Technische nota van het Kenniscentrum van de mobiliteit
in het Brussels Hoofdstedelijk Gewest

Het nut van *big data* voor het mobiliteitsonderzoek in het Brussels Hoofdstedelijk Gewest: belangen, kansen en uitdagingen

Door Thomas Ermans, Céline Brandeleer en Michel Hubert

BRUSSEL MOBILITEIT
GEWESTELIJKE OVERHEIDSDIENST BRUSSEL

De auteurs

Thomas Ermans is geograaf (ULB) en behaalde een aanvullende master in statistische data-analyse (UGent). Als onderzoeker aan het Centrum voor Sociologische Studies van de Université Saint-Louis in Brussel sinds 2014, werkte hij vooral op de stedelijke mobiliteit, meer bepaald binnen het Kenniscentrum van de mobiliteit van het BHG. In 2019 vervoegde hij het Brussels Instituut voor Statistiek en Analyse (BISA – perspective.brussels).
Contact: termans@perspective.brussels

Céline Brandeleer is politicologe (USL-B/UCL). Ze werkt sinds 2014 voor het Centrum voor Sociologische Studies van de Université Saint-Louis in Brussel, waar ze mee de *Katernen* van het Kenniscentrum van de mobiliteit van het BHG opstelt. Haar onderzoeksthema's zijn stedelijke mobiliteit, de analyse van het overheidsoptreden, de mobiliteit van werknemers en sociale ongelijkheid op het vlak van mobiliteit. In 2019 vervoegde ze het Brussels Instituut voor Statistiek en Analyse (BISA – perspective.brussels).
Contact: cbrandeleer@perspective.brussels

Michel Hubert is doctor in de sociologie, gewoon hoogleraar aan de Université Saint-Louis in Brussel, waar hij het Instituut voor interdisciplinair onderzoek over Brussel (IRIB) voorzit, en is daarnaast gastdocent aan het METICES-centrum van de Université Libre de Bruxelles (ULB). Sinds de oprichting in 2006 staat hij ook aan het hoofd van het tijdschrift *Brussels Studies* en is hij vicevoorzitter van het *Brussels Studies Institute* (BSI). Binnen zijn onderzoek bestudeert hij in het bijzonder mobiliteitsgewoontes, en de geschiedenis en de structuur van vervoersnetwerken en hun impact op de stad en haar gebruikers. Michel Hubert coördineert al van bij de oprichting de *Katernen* van het Kenniscentrum van de mobiliteit.
Contact: michel.hubert@usaintlouis.be

De technische nota "Het nut van *big data* voor het mobiliteitsonderzoek in het Brussels Hoofdstedelijk Gewest: belangen, kansen en uitdagingen" werd in september 2018 afgerond.
Daarom stoppen de bronnen waarnaar in het document wordt verwezen op een datum die ruim voor de datum van publicatie ligt. De daarin vervatte beschouwingen blijven echter geldig.

Inhoudsopgave

Inleiding

De explosie van digitale *sporen* binnen een almaar sterker geconnecteerde samenleving heeft geleid tot de ontwikkeling van zogeheten *big data* of massagegevens. Die *big data* staat bol van de opportuniteiten, maar brengt ook een hele rist technische uitdagingen met zich mee en vergt – om er de nuttige informatie uit te filteren – methodes die breken met het 'klassieke' instrumentarium van onderzoekers, administraties, studiebureaus en bij uitbreiding van de geïnteresseerde burger.

Die nieuwe realiteit roept om een (her)positionering van al die partijen, temeer omdat de opkomst van *big data* inspeelt op het grote enthousiasme rond het denkbeeld van *smart cities* (of slimme steden), waarbij dergelijke data duidelijk een streepje voor heeft op de meer klassieke methodes en gegevens. Dat doet een aantal vragen rijzen: Moeten we nog steeds mobiliteitsenquêtes afnemen in een *hypergeconnecteerde* stad die elk *spoor* dat mensen achterlaten, registreert? Welke plaats heeft strategisch onderzoek binnen de *smart city*? enz.

Deze nota oppert dan ook een aantal denksporen omtrent onze positie ten aanzien van *big data*. Zo willen we in het bijzonder nagaan welke plek we dergelijke massale gegevens kunnen toedichten in het methodologische instrumentarium van 'deskundigen', in een heel brede zin van het woord, en op welke manier we ze in het hele maatschappelijke proces kunnen inbedden.

In het bijzonder willen we in deze nota (1) de elementen op een rijtje zetten die de kennisproductie genereren door middel van *big data*; (2) de verhoudingen tussen *big data* en *smart city* uit de doeken doen en toelichten wat ze impliceren, niet enkel op het vlak van kennisproductie, maar ook op het vlak van het overheidsoptreden, meer bepaald wat betreft de modaliteiten om de voortgebrachte kennis om te zetten in *acties* voor ons leven van elke dag; (3) dieper ingaan op de belangen en uitdagingen voor overheidsinstellingen en onderzoekers in het hele mobiliteitsonderzoek, vooral dan op het vlak van relevantie en het contextualiseren van gegevens, eigendom van informatie, technische bekwaamheden en respect voor de privacy; en tot slot (4) die belangen en uitdagingen concreet toelichten aan de hand van drie gegevensbronnen: de *floating mobile data (FMD)*, de *floating car data (FCD)* en de gegevens afkomstig van de automatische ontwaarding van vervoersbewijzen op het openbaar vervoer.

Die drie gegevensbronnen zijn uiteraard niet de enige *big data* geschikt voor mobiliteitsstudies, maar ze behoren wel tot de meeste gebruikte. Ze winnen bovendien enorm aan belang in het Brusselse, nu heel wat private operatoren *big data* verzamelen (denken we aan Proximus, Be Mobile, TomTom enz.), maar ook gelet op de rijkdom aan informatie die uit de Mobib-kaart kan voortkomen (die sinds 2016 gebruikt wordt voor alle MIVB-tariefformules).

©STIB-MIVB

1. *Big data* : nieuwe gegevens voor nieuwe manieren van kennisproductie

1.1. Onverwachte gegevens, *massaal* geproduceerd

In een samenleving die zichzelf meer en meer als '*smart*' bestempelt en steeds sterker geconnecteerd is (via internet, smartphones, *gps* enz.), laat een almaar breder scala aan handelingen zijn sporen na. Dat leidt tot een heuse stroom (en voorraad) aan data die exponentieel toeneemt[1]. Gegevensproductie was lange tijd voorbehouden aan instellingen (overheden, bedrijven, verenigingen) en beperkt tot welbepaalde categorieën (geboortedatum, beroepscategorie ...). Tegenwoordig is ze steeds meer een zaak van individuen of machines die automatisch gegevens verzamelen, en strekt ze zich uit tot domeinen die voorheen niet met een zekere precisie te vatten waren, zoals onze verplaatsingsgewoontes, onze smaken, onze relaties enz. (Cytermann, 2015). De soms wat plotse, maar doorgaans gewenste opkomst van massa databanken ging zo gepaard met de ontwikkeling van oplossingen voor gegevensopslag, organisatie van datastromen, standaardisering van definities en bestandsindelingen, enz.

Door die trend is de oude definitie van *big data* weer helemaal actueel. Ze is opgebouwd rond de 3 V's (Laney 2001), afkomstig van de Engelse termen *volume*, *velocity* en *variety*, stuk voor stuk eigenschappen die *big data* kenmerken. Binnen die definitie staat het volume (*volume*) voor de hoeveelheid belangrijke data, de snelheid (*velocity*) voor het feit dat dit soort gegevens razendsnel wordt aangelegd, vaak onafgebroken, en de verscheidenheid (*variety*) tot slot verwijst naar de gevarieerde bronnen en indelingen van de verzamelde gegevens en duidt vooral op hun uiteenlopende, wanordelijke aard, waardoor ze vaak niet zomaar analytisch verwerkt kunnen worden.

Merk op dat deze definitie voortkomt uit de technische moeilijkheden die analisten uit de e-commerce ondervinden om om te springen met die stortvloed aan gegevens, die hun huidige verwerkingscapaciteiten enorm onder druk zet. Zonder echt goed te weten waar ze nu juist vandaan komt, horen of lezen we vaak de volgende definitie: "[elk gegeven dat niet in een excelbestand past]" (Batty, 2013 : 274). Die definitie, die vooral de nadruk legt op het volume, het *'big'*-aspect, heeft evenwel de neiging om het fenomeen van *big data* in te perken tot een technische kwestie.

Een formulering waar we ons dan ook niet helemaal in kun vinden. Vanaf welke omvang is een steekproef groot genoeg? Vanaf welke dataproductiesnelheid? Met welke verscheidenheid? Het lijdt geen twijfel dat *big data*, als gegevensopslag, bepaalde eigenschappen heeft, maar die lijken minder belangrijk om het verschijnsel te vatten dan de bijzondere manier waarop het kennisproductie creëert. Volgens heel wat auteurs noopt die tot een wijziging in de zienswijze. Om er maar een paar te noemen: Graham en Shelton (2013 : 256) spreken van een "[informaticaparadigma]", terwijl Antoinette Rouvroy (2014 : 413) het heeft over een "nieuw waarheidsstelsel".

1.2. Een uitdaging van formaat: het signaal filteren

Het is interessant om in dit stadium de verschillen tussen enquêtegegevens, administratieve gegevens en *big data* te onderstrepen. Enquêtegegevens zijn doorgaans de gegevens verzameld in het kader van een instrument dat een antwoord wil bieden op voorafbepaalde onderzoeksvragen. De steekproef is bekend en, hoewel de gegevensverzameling aanzienlijk kan zijn, blijft het datavolume over het algemeen toch vrij beperkt[2]. De administratieve gegevens daarentegen worden niet verzameld voor onderzoeksdoeleinden, en hun collectie verloopt minder eenvormig binnen de onderzoekspopulatie (waarvan we doorgaans de steekproef kennen). Het gaat dikwijls om complexe, meerdimensionale en aanzienlijk grotere databanken dan bij enquêtes, die om een specifieke aanpak vragen (Connelly *et al.*, 2016).

Vanuit dat oogpunt lijken twee fundamentelere, of misschien louter synthetischere kenmerken van *big data* de aanzienlijke afstand tussen gegeven en signaal[3], dat we voor een deel terugvinden in het *variety*-aspect van de 3 V's, en het feit dat de gegevens niet a priori worden geproduceerd en gekalibreerd om te kunnen dienen voor de analyse van maatschappelijke verschijnselen[4]. Om dat aspect aan te pakken, spreken sommige auteurs van een 4de V, die staat voor waarheidsgetrouwheid (*veracity* in het Engels) (zie meer bepaald André De Palma, 2017 : 22), om zo de variabele kwaliteit van het gegeven te benadrukken en aan te geven dat het herverwerkt moet worden om bruikbaar te zijn.

In dat opzicht is de uitdaging waar *big data* ons voor stelt niet hoe we een voortdurend groeiende massa aan gegevens sneller en beter verwerken, wel hoe we er toegevoegde waarde uit halen (Floridi, 2012). Deze doelstelling gaat vergezeld van kenmerkende analysetools die volledig deel uitmaken van het afgelijnde *big data*-kader, dat daardoor een bijzondere epistemologische positie krijgt en ze niet voor elke post vervangt door de klassieke instrumenten en analyses.

De gehanteerde technieken in het kader van *big data* vloeien in essentie voort uit een proces van kennisextractie uit databanken[5] (Miller, 2010; Fayyad, 1996), een vaststelling die de eerder genoemde afstand tussen gegeven en signaal onderstreept. We nemen hier de tijd om de voornaamste

[1] Ter indicatie: waar in 2010 het wereldwijde volume aan gestockeerde data nog afklokte op 1,2 zettabyte (1,2.2021 bytes, ofwel 1.200 miljard gigabyte), schatte consultant IDC in 2014 dat dat volume tegen 2020 zal toenemen tot 44 zettabyte. Dat is een vermeerdering met een factor 37, en dat op 10 jaar tijd.

[2] De sociaal-economische enquête van 2001, waarbij de gegevensverzameling de hele beoogde populatie wilde omvatten, telt ongeveer 10 miljoen lijnen voor een honderdtal attributen en geldt in dit kader dan ook als tegenvoorbeeld.

[3] De term *signaal* verwijst hier naar de nuttige, de informatieve data voor een welbepaald onderzoekstarget, in tegenstelling tot de 'ruis' in de gegevens, die het signaal verstoort.

[4] Een eigenschap die het deelt met administratieve gegevens.

[5] In het Engels heet dat *knowledge discovery from databases*.

stappen schematisch uit de doeken te doen (zoals gedefinieerd door Miller (2010 : 189)), enerzijds omdat er weinig twijfel over bestaat dat slechts een fractie hoogopgeleide analisten in staat is om er zich een beeld van te vormen, en anderzijds om te wijzen op het belang van de menselijke factor in de opeenvolgende beslissingen die genomen worden binnen het proces en die het onvermijdelijk een zekere subjectiviteit meegeven:

- Dataselectie: keuze van het subgeheel dat gebruikt zal worden voor de analyse.
- Opschonen van gegevens (dubbels, ontbrekende gegevens enz.) en eventuele aanvulling met bijkomende gegevens.
- Terugdringing van het aantal dimensies en/of projectie van de gegevens in doeltreffendere voorstellingsruimtes.
- Toepassing van technieken voor *datamining*, om herhalingen, patronen (*patterns*), eigenheden verborgen in de gegevens bloot te leggen.
- Interpretatie van de gegevens en conclusierapport.

Binnen dat proces zijn het de verschillende technieken voor *datamining* die de gegevens hun kwalitatieve karakter geven, zodat ze als nuttige informatie gebruikt kunnen worden. Het gaat om aangepaste werktuigen die minder gericht zijn op het verklaren dan op het beschrijven en blootleggen van verborgen structuren binnen de gegevens, door allerhande technieken te hanteren, zoals *clustering* (groepering van vaststellingen, van onderling gelijkaardige objecten), classificatie (het ordenen van objecten in klassen, volgens een eventueel vooraf bepaalde regel), associatie (op basis van de relaties tussen objecten voorspellingen maken van toekomstige waarden; op basis van regressies of beslissingsbomen bijvoorbeeld) en het bestuderen van afwijkingen (individuele objecten die afwijken van de verwachte norm) (Miller, 2010).

Machine learning en *deep learning* (zie onderstaande kadertekst) zijn bijzondere *dataminingtools* die, door zichzelf beslissingsregels aan te leren, beter geschikt zijn voor een dynamische waardering van het ontvangen signaal, dat dan een realtime handeling kan triggeren **(2.2. *Smart city* en *big data*: een nieuw denkbeeld van de stad)**.

Machine learning, deep learning en artificiële intelligentie

Van de *dataminingtechnieken* nemen vooral de concepten en tools voor *machine learning*, *deep learning* en artificiële intelligentie het voortouw, zowel wat hun gebruik als hun media-aandacht betreft.

Van deze drie termen is *machine learning* (machinaal leren in het Nederlands) ongetwijfeld het wijdst verbreid. Het begrip is zowat het gezicht van *datamining* en berust op de uitvoering van algoritmes die een set regels bepalen op basis van een eerste pakket met gegevens (trainingsfase, leerfase). Die regels worden vervolgens toegepast op een tweede datacollectie. Vayatis (2017 : 55) zegt daarover dat we hier rekening moeten houden met "twee algoritmeniveaus: een eerste algoritme A dat het eigenlijke leerproces voor zijn rekening neemt, door contact te maken met een databank (input van algoritme A) die een bepaald criterium (een nutsfunctie) optimaliseert; dat criterium staat voor wat het algoritme beschouwt als een goede beslissing, en leidt tot een beslissingsregel (output van algoritme A). Een tweede algoritme P voert die beslissingsregel van algoritme A dan uit op nieuwe data, om te helpen bij de besluitvorming (naargelang de context spreken we soms van voorspelling)."

De overgang naar het nieuwe millennium luidde de komst in van nieuwe algoritmetechnieken geschikt voor *machine learning*, in het kader van massagegevens. Onder die nieuwe methodes maakt Vayatis (2017) een onderscheid tussen methodes die een hoge mate van transparantie (white box) toelaten in de interpretatie van de resultaten (beslissingsbomen, lineaire parcimonieuze modellen) en methodes die enkel beperkte transparantie (grey box) toelaten (random forests, totaalbenaderingen).

Een andere methode ontleend aan *machine learning* is *deep learning*, waarbij gebruik wordt gemaakt van algoritmes geïnspireerd op de werking van de neuronale netwerken in het menselijk brein, en waardoor ze dus als artificiële intelligentie bestempeld kunnen worden. Deze instrumenten, reeds bedacht (en in praktijk gebracht) in de jaren 80, winnen tegenwoordig opnieuw aan populariteit binnen het hele *big data*-verhaal en gelet op de aanzienlijke verbetering van de rekenkracht. Tegenwoordig spreekt men van diepe algoritmes (*deep learning*) omwille van het gebruik van meerdere lagen met neuronale netwerken, die parallel gegevens verwerken. Ze mogen dan bijzonder performant zijn, hun interpretatie is ondoorzichtig en geeft doorgaans geen inzicht in de functionele drijfveren die de beslissing van het algoritme motiveert. In dat geval spreken we dan ook eerder van een zwarte doos of 'black box' (Vayatis, 2017).

Al deze instrumenten kunnen ingedeeld worden bij de classificatie- en associatiemethodes die verderop in de tekst aan bod komen. Zo zullen we toelichten dat ze aanzetten tot een herconfiguratie van de verhouding tussen onderzoeker (of deskundige, meer algemeen) en instrument, in het kader van een hulpproces voor de besluitvorming (Vayatis, 2017). Terwijl witte- of grijze-doosmethodes een zekere blik van de eerste in de tweede toelaten (augmented experts), dreigt de opkomst van artificiële intelligentie de expert te degraderen tot technisch facilitator.

1.3. Een bijzondere positie in de kennisproductie

Algemeen gesproken zijn deze technieken eerder de vrucht van een inductieve kennisproductie, waarbij de informatie in eerste instantie geherconfigureerd wordt om er de structuren en eigenheden van te ontwaren, door te fungeren als vergrootglas of microscoop op het gegevenscorpus (Miller, 2010; Allemand, 2013). De theoretische formulering van het vastgestelde verschijnsel zou zo pas in tweede instantie gebeuren, in overeenstemming met de blootgelegde structuren die leiden tot de interpretatie. Dit proces verschilt van de 'klassieke' standaarden voor kennisproductie, die voortvloeien uit een doorgaans deductievere aanpak, waarbij de kennis geënt wordt op een theoretisch voorstel dat voorafgaat aan de datacollectie en -analyse, aan de hand van een dispositief om dat initiële voorstel te toetsen. Heel vaak streeft men ernaar om de oorzakelijke verbanden en factoren aan te wijzen die de gemeten verschijnsels beïnvloeden.

Dat onderscheid is evenwel eerder complementair dan eliminerend. De terugkoppeling van informatie en kennis door *big data* is als zodanig interessant en kan uitmonden in een theoretisch kader dat zal dienen als vertrekpunt voor een meer deductieve benadering. Anderzijds is een meer inductieve aanpak onvermijdelijk beperkt tot de beschikbare gegevens, en indien het *big data*-kader toelaat om succesvol in te zoomen op bepaalde problemen, dan bestaat het risico dat andere verwaarloosd worden doordat elke voorafgaande theoretische reflectie wordt weggefilterd.

De positie van *big data* als kennisonthuller wakkert niettemin een zekere positivistische opvatting aan, die tot in het extreme wordt belichaamd door het standpunt van Anderson (2008), die oproept tot "[het einde van de theorie]"[6] (in het bijzonder geciteerd door Graham en Shelton, 2013; Boullier, 2015; Batty, 2013, Boyd en Crawford, 2012) en enthousiast de intrede verwelkomt van een nieuwe manier van kennisproductie, die objectiever is omdat ze de data voor zich laat spreken. Die opvattingen krijgen heel wat tegenwind. Ten eerste kan worden opgeworpen dat er geen zuiver inductief epistemologisch stelsel bestaat. De onderzoeker (in de brede zin van het woord) opereert immers altijd binnen een welbepaald en bestaand gedachtekader, en wordt in zijn handelingen geleid door specifieke doelstellingen en door de aannames en vooronderstellingen die erachter schuilgaan. Die menselijke factor is, zoals we al zagen, duidelijk aanwezig in elke stap van de kennisproductie in het kader van *big data*, waardoor er geen sprake kan zijn van een zekere overkoepelende objectiviteit. Ten tweede opperen meerdere auteurs dat een dergelijk discours niet in de marge van *big data* zou rondwaren, maar wel een centrale plaats zou innemen in dat nieuwe paradigma. Voor Boyd en Crawford (2012 : 663) ageert het aldus als een mythe die de idee in stand houdt dat "[...massale databanken een superieure vorm van inzicht en wijsheid voortbrengen, wat leidt tot een kennis die totnogtoe als onmogelijk werd beschouwd, met een aura van waarheid, objectiviteit en nauwkeurigheid]." Voor Graham en Shelton (2013) komt die epistemologische machtsverhouding tot uiting in "[diezelfde] *big data*", zeg maar de automatische, virale reproductie van een sterk geloof in de superieure aard van nieuwe databronnen en de waarheden die ze onthullen. In die passage wijzen de auteurs op het gevaar van onderschatting dat een dergelijk standpunt inhoudt voor de andere manieren om kennis te produceren, en in het bijzonder voor kritische benaderingen. Verderop lichten we toe op welke manier die symbolische autoriteit, verbonden aan het *big data*-kader, wordt nagebootst door de ontwikkeling van het stedelijk referentiekader van de *smart city*.

***

Om alles wat voorafgaat al even samen te vatten, hebben we een tabel gemaakt (Tabel 1) met wat volgens ons de voornaamste verschillen zijn tussen enquêtegegevens, administratieve gegevens en *big data*.

[6] Chris Anderson schreef in 2008 "The end of theory", een artikel dat verscheen in het magazine Wired, waar hij nu hoofdredacteur is. Dat artikel, uitvoerig geciteerd en becommentarieerd door de wetenschappelijke literatuur, lijkt te zijn uitgegroeid tot baken van de controverse omtrent het gebruik van big data in de sociale wetenschappen en voor de band die deze nieuwe manier van kennis voortbrengen onderhoudt met de traditionele productiewijzen.

Tabel 1. Vergelijking van enquêtegegevens, administratieve gegevens en *big data*

	Enquêtegegevens	Administratieve gegevens	*Big data*
Afstand data – signaal	Laag	Gemiddeld	Ver
Verzameld voor studiedoeleinden	Ja	Neen	Neen
Participatie/werving van individuen (Chen *et al.*, 2016)	Actief	Passief	Passief
Epistemologische benadering	Eerder deductief	Deductief/inductief	Eerder inductief
Hoofddoel van hun gebruik	Oorzakelijk	Oorzakelijk/kenschetsing (en classificatie)	Kenschetsing (en classificatie) en voorspelling
Aura van objectiviteit binnen het *big data*-kader	Neen	Neen	Ja
Kennisproducenten – in het kader van het overheidsbeleid*	Hoofdzakelijk overheidsactoren	Hoofdzakelijk overheidsactoren	Private actoren, overheidsactoren
Tijdgebondenheid van het overheidsoptreden**	Traag/statische verwerking	Traag/statische verwerking	Kort of traag/statische of dynamische verwerking

* Zie **deel 2. *Big data* en het overheidsoptreden: tussen opportuniteit en bevordering in het kader van de *smart city*.**
** Zie **deel 3. De uitdagingen van *big data*.**

2. *Big data* en het overheidsoptreden: tussen opportuniteiten en bevordering in het kader van de *smart city*

2.1. De opportuniteiten die *big data* biedt: volledigheid en ruimtelijk-temporeel detailniveau van de data, en dynamisch beheer

Big data leent zich zowel voor het ramen van datastromen, of het nu gaat om het aantal mensen dat langs een bepaalde plek komt of om individuen die tijdelijk worden gevolgd (mobiele telefoongegevens en data van transporteurs). Vanuit die invalshoek hebben ze als grote voordeel dat ze een heel gedetailleerde ruimte- en tijdschaal mogelijk maken (Chandesris *et al.*, 2017; Ermans *et al.*, 2017; Debusschere *et al.*, 2017).

In de *gps-data* die aanbieders van boordnavigatiesystemen produceren, is die tijdmatige detailgraad uiteraard ook aanwezig, maar heeft hij niet noodzakelijk meerwaarde ten aanzien van vaste sensoren, die de stromen onafgebroken meten. Wat die stromen betreft, bieden de ruimtelijke volledigheid en korrelgrootte (de mate aan detail) van het *gps-gegeven* (autostromen bieden de mogelijkheid om informatie te collecteren over het hele wegennet) het meeste potentieel, door het analysedomein van verkeersopstoppingen en reistijden uit te breiden tot netwerksegmenten waarvoor quasi geen informatie voorhanden is.

Meer bepaald die tijdsdimensie effent het pad naar nieuwe mogelijkheden op het vlak van mobiliteitsstudies en -beheer, en dat op minstens twee vlakken. Allereerst maakt ze analyses mogelijk die tot nu toe buiten de reikwijdte van grote enquêtes vielen. Stromen die heel nauwkeurig in de tijd worden geregistreerd, maken het immers gemakkelijker om bijvoorbeeld de mobiliteit in de daluren ('s avonds, tijdens het weekend, buiten de spits) te bestuderen, een aspect dat minder snel aandacht krijgt in enquêtes die prat gaan op representativiteit en waarvoor de steekproefdiepte vaak beperkt is. Aan de andere kant behelzen dergelijke gegevens de mogelijkheid om grondgebieden en locaties te kenschetsen op basis van hun bezetting, hun gebruik in de tijd. In zekere zin opent dat perspectieven om mobiliteitsstudies vanuit een nieuwe invalshoek te benaderen, door niet langer te onderzoeken hoe individuen zich bewegen over het grondgebied (of het bestuderen van de verplaatsingsgewoontes), maar wel door te kijken naar hoe het gebruik van dat grondgebied door individuen varieert in de tijd (Commenges, 2014).

Ten tweede kan die hoge mate aan tijdmatige nauwkeurigheid, gekoppeld aan het vermogen om data in realtime te verwerken, ertoe aanzetten een dynamische terugkoppeling en een netwerkbeheer te overwegen waarbij heel kort op de bal wordt gespeeld (ingrepen tijdens de spits en op verzadigingspunten enz.). Die visie van een dynamisch overheidsoptreden, in realtime bovendien, is niet nieuw (zie Brandeleer en Ermans (2016) voor een voorbeeld uit het Brussels Hoofdstedelijk Gewest), ze krijgt tegenwoordig bovendien bijval in het kader van de opmars van *big data* en de *smart city*.

2.2. *Smart city* en *big data*: een nieuw denkbeeld van de stad

De definitie van *smart city* ligt niet helemaal vast, en toch komt het begrip tot uiting in diverse andere begrippen: *smart city*, *digitale stad, virtuele stad* enz. Ze opperen echter allemaal een stedelijk denkbeeld waar de leefomgeving en de individuen die er wonen en bewegen, permanent met elkaar verbonden zijn via sensoren en instrumenten uit de informatie- en communicatietechnologie (ICT). Die detecteren al hun acties en vernetten ze. Je hoort ook wel eens spreken van het internet der dingen (*the internet of things*), dat een gigantische hoeveelheid aan data aanlegt. Om *slim* te zijn, moet die onderlinge verbondenheid via algoritmes en functies ook aanleiding geven tot een zekere mate aan rationalisering van acties, verplaatsingen en stromen, die toelaat om "[de efficiëntie, de gelijkheid, de duurzaamheid en de levenskwaliteit van burgers meteen te verbeteren]" (Batty *et al.*, 2012 : 482).

Het *big data*-kader, met al zijn uitdagingen op het vlak van maatschappelijke rechtvaardigheid en bescherming van de privacy, stuurt die visie wat betreft de voorwaarden voor de omschakeling tussen *verbonden stad* en *intelligente stad*, om naar een signaal dat meteen geïnterpreteerd wordt en bruikbaar wordt gemaakt, om te zetten in een actie die een vorm van stedelijk bestuur op heel korte termijn inhoudt. Op mobiliteitsvlak zou het er in wezen om draaien stromen in realtime te optimaliseren, door verkeer en reizigers slimmer te spreiden, door bepaalde toegangen open te zetten of te sluiten, door het transportaanbod en de capaciteit van de wegen te moduleren, door het gebruik van parkings dynamisch te sturen enz. De schijnbare objectiviteit van *big data* (aangezien het het algoritme is dat het signaal onderscheidt van de ruis), geeft ze een belangrijke symbolische relevantie bij besluitvormers. Die strekt zich uit tot de *smart city*, om objectieve, actuele inzichten te beloven, ten gunste van een gestroomlijnd beheer van het stadsleven.

In dat opzicht krijgt de band tussen de empirische opvatting van het stadsbeeld en het stedelijke overheidsoptreden een nieuwe wending. Die empirische benadering uitte zich als vanouds in rapporten en onderzoeken, om indicatoren aan te leggen die moesten dienen om het beeld van de stad en haar evolutie uit te drukken, te beschrijven en te vormen. Die aanpak vloeit voort uit beschrijving, maar ook uit toelichting, uit de zoektocht naar causale verbanden, stuk voor stuk pijlers van het overheidsoptreden (zie bijvoorbeeld van der Loop *et al.* (2017: 10-11).). Het gaat onvermijdelijk om een langetermijnproces, zeker geen onmiddellijk gegeven, gelet op de tijd die nodig is voor onderzoek, beleidsbeslissingen en hun effecten op de realiteit. Welnu, in de visie voorgesteld in voorgaande paragraaf, is dat tijdaspect bijzonder kort en volgt het antwoord quasi meteen. De geproduceerde kennis vloeit minder voort uit de verklaring en meer uit de structurering van de informatie, en dat vergt optimalisatie. Het signaal, de teruggekoppelde kennis, lijkt zo performatief, aangezien het rechtstreeks, op zich, het overheidsoptreden lijkt te sturen.

Maar in werkelijkheid is het instrument nooit neutraal. De politieke besluitvorming en het overheidsoptreden nestelen zich in de voorgeprogrammeerdheid van de antwoorden die geboden moeten worden op de verschillende stimuli van de intelligente stad (welk verkeer wordt er omgeleid? welke gebruiker krijgt voorrang? enz.) (Brandeleer en Ermans, 2016), en zelfs in de keuze van de aantallen die geoptimaliseerd moeten worden door lerende algoritmes. Men kan ook stellen dat de eigenlijke keuze om eerder te opteren voor een vorm van overheidsoptreden die sneller uitvoerbaar is, ten koste van langetermijnbenaderingen die causale mechanismen blootleggen, de mogelijkheden beperkt om mobiliteit als een probleem te benaderen (door de ruimtelijke inrichting van de functies of door te investeren in het openbaar vervoer bijvoorbeeld) en in dat opzicht een beleidskeuze an sich inhoudt. Tot slot houdt die visie op het overheidsoptreden het risico in dat ze enkel geconnecteerde personen bereikt.

Laten we hier concluderen dat indien de digitalisering van onze leefwereld tot nieuwe manieren voor mobiliteitsmeting leidt, ze ook tot nieuwe verplaatsingsgewoontes zal leiden (Chandesris *et al.*, 2107 : 143). Nemen we het voorbeeld van de realtime informatie op voertuigen van de MIVB of van routeplanners. De opkomst van smartphones en navigatiesystemen heeft geleid tot een heuse wildgroei aan mobiliteitsdiensten voor geconnecteerde mensen. Die ontwikkelingen schragen de uitbouw van nieuwe diensten voor pendelaars, te vatten onder de noemer van *Mobility as a Service (MaaS)*. Naast het verstrekken van een routeplanner die alle netwerken bundelt, wordt *MaaS* idealiter opgevat als een portaal om de beschikbare mobiliteitsdiensten te vergelijken en te betalen (voor de klant gaat het erom zowel de prijs als de reistijd te optimaliseren).

Impliciet berust de probleemoplossende kracht van *MaaS* binnen de stedelijke mobiliteit op de eventuele globale samenhang los van gedragingen van individu's, als antwoord op de informatie die intermodale en operatoroverschrijdende routeplanners aanbieden. De goede beheersing van de informatie verstrekt aan particulieren vormt zo een niet-verwaarloosbare uitdaging op het vlak van stedelijk mobiliteitsbeheer.

2.3. De oorsprong van de *smart city*: de ICT-sector

Terwijl het concept van *connected city*, van *smart city* al vijftien jaar in opmars is, hangt de hedendaagse eensluidendheid over de Engelse term *smart city* ongetwijfeld samen met het recente proces om die term in te bedden in het stadsbeleid. We zien dat wereldwijde spelers zich steeds meer roeren in dat stedelijke beleid, door ICT-oplossingen te ontwikkelen en uit te rollen (voor infrastructuur, software), denken we aan IBM, CISCO, Microsoft, Oracle, SAP. Zij dragen allemaal bij tot de uitbouw van een heuse markt voor de digitale stad (Batty *et al.*, 2012 ; Douay en Henriot, 2016). Batty *et al.* (2012) halen het voorbeeld van IBM aan, dat de strategische keuze heeft gemaakt om te investeren in het *smart*-aspect, met de campagne *voor een Smartere planeet* die het bedrijf sinds 2008[7] voert. Het heeft zijn producten en diensten geherpositioneerd, niet enkel om ICT-oplossingen aan te reiken die steden slimmer maken, maar ook door een heel gamma aan adviesdiensten voor lokale overheden en rond allerhande problematieken uit te werken. De bijhorende oplossingen zijn, hoe kan het ook anders, ICT-gebaseerd. Het is in die context dat heel wat steden hebben meegedaan aan de *Smarter Cities Challenge*. In het kader van die challenge stuurt IBM naar elke stad een groep deskundigen, die een rapport moeten opstellen over een vooraf gekozen stedelijk thema (mobiliteit, maar ook leefmilieu, veiligheid, bestuur, sociale diensten, economische ontwikkeling enz.). Aan dat rapport worden allerhande aanbevelingen gekoppeld voor de overheid[8].

De rol van private bedrijven, die door hun samenwerking met de overheid het stedelijke beleid kunnen beïnvloeden ten gunste van het referentiekader van de *smart city*, wordt ook benadrukt door Douay en Henriot (2016), die die wereldwijde trend onder de loep nemen binnen een aantal Chinese steden. Merk tot slot ook nog op dat de auteurs de eigenlijke inhoud van de effecten van de verslimming van steden, die operationeel vaak te wensen overlaat, in vraag stellen. Ze werpen ook heel duidelijk de hypothese op dat de *smart city* vooral bijdraagt tot een proces van *storytelling* (de stad vertelt een verhaal én wordt verteld), vanuit een logica van stadsmarketing waar het *smart*-argument het *duurzaamheidsargument* overstijgt. Ze voegen er nog aan toe dat die omschakeling evenwel een logica zou drijven waarbij *smart cities*, doordat hun volledige samenleving wordt gerationaliseerd, onvermijdelijk uitgroeien tot koolstofarme en dus duurzame steden.

In Brussel komt, naast de deelname aan de *Smarter Cities Challenge*, de ommekeer naar de *smart city* duidelijk tot uiting in het regeerakkoord 2014-2019, dat Brussel wil uitbouwen tot een 'digitale hoofdstad' (Brusselse Regering, 2014 : 25). Een ambitie die navolging kreeg met de aanstelling, in januari 2015, van een Staatssecretaris voor informatie en digitalisering en met de ontwikkeling van een website die specifiek bedoeld is om de slimme aard van Brussel te versterken[9].

Het belang van het CIBG (Centrum voor Informatica van het Brusselse Gewest) in de uitbouw van een visie en een programmering van Brussel als *smart city* moet eveneens onderstreept worden. Deze instelling van openbaar nut ijvert immers voor de promotie van de intelligente stad, op basis van een hele reeks actiepijlers, denken we aan het Witboek 2014-2019 (CIBG, 2014), de organisatie van een *smart city summit*, een *smart city event*, een *smart breakfast* enz. Daarnaast is het als competentiecentrum en beheerder van het telecomnet actief betrokken bij de implementering van de *smart city*, door overheden én burgers bij te staan met ICT-oplossingen, in het bijzonder op het vlak van datacollectie, -beheer en -deling.

Algemener gesproken zien we dat de ontwikkeling van een *smart city*-visie maar een van de aspecten is in het digitaliseringsbeleid van Brussel, naast de oprichting van Impulse (voor werk en onderwijs) en Innoviris (voor innovatie). Deze drie beleidspijlers zijn gebald onder de noemer Digital Brussels (CIBG, n.d.: 1), om de samenhang en de goede coördinatie van het geheel te garanderen.

Zonder exhaustief te willen zijn, zetten we graag een aantal Brusselse initiatieven op een rij die aansluiten op dat uitdijende stedelijk referentiekader. Zo is er sinds 2017 digitYser (www.digitYser.org), dat zich opwerpt als

[7] Zie de webpagina http://www.ibm.com/smarterplanet/us/en/, die opent met de titel 'IBM builds a smarter planet' (IBM bouwt aan een slimmere planeet).

[8] Brussel bijvoorbeeld deed in 2014 al een beroep op IBM voor een mobiliteitsdiagnose. De resultaten zijn te vinden op https://smartercitieschallenge.org/cities/brussels-capital-region-belgium. Het gaat om een samenvatting van de voornaamste bevindingen door de lokale actoren (besturen, universiteiten, studiebureaus), gevolgd door een reeks aanbevelingen, veelal beperkt tot een grondige analyse van de ICT-infrastructuren die geacht worden om op een organische manier baten te genereren op het vlak van vlotte verkeersdoorstroming. De *Smarter Cities Challenge* beperkt zich overigens niet tot mobiliteit, ook talloze andere stadskwesties komen aan bod: leefmilieu, veiligheid, bestuur, sociale diensten, economische ontwikkeling.

[9] http://smartcity.brussels/

een vooraanstaande speler en Brussel op de kaart wil zetten als "digitale hoofdstad van Europa" (DigitYser, 2017). De organisatie, officieel gelanceerd in december 2017, wordt gefinancierd door het Brussels Hoofdstedelijk Gewest en door een hele rist privéactoren (in de eerste plaats investeringsmaatschappij Sofina) en streeft ernaar de ontwikkeling van *IoT*, *big data* en virtual reality te bevorderen, door een gezamenlijke ruimte aan te bieden voor opleidingen, evenementen en start-ups. Ze organiseert ook *hackatons*, wedstrijden waarbij de deelnemers (vaak in team) moeten proberen om binnen een bepaalde tijdspanne (meestal een weekend) een probleem op te lossen aan de hand van de data die ze aangereikt krijgen[10]. Een ander voorbeeld is het initiatief van het platform voor Brusselse ondernemingen *BECI (Brussels Enterprises Commerce and Industry)*, dat sinds april 2018 een *concept store* openhoudt in een *pop-upruimte*[11], met als doel om de Brusselse bedrijven te verenigen en te dienen als katalysator voor *start-ups* rond innovatie: "De bezoekers kunnen er doeltreffende en vernieuwende mobiliteitsproducten en -diensten ontdekken in een tijdelijk en experimentgericht evenementkader"[12].

[10] Hackathon rond de aanmaak van musicalinhoud, bijvoorbeeld uit MIDI-formaten (maart 2018).

[11] Tijdelijke (of toch meestal) handelsruimte.

[12] http://www.beci.be/centre_de_connaissance/mobilite/urban_mobility_pop_up/, pagina geraadpleegd op 2 mei 2018.

3. De uitdagingen van *big data*

3.1. Kwantiteit ten koste van de kwaliteit?

Zoals we al zagen is een van de kenmerken van *big data* dat ze betekenisvolle informatie voortbrengt op basis van een omvangrijke en a priori weinig betekenisvolle dataset. Omvangrijk is echter niet noodzakelijk synoniem met betrouwbaar. Automatische datacollectie is immers geen garantie van exactheid, en de betrouwbaarheid van de gegevens hangt sterk af van de kwaliteit en de lay-out van de gegevenssensoren en van het informatietype dat vergaard kan worden (Rouvroy, 2016). Naar het voorbeeld van administratieve gegevens zijn ook massagegevens onderhevig aan vertekening, die op het moment van hun collectie niet beheerst wordt en die nadien dus geïnterpreteerd en gecorrigeerd moet worden (in tegenstelling tot enquêtes en tellingen).

3.1.1. Representativiteit of repetitiviteit?

Deze twee aspecten zijn niet gewoon maar risico's die een technische oplossing op termijn zou kunnen oplossen. Ze doen de vraag rijzen naar de representativiteit van de informatie die *big data* voortbrengt. Ze maakt het dan wel gemakkelijker om gegevens van bepaalde doelgroepen te verzamelen, dat houdt een risico in bij de extrapolatie van de resultaten naar de volledige populatie. Als men bijvoorbeeld enkel de stromen meet met mensen die een *gps* of smartphone hebben, wordt de stad of mobiliteit in het algemeen enkel benaderd afhankelijk van het gedrag van die personen (Miller, 2010).

Merk op dat die representativiteitsvertekening niet beperkt is tot het *big data*-domein, maar dat ze in het geval van enquêtegegevens doorgaans wel gekend en beheerst wordt van bij de opmaak (Chandesris *et al.*, 2017).

Vayatis (2017) stipt ook de repetitiviteit aan, zeg maar het feit dat alle waardecategorieën regelmatig terugkeren in de gemeten stromen. Die eigenschap is vooral belangrijk bij het gebruik van lerende algoritmen, die enkel die situaties die ze vooraf hebben geleerd, nauwkeurig kunnen indelen en voorspelen.

Representativiteit en repetitiviteit zijn wel in zekere zin tegenstrijdig (een representatieve datastroom zal heel zelden de minder gangbare modaliteiten aan het licht brengen). Voor beschrijvende doeleinden krijgt representativiteit de voorkeur, terwijl classificatie- en voorspellingsoefeningen om de besluitvorming te schragen, baat hebben bij een frequente herhaling van alle gevallen, om zo algoritmes te kunnen 'trainen' voor een grotere verscheidenheid aan situaties.

3.1.2. Het contextdeficit

Een eigenschap van *big data* is het gebrek aan contextuele gegevens. Dat valt allereerst te verklaren door het gebrek aan onderzoeksvraag aan het begin van het collectieproces, maar ook door de keuze van gegevens geschikt voor mathematische modellering. Door die mechanische benadering van datacollectie gaan bepaalde verklarende factoren onvermijdelijk verloren (Boyd en Crawford, 2012). Op mobiliteitsvlak bijvoorbeeld gaat het dan vaak om informatie over individuen (leeftijd, geslacht, sociaaleconomische klasse enz.) of over de eigenlijke verplaatsingen (verplaatsingswijzen, redenen voor en indrukken van een verplaatsing). Door die afbakening blijft *big data* beperkt tot het operationele karakter: het succes van een algoritme wordt afgemeten tegen de snelheid waarmee rationele informatie voor een zo laag mogelijke kostprijs wordt verkregen. Daar gaat een logica van rendement en optimalisatie achter schuil, niet van geldigheid (Rouvroy, 2016).

Zo zou bijvoorbeeld een stroomlijning van de frequentie, de rijtijden en de trajecten van het openbaar vervoer gebaseerd zijn op collectieve belangen afgeleid uit de geolokalisatie van mensen (Rouvroy, 2016). *Big data* zou dan ten dienste staan van een efficiënte stad, waar voor de gemeten problemen dan kort op de bal kan worden gespeeld, met technische bijsturingen (een alternatieve route bij verkeersopstoppingen bijvoorbeeld), zonder echter dieper in te gaan op de oorzaak van die problemen, laat staan ze aan te pakken (de oorzaak van die verkeersopstopping bijvoorbeeld). Dat efficiëntiestreven houdt dus ook het risico in dat bepaalde sociale problemen (sociaaleconomische ongelijkheid, milieukwesties enz.) worden gebagatelliseerd.

Contextuele tekortkomingen (een gebrek aan belangrijke dimensies in de gegevens) zijn evengoed problematisch voor oefeningen met operationele insteek en kunnen leiden tot aanzienlijk vervormde classificaties en voorspellingen. Om die mogelijke gebreken te verhelpen, zijn er in wezen twee mogelijke methodes: ofwel afleiding van de ontbrekende gegevens (bijvoorbeeld het geslacht van de persoon, op basis van zijn of haar e-mailadres), of door kruising met andere databronnen (eventueel van *big data*) (Vayatis, 2017). Bij dergelijke oplossingen is het uiteraard de vraag hoe de personen in kwestie anoniem worden gehouden, en hoe hun persoonlijke levenssfeer wordt gevrijwaard (zie verder).

3.2. Een stug instrument

De opkomst van *big data* is een bestuurlijke uitdaging van formaat voor overheden, in die zin dat het om een instrument gaat dat bijzonder lastig te beheersen valt.

Allereerst vereisen dergelijke gegevens zéér gespecialiseerde technische vakkennis, waardoor ze maar voor een beperkt aantal personen toegankelijk zijn. Samen met de opmars van de *big data* zien we zo ook het ontstaan van een nieuw soort expert: de *data scientist* [13]. Diens legitimiteit berust eerder op de beheersing van een breed scala aan technische instrumenten (programmering, wiskunde, geavanceerde statistiek, *datamining*, *machine learning* enz.) dan op de kennis van een bepaalde discipline (economie, sociologie, politieke wetenschappen enz.). Dat soort profielen is bovendien zeer gegeerd op de arbeidsmarkt, waardoor overheden genoodzaakt zijn om veel geld neer te tellen om ze aan te trekken. Geld dat ze niet altijd hebben.

Ten tweede komen heel wat *big data*-bronnen van particuliere actoren, wat meerdere uitdagingen inhoudt.

Allereerst, zoals we hieronder toelichten, zijn gegevens uit de private sector doorgaans al herbewerkt tot gebruiksklare, klassieke indicatoren. De extractie van signaal, van informatie is dan al gebeurd door de private actor. Dat impliceert dat er geen inzagerecht is in de verrichte handelingen, de gemaakte keuzes en de eventuele grenzen die het afgeleverde eindproduct beïnvloeden.

Bovendien is de overheidsinstantie vaak niet de eigenaar van de gegevens. Door er dan voor te kiezen om haar optreden te enten op een gegevensbron die ze niet zelf bezit, is ze afhankelijk van de productielogica's die schuilgaan achter die gegevens, met inbegrip van de verwerkingen voorafgaand aan kant-en-klare indicatoren. Daar heeft ze geen vat op, waardoor ze zich op elk moment genoodzaakt kan zien haar werkinstrumenten te herzien.

Ten slotte lijdt het geen twijfel dat zowel technologieën die sporen kunnen bijhouden (telecomnetwerken, gsm's, smartphones, *gps-toestellen* enz.) als de bevolkingsgroepen die ze gebruiken om 'geconnecteerd' te zijn, zullen evolueren. Dat houdt dus in dat de verwerkingen bedoeld om het signaal van de verzamelde gegevens te filteren, eveneens moeten worden bijgestuurd. Het gebruik van dergelijke indicatoren om tijdreeksen aan te leggen, komt derhalve sterk onder druk te staan (van der Loop *et al.*, 2017).

3.3. Het respect voor de privacy: een enorme uitdaging

Het bijhouden van persoonlijke gegevens lijkt de onvermijdelijke keerzijde van het gebruik van een veelheid aan digitale toepassingen en toestellen. Door de nieuwe heridentificatiemogelijkheden die ze biedt, doet *big data* het onderscheid tussen anonieme en persoonlijke gegevens vervagen. In het licht daarvan volstaat de anonimisering van gegevens niet langer als voorwaarde om de privacy te vrijwaren (Rouvroy, 2016).

Onderzoekers (De Montjoye *et al.*, 2013) toonden aan, door 15 maanden lang de mobiliteitsgegevens van 1,5 miljoen mensen te bestuderen, dat mobiliteitssporen vrij uniek zijn. In een databank van Proximus, een gevestigde waarde onder de Belgische telecomoperatoren, waar individuele locaties elk uur worden opgeslagen met de detailgraad die het huidige antennenetwerk toelaat, zijn er slechts vier ruimte-tijdpunten nodig om aan 95% van de sporen een individu te koppelen. Twee zijn zelfs al genoeg om meer dan de helft van de gebruikers te identificeren. Nochtans waren de databases anoniem gemaakt, zonder naam, adres of telefoonnummer. Vooral op mobiliteitsvlak houden individu's er vaak unieke schema's op na, die hun verplaatsings- en geolokalisatiegegevens in zekere zin persoonlijker maken dan hun digitale voetafdrukken (die twaalf referentiepunten vereisen om een individu te identificeren) (Grosjean, 2015).

Bovendien zijn individuen door anonimisering onvoldoende afgeschermd van de mogelijkheden om ze te profileren. Profilering is, zeg maar, een vorm van *datamining* waarbij een individu met een zekere mate aan waarschijnlijkheid kan worden ingedeeld in een welbepaalde categorie, om zo zijn of haar beslissingen te bepalen (Grosjean, 2015). Het bekendste voorbeeld van profilering is dat van gerichte internetreclame, afhankelijk van de surfgewoontes en van de websites die een gebruiker bezoekt. Uit dat onlinegedrag kan men niet enkel zijn persoonlijke eigenschappen opmaken (geslacht, leeftijdscategorie, locatie enz.), maar ook zijn voorkeuren en wensen. Op basis van dergelijke gegevens zal bijvoorbeeld een jonge vrouw die door een algoritme als zwanger is bestempeld, zwangerschapsproducten te zien krijgen op de websites die ze bezoekt (Floridi, 2012). Het hele proces doet onvermijdelijk grote ethische vragen rijzen op het vlak van de categorisering en privacy van mensen. Zodra die profilering in zekere zin toelaat om individuele gedragingen te voorspellen, kan de methode ook gebruikt worden voor de bewaking van mensenmassa's (criminele, religieuze, sociale profilering bijvoorbeeld) (Grosjean, 2015).

Vooralsnog is er zeer weinig toezicht op de datacollectie door bedrijven uit de privésector. De huidige Europese wetgeving koppelt de aard van de gegevens aan een reeks principes: een eerlijke en rechtmatige verwerking, een principe van doelbeperking, een principe van evenredigheid, relevantie en juistheid, en een duur die de tijd nodig om de doeleinden te verwezenlijken, niet overschrijdt. Een van de grote knelpunten op het vlak van *big data* is dat het einddoel van een bepaalde verwerking niet noodzakelijk op voorhand geweten is. De relevante privacywetgeving neemt wel de verzameling, de overdracht en de wijziging van persoonsgegevens in acht, maar niet de nieuwe inhoud van gegevens door hun samenvoeging (Bensamoun en Zolynsky, 2015). We moeten vaststellen dat technologische evoluties ontwikkelingen in de regelgeving op snelheid nemen.

[13] Merk op dat het de Franstalige Belgische universiteiten niet is ontgaan dat dergelijke profielen grof wild zijn op de arbeidsmarkt. De meeste onder hen (de Université Catholique de Louvain, de universiteiten van Luik en Namen) bieden sinds het nieuwe academiejaar 2017-2018 dan ook masteropleidingen in *datawetenschap* of *data science* aan.

De General Data Protection Regulation – GDPR (of Algemene verordening gegevensbescherming – AVG), goedgekeurd in 2016, werd ingevoerd als Europees regelgevend kader ter zake. De richtlijn werd in mei 2018 omgezet door de EU-lidstaten en moet leiden tot de oprichting van een uniek Europees loket belast met de effectieve toepassing van de vastgelegde regels, zowel voor de ondernemingen en instellingen binnen de EU als voor hun onderaannemers die verantwoordelijk zijn voor de verwerking van gegevens.

Big data biedt, omwille van de perspectieven die ze biedt dankzij de haarfijne indeling van klanten of de gedetailleerde reclametargeting, een enorme economische meerwaarde (Rouvroy, 2016). Volgens sommige ramingen zou de waarde van de persoonsgegevens van de Europese burgers tegen 2020 oplopen tot 1.000 miljard euro (Jourova, 2016). Europa ziet het aanscherpen van de normen inzake gegevensbescherming dan ook als een kans op commerciële opportuniteiten, en niet als een rem op innovatie.

Deze compromistekst tracht derhalve om mensen beter te beschermen, door rekening te houden met de technologische ontwikkelingen, maar zonder innovaties (en de economische waarde die ze genereren) in de weg te staan (Leonard, 2016).

De voornaamste wijzigingen uit de nieuwe verordeningen zijn (zie Jourova, 2016 en Leonard, 2016):

- Het *recht op vergetelheid*: iedereen kan eisen om zijn gegevens te laten verwijderen, zodat ze niet langer worden verwerkt.

- Het *recht op gegevensoverdraagbaarheid*: de mogelijkheid om gemakkelijk de eigen gegevens in te kijken en gratis over te dragen van de ene naar de andere dienstverlener.

- *Meer transparantie*, wat de verantwoordelijke voor de verwerking verplicht om duidelijke mechanismen te voorzien waarmee de betrokkene zijn of haar rechten kan doen gelden. Wanneer toestemming vereist is, moet die gevraagd worden met een duidelijke handeling. Dezelfde transparantie wordt verwacht voor de manier waarop de gegevens worden verwerkt.

- Het *recht om niet onderworpen te worden aan een volledig automatische beslissing*, waaronder ook profilering valt, tenzij de persoon in kwestie zijn uitdrukkelijke toelating heeft gegeven of om redenen van openbaar belang.

- De principes van *data protection by design* en *data protection by default*: beide principes zijn bedoeld om bedrijven en instellingen die persoonsgegevens te verwerken, nog meer verantwoordelijk en verplicht te maken om rekenschap af te leggen. Zij moeten gepaste maatregelen nemen, zowel in het ontwerp van de verwerking als in de daadwerkelijke uitvoering, zodat die in overeenstemming is met de privacyregels (*by design*). Het *by default*-principe op zijn beurt komt erop neer dat verwerkingsverantwoordelijken zich ertoe verbinden om de verwerking van persoonsgegevens te beperken tot het strikt noodzakelijke.

Het mag duidelijk zijn dat de *AVG* een belangrijke regelgevende stap vooruit is in de bescherming van persoonsgegevens. Indien hun rechten geschonden worden, hebben de betrokkenen recht van verweer tegen de verantwoordelijke van de verwerking of diens onderaannemer. De *AVG* voorziet in een hele reeks sancties bij inbreuken op haar bepalingen. De toezichthouder kan administratieve boetes opleggen, gaande tot 20 miljoen euro of 4% van de jaarlijkse omzet bij de meest ernstige overtredingen (Leonard, 2016).

Een aantal bepalingen blijven echter vaag in hun toepassing. Zo bestaat het hele nut van *big data* er net in om de gegevens te hergebruiken voor doeleinden die vooraf niet gepland waren. Die ambitie druist in tegen de principes van doelbepaling van de collectie, van de evenredigheid van de gegevens verzameld voor de gestelde doeleinden en van de bewaringsduur (Cytermann, 2015). Een van de uitdagingen van de ontwikkeling van *big data* wordt dus om de grens tussen gebruik voor statistische doeleinden en voor profilering duidelijk af te bakenen. Trouwens, hoewel de eerder genoemde toestemming transparenter en explicieter zou kunnen gebeuren, is ze vaak bepalend voor de toegang tot talloze diensten. Het lijkt erop dat heel wat mensen de waarde van hun persoonsgegevens onvoldoende afwegen tegen de beschikbaarheid van een dienst.

4. Enkele gebruiksvoorbeelden van *big data*

In dit hoofdstuk gaan we concreter in op de hinderpalen en de opportuniteiten bij het gebruik van *big data* in mobiliteitsstudies. We bekijken drie soorten gegevens, die heel vaak worden aangehaald: mobiele telefoniegegevens of *floating mobile data (FMD)*, *gps-gegevens* verzameld door aanbieders van boordnavigatiesystemen in motorvoertuigen, ook wel *floating car data* genoemd *(FCD)* en de gegevens met betrekking tot automatische ontwaardingen, afkomstig van de openbaarvervoersmaatschappijen.

Het gebied van de digitale sporen nuttig voor de opmaak van indicatoren voor mobiliteitsstudies is uiteraard veel ruimer. Denken we bijvoorbeeld aan gegevens van Google Maps, die informatie verstrekken over de reistijd naargelang het vervoersmiddel, aan de sporen geregistreerd door Bluetooth-sensoren of door de monitoring van mensen verbonden met een openbaar wifinetwerk, om na te gaan hoe ze zich bewegen door de openbare ruimte, of ook aan de Viapass-gegevens, die de verplaatsingen van vrachtwagens binnen België aangeven. Deze lijst is verre van limitatief.

In de onderstaande uiteenzettingen beperken we het toepassingsgebied van de aangehaalde applicaties tot het statistische gebruik van *big data*, zeg maar een gebruik opgevat in het kader van een overheidsoptreden op lange termijn.

4.1. Gegevens van de mobiele operatoren – *Floating Mobile Data* (*FMD*)

FMD worden geproduceerd door en zijn het eigendom van mobiele providers. Vanzelfsprekend gaat het om een bijproduct van hun kernactiviteit. Ze zijn zo nuttig omdat ze een heel nauwkeurig beeld kunnen geven van de ruimtelijke spreiding van de gebruikers van de telefoonnetwerken, zowel ruimte- als tijdmatig, en dat voor een heel concurrentiële prijs en termijn, vergeleken met bijvoorbeeld volkstellingen of verplaatsingsenquêtes.

In België zijn de voornaamste dataleveranciers Proximus, Orange en Base. Brussel Mobiliteit gebruikt sinds kort de gegevens van Proximus om verplaatsingen binnen het Gewest te analyseren op een heel gedetailleerde ruimte-tijdschaal (Determe, 2018).

4.1.1. Van het ruwe gegeven naar de verplaatsingen

a) Signaleringsgegevens versus facturatiegegevens

We onderscheiden twee soorten mobiele telefoondata die gebruikt kunnen worden om verplaatsingsgewoonten te schetsen: transactiegegevens en *passieve* gegevens. Het eindproduct is sterk afhankelijk van de keuze voor een van beide types of, eventueel, van de bepalingen voor hun koppeling.

Transactiegegevens omvatten de transacties waarvoor een facturatie nodig is, zoals oproepen of sms'jes, hetzij de gedetailleerde registratie van oproepen of *call detail records (CDR's)* in het Engels. Ook de uitwisseling van gegevens via internet past in die categorie. Die zijn heel precies te situeren in de tijd (op het moment van de verzending, bij het begin van de oproep) en kunnen gelokaliseerd worden in de ruimte met behulp van de zone of cel die het communicatienetwerk bestrijkt[14]. In dit geval wordt de gebruiker geïdentificeerd in het midden van de cel.

Er worden ook gegevens verzameld die losstaan van de communicatie van elke gebruiker. Die hangen samen met de verbinding die het netwerk tot stand brengt met de mobiele telefoons van de gebruikers, los van hun handelingen. Ze worden doorgaans passieve gegevens genoemd (*sighting data* (Chen *et al.*, 2016) of *signaling data* (Bonnel *et al.*, 2015)). Het gaat om communicaties die bedoeld zijn om op elk moment de positie van een mobiele telefoon op het net te kennen, om hem zo snel mogelijk te kunnen lokaliseren bij een oproep, de verzending van een sms of een andere vorm van communicatie via het internet. De vereiste ruimtelijke schaal voor die lokalisatie is evenwel veel grover dan bij factureerbare transacties. Hier spreken we dan van locatiegebieden (*location areas*), die meerdere cellen omvatten[15]. Die passieve uitwisseling tussen mobiel toestel en netwerk wordt tot stand gebracht bij het aan- en uitzetten van het toestel, wanneer de telefoon van *locatiegebied* wisselt, en wanneer de telefoon op dezelfde plek blijft, wordt zijn positie op gezette intervallen bijgewerkt.

[14] Elke zone is gekoppeld aan een basisstation, verbonden met een antenne. Die stations staan echter niet noodzakelijk centraal in elke cel en kunnen zich er zelfs buiten bevinden en meerdere cellen tegelijk bestrijken.

[15] Om een idee te geven van de grootteorde: Bonnel *et al.* (2015) tellen 32 locatiegebieden voor 10.000 basisstations in de regio Île-de-France.

Figuur 1. Voorbeelden van telefoniegegevens: *call detail records* (bovenaan) en signaleringtransacties (onderaan)
Bron: Chen *et al.* 2016

Table 1
Sample records in CDR data. [a]

X	Y	ID	Time	Duration (sec)
195925	32464	J000001	82141	81
195925	32464	J000001	82456	75
195018	31555	J000002	82100	140

[a] XY coordinates are transferred from geographical coordinate system. A conversion can be made to convert them into the absolute latitude and longitude coordinates.

Table 2
An example of the sightings data.

ID	Time [a]	Location [b]
3X35E90	1319242582	34.044162|–112.454400
3X35E90	1319242583	34.044059|–112.455550
3X35E90	1319301785	34.044392|–112.453519

[a] Time is Unix timestamp-defined as the number of seconds that have elapsed since 00:00:00 Coordinated Universal Time, Thursday, 1 January 1970.
[b] Location is the longitude and latitude coordinates of mobile phones.

b) Het pingpongeffect

Ten tweede bevinden de gebruikers zich niet altijd daadwerkelijk in de cel die de communicatie verzorgt. Hier kunnen heel wat stoorzenders in het spel zijn. Zo kan een station bij druk telefoonverkeer een deel van haar belasting afwentelen op de naburige stations. En ook de weersomstandigheden of het terrein kunnen zorgen voor een reorganisatie in de verwerking van de stromen. Als we dat dan bekijken vanuit het oogpunt van de interpretatie van gegevens, dan leiden die storende factoren tot onnauwkeurigheid in de eigenlijke locatie van de gebruiker, en kunnen ze ook aanleiding geven tot een oscillatie-effect (Chen *et al.*, 2016) of een pingpongeffect (Bonnel *et al.*, 2015), waarbij het lijkt dat de gebruiker aan hoge snelheid voortdurend van cel wisselt. In dat laatste geval zijn er verschillende manieren om de metingen bij te sturen, vooral dan om de verplaatsingen niet te gaan overschatten[16].

c) Ruimtelijke lokalisatie van mobiele apparatuur

Het mobiele telefoonnet zien als een hiërarchische structuur opgedeeld in cellen, zo ver mogelijk uitgesplitst en ingebed in overkoepelende locatiegebieden (*location areas*), is een simplistische voorstelling van de werkelijkheid, aangepast aan de noden van de oefening. Het netwerk is immers complexer dan dat: het omvat zowel een 2G-architectuur (voor het loutere gsm- en sms-verkeer), een 3G-infrastructuur (dat ook de uitwisseling van data toelaat en internetcommunicatie mogelijk maakt) en sinds kort ook een 4G-infrastructuur (om ook nieuwe objecten te laten communiceren met het netwerk). Het komt er in eerste instantie dus op aan het ruimtelijk referentiekader[17] te vereenvoudigen. Dat krijgt doorgaans de vorm van een Voronoï-diagram (Figuur 2) dat de ruimte zodanig ordent dat elk punt binnen elke veelhoek zich dichter bij het basisstation bevindt waarvan het afhangt dan bij elk ander station.

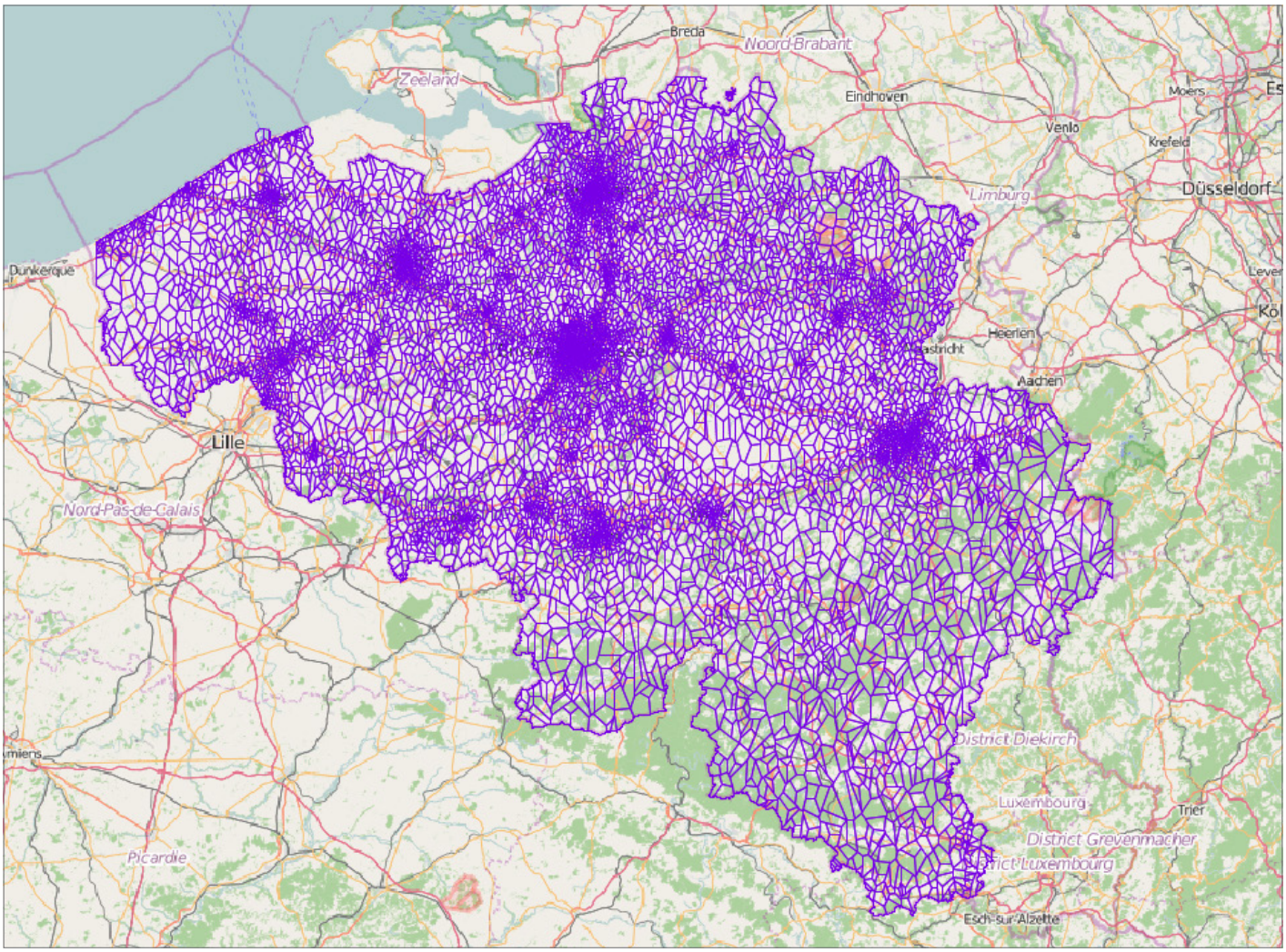
We zien meteen dat de maaswijdte van die secties sterk uiteenloopt. De dekkingsgraad van het netwerk hangt immers nauw samen met de bevolkingsdichtheid, de mazen worden dan ook breder naarmate die dichtheid afneemt. Algemeen genomen is het de intensiteit van transacties die de dimensionering van het netwerk bepaalt. Dat systeem, aanpasbaar in de tijd, zal dus niet even nauwkeurig zijn in de daluren als op momenten met veel activiteit (Ricciato *et al.*, 2015).

[16] Eigenlijk komt het erop neer dat schijnverplaatsingen worden geëlimineerd op basis van een snelheidslimiet, op basis van de perceptie van opeenvolgende heen-en-weerbewegingen of door een combinatie van de twee methodes.

[17] Debusschere *et al.* (2017) hanteren het concept van de 'technology-agnostic cell sector' (TACS), zeg maar een celsector die losstaat van de onderliggende technologie. Elke TACS bestaat uit alle cellen met hetzelfde azimut, ongeacht de gebruikte technologie.

Figuur 2. Indeling van het Belgische grondgebied in een Voronoï-diagram opgebouwd rond de basisstations van het Proximus-netwerk

Bron: Ermans *et al.*, 2017



d) Van het gebruik van het grondgebied naar de verplaatsingen

Om de verplaatsingen te bepalen, moet er een onderscheid gemaakt kunnen worden tussen de ruimte-tijdposities in stilstand ('*staying points*') en die in beweging ('*transit points*'). Twee of meerdere opeenvolgende metingen worden als stationair beschouwd en maken deel uit van hetzelfde 'staying point' indien ze zich binnen dezelfde maas bevinden en de duur tussen de eerste en de laatste waarneming binnen een gezette minimale tijdsduur-drempel ligt. Voor de gegevens van Proximus baseert de operator zich op een tijdsdrempel van een uur (Ermans *et al.*, 2017). In het andere geval worden de waarnemingen gelabeld als 'transit points'.

De verplaatsingen worden aldus gedefinieerd tussen twee opeenvolgende *staying points*, en hun trajecten kunnen nog gedetailleerder worden beschreven (in tijd en ruimte) indien men ook over tussenliggende transit points beschikt. Elke verplaatsing binnen een maas of binnen de vastgestelde tijdsdrempel wordt dus niet geregistreerd. Bonnel *et al.* (2015) toonden al aan dat de resultaten (herkomst-bestemmingsmatrices) sterk afhangen van de gekozen waarde.

4.1.2. Representativiteit en intransparantie: de voornaamste beperkingen voor het gebruik van *FMD*

Het grootste probleem van *FMD*, wanneer ze aangewend worden voor doeleinden van ruimtelijke ordening, hangt samen met hun representativiteit. In dat verband kunnen we een aantal factoren aanstippen (Chen *et al.*, 2016):

- De penetratiegraad van de operator, die sterk verschilt naargelang de bevolkingscategorie.
- Niet iedereen heeft een mobiele telefoon.
- De gebruiksintensiteit van de telefoons, die eveneens sterk verschilt naargelang de gebruiker en het type telefoon (smartphones die met internet verbonden zijn, zijn op elk moment detecteerbaar).
- Sommige gebruikers hebben meer dan één mobiele telefoon.

Bonnel *et al.* (2015) hebben bijvoorbeeld al getracht om herkomst-bestemmingsmatrices te bepalen voor passieve gegevens binnen het grondgebied van de regio Île-de-France. De vergelijking met gegevens over het pendelgedrag (werk en school), gehaald uit de telling, mondt uit in een gemengd resultaat, met aanzienlijke verschuivingen in de structuur van de matrices. De vergelijking met de Global Transport Survey (2010), die alle beweegredenen voor verplaatsingen behelst, levert betere resultaten op als het gaat om die matrixstructuur. En globaal gezien geeft het gebruik van *CDR* nuttigere resultaten.

De verkregen matrices kunnen bijgeschaafd worden met aanvullende gegevens, zoals volkstellingen, om de matrices in te stellen met een schaalfactor die toelaat om de daadwerkelijke verplaatsingsvolumes beter te benaderen (Chen *et al.*, 2016). In de producten die Proximus aanbiedt, stuurt de operator de informatie bij (de eindgebruiker kan uitgaan van een vlakfactor gekoppeld aan elke stroom, in het kader van een H/B-matrix), maar hij informeert niet over de aard van de verwerkingen die ten grondslag liggen aan die bijsturing.

De geleverde, kant-en-klare producten zijn dubbel ondoorzichtig, enerzijds wat betreft het proces voor datacollectie, en anderzijds met betrekking tot de herverwerking van de sporen alvorens de bestanden te verstrekken, die bovendien de nodige problemen met zich meebrengen om tijdreeksen aan te maken. Mobiele telefonie is immers een technologie die voortdurend beweegt, en zowel mobiele apparatuur als het netwerk evolueren jaar na jaar. Bovendien hanteert de operator targets om zijn netwerken te stroomlijnen, en die kunnen de netwerkdekking beïnvloeden los van de infrastructuur. Zonder duidelijke metadata en validatieprocessen (met eventuele correcties) door gegevens van buitenaf, is het gebruik van *FMD* om trends te ontwaren een gevaarlijk spelletje.

Op privacygebied zagen we al dat de toegang tot individuele sporen de deur zou openzetten naar eenvoudige heridentificatie van individuen. Voor de producten die Proximus aflevert zonder eerst een aanvraag in te dienen bij de Commissie voor de bescherming van de persoonlijke levenssfeer, worden de gegevens samengevoegd in herkomst-bestemmingsparen. Paren met minder dan 30 sporen worden niet gerapporteerd.

4.1.3. De concrete toepassingen van *FMD*

We kunnen het gebruik van *FMD* indelen in drie categorieën.

We kunnen *FMD*-producten opsplitsen in drie hoofdcategorieën, eigen aan het mobiliteitsonderzoek. De voornaamste categorie omvat herkomst-bestemmingsmatrices van de verplaatsingen (zie Chen *et al.* (2016) en Bonnel *et al.* (2015) bijvoorbeeld). Die matrices kunnen gebruikt worden om plaatsen te kenschetsen (bijvoorbeeld het uurgebaseerd gebruiksprofiel van een bepaalde vervoershub) of ruimtes te karakteriseren (Debusschere *et al.*, 2017).

Daarnaast zijn *FMD* prima geschikt voor de longitudinale tracering van individuen in tijd en ruimte. Zo wordt het mogelijk om verplaatsingsprofielen naar tijdstip te ontwaren of ruimte-tijdgebonden bezettingssequenties te bepalen (Bayir *et al.*, 2010).

En op basis van een voldoende lange monitoring van individuen wordt het ook mogelijk om af te leiden welke activiteit ze beoefenen op de plaatsen die ze het vaakst aandoen. Doorgaans zijn de woonplaats en de locatie van de hoofdactiviteit (werk of onderwijs) vrij eenvoudig te achterhalen (Chen *et al.*, 2016, Debusschere *et al.*, 2017).

4.2. Gegevens van providers van boordnavigatiesystemen – *Floating car data* (*FCD*)

De gegevens waar we het hier over hebben, zijn bijproducten van de boordnavigatie-instrumenten in voertuigen. Hun eigenaars verzamelen ze en voegen ze samen, om hun klanten informatie te verstrekken over de verkeerssituatie en hen, in de meeste gevallen, een routeplanner aan te bieden waarmee ze desgevallend hun reistijd tot het minimum kunnen beperken. In België zijn de voornaamste verstrekkers van dergelijke gegevens (en eventueel van de bijhorende diensten) TomTom, Inrix, Be-Mobile, Waze en ook Coyote.

In tegenstelling tot mobiele telefoongegevens meten *FCD*-gegevens vooral de verkeerssituatie. In het Nederlands gebruiken we ook wel 'mobiele data'[18], om aan te geven dat de sensoren ingebouwd zijn in de rijdende voertuigen, in tegenstelling tot de vaste sensoren op, onder, naast of boven de weg. Die gegevens kunnen bijvoorbeeld dienen om de verkeersdrukte te meten, maar hun gebrekkige representativiteit maakt het momenteel nog te complex om ze op een betrouwbare, waterdichte manier te gebruiken.

4.2.1. Van mobiele sporen naar reistijden

De *gps-data* van boordsystemen is heel uiteenlopend. Dit zijn de vaakst genoemde sensoren (men heeft het dikwijls over '*probe network*', of een netwerk van 'voelers'):

- Boordnavigatiesystemen van persoonlijke voertuigen.
- Verkeersapps voor de smartphone.
- Navigatiesystemen geïntegreerd in het fleet management (van bedrijven, meer bepaald voor goederen- en personenvervoer enz.).
- Sommige providers zouden gebruik maken van de gegevens van vaste sensoren langs de weg (Ermans *et al.*, 2017).

De *gps-locatiegegevens* worden aan de operator doorgegeven via het mobiele telefoonnet en vervolgens samengevoegd door algoritmes, die om evidente commerciële redenen niet worden prijsgegeven en eigen zijn aan elke provider. Het uiteindelijke doel is om naar elk verbonden voertuig actuele navigatie-informatie te sturen. In realtime is de voertuigstroom in heel wat segmenten echter onvoldoende om betrouwbare ramingen te kunnen opmaken. In dat geval wordt gebruik gemaakt van historische gegevens. De betrouwbaarheid van actuele data varieert dan ook sterk naargelang het segment of het tijdsbestek. Zo vermeldt TomTom een betrouwbaarheidswaarde voor elk gegeven, maar het proces voor de aanleg van die waarden blijft eerder intransparant (Ermans *et al.*, 2017).

Voor onderzoekers en experts krijgt die dataproductie de vorm van indicatoren voor de reistijden (of verplaatsingssnelheden) per routesegment dat door de operator wordt gedekt, doorgaans samengevoegd in groepjes van vrij korte intervallen (5, 15 of 30 minuten bijvoorbeeld). Dat zijn dan gegevens die hetzij overeenstemmen met een bepaalde dag en moment tijdens die dag, ofwel zogeheten historische gegevens, die worden samengevoegd in de tijd.

4.2.2. Intransparantie en representativiteit: de voornaamste beperkingen voor het gebruik van *FCD*

Het gebrek aan transparantie, zowel over de samenstelling van de sensoren als over de algoritmen die gebruikt worden om de informatie aan te leggen, is een aanzienlijke hinderpaal voor eindgebruikers van die data, omdat ze zo geen duidelijk beeld krijgen van de desbetreffende populatie of van de gemaakte methodologische keuzes. En zo is het lastig om uit te maken in hoeverre de onvermijdelijke vertekening van elke indicator beheerst wordt (Van Der Loop *et al.*, 2018).

Allereerst is het onmogelijk om de representativiteitsgraad van de verbonden sensoren te beoordelen (het geconnecteerd verkeer maakt 5% uit van alle voertuigen in Nederland (Van Der Loop *et al.*, 2018), in Île-de-France is dat 8 tot 10% (Remesy en Belloche, 2018)). Dat maakt elke meting van het verkeersvolume heikel.

Ten tweede bestaat er onzekerheid over de toetsbaarheid van de gegevens in de tijd. Operatoren hebben immers de neiging om regelmatig wijzigingen aan te brengen in hun dataextractiemethodes (net zoals in de modellering van het wegennet), om zo hun realtime dienstverlening te stroomlijnen. En dat vertekent vergelijkingen in de tijd: komen de waargenomen evoluties voort uit wijzigingen van de verkeerssituatie, uit voorbewerkingen door de provider of uit de samenstelling van de sensoren? Eén oplossing, niet altijd toepasbaar en onvermijdelijk beperkt, kan zijn om de gegevens te corrigeren door ze te koppelen aan de data van vaste sensoren (Van Der Loop *et al.*, 2018).

4.2.3. De concrete toepassingen van *FCD*

FCD concurreren voornamelijk met vaste sensoren die langs de wegen zijn geïnstalleerd om verkeersparameters te meten. In vergelijking met die parameters bieden *FCD* het grote voordeel dat ze een heel groot deel van het wegennet bestrijken, terwijl vaste sensoren doorgaans geplaatst worden op de belangrijke verkeersaders (snel- en hoofdwegen). In tegenstelling tot die laatste zijn ze daarentegen niet geschikt om verkeersvolumes te berekenen, althans niet in hun huidige staat. De reistijden en snelheden zijn nogal volatiel, waardoor hun gebruik in absolute waarden met de nodige omzichtigheid moet worden aangepakt.

Derhalve zijn de meest interessante toepassingen van deze gegevens de volgende:

Allereerst kunnen ze dienen om de verzadiging van de wegen te beoordelen, of het nu gaat om de detectie van zwarte punten, in realtime of over een langere tijdspanne (Van Der Loop *et al.*, 2018; Trotta, 2016 bijvoorbeeld), en vooral dan voor vrij betrouwbare segmenten van het wegennet. Zo is het

[18] De term slaat ook op mobiele telefoniegegevens.

mogelijk om de evolutie van de verkeersverzadiging in een wijk of rond een groot kruispunt te tonen, om zo bloot te leggen welke lokale dynamieken er schuilgaan achter de verspreiding van die verzadiging (Van Der Loop *et al.*, 2018). Een andere toepassing bestaat erin om de gemiddelde reistijden en de vertragingen door verkeersopstoppingen op welbepaalde routes te berekenen. Zo wordt het mogelijk om routes onderling te vergelijken of, door een specifiek segment of traject in aanmerking te nemen, om snelheids- en congestieprofielen op te stellen naargelang het tijdstip (Ermans *et al.*, 2017).

De aanleg van dergelijke opstoppingsindicatoren is doorgaans gebaseerd op een vergelijking van de waargenomen snelheden (of reistijden) binnen een segment van het net of voor een traject binnen dat net, vergeleken met een referentiesnelheid (of -reistijd). Aan de basis van die referentiewaarde kunnen verschillende congestiebenaderingen liggen. Die waarde kan dan overeenstemmen met een snelheid die willekeurig als aanvaardbaar wordt beoordeeld (de arbitraire benadering), met de snelheid bij maximale bezetting van de weg (de technische aanpak, de 'ingenieursbenadering') of ook met de optimale bezetting van de weg (economische benadering), anders gezegd een situatie waarbij bij maximaal debiet de vlotte doorstroming toch nog gegarandeerd is (Reymond, 2005). Die laatste benadering krijgt vaak de voorkeur.

Merk ook op dat er een onderscheid moet worden gemaakt tussen de opstoppingsindicatoren die de operatoren zelf verspreiden en de indicatoren die onderzoekers en experts vaststellen op basis van de onderliggende gegevens die diezelfde operatoren prijsgeven. Zo publiceren Inrix en TomTom regelmatig congestiegegevens voor de steden (Inrix index, TomTom traffic index). Bovenop het geringe vertrouwen dat we in tijdsvergelijkingen moeten stellen, komt de weinig eensluidende definitie (meer bepaald met betrekking tot dichtheids- en verstedelijkingscriteria) van het begrip stad, die ook bij ruimtelijke vergelijkingen tot een zekere terughoudendheid noopt (Van Der Loop *et al.*, 2018). Bovendien wordt de congestiegraad beoordeeld op basis van een standaard inkomende route van een pendelaar, de categorie dus die het meest met verkeersopstoppingen te maken krijgt (Ermans en Brandeleer, 2016). Dat onderstreept ongetwijfeld het belang van marketingcommunicatie bij de publicatie van die indicatoren (Van Der Loop *et al.*, 2018).

Ten tweede komt de kennis van de reistijden van pas om toegankelijkheidsindicatoren op te maken, naar het voorbeeld van Lebrun (2018) bij zijn analyse van het openbaar vervoer in het Brussels Hoofdstedelijk Gewest. Op die manier kunnen we de bereikbaarheid van een locatie beoordelen door een samenvattende positie-indicator (gemiddeld of meestal mediaan) te berekenen voor alle reistijden tot aan die plek (toegankelijkheid van de bestemming) of van die plek naar elke andere locatie (toegankelijkheid vanaf de herkomst) voor een welbepaald deel van het grondgebied. Daarnaast is het ook mogelijk om slechts een bepaald aantal plaatsen uit te kiezen die belangrijk zijn gezien de functies die ze huisvesten (woningen, werk, school, cultuur enz.).

Reistijden kunnen ook van pas komen om verplaatsingsmodellen te maken[19]. Ze kunnen in beschouwing worden genomen bij de verdeling van de vraag tussen plaats van herkomst en bestemming, bij de attributie van de vervoerskeuzes of ook bij de toewijzing van de vraag op het net. Ze worden dan vooral gebruikt als maatstaf voor de afstand, voor de weerstand tegen een verplaatsing (we spreken hier van 'impedantie') tussen twee beschouwde plaatsen (Ermans *et al.*, 2017).

Tot slot, afgezien van het feit dat het gebrek aan vergelijkbaarheid in de tijd (in het bijzonder van jaar tot jaar) nog steeds een obstakel vormt, kunnen *FCD* handig zijn om maatregelen of infrastructuurwijzigingen ex-post te evalueren. Die aanpak werd onlangs nog uitgeprobeerd met de afbraak van het Reyersviaduct (Servonnat, 2017).

4.3. Ticketinggegevens van de openbaarvervoermaatschappijen

Onder ticketinggegevens verstaan we hier alle data die de exploitanten verzamelen wanneer gebruikers hun vervoersbewijs ontwaarden op het net via een automatisch ontwaardingssysteem met *RFID*-technologie (*RFID* staat voor *Radio Frequency IDentification*). Het eerste doel van dergelijke systemen is uiteraard om vervoersbewijzen te ontwaarden[20], maar daarnaast produceren ze ook voortdurend een massa gegevens die gebruikt kunnen worden om de mobiliteit van mensen op het net te onderzoeken en de exploitant te helpen om zijn dienstverlening te verbeteren.

Voor de eindgebruiker, doorgaans de exploitant zelf, bieden de gegevens heel wat mogelijkheden. Omdat hij immers eigenaar is van de gegevens, zijn de bijkomende kosten (naast de aanzienlijke investeringen om de infrastructuur op poten te zetten en draaiende te houden) aan de lage kant. Bovendien kan de exploitant de verzamelde sporen en verwerkingen voorafgaand aan de analyse helemaal zelf beheersen. Vanuit dat oogpunt gaat het om een configuratie die nogal sterk verschilt van die voor *FMD* en *FCD*.

Ticketinggegevens hebben een nut dat te vergelijken is met dat van *FMD:* ze bieden de mogelijkheid om de gebruikers op het net (per station en per halte) heel nauwkeurig te onderscheiden in de tijd en voor een heel breed registratiebestek (terwijl het netwerk opereert in feite). Dankzij de unieke identificatiecode op elke kaart kan elke gebruiker van het net gevolgd en gemonitord worden. Dit soort gegevens is uiteraard beperkt tot de verplaatsingen van de beschouwde openbaarvervoergebruikers.

In het Brussels Hoofdstedelijk Gewest hebben de MOBIB-kaarten sinds 2016 een *RFID*-chip, die dient als drager bij de aankoop en ontwaarding van alle soorten vervoersbewijzen die de MIVB uitgeeft[21]. De gegevens die via die chip verzameld worden, houden een aanzienlijk potentieel in op het vlak van verplaatsingsanalyses, maar hun gebruik vertoont nog wat kinderziektes. We denken dan bijvoorbeeld aan de gebrekkige ontwaarding in stations en aan haltes zonder poortjes, een probleem dat zich vooral bij het verlaten stelt (omdat je je vervoersbewijs dan niet verplicht moet ontwaarden en zelden eerst terug door een poortje moet), maar ook bij het opstappen. Enkel op het ondergrondse net moet je immers altijd eerst door poortjes en ben je dus verplicht om je vervoersbewijs te ontwaarden.

[19] Modellen van uiteenlopende schaal en toepassing die bedoeld zijn om de verplaatsingsvraag binnen een gegeven grondgebied in te schatten en, naargelang het geval, die toe te wijzen aan een transportnet, volgens uiteenlopende modaliteiten voor vervoerswijzeparameters en verplaatsingsredenen. Zie ook het Franse *Centre d'Etudes sur les Réseaux* (2003) voor een discussie over dit thema.

[20] Zie Pelletier *et al.,* 2011 voor een overzicht van de voor- en nadelen die de wetenschappelijke literatuur opwerpt voor dit systeem.

[21] De MIVB is de grootste openbaarvervoersmaatschappij binnen het Brussels Hoofdstedelijk Gewest, maar zeker niet de enige: ook De Lijn, de TEC en de NMBS hebben hun aandeel in het globale aanbod.

4.3.1. Van ticketspoor tot verplaatsing: bestemmingen en aansluitingen inschatten

Je vervoersbewijs ontwaarden bij het uitstappen of bij het verlaten van het station is doorgaans niet verplicht, waardoor er vaak niet meteen informatie beschikbaar is over de plaats en het tijdstip van beëindiging van een openbaarvervoerstraject. Om de bestemmingen op het metronet in Rennes in te schatten, namen Briand *et al.* (2017), daarbij geïnspireerd door de oplossing voorgesteld door Trépanier *et al.* (2007) als bestemming van een welbepaald traject steeds het dichtstbijzijnde station, uit een lijst van mogelijke bestemmingen, gebruikt bij de eerstvolgende ontwaarding (bij het binnengaan van dat station dus). Indien de afstand tussen beide stations meer dan 500 meter bedroeg, gingen ze ervan uit dat het station van bestemming niet bekend is. Voor de bestemming van de laatste ontwaarding van elke dag wordt het eerste ontwaardingsstation tijdens diezelfde dag in aanmerking genomen. De hypothese die aan de basis van dit algoritme ligt, is dat gebruikers die zich op het net verplaatsen afstanden die ze met het openbaar vervoer kunnen afleggen, niet te voet (of op een andere manier) doen.

Zodra de op- en afstapketen van een voertuig dat rondrijdt op het net, bepaald is, moet er enkel nog een onderscheid worden gemaakt tussen de bewegingen om een overstap te maken en de bewegingen die het einde van een verplaatsing inluiden. In Briand *et al.* (2017) spreekt men van een overstap wanneer het verschil tussen het tijdstip van aankomst in het afgeleide station van bestemming en het tijdstip van vertrek in het volgende station (met vastgestelde ontwaarding) minder dan 30 minuten bedraagt. Als die tijdspanne langer is, dan gaat men ervan uit dat het om twee verschillende verplaatsingen gaat, met hun eigen en aparte reden. Dezelfde logica wordt gehanteerd in andere voorbeelden van het gebruik van variabele tijdsintervallen (Ma *et al.*, 2013 ; Ma *et al.,* 2017). Net zoals bij de *FMD* bestaat ook hier het risico dat eventuele korte ritten worden genegeerd.

Merk tot slot nog op dat het ook mogelijk is om de stations het dichtst bij de woon- en de werkplaats (desgevallend) af te leiden aan de hand van de ontwaardingsfrequenties en -tijdstippen.

4.3.2. Technisch beheer en verwerking van indicatoren: de voornaamste beperkingen voor het gebruik van ontwaardingsdata

Allereerst is het beheer van heel grote datavolumes niet zomaar vanzelfsprekend voor openbaarvervoersmaatschappijen, *in het bijzonder* bij netwerken met meerdere operatoren, die allemaal ook nog eens hun eigen collectieprocessen en -formaten gebruiken (zie Chandesris *et al.* (2017) voor het voorbeeld van het openbaarvervoersnet in Île-de-France). De technische antwoorden op die uitdagingen passen niet altijd in de alledaagse core business van de exploitant, waardoor het hem soms ook ontbreekt aan de juiste middelen. Ten aanzien van *FMD* en *FCD* mag de totale beheersing van ontwaardingsgegevens dan wel de transparantie en de betrouwbaarheid ten goede komen, ze vereist ook een aanzienlijke investering in het juiste personeel.

Ten tweede zitten er vaak lacunes in de ticketingdata, wat de representativiteit van de indicatoren vertekent. Die lacunes hebben te maken met verschillende factoren, denken we aan vervoersfraude, het feit dat niet alle stations over volwaardige ontwaardingsinfrastructuur beschikken, of ook de mogelijkheid die de gebruiker heeft om het net te verlaten zonder zijn vervoersbewijs te moeten aanbieden. In bepaalde gevallen bestaan de automatische ontwaardingssystemen naast manuele validatiesystemen. In dat laatste geval wordt een deel van de verkeerssporen dus niet geregistreerd.

Op het niveau van de totale reizigersstromen, samengevoegd per herkomst- en bestemmingspaar, analyseren Chandesris *et al.* (2017) de situatie van het transportnet in Île-de-France (SNCF, RATP, Optile), waarbij ze stuiten op het probleem dat niet alle trein- en metrostations bij de in- en de uitgang zijn uitgerust met sensoren. Om dat probleem te verhelpen, corrigeren ze de geregistreerde trajecten op basis van een voorafgaand bemonsteringsplan ("Alles gebeurt alsof de gegevens werden verzameld volgens dat plan") met een gelaagde structuur op meerdere niveaus, om de complexe ruimte- en tijdsverdeling in rekening te brengen. Vervolgens worden de gegevens gecorrigeerd door de matrixelementen af te stemmen op de marges, waarbij een beroep wordt gedaan op aanvullende bronnen om het algemene gebruik van het vervoersnet te ramen (tellingen, ticketverkoop enz.).

Ten derde ontbreekt in de verzamelde gegevens elke context over de eigenschappen van individu's, meer bepaald vanuit sociaal-economisch standpunt (leeftijd, geslacht, socio-professionele status enz.). Om de analyses te verrijken, kunnen die gegevens worden toegerekend op basis van een toetsing met commerciële databanken (het abonnementtype bijvoorbeeld), bij afleiding op basis van de gedragingen op het netwerk, of via de twee tegelijk (Briand *et al.*, 2017). Het afleiden van ontbrekende contextgegevens heeft wel onvermijdelijk tot gevolg dat hun betrouwbaarheid niet ten volle gegarandeerd kan worden. Voor handelingen die het heridentificatiepotentieel van individuen vergroten, moeten de bevoegde privacyautoriteiten hoe dan ook vooraf hun toestemming geven.

Gegevens over de verplaatsingsredenen ontbreken uiteraard ook. Het is mogelijk om de functie van bepaalde bestemmingen af te leiden aan de hand van de verplaatsingsfrequentie en het tijdstip van verplaatsing (meestal gaat het dan om de woon- of werkplaats), wat toelaat om aan die bestemmingen een beweegreden te koppelen (naar huis of naar het werk gaan). Kusakabe en Asakura (2014) stellen ook een methode voor om verplaatsingspatronen opgemaakt uit ticketinggegevens te kwalificeren, door de ticketinggegevens te koppelen met data uit enquêtes naar de vervoersgewoontes[22].

[22] Die koppeling gebeurt voor de vertrek- en aankomsttijdstippen en -stations op het net en de afgeleide redenen van de verdeling vastgesteld in de enquête naar de vervoersgewoontes.

4.3.3. De concrete toepassing van ticketinggegevens

De analyses die mogelijk worden dankzij die gegevens, leunen logischerwijs nauw aan bij de analyses op basis van mobiele data. We kunnen ze indelen in drie verschillende categorieën.

Eerst zijn er de herkomst-bestemmingmatrices naargelang het moment van de dag. Mochten we beschikken over aanvullende contextgegevens, dan wordt het ook mogelijk om die matrices op te stellen naar gebruikerstype (student, werknemer, scholier), naargelang de beschikbaarheid van aanvullende variabelen.

Ten tweede zijn er de standaardprofielen van de gebruikers afhankelijk van hun gebruik van het net (verplaatsingsfrequentie en -tijdstip) tijdens een gegeven dag (Briand *et al.*, 2017), eventueel naargelang de aard van de dag (werkdag, zaterdag, zondag), of ook voor een periode van een week (Ma *et al.,* 2013) of een maand (Ma *et al.*, 2017). Zo gebruiken Ma *et al.* (2013) de ticketinggegevens van het openbaarvervoersnet in Beijing om te achterhalen welke routes frequent gebruikt worden, en met welke regelmaat die routes gebruikt worden (in termen van vertrekuren, dagelijkse drukte enz.).

Ten derde is het mogelijk om de stations te kenschetsen aan de hand van hun gebruik in de tijd.

Volgens Pelletier *et al.* (2011) kunnen de analyses van ontwaardingsgegevens in opdracht van openbaarvervoersexploitanten worden ingedeeld in drie groepen: (1) strategische studies voor besluitvorming op lange termijn (uitbreiding/wijziging van het net, voorspelling van de vraag, inschatting van evoluties in de gewoontes van gebruikers); (2) tactische studies, bedoeld om op verzoek diensten bij te sturen (wijziging van de dienstregeling, aanpassing van de frequentie, aanpassing van aansluitingen enz.); (3) operationele baten (prestatie-indicatoren (zoals regelmaat, snelheid, stiptheid), tariefflexibiliteit, verbetering van het ticketsysteem enz.). Aan die lijst kunnen we nog de mogelijkheid toevoegen om de impact van planningmaatregelen op de vraag te evalueren. Briand *et al.* (2017) evalueren bijvoorbeeld het effect van de afvlakking van de ochtendspits, toe te schrijven aan het latere startuur van de studenten aan de universiteit van Rennes.

De nieuwe analyseperspectieven die ontwaardingsgegevens bieden, worden zeker op gejuich onthaald in de vakliteratuur, vooral dan omdat ze een onafgebroken monitoring mogelijk maken. Maar heel wat auteurs benadrukken dat naast die nieuwe informatiebronnen er ook nog steeds klassieke verplaatsingsenquêtes afgenomen moeten worden, enkel zo is een omvattende meerdimensionale analyse immers mogelijk (Briand *et al.*, 2017). Globaal genomen blijven parallelle gegevensbronnen nodig, zowel om ontbrekende gegevens af te leiden of te ramen als voor de validering van de teruggekoppelde informatie.

4.4. Vergelijking van de aangehaalde exploitatievoorbeelden

Om deze paragraaf te besluiten, zetten we nog een aantal eigenschappen op een rij die de drie aangehaalde gegevensbronnen lijken te kenschetsen of te onderscheiden. Allereerst zetten ze allemaal een signaalterugkoppeling in gang die maakt dat de omschakeling van spoor naar indicator een uiterst complex proces is, waarvoor aanvullende gegevens nodig zijn om eventuele correcties door te voeren en, vooral, het hele proces te valideren.

Ten tweede is het belang van indicatoren voor de exploitant die ze aanlegt en voor zijn *core business* heel uiteenlopend, en dat kan gevolgen hebben voor de kwaliteit. Zo hebben mobiele telefonieproviders *a priori* niets aan gedetailleerde kennis van de verplaatsingen van personen als het gaat om de levering van hun belangrijkste diensten. Omgekeerd maken *FCD* het hart uit van boordnavigatiediensten, ook al vallen de doelstellingen van de aanbieders ervan niet helemaal samen met die van de onderzoekers (zoektocht naar de optimale route vanuit de invalshoek van de trajectduur vs zo getrouw mogelijke weergave van de reistijden op het wegennet). Tot slot gaat het openbaarvoersmaatschappijen er in de eerste plaats om hun vervoersaanbod te beheren en te plannen op verschillende niveaus (strategisch, tactisch, operationeel), waardoor ze dus ook aandachtig zijn voor de geldigheid ervan.

Ten derde lijkt het erop dat het proces voor de aanleg van indicatoren wat minder gedocumenteerd is voor particuliere actoren, vooral dan met het oog op het bewaren van bedrijfsgeheimen. Dat maakt het lastiger om de geldigheid van data te verzekeren, en vooral dan om ze te gebruiken om trends te ontwaren.

	FMD	FCD	Ontwaardingsgegevens
Eigenaar	Privaat	Privaat	(para)publiek
Rol van de indicatoren aangelegd voor de activiteiten van de exploitant	Beperkt	Commerciële diensten	Beheer van het net op strategisch, tactisch en operationeel niveau
Datacollectie en -verwerking	Ondoorzichtig	Ondoorzichtig	Transparant
Validatie vereist	Ja	Ja	Ja
Tijdgebonden vergelijkbaarheid	-	-	+/-

Algemene conclusie

We hebben aangetoond dat de opkomst van *big data* leidt tot een paradigmaverschuiving in de productie van kennis. Die verschuiving hangt in de eerste plaats samen met de kenmerken van dergelijke gegevens: (1) massale gegevens (de 3 V's), die een aanzienlijke technische beheersing vergen en (2) de afstand tussen ruwe data en informatie, het signaal dat eruit gefilterd kan worden. Die informatie vereist een verwerking voorafgaand aan de analyse, die zowel breed als gericht is en een eerder inductieve als deductieve richting volgt.

Voorts leidt *big data* ook tot een soms positivistisch enthousiasme en een heel specifieke nabootsing, vooral dan in de uitbouw van *smart cities*. In dat opzicht onderscheidt *big data* zich misschien minder als nieuw gegevenstype, maar eerder als opkomend kader voor kennisproductie, voor nieuwe manieren om *de wereld betekenisvoller te maken* (Rouvroy, 2014). Het belang van methodes voor signaalterugkoppeling en vergrotingseffect kunnen een natuurlijk objectiviteitsgevoel voeden dat we met een korrel zout moeten nemen, enkel zo kunnen we immers een verarming van de datadiversiteit en de analysemethodes vermijden.

Daarnaast zet de promotie van *big data* binnen het denkbeeld van de *smart city* aan tot een dynamisch overheidsoptreden, waar de prikkels waargenomen door *het internet of things* meteen geïnterpreteerd zouden moeten worden en omgezet in acties, vanuit de ambitie om de stedelijke mobiliteit in realtime te optimaliseren (herverdeling van het verkeer, sturing van het vervoersaanbod en van de wegcapaciteit enz.). In dat opzicht behelst de hele *big data*-problematiek ook de vraag hoe we al die geproduceerde kennis vertalen naar reële *acties*. Het lijkt ons dan ook belangrijk om tegengewicht te bieden door ook een kennisproductie aan te houden die van nut kan zijn voor programmering op langere termijn (bijvoorbeeld in het kader van strategische studies) en die daarnaast toelaat om, los van de kortetermijnrationalisering van verplaatsingsstromen, de mobiliteitsproblematiek op elk moment in vraag te stellen.

Het gebruik van *big data* kan uiteraard op dezelfde manier worden beschouwd als de 'traditionele' instrumenten en indicatoren (enquêtes, tellingen, administratieve gegevens) in het kader van een kennisproductie met een eerder strategisch opzet. Door *big data* terug zodanig te herdimensioneren dat ze een instrument wordt, naast andere in het hele arsenaal aan methodologische tools, biedt ze mogelijkheden voor mobiliteitsstudies die in essentie berusten op de exhaustiviteit en de zeer fijne korrelgrootte van de ruimte-tijdsdekking, die doorgaans niet gehaald wordt met enquêtes of vaste sensoren (in het bijzonder voor verplaatsingen tijdens de daluren of parameters voor het autoverkeer over het hele wegennet), vaak weliswaar om budgettaire redenen.

Afgezien van die mogelijkheden vervangt *big data* evenwel niet de klassiekere databronnen. Allereerst ontbreekt het immers aan contextdata (leeftijd, geslacht, sociaal-economische eigenschappen als het om individuen gaat; wijzen, redenen en indrukken als het om verplaatsingen gaat), een leemte die met andere bronnen moet worden opgevuld. Daardoor lenen ze zich minder tot oorzakelijkheidsanalyses en zijn ze eerder geschikt voor het beschrijven en kenschetsen van stromen en verplaatsingen.

Ten tweede schuilt de complementariteit met de klassiekere gegevensbronnen eveneens in de onontbeerlijke validatie en de eventuele bijsturing van de gegevens, voor zover ze nog een aantal verwerkingen ondergaan na hun collectie (vermoedensgebaseerde aanvulling van ontbrekende gegevens, van afwezige variabelen, aanpassingen enz.). Vanuit dat oogpunt lijkt het gebruik van *big data* in zekere zin onlosmakelijk verbonden met het bestaan van ijkpunten die het mogelijk maken om die massale gegevens te valideren. Dat zet ons ertoe aan de betaalbaarheid (in deze fase alleszins!) van kant-en-klare *big data* van privéoperatoren toch wat te relativeren, aangezien ook de kostprijs voor de validatie (werkuren, verwerking en productie van vergelijkingsdata enz.) in rekening moet worden gebracht.

Ten derde lijdt het geen twijfel dat zowel technologieën om sporen te registreren (telecomnetwerken, *gps-systemen* enz.) als de mensen die ze gebruiken om 'geconnecteerd' te zijn, zullen blijven evolueren. De verwerkingen bedoeld om het signaal te filteren uit alle verzamelde data, zullen dus ook aangepast moeten worden. Dat houdt in dat de validatieprocedures geregeld vernieuwd moeten worden. Die vluchtigheid in het hele productieproces van *big data*-indicatoren doet ook vragen rijzen over hun bruikbaarheid voor de aanleg van tijdgebonden datasets die de mogelijkheid bieden om trends bloot te leggen.

Die complexe aard van het productieproces is nog problematischer bij private dataproducenten die weinig vertellen over hun productieproces, vooral dan om hun bedrijfsgeheimen niet prijs te geven, waardoor we moeten afvragen hoe de verhoudingen liggen tussen overheidsinstanties en private operatoren. Vanuit die invalshoek ziet het ernaar uit dat de eerste er alle belang bij hebben om te overleggen met de tweede, om hun noden kenbaar te maken en samen een dataproductieproces op te zetten. Voor overheden komt het erop aan de kwaliteit en de goede beheersing van de indicatoren te verbeteren en tegelijk voor haar werkmethodes en middelen zo min mogelijk afhankelijk te zijn van de productielogica's van de privépartij. Voor die laatste kan het een uitgelezen kans zijn om haar statistische producten te laten valideren door een openbare instelling, wat de zichtbaarheid en de legitimiteit ervan ten goede komt.

Bibliografie

Allemand L. (2013), "Dossier – Les promesses du big data", *La Recherche*, december 2013, pp. 27-42.

Anderson C. (2008), "The end of theory?: The data deluge makes the scientific method obsolete", *Wired Magazine*, vol. 16, n°7.

Batty M. (2013), "Big data, smart cities and city planning", *Dialogues in Human Geography*, vol. 3, n°3, pp. 274-279.

Batty M., Axhausen K.W., Giannotti F., Pozdnoukhov A., Bazzani A., Wachowicz M., Ouzounis G. en Portugali Y. (2012), "Smart cities of the future", *The European Physical Journal Special Topics*, vol. 214, n°1, pp. 481-518.

Bayir M.A., Demirbas M. en Eagle N. (2010), "Mobility profiler : A framework for discovering mobility profiles of cell phone users", *Pervasive and Mobile Computing*, vol. 6, n°4, pp. 435-454.

Bensamoun A., Zolynsky C. (2015), "Cloud computing et big data. Quel encadrement pour ces nouveaux usages des données personnes ?", *Réseaux*, n°189, pp. 103-121.

Bonnel P., Hombourger E., Olteanu-Raimond A.-M. en Smoreda Z. (2015), "Passive Mobile Phone Dataset to Construct Origin-destination Matrix : Potentials and Limitations", *Transportation Research Procedia*, vol. 11, pp. 381-398.

Boullier D. (2015), "Les sciences sociales face aux traces du big data ? Société, opinion et répliques", *FMSH*, WP, n°88.

Boyd D. en Crawford K. (2012), "Critical questions for Big Data : Provocations for a cultural, technological, and scholarly phenomenon", *Information, Communication & Society*, vol. 15, n°5, pp. 662-679.

Brandeleer C. en Ermans T. (2016), "Quand gérer des feux de circulation préfigure des choix de mobilité : les enjeux stratégiques d'un outil technique", *Brussels Studies*, n° 103.

Briand A.-S., Côme E., Coulombel N., El Mahrsi M.K., Munch E., Richer C. en Oukhellou L. (2017), "Projet MOBILLETIC, Données billettiques et analyses des mobilités urbaines : le cas de Rennes", *in* André De Palma en Sophie Dantan (dir.), *Big data et politiques publiques dans les transports*, Parijs, Economica, pp. 174-196.

Brusselse Hoofdstedelijke Regering (2014), "Ontwerp van meerderheidsakkoord 2014-2019", Brussel, Be.Brussels.

CERTU (2003), "Modélisation des déplacements urbains de voyageurs : guide des pratiques". Centre d'études sur les réseaux, les transports, l'urbanisme et les constructions publiques, Ministère de l'équipement, des transports, du logement, du tourisme, et de la mer, Direction des transports, Lyon, 244 p.

Chandesris M., Ganansia F. en Remy A. (2017), "Les données massives au service des mobilités de demain", *in* André De Palma en Sophie Dantan (dir.), *Big data et politiques publiques dans les transports*, Parijs, Economica, pp. 138-169.

Chen C., Ma J., Susilo Y., Liu Y. en Wang M. (2016), "The promises of big data and small data for travel behavior (aka human mobility) analysis", *Transportation Research Part C : Emerging Technologies*, vol. 68, pp. 285-299.

CIBG (2014), "Witboek 2014-2019", Centrum voor Informatica voor het Brusselse Gewest, Brussel, 71 p.

CIBG (z.d.), "Smart city strategie", Centrum voor Informatica voor het Brusselse Gewest, Brussel, 4 p.
https://cibg.brussels/nl/bestanden/brussels-smart-city-strategie

Commenges H. (2014), "La mobilité comme variabilité temporelle de la présence spatiale", *Flux*, 2014, 1(95), pp. 41-55.

Connelly R., Playford C. J., Gayle V., Dibben C. (2016), "The role of administrative data in the big data revolution in social science research", *Social Science Research*, n°59, pp. 1-12.

Cytermann L. (2015), "Promesses et risques de l'open et du big data : les réponses du droit", *Informations sociales*, n°191, pp. 80-90.

Debusschere M., Lusyne P., Dewitte P., Baeyens Y., De Meersman F., Seynaeve G., Wirthmann A., Demunter C., Reis F. en Reuter H.I. (2017), "Big data et statistiques. Un recensement tous les quarts d'heure...", Brussel, Algemene directie Statistiek – Statistics Belgium, 22 p.

De Montjoye Y.-A., Hidalgo C., Verleysen M., Blondel V. (2013), "Unique in the crowd : the privacy bounds of human mobility", *Scientific reports*, vol. 23, n°1376, pp. 1-5.

De Palma A. (2017), "Tour d'horizon et repérages", *in* André De Palma en Sophie Dantan (dir.), *Big data et politiques publiques dans les transports*, Parijs, Economica, pp. 15-35.

Determe K. (2018), "Analyse macro des flux de mobilité grâce aux données de téléphonie mobile". Voorgesteld tijdens de 67ste vergadering van de gewestelijke mobiliteitscommissie, Brussel, 23 april.

DigitYser (2017), "DigitYser opent "digitaal clubhuis" voor Nextondernemers in hartje Brussel", 18 december 2017.

Douay N. en Henriot C. (2016), "La Chine à l'heure des villes intelligentes", *L'Information géographique*, vol. 80, n°3, pp. 89-102.

Ermans T., Haynes J., Kluyskens E. en Servonnat G. (2017), "Inventaire et possibilités de croisements de sources de données statistiques sur la mobilité des personnes en Belgique", Brussel, FOD Mobiliteit en Vervoer, 111 p.

Fayyad U., Piatetsky-Shapiro G. en Smyth P. (1996), "From data mining to knowledge discovery in databases", *AI magazine*, 1996, pp. 37-54.

Floridi L. (2012), "Big Data and Their Epistemological Challenge", *Philosophy & Technology*, vol. 25, n°4, pp. 435-437.

Graham M. en Shelton T. (2013), "Geography and the future of big data, big data and the future of geography", *Dialogues in Human Geography*, vol. 3, n°3, pp. 255-261.

Grosjean A. (2015), "Le profilage : un défi pour la protection des données à caractère personnel", *in* Grosjean (dir.) *Enjeux européens et mondiaux de la protection des données personnes*, Brussel, Ed. Larcier, pp. 277-310.

IDC (2014), "The digital universe of opportunities – Rich data and the increasing value of the internet of things", EMC, 17 p.

Jourova V. (2016), "La réforme de la protection des données dans l'UE et les mégadonnées", technische fiche, Europese Commissie, directoraat-generaal Justitie en Consumentenzaken, 4 p.

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

Laney D. (2001), "3D data management : controlling data volume, velocity, and variety", Meta Group, 4 p.

Lebrun K. (2018), "Temps de déplacements en transport public à Bruxelles: l'accessibilité des pôles d'activités", *Brussels Studies*, n°123.

Leonard T., Chaumont D. (2016), "Commentaire général du GDPR", Ulys, Brussel, 19 p.

Ma X., Liu C., Wen H., Wang Y. en Wu Y.-J. (2017), "Understanding commuting patterns using transit smart card data", *Journal of Transport Geography*, vol. 58, pp. 135-145.

Ma X., Wu Y.-J., Wang Y., Chen F. en Liu J. (2013), "Mining smart card data for transit riders' travel patterns", *Transportation Research Part C : Emerging Technologies*, vol. 36, pp. 1-12.

Malaurent J. (2017), "Big data : enjeux et applications pour appréhender la mobilité", *in* André De Palma en Sophie Dantan (dir.), *Big data et politiques publiques dans les transports*, Parijs, Economica.

Miller H.J. (2010), "The data avalanche is here. Shouldn't we be digging ?", *Journal of Regional Science*, vol. 50, n°1, pp. 181-201.

Pelletier M.-P., Trépanier M. en Morency C. (2011), "Smart card data use in public transit : A literature review", *Transportation Research Part C: Emerging Technologies*, vol. 19, n°4, pp. 557-568.

Remesy R. en Belloche S. (2018), "Floating car data : quel bilan pour la gestion du trafic ? ", *TEC*, april, n°237, pp. 38-39.

Reymond M. (2005), "La tarification de la congestion automobile : Acceptabilité sociale et redistribution des recettes du péage", Thèse de doctorat, Montpellier, Université Montpellier 1, 339 p.

Ricciato F., Widhalm P., Craglia M. en Pantisano F. (2015), *Estimating population density distribution from network-based mobile phone data*, Luxemburg, Publications Office of the European Union, 70 p.

Rouvroy A. (2016), "Des données et des hommes. Droits et libertés fondamentaux dans un monde de données massives", Rapport de recherche, Bureau du comité consultatif de la convention pour la protection des données à l'égard du traitement automatisé des données à caractère personnel, Conseil de l'Europe, 45 p.

Rouvroy A. (2014), "Des données sans personne : le fétichisme de la donnée à caractère personnel à l'épreuve de l'idéologie des Big Data", *in* Conseil d'Etat (dir.) *Le numérique et les droits fondamentaux,* La Documentation française, pp. 407-421.

Servonnat G. (2017), "Big data – The theory of chaos", Results Presentation BCUS Civil Society Fellowship 2017, Brussel, 19 december 2017.

Trépanier M., Tranchant N. en Chapleau R. (2007), "Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System", *Journal of Intelligent Transportation Systems*, vol. 11, n°1, pp. 1-14.

Trotta M. (2016), Wat vertellen gps-data over de snelheid op onze wegen? – Gedragsmeting: snelheid buiten de bebouwde kom 2015", Onderzoeksrapport nr. 2016-R-03-NL, Brussel, Belgisch Instituut voor de Verkeersveiligheid – Kenniscentrum Verkeersveiligheid, 50 p.

van der Loop H., Francke J., Jorritsma P. en Moorman S. (2017), *Bruikbaarheid van floating car data voor beleidsonderzoek*, Den Haag, Kennisinstituut voor Mobiliteitsbeleid (KiM), 46 p.

Vayatis N. (2017), "La décision par algorithme", *in* André De Palma en Sophie Dantan (dir.), *Big data et politiques publiques dans les transports*, Parijs, Economica, pp. 51-70.

Verantwoordelijke uitgever: Camille Thiry – Vooruitgangstraat 80 – 1035 Brussel

Redactie: Thomas Ermans, Céline Brandeleer en Michel Hubert

Plannen van het BHG: Brussels UrbIS® © CIBG

Foto's: GOB - p. 4: STIB-MIVB

Lay-out en productie: Altavia ACT* - www.altavia-act.com

BRUSSEL MOBILITEIT
GEWESTELIJKE OVERHEIDSDIENST BRUSSEL