Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Notat

Trafikkstatistikk for trikk

Notatnr SAMBA/26/12
Forfattere Magne Aldrin
Egil Ferkingstad
Ola Haug

Dato 18. juni 2012

Norsk Regnesentral

Norsk Regnesentral (NR) er en privat, uavhengig stiftelse som utfører oppdragsforskning for bedrifter og det offentlige i det norske og internasjonale markedet. NR ble etablert i 1952 og har kontorer i Kristen Nygaards hus ved Universitetet i Oslo. NR er et av Europas største miljøer innen anvendt statistisk-matematisk modellering og har et senter for forskningsdrevet innovasjon, Statistics for Innovation – $(sfi)^2$, med finansiering fra Norges forskningsråd. Det jobbes med et bredt spekter av problemstillinger, for eksempel finansiell risiko, jordobservasjon, estimering av fiskebestander og beskrivelse av geologien i petroleumsreservoarer. NR er ledende i Norge innen utvalgte deler av informasjons- og kommunikasjonsteknologi. Innen IKT-området har NR innsatsområdene e-inkludering, informasjonssikkerhet og smarte informasjonssystemer.

NRs visjon er forskningsresultater som brukes og synes.

Norsk Regnesentral
Norwegian Computing Center
Postboks 114, Blindern
NO-0314 Oslo, Norway

Besøksadresse
Office address
Gaustadalléen 23
NO-0373 Oslo, Norway

Telefon · telephone
(+47) 22 85 25 00
Telefaks · telefax
(+47) 22 69 76 60

Internett · internet
www.nr.no
E-post · e-mail
nr@nr.no

Tittel	**Trafikkstatistikk for trikk**
Forfattere	**Magne Aldrin** <magne.aldrin@nr.no> **Egil Ferkingstad** <egil.ferkingstad@nr.no> **Ola Haug** <ola.haug@nr.no>
Dato	18. juni 2012
Publikasjonsnummer	SAMBA/26/12

English summary

The Norwegian Computing Center has been engaged by Ruter AS, the common management company for public transport in Oslo and Akershus, Norway, to develop a method for estimating tram passengers on all trips based on passenger counts on a sample of trips. The method estimates or predicts the number of passengers entering and leaving the tram at every stop, and the number of passengers who stay on the tram between each stop is therefore implicitly estimated as well. All basic predictions are made at stop level on every single trip, and all types of aggregated quantities can therefore be calculated by summing over all single stops of interest, where real data are used when counts are available. Important aggregated quantities include the sum of passengers per year, per month and per day in week, separate for each tram and in total over all lines. Passenger growth from one year to another can also be calculated. All estimates are given with a quantified uncertainty.

The method consists of one model for the number of passengers entering the tram, and another model for passengers leaving the tram. In the latter, the number of passengers leaving the tram is modelled as a proportion of the passengers that stay on the train, which ensures consistency in the predictions. The basic structure of the models takes care of the systematic variation in the passenger data, with multiple seasonality over day, week and year, and typically with similarities between stops. The method is fitted to data from Oslo, but is general in nature, and thus potentially transferable to other regions and similar means of transport.

Emneord	
Målgruppe	Ruter
Tilgjengelighet	Åpen
Prosjekt	TrikkStat
Prosjektnummer	220530
Satsningsområde	Teknologi, industri og forvaltning
Antall sider	40

Innhold

1 Innledning

Dette notatet beskriver NRs arbeid med en statistisk metodikk for analyse av trafikktellinger for trikken i Oslo. Hensikten med prosjektet er å utarbeide en prototype for en metodikk for å kunne beregne antall påstigende og avstigende passasjerer for ethvert stopp, samt last mellom hvert stopp. Dette kan deretter aggregeres opp til blant annet tur/strekning, dag, måned eller år. Per i dag er maskinelt telleutstyr installert i om lag halvparten av trikkene. Hovedoppgaven er dermed å beregne antall på- og avstigende passasjerer for stopp på de turer hvor det ikke foreligger maskinelle tellinger. kapittel 2 gir først en gjennomgang av ulike aspekter ved selve dataene, og deretter presenteres en oversikt over metodikken i kapittel 3. Eksempler på resultater av metodikken (i form av ulike tabeller og figurer) gis i kapittel 4. Appendiks A gir til slutt en detaljert beskrivelse av metodikk og algoritmer.

2 Datakvalitet og datahåndtering

2.1 Datakilde

Datamaterialet til prosjektet er hentet fra Trafikantens sanntids-informasjonssystem (SIS). Tallene i denne databasen er en kvalitetssikret dump fra balanseringsproduktet i Mobile Statistics levert av INIT (Innovation in Traffic Systems AG). Data er levert for trikkelinjene 11, 12, 13, 17, 18 og 19 for årene 2009 - 2011. Grunnet uregelmessigheter og avvik i trikketrafikken i forbindelse med omleggingen av Jernbanetorget, er analysene begrenset til perioden fra 26. april 2009 til 31. desember 2011.

Enkelte av trikkene har installert automatisk telleutstyr som registrerer passasjerflyten gjennom hver enkelt tur. Per september 2011 finnes slikt utstyr i 35 av 70 trikker. Totalt gjennom analyseperioden er telledekningen på om lag 31% (av totalt 968083 turer er 302596 turer med tellinger).

2.2 Dataformater

Dataene fra Trafikanten er levert som rene tekstfiler (ASCII) med TAB som skilletegn. Datamengden er fordelt på separate filer ut fra trikkelinje og retning. Hver fil inneholder all loggført aktivitet på Trikkens rutegående vogner.

Originalfilene fra Trafikanten er organisert som matriser hvor hver rad angir en av hendelsene: "stopp på holdeplass", "passering av holdeplass uten å stoppe" eller "stopp og åpning av dør utenfor holdeplass". For hver rad er det via ulike kolonner registrert en rekke parametre så som posisjon, kjøretøynummer og avgangstid. Utfra originalfilene kan vi skille mellom trikketypene SL79 og SL95 og buss for trikk, men ikke mellom busstyper. I vogner med telleutstyr registreres passasjerflyten som antall påstigende og avstigende per stopp.

NR har funnet det hensiktsmessig å organisere dataene på et nytt matriseformat hvor radene representerer enkeltturer. Informasjon som er felles for alle holdeplasser langs turen er satt av via et fast kolonnesett lengst mot venstre i matrisa, og så følger kolonner med spesifikke tall for hver enkelt holdeplass langs den aktuelle trikkelinja og retningen. Med holdeplasser som er spesifikke for den enkelte trikkelinje og retning, innebærer dette at dataene for hver linje og retning er samlet på separate filer. De spesifikke holdeplasskolonnene er organisert slik at de fra venstre mot høyre representerer holdeplasser som alltid ligger framover i avgangstid.

2.3 Feil og mangler i rådata

Feil og mangler i datamaterialet kommer både som en følge av manuelle (menneskelige) og maskinelle feil. Selv om mye av prosessen rundt trafikkavviklingen for trikkene er styrt av maskinelle rutiner, er det fortsatt mennesker som styrer og overvåker de beslutninger som tas. Spesielt er den menneskelige faktoren viktig i avvikssituasjoner. I tillegg er linjenettet/infrastrukturen lagt opp slik at på enkelte steder går linjer over i hverandre. Dette introduserer feil i forhold til hvilken linje av- og påstigende blir registrert på, og føreren har ingen mulighet til å påvirke dette.

Det automatiske telleutstyret som er plassert i deler av vognparken, logger antall på- og avstigende passasjerer på alle stopp langs ruta hvor dørene åpnes. På bakgrunn av dette beregnes en trafikklast som angir hvor mange passasjerer som til enhver tid befinner seg ombord. Registreringen av på- og avstigende passasjerer skjer gjennom sensorer som er plassert i dørene. Trikken har kontrollert tellingene på mange vogner og funnet få avvik. Det kan likevel være enkelte feil på sensorer/tellecomputer. Den interne balanseringsalgoritmen som er bygget inn i Mobile Statistics søker å rette på dette gjennom å fordele passasjeravviket utover stoppene langs ruta etter nærmere angitte kriterier. Algoritmen gir likevel ikke konsistente passasjertall, og NR har derfor utviklet en supplerende metode for beregning av balanserte lastdata, se avsnitt 2.5.

Nedenfor følger eksempler på feil- og avvikssituasjoner som vi har observert i dataene:

- Urimelige odometerverdier: Det er et eksempel fra linje 11 (TRIP-ID=592655415) på at odometeret har registrert en tilbakelagt strekning på 101 meter i løpet av 4 sekunder. En mulig forklaring på en slik åpenbar feil er ifølge Trikken at registreringshjulet spinner ved glatt underlag.
- For enkelte delstrekninger langs en gitt linje og retning, hvor det ikke er mulighet for å snu eller kjøre av, og hvor det ikke forekommer start eller stopp av turer annet enn i ytterholdeplassene av delstrekningen, skal i teorien totalt antall stopp eller passeringer av holdeplasser være like. Dette stemmer ikke alltid. Et eksempel på en slik delstrekning er holdeplassene Schultz' gate, Rosenborg og Homansbyen på linje 11. For trafikk i retning 2 (mot Majorstua) er det totalt i dataperioden registrert 65087 stopp og passeringer på Homansbyen, 66562 på Rosenborg og 66522 på Schultz' gate. Mulige forklaringer på disse forskjellene kan ifølge Trikken være at turen av ulike grunner må avbrytes innenfor delstrekningen, eller at det oppstår irregulariteter i SIS-systemet.
- Dupliserte holdeplassdata i SIS-filene fra Trafikanten: Dette tilskrives feil i plandataene som så er videreført til SIS-basen. Avviket er håndtert ved å utelate den ene av registreringene.
- Stopp utenfor holdeplass: Mange av stoppene med åpning av dør utenfor holdeplass skjer langt fra ordinær holdeplass både i avstand og tid. Dette forekommer typisk når en tur fortsetter med stopp etter endeholdeplass, eventuelt når det registreres stopp før turens første ordinære holdeplass. Et eksempel på førstnevnte er fra linje 11 den 30. april 2009 (TRIP-ID=650209078). Her fortsetter turen etter siste ordinære holdeplass som er Disen, og tilbakelegger ifølge odometeret en strekning på vel 26 kilometer over en periode på omlag to timer før turen avsluttes.
- Turer som ikke loggføres. Av ukjente årsaker har det hendt at faktisk kjørte turer – både med og uten telleutstyr – ikke har blitt logget av SIS. Eksempler på dette har vi i uke 4 og 5 2011. Se også tekst til tabell 3 og figur 18.

I noen få tilfeller gir de automatiske tellingene opplagt urealistisk mange passasjerer på trikken (ekstrem last). For trikketype SL79 er det f.eks. to turer med maksimal last på henholdsvis 355 og 400, og for SL95 er det to turer med maksimal last på henholdsvis 1164 og 297. I samråd med

Ruter er det satt grenser for maksimal last på henholdsvis 170, 220, og 120 for henholdsvis SL79, SL95 og leddbuss. Dette tilsvarer 4,5 stående pr kvm. Turer med minst én holdeplassregistrering over grensen for maksimal last settes til "ikke talt". Disse utgjør kun 40 av totalt 968038 turer, dvs 0.004%. Merk at beregningen av maksimal last utføres etter kjøring av balanseringsalgoritmen beskrevet i kapitlene 2.5 og A.1.

2.4 Definisjon av holdeplasser

For hver trikkelinje og retning kjøres det i utgangspunktet etter et forhåndsdefinert sett av holdeplasser. Det forekommer imidlertid en del avvik fra denne normalsituasjonen. Disse kan være planlagt ved at enkelte turer har et alternativt startpunkt eller følger en annen trase, for eksempel på bestemte ukedager eller i forbindelse med spesielle arrangementer. I andre situasjoner kan det være behov for at trikken fraviker sitt planlagte kjøremønster på grunn av akutte hendelser i trafikkbildet. Dette kan være omdirigering etter ulykker eller forsinkelser, eller for å omgå kjøretøy som sperrer skinnegangen. I forlengelsen av dette har vi perioder med stengning av skinnegangen, for eksempel i forbindelse med vedlikeholdsarbeid. I slike tilfeller settes det gjerne inn erstatningsbusser for trikken. Disse følger ikke nødvendigvis trikkens trase men trafikkerer i stedet nærliggende holdeplasser for buss.

Enkelte holdeplasser er plassert slik i forhold til hverandre at det er naturlig å betrakte dem som en fysisk enhet. Dette er typisk holdeplasser i sentrum med kort avstand til hverandre. For eksempel slås ulike holdeplasser på og i nærheten av Stortorvet sammen til én samleholdeplass, og det samme er tilfellet med holdeplasser på Jernbanetorget og i Dronningens gate. Tanken bak sammenslåingen er at det fra den reisendes ståsted ikke spiller noen rolle hvilken av holdeplassene den aktuelle avgangen benytter seg av. Den reisende vil uansett gå på eller av trikken ved den holdeplassen som blir brukt på den faktiske avgangen. Vurderingen av hvilke holdeplasser som skal slås sammen er gjort av Ruter.

I dataene som analyseres inngår holdeplasser både fra normal- og avvikssituasjoner. Holdeplasser med registrerte stopp, men som ikke anses relevante for passasjertrafikken langs en linje og retning, er fjernet (for eksempel holdeplass Grefsen vognhall på linje 13). Slike holdeplasser er identifisert av Ruter.

I tillegg er passasjerflyten på stopp hvor døren åpnes utenfor holdeplass (STOPTYPE = 2) tatt med i datasettet dersom stoppet skjer nærmere enn 50 meter fra en reell holdeplass. Bakgrunnen for dette er todelt. For det første kan trikken ved stopp på holdeplass plassere seg på en slik måte i forhold til utstrekningen av holdeplassen at stoppet blir registrert som utenfor holdeplass. I andre tilfeller kan en holdeplass være midlertidig flyttet. I begge situasjoner ønsker vi å fange opp passasjerflyten som om trikken hadde stoppet innenfor utstrekningen av den ordinære holdeplassen. Dette gjøres ved å legge til eller flytte på- og avstigninger fra stoppet med STOPTYPE = 2 til det ordinære stoppet.

2.5 Balansering av på- og avstigende

For en virkelig trikketur vil påstigende, avstigende og last ved avgang nødvendigvis være konsistente i følgende forstand:

1. Last ved avgang fra første holdeplass er lik påstigende på første holdeplass.
2. Antall avstigende på holdeplass $i > 1$ kan ikke være større enn last ved avgang fra holdeplass $i - 1$.
3. Last ved avgang fra holdeplass $i > 1$ er lik last ved avgang fra forrige holdeplass $i - 1$ pluss differansen mellom antall på- og avstigende på holdeplass i.

4. Antall avstigende på siste holdeplass er lik last ved avgang fra nest siste holdeplass. Antall påstigende på siste holdeplass er null.

Selv om dataene levert fra SIS-systemet skulle vært balansert, så tilfredstiller de ikke disse logiske konsistenskriteriene, og vi har derfor sørget for å gjøre dataene konsistente ved bruk av en supplerende balanseringsalgoritme som bygger på kriteriene over. Noen resultater av denne algoritmen er vist i tabell 1, og algoritmen er beskrevet i appendiks A.1.

	11, 1	11, 2	12, 1	12, 2	13, 1	13, 2	17, 1	17, 2	18, 1	18, 2	19, 1	19, 2	Totalt
A	20.6	20.8	18.3	13.6	12.8	16.4	17.6	16.3	17.7	17.3	21.0	19.4	17.6
B	100.4	99.6	99.3	99.1	99.7	100.7	101.5	98.3	99.8	100.3	100.2	101.2	99.9
C	98.5	98.1	98.5	100.6	99.3	98.6	96.4	100.4	98.1	98.5	94.8	95.8	98.3
D	98.9	97.7	97.8	99.7	99.0	99.2	97.8	98.6	97.9	98.8	95.0	96.9	98.2

Tabell 1. Resulater av balanseringsalgoritme for hver linje/retning og totalt, for hele dataperioden. Rad A: prosent turer der justering er nødvendig; Rad B: ujustert sum avstigende som prosent av ujustert sum påstigende; Rad C: justert sum avstigende som prosent av ujustert sum avstigende; Rad D: justert sum påstigende som prosent av ujustert sum påstigende.

2009-2	11,1	11,2	12,1	12,2	13,1	13,2	17,1	17,2	18,1	18,2	19,1	19,2	Totalt
A	20.82	19.56	15.83	12.15	7.18	14.98	30.78	27.20	18.27	19.22	19.98	24.63	19.68
B	100.91	99.40	99.86	99.19	99.83	101.73	103.05	94.11	100.58	100.62	100.55	103.39	100.10
C	97.52	97.86	96.91	100.42	99.59	97.49	94.46	104.80	96.79	98.02	91.06	91.09	97.51
D	98.41	97.26	96.77	99.60	99.42	99.18	97.34	98.63	97.35	98.63	91.56	94.17	97.61
2010-1	11,1	11,2	12,1	12,2	13,1	13,2	17,1	17,2	18,1	18,2	19,1	19,2	Totalt
A	16.61	19.56	16.87	11.50	6.93	12.27	16.19	16.24	14.30	17.85	20.06	24.55	16.03
B	100.75	99.51	99.66	99.18	99.98	100.38	102.46	96.89	100.18	100.24	100.35	103.24	100.11
C	98.03	97.77	97.08	100.43	99.11	98.63	95.42	102.34	97.71	98.98	91.87	92.20	97.68
D	98.77	97.29	96.75	99.60	99.09	99.01	97.78	99.16	97.89	99.21	92.19	95.19	97.79
2010-2	11,1	11,2	12,1	12,2	13,1	13,2	17,1	17,2	18,1	18,2	19,1	19,2	Totalt
A	15.63	18.82	16.82	12.22	17.94	13.51	9.34	10.46	14.54	15.10	19.39	17.28	15.07
B	100.73	99.52	99.07	99.40	101.95	99.88	99.73	99.65	99.61	99.85	100.16	100.71	99.95
C	98.12	98.16	98.85	100.29	95.38	99.10	98.76	99.46	98.67	99.51	95.56	96.54	98.34
D	98.85	97.70	97.92	99.68	97.24	98.98	98.49	99.12	98.28	99.36	95.71	97.23	98.29
2011-1	11,1	11,2	12,1	12,2	13,1	13,2	17,1	17,2	18,1	18,2	19,1	19,2	Totalt
A	15.79	20.64	13.25	11.02	6.33	16.70	7.03	5.26	10.45	6.79	18.61	9.94	11.91
B	100.52	99.75	99.60	99.18	99.95	100.78	99.84	99.65	99.57	99.75	99.96	99.69	99.83
C	98.59	98.01	99.28	100.64	99.58	98.68	98.83	99.33	99.12	99.55	97.33	98.87	99.05
D	99.10	97.76	98.88	99.81	99.53	99.45	98.67	98.98	98.70	99.29	97.30	98.57	98.88
2011-2	11,1	11,2	12,1	12,2	13,1	13,2	17,1	17,2	18,1	18,2	19,1	19,2	Totalt
A	32.48	25.73	30.00	21.55	25.06	26.61	19.85	18.53	29.12	25.72	26.24	18.26	24.73
B	99.53	99.72	98.57	98.92	98.09	101.04	99.08	100.79	98.63	100.55	100.01	99.51	99.47
C	99.68	98.31	100.00	100.72	101.44	98.26	98.84	97.08	98.86	97.70	97.39	99.22	99.10
D	99.21	98.04	98.57	99.63	99.50	99.28	97.93	97.85	97.50	98.24	97.40	98.73	98.57

Tabell 2. Resulater av balanseringsalgoritme for hver linje/retning og totalt, for hvert halvår separat. 2009-2 er andre halvår (jul-des) i 2009, 2010-1 første halvår (jan-jun) i 2010, osv. Rad A: prosent turer der justering er nødvendig; Rad B: ujustert sum avstigende som prosent av ujustert sum påstigende; Rad C: justert sum avstigende som prosent av ujustert sum avstigende; Rad D: justert sum påstigende som prosent av ujustert sum påstigende.

Som vi ser av tabell 1 er altså justering nødvendig for 17.6% av alle turer. Før justering er totalt antall avstigende passasjerer omtrent det samme som antall påstigende (99.9%), men for den enkelte linje/retning kan antall avstigende være opptil 1.5% (101.5−100) høyere eller 1.7% (98.3−100) lavere enn antall påstigende. Dette kan delvis være et resultat av at en trikk skifter fra en linje til en annen ved noen endeholdeplasser uten at trikken tømmes for passasjerer, men kan også skyldes problemer med telleutstyret. Justeringen har den positive bieffekt at en del tilfeller med urimelig stor last blir justert til mer rimelige verdier. Justeringen fører til at totalt antall avstigende og påstigende passasjerer blir justert ned med henholdsvis 1.7% og 1.8%. Dette er verdt å merke seg

når vi senere presenterer estimater for trafikkvolum med usikkerheter, for de usikkerhetsanslagene tar ikke hensyn til usikkerheten i selve tellingene. Tabell 2 viser tilsvarende resultater for hvert halvår separat.

3 Oversikt over metodikk

3.1 Hovedidéer

Metoden går ut på å estimere antall påstigende og avstigende passasjerer, samt last (antall passasjerer på trikken) for hver enkelt holdeplass på alle turer eller avganger der det ikke foreligger tellinger. Dette gjøres basert på data for talte turer. I tillegg trengs opplysninger om turene hvor det ikke foreligger tellinger, og som angir når og hvor trikken har kjørt, f.eks. når trikken har vært på den enkelte holdeplass på enhver tur. Alle akkumulerte størrelser, slik som antall passasjerer per time, døgn, uke eller år; eller per linje eller totalt over alle linjer; beregnes ved å summere over alle talte eller beregnede (kun for turer uten tellinger) passasjertall i perioden. Usikkerheten i estimatene kan også beregnes.

Passasjertrafikk er preget av systematisk variasjon. Det er sesongvariasjon over året som gjentar seg år for år. Begrepet sesongvariasjon brukes om variasjon som gjentar seg systematisk med en viss periode, som ikke trenger å være et år. Det er derfor også sesongvariasjon over ukedagene som i hovedsak gjentar seg uke for uke. Og til slutt er det sesongvariasjon over døgnet. Videre er det slik at ulike holdeplasser i hovedsak har det samme trafikkmønsteret. I tillegg til den systematiske variasjonen er det annen variasjon som er vanskelig å modellere, og som betraktes som tilfeldig variasjon. Et eksempel er variasjon i passasjertallene som skyldes værforhold.

På bakgrunn av dette har vi konstruert en statistisk modell hvor det legges vekt på å beskrive mest mulig av den systematiske variasjonen, slik at den tilfeldige, ikke-predikerbare variasjonen blir minst mulig. Hver holdeplass på en linje og retning modelleres for seg. Men fordi modellen inneholder mange parametre (ukjente størrelser) som må estimeres (tallfestes) fra dataene tar vi også hensyn til data fra andre holdeplasser langs samme linje og retning når vi estimerer modellen for den enkelte holdeplass. Dette kalles å låne styrke på tvers av holdeplassene.

Modellen tar hensyn til en mulig langtidstrend, en sesongeffekt over år, en sesongeffekt over uke og en sesongeffekt over døgn. Videre er det slik at sesongeffekten over uke varierer glatt over året. Det vil si at to påfølgende uker har ganske likt mønster, men ikke helt likt. Uker som ligger lenger fra hverandre, f. eks. en vinteruke og en sommeruke, kan ha mer ulikt trafikkmønster. Tilsvarende kan døgnvariasjonen variere glatt over året. To påfølgende mandager vil ha et ganske likt variasjonsmønster, mens to mandager som har noen måneders mellomrom kan være mer ulike. I tillegg tar modellen hensyn til spesialdager som påskedagene, fridager i mai eller dager omkring jula. For å modellere alt dette brukes nær 360 forklaringsvariable, som betyr at det er svært mange parametre som skal estimeres. Derfor bruker vi ikke data kun for en holdeplass av gangen, men alle data for en linje og retning simultant.

Det vi har beskrevet over er kun knyttet til tidspunkt for en tur. I tillegg tar vi hensyn til noen flere forklaringsvariable. Vi tar blant annet hensyn til om det har vært spesielt mange passasjerer på turene med tellinger før og etter den turen vi skal predikere for. Vi tar også hensyn til hvor mange holdeplasser det er igjen på turen. Det er viktig for holdeplasser på linjer hvor endeholdeplassen varierer. Ta Storo på linje 11, retning mot Kjelsås, som eksempel: hvis turen ender på Disen, som er neste holdeplass, vil få passasjerer gå på på Storo, men om trikken går til Kjelsås vil det typisk bli flere påstigende passasjerer på Storo. I prinsippet kan metoden også ta hensyn til forsinkelse

på den aktuelle turen, og om foregående tur var forsinket. Vi har imidlertid ikke tatt dette med i modellen vi presenter resultater for her. Grunnen er at vi ikke har fått undersøkt godt nok hvordan en best kan ta hensyn til forsinkelser, og dette er noe det med fordel kan arbeides noe mer med i eventuelt videre arbeid.

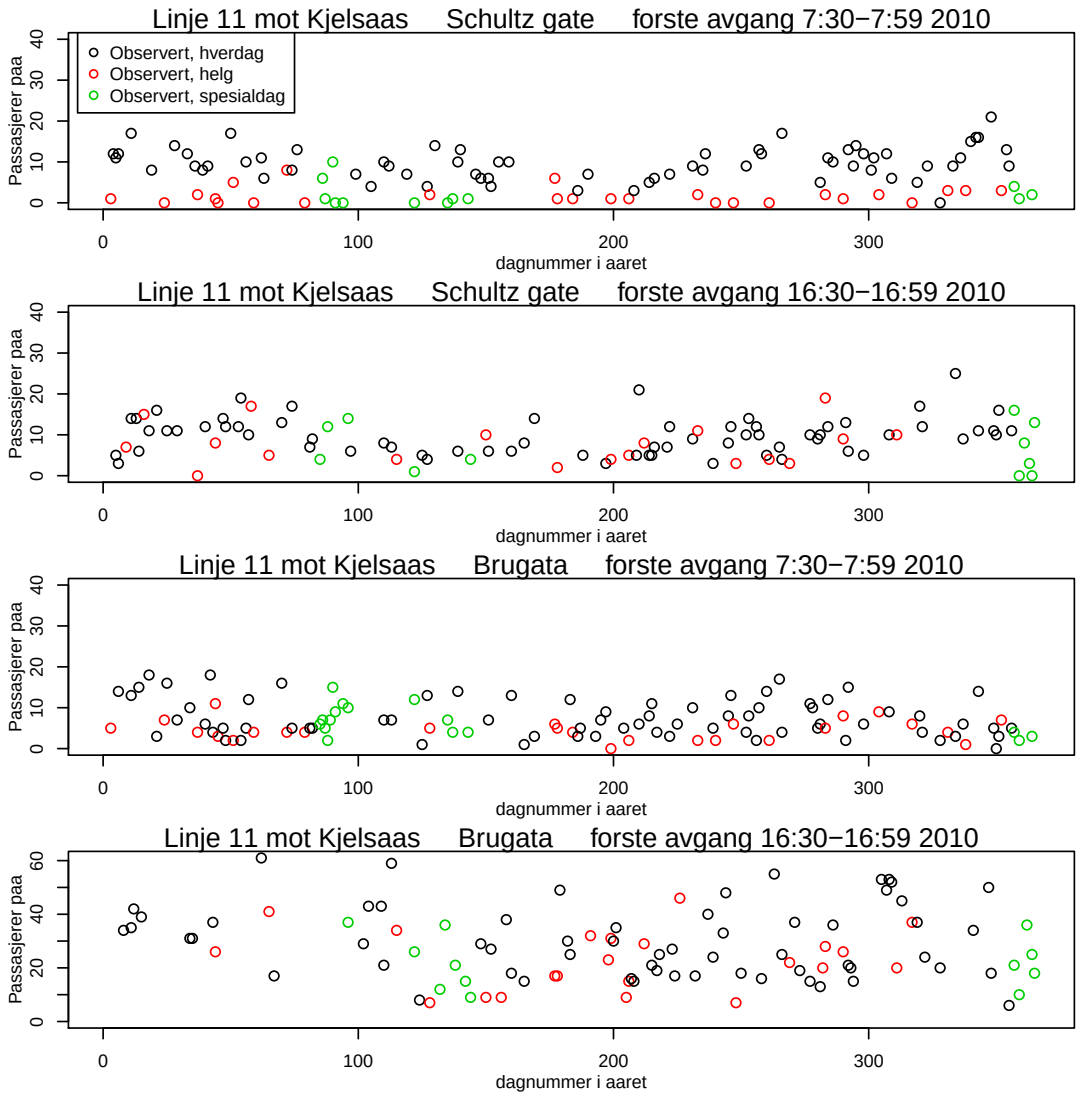
Det er én modell for antall påstigende passasjerer, og en annen for antall avstigende. Last etter avgang fra en holdeplass, det vil si antall passasjerer på trikken, beregnes som "last før holdeplass + antall påstigende – antall avstigende". Modellene for påstigende og avstigende er ganske like, men med en vesentlig forskjell. Modellen for påstigende modellerer antall påstigende direkte. I modellen for avstigende modelleres først *andel* avstigende som en andel av last, og antall avstigende fås ved å gange med last. Dette sikrer at forholdet mellom påstigende, avstigende og last alltid er konsistent.

Vi illustrerer hvordan modellen for påstigende virker ved å ta for oss noen konkrete eksempler fra 2010 for linje 11, retning mot Kjelsås, som vist i figurene 1-4. Figur 1 har fire paneler, hvor de to øverste gjelder Schultz gate og de to nederste Brugata. Øverste panel viser antall påstigende passasjerer på første tur med avgang etter 7:30, men før 8:00, fra Majorstua, men kun for de ca. 100 turene med tellinger. Hverdager (mandag-fredag) er vist med svart, helgedager (lørdag-søndag) med rødt, og spesielle dager (påske, fridager i mai, jul) med grønt. Panelet under viser tilsvarende antall påstigende passasjerer for første avgang etter 16:30, men før 17:00. De to nederste panelene tilsvarer de to øverste, men gjelder for Brugata. Oppgaven går altså ut på å beskrive hva som er systematisk variasjon i disse dataene.

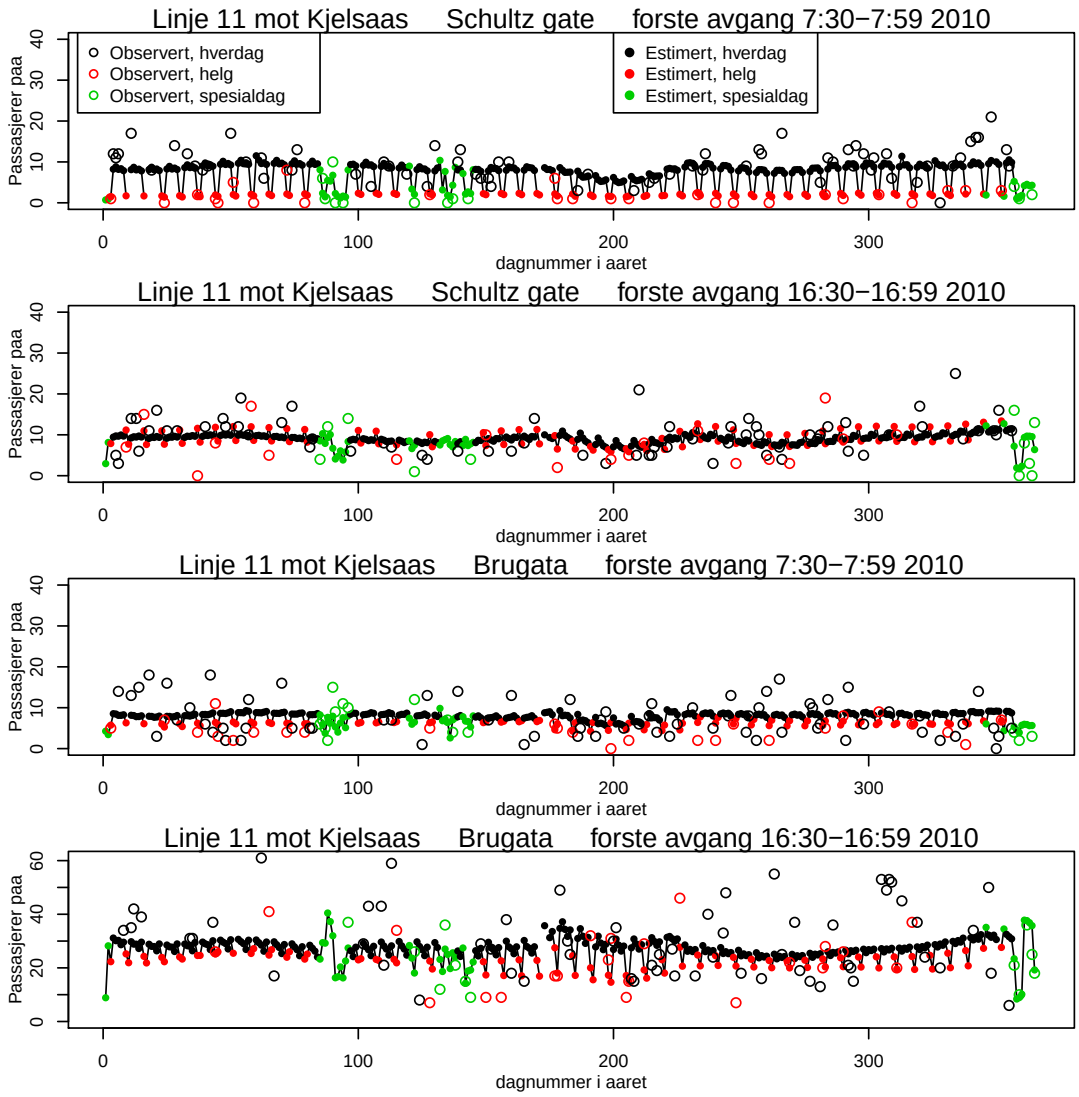
Figur 2 viser hva som ut fra modellen er det typiske eller normale antall påstigende på disse morgen- og ettermiddagsavgangene over året. For Schultz gate om morgenen (øverste panel) er det en nedgang i trafikken om sommeren, og mindre trafikk på helgedager enn på hverdager. Vi ser også en glidende overgang i uke-mønsteret gjennom året. Om vi ser på Schultz gate om ettermiddagen, så er det faktisk vanlig med mer trafikk på lørdagene enn på hverdagene. Mønsteret for Brugata er videre litt forskjellig fra det for Schultz gate.

Figur 3 viser antall påstigende passasjerer i Brugata på alle turer med tellinger de fire dagene fra fredag 3/2-2011 til søndag 6/2-2011. Figur 4 viser det samme, men sammen med modellprediksjonene for alle avganger i perioden. Modellprediksjonene er vist med fylte svarte sirkler med linjer mellom, og disse ligger tettest på dagtid fordi det da er flest avganger. I denne retningen er det relativt få passasjerer om morgenen både på hverdager og helgedager. Vi kan se ettermiddagsrushet klart både torsdag og fredag, og dette kommer tidligere på en fredag enn en torsdag. Videre er Brugata en holdeplass hvor det er jevnt med passasjerer utover kvelden både fredag og lørdag.

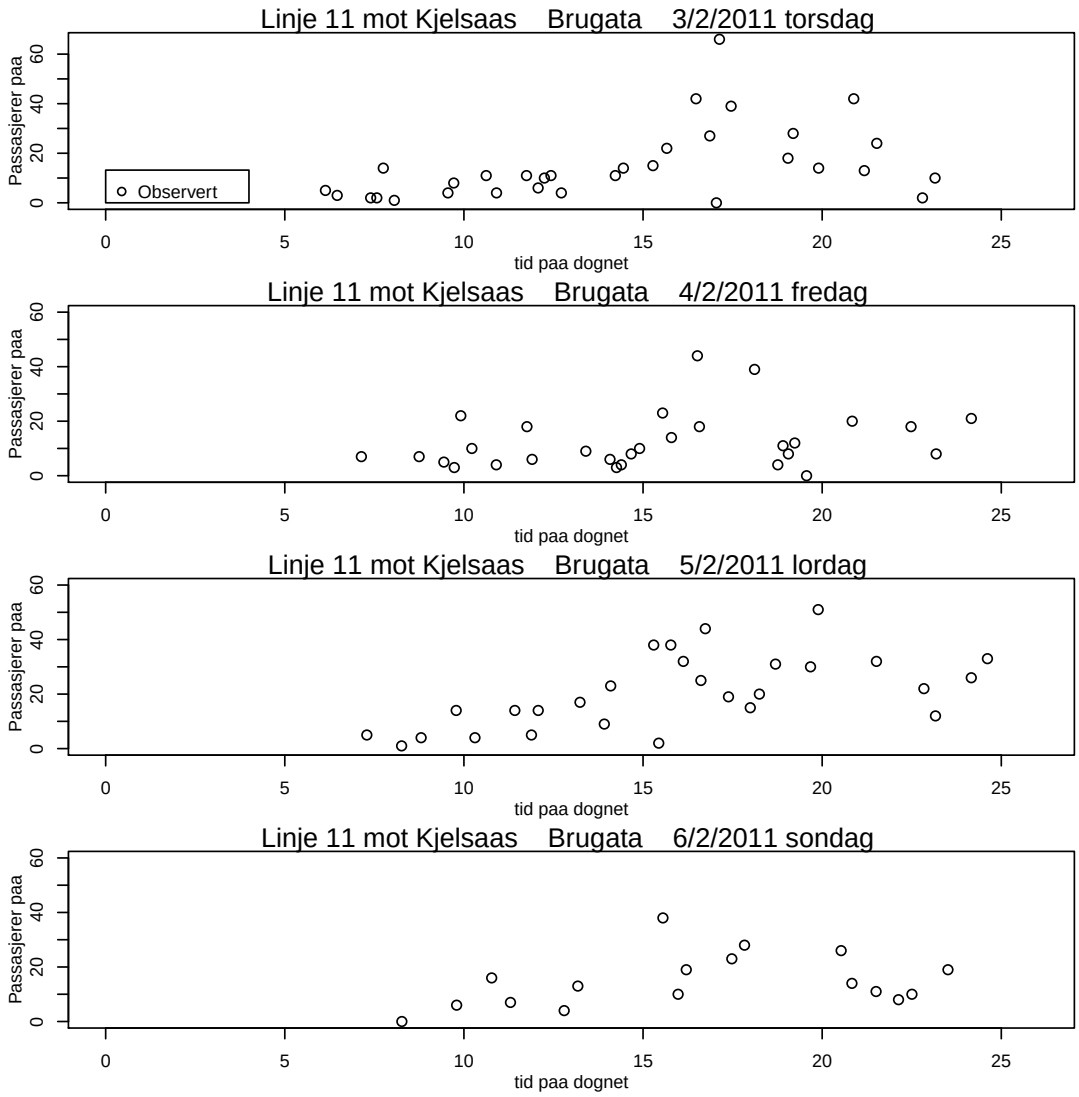
Den egentlige metoden er altså slik at for en tur uten telling brukes modellprediksjonen, mens for en tur med telling brukes den observerte verdien. I figurene 2 og 4 er modellprediksjonene gjengitt også for de avgangene hvor det foreligger tellinger. De er videre forenklet noe for å gi mer lesbare figurer. I tillegg til hva som er vist i figurene justeres modellprediksjonene som noe opp eller ned avhengig av om turene med tellinger før og etter har flere eller færre passasjerer enn normalt.



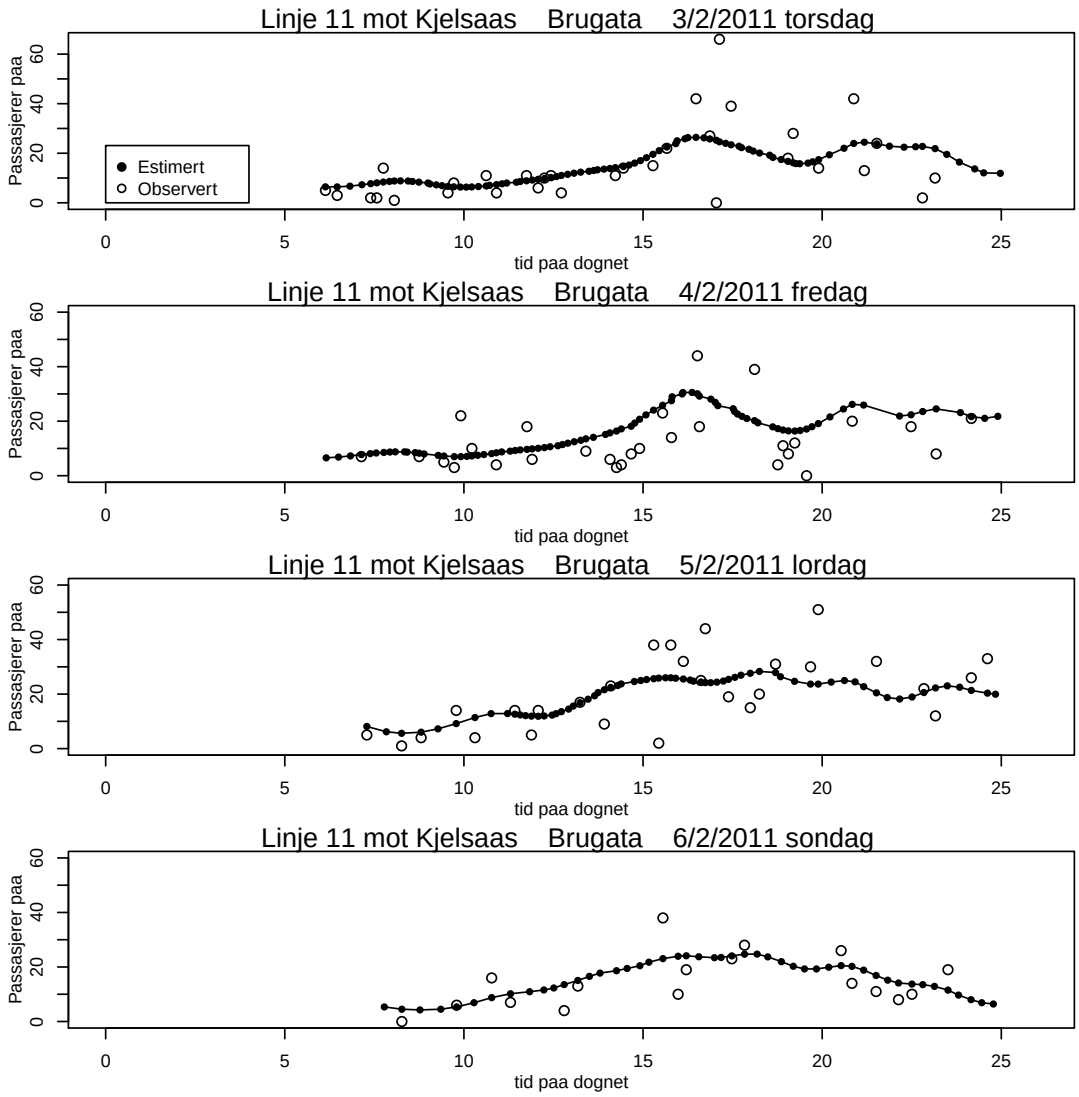
Figur 1. Observert antall påstigende gjennom året for morgen- og ettermiddagsavganger på linje 11, retning mot Kjelsås, fra Schultz gate og Brugata.



Figur 2. Observert og estimert antall påstigende gjennom året for morgen- og ettermiddagsavganger på linje 11, retning mot Kjelsås, fra Schultz gate og Brugata.



Figur 3. Observert antall påstigende gjennom gjennom døgnet for fire døgn i februar 2011 på linje 11, retning mot Kjelsås, på Brugata.



Figur 4. Observert og estimert antall påstigende gjennom gjennom døgnet for fire døgn i februar 2011 på linje 11, retning mot Kjelsås, på Brugata.

3.2 Regresjonsanalyse og interpolasjon

Hovedidéen bak metodikken kan kort og uformelt beskrives som å "fylle inn" (interpolere) manglende telledata ved hjelp av å utnytte systematiske variasjoner i dataene. For noen (ca 31%) av trikketurene har vi automatiske tellinger av på- og avstigende, mens vi for *alle* trikketurene i tillegg blant annet har følgende informasjon:

- hvor trikken har kjørt
- når trikken skulle ha vært på hvert stopp i henhold til rutetabellen
- når trikken faktisk kjørte på hvert stopp

Vi kan dermed bruke denne informasjonen til å anslå hvor mange som gikk på og av på alle holdeplasser for turer som er kjørt, men som det ikke foreligger tellinger for. For å kunne gjøre dette automatisk trenger vi en statistisk modell som beskriver sammenhengen mellom *forklaringsvariable* basert på punktene ovenfor, og passasjertellingene (*responsvariabelen*). Denne type modell kalles en *regresjonsmodell*. Vi tilpasser regresjonsmodellen til data der vi både har observert forklaringsvariablene og responsvariabelen, og bruker den tilpassede modellen til å predikere (anslå) hvor mange som gikk på og av trikken der vi bare har observert forklaringsvariablene. Ved hjelp av å *simulere* fra den estimerte regresjonsmodellen kan vi også anslå usikkerheten i de predikerte trafikktallene. Med "simulering" menes her såkalt Monte Carlo-simulering eller stokastisk simulering, dvs å trekke tilfeldig fra en sannsynlighetsfordeling. Kapittel 3.3 beskriver forklaringsvariablene, kapittel 3.4 beskriver regresjonsmodellen, mens kapittel 3.5 forklarer hvordan vi har simulert for å anslå usikkerheten i de predikerte trafikktallene.

3.3 Forklaringsvariable

Basert på punktene beskrevet i kapittel 3.2, dvs når og hvor trikken har/skulle ha kjørt, kan vi definere mange ulike forklaringsvariable. Strategien i vår metodikk er å bruke ganske mange forklaringsvariable, og kombinere disse gjennom en såkalt krympemetodikk som "automatisk" velger ut kombinasjoner som gir stor forklaringskraft. Det at vi tar med en forklaringsvariabel i modellen betyr at vi tenker oss at variabelen *potensielt* kan være viktig, men ikke at den nødvendigvis har stor effekt i alle tilfeller. Krympemetodikken vil uansett sørge for at uviktige forklaringsvariable ignoreres i modellen. En detaljert liste over alle forklaringsvariablene er gitt i tabell A.1 og A.2, her gis en oversikt.

De fleste av forklaringsvariablene er på tur-nivå, dvs vi har en verdi av forklaringsvariabelen pr tur med trikken, mens tre av forklaringsvariablene er på holdeplass-nivå, dvs at de kan variere fra holdeplass til holdeplass. Vi starter med å beskrive variablene på tur-nivå. Alle disse er basert på faktisk starttidspunkt (tid og dato) for hver trikketur. Vi kan bruke disse tidspunktene til å beskrive systematisk variasjon i tellingene. For det første kan vi se for oss at det er en trend (kanskje antall passasjerer øker over tid pga befolkningsvekst i Oslo?), så ulike tids-trendledd er tatt med. I tillegg vil det være periodisk variasjon over året og over døgnet, og dette tas med ved å bruke ulike sinus- og cosinus-funksjoner av dagnummer i året og tid på døgnet. Ukedag (mandag-søndag) og spesielle dager som julaften og 17. mai (en fullstendig liste er i tabell A.3) er også tatt med som forklaringsvariable. I tillegg er det tatt med en egen sommereffekt, se tabell A.1 for definisjon. På holdeplassnivå har vi med tre ulike forklaringsvariable. Den ene er antall holdeplasser som er igjen på turen (oppad begrenset til 5), fordi det å ha få holdeplasser igjen normalt vil gi færre påstigende. De to andre forklaringsvariable på holdeplassnivå beskriver avhengigheter mellom henholdsvis nærliggende (tidligere) holdeplasser på samme tur og avhengigheter mellom nærliggende turer (i tid) for samme holdeplass. (Hvis det har vært uvanlig mange påstigende på tidligere holdeplasser for en gitt tur og holdeplass, så vil det være en tendens til at det også blir uvanlig mange påstigende på den gitte holdeplassen.)

I tillegg tar vi med en del ulike *interaksjoner* mellom forklaringsvariablene nevnt ovenfor. Interaksjon betyr i denne sammenhengen at to eller flere forklaringsvariable får en egen tilleggseffekt av å opptre samtidig. For å ta hensyn til dette kan vi ta med produktet av to eller flere forklaringsvariable som en egen forklaringsvariabel. For eksempel kan vi tenke oss at døgnvariasjonen er forskjellig for hver ukedag (eksempelvis forskjøvet fredagsrush), og for å ta hensyn til dette tar vi med interaksjonen mellom forklaringsvariablene for døgnvariasjon og ukedag. I modelleringen av en gitt holdeplass på en gitt tur tas det også hensyn til antallet påstigende på turene før og etter, og på ulike holdeplasser på samme tur.

Vi presiserer igjen at det at vi tar med en forklaringsvariabel i modellen ikke betyr at vi antar at variabelen nødvendigvis er viktig, bare at den potensielt har betydning. I tillegg antas intet på forhånd om størrelse og retning på effektene av variablene; dette estimeres som en del av modellen. Sinus/cosinus-funksjonene for sesongvariasjon gir stor fleksibilitet til å beskrive ulike sesongmønstre både for døgn- og årsvariasjon.

3.4 Modeller for antall påstigende og andel avstigende

Dette avsnittet gir en uformell beskrivelse av regresjonsmodellene for på- og avstigende. Begge disse responsvariablene modelleres separat for hver linje og retning. Antall påstigende modelleres direkte. For avstigende modellerer vi den forventede andelen p som går av: det vil si at det er en sannsynlighet p for at en passasjer går av, det er n passasjerer, og forventet antall avstigende er pn. Dette gjøres slik blant annet for å sikre at det aldri kan være flere avstigende enn det til enhver tid er på trikken.

Regresjonsmetoden som brukes er en kombinasjon av to statistiske metodikker:

1. redusert rang-regresjon (RRR)

2. generaliserte lineære modeller (GLM)

RRR er en såkalt krympemetode som brukes til å lage nye, "smarte" kombinasjon av forklaringsvariablene beskrevet i kapittel 3.3. Denne metoden fungerer slik at forklaringsvariable som har stor forklaringskraft automatisk vektes opp, og vi ender opp med et håndterlig antall (typisk rundt ti) nye variable. Dette gjøres på en slik måte at vi bruker informasjon på tvers av holdeplasser innen samme linje og retning, og vi drar nytte av at variasjon over f.eks. døgnet er forholdsvis lik fra den ene holdeplassen til den andre. Dette er særlig nyttig for holdeplasser som ligger utenom den ordinære ruta for en linje, f.eks. når linje 11 unntaksvis kjører en delstrekning av turene langs traseen til linje 19. GLM er en klasse av regresjonsmodeller som håndterer responsvariable som ikke er kontinuerlige. Vanlig lineær regresjon kan kun brukes når responsen er kontinuerlig og nær normalfordelt, mens modeller i GLM-klassen også takler responsvariable som er antall (påstigende) og andeler (avstigende). GLM er en kjernemetodikk innen moderne statistikk, og det eksisterer en velutviklet teori som sikrer at dette fungerer.

Stegene i vår modell kan (noe forenklet) beskrives slik:

1. Nye "smarte" forklaringsvariable lages ved hjelp av en innledende RRR

2. GLM-regresjon separat for hver holdeplass med de nye forklaringsvariablene. Antall variable som tas med avhenger av antall observasjoner

Denne estimeringen resulterer i parametre som beskriver sannsynlighetsfordelingene for på- og avstigende.

3.5 Simulering av påstigende, avstigende og last; anslag for usikkerhet

Etter at modellene for på- og avstigende er tilpasset som beskrevet i kapittel 3.4, kan vi simulere fra de tilpassede modellene. Som nevnt mener vi her alltid Monte Carlo-simulering (stokastisk simulering) når vi snakker om "simulering", dvs simulering betyr her tilfeldig trekning fra en sannsynlighetsfordeling. Simuleringene foregår på følgende måte: Betegn holdeplassene med $1, 2, \ldots, n$ og la P_i, A_i og L_i betegne henholdsvis påstigende, avstigende og last ved holdeplass $i = 1, 2, \ldots, n$. For holdeplass 1, trekk P_1 fra den estimerte sannsynlighetsfordelingen, sett A_1 til 0 og sett L_1 lik P_1. For holdeplass $2, 3, \ldots, n-1$, simuler P_i. Trekk A_i fra modellen for antall avstigende. Denne modellen avhenger av L_{i-1} og den forventende andelen avstigende. Deretter, oppdater last ved å sette $L_i = L_{i-1} + P_i - A_i$. For siste holdeplass n settes $A_n = L_{n-1}$, slik at alle går av trikken på siste holdeplass.

Ved å simulere N ganger (med f.eks. $N = 1000$), kan vi på denne måten si noe både om forventede antall påstigende, avstigende og last for alle linjer, holdeplasser og turer, og også om usikkerheten i disse tallene. Merk at alt simuleres på holdeplassnivå, dvs vi simulerer antallene for hver holdeplass, tur, linje og retning. Hvis man ønsker resultater for aggregerte størrelser, f.eks. alle morgenavganger i august 2011 tilsammen for linje 13, så kan aggregeringen gjøres for hver av de N simuleringene, slik at vi sitter igjen med N simuleringer også for den aggregerte størrelsen. Merk at vi kun simulerer for de turene hvor vi ikke har tellinger.

Punktestimatet beregnes ved å ta gjennomsnittet av alle simuleringene. Usikkerheten bestemmes i utgangspunktet ved å ta kvantiler av de simulerte tallene. Hvis vi f.eks. ønsker et 95% konfidensintervall, tar vi 2.5%-kvantilen som nedre grense og 97.5%-kvantilen som øvre grense, slik at 95% av simuleringene vil ligge innenfor intervallet (p-prosent kvantilen er definert ved at p prosent av simuleringene er mindre enn eller lik p). Imidlertid viser det seg at denne metoden gir noe for smale intervaller for aggregerte størrelser, på grunn av at vi ikke tar tilstrekkelig hensyn til avhengighet mellom holdeplassene, og mellom turer som er nære i tid. Dette korrigeres for ved å utvide usikkerhetsintervallene med en faktor som avhenger av antallet datapunkter som inngår i den aggregerte størrelsen. Korrigerte intervaller for ulike summer over aggregerte størrelser er vist i tabell 3.

Det kan også være av interesse å se på usikkerhet av differanser, særlig for å teste for endringer over tid, f.eks. om det har vært en endring i antall påstigende fra januar 2010 til januar 2011. Dette kan testes ved å simulere differansene, og se etter om konfidensintervallet dekker null. Hvis intervallet dekker null, så er det ikke statistisk grunnlag for å si at det er en endring. Også i dette tilfellet må intervallene korrigeres (utvides), på en lignende måte som for summer over aggregerte størrelser.

4 Resultater

Dette kapitlet gir noen eksempler (i form av ulike tabeller og plott) på hvilken type informasjon man kan få ut av metoden. Tabell 3 viser totalt antall påstigende (med 95% usikkerhetsintervall) for ulike utvalg av turer, både for hver linje/retning separat og totalt for alle linjer/retninger. Beregningene forutsetter at vi har fått opplysninger om alle kjørte avganger i dataperioden.

Vi har også testet for om det er en økning i total sum påstigende fra 2010 til 2011. Utfra tabell 3 kan det se ut som om det er en økning på 387000 påstigende (dvs ca 1% økning). For å teste om denne økningen er statistisk signifikant har vi simulert 1000 verdier for differansen mellom sum påstigende i 2011 og sum påstigende i 2010, og beregnet et korrigert 95% konfidensintervall for

denne differansen. Dette konfidensintervallet blir (−0.8%, 2.8%). Siden dette intervallet dekker null, så er økningen ikke statistisk signifikant.

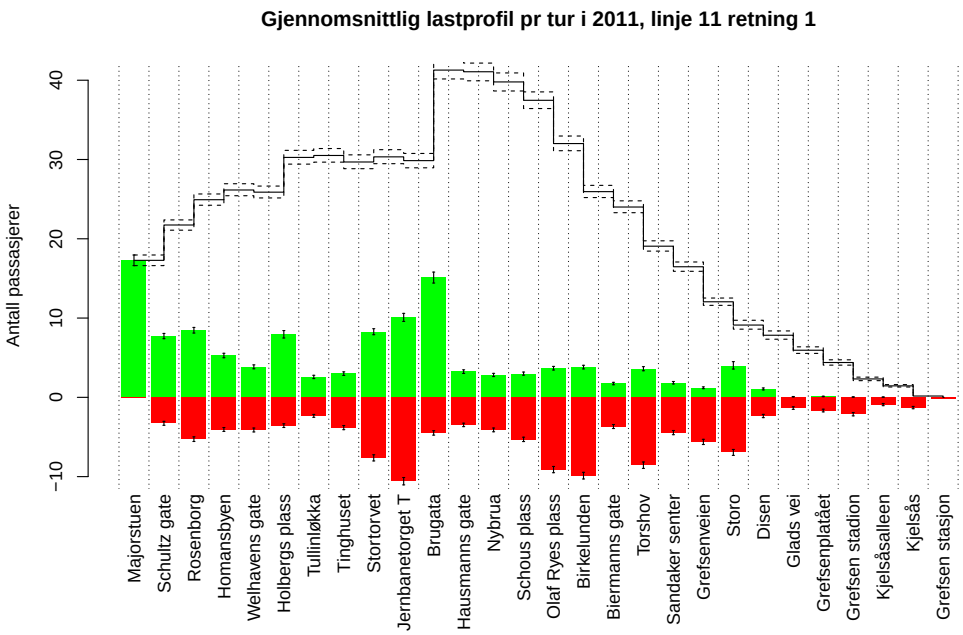
Figurene 5–17 viser lastprofiler med på- og avstigende for hver linje og retning (pr avgang for hele 2011). Kun holdeplassene på den vanlige ruta for disse linjene er vist på disse figurene, men beregningene er likevel gjort for alle holdeplasser.

Figur 18 viser årsvariasjonen som daglig antall påstigende passasjerer for hver uke for alle linjer i 2011, mens figur 19 viser årsvariasjonen som antall påstigende per tur. I figur 18 ser vi som ventet at antall passasjerer er mindre om sommeren, og vi ser tydelig at det er færre passasjerer i uke 16, som er påskeuka. Men vi ser også et kraftig og uventet fall i antall passasjerer i uke fire og fem. Det skyldes at det for linjene 13, 17 og 18 er betydelig færre turer i denne perioden i følge de dataene vi har fått fra SIS-databasen. Figur 20 illustrerer dette. Øverste panel viser antall registrerte turer per dag i 2011 for linje 13, retning 1, og det er tydelig færre turer i en periode tidlig på året. Det nederste panelet viser hvor stor andel av turene som det foreligger tellinger for. Andelen øker på slutten av året som en følge av flere trikker med telleutstyr, men fram til omtrent dag 150 ligger den stabilt på omkring 0.2, også i perioden med få turer. Om dataene er riktige, og det reellt er færre turer i denne perioden tidlig på året, blir våre beregninger korrekte. Om dette derimot skyldes at vi av en eller annen grunn ikke har fått informasjon om alle turer fra SIS-databasen, vil beregningene gi feil resultat. (Vi har fått opplyst fra Trafikanten at det kan ha vært problemer med å få lastet ned data fra disse trikkene til SIS-databasen i denne perioden.)

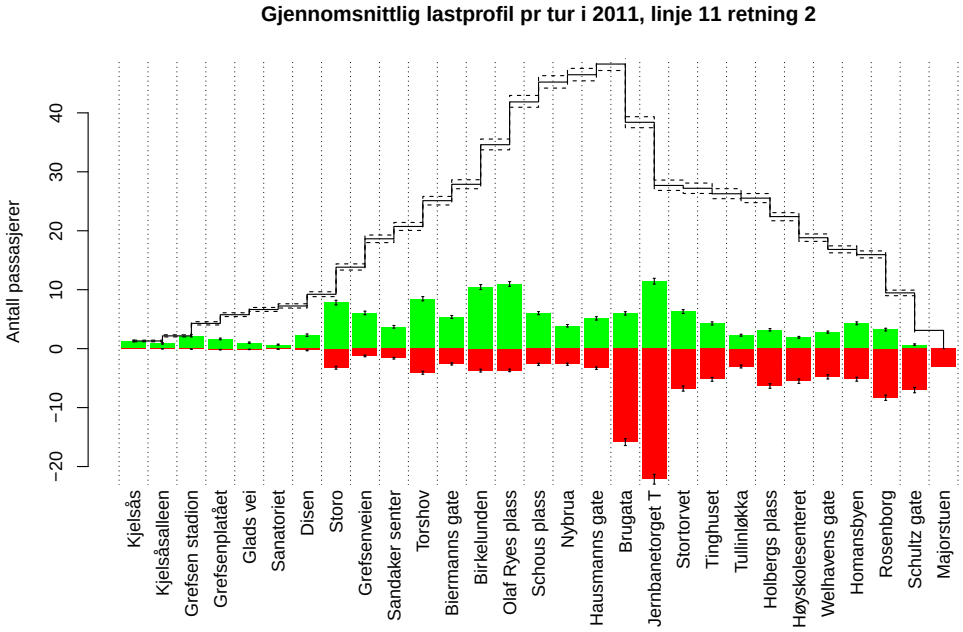
Figurene 21-24 viser henholdsvis døgn- og ukevariasjon pr tur og pr dag. For ukevariasjonen ser vi at det er færre passasjerer pr dag i helgene. Fordi det også er færre avganger i helgene, så blir det ikke så mange færre passasjerer *pr tur*, men det er noe færre. I døgnvariasjonen ser vi tydelig rushtidene på morgen (7:00-8:59) og ettermiddag (15:00-16:59).

Utvalg / linje	11, 1	11, 2	12, 1	12, 2	13, 1	13, 2	
Hele 2010	**3245** (3179, 3313)	**3367** (3306, 3427)	**4381** (4277, 4488)	**4328** (4252, 4411)	**3179** (3108, 3252)	**3102** (3030, 3168)	
Hele 2011	**3480** (3416, 3552)	**3490** (3436, 3550)	**4233** (4140, 4329)	**4387** (4314, 4464)	**3312** (3241, 3386)	**3124** (3066, 3186)	
Man-fre 2010	**2572** (2518, 2631)	**2687** (2634, 2738)	**3456** (3369, 3542)	**3429** (3365, 3495)	**2543** (2483, 2602)	**2500** (2440, 2560)	
Man-fre 2011	**2762** (2705, 2820)	**2787** (2739, 2837)	**3299** (3222, 3381)	**3428** (3366, 3493)	**2672** (2615, 2736)	**2528** (2477, 2580)	
Hverdager 2010	**2260** (2211, 2315)	**2346** (2299, 2396)	**2953** (2877, 3029)	**2953** (2899, 3015)	**2210** (2156, 2265)	**2133** (2080, 2185)	
Hverdager 2011	**2428** (2376, 2483)	**2447** (2404, 2492)	**2817** (2747, 2889)	**2937** (2883, 2993)	**2327** (2269, 2384)	**2192** (2147, 2239)	
Lørdager 2010	**396** (380, 415)	**381** (367, 395)	**518** (492, 546)	**517** (497, 537)	**354** (336, 372)	**338** (322, 355)	
Lørdager 2011	**418** (401, 433)	**398** (386, 413)	**499** (474, 524)	**526** (508, 545)	**350** (336, 368)	**325** (312, 338)	
Søndager 2010	**272** (259, 286)	**297** (285, 309)	**408** (384, 434)	**380** (364, 399)	**284** (267, 303)	**265** (252, 279)	
Søndager 2011	**284** (272, 297)	**290** (278, 301)	**424** (399, 449)	**420** (403, 438)	**280** (264, 297)	**262** (250, 275)	
Hverdager juli 2010	**140** (132, 150)	**160** (152, 169)	**268** (250, 286)	**242** (229, 255)	**151** (142, 160)	**194** (183, 207)	
Hverdager juli 2011	**145** (138, 153)	**150** (144, 157)	**223** (209, 238)	**227** (217, 237)	**148** (139, 157)	**148** (140, 156)	
	17, 1	17, 2	18, 1	18, 2	19, 1	19, 2	Sum
Hele 2010	**3552** (3475, 3631)	**3396** (3319, 3477)	**2904** (2834, 2980)	**2810** (2735, 2884)	**2070** (2018, 2125)	**2167** (2118, 2214)	**38500** (38128, 38853)
Hele 2011	**3468** (3396, 3540)	**3265** (3195, 3343)	**2840** (2771, 2911)	**2823** (2746, 2899)	**2167** (2119, 2220)	**2298** (2250, 2348)	**38887** (38542, 39250)
Man-fre 2010	**2956** (2885, 3030)	**2850** (2779, 2921)	**2461** (2396, 2531)	**2371** (2301, 2439)	**1682** (1638, 1728)	**1778** (1736, 1823)	**31285** (30923, 31612)
Man-fre 2011	**2879** (2816, 2944)	**2716** (2650, 2785)	**2405** (2343, 2470)	**2372** (2302, 2438)	**1749** (1706, 1796)	**1879** (1836, 1922)	**31474** (31193, 31779)
Hverdager 2010	**2618** (2555, 2688)	**2516** (2450, 2583)	**2186** (2123, 2248)	**2100** (2034, 2164)	**1467** (1427, 1508)	**1564** (1524, 1607)	**27307** (26990, 27620)
Hverdager 2011	**2536** (2478, 2597)	**2385** (2325, 2448)	**2138** (2084, 2198)	**2100** (2039, 2159)	**1532** (1494, 1575)	**1651** (1612, 1692)	**27489** (27211, 27770)
Lørdager 2010	**347** (332, 363)	**317** (302, 334)	**236** (223, 248)	**223** (211, 235)	**228** (217, 240)	**225** (215, 236)	**4080** (3998, 4162)
Lørdager 2011	**329** (315, 343)	**303** (289, 319)	**223** (212, 234)	**229** (217, 241)	**236** (225, 246)	**236** (227, 247)	**4072** (3993, 4153)
Søndager 2010	**248** (236, 262)	**227** (213, 239)	**212** (201, 224)	**222** (209, 234)	**158** (149, 169)	**163** (154, 172)	**3137** (3063, 3210)
Søndager 2011	**248** (236, 261)	**237** (225, 250)	**210** (199, 225)	**218** (207, 231)	**174** (164, 186)	**175** (166, 184)	**3222** (3149, 3298)
Hverdager juli 2010	**149** (140, 159)	**142** (133, 152)	**119** (110, 127)	**116** (108, 126)	**105** (97, 113)	**100** (93, 106)	**1886** (1833, 1936)
Hverdager juli 2011	**147** (138, 158)	**139** (129, 149)	**109** (101, 117)	**113** (104, 122)	**97** (90, 105)	**103** (97, 109)	**1749** (1703, 1799)

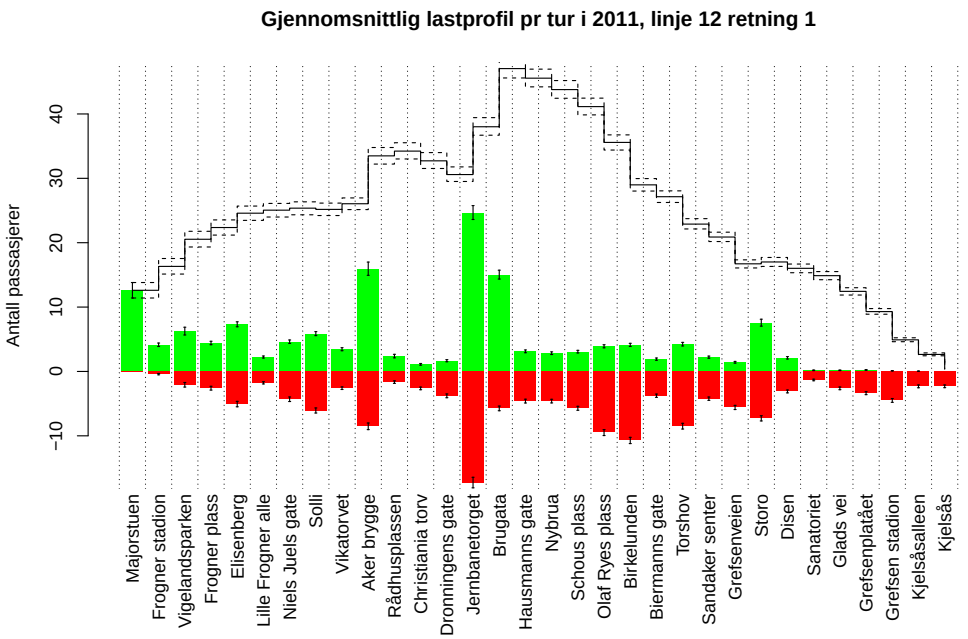
Tabell 3. Estimert sum påstigende (i tusener) med 95% usikkerhetsintervaller for ulike utvalg. “Man-fre” er alle dager fra mandag til fredag (selv om de skulle være helligdager), mens “hverdager” er vanlige hverdager utenom juli. *Merknad: Sent i oppdraget oppdaget vi en uforutsett feil i dataleveransene. Denne går ut på at ikke alle faktisk kjørte turer er registrert i SIS i uke 4 og 5 i 2011. Dette gjelder turer både med tellinger og uten tellinger. Den etablerte predikeringsmetoden forutsetter at alle faktisk kjørte turer er registrert i SIS. Når det ikke er tilfelle, beregnes det for lave trafikktall for de aktuelle dagene. Dette påvirker også alle aggregeringer hvor ukene 4 og 5 (2011) inngår. Av disse grunner har ovenstående tabell (figur) for lave tall for uke 4 og 5, og tallene som sådan har ingen verdi. Vi har likevel beholdt tabellen (figuren) for å illustrere viktige tema ved resultatpresentasjonen. Vi understreker også at denne type feil i dataleveransene ikke har betydning for godheten i de metodene som presenteres i denne rapporten. Men metoden tar ikke høyde for denne type feil i dataleveransene, som i utgangspunktet var opplyst ikke skulle kunne skje. Dersom man ønsker riktige tall for uke 4 og 5 må man enten få ny leveranse med komplette data, eller man må endre beregningsmetodene slik at man forholder seg til planlagte avganger i de tilfeller der man ikke har riktige data fra SIS om faktisk kjørte avganger. Det har det ikke vært rom for å gjøre i dette prosjektet.*



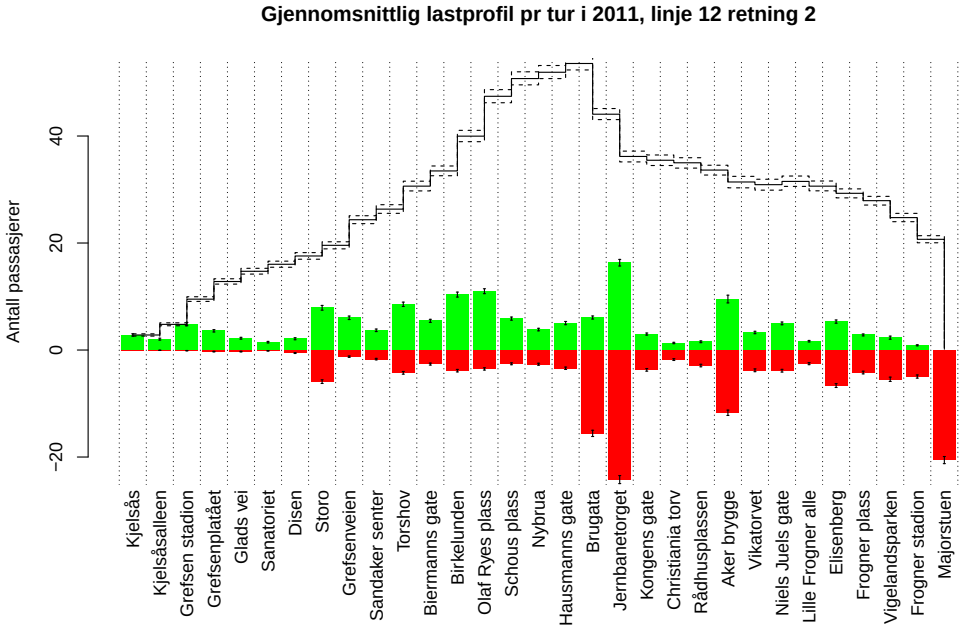
Figur 5. Lastprofil (pr tur) for antall påstigende (grønt), antall avstigende (rødt) og antall ombord (heltrukken sort) med 95% usikkerhetsintervaller, 2011, linje 11 retning 1



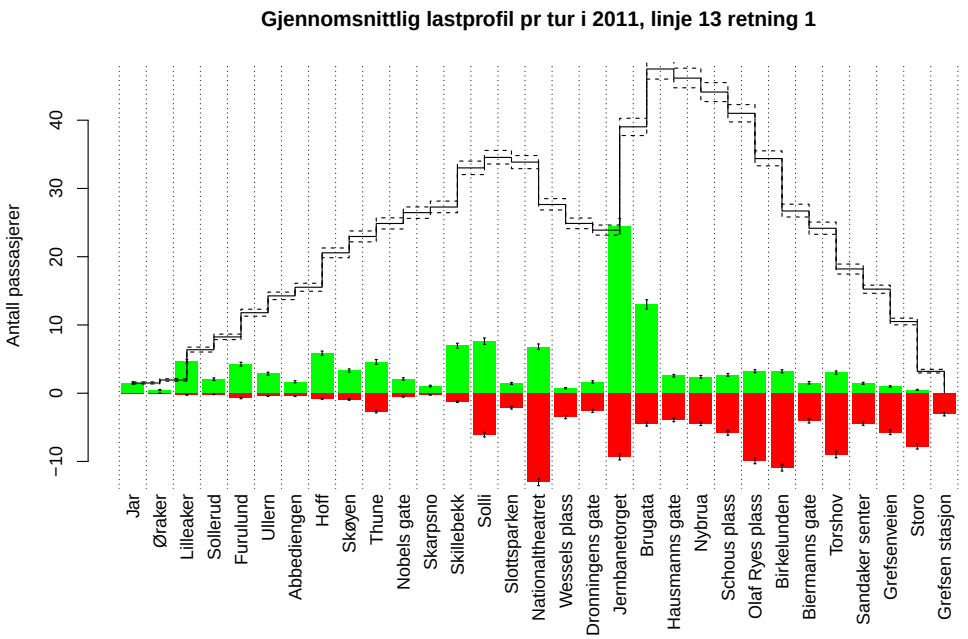
Figur 6. Lastprofil (pr tur) for antall påstigende (grønt), antall avstigende (rødt) og antall ombord (heltrukken sort) med 95% usikkerhetsintervaller, 2011, linje 11 retning 2



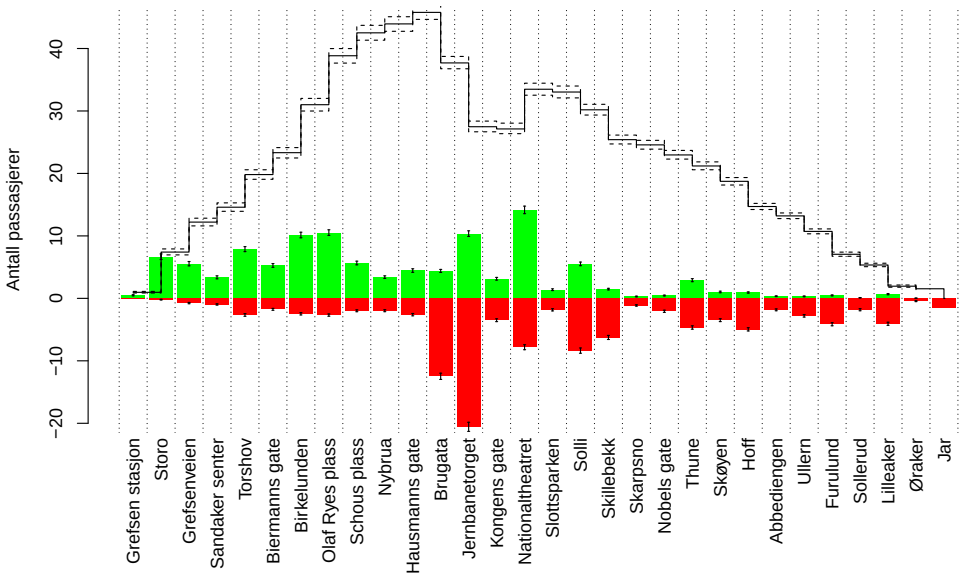
Figur 7. Lastprofil (pr tur) for antall påstigende (grønt), antall avstigende (rødt) og antall ombord (heltrukken sort) med 95% usikkerhetsintervaller, 2011, linje 12 retning 1



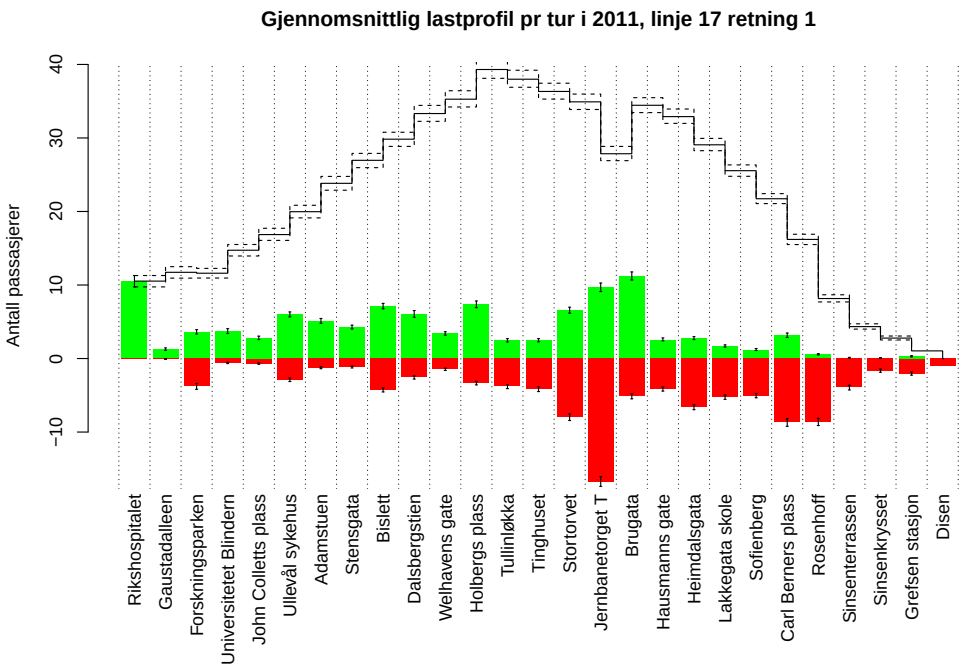
Figur 8. Lastprofil (pr tur) for antall påstigende (grønt), antall avstigende (rødt) og antall ombord (heltrukken sort) med 95% usikkerhetsintervaller, 2011, linje 12 retning 2



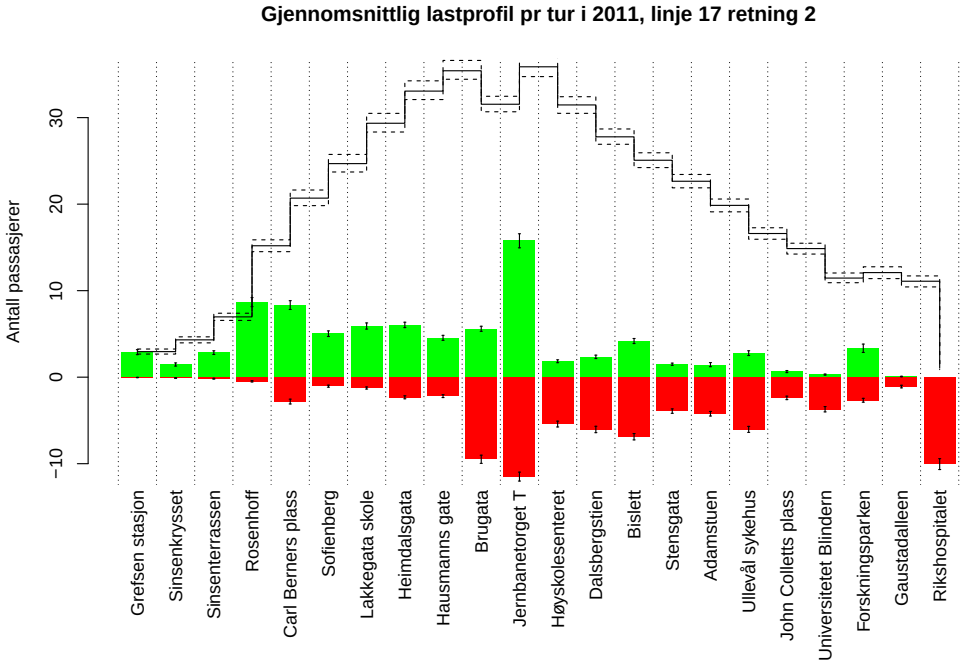
Figur 9. Lastprofil (pr tur) for antall påstigende (grønt), antall avstigende (rødt) og antall ombord (heltrukken sort) med 95% usikkerhetsintervaller, 2011, linje 13 retning 1



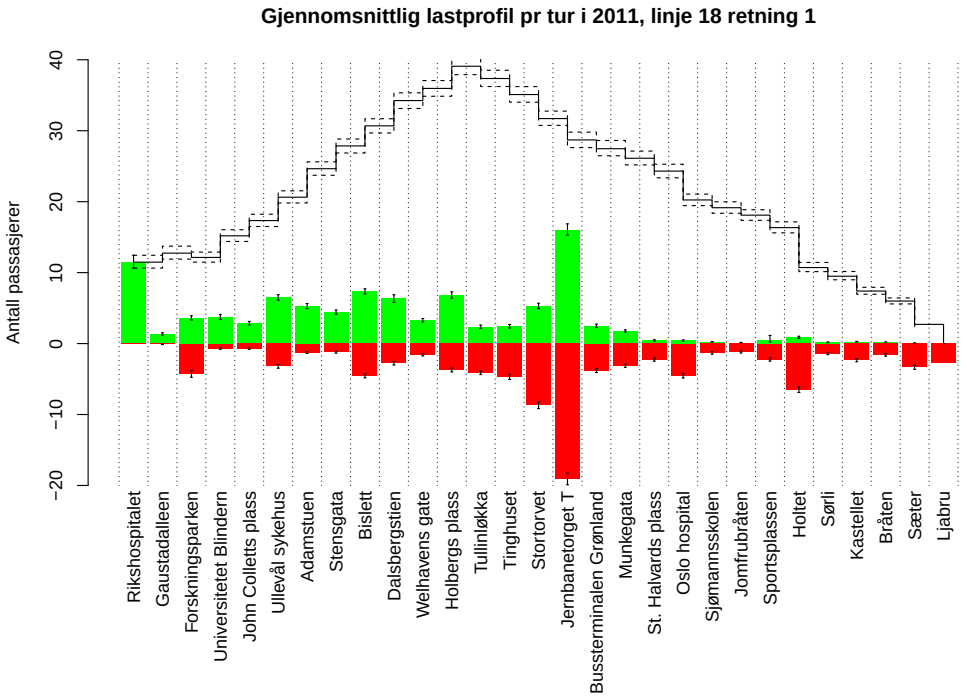
Figur 10. Lastprofil (pr tur) for antall påstigende (grønt), antall avstigende (rødt) og antall ombord (heltrukken sort) med 95% usikkerhetsintervaller, 2011, linje 13 retning 2



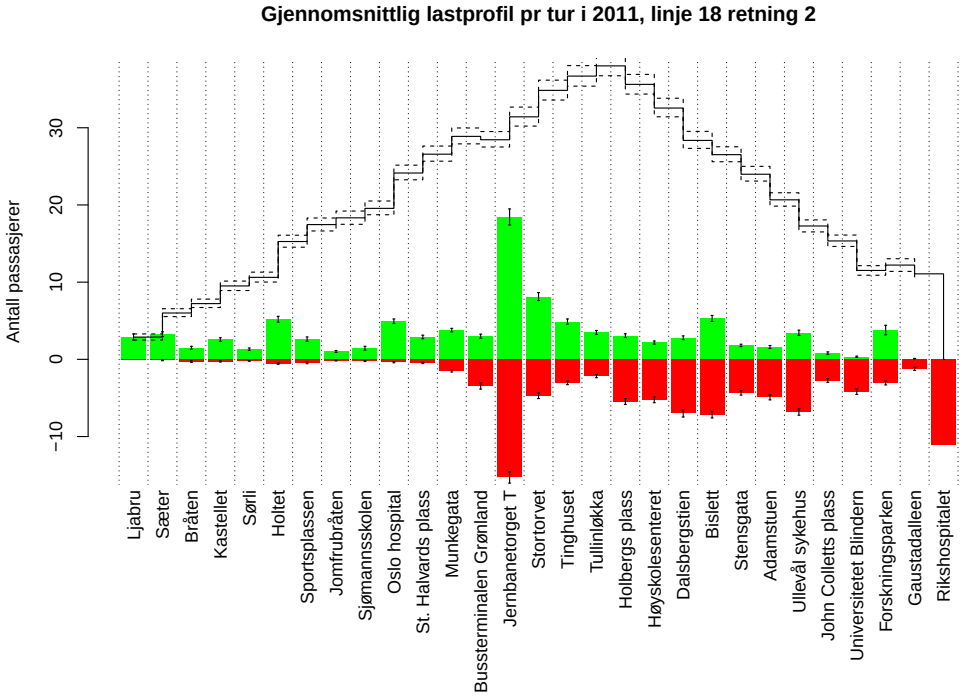
Figur 11. Lastprofil (pr tur) for antall påstigende (grønt), antall avstigende (rødt) og antall ombord (heltrukken sort) med 95% usikkerhetsintervaller, 2011, linje 17 retning 1



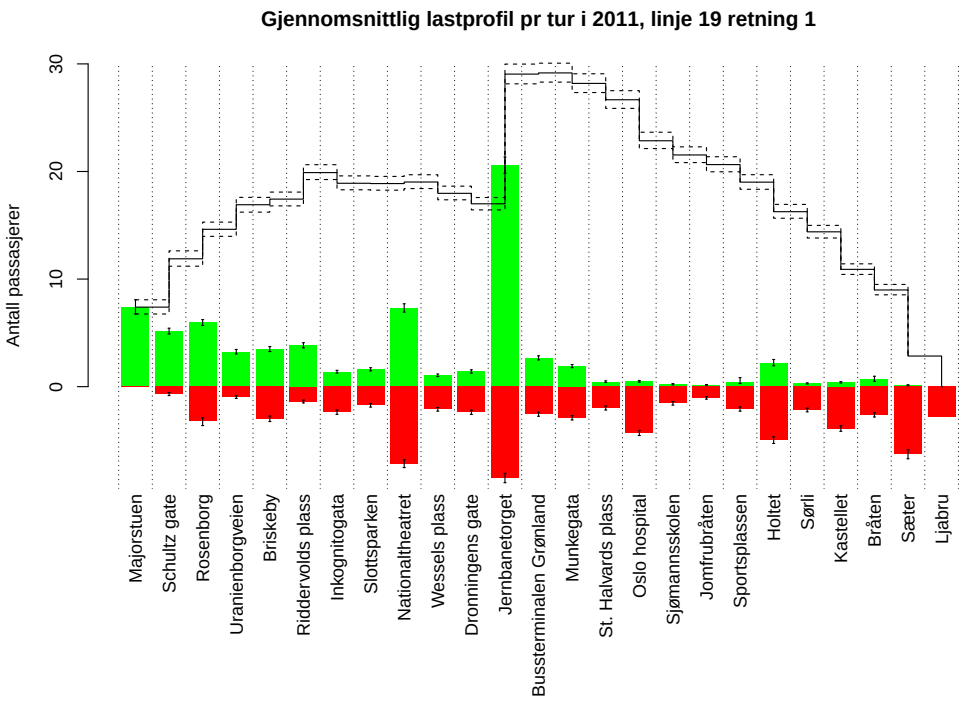
Figur 12. Lastprofil (pr tur) for antall påstigende (grønt), antall avstigende (rødt) og antall ombord (heltrukken sort) med 95% usikkerhetsintervaller, 2011, linje 17 retning 2



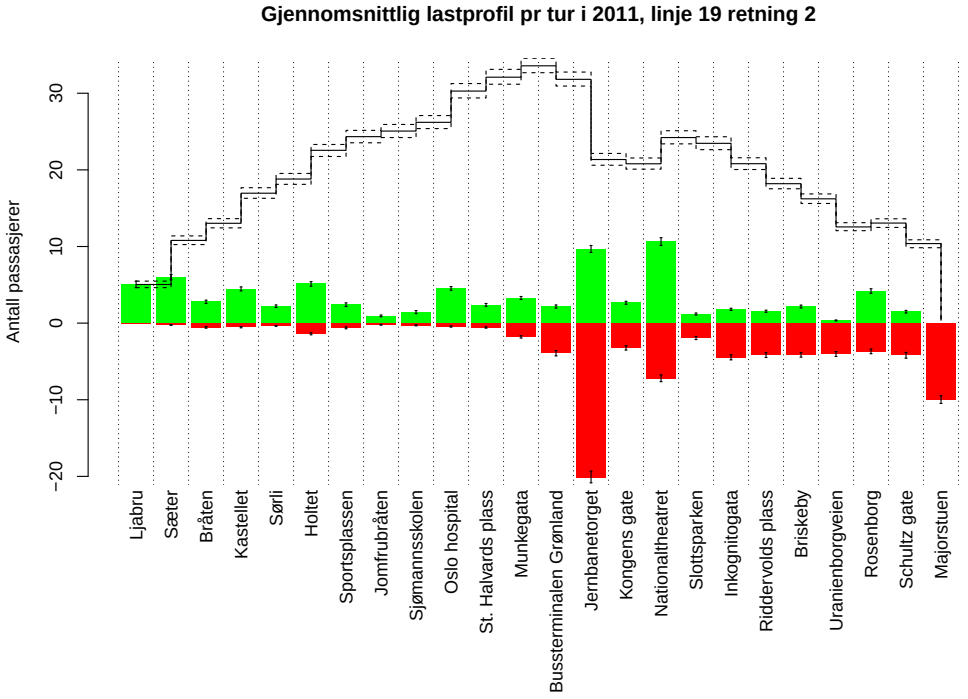
Figur 13. Lastprofil (pr tur) for antall påstigende (grønt), antall avstigende (rødt) og antall ombord (heltrukken sort) med 95% usikkerhetsintervaller, 2011, linje 18 retning 1



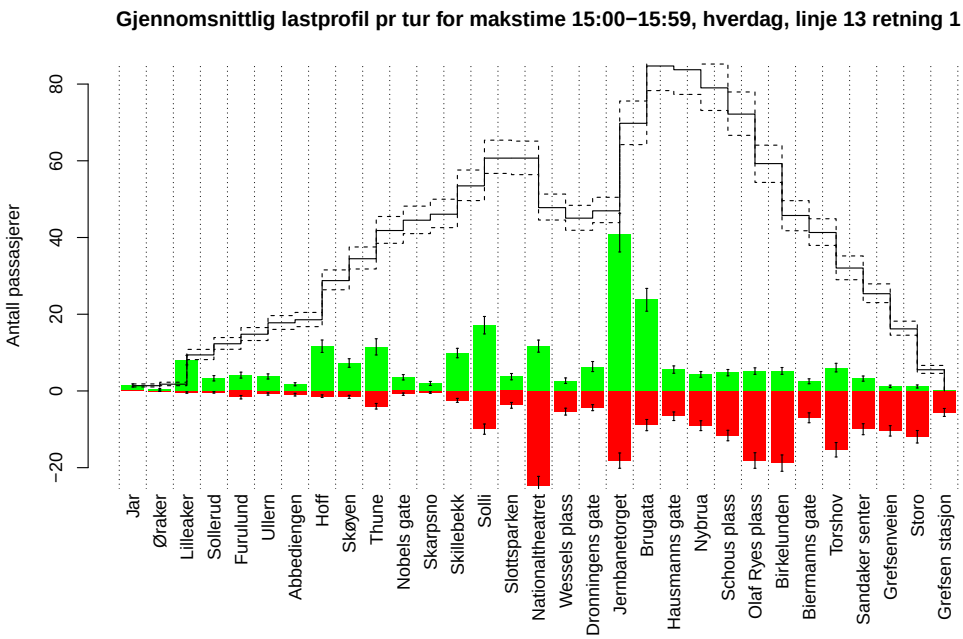
Figur 14. Lastprofil (pr tur) for antall påstigende (grønt), antall avstigende (rødt) og antall ombord (heltrukken sort) med 95% usikkerhetsintervaller, 2011, linje 18 retning 2



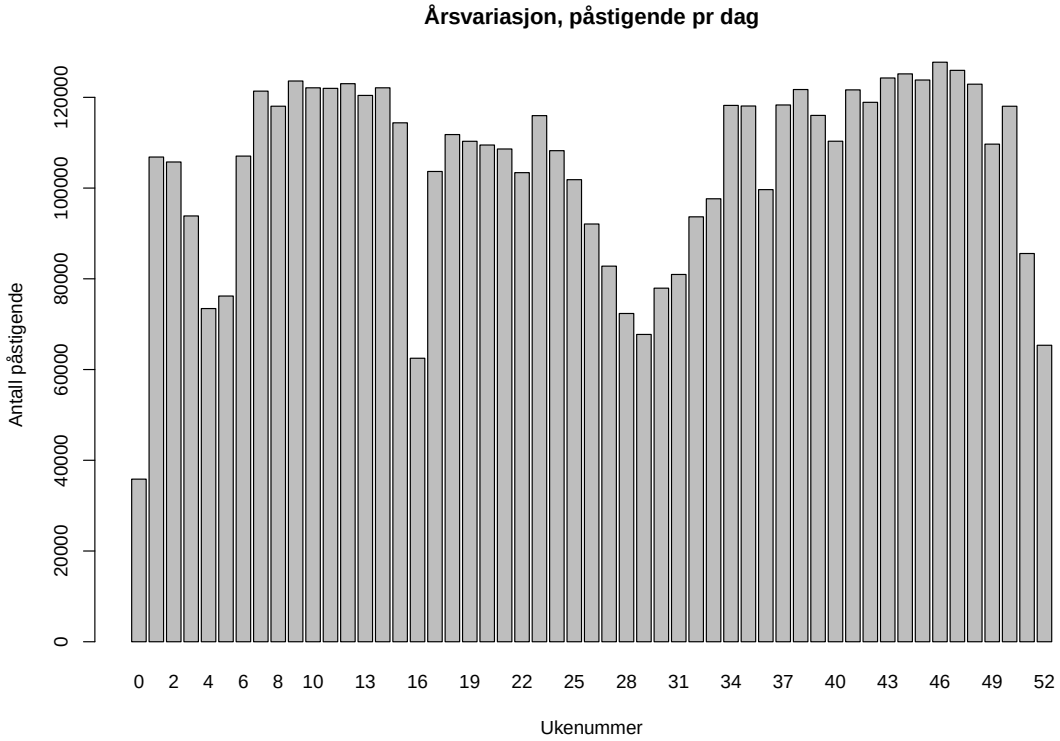
Figur 15. Lastprofil (pr tur) for antall påstigende (grønt), antall avstigende (rødt) og antall ombord (heltrukken sort) med 95% usikkerhetsintervaller, 2011, linje 19 retning 1



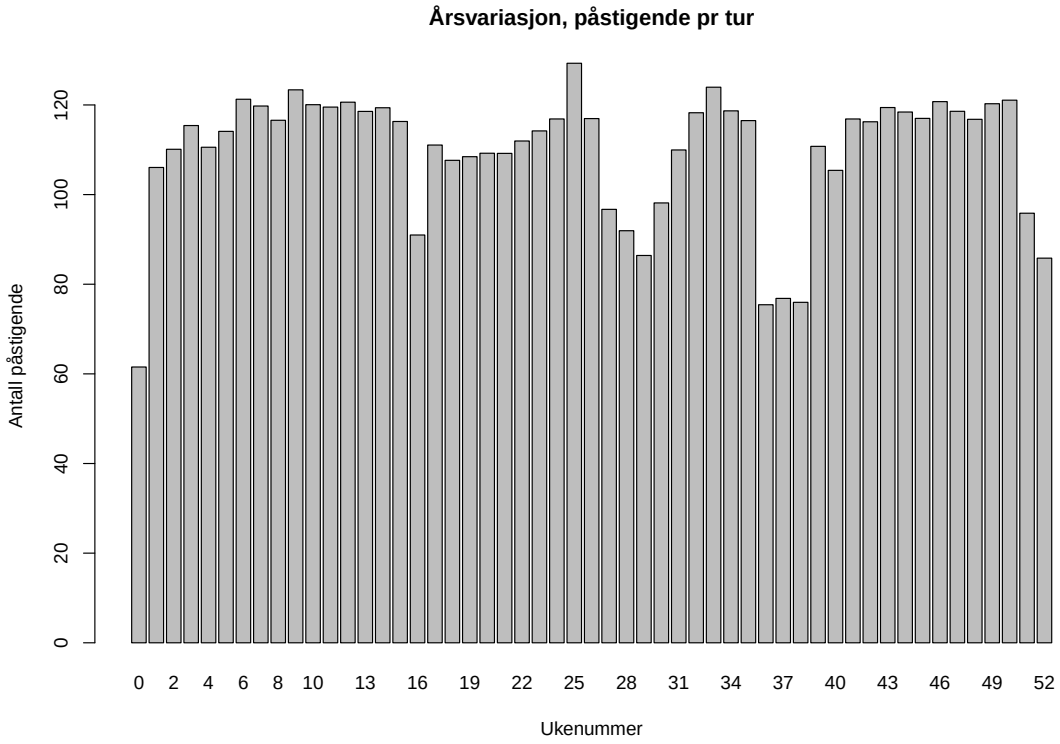
Figur 16. Lastprofil (pr tur) for antall påstigende (grønt), antall avstigende (rødt) og antall ombord (heltrukken sort) med 95% usikkerhetsintervaller, 2011, linje 19 retning 2



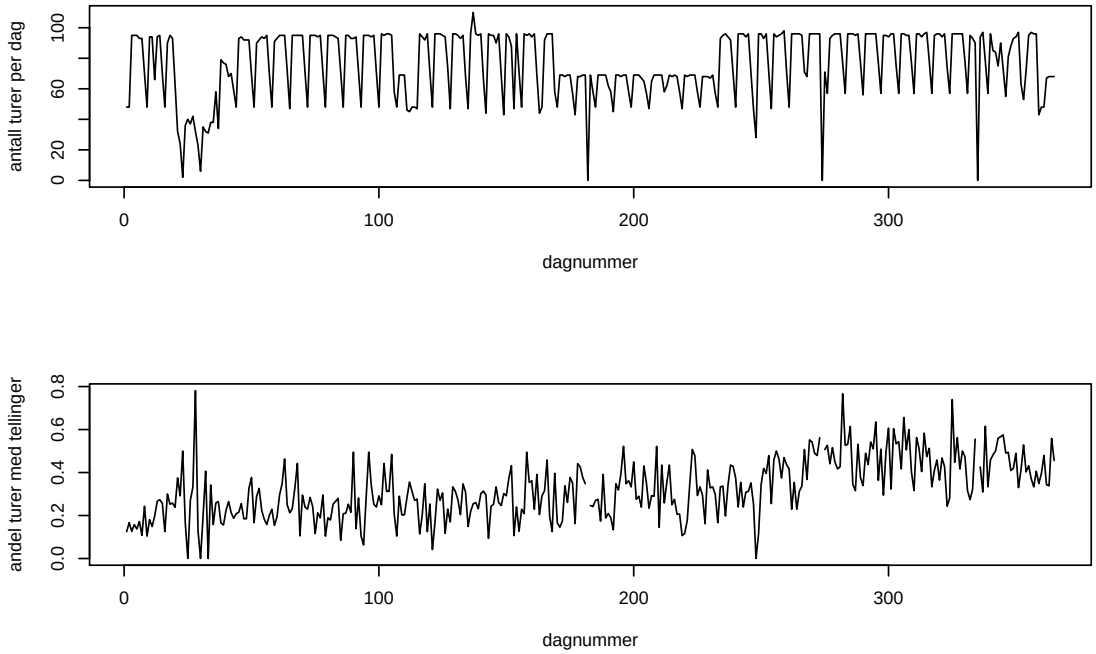
Figur 17. Lastprofil (pr tur) for antall påstigende (grønt), antall avstigende (rødt) og antall ombord (heltrukken sort) med 95% usikkerhetsintervaller for makstime 15:00-15:59, linje 13 retning 1



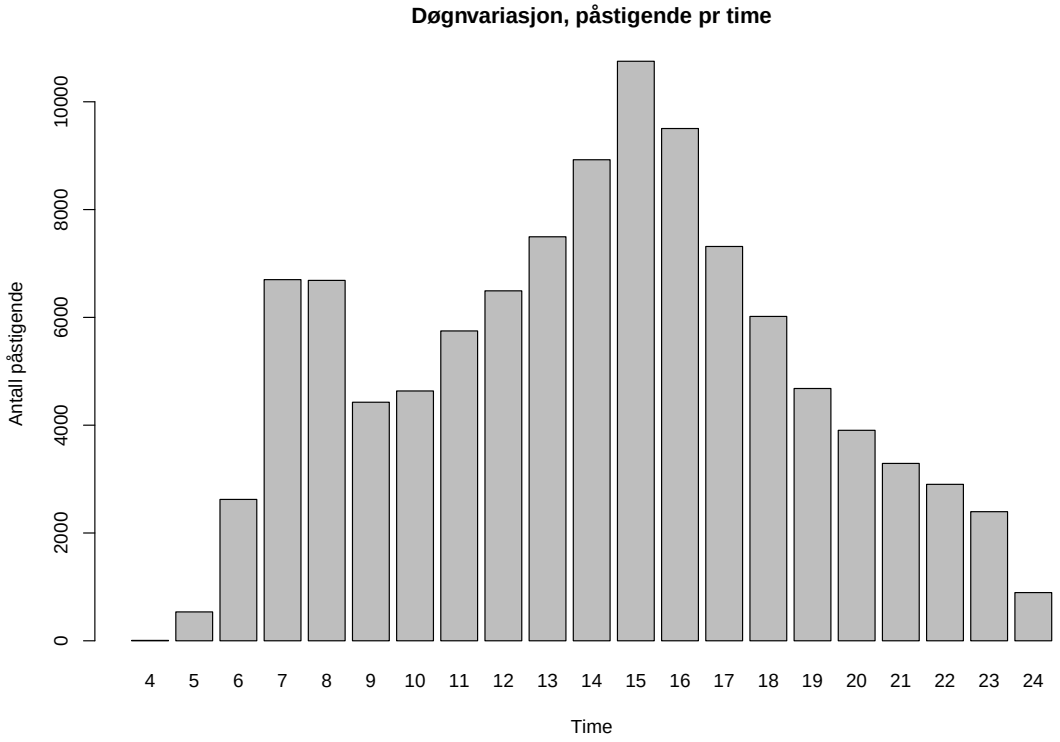
Figur 18. Årsvariasjon for daglig antall påstigende pr uke. Uke 0 er 1.-2. januar, mens uke 52 er 26.-31. desember. *Merknad: Sent i oppdraget oppdaget vi en uforutsett feil i dataleveransene. Denne går ut på at ikke alle faktisk kjørte turer er registrert i SIS i uke 4 og 5 i 2011. Dette gjelder turer både med tellinger og uten tellinger. Den etablerte predikeringsmetoden forutsetter at alle faktisk kjørte turer er registrert i SIS. Når det ikke er tilfelle, beregnes det for lave trafikktall for de aktuelle dagene. Dette påvirker også alle aggregeringer hvor ukene 4 og 5 (2011) inngår. Av disse grunner har ovenstående figur for lave tall for uke 4 og 5, og tallene som sådan har ingen verdi. Vi har likevel beholdt figuren for å illustrere viktige tema ved resultatpresentasjonen. Vi understreker også at denne type feil i dataleveransene ikke har betydning for godheten i de metodene som presenteres i denne rapporten. Men metoden tar ikke høyde for denne type feil i dataleveransene, som i utgangspunktet var opplyst ikke skulle kunne skje. Dersom man ønsker riktige tall for uke 4 og 5 må man enten få ny leveranse med komplette data, eller man må endre beregningsmetodene slik at man forholder seg til planlagte avganger i de tilfeller der man ikke har riktige data fra SIS om faktisk kjørte avganger. Det har det ikke vært rom for å gjøre i dette prosjektet.*



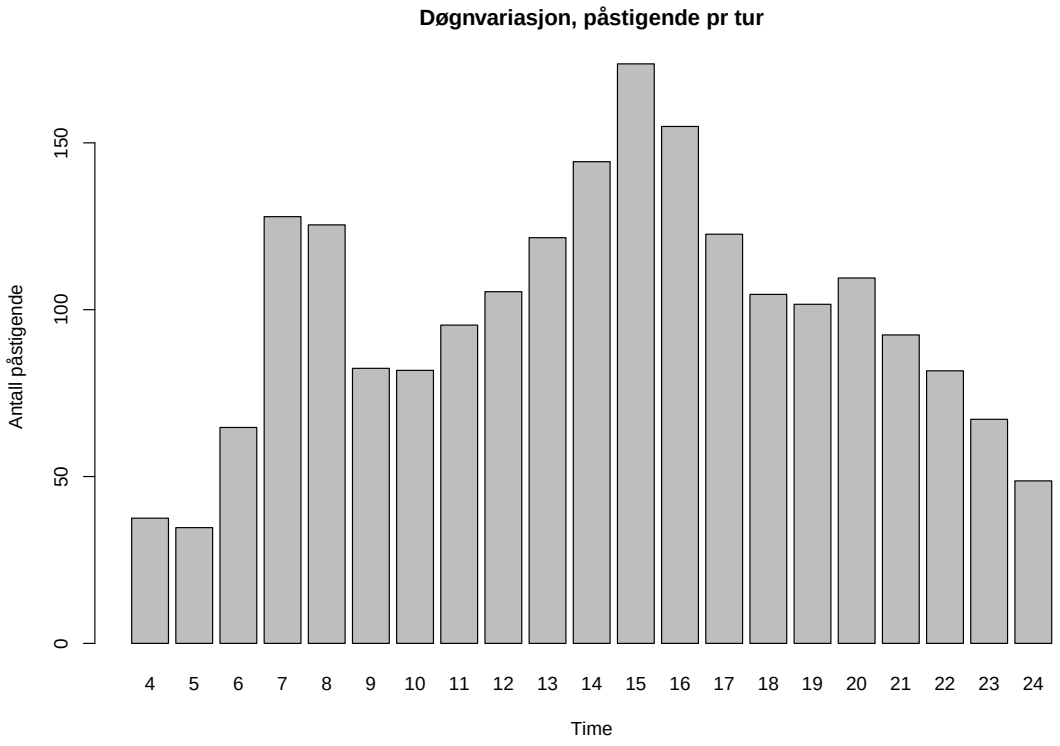
Figur 19. Årsvariasjon i påstigende pr tur. Uke 0 er 1.-2.1., mens uke 52 er 26.-31.12.



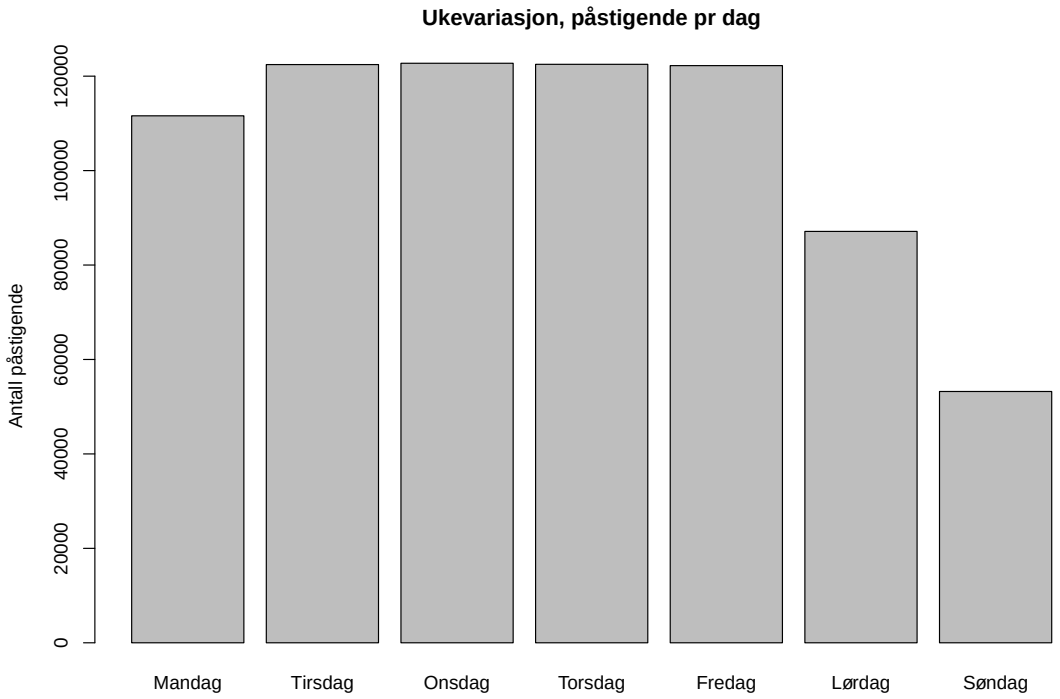
Figur 20. Turer linje 13, retning 1 over året 2011. Det øverste panelet viser antall turer pr dag, mens det nederste panelet viser andel turer med tellinger.



Figur 21. Døgnvariasjon i påstigende pr time, samlet for alle linjer og retninger. Time 4 er 4:00-4:59, time 5 er 5:00-5:59 osv. *Merknad: Sent i oppdraget oppdaget vi en uforutsett feil i dataleveransene. Denne går ut på at ikke alle faktisk kjørte turer er registrert i SIS i uke 4 og 5 i 2011. Dette gjelder turer både med tellinger og uten tellinger. Den etablerte predikeringsmetoden forutsetter at alle faktisk kjørte turer er registrert i SIS. Når det ikke er tilfelle, beregnes det for lave trafikktall for de aktuelle dagene. Dette påvirker også alle aggregeringer hvor ukene 4 og 5 (2011) inngår. Av disse grunner har ovenstående figur for lave tall for uke 4 og 5, og tallene som sådan har ingen verdi. Vi har likevel beholdt figuren for å illustrere viktige tema ved resultatpresentasjonen. Vi understreker også at denne type feil i dataleveransene ikke har betydning for godheten i de metodene som presenteres i denne rapporten. Men metoden tar ikke høyde for denne type feil i dataleveransene, som i utgangspunktet var opplyst ikke skulle kunne skje. Dersom man ønsker riktige tall for uke 4 og 5 må man enten få ny leveranse med komplette data, eller man må endre beregningsmetodene slik at man forholder seg til planlagte avganger i de tilfeller der man ikke har riktige data fra SIS om faktisk kjørte avganger. Det har det ikke vært rom for å gjøre i dette prosjektet.*



Figur 22. Døgnvariasjon i påstigende pr tur, samlet for alle linjer og retninger. Time 4 er 4:00-4:59, time 5 er 5:00-5:59 osv.



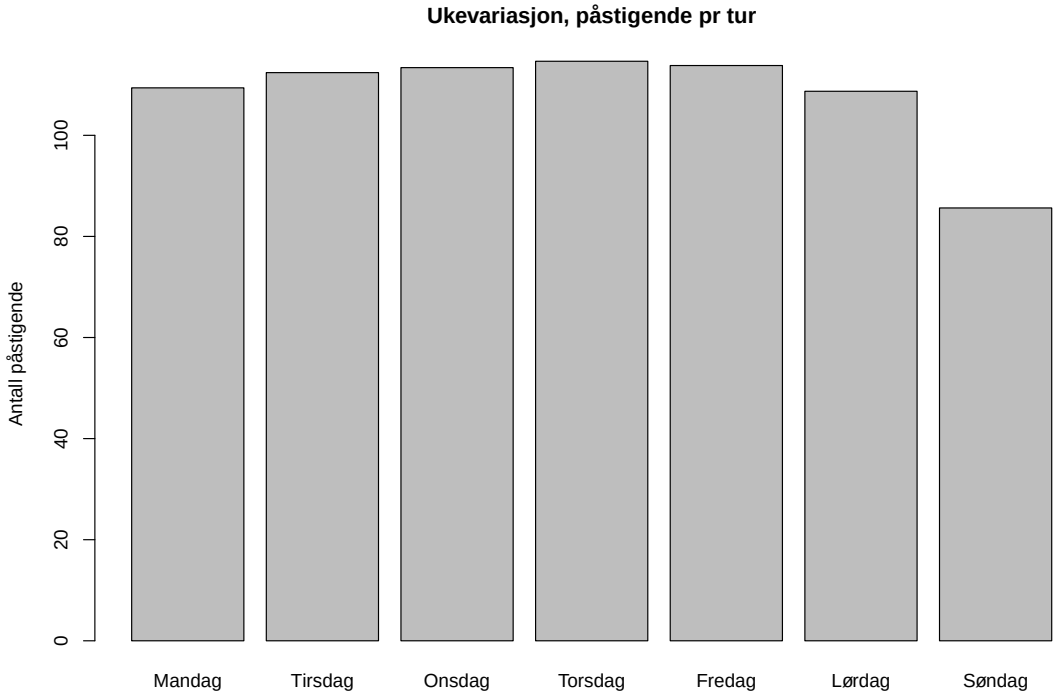
Figur 23. Ukevariasjon i daglig antall påstigende, samlet for alle linjer og retninger. *Merknad: Sent i oppdraget oppdaget vi en uforutsett feil i dataleveransene. Denne går ut på at ikke alle faktisk kjørte turer er registrert i SIS i uke 4 og 5 i 2011. Dette gjelder turer både med tellinger og uten tellinger. Den etablerte predikeringsmetoden forutsetter at alle faktisk kjørte turer er registrert i SIS. Når det ikke er tilfelle, beregnes det for lave trafikktall for de aktuelle dagene. Dette påvirker også alle aggregeringer hvor ukene 4 og 5 (2011) inngår. Av disse grunner har ovenstående figur for lave tall for uke 4 og 5, og tallene som sådan har ingen verdi. Vi har likevel beholdt figuren for å illustrere viktige tema ved resultatpresentasjonen. Vi understreker også at denne type feil i dataleveransene ikke har betydning for godheten i de metodene som presenteres i denne rapporten. Men metoden tar ikke høyde for denne type feil i dataleveransene, som i utgangspunktet var opplyst ikke skulle kunne skje. Dersom man ønsker riktige tall for uke 4 og 5 må man enten få ny leveranse med komplette data, eller man må endre beregningsmetodene slik at man forholder seg til planlagte avganger i de tilfeller der man ikke har riktige data fra SIS om faktisk kjørte avganger. Det har det ikke vært rom for å gjøre i dette prosjektet.*

5 Mulige forbedringer og utvidelser

Metodikken vi har beskrevet er utviklet for trikken i Oslo. Det er å betrakte som en prototyp metode, på den måten at det fortsatt fins rom for forbedringer. Dette er i tråd med oppdraget Norsk Regnesentral har fått av Ruter, hvor utvikling av metodikken har vært prioritert, men hvor den endelige løsning ikke nødvendigvis er funnet på alle punkter. Vi mener at metoden i hovedsak er god nok til at den kan implementeres relativt snart, men vi anbefaler at det gjøres noe mer arbeid i forbindelse med momentene vi nevner under.

5.1 Datakvalitet

Dataene vi har arbeidet med var såkalte balanserte data fra SIS-databasen. De var likevel ikke fullstendig balanserte, og vi brukte derfor en egen balanseringsalgoritme for en supplerende ba-



Figur 24. Ukevariasjon i påstigende pr tur, samlet for alle linjer og retninger

lansering av antall påstigende med antall avstigende per tur. Dette ga konsistente data, men hadde den ulempen at både totalt antall påstigende og totalt antall avstigende blei justert ned med ca. 1.7% som er langt større enn usikkerheten i beregningene for flere av størrelsene som er aggregert over en linje og over år. Det er dermed viktig å arbeide mer med dette. Metodikken vi bruker kan ta høyde for at det ved start av en tur allerede er passasjerer på trikken, og at noen passasjerer blir på trikken ved endeholdeplassen (f.eks. ved overgang mellom linje 11 og linje 12 på Majorstua). Det kan derfor være fornuftig å undersøke ubalanserte data fra SIS-databasen, for å se om det er mulig å bruke de dataene med mindre justeringer.

Ved noen planlagte avvik betjenes trikkelinjer av "buss for trikk", dvs. at buss erstatter trikken. Basert på dataene vi har fått, har det vært mulig å skille mellom erstatningsbuss og trikk, og mellom de to trikketypene. Det har imidlertid ikke vært mulig å skille mellom leddbuss og andre busser. Hvis en klarer å skille busstypene fra hverandre, vil en kunne identifisere turer på vanlige busser hvor lasten er urimelig høy, og luke ut disse som feilaktige tellinger.

Det har også vært en forutsetning for vår metodikk at vi har informasjon om de turene det skal predikeres passasjertall for. Turer som ikke logges havner ikke i SIS-databasen, men det er antatt at dette gjelder et lite antall turer. Det bør undersøkes om dataene vi har arbeidet med har vært relativt komplette med hensyn på kjørte turer, jfr. diskusjonen omkring antall turer i uke 4 og 5 2011 på linjene 13, 17 og 18. Dette har ikke konsekvenser for bruk av metoden videre, eller for prediksjonsmodellen, men har betydning for de beregningene vi har presentert som er akkumulert over turer, men ikke normert per tur.

5.2 Usikkerhet

Vi har tatt i bruk en kompleks modell, men på et vis ikke kompleks nok. Det vil typisk være positiv korrelasjon i dataene; i) mellom ulike holdeplasser på samme linje på samme tur, men

også innen samme dag og uke, ii) mellom ulike turer på samme holdeplass, så lenge turene er nær hverandre i tid, og 3) mellom ulike linjer (f.eks. er de påvirket av samme værforhold). Vi har ignorert mye av denne samvariasjonen, rett og slett fordi det blir for komplekst. Men når vi summerer over størrelser som er positivt korrelert blir usikkerheten til summen større enn om størrelsene var ukorrelert. For å ta høyde for dette beregner vi usikkerheten til estimatene i to trinn. Først beregner vi et foreløpig usikkerhetsintervall, og det vil typisk være for trangt (for lite usikkerhet). I neste trinn korrigerer vi usikkerhetsintervallet. Dette er foreløpig gjort kun for et 95% usikkerhetsintervall, og kun for antall påstigende. Det bør undersøkes nærmere om samme korreksjon kan brukes for antall avstigende og for last, og for andre usikkerhetsgrenser, eller om korreksjonsfaktoren bør reberegnes. Det finnes også en annen, og potensielt bedre og mer generell, metode å sørge for riktig usikkerhet på, ved at en faktisk modellerer korrelasjonen i dataene og tar hensyn til det i simuleringene. En slik løsning vil imidlertid være mer arbeidskrevende.

5.3 Optimal modell-kompleksitet

Metoden er laget slik at den skal fungere for holdeplasser med mange data, f.eks. data for 30000 turer, og for holdeplasser med langt færre data. Ved noen anledninger kjører for eksempel linje 11 langs ruta til linje 19, og det er i datasettet ca. 600 observasjoner tilgjengelig for disse stoppene. Antall frie parametre som skal estimeres må begrenses når det er få data. Vi har utviklet beregningsregler som avgjør hvor kompleks modellen kan være (hvor mange parametre den har) som funksjon av datamengde. Disse beregningsreglene kan trolig forbedres noe for å gi optimale prediksjoner i det lange løp.

5.4 Flere forklaringsvariable, inkludert forsinkelse

Det er rimelig å anta at forsinkelse har betydning for passasjertall, og forsinkelse på både nåværende og foregående tur kan ha betydning, samt forholdet mellom disse. Vi har undersøkt noen av disse mulighetene, men har i dette prosjektet ikke rukket å konkludere med hvordan en best skal ta hensyn til forsinkelse. I modellen vi har brukt til beregningene vi har vist, er effekt av forsinkelse derfor ignorert, men metodisk sett er det ingen ting i veien for å inkludere forsinkelse på en hensiktsmessig måte. Vi mener dette bør undersøkes noe nærmere.

Vi har også undersøkt om det var mulig å se en avvisningseffekt, dvs. om det er en tendens til at antall påstigende passasjerer blir færre når trikken er full. Det er opplagt at en slik effekt finnes i virkeligheten, men vi har så langt ikke funnet ut hvordan den best kan inkluderes i modellen. Dette kan også være verdt å undersøke nærmere.

5.5 Noen momenter med tanke på implementering

Et vesentlig element i implementeringa vil være dataflyt, tilrettelegging av data for modellering, og ikke minst automatisk eller kombinert automatisk og manuell kvalitetssjekk av data. Videre, hvis man tillater at passasjerer forblir på trikken ved endeholdeplassen, og neste turer starter med disse passasjerene på, er det viktig å etablere et godt bokholderi som tar vare på vognidentiteten. Et annet mindre moment er at når beregninger (som inkluderer stokastisk simulering og trekking av tilfeldige tall) først er gjort for en dag bør disse «låses», slik at disse ikke endres når nye beregninger gjøres for de samme turene for eksempel dagen eller måneden etter. Eventuelt kan det spares på flere versjoner av beregningene.

5.6 Bruk av metoden for buss og t-bane

Metoden er utviklet for trikk. I hovedsak bør den kunne brukes også for T-bane og buss, men metoden må prøves ut på data for hver av disse. T-banenettet er fast, og det er derfor trolig lettest å overføre metodikken til T-bane. Metodikken vil fint kunne tas i bruk selv om det er tellinger på

f.eks. kun 10% av turene. For busser vil rutenettet være mer ustabilt. Det vil kreve mer bokholderi, men i prinsippet bør metoden kunne brukes som før. I den forbindelse kan nevnes følgende: Metoden består først av å beregne beregne et mindre antall "smarte" forklaringsvariable, maksimalt 20, som er funksjoner av de opprinnelige ca. 360 variable. Det gjøres ved den såkalte redusert rang regresjonen hvor data for alle holdeplasser på en linje og en retning brukes samtidig. Deretter behandles hver holdeplass for seg i neste trinn. For busser kan vi tenke oss at noen holdeplasser, eller turer er så pass avvikende fra det normale at de ikke bør tas med i det første trinnet. Likevel kan de "smarte" forklaringvariablene brukes når antall passasjerer på en "unormal" holdeplass modelleres.

5.7 Turneringsplan for vogner med telleutstyr

Metodikken er ikke avhengig av at turer med telleutstyr er jevnt eller tilfeldig fordelt, for metoden korrigerer for om vogner med telleutstyr er overrepresentert på enkelte linjer eller enkelte dager. Om det er få vogner med telleutstyr kan det likevel være gunstig å spre disse mest mulig. For trikken er det imidlertid installert telleutstyr i så mange vogner at det ikke har noen hensikt å etablere egne turneringsplaner for trikker med telleutstyr.

A Appendiks: Detaljert teknisk beskrivelse av metodikk og algoritmer

A.1 Algoritme for balansering av på/avstigende og last

Følgende algoritme benyttes for å gjøre tellingene konsistente (se kapittel 2.5):

1. La P_i, A_i og L_i betegne påstigende, avstigende og last ved avgang for holdeplass $i = 1, 2, \ldots, n$.
2. Sett $L_1 = P_1$ og $A_1 = 0$.
3. For hver $i = 2, 3, \ldots, n-1$:
   a. Sett $D_i = P_i - A_i$
   b. Sett $A_i = \min(L_{i-1}, A_i)$
   c. Sett $P_i = \min(D_i + A_i, 0)$
   d. Sett $L_i = L_{i-1} + P_i - A_i$
4. Sett $P_n = 0$, $A_n = L_{n-1}$ og $L_n = 0$.

A.2 Redusert rang-regresjon (RRR)

Redusert rang-regresjon (RRR) er en multivariat lineær regresjonsmetode der flere responsvariable er relatert til samme utvalg av forklaringsvariable, og der den estimerte matrisen av regresjonskoeffisienter har redusert rang (en matrises rang er definert som antall lineært uavhengige kolonner i matrisen), og dermed færre frie parametre som må estimeres.

En vanlig multivariat lineær regresjonsmodell kan skrives som

$$\boldsymbol{y} = \boldsymbol{b}_0 + \boldsymbol{B}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{\epsilon},$$

der $\boldsymbol{y}$ er en kolonnevektor med q responsvarible, $\boldsymbol{x}$ er en p-dimensjonal kolonnevektor med prediktorer, og $\boldsymbol{\epsilon}$ er en q-dimensjonal kolonnevektor med feilledd med forventningsverdi $\mathbf{0}$ og kovariansmatrise $\boldsymbol{\Sigma}$. Matrisen $\boldsymbol{B}$ kan estimeres med vanlig minste kvadraters metode, og estimatet $\hat{\boldsymbol{B}}^{\mathrm{MKM}}$ vil da ha full rang lik $m = \min(p, q)$. I redusert rang-regresjon blir estimatet $\hat{\boldsymbol{B}}^{\mathrm{RRR}}$ for $\boldsymbol{B}$ restriktert til å ha redusert rang $r \leq m$.

Anta at vi har n multivariate observasjoner av respons- og prediktorvariablene, organisert i matriser $\boldsymbol{Y}$ og $\boldsymbol{X}$ med dimensjoner $n \times q$ og $n \times p$, med en rad for hver observasjon. Anta videre at disse datamatrisene er sentrert, slik at gjennomsnittet av hver kolonne er null. Estimatet beregnes med minste kvadraters metode; ved å minimere

$$Q(\hat{\boldsymbol{B}}) = \mathrm{trace}\left[(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{B}})^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{B}})\right] \tag{A.1}$$

med hensyn på $\hat{\boldsymbol{B}}$ under restriksjonen $\mathrm{rang}(\hat{\boldsymbol{B}}) = r$.

Det kan vises at estimatet $\hat{\boldsymbol{B}}$ kan skrives som en sum av matriser $\hat{\boldsymbol{B}}_k$ med rang lik én, på følgende måte:

$$\hat{\boldsymbol{B}} = \sum_{k=1}^{r} \hat{\boldsymbol{B}}_k = \sum_{k=1}^{r} \boldsymbol{\alpha}_k \boldsymbol{\beta}_k^{\mathrm{T}},$$

der hver $\boldsymbol{\alpha}_k$ er en $p \times r$-matrise og hver $\boldsymbol{\beta}_k$ er en $q \times r$-matrise. Her er $\hat{\boldsymbol{B}}_1$ rang 1-matrisen som forklarer maksimalt av variansen til $\boldsymbol{Y}$, mens $\hat{\boldsymbol{B}}_2$ er rang 1-matrisen som forklarer maksimalt av

den gjenværende variansen, osv. Videre får vi at

$$\hat{\boldsymbol{B}}^{\mathrm{T}}\boldsymbol{x} = \sum_{k=1}^{r} \boldsymbol{\beta}_k(\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{x}) = \sum_{k=1}^{r} \boldsymbol{\beta}_k \boldsymbol{z}_k, \tag{A.2}$$

der $\boldsymbol{z}_k = \boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{x},\ k = 1, \ldots, r$. Dermed kan lineærkombinasjonene $\boldsymbol{z}_k$ ses på som nye forklaringsvariable i redusert-rang modellen. Disse er sortert slik at $\boldsymbol{z}_1$ er viktigst. Med "viktigst" menes den lineærkombinasjonen av x-ene som forklarer totalvariasjonene i y-ene best. Videre er $\boldsymbol{z}_2$ nest viktigst osv.

$\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3$ osv kan altså oppfattes som nye forklaringsvariable som kan brukes til å forklare en enkelt y, og fordi disse er sortert etter viktighet vil det ikke være nødvendig å bruke så mange z-er som det er x-er. På denne måten kan man bruke informasjon på tvers av holdeplassene når man modellerer antall på- eller avstigende på en enkelt holdeplss.

A.3 Generaliserte lineære modeller (GLM) for på- og avstigende

For antall påstigende og avstigende brukes generaliserte lineære modeller (GLM). Generelt kan GLM beskrives som en utvidelse av den vanlige lineære regresjonsmodellen til tilfeller der responsen ikke er normalfordelt, f.eks. der responsen er antall (som i dette tilfellet). I GLM antar vi at responsen, si y, er observert uavhengig på fikserte verdier av forklaringsvariable, si $x_1, \ldots, x_p$. Forklaringsvariablene påvirker fordelingen til y gjennom en lineær funksjon som kalles den lineære prediktoren $\eta = \beta_1 x_1 + \cdots \beta_p x_p$, der middelverdien μ av y er gitt som en funksjon av η, dvs $\mu = f(\eta)$. Den inverse funksjonen $l(\cdot) = f^{-1}(\cdot)$ kalles lenkefunksjonen, og ulike lenkefunksjoner er aktuelle for ulike typer GLM.

I vår modell brukes en negativ binomialfordelt GLM for påstigende og en betabinomisk GLM for avstigende. Dermed antar vi at antall påstigende er negativ binomialfordelt, mens antall avstigende er betabinomisk fordelt. Som lenkefunksjoner er henholdsvis logaritmefunksjonen $\log(x)$ og funksjonen $\mathrm{logit}(x) = \log\left(\frac{x}{1-x}\right)$ valgt (dette er de mest vanlige valgene av lenkefunksjoner for disse modellene).

Vi vil nå gå nærmere inn på bakgrunnen for disse valgene. For påstigende er det mest grunnleggende at vi skal modellere et antall hendelser (dvs påstigninger) som foregår i et gitt tidsrom. Hvis disse påstigningene antas å skje som en tilfeldig prosess, og påstigninger i ulike tidsrom er uavhengige av hverandre, så er det av flere grunner naturlig å anta at antallene er såkalt Poisson-fordelte, dvs antallet $Y \sim Poisson(\mu)$, der μ vil avhenge av den lineære prediktoren η for en Poisson-GLM (vanligvis ved at $\eta = \log(\mu)$). I en Poisson-modell er både middelverdien og variansen til Y lik μ, men i noen tilfeller ser vi i dataene at variansen er større enn middelverdien. Dette fenomenet kalles overdispersjon, og oppstår gjerne på grunn av at avhengighet mellom tellingene ikke er tatt hensyn til i modellen. En måte å ta hensyn til slik overdispersjon er å bruke en negativ binomial-modell i stedet for Poisson-modellen. Negativ binomial-modellen har to parametre μ og θ, der forventning fortsatt er lik μ, mens variansen er gitt ved $\mu + \mu^2/\theta$. Dermed gis det en mulighet til å ha høyere varians enn forventning. For avstigende må man også ta hensyn til at det ikke kan være flere avstigende enn antall personer L som til en hver tid er på trikken, dvs lasten. Dette gjøres ved å modellere andelen passasjerer som går av, dvs sannsynligheten p for at en gitt passasjer går av, og forventet antall avstigende er Lp. Hvis passasjerene går av fullstendig uavhengig av hverandre, hver med sannsynlig p, har vi et såkalt binomisk forsøk, og antall avstigende vil da være binomisk fordelt med parametre L og p. Imidlertid vil vi også her ha overdispersjon på samme måte som for påstigende, så derfor brukes i stedet en såkalt beta-binomisk modell: I den binomisk modellen er forventningsverdien Lp og variansen $Lp(1-p)$, mens vi i den beta-binomiske modellen har en ekstra parameter ϕ, og variansen er gitt ved $Lp(1-p)[1 + (L-1)\phi]$. Parametrene θ og ϕ estimeres separat på hver holdeplass.

A.4 Modelltilpasning ved kombinasjon av RRR og GLM

De komplette modellene for på- og avstigende benytter en kombinasjon av RRR- og GLM-metodene forklart i kapittel A.2 og A.3. Dette foregår i følgende to trinn:

1. Nye forklaringsvariable genereres fra RRR
2. GLM (henholdsvis negativ binomial eller betabinomisk) kjøres med de nye forklaringsvariable

La P være antall påstigende for en gitt holdeplass, A være antall avstigende, og L være last ved ankomst. For påstigende skjer trinn 1 ovenfor ved at vi beregner $Y = \log(P+c)$, der c er en positiv konstant som legges til for å unngå problemer hvis $P = 0$ (merk at $\log(0) = -\infty$), og kjører RRR med Y som respons og forklaringsvariablene gitt i tabellene A.1 og A.2. Vi har brukt $c = 1$. For avstigende beregnes først $Q = \frac{A+k}{L+2k}$ der $k > 0$ er en konstant som legges til for å unngå null i nevner hvis $L = 0$. Vi har brukt $k = 0.1$. $Y = \log\left(\frac{Q}{1-Q}\right)$ brukes som respons i RRR for avstigende.

Som forklart i kapittel A.2 resulterer RRR-metoden i ett sett av nye forklaringsvariable $\boldsymbol{z}_k,\ k = 1, \ldots, r$. Rangen r bestemmes av mengden data, på følgende måte: La n være antall observasjoner (tellinger av på- eller avstigende) på en gitt holdeplass. Rangen r settes da lik $r = \lfloor n^\gamma \rfloor - 1$, der vi har brukt $\gamma = 0.25$ og $\lfloor x \rfloor$ er definert som det høyeste heltallet som er mindre enn eller lik x. Logikken bak dette er at når vi har mer data, så kan vi ta med flere forklaringsvariable. I tillegg trunkeres r til maksimalt $r = 20$ i modellen for påstigende og maksimalt $r = 10$ i modellen for avstigende. I trinn 2 ovenfor brukes disse som forklaringsvariable i GLM.

Vi har også variable separat for hver holdeplass. Den første av disse er antall stopp igjen til siste holdeplass (oppad begrenset til fem). I tillegg til dette tar vi hensyn til avhengighet mellom påstigende på tidligere holdeplasser på samme tur, og avhengighet mellom påstigende på samme holdeplass for turer som er nær i tid. Dette gjøres ved at vi beregner GLM-modellen i to trinn. Først ett trinn som beskrevet ovenfor. Dette gir oss "normalt" antall påstigende for hvert stopp, kalt Y_{est}. Deretter beregnes GLM en gang til med to ekstra forklaringsvariable X_1 og X_2 som defineres (separat for hver tur og holdeplass) ved $X_i = \log(Y_{obs}W_i + \epsilon) - \log(Y_{est}W_i + \epsilon),\ i = 1, 2$, der Y_{obs} er en matrise med observert antall påstigende for hver tur og holdeplass, Y_{est} er en matrise med tilsvarende tilpassede verdier fra modellen, W_i er vektmatriser som avhenger av avstanden enten i antall holdeplasser (for $i = 1$) eller i tid (for $i = 2$), og $\epsilon = 0.1$ (setter $\epsilon > 0$ for å unngå numeriske problemer). Elementet i rad j og kolonne k i W_1, W_{1jk}, er gitt ved $1/d_1^{\theta_1}$, der d_1 er avstand i antall holdeplasser (og $W_{1jk} = 0$ hvis $k > j$, dvs at vekten for holdeplasser senere på turen settes til null), og $\theta_1 > 0$. Tilsvarende er W_2 bestemt ved avstand i tid (men vekten settes til null utenfor samme døgn, slik at det bare tas hensyn til avhengighet mellom turer innenfor samme døgn), ved $1/d_2^{\theta_2}$, der d_2 er avstand i timer. Ved å optimalisere likelihood for ulike θ_1, θ_2 har vi valgt $\theta_1 = 1$ og $\theta_2 = 0.3$.

For enkelte stopp kan det være for lite data til at det er hensiktsmessig/mulig å ta med alle tre variablene som er separat pr stopp. En variabel tas med etter kriterier som at variabelen kan ikke være konstant over alle turer og at det ikke kan være for få observasjoner av variabelen.

A.5 Simulering av påstigende, avstigende og last

For å kunne estimere usikkerhet både for enkelte tellinger og for aggregerte størrelser (f.eks. alle avganger mellom klokka 8 og 9 på fredagsmorgener i april 2010, eller alle avganger i 2011) benyttes simuleringer fra modellen beskrevet i A.1-A.4.

Simuleringene foregår sekvensielt for holdeplasser $1, 2, \ldots, n$ på følgende måte, der P_i, A_i, L_i betegner henholdsvis påstigende, avstigende og last ved avgang på holdeplass i. Simuleringene

gjøres kun for turer uten tellinger, for turer med tellinger brukes de faktisk observerte verdiene, siden disse er kjent (ingen usikkerhet).

1. For første holdeplass $i = 1$, trekk P_1 fra estimert negativ binomial-modell for holdeplass 1, og sett $A_1 = 0$ og $L_1 = P_1$.
2. For holdeplass $i = 2, \ldots, n-1$, trekk P_i fra estimert negativ binomial-modell for holdeplass i, og trekk A_i fra estimert betabinomisk fordeling (som avhenger av L_{i-1}). Sett $L_i = L_{i-1} + P_i - A_i$.
3. For siste holdeplass $i = n$, sett $P_n = 0$, $A_n = L_{n-1}$ og $L_n = L_{n-1} + P_n - A_n = 0$.

Merk at simuleringene antar at parametrene i GLM-modellene er kjent, dvs det tas ikke hensyn til parameterusikkerhet.

A.6 Estimering av usikkerhet

I utgangspunktet kan vi bruke 2.5%- og 97.5%-kvantilene fra simuleringene som nedre og øvre grense for 95% konfidensintervall. Imidlertid vil det oppstå problemer når vi ser på summeringer over aggregerte nivåer, pga at ikke all avhengighet mellom holdeplassene er tatt hensyn til i modellen, og derfor vil konfidensintervaller direkte fra simuleringene kunne være for smale.

For å korrigere for at de opprinnelige intervallene blir for smale på aggregerte nivåer har vi brukt følgende metode. La L og U betegne nedre og øvre grenser for konfidensintervall fra simuleringene, dvs 2.5%- og 97.5%-kvantilene fra de simulerte dataene, og M betegne middelverdien fra simuleringene. Videre, for et gitt aggregeringsnivå (f.eks. dag, uke eller måned), la N betegne antall predikerte tellinger som summen på dette nivået er tatt over.

For korrigering av summer på ulike aggregerte nivåer gjør vi korrigering på logaritmisk skala (siden summene alltid vil være positive), på følgende måte: La $l = \log(L)$, $m = \log(M)$ og $u = \log(U)$. Videre, la $\theta = N^\alpha$ for $0 < \alpha < 0.5$. Intervallet (l, u) utvides på følgende måte: Sett

$$l' = m - \theta(m - l)$$

og

$$u' = m + \theta(u - m).$$

Merk at $l' = l$ og $u' = u$ hvis $\theta = 1$, ellers vil $l' < l$ og $u' > u$. La $L' = \exp(l')$ og $U' = \exp(u')$. Det korrigerte intervallet er da gitt som (L', U'). Denne korreksjonen har én ukjent parameter, nemlig α. Vi har fastsatt α ved å simulere turer der antall påstigende faktisk er observert, og for ulike α berereregne dekningsgrad for det korrigerte intervallet, dvs hvor stor andel av observerte turer som blir liggende innenfor konfidensintervallet (dette bør idéelt sett ligge på 95%). Vi har funnet at $\alpha = 0.16$ gir tilfredstillende dekningsgrad for ulike aggregeringsnivåer som tur, dag, uke og måned. Bakgrunnen for valget av den parametriske formen N^α er at for en sum av uavhengige normalfordelte variable, så vil standardfeilen (og dermed vidden på et konfidensintervall) synke proporsjonalt med $N^{0.5}$. For avhengige variable vil det være naturlig å tenke seg at standardfeilen synker en del saktere, men ellers på en lignende måte, og det vil da være naturlig å utvide konfidensintervallene med N^α der $\alpha < 0.5$. Logaritmetransformasjonen tas for å komme nærmere normalfordeling.

Det kan også være av interesse å se på usikkerhet for differanser, hvis vi f.eks. ønsker å teste statistisk om det er en endring over tid i passasjertrafikken. Eksempelvis kan vi ønske å teste på 5% signifikansnivå om det har vært en endring i totalt antall påstigende fra februar 2010 til februar 2011. Man kan da ta utgangspunkt i simulerte realisasjoner av differanser fra februar 2010 til februar 2011. Dette usikkerhetsintervallet korrigeres på lignende måte som for summer,

med samme begrunnelse. Imidlertid vil differansene kunne være negative, så det gir ikke mening å log-transformere, slik vi har gjort for summer. Vi korrigerer dermed direkte på L og U, dvs setter $L' = M - \theta(M - L)$ og $U' = M + \theta(U - M)$. Vi bruker fortsatt formen $\theta = N^{\alpha}$, og har funnet at $\alpha = 0.17$ gir tilfredstillende dekningsgrad for observerte differanser av påstigende.

A.7 Definisjon av starttidspunkt for en tur

Alle forklaringsvariablene på tur-nivå avhenger av datoen og/eller starttidspunktet t for turen. Starttidspunktet t er start ved første holdeplass langs linja, målt i timer med desimaler (slik at f.eks. kvart på fire betegnes som 15.75). Hvis en tur ikke starter på første holdeplass langs linja, beregnes et hypotetisk starttidspunkt t. Dette gjøres for å kunne ha en entydig tidsordning av turene. Merk også at driftsdøgnet for trikken er definert fra første morgenavgang til siste avgang sen kveld/natt, også dersom avgangene er etter midnatt. Dermed vil en avgang klokka ett natt til lørdag defineres som en fredagsavgang, og ha $t = 25$.

A.8 Liste over alle forklaringsvariable

Tabell A.1 viser en liste over alle grupper av forklaringsvariable på tur-nivå, med antall forklaringsvariable hver gruppe består av. Totalt er det 359 forklaringsvariable på tur-nivå. Tabell A.2 viser forklaringsvariablene som er på holdeplass-nivå. Tabell A.3 viser en liste over spesielle dager som tas hensyn til i modellen. Her kunne vi i prinsippet også tatt hensyn til f.eks. skolens høst- og vinterferier.

Navn	Forklaring	Antall
trend	lineære, kvadratiske og kubiske trender: d, d^2 og d^3 for dag d	3
sinyear	årlig periodisk effekt: $\sin(2\pi id/365), i = 1, \ldots, 6$, dag d	6
cosyear	årlig periodisk effekt: $\cos(2\pi id/365), i = 1, \ldots, 6$, dag d	6
sin0summer	sommereffekt: $\sin(2\pi(u-24)/18)$, uke $u = 25, \ldots, 32$	1
sinsummer	sommereffekt: $\sin(2\pi i(u-24)/8), i = 1, 2$, uke $u = 25, \ldots, 32$	2
cossummer	sommereffekt: $\cos(2\pi i(u-24)/8), i = 1, 2$, uke $u = 25, \ldots, 32$	2
days.in.week	ukedagseffekt (konstantledd for tirsdag, onsdag, ..., søndag)	6
days.in.week.season	interaksjon: hverdag/lørdag/søndag $\times$ periodiske effekter (årlig/sommer)	26
days.in.week.trend	interaksjon: hverdag/lørdag/søndag $\times$ lineære/kvadratiske trender	4
special.days	effekt av spesielle dager, bl.a. helligdager (se egen tabell)	37
sinday	daglig periodisk effekt: $\sin(2\pi it/24), i = 1, \ldots, 10$, starttid t	10
cosday	daglig periodisk effekt: $\cos(2\pi it/24), i = 1, \ldots, 10$, starttid t	10
days.in.week.sinday	interaksjon: days.in.week $\times$ sinday (se ovenfor)	36
days.in.week.cosday	interaksjon: days.in.week $\times$ cosday (se ovenfor)	36
special.days.sinday	interaksjon: special.days $\times$ sinday (se ovenfor)	37
special.days.cosday	interaksjon: special.days $\times$ cosday (se ovenfor)	37
days.in.week.season.sinday	interaksjon: days.in.week.season $\times$ sinday (se ovenfor)	26
days.in.week.season.cosday	interaksjon: days.in.week.season $\times$ cosday (se ovenfor)	26
sinyear.sinday	interaksjon: sinyear $\times$ sinday (se ovenfor)	12
sinyear.cosday	interaksjon: sinyear $\times$ cosday (se ovenfor)	12
cosyear.sinday	interaksjon: cosyear $\times$ sinday (se ovenfor)	12
cosyear.cosday	interaksjon: cosyear $\times$ cosday (se ovenfor)	12
Totalt antall		359

Tabell A.1. Forklaringsvariable på tur-nivå

Navn	Forklaring
x.sep1	$\max\{$antall holdeplasser igjen før siste holdeplass$, 5\}$
x.sep2	X_1 definert i siste avsnitt i appendiks A.4
x.sep3	X_2 definert i siste avsnitt i appendiks A.4

Tabell A.2. Forklaringsvariable på holdeplass-nivå

1	1. januar
2	Dag etter 1. januar
3	Fredag før palmesøndag
4	Lørdag før palmesøndag
5	Palmesøndag
6	Mandag i påskeuka
7	Tirsdag i påskeuka
8	Onsdag i påskeuka
9	Skjærtorsdag
10	Langfredag
11	Påskeaften
12	1. påskedag
13	2. påskedag
14	Tirsdag etter påske
15	Dag før 1./17. mai
16	1. mai
17	Dag etter 1./17. mai
18	17. mai
19	Onsdag før Kristi Himmelfartsdag
20	Kristi Himmelfartsdag
21	Fredag etter Kristi Himmelfartsdag
22	Lørdag etter Kristi Himmelfartsdag
23	Søndag etter Kristi Himmelfartsdag
24	Fredag før pinse
25	Lørdag før pinse
26	1. pinsedag
27	2. pinsedag
28	Tirsdag etter pinse
29	Nest siste søndag før jul
30	Siste søndag før jul
31	Lille julaften
32	Julaften
33	1. juledag
34	2. juledag
35	Søndag i romjul
36	Hverdag i romjul
37	Nyttårsaften

Tabell A.3. Spesielle dager