Inleidende Methodebeschrijving Pilot Mobiele telefoniedata

Behorend bij de technische handleiding "Estimating Hourly Population Flows"

Offermans, M.; Gootzen, Y.; de Jonge, E.; van der Laan, J.; Pijpers, F. ; Shah, S.; Tennekes, M.; de Wolf, P.-P.

Publication date
2021

Document Version
Final published version

CBS Publicatie

Auteur: May Offermans, Yvonne Gootzen, Edwin de Jonge, Jan van der Laan, Frank Pijpers, Shan Shah, Martijn Tennekes, Peter-Paul de Wolf.
Publicatiedatum: 12-1-2021 16:25

Inleidende Methodebeschrijving Pilot Mobiele telefoniedata

1. Inleiding

Uit geanonimiseerde geaggregeerde mobiele telefoondata kunnen indirecte schattingen gemaakt worden van het aantal personen dat in een gemeente aanwezig is en van het aantal bezoekers in deze gemeente. Dit is de conclusie die uit het pilot-onderzoek kan worden getrokken. Deze schattingen kunnen ook in officiële statistieken worden gebruikt.

Ook door EuroSTAT worden mobiele telefoniedata inmiddels gezien als een belangrijke bron om officiële statistieken te maken of te verbeteren voor toerisme, economie, milieu, drukte en mobiliteit. Dat is ook de reden dat het CBS sinds 2009 onderzoek doet naar het gebruik van geaggregeerde geanonimiseerde mobiele telefoniedata voor statistieken. Het CBS is daarin geen uitzondering. Overal in de wereld zijn organisaties als universiteiten, nationaal statistische bureaus, maar ook commerciële organisaties bezig met het ontwikkelen van methoden en software voor het analyseren en verwerken van telefoniegegevens voor statistiek, wetenschap en beleid.

Ruwe en onvoldoende geaggregeerde mobiele netwerkdata of verkeersgegevens zijn uiterst privacygevoelig. Bij het ontwikkelen van een verwerkingsmethode is niet alleen rekening gehouden met bestaande wetgeving die verlangt dat de data geanonimiseerd zijn. Er is van te voren ook rekening gehouden met een reeks van risico's en er zijn randvoorwaarden opgesteld om het onderzoek te kunnen doen. Deze worden in hoofdstuk 3 en 4 besproken. Hierin wordt beschreven hoe voorafgaand aan het onderzoek statistische beveiligingsmaatregelen in het ontwerp van de methode zijn geïntegreerd. Het toepassen van slechts één maatregel is niet voldoende. Daarom zorgt een combinatie van meerdere maatregelen als geheel voor de statistische beveiliging waarmee de privacy wordt gewaarborgd.

Deze notitie vormt een inleiding op het technische rapport [1]. Hierin is de wiskunde beschreven zoals die plaatsvindt bij het telecombedrijf. Het CBS zorgt als statistisch instituut voor onafhankelijkheid, continuïteit (garantie voor levering), internationale vergelijkbaarheid, kwaliteitskaders en transparantie van de algoritmen. D.w.z. de algoritmes en methoden waarop de statistieken zijn gebaseerd dienen voor 100% uitlegbaar en transparant te zijn. De rapportages van het onderzoek zijn daarom ook openbaar. Deze pilot is in 2019 afgesloten.

Dit rapport is uit een aantal hoofdstukken opgebouwd. In hoofdstuk 2 komen kort de basis principes van de methode aan bod en is globaal het verwerkingsproces beschreven. In hoofdstuk 3 worden alle stappen van het proces in detail beschreven. Tenslotte wordt in hoofdstuk 4 ingegaan op de privacyaspecten van de methode aan de hand van een aantal internationale publicaties waarbij er onthullingen van personen zijn gedaan uit geaggregeerde datasets met herkomst bestemmingsinformatie.

CBS Publicatie

Auteur: May Offermans, Yvonne Gootzen, Edwin de Jonge, Jan van der Laan, Frank Pijpers, Shan Shah, Martijn Tennekes, Peter-Paul de Wolf.
Publicatiedatum: 12-1-2021 16:25

Inleidende Methodebeschrijving Pilot Mobiele telefoniedata

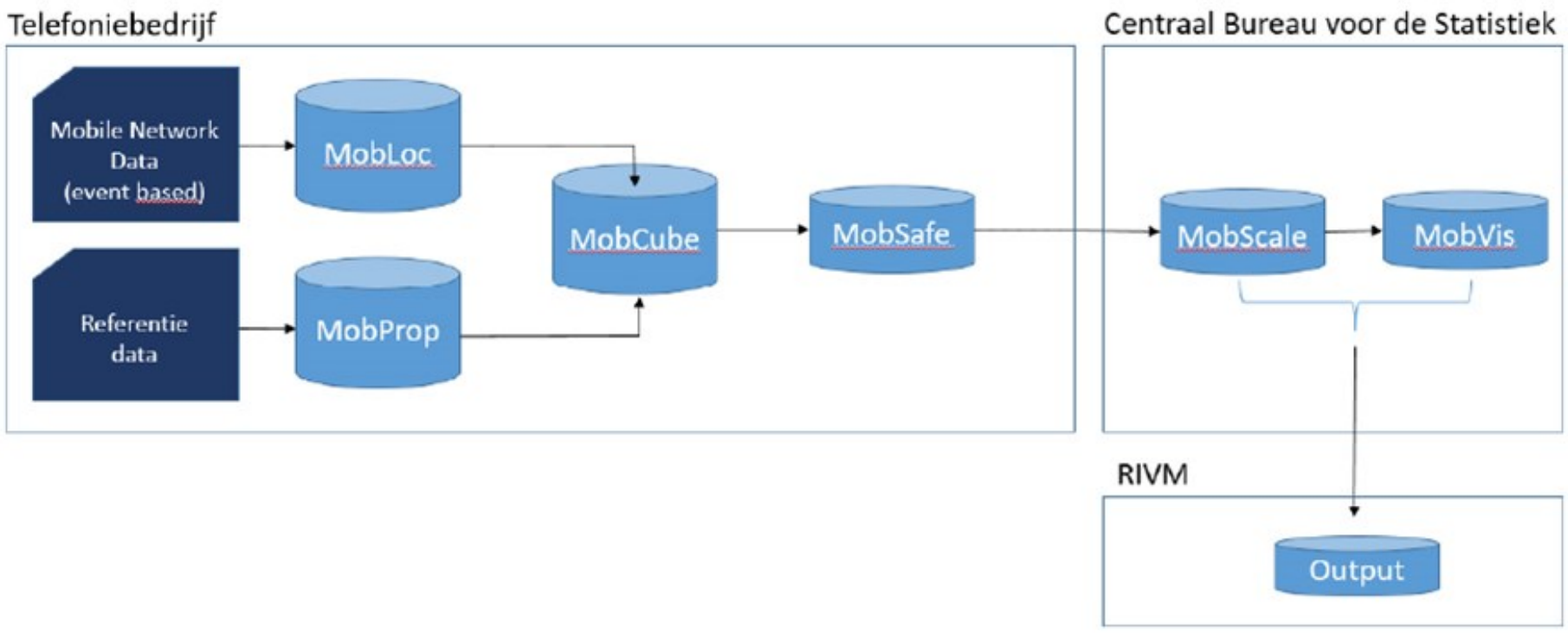
2. Basisprincipes van het proces

Het proces om signallingdata te verwerken tot statistische informatie bestaat uit 9 stappen die in hoofdstuk 3 in detail worden toegelicht. Het betreft een geautomatiseerde verwerking die uiteindelijk leidt tot *geaggregeerde geanonimiseerde informatie*.

De brondata worden niet rechtstreeks aan het CBS geleverd zoals bij andere statistische processen van het CBS gebruikelijk is, maar de data worden reeds bij het telefoonbedrijf geanonimiseerd en geaggregeerd. Het CBS krijgt daardoor geen herleidbare verkeersgegevens, maar alleen geanonimiseerde en geaggregeerde informatie uit deze data.

In figuur 2.1 zijn de software modules weergegeven die bij het proces met de 9 stappen horen. Deze zijn zodanig vormgegeven dat ze zoveel mogelijk aansluiten bij internationale standaarden die momenteel wereldwijd worden ontwikkeld. In hoofdstuk 3 wordt naar deze modules verwezen.

Figuur 2.1 Module-indeling voor software-ontwikkeling



CBS Publicatie

Auteur: May Offermans, Yvonne Gootzen, Edwin de Jonge, Jan van der Laan, Frank Pijpers, Shan Shah, Martijn Tennekes, Peter-Paul de Wolf.
Publicatiedatum: 12-1-2021 16:25

Inleidende Methodebeschrijving Pilot Mobiele telefoniedata

3. Methodebeschrijving in 9 stappen

3.1. Verwerkingsstappen die plaatsvinden bij het telefoonbedrijf

3.1.1 Stap 1 Beschrijving van brondata die al bij het telefoonbedrijf aanwezig zijn en pseudonimisering

Brondata

De brondata voor de methode bestaan uit *signalling* data. Dat zijn data over gebeurtenissen (hierna: events) die plaatsvinden tussen het toestel en het netwerk die worden vastgelegd in de telefooncentrale van het telecombedrijf. De *signalling* data worden gegenereerd doordat telefoons contact maken met het mobiele telefonienetwerk, bijvoorbeeld vanwege een telefoongesprek, een dataverbinding die wordt gestart, of een sms die wordt ontvangen. Dit wordt geregistreerd bij de telefooncentrale en voor maximaal 6 maanden bewaard. Het gaat niet om de inhoud van de communicatie maar het tijdstip, de duur en de *cell* waaraan dat betreffende event wordt toegekend. Een *cell* is een onderdeel van een antenne; een antenne bestrijkt een gebied dat bestaat uit één of enkele cellen.

Er zijn diverse typen van event-gebaseerde brondata, afhankelijk van welke gebeurtenissen zijn opgeslagen. Naast *signalling* data zijn dat bijvoorbeeld Event Data Records (EDR) en Call Detail Records (CDR). *Signalling* data bevatten de meeste gebeurtenissen. Vaak worden deze data apart gegenereerd voor 2G, 3G, 4G, en 5G netwerken. Dat maakt allemaal voor de methode niet uit. Het gaat er om dat er per toestel voldoende van dergelijke datapunten (dus events die worden vastgelegd) beschikbaar zijn. Hoe meer events er zijn hoe beter dat is voor de schattingen. Tabel 3.1 is geleend uit publicatie [2] en laat de verschillen in aantallen events zien tussen verschillende type data.

Tabel 3.1 Overzicht tussen verschillende type "event" gebaseerde mobiele telefoniedata [2]. MNO staat voor Mobile Network Operator; dit is het telefoonbedrijf

Data type	Availability in MNO	Approx. Avg number of records per subscriber per day
CDR outgoing	Always stored (billing)	2-3
CDR incoming	If MNO has decided to store	2-3
IPDR	If MNO has decided to store	0…200
LA update	If MNO has decided to store	12
Signalling	Requires additional network hardware installation + vast storage amounts (Hadoop HDFS)	200…2000

Van de *signalling* data worden uitsluitend een dubbel gepseudonimiseerd ID, de tijd en het antenne-ID (nummer van de gebruikte antenne) van elke actieve verbinding gebruikt als brondata voor de methode. Daarbij komen aansluitend statische referentiedata als het *cell*plan (waar de antennes staan) bij.

Overigens kan de methode ook overweg met Timing Advance data als speciale vorm van een event, maar deze vallen uitdrukkelijk buiten de pilot. Ook speciale locatiebepalingstechnieken zoals deze voor 4G en 5G netwerken door netwerkleveranciers worden aangeboden vallen buiten de scope van deze methode (en worden dus niet gebruikt).

Opslag en pseudonimisering van de brondata

De methode begint bij het verwerken en opslaan van de *signalling* data met betrekking tot abonnees en gebruikers van de communicatiediensten van het telefoonbedrijf. Dit is een bestaand proces bij het telefoonbedrijf. Alle vervolgstappen zijn hiervan afgeleid. Zoals hierboven reeds is aangegeven, is de wettelijke bewaartermijn van de *signalling* data 6 maanden.

Signalling data worden gespeudonimiseerd en voorzien van een identificatienummer. Dit is een bestaande interne beveiligingsprocedure. Het betreft een gebruikelijke interne handeling in dataverwerking die ook door andere dataverwerkers in andere sectoren dan telefonie gebruikt wordt om directe herleidbaarheid van personen binnen de organisatie tegen te gaan. Deze pseudonimisering zorgt ervoor dat alle direct herleidbare informatie uit de data wordt gehaald (telefoonnummer, IMSI etc.). De medewerkers van het telecombedrijf hebben getekend voor geheimhouding en mogen geen activiteiten ontplooien die leiden tot onthulling.

3.1.2 Stap 2 Tweede pseudonimisering

Als tweede stap pseudonimiseert het telefoonbedrijf de signallingdata opnieuw, waarbij de sleutels steeds veranderen en ook niet bewaard kunnen worden. Iedere 30 dagen wordt dit proces voor alle gepseudonimiseerde nummers herhaald. Dit gaat onder andere groepsonthulling (door onderzoekers zelf binnen de operator) tegen van groepen personen (vaak VIPS met beveiliging) waarvan de agenda publiek bekend is. Veel belangrijker is nog dat dit proces onomkeerbaar is. De medewerkers van het telecombedrijf die de data verwerken kunnen niet terug naar de oorspronkelijke data.

Verderop in het proces worden aanzienlijk meer waarborgen voor de privacy ingebouwd. Toch is ook in deze processtap al gekeken naar risico's van herleidbaarheid van de data binnen het telefoonbedrijf. Gekeken is naar de risico's op herleiding van de gespeudonimiseerde gegevens door de onderzoekers zelf. We zien twee risico's. Het eerste risico bestaat uit het kraken van de sleutel. Het tweede betreft het in het advies van de Article 29 Working Party (hierna: WP29) aangegeven risico op onthulling op basis van *signalling* data.

Het kraken van de sleutel is lastig en kost veel rekenkracht. Het herleiden van individuen door het schatten van de locatie uit de *signalling* data zoals aangegeven in het Europese advies van de WP29 is ook zeer moeilijk geworden doordat er een tweede pseudonimisering heeft plaatsgevonden waarvan geen sleutel beschikbaar is. In de literatuur wordt genoemd dat onthulling in dit geval gemakkelijk mogelijk is als men enkele locaties van een persoon kent. Die theorie is echter niet zomaar toe te passen op deze praktijk.

Stel dat iemand moedwillig een persoon (toestel) – tegen de regels in – wil onthullen op basis van geschatte geolocaties uit signalling data.

1. Ten eerste moet een sortering worden gedaan uit een dataset van 40 tot 200 miljard records per maand. Dat is zeer lastig, immers het IMSI of een andere variabele is niet bekend.
2. Om die sortering te doen op gepseudonimiseerde data binnen de operator moet extra informatie worden verzameld uit een agenda van iemand met minstens 4 nauwkeurige locaties per dag en daarbij ook het tijdstip. Een groot deel van de route moet vooraf bekend zijn om die herleiding te kunnen uitvoeren. Bekendheid van de werkplaats en woonplaats alleen is niet voldoende. In totaal zijn 4 locaties nodig om een slagingspercentage van 95% te bereiken.
3. Vervolgens moeten de bekende punten (plaats en tijd) worden vertaald naar een route van cellen en antennes die terug te vinden zijn in het bronbestand. Er zullen fout positieve matches zijn die leiden tot de verkeerde antenne en dus ook tot het verkeerde bijbehorende versleutelde en daarmee gepseudonimiseerde IMSI. Dat maakt het proces behoorlijk arbeidsintensief. Bovendien zal er extra software geschreven moeten worden. Dit is niet mogelijk met de beschikbare algoritmen vanuit het proces. In het hele proces wordt nergens gerekend met "routeinformatie". Er zal dus een illegaal geautomatiseerd proces moeten worden ingericht binnen de muren van het telefoonbedrijf om de route van een individu samen te stellen. Die route is niet vastgesteld op basis van precieze punten zoals dat bij GPS data kan, maar gebieden die verschillen per locatie. Deze exercitie is vele malen complexer en moeilijker dan in een "gewone" dataset zoals die van een survey.

Onze inschatting is dat een poging tot herleiding meer dan 4-5 dagen werk kost en opvallend veel rekenkracht binnen de IT-voorzieningen kost. Zeer weinig medewerkers van het telefoonbedrijf hebben toegang tot de originele signalling data. Zij hebben allen een geheimhoudingsverklaring getekend. Daarnaast is een samenspanning van deze medewerkers, de analisten bij het telefoonbedrijf en de medewerkers van de IT-security nodig. Immers, binnen de IT-voorziening van het telefoonbedrijf's zijn er beveiligingen die de inzet van een grote hoeveelheid rekenkracht signaleren en er is sprake van logging van handelingen van medewerkers.

Uiteraard is het inherente restrisico van moedwillige re-identificatie zeker niet nul is, maar deze is vele malen moeilijker dan in een reguliere dataset zoals belastingdienstgegevens of medische dossierinformatie. De moeite die onthulling kost in combinatie met de maatregelen die zijn genomen, leidt dan ook tot een zeer klein risico. Herleiding in andere meer traditionele datasets is eenvoudiger en dat komt veelal door de combinatie met unieke variabelen als leeftijd of geslacht. Deze ontbreken hier geheel.

3.1.3 Stap 3 Locatieschattingen maken (module MobLoc)

In het onderdeel MobLoc wordt de locatie geschat. Die locatie wordt aan de hand van het dekkingsgebied van een cel geschat met een Bayesiaans model. Om de geaggregeerde schattingen zo representatief mogelijk te krijgen worden de toestellen herverdeeld en verspreid. Er is dus geen sprake van een exacte bepaling zoals bij GPS-data, maar een schatting waarbij met kansen gerekend wordt. Het doel van deze stap is om een schatting te maken van aantallen toestellen per gemeente op basis van de gespeudonimiseerde signalling data. Hierbij wordt ook gebruik gemaakt van de antennekaart (*reference data*) en kan er publiek beschikbare hulpinformatie worden voorgesteld door het CBS om uitkomsten te verbeteren, zoals landgebruik en hoogtekaarten. Dit is statische informatie.

Het algoritme dat gebruikt wordt voor de locatieschatting werkt als volgt. Het land wordt opgedeeld in tegels van 100x100 meter. Dit zijn vakjes die (vergelijkbaar met pixels) goed te vertalen zijn naar de vloeiende grenzen van gemeenten.[1)] In het geval een toestel verbinding maakt met een *cell* worden voor de dichtbij gelegen tegels de kansen geschat dat het toestel zich in de desbetreffende tegel bevindt. Hierbij wordt gebruik gemaakt van data uit de antennekaart, zoals locatie van de antennes/cellen en ook de fysieke eigenschappen zoals hoogte en richtingshoek. Deze kansen worden vermenigvuldigd met prior kansen. Dit zijn genormaliseerde schattingen van het aantal toestellen per tegel die alleen gemaakt zijn op basis van hulpinformatie, zoals stedelijkheidsgraad. De geschatte kansen worden posterior kansen genoemd.

De *cell*-ID's uit de data worden gekoppeld aan deze posterior kansen, zodat per gemeente een schatting wordt gemaakt van het aantal toestellen. De locatie is dus niet heel precies: er wordt in deze stap statistische ruis toegevoegd. Een bijkomend effect van deze methode is dat aantallen toestellen worden "verspreid" in de ruimte en daarmee worden herverdeeld. Ook in de tijd is er sprake van een verspreiding. Het model werkt in een batch van een uur. Als er maar één meting is in een specifiek uur dan worden deze gegevens geëxtrapoleerd over andere minuten binnen dat uur. Er vindt geen spreiding plaats tussen de uren als tijdseenheid. De verspreiding in ruimte en tijd veroorzaakt een extra ruis die elk uur verandert van samenstelling. Dit betekent dat de relatie tussen de tellingen per gebied en de tellingen per mast een systeem van lineaire vergelijkingen is, met minder vergelijkingen dan onbekenden en daarom niet uniek inverteerbaar.

We illustreren dit met een voorbeeld. Stel dat we voor een toestel met nummer 1 in het uur 8:00–9.00 over gemeenten A, B en C de respectievelijke kansen 0,5, 0,3 en 0,2 verkrijgen. We achtten de kans dus groter dat het toestel zich in gemeente A bevond, maar laten deze kansverdeling zoals hij is – wij kennen hem niet definitief toe aan A. Stel dat wij verder voor een ander toestel 2 in hetzelfde uur een kansverdeling over gemeente A, B en C deze op 0, 1, en 0 schatten. Oftewel, wij vermoeden dat het toestel zich in dat uur volledig in gemeente B bevond. Dan tellen wij de kansverdelingen voor beide toestellen op: 0,5, 1,3 en 0,2. **Wij schatten dus dat zich 1,3 toestellen in gemeente B bevonden.**

3.1.4 Stap 4 Het toevoegen van de **woongemeente** (module MobProp(erties)

In deze stap wordt de woongemeente afgeleid. Een indicatie voor een woonverblijf is in deze pilot dat een telefoon het vaakst in de buurt is van een bepaalde antenne. Ook dit betreft geen pinpoint en maakt gebruik van informatie uit Mobloc, een schatting van de locatie. De "woongemeente" –tellingen worden tenslotte ook geschat. Ieder toestel krijgt een zogenaamde woonmast toegewezen. Dit is in het verwerkingsproces de zendmast waarmee het toestel het langst verbonden is geweest gedurende de observatieperiode. De massa van het toestel wordt vervolgens verspreid over de gemeenten die in het bereik van de woonmast liggen. Over het algemeen wordt dus aan een toestel niet één enkele woongemeente toegekend – het gaat om een kansverdeling over één of meer gemeenten.

3.1.5 Stap 5 Automatische dataverwerking en indikking (module MobCube)

Indikking van de dataset vindt plaats door een selectie te maken van een gebied en een tijdsperiode en in MobCube te verwerken. Deze laatste stap is **onomkeerbaar** omdat er zeer veel informatie vernietigd wordt (80-90%). Na deze stap is er alleen nog sprake van een statistische geaggregeerde telling. **In feite is één geschat toestel al een vorm van een aggregaat.**

De data uit MobLoc en MobProp worden geautomatiseerd verwerkt en geaggregeerd tot statistische informatie. Feitelijk komt er een telling tot stand. Een voorbeeld van een dergelijke telling is beschreven in figuur 3. Als een persoon, woonachtig is in Maastricht en om 12 uur in Utrecht is, dan wordt deze geteld in de tabelcel (kolom Woonplaats) Maastricht en (rij) Utrecht. Als dezelfde persoon om 11 uur in Eindhoven is, dan wordt deze geteld in de (kolom Woonplaats) Maastricht en de (rij) Eindhoven. Er is geen koppeling tussen de verschillende uurtellingen en er kan dus met deze uitkomsten geen route (Maastricht-Eindhoven-Utrecht) berekend worden.

Figuur 3.1 Format van de output

Tijdstip: 12:00

Tijdstip: 11:00

Tijdstip: 10:00

Tijdstip 12:00		Afkomstig uit woongemeente		
		Eindhoven	Utrecht	Maastricht
Aanwezig in gemeente	Eindhoven	333.050	33.400	44.400
	Utrecht	333.250	100.000	200
	Maastricht	55.300	240.000	399.300

3.1.6 Stap 6 Outputcontrole op geanonimiseerde data en verzending (module MobSafe)

Na MobCube zitten er vrijwel geen herleidbare gegevens meer in de dataset. Wel kan het voorkomen dat sommige tabelcellen zeer kleine aantallen tellingen bevatten, bijvoorbeeld het aantal personen dat in een kleine gemeente woont en aanwezig is in een andere gemeente die niet in de buurt ligt. Om elk risico op onthulling te vermijden vindt er daarom een definitieve statistische beveiliging in MobSafe plaats waarbij alle output met een celvulling lager dan 15 toestellen definitief wordt verwijderd. Deze tabelcellen blijven leeg.

In deze stap gaat dus ook weer informatie verloren. Het telefoonbedrijf doet uiteindelijk ook handmatig een controle om te zien of het proces goed is uitgevoerd. De uitkomst van dit proces bij het telefoonbedrijf betreft geaggregeerde en geanonimiseerde informatie die aan het CBS wordt geleverd. Qua format is dit dezelfde tabel als in figuur 3 is beschreven. Alleen is door de N>15 regel wederom een groot deel van de data vernietigd. Dat is ook in deze tabel meer dan 80% van de cellen het geval (zie ook hoofdstuk 4). Deze krijgen een missing value (een puntje of een streepje). Het resultaat van de in stap 1 t/m 6 toegepaste techniek is permanent. Door alle genomen maatregelen bestaat geen mogelijkheid tot deduceerbaarheid. Ook koppeling met externe datasets zou niet tot onthulling kunnen leiden. Het herleiden van de geaggregeerde en geanonimiseerde informatie tot individuele persoonsgegevens uit deze tabel is dan ook onmogelijk.

De gecontroleerde geanonimiseerde geaggregeerde data gaan over een beveiligde verbinding naar het CBS. De informatie wordt als uiterst bedrijfsgevoelig behandeld, omdat deze op gemeenteniveau inzicht geeft in de verdeling van het aantal abonnees van een individueel telefoonbedrijf over het land. De informatie wordt daarom versleuteld alvorens deze te verzenden.

3.2 Verdere verwerking bij het CBS

3.2.1 Stap 7 Correctie op data (MobCal(ibration)

Bij het CBS vindt een correctie plaats om van de verkregen gegevens representatieve gegevens te maken over de Nederlandse bevolking. Doel is immers om iets te zeggen over de bevolking en niet over toestellen of de populatie van de telefoonbedrijven. Die correctie vindt plaats op basis van registerdata uit de Basisregistratie Personen (BRP) volgens een weegmodel. Ook kan de data op gemeenteniveau (dus niet persoonsniveau) worden verrijkt met andere CBS data, zoals het vakantieonderzoek of andere registerinformatie.

3.2.2 Stap 8 Verwerking naar relatieve waarden

Er komt na de weging uit stap 7 een aangepaste telling tot stand. De tellingen zijn in deze pilot nog niet gevalideerd. Dat wil zeggen dat onzeker is of de aantallen de juiste aantallen zijn voor alle gemeenten. Hierop is nog geen controle gedaan. Voor de visualisatie in het beta-product worden de waarden daarom vertaald naar relatieve verschillen. Dat wil zeggen dat de aantallen worden vertaald naar percentages van en naar een gemeente. Bijvoorbeeld om 12:00 uur is 8% van de bewoners uit Amsterdam vertrokken naar Almere. Dit heeft vooral te maken met de kwaliteit van de uitkomsten maar is ook te zien als extra maatregel tegen groepsonthulling met informatie van buiten. Het format blijft gelijk als in figuur 3 beschreven, maar het betreft dus ratio's/percentages die in de visualisatie kunnen worden afgelezen straks (MobVis).

3.2.3 Stap 9 Publicatie

De uitkomsten zijn openbaar gepubliceerd in een nieuwe visualisatie (MobVis) die speciaal hiervoor ontwikkeld is.

[1] De ouput aan het einde is op gemeenteniveau.

CBS Publicatie

Auteur: May Offermans, Yvonne Gootzen, Edwin de Jonge, Jan van der Laan, Frank Pijpers, Shan Shah, Martijn Tennekes, Peter-Paul de Wolf.
Publicatiedatum: 12-1-2021 16:25

Inleidende Methodebeschrijving Pilot Mobiele telefoniedata

4. Privacy

De Telecomwet vereist dat de gegevens die het telecombedrijf verlaten anoniem zijn. In dat kader is in een advies van de AP verwezen naar een wetenschappelijk artikel [2] dat een methode beschrijft om uit geaggregeerde telefoniedata de trajecten te herleiden van mobiele telefoons. In dit hoofdstuk wordt dieper op de in dit artikel beschreven methode ingegaan en wordt de vraag beantwoord in hoeverre dit gevolgen heeft voor de anonimiteit van de data.

Uitgangspunt is de dataset zoals deze vanaf de telefoonprovider geleverd wordt aan het CBS. Dit zijn geaggregeerde data met daarin geschatte tellingen hoeveel toestellen zich in een gegeven uur vanuit een gegeven woongemeente zich in een gegeven gemeente bevinden (de zogenaamde flowkubus). Dit hoofdstuk is als volgt ingedeeld:

1. In paragraaf 4.1 wordt het inverse probleem uitgelegd en hoe dat in het proces veroorzaakt wordt. Daarmee wordt gekeken in hoeverre het technisch wiskundig mogelijk is om op basis van de laatste kennis, ondanks alle maatregelen rechtstreeks terug te rekenen tot een individu. Hierin wordt de anonimiteit wiskundig aangetoond. In dat kader is in een advies van de AP verwezen naar een wetenschappelijke artikelen [3] [4] uit China dat een methode beschrijft om uit geaggregeerde telefoniedata de trajecten te herleiden van mobiele telefoons. De in het artikel aangehaalde methode om uitkomsten te ''kraken" tot herleidbare uitkomsten is vrij nieuw maar loopt ook tegen dit inverse probleem aan. Het artikel beschrijft een complexe methode die de nodige wiskundige analyse vereist om deze goed te kunnen duiden. In paragraaf 4.1.3 wordt als aanvulling op de wiskundige benadering ook dieper op deze Chinese artikelen ingegaan en waarom deze niet kunnen werken op de methode die het CBS heeft gebruikt in de pilot.
2. In paragraaf 4.2 wordt ingegaan op het hypothetisch geval waarbij aangenomen wordt dat de aanvaller (dit kan binnen het CBS zelf zijn, of een externe hacker met toegang tot data van bijvoorbeeld een "*big tech*" bedrijf) beschikt over een zeer grote hoeveelheid informatie (het meest extreme scenario) is het dan mogelijk om toch de locatie van het toestel van iemand te bepalen (d.m.v. een flowkubus)? Deze aanpak wordt in paragraaf 4.2 beschreven. Dit betreft een geavanceerde aanval die verder gaat dan het wetenschappelijk artikel uit China uit paragraaf 4.1.
3. In paragraaf 4.3 wordt tenslotte het fenomeen groepsonthulling beschreven. Het betreft de mogelijkheid om groepen te kunnen identificeren uit de output data en welke waarborgen er worden genomen om het risico hierop zo veel mogelijk te beperken. Strikt genomen heeft dat niets met anonimiteit van het individu te maken. Het monitoren van groepen is juist een belangrijke toepassing voor de bestrijding van COVID. Toch is het belangrijk ook hier aandacht aan te besteden.

4.1 Wiskundige basisprincipes bij de anonimiteit van de CBS-methode

4.1.1 Het klassieke inverse probleem

Het wiskundige grondprincipe dat het CBS hanteert bij databeveiliging komt uit het vakgebied van algebra.

> *Wanneer een consistent systeem van onafhankelijke lineaire vergelijkingen over de reële getallen een kleiner aantal vergelijkingen heeft dan onbekenden, dan is de oplossingsverzameling oneindig groot.*

Bijvoorbeeld als men een optelling doet van 5 getallen en alleen de som publiceert kan men onmogelijk de oorspronkelijke getallen achterhalen. Dit principe is de kern waarom de gegevens die door de mobiele telefoonbedrijven verstrekt worden anoniem zijn. Er kan dus wiskundig gezien niet worden teruggerekend, met welke methode of techniek dan ook.

De relatie tussen individuele objecten, zoals mobiele telefoons of vergelijkbare elektronische toestellen, en een meetgegeven als een telling van die objecten is een wiskundige vergelijking. De belangrijkste en onomkeerbare procedure in iedere deelstap van een verwerkingsproces is daarom om bewust bepaalde tellingen niet uit te voeren. Deze onomkeerbare informatievernietiging garandeert dat geen van de deelstappen omgedraaid kan worden om naar een individueel device te leiden. Dat geldt dus a fortiori voor het totale proces: van signalling data naar de geaggregeerde geanonimiseerde dataset die door het telecombedrijf wordt verstrekt aan het CBS.

4.1.2 Het inverse probleem dat ontstaat in het proces

Voor een precieze en volledige beschrijving van de processtappen wordt verwezen naar hoofdstuk 3 en de technische beschrijving [1]. In deze paragraaf wordt ingegaan op een aantal specifieke stappen die belangrijk zijn in het proces om te komen tot anonieme telefoondata.

1. Stap 1 t/m 6: De verwerking door het telecombedrijf (die plaatsvindt binnen de infrastructuur van het telecombedrijf)

Iedere 30 dagen worden alle gepseudonimeerde data bij de operator opnieuw gespeudonimiseerd waarbij er geen sleutel wordt gegenereerd (en dus ook niet bewaard kan worden). De dubbel gepseudonimiseerde data bevatten koppelingen tussen een mast en een device op een tijdstip van uitwisseling van een contactsignaal of data. De koppelingen tussen mast en device worden nog bewerkt. Daar zijn twee redenen voor:

1. Een inhoudelijke reden. Tussen welke antenne en device er op een tijdstip een koppeling wordt gemaakt, is niet uitsluitend afhankelijk van fysieke nabijheid, maar hangt ook af van signaalsterkte die niet uniform is rond een antenne, en beschikbaarheid van verwerkingscapaciteit van signalen door de antenne. Binnen een tijdsinterval van een uur wisselen toestellen vaak en in willekeurige volgorde met meerdere masten signalen uit. Het is daarom statistisch meer zuiver om ieder van de individuele mast-device koppelingen gewogen te verdelen over een aantal masten.
2. Een beveiligingsreden. Doordat koppelingen opgedeeld worden over een aantal antennes, en vervolgens aantallen fractionele koppelingen op te tellen per mast wordt de ***1-op-1 relatie tussen een individuele mast en device onherstelbaar verstoord.***

De optelling per mast moet altijd gewogen worden verdeeld over gemeenten, omdat het gevoeligheidsgebied van iedere mast meerdere gemeenten bestrijkt. Voor deze data geldt dus dat er geen unieke, omkeerbare, relatie bestaat tussen de tellingen per gebied of mast en de aantallen betrokken toestellen per gebied of mast: er zijn meer onbekenden dan er meetgegevens zijn en dus geldt het wiskundige grondprincipe dat de onbekenden niet meer kunnen worden bepaald. Er is bewust informatie vernietigd om identificatie te voorkomen. In hoofdstuk 3 staat de exacte werking van het algoritme van de locatietelling vermeld als volgt:

- *"….De locatie is dus niet heel precies: er wordt in deze stap statistische ruis toegevoegd. Een bijkomend effect van deze methode is dat aantallen toestellen worden "verspreid" in de ruimte en daarmee worden herverdeeld. Ook in de tijd is er sprake van een verspreiding. Het model werkt in een batch van een uur. Als er maar één meting is in een specifiek uur dan worden deze gegevens geëxtrapoleerd over andere minuten binnen dat uur. Er vindt geen spreiding plaats tussen de uren als tijdseenheid. De verspreiding in ruimte en tijd veroorzaakt een extra ruis die elk uur verandert van samenstelling.* ***Dit betekent dat de relatie tussen de tellingen per gebied en de tellingen per mast een systeem van lineaire vergelijkingen is, met minder vergelijkingen dan onbekenden en daarom niet uniek inverteerbaar.***
- *We illustreren dit met een voorbeeld. Stel dat we voor een toestel met nummer 1 in het uur 8:00—9.00 over gemeenten A, B en C de respectievelijke kansen 0,5, 0,3 en 0,2 verkrijgen. We achtten de kans dus groter dat het toestel zich in gemeente A bevond, maar laten deze kansverdeling zoals hij is – wij kennen hem niet definitief toe aan A. Stel dat wij verder voor een ander toestel 2 in hetzelfde uur een kansverdeling over gemeente A, B en C deze op 0, 1, en 0 schatten. Oftewel, wij vermoeden dat het toestel zich in dat uur volledig in gemeente B bevond. Dan tellen wij de kansverdelingen voor beide toestellen op: 0,5, 1,3 en 0,2. Wij schatten dus dat zich 1,3 toestellen in gemeente B bevonden."*

Aanvullend dient gemeld te worden dat het bereik van antennes sterk varieert, van 5 meter tot 60 kilometer over gemeentegrenzen heen. Ook de hoek kan variëren van 2 graden tot 360 graden. Zoals voorgeschreven hanteert het telecombedrijf de regel dat wanneer de tellingen in een bepaald gebied binnen een bepaald tijdvak onder de ondergrens van 15 vallen, die data onderdrukt en daarmee vernietigd wordt. Het totale aantal toestellen per tijdsinterval in de uiteindelijke dataset die naar het CBS komt fluctueert sterk. Immers, als een toestel in een uur geen verbinding heeft gemaakt met het netwerk voor een telefoongesprek of sms bestaat voor dat toestel geen registratie voor dat uur, en wordt dat toestel in dat uur niet meegeteld in de dataset die aan het CBS wordt verstrekt. De populatie is dus niet constant in grootte en al helemaal niet in samenstelling. De fluctuaties worden nog versterkt door de onderdrukking van kleine tellingen zoals hierboven genoemd. Het is belangrijk dit hier te vermelden omdat het voor bijvoorbeeld het paper[3] een essentiële vereiste is dat de populatie van toestellen constant is in samenstelling en grootte.

Bij binnenkomst van deze tabellen bij het CBS zijn er dus verschillende onomkeerbare handelingen verricht die informatie vernietigd hebben, zodat er zelfs met de data en resources die het CBS zelf ter beschikking heeft geen identificatie meer mogelijk is. De data die het CBS ontvangt zijn dus anoniem.

Stappen 7-9 Statistische output

Het aantal verschillende tellingen dat wordt gemaakt (per gemeente) heeft een ingewikkelde en niet omkeerbare relatie met de aantallen getelde toestellen. Het aantal onbekenden is weer groter dan het aantal vergelijkingen: de tellingen op gemeenteniveau. De "standaard" CBS statistische beveiliging voor CBS publicaties wordt hier nogmaals toegepast. Als extra maatregel zijn de te publiceren aantallen aangepast naar relatieve waarden. Er wordt dus in deze stap opnieuw en bewust informatie vernietigd om het onmogelijk te maken voor een externe partij om zelfs maar tellingen per mast te reconstrueren, laat staan om bewegingen van individuen te herleiden.

4.1.3 Bespreking academische publicaties

Op basis van wat in 4.1.1 en 4.1.2 beschreven is identificatie onmogelijk omdat dit een schending is van de genoemde wiskundige grondprincipe. Behalve het in de inleiding geciteerde paper[3] zijn er ook andere papers[4],[5] gepubliceerd waarin technieken worden besproken die zich erop richten om uit het pad dat een device volgt iets te kunnen onthullen over de drager of eigenaar van dat device. Er is beduidende academische interesse in dit onderwerp en er zijn dus veel meer papers op dit gebied die al gepubliceerd zijn of gepubliceerd zullen worden. Het is niet haalbaar om alle bestaande literatuur op dit gebied individueel te bespreken.

Echter, al dergelijke methoden moeten, om te kunnen werken, toegepast kunnen worden op data in een proces waar het eerder genoemde wiskundige grondprincipe expliciet of impliciet niet van toepassing is. Met andere woorden, direct terugrekenen om te onthullen kan niet. Deze papers zijn veelal gericht op benadering van het individu met een zekere waarschijnlijkheid.

De meeste methoden blijken niet toepasbaar op uitkomsten zoals die volgens de hiervoor beschreven werkwijze tot stand komen. Het ontwerp van de methode maakt een aantal kraakmethoden daardoor niet toepasbaar. In een deel van de gepubliceerde papers wordt aangenomen dat er 'tracks' beschikbaar zijn waarin de opeenvolgende masten waarmee toestellen in contact treden worden opgeslagen en als geheel verwerkt.

De fundamentele aanname voor die methodes is dat de 1-op-1 relatie tussen device en mast of gebied dus behouden blijft. Dit is in strijd met de door de telecomaanbieder in stap 1 (en stap 3) gehanteerde methode. De conclusies van elk paper dat gebruik maakt van dergelijke tracks kunnen daarom niet toegepast worden op de methode die door het CBS wordt gebruikt. De module MobLoc is hiervoor essentieel. Daarnaast blijft het noodzakelijk voor dit soort methoden om externe individueel identificerende informatie te gebruiken zodat een bepaald track van een device aan een persoon verbonden kan worden.

De door de CBS-methode gegenereerde data is van een heel ander type dan de data in het paper [3]. De belangrijkste aanname in het paper [1] is dat in opeenvolgende tijdsintervallen er identiek 1-op-1 dezelfde toestellen in de dataset zitten, waar bovendien ieder device 1-op-1 aan een mast gekoppeld is. Het 1-op-1 gekoppeld zijn van mast en device wordt in stap 1 van het hierboven besproken proces expliciet voorkomen.

Het is onduidelijk of de in het paper beschreven procedure überhaupt hierop aan te passen is. Hiervoor is aanvullend onderzoek nodig. In dit artikel wordt een relaxatie voor het algoritme toegepast op basis van een aantal aannames om de berekening mogelijk te maken. Het is waarschijnlijk dat een relaxatie voor de situatie met miljoenen toestellen die moet leiden tot een aanvaardbare rekentijd zal leiden tot een verlaging van de nauwkeurigheid waardoor er meer "onjuiste" routes gevonden zullen worden. Ook al is de rekenkracht aanwezig, dan nog kan de gewenste nauwkeurigheid die moet leiden tot identificatie niet bereikt worden.

Het eerstgenoemde paper[3] gebruikt ook tracks, maar neemt die niet als startpunt. In plaats daarvan worden tellingen per mast in opeenvolgende tijdsintervallen aan elkaar gekoppeld zodat tracks geconstrueerd kunnen worden. Hierbij worden modelaannames gebruikt voor typische bewegingen van toestellen (patronen). Immers, van verplaatsingsgedrag is bekend dat mensen vaak dezelfde patronen hebben. Om dit probleem op te lossen wordt het zogenaamde 'Hongaars algoritme' gebruikt. Er zijn meerdere aannames inherent aan de methode zoals beschreven in dat paper, waardoor de conclusies niet van toepassing kunnen zijn op de processtappen die zijn gevolgd door het CBS. Het aantal toestellen in het paper[3] is vele malen kleiner dan bij de data op grond van de analyses die in het kader van pilot voor het betaproduct zijn gedaan (100 duizend versus 21 miljoen) terwijl het aantal locaties waarvoor aggregaten bepaald worden in dat paper vele malen groter is dan de data (uitsluitend op gemeenteniveau) die aan het CBS zal worden verstrekt (8,600 versus 400). Het verschil tussen het aantal vergelijkingen en het aantal onbekenden is daarom voor de data die aan het CBS wordt verstrekt nog vele malen groter dan het geval is in het paper. Dat zou neerkomen op een orde van grootte van $1*10^{18}$. Dat maakt de verwerking van deze procedure met de huidige beschikbare computerkracht van het CBS onmogelijk. Zover bekend is het uitvoeren van de in het artikel beschreven procedure op een dergelijke dataset slechts door een viertal supercomputers in de wereld mogelijk (exa-scale computing).

Laten we kijken naar het volgende gedachtenexperiment behorende bij de kraakmethode uit het artikel van Tu [3] Daarbij nemen we voor het gemak even aan dat een toestel zich gedurende een uur slechts in één gemeente kan bevinden[2)]. Beschouw de werkelijke trajecten van twee toestellen met dezelfde thuisgemeente, en wijzig deze door voor een bepaald uur de twee gemeentes in hun trajecten [3] te verwisselen. Dan zijn de geaggregeerde flowkubussen[3)] voor en na de verwisseling identiek. Dit eenvoudige voorbeeld toont al aan dat meerdere trajectendatasets na aggregatie tot dezelfde flowkubus kunnen leiden. Uit dit argument is echter niet op te maken of, en in welke mate de flowkubus kan bijdragen aan identificatie van individuen. Het zou namelijk kunnen dat sommige combinaties van paden plausibeler zijn dan andere. Deze plausibiliteitsinformatie zou een aanvaller kunnen toevoegen aan het probleem om een betere schatting van de oorspronkelijke paden te maken. Dus stel dat een aanvaller beschikt over extra data (zoals algemene mobiliteitspatronen, of data van een specifiek individu) naast de flowkubus. In hoeverre is het dan mogelijk om delen van de vergelijkingen toch op te lossen?

De aanname dat alle toestellen op ieder moment in de data aanwezig zijn is een geïdealiseerde situatie die in echte datasets, zoals de aan het CBS te verstrekken dataset, niet optreedt. Dit is echter wel cruciaal voor het Hongaars algoritme (een optimalisatiemethode) als in [3] beschreven om te kunnen werken. Zodra er zelfs maar 1 device in een tijdsinterval verdwijnt of verschijnt kan dat algoritme niet geoperationaliseerd worden. In het paper[3] worden daarom toestellen geheel weg gelaten uit de populatie of wordt er geïnterpoleerd. De in het paper[3] beschreven academische methode is om die reden niet toepasbaar op de dataset die het CBS in de pilot heeft verwerkt.

Ook wordt in het paper[3] aangegeven dat extra informatie nodig is om een gereconstrueerd traject aan een herkenbaar individu te koppelen. In het paper is een gereconstrueerd traject “unique” als het op basis van n toplocaties te koppelen is aan een uniek traject binnen de “ground truth”. Vervolgens wordt er van uit gegaan dat een uniek traject binnen de “ground truth” aan een uniek persoon in de populatie te koppelen is. Voor die laatste stap is dus extra informatie nodig, die ook nog eens uniek aan een herkenbaar persoon gekoppeld moet kunnen worden.

In het algemeen kan gesteld worden dat ieder algoritme of methode dat een uniek device kan identificeren, er vanuit moet gaan dat iedere stap in het verwerkingsproces van de ruwe telecomdata uniek inverteerbaar is. Het wiskundige grondprincipe zorgt er juist voor dat iedere verwerkingsstap niet inverteerbaar is, en onthulling dus niet slechts onwaarschijnlijk of zeer rekenintensief, maar wiskundig onmogelijk is. De vervolgvraag is dan of een aanvaller met extra informatie toch niet de waarheid kan benaderen. In paragraaf 4.2 wordt hier dieper op in gegaan d.m.v. een ander extreem aanvalsscenario.

Voor de volledigheid zijn er nog onderstaande overige kritische punten bij dit artikel [3]:

1. Door uit te gaan van een N>15 vervallen alle unieke routes die men snel zou kunnen samenstellen op basis van lage aantallen uit de data.
   Hoe uniek een route is wordt in het artikel niet genoemd, het beschrijft alleen hoe groot het percentage gevonden unieke trajecten is op basis van de TopK (meest bezochte) locaties *“...uniqueness measures the possibility that he can uniquely distinguish victims' recovered trajectories. Therefore, we define the uniqueness as the percentage of recovered trajectories that can be uniquely distinguished by their most frequent k locations (Top-K)”*. Vervolgens is externe informatie nodig om dat gevonden traject te koppelen aan een individu; *“Therefore, the results indicate that the recovered trajectories are very unique and vulnerable to be reidentified with little external information.”* Welke externe informatie nodig is, wordt niet genoemd, maar het is aannemelijk dat ze bedoelen dat je weet welke TopK locaties bij een persoon horen. Of het daadwerkelijk de juiste persoon is (is die persoon de enige met die TopK locaties) wordt niet bekeken, omdat men aanneemt dat een persoon/toestel kan worden geïdentificeerd door een unieke TopK . Dus een persoon met een TopK aantal locaties klopt met de route, maar dat mag je niet omkeren naar de opmerking dat je die aan een uniek persoon kunt koppelen. Juist ook omdat in de hoofdstuk 3 beschreven methoden de gebieden erg groot zijn (en geen cellen). De vraag is dan ook of die koppeling de juiste is.
2. In de studie worden cellen (“area covered by base station”) gebruikt als geografische afbakening. Wij gebruiken gemeenten in plaats van gebiedscellen. Dat maakt de koppeling van een route aan een toestel veel moeilijker. In het artikel wordt gesproken over 8.000 base stations. Dit betekent dat er 511.808.016.000 unieke varianten van een top 3 van meest bezochte base stations mogelijk zijn. In onze dataset met 355 gemeenten zijn er 44.361.510 unieke varianten van een top 3 van meest bezochte gemeenten van elk persoon. Dit is een factor van ongeveer 11.500 lager om een top 3 aan een toestel te koppelen. Het kleinere aantal unieke varianten van een top 3 wordt in ons geval ook nog eens over 2 tot 4 miljoen toestellen verdeeld in plaats van over 100 duizend toestellen. Daarmee wordt de “uniciteit” van een top 3 combinatie (“being different from others” in het artikel genoemd) 200.000 tot 400.000 keer zo klein.

De "recovery accuracy" zal vele malen lager zijn doordat niet 100.000 maar 4 miljoen devices in de dataset zitten (zie figuur 11 a voor afname accuracy bij toename van aantal toestellen). Daardoor is de kans groot dat indien een traject met de correcte TopK locaties aan een persoon gekoppeld kan worden de route niet juist is. Daarmee wordt de kennis over die persoon dus niet verrijkt.

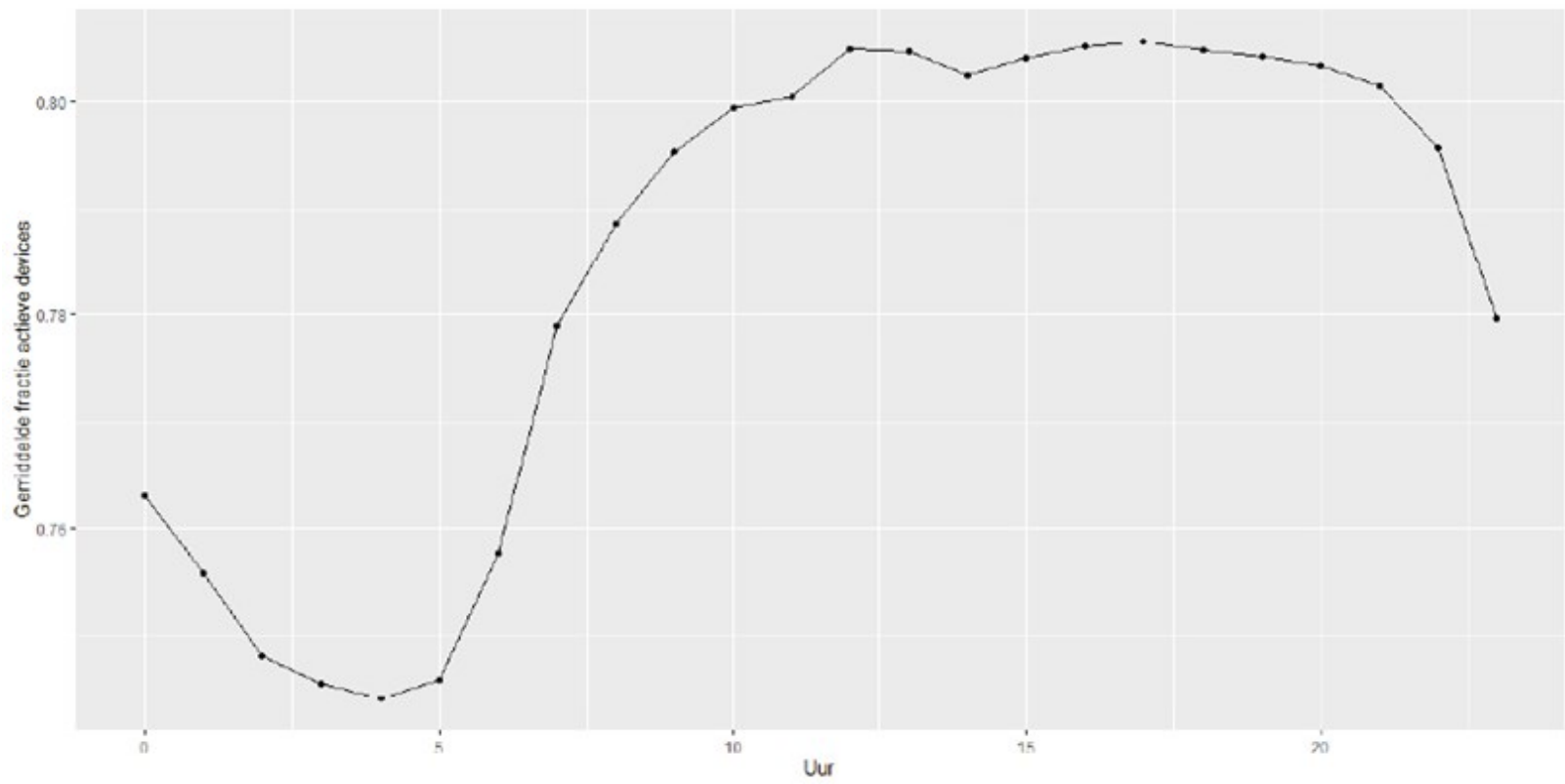
4.2 Aanvalsscenario

In paragraaf 4.1 is aangetoond dat het oplossen van vergelijkingen niet werkt om tot onthulling te komen. Wat als iemand, een aanvaller of hacker over extra aanvullende informatiebronnen kan beschikken? Wat zijn dan de gevolgen? Stel, de aanvaller weet voor een bepaalde afgeleide herkomstgemeente van alle toestellen van een telecombedrijf op één na waar ze zijn gedurende periode t (van één uur). In de periodes daarvoor en daarna weet de aanvaller ook van dit ene toestel waar het is. Het doel van de aanvaller is om te weten te komen waar het toestel gedurende periode t was. De aanvaller weet de locatie van het device in t-1 en in t+1, en gezien de maximale reissnelheid van het toestel blijven er een aantal regio’s over waarin het toestel in de tussentijd geweest kan zijn. De vraag is nu in hoeverre de flowkubus de aanvaller kan helpen met dit probleem. Laten we aannemen dat er op tijdstip t sprake is van meer dan één mogelijke regio aangezien de aanvaller anders de flowkubus sowieso niet nodig heeft. Die data onthult dan niets wat de aanvaller al niet wist.

Aangezien de aanvaller de locaties van alle andere toestellen weet zou de locatie simpel te vinden moeten zijn: hij neemt de aantallen uit de flowkubus; trekt de bekende aantallen ervan af en als het goed is, is er één regio over waar een verschil van één overblijft. Er zijn echter een aantal eigenschappen van de signalling data en de methodologie waarmee de flowkubus is geproduceerd waardoor deze rekensom veel lastiger is om te maken dan men in eerste instantie zou denken. Dit heeft de volgende oorzaken:

1. Niet alle toestellen zijn altijd actief. Dus hoewel de locaties van alle toestellen te achterhalen zijn op basis van de brondataset, is niet bekend welke daarvan ook daadwerkelijk actief waren gedurende periode t. Voor de zekerheid wordt aangenomen dat het toestel waarvan de locatie gezocht wordt wel actief was en dat dat bekend is bij de aanvaller. Zoals in figuur 4.1 te zien is, is gedurende de dag per uur circa 20-25 procent van de toestellen niet actief. Gemiddeld is 78% om 7:00 in de ochtend actief.

Figuur 4.1 Gemiddelde fractie actieve toestellen als functie van uur van de dag



2.

3. De aanvaller weet waar de apparaten zich bevinden, maar weet niet zeker aan welk gebied deapparaten zijn toegewezen. Toewijzing van apparaten aan gebieden gebeurt via de zendmast[4)]. Voor veel zendmasten geldt dat ze in meerdere gebieden dekking bieden. Met behulp van een model wordt voor een gegeven zendmast één toestel verdeeld over alle gebieden waar de zendmast dekking geeft (de fracties tellen op tot één). Het is de aanvaller ook niet bekend met welke zendmast een toestel verbonden is geweest (als hij al verbonden was; zie vorig punt).

Voor veel zendmasten geldt dat een connectie met deze zendmast vervolgens ook niet 1 op 1 vertaald kan worden naar een gemeente. Dus zelfs als de aanvaller gegevens heeft over de werkelijke locatie van elk toestel, bestaat er onzekerheid over in welke gemeente dit toestel wordt opgenomen door de schattingsmethode. Wanneer een zendmast meerdere gemeenten bedient dan wordt een toestel in gewogen fracties meegeteld in alle gebieden waar de zendmast dekking geeft. Een waarde van 100 toestellen in de data kan bijvoorbeeld dus bestaan uit 200 maal een bijdrage van 0.5.

Van regio's waarin minder dan 15 toestellen geschat worden is de waarde niet aanwezig in de flowkubus. Er zijn dus heel veel regio's waarvoor de som sowieso niet uitgerekend kan worden. Het aantal cellen dat wordt onderdrukt is wel sterk tijdstip afhankelijk. 's Nachts gaat het om veel meer tabelcellen maar wel veel minder toestellen en vice versa.

Ieder van bovenstaande punten maakt het onmogelijk voor de aanvaller om de locatie van het toestel met redelijke zekerheid te bepalen. Stel er worden in een regio 0 (nul) toestellen geschat. Circa 80% van de toestellen wordt waargenomen. Het aantal waargenomen toestellen zal daarom een binomiale verdeling volgen (met p = 0.8). Dit betekent dat er met 90% waarschijnlijkheid zich ook één toestel in die regio kan bevinden. Als er meer toestellen worden geschat wordt het waarschijnlijkheidsinterval alleen nog maar groter. Bij 10 waargenomen toestellen, kunnen er zich in werkelijkheid tot 16 toestellen in de regio bevinden. Aangezien de onzekerheid in het aantal werkelijke toestellen in de regio al groter of gelijk is aan één wordt de locatie bepalen van het ene toestel erg lastig. Dit voorbeeld wordt in bijlage 1 verder uitgewerkt aangezien de bekende aantallen toestellen wel extra informatie geven.

Het effect van punt 2 varieert ook sterk per gebied. In stedelijk gebied zijn veel meer zendmasten waardoor zendmasten minder vaak een groot gebied bestrijken terwijl in landelijk gebied de masten vaak grotere gebieden bestrijken. Ook als veel zendmasten overlappen in dekkingsgebied wordt dit probleem complexer. Exacte cijfers hierover zijn op dit moment echter niet beschikbaar.
Zoals in punt 3 wordt aangegeven, worden voor veel regio's de cijfers onderdrukt. Over deze regio's is niets bekend anders dan dat het aantal geschatte toestellen kleiner is dan 15.

4.3 Groepsonthulling en nuancering van de N>15 regel bij stap 6

De output die het telecombedrijf levert bevat een tabel waarvan alle cellen meer dan 15 geschatte toestellen bevatten. Het is belangrijk te realiseren dat dit geen echte directe tellingen zijn. Het is een uitkomst van een Baysiaanse schatting. Dus zelfs een uitkomst van één toestel is in werkelijkheid niet uniek en daarmee anoniem. Als er een identificatie plaatsvindt dan is dat puur op basis van toeval en is een onthuller er dus ook niet zeker van. Echter, ook onterechte (maar als juist geïnterpreteerde) onthulling is onwenselijk; vandaar dat er wel een ondergrens is gesteld. Dat is de voornaamste reden voor deze grens.

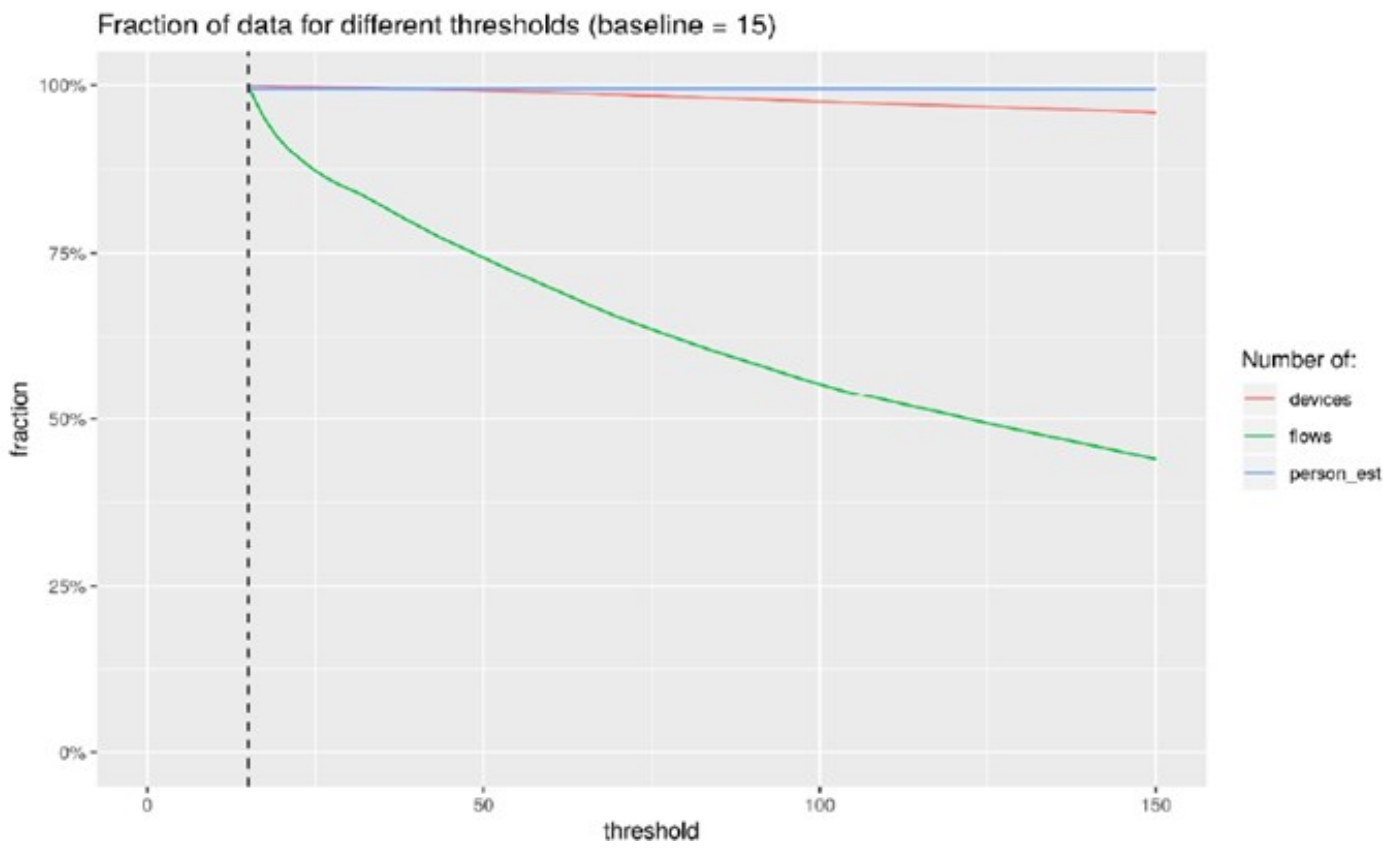
Het denkvoorbeeld is een bus studenten die tussen 2:00 en 3:00 uur van de gemeente Groningen naar een andere gemeente gaat. Als er tussen deze studenten meer dan 15 dezelfde provider hebben en deze groep van studenten allemaal in eenzelfde uur in een zelfde gemeente actief zijn (er zijn events van die toestellen) en toevallig door het geolocatiealgoritme in dezelfde gemeente worden toegewezen én er is geen enkele andere beweging van toestellen met dezelfde herkomstgemeente Groningen naar die gemeente dan kan een dergelijke groep herkend worden in de aan het CBS geleverde data.

Dat lijkt op onthulling, maar in feite wist men dat men op zoek moest gaan naar deze groep. Agenda informatie (bijvoorbeeld een programmapublicatie op het internet) was noodzakelijk. Er is dus uit deze databron geen extra informatie onthuld in dit geval, wel is er sprake van een eenmalige herkenning (of validatie) van de uitkomsten. Ook kan men binnen die groep geen individuen herkennen. Wel kan men aan die groep bepaalde kenmerken toevoegen zoals dat deze via een andere gemeente is gereisd.

Bij een hogere grens N zou de kans op groepsherleiding nóg kleiner worden, maar dat geldt dan alleen binnen de infrastructuur van het CBS. De grens van N>15 is echter nauwkeurig onderzocht. Daarbij is er gestreefd naar data-minimalisatie in relatie tot de toepassing: de data moeten wel bruikbaar blijven. Figuur 4 en 5 laten zien dat het verder verhogen van de grens in stap 6 leidt tot uiteindelijk uitkomsten die onbruikbaar zijn. In figuur 4.2 is te zien dat de dynamiek/stromen van de bevolking tussen verplaatsingen van gebieden (groene lijn) sterk verdwijnt bij het verhogen van de grens/"treshold", terwijl die informatie juist relevant is voor het maken van statistiek.

Wetende dat er onder zeer uitzonderlijke omstandigheden een risico op groepsherleiding kan bestaan zijn er diverse waarborgen genomen om te voorkomen dat deze informatie zou worden gebruikt om groepen personen te identificeren in de data die aan het CBS worden geleverd. Wettelijk is het voor het telefoonbedrijf verboden om herleidbare gegevens van individuen te verstrekken en voor het CBS is het wettelijk verboden om herleidbare gegevens openbaar te maken. Dit verbod wordt ook met organisatorische en technische maatregelen ondersteund. Zo is de toegang tot deze data bij het CBS afgeschermd en kunnen slechts enkele medewerkers van het CBS bij deze data.

Figuur 4.2 Percentage stromen en geschatte toestellen en geschatte personen dat afneemt met het verder verhogen van de grens van N>15



2) In de praktijk wordt een toestel over meerdere gemeenten uitgesmeerd.
3) In het wetenschappelijk artikel worden trajecten samengesteld uit de data. De in de pilot gebruikte methode kan niet gebruikt worden om routes samen te stellen, maar er kan aangenomen worden dat uit de tellingen per uur misschien een route kan worden samengesteld.
4) Hiermee wordt een woonplaats bestemmingstabel bedoeld met daarin tellingen in de tijd. Bijvoorbeeld in Amsterdam zijn om 12:00 uur 300.000 mensen uit Amsterdam en 500.000 mensen uit omliggende gemeenten. Door de uren naast elkaar te leggen kan men in de tellingen de "stromen"/ "flow" waarnemen. De derde dimensie is dan de tijd, vandaar de naam flowkubus, een herkomstbestemmingsmatrix in de tijd.
5) Dus de aanvaller weet wel de locaties van alle andere toestellen, maar dat aftrekken van de flowkubus aantallen zal verkeerde (en mogelijk negatieve) getallen opleveren. Dus de aanvaller zou ook moeten weten welke andere toestellen allemaal actief waren.

CBS Publicatie

Auteur: May Offermans, Yvonne Gootzen, Edwin de Jonge, Jan van der Laan, Frank Pijpers, Shan Shah, Martijn Tennekes, Peter-Paul de Wolf.
Publicatiedatum: 12-1-2021 16:25

Inleidende Methodebeschrijving Pilot Mobiele telefoniedata

5. Conclusie

De pilot heeft aangetoond dat het mogelijk is goede schattingen te maken uit geanonimiseerde geaggregeeerde mobiele telefoniedata. Deze zouden in de toekomst belangrijke informatie kunnen leveren voor beleid en wetenschap op terreinen als toerisme, mobiliteit en milieu. Ook is er uitvoerig naar privacy gekeken. Uit dat onderzoek blijkt dat met deze methode het niet mogelijk is om de locatie van een toestel met redelijke zekerheid te bepalen, zelfs niet in de situatie dat de aanvaller de beschikking heeft over een onrealistische hoeveelheid informatie.

Met de geaggregeerde informatie die uit deze methode wordt gepubliceerd is het onmogelijk om de locatie van individuele toestellen met zekerheid te bepalen. Deze informatie is dus voor een aanvaller, ook als deze over zeer veel hulpinformatie beschikt, anoniem.

CBS Publicatie

Auteur: May Offermans, Yvonne Gootzen, Edwin de Jonge, Jan van der Laan, Frank Pijpers, Shan Shah, Martijn Tennekes, Peter-Paul de Wolf.
Publicatiedatum: 12-1-2021 16:25

Inleidende Methodebeschrijving Pilot Mobiele telefoniedata

Referenties

[1] CBS (april 2020), "Estimating hourly population flows in the Netherlands (https://www.cbs.nl/-/media/cbs/overons/innovatie/rapport.pdf)".
[2] Positium (https://unstats.un.org/bigdata/events/2019/tbilisi/presentations/Session%201/Mobile%20Phone%20Data%20-%20Introduction%20Siim.pdf).
[3] Tu, Z., Xu, F., Li, Y., Zhang, P., Jin, D. (2018), "A new privacy breach : user trajectory recovery from aggregated mobility data (https://ieeexplore.ieee.org/document/8356232)", IEEE/ACM Trans. on Networking, 26, 3, 1446-1459.
[4] Zang, H. Bolot, J. (2011), "Anonymization of location data does not work: A large-scale measurement study", In *Proceedings of the 17th annual international conference on Mobile computing and networking (https://dl.acm.org/doi/10.1145/2030613.2030630)*, 145-156.
[5] De Montjoye, Y.A., Hidalgo, C.A., Verleysen, M., Blondel, V.D. (2013), "Unique in the crowd: The privacy bounds of human mobility (https://www.nature.com/articles/srep01376)", Nature *Scientific reports*, 3, p.1376-1381.

CBS Publicatie

Auteur: May Offermans, Yvonne Gootzen, Edwin de Jonge, Jan van der Laan, Frank Pijpers, Shan Shah, Martijn Tennekes, Peter-Paul de Wolf.
Publicatiedatum: 12-1-2021 16:25

Inleidende Methodebeschrijving Pilot Mobiele telefoniedata

Bijlage

Verdere uitwerking van punt 1: niet alle toestellen zijn ieder uur actief

De aanvaller wil de regio *x* van het toestel *i* bepalen.

- Van alle toestellen op *i* na is de regio bekend.
- Van toestel *i* weten we dat hij of in regio A zit of in regio B doordat we de locatie hebben van *i* in t-1 en t+1. We beperken ons hier dus tot twee mogelijke regio's. Meer mogelijke regio's maakt de locatie alleen maar onzekerder.
- We nemen een fractie *f* van de toestellen waar.

Gegeven de tellingen per regio, n_A en n_B, en de bekende aantallen per regio, N_A en N_B, wat is de kans dat device *i* in regio A zit en wat is de kans dat het device in regio B zit. De laatste volgt automatisch uit de eerste omdat het device of in A of in B zit.

$$\Pr(n_A, N_A | x = A) = Binom(k = n_A; n = N_A + 1; p = f)$$

En

$$\Pr(n_A, N_A | x = B) = Binom(k = n_A; n = N_A; p = f)$$

Het doel is om $\Pr(x = A | n_A, N_A, n_B, N_B)$ te bepalen. Gebruik makend van de regel van Bayes en conditionele onafhankelijkheid tussen n_A en n_B:

$$\Pr(x = A | n_A, N_A, n_B, N_B)$$
$$= \frac{\Pr(n_A, N_A, n_B, N_B | x = A) \Pr(x = A)}{\Pr(n_A, N_A, n_B, N_B | x = A) \Pr(x = A) + \Pr(n_A, N_A, n_B, N_B | x = B) \Pr(x = B)}$$
$$= \frac{\Pr(n_A, N_A | x = A) \Pr(n_B, N_B | x = A) \Pr(x = A)}{\Pr(n_A, N_A | x = A) \Pr(n_B, N_B | x = A) \Pr(x = A) + \Pr(n_A, N_A | x = B) \Pr(n_B, N_B | x = B) \Pr(x = B)}$$

Onder aanname dat de aanvaller geen extra informatie heeft om te kiezen tussen A en B:

$$\Pr(x = A | n_A, N_A, n_B, N_B) = \frac{\Pr(n_A, N_A | x = A) \Pr(n_B, N_B | x = A)}{\Pr(n_A, N_A | x = A) \Pr(n_B, N_B | x = A) + \Pr(n_A, N_A | x = B) \Pr(n_B, N_B | x = B)}$$

Als hier de formules voor de binomiale verdeling ingevuld worden kan dit verder uitgewerkt worden tot:

$$\Pr(x = A | n_A, N_A, n_B, N_B) = \left\{1 + \frac{(N_B + 1)(N_A + 1 - n_A)}{(N_B + 1 - n_B)(N_A + 1)}\right\}^{-1}$$

De figuur hieronder geeft voor verschillende waardes van n_A en n_B de kans dat i in A zit. N_A en N_B zijn een factor 1/0.8 hoger gekozen dan n_A en n_B, de fractie f is gelijk aan 0.8. De verticale stippellijn geeft de grens aan waaronder de cijfers van een regio worden onderdrukt. De horizontale stippellijn geeft een kans van 0.1 aan: onder deze lijn weten we met 90% zekerheid dat i zich in regio B bevindt. Voor waardes van n_A en n_B groter dan 15, is iedere regio min of meer even waarschijnlijk, de data in de flowkubus geeft dus nauwelijks extra informatie over de locatie van het toestel.

Figuur b1 Kans dat toestel *i* zich in regio *A* bevind voor verschillende aantallen geschatte toestellen in regio's *A* en *B*