

DATA FRA JERNBANEDRIFT
(Railway Operation Data)
Masteroppgave
Høst 2005

Stud. Techn. Øystein Luktvaslimo
Institutt for Produksjons- og
kvalitetsteknikk, NTNU

Innlevert 10.1.2006

NTNU
Norges teknisk-naturvitenskapelige
universitet
Fakultet for ingeniørvitenskap og teknologi
Linjen for produktutvikling og produksjon
erklæring

MASTEROPPGAVE

Høstsemesteret 2005

for

Stud.techn. NYSTEIN LUKTJASSIMO.....

Opgavetittel: DATA FRA JERNBANEDRIFT
RAILWAY OPERATIONS DATA
.....

ERKLÆRING

Jeg erklærer herved på ære og samvittighet at jeg har utført ovennevnte masteroppgave selv og uten noen som helst ulovlig hjelp.

Tordheim..... 19.11.2006

Arne Luktjassimo

De innleverte besvarelser med bilag blir i henhold til reglement for sivilarkitekt- og sivilingeniørstudiets § 3.5.5 universitetets eiendom, og kan av universitetet fritt benyttes til undervisnings- og forskningsformål. Arbeidene kan ikke nyttes til andre formål, f.eks. økonomiske, uten etter avtale mellom universitetet og vedkommende student.



MASTEROPPGAVE
Høsten 2005
for
stud. techn. Øystein Luktvaslimo

DATA FRA JERNBANEDRIFT
(Railway Operation Data)

Oppgaven rettes mot oppfølging av jernbanedrift. Det tas utgangspunkt i analyse og sammenligning av data fra ulike kilder. Fokus rettes mot oppfølging av punktlighet og relaterte faktorer, som kan være kjøretid, stasjonsopphold og egenskaper ved det rullende materiellet.

Oppgaven utføres i samarbeid med NSB og forskningsprosjektet PEMRO.

I oppgaven skal kandidaten mer spesifikt:

1. Gjennomføre et litteraturstudium relatert til datakilder som skal brukes som grunnlag for faktabaserte beslutninger. Et sammendrag av dette skal presenteres.
2. Beskrive/belyse toggangen for en utvalgt strekning ved hjelp av ulike datakilder over en viss tidsperiode.
3. Analysere de ulike datasettene og presentere resultatene for stasjonsopphold på noen stasjoner med spesiell fokus på å vurdere ulike forklaringsfaktorer.
4. Basert på de foregående punktene, beskrive styrker og svakheter, likheter og ulikheter ved de ulike datasettene og analyseformene.

Oppgaveløsningen skal basere seg på eventuelle standarder og praktiske retningslinjer som foreligger og anbefales. Dette skal skje i nært samarbeid med veiledere og fagansvarlig. For øvrig skal det være et aktivt samspill med veiledere.

Innen tre uker etter at oppgaveteksten er utlevert, skal det leveres en forstudierapport som skal inneholde følgende:

- En analyse av oppgavens problemstillinger.
- En beskrivelse av de arbeidsoppgaver som må gjennomføres for løsning av oppgaven. Denne beskrivelsen skal munne ut i en klar definisjon av arbeidsoppgavenes innhold og omfang.
- En tidsplan for fremdriften av prosjektet. Planen skal utformes som et Gantt-skjema med angivelse av de enkelte arbeidsoppgavenes terminer, samt med angivelse av milepæler i arbeidet.

Forstudierapporten er en del av oppgavebesvarelsen og skal innarbeides i denne. Det samme skal senere fremdrifts- og avviksrapporter. Ved bedømmelsen av arbeidet legges det vekt på at gjennomføringen er godt dokumentert.

Besvarelsen redigeres mest mulig som en forskningsrapport med et sammendrag både på norsk og engelsk, konklusjon, litteraturliste, innholdsfortegnelse etc. Ved utarbeidelsen av teksten skal kandidaten legge vekt på å gjøre teksten oversiktlig og velskrevet. Med henblikk på lesning av besvarelsen er det viktig at de nødvendige henvisninger for korresponderende steder i tekst, tabeller og figurer anføres på begge steder. Ved bedømmelsen legges det stor vekt på at resultatene er grundig bearbeidet, at de oppstilles tabellarisk og/eller grafisk på en oversiktlig måte og diskuteres utførlig.

Materiell som er utviklet i forbindelse med oppgaven, så som programvare eller fysisk utstyr er en del av besvarelsen. Dokumentasjon for korrekt bruk av dette skal så langt som mulig også vedlegges besvarelsen.

Eventuelle reiseutgifter, kopierings- og telefonutgifter må bæres av studenten selv med mindre andre avtaler foreligger.

Hvis kandidaten under arbeidet med oppgaven støter på vanskeligheter, som ikke var forutsett ved oppgavens utforming, og som eventuelt vil kunne kreve endringer i eller utelatelse av enkelte spørsmål fra oppgaven, skal dette straks tas opp med instituttet.

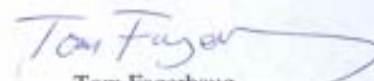
Besvarelsen skal innleveres i 3 eksemplar (innbundne) til instituttet. I tillegg leveres minst et innbundet eksemplar og en CD til henholdsvis veileder og NSB.

Ansvarlig faglærer: Førsteamanuensis Tom Fagerhaug
E-post: tom.fagerhaug@ntnu.no
Tlf.: 73 59 71 24
Mobiltlf.: 909 86 854

Veileder ved SINTEF,
Produktivitet og
prosjektledelse: Nils Olsson
E-post: nils.olsson@sintef.no
Mobiltlf.: 977 13 628

**INSTITUTT FOR PRODUKSJONS-
OG KVALITETSTEKNIKK**


Asbjørn Rolstadås
professor/instituttleder


Tom Fagerhaug
faglærer

**FAKULTET FOR
INGENIØRVITENSKAP
OG TEKNOLOGI
MASTEROPPGAVEN**

Utlevert : 16.8.2005
Innleveres senest : 10.1.2006

FORORD

Under arbeidet med denne oppgaven har jeg nok en gang fått uvurderlig tilbakemelding og støtte fra min veileder Nils Olsson ved SINTEF. I tillegg vil jeg takke Hans Haugland og Abdulrahim Alkadi i NSB Drift og Arne Hovland i Jernbaneverket for all hjelp og innspill.

Takk også til Anne for korrekturlesning samt hele Mellomvn12B for alltid å holde døra åpen.

Trondheim, 10. januar 2006.

Øystein Luktvasllimo

SAMMENDRAG

Denne oppgaven er resultatet av den avsluttende masteroppgaven ved sivilingeniørutdanningen ved NTNU. Oppgaven er skrevet og levert ved Institutt for Produksjons- og Kvalitetsteknikk i samarbeid med forskningsprosjektet PEMRO, som er et forskningsprosjekt som fokuserer på bruk og oppfølging av prestasjonsindikatorer i jernbanedrift.

Oppgaven omhandler data i jernbanedrift; mer spesifikt å diskutere kvaliteten av to datakilder AnnaLyse og TELOC. Diskusjonsgrunnlaget for datakvaliteten er en litteraturstudie som munner ut i definisjonen av noen ulike måleparametere for datakvaliteten og en analyse av strekningen Drammen – Eidsvoll over en 15 dagers periode høsten 2005. Oppgaven er skrevet med utgangspunkt i NSB Drifts situasjon.

Konklusjonen som trekkes på grunnlag av litteraturstudiet er at det kan hjelpe struktureringen av produksjonen av kvalitetsdata å adoptere en prosessanalogi. Denne analogien hjelper oss til å dele produksjonen i 3 roller: datainnsamler, databasebestyrer og dataforbruker, eller som 3 underprosesser: innsamling av data, prosessering av data og forbruk av data. Videre er det essensielt at den programmerte datakvaliteten i høyest mulig grad imøtekommer den oppfattede kvaliteten. Dette kan oppnås ved å kartlegge bruksområdet for dataene best mulig. For NSB Drifts del bør dataforbrukerne kravspesifisere sitt bruksområde i form av ønskede metrikker og karakteristikker ved dataproduktet. Disse spesifikasjonene kan så bringes videre bakover i prosesskjeden slik at man tilpasser målingene til dataforbrukerens behov. En slik forbedret innfrielse av dataforbrukerens behov vil være med på å styrke den subjektive dimensjonen av datakvaliteten. Eventuelle endringer i dataproduksjonsprosessen bør gjøres så tidlig så mulig av hensyn til ressursbruk og fleksibilitet.

Den objektive dimensjonen av datakvaliteten kan styrkes ved å gjøre målinger i henhold til de 6 foreslåtte parametrene: antall observasjoner, oppløsning, nøyaktighet, fullstendighet, ensartethet og tidsriktighet. Data som brukes i NSB drift bør ledsages av verdier for disse parametrene for å kunne si noe om påliteligheten av de. Dette kan være med på å hjelpe en beslutningstaker å forstå gyldigheten av sin beslutning. Men å styrke den objektive datakvaliteten vil ikke nødvendigvis være hensiktsmessig i alle tilfeller; en slik økning må balanseres med behovet til forbrukeren for å sikre seg mot unødvendig ressursforbruk.

Data fra AnnaLyse systemet og TELOC kan brukes til å beregne gjennomsnittlige verdier for oppholdstiden, kjøretiden og ankomstforsinkelsen. Datakildene er derimot uegnet slik de framstår i dag, til å gjøre en direkte sammenligning av korresponderende verdier. Dette skyldes at man som dataforbruker ikke sitter med nok kunnskap om måleprosessen, og dermed om den relative kalibreringen av måleutstyret.

Analysen av strekningen Drammen – Eidsvoll tyder på at oppholdstiden fra TELOC generelt viser en tendens til å ligge over det planlagte. Ankomstforsinkelsen ser ut til å avta mot slutten av pendelen. Dette kan i viss grad sees i sammenheng med at den faktiske kjøretiden er mindre enn den planlagte mot slutten, noe som kan antyde at man kjører inn forsinkelser mot slutten. Oppholdstiden fra AnnaLyse består ikke nødvendigvis bare av komponenter som skyldes forholdet ved stasjonene. Dette tyder det store avviket mellom fra oppholdstiden fra TELOC og AnnaLyse på.

Oppholdstiden ved en stasjon består av mange komponenter og påvirkes igjen av mange faktorer. I denne oppgaven er ikke det tilgjengelige datamaterialet; verdier for oppholdstiden fra TELOC og AnnaLyse, passasjerantall og enkelt/dobbeltsett; kvalitetsmessig bra nok for å kunne konstruere regresjonsmodeller for å undersøke dette forholdet. Da det tilgjengelige datamaterialet ikke var nok til å belyse dette forholdet anbefales det at man søker nye data hvis man ønsker å belyse dette forholdet.

Både AnnaLyse og TELOC kommer relativt bra ut i forhold til de foreslåtte parametrene for datakvaliteten, men det ligger en utfordring i å tilpasse dataproduktet til brukeren. For TELOC kan man vurdere om det er mulig å hente andre typer data fra samme kilde. TELOC produserer i dag ikke data med tanke på en NSB Drift ansatts behov. Ved å samarbeide tettere med datainnsamlere om måleprosessen kunne man tilpasset målingene bedre til alle parters behov.

AnnaLyse trenger å videreutvikles som system med tanke på bruksområdet. Med det menes det at forbedringsinnsatsen blir fokusert på det siste leddet i dataproduksjonsprosessen. Systemet henter data fra en rikholdig database; utfordringen ligger i å levere et dataprodukt bedre tilpasset forbrukeren.

SUMMARY

This paper is a dissertation at the Department of Production and Quality Engineering at the Norwegian University of Science and Technology (NTNU). The paper is written in collaboration with PEMRO, a research programme focusing on the use and monitoring of performance indicators in railway operation.

The paper deals with data in railway operation; more specifically discussing the quality of two different data sources namely: AnnaLyse and TELOC. The basis of this discussion is a review of relevant literature providing different parameters to measure the quality of data, and the analysis of the railway line Drammen - Eidsvoll during a 15 days period early autumn 2005. The view from which this paper is written from is *NSB Drift*, a subdivision of NSB AS (the Norwegian State Railways).

The conclusion that is drawn from the review of the literature is that it may help the structuring of quality data production to adopt a process analogy. This analogy helps us divide the process into three roles: data collector, database manager, and data consumer, or three sub processes: collection of data, processing of data, and consummation of data. Furthermore, it is essential that the programmed data quality matches the one perceived by the consumer/customer. This may be achieved by identifying the range of use. For NSB Drift the data consumers should specify their area of use with respect to relevant metrics and characteristics of the data product. Those specifications might be brought back in the process chain, thus eventually adjusting the measurements to the need of the consumer. A such improved satisfaction of the consumers needs will help improve the subjective dimension of data quality.

The objective dimension of data quality may be strengthened by measuring the quality of the data using the six suggested parameters in this paper: number of observations, resolution, accuracy, completeness, homogeneity, and timeliness. Data consumed in NSB Drift should be accompanied by values for these parameters in order to give guidance to their reliability. This might help the decision maker understand the validity of his or her decision.

Data from AnnaLyse and TELOC might be used to calculate mean values for station stops, travel time, and delays. But the data sources are unsuited for comparisons of corresponding values owing to the fact that they are not equally calibrated with respect to time, and this deviance is unknown.

The analysis of the Drammen – Eidsvoll indicates that the station stops stemming from TELOC tend to be longer than planned. Delays seem to decrease in frequency as the train approaches its final destination. A large deviance from the values for station stops from AnnaLyse and TELOC could tell us that AnnaLyse is not apt for calculating those values.

The station stops consist of many different components and is again affected by many different factors. In this paper the data material is not sufficient to draw any conclusions regarding those factors using multiple regression analysis.

Both AnnaLyse and TELOC might be labelled high quality data sources, but there is a challenge in adapting the product the consumers needs. AnnaLyse needs to be further developed to with regards to the consumers needs.

INNHALDSFORTEGNELSE

FORORD	VI
SAMMENDRAG	VII
SUMMARY	IX
INNHALDSFORTEGNELSE	X
FIGURER	XIII
1 INTRODUKSJON	14
1.1 OPPGAVENS PROBLEMSTILLING	14
1.2 OPPGAVENS BEGRENŚING	14
1.3 OPPGAVENS STRUKTUR	14
2 TEORETISK BAKGRUNN	16
2.1 OVERSIKT	16
2.2 BEGREPER OG TERMINOLOGI	17
2.2.1 MÅLING	17
2.2.2 DATA, INFORMASJON OG KUNNSKAP	18
2.2.3 METRIKKER	19
2.2.4 DATATYPER	19
2.3 KVALITET OG DATAKVALITET	20
2.3.1 KVALITET	20
2.3.2 DATAKVALITET	22
2.3.3 DIMENSJONER AV DATAKVALITET	24
2.4 DATA SOM EN PRODUKSJONSPROSESS	26
2.4.1 TDQM	28
2.5 DATAKVALITET OG BESLUTNINGER	28
2.6 HVORDAN SKAPE HØYKVALITETSDATA	29
2.7 MÅLING AV DATAKVALITET OG FEILKORRIGERING	30
2.7.1 OBJEKTIVE PARAMETERE	30

2.7.2 SUBJEKTIVE PARAMETERE	32
2.7.3 MODELL FOR DATAKVALITETSPARAMETERE.....	33
2.8 SAMMENDRAG	34
<u>3 METODOLOGI OG DATAKILDER.....</u>	<u>36</u>
3.1 METODEVALG.....	36
3.1.1 FEILKILDER KNYTTET TIL METODEVALG	36
3.2 DATAINNSAMLING	37
3.2.1 TEORI	37
3.2.2 EMPIRI.....	37
<u>3.3 SAMMENDRAG</u>	<u>44</u>
<u>4 ANALYSEDEL.....</u>	<u>46</u>
4.1 BESKRIVELSE AV STREKNINGEN	46
4.1.1 DESKRIPTIV STATISTISK FREMSTILLING	47
4.1.2 DIFFERANSE MELLOM OPPHOLDSTIDER.....	55
4.1.3 SAMMENDRAG	56
4.2 FORKLARINGSPARAMETERE FOR OPPHOLDSTID	57
4.2.1 EMPIRISKE DATA	58
4.2.2 MODELL	61
4.2.3 ESTIMERING	63
4.2.3.1 MODELL 1: AVHENGIG VARIABEL: OPPHOLDSTID ANNALYSE	65
4.2.3.2 MODELL 2: AVHENGIG VARIABEL: OPPHOLDSTID TELOC	67
4.2.3 DISKUSJON AV MODELLER	68
4.3 SAMMENDRAG	69
<u>5 DISKUSJON AV KVALITET PÅ KILDENE.....</u>	<u>71</u>
5.1 TYPER DATA OG MÅLINGER	71
5.2 FORSLAG TIL MÅLEPARAMETERE	71
5.2.1 OBJEKTIVE.....	71
5.2.2 SUBJEKTIVE	72
5.2 TEST AV MÅLEPARAMETERE	72
5.2.1 ANNA LYSE	72

5.2.2 TELOC	74
5.3 ER DATAENE HENSIKTSMESSIGE?	77
5.4 SAMMENDRAG	78
<u>6 KONKLUSJON</u>	<u>79</u>
6.1 OPPGAVENS KONKLUSJON.....	79
6.2 FEILKILDER OG BEGRENSENINGER	80
6.3 FORSLAG TIL VIDERE STUDIER	81
<u>7 REFERANSER.....</u>	<u>82</u>
<u>8 VEDLEGG.....</u>	<u>85</u>
8.1 LOGG OVER RULLENDE MATERIELL.....	85
8.2 RUTEHÅNDBOK	86
8.3 PASSASJERTELLING.....	87
8.4 TRAINPLAN	87
8.5 ANNALYSE	89
8.5.1 TOGGANG OG DATAGRUNNLAG ANNALYSE	89
8.5.2 HISTOGRAMMER OPPHOLDSTID ANNALYSE	89
8.5.3 HISTOGRAMMER KJØRETID ANNALYSE.....	96
8.6 TELOC	103
8.6.1 HISTOGRAMMER OPPHOLDSTID TELOC	103
8.6.2 HISTOGRAMMER KJØRETID TELOC	111
8.7 FORSTUDIERAPPORT.....	120

FIGURER

Figur 1 Oppgavens struktur.....	15
Figur 2 Forholdet mellom programmert, produsert og oppfattet kvalitet.....	23
Figur 3 En data produksjonsprosess.	27
Figur 4 En analogi mellom fysiske produkt og dataprodukt [18]).....	27
Figur 5 Datakvalitetsdimensjoner og måletype.	34
Figur 6 Prosess for å ekstrahere data fra TELOC.....	39
Figur 7 Datakilder, flyt, sammenhenger og oppløsning.....	40
Figur 8 Brukergrensesnitt AnnaLyse.....	41
Figur 9 Variabler punktlighetsstatistikk.....	41
Figur 10 Målinger og parametere fra AnnaLyse og TELOC.....	42
Figur 11 Ruteplansystemet og Trainplan (Kilde: Jernbaneverket).....	43
Figur 12 Gjennomsnittlig passasjerstrøm over pendelen.	47
Figur 13 Total reisetid og oppholdstid.	48
Figur 14 Komponenter reisetid.....	48
Figur 15 Avvik i sekunder fra planlagt oppholdstid (trimmet snitt 20 %)	53
Figur 16 Avvik fra spesifisert kjøretid i TrainPlan i forhold til TELOC (trimmet snitt 20 %).....	54
Figur 17 Faktorer som kan påvirke oppholdstiden.....	58
Figur 18 Venn diagram som illustrer overlappende varians.....	64
Figur 19 Metode, avhengig og uavhengige variabler; AnnaLyse.....	65
Figur 20 Modellens forklaringskraft; AnnaLyse.....	66
Figur 21 Anova; AnnaLyse.	66
Figur 22 Koeffisienter i modellen; AnnaLyse.....	66
Figur 23 Metode, avhengig og uavhengige variabler; TELOC.....	67
Figur 24 Modellens forklaringskraft; TELOC.....	68
Figur 25 Anova; TELOC.....	68
Figur 26 Koeffisienter i modellen; TELOC.	68
Figur 27 Logg over rullende materiell, Tog ID 72-33.	85

1 INTRODUKSJON

Equation Chapter 1 Section 1

1.1 OPPGAVENS PROBLEMSTILLING

Oppgavens problemstilling, som gitt av Institutt for Produksjons- og Kvalitetsteknikk ved NTNU 22.8.2005, er delt i fire:

- (1) Gjennomføre et litteraturstudium relatert til datakilder som skal brukes som grunnlag for faktabasert beslutninger. Et sammendrag av dette skal presenteres.
- (2) Beskrive/belyse toggangen for en utvalgt strekning ved hjelp av ulike datakilder over en viss tidsperiode.
- (3) Analysere de ulike datasettene og presentere resultatene for stasjonsopphold på noen stasjoner med spesiell fokus på å vurdere ulike forklaringsfaktorer.
- (4) Basert på de foregående punktene, beskrive styrker, likheter og ulikheter ved de ulike datasettene.

Jeg har på grunnlag av oppgavens opprinnelige problemstilling, ressurser tilgjengelige, og tidsbegrensning valgt å spesifisere min problemstilling på følgende måte for å best mulig svare på oppgaven innenfor de gitte rammer. Denne spesifiseringen er gjort i samråd med veileder ved NTNU.

A å beskrive og diskutere ulike perspektiv og måleparametere for datakvalitet, samt datakvalitetens eventuelle betydning for beslutninger.

B å beskrive toggangen mellom Drammen og Eidsvoll i perioden 22.08-05.09.2005 med hensyn på oppholdstid, kjøretid, og ankomstforsinkelse.

C å lage en regresjonsmodell med en avhengig variabel (oppholdstid), og to uavhengige variabler (enkelt/dobbeltsett og passasjerantall) for Asker stasjon, med den hensikt å se på forklaringsparametere for oppholdstiden.

D å foreslå noen måleparametere for datakvalitet, subjektive og objektive, basert på bruk av to datakilder og teoretisk grunnlag, samt gjøre et anslag for kvaliteten av datakildene.

E å diskutere resultatene og kommer med forslag til videre studier.

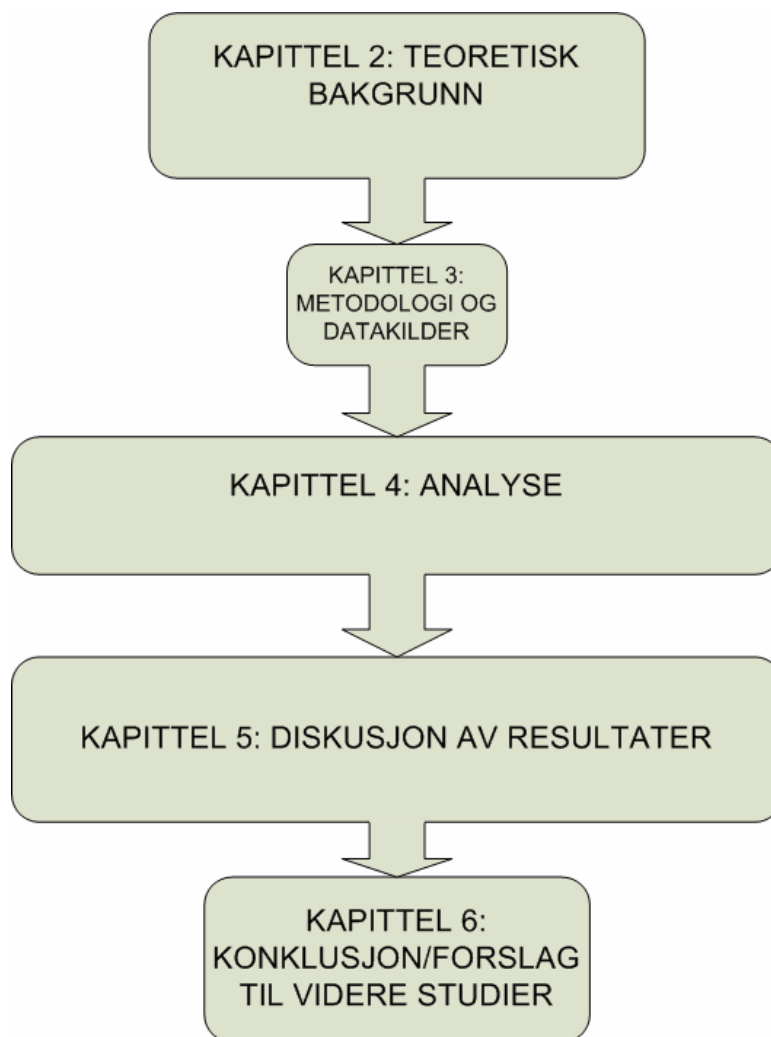
1.2 OPPGAVENS BEGRENŚING

Oppgavens begrensning er først og fremst at den bruker data for en begrenset tidsperiode, det vil si to uker kontinuerlig i tid. En mer realistisk Case ville vært å sammenligne ulike "stikkprøver", men det faller utenfor rammene av denne oppgaven på grunn av begrensinger i tid. Videre brukes kun to datakilder som hovedgrunnlag for analysene, noe som kan føre til en viss begrensning i konklusjonene som kan trekkes.

1.3 OPPGAVENS STRUKTUR

Figur 1 på neste side viser hvordan oppgaven er bygget opp. I kapittel 2 utvikles et teoretisk fundament og parametere for å kunne si noe kvaliteten på de ulike datakildene som benyttes i oppgaven, og i kapittel 4 forsøker man å bruke de tilgjengelige datakildene til å beskrive en togstrekning. Det teoretiske fundamentet og det praktiske erfaringsgrunnlaget danner sammen en basis for å kunne diskutere datakvaliteten utførlig i kapittel 5. Herunder blir det diskutert om dataene er hensiktsmessige; det vil si om de er pålitelige og gyldige.

Opgavens konklusjon samt forslag til videre studier, blir til slutt presentert i kapittel 6. Kapittel 3 gir en oversikt over den metodiske tilnæringsmåten i denne oppgaven samt en nærmere beskrivelse av datakilder som ble benyttet.



Figur 1 Oppgavens struktur

2 TEORETISK BAKGRUNN

Hovedfokuset i denne oppgaven er å undersøke bakgrunnen til ulike datakilder, anvende dem til å produsere ny kunnskap og diskutere kvaliteten av de ulike datakildene. Dette kapitlet gir det teoretiske fundamentet for kunne gjøre slike betraktninger rundt kildekvaliteten.

”Det som kommer ut kan aldri bli bedre det som kommer inn”. Selvfølgelig finnes det en mulighet for å ”dekke over”¹ det svake innholdet, men hvis man preller bort denne fasaden er det veldig vanskelig å ikke innrømme at denne påstanden medfører sannhet. Fisher & Kingma [1] beskriver i en artikkel fra tidsskriftet *Information & Management* hvordan kvaliteten på data er kritisk gjennom å gå nærmere inn på to hendelser som begge medførte tap av menneskeliv og har blitt viet stor oppmerksomhet verden over, nemlig Challenger ulykken og nedskytingen av et iransk Airbusfly fra USS Vincennes. I denne artikkelen fokuseres det på kvaliteten av data som viktig for hendelsesforløpet i motsetning til andre artikler som tradisjonelt har diskutert feil i beslutninger som avgjørende for utfallet. De sier videre at: *”datakvalitet er et av de kritiske problemene som organisasjoner står overfor i dag”* [1, s. 111, min oversettelse]. Ettersom organisasjoner blir mer og mer avhengig av informasjonssystem for å nå deres mål, blir datakvalitet et stadigere viktigere spørsmål i organisasjonen [1].

Informasjon eller kunnskap ligger til grunn for beslutninger som tas i organisasjoner. Man liker å tro at man tar beslutninger på grunnlag av perfekt informasjon, at ulike alternativer vurderes og til slutt velges den beste løsningen. I virkeligheten er verden mye mer kompleks enn som så. For det første har man ikke bestandig tilgang til all informasjonen man trenger for å ta en beslutning. For det andre er det ikke sikkert fordi om man får framlagt informasjon om tre ulike alternativer: at datagrunnlaget som ligger bak er korrekt. Å produsere informasjon handler i mange tilfeller om å fortolke det man observerer i forhold til noe som er forutsatt, selv om det er en såkalt objektiv måling,

Uansett om man tar høyde for at det er vanskelig å framstille fullstendig objektiv og sann informasjon, vil man komme langt på vei i å kunne ta en beslutning basert på rett grunnlag hvis man etterstreber at datakvaliteten er høyest mulig. I dette kapitlet diskuteres det hvordan datakvalitet tradisjonelt blir sett på i litteraturen og hvordan man kan forsøke å oppnå høyest mulig datakvalitet.

2.1 OVERSIKT

Dette teorikapitlet er tredelt: i første del gjøres en begrepsavklaring (avsnitt 2.2), i andre del diskuteres den aktuelle litteraturen rundt datakvalitet (avsnitt 2.3-5), og i siste del gis en presentasjon av noen forslag til noen ulike måleparametere for datakvaliteten (avsnitt 2.6).

¹ Professor ved Yale i USA Edward R. Tuftes påstår for eksempel at presentasjonsverktøyet Microsoft PowerPoint kan virke fordømmende og sier at intelligensen i det man sier blir ikke høyere av å bruke PowerPoint [2].

2.2 BEGREPER OG TERMINOLOGI

I denne delen gjøres det en begrepsavklaring og en terminologi rundt den prosessen som bringer oss fram til ny kunnskap etableres. Videre blir det forsøkt gjort en grenseoppgang mellom det objektive og det subjektive.

2.2.1 MÅLING

Målinger er det som blir benyttet for å skape data og er et begrep som de fleste har et forhold til. Begrepet måling kan både referere til selve kvantifiseringsprosessen i form av verbet *å måle*, og enheten eller standarden for måling. I følge Aschehoug & Gyldendals Store Norske Leksikon defineres *å måle* som: ”*Å måle en størrelse eller egenskap vil si å sammenligne den med en annen størrelse eller egenskap*” [3]. Å måle er altså å sammenligne med en definert størrelse for å kunne si noe om forholdet mellom det vi måler og den definerte størrelsen. Et eksempel kan være en tommestokk som har definerte enheter med en viss størrelse. Når vi måler noe, måler vi hvor stort det er i forhold til disse forhåndsdefinerte enhetene.

Bredrup [4] sier at: ”*all general theories measurement explicitly or implicitly rest in the presupposition of a homomorphism between some empirical relational system and a certain numerical relational system*” [4, s. 54]. Med andre ord er det en forutsetning for å drive måling at det vi observerer empirisk kan omformes til meningsfulle relevante størrelser; at man har et system som samsvarer med det vi observerer.

Diskusjonen rundt samspillet mellom det objektive og det subjektive kommer til å gå igjen i hele dette kapitlet. Det strebes ofte etter å gjøre målinger mest mulig objektive, men det er i de fleste tilfeller umulig å ekskludere det subjektive elementet ved måling. Som nevnt over kan måling sees på som en sammenligningsprosess med en allerede definert størrelse. I denne prosessen vil man svært ofte ha en subjektiv fortolking av målingen. Selv i automatiserte målesystem vil man ha en grad av subjektivitet grunnet i måten systemet er bygd opp. Systemet er alltid lagd av mennesker, som nødvendigvis har hatt en viss forutinntatt subjektivitet i forhold til hvordan sammenligningen, eller målingen skulle gjøres.

I andre sammenhenger som ved måling av for eksempel kundetilfredsstillelse (KTI) er det rett og slett ikke hensiktsmessig å utvikle objektive målinger [4].

Det finnes mange klassifikasjoner av målinger, men i denne oppgaven kommer man til å nøye seg med å skille mellom kvantitative og kvalitative målinger. Kvantitative målinger er det som tidligere er diskutert som objektive målinger, kvalitative målinger er subjektive målinger, og de har de karakteristika som følger:

Kvantitative målinger kjennetegnes ved at de:

- Relaterer seg til kvantitative egenskaper ved objektet som blir målt.
- Kan kvantifiseres ikke bare i grad, men også i størrelse.
- Er resultatet av metoder som veiing, telling, tidtaking etc.
- Tillater en stor grad av nøyaktighet.

Kvalitative målinger kjennetegnes ved at de:

- Relaterer seg til kvalitative egenskaper ved objektet som blir målt.
- Forsøker å måle egenskaper som kvalitet, utseende, brukervennlighet etc.
- Finnes ingen universell kvantitative metode for å måle de.

- Tillater ikke en stor grad av nøyaktighet.

2.2.2 DATA, INFORMASJON OG KUNNSKAP

Ord som *data*, *informasjon* og *kunnskap* brukes ofte for å brukes for mer eller mindre det samme. Det kan derfor være nyttig å prøve å klare opp dette området og forsøke seg på noen definisjoner for data, informasjon og kunnskap.

Redman [5] nevner at det finnes flere definisjoner på data, men han velger å gjøre en todeling i sin definisjon:

”...’data’ (or a collection of data,” etc.) consist of two interrelated components, ‘data models’ and ‘data values.’ ‘Data models’ define what the data are all about. Generally data describe an ‘entity’, some real-world object or abstraction, such as CUSTOMER, EMPLOYEE, or SALE. Attributes and relationships describe pertinent features of the entities. NAME, BIRTH DATE, and ITEM DESCRIPTION, respectively, may be features of interest. ‘Data values’ are assigned to attributes in the data model for specified entities. The ‘September 7, 1954’ in EMPLOYEE BIRTH DATE=September 7, 1954, is a data value for a specified employee” [5, s. 71].

Hvis man ser på definisjonen over, kan man observere at Redman benytter seg av en relativt praktisk definisjon. Han retter først og fremst sin definisjon mot praktisk anvendelse, som for eksempel en database.

Burton-Jones [6] bruker en mer generell angrepsvinkel for å definere data, informasjon og kunnskap:

”data are defined as any signals which can be sent by an originator to a recipient- human or otherwise” (...) “Information is defined as data which as data which are intelligible to the recipient” (...) “Finally, knowledge is defined as the cumulative stock of information and skills derived from use of information by the recipient” [6, s. 5]

Disse to definisjonene er bra, men de virker også, som alle definisjoner, avgrensede. Redman’s definisjon har den svakhet etter min mening, at den trekker ingen klar grense mellom *data* og *informasjon* i den forstand at han behandler alt som data. Derfor mener jeg at Burton-Jones foreslår en tydeligere grenseoppgåing mellom de forskjellige begrepene selv om denne definisjonen avgrensar seg selv ganske kraftig for å kunne skille begrepene.

Det vil gå utenfor denne oppgaven å diskutere mer utførlig hva forskjellen mellom data og informasjon er, men det kan være interessant å merke seg at man ofte beskriver forskjellen mellom de to som at informasjon er ”sannhet” og kan derfor danne en basis for handling [6]. I følge denne tenkemåten vil det være opplagt at kvaliteten av datamaterialet er kritisk. Som man kommer tilbake til senere, vil det være umulig å skape noe bedre enn det materialet man har. Med andre ord: informasjonen kan ikke bli mer ”sann” enn de data man bruker som grunnlag; eller sagt på en mer folkelig måte: man kan ikke lage gull av gråstein. Redman [5] sier at dataverdier ikke nødvendigvis er korrekte. Det ville derfor være feil å kalle data “fakta” eller “sannhet”. For å fremskaffe informasjon er man avhengig av en bearbeiding av dataene, noe som ofte innebærer korrigerer av feil eller tilpasning til brukeren.

Det vil i denne oppgaven i de fleste tilfeller når man snakker om data, snakke om tall fra to forskjellige kilder: TELOC og ANNALyse². For den ene datakilden, den som blir nærmere beskrevet senere i oppgaven som *TELOC*, vil man ha tilgang mer eller mindre direkte til kilden, altså noe som kunne passe relativt bra å kalle data etter definisjonen nevnt over. Mens for den andre kilden, ANNALyse, har man ikke tilgang på kilden, men må derimot betrakte allerede bearbeidet data, altså informasjon, etter definisjonen som data. Derfor for å avklare ting, kommer man i denne oppgaven til å betrakte de to kildene som data, eller datakilder, og produktet av bearbeiding og analysering av disse dataene som informasjon. Den eventuelle kunnskapen blir da det som leseren måtte sitte igjen med etter å ha prosessert dette informasjonsproduktet.

Et annet moment som kan nevnes er at det er i mange tilfeller vanskelig å kunne bruke datagrunnlaget, eller informasjonen, på en nyttig måte hvis man ikke innehar noen kunnskap om bruksområdet. Man kan derfor på en generell basis si at en bruker som allerede innehar kunnskap om bruksområdet vil ha et stort fortrinn når det kommer til fortolkning av data. Lee & Strong [7] utforsket dette temaet i en studie. De stilte seg spørsmålet: *vil bedre kunnskap om dataproduksjonsprosessen³ også føre til bedre datakvalitet for dataforbrukeren?* Mer spesifikt så de på tre typer kunnskap: *vite-hva* (definert som forståelsen av de aktiviteter involvert i en data produksjonsprosess, *vite-hvorfor* (definert som forståelsen av de prosedyrer som finnes for å takle kjente dataproblem) og *vite-hvordan* (definert som egenskapen til å analysere underliggende prinsipper og oppdage tidligere ukjente datakvalitets problemer eller løsninger). Studien viste at en datainnsamlers *vite-hvorfor* var den mest kritiske forutsetningen for høy datakvalitet gjennom hele dataproduksjonsprosessen [7, s 30].

2.2.3 METRIKKER

En metrikk kan defineres som: ”*en karakteristikk av et produkt eller en prosess (...) Metrikkdata er data som representerer verdier knyttet til en metrikk.*” [8, s. 5]. Metrikker kan brukes til å beskrive for eksempel produktivitet, karakteristikk eller rett og slett den direkte tolkingen i forhold til distanse (meter) [4].

Et enkelt eksempel på en objektiv metrikk kan være: ”Antall timeverk brukt til vedlikehold på tog 72-33”. Man kan samle objektive data (telle antall timer) og man kan relatere dette som et forholdstall, eller ratio, til et annet antall. Hvis man derimot omformulerte metrikk til ”vedlikeholdsevnen til tog 72-33”, må man nødvendigvis inkludere en subjektiv måling av togsettets vedlikeholdsevne. Dette kan gjøres ved at man for eksempel lar personellet vurdere vedlikeholdsevnen i forhold til en skala.

2.2.4 DATATYPER

På samme måter som for mål og målinger kan man dele inn data i to forskjellige typer: subjektive og objektive data. Nå har det blitt diskutert tidligere i denne oppgaven at de fleste former for målinger ofte innebærer en eller annen form for subjektivitet. Det er likevel nødvendig for oversikten i denne oppgaven, å dele innsamlede data i to på denne måten. Uten en slik inndeling vil det være umulig å kategorisere, klassifisere og dermed gjøre vurderinger av datakildene.

² Datakildene kommer til å bli beskrevet nærmere senere i oppgaven.

³ Begrepet *dataproduksjonsprosess* blir diskutert mer utførlig senere i dette kapitlet.

2.2.4.1 OBJEKTIVE DATA

Objektive data kan sies å være de dataene som framkommer av kvantitative målinger; eller med andre ord produktet av det som ble definert som kvantitative målinger i avsnitt 2.2.1. En mulig definisjon kan være det følgende: ”Objektive data er data som kan måles slik at alle som utfører målingen alltid får samme resultat. Det er ingen personlige vurderinger involvert. Objektive data medfører ofte at man teller noe og at måten man teller på er entydig definert.” [8, s. 6]. Det som er viktig å merke seg er forskjellen mellom måling og data: måling er prosessen og data er produktet av prosessen.

Man velger i denne oppgaven å betrakte objektive data som:

- Data som er samlet inn ved hjelp av automatiserte verktøy, eller automatiserte metoder der det ikke inngår noen form for vurdering.
- Data som er telt og kategorisert i henhold til en universell anvendbar kategorisering⁴.

2.2.4.2 SUBJEKTIVE DATA

I motsetning til objektive data inneholder subjektive data en eller annen form for subjektiv vurdering i måleprosessen. Man kan derfor tenke seg et resultat som er avhengig av erfaring og holdinger hos datainnsamleren. Denne subjektiviteten kan komme til syne både i tolkning av hvordan målingen skal foregå (regler) og i rene subjektive vurderinger av det man observerer.

Wedde [8] nevner at slike subjektive data ofte framkommer når man ikke kan måle metrikken direkte, eller det er svært ressurskrevende å måle metrikken. Man kan da tenke seg en situasjon der man ber noen subjektiv anslå verdier for metrikken.

2.3 KVALITET OG DATAKVALITET

Denne oppgaven handler om data i jernbanedrift og målet med dette kapitlet er som tidligere nevnt å få på plass et teoretisk fundament for å kunne vurdere datakvaliteten på ulike kilder i jernbanedrift. Til nå har man sett på begreper og terminologi knyttet til målinger og innsamling av data. Oppgaven beveger seg nå over til å se på hvordan man kan vurdere kvaliteten av de innsamlede dataene og hvilken innvirkning det har på beslutninger i en organisasjon.

2.3.1 KVALITET

Før man begynner å snakke konkret om datakvalitet er det på sin plass å se på kvalitetsbegrepet, som er av natur u håndgripelig.

Vanlige og klassiske måter å definere kvalitet på relaterer seg ofte til evnen et produkt har til å tilfredsstille kundenes krav og forventninger, eller produktets brukervennlighet [9,10,11]. I den forberedende prosjektoppgaven til denne masteroppgaven ble det i litteraturstudiet utførlig diskutert kvalitetsbegrepet og derfor kommer det ikke til å gås i detalj på det feltet, men det bør nevnes at kvalitet i hovedsak dreier seg om å tilfredsstille behovet til brukeren [12].

Å gi en kortfattet, konsis definisjon av begrepet kvalitet ville være en svært vanskelig oppgave. Garvin [13] nevner for eksempel fem forskjellige mulige angrepsmåter for å

⁴ Dette inkluderer data som er målt med etter forskjellige enheter så lenge de kan konverteres til samme enhet.

definere kvalitetsbegrepet: *opphøyd, produktbasert, brukerbasert, produksjonsbasert og verdibasert*. Videre definerer han åtte forskjellige dimensjoner av kvalitetsbegrepet: *ytelse, egenskaper, pålitelighet, overensstemmelse, bestandighet, tilgjengelighet, estetikk og oppfattet kvalitet*.

Aune [9] holder seg i motsetning til Garvin [13] til tre forskjellige angrepsmåter for kvalitetsbegrepet:

- ”1 Produkt- og/eller brukerbasert: Kvaliteten beskrives av produktegenskaper (som underforstått tilfredsstillende brukerens behov).
- 2 Produksjonsbasert: Kvalitet vil si fravær av feil – overensstemmelse med gitte spesifikasjoner.
- 3 Følelsesbasert: Kvalitet innebærer noe ettertraktelsesverdig og udefinert, som du imidlertid kjenner igjen når du ser det – noe luksuriøst, noe som koster mer”. [9, s.17]

Videre gjør Aune en definisjon av begrepene behov, forventning, krav og spesifisering som vist i Tabell 1.

Begrep	Definisjon	Klassifisering	Kommunikasjonsmåte
Behov	Fravær av noe som er krevd, ønsket eller nyttig; en tilstand som krever anskaffelse eller avhjelping	Objektivt definerbart	Uformell, oversikter, rapporter
Forventning	Forutsigelse av tingenes framtidige tilstand, av framtidige fordeler eller av måter behov vil bli dekket på	Subjektiv	Uformell, verbal
Krav	Formell beskrivelse av behov og forventet måte de skal bli dekket på	Mottakers (kundes ellers brukers) syn på produkt/tjeneste	Dokument
Spesifisering	Formell beskrivelse av vare/tjeneste og planlagt måte å frambringe den på	Leverandørs syn på vare/tjeneste	Dokument

Tabell 1 Definisjon av begrepene behov, forventning, krav og spesifisering (fritt etter Aune [9])

Allerede nå kan man ane at begrepet kvalitet er særdeles kompleks av natur. For det første virker det å ligge en kompleksitet i det at man i begrepet kvalitet legger en subjektiv dimensjon; hvordan kan man måle, standardisere og spesifisere en subjektiv dimensjon? For det andre virker som om det som vi oppfatter som kvalitet, å være forskjellig fra person til person, fra bruksområde til bruksområde, selv om man kunne spesifisere sine krav objektivt. Et produsert rør kunne for eksempel oppfylle kvalitetskravene innenfor ett bruksområde, men ligge utenfor toleransegrensene for et annet bruksområde.

For å foreslå noen andre definisjoner av kvalitet kan man ta med NS – EN ISO 2000:9000 [11], som definerer begrepet kvalitet på følgende måte:

”I hvilken grad en samling av iboende egenskaper* oppfyller krav** ” [11, s13]

* Behov eller forventning som er angitt, vanligvis underforstått eller obligatorisk

** Særpreg som gjør det mulig å skjelne

fysiske (for eksempel mekaniske, elektriske, kjemiske eller biologiske egenskaper)

sensoriske (for eksempel forbundet med lukt, berøring, smak, syn, hørsel)

atferdsmessige (for eksempel vennlighet, ærlighet, sannferdighet)

tidsmessige (for eksempel punktlighet, pålitelighet, tilgjengelighet)

ergonomiske (for eksempel fysiologisk egenskap, eller som vedrører sikkerhet for mennesker)

funksjonelle (for eksempel et flys maksimale hastighet)

Juran [14] definerer kvalitet på følgende korte måte: ”*Quality is fitness for use*” [14, s. 2-2].

En mer filosofisk, og kanskje enda mer forvirrende, definisjon finner vi hos Pirsig [15]: ”*Quality isn’t either mind nor matter, but a third entity independent of the two (...) even though Quality cannot be defined, you know what it is.*” [15, s. 185, 213]

For denne oppgaven kommer man til å fokusere på det aspektet ved kvalitet som innebærer feilfrihet. Dette er min mening det som man kan se på som objektivt ved kvalitet. Det subjektive ligger da i kundens, eller brukerens, vurdering av om produktet tilfredsstillende de krav og forventninger han måtte ha. Men så lenge man tilstreber en høyest mulig grad av feilfrihet vil det danne et solid grunnlag for bruk.

2.3.2 DATAKVALITET

I det foregående avsnittet er det forsøkt å trekke opp noen skillelinjer mellom data og informasjon, samt å forklare kompleksiteten i kvalitetsbegrepet. Hensikten har med dette vært å kunne ha en viss plattform når man nå forsøker å introdusere begrepet *datakvalitet*.

Innen området datakvalitet har det hovedsaklig blitt forsket på tre hovedområder: *definisjoner* (herunder dimensjoner av datakvalitet), *modellering* og *kontroll*. Videre kan det nevnes at datakvalitet ofte blir behandlet som et *indre konsept* [16]: kontekstuavhengig. I følge det som har blitt nevnt tidligere i forhold til subjektivitet og objektivitet, stiller jeg meg kritisk til å behandle datakvaliteten som strengt kontekstuavhengig. Datakvaliteten er nødvendigvis, som det også kommes inn på senere i oppgaven, ett produkt av konteksten i den forstand at et helhetlig bilde av datakvaliteten inkluderer en subjektiv dimensjon.

Datakvalitet har som område blitt relativt grundig studert og referer ofte til i hvilken grad data tilfredsstillende brukerens krav, eller er passende for en spesiell prosess [17]. For en mer grundig gjennomgang av forskningen som er gjort på datakvalitet enn det som blir gjort i denne oppgaven, og et mulig rammeverk for å sortere de ulike bidragene på området, vises det til for eksempel Wang et al [18].

Et mulig forslag til definisjon av datakvalitet er gitt av Redman [5]:

”Data are of high quality if they are fit for their intended uses in operations, decision making and planning. Data are fit for use if they are free of defects and possess desired features” [5, s73].

Dette er en definisjon som er helt i tråd med det som har vært tidligere nevnt i denne oppgaven rundt begrepene data og kvalitet. Definisjonen tar både hensyn til at dataene skal være minst mulig feilfri, og at de er hensiktsmessig til sitt bruk.

Strong, Lee, og Wang [16] definerer høy datakvalitet på følgende måte: ”*We define high-quality data as data that is fit for use by data consumers*” [16, s 104].

En annen tilnærming, foreslått av Orr [19], til definisjon av datakvalitet kan være som et prosentvis avvik mellom den virkelige verdenen og den informasjonen som dataene gir. Datakvalitet på 100 % vil da for eksempel indikere at dataene gir oss informasjon som sammenfaller fullstendig med vår oppfatning av den virkelige verden, mens 0 % vil indikere ingen overensstemmelse. I praksis vil ikke noe informasjonssystem ha en datakvalitet på 100

% . En slik tilnærming kan minne om det diskutert under avsnitt 2.2.1 om måling, der man legger til grunn for måling at det er en sammenlikning mellom det vi observerer og noe allerede definert. Avviket burde da være feilen i målingen; med andre høyest mulig datakvalitet innebærer minst mulig avvik.

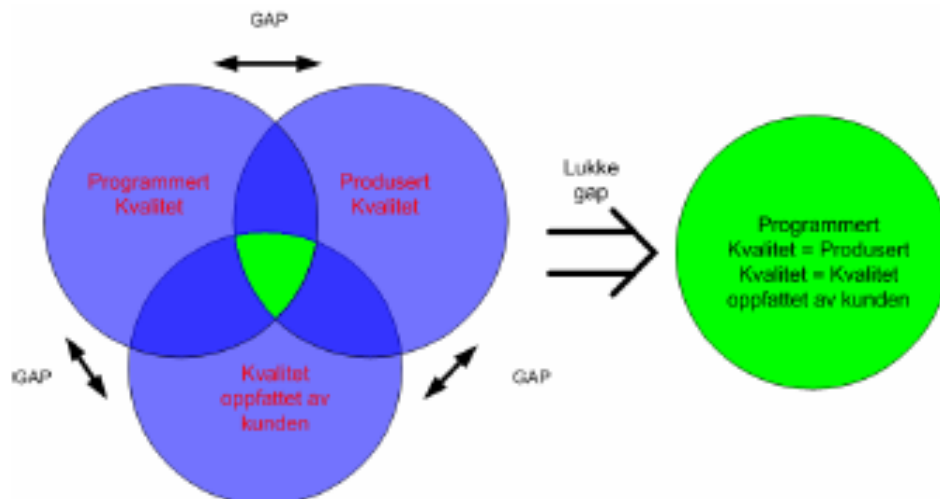
Med bakgrunn i Redman's [5] definisjon og det som har blitt diskutert rundt begrepene kvalitet, måling og data, foreslås følgende definisjon for kvalitetsdata til bruk i denne oppgaven:

Kvalitetsdata kjennetegnes ved at de:

- Har minst mulig feil (objektivt, kvantitativt).
- Tilfredsstillter kundens/brukerens behov (subjektivt, kvalitativt).

Denne foreslåtte definisjonen inkluderer både et indre perspektiv på datakvalitet (datakvalitet er kontekstuavhengig), og et ytre perspektiv som tar hensyn til at datakvalitet ikke bare handler om minst mulig feil, men er i stor grad avhengig av datakonsumentens bruksområde.

Umiddelbart etter å ha foreslått en slik definisjon kommer det opp to nye spørsmål. For det første: *hvordan kan vi vite at datamaterialet har minst mulig feil*, og for det andre: *hva er behovet til kunden/brukeren*. Når det gjelder det første spørsmålet kommer det til å bli foreslått noen måleparametere senere i dette kapitlet som man kan bruke som indikasjon på nettopp dette. Det andre spørsmålet handler om å kjenne bruksområdet for i dette tilfellet, datamaterialet. Ved å best mulig kartlegge behovet til brukeren, er man nærmere å kunne få høy datakvalitet. Her kan det være naturlig å trekke en analogi til generell kvalitetsteori. Bustinduy [20] sier at en kvalitetstilnærming handler om å lukke gapet mellom den produserte og den oppfattede kvaliteten (jfr Figur 2).



Figur 2 Forholdet mellom programmert, produsert og oppfattet kvalitet.

Man kan tenke seg at å kartlegge behovet til dataforbrukeren innebærer å minske gapet mellom den produserte og den oppfattede kvaliteten, men som med en generell kvalitetstankegang er det viktig å tilstrebe minst mulig feil fordi det ofte er en viktig faktor for forbrukeren.

Utfordringen ligger som sagt i å kunne ha tilgjengelig rett data til rett bruk. For å kunne ha tilgjengelig rett data til rett bruk trenger man nødvendigvis å kjenne behovet til brukeren. Wand & Wang [21] argumenterer for at data- eller informasjonskvaliteten, avhenger av hvordan dataene blir brukt. Det som kan betraktes som god data av en bruker eller

bruksområde, er ikke nødvendigvis bra nok for et annet formål. Dilemmaet med ”*relativity of quality*” [21] ser de på som et problem, og de mener løsningen bør ligge i å *ontologisk* forankre dimensjonene av datakvalitet som en guide for systemdesignere når det gjelder spørsmål rundt datakvalitet. Målet er å kunne støtte spesifikasjoner av datakvalitetskrav allerede i designprosessen av systemet. Dimensjonene av datakvalitet blir definert ut fra avvik mellom en bruker oppfatning av informasjonen fra systemet og den virkelige verden; altså samme tankegang som Orr [19]. Man tar med andre ord utgangspunkt i at informasjonssystemet gir en representasjon av den virkelige verden, og at brukeren har en oppfatning av hva som er virkelig. Med disse nevnte forutsetningene blir fire dimensjoner av datakvalitet definert: *riktighet, utvetydighet, fullstendighet og meningsfullhet*⁵ [21]. Det er viktig å merke seg at disse fire nevnte dimensjonene behandler datakvaliteten som et indre konsept, uavhengig av konteksten som data blir produsert og brukt i.

2.3.3 DIMENSJONER AV DATAKVALITET

På samme måte som for begrepet kvalitet, har datakvalitet også mange forskjellige dimensjoner ut fra en slik definisjon som ble foreslått i forrige avsnitt. I forrige avsnitt ble det allerede foreslått fire dimensjoner: riktighet, utvetydighet, fullstendighet og meningsfullhet. Wand & Wang [21] sier at hvordan man velger disse dimensjonene er primært basert på intuitiv forståelse, erfaring fra industrien, eller litteraturstudier. Men en litteraturstudie utført av Wang et al [18], har vist at det eksisterer ingen generell konsensus rundt dimensjonene av datakvalitet.

Wand&Wang [21] sier at: ”*In short, despite frequent use of certain terms to indicate data quality, there does not exist a rigorously defined set of data quality dimensions.*” [21, s. 87]. Det kan likevel sies at det finnes en god del litteratur når det gjelder de forskjellige dimensjonene av datakvalitet, og en mulig forklaring på at det hersker en viss uoverensstemmelse kan være nettopp at man må nødvendigvis involvere begrepet kvalitet.

Redman[5] forsøker seg til tross for at det som tidligere nevnt ikke finnes noen konsensus rundt dimensjonene for datakvalitet, på å utforske de ulike dimensjonene. Redman’s forslag til tabellmessig oppsummering av 27 forskjellige dimensjoner for datakvalitet er gitt i Tabell 2. Dimensjonene er delt inn i 3 hovedkategorier: begrepsmessig syn, verdier og representasjon.

⁵ Henholdsvis: *Correctness, unambiguous, completeness, and timeliness.*

Begrepsmessig syn

Innhold	Relevans	Tilgjengelighet	Klarhet i definisjon
Omfang	Omfattende kunnskaper	Vesentlighet	
Detaljnivå	Detaljnivå egenskapsdata	Målenøyaktighet	
Sammensetning	Naturlighet Homogenitet	Unikt identifiserbar Minimum av unødvendig data	
Ensartethet	Semantisk ensartethet	Strukturell ensartethet	
Endring	Robusthet	Fleksibilitet	
Verdier			
	Nøyaktighet	Fullstendighet	
	Overensstemmelse	Livssyklus	
Representasjon			
Formater	Egnethet Fortolkbarhet Flyttbarhet	Presisjon Fleksibilitet i format Evne til å representere null verdier	Effektiv bruk av lagringsplass
Data verdier	Ensartethet i representasjon		

Tabell 2 27 kvalitetsdimensjoner for datakvalitet (fritt etter Redman[5])

Tabell 2 gir en relativt omfattende beskrivelse av forskjellige dimensjoner av datakvalitet og ikke alle dimensjonene gjengitt i tabellen er like relevante for denne oppgaven. Tabellen kan likevel være en god indikasjon på de forskjellige innfallsvinklene det finnes til begrepet datakvalitet og gir kanskje en pekepinn på hvorfor det ikke hersker en samlet enighet om hvilke dimensjoner begrepet omfatter.

En litteraturstudie utført av Wand & Wang [21] publisert i 1996 viste at nøyaktighet, pålitelighet, tidsriktighet, relevans, og kompletthet var de mest siterte kvalitetsdimensjonene med henholdsvis 25,22,19,16, og 15 siteringer. På bunnen fant man informativhet, detaljnivå, mengde, omfang, og forståelighet alle med 2 siteringer⁶.

Lee et al. [22] deler *informasjonskvalitet*⁷ inn i fire kategorier: reell kvalitet, kontekstavhengig kvalitet, representerbar kvalitet og kvalitet knyttet til tilgjengelighet. Representerbar kvalitet

⁶ Det er nødvendig med en stor grad av ”fornorskning” av etablerte engelske begreper og det vises til kildelitteraturen for korrekte begrep. I dette tilfellet: Wand&Wang [21].

⁷ Informasjonskvalitet og datakvalitet er begreper som i stor grad overlapper hverandre i litteraturen.

og kvalitet knyttet til tilgjengelighet, dreier seg om systemet for lagring og adgang til informasjonen. Reell kvalitet handler om at informasjonen har kvalitet uavhengig av konteksten. Kontekstavhengig kvalitet utdyper det at informasjonen bestandig må sees i forhold til sitt bruksområde; den må være *relevant*, *tidsriktig*, *fullstendig* og *passende i mengde* [22]. Et forslag til sortering av de ulike dimensjonene for datakvalitet etter denne kategoriseringen foreslått av Lee et al. [22], er gitt av Strong, Lee og Wang [16] og gjengitt her i Tabell 3.

Kategori av datakvalitet	Dimensjon av datakvalitet
Indre datakvalitet	Nøyaktighet, Objektivitet, Trolighet, Ry
Tilgjengelighet datakvalitet	Tilgjengelighet, Tilgangssikkerhet
Kontekstuell datakvalitet	Relevans, Merverdi, Tidsriktighet, Kompletthet, Mengde data
Representerbar datakvalitet	Fortolkbarhet, Forståelighet, Konsis representasjon, Konsekvent representasjon

Tabell 3 Datakvalitet – kategorier og dimensjoner (fritt etter Strong, Lee og Wang [16]).

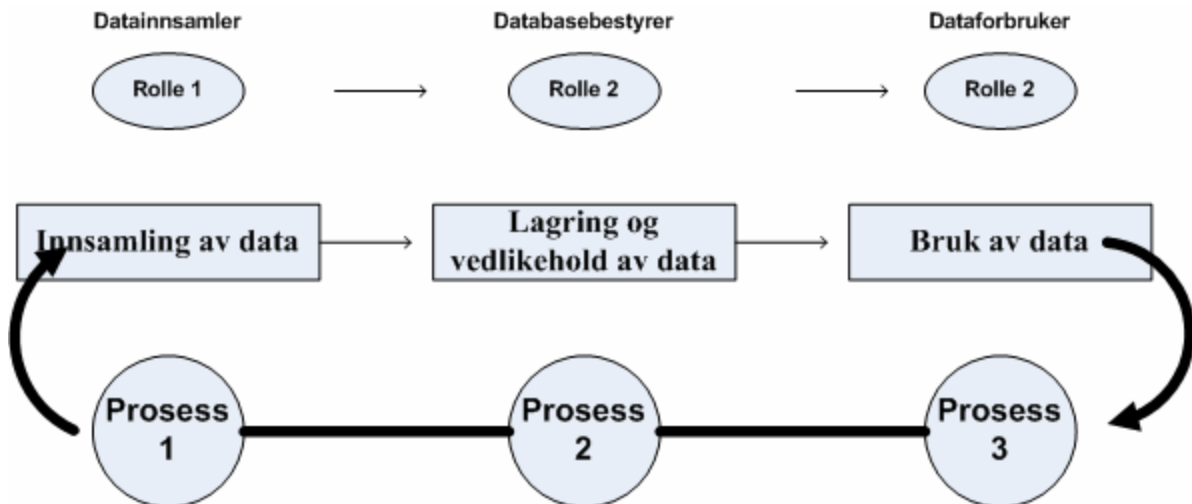
Ballou & Pazer [23] nevner noen etter deres mening nøkkeldimensjoner ved datakvalitet: *nøyaktighet* (den registrerte verdien er i overensstemmelse med virkelige verdien), *tidsriktighet* (den registrerte verdien er ikke gått ut på dato), *fullstendighet* (alle verdier for en variabel er registrert), og *ensartethet* (fremstillingen av data verdien er den samme for alle tilfeller). De behandler ikke direkte de mer subjektive dimensjonene ved datakvalitet selv om de anerkjenner gapet som finnes mellom forventningene til brukeren og hva informasjonssystemet faktisk leverer [22].

Man kan derfor med rimelighet si at kvaliteten på data er avhengig av hva datamaterialet skal brukes til siden kvalitet er som nevnt knyttet til hva som blir oppfattet av brukeren, eller kunden om man vil.

2.4 DATA SOM EN PRODUKSJONSPROSESS

Styring av *produktkvaliteten* er et område som har eksistert i mange år og typiske måter å kontrollere kvaliteten på har vært ved hjelp av statistisk prosesskontroll og pålitelighetsstudier[9,10]. Likevel går det an å trekke mange paralleller mellom kvalitet på et produkt og kvalitet av data.

Som i produksjonen av et fysisk produkt kan en dataproduksjonsprosess også deles inn i karakteristiske arbeidsprosesser: *innsamlings-*, *lagrings-*, og *bruksarbeidsprosesser* [16]. Man kan videre identifisere tre roller i en data produksjonsprosess: *datainnsamlere* (personer, grupper, eller andre kilder som genererer informasjon), *data(base)bestyrere* (personer som administrerer beregningsressurser for lagring og prosessering av data), og *dataforbrukere* (personer eller grupper som bruker data) [7]. Skjematisk kunne man tenke seg denne dataproduksjonsprosessen noe sånn som vist i Figur 3.



Figur 3 En data produksjonsprosess.

Wang, Storey og Firth [18] støtter seg på denne analogien mellom et produksjonssystem og et informasjonssystem i sitt forsøk på å etablere et rammeverk for de studiene som har blitt gjort inne området datakvalitet. I denne analogien ses data som råmaterialet (input) og informasjon eller dataproduktet, som output. En typisk skjematisk oppstilling av denne analogien kunne se ut som vist i Figur 4.

	Vareproduksjon	Data produksjon
Input	Råmateriale	Rå data
Prosess	Bearbeiding av materiale	Bearbeiding av data
Output	Fysisk produkt	Data produkt

Figur 4 En analogi mellom fysiske produkt og dataprodukt [18]).

Som man kan se av Figur 4 kommer det i et tradisjonelt produksjonssystem et materiale som *input*, blir bearbeidet, og man får et fysisk produkt som *output*. I et informasjonssystem kan man se dette på samme måte med den forskjellen at man ser på rådata som råmaterialet. Informasjonssystemet bruker rådata (for eksempel tall, en fil, et regneark, eller en rapport) som *input* for å produsere data (for eksempel en sortert fil, eller en korrigert e-postliste). Produserte data kan igjen brukes som rådata i et annet data produksjonssystem [18].

2.4.1 TDQM

Analogien mellom produksjon av et fysisk produkt og produksjon av data, har blitt utviklet videre fra den til dels omfattende litteraturen og metodene rundt det som i dag kalles TQM (Total Quality Management), til det som går under betegnelsen TDQM (Total Data Quality Management).

TQM, eller TKL (Total KvalitetsLedelse) på norsk, kan defineres som:

”En ledelsesform i en organisasjon, fokusert på kvalitet, som baseres på medvirken fra alle medarbeidere og der langsiktig suksess tilstrebes ved å oppnå kunders tilfredshet og fordeler for alle medarbeidere og for samfunnet.” [24].

Aune [9] peker på at i en kvalitetsstyrt bedrift omfatter TKL tre hovedoppgaver:

- Kvalitetsvedlikehold omfatter kvalitetsstyring og – sikring og dreier seg om utførelse av tildelte oppgaver etter fastlagte standarder i eksisterende prosesser. Effektivt kvalitetsvedlikehold krever kvalitetsoppmerksomme, fagutdannede medarbeidere. Kvalitetsvedlikehold er systemorientert og representerer ”ledelse for status quo”.
- Kvalitetsforbedring (kaizen) omfatter mindre forandringer i eksisterende standarder, prosesser og produkter, og krever i tillegg skoling i kvalitetstenkning og problemløsning. Antallet gjennomførte forbedringsprosjekter bestemmer forbedringstakten, og kostnadene er små.
- Kvalitetsfornyning (kairoyo) omfatter større forandringer i standarder, prosesser og produkter og ny teknologi, og er helt avhengig av et tillitsfullt og kreativt miljø, og kostnadene kan bli store.

TDQM er et område som er under utvikling og det gjenstår fortsatt mange terreng som må utforskes for å kunne utvikle en selvstendig metodologi for dette området [25]. Noen av begrensingene som Wang [25] beskriver i en artikkel fra *Communications of the ACM*, er at det er vesentlige forskjeller mellom råmateriale og rådata, og dermed visse svakheter ved bruk av denne analogien. Rådata kan for eksempel brukes av mange brukere uten at man utarmer ressursen, eller forringer det på noen måte. Et råmateriale derimot kan bare brukes til ett fysisk produkt [25]. Andre svakheter ved denne analogien mellom rådata og råmateriale knyttes også i forhold til de parametrene som man tradisjonelt anvender for å beskrive datakvalitet.

2.5 DATAKVALITET OG BESLUTNINGER

Denne oppgaven handler i hovedsak om data i jernbanedrift og kvaliteten på datakilden, men disse dataene er grunnlaget for informasjon, som igjen brukes som beslutningsunderlag. Tidligere i dette kapitlet har det blitt diskutert hvordan datakvalitet er et særdeles kompleks og ofte subjektivt begrep. Hvordan informasjonen blir oppfattet er også i mange tilfeller avhengig av hvordan den blir presentert. Det er derfor rimelig å stille spørsmål ved begrepet *faktabasert styring*. Finnes fakta? På hvilket grunnlag har disse faktaene kommet fram?

En organisasjon eller bedrift fatter beslutninger på mer eller mindre kontinuerlig basis på forskjellige områder. Gripsrud et al. [26] deler disse beslutningene inn i to kategorier: *rutinemessige* og *strategiske*. Rutinemessige beslutninger følger en fast prosedyre for hva som skal gjøres når noe inntreffer. Et eksempel på en rutinemessig beslutning kan være å ta inn et togsett til vedlikehold hvis man passerer en viss kjørelengde, eller starte et forbedringsprosjekt på en strekning hvis punktligheten tipper under 90 % i løpet av en 6 måneders periode. Man kan si at rutinemessige beslutninger har klare rammer.

Strategiske beslutninger er forbundet med mye høyere grad av usikkerhet siden det kan dreie seg om inngåelse av samarbeid, omorganisering av vedlikeholdet, satsing på nye produktområder og så videre. Usikkerheten i disse beslutningene kommer av at de involverer omgivelsene, som igjen kan endre seg [26].

Når det gjelder en organisasjons beslutningsprosess, har det blitt gjort studier både innen psykologi og ledelsesteori og kan veldig grovt deles i to: *normative* modeller, det vil si hvordan beslutninger bør treffes, og *deskriptive* modeller, det vil si hvordan beslutninger faktisk treffes [26]. Det faller utenfor denne oppgavens rammer å oppsummere den til dels omfattende litteraturen på dette området. Eisenhardt og Zbaracki [27] gjør en oppsummering av litteraturen innen strategisk beslutningstaking. De deler forskningen innen strategisk beslutningsfatning i tre paradigmer:

- Beslutningsprosessen er preget av rasjonalitet eller begrenset rasjonalitet.
- Beslutningsprosessen er preget av politikk og maktutøvelse.
- Beslutningsprosessen er preget av uklare mål og flytende deltakelse.

Det viktige å legge merke til er forskjellen mellom på den ene side modellen med den rasjonelle beslutningstaker som har perfekt informasjon om alle mulig alternativer, kan vurdere alle mulige konsekvenser av disse alternativene, kan rangere alternativer på basis av konsekvensene og kan velge det alternativet som rangerer høyest. På den andre siden har man Cohen et al [28] sin søppelkassemodell, som fremstiller beslutningsprosesser som ustrukturerte, avhengig av tilfeldigheter og uten klare mål. Man har forskjellige personer som deltar til enhver tid i prosessen, og disse er ofte ute etter å kun treffe beslutninger. Dette sår med andre ord tvil om beslutningsprosessen virkelig er så strukturert og formalisert som i en rasjonell tankegang, men det ligger likevel et viktig poeng i at høykvalitetsdata og korrekt informasjon bidrar til å lette beslutningsprosessen i den forstand at det blir enklere å gjøre et valg mellom de alternativene som foreligger.

2.6 HVORDAN SKAPE HØYKVALITETSDATA

Et mål for enhver organisasjon burde være å forsøke forsyne brukeren med data av høyest mulig kvalitet. Som nevnt tidligere er det rimelig å anta at datakvalitet er kontekstavhengig; det vil si at det er avhengig av bruksområdet. Det er mulig å tenke seg en analogi til forholdet mellom kvalitet på servicetjenester og produkter. Servicetjenester er av natur uåtgripelige og heterogene, mens produkter er i mange tilfeller (i alle fall for vareproduserende industri) homogene og åtgripelige [29]. Kompleksiteten i å drive kvalitetskontroll på informasjonsprodukter ligger derfor nettopp i dette forholdet. Fordi om det har blitt hevdet at man kan bruke de samme metoder for statistisk prosesskontroll på tjenester som for produkter [30], gir de objektive måleparametrene som ved statistisk prosesskontroll ikke nødvendigvis noen hjelp for å tilfredsstille brukerens subjektive behov. For å sitere Wang et al: "*Even accurate data, if not interpretable and accessible for the user, is of little value*" [31, s. 670].

En interessant studie er blitt gjort av Kahn et al [32] der de ser på en mulig metodologi for å *benchmarke*⁸ i hvilken grad en organisasjon utvikler informasjonsprodukter og leverer informasjon til konsumenter. En benchmarkingsmodell, PSP/IQ, ble konstruert bygd på

⁸ *Benchmarking* betyr i praksis prestasjonsmåling på norsk, men det engelske begrepet blir for det meste brukt. Se for eksempel: [35,36,37] for en nærmere beskrivelse av konseptet.

tankegangen om data som en produksjonsprosess vist i Figur 4. De dimensjonene som ble brukt som benchmarkingsmål var: *anvendelig* (kvalitet på service møter/overgår kundens forventninger), *ekte* (kvaliteten på produktet er i overensstemmelse med spesifikasjoner), *pålitelig* (kvaliteten på service er i overensstemmelse med spesifikasjoner), og *nyttig* (kvaliteten på produktet møter/overgår kundens forventninger). Andre konklusjoner som kom ut av studien var at dataprodusenten, databasebestyreren, og dataforbrukeren alle er essensielle for å levere kvalitetsinformasjon som produkt og service [32, s. 192].

2.7 MÅLING AV DATAKVALITET OG FEILKORRIGERING

Måling av datakvalitet kan gi en pekepinn på i hvor stor grad man leverer kvalitetsinformasjon. En slik måling må nødvendigvis ta hensyn til både de objektive og de subjektive dimensjonene av datakvalitet for å gi et helhetlig bilde av informasjonskvaliteten. Subjektiv datakvalitet reflekterer behovene og erfaringene til produsent, bestyreren, og forbrukeren av et informasjonsprodukt [33,34]. Det finnes til dags dato ikke et fullstendig sett av måleparametere for datakvalitet som inkorporerer både de objektive og de subjektive måleparametere, og de som finnes har i stor grad blitt utviklet *ad hoc* tilpasset den enkelte bedrift. Faktisk er det slik at de fleste mål på datakvalitet blir utviklet for å løse spesifikke problem og har i det en iboende subjektiv dimensjon [17]. Det vil være nødvendig å utvikle objektive og klart definerte måleparametere for de ulike (objektive) dimensjonene av datakvalitet.

Tidligere i denne oppgaven har det blitt gjort rede for en del dimensjoner av datakvalitet man kan støte på i litteraturen. Fordi om det kan herske uenighet om definisjoner har man identifisert i dette litteraturstudiet at det har blitt foreslått noen ulike parametere som uttrykke disse forskjellige dimensjoene. Jeg kommer i denne oppgaven til å foreslå *antall observasjoner, oppløsning, nøyaktighet, fullstendighet, ensartethet, tidsriktighet, tolkbarhet, og tilgjengelighet* som de viktigste dimensjonene basert på den litteraturen som er blitt gjennomgått i arbeidet med denne oppgaven. I litteraturen kan man finne algoritmer for eksempel nøyaktighet, fullstendighet, ensartethet, og tidsriktighet, mens for parametre som tolkbarhet og tilgjengelighet lar det seg vanskelig gjøre å definere objektive mål. Det betyr at de er i stor grad basert på brukerens vurdering [17].

I det følgende blir de forskjellige måleparametrene presentert og diskutert. Der det er mulig vil det blir foreslått kvantitative mål for disse parametrene. Det blir i oppgaven benyttet en todeling; henholdsvis objektive og subjektive parametere. Disse parametrene er forfatterens forslag på grunnlag av den gjennomgåtte litteraturen og hensynet til de datakildene som benyttes.

2.7.1 OBJEKTIVE PARAMETERE

De objektive parametere referer i hovedsak til de dimensjonene av datakvaliteten som man kan sette et mål på uten å nødvendigvis ta hensyn til bruksområdet. I praksis vil slike parametere også måtte ta hensyn til bruksområdet og gir ofte mer mening hvis man sammenligner med andre datakilder. Likevel skal man kunne beregne verdier for disse objektive parametrene som er uavhengige av bruksområdet.

2.7.1.1 ANTALL OBSERVASJONER

Antall observasjoner referer til hvor mange observasjoner man kan velge blant. Dette antallet må nødvendigvis relateres til en ramme; for eksempel tid. Det er naturlig å tenke seg at flere

observasjoner vil representere et bedre utvalg, men det er ikke nødvendigvis sant da det viktigste er å ha de rette observasjonene. Likevel er man i praksis ofte interessert i å bruke datamaterialet til å se på et fenomen innefor en viss ramme. Da er det viktig at man har et utfyllende datamateriale å velge blant.

Målet for antallet observasjoner kan oppgis som en ratio mellom to datakilder for å kunne si noe om kvaliteten på den ene kilden i forhold til den andre. Hvis for eksempel datakilde A gir oss 100 observasjoner over en tidsperiode og datakilde B gir oss 50 observasjoner vil forholdstallet være 2:1. Merk her at man snakker om det faktiske antallet observasjoner og ikke det teoretiske antallet.

2.7.1.2 OPPLØSING

Oppløsningen antyder hvor stor målenøyaktigheten er. Umiddelbart vil det være et ønske at oppløsningen på data er så høy som mulig, men i mange tilfeller er det ikke hensiktsmessig verken i forhold til ressursbruk eller bruksområde. I andre tilfeller tillater ikke målemetoden høyere oppløsning. Oppløsningen kan også si noe om hvilke forhåndsregler man bør ta ved bruk av datamaterialet. Hvis oppløsningen er veldig lav, kan datamaterialet allerede ha vært utsatt for avrunding noe som kan gi store feilkilder ved bruk. Det bør være mål at dataforbrukeren får den oppløsningen på datamaterialet som er mest mulig hensiktsmessig, og at avrunding er opp til brukeren.

Objektive mål for oppløsningen gir i for seg ingen mening uten at det målet relateres i forhold til en annen datakilde, eller til en ønsket oppløsning. Man kunne tenke seg at mål kunne være antall desimaler bak komma for en gitt enhet i forhold til det samme for en annen datakilde.

I denne oppgaven bruker man begrepet oppløsning om antall desimalers nøyaktighet på en måling. En annen mulig måte å se det er som frekvensen av målinger over for eksempel en strekning eller tid. Dette betyr at denne parameteren er tilpasset bruksområde til dataene. Hvis man for eksempel ønsket å se på kjøreprofilen over hele strekningen, hadde det vært mer hensiktsmessig å snakke om oppløsning over strekning eller tid som antall målinger. Forskjellen mellom de to datakildene med hensyn på denne parameteren vil komme tydeligere fram i kapittel 3, der man diskuterer kildene mer i detalj.

2.7.1.3 NØYAKTIGHET

Nøyaktigheten i datamaterialet gjenspeiler i hvor stor grad man er nødt til korrigere verdier som er feilregistreringer. Hvilke verdier man korrigerer er i stor grad avhengig av bruksområdet til datamaterialet. Hvis man for eksempel ønsker å beregne et gjennomsnittlig tall basert på et stort antall observasjoner er det i mange tilfeller hensiktsmessig å fjerne ekstreme verdier. Ønsker man derimot å granske årsaker til avvik fra gjennomsnitt kan disse ekstreme verdiene være svært viktige.

Et annet moment ved å bestemme nøyaktigheten er at det kan være vanskelig å identifisere hvilke verdier som er feilregistreringer og hvilke som er ekstreme verdier, men korrekt registrert. Denne parameteren kan derfor betegnes som både subjektiv og objektiv.

Et typisk mål for nøyaktighet i prosent, kan være:

$$(1.1) N = \left(1 - \frac{\text{KorrigerteVerdier}}{\text{AntallVerdier}}\right) * 100$$

Denne parameteren gjenspeiler i stor grad påliteligheten av datamaterialet.

2.7.1.4 FULLSTENDIGHET

Denne parameteren sier noe om hvor godt datamaterialet samsvarer med det teoretiske antallet observasjoner. Ved enhver måling vil man oppleve at verdier ikke blir registrert på grunn av feil i registrering, måleutstyr etc. Fullstendigheten gjenspeiler da avviket mellom det faktisk antallet observasjoner og det teoretiske antallet observasjoner.

Et kvantitativt mål for fullstendigheten kan være:

$$(1.2) F = \frac{\text{AntallObservasjoner}}{\text{TeoretiskAntallObservasjoner}} * 100$$

2.7.1.5 ENSARTETHET

Ensartetheten i datamaterialet antyder om man har dataene har gjennomgående samme format eller om samme type data kan være målt i forskjellige enheter, eller på forskjellig måte, men blir behandlet som samme type data. Det bør være et mål at dataene er mest mulig ensartet og kategorisert i forhold til format og målemetode.

2.7.1.6 TIDSRIKTIGHET

Med tidsriktighet menes å antyde to komponenter ved datakilden:

- Om man har tilgang til oppdaterte data: at de er relevante for den tidsrammen man ønsker å undersøke.
- Referansetid: Om alle målinger er gjort i henhold til en entydig referansetid og om denne referansetiden er sammenlignbar med andre målinger av samme type.

Denne parameterens viktighet er nødvendigvis avhengig av bruksområdet for dataene, men kan vurderes rent objektivt. Ønsker man for eksempel å beregne ankomstforsinkelsen ved hjelp av en datakilde på den måten at man sammenligner med en annen datakilde (for eksempel publikumsrute), må begge tidsreferansene være synkroniserte.

2.7.2 SUBJEKTIVE PARAMETERE

For de subjektive parametrene som er listet her, er det vanskelig å utvikle objektive mål; målingene må nødvendigvis basere seg på en viss grad av vurdering for å kunne utvikle en verdi. Videre er det vanskelig å måle disse parametrene uten å ta hensyn til bruksområdet.

Siden slike parametre ikke kan måles som for eksempel nøyaktigheten, er man nødt til å trekke inn personlige vurderinger. Tayi & Ballou [38] sier at en mulig måte å måle hvor godt en organisasjon oppfyller de subjektive datakvalitetsdimensjonene er ved hjelp av et spørreskjema. Mange organisasjoner i USA inne helseomsorg, finans, og forbruksvarer har brukt et spørreskjema for å vurdere disse subjektive dimensjonene av datakvalitet.

I denne oppgaven blir det foreslått to parametere som kan brukes som indikatorer på den subjektive dimensjonen av datakvaliteten.; tolkbarhet og tilgjengelighet.

2.7.2.1 TOLKBARHET

Tolkbarheten til datamaterialet er ment å si noe om følgende to ting:

- Egnethet; er dataene egnet til sitt bruksområde, har de et forståelig og universalt sammenlignbart format.
- Manipulerbarhet; er dataene egnet for flere bruksområder.

Denne parameteren kan også til en viss grad si noen om validiteten av målingene; om dataene er gyldige.

2.7.2.2 TILGJENGELIGHET

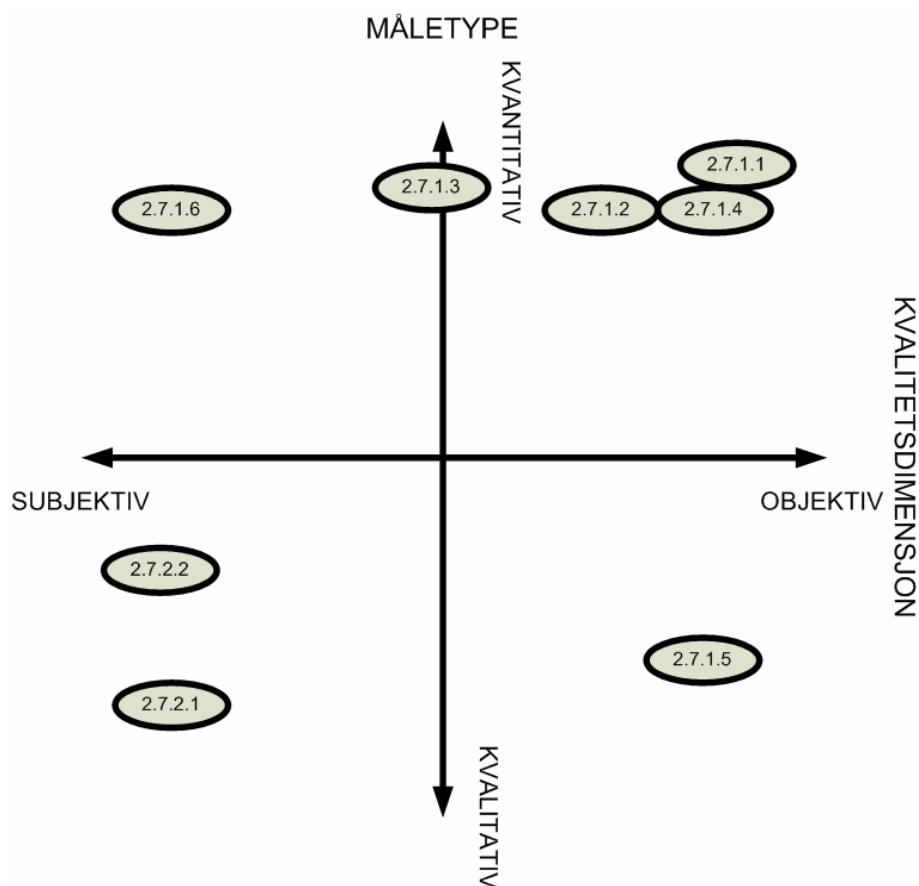
Tilgjengeligheten sier rett og slett noe om hvor tilgjengelige dataene er for dataforbrukeren. Noen spørsmål man kunne stille seg for å estimere denne parameteren ville være:

- Hvor lett tilgjengelig er dataene?
- Hvor stor frihetsgrad har man i å få tilgang til akkurat de dataene man ønsker?
- Innebærer det en stor grad av prosessering å kunne benytte dataene?

2.7.3 MODELL FOR DATAKVALITETSPARAMETERE

Figur 5 viser sammenhengen mellom måletype; kvantitativ eller kvalitativ; og parameteren befinner seg i den subjektive eller den objektive dimensjonen av datakvalitet. Nummereringen i ellipsene refererer til avsnittsnummer i denne oppgaven.

Hensikten med å prøve å lage en slik modell er å antyde at man kan ha parametere som kan måles rent kvantitativt, men som for å gi mening nødvendigvis må relateres til et annet mål. Det må med andre ord gjøres en subjektiv vurdering av den kvantitativt målte parameteren for at den skal gi noen mening.



Figur 5 Datakvalitetsdimensjoner og måletype.

2.8 SAMMENDRAG

Det er i dette kapitlet bestrebet å skape et teoretisk fundament for å kunne si noe om kvaliteten på de datakilder som danner en basis for den informasjonen beslutninger blir tatt på grunnlag av. Kapitlet kan noe grovt sies å ha fulgt en tredeling, der den første delen av kapitlet var en begrepsavklaring, den andre delen en kort oversikt over forskningsområdet *datakvalitet* og den tredje, og siste, delen et forsøk på å definere noen ulike parametere for å måle datakvaliteten.

Denne oppgaven adopterer en prosessanalogi mellom produksjonen av data og produksjonen av et fysisk produkt. En slik analogi er virkningsfull for å dra en parallell til tradisjonell styring av produktkvaliteten; et område som har eksistert i mange år. På den andre siden åpnet en slik analogi opp døren til store mangler ved å gjøre en slik sammenligning. Rådata er alltid resultatet av målinger som enten kan enten være kvantitative eller kvalitative, mens et råmateriale derimot kan være urørt i det produksjonsprosessen starter. Videre kan rådata brukes flere ganger til å produsere flere ulike informasjonsprodukter, i motsetning til råmaterialet som blir forringet ved prosessering. Dette leder oss til den slutning at hvis man skal snakke om kvaliteten på dataene, er det vanskelig ikke å ta hensyn til bruksområdet.

På samme måte som for kvaliteten av et produkt, har også kvaliteten av data ulike dimensjoner. I tidligere forskning på datakvalitet har det vært vanlig å skille mellom den subjektive og den objektive dimensjonen. Det hersker ingen generell konsensus rundt hvilke parametere man kan bruke for å beskrive datakvaliteten, men basert på en gjennomgang av litteraturen ble det i dette kapitlet foreslått følgende 8: *antall observasjoner, oppløsning,*

nøyaktighet, fullstendighet, ensartethet, tidsriktighet, tolkbarhet og tilgjengelighet. De 6 første ble definert som objektive, mens de 2 siste som subjektive.

Beslutningsprosessen i en organisasjon er ofte kompleks og muligens ikke så strukturert og rasjonell som man ønsker å tro. Likevel vil man kunne komme langt med å etterstrebe to ting: 1. Et mest mulig feilfritt datamateriale. 2. At dataene er hensiktsmessig til sitt bruk. Punkt 1 kan man forsøke å oppnå uten å ta hensyn dataens bruksområde. Punkt 2 derimot krever at man undersøker bruksområdet og ser om dataene er hensiktsmessige. I kapittel 4 belyses nettopp dette aspektet av datakvaliteten ved å bruke dataene til å beskrive en togstrekning analytisk. Sammen med de definerte parametrene for vurdering av datakvaliteten er håpet at denne anvendelsen bedre kan forklare det totale aspektet ved kvaliteten på data.

Det neste kapitlet beskriver metodologien, samt en oversikt over de ulike datakildene brukt i denne oppgaven.

3 METODOLOGI OG DATAKILDER

Dette kapitlet beskriver den valgte forskningsmetodikken for denne masteroppgaven og motivasjon for valgte metode. En oversikt over de ulike datakildene brukt blir også presentert.

3.1 METODEVALG

Det er vanlig å skille mellom to overordnede typer metoder, kvalitative og kvantitative metoder. I følge Holme & Solvang [39] finnes det ikke noe absolutt skille mellom kvalitative og kvantitative metoder: *”De er alle arbeidsredskaper som i skiftende grad tar i bruk de ulike metodeprinsippene: det analytiske prinsippet, systemprinsippet og aktørprinsippet”* [39, s 33]. Videre nevnes det at grovt og enkelt kan forskjellen mellom de to oppsummeres som at den kvantitative metoden omformer tall og data til mengestørrelser, mens innenfor kvalitative metoder er det forskerens forståelse eller tolkning av informasjonen som står i forgrunnen, for eksempel tolkning av meningsrammer, motiver, sosiale prosesser eller sammenhenger. I den kvalitative metoden bør, eller kan resultatet ikke tallfestes.

Denne oppgaven bruker i all hovedsak kvantitativ metode for å besvare oppgavens problemstilling på best mulig måte. Grunnen til at en slik framgangsmåte er valgt er at oppgaven prøver å bruke mange observasjoner til å utvikle et generelt mål, og ikke et typisk casestudie der man tar utgangspunkt i én observasjon, eller ett case. Ett casestudium kan kjennetegnes på følgende måte: få enheter av observasjoner, mange variabler og hvor man forsøker å beskrive et fenomen i dets kontekst eller som et komplekst sosialt fenomen [40]. Dette kan beskrives som det motsatte av den tilnærmingen som blir brukt i denne oppgaven.

I tillegg til oppgavens problemstilling, og det overnevnte, har to andre faktorer for metodevalg også spilt inn: 1. oppgavens begrensninger og 2. forfatterens egen bakgrunn. Innenfor de tids- og ressursrammer som denne oppgaven opererer ville det vært vanskelig å kunne bruke en kvalitativ tilnæringsmåte. En slik tilnæringsmåte ville krevd et lengre tidsperspektiv på oppgaven, samt en nærmere tilknytning til NSB og bedre kjennskap til bedriften og dens interessenter. Videre har forfatterens egen bakgrunn og interesser gjort at en kvantitativ tilnærming er lettere å rettfærdiggjøre.

Likevel må det påpekes at et kvalitativt studie på mange måter kunne vist seg svært nyttig i forhold til å for eksempel undersøke faktorer for stasjonsopphold mer i dybden. Å belyse en problemstilling fra flest mulig vinkler og utvikle egne data kunne ha åpnet opp nye perspektiver både for NSB Drift og forfatteren.

3.1.1 FEILKILDER KNYTTET TIL METODEVALG

Som nevnt i forrige avsnitt følger denne oppgaven en kvantitativ tilnærming. I det ligger det selvfølgelig både fordeler og ulemper i denne valgte framgangsmåten. En ting kan være at man stoler blindt på datagrunnlaget ut fra en tankegang om at de representerer en ”objektiv sannhet”. Det er viktig å være klar over at selv om man forsøker å holde en objektiv avstand til datagrunnlaget og dermed prøver å utvikle objektive parametere, vil de til slutt alltid utsettes for en viss grad av tolkning i den form de presenteres. Data er sjelden feilfrie, og det kan genereres nye feil ved bearbeiding av dataene, som diskutert i det teoretiske grunnlaget til oppgaven.

Videre kan det at man skaper en avstand mellom empiri og observatør føre til et manglende erfaringsgrunnlag. Avstanden hjelper kanskje observatøren til å se ting fra andre perspektiver, men i mange tilfeller er det svært viktig at erfaringsdata blir brukt sammen med kvantitative data slik at man lettere kan unngå feiltolkninger og at det fokuseres på feil ting.

3.2 DATAINNSAMLING

Dette avsnittet beskriver datagrunnlaget som er blitt brukt til henholdsvis det teoretiske fundamentet og det den empiriske analysen.

3.2.1 TEORI

Det teoretisk fundamentet for denne oppgaven er hentet fra i hovedsak bøker med fokus på bakgrunnskunnskap som statistiske analyser og programmer, og artikler med fokus på datakvalitet. Alt dette materialet er gjort tilgjengelig ved hjelp av Universitetsbiblioteket i Trondheim og ulike artikkeldatabaser som NTNU abonnerer på.

Litteratursøket har startet med at man har søkt i databaser på begrepet "data quality". Videre har man identifisert 4-5 artikler som "kjerneartikler". De respektive referansene i kjerneartiklene har blitt brukt til å til å skape et nettverk av artikler og bøker, som til syvende og sist har blitt brukt som grunnlag for litteraturstudiet i kapittel 2. For en nærmere beskrivelse av litteraturen benyttet i arbeidet med denne oppgaven henvises det til referanselisten i kapittel 7.

3.2.2 EMPIRI

Denne oppgaven kommer til å fokusere på data i jernbanedrift. Hvis man skal bruke data i noen som helst form og sammenheng, står man vanligvis ovenfor to strategier: 1. Samle inn nye data 2. Bruke eksisterende data som allerede er samlet inn. I arbeidet med denne oppgaven er det valgt å følge strategi nummer 2. Motivasjonen for dette har først og fremst vært for å se på hva slags data som finnes tilgjengelig, men også å kunne komme med betraktninger i forhold til kvaliteten på de allerede eksisterende datakildene. Man kan dermed si med bakgrunn i Figur 3 at forfatteren har hatt både rollen som databasebestyrer og dataforbruker i arbeidet med datakildene. Datakildene som er blitt brukt, og forholdet mellom dem, blir nærmere beskrevet i de påfølgende avsnitt.

Da denne oppgaven har fulgt en kvantitativ tilnærming, har det ikke vært naturlig å drive formelle intervjuer av ressurspersoner i jernbanemiljøet. Det er likevel ikke til å unngå, og høyst nødvendig, at man har gjennom samtaler med både veileder og kontaktpersoner både i NSB og Jernbaneverket, tilegnet seg en god del kunnskap om for eksempel togframføringen på den aktuelle strekningen og dataflyten.

3.2.2.1 DATAKILDER

I dette avsnittet blir datakildene som det empiriske datagrunnlaget er hentet fra diskutert og presentert utfyllende. Det gjøres også et forsøk på å se sammenhengen mellom de ulike datakildene og hvordan dataflyten er.

TELOC

TELOC er et produkt levert av HasslerRail AG Bern som elektronisk⁹ lagrer ulike datasignaler fra et tog. Disse datasignalene kan evalueres i etterkant ved behov. En god sammenligning av hva dette produktet faktisk er kan være et fly's såkalte *black box*, eller ferdskriver på norsk. De metrikkene som er mulig å hente fra TELOC med den gjeldende konfigurasjonen¹⁰ er: hastighet[km/t], distanse(målt i kilometer fra start registrering) og tidspunkt relativt til dens egen klokke. Oppløsingen er henholdsvis 2 desimaler, 5 desimaler og 2 desimaler (hundredeler). Registrering skjer ikke med faste intervaller og kan variere mellom hver tidel, hvis man har endring i hastigheten, og hvert halve minutt, hvis toget holder konstant hastighet. Et forsøk på å måle et gjennomsnittlig registreringsintervall ga det resultat at TELOC registrerer data i gjennomsnitt i underkant av hvert 2. sekund.

Data fra TELOC hentes under normale omstendigheter ut i forbindelse med periodisk vedlikehold på et togsett. For at NSB skal få tilgang til disse dataene må de bestilles fra Mantena som er ansvarlige når et togsett er inne til vedlikehold. Med andre ord er dette et datamateriale som befinner seg utenfor NSBs regulære dataflyt.

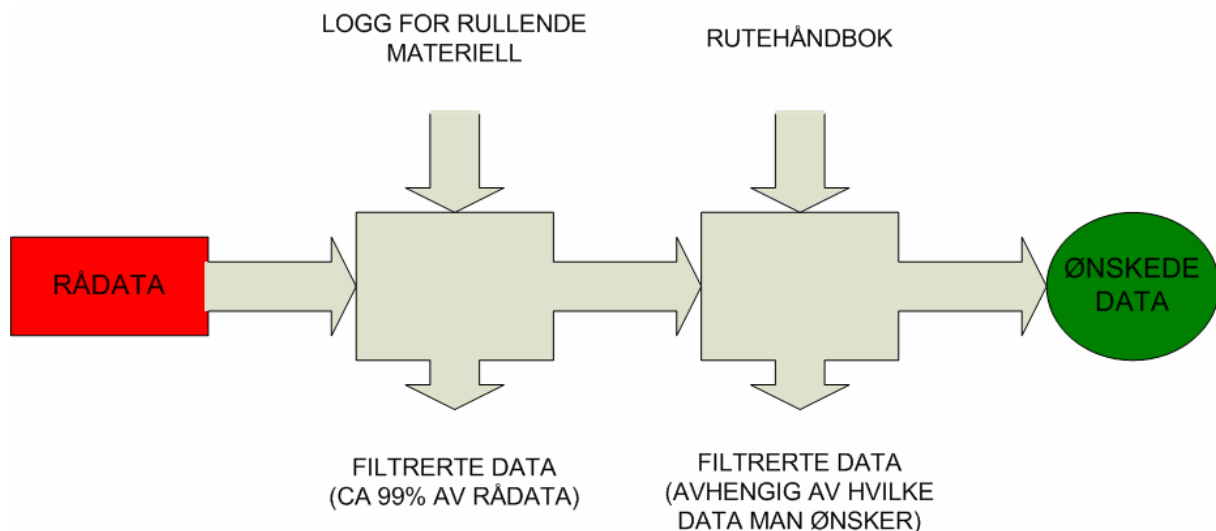
Som tidligere nevnt er altså TELOC noe som befinner seg fysisk fastmontert på et togsett, i dette tilfellet 72-33, som er togets sett ID og betegner henholdsvis togtype og identifikasjonsnummer. Det betyr at man kan ikke be om å få tilgang til kjørelaggen for bestemte tognummer¹¹. Man må derimot plukke ut de dataene man ønsker fordi en slik kjørelagge inneholder alle bevegelser for dette togsettet. Framgangsmåten som er benyttet for å trekke ut de datagrunnlaget er å sammenholde *logg for rullende materiell* for den aktuelle tidsperioden med det aktuelle togsettets kjørelagge. Dette begrenser i stor grad friheten for å hente ut de data man ønsker med mindre man er interessert i å analysere et togsetts bevegelser. Et eksempel på hvordan en slik logg for rullende materiell ser ut er vist i vedlegg 8.1.

Kjørelaggen fra TELOC må videre behandles av en programvare, Hassler, for å kunne hente ut de ønskede data på en presentabel måte. I forbindelse med denne oppgaven har rådata blitt eksportert til Microsoft Excel ved hjelp av programvaren Hassler. Dataene har videre blitt prosessert for å kunne isolere kun stopp og start i togets bevegelser, siden dette er hva man var interessert i denne forbindelse. Med gjennomsnittlig registrering hvert 2. sekund blir det en stor mengde uinteressante data man må filtrere bort, enten automatisk eller manuelt, for å fange de interessante dataene.

⁹ Man snakker i denne sammenheng om en versjon av produktet som lagrer informasjonen elektronisk og ikke en analog versjon som registrerer tid og hastighet på en papirrull. Se for eksempel kapittel 3 i utredningen etter Åsta ulykken tilgjengelig her: http://odin.dep.no/jd/norsk/dok/andre_dok/nou/012001-020007/hov003-bn.html der en analog versjon av TELOC ble brukt til å beskrive hendelsesforløpet.

¹⁰ TELOC har mulighet til å registrere andre signaler enn det som blir registrert per dags dato, men det krever i tilfelle en endring av det nåværende oppsettet og konfigurasjonen. Disse eventuelle endringene kan igjen ha implikasjoner for andre områder siden det eksisterer strenge rutiner i forhold til sikkerhet hvis man endrer et slikt oppsett.

¹¹ Med tognummer menes det nummeret som toget har i publikumsruta. I denne oppgaven har man ønsket å se på oddetallstognumrene 1605-1641. Det betyr at man ønsker i utgangspunktet data som er sortert i henhold til tognummer, ikke togets sett ID.



Figur 6 Prosess for å ekstrahere data fra TELOC.

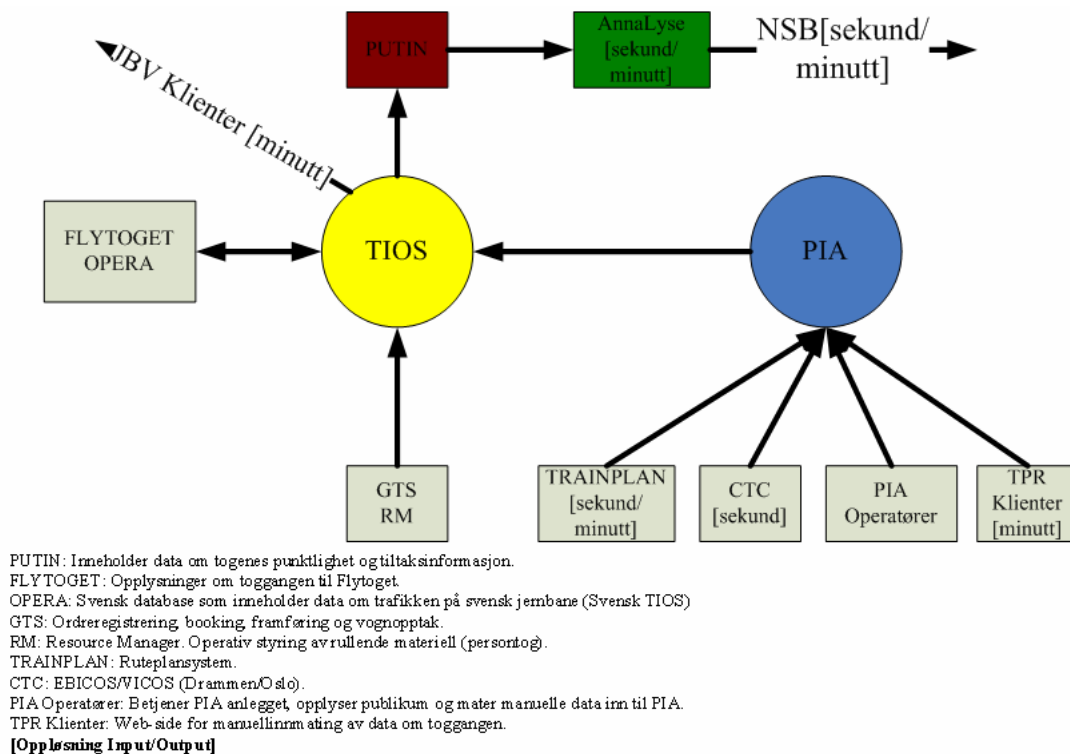
Figur 6 viser skjematisk hvordan prosessen for å hente ut data fra TELOC for den aktuelle strekningen har foregått. På ett døgns målinger må om lag 99 % uinteressante og blanke registreringer filtreres bort. Logg for rullende materiell hjelper å plukke ut de dagene hvor togsettet kjørte som de interessante tognumrene. Videre må man identifisere hvor de ulike registreringene er gjort. Dette gjøres ved å sammenholde avstand fra stasjon til stasjon i *rutehåndboken* (nærmere beskrevet senere i dette kapitlet) med avstandsmåleren i TELOC. Avhengig av hvilke tog man vil ha for den dagen blir et visst antall registreringer filtrert bort. Man står da igjen med en relativt liten brøkdel av rådataene igjen. Som eksempel kan det nevnes at rådataene som ble brukt i forbindelse med denne oppgaven inneholdt data fra perioden 4.8.2005 til 22.9.2005, der togsettet hadde kjørt på mange ulike strekninger og som mange ulike tognummer, mens det kun var interessant å se på tog som hadde gått på strekningen Kongsberg – Eidsvoll i perioden 22.8.2005 – 5.9.2005.

AnnaLyse

Det som refereres til som *AnnaLyse* i denne oppgaven, er NSBs nye (lansert sommeren 2005) registreringssystem som henter data fra Jernbaneverkets (JBV)s database TIOS (Trafikk Informasjon og OppfølgingsSystem). Oppløsingen er ett sekund på de dataene som stammer fra signalanlegget og ett minutt på manuelt registret data. Alle data fra TIOS som er blitt brukt i denne oppgaven er hentet fra signalanlegget med en oppløsning på ett sekund. I JBV's rapportering som henter data fra TIOS, rundes alle data av til hele minutt. Dette gir en ny mulighet for NSB å jobbe med for eksempel kontinuerlig punktlighetsovervåking med en høyere oppløsning enn tidligere.

Som nevnt tidligere baserer dataene i denne oppgaven seg på signalanlegget ved hver stasjon noe som igjen betyr at målingene er fra når toget er i bevegelse; når toget kjører inn og ut av en stasjon. Konsekvensen dette får for bruk av dataene er at man kan ikke snakke om for eksempel oppholdstid i den forstand at toget står i ro, men derimot oppholdstid med tanke på hvor lang tid toget belegger et spor på den aktuelle stasjonen.

AnnaLyse er igjen en del av et større system av datamateriale som stammer fra både NSB og Jernbaneverket. Figur 7 viser et oversiktskart over dataflyten som ligger til grunn for AnnaLyse-systemet.



Figur 7 Datakilder, flyt, sammenhenger og oppløsning.

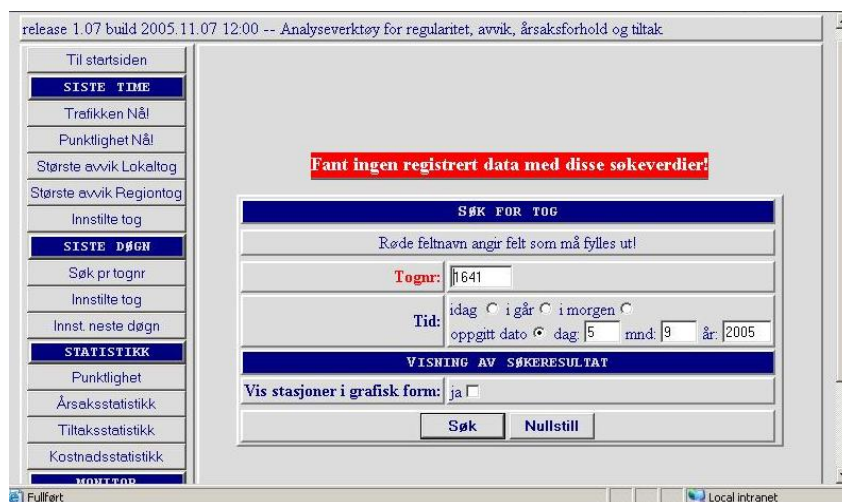
En viktig ting å legge merke til i Figur 7 er hvordan det både kommer inn målinger gjort på sekunder, CTC (Drammen/Oslo området), og minutter, TPR klienter (alle andre stasjoner). Man vil dermed få ut data fra AnnaLyse systemet med både sekund og minutt oppløsning selv om det er sett på en strekning som ligger innefor signalanleggsområdet.

Signalene som kommer inn fra signalanlegget (CTC) til PIA blir sendt når toget belegger en ny blokkstrekning; for eksempel når det kjører inn på en stasjon vil det belegge den blokkstrekningen; stasjonen; og når det kjører ut vil det sendes et signal om at det belegger en ny blokkstrekning. En opplagt feilkilde i datamaterialet vil være om signalanlegget opererer med en korrekt klokke. Hvis for eksempel tiden på signalanlegget er feiljustert med ett minutt i forhold til "virkelig tid", vil man bestandig rapportere ett minutts avvik. Når man opererer med en oppløsning på sekundnivå kan dette være til dels store feilkilder i datamaterialet. Et annet moment ved registreringen av data fra signalanlegget, er forsinkelsen mellom det tidspunktet toget passerer et punkt og til det er registrert i databasen. Nå er det ikke logisk å tenke seg at en slik forsinkelse vil være av særlig av stor grad, men det bør nevnes, særlig med tanke på at man ikke kjenner den bestemte algoritmen bak målingene.

Informasjonen fra PIA – databasen vil så bli tatt videre til jernbaneverkets database TIOS. TIOS gir klienter i JBV sitt nett, mulighet til å hente ut tabeller, oversiktskart og statistikk for blant annet punktligheten. Data som blir hentet ut fra denne databasen kommer ut i hele minutter og følger en regel om at sekunder strykes. Dette gjelder både om det måles på ankomst og avgang. Med andre ord vil to tog som ankommer henholdsvis 54 sekunder (-54 i AnnaLyse) og 63 sekunder (-63 i AnnaLyse) for tidlig, begge fremstå som om de kom 1 minutt for tidlig på grunn av at man stryker sekundene. Avvikene måles i forhold til publikumsruten.

Data fra TIOS blir tatt videre til NSBs database, PUTIN, der man kan hente ut data ved hjelp av AnnaLyse systemet med sekundoppløsning. Dette er altså de samme data som ligger i TIOS, men forskjellen er at AnnaLyse ikke stryker sekunder.

I praksis er AnnaLyse en database med et brukergrensesnitt ut mot de ansatte med brukerrettigheter i NSB. Systemet kan gi informasjon om alle stasjoner med signalanlegg som rapporterer inn til JBV og de som rapporterer manuelt. Et typisk skjermbilde for en bruker vil se ut som vist i Figur 8.



Figur 8 Brukergrensesnitt AnnaLyse.

AnnaLyse kan altså, som man ser av Figur 8, gi oss informasjon om punktlighet, årsaksstatistikk, tiltaksstatistikk og kostnadsstatistikk både i grafisk- og tabellform. I denne oppgaven blir det brukt systemets mulighet til å gi punktlighetsstatistikk i tabellform. For å få tilgang til denne informasjonen må man da spesifisere både dato og tognummer/stasjon avhengig av hvilken variabel man er interessert i, noe som er svært lite hensiktsmessig hvis man ønsker å se på en lengre tidsperiode eller et stort antall tog.

STASJON	RUTETID ANKOMST	ESTIMERT ANKOMST	FAKTIISK ANKOMST	AVVIK SEK	RUTETID AVGANG	ESTIMERT AVGANG	FAKTIISK AVGANG	AVVIK SEK
Kongsberg	22:53				22:53		22:53	2
Kapermoen	22:55				22:55			
Skollenborg	22:58		22:56	-64	22:58		22:57	-15
Grosvold	23:00				23:00			
Teigen	23:02				23:02			
Krekling	23:04		23:01	-143	23:04		23:02	-111
Darbu	23:07		23:05	-81	23:07		23:06	-1
Flesaker	23:10				23:10			
Vestfossen	23:13		23:10	-163	23:15		23:15	6

Figur 9 Variabler punktlighetsstatistikk.

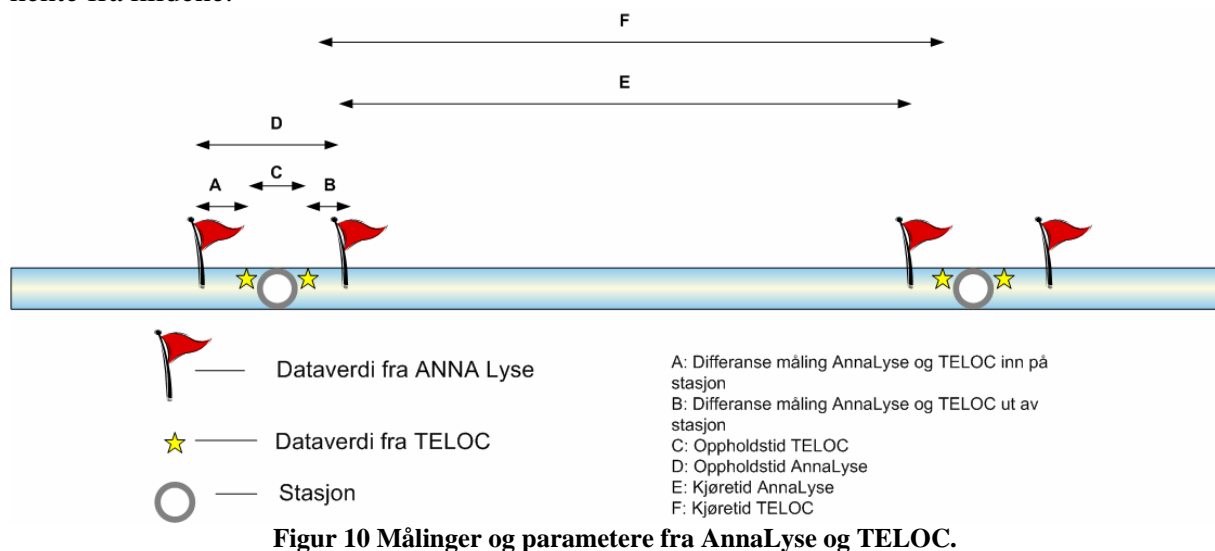
Etter å ha spesifisert tognummer og dato vil systemet produsere følgende informasjon med hensyn på hver stasjon det aktuelle toget passerer eller stopper på: faktisk ankomst[minutter], avvik ankomst[sekunder], faktisk avgang[minutter] og avvik avgang[sekunder]. Rutetid

ankomst og avgang er hentet fra publikumsruten. Et eksempel på hvordan dette ser ut for dataforbrukeren er vist i Figur 9.

Systemet har i tillegg til de funksjonene nevnt over, en eksporteringsfunksjon som gir mulighet for å eksportere til Microsoft Excel, men den var ikke tjenlig for denne oppgavens formål. Dataene fra AnnaLyse måtte derimot behandles, sorteres og bearbejdes i stor grad manuelt i etterkant for å kunne hente ut de ønskede metrikkene: kjøretid, oppholdstid og ankomstforsinkelse. Systemet fremstår per dags dato som lite brukervennlig med tanke på å hente ut slike data over en viss tidsperiode og strekning, men denne vurderingen er en subjektiv vurdering avhengig hvilke data man er interessert i.

SAMMENHENG MELLOM TELOC OG AnnaLyse

Etter denne korte presentasjonen av de to sentrale datakildene i denne oppgaven er det på sin plass å si noe om sammenhengene og forskjellene, mellom de ulike parametrene man kan hente fra kildene.



Figur 10 viser hvilke målinger man kan hente ut fra AnnaLyse og TELOC, og hvilke metrikker man dermed kan utvikle på grunnlag av disse målingene. AnnaLyse bruker som nevnt tidligere data fra når et tog kjører inn og ut av en stasjon. TELOC sine målinger indikerer det øyeblikket toget står helt i ro og det øyeblikket det er i bevegelse igjen¹².

Det er også mulig å beregne parametrene for tidsdifferansen mellom det øyeblikket et tog kjører inn på en stasjon og til det står helt i ro, og på samme måte når det kjører ut av en stasjon (jfr. A og B i Figur 10). Men de eventuelle målingene man da bruker som grunnlag, vil være avhengig av at signalanlegget og TELOC har synkronisert tid. Man kan tenke seg en situasjon der kjøreløkken viser at toget stopper ved en stasjon 12:03:56. Data fra AnnaLyse viser at det samme toget ankommer stasjonen 12:03:43. Dette skulle indikere at differansen mellom toget ankommer stasjonen og til det faktisk står i ro, var $12:03:56 - 12:03:43 = 13 \text{ sekunder}$, men det forutsetter at begge klokkene, både i TELOC og signalanlegget, er synkronisert. Hvis ikke vil denne målingen være verdiløs på grunn av

¹² Kjøreløkken fra TELOC inneholder også mange registreringer fra når toget er i bevegelse, men for å se på parametrene kjøretid, oppholdstid og ankomstforsinkelse, er man kun interessert i det øyeblikket et tog stopper og starter på en stasjon.

avvik i klokken. Det som derimot kan beregnes med nøyaktighet fra AnnaLyse og TELOC, er parametrene A + B fra Figur 10 siden man kan bruke absolutte tidsmålinger.

I tillegg til de parametrene som er vist i Figur 10, kan man også hente ut data for ankomstforsinkelsen fra de to datakildene. Nok en gang støter man på problemet med tidssynkronisering ved sammenligning av de to ankomstforsinkelsene, siden de to kildene ikke nødvendigvis følger samme tidsreferanse.

PASSASJERTELLINGER

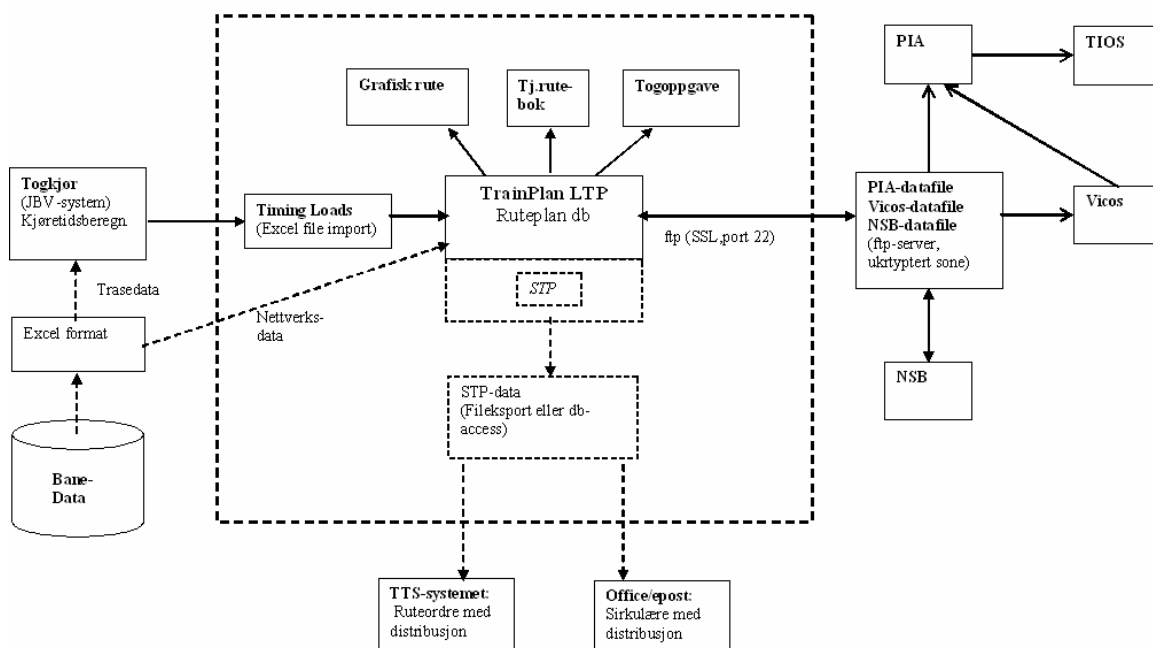
Passasjertellinger blir gjort to ganger i året, vår og høst, og presenteres som et maksimalt antall påstigende, avstigende og passasjerer ombord for tog og stasjoner. Målingene blir gjort for en hverdag, en lørdag og en søndag. Et eksempel på hvordan dette datamaterialet ser ut i praksis for strekningen for pendelen Kongsberg – Eidsvoll er vist i vedlegg 8.3.

Disse passasjertellingene har den svakhet at de er kun et anslag for passasjermengden. De gir ingen eksakte verdier for passasjertallet noe som gjør at de kan både sies å ha lav tidsriktighet og oppløsning. I tillegg har man ikke datagrunnlag for spesifikke tidspunkt. Målingene blir, som nevnt, kun utført som overslag hver vår og høst. Dette fører igjen til at de er lite egnet til annet enn å beskrive passasjerstrømmen på generelt basis; ikke stilles i relasjon til andre variabler for å se sammenhenger.

TRAINPLAN

Trainplan er et it-verktøy som holder på rutedata og brukes blant annet til å utvikle publikumsruta: den rutetabellen med tidspunkt man som passasjer med NSB forholder seg til. Trainplan er en del av et større system for ruteplanlegging i JBV. Figur 11 viser hvordan Trainplan henger sammen med de andre komponentene i systemet og PIA.

Systemkart_TP_JBV_20051129 : Stiplet ramme rundt RPS-systemet med TrainPlan sentralt.



Figur 11 Ruteplansystemet og Trainplan (Kilde: Jernbaneverket).

Systemet avgrenset av den stiplede linjen i Figur 11 indikerer det som benevnes "RPS" i Figur 7. På grunnlag av Trainplan lager man også Tjenesterutehåndboken som benyttes av lokførere, konduktører, togledelse og togekspeditører. I denne håndboken finnes det informasjon det blant annet informasjon om følgende ting: om det er enkeltspor eller dobbeltspor der toget kjører, om togekspeditør skal gi signal "kjøretillatelse" til toget, hvilke typer stillverk det er på stasjoner på ikke fjernstyrt strekning, hvilken type sikringsystem det er på de ulike strekningene, hvilket spor toget kjører på stasjonene og når stasjonen er betjent for toget.

En utskrift fra Trainplan for strekningen Kongsberg-Eidsvoll er vist i vedlegg 8.4.

RUTEHÅNDBOK

Rutehåndboka inneholder i stor grad samme type informasjon som utskriften fra Trainplan, men fordelene er at den er allment tilgjengelig hos NSB drift. Da det i denne oppgaven er brukt Rutehåndboka til å bestemme hvor i pendelen målingene i TELOC er blitt gjort ved å sammenligne kilometertallet, er problematikken ved å hente data fra Trainplan eksternt for å gjøre denne sammenligningen unngått.

En utskrift fra rutehåndboka er vist i vedlegg 8.2.

LOGG OVER RULLENDE MATERIELL

Logg over rullende materiell viser som hvilket tognummer et aktuelt togsett kjørte med hensyn på dato. Denne loggen ble brukt til å identifisere hvilke målinger fra TELOC man var interessert i.

Loggen brukt i denne oppgaven er vist i vedlegg 8.1

3.3 SAMMENDRAG

Dette kapitlet har foreslått og begrunnet en kvantitativ tilnærming til å besvare denne oppgavens problemstilling på en best mulig måte. Hovedmotivasjonen for valget av en slik metodologi har vært ønsket om å bruke et relativt stort antall målinger til å bygge gjennomsnittsmål og gjøre generelle betraktninger om datakvaliteten. Faremomentene ved en slik tilnærming, i motsetning til en kvalitativ en, kan være at man har et ensidig syn på problemstillingen og dermed ikke belyser den fra et tilfredsstillende antall vinkler.

Det empiriske datagrunnlaget i denne oppgaven stammer i hovedsak fra to kilder; TELOC og ANNALyse; begge beskrevet i dette kapitlet. Den fundamentale forskjellen mellom disse to datakildene kan sies å være to ting:

1. Etter det som ble diskutert i kapittel 2.4, kan man si at data fra TELOC kan betraktes som rådata og data fra AnnaLyse som et dataprodukt.
2. Fra NSBs ståsted er det relativt enkelt å få tilgang til data fra AnnaLyse; dette gjelder ikke data fra TELOC som ligger utenfor NSBs regulære dataflyt.

For å kunne prosessere rådata fra TELOC til et dataprodukt har det vært nødvendig å inkludere en del andre datakilder som: Rutehåndbok, Logg over rullende materiell og Trainplan. I arbeidet med denne med denne oppgaven har det vist seg at det har vært nødvendig med en viss grad av prosessering av data fra AnnaLyse for å tilpasses oppgavens

problemstilling. Dermed kan man på mange måter kalle data fra AnnaLyse rådata i denne forbindelse.

I det neste kapitlet blir de ulike datakildene presentert her benyttet for å beskrive strekningen Drammen – Eidsvoll i perioden 22.8.2005 til 5.9.2005.

4 ANALYSEDEL

Som nevnt i kapittel 2.8, vil oppgaven ved hjelp av å bruke de to datakildene TELOC og AnnaLyse til å beskrive toggangen mellom Drammen og Eidsvoll i en 15 dagers periode, å kunne utvikle et fundament for å si noe om de ulike datakildene er hensiktsmessige til det man ønsker å bruke de til.

Denne analysedelen av oppgaven er igjen delt opp i to deler: en som forsøker å beskrive toggangen mellom Drammen og Eidsvoll ved hjelp av oppholdstid, kjøretid, og ankomstforsinkelse, og en som forsøker å undersøke hvilke variabler som kan påvirke oppholdstiden.

4.1 BESKRIVELSE AV STREKNINGEN

Strekningen Drammen - Eidsvoll er om lag 120 km lang¹³. I denne oppgaven konsentrer man seg om østgående 1600 tog: det vil si alle oddetallstognummer fra 1605 til 1641. Stasjoner, tognummer og dager er gitt i Tabell 4, som i tillegg viser det teoretisk antall observasjoner for hver stasjon man kan få over den 15-dagers perioden¹⁴ det er sett på.

Stasjon	Mandag - fredag	Lørdag	Søndag
Drammen (275)	Alle tog (19,209)	Ikke 1605,1639 (17, 34)	Ikke 1605,1607,1639 (16, 32)
Brakerøya (275)	Alle tog (19,209)	Ikke 1605,1639 (17, 34)	Ikke 1605,1607,1639 (16, 32)
Lier (275)	Alle tog (19,209)	Ikke 1605,1639 (17, 34)	Ikke 1605,1607,1639 (16, 32)
Asker (275)	Alle tog (19,209)	Ikke 1605,1639 (17, 34)	Ikke 1605,1607,1639 (16, 32)
Sandvika(275)	Alle tog (19,209)	Ikke 1605,1639 (17, 34)	Ikke 1605,1607,1639 (16, 32)
Lysaker (275)	Alle tog (19,209)	Ikke 1605,1639 (17, 34)	Ikke 1605,1607,1639 (16, 32)
Skøyen (277)	Alle tog (19,209)	Ikke 1605,1639 (17, 34)	Ikke 1605,1639 (17, 34)
Nationaltheatret (277)	Alle tog (19,209)	Ikke 1605,1639 (17, 34)	Ikke 1605,1639 (17, 34)
Oslo (277)	Alle tog (19,209)	Ikke 1605,1639 (17, 34)	Ikke 1605,1639 (17, 34)
Lillestrøm (277)	Alle tog (19,209)	Ikke 1605,1639 (17, 34)	Ikke 1605,1639 (17, 34)
Leirsund (277)	Alle tog (19,209)	Ikke 1605,1639 (17, 34)	Ikke 1605,1639 (17, 34)
Frogner (277)	Alle tog (19,209)	Ikke 1605,1639 (17, 34)	Ikke 1605,1639 (17, 34)
Lindeberg (277)	Alle tog (19,209)	Ikke 1605,1639 (17, 34)	Ikke 1605,1639 (17, 34)
Kløfta (277)	Alle tog (19,209)	Ikke 1605,1639 (17, 34)	Ikke 1605,1639 (17, 34)
Gardermoen (277)	Alle tog (19,209)	Ikke 1605,1639 (17, 34)	Ikke 1605,1639 (17, 34)
Eidsvoll Verk (277)	Alle tog (19,209)	Ikke 1605,1639 (17, 34)	Ikke 1605,1639 (17, 34)
Eidsvoll (277)	Alle tog (19,209)	Ikke 1605,1639 (17, 34)	Ikke 1605,1639 (17, 34)

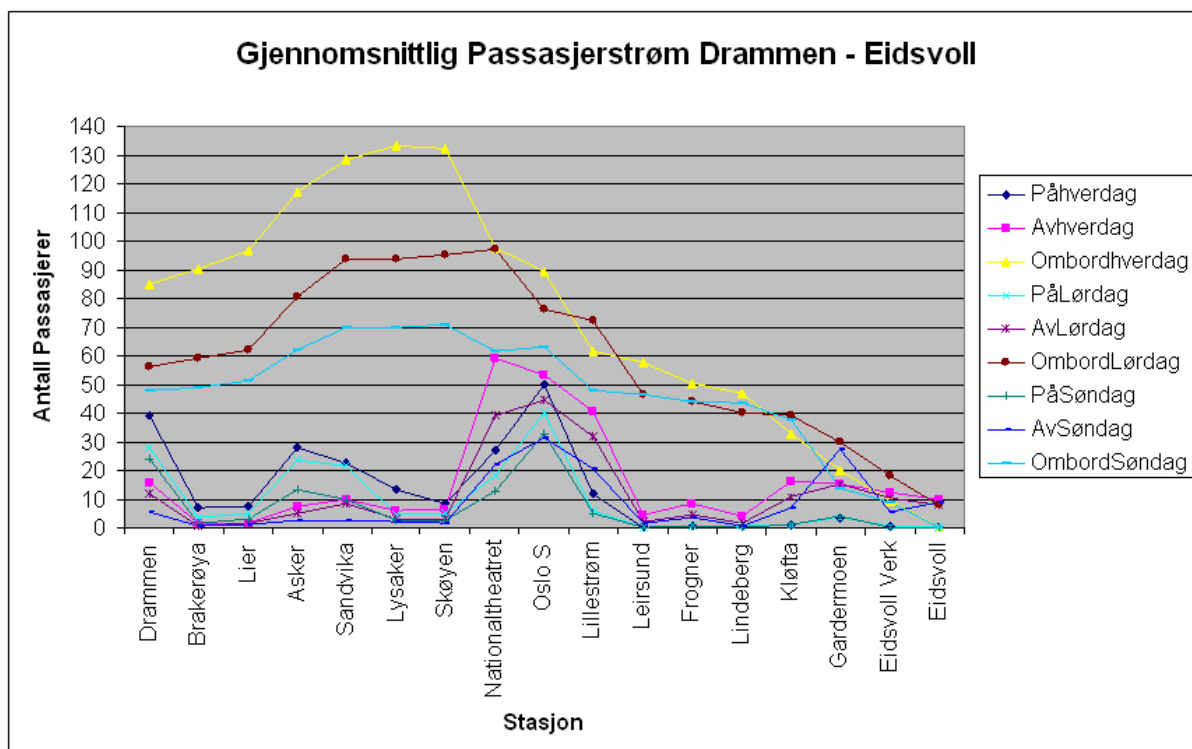
Tabell 4 Toggang Drammen – Eidsvoll; (antall tog, teoretisk antall observasjoner).

Den første kolonnen i Tabell 4 viser stasjonen og i parentes det teoretiske antallet observasjoner man kan ha i løpet av de 15 dagene. Den andre, tredje og fjerde kolonnen viser i parentes hvor mange observasjoner man har i løpet av én slik dag og det totale antallet observasjoner for denne typen dager over hele perioden.

Passasjerstrømmen, basert på passasjertellingene vist i vedlegg 8.3, over strekningen er illustrert i Figur 12. Grafene viser tre ”topper”; henholdsvis Asker, Oslo S og Gardermoen noe som kan indikere at disse er de travleste stasjonene.

¹³ Se vedlegg 8.2.

¹⁴ Perioden er fra 22.8.2005 til 5.9.2005 og inneholder 11 ukedager, to lørdager og to søndager.



Figur 12 Gjennomsnittlig passasjerstrøm over pendelen.

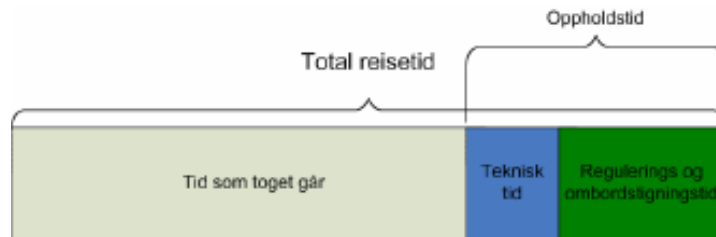
En annen ting som er verdt å merke seg er hvordan passasjerer ombord viser en klar tendens til å avta fra Skøyen og utover i pendelen. Nå er dette i seg selv ikke så oppsiktsvekkende, men kan være interessant hvis man beregner punktligheten ut fra endestasjon. Selv om toget er punktlig ved endestasjon vil passasjerene på foregående stasjoner i pendelen ikke oppleve det som punktlig.

4.1.1 DESKRIPTIV STATISTISK FREMSTILLING

I det følgende beskrives det hvordan toggangen fortoner seg *gjennomsnittlig* mellom Drammen og Eidsvoll. For å gjøre det er det benyttet data fra AnnaLyse systemet for perioden 22.8.2005 til 5.9.2005 og data fra TELOC for et togsett 72-33 som gikk på denne strekningen blant annet i denne aktuelle perioden.

4.1.1.1 GENERELT OM OPPHOLDSTIDEN FRA DE TO FORSKJELLIGE KILDENE

Det vil være på sin plass å si noe generelt om hvilke ulike faktorer som inngår i det som kan betegnes oppholdstid. Heinz [41] gjorde i 2000 et relativt omfattende studie av reisendes av- og påstigningstider på tog. Der definerte hun den totale reisetiden for tog bestående av tid som toget går, teknisk tid og regulerings- og ombordstigningstid, der de to siste komponentene kan danne det som kan kalles *oppholdstid*. Sammenhengen mellom de forskjellige komponentene er vist i Figur 13.

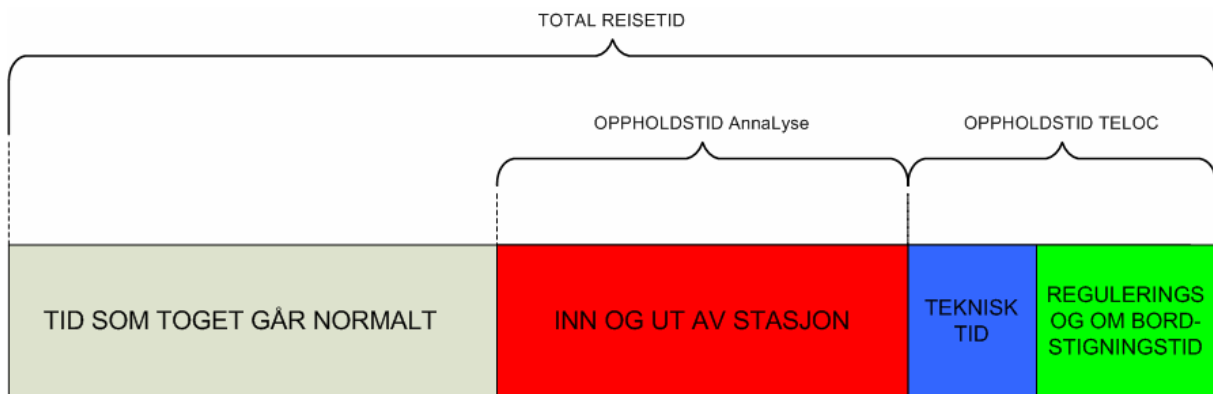


Figur 13 Total reisetid og oppholdstid.

Denne oppholdstiden kan i følge Heinz [41] igjen deles opp i 8 forskjellige komponenter:

1. Tog stopper
2. Kontrollprosedyre (kan for eksempel være en kontroll for om hele toget er inne på plattformen)
3. Dørene åpner
4. Av- og påstigning
5. Eventuell ventetid før avgang i henhold til rutetabellen
6. Kontroll prosedyre (for eksempel å forsikre seg om at ingen dører er blokkert)
7. Dørene stenger
8. Endelig klarering og avgang

Oppholdstiden slik som den framkommer fra TELOC er den som omfatter alle 8 komponentene nevnt over. Oppholdstiden fra AnnaLyse derimot vil i tillegg betegne en del av tiden som toget går etter Figur 13. Det blir i denne oppgaven derfor nødvendig å presentere en ny, alternativ, modell som tar hensyn til komponenten fra AnnaLyse:



Figur 14 Komponenter reisetid.

Denne alternative modellen vist i Figur 14 gir oss mulighet til å beskrive følgende komponenter:

1. Tog passerer signal på vei inn på stasjon
2. Tog stopper
3. Kontrollprosedyre
4. Dørene åpner
5. Av- og påstigning
6. Eventuell ventetid før avgang i forhold til rutetabellen
7. Kontrollprosedyre
8. Dørene stenges
9. Endelig klarering og avgang
10. Tog passerer signal på vei ut av stasjon

4.1.1.2 AnnaLyse

I det følgende beskrives henholdsvis oppholdstiden, kjøretiden og ankomstforsinkelsen basert på data fra AnnaLyse systemet. Toggang og datagrunnlag for dataene fra AnnaLyse er vist i vedlegg 8.5.1.

OPPHOLDSTID

Som tidligere nevnt i kapittel 3, gir AnnaLyse systemet oss en oppholdstid som betegner hvor lang tid toget belegger en blokkstrekning; i dette tilfelle en stasjon (se Figur 10). Denne oppholdstiden kan ikke direkte sammenlignes med noen referanse, men det er en forutsetning at den skal være større enn oppholdstiden fra TELOC som er når toget står i ro.

Tabell 5 viser oppholdstiden slik den framkommer fra AnnaLyse med hensyn på de ulike stasjonene det finnes data for. Det er også tegnet histogrammer med normalkurver for dataene for alle stasjoner. Dette er vist i vedlegg 8.5.1.

Stasjon	Antall	Min	Max	Gjennomsnitt	Trimmet snitt 20 %	Trimmet snitt 50 %	Skjevfordeling
Brakerøya	238	01:32	20:16	02:48	02:13	01:57	4,16
Lier	239	00:38	44:40	03:11	02:38	02:35	9,68
Asker	232	01:01	45:07	03:15	02:38	02:33	8,34
Sandvika	245	00:01	40:41	02:40	02:07	02:03	8,20
Lysaker	241	01:11	43:12	03:33	02:48	02:38	5,96
Skøyen	258	00:40	41:48	03:23	02:24	02:03	4,78
Nationaltheatret	258	00:02	40:53	03:03	02:11	02:00	5,58
Oslo	223	03:16	22:12	05:14	04:37	04:36	4,16
Lillestrøm	243	02:22	19:32	04:01	03:40	03:37	5,20
Frogner	259	01:34	02:41	01:55	01:54	01:54	0,85
Kløfta	257	01:52	36:40	04:02	03:07	02:51	4,37
Gardermoen	256	01:01	32:59	03:47	02:58	02:54	4,57

Tabell 5 Oppholdstid Anna Lyse

Kolonne 2, 3, og 4 viser henholdsvis antall observasjoner, minimum og maksimum. Kolonne 5 viser gjennomsnittet av observasjonene, kolonne 6 og 7 er et såkalt trimmet¹⁵ gjennomsnitt av observasjonene for å ekskludere ekstreme verdier. Den siste kolonnen, 8, er et mål på hvordan verdiene er fordelt i forhold til middelveien beregnet i kolonne 5. Skjevfordelingen karakteriserer fordelingen rundt dens middelvei. Positiv skjevfordeling indikerer en fordeling med en asymmetrisk side som heller mot positive verdier. Negativ skjevfordeling indikerer en fordeling med en asymmetrisk side som heller mot negative verdier og beregnes etter følgende formel:

$$(1.3) \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

Kolonne 8 kan dermed fortelle oss hvordan verdiene er fordelt rundt middelveien og om det ville være hensiktsmessig å fjerne ekstreme verdier for å gi en bedre normalfordeling. Sammenhengen kan observeres ved å se hvordan gjennomsnittet går ned ved å trimme det i forhold til middelveien ved en positiv skjevfordeling, og omvendt.

¹⁵ Beregner middelveien ved å ekskludere en viss prosent av datapunkt fra toppen og bunnen av datasettet. Hvis prosent for eksempel er 20 %, vil 4 punkt bli trimmet fra et datasett på 20 punkt (20x0,2): 2 fra toppen av settet og 2 fra bunnen av settet.

Som tidligere nevnt er det ikke mulig med det datamaterialet som var tilgjengelig i denne oppgaven å sammenligne oppholdstiden fra AnnaLyse med noen planlagt verdi. Det er av den grunn vanskelig å kunne gjøre noen betraktninger om verdiene fra Tabell 5 og deres størrelse. Man kunne tenke seg at da det ikke finnes noen planlagt verdi å sammenligne med, hadde det vært mer interessant å sett på hvordan oppholdstiden varierer i forhold til tognummer og tidspunkt, enn å utvikle en gjennomsnittverdi for hele perioden. En slik analyse kunne eventuelt identifisert interessante trender i belegging av en blokkstrekning.

KJØRETID

I Tabell 6 ser man kjøretiden basert på data fra AnnaLyse-systemet. Det har blitt tegnet histogrammer med normalkurver for dataene, vist i vedlegg 8.5.3. Kjøretiden er fra foregående stasjon det finnes data for, til den stasjonen indikert i tabellen. Eksempelvis vil det for Lier være kjøretiden fra foregående stasjon, som er Brakerøya.

Stasjon	Antall	Min	Max	Gjennomsnitt	Trimmet snitt 20 %	Trimmet snitt 50 %	Skjevfordeling
Brakerøya	248	00:36	24:42	01:55	01:29	01:21	6,40
Lier	254	00:22	03:57	00:52	00:51	00:51	13,10
Asker	254	05:55	09:58	07:11	07:07	07:06	1,02
Sandvika	258	04:50	09:14	06:13	06:11	06:11	0,92
Lysaker	255	05:13	08:38	05:48	05:44	05:43	3,18
Skøyen	260	01:32	06:21	01:49	01:44	01:44	7,17
Nationaltheatret	272	01:21	03:58	01:40	01:38	01:37	5,71
Oslo	272	00:55	08:09	01:26	01:19	01:18	7,09
Lillestrøm	258	10:00	14:41	11:28	11:25	11:26	0,82
Frogner	268	05:39	09:47	06:32	06:24	06:21	2,06
Kløfta	259	01:35	32:58	08:12	07:11	06:52	2,55
Gardermoen	273	06:37	14:54	08:38	08:37	08:37	2,04
Eidsvoll	271	08:50	15:51	10:17	10:11	10:09	2,50

Tabell 6 Kjøretid Anna Lyse

Tabell 6 følger det samme mønsteret for kolonnene som Tabell 5.

På samme måte som for oppholdstiden, kan heller ikke den beregnede gjennomsnittlige verdien for kjøretiden vist i Tabell 6 sammenlignes med en planlagt verdi. Avhengig av hva man er ute etter å vite gir slike gjennomsnittverdier mye eller lite mening. Disse verdiene i Tabell 6 gir først og fremst et gjennomsnittlig bilde av belegging av blokkstrekningen mellom stasjonene. De kan ikke vise hvordan beleggingen varierer over for eksempel tid.

ANKOMSTFORSINKELSE

Generelt vil Anna Lyse systemet oppgi et avvik, i sekund, fra planlagt ankomst og avgang til en stasjon. Denne planlagte ankomsttiden er oppgitt i minutter. For de stasjonene der kun avgangstiden er angitt i rutehåndboken, vil begge disse avvikene angis i forhold til nettopp avgangstiden. Tabell 7 viser ankomstforsinkelsen man kan hente ut fra Anna Lyse, der man har en ankomsttid for den aktuelle stasjonen oppgitt i rutehåndboken.

Stasjon	Antall	Min	Max	Gjennomsnitt	Trimmet snitt 20 %	Trimmet snitt 50 %	Skjevfordeling
Asker	243	00:01	44:07	02:14	01:38	01:32	8,32
Oslo S	231	00:01	19:12	02:12	01:36	01:35	4,23
Lillestrøm	255	00:01	17:32	02:01	01:40	01:38	5,10
Kløfta	268	00:52	35:40	03:17	02:16	01:56	3,78
Gardermoen	266	00:01	31:59	02:53	02:00	01:55	4,23
Eidsvoll	61	00:02	14:36	02:45	02:20	01:52	1,38

Tabell 7 Ankomstforsinkelse: AnnaLyse.

Det er valgt å ta med alle positive avvik (forsinkelser) fra ankomsttiden. Grunnen er at målet for NSB burde være et avvik på null, sammen med det faktum at kunden alltid vil oppleve et avvik som en forsinkelse selv om dette skulle ligge under den definerte grensen på 3 minutter. Det betyr at man kan observere at gjennomsnittlig ankomstforsinkelse vil endre seg en del både i forhold til å trimme middelveien til 20 og 50 prosent. Dette kommer av at man har som man kan se av Tabell 7 en veldig stor spennvidde i målingene.

En annen ting som er verdt å merke seg er hvordan det er et stort antall tog som er i forsinket tidlig i pendelen, men et mye mindre antall tog er forsinket ved ankomst endestasjon. Forutsatt at man måler punktligheten ved endestasjonen i pendelen som andel tog innenfor et godkjent avvik, vil man komme ut med et relativt positivt inntrykk. Men derimot som reisende "innenfor" pendelen vil man oppleve mange flere avvik fra ankomsttiden. Grunnen til at man ser en slik oppførsel i ankomstforsinkelsen over strekningen er at man har lagt inn slakk i kjøretiden for å kunne ta høyde for eventuelle forsinkelser langs med pendelen. Dermed vil man ha større mulighet for å kjøre inn eventuelle forsinkelser utover i pendelen.

For å illustrere dette momentet med et enkelt eksempel for en valgt stasjon, kan man bruke passasjertellinger for avstigende passasjerer ved hjelp av følgende formel:

$$(1.4) \frac{\text{AntallForSeneTog}}{\text{TotaltAntallTog}} * \text{AntallAvstigende}$$

Eksempelvis vil dette for Kløfta stasjon gi at gjennomsnittlig vil $96,8\% * 4425 = 4281$ passasjerer, over perioden det er sett på, oppleve en forsinket ankomst til sin destinasjon. Dette utgjør 96,8 % av alle tog og avstigende passasjerer. Hvis man antok samme prosentandel forsinkede tog ved Eidsvoll stasjon som ved Kløfta stasjon, ville i gjennomsnitt over perioden $96,8\% * 2794 = 2703$ passasjerer, oppleve en forsinket ankomst til sin destinasjon. Selv om man kommuniserer utad at punktligheten var bra fordi man måler punktligheten ved endestasjon vil passasjerene ikke oppleve det samme på grunn av at antall avstigende passasjerer avtar utover i pendelen mot endestasjon.

4.1.1.3 TELOC

I det følgende blir oppholdstiden, kjøretiden og ankomstforsinkelsen basert på data fra TELOC beskrevet. De målepunktene dataene stammer fra er vist i Figur 10.

Forskjellen mellom dataene fra AnnaLyse og TELOC er at de sistnevnte kan sammenlignes med planlagte verdier fra Trainplan. Dette gir oss muligheten til å komme med noen betraktninger om størrelsen på verdiene og eventuelle avvik fra det planlagte.

OPPHOLDSTID

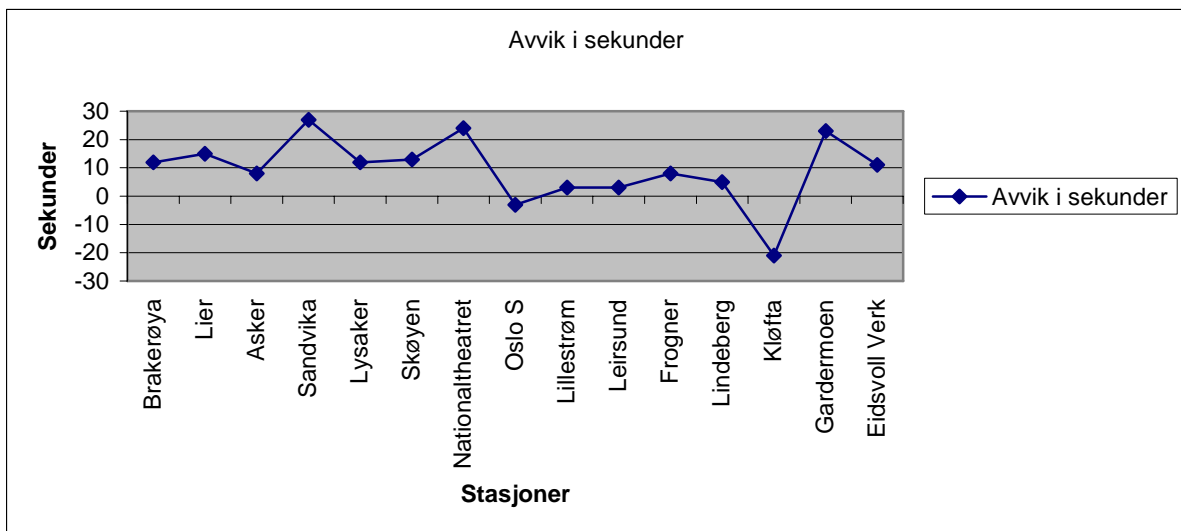
I Tabell 8 viser oppholdstiden hentet fra TELOC på de forskjellige stasjonene i pendelen sammenlignet med det som er spesifisert i *Trainplan*. Histogrammer for hver stasjon med tegnet normalkurve er vist i vedlegg 8.6.1.

Stasjon	Spesifisert	Antall	Min	Max	Gjennomsnitt	Trimmet snitt 20 %	Trimmet snitt 50 %	Skjevfordeling
Drammen								
Brakerøya	00:20	30	00:19	01:09	00:33	00:32	00:32	2,19
Lier	00:20	29	00:22	01:13	00:36	00:35	00:34	2,10
Asker	01:01	28	00:31	02:12	01:10	01:09	01:06	0,61
Sandvika	00:30	31	00:30	01:37	00:59	00:57	00:56	0,68
Lysaker	00:30	27	00:25	01:08	00:43	00:42	00:40	0,97
Skøyen	00:30	31	00:24	01:22	00:45	00:43	00:41	1,30
Nationaltheatret	00:30	31	00:30	01:37	00:56	00:54	00:53	0,84
Oslo S	03:00	31	00:59	05:01	02:55	02:57	02:59	-0,17
Lillestrøm	02:00	31	00:05	03:12	02:00	02:03	02:06	-0,46
Leirsund	00:30	30	00:23	00:59	00:35	00:33	00:33	1,37
Frogner	00:30	31	00:24	01:05	00:39	00:38	00:38	1,19
Lindeberg	00:30	31	00:22	01:02	00:36	00:35	00:34	1,05
Kløfta	01:01	30	00:23	02:41	00:44	00:40	00:39	4,31
Gardermoen	01:01	31	00:43	02:23	01:25	01:24	01:23	0,36
Eidsvoll Verk	00:30	30	00:18	01:23	00:43	00:41	00:39	1,07
Eidsvoll								

Tabell 8 Oppholdstider TELOC

Kolonne 2 i Tabell 8 angir stasjonsoppholdet slik det er spesifisert i *Trainplan*. Det er viktig å merke seg at denne oppholdstiden ikke nødvendigvis er teoretisk mulig. Avviket mellom hva som er teoretisk mulig og hensynet til at et tog ikke kan (planlagt) forlate en stasjon på annet enn hele minutter, blir justert for i kjøretiden. Publikum forholder seg uansett, i de tilfeller hvor oppholdstiden er angitt i rutetabellen, til dette angitte stasjonsoppholdet. Et eksempel kan være Brakerøya stasjon der en kunde vil gjennomsnittlig oppleve at toget er 20 sekunder forsinket ved avgang i henhold til *Trainplan*. For en nærmere forklaring av hvordan disse justeringene gjøres i praksis vises det til vedlegg 8.4.

Som man kan se viser den faktiske oppholdstiden, med visse unntak, tendensen til å ligge over den planlagte oppholdstiden fra *Trainplan*. Den samme tendensen er vist i Figur 15.



Figur 15 Avvik i sekunder fra planlagt oppholdstid (trimmet snitt 20 %)

Det er vanskelig å trekke noen konklusjoner om hvorfor de faktiske oppholdstidene er lengre enn de planlagte. Man kan derimot spekulere i om det ligger såpass mye slakk i kjøretidene at det spiller ingen rolle for ankomstpunktligheten om togene står lengre enn planlagt. På en generell basis bør det derimot være et mål at man følger det som er planlagt.

KJØRETID

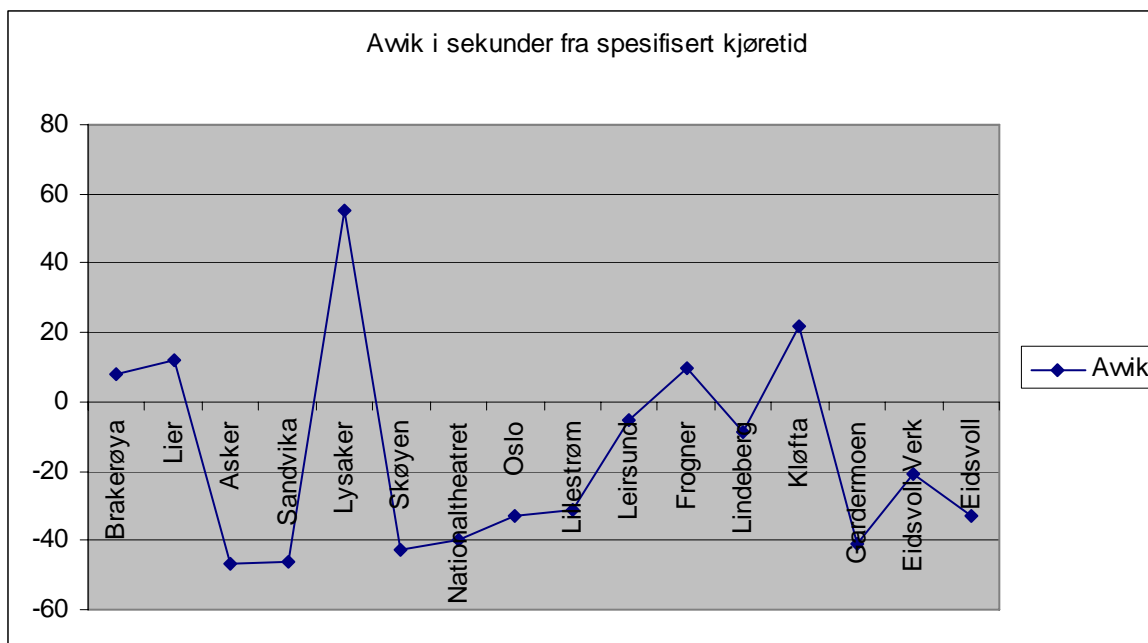
Fra TELOC kan man beregne kjøretidene fra det tidspunkt toget starter til det står i ro igjen.

Stasjon	Spesifisert	Antall	Min	Max	Gjennomsnitt	Trimmet snitt 20 %	Trimmet snitt 50 %	Skjevfordeling	Avvik [sekund]
Brakerøya	02:10	30	02:03	03:12	02:20	02:18	02:18	2,11	8
Lier	02:38	29	02:37	03:26	02:51	02:50	02:49	1,43	12
Asker	07:59	27	06:35	08:46	07:16	07:12	07:10	1,65	-47
Sandvika	07:30	28	05:37	07:37	06:44	06:44	06:46	-0,27	-46
Lysaker	05:30	29	05:23	08:43	06:28	06:25	06:27	0,64	55
Skøyen	03:30	29	02:31	04:28	02:50	02:47	02:46	3,79	-43
Nationaltheatret	03:30	31	02:34	03:12	02:50	02:50	02:50	0,28	-40
Oslo	03:00	31	02:06	03:48	02:30	02:27	02:27	2,61	-33
Lillestrøm	11:00	31	09:03	17:32	10:40	10:29	10:28	3,83	-31
Leirsund	04:30	30	03:56	06:12	04:29	04:25	04:24	1,80	-5
Frogner	02:30	30	02:26	02:58	02:41	02:40	02:39	0,54	10
Lindeberg	02:30	31	02:07	02:59	02:22	02:21	02:20	1,01	-9
Kløfta	02:59	30	02:48	06:43	03:30	03:21	03:18	2,80	22
Gardermoen	09:59	30	07:32	10:16	09:17	09:18	09:19	-0,71	-41
Eidsvoll Verk	06:30	30	05:28	07:03	06:10	06:09	06:10	0,23	-21
Eidsvoll	05:00	27	03:40	05:30	04:27	04:27	04:26	0,32	-33

Tabell 9 Kjøretid TELOC.

Tabell 9 viser kjøretidene fra foregående stasjon i pendelen til stasjonen oppgitt i tabellen. Kolonne 2, "spesifisert" er kjøretiden slik den er spesifisert i *Trainplan*. Det man kan observere er at kjøretiden både avviker positivt og negativt fra det planlagte.

I Figur 16 ser man hvordan avviket utvikler seg utover i pendelen. Avviket er beregnet i forhold til det trimmede snittet på 20 %.



Figur 16 Avvik fra spesifisert kjøretid i TrainPlan i forhold til TELOC (trimmet snitt 20 %).

Det er vanskelig på grunnlag av datagrunnlaget å trekke noen entydige konklusjoner om kjøretiden. Men det man kan merke seg er hvordan grafen for kjøretiden støtter opp om det som ble diskutert for ankomstforsinkelsen under 4.1.1.2; det ser ut som om det er nok slakk i den planlagte togframføringen til å kunne kjøre inn eventuelle forsinkelser mot slutten av pendelen. Dette kan observeres fra hvordan kjøretiden ligger gjennomsnittlig under det planlagte for de tre siste stasjonene i pendelen.

En annen interessant ting å merke seg er hvordan avviket fra den planlagte kjøretiden ikke viser noen tendens til å følge størrelsen på selve kjøretiden. Hvis man for eksempel sammenligner den planlagte kjøretida fra Nationaltheatret til Oslo S (3 minutter) med den faktiske fra TELOC (2 minutter og 27 sekunder) er dette et avvik på om lag 17 %. Den faktiske kjøretiden fra Oslo S til Lillestrøm viser om lag det samme avviket fra den planlagte som fra Nationaltheatret til Oslo S, men her er den planlagte kjøretiden hele 11 minutter, noe som betyr at det prosentmessige avviket er mye mindre.

ANKOMSTFORSINKELSE

Tabell 10 viser avvikene i ankomstene hentet fra TELOC sammenlignet med det som er oppgitt i rutehåndboken. Alle avvikene fra den planlagte ankomsten er angitt i tabellen. Som man kan observere, går det gjennomsnittlige avviket mot null for endestasjonen Eidsvoll.

Stasjon	Antall	Min	Max	Gjennomsnitt	Trimmet snitt 20 %	Trimmet snitt 50 %	Skjevfordeling
Asker	28	-69	229	38	34	30	0,79
Oslo	31	-44	609	62	39	30	2,97
Lillestrøm	30	-67	414	38	12	11	2,36
Kløfta	30	5	283	79	65	56	1,68
Gardermoen	31	-67	343	18	5	-2	2,25
Eidsvoll	28	-116	298	8	1	-3	1,82

Tabell 10 Ankomstavvik alle avvik: TELOC [sekunder].

Tabell 11 viser kun ankomstforsinkelsen: det vil si alle tog som ankom for sent i forhold til tidspunkt oppgitt i rutehåndboken med hensyn på stasjon. Man kan observere at også her viser forsinkelsen en tendens til å ned mot slutten av pendelen. Ingen av de gjennomsnittlige ankomstforsinkelsene ligger over den aksepterte grensen på 180 sekunder, men det finnes derimot enkeltverdier som gjør det.

Stasjon	Antall	Min	Max	Gjennomsnitt	Trimmet snitt 20 %	Trimmet snitt 50 %	Skjevfordeling
Asker	20	1	229	67	60	58	1,17
Oslo	19	7	609	115	92	77	2,85
Lillestrøm	17	1	414	93	77	45	1,96
Kløfta	30	5	284	79	65	56	1,69
Gardermoen	13	3	343	86	70	65	2,12
Eidsvoll	14	5	298	61	46	41	2,51

Tabell 11 Ankomstforsinkelse TELOC [sekunder].

Generelt kan man si om dataene for ankomstforsinkelsen at de har den svakhet at de relaterer til seg et klokkeslett; de er ikke absolutte differanser mellom to verdier. For TELOC sin del er avvikene beregnet i forhold til dens interne klokke. Det vil med andre ord si at størrelsen på avvikene nødvendigvis må korrigeres for om TELOCs klokke er korrekt.

4.1.2 DIFFERANSE MELLOM OPPHOLDSTIDER

De to datakildene gir oss muligheten til, som man kan se av Figur 10, å beregne differansen mellom oppholdstiden fra TELOC; hvor lenge toget står i ro; og oppholdstiden fra AnnaLyse; hvor lenge toget belegger en blokkstrekning.

Ideelt sett hadde det vært ønskelig å beregne to parametere; en fra toget passerer signalet inn på stasjonen til det stopper, og en fra toget starter til det passerer signalet på vei ut av stasjonen. Men på grunn av usikkerheten knyttet til om TELOC og AnnaLyse forholder seg til samme klokke, noe som vil gi store utslag ved små differanser og oppløsning på sekundnivå, kommer jeg til kun til å beregne den samlede differansen. Det som beregnes er med andre ord altså A + B fra Figur 10.

Stasjon	TELOC	AnnaLyse	Differanse
Drammen	Trimmet snitt 20 %	Trimmet snitt 20 %	
Brakerøya	00:32	02:13	01:41
Lier	00:35	02:38	02:03
Asker	01:09	02:38	01:29
Sandvika	00:57	02:07	01:10
Lysaker	00:42	02:48	02:06
Skøyen	00:43	02:24	01:41
Nationaltheatret	00:54	02:11	01:17
Oslo S	02:57	04:37	01:40
Lillestrøm	02:03	03:40	01:37
Leirsund	00:33		
Frogner	00:38	01:54	01:16
Lindeberg	00:35		
Kløfta	00:40	03:07	02:27
Gardermoen	01:24	02:58	01:34
Eidsvoll Verk	00:41		
Eidsvoll			

Tabell 12 Differanse i oppholdstid TELOC og AnnaLyse.

Tabell 12 viser den totale tiden toget bruker mellom signaler til det starter eller stopper, for hver stasjon det finnes data fra begge kildene for. Det som kan observeres fra tabellen er at differansen mellom oppholdstida fra AnnaLyse og TELOC er stor. Dette stiller spørsmål ved om oppholdstida fra AnnaLyse virkelig er et godt mål på noe som har med et stasjonsopphold å gjøre, eller om denne verdien får sin størrelse fra om toget venter på signal. Ideelt sett er denne differansetiden et uttrykk for akselrasjons- og retardasjonstid, men hvis tidsrommet derimot inneholder en stor grad av venting på signal vil det ikke gi mening å si at verdien påvirkes av noe som skjer inne på selve perrongen; for eksempel passasjerantall eller høyde på plattform.

4.1.3 SAMMENDRAG

Det er i denne delen av kapitlet brukt datakildene AnnaLyse og TELOC, til å beregne henholdsvis oppholdstid, kjøretid og ankomstforsinkelse. I tillegg er det sett på differansen mellom oppholdstidene for de to kildene.

Når det gjelder AnnaLyse finnes det ingen planlagte verdier man kan sammenligne med. Det er derfor vanskelig å trekke noen konklusjoner om den relative størrelsen på de beregnede gjennomsnittlige verdiene. For TELOC hadde man derimot sjansen til å sammenholde verdiene med planlagte verdier. Verdiene for oppholdstiden og kjøretiden kan sies å være absolutte i den forstand at de er mål for differansen mellom to tidspunkt. De beregnede ankomstforsinkelsene derimot, har den svakhet at de måles relativt til et klokkeslett. Av den grunn er det problematisk å diskutere ankomstforsinkelsens størrelse, selv om trender likevel kan identifiseres.

Dataene fra TELOC for oppholdstiden viste at den faktiske oppholdstiden hadde en generell tendens til å ligge over den planlagte. Årsaken til dette er ukjent, men det kan skyldes at den planlagte oppholdstiden fra Trainplan ikke er praktisk mulig, men er derimot satt slik av hensyn til publikumsruta. En annen årsak kan være at disse planlagte oppholdstidene ikke blir

aktivt brukt på andre stasjoner enn der man har ei ankomsttid oppgitt i publikumsruta, og at man isteden forholder seg kun til avgangstid.

Verdiene for kjøretiden og ankomstforsinkelsen viste at de begge demonstrerte en tendens til å avta mot slutten av pendelen. Dette kan antyde at man har større slakk i kjøretiden når man nærmer seg endestasjonen, noe som igjen kan indikere et ønske om å ha flest mulig tog punktlig ved endestasjon. Det bør likevel påpekes at dette blir kun spekulasjoner da forholdet ikke er undersøkt for forholdet nærmere.

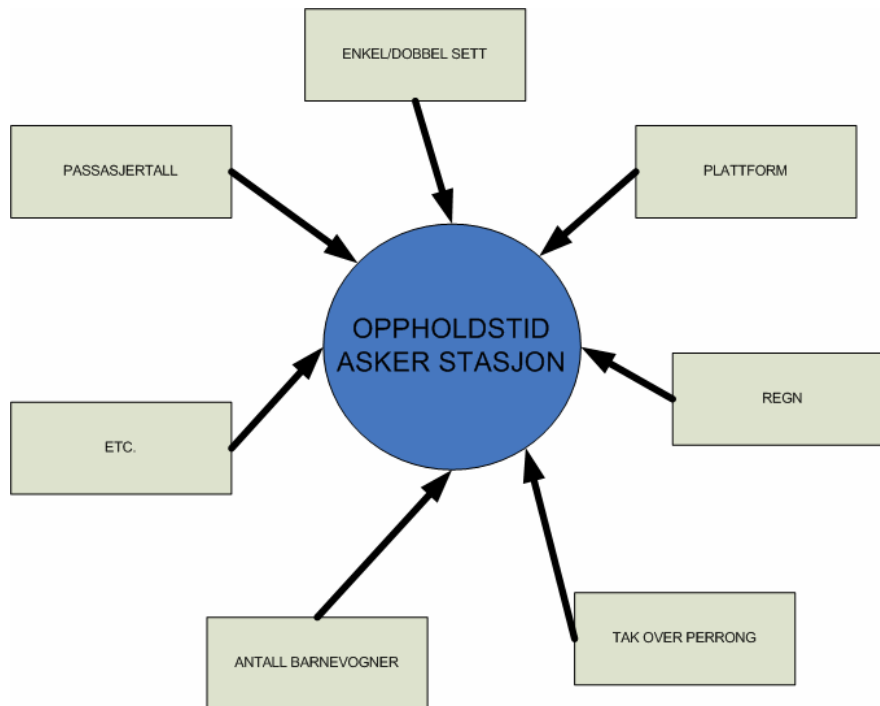
Datakildene har til tross for at de begge har krevd en relativ stor grad av prosessering, vært egnet til å produsere de parametrene man har sett på i denne delen av oppgaven. Hovedproblemet har som tidligere nevnt ligget i å sammenligne data fra de to kildene direkte. Kvaliteten av datakildene vil bli nærmere diskutert i kapittel 5.

4.2 FORKLARINGSPARAMETERE FOR OPPHOLDSTID

I denne delen av kapitlet undersøkes sammenhengen mellom oppholdstiden på Asker stasjon og to variabler: om toget kjører som dobbelt- eller enkeltsett, og ombordstigende passasjerer. Regresjonsanalyse blir brukt for å se på dette forholdet. Det blir etablert to modeller: én med oppholdstiden fra AnnaLyse som avhengig variabel og én med oppholdstiden fra TELOC som avhengig variabel.

Et sekundært forhold som ønskes belyst er om datagrunnlaget som blir brukt i arbeidet med denne oppgaven er hensiktsmessig til bruk i en slik regresjonsanalyse.

Det er rimelig å anta at det finnes mange andre variabler enn passasjerantallet og enkelt/dobbeltsett som påvirker oppholdstiden på Asker stasjon og man kommer aldri til å kunne lage en modell som inkluderer alle variablene og fullstendig forklare variasjonen i oppholdstiden. Egen erfaring har vist at oppholdstiden på en stasjon kan være avhengig av et vidt spekter av variabler som for eksempel: høydeforskjell plattform - tog, hvilken plattform toget kjører inn på, antall barnevogner på perrongen, regn og så videre. En illustrasjon av de mulige faktorene som kan påvirke oppholdstiden på Asker stasjon er vist i Figur 17.



Figur 17 Faktorer som kan påvirke oppholdstiden

4.2.1 EMPIRISKE DATA

Under følger en beskrivelse av det empiriske datamaterialet som ligger til grunn for analysen.

4.2.1.1 OPPHOLDSTID

Som avhengig variabel benyttes de to datasettene for oppholdstiden på Asker stasjon: AnnaLyse som viser den totale oppholdstiden mellom signaler, og TELOC som gir den tiden toget faktisk står i ro på stasjonen. Videre vil det konstrueres to regresjonsmodeller for de to forskjellige avhengige variablene, men med samme uavhengige variabler.

Tabell 13 viser oppholdstidene for Asker stasjon fra de to datasettene AnnaLyse og TELOC i perioden 22. til 28. august. Oppholdstidene for hverdag fra AnnaLyse er angitt som gjennomsnittsverdier for det aktuelle tognummeret over den samme perioden, mens verdiene fra TELOC er de faktiske registrerte verdiene for det aktuelle tog og dato.

Tog	Tidspunkt	Hverdag		Lørdag		Søndag	
		AnnaLyse	TELOC	AnnaLyse	TELOC	AnnaLyse	TELOC
1605	05:53	167	63				
1607	06:53	146	45	175			
1609	07:53	144	57	137		697	
1611	08:53	152		172		131	
1613	09:53	188	108	117	52	223	
1615	10:53	186		122		172	
1617	11:53	165	132	142		61	
1619	12:53	129	75	531		204	114
1621	13:53	151	52	143		147	
1623	14:53	182	44	141	53	138	
1625	15:53	219		185		543	
1627	16:53	205				370	
1629	17:53	117	81	173		140	
1631	18:53	143	79	186		141	
1633	19:53	137		131	72	229	
1635	20:53	170	90	93		112	
1637	21:53	160		111		245	
1639	22:53		50				
1641	23:53	203	43	83		95	

Tabell 13 Oppholdstider Asker stasjon: AnnaLyse og TELOC [sekunder].

4.2.1.2 PASSASJERTALL

Den første uavhengige variabelen som blir brukt kommer fra data for passasjertellinger. Passasjertellinger blir gjort en ukedag, en lørdag og en søndag, to ganger pr år for alle tog, og blir oppgitt som maksimalt antall passasjerer. De parametrene som blir oppgitt er: avstigende passasjerer, ombordstigende passasjerer og passasjerer om bord i toget. Et eksempel på disse passasjertellingene er gitt i vedlegg 8.3.

Med bakgrunn i det som ble diskutert i kapittel 2 kan det med ren betraktning foreslå at disse passasjertellingene ville komme relativt dårlig ut hvis man testet de i forhold til måleparametrene foreslått i kapittel 2.7. Det skal ikke diskuteres i detalj hvordan datakilden ligger an i forhold til hver parameter, men tidsriktigheten på dataene er særlig et problem i dette tilfellet; dataene gjelder ikke for det samme tidsrommet som man har data fra AnnaLyse og TELOC fra. Likevel bør man kunne anta at passasjertellingene kan gi et bilde av passasjerstrømmen og dermed brukes som sammenligningsgrunnlag med andre trenddata.

Tabell 14 viser maks antall reisende på Asker stasjon for 1600-tog i østgående retning samt beleggsprosent. Tall i rødt er dobbeltsett. Beleggsprosenten er beregnet ut fra antall reisende delt på maks antall seter, 300, for øvrig samme måte som i Olsson et. al. [42, s 70]. Med tanke på det som har blitt diskutert i forhold til kvaliteten på passasjertellingene, kommer alle nye parametre som beregnes med basis i disse dataene, for eksempel beleggsprosenten, bære med seg den samme svakheten. Kolonne 4 i Tabell 14 bør derfor betraktes på samme måte som kolonne 3; en trend over tid; ikke eksakte målinger.

Tog	Tidspunkt	Antall reisende hverdag	Belegg	Antall reisende lørdag	Belegg	Antall reisende søndag	Belegg
1605	05:53	111	37 %	-	-	-	-
1607	06:53	444	148 %	35	12 %	19	6 %
1609	07:53	574	191 %	167	56 %	32	11 %
1611	08:53	158	53 %	84	28 %	19	6 %
1613	09:53	73	24 %	30	10 %	39	13 %
1615	10:53	67	22 %	77	26 %	41	14 %
1617	11:53	79	26 %	102	34 %	74	25 %
1619	12:53	84	28 %	94	31 %	87	29 %
1621	13:53	77	26 %	67	22 %	46	15 %
1623	14:53	84	28 %	74	25 %	93	31 %
1625	15:53	128	43 %	83	28 %	115	38 %
1627	16:53	82	27 %	124	41 %	75	25 %
1629	17:53	64	21 %	98	33 %	86	29 %
1631	18:53	39	13 %	107	36 %	100	33 %
1633	19:53	31	10 %	73	24 %	80	27 %
1635	20:53	32	11 %	77	26 %	47	16 %
1637	21:53	57	19 %	33	11 %	83	28 %
1639	22:53	22	7 %	-	-	-	-
1641	23:53	18	6 %	45	15 %	17	6 %

Tabell 14 Antall reisende, tidspunkt og beleggsprosent Asker stasjon.

Tabell 15 viser data for påstigende og avstigende passasjerer sortert med hensyn på tognummer og tidspunkt for Asker stasjon. Dobbelsett er angitt i rødt.

Tog	Tidspunkt	Hverdag		Lørdag		Søndag	
		Påstigende	Avstigende	Påstigende	Avstigende	Påstigende	Avstigende
1605	05:53	10	1	-	-	-	-
1607	06:53	100	14	17	3	10	6
1609	07:53	100	30	13	4	11	2
1611	08:53	40	20	24	7	4	1
1613	09:53	12	5	8	2	6	2
1615	10:53	7	2	11	4	19	1
1617	11:53	17	3	22	6	17	0
1619	12:53	15	7	29	6	25	2
1621	13:53	40	3	20	3	7	2
1623	14:53	26	12	26	6	19	2
1625	15:53	30	20	21	2	25	2
1627	16:53	43	9	27	6	15	4
1629	17:53	30	3	40	12	18	1
1631	18:53	10	5	40	3	19	3
1633	19:53	7	1	35	10	9	5
1635	20:53	14	3	44	3	14	4
1637	21:53	16	3	9	1	6	4
1639	22:53	10	1	-	-	-	-
1641	23:53	5	0	9	2	2	0

Tabell 15 Påstigende og avstigende passasjerer Asker stasjon.

Som grunnlag for passasjertall i modellen benyttes det samlede antall avstigende og påstigende passasjerer.

4.2.1.3 ENKELT/DOBBELTSETT

Det andre datasettet som benyttes som uavhengig variabel, er om toget er et enkelt eller dobbeltsett. Tabell 16 viser avgangstidspunktene for de ulike togene og om de var enkelt eller dobbeltsett. Dette er det som virkelig ble kjørt i perioden 22.8 – 28.8.2005.

Tog	Tidspunkt	Hverdag	Lørdag	Søndag
1605	05:53	Enkel		
1607	06:53	Dobbel	Enkel	Enkel
1609	07:53	Dobbel	Enkel	Enkel
1611	08:53	Enkel	Enkel	Enkel
1613	09:53	Enkel	Enkel	Enkel
1615	10:53	Enkel	Enkel	Enkel
1617	11:53	Enkel	Enkel	Enkel
1619	12:53	Dobbel	Enkel	Enkel
1621	13:53	Dobbel	Enkel	Enkel
1623	14:53	Enkel	Enkel	Enkel
1625	15:53	Enkel	Enkel	Enkel
1627	16:53	Enkel	Enkel	Enkel
1629	17:53	Enkel	Enkel	Enkel
1631	18:53	Enkel	Enkel	Enkel
1633	19:53	Enkel	Enkel	Enkel
1635	20:53	Enkel	Enkel	Enkel
1637	21:53	Enkel	Enkel	Enkel
1639	22:53	Enkel		
1641	23:53	Enkel	Enkel	Enkel

Tabell 16 Enkelt eller dobbeltsett.

Som tabellen viser er det liten variasjon i om det kjøres enkelt eller dobbeltsett. I helgene kjøres det ikke dobbeltsett i det hele tatt.

4.2.2 MODELL

Med bakgrunn kan i det empiriske datamaterialet kan man nå sette opp en tabell for datagrunnlaget som skal brukes for de to modellene:

Tognummer	Avhengig variabel		Uavhengig variabel	Uavhengig variabel
	Oppholdstid AnnaLyse	Oppholdstid TELOC	Passasjerstrøm	Type Sett
1605	167	63	11	Enkel=0
1607	146	45	114	Dobbel=1
1609	144	57	130	Dobbel=1
1611	152		60	Enkel=0
1613	188	108	17	Enkel=0
1615	186		9	Enkel=0
1617	165	132	20	Enkel=0
1619	129	75	22	Dobbel=1
1621	151	52	43	Dobbel=1
1623	182	44	38	Enkel=0
1625	219		50	Enkel=0
1627	205		52	Enkel=0
1629	117	81	33	Enkel=0
1631	143	79	15	Enkel=0
1633	137		8	Enkel=0
1635	170	90	17	Enkel=0
1637	160		19	Enkel=0
1639		50	11	Enkel=0
1641	203	43	5	Enkel=0
Antall	18	13	19	19

Tabell 17 Datagrunnlag for modell.

Modellene er relativt små med kun 18 observasjoner for oppholdstiden fra AnnaLyse og 13 fra TELOC. Det finnes flere observasjoner man kunne brukt, som man har sett fra kapittel 5.1, men type togsett varierer ikke mer over fem hverdager enn én. På lørdag og søndag kjøres det normalt kun enkeltsett, med unntak av 1641 på søndag, dermed vil disse dagene være uinteressante fordi de ikke varierer. I modellen blir det antatt en verdi 0 for enkeltsett og 1 for dobbeltsett.

For å undersøke den eventuelle sammenhengen mellom variablene Oppholdstid (OT), Passasjertall(PT) og Enkelt/Dobbelsett(ED) kan man bruke det som kalles regresjonsanalyse. Denne analyseteknikken kan brukes til å studere sammenhengen mellom en eller flere såkalte uavhengige variabler $X_1, X_2, X_3, \dots, X_k$ og en avhengig kontinuerlig variabel Y . Det man ønsker å studere med regresjonsanalyse er hvordan endringer i de uavhengige variablene forklarer endringer i den avhengige variabelen [26]. En generell måte å uttrykke dette forholdet på er:

$$(1.5) y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + \varepsilon_i$$

Hvor Y er den avhengige variabelen og X - ene representerer de k uavhengige variablene, ε_i er feilleddet og representerer den uforklarlige delen av modellen, og i - ene referer til den i -te observasjonen. β_0 kalles konstantleddet, og angir oppholdstiden når alle variablene antar verdien null. $\beta_i (i=1, 2, 3, \dots, k)$ kalles regresjonsparametrene og angir den "isolerte" effekten som hver forklaringsvariabel har på oppholdstiden. For eksempel vil β_1 fortelle oss hvor stor effekts en enhetsendring av PT vil ha på oppholdstiden.

I vårt tilfelle betraktes ED og PT som uavhengige variabler og OT som den avhengige variabelen. Med andre kan man utrykke forholdet mellom det man ønsker å undersøke på følgende måte:

$$(1.6) \text{Oppholdstid} = F(PT, ED, \dots)$$

Eller hvis man antar at funksjonen er lineær på samme måte som i ligning (1.5):

$$(1.7) \text{Oppholdstid} = \beta_0 + \beta_1 PT + \beta_2 ED + \varepsilon$$

4.2.3 ESTIMERING

Ved hjelp av de to modellene er ønsket å kunne bekrefte eller avkrefte om passasjerstrøm og type togsett påvirker oppholdstiden ved Asker stasjon. Før man starter å beregne modellene er det nødvendig å etablere en nullhypotese og en alternativ hypotese til modellen:

$$(1.8) \text{Oppholdstid} = \beta_0 + \beta_1 PT + \beta_2 ED + \varepsilon$$

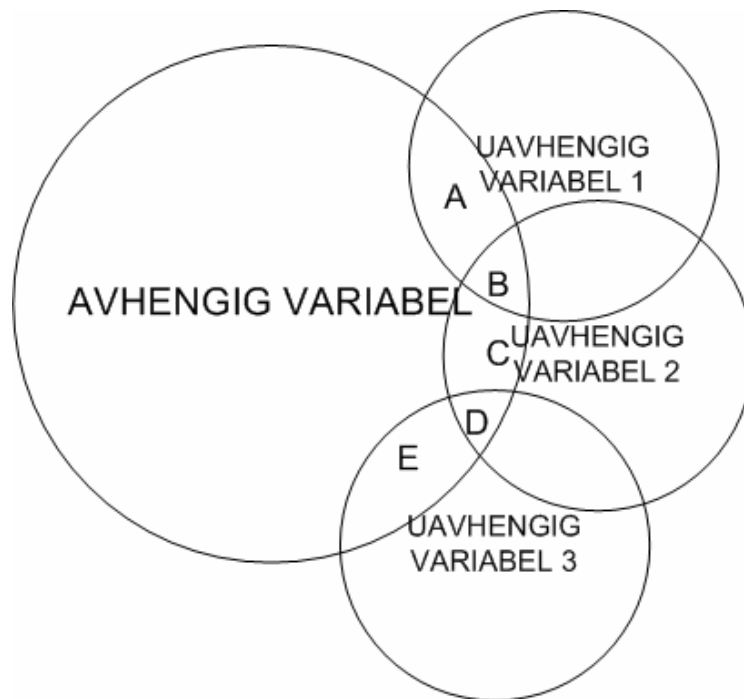
Nullhypotesen er: $H_0 : \beta_i = 0$; og den sier at verken passasjerstrømmen eller enkelt/dobbelsett påvirker oppholdstiden på Asker stasjon.

Den alternative hypotesen er at $H_1 : \beta_i > 0$; og sier at både passasjerstrømmen og om det kjøres med dobbelsett kan påvirke oppholdstiden på Asker stasjon.

Ved hjelp av regresjonsanalyse bekrefte eller avkrefte nullhypotesen, og ensidige t-tester benyttes for å finne ut om parametrene har signifikante verdier som er forskjellig fra null.

For å estimere parametrene benyttes en statistisk programvare som heter SPSS. Metoden som benyttes er standard multipel regresjonsanalyse, eller kjent som ENTER i SPSS. I denne metoden blir alle de uavhengige variablene tatt inn i modellen samtidig, men hver uavhengig variabel blir vurdert som om den hadde blitt tatt inn i modellen etter alle de andre uavhengige variablene har blitt vurdert. Hver og en blir vurdert i forhold til hva de tilføyer til forutsigelsen av den avhengige variabelen. Med andre ord blir forholdet mellom hver uavhengig variabel og den avhengige variabelen etablert med bidragene fra de andre uavhengige variablene delt ut. Områder av den avhengige variabelen som overlappes av mer enn en uavhengig variabel blir tatt med i beregningen av den totale variansen forklart av settet av uavhengige variabler, men ikke tildelt en spesiell uavhengig variabel [43, s. 185].

Sammenhengen kan beskrives ut fra diagrammet i Figur 18.



Figur 18 Venn diagram som illustrer overlappende varians.

I følge Figur 18 blir altså uavhengig variabel 1 kreditert for å forklare A, uavhengig variabel 2 for å forklare C og uavhengig variabel 3 for å forklare E; B og D blir tatt med for å forklare den totale variansen, men ikke tildelt noen spesiell uavhengig variabel. I disse to modellene er det kun to uavhengige variabler, men prinsippet er det samme som vist over.

Før man går over til resultatet av den statistiske analysen er det nødvendig å se på hvilke parametere som kommer ut av en multippel regresjonsanalyse i SPSS.

Multiple R

Denne koeffisienten reflekterer styrken i relasjonen mellom den avhengige variabelen og de kombinasjonene av uavhengige variabler som har blitt vektet i henhold til regresjonslikningen: $y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + \varepsilon_i$. Koeffisienten antar verdier mellom 0 og 1, der 0 indikerer et det ikke finnes noe lineært forhold mellom de uavhengige variablene og den avhengige. Høyere R- verdi antyder en sterkere lineær relasjon.

R^2

I multippel regresjonsanalyse betegner denne koeffisienten mengden av varians i den avhengige variabelen som er delt av kombinasjonen av de vektete uavhengige variablene. R^2 kan brukes som et mål på godheten av modellen i den forstand at hvis man trekker verdien for R^2 fra 1 og ganger med 100 vil man ha et prosentmessig mål på hvor mye som ikke er forklart av modellen.

Standard error

Med mindre modellen er perfekt ($R^2=1$), vil man ha en differanse mellom de estimerte og de virkelige observasjonene. Disse forskjellene betegnes i SPSS *standard error*, og reflekterer feil i modellen i form av standardavviket til feilene. En god modell vil gi en lav standard error og gir en generell pekepinne på den gjennomsnittlige feilen i estimatet.

B (stigningstallet til regresjonslinja)

B er den partielle regresjonskoeffisienten; der partielle indikerer at denne effekten blir beregnet etter innflytelsen til alle de andre uavhengige variablene har blitt statistisk kontrollert. En B – verdi på for eksempel -2 vil indikere at for hver økning av en enhet av den uavhengige variabelen vil gi en nedgang på 2 enheter i den avhengige variabelen.

β (standardisert stigningstall av regresjonslinja)

Med flere uavhengige variabler som måles på forskjellige skalaer oppstår det et problem med å sammenligne deres relative innflytelse. β - koeffisienten unngår denne problematikken ved å sammenligne standardavvikene for de uavhengige variablene og den avhengige variabelen. Hvis for eksempel β har en verdi på 0,5 for en uavhengig variabel, vil ett standardavviks økning i den uavhengige variabelen føre til 0,5 økning i standardavviket for den avhengige variabelen. Man bør likevel være forsiktig med å tolke β - verdier da de ikke er absolutte, men avhenger av de andre uavhengige variablene i likningen.

R^2 change

Når det legges til nye uavhengige variabler i en regresjonsanalyse vil det føre til en økning i R^2 , men det betyr ikke at det er økning i graden regresjonsmodellen passer; *standard error* kan faktisk øke. R^2 change er rett og slett forskjellen i verdien av R^2 før og etter man legger til en ny variabel.

4.2.3.1 MODELL 1: AVHENGIG VARIABEL: OPPHOLDSTID AnnaLyse

Den første modellen som estimeres i SPSS, er med oppholdstiden fra AnnaLyse som avhengig variabel.

For å oppsummere har man altså følgende modell:

Avhengig variabel: Oppholdstid (AnnaLyse) Asker stasjon

Uavhengig dummy variabel: Om toget er enkelt eller dobbeltsett, hvor 0 er enkel og 1 dobbel.

Uavhengig variabel: Passasjerantall

H_0 : verken passasjerstrømmen eller enkelt/dobbelsett påvirker oppholdstiden på Asker stasjon.

H_1 : både passasjerstrømmen og om det kjøres med dobbeltsett påvirker oppholdstiden på Asker stasjon.

Signifikansnivået er satt til å være 5 %.

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	TypeSett, Passasjerstrøm	.	Enter

a. All requested variables entered.

b. Dependent Variable: OppholdstidAnna

Figur 19 Metode, avhengig og uavhengige variabler; AnnaLyse.

Figur 19 viser en oversikt over metoden som ble brukt og hva som var den avhengige og den uavhengige variabelen. Tallet i kolonnen "Modell" referer til antall steg. Siden det er brukt metoden ENTER der alle variablene kommer inn simultant, er det kun ett steg.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,458 ^a	,210	,105	26,475

a. Predictors: (Constant), TypeSett, Passasjerstrøm

Figur 20 Modellens forklaringskraft; AnnaLyse.

I Figur 20 gir en oversikt over modellens forklaringskraft. Figuren viser at R^2 er 0,210, dette betyr at 21 % av variasjonen i oppholdstid er relatert til endringer i passasjerstrømmen og om det kjøres enkelt eller dobbeltsett. Estimater har videre, som man kan observere av figuren en høy *standard error*. Det er på grunnlag av dette at modellen har en svært lav forklaringskraft for det utvalget som er testet. Videre ser man at forskjellen mellom "R square" og "Adjusted R square" er relativt stor (0,105).

Figur 21 viser testing av nullhypotesen om at det ikke eksisterer noen lineær relasjon mellom anslagene og den avhengige variabelen.. F er forholdet mellom det gjennomsnittlige kvadratet for regresjonen og det gjennomsnittlige kvadratet for residualen.

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2792,358	2	1396,179	1,992	,171 ^a
	Residual	10513,642	15	700,909		
	Total	13306,000	17			

a. Predictors: (Constant), TypeSett, Passasjerstrøm

b. Dependent Variable: OppholdstidAnna

Figur 21 Anova; AnnaLyse.

Figur 21 viser at signifikansnivået forbundet med den observerte verdien av F er 0,171. Det betyr at man kan ikke uten videre forkaste nullhypotesen.

Figur 22 viser koeffisientene i modellen. Den forteller oss at ingen av variablene er signifikante på det valgte nivået.

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	167,330	9,254		18,082	,000		
	Passasjerstrøm	,145	,236	,182	,615	,548	,600	1,667
	TypeSett	-36,041	19,378	-,551	-1,860	,083	,600	1,667

a. Dependent Variable: OppholdstidAnna

Figur 22 Koeffisienter i modellen; AnnaLyse

Siste kolonne i Figur 22 tester for multikolaritet: om det finnes en innbyrdes korrelasjon mellom variablene. I følge Gripsrud et al. [26, s. 302] er det en tommelfingerregel at denne verdien skal være under 5. Som figuren viser er den det.

SAMMENDRAG

Regresjonsanalysen ga ingen entydig konklusjon annet at det ikke er grunnlag for å forkaste nullhypotesen. Verken ED, som ble brukt som dummy variabel, eller PT var signifikante på det valgte nivået.

Datamaterialet virker generelt ikke å inneholde et nødvendig antall observasjoner eller være nok tidsriktig til å kunne brukes som grunnlag for en slik modell.

4.2.3.2 MODELL 2: AVHENGIG VARIABEL: OPPHOLDSTID TELOC

Den andre modellen som er forsøkt estimert i SPSS er med oppholdstiden fra TELOC som avhengig variabel.

Man har altså følgende modell:

Avhengig variabel: Oppholdstid (TELOC) Asker stasjon

Uavhengig dummy variabel: Om toget er enkelt eller dobbeltsett, hvor 0 er enkel og 1 dobbel.

Uavhengig variabel: Passasjerantall

H_0 : verken passasjerstrømmen eller enkelt/dobbelsett påvirker oppholdstiden på Asker stasjon.

H_1 : både passasjerstrømmen og om det kjøres med dobbeltsett påvirker oppholdstiden på Asker stasjon.

Signifikansnivået er satt til å være 5 %.

Figur 23 viser på samme måte som Figur 19, regresjonsmetoden som ble brukt (ENTER), med avhengige og uavhengige variabler. ED ble også her brukt som dummy variabel der 0 antydte om toget var et enkeltsett og 1 om toget var et dobbeltsett.

Model	Variables Entered	Variables Removed	Method
1	TypeSett, Passasjerstrøm	.	Enter

a. All requested variables entered.

b. Dependent Variable: OppholdstidTE

Figur 23 Metode, avhengig og uavhengige variabler; TELOC.

Figur 24 viser modellens forklaringskraft. Som modellen avslører kan de uavhengige variablene forklare 13,2 % av variasjonen i oppholdstiden. *Standard error* er veldig høy. Man kan med andre ord si at modell 2 har, på samme måte som modell 1, svært lav forklaringskraft.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,363 ^a	,132	-,042	27,771

a. Predictors: (Constant), TypeSett, Passasjerstrøm

Figur 24 Modellens forklaringskraft; TELOC.

Den neste figuren, Figur 25, viser oss at den observerte F verdien ikke er signifikant og man derfor ikke uten videre forkaste nullhypotesen.

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1168,448	2	584,224	,758	,494 ^a
	Residual	7712,321	10	771,232		
	Total	8880,769	12			

a. Predictors: (Constant), TypeSett, Passasjerstrøm

b. Dependent Variable: OppholdstidTE

Figur 25 Anova; TELOC.

Figur 26 viser at ingen av de uavhengige variablene er signifikante. Test for kolinearitet viser at den ligger under 5. Det er dermed ikke noen grunn til å anta at det eksisterer kolinearitet mellom variablene.

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	78,817	10,694		7,370	,000		
	Passasjerstrøm	-,116	,289	-,169	-,402	,696	,493	2,030
	TypeSett	-12,614	23,777	-,223	-,531	,607	,493	2,030

a. Dependent Variable: OppholdstidTE

Figur 26 Koeffisienter i modellen; TELOC.

SAMMENDRAG

Modell 2 har ikke gitt oss noe grunnlag for å påstå at det finnes en sammenheng mellom oppholdstiden fra TELOC, passasjerstrømmen og om toget kjører enkelt eller dobbeltsett. Det vil derfor være naturlig, forutsatt at man bruker korrekt datamateriale, å beholde nullhypotesen om at det finnes ingen sammenheng mellom den avhengige og de uavhengige variablene.

4.2.3 DISKUSJON AV MODELLER

Målet med å lage disse to modellene var som påpekt under avsnitt 4.2, todelt: For det første ønsket man å undersøke om passasjerstrømmen og om et tog kjører enkelt eller dobbeltsett

påvirker lengden på stasjonsoppholdet. For det andre ville man se om det tilgjengelige datamaterialet var hensiktsmessig til bruk i en slik analyse.

Modellene i seg selv har ikke gitt oss noen grunn til å anta at det finnes noen sammenheng mellom de avhengige og de uavhengige variablene. Men som man diskuterte i kapittel 2 er det vanskelig at resultatet blir noe bedre enn det man putter inn. For å undersøke dette forholdet bedre burde man ha hatt korresponderende målinger i tid og lagd en tidsserie. Men med det datamaterialet som har blitt brukt i forbindelse med disse modellene, har dette ikke vært mulig.

I disse modellene har det først og fremst vært passasjertellingene som har vist seg å være det svake punktet, og da spesielt med hensyn på tidsriktighet. For å kunne sammenligne med dataene fra de andre datakildene er det nødvendig at observasjonene sammenfaller i tid; hvis ikke kan man undersøke sammenhenger som ikke eksisterer. Det er også et problem knyttet til oppløsningen på dataene; fordi om man har oppgitt målingene ned på enkeltpassasjernivå er de likevel bare anslag for antall passasjerer og ikke nøyaktige målinger. Det er feil å angi høyere oppløsning enn det som er tilfelle. Hvis passasjertellingene er bare anslag burde de oppgis som det; for eksempel: mindre enn 20, mellom 20 og 40 og så videre.

Hvis man ser på dataene fra TELOC og AnnaLyse, har dataene fra AnnaLyse vist seg å være foretrukket siden de både har et høyere antall observasjoner og høyere grad av fullstendighet relativt til bruksområdet. Dette gjør at AnnaLyse er mer tilgjengelig som datakilde i og med at man kan velge akkurat de dataene man ønsker. Men det bør nevnes at man kan i prinsippet få de dataene man ønsker fra TELOC.

Dataene for om togene kjørte som enkelt eller dobbeltsett kan sies å ligge utenfor NSB Drifts generelle dataflyt, men det er likevel lett tilgjengelige.

For å ha kunnet undersøkt ulike variabler som påvirker oppholdstiden på Asker stasjon grundigere vil det være nødvendig å samle inn egne data. Datagrunnlaget som ble brukt til å konstruere disse modellene har for dårlig kvalitet til å kunne brukes til en slik analyse og de gjenspeiler ikke nok mange av variablene som kan påvirke oppholdstiden på en stasjon.

4.3 SAMMENDRAG

Det er i dette kapitlet først brukt datakildene AnnaLyse og TELOC til utvikle gjennomsnittlige parametere for å beskrive strekingen Drammen – Eidsvoll over en 15 dagers periode. Dernest har man tatt i bruk data for passasjertellinger og om det ble kjørt enkelt eller dobbeltsett, for å undersøke hva som påvirker oppholdstiden på Asker stasjon.

De gjennomsnittlige parametrene som ble beregnet ut fra dataene fra AnnaLyse og TELOC var oppholdstid, kjøretid og ankomstforsinkelse. I tillegg beregnet man differansen mellom oppholdstiden fra TELOC og AnnaLyse. Dette ga en gjennomsnittlig verdi for hver stasjon, for hvor lang tid toget bruker mellom signaler på stasjonene og start/stopp tilstand. Til tross for en stor del av databearbeiding var datagrunnlaget tilfredsstillende ut fra et kvalitetsmessig hensyn som grunnlag for disse nevnte parametrene. Et problem oppsto derimot i forhold til tidsriktigheten hvis man ønsket å sammenligne målinger fra de to ulike kildene med hverandre. Dataene fra TELOC kunne i motsetning til de fra AnnaLyse sammenholdes med planlagte verdier. Dette ga et grunnlag for å kunne si noe om forholdet mellom det faktiske og det planlagte.

Oppholdstiden basert på data fra TELOC viste en generell tendens til å ligge over det som var planlagt ut fra Trainplan. Verdiene for ankomstforsinkelsen og kjøretiden basert på begge datakildene hadde en tendens til å avta mot slutten av pendelen.

Modellene for oppholdstiden ved Asker stasjon kunne ikke vise at oppholdstiden ved Asker stasjon var avhengig av verken passasjerstrømmen, eller om det ble kjørt enkelt eller dobbeltsett. Datagrunnlaget var heller ikke egnet til å gjøre en slik analyse, hovedsakelig knyttet til passasjertellingenes svake datakvalitet. For å kunne undersøke en slik sammenheng ble det foreslått å samle inn nye data.

I det neste kapitlet diskuteres kvaliteten på de to datakildene AnnaLyse og TELOC med bakgrunn i de foreslåtte måleparametrene i kapittel 2 og bruksområdet i kapittel 4.

5 DISKUSJON AV KVALITET PÅ KILDENE

I denne delen av oppgaven diskuteres kvaliteten på de to hovedkildene, TELOC og AnnaLyse, som er blitt brukt under arbeidet med denne oppgaven. Både objektive og subjektive måleparametere blir foreslått som diskusjonsgrunnlag for datakvaliteten. Videre ønsker man med bakgrunn i kapittel 4 å diskutere om datakildene er hensiktsmessige til sitt bruk; det vil si om de har mulighet til å gi oss den informasjonen vi er ute etter.

5.1 TYPER DATA OG MÅLINGER

Begge datakildene kan etter det som blir diskutert i avsnitt 2.2.1, betraktes som resultatet av objektive målinger; måleprosessen er automatisert og inneholder ingen subjektive vurderinger fram til det ferdige dataproduktet. Subjektive vurderinger ligger i tilfelle i hvordan måleinstrumentene er kalibrert i forhold til å måle de ulike parametrene, og hvordan målingene blir tolket.

Hovedforskjellen mellom de to kildene kan sies å ligge i hvor man befinner seg i dataproduksjonsprosessen.

Hvis man ser på Figur 4 og analogien til vareproduksjon, har begge kildene har krevd en viss grad av bearbeiding av dataene for å kunne bruke de til det formålet som blir gjort i denne oppgaven. Men isolert sett er AnnaLyse brukergrensesnittet i en database for å presentere blant annet punktligheten for ulike tog. Dataene man henter ut av AnnaLyse kan brukes til det rette formålet. TELOC derimot er rådata som trenger en eller annen form for bearbeiding ofte i form av en programvare, for å kunne lese noe meningsfullt ut av disse dataene. Derfor kan det være rimelig å betrakte TELOC som rådata og AnnaLyse som et dataprodukt.

Videre kan man si at dataproduktet fra AnnaLyse stammer fra en rekke prosesseringer av rådata utenfor brukerens kontroll. Uansett om man benytter dataene fra AnnaLyse til å lage nye dataprodukter, slik som tilfellet er for denne oppgaven, vil man fortsatt ikke kunne betrakte dataene som rådata fordi de allerede er prosessert. For å kunne jobbet med rådataene måtte man i tilfelle ha gått helt tilbake til signalsystemet og ”tappet” det på samme måte som for TELOC. Man befinner seg med andre ord på to forskjellige plasser i dataproduksjonskjeden: ved input for TELOC og ved output for AnnaLyse.

5.2 FORSLAG TIL MÅLEPARAMETERE

I avsnitt 2.6.1 ble det foreslått på grunnlag av teorien noen ulike måleparametere for datakvaliteten. Disse parametrene ble kategorisert i forhold til om de reflekterte den subjektive eller den objektive dimensjonen av datakvaliteten. Disse parametrene blir listet i det følgende.

5.2.1 OBJEKTIVE

- Antall observasjoner
- Oppløsning
- Nøyaktighet
- Fullstendighet
- Ensartethet

- Tidsriktighet

5.2.2 SUBJEKTIVE

- Tolbarhet
- Tilgjengelighet

5.2 TEST AV MÅLEPARAMETERE

I det følgende diskuteres kvaliteten av datakildene i forhold til måleparametrene som ble foreslått.

5.2.1 ANNA Lyse

5.2.1.1 ANTALL OBSERVASJONER

Antallet observasjoner som var tilgjengelig fra AnnaLyse i den analyserte tidsperioden varierte i forhold til hvilken stasjon man betraktet. Nærmere spesifisering av antallet observasjoner med hensyn på de ulike stasjonene finnes i Tabell 5 og Tabell 6.

5.2.1.2 OPPLØSING

Oppløsingen fra AnnaLyse er på sekundnivå. Denne oppløsningen forutsetter at man benytter data fra de stasjonene med automatiske signalanlegg; hvis ikke vil man måtte benytte data som har en oppløsning på minutt nivå, manuelt registrert (jfr. Figur 7).

5.2.1.3 NØYAKTIGHET

Som diskutert under avsnitt 2.6.2.3 er nøyaktigheten et forholdstall mellom korrigerede verdier og det totale antallet verdier. Datamaterialet ble i denne oppgaven brukt til å beregne gjennomsnittstørrelser for oppholdstiden, kjøretiden og ankomstforsinkelsen. Det var derfor ønskelig å ekskludere både de verdiene som var åpenbare feilregistreringer, samt de ekstreme verdiene som var utslag av unormale hendelser. For å oppnå et normalisert gjennomsnitt benyttet man seg av et trimmet gjennomsnitt. Fra Tabell 5 og Tabell 6 kan man se at man bør minst fjerne 20 % av de ytterliggende verdiene for å få et gjennomsnitt under normale driftsomstendigheter.

På grunnlag av det nevnt over foreslår jeg at nøyaktigheten til dataene fra AnnaLyse, forutsatt at man ønsker å bruke de til kvantitative analyser, er **80 %**.

5.2.1.4 FULLSTENDIGHET

Fullstendigheten er forholdet mellom antall observasjoner, gjengitt i Tabell 5 og Tabell 6, og det teoretiske antallet observasjoner gjengitt i Tabell 22. Denne parameteren vil i tillegg variere med stasjonen man ser på. Tabell 18 viser fullstendigheten og hvordan den varierer i forhold til stasjon.

Stasjon	Teoretisk Antall Observasjoner	Antall Observasjoner Kjøretid	Antall Observasjoner Oppholdstid	Fullstendighet Kjøretid	Fullstendighet Oppholdstid
Brakerøya	275	248	238	90 %	87 %
Lier	275	254	239	92 %	87 %
Asker	275	254	232	92 %	84 %
Sandvika	275	258	245	94 %	89 %
Lysaker	275	255	241	93 %	88 %
Skøyen	277	260	258	94 %	93 %
Nationaltheatret	277	272	258	98 %	93 %
Oslo	277	272	223	98 %	81 %
Lillestrøm	277	258	243	93 %	88 %
Leirsund					
Frogner	277	268	259	97 %	94 %
Lindeberg					
Kløfta	277	259	257	94 %	93 %
Gardermoen	277	273	256	99 %	92 %
Eidsvoll Verk					
Eidsvoll	277	271		98 %	
Gjennomsnitt	276	262	246	95 %	89 %

Tabell 18 Fullstendighet AnnaLyse.

Som tabellen viser vil den gjennomsnittlige fullstendigheten ligge på i overkant av 90 % for datamaterialet.

5.2.1.5 ENSARTETHET

AnnaLyse tar inn data fra flere forskjellige kilder både automatisk og manuelt registrerte uten å kategorisere de forskjellige målingene. Manuelt registrerte data ved stasjoner for avvik fra publikumsruten blir presentert sammen med automatiske registrerte data fra signalanlegget med samme oppløsning selv om de er registrert både med forskjellig metode og format.

Det er ugunstig at man samler data som er av forskjellige type og at man ikke kategoriserer de ulike målingene med hensyn på målemetode. En uerfaren bruker av systemet vil kunne ende opp med å betrakte forskjellige målinger som det samme.

5.2.1.6 TIDSRIKTIGHET

Når det gjelder å få tilgang til oppdaterte data, er tidsriktigheten i AnnaLyse-systemet bra. Man kan få tilgang til toggangen så tidlig som dagen etter, eller mer eller mindre simultant for data fra signalanlegg. Dataene for denne oppgaven ble hentet en uke etter dato for de siste registreringene.

Avvikene som er blitt brukt for å beregne oppholdstid og kjøretid i denne oppgaven blir målt i forhold til publikumsruten. Siden oppholdstiden og kjøretiden beregnes som forholdet mellom de to avvikene er det ikke kritisk at referansetiden samsvarer med en korrekt klokke; dette er bare et relativt avvik.

Det oppstår derimot et problem hvis man ønsker å sammenligne klokkeslett fra TELOC og AnnaLyse med hverandre. AnnaLyse oppgir for eksempel at toget gikk 53 sekunder for sent i

forhold til avgangen i publikumsruten som var 12:03. Dette skulle med andre ord bety at toget passerte signalet på vei ut av stasjonen 12:03:53. En måling for samme tog fra TELOC viser at toget begynte å gå igjen 12:03:24. Forutsatt at begge målingene ble gjort med samme klokke ville man kunne si at toget brukte 29 sekunder fra det var i bevegelse til det passerte signalet på vei ut av stasjonen. Men det finnes ikke noe grunnlag for å si at begge systemene følger samme klokke. Dermed kan man ikke gjøre slike sammenligninger med de to, men må holde seg til parametere som er absoluttverdier. Dette er en klar svakhet hvis man ønsker å bruke AnnaLyse som sammenligningsgrunnlag.

5.2.1.7 TOLKBARHET

Det er i utgangspunktet vanskelig å vurdere en subjektiv parameter som tolkbarheten siden den må sies å være ekstremt avhengig av hvilke forventninger man har.

I 2.7.2.1 blir det nevnt to hensyn man kan ta for å vurdere tolkbarheten av dataene: manipulerbarhet og egnethet. Selv om dataene slik de blir presentert i AnnaLyse systemet kan gi informasjon om for eksempel punktligheten og årsaker, har det i denne oppgaven vært nødvendig å manipulere dataene for å beregne parametrene i denne oppgaven. Data fra AnnaLyse kan dermed kalles manipulerbare. Egnetheten til oppgavens bruksområde er og god.

5.2.1.8 TILGJENGELIGHET

Tilgangen til dataene fra AnnaLyse må kunne sies å være relativt bra. Denne oppgaven er skrevet med utgangspunkt i hvilke data som er tilgjengelig fra NSB Drifts ståsted. AnnaLyse systemet gir brukere med rettigheter i denne avdelingen, tilgang til data som i stor grad blir samlet inn av JBV. Dataene er presentert på en form som gjør de lettfattelig for personer uten nødvendigvis inngående kunnskap om toggangen. På den måten kan man si at AnnaLyse systemet bidrar til å øke tilgjengelighet for NSBs ansatte til disse dataene.

For å kunne bruke dataene til å beregne de parametrene som er blitt gjort i denne oppgaven må man derimot både ha en viss kunnskap om hvordan de er framkommet samt prosessere de videre for å få de på ønsket format. Systemet kan heller ikke sies å ha alle de ønskede funksjonene særlig med tanke på eksportering av data. En videreutvikling av systemet, gjerne ved hjelp av brukerundersøkelser, kunne ha gjort AnnaLyse mer fleksibelt og tilgjengelig for alle formål.

5.2.2 TELOC

5.2.2.1 ANTALL OBSERVASJONER

Antallet observasjoner er på samme som for AnnaLyse avhengig av stasjonen. I Tabell 8 og Tabell 9 kan man se det antallet observasjoner som var tilgjengelig fra TELOC for den perioden det ble jobbet med i denne oppgaven. Siden hver eneste kjøreløgg kun gir bevegelsen til ett togsett er det begrenset for hvor mange observasjoner man kan få tilgang til fra kun én kjøreløgg; derfor det noe beskjedene antallet observasjoner i forhold til AnnaLyse.

Fordelen med TELOC i forhold til AnnaLyse, er at den inneholder data for stasjonsopphold for alle stasjoner i pendelen der toget stopper. I AnnaLyse er det ikke alle stasjoner som

rapporterer inn data, mens for TELOC vil man ha tilnag til alle start og stopp forutsatt at systemet fungerer normalt.

5.2.2.2 OPPLØSING

Oppløsningen i TELOC er meget bra; ned til tiendedels sekund på tidspunkter. Det er her nøydt seg med å jobbe med sekunder og har derfor rundet av til nærmeste hele sekund. Farten oppgis i km/t med to desimalers nøyaktighet og avstanden i kilometer med 5 desimalers nøyaktighet.

Generelt kan man si at oppløsningen i TELOC gir brukeren stor fleksibilitet i forhold til bruksområde. Dette skyldes det som man diskuterte under avsnitt 2.7.2.1 om at oppløsningen kan ha flere tolkninger. Hvis man så på oppløsningen som antall målinger over tid eller distanse, ville TELOC komme veldig bra ut også i forhold til en slik definisjon. AnnaLyse derimot har svært lav oppløsning i forhold til et slikt syn da systemet kun gir målinger på faste punkt.

5.2.2.3 NØYAKTIGHET

For å kunne bruke TELOC som beregningsgrunnlag for gjennomsnittlige størrelser er det som for AnnaLyse nødvendig å ekskludere ekstreme verdier som indikerer noe unormalt. Som man kan se av Tabell 8 og Tabell 9 er nødvendig å fjerne opp mot 20 % av verdiene.

Hvis man derimot ønsker å analysere de unormale verdiene, er TELOC svært nøyaktig og inneholder få eller ingen feilregistreringer.

5.2.2.4 FULLSTENDIGHET

Tabell 19 viser det teoretisk antallet observasjoner fra den tilgjengelige kjøreløgen fra TELOC og fullstendigheten. Fullstendigheten varierer med stasjonen og gjennomsnittverdien kan sies å være henholdsvis 95 % for kjøretiden og 97 % for oppholdstiden.

Stasjon	Teoretisk Antall Observasjoner	Antall Observasjoner Kjøretid	Antall Observasjoner Oppholdstid	Fullstendighet Kjøretid	Fullstendighet Oppholdstid
Brakerøya	31	30	30	97 %	97 %
Lier	31	29	29	94 %	94 %
Asker	31	27	28	87 %	90 %
Sandvika	31	28	31	90 %	100 %
Lysaker	31	29	27	94 %	87 %
Skøyen	31	29	31	94 %	100 %
Nationaltheatret	31	31	31	100 %	100 %
Oslo	31	31	31	100 %	100 %
Lillestrøm	31	31	31	100 %	100 %
Leirsund	31	30	30	97 %	97 %
Frogner	31	30	31	97 %	100 %
Lindeberg	31	31	31	100 %	100 %
Kløfta	31	30	30	97 %	97 %
Gardermoen	31	30	31	97 %	100 %
Eidsvoll Verk	31	30	30	97 %	97 %
Eidsvoll	31	27		87 %	
Gjennomsnitt	31	30	30	95 %	97 %

Tabell 19 Fullstendighet TELOC.

5.2.2.5 ENSARTETHET

Dataene fra TELOC er målt på samme måte for et togsett over en bestemt tidsperiode. De er med andre ord veldig ensartede. Man kunne tenkt seg en situasjon der man hadde kombinert flere kjøreløgger. Man måtte i det tilfellet tatt hensyn at man hadde forskjellig måleutstyr.

5.2.2.6 TIDSRIKTIGHET

Data fra TELOC er alltid avhengig at man har tilgang til en kjøreløgg. Som det er forklart tidligere i denne oppgaven er dataene tilgjengelig ved periodisk vedlikehold av et togsett. Tidsriktigheten kan dermed sies å være lav i forhold til for eksempel AnnaLyse der man kan hente ut de dataene man ønsker til enhver tid.

5.2.2.7 TOLKBARHET

Manipulerbarheten ved dataene fra TELOC er god, selv om det krever en god del innsats. Dataene er ikke lagd for bruksområdet til denne oppgaven, men kan likevel manipuleres slik at egnetheten økes. Uten bearbeiding krever dataene en viss grad av forhåndskunnskap for å kunne gi noen mening.

5.2.2.8 TILGJENGELIGHET

Data fra TELOC stammer som nevnt flere ganger tidligere fra kjøreløggen til ulike tog. Sett fra et NSB Drift ståsted er ikke dataene veldig tilgjengelige. Kjøreløggen kan hentes ut i sammenheng med vedlikehold og må bestilles fra Mantena. Det er vanskelig å få eksakt de dataene man vil ha da et togsett kan ha beveget seg på mange strekninger og må kryssjekkes etterpå mot logg for rullende materiell for å finne ut hvilke tognummer toget kjørte som. Man

dermed ikke på samme måte som for AnnaLyse hente ut akkurat de dataene man vil ha for spesifikke tog, strekninger, tidspunkt stasjoner osv.

Videre kompliseres tilgangen ved at man er avhengig av andre data; materiellturnering; for å kunne bruke dataene. Henting av eksterne data kan være vanskelig både med tanke på at man involverer andre avdelinger og at man ikke nødvendigvis er klar over hvilke data som finnes.

Dataene fra TELOC kan også sies å kreve en stor del prosessering før de kan brukes som analysegrunnlag. Denne prosessen er svært arbeidskrevende og kan generere feil i datamaterialet.

Den største innvendingen mot å bruke TELOC slik den er blitt brukt i denne oppgaven er at den er rett og slett ikke egnet til formålet. Datakilden kan gi oss svært nøyaktige og høyoppløselige data om både oppholdstid og kjøretid, men det finnes ikke noen egnet programvare for å kunne ta ut disse dataene. Dermed er man avhengig av en stor grad manuelt arbeid som ofte skaper feil.

5.3 ER DATAENE HENSIKTSMESSIGE?

I dette avsnittet diskuteres gyldigheten og påliteligheten av dataene.

Ingen datakilde vil kunne betraktes som særlig god hvis den ikke er stand til å gi oss den informasjonen vi søker. I det forrige avsnittet ble karakteristika ved AnnaLyse og TELOC identifisert i forhold til noen måleparametere for datakvaliteten, for å kunne si noe på generelt basis om godheten av de to datakildene. Man ønsker nå med bakgrunn i bruksområdet i kapittel 4, å diskutere om dataene virkelig er egnet til å gi oss den informasjonen vi søker.

AnnaLyse som system vil ikke direkte produsere den informasjonen vi er ute etter å finne i kapittel 4. Det kreves en del bearbeiding før man har de på den formen man ønsker. Men på en generell basis bør man betrakte data fra AnnaLyse som godt egnet som rådata for produksjon av den informasjonen vi ønsket. Dette skyldes i hovedsak at datakilden kommer relativt godt ut i forhold til de ulike dimensjonene av datakvalitet som er diskutert i det forrige kapitlet. Problemet ligger som sagt i AnnaLyse-systemets begrensning på prosesseringen av dataene; det vil si en dataforbruker er, hvis han ønsker å beholde sin rolle som dataforbruker, avhengig av de prosesseringene AnnaLyse og eventuelt andre foregående systemer har gjort med rådataene. Ønsker man å manipulere og prosessere dataene videre må man nødvendigvis forholde seg til dette faktum. Et eksempel på dette paradokset kan være forholdet mellom TIOS og AnnaLyse, der de er basert på samme rådata, men systemet prosesserer dataene på forskjellig måte med forskjellig oppløsning.

TELOC er en datakilde som er konstruert med en helt annen hensikt enn det man benytter den til i denne oppgaven: den skal gi en oversikt over togets bevegelser. At man kan bruke kjøreløgen som grunnlag for å beregne oppholdstid, ankomstforsinkelse og kjøretid er å betrakte som et biprodukt. Dette ligger til grunn for alt arbeid med datakilden i denne oppgaven; det kreves en stor grad bearbeiding for å få ut akkurat de dataene man ønsker. På den måten kan påstå at TELOC som datakilde ikke er veldig hensiktsmessig i den betydningen at den ikke er konstruert for å kunne brukes til vårt bruksområde. Men på en annen side vil den ved bearbeiding produsere høyoppløselige, nøyaktige, fullstendige og ensartede data som er sammenlignbare med planlagte verdier.

I denne oppgaven har man ikke hatt mulighet til å beregne parametere fra de to datakildene som sammenlignes direkte. Dette skyldes eventuelle forskjeller i referansetiden som det blir målt i forhold til. For å kunne gjøre slike sammenligninger ville det vært nødvendig å forsikre seg om målingene virkelig var gjort ved samme tidspunkt. Man kan dermed si at de to datakildene er uegnet for en slik sammenligning. Et annet moment ved TELOC er at dataene vil være hentet fra en kjøreløgg som begrenser antall tog man kan hente fra for eksempel en strekning. Det betyr at det er vanskelig å bruke data fra TELOC som trenddata uten å hente inn veldig mange kjøreløgger, noe som igjen er veldig ressurskrevende.

5.4 SAMMENDRAG

Dette kapitlet har diskutert kvaliteten av datakildene TELOC og AnnaLyse i forhold til 8 forskjellige dimensjoner av datakvalitet; antall observasjoner, oppløsning, nøyaktighet, fullstendighet ensartethet, tidsriktighet, tolkbarhet og tilgjengelighet; samt om dataene er hensiktsmessige til sitt bruksområde.

AnnaLyse-systemet gjør et bredt datagrunnlag lett tilgjengelig fra et NSB Drift ståsted. Dataene kan kjennetegnes ved at de inneholder et stort antall observasjoner, har oppløsning på sekunds nivå for stasjoner med automatisk registrering fra signalanlegg, har en anslått nøyaktighet på 80 % og fullstendighet på i overkant av 90 %. En rimelig antakelse er derfor at datakilden kan betegnes som å ha høy kvalitet.

Systemets svakhet ligger i at det blandes forskjellige typer data fra manuelle og automatiske registreringer. Dataforbrukeren må dermed inneha en del kunnskap om måleprosessen for å kunne benytte seg av dataene. Videre må man ta høyde for forskjellige måletidspunkt ved sammenligning med andre datakilder.

Hensiktsmessigheten til dataene fra AnnaLyse har for denne oppgavens bruksområde vært god, men AnnaLyse som system produserer i dag ikke de parametrene man har analysert i denne oppgaven. Det betyr at dataene måtte prosesseres fra rådata til det ønskede produktet. Dette kan være med på å generere feil samt at det krever en dypere forståelse av dataenes opprinnelse.

Dataene fra TELOC ga grunnlag for et noe lavere antall observasjoner enn AnnaLyse. Dette kommer av at en kjøreløgg som dataene stammer fra, følger et togsett som kan gå på mange ulike strekninger, noe som medfører at man ikke har muligheten til å hente data for akkurat de tidsperiodene man ønsker. Fullstendigheten ble estimert til å være noe høyere enn for TELOC; om lag 96 %. Oppløsningen og ensartetheten var meget god på de dataene som ble benyttet.

TELOC ligger utenfor NSB Drifts dataflyt og må kombineres med andre kilder som også ligger på utsida av den regulære datastrømmen. Videre er dataene avhengig av en god del prosessering før de kan betraktes som egnet for denne oppgavens bruksområde.

6 KONKLUSJON

I dette kapitlet blir oppgavens konklusjon presentert i forhold til oppgavens problemstilling slik den ble definert og avgrenset i kapittel 1.1. Videre vil man diskutere feilkilder og begrensninger ved oppgaven, grad av måloppnåelse og forslag til videre studier.

6.1 OPPGAVENS KONKLUSJON

Opgavens problemstilling ble i kapittel 1.1 spesifisert til følgende 5 punkter:

A å beskrive og diskutere ulike perspektiv og måleparametere for datakvalitet, samt datakvalitetens eventuelle betydning for beslutninger.

B å beskrive toggangen mellom Drammen og Eidsvoll i perioden 22.08-05.09.2005 med hensyn på oppholdstid, kjøretid, og ankomstforsinkelse.

C å lage en regresjonsmodell med en avhengig variabel (oppholdstid), og to uavhengige variabler (enkelt/dobbeltsett og passasjerantall) for Asker stasjon, med den hensikt å se på forklaringsparametere for oppholdstiden.

D å foreslå noen måleparametere for datakvalitet, subjektive og objektive, basert på bruk av to datakilder og teoretisk grunnlag, samt gjøre et anslag for kvaliteten av datakildene.

E å diskutere resultatene og kommer med forslag til videre studier.

Det kommer i det følgende til å bli diskutert de relevante funn i forhold til problemstillingen og eventuelle konklusjoner som kan trekkes.

Det kan hjelpe struktureringen av produksjonen av kvalitetsdata å adoptere en prosessanalogi. Denne analogien hjelper oss til å dele prosessen i 3 roller: datainnsamler, databasebestyrer og dataforbruker, eller som 3 underprosesser: innsamling av data, prosessering av data og forbruk av data. Videre er det essensielt at den programmerte datakvaliteten i høyest mulig grad imøtekommer den oppfattede kvaliteten. Dette kan oppnås ved å kartlegge bruksområdet for dataene best mulig. For NSB Drifts del bør dataforbrukerne kravspesifisere sitt bruksområde i form av ønskede metrikker og karakteristikker ved dataproduktet. Disse spesifikasjonene kan så bringes videre bakover i prosesskjeden slik at man tilpasser målingene til dataforbrukerens behov. En slik forbedret innfrielse av dataforbrukerens behov vil være med på å styrke den subjektive dimensjonen av datakvaliteten. Eventuelle endringer i dataproduksjonsprosessen bør gjøres så tidlig så mulig av hensyn til ressursbruk og fleksibilitet.

Den objektive dimensjonen av datakvaliteten kan styrkes ved å gjøre målinger i henhold til de 6 foreslåtte parametrene: antall observasjoner, oppløsning, nøyaktighet, fullstendighet, ensartethet og tidsriktighet. Data som brukes i NSB drift bør ledsages av verdier for disse parametrene for å kunne si noe om påliteligheten av de. Dette kan være med på å hjelpe en beslutningstaker å forstå gyldigheten av sin beslutning. Men å styrke den objektive datakvaliteten vil ikke nødvendigvis være hensiktsmessig i alle tilfeller; en slik økning må balanseres med behovet til forbrukeren for å sikre seg mot unødvendig ressursforbruk.

NSB Drift bør søke nye datakilder utenfor den regulære datastrømmen og etterstrebe at disse gjøres lett tilgjengelig. Det bør etableres rutiner som sikrer tilgangen til alle de ulike datakildene som er brukt i arbeidet med denne oppgaven. Det kan med fordel kartlegges

måleprosessen som ligger til grunn for datakildene bedre for å øke kunnskapen om påliteligheten. Informasjon om hvordan dataene kom til hjelper dataforbrukeren i sitt arbeid.

Data fra AnnaLyse systemet og TELOC kan brukes til å beregne gjennomsnittlige verdier for oppholdstiden, kjøretiden og ankomstforsinkelsen. Datakildene er derimot uegnet slik de framstår i dag, til å gjøre en direkte sammenligning av korresponderende verdier. Dette skyldes at man som dataforbruker ikke sitter med nok kunnskap om måleprosessen, og dermed om den relative kalibreringen av måleutstyret. Som nevnt over kan man ved å inkludere informasjon om dataens opprinnelse eventuelt eliminere denne svakheten.

Analysen av strekningen Drammen – Eidsvoll viste at oppholdstiden fra TELOC generelt viste en tendens til å ligge over det planlagte. Ankomstforsinkelsen så ut til å avta mot slutten av pendelen. Dette kan i viss grad sees i sammenheng med at den faktiske kjøretiden er mindre enn den planlagte mot slutten, noe som kan antyde at man kjører inn forsinkelser mot slutten. Disse forholdene bør undersøkes nærmere. Oppholdstiden fra AnnaLyse består ikke nødvendigvis bare av komponenter som skyldes forholdet ved stasjonene. Dette tydet det store avviket fra oppholdstiden fra TELOC på. Det anbefales at man gjør andre målinger hvis man ønsker å få pålitelige verdier for oppholdstiden.

Oppholdstiden ved en stasjon består av mange komponenter og påvirkes igjen av mange faktorer. I denne oppgaven var ikke det tilgjengelige datamaterialet; verdier for oppholdstiden fra TELOC og AnnaLyse, passasjerantall og enkelt/dobbeltsett; kvalitetsmessig bra nok for å kunne konstruere regresjonsmodeller for å undersøke dette forholdet. Man anbefaler på grunnlag av arbeidet med denne oppgaven, å samle inn nye data hvis man ønsker å undersøke nærmere variablene som innvirker på oppholdstiden. En kombinert kvantitativ og kvalitativ metode bør anvendes. Det tilgjengelige datamaterialet var ikke nok til å kunne belyse dette forholdet på en tilfredsstillende måte.

Både AnnaLyse og TELOC kommer relativt bra ut i forhold til de foreslåtte parametrene for datakvaliteten, men det ligger en utfordring i å tilpasse dataproduktet til brukeren. For TELOC kan man vurdere om det er mulig å hente andre typer data fra samme kilde. TELOC produserer i dag ikke data med tanke på en NSB Drift ansatts behov. Ved å samarbeide tettere med datainnsamlere om måleprosessen kunne man tilpasset målingene bedre til alle parters behov. Prosesseringen og lagringen av data bør standardiseres og rutineres for å unngå subjektive tolkninger av målinger.

AnnaLyse trenger å videreutvikles som system med tanke på bruksområdet. Med det menes det at forbedringsinnsatsen blir fokusert på det siste leddet i dataproduksjonsprosessen. Systemet henter data fra en rikholdig database; utfordringen ligger i å levere et dataprodukt bedre tilpasset forbrukeren.

6.2 FEILKILDER OG BEGRENSINGER

Det er mulig å oppleve mange kilder til feil i arbeidet med en masteroppgave. Noe grovt er det mulig å dele disse feilkildene inn i to hovedkategorier: 1. Feil knyttet til metode 2. Feil knyttet til gjennomføring.

Det første punktet innbefatter momenter som: Har man valgt rett problemstilling? Har man valgt rett metode i forhold til problemstillingen? Bruker man representative data? I denne oppgaven har både spesifiseringen av problemstillingen og metodetilnærming i stor grad blitt

bestemt av hvilke data man hadde til rådighet. For å kunne diskutere kvaliteten på datakildene var det nødvendig å benytte allerede eksisterende målinger. Videre gir den begrensede tidsrammen på oppgaven (20 uker) føringer på valg av metode. Men som det har blitt diskutert underveis i oppgaven kunne en kombinasjon av kvantitativ og kvalitativ tilnærming gitt bedre innsyn i faktorer som påvirker oppholdstida.

Feil knyttet til gjennomføring kan være om for eksempel om man har gjort metodefeil. Disse feilene kan elimineres ved å sammenligne resultater med andre verdier for så sjekke deres gyldighet. I denne oppgaven har det blitt forsøkt å gjøre slike sammenligninger i den grad det var mulig for å kunne eliminere slike feilkilder. Hvis andre sammenlignbare resultater ikke eksisterer er man avhengig av å validere resultatene med egne eller andres erfaring. Forfatterens manglende innsikt i enkelte områder av jernbanedriften kan ha ført til at slike feil ble stående ukorrigert.

I tillegg til det som ble nevnt over knyttet det også en feilkilde til tid. Denne oppgaven er gjennomført innefor en begrenset tidsramme noe som kan gi utslag i at man ikke har mulighet til å se utvikling over lang tid.

Generelt anbefales det at resultatene fra denne oppgaven leses med de nevnte feilkildene som bakgrunn for å ha et kritisk blikk.

6.3 FORSLAG TIL VIDERE STUDIER

Fra litteraturstudiet i kapittel kan det foreslås at det utforskes nærmere området TDQM. Dette er en spennende analogi til TQM som allerede er et velkjent og anerkjent konsept. NSB kan med fordel studere nærmere hvilke parametere for datakvaliteten de finner mest relevante og forsøke å utvikle mål for disse. Slike parametere bør være overførbare uavhengig av bransje da de relaterer seg til datakvalitet; ikke produktkvalitet. En nærmere studie av hvilke parametere andre organisasjoner benytter kunne vært svært nyttig.

En brukerundersøkelse hadde vært tjenlig å gjennomføre for å kartlegge behovet til dataforbrukerne bedre. Med utgangspunkt i resultatet fra denne undersøkelsen kunne implikasjonene for prosessene i dataproduksjonsprosessen evalueres.

Hvilke faktorer som påvirker oppholdstida ved stasjonene bør studeres ved hjelp av en kombinasjon av kvalitativ og kvantitativ metode. Det datagrunnlaget som lå til grunn for denne oppgaven var ikke nok til å avdekke hvilke variabler som påvirker oppholdstida ved Asker stasjon, men det betyr ikke at det ikke kan brukes til å bevise andre sammenhenger. Men for å kunne gi et mest mulig helhetlig bilde bør en viss kvalitativ tilnærming adopteres.

Det bør undersøkes hvordan forskjellen i kalibrering er for de forskjellige målingene. Hvis man kunne etablert generelle verdier for avvik i kalibreringen ville det vært mulig å bruke det eksisterende datagrunnlaget for å gjøre sammenligninger. Videre kan man måle oppholdstida fra AnnaLyse med andre instrumenter for å bruke som referanse.

7 REFERANSER

1. Fisher CW & Kingma BR. Criticality of data quality as exemplified in two disasters. *Information & Management*. 2001; 39: 109-116.
2. Dagbladet Nett [hjemmeside på internett]Oslo: Dagbladet[oppdatert 15. august 2005; hentet 5. september 2005]. Tilgjengelig fra: <http://www.db.no>.
3. Aschehoug & Gyldendals Store Norske Leksikon, Online.
4. Bredrup H. 1995: *Performance Measurement in a Changing Competitive Industrial Environment: Breaking the Financial Paradigm*, Doktor Ingeniøravhandling, IPK, NTNU, Trondheim.
5. Redman TC. *Data Quality: the field guide*. 1. utg. USA: Digital Press; 2001.
6. Burton-Jones A. *Knowledge Capitalism: Business, Work, and Learning in the New Economy*. Oxford: Oxford University Press; 1999.
7. Lee Y W & Strong D M. Knowing-Why about data processes and data quality. *Journal of Management information Systems*. 2003-4 Vinter; 20(3): 13-39.
8. Wedde KJ. 1997: *Innsamling og vurdering av måledata*. SPIQ, Tilgjengelig fra: www.geomatikk.no/spiq/pulikasjoner/teknikknotater/innsamling1.pdf.
9. Aune A. *Kvalitetsdrevet ledelse kvalitetsstyrte bedrifter*. Oslo: Gyldendal Forlag; 2001.
10. Kvalfors T. *Kvalitetsutvikling i Bedrifter*. Oslo: Cappelen Akademiske Forlag; 1998.
11. NS – EN ISO 2000:9000.
12. Luktvaslimo Ø. *Faktabasert styring i jernbanedrift* [prosjektoppgave]. IPK: NTNU; 2005.
13. Garvin D A. *Managing Quality: the Strategic and Competitive Edge*. New York. Free Press; 1998.
14. Juran J M. *Quality Control Handbook*. 3. utg. New York: McGraw-Hill; 1974.
15. Pirsig R M. *Zen and the Art of Motorcycle Maintenance*. New York. Bantam Books; 1974.
16. Strong D M, Lee Y W & Wang R Y. Data Quality in context. *Communications of the ACM*. 1997 Mai; 40(5): 103-110.
17. Capiello C, Francalanci C og Pernici B. Data Quality Assesment from the User's Perspective. *Proceedings of the 2004 international workshop on Information Quality in Information Systems*. 2004: 68-73.
18. Wang R Y, Storey V C & Firth C P. A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering*. 1995 August; 7(4): 623-640.
19. Orr K. Data quality and systems. *Communications of the ACM*. 1998 Februar; 41(2): 66-71.
20. Bustinduy J. More quality in regional transport, *Public Transport '95, 51st International Congress Paris 1995*: International Commision on Regional Transport.
21. Wand Y & Wang R Y. Anchoring data quality dimensions in onthological foundations. *Communications of the ACM*. 1996 November; 39(11): 86-95.
22. Lee YW, Strong DM, Kahn BK & Wang RY. AIMQ: a methodology for information quality assessment. *Information & Management*. 2002; 40: 133-146.
23. Ballou D P & Pazer H L. Cost quality tradeoffs for control procedures in information Systems. *International Journal of Management Science*. 1987; 15(6): 509-521.
24. NS – ISO 8402.
25. Wang R Y. A product perspective on total data quality management. *Communications of the ACM*. 1998 February; 41(2): 58-65.

26. Gripsrud, G. *Metode og Dataanalyse: med Fokus på Beslutninger i Bedrifter*: Høyskoleforlaget, Kristiansand; 2004.
27. Eisenhardt KM & Zbaracki MJ. Strategic Decision Making. *Strategic Management Journal*. 1992; 13: 17-37.
28. Cohen MD, March JG and Olsen JP. Garbage Can Model of Organizational Choice. *Administrative Science Quarterly*. 1972; 17: 1-25.
29. Zeithaml AV, Parasuraman A, Berry L. Delivering quality service: balancing customer perceptions and expectations. New York. The Free Press; 1990.
30. Rosander AC. *Applications of Quality Control in the Service Industries*. New York. MARCEL DEKKER INC ASQC Quality Press; 1985.
31. Wang R Y, Kon H B & Madnick S E. Data quality requirements analysis and modelling. *IEEE Transactions on Knowledge and Data Engineering*. 1993; : 670-677.
32. Kahn B, Strong DM og Wang RY. Information Quality Benchmarks: Product and Service Performance. *Communications of the ACM*. 2002 april; 45(4): 184:92.
33. Wang RY & Strong DM. Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems*. 1996; 12(4): 5–34.
34. Ballou, D., Wang, R., Pazer, H., and Tayi, G. Modeling information manufacturing systems to determine information product quality. *Management Science*. 1998 April; 44(4): 462–84.
35. Karlöf, B.: Benchmarking: veiviser til forbedret produktivitet og kvalitet, Gyldendal, Oslo, 1993.
36. Andersen, B. & Pettersen, Per – Gaute. *Benchmarking: en praktisk håndbok*: TANO, Oslo; 1995.
37. Fearnley, N. 2004: ”Benchmarking European Sustainable Transport – Dokumentasjon av prosjektene BOB og BEST samt TØIs deltakelse”, Transportøkonomisk institutt, Oslo.
38. Tayi K G & Ballou D P. Examining data quality. *Communications of the ACM*. 1998 February; 41(2): 54-57.
39. Holme, IM & Solvang, BK. *Metodevalg og Metodebruk*: TANO, Oslo; 1991.
40. Yin, RK *Case Study Research - Design and Methods*: SAGE Publications, London; 2003.
41. Heinz, W. 2000: *Pasagerutbyte i tåg. Mätningar av av-och påstigningstider samt ansats til modell för att beskriva samband*, TRITA IP AR 00-86. Royal Institute of Technology, Stockholm.
42. Olsson, O. E. Nils & Sætermo, F. Inger – Anne & Røstad, C. Ch. 2002: SINTEF Rapport: *Konsekvensvurdering av anleggsarbeid i Vestkorridoren*, SINTEF Teknologiledelse.
43. Kerr AW, Hall HK, and Kozub, SA. *Doing Statistics with SPSS*: SAGE Publications, London; 2003.
44. Quesenberry CP. *SPC Methods for Quality Improvement*. 1. utg. Canada: Wiley & Sons INC.; 1997.
45. Zarkovich SS. *Quality of Statistical Data*. Roma: Food and Agriculture Organization of the United Nations; 1966.
46. Montgomery DC, & Runger GC. *Applied Statistics and Probability for Engineers*. 3. utg. New York: Wiley; 2003.
47. Naus JI. *Data Quality Control and Editing*. 1. utg. New York: Marcel Dekker; 1975.
48. Liepins GE & Uppuluri VRR. *Data Quality Control: Theory and Pragmatics*. 1. utg. New York: Marcel Dekker; 1990.

49. Spirer HF, Spirer L & Jaffe AJ. *Misused Statistics*. 2. utg. New York: Marcel Dekker; 1998.
50. Hines WW, Montgomery DC, Goldsman DM & Borror CM. *Probability and Statistics in Engineering*. 4. utg. Hoboken, N.J.: Wiley; 2003.
51. Chakravarti IM, Laha RG & Roy J. *Handbook of Methods of Applied Statistics: Techniques of Computation, Descriptive Methods and Statistical Interference*. USA: Wiley & Sons INC; 1967.
52. Wang YR, Reddy MP & Kon HB. Toward quality data: An attribute-based approach. *Decision Support Systems*. 1995; 13: 349-372.
53. Ragthunathan S. Impact of information quality and decision-maker quality on decision quality: a theoretical model and simulation analysis. *Decision Support Systems*. 1999; 26: 275-286.
54. Levitin AV & Redman TC. A model of the data (life) cycles with application to quality. *Information and Software Technology*. 1993 April; 35(4): 217-223.
55. Chakraborty S & Tah D. Real time statistical process advisor for effective quality control. *Decision Support Systems*. 2005; Under publising.
56. Klein BD. Detecting errors in data: clarification of the impact of base rate expectation and incentives. *Omega*. 2001; 29: 391-404.
57. Winkler WE. Methods for evaluating and creating data quality. *Information Systems*. 2004; 29: 531:550.
58. Nishisato S. Graphical representation of quantified categorical data: its inherent problems. *Journal of Statistical Planning and Inference*. 1995; 43; 121-132.
59. Chengular-Smith IN, Ballou DP & Pazer HL. The Impact of Data Quality Information on Decision Making: An Exploratory Analysis. *IEEE Transactions on knowledge and data engineering*. 1999 November-December; 40(6): 853-864.
60. Pierce EM. Assessing data quality with control matrices. *Communications of the ACM*. 2004 Februar; 47(2): 82-86.
61. Au G & Choi I. Facilitating implementation of total quality management through information technology. *Information & Management*. 1999; 36: 287-299.
62. Shankaranarayanan G. Supporting data quality management in decision-making. *Decision Support Systems*. 2005; Under publising.
63. Brodie M L. Data quality in information Systems. *Information and Management*. 1980; 3: 245-258.
64. Redman T C. The impact of poor data quality on the typical enterprise. *Communications of the ACM*. 1998 Februar; 41(2): 79-82.
65. Laudon K C. Data quality and due process in large interorganizational record systems. *Communications of the ACM*. 1986 Januar; 29(1): 4-11.
66. Morgan G & Smircich L. The Case for Qualitative Research. *Academy of Management Review*. 1980; 5(4): 491-500.
67. Wand Y & Weber R. On the Deep Structure of Information Systems. *J. Info. Syst*. 1995; 16(11): 203:223.
68. Wand Y & Weber R. an Ontological Model of an Information System. *IEEE Trans. Soft. Eng.*. 1990; 16(11): 1282:1292.
69. Leo PL, Lee YW, og Wang RY. Data Quality Assessment. *Communications of the ACM*. 2002 April; 45(4): 211-8.
70. Zarkovich SS. *Quality of Statistical Data*. Roma: Food and Agriculture Organization of the United Nations; 1966.

8 VEDLEGG

8.1 LOGG OVER RULLENDE MATERIELL

Logg over rullende materiell

august 2005

Set Number: 72-33

01	504-1605-1608-509-520-1617-1620-521-532-1629-1632-537-544-1641	1077 km
02	1604-505-516-1613-1616	415 km
03	545-508-1607-1610-511-522-1619-1622	615 km
04	12283-562-2102-1001-1006-11006-11590-10177-177-180-1025-1030-11030	431 km
05	10163-163-166-1011-1016-175-178-1023-1028-191-190-12241-112241	613 km
06	SUD	
07	SUD	
08	U12283-12283-562-2102-1001-1006-11006-11590-10177-177-180-1025-1030-11030	433 km
09	10163-163-166-1011-1016-175-178-1023-1028-191-190-12241	611 km
10	12283-562-2102-1001-1006-11006-11590-10177-177-180-1025-1030-11030	431 km
11	12209-1105-1114-1113-1122-1121-1130-1133	712 km
12	1102-1107-1116-1115-1124-1125-1132-1135-1140-2298	908 km
13	1145	59 km
14	MOS	
15	1102-1107-1116-1115-1124-1125-1132-1135-1140-2298-11581	915 km
16	12209-1105-1154-11154-12225-12236	274 km
17	12209-1105-1154-11154-12225-12225	237 km
18	112263-12263-514-1611-1614-515-524-1623-1626-531-538-1635-1638-543	927 km
19	12284-12253-512-1609-1612-513-524-1621-1624-527-12262	711 km
20	545-12252	93 km
21	12291-514-1611-1614-515-524-1621-1624-527-534-1631-1634-539-12296-541	886 km
22	504-1605-1608-509-520-1617-1620-521-532-1629-1632-537-544-1641	1077 km
23	1604-505-516-1613-1616-517-528-1625-1628-533-13310	815 km
24	12263-514-1611-1614-515-526-1623-1626-531-538-1635-1638-543	925 km
25	12284-12253-512-1609-1612-513-524-1621-1624-527-12262	711 km
26	545-508-1607-1610-511-522-1619-1622-523-534-1631-1634-539-542-1639	1035 km
27	1606-507-516-1613-1616-517-526-1623-1626-531-536-1633-1636-12240	1033 km
28	11550-1606-1609-1612-513-522-1619-1622-523-532	628 km
29	12997-3078	619 km
30	3005-3012-3017-3024-3029-3036-3041-3048-3053-3060-3065-3072-3077	951 km
31	3010-3211	75 km

Logg over rullende materiell

september 2005

Set Number: 72-33

01	3010-3211-3242-3043-3050-3051-3058-3059-3066-3067-3074-3075	450 km
02	13206	546 km
03	DRM	
04	DRM	
05	DRM	
06	DRM	
07	DRM	
08	DRM	
09	DRM	
10	DRM	
11	DRM	

Figur 27 Logg over rullende materiell, Tog ID 72-33.

8.2 RUTEHÅNDBOK

	Ankomst	Avgang	Km		
Drammen		38	52,86		
Brakerøya		40	50,76	2,1	
Lier		44	46,84	3,92	
Eriksrud		48	42,99	3,85	
Asker		53	23,83	19,16	
Hvalstad		56	20,19	3,64	
Billingstad		58	17,62	2,57	
Sandvika		2	14,14	3,48	
Høvik		5	10,72	3,42	
Stabekk		6	8,99	1,73	
Lysaker		7	7	1,99	
Skøyen		11	4,38	2,62	
Nationaltheatret		15	1,4	2,98	
Oslo	18	21	0,27	1,13	
Hellerud		26	6,2	5,93	
Lillestrøm	32	34	20,95	14,75	
Lillestrøm N		37	25,18	4,23	
Leirsund		39	26,94	1,76	
Frogner		42	29,8	2,86	
Lindeberg		45	32,28	2,48	
Kløfta	48	49	36,38	4,1	
Asper		52	40,3	3,92	
Langeland		54	42,22	1,92	
Gardermoen	59	0	51,85	9,63	
Bekkedalshøgda		6	62,35	10,5	
Eidsvoll Verk		7	63,3	0,95	
Venjar		9	65,74	2,44	
Eidsvoll		12	67,86	2,12	120,72

Tabell 20 Rutehåndbok

8.3 PASSASJERTELLING

Stasjon	Påhverdag	Avhverdag	Ombordhverdag	PåLørdag	AvLørdag	OmbordLørdag	PåSøndag	AvSøndag	OmbordSøndag
Drammen	39	16	85	28	12	56	24	5	48
Brakerøya	7	1	90	3	1	59	2	0	49
Lier	7	1	97	5	2	62	3	1	51
Asker	28	7	117	23	5	81	13	2	62
Sandvika	22	10	128	21	9	94	10	2	70
Lysaker	13	6	133	5	3	94	2	2	70
Skøyen	8	6	132	5	3	95	3	2	71
Nationaltheatret	27	59	98	18	39	97	13	22	62
Oslo S	50	53	89	40	44	76	33	31	63
Lillestrøm	12	40	62	6	32	72	5	20	48
Leirsund	0	5	58	0	2	46	0	2	46
Frogner	1	8	50	1	5	44	1	3	44
Lindeberg	0	4	47	1	2	40	0	1	43
Kløfta	1	16	33	1	11	39	1	7	37
Gardermoen	3	15	20	3	15	30	4	27	14
Eidsvoll Verk	0	12	9	0	10	18	0	5	9
Eidsvoll	0	10	0	0	8	8	0	9	0

8.4 TRAINPLAN

Stasjon	Ankomst	Avgang	Spesifisert	Spor	Just	Kjøretid
Drammen		10:38		3	-10	01:14
Holmen		10:39:04			-20	00:56
Brakerøya	10:39:40	10:40:00	00:00:20	2	1	01:44
Huseby		10:41:45			61	00:54
Lier	10:43:40	10:44	00:00:20		14	00:57
Sørumsåsen		10:46:11			30	01:19
Eriksrud		10:48		2	15	01:42
Solberg		10:49:57			15	01:47
Asker	10:51:59	10:53	00:01:01	3		00:40
H sign 4604		10:54:32				00:41
Skaugum		10:56:13				00:44
Åstaddalen		10:57:57				00:39
Lagerud		10:58:36				00:38
Taunumåsen		10:59:14				01:16
Sandvika	11:00:30	11:01	00:00:30	3	3	01:03
Engervannet		11:02:06			5	00:29

Blommenholm		11:02:40		2	5	00:05
Blommenholm BP		11:02:50			11	00:59
Høvik		11:04		2	-7	01:07
Stabekk		11:05		3	3	01:27
Lysaker	11:06:30	11:07	00:00:30	1	63	02:27
Skøyen	11:10:30	11:11	00:00:30	4	10	01:06
H sign 118		11:12:16			10	00:32
Elisenberg		11:12:58			10	00:27
Inkognigt		11:13:35			21	00:34
Nationaltheatret	11:14:30	11:15	00:00:30	4	14	00:18
H sign 130		11:15:32			20	02:08
Oslo S	11:18:00	11:21	00:03:00	10	12	01:21
H sign 249		11:22:33			20	00:28
H sign 197		11:23:21			20	00:43
H sign 191		11:24:24			20	01:16
Hellerud		11:26		1		01:13
Kjerinngmyrene		11:27:13			-4	00:50
Røykås		11:27:59			-4	00:50
Fjellsrud		11:28:45			-4	00:49
H sign 1301		11:29:30			-4	00:49
H sign 1321		11:30:15			-4	00:40
H sign 1341		11:30:51			-4	01:13
Lillestrøm	11:32	11:34	00:02:00	4	4	01:47
General Motor		11:35:51				01:09
Lillestrøm N		11:37		14	16	01:14
Leirsund	11:38:30	11:39	00:00:30		12	02:18
Frogner	11:41:30	11:42	00:00:30	2	24	02:06
Lindeberg	11:44:30	11:45	00:00:30	2	12	02:47
Kløfta	11:47:59	11:49	00:01:01	3	38	02:22
Asper		11:52		2	56	01:04
Langeland		11:54				01:01
H sign 16255		11:55:01				00:44
h sign 1623		11:55:45				00:41
Skåntjern		11:56:26				00:32
Olaløkka		11:56:58				00:33
H sign 1701		11:57:31				01:28
Gardermoen	11:58:59	12:00	00:01:01	1	10	02:00
H sign 1703		12:02:10			10	01:12
Rismyr		12:03:32			20	02:08
Bekkedalshøgda		12:06		1	-19	00:49
Eidsvoll Verk	12:06:30	12:07	00:00:30	1	27	01:33
Venjar		12:09		1	41	02:19
Eidsvoll	12:12			2		

Tabell 21 Utskrift fra Trainplan

8.5 AnnaLyse

8.5.1 TOGGANG OG DATAGRUNNLAG AnnaLyse

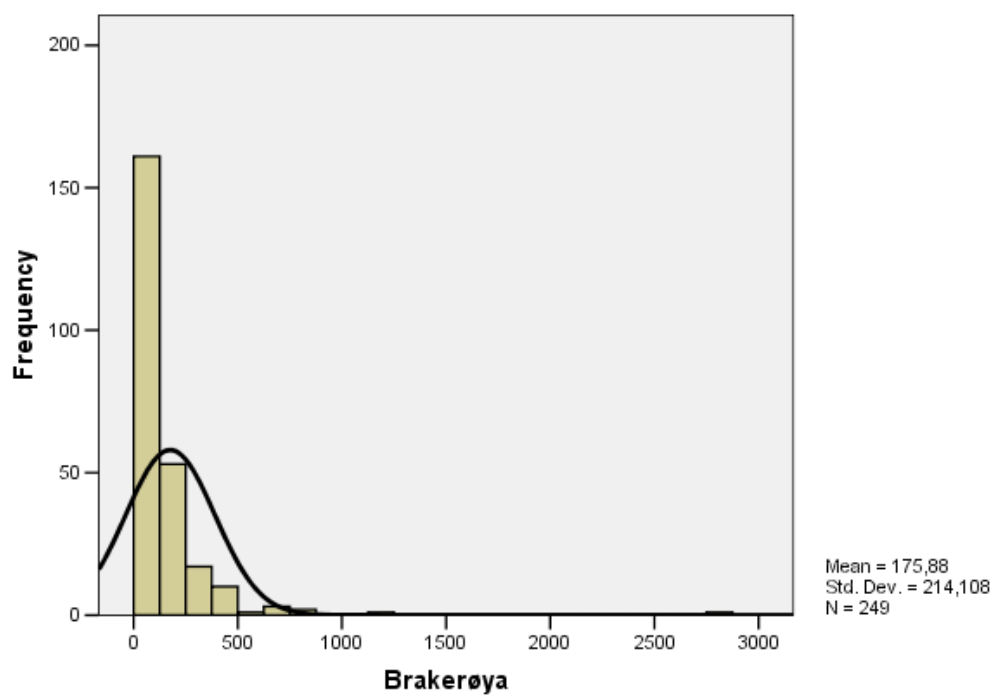
Stasjon	Mandag - Fredag	Lørdag	Søndag	Teoretisk Antall Observasjoner
Brakerøya	Alle tog	Ikke 1605,1639	Ikke 1605,1607,1639	275
Lier	Alle tog	Ikke 1605,1639	Ikke 1605,1607,1639	275
Asker	Alle tog	Ikke 1605,1639	Ikke 1605,1607,1639	275
Sandvika	Alle tog	Ikke 1605,1639	Ikke 1605,1607,1639	275
Lysaker	Alle tog	Ikke 1605,1639	Ikke 1605,1607,1639	275
Skøyen	Alle tog	Ikke 1605,1639	Ikke 1605,1639	277
Nationaltheatret	Alle tog	Ikke 1605,1639	Ikke 1605,1639	277
Oslo	Alle tog	Ikke 1605,1639	Ikke 1605,1639	277
Lillestrøm	Alle tog	Ikke 1605,1639	Ikke 1605,1639	277
Leirsund	Ikke data	Ikke data	Ikke data	0
Frogner	Alle tog	Ikke 1605,1639	Ikke 1605,1639	277
Lindeberg	Ikke data	Ikke data	Ikke data	0
Kløfta	Alle tog	Ikke 1605,1639	Ikke 1605,1639	277
Gardermoen	Alle tog	Ikke 1605,1639	Ikke 1605,1639	277
Eidsvoll Verk	Ikke data	Ikke data	Ikke data	0
Eidsvoll	Alle tog	Ikke 1605,1639	Ikke 1605,1639	277

Tabell 22 Toggang og datagrunnlag AnnaLyse.

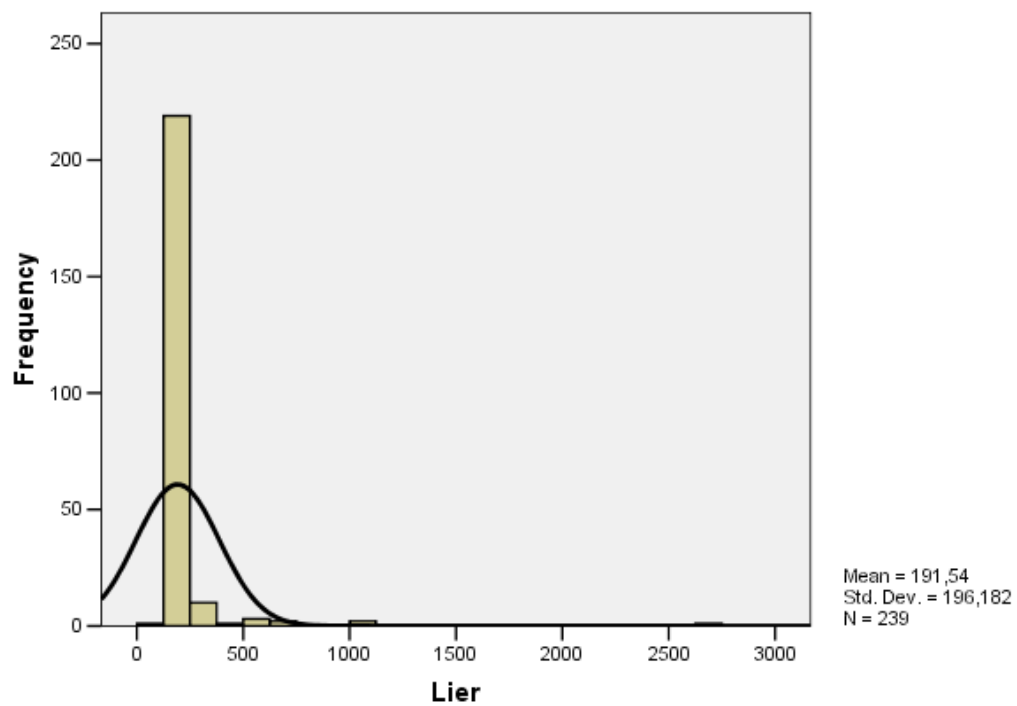
8.5.2 HISTOGRAMMER OPPHOLDSTID AnnaLyse

De følgende histogrammene viser oppholdstiden med hensyn på hver stasjon som det finnes data for i pendelen (jfr. Tabell 22).

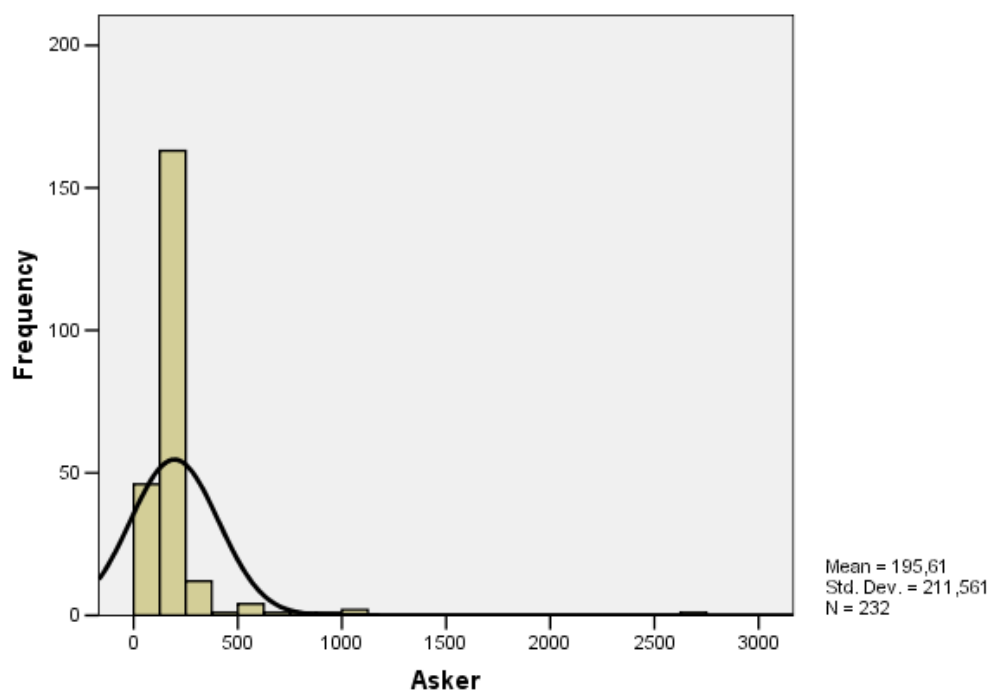
Histogram



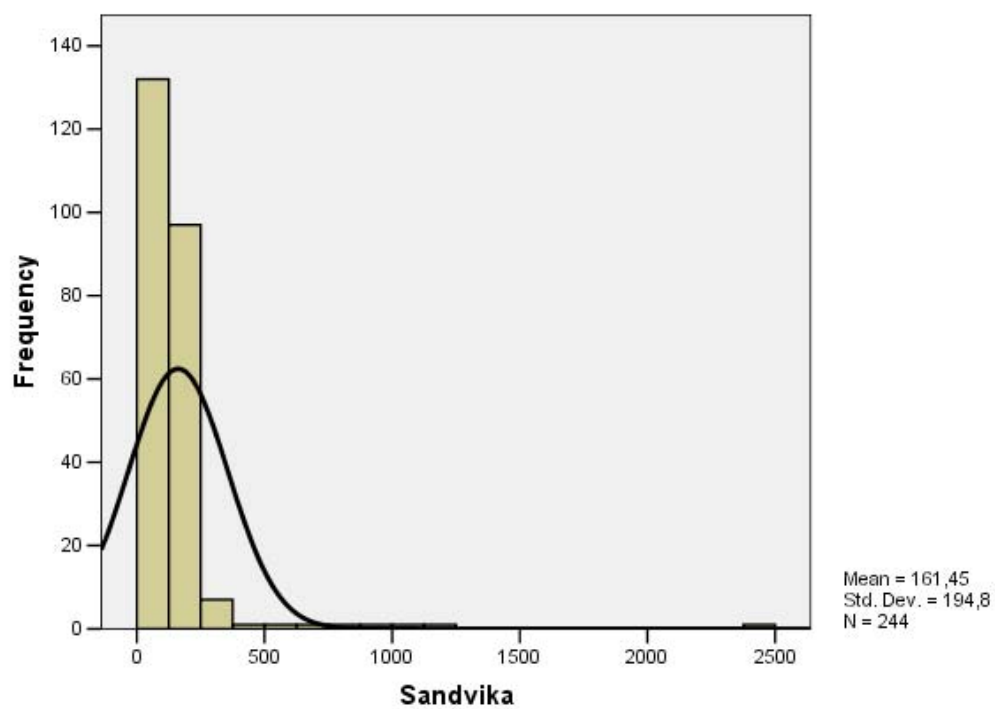
Histogram



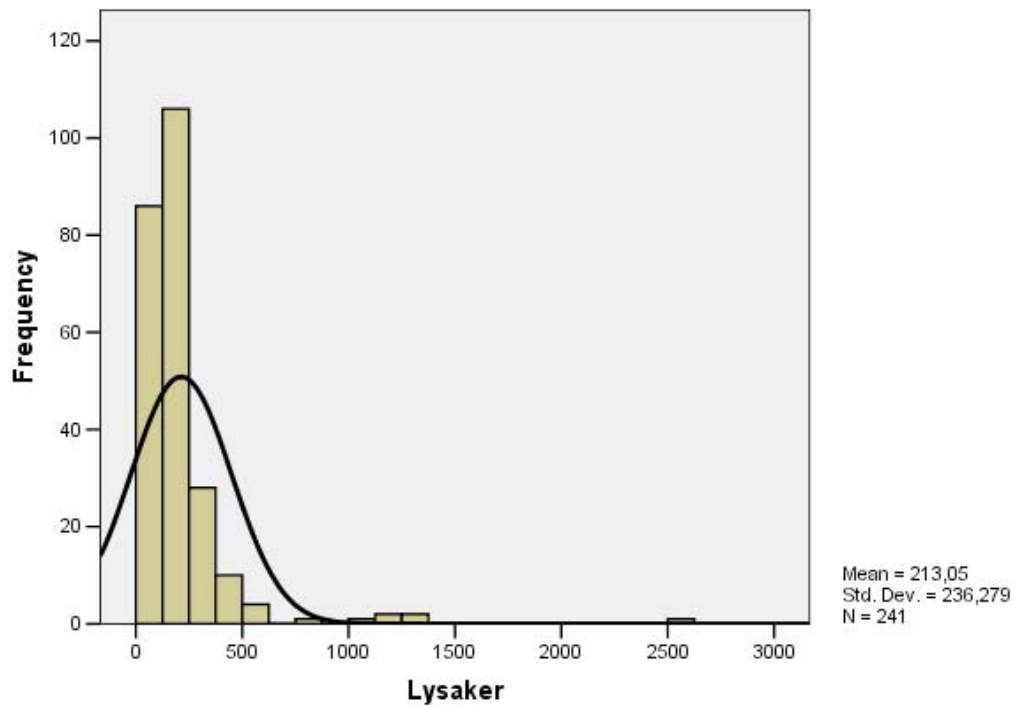
Histogram



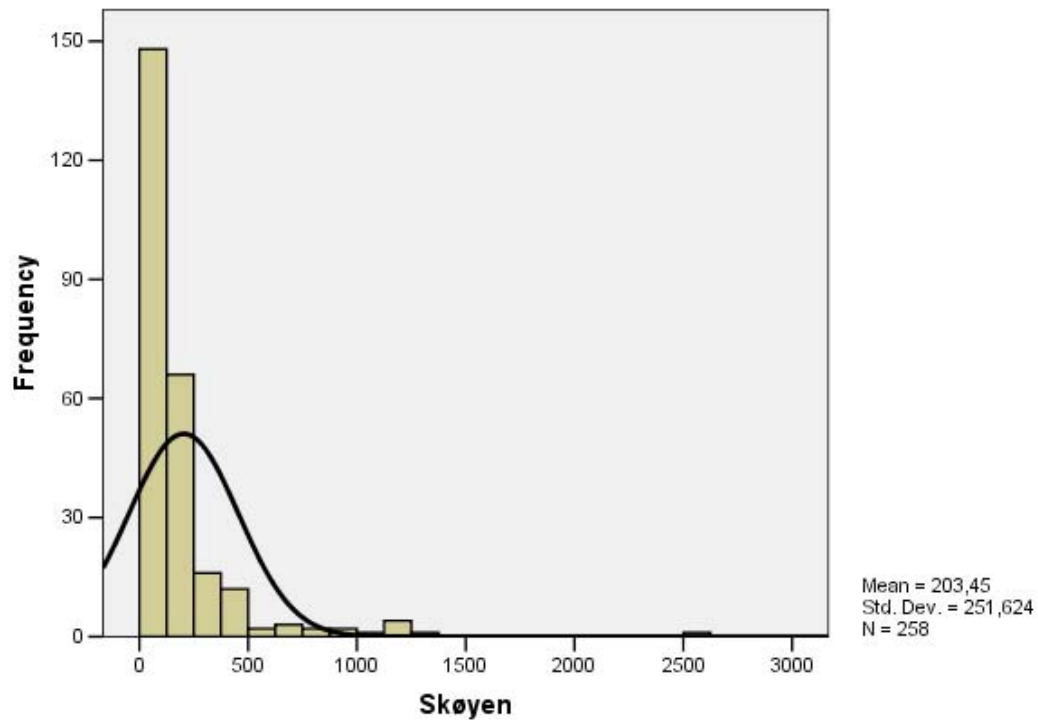
Histogram



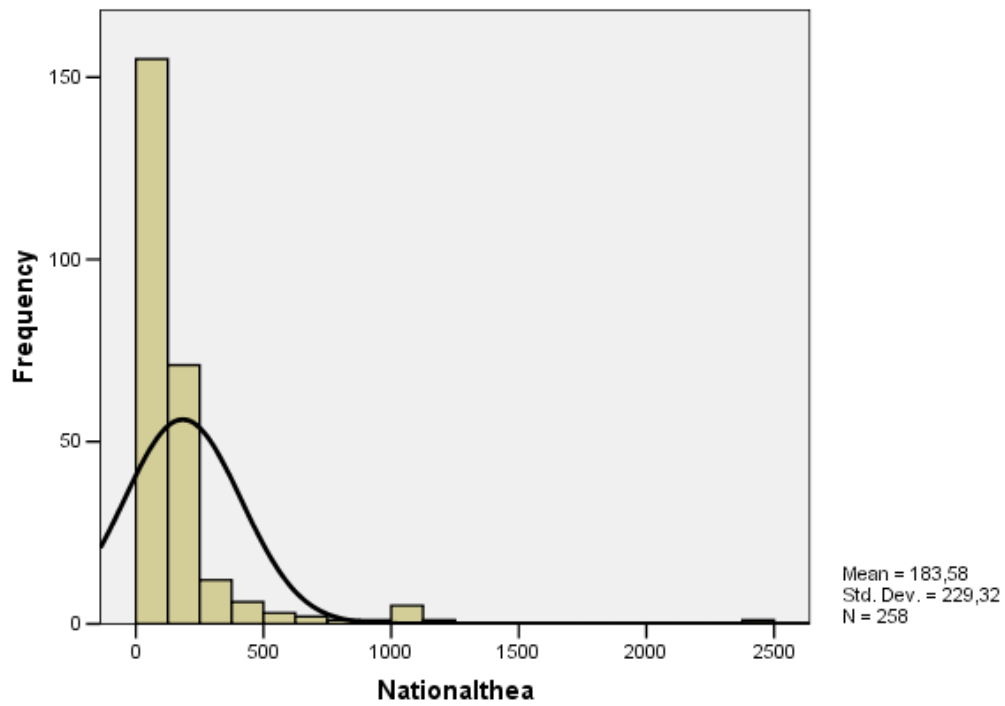
Histogram



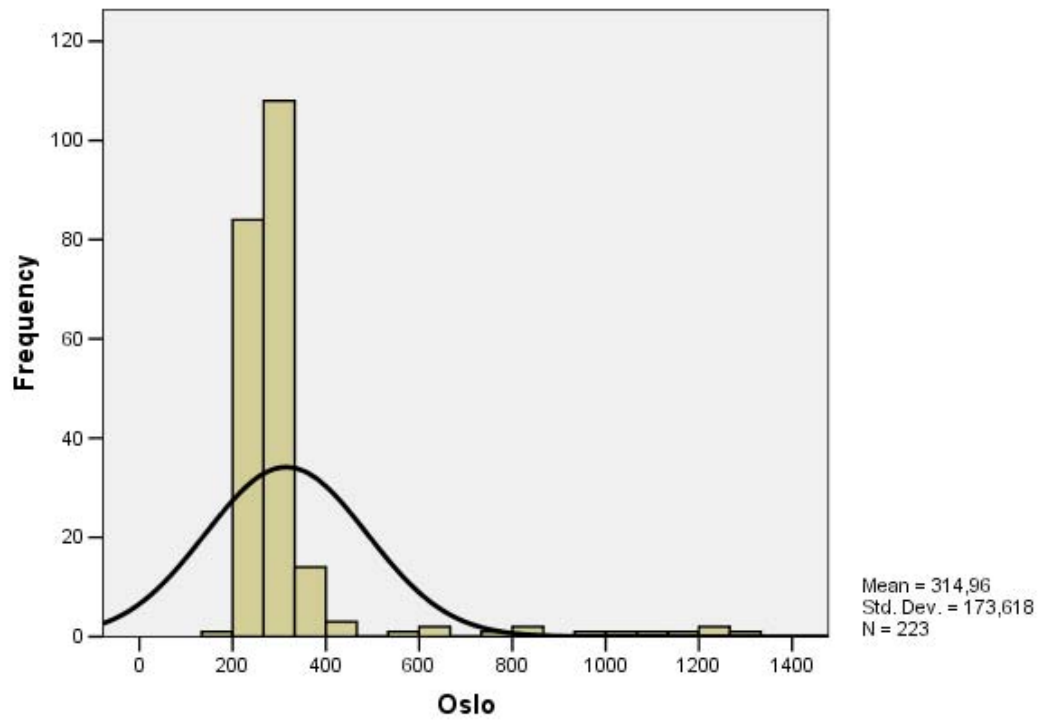
Histogram



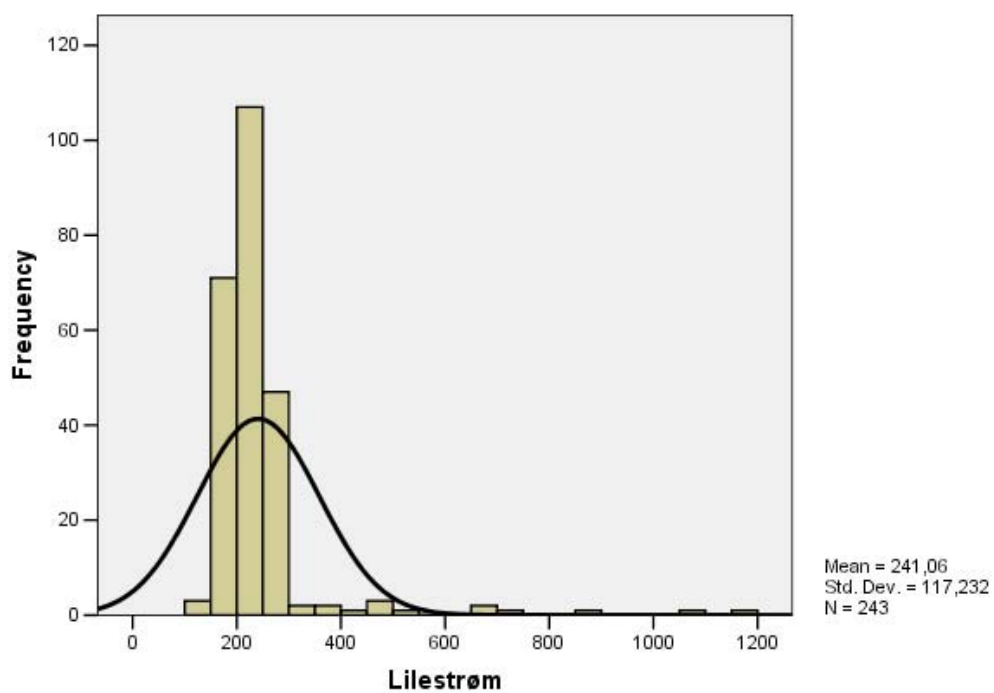
Histogram



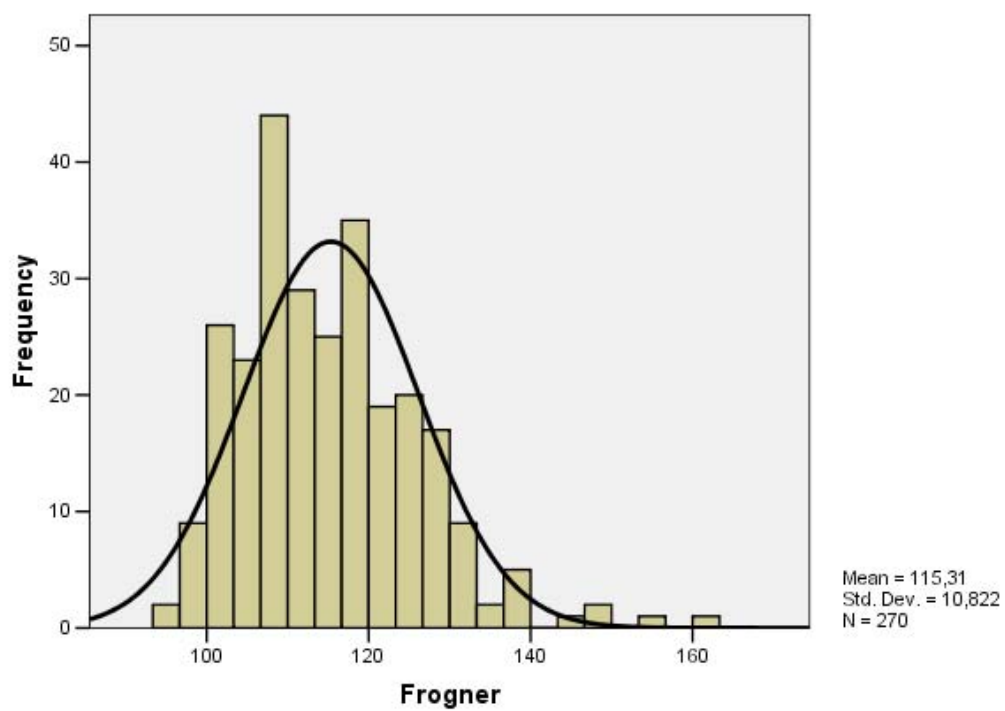
Histogram



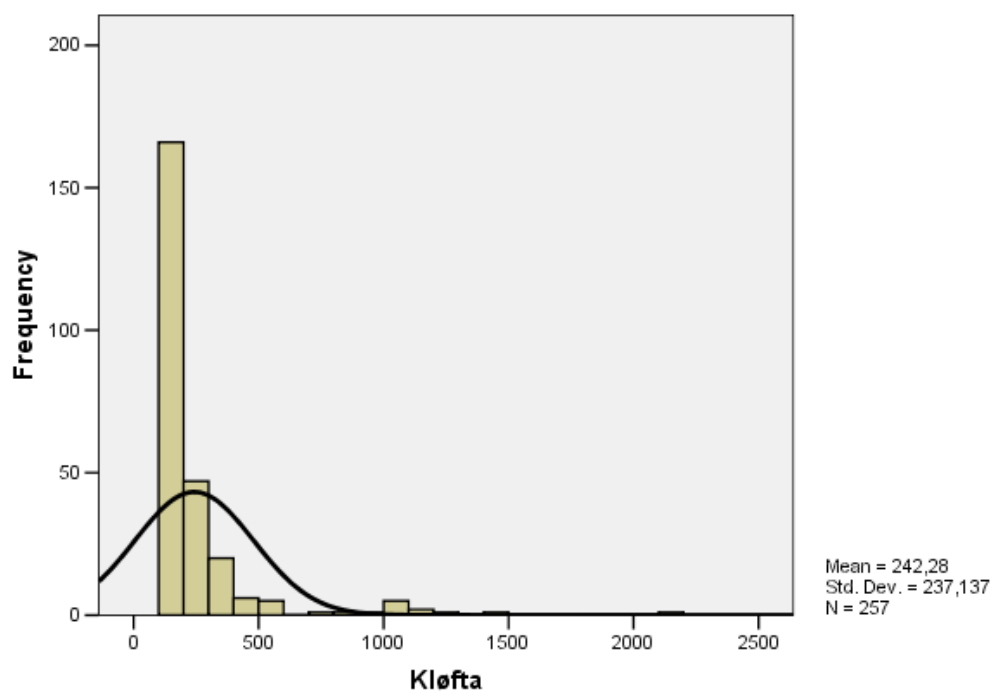
Histogram



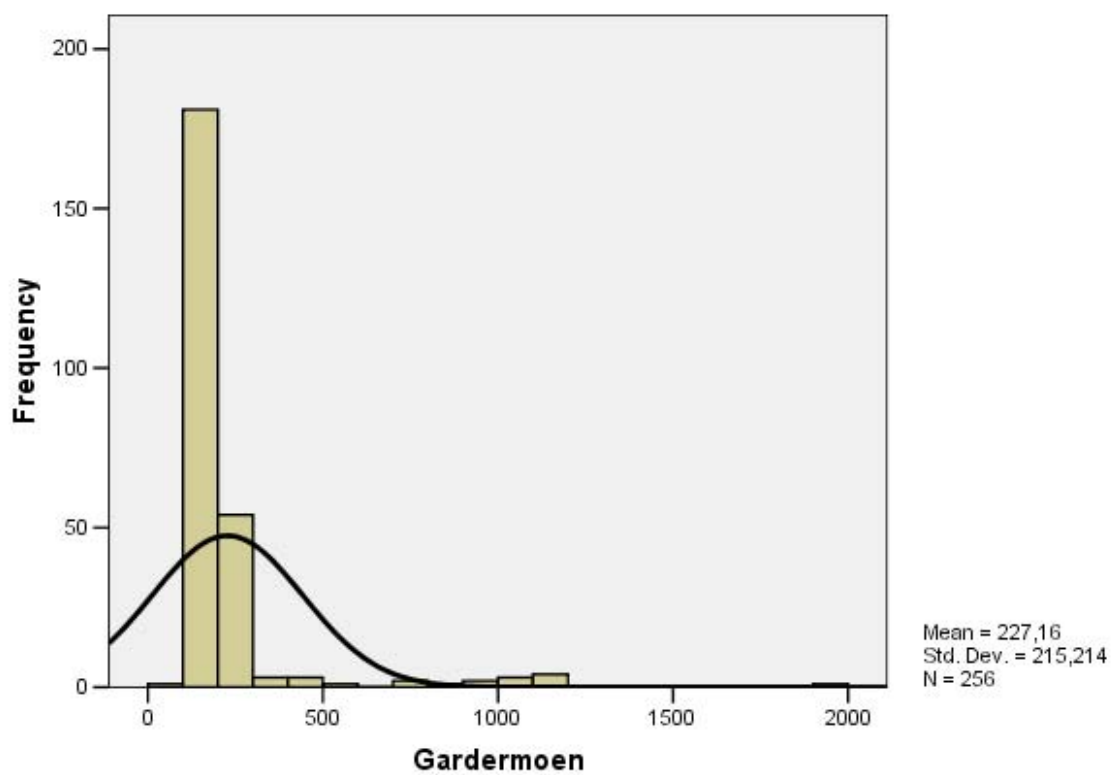
Histogram



Histogram

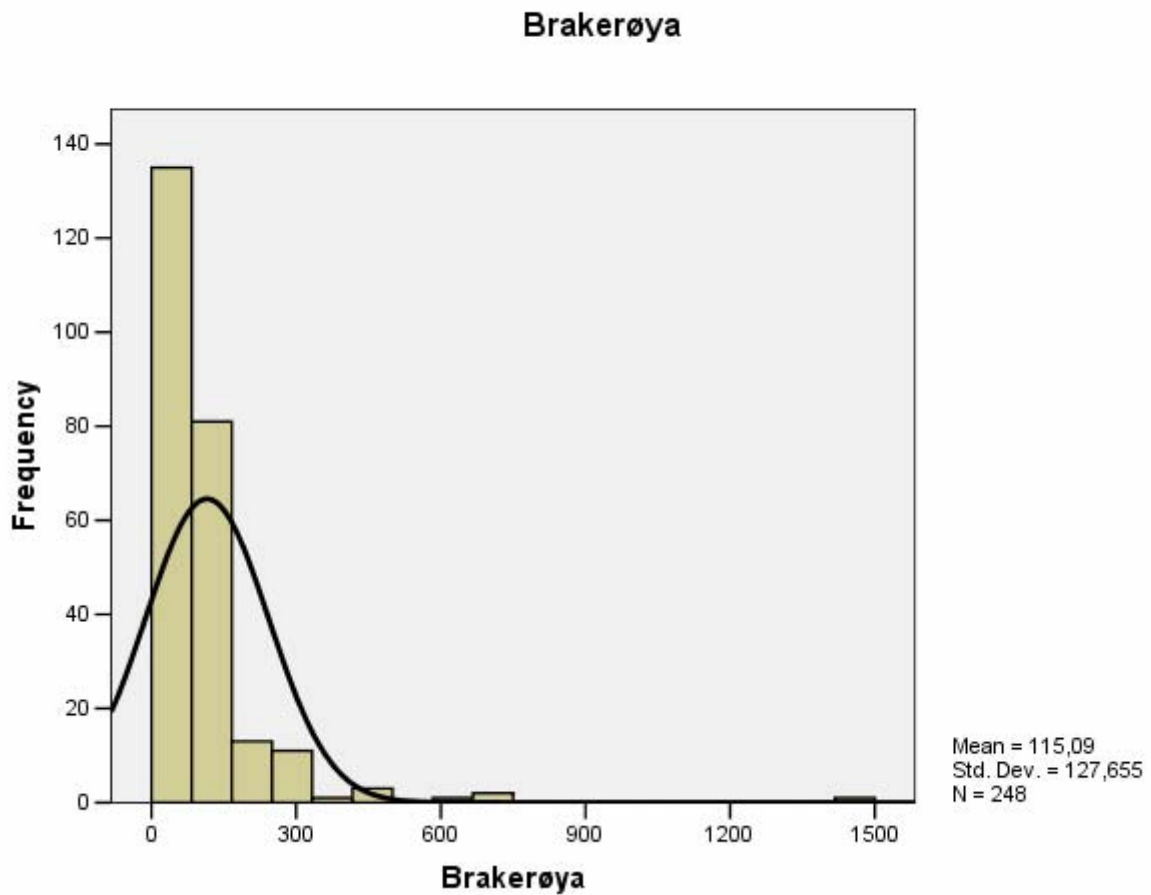


Histogram

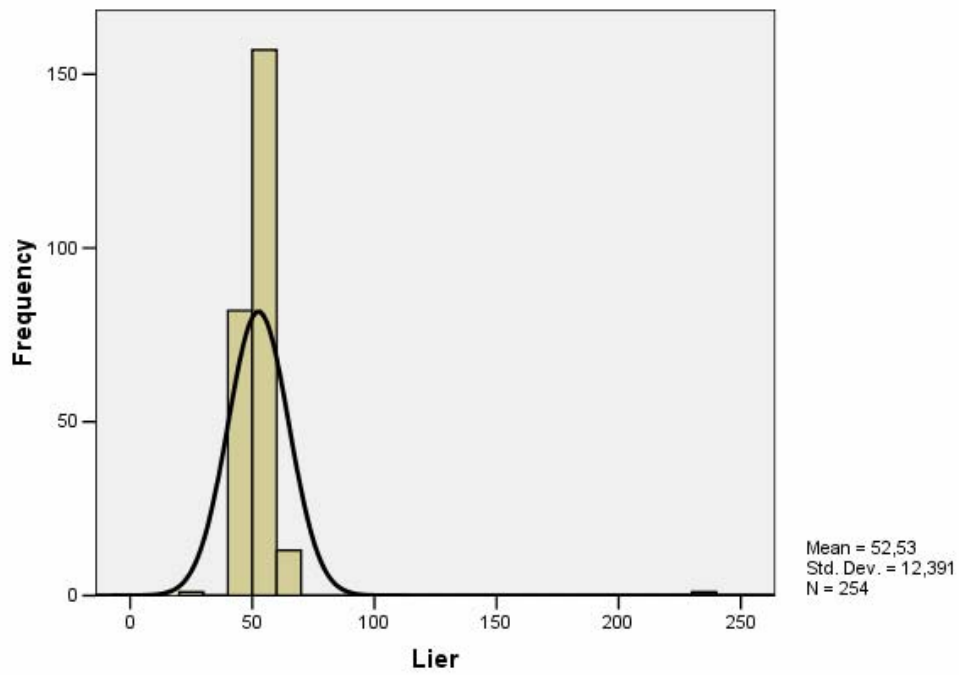


8.5.3 HISTOGRAMMER KJØRETID AnnaLyse

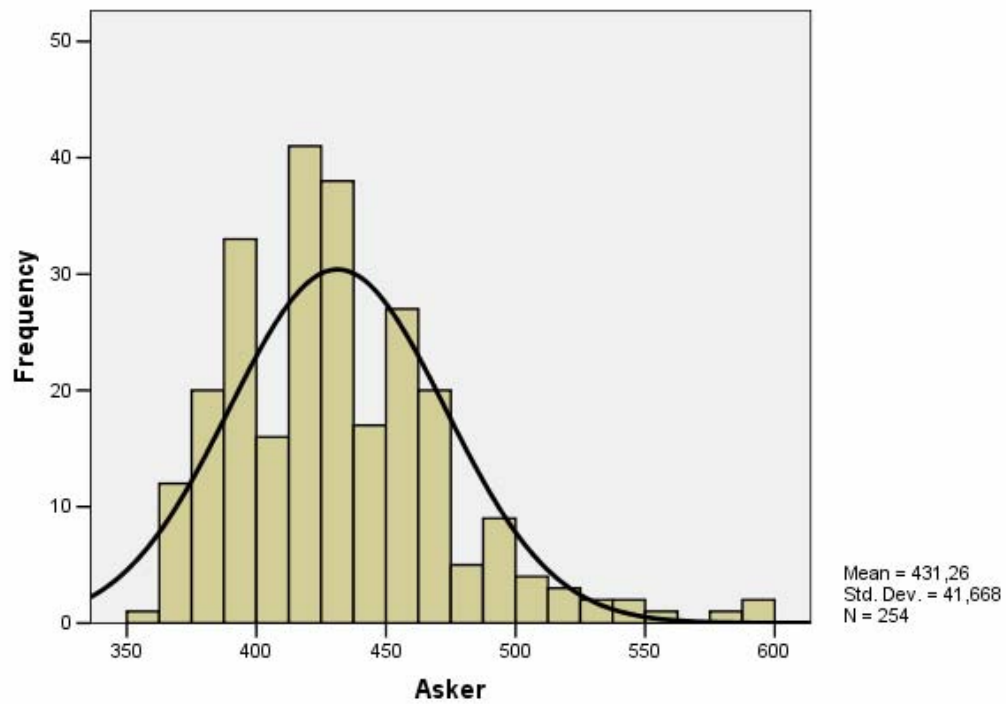
De følgende histogrammene visere kjøretiden til den indikerte stasjonen fra den foregående stasjonen som det finnes registreringer for i AnnaLyse systemet (jfr. Tabell 22).



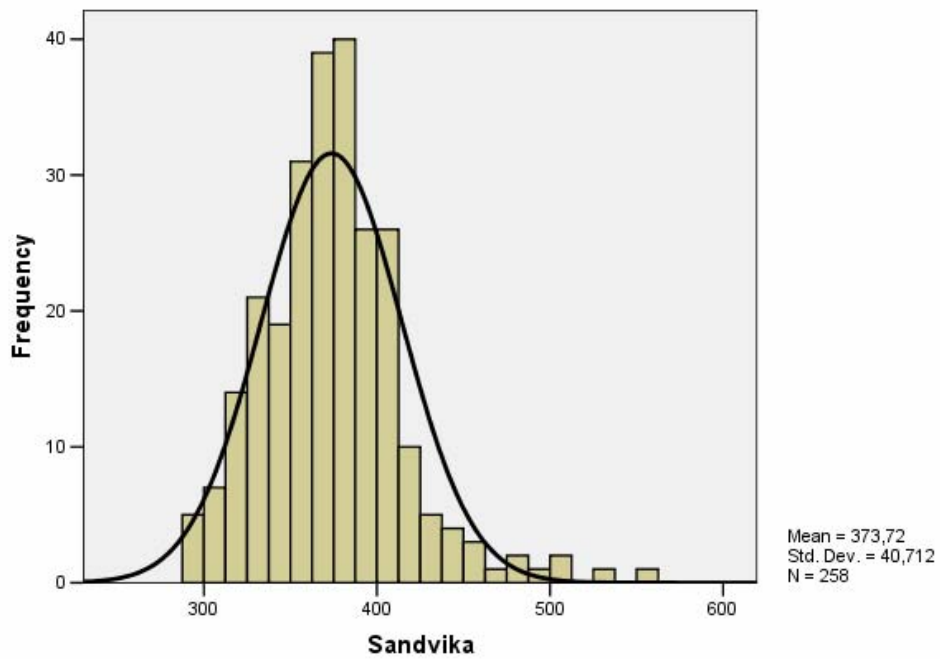
Lier



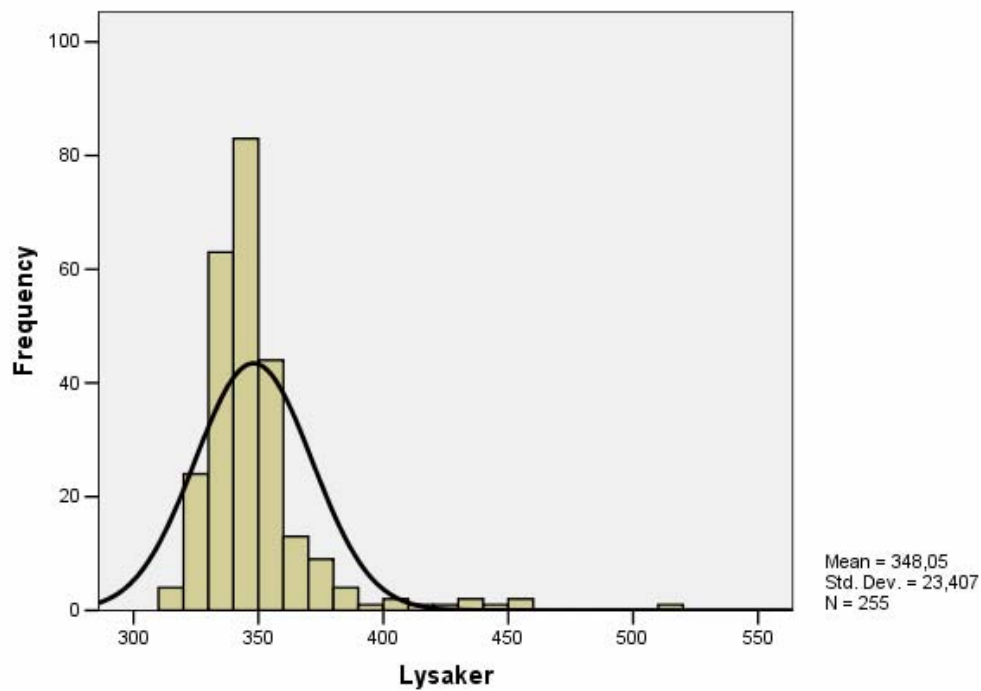
Asker



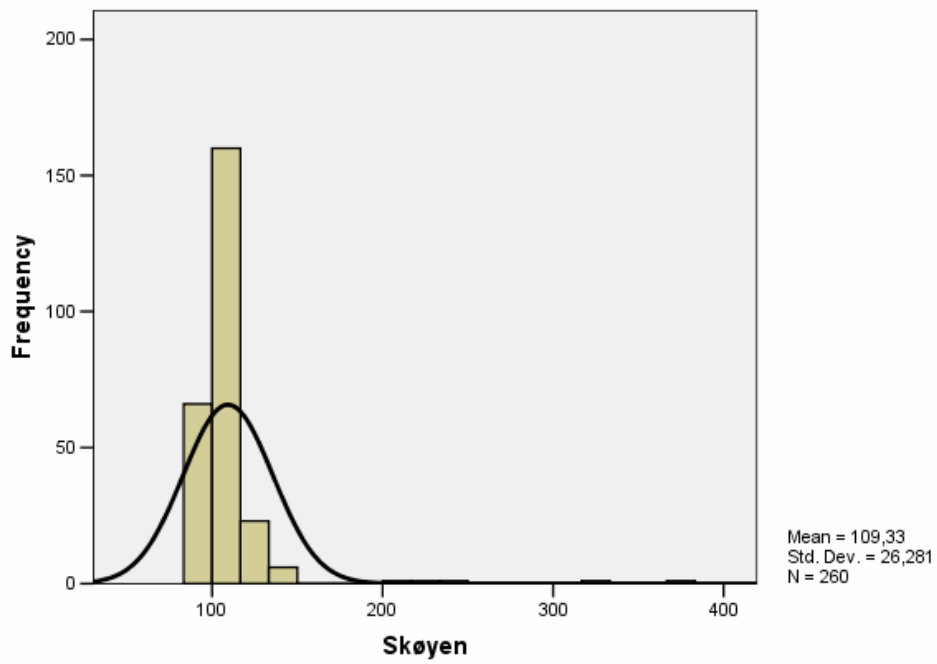
Sandvika



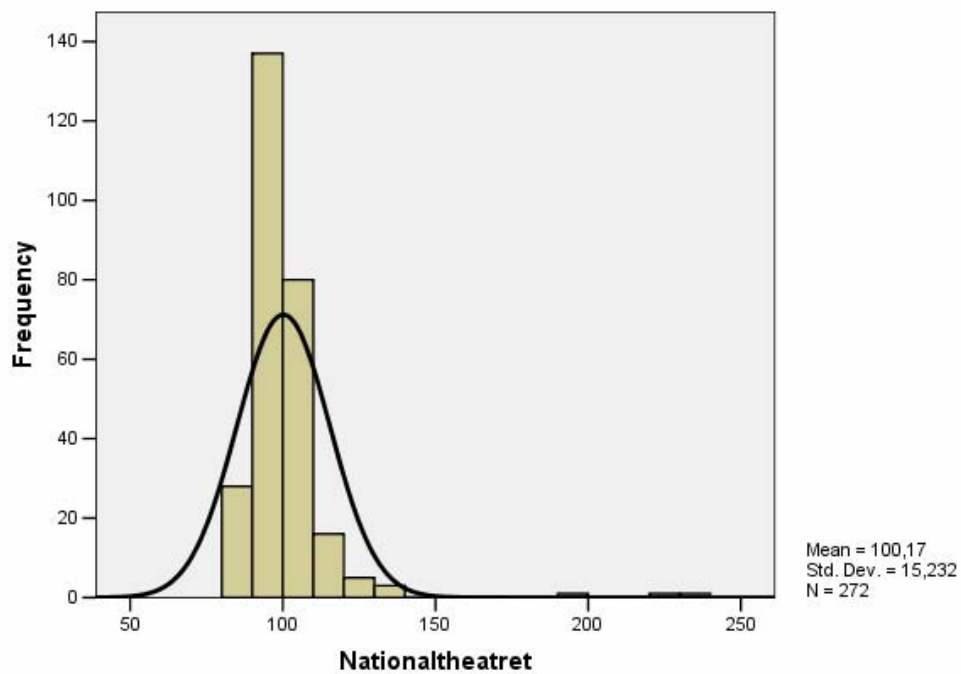
Lysaker



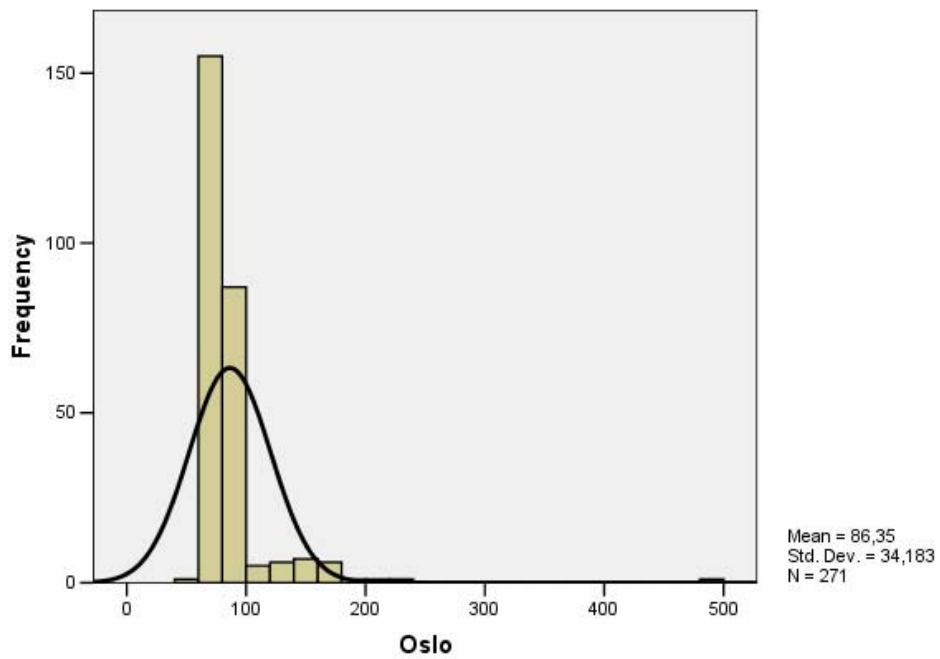
Skøyen



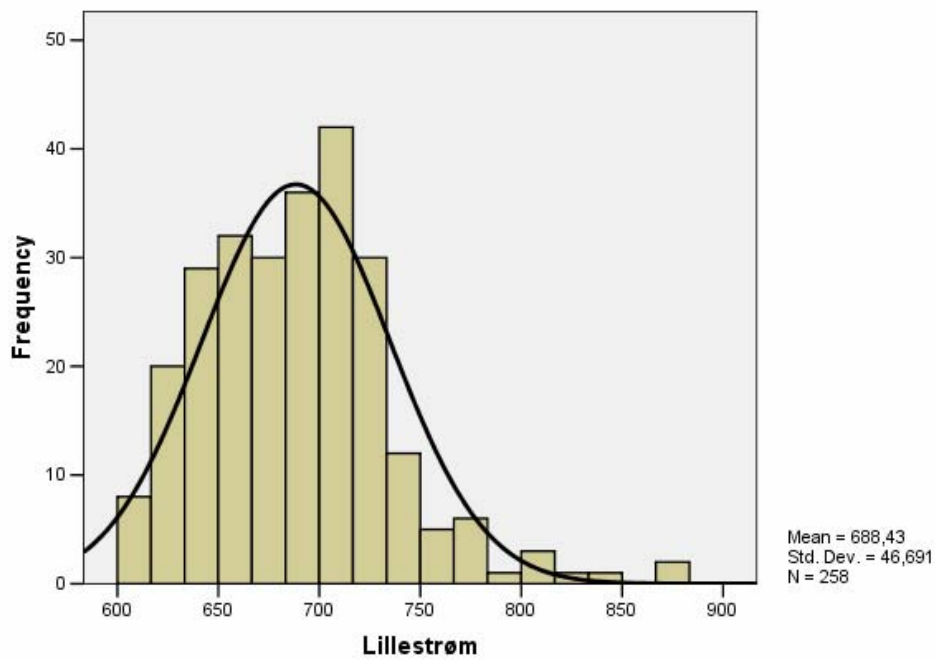
Nationaltheatret



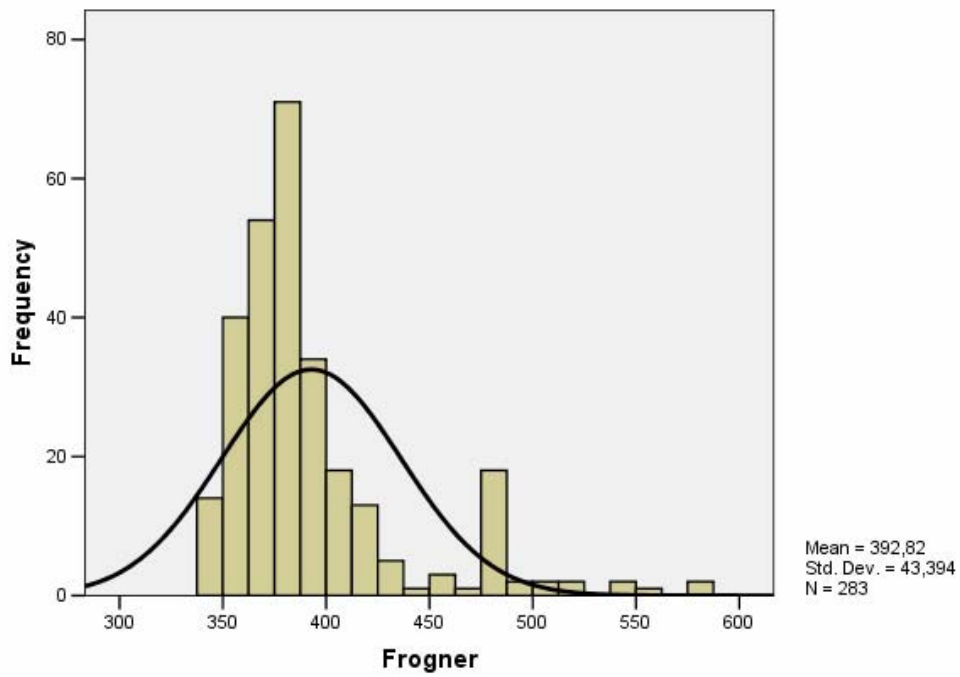
Oslo



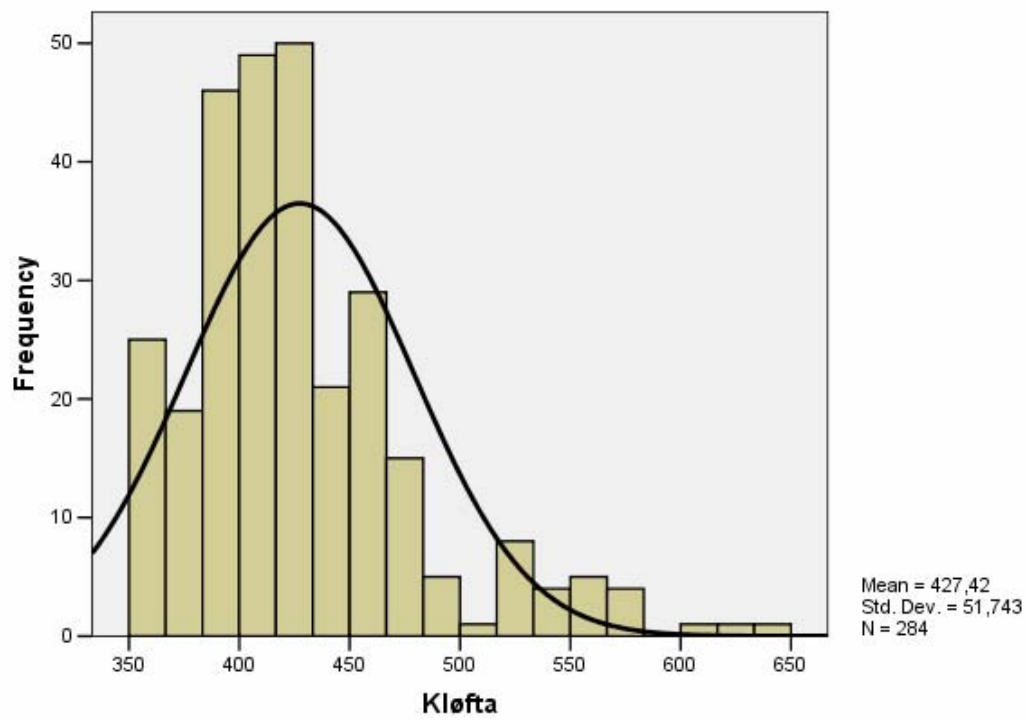
Lillestrøm



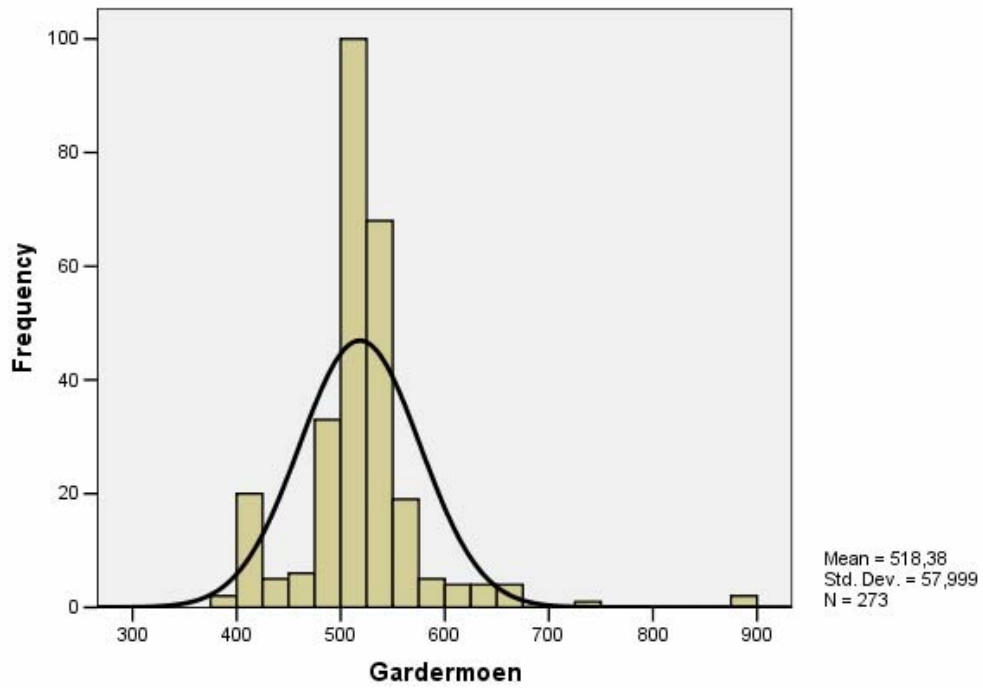
Frogner



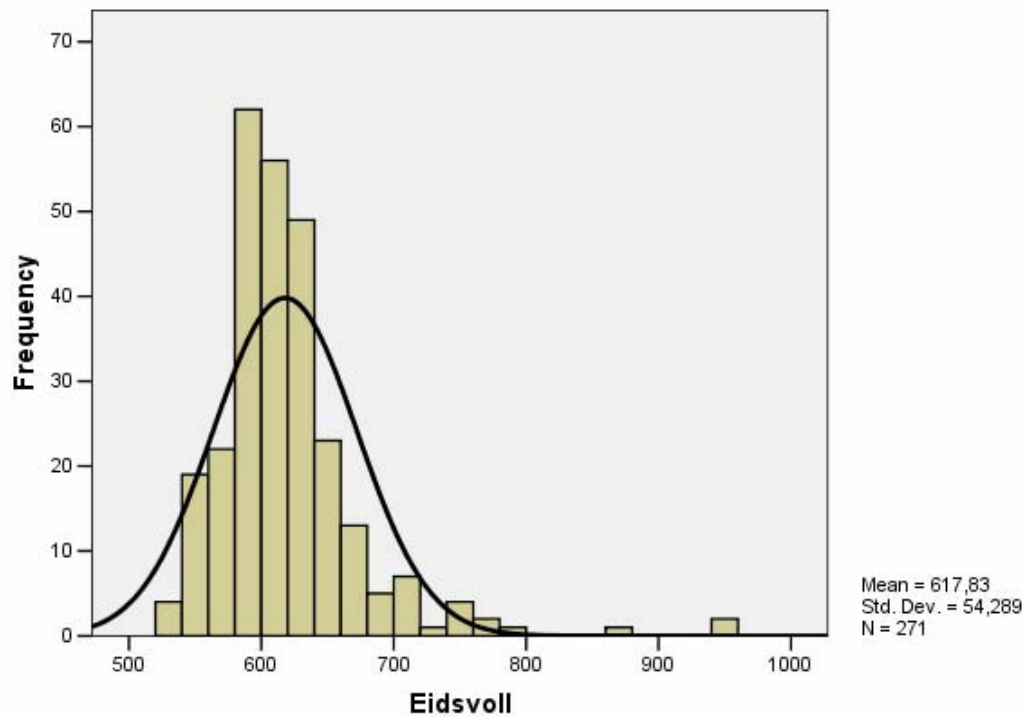
Kløfta



Gardermoen



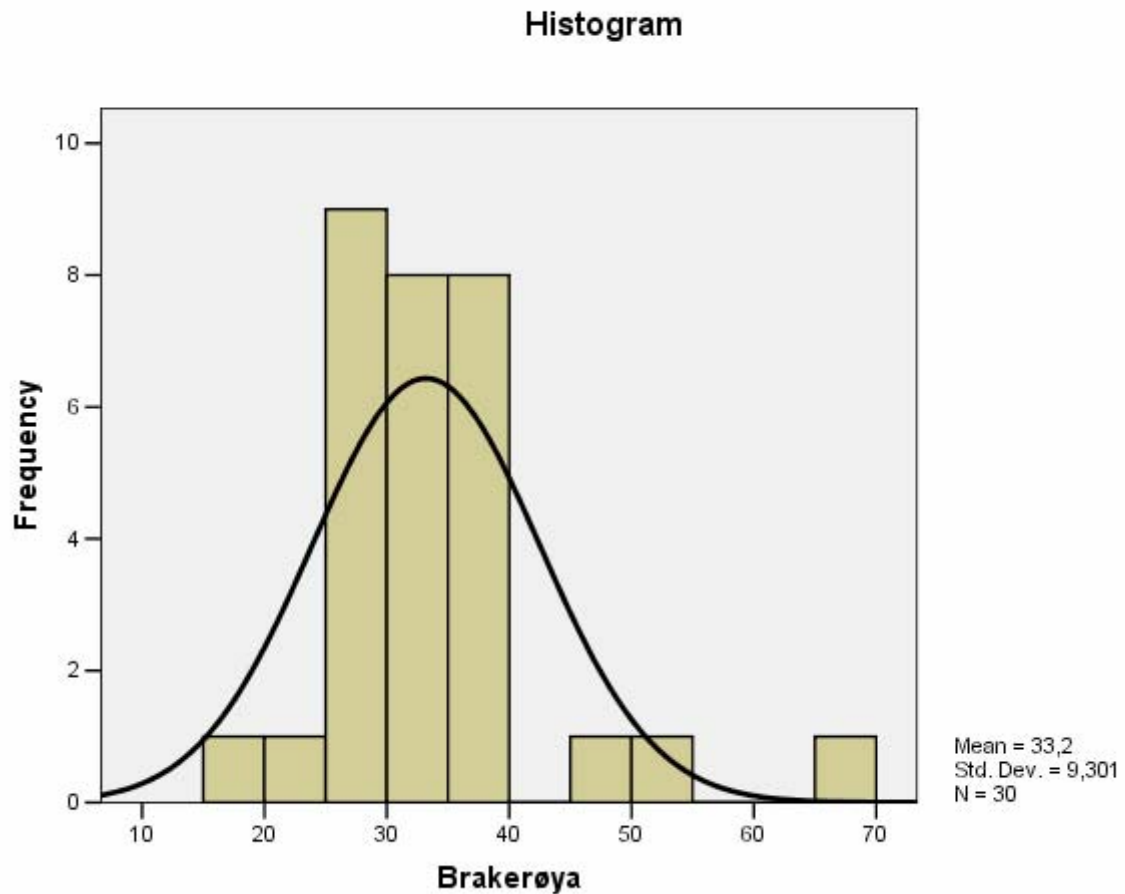
Eidsvoll



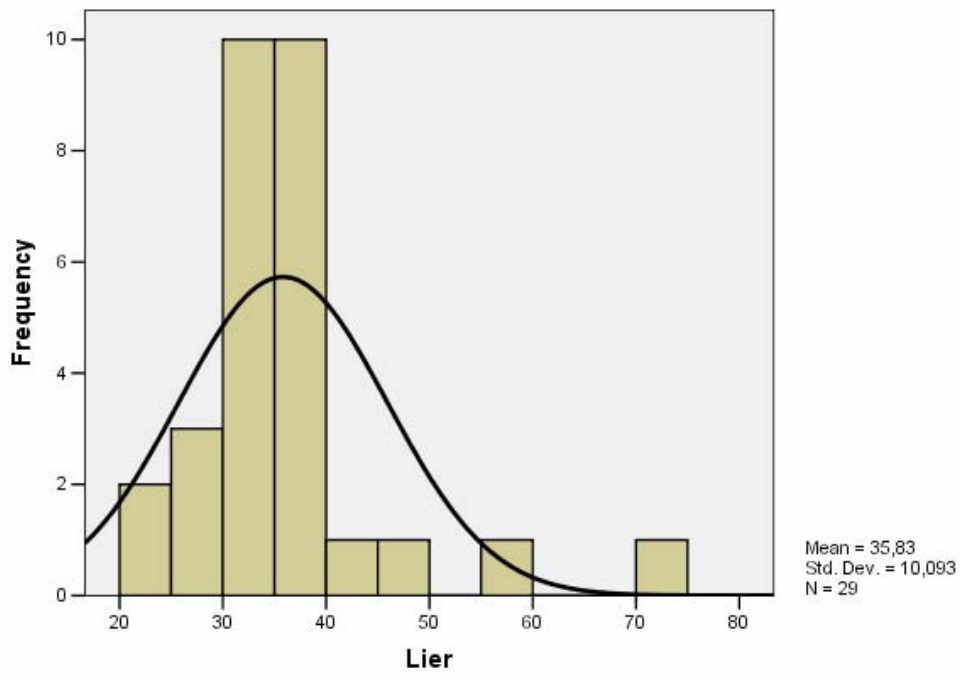
8.6 TELOC

8.6.1 HISTOGRAMMER OPPHOLDSTID TELOC

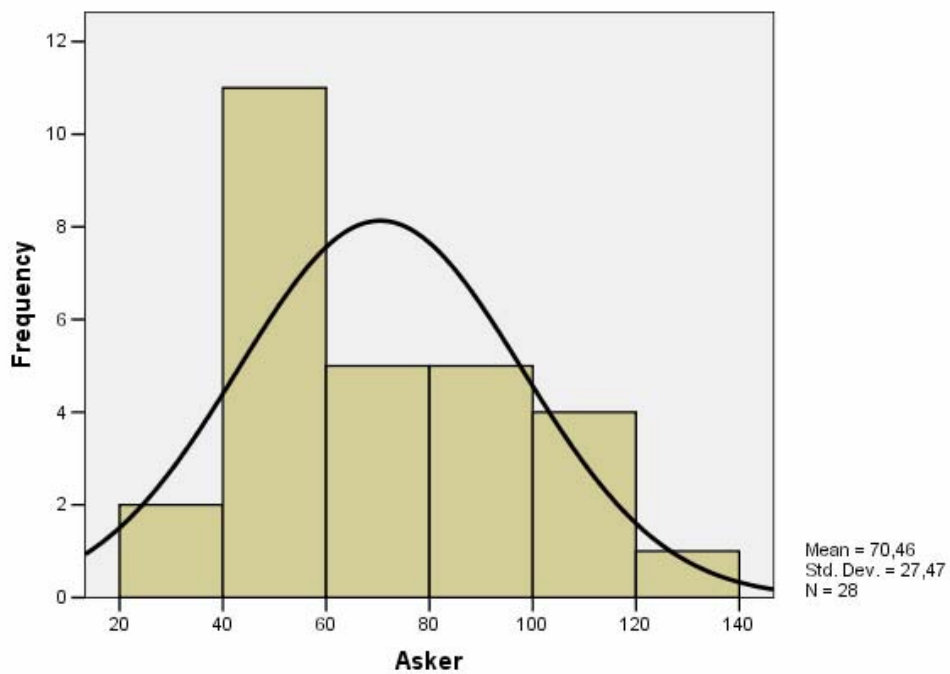
Histogrammene viser oppholdstida fra TELOC i sekunder, alle observasjoner.



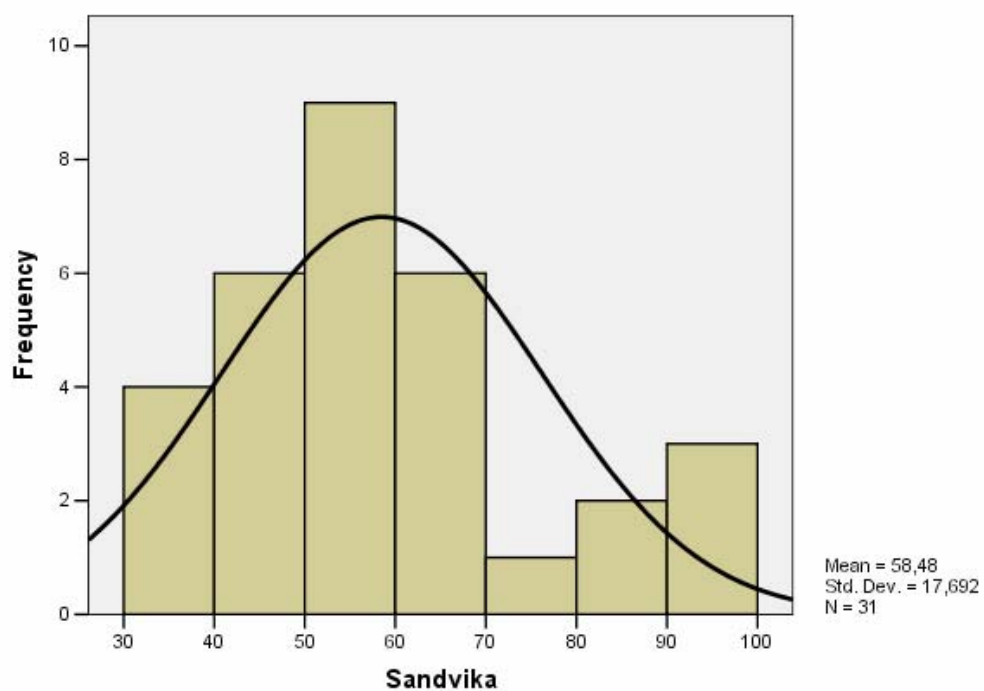
Histogram



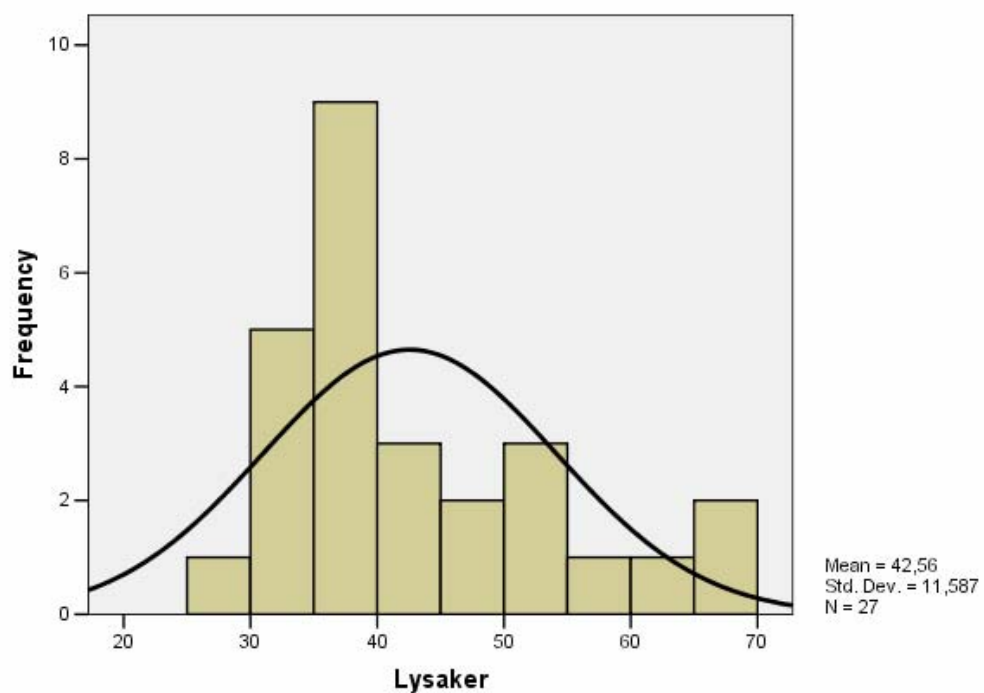
Histogram



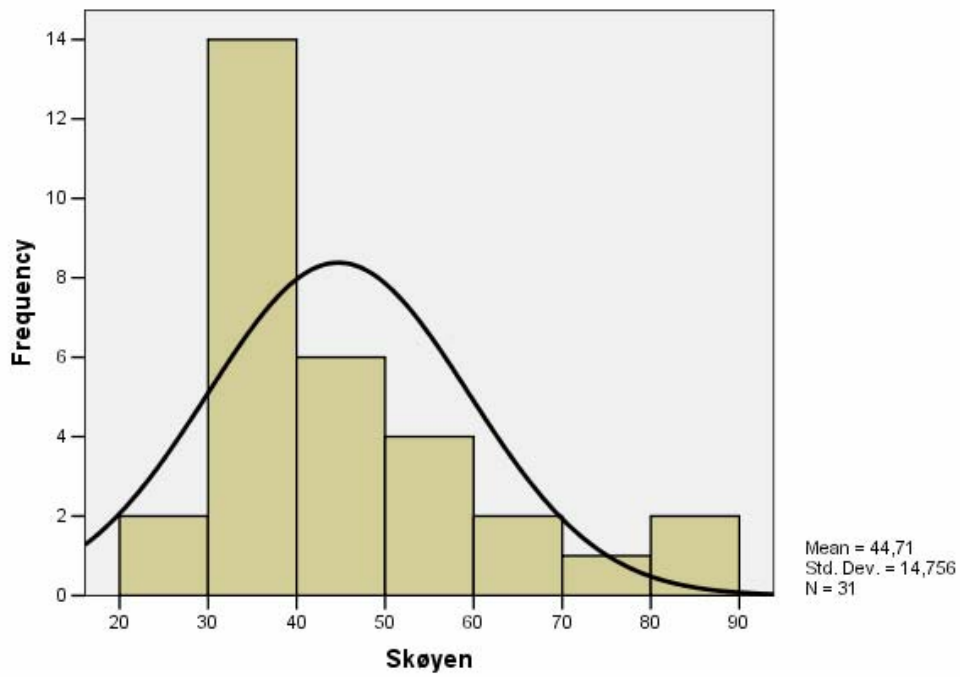
Histogram



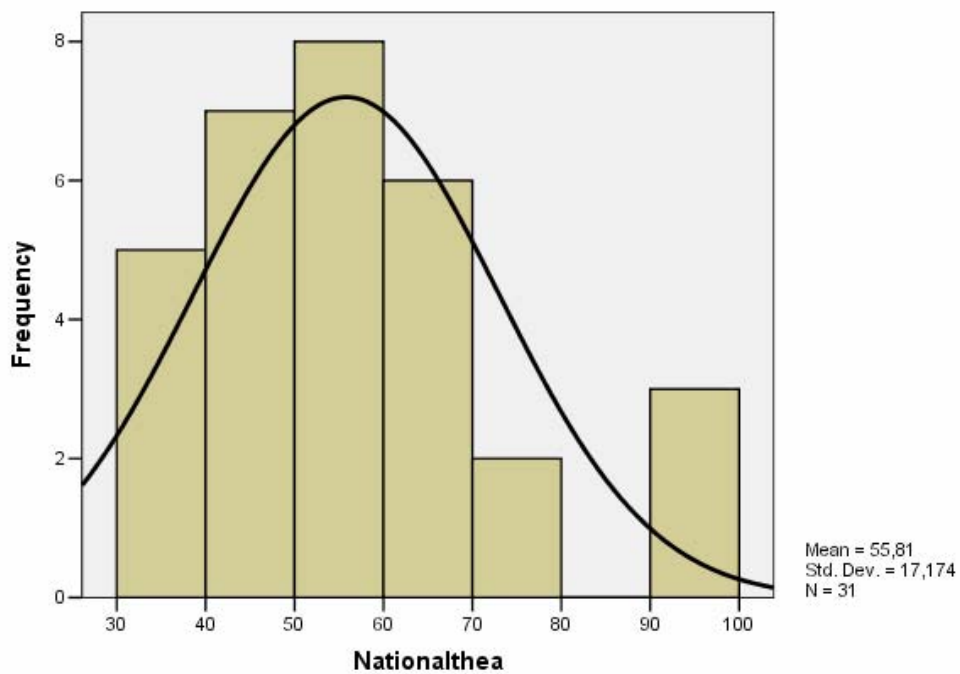
Histogram



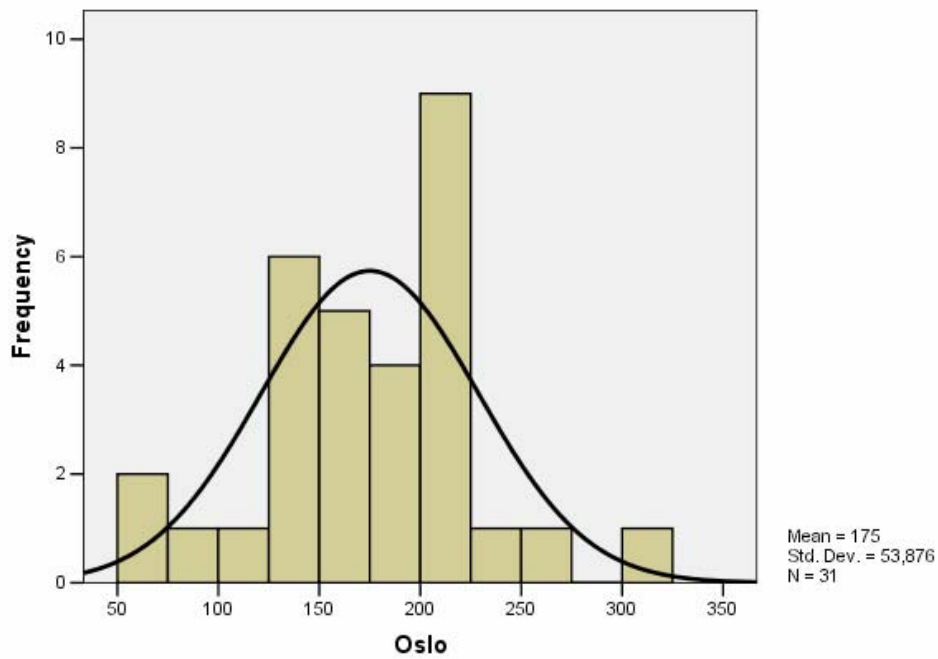
Histogram



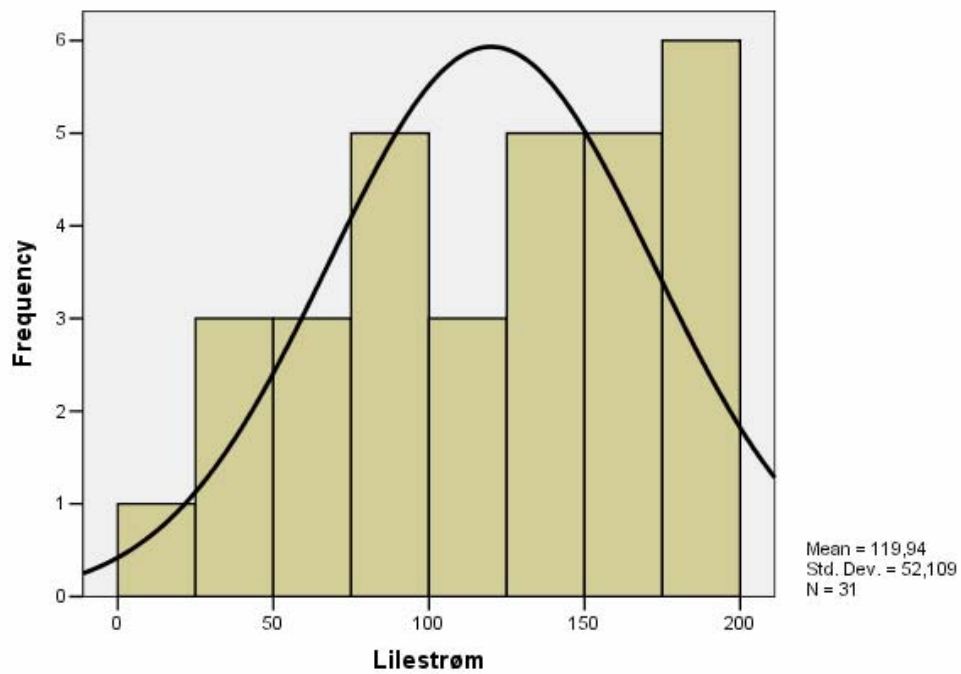
Histogram



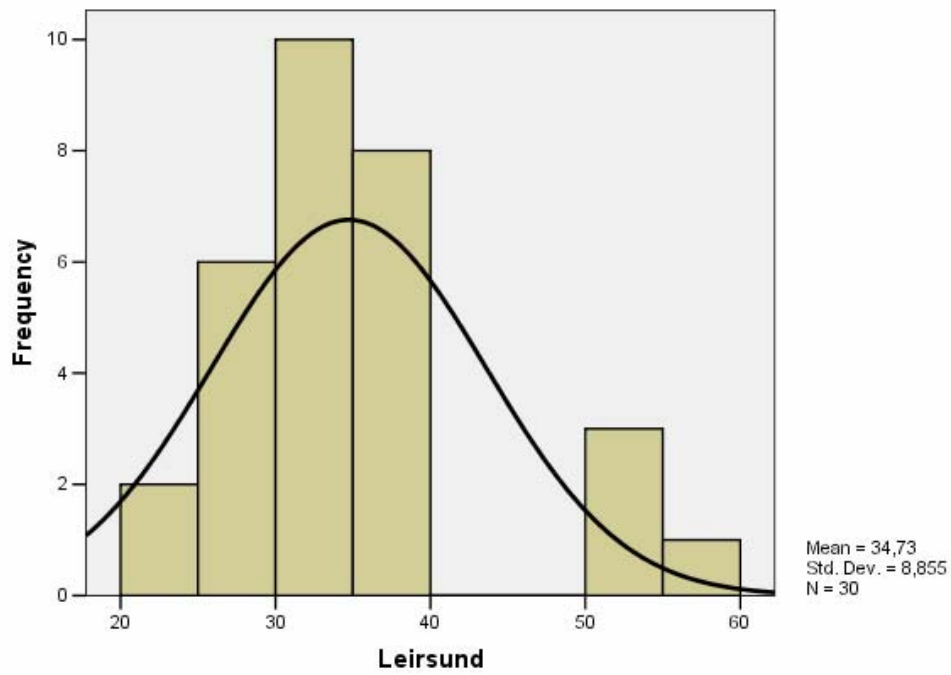
Histogram



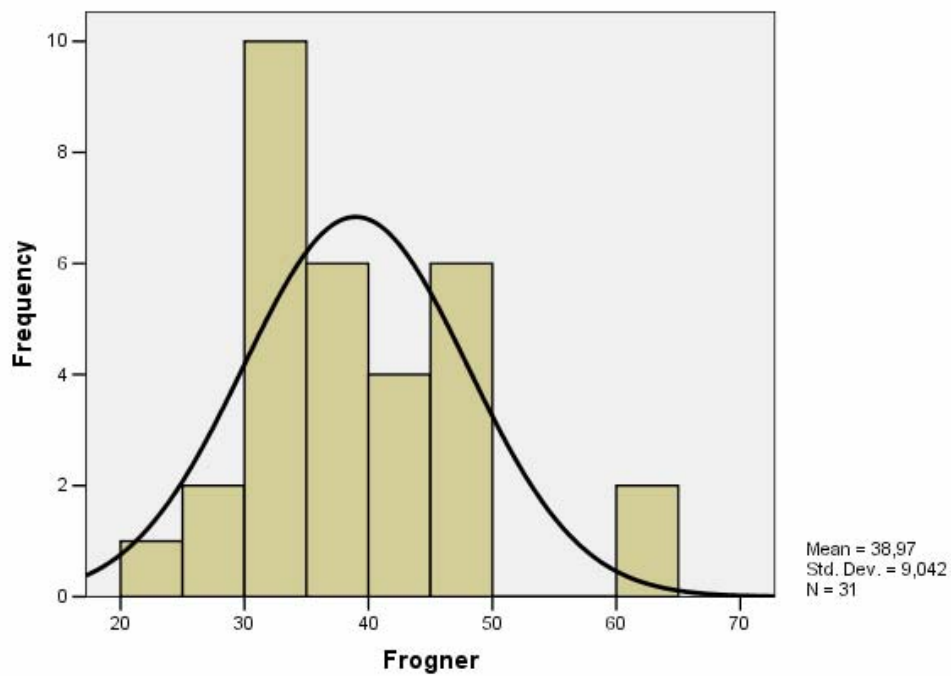
Histogram



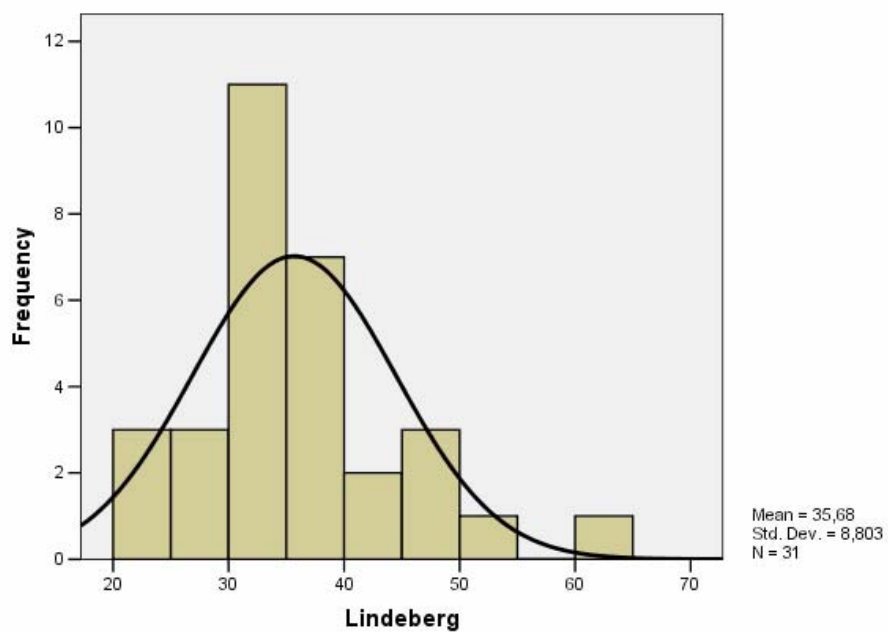
Histogram



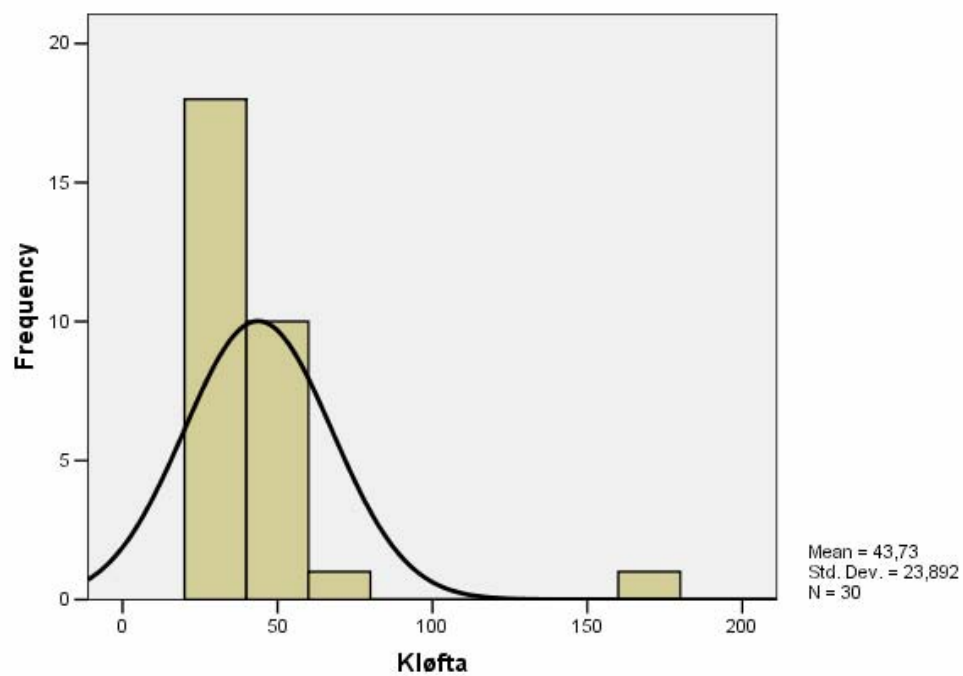
Histogram



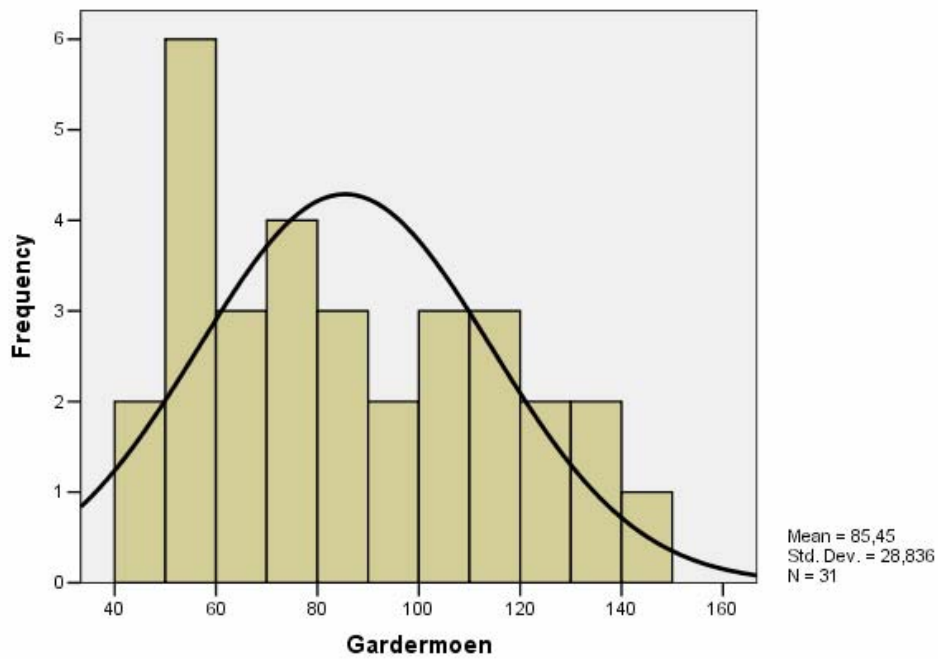
Histogram



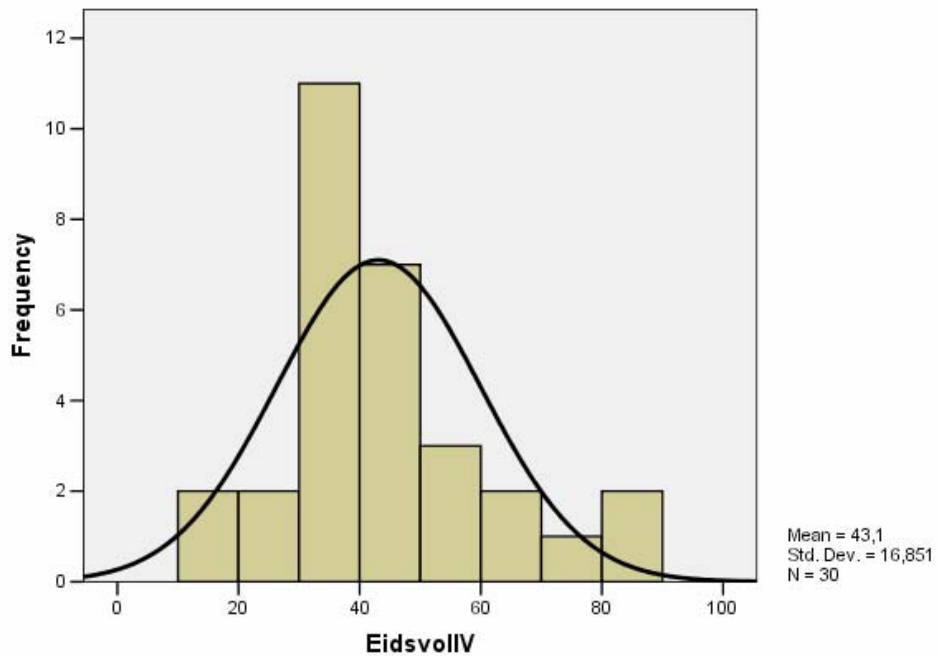
Histogram



Histogram

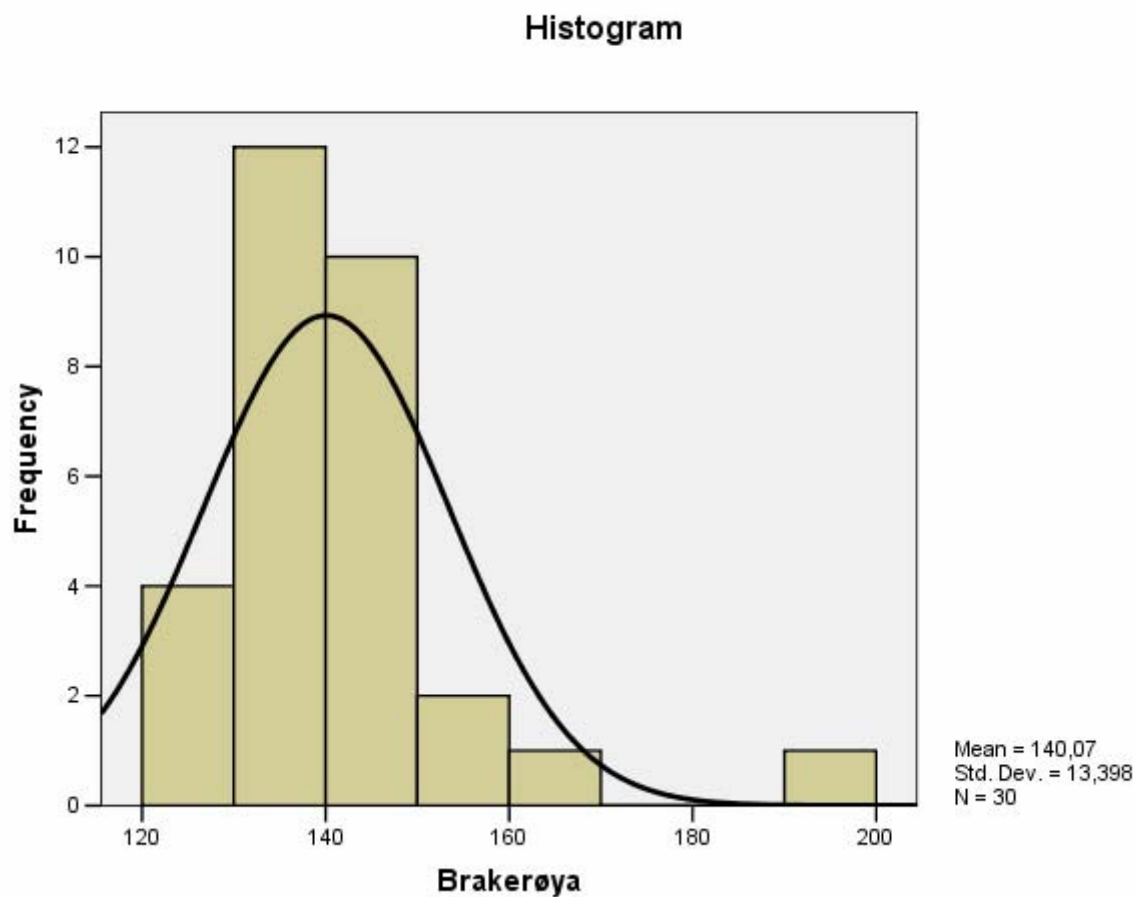


Histogram

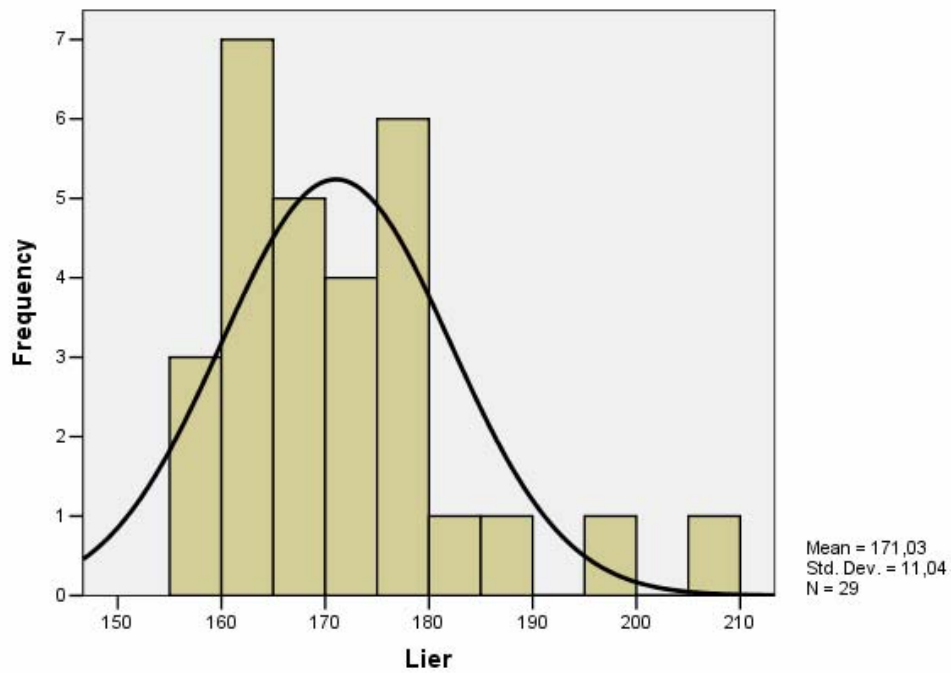


8.6.2 HISTOGRAMMER KJØRETID TELOC

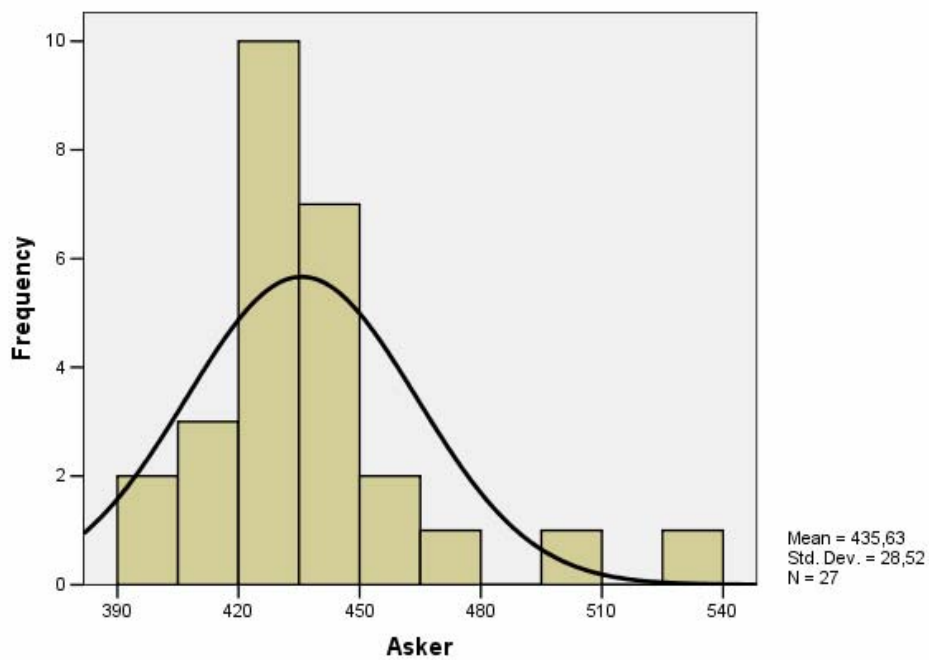
Histogrammene viser kjøretid i sekunder fra foregående stasjon til angitt stasjon i sekunder.



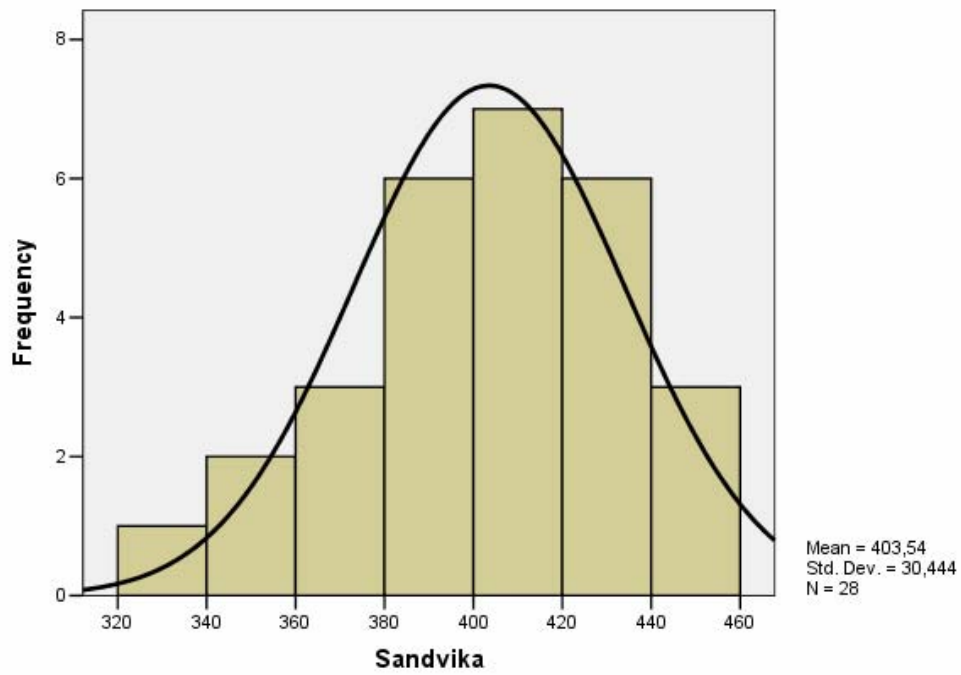
Histogram



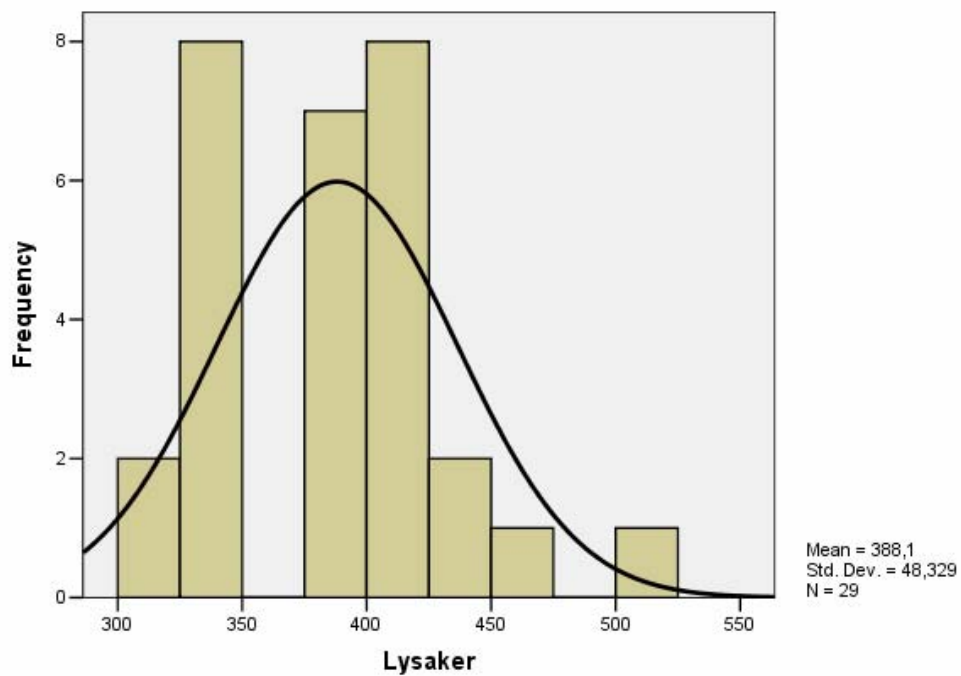
Histogram



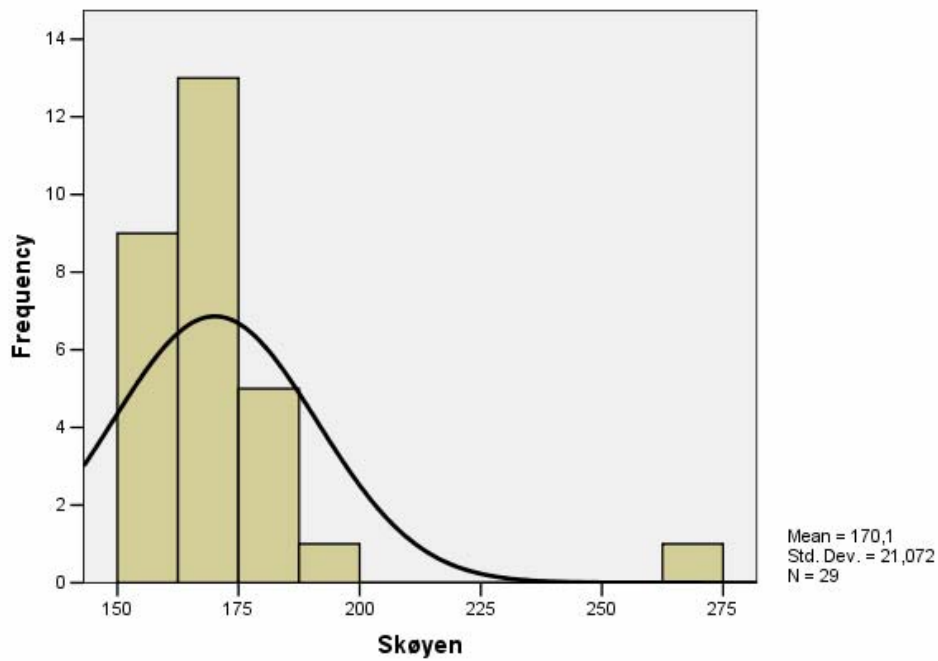
Histogram



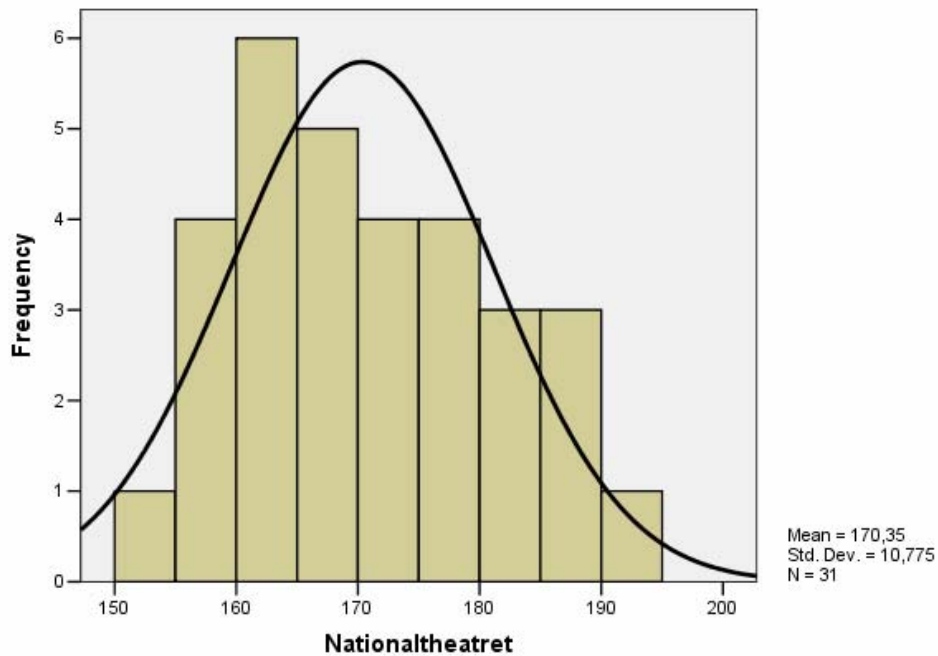
Histogram



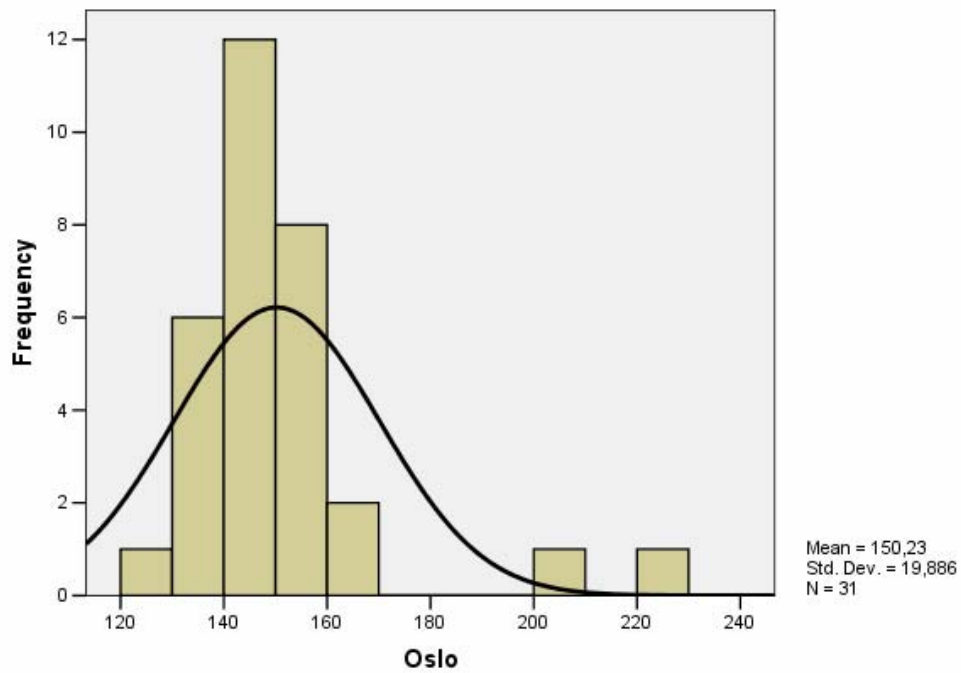
Histogram



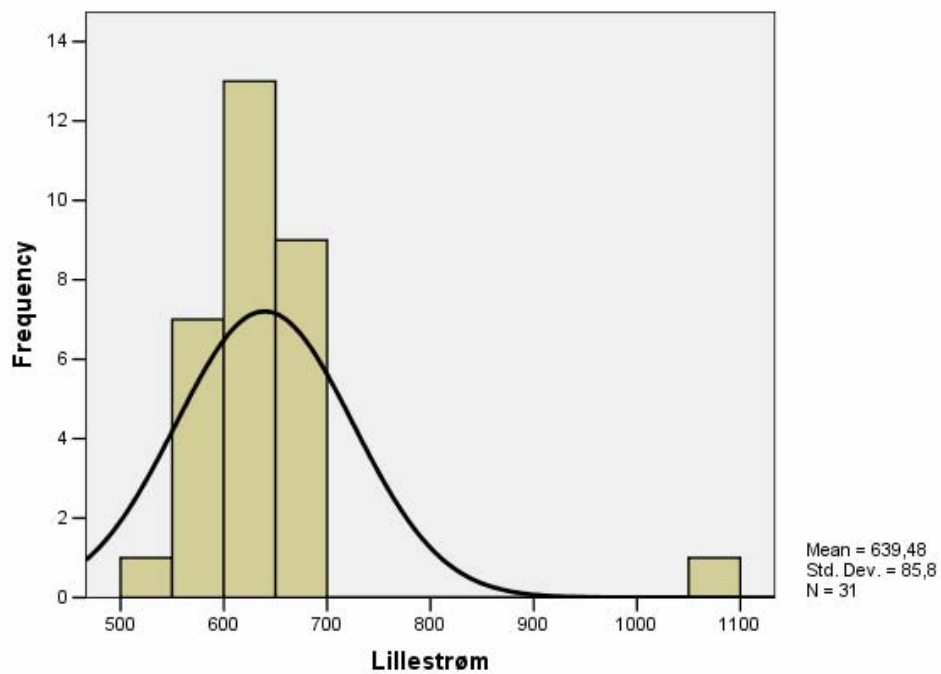
Histogram



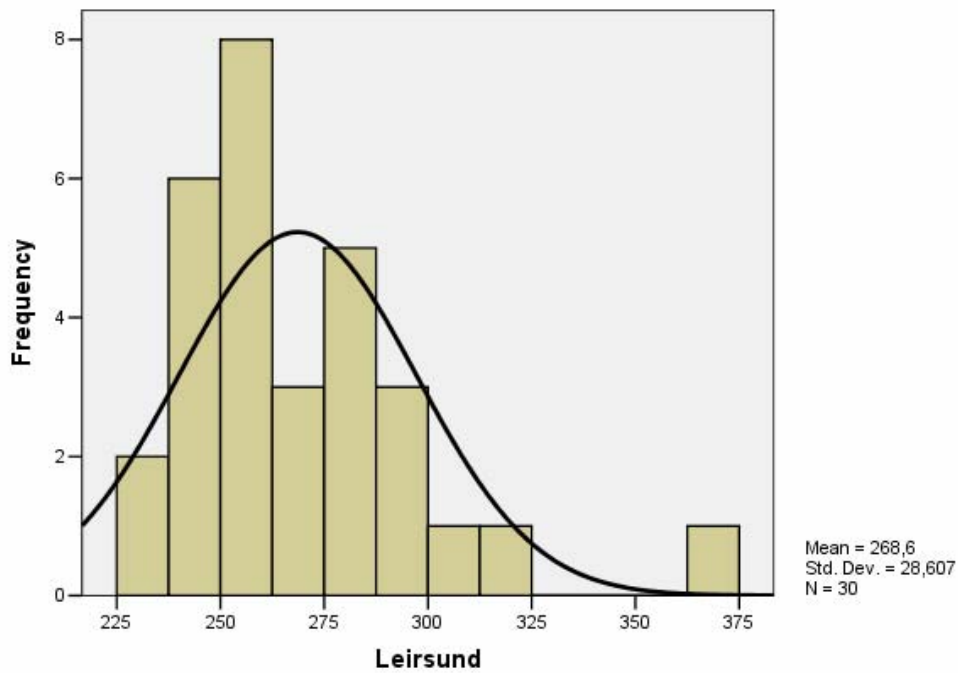
Histogram



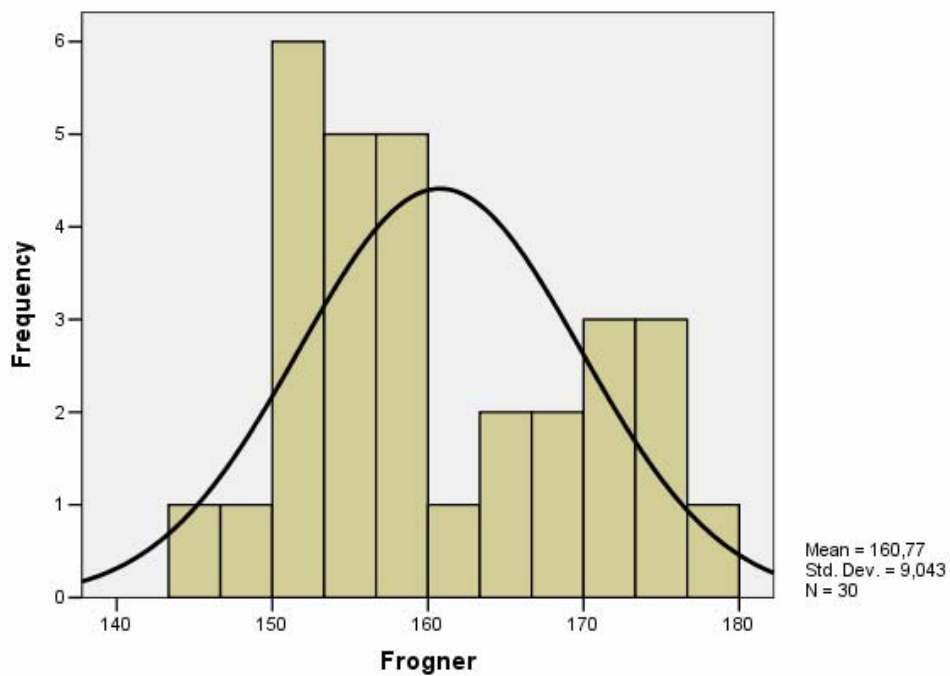
Histogram



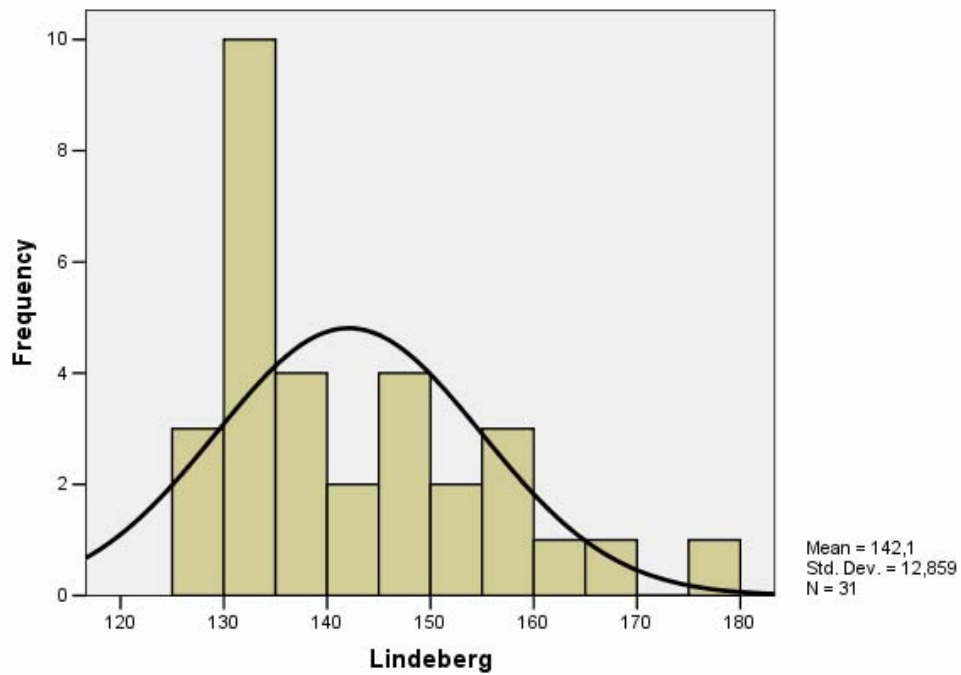
Histogram



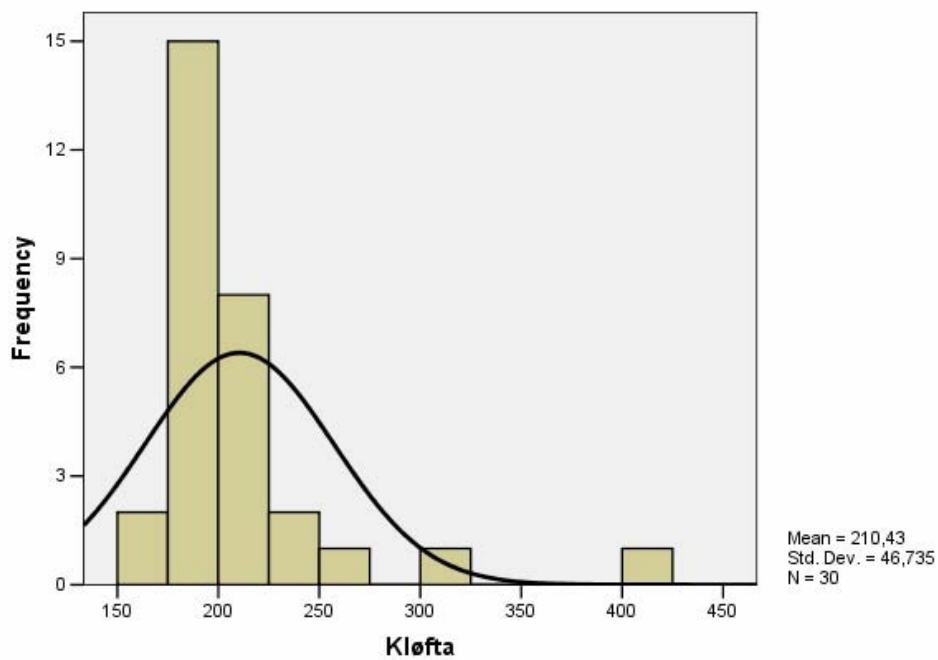
Histogram



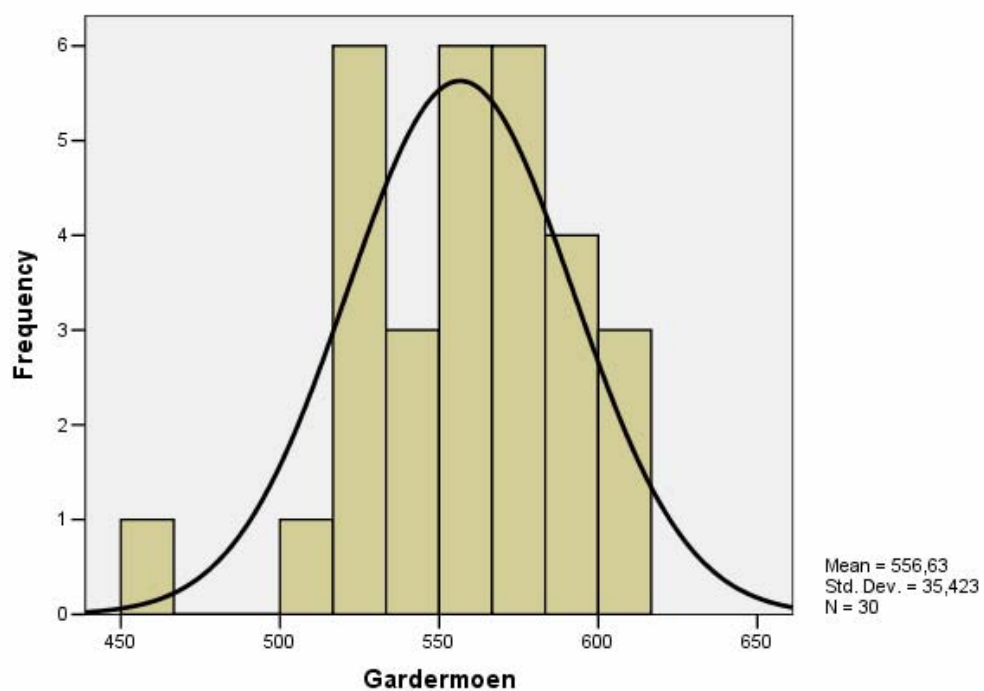
Histogram



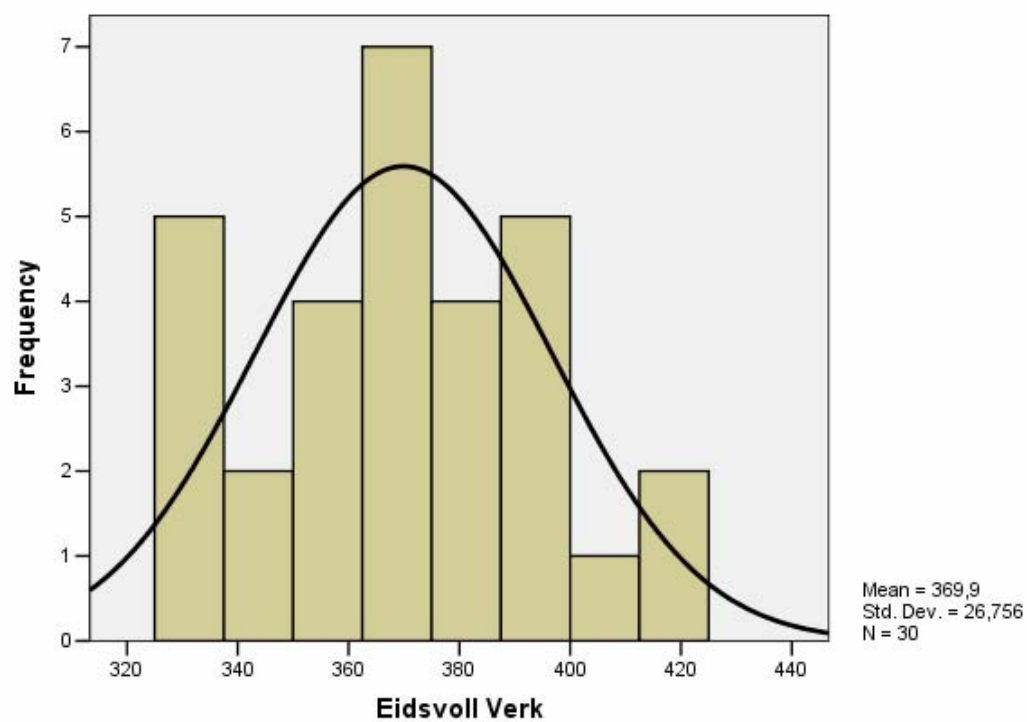
Histogram



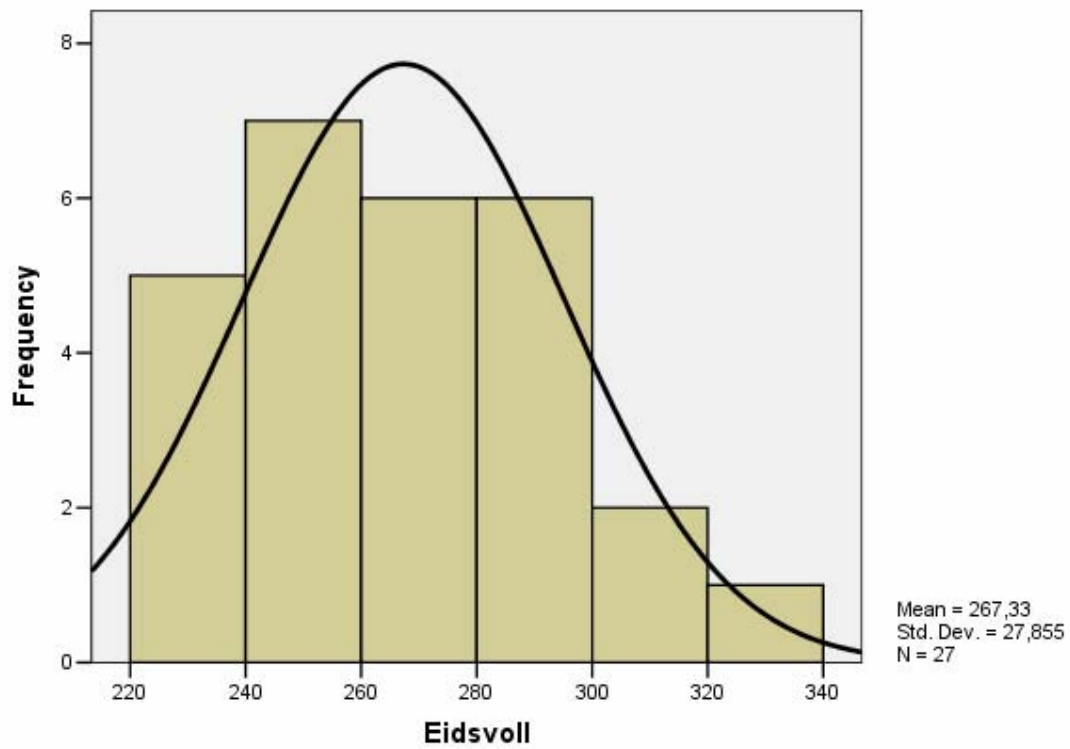
Histogram



Histogram



Histogram



8.7 FORSTUDIERAPPORT



Fakultet for ingeniørvitenskap og teknologi
INSTITUTT FOR PRODUKSJONS- OG
KVALITETSTEKNIKK

FORSTUDIERAPPORT

DATA FRA JERNBANEDRIFT
(Railway Operation Data)
Masteroppgave
Høst 2005

Stud. Techn. Øystein Luktvaslimo
Institutt for Produksjon- og Kvalitetsteknikk
NTNU

Innholdsfortegnelse

Innholdsfortegnelse	2
Prosjektbeskrivelse	3
Bakgrunn for prosjektet.....	3
Problembeskrivelse	3
Målsetting og leveranser	3
Presisering	3
Prosjektstyring.....	5
Aktiviteter med tildelninger	5
Vedlegg	I

Prosjektbeskrivelse

Bakgrunn for prosjektet

Problembeskrivelse

Oppgaven rettes mot oppfølging av jernbanedrift. Det tas utgangspunkt i analyse og sammenligning av data fra ulike kilder. Fokus rettes mot oppfølging av punktlighet og relaterte faktorer, som kan være kjøretid, stasjonsopphold og egenskaper ved det rullende materiellet.

Oppgaven utføres i samarbeid med NSB og forskningsprosjektet PEMRO.

I oppgaven skal kandidaten mer spesifikt:

1. Gjennomføre et litteraturstudium relatert til datakilder som skal brukes som grunnlag for faktabaserte beslutninger. Et sammendrag av dette skal presenteres.
2. Beskrive/belyse toggangen for en utvalgt strekning ved hjelp av ulike datakilder over en viss tidsperiode.
3. Analysere de ulike datasettene og presentere resultatene for stasjonsopphold på noen stasjoner med spesiell fokus på å vurdere ulike forklaringsfaktorer.
4. Basert på de foregående punktene, beskrive styrker og svakheter, likheter og ulikheter ved de ulike datasettene og analyseformene.

Oppgaveløsningen skal basere seg på de eventuelle standarder og praktiske retningslinjer som foreligger og anbefales. Dette skal skje i nært samarbeid med veiledere og fagansvarlig. For øvrig skal det være et aktivt samspill med veiledere.

Målsetting og leveranser

Presisering

I det følgende vil man forsøke å utdype, presisere og beskrive nærmere de nødvendige arbeidsoppgavene for å nå de satte målene, og løse oppgaven på en best mulig måte. Som man kan se av problembeskrivelsen er oppgaven delt inn i fire arbeidsoppgaver. Disse arbeidsoppgavene er listet under med utdyping av de problemstillingene studenten mener han står ovenfor ved arbeidet med disse, og hvilke nye arbeidsoppgaver han ønsker å fokusere på og som sees nødvendige. Arbeidsoppgavene blir delt opp og nummerert.

1. Gjennomføre et litteraturstudium relatert til datakilder som skal brukes som grunnlag for faktabaserte beslutninger. Et sammendrag av dette skal presenteres.
Arbeidsoppgaver:
Kategorisere ulike typer datakilder: hvilke krav og egenskaper stilles til ulike datakilder relativt til sitt bruk?
Finne ut hva som er bruksområdet til ulike typer datakilder.
Definere hva som kjennetegner en faktabasert beslutning.
Litteratur:
Quesenberry, C.P. 1997, *SPC Methods for Quality Improvement*, Wiley & Sons INC., Canada.
Redman, Th.C. 2001, *Data Quality: the field guide*, Digital Press, USA.

Zarkovich, S.S. 1966, *Quality of Statistical Data*, Food and Agriculture Organization of the United Nations, Roma.
Montgomery, D.C., & Runger, G.C. 2003, *Applied Statistics and Probability for Engineers*, Wiley, New York.
Naus, J.I. 1975, *Data Quality Control and Editing*, Marcel Dekker, New York.
Liepins, G. E. & Uppuluri, V.R.R., 1990, *Data Quality Control: Theory and Pragmatics*, Marcel Dekker, New York.
Burlington, R.S. 1970, *Handbook of Probability and Statistics with Tables*, McGraw-Hill, New York.
Spirer, H.F., Spirer, L., & Jaffe, A.J. 1998, *Misused Statistics*, Marcel Dekker, New York.
Hines, W.W., Montgomery, D.C., Goldsman, D.M., & Borror, C.M. 2003, *Probability and Statistics in Engineering*, Wiley, Hoboken, N.J.

2. Beskrive/belyse togangen for en utvalgt strekning ved hjelp av ulike datakilder over en viss tidsperiode.
Arbeidsoppgaver:
Hente data fra NSB for en gitt tidsperiode (22. august til 5. september) for en strekning.
Datakilder: TELOC, ANNALyse, manuelle målinger gjort sommer 2005, passasjerantall, stasjonsopphold fra signalanlegg.
Visualisere togangen ved hjelp av ulike diagrammer med hensyn på stasjonsopphold, kjøretid og ankomstforsinkelse.
3. Analysere de ulike datasettene og presentere resultatene for stasjonsopphold på noen stasjoner med spesiell fokus på å vurdere ulike forklaringsfaktorer.
Arbeidsoppgaver:
Hente statistisk data fra SBB for befolkningssammensetning, pendlemønster, yrkessammensetning etc.
Konstruere og sammenligne to ulike stasjoner med to statistiske modeller som inkluderer noen utvalgte forklaringsparametere for lengde på stasjonsoppholdet.
4. Basert på de foregående punktene, beskrive styrker og svakheter, likheter og ulikheter ved de ulike datasettene og analyseformene.
Sammenligne de ulike datakildene som er brukt i oppgaven. Gjøre et forsøk på å kvalitetsmessig vurdere dem opp mot teori fra del 1 av oppgaven.
Sammenligne de ulike analyseformene som er blitt brukt på datamaterialet.

Prosjektstyring

Prosjektets startdatum: tis 16.08.05

Prosjektets slutdatum: tis 10.01.06

Aktiviteter med tildelninger

ID	Aktivitet	Arbete	Varaktighet	Start	Slut
2	Forstudierapport	0 tim	22 dagar?	tis 16.08.05	tis 06.09.05
3	Ferdig utkast del 1	0 tim	0 dagar	fre 07.10.05	fre 07.10.05
4	Del 1	0 tim	53 dagar?	tis 16.08.05	fre 07.10.05
5	1.1	0 tim	53 dagar?	tis 16.08.05	fre 07.10.05
6	1.2	0 tim	53 dagar?	tis 16.08.05	fre 07.10.05
7	1.3	0 tim	53 dagar?	tis 16.08.05	fre 07.10.05
8	Ferdig utkast del 2	0 tim	0 dagar	mån 31.10.05	mån 31.10.05
9	Del 2	0 tim	54 dagar?	ons 07.09.05	mån 31.10.05
10	2.1	0 tim	54 dagar?	ons 07.09.05	mån 31.10.05
11	2.2	0 tim	54 dagar?	ons 07.09.05	mån 31.10.05
12	2.3	0 tim	54 dagar?	ons 07.09.05	mån 31.10.05
13	Ferdig utkast del 3	0 tim	0 dagar	mån 28.11.05	mån 28.11.05
14	Del 3	0 tim	52 dagar?	fre 07.10.05	mån 28.11.05
15	3.1	0 tim	52 dagar?	fre 07.10.05	mån 28.11.05
16	3.2	0 tim	52 dagar?	fre 07.10.05	mån 28.11.05
17	Ferdig utkast del 4	0 tim	0 dagar	fre 16.12.05	fre 16.12.05
18	Del 4	0 tim	38 dagar?	ons 09.11.05	fre 16.12.05
19	4.1	0 tim	38 dagar?	ons 09.11.05	fre 16.12.05
20	4.2	0 tim	38 dagar?	ons 09.11.05	fre 16.12.05
21	Ferdigstille prosjektet	0 tim	26 dagar?	fre 16.12.05	tis 10.01.06
22	Innlevering	0 tim	0 dagar	tis 10.01.06	tis 10.01.06

