Masteroppgåve i kjemometri

Multivariat modellering og prediksjon av brennverdi og tettleik i naturgass

Liv Kjersti Skartland

Kjemisk institutt

Universitetet i Bergen

Februar 2010

Forord

Denne oppgåva er utført ved kjemisk institutt, Universitetet i Bergen i samarbeid med Christian Michelsen Reaserch (CMR), Instrumetation.

Eg vil takke min veiledar Bjørn Grung for god og konstruktiv veiledning og nyttige tilbakemeldingar.

Eg vil også takke Kjell-Eivind Frøysa ved CMR Instrumentation for tillit til å gjennomføre prosjektet og for rask tilbakemelding både når det gjeld simulering av datamateriale og svar på spørsmål.

Ellers vil eg takke Inga-Cecilie Sørheim for gjennomlesing, råd til oppbygging og god hjelp i skriveprosessen.

Takk til familie, venner og medstudentar for oppmuntring og støtte undervegs.

Innhaldsliste

Samandrag

Denne oppgåva er utført i samarbeid med CMR Instrumentation, Bergen.

Problemstillinga for oppgåva var å bruke multivariate metodar for å lage modellar som er i stand til å predikere tettleik og brennverdi i naturgass. Det er tatt utgangspunkt i at ein ikkje kjenner den kjemiske samansetjinga i naturgassen, såkalla BCA1 [8].

Innsendte variablar i modellane er trykk, temperatur og lydhastigheit samt eit estimat av den kjemiske samansetjinga i naturgassen. Ved å nytte delvis minste kvadraters metode (PLS) har ein i denne oppgåva laga ulike modellar som kan predikere den kjemiske samansetjinga i naturgassen frå dei målte variablane trykk, temperatur og lydhastigheit. I forsøket på å finne dei beste prediksjonane har ein nytta ulike tilnærmingar til prediksjon av den kjemiske samansetjinga. Ein har nytta både prediksjon i Sirius og iterativ konsentrasjonsbestemming ved alternerande regresjon i MATLAB.

Det viser seg at den kjemiske samansetjinga er viktigare i prediksjonen av brennverdi enn det den er for tettleik. Tettleik i naturgassen er svært avhengig av trykk og temperatur og ein ser store prediksjonsfeil i situasjonen låg temperatur og høgt trykk. Ein har også forsøkt å predikere brennverdi med tettleik som ein del av **X** utan at dette gav betre prediksjon av brennverdi.

Resultata viser at prediksjonsfeilen for tettleik blir lågast ved bruk av modellar der ein held konstant trykk og temperatur. For brennverdi får ein lågast prediksjonsfeil ved revers prediksjons av den kjemiske samansetjinga.

1. Innleiing

1.1 Problemstilling

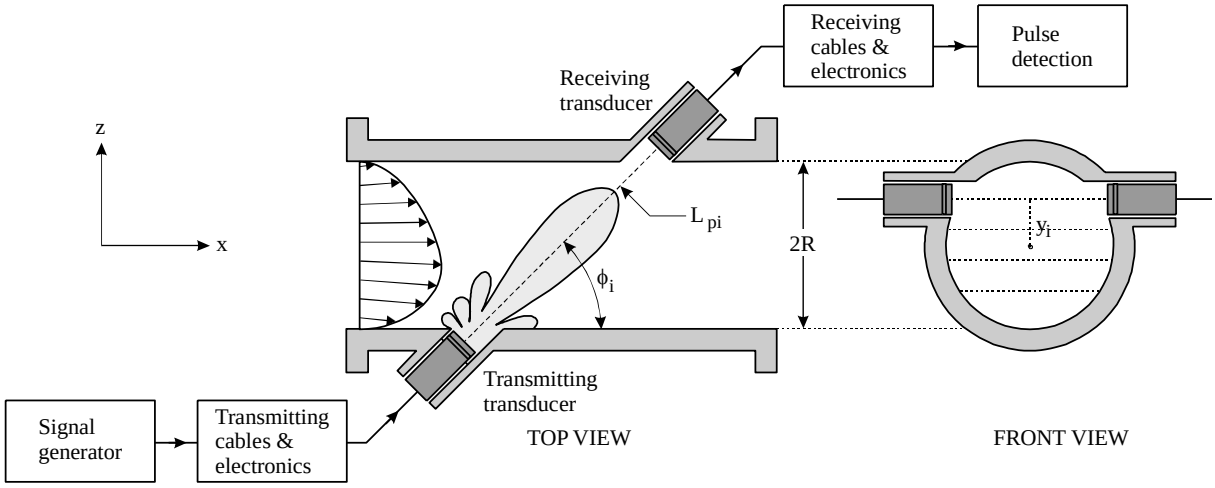
Målet med prosjektet er å nytte multivariate metodar til å utvikle modellar for prediksjon av brennverdi og tettleik i naturgass. Innsendte variablar i modellane er dei målte variablane trykk, temperatur og lydhastigheit samt ei predikert kjemisk samansetjing. Den kjemiske samansetjinga skal predikerast frå trykk, temperatur og lydhastigheit og denne oppgåva vil belyse ulike måtar å komme fram til denne på.

1.2 Bakgrunn for oppgåva

Naturgass er ein brennbar gass som består hovudsakleg av metan, men der også hydrokarbon med opp til seks karbon er tilstades i mindre mengder. I tillegg til dei brennbare hydrokarbona finn ein også komponentar som karbondioksid, nitrogen og helium tilstades i små mengder. Naturgass finn ein vanlegvis i petroleumreservoar, men der finst også reservoar som inneheld berre naturgass [1]. Gass frå ulike felt vil ha varierande konsentrasjonar av dei ulike kjemiske komponentane og energiinnhaldet vil variere betydeleg. Variasjonar opp til 20 % er vanleg. Det er viktig både for leverandør og kundar å kjenne innhaldet i slike gassblandingar slik at berekningar av energiinnhaldet og dermed også forbrenningsmoglegheitene og verdien av gassen vil vere mogleg [2].

Akustiske måleprinsipp vert delt inn i aktive og passive. Med passivt akustisk måleprinsipp meinar ein at prosessen emitterer eit akustisk signal, dette vert kalla akustisk emisjon (AE). Ved aktiv måling vert prosessen påverka av impulsar sendt frå ei akustisk kjelde, dette ser ein døme på i ultralyd i medisinsk diagnostikk og innanfor målingar av gass – og væskestraumar. PLS i samanheng med måling av fleirfase straumar vil kunne identifisere kjemiske samansetjingar [3]. Kjemometri og akustiske data vil i mange tilfelle vere ein svært bra kombinasjon, sidan bruk av kjemometriske teknikkar kan gjere tolking av slike signal enklare [4].

Kontinuerleg overvaking av gassen er nyttig i forhold til prosesskontroll. Jo betre sensorar ein har, jo betre blir kvaliteten på prosessen. Ultralydsensorar bidrar til forbetring av kvaliteten.



Figur 1.1 Prinsipp for ultralydmåling i gass. Figur lånt med løyve frå CMR Instrumentation.

Grunnlaget for ultralydmålingar er ganske enkelt: sensorane sender ut akustiske bølgjer og tek imot dei igjen etter at bølgjene har passert gjennom den prosessen ein ynskjer å undersøkje slik det er vist i figur 1.1. Mottakaren registrerer signal som inneheld viktig informasjon om parametrar ein måler på. Ultralyd dekkjer eit område på 20 kHz til 1 GHz [5]. Når ultralyd passerer gjennom ein gass kan blant anna lydhastigheit observerast. Lydhastigheita er avhengig av frekvens så vel som viskøse, molekylære og termiske samanhengar som oppstår når molekyla kolliderer [2].

Den tradisjonelle måten å bestemme kvaliteten på naturgass er ved bruk av Wobbe indeks. [6]. Wobbe indeks (W) for ein gass vert definert slik:

$$W = \frac{B}{\sqrt{d}} \qquad (1.1)$$

Der B er brennverdien per eining volum (MJ m^{-3}) og d er relativ tettleik for gassen [7]. Wobbe indeks er altså om lag det same som brennverdi, det vil seie brennverdi dividert på relativ tettleik ved standard føresetnadar (15 ˚C og 1 atm).

Multipath ultrasonic transit-time meters (USM) er blitt eit konkurransedyktig alternativ til meir konvensjonelle måleteknologiar for blant anna målingar knytt til skatteberekning av naturgass. Gass vert ikkje selt på bakgrunn av volum, men i måleeiningar som masse (til dømes i Noreg) eller energi (typisk i europeiske land). Når ein nyttar slike einingar treng ein kjenne parameter som tettleik og brennverdi i tillegg til den volumetriske strømningshastigheita. Slike målingar vert tradisjonelt utført med densiometer, gasskromatograf eller kalorimeter. For å redusere kostnadane og også tidsbruken under målingane ynskjer ein å utvikle alternativ til desse målemetodane. USM måler hastigheit, volumetrisk strømningshastigheit og lydhastigheit i gassen.

I løpet av dei siste åra har ein utvikla metodar for å berekne tettleik og brennverdi frå målt lydhastigheit ved bruk av ultralydmålingar på gassen. Dette arbeidet inkluderer både CMR og andre [8].

Lydhastigheita (c) for ei blanding av gassar vert definert slik:

$$c^2 = \frac{\gamma_{mix}\, RT}{Mw_{mix}} \qquad (1.2)$$

Der γ_{mix} er eit uttrykk for ratioen av spesifikk varme for gassblandinga, R er den universale gasskonstanten, T er temperatur og M_{mix} er molekylvekta til blandinga.

Motivasjonen for å nytte lydhastigheit til å bestemme kvaliteten på naturgass var at lydhastigheten kunne nyttast til å bestemme mengda metan i gassen sidan metan har svært låg molekylvekt samanlikna med dei andre komponentane i naturgass. Ratioen av spesifikk varme vil ikkje variere så mykje frå ein gasskomponent til ein annan, dermed vil mengda metan i gassen påverke lydhastigheita i stor grad gjennom molekylvekta, slik ein kan sjå i formel (1.2) [6].

Masse- og energimålingar av naturgass er i dag mogleg ved å bruke USM i kombinasjon med trykk- og temperaturmålingar, ei typisk hydrokarbonsamansetjing og estimat av molare fraksjonar av CO_2 og N_2. Det vil seie at ein kan utføre desse målingane utan å nytte densiometer, gasskromatograf eller kalorimeter som vart nytta i tradisjonelle målestasjonar. Metodar er utvikla frå lydhastigheita som kan nyttast til å rekne ut tettleik og brennverdi. Ved CMR Instrumentation er det utvikla algoritmar som er meir robuste i forhold til dei variasjonane i kjemisk samansetjing ein har i naturgassen. Denne metode for berekning av

brennverdi og tettleik er altså utvikla ut frå målingar av trykk, temperatur og lydhastigheit [9]. I utviklinga av denne nye algoritmen stod ein framføre fleire utfordringar slik som til dømes effekten av andre hydrokarbon med høgare molekylvekt enn metan og etan og effekten av dei inerte gassane CO_2 og N_2.

Denne algoritmen kan nyttast i ulike situasjonar:

1. ”Blind composition approach” (BCA1) der ein ikkje har noko kunnskap om den kjemiske samansetjinga i gassen i det heile tatt.
2. ”Blind composition approach” (BCA2) der ein ikkje kjenner samansetjinga av hydrokarbon, men har kjennskap til typiske innhald av CO_2 og N_2.
3. ”Typical composition approach” (TCA) der ein har kjennskap til både den kjemiske samansetjinga og typiske innhald av CO_2 og N_2.
4. ”Continuous composition approach” (CCA) der ein kjenner den kontinuerlege gassamansetjinga.

BCA1 og BCA2 vert nytta i dei tilfella der ein ikkje kjenner samansetjinga av naturgassen og TCA vert nytta når ein kjenner den typiske gassamansetjinga, men ikkje kjenner dei daglege variasjonane. Usikkerheita i tettleik og brennverdi er høgare ved BCA1 og BCA2 enn ved TCA. Målte verdiar i algoritmen er trykk, temperatur, lydhastigheit og gassamansetjing. Predikerte verdiar er tettleik og brennverdi for gassen. Algoritmen reknar ut tettleik og brennverdi frå målt lydhastigheit med avvik frå referanseverdien på mindre enn 0.5 – 1.0 % i dei tilfella der ein har kjennskap til både den kjemiske samansetjing og typiske innhald av CO_2 og N_2 (TCA). [8].

Usikkerheita i metoden når den er nytta på reelle data består av fleire deler, som usikkerheit i sjølve algoritmen, usikkerheit i dei underliggjande matematiske modellane og usikkerheit i variablane i **X**, det vil seie trykk, temperatur, lydhastigheit og kjemisk samansetjing av naturgassen.

Algoritmen for berekning av tettleik og brennverdi er testa for eit stort område ulike gassamansetjingar og over eit stort område av temperatur og trykk. Det er hevda at for å

kunne berekne tettleik i naturgass nøyaktig treng ein kunnskap om den kjemiske samansetjinga i gassen [8], [10].

Ved testing på gass frå fleire felt i Nordsjøen [11] er det funne at så lenge det typiske CO_2- og N_2innhaldet i gassen er kjend, altså for situasjonane BCA2, TCA og CCA, er det usikkerheita i algoritmen og usikkerheita i den målte lydhastigheita som står for det største bidraget når det gjeld usikkerheit. Det er vist at usikkerheita i algoritmen er låg og kan i tilfeller som TCA truleg vere ein systematisk feil. Dette gjeld i tilfeller der ein har funne usikkerheita i algoritmen til å liggje mellom 0.1 % og 0.6 % for trykk i området 140-160 bar. Dersom trykket er lågare vil også usikkerheita i algoritmen vere lågare. Det er vist at dersom usikkerheita i måling lydhastigheit er lita, vil også avviket for tettleik og brennverdi i forhold til referanseverdiar vere lite [11]. På den andre sida, i dei tilfella der usikkerheita i målt lydhastigheit er stor vil også feilen i tettleik og brennverdi vere stor. Dette peikar mot at målinga av lydhastigheit er eit nøkkelparameter i denne metoden. Avvik frå referanseverdien for lydhastigheit opp mot 2 m/s har blitt vist, noko som resulterer i avvik på om lag 1.2 % for tettleik og på om lag 0.6 % for brennverdi. Desse tala er funne ved bruk av TCA. Data frå BCA1 frå felt i Nordsjøen viser eit avvik frå referanseverdi på 1 – 2 % for tettleik og 0 – 1 % for brennverdi (Gullfaks november 2005). Her er varierer trykket i området 156 – 170 bar medan temperaturen er ca 47 ˚C. I dette tilfellet er det også vist at det ikkje er nødvendig å kjenne den nøyaktige gassamansetjinga. For data frå Draupner ser ein feil i området -1.5 – 0 % for tettleik og mellom 0.5 % og i overkant av 1 % for brennverdi. Trykket er her 10 bar og temperaturen varierer mellom – 5 og 10 ˚C. I dette tilfellet er prediksjonen lite avhengig av kunnskap om det kjemiske samansetjinga i gassen.[11].

Bruk av delvis minste kvadraters metode (PLS) for prediksjon av molare fraksjonar i naturgass har tidlegare vist å ha høgare prediktiv evne enn pseudo-inverse eller prinsipale komponent regresjon (PCR) teknikkar. Årsaka til dette er at sjølv om PCR bevarar så mykje som mogleg av variasjonen i **X**-blokka, så vil PLS beskrive kovariansen mellom **X** - og Y blokka [2].

1.3 Arbeidsteknikkar

Datagrunnlaget for denne masteroppgåva er basert på data frå simuleringar utført med eit internt dataprogram på Christian Michelsen Reaserch (CMR Instrumentation). Kalibreringsdatasettet består av ei simulert datamatrise med 4000 ulike situasjonar der trykk, temperatur og samansetjing av naturgassen er variert innanfor bestemte, logiske grenser. Variablane lydhastigheit, brennverdi og tettleik er simulert frå innsende verdiar for trykk, temperatur og kjemisk samansetjing. I utgangspunktet er trykk, temperatur og lydhastigheit kjende **X**-verdiar medan brennverdi og tettleik er **y**-verdiar som ein ynskjer å predikere ut frå modellen. Når det gjeld kjemisk samansetjing av naturgassen kan desse parametrane både fungere som **X** og **Y**. Ein kan nytte kjemisk samansetjing til å predikere brennverdi og tettleik og ein kan predikere kjemisk samansetjing ut frå trykk, temperatur og lydhastigheit.

Målet er ved bruk av multivariate metodar å lage modellar som kan nytte trykk, temperatur og lydhastigheit til å predikere kjemisk samansetjing i naturgassen, slik at ein ut frå den predikerte kjemiske samansetjinga vere i stand til å predikere brennverdi og tettleik. Ein ynskjer å undersøkje kor følsame modellane er for feilestimering i den kjemiske samansetjinga. Usikkerheita i modellen bør vere < 1 %, og ideelt sett 0,3 % eller betre. Særlege utfordringar er knytt til området der ein har høgt trykk og låg temperatur, her vil gassmolekyla i stor grad påverke kvarandre. Ein vil også nytte alternerande regresjon til raffinering av modellane, der dei kjemiske komponentane i naturgassen skiftar på å vere prediktor og respons. I denne samanhengen vil ein også sjå på kor sterke føringar må ein leggja på løysingane for å konvergera til eit meiningsfullt resultat. Til utføring av alternerande regresjon skal det lagast eit eige program i MATLAB som i tillegg til å utføre iterativ konsentrasjonsbestemming av dei kjemiske komponentane i naturgassen ved bruk av alternerande regresjon også skal rekne ut RSD for objekta. I denne oppgåva vil ein i all hovudsak nytte PLS modellering av kjemisk samansetjing ved BCA1-tilnærminga, prediksjon av kjemisk samansetjing med det utgangspunktet at ein ikkje kjenner samansetjinga av gassen. Modellane som er nytta til prediksjon er bygd ut frå typisk kjemisk samansetjing definert ved CMR Instrumentation. Ved prediksjon av dei ulike kjemiske komponentane tek ein utgangspunkt i at dei einaste kjende variablane er trykk, temperatur og lydhastigheit.

2. Teori og metode

2.1 Statistiske metodar

Standardavvik og varians

Variansen i datamaterialet vert målt med prøve variansen som er definert slik:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \qquad (2.1)$$

Variansen er summen av dei kvadrerte avvika mellom kvar prøve og prøvegjennomsnittet, dividert på størrelsen på prøvematerialet minus 1. Dersom det ikkje er varians tilstades i prøvematerialet vil kvar observasjon vere $x_i = \bar{x}$ og variansen $s^2 = 0$. Generelt kan ein seie at jo høgare verdi variansen har, jo større variasjon er det i datamaterialet.

Eininga til variansen blir den kvadrerte eininga til det opphavlege datamaterialet. Dette kan vere vanskeleg å tolke, ein nyttar difor ofte standardavviket som eit mål på variasjon [13]:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad (2.2)$$

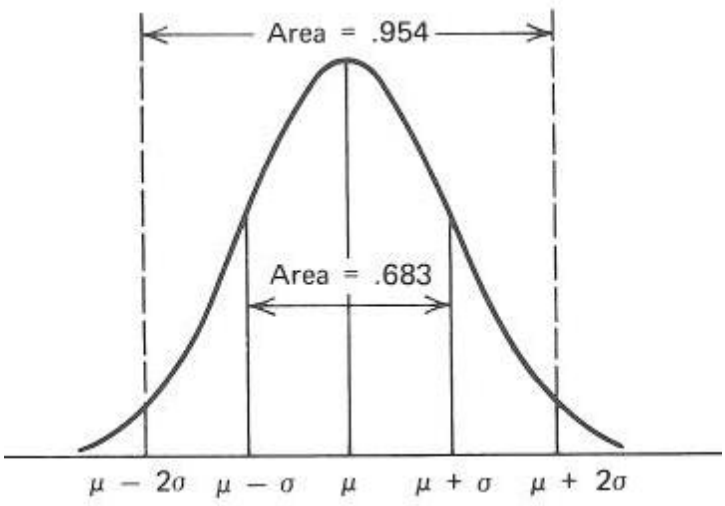
Normalfordeling

Normalfordelinga vert i mange tilfelle sett på som den viktigast fordelinga både når det gjeld teori og praksis innan statistikken. Dersom x er ein tilfeldig variabel vil sannsynsfordelinga til x vere definert som

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad -\infty < x < \infty \qquad (2.3)$$

Gjennomsnittet for normalfordelinga er μ $(-\infty < \mu < \infty)$ og variansen er $\sigma^2 > 0$

[14].



Figur 2.1 Normalfordeling, figur frå [15].

Visuelt sett er normalfordelinga ei klokkeforma kurve med gjennomsnitt μ slik det er vist i figur 2.1. Sannsynet for at ein tilfeldig variabel fell innanfor μ ± σ er 0.683 og sannsynet for at ein tilfeldig variabel fell innanfor μ ± 2σ er 0.954 [15].

Dersom gjennomsnittet er 0 og standardavviket er 1 har ein spesialtilfellet standard normalfordeling. Dette skriv ein som N(1,0).

Sentralgrenseteoremet

Sentralgrenseteoremet seier at summen av n uavhengige tilfeldig fordelte variablar er tilnærma normal, uavhengig av fordelinga til dei individuelle variablane. [14].

Dersom $x1, x2, \dots, xn$ er uavhengige tilfeldige variablar med gjennsomsnitt μ_i og varians σ^2_i og dersom $y = x1 + x2 + \dots + xn$, så vil fordelinga

$$\frac{y - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}} \qquad (2.4)$$

gå mot standard normalfordeling N(0,1) når n går mot uendeleg.

I mange tilfelle vert normalfordelinga sett på som den passande sannsynsmodellen for ein tilfeldig variabel, og sentralgrenseteoremet blir nytta som bevis for dette.

Kor stor n må vere vil variere, men generelt sett kan ein seie at normaltilnærminga vil vere bra dersom n ≥ 30. I tilfeller der n < 30 vil tilnærminga vere god berre dersom fordelinga

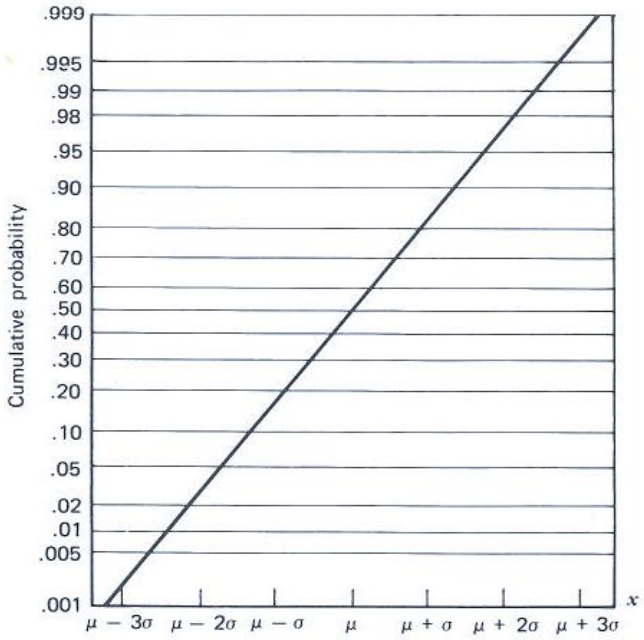
ikkje er så ulik normalfordelinga i utgangspunktet. Dersom populasjonen er kjent normalfordelt vil fordelinga av $\bar{X}$ følge normalfordelinga uavhengig av storleiken på prøvematerialet [16].

Normalfordelingsplott

Normalfordelingsplott er ei grafisk framstilling av data for å undersøkje om datamaterialet passar inn under normalfordelinga. Plottinga vert gjort på ein bestemt type grafpapir som omdannar grafen P[X≤x] til ei rett linje. Dette er vist i figur 2.2. Tolkinga av resultata er visuell og subjektiv. For ein tilfeldig prøve vil ein forvente at dei kumulative relative frekvensane liknar oppførselen til dei kumulative sannsyna. Ein treng helst 15-20 observasjonar for å få eit meiningsfullt plott.

Konstruksjon av normalplott:

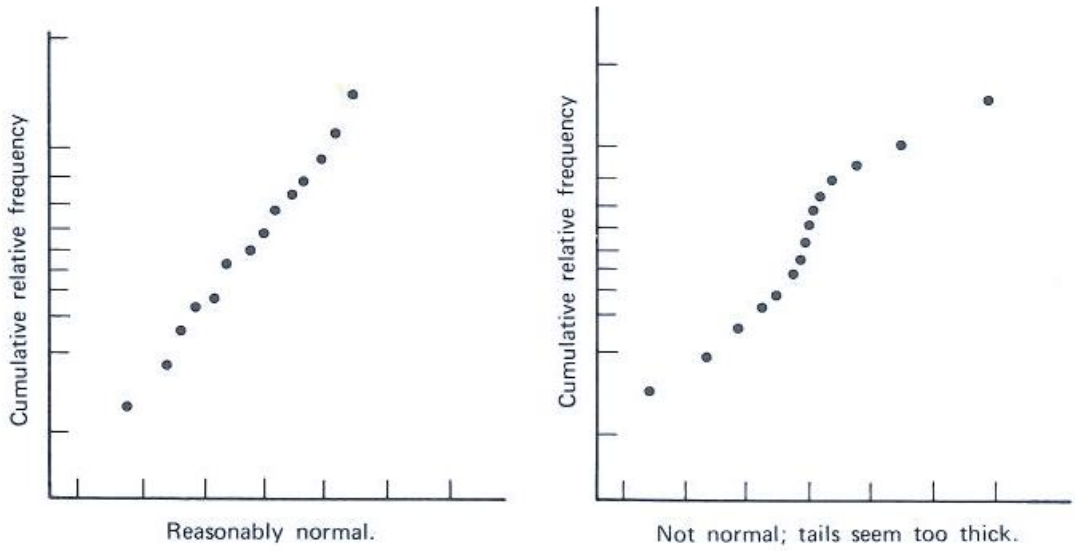
1. Ordne dei n observasjonane frå lågast verdi til høgast verdi
2. Finn ein skala på den horisontale aksen som er tilpassa alle observasjonane
3. Plott dei modifiserte kumulative relative frekvensane $\left(i - \frac{1}{2}\right)/n$ på den vertikale skalaen mot verdiane for den $i$te ordna observasjonen på den horisontale skalaen.
4. Undersøk plottet for avvik frå eit rettlinja mønster. Systematiske avvik tyder på avvik frå normalfordelinga [17].



Figur 2.2 Illustrasjon av normalfordelingsgrafpapir som viser ein graf av P[X≤x] når X er normalfordelt N(μ, σ). Figur frå [17].

Tolking av normalfordelingsplott:

Når punkta ligg nær ei rett linje kan ein anta normalfordeling. Dersom eit eller fleire objekt avvik frå ei rett linje er det grunn til å tru at ein her ikkje har normalfordeling (figur 2.3). I tillegg vil mønsteret i dei avvikande objekta kunne seie noko om årsaka til avviket frå normalfordelinga og ein kan utføre korrigerande handlingar [19]. Kryssingspunktet med y-aksen er eit estimat av populasjonens gjennomsnitt og stigningstalet er eit estimat av standardavviket σ. Eit eventuelt avvik frå normalitet kjem då tydeleg fram ved å studere forma på plottet. Asymmetri i datamaterialet vil resultere i endringar i stigningstalet i plottet [18].



Figur 2.3 Viser plott på normalfordelingspapir. Plottet til venstre viser normalfordelt datasett, plottet til høgre viser datasett med avvik frå normalfordelinga. Figur frå [17].

Standardisering og sentrering

Standardisering og sentrering er vanlege måtar å førebehandle eit datasett på ved kjemometriske metodar. Ved førebehandling oppnår ein å fjerne effektar som ikkje speglar den kjemiske samansetjinga [20].

Sentrering vert utført ved at matrisa **X** vert sentrert til matrisa **X**c ved at gjennomsnittet til alle prøvane $\bar{\boldsymbol{x}}$ vert trekt frå kvar enkelt prøve:

$$\boldsymbol{X}c = \boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{x}}^T \qquad (2.5)$$

Der **1** er ein vektor med dimensjonen N x 1 . Sentrering fører til at origo i koordinatsystemet vert flytta til gjennomsnittet av alle prøvane, [16] og ulikskapar mellom nivå på variablane vert fjerna [20].

Ved standardisering vert kvar kolonne dividert med kolonnen sitt tilhøyrande standardavvik og dette fører til at ulikskapar i variasjonsbreidde for variablane vert fjerna. Effekten av sentrering og standardisering vert då at alle variablane vil ha like stor påverknad på analysen [20].

I denne oppgåva er datasetta sentrert og standardisert før modellering.

Normalisering

Ved normalisering av eit datamateriale oppnår ein å gi objekta same relative eller absolutte størrelse.

Blokk normalisering går ut på dividere alle dei valde variablane med summen deira og dermed oppnår ein den relative fordelinga av variabelen i kvart objekt. Denne prosedyren vert også kalla normalisering til konstant sum.

$$100 * \frac{X_{NM}}{\sum X_{NM}} \rightarrow X_{NM} \qquad (2.6)$$

Der N representerer objekta og M representerer variablane.

Etter å ha utført denne transformasjonen vil den relative konsentrasjonen for kvart objekt summere til 100 [21].

Kovarians og korrelasjon

Å bestemme kovariansen mellom to variablar er ein måte å bestemme kor tett dei fylgjer dei same trendane. Kovariansen mellom **X** og **Y** er definert slik:

$$Cov(\boldsymbol{X}, \boldsymbol{Y}) = E[(\boldsymbol{X} - \mu_{\boldsymbol{X}})(\boldsymbol{Y} - \mu_{\boldsymbol{Y}})] = E(\boldsymbol{XY}) - \mu_{\boldsymbol{X}}\mu_{\boldsymbol{Y}} \quad (2.7)$$

Dette kan tolkast slik at dersom **X** og **Y** varierer i same retning er det truleg at store verdiar for **X** er assosiert med store verdiar for **Y** og små verdiar for **X** er assosiert med små verdiar for **Y**. Då vil både $(\boldsymbol{X} - \mu_{\boldsymbol{X}})$ og $(\boldsymbol{Y} - \mu_{\boldsymbol{Y}})$ vere enten store positive eller store negative og produktet av desse blir både stort og positivt. Dersom **X** og **Y** varierer i kvar sin retning vil ein derimot få ulikt forteikn på $(\boldsymbol{X} - \mu_{\boldsymbol{X}})$ og $(\boldsymbol{Y} - \mu_{\boldsymbol{Y}})$ og produktet vert negativt. Slik vil både forteiknet og størrelsen på kovariansen seie noko om forholdet mellom **X** og **Y**.

Ved å dividere kovariansen på standardavviket (σ) til **X** og **Y** får ein eit uttrykk for korrelasjonskoeffisienten:

$$Corr(\boldsymbol{X}, \boldsymbol{Y}) = \frac{Cov(\boldsymbol{X}, \boldsymbol{Y})}{\sigma_{\boldsymbol{X}}\sigma_{\boldsymbol{Y}}} \qquad (2.8)$$

Korrelasjonskoeffisienten har ein verdi mellom +1 og -1. Dersom verdien er +1 er variablane perfekt positivt korrelerte, og dersom verdien er -1 har ein perfekt negativ korrelasjon mellom variablane. Dersom dette er tilfelle kan ein predikere **Y** eksakt dersom ein kjenner **X**.

Jo nærare korrelasjonskoeffisienten ligg 0, jo vanskelegare er det å nytte den eine variabelen til å predikere den andre [22]. Korrelasjonskoeffisienten endrar seg ikkje sjølv om ein adderer eller multipliserer variablane med konstantar [23].

Histogram

Eit histogram er ei kompakt oppsummering av datamateriale. For å konstruere eit histogram for kontinuerlege data deler ein opp datamaterialet i intervall. Det er vanleg å finne mellom fem og 200 intervall i eit histogram og talet på intervall bør auke med aukande tal på observasjonar, n. Det fungerer ofte bra å nytte $\sqrt{n}$ intervall. Ein lar så den horisontale aksen representere skalaen for observasjonane og den vertikale aksen representere talet på observasjonar, også kalla frekvensen [24]. I mange tilfeller kan ein sjå at histogrammet avslører det som er den vanlegaste trenden, men ikkje alltid det som er den viktigaste trenden i datamaterialet.

2.3 Multivariate metodar

Kjemometri kan definerast som bruk av anvendt matematikk, statistikk og informasjonsvitskap til å velje optimale prosedyrar for å generere data og ekstrahere optimal informasjon frå data [12].

2.3.1 Prinsipal komponent analyse

Prinsipal komponent analyse (PCA) er ein av dei viktigaste multivariate metodane innan kjemometrien, metoden som i stor grad har endra kjemikaren sin måte å analysere data på. [25].

Når hovudmålet er å tolke data, er ein interessert i å beskrive systematisk variasjon ved hjelp av latente variablar, og då så få som mogleg. Det er også viktig å framstille data slik at dei er lette å tolke, slik ein kan gjere ved å nytte grafisk framstilling. PCA er eit godt verktøy til dette. Dei latente variablane i PCA vert kalla prinsipalkomponentar [26].

Ved å nytte PCA kan ein redusere talet på variablane frå mange hundre til relativt få utan å miste vesentlege deler av informasjonen i datamaterialet. Denne metoden har fått mykje å seie for mange område innan kjemometrien, og den er viktig av fleire grunnar. Ved å redusere dimensjonen på datamaterialet gjer ein det lettare å tolke resultata og finne strukturar og tendensar både når det gjeld likskapar og ulikskapar i datamaterialet. PCA finn nye, ortogonale latente variablane som er basert på dei kolineære målte variablane. Denne høge korrelasjonen mellom variablane blir då redusert dersom talet på komponentar er mindre eller lik rangen i det opphavlege datamaterialet. Dermed kan ein nytte datamaterialet i andre samanhengar, som til dømes regresjon. Det kan vere nyttig å redusere talet på variablar også når ein skal bruke andre analyser som klassifisering og regresjonsanalyse. PCA er også eit viktig verktøy når det gjeld å avsløre uteliggjarar i datamaterialet [27].

Ved å nytte PCA finn ein kombinasjonar av variablar eller faktorar som beskriv hovudtrendane i datasettet [28].

Matematisk grunnlag:

PCA byggjer på eigenvektordekomposisjon av kovarians- eller korrelasjonsmatrisa til prosessvariablane. For ei gitt datamatrise **X** med n rekker som representerer prøvane og m kolonnar som representerer variablane er kovariansmatrisa til **X** definert som:

$$cov(\boldsymbol{X}) = \frac{\boldsymbol{X}^T\boldsymbol{X}}{n-1} \qquad (2.9)$$

Likning (2.9) føresett at kolonnane i **X** har blitt sentrert før analysen er utført. Dersom kolonnane i **X** har blitt autoskalert, det vil seie både standardisert og sentrert, gir likning (2.9) korrelasjonsmatrisa til **X** [28].

Ekstrahering av prinsipalkomponentane (PC) går føre seg ved at ein først finn lineærkombinasjonen av dei opphavlege variablane som forklarer mest mogleg av variansen i **X**. Dette vert kalla prinsipalkomponent 1 (PC1). PC2 finnes ved at variansen som blir forklart av PC1 vert fjerna og PC2 vert ekstrahert. PC2 forklarer då mest mogleg at den variansen som PC1 ikkje forklarte. PC2 er ortogonal til PC1. Slik fortset ein å ekstrahere prinsipalkomponentar til den øvre grensa som er lik rangen av den opphavlege matrisa **X** [26].

PCA dekomponerer datamatrisa **X** slik at summen av ytreproduktet av vektorane $\mathbf{t}_i$ og $\mathbf{p}_i$ pluss eit residual **E** kan definerast slik:

$$\boldsymbol{X} = \boldsymbol{t}_1\boldsymbol{p}_1^T + \boldsymbol{t}_2\boldsymbol{p}_2^T + \ldots + \boldsymbol{t}_k\boldsymbol{p}_k^T + \boldsymbol{E} \qquad (2.10)$$

Der k må vere mindre eller lik dimensjonen av **X**. $\mathbf{t}_i$ er skårevektoren som seier noko om samanhengen mellom prøvane og $\mathbf{p}_i$ er ladningsvektoren som seier noko om samanhengen mellom variablane. [28].

Prinsipale komponentar er avhengige av kva skala ein har på variablane og dersom ein endrar skala vil dette påverke heile analysen. Dette fører til at dersom det er store skilnader på eining og variasjon kan det i mange tilfelle vere fornuftig å standardisere variablane før ein nyttar PCA [27].

Tolking av resultat etter utført PCA:

Koordinatverdiane for dei enkelte prøvane for kvar prinsipalkomponent vert kalla skårer (**t**). Dersom ein plottar to eller tre skårer i eit koordinatsystem vil ein få ei oversikt over korleis hovudvariasjonen i datamaterialet fordeler seg. I dei fleste tilfella vil det vere relevant å plotte dei komponentane som skildrar størst varians, altså dei første som blir ekstrahert. Dersom ein plottar $\mathbf{t}_2$ mot $\mathbf{t}_1$ og $\mathbf{t}_3$ mot $\mathbf{t}_1$ vil dette gje ei oversikt over korleis prøvane fordeler seg i forhold til kvarandre innanfor det reduserte koordinatsystemet. Skåreplott vil dermed gje viktig informasjon om relasjonar, likskapar, ulikskapar, grupperingar og ikkje minst kva prøvar som har medverka betydeleg til variasjonen langs dei ulike aksane.

Ladningane beskriv relasjonen mellom dei opphavlege variablane og kvar enkelt prinsipalkomponent. Ein kan plotte ladningane på same måte som beskrive for skårene i avsnittet over. Dersom ein variabel har høge ladningsverdiar tyder det at prøvane har variert mykje for denne variabelen langs den komponenten ein ser på. Dersom ein plottar skårer og ladningar i same plott kan ein finne ut kva variablar som er relatert til kvarandre og kva type variasjon mellom prøvane desse variablane kan knytast til. Dette vert kalla biplott.

Eigenverdiane viser kor mykje av den totale variansen kvar enkelt prinsipalkomponent har forklart. Variansane kan akkumulerast slik at summen av nokre valde eigenverdiar vil seie kor mykje dei tilsvarande prinsipale komponentane har forklart av den totale variansen.

X-residuala beskriv den variasjonen som ikkje er modellert av dei komponentane som er berekna, og vert kalla restvariasjonen. Plott av residuala vil kunne avsløre kva prøvar eller variablar som ikkje har latt seg modellere så godt som resten, og residuala er difor eit viktig verktøy for å avsløre uteliggjarar både når det gjeld prøvar og variablar. [27].

2.3.2 Partial Least Square - delvis minste kvadraters metode

Partial Least Square (PLS) vert ofte kalla hovudregresjonsteknikken for multivariate data. [29].

PLS er ein mykje nytta teknikk dersom ein ynskjer å sjå på samanhengen mellom to sett variablar, **X** og **Y**. I slike tilfeller bør ein ikkje la dei latente variablane forklare så mykje som mogleg av variansen i **X**, ein bør i staden forsøke å trekkje ut den informasjonen i **X** som samsvarar med informasjonen i **Y**. Ein vil då prøve å trekkje ut dei latente variablane som best mogleg beskriv **Y**-rommet ved å dekomponere **X**-rommet [26]. PLS skil seg frå PCR ved å nytte **Y**-variablane aktivt under den bilineære dekomponeringa av **X**-matrisa. Ved å balansere **X**- og **Y**-informasjonen reduserer metoden påverknaden av store og irrelevante X-variasjonar i kalibreringsmodellen [30].

Modellen for PLS kan uttrykkast slik:

$$\boldsymbol{X} = 1 * \overline{\boldsymbol{X}} + \boldsymbol{TP}^{'} + \boldsymbol{E} \qquad (2.11)$$

$$\boldsymbol{Y} = 1 * \overline{\boldsymbol{Y}} + \boldsymbol{UC}^{'} + \boldsymbol{F} \qquad (2.12)$$

$$\boldsymbol{U} = \boldsymbol{T} + \boldsymbol{H} \qquad (2.13)$$

Der **T** er skårematrisa for **X**-variablane, **P** er ladningane til **X**-variablane, **U** er skårematrisa til **Y**, **C** er ei matrise som viser forholdet mellom **Y** og **T**, **E**,**F**, og **H** er residual. I PLS algoritmen finn ein også eit sett ladningar kalla vekter (**W**). Desse er eit uttrykk for korrelasjonen mellom **U** og **X** og vert nytta til å rekne ut **T**. Modellen slik den er vist i likningane (2.11) – (2.13) vil geometrisk svare til å tilpasse ei linje, eit plan eller eit hyperplan til både **X** og **Y**. Denne tilpassinga vil vise seg som punkt i det multidimensjonale rommet der målet er å tilpasse modellen til dei originale datatabellane **X** og **Y** samt maksimere kovariansen [38].

Hovudbruksområdet for PLS er kaliberering [26]. PLS-modellar kan vere svært robuste når ein ynskjer å predikere framtidige prøvar som har lik samansetjing som det opphavlege datamaterialet. Det er viktig å ta omsyn til at bruk av PLS ikkje kan kompensere for dårleg designa eksperiment eller utilstrekkelege eksperimentelle data [29].

PLS er i slekt med både prinsipal komponent regresjon (PCR) og multippel lineær regresjon (MLR). I PCR ligg fokuset på å finne faktorar som beskriv mest mogleg av variansen i prediktorvariablane. I MLR prøver ein å finne ein enkelt faktor som best korrelerer prediktorvariablane med predikerte variablar. I PLS derimot, vil ein søkje å finne faktorar som gjer begge deler, både beskriv varians og oppnår korrelasjon. PSL har til hensikt å maksimere kovariansen og PLS regresjon vert betrakta som eit medlem av dei bilineære metodane. Det som kjenneteiknar slike metodar er at dei er kraftige og fleksible tilnærmingsmåtar til multivariat kalibrering. Denne typen modellering kan gje informative og presise prediktorar ved å projisere mange variablar på få variablar [30]. Den delen av datamaterialet som ikkje blir forklart av modellen vert kalla residaual. Store Y-residual tyder på at modellen ikkje er tilfredsstillande og eit normalfordelingsplott av residuala vil vere nyttig for å identifisere eventuelle uteliggjarar. I PLS har ein også **X**-residual og desse er nyttige i forhold til å identifisere uteliggjarar i X-rommet [32].

Ein metode for å kalkulere PLS-modell parametrane er non-iterativ partial least square (NIPALS). I denne metoden reknar ein ut skårer **T** og ladningar **P** og i tillegg eit sett vekter **W**. **W** er nødvendig for at ein skal oppretthalde ortogonale skårer. Denne algoritmen kan ein også nytte dersom det er meir enn ein predikert variabel **Y**. I slike tilfeller blir skårer **U** og ladningar **Q** rekna ut for **Y**-blokka. Ein vektor som seier noko om koeffisientane for det indre forholdet vert kalla b og relaterer **X**- og **Y**-blokk skårene [28].

Ein annan algoritme for utrekning av PLS-modellar er SIMPLS, utvikla av Sijmen de Jong [28]. Det er hevda at denne algoritmen har fleire fordeler i forhold til NIPALS: faktorane vert rekna ut frå originale datasettet og vektene vert difor lettare å tolke og kan finnast utan utrekninga av den inverse til matrisa. Algoritmen involverer heller ikkje oppdeling av **X**-matrisa og er dermed raskare. SIMPLS er heilt ekvivalent til PLS1 for modellar med berre ein respons [31].

Den lineære PLS-modellen finn nokre få "nye" variablar som er estimat av dei latente variablane eller deira rotasjon. Desse nye variablane vert kalla **X**-skårer og skrives som $\mathbf{t}_i$ (i= 1,2,..,I). **X**-skårene er gode prediktorar for **Y**, og modellerer også **X**, dermed blir både **X** og **Y**, i alle fall delvis, modellert av dei same latente variablane. **X**-skårene blir estimert som lineære kombinasjonar av dei originale variablane $\mathbf{x}_k$, som har koeffisientane, eller vektene, $\boldsymbol{w}_{ki}$ (i = 1,2,..,I) [32].

Som nemnt tidlegare skil PLS seg frå PCR ved at ein her nyttar **Y**-variablane aktivt når ein utfører den bilineære dekomponeringa av **X**. Ved å balansere **Y**- og **X**-informasjonen vil metoden redusere påverknaden av store, men irrelevante, **X**-variasjonar i kalibreringsmodelleringa. Eit minus ved PLS er at for støy i **Y** har PLS ein større tendens til overtilpassing enn PCR, noko som gjer at valideringa blir særleg viktig [30]. Den vanlegast tilnærminga til PLS vert ofte kalla PLS1 [29], og Sirius nyttar denne noniterative metoden.

2.4 Alternerande regresjon

Alternerande regresjon (AR) er ein metode som vart utvikla for å finne narkotiske stoff i dopinganalyser [33]. Metoden vart først utvikla på empirisk basis, og det teoretiske grunnlaget kom seinare. Metoden høyrer til gruppa av raske lokale metodar på same måte som metodane basert på faktoranalyse.

Den grunnleggjande prinsippet for alternerande regresjon er lineær regresjon. Ein har redusert talet på variablar frå den opprinnelege **X**-matrisa og sit no igjen med betydeleg færre uavhengige variablar til å utføre lineær regresjon med [33]. I første del nyttar ein lineær regresjon for å finne konsentrasjonar for dei ulike stoffa. Deretter blir dei predikerte verdiane brukt som startpunkt [34]. I AR vert altså ein syklus av to eller fleire problem repetert til ein når eit konvergenskriterium [35]. Startpunktet kan i mange tilfeller vere ei matrise fylt med tilfeldige tall [34]. Det kan også leggjast ulike føringar inn i algoritmen, som til dømes at alle konsentrasjonar må vere positive.

Den alternerande regresjonsalgoritmen er ein algoritme som arbeider raskt. Iterasjonane i seg sjølv er raske å rekne ut og ein treng få iterasjonar.

Matrisa vert definert slik:

X = konsentrasjonsmatrise, består av variablar nytta til prediksjon.

Stega i algoritmen kan ein definere på følgjande måte:

1. Anta startverdiar i **X**. I denne oppgåva har ein nytta både middelverdiar i dei definerte områda og predikerte verdiar for dei kjemiske komponentane.
2. Utfør lineær regresjon for den første av dei kjemiske komponentane for det første objektet i **X**.
3. Erstatt startverdien for den første kjemiske komponenten med den predikerte verdien funnen i punkt 2.
4. Oppdater **X** slik at både andregradsledd og vekselverknadsledd inneheld den predikerte verdien funnen i 2.
5. Gjenta steg 2 og 4 til alle dei kjemiske komponentane er predikerte.
6. Rekn ut RSD for objektet.
7. Gjenta steg 2 til 6 til ein når konvergenskriteriet.
8. Gjenta steg 2 til 6 for alle objekta i **X**.

Validering av resultat frå Alternerande Regresjon

Resultata bør sjekkast ut frå både eit kjemisk og eit matematisk utgangspunkt. Evna resultata har til å repeterast matematisk må sjekkast. Dette kan ein gjere ved at ein utfører alternerande regresjon med ulike startverdiar og ser om det fører til statistisk signifikante endringar i resultata. Den totale evna til repetisjon i ein alternerande regresjonsprosess er ein indikasjon på at komponentane er funne på ein konsekvent måte. Dei kjemiske resultata er det einaste i alternerande regresjon som vert kontrollert av analytikaren, resultata av desse bør vurderast ut frå ein kjemisk ståstad [35].

2.5 SIMCA RSD og identifisering av uteliggjarar

Residual standard avvik (RSD) utgjer forskjellen mellom det originale standardavviket og det standardavviket det er tatt høgde for i modelleringa. RSD refererer til **X**-blokka [25].

I Soft Independent Modelling of Class Analogies (SIMCA) terminologien vert s_k kalla residual standard avvik (RSD) til objektet k, og s_k er relatert til avstanden frå eit objekt til modellen slik det er vist i formel (2.14):

$$S_k^2 = \boldsymbol{e}_k{}'\boldsymbol{e}_k/(M-A) \qquad (2.14)$$

Der $\mathbf{e}_k$ er residualet til objektet, M er talet på variablar og A er talet på prinsipalkomponentar i modellen.

Ved å dividere på (M-A) får ein eit avstandsmål som er uavhengig av talet på variablar (M), og korrigert for tap av fridomsgrader i forhold til tilpassing av A prinsipalkomponentar.

RSD for objekta kan samlast i ein avstandsvektor **s** som har dimensjonen Nx1. Den kvadrerte avstandsvektoren **s** definerer gjennomsnittleg residual standardavvik av klassen:

$$\boldsymbol{S}_c^2 = \boldsymbol{s}'\boldsymbol{s}/(N-A-1) \qquad (2.15)$$

Der N er talet på objekt.

Ved å dividere på (N-A-1) gir det ein skala som er uavhengig av talet på objekt og korrigert for tap av fridomgrader i forhold til kolonnesentrering og tilpassing av A prinsipalkomponentar.

Samanlikning av RSD for objektet k med gjennomsnittleg RSD for klassen gir eit direkte mål på kor likt objektet er klassemodellen. Wold introduserte F-testobservatoren for samanlikning av $\mathbf{s}_k^2$ og $\mathbf{s}_c^2$ i eit forsøk på å finne ein kvantitativ basis for denne testen. Talet på fridomsgrader ein nyttar til å finne den kritiske F-verdien er (M-A) for $\mathbf{s}_k^2$ og (N-A-1) for $\mathbf{s}_c^2$. F-testen kan nyttast til å rekne ut ei øvre grense for RSD for objekt som høyrer til klassen:

$$\boldsymbol{s}_{max}^2 = \boldsymbol{s}_c^2 F_{crit} \qquad (2.16)$$

F_{crit} vert vanlegvis bestemt ved signifikansnivået p=0.05 eller p=0.01.

Når ein nyttar F-testen til å rekne ut øvre grenser for akseptabel resiudal avstand kan ein seie at ein har lukka SIMCA modellen rundt prinsipalkomponentane. Modellen er derimot framleis open langs kvar prinsipalkomponent og Wold nytta difor scoreverdiane $\mathbf{t}_{min,a}$ og $\mathbf{t}_{maxa}$ og deira spreiing $s_{t,a}$ langs kvar komponent til å definere nedre og øvre grense for skårene:

$$\boldsymbol{t}_{lower,a} = \boldsymbol{t}_{min,a} - \frac{1}{2}\boldsymbol{s}_{t,a} \quad (2.17)$$

$$\boldsymbol{t}_{upper,a} = \boldsymbol{t}_{max,a} + \frac{1}{2}\boldsymbol{s}_{t,a} \quad (2.18)$$

der

$$\boldsymbol{s}_{t,a}^2 = \boldsymbol{t}_a{}'\boldsymbol{t}_a/N = \boldsymbol{g}_a/N \quad (2.19)$$

No er SIMCAmodellen stengt i alle retningar.

Det er blitt påpeika at desse grensene ikkje er grunngjevne teoretisk, men erfaring viser at dei gir fornuftige resultat.

Bruk av F-test for å bestemmer øvre grenser for residualavstanden er ikkje utan problem. F-testobservatoren gir smale konfidensintervall når talet på variablar er stort i forhold til talet på objekt. Dette ser ein også når talet på objekt er relativt stort, så lenge det er sterk korrelasjon mellom variablane. Ei løysing på dette problemet er å nytte leave-out-one-block-of-samples prosedyren i validering av modellen. Denne prosedyren vil gje større konfidensintervall rundt modellen. Ein kan også rekne ut F_{crit} ved å nytte eit lågare tal på fridomsgrader [37].

Leverage

Leverage er eit mål på påverknaden eit objekt har på ein modell. Dersom eit objekt har høg leverage kan det ha stor påverknad på modellen og høg leverage for eit objekt kan indikere at objektet er ein uteliggjar som bør fjernast frå modellen [38].

SIMCA reknar ut leverage for objekta i X og Y rommet som diagonalelementa til matrisene $\mathbf{H_o}$ og $\mathbf{H_Y}$:

$$\boldsymbol{H_0} = \boldsymbol{T}(\boldsymbol{T}'\boldsymbol{T})^{-1}\boldsymbol{T}' \quad (2.20)$$

$$\boldsymbol{H_Y} = \boldsymbol{U}(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}' \quad (2.21)$$

Der **T** er skårematrisa til **X** og **U** er skårematrisa til **Y**.

Leverage har verdi mellom 0.0 og 1.0. Høg leverage i kombinasjon med låg RSD for eit objekt indikerer at objektet er ein uteliggjar [39].

Identifisering av uteliggjarar

Feil og uventa fenomen er uunngåelege i den verkelege verda. Dette gjeld både i forsking og i rutine laboratorieundersøkingar. Identifisering av slike unormale observasjonar, uteliggjarar, er difor viktig. I forhold til den tradisjonelle univariate kalibereringskurva vil multivariat kalibrering gi eit mykje betre verktøy for å automatisk identifisere uteliggjarar. Det er mange typar feil som kan avslørast og behandlast, til dømes feil i **X**, feil i **Y**, feil i kalibreringsdata og feil i framtidige predikerte data. I tillegg kan unormale resultat innehalde viktig informasjon [40].

Generelt representerer uteliggjarar dataelement som enten er irrelevante, svært feil eller unormale i forhold til resten av datamaterialet samanlikna med hovudtyngda av datamaterialet. For å sikre optimal bruk at datamaterialet er det viktig å identifisere slike uteliggjarar. Både objekt, variablar og enkelte dataelement kan oppføre seg som uteliggjarar. Uteliggjande objekt er dei viktigaste av uteliggjarane og er eit objekt der datavektorane $\mathbf{x}_i$ og/eller $\mathbf{y}_i$ avviker frå majoriteten av dei objekta det er kalibrert for, enten fordi det er feil i data eller fordi det er feil i samansetjinga eller fysisk tilstand for det objektet som er analysert. For å forsikre seg om korrekt bruk av moderne analytiske instrument bør identifisering av uteliggjarar inkluderast som eit automatisk og viktig del av prosedyren [40].

Ved kalibrering er det viktig å avsløre og kanskje fjerne eller korrigere data frå objekt eller variablar som elles kunne setje ned prediksjonsmoglegheitene til dei estimerte kalibreringskoeffisientane. Ved predikering av ukjente objekt er det viktig å ha metodar for å avsløre unormale situasjonar, noko som aukar konfidensen for dei predikerte konsentrasjonsresultata [40].

Uteliggarar er viktige og uunngåelege i den vitskapelege prosessen. Det er difor viktig å prøve å lære av dei. Uteliggarar må altså behandlast kritisk: når dei er identifiserte må ein korrigere eller ignorere dei dersom dei ser ut til å øydeleggje modellen. Det er også viktig å prøve å forstå årsaka til kvar uteliggjar. Ein skil mellom ”gode” og ”dårlege” uteliggjarar. Eit

døme på ein god uteliggjar er det objektet med høgast eller lågast analytisk nivå. Dersom responsen er relativt lineær vil slike uteliggjarar vere svært informative og bør behaldast i kalibereringsdatasettet. Dårlege uteliggjarar er som regel dei ein ikkje kan forstå og forklare og som har sterk påverknad på modellen. Dersom ein ikkje er sikker på kva for ein av desse kategoriane ein uteliggjar tilhøyrer kan ein slette den aktuelle uteliggjaren frå datasettet og repetere kalibreringa for å vere sikker på at uteliggjaren ikkje influerer eller øydelegg predikeringa. Denne strategien er særleg god dersom uteliggjaren skuldast feil på utstyret eller feil hos operatøren. Denne strategien bør i utgangspunktet berre nyttast når ein har eit stort datasett og få uteliggjarar [40].

2.6 Validering

Validering av modellar skjer alltid ved bruk av nye prøvar og ikkje med dei prøvane modellen er bygd på bakgrunn av. Unntak frå dette er intern validering.

Validering er ein svært viktig del av kalibreringsmetodane. Døme på spørsmål som ein må svare på er:

- Kor mange signifikante komponentar treng ein for å beskrive datasettet?
- Kor bra blir dei ukjente prøvane predikert?
- Kor representativt er datamaterialet som er nytta til å bygge modellen?
  [41].

Kvaliteten på den multivariate kalibreringa er viktig i fleire ledd:

- Brukaren av modellen for å predikere **Y** frå **X** treng mange typar statistisk informasjon. Særleg er det viktig med maksimumsfeil og gjennomsnittleg forventa presisjon. Brukaren må også kjenne området for bruk av ein gitt kalibreringslikning. Usikkerheita for kvar individuelle framtidig prediksjon er også viktig å kjenne.

- Personen som kaliberer instrumentet vil ofte vere interessert i dei statistiske moglegheitene for ulike kalibreringsmetodar. Dette vil vere nyttig i arbeidet med å velje ut kalibreringsdata og kalibreringsmetode i forhold kva behov ein har definert. Kalibreringsmetoden må vere tilpassa den gitte problemstillinga, men statistisk innsikt vil også vere nødvendig sidan det er lett å gjere grove statistiske feil.
- Produsenten av instrumenta ynskjer å optimalisere instrumentet sin prestasjon når det gjeld forventa kriterium kunden måtte ha i dei ulike tilfella. Produsenten bør også vere i stand til å beskrive predikerte moglegheiter for gitte situasjonar.
  [42].

Ulike typar prediksjonsfeil:
Det er mange ulike faktorar som kan verke inn på kvaliteten av ei kalibreringslikning:

- Modellfeil. Dette omhandlar tilpassing av kalibreringsmodellen. Dette er viktig fordi dersom modellen er dårleg tilpassa datamaterialet vil kalibreringslikninga aldri gi presise resultat. Sidan kalibreringsmodellering vanlegvis omhandlar lokal tilpassing til ein ukjent vil alle kalibreringsmodellar truleg innehalde noko feil i modellen.
- Kalibreringsdatasettet er ikkje representativt. Dette kan bli tilfelle dersom kalibreringsdatasettet ikkje inneheld det totale området av variabilitet i populasjonen i forhold til framtidige objekt. I slike tilfeller står ein i fare for dårleg prediksjon for objekt som ligg utanfor kalibreringsområdet. Ein kan då sjå god prediksjon i det området kalibreringsdatasettet gjeld, men ein kan ikkje ekstrapolere og forvente god prediksjon i området utanfor kalibreringsdatasettet.
- Tilfeldig støy i kalibrering og predikerte data. Støy i kalibreringsdata kan påverke estimeringa av kalibreringsparametrane i negativ grad, noko som resulterer i låge prediksjonseigenskapar for modellen.
  [42].

Intern validering - Kryssvalidering

Intern validering omhandlar validering frå sjølve kalibreringsdatasettet. Vurdering av modellen basert på intern validering og er ikkje det same som å teste kor bra modellen predikerer nye objekt. [44].

Kryssvalidering er eit viktig verktøy innan kjemometrien. Kryssvalidering går føre seg slik at ein sjekkar den prediktive evna ein modell har ved at ein del av datasettet predikerer den gjenståande delen av datasettet. [41]. I full kryssvalidering blir kalibreringa repetert i gonger der kvar repetisjon behandlar ein $i'te$ del av heile kalibreringsdatasettet som prediksjonsobjekt. Til slutt, når alle kalibreringsobjekta er blitt behandla som prediksjonsobjekt, kan den estimerte kvadrert gjennomsnittsfeilen (mean square error, MSE) for kryssvalidering (MSECV) reknast ut. Sidan full kryssvalidering er basert på repeterande kaliberingar, noko som kan vere tidskrevjande for datamaskina, er eit alternativ å utføre kryssvalidering ved å splitte kalibreringsdatasettet i G (G<i) deler og dermed kalibrere G gonger, kvar gong testar ein ca (1/G)del av kalibreringsdatasettet. [44].

Målet med kryssvalidering er å bestemme rett tal på komponentar i ein modell. Komponentar langt nede i modellen vil ofte representere støy og desse bør ikkje inkluderast i modellen dersom denne blir nytta til å predikere ukjende objekt. Kryssvalidering kan også nyttast som ein realistisk feilestimator for den prediktive evna til ein modell. [41].

Denne typen intern validering vil på same måte som ekstern validering tilnærme seg å validere kalibreringsmodellen på uavhengige data. I motsetning til ekstern validering vil denne metoden ikkje kaste bort data berre på testing.

Sirius gir ut verdien kryssvaliderings-standardavvik (CsvSD) som er eit uttrykk for ratioen mellom den totale prediksjonsfeilen til ein modell etter at ein har inkludert ein ny komponent og det totale residuelle standardavviket før denne komponenten vart inkludert. Generelt sett skal CsvSD for den aktuelle komponenten vere < 1 for at denne skal kunne inkluderast i modellen.

Prediksjonseigenskapar

Når ein predikerer ein verdi $\hat{y}$ i eit framtidig ukjent objekt antar ein at den korresponderande "ekte" y eksisterer sjølv om ein ikkje har målt denne og heller aldri kjem til å vere i stand til å måle den med 100 % tryggleik. Ein ynskjer at differansen mellom $\hat{y}$ og y skal vere så liten som mogleg, og for å ta høgde for både positive og negative forskjellar så ynskjer ein å ha $(y - \hat{y})^2$ så liten som mogleg. Ein ynskjer også å minimere gjennomsnittleg prediksjonsfeil for alle objekta som kalibreringa omfattar. Dette kan ein beskrive ved hjelp av den statistiske størrelsen mean squared error (MSE) i formel (2.22):

$$MSE = E(\boldsymbol{y} - \hat{\boldsymbol{y}})^2 \qquad (2.22)$$

MSE må estimerast på bakgrunn av datasettet, og dette kan gjerast på to måtar. I enklare tilfelle kan ein estimere MSE basert på teoretiske formlar frå lineær regresjonsteori. For dei kraftigare og meir fleksible kalibreringane som til dømes PCA og PLS er den statistiske teorien ikkje godt nok utvikla til at ein kan nytte slike formlar bortsett frå i svært enkle tilfelle. MSE må difor estimerast frå ein reell samanlikning av data frå y og $\hat{y}$ frå eit begrensa tal objekt frå kvart av kalibreringssetta (intern validering) eller frå ein separat test eller testdatasett (ekstern validering). Desse involverer også statistiske fordelingsvurderingar: for å oppnå god estimering av gjennomsnittlege prediksjonseigenskapar må settet av testobjekt dette er basert på vere representativt for heile populasjonen av framtidige ukjende objekt. Dersom dette ikkje er tilfelle vil den estimerte MSE vere svært villeiande. [43].

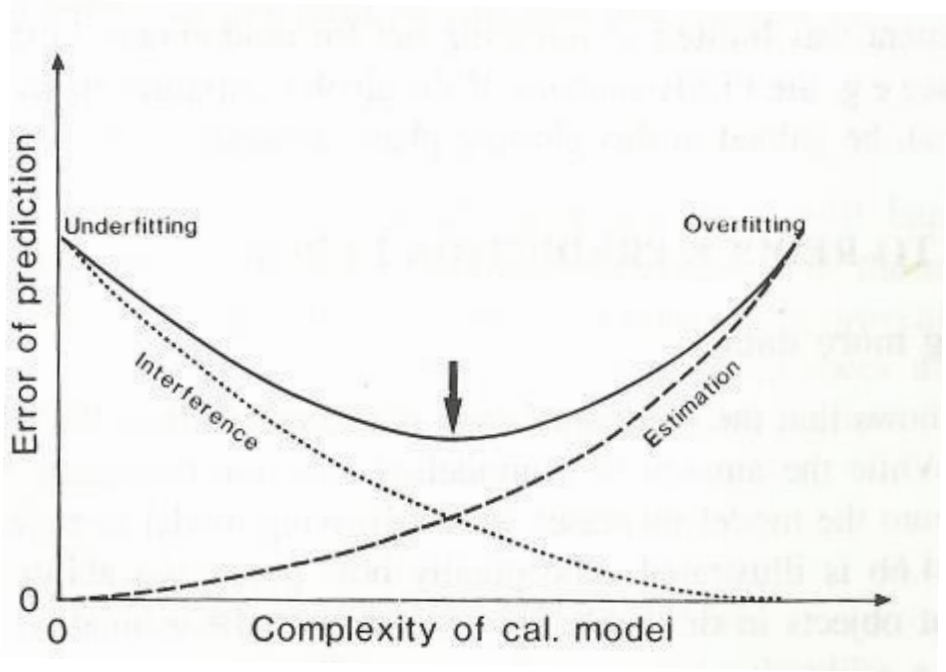
Forklart og predikert varians

R^2 kan definerast som den forklarte variansen og seier noko om modellen si tilpassing. R^2 varierer mellom 0 og 1, der 1 betyr perfekt tilpassing til modellen og 0 betyr ingen tilpassing i det heile tatt. Q^2 kan definerast som den predikert variansen og seier noko om den prediktive evna til modellen. I PCA viser R^2 og Q^2 til **X**, men i PLS viser R^2 og Q^2 oftast til **y**. R^2 vil raskt gå mot 1 når modellen blir kompleks, til dømes dersom mange komponentar vert inkludert. Dette er derimot ikkje tilfelle for Q^2. Ved evaluering av ein modell er det nyttig å sjå på desse verdiane og generelt kan ein seie at utan høg R^2 er det ikkje mogleg å få høg Q^2. $Q^2 > 0.5$ vert sett på som bra, medan $Q^2 > 0.9$ vert sett på som svært bra. Differansen

mellom R^2 og Q^2 bør ikkje vere for stor, helst ikkje over 0.2 – 0.3. I tilfeller der differansen overstig 0.3 kan det tyde på at modellen inneheld for mange irrelevante ledd eller at det er uteliggjarar tilstades. [45].

Bruk av predikert feil for å velje kalibereringsmodell

Ei av utfordringane ved å velje rett kalibreringsmodell er korleis bestemme rett tal på komponentar i modellen som seinare skal nyttast til prediksjonar. Målet med multivariat kalibrering er nettopp å redusere prediksjonsfeilen ved å modellere kjemisk og fysisk interferens som elles ville kunne øydeleggje konsentrasjonsbestemmingane.



Figur 2.4 Illustrasjon av korleis prediksjonsfeilen endrar seg med aukande tal på komponentar inkludert i modellen. Figur frå [46].

Det er hovudsakleg to bidrag til den predikerte feilen: den gjenståande interferensfeilen og den estimerte feilen. Den første er den systematiske feilen som er forårsaka av umodellert interferens i datamaterialet og den siste er forårsaka av ulike typar tilfeldig målestøy. Desse to bidraga har motsatt trend ved stigande tal på komponentar inkludert i modellen. Interferensfeilen minkar med aukande tal på komponentar og den estimerte feilen aukar

med aukande tal på komponentar slik det er vist figur 2.4. Optimalt tal på parameter ligg då i skjeringspunktet mellom desse kurvene. Dette er viktig fordi det er grenser for kor mange parameter ein kan estimere med høg presisjon frå eit gitt sett med kalibreringsdata. Bruk av for få komponentar vert kalla undertilpassing og modellering av for mange komponentar vert kalla overtilpassing av modellen. [46].

Ein teknikk for å velje rett tal på komponentar som skal inkluderast i modellen er å rekne ut prediksjonsmoglegheita for det ulike talet på faktorar ved anten prediksjonstesting eller intern validering av kalibreringsdata. Visuell tolking av plott er også eit viktig instrument i val av modell. [46].

Eit viktig spørsmål vil vere korleis ein kan redusere prediksjonsfeilen. Først og fremst kan ein bruke større datasett. Predikert feil kan kallast summen av effektane av undertilpassing og overtilpassing. Estimert feilkurve blir betre for store kalibereringssett. Ein kan også bruke betre data. Dette omhandlar støynivået til instrumentet som vert nytta, men også tillaging av objekta. Ved forenkling av populasjonen for kalibrering kan ein i enkelte tilfeller kan det få betre resultat dersom kalibreringspopulasjonen blir delt opp i fleire mindre underpopulasjonar slik at til dømes lineær approksimasjon blir meir tilfredsstillande. Til slutt kan ein forsøke bruk av ein betre modell. Årsaka til dårlege resultat kan vere at modellen er upassande, det kan då vere nødvendig å sjekke tilpassinga av modellen. [46].

OPPSUMMERING VALIDERING

For dei ulike variablane er det laga ei rekke modellar. Når det gjeld å velje den rette modellen i kvart enkelt tilfelle er det ein del ulike punkt som vert vurdert:

- Høg forklart varians i y
  Når ein byggjer modellar i Sirius aukar forklart varians i y med talet på komponentar. Ved å inkludere mange komponentar får ein dermed høg grad av forklart varians i y. I denne oppgåva vert modellane bygd av simulerte data som i utgangspunktet ikkje inneheld støy. Dermed vil alle bidrag til modellen vere signifikante. Generelt sett vil det vere fare for overtilpassing dersom ein inkluderer for mange komponentar i modellen.

- Kryssvalidering

  CsvSD skal vere mindre enn 1.00 for at komponenten skal kunne inkluderast i modellen.

- Normalfordelte responsresidual

  Ein god modell skal ha tilnærma normalfordelte responsresidual. Dersom normalfordelingskurva er S-forma er dette eit teikn på at modellen ikkje er bra. Dette inneber at nokre objekt har høgare responsresidual enn forventa verdi og dermed ligg lenger vekk frå modellen enn forventa. I denne oppgåva er datamaterialet utan støy. Alle avvik vil difor vere systematiske, og ein krev ikkje normalfordelte responsresidual for å vurdere ein modell som god. Likevel, sidan ein ikkje kjenner detaljane rundt simuleringa veit ein heller ikkje noko om eventuelle svakheitar ved algoritmen og ein kan dermed ikkje utelukke at normalfordelingsplotta kan vere av interesse.

- Testing av modellar: R^2-verdi, kumulativt forklart variasjon

  R^2-verdi nær 1,00 for den siste komponenten tyder på ein god modell, men høg R^2-verdi kan også skuldast overtilpassing.

- Q^2-verdien er eit mål på intern prediktiv evne for regresjonsmodellen.

  Dette er eit betre uttrykk for modellen sin prediksjonsevne enn R^2-verdien. Prediksjonsfeil vert vurdert ved prediksjon av nye objekt og vert målt i einingane til den målte variabelen og i prosent I denne oppgåva er prediksjonsfeilen eit uttrykk for gjennomsnittsverdien av residuala uttrykt i absoluttverdi. Prosent prediksjonsfeil er rekna ut etter denne formelen:

$$\%\ prediksjonsfeil\ = \frac{prediksjonsfeil\quad i\ absoluttverdi}{Gjennomsnittsverdi\quad for\ det\ aktuelle\ \ omr\aa det}\ x\ 100 \quad (2.23)$$

- Plott av responsresidual mot RSD

  Når ein plottar responsresidual mot RSD vil det vere eit bra teikn at det er aukande RSD-verdi med aukande responsresidual. Dette tyder på at dei objekta som har høgast avvik i **X** også har høgast avvik i **Y**, altså at dei objekta som ligg lengst frå modellen også vert predikert langt frå sann verdi.

2.7 Selektivitetsratio og targetrotasjon

Når ein utfører PCA, PLS eller andre multivariate metodar kan den resulterande dekomposisjonen bestå av mange komponentar, noko som fører til at det kan vere vanskeleg å tolke modellen. For å hjelpe til med å løyse dette problemet kan ein nytte target rotasjon eller target projeksjon (TP) [47]. Target rotasjon eller target projeksjon produserer ein enkelt prediktiv komponent ved å projisere dei dekomponerte latente variablane på responsvariabelen. Hovudføremålet med targetprojeksjon er å overvinne dei tolkingsproblema som oppstår i samanheng med den ortogonale variasjonen. Ved hjelp av desse metodane kan ein modellere kovariansen mellom dei instrumentelle variablane og responsen [48].

Selektivitetsratioen (SR) kan hjelpe til med å avgjere kva variablar som verkeleg betyr noko i ein modell. Mean Correct Classification Rate (MCCR) gir ein statistisk funnen terskelverdi for val av variabel. Dersom ein plottar MCCR mot SR får ein det som vert kalla DIVA-plott, eit nytt kvantitativt plott som kan hjelpe til med tolking og val av variabel. DIVA test og SR-plott er utvikla som kvantitative verktøy for å avsløre dei variablane som skil best mellom to grupper av prøvar. Visuelt sett er eit SR-plott likt eit spekter eller kromatogram, men der dei mest intense regionane korresponderer til dei variablane som skil gruppene best.

Prinsippet for targetrotasjon

Ein kan tenkje seg at all informasjon i eit datasett er samla i ein rådata informasjonskube. Denne kuben inneheld både nyttig og viktig informasjon, men også uønskja informasjon. Målet med targetrotasjonen er å fjerne denne uønskja informasjonen slik at ein sit igjen med ei såkalla reinsa informasjonsblokk. Føresetnaden som er knytt til targetrotasjonen er at den uønskja informasjonen er knytt til spesifikke variablar.

Det som skil targetrotasjon frå vanleg PC dekomponering er at den variabelen som er knytt til den uønskja variasjonen får endra verdi til 1,000 i ladningsvektoren for targetrotasjon. Residualet som er knytt til denne variabelen kjem då ut som 0,000 i residualmatrisa. All variasjon som er knytt til endring i den aktuelle variasjon ligg då i targetrotasjonen si modellmatrise. Residualmatrisa inneheld dermed ingen informasjon som er knytt til variansen til denne variabelen [48].

Regresjonskoeffisienten for ein responsvariabel representerer den beste tilpassa linja i det reproduserte variabelrommet for prediktorar. Prøveskårer og variabelladningar som korresponderer til rotasjon nyttar vektoren for målt respons som target, og kan finnast ved projisering av det reproduserte prediktorrommet på den normaliserte regresjonsvektoren. Dei targetroterte skårene er vist å vere proporsjonale med prediktorresponsen, medan den forklarte variansen av kvar prediktor med ein respons er proporsjonal med den kvadrerte targetroterte ladning. Skårene og ladningane oppnådd ved projisering på den normaliserte regresjonsvektoren viser den prediktive evna av ein latent variabel regresjonsmodell og viktigheita av kvar enkelt prediktor for respons. Dette kan visast samtidig i eit biplott. Biplott-presentasjon kan gi viktige teikn på korleis interferens påverkar prediksjonen og om og korleis ein kan ta omsyn til slik interferens i prediksjonen. Denne informasjonen er like viktig som avsløring og korreksjon av umodellert interferens [49].

Skårene og ladningane for targetprojeksjonsmodellen kan reknast ut frå følgjande formel:

$$\boldsymbol{X} = \boldsymbol{TP}^{'} + \boldsymbol{E} = \boldsymbol{t}_1\boldsymbol{p}_1{}^{'} + \boldsymbol{t}_2\boldsymbol{p}_2{}^{'} + \ldots + \boldsymbol{t}_i\boldsymbol{p}_i{}^{'} \qquad (2.24)$$

Targetprojeksjonsmodellen kan skrivast:

$$\boldsymbol{X} = \widehat{\boldsymbol{X}}_{TP} + \boldsymbol{E}_{TP} = \boldsymbol{t}_{TP}\boldsymbol{p}_{TP}^{'} + \boldsymbol{E}_{TP} \qquad (2.25)$$

Der TP viser til den latente variabelen som er funnen ved targetprojeksjon.

Frå likning (2.25) kan ein rekne ut forklart varians ($v_{expl,i}$) og residual varians ($v_{res,i}$) for kvar variabel i TP-modellen. Ut frå dette definerer ein selektivitetsratioen (SR) for kvar variabel som:

$$SR_i = \frac{v_{expl\,,i}}{v_{res\,,i}} \quad i = 1,2,3.. \qquad (2.26)$$

Jo høgare verdi for SR, jo betre skil variabelen mellom to grupper av prøvar. Dermed kan selektivitetsratioen nyttast til å kvantitativt rangere variablar i forhold til evne til å skilje mellom grupper. DIVA-plottet gir då ein moglegheit til å objektivt velje ein terskel for evne til å separere grupper som balanserer risikoen for å miste viktige variablar mot risikoen for å velje mange variablar som berre er eit resultat av tilfeldig korrelasjon. Frå den ikkje-parametriske DIVA-testen kan ein finne sannsynsbaserte grenser for SR-plottet [47].

3. Eksperimentelt

3.1 Simulering

I all datainnsamling må ein vurdere kor stort datamateriale ein treng i forhold til kva ressursar ein må setje inn for å skaffe seg desse data. I denne oppgåva kjem kalibreringsdatasett og testdatasett frå simuleringar og det var dermed ingen ressursmessige avgrensingar på kor store datasetta kunne vere. Likevel ville det vore lite hensiktsmessig å laga datasetta så store at ein kunne fått problem med reknekapasitet og lagringskapasitet. Kalibreringsdatasettet vart difor bestemt til å innehalde 4000 objekt og testdatasettet til å innehalde 400 objekt. Eit av objekta i kalibreringsdatasettet gav uforklarlege verdiar langt frå dei sannsynlege verdiane under simuleringa og vart etter råd frå CMR Instrumentation fjerna frå datasettet. Kalibreringsdatasettet består difor av 3999 objekt.

Kalibreringsdatasettet vart generert ved at ein valde verdiar innanfor definerte grenser for kvar variabel og kombinerte desse i ulike kombinasjonar. Øvre og nedre grenseverdi for variablane vart bestemt i samråd med CMR. Datasettet vart så sendt til CMR og simulerte verdiar for brennverdi, tettleik og lydhastigheit vart returnert. Datasettet er i utgangspunkt utan støy.

Tabell 3.1.1 Innsendte variablar i kalibreringsdatasettet

Variabel	Verdiar	Måleeining
Trykk	10, 35, 60, 80, 100, 120, 140, 160, 180, 200	Bar
Temperatur	-10, 0, 12, 24, 36 48, 60, 74, 86, 100	˚C
Metan	72.0 – 100.0	%
Etan	0.0 – 15.0	%
Propan	0 .0- 5.0	%
Butan	0.0 - 2.0	%
CO_2	0.0 - 3.0	%
N_2	0.0 – 3.0	%

Tabell 3.1.1 viser innsendte variablar i det genererte datasettet og grenseverdiane for kvar av dei. For variablane trykk og temperatur er alle nivå presiserte i tabellen.

Til å teste modellane vart det laga eit nytt datasett med punkt som ikkje skulle vere samanfallande med dei som ein hadde i kalibreringsdatasettet i tabellen ovanfor. Ein bestemte difor nye verdiar for trykk og temperatur, og desse skulle haldast innanfor dei same grenseverdiane som definert for kalibreringsdatasettet. Verdiane for objekta i datasettet er presentert i tabell 3.1.2

Tabell 3.1.2 Innsendte variablar i testdatasettet

Variabel	Verdiar	Måleeining
Trykk	22, 40, 52, 70, 90, 110, 130, 150 ,170, 190	Bar
Temperatur	-5, 3, 10, 17, 30, 40, 52, 67, 80, 92	˚C
Metan	0.72 - 1.00	%
Etan	0.0 - 15.0	%
Propan	0.0 - 5.0	%
Butan	0.0 - 2.0	%
CO_2	0.0 - 3.0	%
N_2	0.0 - 3.0	%

Det nye datasettet som er presentert i tabell 3.2.1 består av 400 objekt og vert kalla testdatasett.

3.2 Eksperimentell utføring

3.2.1 Modellering og prediksjon i Sirius

PLS modellering

Før modellering vart datasetta standardisert og sentrert i Sirius.

Det vart utført PLS-modellering av kjemisk samansetjing. Kvar kjemiske komponent vart modellert som ein funksjon av trykk, temperatur, lydhastigheit og dei resterande kjemiske komponentane. På same måte som for brennverdi og tettleik vart ulike kombinasjonar av førstegradsledd, andregradsledd og vekselverknadsledd testa ut til ein fann den beste modellen for kvar einskild kjemiske komponent.

Det vart også utført PLS-modellering i Sirius av tettleik og brennverdi ut frå kalibreringsdatasettet. Modellane vart laga med ulike kombinasjonar av førstegradsledd, andregradsledd og vekselverknadsledd for **X**, og logaritme- og kvadratrottransformasjonar av **y**.

Testing av modellar

Modellane vart testa med testdatasettet. Testdatasettet vart førebehandla på same måte som den aktuelle modellen, og dei 400 objekta vart predikert med den aktuelle modellen i Sirius.

Modellering av metan ut frå butankonsentrasjon

I tillegg til å iterere på alle dei kjemiske komponentane vart det også utført iterasjon for berre metan og etan. Dei andre kjemiske komponentane i naturgassen vart haldne utføre. Residuala for metan og butankonsentrasjonen for objekta etter iterasjon viste tydeleg fem grupperingar som igjen kunne relaterast til butankonsentrasjonen. Datasettet vart difor delt inn i fem grupper etter butankonsentrasjon og metan vart modellert på nytt for kvar gruppe. På den måten fekk ein fem lokale modellar for metan. Alle modellane består berre av førstegradsledd og har mellom fem og sju komponentar.

Prediksjon av kjemisk samansetjing

Ved å predikere den kjemiske samansetjinga i naturgassen for ein og ein komponent der den predikerte komponenten som i utgangspunktet var **y** inngjekk som ein del i **X** i prediksjon av neste kjemiske komponent ved ville ein forsøke å oppnå ei predikert samansetjing for naturgassen ut frå å kjenne berre trykk, temperatur og lydhastigheit, slik det er vist i tabell 3.2.1. Modellane for alle dei kjemiske komponentane vart laga frå kalibreringsdatasettet. Sidan metan er den bestanddelen som det er mest av i naturgass, starta ein ut med å predikere metan i testdatasettet som **y** ved å bruke modell frå kalibreringsdatasettet. For metan og etan vart det utført iterativ konsentrasjonsbestemming og dei itererte verdiane inngår i **X** ved prediksjon av dei resterande komponentane. Som startverdiar for iterasjonen nytta ein predikerte verdar for metan og etan, slik vist i tabell 3.2.1. Innsendte målte variablar var då trykk (bar) , temperatur (˚C) og lydhastigheit (m/s).

Tabell 3.2.1 Modellering og prediksjon av kjemisk samansetjing

Modellar Kalibreringsdatasett	Prediksjonar Testdatasett	Iterasjon Testdatasett
Metan=f(p,T,c)	Metan_pred=f(p,T,c)	
Metan=f(p,T,c,etan)		Metan_it=f(p,T,c, etan_pred)
Etan=f(p,T,c)	Etan_pred=f(p,T,c)	
Etan=f(p,T,c,metan)		Etan_it=f(p,T,c, metan_pred)
Propan=f(p,T,c,metan, etan)	Propan_pred=f(p,T,c,metan_it, etan_it)	
CO_2=f(p,T,c,metan,etan, propan)	CO_2_pred=f(p,T,c,metan_it, etan_it, propan_pred)	
N_2=f(p,T,c,metan,etan, propan,CO_2)	N2_pred=f(p,T,c,metan_it,etan_it, propan_pred, CO_2_pred)	
Butan=f(p,T,c,metan,etan, propan,CO_2,N_2)	Butan_pred=f(p,T,c,metan_it,etan_it, propan_pred,CO_2_pred,N_2_pred	

p er trykk (bar), T er temperatur (˚C) og c er lydhastigheit (m/s)

Objekta i testdatasettet vart predikert komponent for komponent inntil ein har predikert heile den kjemiske samansetjinga for dei 400 objekta. Deretter vert brennverdi og tettleik predikert frå den nye predikerte kjemiske samansetjinga og dei målte variablane trykk, temperatur og lydhastigheit.

Tabell 3.2.2 Modellering og revers prediksjon av kjemisk samansetjing

Modellar Kalibreringsdatasett	Prediksjonar Testdatasett
Butan=f(p,T,c)	Butan_pred=f(p,T,c)
N_2=f(p,T,c,butan)	N_2_pred=f(p,T,c,butan_pred)
CO_2=f(p,T,c,butan, N_2)	CO_2_pred=f(p,T,c,butan_pred, N_2_pred)
Propan=f(p,T,c,butan,N_2,CO_2)	Propan_pred=f(p,T,c,butan_pred, N_2_pred,CO_2_pred)
Etan=f(p,T,c,butan, N2,CO_2,propan)	Etan_pred=f(p,T,c,butan_pred,N_2_pred, CO_2_pred,propan_pred)
Metan=f(p,T,c,butan, N_2,CO_2,propan, etan)	Metan_pred=f(p,T,c,butan_pred,N_2_pred, CO_2_pred,propan_pred, etan_pred)

Ein utfører også revers prediksjon av den kjemiske samansetjinga der ein startar med butan som er den komponenten med lågast konsentrasjon i naturgassen, predikerer denne som ein funksjon av trykk, temperatur og lydhastigheit, deretter N_2 som ein funksjon av trykk, temperatur, lydhastigheit og butan_pred og vidare til slutt metan som ein funksjon av trykk, temperatur, lydhastigheit, butan_pred, N_2_pred, CO_2_pred, propan_pred og etan_pred slik det er vist i tabell 3.2.2 Modellen for kvar kjemiske komponent vart valt på bakgrunn av forklart varians i y og maksimumsfeil og prediksjonsfeil for prediksjon av nye objekt. Brennverdi og tettleik vart så predikert med den nye kjemiske samansetjinga ein kom fram til.

Modellering av lydhastigheit

Det viste seg at ein truleg har behov for å kjenne den kjemiske samansetjinga i naturgassen til ein viss grad for å kunne predikere tettleik og brennverdi, og ein modellerte difor lydhastigheit ut frå variablane trykk, temperatur og kjemisk samansetjing. Det vart også laga modellar for lydhastigheit i desse områda:

Tabell 3.2.3 Oversikt over området for modellering av lydhastigheit

Namn	Temperatur, ˚C	Trykk ,bar	Komponentar i modellen
1a) Høg temp-lågt trykk	100	10	8
1b) Høg temp-høgt trykk	100	200	8
1c) Låg temp-lågt trykk	-10	10	8
1d) Låg temp-høgt trykk	-10	200	9
1e) Senterpunkt trykk-senterpunkt temp	60	100	8

Modellering av kjemisk samansetjing og tettleik og brennverdi for konstant trykk og temperatur

I senterpunktområdet, der trykket er 60 bar og temperaturen 100 ˚C (tabell 3.2.3), vart den kjemiske samansetjinga predikert ved å predikere ein og ein variabel på same måte som presentert ovanfor. I desse modellane sendte ein inn objekt frå testdatasettet med om lag like verdiar for trykk og temperatur: 90 og 110 bar og 67 ˚C. Dei predikerte verdiane av kjemisk samansetjing vart nytta som startpunkt for neste steg. Til slutt hadde ein eit datasett bestående av lydhastigheit og ei predikert kjemisk samansetjing. Denne vart nytta til å predikere brennverdi og tettleik i det aktuelle området. Sidan modellane for brennverdi og tettleik innehaldt trykk og temperatur vart simulerte verdiar for desse variablane inkludert i datasettet med predikert samansetjing. Ein forsøkte også å predikere kjemisk samansetjing med objekt henta frå kalibreringsdatasettet. Desse objekta var med i datasettet nytta til modellering, men sidan den kjemiske samansetjinga vart predikert ut frå lydhastigheita fekk ein i så måte ”nye” objekt. For å undersøke kor stor effekt trykk og temperatur har på modellane testa ein ved å predikere alle objekta frå heile datasettet, både

kalibreringsdatasett og testdatasett med modellane for senterpunktområdet. Etter å ha testa desse modellane med heile datasettet viser det seg at ein treng å kjenne trykk og temperatur i tillegg til den kjemiske samansetjinga.

Modellering av metan, etan og addert kjemisk samansetjing (aks)

Det viste seg å vere vanskeleg å utføre alternerande regresjon på heile seks likningar. Ein samla difor dei fire minste (i %-størrelse) komponentane i ein komponent kalla addert kjemisk samansetjing (aks). Ein laga nye datasett og nye modellar for dei tre kjemiske komponentane metan, etan og aks. Variablane trykk, temperatur og lydhastigheit var uendra.

Ein forsøker så å iterativt konsentrasjonsbestemme den kjemiske samansetjinga bestående av metan, etan og aks ved hjelp av AR og dei nye modellane.

3.2.2 MATLAB

Kode for alternerande regresjon

Det vart laga eit eige program i MATLAB som skulle utføre alternerande regresjon med det føremålet å predikere kjemisk samansetjing ut frå dei målte variablane trykk, temperatur og lydhastigheit. Den predikerte kjemiske samansetjinga skulle igjen nyttast til å predikere brennverdi og tettleik av naturgassen.

Sidan den kjemiske samansetjinga består av seks komponentar ville ein prøve alternerande regresjon med seks likningar. Ein sendte først inn middelverdien i det definerte området for kvart kjemiske stoff, starta med å predikere det eine stoffet, nytta denne prediksjonen i staden for middelverdien for å predikere neste komponent og gjentok denne syklusen til alle seks stoffa er predikert og ein har oppnådd konvergens i forhold til gitt terskelverdi.

Koden består av eit hovudprogram (*hovud_kjemi*) som kallar opp eit anna program (*pred_runde_test*) så mange gonger som nødvendig for å oppnå konvergens i forhold til gitt terskelverdi. Dette programmet kallar igjen opp programma *pred, oppdater* og *rsd*. I *pred* vert prediksjonen av den kjemiske komponenten utført. I *oppdater* vert andregradsledd og

vekselverknader for dei aktuelle komponentane oppdatert, slik at desse ledda blir korrekte i forhold til predikert verdi og til vidare bruk i nye prediksjonar. I *rsd* reknar ein ut RSD for kvart objekt. Dersom programmet predikerte utanfor det definerte området for komponenten måtte resultatet rettast på slik at det var innanfor dei definerte grensene.

RSD for kvart objekt vart rekna ut og talet på iterasjonar vart talt opp for kvart objekt.

Det vart også sett på om det vart forskjell i prediksjonane avhengig av innsendt startverdi.

3.3 Dataprogram

I denne oppgåva vart følgjande program brukt:

MATLAB The Language of Technical Computing, version R2007a, Copyright 1998 – 2007, The MathWorks, Inc vart brukt til å utføre alternerande regresjon for å iterativt bestemme konsentrasjonen av kjemiske komponentar i naturgass.

Sirius versjon 7.0, Copyright 1995 – 09, The Pattern Recognition System as vart nytta til modellering og prediksjon av kjemisk samansetjing, brennverdi og tettleik i naturgassen.

Microsoft Office Word 2007 vart nytta til skrivearbeidet og Microsoft Office Excel vart nytta til behandling av datasett før iterativ konsentrasjonsbestemming i MATLAB.

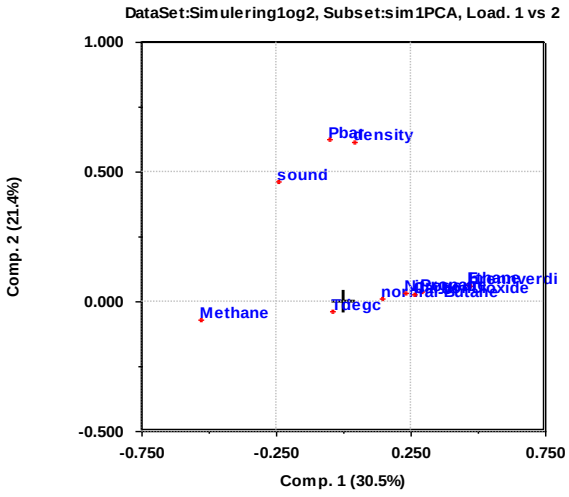
Internt simuleringsprogram, CMR vart nytta til simulering av kalibreringsdatasett og testdatasett.

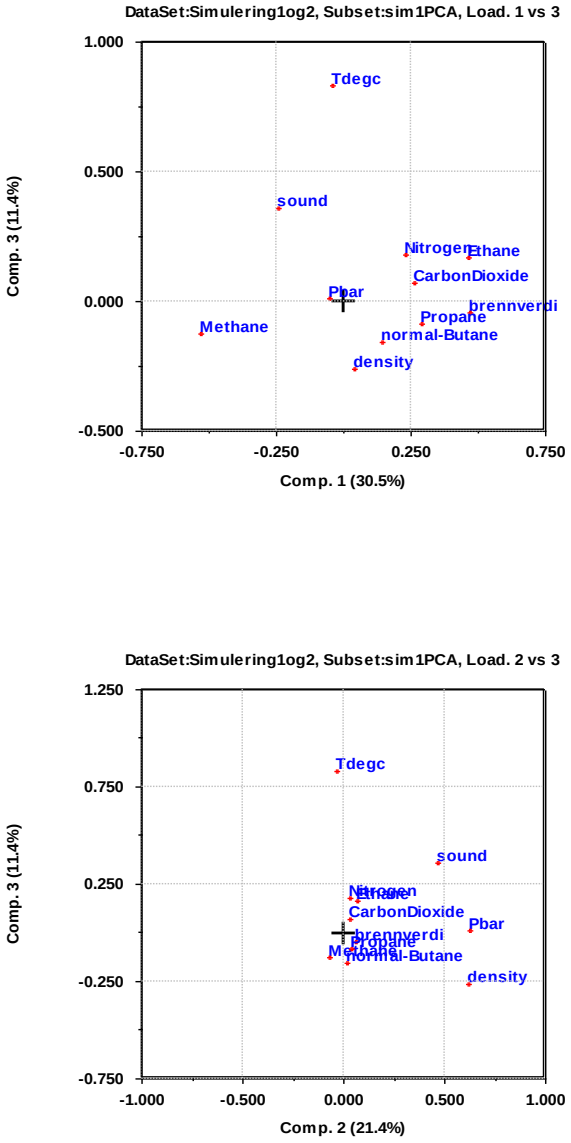
4. Resultat og diskusjon

4.1 Kjemisk samansetjing modellert og predikert frå heile datasettet

Først av alt må det undersøkast om den kjemiske samansetjinga lar seg modellere ut frå heile datasettet, altså kvar kjemiske komponent modellert som ein funksjon av trykk, temperatur, lydhastigheit og resten av den kjemiske samansetjinga. Dette er viktig fordi det vil gje signal om meir avanserte modellar også kan fungere. Ved å modellere på denne måten vil ein teste om iterativ konsentrasjonsbestemming ved bruk av alternerande regresjon lar seg utføre på seks kjemiske komponentar.

Ein startar ut med PCA for kalibreringsdatasettet for å sjå korleis variablane er korrelerte. Modellen har berre førstegradsledd inkludert, og resultatet blir ein trekomponentmodell med forklart varians i y ca 63 %. Ladningsplotta i figur 4.1.1 viser at trykk og tettleik er positivt korrelerte, tettleiken i gassen aukar altså med aukande trykk. Desse to variablane er også til ein viss grad positivt korrelerte med lydhastigheita. Metan og brennverdi er negativt korrelerte. Metan og etan er negativt korrelert, og ein ser at etan er positivt korrelert med brennverdi. Det betyr også at jo mindre metan ein har i naturgassen, jo høgare blir brennverdien og jo høgare etankonsentrasjon, jo høgare brennverdi. Trykk og temperatur er ukorrelerte.





Figur 4.1.1 Ladningsplott for dei tre komponentane generert ved PCA av kalibreringsdatasettet.

Det er bygd mange ulike modellar for kvar kjemiske komponent, berre dei beste modellane er presentert her. Oversikt over modellar finst i appendiks.

Kvar kjemiske komponent er modellert som ein funksjon av trykk, temperatur, lydhastigheit og dei resterande kjemiske komponentane. Regresjonskoeffisientane til den beste modellen for kvar kjemiske komponent er sett saman til ei B-matrise som er nytta til å utføre alternerande regresjon.

Tabell 4.1.1 Validering av modellar for dei kjemiske komponentane

Valideringsparameter	Metan	Etan	Propan	Butan	CO_2	N_2
Forklart varians i y, %	99.97	99.95	99.62	99.44	99.85	99.82
Kryssvalidering siste inkluderte komponent, CsvSD	0.509	0.741	0.984	0.954	0.105	0.391
Komponentar inkludert i modellen	4	10	13	13	6	6
RSD >	1.457	0.461	0.299	0.190	1.706	1.642
R^2	1.000	0.999	0.996	0.995	0.999	0.999
Q^2	1.000	0.999	0.996	0.995	0.999	0.999
Prediksjonsfeil	0.001	0.001	0.001	0.000	0.000	0.000
Prediksjonsfeil %	0.12	1.3	4	-	-	-

I tabell 4.1.1 er prediksjonsfeilen henta frå Sirius for den siste inkluderte komponenten i modellen. Prosent prediksjonsfeil er då rekna ut i forhold til gjennomsnittsverdien for variabelen.

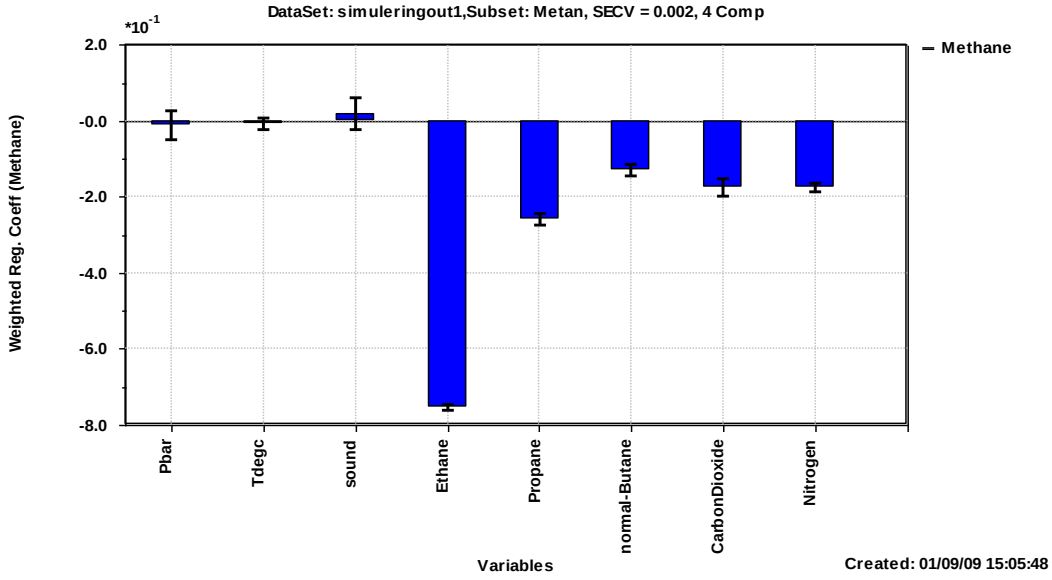
4.1.1 Metan

Metan, CH_4, er den kjemiske komponenten som det er klart mest av i naturgass. I denne oppgåva er området for metankonsentrasjon sett til 72.0 – 100.0 %.

$$Metan = f(p, T, c, etan, propan, butan, CO_2, N_2) \qquad (4.1.1)$$

Der p er trykk (bar), T er temperatur (˚C) og c er lydhastigheit (m/s).

Den beste modellen for metan inneheld førstegradsledd i **X** og ingen transformasjon av y. Fire komponentar vart inkludert i modellen. Etan er den variabelen som har det største bidraget i modellen for metan. Verken trykk, temperatur eller lydhastigheit har signifikante bidrag til modellen (figur 4.1.2). Den kjemiske samansetjinga summerer til 100 % for kvart objekt så det er ikkje overraskande at ein komponent i eit system kan predikerast frå dei andre.

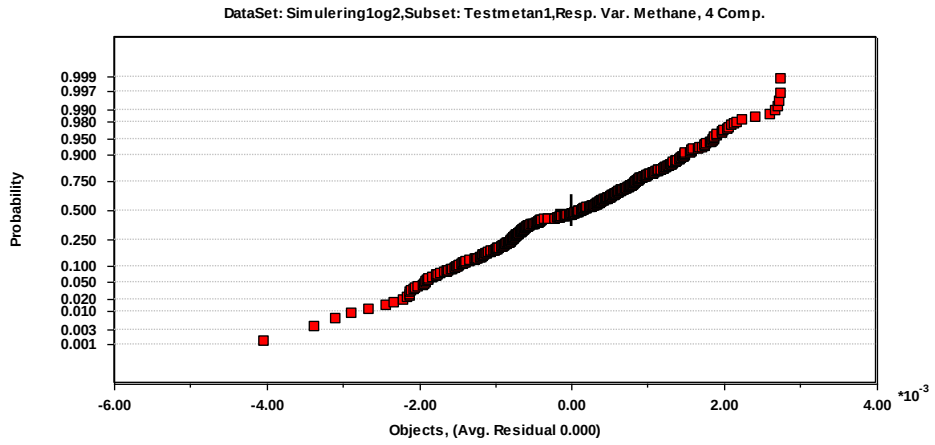


Figur 4.1.2 Grafisk framstilling av dei vekta regresjonskoeffisientane for variablane som er inkludert i modellen for metan.

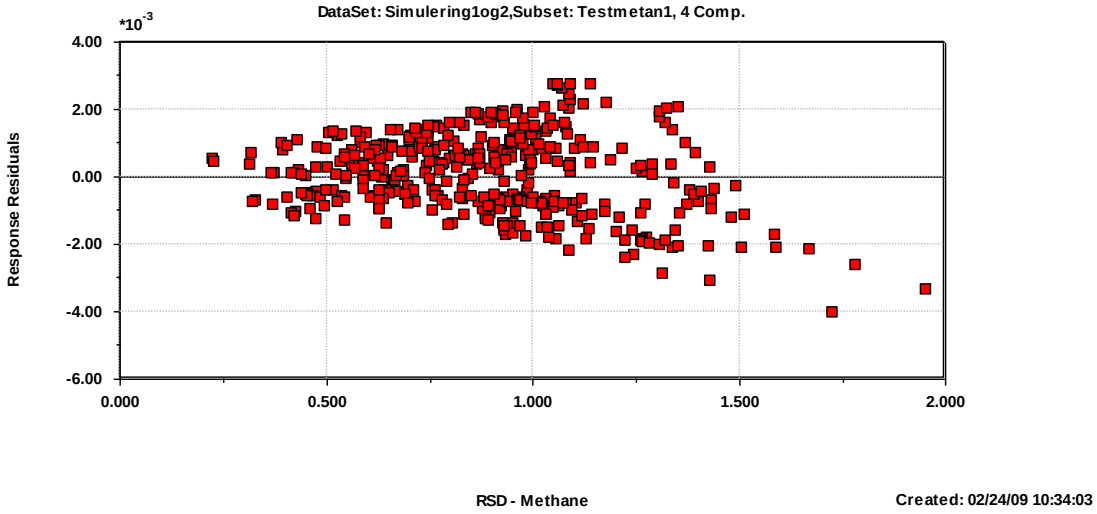
Testing og validering av modellen

Modellen har høg grad av forklart varians i y, heile 99.97 % slik det er vist i tabell 4.1.1. Standardavviket til kryssvalideringa (CsvSD) indikerer at den fjerde komponenten kan inkluderast. Responsresidual er normalfordelte, slik det er vist i figur 4.1.3. Kriteriet for om eit objekt kan avvisast som uteliggjarar er for denne modellen RSD > 1.457. Resultata frå kvart enkelte objekt frå testdatasettet viser at dei aller fleste objekta har RSD-verdiar under denne grensa. Nokre er likevel over, men den høgaste har RSD på 1.950, altså ikkje langt over grensa. Både R^2 – verdien og Q^2 – verdien er høg, og dette peikar mot at modellen er til å stole på. Prediksjonsfeilen er 0.001, noko som svarar til 0.12 % feil. Dette er bra innanfor grensa på 1 % (tabell 4.1.1). Det er ein samanheng mellom aukande responsresidual og aukande RSD. Dette er resultat som forventa sidan RSD-verdien er eit mål på kor langt frå modellen eit objekt er. Høg RSD vil altså tyde på at objektet er langt frå modellen medan eit høgt responsresidual vil tyde på dårleg prediksjon, ein samanheng mellom desse to parametrane vil altså ikkje vere uventa. Ser også at ei gruppe objekt med store negative residual har høge RSD-verdiar. (Figur 4.1.4). Dette er objekt med låg temperatur og høgt trykk.

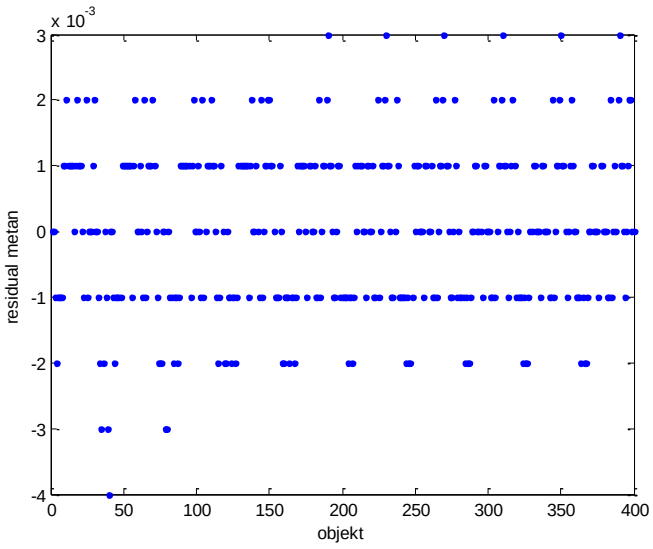
Aller størst residual finn ein for objektet med temperatur -5 ˚ C og trykk 190 bar. (Figur 4.1.5). Når ein testar modellen ser ein frå plottet i figur 4.1.6 at det er lineær samanheng mellom målte og predikerte verdiar.



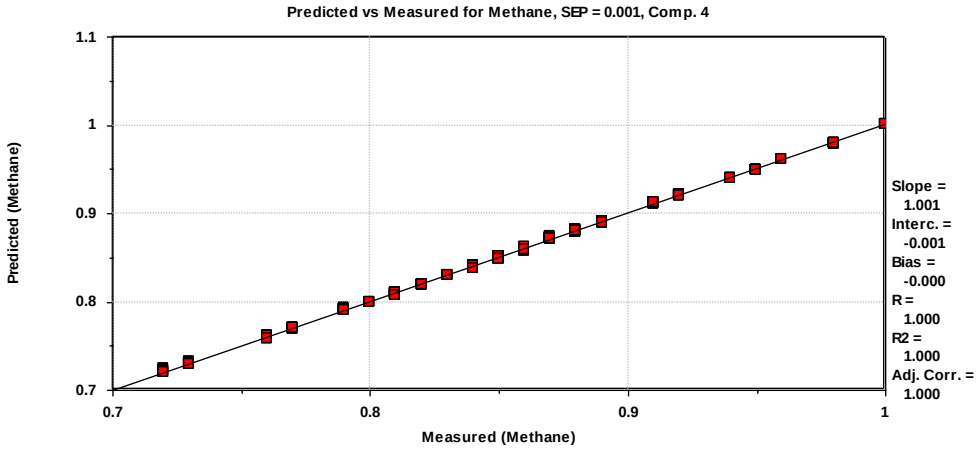
Figur 4.1.3 Normalfordelingsplott av responsresiudala for testing av gjeldande modell for metan.



Figur 4.1.4 Plott av responsresidual mot RSD av objekta i testdatasettet for gjeldande modell av metan.



Figur 4.1.5 Plott av residuala for metan etter testing med testdatasett



Figur 4.1.6 Plott av predikert verdi mot målt verdi for objekt i testdatasettet ved bruk av gjeldande modell for metan.

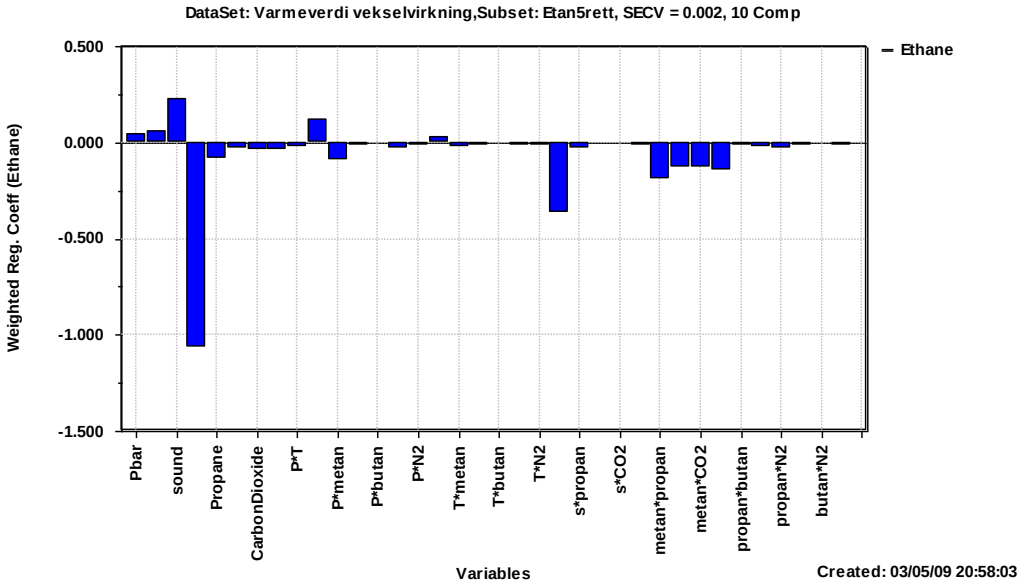
4.1.2 Etan

Etan, C_2H_6, er den nest største bestanddelen i naturgass. I denne oppgåva er området for etan definert til 0.0 – 15.0 %.

$$Etan = f(p, T, c, metan, propan, butan, CO_2, N_2) \qquad (4.1.2)$$

Der p er trykk (bar), T er temperatur (˚C) og c er lydhastigheit (m/s).

Modellen for etan inneheld førstegradsledd og vekselverknadsledd i **X** og ingen transformasjon av y. Ut frå kryssvalidering og SECV-plott valde ein å inkludere 10 komponentar i modellen for etan. Metan er den variabelen som heilt klart står for det største negative bidraget i modellen for etan, og metan står også for det største bidraget totalt. Vekselverknadsleddet lydhastigheit x metan markerer seg også med stort bidrag i negativ retning. (Figur 4.1.7)

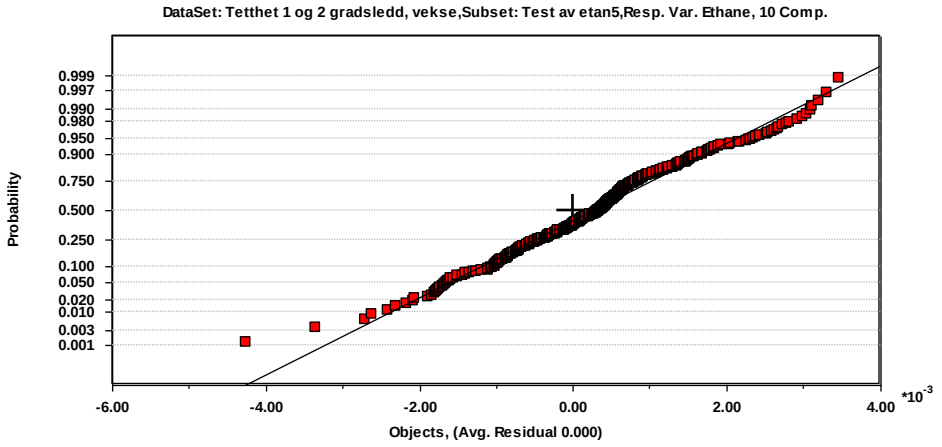


Figur 4.1.7 Plott av dei vekta regresjonskoeffisientane i modellen for etan.

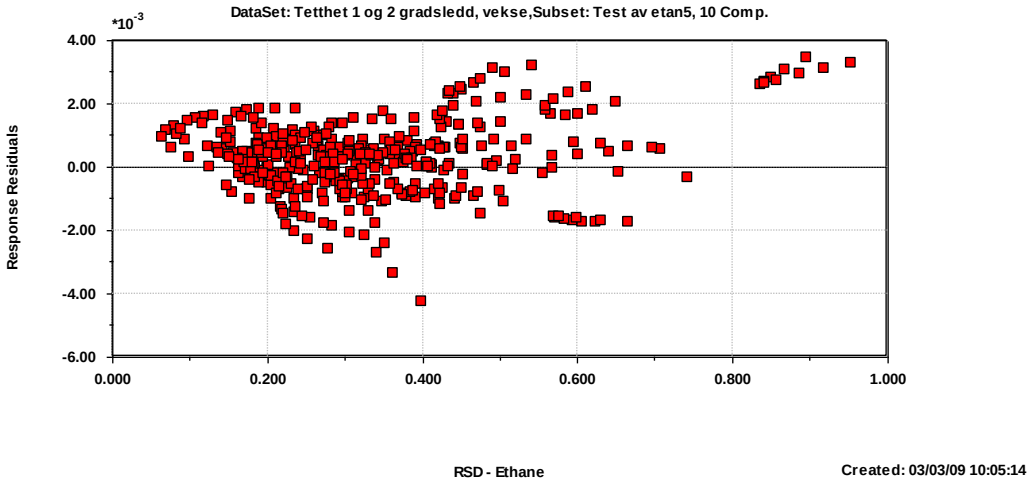
Test av modell for etan

Som ein kan sjå frå tabell 4.1.1 har modellen høg grad av forklart varians i y, heile 99.95 % av y er forklart av dei 10 komponentane som er inkludert. Responsresiduala for objekta etter testing av modellen for etan er normalfordelte slik det er vist i figur 4.1.8. CsvSD-verdien for

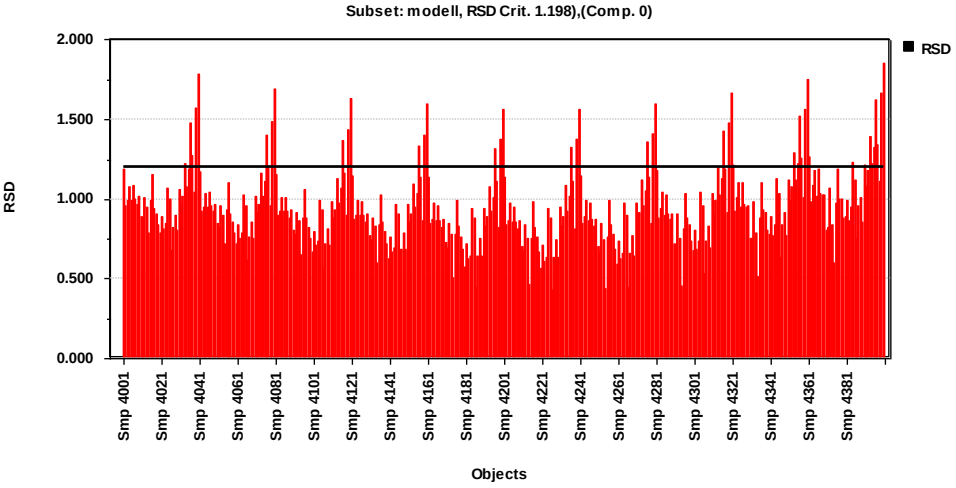
den siste komponenten peikar mot at denne kan inkluderast i modellen. Ein kan også sjå at både R^2-verdien og Q^2-verdien er 0.999 og dette peikar mot ein svært bra modell. Prediksjonsfeilen er 0.001. Dette tilsvarar 1.3 % feil, men her må ein ta omsyn til at verdiane er små og avrunding vil få store konsekvensar for utrekning av prosent feil. Avvisningskriteriet for uteliggjarar er RSD > 0.461. Resultata frå testinga viser at heile 69 av objekta har RSD over grensa for uteliggjarar. Dei fleste ligg likevel rett over, den høgaste er ca 0.9. Det er tydeleg samanheng mellom aukande responsresiudal og aukande RSD. Det er størst tettleik av objekt med RSD-verdi 0.2 – 0.4 og responsresidual - 0.1 – 0.1. Det ligg også ei gruppe objekt for seg sjølv oppe til venstre i plottet. (Figur 4.1.9). Felles for desse objekta er at det har trykk 190 bar, når det gjeld temperatur varierer dei frå -5 ˚C til 92 ˚C. Det er sjølvsagt mange fleire objekt tilstades som har trykk 190 bar. Plottet i figur 4.1.10 viser at det er objekta med høgast trykk som har høgast responsresidual og dei aller høgaste finn ein når høgt trykk er kombinert med enten dei lågaste temperaturane eller dei høgaste temperaturane. Det kan sjå ut som om objekta i midten av kvar gruppe har lågast RSD, her finn ein altså dei objekta med trykk rundt 100 bar og temperatur rundt 50 ˚C. Det er god samanheng mellom predikert og målt verdi for modellen for etan. (Figur 4.1.11).



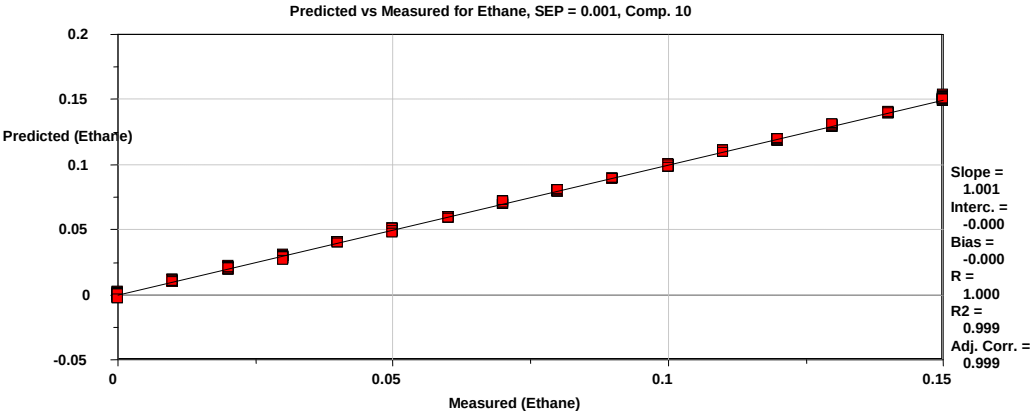
Figur 4.1.8 Normalfordelingsplott av responsresiduala ved testing av gjeldande modell for etan.



Figur 4.1.9 Plott av RSD mot responsresidual etter testing av gjeldande modell for etan.



Figur 4.1.10 Plott av RSD for objekta i testdatasettet



Figur 4.1.11 Plott av predikert verdi mot målt verdi for gjeldande modell for etan.

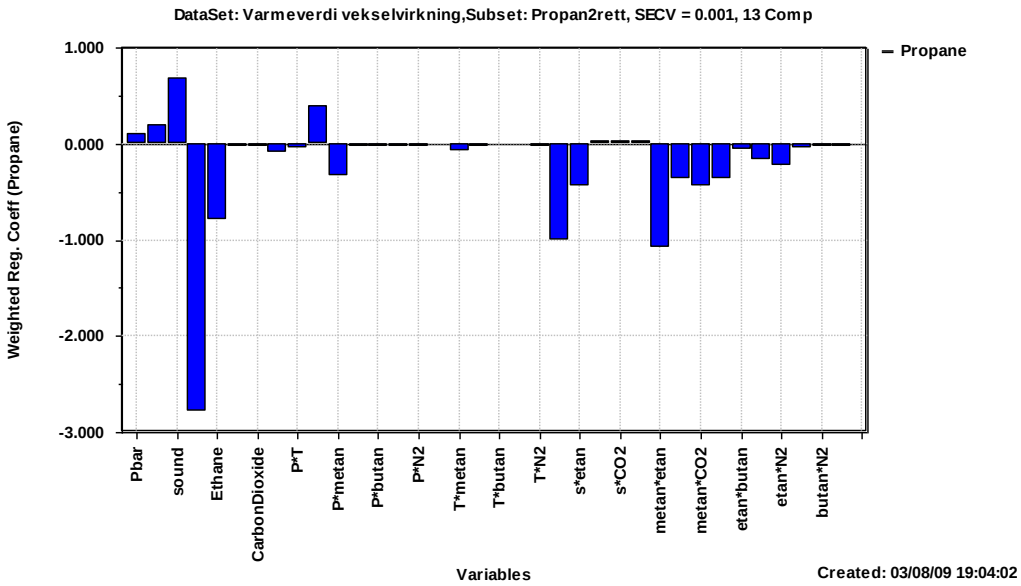
4.1.3 Propan

I denne oppgåva er propan, C_3H_8, definert til å utgjere mellom 0.0 og 5.0 % av den kjemiske samansetjinga i naturgass.

$$Propan = f(p, T, c, metan, etan, butan, CO_2, N_2) \qquad (4.1.3)$$

Der p er trykk (bar), T er temperatur (˚C) og c er lydhastigheit (m/s).

Modellen for propan består av førstegradsledd og vekselverknadsledd i **X** og ingen transformasjon av y. Modellen består av heile 13 komponentar. Metan er den variabelen som har det dominerande bidraget i denne modellen. Lydhastigheit har eit større bidrag i modellen for propan enn det ein har sett i modellane for metan og etan. Vekselverknadsleddet metan x etan har det nest største bidraget til modellen. (Figur 4.1.12).

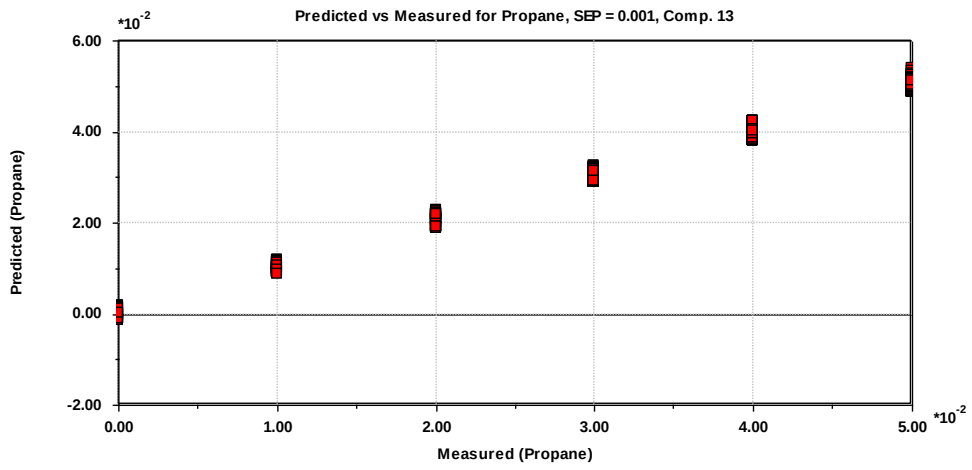


Figur 4.1.12 Plott av vekta regresjonskoeffisientar for variablane i modellen for propan.

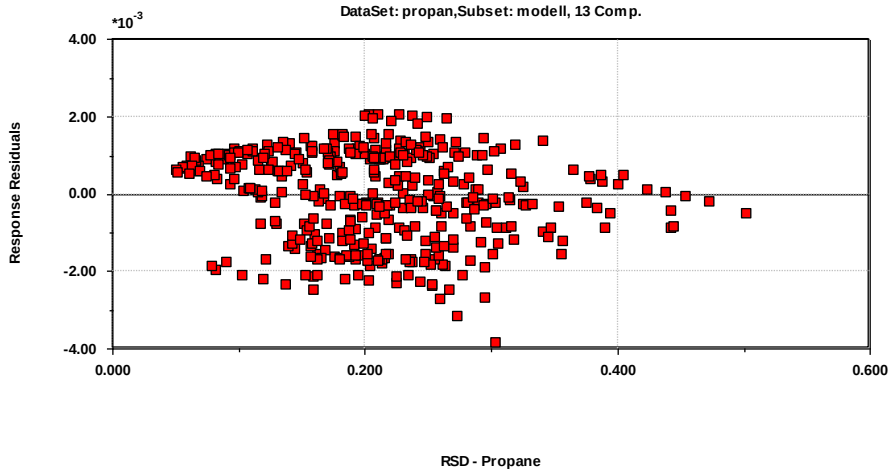
Test av modell for propan

Det er bra samanheng mellom målt og predikert verdi for propan i denne modellen. Dei ulike nivåa for propan blir predikert med om lag like store residual. Propanverdien 0.05 har noko overvekt av for høge prediksjonar slik ein kan sjå frå plottet i figur 4.1.13. I tabell 4.1.1 kan ein sjå at graden av forklart varians i y i modellen for propan er 99.62 %. Prediksjonsfeilen er 0.001, dette tilsvarar 4 %, men igjen er det små verdiar som gir store utslag når det blir rekna om til prosent. Den 13. komponenten har CsvSD-verdi 0.984, noko som indikerer at

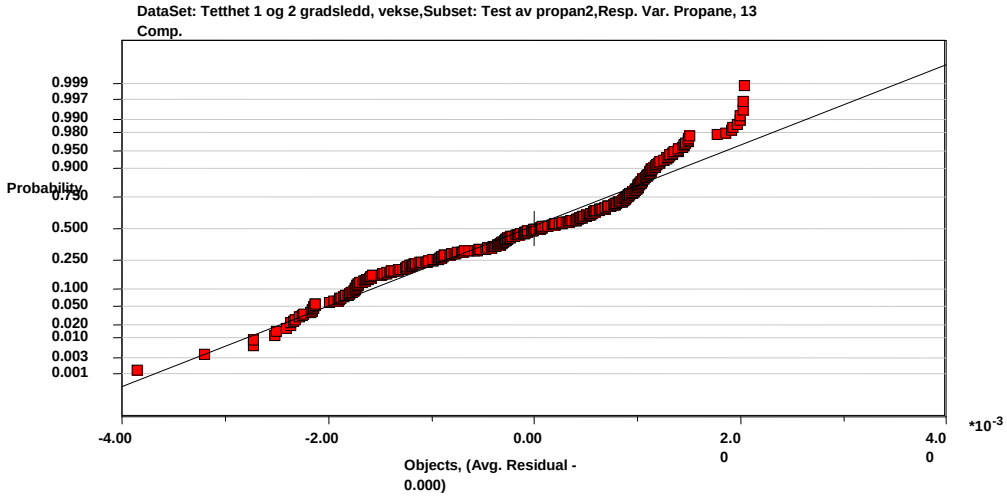
komponenten kan inkluderast i modellen. R^2-verdien er 0.996, dette er ein høg verdi som også kan tyde på at modellen er tilfredsstillande. I og med at denne modellen for etan har heile 13 komponentar inkludert kan den høge R^2-verdien også tyde på overtilpassing av modellen. Datasettet som ein har nytta i denne oppgåva er simulert og tilsynelatande utan støy og dermed vil alle bidrag til modellen vere signifikante. Q^2 –verdien er også 0.996 og sidan dette er eit uttrykk for intern prediksjonsevne i modellen vil dette peike mot at modellen er bra. Det er samanheng mellom auke i responsresidual og auke i RSD (figur 4.1.14) , dette indikerer at dei objekta med størst residual i **X** og så har størst responsresidual i y. Variasjonen i responsresidual er størst for dei objekta som har RSD-verdi mellom 0.200 og 0.300. I dette området finn ein også dei største residuala. For høgare RSD-verdiar ser ein at ein har både færre objekt, mindre variasjon i residuala og lågare residual. Det er interessant å merke seg at dei objekta som har aller høgast RSD-verdi har residual i størrelsesorden 0. Avvisningskriteriet for uteliggjarar i denne modellen er RSD > 0.299. Av dei 400 objekta i testdatasettet er det 46 objekt som har RSD-verdi over denne grensa. Dei fleste objekta ligg likevel like over, det objektet med høgast RSD-verdi har RSD=0.503. Normalfordelingsplottet i figur 4.1.15 viser at responsresiduala etter testing av modellen for propan er relativt bra normalfordelte. Ei lita gruppe objekt oppe til høgre i plottet har noko lågre responsresidual enn venta.



Figur 4.1.13 Plott av predikert verdi mot målt verdi for propan etter testing med testdatasett.



Figur 4.1.14 Plott av responsresidual mot RSD.



Figur 4.1.15 Normalfordelingsplott av responsresiudala etter tesing av gjeldande modell for propan.

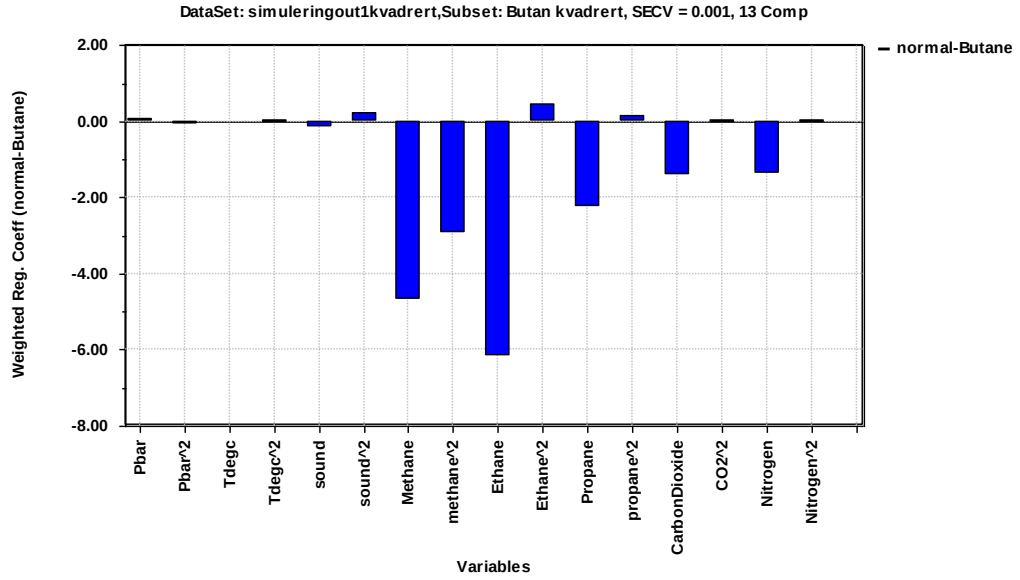
4.1.4 Butan

Butan, C_4H_{10}, utgjer 0.0 -2.0 % av den kjemiske samansetjinga i naturgass i denne oppgåva.

$$Butan = f(p, T, c, metan, etan, propan, CO_2, N_2) \qquad (4.1.4)$$

Der p er trykk (bar), T er temperatur (˚C) og c er lydhastigheit (m/s).

Den beste modellen for butan inneheld førstegradsledd og andregradsledd i **X** og ingen transformasjon av y. 13 komponentar er inkludert i denne modellen. Det er variablane etan, metan og kvadrert metan som har mest å bety for modellen. I motsetning til i modellen for propan har etan eit større bidrag enn metan i modellen for butan. Variablane trykk, temperatur og lydhastigheit har svært små bidrag til denne modellen. (Figur 4.1.16).

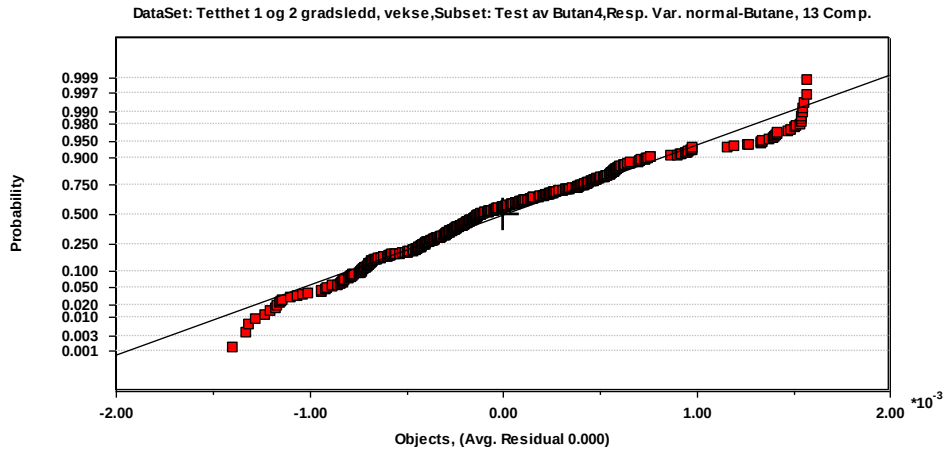


Figur 4.1.16 Grafisk framstilling av vekta regresjonskoeffisientane for variablane i modellen for butan

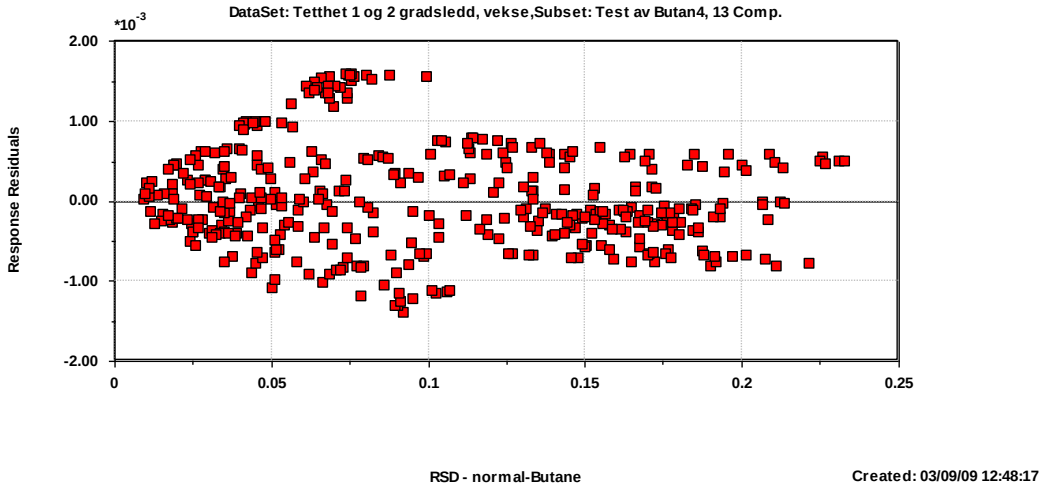
Testing av modellen

Som ein kan sjå i tabell 4.1.1 er graden av forklart varians i y 99.44 %. CsvSD for den siste komponenten som er inkludert er 0.954, noko som indikerer at det er fornuftig å inkludere denne komponenten. Prediksjonsfeilen for modellen vert i Sirius gitt som 0.000 for den siste inkluderte komponenten. R^2-verdien er 0.954 som tyder på at modellen er svært god, særleg når ein kombinerer det med Q^2 som er 0.995 for modellen. Responsresiudala for objekta er relativt bra normalfordelt etter testing av modellen, dette er vist i figur 4.1.17. Ei gruppe

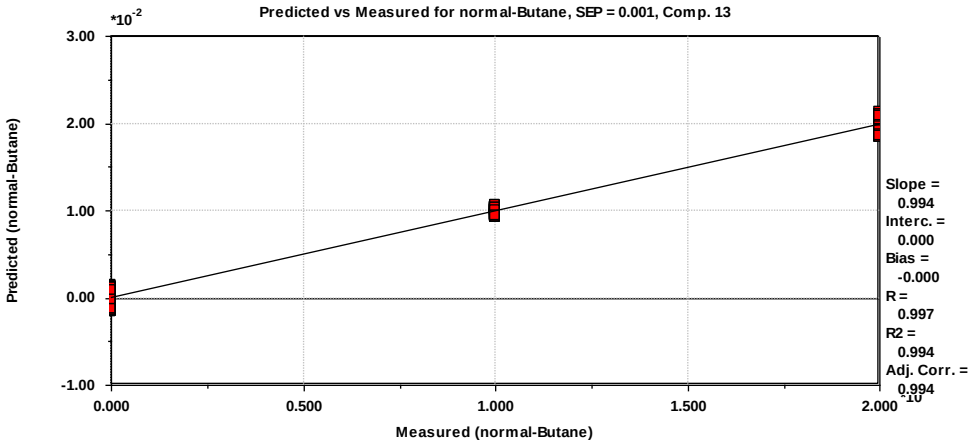
residual oppe til høgre i plottet har noko høgare responsresidual enn forventa ut frå normalfordelingsprinsippet. Også for butan er det samanheng mellom aukande responsresidual og aukande RSD, i alle fall til ein kjem opp mot RSD-verdiar på ca 0.1. (Figur 4.1.18). For høgare RSD-verdiar er det ingen er samanheng mellom høge responsresidual og RSD. Avvisningskriteriet for uteliggjarar er RSD > 0.190. 29 av dei 400 objekta i testdatasettet for denne modellen har RSD-verdi over denne grensa. Objektet som har høgast RSD-verdi har verdien 0.231. Det er tilfredsstillande samanheng mellom målt og predikert verdi for butan etter testing med objekta i testdatasettet. Samanhengen mellom målt og predikert verdi er best for målt verdi 0.01, altså for konsentrasjonen 1 %. For 0.00 og 0.02 er det litt større spreiing i resultata. Ein hadde berre tre nivå for butan i datasettet. (Figur 4.1.19).



Figur 4.1.17 Normalplott av responsresiduala etter testing av modell for butan.



Figur 4.1.18 Plott av responsresidual mot RSD for alle objekta i testdatasettet.



Figur 4.1.19 Plott av målt verdi mot predikert verdi etter testing av modellen for butan.

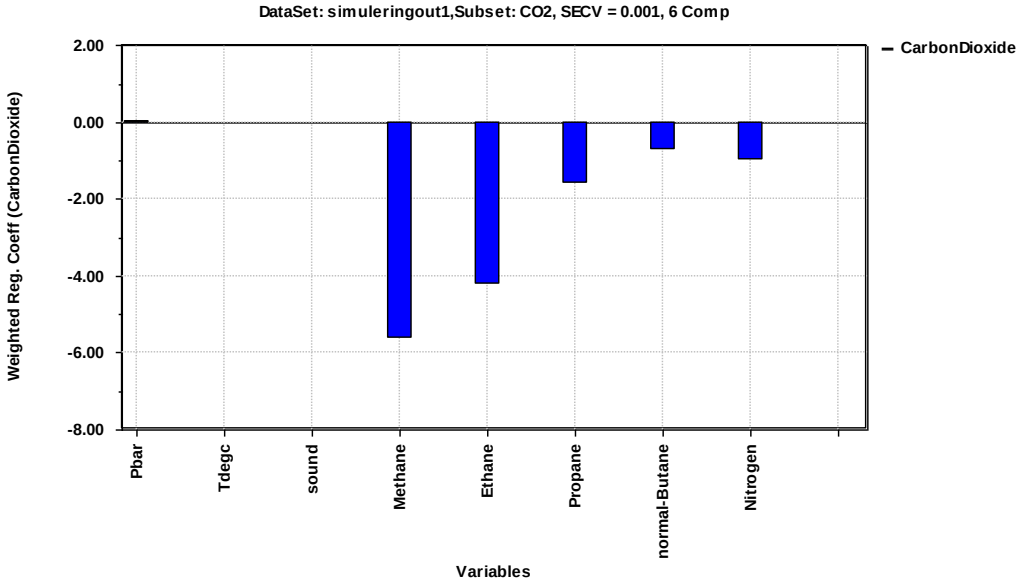
4.1.5 CO_2

I denne oppgåva er CO_2-konsentrasjonen definert i området 0.0 – 3.0 % av naturgassen.

$$CO_2 = f(p, T, c, metan, etan, propan, butan, N_2) \qquad (4.1.5)$$

Der p er trykk (bar), T er temperatur (˚C) og c er lydhastigheit (m/s).

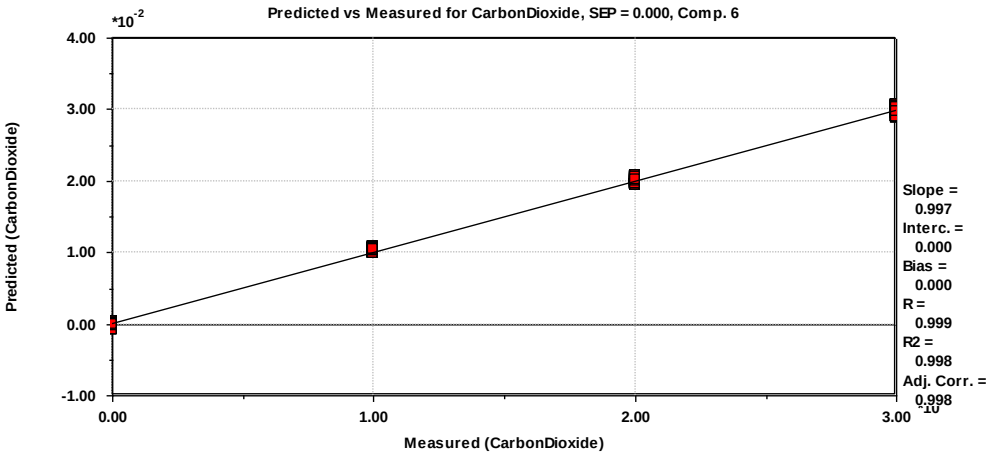
Modellen for CO_2 består av førstegradsledd i **X** og ingen transformasjon av y. Modellen inneheld seks komponentar. Metan er den variabelen som har det største bidraget i modellen for CO_2. Etan fylgjer etter og har det nest største bidraget. Alle dei kjemiske komponentane trekker i negativ retning i modellen. Trykk, temperatur og lydhastigheit bidrar ikkje til modellen. (Figur 4.1.20).



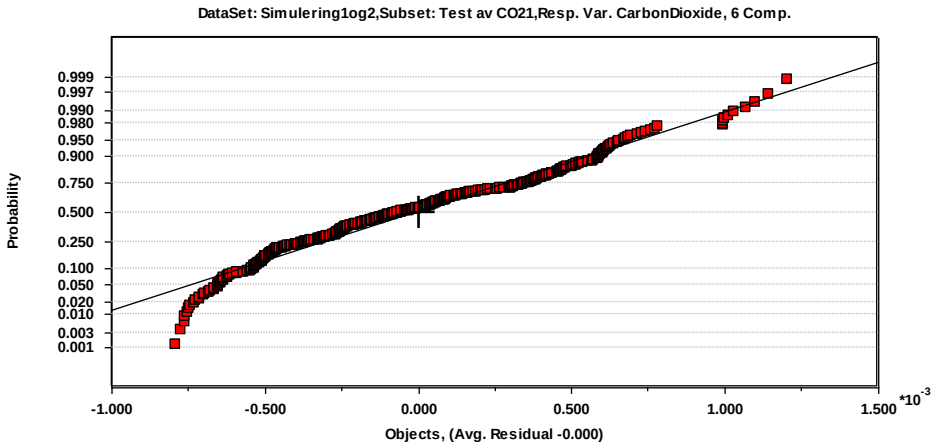
Figur 4.1.20 Grafisk framstilling av vekta regresjonskoeffisientar for variablane i modellen for CO_2.

Test av modellen

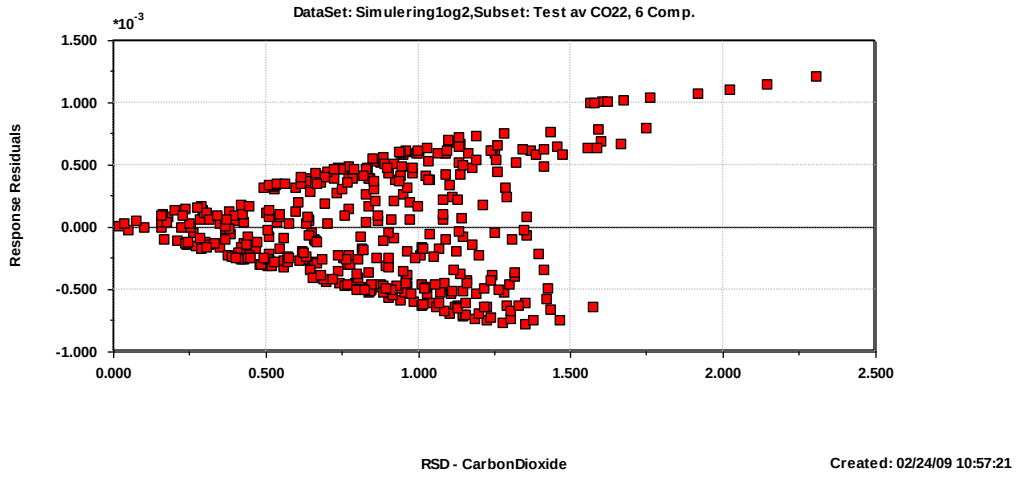
Det er høg grad av samanheng mellom predikert og målt verdi for CO_2. Den målte verdien 0.01 ser ut til å ha dei mest presise prediksjonane slik ein ser i figur 4.1.21. Responsresiduala for objekta etter testing er relativt normalfordelte, men ein kan sjå at det ligg ei lita gruppe objekt for seg sjølv oppe til høgre i plottet. (Figur 4.1.22). Dette kan tyde på at det her dreier seg om to fordelingar. Tabell 4.1.1 viser at forklart varians i y for modellen er høg, 99.85%. R^2 og Q^2 har begge verdiane 0.999, noko som peikar mot at dette er ein tilfredsstillande modell. Prediksjonsfeil for modellen er 0.000. CsvSD-verdien er 0.105 for den sjette komponenten i modellen, og denne verdien indikerer at det var riktig å inkludere komponenten i modellen. Det er tydeleg samanheng mellom auke i responsresidual og auke i RSD for objekta. (Figur 4.1.23).Det ligg også nokre objekt oppe til høgre i plottet, desse har både høge RSD-verdiar og høge responsresidual. Avvisningskriteriet for uteliggarar i denne modellen er RSD > 1.706. Det er berre seks objekt som har høgare RSD-verdiar enn denne grensa, det objektet som har høgast RSD-verdi har verdien 2.148. Fem av desse objekta er dei same som beskrive i kapittel 4.1.2 Etan. Det sjette objektet derimot, har trykk 40 bar. Det er det siste objektet i kvar gruppe, altså det objektet med høgast trykk som har mykje høgare RSD enn dei andre objekta i gruppa. Felles for desse fem objekta er altså trykk 190 bar, temperaturar mellom -5 og 30 ˚ C og alle har dei metankonsentrasjon 0.87. Det kan altså sjå ut som om objekt som har høgt trykk og kombinasjon med låg temperatur og metankonsentrasjon på 87 % som skapar problem for prediksjonen.



Figur 4.1.21 Plott av målt verdi mot predikert verdi etter testing av modell for CO_2.



Figur 4.1.22 Normalplott av responsresiduala etter testing av modell for CO_2



Figur 4.1.23 Plott av responsresidual mot RSD etter testing av modell for CO_2

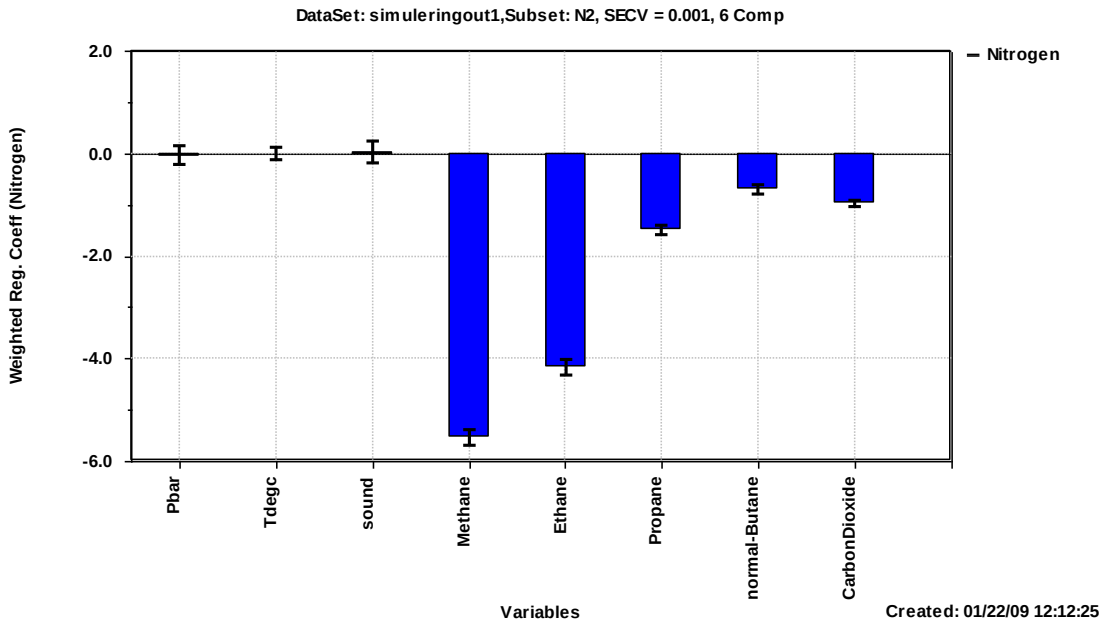
4.1.6 N_2

I denne oppgåva utgjer N_2 0.0 -3.0 % av naturgassen.

$$N_2 = f(p, T, c, metan, etan, propan, butan, CO_2) \qquad (4.1.6)$$

Der p er trykk (bar), T er temperatur (˚C) og c er lydhastigheit (m/s).

Modellen for N_2 har førstegradsledd i **X** og ingen transformasjon av y. Seks komponentar er inkludert i modellen for N_2. Metan og etan er dei variablane som har størst bidrag til modellen. (Figur 4.1.24).

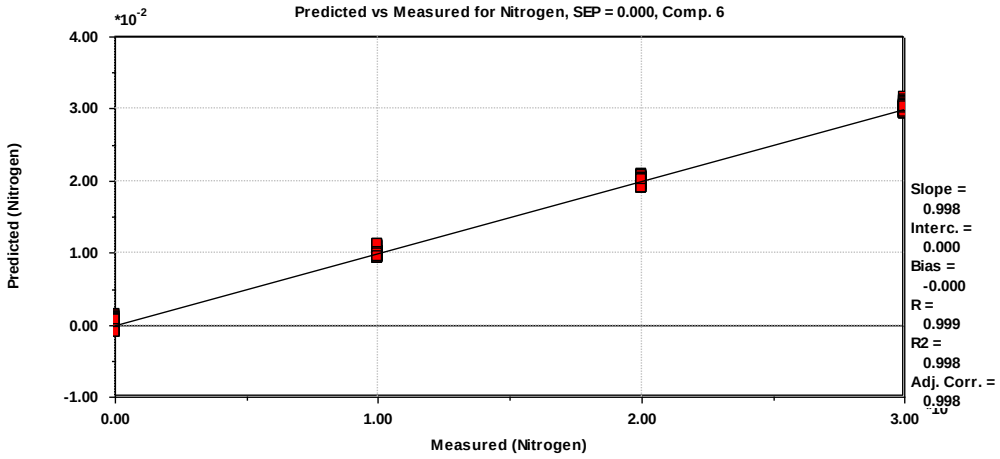


Figur 4.1.24 Grafisk framstilling av vekta regresjonskoeffisientene i modell for N_2.

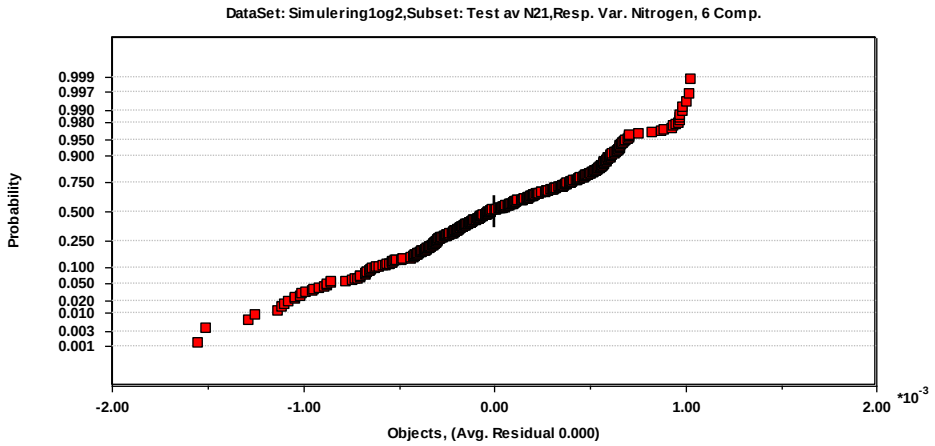
Test av modellen

Det er høg grad av samanheng mellom målte og predikerte verdiar etter testing med nye objekt. Det er større variasjon i prediksjonar på den positive sida av den målte verdien enn på den negative sida. (Figur 4.1.25). Figur 4.1.26 viser at responsresiduala etter testing av modellen er relativt bra normalfordelte. Tabell 4.1.1 viser at graden av forklart varians y i modellen er 99.82 %, dette er tilfredsstillande grad av forklart varians i y. Kryssvalideringa indikerer også at den sjette komponenten kan inkluderast med ein CsvSD-verdi på 0.391. Prediksjonsfeilen er 0.000. Når det gjeld kumulativ forklart varians, R^2 og intern prediktiv evne for modellen, Q^2, har begge desse parametrane verdien 0.999 som er svært høgt, dette

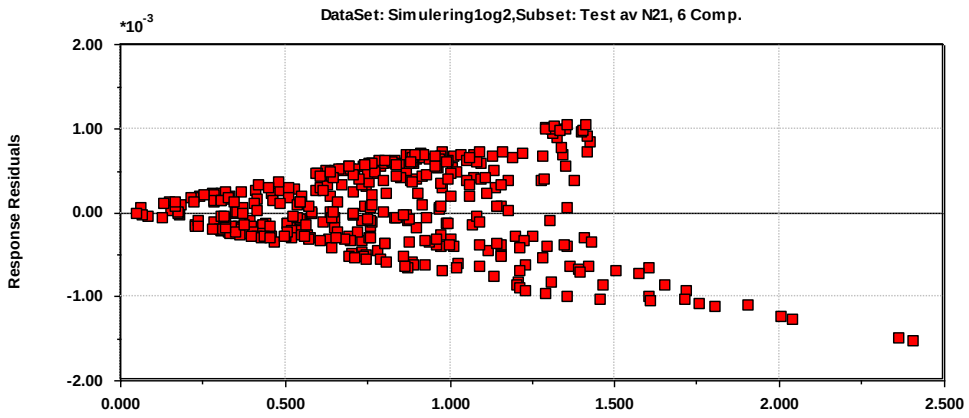
peikar mot at dette er ein bra modell. Også for N_2 er det ein klar samanheng mellom aukande responsresidual og aukande RSD. For CO_2 finnes ei gruppe objekt med store positive residual og høg RSD, men her for N_2 ser ein ei gruppe med store negative residual og høg RSD (Figur 4.1.27). Dersom ein ser bort frå desse objekta ser det ut som det er om lag lik fordeling mellom positive og negative residual sett i samanheng med størrelse på RSD-verdien. Avvisningskriteriet for uteliggjarar i denne modellen er RSD > 1.642. 10 objekt har RSD-verdi over denne grensa, dei fleste objekta ligg like over grenseverdien. Det objektet med høgast RSD-verdi har RSD = 2.410. Blant desse 10 objekta har åtte av dei trykk på 170 - 190 bar og temperaturar mellom -5 og 10 ˚C. Metankonsentrasjonen varierer mellom 72 % og 87 %. Også her ser det altså ut som om det er dei objekta med høgt trykk og låg temperatur som skapar utfordringar for prediksjonen.



Figur 4.1.25 Plott av predikert verdi mot målt verdi av modellen for N_2



Figur 4.1.26 Normalplott av responsresiduala etter testig av modell for N_2.



Figur 4.1.27 Plott av responsresidual mot RSD etter testing av modell for N_2

Alle dei kjemiske komponentane i naturgassen lar seg modellere tilfredsstillande frå variablane trykk, temperatur, lydhastigheit og resten av den kjemiske samansetjinga. Modellane får høg forklart varians i y og har tilfredsstillande prediktiv evne slik ein kan sjå frå tabell 4.1.1. Modellane skal nyttast til iterativ konsentrasjonsbestemming ved bruk av alternerande regresjon.

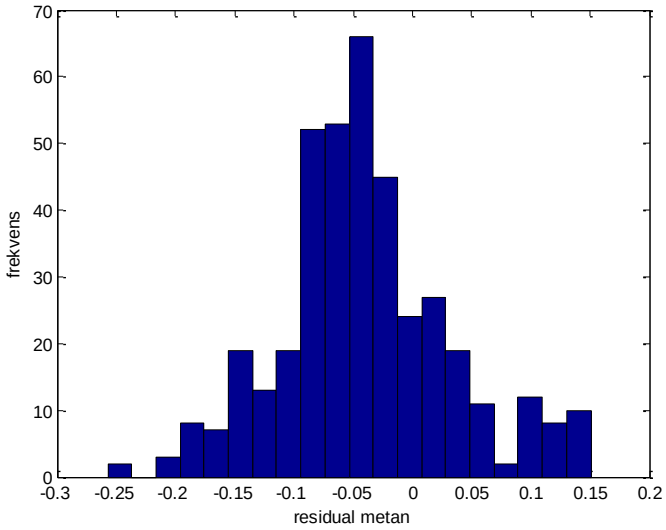
4.1.7 Iterativ konsentrasjonsbestemming av kjemisk samansetjing

Ein utfører no iterativ konsentrasjonsbestemming ved AR med regresjonskoeffisientane for dei kjemiske komponentane som er funne i modellane presentert i kapittel 4.1. Metan er predikert først, deretter dei kjemiske komponentane i denne rekkjefølgja: etan, propan, CO_2, N_2 og butan.

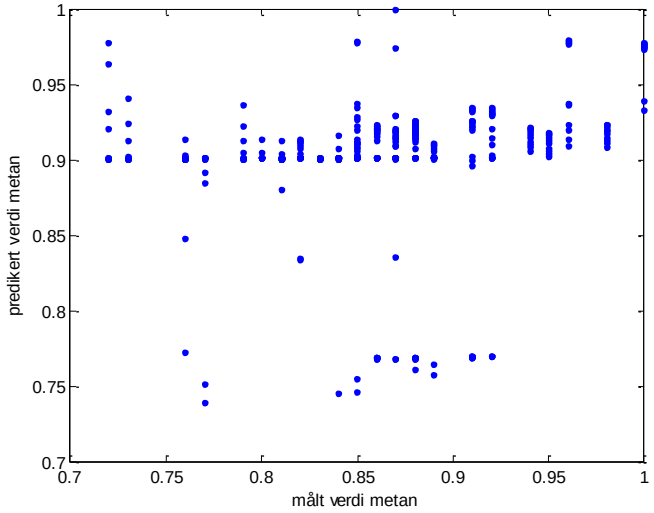
Tabell 4.1.2 Resultat av prediksjon av kjemisk samansetjing ved iterativ konsentrasjonsbestemming

Kjemisk komponent	Maksimumsfeil negativ retning	Maksimumsfeil positiv retning	Prediksjonsfeil Absolutt verdi	Prediksjonsfeil, %
Metan	- 0.257	0.151	0.070	8.1
Etan	- 0.130	0.150	0.078	101.3
Propan	- 0.050	0.050	0.021	84.0
CO_2	- 0.030	0.030	0.019	118.8
N_2	- 0.030	0.030	0.013	86.7
Butan	- 0.020	0.020	0.009	90.0

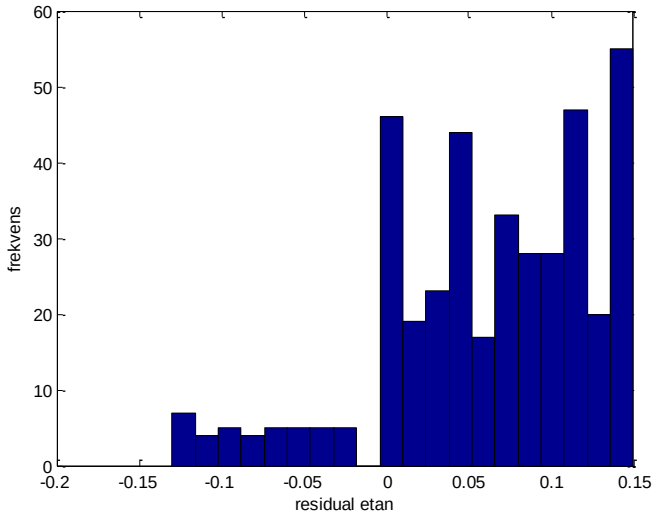
Som ein kan sjå i tabell 4.1.2 er det store skilnader når det gjeld kor godt dei ulike kjemiske komponentane vert predikert når ein nyttar alternerande regresjon. Metan er den kjemiske komponenten som vert predikert best. Metan er den klart største bestanddelen i naturgass og difor vil det vere viktigast med god prediksjon for denne. Etan har svært stor prediksjonsfeil. Figur 4.1.28 viser histogram av residuala for metanprediksjonen og ein kan sjå at det er overvekt av negative residual, plottet har topp ved ca -0.05. Ein ser av plottet i figur 4.1.29 at det er ingen samanheng mellom målt og predikert verdi for metan. Predikert verdi for etan har stor overvekt av positive residual slik ein kan sjå av plottet i figur 4.1.30. Predikert verdi for propan har ein overvekt av negative residual (figur 4.1.31). For predikert verdi av CO_2 ser ein overvekt av positive residual, og når ein framstiller residuala i histogram ser ein fem toppar. Det same ser ein også for residuala for N_2prediksjonen, men her er residuala i all hovudsak negative. For predikert verdi for butan ser ein tydeleg fem toppar, og her er også dei fleste residuala negative. (Appendiks, figurane A.1.1 – A.1.3). Dette kan tyde på at det fem fordelingar tilstades. Propan er den einaste variabelen som har fem nivå men denne inndelinga har ikkje nokon samanheng med nivå av propan. Det er ingen årsaker som utpeikar seg i forhold til denne inndelinga.



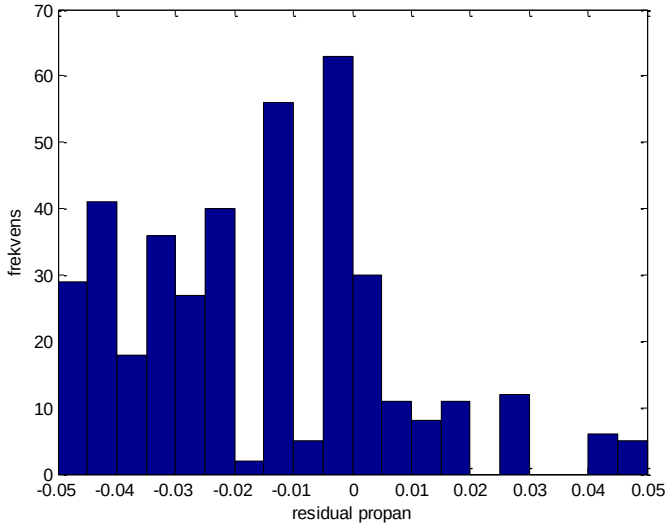
Figur 4.1.28 Histogram av residuala for metan etter iterativ konsentrasjonsbestemming ved AR.



Figur 4.1.29 Plott av predikert verdi mot målt verdi for metan



Figur 4.1.30 Histogram av residuala for etan etter iterativ konsentrasjonsbestemming ved AR.



Figur 4.1.31 Histogram av residuala for propan etter iterativ konsentrasjonsbestemming ved AR.

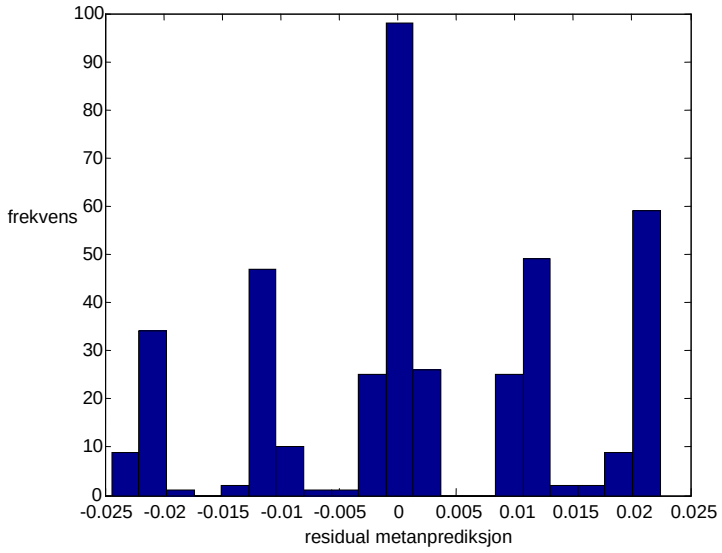
Iterativ konsentrasjonsbestemming av metan og butan

Sidan iterativ konsentrasjonbestemming ved AR av alle dei seks kjemiske komponentane samtidig ikkje såg ut til å gi tilfredsstillande resultat forsøker også å iterativt predikere berre metan og butan. Sender inn middelverdiane i dei definerte områda for metan (0.85) og butan (0.01). For dei resterande variablane er det ikkje gjort endringar i forhold til originalt kalibreringssett og desse variablane er heller ikkje tatt med i prediksjonen. Metan er predikert først, deretter butan.

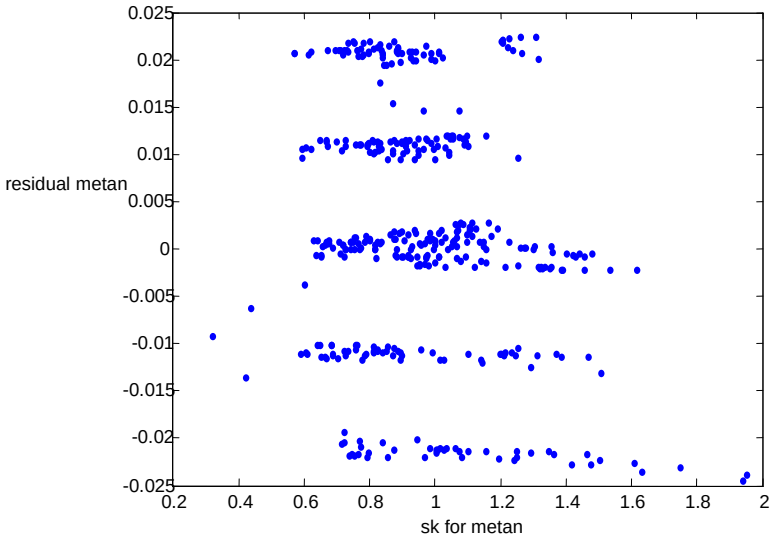
Tabell 4.1.3Maksimumsfeil i positiv og negativ retning for prediksjon av metan og butan for objekta i kalibreringsdatasettet og objekta i testdatasettet.

	Datasett	Kalibreringsdatasett	Testdatasett
Metan	Maksimumsfeil positiv retning	0.023	0.022
	Maksimumsfeil negativ retning	-0.025	-0.022
	Maksimumsfeil, %	2.7	2.6
Butan	Maksimumsfeil positiv retning	0.02	0.02
	Maksimumsfeil negativ retning	-0.02	-0.02
	Maksimumsfeil, %	200	200

Som ein kan sjå frå tabell 4.1.3 er det ingen stor endring i maksimumsfeilen for metan avhengig av om ein predikerer på kalibreringsdatasettet eller om ein predikerer på nye objekt. Det er litt overraskande i og med at det ville vore naturleg med mindre feil for dei objekta modellen faktisk var bygd av. Maksimumsfeilen i prosent er rekna ut frå absoluttverdien av maksimumsfeilen og viser at det er liten skilnad mellom feil i prediksjonen for data henta frå testdatasettet og data henta frå kalibreringsdatasettet. Histogrammet i figur 4.1.32 viser at det etter prediksjon er fem grupper residual for metan tilstades. Metan har konsentrasjonar mellom 72 og 100 % og mange fleire nivå enn fem. Det er ingen opplagte årsaker til at ein skal få denne fordelinga av residual. Som ein kan sjå ut frå figur 4.1.33 ser ein tydeleg fem grupper også når ein plottar residual mot RSD for objekta og det er objekta med dei største negative residuala som har høgast RSD-verdi.



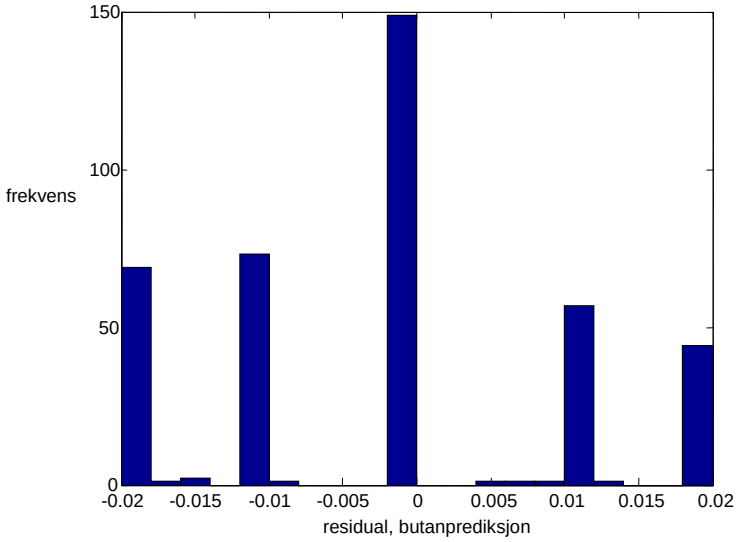
Figur 4.1.32 Histogram over residuala for dei predikerte metankonsentrasjonane for objekta i testdatasettet.



Figur 4.1.33Plott av RSD-verdiar mot residual for metan for objekta frå testdatasettet.

Tabell 4.1.8 viser at maksimumsfeilen for butan er svært stor både i positiv og negativ retning, heile 200 %. Sidan butan utgjer ein svært liten bestanddel av naturgassen, 0.0 -3.0 %, kan stor prediksjonsfeil for denne kjemiske komponenten vere mindre viktig for vidare prediksjon av tettleik og brennverdi. Små prediksjonsfeil vil også få store utslag når

ein reknar dei om til prosent. For butanprediksjonen ser ein også tydeleg at residuala fordeler seg på fem grupper i histogrammet i figur 4.1.34. Butan har konsentrasjonar mellom 0.0 og 2.0 % og fordeler seg på tre nivå. Det er heller ikkje her opplagte årsaker til at ein får ei slik fordeling.



Figur 4.1.34 Histogram over residuala for dei predikerte butankonsentrasjonane for objekta i testdatasettet.

4.1.8 Fem nye modellar for metan

Sidan residuala til dei predikerte metanverdiane fordelte seg i fem grupper vart det laga lokale modellar for desse gruppene. Datasettet vart delt inn etter residual og nye modellar vart laga. Alle modellane inneheldt berre førstegradsledd.

Tabell 4.1.4 Oversikt over lokale modellar for metan

Gruppe	Objekt i modellen	Komponentar i modellen	Forklart varians i y, %
1	383	5	100.00
2	567	5	100.00
3	1496	7	100.00
4	783	5	99.94
5	760	5	99.97

Slik det er vist i tabell 4.1.4 kan ein for alle dei fem modellane sjå høg grad av forklart varians i y. Alle modellane har tilfredsstillande normalfordeling av responsresiduala, og etan var for alle modellane den komponenten med størst bidrag til modellen.

4.2 Oppbygging av kjemisk samansetjing ved prediksjon

I denne oppgåva ynskjer ein å finne modellar som kan predikere tettleik og brennverdi ut frå den predikerte kjemiske samansetjinga samt dei målte variablane trykk, temperatur og lydhastigheit. I dette kapitlet presenterer ein den kjemiske samansetjinga ved ein og ein komponent predikert. Modellane er validert ut frå testing med nye objekt, til dette har ein nytta testdatasettet som består av 400 objekt. Modellen er så nytta til å predikere ein og ein kjemisk komponent der predikert verdi av førre komponent inngår i prediksjon av neste komponent. Slik har ein bygd opp datasettet som seinare er nytta til prediksjon av tettleik og brennverdi.

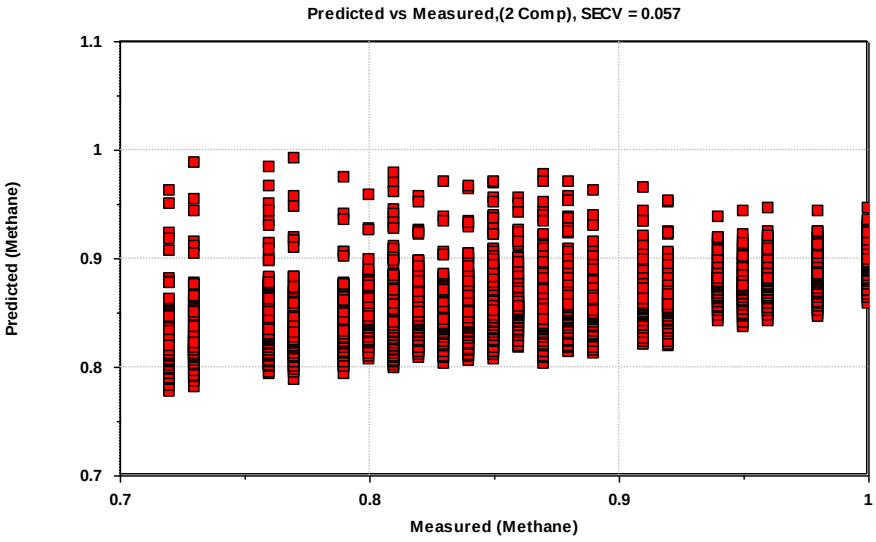
Tabell 4.2.1 Oversikt over modellar for oppbygging av kjemisk samansetjing bygd av kalibreringsdatasettet

Modellering			Validering			Prediksjon	
Modell	Forklart varians y %	Komp	RSD >	R^2	Q^2	Feil	Feil %
Metan=f(c,p,t)	20.68	2	1.953	0.185	0.180	0.050	5.8
etan=f(c,p,t,metan)	84.07	3	1.953	0.836	0.835	0.036	48.0
Etan=f(c,p,t,metan, andregradsledd og vekselverknadsledd	89.11	12	0.049	0.890	0.888	0.047	62.6
Metan iterert ved AR						0.040	46.5
Etan iterert ved AR						0.036	48.0
propan=f(c,p,t, metan,etan)	56.85	4	1.969	0.575	0.571	0.015	60.0
CO_2=f(c,p,t,metan, etan,propan)	44.05	5	1.969	0.430	0.425	0.009	60.0
N_2=f(c,p,t,metan, etan,propan,CO_2)	70.07	6	1.969	0.704	0.703	0.001	6.7
butan=f(c,p,t,metan, etan,propan,CO_2,N_2)	99.62	6	1.719	0.992	0.989	0.007	70.0

I resultata under har ein først predikert metan ut frå trykk, temperatur og lydhastigheit, deretter nytta den predikerte verdien av metan til å predikere etan ut frå trykk, temperatur, lydhastigheit og predikert metankonsentrasjon. Denne prosedyren er utført for alle dei kjemiske komponentane og ein har dermed til slutt fått eit datasett bestående av trykk, temperatur, lydhastigheit og ei predikert kjemisk samansetjing.

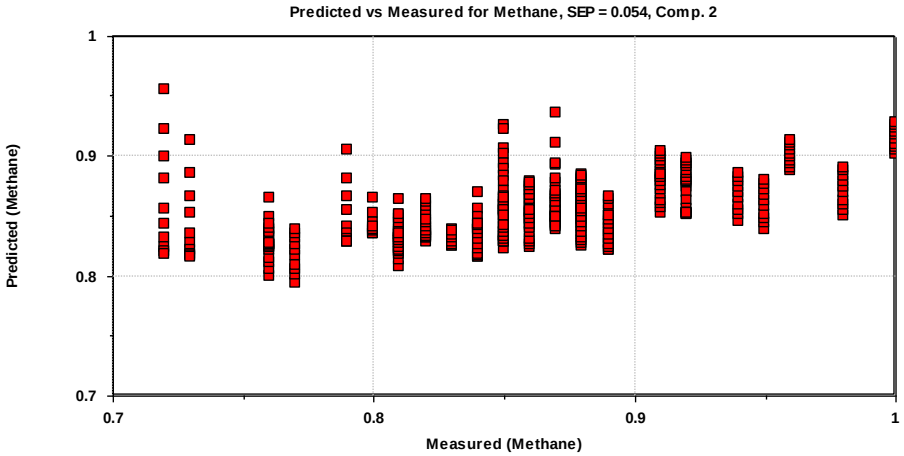
METAN

Som ein kan sjå frå tabell 4.2.1 har modellen for metan forklart varians i y på berre 20.68 %. Modellen har to komponentar. Samanhengen mellom målt og predikert verdi for modellen er svært lite tilfredsstillande (figur 4.2.1).

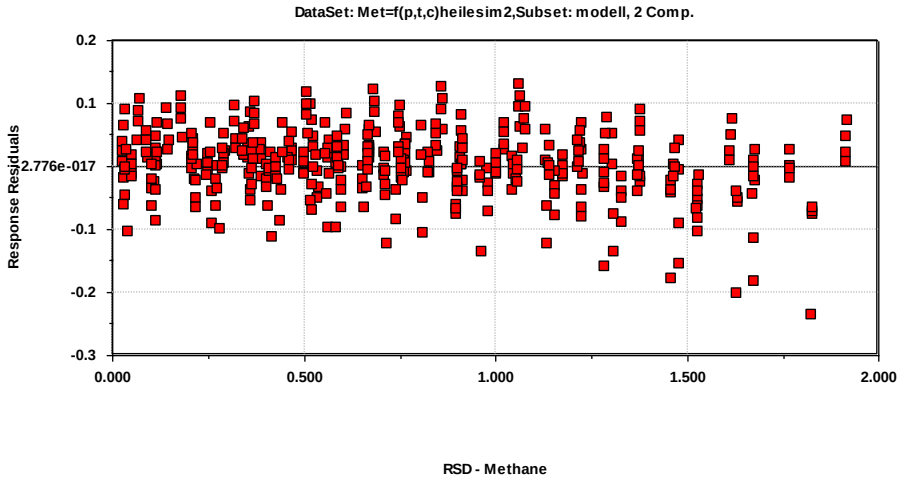


Figur 4.2.1 Plott av predikert verdi mot målt verdi for metan

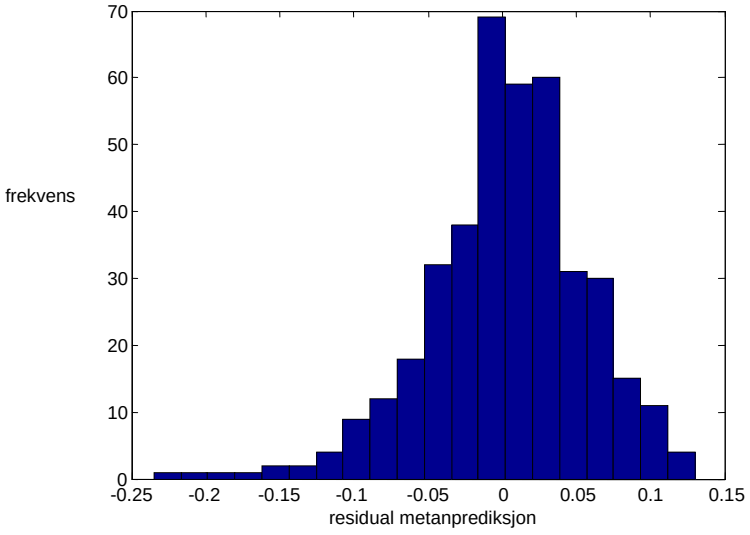
Det er heller ikkje tilfredsstillande samanheng mellom predikert og målt verdi for metan når ein testar modellen (figur 4.2.2). For dei lågaste og høgaste verdiane av metan treff ikkje modellen i det heile tatt. Best prediksjon ser ein for dei objekta som har verdiar i området 0.85. R^2 for modellen er 0.185 og Q^2 er 0.180, det er begge låge verdiar som bekreftar at modellen ikkje er tilfredsstillande. Det er ingen samanheng mellom høg RSD-verdi og høge responsresidual for objekta (figur 4.2.3). Avvisningskriteriet for uteliggjarar i denne modellen er RSD > 1.953. Ingen av objekta i testdatasettet har RSD-verdi over denne grensa. Dei objekta med dei største negative residual har høg RSD-verdi. Ut frå histogrammet i figur 4.2.4 som viser residuala for predikert metan for objekta i testdatasettet kan ein sjå at maksimumsfeilen er størst i negativ retning, men at hovudtyngda av residuala fordeler seg mellom -0.10 og 0.12. Ein ser størst residual for objekta med låg temperatur og høgt trykk (figur 4.2.5). Tabell 4.2.1 viser at prediksjonsfeilen er 0.05, dette tilsvarar 5.8 %.



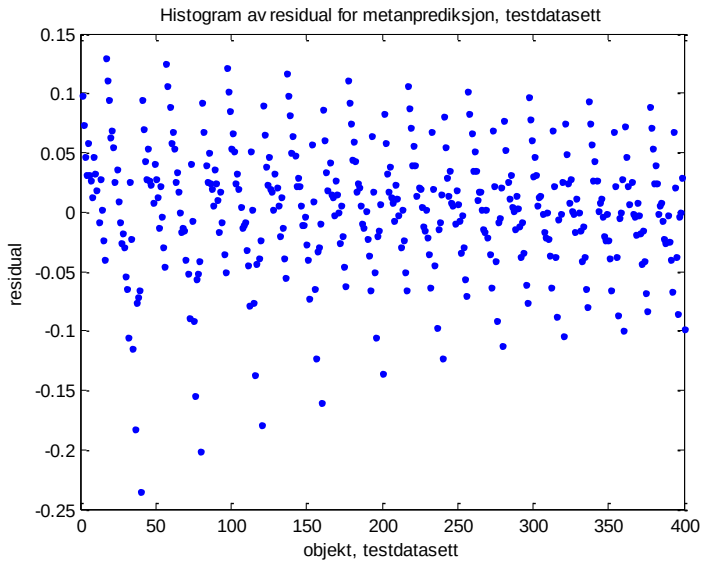
Figur 4.2.2 Plott av predikert verdi mot målt verdi for objekt frå testdatasettet.



Figur 4.2.3 Plott av responsresidual mot RSD for objekta som vert nytta til prediksjon.



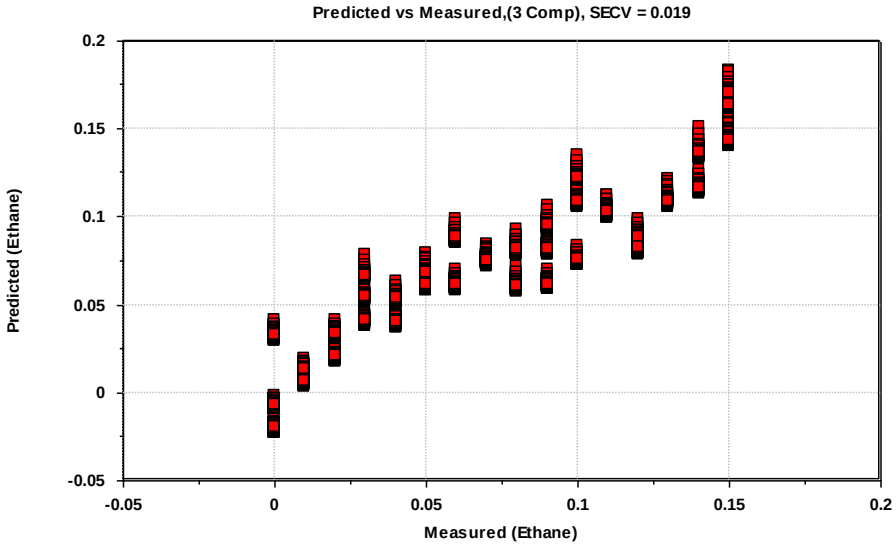
Figur 4.2.4 Histogram over residual for prediksjon av objekt frå testdatasettet.



Figur 4.2.5 Plott av residuala for metan for objekta frå testdatasettet.

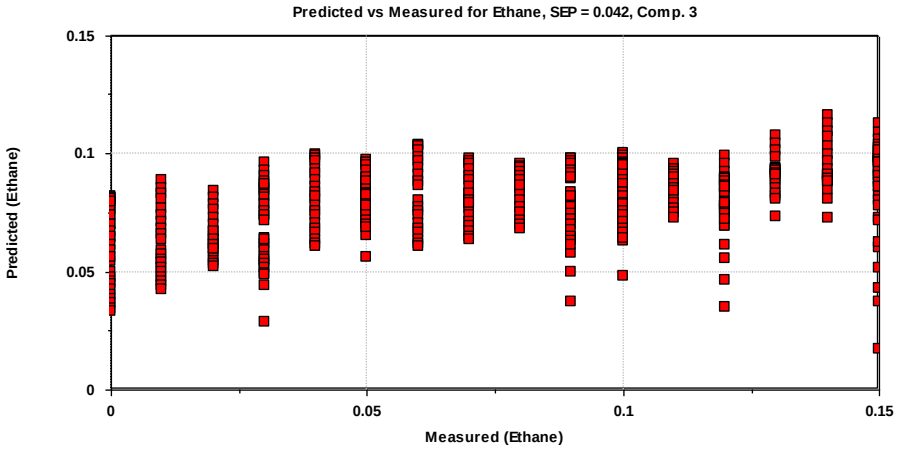
ETAN

Modellen for etan har tre komponentar og forklarer 87.07 % av variansen i y (tabell 4.2.1). Det er ein viss grad av samanheng mellom målt og predikert verdi sjølv om ein ser store residual for mange av verdiane (figur 4.2.6).

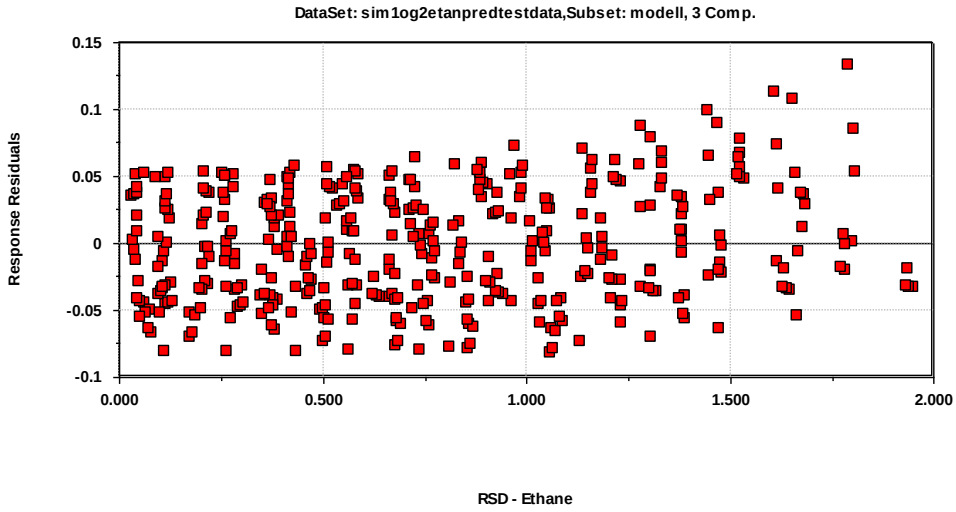


Figur 4.2.6 Plott av predikert verdi mot målt verdi i modellen for etan

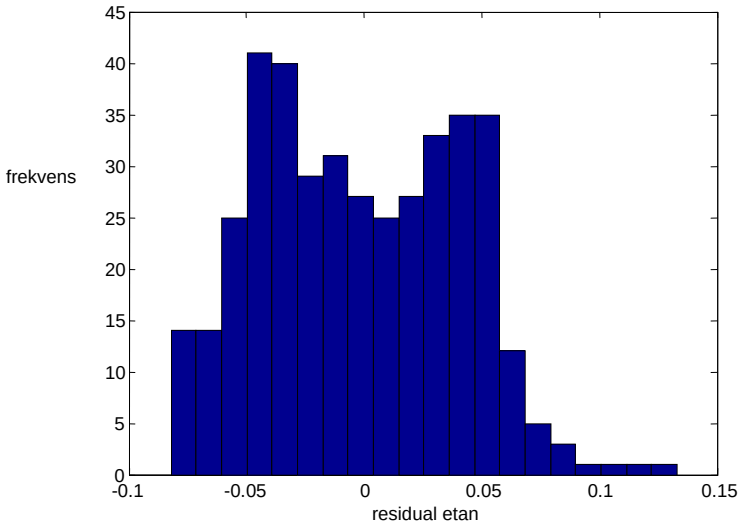
Som ein kan sjå ut frå tabell 4.2.1 er prediksjonsfeilen 0.036. Etan har verdiar i området 0.0 til 15.0 %, så er dette svarar til ein prediksjonsfeil på nesten 50 %. R^2-verdien er 0.836 og Q^2-verdien er 0.835, noko som tyder på at modellen har god prediktiv evne. Det er likevel ikkje tilfredsstillande samanheng mellom predikert og målt verdi for etan (figur 4.2.7), dette indikerer at modellen ikkje er til å stole på. Avvisningskriteriet for uteliggjarar i denne modellen er RSD > 1.953. Ingen av objekta i testdatasettet som vert nytta til prediksjon har RSD-verdiar over denne grensa. Dei objekta som har høgast positive residual har også høge RSD-verdiar, samtidig som det finnes objekt med responsresidual rundt 0 med like høge RSD-verdiar, dette er vist i figur 4.2.8 . Ut frå histogrammet i figur 4.2.9 kan det sjå ut som om det er to overlappande fordelingar av etan, ein med topp for residuala ved -0.05 og ein med topp ved 0.05.



Figur 4.2.7 Plott av predikert verdi mot målt verdi for etan, objekt frå testdatasett

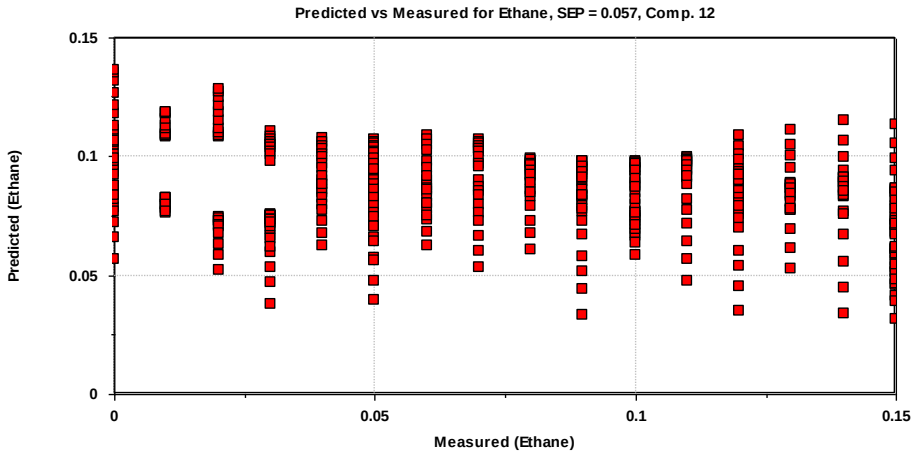


Figur 4.2.8 Plott av responsresidual mot RSD for objekta som nyttast til prediksjon

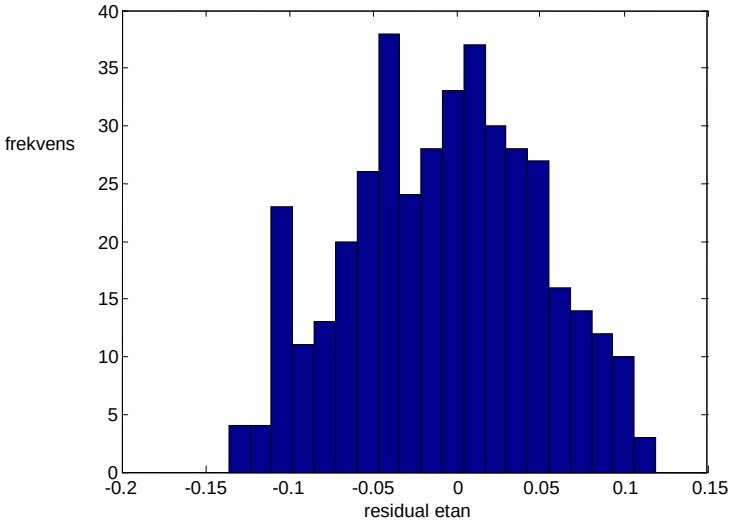


Figur 4.2.9 Histogram av residuala for prediksjon av etan.

Det vart også laga ein modell for etan som inneheld førstegradsledd, andregradsledd og vekselverknadsledd. Ein kan sjå frå tabell 4.2.1 at denne modellen inneheld 12 komponentar og har forklart varians i y på 89.11 %. Ser likevel at prediksjonsfeilen aukar frå 40 % til heile 60 %. og ein kan sjå ut frå histogrammet i figur 4.2.11 at også maksimumsfeilen aukar i negativ retning. Modellen har ikkje tilfredsstillande samanheng mellom målt og predikert verdi, slik ein kan sjå i figur 4.2.10. Dersom ein samanliknar med situasjonen der modellen inneheld berre førstegradsledd ser ein at det er enda dårlegare samanheng mellom predikert og målt verdi for modellen for etan som inkluderer førstegradsledd, andregradsledd og veksleverknadsledd. R^2 er 0.890 og Q^2 er 0.888 for modellen, dette er litt betre enn for førstegradsmodellen, men ikkje så mykje at denne auken vil vere avgjerande for å velje ein modell med andregradsledd og vekselverknadsledd inkludert. Vel difor å gå vidare med prediksjonane frå førstegradsmodellen.



Figur 4.2.10 Plott av predikert verdi mot målt verdi for etan, modell med andregradsledd og vekselvirkningsledd.



Figur 4.2.11 Histogram av residual for etan etter test med objekt frå testdatasett, modellen inneheld andregradsledd og vekselverknadssledd.

ITERATIV KONSENTRASJONSBESTEMMING AV METAN OG ETAN VED AR

Ynskjer å sjå om AR kan føre til at prediksjonane av metan og etan kjem nærare referanseverdiane. Ein utfører difor iterativ konsentrasjonsbestemming ved AR for metan og etan der ein nyttar regresjonskoeffisientane frå modellane (4.2.1) og (4.2.2)

$$metan = f(p, T, c, etan) \quad (4.2.1)$$

$$etan = f(p, T, c, metan) \quad (4.2.2)$$

Startverdiane for iterasjonen er dei predikerte verdiane ein har funne i Sirius ved å nytte modellane (4.2.3) og (4.2.4)

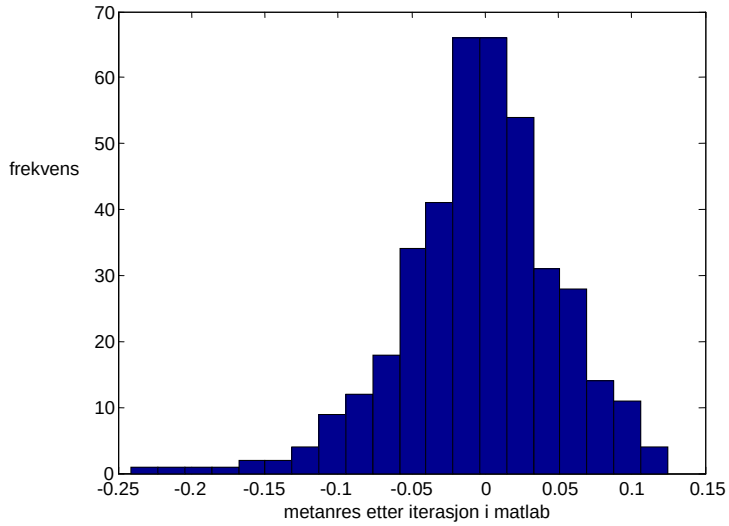
$$metan = f(p, T, c) \quad (4.2.3)$$

$$etan = f(p, T, c) \quad (4.2.4)$$

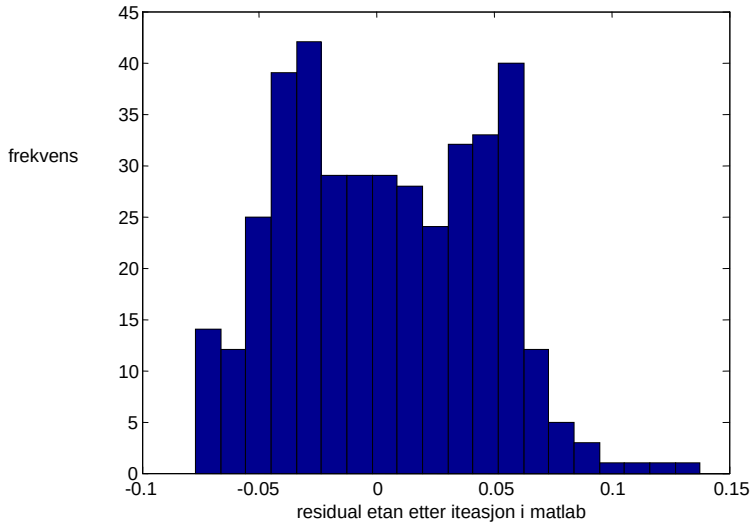
Der p er trykk (bar) , T er temperatur (˚C) og c er lydhastigheit (m/s).

MATLAB-programmet gjennomførte 14-16 iterasjonar for objekta i testdatasettet. Residuala er rekna ut frå målte verdiar i testdatasettet og ikkje frå startpunktet i iterasjonen.

Etter iterativ konsentrasjonsbestemming ser ein at residuala fordeler seg om lag likt for metan som ved prediksjon i Sirius. Tabell 4.2.1 viser at prediksjonsfeilen no er redusert frå 0.05 til 0.04, noko som tilsvarar ein reduksjon frå 5.8 % feil til 4.7 % feil. Maksimumsfeilen for metan endrar seg ikkje etter at ein har utført AR, histogram over residuala er vist i figur 4.2.12. For etan endrar ikkje prediksjonsfeilen seg etter iterativ konsentrasjonsbestemming. Prediksjonsfeilen utgjer heile 48 %, noko som er svært høgt. Etan har verdiar i området 0.00 – 0.15 og dermed vil ein få store prosentvise feil sjølv om den reelle feilen ikkje er så stor. Histogrammet i figur 4.2.13 viser at maksimumsfeilen heller ikkje har endra seg for etan etter at ein har utført AR. Ein ser også her at det er to toppar i plottet, det eine rundt -0.05 og det andre rundt 0.05. For metan kan det vere lønnsamt å gå vegen via iterasjon, men for etan har det ingen effekt på prediksjonen. Ein går vidare med metan og etan predikert frå Sirius. Dette vert grunngjeve med at sjølv om den prediksjonsfeilen rekna ut frå formel (2.23) går ned endrar ikkje maksimumsfeilen seg. Modellen for metan er uansett ikkje tilfredsstillande verken når det gjeld forklart varians og prediktiv evne.



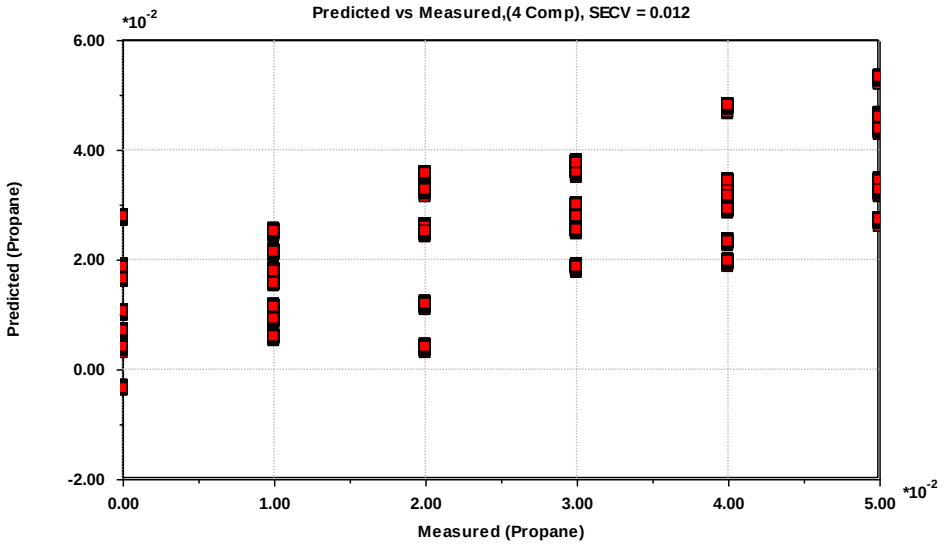
Figur 4.2.12 Residual for metan etter iterativ konsentrasjonsbestemming ved AR



Figur 4.2.13 Residual for etan etter iterativ konsentrasjonsbestemming i AR.

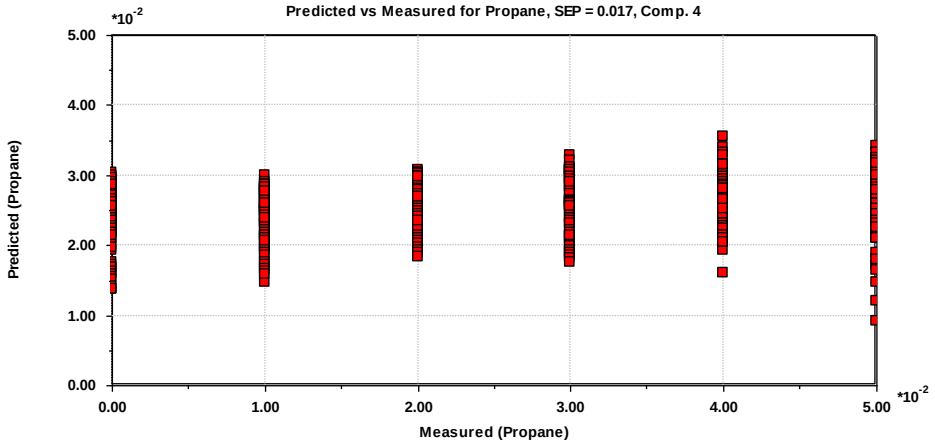
PROPAN

Som ein kan sjå frå tabell 4.2.1 består modellen for propan av fire komponentar og har forklart 56.85 % av variansen i y. Samanhengen mellom predikert og målt verdi for objekta er ikkje tilfredsstillande. Ein ser at for dei objekta med låg målt verdi er den predikerte verdien for høg, og for dei objekta som har høg målt verdi er prediksjonane for låge, dette er vist i figur 4.2.14.

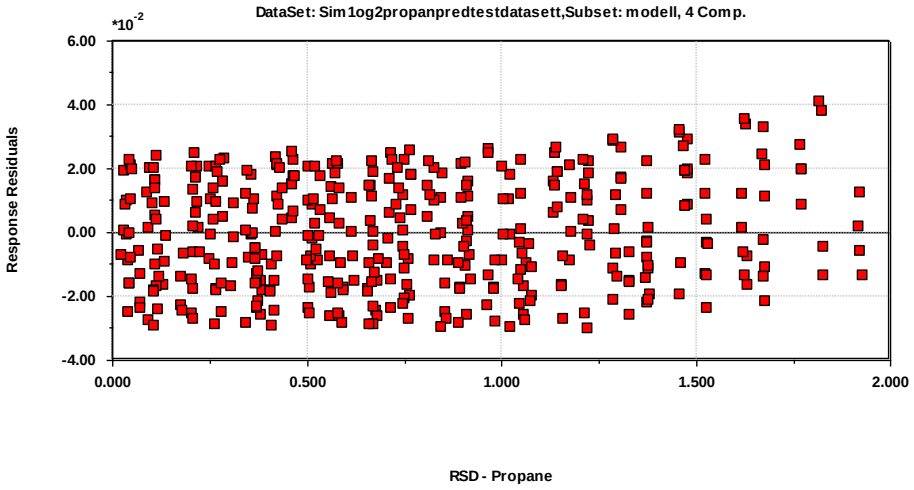


Figur 4.2.14 Plott av predikert verdi mot målt verdi for modell for propan.

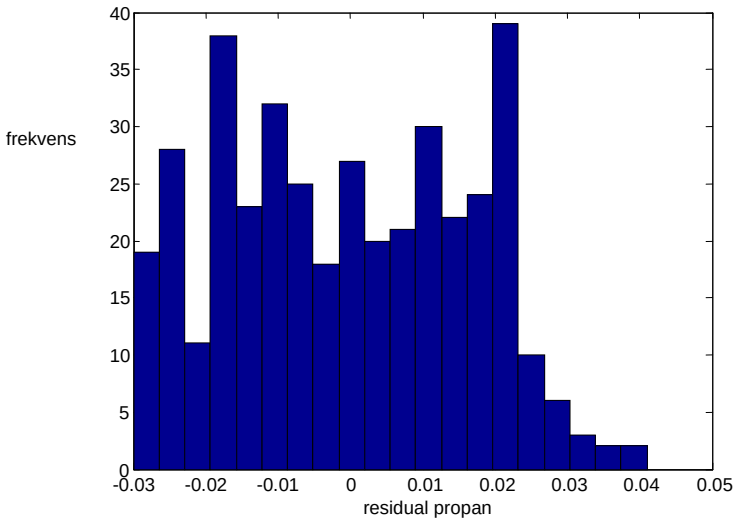
Tabell 4.2.1 viser at prediksjonsfeilen for propan er 0.015, noko som svarar til heile 60 %. R^2 for modellen er 0.575 og Q^2 for modellen er 0.571, dette vert i følgje [44] sett på som bra, men det samsvarar ikkje med resultata funne i denne oppgåva, ein ser at modellen ikkje predikerer bra på nye objekt (figur 4.2.15). Avvisningskriteriet for uteliggjarar i denne modellen er RSD > 1.969. Det er ingen av objekta nytta til prediksjon som har RSD-verdi over denne grensa. Det er ingen tydeleg samanheng mellom responsresidual og RSD, men ser at dei objekta med høgast positive residual har høge RSD-verdiar, dette er vist i figur 4.2.16. Når ein ser på histogrammet i figur 4.2.17 ser ein ei overvekt av negative residual. Residuala fordeler seg ikkje i retning normalfordeling og også her ser ein omrisset av to toppar, ein for verdiar nær -0.02 og ein for verdiar nær 0.02.



Figur 4.2.15 Plott av predikert mot målt verdi for propan, testdatasett



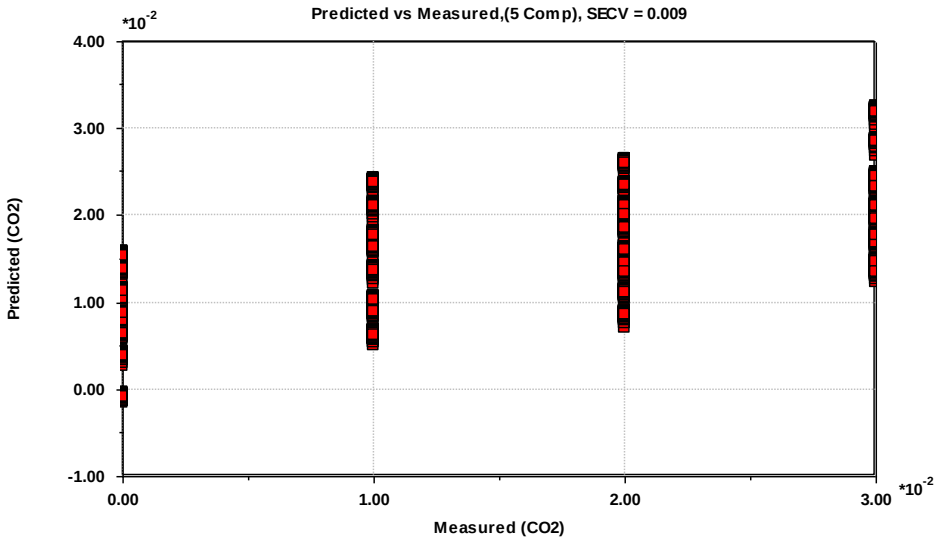
Figur 4.2.16 Plott av responsresidual mot RSD for objekta nytta til prediksjon.



Figur 4.2.17 Histogram av residual for prediksjon av propan.

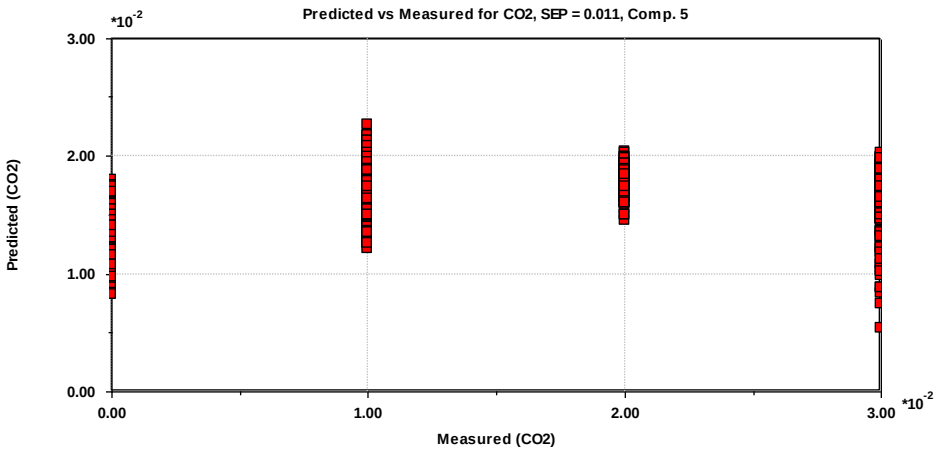
CO_2

Som ein kan sjå i tabell 4.2.1 har modellen for CO_2 inkludert fem komponentar og forklart varians i y er 44.05 %. Det er ikkje bra samanheng mellom predikert og målt for verdi for objekta slik ein kan sjå i plottet i figur 4.2.18.

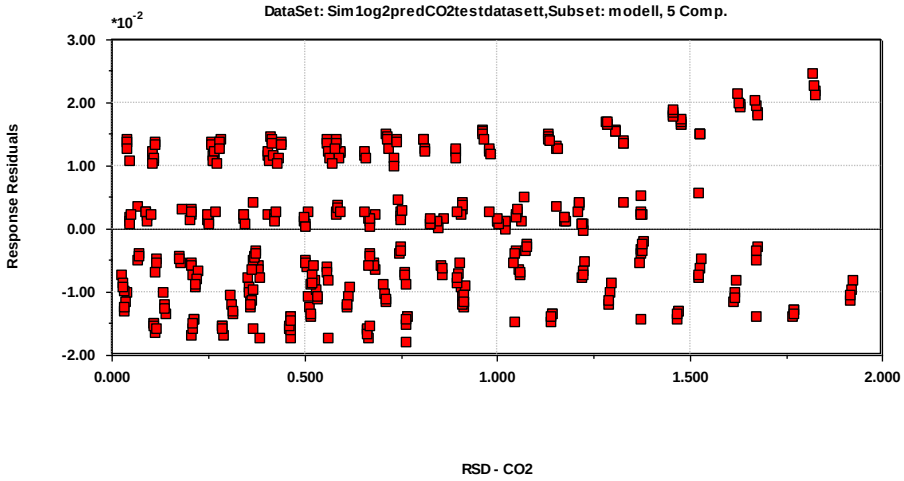


Figur 4.2.18 Plott av predikert verdi mot målt verdi i modellen for CO_2

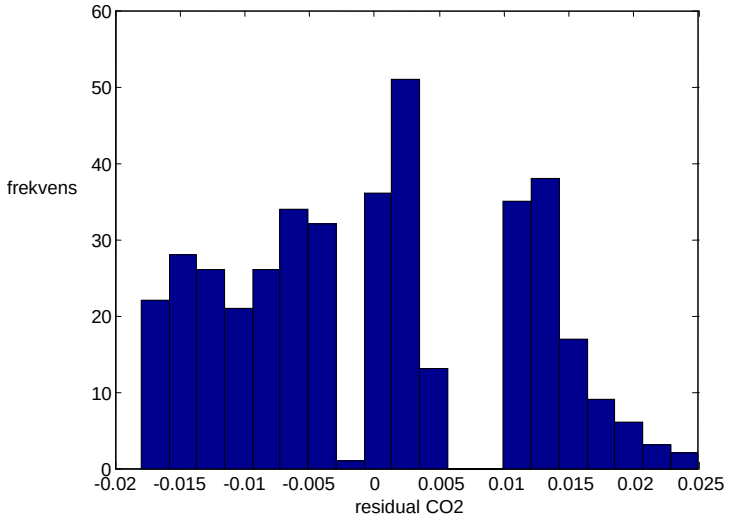
Prediksjonsfeilen for CO_2 når ein testar med nye objekt er 0.009, dette tilsvarar 45 %. Dette er eit høgt tal, men CO_2 har målte verdiar i området 0.00 til 0.03, dette vil også her føre til at små feil utgjer store prosenttal. R^2 verdien for modellen er 0.430 og Q^2 verdien er 0.425. Dette indikerer at den prediktive evna til modellen ikkje er bra og ved å plotte predikert verdi mot målt verdi ser ein at modellen predikerer svært dårleg (figur 4.2.19). Avvisningskriteriet for uteliggjarar i denne modellen er RSD > 1.969. Ingen av objekta som vert nytta til prediksjon har RSD over denne grensa. Det kan sjå ut som det er tre grupper i plottet av responsresidual mot RSD (figur 4.2.20). Ei fordeling med responsresidual mellom -0.02 og 0, ei fordeling med responsresidual rundt 0 og ei fordeling med responsresidual rundt 0.01 – 0.02. Histogrammet i figur 4.2.21 viser at residuala er noko ujamt fordelt rundt 0. Maksimumsfeilen er størst i positiv retning.



Figur 4.2.19 Plott av predikert verdi mot målt verdi for CO_2 etter prediksjon



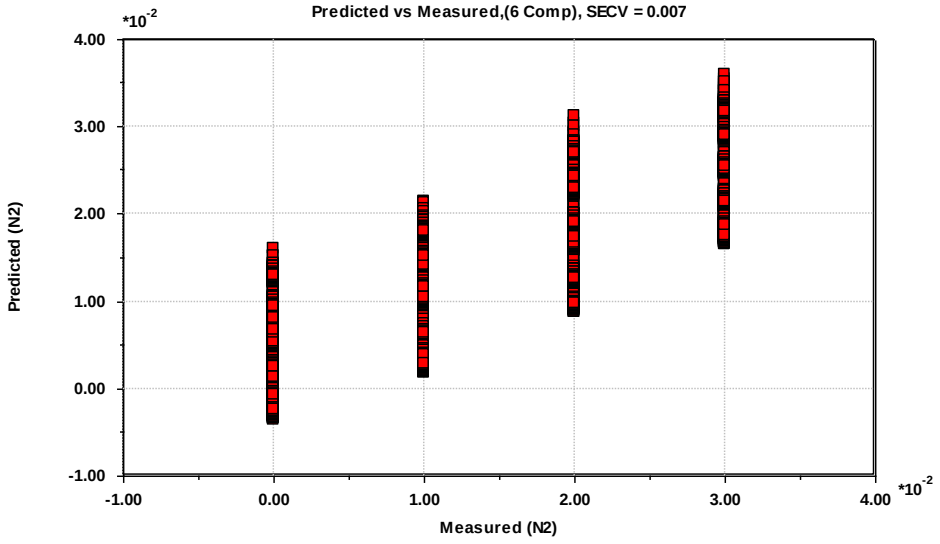
Figur 4.2.20 Plott av responsresidual mot RSD for objekta brukt til prediksjon



Figur 4.2.21 Histogram av residual for CO_2

N_2

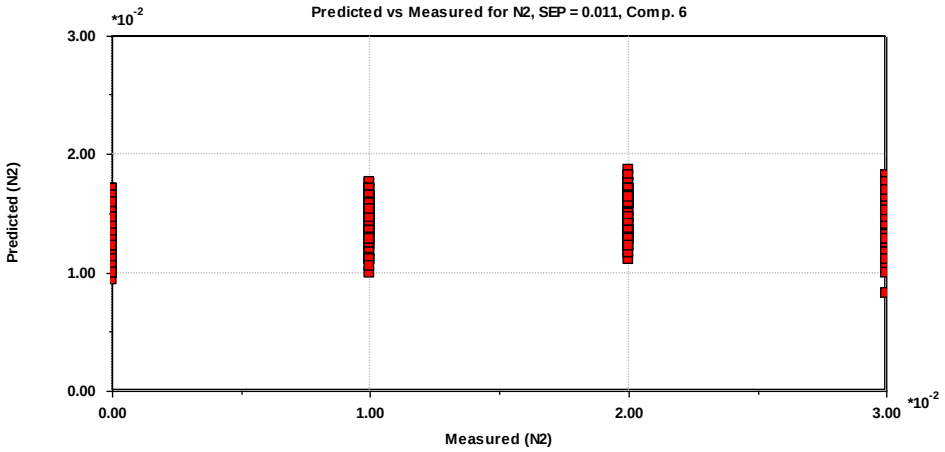
Modellen for N_2 inneheld seks komponentar og har forklart varians i y på 70.07 %, slik ein kan sjå i tabell 4.2.1. Forklart varians i y har altså auka betrakteleg frå modellen for CO_2 til modellen for N_2. Ser også at det er ein viss grad av samanheng mellom målt og predikert verdi sjølv om residuala er store, dette er vist i figur 4.2.22.



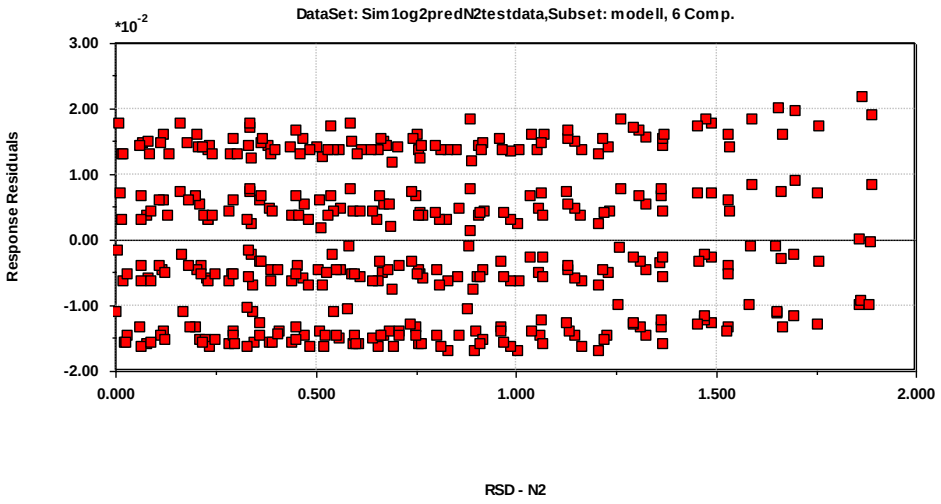
Figur 4.2.22 Plott av predikert verdi mot målt verdi for N_2.

Prediksjonsfeilen for N_2 for nye objekt er også mykje lågare enn for CO_2. Verdien er 0.001, dette tilsvarar prediksjonsfeil i størrelsesorden 6,7 %. (Tabell 4.2.1). Samanhengen mellom predikert og målt verdi er ikkje tilfredsstillande. Av plottet i figur 4.2.23 ser ein at for låge verdiar av N_2 vert predikert for høgt og for høge verdiar av N_2 vert predikert for lågt. R^2 og Q^2 for modellen er 0.704 og 0.703, noko som peikar mot at modellen skal vere i stand til å predikere relativt bra sjølv om dette ikkje er tilfelle for denne modellen. Avvisningskriteriet for uteliggjarar i denne modellen er RSD > 1.969. Ingen av objekta som vert nytta til testing har RSD-verdiar over denne grensa. Ved å plottet responsresidual mot RSD ser ein igjen dei fire gruppene med residual som er omtala ovanfor (figur 4.2.24). Histogrammet i figur 4.2.25 viser at maksimumsfeilen er størst i positiv retning. Ut frå histogrammet ser det ut som om det er fire grupper med residual. Desse gruppene ser ut til å vere om lag like store. I slike tilfeller er det naturleg å tenke på om lokale modellar er ein moglegheit. I figur 4.2.26 ser ein korleis residuala endrar seg. Ein kan sjå eit gjentakande mønster for kvart 40. objekt, altså for kvar gong temperaturen aukar. Inne i kvar gruppe endrar trykket seg for kvart fjerde

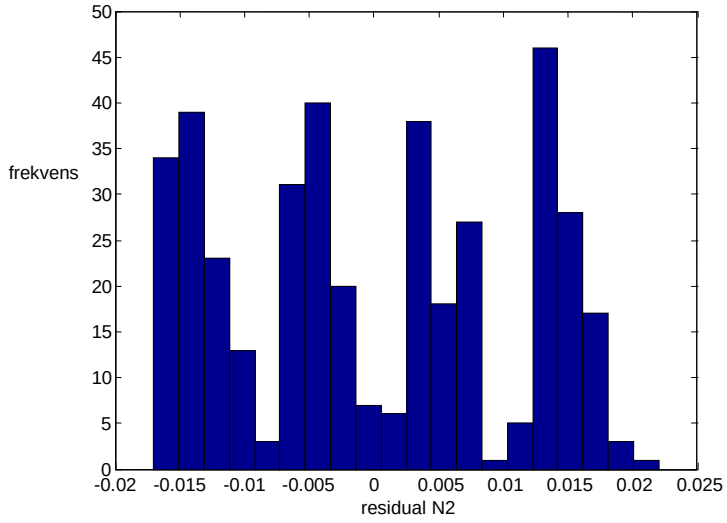
objekt og ein ser at inne i desse gruppene på fire objekt som har same trykk og temperatur har ein stor variasjon i residual. Nitrogen har fire nivå for målte verdiar, 0.00, 0.01, 0.02 og 0.03.



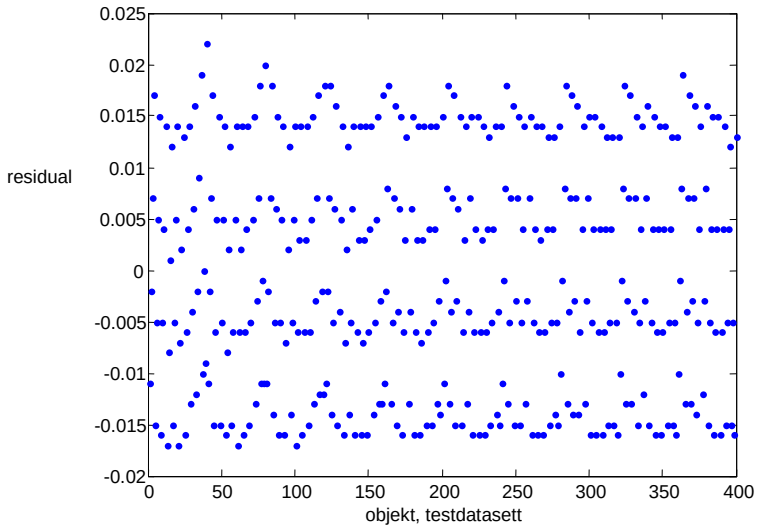
Figur 4.2.23 Plott av predikert mot målt verdi for N_2 etter prediksjon



Figur 4.2.24 Plott av responsresidual mot RSD for objekta nytta til prediksjon av N_2



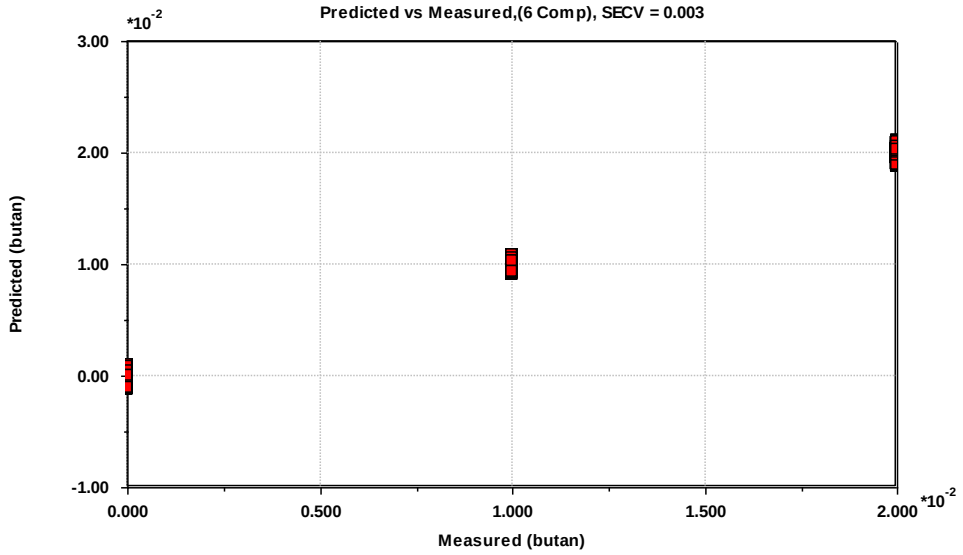
Figur 4.2.25Histogram av residual for predikert N_2, objekt frå testdatasett



Figur 4.2.26 Plott av residual mot objekt for N_2 for objekta frå testdatasettet.

BUTAN

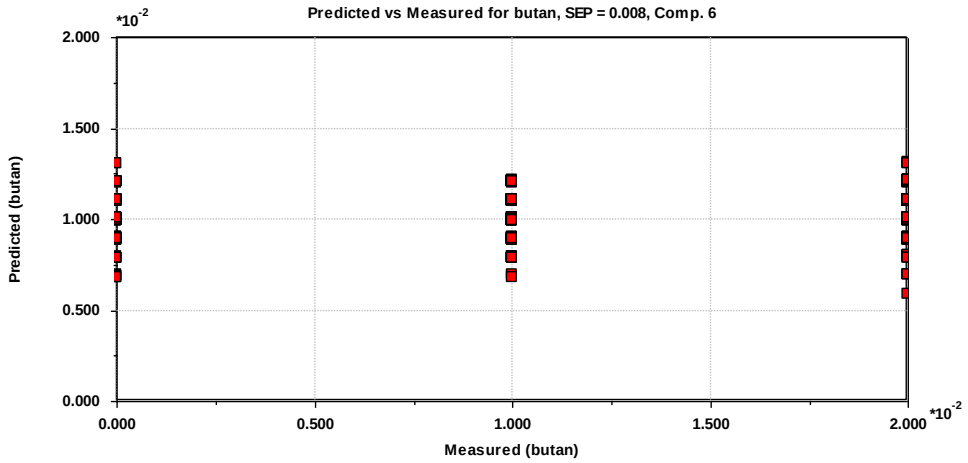
Modellen for butan inneheld seks komponentar og har forklart varians i y på 99.62 %. (Tabell 4.2.1). Sidan ein no har inkludert heile den kjemiske samansetjinga i tillegg til variablane trykk, temperatur og lydhastigheit er det forventa at forklart varians i y blir god. Det er god samanheng mellom predikert og målt verdi i modellen slik ein kan sjå frå plottet i figur 4.2.27.



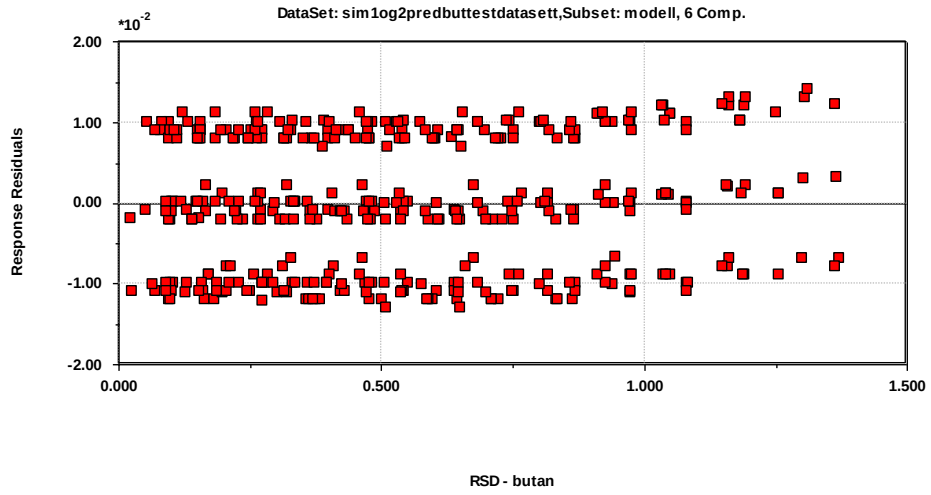
Figur 4.2.27 Plott av predikert mot målt verdi for butan.

Feilen etter prediksjon av nye objekt er 0.007 slik det er vist i tabell 4.2.1. Dette tilsvarar 70 % feil. For låg butanverdi er prediksjonane for høge, og for høg butanverdi er prediksjonane for låge. Dette samsvarar dårleg med at R^2 og Q^2 for modellen er 0.992 og 0.989 noko som i utgangspunktet peikar mot at modellen har svært bra intern prediksjonsevne. Det er likevel ikkje tilfredsstillande samanheng mellom predikert og målt verdi for objekta frå testdatasettet (figur 4.2.28). Avvisningskriteriet for uteliggjarar i denne modellen er RSD > 1.719. Ingen av objekta nytta til prediksjon har RSD-verdi over denne grensa. Ved å plotte responsresidual mot RSD slik det er vist i figur 4.2.29 ser ein tydeleg at det er tre grupper residual tilstades, inne i desse gruppene er RSD-verdiane om lag likt fordelt. Dei objekta med det høgaste residualet i positiv retning har høgast verdiar for RSD. Histogrammet i figur 4.2.30 viser også tydleg at det er tre grupper residual tilstades. Butan har berre tre nivå for målte verdiar: 0.00, 0.01 og 0.03. Ved å inspisere datasettet ser ein at objekta med målt verdi 0.00 har negative residual, objekta med målt verdi 0.01 har residual omkring 0 og

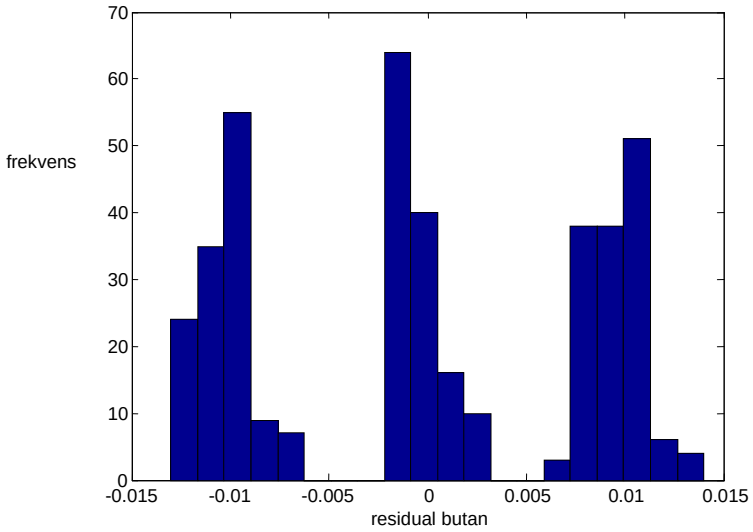
objekta med målt verdi 0.02 har positive residual. Toppane i histogrammet er altså knytt opp mot målte verdiar av butan. Dette samsvarar med det som er sagt tidlegare om å vurdere lokale modellar.



Figur 4.2.28 Plott av predikert mot målt verdi for butan, objekt frå testdatasett.



Figur 4.2.29 Plott av responsresidual mot RSD for objekta nytta til prediksjon



Figur 4.2.30 Histogram av residual for butan, objekt frå testdatasett

Ein har no forsøkt å predikere heile den kjemiske samansetjinga ut frå dei kjende variablane trykk, temperatur og lydhastigheit. Resultata av prediksjonane er svært varierande og ein merkar seg særleg at modellen for metan, som jo er hovudbestanddelen i naturgass, er svært lite tilfredsstillande. Kor mykje dette påverkar prediksjon av tettleik og brennverdi er presentert i kapittel 4.6 og 4.7.

4.3 Oppbygging av kjemisk samansetjing ved revers prediksjon

Sidan modellane blir betre jo fleire kjemiske komponentar ein inkluderer prøver ein her å starte med å modellere den kjemiske komponenten det er minst av, slik at dei største bestanddelane i naturgassen (metan og etan) skal få gode modellar til prediksjonen.

Tabell 4.3.1 Oversikt over oppbygging av modellar og revers prediksjon av kjemisk samansetjing

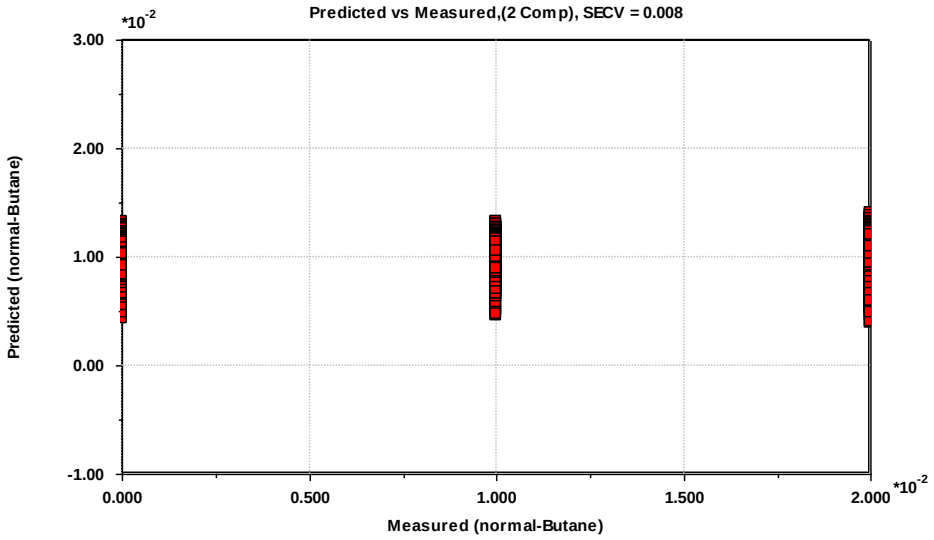
Modellering			Validering			Prediksjon	
Modell	Forklart varians y %	Komp	RSD>	R^2	Q^2	Feil	Feil, %
Butan=f(p,t,c)	2.73	2	1.953	0.020	0.015		
N_2=f(p,t,c,butan)	2.89	3	1.953	0.035	0.031	0.010	66.7
CO_2=f(p,t,c,butan,N_2)	7.72	3	1.727	0.075	0.071	0.009	60.0
Propan=f(p,t,c,butan, N_2,CO_2)	7.58	4		0.076	0.075	0.015	60.0
Prop inkl 2.gradsledd	13.61	11		0.122	0.109	0.014	56.0
Prop inkl 2.gradsledd og vekselverknadsledd	36.78	19	0.233	0.354	0.341	0.004	16.0
Etan=f(p,t,c,butan,N_2, CO_2,propan)	32.81	4		0.316	0.313	0.036	48.0
Et inkl 2.gradsledd og vekselverknadsledd	79.57	19	0.332	0.785	0.773	0.028	37.3
Metan=f(p,t,c,butanN_2, CO_2,propan,etan)	99.96	4	1.457	1.000	1.000	0.026	3.0

I resultata presentert under har ein først predikert butan ut frå trykk, temperatur og lydhastigheit, deretter nytta den predikerte verdien av butan saman med trykk, temperatur og lydhastigheit til å predikere N_2. Den same prosedyren er utført for dei resterande kjemiske komponentane slik det er vist i tabell 4.3.1. Dermed har ein til slutt eit datasett bestående av trykk, temperatur, lydhastigheit og ei predikert kjemisk samansetjing.

Modellane er validert ut frå testing med nye objekt, til dette har ein nytta testdatasettet som består av 400 objekt. Ein har så nytta modellane til å predikere den kjemiske samansetjinga som beskrive.

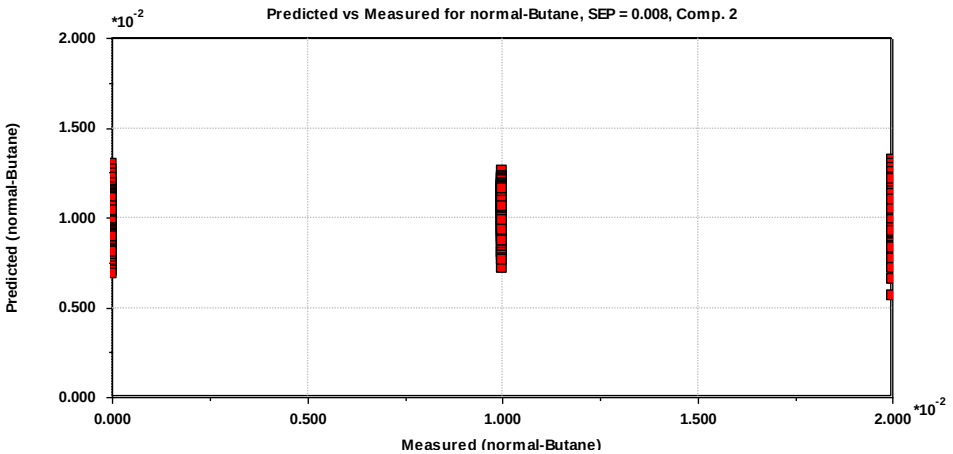
BUTAN

Som ein kan sjå frå tabell 4.3.1 har modellen for butan forklart varians i y på berre 2.73 %. Modellen inneheld to komponentar. Det er ikkje tilfredsstillande samanheng mellom målt og predikert verdi for objekta i modellen (figur 4.3.1).

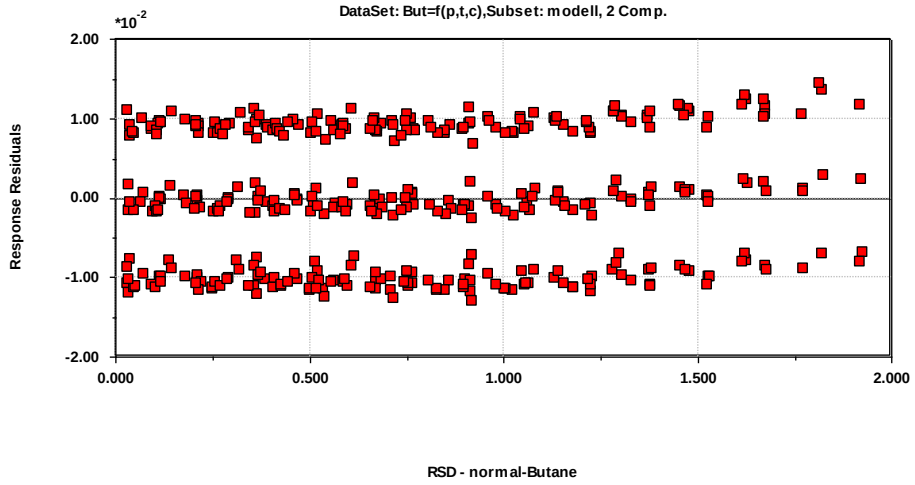


Figur 4.3.1 Plott av predikert mot målt verdi for objekta i modellen for butan

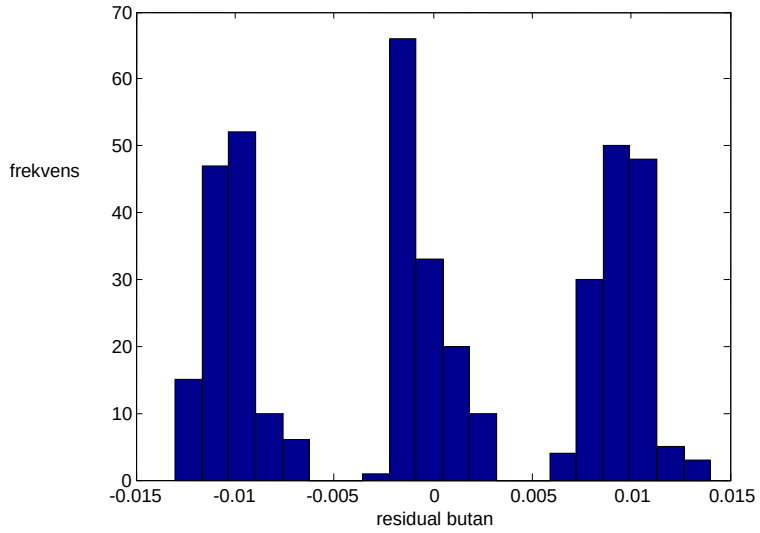
Ved validering av modellen ser ein heller ikkje tilfredsstillande samanheng mellom målt og predikert verdi (figur 4.3.2). R^2 og Q^2 er 0.020 og 0.015, noko som også indikerer svært dårleg kumulativ forklart varians og intern prediktiv evne. Når deg gjeld forholdet mellom responsresidual og RSD er det i dei tre gruppene om lag lik fordeling av RSD-verdiar (figur 4.3.3). Avvisningskriteriet for uteliggjarar i denne modellen er RSD > 1.953. Ingen objekt ligg over denne grensa. Histogrammet i figur 4.3.4 viser at det er tre grupper residual tilstades. Desse er knytt opp mot nivået av målt butan. Objekt med målt butanverdi 0.00 har negative residual, objekt med målt butanverdi 1.00 har residual rundt 0.00 og objekt med målt butanverdi 0.01 har positive residual.



Figur 4.3.2 Plott av predikert mot målt verdi for butan, data frå testdatasettet.



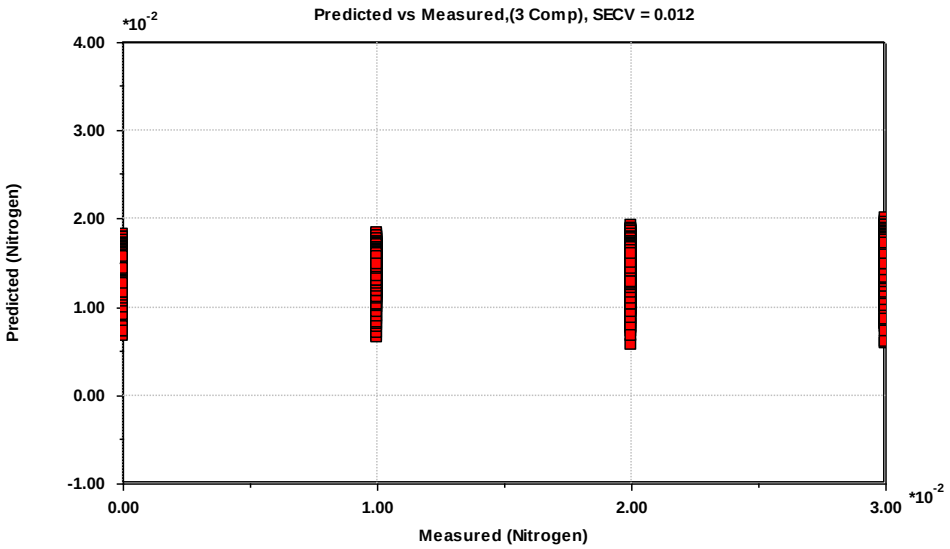
Figur 4.3.3 Plott av responsresidual mot RSD for butan



Figur 4.3.4 Histogram av residual for butan

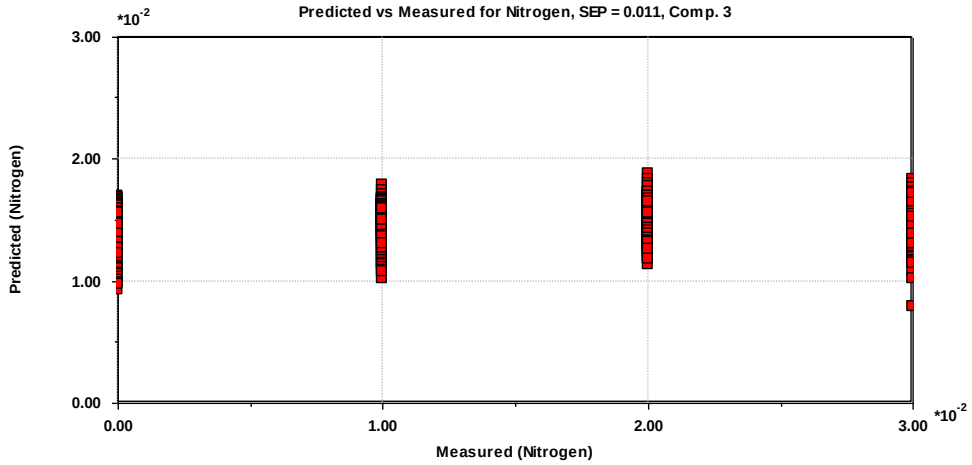
N_2

Modellen for N_2 har tre komponentar og forklart varians i y på berre 2.89 % (tabell 4.3.1). Dette er i utgangspunktet ein lite tilfredsstillande modell sidan den forklarer så liten del av datamaterialet. Ein ser svært liten samanheng mellom målt og predikert verdi (figur 4.3.5).

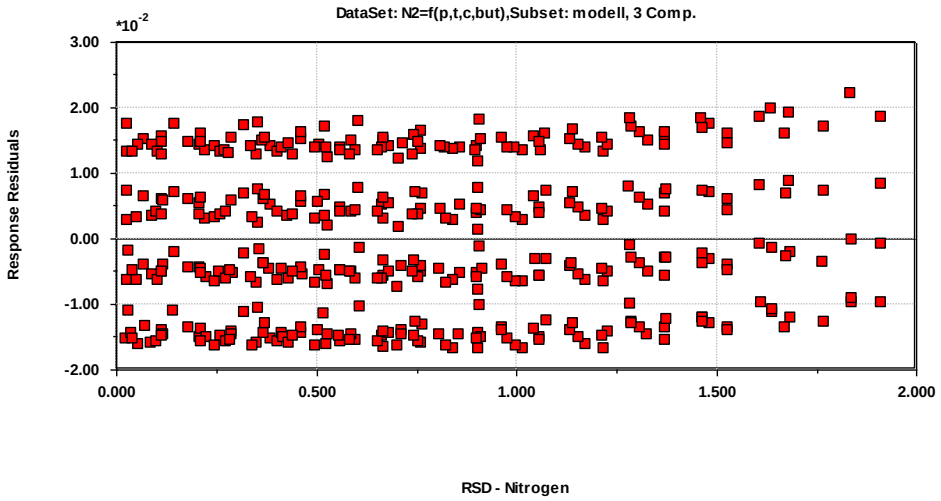


Figur 4.3.5 Plott av predikert verdi mot målt verdi i modellen for N_2

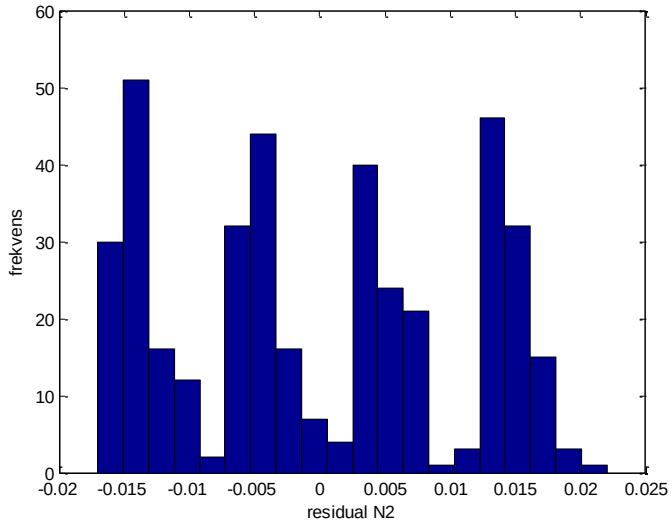
Når ein testar modellen med nye prøvar frå testdatasettet ser ein at det heller ikkje no er samanheng mellom målt og predikert verdi (figur 4.3.6). I tabell 4.3.1 kan ein sjå at R^2 er 0.035 og Q^2 er 0.031, med så låge verdiar for både kumulativ forklart varians og intern prediktiv evne tyder det på at den prediktive evna til modellen ikkje er forventa å vere tilfredsstillande. Prediksjonsfeilen er 0.010, noko som tilsvarar heile 67 % feil i forhold til gjennomsnittsverdien av N_2. Det om lag lik fordeling av RSD-verdiar for objekta inne i dei fire gruppene (figur 4.3.7). Avvisningskriteriet for uteliggjarar i denne modellen er RSD > 1.953. (Tabell 4.3.1) Ingen objekt har RSD-verdi over denne grensa. Histogrammet i figur 4.3.8 viser at residuala for prediksjon av N_2 deler seg i fire grupper. Der N_2 har målt verdi 0.00 ser ein dei største negative residuala, for dei objekta der N_2 har målt verdi 0.01 ser ein residual med topp rundt -0.05, der N_2 er målt til 0.02 er residuala fordelt rundt 0.05 og der N_2 er målt til 0.03 ser ein dei største positive residuala, med topp rundt 0.015. Maksimumsfeilen er størst i positiv retning. Dette tyder på at ein kunne vurdert lokale modellar for N_2.



Figur 4.3.6 Plott av predikert verdi mot målt verdi for predikert N_2



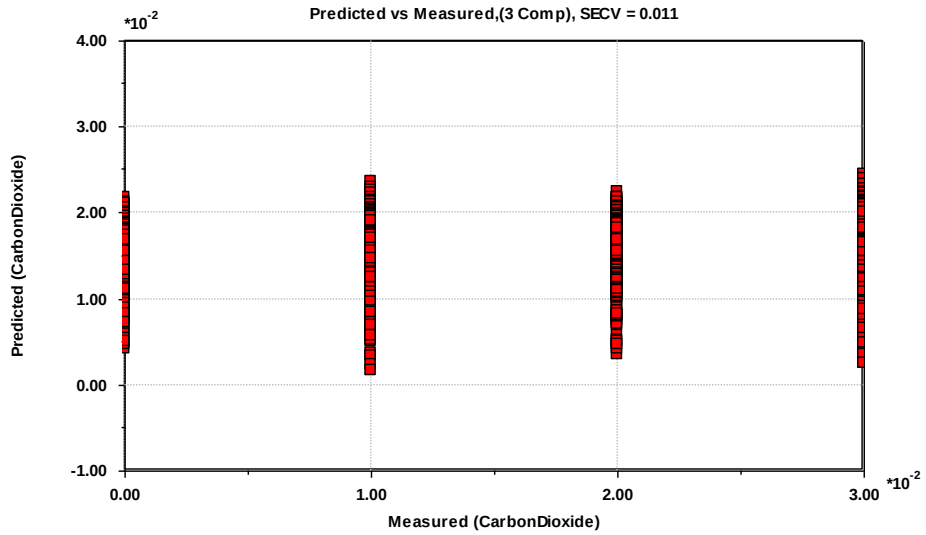
Figur 4.3.7 Plott av responsresidual mot RSD for predikert N_2



Figur 4.3.8 Histogram over residuala i prediksjon av N_2, objekt frå testdatasett

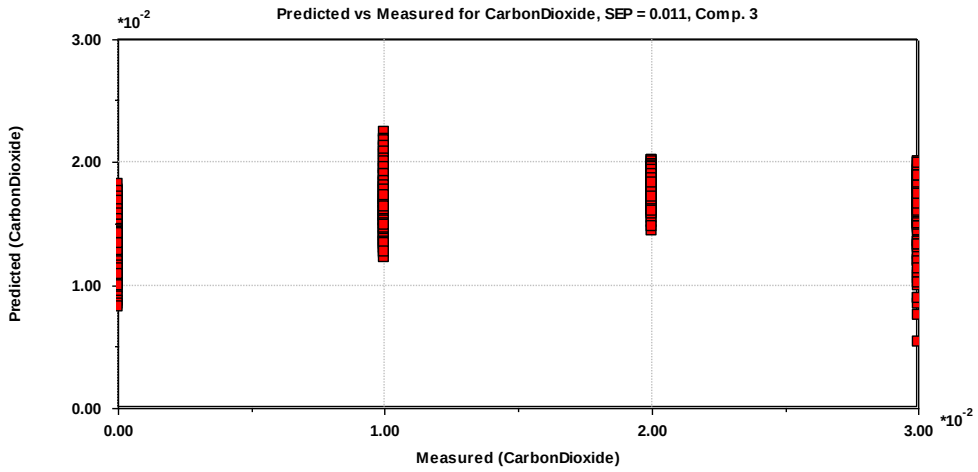
CO_2

Modellen for CO_2 er ein trekomponentsmodell som forklarer 7.72 % av y. (Tabell 4.3.1). Plottet i figur 4.3.9 viser at det ikkje er tilfredsstillande samanheng mellom målt og predikert verdi for objekta.

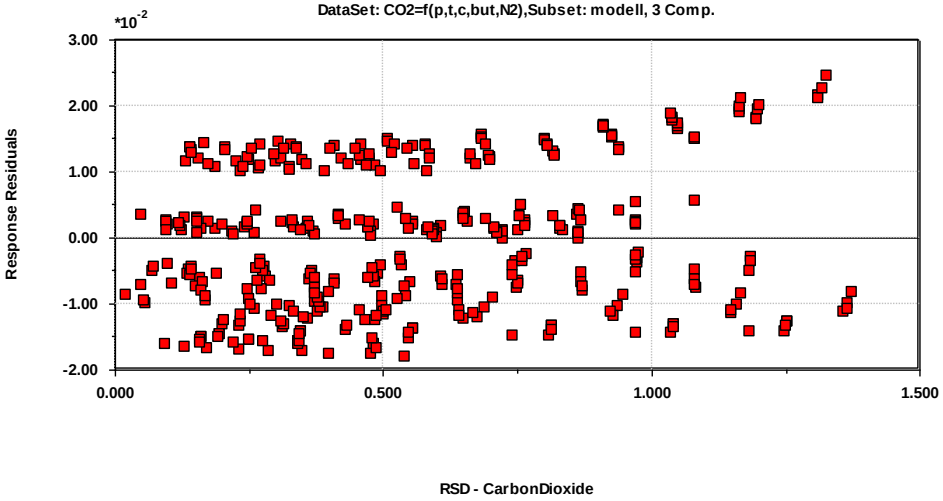


Figur 4.3.9 Plott av predikert mot målt verdi for CO_2, data frå kalibreringsdatasettet

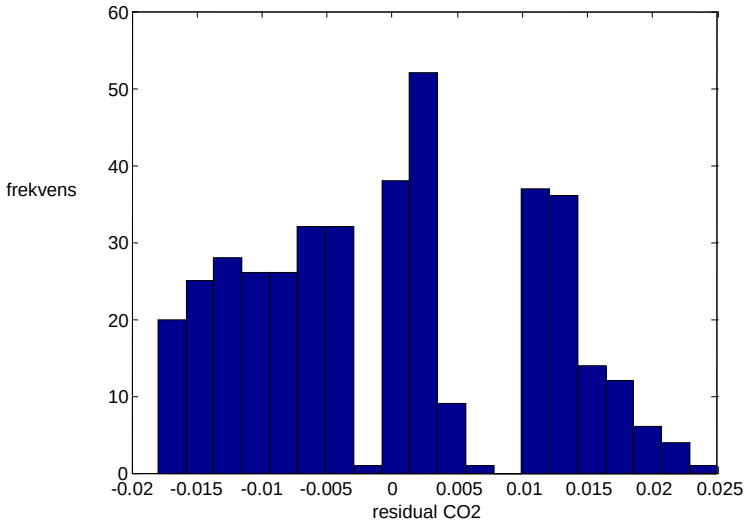
Tabellen 4.3.1 viser også at den kumulative forklarte variansen, R^2, er 0.075 og intern prediktiv evne for modellen, Q^2, er 0.071. Desse verdiane peikar mot at modellen ikkje predikerer godt, og det er heller ingen samanheng mellom målt og predikert verdi når ein testar nye objekt frå testdatasettet (figur 4.3.10). Prediksjonsfeilen for nye objekt er 0.009, dette tilsvarar heile 60 %. CO_2 blir altså predikert med stor usikkerheit. Samanhengen mellom responsresidual og RSD i figur 4.3.11 viser at objekta som har store positive residual også har høg RSD-verdi, men denne samanhengen gjeld ikkje for dei objekta med høge negative residual. Avvisningskriteriet for uteliggjarar i denne modellen er RSD > 1.727. Det er ingen objekt som har så høge RSD-verdiar for CO_2 i dette datasettet. I histogrammet i figur 4.3.12 ser ein korleis residuala fordeler seg og ein ser at det er hol i histogrammet, noko som tyder på at det er fleire fordelingar tilstades. Maksimumsfeilen er størst i positiv retning, men det er overvekt av objekt som har store negative residual.



Figur 4.3.10 Plott av predikert mot målt verdi for CO_2, predikerte data



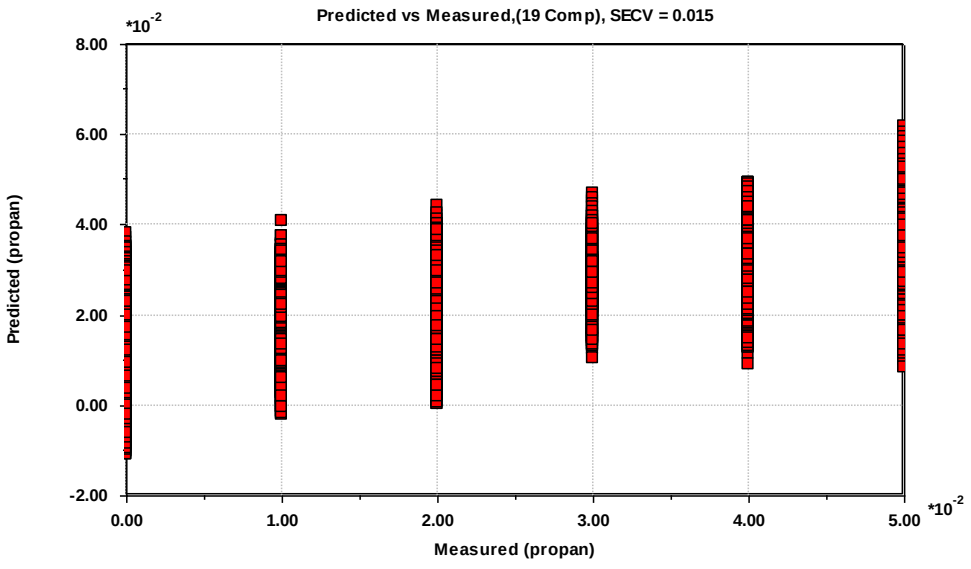
Figur 4.3.11 Plott av responsresidual mot RSD for CO_2.



Figur 4.3.12 Histogram over residuala for CO_2

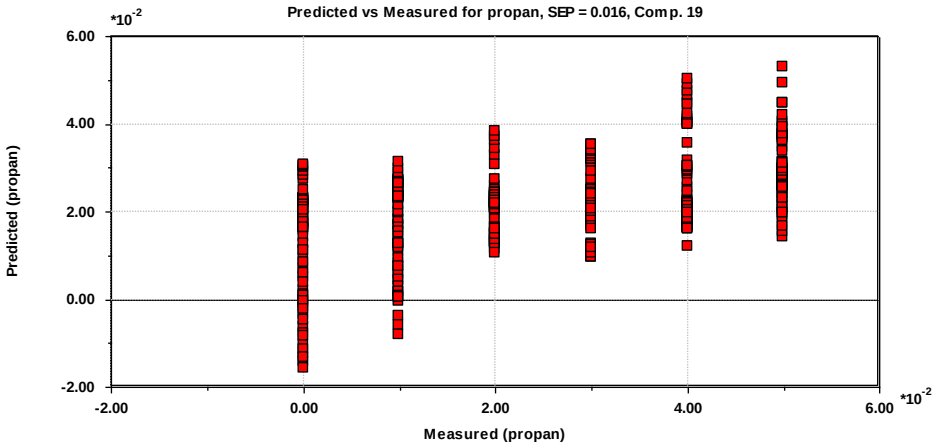
PROPAN

For propan vart det forsøkt ulike variantar for å finne den beste modellen. Det er modellen som har inkludert førstegradsledd, andregradsledd og vekselverknadsledd i **X** som klart har høgast forklart varians i y, 36.78 %, i forhold til berre 7.58 % for førstegradsleddmodellen og 13.61 % for modellen med førstegradsledd og andregradsledd inkludert. Prediksjonsfeilen er også redusert frå 60 % for førstegradsleddmodellen til 16 % for modellen med førstegradsledd, andregradsledd og vekselverknadsledd. Verdiane for R^2 og Q^2 aukar også når ein inkluderer andregradsledd og vekselverknadsledd. (Tabell 4.3.1). Modellen som vert presentert her og nytta til vidare prediksjonar er difor sistnemnte. Figurane for modellane for propan som ikkje er nytta til vidare prediksjonar er samla i appendiks, figur A.3.1 – A.3.3. For modellen med førstegradsledd og modellen med første - og andregradsledd er det mange fleire objekt som får store residual enn det som er tilfelle med modell som inkluderer førstegrad -, andregrad - og vekselverknadsledd. R^2 og Q^2 for modellen er 0.354 og 0.341. Med så låge verdiar forventar ein ikkje gode prediksjonar. Det er ikkje tilfredsstillande samanheng mellom predikert og målt verdi i modellen eller ved testing med objekta i testdatasettet, dette er vist i figur 4.3.13.

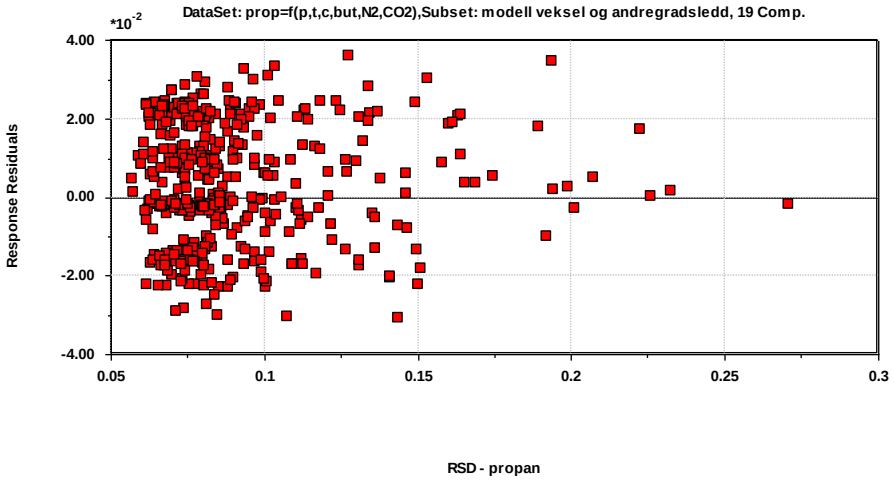


Figur 4.3.13 Plott av predikert mot målt verdi i modellen for propan

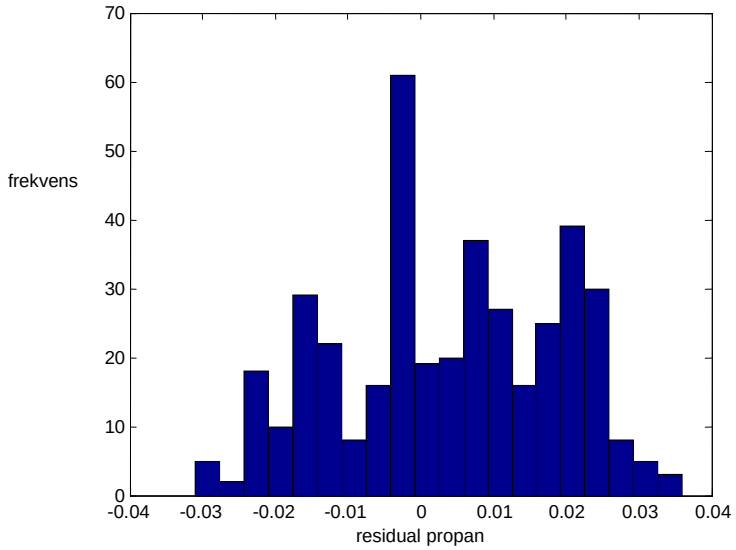
Ved prediksjon av nye objekt er likevel samanhengen mellom predikert og målt verdi noko betre, men den prediktive evna til modellen er langt frå tilfredsstillande (figur 4.3.14). Avvisningskriteriet for uteliggjarar i denne modellen er RSD > 0.233. Eit objekt har RSD over denne grensa, her er verdien 0.271. Dei fleste objekta har RSD < 0.1, og dei objekta som har høgast RSD har relativt låge residual (figur 4.3.15). Histogrammet i figur 4.3.16 viser at maksimumsfeilen er størst i positiv retning og at ein har ein tydeleg topp av residual rundt 0.



Figur 4.3.14 Plott av predikert verdi mot målt verdi av prediksjon av propan.



Figur 4.3.15 Plott av responsresidual mot RSD for predikert propan



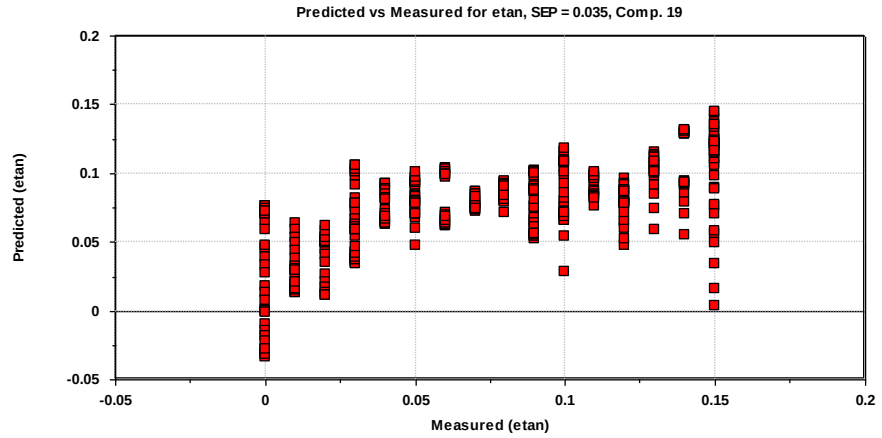
Figur 4.3.16 Histogram av residual for prediksjon av propan

Ein går vidare med dei propanprediksjonane som vart funne ved bruk av modell med andregradsledd og vekselverknader inkludert.

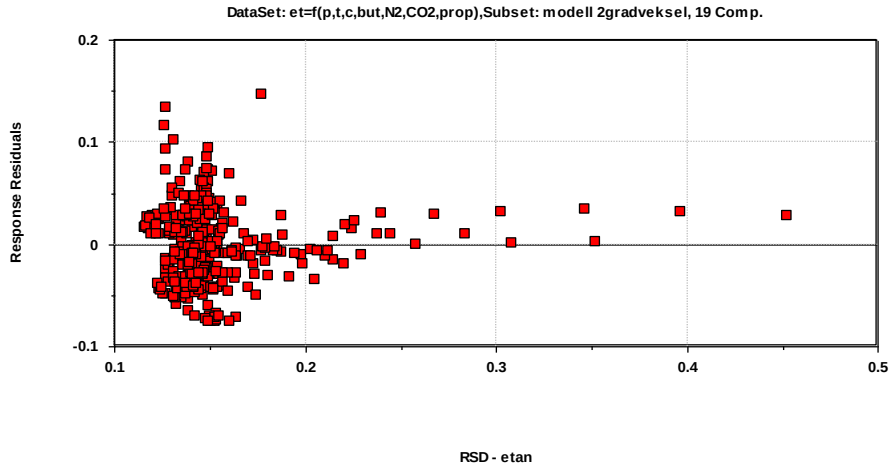
ETAN

For etan vart det laga to modellar. Den første er ein firekomponentsmodell som inneheld berre førstegradsledd og har forklart varians i y på 32.81 %. Her er gjennomsnittleg prediksjonsfeil 0.036, noko som tilsvarar 48 %. Modellen har også dårleg kumulativ forklart varians, berre 0.316, og den interne prediksjonsevna for modellen, Q^2, er 0.313. Figurar for denne modellen er samla i appendiks, figur A.3.4 og A.3.5. Det vart også bygd ein modell for etan som har både førstegradsledd, andregradsledd og vekselverknader inkludert. Denne modellen har 19 komponentar inkludert og forklart varians i y er no 79.57 %. Prediksjonsfeilen er 37 %, altså framleis ganske høg. Kumulativt forklart varians er 0.758 og intern prediksjonsevne for modellen er 0.773 (tabell 4.3.1). Dette indikerer at modellen er god. Samanhengen mellom predikert og målt verdi er vist i figur 4.3.17 og samanhengen er litt betre for den aktuelle modellen samanlikna med modellen med berre førstegradsledd. Avvisningskriteriet for uteliggjarar i denne modellen er RSD > 0.332. Fire objekt har RSD-verdi over denne grensa. Det er ingen samanheng mellom auke i responsresidual og auke i RSD. Dei objekta som har høg RSD-verdi har residual rundt 0.00 (figur 4.3.18). Histogrammet

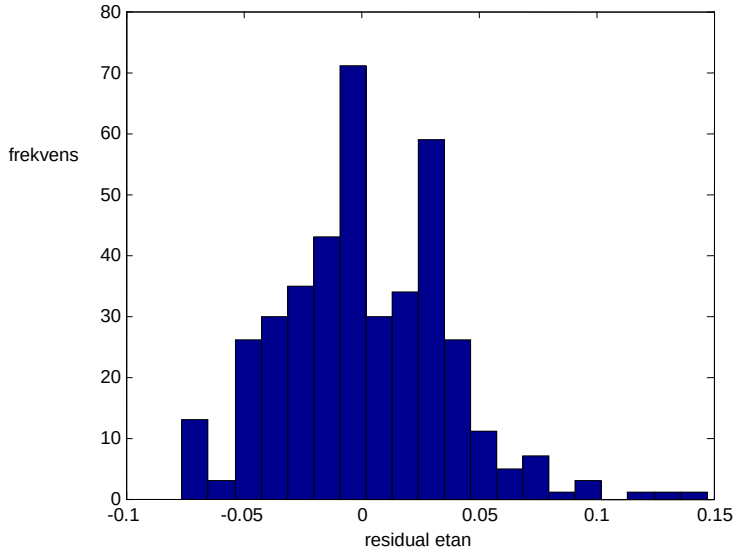
i figur 4.3.19 viser at maksimumsfeilen er størst i positiv retning. Maksimumsfeilen er litt større for denne modellen enn for modellen med berre førstegradsledd. For etan ser ein at 12 av objekta får predikert negativ verdi, noko som sjølvsagt er umogleg i ein verkeleg situasjon. Desse feila blir likevel ikkje retta på. Predikert verdi for desse objekta er i området -0.002 - -0.033 og alle objekta har målt verdi 0.000. Ved å rette desse objekt til 0.000 og ikkje tillate negative predikerte konsentrasjonar vil ein behandle feil i nedre del av området annleis enn ein behandlar feil i andre deler av området der ein tillet både positive og negative residual. Det er heller ikkje prediksjon av etan som er hovudmålet med denne prosedyren, det heile handlar om å til slutt predikere tettleik og brennverdi i naturgassen best mogleg.



Figur 4.3.17 Plott av predikert mot målt verdi for predikert etan



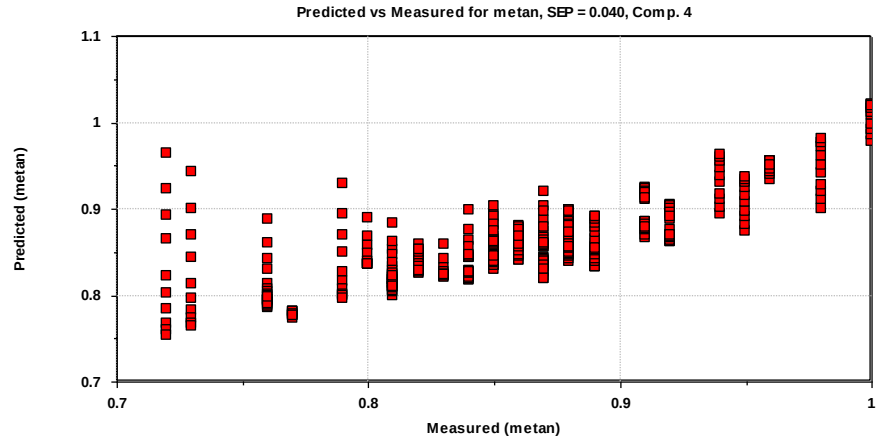
Figur 4.3.18 Plott av responsresidual mot RSD for predikert etan.



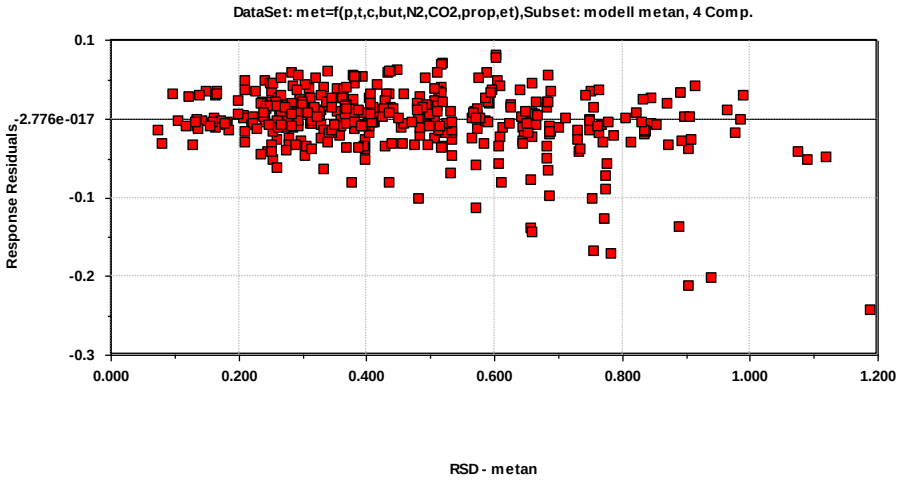
Figur 4.3.19 Histogram av residual for prediksjon av etan

METAN

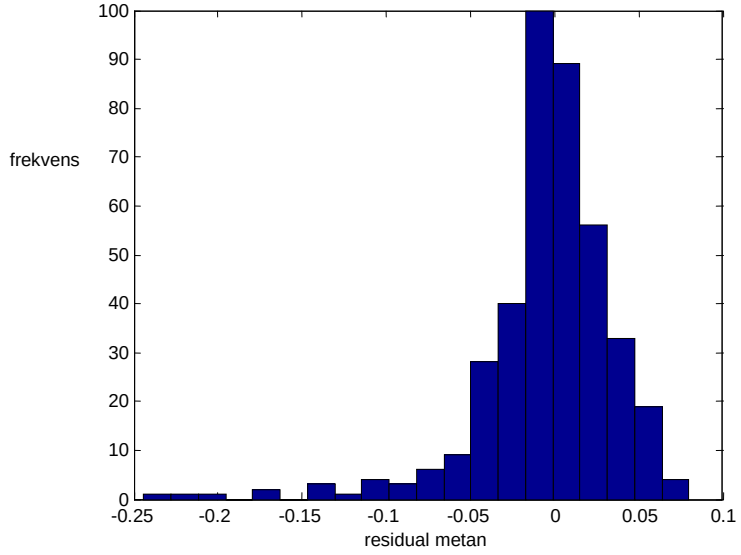
Modellen for metan er den same som er presentert i kap 4.1.3 metan. Modellen har som vist i tabell 4.3.1 fire komponentar inkludert og forklart varians i y på 99.96 %. Maksimumsfeilen ved prediksjon av nye objekt er 0.026, dette svarar til 3 %. Ein ser at plottet i figur 4.3.20 at det er ein viss samanheng mellom predikert og målt verdi ved prediksjon av nye objekt. Objekta med målt verdi i området 0.72 til 0.80 har likevel for høge predikerte verdiar. Avvisningskriteriet for uteliggjarar i denne modellen er RSD > 1.457. Ingen av objekta har RSD-verdi over denne grensa. For ein liten del av objekta er det ein samanheng mellom auke i responsresidual i negativ retning og auke i RSD, slik ein ser i figur 4.3.21. Det er ikkje tilsvarande samanheng mellom store positive residual og RSD. Ein ser av histogrammet i figur 4.3.22 at maksimumsfeilen er størst i negativ retning. Residuala ser ut til å ha tilnærma normalfordeling.



Figur 4.3.20 Plott av predikert mot målt verdi for predikert metan



Figur 4.3.21 Plott av responsresidual mot RSD for metan



Figur 4.3.22 Histogram av residual for prediksjon av metan

Ein har no forsøkt å predikere den kjemiske samansetjinga i naturgassen ved å starte med den komponenten det er minst av i gassen, nemleg butan. Deretter har ein nytta butanprediksjonen til å predikere N_2 og til slutt enda opp med ei kjemisk samansetjing der metan i dette tilfellet blir predikert med best mogleg modell i håp om at dette kan føre til betre prediksjon av tettleik og brennverdi. Ein ser at ein lukkast i å predikere metan betre ved denne framgangsmåten enn med den presentert i kapittel 4.2. Etan er den klart viktigaste variabelen i modellen for metan, dette er vist i figur 4.1.2. Ein er altså i liten grad avhengige av nøyaktige prediksjonar av konsentrasjonen av propan, butan, CO_2 og N_2 for å predikere metan tilfredsstillande.

4.4 Konstant trykk og temperatur

Mange ulike kjemiske samansetjingar kan gi dei same fysiske målingane, eit døme på dette er at mange ulike gassblandingar til dømes vil ha same tettleik sjølv om den kjemiske samansetjinga er ulik. Informasjon om den kjemiske samsetjinga i gassen er difor viktig for korrekt prediksjon av tettleik og brennverdi. Trykk og temperatur er fysiske variablar som er gitt og som ein kan endre sjølv. Lydhastigheita skil seg ut i den forstand at objekt som har identisk trykk og temperatur kan og vil ha heilt ulik lydhastigheit og kjemisk samansetjing.

Tettleik er ein funksjon av trykk og temperatur, men også av kjemisk samansetjing. Den grafiske framstillinga av dei vekta regresjonskoeffisientane for modellen for tettleik i figur 4.6.1 viser at trykk og kvadrert trykk er variablar med store bidrag i modellen for tettleik. Ein ser også eit moderat bidrag frå kvadrert temperatur. Dei kjemiske komponentane har ingen bidrag når dei opptrer som første - og andregradsledd, men når dei inngår i vekselverknader med trykk og temperatur får dei større bidrag til modellen.

Brennverdi er eit uttrykk for kor mykje energi ein får ut av gassen og er ein funksjon av den kjemiske samansetjinga. Brennverdien vil ikkje vere avhengig av variablane trykk og temperatur. Dette kjem klart fram frå den grafiske framstillinga av dei vekta regresjonskoeffisientane i figur 4.7.1.

Lydhastigheita er som ein kan sjå av likning (4.41) ein funksjon av den kjemiske samansetjinga og av variablane trykk og temperatur.

$$c = f(Cn, T, p) \qquad (4.4.1)$$

Der c er lydhastigheita, Cn er den kjemiske samansetjinga (ein homolog serie), T er temperatur (˚C) og p er trykk (bar).

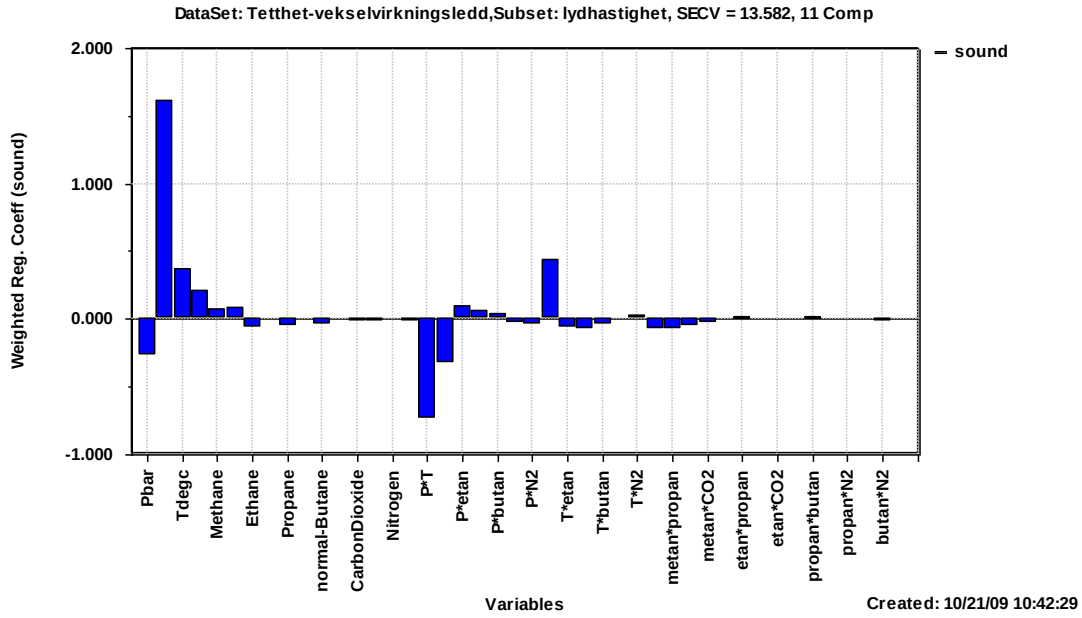
4.4.1 Modellering av lydhastigheit i Sirius

Tabell 4.4.1 Oversikt over modellar for lydhastigheit

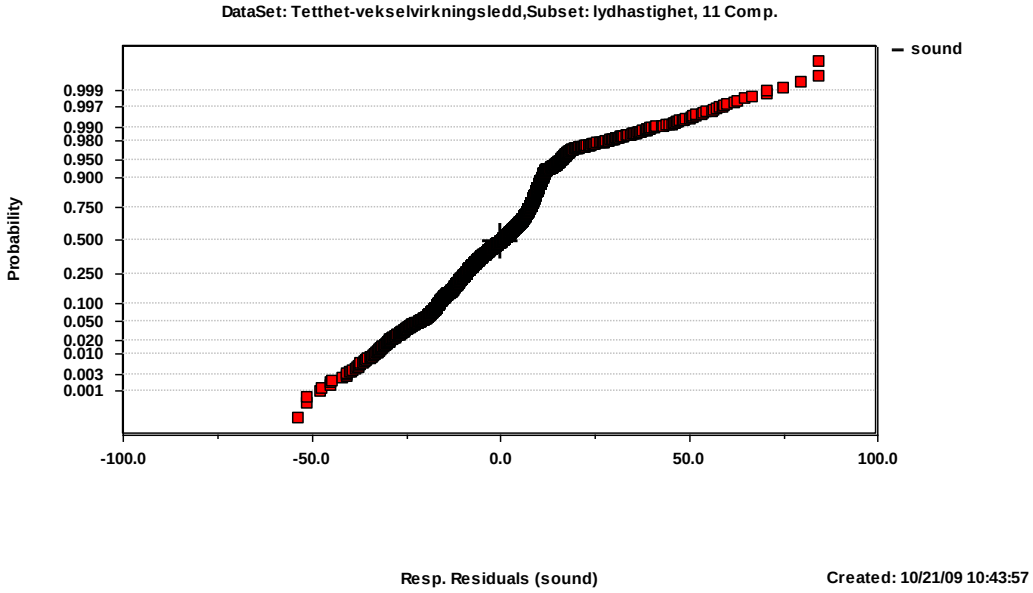
Namn	Temperatur, ˚C	Trykk, bar	Komp i modell	Forklart varians y , %
Heile området	-10 - 100	0 - 200	11	91.81
1a) Høg temperatur – lågt trykk	100	10	8	99.99
1b) Høg temperatur - høgt trykk	100	200	8	99.99
1c) Låg temperatur - lågt trykk	-10	10	8	99.99
1d) Låg temperatur - høgt trykk	-10	200	9	99.96
1e) Senterpunkt trykk- temp	60	100	8	99.99

Modellering av lydhastigheit ut frå heile kalibreringsdatasettet gir som ein kan sjå i tabell 4.4.1 ein modell med 11 komponentar og forklart varians i y på 91.81 %. Modellen inneheld både førstegradsledd, andregradsledd og vekselverknadsledd. Ser ut frå figur 4.4.1 at det er kvadrert trykk som er den variabelen med klart størst bidrag til modellen. Også vekselverknadane trykk x temperatur og temperatur x metan har bidrag, men langt frå i den grad som kvadrert trykk. Responsresiduala for objekta i modellen ikkje er heilt normalfordelte. Ein ser at plottet har ein knekk i området rundt residual på 25 m/s. Øvst til høgre i plottet finn ein nokre objekt med responsresidual som fell utanfor normalfordelinga, desse objekta har litt lågare residual enn forventa i forhold til normalfordelinga(figur 4.4.2). Resiudala er relativt store. Lydhastigheit har verdiar rundt 350-550 m/s og med residual på 50 m/s svarar dette til prediksjonsfeil på 11 %. Plottet i figur 4.4.3 viser predikert mot målt verdi for lydhastigheit. Det kan sjå ut som det er ein tendens til at prediksjonsfeilen aukar

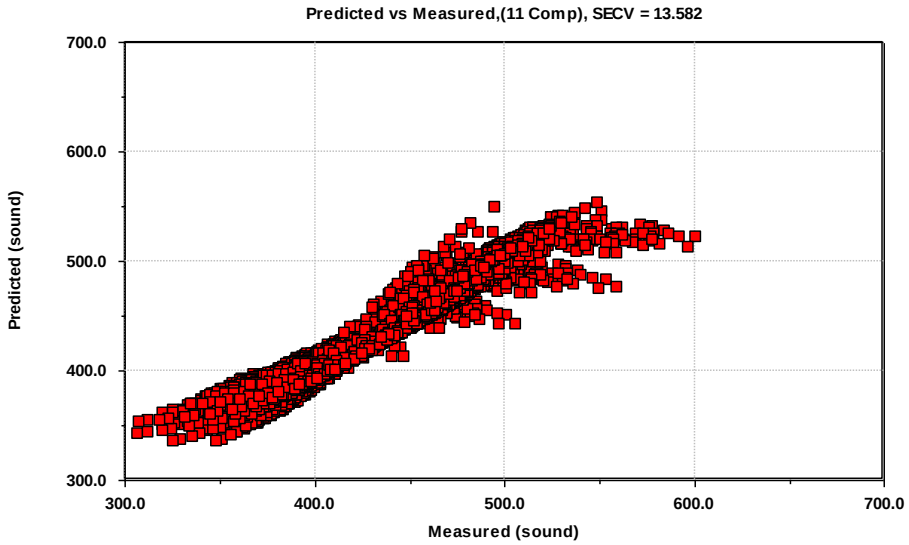
med aukande målt verdi sidan det er større spreiing i høgre del av plottet. Ut frå plottet ser det ut som det er målte verdiar på rundt 400 m/s som har dei beste prediksjonane. Objekta med lydhastigheit over 500 m/s er i mange tilfelle predikert for lågt.



Figur 4.4.1 Grafisk framstilling av dei vekta regresjonskoeffisientane i modellen for lydhastigheit.

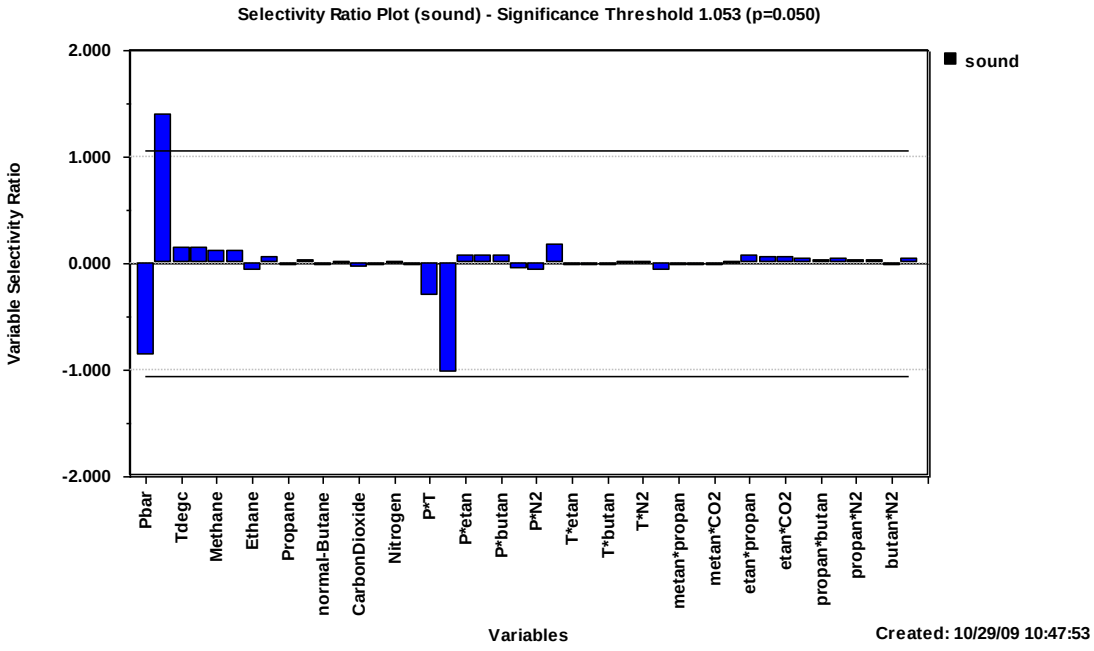


Figur4.4.2 Normalfordelingsplott for responsresiduala i modellen for lydhastighet.



Figur 4.4.3 Plott av målt verdi mot predikert verdi for lydhastigheit

Det vart utført targetprojeksjon av modellen for lydhastigheit. Selektivitetsratioplottet vist i figur 4.4.4 viser at det er kvadrert trykk som er den viktigaste variabelen i positiv retning. Trykk og vekselverknadsleddet trykk x metan er størst i negativ retning. Her er terskelen sett til litt over 1.000 av Sirius. Det er altså kvaderert trykk som er den viktigaste variabelen for modellen.



Figur 4.4.4 Selektivitetsratioplott for lydhastigheit

Modellen for lydhastigheit er ikkje nytta til prediksjon. Den er med for å illustrere korleis lydhastigheita kan forklarast ut frå datasettet. Dette er interessant fordi ein då kan få kunnskap om kva variablar som i størst grad påverkar lydhastigheita i naturgassen. Resultata frå modellering av lydhastigheit viser at variablane kvadrert trykk og trykk er dei viktigaste variablane i forhold til lydhastigheit.

Som tabell 4.4.1 viser er blir det gode modellar for lydhastigheit i alle områda som er testa til no. Alle modellane har inkludert førstegradsledd, andregradsledd og vekselverknadsledd.

Under ultralydmålingar av lydhastigheita vil ein offshore i norsk sektor ofte finne trykk på over 100 bar, i enkelte tilfeller også over 200 bar. I Europa er derimot trykket ofte noko lågare, her ligg trykket ofte rundt 60-70 bar. Når det gjeld temperatur ser ein ofte at denne ligg mellom 0 og 60 ˚ C, men også her finst unntak.

Ein går vidare med undersøkingar av senterpunktområdet som har temperatur 60 ˚C og trykk 100 bar. Modellen for lydhastigheit i dette området inneheld både førstegradsledd, andregradsledd og vekselverknadsledd. Ein såg under modellering at ein fekk høg forklart varians i y med modell som berre inneheld førstegradsledd, men responsresiduala vart ikkje normalfordelte. Kalibreringsdatasettet som er brukt til modellering inneheld ikkje støy. Som nemnt tidlegare er det dermed ikkje nødvendigvis normalfordelte responsresidual viktige for validering av modellen.

4.4.2. Modellering av kjemisk samansetjing i Sirius

I ein reell målesituasjon kan ein kan tenkje seg ein utgangspunkt der ein måler lydhastigheita medan trykket og temperaturen vert halde konstant. Ein slik situasjon er simulert her ved at ein i modelleringa og prediksjonane av den kjemiske samansetjinga har nytta senterpunktområdet beskrive i kapittel 4.4.1 (trykk 100 bar og temperatur 60 ˚C). Iterativ konsentrasjonsbestemming ved AR er utført med åtte objekt frå testdatasettet. Trykk og temperatur i testdatasettet er ikkje nøyaktig likt som for dei objekta modellane er bygd på. Trykk for objekta i testdatasettet var 90 bar og 110 bar. Temperatur for objekta i testdatasettet var 67 ˚C. Testdatasettet består av dei objekta som har desse aktuelle verdiane, totalt 16 objekt.

Modellering av dei ulike kjemiske komponentane er samla i tabellane A.4.1 – A.4.7. Den totale forklarte variansen i y for modellen for metan aukar når ein inkluderer fleire kjemiske komponentar. Maksimumsfeilen etter testing i Sirius med data frå testdatasettet minkar når ein legg til etan og held seg om lag lik når ein også legg til propan og CO_2. Den minkar så når ein inkluderer N_2 og butan. Maksimumsfeilen etter iterativ konsentrasjonsbestemming ved AR aukar når ein skal predikere på fleire komponentar samtidig. For modellen som inneheld andregradsledd og vekselverknadsledd ser ein at forklart varians i y i modellen aukar, men feilen etter iterativ konsentrasjonsbestemming ved AR blir større. Desse resultata peikar mot at det ikkje svarar seg å inkludere andregradsledd og vekselverknadsledd i modellane.

Den totale forklarte variansen i y for modellen for etan aukar mykje ved å inkludere metan i modellen. Maksimumsfeilen blir litt større i negativ retning, men området for residuala blir mindre jo fleire kjemiske komponentar som blir inkludert i modellen. I den siste modellen ser ein at residuala er blitt positive. Ved iterativ konsentrasjonsbestemming ved AR ser ein at maksimumsfeilen aukar når ein inkluderer propan i tillegg til lydhastigheit og metan. Også for etan aukar den forklarte variansen i y for modellen når ein inkluderer andregradsledd og vekselverknadsledd, men feilen blir større.

Forklart varians i y for modellen for propan aukar når ein inkluderer fleire komponentar. Området som feilen etter testing i Sirius er i minkar også når ein inkluderer fleire komponentar. Alle objekta predikerer til fast verdi når ein forsøker iterativ konsentrasjonsbestemming ved AR.

I modellen for CO_2 aukar den forklarte variansen i y mykje ved å inkludere N_2 i modellen. Området for feilen etter test i Sirius blir også betrakteleg mykje mindre. Ein får også ein betre modell ved å inkludere butan. Iterativ konsentrasjonsbestemming ved AR går frå å predikere alle objekta til 0.00 til å predikere alle objekta til 0.03 og tilbake til å predikere alle objekta til 0.00.

Forklart varians i y for modellen for N_2 aukar litt når ein inkluderer butan. Området for feilen etter test i Sirius blir også mindre. Alle objekta blir predikert til 0.00 etter iterativ konsentrasjonsbestemming ved AR, ein ser difor ikkje noko endring i maksimumsfeil etter AR når ein inkluderer butan i modellen.

Tabell 4.4.2 Oversikt over modellar nytta til prediksjon av den kjemiske samansetjinga, data frå testdatasett.

Modell	Forklart varians i y %	Maksimumsfeil etter test i Sirius		Maksimumsfeil etter AR		AR
		pos	neg	pos	neg	
Metan=f(c)	89.69	0.01	-0.036			
Etan=f(c)	61.67	0.017	-0.023			
Metan=f(c,etan)	97.40	-0.005	-0.025	-0.02	-0.07	
Etan=f(c,metan)	90.35	-0.007	-0.035	0.043	0	
Propan=f(c,metan, etan)	62.30	0.019	-0.017	0.05	0	0.00
CO_2=f(c,metan,etan, propan)	46.54	0.013	-0.003	0.02	0.01	0.00
N_2=f(c,metan,etan, propan, CO_2)	98.09	-0.003	-0.012	0.03	0	0.00
Butan=f(c,metan,etan, propan,CO_2,N_2)	96.95	0.01	0.003	0	-0.02	0.02

Ut frå resultata vist i tabellane 4.4.2 ser ein at ved å utføre iterativ konsentrasjonsbestemming ved AR blir området for maksimumsfeil større enn det vert ved å predikere i Sirius. Ein ser også at når ein skal iterere på fleire enn to komponentar får alle objekta same verdi for dei enkelte komponentane. Når ein sender inn middelverdiar som startverdiar for alle dei kjemiske komponentane og utfører iterativ konsentrasjonsbestemming ved AR med seks likningar får ein desse konsentrasjonane: Metan: 1.00 for alle objekta, etan: 0.15 for alle objekta, propan: 0.00 for alle objekta, CO_2: 0.00 for alle objekta, N_2: 0.00 for alle objekta, butan: 0.02 for alle objekta. Dette har samanheng med størrelsen og forteiknet på regresjonskoeffisientane som er henta frå modellering i Sirius. Dette viser at iterativ konsentrasjonsbestemming i AR ikkje fører til betre prediksjonar og difor er unødvendig i denne samanheng.

4.4.3 Prediksjon av kjemisk samansetjing ved konstant trykk og temperatur

Den kjemiske samansetjinga ved konstant trykk og temperatur blir funnen ved å predikere ein og ein komponent, slik det er vist i kolonne 1 i tabell 4.4.3.

Tabell 4.4.3 Residual og gjennomsnittfeil for kjemiske komponentar funne ved prediksjon.

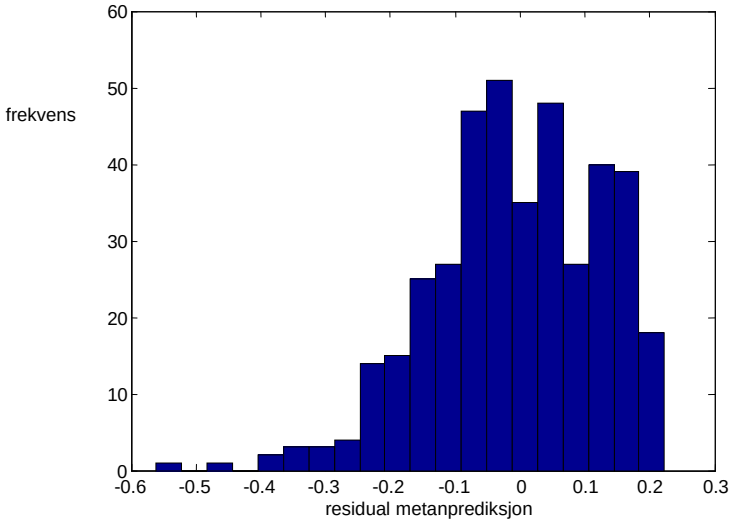
	Maksimumsfeil		Predik-sjonsfeil	feil %
	pos	neg		
Metan=f(c)	0.251	-0.162	0.095	11.0
Etan_pred=f(c, metan_pred)	0.065	-0.046	0.019	27.1
Propan_pred=f(c,metan_pred, etan_pred)	0.019	-0.027	0.011	44.0
CO_2_pred=f(c,metan_pred, etan_pred,propan_pred)	0.013	-0.016	0.008	66.7
N_2_pred=f(c,metan_pred, etan_pred,propan_pred,CO_2_pred)	0.030	0	0.016	100.0
Butan_pred=f(c,metan_pred, etan_pred,propan_pred,CO_2_pred,N_2_pred)	0.009	-0.015	0.007	46.7

Metan blir predikert først og ein ser frå tabell 4.4.3 at prediksjonsfeilen er 11.0 %. For etan er prediksjonsfeilen i størrelsesorden 27.1 %. Prediksjonen av dei kjemiske komponentane blir altså ikkje spesielt gode ved å halde trykk og temperatur konstant. Når ein ser på modellar av kjemiske komponentar slik som er presentert i kapittel 4.1 ser ein at trykk og temperatur har svært små bidrag til modellen. Det som er viktig for modellane for dei kjemiske komponentane er bidraga frå dei andre kjemiske komponentane, noko som er naturleg sidan dette er eit lukka system der den kjemiske samansetjinga i gassen summerer til 100 %. Når metan blir predikert med stor feil vil denne feilen forplante seg vidare. Dette ser ein skjer i tabell 4.4.3 der prediksjonsfeil uttrykt i prosent aukar etter kvart som fleire komponentar blir predikert. Butan er unntaket frå dette. Viktig også å merke seg her at små prediksjonsfeil gir store utslag når ein reknar dei om til prosent.

Prediksjon av kjemisk samansetjing for heile datasettet ut frå modellar med konstant trykk og temperatur

For å undersøke kor stor effekt variablane trykk og temperatur har på modellane testa ein ved å predikere alle objekta frå både kalibreringsdatasett og testdatasett med modellane for senterpunktområdet.

Histogrammet i figur 4.4.5 viser residuala for objekta i testdatasettet etter prediksjon av metan med modell med konstant trykk og temperatur, og ein ser tydeleg at det er større spenn i dei negative residuala enn dei positive. Hovudtyngda av residual finn ein mellom -0.2 og 0.2. Dette er høge residual når målt verdi er mellom 0.72 og 1.00. Konklusjonen blir her at sidan trykk og temperatur har noko å seie for lydhastigheita vil desse variablane også påverke prediksjonane, ein må difor inkludere trykk og temperatur i modellane.



Figur 4.4.5 Histogram over residuala etter prediksjon av metan

4.5 Metan, etan og addert kjemisk samansetjing (aks)

4.5.1 Modellering og validering av metan, etan og aks

Ein ynskjer å undersøkje om iterativ konsentrasjonsbestemming ved AR kan fungere betre dersom ein har færre variablar tilstades. Ein slår difor saman dei fire kjemiske komponentane som utgjer dei minste bestanddelane i naturgassen til ein variabel kalla aks. Variablane nytta i modelleringa er trykk, temperatur og lydhastigheit samt dei kjemiske komponentane metan, etan og aks, der variabelen aks er summen av propan, butan, CO_2 og N_2.

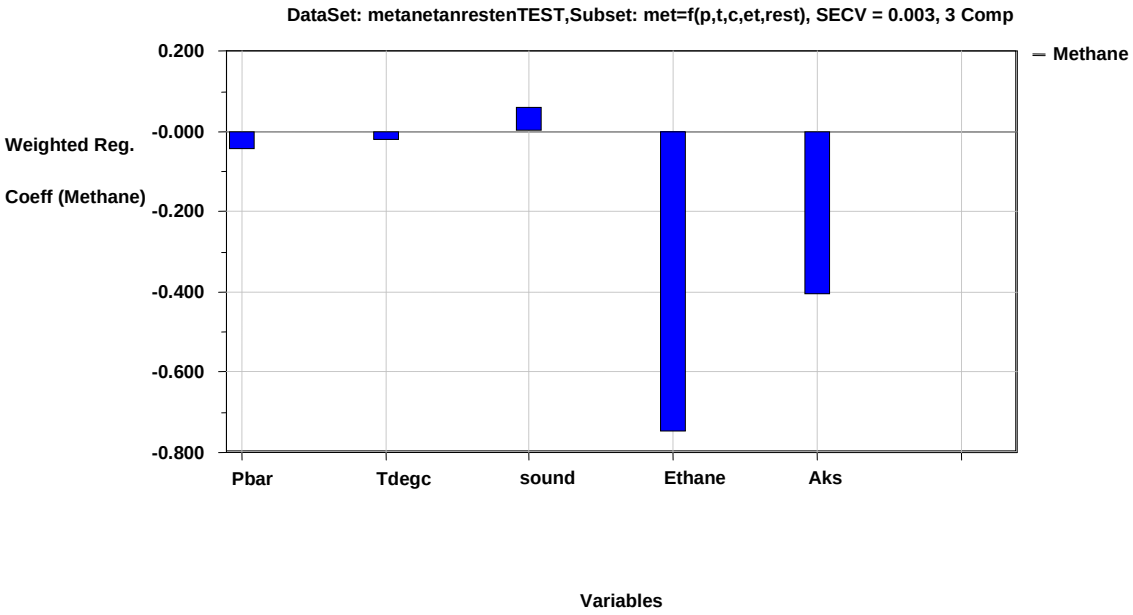
Tabell 4.5.1 viser dei ulike modellane for metan, etan og aks

Modell	Komp	Forklart varians y, %	RSD >	Feil	Feil %	R^2	Q^2
Metan=f(p,t,c,etan,aks)	3	99.84	1.968	0.002	0.23	0.998	0.998
Metan=f(p,t,c,etan,aks)	4	100.00					
Etan=f(p,t,c,metan, aks	3	99.83	1.965	0.001	1.33	0.999	0.999
Aks =f(p,t,metan,etan)	3	92.87	1.333	0.005	7.69	0.927	0.925

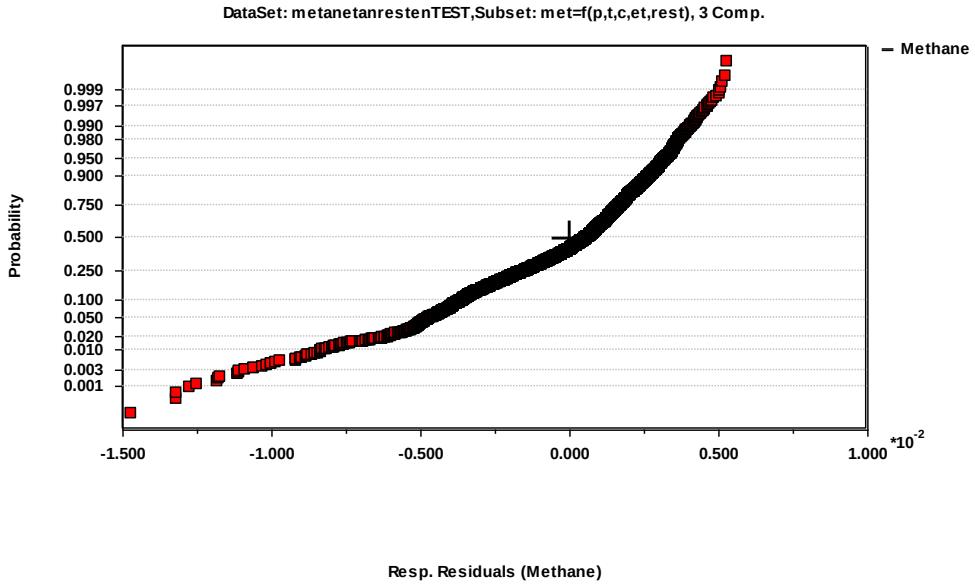
METAN

Kryssvalidering og SECV-plott indikerte at ein kunne inkludere fire komponentar i modellen for metan, då får ein forklart 100.00 % av variansen i y. Det har likevel vist seg tidlegare at å ha 100.00 % forklart varians i y kan skape problem med prediksjonen då regresjonskoeffisientane Sirius gir ut får verdiane 1.000 for både etan og aks. Difor vel ein å gå ned i graden av forklart varians i y og vel ein modell som har tre komponentar og forklarer 99.84 % av variansen i y. Dermed blir regresjonskoeffisientane meir reelle og kan nyttast til prediksjon og iterasjon ved bruk av alternerande regresjon. Framleis har ein høg grad av forklart varians i for y i modellen. Etan og aks er dei variablane som har mest å bety for modellen, dette er vist i figur 4.5.1. Trykk, temperatur og lydhastigheit har beskjedne bidrag. Normalfordelingsplottet av responsresiduala har ein hale nede til venstre i plottet slik ein kan sjå i figur 4.5.2. Det vil seie at ein har eit avvik frå normalfordelinga. Det er lagt lite vekt

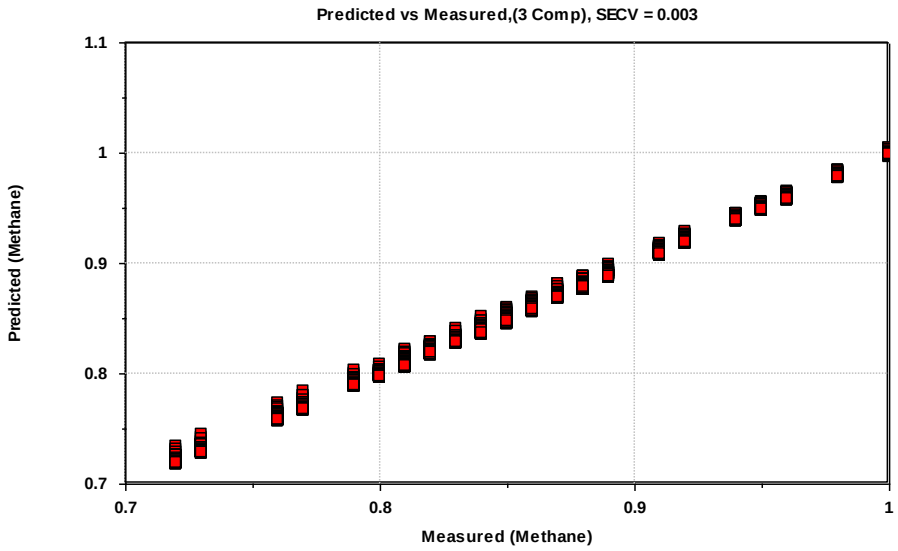
på normalfordelingsplotta i denne samanhengen sidan det ikkje er støy tilstades i datamaterialet. Det er god samanheng mellom målt og predikert verdi for metan og dette er eit betre mål på kor god modellen er enn normafordelingsplott, figur 4.5.3.



Figur 4.5.1 Grafisk framstilling av vekta regresjonkoeffisientar for modell for metan

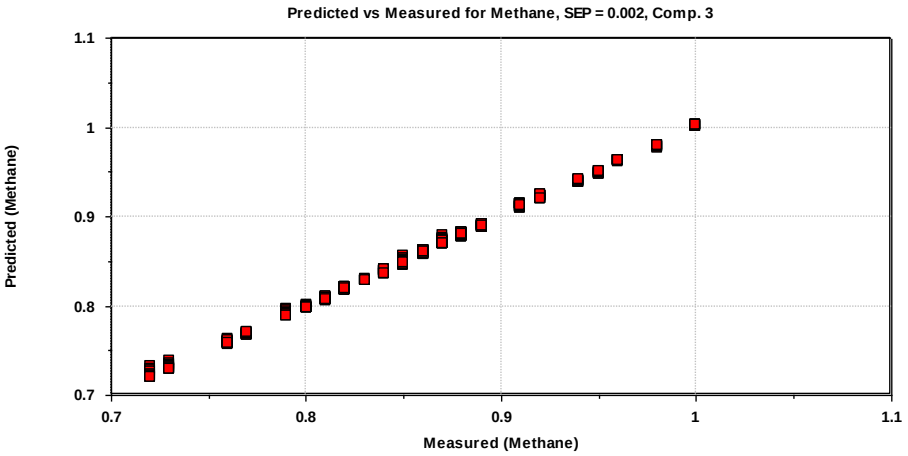


Figur 4.5.2 Normalfordelingsplott av responsresiduala i modellen av metan.

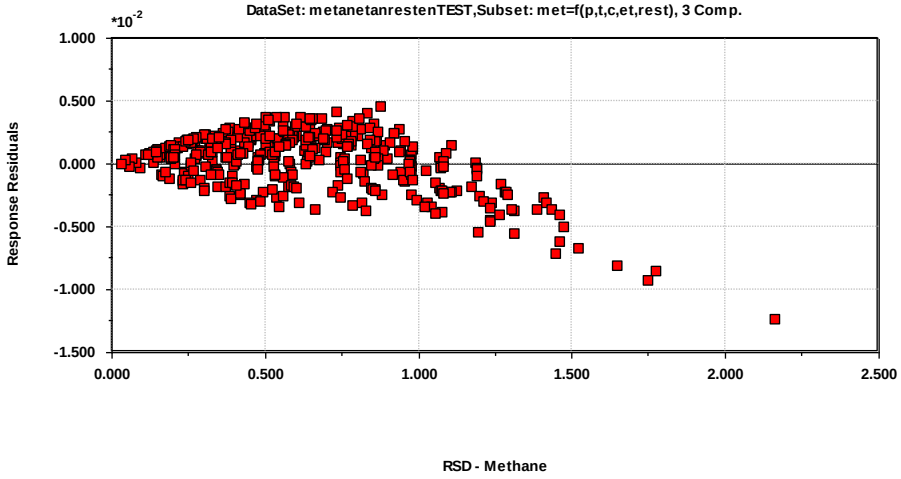


Figur 4.5.3 Plott av predikert verdi for metan mot målt verdi for metan.

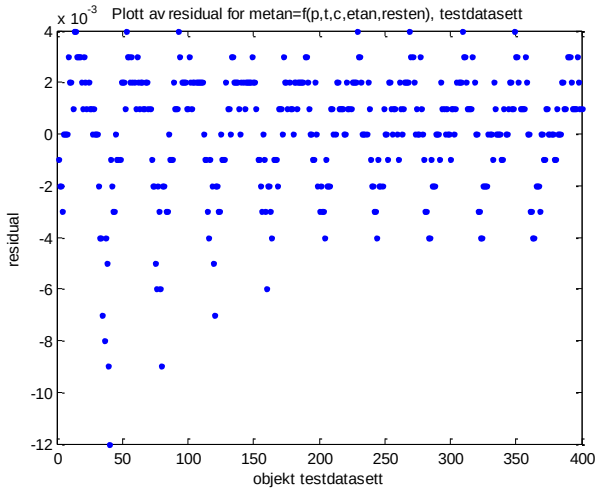
Residuala etter testing av modellen er ganske små, hovudtyngda av residual finn ein i området -0.003 til 0.003. Det er også svært bra samanheng mellom målt og predikert verdi for metan når ein testar med objekt i testdatasettet, dette er vist i figur 4.5.4. Ein ser frå figur 4.5.5 at det er samanheng mellom aukande RSD og aukande responsresidual, både i positiv og negativ retning. Dette er ikkje uventa sidan RSD beskriv avstanden objektet har til modellen og responsresidualet beskriv avstanden mellom predikert og målt verdi for objektet. Stor avstand til modellen vil ha samanheng med stor avstand til målt verdi for objektet. Kriteriet for at eit objekt skal avvisast som uteliggarar i denne modellen er RSD > 1.968. Det er berre eit objekt som fell utanfor dette kriteriet. Når RSD-verdiane overstig omlag 1.1 er det berre negative residual tilstades. Plottet viser at dei objekta som har høgast negative residual har høgast RSD-verdiar. Residuala er størst for dei objekta som har låge trykk og høge temperaturar, dette ser ein av plottet i figur 4.5.6. R^2-verdien for den tredje komponenten er 0.998 og det same er Q^2 verdien. (Tabell 4.5.1). Desse resultata indikerer at denne modellen er bra sjølv om responsresiduala for modellen ikkje er heilt normalfordelte. Prediksjonsfeil for objekta i testdatasettet er 0.23 %, noko som er svært bra.



Figur 4.5.4 Plott av predikert verdi mot målt verdi for metan



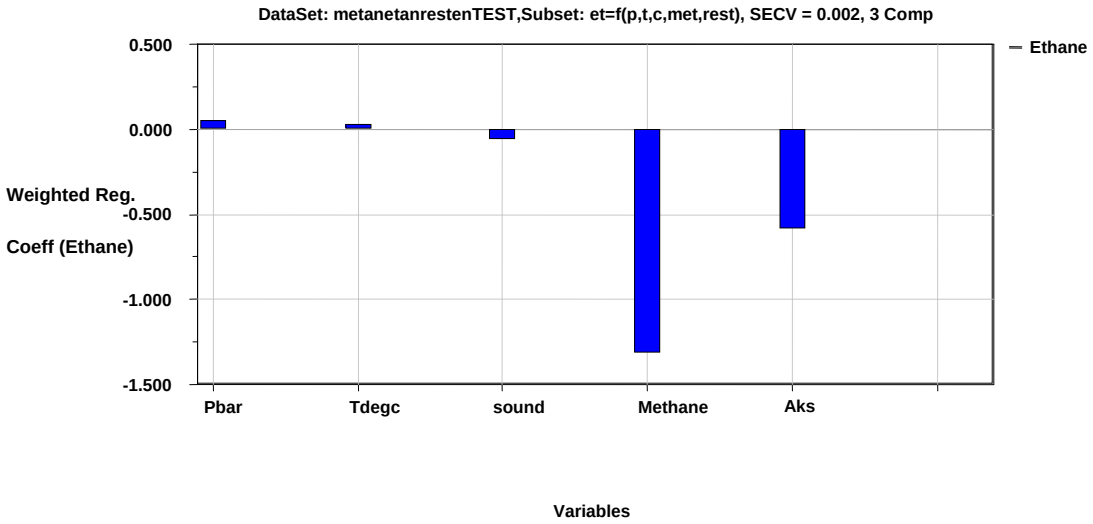
Figur 4.5.5 Plott av RSD for metan mot responsresidual for metan



Figur 4.5.6 Plott av residuala mot objekta i testdatasettet.

ETAN

For etan kunne ein, på same måte som for metan, oppnådd forklart varians i y på 100.00 % dersom ein hadde inkludert fire komponentar i modellen. På same grunnlaget som for metan valde ein å laga ein modell for etan som inkluderte tre komponentar. Det er variablane metan og aks som har mest å bety for modellen. Variablane trykk, temperatur og lydhastigheit har svært liten innverknad på modellen for etan slik ein kan sjå ut frå den grafiske framstillinga av dei vekta regresjonskoeffisientane i modellen for etan i figur 4.5.7.

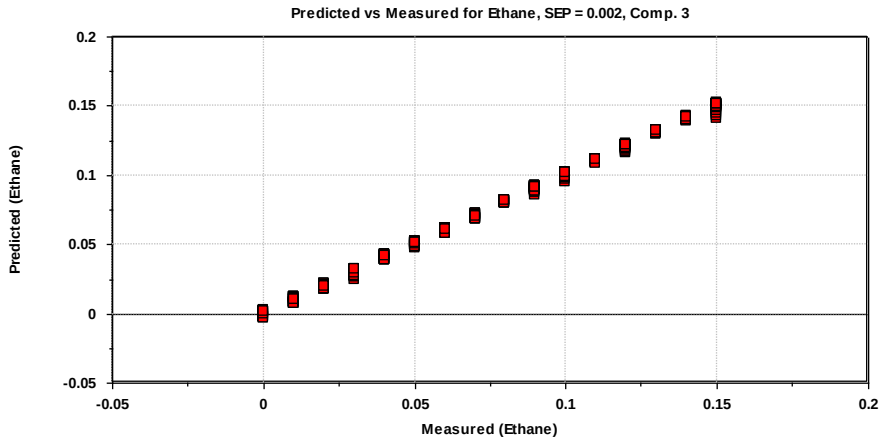


Figur 4.5.7 Grafisk framstilling av regresjonskoeffisientane i modellen for etan.

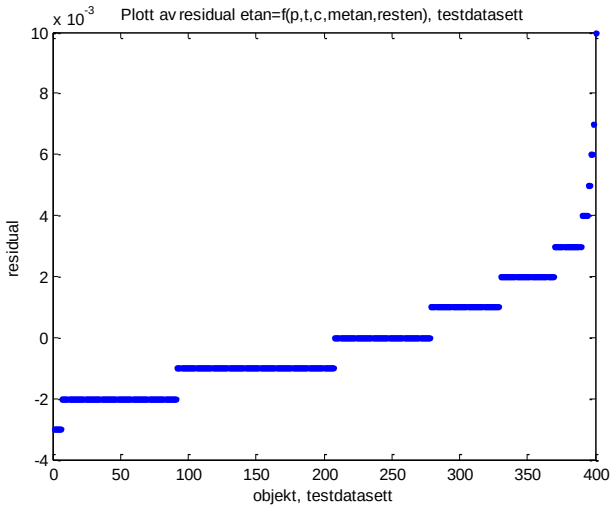
Ein ser av tabell 4.5.1 at forklart varians i y i modellen er 99.83 %, dette er ein høg grad av forklart varians. Når ein testar med nye objekt frå testdatasettet ser ein at det er svært bra samanheng mellom predikert og målt verdi (figur 4.5.8). Hovudtyngda av residuala for etanprediksjon er i området -0.003 til 0.003. Det finst enkelte store positive residual tilstades, på same måte som ein for metan kunne sjå enkelte store negative residual. Residuala aukar med aukande temperatur. Det er ei også ei auke i residual med aukande trykk, dette er vist i figur 4.5.9. Det er tydeleg samanheng mellom aukande RSD og aukande responsresidual slik ein kan sjå av plottet i figur 4.5.10. Når ein kjem til objekt med RSD høgare enn ca 1 ser ein at objekta har utelukkande positive residual. Tabell 4.5.1 viser at kriteriet for at eit objekta kan avvisast som uteliggjarar i denne modellen er RSD > 1.965. Ser på resultata og ser at berre fire objekt har RSD > 1.965 og dei verdiane ein ser er rett over

grensa. Både R^2 og Q^2 for modellen er 0.999 dette er også faktorar som peikar mot at modellen er god. Prediksjonsfeil for objekta i testdatasettet er 1.33 %.

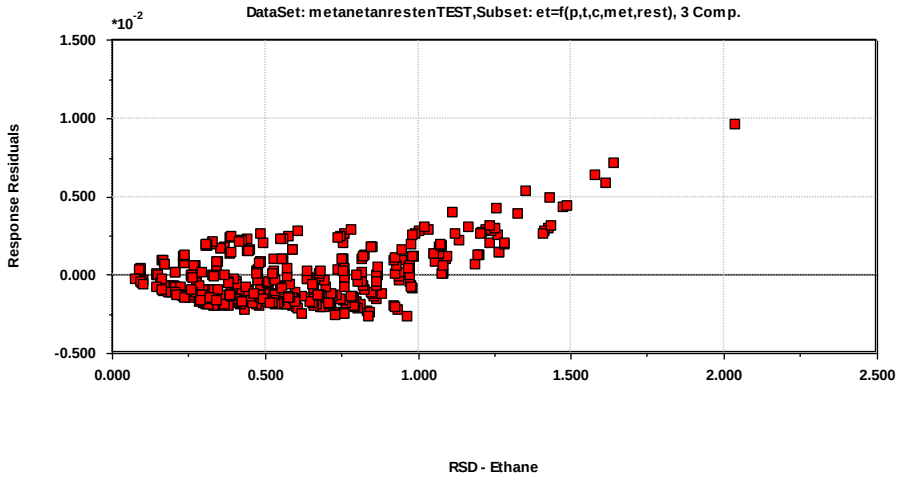
Etan vart også modellert med vekselverknadsledd inkludert, men sidan både forklart varians i y og testresultat vart likt som for modellen over som berre inneheld førstegradsledd, valde ein å behalde den enklaste modellen, nemleg den med berre førstegradsledd inkludert.



Figur 4.5.8 Plott av predikert verdi for etan mot målt verdi for etan.



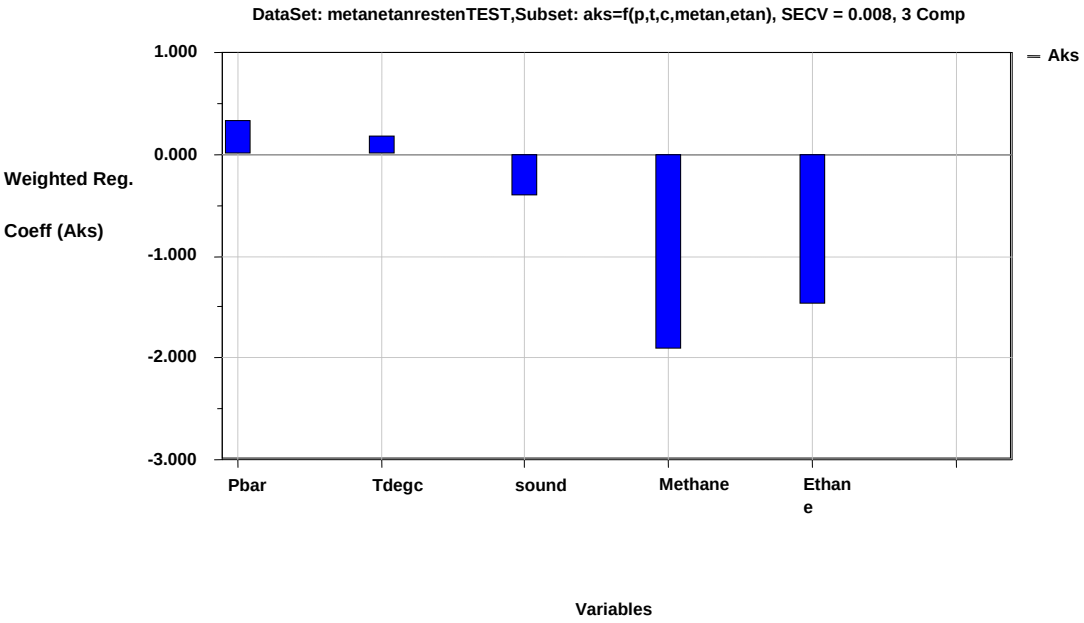
Figur 4.5.9 Plott av residual for etan mot objekt



Figur 4.5.10 Plott av RSD mot responsresidual for etan.

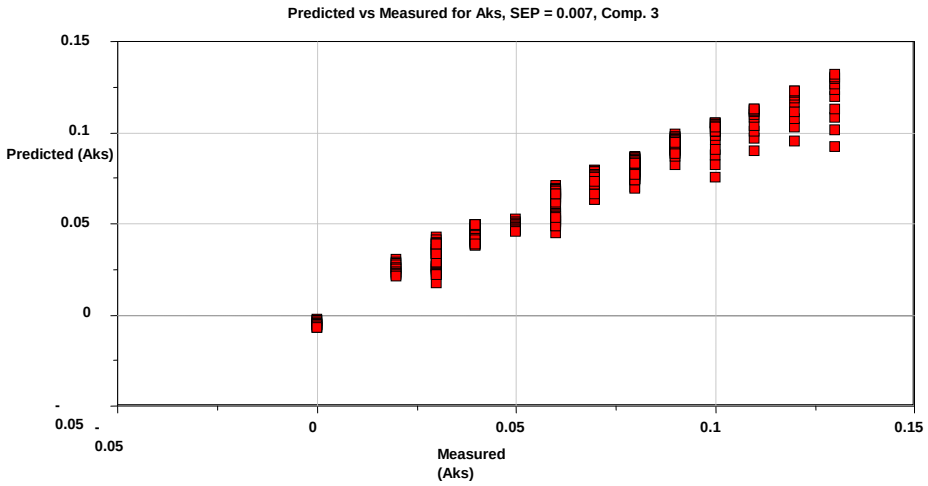
AKS

Variabelen aks består av summen av variablane propan, butan, CO_2 og N_2. Variabelen har verdiar i området 0.00 til 0.13, den utgjer altså mellom 0 og 13 % av naturgassen. Modellering av aks gav den dårlegaste modellen av dei tre som er omtalt i dette kapittelet, som ein kan sjå av tabell 4.5.1 er forklart varians i y nede i 92.87 %. Dersom ein hadde inkludert ein komponent til i modellen ville forklart varians i y auka til 100.00 %. Då ville ein fått same problem som beskrive for metan, nemleg at regresjonskoeffisientane får verdiar 1.000 for både metan og etan. Ein beheld difor modellen med tre komponentar. Det er variablane metan og etan som har mest å bety for modellen, begge bidreg i negativ retning. Trykk og temperatur bidreg i positiv retning og lydhastigheit har negativt bidrag, men alle desse tre har små bidrag når ein samanliknar med metan og etan (figur 4.5.11).

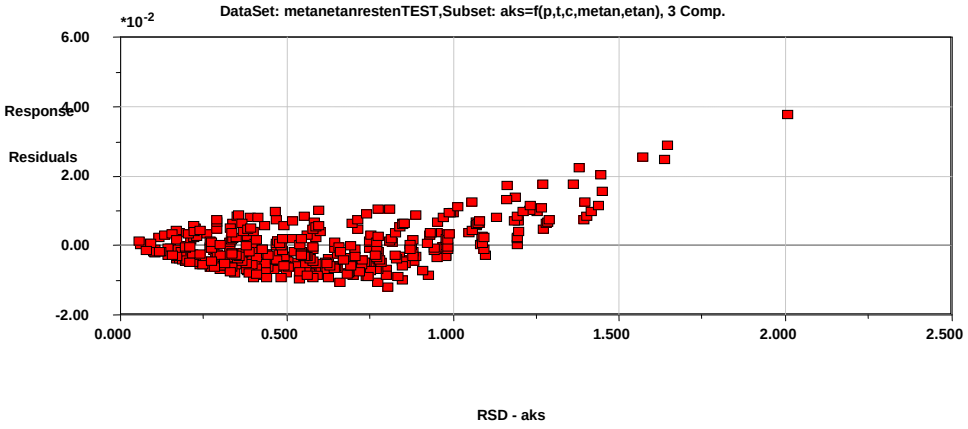


Figur 4.5.11 Grafisk framstilling av regresjonskoeffisientane for modellen for aks.

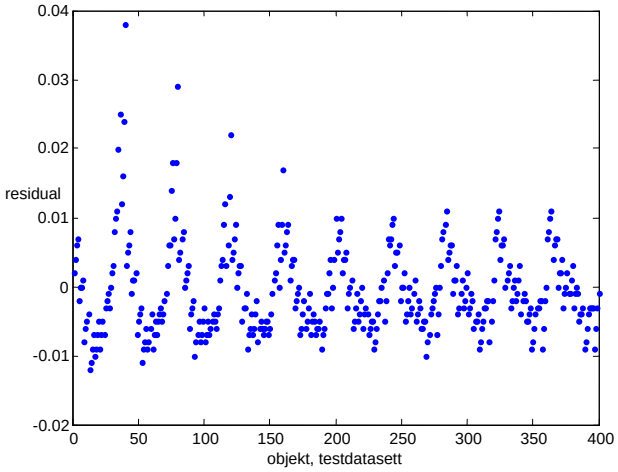
I denne modellen ser ein at det er relativt bra samanheng mellom målt og predikert verdi (figur 4.5.12). Enkelte av verdiane har likevel større residual enn andre og det ser ut til å vere ein tendens til at feilen blir større for høgare verdiar av aks. Det er objekta med låg temperatur og høgt trykk som har dei største residuala, dette er vist i figur 4.5.14. Hovudtyngda av residuala for aks finn ein i området -0.01 til 0.01. Det er ein klar samanheng mellom aukande responsresidual og aukande RSD også for denne variabelen (figur 4.5.13), slik ein kan sjå for både metan og etan. Avvisningskriteriet for uteliggjarar i denne modellen er RSD > 1.333. 13 objekt ligg over denne grensa, dei fleste ligg rett over. Det objektet som har høgast RSD har verdien 2.010. Tabell 4.5.1 viser at R^2 for modellen er 0.927 og Q^2 er 0.925, dette indikerer at dette ein relativt bra modell når det gjeld prediktiv evne. Prediksjonsfeil for variabelen aks når ein testar på objekta i testdatasettet i størrelsesorden 7.7 %. Det er altså klart størst prediksjonsfeil for aks samanlikna med metan og etan. Dette heng saman med at aks har dårlegast forklart varians i y for modellen og at R^2 og Q^2 er lågare for modellen for aks enn for modellane for metan og etan.



Figur 4.5.12 Plott av målt verdi for resten mot predikert verdi for aks.



Figur 4.5.13 Plott av responsresidual mot RSD for variabelen aks



Figur 4.5.14 Plott av residual for aks for objekt frå testdatasett

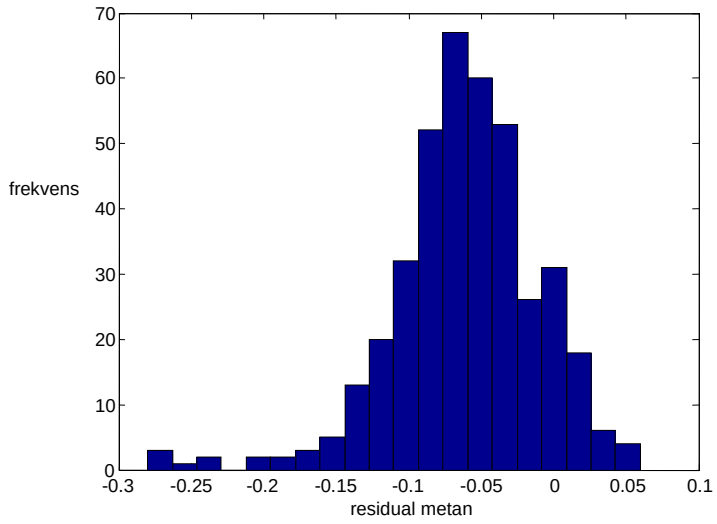
4.5.2 Iterativ konsentrasjonsbestemming av kjemisk samansetjing ved AR

Modellane som er presentert i kapittel 4.5.1 skal no nyttast til iterativ konsentrasjonsbestemming ved AR for å finne ei predikert kjemisk samansetjing som igjen skal nyttast til prediksjon av tettleik og brennverdi i naturgass.

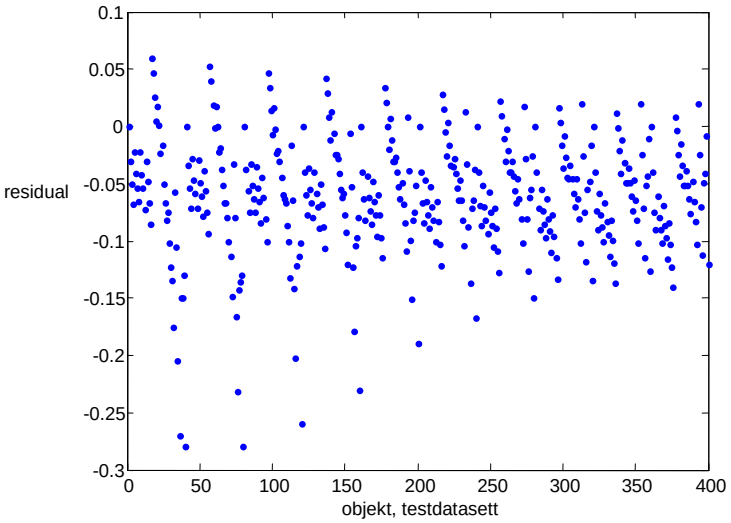
Tabell 4.5.2 Resultat for iterativ konsentrasjonsbestemming av kjemisk samansetjing ved AR

	Maksimumsfeil pos	Maksimumsfeil neg	Prediksjonsfeil	Prediksjonssfeil, %
Metan	0.06	-0.28	0.07	8.1
Etan	0.15	0	0.08	106.7
Aks	0.13	-0.09	0.03	42.9

Ein utfører AR med tre likningar og regresjonskoeffisientane for modellane som er presentert i kapittel 4.5.1, samt ei **X**-matrise frå testdatasettet der ein har trykk, temperatur og lydhastigheit og med middelverdiar i dei definerte områda for dei kjemiske komponentane. Startverdiane for AR er altså 0.85 for metan, 0.07 for etan og 0.06 for aks. Metan vert predikert først, så etan og til slutt aks før ein startar frå toppen igjen med dei predikerte verdiane som startverdiar. Denne prosedyren vert gjentatt til ein har nådd eit gitt konvergenskriterium. I denne oppgåva er dette kriteriet sett til at absoluttverdien av differansen mellom ny prediksjon og førre prediksjon for kvar variabel må vere mindre enn 0.001. På denne måten utfører ein iterativ konsentrasjonsbestemming av metan, etan og aks ved AR for alle objekta i testdatasettet. Ut frå histogrammet i figur 4.5.15 ser ein at det er overvekt av negative residual for prediksjon av metan. Det betyr at prediksjonen har vore for høg i forhold til målt verdi for mange av objekta. Hovudtyngda av residuala ligg i området -0.15 til 0.05 og plottet har ein topp ved -0.06. I tillegg er det tilstades ein del objekt med store negative residual. Figur 4.5.16 viser at objekta med lågast temperatur har dei høgaste residuala.



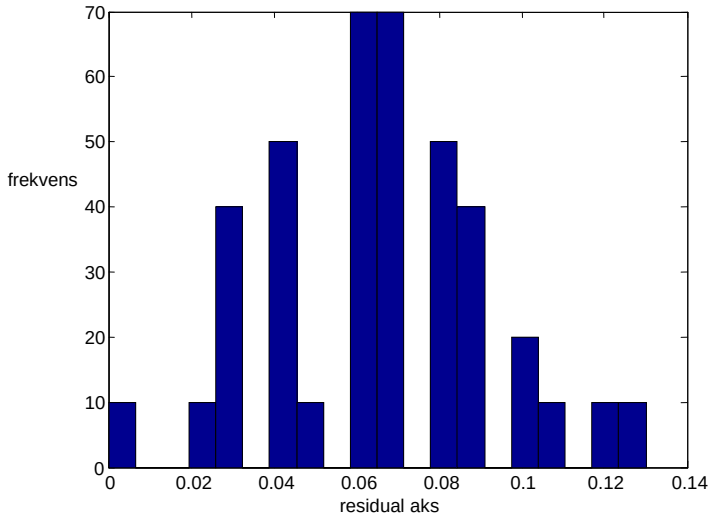
Figur 4.5.15 Histogram av residuala etter prediksjon av metan ved AR.



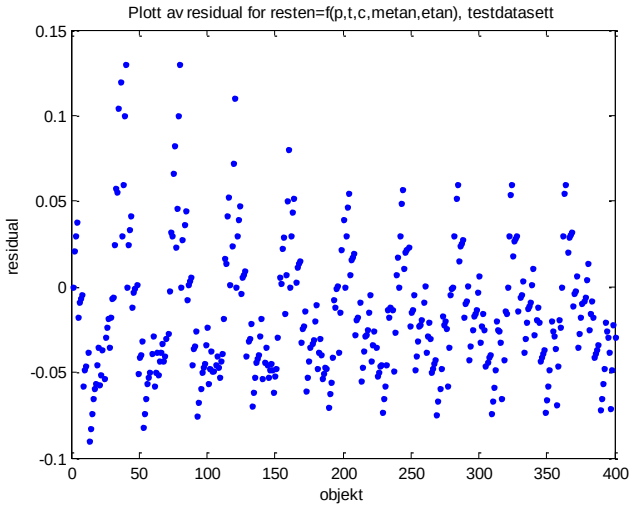
Figur 4.5.16 Plott av residual mot objekt for metan.

For etan ser ein at alle objekta får konsentrasjonen 0 etter iterasjon ved AR. Residuala tilsvarar då avstanden frå den opphavlege målte verdien til 0 og gir ikkje særleg nyttig informasjon, figuren som viser histogram over residuala for etan finst i appendiks, figur A.5.1.

Histogrammet av residuala etter iterativ konsentrasjonsbestemming ved AR av variabelen aks (figur 4.5.17) viser at alle residuala er positive. Ein finn residuala seg i området 0.00 til 0.13 og det er samanfallande med området variabelen er definert for. Dei største residuala finn ein også her for objekta som har kombinasjonen låg temperaturar og høgt trykk (figur 4.5.18).



Figur 4.5.17 Histogram av residual for aks etter prediksjon med AR



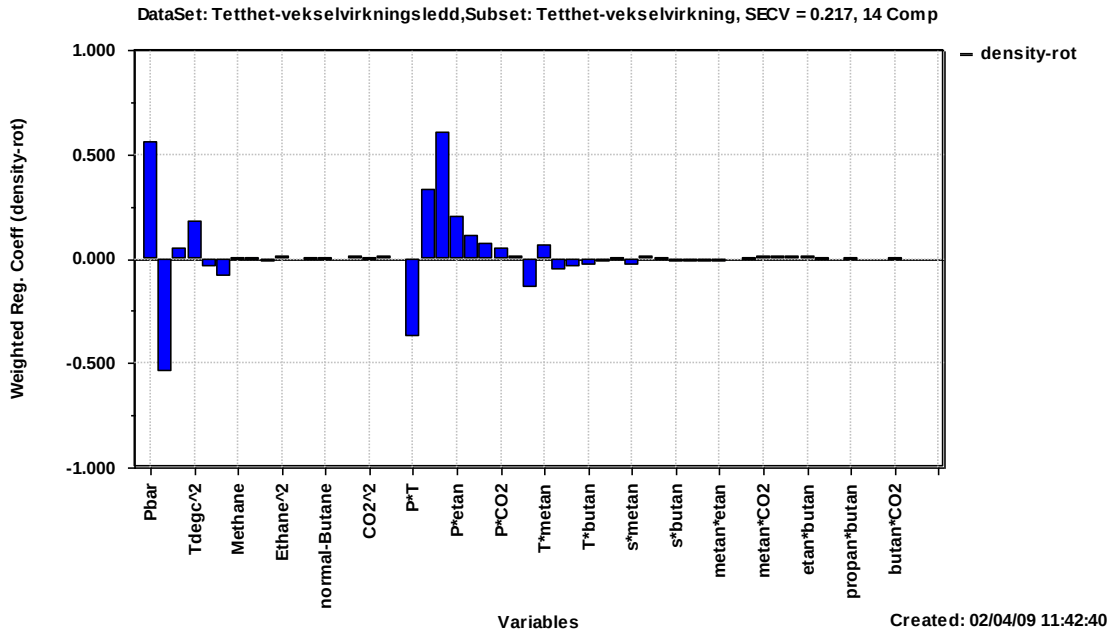
Figur 4.5.18 Plott av residual for resten etter AR.

4.6 Tettleik

4.6.1 Modellering og validering av tettleik frå kjemisk samansetjing

Ein forsøker først å modellere tettleik frå variablane trykk, temperatur, lydhastigheit og den kjemiske samansetjinga i naturgassen.

Ulike modellar for tettleik vart forsøkt og den modellen som gav høgast forklart varians i y i tillegg til å ha normalfordelte responsresidual var ein modell som inneheld førstegradsledd, andregradsledd og vekselverknadsledd i **X** og kvadratrottransformert respons. 14 komponentar vart inkludert i modellen. Oversikt over modellar som vart forsøkt finst i appendiks. I datasettet har tettleik målte verdiar i området 7.53 kg/m^3 - 159.26 kg/m^3. Når det er stort spenn i ein variabel kan ein i mange tilfelle få ein betre modell dersom ein komprimerer området, slik som blir konsekvensen ved å ta kvadratrota av tettleik for alle objekta. Talverdien for tettleik vert då redusert til å ligge i området 2.74– 12.62.



Figur 4.6.1 Grafisk framstilling av vekta regresjonskoeffisientar for variablane som er inkludert i modellen for tettleik.

Som ein kan sjå frå figur 4.6.1 er det variabelen trykk saman med veksleverknadsleddet trykk x metan som har det største positive bidraget til denne modellen. Kvadrert trykk har saman med vekselverknaden trykk x temperatur det største bidraget til modellen i negativ retning. Den kjemiske samansetjinga har ubetydelege bidrag til modellen. I dei tilfella der kjemiske

komponentar bidrar til modellen er det når dei inngår i vekselverknadsledd med trykk. Responsresiduala for modellen for tettleik er relativt normalfordelte. (Appendiks, figur A.6.1). Det ser altså ikkje ut som om kunnskap om det kjemiske samansetjinga er naudsynt for nøyaktig prediksjon av tettleik, slik det er hevda i tidlegare arbeid [10].

Det var også forsøkt å lage ein modell for tettleik som inkluderte førstegradsledd, andregradsledd og vekselverknadsledd i **X** og ingen transformasjon av y. Denne modellen gav forklart varians for y på 99,26 %, men responsresiduala var svært store. Ein held difor fast på at det er riktig å kvadratrottransformere responsen i modelleringa av tettleik.

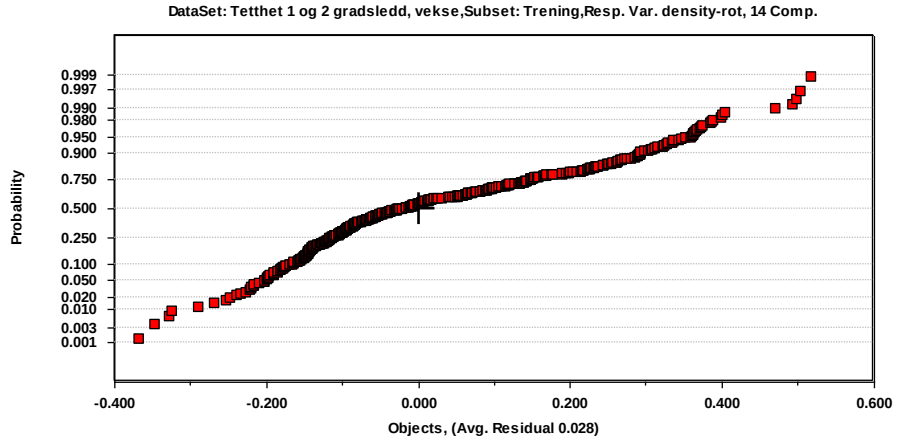
Modellen er testa med testdatasettet og validert på bakgrunn av desse resultata.

Tabell 4.6.1 Valideringstabell for modell for tettleik

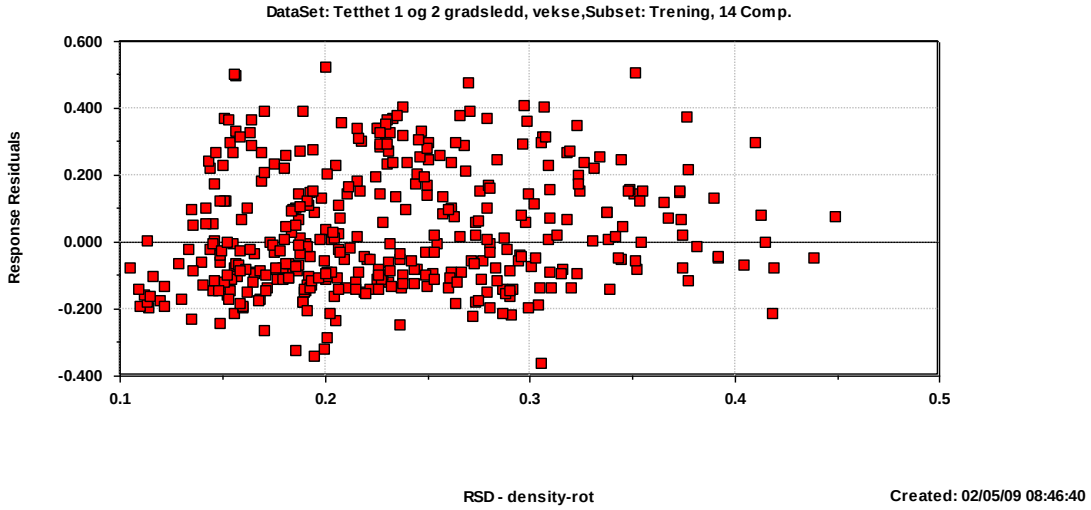
Valideringsparameter	Verdi
Forklart varians i y, %	99.66
Talet på inkluderte komponentar	14
Kryssvalidering, CsvSD	0.972
RSD >	0.207
R^2	0.997
Q^2	0.996
Prediksjonsfeil, kg/m^3	0.142
Prediksjonsfeil %	0.14

Som ein ser frå tabell 4.6.1 forklarar modellen heile 99.66 % av variansen i y. Kryssvalideringa viser at CsvSD for den fjerde komponenten er under 1.000, dette peikar mot at den kan inkluderast. Med så mange komponentar inkludert i modellen vil det vere fare for overtilpassing, og R^2 verdien på 0.997 kan både tyde på at dette er tilfelle, men det kan også tyde på at modellen er god. . Ein ser god intern prediktiv evne med Q^2-verdi på 0.996. Responsresiduala for objekta frå testdatasettet er relativt normalfordelte. Nokre objekt ligg for seg sjølv oppe til høgre, desse har ikkje mykje større residual enn venta ut frå normalfordelingsplottet, dette er vist i figur 4.6.2. Prediksjonsfeilen er 0.142 kg/m^3 og ein vil dermed anta at dette er ein tilfredsstillande modell for tettleik. Som ein kan sjå av figur 4.6.3

er det ingen tydeleg samanheng mellom responsresiduala og RSD for objekta. Det ser ut som om dei positive responsresiduala har eit litt større spenn enn dei negative. Tilfeldig fordeling i dette plottet peikar også mot ein god modell. Ein merkar seg at responsresiduala i dette plottet svarar til dei kvadratrottransformerte verdiane av tettleik. Forholdet mellom RSD og responsresidual er likevel det same.

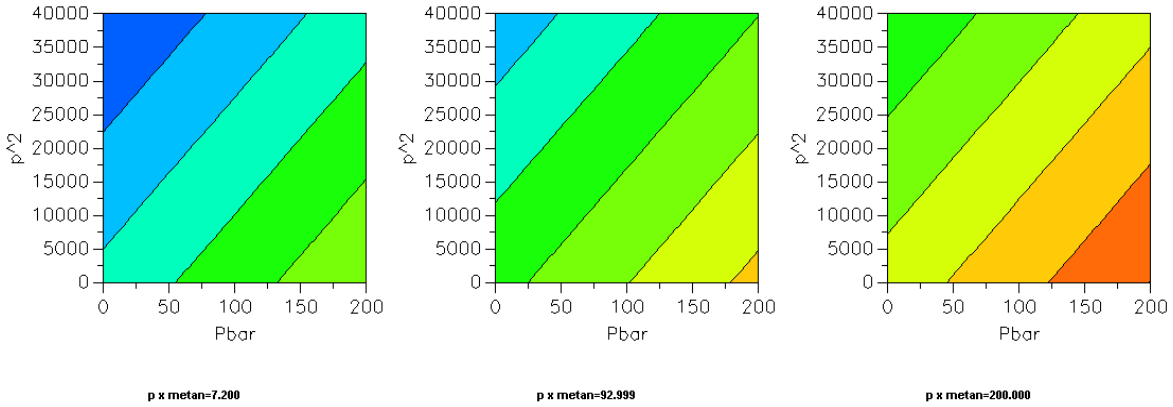


Figur 4.6.2 Normalplott av responsresidual mot objekt



Figur 4.6.3 Plott av responsresidual mot RSD. NB: denne figuren viser responsresiduala for kvadratrota av dei opphavlege verdiane.

Konturplott av tettleik



Figur 4.6.4 Konturplott for tettleik-kvadratrot

Som ein kan sjå av figur 4.6.4 har ein høgast verdi for kvadratrota av tettleik, og dermed også tettleik, når verdien for trykk x metan er høg, kvadrert trykk er låg og trykk er høgt. Lågast respons ser ein når trykk x metan er låg, kvadrert trykker høg og trykket er lågt.

Ein har forsøkt fleire ulike modellar for tettleik. Resultatet vart ein modell med førstegradsledd, andregradsledd og vekselverknadsledd i **X** og kvadratrottransformasjon av y.

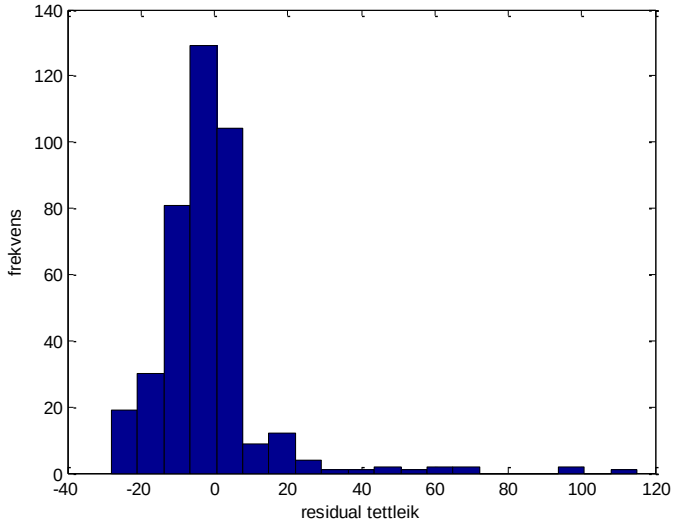
4.6.2 Prediksjon av tettleik frå iterativ konsentrasjonsbestemming av kjemisk samansetjing ved AR

Tettleik vart predikert frå modellen som er presentert i kapittel 4.6.1. Kjemisk samansetjing er funnen ved iterativ konsentrasjonsbestemming ved AR, dette er presentert i kapittel 4.1. Residuala er rekna ut frå dei opphavlege simulerte verdiane.

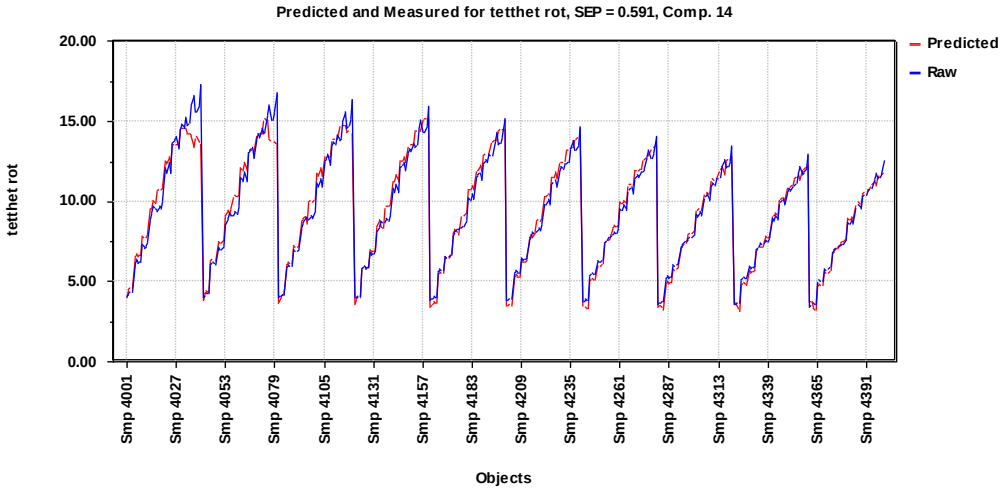
Tabell 4.6.2 Resultat av prediksjon av tettleik frå iterativ konsentrasjonsbestemming av kjemisk samansetjing ved AR

	Maksimumsfeil neg, kg/m^3	Maksimumsfeil pos, kg/m^3	Prediksjonsfeil, kg/m^3	Prediksjonsfeil, %	Talet på objekt
Tettleik	-28	115	9.0	9.0	400

Som ein kan sjå i tabell 4.6.2 er prediksjonsfeilen for tettleik i størrelsesorden 9 % . Histogrammet i figur 4.6.5 viser at ein finn hovudtyngda av residuala mellom -20 og 20 kg/m^3. Ein ser også at plottet har topp rundt 0. Det er nokre objekt tilstades som har svært store positive residual, noko som betyr at for desse objekta er prediksjonen for låg. Ved å plotte predikert og målt verdi for alle objekta i same plott slik det er gjort i figur 4.6.6 ser ein at det er objekta med kombinasjonen låg temperatur og høgt trykk som har desse store positive residuala. I denne figuren er det dei kvadratrottransformerte verdiane som er plotta, men den avslører likevel tydeleg kva objekt som har størst residual. Når temperaturen aukar ser ein at skilnadane mellom målt og predikert verdi er mykje mindre, dette gjeld også for dei objekta som har høgt trykk.



Figur 4.6.5 Histogram av residual for tettleik etter iterativ konsentrasjonsbestemming ved AR.



Figur 4.6.6 Plott av prediket og målt verdi av tettleik. NB: data i denne figuren er kvadratrottransformert.

Ein har i dette kapitlet predikert tettleik der den kjemiske samansetjinga i **X** er funnen ved iterativ konsentrasjonsbestemming ved AR. Resultatet vart prediksjonsfeil i størrelsesorden 9 %. I tillegg var det liten samanheng mellom målt og predikert verdi for dei objekta som har kombinasjonen låg temperatur og høgt trykk.

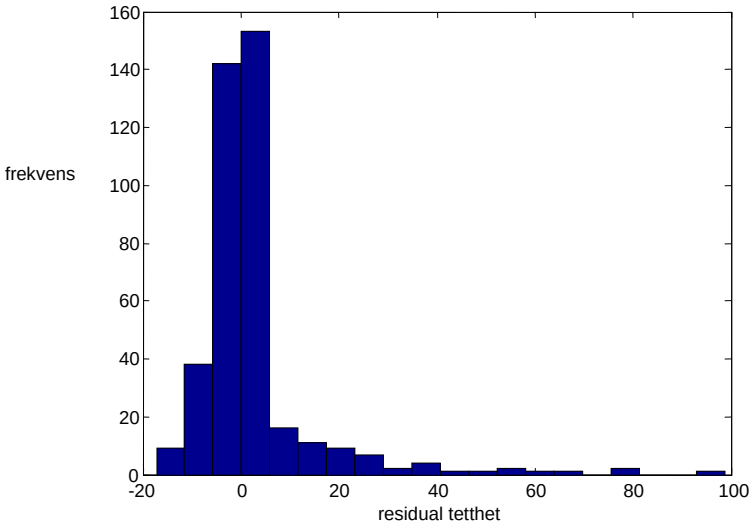
4.6.3 Prediksjon av tettleik ved oppbygging av kjemisk samansetjing ved prediksjon

I denne oppgåva har ein forsøkt å nytte den predikerte kjemiske samansetjinga ein har funne i kapittel 4.2 til å predikere tettleik i naturgassen. Modellen som er nytta til prediksjon av tettleik er presentert i kapittel 4.6.1.

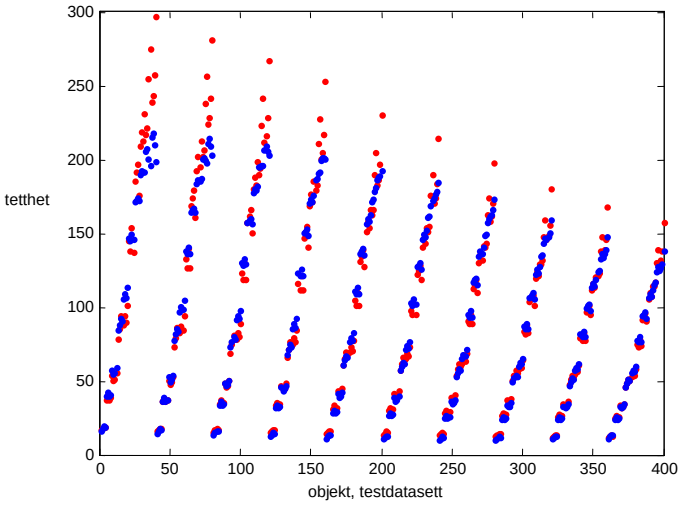
Tabell 4.6.3 Resultat av prediksjon av tettleik ved kjemisk samansetjing oppbygd ved prediksjon

	Maksimumsfeil neg, kg/m^3	Maksimumsfeil pos, kg/m^3	Prediksjonsfeil, kg/m^3	Prediksjonsfeil, %	Objekt predikert
Tettleik	-18	98	6.5	6.5	400

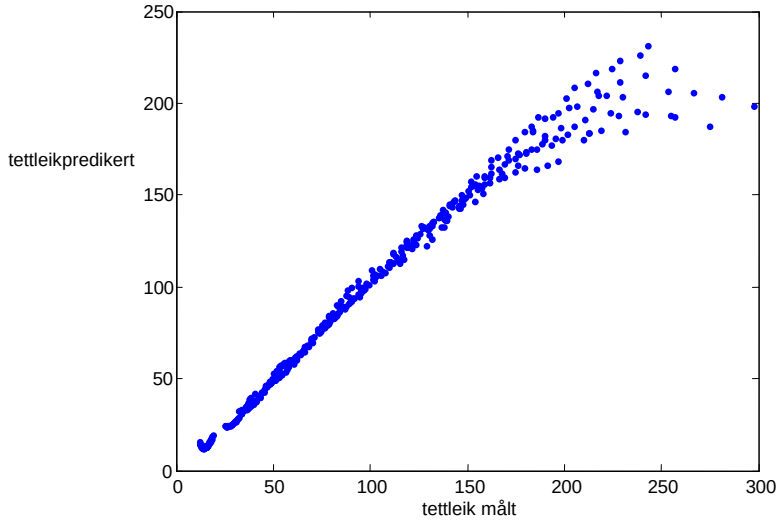
Tabell 4.6.3 viser at prediksjonsfeilen er ganske høg og at maksimumsfeilen i positiv retning er svært høg, heile 98 kg/m^3. Dette kjem også fram av figur 4.6.7 som viser histogram av residuala ved prediksjon av tettleik. Når ein ser på dei 14 objekta med høgast residual ser ein at felles for desse er at dei har høgt trykk ($p \geq 150$ bar) og låg temperatur ($t \leq 30$ ˚C). Alle objekta har også tettleik ≥ 225 kg/m^3. Dette viser at det er vanskeleg å predikere tettleik bra i situasjonar der ein har både høge trykk og låge temperaturar. Dette ser ein tydeleg i figur 4.6.8. I denne figuren er det tydeleg at dei predikerte verdiane ikkje når høgt nok i situasjonar der ein har låge temperaturar, og ein får store positive residual. Ved å sjå på plottet i figur 4.6.9 ser ein at når ein plottar prediket tettleik mot målt tettleik er samanhengen lineær opp til verdiar rundt 150, deretter bøyer plottet av. Ein har her tydelege problem med å predikere dei høgaste verdiane for tettleik og når tettleiken er høg er også trykket høgt og temperaturen låg.



Figur 4.6.7 Histogram av residuala ved prediksjon av tettleik



Figur 4.6.8 Plott av predikert (blå) og målt (raud) verdi for tettleik for objekta i testdatasettet.



Figur 4.6.9 Plott av predikert verdi for tettleik mot målt verdi for tettleik.

I tilfellet som er presentert i dette kapitlet er den største (i prosent) kjemiske komponenten predikert først. Modellen ein har nytta har låg grad av forklart varians for y og ville normal sett ikkje vore vurdert som ein god modell. Årsaka til at ein valde dette startpunktet var fordi ein tenkte seg at prediksjonsfeilen ville forplante seg utover etter kvart som fleire og fleire kjemiske komponentar vart predikert, og stor feil i dei minste (i prosent) komponentane ville ha mindre å seie enn stor feil i dei komponentane som utgjer hovudtyngda av naturgass. Ulempa med å starte med metan var at modellen for metan vart lite tilfredsstillande, og modellane for dei resterande kjemiske komponentane vart stadig betre. Ein ville difor undersøkje om resultata av prediksjon av tettleik vart best med å starte med ein dårleg modell for metan, eller å predikere metan ut frå ein god modell men med eit datasett som kanskje ville innehalde feil av varierande storleik for dei kjemiske komponentane. Ved å starte prediksjonen i andre enden, med den minste komponenten først vil ein få betre modell til å predikere metan med. Hovudføremålet er likevel å undersøkje kor godt ein kan predikere tettleik og brennverdi ut frå den predikerte kjemiske samansetjinga.

Ein har no sett at prediksjon av tettleik med kjemisk samansetjing slik presentert i dette kapitlet gir prediksjonsfeil i størrelsesorden 6.5 % og dårleg samsvar mellom målt og predikert verdi for objekt med tettleik over 150 kg/m^3.

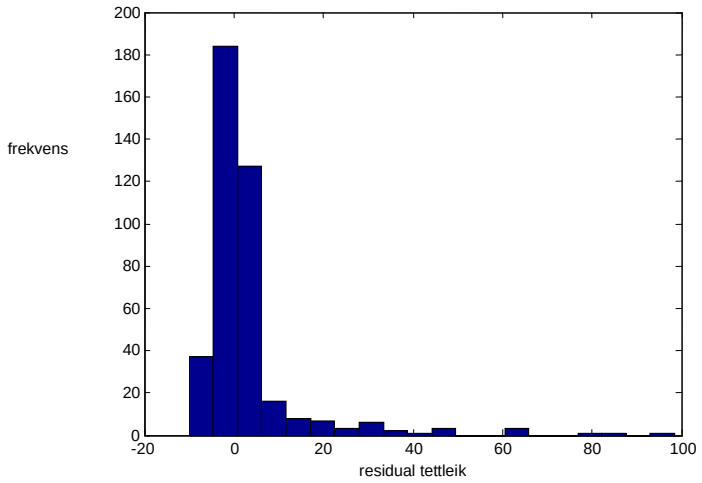
4.6.4 Prediksjon av tettleik ved revers prediksjon av kjemisk samansetjing

Ein ynskjer no å undersøkje om revers prediksjon av kjemisk samansetjing slik det er presentert i kapittel 4.3 kan gi betre prediksjonar av tettleik. Modellen som er nytta til prediksjon av tettleik er presentert i kapittel 4.6.1.

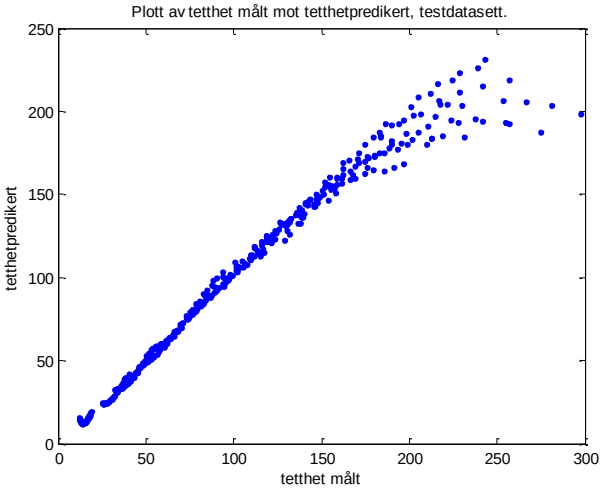
Tabell 4.6.4 Resultat for prediksjon av tettleik ved revers prediksjon kjemisk samansetjing

	Maksimumsfeil pos, kg/m^3	Maksimumsfeil neg, Kg/m^3	Prediksjonsfeil kg/m^3	Prediksjonsfeil %	Objekt predikert
Tettleik	98	-10	5.5	5.5	400

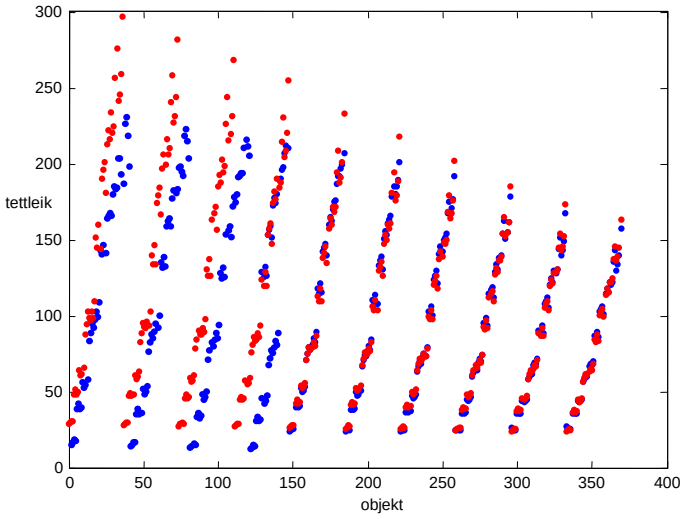
I tabell 4.6.4 ser ein at maksimumsfeilen i negativ retning er -10 kg/m^3 og i positiv retning heile 98 kg/m^3. Prediksjonsfeilen er i størrelsesorden 5.5 kg/m^3, noko som tilsvarar ein feil på 5.5 %. Frå histogrammet i figur 4.6.10 ser ein at hovudtyngda av residuala er fordelt mellom -10 kg/m^3 og 10 kg/m^3. Objekta som har låg temperatur og høgt trykk er også i denne situasjonen dei som får dei største prediksjonsfeila (figur 4.6.12). Figur 4.6.11 viser at også her er det bra samanheng mellom målt og predikert verdi opp til verdiar for tettleik på om lag 150 kg/m^3. Deretter bøyer plottet av og prediksjonane blir for låge.



Figur 4.6.10 Histogram av residual for prediksjon av tettleik.



Figur 4.6.11 Plott av predikert verdi for tettleik mot målt verdi for tettleik, revers prediksjon av kjemisk samansetjing



Figur 4.6.12 Plott av predikert (blå) og målt (raud) verdi for tettleik, revers prediksjon av kjemisk samansetjing.

Resultata presentert i dette kapitlet viser at når den kjemiske samansetjinga er funnen ved revers prediksjon får ein prediksjonsfeil for tettleik i størrelsesorden 5.5 %. Også her er det tilfredsstillande samanheng mellom målt og predikert verdi tettleik opp til om lag 150 kg/m^3, deretter ser ein ikkje tilfredsstillande samanheng. Det er også vist at objekt med kombinasjonen høgt trykk og låg temperatur har dei største residuala i prediksjonen av tettleik.

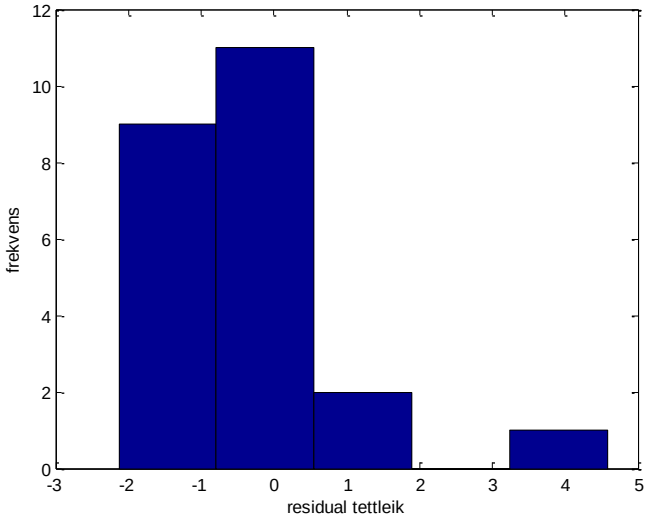
4.6.5 Prediksjon av tettleik ved konstant trykk og temperatur

I denne delen av oppgåva vil ein undersøkje kor stor effekt det har på prediksjon av tettleik å halde trykk og temperatur konstant.

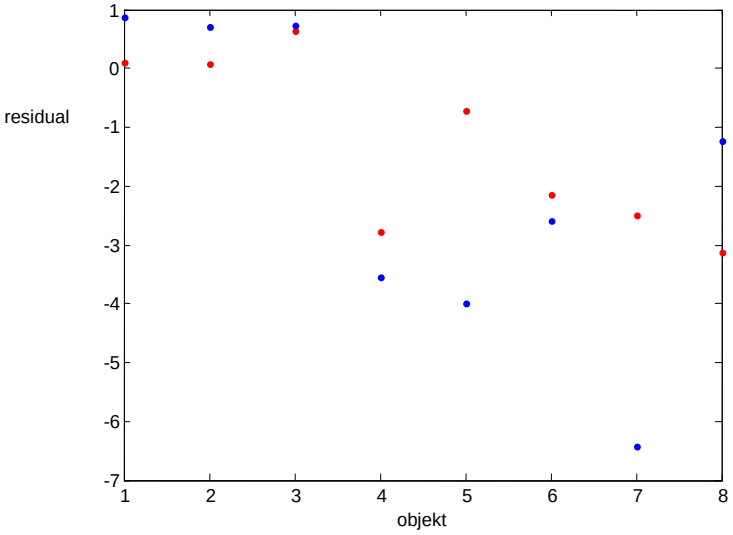
Tabell 4.6.5 Resultat frå prediksjon av tettleik ved konstant trykk og temperatur

	Maksimumsfeil neg, kg/m^3	Maksimumsfeil pos, kg/m^3	Prediksjonsfeil, kg/m^3	Prediksjonsfeil, %	Objekt predikert
Tettleik	4.6	-2.1	1.0	1.3	23

Når trykk og temperatur vert haldne konstante blir prediksjonane av tettleik svært mykje betre enn det ein har sett i dei andre situasjonane som er presentert. Trykk og temperatur vil påverke prediksjonen i stor grad. Maksimumsfeilen er no berre -2.1 kg/m^3 i negativ retning og 4.6 kg/m^3 i positiv retning. Prediksjonsfeilen er nede i størrelsesorden 1.0 kg/m^3, noko som tilsvarar 1.3 % feil. Dette er vist i tabell 4.6.5 og histogrammet i figur 4.6.13. Objekta som er predikert har trykk og temperatur i nærleiken av det området der modellane for dei kjemiske komponentane er bygd. Prediksjonen av tettleik er gjort med modellen som er presentert i kapittel 4.6.1. Den kjemiske samansetjinga for objekta henta frå kalibreringsdatasettet summerte ikkje til 100 % etter at alle dei kjemiske komponentane i naturgassen var predikert. Ein ville difor undersøkje om dette hadde mykje å bety for prediksjonen av tettleik. Objekta vart difor normalisert til konstant sum 100 % og tettleik vart predikert på nytt med den normaliserte kjemiske samansetjinga. Som ein kan sjå av plottet i figur 4.6.14 får ein ikkje betre predikert verdi av tettleik ved å normalisere den kjemiske samansetjinga i datasettet. Det er altså ikkje summen av den kjemiske samansetjinga som har noko å bety for kor god prediksjonen blir. Ein har tidlegare vist at dei kjemiske komponentane har små bidrag til modellen for tettleik (figur 4.6.1), det er difor ikkje uventa at normalisering ikkje har effekt på prediksjon av tettleik.



Figur 4.6.13 Histogram av residuala ved prediksjon av tettleik med 23 objekt innanfor same område



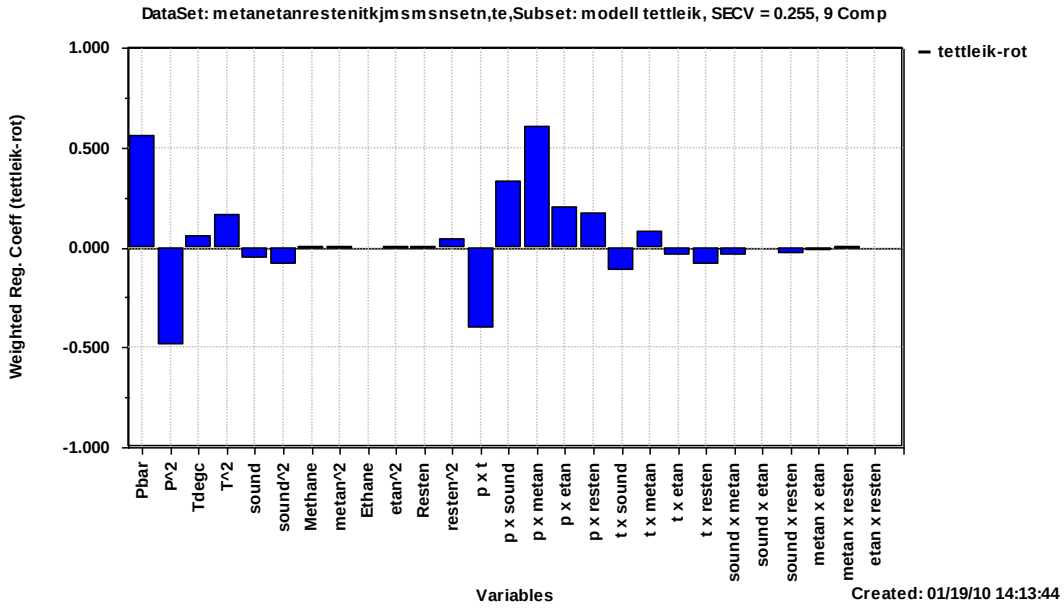
Figur 4.6.14 Plott av residual for predikerte verdiar for tettleik. Raud er predikerte objekt, blå er predikerte og normaliserte objekt.

Ein har no sett at ved å halde trykk og temperatur konstant får ein stor reduksjon i prediksjonsfeilen for tettleik. No ser ein predikert feil i størrelsesorden 1.3 %. Normalisering av den kjemiske samansetjinga førte ikkje til betre prediksjon av tettleik.

4.6.5 Modellering av tettleik frå metan, etan og aks

Tettleik vert no modellert som ein funksjon av variablane trykk, temperatur, lydhastigheit, metan, etan og addert kjemisk samansetjing (aks). Aks består av summen av propan, butan, CO_2 og N_2.

For denne modellen for tettleik er det vekselverknadsleddet trykk x metan samt førstegradsleddet trykk som har dei største bidraga i positiv retning. I negativ retning er det er kvadrert trykk og trykk x temperatur som har dei største bidraga (figur 4.6.15).



Figur 4.6.15 Grafisk framstilling av vekta regresjonskoeffisientar for tettleik modellert frå metan, etan og aks.

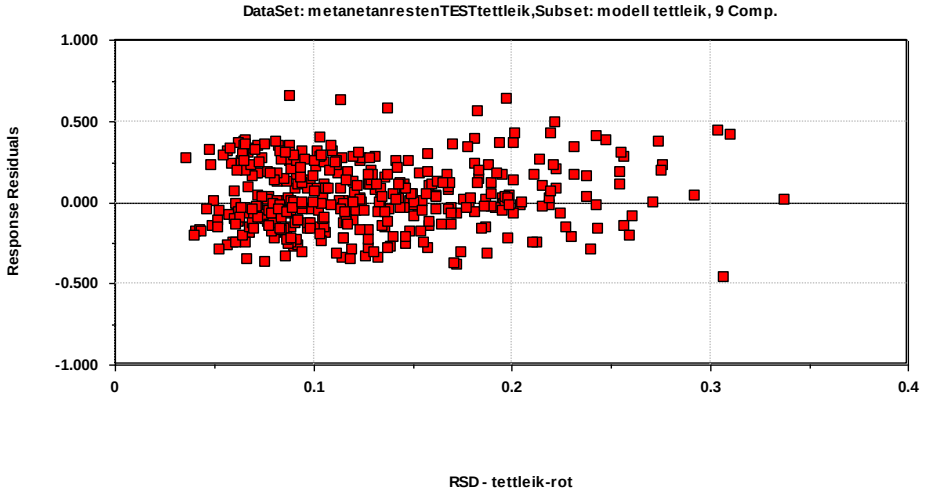
Validering av modellen

Tabell 4.6.6 Validering av modell for tettleik ut frå metan, etan, aks

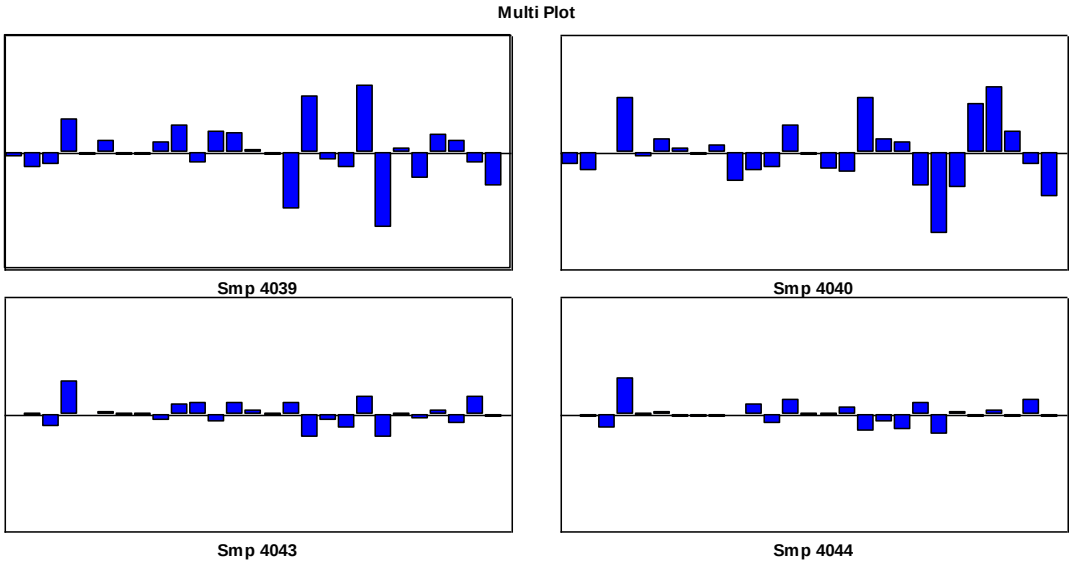
Valideringsparameter	Verdi
Forklart varians i y, %	99.55
Talet på inkluderte komponentar	9
Kryssvalidering siste inkluderte komponent, CsvSD	0.909
RSD >	0.207
R^2	0.996
Q^2	0.996
Prediksjonsfeil, kg/m^3	0.19
Prediksjonsfeil %	0.2

I tabell 4.6.6 ser ein at modellen for tettleik forklarer heile 99.55 % av variansen i y. Ni komponentar er inkludert i modellen. Den siste komponenten kan inkluderast då CsvSD for denne gir verdien 0.909. Både R^2 og Q^2 har høge verdiar og saman med ein prediksjonsfeil på 0.19 kg/m^3 som svarar til 0.2 % indikerer dette at modellen er til å stole på. Det er ingen tydeleg samanheng mellom auke i responsresidual og auke i RSD utover at det er ein større tettleik av objekt med RSD-verdiar i området 0.5 – 1.5 enn med høgare RSD-verdiar (figur 4.6.16). Avvisningskriteriet for uteliggarar i denne modellen er RSD > 0.207. 42 objekt har RSD over denne grensa. Det objektet som har høgast RSD har RSD = 0.338. Ved å samanlikne **X**-residuala for to objekt med RSD < 0.207 og to objekt med RSD > 0.207 ser ein at det er tydelege forskjellar mellom størrelsen på residuala for dei ulike variablane. Særleg skil variabelen aks og vekselverknadsledd med aks inkludert seg ut med store **X**-residual for dei to objekta med RSD > 0.207. Dette er vist i figur 4.6.17. RSD-verdien seier noko om objekta i testdatasettet ligg langt frå modellen, og i dette tilfellet gjer ein del av dei det, men det er ikkje sikkert dette fører til dårlegare prediksjonar av tettleik. Ved å studere plottet i figur 4.6.19 ser ein at det er bra samsvar mellom predikert og målt verdi for tettleik når ein testar modellen. Ein merkar seg at i dette plottet er det dei kvadratrottransformerte verdiane som er plotta.

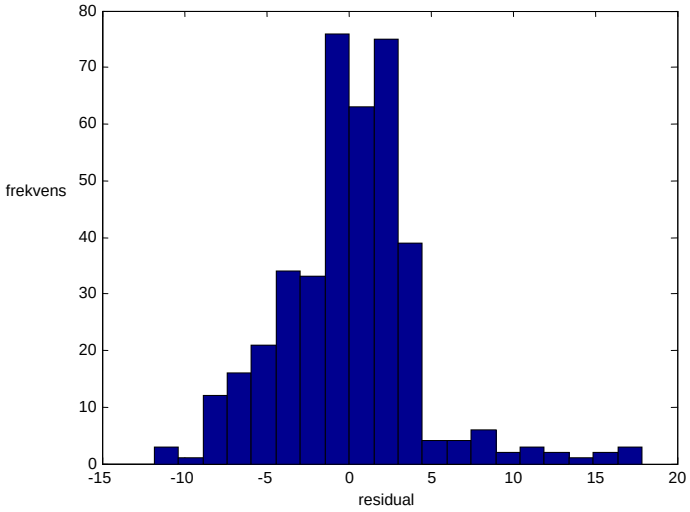
I histogrammet i figur 4.6.18 kan ein sjå at tyngdepunktet av residuala er å finne i området -8 kg/m^3 til 5 kg/m^3. Maksimumsfeilen er større i positiv retning enn i negativ retning.



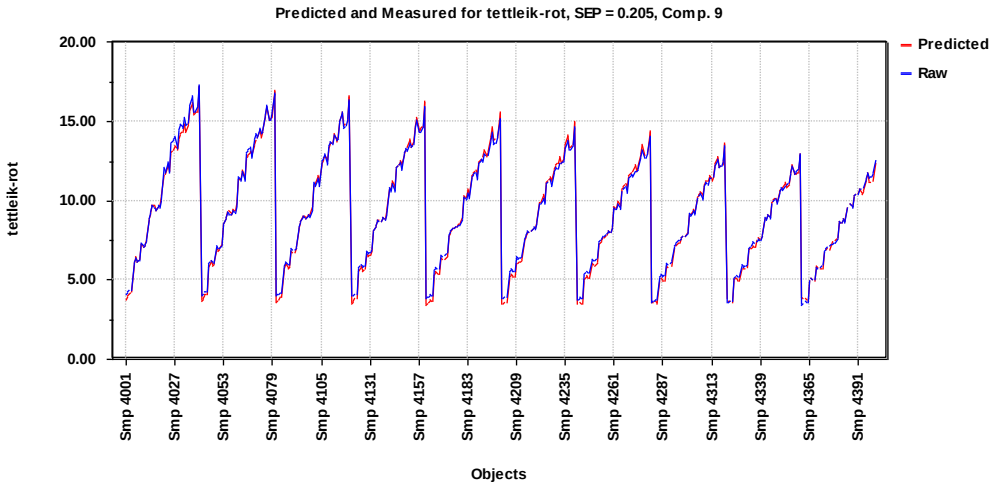
Figur 4.6.16 Plott av responsresidual mot RSD for test av tettleik



*Figur 4.6.17 Plott av **X**-residuala for to prøvar med RSD < 0.207 (spm 4043 og 4044) og to prøvar med RSD > 0.207 (smp 4039 og 4040)*



Figur 4.6.18 Histogram av residual etter testing av modell for tettleik



Figur 4.6.19 Plott av predikert og målt verdi for kvadratrota av tettleik

Modellen for tettleik presentert i dette kapitlet har bra forklart varians i y og ein ser god samanheng mellom målte og predikerte verdiar. Ein del av objekta fell i kategorien uteliggjarar i forhold til RSD-verdien. Det er likevel ikkje sikkert dette påverkar prediksjonen av tettleik.

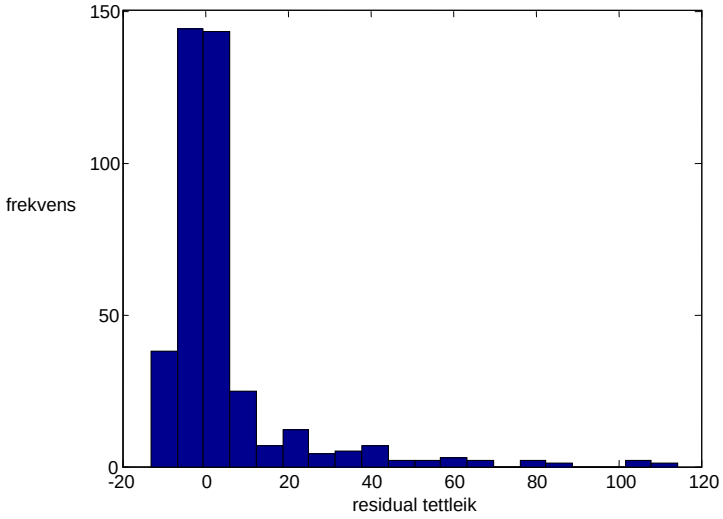
4.6.6 Prediksjon av tettleik ved metan, etan og aks

Modellen som er presentert i kapittel 4.6.6 vart så nytta til prediksjon av tettleik der den kjemiske samansetjinga er funnen ved iterativ konsentrasjonsbestemming ved AR for variablane metan, etan og aks.

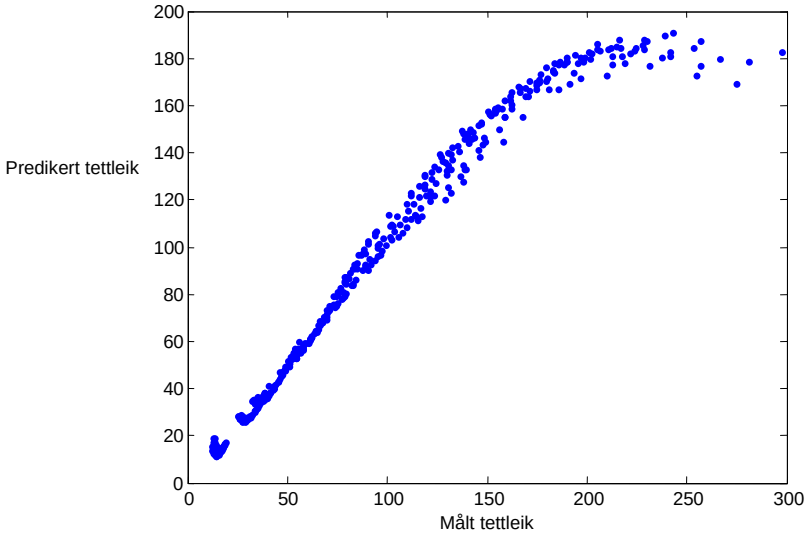
Tabell 4.6.7 Resultat av prediksjon av tettleik frå metan, etan og aks

	Maksimumsfeil pos, kg/m^3	Maksimumsfeil neg, kg/m^3	Prediksjonsfeil, kg/m^3	Prediksjonsfeil, %	Objekt predikert
Tettleik	114	-13	8.2	8.3	400

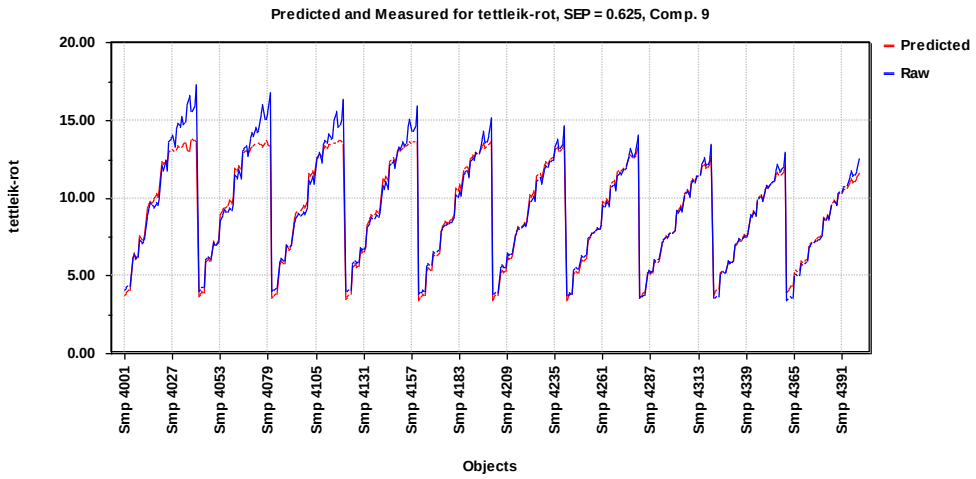
I modellen for tettleik har ein kvadratrottransformert responsen av same årsaker som diskutert i kapittel 4.6.1. Sidan det i utgangspunktet er stort spenn variabelen er det truleg at modellen blir betre dersom ein nyttar kvadratrottransformasjon av y. Alle resultata etter prediksjonen er difor kvadrert og residuala er rekna ut frå den opphavlege målte verdien av tettleik. Tabell 4.6.7 viser at prediksjonen av tettleik har ein svært høg maksimumsfeil i positiv retning. Når ein ser på histogrammet i figur 4.6.20 ser ein at hovudtyngda av residuala ligg mellom -18 og 16 kg/m^3. Det er få objekt som har høgare residual enn dette. Prediksjonsfeilen er 8.25 kg/m^3, dette svarar til ein prediksjonsfeil i størrelsesorden 8.3 %. Plottet i figur 4.6.21 viser at det er samsvar mellom predikert og målt verdi for objekt som har låg tettleik. Når verdiane blir høgare bøyer plottet av og ein ser at for høge verdiar av tettleik er prediksjonen for låg. Trykk har stort positivt bidrag til modellen, det vil seie at høge verdiar av tettleik er assosiert med høge trykk. I datasettet har etan som nemnt tidlegare verdien 0.00 for alle objekta. Etan inngår i vekselverknadsledd med trykk og dette leddet har eit positivt bidrag til modellen. No vil alle desse verdiane vere 0 og dermed ikkje bidra til modellen noko som også kan vere med å forklare kvifor prediksjonane blir for låge. Figur 4.6.21 viser at når ein plottar predikert verdi av tettleik saman med målt verdi av tettleik får ein best resultat for objekt med høge temperaturar. Ein ser at prediksjonen er dårlegast når ein har kombinasjonen høgt trykk og låg temperatur.



Figur 4.6.20 Histogram av residual for prediksjon av tettleik



Figur 4.6.21 Plott av predikert verdi for tettleik mot målt verdi for tettleik.



Figur 4.6.21 Plot av predikert og målt verdi for kvadratrota av tettleik, kjemisk samansetjing er predikert.

Ein ser at ved å predikere tettleik med variablane metan, etan og aks får ein prediksjonsfeil i størrelsesorden 8.3 %. Samanhengen mellom målt og predikerte verdi for tettleik er tilfredsstillande for låge verdiar, men når målt verdi passerer om lag 150 kg/m^3 vert prediksjonane svært upresise. Dei største residuala ser ein for objekta med kombinasjonen høgt trykk og låg temperatur.

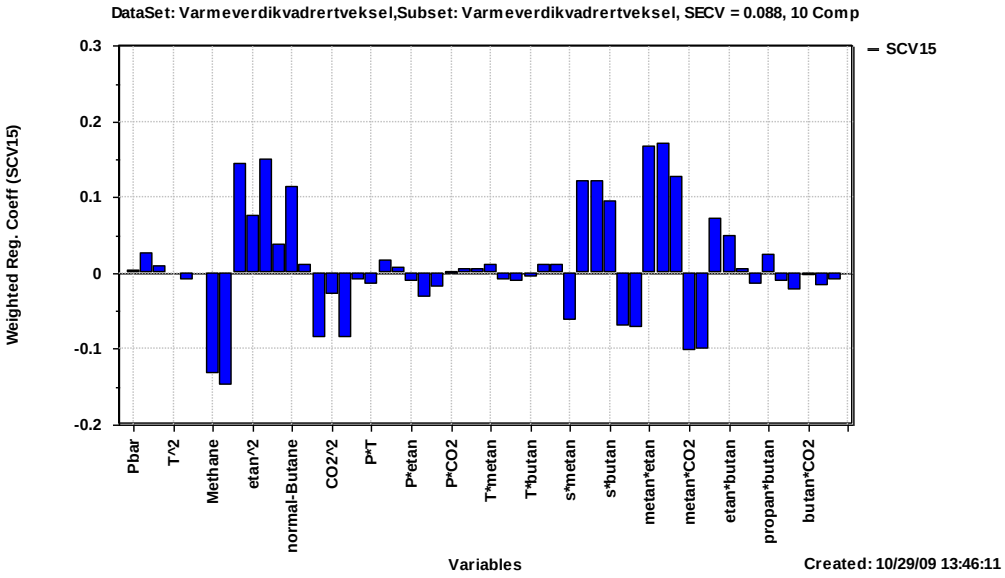
4.7 Brennverdi

4.7.1 Modellering og validering av brennverdi frå kjemisk samansetjing

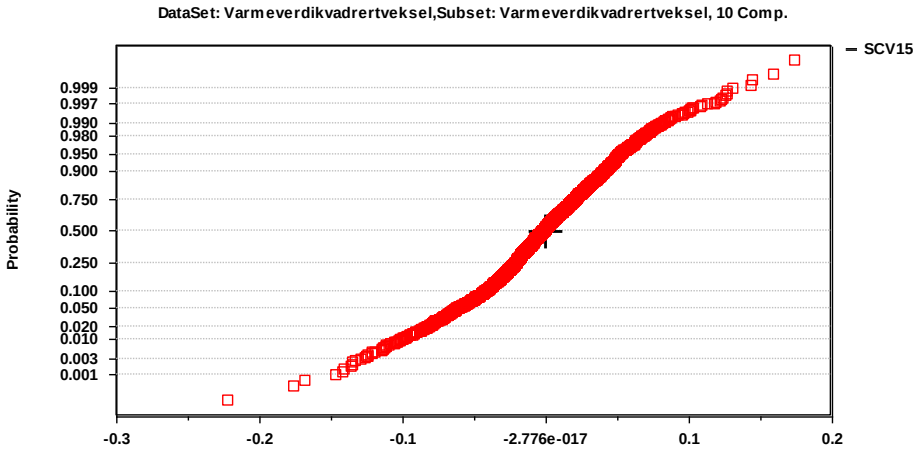
Brennverdi vart modellert på fleire ulike måtar og den modellen som gav høgast forklart varians i y er modellen som inneheld førstegradsledd, andregradsledd og vekselverknadsledd i **X** og ingen transformasjon av y. Ei oversikt over alle modellane finst i appendiks. I enkelte av figurane er brennverdi skriven som SCV 15.

Ut frå kryssvalidering og SECV-plott valde ein å inkludere 10 komponentar i modellen.

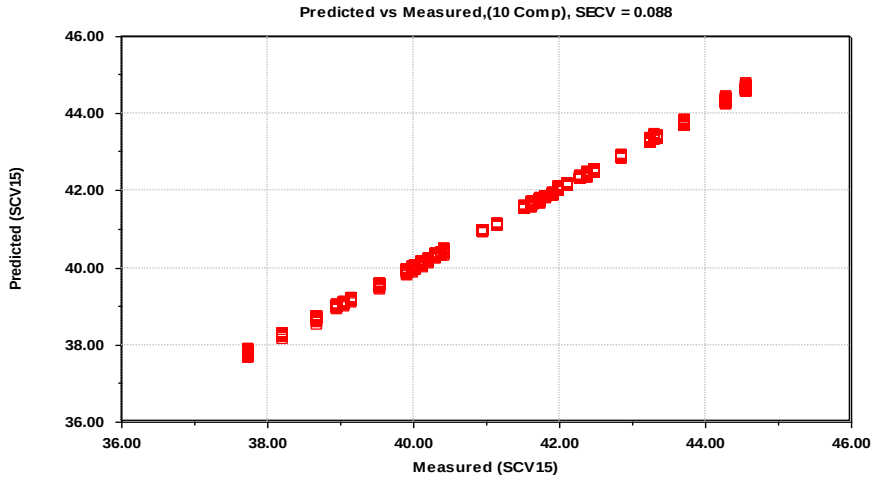
Som ein kan sjå i figur 4.7.1. er kvadrert metan og metan dei variablane som bidrar mest til modellen i negativ retning. I positiv retning er det fleire variablar som utmerkar seg med stort bidrag. Den som bidrar mest er vekselverknadsleddet metan x propan. Metan x etan har også eit stort positivt bidrag til denne modellen. Trykk, temperatur og lydhastigheit har nærmast ingenting å seie for denne modellen når ein samanliknar med bidraga frå den kjemiske samansetjinga. Brennverdien er eit mål på kor mykje energi det er mogleg å få ut av gassen og det er difor naturleg å tru at dette har samanheng med den kjemiske samansetjinga og ikkje i så stor grad med trykk og temperatur som jo er variablar ein kan kontrollere i prosessen. Hovudtyngda av responsresiduala er bra normalfordelte. I kvar ende av plottet ser ein nokre få objekt som fell utanfor normalfordelinga. Desse ligg lengre borte frå modellen enn dei skulle gjort dersom dei hadde falle innanfor normalfordelinga (figur 4.7.2). I denne modellen er det svært bra samanheng mellom predikert verdi og målt verdi (figur 4.7.3). Det er viktig med god samanheng mellom målt og predikert verdi, og i mange tilfeller vil ein leggje meir vekt på denne parameteren enn at modellen har normalfordelte responsresidual og høg forklart varians i y. Det vil likevel vere valideringa som avgjer om ein modell kan nyttast eller ikkje.



Figur 4.7.1 Grafisk framstilling av vekta regresjonskoeffisientar modellen for brennverdi.



Figur 4.7.2 Normalplott av responsresiduala for modellen for brennverdi.



Figur 4.7.3 Plott av predikert verdi mot målt verdi.

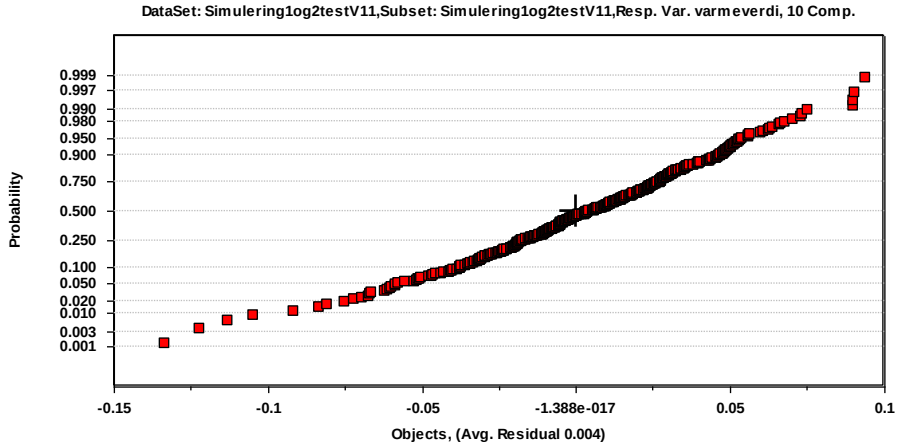
Modellen er testa ved å predikere brennverdi for 400 nye objekt i testdatasettet.

Tabell 4.7.1 Valideringstabell, modell for brennverdi

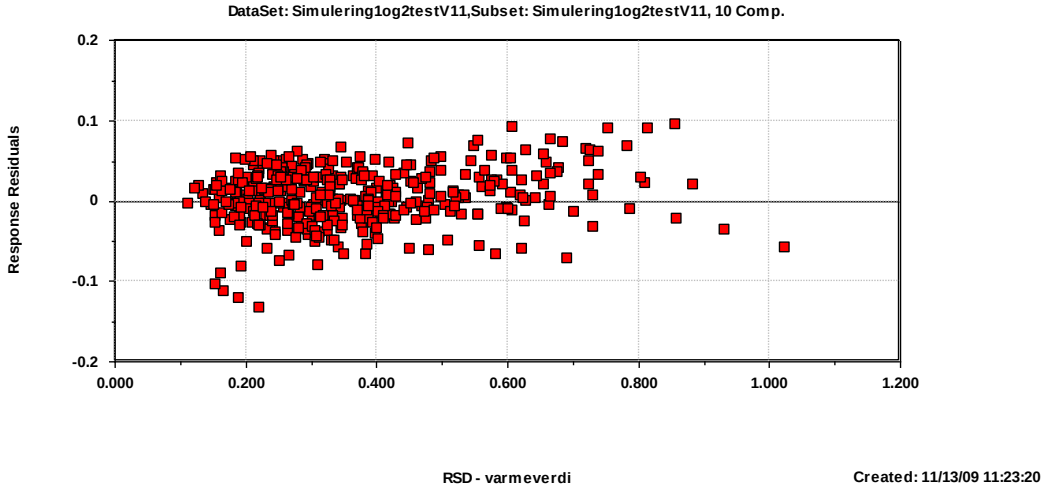
Valideringsparameter	Verdi
Forklart varians i y, %	99.96
Komponentar inkludert	10
Kryssvalidering, siste inkluderte komponent, CsvSD	0.853
RSD >	0.523
R^2	0.999
Q^2	0.999
Prediksjonsfeil, MJ/Sm^3	0.004
Prediksjonsfeil %	0.004

Responsresiduala for dei nye objekta i testdatasettet er bra normalfordelte. Ein ”hale” i plottet nede til venstre tyder på at nokre av objekta ligg litt for langt frå modellen sjølv om normalfordelinga er ivaretatt, dette er vist i figur 4.7.4. Validering av denne modellen for brennverdi er oppsummert i tabell 4.7.1. For modellen er forklart varians for y heile 99.96 %. Ein har inkludert 10 komponentar i modellen. Det vil alltid vere fare for overtilpassing når ein inkluderer så mange komponentar i modellen. R^2 verdien for denne modellen er høg, heile 0.999. Høg kumulativt forklart verdi kan tyde på at modellen er bra, men høg R^2 kan også skuldast overtilpassing. I dette datasettet har ein berre simulerte verdiar og dermed ikkje støy tilstades Dermed vil alle bidrag til modellen vere signifikante og ein antar at ein kan inkludere 10 komponentar utan at overtilpassing førekjem. Intern prediktiv evne for modellen, Q^2, er eit betre uttrykk for prediktiv evne enn R^2. For den aktuelle modellen ser ein at Q^2-verdien er 0.999, dette viser på at modellen har svært god intern prediktiv evne. Prediksjonsfeilen for modellen er 0.0042 MJ/Sm^3, ein verdi som svarar til omlag 0.004 % feil. Avvisningskriteriet for uteliggjarar i denne modellen er RSD > 0.523. 72 av objekta i testdatasettet har RSD-verdi over grensa, objektet med høgast RSD-verdi har RSD = 1.024. Ein del objekt kan altså definerast som uteliggjarar ut frå dette kriteriet. Det er likevel ikkje sikkert høge verdiar for RSD har noko å bety for vidare prediksjonar. Det er ingen tydeleg samanheng mellom auke i RSD og auke i responsresidual. Responsresiduala er ganske likt

fordelt rundt 0, men det er ei lita overvekt av positive residual som har høg RSD, dette er vist i plottet i figur 4.7.5. Basert på vurderinga av desse faktorane kan ein konkludere med at denne modellen er ein tilfredsstillande modell for prediksjon av brennverdi.

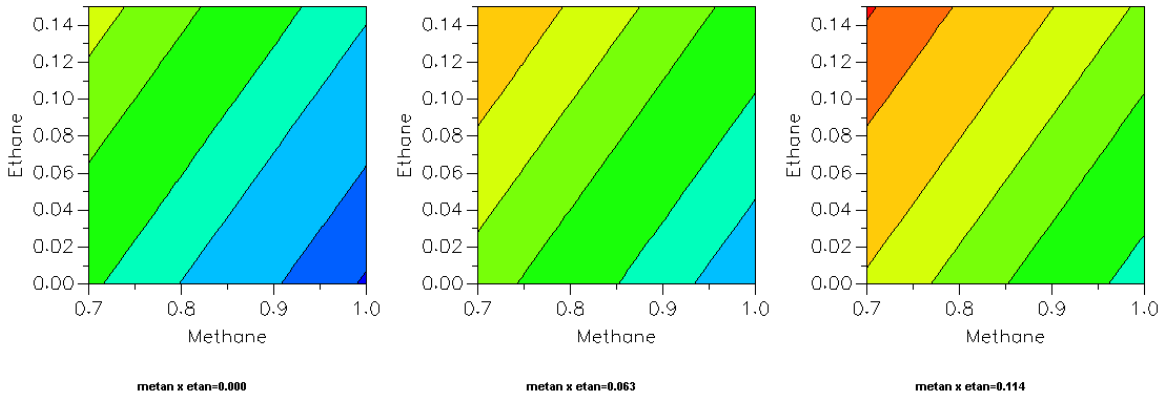


Figur 4.7.4 Normalfordelingsplott av responsresidual mot objekt



Figur 4.7.5 Responsresidual plotta mot RSD

Konturplott for brennverdi



Figur 4.7.6 Konturplott for brennverdi

Som ein kan sjå i figur 4.7.6 har ein høgast verdi for brennverdi når vekselverknadsleddet metan x etan har høg verdi i kombinasjon med låge verdiar for metan og høge verdiar for etan. Lågast verdi for brennverdi har ein når metan x etan er 0, etanverdien er låg og metanverdien er høg.

Modellen for brennverdi har høg grad av forklart varians i y og tilsynelatande god prediktiv evne.

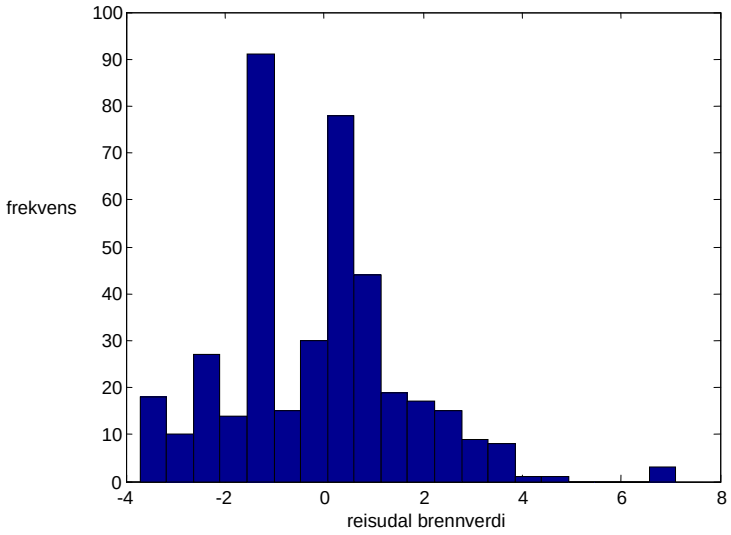
4.7.2 Prediksjon av brennverdi ved iterativ konsentrasjonsbestemming av kjemisk samansetjing ved AR

Modellen for brennverdi som er presentert i kapittel 4.7.1 skal no nyttast til prediksjon av brennverdi ved kjemisk samansetjing funnen ved iterativ konsentrasjonsbestemming ved AR.

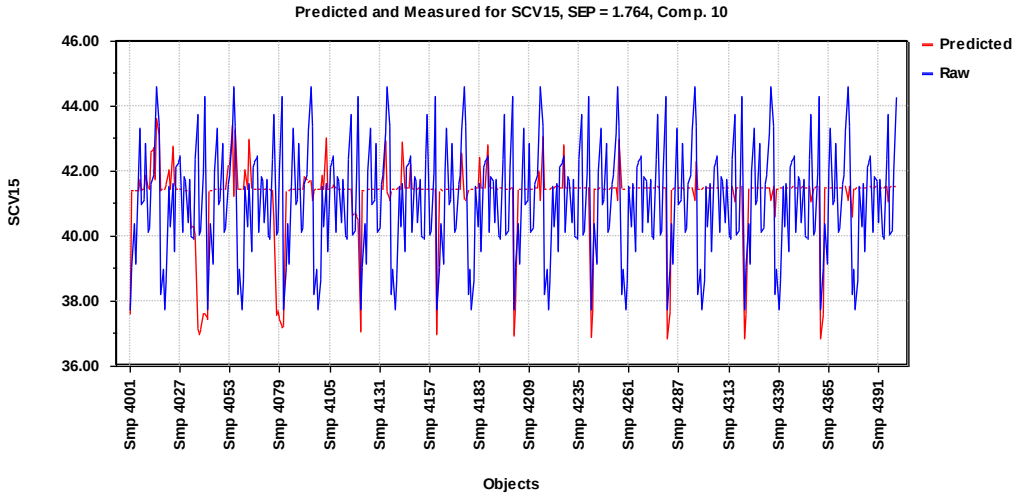
Tabell 4.7.2 Resultat frå prediksjon av brennverdi ved iterativ konsentrasjonsbestemming av kjemisk samansetjing ved AR

	Maksimumsfeil pos, MJ/Sm^3	Maksimumsfeil neg ,MJ/Sm^3	Prediksjonsfeil, MJ/Sm^3	Prediksjonsfeil, %	Objekt predikert
Brennverdi	- 3.7	7.1	1.4	3.4	400

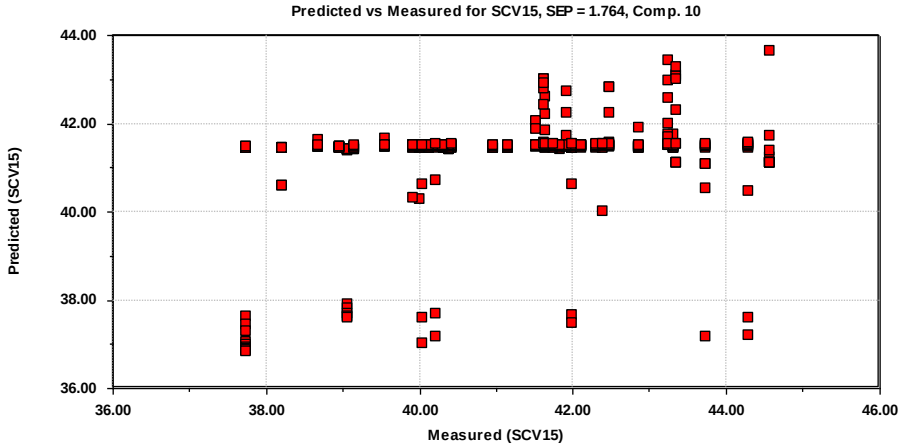
Prediksjon av brennverdi for dei 400 objekta i testdatasettet er utført med modellen som er presentert i kapittel 4.7.1 Den kjemiske samansetjinga er funne ved iterativ konsentrasjonsbestemming ved AR på seks likningar. Maksimumsfeilen er -3.7 MJ/Sm^3 i negativt retning og 7.1 MJ/Sm^3 i positiv retning. Prediksjonsfeilen er 1.4 MJ/Sm^3, dette svarar til ein prediksjonsfeil i størrelsesorden 3.4 %. (Tabell 4.7.2). Histogrammet i figur 4.7.7 viser at residuala fordeler seg med to toppar, ein ved ca -1 MJ/Sm^3 og ein rett over 0 MJ/Sm^3. Prediksjonane av brennverdi er, som ein kan sjå frå plottet i figur 4.7.8, svært upresise. Det er ingen samanheng mellom målt og predikert verdi (figur 4.7.9). Prediksjonane når aldri opp til dei høgaste målte verdiane. Ein kan her få inntrykk av at prediksjonane er tilfeldige i forhold til målt verdi.



Figur 4.7.7 Histogram av residuala ved prediksjon av brennverdi med iterativ konsentrasjonsbestemming ved AR av kjemisk samansetjing



Figur 4.7.8 Plott av predikert og målt verdi for brennverdi



Figur 4.7.9 Plott av predikert mot målt verdi for brennverdi.

Ein har no forsøkt å predikere brennverdi frå trykk, temperatur, lydhastigheit og ei predikert kjemisk samansetjing funnen ved iterativ konsentrasjonsbestemming ved AR. Ein ser at ein får svært dårleg samanheng mellom målt og predikert verdi for brennverdi og at prediksjonsfeilen er i størrelsesorden 3.4 %.

4.7.3 Prediksjon av brennverdi ved oppbygging av kjemisk samasetning ved prediksjon

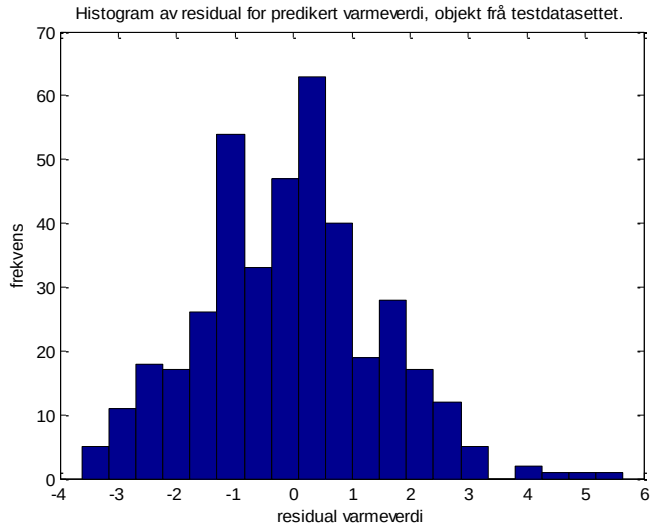
Ein vil no nytte den kjemiske samansetjinga ein har funne ved prediksjon til å predikere brennverdi.

Det nye datasettet med dei predikerte kjemiske samansetjingane er behandla slik at andregradsledd og vekselverknadsledd er inkludert. Deretter er brennverdi predikert ved bruk av modellen som er beskriven i kapittel 4.7.1.

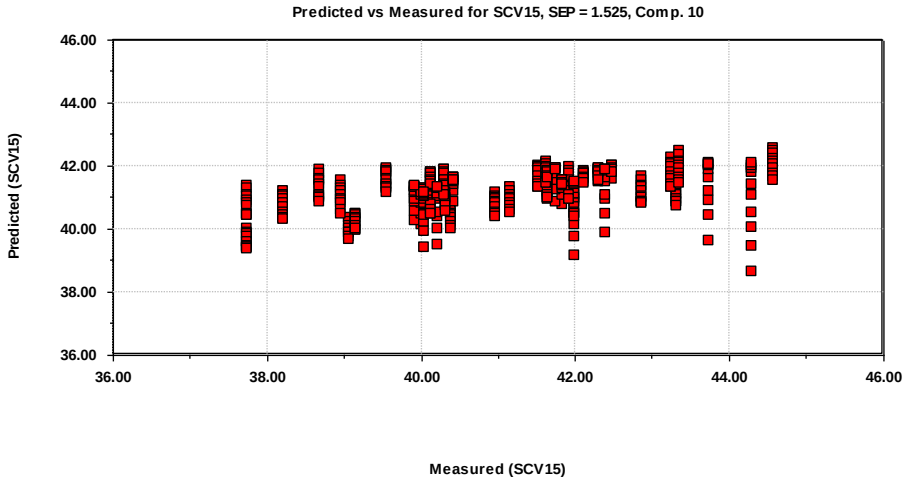
Tabell 4.7.3 Resultat av prediksjon av brennverdi

	Maksimumsfeil pos, MJ/Sm^3	Maksimumsfeil neg, MJ/Sm^3	Prediksjonsfeil, MJ/Sm^3	Prediksjonsfeil, %	Objekt predikert
Brennverdi	5.65	-3.62	1.2	3.0	400

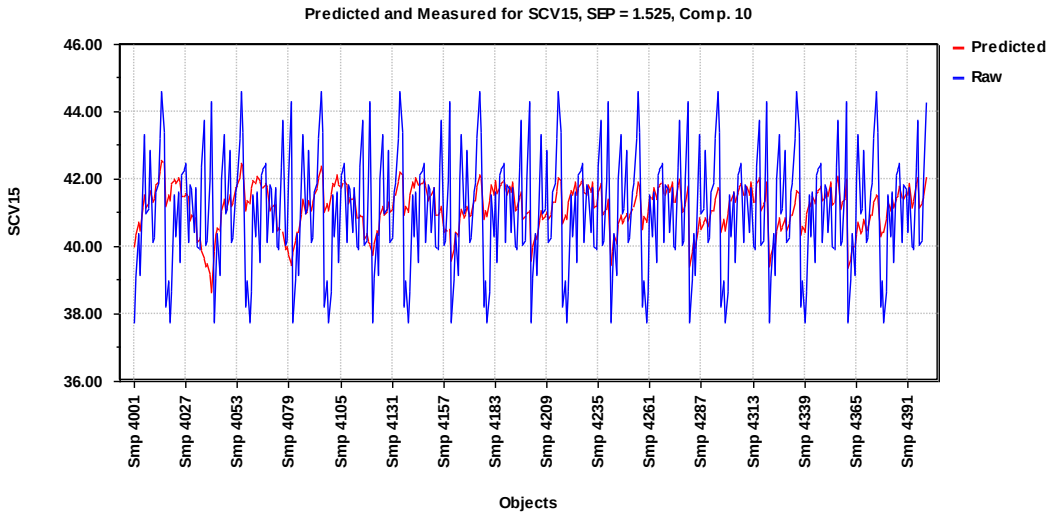
I tabell 4.7.3 ser ein at gjennomsnittfeilen for prediksjon av brennverdi er 1.2 MJ/Sm^3, noko som tilsvarar ein feil på 3.0 %. Gjennomsnittsverdien for brennverdi i testdatasettet er 41.10 MJ/Sm^3. Maksimumsfeilen i prediksjonen er større i positiv retning enn i negativ retning. Histogrammet i figur 4.7.10 viser at residuala er fordelt omlagt likt rundt 0. Dei største positive residuala finn ein for objekt som har høgt trykk og låg temperatur og dei største negative residuala finn ein også for objekta med låg temperatur. Det er dårleg samanheng mellom predikert verdi og målt verdi. (Figur 4.7.11). Ein ser også at dei predikerte verdiane har mindre spenn enn dei målte verdiane og aldri når verken høgt nok eller lågt nok i forhold til målt verdi, dette er vist i figur 7.7.12.



Figur 4.7.10 Histogram av residual for predikert brennverdi



Figur 4.7.11 Plott av predikert verdi mot målt verdi for brennverdi



Figur 4.7.12 Plott av predikert og målt verdi for brennverdi mot objekt

Ein har altså forsøkt å nytte den predikerte kjemiske samansetjinga saman med trykk, temperatur og lydhastigheit til å predikere brennverdi for dei 400 objekta i testdatasettet. Også her blir resultatet at ein får dårleg samanheng mellom målt og predikert verdi. Prediksjonsfeilen er i størrelsesorden 3.0 %.

4.7.4 Prediksjon av brennverdi med oppbygging av kjemisk samansetjing ved revers prediksjon

Ein vil no forsøke å predikere brennverdien i naturgass frå variablane trykk, temperatur og lydhastigheit saman med den kjemiske samansetjinga ein har funne ved revers prediksjon.

Fordelen med revers prediksjons slik det er presentert i kapittel 4.3 er at ein får ein betre modell til å predikere metan med. Metan er den største bestanddelen i naturgassen, og ein antar difor at det er viktig at denne blir predikert best mogleg. Ulempa med denne metoden vil vere at det er relativt store prediksjonsfeil i dei andre komponentane som vil påverke prediksjonen av metan. Føremålet er heile tida å predikere brennverdi og tettleik med minst mogleg feil, prediksjon av den kjemiske samansetjinga er berre eit steg på vegen dit. Feilen i den kjemiske samansetjinga er difor underordna feilen i prediksjonen av brennverdi og tettleik. Prediksjonane er utført med data frå testdatasettet.

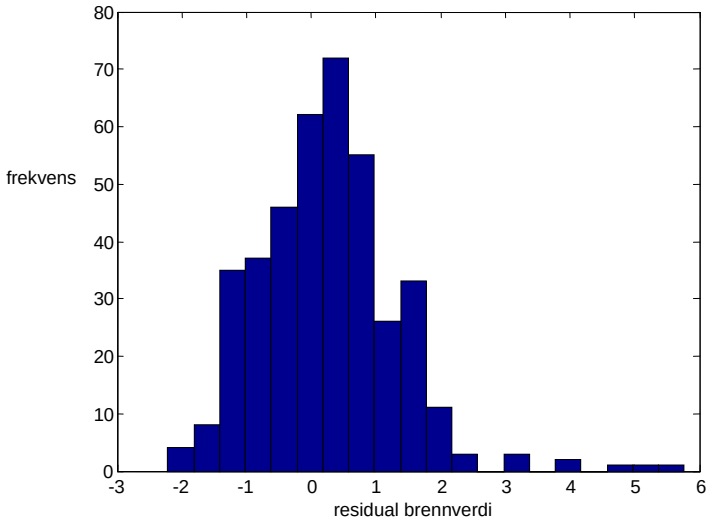
Modellane som er nytta til prediksjon av brennverdi er presentert i kapittel 4.7.1.

Tabell 4.7.4 Resultat av prediksjon av brennverdi, revers prediksjon av kjemisk samansetjing

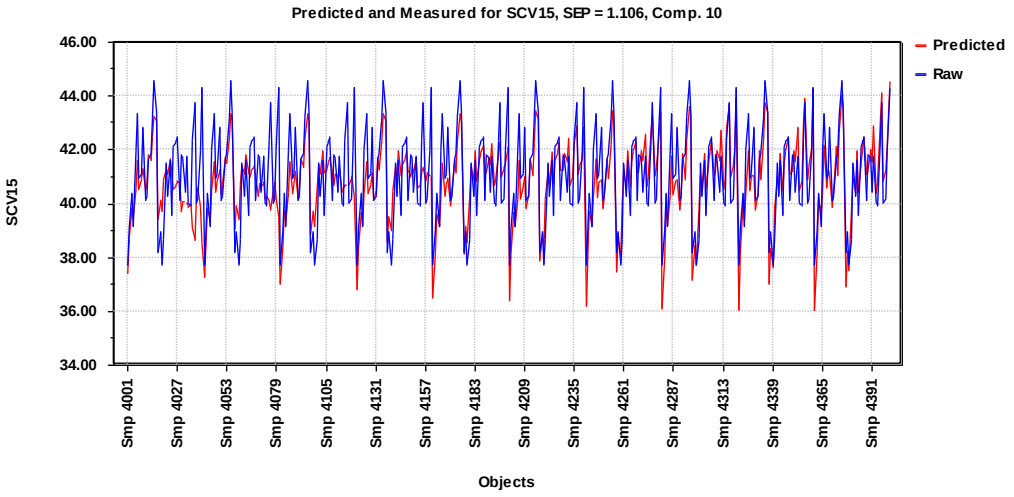
	Maksimumsfeil pos, MJ/Sm^3	Maksimumsfeil neg, MJ/Sm^3	Prediksjonsfeil, MJ/Sm^3	Prediksjonsfeil %	Objekt predikert
Brennverdi	5.7	-2.2	0.8	2.0	400

I tabell 4.7.4 ser ein at maksimumsfeilen i negativ retning er -2.2 MJ/Sm^3. Prediksjonsfeilen er 0.8 MJ/Sm^3, ein verdi som tilsvarar prediksjonsfeil i størrelsesorden 2.0 %. Histogrammet i figur 4.7.13 viser at hovudtyngda av residuala ligg i området -2 MJ/SM^3 til 2 MJ/Sm^3 med ein topp rundt 0.5 MJ/Sm^3. Frå plottet i figur 4.7.14 ser ein at det er samanheng mellom predikert og målt verdi. Ved å studere plottet i figur 4.7.15 ser ein at dei største residuala ser ein for objekta med høg brennverdi. Ein veit at temperaturen aukar utover i datasettet og

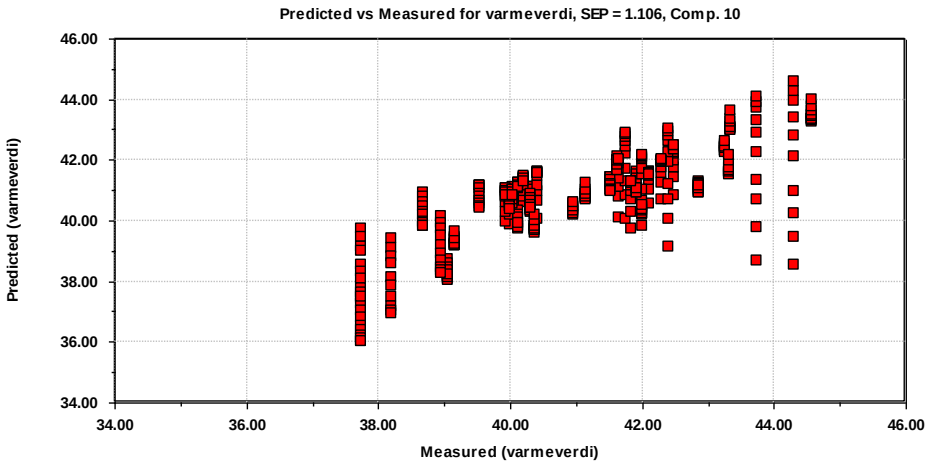
ein ser også frå plottet i figur 4.7.15 at dei predikerte verdiane blir lågare jo lenger utover i datasettet ein kjem. Dette fører til at det er dei objekta som har høg brennverdi får dei største positive residuala, noko ein også kan sjå av plottet i figur 4.7.14.



Figur 4.7.13 Histogram av residuala for brennverdi



Figur 4.7.14 Plott av predikert og målt verdi for brennverdi mot objekt



Figur 4.7.15 Plott av predikert brennverdi mot målt brennverdi

Ein har no predikert brennverdi ut frå variablane trykk, temperatur, lydhastigheit og ei kjemisk samansetjing funnen ved revers prediksjons. Ein ser at det er betre samanheng mellom målt og predikert verdi enn det ein har sett i dei to føregåande kapitla. Prediksjonsfeilen for brennverdi er her nede i størrelsesorden 2 %.

4.7.5 Prediksjon av brennverdi ved konstant trykk og temperatur

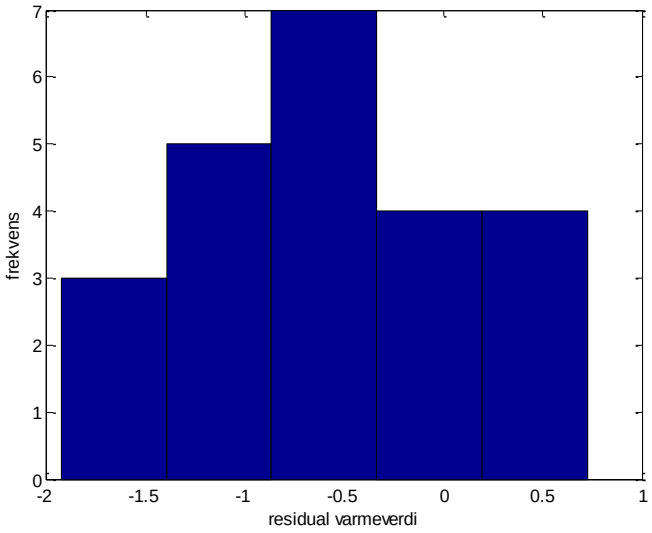
Ein vil no forsøke å nytte den kjemiske samansetjinga ein har funne for senterpunktområdet og saman med trykk, temperatur og lydhastigheit predikere brennverdi i naturgassen for dei 23 objekta i dette datasettet. Objekta i dette datasettet har trykk i området 90 – 120 bar og temperatur i området 60 – 67 ˚C.

Tabell 4.7.5 Resultat frå prediksjon av brennverdi, konstant trykk (p) og temperatur (T).

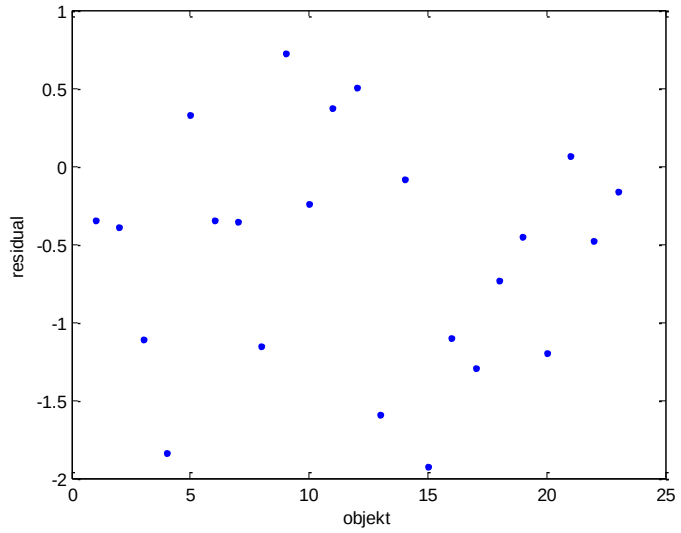
	Maksimums feil pos, MJ/Sm^3	Maksimums feil neg, MJ/Sm^3	Prediksjons feil, MJ/Sm^3	Prediksjonsfeil %	Objekt testa	Gj sn brennverdi MJ/Sm^3
Konstant p og T	0.725	-1.928	0.73	1.8	23	41.66

Prediksjon av brennverdi vart utført med åtte objekt henta frå testdatasettet og 15 objekt henta frå kalibreringsdatasettet. Den kjemiske samansetjinga er predikert ved konstant trykk

og temperatur. Ved prediksjon av brennverdi er trykk og temperatur tatt med sidan desse variablane inngår i modellen for brennverdi. Modellen som er nytta til prediksjon er presentert i kapittel 4.7.1. Som ein kan sjå frå tabell 4.7.5 og histogrammet i figur 4.7.16 er maksimumsfeilen klart størst i negativ retning. Berre 23 objekt er testa, og prediksjonsfeilen på 0.730 MJ/Sm^3 tilsvarar ein feil i størrelsesorden 1.8 %. Plottet i figur 4.7.17 viser at for prediksjon av brennverdi er overvekt av negative residual. Det er ingen skilnad på om objektet er henta frå testdatasettet eller kalibreringsdatasettet når det gjeld størrelse og forteikn på residuala i prediksjon av brennverdi. Den kjemiske samansetjinga er prediket for alle objekta, det er difor berre trykk, temperatur og lydhastigheit som er igjen av objekta i forhold til opphavlege verdiar.



Figur 4.7.16 Histogram over residuala for prediksjon av brennverdi

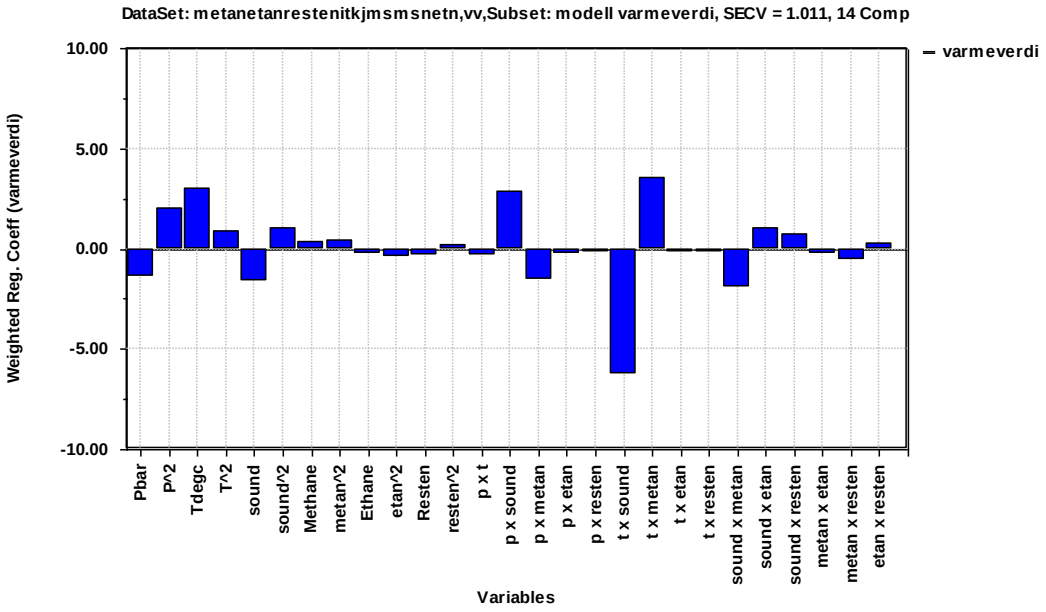


Figur 4.7.17 Plott av residuala for brennverdi

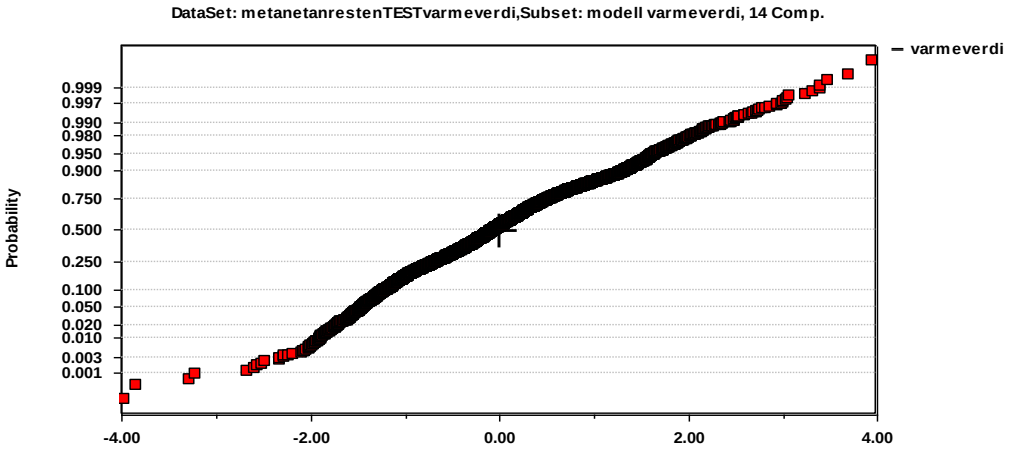
Ein har no forsøkt å predikere brennverdi i eit mindre område av datasettet der trykket i naturgassen er mellom 90 og 110 bar og temperaturen er mellom 60 og 67 ˚C. Prediksjonsfeilen er no redusert til 1.8 %. Ein ser likevel at sjølv om ein nyttar lokale modellar for trykk og temperatur blir ikkje prediksjonen av brennverdi så veldig mykje betre slik ein kunne sjå var tilfelle for tettleik.

4.7.6 Modellering av brennverdi for metan, etan og aks

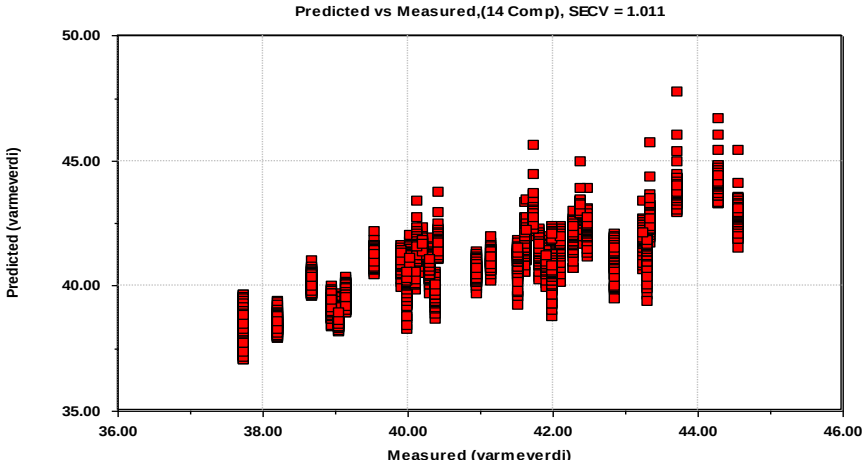
For å kunne predikere brennverdi med utgangspunkt i den kjemiske samansetjinga ein har funne ved iterativ konsentrasjonsbestemming ved AR for variablane metan, etan og aks må ein først lage ein tilfredsstillande modell for brennverdi som inneheld variablane trykk, temperatur, lydhastigheit, metan, etan og aks. Modellering av brennverdi frå variablane trykk, temperatur, lydhastigheit, metan, etan og aks gav ein modell som forklarer 69.70 % av variansen i y og har 14 komponentar. Dette er vist i tabell 4.7.6. Temperatur x metan og temperatur har dei største positive bidraga til modellen. Temperatur x sound har størst bidrag i negativ retning (figur 4.7.18). Responsresiduala er normalfordelte (figur 4.7.19), men samanhengen mellom predikert og målt verdi er ikkje tilfredsstillande (figur 4.7.20). Residuala er store både for objekta med låg brennverdi og objekta med høg brennverdi.



Figur 4.7.18 Grafisk framstilling av vekta regresjonskoeffisientar i modell for brennverdi.



Figur 4.7.19 Normalfordelingsplott av responsresiduala i modell for brennverdi



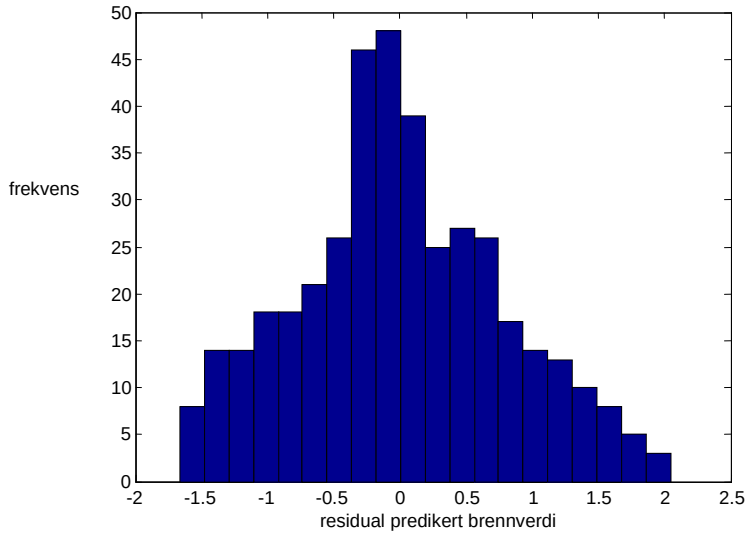
Figur 4.7.20 Plott av predikert verdi mot målt verdi for brennverdi

Validering av modell

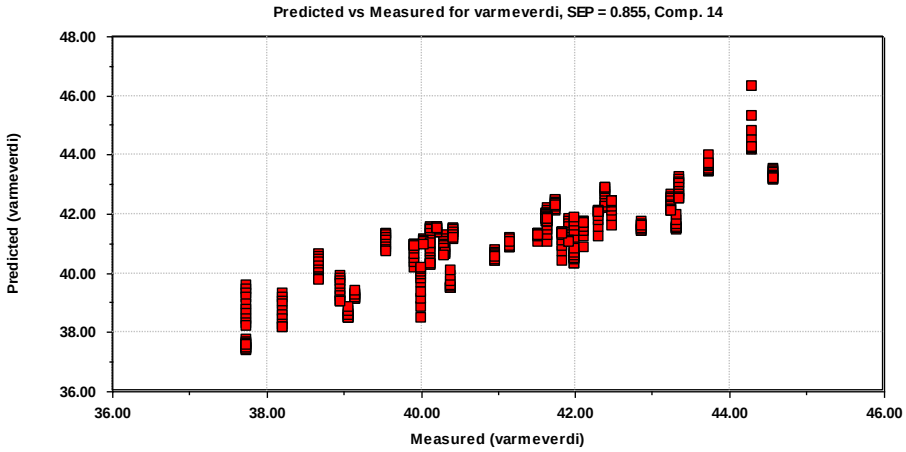
Tabell 4.7.6 Validering av modell for brennverdi

Valideringsparameter	Verdi
Forklart varians i y, %	69.70
Talet på komponentar inkludert i modellen	14
Kryssvalidering siste komponent CsvSD	0.993
RSD >	0.123
R^2	0.702
Q^2	0.690
Prediksjonsfeil MJ/Sm^3	0.74
Prediksjonsfeil %	1.8

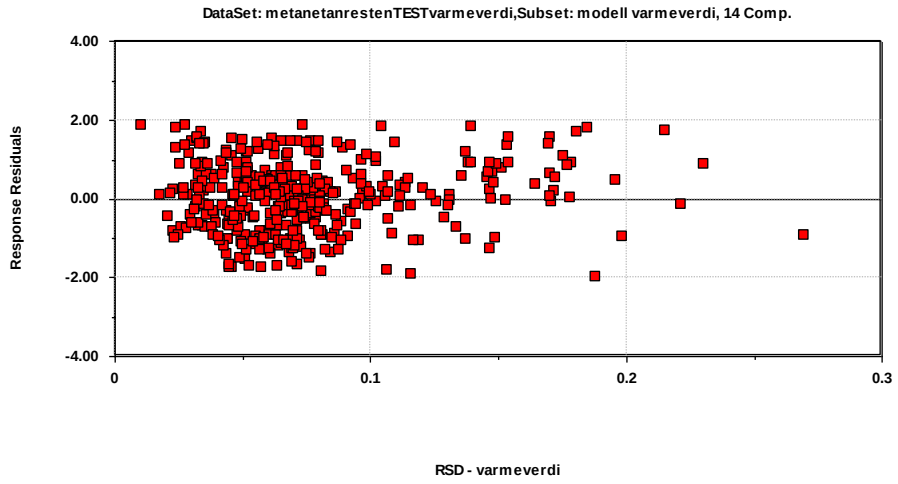
14 komponentar er inkludert i denne modellen og den siste komponenten har CsvSD-verdi 0.993. Det er like innafor grensa på 1.000 og ein kan difor inkludere denne komponenten i modellen. R^2-verdien er 0.702 og Q^2-verdien er 0.690. Dette er verdiar som peikar mot at modellen har god prediktiv evne. Prediksjonsfeilen er 0.74, dette svarar til prediksjonsfeil i størrelsesorden 1.8 %. (Tabell 4.7.6). Det er ganske bra forhold mellom predikert verdi og målt verdi (figur 4.7.22). Det ut som om det er betre samanheng mellom målt og predikert verdi etter testing med objekt frå testdatasettet enn ved testing av sjølve modellen slik plottet i figur 4.7.20 viser. Histogrammet i figur 4.7.21 viser at residuala ligg ganske jamt fordelt rundt 0. Maksimumsfeilen er litt større i positiv retning enn i negativ retning. Det ser ikkje ut til å vere nokon samanheng mellom RSD og responsresidual for objekta i modellen for brennverdi (figur 4.7.23).



Figur 4.7.21 Histogram av residual for prediksjon av brennverdi



Figur 4.7.22 Plott av prediket mot målt verdi for brennverdi



Figur 4.7.23 Plott av responsresidual mot RSD for objekt i testdatsettet.

Ein har no modellert brennverdi som ein funksjon av variablane trykk, temperatur, lydhastigheit, metan, etan og variabelen addert kjemisk samansetjing (aks). Modellen har forklart varians for y på 69.7 %, men modellen viser likevel relativt god prediktiv evne. Denne modellen skal brukast til prediksjon slik det er presentert i kapittel 4.7.7.

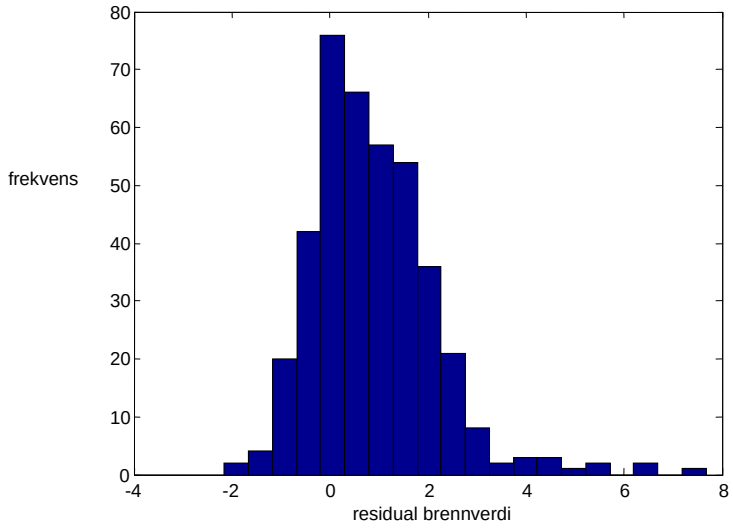
4.7.7 Prediksjon av brennverdi for situasjonen med aks

Den kjemiske samansetjinga som er funnen ved iterativ konsentrasjonsbestemming ved AR for variablane metan, etan og aks vert no nytta til prediksjon av brennverdi. Til prediksjonen har ein nytta modellen som er presentert i kapittel 4.7.6.

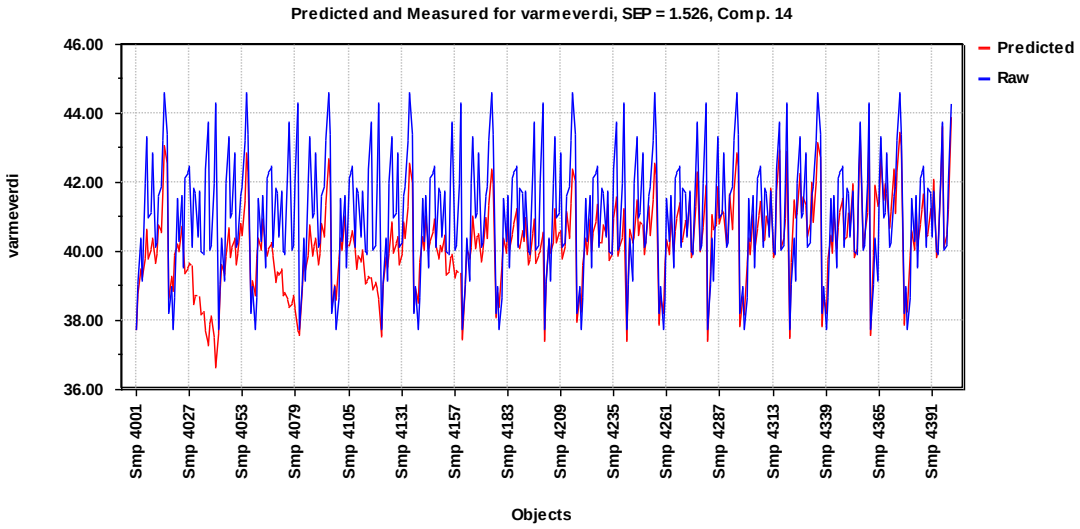
Tabell 4.7.7 Resultat av prediksjon av brennverdi

	Maksimumsfeil pos, MJ/Sm^3	Maksimumsfeil neg, MJ/Sm^3	Prediksjonsfeil MJ/Sm^3	Prediksjonsfeil %	Objekt predikert
Brennverdi	7.7	-2.2	1.1	2.7	400

I tabell 4.7.7 ser ein at maksimumsfeilen i positiv er ganske høg retning for brennverdi. Histogrammet i figur 4.7.13 viser at hovudtyngda av residuala ligg mellom -2.2 MJ/Sm^3 og ca 3 MJ/Sm^3. Det er berre få objekt som har høgare residual. Prediksjonsfeilen er 1.1 MJ/Sm^3, dette svarar til 2.7 %. Ein skulle kanskje forventa at resultata for brennverdi vart enda dårlegare sidan forklart varians for y ikkje er høgare enn 69.70 %. Innsendt verdi i prediksjonen er 0 for etan og dermed 0 også for alle vekselverknadsledd etan inngår i. Etan og vekselverknadsledd med etan har likevel så lite å bety for modellen at dette sannsynlegvis påverkar prediksjonen i liten grad. I figur 4.7.25 ser ein at objekt med kombinasjonen låg temperatur og lågt trykk har klart dårlegast predikert verdi for brennverdi. Dei beste prediksjonane ser ein til høgre i plottet der temperaturen er høg.



4.7.24 Histogram over residual for prediksjon av brennverdi

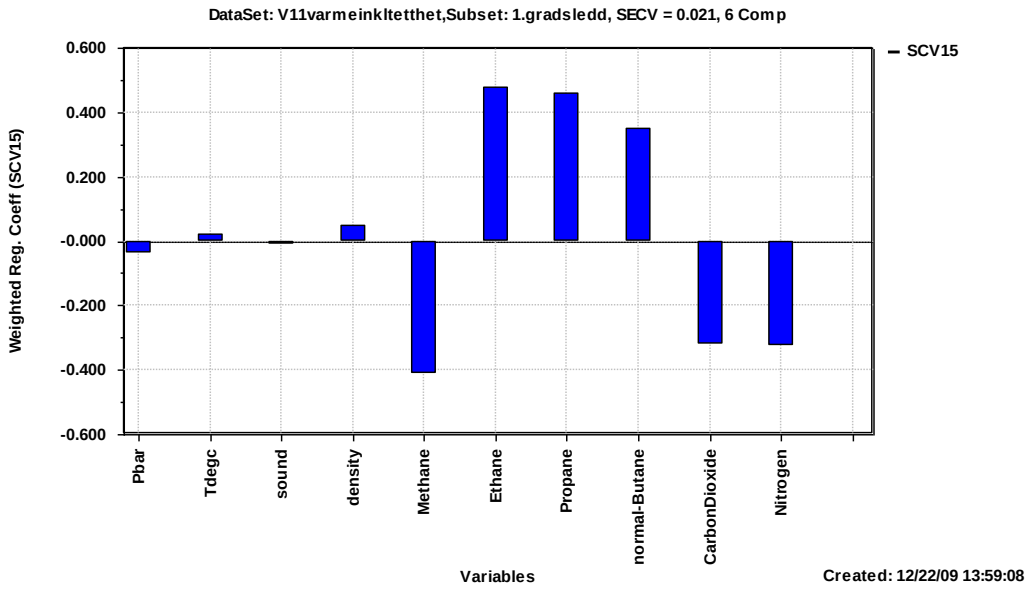


Figur 4.7.25 Plott av predikert og målt verdi for brennverdi

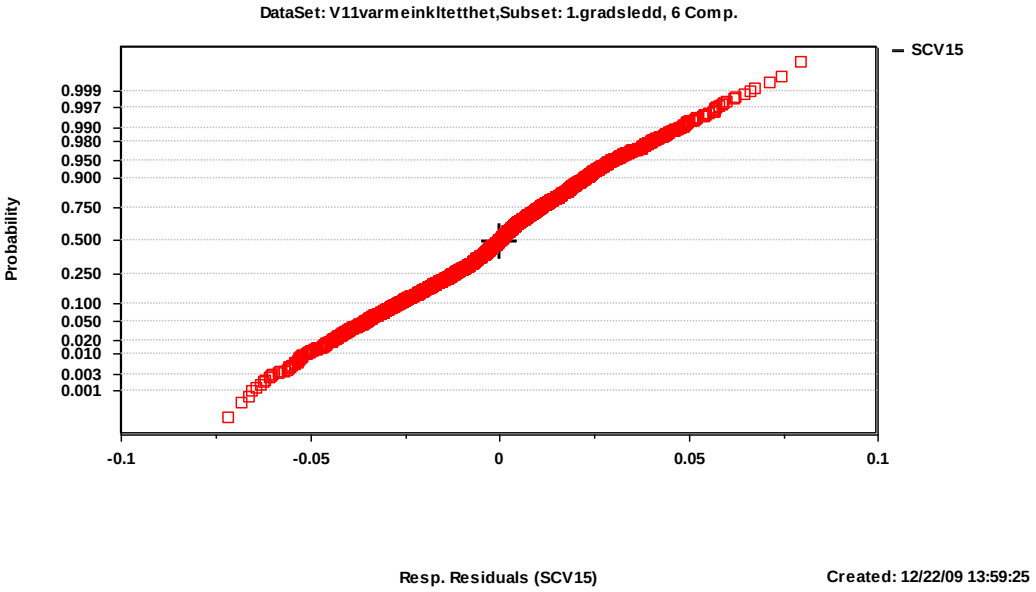
Ein har no predikert brennverdi i situasjonen der ein berre har tre kjemiske komponentar tilstades, metan, etan og addert kjemisk samansetjing som består av summen av propan, butan, CO_2 og N_2. Ein ser at prediksjonsfeilen er i størrelsesorden 2.7 % og at samanhengen mellom målt og predikert verdi er best for objekta med høge temperaturar.

4.7.8 Modellering av brennverdi med tettleik som ein del av X

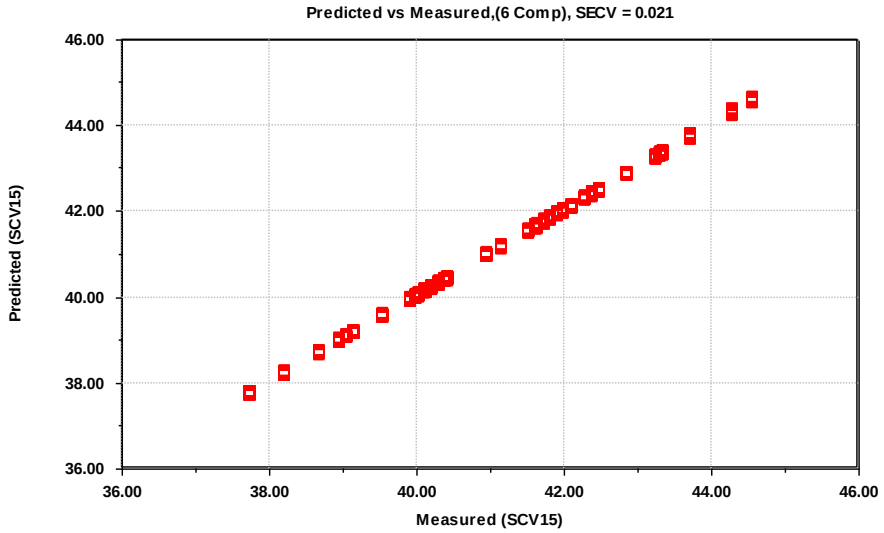
Det vil i mange tilfeller vere mogleg å inkludere ein sensor som kan måle tettleik i naturgassen på installasjonane som måler lydhastigheit, trykk og temperatur. Med nøyaktige målingar av tettleik kan denne parameteren inkluderast i modellen for brennverdi. Tettleik vart difor inkludert i **X**-matrisa og ein modell med førstegradsledd vart bygd. Denne modellen inneheld seks komponentar og forklarer heile 99.99 % av variansen i y. Det er framleis dei kjemiske komponentane som dominerer i forhold til kva variablar som har store bidrag til modellen (figur 4.7.26). Bidraget frå tettleik er svært lite, og det same gjeld for variablane trykk, temperatur og lydhastigheit. I modellen ser ein normalfordelte responsresidual (figur 4.7.27) og svært tilfredsstillande samanheng mellom målt og predikert verdi (figur 4.7.28).



*Figur 4.7.26 Grafisk framstilling av vekta regresjonskoeffisientar i modell for brennverdi der tettleik er ein del av **X**.*



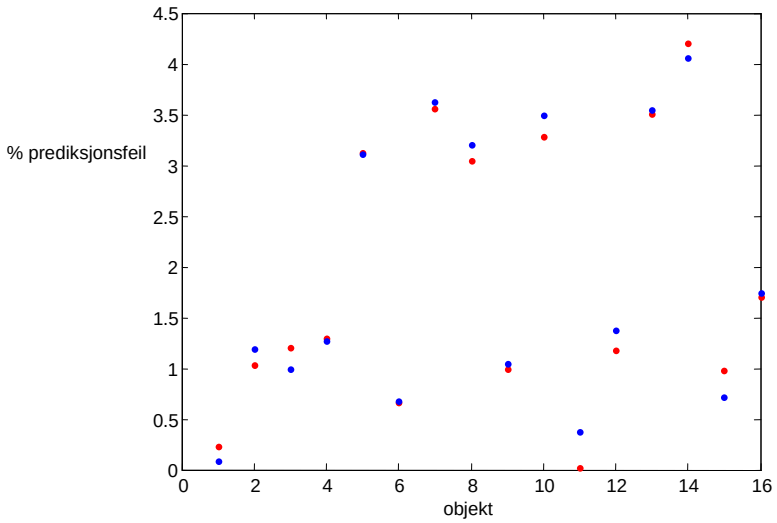
*Figur 4.7.27 Normalfordelingsplott av responsresidual i modell for brennverdi der tettleik er ein del av **X**.*



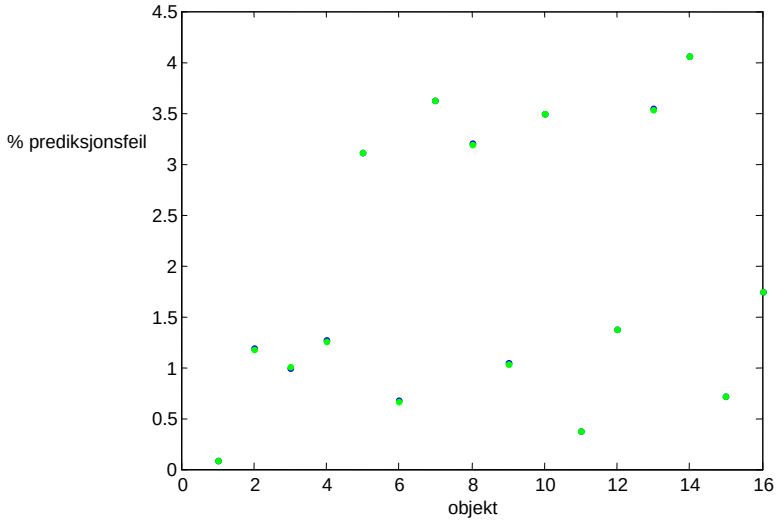
*Figur 4.7.28 Plott av predikert verdi mot målt verdi for modell for brennverdi med tettleik som ein del av **X**.*

Prediksjon av brennverdi vart utført ved konstant trykk (90 – 110 bar) og temperatur (67˚C). Den kjemiske samansetjinga vart predikert ved å predikere ein og ein komponent. Ein starta med prediksjon av metan og deretter predikerte ein i denne rekkjefølgja: etan, propan, CO_2, N_2, butan på same måte som beskrive i kapittel 4.2. Tettleik vart predikert ut frå modellen presentert i kap 4.6.1. Testdatasettet bestod av 16 objekt. Ser i figur 4.7.29 at ut frå testing med dette datasettet blir prediksjonane av brennverdi ikkje betre sjølv om ein har tettleik inkludert i modellen. Ved å bytte ut dei predikere verdiane med simulerte verdiar for tettleik

får ein heller ikkje betre verdiar for brennverdi. (Figur 4.7.30). Det er altså ingenting å vinne på å inkludere tettleik i modellen for brennverdi basert på desse resultata. Dette forklarar ein ut frå plott av regresjonskoeffisientar for modell for brennverdi med tettleik inkludert, her ser ein at tettleik har svært lite bidrag til modellen. I ladningsplottet for komponent 1 mot komponent 2 og komponent 1 mot komponent 3 i kapittel 4.1.1 ser ein at det er om lag 90 grader mellom tettleik og brennverdi, desse variablane er altså i utgangspunktet ukorrelerte.



Figur 4.7.29 Plott av % feil for prediksjon av brennverdi. Raud er modell for brennverdi utan tettleik inkludert, blått er modell for brennverdi med tettleik inkludert.



Figur 4.7.30 Plott av % feil for prediksjon av brennverdi. Blått er predikert verdi for tettleik, grønt er målt verdi for tettleik.

Ein har no undersøkt om prediksjonen av brennverdi vert betre dersom ein inkluderer tettleik som ein del av **X**. Prediksjonen er utført med konstant trykk og temperatur. Resultata som er presentert her viser at ein ikkje oppnår betre prediksjonar for brennverdi ved å inkludere tettleik i **X**. Dette forklarar ein med at tettleik har svært lite bidrag til modellen slik det er vist i figur 4.7.28.

4.8 Samanlikning av prediksjonar av tettleik og brennverdi

I dette kapitlet vert resultata av prediksjonane av tettleik og brennverdi oppsummert og samanlikna.

Tabell 4.8.1 Samanfatning av prediksjonar av tettleik og brennverdi ved ulike måtar å predikere den kjemiske samansetjinga

		Iterativ konsentrasjons bestmming ved AR	Oppbygging ved prediksjon	Revers prediksjon	Konstant p og T	Metan, etan, aks
Metan	Prediksjonsfeil	0.070	0.050	0.026	0.095	0.066
	Prediksjonsfeil %	8.1	5.8	3.0	11.0	8.1
Etan	Prediksjonsfeil	0.078	0.036	0.028	0.019	0.077
	Prediksjonsfeil %	101.3	48	37.3	27.0	102.7
Tettleik	maksimumsfeil neg	-28	-18	-10	-2.1	-13
	maksimumsfeil pos	115	98	98	4.6	114
	Prediksjonsfeil	9.0	6.4	5.5	1.0	8.2
	Prediksjonsfeil %	9.0	6.5	5.5	1.3	8.3
Brennverdi	maksimumsfeil pos	7.1	5.8	5.8	0.7	7.7
	maksimumsfeil neg	-3.7	-3.6	-2.2	-1.9	-2.2
	Prediksjonsfeil	1.4	1.2	0.8	0.7	1.1
	Prediksjonsfeil %	3.4	2.9	2.0	1.8	2.7

I tabell 4.8.1 ser ein at det er ganske store skilnader på prosent feil ved prediksjon av brennverdi og tettleik avhengig av korleis ein har predikert den kjemiske samansetjinga. Predikert feil er her eit utrykk for summen av absoluttverdien av responsresiduala dividert på gjennomsnittsverdien i det området modellen gjeld (formel (2.23)). I denne oppgåva har ein hovudsakleg modellert og predikert med heile datasettet. Ein har også sett at kor godt ein er i stand til å predikere kjemiske samansetjinga er avhengig av variablane trykk og temperatur. I områder der ein har kombinasjonen høg temperatur og lågt trykk ser ein store

residual. Dette vil sjølvsagt påverke prediksjonen av brennverdi og tettleik. Den kjemiske samansetjinga har meir å bety for brennverdi enn for tettleik, og feil i prediksjon av kjemisk samansetjing vil truleg ha større utslag for brennverdi enn for tettleik. I tabellen ser ein at prediksjonsfeilen for tettleik går dramatisk ned når ein nyttar lokal modell, altså ein modell der trykk og temperatur er haldne konstante. Her er prediksjonsfeilen for tettleik berre 1.3 %. For brennverdi ser ein også at prediksjonsfeilen er lågare i dette området, 1.8 %. Skilnaden er likevel tydelegast for tettleik. Metan og etan i tabellen over representerer den kjemiske samansetjinga i dei ulike situasjonane. Desse to kjemiske komponentane er hovudbestanddelane i naturgassen og prediksjonane av desse er dermed viktigast. For metan er det prediksjon ved revers oppbygging som gjer den lågaste feilen uttrykt i prosent, men for etan er det situasjonen med konstant trykk og temperatur som gir best prediksjon. Både for tettleik og brennverdi ser ein best prediksjon i situasjonen der ein har halde trykk og temperatur konstant. Dette peikar mot at lokale modellar vil gje best resultat når det gjeld å predikere brennverdi og tettleik frå ei predikert kjemisk samansetjing. Ved revers predikert oppbygd kjemisk samansetjing er prediksjonsfeilen for brennverdi i størrelsesorden 2.0 %. For tettleik er talet heile 5.5 %. Her må ein ta omsyn til at heile den kjemiske samansetjinga er predikert, situasjonen er altså tilsvarande BCA1. Prediksjonsfeilen for metan er 3.0 % i situasjonen med revers prediksjon, det er altså i denne situasjonen at metan vert predikert best. Ein har tidlegare sett at den kjemiske samansetjinga ikkje er viktig i modellen for tettleik, men at den spelar viktig rolle i modellen for brennverdi. Det er årsaka til at brennverdi blir predikert så mykje betre ved revers prediksjon enn tettleik. Prediksjonsfeilen går frå 1.8 % ved konstant trykk og temperatur til 2.0 % ved revers oppbygging. For tettleik går prediksjonsfeilen opp frå 1.3 % ved konstant trykk og temperatur til heile 5.5 % for revers oppbygging. Denne høge feilen er knytt til dei store problema ein får i situasjonen med høgt trykk og låg temperatur. Dette unngår ein i tilfellet der ein har modellar med konstant trykk og temperatur. Dei største prediksjonsfeila for både metan, tettleik og brennverdi ser ein i situasjonen der utfører iterativ konsentrasjonsbestemming ved AR med gjennomsnittsverdiar for dei kjemiske komponentane som startverdiar.

5 Konklusjon og forslag til vidare arbeid

5.1 Konklusjon

Målet med denne oppgåva var å undersøkje om det er mogleg å modellere og predikere brennverdi og tettleik i naturgass utan å kjenne den kjemiske samansetjinga i naturgassen. Denne situasjonen vert kalla BCA1 [8]. Det er forsøkt ulike innfallsvinklar for å modellere den kjemiske samansetjinga. Det viste seg at det som gav best resultata i forhold til prediksjon av brennverdi og tettleik frå eit område med trykk 0 – 200 bar og temperatur -10 – 100 ˚C var å byggje opp en kjemiske samansetjinga ved revers prediksjon av ein og ein kjemisk komponent. Ein såg også at ved å nytte lokale modellar for senterpunkta for trykk (100 bar) og temperatur (60 ˚C) fekk ein forbetra prediksjonane av tettleik betydeleg. Dette skuldast at prediksjon av tettleik er meir kjenslevar for endringar i temperatur og trykk enn det ein ser for prediksjon av brennverdi. Modellen for brennverdi er i større grad avhengig av gode prediksjonar av den kjemiske samansetjinga. Basert på resultata funne i denne oppgåva fekk ein ikkje betre prediksjon av brennverdi ved å inkludere tettleik i modellen. I tidlegare arbeid utført ved CMR [9], [11] ser ein at det er observert feil i området -1.5 – 2.0 % for tettleik og 0 – 1 % for brennverdi i situasjonen BCA1. I desse tilfella er trykket og temperaturen haldne relativt konstant. Ved å modellere på heile datasettet ser ein i denne oppgåva at ein oppnår prediksjonsfeil i størrelesesorden 5.6 % for tettleik og 2.0 % for brennverdi. Dersom ein derimot opererer med lokale modellar slik det er gjort når temperatur og trykk er haldne konstante innanfor eit visst område ser ein at prediksjonsfeilen for tettleik fell til 1.3 % og prediksjonsfeilen for brennverdi vert redusert til 1.8 %. Ved bruk av lokale modellar ser ein størst reduksjon i feilen for tettleik sidan modellen for denne parameteren i stor grad er avhengig av trykk og temperatur.

5.2 Forslag til vidare arbeid

Det vil først og fremst vere nyttig med vidare undersøkjingar med bruk av lokale modellar. Dette gjeld særleg for prediksjon av tettleik der ein ser at modellen i stor grad er avhengig av trykk og temperatur. Ein ser at prediksjonane for tettleik får store residual ved situasjonen låg temperatur og høgt trykk. Dette gjeld også ved modellering av kjemiske komponentar og brennverdi.

Når det gjeld forslaget om å slå saman dei kjemiske komponentane med små konsentrasjonar vil det kunne vere interessant å undersøkje kor stor innverknad dei inerte gassane har på predikert verdi av tettleik og brennverdi. Korleis ville resultata blitt dersom ein ikkje hadde inkludert CO_2 og N_2 i aks?

Det vil også vere interessant å undersøkje om ikkje-lineære samanhengar i modellane vil kunne gje enda betre prediksjonar. Ein kunne til dømes sett på eksponensielle samanhengar.

Ein kan forsøkje å nytte alternerande regresjon til raffinering av modellane, der variablane lydhastigheit, tettleik og brennverdi skiftar på å vera prediktor og respons i ein syklus inntil konvergens.

Det vil til slutt også vere interessant å teste modellane på reelle data frå ulike felt i Nordsjøen.

Referanseliste

[1]. Speight, J.G., *Historical perspectives* i *The Chemistry and Technology of Petroleum, Fourth Edition*. 2007, CRC Press Taylor & Francis Group, Boca Raton, s 20 – 22.

[2]. Carlson, J.E. og Carlson, R*., Prediction of Molar Fractions in Two-Component Gas Mixture Using Pulse- Echo Ultrasound and PLS Regression*. 2006, IEEE Transactions on Ultrasonic, Ferroelectrics, and Frequent Control, 53, s 606 – 613.

[3]. Westad, F., et al., *Akustikk* i *Anvendelse av kjemometri innen forskning og industri*, Nordtvedt, R., et al., (Redaktørar). 1996, Tidsskriftforlaget Kjemi AS: Bergen, s 297 - 303.

[4]. Wade, A.P., *Acoustic Emission: Is Industry Listening?*. 1990, Chemometrics and Intelligent Laboratory Systems, 8, s 305 – 310.

[5]. Hauptmann, P., et al., *Application of ultrasonic sensors in the prosess industy*. 2002, Meas. Sci. Technol., 13, s R73 – R 83.

[6]. Lueptow, R.M. og Phillips, S., *Acoustic sensor for determining combustion properties of natural gas*. 1994, Meas. Sci. Technol., 5, s 1375 – 1381.

[7]. Florisson, O. og Burrie, P.H., *Rapid determination of the Wobbe index of natural gas*. 1989, Phys. E: Sci. Instrum., 22, s 12 – 128.

[8]. Frøysa, K.E., et al., *Mass and energy measurement of natural gas using ultrasonic flow meters. Results from testing on various North Sea gas field data*. 2006, 29 th Scandinavian Symposium on Physical Acoustics, Ustadoset, Norway, 29 January – 1 February.

[9]. Frøysa, K.E. og Lunde, P., *Density and calorific value measurement in natural gas using ultrasonic flow meters.* 2005, 23rd International North Sea Flow Measurement Workshop, Tønsberg, Norway 18 – 21 October.

[10]. Beecroft, D., Is *a wet gas (multiphase) mass flow meter just å pipe dream?*. 1998, 16th International North Sea Flow Measurement Workshop, Glenagles Hotel, Perthshire, Scotland, 26 – 29 October.

[11]. Frøysa, K.E., et al., *Density and calorific value measurement in natural gas using ultrasonic flow meters. Results from testing on various North Sea gas field data*, 24 th International Nort Sea Flow Measurement Workshop, 24 th – 27 th October 2006.

[12]. Kvalheim, O., *Fra data til informasjon* i Anvendelse *av kjemometri innen forskning og industri*, Nordtvedt, R., et al., (Redaktørar). 1996, Tidsskriftforlaget Kjemi AS: Bergen, s 53.

[13]. Montgomery, D.C., *Describing Variation*, i *Introduction to Statistical Quality Control*. 2005, John Wiley & Sons, Inc, s 47 – 50.

[14]. Montgomery, D.C., *Important continuous distribution*, i *Introduction to Statistical Quality Control*. 2005, John Wiley & Sons, Inc, s 61 – 66.

[15]. Bhattacharyya, G.K. og Johnson, R.A., *The Normal Distribution*, i *Statistical Concepts and Methods*. 1977, John Wiley & Sons, s 193 – 194.

[16]. Walpole, et al., *Sampling Distributions of Means* i *Probability & Statistics for engineers and scientists*. 2002,Prentice Hall, Upper Saddle River, s 208 – 214.

[17]. Bhattacharyya, G.K. og Johnson, R.A., *Checking the assumptions of a normal population*, i *Statistical Concepts and Methods*. 1977, John Wiley & Sons, s 220 – 223.

[18]. Walpole, et al., *Data Display and Graphical Methods* i *Probability & Statistics for engineers and scientists*. 2002, Prentice Hall, New Jersey. s 204 – 208.

[19]. Johnson, R.A. og Wichern, D.W, *The Multivariate Normal Distribution* i *Applied multivariate statstical analysis, sixth edition*. 2007, Pearson Prentice Hall, Upper Saddle River s 150 – 156.

[20]. Karstang, T. V., *Forbehandling av data* i *Anvendelse av kjemometri innen forskning og industri*, Nordtvedt, R., et al., (Redaktørar). 1996, Tidsskriftforlaget Kjemi AS: Bergen, s 129 - 144.

[21]. Sirius versjon 7.0, Copyright 1995 – 09, The Pattern Recognition System–help – normalisation

[22]. Brereton, R.G., *Basic Statistical Concepts* i *Chemometrics Data Analysis for the Laboratory and Chemical Plant*. 2003, John Wiley & Sons, Ltd, s 417 – 419.

[23]. Bhattacharyya, G.K. og Johnson, R.A., *Covariance and correlation*, i *Statistical Concepts and Methods*, 1977, John Wiley & Sons, s 129 - 131.

[24]. Montgomery, D.C., *Describing Variation* i *Introduction to Statistical Quality Control*. 2005, John Wiley & Sons, Inc, s 44 – 45.

[25]. Brereton, R.G., *Pattern Recognition* i *Chemometrics Data Analysis for the Laboratory and Chemical Plant*. 2003, John Wiley & Sons, Ltd, s 183 - 184.

[26]. Grung, B., *Det matematiske grunnlaget for latent variabelmetoder* i *Anvendelse av kjemometri innen forskning og industri*, Nordtvedt, R. et al., (Redaktørar). 1996, Tidsskriftforlaget Kjemi AS: Bergen, s 121 - 128.

[27]. Isaksson, T. og Næs, T., *Prinsipal komponent analyse* i *Anvendelse av kjemometri innen forskning og industri*, Nordtvedt, R., et al., (Redaktørar). 1996, Tidsskriftforlaget Kjemi AS: Bergen, s 145 - 149.

[28]. MATLAB The Language of Technical Computing, version R2007a, Copyright 1998 – 2007, The MathWorks, Inc ,PLS Toolbox MATLAB

[29]. Brereton, R.G., *Partial Least Square* i *Chemometrics Data Analysis for the Laboratory and Chemical Plant*. 2003, John Wiley & Sons, Ltd, s 297 - 303.

[30]. Martens, H. og Næs, T., *Partial Least Square Regression (PLSR)* i *Multivariate Calibration*. 1989, John Wiley & Sons, Ltd, s 116 – 125.

[31]. de Jong, S., SIMPLS*: an alternative approach to partial least square regression*. 1992, Chemometrics and Intelligent Laboratory Systems, 18, s 251 – 263.

[32]. Wold, S. et al*.*, *PLS-regression: a basic tool of chemometrics*. 2001, Chemometrics and Intelligent Laboratory System, 58, s 109 – 130.

[33]. Karjalainen, E. J., Karjalainen, U.P., *Component reconstruction in the primary space of spectra and concentrations. Alternatring regression and related direct methods*. 1991, Analysis Chimica Acta, 250, s 169-179.

[34]. Karjalainen, E.J., *The spectrum reconstruction problem. Use of alternating regression for unexpected spectral components in two-dimensional spectroscopies*. 1989. Chemometrics and Intelligent Laboratory Systems, 7, s 31-38.

[35]. Shinzawa, H., et al., *A convergence criterion in alternating least squares (ALS)*. 2008, Journal of Molecular Structure 883-884, s 73-78.

[36]. Sirius versjon 7.0, Copyright 1995 – 09, The Pattern Recognition System– help- RSD

[37]. Kvalheim, O.M og Karstang, T.V, *SIMCA-Classification by means of disjoint cross validated principal component models* i *Latent-variable modelling of multivariate data*. Department of Chemistry, University of Bergen.

[38]. Eriksson, E., et al., *Appendix II: Statistical Notes* i *Multi- and Megavariate Data Analysis Part I.* 2006, Umetrics AB, s 387 - 397.

[39]. Sirius versjon 7.0, Copyright 1995 – 09, The Pattern Recognition System-help-leverage

[40]. Martens, H. og Næs, T., *Outlier Detection* i *Multivariate Calibration*. 1989, John Wiley & Sons, Ltd, s 267 – 274

[41]. Brereton, R.G., *Model Validation* i *Chemometrics Data Analysis for the Laboratory and Chemical Plant*. 2003, John Wiley & Sons, Ltd, s 313 - 323.

[42]. Martens, H. og Næs, T., *Good and Bad Calibration* i *Multivariate Calibration*. 1989, John Wiley & Sons, Ltd, s 237 – 246.

[43]. Martens, H. og Næs, T., *Prediction Ablity* i *Multivariate Calibration*. 1989, John Wiley & Sons, Ltd, s 246 – 250.

[44]. Martens, H. og Næs, T., *Validation in Practice: estimation of MSE* i *Multivariate Calibration*. 1989, John Wiley & Sons, Ltd, s 250 - 259.

[45]. Eriksson, E., et al., *Additional PLS diagnostics* i *Multi- and Megavariate Data Analysis Part I*. 2006, Umetrics AB, s 95 – 97.

[46]. Martens, H. og Næs, T., *Using prediction error to select the calibration model* i *Multivariate Calibration*. 1989, John Wiley & Sons, Ltd, s 261 - 263.

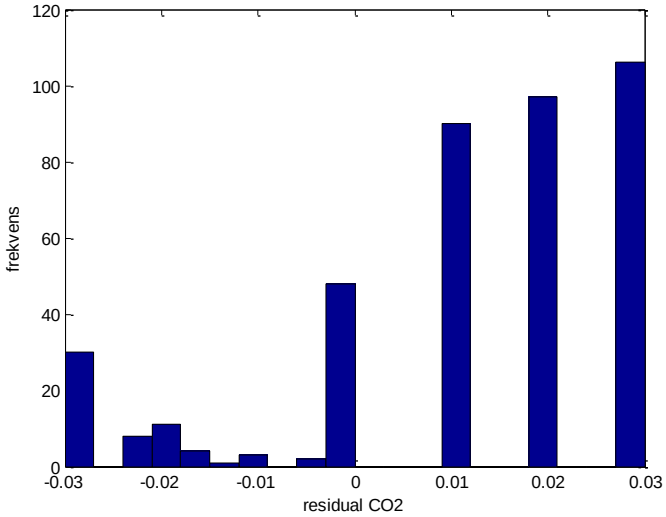
[47]. Rajalahti, T., et al., *Discriminating Variable Test and Selectivity Ratio Plot: Quantitative Tools for Interpretation and Variable (Biomarker) Selection in Complex Spctral or Chromatographic Profiles*. 2009, Anal. Chem, 81, s 2581-2590.

[48]. Christie, O.H.J., *Oljeindustri* i *Anvendelse av kjemometri innen forskning og industri*, Nordtvedt, R., et al., (Redaktørar). 1996, Tidsskriftforlaget Kjemi AS: Bergen, s 401 - 409.

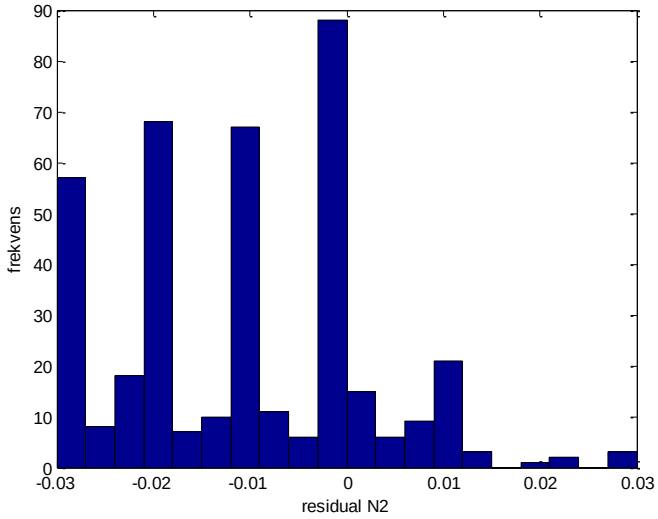
[49]. Kvalheim, O.M. og Karstang, T.M., *Interpretation of Latent-Variable Regression Models*. 1989, Chem and Intell Lab systems, 7, s 39-51.

Appendiks

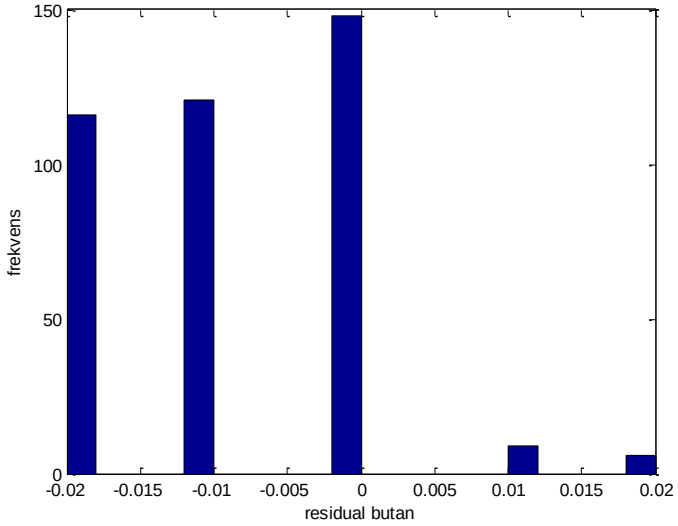
Prediksjon av kjemisk samansetjing ved iterasjon i matlab



Figur A.1.1 Histogram av residuala for CO_2 etter iterativ konsentrasjonsbestemming ved AR

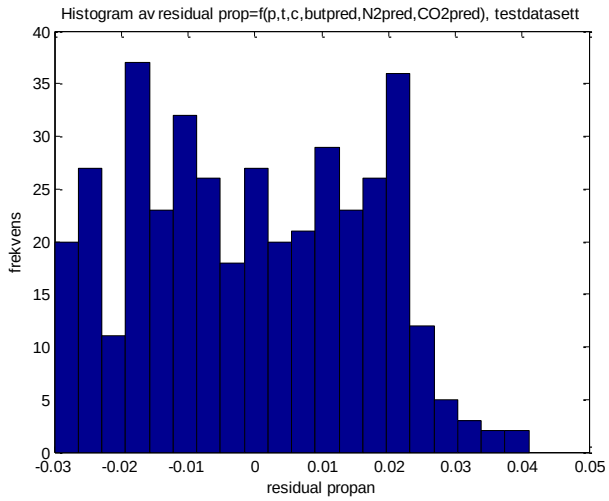


Figur A.1.2 Histogram av residuala for N_2 etter iterativ konsentrasjonsbestemming ved AR

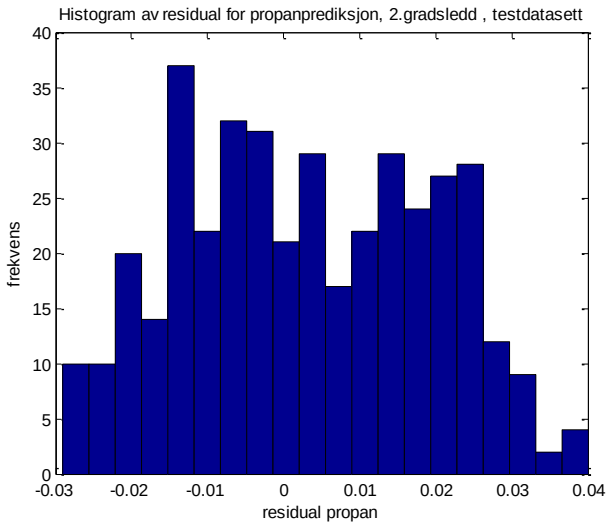


Figur A.1.3 Histogram av residuala for butan etter iterasjon konsentrasjonsbestemming ved AR.

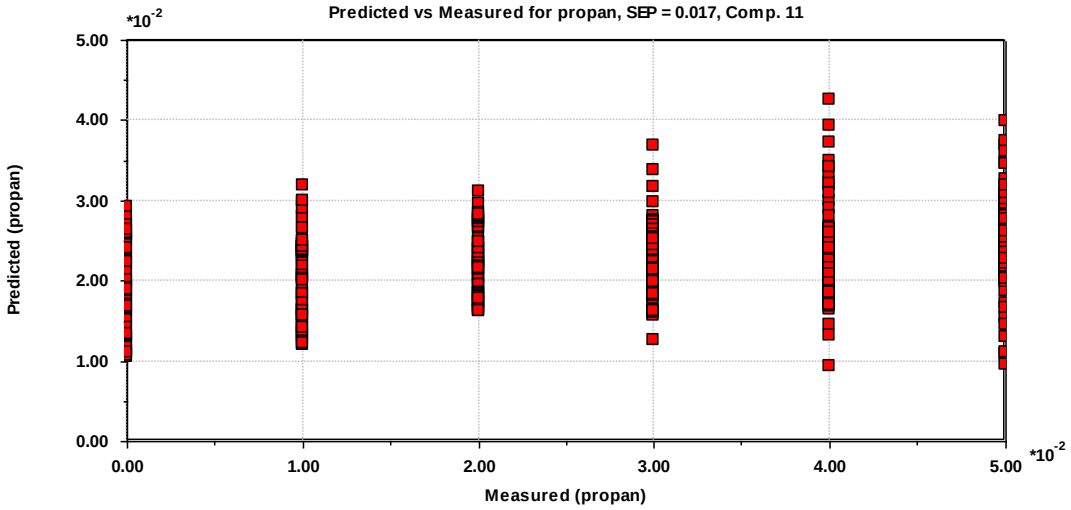
Kapittel 4.3 Oppbygging av kjemisk samansetjing ved revers prediksjon



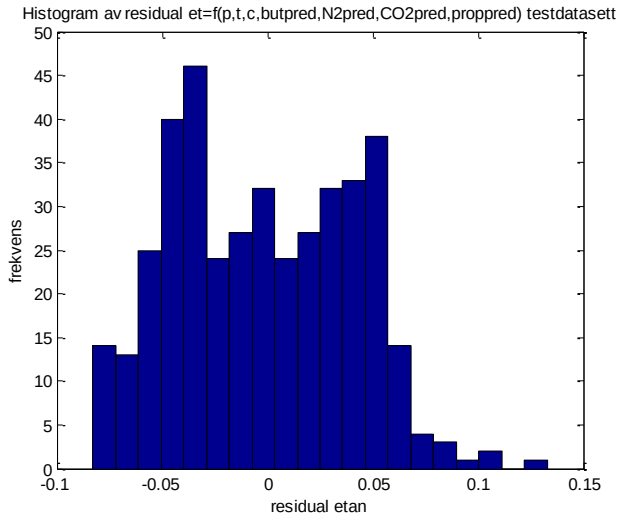
Figur A.3.1Histogram av residual for prediksjon av propan med modell med kun førstegradsledd



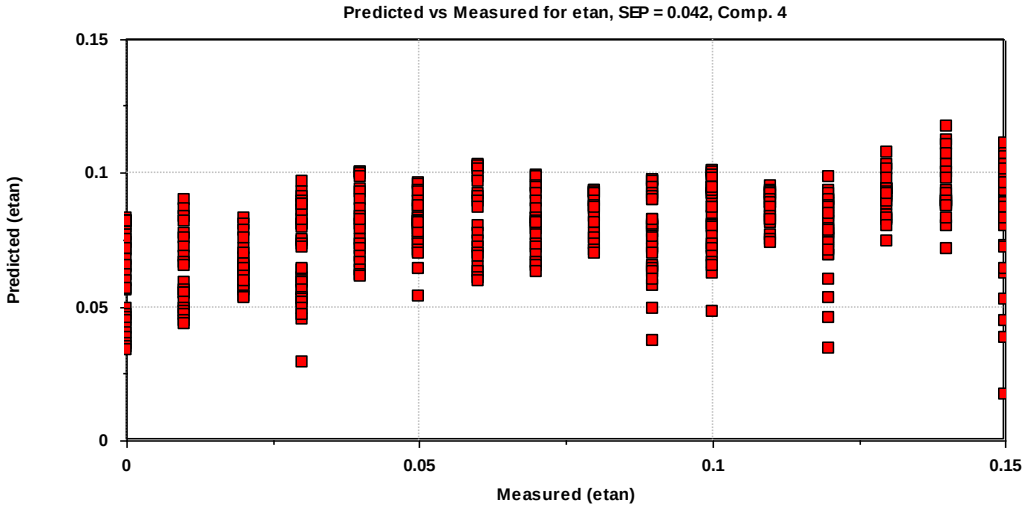
Figur A.3.2 Histogram av residual for predikert verdi for propan, første og andregradsledd



Figur A.3.3 Plott av predikert mot målt verdi for propan, modell med førstegradsledd og andregradsledd.



Figur A.3.4 Histogram av responsresiduala for etan, førstegradsmodell.



Figur A.3.5 Plott av predikert mot målt verdi for predikert etan, modell med berre førstegradsledd

TABELLAR OVER MODELLERING AV KJEMISK SAMANSETNING I SIRIUS:

Tabell A.4.1 forklart varians i y for modell og maksimumsfeil etter testing i Sirius og MATLAB for metan.

Modell	Total forkl. grad modell %	Maksfeil etter test i Sirius Nedre og øvre grense	Maksfeil etter AR.nedre og øvre grense	Kommentar
Met=f(lyd)	89.69	-0.036 0.01		
Met=f(lyd,lyd^2)	90.09	-0.037 0.049		
Met=f(lyd,et)	97.40	-0.025 -0.005	-0.07 -0.02	
Met=f(lyd,lyd^2,et,et^2,lyd*et)	97.79	-0.018 0.023		
Met=f(lyd,et,prop)	97.37	-0.028 -0.003	-0.16 -0.02	Alle objekt pred til 1.00 etter AR
Met=f(lyd,et,prop,CO_2)	97.62	-0.026 -0.001	-0.16 -0.02	Alle objekt pred 1.00 etter AR
Met=f(lyd,et,prop,CO_2,N_2)	99.95	-0.011 -0.03	0.12 0.26	Alle objekt pred til 0.72 etter AR
Met=f(lyd,et,prop,CO_2,N_2,but)	99.98	-0.005 -0.002	-0.16 -0.02	Alle objekt pred til 1.00 etter AR

Tabell A.4.2 forklart varians i y for modell og maksimumsfeil etter testing i Sirius og MATLAB for etan.

Modell	Total forkl. grad modell %	Maksfeil etter test i Sirius. nedre og øvre grense	Maksfeil etter AR. nedre og øvre grense	Kommentar
et=f(lyd)	61.67	-0.023 0.017		
Et=f(lyd,lyd^2)	61.78	-0.07 0.052		
et=f(lyd,met)	90.35	-0.035 -0.007	0 0.043	
Et=f(lyd,lyd^2,met,met^2, lyd*met)	90.94	-0.032 0.028		
et=f(lyd,met,prop)	92.62	-0.031 -0.005	0 0.07	Alle objekt pred til 0.00 etter AR
et=f(lyd,met,prop,CO_2)	93.79	-0.028 0.001	-0.15 -0.064	
et=f(lyd,met,prop,CO_2,N_2)	99.78	-0.016 -0.005	-0.15 -0.08	Alle objekt pred til 0.15 etter AR
et=f(lyd,met,prop,CO_2,N_2,but)	99.90	0.003 0.011	-0.15 -0.08	Alle objekt pred til 0.15 etter AR

Tabell A.4.3 forklart varians i y for modell og maksimumsfeil etter testing i Sirius og MATLAB for propan.

Modell	Total forkl. grad modell %	Maksimumsfeil etter test i Sirius,nedre og øvre grense	Maksimumsfeil etter AR,nedre og øvre grense	Kommentar
Prop=f(lyd)	33.16	-0.019 0.024		
Prop=f(lyd,met)	50.71	-0.011 0.029		
Prop=f(lyd,met,et)	62.30	-0.017 0.019	0 0.05	Alle objekt pred til 0.00 etter AR
Prop=f(lyd,met,et,CO_2)	70.97	-0.015 0.022	0 0.05	Alle objekt pred til 0.00 etter AR
Prop=f(lyd,met,et,CO_2,N_2)	94.58	-0.026 -0.007	-0.05 0	Alle objekt pred til 0.05 etter AR
Prop=f(lyd,met,et,CO_2,N_2,but)	98.90	0.005 0.014	0 0.05	Alle objekt pred til 0.00 etter AR

Tabell A.4.5 forklart varians i y for modell og maksimumsfeil etter testing i Sirius og MATLAB for CO_2.

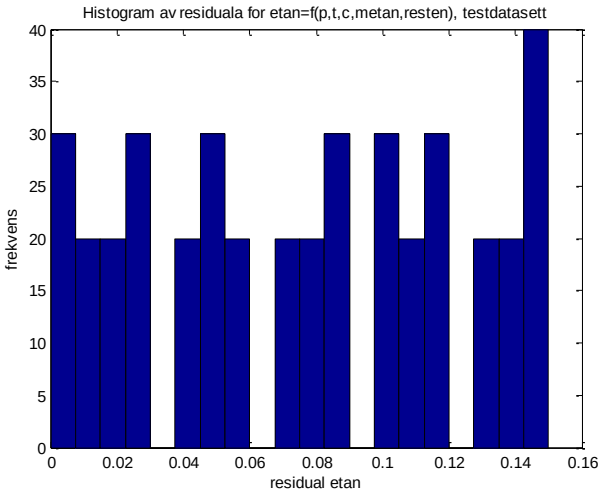
Modell	Total forkl. grad modell %	Maksimumsfeil etter test i Sirius. nedre og øvre grense	Maksimumsfeil etter AR. nedre og øvre grense	Kommentar
CO_2=f(lyd,met,et,prop)	46.54	-0.003 0.013	0.01 0.02	Alle objekt pred til 0.00 etter AR
CO_2=f(lyd,met,et,prop,N_2)	94.37	-0.018 -0.004	-0.02 -0.01	Alle objekt pred til 0.03 etter AR
CO_2=f(lyd,met,et,prop,N_2,but)	97.22	0.004 0.014	0.01 0.02	Alle objekt pred til 0.00 etter AR

Tabell A.4.6 forklart varians i y for modell og maksimumsfeil etter testing i Sirius og MATLAB for N_2.

Modell	Total forkl. grad modell %	Maksimumsfeil etter test i Sirius, nedre og øvre grense	Maksimumsfeil etter AR, nedre og øvre grense	Kommentar
N_2=f(lyd,met,et,prop,CO_2)	98.09	-0.012 -0.003	0 0.03	Alle objekt pred til 0.00 etter AR
N_2=f(lyd,met,et,prop,CO_2,but)	99.44	-0.006 -0.002	0 0.03	Alle objekt pred til 0.00 etter AR

Tabell A.4.7 forklart varians i y for modell og maksimumsfeil etter testing i Sirius og MATLAB for butan.

Modell	Total forkl. grad modell %	Maksimumsfeil etter test i Sirius. nedre og øvre grense	Maksimumsfeil etter AR, nedre og øvre grense	Kommentar
but=f(lyd,met,et,prop,CO_2,N_2)	96.95	0.01 0.003	-0.02 0	Alle objekt pred til 0.02 etter AR



Figur A.5.1 Histogram av residuala for etan etter AR.

MATLABkode Iterativ konsentrasjonsbestemming ved alternerande regresjon

HOVED

```
function
[y_alle,but_sk_alle,met_sk_alle,et_sk_alle,propan_sk_alle,CO2_sk_alle,N2_sk
_alle,but_pred_alle,CO2_pred_alle,N2_pred_alle,prop_pred_alle,et_pred_alle,
metan_pred_alle,sk_siste_met]=hovedkjemi(X,B)

%hovedfunksjonen skal styre input til iterasjonen og kalle
%opp pred_runde så mange ganger som nødvendig før
%konvergens.X er her heile vektoren som
%gjeld for alle dei kjemiske stoffa. "Utplukkinga" skjer i
%pred_runde. Må også sende i regresjonskoeffisientane b, desse
%er spesifikke for kvar modell og må sendast inn for kvart stoff.
%Denne er forma som ei matrise, der elementa ikkje er med er
%regresjonskoeffisienten sett til 0.

X=xlsread(X);
B=xlsread(B);

xlsread Xbut.xls;
Xbut=ans;
Xcalbut=Xbut(1:3999,:);
middbut=mean(Xcalbut);
standbut=std(Xcalbut);
Xcalstdbut=(Xcalbut-ones(3999,1)*middbut)./(ones(3999,1)*standbut);

Xblokkbut=[Xcalstdbut(:,1:12) Xcalstdbut(:,14:17)];
ybut=Xcalstdbut(:,13);
A=13;

[reg,ssq,xlds,ylds,wts,xscrs,yscrs,bin] = nippls(Xblokkbut,ybut,A);
wtsbut=wts;
xldsbut=xlds;

xlsread XCO2.xls;
XCO2=ans;
XcalCO2=XCO2(1:3999,:);
middCO2=mean(XcalCO2);
standCO2=std(XcalCO2);
XcalstdCO2=(XcalCO2-ones(3999,1)*middCO2)./(ones(3999,1)*standCO2);

XblokkCO2=[XcalstdCO2(:,1:7) XcalstdCO2(:,9)];
yCO2=XcalstdCO2(:,8);
A=6;

[reg,ssq,xlds,ylds,wts,xscrs,yscrs,bin] = nippls(XblokkCO2,yCO2,A);
wtsCO2=wts;
xldsCO2=xlds;

xlsread XN2.xls;
XN2=ans;
XcalN2=XN2(1:3999,:);
middN2=mean(XcalN2);
standN2=std(XcalN2);
XcalstdN2=(XcalN2-ones(3999,1)*middN2)./(ones(3999,1)*standN2);
```

```
XblokkN2=[XcalstdN2(:,1:8)];
yN2=XcalstdN2(:,9);
A=6;

[reg,ssq,xlds,ylds,wts,xscrs,yscrs,bin] = nippls(XblokkN2,yN2,A);
wtsN2=wts;
xldsN2=xlds;

xlsread Xpropan.xls;
Xpropan=ans;
Xcalpropan=Xpropan(1:3999,:);
middpropan=mean(Xcalpropan);
standpropan=std(Xcalpropan);
Xcalstdpropan=(Xcalpropan-
ones(3999,1)*middpropan)./(ones(3999,1)*standpropan);

Xblokkpropan=[Xcalstdpropan(:,1:5) Xcalstdpropan(:,7:37)];
ypropan=Xcalstdpropan(:,6);
A=13;

[reg,ssq,xlds,ylds,wts,xscrs,yscrs,bin] = nippls(Xblokkpropan,ypropan,A);
wtspropan=wts;
xldspropan=xlds;

xlsread Xmetan.xls;
Xmet=ans;
Xcal=Xmet(1:3999,:);
middmet=mean(Xcal);
standmet=std(Xcal);

Xcalstd=(Xcal-ones(3999,1)*middmet)./(ones(3999,1)*standmet);
A=4;

Xblokkmet=[Xcalstd(:,1:3) Xcalstd(:,5:9)];
ymet=Xcalstd(:,4);

[reg,ssq,xlds,ylds,wts,xscrs,yscrs,bin] = nippls(Xblokkmet,ymet,A);
wtsmet=wts;
xldsmet=xlds;

xlsread Xetan.xls;
Xet=ans;
Xcal=Xet(1:3999,:);
Xpre=Xet(4000:4399,:);
middet=mean(Xcal);
standet=std(Xcal);

Xcalstd=(Xcal-ones(3999,1)*middet)./(ones(3999,1)*standet);
A=10;

Xblokket=[Xcalstd(:,1:4) Xcalstd(:,6:37)];
yet=Xcalstd(:,5);

[reg,ssq,xlds,ylds,wts,xscrs,yscrs,bin] = nippls(Xblokket,yet,A);
wtset=wts;
xldset=xlds;
```

```
s=struct('met',{{middmet, standmet, wtsmet, xldsmet}},'but',{{middbut,
standbut,wtsbut,xldsbut}},'et',{{middet,
standet,wtset,xldset}},'propan',{{middpropan,standpropan,wtspropan,xldsprop
an}},'CO2',{{middCO2,standCO2,wtsCO2,xldsCO2}},'N2',{{middN2,standN2,wtsN2,
xldsN2}});

maxiterasjon = 1000;
terskel = 0.001;
ferdig = 0;
teller = 1;
tellerbutan=0;
objects=1;
N=400;

but_pred_alle=zeros(N,maxiterasjon);
metan_pred_alle=zeros(N,maxiterasjon);
et_pred_alle=zeros(N,maxiterasjon);
propan_pred_alle=zeros(N,maxiterasjon);
CO2_pred_alle=zeros(N,maxiterasjon);
N2_pred_alle=zeros(N,maxiterasjon);
but_sk_alle=zeros(N,maxiterasjon);
met_sk_alle=zeros(N,maxiterasjon);
et_sk_alle=zeros(N,maxiterasjon);
propan_sk_alle=zeros(N,maxiterasjon);
CO2_sk_alle=zeros(N,maxiterasjon);
N2_sk_alle=zeros(N,maxiterasjon);
sk_siste_met=zeros(N,1);
sk_siste_but=zeros(N,1);

%går så inn i ei løkke der eg har definert terskelverdien.
%Kallar opp pred_runde_kjemi og reknar ut differansen mellom oppdatert
%og gammal vektor. Dersom denne differansen er mindre enn
%terskelverdien er ein ferdig. Dersom den er større blir gammal
%vektor erstatta med ny og ein køyrer ein runde til. Dette skjer
%heilt til ein oppnår terskelverdi.


for i = 1:N
    x=X(i,:);

while (~ ferdig) & (teller < maxiterasjon)


[y,rsd_objectbut,rsd_objectmet,rsd_objectet,rsd_objectpropan,rsd_objectCO2,
rsd_objectN2,but_predikert,met_predikert,et_predikert,prop_predikert,CO2_pr
edikert,N2_predikert]=pred_runde_kjemi(x,B,s);

    if teller>1
        but_sk = [but_sk,rsd_objectbut];
    else
        but_sk = rsd_objectbut;
    end

    if teller>1
        but_pred = [but_pred, but_predikert];
    else
        but_pred = but_predikert;
    end
```

```
 if teller>1
     met_sk = [met_sk,rsd_objectmet];
 else
     met_sk = rsd_objectmet;
 end

 if teller>1
     metan_pred = [metan_pred, met_predikert];
 else
     metan_pred = met_predikert;
 end

 if teller > 1
     et_pred = [et_pred, et_predikert];
 else
     et_pred = et_predikert;
 end

 if teller > 1
     et_sk = [et_sk, rsd_objectet];
 else
     et_sk = rsd_objectet;
 end

 if teller > 1
     prop_pred = [prop_pred, prop_predikert];
 else
     prop_pred = prop_predikert;
 end

if teller > 1
    propan_sk = [propan_sk, rsd_objectpropan];
else
    propan_sk = rsd_objectpropan;
end

if teller > 1
    CO2_pred = [CO2_pred, CO2_predikert];
else
    CO2_pred = CO2_predikert;
end

if teller > 1
    CO2_sk=[CO2_sk, rsd_objectCO2];
else
    CO2_sk=rsd_objectCO2;
end

if teller > 1
    N2_pred = [N2_pred, N2_predikert];
else
    N2_pred = N2_predikert;
end

if teller > 1
    N2_sk = [N2_sk, rsd_objectN2];
else
    N2_sk = rsd_objectN2;
end
```

```
    differanse = abs(x-y);
    if max(differanse) < terskel;
        ferdig = 1;
    else x = y;

    end
    teller = teller + 1;

end

ferdig=0;
teller=1;
objects =objects + 1

lengde = length(but_pred);
but_pred_alle(i,1:lengde) = but_pred;
metan_pred_alle(i,1:lengde)=metan_pred;
et_pred_alle(i,1:lengde)=et_pred;
prop_pred_alle(i,1:lengde)=prop_pred;
CO2_pred_alle(i,1:lengde)=CO2_pred;
N2_pred_alle(i,1:lengde)=N2_pred;

but_sk_alle(i,1:lengde)=but_sk;
met_sk_alle(i,1:lengde)=met_sk;


y_alle(i,:)=y;

index=find(met_sk_alle(i,:) ~=0);
lengdesk=length(index);
sk_siste_met(i)=met_sk(lengdesk);

index=find(but_sk_alle(i,:) ~=0);
lengdesk=length(index);
sk_siste_but(i)=but_sk(lengdesk);

end
teller
```

PRED RUNDE KJEMI

```
function [y, but_sk, met_sk, et_sk, propan_sk, CO2_sk, N2_sk, but_pred,
metan_pred, et_pred, prop_pred, CO2_pred, N2_pred]=pred_runde_kjemi(x,B,s)
%y - ny predikert variabel etter ein runde med prediksjon. Til slutt den
%endelege predikerte verdien for akutelt stoff.
%but_pos indikerer posisjonane til butan i X. b_but er
regresjonskoeffisientane i modellen for butan.
%Tilsvarande for dei andre stoffa. Først skal koden predikere butan på
%bakgrunn av T,p,c og gjennomsnittsverdiane for N2,CO2,propan, etan og
%metan. Deretter skal den predikerte verdien for butan nyttast saman med
%T,p,c og gjennomsnittsverdien til CO2, propan, etan og metan til å
%predikere N2 osv til alle er predikert. Dette skal så gjentas til
%(forhåpentlegvis)konvergens. Prediksjonen skal skje ved å bruke
%regresjonskoeffisientane frå modellane som er laga i sirius.

%tabell for butan
```

```
posisjon_but =[14 15 25 32 38 43 47 50 53 54]';
flagg=[0 2 1 1 1 1 1 1 1 1]';
veksel_pos=[0 0 2 4 6 8 10 12 16 18]';
pos_tabell_but=[posisjon_but flagg veksel_pos];

%tabell for N2
posisjon_N2=[18 19 27 34 40 45 49 52 54 55]';
flagg=[0 2 1 1 1 1 1 1 1 1]';
veksel_pos=[0 0 2 4 6 8 10 12 14 16]';
pos_tabell_N2=[posisjon_N2 flagg veksel_pos];

%tabell for CO2:
posisjon_CO2=[16 17 26 33 39 44 48 51 53 55]';
flagg=[0 2 1 1 1 1 1 1 1 1]';
veksel_pos=[0 0 2 4 6 8 10 12 14 18]';
pos_tabell_CO2=[posisjon_CO2 flagg veksel_pos];

%tabell for propan
posisjon_prop=[12 13 24 31 37 42 46 50 51 52]';
flagg=[0 2 1 1 1 1 1 1 1 1]';
veksel_pos=[0 0 2 4 6 8 10 14 16 18]';
pos_tabell_prop=[posisjon_prop flagg veksel_pos];

%tabell for etan
posisjon_etan=[10 11 23 30 36 41 46 47 48 49]';
flagg=[0 2 1 1 1 1 1 1 1 1]';
veksel_pos=[0 0 2 4 6 8 12 14 16 18]';
pos_tabell_etan=[posisjon_etan flagg veksel_pos];

%tabell for metan
posisjon_metan=[8 9 22 29 35 41 42 43 44 45]';
flagg=[0 2 1 1 1 1 1 1 1 1]';
veksel_pos=[0 0 2 4 6 10 12 14 16 18]';
pos_tabell_metan=[posisjon_metan flagg veksel_pos];

%For metan:
Xmetan=x;
y=x;

Xmetan(posisjon_metan)=[];
b_metan=B(6,1:55);
b_metan(posisjon_metan)=[];

metan_pred=pred(Xmetan,b_metan);

metan_pred;
if metan_pred > 1.00;
    metan_pred = 1.00;
elseif metan_pred < 0.72;
    metan_pred = 0.72;
end
y(posisjon_metan(1,1))=metan_pred;
%så må vi oppdatere y - de gamle metanverdiane skal erstattes
y = oppdater(y,pos_tabell_metan);

metx_sk=[y(:,2),y(:,4),
y(:,6),y(:,8),y(:,10),y(:,12),y(:,14),y(:,16),y(:,18)];
metx_sk_stand=(metx_sk-s.met{1})./(s.met{2});
metx_sk_stand_tilrsd=[metx_sk_stand(:,1:3),metx_sk_stand(:,5:9)];
```

```
%skal så rekne ut sk for det predikerte objektet
met_sk=rsd(metx_sk_stand_tilrsd,s.met{3},s.met{4});

%For etan
Xetan=y;
Xetan(posisjon_etan)=[];
b_etan=B(5,1:55);
b_etan(posisjon_etan)=[];

%nå kalles pred opp for etan
et_pred=pred(Xetan,b_etan);

%er nå etan innanfor området?
et_pred;
if et_pred > 0.15;
    et_pred = 0.15;
elseif et_pred < 0;
    et_pred = 0;
end
y(posisjon_etan(1,1))=et_pred;
%så må vi oppdatere y - de gamle etanverdiane skal erstattes
y = oppdater(y,pos_tabell_etan);

etx_sk=[y(:,2),y(:,4),
y(:,6),y(:,8),y(:,10),y(:,12),y(:,14),y(:,16),y(:,18),y(:,20:22),y(:,24:29)
,y(:,31:35),y(:,37:40),y(:,42:45),y(:,50:55)];
etx_sk_stand=(etx_sk-s.et{1})./(s.et{2});
etx_sk_stand_tilrsd=[etx_sk_stand(:,1:4),etx_sk_stand(:,6:37)];

%skal så rekne ut sk for det predikerte objektet
et_sk=rsd(etx_sk_stand_tilrsd,s.et{3},s.et{4});

%For propan:
Xprop=y;
Xprop(posisjon_prop)=[];
b_prop=B(4,1:55);
b_prop(posisjon_prop)=[];

%nå kalles pred opp for propan
prop_pred=pred(Xprop,b_prop);

%er nå propan innanfor området?
if prop_pred > 0.05;
    prop_pred = 0.05;
elseif prop_pred < 0;
    prop_pred = 0;
end
y(posisjon_prop(1,1))=prop_pred;
%så må vi oppdatere y - de gamle propanverdiane skal erstattes
y = oppdater(y,pos_tabell_prop);

propanx_sk=[y(:,2),y(:,4),y(:,6),y(:,8),y(:,10),y(:,12),y(:,14),y(:,16),y(:
,18),y(:,20:23),y(:,25:30),y(:,32:36),y(:,38:41),y(:,43:45),y(:,47:49),y(:,
53:55)];
propanx_sk_stand=(propanx_sk-s.propan{1})./(s.propan{2});
propanx_sk_stand_tilrsd=[propanx_sk_stand(:,1:5),propanx_sk_stand(:,7:37)];

%skal så rekne ut sk for det predikerte objektet
```

```
propan_sk=rsd(propanx_sk_stand_tilrsd,s.propan{3},s.propan{4});

%For CO2:
XCO2=y;
XCO2(posisjon_CO2)=[];
b_CO2=B(3,1:55);
b_CO2(posisjon_CO2)=[];

%nå kalles pred opp for CO2
CO2_pred=pred(XCO2,b_CO2);

CO2_pred;
if CO2_pred > 0.03;
    CO2_pred = 0.03;
elseif CO2_pred < 0;
    CO2_pred = 0;
end
y(posisjon_CO2(1,1))=CO2_pred;
%så må vi oppdatere y - de gamle CO2verdiane skal erstattes
y = oppdater(y,pos_tabell_CO2);

CO2x_sk=[y(:,2),y(:,4),y(:,6),y(:,8),y(:,10),y(:,12),y(:,14),y(:,16),y(:,18
)];
CO2x_sk_stand=(CO2x_sk-s.CO2{1})./(s.CO2{2});
CO2x_sk_stand_tilrsd=[CO2x_sk_stand(:,1:7),CO2x_sk_stand(:,9)];

%skal så rekne ut sk for det predikerte objektet
CO2_sk=rsd(CO2x_sk_stand_tilrsd,s.CO2{3},s.CO2{4});

%For N2:
XN2=y;
XN2(posisjon_N2)=[];
b_N2=B(2,1:55);
b_N2(posisjon_N2)=[];
%nå kalles pred opp for nitrogen
N2_pred = pred(XN2,b_N2);

N2_pred;
if N2_pred > 0.03;
    N2_pred = 0.03;
elseif N2_pred < 0;
    N2_pred = 0;
end
y(posisjon_N2(1,1))=N2_pred;
%så må vi oppdatere y - de gamle N2verdiane skal erstattes
y = oppdater(y,pos_tabell_N2);

N2x_sk=[y(:,2),y(:,4),y(:,6),y(:,8),y(:,10),y(:,12),y(:,14),y(:,16),y(:,18)
];
N2x_sk_stand=(N2x_sk-s.N2{1})./(s.N2{2});
N2x_sk_stand_tilrsd=[N2x_sk_stand(:,1:8)];

%skal så rekne ut sk for det predikerte objektet
N2_sk=rsd(N2x_sk_stand_tilrsd,s.N2{3},s.N2{4});


%For butan:
Xbutan=y;
```

```
Xbutan(posisjon_but)=[];
b_but=B(1,1:55);
b_but(posisjon_but)=[];

%nå kalles pred opp for butan
but_pred = pred(Xbutan,b_but);

but_pred;
if but_pred > 0.02
    but_pred = 0.02;
elseif but_pred < 0
    but_pred = 0;
end
y(posisjon_but(1,1))=but_pred;
%så må vi oppdatere x - de gamle butanverdiene skal erstattes
y = oppdater(y,pos_tabell_but);

butx_sk=[y(:,2:14),y(:,16:19)];
butx_sk_stand=(butx_sk-s.but{1})./(s.but{2});
butx_sk_stand_tilrsd=[butx_sk_stand(:,1:12),butx_sk_stand(:,14:17)];

%skal så rekne ut sk for det predikerte objektet
but_sk=rsd(butx_sk_stand_tilrsd,s.but{3},s.but{4});

end
```

PRED

```
function yp=pred(x,b)

%output i denne funksjonen er predikert verdi for y. Koden skal bruke dei
%predikerte verdiane frå pred_runde og rekne ut predikert verdi for kvar av
dei kjemiske stoffa.
%husk at b har konstantledd b0, x0 må difor alltid vere 1

q=x.*b;
yp=sum(q);
```

OPPDATER

```
function y = oppdater(x,pos_tabell);
%denne fuksjonen skal oppdatere tabellen som inneheld alle variablane inkl
%vekselverknadsledd og andregradsledd. Desse treng ein vidare i prediksjon
%av dei andre stoffa. Oppdateringa skjer etter at den endelege prediksjonen
%for det aktuelle stoffet er foretatt.
%pos_tabell=(posisjonsvektor for aktuelt stoff,flagg,kode for
%vekselvirkningsledd)

n = size(pos_tabell,1); %lengden av første kollonne i posisjonstabellen

for i = 2:n;
    if (pos_tabell(i,2) == 1);
        %fiks vekselvirkningen
        %då skal den tilhøyrande posisjonen i x få ein ny verdi
        x(pos_tabell(i,1))=x((pos_tabell(1,1)))*x((pos_tabell(i,3)));
    elseif (pos_tabell(i,2) == 2);
        %fiks andregradsledd
```

```
        x(pos_tabell(i,1))=x(pos_tabell(1,1))^2;
    else
        disp('FEIL!')
    end
end
y=x;
```

RSD

```
function sk=rsd(x,w,p)
%function sk=rsd(x,w,M,A) skal rekne ut
%x-residuala for objekta slik at ein kan
%samanlikne desse residuala med rejection
%criteria som sirius reknar ut.
%x er objektet som er predikert
%w er vektene, desse importeres frå sirius
%p er ladningane
%M er antall variabler
%A er antall komponentar i modellen


M=size(w,1);
A=size(w,2);
tk=x*w; %1xA
ek=x-(tk*p');
%ek=ek';
sk2=(ek*ek')/(M-A);
sk=sqrt(sk2);
```

Samanfatning for modellar for kjemiske komponentar

Modell	Responsvariabel	Komp	Forkl.grad %	X			y			Normford Respres		TESTA
				1.gr.ledd	2.gr.ledd	Vekselvirkn	1.gr.ledd	Kvadratrot	ln	JA	NEI	
M1	Metan	4	99,97	x			x			X		x
M2	Metan	5	100,00	x			x				X	
M3	Metan	9	99,82	x		x	x			x		
E1	Etan	4	99,79	x			x				X	
E2	Etan	4	98,97	x	x		x				X	
E3	Etan √	4	94,57	x	x			x			X	
E4	Etan ln		29	x					x		x	
E5	Etan	10	99,95	x		x	x			x		x
P1	Propan	6	99,98	x			x				x	
P2	Propan	13	99,62	x		x	x			x		x
B1	Butan	7	100,00	x			x				x	
B2	Butan	10	99,14	x	x		x			x		
B3	Butan	14	99,11	x		x	x			x		x
B4	Butan	13	99,44	x	x		x			x		x
CO1	CO_2	6	99,85	x			x			x		x
CO2	CO_2	15	99,46	x		x	x			x		
N1	N_2	6	99,82	x			x			x		x
N2	N_2	13	99,37	x		x	x			x		

Samanfatning for modellar for brennverdi (B)og tettleik (T)

Modell	Forkl. Grad %	Komp	X					y				Norm.ford Resp.res	
			1.gr. ledd	2.gr. ledd			Veksel virkning	1.gr. ledd	2.gr. ledd	Kvadrat rot	Log	JA	NEI
T1	94,97	2	x					x				x	
T2	96,17	7	x	x				x				x	
T3	94,86	4	x							x		x	
T4	98,61	10	x	x						x		x	
T5	90,07	4	x								x		x
T6	99,66	14	x	x			x			x		x	
T7	99,26	10	x	x			x	x				x	
B1	99,63	3	x					x				x	
B2	99,94	4	x					x					x
B3	98,49	3	x	x				x				x	
B4	99,93	4	x							x			x
B5	99,99	6	x							x			x
B6	99,87	4	x								x		x
B7	99,95	6	x								x		x
B8	99,96	8	x	x				x					x
B9	99,98	5	x					x				Tja	
B10	99,92	10	x				x	x				x	
B11	99,96	10	x	x			x	x				x	