Privacy-bewust analyseren van ov-chipkaartdata

Dr. Ir. H.J.W. van Heerde
Irias Informatiemanagement
h.vheerde@irias.nl

Ir. J.J.F. Hoogenboom
BonoTraffics bv
jhoogenboom@bonotraffics.nl

Bijdrage aan het Colloquium Vervoersplanologisch Speurwerk
20 en 21 november 2014, Eindhoven

Samenvatting

Privacy-bewust analyseren van ov-chipkaartdata

Het gebruik van ov-chipkaartdata voor analysedoeleinden biedt veel mogelijkheden. Echter, de ov-chipkaartdata ligt gefragmenteerd opgeslagen bij vervoerders, waardoor de beschikbaarheid van dergelijke data nog beperkt is. Belangrijke informatie, zoals het aantal overstappers tussen trein en bus, maar ook de herkomst-bestemmingsrelaties, is daardoor niet of moeilijk te achterhalen. Het open beschikbaar stellen van de data kan daarom leiden tot betere analysemogelijkheden waardoor bijvoorbeeld overheidsgeld beter besteed kan worden. Daarnaast kan het open beschikbaar stellen leiden tot betere marktwerking bij het meedingen naar concessies. Eén van de argumenten tegen het openbaarstellen van ov-chipkaartdata zijn zorgen over de privacy.

Het doel van dit artikel is om het begrip privacy in de context van openbaar vervoer, en specifieker de ov-chipkaart, inzichtelijker te maken. Hierdoor kan er op een privacy-bewuste manier omgegaan worden met de analysemogelijkheden van ov-chipkaartdata. Er is veel publieke aandacht besteed aan onder meer de beveiliging van de ov-chipkaart, waardoor er veel wantrouwen is ontstaan onder gebruikers. Het is daarom van belang onderscheid te maken tussen beveiliging van de kaart zelf, en het op een privacy-bewuste manier omgaan met de gegevens die als bijproduct van de ov-chipkaart worden verzameld. Wij richten ons op het laatste, met concrete voorbeelden van anonimisatietechnieken. We laten zien hoe ov-chipkaartgegevens kunnen worden geanonimiseerd, maar ook welke problemen hierbij een rol spelen. We laten zien dat er altijd een balans moet worden gevonden tussen de mate van privacy voor individuen en de bruikbaarheid van de gegevens voor analyse.

Door gericht vooraf te bepalen voor welke analysedoeleinden de data gebruikt moet kunnen worden, kunnen openbare datasets beperkt worden tot de gegevens die er echt toe doen, en daarmee de privacy-gevoeligheid worden verminderd. Dit geldt ook voor ov-chipkaartdata. Daarmee wordt de deur opengezet tot het veilig en bewust verspreiden van ov-chipkaartdata, zodat grote maatschappelijke besparing dankzij gerichte analyses kunnen worden gerealiseerd.

1. Inleiding

Vanaf 2005 hebben reizigers kennis kunnen maken met de ov-chipkaart. Sinds 2012 is het systeem landelijk ingevoerd en kan de kaart in het hele openbaar vervoer[1] worden gebruikt. Veruit de meeste reizen worden inmiddels met behulp van een ov-chipkaart gemaakt. Daarmee zijn de data die alle chipkaarten in Nederland generen een potentieel interessante bron voor onderzoekers en adviseurs. Echter, onderstaande tabel laat zien wie in de huidige situatie toegang heeft tot welke data en met welk detailniveau. Zoals uit het overzicht blijkt is op dit moment Trans Link Systems (TLS) de enige partij in Nederland die de beschikking heeft over alle reizigersdata.

Wie?	**Toegang tot welke data?**	**Wat is het detailniveau?**
Reizigers	Van eigen chipkaarten	Alle afzonderlijke transacties
Vervoerder	Van de reizen met de betreffende vervoerder	Alle afzonderlijke transacties
Concessieverlener	Van de reizen binnen de concessie	Geaggregeerd naar bv uurblokken, maandtotalen, hele concessie
Trans Link Systems	Van alle chipkaarten	Alle afzonderlijke goedgekeurde transacties

De data is dus niet vrij beschikbaar en wordt ten behoeve van onderzoek of analyses gefragmenteerd per vervoerder aangeleverd. Belangrijke informatie, zoals het aantal overstappers tussen trein en bus, maar ook de herkomst-bestemmingsrelaties, is daardoor niet te achterhalen. Het open beschikbaar maken van deze data kan dan ook veel voordelen bieden:

- Er ontstaat een compleet beeld van het ov-systeem;
- Betere marktwerking bij meedingen naar concessies;
- Maatschappelijke waarde.

De openbaar vervoer markt is onderverdeeld in meerdere concessies, waar verschillende vervoerders actief zijn. Omdat de data tussen vervoerders en concessies van verschillende opdrachtgevers gescheiden is, is er geen beeld van de hele reis die iemand binnen het hele ov-systeem maakt. Door de fragmentatie van de data zijn het teveel puzzelstukjes, die niet op elkaar gepast kunnen worden. Het open beschikbaar stellen van de data zorgt voor een beter beeld van het hele ov-systeem waar zowel de overheid als de vervoerder baat bij hebben. Als overheid kun je vervolgens integraal het netwerk analyseren en verbeteren in plaats van telkens lokaal een probleem aan te pakken. Dit maakt gerichte investeringen mogelijk, waardoor overheidsgeld effectiever wordt besteed.

Niet alleen de overheden maar ook de vervoerder profiteren van het delen van reizigersinformatie. Wanneer alle vervoerders inzicht hebben in elkaar reizigersstromen is het eenvoudiger om het aanbod en netwerk op elkaar af te stemmen. Reizigers rijden nu

[1] Met uitzondering van de Waddeneilanden (wel op Texel) en een aantal buurtbussen

mogelijk onnodige om omdat iedere vervoerder gefocust is op zijn eigen schakel in de reisketen. Inzicht in de reizigersstromen van begin tot eindpunt maakt een geoptimaliseerd netwerk mogelijk waarbij de kosten dalen en het aantal reizigers toeneemt. Hiervan profiteren alle vervoerders en ook de maatschappij.

Het open beschikbaar stellen van de data leidt ook tot betere markwerking bij het meedingen naar concessies. Nu heeft de zittende vervoerder een enorme voorsprong op de concurrent omdat hij gedetailleerd inzicht heeft in de reizigersstromen. Ook heeft hij inzicht in de ontwikkelingstrend van het aantal reizigers op allerlei trajecten en herkomst-bestemmingsrelaties. Omdat een nieuwe vervoerder geen cijfers tot zijn beschikking heeft van de reizigersstromen in een concessiegebied, zal hij onnodig risico's moeten nemen om toch een kans te maken. De vervoerder rekent zich rijk met reizigers die er niet zijn of schat het aantal reizigers veel te laag in. Het aanbod zal dan niet overeenstemmen met de vraag. In beide gevallen zijn hier kosten mee gemoeid die direct of met een omweg weer bij de opdrachtgever, de overheid, terecht komen. Open reizigersdata biedt een gelijker speelveld waardoor de kansen voor alle vervoerders min of meer gelijk zijn en concurrentie kan plaatsvinden op kwaliteit en innovatie.

Ook de maatschappij heeft belang bij open chipkaartdata. Het openbaar vervoer wordt deels bekostigd met belastinggeld. Zowel direct via de Brede Doelen Uitkering (BDU) als indirect via bijvoorbeeld de ov-studentenkaart. Als iedereen zou kunnen zien waartoe deze overheidsuitgaven leiden, kan het publieke debat over deze uitgaven beter worden gevoerd. Is het anno 2014 nog verantwoord om de ontsluiting van het platteland middels openbaar vervoer te bekostigen met overheidsgeld of zijn er innovatieve en goedkopere oplossingen denkbaar die hetzelfde doel dienen?

Waarom is ov-chipkaartdata dan nog niet openbaar? Uit de eerder genoemde voordelen blijkt dat het openbaar beschikbaar stellen van chipkaartdate potentieel veel voordeel kan bieden. Waarom gebeurt het dan nog niet?

De overheid heeft meer dan 1 miljard euro in de ov-chipkaart gestoken. Hoewel een groot deel daarvan op is gegaan aan harde infrastructurele maatregelen zoals poortjes en paaltjes is een substantieel deel besteed aan de ontwikkeling en invoering van de kaart zelf (Kamerstukken II, 2007/08 3866). In veel landen zou de ov-chipkaartdata per definitie in het publieke domein zitten als het uit publieke middelen is gefinancierd.

Een belangrijke reden waarom ov-chipkaartdata nog niet open beschikbaar is, is omdat zittende vervoerders hun grootste voordeel verliezen bij aanbestedingen. Vervoerders verliezen het voordeel dat ze hebben op nieuwe vervoerder. De meeste concessies laten weinig ruimte open voor innovatie waardoor vervoerders vooral op prijs met elkaar moeten concurreren. Als jij als zittende vervoerder gedetailleerd zicht hebt in de reizigersstromen kun je waarschijnlijk een beter bod indienen dan de concurrent. Zoals in de voordelen voor het open stellen van de data benoemd is, is dit niet in het belang van zowel de maatschappij als de vervoerder omdat het leidt tot ongelijkheid en daardoor suboptimale marktwerking.

Privacy is een ander veel genoemd argument. Er heerst een algemene 'angst' binnen de samenleving, en onbegrip bij gebruikers en analisten van ov-chipkaartdata. Doel van dit

artikel is onder meer om die angst weg te nemen, en analisten bewust maken van de privacy-issues. Door privacy-bewust om te gaan met ov-chipkaartdata kan deze data als open data verspreid worden en zou privacy geen rol moeten spelen in het wel of niet openbaar maken van de data. De open data leidt vervolgens tot nieuwe analysemogelijkheden en toepassingen die het openbaar vervoer efficiënter, kwalitatief beter en transparanter laten functioneren dan nu het geval is.

In het vervolg van dit artikel gaan wij in op deze privacy-aspecten van ov-chipkaartdata. Met simpele voorbeelden laten we zien wat begrippen als privacy, anonimisatie en aggregatie inhouden en betekenen voor de analysemogelijkheden van ov-chipkaartdata. Daarna, in sectie 3, gaan we in op die analysemogelijkheden, en laten zien wat nu al in de praktijk mogelijk is. Sectie 4 sluit af met een conclusie.

2. Privacy

Al vanaf de introductie van de ov-chipkaart is er discussie over 'de privacy'. Onder andere het hacken van de ov-chipkaart door onderzoeksjournalist Brenno de Winter (Wikipedia 2014) is veel in de publiciteit geweest, wat het vertrouwen in de beveiliging van de ov-chipkaart geen goed gedaan heeft. Hierdoor is, al dan niet terecht, ook het vertrouwen in de waarborging van privacy gedaald. Dit kan tot gevolg hebben dat innovaties geremd worden. Vaak wordt privacy als reden genoemd om bijvoorbeeld data niet te verzamelen. Nog vaker wordt privacy in zijn geheel genegeerd en veel meer verzameld dan nodig is, met vervelende privacy-schandalen tot gevolg.

De discussie rond privacy is een moeilijke. Privacy is meestal niet tastbaar en concreet uit te drukken in waarde, terwijl de mogelijkheden van innovaties en nieuwe producten of diensten dat meestal wél zijn. Vaak leidt de discussie rond privacy ook tot onbegrip: voor onderzoekers en wetenschappers is het vaak niet duidelijk waar de privacy-gevaren liggen, en voelen zij zich onnodig geremd in hun werk om, bijvoorbeeld, het openbaar vervoer beter te maken.

2.1 Privacy in de context van openbaar vervoer

Privacy is een breed begrip, dat zich moeilijk laat definiëren. Een van de gangbare definities is (Tsukada 2009):

"Privacy is het recht om te verbergen wat heeft plaatsgevonden."

Dit recht is een onbetwistbaar recht, dat wil zeggen, een individu hoeft geen reden op te geven *waarom* hij iets wil verbergen. Dit recht zit over het algemeen ook stevig verankerd binnen bestaande wetgevingen, al zitten er grenzen aan het 'wat'. Iemand heeft in Nederland bijvoorbeeld niet het recht om zijn inkomen te verbergen voor de belastingdienst.

De praktijk is weerbarstiger. Letterlijk vertaald naar het domein van openbaar vervoer houdt deze definitie van privacy namelijk in dat een reiziger het recht heeft om het feit dat hij een bepaalde reis heeft afgelegd te verbergen. Dit staat haaks op het belang van de vervoerders, immers, een vervoerder wil een reis kunnen afrekenen en daarvoor moet

de vervoerder op zijn minst weten dat en hoever een reiziger gereisd heeft. Daarom doet een reiziger afstand van zijn recht om zijn reis verborgen te houden op het moment dat hij akkoord gaat met de algemene voorwaarden en het daaraan gekoppelde privacy-beleid van de vervoerder op het moment dat hij een ov-chipkaart aanschaft. Zolang dit privacy-beleid getoetst en goedgekeurd wordt door het College Bescherming Persoonsgegevens (CBP) is er in principe niets aan de hand.

Het feit dat een reiziger ooit akkoord is gegaan met de algemene voorwaarden van een vervoerder betekent niet direct dat de vervoerders vervolgens een vrijbrief hebben om onbeperkt reisgegevens op te slaan en te analyseren. Een gangbaar privacy-beleid houdt zich aan de volgende richtlijnen:

- Gegevens mogen alleen opgeslagen en verwerkt worden als er een specifiek geldig doel voor is (en de gegevensverstrekker, de reiziger, moet hier ondubbelzinnig mee akkoord gaan);
- Gegevens mogen slechts beperkte tijd worden opgeslagen.

Het College Bescherming Persoonsgegevens heeft in haar visie op privacy met betrekking tot de ov-chipkaart gesteld dat de verwerking van gegevens over reisbewegingen beperkt moet blijven tot hetgeen nodig is om de *vervoersfunctie* van de vervoerders te vervullen (CBP 2005). Het CBP onderkent hierbij de noodzaak om specifieke reizen gedurende een periode te bewaren, zodat eventuele claims van reizigers wat betreft onterecht gefactureerde bedragen afgehandeld kunnen worden. Het zondermeer voor langere tijd opslaan van tot op de persoon herleidbare reisbewegingen is dan ook niet toegestaan.

De vraag is dan ook in hoeverre privacywetgeving het publiceren van reisbewegingen als open data in de weg staat. Immers:

- Het doel 'analyseren van ov-chipkaartdata' is vaag en te breed, en voldoet daarom niet aan de eis van het CBP;
- Het kunnen analyseren van ov-chipkaartdata is geen vereiste om de vervoersfunctie mogelijk te maken, en is daarom geen reden om reisbewegingen op te slaan;
- Voor historische analyses zijn historische gegevens nodig. Reisbewegingen mogen echter maar voor korte tijd worden opgeslagen.

Gelukkig is er een oplossing: het anonimiseren van gegevens. De wetgeving zoals hierboven beschreven gaat over het beschermen van *persoonsgegevens*. Op het moment dat ov-chipkaartdata niet meer te herleiden is tot personen, is er geen sprake meer van persoonsgegevens, en dus ook niet van privacy-gevoelige data.

2.2 Anonimiseren van data

Anonimisatie van gegevens is een gangbare methode om privacy-gevoelige gegevens te kunnen publiceren zonder privacywetgeving te schenden (Sweeney 2002). Anonimiteit is wezenlijk anders dan privacy: waar het bij privacy gaat om het beschermen van *wat* een specifiek iemand heeft gedaan, gaat het bij anonimiseren om het verbergen van *wie* iets gedaan heeft. Anonimiseren wordt vaak verward met *aggregeren*. Aggregeren is het samenvatten van een groep gegevens tot één kenmerkend gegeven, zoals het gemiddelde of de som van de gegevens, of het aantal in de groep. Aggregeren kan een

techniek zijn die gebruikt wordt bij het anonimiseren. Het aggregeren van data garandeert echter geen anonimiteit, en anonimiteit kan bereikt worden zonder gegevens te aggregeren.

Voorbeeld van aggregatie:

Reiziger	**Reis**
Piet	A naar B naar C
Jan	A naar B naar C
Kees	A naar E
Wim	B naar D
Daphne	A naar C
Yvonne	A naar C
Anouk	B naar D

Tabel 1: voorbeeld van brontabel

wordt

Aantal reizigers	**Reis**
2	A naar B naar C
1	A naar E
2	B naar D
2	A naar C

Tabel 2: voorbeeld van geaggregeerde tabel

Voorbeeld van anonimisatie:

Reiziger	**Reis**
Man	A naar B naar C
Man	A naar B naar C
Man	A naar E
Man	B naar D
Vrouw	A naar C
Vrouw	B naar C
Vrouw	B naar D

Tabel 3: voorbeeld van geanonimiseerde

Bij het analyseren van ov-chipkaart data gaat het niet om welk individu iets gedaan heeft, het gaat om de reisbewegingen die gemaakt zijn. Anonimisatie (eventueel geholpen door het aggregeren van gegevens) is dus in principe een geschikte techniek om wel beschikking te kunnen hebben over ov-chipkaartdata, zonder dat de privacy van individuen in het geding komt. Echter, het dusdanig anonimiseren zodat geen enkele data te herleiden is tot een individu, is extreem moeilijk (van Heerde 2010). Met name als men in ogenschouw neemt dat individuele reizen identificerend kunnen zijn, en er gebruik gemaakt kan worden van 'steunbewijs' uit andere datasets, wordt het met anonimisatietechnieken moeilijk om volledige privacy te garanderen. Volledige privacy kan, maar dan zal de overgebleven data vrijwel nutteloos zijn.

Om toch gebruik te kunnen maken van ov-chipkaartdata moeten er dus compromissen gesloten worden tussen de bruikbaarheid van data en de waarschijnlijkheid dat een specifiek gegeven, zoals een reisbeweging, terug te voeren is tot een individu.

Voorbeeld van gebrekkige anonimisatie

Reiziger	Woonplaats	Werkadres
Piet	A	C
Jan	A	C
Kees	A	E
Wim	B	D
Daphne	A	C
Yvonne	A	C
Anouk	B	D
Marieke	B	D

Tabel 4: tabel met woon/werk-gegevens (voorbeeld)

Stel, de woonplaats en werkadres van iedere mogelijke reiziger is bekend (tabel 4). Stel we weten dat de geanonimiseerde tabel 4 een lijst met woon-werk reisrelaties bevat. Uit de combinatie van tabel 4 en 5 is met grote mate van waarschijnlijkheid (in dit vereenvoudigde voorbeeld) te herleiden dat de man die van A naar E reisde, in werkelijkheid Kees is geweest. Bovendien weten we hoewel er twee individuen van B naar D reizen, er slechts één man in B woont en in D werkt. We kunnen dus ook veronderstellen dat Wim de man is die van B naar D heeft gereisd. We kunnen niet zeker weten of het Anouk of Marieke is geweest die van B naar D heeft gereisd.

2.3 Anonimisatie van ov-chipkaartdata in de praktijk

Om inzicht te geven hoe er momenteel in de praktijk omgegaan wordt met reisgegevens vanaf de bron tot de analyse zoals in sectie 2 worden beschreven, laten we stap voor stap zien hoe de data verwerkt wordt tot steeds minder privacy-gevoelig. De stappen zijn sterk vereenvoudigd, en kunnen afwijken van hoe vervoerders met de data omgaan.

Als voorbeeld nemen we aan dat de data als volgt wordt opgeslagen:

Reiziger	Kaartnummer	Tijdstip	Reis
Piet	1	2014-09-04 07:31	A naar B naar C

Stap 1: splitsen van persoonsgegevens en reisgegevens

De meest voor de hand liggende vorm van anonimiseren is het simpelweg loskoppelen van reiziger en kaartnummer (Connexxion 2014).

Reiziger	Kaartnummer
Piet	1

Kaartnummer	Tijdstip	Reis
1	2014-09-04 07:31	A naar B naar C

Het idee hierachter is dat zolang iemand alleen toegang heeft tot de reisbewegingen, deze niet direct gekoppeld is aan de persoonsgegevens van de reiziger, en privacy dus gewaarborgd zou zijn. Hier worden echter twee belangrijke aannamen gedaan:

- Het kaartnummer is nooit te herleiden tot een persoon;
- Het reispatroon van een persoon is niet identificerend

Beide aannamen zijn zwak. Ten eerste, systemen worden vrijwel altijd vroeg of laat gehackt of op andere wijze ontsloten naar de buitenwereld (Van Heerde 2010). Op het moment dat een lijst met persoonsgegevens en kaartnummers wordt openbaard door wat voor reden dan ook, zijn alle reisbewegingen dit tot dat moment als anoniem werden gepubliceerd in één klap ge-de-anonimiseerd. Ten tweede zijn reispatronen in sommige gevallen niet uniek, vooral niet als er relatief 'complexe' reizen worden gemaakt met meerdere overstappen. Een enkele tot de persoon terug te herleiden reis maakt gelijk alle reizen die op dezelfde kaart gemaakt zijn herleidbaar.

Stap 2: het anonimiseren van het kaartnummer

Door het kaartnummer van de reisgegevens te anonimiseren is het niet meer mogelijk om reisgegevens direct aan persoonsgegevens te koppelen:

Kaarttype	**Tijdstip**	**Reis**
Ov-jaarkaart	2014-09-04 07:31	A naar B naar C

Echter, ook hier is het gevaar dat specifieke reizen nog steeds identificerend zijn. Als zowel herkomst als bestemming in een dunbevolkt gebied liggen kan het specifieke tijdstip van de reis de doorslag geven. Vooral de koppeling met externe datasets maakt het mogelijk om individuele reizen te herleiden tot een persoon, gezin, of beperkte groep mogelijke personen. Dergelijke externe datasets kunnen van alles zijn, zoals bijvoorbeeld (openbare) deelnemerslijsten van bijeenkomsten. Veel instanties, met name de overheid, maar ook bijvoorbeeld banken hebben toegang tot gegevens waaruit blijkt dat het een specifiek persoon ergens op een bepaald moment is geweest. Door kruisverbanden te leggen met, weliswaar geanonimiseerde, reisgegevens kan meer worden herleid dan de bedoeling was bij het anonimiseren van de ov-chipkaartdata.

Sommige bronnen zijn voor iedereen toegankelijk, en met de opkomst van open data worden deze bronnen alleen maar talrijker. In de basisregistraties adressen en gebouwen (BAG), welke openbaar beschikbaar is, staan alle adressen in Nederland en hun geografische ligging. In de datasets van Grenzeloze Openbaar Vervoer Informatie (GOVI) staan alle locaties van haltes in Nederland. In het telefoonboek staan zeer veel adressen gekoppeld aan personen. Het zal weliswaar niet mogelijk zijn om voor grote hoeveelheden reisgegevens exact te bepalen wie deze reis heeft uitgevoerd; het *aannemelijk* maken dat een specifiek persoon een specifieke reis heeft afgelegd kan al een schending van privacy zijn.

Stap 3: het aggregeren van reisgegevens

De laatste stap die wij hier behandelen is het aggregeren van de reisgegevens. Hierbij wordt per uur van de dag, per dag van de week, per maand het aantal reizen tussen herkomsten en bestemmingen geteld. Indien het aantal kleiner dan 5 is, wordt het aantal verborgen.

Aantal	**Uur**	**Dagtype**	**Maand**	**Reis**
6	7	Maandag	September 2014	A naar B naar C
* (< 5)	7	Zondag	September 2014	E naar D naar C

Het wordt steeds moeilijker en onwaarschijnlijker dat dergelijke gegevens herleid kunnen worden tot individuen. Toch geldt ook hier dat hoe kleiner en specifieker de groepen worden, hoe groter de kans is dat externe databronnen er aan kunnen bijdragen dat de data kan worden ge-de-anonimiseerd.

Daarbij bestaat het gevaar dat dergelijke, weliswaar stevig geanonimiseerde data, toch gebruikt kan worden door iemand die echt moeite wil doen. Als iemand daadwerkelijk waargenomen wordt tijdens het instappen, of dat bijvoorbeeld een vriend of familielid weet dat een te volgen persoon in een specifiek uurblok is ingestapt bij een bushalte, kan in de ov-chipkaartdata worden teruggezocht welke eindbestemmingen bij die herkomst voor het waargenomen uurblok horen, met een mogelijke privacy-schending tot gevolg. Als bijvoorbeeld de verwachte of afgesproken eindbestemming niet in de data voorkomt, kan dit leiden tot verdachtmakingen.

Bovenstaand probleem kan worden weggenomen door *alle* mogelijke herkomst/bestemming-combinaties een aantal op te nemen, ook al is dit aantal 0 (wat vervolgens wordt genoteerd als < 5). Daarnaast kunnen de groepen waarnaar toe wordt geaggregeerd groter gemaakt worden (bijvoorbeeld grotere uurblokken, of geen onderscheid maken tussen dagtypen).

Dergelijke maatregelen zullen er echter altijd toe leiden dat de kwaliteit en analyseerbaarheid van data afnemen. Er zal daarom een balans gevonden moeten worden tussen welke gegevens nuttig zijn voor analyse, en hoever men wil gaan in het beschermen van privacy. Over deze analyseerbaarheid wordt ingegaan in sectie 3.

3. Analysebehoeften

Aangezien er verschillende methodes zijn voor het anonimiseren van data waarbij elke optie gevolgen heeft voor de analysemogelijkheden, is het belangrijk om van te voren goed na te denken over het doel van de analyse. Het is dan mogelijk om voor verschillende analysebehoefte specifieke datasets te generen. De verschillende analyse- en informatiebehoeften kunnen eenvoudig samengevoegd worden tot drie categorieën van datasets :

- reizigers op halteniveau;
- reizigers op lijnniveau;
- reizigers op relatieniveau.

Met deze drie geanonimiseerde datasets zouden veruit het grootste deel van de vraagstukken beantwoord moeten kunnen worden.

3.1 Analyse naar reizigers op halteniveau

Vaak is er een informatiebehoefte naar reizigersstromen op halteniveau. Het gaat om vragen als:

- Hoeveel reizigers maken gebruik van deze halte?
- Wanneer wordt deze halte het vaakst gebruik?
- Hoe is op deze halte de verdeling van het aantal in- en uitstappers over de verschillende lijnen?
- Wat zijn de belangrijkste overstaphaltes in gemeente x?
- Naar welke buslijnen stappen reizigers over op het treinstation?
- Welke haltes worden niet of nauwelijks gebruikt?

Stel de wegbeheerder heeft een beperkt budget beschikbaar om faciliteiten op bushaltes te verbeteren. De wegbeheerder wil het geld gebruiken om abri's (wachtruimtes) te plaatsen zodat reizigers beschermd tegen weersinvloeden kunnen wachten op de bus. De wegbeheerder wil het geld zo nuttig mogelijk besteden en wil inventariseren op welke haltes de meeste reizigers in- en/of overstappen, zodat de meeste reizigers profiteren van de nieuwe abri's. De wegbeheerder heeft geld beschikbaar om 50 haltes te verbeteren en wil daarom een lijst met de top 50 drukste haltes o.b.v. in- en overstappers.

Hoe zou de data eruit moeten zijn waar de weggebruiker zijn analyse mee kan uitvoeren?

Halte	Maand	Dag	tijd	Lijn	In	Uit	Overstappers
A	Januari 2014	Maandag 6	7:01	5	*(<5)	*(<5)	*(<5)
A	Januari 2014	Maandag 6	7:02	5	*(<5)	*(<5)	*(<5)

Tabel 5: Voorbeeld gedetailleerde geanonimiseerde data

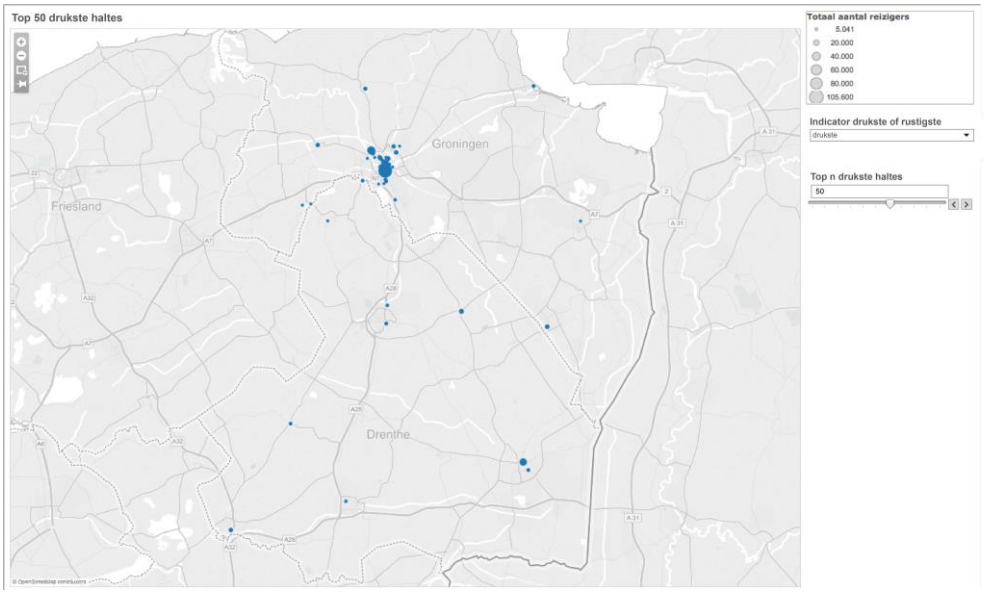
Doordat zowel de dag als het tijdstip te nauwkeurig zijn, kan uit tabel x geen nuttige informatie worden herleidt. Het aantal in-, uit- of overstappers kan immers op elk moment 0, 1, 2, 3 of 4 zijn.

Halte	Maand	Dag	tijd	Lijn	In	Uit	Overstappers
A	Januari 2014	Maandag	7	5	564	78	*(<5)
A	Januari 2014	Maandag	8	5	786	124	*(<5)

Tabel 6: Voorbeeld van aggregeerde data waardoor data anoniem en bruikbaar is.

Tabel 6 bevat nog steeds velden met een waarden kleiner dan 5. Dat het aantal overstappers op deze locatie kleiner is dan 5 is echter ook waardevolle informatie. Hier is immers uit af te leiden dat weinig tot geen reizigers op deze betreffende halte overstappen.

Wanneer alle data van alle haltes binnen het beheergebied van de wegbeheerder wordt gecombineerd kan een top 50 van drukste haltes worden samengesteld.
In figuur 1 is een voorbeeld te zien op basis van fictieve data van de top 50 drukste haltes in de provincies Groningen en Drenthe gevisualiseerd met behulp van de data-analysetool *ovit*.



Figuur 1: voorbeeld van de top 50 drukste haltes in Groningen/Drenthe (op basis van fictieve data)

3.2 Analyse naar reizigers op lijnniveau

Ook op lijnniveau is er een bepaalde informatiebehoefte. Wanneer men wil toetsen of het bestaande ov-netwerk optimaal functioneert dan wil men bijvoorbeeld de volgende vragen kunnen beantwoorden:

- Hoeveel reiziger reizen op een corridor tussen halte A en B?
- Wat is de gemiddelde bezetting op lijn x in de ochtendspits?

Mocht nu uit de analyse naar voren komen dat het netwerk op meerdere punten niet optimaal functioneert, dan kunnen maatregelen getroffen worden om knelpunten te verhelpen. Om te bepalen welke maatregel de hoogste prioriteit heeft is het belangrijk om te weten hoeveel reizigers gebruik van het traject gebruik maken waar de maatregel getroffen wordt. De vraag is dan:

- Hoeveel reizigers profiteren van een (infra)maatregel?

Om alle bovenstaande vragen te kunnen beantwoorden is informatie nodig over het aantal reizigers op een lijn. Momenteel is deze informatie af te leiden uit de MIPOV-tabellen (Model InformatieProfiel Openbaar Vervoer; een gestandaardiseerde tabel met herkomst-bestemmingsinformatie op lijnniveau). Deze data bevat echter wel privacygevoelige informatie. Als er bijvoorbeeld in onderstaand figuur maar 1 reiziger in een maand van A naar C reist dan zou de identiteit van deze persoon eenvoudig achterhaald kunnen worden. In de MIPOV-tabellen wordt normaal gesproken geen ondergrens gehanteerd en komen relaties met 1, 2, 3, enz. aantallen reizigers gewoon voor.

Figuur 2 Schematische weergave van een buslijn

Van halte	Naar halte	Week	Dag	tijd	Lijn	Aantal
A	B	2-2014	Maandag	7	401	2
A	D	2-2014	Maandag	7	401	43
B	C	2-2014	Maandag	7	401	4
B	D	2-2014	Maandag	7	401	7
C	D	2-2014	Maandag	7	401	3

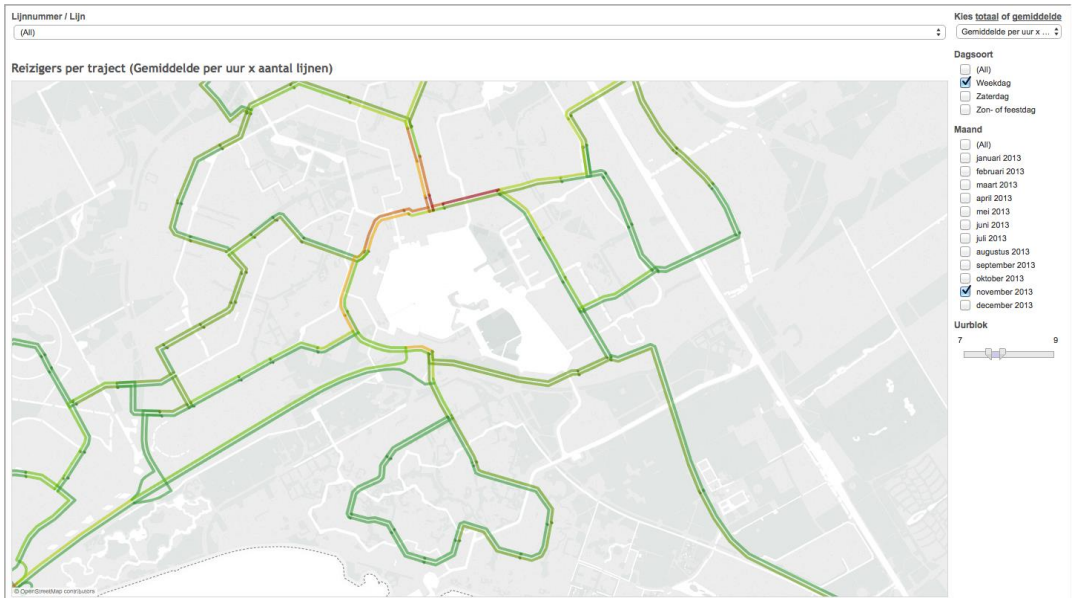
Tabel 7 Gedetailleerde niet geanonimiseerde data van herkomst-bestemmingsrelaties op een lijn (MIPOV-tabellen)

Om aan de informatiebehoefte op lijnniveau te voldoen, is kennis van het aantal reizigers tussen twee opeenvolgende haltes ook voldoende (zie tabel 8). Hierdoor is het vrijwel onmogelijk geworden te herleiden van waar naar waar iemand is gereisd. Door de data te combineren is het aantal in- en uitstappers op een halte niet meer te herleiden. Dat is ook niet nodig, want hier kan een aparte dataset voor worden opgesteld.

Van halte	Naar halte	Week	Dag	tijd	Lijn	Aantal
A	B	2-2014	Maandag	7	401	45
B	C	2-2014	Maandag	7	401	54
C	D	2-2014	Maandag	7	401	53

Tabel 8: Geanonimiseerde data op lijnniveau door combineren van data tot opvolgende halteparen

Door deze data vervolgen te visualiseren met ovit, valt direct op waar de drukke corridors zich bevinden (zie figuur 3). Met handige filters en selectiemethode kan de gebruiker eenvoudig verschillende momenten op de dag, over het jaar of zelfs analyses per lijn uitvoeren.



Figuur 3 voorbeeld van gemiddeld aantal reiziger per uur (op basis van fictieve data)

3.3 Analyse naar reiziger op relatieniveau

De derde categorie van veel voorkomende vragen spelen zich af op het niveau van de herkomst-bestemmingsrelatie. Voorbeelden van vragen in deze categorie zijn:

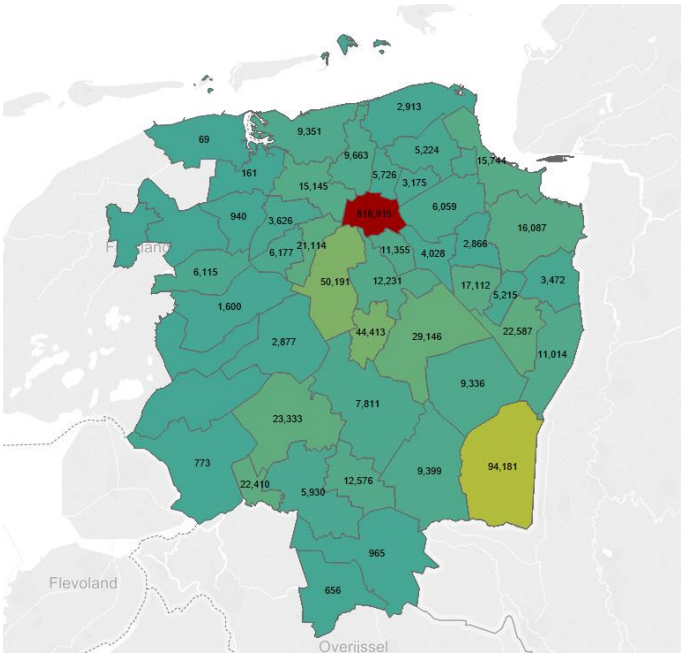
- Wat zijn de belangrijkste bestemmingen voor reizigers uit wijk A?
- Waar komen de reizigers vandaan die naar het centrum van de stad gaan?

De informatiebehoefte binnen deze categorie bevindt zich meestal op een hoger schaalniveau. Het gaat niet meer om haltes of lijnen, maar om gebieden. Dat is maar goed ook want herkomst-bestemmingsrelaties van halte naar halte zijn vaak zo uniek dat het beschikbaar stellen van deze dataset op halteniveau of de privacy in het geding brengt of geen informatie meer bevat.

Door de data te aggregeren naar grotere geografische zones, zoals postcode 4 gebieden, buurten of wijken, is de privacy zelfs bij kleine aantallen gewaarborgd, maar is uit de data wel de benodigde informatie te achterhalen.

Van postcode	Naar postcode	Maand	Dag	tijd	Aantal
1012	1015	Januari 2014	Maandag	7	415
1012	1034	Januari 2014	Maandag	7	287
1012	9725	Januari 2014	Maandag	7	6

In figuur 4 is te zien hoe met behulp van ovit de herkomst-bestemmingsrelaties gevisualiseerd kunnen worden.



Figuur 4 Voorbeeld van herkomst van reizigers met bestemming A geaggregeerd naar gemeenten (op basis van fictieve data)

4. Conclusie

Het breder verspreiden van geanonimiseerde ov-chipkaart biedt veel mogelijkheden. Hierbij moet bewust omgegaan worden met het feit dat anonimiseren van gegevens niet eenvoudig is. Ondanks dat de voorbeelden in dit artikel sterk vereenvoudigd zijn, zijn ze wel gestoeld op wat in de praktijk met moderne technieken mogelijk is. Telkens zal dan ook de afweging gemaakt moeten worden wat het gebruik van ov-chipkaart de maatschappij kost aan het inleveren van privacy, en het oplevert in het kader van het efficiënter maken of verbeteren van het openbaar vervoer. De overheid heeft hierin een belangrijke rol: zij als rol van opdrachtgever en grootste financierden van ov-chipkaartdata kan druk uitoefenen op vervoerders om mee te werken aan het publiceren van de gegevens. Het argument privacy hoeft hierbij, zoals dit artikel laat zien, niet in de weg te staan.

Referenties

Tsukada, Y et al. (2009). *Anonymity, privacy, onymity, and identity: A modal logic approach.* IEEE International Conference on Computational Science and Engineering.

CBP (2005). *Privacy en de ov-chipkaart. De visie van het College bescherming persoonsgegevens* (CBP). http://www.cbpweb.nl/Pages/uit_z2004-0850.aspx

Heerde, H.J.W. van (2010) *Privacy-aware data management by means of data degradation : making private data less sensitive over time*. University of Twente, CTIT Ph.D. thesis series.

Sweeney, L. (2002*). k-anonymity: a model for protecting privacy*. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

Connexxion (2014), *Privacy ov-chipkaart*, http://www.connexxion.nl/privacy-ov-chipkaart/577

Wikipedia-bijdragers (2014), Brenno de Winter, Wikipedia, de vrije encyclopedie. Opgehaald september 2014 van http://nl.wikipedia.org/w/index.php?title=Brenno_de_Winter&oldid=41283814