LTB. Den största avvikelsen mellan det predikterade och det uppmätta värdet (alltså residualen) blir runt fyra procentenheter, för icke standardiserat LTB, då PLSR används med två faktorer. Det är dock inte helt samma variabler som är mest signifikanta nu jämfört med då bara en PC användes. För PCR-analysen kan regressionskoefficienterna ses i tabell 13. För PLS-regressionen kan regressionskoefficienterna ses i tabell 14. I figur 32, 33 och 34 kan score-plottar och laddnings-plotttar för en PCR, med två PC, ses. I figur 35, 36 och 37 kan motsvarande för en PLSR, med två faktorer, ses.

Det kan noteras att både score och laddning är nästan identiska för PLSR och PCR. Detta oavsett om en eller två faktorer används.

4.9 Frekvenstabeller för KK2 och KK3

Då det verkar vara svårt att hitta någon modell som med stor säkerhet kan prediktera LTB-värdet gör vi ett försök att göra prediktioner av typen ”stor risk för lågt LTB”. För att undersöka effektiviteten för en prediktion av denna typ görs frekvenstabell där LTB respektive predikterat LTB klassificeras som lågt respektive normalt. Predikterat LTB beräknas med PLSR och korsvalidering.

4.9.1 Frekvenstabeller för KK2

I data för KK2 fås antalet faktorer till tre stycken. Från tabell 15 kan det ses att denna typ av prediktion verkar fungera väldigt bra. Ett allvarligt problem är dock att nästan alla låga LTB-värden kommer från samma tidsperiod (se avsnitt 4.6). För att undersöka hur denna typ av analys fungerar då de låga LTB-värdena är mer spridda görs också frekvenstabeller för LTB-värdet i KK3.

4.9.2 Frekvenstabeller för KK3

I data för KK3 är de låga LTB-värden mycket mer spridda än i data för KK2. Då KK3 är ett helt annat pelletsverk är därför inte processerna helt desamma och vissa förklarande variabler kommer därför också att skilja sig. För KK3 finns 56 stycken fullständiga observationer på KPBO, medelvärdet på LTB för dessa är runt 75.2 och standardavvikelsen är runt 5.4%. Motsvarande frekvenstabeller för KK3 data kan ses i tabell 16, 17 och 18. Från dessa tabeller kan det ses att andelen fel blir större då data från KK3 används. Det kan också påpekas att LTB-värdet för KPBO i KK3 ligger så mycket lägre än det gör i KK2 så att 80% gränsen fungerar åt motsatt håll. För KK3 har några observationer högre LTB än 80% och för KK2 har några observationer lägre LTB än 80%.

4.10 Analys av LTB-mätmetoden

För att undersöka LTB-mätmetodens varians utnyttjas att det har körts dubbelprover (två replikat) i LTB-ugnarna. Alla data i detta stycke kommer från KK2. Det har körts 27 dubbelprover i ugn 1 och 26 dubbelprover i ugn 0 och 3. För att uppskatta de olika ugnarnas varianser görs variansanalyser med mättillfälle som klassvariabel, se tabell 19, 20 och 21. Det intressanta i denna analys är att jämföra variansskattningarna för de olika ugnarna. För ugn 1 och 3 blir standardavvikelsen runt 1.3 och för ugn 0 blir den runt 1.08. För att en F-fördelad kvot mellan två variansskattningar med 26 frihetsgrader ska vara signifikant på 5%-nivån skall variansskattningarna skilja sig minst med en faktor av 1.93. Det finns således ingen anledning att tro att ugnarnas varians skiljer sig åt. Medelvärdet av ugnarnas varianser

ger en standardavvikelse på runt 1.24. För normalfördelningen ligger cirka 95% av massan inom $\pm 2\sigma$. Om LTB-värdet vore normalfördelat förväntas alltså mätmetoden inducera en störning mindre eller lika med ± 2.48 95% av tiden. Enligt ISO beskrivningen ska skillnaden mellan två replikat inte överstiga 3.0 för dessa LTB-observationer.

För ugn 1 och 0 hör 12 av dessa dubbelprover parvis samman. Dessa 12 observationer kan användas för att testa hypotesen att det inte är någon skillnad mellan ugnarna (variansanalys tvåsidig indelning, med observationstillfälle och ugn som klassvariabler, och två observationer per cell). Eftersom två observationer finns per cell kan även samspelstermen inkluderas i modellen, se tabell 22. Det fås att det är en signifikant skillnad mellan ugnarna (se figur 38 för en plott av LTB-värdena). Att samspelstermen inte är signifikant visar att skillnaden är tämligen konstant under den studerade tidsperioden.

En variansanalys med dygn och position inom dygn som klassvariabler ger att position inom dygn inte är signifikant. Det betyder att vi kan utgå från att variationen inom dygn är rent slumpmässig.

4.11 Variation inom respektive mellan dygn

Att dygnsmedelvärden används på LTB spelar inte någon roll så länge det är lika många observationer per dygn. Ett problem är dock att n_i kan variera ibland, där n_i är antal observationer för dygn i. Vissa dygn görs bara en LTB-observation, det kan vara orsakat av t.ex. produktionstopp. En vägd regression med $1/n_i$ som vikt går inte heller i detta fall eftersom flera observationer under ett dygn bara minskar varianskomponenten inom dygn. Vi har att $\sigma^2_{total} = \sigma^2_{mellan} + \sigma^2_{inom}/n_i$ och vi känner inte de olika varianskomponenternas storlek. En oviktad regressionsanalys kan också vara att föredra framför en viktad regressionsanalys om inte varianserna uppvisar stor heterogenitet ([10] sid 18). Vi kan dock skatta inomdygnsvariansen genom att beräkna variansen för alla dygn med tre observationer och sedan poola ihop dessa (dvs. medelvärdesbilda). För båda verken blir $\hat{\sigma}^2_{inom}$ runt 14 och $\hat{\sigma}^2_{mätmetoden}$ blir runt 1.5 (se avsnitt 4.10).

Om vi visste att samma LTB-ugn använts för alla mätningar inom ett dygn kunde vi dra slutsatsen att mätmetoden står för en relativt liten del av variationen inom dygn. När vi inte vet vilka mätningar som är gjorda med samma ugn kan vi inte dra några slutsatser av denna typ eftersom det enligt variansanalysen är en signifikant skillnad mellan LTB-ugnarna.

Variansen för residualerna i en PLSR-modell där korsvalidering har använts blir för KK2 runt 8.8 (tre faktorer) och för KK3 runt 17.2 (en faktor). Den variansskattning som vanligtvis används är $(1/(n-m))\sum_{i=1}^{n} r_i^2$ då den är

unbiased, där m är antal parametrar i modellen och r_i är enligt ekvation (21). Skillnaden är dock liten från $Var(r_i)$ då $m << n$. I en bra modell borde $\sigma^2_{inom}/3$ vara av samma storleksordning som $Var(residualerna)$ pga. medelvärdesbildningen över dygn, om det finns tre observationer de flesta dygnen. För KK2 har runt 79% av de fullständiga observationerna tre mätningar per dygn och motsvarande siffra för KK3 är runt 68%. För KK2 är $\sigma^2_{inom}/3$ runt 4.9 och för KK3 är den runt 4.6. Orsaken till att variansen för residualerna i modellen för KK3 är större än i modellen för KK2 måste ha att göra med att det i princip bara finns en avgränsad period med låga LTB i KK2, medan de låga LTB-värdena är mer spridda i KK3. Modellens uppgift är att försöka förklara σ^2_{mellan}. Residualerna indikerar att KK3 har betydligt större residualvarians än KK2, men ungefär samma σ^2_{inom}. Således lyckas modellen för KK3 sämre med att förklaringa variationen mellan dygn. För KK2 finns inte så mycket variation kvar som skulle kunna förklaras av de förklarande variablerna. För KK3 däremot finns större potential till förbättring. Modellens oförmåga att förklara en större del av residualvariansen indikerar dock att de tillgängliga förklarande variablerna inte lyckas förklara låga LTB-värden (det behövs dock en djupare analys för att dra säkrare slutsatser). Modellen för KK2 förklarar 76% av variationen i LTB med hjälp av 56% av variationen i de förklarande variablerna (PLSR med tre faktorer). Motsvarande siffror för KK3 är att 41% av variationen i LTB förklaras med hjälp av 22% av variationen i de förklarande variablerna (PLSR en faktor).

5 Diskussion och slutsatser

Detta examensarbete går ut på att göra en pilotstudie angående möjligheten att prediktera metallurgiska kvalitetsparametrar. Målsättningen med examensarbetet har lyckats ur den synpunkten att en inledande studie angående möjligheten att prediktera LTB har gjorts. Resultatet av denna inledande studie är dock att det troligen inte är möjligt att göra numeriska prediktioner av LTB med acceptabelt prediktionsfel, i alla fall inte med den datamängd som finns tillgänglig för detta examensarbete. Rekommendationen efter utförandet av detta examensarbete är att fortsätta med att försöka göra prediktioner av typen "stor risk för lågt LTB". Vidare var en del av målsättningen att försöka bestämma vilka variabler som påverkar LTB-värdet mest. Med hjälp av både data från KK2 och KK3 kan vi med stor säkerhet peka ut några variabler som påverkar LTB-värdet.

Då nästan alla observationer med lågt LTB finns inom en avgränsad tidsperiod för KK2 innebär det att det är svårt att dra några generella slutsatser om vilka variabler som påverkar så att vi får låga LTB-värden. Nedan följer två exempel från KK2 som visar på svårigheten då nästan alla låga LTB-observationer är tidsmässigt nära varandra.

Effekten på fläkt 053 ($Ovr44$) har enligt våra analyser stor påverkan på LTB-värdet och dess påverkan är positiv. Vi vill därför ha stora värden på denna variabel. För att kunna dra säkrare slutsatser skulle det dock behövas fler observationer med lågt LTB-värde som är tidsmässigt avskilda eftersom vi tidigare sett att de observationer med lågt LTB som kommer från juni 04 inte är konsekvent med den låga LTB-observation som inte kommer från juni 04. För åtta av de nio lägsta LTB-observationerna är det låg effekt på denna fläkt. Dessa åtta observationer är dock tidsmässigt nära varandra. För den låga LTB-observationen som är tidsmässigt avskild är det normal effekt på denna fläkt. Det skulle kunna betyda att det bara råkar vara låg effekt på denna fläkt i juni 04 och att fläkten inte påverkar LTB-värdet.

Här är ett till exempel på tolknings svårigheter då vi har observationer som är tidsmässigt nära varandra. Vi har i den här studien sett att $Ovr43$ verkar ha ett samband med LTB-värdet. Det skulle därför vara lätt att dra slutsatsen att LTB-värdet påverkas av $Ovr43$. $Ovr43$ är dock utomhustemperaturen och med tanke på att de flesta låga LTB-observationer kommer från juni 04 blir bilden mer osäker (juni har ju högre medeltemperatur än normalt). Detta exempel visar också på svårigheten att dra några slutsatser när de låga LTB-observationerna är så tidsmässigt nära varandra.

För att kunna dra säkrare slutsatser jämför vi de olika variablernas påverkan på LTB från modeller från både KK2 och KK3 data. Från de analyser vi gjort på KK2 och KK3 data är sikten alltid signifikant och vi vill ha en

stor andel kulor i $sikt2$ och $sikt3$ (alltså i storleken 5-12.5 mm). Sikten beskiver viktandelen pellets i olika storlekar. I KK3 finns också en variabel som beskriver volymandelen pellets i olika storlekar. Från denna variabel fås att vi vill ha stor andel pellets i storleken 9-10 mm ($bild4$=variabel nummer fyra i bildanalysen) men även i storleken 10-11.2 mm ($bild5$=variabel nummer fem i bildanalysen). Den kemiska sammansättningen verkar också ha betydelse för LTB-värdet. Vi vill ha liten andel järnoxid samt liten andel magnesiumoxid. I vissa analyser påverkar järnhalten LTB-värdet bara marginellt och i andra analyser har den stor påverkan. Den påverkar dock alltid LTB-värdet positivt i våra analyser, vi vill därför ha en hög järnhalt.

Vi har tidigare sett att det vore bra att veta i vilken LTB-ugn de olika mätningarna är gjorda. Det vore också bra om det framgick vilka observationer som är tagna under extrema förhållanden. Med det menas obsevationer som är tagna då vi har processavvikelser som inte mäts.

Vissa av våra variabler är bara sparade som dygnsmedelvärden medan andra sparas kontinuerligt. LTB-värdet sparas vanligen efter varje skift, detta ger tre gånger per dygn. Vi skulle kunna göra en analys utan att medelvärdesbilda över dygn men då måste vi anta att de variabler som sparas som dygnsmedelvärden är konstanta över dygn samt medelvärdesbilda de kontinuerliga variablerna över skift. Eftersom minst $\sigma^2_{inom}/3$ (då vi inte har tre observationer för alla dygn) procentenheter av variationen är omöjlig att komma åt då vi medelvärdesbildar över dygn skulle det vara intressant att göra liknande analys utan medelvärdesbildningen över dygn. Det är dock omöjligt att säga på förhand hur stor del av inomdygnsvariationen som vi skulle kunna förklara på detta sätt.

Detta kapitel får avslutas med några varnande ord. Om det finns variabler som är starkt korrelerade med responsen men som inte mäts medför det att tillfredställande modeller kan vara omöjliga att hitta. Om observationerna är insamlade under en lång tidsperiod kan också orsaker som slitage och byte av mätinstrument orsaka variation som inte mäts. Vi kan också bara dra slutsatser inom de intervall där vi har observationer. Variabler som är nära konstant ger t.ex. dåliga förutsättningar att upptäcka om de påverkar responsen.

6 Akronymlista

KK2(3)	Kiruna Kulsinterverk 2(3)
KPBA	Kiruna Pellets Blastfurnace Acid
KPBO	Kiruna Pellets Blastfurnace Olivin
LTB	Low Temperature Breakdown
PC	Principal Component
PCA	Principal Components Analysis
PCR	Principal Components Regression
PLSR	Partial Least Squares Regression
$PRESS$	Predicted REsidual Sum of Squares
VIP	Variable Importance for the Projection

Referenser

[1] J. Aitchison. *The statistical analysis of compositional data.* Chapman and Hall, 1986.

[2] P. Bastien, V. E. Vinzi, M. Tenenhaus. *PLS generalised linear regression.* Computational statistiscs & data analysis 48 17-46, 2005.

[3] P. J. Brown. *Measurement, regression, and calibration.* Clarendon press, 1993.

[4] D. Causeur, G. Daumas, T. Dhorne, B. Engel, M. Font I Furnols, S. Hojsgaard. *Statistical handbook for assessing pig classification methods: recommendations from the "EUPIGCLASSproject group.*

[5] L. Eriksson, E. Johansson, N. Kettaneh-Wold, S. Wold. *Multi- and megavariate data analysis, principles and applications.* Umetrics academy, 2001.

[6] J. Hinkle, W. Rayens. *Partial least squares and compositional data: problems and alternatives.* Chemometrics and intelligent laboratory systems 30 159-172, 1995.

[7] Influence Diagnostics. *SAS help and documentation.*

[8] R. A. Johnson, D. W. Wichern. *Applied multivariate statistical analysis.* 5th ed. Prentice Hall, 2002.

[9] E. C. Malthouse. *Nonlinear partial least squares.* PhD thesis, Northwestern university, 1995.

[10] T. J. Robinson. *Dual model robust regression.* PhD thesis, Virginia polytechnic institute and state university, 1997.

[11] SAS institute inc. *Examples using the PLS procedure.*

[12] R. Sundberg. *Kompendium i tillämpad matematisk statistik.* 1997.

A Variabelförklaring

	Grate		
Ovr1	Produktion		
Ovr2	Avsiktat		**Kiln**
Ovr3	Bäddhöjd	Ovr27	Tegel temp
Ovr4	Tryck ob UDD	Ovr28	Material temp
Ovr5	Tryck ob DDD	Ovr29	Oljeflöde
Ovr6	Temp ob DDD	Ovr30	Kolflöde
Ovr7	Tryck ob TPH		**Kylare**
Ovr8	Temp ob TPH	Ovr31	Temp PH
Ovr9	Tryck ob PH	Ovr32	Tryck PH
Ovr10	Temp ob PH	Ovr33	Temp TPH
Ovr11	Tryck ub UDD	Ovr34	Tryck TPH
Ovr12	Temp ub UDD	Ovr35	Temp DDD
Ovr13	Tryck ub DDD	Ovr36	Tryck DDD
Ovr14	Temp ub DDD	Ovr37	Temp UDD
Ovr15	Tryck ub TPH	Ovr38	Temp radar kylvent
Ovr16	Temp ub TPH	Ovr39	Tryck radar kylvent zon1-2
Ovr17	Tryck ub PH	Ovr40	Tryck radar kylvent zon3-4
Ovr18	Temp1 ub PH	Ovr41	Effekt fläkt 050
Ovr19	Temp2 ub PH	Ovr42	Varv fläkt 050
Ovr20	Temp före filter	Ovr43	Temp ute
Ovr21	Flöde FL009	Ovr44	Effekt fläkt 053
Ovr22	Flöde FL013	Ovr45	Varv fläkt 053
Ovr23	Flöde FL019c	Ovr46	KPC1 yta
Ovr24	Temp FL019	Ovr47	TAKA1 skit<45
Ovr25	Flöde FL023		
Ovr26	Effekt		

B Tabeller

Sikt1	Sikt2	Sikt3	Sikt4	Sikt5
0,8	3,8	83,2	12,2	0
0,4	2,2	76,5	20,3	0,6
0,2	3,6	80,7	15,4	0
0,8	3,8	78,1	17,2	0,2
0,4	4	84,5	11	0
0,6	4,4	83,6	11,5	0
0,4	3,4	84,3	11,8	0
0,4	2,6	81,3	15,8	0
0,2	0,6	20,9	70,9	7,4
0,4	2	72,5	25,1	0
0,4	1	75	23,6	0

Tabell 1: Typiska värden för siktvariablerna.

The PLS Procedure

Split-sample Validation for the Number of Extracted Factors

Number of Extracted Factors	Root Mean PRESS	Prob > PRESS
0	1.063119	<.0001
1	0.785456	0.0060
2	0.666867	0.0050
3	0.633894	0.0640
4	0.554214	0.1920
5	0.535008	1.0000
6	0.560893	0.0870
7	0.598973	0.0070
8	0.674316	0.0030
9	0.718612	0.0010
10	0.796834	0.0010
11	0.856525	<.0001
12	0.928324	<.0001
13	0.969489	<.0001
14	0.956905	<.0001
15	0.963843	<.0001

Minimum root mean PRESS	0.5350
Minimizing number of factors	5
Smallest number of factors with p > 0.1	4

Tabell 2: $PRESS$ för olika stora modeller.

The PLS Procedure

Percent Variation Accounted for by Partial Least Squares Factors

Number of Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total
1	22.9343	22.9343	56.7160	56.7160
2	16.9727	39.9070	20.1226	76.8386
3	20.2597	60.1667	7.2054	84.0441
4	7.1879	67.3547	4.5862	88.6302

Tabell 3: Resultatet av en PLSR med fyra faktorer.

X_VAR	B1	VIP			
P1	-0.08174	0.97930	OVR20	0.03143	0.76477
P2	0.09566	0.82987	OVR21	-0.10050	1.18077
KEMFE	0.17749	1.54327	OVR22	-0.00755	0.54584
KEMP	-0.07476	0.65700	OVR23	-0.00275	0.95692
KEMFEO	-0.10834	1.08900	OVR24	-0.01894	0.79385
KEMSIO2	-0.09686	0.94866	OVR25	0.02640	1.15672
KEMCAO	-0.02805	0.89117	OVR26	0.01141	0.80035
KEMMGO	-0.18658	1.55499	OVR27	-0.04296	0.70909
OVR1	-0.02379	0.70161	OVR28	0.00149	0.57166
OVR2	0.02533	0.56976	OVR29	-0.04842	0.55375
OVR3	-0.06332	0.74182	OVR30	-0.02062	0.72241
OVR4	-0.06307	0.68773	OVR31	-0.01120	0.92665
OVR5	-0.02016	1.01707	OVR32	0.09524	0.73469
OVR6	0.04959	0.61340	OVR33	-0.04222	1.08394
OVR7	0.05240	1.02813	OVR34	0.04748	0.92078
OVR8	-0.03411	0.80796	OVR35	0.08298	0.70322
OVR9	-0.01872	1.21389	OVR36	-0.05823	1.20901
OVR10	-0.04451	0.77897	OVR37	-0.03115	1.37469
OVR11	-0.03196	0.82086	OVR38	0.15776	1.85865
OVR12	0.02179	0.67812	OVR39	0.01191	1.13113
OVR13	0.04425	0.96104	OVR40	0.07473	1.31649
OVR14	0.02807	0.73893	OVR41	-0.00124	1.06152
OVR15	-0.04047	1.15026	OVR42	-0.04886	0.98053
OVR16	0.03446	0.79005	OVR43	-0.10089	1.46353
OVR17	0.00402	0.98974	OVR44	0.11278	1.57471
OVR18	-0.02002	0.68810	OVR45	0.057258	1.19626
OVR19	-0.06540	0.86490	OVR46	-0.051332	0.70842
			OVR47	-0.089707	1.26942

Tabell 4: Regressionskoefficienter och VIP-värden, där VIP beräknas enligt ekvation (31).

The PLS Procedure

Percent Variation Accounted for by Partial Least Squares Factors

Number of Extracted Factors	Model Effects Current	Model Effects Total	Dependent Variables Current	Dependent Variables Total
1	29.5104	29.5104	58.1549	58.1549
2	16.5080	46.0184	20.5849	78.7399
3	13.3466	59.3651	6.9663	85.7061
4	8.4966	67.8617	2.4194	88.1256

Tabell 5: Resultatet av en PLSR efter att vissa variabler eliminerats.

	KemFE	KemP	KemFeO	KemSiO2	KemCaO	KemMgO
medel	66,68389	0,024422222	0,372778	2,037333333	0,468888889	1,506888889
medel juni 04	66,693	0,0249	0,561	2,043	0,49	1,524
medel juni 04/medel	1,000137	1,019563239	1,504918	1,002781414	1,045023697	1,011355257
	Ovr1	**Ovr2**	**Ovr3**	**Ovr4**	**Ovr5**	**Ovr6**
	487,7811	16,94907778	19,80194	-12,00531111	-22,32063333	428,0420889
	474,17	16,1807	18,3537	-10,9601	-19,3177	422,2468
	0,972096	0,954665511	0,926864	0,912937607	0,865463794	0,986460937
	Ovr7	**Ovr8**	**Ovr9**	**Ovr10**	**Ovr11**	**Ovr12**
	-18,8252	930,891	-10,2988	1124,714667	239,9067333	125,3409111
	-24,6226	949,549	-5,9565	1150,118	237,7152	133,5767
	1,30796	1,020043163	0,578367	1,022586469	0,990865061	1,065707109
	Ovr13	**Ovr14**	**Ovr15**	**Ovr16**	**Ovr17**	**Ovr18**
	-198,235	83,19481111	-235,612	116,6994333	-173,0482	290,1600222
	-171,152	87,5648	-185,032	135,9541	-130,433	322,8223
	0,863378	1,052527181	0,785323	1,16499366	0,753737976	1,11256643
	Ovr19	**Ovr20**	**Ovr21**	**Ovr22**	**Ovr23**	**Ovr24**
	423,2073	261,5183222	377,6519	195,3002111	438,6581222	133,2860111
	459,6502	274,6845	453,2529	189,6096	388,5288	165,5693
	1,086111	1,050345145	1,200187	0,970862238	0,885721204	1,242210631
	Ovr25	**Ovr26**	**Ovr27**	**Ovr28**	**Ovr29**	**Ovr31**
	292,4656	13,81637778	1350,692	1234,344444	18,47091111	1121,39
	214,5	12,3573	1377,86	1236,06	29,497	1170,177
	0,733419	0,894395058	1,020114	1,001389852	1,596943422	1,043505828
	Ovr32	**Ovr33**	**Ovr34**	**Ovr35**	**Ovr36**	**Ovr37**
	-2,64149	928,8753333	-9,57188	614,4350778	-14,0944	154,7670111
	-2,4154	1042,112	-13,3632	568,0471	-9,2353	221,6089
	0,914409	1,121907281	1,39609	0,924503044	0,655246055	1,431887186
	Ovr38	**Ovr39**	**Ovr40**	**Ovr41**	**Ovr42**	**Ovr43**
	1079,485	301,7257222	287,0287	461,3055778	655,6933333	2,870233333
	932,703	233,0723	207,5865	329,031	611,34	8,8658
	0,864026	0,772464138	0,723226	0,713260398	0,932356589	3,088877792
	Ovr44	**Ovr45**	**Ovr46**	**Ovr47**	**p1**	**p2**
	344,3289	619,8855556	9744,692	64,22655556	1,341793111	3,156031667
	170,2809	511,72	9949,332	76,161	3,047351	2,970683
	0,49453	0,825507217	1,021	1,185817912	2,271103477	0,941271607

Tabell 6: Medelvärden av de förklarande variablerna under hela tidsperioden samt under juni 04.

e1			
0.053405	**KemFe**	0.01727	**Ovr22**
-0.10168	**KemP**	0.12924	**Ovr23**
-0.2831	**KemFeO**	-0.17367	**Ovr24**
0.045618	**KemSiO2**	0.22825	**Ovr25**
-0.089521	**KemCaO**	0.10089	**Ovr26**
-0.14378	**KemMgO**	-0.067577	**Ovr27**
0.027157	**Ovr1**	0.0006886	**Ovr28**
0.1479	**Ovr2**	-0.046899	**Ovr29**
0.14342	**Ovr3**	-0.078864	**Ovr31**
-0.12755	**Ovr4**	-0.061531	**Ovr32**
-0.14105	**Ovr5**	-0.14487	**Ovr33**
0.022989	**Ovr6**	0.12321	**Ovr34**
0.1511	**Ovr7**	0.10129	**Ovr35**
-0.023038	**Ovr8**	-0.1754	**Ovr36**
-0.18405	**Ovr9**	-0.23441	**Ovr37**
-0.067973	**Ovr10**	0.21108	**Ovr38**
-0.0041124	**Ovr11**	0.15008	**Ovr39**
-0.086406	**Ovr12**	0.17699	**Ovr40**
-0.10301	**Ovr13**	0.12218	**Ovr41**
-0.078595	**Ovr14**	0.061199	**Ovr42**
-0.1593	**Ovr15**	-0.16703	**Ovr43**
-0.15485	**Ovr16**	0.245	**Ovr44**
-0.14911	**Ovr17**	0.16763	**Ovr45**
-0.12782	**Ovr18**	-0.14596	**Ovr46**
-0.12181	**Ovr19**	-0.14959	**Ovr47**
-0.1059	**Ovr20**	-0.22995	**p1**
-0.075084	**Ovr21**	0.047541	**p2**

Tabell 7: Första egenvektorn till T (låt den betecknas e_1), där T är enligt ekvation (32).

LTB stdize	c1	LTBHat stdize	LTBRes stdize		LTB	LTBHat
-1,93874	-2,109	-0,9891	-0,94964		75,9	81,64454
-2,01594	-5,671	-2,66002	0,64408		75,433	71,53681
-2,93061	-4,427	-2,07632	-0,85429		69,9	75,06773
-2,16191	-4,718	-2,21277	0,05086		74,55	74,24231
-3,06831	-4,883	-2,2905	-0,77781		69,067	73,77211
-3,75716	-5,335	-2,50242	-1,25474		64,9	72,49016
-2,14819	-5,698	-2,67287	0,52468		74,63	71,45907
-1,845	-5,663	-2,65645	0,81145		76,467	71,5584
-1,6301	-5,378	-2,5224	0,8923		77,767	72,3693

Tabell 8: Uppmätt och predikterat LTB, där stdize betyder att värdet är beräknat på de standardiserade variablerna ($c1$ är regressionsparametern som fås som första kolumnen i C, där C fås från ekvation (33)).

Student Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS	-DFBETAS- c1
-1.111	-1.1306	0.0198	0.9859	-0.1608	0.1608
0.806	0.7870	0.1434	1.2258	0.3220	-0.3220
-1.036	-1.0417	0.0874	1.0842	-0.3223	0.3223
0.0621	0.0581	0.0993	1.2682	0.0193	-0.0193
-0.953	-0.9473	0.1063	1.1336	-0.3268	0.3268
-1.556	-1.7430	0.1269	0.9128	-0.6646	0.6646
0.657	0.6323	0.1448	1.2642	0.2602	-0.2602
1.016	1.0180	0.1430	1.1616	0.4159	-0.4159
1.108	1.1263	0.1290	1.1108	0.4334	-0.4334

Tabell 9: Inflytelse diagnostik för regressionsmodellen med $c1$ som förklarande variabel (c1 är regressionsvariabeln som fås som första kolumnen i C, där C fås från ekvation (33)).

Dependent Variable	Predicted Value	95% CL Predict	
-1.9387	-0.9891	-2.9989	1.0207
-2.0159	-2.6600	-4.7881	-0.5320
-2.9306	-2.0763	-4.1516	-0.001073
-2.1619	-2.2128	-4.2993	-0.1262
-3.0683	-2.2905	-4.3838	-0.1972
-3.7572	-2.5024	-4.6151	-0.3898
-2.1482	-2.6729	-4.8022	-0.5435
-1.8450	-2.6565	-4.7842	-0.5288
-1.6301	-2.5224	-4.6370	-0.4078

Tabell 10: Prediktionsintervall för regressionsmodellen med $c1$ som förklarande variabel (c1 är regressionsvariabeln som fås som första kolumnen i C, där C fås från ekvation (33)).

The PLS Procedure

Percent Variation Accounted for
by Partial Least Squares Factors

Number of Extracted Factors	Model Effects Current	Model Effects Total	Dependent Variables Current	Dependent Variables Total
1	62.7265	62.7265	92.6553	92.6553

Tabell 11: Resultatet av en PLSR på de nio observationer med lägst LTB, se avsnitt 4.8.

KemFE	KemP	KemFeO	KemSiO2	KemCaO	KemMgO	Ovr1	Ovr2
0,1377	-0,10947	-0,26045	-0,0259	-0,12372	-0,19283	0,01803	0,16198
Ovr3	Ovr4	Ovr5	Ovr6	Ovr7	Ovr8	Ovr9	Ovr10
0,14428	-0,13473	-0,13125	0,03751	0,17255	-0,01271	-0,18545	-0,07515
Ovr11	Ovr12	Ovr13	Ovr14	Ovr15	Ovr16	Ovr17	Ovr18
-0,00736	-0,07735	-0,09585	-0,07192	-0,15314	-0,16008	-0,14218	-0,1331
Ovr19	Ovr20	Ovr21	Ovr22	Ovr23	Ovr24	Ovr25	Ovr26
-0,12648	-0,10405	-0,08082	0,0077	0,11825	-0,17926	0,212	0,07845
Ovr27	Ovr28	Ovr29	Ovr31	Ovr32	Ovr33	Ovr34	Ovr35
-0,07038	-0,00214	-0,02771	-0,08264	-0,07126	-0,14584	0,13077	0,12535
Ovr36	Ovr37	Ovr38	Ovr39	Ovr40	Ovr41	Ovr42	Ovr43
-0,17837	-0,20708	0,1997	0,14742	0,17364	0,11936	0,05348	-0,15746
Ovr44	Ovr45	Ovr46	Ovr47	p1	p2		
0,24747	0,163	-0,13824	-0,15317	-0,23565	0,05184		

Tabell 12: De förklarande variablernas vikter i en PLSR med en faktor på de nio observationer med lägst LTB, se avsnitt 4.8.

p1	-.1133603607	Ovr23	0.0507673150
p2	0.0404699557	Ovr24	-.0820705235
KemFE	0.1827986536	Ovr25	0.0777188951
KemP	-.0960546785	Ovr26	0.0172852900
KemFeO	-.1117968908	Ovr27	-.0247201868
KemSiO2	-.1216811927	Ovr28	-.0005332202
KemCaO	-.0776145583	Ovr29	0.0454582216
KemMgO	-.1573188041	Ovr31	-.0393295070
Ovr1	0.0017565998	Ovr32	-.0494219448
Ovr2	0.1020595472	Ovr33	-.0646364222
Ovr3	0.0539325918	Ovr34	0.0661782272
Ovr4	-.0911016405	Ovr35	0.0837884304
Ovr5	-.0565235377	Ovr36	-.0866878621
Ovr6	0.0291794926	Ovr37	-.0703292268
Ovr7	0.0982473712	Ovr38	0.0813618129
Ovr8	0.0051391962	Ovr39	0.0690952297
Ovr9	-.0840510279	Ovr40	0.0785087044
Ovr10	-.0474579723	Ovr41	0.0544905753
Ovr11	-.0093387541	Ovr42	0.0215814910
Ovr12	-.0308546590	Ovr43	-.0774338566
Ovr13	-.0373308121	Ovr44	0.1155529707
Ovr14	-.0259362179	Ovr45	0.0723284525
Ovr15	-.0683971129	Ovr46	-.0754692702
Ovr16	-.0711192888	Ovr47	-.0703397154
Ovr17	-.0531490839		
Ovr18	-.0618510024		
Ovr19	-.0592884527		
Ovr20	-.0440027144		
Ovr21	-.0423157625		
Ovr22	-.0046312954		

Tabell 13: Regressionskoefficienterna i en PCR med två faktorer.

p1	-.1235180686	Ovr22	-.0127701077
p2	0.0295848946	Ovr23	0.0368945354
KemFE	0.2034646982	Ovr24	-.0967899330
KemP	-.0582090807	Ovr25	0.0750726510
KemFeO	-.0827101128	Ovr26	-.0006669647
KemSiO2	-.1268755269	Ovr27	-.0407318553
KemCaO	-.1213701023	Ovr28	-.0066232150
KemMgO	-.1729078263	Ovr29	0.0121838556
Ovr1	-.0072425467	Ovr31	-.0467372937
Ovr2	0.0993021416	Ovr32	-.0498390650
Ovr3	0.0732317855	Ovr33	-.0725616435
Ovr4	-.0739796812	Ovr34	0.0760312491
Ovr5	-.0457155016	Ovr35	0.1012727663
Ovr6	0.0433824424	Ovr36	-.0906626926
Ovr7	0.1199890253	Ovr37	-.0522799146
Ovr8	0.0114307352	Ovr38	0.0766593278
Ovr9	-.0923537852	Ovr39	0.0658822248
Ovr10	-.0473745832	Ovr40	0.0777765643
Ovr11	-.0088106416	Ovr41	0.0523550192
Ovr12	-.0204465602	Ovr42	0.0117587841
Ovr13	-.0341073138	Ovr43	-.0563174273
Ovr14	-.0228711698	Ovr44	0.1237955103
Ovr15	-.0626511600	Ovr45	0.0704487592
Ovr16	-.0871931014	Ovr46	-.0479394420
Ovr17	-.0576833012	Ovr47	-.0806388889
Ovr18	-.0739528328		
Ovr19	-.0696802884		
Ovr20	-.0470291898		
Ovr21	-.0489020280		

Tabell 14: Regressionskoefficienterna i en PLSR med två faktorer.

KK2 normalt>80, PLS 3 faktorer

		LTB	
		normalt	lågt
LTHat	normalt	79	1
	lågt	1	9

Tabell 15: Frekvenstabell för KK2, där normalt>80.

KK3 normalt>80, PLS 1 faktor

		LTB	
		normalt	lågt
LTHat	normalt	5	0
	lågt	6	45

Tabell 16: Frekvenstabell för KK3, där normalt>80.

KK3 normalt>75, PLS 1 faktor

		LTB	
		normalt	lågt
LTHat	normalt	23	6
	lågt	8	19

Tabell 17: Frekvenstabell för KK3, där normalt>75.

KK3 normalt>70, PLS 1 faktor

		LTB	
		normalt	lågt
LTHat	normalt	46	6
	lågt	2	2

Tabell 18: Frekvenstabell för KK3, där normalt>70.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Model	26	62.5633	2.4063	1.37	0.2125
Error	27	47.5450	1.7609		
C Total	53	110.1083			

Tabell 19: Variansanalys på ugn 1.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Model	25	81.0508	3.2420	2.80	0.0057
Error	26	30.1400	1.1592		
C Total	51	111.1908			

Tabell 20: Variansanalys på ugn 0.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Model	25	90.9877	3.6395	2.12	0.0313
Error	26	44.6700	1.7181		
C Total	51	135.6577			

Tabell 21: Variansanalys på ugn 3.

Analysis of Variance

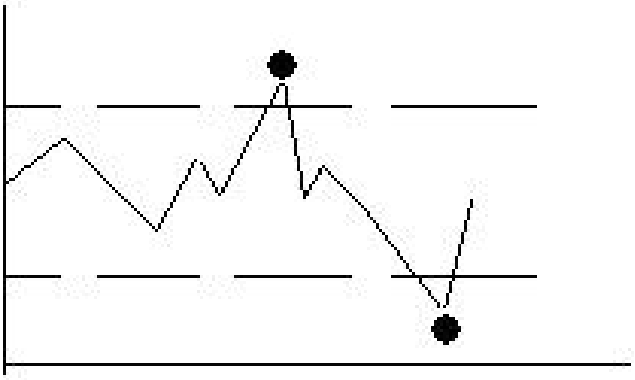
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Model	23	82.5367	3.5886	1.92	0.0591
Error	24	44.7600	1.8650		
C Total	47	127.2967			

Type III Tests

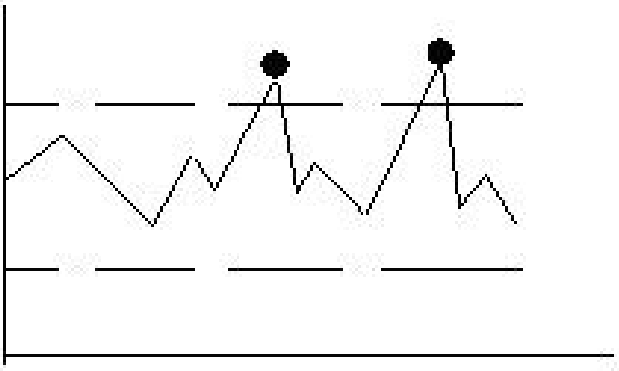
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
obs_tillf	11	30.6467	2.7861	1.49	0.1979
ugn	1	17.0408	17.0408	9.14	0.0059
obs_tillf*ugn	11	34.8492	3.1681	1.70	0.1343

Tabell 22: Variansanalys med ugn (0 eller 1), observationstillfälle och samspelstermen som förklarande variabler.

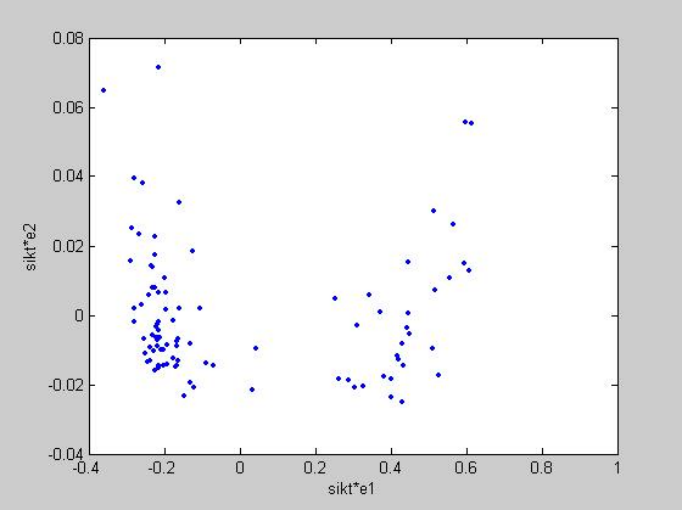
C Figurer



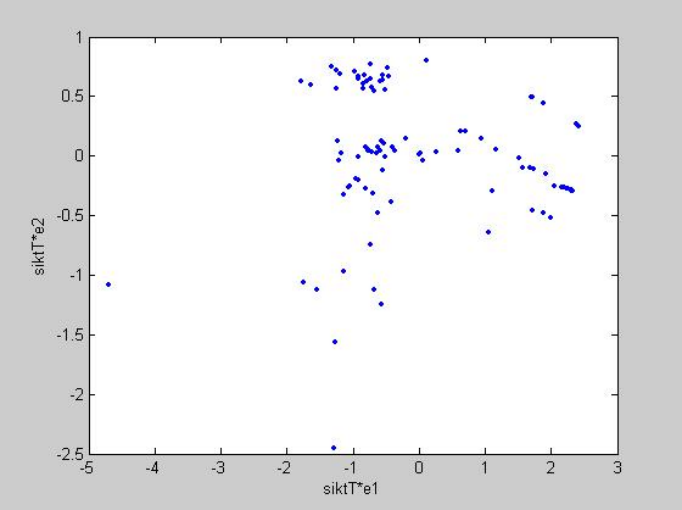
Figur 1: I detta fall kommer bara de två svarta prickarna att sparas.



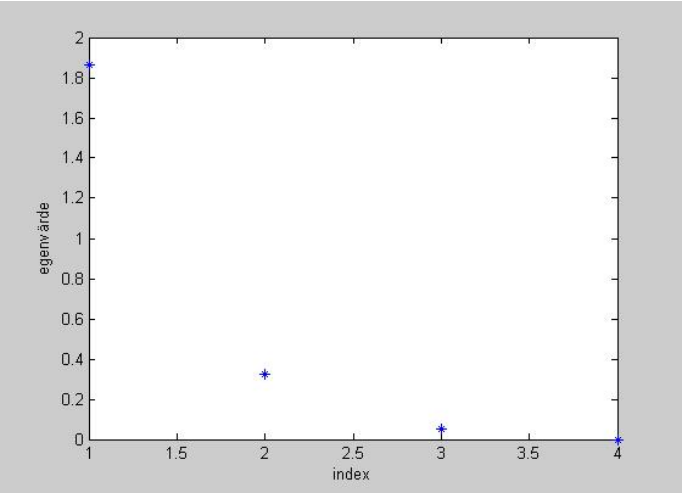
Figur 2: I detta fall kommer det sparade medelvärdet att kraftigt avvika från det sanna.



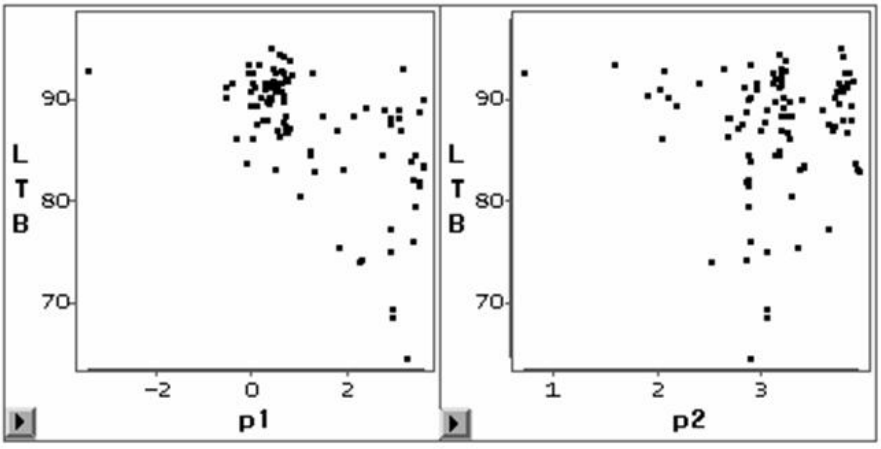
Figur 3: $S * e_2$ plottat mot $S * e_1$.



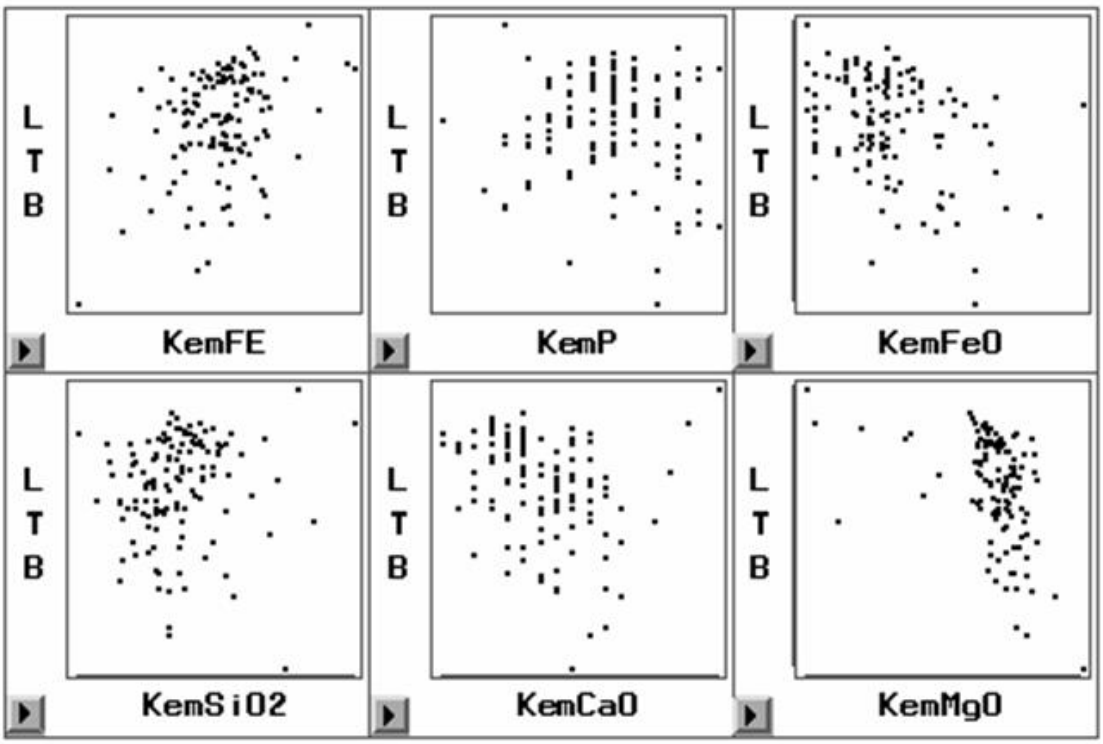
Figur 4: $log(Sny) * e_2$ plottat mot $log(Sny) * e_1$.



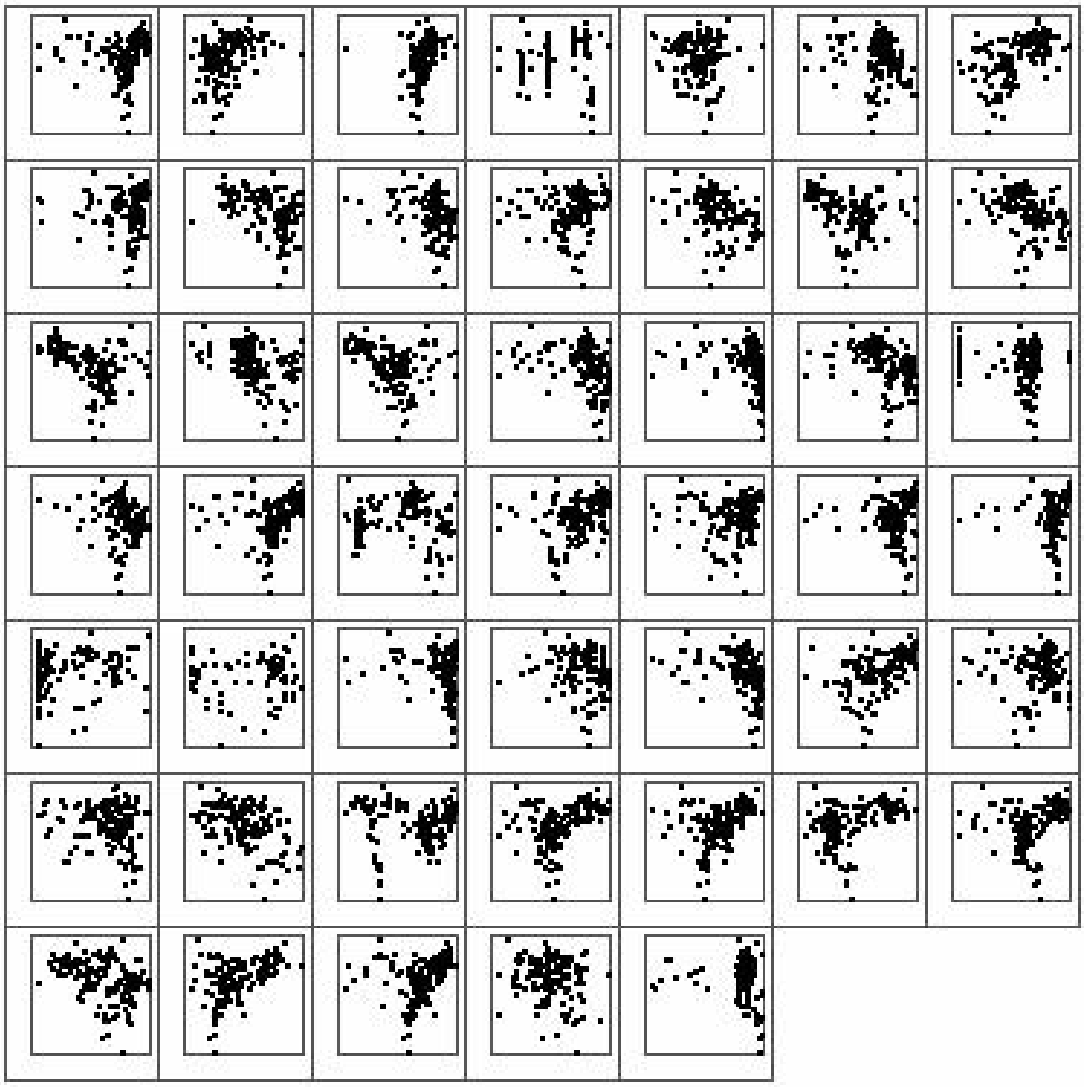
Figur 5: Scree plott av egenvärdena till $Cov(Sny^*)$.



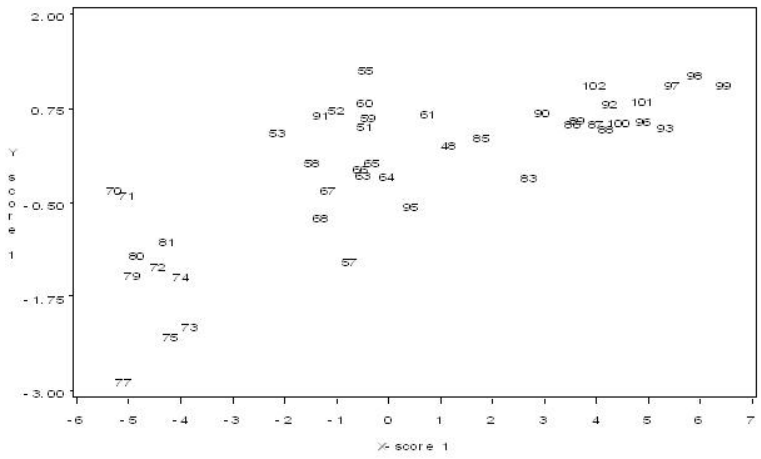
Figur 6: LTB mot transformerade siktvariabler.



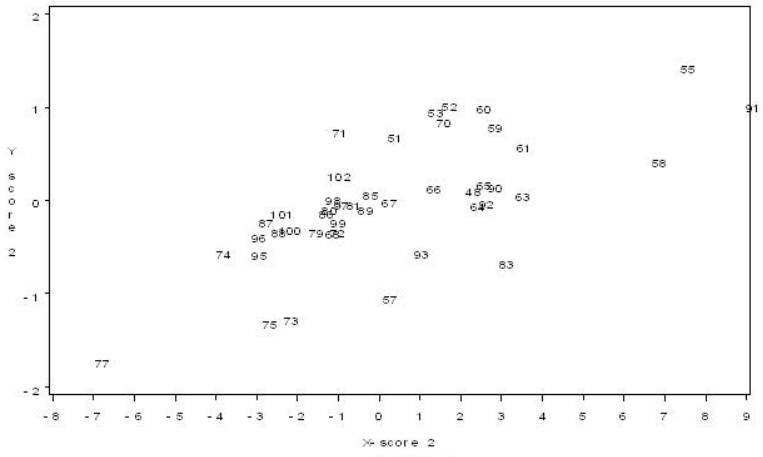
Figur 7: LTB mot kemiska sammansättningen.



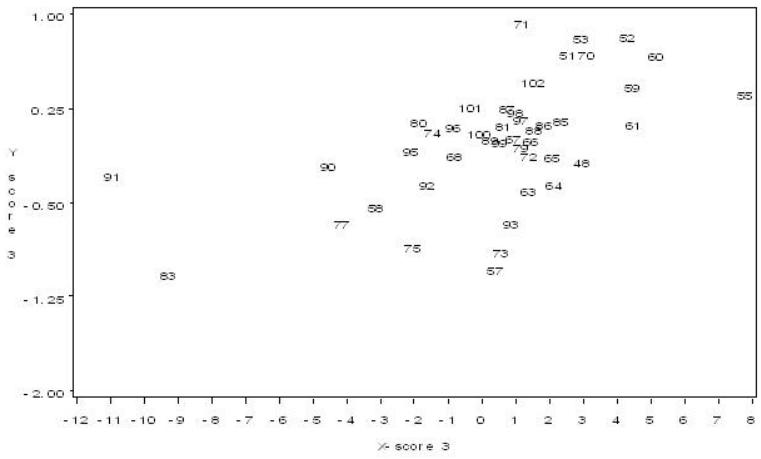
Figur 8: LTB mot övriga processparametrar (från vänster uppe $Ovr1$-$Ovr47$, se appendix A för förklaring).



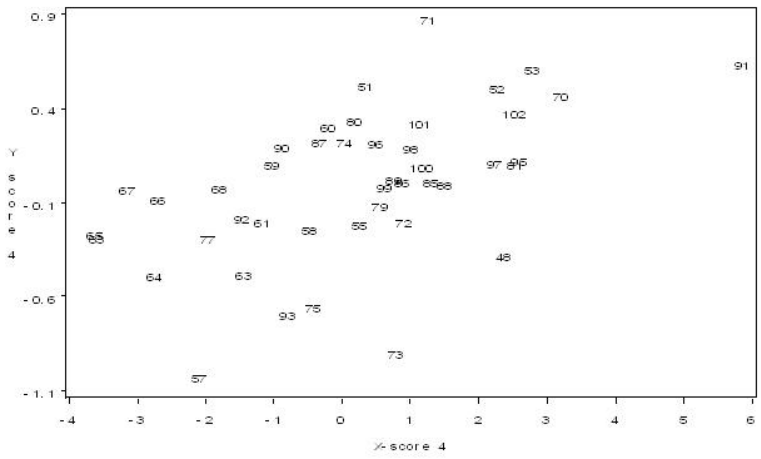
Figur 9: Y-score1 plottat mot X-score1.



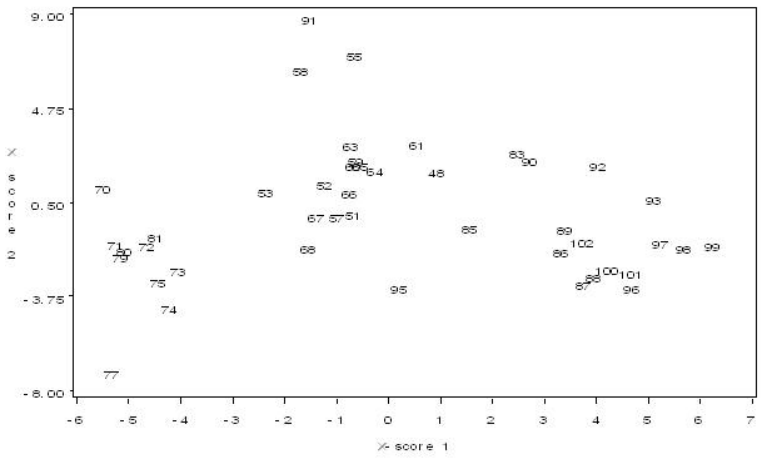
Figur 10: Y-score2 plottat mot X-score2.



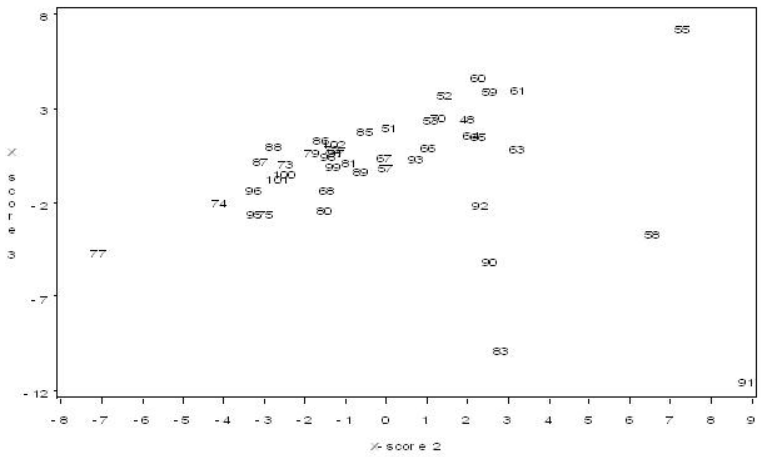
Figur 11: Y-score3 plottat mot X-score3.



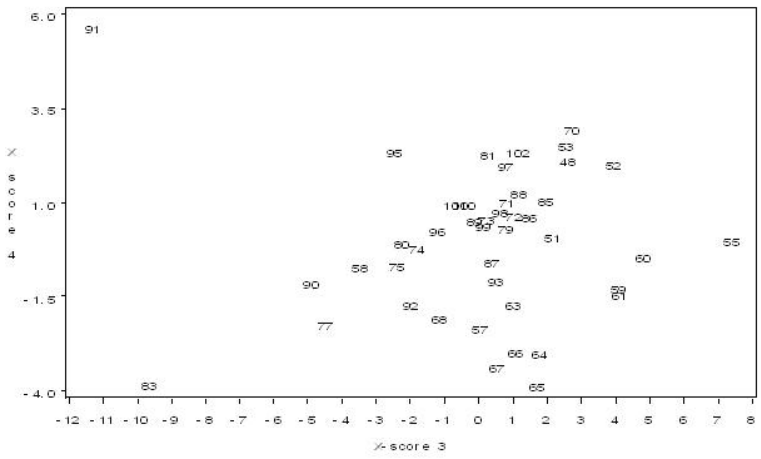
Figur 12: Y-score4 plottat mot X-score4.



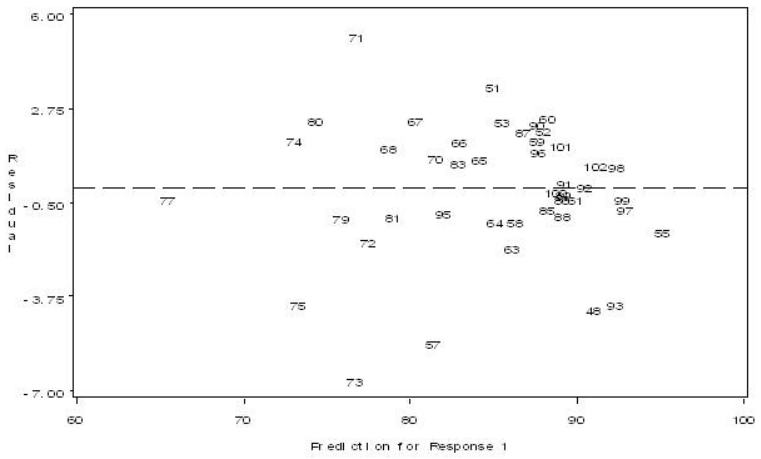
Figur 13: X-score2 plottat mot X-score1.



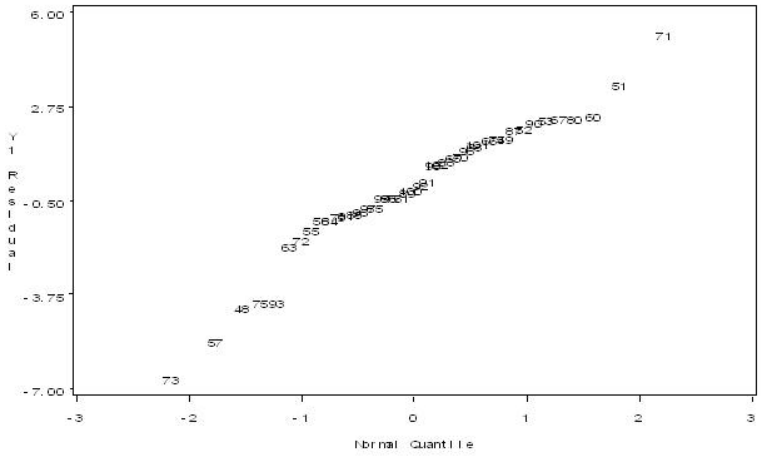
Figur 14: X-score3 plottat mot X-score2.



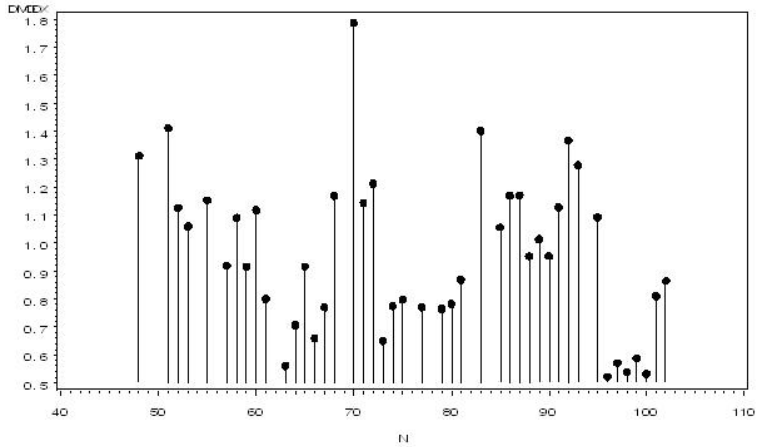
Figur 15: X-score4 plottat mot X-score3.



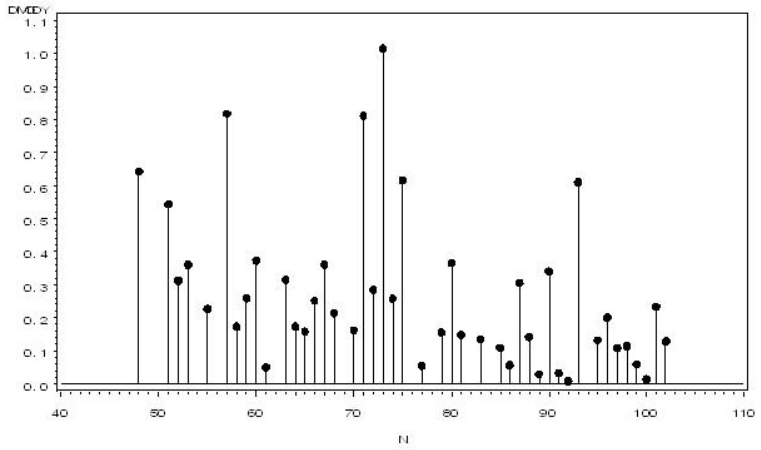
Figur 16: Residualerna plottat mot predikterade LTB-värden.



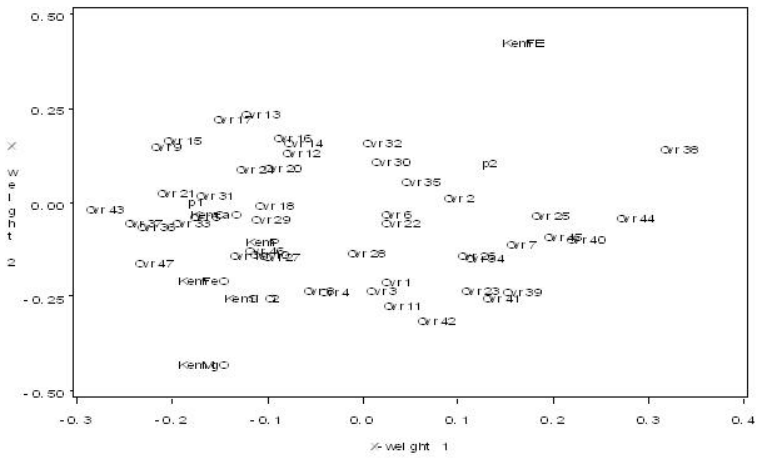
Figur 17: LTB-residualerna plottat mot kvantilerna i en standardiserad normalfördelninen.



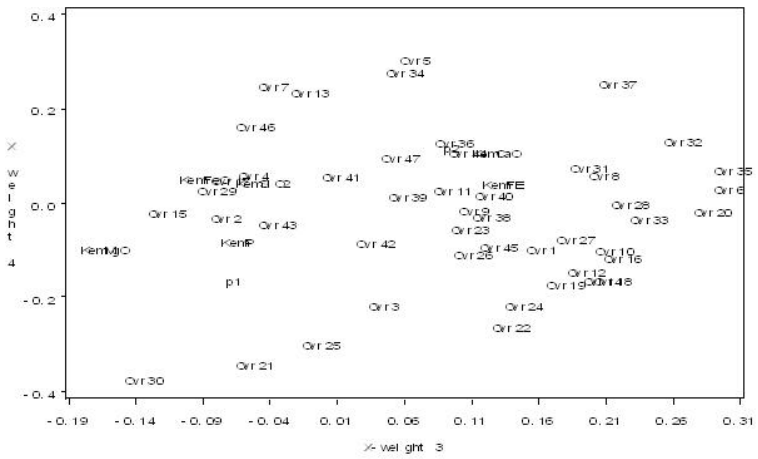
Figur 18: Euklidiska avståndet för de förklarande variablerna till modellen för varje observation.



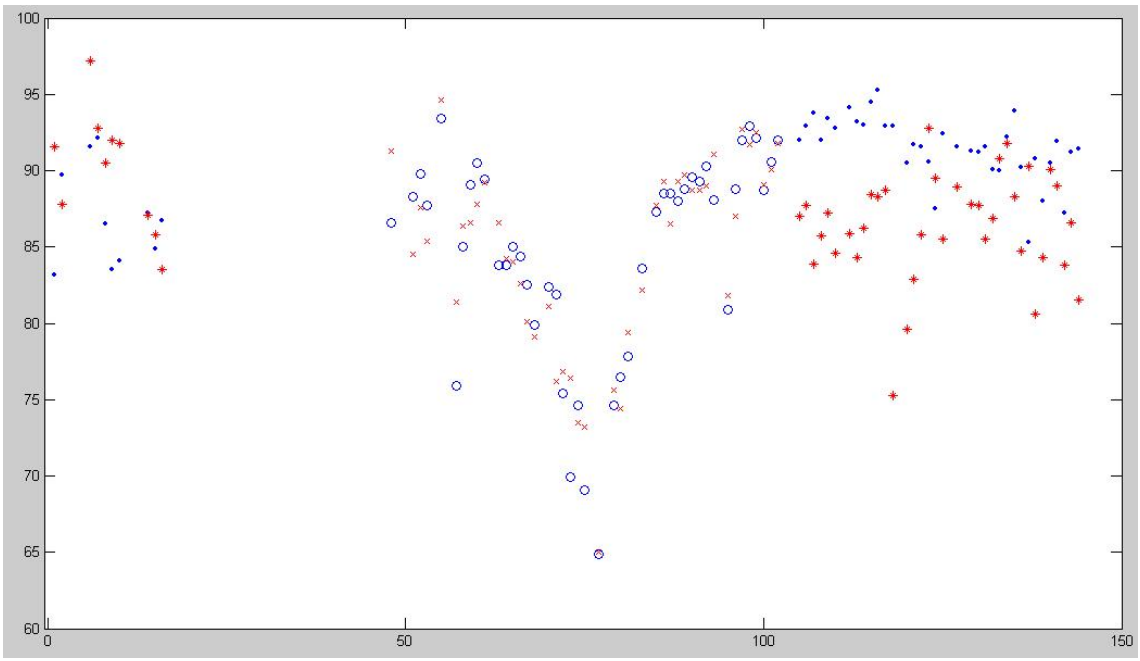
Figur 19: Euklidiska avståndet för LTB till modellen för varje observation.



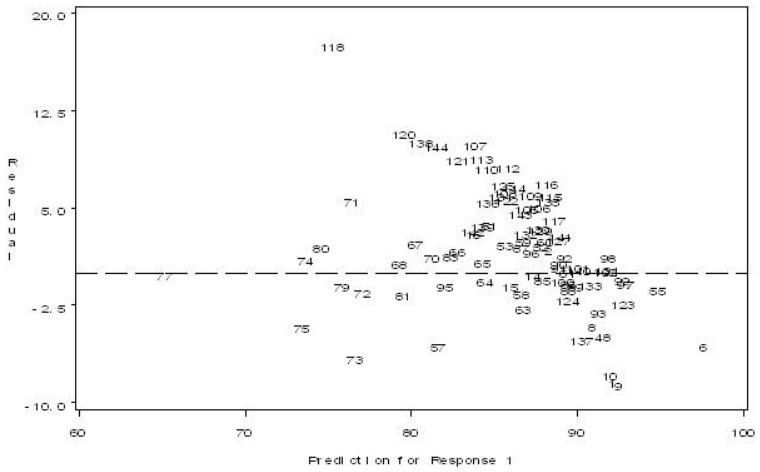
Figur 20: X-weight2 plottat mot X-weight1.



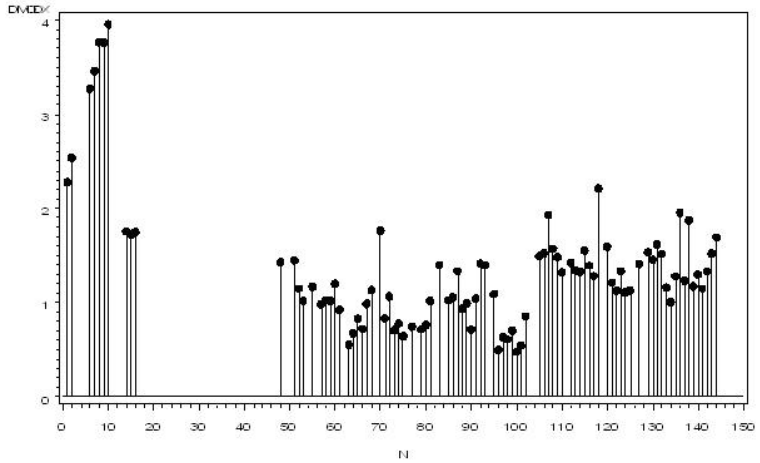
Figur 21: X-weight4 plottat mot X-weight3.



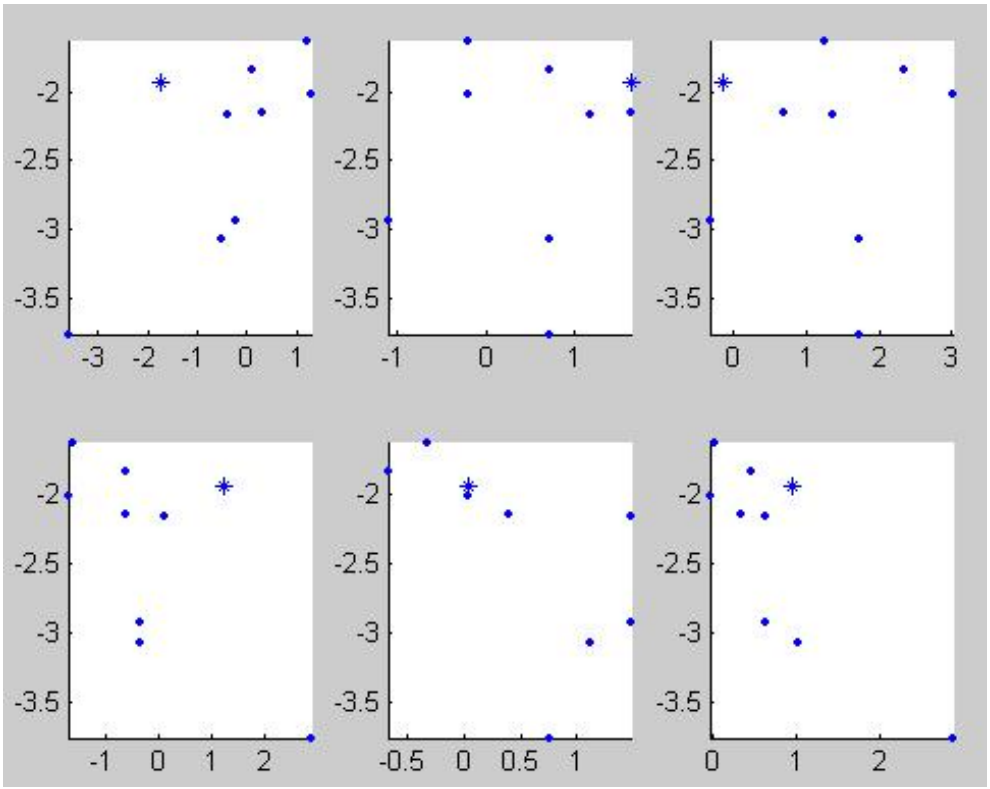
Figur 22: LTB samt predikterat LTB. LTB på observationer som inte har använts för att bygga modellen betecknas med $\cdot$. Predikterat LTB för dessa observationer betcknas med *. LTB på observationer som har använts för att bygga modellen betecknas med ^{o}. Predikterat LTB för dessa observationer betcknas med ^{x}.



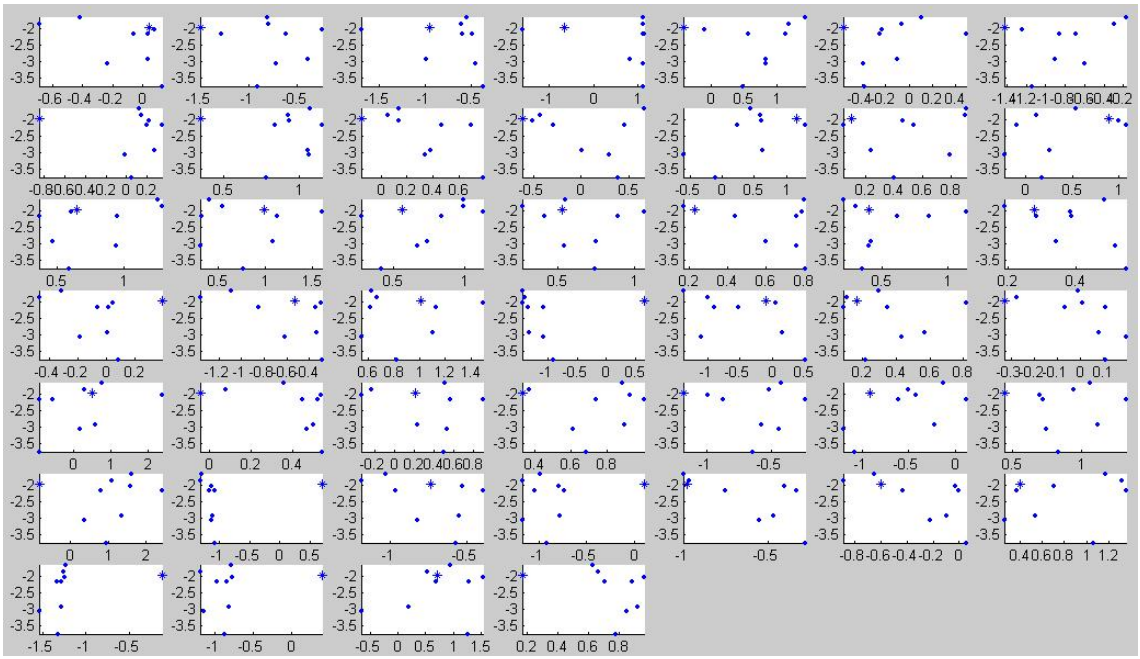
Figur 23: Residualerna och prediktionsfelen för LTB plottat mot predikterat LTB för den reducerade modellen.



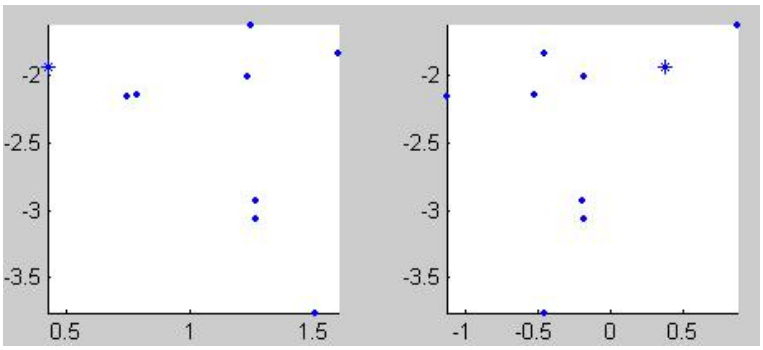
Figur 24: Euklidiska avståndet från observationerna på de förklarande variablerna till den reducerade modellen (gruppen från 48 till 102 har använts för modellanpassningen).



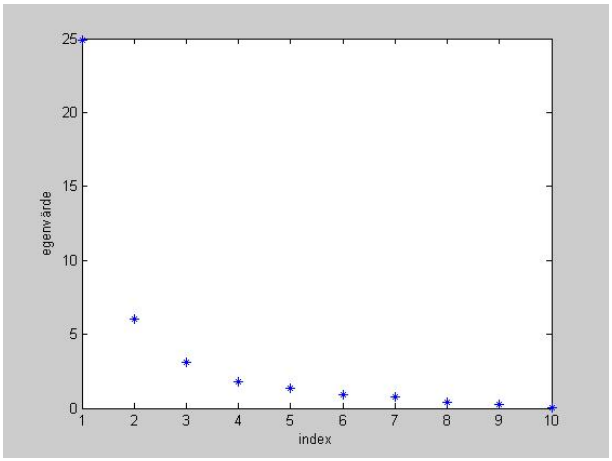
Figur 25: Standardiserade LTB mot standardiserade kemiska sammansättningen, där de översta figurerna visar från vänster $KemFe$, $KemP$ och $KemFeO$ och de understa $KemSiO2$, $KemCaO$ och $KemMgO$. * är den observation som inte kommer från samma avgränsade tidsperiod som övriga.



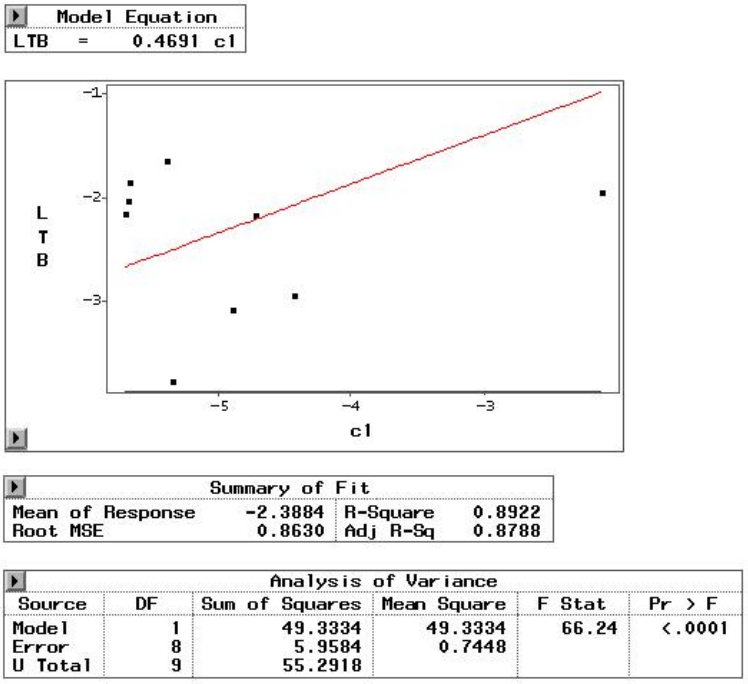
Figur 26: Standardiserade LTB mot standardiserade övriga process variabler (från vänster uppe $Ovr1$-$Ovr29$ och $Ovr31$-$Ovr47$, se appendix A för förklaring). * är den observation som inte kommer från samma avgränsade tidsperiod som övriga.



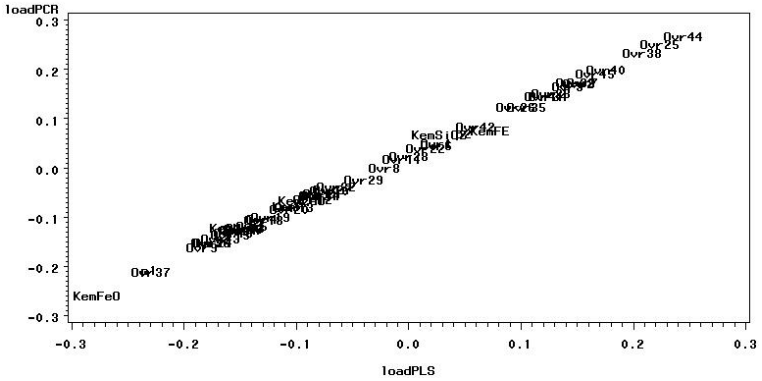
Figur 27: Standardiserade LTB mot standardiserade transformerade siktvariabler, där transformationerna är gjorda enligt ekvation (28) och (29). * är den observation som inte kommer från samma avgränsade tidsperiod som övriga.



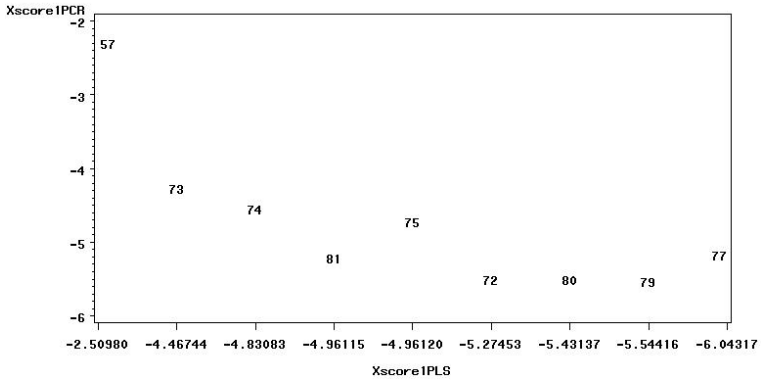
Figur 28: En scree plott av egenvärdena till T, där T är enligt ekvation (32).



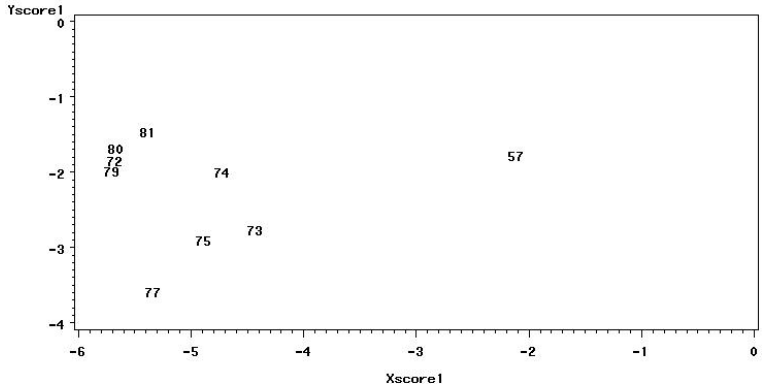
Figur 29: Resultatet av en regressionsanalys med första PC som förklarande variabel, alltså första kolumnen i C (se ekvation (33)).



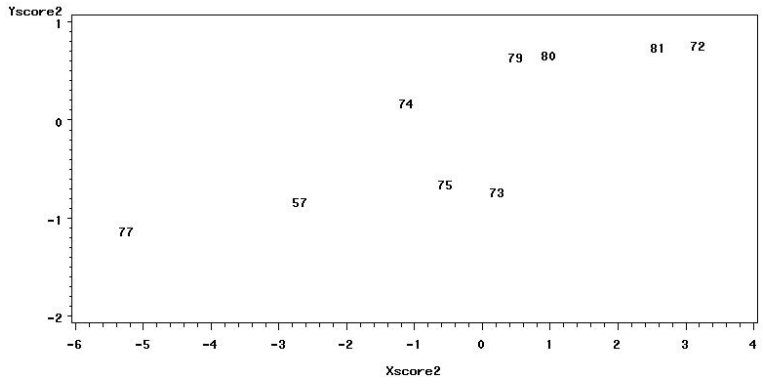
Figur 30: X-laddning för en PCR, med en PC, plottat mot X-laddning för en PLSR, med en faktor.



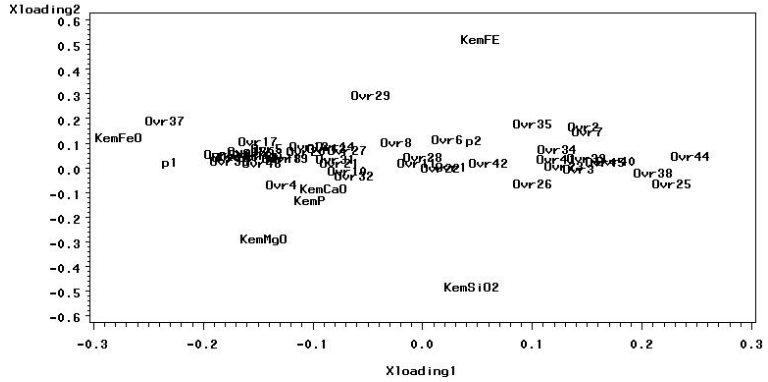
Figur 31: X-score för en PCR, med en PC, plottat mot X-score för en PLSR, med en faktor.



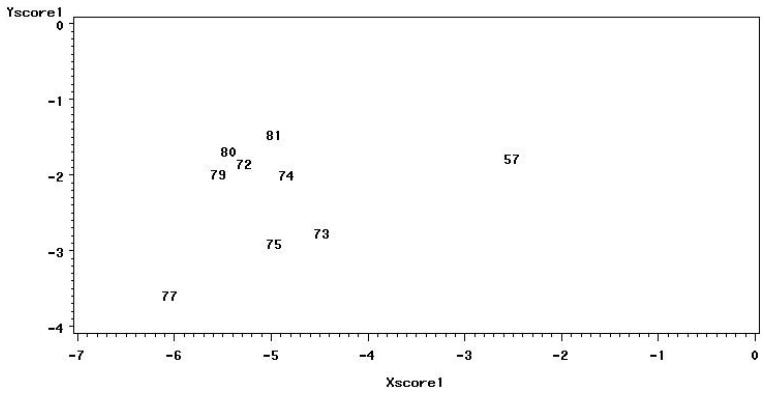
Figur 32: Y-score1 plottat mot X-score1 för en PCR, med två PC.



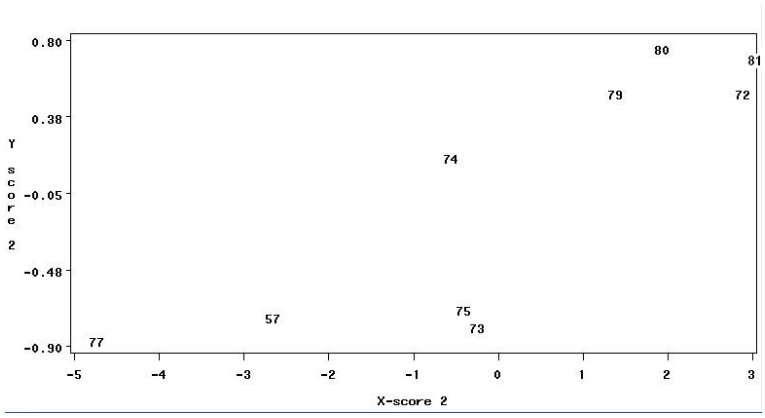
Figur 33: Y-score2 plottat mot X-score2 för en PCR, med två PC.



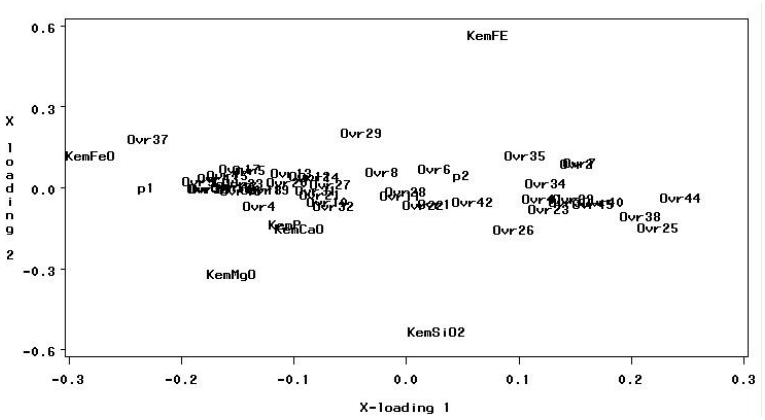
Figur 34: X-laddning2 plottat mot X-laddning1 för en PCR, med två PC.



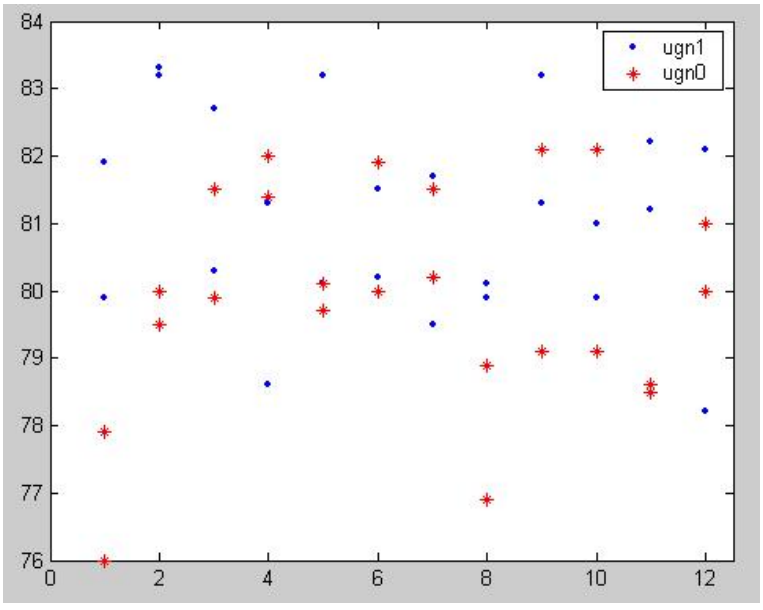
Figur 35: Y-score1 plottat mot X-score1 för en PLSR, med två faktorer.



Figur 36: Y-score2 plottat mot X-score2 för en PLSR, med två faktorer.



Figur 37: X-laddning2 plottat mot X-laddning1 för en PLSR, med två faktorer.



Figur 38: *LTB* mätningar för ugn 0 och 1.