

Prosjektnotat

Mobility Analytics

Undersøkelse av mobildata for Ruter AS

VERSJON

1.0

DATO

2019-12-16

FORFATTER(E)

Petter Arnesen
Andreas D. Landmark
Odd A. Hjelkrem
Carl Södersten

OPPDRAGSGIVER(E)

Ruter AS

OPPDRAGSGIVERS REF.

199380

PROSJEKTNR

102016932-2

ANTALL SIDER OG VEDLEGG:

38

SAMMENDRAG

Hovedhensikten i dette prosjektet har vært «å utforske analysemulighetene dataen fra teleoperatørene kan gi [Ruter], og sammenlikne kvaliteten på datagrunnlaget med dagens datagrunnlag fra reisevaneundersøkelsene (MIS)».

Våre hovedkonklusjoner fra prosjektet er:

- Å bruke mobildata til å undersøke/understøtte relative endringer vil være av verdi.
- Mobildata representerer den datakilden der absolutte reisetall telles mest direkte, kun oppskalering fra markedsandel(er) for mobilabonnement benyttes.
- Det er utfordrende å sammenligne tellinger fra mobildata med mer tradisjonelle datakilder som RVU og MIS, primært knyttet til ulik definisjon av ev reise.
- Å bestille mobildata krever mye innsikt i konkrete problemer man skal undersøke. To ulike problemstillinger vil i de alle fleste tilfeller kreve to separate bestillinger.

UTARBEIDET AV

Petter Arnesen

GODKJENT AV

Terje Reitaas

PROSJEKTNOTAT NR

N-07/19

SIGNATUR



SIGNATUR



GRADERING

Åpen

Historikk

VERSJON	DATO	VERSJONSBEKRIVELSE
0.9	2019-12-02	Versjon til kvalitetssikring

1.0	2019-12-04	Endelig versjon
-----	------------	-----------------

Innholdsfortegnelse

1	Introduksjon og bakgrunn	4
1.1	Notatets struktur	6
2	Mobiltelefonidata som datakilde i transportanalyse.....	7
3	Bestilling	9
3.1	Områder, tidsvinduer og aggregering.....	9
3.2	Beskrivelse av data.....	10
3.3	Definisjon av reiser	12
4	Validering.....	15
4.1	Sensurering av data	15
4.2	Stabilitet.....	17
4.3	Likevekt.....	18
4.4	Sammenligning mot befolkningstall	19
4.5	Sammenligning mot passasjertellinger	20
5	Analyser	25
5.1	Antall reiser som gjennomføres i Oslo og Akershus	25
5.2	Sammenligning reisevaneundersøkelse/MIS og mobildata.....	25
5.3	Reisestrømmer i de tre korridorene (Vest, Nord-øst, Sør)	27
5.4	Hvor mange reiser over bomringen.....	29
5.5	Reiser Asker Sentrum.....	30
5.6	Reiser med tog.....	32
6	Mobiltelefonidata som datakilder i transportanalyser	35
7	Konklusjoner	37
8	Kilder	38

1 Introduksjon og bakgrunn

Dette notatet oppsummerer arbeidet som er gjennomført i prosjektet Mobility Analytics for Ruter. Hovedmålsetningen i dette prosjektet har vært «å utforske analysemulighetene dataen fra teleoperatørene kan gi [Ruter], og sammenlikne kvaliteten på datagrunnlaget med dagens datagrunnlag fra reisevaneundersøkelsene (MIS)».

Basert på målsetningen ble det utformet et pilotprosjekt hvor mulig kjøp av data fra Telenor og Telia ble undersøkt. Det ble avgjort å kjøpe data fra Telia i dette prosjektet, primært basert på to av deres eksisterende algoritmer for å beskrive mobilitetsmønster, i tillegg til at utviklingsarbeid på reisemiddel-deteksjon av togreiser ble gjennomført i samarbeid med Telia. Det ble kjøpt data for fire onsdager i februar, inndelt i tre datasett: ett som er rene tellinger i de 40 bydel/kommunene som utgjør Ruters fire markedsområder, vist i Figur 1, ett som er OD-matriser (Origin-Destination) for de samme 40 bydel/kommunene, og ett som er OD-matrise fra/til Akser Sentrum fra/til alle andre grunnkretser som utgjør Ruters fire markedsområder. Alle disse datasettene ble levert med tellinger time for time for de fire dagene.

Main marked area	Sub-marked areas	Bydel/Kommune	
Indre By	Indre by 0	Oslo Sentrum	
		Gamle Oslo	
		Grünerløkka	
		Sagene	
		St. Hanshaugen	
Nordøst	Nordøst 1	Frogner	
		Bjerke	
		Grorud	
		Stovner	
	Nordøst 2	Alna	
		Rælingen	
		Lørenskog	
		Skedsmo	
	Nordøst 3	Nittedal	
		Gjerdrum	
		Fet	
		Sørum	
	Nordøst 4	Ullensaker	
		Nes	
		Aurskog-Høland	
		Nannestad	
Eidsvoll			
Hurdal			
Vest		Vest 1	Ullern
			Vestre Aker
	Vest 2	Nordre Aker	
		Bærum	
		Asker	
Vest 3	Hurum		
	Røyken		
Sør	Sør 1	Østensjø	
		Nordstrand	
		Søndre Nordstrand	
	Sør 2	Nesodden	
		Ski	
		Oppegård	
	Sør 3	Enebakk	
		Frogn	
		Ås	
		Vestby	

Figur 1: Fire markedsområder i Oslo og Akershus og deres oppdeling- Ruter

Basert på disse datakildene, koblet og sammenlignet med andre datasett, har vi forsøkt å beskrive kvaliteten på mobiltelefoni som kilde for mobilitet og trukket opp noen betraktninger i hvilke hensikter og metoder dette vil være relevant for Ruter.

Generelt kan man si at reiseidentifikasjon fungerer godt når det er definerte O-er (origins) og D-er (destinations), i denne studien er Oslo og Akershus i 40 separate geografiske områder. Hovedutfordringene ved datakilden ligger i at det ikke er et godt sammenfall mellom definisjonene man forventer i det tradisjonelle transportanalysefaget og hva som er mulig å legge til grunn for algoritmene som benyttes for å trekke ut og definere reiser ifra mobiltelefonidata. Dette utfordrer hvordan man skal tolke resultater, hvordan mobildata kan sammenlignes med kjente datakilder som RVU og MIS, men også hvilke problemstillinger som er relevante.

Reisevaneundersøkelser er av natur ganske omfattende å gjennomføre, så alternative datakilder til svar på én eller flere av de typer problemstillinger som svares ut av RVUer vil kunne gjøre at Ruter kan korte tiden fra spørsmål til svar – og gjør databaserte beslutninger på mindre problemstillinger eller hyppigere enn oppdatering av MIS og de nasjonale RVUene. Ruter gjennomfører ca. 9000 personintervjuer hvert år for å kartlegge og forstå befolkningens reisevaner i Oslo og Akershus. Her telles stedfestede reiser, dvs. reiser identifiseres som fra/til grunnkretser og aggregeres til OD-matriser på ulike nivå. Her kan også detaljeringsnivået settes høyt, f.eks. kan reisemiddel, tid på døgnet og tidsbruk registreres. Den åpenbare fordelene med analyser av mobiltelefonidata synes å være volumet det tilbyr, i prinsippet hentes data inn for alle mobile enheter i et definert området, men på hvilket detaljeringsnivå og til hvilken kvalitet?

Utfordringen er å bygge et metodeapparat eller en «verktøykiste» for å jobbe med slike nye kilder, og ikke minst hvordan resultater fra slike kilder skal tolkes i forhold til den kunnskapen man allerede har. Er mobiltelefoni et supplement eller et blindspor?

Metodisk så har man jobbet tett med Ruter og Telia i dette prosjektet. Funn og hypoteser har blitt drøftet med begge parter som et ledd i å utvikle og utfordre metodikken og praksisfeltet. Det som oppsummeres i dette notatet bør leses som et øyeblikksbilde av en diskusjon som har pågått over noe tid, og som kommer til å pågå i lang tid fremover. Mobiltelefonidata er en ung datakilde hvor det fulle potensialet fremdeles er ukjent. Det ligger derfor flere drøftinger og diskusjoner forut for det som presenteres her.

Resultatene presentert i dette notatet bør leses som en samlet beskrivelse av datakilden teleoperatørene kan tilby i dag, mer enn som enkeltstående analyser som forsøker å konkludere rundt for eksempel antall reisende. I dette notatet fokuserer det på mulighetsrommet, kvaliteten, begrensninger og nytten slike ved slike data, eksemplifisert gjennom utforskende analyser.

Følgende hovedoppgaver ble definert i oppdragsbeskrivelsen:

1. Utarbeide og gjennomføre bestilling av nødvendig rådata fra teleoperatør
2. Vaske, kvalitetssikre og tilrettelegge data mottatt fra teleoperatør
3. Koble data fra teleoperatør med andre relevante kilder (eksempelvis data fra Fjellinjen, Ruter, SSB etc.)
4. Hele oppdraget skal dokumenteres i et prosjektnotat.

De konkrete analysene nevnt i oppdragsbeskrivelsen som man i dialog med Ruter ble enige om å gjennomføre var:

- Antall reiser som gjennomføres i Oslo og Akershus → Gjennomsnittlig antall reiser per person per dag
- Reisestrømmer i de tre korridorene (Vest, Nord-øst, Sør)
- Reiser med tog (vaske vekk bilreiser som går langs toget)

- Hvor mange reiser over bomringen
- Sammenligne mobildata med passasjertellinger fra Ruter
- Alle reiser inn til /ut av til Asker sentrum i morgenrush og ettermiddagsrush - Antall og hvor kommer de fra eller hvor skal de til?
- Sammenligning av MIS med mobildatatelefon
- Undersøkelse av reiser til og fra Kværnerbyen - Antall og hvor kommer de fra eller hvor skal de til?

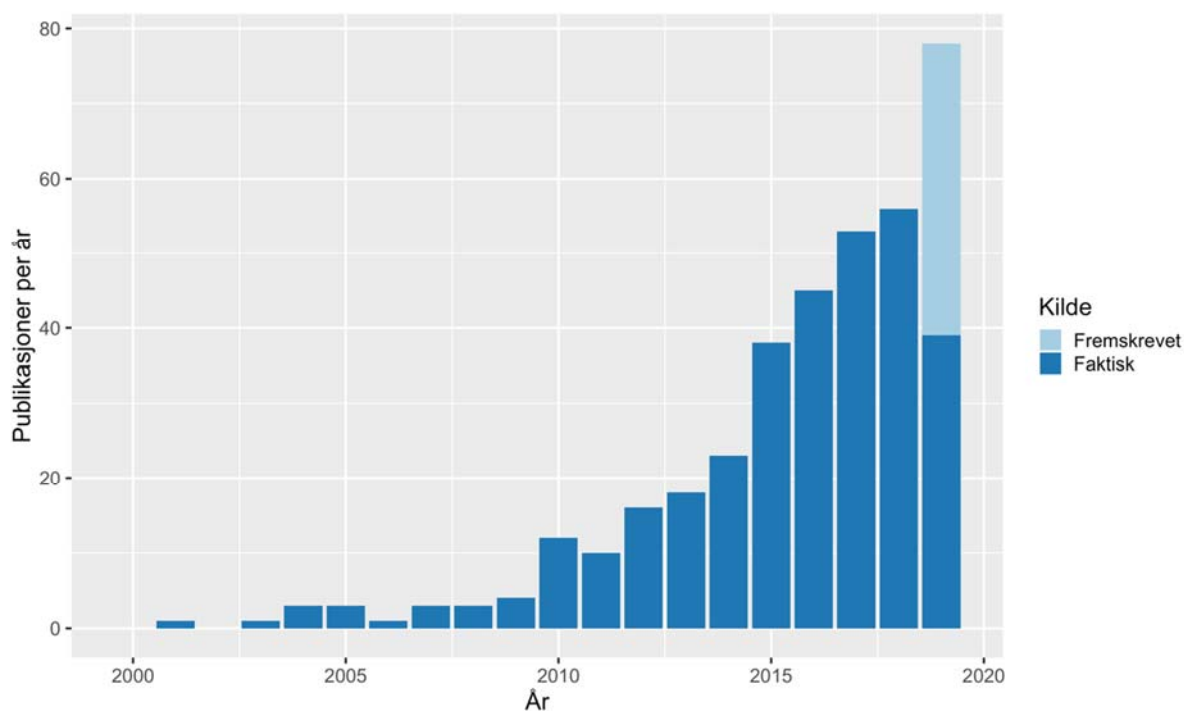
1.1 Notatets struktur

Notatet er så langt det lar seg gjøre bygget opp den kronologiske rekkefølgen som prosjektet ble gjennomført i. Dette er i seg selv et viktig poeng, ettersom selve forståelsen av hvordan man bestiller, hva man bestiller, og hvilke analyser og konklusjoner man gjennomføre både var en stor del av oppdragsbestillingen og i prosjektgjennomføringen.

Notatet begynner med en kort beskrivelse av forskningsfeltet rundt mobiltelefonidata i transportanalyser (Kapittel 2). Deretter beskrives prosessen frem til at man har et datasett med de avveiningene som ble gjort der (Kapittel 3). De etterfølgende kapitlene 4 og 5 inneholder en beskrivelse av kvalitetsvurdering av data og resultater fra de analyser som er gjennomført. Dette leder inn til en avsluttende drøfting og oppsummering av analysemuligheter og kvalitetsaspekter ved denne typen data i transportanalyse (Kapittel 6), samt noen avsluttende konklusjoner i kapittel 7.

2 Mobiltelefonidata som datakilde i transportanalyse

Det er generelt høye forventninger til mobiltelefonidata som kilde til mobilitetsanalyser. I transportanalysefeltet er dette en ny kilde som søker å avdekke *revealed preference* i en skala og kostnadsbilde som gjør det veldig interessant for mobilitets- og transport-analyser. Ved å søke i de store tidsskriftene i familien *Transportation Research* så ser man tydelig en utvikling, se Figur 2



Figur 2: Utvikling i antall papers som omhandler mobildata hos Transportation Research

Eksempler på titler fra 2019:

- “Inferring dynamic origin-destination flows by transport mode using mobile phone data” (Bachir et al, 2019)
- “Transport mode detection based on mobile phone network data: A systematic review” (Huang et al, 2019)
- “Mobile phone records to feed activity-based travel demand models: MATSim for studying a cordon toll policy in Barcelona” (Bassolas et al, 2019)
- “Fusing mobile phone data with other data sources to generate input OD matrices for transport models” (Montero et al, 2019)
- “Optimizing the deployment of electric vehicle charging stations using pervasive mobility data” (Vazifeh et al, 2019)
- “Multimodal divide: Reproduction of transport poverty in smart mobility trends” (Groth, 2019)

Dette er ikke en systematisk kartlegging, men ut ifra publikasjonsraten kan det se ut som et felt som har høy akademisk interesse. Ved å skue til titlene så kan det se ut som om det fortsatt forskes på hvordan man skal jobbe med å vri denne typen data om til noe som kan benyttes som OD-matriser.

Om allokering av mobiltelefonidata til reisemode (en av de mer sentrale temaene) så er (Huang et al, 2019) en grei systematisk oppsummering av feltet:

The rapid development in telecommunication networks is producing a huge amount of information regarding how people (with their mobile devices) move and behave over space and time. While GPS data, typically collected by smartphone apps, are restricted to rather small samples of the population, mobile phone network data, routinely collected by mobile network operators, potentially allow to analyze travel behaviors and social interaction of the whole population, with full temporal (e.g., longitudinal) coverage at a comparatively low cost.

Therefore, recent years have seen an increasing interest in using such data for human mobility studies. However, due to their noisy and temporally infrequent/irregular nature, extracting mobility information such as transport modes from these data is particularly challenging. This paper provides an in-depth, systematic review of transport mode detection based on mobile phone network data. **The results of the review show that existing studies tend to identify easy-to-detect modes (e.g., train or metro), or aggregate fine-grained modes into more general groups (e.g., public versus private transport).** Rule-based methods making use of geographic data were often developed. **More importantly, due to the lack of ground truth data, evaluation of the proposed methods was seldom done and reported.** Finally, we identify a list of research gaps currently being faced in this field, particularly regarding robust and real-time data cleaning and mode detection methods, "benchmark" datasets and metrics allowing the comparison of different mode detection methods, as well as privacy and bias issues.

(Sammendrag fra Huang et al, 2019), vår utheving.

Hvis man skal sortere modenheten til metodene inn i «Proof of concept», «Proof of Value», og «Proof of Use» altså i stigende grad av modenhet for en forretningshensikt – så er det i vår tolkning *få eller ingen resultater* som reelt sett er over noe særlig over nivået for «Proof of Concept».

I vårt begrensede litteratursøk mangler det stort sett alltid en kombinasjon av:

- Validering mot andre (pålitelige) kilder
- Omforente definisjoner, bla. på reise
- Demonstrasjon av (mer)verdi

Forskningen dreier seg stort sett fortsatt primært om metoder og ikke demonstrasjon av at man dekker forretningsmessige behov.

3 Bestilling

I oppdraget første del "Utarbeide og gjennomføre bestilling av nødvendig rådata fra teleoperatør" ble begge mobildataselskapene som tilbyr mobiltelefonidata kontaktet og individuelle møte ble gjennomført for å kartlegge bestillingsmulighetene for analysene definert i kapittel 1. I første ledd presenteres mobiloperatørens standardtjenester for ulike dataformat, for eksempel tellinger av håndsett for definerte områder, ODM matriser, eller "hvor kommer folk fra og hvor reiser de til"-analyser i forbindelse med større arrangementer. Mobiltelefonidata er i sin råeste form skapt igjennom identifikasjon av når og med hvilke basestasjoner hvert enkelt mobilabonnement har registret aktivitet. På dette nivået er mobildata virkelig stordata, med utfordringer knyttet til nøyaktighet, hyppighet på registreringer og personvern. Det som imidlertid gjøres tilgjengelig for kjøp er aggregerte data, på et annet detaljeringsnivå enn det som beskrives over.

Ved innhenting av tilbud for kjøp av mobildata legges stor vekt på personvern og hvilke begrensninger dette setter for uttrekk av data, se mer presis beskrivelse av kapittel 3.1, men det kan være utfordrende å få forståelse, eller hjelp til å forstå, hvilke konsekvenser dette vil ha for hvert enkelt datauttak. Generelt står man som kunde i denne fasen av bestillingsprosessen litt alene med å beskrive og forstå hvilke av standardtjenestene som vil gi rett innsikt i forhold til problemet. Det meste av denne avstanden mellom tilbyder og bestiller kan tilskrives den faglige avstanden mellom transportanalytikere og mobiltelefonidata-tilbydere, se kapittel 3.3 for beskrivelse av forskjellen mellom definisjon på en reise.

Når man har definert et ønsket datasett innfor de rammer som mobilselskapene kan tilby må data, med unntak av håndsett-tellinger, bestilles og samles for en dato fremover i tid. Mobilselskapene lagrer ikke mobildata i sin råeste form med mindre man har definert et tidsrom for å gjøre dette på forhånd.

Etter vurdering av de to mobilselskaperens tilbud endte man opp med å bestille data fra Telia, nøyere beskrevet i kapittel 3.2. I prinsippet var det to forhold som avgjorde dette:

1. Pris og beskrivelse av datasett
2. Tilbud om å jobbe tett sammen med en "data scientist" med tilgang til rådata for å bistå utviklingen med en algoritme for å identifisere og telle reisende med tog.

3.1 Områder, tidsvinduer og aggregering

Et av de viktigste elementene i bestillingen av data var utformingen av områdene. Selv med utgangspunktet at man ønsker seg så finkornede data som mulig, så er begrensningen at man må være innenfor lovlig og akseptert nivå av personvern. Prinsippet er at det ikke skal være mulig å reidentifisere enkeltpersoner basert på deres aktivitetsmønster. Dette løses ved to teknikker:

1. Mellom to områder og innenfor et gitt tidsvindu må man ha *mer enn* en kritisk¹ mengde personer som har *samme aktivitetsmønster*. Dette er det som refereres til som «k-verdi» eller «k-anonymitet». For Telia så er k-verdien satt til 5 personer ($k=5$), *utført etter skalering til befolkningstall*.
2. Alle mobiltelefoner får en ny anonymisert identifikasjonsnøkkel knyttet til seg ved midnatt. Dette resulterer i at det i prinsippet ikke er mulig å spore en eller flere reisende på tvers av to døgn.

Det vil si at dersom tre Telia-kunder har reist fra Hurum til Eidsvoll mellom klokken 08:00-09:00, og Telia har en markedsandel på $\approx 50\%$ så vil det tilsi at man får 6 reisende mellom Hurum og Eidsvoll. Hadde det derimot vært to Telia-kunder, så vil tallet bli sensurert (eller for alle formål satt til «null») selv om det kanskje er fire reisende til med Telenor-telefoner.

Ved å øke områdestørrelsen så øker man sjansen for å få flere reisende med like aktivitetsmønstre. Det er for eksempel flere som reiser mellom bydeler i Oslo, enn det er mellom de individuelle grunnkretsene. Tilsvarende så vil det å øke tidsvindueene gjøre det mer sannsynlig at flere har like aktivitetsmønstre. For eksempel er det større sannsynlighet for at mer enn 5 mennesker reiser mellom to bydeler i Oslo innenfor samme time, enn innenfor hvert kvarter – spesielt på kveld og nattestid.

Selv om prinsippene over nødvendigvis må etterleves, kan det ha noen uheldige konsekvenser for bestiller av data. Bestilling og analyse av data kan naturligvis åpne flere spørsmål og nye behov: For eksempel kan et interessant resultat fra par i en OD-matrise mellom større områder gi et behov for å undersøke reisemønstret for alle grunnkretser i det ene området til alle grunnkretser i det andre. Det vil med dagens system for bestilling av data da være behov for en formel bestilling av et nytt datasett, men tilhørende pristilbud og leveringstid. Dette kan oppleves som et noe rigid og kostbart system for den utforskende transportdataanalytiker.

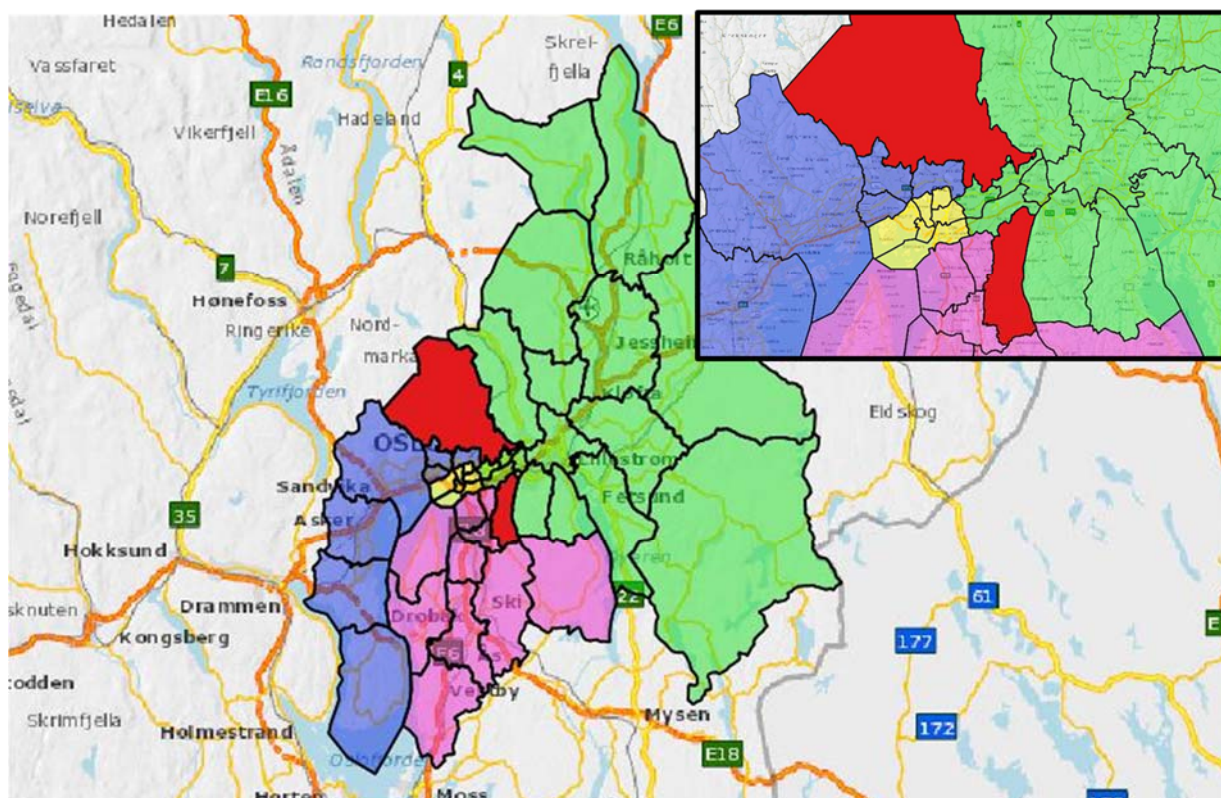
3.2 Beskrivelse av data

Telia har levert tre datasett:

1. OD-matriser for reiser til/fra de 40 definerte områdene bydel/kommunene
2. Håndsett-tellinger for de samme dagene innenfor de samme områdene (minus Oslo Sentrum, som ved en feil ble utelatt fra første bestilling)
3. OD-matriser for reisende til/fra grunnkretsene i Akser Sentrum til/fra alle andre grunnkretser som definerer opp de øvrige delene av de 4 markedsområdene til Ruter.

Alle datasettene er aggregert opp til én times-oppløsning, det vil si tellinger av antall reiser eller håndsett gjøres per time, og opp til de definerte områdene i geografi. Algoritmene som Telia har kjørt har tilgang på mer detaljerte data, slik som sporinger av enkelthåndsett med kontinuerlig tidsregistreringer av aktivitet på mobiltelefoner gjennom tilknytning til mobilceller (området dekket av en eller flere basestasjoner) som romlig oppløsning. Mobilceller har ulik størrelse avhengig av teknologi, type bebyggelse, antall mennesker som trenger mobiltjenester i området etc. I tillegg er begge datasettene skalert opp fra Telias markedsandel til befolkningstall. Denne metoden har vi ikke full innsikt i, men vi har undersøkt kvaliteten nærmere i vårt valideringsarbeid. Det ble bestilt data for de fire onsdagene i februar 2019, der den ene av disse (20. februar) var inneholdt i vinterferieuken.

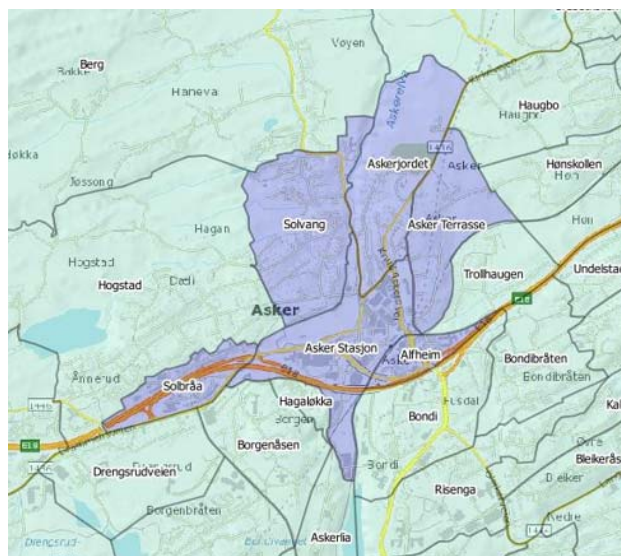
I Figur 3 vises de 40 områdene som danner grunnlaget for alle de tre bestillingene fra Telia.



Figur 3: Ruters fire definerte markedsområdet: Grønn (Nordøst), blå (Vest), rosa (Sør), gul (Indre by), samt oppdeling i de 40 bydel/kommunene. Røde områder er definert ut (Marka). Innfelt vises indre by i mer detalj.

For 39 av områdene ble det levert håndsett-tellinger, dvs. antall personer som har oppholdt seg en viss minimumstid (dwelling) innenfor hvert område. Disse tellingene ble levert per time. Dette er en parameter Telia lar kunden bestemme selv. I tillegg estimeres det i dette datasettet hvor mange av de opptelte håndsettene som oppholder seg i hvert av områdene som til enhver tid er der «på jobb» (altså ikke er hjemmehørende der ut ifra hvor de befinner seg på natten).

I tillegg har vi fått oversendt OD-data, også på timesnivå for de 40 områdene, det vil si en ODM matrise per time for alle de fire onsdagene. For det siste datasettet defineres «Asker sentrum» ved de seks grunnkretsene "Askerjordet", "Asker Terrasse", "Solvang", "Asker Stasjon", "Solbråa" og "Alfheim", se Figur 4.

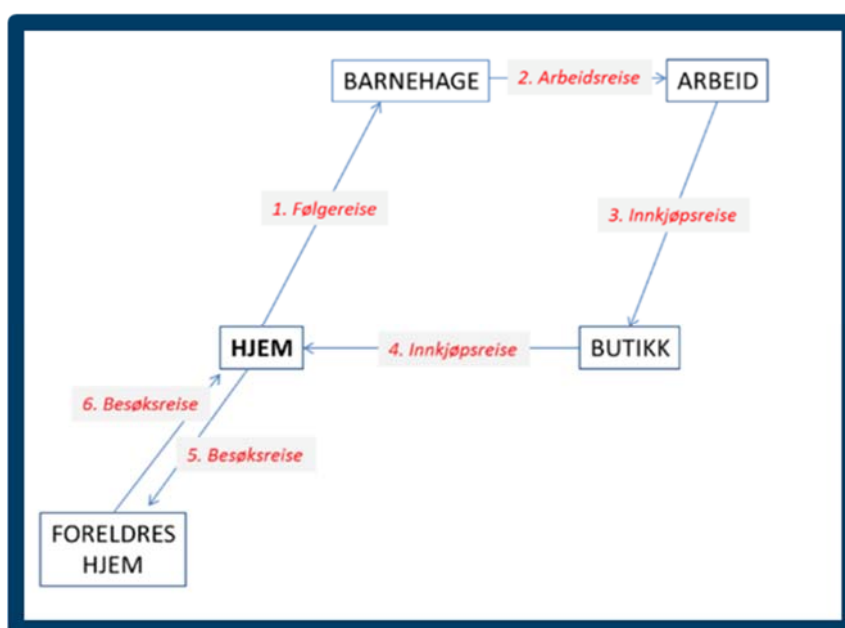


Figur 4: Grunnkretser definert i dette prosjektet som Asker Sentrum

Dette datasettet inneholder tellinger av antall reisende til og fra Asker Sentrum fra og til alle andre grunnkretser som utgjør de resterende områdene av de 40 hovedområdene. Dette utgjør dermed en høyere romlig detaljeringsnivå enn de to andre datasettene, men med tilhørende større risiko for sensurering, se kapittel 5.5 for detaljer.

3.3 Definisjon av reiser

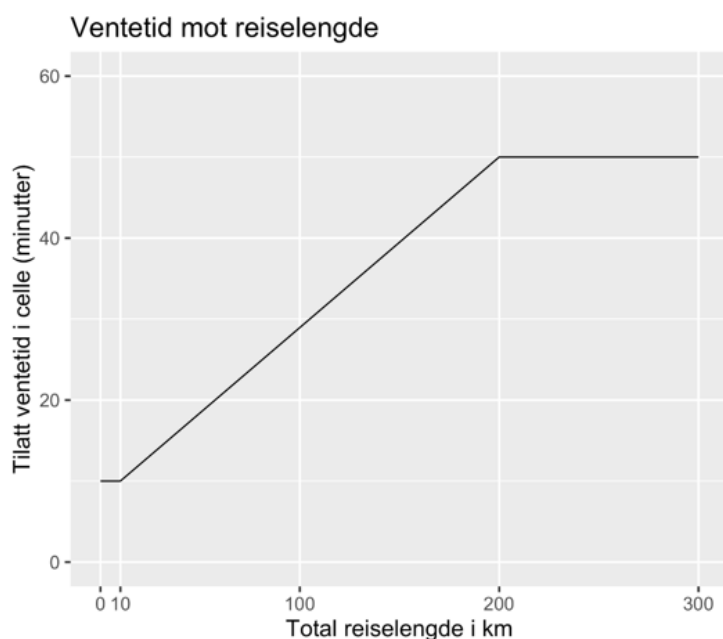
For å forstå mobildata på det nivået som tilbys fra mobiloperatørene er det viktig å ha klart for seg hva som defineres som en reise i denne metodikken og hvordan den skiller fra den som brukes i tradisjonelle reisevaneundersøkelser og i transportanalysefaget generelt. I Figur 5 vises en transportanalytikers definisjon, der: "En reise avgrenses og defineres ut fra formålet på bestemmelsesstedet. Når man har kommet fram til stedet for formålet med reisen, avsluttes reisen."



Figur 5: Definisjon av reiser fra Hjorthol et al. (2014)

Denne definisjonen av en reise er ikke mulig for mobildata å kopiere av den enkle grunn at man generelt sett ikke får noe informasjon om formålet ved noen forflytninger registrert ved en mobiltelefon eller annen enhet med SIM-kort. Forflytning av en mobiltelefon registreres av mobiloperatørene ved å observere et eller flere skifter mellom oppkoblingen fra en mobilcelle med tilhørende basestasjon(er) til en annen. Størrelsen på mobilceller variere fra noen få hundre meter til flere km, avhengig av teknologien som benyttes og etterspørselen etter mobiltjenester. Typisk vil mobilcellene stå tettere i tettbebygde strøk enn mobilcellene som dekker området enn utenfor. Mobildata benytter i hovedsak to prinsipper for å definere en reise:

1. Dersom en mobiltelefon har forflyttet seg mellom to eller flere mobilceller avsluttes reisen når enheten har hatt oppholdstid (dwelling) større enn en gitt terskelverdi i en og samme mobilcelle. Terskelverdien kan selv være en funksjon av lengde på reisen, og funksjonen benyttet hos Telia i dette prosjektet er vist i Figur 6.

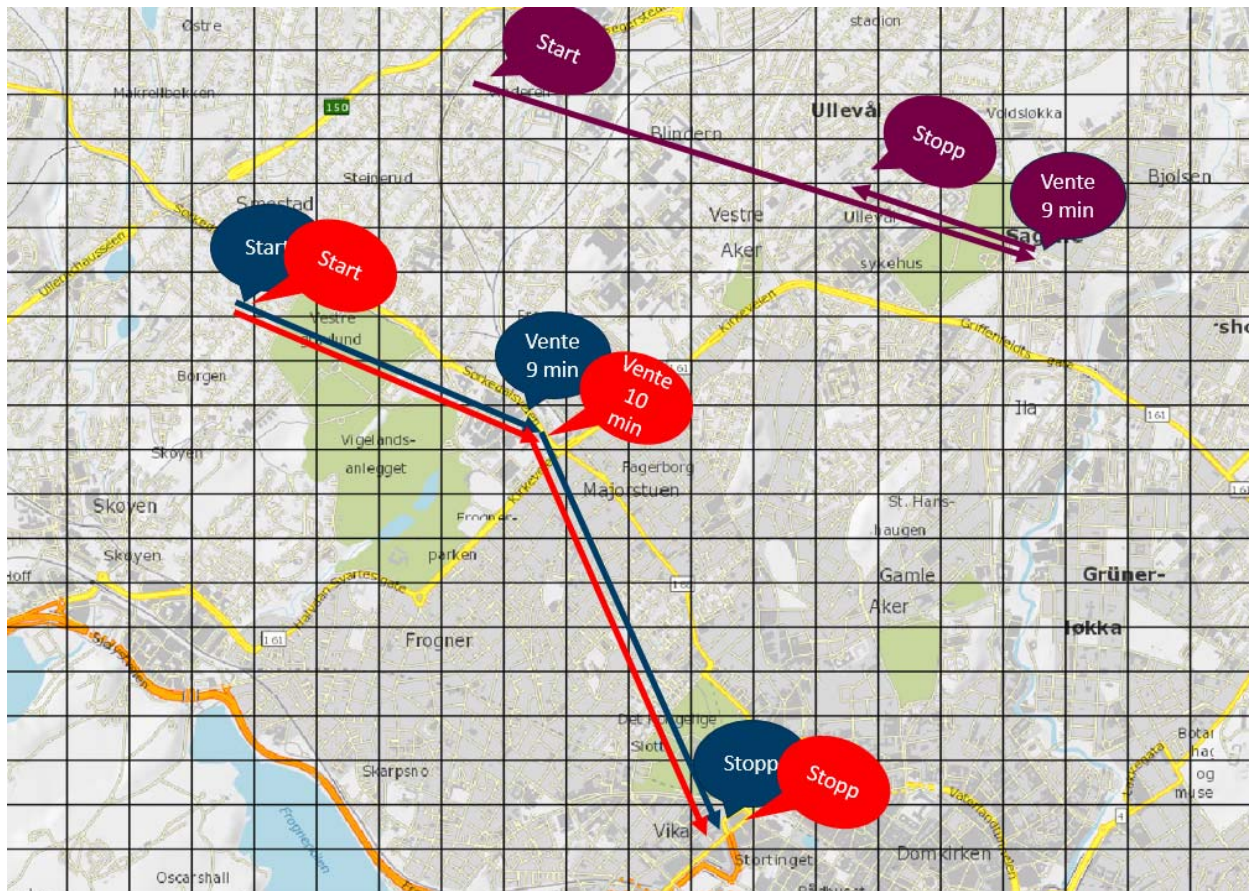


Figur 6: Oppholdstid i en mobilcelle før en reise defineres som avsluttet

Telia tilbyr å endre på parameterne som definerer denne funksjonen, men det vil kreve mer jobb å endre selve funksjonsdefinisjonen. Av denne figuren kan man lese at for reise mindre enn 10 km aksepteres en ventetid opp til 10 minutter i en mobilcelle før reisen avsluttes. For reiser over 200 km er ventetiden som aksepteres 50 minutter, mens mellom 10 min og 50 min er dette antatt å være et lineært avhengig forhold.

2. Dersom en mobiltelefon har beveget seg i en retning for så å snu 180 grader, det vil si begynne å bevege seg i motsatt retning, vil reisen regnes som avsluttet.

I Figur 7 illustreres hvordan reise defineres i mobildatagrunnlaget. F.eks. ser vi hvordan to geografisk identiske forflytninger kan resultere i en eller to turer basert på om ventetiden (f.eks. på buss- eller t-baneholdplass, i kiosk eller i barnehage) er mer eller mindre enn 10 min.



Figur 7: Illustrasjonseksempel som viser hvordan reiser med mobildata telles ved å registrere bytter mellom mobilceller (grid). Rød forflytning inneholder en stopp på 10 minutter som resulterer i to reiser, blå forflytning telles som en reise fordi tilsvarende stopp er på 9 minutter, og lilla forflytning telles som to reiser fordi man registrerer en endring i retning på 180 grader.

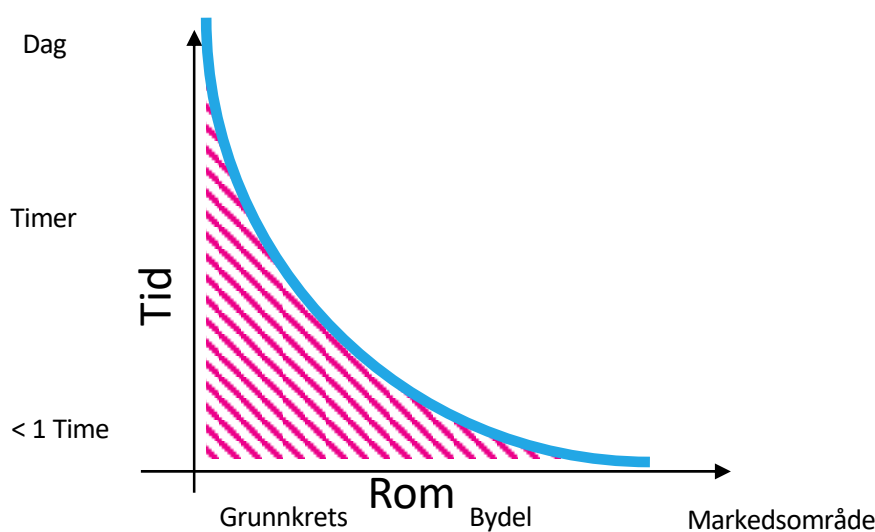
4 Validering

Det finnes ingen etablert/standardisert metode for å vurdere kvaliteten på denne type data, men vi vil i dette kapitlet forsøke å undersøke konsistens og validitet på ulike måter i de to datasettene som er benyttet i denne analysen. Dette er på ingen måte en uttømmende analyse av datakvalitet eller validitet for bruk av mobiltelefonidata generelt i transportanalyser, men gir en mulighet til å diskutere en ramme for hvor metode og data synes valide («face validity») og i hvilke typer analyser eller områder man bør trå mer varsomt.

4.1 Sensurering av data

Uttrekkene har alle vært gjennom en anonymiseringsprosess (k-anonymitet) som skal garantere at man ikke kan identifisere enkeltpersoner ved at man sørger for at det er k antall like reiser. Telia har operert med en $k=5$ (etter skalering til befolkning). Det vil si at alle uttrekk har sensurert bort alle relasjoner hvor det har vært 1-4 reiser i tidsperioden. Utfordringen med k-anonymitet og derfor også å anslå sensureringsgraden fra datasettet, er at man ikke robust kan skille mellom de som er under k-verdi (antall reiser = 1-4) og hvor det faktisk ikke har skjedd noen reiser i tidsvinduet (antall reiser = 0, som også er sannsynlig for mange av relasjonene i utvalget).

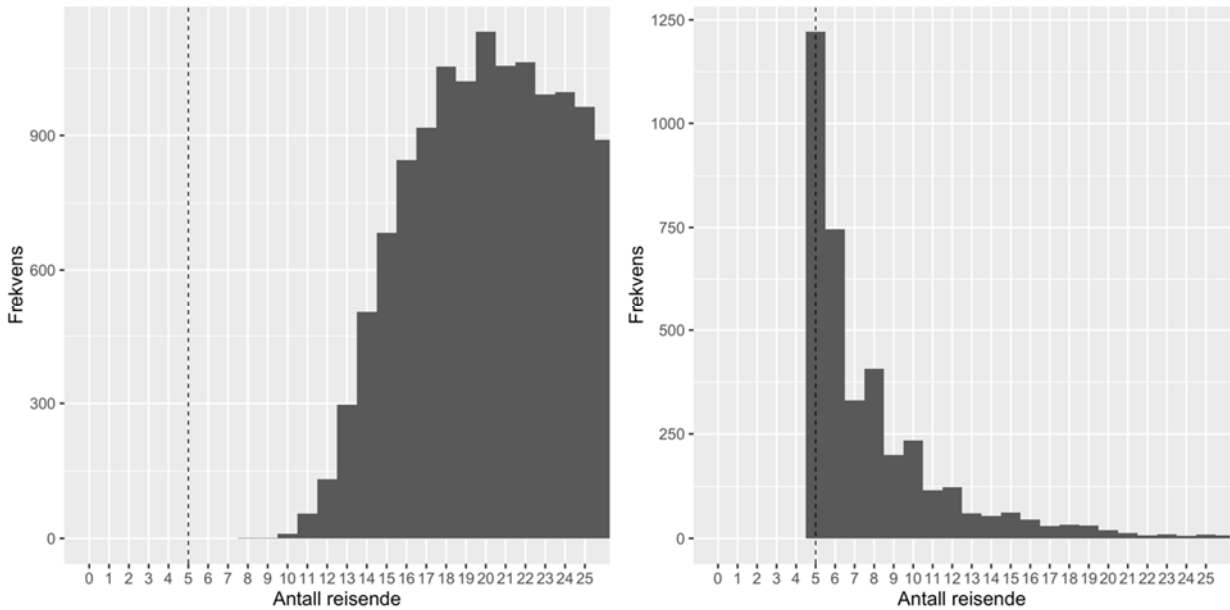
Så i definisjonen av uttrekk når man gjør en bestilling så gjør man egentlig to kritiske valg: *tidsoppløsning* og *romlig oppløsning*. Dette blir en avveining mellom tid og rom (hvor rom *egentlig* representerer befolkningen mer enn areal), slik at man ender opp med en sammenheng slik som vist i Figur 8.



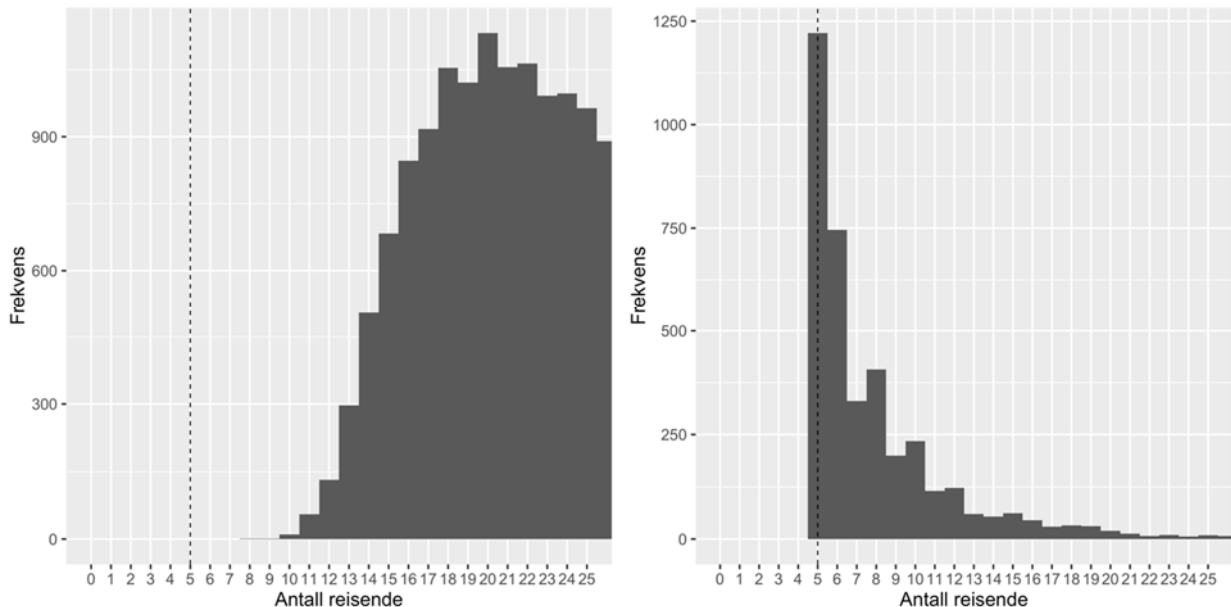
Figur 8: Prinsipiell skisse mellom tid og rom med risiko for sensurering

De skraverte områdene i figuren representerer «høy» risiko for sensur. De to datasettene vist i Figur 9, representerer samme tidsoppløsning, men med ulik romlig oppløsning, slik at datasettet til høyre i figuren ramler inn i det skraverte området, mens det andre ikke gjør det.

Hvis man plotter antall reiser i de relasjonene vi har data, vil man i et kraftig sensurert datasett se en skarp kant ved antall reiser = 5.



Figur 9 viser hvordan dette ser ut for OD-matrise datasettet på bydel/kommune-nivå og OD-matrise datasettet grunnkrets-nivå. På bydel/kommune-nivå (venstre i figuren) så ser vi ikke en skarp kant ned mot 5, men heller en opptrapping fra omtrent 11 reiser (med noe data i 8-10 reiser). Dette peker mot at datasettet på bydel/kommune ikke er kraftig påvirket av k-anonymitetssensureringen. På grunnkrets-nivå (høyre i figuren) så ser vi derimot helt klart at fordelingen helt klart er kappet ved reiser = 5, og at hvis man antar en omtrentlig lik venstre flanke som i figuren til venstre, så er det reisemønstre som er sensurert.

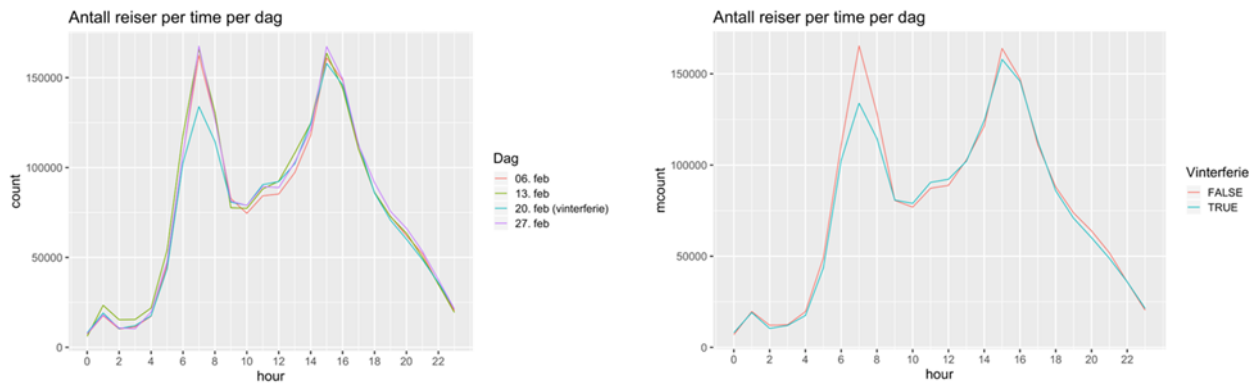


Figur 9: Utdrag av plott av antall reiser i hver relasjon for bydel/kommune-oppløsning (venstre) og grunnkretsoppløsning (høyre). Den vertikale akse viser antall OD-kombinasjoner som resulterte i de antall turene som er oppgitt på den horisontale akse. OD-kombinasjoner med mindre en 5 turer filtreres vekk, det vil si settes til 0.

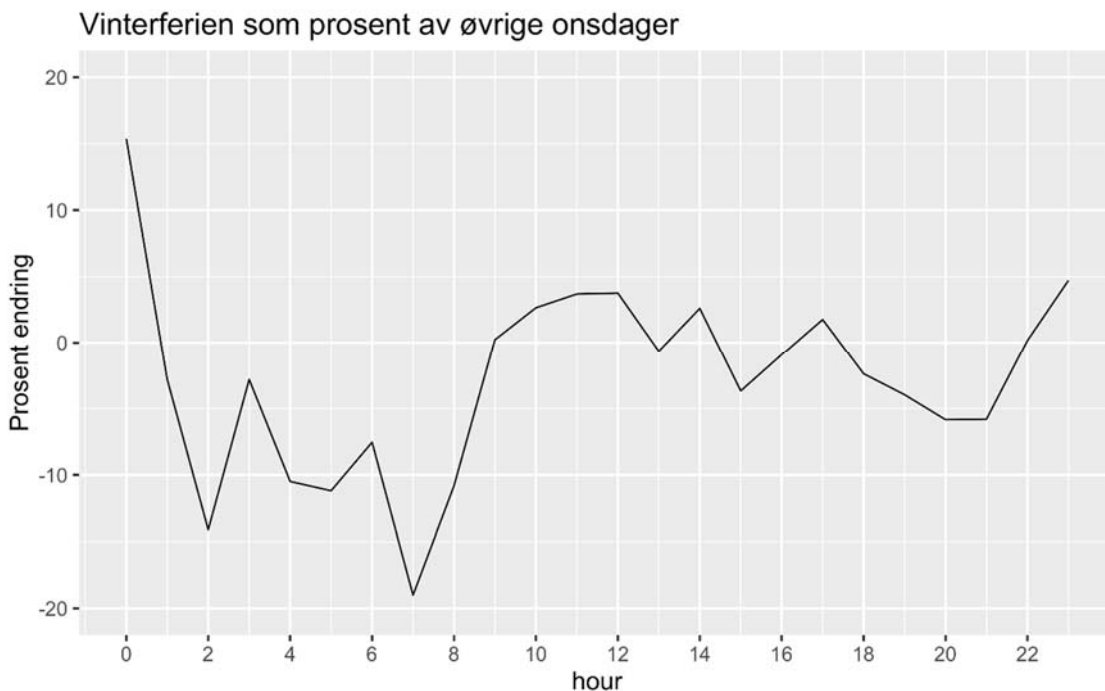
Dette tyder på at grunnkrets (med det forbeholdet om at grunnkretser her er i Asker kommune, noen med veldig lav befolkning) er en for liten romlig oppløsning når man arbeider med times-oppløsning i tid. Origin-Destination Matrisene

4.2 Stabilitet

Det ble bestilt data fra fire onsdager i februar 2019. Dette ble gjort ut ifra en antagelse om at onsdag er en «gjennomsnittlig» arbeidsdag, men februar har en uke vinterferie. Vi valgte med vilje å inkludere vinterferien også for å få med en eventuell forskjell på de onsdagene. I Figur 10 - Figur 11, så viser vi forskjellen på de ulike dagene og isolert sett også opp mot vinterferien ved å telle antall reiser i OD-datasettet på bydel/kommune nivå.



Figur 10: Venstre: Reiser per time per dag i februar. Høyre: Gjennomsnittet for de dagene som ikke er vinterferien opp mot onsdagen i vinterferieuken.

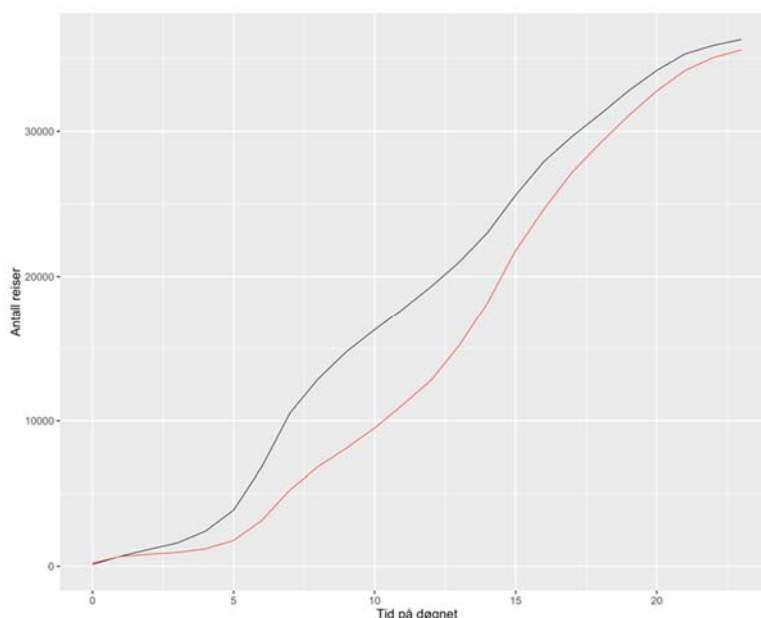


Figur 11: Effekten av vinterferie som prosent av reiser mot de øvrige ukene

Som vist er det lav variasjon over de tre «vanlige» onsdagene, mens vinterferien viser noe lavere (opp imot 20% reduksjon) i antall bevegelser i morgenrush, noe høyere reiseaktivitet på dagtid og så er reduksjon i ettermiddag/kveld. Det kan tenkes at onsdag også for mange er den siste «ordinære» arbeidsdagen i vinterferien før man tar en lengere helg på grunn av skolefri, men veldig stor endring i reisemønster er det ikke. I resten av notatet har vi benyttet data fra samtlige dager uten spesielt hensyn til at én av de er skolens vinterferie.

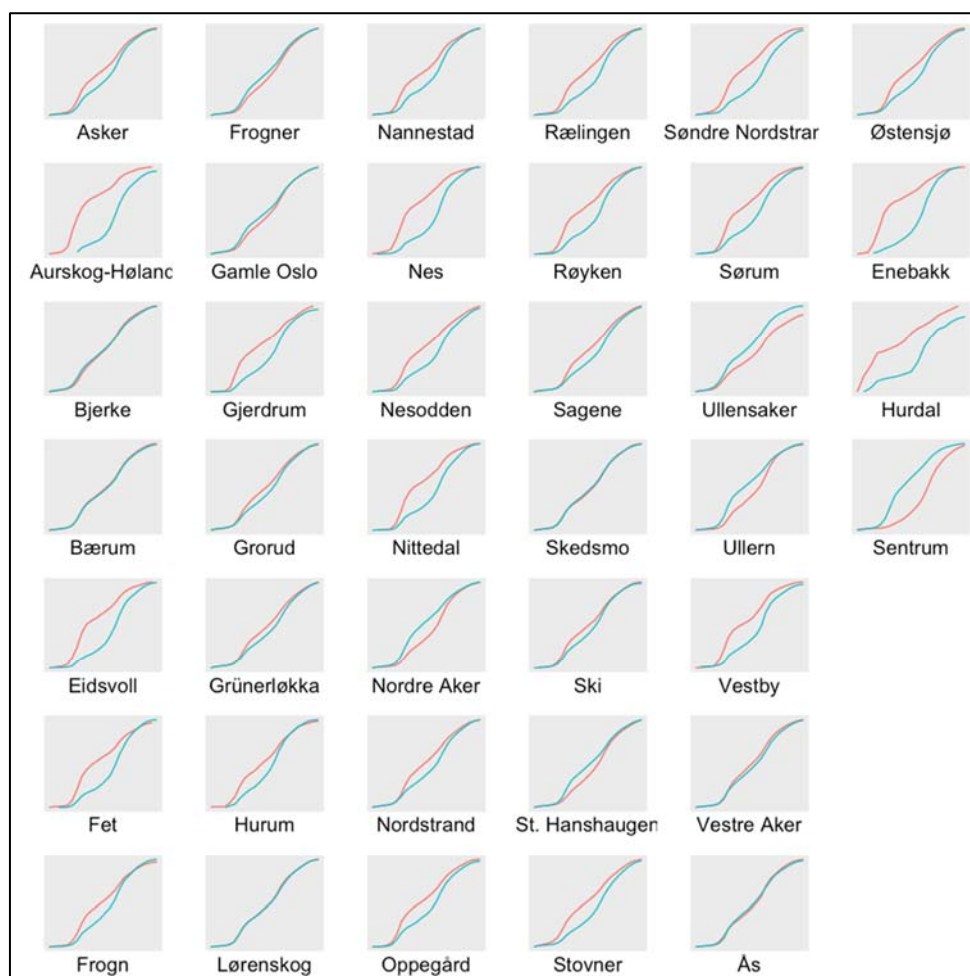
4.3 Likevekt

Som en sjekk på Telias ODM-algoritmes indre validitet, så har vi lagt til grunn en antagelse om likevekt for datasettet, altså i løpet av et døgn så skal samlet netto endring i befolkningstall være omtrent null. Det vil si at det bør være omtrent like mange reiser til et område som fra et område. Dette kan visuelt fremstilles som i Figur 12, hvor reiser ut er den sorte linjen, og reiser inn er den røde. Her ser vi at det er noe mer reiser ut på (arbeidsreiser *ut* av området), men at de returnerer og ved kvelden så er det omtrent like mange reiser inn og ut.



Figur 12: Illustrasjon over reiser til fra et område (i likevekt, sort er reiser ut, rød er reiser inn).

I håndsett-datasettet for bydel/kommuner, så ser det generelt ut som om dette stemmer relativt godt. Figur 13 viser slike plott for de 39 bydel/kommunene i dette datasettet.

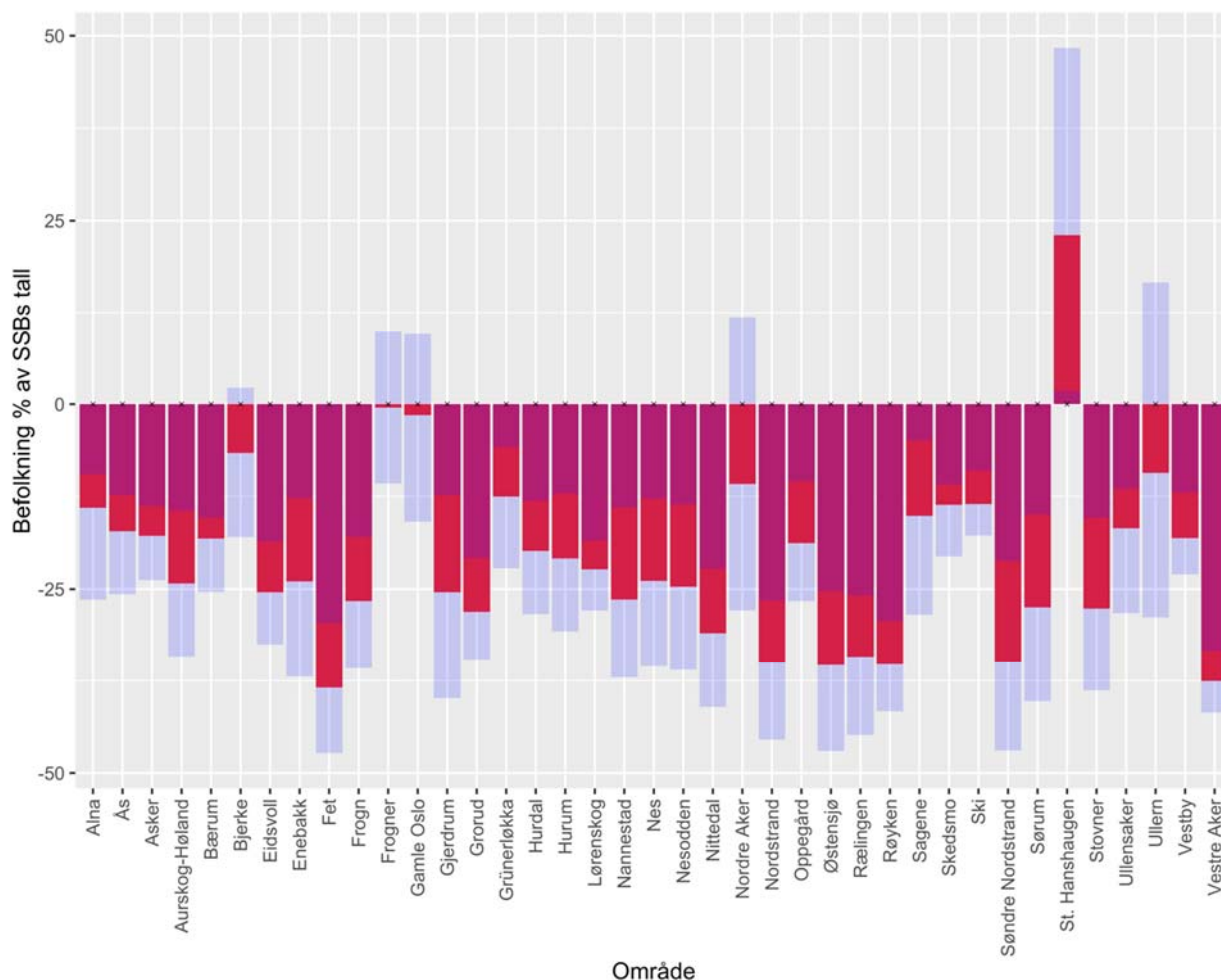


Figur 13: Likevekts-illustrasjon for 39 bydel/kommunene i dette datasettet.

4.4 Sammenligning mot befolkningstall

Her er det tatt utgangspunkt i SSBs befolkningstall per 1.1.2019, på kommune og bydelsnivå. Disse er så sammenholdt med befolkningstallet ut fra håndsett-registreringen fra Telia (datasett 2). Med unntak av 1 bydel (St. Hanshaugen) så gir dette et systematisk avvik hvor SSBs befolkningstall er høyere enn hva aktivitetsregistreringen tilsier. Dette er oppsummert i Figur 14, hvor SSBs befolkningstall er 0, og %-verdier for maks/min (lys blå) og gjennomsnitt (rød) antall «personer» vises på y-aksen.

For gjennomsnittstallet (rød søyle) så varierer avviket fra -38% for Fet (altså tilsier aktivitetsregistreringen omtrent 62% av SSBs befolkningstall), til +23% for St. Hanshaugen.

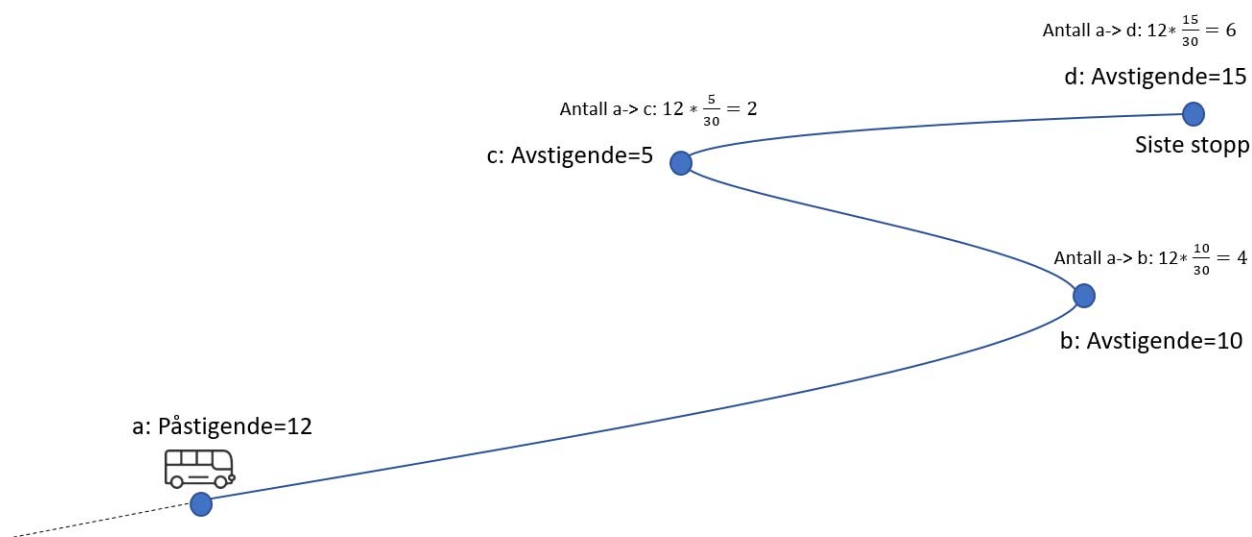


Figur 14: Håndsett-registrering (datasett 2) og SSBs befolkningstall

Det er her verdt å merke seg at det er mange feilkilder som bidrar. SSBs definisjon er personers bostedsregistrering, og ikke døgnhvile – som for eksempel tilsier at «ugifte studenter ifølgje reglane skal stå som busette i heimen til foreldra». SSB oppgir selv at undersøkelser har vist at 5,5% av befolkningen blir plassert på feil bosted (Schjalm, 1996).

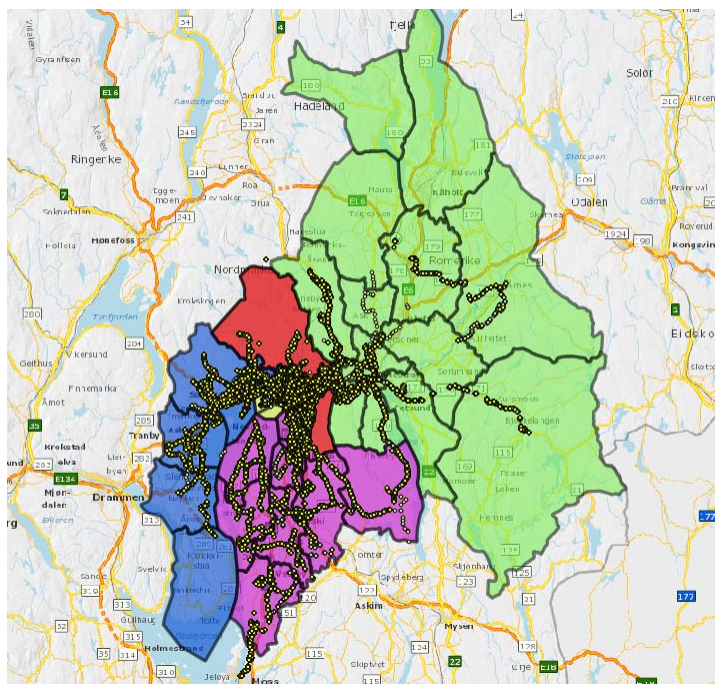
4.5 Sammenligning mot passasjertellinger

Ruter foretar tellinger av passasjerer på sine kollektivruter. Det gjøres ved hjelp av sensorer i dørene som teller antall påstigende og avgående passasjerer. Ved å ta differensen mellom disse kan man grovt sett vite hvor mange som er på en buss eller trikk. I dette prosjektet ønsket man å sammenligne passasjertellingene fra Ruter med mobildata. Ettersom passasjertellingene ikke gir noen direkte informasjon om hvilke fra/til reiser folk foretar må dette estimeres. Det finnes en naiv metodikk for dette som baserer seg på å fordele antall påstigende på en holdeplass som avstigende utover fremtidige holdeplasser vektet med de faktiske avstigningstallene. Altså at andelen avstigende oppstrøms benyttes som vektning for å estimere hvor de som gikk på går av. Denne metodikken er illustrert i Figur 15



Figur 15: Illustrasjon som viser hvordan man kan naivt estimere antall fra til reiser basert på passasjertellinger (av- og påstigninger). På punkt a kommer 12 passasjerer på. Disse må gå av senest på siste stopp (punkt d). Avstigende oppstrøms punkt a er 10 passasjerer (punkt b), 5 passasjerer (punkt c) og 15 passasjerer (punkt d). Vi antar de 12 påstigende passasjerene fordeler seg som avstigende basert på disse andelene, det vil si 4, 2 og 6 personer av de 12 påstigende går av på henholdsvis punkt b, c og d.

Data ble mottatt fra Ruters passasjertellingssystem for de samme fire onsdagene i februar som for mobildataene. Ved hjelp av metodikken beskrevet overfor kan passasjertellingene oversettes til fra/til reiser som alle starter eller slutter ved holdeplasser langs Ruters rutetilbud. Figur 16 viser alle punktene i det oversendte datasettet fra Ruter hvor passasjertellinger er tilgjengelig fra de fire onsdagene i februar. Som vi kan se er det ikke alle soner som er like godt dekket med hensyn til slike tellinger, og hvor usikkerheten ved sammenligning med mobildata øker tilsvarende.



Figur 16: Alle registrerte holdeplasseringer (gule punkter) for de fire onsdagene i februar 2019, der det er registrert passasjertellinger.

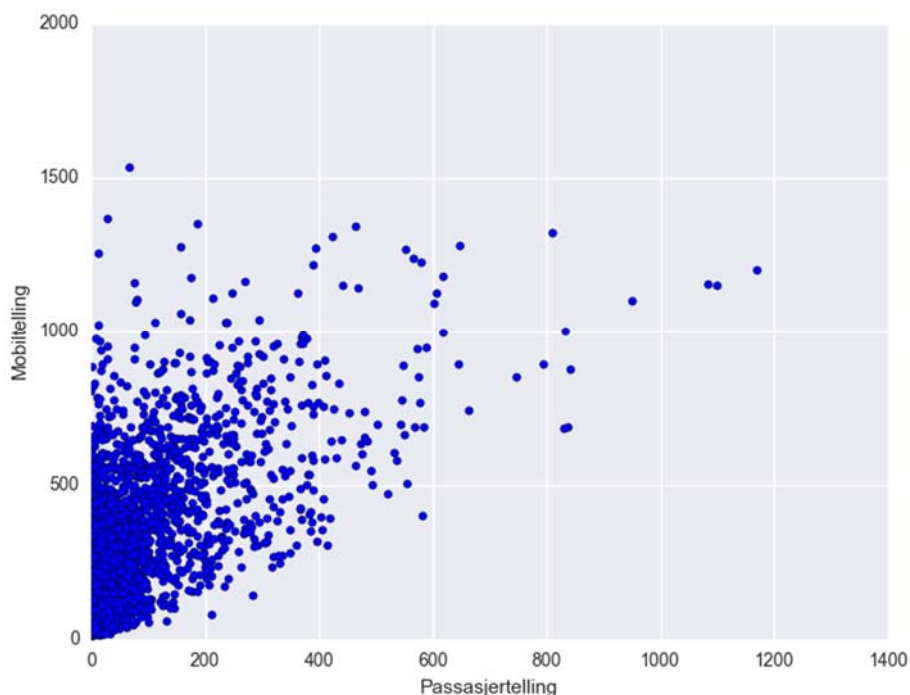
Deretter er hver holdeplass koblet geografisk til bydel/kommunene som er brukt i dette prosjektet. Totalt er det 4 249 511 linjer i tekstfilen med passasjertellinger. Hver av disse tekstlinjene blir behandlet med følgende algoritme:

1. Gjør om bydel/kommune til polygon med gitt geografisk utstrekning

For hver tekstlinje i filen:

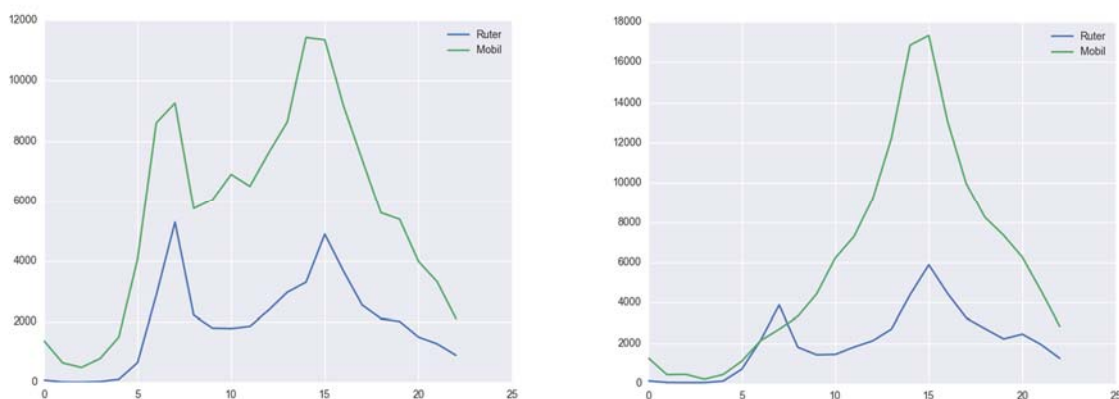
2. Bestem geografisk plassering til startholdeplass
3. Finn det polygonet som inneholder punktet med startholdeplass
4. Skriv id på bydel/kommune til startholdeplass på slutten av hver tekstlinje.
5. Gjenta steg 2 til 4 for sluttholdeplass

Basert på tekstfilen blir det opprettet en OD-matrise basert på passasjertellinger over de samme 40-bydel/kommunene i dette prosjektet (Figur 3). Det tas da utgangspunkt i en bydel/kommune om gangen, og summert opp alle reisende til alle andre bydel/kommuner hver for seg for hver time. Resultatet er en matrise som inneholder antall reisende mellom bydel/kommunene, slik at det er mulig å sammenligne med data fra Telia, det vil si OD-matrise datasettet fra bydel/kommune. En oppsummerende sammenligning av alle relasjoner mellom bydel/kommuner er vist i Figur 17. Her ser vi at de aller fleste observasjoner ligger over diagonalen, som viser at det er observert flere reisende med mobiltellinger enn med passasjertellinger. Dette er som forventet på grunn av at mobiltellinger inkluderer andre reisemodi enn bare kollektivtransport, som er det man får fra passasjertellinger.



Figur 17: Sammenligning av antall reisende mellom bydel/kommuner. Hvert punkt er en relasjon mellom to bydel/kommuner, med antall reisende fra henholdsvis mobiltelling på y-aksen og passasjertelling på x-aksen.

Noen konkrete eksempler på sammenligninger for relasjoner mellom bydel/kommuner er vist i Figur 18, hvor antall reisende med enten Frogner eller Sentrum som destinasjon er presentert, time for time.



Figur 18: Venstre: Frogner. Høyre: Sentrum. Antall reisende med bydel/kommune som destinasjon, per time.

Generelt sett ser det ut som god korrelasjon mellom mobildata-reisemønster og det vi kan estimere ut fra passasjertellinger. Det er også fornuftig at mobildata registrere flere reiser enn Ruter, se kommentar over, noe som også er synlig i denne figuren. For Sentrum har vi allikevel et tilfelle rundt kl. 7 på morgenen det syns å være flere reisende med buss en reisende totalt. Det er vanskelig å peke på nøyaktig hvorfor slike ulogiske observasjoner oppstår, men det skyldes trolig både usikkerhet i mobildata og i

estimeringsmetodikken for å konvertere passasjertellinger til OD-matrise. Spesielt tar ikke den siste metodikken hensyn til at svært korte eller lange reiser med kollektiv transport kan være mindre sannsynlig enn mellomstore reiser.

5 Analyser

5.1 Antall reiser som gjennomføres i Oslo og Akershus

Datagrunnlaget gir i utgangspunktet godt grunnlag for å beregne antall reiser per dag, men med *to vesentlige begrensninger*:

- Definisjonen av reise er «uvariant» i transportanalyser
- Reiser innad i en og samme området (f.eks. bydel/kommune) telles ikke, ei heller reiser utenfor Oslo og Akershus.

Når vi summerer gjennomsnittlig antall reiser per time: så blir det totalt: 1 890 972 reiser på i en matrise for en «gjennomsnittlig» onsdag basert på OD-matrise datasettet for bydel/kommune. SSBs befolkningstall sier 1 334 883 (eventuelt kan man benytte 1 083 578 som er gjennomsnittlig summen av tellingen av mobiltelefoner fra datasettet for håndsett-registrering).

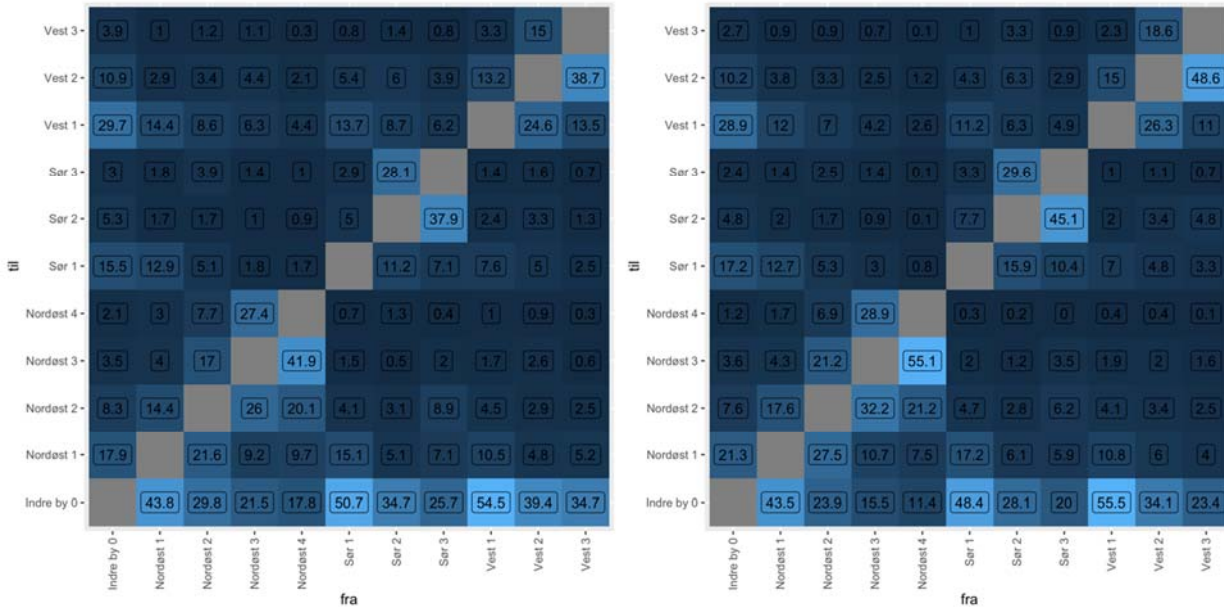
Dette gir 1,4 til 1,7 reiser mellom områder per person per dag.

Til sammenligning så sier Den nasjonale reisevaneundersøkelsen 2013/2014 (Hjorthol et al, TØI 1383/2014) 3,3 reiser per person. Gitt begrensningene (inkl. ulik definisjon) så er det unaturlig å sammenligne disse tallene direkte.

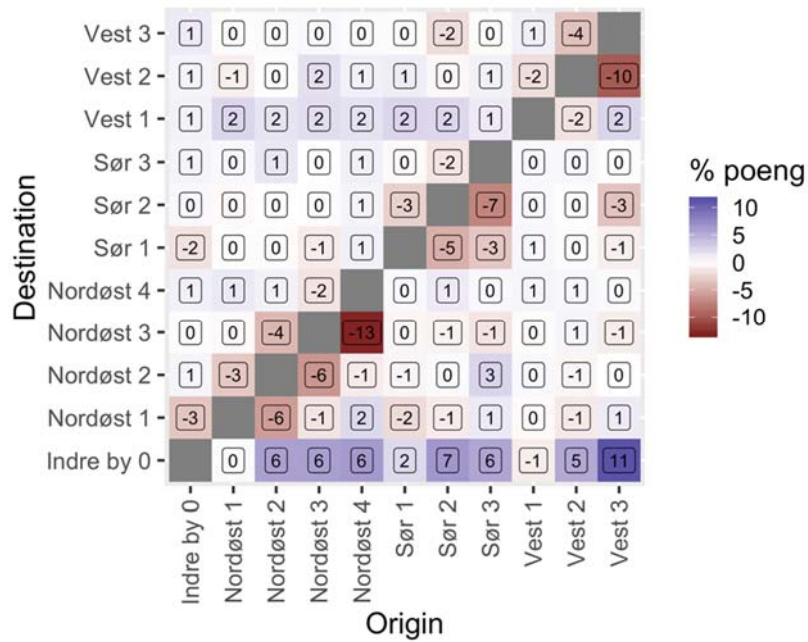
5.2 Sammenligning reisevaneundersøkelse/MIS og mobildata

For å få et inntrykk av hvor godt mobildata ODM representerer den samme virkeligheten som målt i Ruters egen reisevaneundersøkelse har vi trukket ut data fra MIS på reiser fra og til de ulike delmarkedsområdene (skalert) og sammenstilt de samme områdene fra mobildata, der OD-datasettet for bydel/kommuner er lagt til grunn. For å få sammenlignbare tall så har vi fjernet reiser innad i områdene – og ignorert delmarkedsområde Vest 4 (som ikke var med i RVU).

Data er beregnet i prosenter av radene i ODM-matrisen. Figur 19 viser de to reisematrixene i prosent, med relativt sett godt samsvar mellom fordelingene. Figur 20 oppsummerer disse i en differansematrise, som viser differansen mellom de to matrixene. Differansene i matrixen ligger i omfanget [-13, +11] %-poeng. Det kan tolkes dithen at mobildata og reisevaneundersøkelsene viser en god grad av samsvar, med 50% av relasjonene imellom [-1, 1] %-poeng.



Figur 19: Reiser mellom delmarkedsområdene innen prosent av fra. Ruters MIS til venstre, Mobildata til høyre



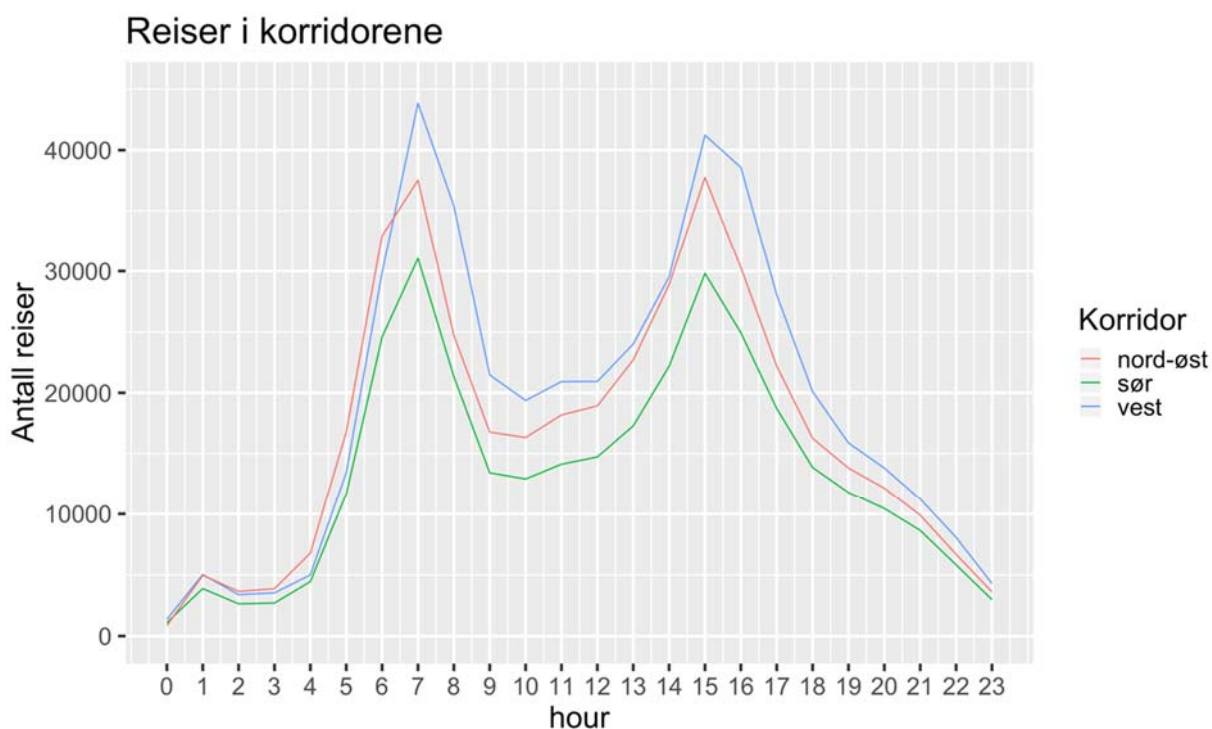
Figur 20: Differanse-ODM, MIS - Mobildata i prosentpoeng

5.3 Reisesstrømmer i de tre korridorene (Vest, Nord-øst, Sør)

Vi har benyttet OD-matrisene fra datasettet på bydel/kommune-nivå direkte for å beregne reisesstrømmer i de tre korridorene.

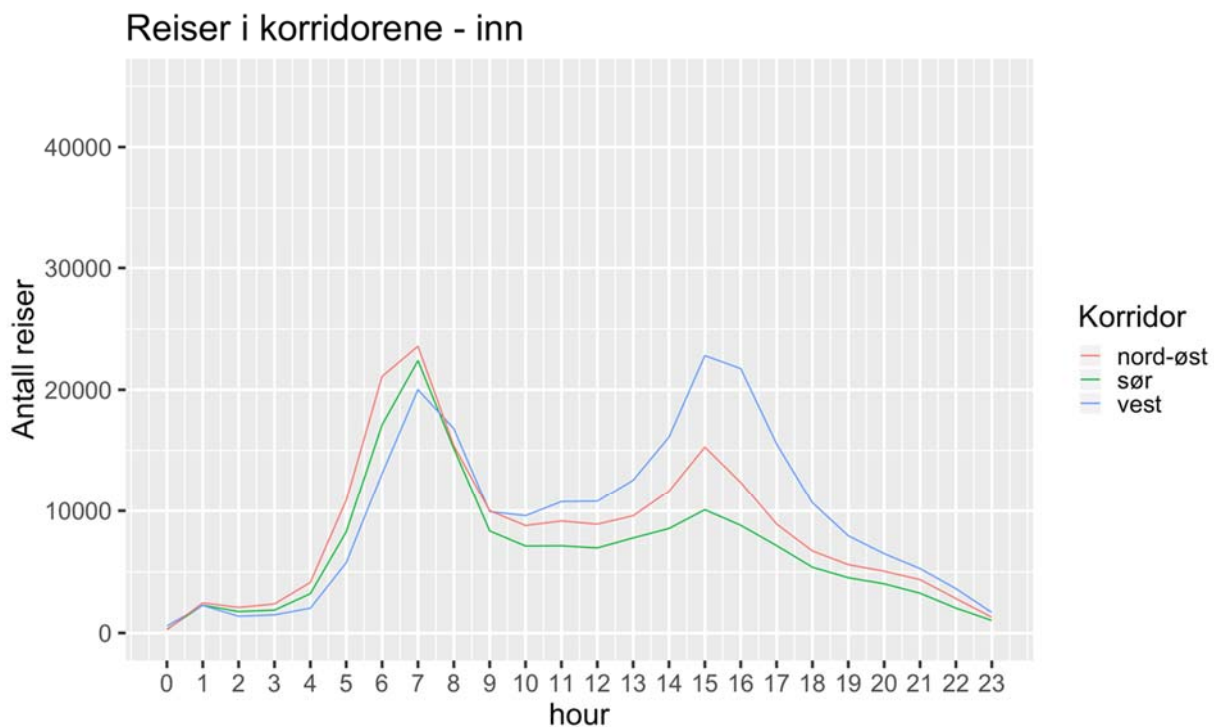
Da har vi lagt til grunn at en reise i en korridor har utgangspunkt (O) i en av markedsområdene som hører hjemme i korridoren og har en destinasjon (D) som er utenfor korridoren. Her har vi snittet datagrunnlaget over de fire onsdagene, slik at det er beregnet gjennomsnittlig antall for hver relasjon før vi måler reisesstrømmen.

Nord-Øst	Bjerke, Grorud, Stovner, Alna, Rælingen, Lørenskog, Skedsmo, Nittedal, Gjerdrum, Fet, Sørum, Ullensaker, Nes, Aurskog-Høland, Nannestad, Eidsvoll, Hurdal
Vest	Ullern, Vestre Aker, Nordre Aker, Bærum, Asker, Hurum, Røyken
Sør	Østensjø, Nordstrand, Søndre Nordstrand, Nesodden, Oppegård, Ski, Enebakk, Frogn, Ås, Vestby

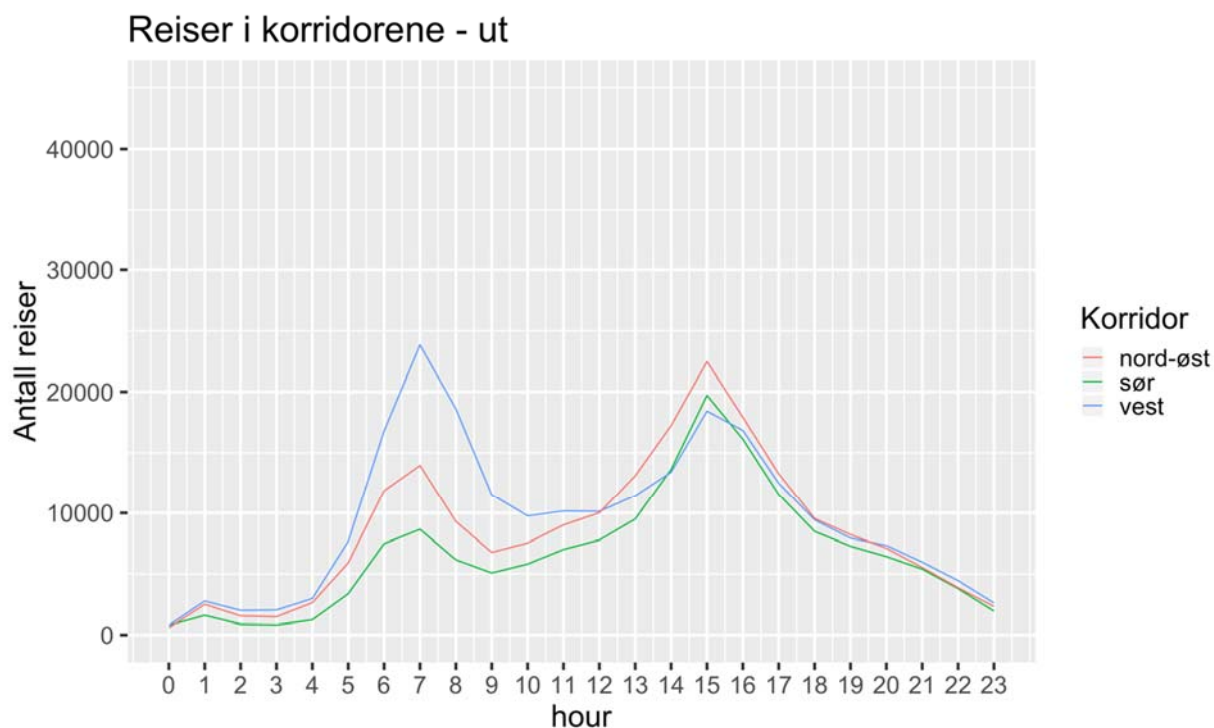


Figur 21: Totalt reiser i korridorene over døgnet

Figur 21 viser totalt antall reiser delt i korridorer slik som tabellen over viser. Jevnt over ser vi at volumene i de tre korridorene er litt forskjellig (som antageligvis har mest med befolkningstall å gjøre), men også at tidspunktet for reise kan synes noe forskjøvet. De som reiser i korridoren Nord-øst ser ut til å starte noe tidligere på morgenen enn de som reiser i Vest, som igjen starter før de som reiser i korridoren Sør.



Figur 22: Reiser langs de tre korridorene mot sentrum over døgnet

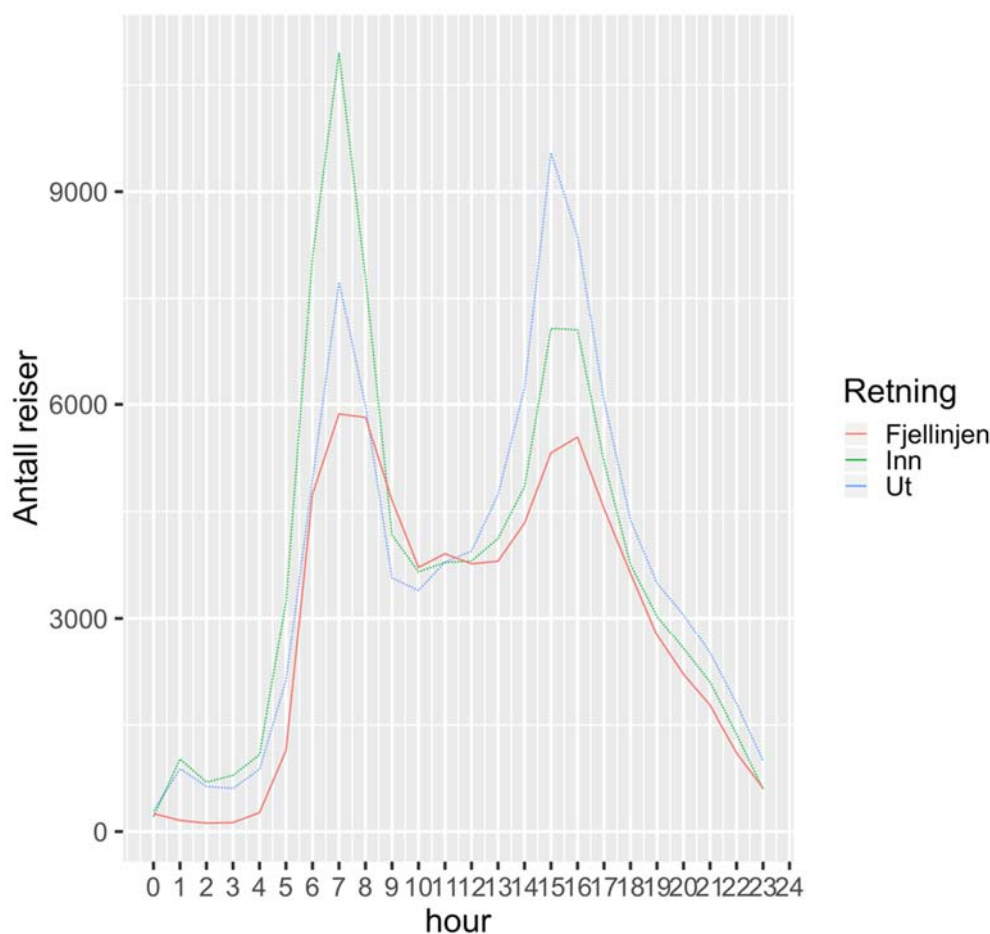


Figur 23: Reiser langs de tre korridorene fra sentrum over døgnet

Når vi deler dette i retningsbestemte reiser (Figur 22 og Figur 23), så ser vi større forskjell hvor korridor Vest er den som visuelt skiller seg ut fordi den har mye høyere grad av reiser *ut* i morgenrush (altså fra Oslo sentrum), mens de andre tre mye tydeligere har et reisemønster som tilsier arbeidsreiser mot Oslo.

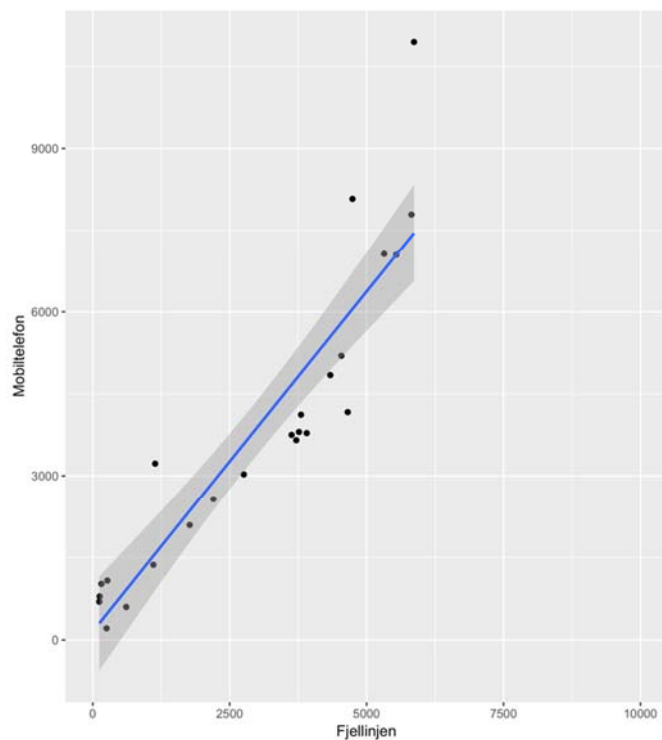
5.4 Hvor mange reiser over bomringen

Fra Fjellinjen mottok vi kjøretøystellinger for bomsnittet fra vest og inn i Oslo. Dette bomsnittet er enveis fra Bærum og inn i Oslo, og dataene ble levert med en times oppløsning for alle de fire onsdagene i februar. Tilsvarende passeringer i dette bomsnitt kan også beregnes ved bruk av mobildata, der samme metodikk som i kapittel 5.3 over blir benyttet. Et gjennomsnittresultat over de fire onsdagene fra Fjellinjen plottet i sammen som mobildata aggregert over dette snittet, gir fordelingen per time som vist Figur 24. OD-datasettet på bydel/kommunenivå er lagt til grunn her.



Figur 24: Antall reiser over bomringen fra Bærum, Mobildata og Fjellinjens tall (Fjellinjen teller kun i samme retning som mobildatas 'inn')

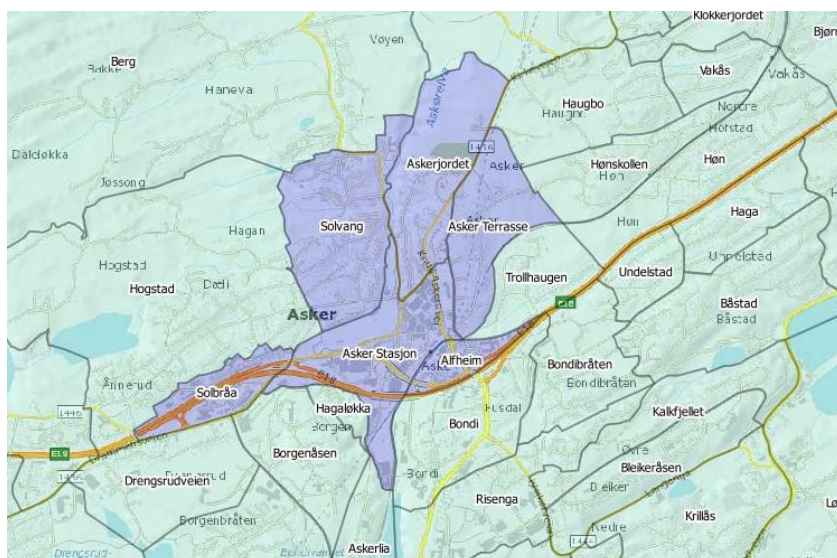
Mobiltelefonidata viser nesten systematisk et høyere antall reiser, men det *kan* forklares i at man her ikke vet om det er én eller flere telefoner i hvert kjøretøy (eller om reisen foregikk med en transportmodi som ikke telles av Fjellinjen). I Figur 25 vises et korrelasjonsplott mellom de to datakildene, som stort sett gir godt samsvar, bortsett fra store avvik i rushtid (hvor Fjellinjen har et betraktelig lavere antall passeringer).



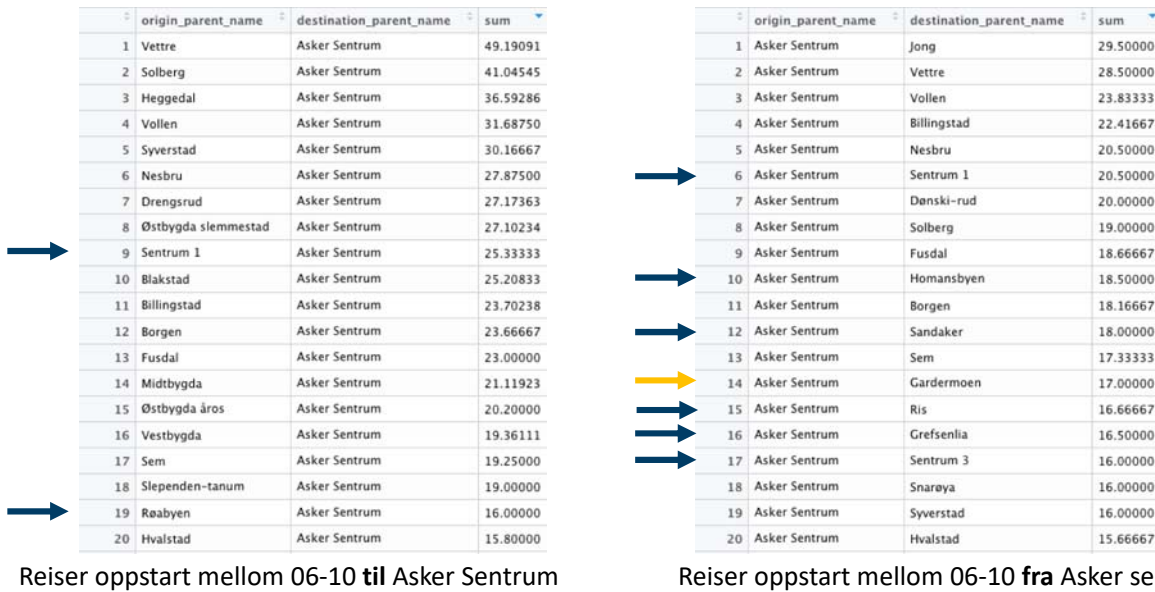
Figur 25: Korrelasjonsplott mellom Fjellinjen og mobildata i samme retning

5.5 Reiser Asker Sentrum

Det ble gjort en egen bestilling og øvelse på analyse av reiser til/fra Asker Sentrum, der sentrum ble definer som i Figur 26. Det vil si datasett 3 fra kapittel 3.2. Dette ble gjort på grunnkrets heller enn bydel/kommune (ettersom delmarkedsområde Vest 3 i praksis er *hele* Asker kommune).



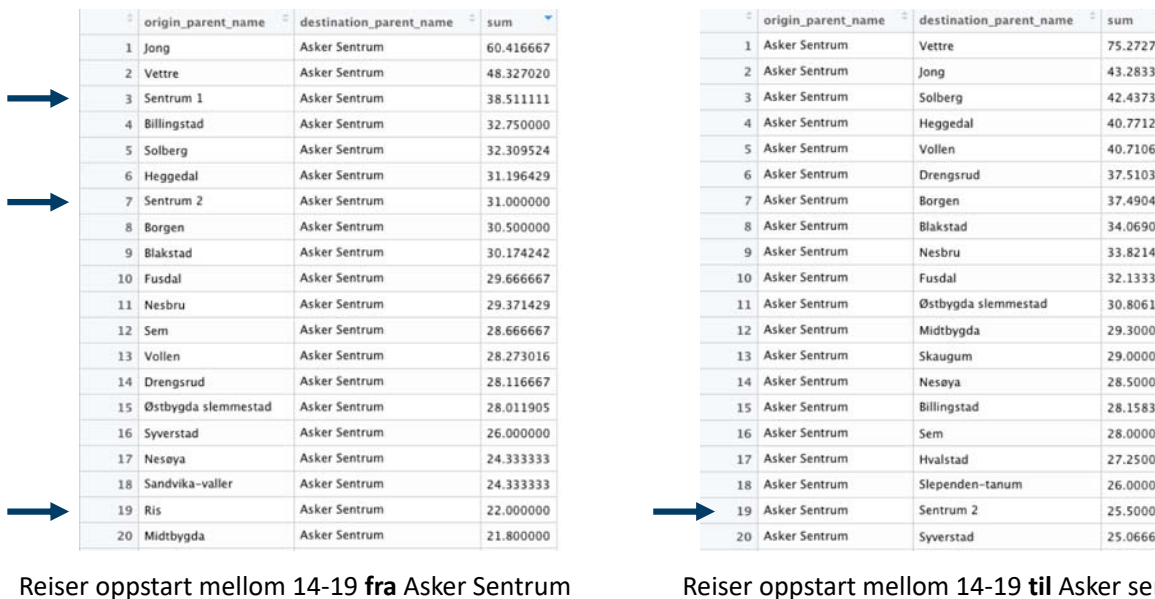
Figur 26: Oversiktskart over Asker sentrum (markert i blått) og omliggende områder



Reiser oppstart mellom 06-10 til Asker Sentrum

Reiser oppstart mellom 06-10 fra Asker sentrum

Figur 27: Reiser til (venstre) og fra (høyre) i morgenrush (06-10), blå piler markerer grunnkretser i Oslo sentrum.



Reiser oppstart mellom 14-19 fra Asker Sentrum

Reiser oppstart mellom 14-19 til Asker sentrum

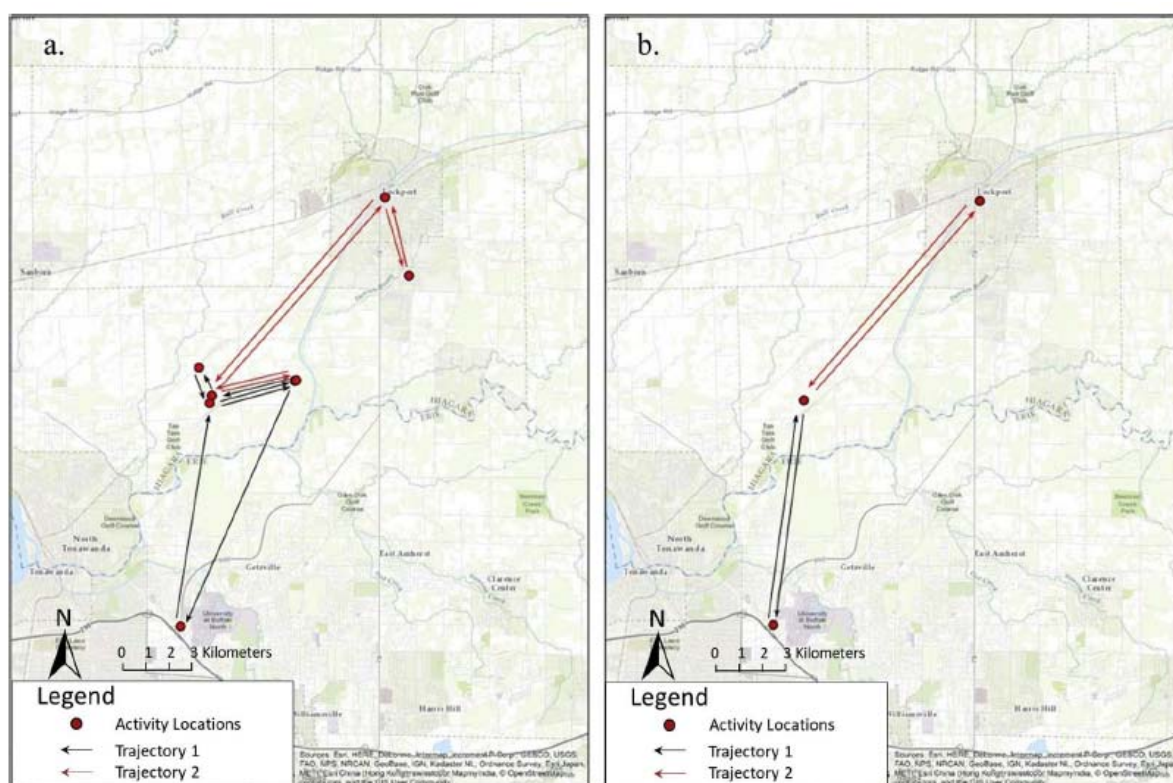
Figur 28: Reiser fra (venstre) og til (høyre) i ettermiddagsrush (14-19), blå piler markerer grunnkretser i Oslo sentrum.

Figur 27 og Figur 28 viser en rangert fremstilling over hvilke grunnkretser som hyppigst er origin og destinasjon i relasjonene ved reiser til/fra Asker sentrum i rushtid. Dette bildet domineres av lokale reiser i pendling inn til Asker sentrum, mens de som bor i Asker sentrum og reiser ut av sentrum hvis de reiser «langt» pendler inn til Oslo sentrum og de store arbeidsplassene rundt sentrum (Ris, Homansbyen, Sandaker).

5.6 Reiser med tog

Som en ekstra utfordring har det blitt jobbet spesielt med å få fordelt reisemodi mellom jernbane og «øvrige» transportformer. Her har vi i SINTEF samarbeidet med Telia: Telia har implementert og videreutviklet noen av algoritmene som er oppsummert i kapittel 2, og SINTEF har bidratt med kvalitetssikring og evaluering av resultater.

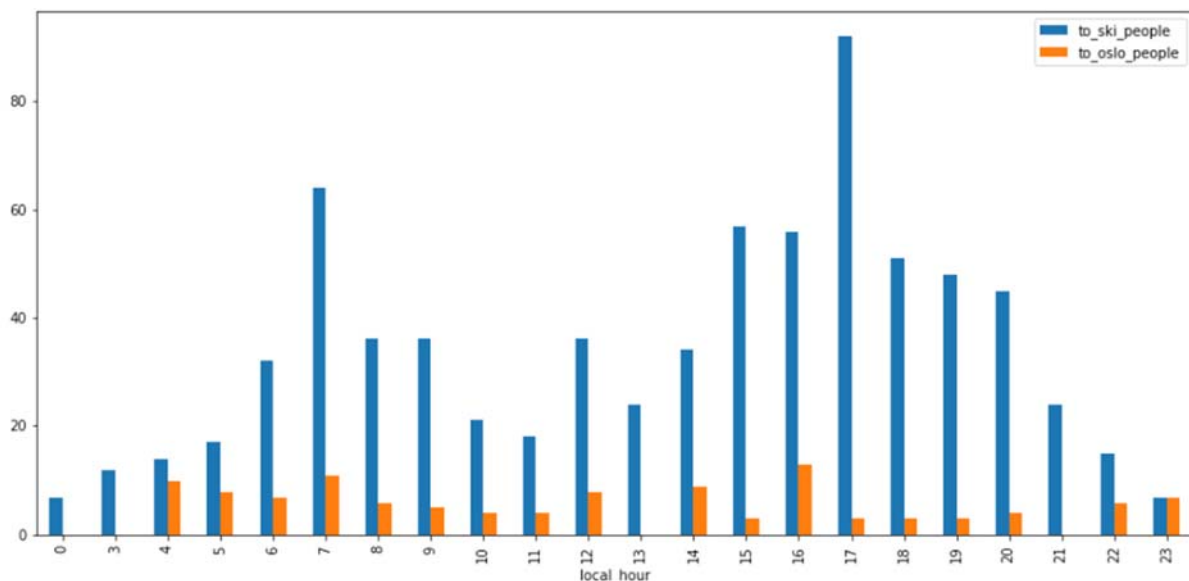
Oppsummert så prøver dagens algoritme seg på at man har forsøker å innpasse observasjonene fra mobilmaster med en «mest sannsynlig» reisevei. Der hvor den mest sannsynlige passer bedre med jernbanespor enn andre alternativer så antar man at sporet har reist med jernbane. Dette er en utfordrende øvelse først og fremst med hensyn til usikkerheten til hvilke mobilmaster en spesifikk telefon er koblet til ved et hvert registreringstidspunkt, se Figur 29 for en illustrasjon.



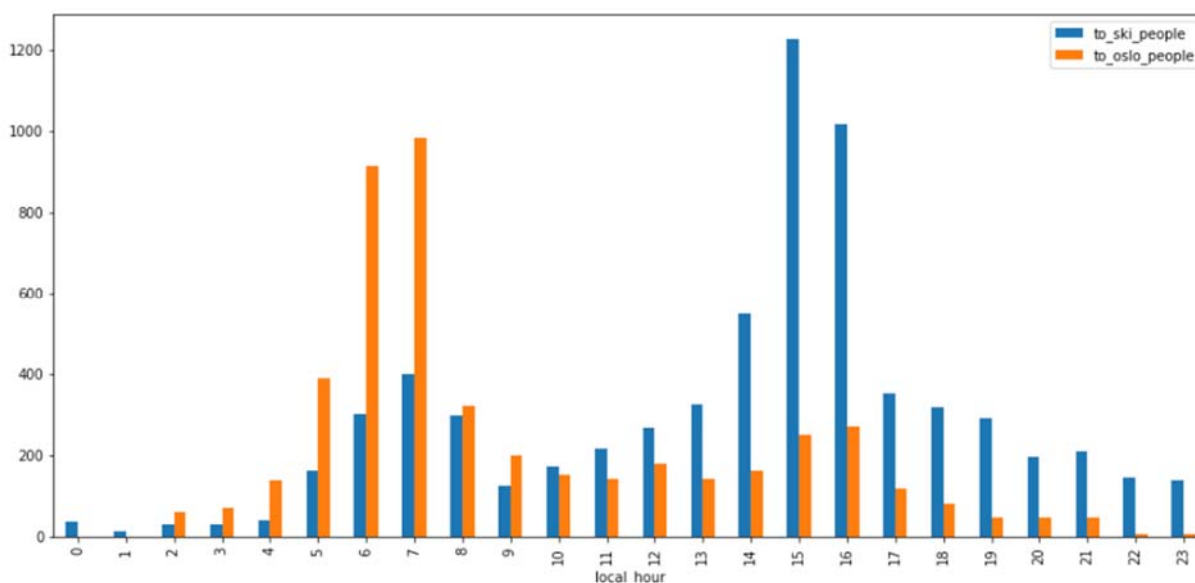
Figur 29: Illustrasjon fra Wang and Chen (2018) for viser utfordringen med mobildata og oppkobling til basestasjoner som nødvendigvis ikke er nærmeste. Høyre: Rådata der oppkobling til ulike mobilmaster vises. Venstre: Behandlet data som viser det faktiske reisemønsteret med hensyn til nærmeste basestasjon.

Dette kan igjen benyttes for å gi en fordeling av reisende over døgnet eller oppdeling av ODM-matriser i jernbanereisende og resten (for eksempel som prosent). Dette er antatt tilgjengelig i Telias infrastruktur fra medio November 2019.

Et eksempel på en type utprøving som er gjort, er å utnytte «naturlige eksperimenter» som når banestrekninger uplanlagt har måtte stenge. Et eksempel på dette er vist i Figur 30 - Figur 31, hvor strekningen Ski – Oslo var stengt for trafikk (29. august), mens en uke senere 5. september var strekningen åpen for vanlig trafikk.



Figur 30: Togreisende Ski-Oslo 29.08.2019



Figur 31: Togreisende Ski-Oslo 05.09.2019

Som vist i Figur 30, er det en åpenbar utfordring med en slik modell-basert tilnærming til konstruksjon av reiser, nemlig at når det ikke går tog – så vil den lille andelen av feil i modellen fortsatt finne reisende på toget som en kombinasjon av at man reiser veldig nært tog, med en hastighet som ikke er ulik togets. I dette plottet skulle de oransje søylene gått til null fordi vi vet at det er ikke var togtrafikk der i det hele tatt. Forskjellen på de to figurene (legg merke til ulike y-akser også), viser at modellen er sensitiv for så store endringer som det var her, men nøyaktigheten for mindre variasjon vet vi mindre om.

Det finnes heller ikke noen «gullstandard» å sammenligne med, da omregning fra dørtelling til reisende på strekningen også er en modellbasert tilnærming slik at man sammenligner to kilder med litt usikker feildistribusjon.

Samtidig med dette arbeidet har Telia ferdigstilt en modell for nedbryting i reisemode. Denne vil etter sigende bli kommersielt tilgjengelig fra 2020, og vil gi OD-matriser lignende det som har vært brukt i dette prosjektet med brutt ned på reisemode slik som eksempelet vist i tabellen under:

Fra	Til	Tidspunkt	Antall	Mode
A	B	10:00-11:00	20	Tog
A	B	10:00-11:00	30	Vei
A	B	10:00-11:00	50	Total
B	C

Dette vil gi en andel av reisende som velger tog (og noen andre reisemodi). I skrivende stund er dette på vei i produksjon så vi har ikke vurdert kvaliteten på denne metoden direkte. Det er sannsynlig at det vil gjelde de samme forbeholdene som dette notatet for øvrig – at metoden vil være internt konsistent, men at det å sammenligne den med andre datakilder vil kunne være problematisk. Ikke fordi den nødvendigvis er feil, men fordi forutsetningene er ulike og feilmarginene også vil slå ulikt ut.

6 Mobiltelefondata som datakilder i transportanalyser

Det er bygget høye forventninger til mobiltelefondata som kilde til mobilitetsanalyser. I feltet er dette en ny kilde som søker å avdekke *revealed preference* i en skala og kostnadsbilde som gjør det veldig interessant for mobilitet/transportanalyser. I utgangspunktet så er det en datainnsamlingsmekanisme som er mindre beheftet med de tradisjonelle seleksjonsbias-utfordringene som hukommelsesbias, respons/frafall og andre former for seleksjonsbias. Det betyr ikke at metoden er fri for metodiske utfordringer.

Mobiltelefonidata som utgangspunkt for mobilitetsdata er sekundærbruk² av kilden, og det betyr også noen utfordringer. Vi beskriver her kort noen av de utfordringene vi har sett, og ser med denne typen datakilde. Det betyr ikke at datakilden er ubrukelig, men at den må innføres og benyttes som et supplement hvor den er god – og ikke erstatte eksisterende kilder hvor de er bedre:

Utfordring: Samplingsraten på stedfesting er ujevn. Det vil si at man ikke har faste intervall på når man stedfester en telefon, men at stedfestingen skjer basert på mobiltelefonens trafikk (typisk minimum hvert 10. minutt), men hyppigere når telefonen er i bruk (både tale- og datatrafikk) eller når det forflytter seg mellom celler. Typiske tall oppgitt fra Telia er at man har 500-700 stedfestinger per døgn i gjennomsnitt i Norge. Til sammenligning så har GPS-baserte studier vist at man bør ned mot/under 2 minutter samplingsrate for å pålitelighet (i for eksempel reisemodedeteksjon), altså omtrent 700 stedfestinger i døgnet (uten å ta hensyn til at de fleste er stasjonære når de sover).

Når man skal gjenskape en faktisk reise, så må det benyttes rekonstruksjonsteknikker som har en eller annen forforståelse for hvordan mennesker beveger seg – og i dette ligger det at det er et estimat med usikkerhet på både spesifikk reiseveg og -modi. Ikke ulikt hva man gjør for å gå fra dørtelling til reiser.

Utfordring: Stedfestingen er basert på en antagelse om proksimitet, altså nærhet til en basestasjon. Utforming av dekning og forhold for radiobølger gjør at dette er en antagelse beheftet med usikkerhet, refleksjoner i bygninger, vann, og lignende gjør at signaler kan plukkes opp av en basestasjon som i utgangspunktet ikke er nærmest forbrukeren – og i mobilitetshensyn så kan for eksempel innslag på basestasjon over åpent vann plassere inn håndsettet på feil side av innsjø og lignende. Dette er en vesentlig utfordring når man skal gjøre rekonstruksjon til reiser, hvor forståelsen har veldig mye å si for «routing»-algoritmen som allokere en reise. Nærhet til kollektivtransport vil for eksempel dreie allokering reiser til reisemodi, mens en feilplassering her vil kunne gi et vesentlig bias i allokeringsalgoritmen.

Utfordring: «Forventning» om at definisjoner fra transportfaget skal holde. Når datakilden varierer veldig fra hva man tradisjonelt sett bruker, så er det å holde på definisjonene fra mer tradisjonelle kilder potensielt farlig. For eksempel er definisjonen av «en reise» fra transportplanlegging typisk varianter over «enveis bevegelse fra én origin til én destinasjon for én hensikt»³. Som mange andre kilder så gir ikke mobiltelefonidata noe innsyn i hensikten. Hvis man i fortolkning av resultater holder på en slik definisjon, mens man i realiteten i prosessering fra rådata opp til OD-matrise i utgangspunktet beregner med en annen definisjon – så er definisjonskonflikten et faktum.

Utfordring: Mobiltelefonidata har ikke etablerte og anerkjente kvalitetsmetriker. Når man benytter mer tradisjonelle metoder (som for eksempel «stated preference»-undersøkelser) så har man etablerte kvalitetsmål av typen svarprosent, konfidensintervaller/feilestimatorer som kan nyansere fortolkningen. Slike

² Altså at primærhensikten med denne datainnsamlingen *ikke* er å beskrive mobilitet, men noe helt annet. All annen bruk som man indirekte kan benytte dataen til må innrette seg etter prioriteringene fra primærbruken.

³ Akkurat denne er hentet fra (Ogden, 1979), hvor det helt riktig påpekes at denne definisjonen er av mange ansett som en overforenkling av virkeligheten.

mål, med «tolkningsregler» finnes ikke for mobiltelefonidata ennå. Det blir litt opp til hver enkelt analyse å sannsynliggjøre kvaliteten.

Gitt alt dette, så ser vi også noen fordeler:

Fordel: Mobiltelefonidata gir veldig rask vei til stor grad av dekning for store populasjoner – som med manuelle metoder vil være tilnærmet umulig å oppnå. I dette notatet har vi sett på mobilitetsmønsteret til i overkant 1,3 millioner mennesker – naivt beregnet til å være basert på rådata om bevegelsen til 600 000 mobiltelefoner⁴. Å gjennomføre en tradisjonell spørreundersøkelse på en brøkdel av denne populasjonen ville være mye mer kostnads- og tidskrevende enn å benytte mobiltelefonkilder (med alle de seleksjonsutfordringene det også vil gi).

Fordel: Supplement til andre automatiske kilder:

Vi ser at det er relativt sett godt samsvar mellom de store reisemønstrene når man sammenligner delmarkedsområde til delmarkedsområde mellom reisevaneundersøkelser med intervju og mobildata (se kapittel 5.2). Dette betyr at dette kan være en utfyllende kilde til både manuelle og automatiske kilder, hvor mobildata kan fylle ut hull eller svakheter hvor man har dårlig dekning i automatiske kilder (passasjertelling eller RVU).

⁴ Under en antagelse om at Telia har en markedsandel på ~50%. Det er en forenkling.

7 Konklusjoner

Reise-identifikasjon fungerer godt når det er definerte og relativt store O-er og D-er, slik som i utgangspunktet for denne analysen hvor man hadde delt inn Oslo- og Akershus i 39 separate geografiske områder. Da er mobiltelefoni veldig robust og pålitelig i å fange transport fra et område til et annet. Dette betyr at man må ha en til dels utviklet problemformulering før man bestiller data, og at man i analysen etterpå blir begrenset til den originale problemformuleringen sin i form av at man kan ikke si noe om andre reisemønstre enn de områdene man har definert, og har egentlig liten mulighet til i datasettet direkte bedømme om man har stilt korrekt problemformulering.

Analysen på grunnkrets nivå, bærer mye høyere preg av sensurer/utfall av data. Derfor bør det benyttes mer forsiktig. Grunnkretser har mange steder, selv i befolkningsrike kommuner, lave befolkningstall og kan derfor få veldig få like reiser for å komme seg over grensen for sensurering. Til eksempel har Oslo kommune 616 grunnkretser, hvorav 43 har under 100 mennesker i SSBs befolkningsoversikt. Dette begrenser anvendelsen av mobildata til å se på de større trekkene, heller enn detaljert planlegging av små tiltak.

De mer avanserte anvendelsene av mobiltelefonidata, både som vi ser konturene av i denne studien og i litteraturen, behøver berikelse med andre datasett. Som et minimum så bør man berike mobiltelefoni-kildene med georefererte datasett på rutenett/transportmønstre og gjerne tidfestede transporter for å støtte og styrke allokeringen inn i reiser og reisemønstre.

Erfaringen med bruk av mobildata til transportanalyse som i denne studien mener vi oppsummeres best med at dette er en umoden kilde. Det vil si at vi regner det som et stort gap mellom tradisjonelle metoder i transportanalysefaget og samfunnsøkonomiske modeller og det som kommer ut av mobildatafangst. Begge metodene *resulterer* i OD-matriser, som på et litt overordnet nivå ser ut til å sammenfalle – men i detaljene så er det betydelig forskjell i definisjoner som man bør ta hensyn til.

Til isolerte oppgaver så er mobildata en ekstra kilde til innsikt, som for eksempel hvis man raskt skal skaffe seg et overblikk over større områder og større befolkningsstrømmer. Det er også mulig å benytte denne typen analyse til å se på trender, da vi innsamlingsmetoden synes internt valid – slik at endring over tid antageligvis vil representere en underliggende endring. Absoluttverdiene derimot vil variere mot andre kilder.

8 Kilder

Bachir, D., Khodabandelou, G., Gauthier, V., El Yacoubi, M., & Puchinger, J. (2019). Inferring dynamic origin-destination flows by transport mode using mobile phone data. *Transportation Research Part C: Emerging Technologies*, 101, 254-275.

Bassolas, A., Ramasco, J. J., Herranz, R., & Cantú-Ros, O. G. (2019). Mobile phone records to feed activity-based travel demand models: MATSim for studying a cordon toll policy in Barcelona. *Transportation Research Part A: Policy and Practice*, 121, 56-74.

Groth, S. (2019). Multimodal divide: Reproduction of transport poverty in smart mobility trends. *Transportation Research Part A: Policy and Practice*, 125, 56-71.

Hjorthol, R., Ø. Engebretsen og T.P. Uteng (2014) Den nasjonale reisevaneundersøkelsen 2013/14 – nøkkelrapport, TØI-rapport 1383/2014, Oslo: Transportøkonomisk institutt

Huang, H., Cheng, Y., & Weibel, R. (2019). Transport mode detection based on mobile phone network data: A systematic review. *Transportation Research Part C: Emerging Technologies*.

Montero, L., Ros-Roca, X., Herranz, R., & Barceló, J. (2019). Fusing mobile phone data with other data sources to generate input OD matrices for transport models. *Transportation research procedia*, 37, 417-424.

Schjalm, A. (1996) Kvalitetsundersøkelsen for Folke- og bolig telling 1990. SSB 96/10
https://www.ssb.no/a/histstat/rapp/rapp_199610.pdf

Vazifeh, M. M., Zhang, H., Santi, P., & Ratti, C. (2019). Optimizing the deployment of electric vehicle charging stations using pervasive mobility data. *Transportation Research Part A: Policy and Practice*, 121, 75-91.

Wang, F., & Chen, C. (2018). On data processing required to derive mobility patterns from passively-generated mobile phone data. *Transportation Research Part C: Emerging Technologies*, 87, 58-74.



Teknologi for et bedre samfunn

www.sintef.no