SINTEF
Postadresse:
Postboks 4760 Torgarden
7465 Trondheim

Sentralbord: 40005100

info@sintef.no

Foretaksregister:
NO 919 303 808 MVA

Prosjektnotat

Mobiltelefonitellinger

Bekrivelse og gjennomgang av data

VERSJON	DATO
1	2021-03-10

FORFATTERE
Andreas Dypvik Landmark
Petter Arnesen

OPPDRAGSGIVER	OPPDRAGSGIVERS REF.
Ruter AS	Lene Jahnsen (18/00776)

PROSJEKTNR	ANTALL SIDER OG VEDLEGG:
102016932	9 + 1 vedlegg

SAMMENDRAG

Notatet sammenfatter datasettene som er anskaffet fra Telia i dette prosjektet med en beskrivelse av metadata og innhold, samt en kortfattet kvalitetssjekk av dataene.

SINTEFs konklusjon er at størrelse på områder og aggregering ser ut til å være i god balanse med graden av sensurering i datasettet. Det vil si, dataene er på ønsket detaljnivå uten for mye sensurering, og dette virker som et godt grunnlag for videre analyser.

UTARBEIDET AV	SIGNATUR
Andreas Dypvik Landmark	

GODKJENT AV	SIGNATUR
Odd André Hjelkrem	

PROSJEKTNOTAT NR	GRADERING
N-10/20	Åpen

Innholdsfortegnelse

SINTEF

1 Introduksjon

Hensikten med dette notatet er å beskrive datasettet fra Telia samt gjøre en innledende kvalitetssikring av datasettet. Gjennom dette arbeidet har det blitt avdekket noen små uoverensstemmelser som har blitt korrigert av Telia slik at datasettet slik det foreligger nå ikke har noen kjente avvik fra det som ble bestilt.

Notatet benytter omtrentlig samme metoder som notatet «N-07/19 Mobility Analytics Undersøkelse av mobildata for Ruter AS». For utdypelser av kunnskapsgrunnlaget og en lengre utredning av kvalitetstriangelet «Sted – Tid – K-anonymitet» så henvises det til N-07/19.

2 Beskrivelse av datasettet

Bestillingen var «OD-matriser med reisestrømmer fra alle delområder/delbydeler til alle delområder/delbydeler for hver time og type dag basert på mobiltelefontellinger». Det var ønsket å bestille data fra hverdag og helg. For å motvirke en stor andel sensureringer i datasettet ble alle reiser for totalt 8 onsdager og 8 lørdager slått sammen til en summert onsdag og en summert lørdag, se kapittel 3.1 for detaljer. Ukene 42-45 i 2019 og uke 3-6 i 2020 ble lagt til grunn.

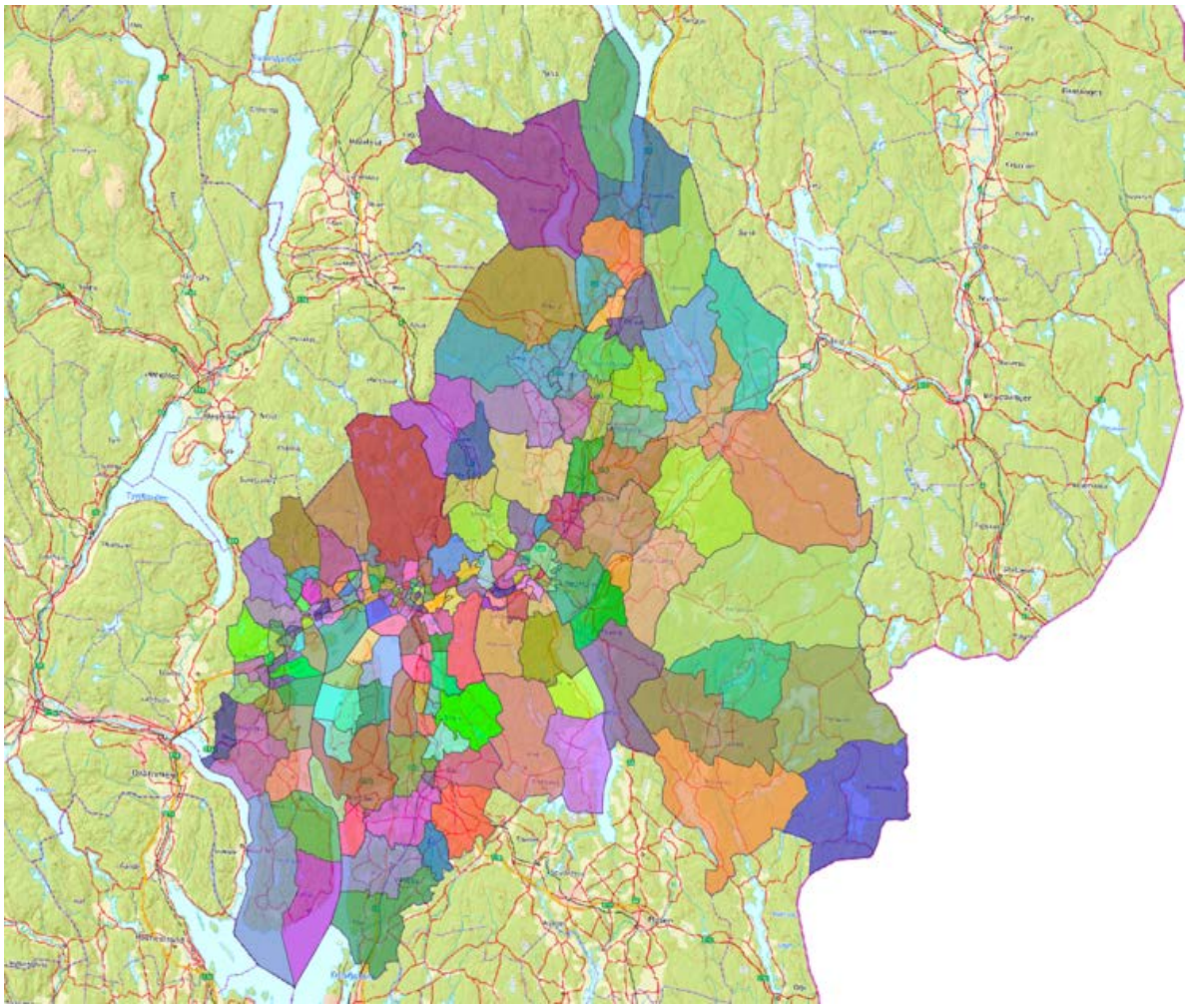
Uttrekket er basert på samme definisjon av reiser og begrensninger som beskrevet tidligere (se N-07/19). Det er i praksis levert *to* datasett, ett som inneholder data for onsdager – og ett for lørdager. Disse to settene har identisk metadata og følgende kolonner:

Variabel	Beskrivelse
utc_hour	Tidspunkt for målingen (merk i UTC, altså én time før norsk vintertid, UTC har ikke sommertid).
origin_al3_code	Avreiseområde som områdekode (4+2+00). Hvis odde antall siffer så er den første 0 utelatt (altså område 301xx er egentlig 0301xx – altså Oslo kommune). **NB! Telia benytter kommunenummer fra før kommunereformen.**
origin_al3	Avreiseområde med navn. **NB! Denne alene, uten tall vil ikke være unik. For eksempel stedet «Ås» finnes i flere steder.**
destination_al3_code	Destinasjonsområdet, se for øvrig over.
destination_al3	Destinasjonsområdet, se for øvrig over.
people	Antall mennesker på kombinasjonen Orgin, Destination og utc_hour.

Det vil si at for hver time gjennom døgnet (`utc_hour`) så finnes det en oppføring av avreisested (`origin_al3_code` og `origin_al3`) og destinasjon (`destination_al3_code` og `destination_al3`) og hvor mange (`people`) som har gjennomført akkurat denne reisen med oppstart fra avreise i denne timen. Data i uttrekket er allerede skalert opp til befolkning, og `people`-kolonnen er summen av alle åtte dagene – så et estimat for én gjennomsnittlig onsdag eller lørdag vil være 1/8 av verdien.

Det er verdt å merke seg at der hvor data er sensurert (eller det ikke finnes noen reisende) så er linjen utelatt. Det vil si at for et område hvor det er mindre enn 5 reiser i en gitt time (etter at 8 dager er lagt sammen) så

denne relasjon ikke finnes i datasettet. Områdene, totalt 200, definert for de to uttrekkene er vist i Figur 1 under.



Figur 1 Oversiktskart over områder

Oversikt over områder ved navn:

Alfaset	Haneborglia	Nannestad	Svartskog
Algarheim	Hasle	Nesbru	Sydøstre hurum
Aurskog	Haslum	Nesøya	Syverstad
Bekkelaget	Hebekk	Nordby	Såner
Berger	Heggedal	Nordby	Søndre hakadal
Billingstad	Holmen	Nordbygda	Søndre høland
Bjerke	Holmenkollen	Nordkisa	Sørkedalen
Bjerke	Holter	Nordmarka	Sørum
Bjølsen	Homansbyen	Nordre hakadal	Sørumsand
Bjørkelangen	Hosle nord	Nordstrand	Tangen
Bjørnemyr	Hosle sør	Nordøstre hurum	Torshov
Bjørnholt-kurland	Hovinhøgda	Retten	Tårnåsen
Blaker	Hurdal	Riddersand	Tåsen
Blakstad	Huseby	Ris	Tøyen
Blystadlia	Hvalstad	Rodeløkka	Udnes
Borgen	Hvam	Rotnes	Ullern

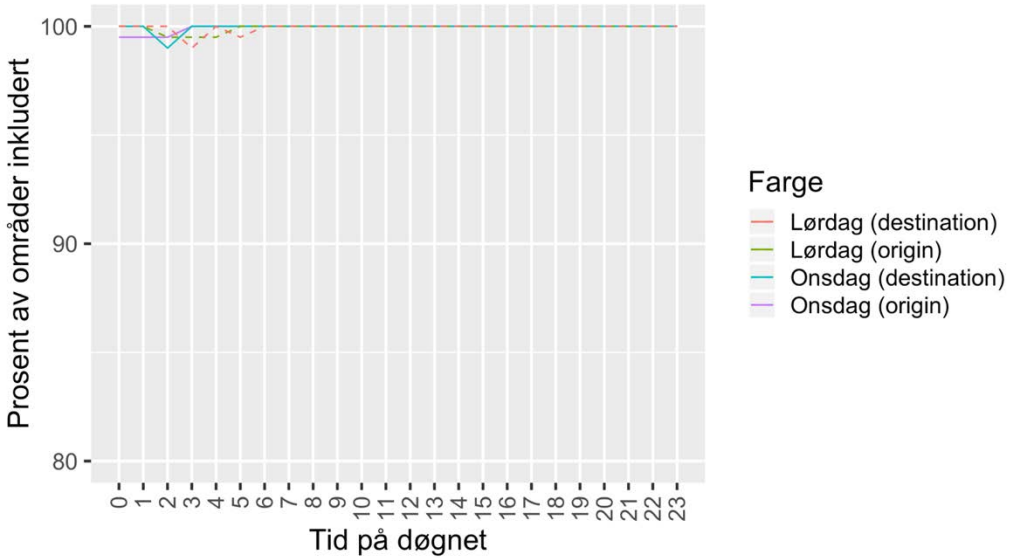
Borgen
Borgen
Bygdøy
Bærums verk
Dal
Dalen
Drengsrud
Drøbak
Dønski-rud
Eidsvoll verk
Fagerborg
Feiring
Fenstad
Filipstad
Finstad
Finstad-losby
Fjellstrand
Fjerdingby
Flateby
Fossum
Frogn nord
Frogn syd
Frogner
Frogner
Fusdal
Gamle aker
Gamlebyen
Gardermoen
Gjedsjø/Kråkstad
Gjelleråsen
Gjelleråsen
Gjerdrum
Grav
Grefsenlia
Grefsenmarka
Greverud
Grorud
Grønland
Grünerløkka
Høvik
Ila
Jaer
Jar
Jessheim
Jong
Kampen
Kirkebygda
Kirkerud-sollihøgda
Kjeller
Kjenn-fjellhamar
Kløfta
Kolbotn
Kolsås
Kroer
Langerud
Langhus
Leirsund
Lilleaker
Lindern
Ljansbyen
Loenga
Lommedalen
Lysaker
Lysås-løken
Løkeberg-blommenholm
Løken
Løvenstad
Majorstuen
Manglerud
Marienlyst
Midtbygda
Mogreina
Myklerud
Rud
Rudene
Rykkinn
Råholt
Røabyen
Rømskog
Sagene
Sand
Sandaker
Sandvika-valler
Sem
Sentrum
Sentrum 1
Sentrum 2
Sentrum 3
Setskog
Siggerud
Sinsen
Skaugum
Skedsmokorset
Ski øst
Skillebekk
Skjetten
Skogbygda
Skårer
Skøyen
Slattum
Slependen-tanum
Sofiemyr
Solberg
St.hanshaugen
Stabekk
Stalsberg
Stortorget
Strandsåsen
Ulsholt
Ulven
Uranienborg
Vestby
Vestbygda
Vestre hurum
Vettre
Vigernes
Voll
Volla
Vollen
Vormnes
Vålerenga
Ytre enebakk
Årnes
Ås
Ås
Åsen
Åsenhagen
Østbygda slemmestad
Østbygda åros
Østensjø
Østerås-eiksmarka
Østmarka
Østre bærumsmarka
Østsida
Øyene

3 Kvalitetssikring med deskriptiv statistikk

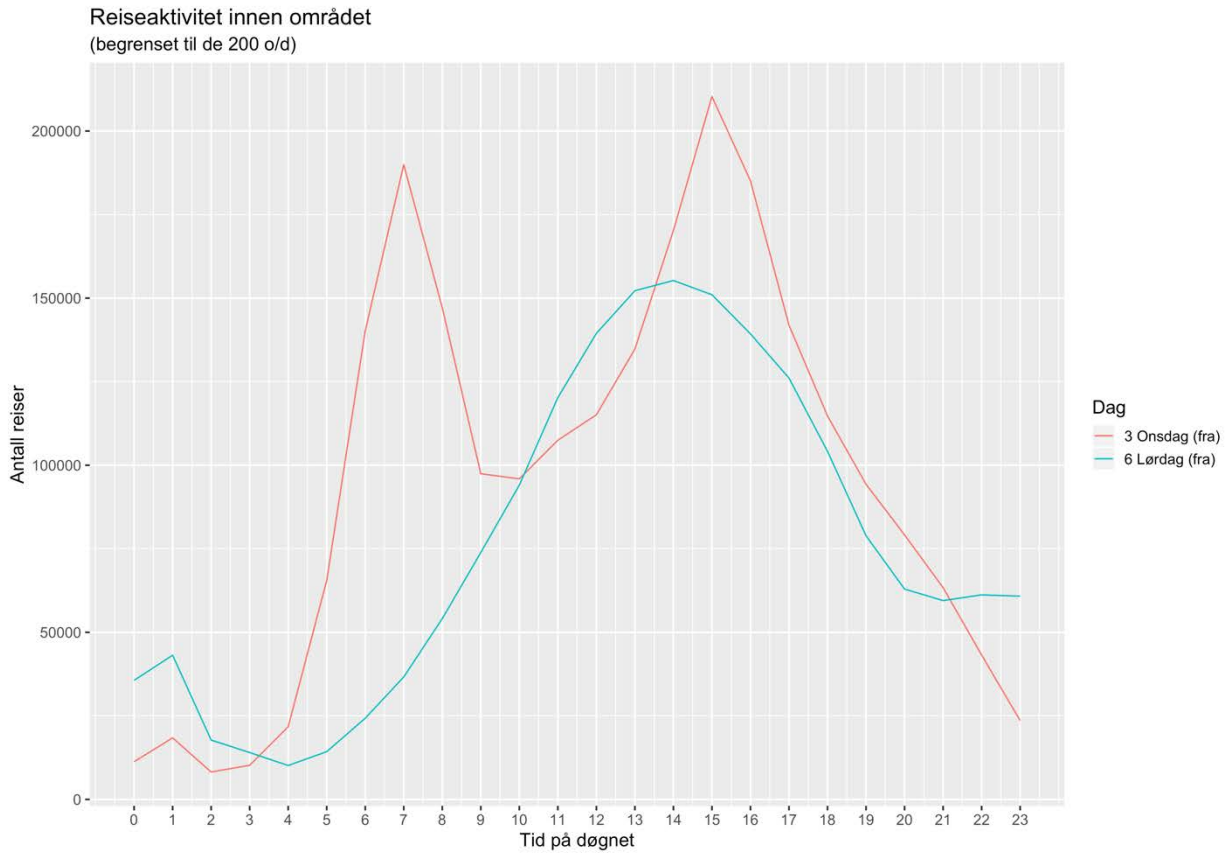
For datasettet som representere hverdager (8 onsdager) er det 1 431 unike områder i O+D. I datasettet for helg (8 lørdager) er det 1260 unike områder i O+D. Antallet unike områder overstiger 200, fordi reiser som starter eller slutter utenfor de 200 predefinere områder telles. Datasettene inneholder altså reiser som har opprinnelse innenfor de 200 områdene, men slutter et annet sted i landet – og motsatt.

Hvis vi ser på de 200 områdene i Shapefilen `Delområder_Bosatte2018_GmlBykommunenr_OA`, så er alle 200 områdene med som både O og D i begge datasettene, når man ser døgnet under ett.

Hvis man deler det opp time-for-time, så får man dekningsgradene vist i Figur 2 for hver time. Her ser vi at bortsett fra natt (perioden 23-6) så er så godt som alle områdene representert både som origin og destinasjon i matrisen gjennom hele døgnet. Det betyr ikke at matrisen er «fullkoblet» (altså at det er reiser mellom alle relasjoner), men at alle relasjonene er representert med med én eller flere origin/destinasjoner.



Figur 2 Dekning i de 200 områdene time for time (stiplet linje er lørdag)



Figur 3 Reiseaktivitet i de 200 områdene gjennom døgnet

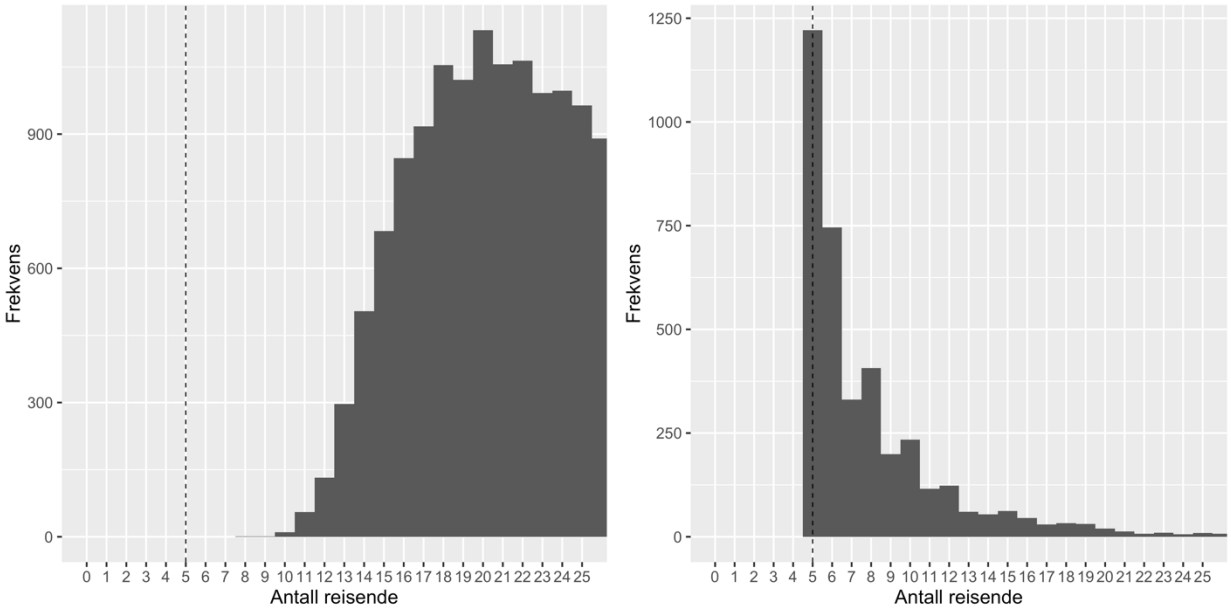
Figur 3 viser total reiseaktivitet (her beregnet fra avreise) gjennom døgnet – fordelt på en strek for onsdag (rød) og en for lørdag (grønn). Vi ser klart at onsdagsgrafen er bimodal med rushtider, mens lørdagsgrafen er tilnærmet unimodal med hovedvekt av reiseaktivitet rundt klokken 14:00. Dette er som forventet, i tråd med tidligere resultater.

Som en kvalitetssjekk på om disse to datasettene fanger variasjonen mellom ukedag og helg, så kan vi sammenligne med «Reisevaner i Osloområdet. En analyse av den nasjonale reisevaneundersøkelsen 2013/14» (PROSAM rapport 218, http://www.prosam.org/index.php?page=report&nr=218) som sier at gjennomsnittlig antall reiser på lørdager er 3,1 mot 3,6 på onsdager.

Det gir et forholdstall på 3,1/3,6 ≈ 0,86. Hvis man bare summerer opp antall i de to datasettene så får man et forholdstall på ≈ 0,84. Noe som må sies å være ganske godt samsvar.

3.1 Sensurering

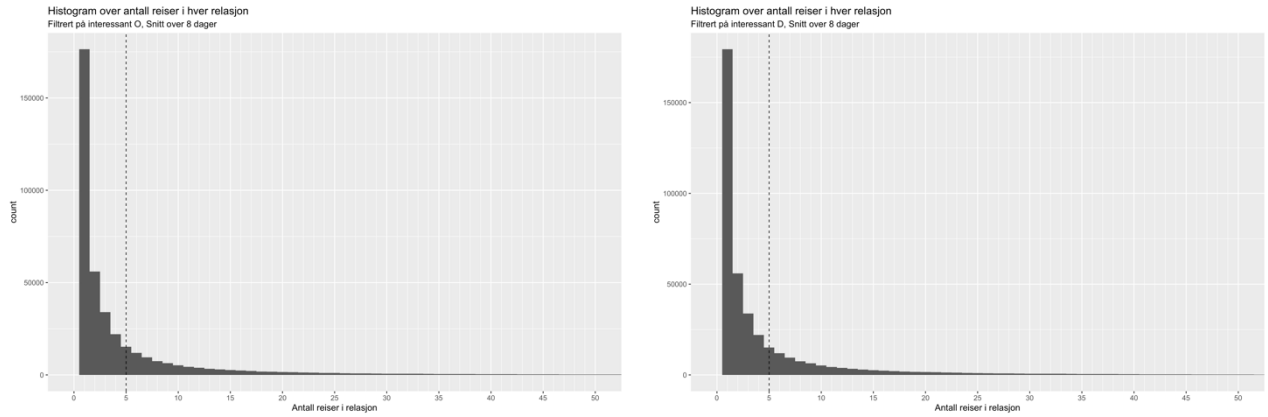
Tidligere har vi benyttet histogrammer med *fordelingen* av antall reisende i matrisen – altså hvor mange som står i hver enkelt celle. Hvis man henter ut data fra én dag så kan man se på hvordan fordelingens venstre flanke ser ut, vil man i et kraftig sensurert datasett se en skarp kant ved antall reiser = 5.



Figur 4 Eksempel på tetthetsplott for OD-matrise. Frekvens på y-aksen henviser til antall OD-par observert med et gitt antall reisende (x-aksen) i datasettet.

I Figur 4 ser vi på et eksempel for tetthetsplott for en OD-matrise, og den venstre del av figuren viser en tetthetsfordeling hvor det ikke er noe markant «kutt» i nedre del – som gjør at man kan anta at man ikke har fått en kraftig sensurert OD-matrise (altså bias mot få reisende). Høyre del av figuren viser derimot et kuttpunkt hvor det ser ut som om tettheten (figurens «mode») ligger innenfor det avkuttede området (vist med stiplet linje). Det kan tilsi at matrisen har en bias hvor relasjoner med få reisende er sensurert.

For å unngå denne type bias, så har Telia lagt 8 dager oppå hverandre og summert antall reisende i hver relasjon *før* man gjennomfører sensurering for relasjoner med mindre enn fem reisende. I tolkningen så kan man jo si at for å skalere tilbake til én dag så deler man matrisen på 8 (til eksempel så vil altså 16 reisende i matrisen tilsvarer 16/8=2 reisende per dag). Da får man ikke lengre heltall i matrisen, og minste-tallet på 5 blir nå i praksis 5/8, se Figur 5 for tetthetsplot i dette tilfellet.

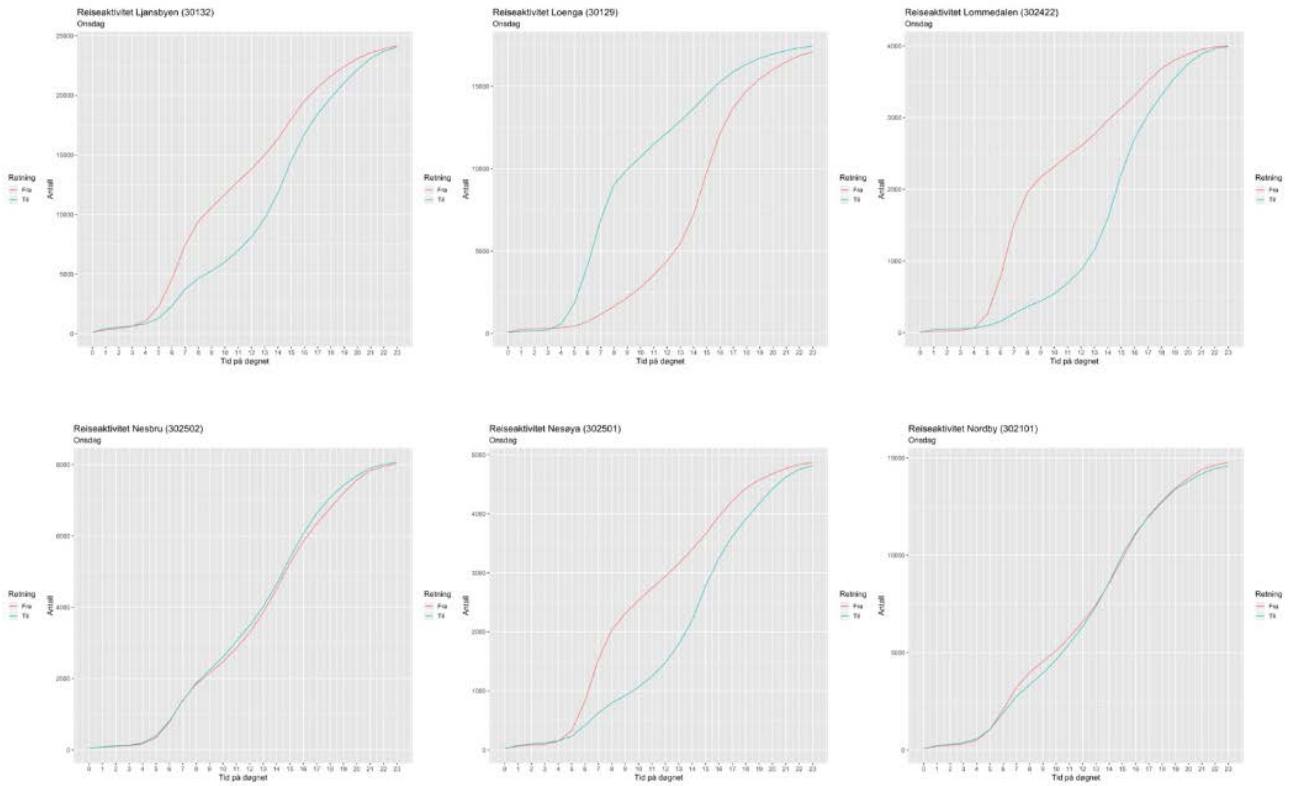


Figur 5 Tetthetsplott for datauttrekket på 200 områder, filtrert for hhv interessant O (venstre) og D (høyre)

Dette gjør det vanskeligere å bedømme hvordan k-anonymitet treffer, fordi det er også ikke er ufornuftig å anta at den hyppigste antall reisende er få (for eksempel gjennom hele natten). Men sammenholdt med dekningsgraden vist i Figur 2, som viste at det stort sett er én eller flere relasjoner gjennom døgnet så antar vi at vi ikke har noen betydelig grad av sensurering.

3.2 Likevekt

Vi har tidligere også benyttet at innenfor et døgn så antar vi at antallet reiser inn og ut av et område er omtrent symmetrisk – altså at det netto ikke er en stor endring. Vedlagt til dette notatet finnes høyoppløselige figurer for alle de inkluderte 200 områdene, med et eksempel for noen områder vist i Figur 6. Ved visuell inspeksjon av disse figurene ser man at dette ser ut til å stemme godt også for disse datasettene, der de røde kurvene for akkumulert antall reiser ut av områdene og de blå kurvene for akkumulert antall reiser inn i områdene ender på omtrent samme verdi etter 24 timer.



Figur 6 Akkumulert reiseaktivitet for samtlige områder (onsdag) (bilde med høy oppløsning vedlagt)

Teknologi for et bedre samfunn

www.sintef.no