Waar liggen kansen voor OV: datafusie GSM en chipkaart

Karin de Regt – Goudappel Coffeng – KdRegt@goudappel.nl
Niels van Oort – Goudappel Coffeng / TU Delft– NvOort@goudappel.nl
Oded Cats – TU Delft– O.Cats@tudelft.nl

Bijdrage aan het Colloquium Vervoersplanologisch Speurwerk
24 en 25 november 2016, Zwolle

Samenvatting

De grootste uitdaging van de openbaar vervoer sector is om tegemoet te komen aan de verscheidenheid aan reispatronen, en de bijbehorende behoeften en preferenties, van reizigers. Het beter matchen van vraag en aanbod levert zowel een kwaliteitssprong als kostenreductie op en heeft dus alle aandacht. Bestaande databronnen helpen, maar zijn nog niet afdoende. De combinatie van nieuwe bronnen biedt echter hoopgevende resultaten. Door een innovatieve methodiek kunnen GSM- en anonieme chipkaartdata gecombineerd worden om de OV potentie in kaart te brengen.
Bestaande onderzoeken (zoals OViN) geven informatie over de totale reisbehoefte en de ruimtelijke spreiding hiervan. Deze huishoudsurveys bieden veelal echter geen inzicht in de spreiding van deze reisbehoefte over de tijd. Een nieuwe methodiek om GSM- met anonieme OV chipkaartdata te koppelen, geeft die inzichten wel. Door middel van deze datafusie kunnen zowel de ruimtelijke als temporele patronen van OV gebruik vergeleken worden met de totale ruimtelijke en temporele reispatronen. Dit geeft inzicht in de (mis)match van vraag en aanbod in ruimte én tijd. Ideaal dus als eerste stap voor het verbeteren van deze match: OV potentie kan zo worden opgespoord.
Deze methode is toegepast in een case study in Rotterdam om te onderzoeken of het huidige OV bedieningsinterval voldoende aansluit bij de latente vraag. De resultaten demonstreren dat er potentie is om het OV bedieningsinterval zowel in de vroege ochtend als in de late avond uit te breiden. In de vroege ochtend, juist voordat het OV wordt opgestart, kan een uur-tot-uur toename in bezoekersaantallen van 33% tot zelfs 88% worden waargenomen in diverse delen van de Rotterdamse regio. Dit illustreert de potentiële vraag voor extra openbaar vervoer aanbod in de vroege ochtend. Op vergelijkbare wijze is deze analyse uitgevoerd voor het OV aanbod in de late avond.
Deze innovatieve methode van datafusie is van grote toegevoegde waarde te zijn gebleken ter ondersteuning van OV planning. Deze datafusie methode kan ook eenvoudig worden toegepast op andere herkomst-bestemmingsdata.

1. Introductie

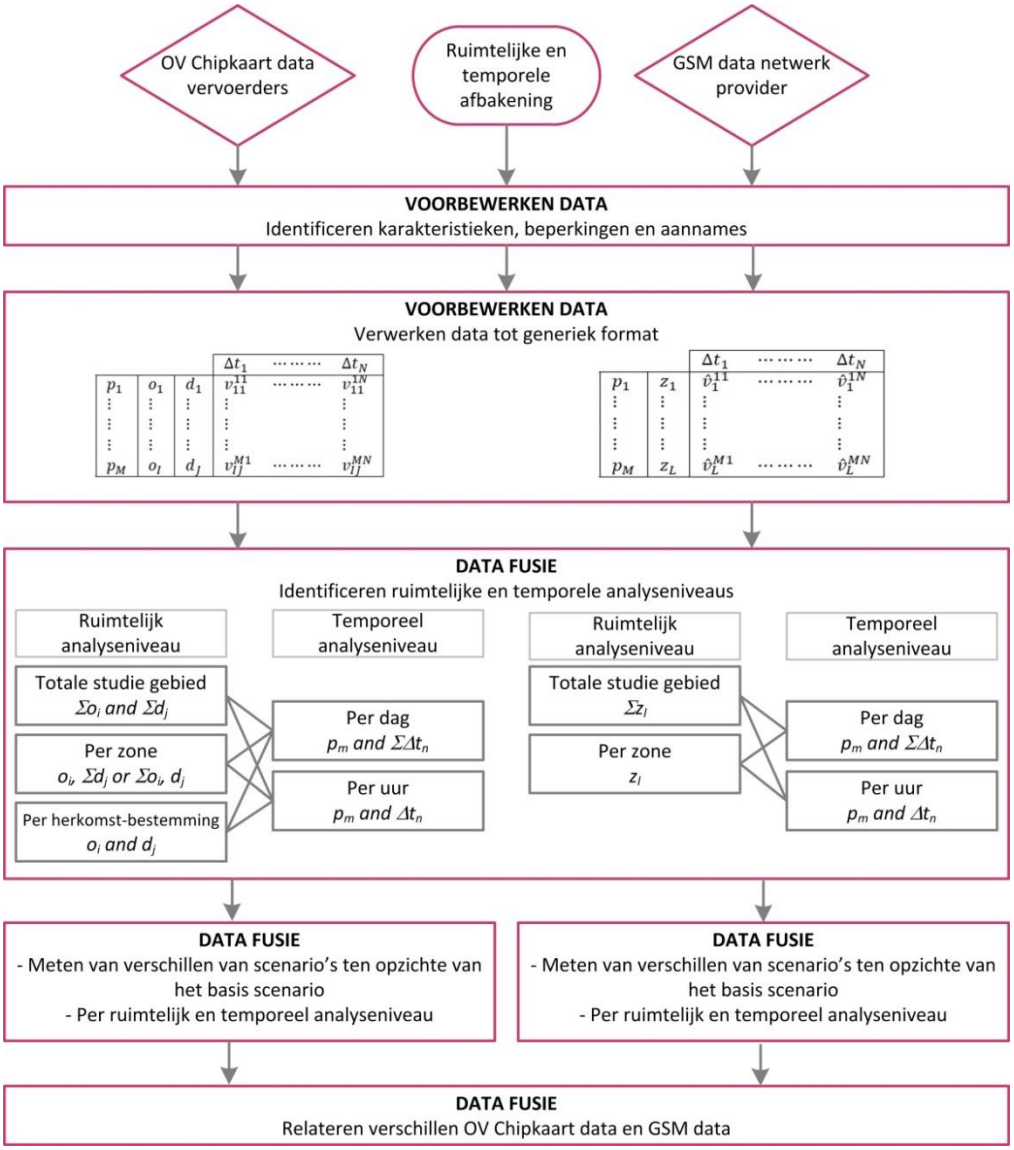
Zowel reizigers als overheden vragen om een efficiënt openbaar vervoer (OV) systeem met hoge kwaliteit tegen lage kosten. Dit OV systeem moet georiënteerd zijn op de reiziger, en daarmee tegemoet komen aan de behoeften en preferenties van reizigers *(Guedes et al. 2012)*. Echter, niet alle reizigers hebben dezelfde reispatronen met bijbehorende behoeften en voorkeuren. Er is niet alleen sprake van ruimtelijke variatie in vervoersvraag, maar ook van een spreiding in tijd, wat resulteert in een dynamische omgeving *(Cats et al. 2015; Gutierrez & Garcia-Palomares 2007)*. Om een OV netwerk te ontwerpen binnen deze dynamische omgeving, wordt veelal gebruik gemaakt van anonieme OV Chipkaart data om reispatronen te analyseren *(Oort et al. 2015b)*. OV Chipkaart data geeft echter alleen informatie over de OV vervoersvraag en neemt de totale vervoersvraag niet in beschouwing. Traditioneel worden zogenaamde huishoud surveys gebruikt voor dataverzameling om de totale vervoersvraag te schatten en analyseren *(Durand et al. 2016; Long & Thill 2015)*. Huishoud surveys worden gebruikt om wensen en preferenties van reizigers te analyseren per vervoerwijze, reismotief en reisattribuut *(Durand et al. 2016; Del Castillo & Benitez 2012)*. Het verzamelen van data voor huishoud surveys is echter tijdsintensief en duur, voornamelijk als gevolg van het arbeidsintensieve proces voor het verkrijgen en verwerken van de data. Hierdoor worden deze surveys over een lange periode afgenomen, met als doel om reispatronen voor een gemiddelde (werk)dag te representeren *(Frias-Martinez et al. 2016)*. Omdat alleen een gemiddelde werkdag wordt gehanteerd, is het niet mogelijk om met deze surveys de dynamiek van reispatronen over tijd te onderscheiden. Dit vraagt derhalve om de ontwikkeling van methoden die inzicht geven in zowel de ruimtelijke als temporele dynamiek van reispatronen van OV reizigers in relatie tot de totale vervoervraag.

Behalve anonieme OV Chipkaart data en huishoud surveys worden diverse andere databronnen gebruikt om informatie te verkrijgen over reispatronen en om OV netwerken te verbeteren. Voorbeelden hiervan zijn automatische voertuig locatie systemen (GOVI), Wi-Fi, Bluetooth, social media en GSM data *(Van Oort et al. 2015a)*. De grootste uitdaging hierbij is om deze data dusdanig te verwerken dat het bruikbaar wordt om het ontwerp van OV netwerken te verbeteren. Hoewel GOVI data gebruikt wordt om de prestatie van de OV voertuigen te monitoren, geeft het geen informatie over de effectiviteit van het OV aanbod. Data afkomstig van Wi-Fi, Bluetooth en social media wordt pas recentelijk gebruikt in de transport sector. Deze databronnen bieden zeer gedetailleerde informatie over een kleine steekproef van de gehele populatie, en in geval van social media data is daarnaast een complexe semantische analyse nodig *(Van Oort et al. 2015a)*. Deze databronnen geven daarom geen informatie over de totale vervoersvraag, maar geven vooral complementaire informatie. GSM data wordt in toenemende mate gebruikt om reizigersvraag te analyseren. De mate waarin GSM data beschikbaar is, en op welk ruimtelijk en temporeel niveau de data wordt verstrekt, verschilt sterk per land. GSM data wordt afgeleid uit gedetailleerde bel-records die door een provider verstrekt worden *(Van der Mede 2014)*. De drie voornaamste toepassingen van GSM data in onderzoek in de transport sector zijn het schatten van Herkomst-Bestemming matrices, het detecteren van events op basis van drukte, en het identificeren van de gebruikte vervoerwijze *(Aguilera et al. 2014; Calabrese et al. 2013;*

Iqbal et al. 2014). Deze laatste toepassing is nog niet beschikbaar in Nederland. Daarom is het voor dit studiedoel niet voldoende om alleen gebruik te maken van GSM data.
Het combineren van databronnen, datafusie, biedt een veelbelovende route om informatie te verkrijgen over OV reispatronen in relatie tot de totale reispatronen en de spreiding in ruimte en tijd hiervan. Verschillende datafusie studies combineren ofwel OV Chipkaart data, ofwel GSM data met data van huishoud surveys om reismotieven te schatten *(Long & Thill 2015; Kusakabe & Asakura 2014)*. In Emmen is een pilot uitgevoerd, waarbij OV Chipkaart en GSM data gefuseerd zijn om gebieden te vinden met potentie voor extra OV aanbod *(Elfrink et al. 2015)*. Ook in een studie in Singapore is de combinatie van chipkaart en GSM data verkend om zwakke OV verbindingen te identificeren *(Holleczek et al. 2014)*. Met beide studies wordt de hypothese ondersteund dat datafusie van OV Chipkaart en GSM data synergie effecten biedt, welke resulteren in nieuwe inzichten *(Elfrink et al. 2015)*. OV Chipkaart data biedt informatie over OV reizigers die met een specifieke vervoerder reizen. GSM data biedt juist inzichten in de totale vervoersvraag (op basis van een zeer grote steekproef), met de bijbehorende spreiding in ruimte en tijd. Hoewel alle vervoerwijzen hierin worden meegenomen, is het niet mogelijk om onderscheid te maken tussen de verschillende vervoerwijzen. Derhalve bevatten beide databronnen informatie over ruimtelijke en temporele variaties.
Het doel van deze studie is om de potentie te verkennen van fusie van anonieme OV Chipkaart- en GSM data om informatie te verkrijgen over OV reispatronen in relatie tot de totale vervoersvraag, waarbij zowel spreiding in ruimte als tijd expliciet wordt meegenomen. Deze aanpak kan voor diverse doeleinden gebruikt worden. Diverse vervoerders exploiteren een speciaal nachtnet, waarbij de tijden waarop de transitie van dag-netwerk naar nacht-netwerk en andersom plaats vindt moeten worden vastgesteld *(Duff-Riddell & Bester 2005)*. Ook dient de vervoersvraag gedurende de nacht in beschouwing te worden genomen, zodat OV netwerk en OV bedieningsinterval overeenkomen met de vraag tijdens de nacht. We passen deze data fusie toe in een case study in Rotterdam, waarbij zowel OV gebruik als de totale vervoersvraag voor de late avond en vroege ochtend geanalyseerd is voor 5 verschillende dagsoorten. Het doel is om te identificeren in hoeverre het huidige aanbod van stedelijk OV voldoende aansluit bij de vervoersvraag in deze uren, en in welke mate dit verschilt over de verschillende dagsoorten. De resultaten van deze studie kunnen besluitvormers ondersteunen tijdens de evaluatie van het huidige netwerk en de huidige dienstregeling, en om potentiële verbeteringen hierin te identificeren.
De opbouw van deze paper is als volgt: het volgende hoofdstuk zet de voorgestelde datafusie methode uiteen. In hoofdstuk 3 wordt de methode toegepast in de case study betreffende nachtelijk OV aanbod in Rotterdam. Tenslotte worden in hoofdstuk 4 conclusies en aanbevelingen voor verdere toepassingen geformuleerd.

2. Methodologie fusie GSM en chipkaart data

Dit hoofdstuk geeft een overzicht van de ontwikkelde datafusie methodologie. Eerst wordt een overzicht van de volledige methode structuur gegeven, waarna de verschillende stappen in detail worden beschreven. De analyse methode is gevisualiseerd in Figuur 1.



Figuur 1: Flowchart van de ontwikkelde methodologie

De input voor de methodologie is geanonimiseerde OV Chipkaart en GSM data, met een bijbehorende afbakening in ruimte en tijd. Om OV reispatronen te analyseren in relatie tot de totale vervoersvraag worden scenario's ontwikkeld die vergeleken worden met een vooraf bepaald basis scenario. Het basis scenario voor zowel de OV Chipkaart als de GSM data wordt geconstrueerd afhankelijk van het doel van het onderzoek; hierbij kan worden gedacht aan een basis scenario dat een gemiddelde dag weergeeft, of een basis scenario met een dynamisch referentie punt (bijvoorbeeld het vorige uur). Om deze scenario's te construeren moet de input data worden voorbewerkt. Het voorbewerken van de data bestaat uit twee aspecten: identificeren van de karakteristieken, beperkingen en aannames per dataset, en het verwerken van de data tot een generiek format. Na het voorbewerken van de data, kan de volgende stap worden uitgevoerd. In de data fusie stap worden eerst verschillende ruimtelijke en temporale analyseniveaus geïdentificeerd met behulp van

differentiatie en aggregatie in ruimte en/of tijd. Per dataset en per analyseniveau worden de verschillen tussen de data van de scenario's ten opzichte van het basis scenario kwantitatief gemeten. De daadwerkelijke datafusie wordt vervolgens geconstrueerd door de resultaten van de OV Chipkaart per scenario te relateren aan de resultaten van de GSM data voor datzelfde scenario. Data fusie kan hiermee worden bewerkstelligd voor elk scenario en voor elk beschikbaar ruimtelijk en temporeel analyseniveau. De methodologie ontwikkeld in deze studie kan worden gebruikt voor verschillende datasets die informatie bevatten over herkomsten en bestemmingen in transportnetwerken.

2.1 Voorbewerken van de data

De OV Chipkaart data en de GSM data hebben verschillende karakteristieken, beperkingen en aannames. Beide datasets worden daarom afzonderlijk besproken. Meer informatie aangaande de gebruikte data en de verschillende formats is te vinden in *(De Regt 2016).*

OV Chipkaart data
De OV Chipkaart data die voor dit onderzoek is gebruikt is geanonimiseerd. In Nederland wordt er met de OV Chipkaart ingecheckt bij het aan boord gaan van een voertuig, en uitgecheckt wanneer een reiziger het voertuig verlaat. Dit leidt tot een registratie van herkomst i naar bestemming j, beide op halte niveau, inclusief een tijdsregistratie. De tijdsregistratie wordt geaggregeerd per dag m en per tijdsinterval n. Deze aggregatie leidt tot een passagiersvolume, v_{ij}^{mn}, die reizen tussen herkomst i en bestemming j op een specifieke dag m per tijdsinterval n.

GSM data
De GSM data voor dit onderzoek is beschikbaar gesteld door DAT.Mobility, die op hun beurt de GSM data ontvangen van een netwerk provider in Nederland (Vodafone), met een marktaandeel van ongeveer 33%. De ontvangen data was al volledig geanonimiseerd: individuen kunnen niet worden onderscheiden *(Elfrink et al. 2015).* De GSM data bevat informatie over het aantal toestellen van Vodafone in de ruimtelijke en temporele afbakening. Met behulp van een algoritme wordt de data opgehoogd tot de gehele populatie. De resultaten van dit algoritme zijn gevalideerd door zowel DAT.Mobility als het Centraal Bureau voor de Statistiek, waarbij geconcludeerd is dat de resultaten betrouwbaar zijn *(ViewDAT 2015).*

Elke keer dat een toestel verbinding maakt met een netwerk, wordt het gedetecteerd en geregistreerd in de database. Elk device dat aan staat, maakt ten minste 20 keer per dag verbinding met het netwerk, ook als het niet actief wordt gebruikt. Wordt een toestel wel actief gebruikt, wordt deze ook vaker gedetecteerd en daardoor geregistreerd. De locatie wordt per device bepaald, op basis van de antenne waarmee verbinding wordt gemaakt. Antennes hebben echter een bereik in meerdere richtingen en ook het bereik van meerdere antennes kan overlappen *(Iqbal et al. 2014)*. Hierdoor treedt er een zogenaamde lokalisatie-fout op, wanneer de exacte locatie van een device wordt geschat *(Calabrese et al. 2013)*. In verband met de betrouwbaarheid van de data, zijn ruimtelijk zones gedefinieerd, waar

devices aan worden gelinkt. Zones gedefinieerd in de GSM data zijn groter dan het halte niveau in de OV Chipkaart data. Echter, de grootte van de zones in de GSM data kan sterk variëren; deze worden gedefinieerd op basis van een of meerdere postcode zones. Er zijn bijvoorbeeld zones van 6km^2, maar er zijn ook zones van 30km^2.

De GSM data beschikbaar voor dit onderzoek is geaggregeerd in tijd in vooraf gedefinieerde tijdsintervallen. Het lokalisatie algoritme maakt hierbij een schifting naar unieke devices per tijdsinterval. Als een device is gedetecteerd in meerdere zones in een enkel tijdsinterval, wordt het device verbonden aan de zone waar deze het langste is gedetecteerd gedurende dat tijdsinterval. Daarnaast wordt er onderscheid gemaakt tussen inwoners en bezoekers. Om te bepalen of een device bij een inwoner of bezoeker hoort, wordt de woonplaats van een device bepaald op basis van nachtelijke verbinding. De zone waarin de device gedurende een merendeel van de nachten per maand wordt gedetecteerd, wordt geregistreerd als de woonplaats van de device. De woonplaats wordt maandelijks gedefinieerd, aangezien de data per maand is versleuteld. Als een device wordt gedetecteerd in zijn woonplaats, wordt een registratie van een inwoner gemaakt; wordt de device in elke andere zone gedetecteerd, wordt deze geregistreerd als bezoeker. Door het geografische aggregatie niveau, is het niet mogelijk om onderscheid te maken tussen een inwoner die thuis is gebleven, of zich heeft bewogen binnen zijn woonplaats zone. Bezoekers daarentegen, hebben zich verplaatst van hun woonplaats naar een andere zone, en daarmee een vraag naar transport laten zien. Vanwege het doel van deze studie, worden daarom alleen de bezoekers meegenomen in de GSM data.

GSM drukte data bevat informatie over $\hat{v}_l^{mn}$, het aantal bezoekers gedetecteerd in een zone $l \in L$ gedurende dag m en tijdsinterval n. L is de set van zones gedefinieerd binnen de ruimtelijke afbakening van de case study. De woonplaats van de bezoekers wordt niet meegenomen; het is onbekend waar de bezoekers vandaan komen. Daarnaast is het verschil tussen twee opeenvolgende uren een netto verandering in de drukte per zone: de aankomst-vertrek ratio kan hier niet uit worden afgeleid. De vraag naar transport volgend uit de GSM data wordt onderzocht aan de hand van de netto verandering in bezoekers per tijdsinterval. De absolute vraag naar transport kan niet worden afgeleid uit de data.

2.2 Datafusie

Analyseniveaus in tijd en ruimte

Om consistentie in de data te waarborgen, is de OV Chipkaart data geaggregeerd tot de ruimtelijke zone afbakening in de GSM data. Voor elke zone, OV Chipkaart transacties van haltes gelegen in die zone worden gesommeerd. Verschillende ruimtelijke en temporele analyseniveaus kunnen vervolgens worden onderscheiden door aggregatie en differentiatie van zowel ruimtelijke als temporele aspecten. Scenario's kunnen geanalyseerd worden per analyseniveau. De ruimtelijke analyseniveaus die kunnen worden onderscheiden zijn het hele case study gebied, per zone of per herkomst-bestemming relatie. Herkomst-bestemming relaties worden alleen onderscheiden in de OV Chipkaart data, en is hierdoor niet mogelijk in de data fusie. Temporele analyseniveaus zijn per dag of per uur. Een

combinatie van de verschillende ruimtelijke en temporele analyseniveaus leidt tot de volgende vier combinaties: totaal per dag, totaal per uur, per zone per dag en per zone per uur. Het analyseniveau totaal per dag geeft hierbij high-level informatie over de data, waarbij elk van de volgende analyseniveaus een gedetailleerder ruimtelijk danwel temporeel analyseniveau meeneemt. Deze zogenaamde top-down benadering is vaak gebruikt om mobiliteitspatronen in (openbaar) vervoer te analyseren *(Elfrink et al. 2015; Liu et al. 2009; Nishiuchi et al. 2013)*.

Verschillen meten per dataset

Per dataset en per analyseniveau worden de verschillen gemeten van een scenario in verhouding tot een basis scenario. Genormaliseerde verschillen worden berekend, zodat de resultaten van de twee verschillende datasets aan elkaar gerelateerd kunnen worden. Niet alleen de normalisatie van de verschillen is van belang, de grootte en richting van de verschillen spelen ook een belangrijke rol. Om deze drie aspecten mee te nemen is er besloten om de Mean Percentage Error (MPE) te gebruiken om verschillen te kwantificeren. De formules voor de MPE verschillen per analyseniveau, evenals de waarden die worden meegenomen in de OV Chipkaart en GSM data set (respectievelijk v_{ij}^{mn} en $\hat{v}_l^{mn}$). Voor de OV Chipkaart data kan er binnen het analyseniveau per zone per uur, onderscheid worden gemaakt tussen aankomsten en vertrekken per zone. Vergelijkingen (1)-(3) geven de MPE definities voor de OV Chipkaart data weer, waarbij vergelijkingen (4)-(5) de MPE definities voor de GSM data weergeven. Voor beide datasets geldt dat alleen het temporele analyseniveau per uur is meegenomen.

$$Totaal\ per\ uur\ MPE_{OV,n} = \frac{1}{I \cdot J} \cdot \frac{\left(\sum_i \sum_j v_{ij}^{[scenario]n} - \sum_i \sum_j v_{ij}^{[basis]n}\right)}{\sum_i \sum_j v_{ij}^{[basis]n}} \quad (1)$$

$$Per\ zone\ per\ uur\ MPE_{OV,jn} = \frac{1}{I} \cdot \frac{\left(\sum_i v_{ij}^{[scenario]n} - \sum_i v_{ij}^{[basis]n}\right)}{\sum_i v_{ij}^{[basis]n}} \quad (2)$$

$$Per\ zone\ per\ uur\ MPE_{OV,in} = \frac{1}{J} \cdot \frac{\left(\sum_j v_{ij}^{[scenario]n} - \sum_j v_{ij}^{[basis]n}\right)}{\sum_j v_{ij}^{[basis]n}} \quad (3)$$

$$Totaal\ per\ uur\ MPE_{GSM,n} = \frac{1}{L} \cdot \frac{\left(\sum_l \hat{v}_l^{[scenario]n} - \sum_l \hat{v}_l^{[basis]n}\right)}{\sum_l \hat{v}_l^{[basis]n}} \quad (4)$$

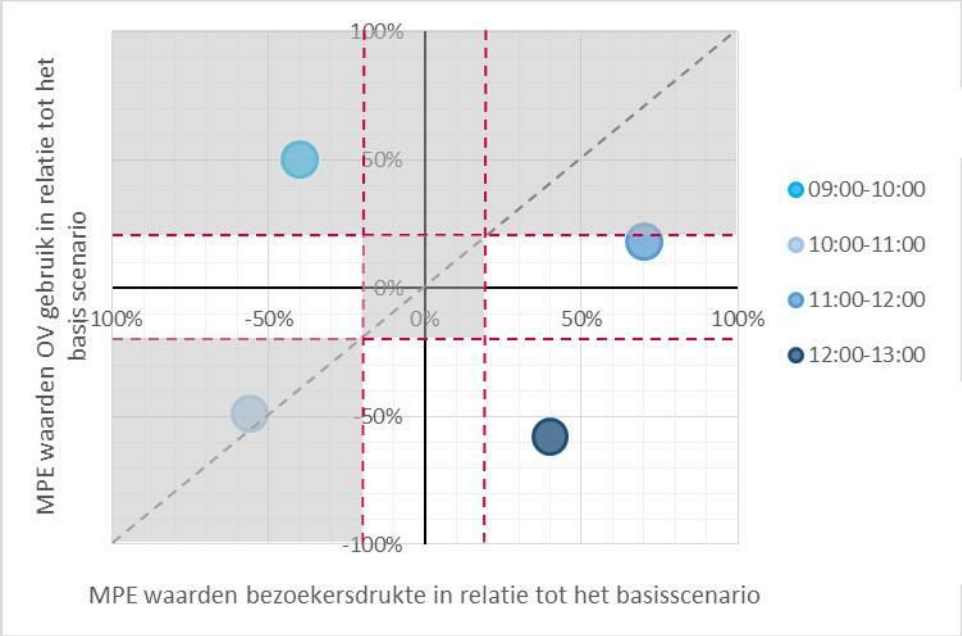
$$Per\ zone\ per\ uur\ MPE_{GSM,ln} = \frac{\left(\hat{v}_l^{[scenario]n} - \hat{v}_l^{[basis]n}\right)}{\hat{v}_l^{[basis]n}} \quad (5)$$

De MPE waarden bevinden alleen in de het bereik $[-1,\infty)$. Als de MPE waarde zich in het bereik $[-0.2,0.2]$ bevindt, dan wordt de geanalyseerde waarde niet significant verschillend van het basis scenario bevonden.

Relateren gemeten verschillen OV Chipkaart data en GSM data

De laatste stap in het data fusie proces is het relateren van de berekende OV Chipkaart waarden voor een bepaald scenario op een specifiek analyseniveau aan de berekende GSM data in diezelfde situatie. De MPE waarden van beide datasets worden aan elkaar gerelateerd met behulp van een grafiek waarbij de MPE waarden van de datasets worden

uitgezet op de assen, zoals gevisualiseerd in Figuur 2. De grenswaarden worden weergeven met behulp van roze stippellijnen. Als de stippen zich op of nabij de grijze stippellijn bevinden, betekent dit dat de relatieve MPE waarden van OV gebruik in dezelfde orde grootte zijn als de bezoekersdrukte. De niet-gearceerde gebieden zijn het meest interessant voor OV vervoerders. In het tijdsinterval 11:00-12:00 bijvoorbeeld, is de bezoekersdrukte significant gestegen ten opzichte van het basisscenario, terwijl het OV gebruik significant is gedaald ten opzichte van het basisscenario in diezelfde periode. Voor een vervoerder is het zeker van belang om te onderzoeken waarom OV gebruik is gedaald terwijl de algehele vraag naar transport is gestegen voor dit gebied en tijdsinterval.



Figuur 2: Illustratie van het relateren van de MPE van OV gebruik en bezoekersdrukte voor een bepaald gebied in een specifieke tijdsperiode wanneer deze vergeleken wordt met een basisscenario.

3. Case study: dagranden in Rotterdam

3.1 Case studybeschrijving

Wij hebben onze methodologie toegepast op twee case studies: (a) speciale evenementen (bijvoorbeeld festivals en verstoringen) in Amsterdam met bijbehorende mobiliteitspatronen en OV reispatronen; (b) het OV bedieningsinterval in de vroege ochtend en late avond in Rotterdam. Alleen de laatste wordt behandeld in verband met de ruimte limiet. Een gedetailleerde beschrijving van de case study in Amsterdam inclusief de resultaten is beschikbaar in *(De Regt 2016).*

Rotterdam is de op een na grootste stad in Nederland, met ongeveer 600.000 inwoners. RET is de OV vervoerder in de stad en de omgeving, en exploiteert zowel bus, tram en metro.

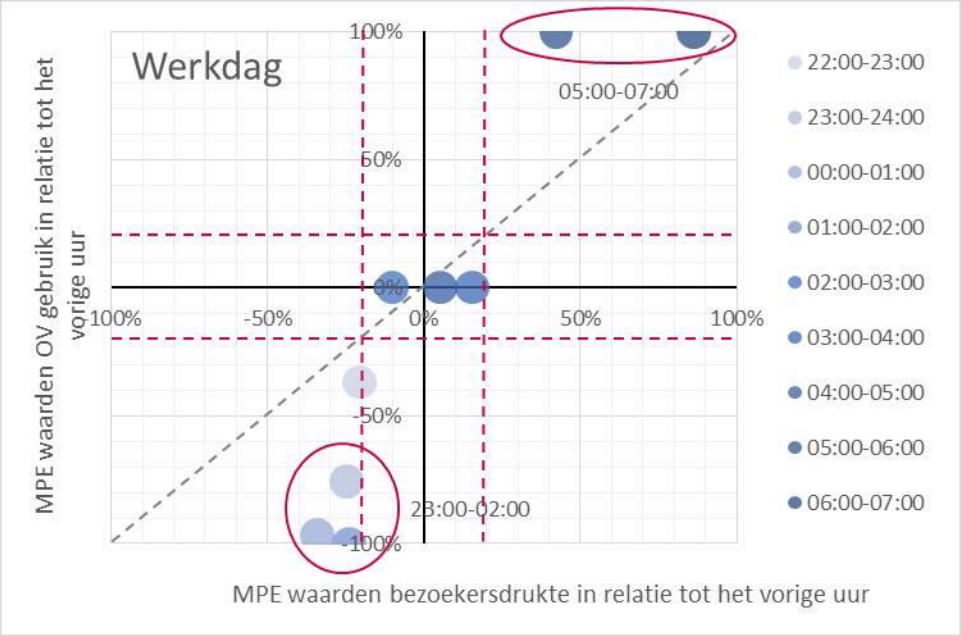
Jaarlijks worden er ongeveer 160 miljoen passagiersritten uitgevoerd door RET *(RET 2016).* Het case study gebied beslaat 34 zones, gebaseerd op de beschikbaarheid van stedelijk OV in de late avond en vroege ochtend (Figuur 3).

Figuur 3: Ruimtelijke afbakening van de case study in Rotterdam

Het doel van de case study is om te onderzoeken of het huidige OV bedieningsinterval voldoende aansluit bij de latente vraag; of het OV aanbod in de late avond en vroege ochtend overeenkomt met een respectievelijke daling danwel stijging in de algehele vraag naar vervoer. Hieruit kan bijvoorbeeld worden geconcludeerd dat volgens de daling danwel stijging in de algehele vraag maar transport voor een specifiek dagtype het nuttig is om later op de avond of vroeg in de ochtend het OV bedieningsinterval uit te breiden. Alle werkdagen tussen 5 januari en 31 mei 2015 zijn in deze studie meegenomen, met uitzondering van een aantal dagen waarin grote evenementen plaatsvinden. Het begin en einde van het OV aanbod kan verschillen per zone. In de late avond worden de laatste ritten tussen middernacht en 02:00 gereden, en in de vroege ochtend wordt de exploitatie hervat tussen 05:00-07:00. In totaal zijn er 84 nachten meegenomen in de analyse, waarbij de resultaten van zowel de OV chipkaart als de GSM data worden gebaseerd op de gemiddelde mobiliteitspatronen van deze nachten. De resultaten worden gepresenteerd met betrekking tot de relatieve verandering ten opzichte van het vorige uur. In het geval van de bezoekersdrukte, gemeten door de GSM data, betekent een daling ten opzichte van het vorige uur een vraag naar transport uit de zone. Een stijging in bezoekersdrukte daarentegen laat een vraag naar transport naar de zone zien. De resultaten voor de analyseniveaus totaal per uur en per zone per uur worden gepresenteerd in de hierop volgende paragrafen.

3.2 Resultaten analyseniveau totaal per uur

De MPE waarden voor een werkdag voor het totale case study gebied per uur zijn geplot in Figuur 4. In de late avond tot 02:00, nemen zowel de bezoekersdrukte als het OV gebruik significant af per uur, waarbij het OV gebruik sterker daalt. Deze resultaten suggereren dat er potentie is om het OV bedieningsinterval in de late avond uit te breiden, aangezien de daling in bezoekersdrukte geleidelijker plaatsvindt, waarbij ook andere modaliteiten dan alleen OV worden gebruikt. Tussen 02:00-05:00 vinden er geen significante veranderingen plaats in beide type mobiliteitspatronen. In de vroege ochtend kan een extreme stijging in zowel OV gebruik als bezoekersdrukte worden geïdentificeerd. Op dit geaggregeerde ruimtelijke en temporele analyseniveau lijkt de verschuiving van de nacht naar de dagelijkse operatie overeen te komen met de vraag naar transport.



Figuur 4 MPE waarden van OV gebruik en bezoekersdrukte (per uur, voor werkdagen) gerelateerd aan het vorige uur

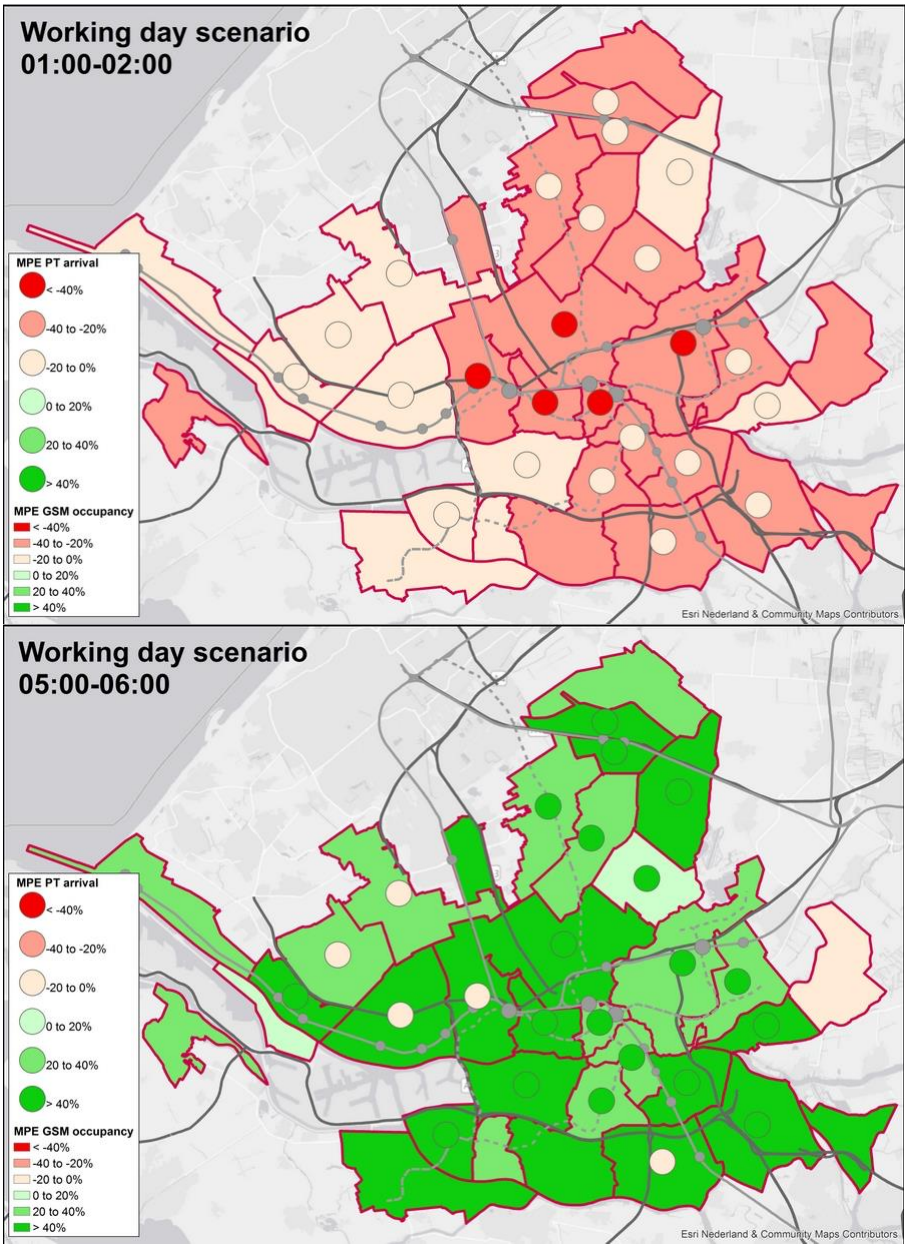
3.3 Resultaten analyseniveau per zone per uur

Binnen het analyseniveau per zone per uur is het mogelijk om op zone niveau mogelijke tekortkomingen van het huidige OV bedieningsinterval te identificeren. In de resultaten valt op dat in enkele zones de bezoekersdrukte nog afneemt, na het einde van het OV bedieningsinterval (01:00-02:00), of alweer significant toeneemt in de ochtend voor de start van het OV bedieningsinterval (05:00-06:00). De resultaten voor beide tijdsintervallen zijn gevisualiseerd in Figuur 5. De achtergrond kleur van elke zone geeft de relatieve MPE waarde van de bezoekersdrukte in relatie tot het vorige uur weer, de cirkel in de zone geeft de relatieve MPE waarde van het OV gebruik weer ten opzichte van het vorige uur. Als er geen cirkel in de zone is, is er geen OV data beschikbaar voor werkdagen. De minimale MPE

waarde volgend uit de OV Chipkaart data is -100%, terwijl dit -37% is voor de GSM data. De maximale MPE waarden voor het OV gebruik zijn te vinden in het eerste uur van operatie in de vroege ochtend, aangezien in het uur ervoor geen OV exploitatie plaatsvond. De maximale MPE waarde volgende uit de GSM data is +88%. In de figuur geven een lichtrode en lichtgroene arcering een respectievelijke daling en stijging ten opzichte van het vorige uur weer die binnen de grenswaarde valt; deze veranderingen zijn niet significant (sectie 2.2). Voornamelijk zones met een tegengestelde achtergrond kleur en cirkel in de zone zijn interessant voor de RET. Voor het tijdsinterval 01:00-02:00 (Figuur 5, boven) zijn de meest interessante zones in het noorden en zuiden van Rotterdam te vinden (donker rode achtergrond, licht rode cirkel). In deze voorsteden en woongebieden is in de late avond nog een significante afname in bezoekersdrukte te vinden, die niet wordt geaccommodeerd door het OV, aangezien het bedieningsinterval al is afgelopen. Voor het tijdsinterval 05:00-06:00 (Figuur 5, beneden) zijn de meest interessante zones in het westen en zuiden van Rotterdam te vinden (donker groene achtergrond, licht rode cirkel). Deze industriële en logistieke zones rondom de Rotterdamse haven ondervinden tussen 05:00-06:00 een significante stijging in bezoekersdrukte, terwijl het OV bedieningsinterval nog niet is begonnen.

De MPE waarden geven de relatieve verandering in de data weer, om twee verschillende databronnen aan elkaar te kunnen linken. De netto verandering in bezoekersdrukte per uur is echter ook belangrijk voor de vervoerder, zodat potentieel OV gebruik kan worden geïdentificeerd. Tabel 1 en 2 geven de relatieve MPE waarden weer voor de bezoekersdrukte voor respectievelijk de tijdsintervallen 05:00-06:00 en 01:00-02:00, ten opzichte van het vorige uur, gebaseerd op de interessante zones geïdentificeerd in Figuur 5. Daarnaast geven Tabel 1 en 2 ook de netto verandering in bezoekersdrukte weer. In de vroege ochtend, tussen 05:00-06:00 wordt een netto drukte verandering van 1.600 bezoekers geïdentificeerd in Barendrecht, een industriële zone (Tabel 1, Figuur 3). In deze zone wordt een grote inkomende bezoekersstroom worden verwacht door de OV vervoerder, en is er OV potentie aanwezig. Voor Maasland en Schipluiden echter, geldt dat OV potentie niet aanwezig is, gezien de lage netto verandering in bezoekersdrukte. Voor andere zones zoals Schiedam en Vlaardingen, kan een relatief hoge netto verandering in bezoekersdrukte worden geïdentificeerd. Voor deze zones geldt ook dat er mogelijk vraag is naar een uitbreiding van het OV bedieningsinterval in de vroege ochtend.

Daarnaast ondervinden de zones net ten zuiden van de Maas, Feyenoord en Ridderkerk (Tabel 2, Figuur 3), een significante daling in drukte van tenminste 1.000 bezoekers in de late avond. Dit aantal bezoekers is het minimum aantal mensen met een vraag naar transport uit die zone gedurende dit uur. Aangezien de verandering een netto verandering is, kan de aankomst-vertrek ratio niet worden afgeleid uit de data. In deze zones is daarom potentie voor een langer OV bedieningsinterval geïdentificeerd.



Figuur 4 MPE waarden gedurende werkdagen voor bezoekersdrukte (achtergrondkleur van de zone) en OV gebruik (kleur van stip in de zone) voor de tijdsintervallen 01:00-02:00 (boven) en 05:00-06:00 (beneden).

Tabel 1: Relatieve en absolute verandering in bezoekersdrukte voor geselecteerd zones gedurende het tijdsinterval 05:00-06:00 op werkdagen

Zone naam	**MPE waarde bezoekersdrukte**	**Netto verandering in bezoekersdrukte**
Barendrecht	+88%	1.600
Maasland	+33%	180
Schiedam	+60%	1,300
Schipluiden	+35%	75
Vlaardingen	+51%	1.000

Tabel 2: Relatieve en absolute verandering in bezoekersdrukte voor geselecteerde zones gedurende het tijdsinterval 01:00-02:00 op werkdagen

Zone naam	**MPE waarde bezoekersdrukte**	**Netto verandering in bezoekersdrukte**
Barendrecht	-27%	550
Bergschenhoek	-23%	200
Berkel & Rodenrijs	-27%	300
Feyenoord	-27%	1.100
IJsselmonde	-23%	550
Pijnacker	-24%	150
Ridderkerk	-37%	1.000
Zoetermeer Midden	-25%	500
Zoetermeer Zuid	-29%	300

4. Conclusies en aanbevelingen

De uitdaging voor de openbaar vervoer sector is om tegemoet te komen aan de verscheidenheid aan reispatronen, en de bijbehorende behoeften en preferenties, van reizigers. Hoewel datafusie de potentie heeft om spreiding van vervoersvraag in zowel ruimte als tijd te onderzoeken, wordt dit nauwelijks toegepast door vervoerders. We ontwikkelden een methode om anonieme OV Chipkaart data en GSM data te fuseren, teneinde OV gebruik in relatie tot de totale vervoersvraag te analyseren. Op basis van de relatie tussen de relatieve verandering in OV gebruik en de bezoekersdichtheid op verschillende analyseniveaus, kunnen interessante ruimtelijke en temporele aspecten gevonden worden. De voorgestelde methode voor datafusie is primair van toegevoegde waarde ter ondersteuning van OV planning en besluitvorming op tactisch niveau.
Als gevolg van verschillende semantiek tussen OV Chipkaart data en GSM data is het niet mogelijk om de beide datasets direct met elkaar te fuseren. Echter, onze methode illustreert een systematische verkenning en analyse van OV gebruik in relatie tot de totale vervoersvraag. Wanneer slechts één dataset gebruikt zou worden, is het niet mogelijk om deze informatie te deduceren. Door het minder gedetailleerde ruimtelijke detailniveau van GSM data is het niet mogelijk om de exacte locaties te bepalen waar een vervoersbehoefte bestaat. Ook zijn herkomst-bestemming relaties onbekend. Echter, de toepassing van de methode in een case study in Rotterdam liet diverse zones zien die interessant zijn voor vervoerders: zones die potentie lieten zien om het OV bedieningsinterval uit te breiden zowel in de late avond als in de vroege ochtend. De potentie van deze latente OV vraag moet echter in meer detail onderzocht worden. Hierbij dienen de mogelijke lijnvoeringen en het OV marktaandeel in beschouwing te worden genomen, aangezien OV aanbod tijdens deze uren niet een volledige modal shift zal realiseren. Om vast te stellen of het uitbreiden van het OV bedieningsinterval kansrijk is, zijn inschattingen voor gebruik en kosten noodzakelijk.
De datafusie methode zoals voorgesteld in deze paper kan worden gebruikt om een grote range aan datasets met (geaggregeerde of gedesaggregeerde) informatie over herkomsten en bestemmingen te fuseren. Beperkingen van deze methode hebben primair betrekking op issues rondom de dataverwerking. Hoewel de zones gebruikt voor de aggregatie van GSM

data in Nederland steeds kleiner worden, bepalen privacy regels dat een aanzienlijke aggregatie in stand gehouden dient te worden *(Calabrese et al. 2013)*. Het in beschouwing nemen van herkomst-bestemming relaties in de GSM data als toekomstige uitbreiding van de methode biedt meer informatie over de richting van potentiële OV vraag. OV Chipkaart data is in Nederland eigendom van de verschillende individuele vervoerders. Fusie van OV Chipkaart data van verschillende vervoerders, inclusief data van het hoofdrailnet, heeft de potentie om overstappen van passagiers tussen OV diensten van verschillende vervoerders te identificeren.

Acknowledgements

Dit onderzoek is tot stand gekomen door een samenwerking van TU Delft, Goudappel Coffeng, DAT.Mobility en RET.

Referenties

Aguilera, V., Allio, S., Benezech, V., Combes, F. and Milion, C. Using cell phone data to measure quality of service and passenger flows of Paris transit system. *Transportation Research Part C: Emerging Technologies,* Vol. 43(2), 2014, pp. 198-9 211.

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., & Ratti, C. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation research part C: emerging technologies*, Vol 26, 2013, pp. 301-313.

Cats, O., Wang, Q., & Zhao, Y. Identification and classification of public transport activity centres in Stockholm using passenger flows data. *Journal of Transport Geography*, Vol. 48, 2015, pp. 10-22.

Del Castillo, J. M., & Benitez, F. G. A methodology for modeling and identifying users satisfaction issues in public transport systems based on users surveys. *Procedia Social and Behavioral Sciences*, Vol. 54, 2012, pp. 1104-1114.

De Regt, K.L. How do spatial and temporal patterns of public transport relate to the overall travel demand? A data fusion method for smart card data and GSM data, *Master thesis, Delft University of Technology,* 2016.

Duff-Riddell, W. R. and Bester, C. J. Network modeling approach to transit network design. *Journal of Urban Planning and Development,* Vol. 131(2), 2005, pp. 87-97.

Durand, C. P., Tang, X., Gabriel, K. P., Sener, I. N., Oluyomi, A. O., Knell, G., Porter, A.K., Hoelscher, D. M. & Kohl, H. W. The association of trip distance with walking to reach public transit: Data from the California Household Travel Survey. *Journal of Transport & Health*, Vol. 3(2), 2016, pp. 154-160.

Elfrink, M., Courtz, M., Metz, S., Ebben, M., and Weppner, J. OV-potentie opsporen 20 door datafusie. *Nationaal Verkeerskundecongres*, 2015.

Frias-Martinez, V., Soguero, C. and Frias-Martinez, E. Estimation of urban commuting patterns using cell phone network data. In *Proceedings of 6th Transport Research Arena, April 18-21, 2016, Warsaw, Poland,* 2016.

Guedes, M. C. M., Oliveira, N., Santiago, S., & Smirnov, G. On the evaluation of a public transportation network quality: Criteria validation methodology. *Research in Transportation Economics*, Vol. 36(1), 2012, pp. 39-44.

Gutierrez, J., & García-Palomares, J. C. New spatial patterns of mobility within the metropolitan area of Madrid: towards more complex and dispersed flow networks. *Journal of transport geography*, Vol. 15(1), 2007, pp. 18-30.

Holleczek, T., Yu, L., Lee, J. K., Senn, O., Ratti, C., & Jaillet, P. Detecting weak public transport connections from cellphone and public transport data. In *Proceedings of the 2014 International Conference on Big Data Science and Computing* (p. 9). ACM, 2014.

Iqbal, M. S., Choudhury, C. F., Wang, P., & González, M. C. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, Vol. 40, 2014, pp. 63-74.

Kusakabe, T. and Asakura, Y. Behavioural data mining of transit smart card data: A data fusion approach. *Transportation Research Part C: Emerging Technologies,* Vol. 46, 2014, pp. 179-191.

Liu, L., Hou, A., Biderman, A., Ratti, C. and Chen, J. Understanding individual and collective mobility patterns from smart card records: A case study in Shenzhen. In *Intelligent Transportation Systems, 2009. ITSC' 09. 12th International IEEE Conference*, p. 1-6, 2009.

Long, Y., & Thill, J. C. Combining smart card data and household travel survey to analyze jobs–housing relationships in Beijing. *Computers, Environment and Urban Systems*, Vol. 53, 2015, pp. 19-35.

Nishiuchi, H., King, J. and Todoroki, T. Spatial-temporal daily frequent trip pattern of public transport passengers using smart card data. *International Journal of Intelligent Transportation Systems Research,* Vol. 11(1), 2013, pp. 1-10.

RET, *Over RET,* http://corporate.ret.nl. Juli 2016.

Van der Mede, P. Over het meten van mobiliteit met GSM-data: mogelijkheden en onmogelijkheden. Contribution for *Colloquium Vervoersplanologisch Speurwerk, Eindhoven, 2014.*

Van Oort, N., Sparing, D., Brands, T. and Goverde, R. M. P. Data driven improvements in public transport: the Dutch example. *Public Transport,* Vol. 7(3), 2015a, pp. 369-389.

Van Oort, N., T. Brands, E. de Romph, Short-Term Prediction of Ridership on public Transport with Smart Card Data, *Transportation Research Record*, No. 2535, 2015b, pp. 105-111.

ViewDAT. *ViewDAT*, http://view.dat.nl/viewdat/. Oktober 2015.