Big Data and Its Uses for Development of Public Transport Mobility Policies

Paul Verhoeven
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
p.l.verhoeven@student.utwente.nl

ABSTRACT
The OV-Chipcard, the Dutch national public transport smartcard, can provide public transport operators and authorities extensive information regarding the behaviour of smartcard owners. However, this information is limited to the data generated between check-in and check-out and contains little to no information regarding public transport users’ travel behaviour outside this ‘checked trip’. In addition, the need for alternative data has arisen in order to expand the possibilities of analysis and increase independency from operators responsible for the OV-Chipcard data.

This research identifies alternative data sources, that can be used in addition to OV-Chipcard data to improve the development of public transport mobility policies. Alternative data sources are proposed using the results of a literature review and interviews with public transport operators and authorities. The interviews are subsequently used to gain more insight in operators’ and authorities’ visions regarding alternative data sources and to find out what questions they consider answerable using big data. Lastly, the contribution of identified big data sources to development of mobility policies sources is analysed.

The proposed data sources, insights from public transport experts and contribution of identified sources can help public transport operators and authorities develop more efficient and effective public transport mobility policies using alternative data sources in addition to existing data.

Keywords
Big Data; Travel Behaviour; Public Transport; Mobility Analysis; Smartcard; Mobility Policies; Public Policy Making; Origin Destination Matrices; OV-Chipcard

1. INTRODUCTION
Between 2005 and 2012 a smartcard, the OV-Chipcard, was introduced in the Dutch public transport system. Its predecessor, which consisted mostly of paper tickets, could barely provide operators and (local) authorities with information regarding the travel behaviour of public transport users. One of the benefits of the OV-Chipcard was that it would solve the mentioned problem by digitalising the existing system and storing the produced public transport data.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

27th Twente Student Conference on IT, July 7th, 2017, Enschede, The Netherlands.
Copyright 2017, University of Twente, Faculty of Electrical Engineering, Mathematics and Computer Science.

by checking in and out in a centralized database. [15]

In addition, Bagchi, M., and White, P.R. came up with several benefits that smartcard data would have to public transport operators [4]:

- A larger availability of personal travel data for analyses
- Ability to link the travel data to an existing card or traveller
- Access to continuous travel data covering longer periods of time
- Ability to identify most frequent public transport users

A traveller using the OV-Chipcard is expected to check in at the start of the trip and check out at the end destination. When changing operator, a traveller is also expected to transfer by doing the same process. These transactions generate data consisting of card information, timestamps and trip information. In Table 1, a simplified overview of data generated with OV-Chipcard transactions is presented.

<table>
<thead>
<tr>
<th>Card number</th>
<th>Check-In time</th>
<th>Check-In location code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>Check-Out time</td>
<td>Check-Out location code</td>
</tr>
<tr>
<td>Product</td>
<td>Control time</td>
<td>Information about used operator</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Information about used line or vehicle</td>
</tr>
</tbody>
</table>

Smartcard data provides operators and authorities with extensive and accurate information, usable for the development of mobility policies. Thus, much research has been conducted in the field of analysing smartcard data.

1.1 Problem Statement
Even though smartcard data can provide operators and authorities with extensive information, it is opposed by some limitations.

Operators are responsible for the data generated by travellers using their part of the public transport sector. [10] Assuming the data is already anonymised and aggregated, operators can, without having to ask for permission, only perform analyses on data generated in their concessions. Furthermore, authorities involved in the tender contracts of those operators are limited to data about which it has made agreements with operators to perform analyses on.

Apart from that, for the usage of OV-Chipcard data to perform analyses, it can only be retrieved with clear
permission from the responsible parties, the operators. [12][29] Only given strictly specified purposes and within the boundaries of said purposes data can be released for specific projects. Third parties cannot use the data independent from the operators. In addition, to operators who are eager to perform analyses on data from other operators, the same restrictions apply.

Apart from the limitations that arise when trying to obtain OV-Chipcard data from the responsible operators or Trans Link, an organisation responsible for managing all the OV-Chipcard data from the Dutch operators, the need for more data has arisen in order to expand the possibilities of analyses. OV-Chipcard data is limited to the ‘checked part’ of the trip. It offers no insight into travellers’ mobility before checking in and after checking out. Research is unable to identify the actual origin and destination of trips using only OV-Chipcard data.

To expand the possibilities of analyses and to increase independency from operators, the need for alternative data to use apart from or in addition to existing data has originated.

1.2 Research Objectives
This paper provides insights into the usage of big data for the purpose of developing mobility policies. Substantiated with a literature review and expert interviews, an analysis of available data sources in combination with the visions of mobility policy makers and those involved leads to an elaboration of how big data can contribute to public transport mobility policies.

1.3 Research Questions
The Problem Statement can be translated into the following research questions:

- **RQ 1** What data sources are under what conditions available for mobility analyses?
- **RQ 2** What questions, according to mobility policy makers, are expected to be answerable using big data?
- **RQ 3** How can the identified big data sources contribute to the development of mobility policies?

1.4 Paper Structure
In Section 2, the research methodology is described. Sections 3 and 4 provide an overview of the available data sources using a literature review, trip model and expert interviews and outline the conditions under which these sources can be used as stated in **RQ 1**. Section 5 provides insights from expert interviews as of **RQ 2**. In section 6, the contribution of big data for mobility policies is explained (**RQ 3**). Section seven provides a discussion of the research and in section eight the paper is concluded. The acknowledgements are present in section nine, followed by the list of references used in this research in section ten. After the references, the Appendices can be found containing more information regarding the literature review and expert interviews.

2. METHODOLOGY
First, a literature study is conducted to gain more insight in the field a public transport and mobility analyses. This study is used to identify data sources that can be used for mobility analyses. For the literature study we use the paper of Webster, J. and Watson, T. [21] To find relevant scientific literature Scopus and Mendeley are primarily used. In addition, Google Scholar is used to find non-scientific literature, such as reports. Further information regarding the literature study can be found in Appendix A.

To gain insights into what experts in the field of public transport think about the alternative data sources and to find out what information they deem answerable using big data, multiple experts in the field of public transport are interviewed. Each interview takes up around thirty minutes and we used semi-structured interviews. Due to this structure, the main ‘framework’ of the interviews was prepared in advance, containing fixed questions. However, during the interviews the structure allows for flexibility to have a discussion or come up with follow-up questions. Further information regarding the interviews’ guidelines and setups can be found in Appendix B.

3. AVAILABLE DATA SOURCES
To identify the available data sources, the process is divided in two phases. In the first phase, a literature study is used to identify useful and researched data sources. In the second phase, a Trip Model is created to be able to gain a clear ‘view’ of the entire process in order to identify new data sources and prevent data sources from being kept out.

3.1 Literature Review
The term Big Data is used in daily life more and more. Never was the availability of data growing as fast as now, and never was so much data generated as before. With this phenomenon, the idea of using Big Data for the purposes of mobility analyses has arisen. Thus, much research in the field of using alternative data sources to develop mobility analyses has been done.

3.1.1 GPS data
Much research has been done using GPS data to mainly understand mobility. GPS data is generated in many ways. Examples of devices connecting to GPS and thus generating GPS data are navigation systems in cars and smart devices such as smartphones and tablets.

GPS data is, regardless of how it can be used for analyses, available in extreme quantities as mentioned by Calabrese, F. et al. [6] Mostly during the last twenty years, an explosion in the use of pervasive systems has occurred, allowing for immense amounts of data. [7][25] The data from pervasive systems, such as GPS devices and mobile phones, exists of digital footprints, telling where people are and at what time. [6]

From the available data sources, GPS data has some important advantages and disadvantages to cover. The GPS infrastructure can be used worldwide, enabling users to always use the infrastructure and thus leave accurate traces. [2] Furthermore, having 24 working satellites, the accuracy of the system is sufficient, leaving useful location history traces. [17] These can support the visualisation of people’s mobility, so that origin-destination matrices can become more accurate and detailed.

However, the technology is also facing some disadvantages. As GPS positions from a particular device are saved as individual mobility datasets, privacy issues occur when collecting data at larger scales. The datasets have the ability to accurately unveil where people live, work, shop, etcetera., which leads to law infringements. [25]
3.1.2 Cellular connection data
Each time devices make connections with beacons from cellular networks, available base transceiver stations can measure their location by using triangulation, as explained by Srinivasan, A. [17] These connections consist of three categories:

- Calling
  When a device is used to either place or receive a call at the beginning and the end of a call.
- Texting
  When a device is used to send or receive a text.
- Internet
  When a device is used to connect with the internet (could be for example to browse the web, use an app with internet connection or send an e-mail).

The usage of data generated by said network connections faces some limitations. The data has a lower resolution than GPS data. From tests, the uncertainty radius is measured to be 320 meters, with a median of 220 meters. [6] However, other research has measured the error at 500 meters in urban location to about 15 kilometres in rural areas. [17] Furthermore, incorrect locations might be generated at peak hours, when one might be transferred to a cellular tower further away due to the congestion of a tower closer by. [6]

3.1.3 Wi-Fi data
During the last years, in many public areas a Wi-Fi infrastructure is installed. Using the data generated by Wi-Fi networks is a hot topic in literature as nowadays many people have Wi-Fi compatible devices that make use of public networks. Data generated by devices connecting with Wi-Fi is often discussed as a useful data source for tracking crowds and estimating crowd densities. [16][3] In order to discover the devices using Wi-Fi, two mechanisms are present as explained by Schauer, L., et al. [16]:

1. When devices are passively scanning for Wi-Fi points, they can be picked up by an access point. In this scenario, the devices are ‘listening’ to messages from access points, resulting the discoverability of the device.
2. Devices can also be actively scanning, when it is for example being used to send messages via the access point.

For this data to be useful to the development of public transport mobilities, some factors should be considered. As there is no standard for scanning for wireless networks, differences amongst operating systems and mobile phones are present. This can cause the results to become inaccurate, as the devices all have different durations per probe request. [3] This difference must be considered when using Wi-Fi data for analyses.

3.1.4 Bluetooth data
In addition to Wi-Fi data, Bluetooth data is the topic in multiple researches. Bluetooth data can be most ideally used to estimate crowd densities [23], however, when having receivers positioned in multiple places, the movements of crowds could be detected as well.

For Bluetooth data to be generated, public transport travellers need to carry mobile devices with them. Two main methods of sensing Bluetooth signals are present: using fixed mobile scanning devices to track travellers or using mobile scanners to do the same. The latter is mostly applicable for the usage on a smaller scale as specialised software has to be installed on the smartphones of test users. [18] For the purpose of mobility analyses, fixed scanners are more suitable, as they function better and no ‘manual’ process is required.

3.1.5 Radio Frequency data
A more traditional form of data is that of Radio Frequency (RF) based technologies. This data can be used to determine the locations of people or crowd densities. [23] There exist two methods to obtain RF data:

1. Mobile method. In this method, certain people need to carry devices such as sensor nodes and RFID tags in order to locate people or estimate crowd densities.
2. Fixed method. For this method, deployed wireless sensors can be used instead. However, in order to use these, they have to be pre-installed.

Both methods allow for data with location-based information. However, it must be noted that radio frequency signals are easily affected by dynamic environments, resulting in potentially incomplete or incorrect data. [24]

3.2 Public Transport Trip Model
Present research is primarily focussed on analysing current technologies regarding public transport mobility. To identify more data sources with potential uses to the development of public transport mobility policies, we designed a model translating the basic processes involved when public transport is used (see Figure 1). The goal of this process is to visualize the relationships between certain parts and incentives involved in a stereotype public transport trip, to support the identification of data sources involved in the process.

The model is made by going step-by-step through the process of using public transport. Its primary goal is to support the identification of new data sources by visualising the process and allowing for a structured view revealing valuable information sources.

The core of the model is formed by the public transport users. We assume that a person is willing to undertake a trip using public transport. This user has personal information, which can contain information regarding the intended trip. Personal information can come in the form of Personal Information Managers such as calendars or e-mail clients, in which for example pre-booked flights or tickets are stored. Furthermore, it can consist of data from social media, on which persons have reported certain mobility.

![Figure 1. Public Transport Trip Model](image)

Furthermore, many users own an OV-Chipcard which is used during the intended trip and a substantial number of users plan their trips in advance using trip planners. However, this differs per user, as daily users travelling for work, have less reason to make use of such travel planners.
The users have a goal when travelling, which can be substantiated with a motive. These motives are divided in three categories: work, education and leisure. Furthermore, the trips (goals) have a certain frequency. Whereas commuters might use the public transport system every day, tourists might only take the train once a year. This information can help identify data sources containing information. Companies have knowledge of what employees use public transport and what route they travel, big events know that a number of visitors will use public transport as a means of transport and educational organisations can estimate when their students and/or employees increase the demand on public transport.

When the intended trip is undertaken, the trip itself contains certain information generated by the OV-Chipcard. Furthermore, during the trip, the mobility is registered by multiple technologies. People have left traces with GPS and network locations and can be tracked in the public transport by means of cameras, Wi-Fi or Bluetooth.

3.2.1 Trip Planner Data

From the model, a potentially interesting data source to analyse is that of data generated by trip planners. In the Netherlands, the largest and most used travel planner, OV9292, produces 1.9 million travel advices per day. [1] Given the fact that more than one million users make use of the Dutch public transport every day, trip planners are a frequently used application for many of the travellers. [5]

When requesting trip advice via a trip planner, the requests are consequently saved in a database. As many trip planners are available to users, a distinction should be made amongst the planners. Naturally, national and regional trip planners generate different data for the purpose of analyses, as local bus advices have a different value than national train advices.

3.2.1.1 9292

Extremely valuable is the data generated by 9292. The application is one of the largest and most used public transport travel planner of the Netherlands. Furthermore, Newcom Research & Consultancy has researched the planner’s brand awareness and came to a percentage of 58%. [11] 9292 can be used to plan trips using public transport. When requesting trip advice using 9292, the request is stored in a database from 9292. These requests contain detailed mobility information as the start- and endpoints of trips are inserted in order to have a trip planned.

An advantage of 9292 data is that it contains clear origin-destination information. This data furthermore is not as privacy sensitive as OV-Chipcard transactions. [29] A downside of using 9292 data to perform mobility analyses is that it remains unsure how many from the 9292 requests have actually resulted in public transport trips. However, this can be deducted when more data sources such as OV-chipcard data and location-based data are combined with the 9292 dataset. Trans Link has already worked on this combination to find out if 9292 and OV-Chipcard data can be combined to increase the total value of the data. [29]

3.2.1.2 Other Operator-Owned Travel Planners

Apart from 9292, many other dedicated public transport trip planners have been made available to the public. Most of these planners are owned and provided by operators active in the Dutch public transport sector. In Table 2, an overview is given from available trip planners that are owned by operators. Note that multiple operators are owned by another operator and thus logically make use of the same trip planner framework (not necessarily the same planner).

From Table 2 it becomes clear that although many operators have implemented their own trip planner, almost all the planners use the 9292 API in the background. Only Bravo, facilitating the public transport in the Dutch province of Noord-Brabant, has decided to use Go About rather than 9292.

As most trip planners are still using 9292 to generate trip advice, the advices given by these trip planners are still stored at 9292 as well. Some operator owned planners do in fact offer additional functionalities to its users, but data-wise no difference is present regarding the trip advices made.

Table 2. List of Trip Planners Provided by Operators

<table>
<thead>
<tr>
<th>Operator</th>
<th>Has trip planner?</th>
<th>Travel advice made possible by</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arriva</td>
<td>Yes</td>
<td>9292</td>
</tr>
<tr>
<td>Bravo</td>
<td>Yes</td>
<td>Go About</td>
</tr>
<tr>
<td>Breng</td>
<td>Yes</td>
<td>9292</td>
</tr>
<tr>
<td>Connexion</td>
<td>Yes</td>
<td>9292</td>
</tr>
<tr>
<td>EBS</td>
<td>Yes</td>
<td>9292</td>
</tr>
<tr>
<td>GVB</td>
<td>Yes</td>
<td>9292</td>
</tr>
<tr>
<td>Hermes</td>
<td>Yes</td>
<td>9292</td>
</tr>
<tr>
<td>HTM</td>
<td>Yes</td>
<td>9292</td>
</tr>
<tr>
<td>OV Regio IJsselmond</td>
<td>Yes</td>
<td>9292</td>
</tr>
<tr>
<td>RET</td>
<td>Yes</td>
<td>9292</td>
</tr>
<tr>
<td>Syntus</td>
<td>Yes</td>
<td>9292</td>
</tr>
<tr>
<td>NS</td>
<td>Yes</td>
<td>9292</td>
</tr>
<tr>
<td>Qbuzz</td>
<td>Yes</td>
<td>9292</td>
</tr>
</tbody>
</table>

3.2.1.3 Route Planners

In addition to trip planners, route planners can nowadays also be used when requesting trip advice. One of the largest planners, Google Maps, offers its users public transport information in addition to trip advice since 2013. Once more, for this functionality Google has collaborated with 9292 to offer accurate, national trip advice. [26] Thus, every request for public transport trip advice is still processed via 9292, thus stored in the same dataset of trip advices from trip planners or 9292 itself.

However, when used, route planners such as Google Maps do generate GPS data as this is required for the planner to work. This GPS data can then be used for mobility analyses as described in section 3.1.1. Another form of route planners making use of GPS data is that of car navigation applications. These applications contain clear GPS trajectories that can be used to gain insight into its owner’s mobility. [8] The downside of using car navigation data is that it only provides mobility information generated when driving a car. However, this information could be especially useful in rural areas, to analyse where public transport can be used in regard with present demand. [27]

3.2.2 Personal Information Managers (PIM)

In the Trip Model, a public transport user is given. This user also brings forth a potential interesting data source: personal information managers. A personal information manager (PIM) can be best described as an information management tool, mostly responsible for managing and keeping track of its owner’s personal information. Some of these PIMs hold...
information regarding the mobility of its owner, useful for analyses.

A calendar consists of highly personalised information. The appointments can be perfectly used to analyse and predict mobility, as these are often associated with a corresponding location. [20] However, there are some limitations towards using calendar data, as described by Tran, L.H. et al. (2012): only a small number of users had a sufficient amount of calendar data and secondly, as the calendar data had to be anonymised, most information appeared as one-time events with few uses to mobility analyses. Only recurring appointments could help predict a mobility.

In theory, e-mail data contains similar information. Only in this case, it is not logically structured in a calendar, but rather scattered. No research is present regarding the analysis of e-mail data to analyse or predict mobility. Naturally, calendar data possesses the most important data from e-mails as users would copy confirmations for events, flights, bookings, etc. to their calendar (or this is done by the e-mail client itself).

3.2.3 Social Media
As stated by Terroso-Saenz, F., et al (2016): “the combination of mobile and social media sensors is foreseen to become a crucial course of action so as to comprehensively capture and understand the movement of people in large spatial regions.” Major social-media companies, such as Facebook and Twitter have implemented location-based capabilities in their applications. Consequently, most of the information generated by the companies can be geo-tagged. These so called soft sensors can combine location information with social-media. [19]

Combining the location information data with corresponding social media data can provide valuable information. The location information shows where a person is at certain times (spatio-temporal), when combined with textual social media data insights into why a person travels from A to B can be gained. [22][19][9]

Even though analyses over this data can be relatively easy performed, the information retrieved has some limitations. There is a great variety of people using social media. Whereas one person might only post about very special occasions, such as going on holidays, another person posts about every single mobility, online checking in on each place. Naturally, the latter person can provide more useful information for mobility analyses. However, these results are sensitive to be not representative for an average traveller. In this case, it should be noted that clear mobility patterns are mostly representative for the younger travellers using social media more intensive.

3.2.4 Administrations
Many public transport trips can be predicted using large administrations. For example, when big events such as festivals or concerts take place, an increase of public transport usage can be expected. Furthermore, large companies such as Randstad and Shell have immense administrations containing addresses of its employees. The companies know what employees use public transport to arrive at work and know what journey they will undertake. This information, when anonymously available to operators, authorities or third parties, could be valuable as demand can be predicted better.

An operator active in the Dutch public transport sector has confirmed that the company greatly values a business-to-business model, a model which allows for tailormade public transport such as the transportation of specific groups of people (elderly, disabled persons). During the last four years, the company has developed a department working on this model. With this model, the company aims to perfectly react on present demand, in order to maximise efficiency. [28]

4. CONDITIONS FOR USAGE OF DATA SOURCES
There are many conditions to the usage of the identified data sources. Especially as this research aims to identify alternative data sources, that can be used in combination with OV-Chipcard data, certain conditions arise when the combination between data sets is made. Therefore, conditions under which OV-Chipcard data can be used, should be considered as well.

4.1 OV-Chipcard Data
OV-Chipcard data is generated when people use the public transport network by checking in and checking out. This information is stored in multiple places. Each operator saves the data from its own concession. However, all the operators have agreed upon sending the data to Trans Link as well. Trans Link is the only organisation managing all the OV-Chipcard data generated in The Netherlands, therefore an interview has been conducted with R. van den Berg, Information Service Director at Trans Link Systems B.V.

Table 2 gives an overview, containing what parties can access what data.

If a third party, such as a local authority, consultancy firm or a university desires to perform analyses on nationwide OV-Chipcard data, Trans Link must be contacted in order to obtain concession-exceeding data. [29] In the case data from only one concession is desired, the request can be send to the involved operator instead.

<table>
<thead>
<tr>
<th>Who?</th>
<th>Access to what data?</th>
<th>How</th>
</tr>
</thead>
<tbody>
<tr>
<td>Travellers</td>
<td>Data generated by own OV-Chipcard</td>
<td>All individual transactions</td>
</tr>
<tr>
<td>Operators</td>
<td>Data generated by all trips made with operator</td>
<td>All individual transactions</td>
</tr>
<tr>
<td>Authorities</td>
<td>Data from within respective concession</td>
<td>Aggregated data</td>
</tr>
<tr>
<td>(concessionaire)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trans Link</td>
<td>Data generated by all OV-Chipcards</td>
<td>All individual (approved) transactions</td>
</tr>
</tbody>
</table>

When Trans Link receives a data related request, the company must identify the involved operators who own the requested data. When identified, Trans Link requests permission from the operator using a generic editorial agreement. For each request, Trans Link must fill in the consent form containing the goal of the research, the involved datasets, the period of use, all involved parties and some other formalities. Only when this permission is granted, Trans Link can be ‘processor’ of the data, granting third parties access to the anonymised and aggregated dataset. [28]

When an operator receives a request from Trans Link, the company must decide if it grants permission or not. Operators are responsible for ensuring privacy regarding the smartcard data. For each request, operators must determine whether the goal of the request is in line with the initial goal of why the company has received and managed the data. Important factors influencing the decision are the traceability of individuals and business confidentiality. [29] It is very
important that with the dataset individuals cannot be traced, or in other words: the dataset should be properly anonymised and aggregated before it is released for analyses. [27]

4.2 Alternative Data Sources
To obtain other data sources, with the idea of combining this with OV-Chipcard data, in general the same methodology applies. With respect to the law, preserving privacy of the user is always the number one priority. When a party has the intention to obtain data in order to process it for the purpose of analyses, ten basic steps should be taken into account (see Table 3). When these steps are considered, and none are violated, in theory the data can be processed.

It is important to note that data, containing personal details, can be ‘edited’ for statistical or scientifical purposes, as mentioned in the Dutch WBP (law for protection of personal information), Article 9, Paragraph 3. The analyses performed by operators or authorities are mostly commercial of nature, leading to more obstacles regarding privacy. Especially in this case, the steps from Table 3 should be carefully considered.

Even though all requirements might be met, the combination of anonymised and aggregated data with other data under the same conditions may cause traceability of individuals. It is very important to prevent traceability, both direct and indirect, from being present in the dataset. For each combination that is made with other datasets, checks should be made to preserve the privacy of individuals.

Table 4. 10 Step-Analysis for Intended Data Processing

<table>
<thead>
<tr>
<th>Step</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Is (special) personal information involved?</td>
</tr>
<tr>
<td>2.</td>
<td>Purpose well defined, expressly defined and justified?</td>
</tr>
<tr>
<td>3.</td>
<td>Is it necessary? Option for data minimalization?</td>
</tr>
<tr>
<td>4.</td>
<td>On what legal basis?</td>
</tr>
<tr>
<td>5.</td>
<td>Who is responsible? Who is/are the editor(s) of the data?</td>
</tr>
<tr>
<td>6.</td>
<td>Written editorial agreement(s)?</td>
</tr>
<tr>
<td>7.</td>
<td>Mandatory notification or exemption? Prior research CBP (authority for personal data) required?</td>
</tr>
<tr>
<td>8.</td>
<td>Obligation to inform involved or exemption?</td>
</tr>
<tr>
<td>9.</td>
<td>Legal obligations concretized and respected?</td>
</tr>
<tr>
<td>10.</td>
<td>Data-export?</td>
</tr>
</tbody>
</table>

From the 25th of May 2018 onwards, the General Data Protection Regulation (GDPR) is applied to the EU, including the Netherlands. [31] As this regulation is EU-wide, from that point onwards the processing of data containing personal information is regulated similarly in all EU states. For processors of data, this can simplify the process, as one general methodology can be used within the European Union. Lastly, the economic value of the data must be considered. The owners and managers of the identified sources have made costs for collecting and managing their data. When this data is used by other parties for the purpose of analysis, the actual owners and managers can request a financial compensation for its usage. Furthermore, they can demand that the ‘added value’ (results of analysis) is shared during and after the process. These economical values must be taken into account when using alternative data for analyses.

4.3 Trans Link Information Management System
Given the extensive procedures that Trans Link is facing when requesting permission to process data and given the fact that, under the law, many measures have to be taken in order to preserve travelers’ privacy, Trans Link is working on an information management system to simplify the process for the future. [13]

In this information management system, all safeguards are considered to enable further research with smartcard data. In this system, anonymized data is saved in response to the increased demand for data by the public transport sector. In this system, no data will be traceable, and data is purely available for analysis purposes. [29]

Furthermore, when the system is finalized, more datasets can be combined relatively easy and within the law, preventing potential trackabilities from occurring given the new data manager. [29] The system can support the mapping of the mobility market, useful for development of policies.

5. ANSWERABLE QUESTIONS USING BIG DATA
To find out what questions public transport operators and authorities deem answerable using big data, interviews with experts in the public transport sector have been conducted to gain more insight in their views regarding big data usage for the development of public transport mobilities. During the interviews, the usage of identified data sources was discussed in order to find out what the interviewees’ opinion on these sources was.

5.1 Interview DOVA
DOVA is the cooperation of the fourteen decentralised public transport authorities in the Netherlands. The organisation supports the authorities in their responsibilities by fulfilling multiple tasks. The goal of the organisation is to realise effective and efficient public transport for the travellers.

From an interview with DOVA it became clear that operators are very eager to gain insights in the travel behaviour of travellers. Not only are they looking for data regarding the behaviour of people in their own concessions, but operators are also interested in the behaviour from people travelling in other concessions.

According to DOVA, (local) authorities are also looking for data regarding the travel behaviour of people in certain concession. With this information, authorities aim to gain more insights to agree upon proper tender contracts. Authorities can base these tender contracts on insights on the travel behaviour of people.

An important part of setting up proper tendering contracts is the insight authorities have in the users of concessions. DOVA notices that local authorities are looking for information in order to find out who a certain concession’s users are. With these insights, local authorities can offer fitting public transport. In line with offering fitting public transport services, authorities can also use big data to find out how certain lines are ‘doing’.

5.2 Interview Operator
The interviewed operator has expressed the wish to remain anonymous and is thus referred to as: ‘operator’. About the operator can be said that the company is active in the Dutch public transport sector.
The operator is very interested in performing mobility analyses using big data. These mobility analyses can be, from the operator’s perspective, divided amongst three categories:

The primary goal for the operator to perform mobility analyses is the development of its product offerings, and thus more commercial of nature. Examples of product offerings can be personalised subscriptions or public transport bicycles. The operator is eager to find out what products are used by what customers. Furthermore, insight into the times at which these products are used would be of great value as well. Apart from the products, the operator explains that insight into traffic flows can be used to explore where their supply of public transport is not fitting or where new users can be tempted to use the network as well. Finally, improving travellers’ frequency or tempting incidental users to increase the usage of public transport is important to the operator. Alternative data sources can be useful when in need of more information to answer these questions.

Secondly, the operator aims to improve its timetable. To do so, mobility analyses are performed using, where possible, big data. However, an important distinction should be made as operators face two types of commissioning authorities. In the case where authorities are responsible for the timetable, operators have fewer intentions to perform analyses themselves, as they are not able to change the timetables. However, in the case operators are expected to either collaborate with local authorities or individually optimise the public transport offer, mobility analyses are performed more extensively.

Lastly, operators use big data for mobility analyses with respect to the company’s continuity in the future. The company is interested in the change of transport modes. Furthermore, it is interested in insights regarding car ownerships and usage of public transport in rural areas. These analyses are important to the company as they are invaluable to its long-term strategy. To some of these questions, identified data sources can help provide answers.

6. CONTRIBUTION TO MOBILITY POLICIES

This section addresses how the identified big data sources can contribute to the development of mobility policies. The OV-Chipcard can provide detailed information about the travel behaviour of travellers. However, the range of this data is fairly limited. It can only register the travel behaviour inside the public transport network. However, no information from outside the network is available, for example how people get to the bus stop or where people go after leaving the train station. [12] The identified big data sources can add information to the present dataset, increasing the range of the data.

The identified data sources can increase the analysis potentials of third parties, operators and authorities. [12] As mentioned before, it can increase the range of existent data but apart from that it also provides more and different data sources. Public transport mobility policies can be based on more and different data sources than only the OV-Chipcard data. In addition to the analysis potentials this also leads to an increased independency. Third parties are no longer solely dependent on operators to provide them with data, rather they can obtain their own independent datasets to potentially use alongside the OV-Chipcard data.

Furthermore, having extra big data sources contribute to the development improves the value of the public transport sector for the travellers. The public transport offer can be improved as big data supports the implementation of a more tailor made public transport network. For example, in rural areas, big data can help identify the present demand and usage of public transport. Consequently, ‘call-busses’, public transport bikes and smaller neighbourhood busses can be deployed more effectively resulting in an improved experience, better offer, lower costs and overall more effective public transport. [27]

The big data sources contain a lot of information and can improve the insights into current public transport. Furthermore, it can stimulate the propositions and mobility innovations in the sector resulting in increased value to the customers, the travellers. [29] As higher political expectations, increased competition between operators and reduced budgets force operators and authorities to improve their efficiency, these innovations can support them to reduce inefficiencies and bottlenecks the companies are facing. [14]

7. DISCUSSION

In this research, we have to the best of our ability, tried to identify all the available data sources. By using a thorough literature review, modelling the processes and interviewing multiple experts in the field of public transport, we tried to minimize the option of having data sources being left out of the research. Future studies can continue the research by expanding the insights into the technical possibilities and potentially expand the number of identified data sources.

Furthermore, parts of this research have had its focus on identifying available data sources and analysing under what conditions these can be used. Future research can expand this study by analysing under what conditions each individual data source can be used and how the identified data sources can be technically implemented. With this knowledge, the research becomes even more detailed and valuable towards operators, authorities and third parties.

For this research, we interviewed as many experts from the field of public transport as possible. To gain the best insights in the sector, we tried to interview experts from all different 'positions', resulting in interviews with an operator, an organisation that represents the fourteen decentralised authorities and a company managing the public transport data. Due to the scope of this research, no other parties were interviewed. In future research, more experts can be interviewed to obtain even more information from the sector, for the improvement of the research.

8. CONCLUSION

OV-Chipcard data can provide operators, authorities and third parties with valuable mobility information. However, to use the data certain limitations arise. In addition, the demand for alternative data to expand the possibilities of analyses and increase the independency from operators has grown during the last years.

This research identified data sources that can be used for mobility analyses and analysed under what conditions those can be used. Furthermore, the research identified questions that mobility policy makers deemed answerable and analysed how big data sources can contribute to the development of mobility policies.

By using a literature study, Trip Model and expert interviews data generated by GPS, cellular connections, Wi-Fi, Bluetooth, Radio Frequency, trip and route planners, personal
information managers, social media and administrations were considered available to analyses. The questions of policy makers regarding travel behaviour, analyses of users, development of product offerings, timetable improvements and future company continuities can be answered with alternative data sources when the privacy conditions are met. The sources contribute to mobility policies by improved analyses possibilities, increased independencies, improved public transport and reduced inefficiencies.

The results from this research contribute to the development of public transport mobility policies and allow for greater analyses possibilities and increased independency.

9. ACKNOWLEDGEMENTS

This paper would not have been possible without the guidance of prof.dr.ir. Bart Nieuwenhuis, professor at the University of Twente and Rob Rijnhout, managing consultant for Hypercube Business Innovation. Additional thanks go to A. Kruijt and R. Mijnans from DOVA, R. van den berg from Trans Link, mr.dr. L. Broos from the University of Twente and a representative from an operator active in the Dutch public transport sector. Lastly, I would like to thank S. Braams for peer-reviewing my proposal and paper, providing valuable suggestions and remarks.

10. REFERENCES

[27] Verhoeven, P. 2017. Interview DOVA.


APPENDIX

A. LITERATURE REVIEW

For the identification of present literature, the paper from Webster, J. and Watson, T. was consulted. [21] The paper suggested a structured approach for the determination of source materials. This approach consists of three steps, which are considered for the literature study.

Before identifying present literature, keywords were defined to be used as search ‘terms’. The defined keywords are listed below:

- Big Data, Travel Behaviour, Public Transport, Mobility Analysis, Smartcard, Mobility Policies, Public Policy Making, Origin Destination Matrices, Tracking

Using these keywords, the first step of the approach was executed. In this step, we used the systems of Scopus, Mendeley and Google Scholar to search for literature from leading journals. For this search, the keywords from above were used both individually and combined. In the next step, the process went backward, as described by Webster and Watson. The reference lists from the previously found literature was used to identify other papers. Lastly, the process went forward. In this step, we looked at the articles that had cited the identified articles from the first two steps.

During this process, certain keywords were added, as more insights into the field of study was gained. Examples of added keywords are:

- GPS, Cellular Networks, Social Media, Bluetooth, Wi-Fi, Personal Information Managers

The selection of the articles was done by scanning the abstracts, introductions and conclusions. In some cases, the introduction was not scanned as the abstract and conclusion provided sufficient information.

B. INTERVIEWS

For this research, interviews were conducted with multiple experts in the field of public transport. To obtain as much information as possible from all the ‘corners’ of the sector, the following experts were interviewed:

- A. Kruijt and R. Mijnans. Policy Employees at DOVA
- R. van den Berg. Information Service Director at Trans Link Systems B.V.
- Representative from an operator active in the Dutch public transport sector

The interviews had a duration of around 30 minutes and were conducted face-to-face or by telephone. Before the start of the interviews, permission was asked to record the conversation to allow for the writing of a detailed transcript. Furthermore, the interviewee was asked whether he/she wished to remain anonymous.

The interviews started with an informal introduction and a brief summary of the research which has led to the interview. For the remainder of the interviews, a semi-structured approach was used. The decision to use semi-structured interviews was made to ensure that certain questions were discussed but to allow room for discussion and potential other topics as well.

The goal of the interviews was to gain insights into what questions mobility police makers deem answerable using big data. However, since the experts all represented different areas within the public transport sector, for each interview a different guideline was prepared. However, the following questions were, depending on the interviewee, at least partially discussed (in addition to interviewee-specific questions):

- What data does [interviewee] possess?
  - Where is this data generated?
- Under what conditions can the data be used for analytical purposes?
- Where does [interviewee] see possibilities for the usage of alternative data sources?
- What data sources does [interviewee] deem usable in addition to OV-Chip card data?
- Is [interviewee] already actively using alternative data sources?
- For what purposes does [interviewee] analyse public transport data?
- To what extent anonymises/aggregates [interviewee] the data before it is analysed?
- What information hopes [interviewee] to obtain when using alternative data sources for mobility analyses?
- The research has identified the following data sources (…), what sources considers [interviewee] usable for analysis?
- What sorts of data are most useful for the development of mobility policies?

In addition to these questions, follow-up questions were asked to obtain more in-depth answers. When the interviews were finished, summarizing transcripts were made and sent to the interviewees. Only after the transcripts were accepted by the interviewees, they were used in the research.