



Centraal Bureau voor de Statistiek

Programma Impact ICT

Onderzoeksrapport nr. 9

Medegefinancierd door het Ministerie van Economische Zaken en Prima

RAPPORTAGE PROJECT IMPACT ICT; MOBIELE TELEFONIE

May Offermans, Alex Priem, Martijn Tennekes

Kennisgeving:

De in dit rapport weergegeven opvattingen zijn die van de auteurs en komen niet noodzakelijk overeen met het beleid van het Centraal Bureau voor de Statistiek.

Datum:

16 april 2013

Samenvatting

Dit onderzoeksproject is een verdieping van analyses op basis van GSM-netwerkgegevens en vormt onderdeel van het innovatieprogramma IMPACT-ICT. GSM-netwerkgegevens zijn een relatief nieuwe 'Big data' bron waarbij geanonimiseerde en geaggregeerde gegevens over locatie, tijdstip, nationaliteit en dergelijke worden verzameld. Bij dit onderzoek is nauw samengewerkt met Vodafone) en Mezero. Nieuwe ontwikkelingen waarbij gegevens worden samengevoegd en geanonimiseerd maken het mogelijk om analyses te doen waarbij de privacy gegarandeerd wordt. In dit project is gekeken naar toepassingsmogelijkheden van deze geanonimiseerde en geaggregeerde GSM-netwerkgegevens voor statistieken.

De resultaten van dit onderzoek laten zien dat met behulp van achtergrondgegevens van de populatie inzichten in representativiteit worden verkregen. Deze populatiegegevens zijn niet gekoppeld aan de GSM-netwerkgegevens. De eerste indruk van de representativiteit is goed. Het is de bedoeling om in een vervolgonderzoek hier dieper op in te gaan. De bel-, sms- en data-intensiteiten in de tijd voor verschillende plaatsen laten zeer regelmatige patronen zien. Verwacht wordt dat subtiele veranderingen in deze patronen interessante resultaten zullen opleveren voor vervolgonderzoek. Tellingen van personen uit het buitenland blijken goed mogelijk te zijn en leveren interessante resultaten op. Het inpassen van deze tellingen binnen de toerismestatistieken lijkt lastig met de huidige dataset, maar biedt wel perspectief voor de toekomst.

In dit onderzoek is verkennend onderzoek gedaan voor het schatten van de 'daytime population', dat wil zeggen het aantal personen in een bepaald gebied in een bepaalde periode overdag (meestal kantooruren). Deze informatie is nieuw en opent nieuwe mogelijkheden op gebieden als binnenlandstoerisme, evenementen, vervoer en rampenbestrijding. Dit in tegenstelling tot bestaande cijfers op basis van de gemeentelijke basisadministratie (GBA). Uit deze gegevensbron is vast te stellen waar mensen 's nachts zijn (nighttime population).

Hoewel er nog verschillende methodologische issues zijn die moeten worden opgelost zijn de resultaten van dit project positief. In korte tijd zijn al zeer informatieve resultaten naar voren gekomen die maatschappelijk relevant kunnen zijn. Het CBS heeft besloten daarom deze gegevens op een structurele wijze verder te onderzoeken voor toepassingen bij nieuwe of bestaande statistieken.

Inhoud

1. Inleiding.....	3
1.1 Aanleiding van het project.....	3
2. Proces.....	3
2.1 Beschikbaarheid van data	3
2.2 Overdracht van gegevens en privacybescherming.....	3
2.3 Dataverwerkingsproces.....	4
2.4 Algemene kenmerken van de data	5
2.4.1 Mastgegevens.....	5
2.4.2 Populatiegegevens	5
2.4.3 Vaststellen van de plaatsbepaling	6
3. Representativiteit	7
3.1 Leeftijdsopbouw	7
3.2 Regionale dekking	9
3.3 Representativiteit mastgegevens.....	9
4. Resultaten	10
4.1 Belintensiteit in Nederland	10
4.1.1 Belpatronen in tijd en ruimte	10
4.2 Toerisme in Nederland.....	12
4.3 Schattingen daytime populatie.....	14
4.3.1 Algoritme.....	14
4.3.2 Resultaten schattingen daytime populatie.....	17
4.3.3 Discussie	19
5. Conclusie	20
6. Bibliografie	21

1. Inleiding

1.1 Aanleiding van het project

Dit onderzoek is een vervolg op het ‘GSM’ (Gebruik voor Statistiek van Mastgegevens) onderzoek dat in september 2010 is afgerond (Jonge, et al., 2012), (Roos, et al., 2010). In het ‘GSM’ onderzoek lag de focus op het ontwikkelen van een methode om de zendmastdata geschikt te maken voor positiebepaling van geanonimiseerde personen.

Gegevens uit mobiele telefoonnetwerken bevatten informatie over gesprekken, sms’jes, data-events¹, de locatie en tijdstippen. Mobiele telefoon netwerkgegevens vormen daarmee een rijke bron die inzicht geeft in gedrag, sociale netwerken en mobiliteit van de bevolking. Studies in een aantal landen hebben dit al aangetoond (Ahas, et al., 2008), (Eagle, et al., 2009), (Palchykov, et al., 2012), (Kwok, 2009). Immers, vrijwel iedere Nederlander is in bezit van één of meerdere toestellen² en in 2011 had 1 op de 3 abonnementen ook een data-abonnement voor internettoegang.

In het huidige project worden toepassingen verkend bij bestaande of nieuwe statistieken. Daarbij is expliciet aandacht voor het proces van dataverwerking waarbij de privacy van telefoongebruikers kan worden gegarandeerd. Het huidige project heeft als doelen: 1) de representativiteit van de dataset onderzoeken, 2) bel-, sms- en data-intensiteiten in de tijd voor verschillende plaatsen onderzoeken, 3) toerismecijfers het maken en vergelijken met bestaande bronnen en 4) de daytime populatie schatten (aantallen mensen schatten overdag). De verwachting is dat aanvullend onderzoek meer toepassingen van deze gegevens zullen opleveren.

2. Proces

2.1 Beschikbaarheid van data

Om over data te kunnen beschikken is het CBS een samenwerkingsovereenkomst aangegaan met Mezero en Vodafone. Deze overeenkomst loopt tot 1 mei 2013. Op basis van deze overeenkomst kan het CBS beschikken over twee aparte datasets, namelijk de gegevens uit het mobiele telefoonnetwerk en populatiegegevens met achtergrondkenmerken. Gegevens uit mobiele telefoonnetwerken bevatten informatie over gesprekken, sms’jes, data-events, de locatie en tijdstippen. Deze gegevens worden mastgegevens genoemd. De populatiegegevens (zie ook paragraaf 2.4.2) geven inzicht in achtergrondkenmerken van personen of bedrijven.

2.2 Overdracht van gegevens en privacybescherming

Privacybescherming is een belangrijk aspect binnen dit project. Bij elke stap in het in dit project is hierover nagedacht door Vodafone, Mezero en het CBS. Alle werkzaamheden zijn zo ingericht dat privacygevoelige informatie afgeschermd is

¹ Data-events ontstaan doordat gebruikers of applicaties dataverkeer genereren (bijv. internet of emailgebruik).

² Volgens de OPTA waren er in Q4 van 2011, 21,8 miljoen actieve aansluitingen.

van ongeoorloofde inzage. De fysieke beveiliging en het proces tussen Vodafone en Mezero is onafhankelijk getoetst door TNO. Alle partijen hebben ieder ook het onderzoek aan bestaande juridische kaders getoetst³.

De gegevens die door Vodafone worden geleverd zijn altijd ingedeeld in aggregaten (klassen) waardoor er geen sprake meer is van persoonsgegevens. Individuele personen zijn niet meer te herleiden. Deze werkwijze biedt als voordeel dat de privacy van personen die gebruik maken van het Vodafone netwerk gegarandeerd wordt en voldoet aan alle privacyrichtlijnen. Deze gegevens bieden geen mogelijkheid om mensen te volgen in de tijd, wat wel had gekund met microdata. De CBS-wet biedt voor onderzoeksdoeleinden aanzienlijk meer ruimte om bijvoorbeeld gegevens te analyseren op microniveau. In dit project is geen gebruik gemaakt van deze juridische status van het CBS. De geaggregeerde dataset was voldoende voor het beantwoorden van de doelstellingen van dit project.

De geaggregeerde data bestaan uit tellingen van toestellen of van telefoongesprekken, sms-en of data-events, uitgesplitst naar tijd, regio, land van herkomst. Deze geaggregeerde data zijn minder dan 2 maanden oud en statistisch beveiligd. Zo mogen de cellen in opgevraagde tabellen niet minder dan 15 unieke toestellen bevatten en mogen gegevens van toestellen niet langer dan 30 dagen achter elkaar gevolgd worden. Voor personen uit het buitenland die in Nederland op het Vodafone netwerk zijn gelden strengere eisen. Hier mogen gegevens van unieke toestellen slechts van 1 dag worden verzameld op geaggregeerde basis.

Naast de gegevensstroom uit het mobiele telefoonnetwerk zijn door Vodafone ook populatiegegevens beschikbaar gesteld. Dit zijn ook geaggregeerde gegevens en zijn hierdoor niet identificeerbaar. Voor dit project was een eenmalige levering met één specifieke peildatum voldoende om inzicht te krijgen in de populatiekenmerken.

2.3 Dataverwerkingsproces

Voor dit project wordt gebruik gemaakt van een specifiek proces om mastgegevens op te vragen. In deze gegevens zit informatie over belgedrag. Het opvragen van gegevens begint doordat het CBS bij Mezero een specifieke informatievraag neerlegt. Dat betekent in de praktijk dat er grote tabellen worden gevraagd aan Mezero met mastgegevens waarbij er verschillende uitsplitsingen worden gemaakt⁴. Mezero vertaalt deze informatievraag naar een query. Deze worden in een datacentrum van Vodafone verwerkt, ontdaan van alle persoonsinformatie, geanonimiseerd en de output wordt via een beveiligde verbinding aan Mezero geleverd. De tabellen worden vervolgens via een beveiligde verbinding naar het

³ De juridische kaders waarbinnen het CBS met privacygevoelige informatie om mag gaan zijn strikt geregeld in de CBS-wet en de Wet bescherming persoonsgegevens (WBP). Aanvullend heeft het CBS een eigen gedragscode. Het CBS heeft een functionaris voor de gegevensbescherming (FG). Deze persoon waakt over de naleving door het CBS van de wettelijke bepalingen inzake de privacybescherming en houdt daarnaast een meldingenregister bij van de verwerkingen van persoonsgegevens.

CBS verstuurd. De unieke persoonsgegevens blijven bij Vodafone en worden hier geanonimiseerd. Mezuro en het CBS kunnen alleen over geaggregeerde data beschikken.

2.4 Algemene kenmerken van de data

2.4.1 Mastgegevens

Per maand worden er bij Vodafone ongeveer 3,4 miljard records van 6-7 miljoen aansluitingen opgeslagen en verwerkt. Deze mastgegevens bestaan uit het gebruikte mobiele nummer (wordt direct vervangen door een geanonimiseerde code), het antennegebied, het tijdstip en het type netwerkgebruik (bellen, sms'en of mobiel internetten) maar ook informatie over het land van de provider indien het om buitenlanders gaat die in Nederland 'roamen'⁵ op het netwerk. In tegenstelling tot de eerdere uitgevoerde studie wordt de plaatsbepaling binnen het systeem van Mezuro gedaan. In paragraaf 2.4.3 wordt hierop dieper ingegaan.

De geaggregeerde tabellen van mastgegevens die gebruikt zijn in dit onderzoek zijn:

- Aantallen uitgesplitst naar bel-, sms- en data-events, per woonplaats, per tijdsinterval van 1 uur, voor de periode 5 december 2012 tot 31 december 2012.
- Aantallen buitenlandse events per woonplaats per tijdsinterval van 4 uur, uitgesplitst naar nationaliteit, voor de periode 3 januari 2013 tot 31 januari 2013.
- Aantallen unieke actieve toestellen per woonplaats per tijdsinterval van 4 uur, uitgesplitst naar inwoners en bezoekers, voor de periode 5 december 2012 tot 31 december 2012.

De tabellen hebben missende waarden wanneer er minder dan 15 waarnemingen zijn. Daarom is voor de tweede en derde dataset een ruimer tijdsinterval gekozen.

2.4.2 Populatiegegevens

De klanten van Vodafone zijn onder te verdelen in drie hoofdgroepen. Dit zijn klanten met een pre-paid, zakelijk of gewoon abonnement (post-paid abonnement). Deze drie groepen zijn qua grootte ongeveer gelijk⁶. Van de abonnementen zijn betrouwbare gegevens zoals geslacht en leeftijd beschikbaar. Van pre-paid is slechts van een klein deel geregistreerd. Van de zakelijke abonnees is geen geslacht of leeftijd bekend⁷.

⁵ Een gebruiker met een buitenlandse provider maakt bij roaming gebruik van het netwerk in Nederland.

⁶ De exacte verhoudingen zijn bedrijfsgevoelige informatie en kunnen derhalve hier niet vermeld worden.

⁷ Het bedrijf is wel bekend, maar het is niet gelukt om gedurende dit project een goede classificatie te realiseren die een aggregatieniveau mogelijk maakt.

2.4.3 Vaststellen van de plaatsbepaling

Uit de mastgegevens die in de telefooncentrale bij Vodafone terechtkomen zijn indirect locaties vast te stellen. Dit proces is tamelijk complex, en zal in deze paragraaf kort worden besproken.

Een GSM-mast bestaat uit meerdere cellen die ieder een bepaalde hoek bestrijken. De meeste GSM-masten hebben drie cellen die ieder 120 graden dekken. Van ieder event (bellen, sms, of data) is de bijbehorende cel-code opgeslagen.

Het is niet eenvoudig om op basis van de cel-codes de locatie te bepalen. Zo kunnen GSM-masten een bereik hebben van 21 kilometer, en andere masten een bereik van tientallen meters hetgeen een nauwkeurigere plaatsbepaling mogelijk maakt. Een bijkomend probleem is dat de dekkingsgebieden elkaar overlappen. Dit betekent dat iemand die zich gedurende een hele dag op dezelfde locatie bevindt (bijvoorbeeld thuis), verbinding kan hebben met meerdere cellen. In de praktijk zullen deze meerdere cellen daadwerkelijk afwisselend worden gebruikt, ook indien iemand zich niet verplaatst. Het proces dat mobiele telefoons aan cellen koppelt is behoorlijk dynamisch. Natuurlijk kunnen toestellen van personen zich ook verplaatsen van cel naar cel.

In een eerdere studie (Roos, et al., 2010) zijn de locaties bepaald door een Voronoi kaart te maken op basis van de mastlocaties. Dit is gedaan door voor ieder geografisch punt in Nederland te bepalen welke mast het dichtstbij is. Hierbij is dus geen rekening gehouden met de verschillende cellen. Ook het afwisselen van masten terwijl de daadwerkelijke locatie gelijk is wordt hiermee niet goed ondervangen.

In deze studie wordt gebruik gemaakt van een toewijzingsalgoritme dat door Mezuro is ontwikkeld. Van ieder toestel wordt de plaats bepaald aan de hand van de reeks cellen waarmee het is verbonden gedurende de dag. Een toestel bevindt zich in een bepaalde plaats, zeg A, zolang de cellen waarmee het zich verbindt (een deel van) het gebied van plaats A dekken. Zodra een cel wordt gebruikt die plaats A niet dekt, dan bevindt het toestel zich niet meer in A, maar in plaats B, de plaats waar de nieuwe cel het meest dekt.

Deze methode is robuust tegen de dynamiek van het toewijzingssysteem. Met andere woorden, als iemand de hele dag thuis is, dan zal zijn of haar toestel zich volgens het algoritme in zijn of haar woonplaats bevinden. Een nadeel van de methode is dat in dunbevolkte gebieden, bijvoorbeeld de provincies Groningen en Friesland, veel kleine plaatsen (dorpen) liggen waar geen events aan worden toegekend. De reden is dat hier masten staan met een groot bereik, waardoor niet duidelijk is in welk dorp het toestel zich bevindt. Momenteel werkt Mezuro aan een verbetering van het algoritme, waarbij de plaatsen proportioneel worden toegekend aan verhoudingen van de dekkingen van de betreffende cel. Ook het CBS zal onderzoeken of er verder verbeteringen mogelijk zijn.

3. Representativiteit

In het klantenbestand van Vodafone staan gegevens over de klanten van Vodafone. Deze gegevens zijn geaggregeerd naar leeftijd, geslacht, postcodeniveau, en het type abonnement (zakelijk of consument). De representativiteit van Vodafone klanten ten opzichte van de Nederlandse bevolking kan worden onderzocht door het klantenbestand naast de gemeentelijke basis administratie (GBA) van 1 januari 2013 te leggen.

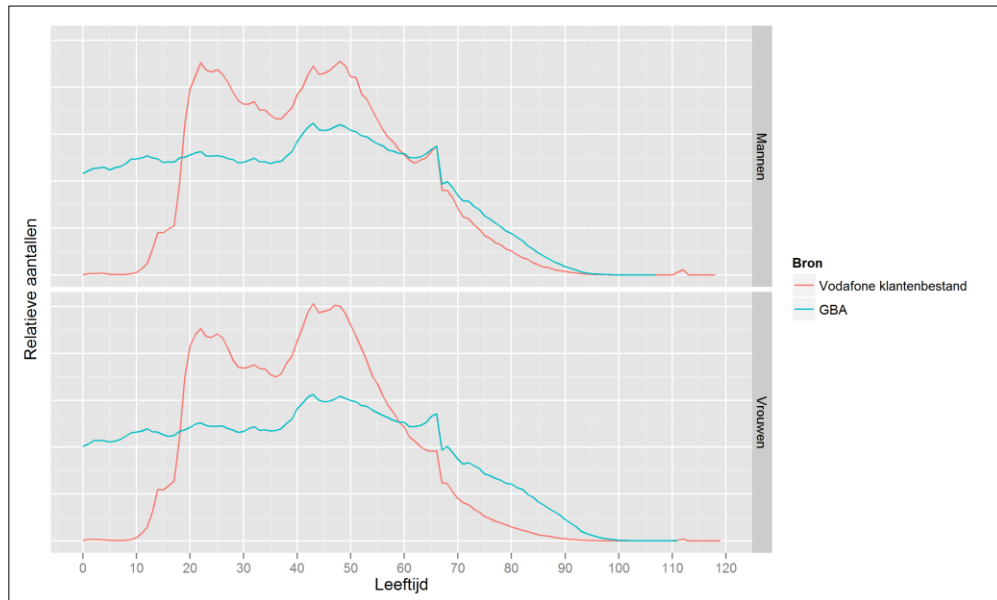
Het klantenbestand dat voor dit onderzoek is gebruikt dateert van 31 januari 2013. Verder zijn alleen de personen meegenomen die minimaal één maal actief waren op het mobiele netwerk (door middel van bel, sms of data-event) in de maand januari.

3.1 Leeftijdsopbouw

In Figuur 3.1 is de leeftijdsopbouw van de Vodafone klanten vergeleken met die van de GBA, uitgesplitst naar geslacht. In deze grafiek vallen enkele aspecten op:

- Kinderen zijn ondervertegenwoordigd, en volwassenen tot een jaar of 60 oververtegenwoordigd. Aannemelijk is dat erg jonge kinderen, geen mobiele telefoon hebben, en dat oudere kinderen een mobiele telefoon hebben die veelal op naam van één van de ouders is geregistreerd. Dit laatste wordt bevestigd door het relatief hoge aantal Vodafone klanten rond de 45 jaar (de rechter kamelenbulten in de rode grafieken bij 45 jaar).
- Behalve bovenstaande verschillen zijn de patronen in de GBA wat betreft leeftijdsopbouw opvallend goed zichtbaar in de Vodafone populatie. Met name de naoorlogse babyboom (66 jaar) is goed zichtbaar, maar ook de lokale demografische patronen rond de 30 en 45 jaar.
- Ouderen zijn ondervertegenwoordigd, met name vrouwen. Het is aannemelijk dat de lokale piek bij 112 jaar is veroorzaakt door onbekende geboortejaren die zijn geïmputeerd met 1900. Het aantal onbekende leeftijden in het register is zeer laag.

In Figuur 3.1 is per leeftijdsjaar het percentage Vodafone klanten ten opzichte van de GBA te zien. Hierbij is de Vodafone populatiedichtheid (rode lijnen in Figuur 3.1) gedeeld door de GBA populatiedichtheid (blauwe lijnen in Figuur 3.1). In deze grafiek is duidelijk te zien dat het percentage dat klant is bij Vodafone het grootst is rond de 20-25, en dat dit percentage bij benadering lineair daalt met de leeftijd.

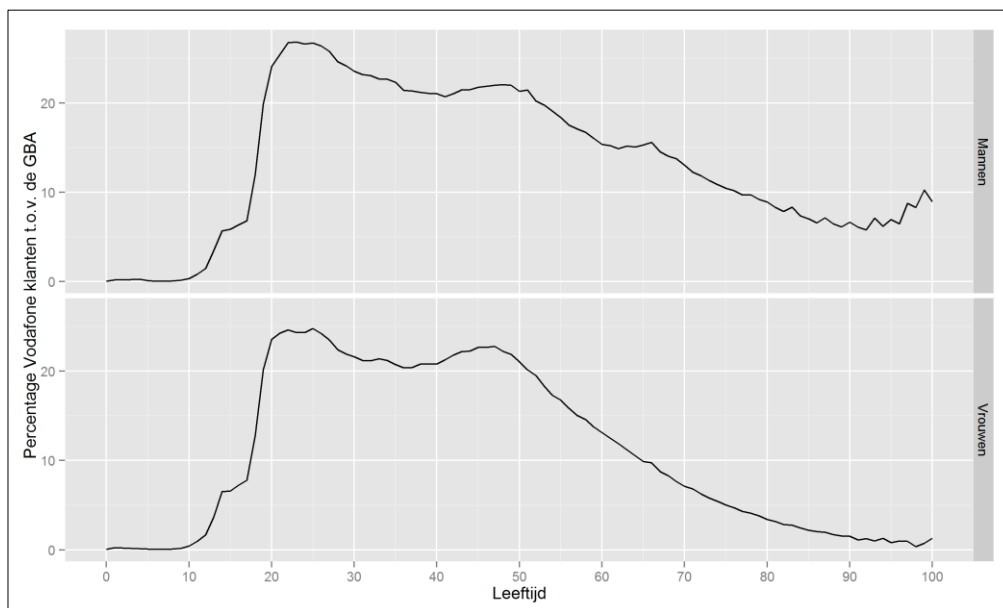


Figuur 3.1: Vergelijking van het klantenbestand met de GBA

Het percentage Vodafone klanten binnen de hoogbejaarde mannen ligt rond de 10%. De grillige staart van de grafiek wordt veroorzaakt door de kleine massa (zowel in de GBA als in de Vodafone klantendatabase) en vermoedelijk door fouten in de data. Voor hoogbejaarde vrouwen ligt dit percentage een stuk later, namelijk rond de 3%.

In Figuur 3.2 is een vergelijking gemaakt van klanten ten opzichte van de GBA naar leeftijd. Ook hier is te zien dat er een lokale piek zit rond de 48 jaar, met name bij vrouwen. Zoals eerder beschreven kan dit worden verklaard doordat deze mensen abonnementen afsluiten voor hun tiener kinderen.

Ondanks dat alleen het klantenbestand van Vodafone is onderzocht, en niet dat van andere grote providers, kan worden aangenomen dat deze globaal gezien dezelfde leeftijdsopbouw zullen hebben. De beschreven patronen, bijvoorbeeld dat ouders abonnementen afsluiten voor hun kinderen, zullen ook bij andere providers aanwezig zijn.



Figuur 3.2: Vergelijking van het klantenbestand met de GBA (vervolg)

Op basis van het Vodafone klantenbestand kan ook de dekking van Vodafone klanten ten opzichte van de populatie worden bepaald op regionaal niveau. Met behulp van thematische kaarten is de dekking van Vodafone klanten onderzocht. Deze kaarten zijn vanwege de bedrijfsgevoeligheid niet opgenomen in dit rapport. Uit dit onderzoek is gebleken dat Vodafone aandeel per gebied behoorlijk kan verschillen.

Belangrijk is dat het postcodegebied van alleen consumenten die een abonnement hebben (post-paid klanten) goed beschikbaar is, omdat hier de facturen naar gestuurd worden. Het opgeven van postcodegegevens is voor pre-paid klanten echter facultatief. Voor pre-paid klanten is daarom het postcodegebied slechts beperkt beschikbaar. De postcodegebieden van zakelijke klanten zijn geen postcodegebieden van bedrijfslocaties, maar (post)locaties van de betreffende bedrijven.

Het is gebleken dat in de gemeenten waar veel bedrijven met een Vodafone contract gevestigd zijn, het totaal aandeel Vodafone klanten sterk wordt beïnvloed.

3.3 Representativiteit mastgegevens

Mastgegevens en populatiegegevens zijn binnen Vodafone gescheiden datastromen. Aan de mastgegevens zijn geen achtergrondkenmerken zoals leeftijd en geslacht gekoppeld. Mastgegevens worden door telefoons gegenereerd, die indirect aan personen kunnen worden gekoppeld. Hierdoor is het mogelijk om de representativiteit te onderzoeken. Door het ontbreken van een directe koppeling is blijft het lastig om uitspraken te doen over de representativiteit van mastgegevens ten opzichte van populatiegegevens.

In het ideale scenario zijn alle Vodafone klanten actief op het netwerk met telefoongesprekken, sms'jes of data-events. Als dit het geval is kan het

klantenbestand worden gebruikt als hulpinformatie bij het schatten van populatieaantallen.

In de praktijk zal het echter zo zijn dat een deel van de klanten de mobiele telefoon niet of nauwelijks gebruikt. Bovendien zal dit deel zeer waarschijnlijk selectief zijn met betrekking tot geslacht, leeftijd, en andere achtergrondvariabelen. In het klantenbestand zijn de personen die in de maand januari helemaal niet actief waren al buiten beschouwing gelaten. Mensen die de mobiele telefoon weinig gebruiken zitten echter nog wel in het klantenbestand.

In de aangeleverde tabellen van mastgegevens, staan het aantallen unieke actieve toestellen voor 1- en 4-uurs tijdsblokken. De representativiteit van de mastgegevens is afhankelijk van hoeveel Vodafone klanten minimaal één keer per 4 uur actief zijn op het netwerk, en de selectiviteit van deze groep.

4. Resultaten

4.1 Belintensiteit in Nederland

4.1.1 Belpatronen in tijd en ruimte

Zoals eerder in paragraaf 2.4.1 is aangegeven zijn de mastgegevens uitgesplitst naar drie verschillende verkeerstypes namelijk: data-, voice- en SMS-events. Deze gegevens blijken een duidelijk, regelmatig dag/nacht-patroon te vertonen, in combinatie met een week/weekeinde-patroon. In nagenoeg alle figuren is te zien dat er meer dataverkeer is dan voice-verkeer, en dat voice-verkeer weer hoger is dan SMS-verkeer.

Bij SMS en voice-verkeer valt op dat het verschil tussen weekdays en weekenddagen veel groter is dan voor dataverkeer. Daarnaast is er nagenoeg geen SMS- en voice-verkeer tussen 2:00 en 4:00 uur 's nachts, zelfs niet in steden als Amsterdam en Rotterdam. Dataverkeer verschilt van de andere verkeerstypes dat het een hoger minimumniveau heeft, en minder gevoelig is voor weekpatronen.

Dit lijkt een hypothese te ondersteunen dat SMS- en voice-verkeer meer gerelateerd is aan (economische) activiteit, en dataverkeer meer correleert met aanwezigheid van mensen (met hun mobiele telefoons). Dit wordt bevestigd door de sterke afname in belverkeer in de periode tussen kerst en nieuwjaar.

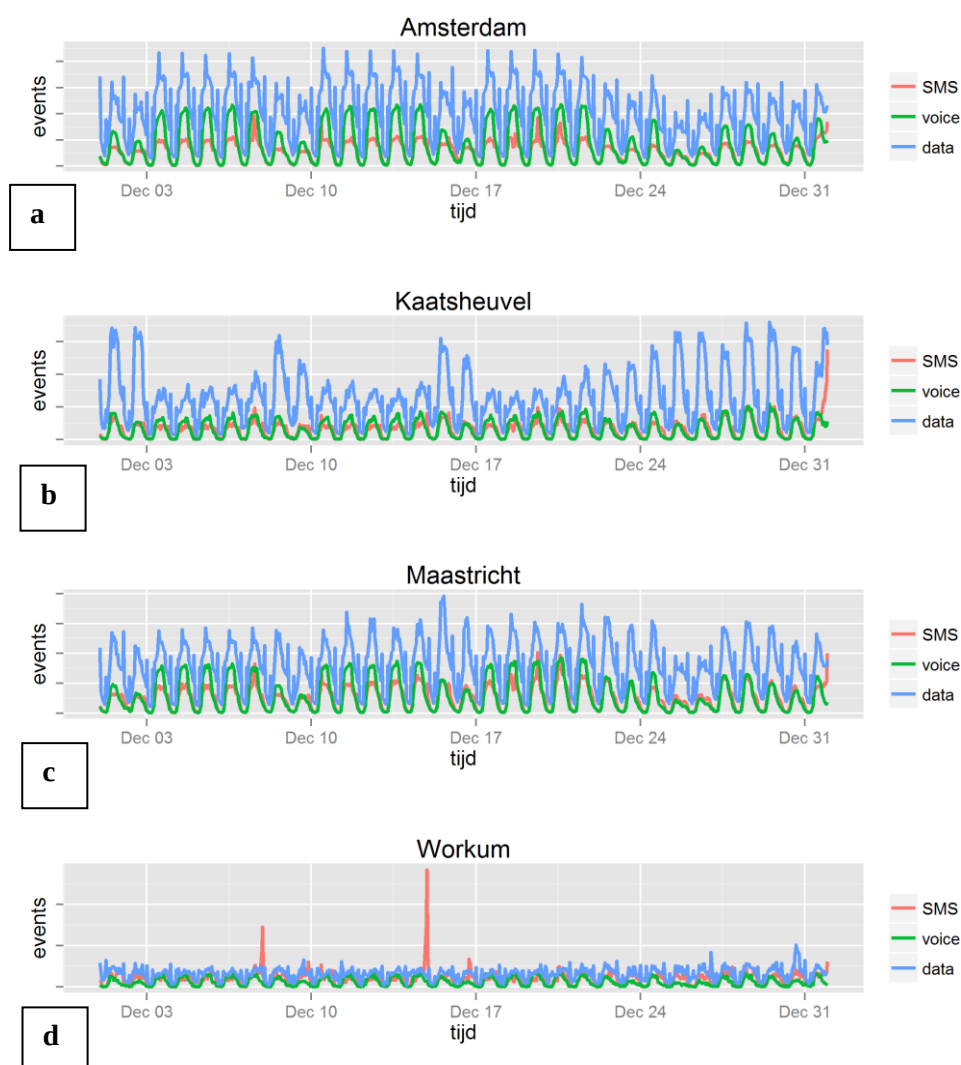
De belpatronen rond Kaatsheuvel in figuur 3.3 (een gebied met een groot pretpark) lijken deze hypothese te ondersteunen: terwijl in andere steden de activiteit van met name voice- en SMS-verkeer in de laatste week van december sterk verminderd, neemt het dataverkeer rond Kaatsheuvel juist toe, maar het voice-en SMS-verkeer niet. Ook is de week/weekeindepatroon van de data-activiteit rond Kaatsheuvel nagenoeg tegengesteld aan dat van bv. Rotterdam. Blijkbaar is het zo dat mensen niet bellen als ze in een attractie zitten, maar wel veel data versturen met de telefoon.

In de belpatronen voor Amsterdam is ook te zien dat de nacht van zaterdag op zondag de hoogste minimumwaarde heeft voor voice-verkeer; dit correleert mogelijk

met horeca-bezoek. Een grotere ruimtelijke resolutie (op wijk/buurtniveau) zou hier meer uitsluitsel over kunnen geven.

Daarnaast laat de data ook andere andere opmerkelijke zaken zien. Zo is in alle datasets rond 31 december 2012 een piek in SMS-activiteit te zien. De verhouding SMS/voice/data lijkt echter voor alle vier gekozen gebieden anders te zijn. Hoewel eindejaars-SMS-jes een duidelijk te herleiden sociaal event zijn, zijn er ook gebeurtenissen die minder eenvoudig te verklaren zijn.

Zo piekt de SMS-activiteit in de omgeving Workum rond 7 en 15 december (zie figuur 4.1d), waarbij het onduidelijk is of dit een technische storing is, of dat er in die periode in die regio door bv. politie dan wel in het kader van een televisieprogramma veel SMS-berichten zijn verstuurd. Andere Friese gebieden laten eenzelfde activiteit zien, maar het middelpunt van de activiteit lijkt zich in Workum te bevinden.

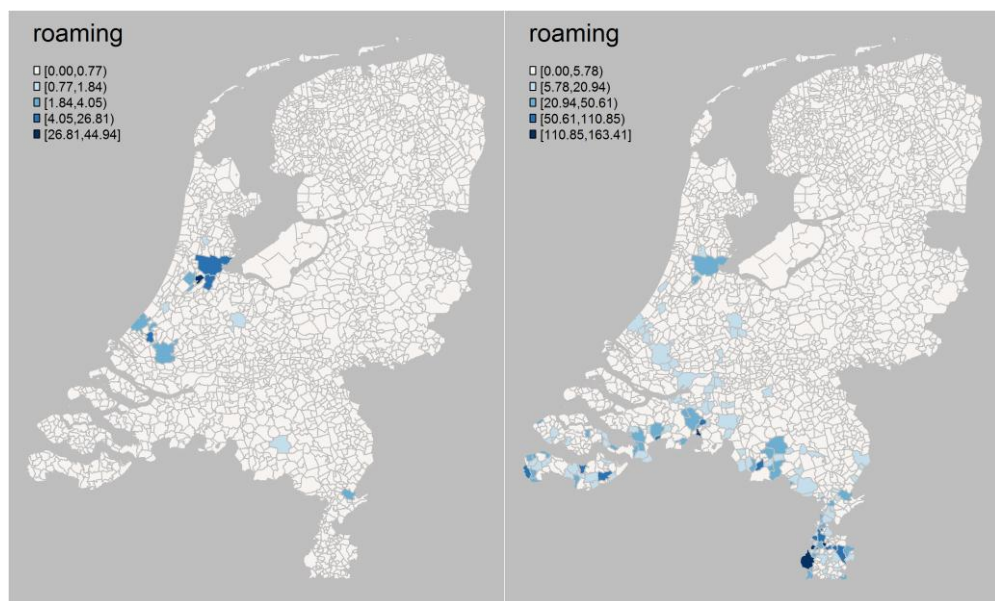


Figuur 4.1: Aantallen SMS, voice en data-events voor verschillende steden: Amsterdam(a), Kaatsheuvel (b), Maastricht (c) en Workum (d).

4.2 Toerisme in Nederland

In de mastgegevens is ook een landscode opgenomen die aangeeft met welke buitenlandse provider het toestel verbinding heeft. Hierdoor zijn toestellen uit het buitenland ook in Nederland te bereiken. Deze landscode wordt bij elke gebeurtenis op het netwerk vastgelegd. Dit maakt het mogelijk een onderscheid te maken tussen mensen uit Nederland en mensen uit het buitenland met een andere provider. Vrijwel iedereen die met een toestel met buitenlandse provider aankomt in Nederland ontvangt een SMS van een Nederlandse provider. Hierdoor is het mogelijk om inzicht te krijgen in aantallen, verblijfslocaties en land van herkomst van buitenlandse toeristen (Raun & Ahas, 2013).

Zoals in 2.4.1 is aangegeven is de roaming-dataset opgedeeld op 4-uurs-blokken naar woonplaats. In figuur 4.2 is voor 11 januari van 12.00 tot 16.00 het aantal roaming-events weergegeven. Hierin is duidelijk te zien dat sommige nationaliteiten vooral aanwezig zijn op Amsterdam/Schiphol en de grote steden (figuur 4.2 links, Azië), terwijl voor roaming-verkeer van toestellen uit België (figuur 4.2 rechts) duidelijk het effect van grensverkeer zichtbaar is.



Figuur 4.2: Mate van roaming-gebruik voor abonnees afkomstig uit Azië (links) resp. België (rechts).

Roaming-data uitgezet tegen de tijd laat voor verschillende nationaliteiten andere patronen zien. Roamingactiviteit in de omgeving Amsterdam/Schiphol voor toestellen uit Azië laat eind januari een verdubbeling zien. Een mogelijke verklaring hiervoor is het Chinees nieuwjaar. Roaming voor Oost-Europese GSM's laat een ander patroon zien, met pieken rond de nieuwjaarsperiode en eind februari voor het gebied rond Amsterdam/Schiphol, terwijl voor Rotterdam juist een conventioneel week/weekeinde-patroon te zien is.

De figuren in 4.2 en 4.3 geven aan dat roamingdata een veel gedetailleerder beeld – zowel in tijd als in ruimte -van toeristische activiteit kan geven dan conventionele statistieken doen. Nader onderzoek is noodzakelijk om te zien in hoeverre deze data gebruikt kan worden om de bestaande toerismestatistieken te verrijken, en welke methodologische problemen daarvoor moeten worden opgelost.



Figuur 4.3: Roamingdata voor verschillende nationaliteiten uitgezet tegen de tijd.

4.3 Schattingen daytime populatie

Met behulp van de door Mezuro geleverde aantallen actieve toestellen per vier uur kan een schatting gemaakt worden van de daytime populatie. Deze paragraaf schrijft het algoritme om dit te doen, de resultaten en een discussie waarin onder andere de gemaakte aannames expliciet worden gemaakt.

4.3.1 Algoritme

Het algoritme bestaat uit drie stappen:

- i. Corrigeren voor regionale onder- of overdekking
- ii. Bepalen relatieve aantallen
- iii. Schatten van daytime populatie

In stap *i* worden de aantallen toestellen per plaats gecorrigeerd voor onder- of overdekking van Vodafone klanten binnen die plaats ten opzichte van het landelijke gemiddelde. Deze dekkingen zijn afkomstig van het klantenbestand, dat in de vorige paragraaf is besproken. Het idee achter deze correctie is dat een bepaald aantal actieve toestellen in een plaats met een lage Vodafone dekking meer mensen vertegenwoordigd als hetzelfde aantal actieve toestellen in een plaats met een hoge Vodafone dekking.

In stap *ii* wordt de (gecorrigeerde) aantallen actieve toestellen gedurende de dag gerelativeerd aan de aantallen actieve toestellen gedurende de referentieperiodes: van 20 uur 's avonds tot 04.00 uur 's nachts. Dit wordt zowel op plaats-niveau als op landelijk niveau gedaan.

In stap *iii* wordt de daytime populatie geschat met behulp van de relatieve aantallen en het GBA. Het komt erop neer dat we de aantallen tijdens de referentieperiodes ijken met de GBA aantallen.

i) Corrigeren voor regionale onder- of overdekking

We beschikken over het aantal toestellen die actief zijn (bel, sms of data-event) per tijdsinterval van vier uur, per plaats. Deze aantallen zijn door Mezuro uitgesplitst naar inwoners r_{dp}^i , en bezoekers v_{dp}^i , waarbij i de plaats is, d de dag en p de vier-uurs periode.⁸ De perioden $p=1,2,\dots,6$ komen overeen met 00.00 tot 04.00 uur, 04.00 tot 08.00 uur, 08.00 tot 12.00 uur, 12.00 tot 16.00 uur, 16.00 tot 20.00 uur en 20.00 uur tot 00.00 uur. De Vodafone dekking in woonplaats i wordt genoteerd als z^i en de landelijke Vodafone dekking met z .

Stel dat plaats A een lage Vodafone dekking heeft van $z^A = 0,1$ en plaats B een hoge Vodafone dekking van $z^B = 0,5$. Als er op een bepaalde dag d en periode $p=4$ 1000 actieve toestellen van inwoners zijn in zowel A als B , met andere woorden,

⁸ De woonplaats van een gebruiker van een toestel is de plaats waar het toestel structureel actief is tussen 20.00 en 07.00 uur. Een gebruiker wordt geteld als een inwoner als zijn of haar toestel in zijn of haar woonplaats actief is, en een bezoeker als het in een andere plaats actief is.

worden $r_{d4}^A = r_{d4}^B = 1000$, dan zal het werkelijke aantal inwoners in plaats A ongeveer 5 maal zo groot zijn als in plaats B. De dekkingen z kunnen worden gezien als insluitkansen, waarbij de inverse als een gewicht kan worden gezien.

Dit voorbeeld kunnen we veralgemeniseren door per plaats i de landelijke Vodafone dekking te delen door de dekking in plaats i :

$$x_{dp}^i = r_{dp}^i \frac{z}{z^i} + v_{dp}^i,$$

voor iedere dag d en periode p . Deze stap kan worden gezien als een weging, waarbij de aantallen toestellen van inwoners gewogen wordt naar Vodafone dekking. De uitkomstwaarden van x kunnen worden geïnterpreteerd als de aantallen actieve toestellen wanneer de Vodafone dekkingen per plaats constant zouden zijn. Onder de bezoekers van een bepaalde plaats is niet bekend welke Vodafone aandeel ze hebben. Daarom wordt verondersteld dat dit gelijk is aan de landelijke dekking z .

ii) Bepalen relatieve aantallen

De waarden x kunnen nog niet direct worden gebruikt om de daytime populatie te schatten. Het zijn namelijk aantallen actieve toestellen. Het is onduidelijk in hoeverre toestellen meer overdag worden gebruikt dan 's avonds en 's nachts. Wel kunnen de plaatselijke dagpatronen vergeleken worden met het landelijke dagpatroon.

Per plaats wordt het aantal actieve toestellen voor iedere periode overdag bepaald ten opzichte van het aantal actieve toestellen 's avonds ($p=6$, dit is van 20.00 tot 00.00 uur) en 's nachts ($p=1$, dit is van 00.00 tot 04.00 uur). Dit wordt ook gedaan voor de landelijke aantallen toestellen. We nemen hierbij aan dat de populatie van plaats i tussen 20.00 uur 's avonds en 04.00 uur 's nachts gelijk is met het GBA aantal y_{GBA}^i . Om de aantallen toestellen in dit tijdsinterval, dat bestaat uit twee vier-uurs perioden ($p=6$ en $p=1$), te ijken met het GBA, zullen deze worden gemiddeld:

$$x_{d6}^i = x_{(d+1)1}^i = \frac{x_{d6}^i + x_{(d+1)1}^i}{2} \text{ voor alle plaatsen } i \text{ en dagen } d.$$

Voor de landelijke aantallen actieve toestellen x_{dp} wordt hetzelfde gedaan:

$$x_{d6} = x_{(d+1)1} = \frac{x_{d6} + x_{(d+1)1}}{2} \text{ voor alle dagen } d.$$

Vervolgens kunnen de relatieve aantallen worden bepaald door de x -aantallen te ijken met de x -aantallen van ijkperioden 1 en 6:

$$a_{dp}^i = \frac{x_{dp}^i}{b_{dp}^i},$$

waarbij

$$b_{dp}^i = \frac{6-p}{5} x_{d1}^i + \frac{p-1}{5} x_{d6}^i,$$

voor alle plaatsen i , dagen d , en perioden p .

Hierbij kan b^i worden gezien als een soort baseline. In de ijkperioden $p=1$ en $p=6$ is b_{dp}^i gelijk aan x_{dp}^i , en in de overige perioden is b_{dp}^i een lineaire combinatie van de x -aantallen in de ijkperioden ervoor en erna.

De uitkomst waarde a_{dp}^i kan worden geïnterpreteerd als relatieve aantal toestellen in periode p ten opzichte van het aantal toestellen in de aangrenzende ijkperiode(n). Wanneer bijvoorbeeld $a_{d3}^i = 4$ in plaats i voor een periode $p=3$ in de middag, dan betekent het dat het aantal actieve toestellen in die periode 4 maal zo hoog is als in de ijkperioden.

Het feit dat de relatieve aantallen in plaats i overdag hoger is dan in de ijkperioden wil natuurlijk niet zeggen dat de populatie overdag groter dan 's avonds en 's nachts. Overdag zal namelijk meer gebruik worden gemaakt van mobiele telefoons dan 's avonds en 's nachts. Wel kunnen we relatieve aantallen vergelijken met de relatieve aantallen op landelijk niveau:

$$a_{dp} = \frac{x_{dp}}{b_{dp}},$$

waarbij

$$b_{dp} = \frac{6-p}{5} x_{d1} + \frac{p-1}{5} x_{d6}$$

voor alle gemeenten i , dagen d , en perioden p .

iii) Schatten van daytime populatie

Door de plaatselijke relatieve aantallen te delen door de landelijke relatieve aantallen, kan een schatting worden gemaakt van de daytime populatie. Voortbordurend op het bovenstaande voorbeeld waarbij $a_{d3}^i = 4$ voor een bepaalde plaats i in periode $p=3$: stel dat landelijk gezien het aantal actieve toestellen in dezelfde periode 3 maal zo hoog als in de ijkperioden, met andere woorden $a_{d3} = 3$. Dan kunnen we, met de gemaakte aannames, concluderen dat woonplaats i overdag $4/3$ maal zoveel mensen zijn als tijdens de ijkperiode. Dit leidt tot de volgende schatting, aangenomen dat de populatie tijdens de ijkperioden gelijk is aan de GBA:

$$\hat{y}_{dp}^i = y_{GBA}^i \frac{a_{dp}^i}{a_{dp}}.$$

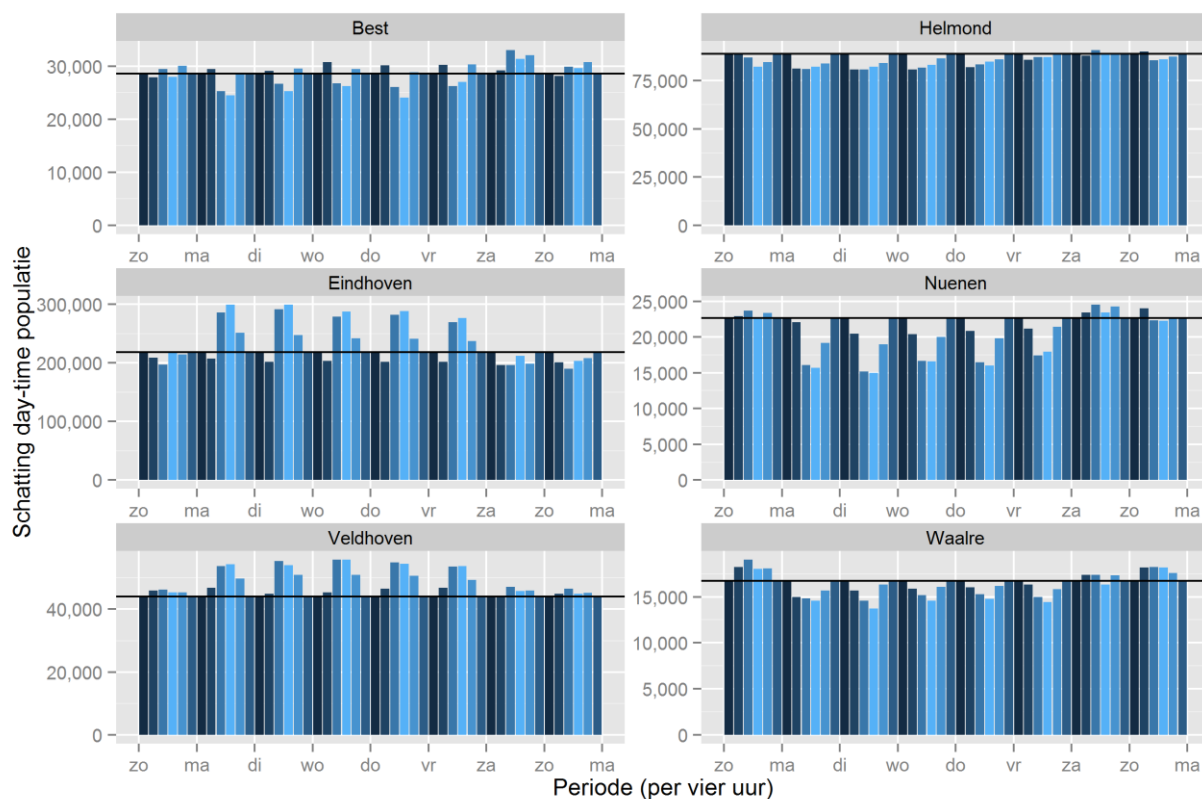
Deze schatting is slechts een puntschatting. Het is in dit stadium van het onderzoek nog niet mogelijk om de bijbehorende varianties te schatten. De nauwkeurigheid van deze schatting is afhankelijk van de kwaliteit onderliggende data, van het plaats-toekenningsalgoritme, en van de gemaakte aannames. Dit zal in paragraaf 4.1.3. nader worden toegelicht.

4.3.2 Resultaten schattingen daytime populatie

Het algoritme is toegepast op de maand januari. Deze maand kent, behalve rond de jaarwisseling, geen bijzonderheden zoals feestdagen of vakantieperiodes. De jaarwisseling zelf, dus periode $p=1$ (00.00 – 04.00 uur) op 1 januari veroorzaakt zeer waarschijnlijk een vertekening van de schattingen van de daytime populaties op 1 januari, omdat deze periode in het algoritme gebruikt wordt als referentieperiode. De aanname dat men tijdens die periode in zijn eigen woonplaats was zal lang niet altijd kloppen.

Als casus is gekeken naar een beschrijving van de resultaten op een normale week zondag 6 januari tot en met zondag 13 januari naar het gebied van Eindhoven en omringende plaatsen. Eindhoven staat bekend als een typische werkstad, waar veel technologische bedrijven gevestigd zijn. Ook is het een belangrijk verkeersknooppunt. Dit houdt in dat er overdag meer mensen te verwachten zijn dan 's avonds. De meeste omringende plaatsen staan te boek als forenzen plaatsen. Alleen in Veldhoven is ook een groot bedrijf gevestigd waar overdag veel mensen werken.

In Figuur 4.4 is de daytime populatie weergegeven van Eindhoven en omstreken. De staven geven de populaties aan voor het betreffende 4-uursblokken. De kleuren van de staven geven de hoeveelheid daglicht aan. De lichtste staven corresponderen met de periode van 12.00 tot 16.00 uur, en de donkerste staven met de periode 00.00 tot 04.00 uur. Deze periode en de periode ervoor (20.00 tot 00.00 uur) zijn geijkt met de GBA aantallen, die zijn weergegeven door horizontale lijnen



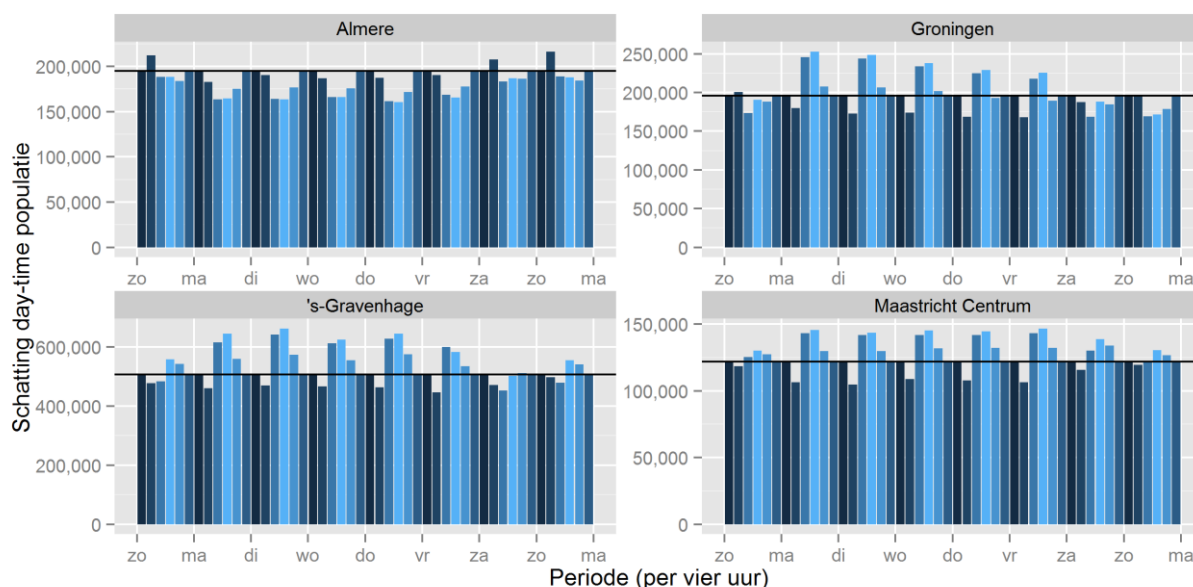
Figuur 4.4: Schatting van de day-time populatie in Eindhoven e.o. van zondag 6 januari t/m zondag 13 januari. Horizontale lijnen zijn GBA aantallen.

In Eindhoven is de geschatte daytime populatie doordeweeks een stuk hoger dan het GBA aantal. Maandags en dinsdags wordt deze rond de 300000 mensen geschat. Overigens liggen de staven van 04.00 tot 08.00 uur doordeweeks structureel iets onder de GBA. Dit duidt waarschijnlijk een vertekening die door het algoritme wordt geïntroduceerd. In het weekend is de daytime populatie iets lager dan de GBA aantallen. De koopzondag van 6 januari springt er nauwelijks uit (zie de meest linkse lichtste staaf, en ter referentie, de meest rechtse staaf).

Ook in Veldhoven is de daytime populatie doordeweeks hoger dan het GBA aantal. Het verschil is ongeveer 10000 mensen, dat deels te verklaren is doordat in de ASML vestiging in Veldhoven circa 6000 mensen werken. Ook het Maxima Medisch Centrum zal veel mensen van buiten Veldhoven aantrekken.

In de overige plaatsen is de geschatte daytime populatie lager dan de GBA, met name in Nuenen. Dit bevestigt het beeld dat deze plaatsen forenzenplaatsen zijn. Tijdens de zondagen zijn er meer mensen in Waalre dan het GBA aantal. Een oorzaak hiervan kan zijn dat beide zondagen koopzondagen waren. Verder valt op dat de daytime populatie in Best op zaterdag hoger is dan het GBA aantal. Een andere, algemene oorzaak van het feit dat de daytime populatie in de omliggende plaatsen tijdens het weekend even groot of groter is dan de GBA, is dat veel studenten uit Eindhoven tijdens het weekend overdags bij hun (schoon)ouders op bezoek gaan.

In Figuur 4.5 is de geschatte daytime populatie weergegeven voor vier andere steden. Ondanks dat Almere de laatste jaren flink aan de weg aan het timmeren is om meer bedrijven naar zich toe te trekken (onder andere middels het project Almere DataCapital), ontpopt het zich toch als forenzenstad. Groningen, Den Haag en Maastricht zijn wel typische werksteden, doordat de geschatte daytime populatie doordeweeks hoger ligt dan de GBA aantallen.



Figuur 4.5: Schatting van de daytime populatie in vier steden van zondag 6 januari t/m zondag 13 januari. Horizontale lijnen zijn GBA aantallen.

4.3.3 Discussie

De beschreven schattingsmethode bevindt zich nog in een verkennend stadium, en kan op diverse vlakken worden verbeterd. Ook de kwaliteit van de aangeleverde geaggregeerde gegevens kan nog worden verbeterd. Met name het bepalen van de locatie van een bel- sms- of data-event blijkt in de praktijk erg lastig te zijn.

Bij het ontwikkelen van het beschreven algoritme zijn de volgende aannames gemaakt:

- Mensen bevinden zich tussen 20.00 uur 's avonds en 04.00 uur 's nachts in hun eigen woonplaats. Deze aanname zal in de praktijk niet altijd opgaan, bijvoorbeeld met betrekking tot het uitgaansleven en tot avondevenementen zoals voetbalwedstrijden.
- Het percentage inwoners dat gebruik maakt van smartphones is per plaats hetzelfde. Deze aanname is belangrijk, omdat smartphones 's nachts veel data-events genereren, terwijl het aantal telefoongespreken en sms'jes 's nachts normaal gesproken erg weinig is. Als bijvoorbeeld in een plaats relatief veel smartphones worden gebruikt, zal de daytime populatie te laag worden geschat.
- Het aandeel Vodafone klanten onder bezoekers van een plaats is gelijk aan het landelijke aandeel Vodafone klanten. Zoals vermeld in paragraaf 4.1.1 is het aandeel Vodafone klanten onder bezoekers niet bekend. Daarom is het landelijke aandeel een goed alternatief.
- Mensen met een zakelijk Vodafone abonnement wonen in dezelfde plaats als de plaats van het postcodegebied van het bedrijf waar ze werken. Deze aanname zal in de praktijk vaak worden geschonden. Mensen wonen namelijk vaak in omliggende plaatsen, en bovendien is het postcodegebied niet altijd de locatie van de vestiging waar men werkt. De consequentie van deze aanname is dat de Vodafone dekking per woonplaats, de z^i -tjes, niet helemaal correct is, wat invloed heeft op de correctiemethode (stap ii).
- Het bel- sms en datagebruik-gedrag onder Vodafone klanten is gelijk aan dat van andere providers. Wat de consequenties zijn van deze aanname is niet eenduidig te verwoorden. Vanzelfsprekend zullen de schattingen nauwkeuriger worden als gegevens beschikbaar komen van meerdere providers.

Er zullen weinig geschikte bronnen zijn om de schattingen mee te vergelijken. Wel kan het aantal werkzame personen per plaats, dat beschikbaar is in de polisadministratie dienen als hulpinformatie. Uit deze bron kunnen ook geaggregeerde tellingen worden gemaakt van een vergelijking van woon- en werkplaats op persoonsniveau.

In de nabije toekomst zal het mogelijk zijn om tabellen te gebruiken met een gedetailleerdere regionale indeling dan gemeente of plaats, bijvoorbeeld op postcode, buurt of wijkniveau. In dat geval zouden kleinedomeinschatters (Rao, 2003) kunnen helpen om de schattingen nauwkeuriger en robuuster te maken.

Hiervoor zijn per regio samenhangende achtergrondkenmerken nodig, zoals stedelijkheid en demografische kenmerken van de inwoners.

5. Conclusie

Mobiele telefoniedata vormen een ‘Big-data’-bron die zeer veel informatie bevat. Het zijn nieuwe data die uniek zijn in verschillende opzichten. Het zijn gegevens die door apparaten gegenereerd worden, maar mensen vertegenwoordigen. Zo kunnen uitspraken worden gedaan over bezoekersaantallen van evenementen, aantallen buitenlandse toeristen, en aantallen personen in een bepaalde plaats overdag (daytime population). Omdat het niet mogelijk is om de mastgegevens direct te koppelen aan personen en hun achtergrondkenmerken is het lastig om uitspraken te doen over representativiteit. Door te kijken naar een andere databron, een klantenbestand, kan hier toch inzicht in verkregen worden. Vanzelfsprekend ontbreken jonge kinderen in de dataset die nog geen mobiele telefoon hebben. Tieners zijn ondervertegenwoordigd, maar dat komt waarschijnlijk omdat de abonnementen op één van hun ouders staan. Globaal gezien zijn volwassenen tussen de 20 en de 50 oververtegenwoordigd, en ouderen ondervertegenwoordigd.

De verkennende analyses op roamingdata (toerisme) geven een goede indruk van tellingen van toeristen die op het eerste oog plausibel zijn. Een belangrijke beperking is echter dat toeristen slechts 1 dag geteld kunnen worden vanwege privacyrichtlijnen. De resultaten laten interessante fenomenen zien. De kaartjes zullen door inhoudelijk experts verder worden onderzocht. In een vervolg onderzoek wordt gekeken hoe deze data kunnen leiden tot concrete toepassingen voor toerismestatistieken.

De resultaten in deze studie laten zien dat het mogelijk is om tellingen van personen te maken. Het nauwkeurig schatten van de daytime population (aantal personen overdag) vormt een rode draad in het onderzoek naar het gebruik van mobiele telefoniedata ten behoeve van statistiekproductie. De huidige schattingen zijn een eerste verkenning. Verder onderzoek is gericht op de nauwkeurigheid en validiteit van de schattingen. Ook zijn er al ideeën om de schattingen nog beter te maken door bijvoorbeeld gebruik te maken van kleinedomeinschatters.

Op basis van de resultaten van dit project is besloten om het onderzoek verder voort te zetten. Het betreft een zeer rijke dataset waarbij het CBS dit jaar verder zal gaan met onderzoek naar toepassingen voor nieuwe of bestaande statistieken.

6. Bibliografie

Ahas, A., Roose, A., Mark, Ü. & Silm, S., 2008. Evaluating Passive Mobile Positioning Data for Tourism Surveys: An Estonian Case Study. *Tourism Management*, pp. 29-469-486..

Eagle, N., Pentland, A. (. & Lazer, D., 2009. Inferring friendship network structure by using mobile phone data. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, p. 15274–15278.

Jonge, E. d., Pelt, v. M. & Roos, M., 2012. Time patterns, geospatial clustering and mobility statistics based on mobile phone network data. *CBS - FCSM.gov*.

Kwok, R., 2009. News Feature Phoning in data. *Nature*, 23 April, p. Vol 458.

Palchykov, V. et al., 2012. *Sex differences in intimate relationships*. Oxford: Nature.

Rao, J. N. K., 2003. *Small Area Estimation*. sl:John Wiley, New York..

Raun, J. & Ahas, R., 2013. *Distinguishing tourism destinations with behavioural data*. Brussel, New Techniques and Technologies for Statistics, NTTS.

Roos, M., Pelt, M. v., Jonge, E. & Tennekes, M., 2010. *Using Mobile phone data for statistiscs. Discussion paper*. Heerlen/Den Haag: CBS (in preparation).