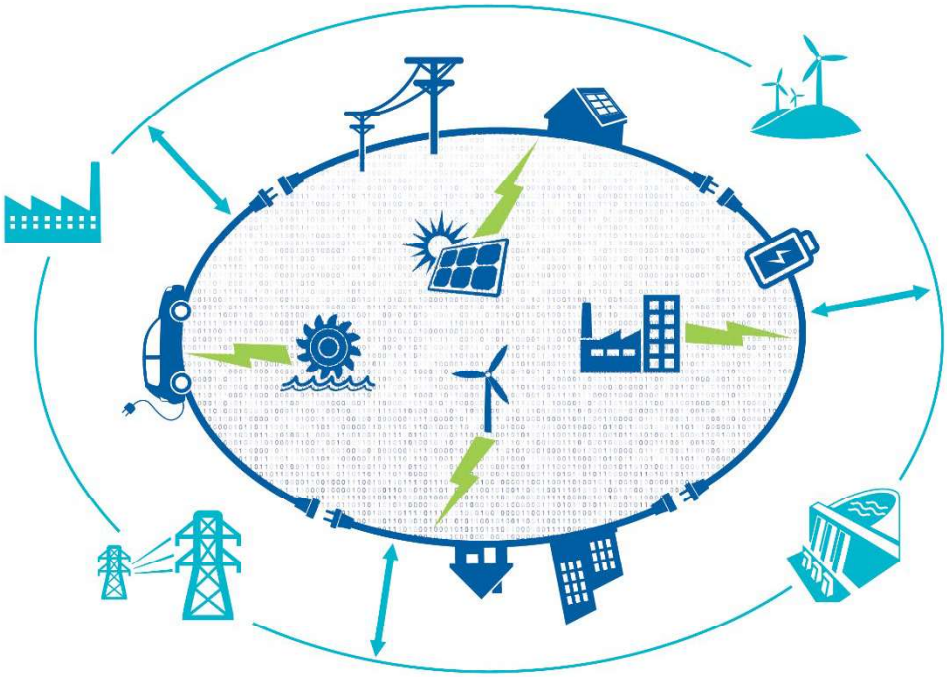
Pilot sluttrapport

Effektanalyse

Forfatter: Per Oddvar Osland



CINELDI

Centre for intelligent electricity distribution
- to empower the future Smart Grid

Centres for Environment-friendly Energy Research

CINELDI - Centre for intelligent electricity distribution

SINTEF and NTNU are the main research partners, with grid operators, technology providers, public authorities and international R&D institutes and universities as partners.

The research centre is financed by the Research Council of Norway and the Norwegian partners through the Centre for Environment-friendly Energy Research (FME) scheme. The FME scheme consists of research centres of limited duration that conduct concentrated, focused and long-term research on a high international level to solve specific challenges related to energy and the environment.

CINELDI

Centre for intelligent electricity distribution
- to empower the future Smart Grid

Centres for
Environment-friendly
Energy Research

Prosjektnotat

TITTEL			
Resultat og erfaringsnotat for Pilot *Effektanalyse*			
WORK PACKAGE	**VERSJON**	**DATO**	**ANTALL SIDER**
WP Pilot	1.0	2024-08-12	11
FORFATTER(E)		**WP-LEDER**	**GRADERING**
Per-Oddvar Osland Per-Oddvar Osland (Oct 7, 2024 22:34 GMT+2)		Maren Istad Maren Istad (Oct 8, 2024 07:39 GMT+2)	Åpen

SAMMENDRAG

Piloten gir innsikt i variasjon for elektrisk effektuttak på ulike punkt i nettet; for sluttkunder og i nettstasjoner.

I lengre tid har Velanders metode vært rådende standard for estimering av maks effektuttak, men i det siste har en stokastisk metode har blitt foreslått i Erling Tønne's doktorgradsavhandling. I første del av dette pilotarbeidet har denne metoden blitt undersøkt, analysert og forbedret. Konklusjonen er at det er tidkrevende å finne en tilpasset sannsynlighetsmodell per kunde, og resultatene i våre analyser fant ingen sannsynlighetsmodell som utpreget seg som «en god tilpasning» for alle.

Videre har ulike andre metoder for estimering av maksimalt effektuttak (timesverdier) har blitt analysert, her fremstår lineær regresjon som den beste metoden. Videre viser piloten hvordan momentan effektpådrag varierer innenfor en klokketime, og at timesverdier her skjuler vesentlige variasjoner. Det demonstreres også at 5-minuttsverdier gir vesentlig mer innsikt i effektvariasjoner enn timesverdier, men at 30-sekundsverdier kun gir marginal gevinst over 5-minuttsverdier.

Innholdsfortegnelse

1 Bakgrunnsinformasjon om pilotprosjektet

Tabell 1: Bakgrunnsinformasjon

	Fra malen "planlegging av pilotprosjekt"	**Viktige endringer i løpet av pilotperioden**
Målsetting	Målet er kunnskap om reell effektvariasjon på ulike nivå i et distribusjonsnett slik at dette kan danne grunnlag for mer relevant modell for dimensjonering av nettkapasitet. Denne modellen kan gjerne være risikobasert, dvs. den kan uttrykke en sannsynlighet for at effektbehovet ikke oversiger kapasiteten.	Målsetningen var å fokusere på « effektvariasjon på ulike nivå i et distribusjonsnett… «. I praksis har det blitt fokusert på effektvariasjon hos sluttkunde og i nettstasjoner. Det har blit gjennomført analyser av effektvariasjon som gir grunnlag for å utvikle ny modell for dimensjonering. Utvikling av selve modellen har ikke vært målsetning for pilotprosjektet.
Problemstilling	Dagens metoder for dimensjonering av nett tar som regel utgangspunkt i empirisk beregningsgrunnlag for effekt, eksempelvis Velanders formel eller brukstid. Dette er velprøvd metodikk som er godt nok i mange tilfeller, men som i enkelte situasjoner gir feilaktige resultat. Hypotesen i denne piloten er at man basert på målinger og egnede modeller kan beskrive og forstå det reelle effektpådraget langt bedre, og dermed danne et mer korrekt grunnlag for dimensjonering. Eksempel på relevante problemstillinger • Det er vanlig å operere med effekt uttrykt i "kWh / h", dvs gjennomsnittlig effektpådrag over en time. Men hvordan varierer effekten innenfor en time? Og hvor fin målesampling trenger man for å observere det reelle effektpådraget med god nok oppløsning? • Når to eller flere kunder deler en ressurs, vil graden av samtidighet påvirke i hvordan man dimensjonerer ressursen. Hvordan kan man modellere "graden av samtidighet" slik at denne inngår i	Det har blitt fokusert på følgende problemstillinger/tema: - Analyse av stokastisk modell for lastmodellering - Analyse av effektvariasjon innenfor en time - Estimering av maksforbruk hos husholdningskunder i Agder Energi Nett via maskinlæringsmodeller og andre algoritmer - Vurdering av verdi av 30-sek data (mot 5-min data) fra nettstasjon Det har blitt brukt statistiske metoder og maskinlæring i gjennomføring av arbeidet.

	en sannsynlighetsmodell for nettkapasitet? • Ved etablering av regelverk for dimensjonering av nett, hvordan kan man gå over til å oppfatte dimensjonerende kapasitet som et mål for at man møter behovet med en gitt sannsynlighet. Uttrykt matematisk: For et etterspurt effektbehov *B* (stokastisk variabel), velg en kapasitet *k* (fast parapeter) slik at $P(B > k) < e$ der *e* er en valg grense som uttrykker risikoviljen til selskapet (eks *e* = 0,005 = 0,5 %)	
Aktiviteter	1. Etablere teoretisk grunnlag for analyse av effektvariasjon. Dette gjøres gjennom litteraturstudier og undersøkelser av "state of the art" på området. 2. Datainnsamling 3. Etablere modeller for effektvariasjon og samtidighet 4. Sammenligne valgt modell opp målte verdier og eksisterende modeller (Velander og brukstid). 5. Evaluering og analyse	
Kostnadsestimat	Timer: 500 timer, herav • Data scientist – 250 timer • Elteknisk fagkompetanse – 100 timer • Utvikler / data engineer – 100 timer • Prosjektstøtte – 50 timer Utstyr • Måleutstyr til innhenting av data i nettstasjon – kr 50 000 IKT-kostnader MS Azure • Det er ikke planlagt maskinlæring som del av denne aktiviteten. Kostnader til lagring av data er begrenset.	
Innovasjonspotensial		Høyt innovasjonspotensiale. Metodikk for estimering av effektuttak har vært basert på Velanders formel i flere tiår. AMS har gitt tilgang på

		detaljerte data for effektuttak, og analyse av disse dataene kan gi grunnlag for vesentlig bedre metoder for estimering av effekttuttak.
Forventet resultat	Rapporter som beskriver observasjoner og analyser.	
Tidsplan	1.1.2020 – 31.12.2021	1.1.2020 – 31.12.2022 Pga begrenset tilgang på ressurser gjennom Covid-perioden ble det nødvendig å forlenge piloten.

2 Om Piloten og fysisk pilotområde

Tabell 2: Piloten og pilotområdet

Pilotområdet	Nettområde til Glitre Nett Sør, dvs tidligere Agder Energi Nett
Måledata og andre data som samles inn og lagres fra Piloten	Data fra husholdningskunder og fritidskunder: - 10-sek verdier for effektforbruk fra ca 20 sluttkunder - Timesverdier (energiforbruk) fra et stort utvalg av sluttkunder Data fra nettstasjoner - 30-sek verdier for effektforbruk fra 3 nettstasjoner - 5-min verdier for effektforbruk fra et stort utvalg av nettstasjoner
Personvern og/eller kraftsensitiv informasjon	Personsensitivt: Måledata fra sluttkunder. Det er innhentet samtykke fra kunder som har bidratt med dataunderlag.
Måle- og kommunikasjons-infrastruktur	AMS-målere i nettstasjoner og hos sluttkunder. Smarthjem-løsninger for uthenting av 10-sek verdier i kundens målepunkt. SafeMon (Safebase måleoppsett) for uthenting av 30-sek verdier fra nettstasjoner.
Use-case-beskrivelser og testplaner	Inngår i rapporter og presentasjoner, se vedlegg.
Regulering og forskrifter	GDPR
Barrierer og løsninger	
Hvem skal eventuelt ta resultater fra Piloten i bruk?	Forskningsmiljø og standardiseringsinstitusjoner som utvikler nye modeller for estimering av maks effektuttak.
Hvem er erfaringene relevant for?	Nettselskap, utstyrs- og programvareleverandører, forskningsmiljø og standardiseringsinstitusjoner. Nye prosjekt som FORSEL vil ha nytte av resultatene.

Hva påvirkes av resultater fra Piloter?	Videre arbeid med utvikling av metode for estimering av maks effektuttak.
Informasjonsdeling mellom aktørene før/underveis/etterpå	Informasjonsdeling har skjedd gjennom CINELDI (presentasjoner og rapporter) og Smartgridsenteret (webinar).
Er det laget planer for videreføring? Skalering/fullskala implementering?	Resultatene er relevante i nye prosjekter og forskningsaktiviteter som omhandler effektvariasjon og dimensjonering.

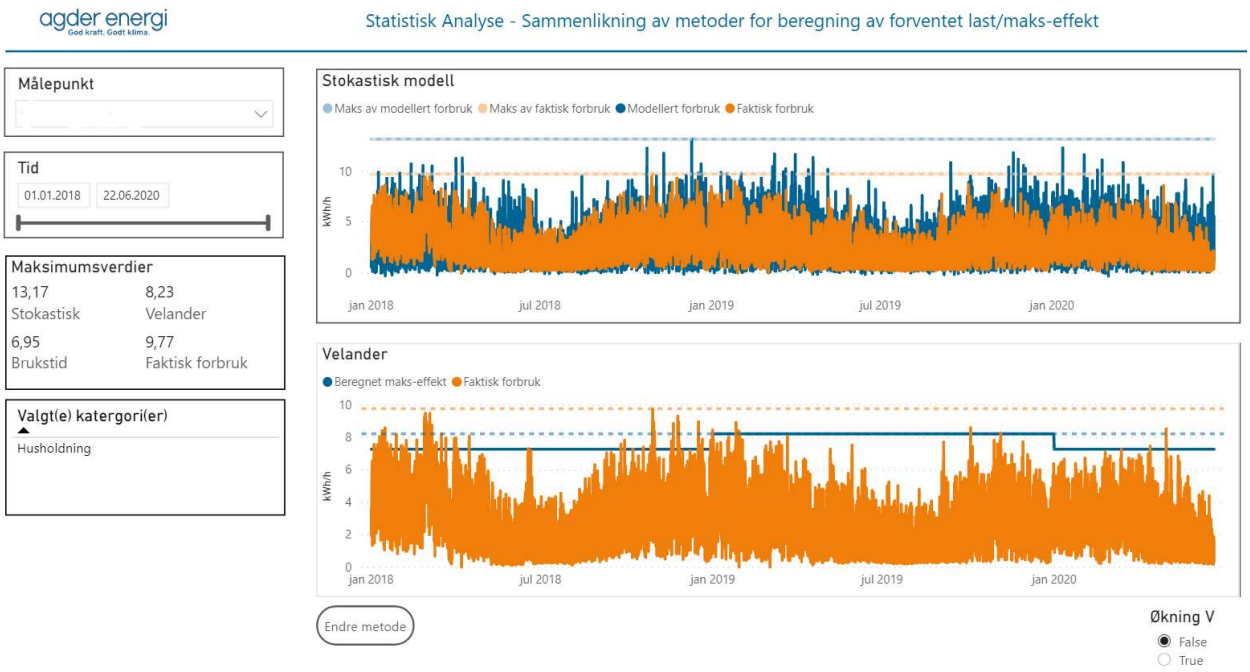
3 Resultater og innovasjoner fra Piloten

3.1 Resultater fra delaktivitet 1: Analyse av stokastisk modell for lastmodellering

Velanders metode har i lang tid vært gjeldende standard for å estimere maks effektforbruk på et gitt punkt i nettet. Dette kan være for en enkelt sluttkunde, eller på et aggregert punkt (for eksempel en nettstasjon). Metoden tar kun forventet årsforbruk som input, og er derfor forholdsvis enkel å bruke. Erfaringsmessig gir Velanders metode akseptable resultat på aggregert nivå, men kan bomme grovt på enkeltkunder.

Med innføring av AMS har man tilgang til forbruksdata time for time for en gitt kunde, og kan dermed angi historisk maksforbruk eksakt. Dette burde gi et godt grunnlag for å kunne modellere forventet maksforbruk på en bedre måte enn med Velanders formel. Erling Tønne har i sin doktorgradsavhandling foreslått en metode der historiske timesverdier bruks som grunnlag for en stokastisk lastmodell. I denne delaktiviteten i piloten analyseres denne metoden.

Vi har undersøkt maks- og gjennomsnittmetode for beregning av variasjonsverdier (verdier fra last-kurver), ulike metoder for vurdering av sannsynlighetsmodell for stokastisk justering, stokastisk justering basert på histogram av historisk avvik mellom variasjonsverdier og faktiske tempteraturjusterte verdier, stokastisk modellering et større antall ganger og analysert maks-verdiene i en slik fremgangsmåte. Vi har benyttet et datagrunnlag som består av historisk forbruk med timesoppløsning for 37 «tilfeldige» kunder. Det er utviklet flere visualiseringer og rapporter i Power BI for å understøtte arbeidet og vise resultat.



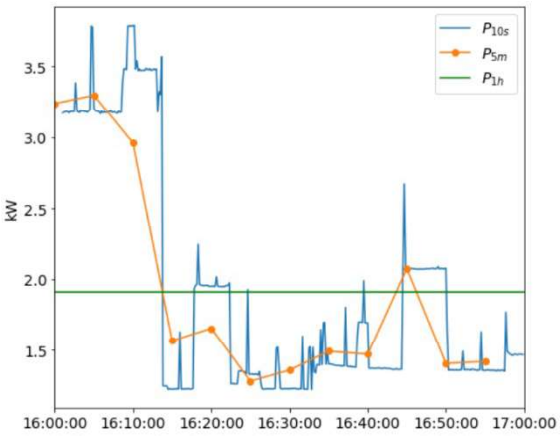
Figur 1 Eksempel på resultat fra sammenligning av Velanders metode, brukstid, stokastisk lastmodell og faktisk forbruk.

Konklusjonen er at det er tidkrevende å finne en tilpasset sannsynlighetsmodell per kunde, og resultatene i våre analyser fant ingen sannsynlighetsmodell som utpreget seg som «en god tilpasning» for alle. En løsning kan derfor være å trekke justeringsverdier til den stokastiske modellen direkte fra et histogram i stedet for fra en sannsynlighetsmodell. En periodeinndeling for de stokastiske justeringene viste seg å være nyttig, hvor man f. eks gjør en to-inndeling av året, og trekker stokastiske justeringer fra avvik som er basert på fra og med april til oktober (sommer) for timesverdier i det intervallet, og gjør tilsvarende for timesverdier fra og med november til april (vinter).

Resultat av arbeidet er oppsummert form av rapport [1] og presentasjon [2].

3.2 Resultater fra delaktivitet 2: Analyse av effektvariasjon innenfor en time

Effektverdier for elektrisk energi er vanligvis målt som gjennomsnitt over en time, dvs uttrykt i form av kWh/h. Samtidig er det velkjent at den momentane elektriske belastningen kan variere mye i løpet av en klokketime. Målet for piloten er å teste hypoteser for effektvariasjon og finne mål på forskjell mellom timesverdi og høyeste effekt innenfor timen.



Som datagrunnlag er det brukt 10-sekundsverdier (P_{10s}) for effekttuttak. Verdiene hentes fra HAN-porten på AMS-målere hos kunder som deltar i pilotprosjekt for uttesting av utstyr for styring av last. 10-sekundsverdier aggregeres i tillegg opp til 5-minuttsverdier (P_{5m}). I tillegg brukes timesverdier (P_{1h}) som hentes direkte fra AMS-måler. Dette gir grunnlag for tre måleserier; Effektverdier basert på 10-sek intervall, 5-min intervall og timesverdier.

Resultater:

Følgende hypoteser ble undersøkt:

Hypotese 1: Husholdnings- og fritidskunder har relativt lik effektvariasjon innenfor en time.
Konklusjon: NEI, mer variasjon hos fritidskunde

Hypotese 2: Husholdnings- og næringskunder har relativt lik effektvariasjon innenfor en time.
Konklusjon: JA, dette viser seg å stemme.

Hypotese 3: Det er ulik variasjon i høylast-, lavlast-, og normallasttimene.
Konklusjon: JA, mindre variasjon ved høy last
Regel: Sikring hos kunde bør tåle 2x maks timesverdi

FoU-spørsmål: Hvilken verdi gir 10-sek effektdata ut over 5-min effektdata?
Konklusjon: $PC_{99}F_{10s}$ ~ $PC_{99}F_{5m}$ i høylast og normallast: 99-prosent percentiler er tilnærmet like for 10-sek data og 5-min data i høylast og normallast. Dette betyr at behovet for å samle inn 10-sek data kun er til stede i spesielle situasjoner som for eksempel ved kundeklager.

Arbeidet er dokumentert i CINELDI-rapport [3] og har blitt presentert på webinar i CINELDI [4].

3.3 Resultater fra delaktivitet 3: Estimering av maksforbruk hos husholdningskunder i Agder Energi Nett via maskinlæringsmodeller og andre algoritmer

Timesverdi for maksimalt effekttuttak / maksforbruk (kwh/h) hos husholdningskunder har frem til innføringen av AMS vært et felt med begrenset reell kunnskap. Velanders formel (historisk maks) har i lang tid blitt brukt til å estimere maks effekttuttak. Med AMS vet man eksakt hvor mye energi som har blitt brukt hver enkelt time, og kan dermed også slå fast maksforbruk for hver enkelt kunde.

Modell	MSE	ME
Historisk maks	2.090	0.608
Lin.Reg.OLS	1.579	-0.0003
Lin.Reg.SGD	1.567	0.0001
Beslutnings tre	2.94	0.00136
CATBoost	1.812	0.007
XGBoost	2.176	-0.004

Det er fortsatt ønskelig å kunne modellere maksimalt effekttuttak. Denne piloten har som formål å vurdere ulike metoder for estimering av maks effekttuttak. Følgende metoder blir vurdert:

- Velanders formel (historisk maks i tabellen til høyre)
- Lineær regresjon (Lin.Reg.OLS og Lin.Reg.SGD)
- Maskinlæring (beslutningstre, CATBoost og XGBoost)

Estimater fra de ulike metodene sammenlignes med målte verdier, der MSE (Means Square Error) og ME (Mean Error) brukes som mål for avvik.

Resultat:

- Lineær regresjon viser seg å gi best resultat, med lavest verdi både for MSE og ME. Dette er positivt da dette er en enklere metode enn maskinlæringsmetodene (krever mindre dataunderlag, forarbeid og regnekraft).
- Maskinlæringsmetoder treffer gjennomsnittlig bra (lav ME), men har stor variasjon i avvikene (forholdsvis høy MSE)
- Velanders metode (historisk maks) har ME på hele 0,608. Dette betyr at i gjennsnitt estimerer denne metoden maksforbruket til å være 0,6 kWh/h lavere enn det faktiske maksforbruket. Dette kan forklares med at maksforbruket hos husholdninger har gått opp de siste årene som følge av mer effektkrevende utstyr, men parameterverdiene i Velanders modell har ikke blitt justert tilsvarende.

Arbeidet er dokumentert i CINELDI-rapport [5] og har blitt presentert på webinar i CINELDI [4].

3.4 Resultater fra delaktivitet 4: Vurdering av verdi av 30-sek data (mot 5-min data) fra nettstasjon

Glitre Nett henter inn 5-min data (strøm og spenning) fra nettstasjoner. Dette gir uvurderlig innsikt i dynamiske egenskaper i strømforbruket, og er godt egnet til å fange opp variasjoner som ellers vil gå tapt dersom man baserer seg på timesmåling. Spørsmålet er om 5-min oppløsning er godt nok, eller om man også går glipp av variasjoner (topper/dipper i strøm og spenning) som man ville han fanget opp basert på målinger med høyrer frekvens.

I dette arbeidet er 5-min verdier sammenlignet med 30-sek verdier for å undersøke hvor mye mer man kan oppdage og avdekke. Resultater viser at 30-sek verdier bidrar med marginalt mer innsikt, men ikke så mye at det vil være kostnadseffektivt å oppgradere alle målinger til 30-sek verdier. Årsaken til dette ligger i dynamiske egenskaper til den underliggende energiflyten; variasjonen i spenning- og strømverdier er ikke så stor at man får vesentlig innsikt ved å gå fra 5-min til 30 sek måleintervall. Dersom man ønsker å gå til høyere dataoppløsning vil det kreve installasjon av utstyr for måling og datainnsamling, og gevinsten med høyere tidsoppløsning er ikke stor nok til å forsvare kostnaden.

Arbeidet er dokumentert CINELDI-rapport .

3.5 Innovasjoner fra Piloten

Tabell 3 Beskrivelse av innovasjoner i forskningsrådets kategorier

Forskningsrådets kategorier	Beskrivelse	Antall
Ferdigstilte nye/bedre metoder/modeller/ prototyper	Piloten gir vesentlig ny innsikt i variasjon og dynamikk i elektrisk effektpådrag. Dette gir grunnlag for å utvikle nye modeller som kan erstatte Velanders formel som metode for estimering av maks effektuttak. Vider gir piloten innspill og underlag til nye prosjekt som for eksempel FORSEL.	1
Bedrifter utenfor FMEen som har innført nye/forbedrede metoder eller modeller eller teknologi		
Bedrifter innenfor FMEen som har innført nye/forbedrede arbeidsprosesser		
Bedrifter innenfor FMEen som har innført nye/ forbedrede metoder eller modeller eller teknologi		
Inngåtte lisensieringskontrakter		
Registrerte patenter		
Ferdigstilte nye/forbedrede produkter		
Ferdigstilte nye/forbedrede prosesser		
Ferdigstilte nye/forbedrede tjenester		

Nye foretak som følge av FME'en		
Nye forretningsområder i eksisterende bedrifter		

4 Tekniske/faglige erfaringer fra Piloten

Datainnsamling:

Det har i stor grad blitt brukt AMS-data i piloten. Fra noe kunder har det blitt hentet inn 10-sek data fra HAN-porten på kundens AMS-måler. Dette har ikke gitt ekstra arbeid for prosjektet da disse kundene alt var med i en annen pilot for uttesting av utstyr for styring av forbruk, og bruk av måledata var sikret gjennom samtykke.

I tillegg ble det brukt 30-sek verdier fra Safemon-installasjoner (fra SafeBase) i tre nettstasjoner. Dette medførte behov for innkjøp og installasjon av Safemon, samt integrasjon mot SafeBase sin skyløsning for overføring av data. Dette bød ikke på større utfordringer.

Datamodellering:

Det har i stor grad vært gode data å jobbe på, med ferdige tidsserier for timesverdier. Bruk av 10-sek, 30-sek data og 5-min data har krev en del aggregering og vask av datasett. Videre har modellering til dels vært krevende, dette gjelder spesielt ved bruk av maskinlæringsmodeller. Utvelging av forklaringsvariable (inputverdier / features) krever mye arbeid, her gjøres det et stort arbeid for å finne ut hvilke variable som har størst relevans og dermed skal inngå i analysen. Det har også vist seg arbeidskrevende å fremskaffe og kvalitetssikre dataunderlag for forklaringsvariable.

5 Kost-/nyttevurderinger basert på resultatene for Piloten

Piloten har i første rekke gitt verdifull ny innsikt i effektvariasjoner og dynamikk i energiforbruk. En rekke resultat har blitt tatt frem, se Kap 3. Den konkrete nytten i resultatene ligger i stor grad i ny innsikt og forståelse av effektvariasjon. Et naturlig neste steg vil være å utvikle en metode som kan erstatte Velanders formel for estimering av maks effektuttak for en sluttkunde, og på aggregert nivå.

5.1 Kostnader

Kostnader har i stor grad vært i form av timer til ressurser som har deltatt i piloten. Det har vært noe kostnader til utstyr, men dette er lite i forhold til pilotkostnadene totalt sett. Det har også vært kostnader til lagring og prosessering av data i Azure (Glitre Nett's løsning for stordata), men det er vanskelig å tallfeste hvor mye dette er siden disse kostnadene ikke kan skilles ut for seg.

5.2 Nyttevurderinger

Nytte kommer i form av ny innsikt:

- Statistiske metoder er best for estimering av maks effektuttak for enkeltkunder
- Velanders formel underestimerer faktisk kapasitet
- Det momentane effektuttaket er rundt regnet 2x timesgjennomsnitt i høylastperioder
- 30-sekundsverdier gir marginal ekstra nytte i forhold til 5- minuttsverdier, og oppgradering til 30 sek måleintervall vil ikke være en aktuell investering for Glitre Nett.

Den direkte nytten av dette er bedre grunnlag for kapasitetsvurderinger ved analyse og planlegging av nye tilknytninger i eksisterende nett og ved bygging av nye nett.

6 Referanser

[1] Rebekka O. Omslandseter, Per-Oddvar Osland: *Stokastisk lastmodellering*. CINELDI-prosjektnotat, 26.11.2020.
[2] Rebekka O. Omslandseter: *Bruk av en stokastisk modell for modellering av forbruk*. Presentasjon 2020
[3] Aksel Holbek Sørbye, Pål Wagner: *Effektvariasjon innenfor en time - AEN Pilot CINELDI 2021*. CINELDI-prosjektnotat, 4.8.2021
[4] Rebekka O. Omslandseter, Julian Gjestvang: *CINELDI Pilot Effektanalyse AE Nett.* Presentasjon på CINELDI-webinar 25.8.2021.
[5] Julian Gjestvang: *Estimering av maksforbruk hos husholdningskunder i Agder Energi Nett via maskinlæringsmodeller og andre algoritmer*, CINELDI-prosjektnotat, 20.8.2021.
[6] Pål Wagner: *Vurdering av verdi av 30-sek data (mot 5-min data) fra nettstasjon*. CINELDI-prosjektnotat, 1.1.2022.

FME CINELDI
Host: SINTEF Energy Research in cooperation with NTNU
Visiting address: Sem Sælands vei 11, N-7034 Trondheim
Post address: P.O.Box 4761 Torgarden, N-7465 Trondheim
Telephone: +47 454 56 000*
E-mail: cineldi@sintef.no
Enterprise/VAT No: NO 939 350 675 MVA
http://www.cineldi.no

CINELDI
Centre for intelligent electricity distribution
- to empower the future Smart Grid

Prosjektnotat

TITTEL			
Stokastisk lastmodellering			
WORK PACKAGE	**VERSJON**	**DATO**	**ANTALL SIDER**
WP1/WP Pilot	1.0	2020-11-26	30
FORFATTER(E)		**WP-LEDER(E)**	**GRADERING**
Rebekka O. Omslandseter Rebekka Olsson Omslandseter (Apr 16, 2021 10:34 GMT+2) Per-Oddvar Osland Per-Oddvar Osland (Apr 28, 2021 10:42 GMT+2)		Oddbjørn Gjerde Oddbjørn Gjerde (Apr 7, 2021 16:30 GMT+2) Maren Istad Maren Istad (Apr 7, 2021 13:05 GMT+2)	CINELDI intern

DISTRIBUSJON		
CINELDI partnere		

SAMMENDRAG

I forskningssamarbeidet ift. CINELDI, har det vært uttrykt et ønske om å analysere og se nærmere på den stokastiske lastmodelleringsmetoden som ble presentert i Erling Tønnes doktorgradsavhandling [1]. Agder Energi Nett (AE Nett) har derfor sett nærmere på denne fremgangsmåten for stokastisk lastmodellering, og denne rapporten sammenfatter hovedpunktene i dette arbeidet. Vi har undersøkt maks- og gjennomsnittmetode for beregning av variasjonsverdier (verdier fra last-kurver), ulike metoder for vurdering av sannsynlighetsmodell for stokastisk justering, stokastisk justering basert på histogram av historisk avvik mellom variasjonsverdier og faktiske tempteraturjusterte verdier, stokastisk modellering et større antall ganger og analysert maks-verdiene i en slik fremgangsmåte. Vi har benyttet et datagrunnlag som består av historisk forbruk med timesoppløsning for 37 «tilfeldige» kunder. Konklusjonen er at det er tidkrevende å finne en tilpasset sannsynlighetsmodell per kunde, og resultatene i våre analyser fant ingen sannsynlighetsmodell som utpreget seg som «en god tilpasning» for alle. En løsning kan derfor være å trekke justeringsverdier til den stokastiske modellen direkte fra et histogram i stedet for fra en sannsynlighetsmodell. En periodeinndeling for de stokastiske justeringene viste seg å være nyttig, hvor man f. eks gjør en to-inndeling av året, og trekker stokastiske justeringer fra avvik som er basert på fra og med april til oktober (sommer) for timesverdier i det intervallet, og gjør tilsvarende for timesverdier fra og med november til april (vinter).

FME CINELDI - Centre for intelligent electricity distribution
Telephone: +47 454 56 000*
cineldi@sintef.no
www.cineldi.no

Host: SINTEF Energy Research in cooperation with NTNU
Visiting address: Sem Sælands vei 11, N-7034 Trondheim
Post address: P.O.Box 4761 Torgarden, N-7465 Trondheim
Enterprise/VAT No: NO 939 350 675 MVA

Innholdsfortegnelse

1 Introduksjon

I dette kapittelet presenteres forutsetningene for dette «stokastisk lastmodellerings»-prosjektet. Prosjektet startet juni 2020 og arbeidet fortsatte utover høsten 2020.

1.1 Bakgrunn

Agder Energi (AE) Nett er samarbeidspartner i CINELDI, og i dette samarbeidet har det blitt etablert tre fokusområder som AE Nett skal se nærmere på: bildeanalyse, prediksjon av belastning og effektanalyse. Stokastisk lastmodellering er en metode som er interessant innenfor hovedområdet effektanalyse og kan være særlig interessant med tanke på fremtidig dimensjonering av strømnettet.

1.2 Motivasjon

For dimensjonering av strømnettet i Norge blir det i hovedsak tatt utgangspunkt i bruk av Velander's formel og brukstid som datagrunnlag for dimensjonering. Det er også mulig å benytte standardiserte variasjonsgrafer, men dette er ikke noe som Agder Energi benytter per i dag. Velander's formel, brukstid og standardiserte variasjonsgrafer er deterministiske metoder. Selv om man i tillegg gjør sensitivitets og risikoanalyser, er metodene fremdeles deterministiske, siden de ikke tar høyde for sannsynlighet for ulike hendelses-scenarioer i sitt datagrunnlag. Velander's formel og brukstid er basert på årlig forbruk, og i et strømnett der kundene får et stadig økende og stokastisk forbruk, er det ikke nødvendigvis slik at disse metodene gir den beste tilpasningen til hvordan forbruket fordeler seg og vil fordele seg fram i tid. Det er imidlertid også viktig å påpeke at metodene som benyttes i dag er basert på kunnskap over lang tid, og at det legges mange faktorer til grunn for en faktisk dimensjonering.

Tidligere hadde nettselskapene liten kjennskap til enkeltkunders timesforbruk. Etter AMS-utrullingen, som var planlagt ferdigstilt i løpet av 2019, fikk nettselskapene mulighet til å analysere enkeltkunders faktiske forbruk med høyere oppløsning (i hovedsak timenivå, men med mulighet for også høyere oppløsning). Ved å bruke innsikten som denne nye ressursen gir oss, kan vi muligens etablere nye metoder for beregningene av datagrunnlaget vi baserer dimensjoneringen av strømnettet på i fremtiden.

I 2016 publiserte Erling Tønne sin doktorgradsavhandling som omhandlet en ny vinkling på lastmodellering. Denne metoden har en sannsynlighetstilnærming, og behandler i større grad lasten på et probabilistisk grunnlag. Dette kan gi et mer nyansert bilde av forbruket, fordi den også gir et bilde av hvordan forbruket vil fordele seg per tidsenhet avhengig av oppløsningen på datagrunnlaget. På denne måten har man mulighet til å analysere hvordan forbruket kunne fordelt seg i forhold til f. eks perioder med høyt forbruk stokastisk justert basert på et historisk avvik mellom variasjonsverdier (fra last-kurver) og faktisk forbruk. Dette gir mer informasjon enn f. eks Velander's formel, som gir oss kun én verdi å forholde seg til. Derfor er denne stokastiske lastmodelleringsmetoden interessant å se nærmere på ift. analyse og ev. dimensjonering.

1.3 Målsetning

I dette prosjektet ønsket vi å implementere og teste ut modellen til Erling Tønne for å få et bedre bilde av hvordan modellen fungerer, og hva som skal til for å f. eks ta den i bruk. Vi ønsket også å analysere denne modellens resultater sett i forhold til Velander's formel og brukstid, spesielt mtp. maksforbruk. Med dette ønsker vi å få en bedre oversikt over metoden og dens ulemper/fordeler.

1.4 Rapportens struktur

I kapittel 2 tar vi for oss de to eksisterende metodene som benyttes for dimensjonering i Agder Energi. I kapittel 3 gjør vi en detaljert gjennomgang av den stokastiske lastmodelleringsmetoden slik den

foreligger i Erling Tønne sin avhandling og slik vi har forstått den. I kapittel 4 viser vi til våre resultater for implementering av en case med 37 kunder, samt videreutviklinger vi har gjort i denne prosessen og deres resultat. I kapittel 5 diskuterer vi resultatene og videreutviklingene, samt peker på en interessante veier videre, før vi konkluderer i kapittel 6.

Notat: Merk at rapporten er forsøkt skrevet kortfattet og på en lett forståelig måte. Eventuelle spørsmål kan rettes til AE Nett v. Per Oddvar Osland.

2 Eksisterende metoder

Dette kapittelet beskriver to velkjente metoder som kan brukes i sammenheng med dimensjonering av strømnettet. Metodene beskrives kort. I [1] og [2] er det beskrevet ytterligere metoder for last-estimering, nemlig bruk av standardiserte og ikke-standardiserte last-kurver. I denne rapporten beskrives ikke bruk av last-kurver fordi metoden(e) generelt ikke benyttes til last-estimering i AE.

2.1 Velander's formel

Velander's formel gir en estimert årlig maks-last for enkeltkunders forbruk basert på summert årsforbruk for kunden. Altså en forventet maks-forbruksverdi (maks-effekt) som kan brukes videre i beregninger ift. dimensjonering.

Velander's formel for maks-last verdi for en enkeltkunde er gitt ved:

$$\hat{P}_{i,y} = k_1 W_{i,y} + k_2 \sqrt{W_{i,y}} \ ,$$

Formel 2-1

hvor $W_{i,y}$ er totalt årsforbruk (summert forbruk) for kunde i for år y. Verdiene k_1 og k_2 er konstanter som er standardisert, og har blitt justert over tid. Disse koeffisientene er bestemt for ulike kundetyper. En liste over eksempler på slike verdier er angitt nederst i dette delkapittelet. Summert forbruk kan beregnes ut ifra følgende formel:

$$\text{Summert forbruk for år } y: \quad W_{i,y} = \sum_{n=1}^{N} w_i(n)\, \alpha_{w_i(n),y} \ ,$$

Formel 2-2

$$\text{hvor } \alpha_{w_i(n),y} = \begin{cases} 1 \ hvis \ w_i(n) \ i \ år \ y. \\ 0 \ ellers. \end{cases}$$

Formel 2-3

Merk at $w_i(n)$ er den totale tidsserien til kunden. Denne tidsserien kan f. eks være for tre år, men når vi analyserer et enkelt år for kunden med Velander's formel, brukes kun verdiene for dette året.

2.1.1 Utledning av Velander's formel:

Velander's formel er basert på en antagelse om at en kundes forbruk er normalfordelt og at forbruket fra en kunde til en annen er uavhengig av hverandre og at kundene er mer eller mindre like. Disse antagelsene er vist å være rimelige under maks-last hendelser [2].

Anta at vi har en samling av kunder som oppfyller kriteriene ovenfor, og anta videre at kundens last er konstant gjennom året og at forbruket til kunde i kan representeres vha. en stokastisk variabel $X_{i,y}$ som er normalfordelt og har et standardavvik $\sigma_{i,y}$. Vi antar videre at den stokastiske variabelens gjennomsnitt er kundens gjennomsnittslast, angitt med $\bar{P}_{i,y}$.

$\hat{P}_{i,y}$ er definert som verdien $X_{i,y}$ ikke vil overstige med en viss sannsynlighet (s), f. eks 0.99 (99%). Denne verdien blir ofte referert til som maks-last verdien, og kan utrykkes som følger:

$$\hat{P}_{i,y} = \overline{P}_{i,y} + k\sigma_{i,y} \, ,$$

Formel 2-4

hor k er en konstant for å oppfylle:

$$\Pr\left(X_{i,y} \leq \hat{P}_{i,y}\right) = s \, ,$$

Formel 2-5

og s er den konfigurerte sannsynligheten for at den stokastiske variabelen for kunde i sin last-etterspørsel, $X_{i,y}$, for år y ikke overstiger $\hat{P}_{i,y}$.

Basert på antagelsene ovenfor og et system som består av totalt C kunder, kan vi si at:

$$\overline{P}_{1,y} = \overline{P}_{2,y} = \cdots = \overline{P}_{i,y} = \cdots = \overline{P}_{C,y} \, ,$$

Formel 2-6

$$\hat{P}_{1,y} = \hat{P}_{2,y} = \cdots = \hat{P}_{i,y} = \cdots = \hat{P}_{C,y} \, ,$$

og

Formel 2-7

$$W_{1,y} = W_{2,y} = \cdots = W_{i,y} = \cdots = W_{C,y} \, .$$

Formel 2-8

Ved dette kan vi videre forstå at:

$$W = CW_1 \rightarrow C = \frac{W}{W_1} \, ,$$

Formel 2-9

hvor W er sammenlagt forbruk for n kunder for et gitt år. Siden forbruket er antatt å være normalfordelt, kan Formel 2-9 også relateres til maks-lasten for et gitt år slik at vi, for samme k, har:

$$\hat{P} = \overline{P} + k\sigma \, , \qquad Pr\left(X \leq \hat{P}\right) = s \, ,$$

Formel 2-10

hvor $\hat{P}$ er maks-last, $\overline{P}$ er gjennomsnitt og σ er standardavviket av last-etterspørselen for «hele systemet» (alle kundene som analyseres). I tillegg har vi X, som er en stokastisk variabel som representer last-etterspørselen for «hele systemet». Videre kan den totale maks-lasten utrykkes som:

$$\hat{P} = C\overline{P}_i + \left(\hat{P}_i - \overline{P}_i\right)\sqrt{C} = C\overline{P}_1 + \left(\widehat{P_1} - \overline{P}_1\right)\sqrt{C} \, .$$

Formel 2-11

Ved å sette Formel 2-9 inn i Formel 2-11, får vi:

$$\hat{P} = \frac{\overline{P_1}}{W_1}W + \frac{\hat{P}_1 - \overline{P_1}}{\sqrt{W_1}}\sqrt{W} \, .$$

Formel 2-12

Dersom vi videre sier at:

$$k_1 = \frac{\overline{P_1}}{W_1} \ , \qquad k_2 = \frac{\hat{P}_1 - \overline{P_1}}{\sqrt{W_1}} \ ,$$

Formel 2-13

kan Formel 2-12 uttrykkes som:

$$\hat{P} = k_1 W + k_2 \sqrt{W} \ .$$

Formel 2-14

På denne måten har vi et uttrykk for maks-last verdien representert ved den årlige totale last-etterspørselen til «systemet» og de såkalte Velander-koeffisientene k_1 og k_2. Utledningen av Velander's formel ovenfor er hentet fra [2].

2.1.2 Velander koeffisienter og analyse av fremtidig/tidligere maks-last verdi

Av den detaljerte avledningen, forstår vi at hvordan Velander-koeffisientene er beregnet er avgjørende for vellykket bruk av den, og at ev. nye beregninger av koeffisienter kan gjøres vha. regresjonsmetodikk for et datasett man ønsker å analysere i mer detalj.

Dersom man ønsker å beregne Velander's verdi for påfølgende år, kan det legges til en forventet årlig økning. Man kan da bruke følgende formel:

$$\hat{A}_y = k_1(1+\varepsilon)W_{i,y-1} + k_2(1+\varepsilon)W_{i,y-1} \ ,$$

Formel 2-15

hvor ε, angir forventet årlig prosentandels-økning i totalt årsforbruk. Denne verdien kan justeres i forhold til erfaring eller innsamlede målinger over flere år for en enkelt kunde.

Dersom man ønsker å se på forventningsverdien for et år tidligere enn det vi har tilgjengelig (antageligvis lite brukt), kan man bruke følgende formel:

$$\hat{A}_y = k_1(1-\varepsilon)W_{i,y-1} + k_2(1-\varepsilon)W_{i,y-1} \ .$$

Formel 2-16

Tabell 1: Eksempel på Velander-konstanter hentet fra [1]. Merk at brukstidsverdiene er angitt for et spesifikt område i Norge, og er heller ikke representativt for koeffisientene brukt i avhandlingen. Antall kunder som er lagt til grunn for kolonnen til høyre er også ukjent for forfatterne av denne rapporten.

Kundetype	k_1	k_2	Brukstid for en enkeltkunde	Brukstid for flere kunder
Enebolig	0,000237	0,0119	3200	4200
Rekkehus	0,000235	0,0116	3100	4250
Leilighetsbygg	0,000264	0,0140	2150	3900
Skole	0,000410	0,1750	1600	2350
Helse og sosial pleie	0,000263	0,0790	3000	3800
Kontor	0,000270	0,0668	3000	3700
Butikk	0,000273	0,0655	2900	3650

For en enebolig bruker Agder Energi koeffisientene $k_1 = 0.00021$ og $k_2 = 0.019$.

2.2 Brukstid

Brukstid er en annen eksiterende metode for beregninger av forventet maks-last verdi (maks-effekt). Denne metoden tar utgangspunkt i totalt årsforbruk for kunden, og etablerte standardverdier for antall brukstimer for den aktuelle kundetypen. Forventet maks-last verdi basert på brukstid kan beregnes ved hjelp av følgende formel:

$$\mathrm{B}_y = \frac{W_{i,y}}{\tau} \ ,$$

Formel 2-17

hvor τ er antatt brukstid for kundetypen som kunde i tilhører. Verdiene for brukstid er standardisert, og følger gitte tabeller som kan være ulikt fra nettselskap til nettselskap. Det er også viktig å merke seg at brukstid ofte ikke benyttes for husholdningskunder, men at det gjerne benyttes til estimater for hytte-kunder. Hytte-kunder har gjerne et varierende forbruk, der kundene ikke er til stede hele tiden, og maks-last verdien kan komme skjevt ut hvis man ikke gjør noen tilpasninger.

For å beregne brukstid et år frem i tid, kan man bruke samme prinsipp som for Velander's formel, der man legger til en antatt prosentvis økning i forbruket fra et år til et annet:

$$\grave{\mathrm{B}}_y = (1+\varepsilon)\frac{W_{i,y-1}}{\tau}.$$

Formel 2-18

På samme måte kan vi beregne et år tidligere:

$$\acute{B}_y = (1-\varepsilon)\frac{W_{i,y+1}}{\tau}.$$

Formel 2-19

I doktorgradsavhandlingen til Erling Tønne er følgende verdier for brukstid listet opp:

Tabell 2: Eksempel på ulike brukstidsverdier hentet fra [1].

Kundekategori	Brukstid (τ)
Husholdning	3600
Skole	2500
Helse og omsorgsboliger	3800
Kontor	3800
Varehandel	4100
Gårdsbruk	3000

I tillegg benytter Agder Energi ofte 1/4 eller 1/8 av antall timer i et år som mål på brukstid for hytter.
Det varierer hvilken av parameterne som benyttes, og dette vurderes gjerne ut ifra hvordan forbruksmønsteret til kunden tidligere har fordelt seg (med en manuell vurdering).

3 Beskrivelse av stokastisk lastmodelleringsmetode

I dette kapittelet beskriver vi Erling Tønne sin modell ift. hvordan vi har forstått den. Vi vil presentere fremgangsmåten stegvis, og hvert delkapittel utgjør en bestand-del av metoden.

3.1 Datagrunnlag

For å benytte stokastisk lastmodellering som metode, må man ha et datagrunnlag som er i samsvar med forventningene til resultatet av modelleringen. Dersom man bare har et år med data, vil man ha et dårligere statistisk grunnlag enn dersom man har flere år med data til rådighet. Hvis man har et år med data, vil man bare ha et eksempel på en januar-måned, mens man har fem eksempler på januar-måned med fem år med data. Likevel er det ikke nødvendigvis slik at mer data automatisk fører til en bedre forståelse av hvordan forbruket er frem i tid. For eksempel kan det ha skjedd en flytting, slik at det datagrunnlaget representerer, ikke lenger er representativt for det som vil skje fremover, eller at man har et «spesial-år» som avviker fra resten og påvirker modellen i negativ retning.

Man kan også ha manglende data. Dette kan håndteres med standard interpoleringsteknikker eller andre kjente metoder for å fylle inn verdier som mangler i datasettet. Man kan også velge å utelate manglende verdier i datasettet. Da minsker man sannsynligheten for at en interpoleringsteknikk der det er mye manglende data påvirker datagrunnlaget på en uønsket måte. Hva som er «beste tilnærming» ift. dette er avhengig av datagrunnlaget og hvilke spesifikasjoner man gjør i modellen.

Hvilken tidsoppløsning man velger/har tilgjengelig for modelleringen, avgjør hvilken tilnærming man kan gjøre til beregningen av variasjonsverdiene og på den måten også hvilke verdier modellen returnerer. Hvis man f. eks har et datagrunnlag med timesoppløsning per målepunkt, vil man kunne lage variasjonsverdier som representerer time for time per målepunkt. Da vil modellen også kunne returnere verdier med timesoppløsning. Hvis man har verdier med 15 minutters-oppløsning, vil man kunne lage variasjonsverdier med 15-minutters oppløsninger etc.

3.2 Steg 1: Temperaturjustering

Første steg i stokastisk lastmodellerings-metoden er å gjøre en temperaturjustering av forbruksdataen. Temperaturjustering gjør datagrunnlaget temperatur-uavhengig. I praksis vil dette gjøre at en høy forbruksverdi når det er veldig kald temperatur, justeres litt ned. Målet er altså at tidsserien skal være representativ uavhengig av temperaturendringer. Dvs. at dersom man ønsker et temperatur-avhengig resultat, må man justere verdiene tilbake til å ta hensyn til temperaturforandringer etter den stokastiske modelleringen.

Tabell 3 Eksempel på ulike verdier for κ [1].

Kundetype	Eldre enn 1950	1951-1970	1971-1988	1989-1998	1999-2008	2009-2011	2012-2014	Passiv-hus
Lite hus	0.75	0.7	0.6	0.5	0.5	0.35	0.30	0.25
Skole	0.65	0.6	0.55	0.5	0.45	0.4	0.35	0.3
Sykehus	0.45	0.4	0.35	0.4	0.35	0.25	0.20	0.15
Hotell	0.55	0.5	0.45	0.45	0.35	0.35	0.3	0.25
Butikk	0.5	0.45	0.4	0.4	0.3	0.25	0.25	0.25
Verksted	0.7	0.65	0.6	0.55	0.55	0.5	0.4	0.35

Temperatur-justeringen gjøres vha. følgende formel:

$$w_i(n) = \omega_i(n) + \left(P_i \kappa \delta (T_n - T_i)\right), \forall n ,$$

Formel 3-1

hvor $w_i(n)$ er den temperaturkorrigerte versjonen av verdi n for kunde i ($\omega_i(n)$), og dersom man har et datagrunnlag med timesverdier, vil $w_i(n)$ altså være den temperatur-korrigerte verdien av time n. Temperatur-avhengighetsandelen er angitt med verdien κ, og eksempler på denne verdien er angitt i Tabell 3. Fordi annet ikke er oppgitt, antar vi at disse verdiene for κ kan brukes for data med timesoppløsning. δ er temperatursensitiviteten, og kan settes til å være 0.05 [1]. T_n er normaltemperaturen for den spesifikke dagen verdi n er hentet fra, og T_i er gjennomsnittlig temperatur for de tre siste dagene, dvs. dagen verdi n er hentet fra, og to dager før dette.

Ifølge [1], kan man hente ut normaltemperatur, T_n , fra Norges Meteorologiske Institutt (MET). Vi har tolket T_n som gjennomsnittlig temperatur på en spesifikk dag i året (f. eks 1. Jan), hentet ut alle temperaturene for denne dagen et antall år bakover i tid og beregnet gjennomsnittet av denne totalen. T_i har vi behandlet som et 72 timers løpende gjennomsnitt (vårt datagrunnlag har timesoppløsning).

3.3 Steg 2: Beregning av variasjonsverdier for etablering av lastkurver

Andre steg handler om å finne ut hvordan årlig forbruk dag for dag og time for time (eller med høyere oppløsning) pleier å se ut, ut ifra en andel av maks-lasten eller gjennomsnittslasten til enkelt-kunden. En last-kurve etableres for et tidsintervall og består av variasjonsverdier[1]. Beregningen av variasjonsverdiene for last-kurvene består av to deler. Første del handler om hvilke perioder last-kurvene består av (som er en vurdering ut ifra størrelsen på datagrunnlaget). Andre delen handler om hvordan verdiene beregnes, hvor man enten bruker det vi har valgt å kalle gjennomsnittsmetoden eller maks-metoden.

I doktorgradsavhandlingen til Erling Tønne er følgende metoder for periodeinndeling av last-kurvene og beregning av variasjonsverdier listet opp:

- Lastkurve-metode 1: En daglig last-kurve (alle timene i et døgn) for hver av dagene i en uke (man, tir, … , søn) per måned.
  Eksempel gjennomsnittsmetode (samme prinsipp for maks-metode – bytt gjennomsnittslast med makslast):
  $$\text{Variasjonsverdi} = \frac{gjennomsnittslast\ for\ mandager\ i\ januar\ klokka\ 13.00}{gjennomsnittslast\ \text{totalt}}.$$
  Dette gir 7x12=84 last-kurver og 24x7x12=2016 variasjonsverdier totalt.

- Lastkurve-metode 2: En daglig last-kurve for ukedager per måned (man-fre) og en daglig last-kurve for helgedager (lør-søn + helligdager) per måned.
  Eksempel gjennomsnittsmetode (samme prinsipp for maks-metode – bytt gjennomsnittslast med makslast):
  $$\text{Variasjonsverdi} = \frac{gjennomsnittslast\ for\ ukedager\ i\ januar\ klokka\ 13.00}{gjennomsnittslast\ \text{totalt}}.$$

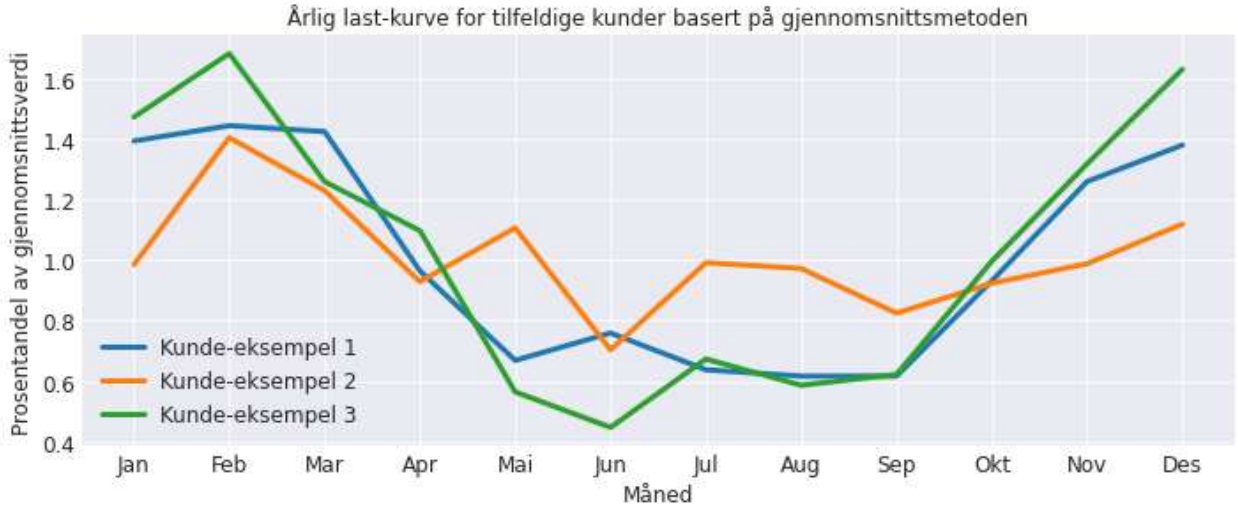
[1] Ofte betegnes disse verdiene som «variasjonskurver» eller «last-kurver» direkte. Vi har valgt å bruke betegnelsen variasjonsverdier, og når man legger variasjonsverdiene «etter hverandre» i en graf/liste utgjør de en «last-kurve».

Dette gir 2x12=24 last-kurver (en lastkurve for ukedager per måned og en last-kurve for helgedager per måned) og 24x2x12=576 variasjonsverdier totalt.

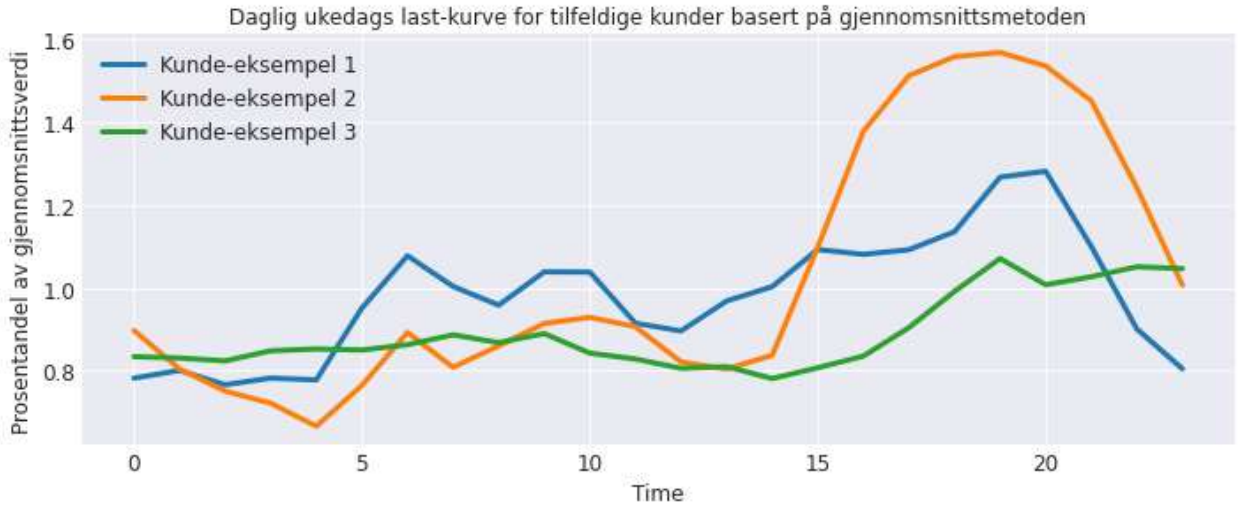
- Lastkurve-metode 3: En kombinasjon av en daglig last-kurve for ukedager (man-fre), en daglig last-kurve for helgedager (lør-søn + helligdager) og en årlig last-kurve (jan, feb, … , des). Eksempel gjennomsnittsmetode (samme prinsipp for maks-metode – bytt gjennomsnittslast med makslast):

$$\text{Variasjonsverdi} = \frac{gjennomsnittslast\ for\ ukedager\ klokka\ 13.00}{gjennomsnittslast\ \text{totalt}}.$$

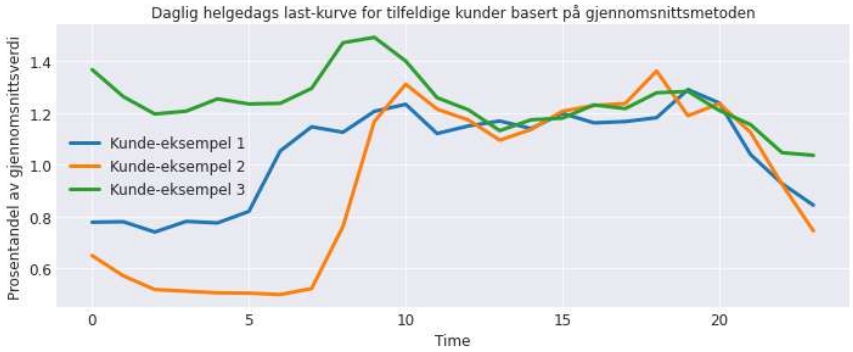
Dette alternativet gir tre last-kurver og (2x24)+12=60 variasjonsverdier totalt.



Figur 1 Eksempler på årlig lastkurve (lastkurve-metode 3 og gjennomsnittsmetoden).



Figur 2 Eksempler på daglig lastkurve for ukedag (lastkurve-metode 3 og gjennomsnittsmetoden).



Figur 3 Eksempler på daglig lastkurve for helgedag (lastkurve-metode 3 og gjennomsnittsmetoden).

I doktorgradsavhandlingen er det argumentert for at Metode 1 er best og mest nøyaktig. Samtidig er det denne metoden som krever mest data. Dersom vi har to år med data, vil vi med Metode 1 basere f. eks gjennomsnittet vårt med gjennomsnittsmetoden på ca. 4 eksempler for mandager i januar klokka 12.00 per år, som gir oss et statistisk grunnlag for mandager i januar klokka 12 på 4x2=8. Dersom vi i stedet deler inn i ukedager/helligdager per måned (Metode 2), har vi ca. 20x2=40 verdier å bruke for å beregne gjennomsnittet for time 12.00 en hverdag i januar. Med Metode 3, har vi enda flere eksempler per variasjonsverdi vi regner ut, hvor vi har ca. 20x12x2=480 eksempler på time 12.00 en hverdag generelt og ca. 24x30x12x2 eksempler for å beregne den årlige last-kurven. Hvilken metode man velger er en vurderingssak ift. datagrunnlag man har tilgjengelig og hvor stort statistisk grunnlag man ønsker å basere seg på.

I forsøkene vi har gjort brukte vi lastkurve-metode 3 fordi vi bare hadde litt under tre år med data og ønsket størst mulig datagrunnlag per variasjonsverdi. Dersom man har få verdier å basere variasjons-verdien på vil man kanskje ikke fange opp den faktiske oppførselen, men spesialtilfeller i stedet.

I de matematiske fremstillingene av gjennomsnittsmetoden og maksmetoden nedenfor har vi brukt lastkurve-metode 3 som grunnlag, men samme prinsipp gjelder for beregning av variasjonsverdier for andre type kurver også.

3.3.1 Gjennomsnittsmetoden

La oss si at vi har en tidsrekke med total lengde N_i for kunde *i*. Denne tidsrekken kan ha verdier som f. eks består av gjennomsnittlig effekt innafor en time (for forbruksverdier med timesoppløsning) angitt med variabelen $w_i(n)$ (temperaturjustert verdi). Hver verdi har et tilhørende tidsstempel angitt med $t_i(n)$, hvor $n \in [1, 2, \dots, N_i]$, og angir element nummer n i den sorterte tidsserien, hvor tidsserien er sortert fra tidligste tidsstempel (først) til seneste tidsstempel (sist) og tidsserien består av N_i verdier totalt. C er, slik definert tidligere, totalt antall kunder.

Steg 1: Beregne variasjonsverdiene for lastkurvene (periodene) som er valgt, basert på gjennomsnittet per periode (nedenfor er lastkurvene for lastkurve-metode 3 listet opp).

Variasjonsverdier for årlig lastkurve:

Gjennomsnittlig forbruk for måned m: $$\overline{M_{i,m}} = \frac{\frac{\sum_{n=1}^{N_i} w_i(n) M_{w_i(n),m}}{\sum_{n=1}^{N_i} M_{w_i(n),m}}}{\frac{\sum_{n=1}^{N_i} w_i(n)}{N_i}}, \forall m \ ,$$

hvor $$M_{w_i(n),m} = \begin{cases} 1 \ hvis \ t_i(n) i \ måned \ m. \\ 0 \ ellers. \end{cases}$$

Formel 3-2

Variasjonsverdier for daglig lastkurve for ukedag:

$$\text{Gjennomsnittlig forbruk for time } h \text{ i en ukedag: } \overline{U_{\iota,h}} = \frac{\frac{\sum_{n=1}^{N_i} w_i(n) U_{w_i(n),h}}{\sum_{n=1}^{N_i} U_{w_i(n),h}}}{\frac{\sum_{n=1}^{N_i} w_i(n)}{N_i}} \ , \forall h \ ,$$

$$\text{hvor } U_{w_i(n),h} = \begin{cases} 1 \ hvis \ t_i(n) \ i \ time \ h \ for \ en \ ukedag \\ 0 \ ellers \end{cases}$$

Formel 3-3

Variasjonsverdier for daglig lastkurve for helgedag:

$$\text{Gjennomsnittlig forbruk for time } h \text{ i en helg/helligdag: } \overline{H_{\iota,h}} = \frac{\frac{\sum_{n=1}^{N_i} w_i(n) H_{w_i(n),h}}{\sum_{n=1}^{N_i} H_{w_i(n),h}}}{\frac{\sum_{n=1}^{N_i} w_i(n)}{N_i}}, \forall h \ ,$$

$$\text{hvor } H_{w_i(n),h} = \begin{cases} 1 \ hvis \ t_i(n) \ i \ time \ h \ for \ en \ helg \ eller \ helligdag. \\ 0 \ ellers. \end{cases}$$

Formel 3-4

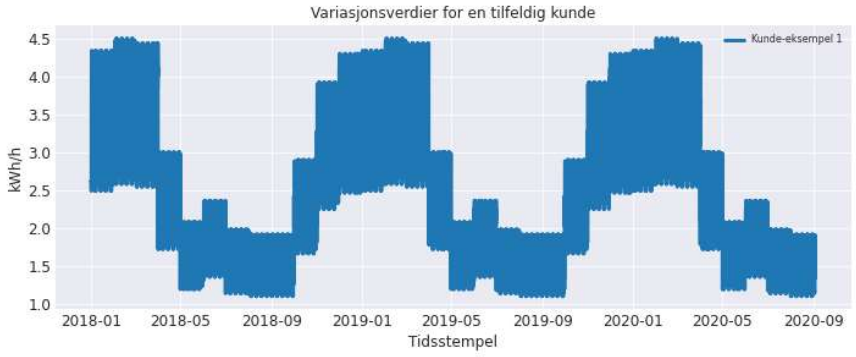
Steg 2: Sammenkoble variasjonsverdiene fra de ulike kurvene:

$$v_i(n) = \left(\frac{\sum_{n=1}^{N_i} w_i(n)}{N_i} \right) M_{w_i(n),m} \overline{M_{\iota,m}} \left(U_{w_i(n),h} \overline{U_{\iota,h}} + \ H_{w_i(n),h} \overline{H_{\iota,h}} \right) \ , \forall n.$$

Formel 3-5

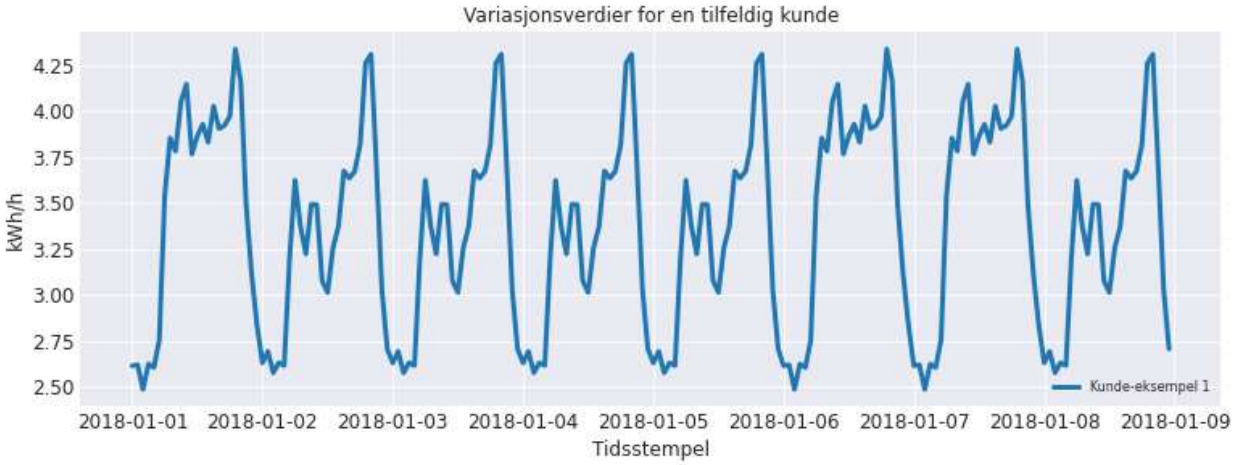
De endelige variasjonsverdiene, betegnet med $v_i(n)$, beregnes for alle n man har tilgjengelig. Slik vi ser av formelen ovenfor multipliseres gjennomsnittsverdien totalt med verdien fra den årlige lastkurven og verdien fra den daglige lastkurven for enten ukedag eller helgedag avhengig av om $t_i(n)$ er en ukedag eller helgedag (samt hvilken måned den tilhører). Eksempel: Vi ønsker å beregne $v_i(n)$ for mandag 01.01.2020 klokka 12.00. Dette er en ukedag i januar-måned. Da multipliserer vi det totale gjennomsnittet for kunden med variasjonsverdien for januar fra årlig last-kurve og variasjonsverdien for klokka 12.00 fra daglig ukedagskurve.

Her har vi antatt at tidsrekken til hver kunde er fullstendig. Dersom man har en verdi/tidsstempel som mangler, kan man gjøre tiltak ift. dette. Fordi man ser på gjennomsnittet over flere verdier, er det ikke nødvendigvis kritisk at en enkeltverdi mangler.



Figur 4 Eksempel på variasjonsverdier for en tilfeldig kunde med lastkurve-metode 3 og gjennomsnittsmetoden. Legg merke til at mønsteret gjentas per år.

Leseren bør notere seg at de variasjonsverdiene vil gjenta seg i et ukedagskurve- og helgedagskurve-mønster som legger seg i ulike høyder ut ifra nivået til den årlige lastkurven de kobles sammen med (og vil være samme for alle år).



Figur 5 Eksempel på variasjonsverdier for en tilfeldig kunde med lastkurve-metode 3 og gjennomsnittsmetoden zoomet inn på en kortere tidsperiode (8 dager). Legg merke til at 1. januar er en helligdag, og at ukedagene og helgedagene har ulik profil.

3.3.2 Maksmetoden

Maksmetoden er lik som gjennomsnittsmetoden, men i stedet for å basere utregningene på gjennomsnittsverdien, ser man på maksverdien.

Steg 1: Beregne variasjonsverdiene for lastkurvene (periodene) som er valgt, basert på maks-verdien per periode (nedenfor er lastkurvene for lastkurve-metode 3 listet opp).

Først må vi finne maks-verdiene per periode

Maksimum last-verdi for måned m: $\widehat{M_{i,m}} = \underset{1 \le n \le N_i}{max} \left(w_i(n) M_{w_i(n),m}\right)$

Formel 3-6

Maksimum last-verdi for time h i en ukedag: $\widehat{U_{i,h}} = \underset{1 \le n \le N_i}{max} \left(w_i(n) U_{w_i(n),h}\right)$

Formel 3-7

Maksimum last-verdi for time h i en helg/helligdag: $\widehat{H_{i,h}} = \underset{1 \le n \le N_i}{max} \left(w_i(n) H_{w_i(n),h}\right)$

Formel 3-8

Deretter må man finne variasjonsverdiene per kurve ved å dele på den totale maks-verdien. Resultatet er en andelsverdi per måned eller time ift. maks-verdien.

Variasjonsverdier for årlig lastkurve:

$$V_{i,m} = \frac{\widehat{M_{i,m}}}{\max(w_i(n))}, \forall\, m\,.$$

Formel 3-9

Variasjonsverdier for daglig lastkurve for ukedag:

$$\dot{V_{i,h}} = \frac{\widehat{U_{i,h}}}{\max(w_i(n))} \ , \forall\, h\,.$$

Formel 3-10

Variasjonsverdier for daglig lastkurve for helgedager:

$$\ddot{V_{i,h}} = \frac{\widehat{H_{i,h}}}{\max(w_i(n))} \ , \forall\, h\,.$$

Formel 3-11

Steg 2: Sammenkoble variasjonsverdiene fra de ulike kurvene:

$$v_i^*(n) = \left(\max_{n\,\in[1,N_i]}(w_i(n)\right) V_{i,\,m} M_{w_i(n),\,m}\left(\dot{V_{i,h}} U_{w_i(n),\,h} + \ \ddot{V_{i,h}} H_{w_i(n),\,h}\right) \ , \forall n.$$

Formel 3-12

3.4 Steg 3: Avviks-analyse og fremstilling av avvik

Etter at man har funnet de endelige variasjonsverdiene, analyseres avviket mellom disse verdiene og de faktiske temperaturjusterte verdiene til kunden. Dette avviket benyttes deretter videre til å justere de endelige variasjonsverdiene til å likne de faktiske verdiene. Dette gjøres ved å trekke verdier fra en sannsynlighetsmodell tilpasset avviket eller fra en histogramframstilling av avviket. Ved å trekke verdier på denne måten er tanken at man utforsker handlingsrommet til kunden ift. hvordan et ev. forbruk kunne sett ut.

I dette steget kan man gå for to ulike tilnærminger. I doktorgradsavhandlingen er det fremgangsmåten med bruk av en tilpasset sannsynlighetsmodell som er presentert. Å finne riktig sannsynlighetsmodell for avviket kan være vanskelig, samt at prosesseringstiden for dette kan være lang (mer detaljert om dette i diskusjonsdelen), og derfor foreslår vi bruk av histogramfremstilling som et alternativ. Derfor, har man i dette steget to muligheter for fremstilling av avviket som skal benyttes videre i den stokastiske modelleringen:

- Tilpasset sannsynlighetsmodell for fremstilling av avviket.
- Histogramfremstilling av avviket.

Fra fremstillingen av avviket trekker man verdier som man bruker i den stokastiske modelleringen. Avviksverdiene trekkes da enten fra et histogram av observert avvik, eller en tilpasset sannsynlighetsmodell for avviket. Disse avviksverdiene benyttes deretter til å justere de endelige variasjonsverdiene funnet i 3.3.

Avviket kan beregnes med følgende tilnærming:

$$\phi_i(n) = \begin{cases} \dfrac{w_i(n) - v_i(n)}{v_i(n)} \ for\ variasjonsverdier\ med\ gjennomsnittsmetoden. \\ \dfrac{w_i(n) - v_i^*(n)}{v_i^*(n)} \ for\ variasjonsverdier\ med\ maksmetoden. \end{cases}$$

Formel 3-13

For å beregne gjennomsnittsavviket kan man bruke følgende formel:

$$\overline{\phi_i(n)} \begin{cases} \frac{1}{N}\sum_{n=1}^{N}\frac{w_i(n)-v_i(n)}{v_i(n)} \; for\ gjennomsnittsmetoden \\ \frac{1}{N}\sum_{n=1}^{N}\frac{w_i(n)-v_i^*(n)}{v_i^*(n)} \; for\ maksmetoden. \end{cases}$$

Formel 3-14

Legg merke til at vi ikke bruker gjennomsnittsavvik direkte i denne modellen (selv om f. eks gjennomsnitt er en parameter for «Gaussian Distribution»). Avviksverdiene er angitt som prosentandeler.

Når man har beregnet avviket, må man fremstille avviket i et histogram. Man deler da inn avviksverdiene mellom minimum og maksimum avvik i λ intervaller. Deretter teller man hvor mange av avviksverdiene som er innenfor hvert intervall η hvor $\eta \in [0, \lambda]$:

$$f_{i,\eta} = \sum_{n}^{N_i} r_{\phi_i(n),\eta} \quad \forall\, \eta, \qquad \eta \in [0, \lambda]\,,$$

$$hvor\ r_{\phi_i(n),\eta} = \begin{cases} 1\ hvis\ \phi_i(n)\ er\ i\ intervall\ \eta. \\ 0\ ellers. \end{cases}$$

Formel 3-15

Resultatet blir at man får et sett med intervaller, og et angitt antall verdier som er innenfor hvert intervall, kalt frekvens, betegnet med $f_{i,\eta}$ (for intervall η og kunde i).

Deretter bruker man enten dette histogrammet til å finne sannsynlighetsmodellen som best passer avviket, eller bruker histogramfremstillingen direkte inn i neste steg. Merk at dersom man f. eks bruker innbygd funksjonalitet for å vurdere en sannsynlighetsmodell sin tilpasning til datagrunnlaget, gjøres ikke dette nødvendigvis fra histogrammet, men f. eks for Chi-Square tilnærmingen er det forskjellen mellom histogrammet og sannsynlighetsmodell-grafen som analyseres.

For histogramfremstilling av avviket foreslår vi å lage histogram fra avvikene basert på en periodeinndeling. Det er ikke nødvendigvis slik at det er like stort avvik mellom variasjonsverdiene og de temperaturjusterte verdiene sommertid som vintertid, eller i januar kontra oktober. Eksempler på periodeinndelinger man kan gjøre er som følger (per kunde):

- Ingen periodeinndeling: Man samler alle avviksverdiene til et histogram, og bruker samme histogrammet for å justere alle variasjonsverdiene i neste steg.
- 12 årlige perioder: Man lager et histogram av avviket per måned. Det vil si at man samler alle avviksverdiene for tidsstemplene i januar uavhengig av år, og lager et histogram av disse verdiene. Deretter trekker man fra dette histogrammet når det gjelder justering av verdier som er i januar-måned.
- 4 årlige perioder: Man lager fire histogrammer med inndelingen: (jan-mar), (apr-jun), (jul-sep), (okt-des). Det vil si at man samler alle avviksverdiene for tidsstemplene i f. eks (jan-mar) uavhengig av år, og lager et histogram av disse verdiene. Deretter trekker man fra dette histogrammet når det gjelder justering av verdier som er i perioden (jan-mar).
- 2 årlige perioder: Man lager to histogrammer med inndelingen (okt-mar) og (apr-sep). Det vil si at man samler alle avviksverdiene for tidsstemplene i f. eks (okt-mar) uavhengig av år, og lager et histogram av disse verdiene. Deretter trekker man fra dette histogrammet når det gjelder justering av verdier som er i perioden (okt-mar), og tilsvarende for (apr-sep).

Man velger altså en periodeinndeling man vil gjøre ut ifra datagrunnlaget man har. Dersom man har mindre data er kanskje en høyere oppløsning en dårligere løsning, men dette er en vurderingssak ift. hvilke forventninger man har for modellen. Slike periodeinndelinger kan for øvrig gjøres for sannsynlighetsfremstillingene av avviket også, men det er ikke noe vi har sett på i vårt arbeid.

I delkapittelet nedenfor er vurderingskriterier for å analysere hvor god en sannsynlighetsmodell sin tilpasning til avviket er diskutert i korte trekk.

3.4.1 Vurderingskriterier for sannsynlighetsmodellens tilpasning

Det finnes ulike metoder for å vurdere hvilken sannsynlighetsmodell som passer avviket best.

Eksempler på parameter for vurdering av en sannsynlighetsmodells-tilpasning:

- Chi-Square
- SIC (Schwarz information criterion, aka Bayesian information criterion BIC)
- AIC (Akaike information criterion)
- HQIC (Hannan-Quinn information criterion)

Når man har funnet den mest passende sannsynlighetsmodellen for avviket, kan man bruke denne videre i steg 4, hvor man justerer de endelige variasjonsverdiene og får det som er resultatet av den stokastiske lastmodelleringen.

I doktorgradsavhandlingen ble det benyttet Excel-tillegget ModelRisk. Dette programmet returnerer SIC, AIC og HQIC for manuell bedømmelse av mest passende sannsynlighetsmodell. Vi har ikke benyttet dette verktøyet, og har i hovedsak benyttet Chi-Square-målet i stedet, selv om vi også testet de andre parameterne.

3.5 Steg 4: Stokastisk lastmodellering

Neste, og i utgangspunkt siste steg av den stokastiske lastmodelleringen, er å justere variasjonsverdiene med en stokastisk verdi (denne verdien er enten trukket fra sannsynlighetsmodellen eller histogrammet funnet i forrige steg). I stedet for å beskrive denne prosessen matematisk, finner man en litt mer detaljert gjennomgang i kapittel 4.

De stokastisk modellerte verdiene/lasten genereres på følgende måte (resultatverdiene/vårt «endelige» svar):

$$s_i(n) = \begin{cases} v_i(n) * (1+\theta) \; for \; gjennomsnittsmetoden, \\ v_i^*(n) * (1+\theta) \; for \; maksmetoden, \end{cases}$$

Formel 3-16

hvor $\theta \in [-\infty, \infty]$ er en verdi som generes fra sannsynlighetsmodellen eller fra histogrammet som ble funnet i forrige steg. Dersom man bruker histogrammet som basis, vil avviksverdiene ikke gå utenfor intervallene som er funnet. Dvs. at $\theta \in [\min(\phi_i(n)), \max(\phi_i(n))]$, og at usannsynlig høye eller lave stokastisk modellerte verdier «lukes ut» automatisk (fordi det ikke kan skje et større eller mindre avvik enn det som er blitt observert). Slik er det ikke dersom man benytter seg av en sannsynlighetsmodell.

F. eks, dersom man har en normaldistribusjon som beste sannsynlighetsdistribusjon for avviket, trekker man en verdi fra denne distribusjonen med de spesifikke parameter som passer til avviket mellom variasjonsverdiene og faktiske verdier, og justerer variasjonsverdiene for denne timen basert på denne verdien. Deretter gjør man det samme for alle timene for den gitte kunden. Dersom man ønsker at disse verdiene skal være bundet innenfor et intervall, og ikke skal kunne ta «hvilken som helst» verdi, må man benytte en sannsynlighetsmodell som enten bare er definert innenfor et visst intervall (f. eks truncnorm), eller manuelt «sette et tak» for hva verdiene ikke kan overstige.

For at de stokastisk modellerte verdiene ikke skal kunne overstige maksimal installert effekt hos kunden, definerte Erling Tønne i sin doktorgradsavhandling at verdier over en viss verdi ble satt til maksimal installert effekt hos kunden. I våre forsøk har vi i noen tilfeller ikke lagt noe bånd på verdiene for å kunne analysere modellen i større detalj, mens vi i andre tilfeller har trukket en ny verdi dersom det viser seg at den stokastiske verdien som ble trukket gjør at den stokastisk modellerte verdien overstiger en viss verdi (f. eks 25 kWh).

3.6 Analyse og evaluering

For å analysere hvordan den stokastiske lastmodelleringen gjør det, finnes det mange metoder man kan benytte. I doktorgradsavhandlingen ble det benyttet følgende konsept for å gjøre en evaluering av den stokastiske lastmodelleringen:

$$\gamma_i(n) = \begin{cases} \dfrac{s_i(n) - w_i(n)}{v_i(n)} \; for\ gjennomsnittsmetoden. \\ \dfrac{s_i(n) - w_i(n)}{v_i^*(n)} \; for\ maksmetoden. \end{cases}$$

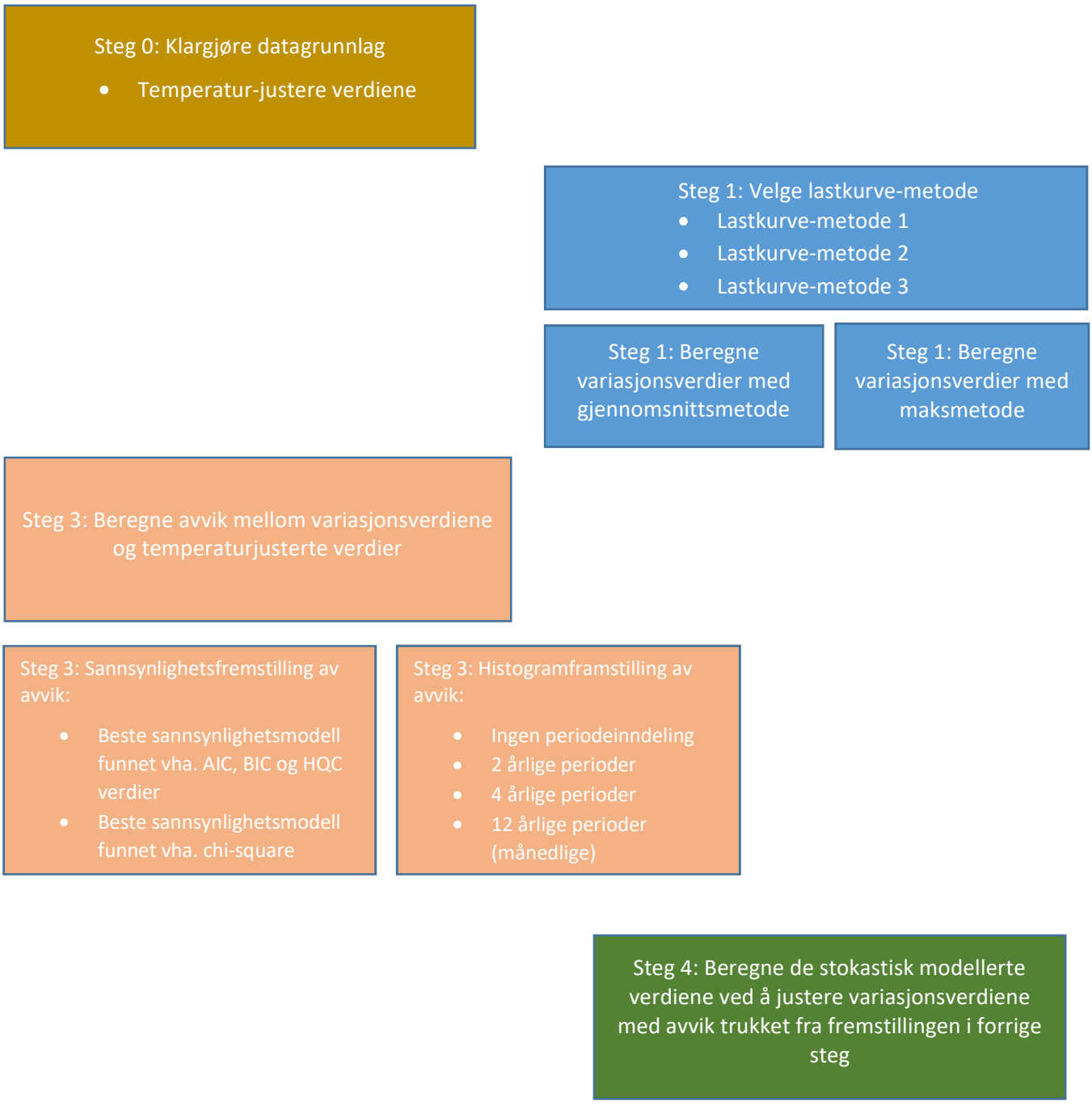
Formel 3-17

I våre analyser valgte vi å benytte «Mean Absolute Percentage Error, forkortet MAPE (selv om den oftest brukes for prediksjoner) og gjennomsnittlig avvik mellom den stokastisk modellerte verdien og de temperaturjusterte verdiene. MAPE er en velkjent parameter, og gir en enkelt-verdi som gjør at man kan sammenlikne den stokastiske modelleringen for en kunde opp mot en annen på en enkel måte. Man kan også se på $\overline{\gamma_i(n)}$ *.*

3.7 Oppsummering av stokastisk lastmodellering

Stokastisk lastmodellering består altså av følgende deler:

- Steg 0: Klargjøre datagrunnlag ved å ev. forbedre datakvalitet og temperatur-justere verdiene.
- Steg 1: Velge lastkurvemetode og enten gjennomsnittsmetode eller maksmetode for beregning av variasjonsverdier.
- Steg 2: Beregne avvik mellom variasjonsverdier og temperaturjusterte verdier, og fremstille dette avviket enten vha. en histogramfremstilling eller sannsynlighetsmodell. Histogrammene kan lages med en periodeinndeling.
- Steg 3: Beregne de stokastisk modellerte verdiene ved å justere variasjonsverdiene med en avviksverdi som trekkes fra histogrammet eller sannsynlighetsmodellen funnet i forrige steg.



4 Testcases og resultater

For å demonstrere stokastisk lastmodellering gjorde vi en testcase med 37 «tilfeldige» kunder som er med i en testpilot hos AE.

4.1 Datagrunnlag

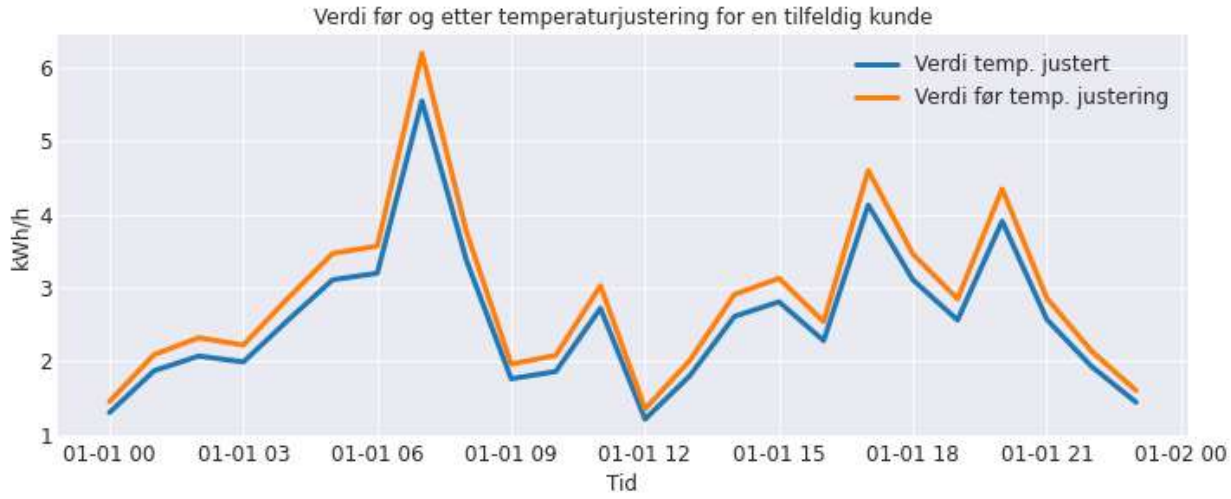
Datagrunnlaget besto av forbruksverdier fra 37 kunder med timesoppløsning. For kundene som er i testpiloten har vi mulighet for høyere tidsoppløsning, men vi valgte å benytte timesoppløsning i våre analyser, fordi det er det som er gjort i doktorgradsavhandlingen. Derfor har vi heller ikke gjort noen analyse på hvilken tidsoppløsning som gir best resultater med den stokastiske lastmodelleringsmetoden.

4.2 Steg 1: Temperaturjustering

Første oppgave var å temperatur-justere forbruksdataen. Vi benyttet følgende konfigurasjon og forenklinger, basert på verdiene oppgitt i [1]:

- Temperaturverdier ble hentet fra lokasjonen Kjevik, selv om kundene er distribuert over flere steder i AE sitt nett.
- For κ satte vi verdien til å være 0.5 (lite hus 1989-2008)
- Temperatursensitiviteten ble satt til å være $\delta = 0.05$.

I Figur 6 ser vi et eksempel på hvordan verdiene så ut for en tilfeldig kunde etter temperaturjustering.



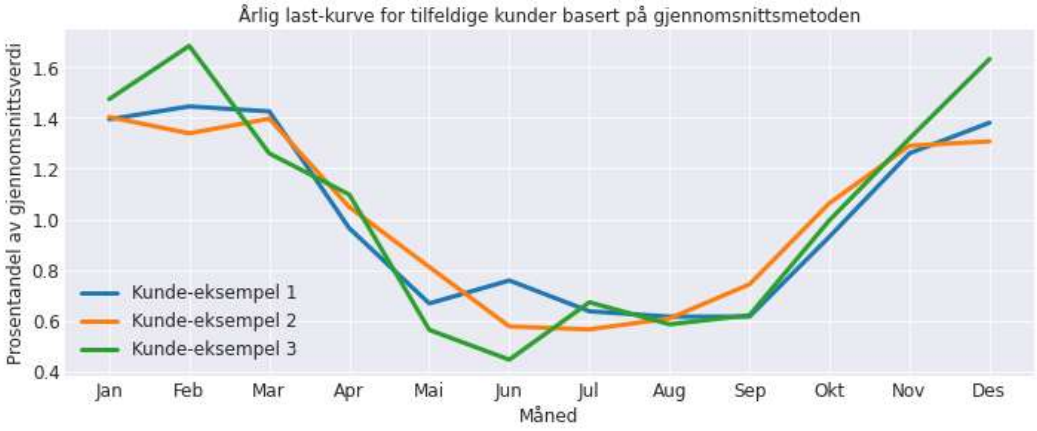
Figur 6 Eksempel på temperaturjustering.

Merk: Videre i rapporten refererer «faktisk verdi» til den temperaturjusterte verdien.

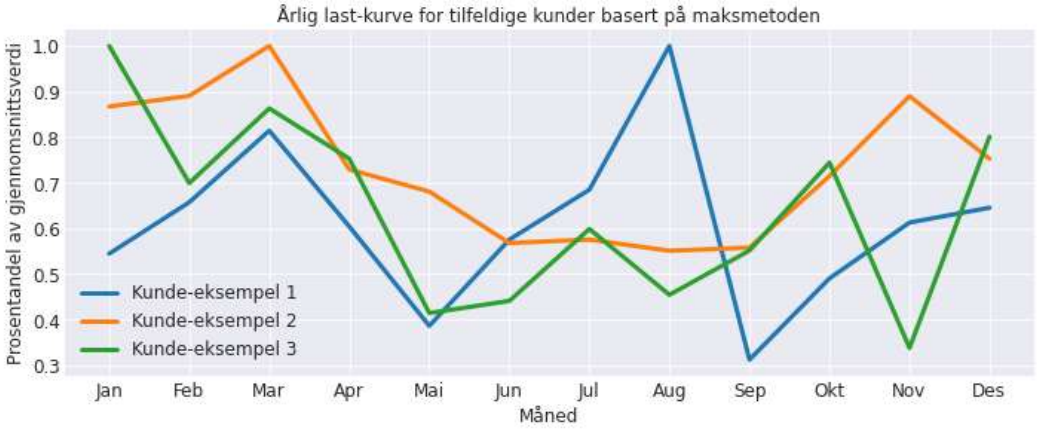
4.3 Steg 3: Variasjonsverdiberegning

For å beregne variasjonsverdiene benyttet vi både gjennomsnitts og maksmetoden. Vi observerte at det var stor forskjell mellom lastkurvene for de ulike metodene. Dette gir utslag i variasjonsverdiene og hvordan avviksverdiene mellom de faktiske verdiene og variasjonsverdiene fordeler seg. Vi valgte å bruke lastkurve-metode 3.

Den klare forskjellen mellom lastkurvene beregnet med gjennomsnitts og maksmetode, ser man tydelig i Figur 7 og Figur 8. Figurene viser resultater for de samme kundene med de ulike metodene.

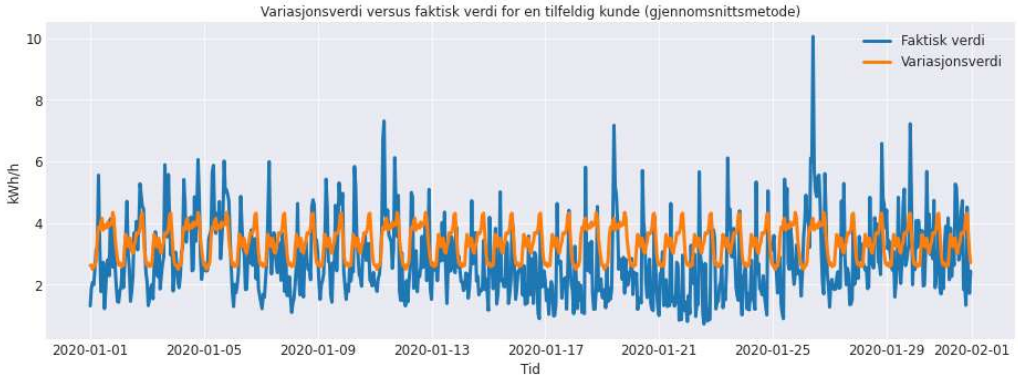


Figur 7 Eksempel på lastkurver for tilfeldige kunder basert på gjennomsnittsmetoden.

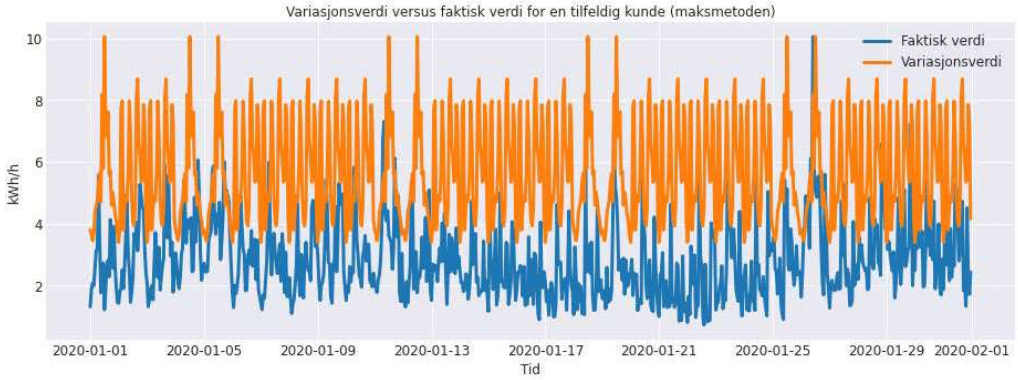


Figur 8 Eksempel på lastkurver for tilfeldige kunder basert på maksmetoden.

Man ser også en distinkt forskjell mellom de beregnede variasjonsverdiene basert på gjennomsnitts og maksmetoden. Denne forskjellen er veldig tydelig i Figur 9 og Figur 10, hvor maks-metoden generelt gir et resultat som ligger over de faktiske verdiene og gjennomsnittsmetoden gir variasjonsverdier som er sentrert mer i midten av de faktiske verdiene. Figurene viser resultater for de samme kundene for de ulike metodene.



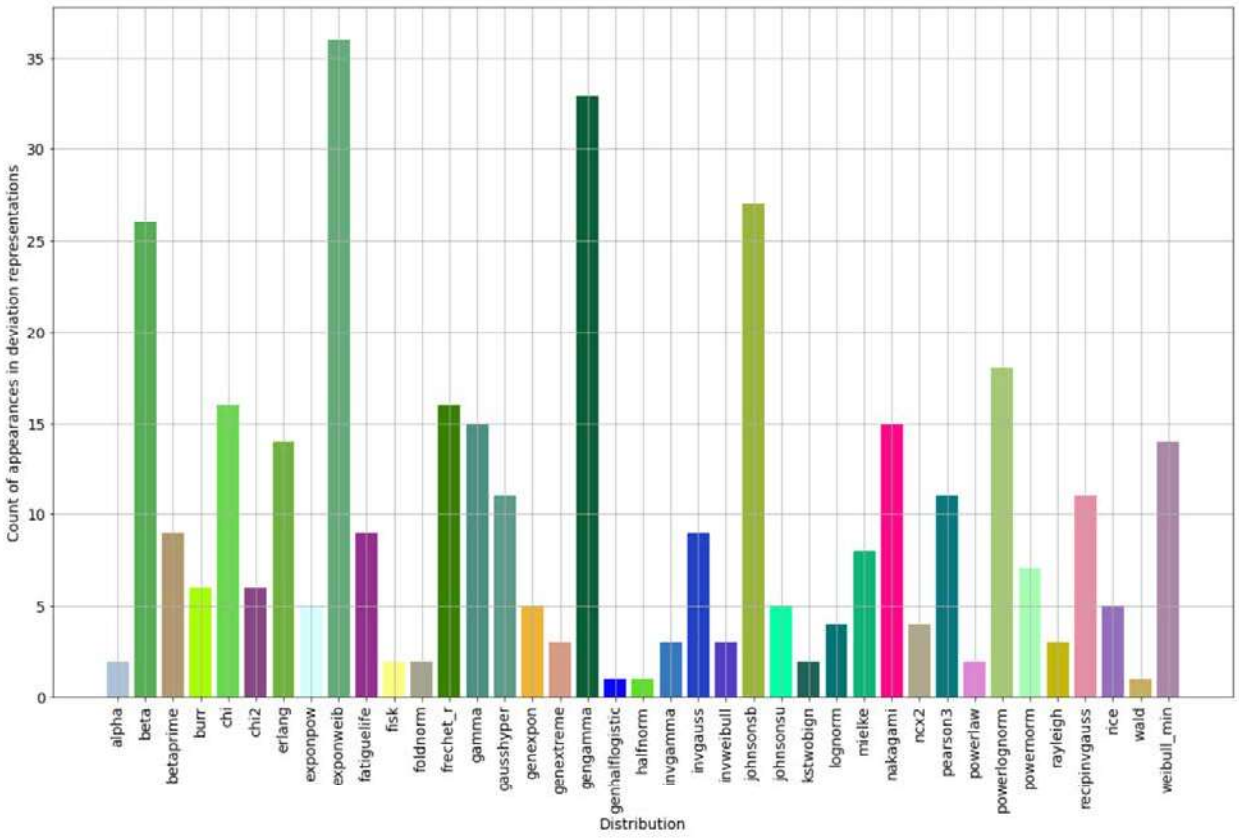
Figur 9 Eksempel på variasjonsverdier for en måned i 2020 for en tilfeldig utvalgt kunde. Variasjonsverdiene er her beregnet med gjennomsnittsmetoden.



Figur 10 Eksempel på variasjonsverdier for en måned i 2020 for en tilfeldig utvalgt kunde. Variasjonsverdiene er her beregnet med maksmetoden.

4.4 Steg 3: Avviks-analyse og fremstilling

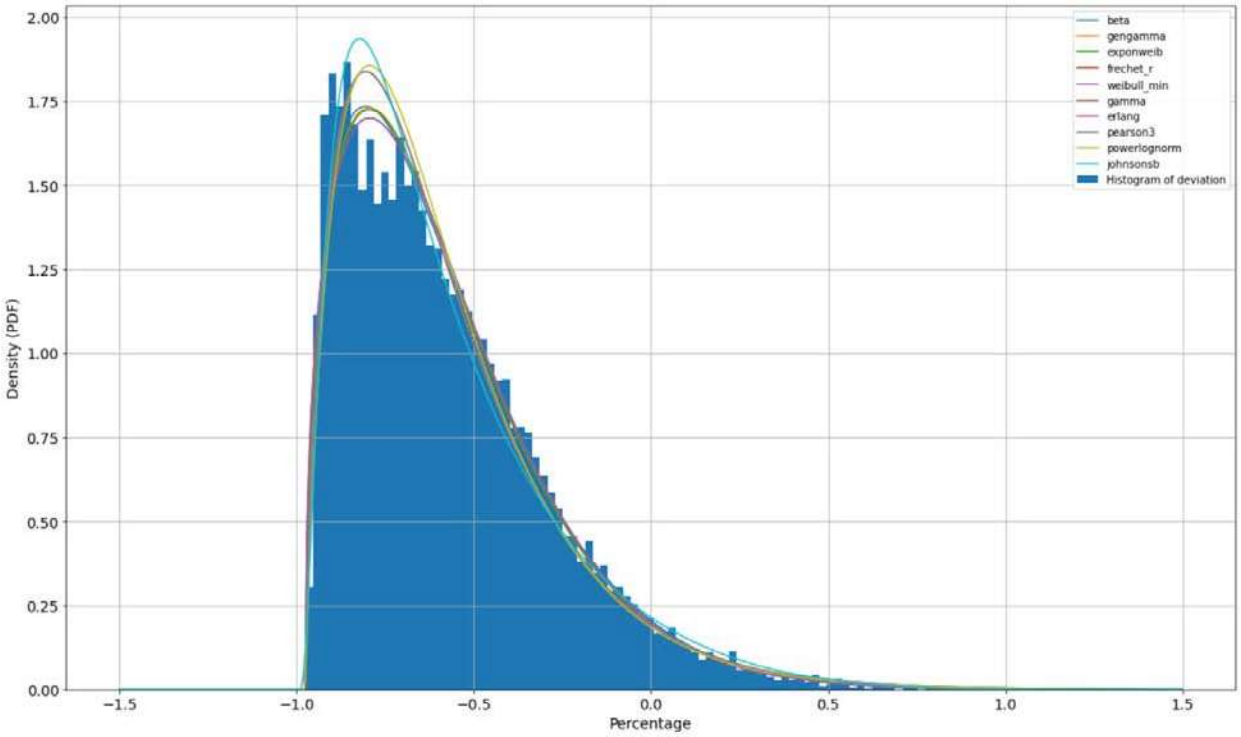
Denne delen av prosessen med stokastisk last-modellering handler om å fremstille avviket på en valgt måte, som det skal trekkes fra i Steg 4. Vi testet ulike metoder for å oppnå dette. Dette steget gir kanskje mest mening når man ser det i sammenheng med resultatene av den stokastiske lastmodelleringen i Steg 4, men i dette delkapittelet diskuterer vi hvordan tilpasningene av avviket til sannsynlighetsmodeller ble, samt viser noen histogramfremstillinger av avviket.



Figur 11 Oversikt over topp 10 distribusjoner for fremstilling av avviket for de 37 kundene.

4.4.1 Sannsynlighetsfremstilling av avvik

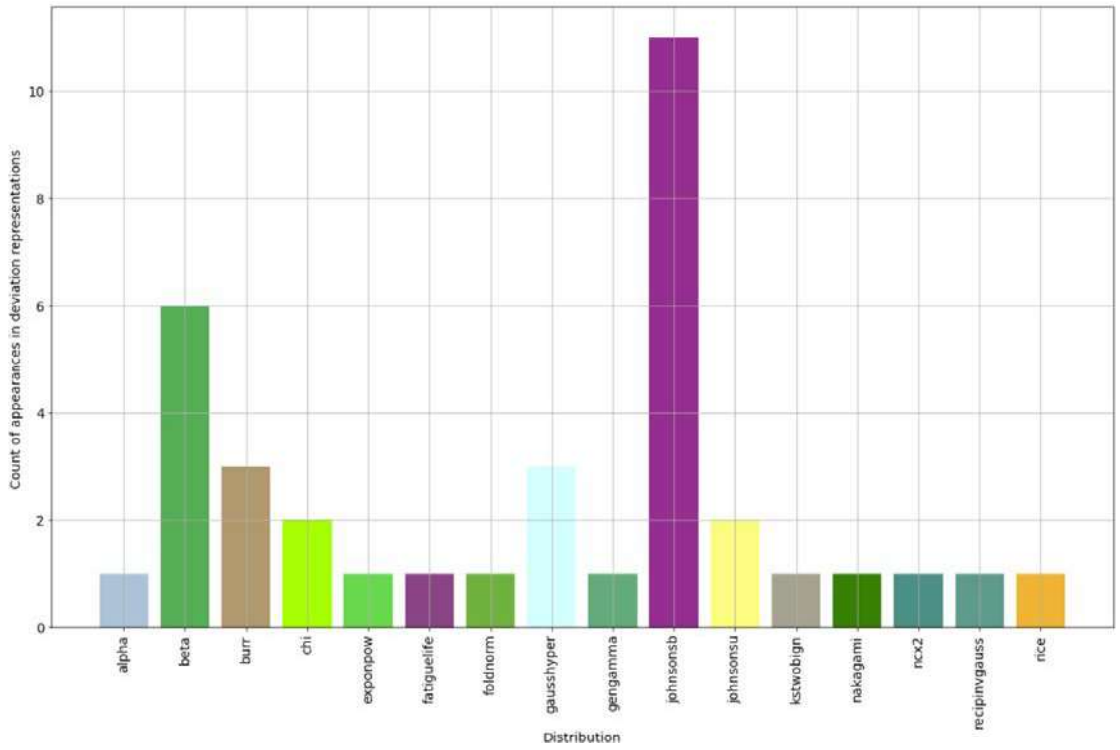
I kapittel 3 påpekte vi at det er flere måter man kan framstille avviket mellom faktiske verdier og variasjonsverdiene. Her presenteres noen av resultatene ift. å bruke en sannsynlighetsfremstilling av avviket. Slik vi ser i Figur 11 er det noen distribusjoner som utmerker seg når man ser på topp 10 distribusjoner for fremstilling av avviket mellom faktiske verdier og variasjonsverdiene for de 37 kundene. F. eks ser vi at «exponential weibull» er på topp 10 listen til nesten alle kundene (36 stykker). Tilpasningen til disse distribusjonene er basert på deres AIC-verdi, og vi har hentet ut de 10 distribusjonene med lavest AIC-verdi for alle kundene, og funnet hvor mange ganger de enkelte distribusjonene forekommer i «topp 10». Dvs. vi har hentet ned en liste med topp 10 distribusjoner for de 37 kundene, funnet de distribusjonene som er unike for kundene totalt sett, og telt opp hvor mange ganger hver enkelt forekommer.



Figur 12 Eksempel på avvikshistogrammet til en tilfeldig kunde og de ti best tilpassede sannsynlighets-distribusjonene basert på AIC-verdi.

I Figur 13 ser vi en fremstilling av beste distribusjon funnet vha. «chi-square» basert på topp 10 distribusjoner basert på AIC-verdi. Dvs. vi beregnet AIC verdien for alle sannsynlighetsdistribusjonene, listet nederst i rapporten, sin tilpasning til avviket, og hentet ut de ti distribusjonene med lavest AIC-verdi for alle kundene. Deretter testet vi disse distribusjonenes tilpasning til avviket vha. «chi-square». Vi ser da at «johnsonsb» distribusjonen var beste tilpasning for flere av kundene (11 stykker).

Vi har ikke tatt noe standpunkt i om det finnes «en» sannsynlighetsdistribusjon man kan benytte som er en god tilpasning for «alle», fordi vårt datagrunnlag var begrenset («bare» 37 kunder). Samtidig observerte vi at sannsynlighetsfremstillinger av avviket kom veldig dårlig ut for noen kunder. F. eks kunder som hadde et histogram med en u-form. Samtidig tok beregningen av distribusjonenes tilpasning lang prosesseringstid. Derfor anså vi histogramfremstilling av avviket som en god, og kanskje bedre tilnærming til avviksfremstilling. Analysene gir likevel en forståelse av avvikets fordeling, som generelt sentrerte seg i nedre sjiktet (nær -1) av avviksverdier for maksmetoden.

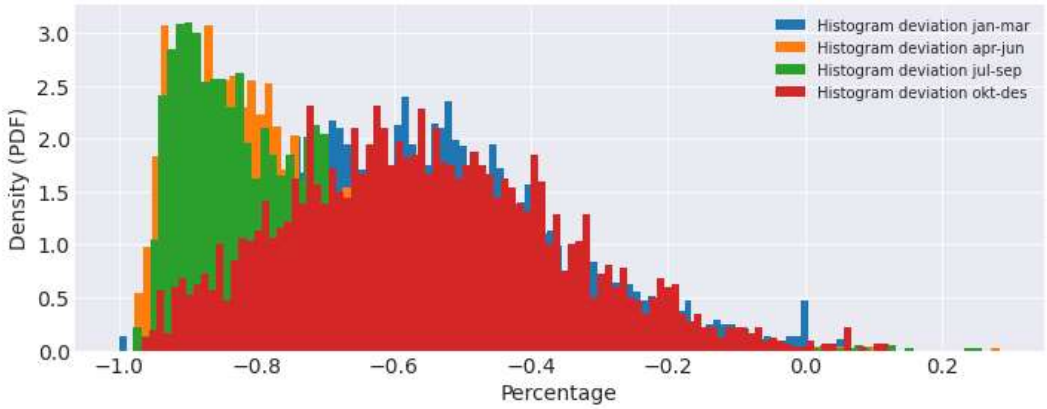


Figur 13 Topp distribusjon og antallet forekomster når vi bruker «chi-square» vurdering i forhold til distribusjonens tilpasning med utgangspunkt i topp 10 distribusjoner funnet vha. AIC.

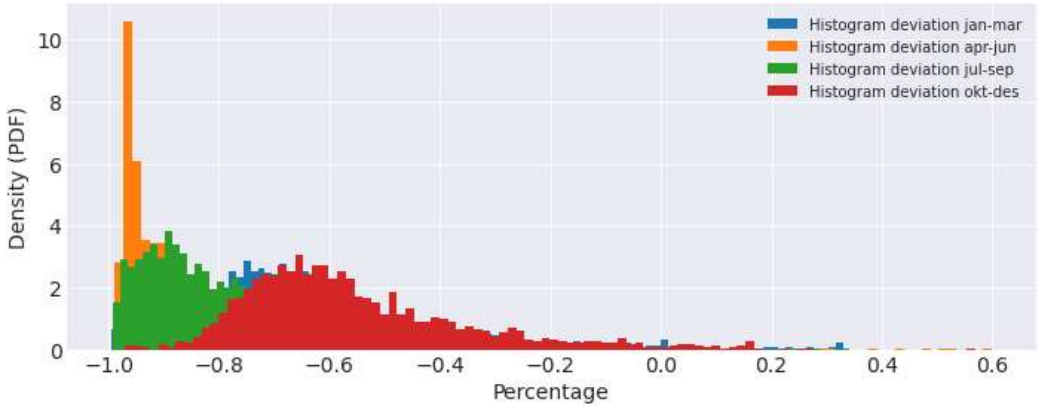
4.4.2 Histogramframstilling av avvik

I stedet for å tilpasse en sannsynlighetsmodell til avviket mellom faktiske verdier og variasjonsverdiene kan man gjøre en histogram-fremstilling av avviket og trekke avviksverdier direkte fra denne fremstillingen for a generere de stokastiske verdiene i steg 4.

I Figur 14 og Figur 15 ser vi hvor forskjellig avviket kan fordele seg mellom de ulike tidsperiodene av et år. Dette taler til fordel for at man i stedet for å trekke avviksverdier fra hele datagrunnlaget, trekker verdier fra histogram som har en periodeinndeling. Dvs. at man f. eks deler inn avviket i en periode fra og med januar til og med mars, og bruker dette histogrammet for å stokastisk modellere verdier som er innenfor dette intervallet. Dette må vurderes ift. hvor stort datagrunnlag man har.



Figur 14 Histogram av avvik mellom faktisk verdi og variasjonsverdi for en tilfeldig kunde inndelt etter perioder.



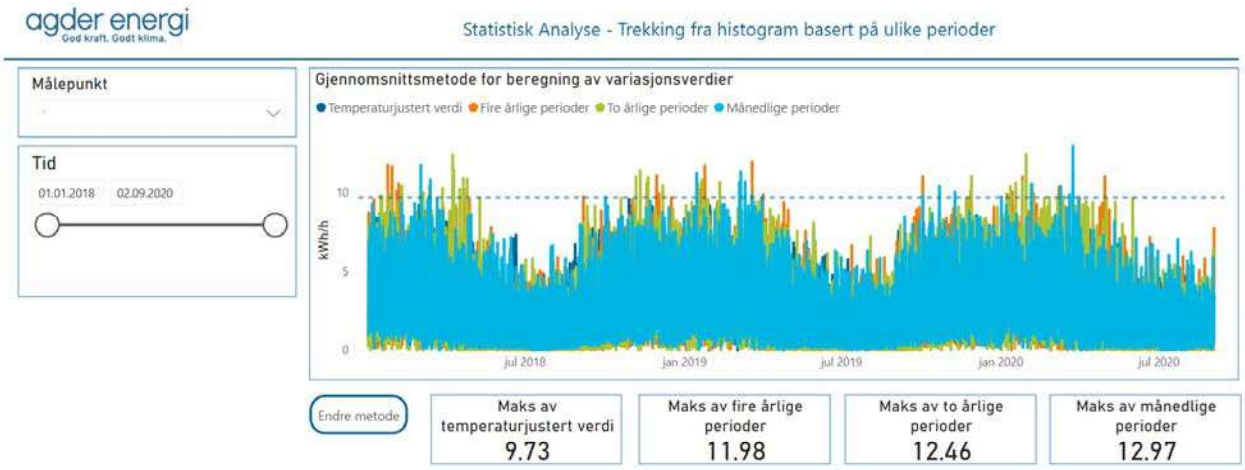
Figur 15 Eksempel på et histogram av avviket til en annen kunde enn i Figur 14. Avviket er også her inndelt i perioder.

Vi har ikke konkludert rundt hvilken periodeinndeling som lønner seg mest. Intuitivt er det nok slik at man vil treffe stadig nøyaktigere dess mindre inndelinger man gjør, men at dette må sees ift. datagrunnlagets størrelse. Dersom man kun har et år med data, vil en månedsvisperiodeinndeling f. eks gi et sparsomt statistisk grunnlag.

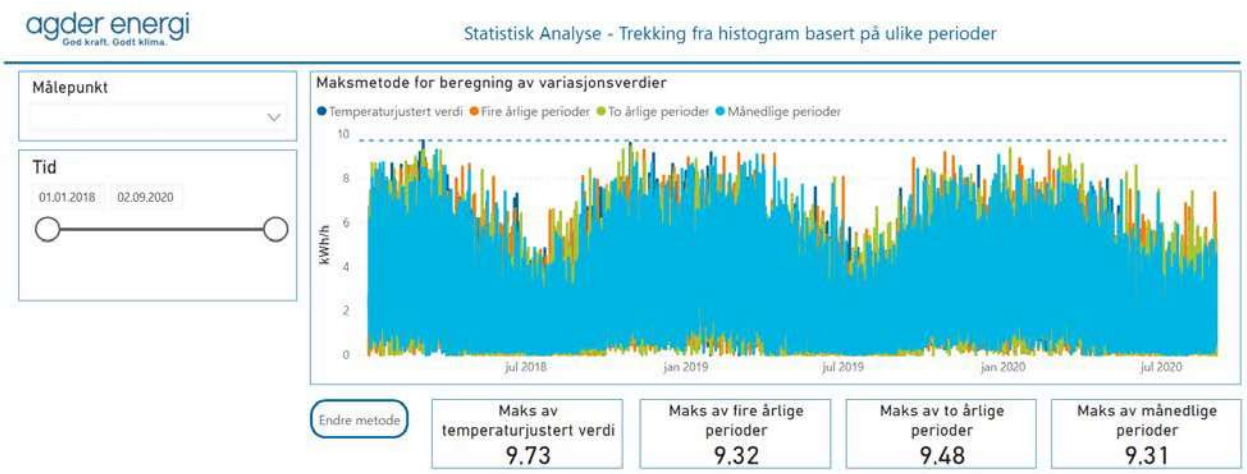
4.5 Steg 4: Stokastisk lastmodellering

Siste steg av metoden er å generere de stokastisk modellerte verdiene. Dette gjøres ved å trekke verdier fra enten sannsynlighetsmodellen eller histogrammet som ble funnet for avviket mellom faktiske verdier og variasjonsverdier. Verdiene som trekkes blir deretter brukt til å justere variasjonsverdiene. På denne måten justeres variasjonsverdiene til å ta høyde for en statistisk usikkerhet, og man kan utforske større deler av handlingsrommet ift. kundens last når man f. eks gjør et stort antall modelleringer.

I Figur 16 og Figur 17 vises resultatet av flere stokastiske modelleringer med ulik periodeinndeling av histogrammet. Følgelig kan man observere hvordan maksverdiene fordeler seg ift. differansen mellom modellert maksverdi og faktisk maksverdi. Figurene viser resultatet fra gjennomsnitts og maksmetoden for en tilfeldig utvalgt kunde. Det er samme kunde som er fremstilt i Figur 16 og Figur 17.



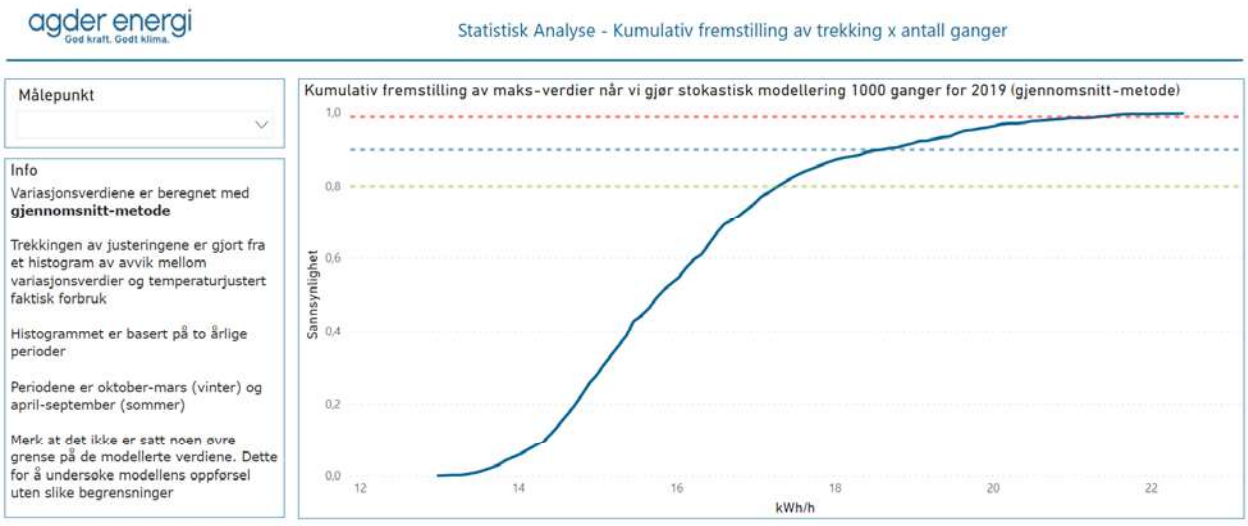
Figur 16 Fra tilhørende Power BI for analyse av ulike periodeinndelinger av avvikshistogrammet man trekker fra. Her er det brukt gjennomsnittsmetode.



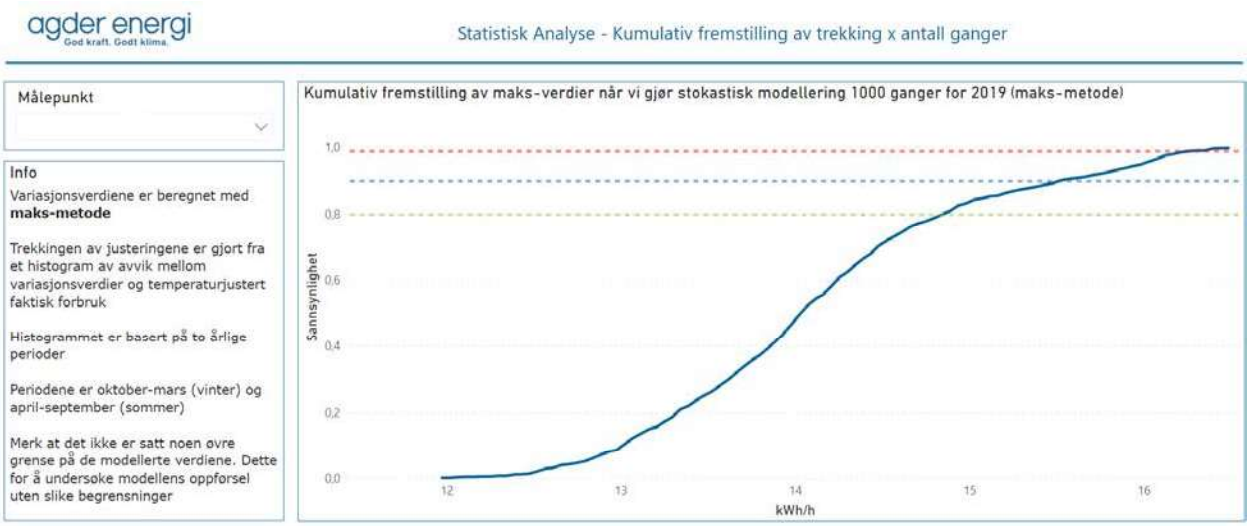
Figur 17 Fra tilhørende Power BI for analyse av ulike periodeinndelinger av avvikshistogrammet man trekker fra. Her er maksmetoden brukt, og visualiseringen er for samme kunde som i Figur 16.

Ved å gjøre den stokastiske modelleringen et større antall ganger kan man utforske handlingsrommet i større grad enn det man får gjort med en modellering. Da kan man hente ut maksverdien per modellering. Deretter kan man fremstille maksverdien som i en kumulativ sannsynlighetsfordeling. Følgelig kan man velge et sannsynlighetsnivå man vil legge seg på ift. risiko. F. eks kan man bestemme at man ønsker å basere videre arbeid på et 50-prosents nivå. Da vil man, med de stokastiske modelleringene lagt til grunn og et visst antall modelleringer, lese av verdien ved 50% sannsynlighet og bruke denne verdien som referanse på at maks-lasten til kunden med 50% sannsynlighet ikke vil være høyere enn denne verdien.

Figur 18 og Figur 19 viser en slik kumulativ sannsynlighetsfremstilling av makslasten når man har modellert 2019 1000 ganger, med gjennomsnittsmetoden og maksmetoden. Slik vi ser av grafene, har gjennomsnittsmetoden en høyere makslast enn det maksmetoden har. Generelt, vha. manuell vurdering, så vi at gjennomsnittsmetoden oftere traff dårligere enn det maksmetoden gjorde på de faktiske verdiene.



Figur 18 Kumulativ sannsynlighetsfremstilling for maksverdi for en tilfeldig kunde med gjennomsnittsmetode. Dette er for 1000 modelleringer av 2019.

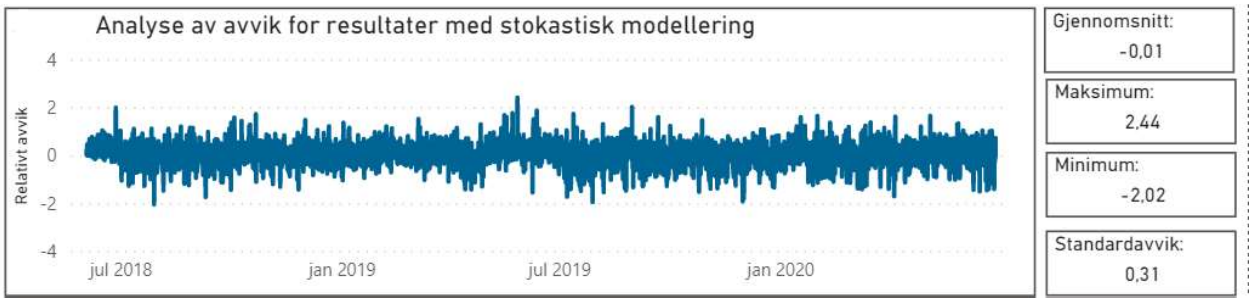


Figur 19 Kumulativ sannsynlighetsfremstilling for maksverdi for en tilfeldig kunde med maksmetode. Dette er for 1000 modelleringer av 2019.

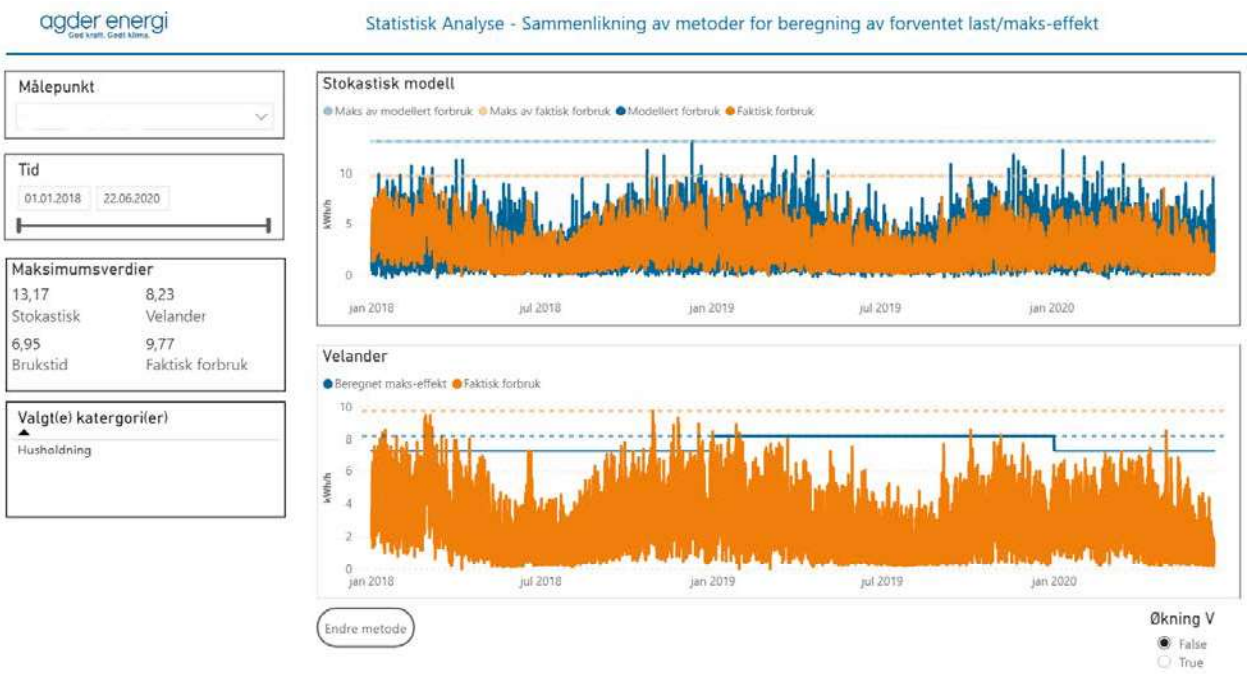
Merk: Figur 18 og Figur 19 er laget uten noen øvre grense på den stokastisk modellerte verdien. Dette er for å analysere hvordan modellen «faktisk» oppfører seg. I et bruks-scenario, ville det være naturlig å sette en limit, slik at de modellerte verdiene ikke overstiger det som er teoretisk mulig for kunden.

4.6 Analyse og evaluering

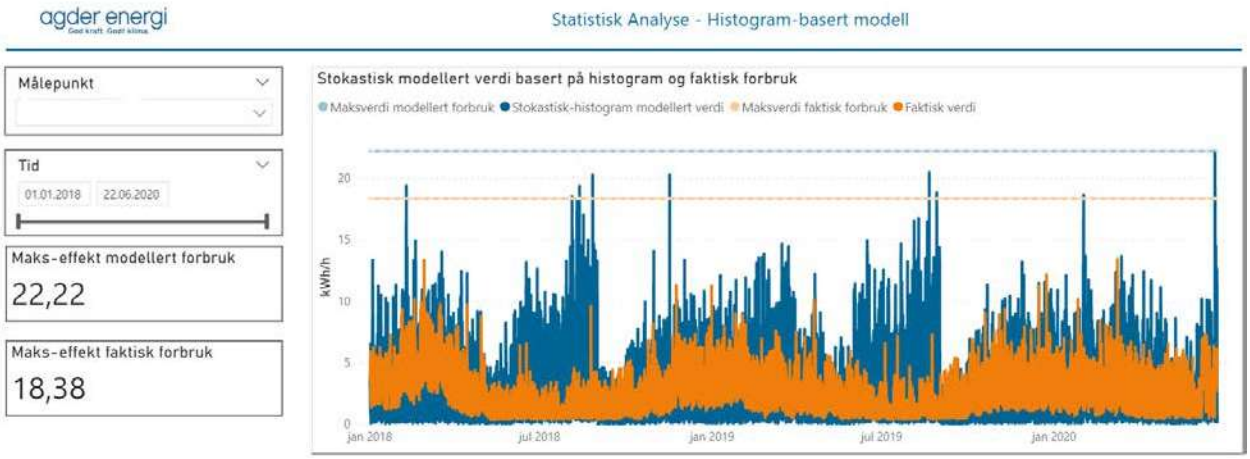
For å evaluere modellen har vi laget ulike Power BI rapporter der vi kan analyse resultatene av den stokastiske modellen. I Figur 21 og Figur 22 kan man se eksempler på slike rapporter. Mesteparten av analysene har vært gjort manuelt vha. disse rapportene og dersom vi hadde hatt mer tid til rådighet i dette prosjektet, ville vi ha gjort enda dypere analyser av hvordan den stokastiske modellen representerte kundens faktiske last. F. eks over flere år. Dvs. at man utelater 1-2 år fra modelleringen, og ser hvordan modelleringen representerer dette forbruket.



Figur 20 Eksempel på relativt avvik beregnet etter Formel 3-17 [1].



Figur 21 Eksempel på resultater med stokastisk modell versus Velander for en stokastisk modellering der 10 tilfeldige distribusjoner fra den totale mengden distribusjoner var testet for å finne beste modell, og avviksverdiene ble trukket fra beste modell ut ifra dette (kun som eksempel).



Figur 22 Eksempel på resultat med histogram-basert avvikstrekking der vi ikke har noen periodeinndeling og heller ikke noen maks-last som kan modelleres. Slik vi ser av resultatet, gir den stokastiske modellen en verdi som er godt over faktisk verdi. I gjennomsnitt en MAPE-verdi på 2.73.

Generelt virket det som at den stokastiske modellen traff bedre enn Velander og brukstid metodene, men at den oftere over-estimerte. Dette ble bedre når vi inndelte avviket i perioder og trakk fra histogram.

5 Diskusjon og videreutvikling

Vi har testet mange kombinasjoner av tilnærminger til stokastisk lastmodellering i vårt arbeid. Under nevnes noen av tilnærmingene og utfordringene og de diskuteres i mer detalj. Forslag til videreutviklinger presenteres også.

5.1 Trekking fra histogram versus trekking fra sannsynlighetsmodell

Noe av utfordringen var prosesseringstiden for å tilpasse en sannsynlighetsmodell til avviket mellom variasjonsverdier og de faktiske forbruksverdiene til kunden. Derfor foreslo vi å trekke stokastiske verdier direkte fra histogrammet av avviket. Dette effektiviserte bruken av den stokastiske lastmodelleringen, fordi man slapp å kjøre igjennom et stort antall mulige sannsynlighetsmodeller per kunde og sammenlikne hvordan hver modell passet til avviket. Samtidig, dersom man har et vanskelig avvik, f. eks et som er todelt, er de fleste sannsynlighetsmodeller uansett en dårlig tilpasning. Ved å trekke direkte fra histogrammet slipper man utfordringen med at selv sannsynlighetsmodellen med best tilpasning til avviket er en dårlig fremstilling.

Det er mulig å gjøre en antagelse om at avviket er normalfordelt, som kan få ned arbeidsmengden med å teste å sammenlikne modeller, men vil kanskje gi en dårligere representasjon av avviket enn det vi oppnår med trekking fra histogram. En sannsynlighetsmodell kan muligens gi større spenn i avviksverdier, og på den måten får utforsket mer av handlingsrommet. Likevel er det slik at i histogram-tilnærmingen trekker vi fra en uniform distribusjon av verdier innenfor intervallene, og avviksverdier man ikke har sett før kan forekomme. På denne måten har man fremdeles en stokastisk tilnærming, men innenfor minimum og maksimum av avviksverdier man «har sett før». Eventuelt kan man legge til et spenn i nedre og øvre sjikt for å øke variabiliteten i hvilke avvik som kommer. Dette er ikke noe vi har testet i vårt arbeid.

5.2 Stokastisk modellering x antall ganger for kumulativ sannsynlighetsfremstilling

Stokastisk lastmodellering benytter sannsynlighet i sin modellering av vha. stokastisk genererte verdier og vil derfor gi litt forskjellige verdier fra gang til gang (modellering til modellering). I den sammenheng kan sannsynlighetsprinsippet videreføres ved at man gjør flere stokastiske modelleringer for samme periode og ser på dette resultatet i forhold til sannsynlighet. Samtidig observerte vi at makslastverdi med den stokastiske modellen var veldig avhengig av verdiene som forelå i variasjonsverdiene. Dersom en høy positiv avviksandel ble lagt til på en veldig høy variasjonsverdi, kunne dette resultere i en, i noen tilfeller, usannsynlig høy verdi. Derfor foreslår vi, i hovedsak, to metoder som kan benyttes for å gjøre modellen mer representativ for forbruket og avviket:

1. Man kan basere maks-forbruket man forventer ut ifra et «verst-tenkelige» scenario når man gjør en stokastisk modellering f. eks 1000 ganger, og altså bruke maks-verdien man får av alle disse modelleringene.
2. Man kan bruke ulike percentiler man ønsker å være innenfor ift risiko. F. eks. Basert på 1000 modelleringer vil vi med 90% sannsynlighet ha en maks-verdi under x kWh/h.

5.3 Sammenlagring av flere målepunkt

«Sammenlagring» (å sammenlikne flere kunder under samme krets eller område) er en interessant case å se videre på. I et dimensjonerings-scenario, vil man ikke bare vurdere et og et målepunkt, men man ønsker å analysere flere på samme tid. En høy korrelasjon i strømforbruk mellom målepunkt under samme nettstasjon, fører til høyere belastningstopper totalt enn det en lav korrelasjon gjør. På denne måten er det interessant å analysere kundenes korrelasjon ift. sammenlagring som en videreutvikling av modellen.

Hvis man analyserer et sannsynlighetsnivå for makslasten med stokastisk modell, så vil man få ut en estimert verdi for maks-forbruk/effekt. Det vil si at man i utgangspunktet kan bruke samme metodikk videre som med f. eks Velander's formel for sammenlagring. Den «eneste» forskjellen er at maks-effekten man bruker videre i sine dimensjoneringsanalyser er basert på stokastisk lastmodellering, og ikke Velander eller brukstid.

5.4 Større datagrunnlag

Ved å teste et datagrunnlag med flere kunder, vil man kunne vurdere metoden sin tilpasning i enda større grad. Det er vanskelig å si noe om modellens faktiske ytelse og oppfyllelse av oppgaven med å modellere en enkeltkundes antagelige last når datagrunnlaget ikke er så stort. Selv om det gir et visst inntrykk, og i forsøkene vi gjorde traff den stokastiske modelleringsmetoden ofte bedre på makslasten enn det Velander og brukstid gjorde i forsøkene våre (brukstid og Velander hadde lavere verdier enn et som var/ble faktisk makslastverdi hos kunden).

6 Konklusjon

I denne rapporten har vi beskrevet fremgangsmåten for stokastisk lastmodellering slik den er forstått fra doktorgradsavhandlingen til Erling Tønne. Vi har også kommet med forslag til forbedringer og metodikker for bruk av modellen. Vi gjorde forsøk på stokastisk lastmodellering med et datagrunnlag på 37 tilfeldig utvalgte kunder, og våre enkle analyser antyder at trekking fra histogram ga bedre resultat enn fra en sannsynlighetsmodell. Ved å gjøre en periodeinndeling av avviket, og trekke fra disse periodene for de samsvarende periodene virket også å være fordelaktig. Ved å gjøre x antall stokastiske modelleringer av samme periode kan man fremstille makslastverdiene i form av en sannsynlighetsfordeling for hvordan makslast kanskje vil se ut i fremtiden, der man kan velge å legge seg på et visst risikonivå. Man får da en antagelse av makslast for en enkeltkunde. Man kan antagelig deretter benytte tradisjonelle metoder for «sammenlagring» av flere kunder basert på denne makslastverdien. Vi vurderer stokastisk lastmodellering som et interessant spor videre, men mer arbeid med metodikken, analyser og tilpasning av modellen må antagelig gjøres før metoden kan tas i bruk.

7 Referanser

[1] E. Tønne, «Planning of the Future Smart and Active Distribution Grids,» Norwegian University of Science and Technology, Trondheim, 2016.

[2] G. Brännlund, «Evaluation of two peak load forecastingmethods used at Fortum,» KTH Royal Institute of Technology School of Electrical Engineering, Stocholm, Sverige, 2011.

Fullstendig liste over sannsynlighetsmodeller vi har testet i arbeidet:

'alpha', 'anglit', 'arcsine', 'beta', 'betaprime', 'bradford', 'burr', 'cauchy', 'chi', 'chi2', 'cosine', 'dgamma', 'dweibull', 'erlang', 'expon', 'exponweib', 'exponpow', 'f', 'fatiguelife', 'fisk', 'foldcauchy', 'foldnorm', 'frechet_r', 'frechet_l', 'genlogistic', 'genpareto', 'genexpon', 'genextreme', 'gausshyper', 'gamma', 'gengamma', 'genhalflogistic', 'gilbrat', 'gompertz', 'gumbel_r', 'gumbel_l', 'halfcauchy', 'halflogistic', 'halfnorm', 'hypsecant', 'invgamma', 'invgauss', 'invweibull', 'johnsonsb', 'johnsonsu', 'ksone', 'kstwobign', 'laplace', 'logistic', 'loggamma', 'loglaplace', 'lognorm', 'lomax', 'maxwell', 'mielke', 'nakagami', 'ncx2', 'ncf', 'nct', 'norm', 'pareto', 'pearson3', 'powerlaw', 'powerlognorm', 'powernorm', 'rdist', 'reciprocal', 'rayleigh', 'rice', 'recipinvgauss', 'semicircular', 't', 'triang', 'truncexpon', 'truncnorm', 'tukeylambda', 'uniform', 'wald', 'weibull_min', 'weibull_max', 'wrapcauchy' (81 stykker)

Merk: Disse distribusjonene var et utvalg av de totalt 100 tilgjengelige distribusjonene i Python biblioteket scipy.stats basert på en «tidligere versjon» hvor ikke like mange distribusjoner var tilgjengelig. Man kunne gjort et smartere utvalg manuelt, fordi det å regne ut deres tilpasning til datagrunnlaget tar mye prosesseringskraft.

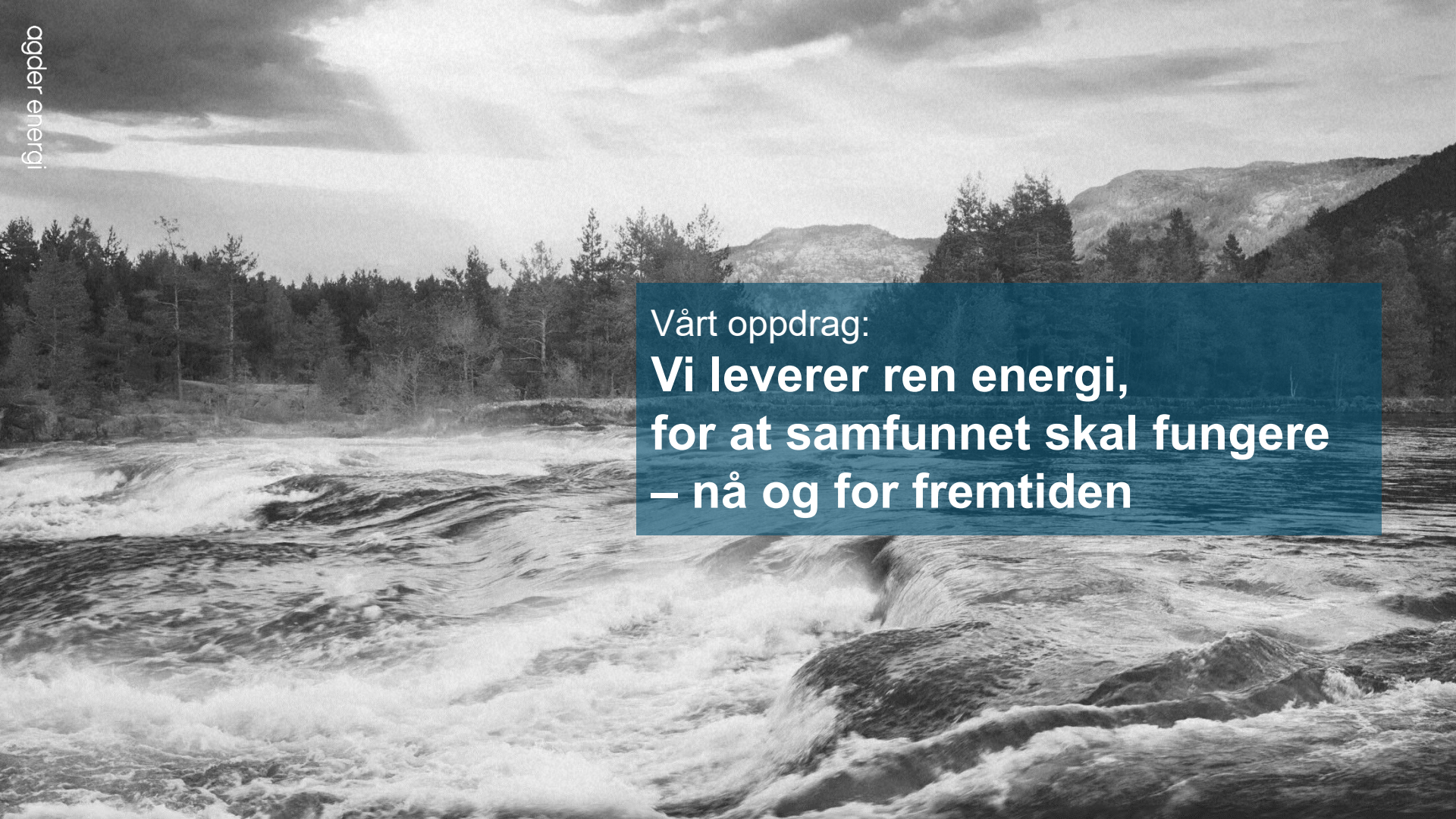
agder energi

Rebekka Olsson Omslandseter

Bruk av en stokastisk modell for modellering av forbruk

Denne presentasjonen er en grundig introduksjon/gjennomgang av dette prosjektet og blir løpende oppdatert med arbeid i relasjon til dette.

agder energi



Vårt oppdrag:

Vi leverer ren energi,
for at samfunnet skal fungere
– nå og for fremtiden

Bakgrunn

Dimensjonering/planlegging av strømnettet – Avhenger av forventet kapasitetsbehov

Forbruket blir stadig mer stokastisk og nye enheter knyttes til nettet

- Strømkrevende enheter skrus av og på hyppig
- Elbil, varmepumper, et økende antall elektriske artikler i husholdninger, etterspørselsordninger etc.
- Økende stabilitetsforventning ift. strømforsyning og spenning

Unngå over-/underdimensjonering – Kostnadseffektivitet

- Ønsker ikke å overdimensjonere – Konsekvens: Høyere kostnad «up-front» for kapasitet som ikke er nødvendig
- Ønsker ikke å underdimensjonere – Konsekvens: Høyere kostnad etter «in-operation» for å øke kapasiteten

Fremtidens strømforsyning vil være mer «aktiv»

→ For å være bedre rustet til å møte disse endringene, kan det tenkes at man kan forbedre metodene man har for beregninger av forventet forbruk og maks-effekt ved å bruke en stokastisk modell

Sensitivity: Internal

Mål

Langtids-planlegging av strømnettet

- Simulere fremtidig strømforbruk og sannsynlig strøm-utvikling for å kunne gjøre gode vurderinger for fremtidig nettstruktur på ethvert spennings-nivå

Korttids-planlegging av strømnettet

- Kan benyttes ift. antatte utfall og overlast

Finne beregnet maks-effekt for å kunne dimensjonere/planlegge for en potensiell «peak» (samt at man kan vite når ev. en slik «peak» er forventet i forhold til denne «enkle» modellen)

Sensitivity: Internal

agder energi

Problemstilling

Tradisjonelt blir deterministiske metoder benyttet

- Velander's formel
- Brukstid
- Standardiserte last-variasjonskurver for ulike «kunde»-kategorier (årlig og daglig variasjon) → Agder Energi benytter ikke denne metoden

Velander og brukstid:

→ Gir oss en verdi for forventet maks-effekt

→ Vi vet ikke når maks-effektuttak er forventet å skje

→ Gir lite innsikt ift. tilgjengelig informasjon

Standard last-variasjonskurver

→ Tar ikke hensyn til forskjellen mellom ukedager og helgedager

→ Ingen info om usikkerhet

→ Ingen håndtering av kunders stokastiske oppførsel

→ Beregnes ut ifra standard-etablerte variasjonskurver, og ikke individuelt

Sensitivity: Internal

Løsning

Doktorgradsavhandling fra 2017 – Erling Tønne

En ny stokastisk metode for å gjøre beregninger på forventet forbruk

Den nye metoden modellerer hver «kunde» individuelt basert på temperatur-korrigerte forbruksverdier med timeoppløsning for de siste 3 til 5 årene

Metoden har ikke noe krav til informasjon om kunders tilkoblede elektriske artikler slik som elbil-lader, varmepumpe etc.

Man får informasjon om forventet forbruk time for time hvor den stokastiske oppførselen er tatt hensyn til

Sensitivity: Internal

agder energi

Overordnet flyt-skjema av metoden

1

Collect hourly measured values for each customers energy consumption [kWh/h]

Temperature-correction:

Adjust and refer every hourly value to the normal temperature for that spesific hour.

$$P_{corr\,i} = P_i + P_i \cdot k \cdot x \cdot (T_i - T_n)$$

Calculate variation curves based on historical data. Find the highest (maximum) value for each time series of measured values

Alt. A: - Yearly variation (per month)
- Daily variations (pr hour)
  - Normal working day
  - Weekend/Holiday

Alt. B: - Daily variations (per hour) for each month
- Normal working day
- Weekend/Holiday

Calculate the expected maximum hourly values by using the variation curves

2

Calculate the relative deviation between the expected maximum hourly values and real measured historical data.

Calculate for several years (e.g. 5 years).

$$\frac{Measured\ value\ -Expected\ maximum\ value}{Expected\ maximum\ value} \cdot 100\%$$

Find the distribution function and its parameters that fits the measured data best.

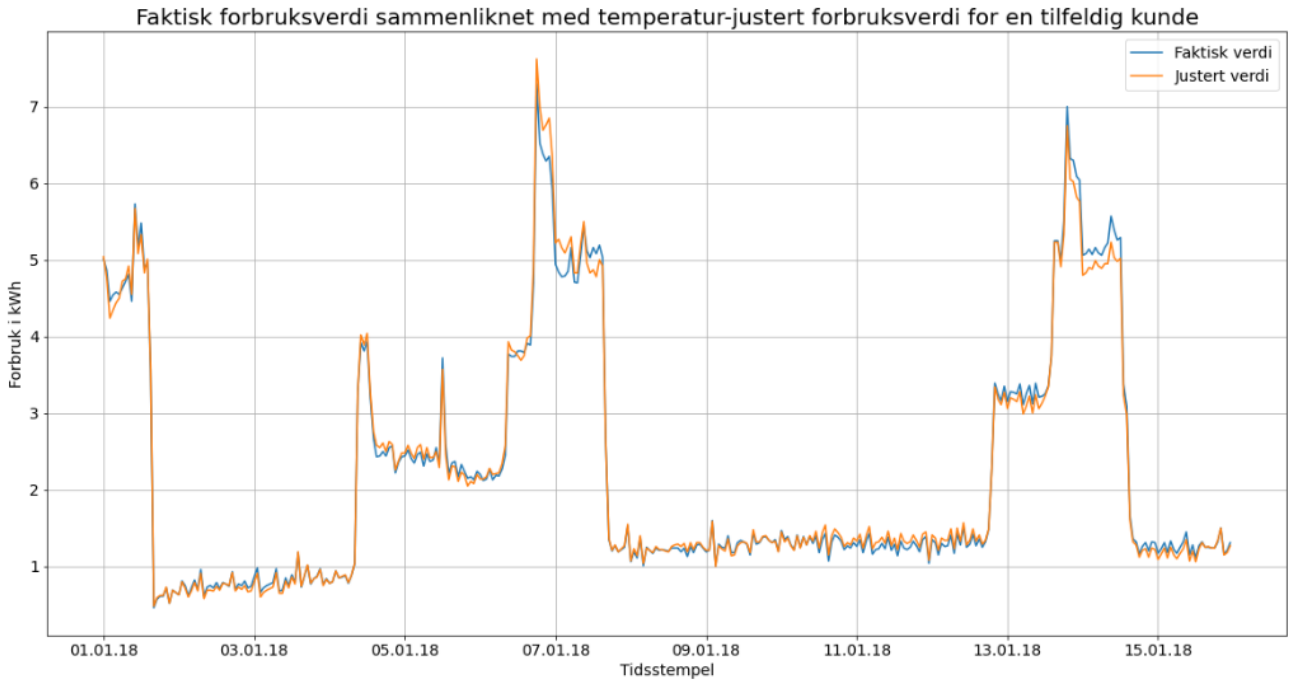
Calculate the expected hourly values :

$$Expected\ value =$$
$$Expected\ maximum\ value\ \cdot\ (1 + Stochastic\ relative\ value)$$

Sensitivity: Internal

Fremgangsmåte: Steg 1

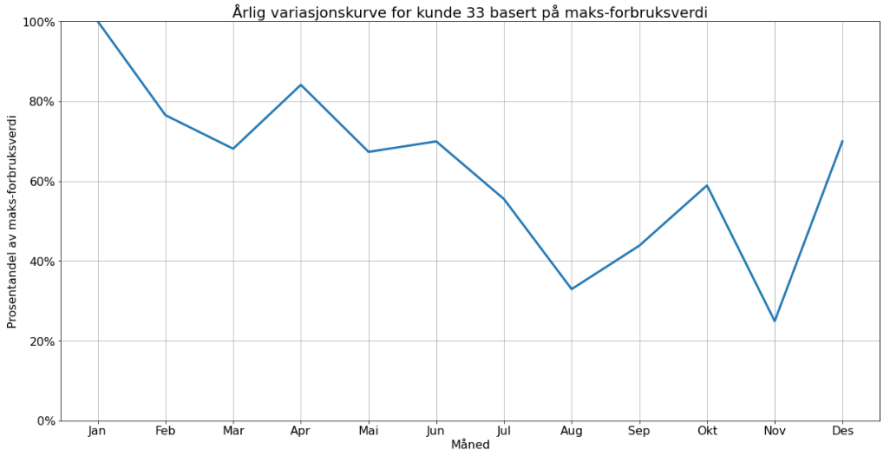
Steg 1: Samlet inn temperaturkorrigert data fra AMS-målere for 37 kunder i effekttariffs-pilot. Disse verdiene hadde timesoppløsning.



Sensitivity: Internal

Fremgangsmåte: Steg 2

Steg 2: Fant variasjonsgrafer og koblet disse sammen for å få beregnede relative maks-verdier med timesoppløsning. Variasjonsgrafene ble beregnet ut ifra høyeste forbruksverdi for gjeldende periode i forhold til høyeste forbruksverdi for hele tidsserien. Vi beregnet en variasjonsgraf for årlig forbruk og to variasjonsgrafer for daglig forbruk – en for ukedager og en for helg/helligdager.

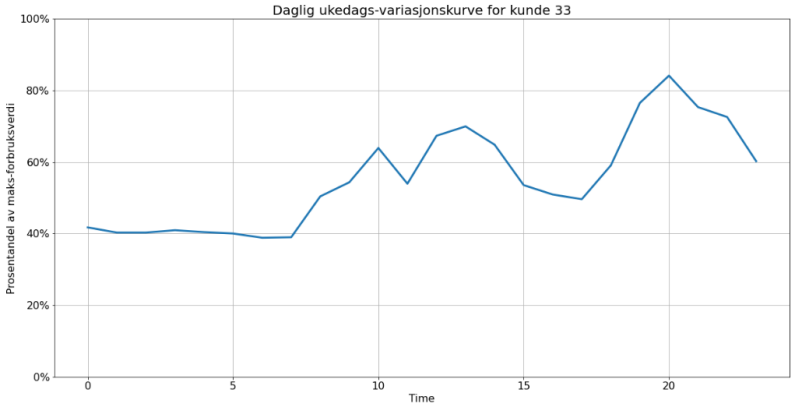


Kurve 1: Årlig variasjonsgraf

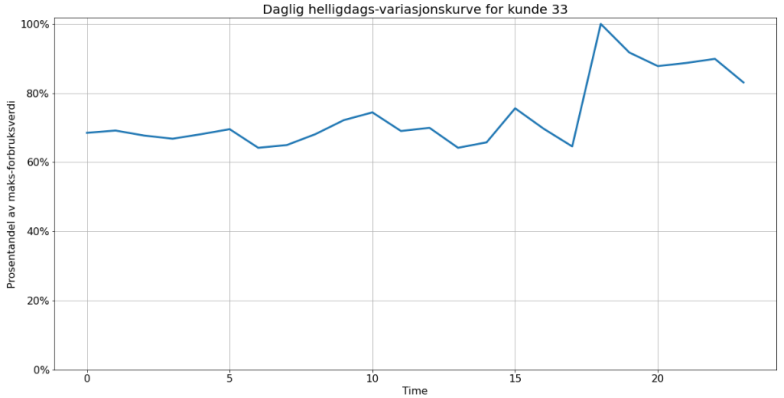
Sensitivity: Internal

agder energi

Fremgangsmåte: Steg 2



Kurve 2: Daglig ukedags-variasjonsgraf



Kurve 3: Daglig helgedags-variasjonsgraf

Sensitivity: Internal

Fremgangsmåte: Steg 2

Vi har en tidsrekke med total lengde N for kunde *i*. Denne tidsrekken består av gjennomsnittlig effekt innafor en time (forbruksverdier med timesoppløsning) angitt med variabelen $w_i(n)$ og hver forbruksverdi har et tilhørende tidsstempel angitt med $t_i(n)$. $n \in [1, 2, \dots, N]$, og angir element nummer n i den sorterte tidsserien, hvor tidsserien er sortert fra tidligste tidsstempel (først) til seneste tidsstempel (sist) og tidsserien består av N verdier totalt. I er totalt antall kunder.

Maksimum forbruksverdi for måned m:

$$\widehat{WM_{i,m}} = \max_{1 \le n \le N} \left(w_i(n)\beta_{w_i(n),m}\right), \text{ hvor } \beta_{w_i(n),m} = \begin{cases} 1 \; hvis \; t_i(n) \; i \; måned \; m \\ 0 \; ellers \end{cases}$$

Maksimum forbruksverdi for time h i en ukedag:

$$\widehat{WU_{i,h}} = \max_{1 \le n \le N} \left(w_i(n)U_{w_i(n),h}\right), \text{ hvor } U_{w_i(n),h} = \begin{cases} 1 \; hvis \; t_i(n) \; i \; time \; h \; for \; en \; ukedag \\ 0 \; ellers \end{cases}$$

Maksimum forbruksverdi for time h i en helg/helligdag:

$$\widehat{WH_{i,h}} = \max_{1 \le n \le N} \left(w_i(n)H_{w_i(n),h}\right), \text{ hvor } H_{w_i(n),h} = \begin{cases} 1 \; hvis \; t_i(n) \; i \; time \; h \; for \; en \; helg/helligdag \\ 0 \; ellers \end{cases}$$

Sensitivity: Internal

Fremgangsmåte: Steg 2

Relativ maks-verdi for årlig forbruk:

$$V_{i,m} = \frac{\widehat{WM_{i,m}}}{\max(w_i(n))}\ , \forall\ m$$

Relativ maks-verdi for daglig forbruk for ukedager:

$$\dot{V_{i,h}} = \frac{\widehat{WU_{i,h}}}{\max(w_i(n))}\ , \forall\ h$$

Relativ maks-verdi for daglig forbruk for helg/helligdager:

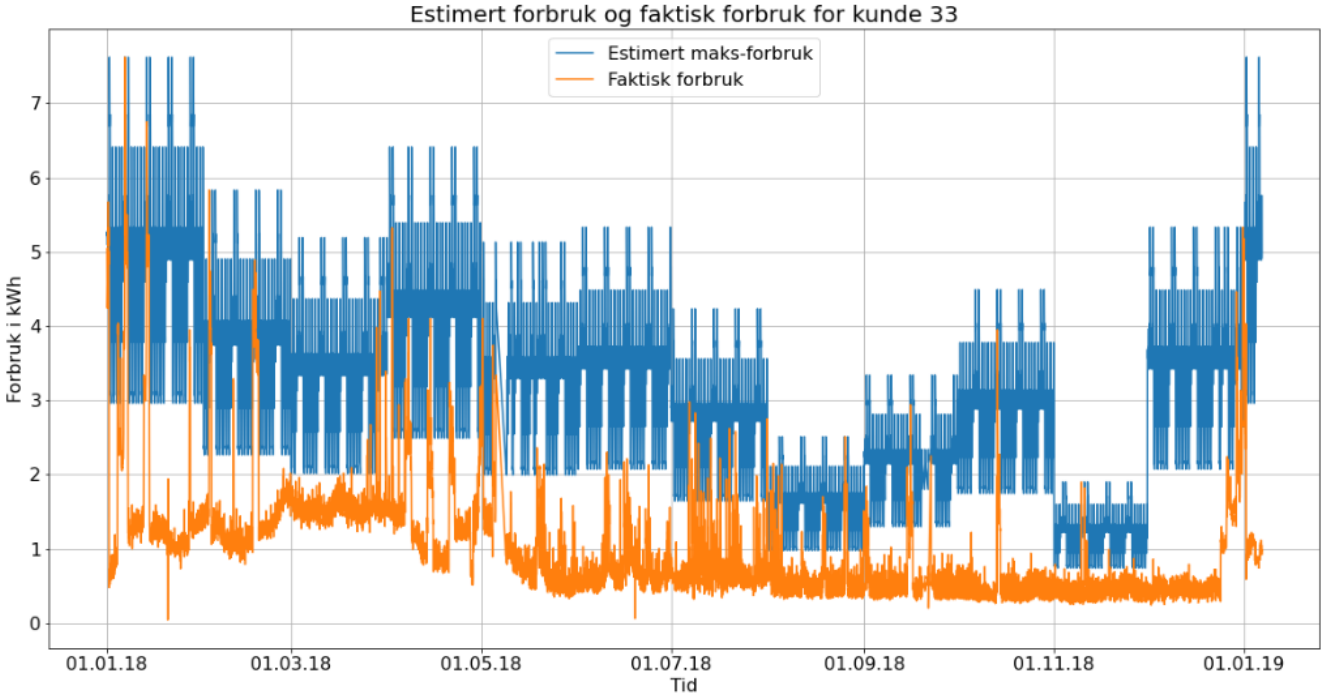
$$\ddot{V_{i,h}} = \frac{\widehat{WH_{i,h}}}{\max(w_i(n))}\ , \forall\ h$$

Sensitivity: Internal

Fremgangsmåte: Steg 2

Deretter beregnes maks-effekt-verdiene etter følgende formel:

$$v_i(n) = \left(\max_{l \in [1,N]}(w_i(l)\right) V_{i,m}\beta_{w_i(n),m}(\dot{V_{i,h}}U_{w_i(n),h} + \ddot{V_{i,h}}H_{w_i(n),h}) \; \forall \; n$$

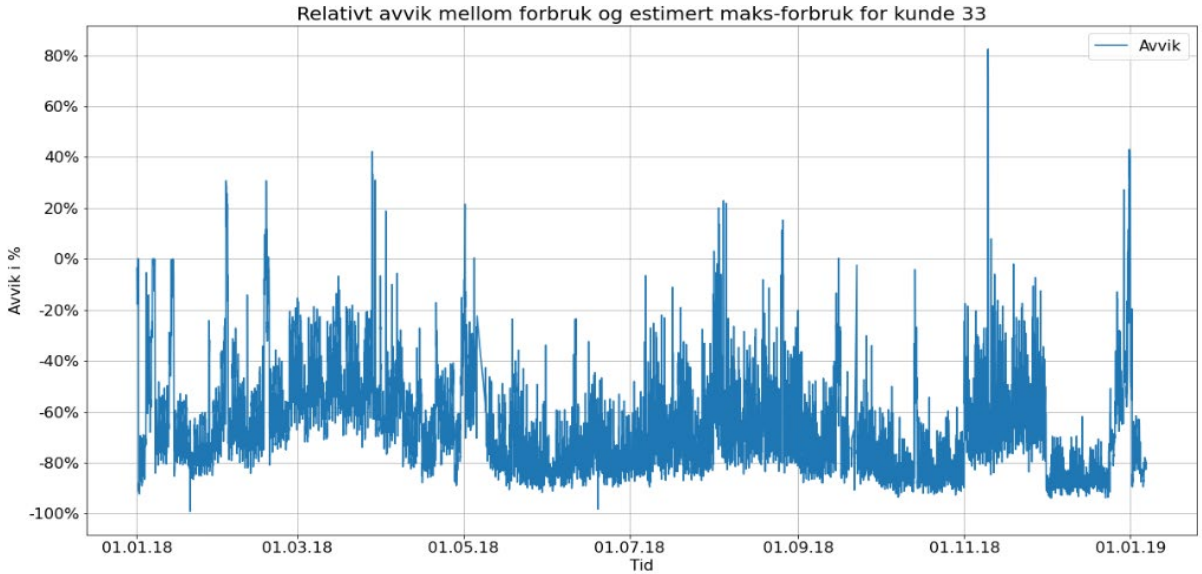


Sensitivity: Internal

Fremgangsmåte: Steg 3

Steg 3: Beregnet avviket mellom beregnede relative maks-verdier og faktisk temperaturkorrigert forbruk. Deretter fant vi den mest passende stokastiske sannsynlighetsmodellen for å representere dette avviket.

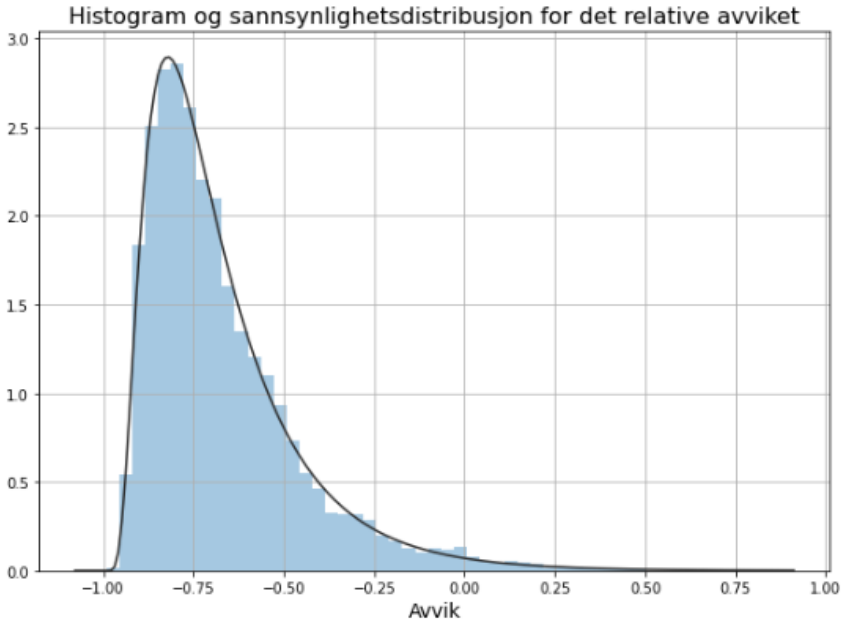
$$\phi_i(n) = \frac{w_i(n) - v_i(n)}{v_i(n)}$$



Sensitivity: Internal

Fremgangsmåte: Steg 3

For å finne mest passende sannsynlighetsmodell for avviket benyttet vi Chi-Square som mål – mer om dette på neste side



Sensitivity: Internal

Fremgangsmåte: Steg 3

NB: Notasjonen på denne siden er uavhengig, og «expected value» er ikke relatert til vår modell/verdier og blir hentet direkte fra sannsynlighetsmodellen for avviket

Chi-Square Formula

This is the formula for Chi-Square:

$$X^2 = \sum \frac{(O - E)^2}{E}$$

- Σ means to sum up (see Sigma Notation)
- O = each **Observed** (actual) value
- E = each **Expected** value

So we calculate $\frac{(O-E)^2}{E}$ for each pair of observed and expected values then sum them all up.

Dette gjøres for intervallene i histogrammet, og ikke for hver enkelt verdi, men prinsippet er det samme

Sensitivity: Internal

Fremgangsmåte: Steg 4

Steg 4: Beregnet de endelige modellerte forbruksverdiene, som var et resultat av relativt maks-forbruk tillagt en faktor av usikkerhet ved å trekke en avviksprosentandel fra sannsynlighetsmodellen per time og justere de endelige forbruksverdiene etter denne

De stokastiske verdiene genereres på følgende måte:

$$s_i(n) = v_i(n) * (1 + \theta) \; \forall \; n$$

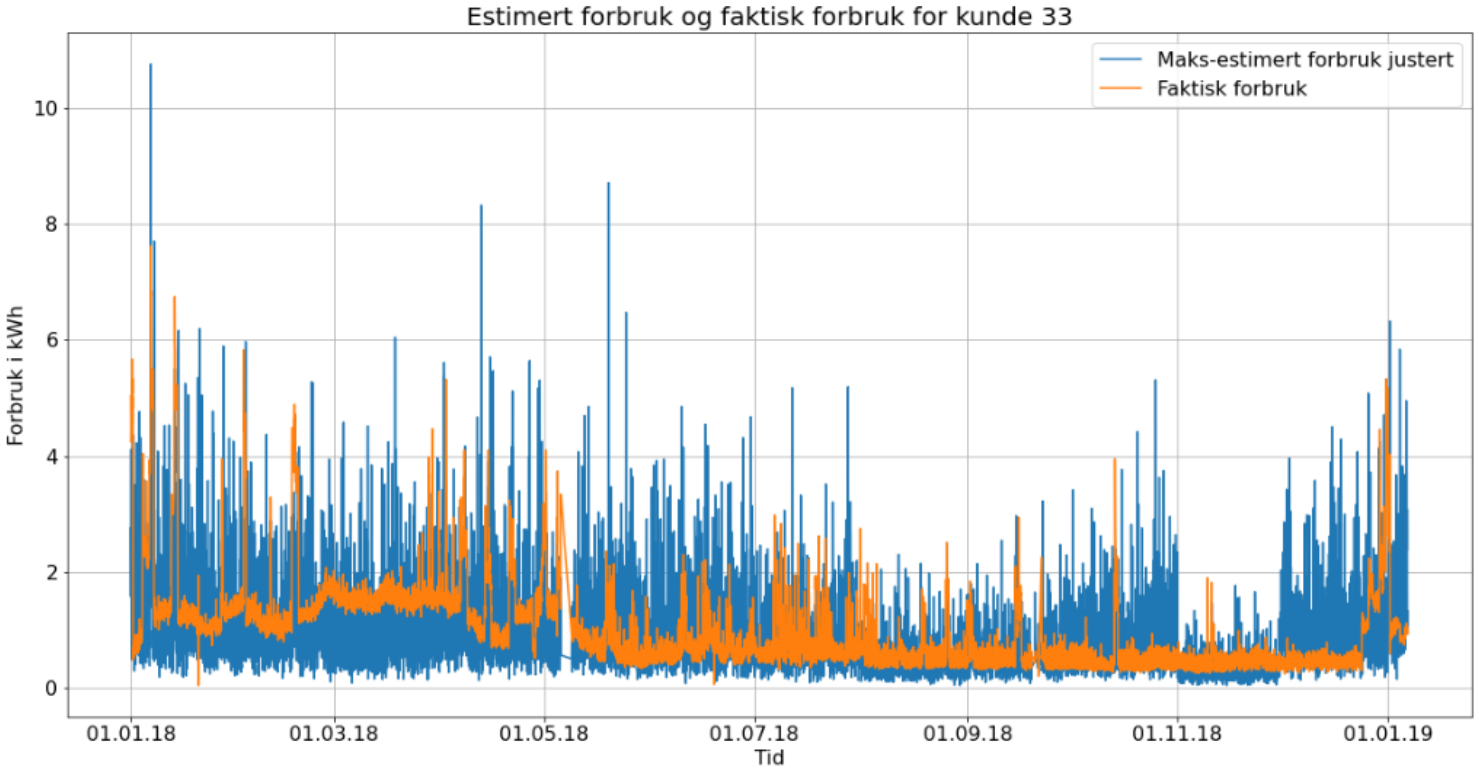
Hvor θ er en stokastisk variabel med distribusjonsfordeling i forhold til hvilken distribusjon som ble funnet i forrige steg. Det vil si at denne verdien kan variere ikke bare mellom -1 og 1, men også verdier over og under disse verdiene.

F. eks, dersom man har en normaldistribusjon som beste sannsynlighetsdistribusjon for avviket, trekker man en verdi fra denne distribusjonen med de spesifikke parameter som passer til dataen, og justerer den modellerte verdien for timen basert på denne. Deretter gjør man det samme for alle timene som man skal modellere.

Sensitivity: Internal

agder energi

Fremgangsmåte: Steg 4



Sensitivity: Internal

Fremgangsmåte: Steg 4

Evaluering av stokastisk modell:

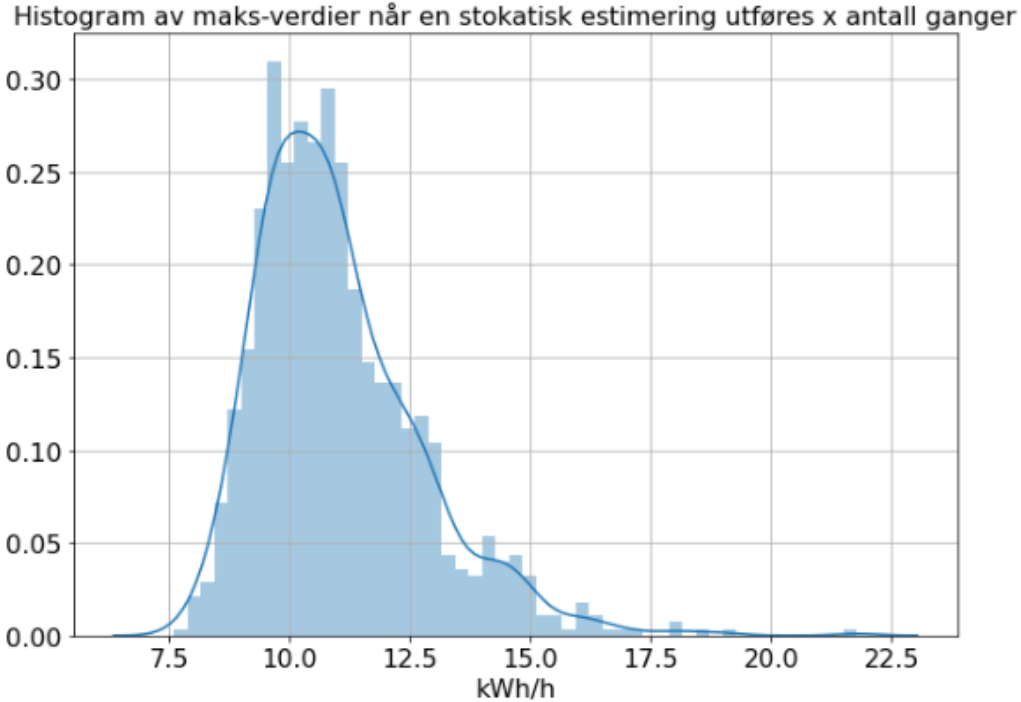
Vi ser på de stokastiske verdiene i forhold til de målte verdiene, men på en slik måte at de justeres i forhold til de verdiene vi hadde før vi avviks-justerte de relative maks-verdiene. På denne måten kan vi sammenlikne avviket mellom de ulike kundene med hverandre selv om det er en ulik skala mellom dem.

$$\varphi_i(n) = \frac{s_i(n) - w_i(n)}{v_i(n)}$$

Sensitivity: Internal

agder energi

Histogram for maksverdien for 1000 eksperimenter med stokastisk modell



→ Vi ser at maks-verdiene som modelleres er ganske varierende. Dette er fordi de er veldig avhengig av dataunderlaget. Dersom et høyt avvik legges på som økning til en allerede høy verdi, får vi høyere maks-verdier enn dersom det motsatte skjer

Sensitivity: Internal

Videreutvikling av stokastisk modell

Beregnet maks-verdi med den stokastiske modellen er veldig avhengig av verdiene som foreligger i dataen fra variasjonsgrafen. Dersom en høy avviksandel legges til på en veldig høy, men det er ikke nødvendigvis alltid tilfellet.

Derfor foreslås det, i hovedsak, to metoder man kan benytte for å gjøre modellen mer representativ for forbruket og avviket frem i tid:

1. Vi kan basere maks-forbruket vi forventer ut ifra «verst-tenkelige» scenario når man gjør en generering av f. eks 1000 stokastisk modellerte tidsrekker, og deretter bruke maks som kommer frem av denne genereringen. Dersom vi ønsker motsatt resultat kan vi gå ut ifra minimums-verdien.
2. Vi kan bruke ulike percentiler vi ønsker å være innenfor. F. eks. Basert på 1000 genereringer, vil vi sannsynligvis ha en maks-verdi under x kWh/h med 90% sannsynlighet.

Vi presenterer samtidig et annet mål for å kunne evaluere avviket til den stokastiske modellen. Fordi man i dimensjonering som regel er interessert i maks-effekt, kan man evaluere denne verdien på en mer spesifikk måte ved å se på estimert maks-effekt (enkeltverdi), relativt til observert maks-effekt.

For den stokastiske modellen ser dette vurderingskriteriet slik ut:

$$\xi_{s_i} = \frac{\max\big(s_i(n)\big) - \max(w_i(n))}{\max(w_i(n))}$$

Dersom man ønsker å vurdere gjennomsnittlig for flere kunder, kan man benytte formelen:

$$\bar{\xi_s} = \frac{1}{I}\sum_{i=1}^{I} \xi_{s_i}$$

Sensitivity: Internal

Andre metoder: Velander

$$A_y = k_1 W_{i,y} + k_2 W_{i,y}$$

Hvor $W_{i,y}$ er totalt årsforbruk for kunde i for år y. Denne verdien kan beregnes ut ifra følgende formel:

Summert forbruk for år y:

$$W_{i,y} = \sum_{n=1}^{N} w_i(n)\, \alpha_{w_i(n),y}$$

$$\text{hvor } \alpha_{w_i(n),y} = \begin{cases} 1 \; hvis \; t_i(n) \; i \; år \; y \\ 0 \; ellers \end{cases}$$

Sensitivity: Internal

Andre metoder: Velander

Dersom man ønsker å beregne Velander's verdi for påfølgende år, kan man følge følgende formel:

$$\grave{A}_y = k_1(1+\varepsilon)W_{i,y-1} + k_2(1+\varepsilon)W_{i,y-1}$$

Dersom man vil beregne Velander's verdi for tidligere år, kan man beregne dette ved hjelp av:

$$\acute{A}_y = k_1(1-\varepsilon)W_{i,y+1} + k_2(1-\varepsilon)W_{i,y+1}$$

Hvor ε, angir forventet årlig prosentandels økning i totalt årsforbruk. Denne verdien kan justeres i forhold til erfaring, eller innsamlede målinger over flere år for en enkelt kunde.

Sensitivity: Internal

Andre metoder: Brukstid

For brukstid har vi benyttet følgende formel:

$$B_y = \frac{W_{i,y}}{T_u}$$

Hvor T_u er antatt brukstid for kundetypen som kunde i tilhører. Verdiene for brukstid er standardisert, og følger gitte tabeller som kan være ulikt fra nettselskap til nettselskap. Det er også viktig å merke seg at brukstid ofte ikke benyttes for husholdningskunder, men at det gjerne benyttes til beregninger for hytter for eksempel.

For å beregne brukstid et år frem i tid, har vi formelen:

$$\grave{B}_y = (1 + \varepsilon)\frac{W_{i,y-1}}{T_u}$$

På samme måte kan vi beregne et år tidligere:

$$\acute{B}_y = (1 - \varepsilon)\frac{W_{i,y+1}}{T_u}$$

Sensitivity: Internal

Andre metoder: Brukstid

I doktorgradsavhandlingen til Erling Tønne er følgende verdier for brukstid listet opp:

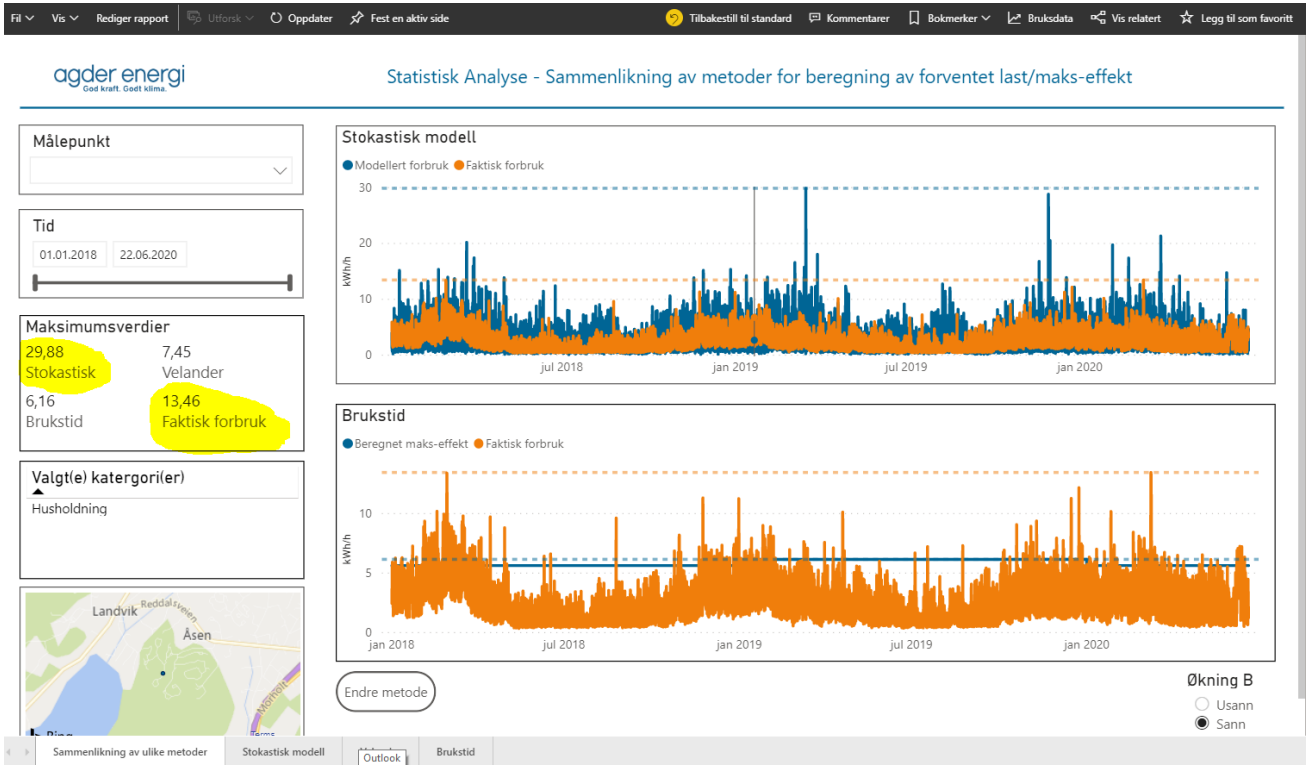
Kundekategori	Brukstid (T_u)
Husholdning	3600
Skole	2500
Helse og omsorgsboliger	3800
Kontor	3800
Varehandel	4100
Gårdsbruk	3000

I tillegg benytter Agder Energi ofte 1/4 eller 1/8 av antall timer i et år som mål på brukstid for hytter. Det varierer hvilken av parameterne som benyttes, og dette vurderes gjerne ut ifra hvordan forbruksmønsteret til kunden tidligere har fordelt seg (som vurderes manuelt).

Sensitivity: Internal

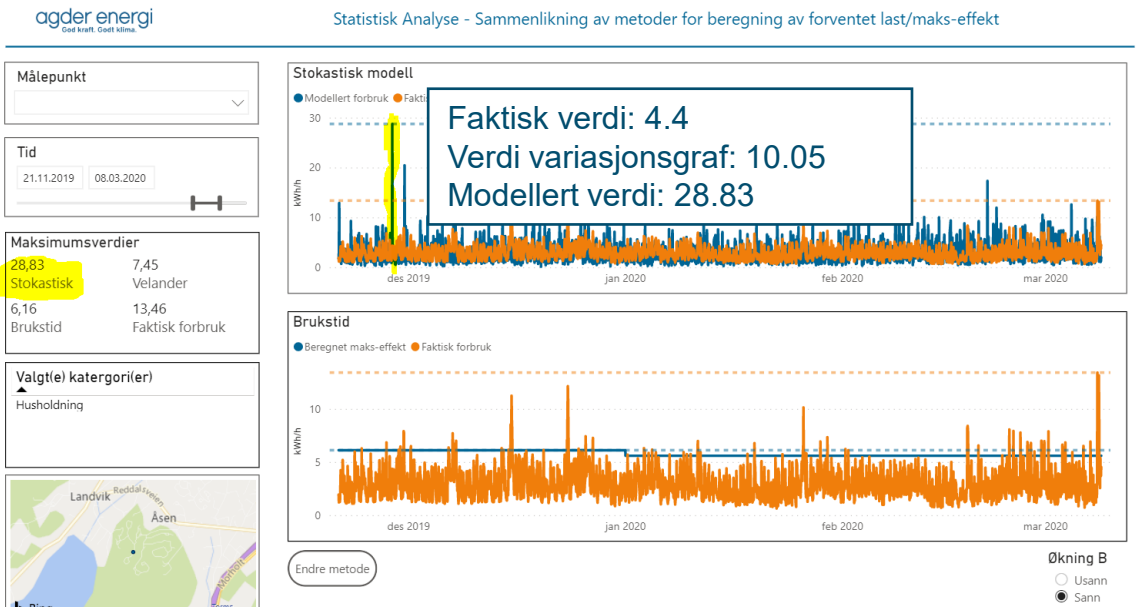
agder energi

Betraktning 1: Hvordan kan noen modellerte verdier bli mer enn dobbelt så høye som maks-verdien av faktisk forbruk?



Sensitivity: Internal

Eksempel ang. Betraktning 1:



Formel for å finne stokastisk modellert verdi:

$$s_i(n) = v_i(n) * (1 + \vartheta) \; \forall \; n$$

$$28.83 = 10.05 \; * (1 + \vartheta)$$

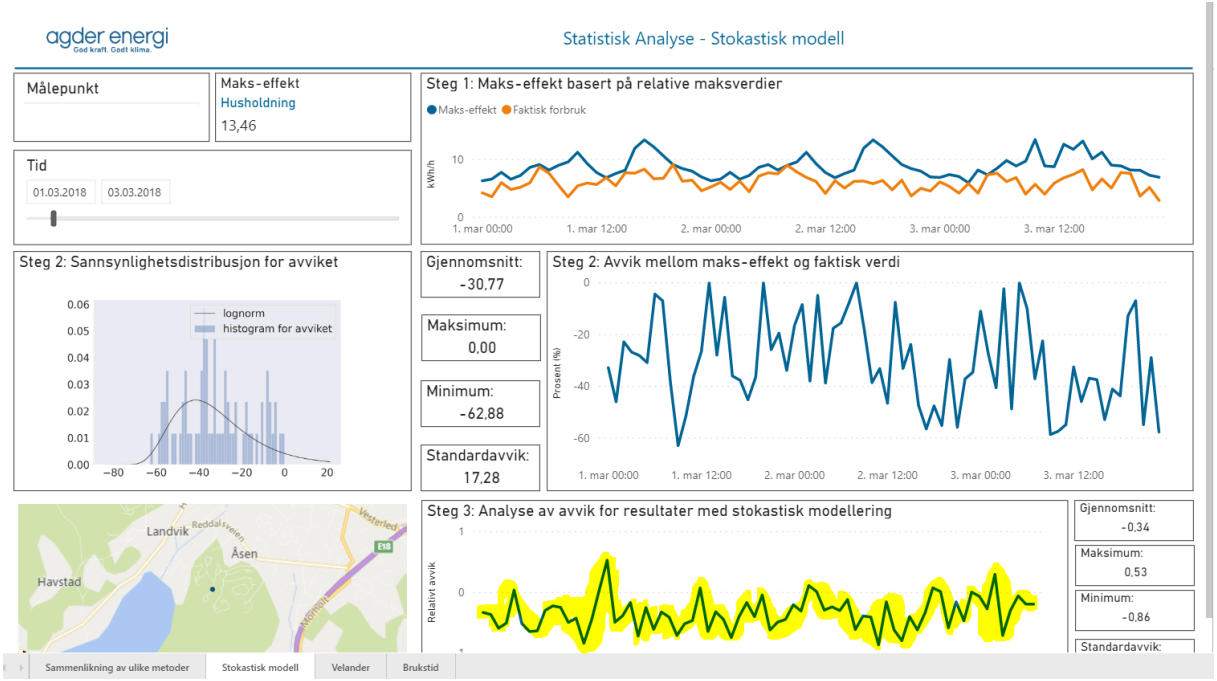
$$\vartheta = \frac{28.83}{10.05} - 1 = 1.87$$

Som kan skje, fordi man trekker verdier fra en sannsynlighetsdistribusjon.

Sensitivity: Internal

agder energi

Betraktning 2: Vi ser at de modellerte verdiene som regel har en maksverdi over det faktiske forbruket. Men i forhold til forventa verdi, så vil denne likne faktisk forbruk. Evalueringsmetoden i doktorgradsavhandlingen kan være vanskelig å analysere. Vi burde derfor finne en ny måte å analysere dette på.



Sensitivity: Internal

Løsning ang. Betraktning 2:

MAPE er et mål på performance når man skal sammenlikne forventede verdier med faktiske verdier (prediksjonsnøyaktighet)

Dette målet kan også benyttes i denne sammenhengen (selv om vi ikke har "predikerte" verdier frem i tid).

$$MAPE = \frac{1}{N}\sum_{n=1}^{N}\left|\frac{w_i(n) - s_i(n)}{w_i(n)}\right|$$

Sensitivity: Internal

Løsning ang. Betraktning 2:

Vi kan også vurdere et enda enklere mål på nøyaktighet med de modellerte verdiene sammenliknet med faktisk forbruk:

$$\gamma_i(n) = s_i(n) - w_i(n)$$

$$\mu_i = \frac{\sum_{n=1}^{N} s_i(n) - w_i(n)}{N}$$

Eller slik antydet i tidligere slide med sammenlikning av maksverdi

Sensitivity: Internal

Betraktning 3: Å finne best passende sannsynlighetsmodell for avviket er tidkrevende, og likevel er det ikke nødvendigvis en god tilpasning

AIC, BIC, HQC til å vurdere tilpasning (likevel lang prosesseringstid for flere av distribusjonene)

De fleste sannsynlighetsmodeller er definert mellom -∞ til +∞

-> Resulterer i at man kan få "umulige" høye/lave verdier

Man har noen typer som er "bounded", f. eks truncnorm – god tilpasning?

Man kan definere en absolutt maks manuelt

Hvor skal man "klippe" distribusjonene? Hva burde være kriteriene?

Løsning: Trekking av avviksverdier fra histogram direkte

Avviket kan da sees på periodisk, slik at man ikke får peaks (store avvik sommerstid) når det ikke er veldig reelt

- Ingen årlig inndeling
- 2 årlige inndelinger (vinter og sommer)
- 4 årlige inndelinger (jan-mar, apr-jun, etc)
- Månedlige inndelinger

Sensitivity: Internal

agder energi

Oppsummering av arbeid – i korte trekk

Brukstid, velander og stokastisk modellering for 37 kunder

Vurdert sannsynlighetsmodeller med Chi-square og AIC, BIC, HQC og i kombinasjon for å få ned prosesseringstid

Testet gjennomsnitt-metode og maks-metode

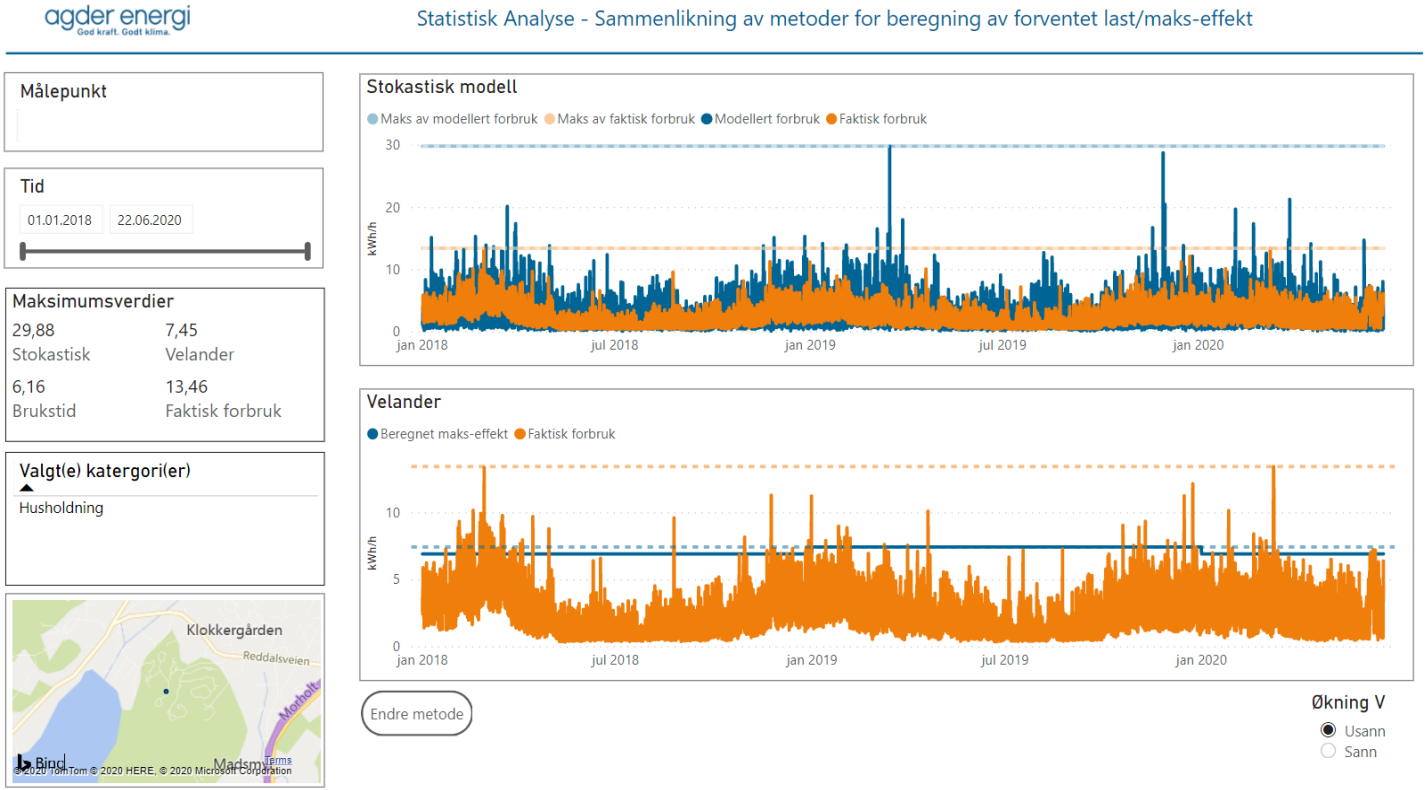
Testet trekking fra histogram i stedet for sannsynlighetsmodell med forskjellige periode-inndelinger

Gjort mange stokastiske modelleringer av samme periode for å analysere hvordan maks-verdiene fordeler seg ift. sannsynlighet og risikoanalyse

Sensitivity: Internal

agder energi

Stokastisk lastmodellering versus andre metoder – Den stokastiske modellen i denne rapporten er ikke med beste tilpassede modell, men beste av 10 tilfeldige av rundt 80



Sensitivity: Internal

AIC, BIC og HQC for distribusjonsvurdering

agder energi

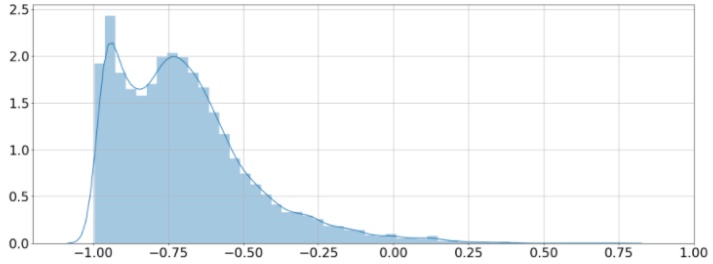
agder energi
God kraft. Godt klima.

Statistisk Analyse - AIC BIC HQC Distribusjonsanalyse

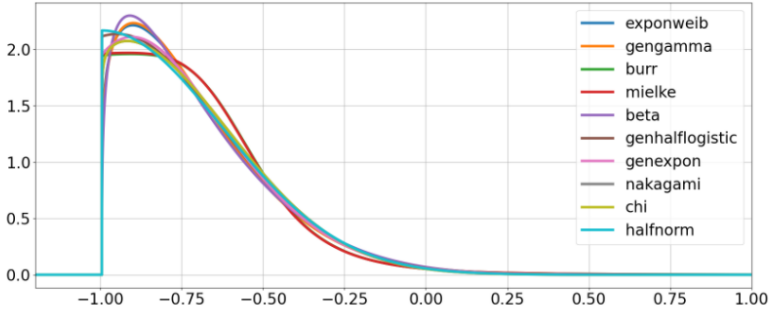
Målepunkt og distribusjon

Dist	UsagePointId	AIC	BIC	HQC
cosine	707057500077881602	NaN	NaN	N
exponweib	707057500077881602	-12 637,52	-12 605,37	-12 627
gengamma	707057500077881602	-12 626,30	-12 594,15	-12 615
burr	707057500077881602	-12 612,75	-12 580,60	-12 602
mielke	707057500077881602	-12 608,80	-12 576,65	-12 598
beta	707057500077881602	-12 604,27	-12 572,12	-12 593
genhalflogistic	707057500077881602	-12 590,28	-12 566,17	-12 582
genexpon	707057500077881602	-12 574,86	-12 534,67	-12 561
nakagami	707057500077881602	-12 557,83	-12 533,72	-12 549
chi	707057500077881602	-12 557,83	-12 533,72	-12 549
halfnorm	707057500077881602	-12 524,60	-12 508,53	-12 519
foldnorm	707057500077881602	-12 522,60	-12 498,49	-12 514
frechet_r	707057500077881602	-12 521,16	-12 497,05	-12 513
weibull_min	707057500077881602	-12 521,16	-12 497,05	-12 513
ncx2	707057500077881602	-12 414,43	-12 382,29	-12 403
gompertz	707057500077881602	-12 350,15	-12 326,04	-12 342
halflogistic	707057500077881602	-12 232,58	-12 216,51	-12 227
erlang	707057500077881602	-12 194,85	-12 170,74	-12 187
gamma	707057500077881602	-12 194,85	-12 170,74	-12 187
pearson3	707057500077881602	-12 194,85	-12 170,74	-12 187
exponpow	707057500077881602	-12 052,27	-12 028,15	-12 044
genpareto	707057500077881602	-11 980,01	-11 955,89	-11 972
johnsonsb	707057500077881602	-11 842,20	-11 810,05	-11 831

Histogram for avviket mellom variasjonsverdier og faktisk forbruk



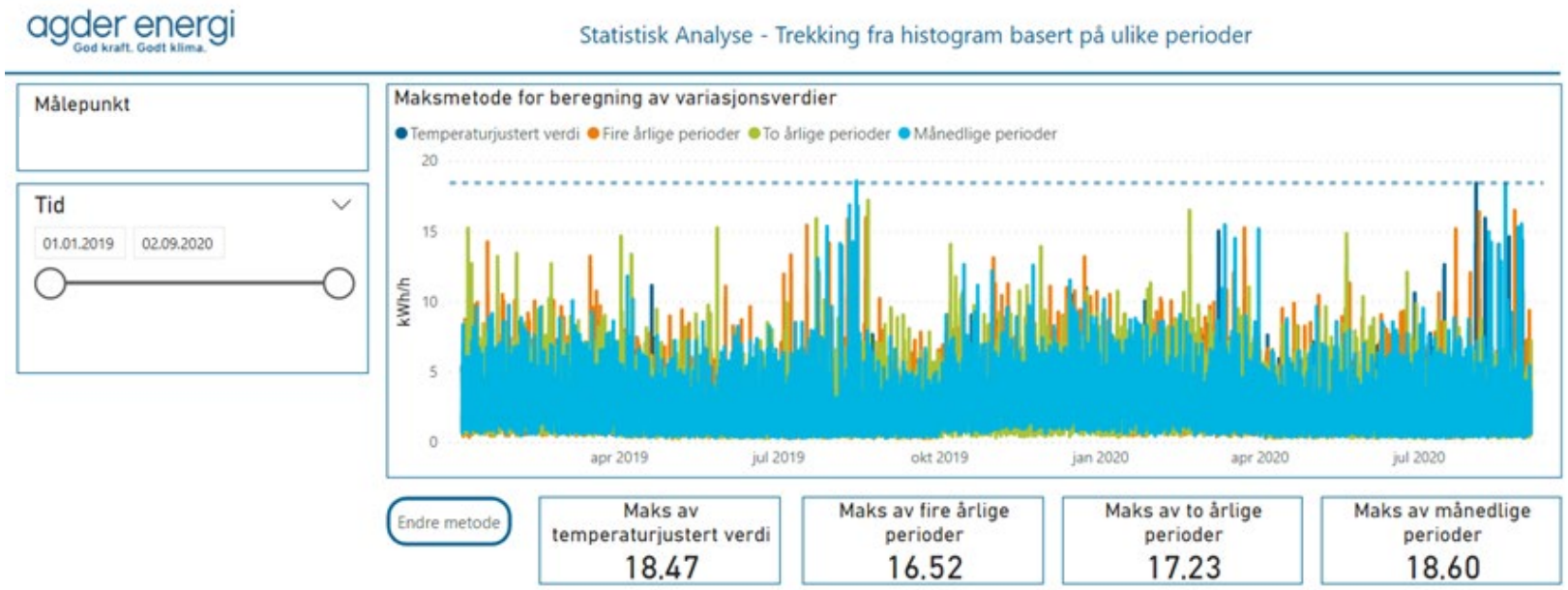
Distribusjon tilpasset avviket



Sensitivity: Internal

agder energi

Stokastisk modellering med periodebaserte histogrammer



Sensitivity: Internal

agder energi

Kumulativ fremstilling av maks ved gjentatte modelleringer

agder energi
God kraft. Godt klima.

Statistisk Analyse - Kumulativ fremstilling av trekking x antall ganger

Målepunkt

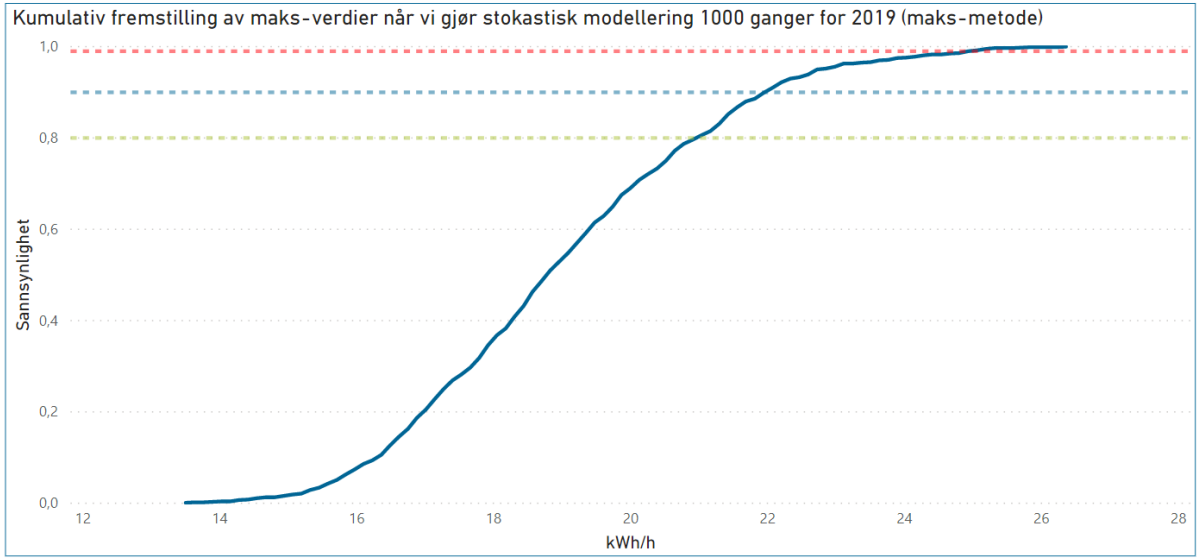
Info

Variasjonsverdiene er beregnet med **maks-metode**

Trekkingen av justeringene er gjort fra et histogram av avvik mellom variasjonsverdier og temperaturjustert faktisk forbruk

Histogrammet er basert på to årlige perioder

Periodene er oktober-mars (vinter) og april-september (sommer)



Endre metode

Sensitivity: Internal


agder energi
God kraft. Godt klima.

Prosjektnotat

TITTEL			
Pilot Effektanalyse: Analyse av effektvariasjon innenfor en time			
WORK PACKAGE	**VERSJON**	**DATO**	**ANTALL SIDER**
WP Pilot	1.0	2021-08-04	33
FORFATTER(E)		**WP-LEDER**	**GRADERING**
Aksel Holbek Sørbye og Pål Wagner Per-Oddvar Osland (Oct 8, 2024 13:28 GMT+2)		Maren Istad Maren Istad (Oct 9, 2024 07:36 GMT+2)	Åpen

DISTRIBUSJON		
CINELDI		

SAMMENDRAG

Analyse av CCD-data og TIBBER-data, bestående av henholdsvis 5-minuttsverdier og 10-sekundsverdier, har ført til både verifisering og falsifisering av hypoteser. Ved å se på fordelinger av forholdsverdier og tilhørende 99. persentiler, ble det dannet et godt grunnlag for å avkrefte hypotesen som sa at husholdnings- og fritidskundene har relativt lik effektvariasjon innenfor en time. Ved bruk av samme metodene, har hypotesen som sa at husholdnings- og næringskundene har relativt lik effektvariasjon innenfor en time, blitt bekreftet. Videre ble det brukt spredningsplott for å plotte 99. persentiler av forholdsverdier, og tilhørende 90. persentiler for å sammenligne effektvariasjon i høylast-, lavlast-, og normallasttimene. Resultatene av denne analysen ga oss rimelig grunnlag for å beholde hypotesen som sa at variasjon i disse ulike forbrukskategoriene er ulik. Til slutt ble det delt inn 20 lastkategorier for Tibberdataen og 99. persentiler av forholdsverdier ble generert for hver av disse for å i detalj analysere hvordan forholdsverdier påvirkes av lastnivåer.

Notasjon

Analyserte effekter

P = Effekt

P_{1h} = Gjennomsnittlig effektforburk i løp av en time, også kalt timesverdi

P_{5m} = Gjennomsnittlig effektforburk i løp av 5 minutter, også kalt en 5-minuttsverdi

P_{10s} = En verdi av gjennomsnittlig effektforburk i løp av 10 sekunder, også kalt en 10-sekundsverdi

Forholdsverdi

$F_{5m} = P_{5m}/P_{1h}$ = Forholdet mellom en 5-minuttsverdi og det gjennomsnittlige effektforbruket i dens tilhørende time

$F_{10s} = P_{10s}/P_{1h}$ = Forholdet mellom en 10-sekundsverdi og dens tilhørende timesverdi

Persentiler

PC_{99} = 99. persentil av en gruppe observasjoner = Verdien som er større enn 99% av observasjonene i gruppen

Matematisk kan persentiler defineres som følger:

Gitt en liste L med N sorterte verdier så vil den k-te persentilen (0<k<=100) være gitt ved:

$$PC_k = L_n$$

Der: $n = \lceil \mathrm{k} \cdot \mathrm{N} \rceil$

$PC_{99,1h}$ = 99.persentil av en gruppe timesverdier

$PC_{99,5m}$ = 99.persentil av en gruppe 5-minuttsverdier

$PC_{99,10s}$ = 99.persentil av en gruppe 10-sekundsverdier

$PC_{99}F_{10s}$ = 99. persentil av en gruppe 10-sekundersforholdsverdier

$PC_{99}F_{5m}$ = 99. persentil av en gruppe 5-minuttsforholdsverdier

Høylast

$PC_{99,10s,Høylast}$ = $PC_{99,1h}$ for de 10% høyeste effektmåleverdiene

$PC_{99}F_{10s,Høylast}$ = $PC_{99}F_{10s}$ for de 10% høyeste effektmåleverdiene

Dataframes

DF_{5m} = Dataframe med 5-minuttsdata.

DF_{1h} = Dataframe med timesdata.

DF_{10s} = Dataframe med 10-sekundsdata.

$DF_{5m\ \&\ 1h}$ = Dataframe med 5-minuttsdata og timesdata.

$DF_{10s\ \&\ 1h}$ = Dataframe med 10-sekundsdata og timesdata.

Ordliste

Frekvens: Tidsoppløsninger på måleverdier i datasettene som analyseres

Høyfrekvent: Original frekvens i datasettene som analyseres

Lavfrekvent: Frekvensen som høyfrekvent data aggregeres til. Synonym til timesoppløsing i denne rapporten

Normallast: Alle effektmålinger i et datasett

Høylast: De 10% største effektmålingene i et datasett.

H_0 = Nullhypotesen. Hypotesen vi starter med

H_1 = Alternativ hypotese. Hypotesen som er det motsatte av H_0. Brukes dersom H_0 må forkastes.

Q_{25} = Median i et intervall [laveste verdi, median i datasett].

Q_{75} = Median i et intervall [median i datasett, høyeste verdi].

IQR = Interquantile range. $Q_{75} - Q_{25}$.

Whisker = Øvre grense. Øverste whisker representerer Q_{75} + 1.5IQR. Nederste whisker representerer Q_{25} – 1.5IQR

Innholdsfortegnelse

1 Problembakgrunn

«Vanlige» strømmålinger hos de fleste nettselskaper viser gjennomsnittlig timesforbruk. Innad i disse timene kan det forekomme høye effekttopper som ikke kommer fram ved et timesgjennomsnitt dersom det også forekommer perioder med lavt forbruk. Strømnettet må dimensjoneres til å dekke de fleste effekttopper som kan forekomme. Derfor er det interessant å analysere effektvariasjonene innad i timer. Analysene som har blitt gjort og presentert i denne rapporten baserer seg på 5-minutts- og 10-sekkundsmålinger. Timesdata ble generert ved å aggregere opp det høyoppløselige datagrunnlaget.

Datagrunnlaget med 5-minuttsmålinger inneholdt mange flere kunder og kundekategorier enn datagrunnlaget med 10-sekkundsmålinger. Dette var grunnen for at dette datagrunnlaget ble brukt for å sammenligne effektvariasjon innen en time for de ulike kundekategoriene. Her ble forholdsverdier av høyoppløselig- og lavoppløseligdata plottet i histogrammer og persentiler ble brukt for å sammenligne effektvariasjonen for ulike kundekategoriene. Videre ble det også sett på effektvariasjon i høylast-, lavlast-, og normallasttimene for de ulike kundekategoriene, og variansen og standardavviket i disse ble sammenlignet, slik at tydelige forskjeller mellom kundekategoriene kunne ses. Dersom effektvariasjonen for de ulike kundekategoriene er relativt lik, kan nettet som forsyner disse kundekategoriene dimensjoneres med like dimensjoneringskriterier, men dersom det er stor variasjon for hver kundekategori, må nettet for de ulike kundekategorier dimensjoneres med ulike kriterier.

Datagrunnlaget med 10- sekundsmålinger, inneholdt mye mer nøyaktige målinger enn datagrunnlaget med 5-minuttsmålinger. Ulempen med dette grunnlaget var at det inneholdt få kunder, og bare en kundekategori. Derfor ble det, i tillegg til plotting av forholdsverdier i histogrammer, utført analyser av hvordan forholdspersentilene fordelte seg blant ulike kunder og ved ulike lastnivåer.

Til slutt sammenlignes 10-sekkundsdata med 5-minuttsdata for å se om det finnes en tydelig forskjell mellom disse. Dersom det ikke gjør det, vil det ikke være nødvendig å beholde 10-sekkundsmålinger, ettersom det krever flere ressurser for å få tilgang til data med så høy oppløsning. Dersom det skulle vise seg at det det er en forskjell mellom datagrunnlagene, bør det utføres kost-nytte-analyse, for å avgjøre hvor viktig det er å beholde 10-sekundsdata..

1.1 Forskningsspørsmål og hypoteser

1.1.1 Hva er forholdet mellom høyfrekvent data og timesoppløst data?

Dette forskningsspørsmålet besvares ved å beregne og plotte forholdsverdiene F_{5m} og F_{10s}, ved å beregne standardavvikene til forholdsverdiene og ved å beregne persentilene $PC_{99}F_{1H}$, $PC_{99,5m}$, $PC_{99}F_{10s}$ og for alle kunder.

1.1.2 Det er samme distribusjonen blant høyfrekvent og lavfrekvente effektmålinger mellom næringsbygg, husholdninger og hytter.

Denne hypotesen besvares ved å plotte fordelingskurver, spredningsplott og boksplott av forholdsverdier og sammenligne variansen i dataen. Dersom variansen i forholdsverdier for de ulike kundekategorier er mindre eller lik 0.15, kan hypotesen bekreftes.

1.1.3 Det er forskjeller i variasjoner mellom høylast og normallast timer

Denne hypotesen besvares ved å sammenlikne $F_{10S,normal\_last}$ og $F_{5m,normal\_last}$ med $F_{10S,Høy\_last}$ og $F_{5m,Høy\_last}$ for individuelle kunder. Videre plottes nevnte forholdsverdier i histogrammer og spredingsplot. Hypotesen bekreftes eller avkreftes ved visuell inspeksjon av plottene.

2 Datagrunnlag

Dataen som analyseres består av 3 verdityper: tidspunkt (TimeStampUtc), målepunktid (UsagePointId) og effektmåling (Value), se Figur 1.

	TimeStampUtc	UsagePointId	Value
0	2020-04-01 00:00:10	707057500	1399.0000
1	2020-04-01 00:00:20	707057500	1400.0000
2	2020-04-01 00:00:30	707057500	1400.0000
3	2020-04-01 00:00:40	707057500	1400.0000
4	2020-04-01 00:00:50	707057500	1399.0000
...	...	...	...

Figur 1: Dataoppsett for 10-sekundsoppløsning

Effektmålingen er gitt i watt og representerer gjennomsnittlig effektforbruk i tidsintervallet som mål. For eksempel, i Figur 1 vil målingen kl. 00:00:10 representere gjennomsnittlig effektforbruk i intervallet [00:00:00, 00:00:10). I rapporten analysers kunder innen husholdning, næring og fritid (hyttekunder). Tabell 1 viser oppsummerende informasjon om datasettene som analyseres.

Tabell 1: Informasjon om datagrunnlaget

Datagrunnlag	**CCD**	**Tibber**
Antall kunder	187	9
Antall datapunkter	10 859 460	4 782 773
Frekvens	5-minutter	10-sekunder
Husholdningskunder	98	9
Fritidskunder	30	0
Næringskunder	59	0
Antall kunder slettet på grunn av for lite data.	6	0
Periode	2020-06-14 til 2021-06-14	2020-04-01 til 2020-06-14

3 Formatering av datagrunnlaget

Datagrunnlaget ble lastet ned fra *Data Lake generation 2* med Spark SQL til *Azure Databricks* og analysen av de ulike datagrunnlagene ble utført ved bruk av programmeringsspråket Python. Det ble brukt ulike innebygde Python-biblioteker, slik som *Pandas* for å formatere og filtrere data og *matplotlib.pyplot* for å plotte ulike grafer. Videre ble programvaren Power BI benyttet til å generere interaktive figurer av relevante resultater.

Noe av datafiltrering og formattering var felles for begge datagrunnlagene. Det første var å slette duplikater fra datasettene. Det andre var at effektverdiene basert på høyfrekvens- og lavfrekvensoppløsning måtte gjøres om fra *objekter* til *floats* (desimaltall). Dette ble gjort for å muliggjøre matematiske utregninger. Det andre var å dele *TimeStampUtc* kolonnen inn i fire ulike kolonner: *år, måned, dag* og *time*, som vist i Figur 2. Dette ble gjort for å muliggjøre korrekt sammenslåing av flere *dataframes* der effektverdiene baserer seg på ulik frekvens.

	Time	UsagePointId	Value	Year	Month	Day	Hour
0	2020-04-01 00:00:00	707057500	1399.0	2020	4	1	0
1	2020-04-01 00:00:10	707057500	1400.0	2020	4	1	0
2	2020-04-01 00:00:20	707057500	1400.0	2020	4	1	0
3	2020-04-01 00:00:30	707057500	1400.0	2020	4	1	0
4	2020-04-01 00:00:40	707057500	1399.0	2020	4	1	0
...	...	...	...	...	...	...	...

Figur 2: Formatert Dataframe

3.1 CCD

Når data ble lastet ned fra databasen, var den ikke komplett, og måtte derfor formatteres og filtreres, slik at nøyaktig analyse kunne bli utført. Det første som ble gjort var å slette alle 5-minuttsverdier som var 0 slik at timesverdier ikke skulle basere seg på feilaktige målinger. Videre ble kunden med mest fullstendig data (referanse) funnet, det vil si flest effektmålinger slik at resten av kundene kunne bli sammenlignet med denne kunden. Kundene som hadde 50% mindre data enn referansen, ble slettet fra datagrunnlaget. Dette ble gjort for å unngå unødvendig sletting av data ved senere filtrering.

Med en oppløsning på 5-minuttsnivå, vil effekten registrert i starten av en bestemt time være det gjennomsnittlige forbruket 5 minutter før den aktuelle timen. Som følge av dette, ble hele datagrunnlaget forskjøvet 5 minutter tilbake i tid, slik at en vilkårlig time startet xx:55 og sluttet yy:55.

For at analysen skulle bli så nøyaktig som mulig, ble det sjekket at hver time inneholdt 12 5-minuttsverdier. Dersom dette ikke var tilfelle, ble den timen med manglende 5-minuttsverdier slettet fra datasettet. Før denne filtreringen, var det 906.144 rader med unike timesverdier, mens etter filtrering ble dette antallet til 904.955 - 1189 rader med timesverdier ble slettet. Denne metoden ble brukt for analyser av fordelingskuver for enkeltkunder. For å utføre analyse av fordelingskurver for flere kunder samtidig, var det nødvendig at effektmålinger for de aktuelle kundene baserte seg på like tidsstempler. For å oppnå dette, ble det brukt en *for-loop* til å gå gjennom hele datasettet, og slette tidsstempler som

ikke var felles. Resultatet ble lagret i en dataframe som ble kun brukt for analyse av fordelingskurver for flere kunder samtidig.

Med riktig antall 5-minuttsverdier innen hver time kunne 5-minuttsverdier aggregeres opp til timesverdier. DF_{5m} ble gruppert etter målepunkt, og alle verdier innen hver time ble summert, delt på antall 5-minuttsverdier (gjennomsnitt) og resultatet ble lagret i DF_{1h}. DF_{5m} og DF_{1h} ble så slått sammen til en felles dataframe, $DF_{5m\ \&\ 1h}$, som inneholdt målepunkt, dato og klokkeslett for registrering av effekt, 5-minutts- og timeseffekt og forholdet mellom disse to effektverdiene. Dette ble gjort ved bruk av en *Pandas* funksjon, *merge*. Ettersom DF_{5m} inneholdt tidsstempel som baserte seg på 5-minuttsverdier, mens DF_{1h} baserte seg på timesverdier, kunne ikke disse slås sammen på kun målepunkt og tidsstempel. Løsningen var å basere sammenslåing på kolonnene for målepunkt, år, måned, dag og time som omtalt tidligere, slik at timesverdien skulle legge seg på riktig 5-minuttsverdi.

3.2 Tibber

Med en oppløsning på 10-sekundsnivå, vil effekten registrert i starten av en bestemt time være det gjennomsnittlige forbruket 10 sekunder før den aktuelle timen. Som følge av dette, ble hele datagrunnlaget forskjøvet 10 sekunder tilbake i slik at en effektmålingene representerte gjennomsnittlig effektforbruk til en spesifikk kunde fra nåværende tidspunkt til neste tidspunkt det skulle blitt foretatt en måling. Altså, etter tidsforskyvingen vil en måling kl. 00:00:10 representere gjennomsnittlig effektforbruk i intervallet [00:00:10, 00:00:20).

For å aggregere tibberdataen opp til timesverdier ble DF_{10s} gruppert etter målepunkt, og alle verdier innen hver time ble summert, delt på antall 10-sekundsverdier(gjennomsnitt) og resultatet ble lagret i DF_{1h}. DF_{10s} og DF_{1h} ble så slått sammen til en felles dataframe, $DF_{5m\ \&\ 1h}$, som inneholdt målepunkt, dato og klokkeslett for registrering av effekt, 10s- og timeseffekt og forholdet mellom disse to effektverdiene. Dette ble gjort ved å bruke *Pandas* funksjonen *merge* til å slå sammen rader med samme id, år, måned dag og time. For å verifisere at timesaggregeringene var korrekte ble det bekreftet at gjennomsnittet til 10-sekundsverdiene var tilnærmet likt gjennomsnittet til de genererte timesverdiene og at utvalgte timesverdier var likt gjennomsnittet at alle 10-sekundsverdier innenfor timene.

Etter aggregeringen ble en kolonne med forholdsverdier generert ved dele 10-sekundersverdiene på timesverdiene. Videre ble en ny dataframe med høylast verdier generert ved å tekke ut rader tilhørende de 10% høyeste timesverdiene. Disse dateframesene ble lastet opp til datalake-en benyttet av Agder energi for videre analyse i Power BI og en dedikert notebook i Databricks for dataanalyse.

3.3 Generering av kumulative fordelingsfunksjoner

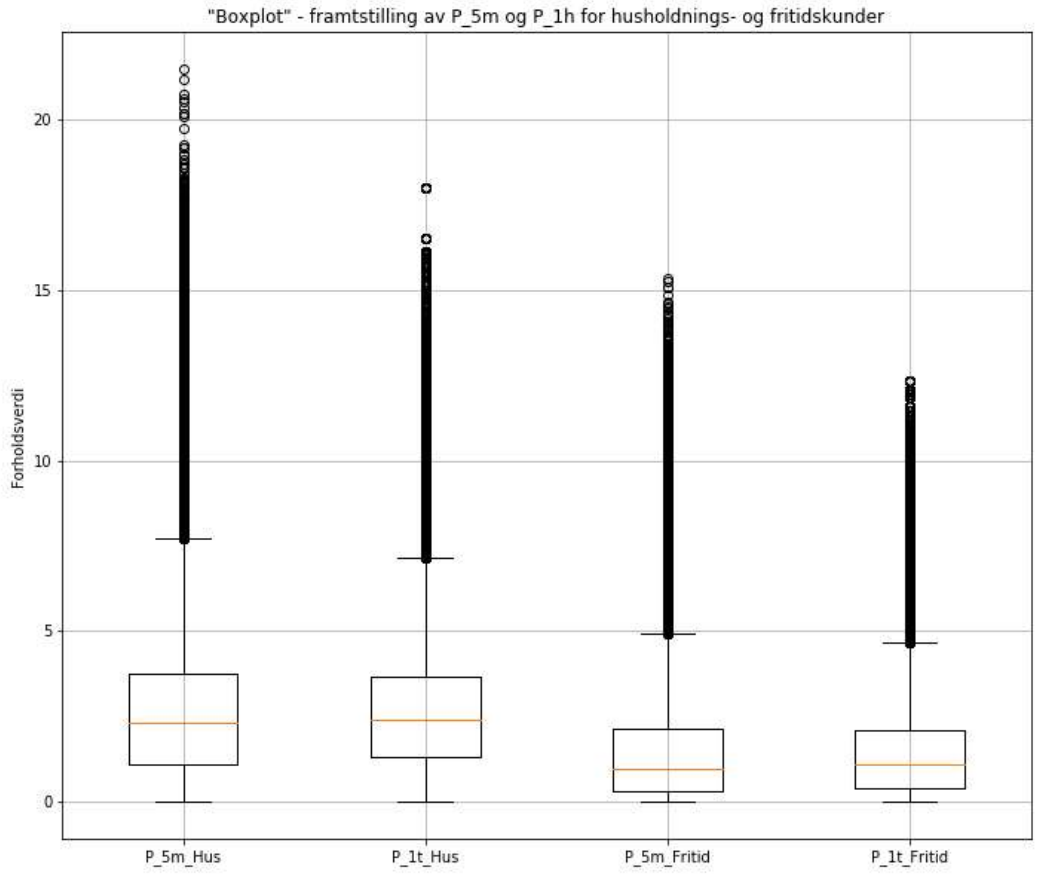
Datagrunnlaget for plotting av kumulative fordelingsfunksjoner for Tibber- og CCD-dataen ble generert i med pandasfunksjonene *groupby* og *rank* i *databricks*. Funksjonene ble brukt til å sortere alle forholdsverdier for hver kunde og rangere verdiene fra 0 til og med 1(0%,100%) der ranken er persentilen til verdien. For eksempel, en forholdsverdi med rank/persentil 0.6 impliserer at 60% av forholdsverdiene til en kunde er mindre eller lik verdien med rank 0.6. Disse kurvene presenteres ikke i denne rapporten, men inngår i Power BI rapportene som tilhører prosjektet

4 Analyse av CCD data

Analysen av datagrunnlaget baserer seg på å beregne forholdsverdier og tilhørende persentilverdier, PC, og å plotte og sammenligne disse verdiene for ulike kundekategorier/forbrukskategorier. I denne analysen har det blitt brukt boksplott for å vise en standardisert fordeling av måleverdier, histogrammer for å sammenligne ulike fordelinger på forholdsverdier og spredningsplott for å visualisere variasjonen i forholdsverdier for hver enkelt kunde.

4.1 Sammenligning av husholdningskunder og fritidskunder

Ut fra Tabell 1 kan det merkes at det er mange flere husholdningskunder enn fritidskunder, noe som vil forårsake delvis unøyaktig analyse grunnet flere målepunkter for den ene kategorien. Figuren under viser hvordan P_{5m} og P_{1h} er fordelt for de to kategoriene.

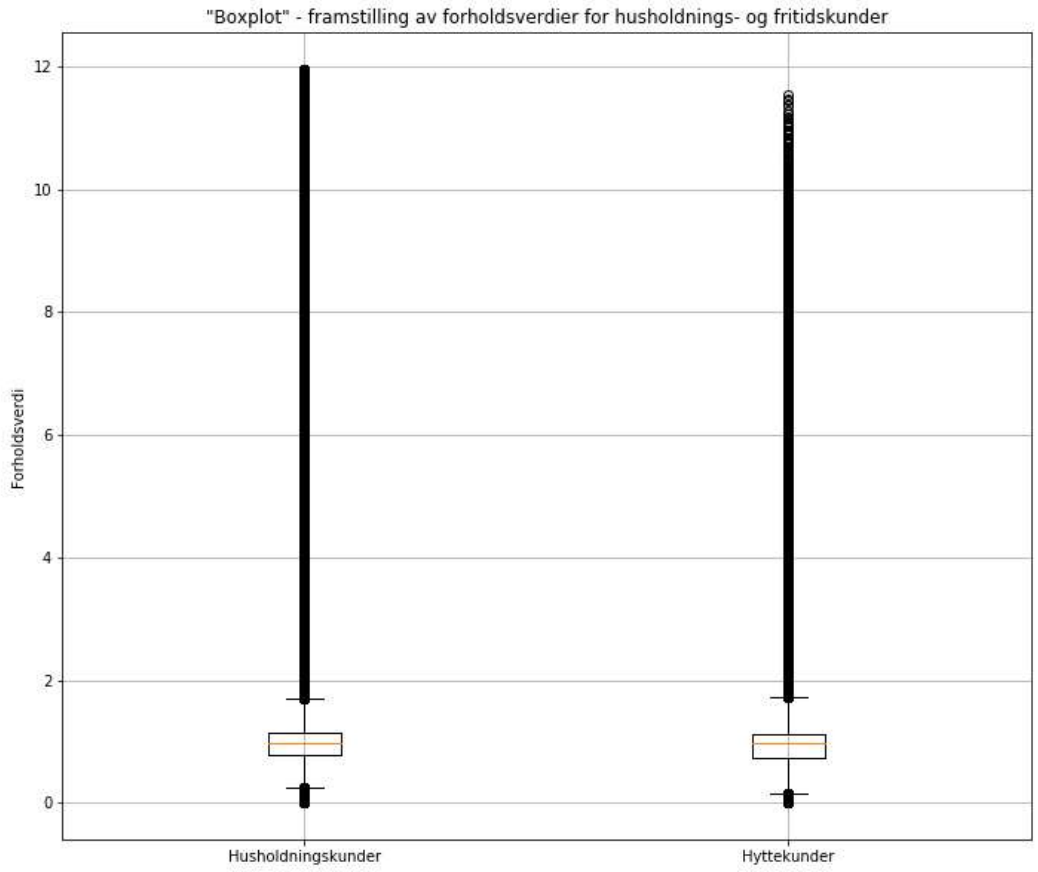


Figur 3: Boxplot av P_{5m} og P_{1h} for husholdnings- og fritidskunder

Av Figur 3 ser vi at P_{5m} for husholdningskunder har verdier opp mot 25 kW, mens timesverdiene er betraktelig lavere. Dette er fordi, som nevnt tidligere, timesverdier baserer seg på gjennomsnittlig P_{5m} verdier innen en gitt time. Nøyaktig samme situasjon observeres for fritidskunder. Den oransje streken i

boksene representerer median, arealet over streken representerer Q_{75} mens arealet under Q_{25}. Whisker på topp har en verdi tilsvarende (Q_{75} + IQR x 1.5), mens whisker på bunn (Q_{25} - IQR x 1.5).

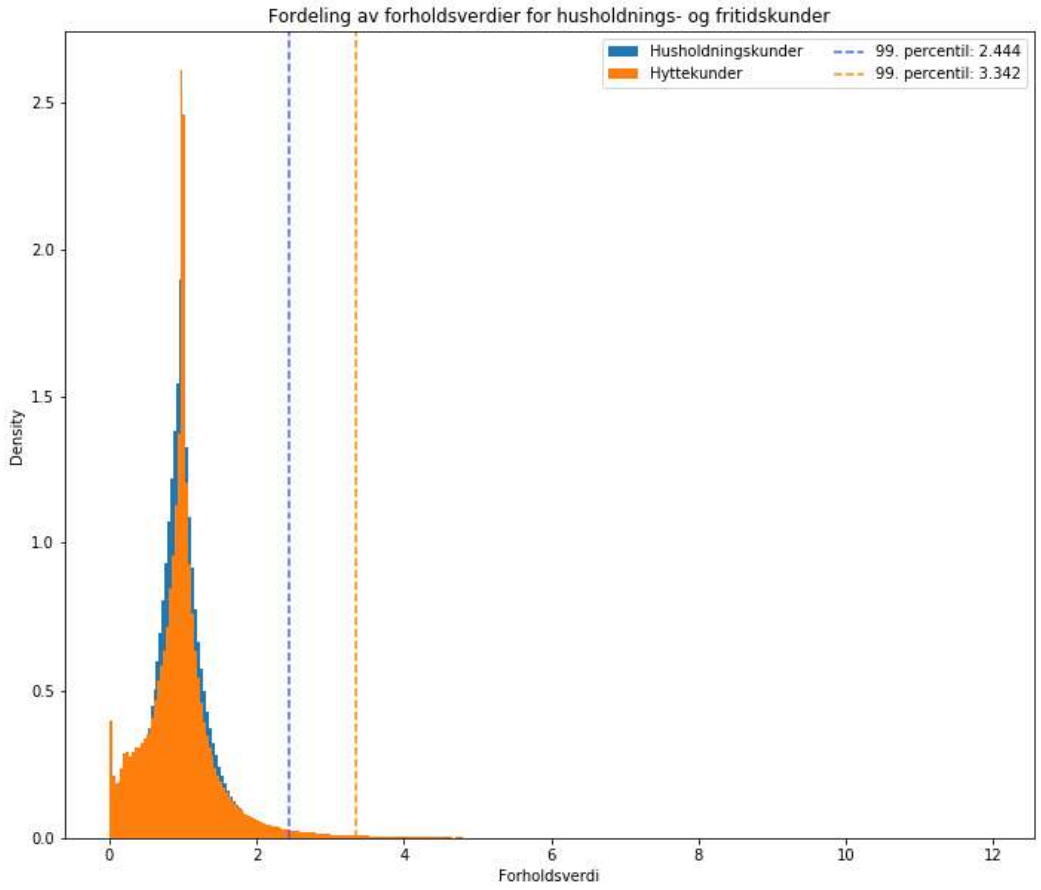
For å sammenligne effektvariasjonen innen en time, kan det først ses på boxplot for å få oversikt over hvordan forholdsverdiene er fordelt, og hvor mange av verdiene kan betraktes som ekstremalverdier:



Figur 4: Boxplot av F_{5m} for husholdnings- og fritidskunder

Figur 4 viser at begge kundekategoriene har ekstremt mange ekstremalverdier, men det er husholdningskundene som har de høyeste forholdsverdier. Medianlinjen i boksene gir inntrykk av at husholdningskundene har en symmetrisk fordeling, mens fritidskundene har en skjev fordeling (Negative Skew), der flesteparten av verdiene har høyest konsentrasjon nærmest grafens hale

Videre ble det brukt histogram, der begge kundekategoriene er vist i samme figur:

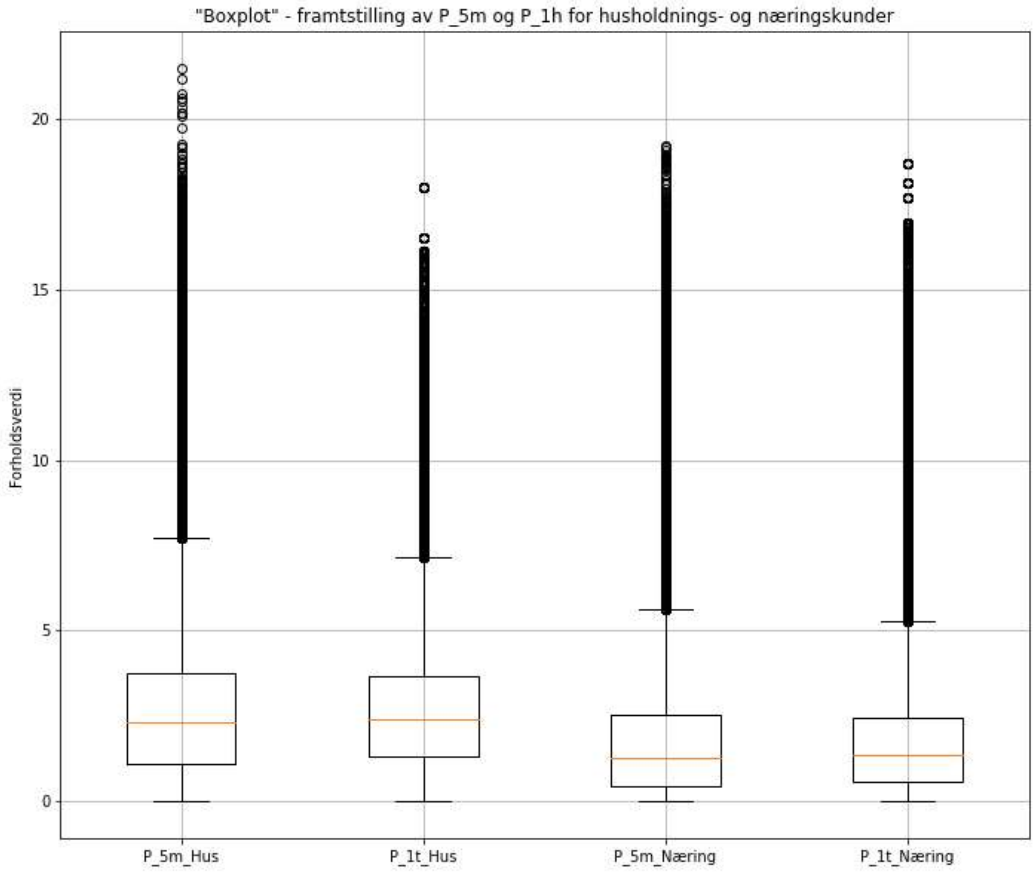


Figur 5: Histogram som viser fordeling av forholdsverdier for husholdnings- og fritidskunder

Av Figur 5, ser man at fordeling av forholdsverdier for begge kundekategoriene ligger tilsynelatende opp på hverandre, men de vertikale, stiplete persentillinjene viser at forskjellen mellom de to grafene er større enn man skulle tro. $PC_{99,5m}$ for husholdningskunder har en verdi på 2.444, mens $PC_{99,5m}$ for hyttekunder har en verdi på 3.342. Dette betyr at 99% av verdiene for husholdningskunder en mindre enn, eller lik 2.444, og 3.342 for fritidskunder. Disse numeriske verdier betyr at vi kan med 99% sikkerhet si at P_{5m} for husholdningskunder er 2.444 ganger større enn P_{1h}, og P_{5m} for fritidskunder er 3.342 ganger større enn P_{1h}.

4.2 Sammenligning av husholdningskunder og næringskunder

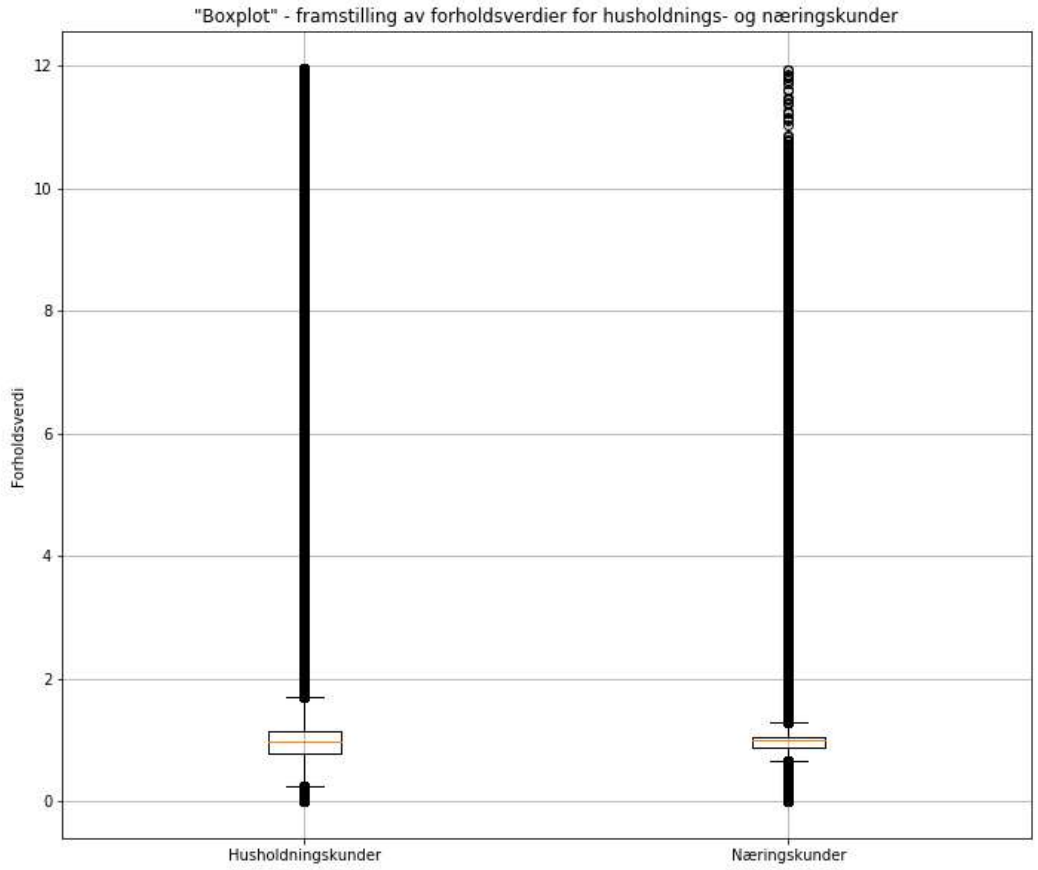
Grunnet lignende datasett, og ønske om å se på de samme tingene, ble det utført samme analyse for å sammenligne husholdning og næring som husholdning og fritid. Figuren under viser hvordan P_{5m} og P_{1h} er fordelt for de to kategoriene som skal analyseres her.



Figur 6: Boxplot av P_{5m} og P_{1t} for husholdnings- og næringskunder

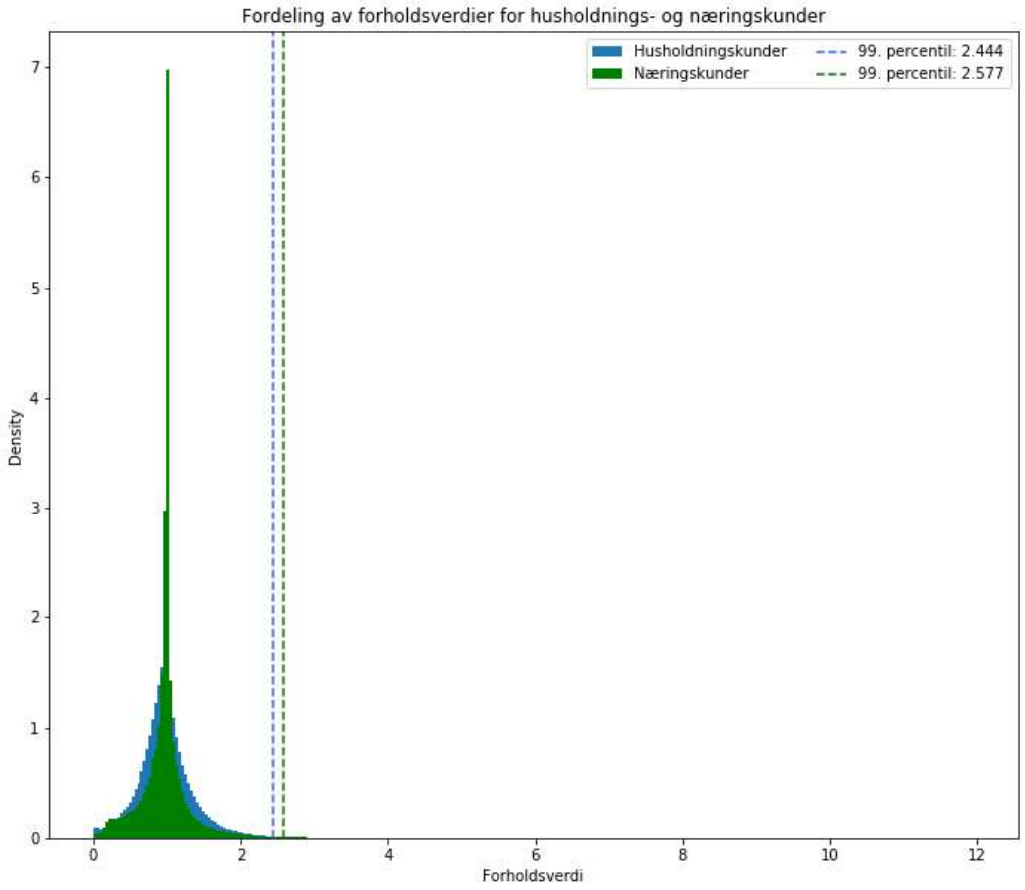
Figur 6 viser at gjennomsnittseffekt for husholdningskundene er høyere enn for næringskunder. Videre ser vi at antall P_{5m} og P_{1h}, som kan anses som ekstremalverdier for næringskundene er veldig likt i motsetning til husholdningskunder. Ved å se på antall verdier over whiskers, kan det merkes at husholdningskundene har mange flere ekstremalverdier enn næringskundene.

Figur 7 viser boxplot av forholdsverdiene for de to kundekategoriene. Man ser umiddelbart at begge kategoriene har forholdsverdier helt opp mot 12, men at fordeling av verdiene er ulik; husholdningskunder har høyest konsentrasjon av forholdsverdier ca. midt på grafen, noe som gir en tilnærmet normalfordeling, mens næringskundene har lik fordeling som fritidskunder, der fordelingens toppunkt er forskjøvet mot høyre. Det er boksen som representerer næringskunder er så smal, indikerer at det er lite spredning i forholdsverdiene.



Figur 7: Boxplot av F_{5m} og F_{1h} for husholdnings- og næringskunder

Figuren under viser fordelingene av forholdsverdier for begge kategoriene i samme plott:



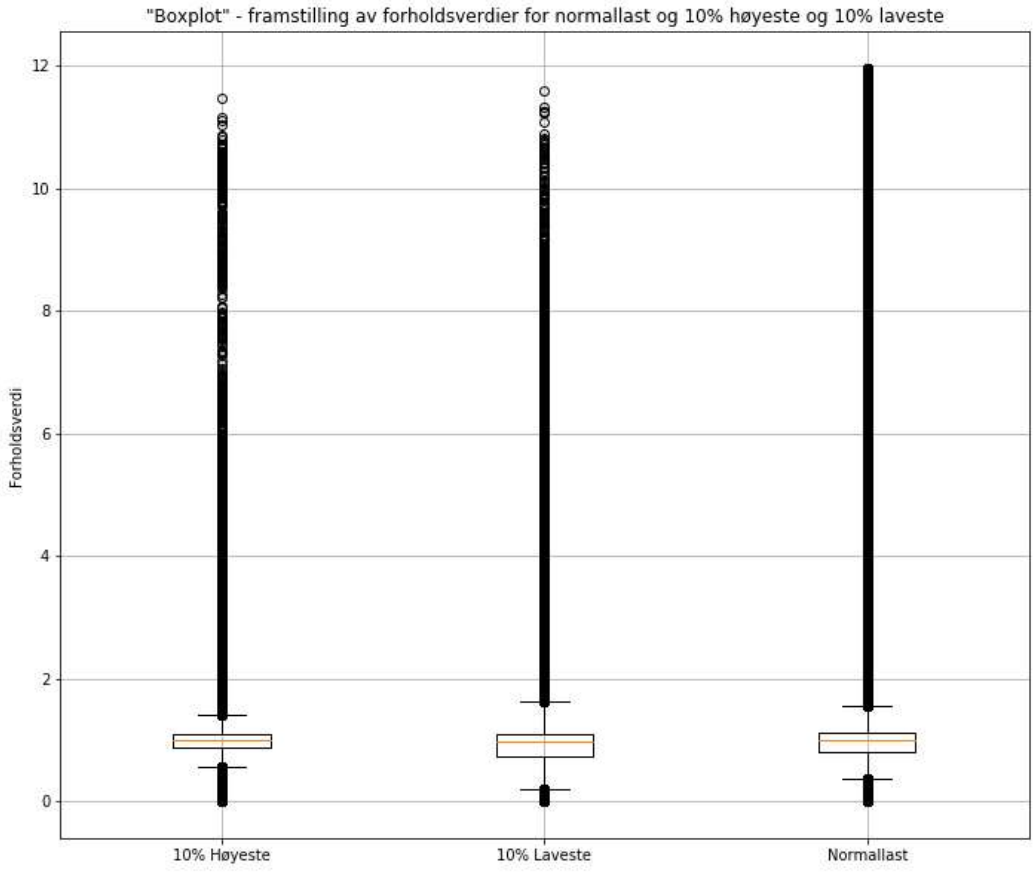
Figur 8: Fordeling av forholdsverdier for husholdnings- og næringskunder

Figur 8 bekrefter det Figur 7 viste: forholdsverdiene for begge kundekategoriene er nokså like. Forskjellen her er at næringskundene har mindre spredning av datapunkter, mens husholdningskundene har en mer symmetrisk fordeling. De blå og grønne, stiplete linjene viser PC_{99} for begge kundekategorier. Av disse er vi at PC_{99} for både husholdningskundene og næringskundene er veldig like; bare 0.133 i differanse.

4.3 Sammenligning av høylasttimer, lavlasttimer og normallast

I denne analysen ble det beregnet antall verdier hver kunde har og hentet forholdsverdiene som tilhører de 10% høyeste og laveste timesverdier. Grunnlaget som ble brukt for normallast er alle forholdsverdier i datasettet. Denne analysen skiller heller ikke mellom kundekategorier.

Som i tidligere delkapitler, begynner vi med å se på boxplot for å få en grunnleggende forståelse av hvordan de ulike forholdsverdiene er fordelt:



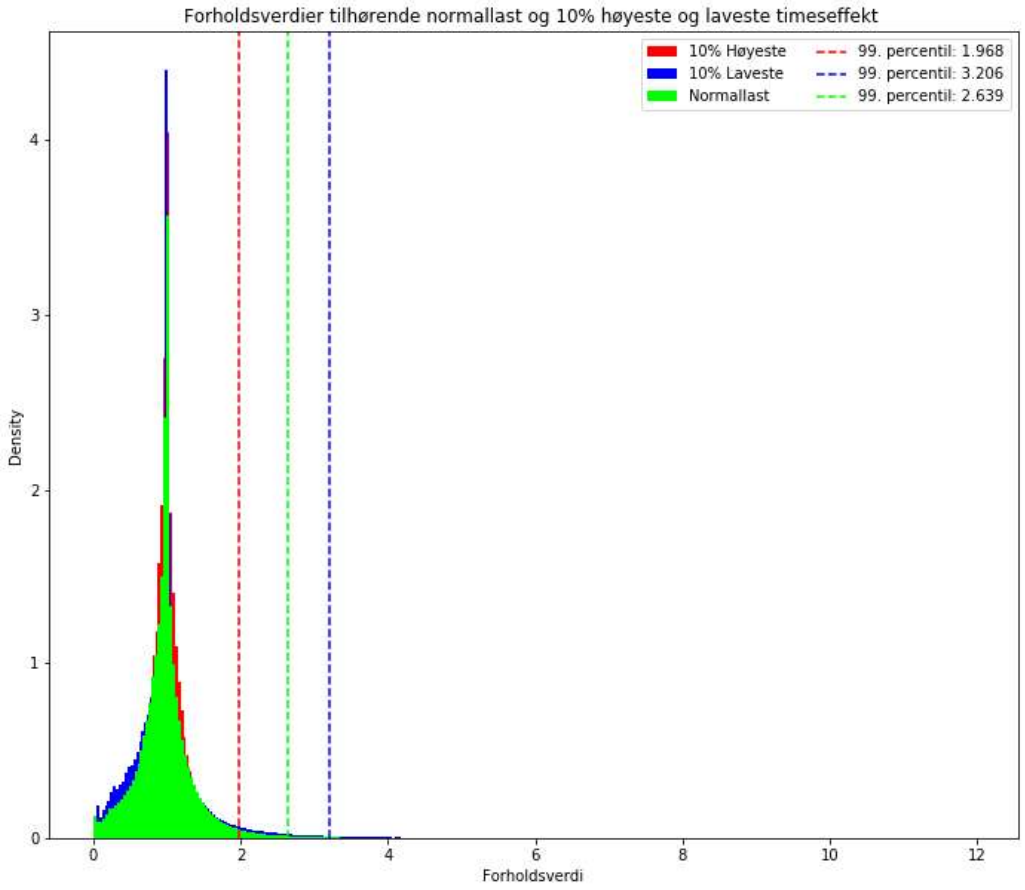
Figur 9: Boxplot av forholdsverdiene til normallast, 10% høyeste og 10% laveste

Av Figur 9 kan det ses at det er normallast og 10% laveste som har de høyeste forholdsverdiene, men det er veldig tett mellom 10% høyeste og 10% laveste. Videre ser vi at 10% høyeste har en fordeling med svært lite spredning, mens 10% laveste med mest spredning. Spredningen av datapunkter er avhengig av standardavviket, som forteller hvor stort avvik det er mellom hver måleverdi og den gjennomsnittlige verdien. Tabellen under oppsummerer standardavvik for de tre situasjonene:

Tabell 2: Standardavvik for de ulike situasjonene

Situasjon	Standardavvik
10% Høyeste	0.3307
10% Laveste	0.5719
Normallast	0.4763

Videre brukes det histogram for å få en mer visuell framstilling av forholdsverdiene:



Figur 10: Fordeling av forholdsverdier for normallast, 10% høyeste og 10% laveste

Figur 10 viser at fordelingene for de tre ulike situasjonene er relativt like. Den største forskjellen ser man på persentillinjene, der differansen mellom PC_{99} for 10% høyeste og 10% laveste er 1.238. Ved å se på standardavvik-verdiene i Tabell 2, kan vi også konkludere med at mindre standardavvik gir tettere fordelingsfunksjon, som igjen fører til redusert PC_{99}-verdi.

4.4 Analyse av kunder med ekstreme forholdsverdier

I denne delen av analysen, blir «ekstreme forholdsverdier» definert som $F_{5m} > 5$, som betyr at P_{5m} er minst 5 ganger høyere enn P_{1h}. Her tas det utgangspunkt i prosentvis differanse, som defineres som:

$$PD_{5m-1h} = \frac{P_{5m} - P_{1h}}{P_{max,5m}}$$

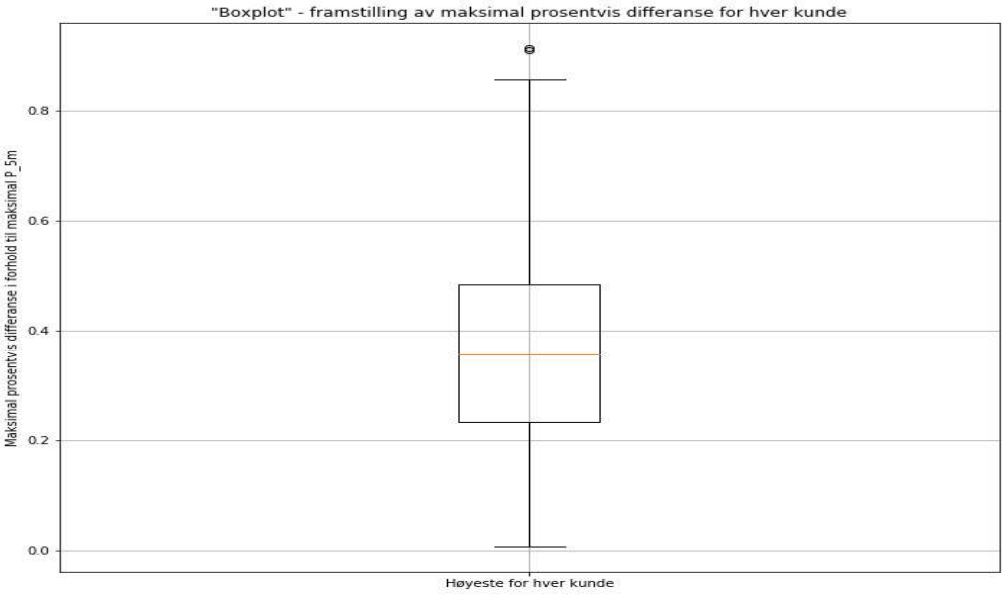
Denne metoden er veldig lik *min/max normaliserings*-metoden, og brukes for å sammenligne datapunker som er i ulik skala. Verdiene for hver kunde vil basere seg på kundens maksimale P_{5m}, noe som vil gjøre PD_{5m-1h} for de ulike kundene sammenlignbare.

I datasettet som ble benyttet for å utføre analyser, var det 111 ulike kunder som hadde ekstreme forholdsverdier. Tabell 3 viser statistikk for PD_{5m-1h}, når vi ser på den høyeste PD_{5m-1h}-verdien per kunde. Dette vil si at datagrunnlaget tabellen er basert på, inneholder en verdi per kunde, og denne verdien er høyeste PD_{5m-1h} for den enkelte kunden (max(PD_{5m-1h}) per kunde). Vi refererer videre til dette datagrunnlaget som *De høyeste PD_{5m-1h}-verdiene*.

Tabell 3: Grunnleggende statistisk informasjon om høyeste PD_{5m-1h} for hver kunde

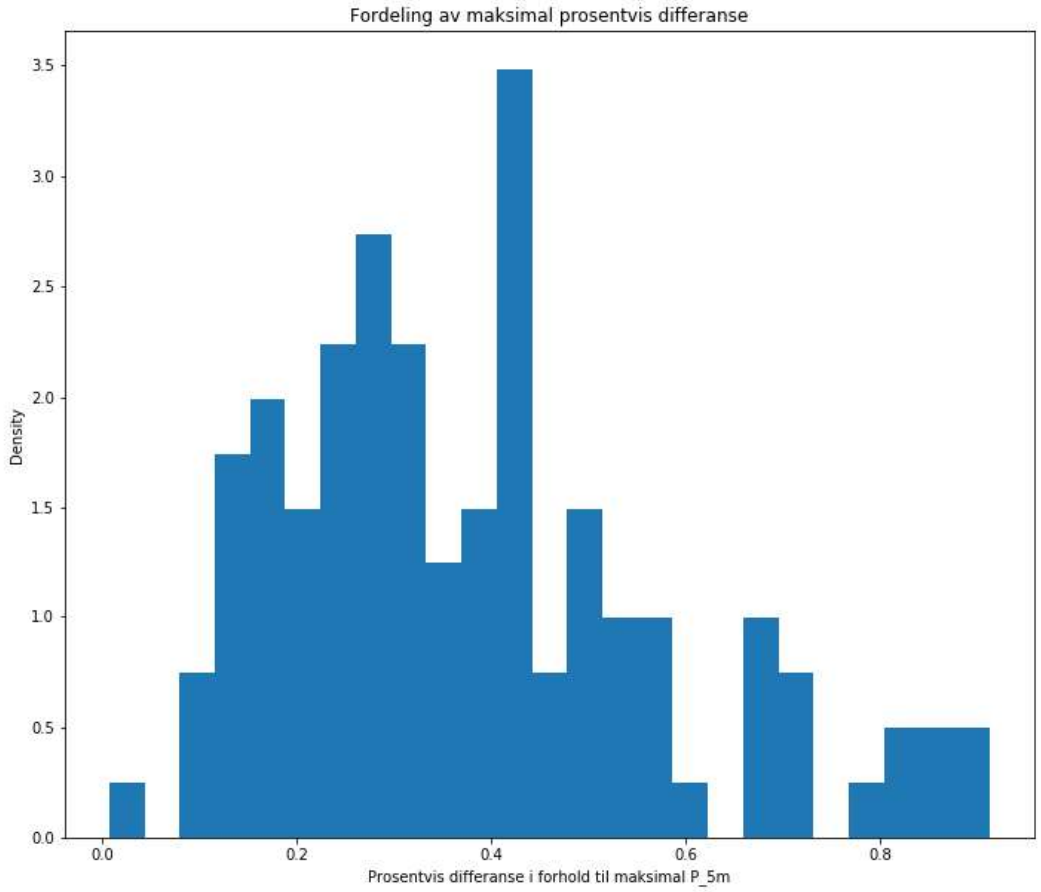
Antall kunder	111
Gjennomsnitt av PD_{5m-1h}	0.379576
Standardavvik av PD_{5m-1h}	0.197511
Laveste verdi av PD_{5m-1h}	0.006841
25. Persentil av PD_{5m-1h}	0.233604
50. Persentil av PD_{5m-1h}	0.356434
75. Persentil av PD_{5m-1h}	0.483370
Høyeste verdi av PD_{5m-1h}	0.912753

Rad 2 i Tabell 3 forteller oss at i snitt, kan kundens maksimale P_{5m} ganges med 0.379576 for å finne differansen $P_{5m} – P_{1h}$. For å få bedre oversikt over fordelingen av de høyeste PD_{5m-1h}, ble datapunktene plottet i en boxplot:



Figur 11: Boxplot av maksimal prosentvis differanse for hver kunde

Figur 11 viser at de maksimale PD_{5m-1h} er tilnærmet normalfordelt, der fleste verdier ligger rundt median. Punktene som ligger over øverste whisker viser at i de verste tilfellene, er enkeltkunders differanse P_{5m}-P_{1h}, over 80% av den maksimale P_{5m}.



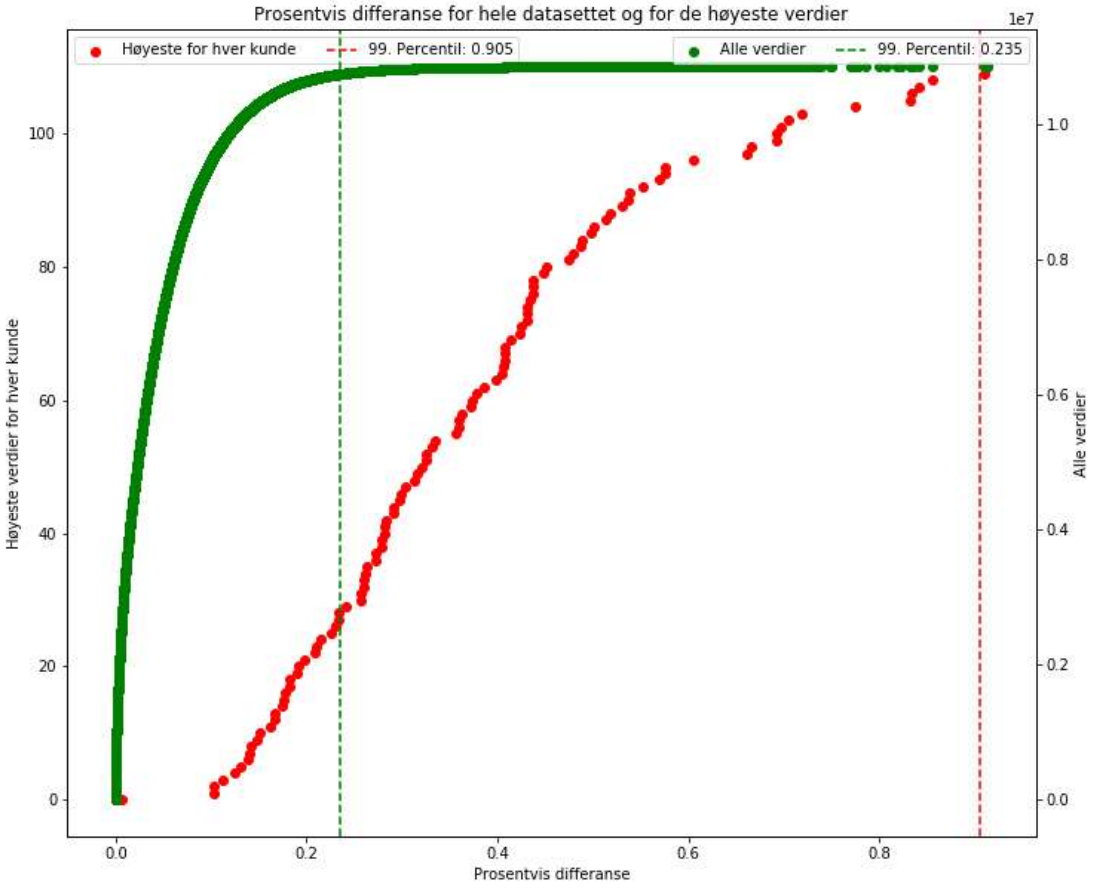
Figur 12: Fordeling av maksimal prosentvis differanse for hver kunde

Figur 12 viser en mer visuell framstilling av fordelingen. Det kan ses at flesteparten av verdiene ligger i intervallet (0.4, 0.45), som betyr at de fleste kundene har en $Diff_{5m-1h}$ som er 40%-45% av kundens maksimale P_{5m}. Videre ser vi at det at enkelte kunder har verdier i intervaller (0, 0.006841) og (0.8, 0.912753). Tabellen under viser gjennomsnittlig PD_{5m-1h}, $P_{5m}/P_{5m,\,max}$ og $P_{1h}/P_{5m,\,max}$ for disse kundene. Grunnen for valg av akkurat disse kundene er for å kunne sammenligne de kundene med høy PD_{5m-1h} (> 0.8) og lav PD_{5m-1h} (< 0.01).

Tabell 4: Oversikt over gjennomsnittlige forholdsverdier for et utvalg av kunder. Kunde 1 tilsvarer kunden med laveste PD_{5m-1h} i Figur 12 (lengst til venstre), og kundene 2-7 tilsvarer kunder med høyeste PD_{5m-1h} i Figur 12 (lengst til høyre)

	Kunde						
Gjennomsnitt	1	2	3	4	5	6	7
PD_{5m-1h}	0.01165	0.00328	0.0899	0.00179	0.05907	0.00128	0.0155
$P_{5m}/P_{5m,\,max}$	0.2362	0.0236	0.26619	0.3332	0.07179	0.00809	0.0141
$P_{1h}/P_{5m,\,max}$	0.2362	0.0236	0.26619	0.3332	0.07179	0.00809	0.0141

For å sammenligne fordelingene for alle og de høyeste PD_{5m-1h}, har alle verdiene blitt sortert i stingende rekkefølge, og plottet i et spredningsplott:



Figur 13: PD_{5m-1h} fra datasettet som inneholder alle målinger og PD_{5m-1h} fra datasettet som inneholder de høyeste PD_{5m-1h} for hver kunde

Figur 13 viser hvordan PD_{5m-1h} fordeler seg når alle verdiene blir sortert i stigende rekkefølge. Verdiene på y-aksene viser antall verdier/datapunkter. Ved å plotte PD_{5m-1h} for alle P_{5m} og P_{1h}, får vi en graf som

ligner kumulativ fordelingsfunksjon, mens grafen for de høyeste $PD_{5m\text{-}1h}$ er tilnærmet lineær. Det kan også merkes at 99% av alle verdier (grønn graf) er mindre eller lik 0.235. Den røde graden indikerer at forholdsverdiene, F_{5m} blir ekstreme, $F_{5m} > 5$, når $PD_{5m\text{-}1h}$ har en verdi på omtrent 0.1, eller 10%.

4.5 F-test

For å teste hvorvidt variansen av to ulike fordelinger er lik, bruker vi F-test. Hypotesene er definert slik:

- H_0: $\sigma_1^2=\sigma_2^2$
- H_1: $\sigma_1^2\neq\sigma_2^2$

Det første som regnes ut i en F-test er F-statistikk, som er forholdet mellom varians i to ulike fordelinger som skal sammenlignes. F-statistikk kan skrives slik:

$$F = \frac{Variasjon\ mellom\ fordelingenes\ gjennomsnitt}{Variasjon\ innen\ fordelingen}$$

Eller:

$$F = \frac{\sigma_1^2}{\sigma_2^2}$$

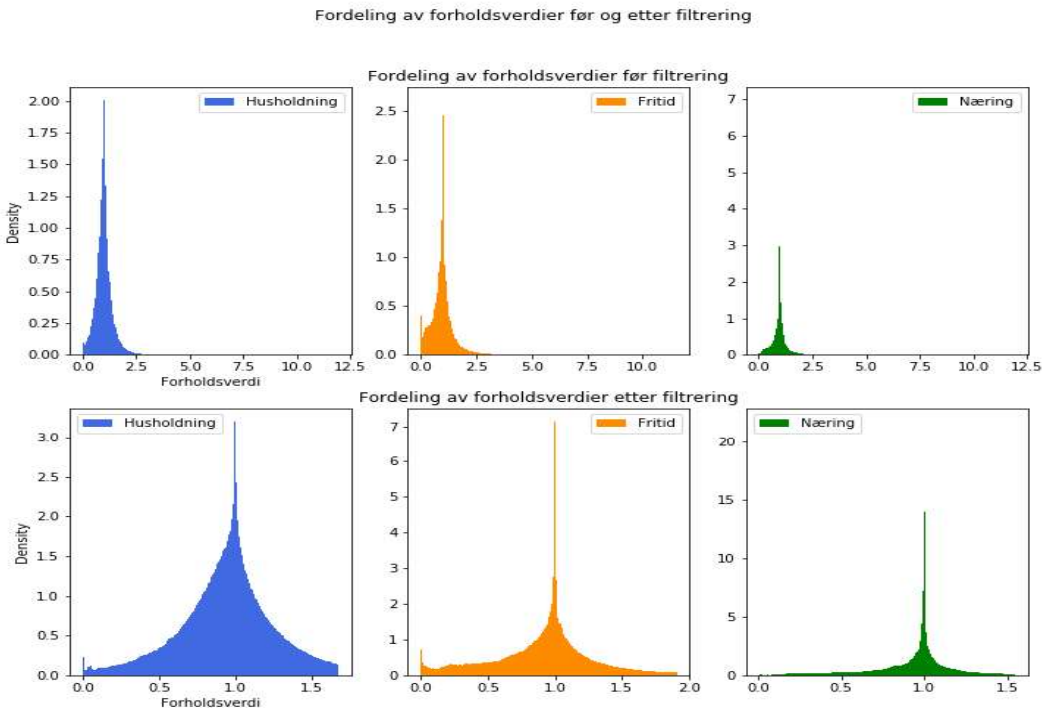
Der σ_1^2 er variansen i fordeling 1, σ_2^2 er variansen i fordeling 2 og $\sigma_1^2 > \sigma_2^2$. Her er det viktig at telleren er større enn nevneren, slik at «*left critical values*» blir unngått, og testen blir tvunget inn i en «*right tailed test*». Videre beregner man «*degrees of freedom*» for både teller (dfn) og nevner (dfd):

$$dfn = L_{fordeling1} - 1$$

$$dfd = L_{fordeling2} - 1$$

Der L står for lengden av fordeling (antall datapunkter). Ved å benytte oss av variablene definert ovenfor, kan p-verdi beregnes. Dersom denne er mindre enn et visst konfidensintervall, er konklusjon at variansen i begge fordelingene er ulike. Utregning av p-verdi baserer seg på sannsynlighetsfordeling til statistikk som det blir testet for, og blir regnet ut fra avvik mellom observerte verdier, og forhåndsvalgt referanse.

Som sett i figurene 5 og 8, er fordelingene for husholdnings-, fritids- og næringskundene tilnærmet normalfordelt, men grunnet mange ekstremverdier, får alle fordelingene lang hale på høye side av fordelingene. Dette skaper problemer når fordelingene testes for lik varians. Som en løsning på dette, tas det utgangspunkt i færre forholdsverdier, noe som gir en fordeling som er mer symmetrisk. Fordelingene filtreres slik at $F_{5m} \leq PC_{95}F_{5m}$, og resultatet kan ses i Figur 14.



Figur 14: Fordeling av forholdsverdier med alle F_{5m} og $F_{5m} \leq PC_{95}F_{5m}$

F-testen ble utført for å sammenligne variansen i fordelingene for husholdnings- og fritidskunder, og husholdnings- og næringskunder. Tabellen under viser resultater av testen:

Tabell 5: Resultater av F-testen

Kundekategori	F-verdi	p-verdi	Beholde H_0
Husholdning og fritid	1.4763	$1.116e^{-16}$	Nei
Husholdning og næring	0.6506	1	Ja

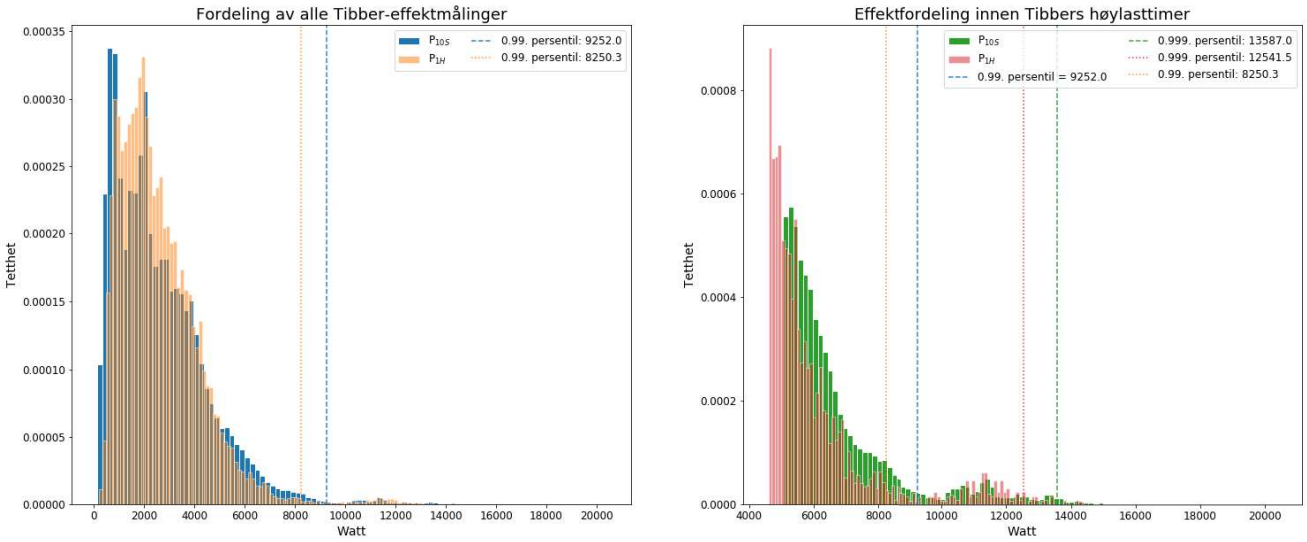
Den lave p-verdien i Tabell 5 gir oss tilstrekkelig med grunnlag for å forkaste H_0, mens den høye p-verdien gir oss tilstrekkelig med grunnlag for å beholde H_0. Ved å sammenligne resultatene av F-testen med fordelingene og PC_{99} i Figurene 5 og 8, er det rimelig å konkludere med at testen er gyldig.

5 Analyse av Tibber data

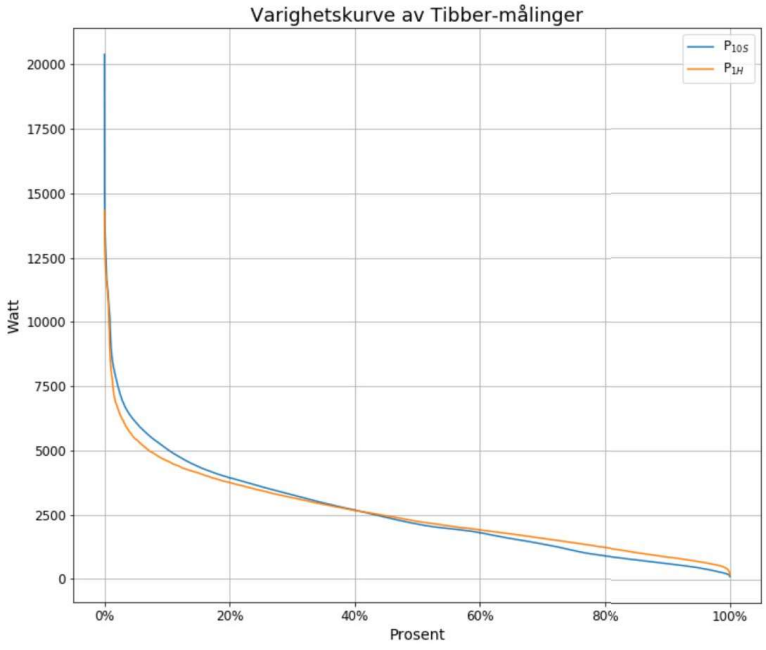
5.1 Måleverdifordelinger

Siden Tibberdataen kun inneholder data for husholdningskunder deles datasettet hovedsakelig inn i normallast og høylastmålinger i denne rapporten. Figur 15 viser fordelingen til Tibber målingene der de blå stolpene inneholder rådata (P_{10s}), mens de hudfargede stolpene representerer hvordan fordelingen hadde vært dersom alle måleverdier hadde blitt erstattet av sine gjennomsnittlige timesverdier (P_{1H}). Ved visuell inspeksjon kommer det fram at fordelingen ikke er normalfordelt og at den inneholder

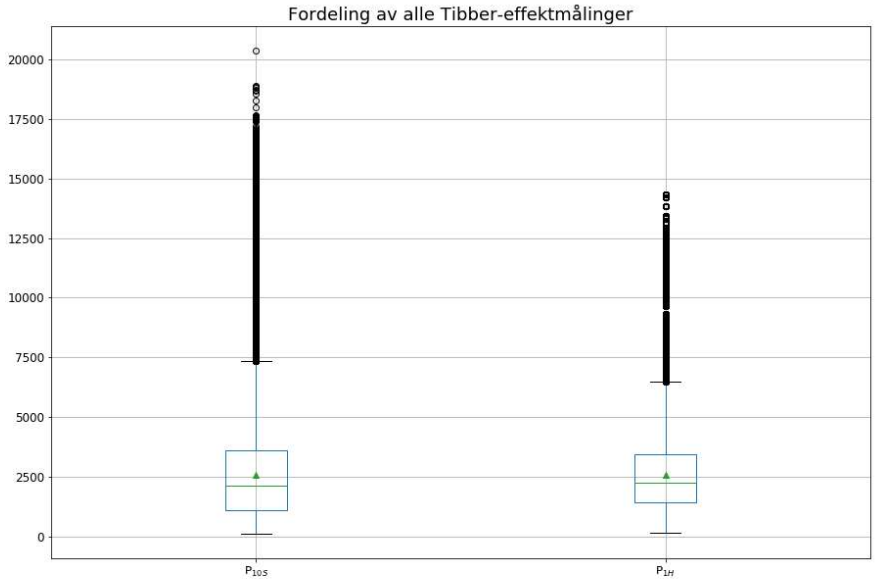
relativt stor varians, med verdier mellom 0 og 20kW. Videre kan man se fra persentillinjene at P_{10s} inneholder en del høye måleverdier som ikke blir registrert dersom målingene aggregeres til timesverdier. Figur 16 og Figur 17 viser alternative framstillinger av måleverdiene som også illustrerer hvordan P_{10s} inneholder en del toppverdier som går tapt når de times-aggregeres. Per definisjon er gjennomsnittsverdien til P_{10s} og P_{1H} like da P_{1H} er gjennomsnitt av P_{10s}-verdier, men medianen er noe høyere for P_{1H}.



Figur 15: Fordelinger av originale og timesaggregerte Tibber-målinger



Figur 16 Varighetskurveframstilling av P_{10s} og P_{1H} sine fordelinger

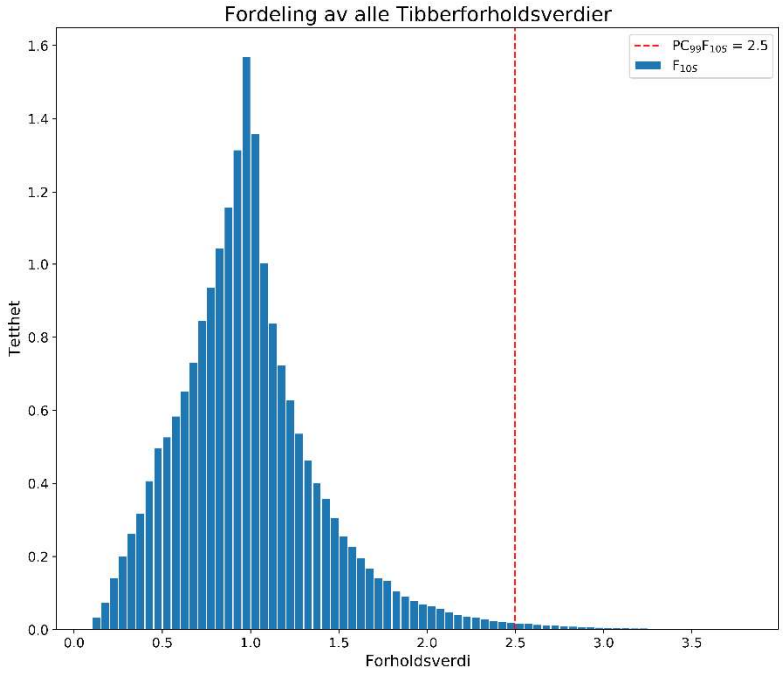


Figur 17 Boxplot framstilling av P_{10s} og P_{1H} sine fordelinger

5.2 Normallastfordelinger

Figur 18 viser fordelingen av F_{10s} (alle 10-sekundsforholdsverdier). Figuren er kuttet av ved forholdsverdi (x-akseverdi) 4, da tettheten til verdiene over dette er for liten til å sees. Ved visuell inspeksjon ser F_{10s} ut til å ha en høyreforskjøvet fordeling sentrert rundt 1. Høyrefordelingen kommer av den teoretiske minimumsverdien til F_{10} er null fordi minimumsverdien til P_{10s} og P_{1H} er null, som betyr at $\min(P_{10s}/P_{1H}) = 0$. Den teoretiske maksverdien derimot er svært stor fordi P_{10s} i teorien kan være mye større enn P_{1H}.

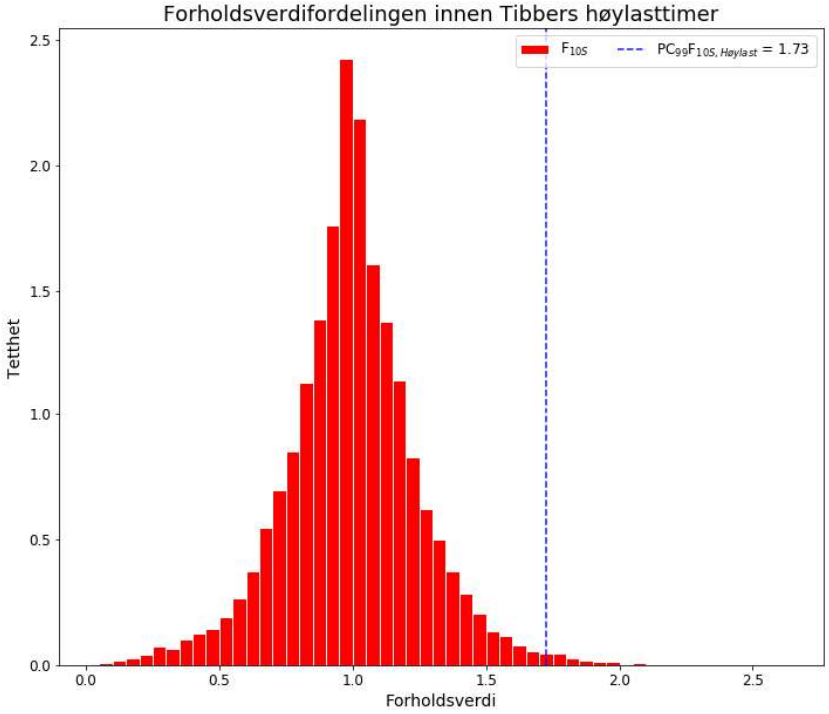
Resultatet som kommer fra figuren er som forventet, da de viser at P_{10s} ofte er relativt lik P_{1h} og at forekomsten av økende forholdsverdier synker tilsynelatende eksponentielt. Videre viser figuren at $PC_{99,10s}$ er 2,5. Altså er 99% av alle normallastmålinger i Tibber-dataen under 2,5 ganger større enn sine tilhørende timesverdier. Med andre ord vil momentaneffektuttak i gjennomsnitt overstige 2,5 ganger gjennomsnittlig timeseffektuttak i 1% av tilfellene.



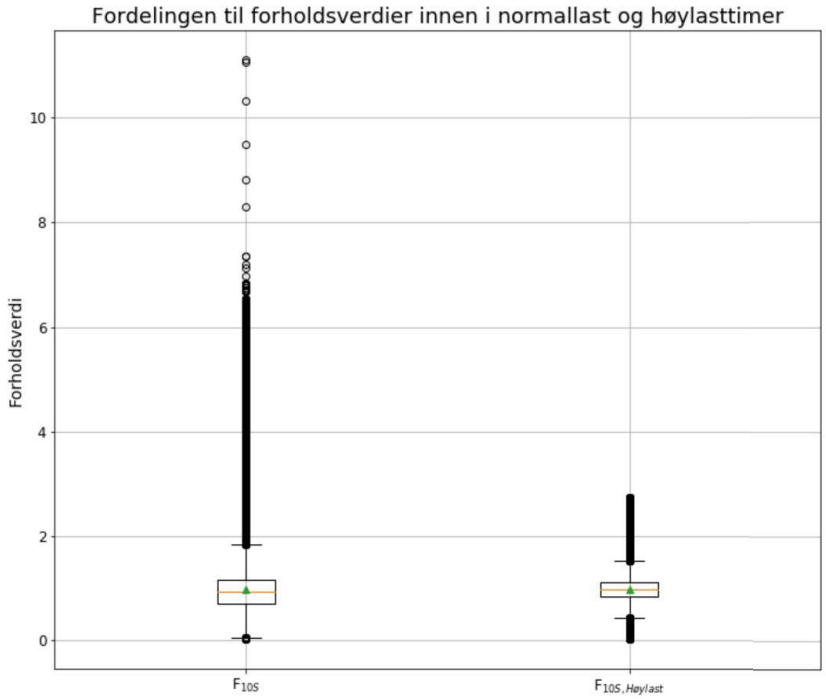
Figur 18: Fordelinger av forholdsverdier mellom orginale og timesaggregerte Tibber-målinger

5.3 Høylastfordelinger

$PC_{99,10s}$ kan være ugunstig å basere seg på, da høye forholdsverdier kan forekomme ved laster som ligger godt under godt under maksforbruket til kunden. Ved dimensjonering av nett er det maksimale effektuttaket til en kunde sentralt for hva slags sikring kunden trenger. Derfor er det spesielt interessant å analysere målinger som inngår i høylasttimer. Figur 18 viser fordelingen av $F_{10s,Høylast}$. Ved visuell inspeksjon kan man se at $F_{10s,Høylast}$ har betydelig mindre varians enn F_{10s}. Videre ser FH_{10s} ut til å være tilnærmet normalfordelt. Det kommer også fram at $PC_{10s,Høylast}$ er 1,73, noe som rundt 30% mindre enn PC_{10s}. Figur 19 er en alternativ framstilling på hvordan effektvariasjonene varierer betydelig mindre i høylasttimer kontra normallasttimer.



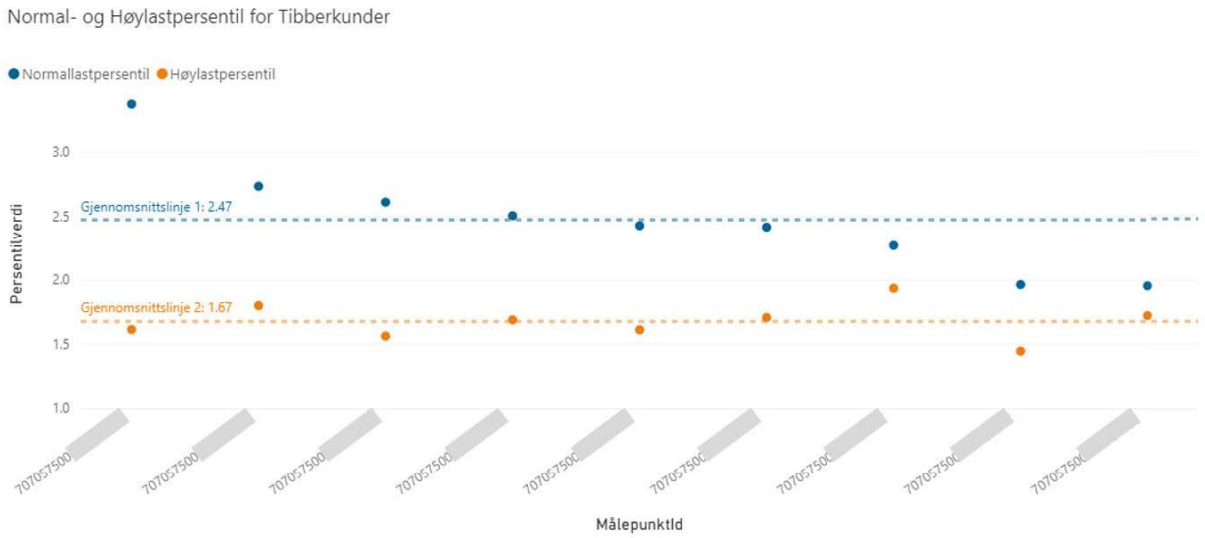
Figur 19: Fordelinger til forholdsverdier av originale og timesaggregerte tibbermålinger i høylasttimer



Figur 20: Boxplot framstilling av fordelinger til F_{10s} og F_{10s} i normallast og høylasttimer

5.4 Individuelle kundeverdier

Hittil har vi sett på alle kundene i Tibber-dataen under ett, men det er også interessant å se på forholdspersentilene (PC_{10s}) for ulike kunder for å analysere individuelle variasjoner. Figur 20 viser PC_{10s} og $PC_{10s,høylast}$ for alle enkeltkunder i datasettet. Her kommer det fram at de fleste kunder har betydelig større effektvariasjoner innen normallasttimer enn innen høylasttimer. Videre kan man se at disse variasjonene varierer mye mellom forskjellige kunder. Et fåtall av kunder har relativt like persentilverdier for normal og høylast. Til sammenlikning er effektvariasjonene innen høylasttimer relativt konsistente blant kundene. Dette indikerer at høylastpersentilverdier for et subsett husholdningskunder generelt kan være representativt for andre husholdningskunder, mens tilsvarende normallastpersentiler kan avvike betraktelig.



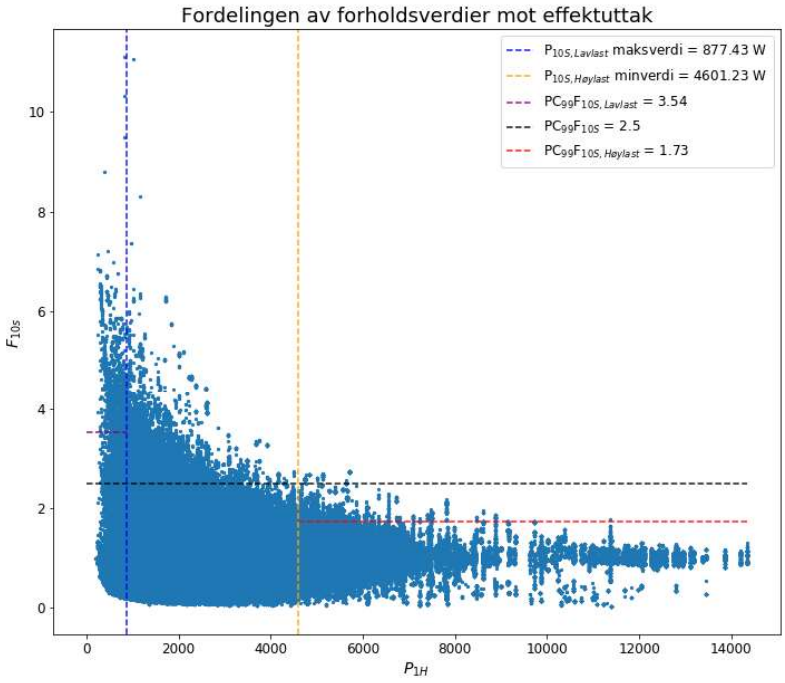
Figur 21: PC_{10s} og $PC_{10s,høylast}$ for alle individuelle kunder i Tibberdataen, samt gjennomsnittsverdien av disse

5.5 Fordelingen til forholdsverdier mot effektuttak

Vi har sett hvordan timeseffektvarisjonen er betydelig mindre i høylast timer enn i normallasttimer. En naturlig fortsettelse av analysen er derfor å studere hvordan de forskjellige forholdsverdiene fordeler seg ved forskjellige effektuttaksverdier. Dette forholdet kommer fram av figur 21. Figuren tar for seg 3 kategorier, lavlast, høylast og normallast. 10% av forholdsverdiene befinner seg under lavlastgrensen på 877,43 Watt. 10% av verdiene befinner seg over terskelen for høylast på 4601,23 Watt. Normallast tar for seg alle lastverdier i datasettet, og må ikke forveksles med området mellom lav og høylastgrensene.

Figuren viser effektvariasjoner synker ved økende lastnivåer. Ved lavlast er 1% av forholdsverdiene over 3,54, som vist av den lilla linja. Der er også ved lavlastgrensen vi ser de største forholdsverdiene på over 10. Merk at forholdsverdiene og normallast og høylast persentilverdiene er de samme som i Figur 18 og 19, bare plottet opp timesverdiene det ble regnet ut ifra, i stedet for i histogramform. Derfor ser vi igjen at 99% av forholdsverdiene er under 2,50 (svart linje) ved normallast og under 1,73 (Rød linje) ved høylast. Dette tyder på at kunder med lavere momentanforbruk generelt har høyere effektvariasjoner

innenfor timer enn kunder høyere momentanforbruk. Med andre vil ord skyldes høye forholdsverdipersentiler hos enkeltkunder at kunden generelt har lavere 10-sekundsverdier enn kunder med lave forholdsverdipersentiler. Denne hypotesen kan bekreftes for kundene i datasettet ved å analysere effektforbruket til enkeltkundene i Power BI rapporten som tilhører dette prosjektet.



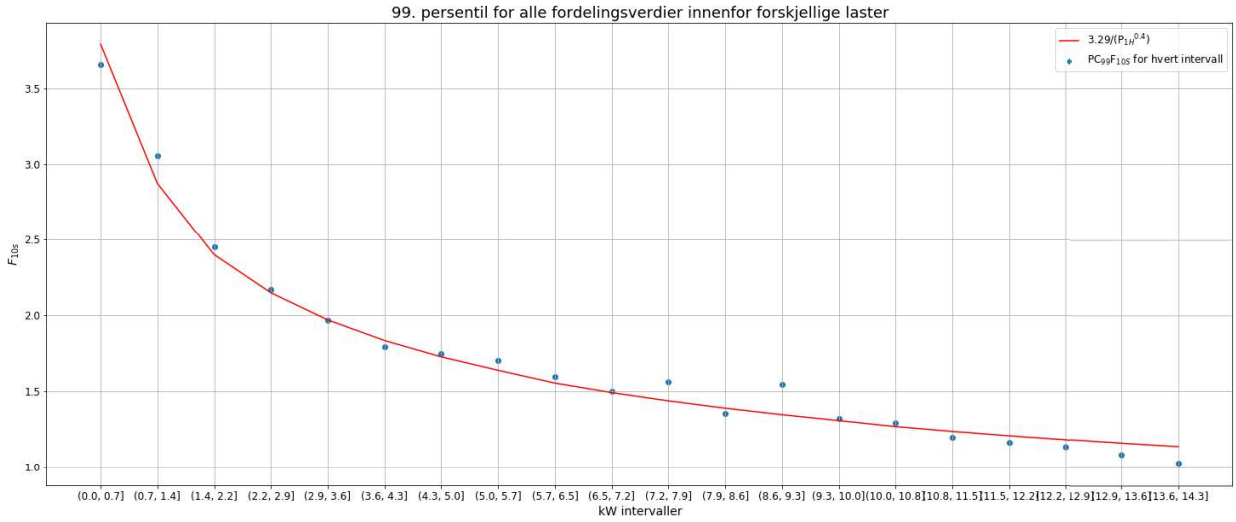
Figur 22: Spredningsplott over hvilke forholdsverdier som befinner seg ved ulike lastnivåer.

Vi kan bygge på Figur 21 og finne $PC_{99}F_{10s}$ ved flere last intervaller. Figur 23 viser hvordan disse forholdsverdiene fordeler seg på ved 20 gjevt spredde lastintervaller. I likhet med Figur 21 ser vi en klar nedgående trend. Det er mange måter å modellere denne trenden på, og det er ikke et mål i denne rapporten å finne en optimal modell for å predikere persentilene. Det er likevel rimelig å anta at $PC_{99}F_{10s}$ til en viss grad kan beskrives med en power law funksjon (se formel) fordi F_{10s} er definert som en høyfrekvent delt på en lavfrekvent måleverdi (P_{10s}/P_{1H}). Når P_{1H} øker kreves det større og større P_{10s}-verdier for å opprettholde høye forholdsverdier. For eksempel, det er mer rimelig at det forekommer en P_{10s} verdi på 5kW når P_{1H} =1kW, enn at det forekommer en P_{10s} verdi på 50kW, dersom P_{1H} =10kW, fordi det skal mye til å bruke 50kW på en gang. Derfor ble en power law funksjon til å modellere persentilverdiene som følger:

$$PC_{99}F_{10s} = \frac{a}{\mathrm{P}_{1\mathrm{H}}^{\mathrm{k}}}$$

Konstantene a og k ble funnet til å bli ved å bruke funksjonen curve_fit fra Pythonbibliteket scipy.optimize, til å tilpasse formelen med non-linear least squares metoden til alle P_{1H} verdier i datasettet. Resultatet kan sees i Figur 23. Merk at den røde linjen er kontinuerlig, men og at den baserer seg på de høyre intervallverdiene. Ved effektmålinger på over 7kW er det færre datapunkter, noe som øker støyen/variansen til persentilene, men ellers passer formelen tilsynelatende godt. Videre så er formelen en veldig enkel tilnærming og er ikke ment for bruk av nøyaktig prediksjon, men den gir et

godt inntrykk av forholdspersentilfordelingen i Tibberdataen og kan brukes som utgangspunkt for videre analyser.



Figur 23: 99. persentiler for forholdsverdier for 20 forskjellige lastintervaller sammen med en «best fit»-linje

6 Resultater CCD data

Som sett i figurene 5 og 8 har de ulike kundekategoriene relativt like fordelinger, men med ulik spredning, noe som fører til ulike verdier for $PC_{99}F_{5m}$. Tabell 6 presenterer de viktigste funnene fra kapittel 5: Analyse av CCD data. Tabellen viser at der er husholdningskunder som har de høyeste P_{5m} og F_{5m} verdier, etterfulgt av næringskunder. Variansen i de ulike kundekategoriene viser seg å være veldig lik, med den maksimale differansen mellom husholdningskunder og fritidskunder. Både varians og standardavvik for husholdnings- og næringskunder er tilnærmet lik, noe som forklarer hvorfor $PC_{99}F_{5m}$ for disse kundekategoriene er så lik. Dette blir også bekreftet av F-testen utført i delkapittel 4.5, som viste at variansforskjellen mellom husholdnings- og næringskundene er tilfeldig.

Tabell 6: Viktigste informasjon fra kapittel 5: Analyse av CCD data

	Husholdning	**Fritid**	**Næring**
Antall Kunder	98	30	59
Høyeste P_{5m}	21.4827	15.3389	19.2111
Høyeste F_{5m}	11.9482	11.5315	11.9184
Varians i F_{5m}	0.1949	0.3582	0.2140
Standardavvik F_{5m}	0.4415	0.5985	0.4627
$PC_{99}F_{5m}$	2.444	3.342	2.577

Det er rimelig å anta at det er flere husholdnings- og næringskunder som er knyttet til samme nettstasjon enn det er husholdning- og fritidskunder, eller fritids- og næringskunder. Med denne

antagelsen, kan man i dimensjonering av nettet som skal forsyne husholdninger og næringer, anta at i 99% av tilfellene, er $P_{5m} \approx 2.577P_{1h}$ for både husholdnings- og næringskundene. $PC_{99}F_{5m}$ for fritidskunder derimot, er ≈30% høyere enn $PC_{99}F_{5m}$ for næringskunder, noe som må tas hensyn til i dimensjoneringen.

Tabell 7: Andel av forholdsverdier innen ulike intervaller for hver kundekategori

		Andel		
		$F_{5m} < 1$	**$1 \leq F_{5m} \leq 5$**	**$F_{5m} > 5$**
Husholdning	Høylast	0.5190	0.4809	0
	Lavlast	0.5968	0.4020	0.0011
	Normallast	0.5562	0.4432	0.0006
Fritid	Høylast	0.5457	0.4542	0
	Lavlast	0.5838	0.4107	0.0054
	Normallast	0.5664	0.4312	0.0023
Næring	Høylast	0.5374	0.4610	0.0015
	Lavlast	0.6013	0.3970	0.0016
	Normallast	0.5615	0.4364	0.0020

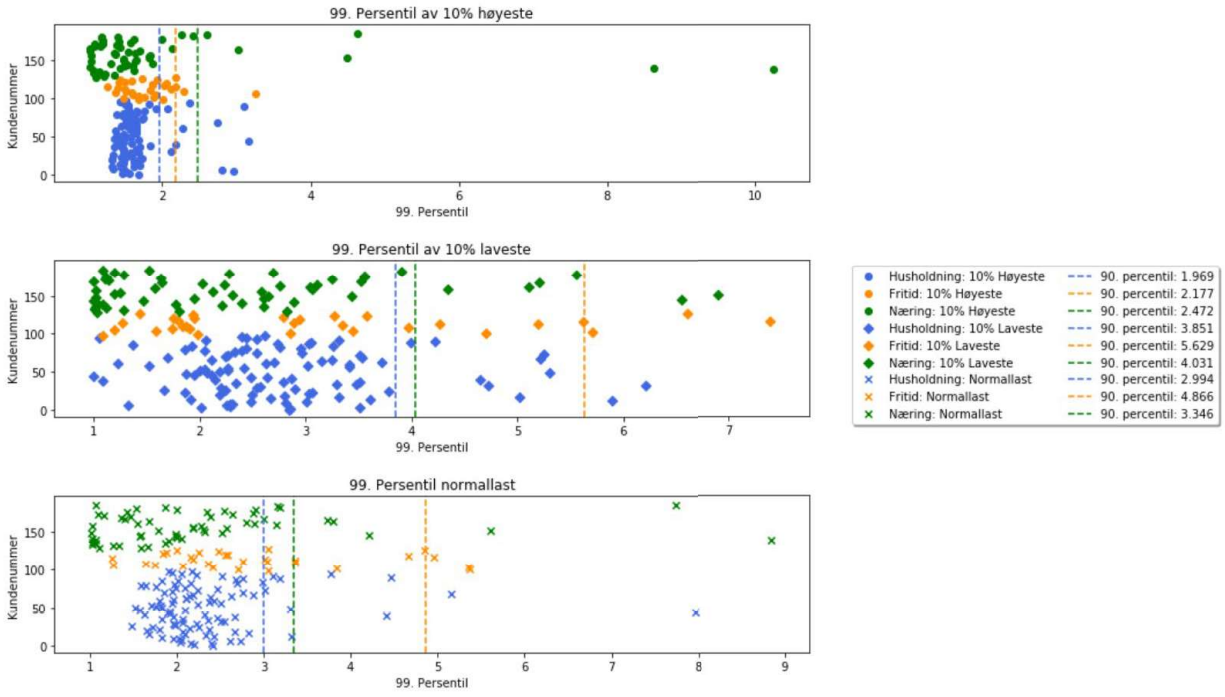
Av Tabell 7 ser vi at over 50% av alle F_{5m}, for alle kunde- og forbrukskategorier, ligger under 1, og at mellom 40% og 48% av F_{5m} ligger mellom 1 og 5. Tar vi utgangspunkt i Tabell 6, ser vi at den høyeste F_{5m} for husholdningskunder er 11.94, men av Tabell 6 ser vi at det er kun mellom 0.1% og 0.6% av F_{5m} som ligger over 5, avhengig av forbukskategori.

I Tabell 8 som viser statistisk informasjon om $PC_{99}F_{5m}$, ser vi at variansen er størst i lavlasttimene, etterfulgt av normallast. Det er spesielt fritidskundene som har høyest variasjon i lavlasttimene.

Tabell 8: Grunnleggende statistisk informasjon for $PC_{99}F_{5m}$

		Max	Gjennomsnitt	Varians	Standardavvik
	Høylast	3.166	1.65	0.128	0.359
Husholdning	Lavlast	6.208	2.77	1	1
	Normallast	7.966	2.35	0.6889	0.83
	Høylast	3.263	1.8	0.1521	0.39
Fritid	Lavlast	7.392	3.06	2.8561	1.69
	Normallast	5.374	2.87	1.3456	1.16
	Høylast	10.249	1.86	2.59	1.61
Næring	Lavlast	6.901	2.35	2.0164	1.42
	Normallast	8.846	2.29	2.1316	1.46

Figur 15 viser spredningsplott av $PC_{99}F_{5m}$ for hver kunde- og forbrukskategori. Spredningsplottet illustrerer variansen presentert i Tabell 8 og viser at generelt sett, er det størst variasjon i lavlasttimene for alle kundekategorier. De vertikale linjene er PC_{90}, og viser at det er fritidskundene som har de høyeste persentilene. Videre kan det ses at $PC_{90}PC_{99}F_{5m}$ for husholdnings- og næringskundene er veldig like, mens fritidskundene skiller seg ut, spesielt i lavlasttimene.



Figur 24: Spredningsplott av $PC_{99}F_{5m}$ for hver kunde- og forbrukskategori

7 Resultater Tibber data

Dette kapittelet oppsummerer de viktigste resultatene fra kapittel 5. Tabell 9 viser hvordan forholdsverdiene mellom timesdata og 10-sekunds data fordeler seg. Den viser hvordan det er mindre forskjeller mellom 10-sekundsmålinger og timesmålinger ved høylast, enn ved normallast. Blant annet ser vi at hvordan differansen mellom nedre og øvre Whisker og differansen mellom Q1 og Q3 er mindre ved høylast enn ved normallast. I tillegg er standardavviket til forholdsverdiene lavere ved høylast, som også innebærer at $F_{10s,Høylast}$ varierer mindre en F_{10s}.

Tabell 9: Oppsummering av forholdsverdifordelingen vist i Figur 19:

	Nedre Whisker	Q_1	Median	Gjennomsnitt	Q_3	Øvre Whisker	Standardavvik
F_{10s}	0,0607	0,733	0,959	1,00	1,18	1,86	0,442
$F_{10s,Høylast}$	0,454	0,86	0,995	1,00	1,13	1,52	0,256

Tabell 10 og 11 viser hvordan forholdsverdipersentilene varierer for normallast og høylast mellom de forskjellige kundene i tibberdataen. I tabell 10 ser vi hvodan det er mye større variasjon mellom forholdspersentiler til foskjellige ved høylast og normalast ved å se på standardavikene og differansen mellom maksimum og minumumnsverdiene for de to lastnivåene. I tabell 11 ser vi blant annet at for datasettet som helhet har vi at:

$PC_{99}F_{10s,Lavlast}=3,54$

$PC_{99}F_{10s}=2,50$

$PC_{99}F_{10s,Høylast}=1,73$

Merk at disse persentilene for hele datasettet er forskjellige fra gjennomsnittet av forholdspersentilene til de forskjellige kundene, noe som kommer av at hver kundepersentil her vektes likt uavhengig av hvor mange datapunkter de har.

Tabell 10: Oppsumering av kundepersentilene vist i Figur 20:

	Gjennomsnitt	Maksverdi	Minverdi	Standardavvik
$PC_{99}F_{10s}$	2,47	3,37	1,95	0,404
$PC_{99}F_{10s,Høylast}$	1,67	1,93	1,44	0,133

Tabell 11 Statistikker for enkeltkunder i Tibberdatasettet

UsagePointId	$PC_{99}F_{10s}$	$PC_{99}F_{10s,Høylast}$	$P_{Gjennomsnitt}$	P_{Median}
707057500	3,37	1,61	1622	1068
707057500	2,73	1,80	2008	1551
707057500	2,61	1,56	2190	1905
707057500	2,50	1,69	2486	2097
707057500	2,43	1,61	2028	1497
707057500	2,41	1,71	2471	2056
707057500	2,27	1,93	3664	2987
707057500	1,96	1,44	2784	2244
707057500	1,95	1,72	3709	3579
Hele datasettet	2,5	1,73	2610	2152

Som vist i figur Figur 23 kan $PC_{99}F_{10s}$ eksempelvis estimeres basert på en spesifikk timesverdi med følgende formel:

$$PC_{99}F_{10s} = \frac{3{,}29}{P_{1H}^{0{,}40}} \quad P_{1H} \in [0{,}7, 14{,}3]$$

Formelen er ikke kalibrert for P_{1Hs}-verdier utenfor intervallet og er ikke ment for å gi nøyaktige estimater, men den gir et godt bilde av hvordan forholdspersentilene avhenger av lasten i Tibberdataen.

8 Konklusjon

8.1 CCD

Figurene 5 og 8 i seksjonene 4.1 og 4.2 viser fordelingene av forholdsverdier for de ulike kategoriene. Av Figur 5 ser vi at fordelingen for husholdnings- og fritidskundene ligger på hverandre, men persentillinjene viser at forskjellen mellom disse fordelingene er større enn man tror. $PC_{99}F_{5m}$ for husholdningskunder er 2.444, mens for fritidskundene 3.342. Her er det en forskjell på 0.898, noe som indikerer at variasjon innenfor en time for husholdnings- og fritidskundene er ulik, og dermed kan vi forkaste nullhypotesen, H_0, presentert i seksjon 1.1. Figur 8 viser det samme som Figur 5, bare for husholdnings- og næringskundene. $PC_{99}F_{5m}$ for husholdningskundene er fortsatt 2.444, mens for næringskundene er den 2.577. Her er det en forskjell på 0.126, noe som indikerer at effektvariasjon innenfor en time for disse kundekategoriene er relativt like, og at vi kan beholde H_0. Resultatene omtalt ovenfor bekreftes av F-testen som ble utført i seksjon 4.5, som indikerer at ulik variansen i forholdsverdier mellom husholdnings- og fritidskundene er ikke tilfeldig, og det motsatte for husholdnings- og næringskundene.

Figur 10 viser høylast-, lavlast- og normallasttimene, der alle kundene blir sett på under ett. Persentilene for denne figuren er veldig ulike, og når vi ser på alle kundene under ett, har vi rimelig grunnlag for å konkludere med at det er ingen likhet mellom de tre forbrukskategoriene. Figur 22 derimot, viser de samme forbrukskategoriene, bare at de at delt opp for hver kundekategori. Her tas det utgangspunkt i $PC_{99}F_{5m}$. Ved bruk av visuell inspeksjon, samt persentilverdiene presentert i figurteksten, kan vi konkludere med at det er høyest effektvariasjon innen en time for alle tre kundekategoriene og at

$PC_{99}F_{5m}$ har laveste verdier i høylasttimene. Ved å analysere figur 22, har vi ingen grunnlag for å konkludere med at de tre kundekategoriene er like i de tre ulike forbrukskategorier.

8.2 TIBBER

I denne rapporten har vi sett hvordan effektforbruket innad en time i snitt ikke overstiger gjennomsnittlig timesverdier med mer enn 150% i 99% av tilfellene. Videre ble det funnet, som forventet, at effektvariasjonene er mindre i høylasttimer en ved normallast timer. Ved høylast er effektforbruket innad timene kun 73% større enn timesverdiene i 1% av tilfellene. Vi har også sett at forholdet mellom timesverdier og 10-sekundsverdier er sentrert rundt 1 og at de er noenlunde normalfordelt ved høylast og høyreforskjøvet eksponensialfordelt ved normallast. Dette vil si at de fleste F_{10s}-verdier er rundt 1 og at sannsynligheten til forekomster av F_{10s}-verdier syker eksponentielt desto lenger F_{10s} er fra 1. Høyreforskyvningen til forholdsverdiene ved normallast innebærer at fordelingen har en lang hale på høyresiden. Disse funnene illustreres iblant annet i Figur 18 og Figur 19.

Når det kommer til individuelle kundeforskjeller ble det igjen observert at forholdsverdipersentilene ved normallast er betydelig større enn ved høylast. Det ble også observert at det var stor variasjon mellom de individuelle forholdsverdipersentilene til kundene ved normallast. For en av kundene var $PC_{99}F_{10s}$=3,37, som betyr at over 1% effektuttakene innenfor timene var over 237% mer enn timesverdiene. Ved høylast derimot var det betydelig mindre varians mellom kundepresentilene, der den største forholdspersentilen var på 1,93. Dataen kan derfor tolkes som at $PC_{99}F_{10s,Høylast}$ for et utvalg husholdningskunder kan være rimelig representativt for kundegruppen som helhet, i motsetning til $PC_{99}F_{10s}$. Disse funnene illustreres i Figur 21.

Til slutt har vi sett hvordan $PC_{99}F_{10s}$ er sterkt avhengig av lastnivået. Normallast inneholder mye høyere forholdsverdier enn høylast. I tillegg synker forekomsten av høye forholdsverdier ved økende timesverdier. Dette indikerer at $PC_{99}F_{10s}$ er lavere for kunder høy forekomst at høye timesverdier, enn for kunder lav forekomst av høye effekttopper.

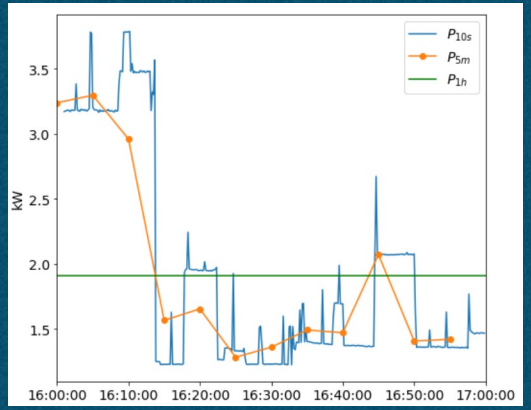
agder energi

Arendal, 25.08.2021

Effektanalyse
Pilot, CINELDI WP1

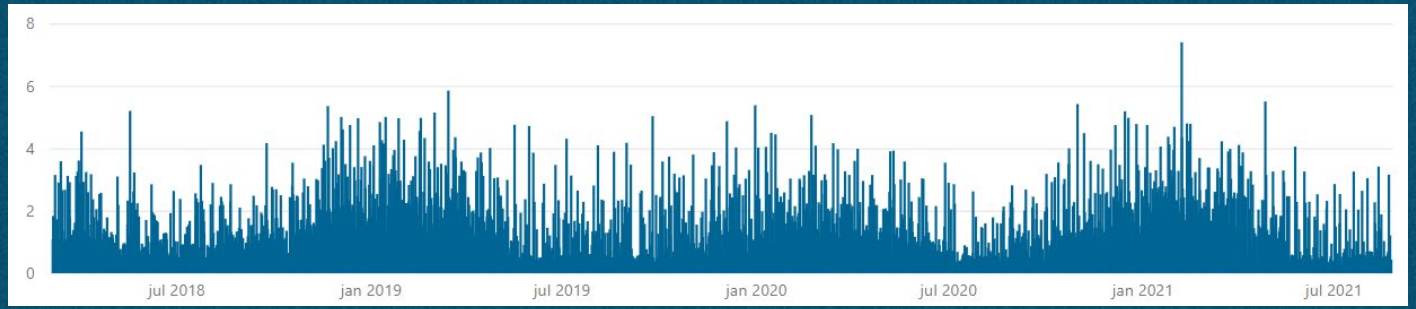
agder energi

Effektvariasjon innenfor en klokketime

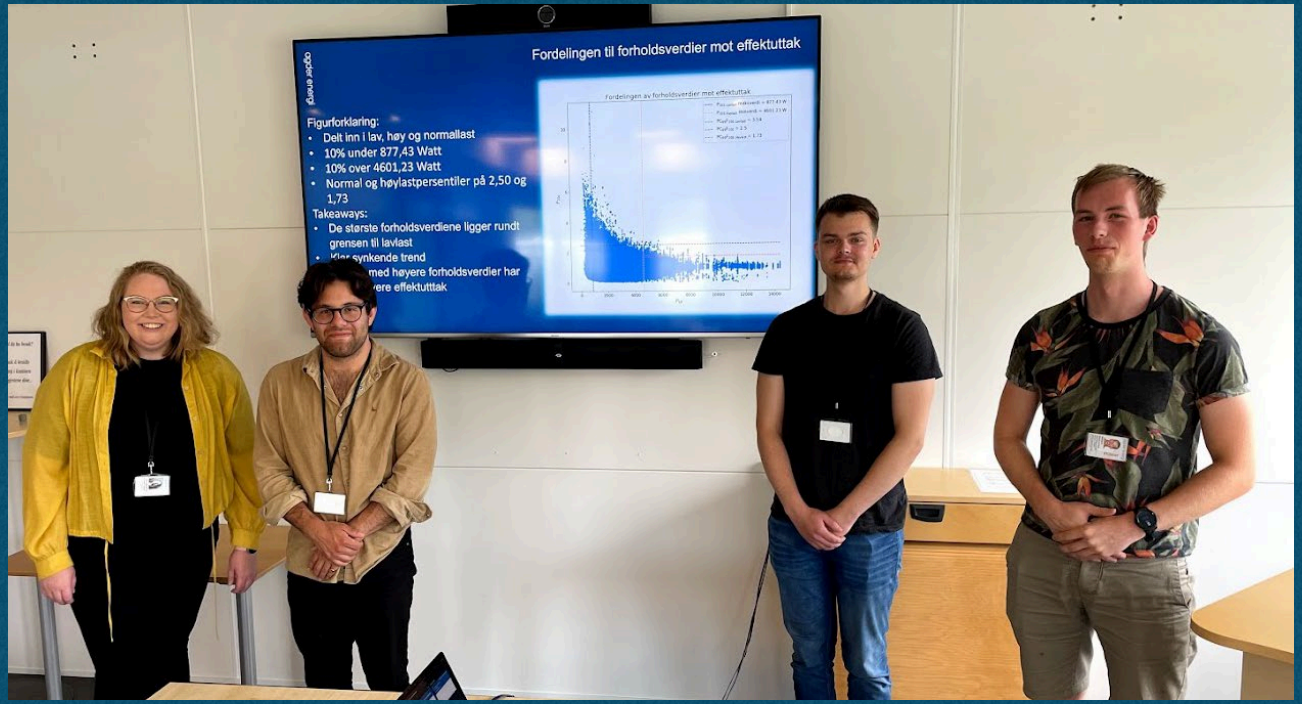


Fokus for arbeidet

Maks timeseffekt pr kunde



agder energi



Rebekka Omslandsæter

Innleid

PhD-student, IKT - UiA

Julian Gjestvang

Trainee

Master Stat/Mat

Pål Wagner

Sommerstudent,

4. år, Power Engineering

NTNU

Aksel Holbek Sørbye

Sommerstudent,

5. år, Energi og Miljø

NTNU

Analyse av effektvariasjon innenfor en time

Innhold

1. Bakgrunn for prosjekt
2. Resultater
3. Oppsummering

Bakgrunn for prosjekt – Generelt

«Analyse av effektvariasjon»
- Effekt kan variere mye innenfor en time.

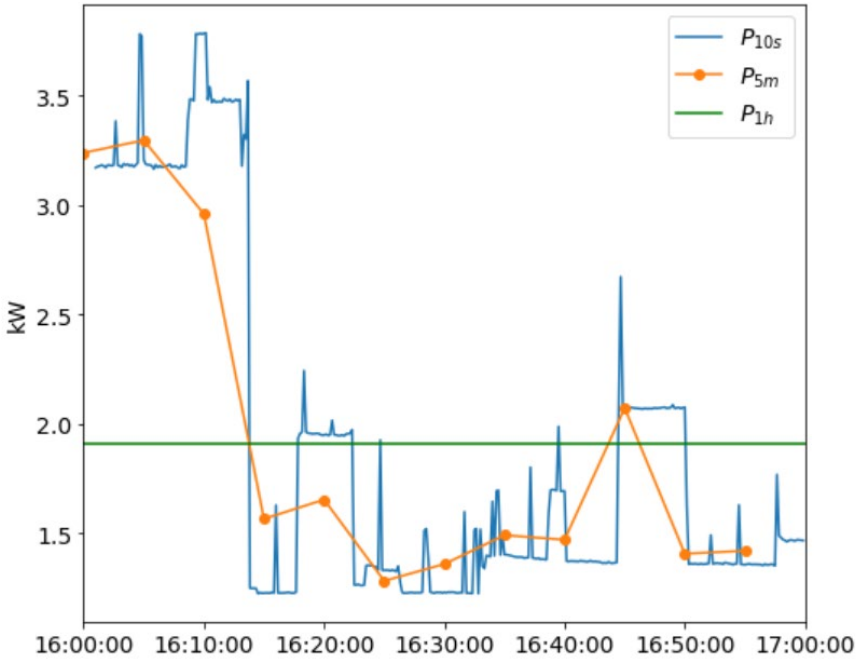
- Agder Energi skal innføre sikringsbasert tariffsystem, effekttopper kan da ha en innvirkning.

- Hvilken sikringsstørrelse burde kunden ha? Nedsikring?

«Mål for prosjekt»
- Teste hypoteser for effektvariasjon og finne mål på forskjell mellom timesverdi og høyeste effekt innenfor timen.

«Gevinster»

- Økt forståelse av effektvariasjon, grunnlag for mer presis dimensjonering av nett.



Sensitivity: Internal

Bakgrunn for prosjekt - Hypoteser

Hypotese 1:

Husholdnings- og fritidskunder har relativt lik effektvariasjon innenfor en time.

Hypotese 2:

Husholdnings- og næringskunder har relativt lik effektvariasjon innenfor en time.

Hypotese 3:

Det er ulik variasjon i høylast-, lavlast-, og normallasttimene

FoU-spørsmål

Hvilken verdi gir 10-sek effektdata ut over 5-min effektdata?

Sensitivity: Internal

Bakgrunn for prosjekt - Notasjon

P_{1h} = Gjennomsnittlig effektforbruk på en time, også kalt timesverdi

P_{5m} = Gjennomsnittlig effektforbruk på 5 minutter

P_{10s} = Gjennomsnittlig effektforbruk på 10 sekunder

$\mathbf{F_{5m}} = P_{5m}/P_{1h}$ = Forholdet mellom 5-minuttsverdi og tilhørende timesverdi

$\mathbf{F_{10s}} = P_{10s}/P_{1h}$ = Forholdet mellom 10-sekundsverdi og tilhørende timesverdi

Sensitivity: Internal

Bakgrunn for prosjekt – Datagrunnlag

Datagrunnlag 5-minuttsoppløsning (P_{5m}):

Datakilde: AMS-måler

Husholdning: 98

Fritid: 30

Næring: 59

Tidsperiode: Juni 2020-Juni 2021

Kundenummer	Tidspunkt	P_{5m} Effekt (kW)
1	2020-06-11 16:00:00	5.274
1	2020-06-11 16:05:00	5.328
1	2020-06-11 16:10:00	5.331
1	2020-06-11 16:15:00	5.354

Datagrunnlag 10-sekundsoppløsning (P_{10s}):

Datakilde: HAN-port / Tibber

Husholdning: 9

Tidsperiode: April 2021-Juni 2021

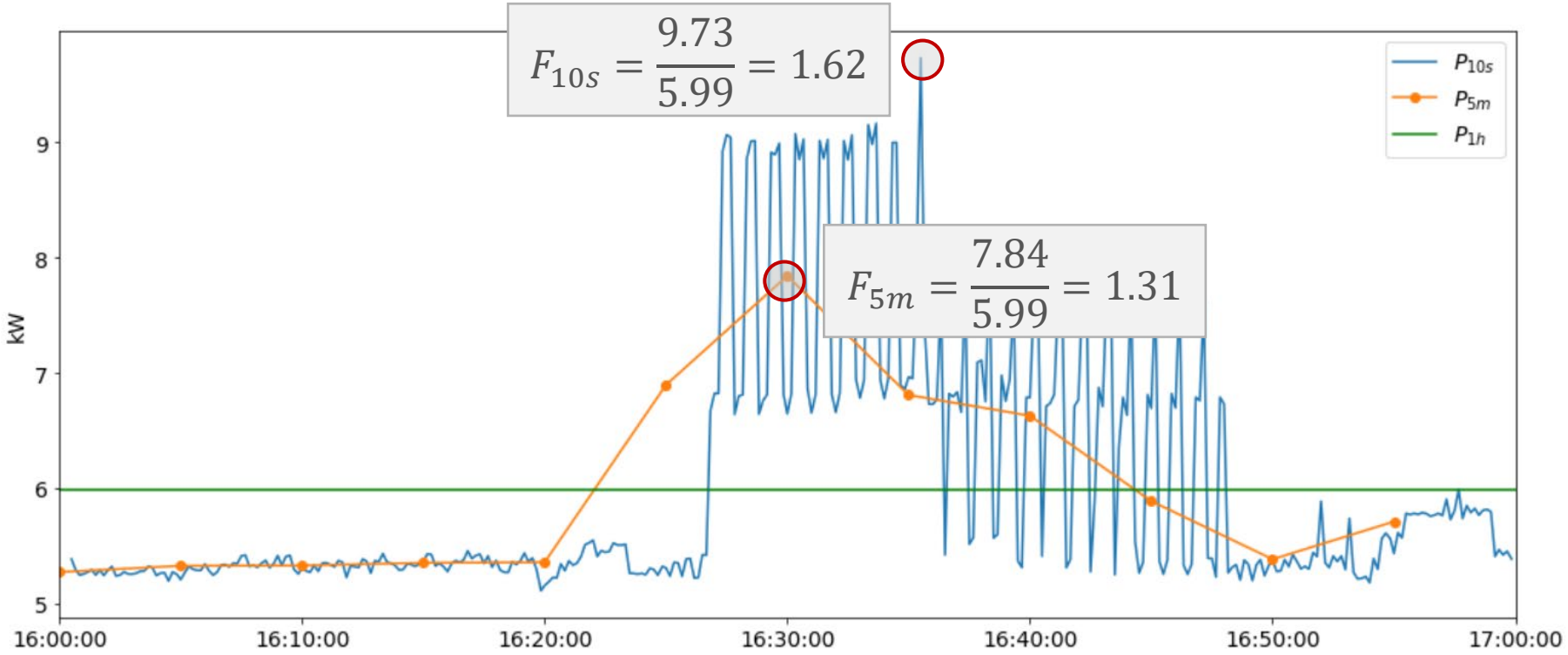
Kundenummer	Tidspunkt	P_{10s} Effekt (kW)
1	2020-06-11 16:00:00	5.387
1	2020-06-11 16:00:10	5.296
1	2020-06-11 16:00:20	5.248
1	2020-06-11 16:00:30	5.258

Merk: Datagrunnlagene er uavhengige (kunder i 10s er ikke i 5m)

Sensitivity: Internal

agder energi

Bakgrunn for prosjekt - Eksempel

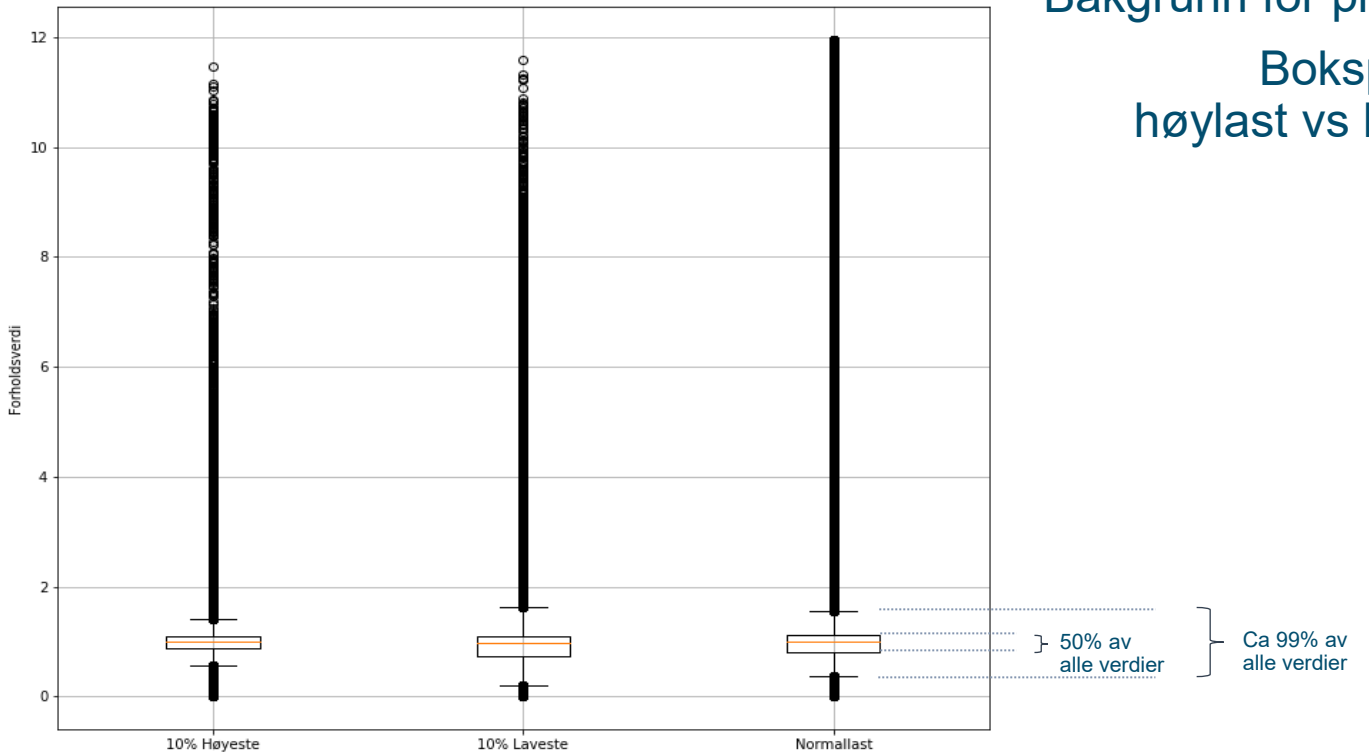


agder energi

Bakgrunn for prosjekt

Boksplott - høylast vs lavlast

$$F_{5m} = P_{5m}/P_{1h}$$



Høylast –10% største forbruksverdier for hver kunde

Lavlast– 10% laveste forbruksverdier for hver kunde

Normallast – Alle forbruksverdier

Sensitivity: Internal

Bakgrunn for prosjekt - Persentiler

«99. persentil»

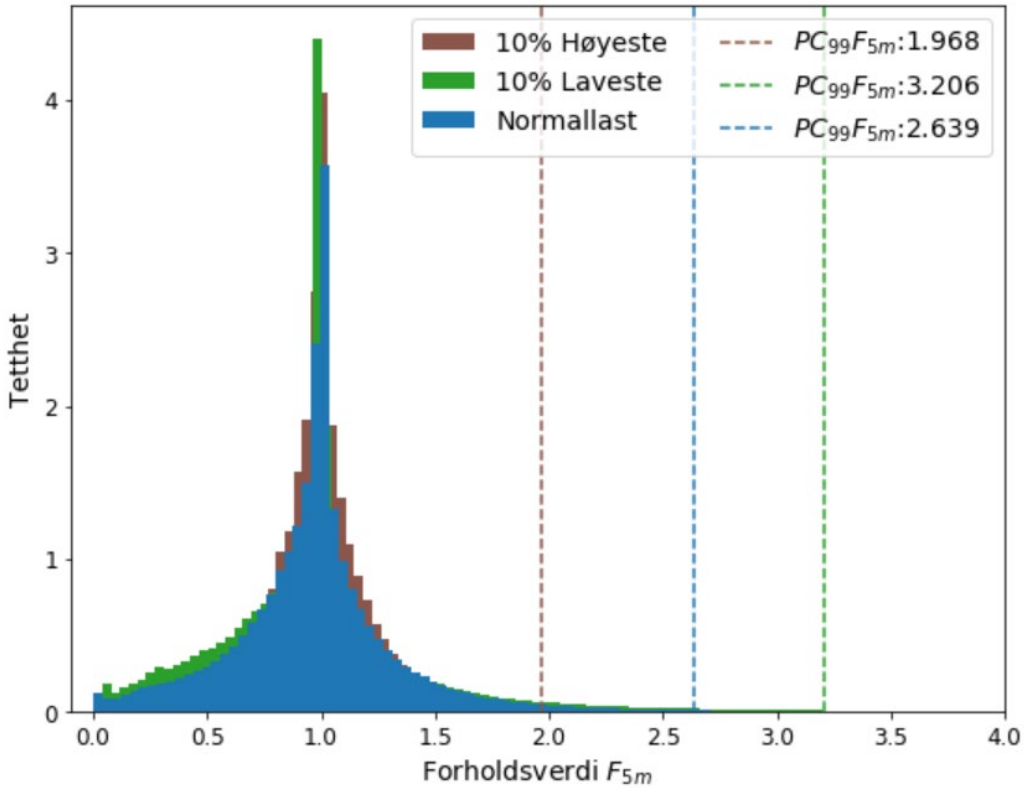
- Vi valgte å fokusere på 99. persentil når vi analyserte forholdsverdier

- 99. persentil er verdien som 99% av forholdsverdiene er mindre eller lik

«Notasjon»

- 99. persentil av forholdsverdier mellom 5-minuttsverdi og timesverdi: $PC_{99}F_{5m}$

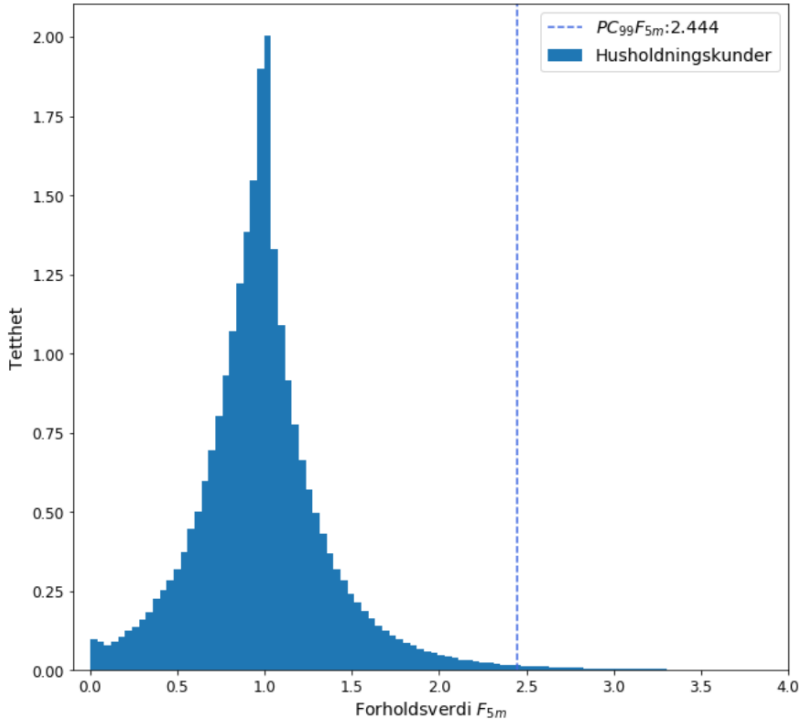
- 99. persentil av forholdsverdier mellom 10-sekundssverdi og timesverdi: $PC_{99}F_{10s}$

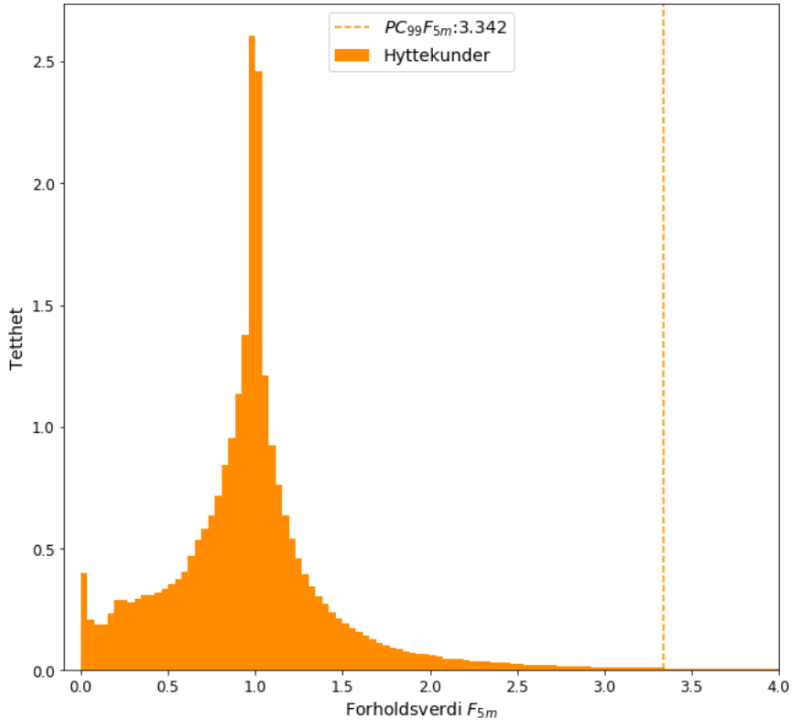


Sensitivity: Internal

Resultater - Hypotese 1

Stor forskjell i 99. persentiler grunnet mange ekstremverdier ->
Husholdningsforbruk og hytteforbruk har ulik effektvariasjon



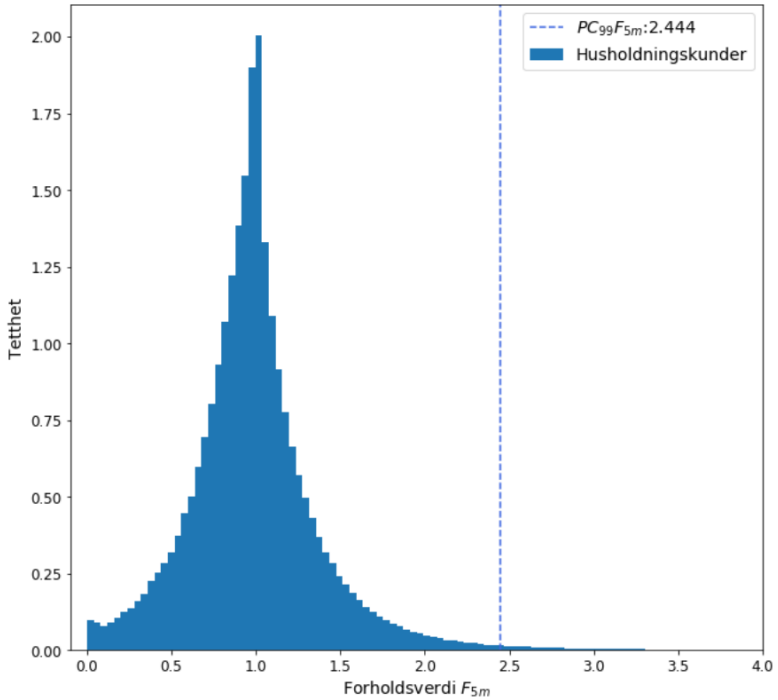


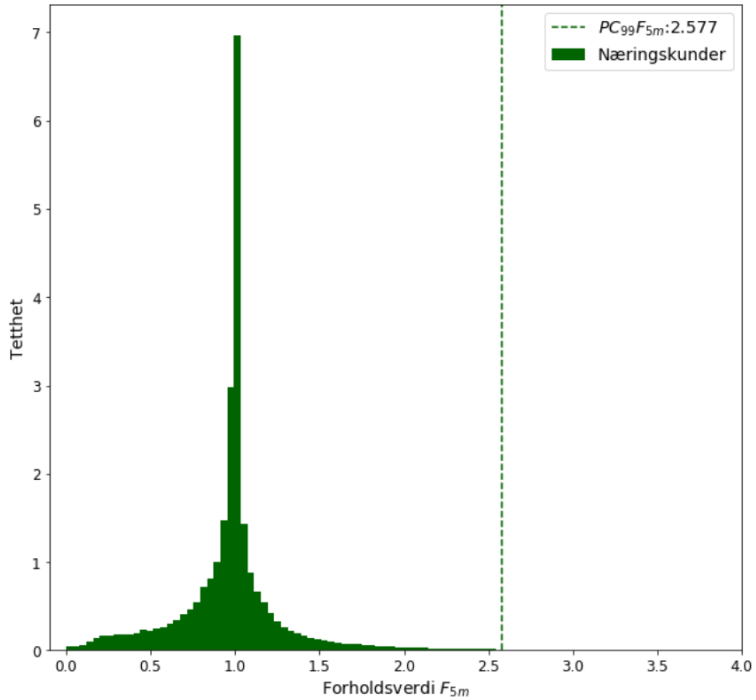
Sensitivity: Internal

Resultater – Hypotese 2

Husholdning og næring: Næringskundene har tilsynelatende mindre varians i fordelingen. 99. persentilene er tilnærmet like for F_{5m}

Husholdning og næring er like når det gjelder høye verdier innefor en time



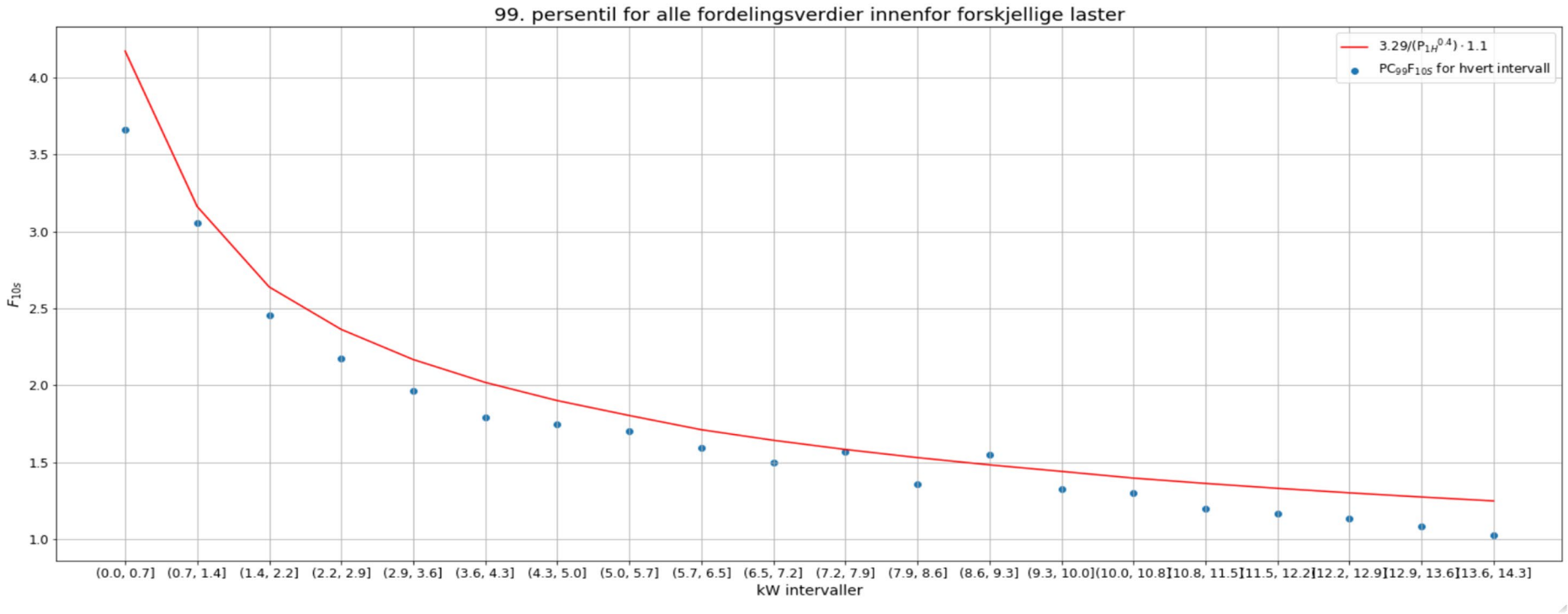


Resultater - Generelt

	Husholdning	Fritid	Næring
Antall Kunder	98	30	59
Høyeste P_{5m}	21.4827	15.3389	19.2111
Høyeste F_{5m}	11.9482	11.5315	11.9184
Varians i F_{5m}	0.1949	0.3582	0.2140
Standardavvik F_{5m}	0.4415	0.5985	0.4627
$PC_{99}F_{5m}$	2.444	3.342	2.577

Sensitivity: Internal

Resultater – 10-sekunds forholdsverdier i mer detalj → $PC_{99}F_{10s}$

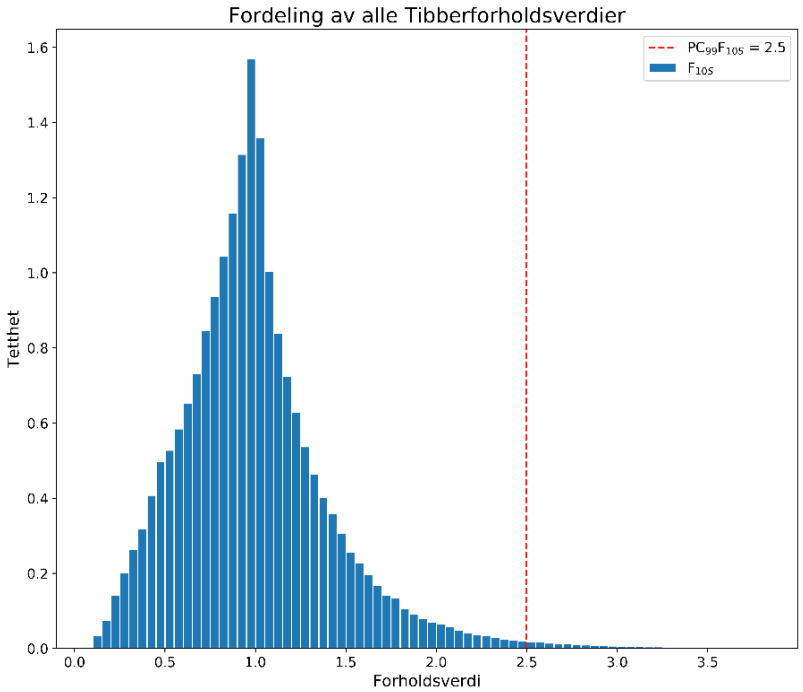


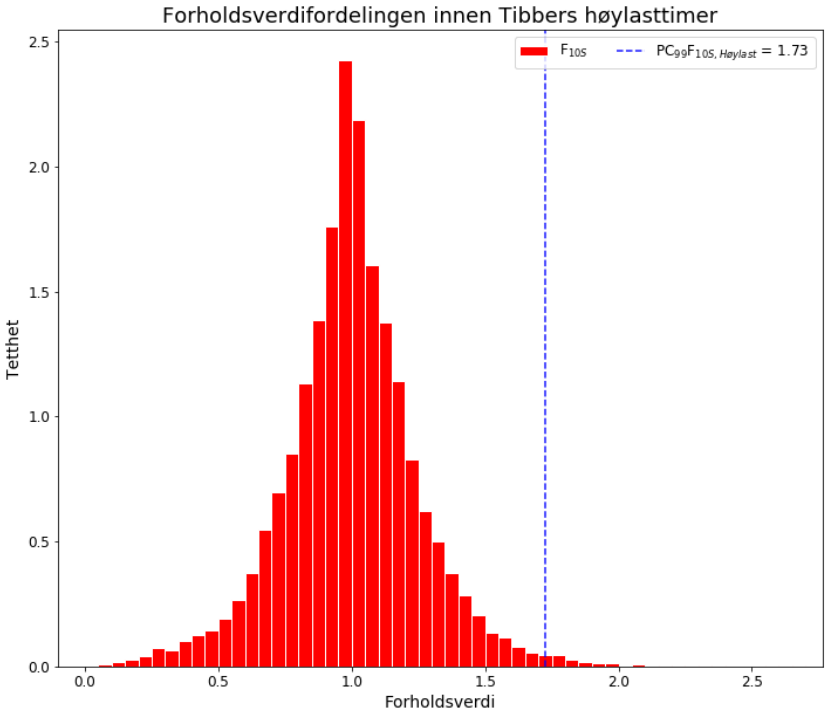
Sensitivity: Internal

agder energi

F_{10s}

Resultater – Normallast versus høylast-timer





Sensitivity: Internal

Oppsummering - Hypoteser

Hypotese 1:

Husholdnings- og fritidskunder har relativt lik effektvariasjon innenfor en time. → **NEI, mer variasjon hos fritidskunde**

Hypotese 2:

Husholdnings- og næringskunder har relativt lik effektvariasjon innenfor en time. → **JA**

Hypotese 3:

Det er ulik variasjon i høylast-, lavlast-, og normallasttimene →
JA, mindre variasjon ved høy last
Regel: Sikring hos kunde bør tåle 2x maks timesverdi

FoU-spørsmål

Hvilken verdi gir 10-sek effektdata ut over 5-min effektdata?

-> Ingen vesentlig ekstra verdi av 10-sek effektdata

($PC_{99}F_{10s}$ ~ $PC_{99}F_{5m}$ i høylast og normallast)

Sensitivity: Internal

Estimering av maksforbruk hos kunder i Agder Energi Nett via maskinlæringsmodeller og andre algoritmer

Julian Gjestvang Data Scientist (Trainee)

Innhold

1. **Grunnlag for prosjekt**
2. **Analyse og resultater**
3. **Oppsummering og veien videre**

agder energi

Grunnlag for prosjekt – Generelt

«Analyse av variasjon"
- Kunder har et unikt forbruksmønster som både er stokastisk og deterministisk.

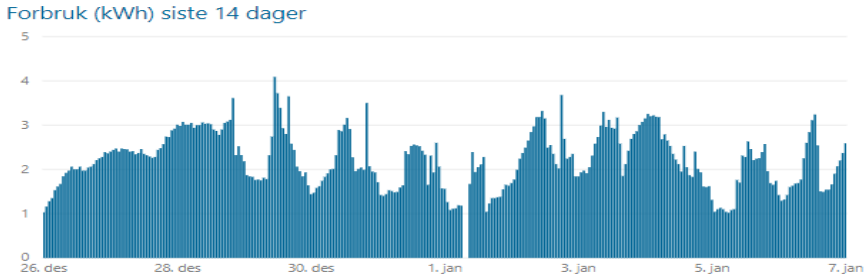
- Forbruksmønster er påvirket av temperatur, tidspunkt, helligdager etc.

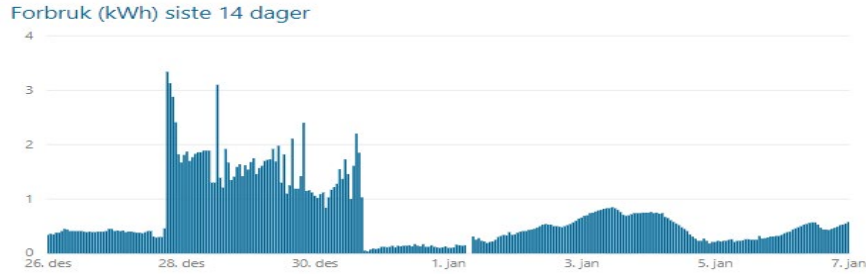
- Summen av kundenes forbruk kan overbelaste komponenter i nettet hvis kundene har et høyt og sammenfallende forbruk.

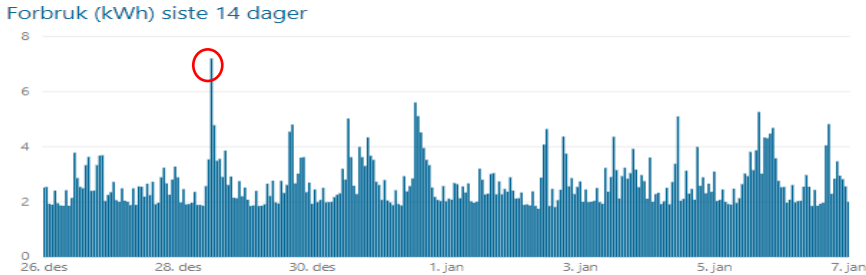
«Mål for prosjekt"
- Sammenligne ulike modeller for å estimere maks forbruk til kunder.

«Gevinster"

- Økt forståelse av forbruk, grunnlag for mer presis dimensjonering av nett – fører til optimal investering og bedre leveransekvalitet.







Sensitivity: Internal

Velanders formel

Per dags dato er Velanders formel brukt for å estimere både individuelle og gruppers maks forbruk. Gir en tidsuavhengig maks basert på data fra år med høyest sum forbruk. Er gitt med følgende:

$$\hat{P}_{i,y} = k_1 W_{i,y} + k_2 \sqrt{W_{i,y}}$$

hvor $W_{i,y}$ er totalt årsforbruk (summert) for kunde i for år y, og $k_1$og k_2 er konstanter som er standard og tilpasset kundesegment.

Styrker

Relativt enkel modell. Effektiv. God på homogene kundegrupper.

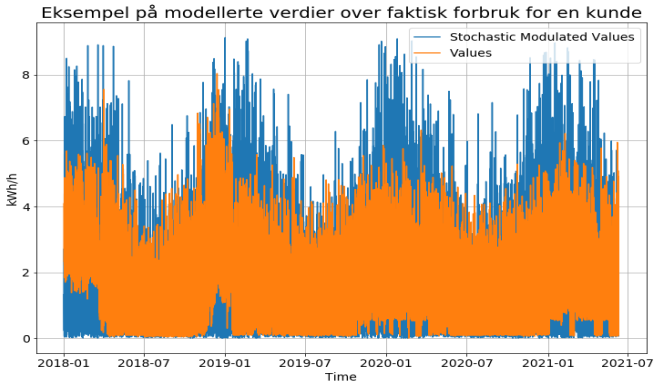
Svakheter

Deterministisk, fanger ikke opp stokastisk oppførsel. Lite dokumentasjon.

Sensitivity: Internal

Stokastisk Lastmodell

Sannsynlighetsmodell (Probabilistic) som baserer seg på alt av data tilgjengelig per kunde. Modellerer et generelt forbruk uavhengig av tidsrammer. Utgangspunkt i Erling Tønnes Phd oppgave.



Styrker

Gir estimat på totalt forbruksmønster per kunde. Plukker opp på individuell variasjon.

Svakheter

Krever mye kapasitet.

Sensitivity: Internal

Maskinlæringsmodeller

Plukker opp et signal i et datasett for å enten predikere eller si noe om forholdet mellom variabler i data. Prosessen som omhandler å plukke om et signal kalles for «læringsprosessen». Denne prosessen er enten *veiledet læring* (supervised) eller *ikke-veiledet læring* (unsupervised). Førstnevnte i dette prosjektet. Gitt data med en respons variabel Y og p forskjellige forklaringsvariabler $X_1, X_2, \dots, X_p$ så antar vi et forhold mellom Y og $X = (X_1, X_2, \dots, X_p)$ som kan skrives følgende:

$$Y = f(X) + \epsilon$$

hvor f er en funksjon av X og ϵ er tilfeldige feilledd med en forventningsverdi på 0 som antas og være uavhengig av X. Maskinlæringsmodeller ønsker å finne et estimat av $f(X)$.

Styrker

Mye informasjon og ressurser tilgjengelig. Bredt spekter av ulike modeller. Plukker opp komplekse mønster i data.

Svakheter

Kan kreve mye data. Blackbox. Sensitiv til endring i datastruktur.

Sensitivity: Internal

Grunnlag for prosjekt – Data

Datasett
- Består av registrert forbruk per time (ikke justert for temperatur) til 1000 husholdningskunder fra 2018-2020.

Forklaringsvariabler
- Standardavviket, gjennomsnittet og sum av forbruk per kunde.

- Basert på tidligere resultater og andre modeller, ønsket også en enkel modell.

Responsvariabel

- Størst registret maks per kunde.

UsagePointId	AvgValue	StdValue	SumValue	MaxValue
947899xxxxxx	3.15433	1.24578	95957.14	9.43
734738xxxxxx	2.21442	1.64248	79547.24	9.03
24889xxxxxx	6.22355	2.76582	201298.78	20.99
……	……	……	……	……

Sensitivity: Internal

agder energi

Grunnlag for prosjekt – Analyse av data

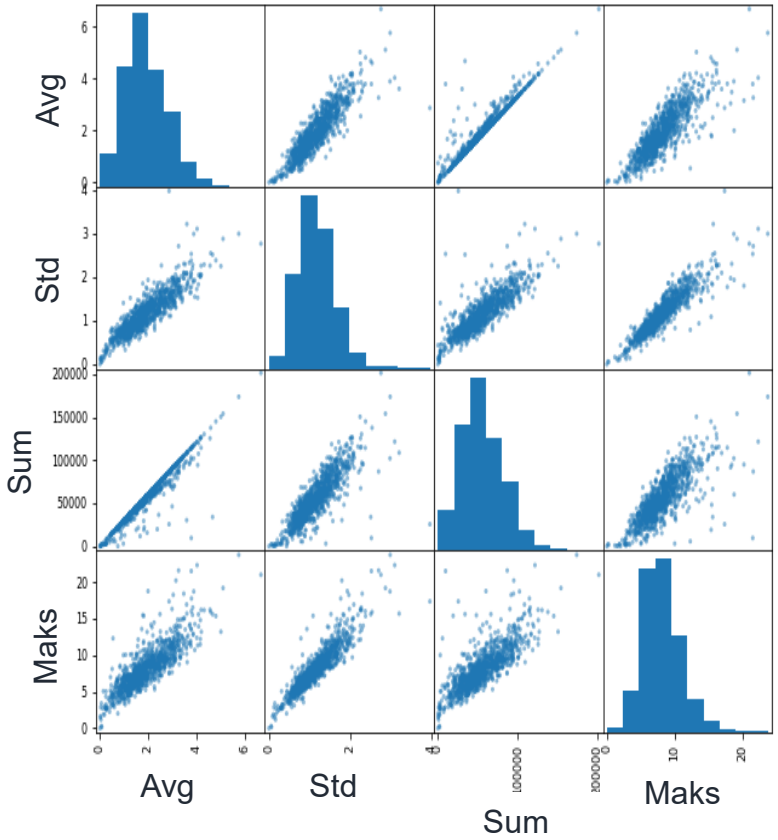
Vi ser at:

- Det er en positiv lineær sammenheng mellom observert maks og forklaringsvariablene.
- Standardavvik noe sterkest korrelert.
- Variablene ser normal fordelt ut.
- Indikerer at en lineær modell vil gjøre en god jobb.

Men :

- Naturen til variablene gir positivt lineært forhold.
- Multikollinearitet (hva driver maks?)

Spredningsplott



Sensitivity: Internal

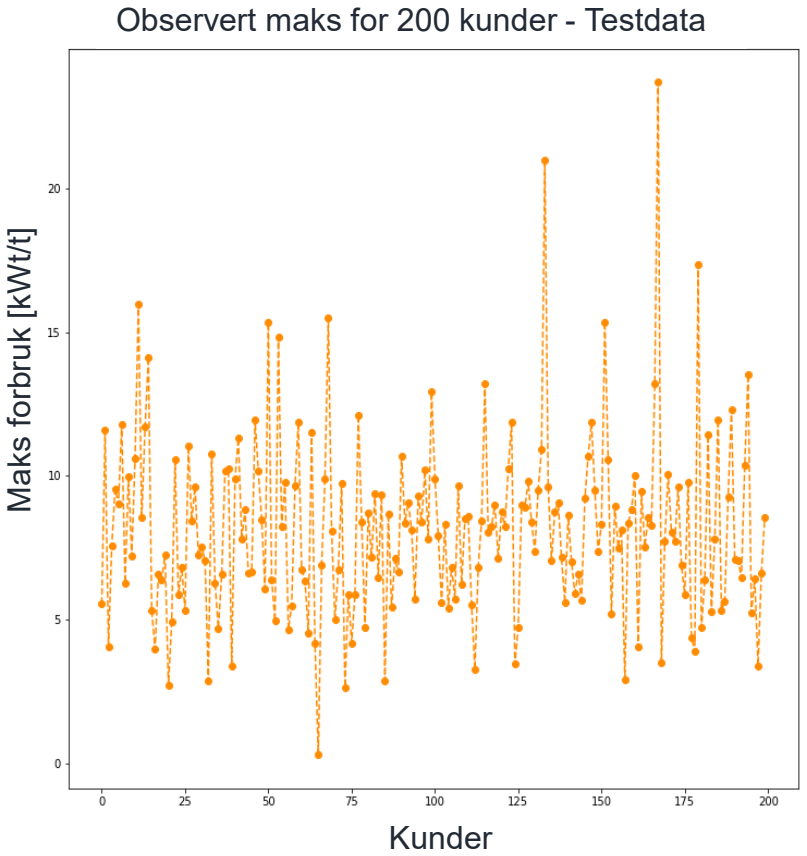
Analyse og resultater – Maskinæringsmodeller

Teknisk:

- Deler opp trening og test i 80/20.
- K-fold kryssvalidering på fem folder, gjentas 10 ganger. Gir et gjennomsnitt av feilleddene.
- Måler feilledd via Mean Squared Error = $\frac{1}{n}\sum_{i=0}^{n}(Y_i - \hat{Y}_i)^2$ og Mean Error = $\frac{1}{n}\sum_{i=0}^{n} Y_i - \hat{Y}_i$
- Grunninnstillinger på alle modeller.

Mål:

- Estimere kurven i plott, estimerer da maks per kunde.



Sensitivity: Internal

agder energi

Analyse og resultater – Historisk maks

Historisk maks:

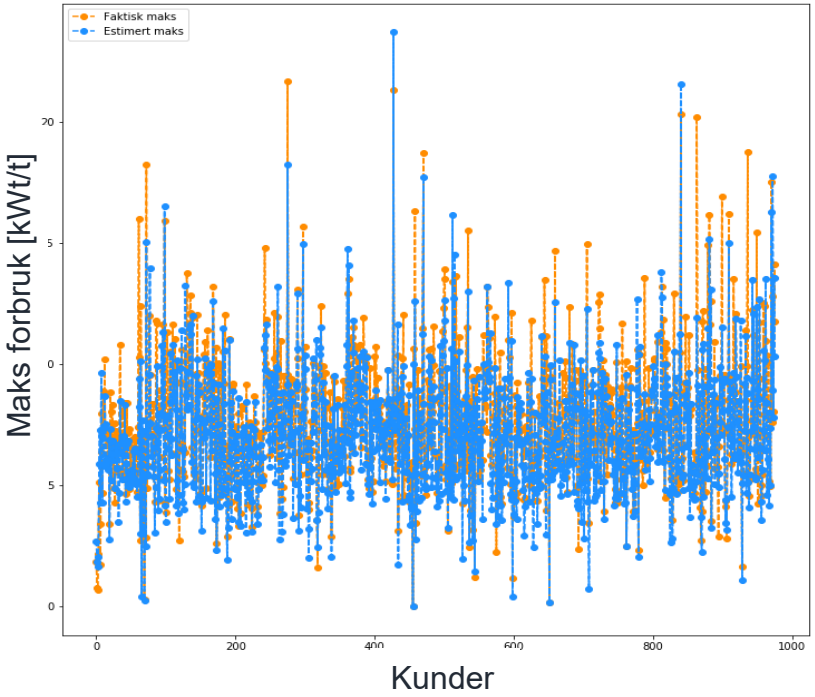
- Bruker maks observert 2018/2019 for hver kunde som estimat for maks 2020.
- Referanse modell

Resultat:

- Baserer seg på 977 kunder (mistet noen kunder pga. deling av data på tid).
- Ser ut til å underestimere.
- Andre modeller bør få en lavere MSE.

Mean Squared Error	Mean Error
2.090	0.608

Maks 2018-2019 for å estimere maks 2020



Sensitivity: Internal

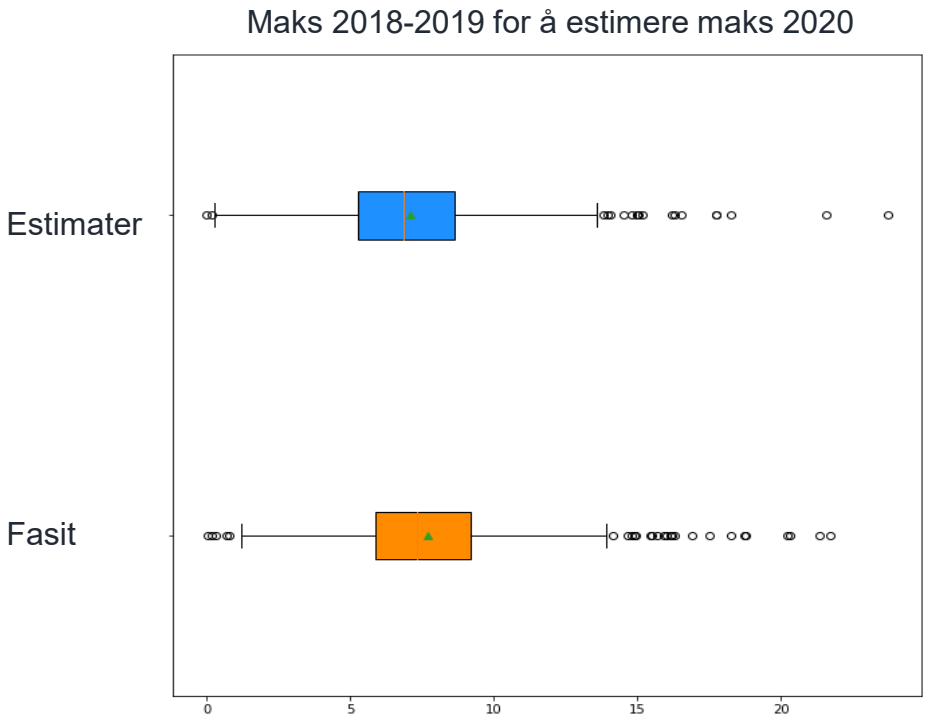
agder energi

Analyse og resultater – Historisk maks

Historisk maks- boks plott:
- Boks plott viser fordelingen til estimatene og faktiske verdier.
- Gjennomsnittet i grønn trekant.
- Basert på 977 kunder.

Resultat:
- Bekrefter at modellen underestimerer.

Mean Squared Error	Mean Error
2.090	0.608



Sensitivity: Internal

Analyse og resultater – Maskinæringsmodeller

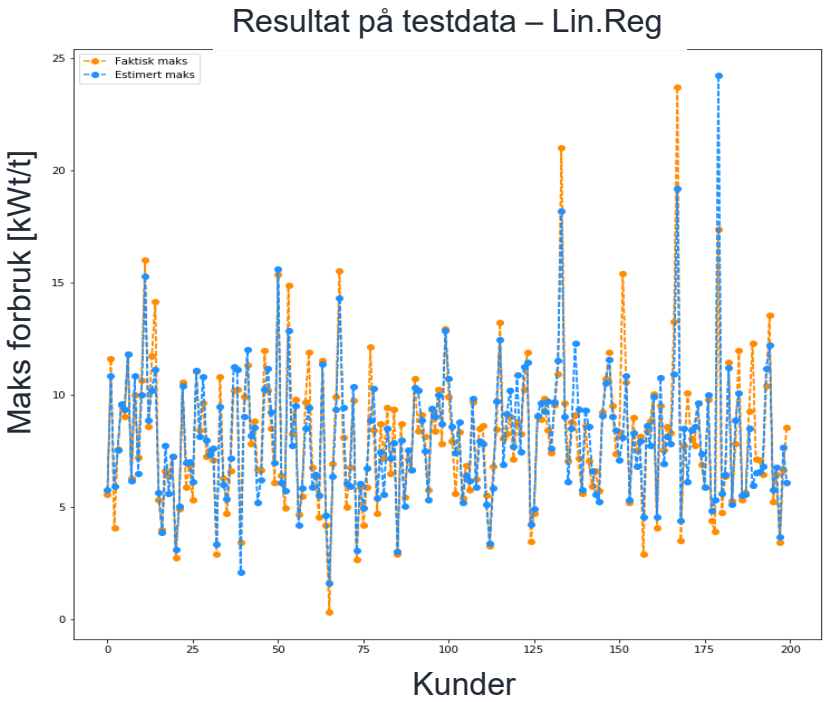
Lineær regresjon:

- Minimerer kvadrert sum mellom estimat og fasit (OLS = Ordinary least square method).
- Parameter innstilt på treningsdata.

Resultat:

- Basert på test data med 200 kunder.
- God jobb sammenlignet med grunnmodell.
- Analyse viste positivt lineært forhold.

Mean Squared Error	Mean Error
1.79	0.065

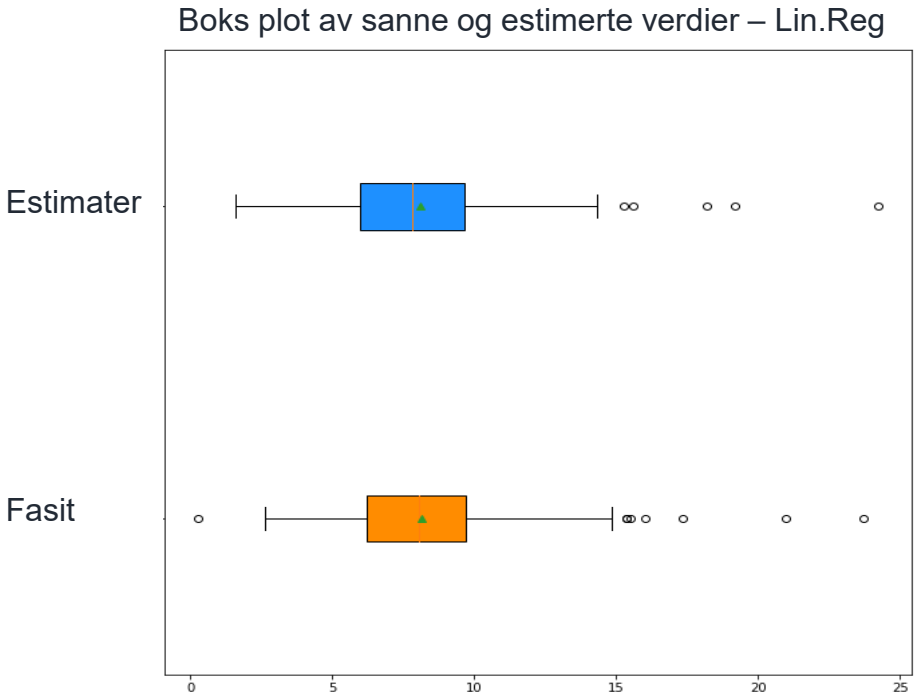


Sensitivity: Internal

Analyse og resultater – Maskinæringsmodeller

Lineær regresjon - boks plott:
- Boks plott viser fordelingen til estimatene og faktiske verdier.
- Gjennomsnittet i grønn trekant.
- Basert på testdata.

Mean Squared Error	Mean Error
1.579	-0.0003



Resultat:
- Basert på kryssvalidering på hele datasettet.
- Estimerer faktisk fordeling på en god måte.
- Lavere MSE sammenlignet med historisk modell.

Sensitivity: Internal

Analyse og resultater – Maskinæringsmodeller

En lineær regresjonsmodell for dette problemet er gitt ved følgende:

$$\hat{Y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$$

Hvor $\hat{Y}$ er estimat for maks, θ_0 er et bias ledd, θ_i er koeffisienter og x_i er forklaringsvariabler. Ved å bruke Lin.Reg.OLS på 800 kunder fikk vi følgende:

$$\hat{Y} = 1.309 + 0.495x_1 + 5.440x_2 - 0.00000743x_3$$

Hvor:

x_1 er gjennomsnitt
x_2 er standardavviket
x_3 er sum av forbruket til en kunde.

Sensitivity: Internal

agder energi

Analyse og resultater – Velanders

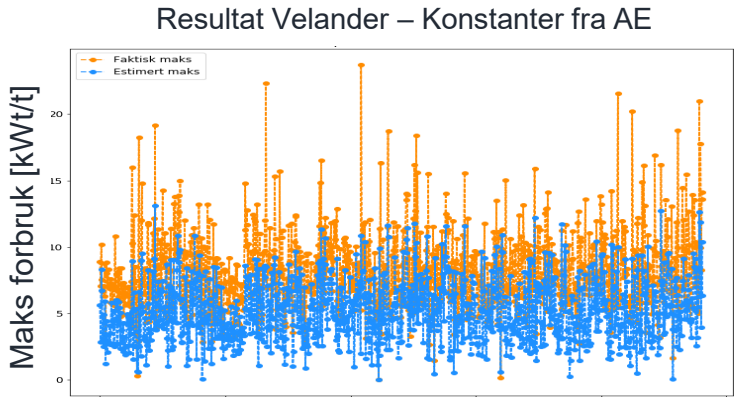
Velanders:

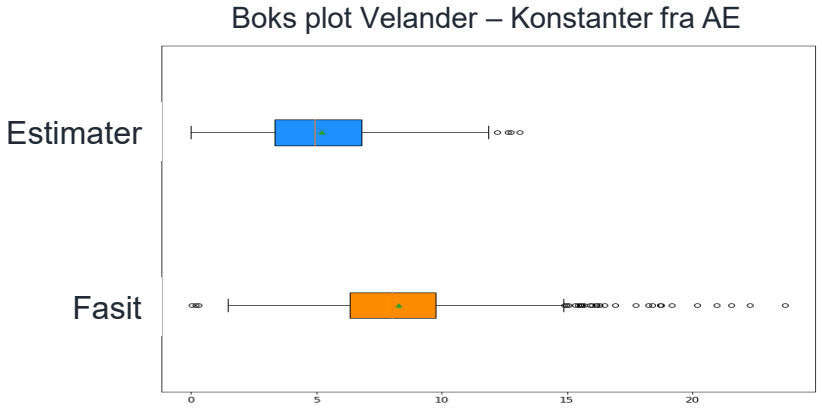
- Bruker konstanter gitt fra Agder Energi for husholdning.
- k_1 = 0.00021 og $k_2 = 0.019$

Resultat:

- Baserer seg på 977 kunder.
- Underestimerer.
- Høy MSE sammenlignet med Naiv.

Mean Squared Error	Mean Error
13.46	3.07





Sensitivity: Internal

Analyse og resultater – Velanders

Velanders:

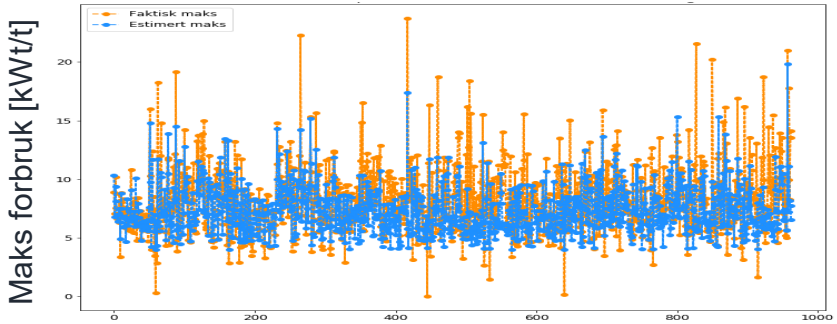
- Bruker konstanter fra en multiple regresjon med sum, kvadrert sum og observert maks for 800 kunder.
- $k_0 = 4.1650$ $k_1 = 0.0003$ og $k_2 = -0.0138$

Resultat:

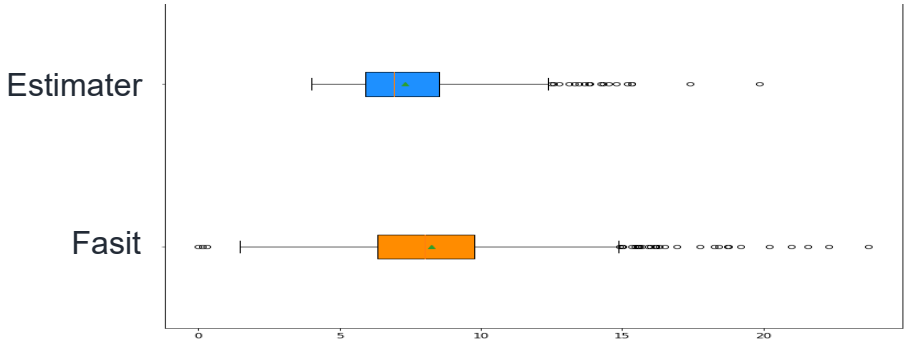
- Baserer seg på 977 kunder.
- Underestimerer.
- Høy MSE sammenlignet med historisk modell.
- Lavere MSE sammenlignet med modell med konstanter fra AE.

Mean Squared Error	Mean Error
5.44	0.93

Resultat Velander – Konstanter fra Lin.Reg



Boks plot Velander – Konstanter fra Lin.Reg



Analyse og resultater – Resultat alle modeller

Resultat:

- Basere seg på kryssvalidering.
- Vanlig Lin.Reg modell ser ut til å få lavest MSE.
- OLS å fortrekke, enklest.
- Avanserte maskinlæringsmodeller ser ut til å få høyere MSE. Mulig overtrener.
- Ulike modeller, men utgangspunkt for å sammenligne.
- Alle ml modellene vektet standardavviket høyest.

Modell	MSE	ME
Historisk maks	2.090	0.608
Lin.Reg.OLS	1.579	-0.0003
Lin.Reg.SGD	1.567	0.0001
Beslutnings tre	2.94	0.00136
CATBoost	1.812	0.007
XGBoost	2.176	-0.004
Velanders m.konstanter AE	13.46	3.07
Velanders m.konstanter Lin.Reg	5.44	0.93
Stok.Last	8.523	-2.228

Lin.Reg	Lineær regresjon
ME	Mean Error
MSE	Mean Square Error
OLS	Ordinary least square
SGD	Stochastic gradient descent
XGBoost	Extreme Gradient Boost

Sensitivity: Internal

Analyse og resultater – Andre resultater

Fjerning av registret maks i datasett

Maks forbruk ligger implisitt i datasett (standardavviket blir påvirket av denne verdien f.eks). Fjernet denne fra datasettet for deretter å kjøre gjennom

Teste modell på data fremover i tid

Ønsket å teste hvor generaliserbar modellene var fremover i tid. For å teste dette splittet vi datasettet opp i 2018-19 og 2020. Vi trente og testet modellene på data fra 18-19 for deretter å måle hvor bra estimatene gjorde det på observert maks for samme kunder i 2020.

Sensitivity: Internal

Oppsummering og veien videre

«En enkel modell er muligens nok "

- Lineær regresjon med OLS gir best resultat for å estimere maks per kunde kunde
- Mer kompliserte maskinlærings modeller får høyere feil verdier
- Velanders har en tendens til å underestimere, og man oppnår bedre estimater ved finne konstanter via lineær regresjon versus standard koeffisienter.
- Stokastisk last modell har en tendens til å overestimere maks per kunde.

«Veien videre og forslag til arbeid "

- Endre mål på feilledd, modeller som treffer høyere eller likt fremtidig maks vektes høyere.
- Bruke maskinlære til å estimere hele forbruksmønster per kunde.
- Finne optimal modell for å estimere en maksverdi for grupper av kunder.
- Se på andre kundegrupper (fritids og næringskunder).

Sensitivity: Internal


agder energi
God kraft. Godt klima.

Prosjektnotat

TITTEL			
Pilot Effektanalyse: Estimering av maksforbruk hos husholdningskunder i Agder Energi Nett via maskinlæringsmodeller og andre algoritmer			
WORK PACKAGE	**VERSJON**	**DATO**	**ANTALL SIDER**
WP Pilot	1.0	2021-08-20	27
FORFATTER(E)		**WP-LEDER**	**GRADERING**
Julian Gjestvang Per-Oddvar Osland (Oct 8, 2024 13:27 GMT+2)		Maren Istad Maren Istad (Oct 9, 2024 07:35 GMT+2)	Åpen

DISTRIBUSJON		
CINELDI		

SAMMENDRAG

Ved å ta i bruk et datasett bestående av 1000 husholdningskunder så vi at en Lineær regresjons modell som baserte seg på minste kvadraters metode kunne estimere en generell maks verdi per kunde med gode resultater. Andre maskinlæringsmodeller og modeller ble også anvendt på samme datasett, men med svakere resultater. Videre så vi at Velanders formel hadde en tendens til å underestimere maks per kunde, men at man kunne oppnå drastisk bedre estimater ved å bruke koeffisienter fra en multiple lineær regresjon basert på sum, kvadrert sum og observert maks. Vi så også at stokastisk lastmodell hadde en tendens til å overestimere maks per kunde.

Innholdsfortegnelse

1 Bakgrunnsinformasjon om pilotprosjektet

Tabell 1: Bakgrunnsinformasjon

	Fra malen "planlegging av pilotprosjekt"	Viktige endringer i løpet av pilotperioden
Målsetting	Estimere maks forbruk per husholdningskunde	
Problemstilling	Finne optimal modell for denne målsettingen	
Aktiviteter	Innlegg på webinar for CINELDI	
Kostnadsestimat	Ca. 250 t	
Innovasjonspotensial	Mer effektiv modell for dimensjonering av nett bl.a.	
Forventet resultat	NA	
Tidsplan	Ca. 6 måneder	

2 Om Piloten og fysisk pilotområde

Tabell 2: Piloten og pilotområdet

Pilotområdet	FoU-Avansert analyse-Ai-Effektanalyse-maks effektuttak
Måledata og andre data som samles inn og lagres fra Piloten	NA
Personvern og/eller kraftsensitiv informasjon	Ja
Måle- og kommunikasjons-infrastruktur	NA
Use-case-beskrivelser og testplaner	NA
Regulering og forskrifter	NA
Barrierer og løsninger	NA
Hvem skal eventuelt ta resultater fra Piloten i bruk?	Ansvarlige for dimensjonering av nett
Hvem er erfaringene relevant for?	Avdeling analyse og avdeling nettstrategi
Hva påvirkes av resultater fra Piloter?	Modeller for estimering av maks effektuttak per kunde
Informasjonsdeling mellom aktørene før/underveis/etterpå	Ja

Er det laget planer for videreføring? Skalering/fullskala implementering?	Ja, kompetanseoverføring fra Data Scientist som hadde ansvar for prosjektet til andre data scientister i analyseavdeling. Eventuelle implementeringer i produksjon må skje på sikt.

3 Resultater og innovasjoner fra Piloten

Per dags dato bruker man to metoder for å beregne størst effektuttak per kunde, Brukstid og Velanders formel. Førstnevnte brukes som oftest på fritidskunder, og sistnevnte på både husholdningskunder og næringskunder. I dette prosjektet har vi fokusert på Velanders formel og husholdningskunder hvor vi har sett om det eventuelt finnes andre modeller som er mer effektive. Vi så da på en modell fra Erling Tønnes sin PhD oppgave, Stokastisk lastmodell, men også ulike Maskinlærings modeller. En av hovedfunnene var at en lineær regresjon basert på minste kvadraters metode oppnådde best resultater. Denne modellen baserte seg på standardavviket, gjennomsnittet og 99-percentilen på registret forbruk per kunde. Dette funnet baserte seg på at vi målte estimatene til hver modell via ulike feil metrikker hvor *mean squared error*, som måler variasjonene til feilleddene, ble vektlagt høyest. Videre så vi at Velanders formel hadde en tendens til å *underestimere* maks effekt per kunde, men at man kan forbedrede modellen ved å hente ut Velanders koeffisienten via en multiple lineær regresjon på sum, kvadrert sum og observert maks per kunde. Vi så også at stokastisk lastmodell hadde en tendens til å *overestimere* maks effekt per kunde, selv når det ble gitt et øvre tak som baserte seg på en teoretisk maks per kunde. Lineær regresjon verken overestimerte eller underestimerte noe særlig og fikk i tillegg den laveste registeret MSE verdien. Resultatene ble også dokumentert i en Power Bi rapport.

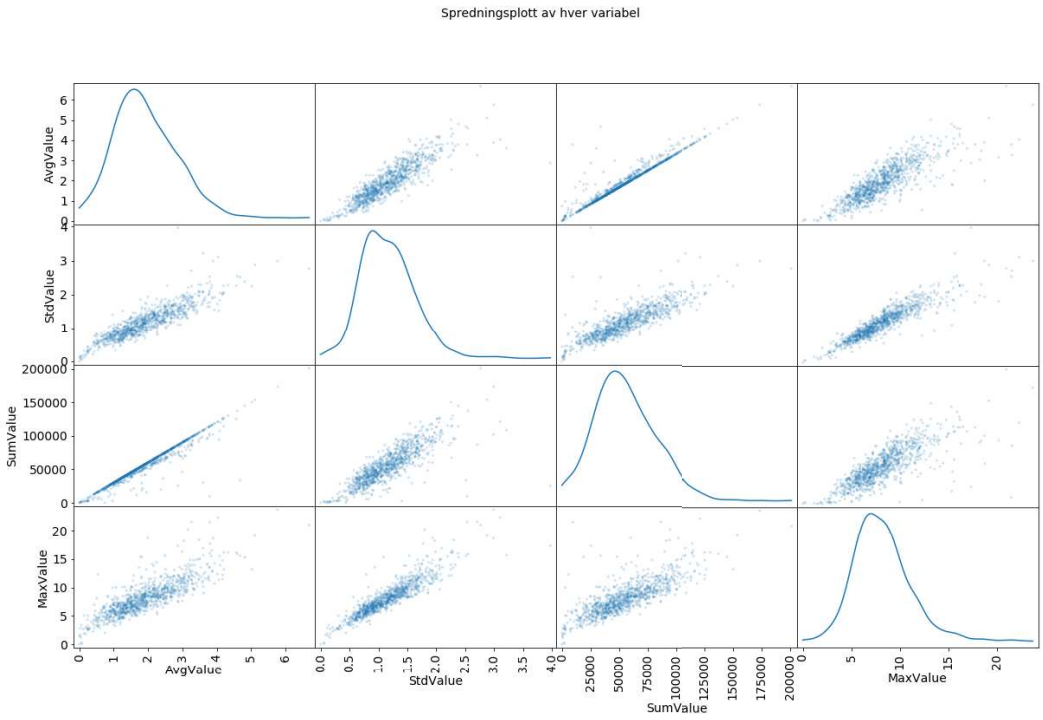
3.1 Datasett - Statistisk analyse

Opprinnelig datasett bestod av 1000 husholdningskunder med registrert forbruk fra 2018 til 2020 (for majoriteten av kundene). Forbruket ble ikke justert for temperatur variasjon. Ved å bruke SQL spørring i DataBricks hentet vi ut gjennomsnittet (AvgValue), standardavviket (StdValue), sum (SumValue) og høyest registret verdi per kunde (MaxValue). Dette datasettet ble da utgangspunktet for maskinlærings modellene i prosjektet. Variablene ble plukket ut fordi vi så i et tidligere stadiet av prosjektet at variasjon av forbruk hang sammen med maks forbruket per kunde. Men, vi tenkte også at det ga mening å inkludere sum forbruk da Velanders formel bruker denne. Vi endte opp med følgende datasett:

Tabell 3 Datasett

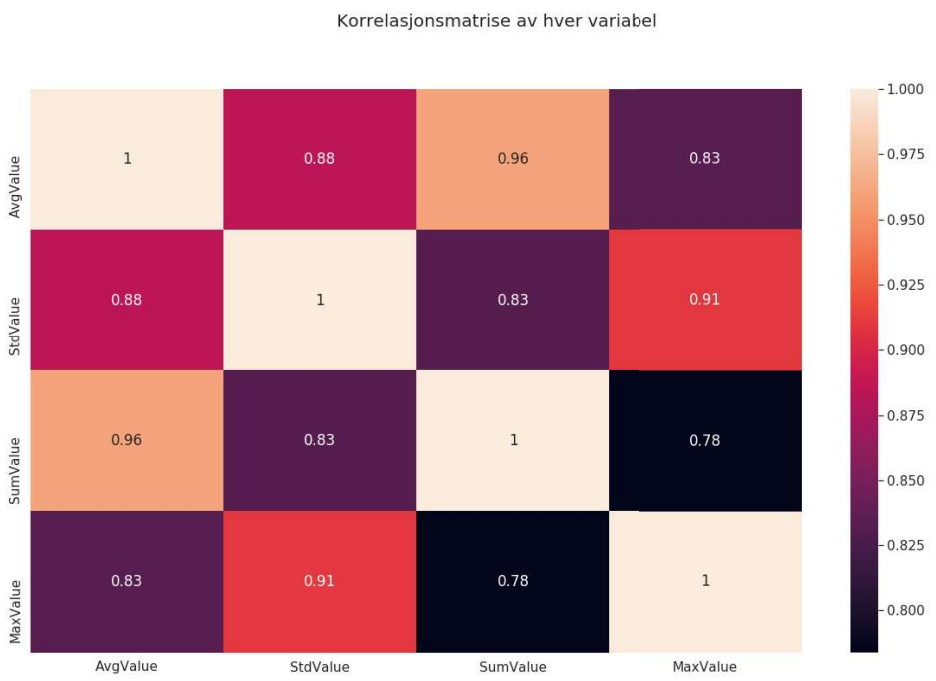
UsagePointId	AvgValue	StdValue	SumValue	MaxValue
947899xxxxxx	3.15433	1.24578	95957.14	9.43
734738xxxxxx	2.21442	1.64248	79547.24	9.03
24889xxxxxx	6.22355	2.76582	201298.78	20.99
……	……	……	……	……

Før vi anvendte modeller på datasettet gjorde vi en statisk analyse for å se etter korrelasjoner og tendenser i datasettet. Denne analysen gir oss et godt bilde på hvordan forbruket til kundene henger sammen, men også hvilke modeller som er hensiktsmessige å bruke. Vi lagde et spredningsplott:



Figur 1 Spredningsplott alle variabler

Spredningsplottet viser oss at observert maks forbruk per kunde har en *positiv lineær sammenheng* med forklaringsvariablene. Den sterkeste sammenhengende ser ut til å være med standardavviket til kundene. Ellers ser vi også at variablene er sterkt positivt korrelert med hverandre. Dette var noe å forventet da variablene har en naturlig positiv sammenheng. For eksempel gir høye sum verdier høy gjennomsnittsverdier. Videre ser vi at variablene også ser ut til å være relativt normalfordelt. En statistisk utfordring med dette datasettet er et fenomen som heter *multikollinearitet*. Det oppstår når det er en sterk lineær sammenheng mellom flere forklaringsvariabler og gir en usikkerhet ift. hva som faktisk driver responsvariablen vår (maks forbruk i dette prosjektet). Fenomenet påvirker ikke modell estimatene. Vi henter ut korrelasjonsmatrisen:



Figur 2 Korrelasjon matrise alle variabler

Matrisen bekrefter hva vi så i spredningsplottet og viser at standardavviket korrelere høyest med observert maks verdi (0.91), deretter følger gjennomsnittet (0.83) og sum av forbruk (0.78). Resultatene vi har sett til nå indikerer at en lineær modell som tar et utgangspunkt i en normalfordelt data vil gjøre en god jobb ved å estimere maks effekt per kunde.

3.2 Maskinlæringsmodeller

En maskinlæringsmodell er i korte trekk en modell som plukker opp et signal i et datasett for å enten predikere eller si noe om forholdet mellom variablene i data. Prosessen som omhandler å plukke om et signal kalles for *læringsprosessen*. Denne prosessen er enten *veiledet* (supervised) eller *ikke-veiledet* (unsupervised). Førstnevnte ble brukt i dette prosjektet. Rent statistisk er det definert slik at gitt data med en respons variabel Y og p forskjellige forklaringsvariabler $X_1, X_2, \dots, X_p$ så antar vi et forhold mellom Y og $X = (X_1, X_2, \dots, X_p)$ som kan skrives på følgende form:

$$Y = f(X) + \epsilon$$

hvor f er en funksjon av X og ϵ er tilfeldige feilledd med en forventningsverdi på 0 som antas å være uavhengig av X. Maskinlæringsmodeller ønsker å finne et estimat av $f(X)$.

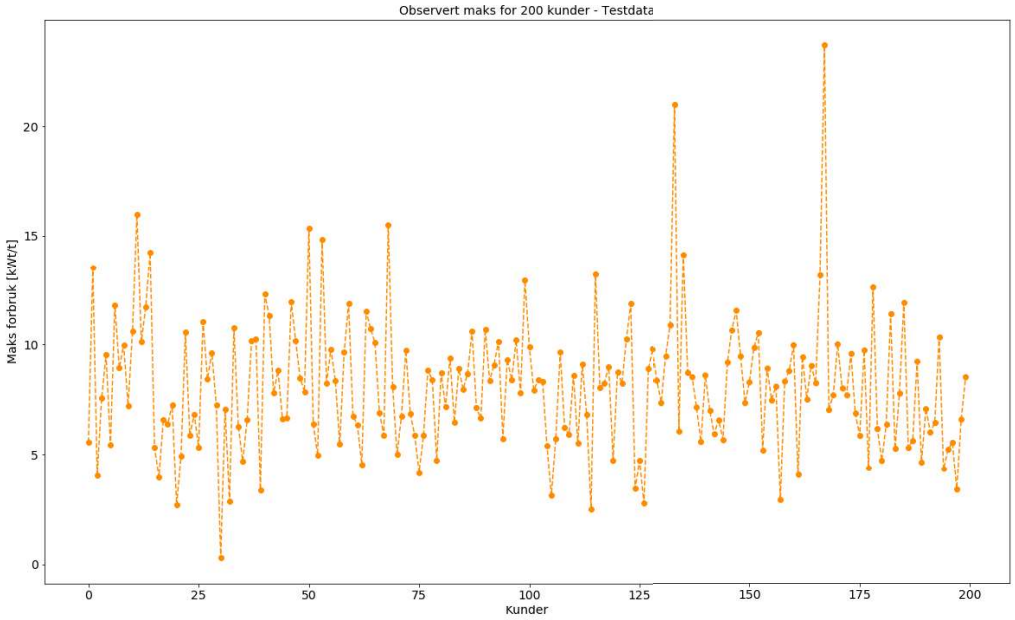
I en slik prosess er det vanlig å dele opp datasettet i trening og test sett for å måle prestasjonen til modellen på usett data. Man ønsker en modell som gir gode resultater på nye datasett. I dette prosjektet valgte vi å trene modellene på *800 kunder for så å teste det på 200 kunder.* Videre anvendte vi en *k-fold kryssvaliderings algoritme* på hver modell. Denne algoritmen gir oss gjennomsnitt av prestasjonen til modellene. Vi brukte 5 folder hvor hele prosessen ble gjentatt 10 ganger. Både resultatene fra test og kryssvalideringen ble tatt med i prosjektet, men sistnevnte ble vektet tyngst. Vi målte prestasjonen til alle modellene i prosjektet, både maskinlærings modellene og de andre, med *Mean Squared Error*:

$$\frac{1}{\mathrm{n}}\sum_{i=0}^{n}(Y_i - \hat{Y}_i)^2$$

hvor Y er observert maks verdi og $\hat{Y}$ er estimert maksverdi for samme kunde. MSE blir da et mål på variasjonen på feilen i estimatene våre. Videre ble også *Mean Error* brukt:

$$\frac{1}{n}\sum_{i=0}^{n} Y_i - \hat{Y}_i$$

som et utgangspunkt for å se om modellene overestimerer eller underestimerer. Men, denne ble også brukt forsiktig da like feilledd med ulike fortegn kan utjevne hverandre. Da dette er et prosjekt som baserer seg på veiledet læring endte vi opp med la maskinlærings modellene lære seg en kurve bestående av observert maks per kunde, dette blir da også det modellene prøver å estimere. Det ser slikt ut:



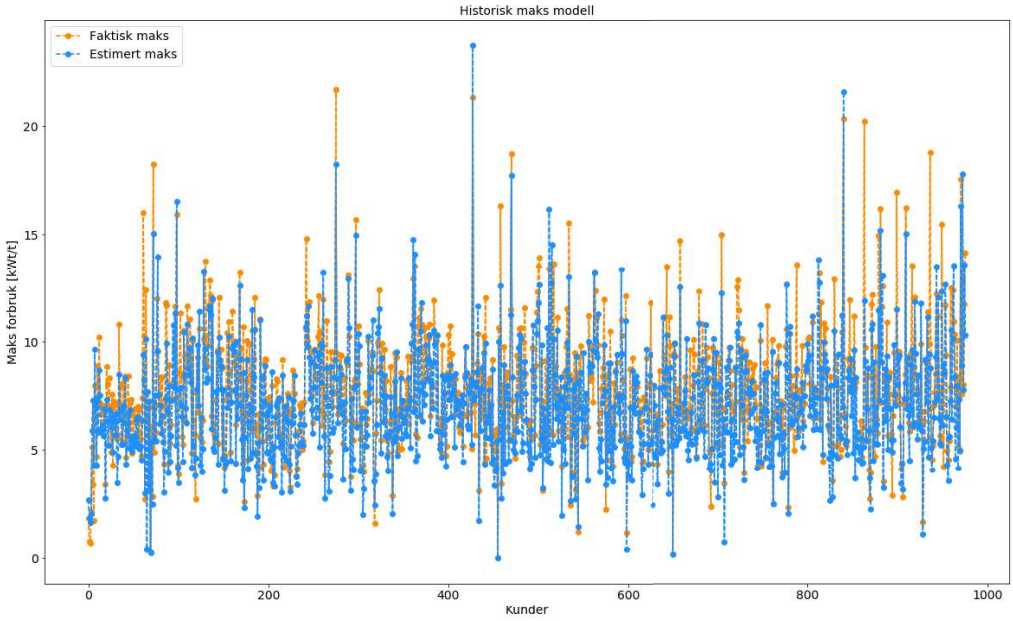
Figur 3 Kurver med observert maks for 200 kunder

3.2.1 Referansemodell

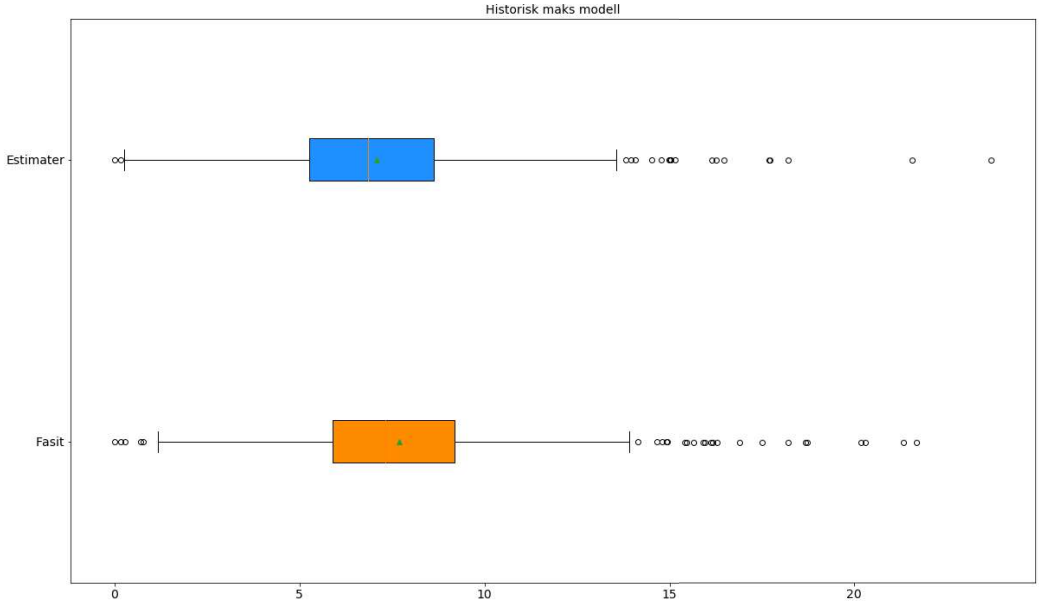
I prosjekter hvor man ønsker å finne en optimal modell for et gitt problem er det hensiktsmessig å ha en *referansemodell* som er relativt lite komplisert. Tanken er at mer kompliserte modeller bør prestere bedre sammenlignet med denne og vi ser i tillegg om det er et signal i datasettet for maskinlæringsmodellene å plukke opp. I dette prosjektet har vi valgt å bruke *historiske observerte maks verdier* som estimat for fremtidige maksverdier per kunde. Vi kaller denne modellen historisk maks og bruker data fra 2018 og 2019 til å estimere maks i 2020. Vi fikk følgende resultater:

Tabell 4 Feilmål historisk maks modell

MSE	ME
2.090	0.608



Figur 4 Historisk maks modell - Estimater



Figur 5 Historisk maks modell – Boks plott av estimater og fasit

Figur 4 og figur 5 viser estimatene fra modellen i blått og faktiske verdier i Orange. Plottene og den positive ME verdien viser oss at historisk maks har en tendens til å underestimere fremtidige maks

verdier. Det er verdt å merke seg at resultatene fra denne modellen vil være avhengig av *hvilket år* man velger hente estimatene sine fra. Et år med høyt gjennomsnitts forbruk vil f.eks. gjøre det bedre med fremtidige estimater sammenlignet med et år med lavt gjennomsnitts forbruk.

3.2.2 Lineær regresjon

Multippel lineær regresjon er en modell hvor man antar at det er et tilnærmet lineært forhold mellom respons variable Y og forklaringsvariabel $X = (X_1, X_2, \dots, X_p)$ slik at man kan finnet et estimat ŷ for Y:

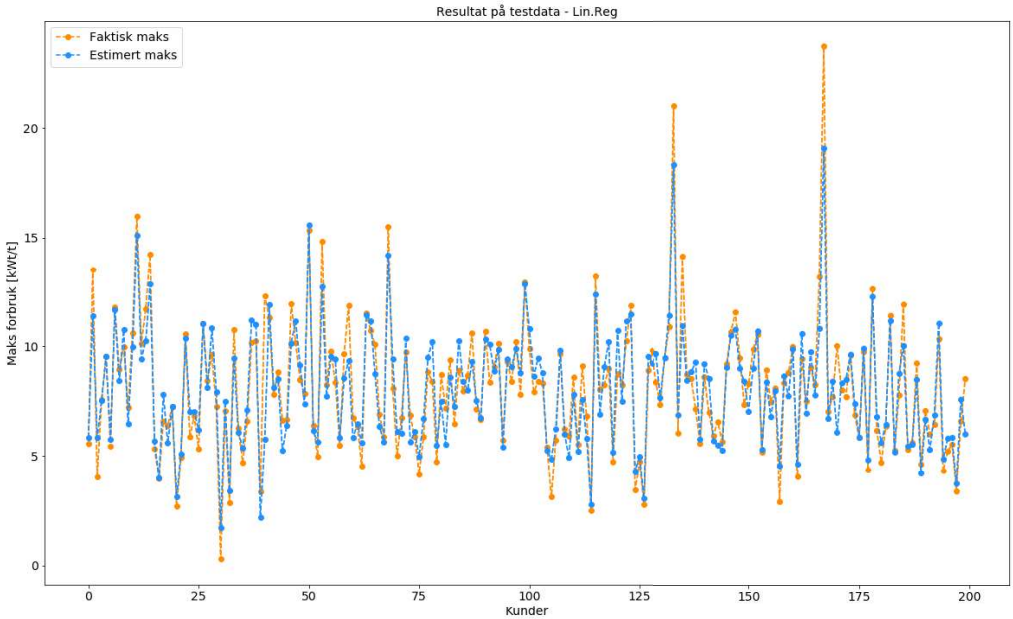
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

hvor $\hat{\beta}$ er koeffisienter som ble estimert via en minste kvadraters metode.

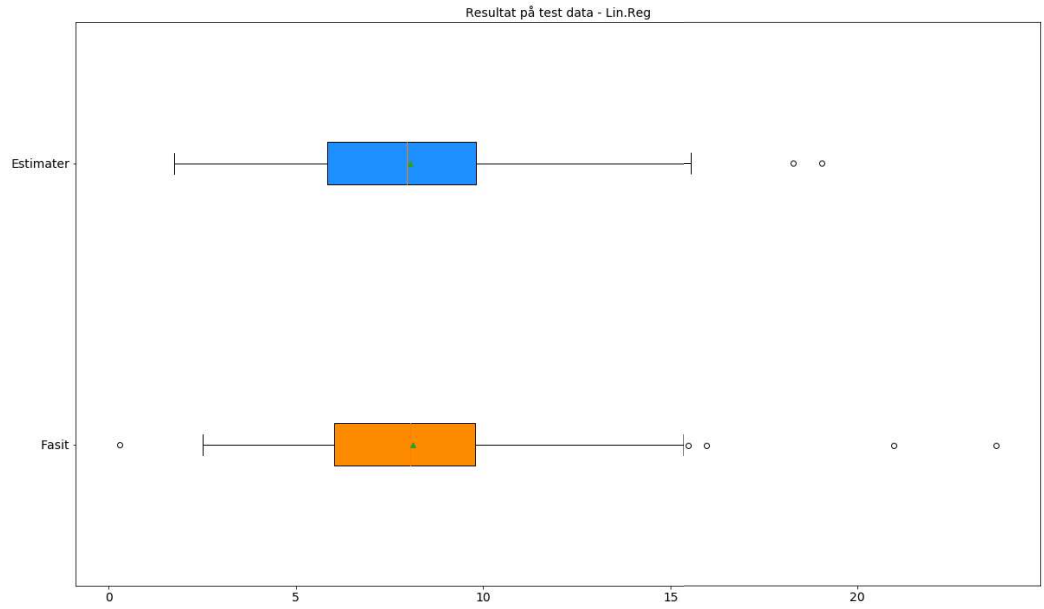
Ved å anvende denne metoden på 800 kunder (trening settet) fikk vi følgende resultater på 200 kunder (test settet):

Tabell 5 Feilmål multippel lineær regresjon

Data	MSE	ME
200 kunder	1.79	0.065
Kryssvalidering	1.579	-0.0003



Figur 6 Multippel lineær regresjon – Estimater over fasit



Figur 7 Multippel lineær regresjons modell – Boks plott av estimater og fasit

Figur 6 viser oss at modellen gjør en relativt god jobb ved å estimere maks for 200 nye kunder. Dette viser oss at modellen er i stand til å plukke opp et signal i datasettet som gjør at den er i stand til å estimere en maks per kunde. Boks plottet viser oss også at fordelingen til estimatene er relativt nærme fordelingen til de faktiske verdiene. Videre ser vi at gjennomsnittet til estimatene, gitt med en grønn trekant i hver boks, ligger tett på gjennomsnittet til de faktiske verdiene. Det betyr at modellen verken overestimerer eller underestimere noe særlig. Til slutt ser vi at MSE verdien fra kryssvalideringen er lavere sammenlignet med referanse modellen. Det viser oss at modellen scorer bedre i gjennomsnitt sammenlignet med referanse modellen. Men det er viktig å presiserer at referansemodellen er noe sensitiv til årene som brukes som grunnlag.

Videre i rapporten presenterer vi kort de andre maskinlærings modellene som ble brukt. Men av hensyn til plass og tid tar vi bare med teorien bak hver modell og scoren fra kryssvaliderings algoritmen.

3.2.3 Stochastic gradient descent

I grove trekk kan man si at denne algoritmen er en itererende metode for å finne et lokalt minimum til en eller annen funksjon som måler feilledd mellom fasit og estimater fra en maskinlæringsmodell. Prosessen ved å finne lokalt minimum oppdaterer da parameterne i funksjonen. Rent matematisk er dette gitt med følgende:

$$\theta = \theta - \delta * \nabla_\theta J(\theta; x^i; y^i)$$

hvor θ er parameterne til kostfunksjonen J , δ er en treningshastighet, ∇_θ er gradienten til J med hensyn på parameterne og x^i og y^i er en observasjon fra trening settet. Parameterne blir i dette tilfelle oppdatert for hver observasjon.

3.2.4 Beslutningstre

Et beslutningstre er en modell som kan brukes til både regresjon og klassifisering av data. Det er en ikke-parametrisk modell. Det betyr at modellen ikke har en antagelse om hvordan funksjonen som beskriver

forholdet mellom forklaringsvariablene og responsvariablen ser ut. I korte trekk så er beslutningstrærne satt sammen av blader, kanter og generasjoner og bruker spørsmål knyttet til forklaringsvariablene for å estimere verdier. Rent matematisk og grovt forenklet kan vi si at:

1. Modellen deler utfallsrommet, det vil si mengden av mulige verdier for forklaringsvariablene $X_1, X_2, \dots, X_p$, inn i J distinkte ikke over-lappende regioner $R_1, R_2, \dots, R_J$
2. For hvert datapunkt som er i en region R_j gir vi samme estimat, nemlig gjennomsnittet av responsvariablene for treningsdatapunktene i R_j

3.2.5 Beslutningskog

Denne typen modell består i hovedsak av mange små mindre kompliserte modeller. En beslutningskog består derfor av mange små beslutningstrær. Et estimat består i dette tilfelle av gjennomsnittet av hvert estimat fra beslutningstrærne i modellen.

3.2.6 XGBoost

Står for Extreme Gradinet Boosting Algoritme og er på mange måter lik en beslutnings skog. Forskjellen er i midlertid at denne algoritmen basere seg på en gradient decent algoritme og iterer seg gjennom et og et tre for deretter å prøve å korrigere feilen etter hvert estimat. Dette gjør den helt til den ikke kan forbedre seg noe nevneverdig.

3.2.7 CATBoost

Relativt lik XGBoost, men stiller f.eks. ikke samme krav om at hvert tree som vokser skal være symmetrisk.

3.2.8 Resultater - Maskinlæringsmodeller

Tabell 5 Feilmål alle maskinlæringsmodeller

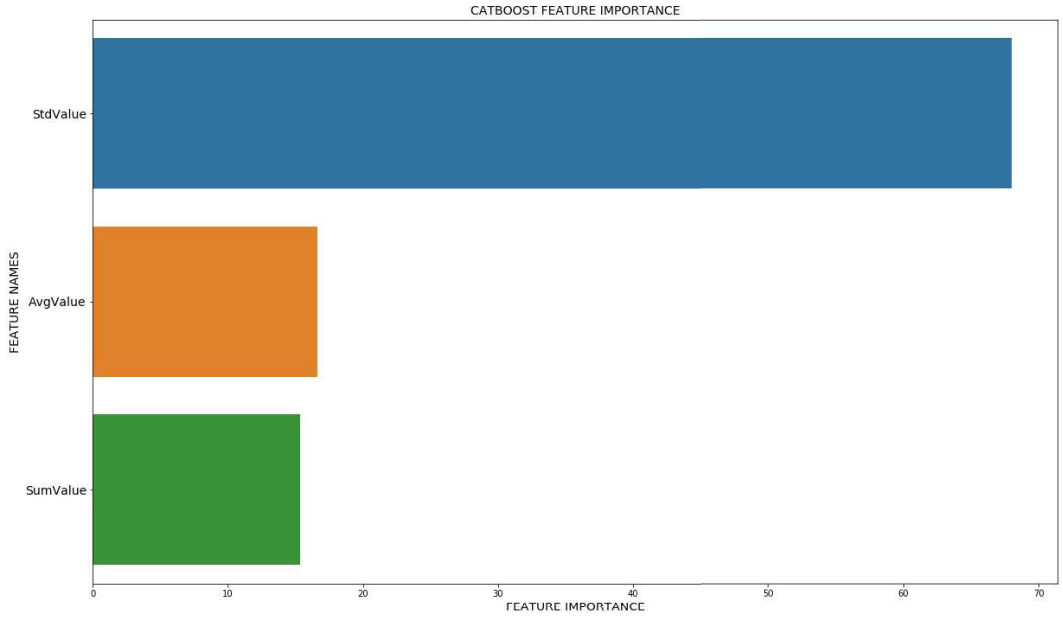
Modell	MSE	ME
Historisk maks	2.090	0.608
Lin.Reg.OLS	1.579	-0.0003
Lin.Reg.SGD	1.567	0.0001
Beslutnings tre	2.94	0.00136
CATBoost	1.812	0.007
XGBoost	2.176	-0.004

Tabell 5 viser oss resultatet fra kryssvalidering for hver modell. Ved å fokusere på MSE verdiene ser vi at modellene som modellerer et lineært forhold mellom forklarings og respons variablene ser ut til å prestere best. Videre ser det ut til å at de mer kompliserte modellene presterer dårligere sammenlignet med referanse modellen, med unntak av CATBoost modellen. Grunnen til dette kan være at disse

modellene *overtrener* dvs. at modellen er for komplisert. Basert på tabellen så vi derfor at en lineær regresjons modell via en minste kvadraters metode løste problemet mest effektivt og med best resultat. Den enkleste modellen er som regel å foretrekke.

3.2.9 Resultater - Maskinlæringsmodeller – Feature Importance

Mange maskinlæringsmodeller har en funksjon som gir et mål på hvilken av forklaringsvariablene som vektes høyest for å gi estimater. Matematikken bak dette varierer fra modell til modell. Vi så i denne delen av prosjektet at *alle* modellene vektet *standardavviket for forbruket* per kunde som viktigst. Dette gir noe mening da variasjon ofte gir informasjon. Til slutt så vi at alle modellene i denne delen av prosjektet vektla *sum for forbruk* lavest. Med hensyn til plass og tid ser vi bare på hvordan feature importance fra CATBoost modellen ser ut (er representativ for de andre modellene):



Figur 8 Feature importance fra CATBoost modell

3.3 Velanders formel

Velanders formel gir en estimert maks-last for enkeltkunders forbruk basert på summert årsforbruk for kunden. I dette prosjektet ble året med høyest sum forbruk tatt i bruk. Velanders formel for maks-last verdi for en enkeltkunde er gitt ved følgende:

$$\hat{P}_{i,y} = k_1 W_{i,y} + k_2 \sqrt{W_{i,y}}$$

hvor $W_{i,y}$ er summert årsforbruk for kunde i for år y. Verdiene k_1 og k_2 er konstanter som er standard hos ulike nettselskaper og har blitt justert over tid. Disse koeffisientene er bestemt for ulike kundetyper. En liste over eksempler på slike verdier er angitt nederst i dette delkapittelet. Summert forbruk kan beregnes ut ifra følgende formel:

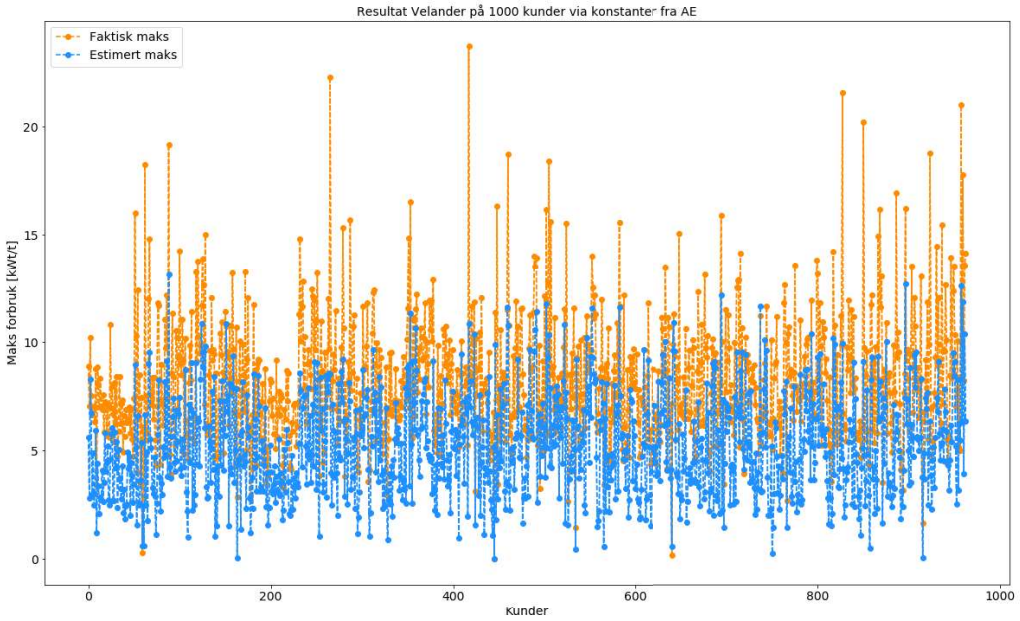
$$W_{i,y} = \sum_{t=1}^{T} x_i(t)\, \alpha_{x_i(t),y} \text{ hvor } \alpha_{x_i(t),y} = \begin{cases} 1 \; hvis \; x_i(t) \; \epsilon \; y. \\ 0 \; ellers. \end{cases}$$

Merk at $x_i(t)$ er den totale tidsserien til kunden.

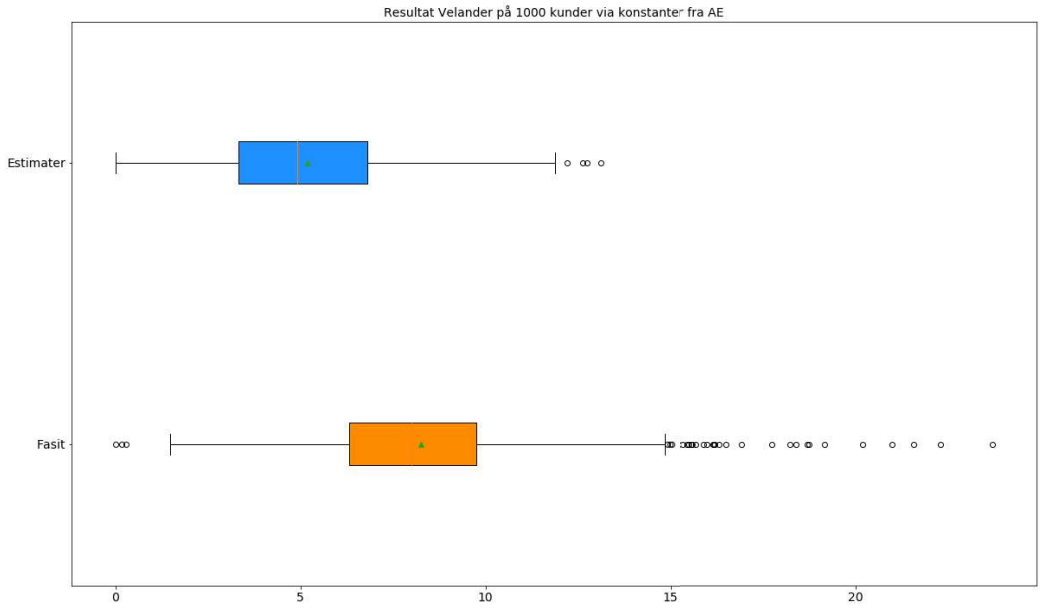
Vi har i dette prosjektet tatt utgangspunkt i data for 2018,2019 og 2020 for å finne året med høyest sum forbruk per kunde. Da Velander gir et generelt estimat for en maks verdi per kunde har vi testet estimatet for observerte maksverdier i 2020. Ved å ta i bruk konstanter gitt fra Agder Energi fikk vi følgende resultater:

Tabell 6 Feilmål Velanders med konstanter fra Agder Energi

MSE	ME
13.46	3.07

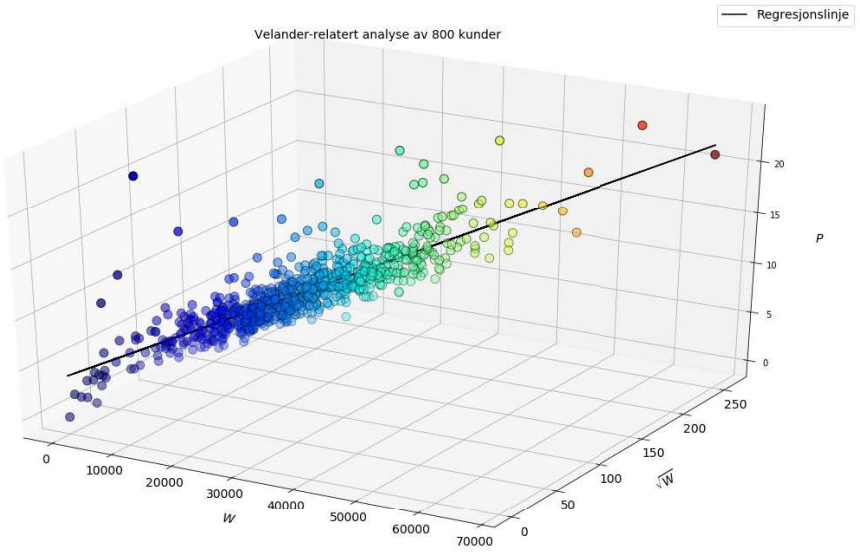


Figur 9 Velander med konstanter fra Agder Energi - Estimater over fasit



Figur 10 Velander med konstanter fra Agder Energi - Boks plott av estimater og fasit

Figur 9 og 10 viser oss estimatene i blått og faktiske verdier i Orange. Figurene viser oss at ved å bruke $k_1 = 0.00021$ og $k_2 = 0.019$, som er konstanter gitt for husholdningskunder i Agder Energi nett, så ender vi opp med at Velanders *underestimerer* maks per kunde. Videre ser vi i tabell 6 at vi får en MSE verdi som er relativ høy sammenlignet med hva vi har sett tidligere. Neste steg kjørte vi samme prosess bare at vi valgte å hente ut konstantene ved via en multippel lineær regresjon basert på sum, kvadrert sum og observert maks for 800 kunder:



Figur 11 Multiple linære regresjon for å finne Velanders konstanter

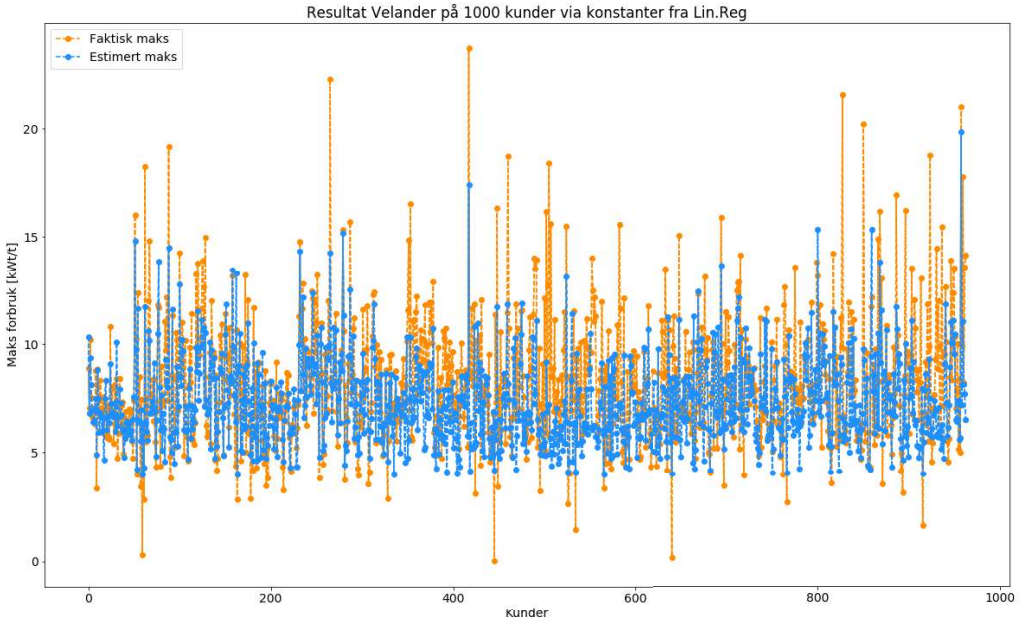
Konstantene vi fikk var $k_0 = 4.1650$, $k_1 = 0.0003$ og $k_2 = -0.0138$. Det gir følgende formel:

$$\hat{P}_{i,y} = k_0 + k_1 W_{i,y} + k_2\sqrt{W_{i,y}} = 4.1650 + 0.0003W_{i,y} - 0.0138\sqrt{W_{i,y}}$$

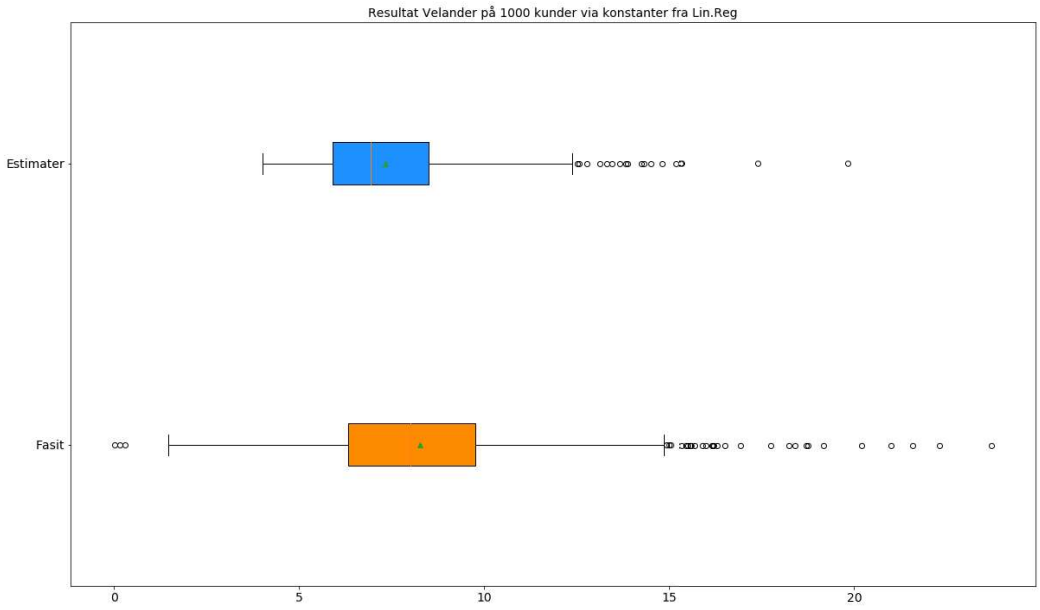
som ga følgende resultater:

Tabell 6 Feilmål Velanders med konstanter fra multiple lineær regresjon

MSE	ME
5.44	0.93



Figur 12 Velander med konstanter fra multiple lineær regresjon - Estimater over fasit

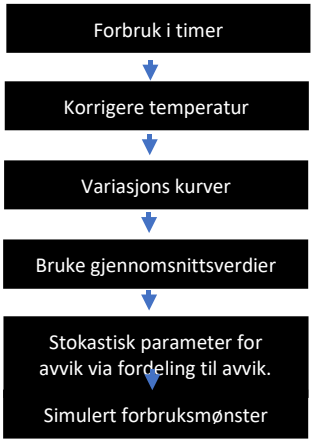


Figur 13 Velander med konstanter fra multiple lineær regresjon - Boks plott av estimater og fasit

Plottene viser oss at vi forbedret estimatene våre ved å hente koeffisienter fra en multiple lineær regresjon. En betydelig forskjell er at vi i dette tilfelle har inkludert et *bias ledd* k_0 som skalerer estimatene på en slik måte at modellen underestimerer mindre. Vi oppnår er reduksjon på 8.02 MSE og boks plottet bekrefter at vi underestimerer i mye mindre grad. På tross av dette ser vi fortsatt at MSE verdien ligger betraktelig høyere sammenlignet med tidligere resultater.

3.4 Stokastisk lastmodellering

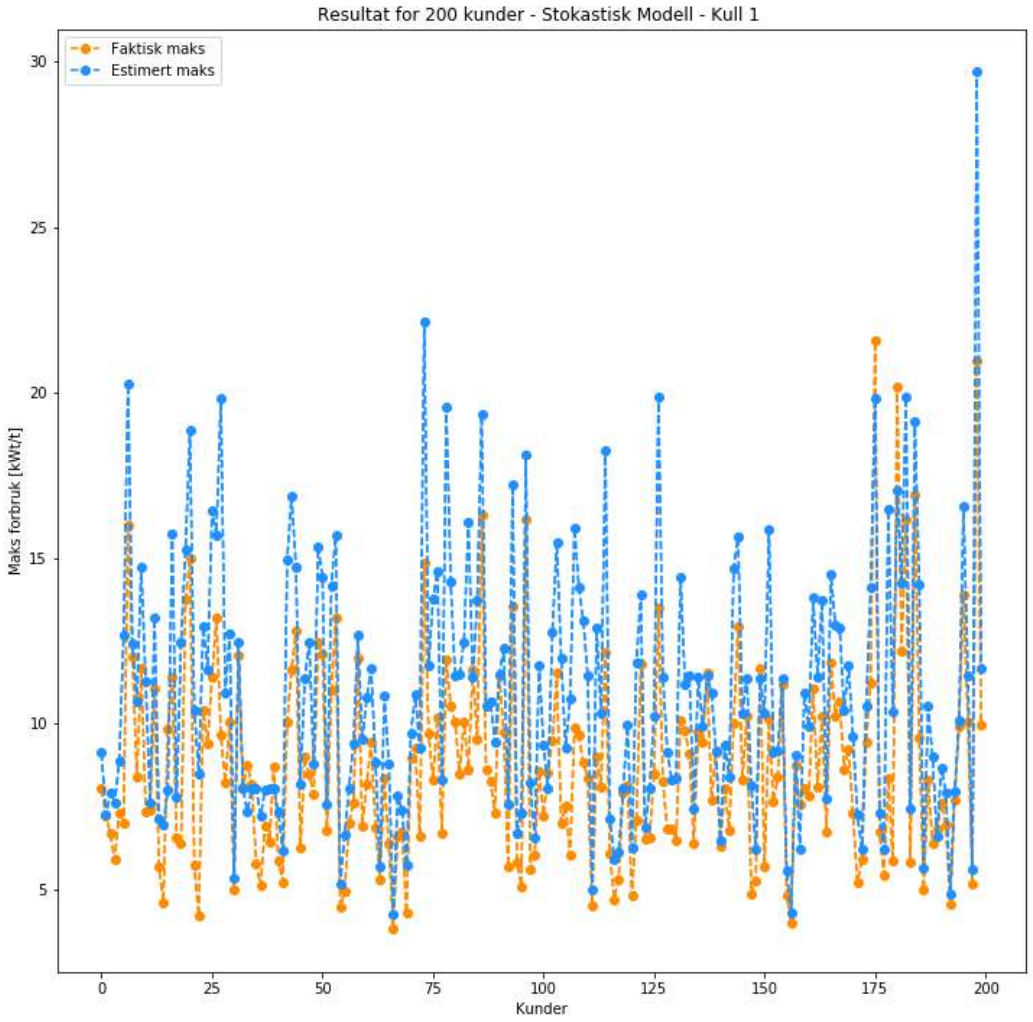
Ved å ta utgangspunkt i Erling Tønne sin PhD oppgave: «Planning of the Future Smart and Active Distribution Grids – With emphasis on probabilistic load and generation modelling based on data from smart meters» tok vi i bruk en stokastisk lastmodelleringsmetode. Kort oppsummert så er dette en type *sannsynlighetsmodell* (probabilistic) som simulerer et generelt forbruk per kunde. Modellen tar et utgangspunkt i alt av tilgjengelig tidsserie data per kunde. Gangen i modellen kan beskrives med følgende punkter:



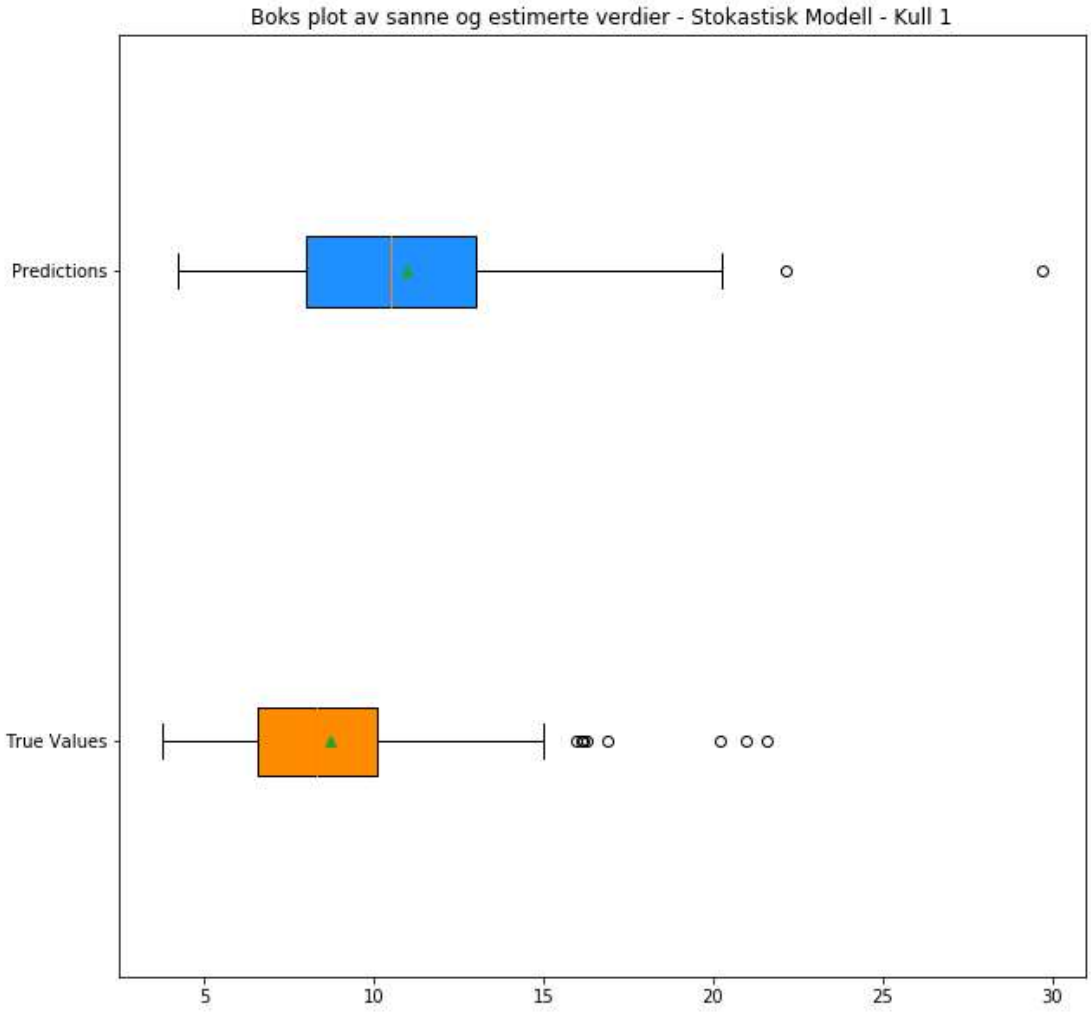
Denne modellen er krevende når det kommer til datakraft og vi endte opp med å anvende den på totalt 800 kunder delt opp i kull på 200. Resultatene er derfor basert på gjennomsnittet av kjøring per kull. Fremgangsmåten som ble brukt var at vi simulerte forbruket til hver kunde for deretter å plukke ut den høyeste simulerte maks verdien. Denne verdien sammenlignet vi med den faktisk observerte høyeste maks verdien. Videre la vi en øvre begrensing for hver kunde som vi kalte teoretisk maks:

$$T_{max} = \sqrt{nPhases} * InstalledMeterVoltage * Maxcurrent$$

og fikk følgende resultat for kull 1:



Figur 14 Stokastisk lastmodell - Estimater over fasit



Figur 15 Stokastisk lastmodell - Boks plott av estimater og fasit

Resultatene fra kull 1, som er representativt for resultatene for de andre kullene, viser oss at modellen i stor grad overestimerer maks forbruk per kunde. Ved å ta gjennomsnittet av kjøringene ender vi opp med følgende:

Tabell 7 Feilmål Stokastisk lastmodell

MSE	ME
8.523	-2.228

Tabellen viser oss at modellen ligger godt over både referanse modellen (historisk maks), maskinlæringsmodellene og Velanders (med konstanter fra Agder Energi) med hensyn på MSE verdier.

ME verdien indikerer også at modellen i gjennomsnitt overestimerer, noe vi også ser via gjennomsnittene i boks plottene i figur 15.

3.5 Resultat – Alle modeller

Tabell 8 Feilmål alle modeller

Modell	MSE	ME
Historisk maks	2.090	0.608
Lin.Reg.OLS	1.579	-0.0003
Lin.Reg.SGD	1.567	0.0001
Beslutnings tre	2.94	0.00136
CATBoost	1.812	0.007
XGBoost	2.176	-0.004
Velanders m.konstanter fra AE	13.46	3.07
Velanders m.konstanter fra Lin.Reg	5.44	0.93
Stokastisk Lastmodellering	8.523	-2.228

Tabellen viser oss først og fremst at vi fikk lavest MSE ved å bruke en modell som er lineær for å estimer maks forbruk per kunde. Da den enkleste modellen som regel er å foretrekke ender vi opp med en lineær regresjons modell basert på minste kvadraters metode. Videre viser det oss at mer kompliserte modeller, da spesielt modeller som baserer seg på beslutningstrær, har en tendens til å få høyere MSE verdier. Det kan tenkes at mer kompliserte modeller overtrener ved å estimere et ikke-lineært forhold mellom variabler som ser ut til å være lineære. Vi så også at alle maskinlæringsmodellene vektla *standardavviket for forbruk* høyest for å estimere en maks per kunde, men også vektla *sum for forbruk* lavest. Videre viser tabellen oss at Velanders formel oppnår en relativt høy MSE verdi sammenlignet med maskinlæringsmodellene og at dette er i stor grad på grunn av den underestimerer. Men, vi ser også at man oppnå bedre resultater ved å hente koeffisientene fra en multiple lineæregresjon basert på sum, kvadrert sum og observert maks per kunde. Til slutt ser vi at stokastisk lastmodellering har en tendens til å overestimere, men også oppnå høyere MSE verdier sammenlignet med de andre modellene (med unntak av Velanders med konstanter fra lin.reg).

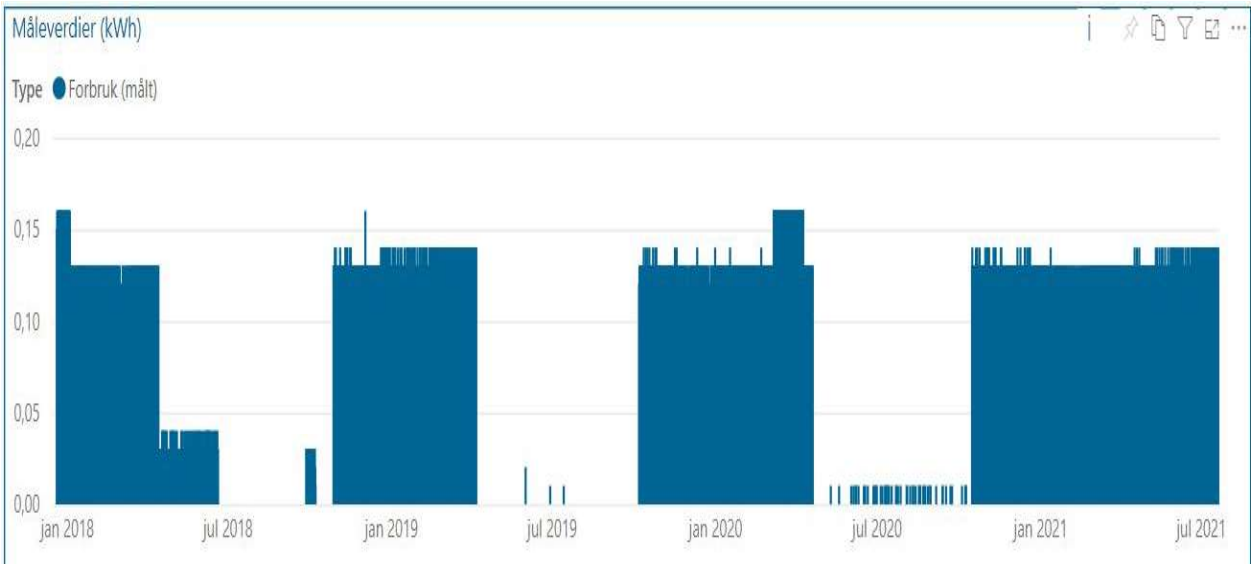
3.6 Andre resultater

Vi har i dette prosjektet testet ut forskjellige metoder og teknikker for gjøre det mest mulig utfordrende for maskinlærings modellene. På grunn av tid har disse forsøkene verdt noe overfladiske og i noen tilfeller bør kanskje mer forskning til. Men, resultatene er fortsatt interessante og blir tatt med da de kan bygges på videre.

3.6.1 Fjerne maks observert fra datasett

I en kontekst av maskinlæring er det hensiktsmessig og «best practice» at datapunktet vi prøver å estimere ikke implisitt er i forklaringsvariablene. I dette prosjektet inneholder forklaringsvariablene implisitt observert maks, f.eks. inneholder standardavviket observert maks for kunden. Selv om dette utgjør en promille tenkte vi det var interessant å se om hvordan modellene reagerte *ved å fjerne observert maks fra datasettet før vi aggregerte opp standardavviket, sum og gjennomsnitt av forbruket*. Med hensyn på tid valgte vi ut lineær regresjon via minste kvadraters metode og CATBoost for å teste dette.

Det første vi la merke til var at noen kunder hadde repeterende maksverdier. For eksempel fant vi kunder som hadde en observert maksverdi 315 ganger. Dette var ekstremt da de fleste kundene som var i denne gruppen lå på 2-8 repeterende maks verdier. Disse type kundene vil maskinlæringsmodellen har mindre problemer med å estimere en maks for da «svaret» på estimatet opptrer flere ganger i forklaringsvariablene.



Figur 16 Kunde med repeterende maksverdier

Videre sammenlignet vi forklaringsvariablene før og etter ved å se på statistikken: Vi kaller statistikken for forklaringsvariablene for opprinnelig dvs. med observert maks og endret dvs. at observert maks er fjernet. Vi fikk følgende tabeller:

Tabell 9 Statistikk for opprinnelig data

Statistikk – Opprinnelig	Gjennomsnitt	Standardavvik	Sum
Gjennomsnitt	1.923507	1.166544	55548.330900
Standardavvik	0.902967	0.464969	26906.673141
Min	0.000000	0.000058	0.010000
25%	1.286214	0.827684	37041.005000
50%	1.825230	1.134932	52456.985000
75%	2.496404	1.437225	72881.235000
Maks	6.684558	3.970537	201298.780000

Tabell 10 Statistikk for endret data

Statistikk - Endret	Gjennomsnitt	Standardavvik	Sum
Gjennomsnitt	1.923259	1.165865	55539.937040
Standardavvik	0.902917	0.464820	26904.459289
Min	0.000000	0.000000	0.000000
25%	1.286075	0.826644	37035.157500
50%	1.824983	1.134524	52448.805000
75%	2.496225	1.436762	72873.212500
Maks	6.684083	3.967792	201277.790000

Vi ser at fordelingen til forklaringsvariablene endrer seg noe, blant annet så faller gjennomsnittet av sum forbruk fra fra 55548.33 til 55539.93. Videre ser vi også at standardavviket på standardavviket faller fra 0.4649 til 0.4648. Vi trener opp de utvalgte maskinlæringsmodellene og får følgende MSE resultater (hvor opprinnelig er modellene med observert maks fortsatt til stedet og endret er uten):

Tabell 11 Feilmål på maskinlæringsmodeller med og uten observert maks

Metrikk	Lin.Reg -Opprin	CatBoost - Opprin	Lin.Reg - Endret	CatBoost - Endret
MSE	1.579	1.812	1.593	1.817

Vi ser at MSE verdien til begge modellene øker noe ved å fjerne observert maks verdi før man aggregerer opp datasettet. Ikke drastisk mye, men nok til at det hadde vært hensiktsmessig å utforske problemstillingen i større grad.

3.6.2 Maskinlæringsmodeller fremover i tid og nye forklaringsvariabler

En av problemstillingene i dette prosjektet har vært å finne en modell som gir et estimat som presterer bra uavhengig av tid. Hvor både Velanders og Stokastisk lastmodell gir tidsuavhengig resultater er det noe mer usikkert hvordan estimatene fra en maskinlæringsmodell lar seg generalisere fremover i tid. Videre har vi også hatt andre ideer til forklaringsvariabler f.eks. 99 percentilen per kunde og 5 forrige registrerte maksverdiene til kunden. Utfordringen med disse er at vi endte med utrolig lave MSE verdier. Dette førte til at vi var usikre på om det kanskje var en *lekkasje* i disse dvs. at forklaringsvariablene implisitt innehar mye informasjon om responsvariabelen. Det vil i tilfelle gjøre det alt for lett for maskinlæringsmodellene. Vi tester både generaliserbarheten, andre forklaringsvariabler og lekkasje problematikken ved å fjerne data fra 2020 datasettet som modellene trener på. Først trener og tester vi modellene på data fra 2018-2019, deretter tar vi estimatene fra testsettet videre og mål feilen på året 2020. Dette er data som modellen ikke har sett og eventuelle lekkasjer bør da dukke opp via unaturlige små feilmålinger.

Oppsettet i denne delen av prosjektet er at man har delt treningen opp i 70/30 slik at man da tester på 300 kunder. Estimatene fra testfasen blir da brukt for å estimere 2020 for de samme kundene. Vi har også lagt til en ny metrikk som heter *Mean absolute error:*

$$\frac{1}{\mathrm{n}}\sum_{i=0}^{n}|Y_i - \hat{Y}_i|$$

Som er et mål på hvor mye modellene i gjennomsnitt bommer i kWt/t. Vi valgte å inkludere denne da den gir god informasjon om prestasjonen til modellene våre.

Med hensyn på tid valgte vi bare ut de mest lovende maskinlæringsmodellene i denne delen av prosjektet. Vi fikk følgende resultater:

Tabell 12 Feil mål med maskinlæringsmodeller med ulike forklaringsvariabler fremover i tid

Modell	Forklaringsvariabler	MSE	MAE
Historisk maks	Tidligere obs.maks	2.090	1.057
Lin.Reg.OLS	Std, Avg og Sum	2.078	1.155
CATBoost	Std, Avg og Sum	2.135	1.179
Lin.Reg.OLS	5 tidligere obs.maks	2.038	1.047

CATBoost	5 tidligere obs.maks	1.933	1.041
Lin.Reg.OLS	Std, Avg og 99percent	1.533	1.006
CATBoost	Std, Avg og 99percent	1.681	1.059

Tabell 12 viser oss at oppnår lavest feilverdi ved å ta i bruk standardavvik, gjennomsnitt og 99 percentilen som forklaringsvariabler. Vi oppnår en feil på 1.006 kWt/t og en variasjon på rundt 1.533 kWt/t på estimatene fremover i tid. Videre ser vi også at lineær regresjon fortsatt ser ut til å score lavest, men at CATboost modellen ligger ganske nært. Begge modellene vektet 99 percentilen høyest i feature importance. Vi ser at ved å bare ta i bruk de 5 tidligere forbrukstoppene til kundene oppnår vi en lavere score sammenlignet med historisk maks, da både med lineær regresjon og CATBoost. Det er også verdt å nevne at CATBoost gjør en bedre jobb sammenlignet med lin.reg i dette tilfelle. Det kan være at det er en ikke-lineær sammenheng mellom de 5 forrige toppene og fremtidig maks. Til slutt ser vi at vi oppnådde svakest resultat ved å kun ta i bruk forklaringsvariablene standardavvik, gjennomsnitt og sum. Basert på disse resultatene er det lite som tyder på lekkasje og estimatene fra maskinlæringsmodellene ser ut til å la seg generalisere fremover i tid.

3.7 Innovasjoner fra Piloten

Tabell 13 Beskrivelse av innovasjoner i forskningsrådets kategorier

Forskningsrådets kategorier	Beskrivelse	Antall
Ferdigstilte nye/bedre metoder/modeller/ prototyper	NA	
Bedrifter utenfor FMEen som har innført nye/forbedrede metoder eller modeller eller teknologi	NA	
Bedrifter innenfor FMEen som har innført nye/forbedrede arbeidsprosesser	NA	
Bedrifter innenfor FMEen som har innført nye/ forbedrede metoder eller modeller eller teknologi	NA	
Inngåtte lisensieringskontrakter	NA	
Registrerte patenter	NA	
Ferdigstilte nye/forbedrede produkter	NA	
Ferdigstilte nye/forbedrede prosesser	NA	
Ferdigstilte nye/forbedrede tjenester	NA	
Nye foretak som følge av FME'en	NA	
Nye forretningsområder i eksisterende bedrifter	NA	

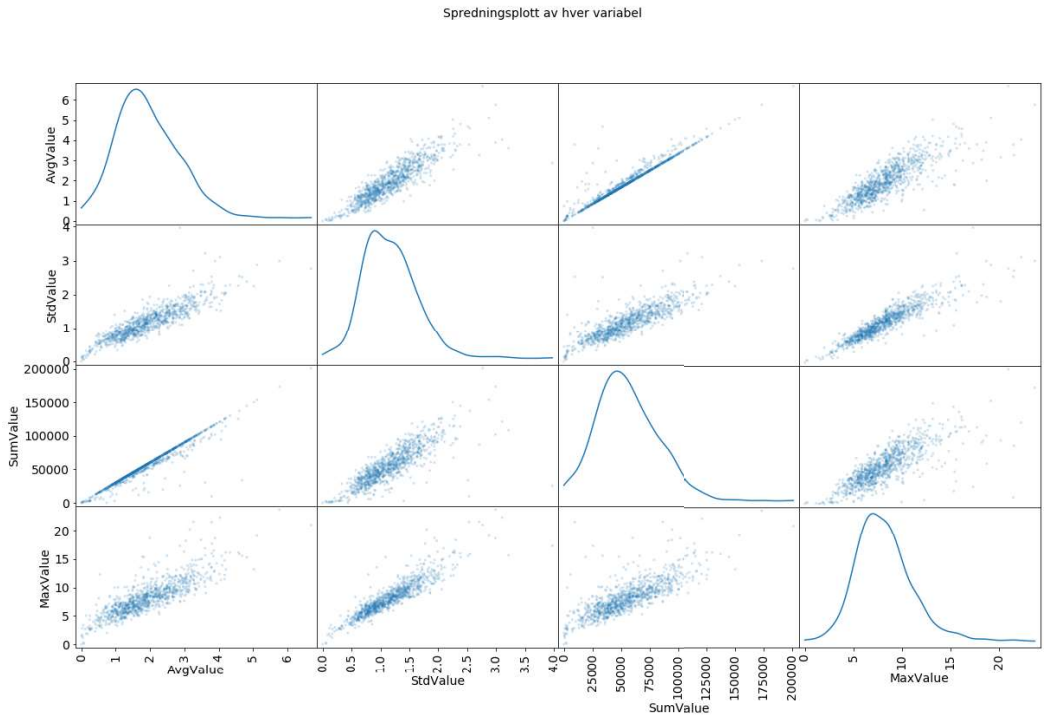
4 Tekniske/faglige erfaringer fra Piloten

4.1 Oppsummering

Vi har i dette prosjektet sett at man kan estimere maks effekt til enkelt kunder effektivt og med gode resultater via en lineær regresjons modell som baserer seg på minste kvadraters metode. Resultatene baserte seg på forklaringsvariablene standardavvik, sum og gjennomsnitt av forbruk per kunde, hvor førstnevnte ble vektes som viktigst for å estimere maks per kunde. Vi så at denne metoden fikk lavest MSE score sammenlignet med andre maskinlæringsmodeller, stokastisk lastmodell og Velanders. Men, vi så også at resultatene kunne forbedres ved å bytte sum ut med 99-percentilen. Vi så dette da vi estimerte maks for kunder «fremover» i tid. Videre så vi også at Velanders hadde en tendens til å underestimere og at stokastisk lastmodell hadde en tendens til å overestimere maks per kunde. Men ved å bruke koeffisienter fra en multiple lineær regresjons modell var det mulig å minke overestimatene og dermed forbedre resultatene for Velanders.

Modell	MSE	ME
Historisk maks	2.090	0.608
Lin.Reg.OLS	1.579	-0.0003
Lin.Reg.SGD	1.567	0.0001
Beslutnings tre	2.94	0.00136
CATBoost	1.812	0.007
XGBoost	2.176	-0.004
Velanders m.konstanter fra AE	13.46	3.07
Velanders m.konstanter fra Lin.Reg	5.44	0.93
Stokastisk Lastmodellering	8.523	-2.228

Vi så også at det var en lineær positiv sammenheng mellom variablene. Dette kan forklare hvorfor en lineær modell gjorde en så god jobb med å estimere maks per kunde.



Vi så også at noen av modellene økte feilleddene sine ved å fjerne observert maks fra datasettet før man aggregerte opp forklaringsvariablene.

Metrikk	Lin.Reg -Opprin	CatBoost - Opprin	Lin.Reg - Endret	CatBoost - Endret
MSE	1.579	1.812	1.593	1.817

Til slutt testet vi også maskinlæringsmodellene fremover i tid hvor vi blant annet byttet ut sum av forbruk med 99 percentilen som forklaringsvariabel. Resultatet indikerte at estimatene lar seg generalisere fremover i tid og at man kan oppnå bedre resultater ved å bytte ut sum med 99-percentilen.

Modell	Forklaringsvariabler	MSE	MAE
Historisk maks	Tidligere obs.maks	2.090	1.057
Lin.Reg.OLS	Std, Avg og Sum	2.078	1.155
CATBoost	Std, Avg og Sum	2.135	1.179
Lin.Reg.OLS	5 tidligere obs.maks	2.038	1.047
CATBoost	5 tidligere obs.maks	1.933	1.041
Lin.Reg.OLS	Std, Avg og 99percent	1.533	1.006
CATBoost	Std, Avg og 99percent	1.681	1.059

4.2 Forslag til videre arbeid

En vei å gå videre hadde vært å bruke maskinlærings modeller til å estimere et forbruk per kunde, mer i tråd med hvordan stokastisk lastmodell fungerer. Det hadde også vært interessant å teste maskinlærings modeller som bare bruker 99-persentilen som forklaringsvariabel – er dette nok for å få gode resultater? Videre hadde det vært interessant å bruke mer data for deretter å stokke om slik at modellene blir testet mer uavhengig av tid. Ellers hadde det også vær interessant å teste modellene opp mot andre modeller som «Brukstid» på andre kundegrupper. Videre gir det også mening å sjekke om resultatene fra dette prosjektet står når man estimerer maks for grupper i stedet for enkelt kunder. Et litt mer omfattende steg man kan ta prosjektet er at man plukker ut tilfeldige husholdningskunder som man ikke har data på fra 2019, grupperer disse utefra et sett med utvalgte kriterier. Grupperingen gir da kundene et estimat på standardavvik, gjennomsnitt og 99-persentil verdi som igjen blir brukt i lineær regresjon modellen fra dette prosjektet. Estimert maks forbruk fra disse stegene blir da sammenlignet med observert maks forbruk i 2021. På denne måten kan man kanskje finne en optimal måte å grupper kunder uten data utefra hvilken gruppering som «treffer» best på maks estimatene.

Prosjektnotat

TITTEL			
Pilot Effektanalyse: Vurdering av verdi av 30-sek data (mot 5-min data) fra nettstasjon			
WORK PACKAGE	**VERSJON**	**DATO**	**ANTALL SIDER**
WP Pilot	1.0	2022-01-01	22
FORFATTER(E)		**WP-LEDER**	**GRADERING**
Pål Wagner Per-Oddvar Osland (Oct 8, 2024 13:28 GMT+2)		Maren Istad Maren Istad (Oct 9, 2024 07:36 GMT+2)	Åpen

DISTRIBUSJON		
CINELDI		

SAMMENDRAG

Formålet med denne rapporten er å vurdere verdien man oppnår ved å samle inn data med 30 sekunds samplingsintervall fra nettstasjoner, og spesielt hvilken ytterligere verdi dette gir ut over å ha data med 5 minutt samplingsintervall. Agder Energi Nett har 5-min målinger på stort sett alle nettstasjoner, og ønsker å vite hvilken tilleggsgevinst man kan oppnå ved å gå ned til 30-sek data.

Rapporten ser på to ulike nettstasjoner som forsyner ulikt antall kunder. Dataen som ble samlet fra disse nettstasjonene var effekt, spenning, strøm og effektfaktor, med oppløsning på 30 sekunder. Denne 30-sekundsdata ble aggregert opp til 5 minutts-målinger, slik at både informasjonstap og informasjonsnytte ved bruk av målefrekvensene nevnt ovenfor kunne analyseres. Effektanalysen som ble gjort tok utgangspunkt i alle årstider, slik at variasjonen i de ulike årstidene kunne bli illustrert og analysert. Årstidene i analysen på spenning, strøm og effektfaktor ble ikke tatt i betraktning, ettersom disse vil variere med effekt. Det ble funnet at forskjellen mellom 30-sekunds og 5-minuttsdata er størst i effekt- og strømdataen, ettersom disse henger sammen. Selv om forskjellen mellom målefrekvensene i disse dataene er merkbar, er ikke forskjellen stor nok for å kunne argumentere for at 30-sekundsdataen gir vesentlig mer verdi, eller at viktig informasjon går tapt når dataoppløsningen er 5 minutter. Forskjellen mellom målefrekvensene i spennings- og effektfaktordataen ble funnet til å være minimal, og variansen mellom målefrekvensene var tilnærmet null.

Innholdsfortegnelse

1 Introduksjon

Innsamling av informasjon om effektforbruk, spenning, strøm og effektfaktor i nettstasjoner gir driftsoperatører muligheten til å kunne drifte nett-systemet så effektivt som mulig. Informasjonsinnsamling kan tillate driftsoperatører estimere levetiden på både komponenter i nettstasjoner og ledere mer nøyaktig, samtidig som at det kan hindre at kundene mister strømforsyningen sin. Informasjon som er relevant kan samles inn med ulike oppløsninger, som for eksempel 30-sekunds, 5-minutts, 15 minutts, timesverdier o.l. Hvilken oppløsning som er valgt, danner grunnlag for hvor mye og hvor nøyaktig informasjon er. Jo høyere oppløsning, desto mer nøyaktig informasjonen blir. Det er også viktig å vurdere kost-nytte av valgt oppløsning, ettersom jo høyere oppløsningen på informasjonsinnsamling er, desto mer koster det å samle denne informasjonen inn.

Denne rapporten ser på to ulike nettstasjoner som begge forsyner ulikt antall kunder. Målingene i disse to nettstasjonene har 30-sekundsoppløsning, mens målingene i andre nettstasjoner har 5-minuttsoppløsning. Målet med denne rapporten er å finne ut hvor stor forskjell det er mellom 30-sekundsoppløsning og 5-minuttsoppløsning ved bruk av statistisk analyse.

2 Datagrunnlag og dataformatering

Dette kapittelet beskriver hvordan de ulike dataene ble lastet ned og formattert, slik at nøyaktig analyse kunne bli utført, samt de ulike Python-bibliotekene som ble brukt. Det som er felles for hele analysen, er at Microsoft Azure Databricks ble brukt for å laste ned data fra SQL-server, og spark.sql-funksjonen ble brukt for å transformere data til Pandas. Pakkene som ble benyttet på tvers av analysen er vist i tabell 1.

Tabell 1: Python-biblioteker brukt i alle analyser

Python-bibliotek	Funksjonalitet
Pandas	Dataanalyse og datamanipulasjon
Matplotlib.pyplot	Grafisk framstilling
Numpy	Matematiske kalkulasjoner
Matplotlib.dates	Manipulering av tidsserier i grafisk framstilling av data

2.1 Effekt

Spark.sql ble brukt for å laste ned stasjons id, tidspunkt for måling, og P_{30s}. Når analysen ble utført, var det to stasjoner som var tilgjengelig, og begge hadde ulik datamengde.

Tabell 2: Informasjon om datagrunnlaget

Navn	Antall målinger	Tidsperiode
4631	1 112 462	2020.09.16 – 2021.10.20
5198	1 133 219	2020.09.09 – 2021.10.20

Pandas ble benyttet for konvertere TimeStampUtc til datetime, P_{30s} til desimaltall, og stasjons id til en streng. Videre ble hele dataen forskjøvet 30 sekunder bak, slik at riktig gjennomsnittsverdi kunne regnes ut. I tillegg ble det opprettet kolonner for TimeStampUtc, år, måned, dag, time og minutt. Videre ble det utført systematisk formattering av datasettet:

1. Groupby-funksjonen ble benyttet for å telle antall P_{30s}-verdier i 5-minuttsintervaller.
2. Groupby-funksjonen ble benyttet for å regne ut gjennomsnittlig P_{5m} basert på P_{30s}.
3. Dataen som inneholdt antall P_{30s} i 5-minuttsintervaller ble lagt inn i DF_{5m}.

4. P_{5m} som inneholdt færre enn 8 P_{30s} ble filtrert bort fra DF_{5m}.
5. Datasettene DF_{30s} og DF_{5m} ble slått sammen til $DF_{30s,5m}$ ved bruk av pandas.merge.
6. Forholdsverdiene, $F_{30s/5m}$ ble regnet ut ved dele kolonnen for P_{30s} på kolonnen for P_{5m}.
7. Det ble loopet over dataframen $DF_{30s,5m}$, kolonnen som inneholdt månedstall, og årstid som tilsvarte månedstallet ble lagt i dataframen $DF_{årstid}$.
8. Pandas.merge ble benyttet for å slå sammen dataframene $DF_{30s,5m}$ og $DF_{årstid}$.

For å utføre analyse av datasettet, ble det benyttet pandas.describe()-funksjon, som gir grunnleggende statistisk informasjon om hver kolonne i et datasett, slik som minimal, maksimal og gjennomsnittlig verdi, 1., 2., og 3. kvartil og standardavvik. Det ble også opprettet fire dataframes for hver stasjon, en for hver årstid:

Tabell 3: Oversikt over dataframes som inneholder informasjon for hver årstid

Dataframe	Funksjonalitet
$DF_{stasjon,w}$	Inneholder data for vintermånedene for en gitt stasjon
$DF_{stasjon,sp}$	Inneholder data for vårmånedene for en gitt stasjon
$DF_{stasjon,sm}$	Inneholder data for sommermånedene for en gitt stasjon
$DF_{stasjon,a}$	Inneholder data for høstmånedene for en gitt stasjon

Videre ble det benyttet ulike metoder for grafisk framstilling for å visualisere dataen i $DF_{30s,5m}$, samt dataframene presentert i tabell 3.

2.2 Spenning, strøm og effektfaktor

Spark.sql ble brukt for å laste ned stasjons id, tidspunkt for måling, P_{30s} og de tre linjespenningene, U1, U2 og U3. Her står U1 for spenning mellom fase A og B, U2 mellom fase B og C, og U3 mellom fase C og A. Når analysen ble utført, var det to stasjoner som var tilgjengelig, og begge hadde ulik datamengde.

Tabell 4: Informasjon om datagrunnlaget for spenning

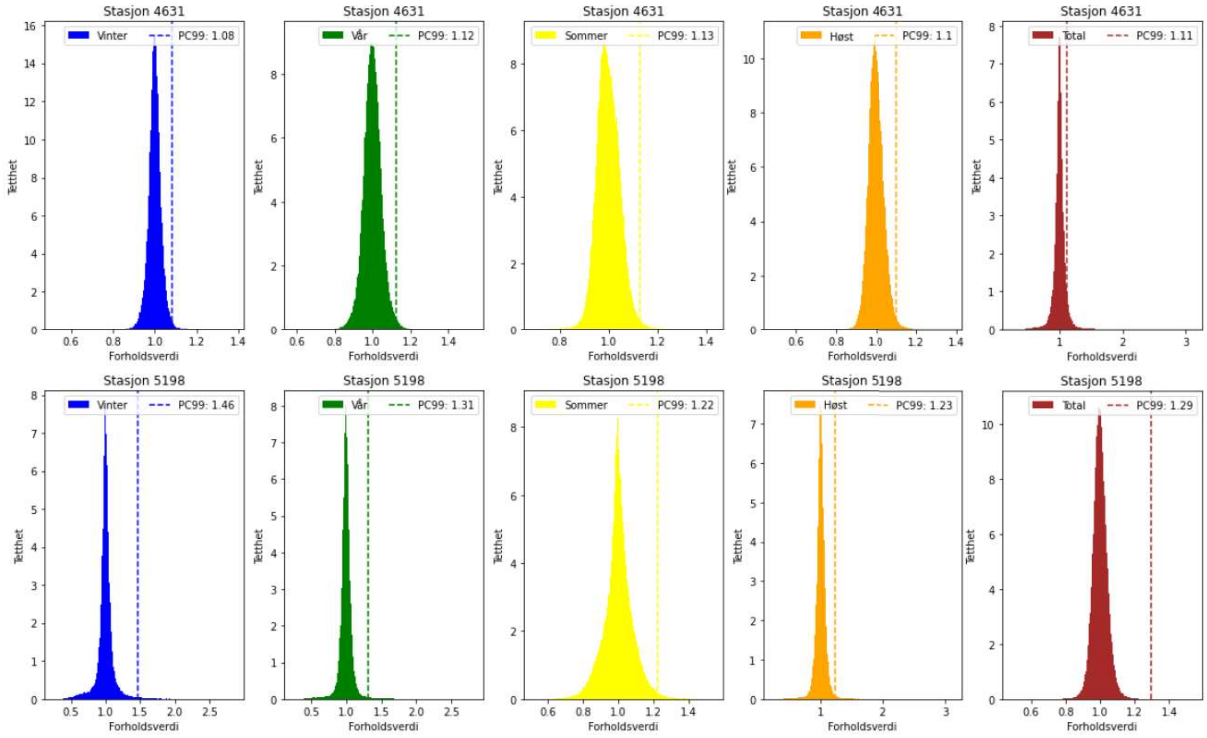
Navn	Antall målinger	Tidsperiode
4631	1 176 792	2020.09.16 – 2021.11.14
5198	1 197 677	2020.09.09 – 2021.11.14

For formattering av dataframen som inneholdt informasjon om spenningsverdier, ble det det brukt samme formattering som presentert i underkapittel 2.1 Effekt. Formattering er presentert punktvis nedenfor.

1. Konvertering av TimeStampUtc til pandas datetime ved bruk av funksjon pd.to_datetime, og endring av P_{30s}, $U1_{30s}$, $U2_{30s}$ og $U3_{30s}$ til float-variabler, og stasjons id til string.
2. Bruk av TimeStampUtc-kolonne til å lage fem ekstra kolonner, som inneholdt år, måned, dag, time og minutt.
3. Utregning av antall 30 sekunds-verdier i 5-minuttsintervaller ved bruk av groupby-funksjon.
4. Utregning av gjennomsnittlig $U1_{5m}$, $U2_{5m}$ og $U3_{5m}$ ved bruk av groupby funksjon.
5. Sammenslåing av 5-minutts-spenningsverdier og informasjon omtalt i punkt 3.
6. Filtrering av 5-minuttsintervaller som inneholdt færre enn 8 30-sekundsverdier.
7. Sammenslåing av 30-sekundsverdier og 5-minuttsverdier i samme dataframe.
8. Utregning av forholdsverdier, U_{30s}/U_{5m} ved å dele kolonnen med 30-sekundsverdier på kolonnen med 5-minuttsverdier.

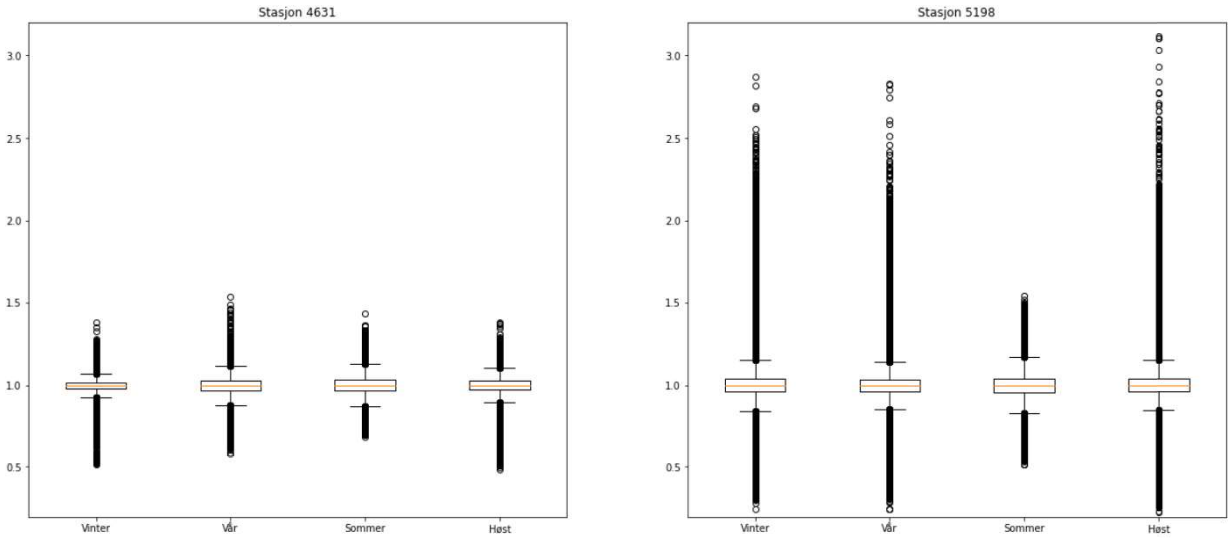
De samme punktene presentert ovenfor ble også brukt for formattering av strøm- og effektfaktordata.

3 Analyse av P_{30s}, P_{5m} og $F_{30s/5m}$



Figur 1: Fordeling for $F_{30s/5m}$ i ulike årstider for de to ulike stasjonene

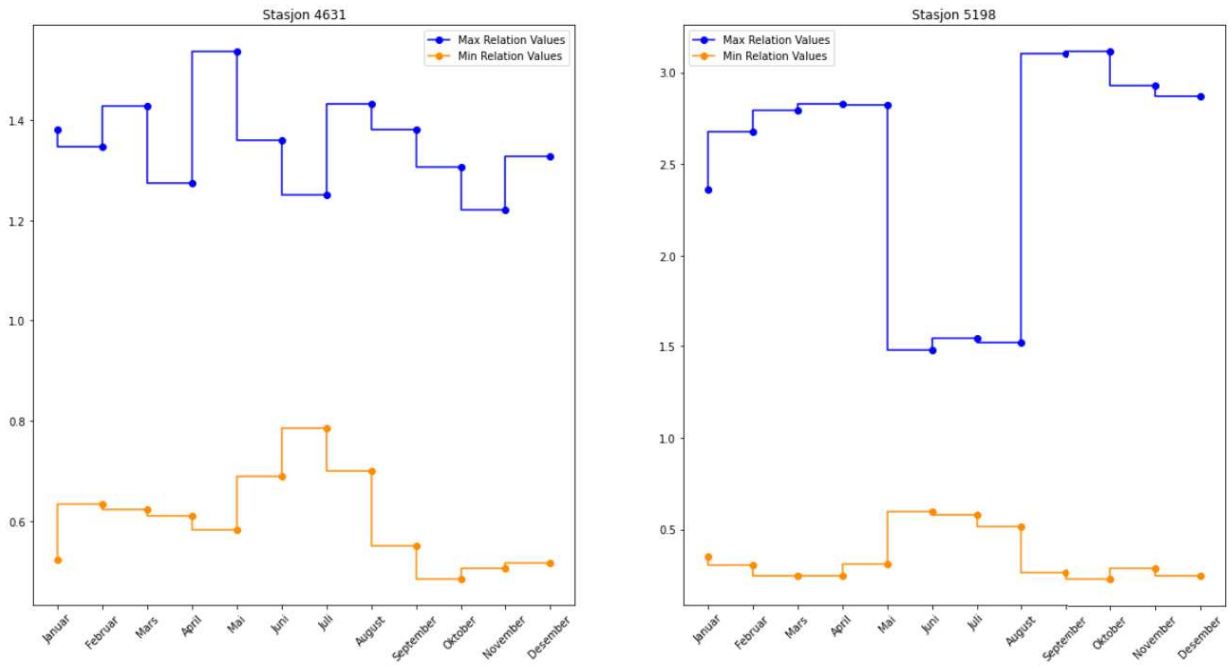
Figur 1 viser fordelingen av forholdsverdier, $F_{30s,5m}$, i ulike årstider for de to stasjonene. Av figuren ser vi at de flesteparten av verdiene er sentrert rundt 1 i alle årstider, for begge stasjonene, men de to stasjonene har relativt ulik varians. Videre kan det merkes at sommermånedene for begge stasjonene opplever svært bred fordelingskurve, noe som betyr at effekten varierer veldig om sommeren. For å se nærmere på fordelingen av forholdsverdier, ble det benyttet boxplot, som vist i figur 2.



Figur 2: Boxplot av forholdsverdier i de ulike årtidene

Figur 2 illustrerer veldig godt hvilken form fordelingen presentert i figur 1 har, men viktigst av alt: antall ekstremalverdier. Av figur 2 kan det ses at stasjon 5198 har svært mange ekstremalverdier i månedene høst-vår; disse verdiene strekker seg helt til 3. Antall verdier som ligger rundt 3 er få, og betyr at $P_{30s} = 3P_{5m}$.

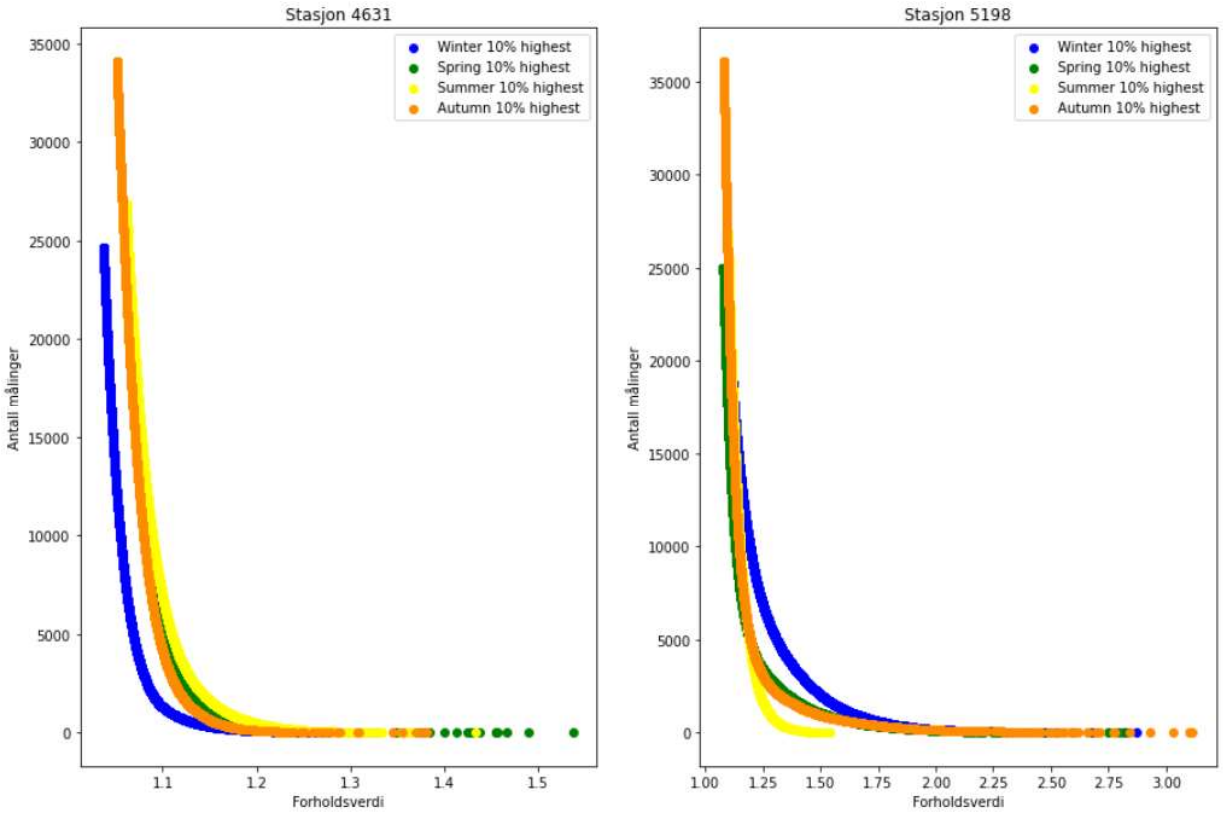
For å sammenligne de to stasjonene med hensyn på maksimale $F_{30s/5m}$, ble det de høyeste forholdsverdiene for hver måned plottet i en linjegraf, som vist i figur 3.



Figur 3: Linjegraf av de største forholdsverdier for hver måned

Når det gjelder stasjon 4631, kan det ses av figur 3 at det er liten variasjon blant de høyeste forholdsverdiene gjennom hele året, og at P_{30s} er i snitt 1.3547 ganger større enn P_{5m}. Når det gjelder stasjon 4198, kan det ses at de største forholdsverdiene stuper i sommermånedene, men er fortsett relativt høye. For stasjon 5198, er P_{30s} i snitt 2.5 ganger større enn P_{5m}.

For å videre illustrere forskjellen mellom stasjonene og årstidene, ble de 10% høyeste forholdsverdiene for begge stasjonene plottet i et scatterplott, som vist i figur 4:

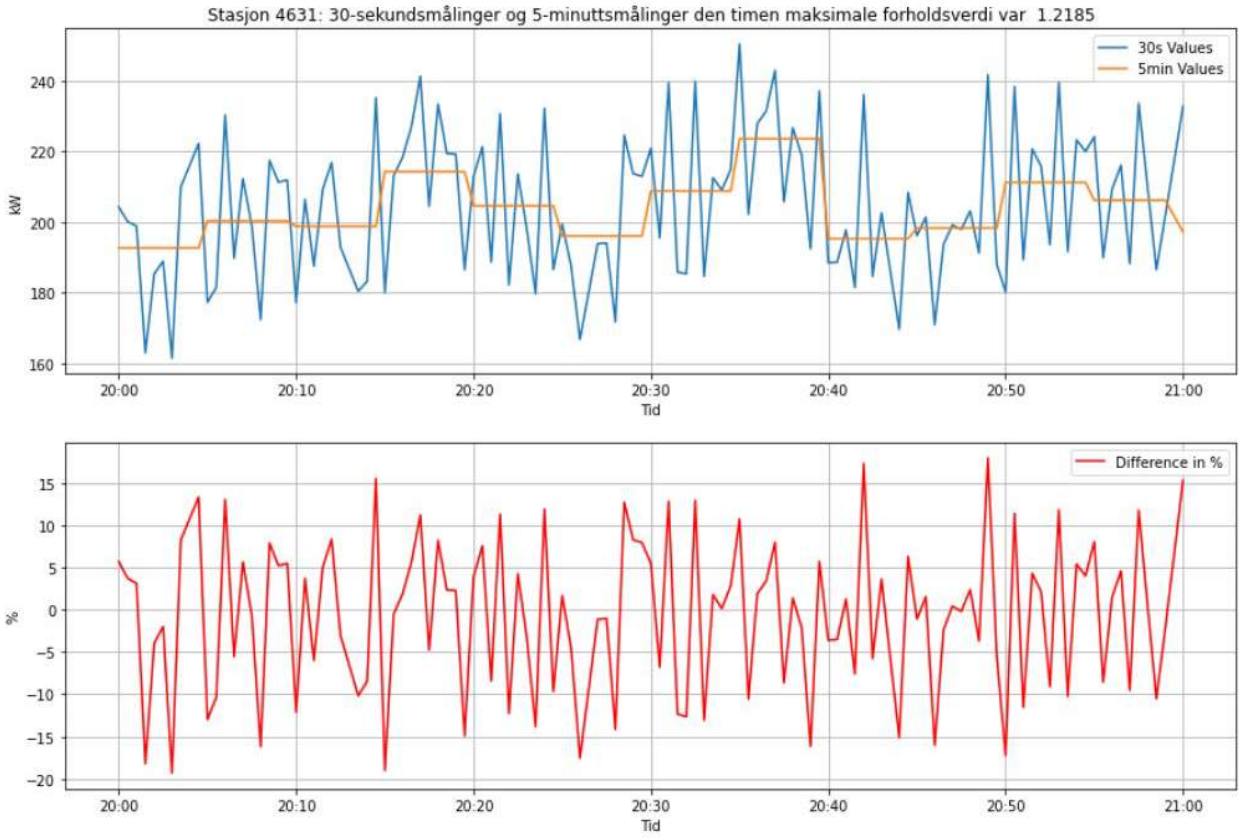


Figur 4: Scatterplott av de 10% høyeste forholdsverdier for stasjonene 4631 og 5198

Figur 4 viser at de 10% høyeste forholdsverdier for stasjon 4631 er relativt like, og har omtrent likt stigningstall i månedene våren til høst. Det kan også legges merke til at det er flest ekstremalverdier om våren, der enkelte verdier går forbi 1.5. Når det gjelder stasjon 5198, kan det ses at månedene våren til høst er relativt like når det kommer til 10% høyeste forholdsverdiene. Det kan også merkes at grafene som illustrerer de ulike månedene synker mye raskere enn stasjon 4631. Den største forskjellen som både denne figuren, og figur 1 viser, er at stasjon 5198 har både langt flere og høyere ekstremalverdier enn stasjon 4631.

Videre ble det funnet konkrete eksempler på forskjellene mellom P_{30s} og P_{5m} i de periodene forholdsverdiene var høye (>1). Dette ble gjort ved å først finne de høye forholdsverdiene, for å

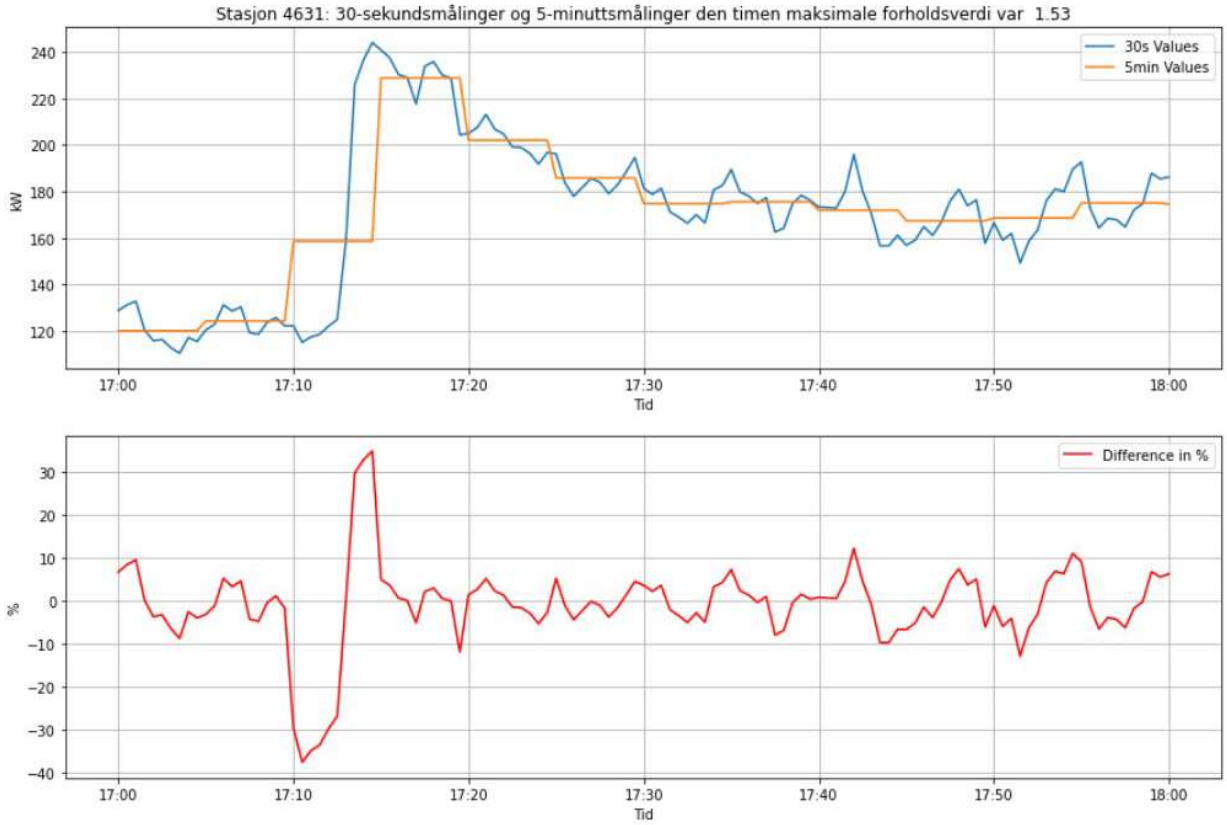
så se på nøyaktig klokkeslett der forholdsverdiene var svært høye. Hele timen som inneholdt den høye forholdsverdien ble plottet i en linjegraf, sammen med en graf som viser differanse i prosent.



Figur 5: P_{30s} og P_{5m} for stasjon 4631 den timen den maksimale forholdsverdien var 1.2185, samt differansen mellom effektfrekvensene

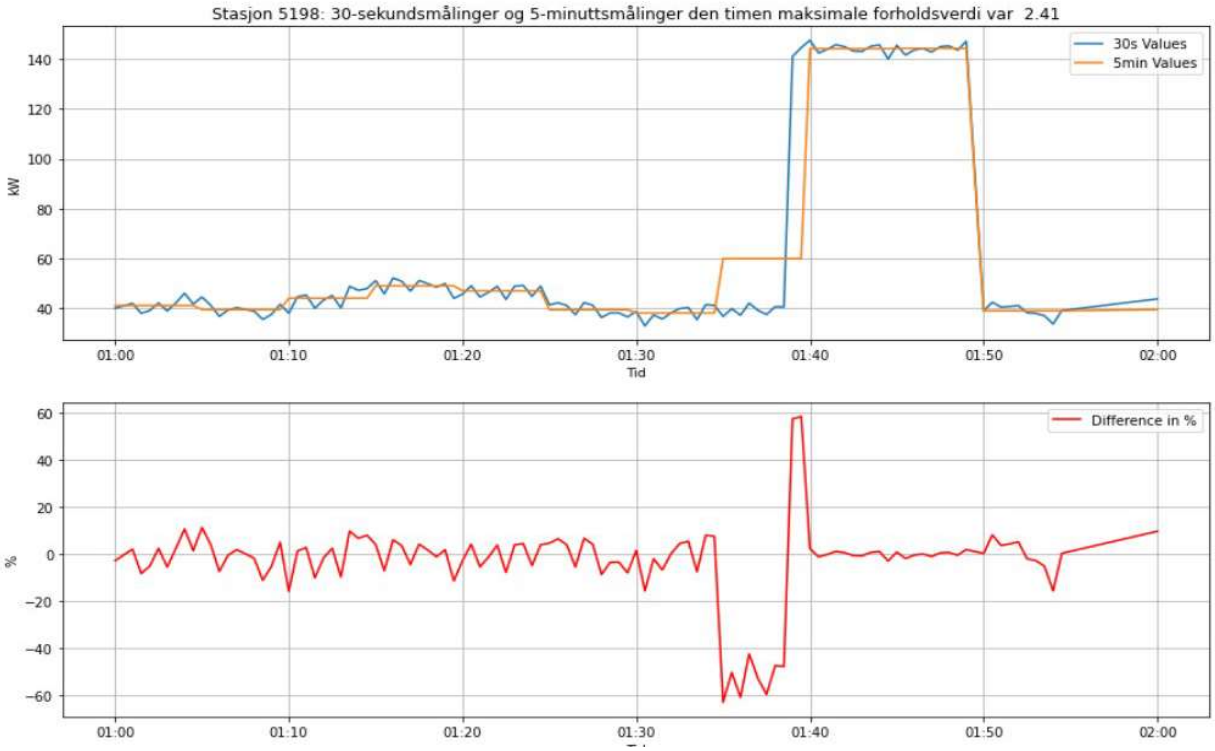
Figur 5 viser variasjon av P_{30s} og P_{5m} for stasjon 4631. Som forventet, er det lite oscillasjoner i grafen som viser P_{5m}, og mye oscillasjoner i grafen som viser P_{30s}. Som vi ser av den nederste grafen som viser differansen i prosent mellom effektfrekvensene, er forskjellen mellom effektene på omtrent 20%.

Det ble også sett på et eksempel der det er både rask og stor endring i effektuttak for stasjon 4631. Dette ble gjort for å undersøke informasjonstap dersom kun lavfrekvente effektmålinger blir benyttet. I eksemplet som gjelder, har det blitt benyttet samme metode som beskrevet ovenfor, og den høyeste forholdsverdien den aktuelle timen var 1.53. Som det kan ses av figur 6, oppstår det en rask effektendring mellom kl. 17:10 og 17:15, der effekthoppet er på over 100 kW. I perioden som er aktuell, kan det ses av den røde grafen at forskjellen mellom effektverdiene er på 39%. Dette kan sammenlignes med figur 5, der P_{5m} holder en relativt stabil verdi gjennom hele timen, og differansen varierer med omtrent 20%.

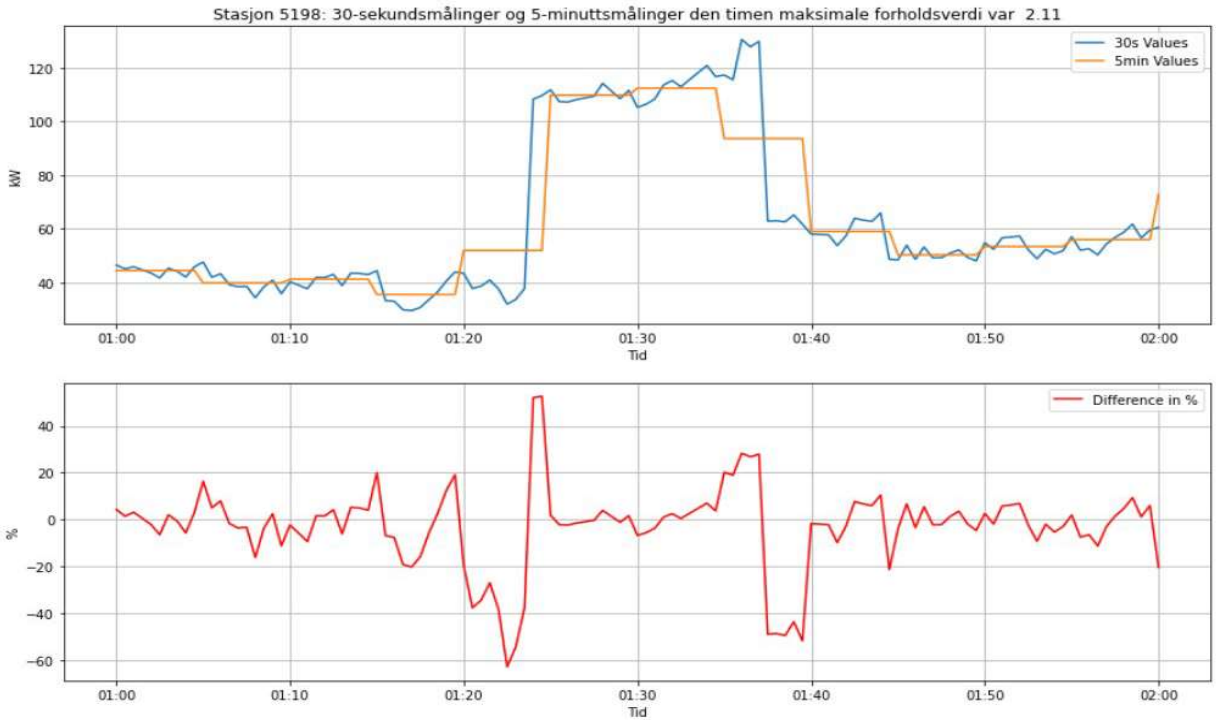


Figur 6: P_{30s} og P_{5m} for stasjon 4631 den timen den maksimale forholdsverdien var 1.53, samt differansen mellom effektfrekvensene

Når det gjelder stasjon 5198, ble det funnet to eksempler som illustrer hvordan P_{30s} er i forhold til P_{5m} når det oppstår en rask endring i effekt, samtidig som eksemplene illustrer hvor liten forskjell det er mellom de to målefrekvensene. Av figur 7 kan det ses at P_{30s} oscillerer rundt P_{5m}, men disse oscillasjonene kan oppfattes som støy. Videre kan det ses at kl. 01:35, ligger P_{5m} omtrent 20 kW over P_{30s}, før det plutselig oppstår stort effektuttak, og P_{30s} øker med 100 kW. Av grafen kan det ses at det er en forsinkelse på omtrent 1 min før P_{5m} stiger til samme nivå som P_{30s}, og derfra ligger grafene på hverandre. I det andre eksempelet, vist i figur 8, ser vi samme trend som i figur 7; når P_{5m} er stabil, vil P_{30s} ligge på P_{5m}, med små oscillasjoner som kan oppfattes som støy. Nå det oppstår en rask og stor endring i P_{30s}, vil P_{5m} følge denne endringen, men med en liten forsinkelse. Det som kan merkes i figur 8, er at når P_{30s} stuper ned, tar det lengere tid for P_{5m} å ta igjen.

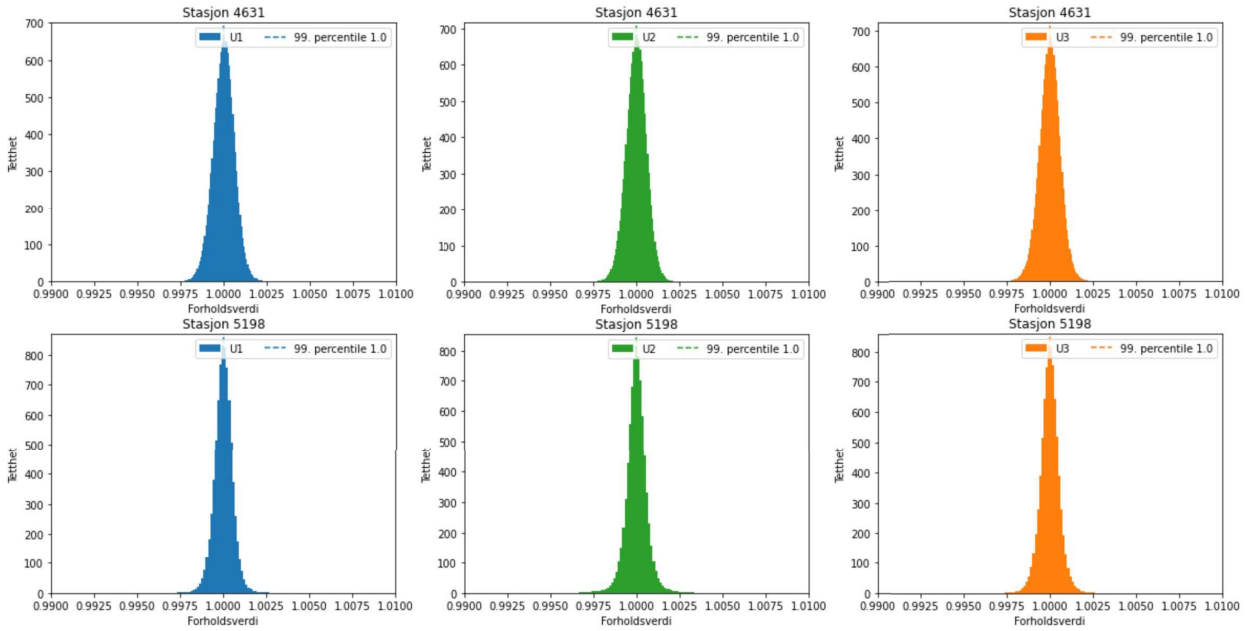


Figur 7: P_{30s} og P_{5m} for stasjon 5198 den timen den maksimale forholdsverdien var 2.41, samt differansen mellom effektfrekvensene



Figur 8: P30s og P5m for stasjon 5198 den timen den maksimale forholdsverdien var 2.11, samt differansen mellom effektfrekvensene

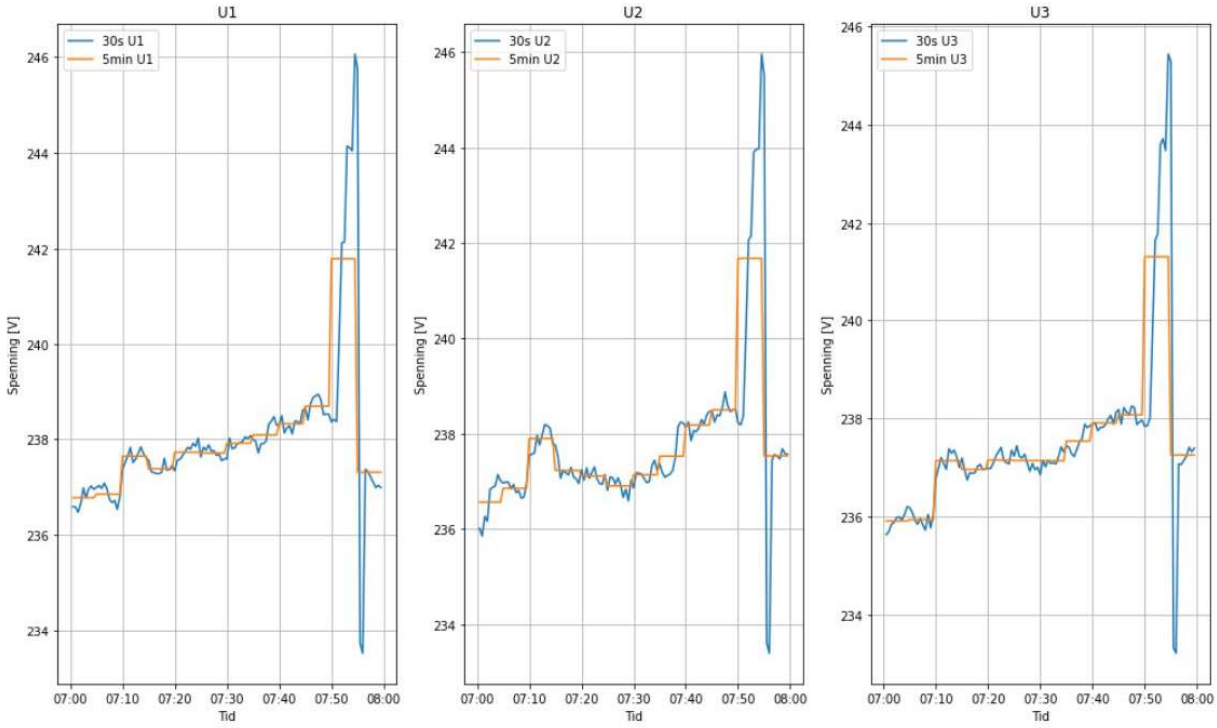
4 Analyse av spenning



Figur 9: Fordeling for U_{30s}/U_{5m} for alle tre spenningene, begge stasjoner

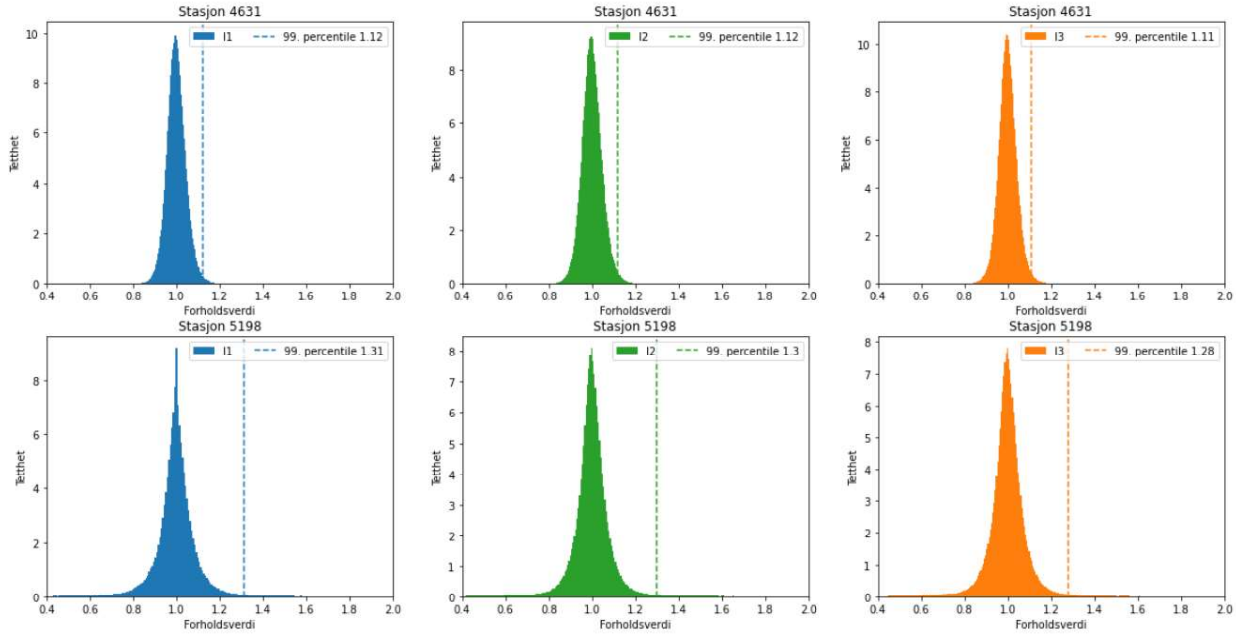
Figur 9 viser fordeling av forholdsverdiene U_{30s}/U_{5m} for alle tre spenningene, for begge stasjonene. Som det kan ses av figuren, ligger forholdstallene rundt 1, som betyr at i de fleste tilfellene, er $U_{30s} = U_{5m}$. Ettersom spenningsmålingene blir tatt i transformatoren, vil det være veldig lite variasjon i spenningsverdier, uavhengig av målefrekvens. Dette er på grunn av omsetningsforholdet på transformatoren – uansett spenningen på primærsiden, vil sekundærsiden alltid prøve å gi 230V. Sekundærspenningen kan ha små variasjoner som følge av lastendring, men denne variasjonen er neglisjerbar, spesielt i godt kompenserte nettverk.

Figur 10 er tatt med for å illustrere forskjellen mellom U_{30s} og U_{5m} når forholds-verdien, U_{30s}/U_{5m} er høy. Som det kan ses, er spenningsverdiene målt med de ulike målefrekvensene så å si like, bortsett fra når det oppstår et sprang, noe som medfører spenningsforskjell på omtrent 4 V.



Figur 10: U_{30s} og U_{5m} den timen den maksimale forholds-verdien var 1.0356

5 Analyse av strøm

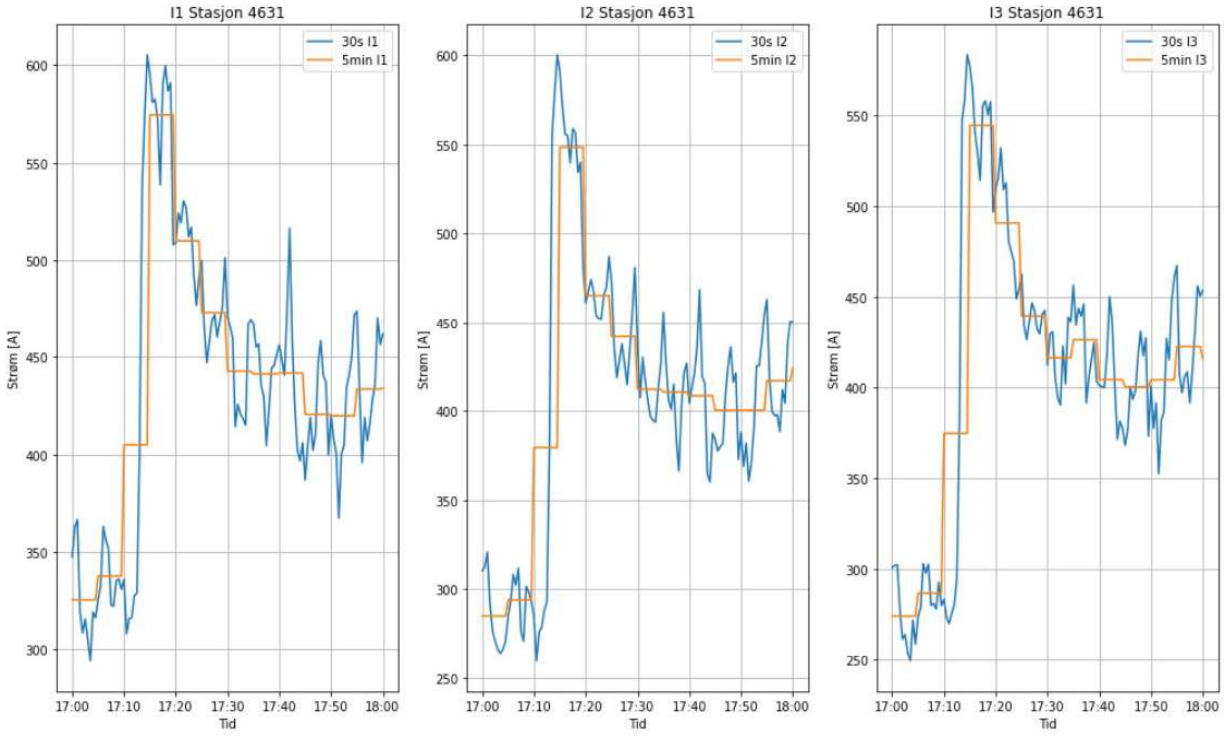


Figur 11: Fordeling for I_{30s}/I_{5m} for alle tre fasestrømmene, begge stasjoner

Figur 11 viser fordelingen av I_{30s}/I_{5m} for alle fasestrømmene for begge stasjonene. Ettersom strøm varierer med effekt, kan det ses større variasjon i figur 11 enn i figur 9. PC_{99} for stasjon 4631

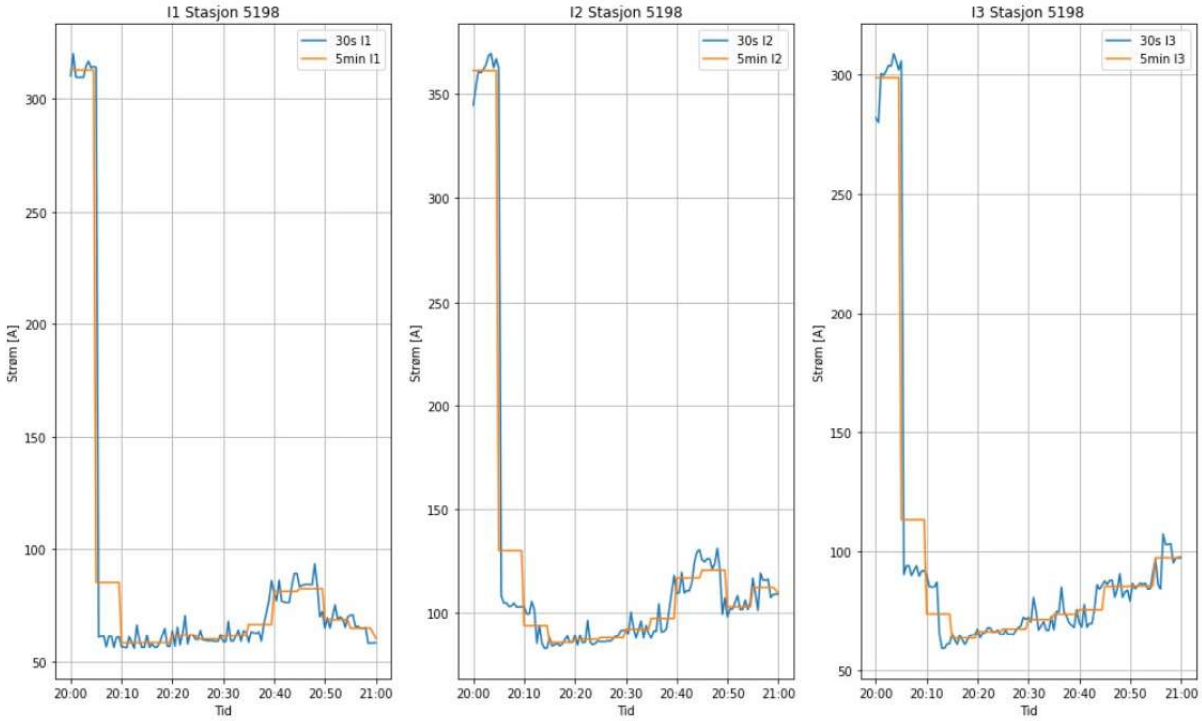
varierer mellom 1.11 og 1.12, mens for stasjon 5198 vil den variere mellom 1.28 og 1.31. Dette er på grunn av større effektvariasjoner som stasjon 5198 er utsatt for. Ellers kan det merkes at fordelingene for hver fase innen hver stasjon er tilnærmet identisk.

Figur 12 viser I_{30s} og I_{5m} for stasjon 4631 den timen den maksimale forholds-verdien var 1.495. Her kan man se at selv om I_{30s} oscillerer rundt I_{5m}, er det veldig liten forskjell mellom disse to målefrekvensene. De gangene forskjellen mellom målefrekvensene er ~100 A, varer ikke denne forskjellen i mer enn 30 sekunder.

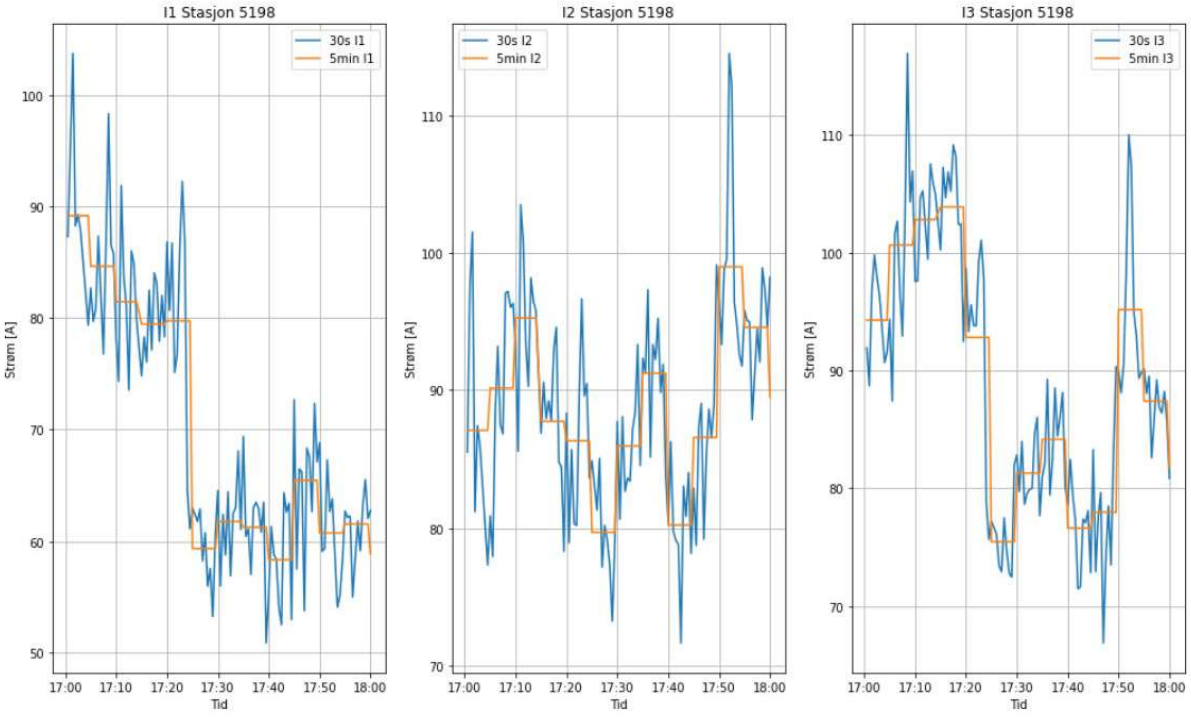


Figur 12: I_{30s} og I_{5m} den timen den maksimale forholds-verdien var 1.495, stasjon 4631

Figur 13 viser I_{30s} og I_{5m} for stasjon 5198 den timen den maksimale forholds-verdien var 3.678. Her ligger begge målefrekvensene opp på hverandre, slik at forskjellen mellom disse er minimal. Figur 14 viser I_{30s} og I_{5m} for stasjon 5198 i et annet tidsintervall, og er tatt med for å vise at det kan oppstå større forskjeller mellom målefrekvensene. Som nevnt tidligere, vil strømmen variere med effekt, slik at forskjellen mellom høyfrekvent og lavfrekvent strøm være tilnærmet tilfeldig.

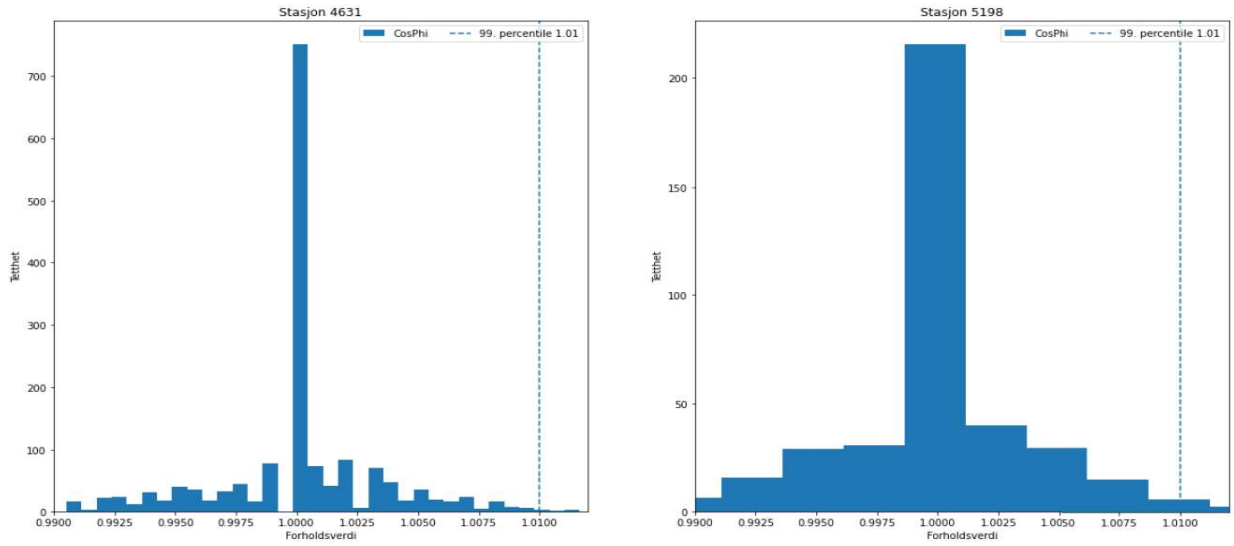


Figur 13: I_{30s} og I_{5m} den timen den maksimale forholds-verdien var 3.678, stasjon 5198



Figur 14: Illustrasjon på at forskjellen mellom I_{30s} og I_{5m} kan være svært ulik, med referanse til figur 13.

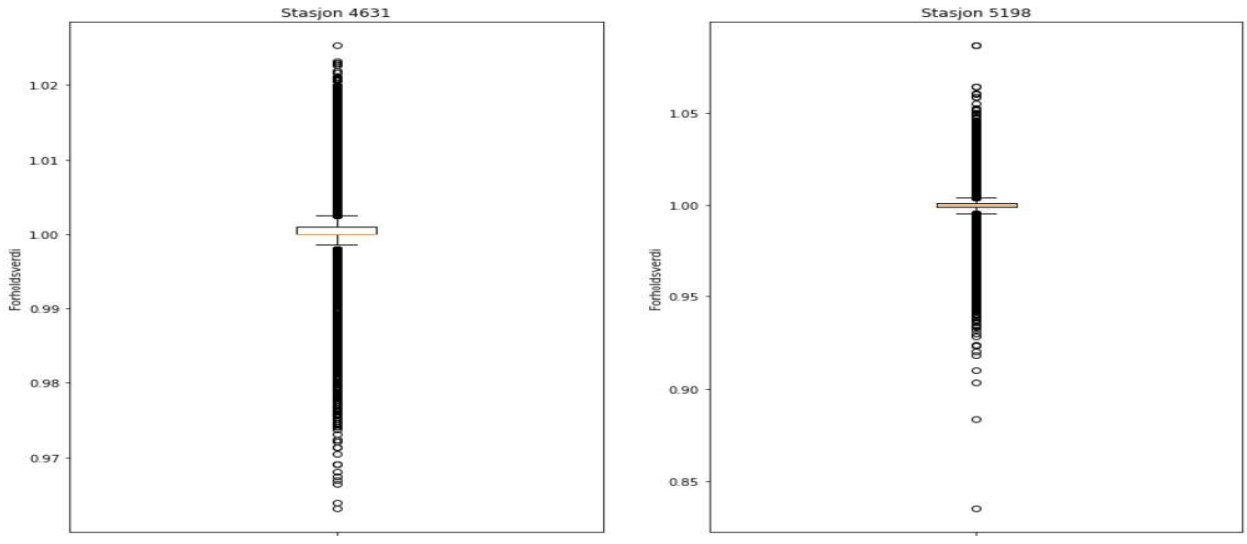
6 Analyse av effektfaktor



Figur 15: Fordeling av Phi_{30s}/Phi_{5min} for begge stasjonene

Figur 15 viser fordeling av forholdsverdiene Phi_{30s}/Phi_{5min} for begge stasjonene, og illustrerer hvor liten forskjell det er mellom de to målefrekvensene. Som persentillinjene illustrerer, er 99% av verdiene vist i figur 15 mindre eller lik 1.01 for begge stasjonene. Det kan også legges merke til at stasjon 4631 har mange flere forholdsverdier som er lik 1. Grunnen for dette er at stasjon 4631 forsyner flere kunder, mens stasjon 5198 bare to.

Figur 16 viser boxplot av Phi_{30s}/Phi_{5min} og bekrefter det som er vist i figur 15 – det er svært liten forskjell mellom målefrekvensene.



Figur 16: Boxplot av Phi_{30s}/Phi_{5min} for begge stasjonene

7 Resultater

7.1 Effekt

Informasjon som standardavvik, varians og PC_{99} presentert i figur 1 er oppsummert i tabell 5, slik at det blir mer oversiktlig:

Tabell 5: Grunnleggende statistikk informasjon presentert i figur 1

Stasjon	**5198**			**4631**		
Periode	Standardavvik	Varians	PC_{99}	Standardavvik	Varians	PC_{99}
Total	0.1	0.01	1.2931	0.0436	0.0019	1.1145
Vinter	0.1285	0.0165	1.46	0.0329	0.00108	1.08
Vår	0.099	0.0098	1.3123	0.049	0.0024	1.1245
Sommer	0.083	0.0069	1.223	0.0498	0.00248	1.13
Høst	0.09	0.0081	1.23	0.04	0.0016	1.105

Av tallene presentert i tabell 5, kan det ses at det er svært liten variasjon når man ser på én stasjon om gangen. For stasjon 5198, er variansen i fordelingskurven størst om vinteren, mens for stasjon 4631 er variansen størst om sommeren. Videre kan det merkes at PC_{99} for stasjon 5198 er svært lik i månedene sommer-høst og vinter-våren, men for stasjon 4631, er persentilene svært like gjennom hele året.

Tabell 6 viser antall og prosentdel ekstremalverdier hver måned for hver stasjon. Tabellen viser at stasjon 4631 har flest ekstremalverdier i månedene mars-oktober, mens stasjon 5198 har flest ekstremalverdier i vintermånedene. Videre kan det merkes at selv om enkelte måneder har noen tusener ekstremalverdier, utgjør disse verdiene en svært liten prosentdel av hele datasettet. Prosentandelene er i intervallet [0.03, 0.328], og det er stasjon 4631 som har størst prosentandel ekstremalverdier (0.328%) i august.

Tabell 6: Antall og prosentandel ekstremalverdier hver måned

Årstid	Måned	Antall ekstremalverdier totalt		Antall ekstremalverdier [%]	
		Stasjon			
		4631	5198	4631	5198
Vinter	Desember	415	2478	0.0375	0.22
	Januar	107	2316	0.0097	0.206
	Februar	774	2095	0.07	0.186
Vår	Mars	2441	1820	0.221	0.1617
	April	1021	1225	0.092	0.1088
	Mai	2568	350	0.232	0.031
Sommer	Juni	1788	111	0.162	0.009
	Juli	1479	368	0.134	0.032
	August	3622	1159	0.328	0.103
Høst	September	1539	379	0.139	0.0337
	Oktober	2173	1090	0.1967	0.0968
	November	428	1582	0.03875	0.14

7.2 Spenning

Informasjon som standardavvik, varians og PC_{99} for hver fasespenning presentert i figur 9 er oppsummert i tabell 7, slik at det blir mer oversiktlig:

Tabell 7: Grunnleggende statistisk informasjon presentert i figur 9

Stasjon	Fase	Standardavvik	Varians	PC_{99}
4631	U1	0.000657	~0	1
	U2	0.000635	~0	1
	U3	0.000646	~0	1
5198	U1	0.000572	~0	1
	U2	0.000640	~0	1
	U3	0.000585	~0	1

Informasjon presentert i både figur 9 og tabell 7 bekrefter at forskjellen mellom målefrekvensene er neglisjerbar.

7.3 Strøm

Informasjon som standardavvik, varians og PC_{99} for hver fasestrøm presentert i figur 11 er oppsummert i tabell 8, slik at det blir mer oversiktlig:

Tabell 8: Grunnleggende statistisk informasjon presentert i figur 11

Stasjon	Fase	Standardavvik	Varians	PC_{99}
4631	I1	0.04516	0.002	1.12
	I2	0.04819	0.0023	1.12
	I3	0.04399	0.0019	1.11
5198	I1	0.104	0.0108	1.31
	I2	0.1	0.01	1.30
	I3	0.0956	0.009	1.28

Som tabell 8 presenterer, er det relativt lite variasjon i fasestrømmene i hver av stasjonene. PC_{99} kolonnen viser at forholdet mellom målefrekvensene i de fleste tilfellene er mindre eller lik 1.12, noe som er akseptabelt, mens for stasjon 5198, kan forskjellen mellom målefrekvensene være opp mot 30%

7.4 Effektfaktor

Informasjon som standardavvik, varians og PC_{99} for effektfaktor presentert i figur 15 er oppsummert i tabell 9, slik at det blir mer oversiktlig:

Tabell 9: Grunnleggende statistisk informasjon presentert i figur 15

Stasjon	Standardavvik	Varians	PC_{99}
4631	0.003479	~0	1.01
5198	0.004438	~0	1.01

Tallene presentert i tabell 9, sammen med figur 15, viser ingen grunn til å anta betydelig forskjell mellom målefrekvensene.

8 Konklusjon

Ettersom analysen ble utført på kun to nettstasjoner, som begge forsyner ulikt antall kunder, ble resultatene for disse nettstasjonene forskjellig. Ved å utføre analyse på effekt- spenning-, strøm- og effektfaktordata, har det blitt vist at den største forskjellen mellom målefrekvensene finner man i effekt- og strømdata, der den 99. persentil av forholdsverdiene kan variere fra 1.1 til 1.3. Forskjellen mellom målefrekvensene i spennings- og effektfaktoranalysen derimot, har blitt vist til å være neglisjerbar, ettersom 99. persentilen for U_{30s}/U_{5m} ble funnet til å være 1 for begge nettstasjonene, og Phi_{30s}/Phi_{5min} til 1.01 for begge nettstasjonene.

Selv om kun to nettstasjoner har blitt analysert i denne rapporten, der den ene nettstasjonen forsyner over dobbelt så mange kunder som den andre, har resultatene vist at det generelt sett er liten forskjell mellom 30 sekundsmålinger og 5 minuttsmålinger. Et utvalg av figurer presentert i analysekapitlet, har vist at det er ingen langvarige effekttopper med stor forskjell mellom målefrekvensene. Med denne informasjon, vil det være rimelig å konkludere med at forskjellen mellom målefrekvensene er for liten for å ta i bruk 30 sekundsmålinger istedenfor 5 minuttsmålinger i nettstasjoner.