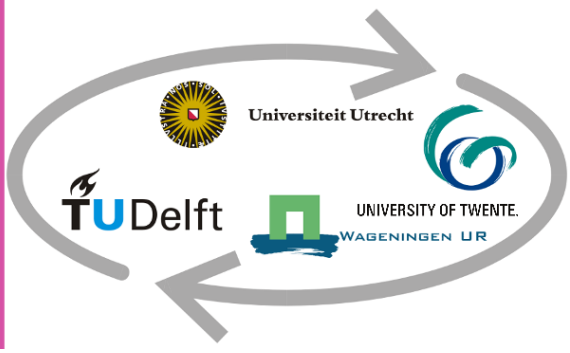
GIMA
Geographical Information Management and Applications

Spatial data quality of public transport data in OpenStreetMap

Author: Martijn van der Putten

Supervisor: John Stuiver

Professor: Arnold Bregt



Preface

This thesis was written as part of the study program Geographical Information Management and Applications. I would like to thank my supervisors at Wageningen University: Arnold Bregt and John Stuiver for their support and constructive talks and Arend Ligtenberg for the technical assistance while creating the script. Also, I want to thank my friends for listening to my struggles, which were probably not always very interesting to listen to and for all the good advices you gave me during the process.

Martijn van der Putten
June 7th, 2017

Abstract

Spatial data quality of Volunteered Geographic Information (VGI) has been subject of research in the recent past. This study investigates the quality of bus route data in OpenStreetMap (OSM), which has not been done before, compared to the quality of bus routes delivered by transport operators.

The spatial data quality of three datasets is assessed. First, a dataset with real-time bus locations was collected. A script was written to make it possible to collect and store vehicle positions for a period of 24 hours. The raw dataset had some issues. For example, train routes had been included. Also, the dataset for transport operator Arriva had some points included which formed strange patterns on the map. Other points were further from the road network than could be expected from normal GPS behaviour. A part of these points represented buses in bus depots or at a parking lot near a station.

After removing trains and a part of the handheld devices, the dataset with real-time bus locations was analysed and prepared for the next step. The distance to the planned bus routes as delivered by the transport operators was calculated. Around 90 percent of the points was within 10 meters from the planned bus routes. Different message types (OffRoute and OnRoute messages) were analysed to try to explain a part of the points which were not within 10 meters from the bus route dataset.

Finally, for every real-time bus location, the distance to the OpenStreetMap (OSM) bus route network was calculated. The distance to the OpenStreetMap bus route network was comparable to the distance to the planned bus routes delivered by the transport operators. The Interface 1 dataset was slightly better than the OpenStreetMap dataset, which is explainable by the way the datasets are created. Interface 1 data is delivered by transport operators, while OpenStreetMap bus routes are created by volunteers.

Keywords: Spatial data quality, Volunteered Geographic Information, VGI, OpenStreetMap, OSM, Public transport, Open data, Spatial accuracy

Table of Content

1. Introduction

Volunteered Geographic Information (VGI) becomes more and more important. People without a formal training now create geographic data (Goodchild 2007). The best-known example of a VGI might be OpenStreetMap (OSM), an initiative to create an open map of the world. OSM has been the subject of studies on data quality issues. For example, OSM data was compared to official topographical maps and it was concluded that overall OSM data is accurate, but on a local level, there are large differences in spatial data quality (Zielstra & Zipf 2010).

Apart from independent initiatives to create an open map such as OSM, several governmental institutions have begun to publish open data. A Dutch example of data published as open data is the availability of public transport data. The Wet Personenvervoer 2000 (Passenger Transport Act 2000) requires transport operators to deliver real-time data to anyone who wants to use this data for a travel information system or application. To facilitate this, the "National Database Public Transport" (NDOV) was founded. Two parties act as NDOV portals: the REISinformatiegroep (a commercial company, best known for the trip planner 9292) and the non-profit organisation Stichting OpenGeo. Both portals make the source data delivered by transport operators available. Operators are required to deliver the position of their bus stops and the planned routes. This data is also included in the NDOV data and for example used by Google Maps. As a result, the route of the bus will be visible on Google Maps when someone plans a trip via the Google Maps transit planner.

Public transport data is also included in the OSM data, which makes it possible to compare NDOV and OSM route data on public transport. Spatial data quality of OSM has been assessed in multiple studies. However, the "second layer" of information, consisting of (bus) routes, has not really been subject to research. Most studies don't even mention this type of information in OSM at all. An example of a study that did take this secondary information into account is the work of Hochmair et al. (2013), investigating the quality of cycle route data in OSM. Public transport routes in OSM have not been the subject of research.

1.1 Research Objectives

The focus of this research will be to investigate the quality of bus routes in OpenStreetMap, the OpenOV NDOV desk, and the REISinformatiegroep NDOV desk.

This research will investigate the spatial data quality of public transport data, as applicable to bus routes in OpenStreetMap, the OpenOV NDOV Portal, and the REISinformatiegroep NDOV Portal.

The objectives of the research are:

- To understand the OpenStreetMap and NDOV data structures;
- To assess the quality of real-time bus locations delivered by transport operators;
- To assess the quality of public transport data delivered by transport operators via both NDOV desks (REISinformatiegroep and OpenOV) and
- To assess the quality of public transport data in OpenStreetMap.

1.2 Scope

Via the NDOV desks, public transport data is delivered for the Netherlands. In OpenStreetMap, data is available for the whole world. Included in the datasets are bus lines, train lines etc. In order to arrive at a dataset that has a manageable size for good quality scientific analysis, the scope of this research will be limited to all bus routes in the province of Gelderland. This area is limited in size and number of bus lines that have to be reviewed, but at the same time has a variety of different public transport operators. Other types of public transport (trains etc.) and other areas in the Netherlands will not be subject to this research.

1.3 Structure of the report

The research will be split into 4 parts, which follow logically from each other. First, existing cases, review of scientific articles and available ISO standards regarding data quality will be reviewed in Chapter 2. After this theoretical outline, Chapter 3 will describe the methodology. Finally, the datasets will be described in chapter 4 and the results of the quality assessment will be presented in chapter 5.

2. Theoretical Framework

The objective of the chapter is to give some theoretical background for the research. First, an introduction to the subjects of Volunteered Geographic Information (VGI), open data and data quality requirements regarding geographic data will be given. Existing studies to the spatial data quality of OpenStreetMap will also be reviewed.

2.1 Volunteered Geographic Information

The concept of Volunteered Geographic Information (VGI) describes the process where citizens without any required formal training create spatial data (Goodchild 2007). One of the most popular examples of VGI is OpenStreetMap. The OpenStreetMap was started in 2004 as an alternative for paid map services and became more and more detailed since this date.

An issue regarding VGI is the quality of the data. One example of several studies that have been conducted on the quality of OpenStreetMap is the work of Haklay (2010). He compared OpenStreetMap data to Ordnance Survey Datasets. He concluded that OpenStreetMap and VGI, in general, can have a very high level of quality. Since OpenStreetMap has many contributors, the data quality can differ from place to place nevertheless. Haklay (2010) mentioned the inconsistency of VGI in terms of quality and he states that this is the price to be paid for "having a loosely organised group of participants". Other examples can be found in the work of Helbich et al. (2012) who compared the positional accuracy of OpenStreetMap and proprietary data and concluded that the accuracy was better in highly populated areas.

2.2 Open data

Besides the growth of available data through VGI, there is a trend to publish all kinds of data as "open data" on the internet. Especially governments publish their data with a license that allows other governments, companies or private individuals to use the data in their own interests. The main reasons given to make data publicly available, are transparency and data re-use (Attard et al. 2015).

In some cases, making data available as open data is mandated by the government. This is, for example, the case for public transport data in the Netherlands.

2.2.1 Public Transport Data in the Netherlands

Because public transport operators in the Netherlands are required to publish their data, it was decided that a national database or public transport had to be established. The REISinformatiegroep showed interest in creating this national database. At the same time, the OpenGeo Foundation (Stichting OpenGeo) was also interested in developing a national database public transport. The OpenGeo Foundation already stimulated and facilitated projects to make better use of publicly available geographic databases, which did make this a logical extension. Two portals[1] were founded, that facilitate publishing public transport data. Thus, public transport operators started to deliver the required data to both offices through the newly created portals. The Ministry of Infrastructure and the Environment then decided that both portals functioned well and there was no need to choose one of the portals as the major portal via a public tender, neither did the ministry consider it necessary to create their own office. Therefore, it was decided to maintain a situation with two NDOV desks, one via 9292.nl and one via the OpenGeo Foundation (Ministerie van Infrastructuur en Milieu 2013).

[1] The portals are referred to in Dutch as Loketten

Via both NDOV desks, several databases have been established, that work through different interfaces[2]. Each interface has an own set of tables, delivered by the transport operator, with several types of information. These databases include:

- Interface 1: planned schedules, including bus routes and timetables;
- Interface 6: real-time locations of vehicles;
- Interface 15: messages which appear on bus, tram and metro stops and
- Interface 17: deviations from the operational process.

Both NDOV portals use the same standard. Public transport operators are required to deliver their datasets conform this standard.

Some principles of standardised open data include:

- Data must be complete: the dataset should include all the available data.
- Data must be primary: the datasets are delivered directly from the source and are not aggregated or modified.
- Data must be timely: in the case of changes, the data should be updated immediately, the most recent data should always be accessible.
- Data must be accessible: it should be easy to access the relevant datasets.

2.2.2 Privacy

Publishing data as open data may conflict with privacy regulations (Kulk & van Loenen 2012). Also, data which does not look like personal data, at first sight, can be considered personal data because personal information can be extracted by cross-linking data of multiple datasets.

For public transport data, privacy issues arise when data on check-ins and check-outs is made publicly available. However, transport operators also prefer to keep this kind of data to themselves. The datasets used in this research are exclusively on bus routes and not personal in any way.

2.3 Public transport networks

Before the datasets will be selected and collected for this research, it is important to understand the basic concepts of road networks, public transport networks and how these relate to the OpenStreetMap and NDOV data models.

Public transport networks are usually very large and consist of lots of elements (Hadas 2013). They consist of both spatial data and temporal data. Examples of spatial data are line elements such as roads and railway routes and point elements such as bus stops. Examples of temporal data are operating times and traveling times.

Since 2006, Google offers a service called Google Transit. With this service, it's possible to plan routes via public transport. To make this possible, Google introduced a standard now known as "General Transit Feed Specification" (GTFS). This standard is for example used for their own Google Transit planner. Interface 1 data is used in this planner, as it provides real-time information regarding public transport in the whole country.

OpenStreetMap does contain public transport data on routes, but data regarding travel times, operating hours and frequency are not included. The public transport routes in OpenStreetMap can

[2] The interfaces are referred to in Dutch as Koppelvlakken and abbreviated as KV

be extracted, but it's not possible to treat the results as a full public transport network. Due to the lack of temporal data in the OpenStreetMap dataset, it is not considered to be a public transport network.

2.4 Elements of spatial data quality

The quality of spatial data is a broad subject. While requirements as availability and timeliness are quite obvious, "quality" is a vague concept. Several types of spatial quality can be defined. A lot of studies to spatial data quality have been performed. An example is the work of Van Oort (2006), who distinguishes 11 elements of spatial data quality, based on different studies to spatial data quality done before. Van Oort distinguishes:

- Lineage, which is used to describe the history of a geographic dataset;
- Positional accuracy, the same as spatial accuracy;
- Attribute accuracy, the same as thematic accuracy;
- Logical consistency;
- Completeness;
- Semantic accuracy, which includes errors;
- Usage, purpose, and constraints, a broader description of "usability" in the ISO standard;
- Temporal quality;
- Variation in quality, which is only relevant if the quality within a dataset differs, and therefore not included in the ISO standard;
- Meta-quality and
- Resolution.

In related work, selections of these elements are used for defining spatial data quality. A more standardized way for defining elements of spatial data quality would be to use the ISO standard for spatial data quality measurement. ISO is the International Organization for Standardization and defines standards which are widely used. The ISO 19157 standard defines how spatial data quality should be assessed and describes the following elements of data quality:

- Temporal quality (e.g. how up to date is the information);
- Spatial accuracy (position of features on the earth);
- Completeness: commission (data is present in dataset) and omission (data is absent);
- Logical consistency (to what extent does the data follow logical rules);
- Thematic accuracy (e.g. how accurate is the classification used) and
- Usability (can the data be used for the intended goal).

Because these six elements are the most important element of spatial data quality they will be examined in more detail in the following part. For every element of spatial data quality a requirement regarding the bus route datasets will be formulated.

2.4.1 Temporal quality

For bus route planning, it is very important to have data which is as up-to-date as possible.

For OpenStreetMap, the dataset can in fact change at any moment. Users can change the dataset at any time and therefore the dataset could theoretically always be up-to-date. A small limitation is the standard "Transport Map" layer on the OpenStreetMap website, which is updated after approximately a day. However, the source data is always changed at the same moment. Only the visibility of this source data on the OpenStreetMap website takes one week.

For the planned timetable data delivered by transport operators and provided via Interface 1 (planned schedules), the changes in the schedule have to be provided at least two weeks before they are to be put in operation. It is therefore possible to insert future changes in the dataset. For example, in the interface 1 file to be used as of January 1st, there might already be a rule for a changing timetable for February 1st.

The NDOV website is checked for the frequency of delivering new datasets. All the transport operators deliver a new dataset at least once a month, most of them deliver new datasets more often.

The planned timetable can – but does not have to – include information on diversions. In the case of temporarily small diversions, lasting some days or even multiple weeks, these changes in the bus route will most likely not be in the datasets. For diversions, lasting multiple months, the changes are more likely incorporated in the dataset, but this is not mandatory. (source Province Gelderland concessievoorwaarden).

In OpenStreetMap, it's not the intention to add diversions and construction works.

The real-time bus data does, of course, show the actual position of the bus, also when the bus is taking a diversion at the moment. The message sent by the bus is an "OffRoute" message instead of an "OnRoute" message and can, therefore, be recognized.

Requirement: bus route should always be up-to-date, with an exception for diversions.

2.4.2 Spatial accuracy

The spatial accuracy can be measured using the technique of Haklay (2010). He used buffers around elements to compare their locations. (Jackson et al. 2013) did the same, but extended this approach to point elements. Other examples can be found in Zandbergen et al. (2011) and Kounadi (2009). Especially the locations of the bus stops should not be far from the real situation. Because of the lack of a reference dataset with bus stops, this will however not be investigated. The spatial accuracy of the routes will be assessed. For a passenger it does not matter if the bus drives via road A or via road B, as long as the bus drives from the right bus stop to the next bus stop. The bus should however drive on a road, so the bus route will have to be within a distance from a road network.

Requirement: bus routes should be within an acceptable distance from the reference road network.

2.4.3 Completeness

Completeness can either be a commission (data is present in the dataset) and omission (data is absent). The datasets with bus routes will be compared with an independent road infrastructure dataset. Proposed is TOP10NL because of the spatial accuracy compared with GPS. + up to date.

Furthermore, a public transport dataset should contain all the available routes. Haklay (2010) compared the road network of OpenStreetMap with Ordnance Survey maps and calculated the complete length of the routes in both datasets to test the completeness. He found an average overlap of 80%. This value can be used to measure completeness: if around 80% of the buffers overlap, the dataset is complete. This technique can be extended with using secondary information about bus routes: a list of the bus routes included in both datasets can be compared to see if all existing bus routes are actually included.

Requirement: all the bus routes in the concession area should be present in the dataset.

2.4.4 Logical consistency

Logical consistency includes conforming to topological rules. Logical consistency also includes connectivity, according to Girres & Touya (2010). For the bus routes is particularly interesting if the routes themselves are one connected line. Especially for the creation of a road network, this situation should be avoided, but also for route planning purposes this is an unwanted situation.

Requirement: the datasets should be topologically correct, to be able to navigate the bus routes.

2.4.5 Thematic accuracy

All bus routes should be tagged in the same way, according to a set of rules. For the bus route datasets, most important are bus line number, combined with origin in the destination, because one line number might exist multiple times in a dataset. Matching tags are essential to be able to compare specific bus lines in both datasets.

OpenStreetMap has a Wiki with a list of rules which users should adhere to. For bus routes the keys “type=route” and “route=bus” are required, ref and network keys are important and some other keys are optional (OpenStreetmap Wiki 2017).

For the planned timetable date, a LinePlanningNumber is a mandatory field (BISON 2015b).

Requirement: specific bus lines in the dataset should be detectable by using corresponding fields.

2.4.6 Usability

The quality of a dataset depends on the usability, as both van Oort and the ISO standard mention it. Van Oort (2006) describes the difference between usage, purpose, and constraints. The intended use of a dataset (the purpose) might not be the same as the usage of the dataset. For example, OpenStreetMap is not primarily meant to be used as a cycle route map. When the dataset is used to plan a cycle route, the purpose is broader than the usage.

An example is the work of Mondzech & Sester (2011) who did a quality analysis based on applications needs, in their case pedestrian navigation. OpenStreetMap data was compared to data from ATKIS, the German topographic dataset. The quality of both datasets was measured by planning the shortest route between arbitrarily chosen points in both datasets, for urban and rural scenarios.

Usability also includes the accessibility of the dataset, which is the most important element for this research.

Requirement: the datasets should be easily accessible and can be used for the intended goal.

Recently, the principal of “Fitness for use” became more important. Instead of a data-centric view, the actual use of the dataset became a key issue in assessing the quality of a dataset.

This research will be data-centric instead of user-centric because the timeframe is limited. However, the idea of “fitness for use” is an interesting subject for further research and will be kept in mind during this study.

3. Methodology

This chapter describes the methodology used for this study. The structure adheres to the research objectives as formulated before:

- To understand the OpenStreetMap and NDOV data structures (Objective 1);
- To assess the quality of real-time bus locations delivered by transport operators (Objective 2);
- To assess the quality of public transport data delivered by transport operators via both NDOV desks (REISinformatiegroep and OpenOV) (Objective 3);
- To assess the quality of public transport data in OpenStreetMap (Objective 4).

Figure 3.1 shows the structure of these objectives and results and will be further explained in this chapter.

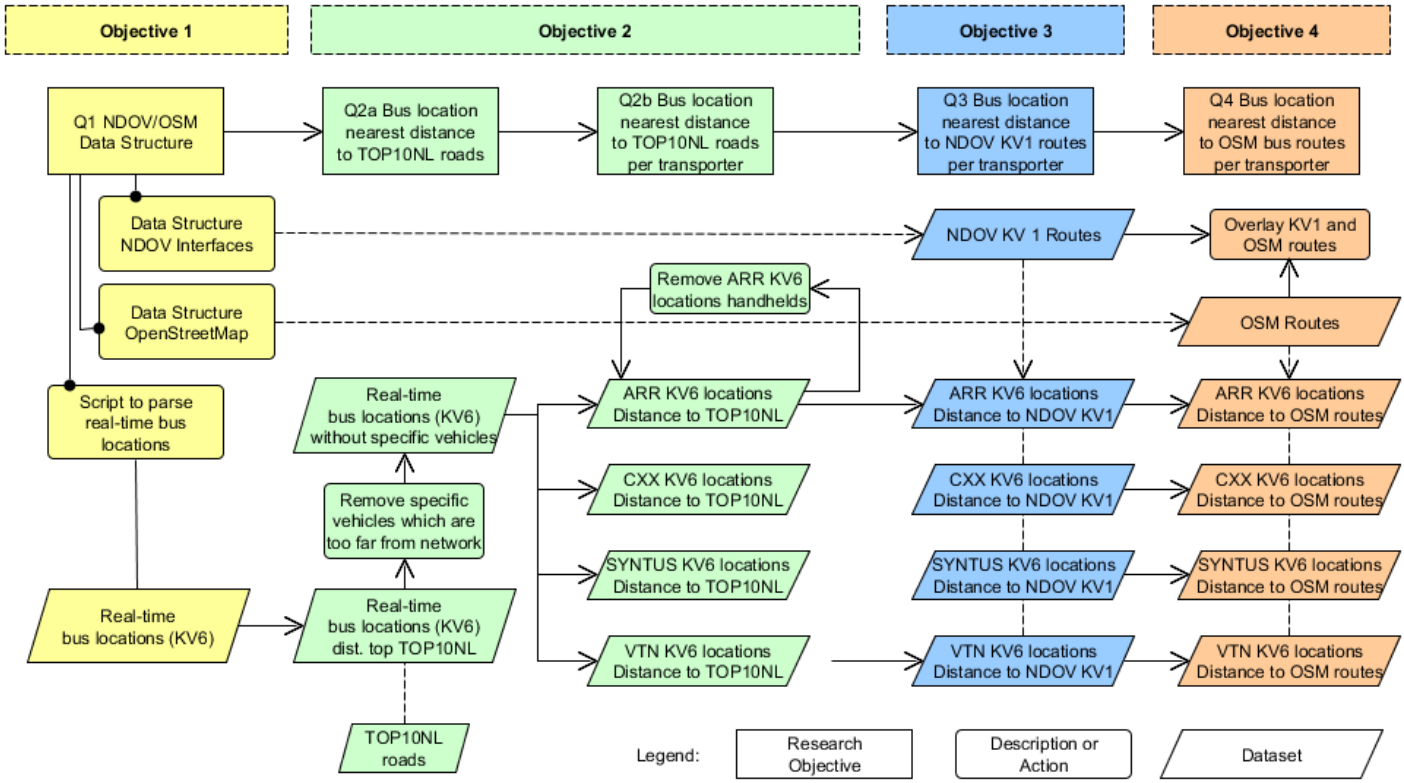


Figure 3.1 Methodology

3.1 OpenstreetMap and NDOV data structures

The first objective is to understand the OpenStreetMap and NDOV data structures. Three main datasets will be used throughout the research: a dataset with the OpenStreetMap bus routes, a dataset with the bus routes as delivered by the transport operators and a reference dataset.

The reference dataset is needed to be able to check the quality of the bus routes delivered by transport operators. PDF files with the actual bus routes are available, but are slightly schematic and probably not suitable to be imported in a GIS. The creation of an own dataset by following the routes of buses with an own GPS device is a possibility, but it's impossible to do this for the whole province given the time frame of the research. Transport operators are required to deliver the real-time

positions of bus vehicles via the NDOV desks. These are available via Interface 6. By collecting this data for a given period of time, a reference dataset can be created.

The planned bus routes delivered by transport operators are also available via the NDOV desk, via the Interface 1. The OpenStreetMap dataset can be downloaded directly from the OpenStreetMap server.

The result of this first objective includes a description of the data model of the source datasets (OpenStreetMap and both NDOV datasets) and the procedure to create the desired dataset. Also a script will be created to make it possible to parse the real-time bus locations.

Outcomes:

- Description of data structure Interface 6 (real-time bus locations)
- Description of data structure Interface 1 (planned bus routes)
- Description of data structure OpenStreetMap (roads and bus routes)
- Script to parse Interface 6 real-time bus locations

These three datasets will be created and serve as input for the other questions.

- A dataset with real-time point locations of buses, provided via NDOV Interface 6
- A dataset with the planned bus routes in NDOV, provided via NDOV Interface 1
- A dataset with bus routes in OpenStreetMap

3.2 Assess the quality of the real-time bus locations

The second objective is therefore to assess the quality of this Interface 6 bus locations delivered by transport operators.

These quality requirements have been defined:

- Bus route should always be up-to-date, with an exception for diversions;
- Bus routes should be within an acceptable distance from the reference road network;
- The datasets should ideally match exactly;
- The datasets should be topologically correct, to be able to navigate the bus routes;
- Specific bus lines in the dataset should be detectable by using corresponding fields;
- The datasets should be easily accessible and can be used for the intended goal.

The reference dataset consists of real-time bus locations and is therefore by definition up-to-date.

Most important is the spatial accuracy of the point data. The GPS accuracy used by transport operator is an important indication of the expected spatial accuracy of the data. The concession terms for transport operators require them to deliver real-time locations of the buses. Specifications for the GPS used cannot be found, but a normal GPS accuracy for a GPS in a car is within 10 meters. However, the signal could be less accurate or even disappear (Longley et al. 2010, p.141)

The real-time point locations of the buses should be within 10 meters from the road network. To test this, the parsed real-time bus locations have to be converted to point features. The nearest distance from a point feature to the road network will be calculated, which results in a specific distance to a road per point.

The point distance to the road network can be used to select points which are too far from the road network. As mentioned before, the average GPS accuracy is around 10 meters. The signal may however be less accurate in for example urban areas. Therefore, a threshold of 20 meters is used. Points further than 20 meters of a road can be considered “not accurate enough”.

The next step is to calculate the average distance to the road network per vehicle. Every combination of transport operator and vehicle number is unique. In this way, it’s possible to select vehicles which are too far from the road network within the whole dataset. This may include vehicles with a broken or inaccurate GPS system.

Finally, the average distance to the road network for all points per transporter can be calculated. This will make it possible to compare the average values from the different transport operators with each other.

Other quality elements besides spatial accuracy will not be tested. The number of point locations in the dataset is variable, therefore it’s not possible to test the completeness of the dataset. The thematic accuracy could be tested by performing basic statistical analysis on fields such as bus line number, but this will not be done because of the limited size of this study.

Outcomes:

- Real-time bus locations with distance to nearest TOP10NL road;
- Real-time bus locations with distance to nearest TOP10NL road with specific vehicles removed;
- Real-time bus locations with distance to nearest TOP10NL road per operator.

3.3 Assess the quality of NDOV bus route dataset

The public transport route data used in for example Google Transit is derived from NDOV. The planned timetable, routes, stops etc. are provided via “Interface 1”. The collection of this dataset is more complicated than the collection of the OpenStreetMap dataset. The data model of Interface 1 is not only much more complex than the OpenStreetMap data model; there are also more choices to make because of different available datasets. The first choice is to pick one of the NDOV desks, as there are two desks operating. The policy of the OpenGeo NDOV desk is no to change anything in the datasets delivered by the transport operators. Transport operators are responsible for the content of their data, the task of the NDOV desk is purely to make this data available for third parties. Because of this policy, the OpenGeo desk will be used.

To assess the quality of the NDOV Interface 1 dataset, the data will be compared with the real-time bus locations. First, Interface 1 data (planned bus routes) will be selected for the same period as the real-time bus locations are collected.

Interface 1 datasets will be created using both NDOV desks. The data delivered by the transport operators should be the same, so the outcomes should also be the same. The datasets can be compared to check if they differ at some point.

Real-time bus locations (the outcome of the previous step) will be selected for the same period as the Interface 1 dataset. The nearest distance from the real-time bus locations to a Interface 1 route will be calculated and reviewed per operator.

Every point is an OffRoute or OnRoute message. This could explain points which are not following a planned route. OffRoute messages too far from the bus routes will be selected to explain these situations.

Outcomes:

- Real-time bus locations with distance to nearest NDOV Interface 1 route per transport operator.

3.4 Assess the quality of the OpenStreetMap dataset

Finally the bus routes in OpenStreetMap have to be collected. First, the data model of OpenStreetMap will be explained, to understand how public transport routes are included in OpenStreetMap. Public transport routes are included in OpenStreetMap as a collection of road sections, which comes with some advantages and disadvantages. Then, the procedure for creating the dataset will be explained.

The completeness, thematic accuracy and logical consistency will be measured in the same way as the Interface 1 dataset. The results will be compared.

The method to assess the quality of the OpenStreetMap dataset is similar to the previous step. The real-time bus locations will be selected for the same period and bus route as the OpenStreetMap dataset and the nearest distance will be calculated.

Finally, the OpenStreetMap and Interface 1 datasets will be compared by overlaying them. The Interface 1 bus routes will get a 10-meter buffer around each route (this is equal to the acceptable distance from the network). Then it will be checked if all of the OpenStreetMap routes are within this buffer. In this way it's possible to find locations where the datasets differ too much.

Outcomes:

- Real-time bus locations with distance to nearest OpenStreetMap bus route per transport operator;
- OpenStreetMap bus routes within 10-meter buffer of Interface 1 bus routes.

3.5 Reference datasets

To select bus routes and points within the province of Gelderland, a vector dataset with the borders of the province of Gelderland is downloaded.

To check the quality of the real-time bus locations an independent roadmap is used. The TOP10NL dataset is an up-to-date road dataset of the whole Netherlands. The scale of this map is 1:10.000. Bus locations are collected with a GPS and can be a little inaccurate. A scale of 1:10.000 is, therefore, sufficient for the research. The dataset can be downloaded as polyline dataset. Both road axis and road sections are collected.

3.6 Area of data collection

To collect the data, an area has to be defined. The research area is the province of Gelderland, which consists of three concession areas for public transport. Every concession area has its own transport company:

- Concession Achterhoek-Rivierenland, currently exploited by Arriva;
- Concession Arnhem Nijmegen, currently exploited by Hermes (Connexxion) and
- Concession Veluwe, currently exploited by Syntus.

Figure 3.2 shows the locations of these concession areas.

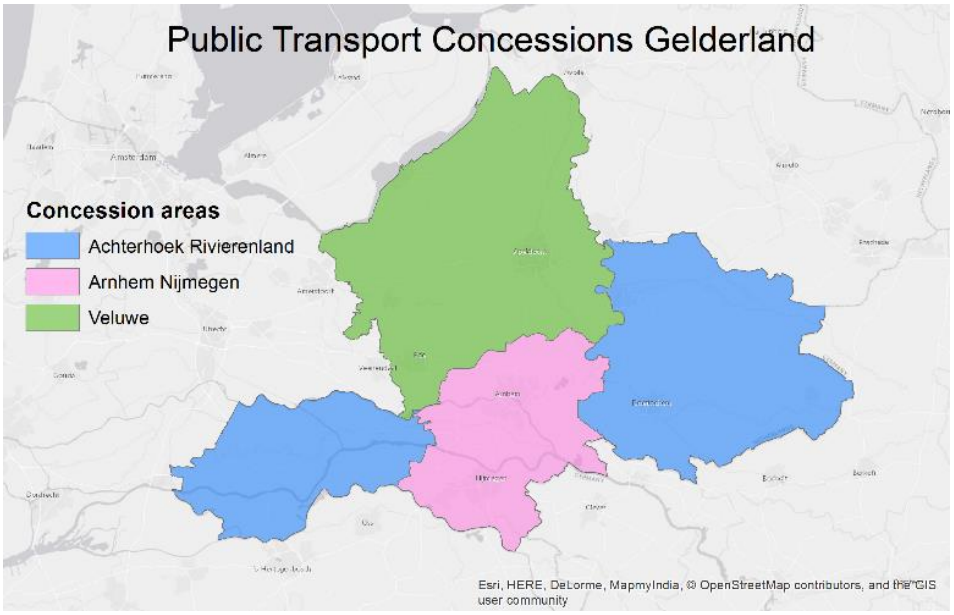


Figure 3.2: Public transport concessions in the study area

The duration of all concessions is until 2020 or later, so the transport companies will not change during the research. The new timetable (and possible new routes of bus lines) starts for every transport company on Sunday, December 13th, 2015. All datasets (OpenStreetMap data, NDOV data, PDF maps of the bus network etc.) will be collected after this date. A list of changes in the bus routes per December 13th is available.

4. Datasets

This chapter describes the structure of the Interface 6 (real-time bus locations), Interface 1 (bus routes delivered by transport operators) and OpenStreetMap datasets. For every dataset, the data model and the procedure to create the dataset are explained.

4.1 Interface 6 dataset

Interface 6 (Dutch: Koppelvlak 6, abbreviated as KV6) is the standard to make real-time positions of vehicles available. This paragraph will explain the data model and creation of this dataset. The information in this paragraph is based on BISON (2015a)[3]

4.1.1 Interface 6 Data model

The Interface 6 data model is a complex model, which can deal with all kinds of situations, such as vehicles departing from a bus stop, driving between two bus stops, arriving at a bus stop, or not driving via the planned route, for example. All the messages sent by buses are delivered as an XML file. Appendix A explains how the XML file and the possible messages are structured.

From the possible messages, 6 are relevant for this research, these are displayed in figure 4.1

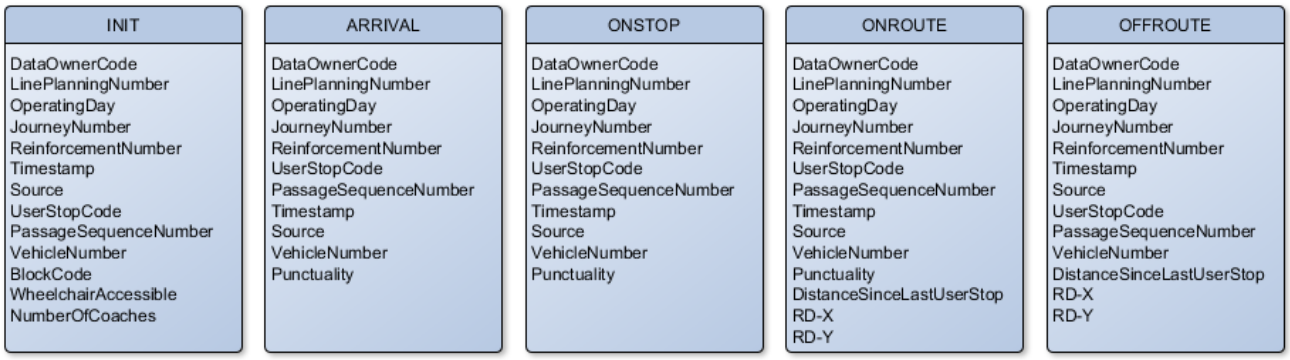


Figure 4.1: Interface 6 data model (selection)

As the focus is on the locations of the bus, only messages where a location can be obtained are relevant. In figure 4.1 can be seen that only “OnRoute” and “OffRoute” messages have RD coordinates included. OnRoute messages are sent if a bus is on the planned route, Offroute messages are sent if a bus is not on the planned route, for example, if there are road works are if the driver accidentally missed a junction.

The “init” (message sent by a bus starting at its first location), “arrival” (a bus arrives at a bus stop) and “onstop” (a bus is standing still at a bus stop) have an *UserStopCode* included. The *UserStopCode* can be linked to a bus stop (delivered via Interface 1) and can therefore also be linked to a location.

4.1.2 Creating the real-time bus location dataset

It is difficult to collect the Interface 6 data per transport operator or concession area. The raw data stream includes the bus locations for the whole country. Therefore, a rectangular window is used to limit the size of the dataset. The coordinates are determined manually, in such a way that the whole province of Gelderland is included (Table 4.1).

[3] BISON is an abbreviation of Beheer Informatie Standaarden OV Nederland (Management Information Standards Public Transport The Netherlands).

X-coordinate minimum (western border)	127711
X-coordinate maximum (eastern border)	256238
Y-coordinate minimum (southern border)	416761
Y-coordinate maximum (northern border)	504921

Table 4.1: Spatial boundaries (RD coordinates)

A script was written to make it possible to collect this data. Via the NDOV an ongoing stream of XML messages is delivered. A python script was created to parse these messages into a comma-separated values file (csv)

```
if message_type == 'ONROUTE':
      routetype = 1
      return parseKV6(message, message_type, required +
      ['userstopcode', 'passagesequencenumber', 'vehiclenumber',
      'punctuality', 'distancesincelastuserstop', 'rd-x', 'rd-y'])
```

Offroute messages are collected in a corresponding way

```
elif message_type == 'OFFROUTE':
      routetype = 2
      return parseKV6(message, message_type, required +
      ['userstopcode', 'passagesequencenumber', 'vehiclenumber', 'rd-
      x', 'rd-y'])
```

Appendix B includes the full script.

4.2 Interface 1 dataset

This paragraph explains the data model and creation of the Interface 1 (Dutch: Koppelvlak 1, abbreviated as KV1) dataset. The information in this paragraph is based on BISON (2015b).

4.2.1 Interface 1 data model

Interface 1 can be described with an Entity Relationship Diagram (ERD) that has more than 20 entities. Together, these entities provide the full bus routes, bus stops, and timetable data. Timetable data is not relevant for this study because time tables are not included in OpenStreetMap and the aim is purely to compare the geographic bus routes. Therefore, only a part of the ERD is of interest. Figure 4.2 shows the relevant part of the ERD, which entities will be described afterward.

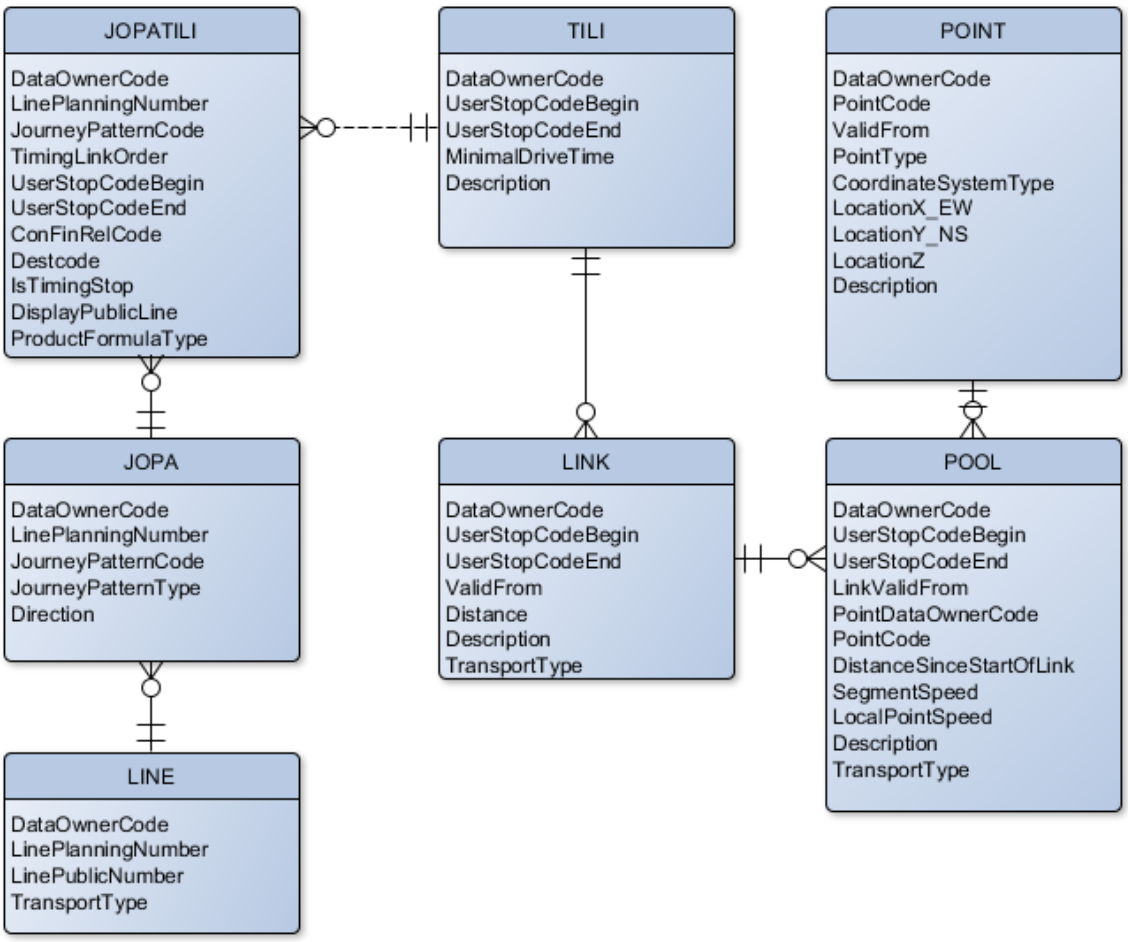


Figure 4.2: Interface 1 data model (selection)

All points exist once in the *Point* table. This includes bus stops, but also points which together form the bus routes. Via the *PointOnLink* table ("POOL") sets of two points are assigned to a *Link*, which forms the route between two bus stops via the road. A link is coupled to a *TimingLink* ("TILI"), which is the logical connection between two bus stops. A *TimingLink* does not include the exact route of the bus, but only the logical connection between two points, including the shortest travelling time.

For every concession area, bus lines are defined in the *Line* table, with an internal *LinePlanningNumber* as well as a *LinePublicNumber*, which is the number that is communicated to the travellers. A bus line can have multiple variants, for example, a different route in the evening or multiple destinations. Also, the route per direction is different and therefore described as a variant of the bus route. Every unique route is defined in the *JourneyPattern* table ("JOPA")

The bus route parts between two bus stops in the *TimingLink* table are connected to the unique route variants in the *JourneyPattern* table via the *JourneyPatternTiminLink* table ("JOPATILI"). Via this table, it is possible to find the route and stops of every specific bus route variant.

4.2.2 Creating the Interface 1 dataset

The datasets have to be created per transport operator/concession area. The Arriva area has been split into two areas (Rivierenland and Achterhoek). Connexxion uses one Interface 1 dataset for the whole country. For every transporter, the data is downloaded from both NDOV desks. Appendix C includes the complete description how the dataset was created.

4.3 OpenStreetMap dataset

In general, the OpenStreetMap data model is less complicated than the Interface 1 and Interface 6 data models. The main reason is that OpenStreetMap does not have data on timetables and fares included. In this paragraph, the data model and creation of this dataset will be explained in more detail.

4.3.1 OpenStreetMap data model

Figure 4.3 is a schematic representation of the OpenStreetMap data model. The OpenStreetMap data model consists of three basic objects: *nodes*, *ways,* and *relations*. Every object can have one or more *tags* to supply information about that object (Bennett 2010)

Point elements, such as a tree or a bus stop are represented by *nodes*. Nodes always have X and Y information included and can have tags. Bus stops do for example have a tag “highway=bus_stop”, while a tree is tagged as “natural=tree”. OpenStreetMap has a wiki which explains for every element how it should be tagged.

Line elements, such as roads or railways are represented as *ways*, which are always lines between one or more *nodes*. Therefore, there is no latitude or longitude included in a line, because this information is already included in the point elements.

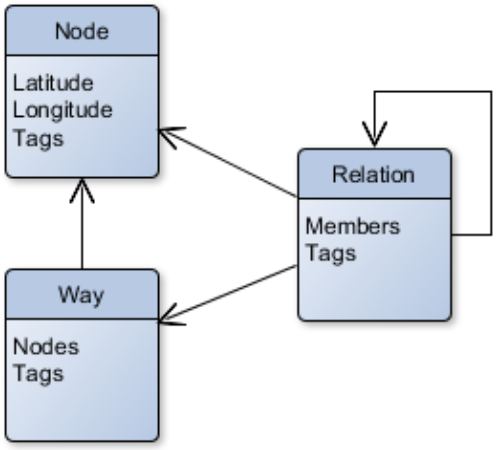


Figure 4.3: Simplified OpenStreetMap data model

There is no special class for polygons in OpenStreetMap. Polygons, such as a building a lake or another area, are represented by *ways* which share the same start and end point. By using the right tags, such as “Building=yes”, a *way* is recognized as a polygon.

Relations consist of multiple *nodes* and/or *ways* which together form another element. One relation can have point and line data at the same time. Some examples of *relations* are:

- Multipolygon, for example, a building with an open inner part. The relation then consists of the polygon which is the outside of the building and the polygon which is the open inner part of the building.
- Route, for example, a bus or cycle route, consisting of multiple roads which together constitute the entire bus route or cycle route. Bus route relations also include the bus stops as point data.
- Boundary; as a boundary will most likely be part of the areas on both sides of the boundary, this type of relation makes sure that only one boundary is needed to describe the boundaries of all involved areas. The boundary between the Netherlands and Belgium will, for example, be part of both the “Boundary of the Netherlands” and the “Boundary of Belgium” relations but is also part of relations which describe the boundary of a province and municipality.

Relations can also purely consist of other relations. For example, bus lines can be collected in a relation consisting of all the bus lines in a specific area. This is of importance for the data collection process.

Keßler et al. (2011) state the number of relations is comparatively very small and do not consider them in their research. Although the number of relations might be relatively small, they do however cover a whole range of information in a “second layer” on OpenStreetMap, especially bus routes, cycle routes, and boundaries.

Barron et al. (2014) developed a framework to evaluate the quality of an OpenStreetMap dataset based on an OSM Full-History-Dump. Their framework makes it possible to investigate data quality without a ground truth dataset. They mention how for future research, relations should be taken into account. It is however not possible to follow their method, as the OSM-History-Importer does not support the import of relations.

4.3.2 Creating the OpenStreetMap dataset

Due to the data model used by OpenStreetMap, public transport routes consist of a substantial number of road sections. This has some advantages and disadvantages. A benefit is a possibility to create a network that matches exactly with a road network. Due to the fact that the road network of OpenStreetMap is used as a basis, it is impossible to have bus routes which go off-road, as long as the road network is considered to be good.

The disadvantage is the high chance to break a bus route when an inexperienced user accidentally deletes or changes a road section.

To create the dataset, first, the members (which are bus lines) of a relation consisting of all the bus lines in a concession area will be downloaded. Every concession area has its own relations. It is, however, possible that not all the bus routes are included in this relation, so caution is needed.

Then the members of the bus lines relations will be downloaded (which are the road sections). The road sections per bus line will be merged into one line per bus line.

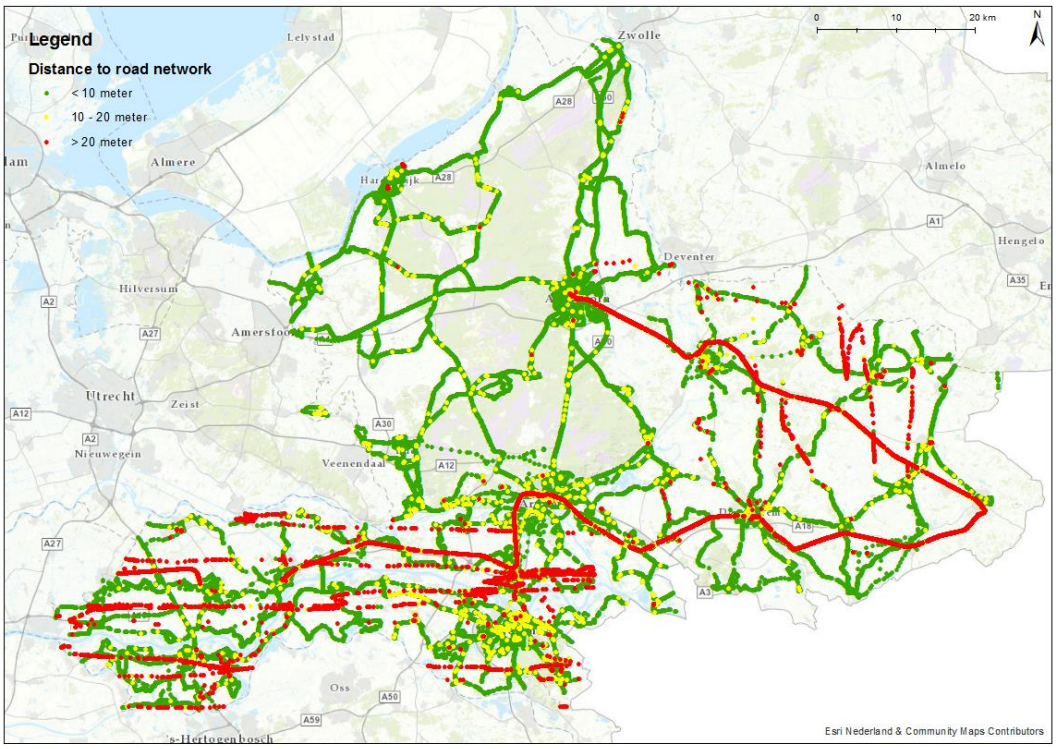
Appendix D describes the process in more detail.

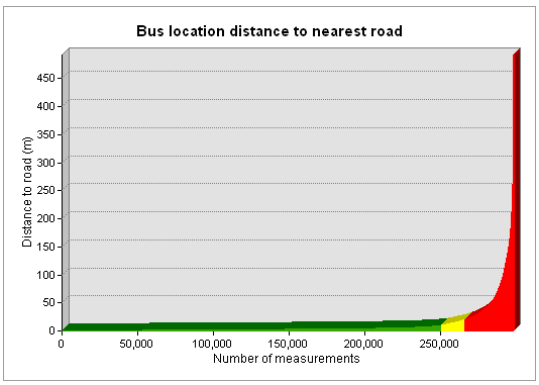
5. Quality Assessment

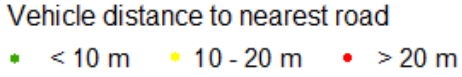
This chapter presents the results of the quality analysis performed. First, all the points are examined: their location to the road network and railroad network are calculated and certain off-route points are filtered out. Then the points are examined per transport operator.
In the second part, the distance of the real-time bus points to the planned bus routes will be examined per operator, in the third part the distance between the real-time locations and the OpenStreetMap dataset is calculated.

5.1 Analysis of the real-time bus locations dataset

The dataset with real-time locations was collected on November 21st, 2016 and consists of 298.204 points. First the distance of the real-time bus locations to the road network has been calculated. The figures below show the distance of the real-time locations to the road network. Over 80 percent of all the points are within 10 meters from the road network. However, a part of the points is further away from the network than acceptable, with distances op to 500 meters.
The map in figure 5.1 shows the spatial distribution of these points.







	< 10 m from road network	10 - 20 m from road network	> 20 m from road network
Points	250816	15351	32037
% of total	84.1%	5.1%	10.7%

Figure 5.1 Bus locations distances to nearest road

Regarding the points further than 20 meters from the road network, two patterns are visible. First, some of the points seem to be clustered around railroads. These can be filtered out by calculating the average distance to a railroad per vehicle. Second, there are some points which seem to form horizontal and vertical lines, apparently regardless of roads or railroads.

5.1.1 Vehicles near railroads

Some of the points further than 10 meters away from the road network seem to be concentrated around railway tracks. Therefore, the distance from the points to the railroad network is calculated. Then, the standard deviation of the distance to the railroad is calculated per vehicle. In this way, the vehicles which are trains can be recognized. Figure 5.2 shows the distance to the nearest railroad, grouped per vehicle.

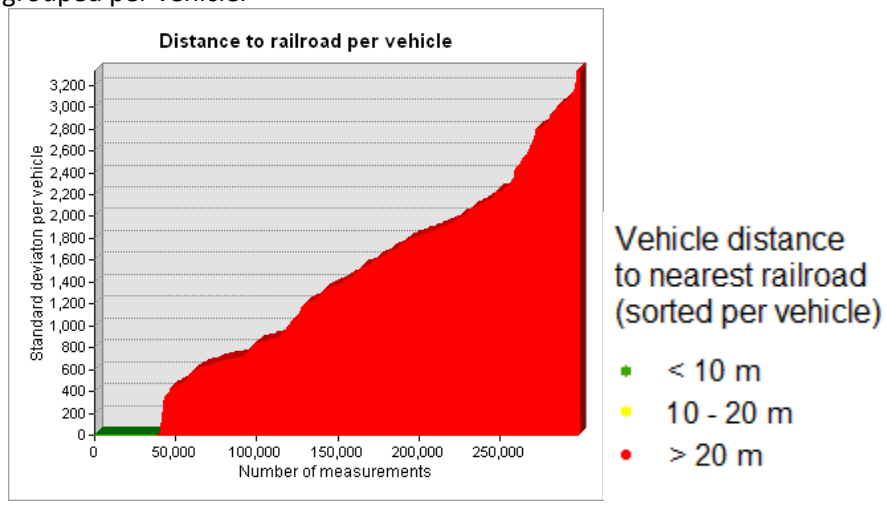


Figure 5.2 Distance between real-time bus locations and railroads

The vehicles which are trains can clearly be recognized by the low standard deviation of distance to the nearest railroad. If the standard deviation of the average distance to a railroad per vehicle is less than 3.0, the vehicle is considered to be a train and is taken out of the dataset. Figure 5.3 shows the points which are deleted from the dataset because they appear to be trains.

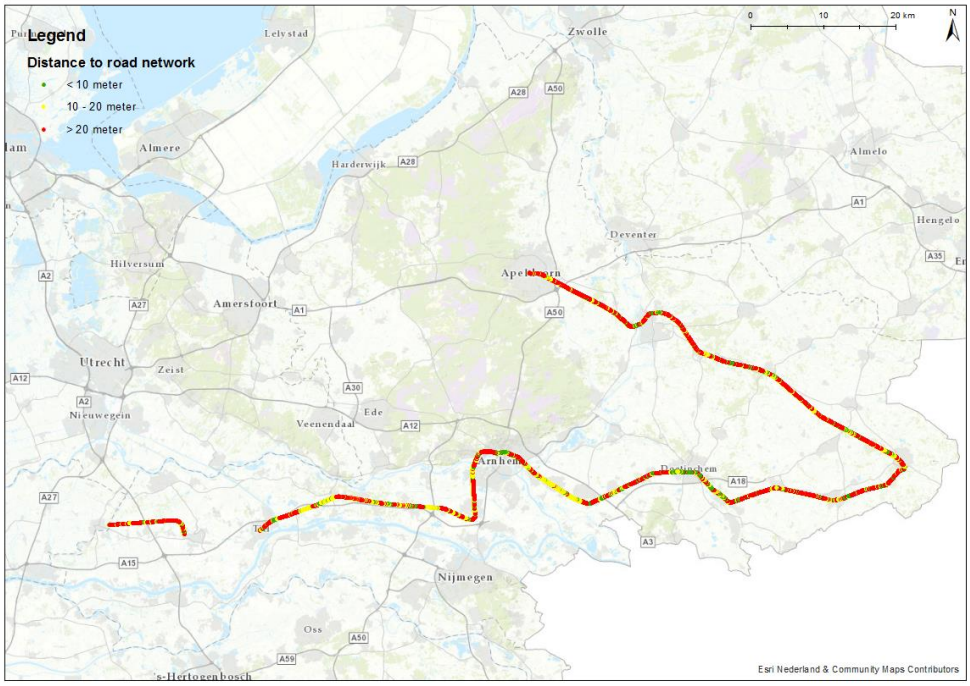
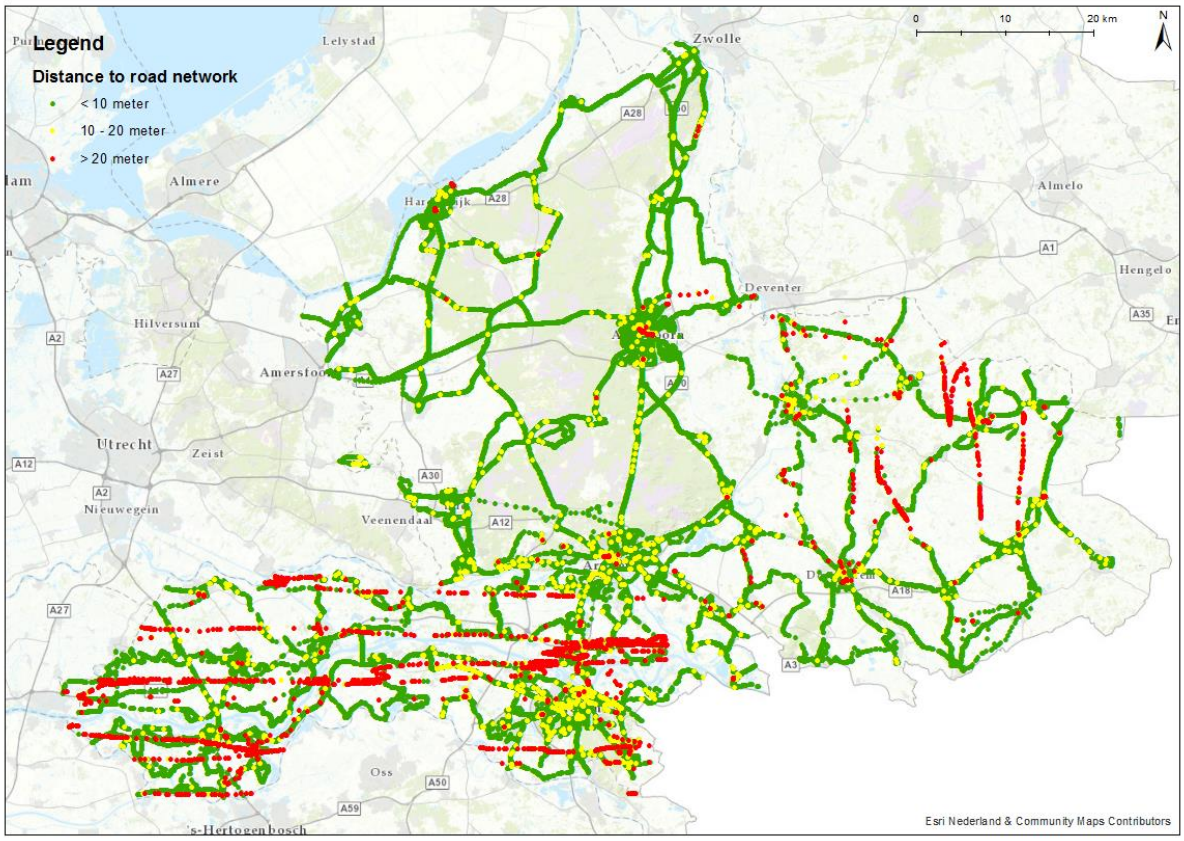
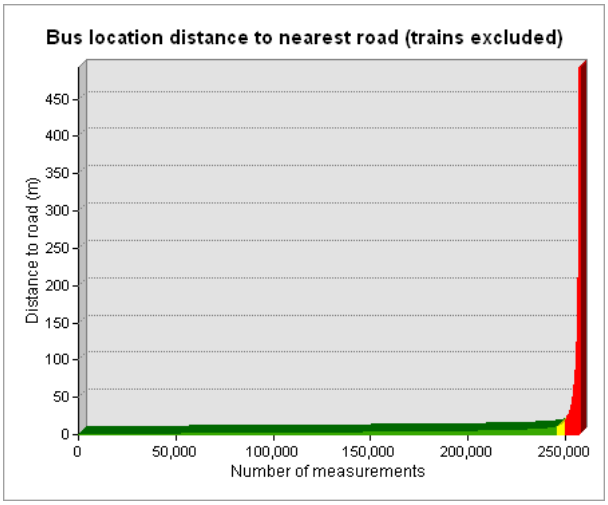


Figure 5.3 Train point location distances to nearest road (Arriva)

A first conclusion is that the NDOV dataset with real-time bus locations does not only have bus locations, but trains are included as well.





Vehicle distance to nearest road

< 10 m 10 - 20 m > 20 m

	< 10 m from road network	10 - 20 m from road network	> 20 m from road network
Points	245761	4669	6580
% of total	95.6%	1.8%	2.6%

Figure 5.4 Bus locations distances to nearest road, trains excluded

Figure 5.4 shows the new dataset, without the trains. Compared to the first figure, the percentage of points within 10 meters from the road network has grown from 84 percent to above 95 percent. The percentage of points further than 20 meters from the road network has decreased from 10 percent to less than 3 percent.

5.1.2 Issues per operator

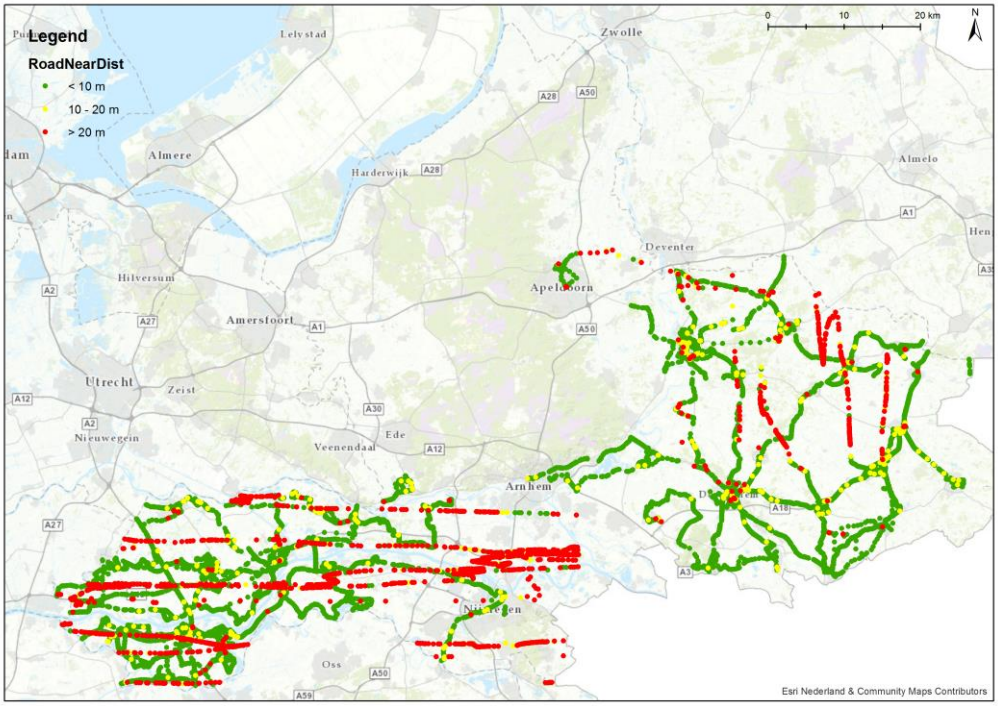
In the next step, the points will be examined per transport operator.
The distance to the road is calculated based on the output of the previous step (trains are excluded). Arriva has the highest average road distance. However, the average distance to the road decreased from 26 to 10 meters after removing the trains. Table 5.1 shows the number of points per transport operator with and without trains.

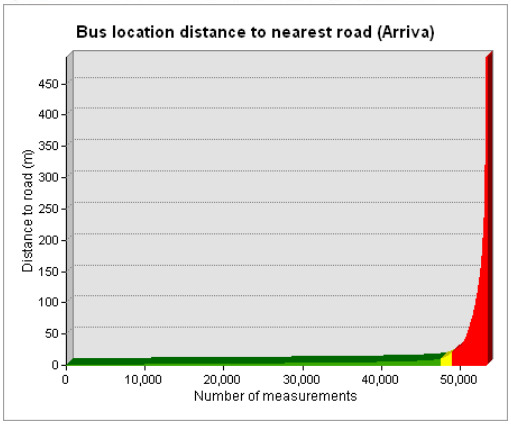
Operator	Number of points (original)	Average road distance (original)	Number of points (trains excluded)	Average road distance (trains excluded)
Arriva	95040	25.62	53856	9.90
Connexxion	77537	3.73	77513	3.72
Syntus	125292	2.81	125292	2.81
Veolia	335	2.60	335	2.60

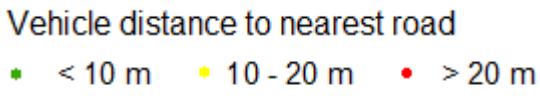
Table 5.1 Number of points and average road distance per transport operator

Arriva

The Arriva dataset has the largest number of points further than 10 meters from the road network. Figure 5.5 shows the distribution of these points.







	< 10 m from road network	10 - 20 m from road network	> 20 m from road network
Points	47644	1526	4687
% of total	88.5%	2.8%	8.7%

Figure 5.5 Bus locations distances to nearest road (Arriva)

Some of the points seem to be in horizontal and vertical lines and cannot be related to routes or roads (figure 5.6). Those points have a vehicle number larger than 22000. These vehicles have handheld GPS devices. These are used in small busses without built-in chip card equipment.

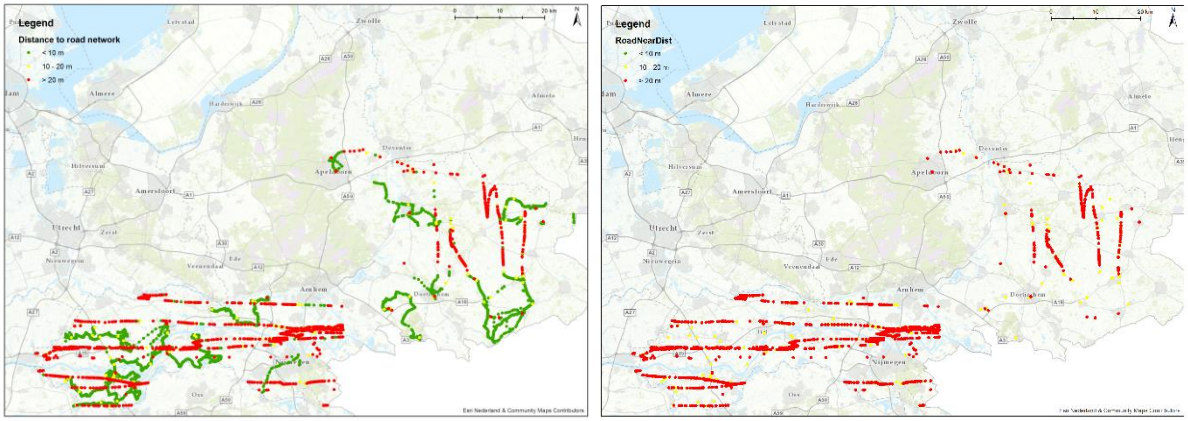
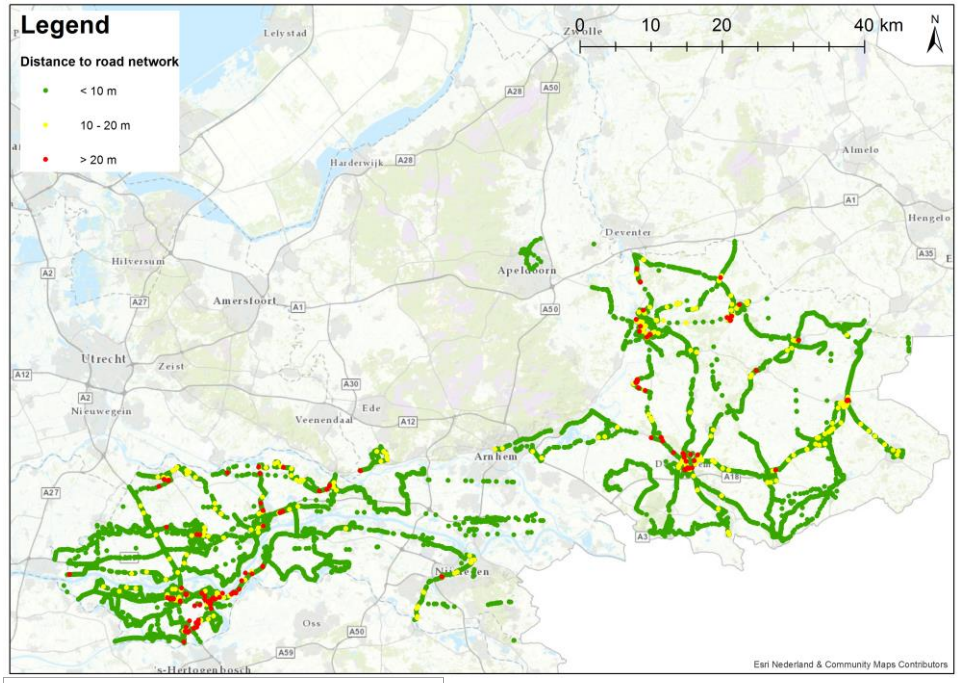
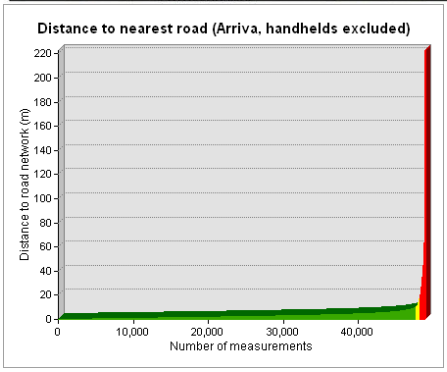


Figure 5.6 Points by handheld devices (left) and handheld devices >10 m from road network (right)

All the handheld devices are selected and the points further than 10 meters from the road are excluded from the dataset. Handheld devices with a location closer than 10 meters from the road network are not excluded. Some of them clearly present correct bus routes as can be seen in the left image above.





Vehicle distance to nearest road

< 10 m 10 - 20 m > 20 m

	< 10 m from road network	10 - 20 m from road network	> 20 m from road network
Points	47,644	416	209
% of total	98.7%	0.9%	0.4%

Figure 5.7 Bus locations distances to nearest road (Arriva), handheld devices excluded

Figure 5.7 shows the result after removing handheld devices further than 10 meters from the road network. 625 points (1.3% of all points) are still too far from the road network. Some of the points are just slightly off the road, other points seem to be clustered around bus stations or depots. Figure 5.8 shows some points near the train station of Zaltbommel (left), a location where buses buffer waiting for a new ride and some points around a bus depot (right).

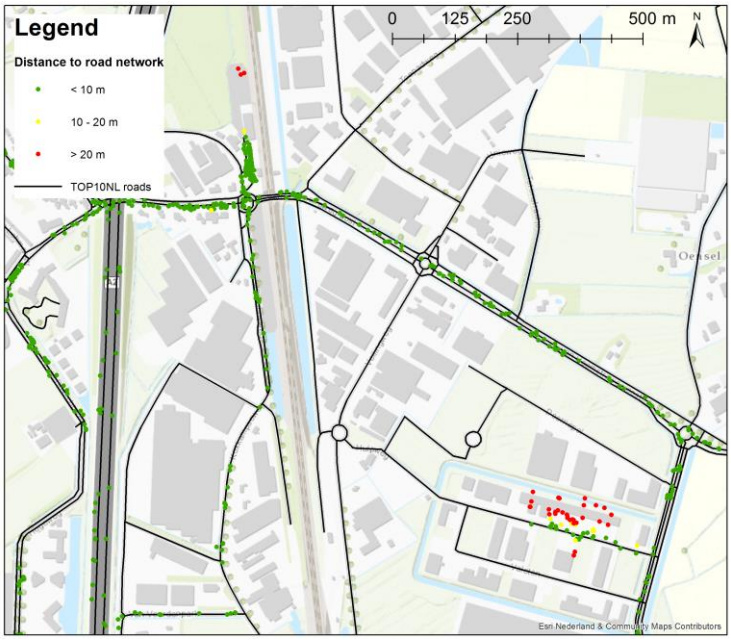
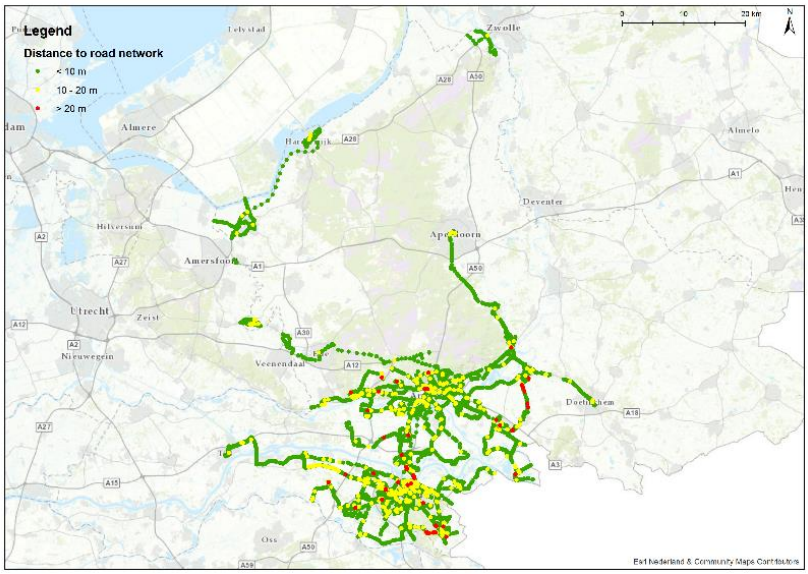


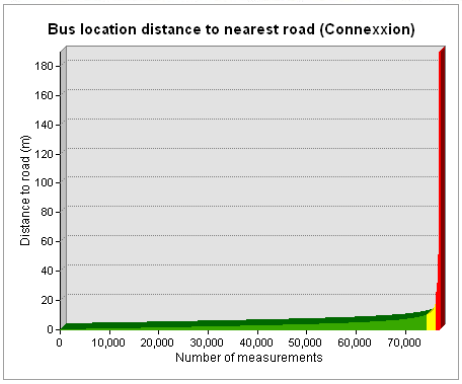
Figure 5.8 Off-road points (example Arriva dataset)

Overall, only 0.4% of the points is further than 20 meters from the road network after removing the handheld devices which are more than 10 meters from the road network. Therefore, the quality of the dataset is sufficient.

Connexxion

The Connexxion point dataset does not have unexplainable point locations like the Arriva dataset. Most of the points are within an acceptable distance from the road network The percentage of points further than 20 meters from the road is slightly larger (0.8% instead of 0.4%) than in the Arriva dataset. Figure 5.9 shows the distribution of these points.





Vehicle distance to nearest road

< 10 m 10 - 20 m > 20 m

	< 10 m from road network	10 - 20 m from road network	> 20 m from road network
Points	74,371	1,816	642
% of total	96.8%	2.4%	0.8%

Figure 5.9 Bus locations distances to nearest road (Connexxion)

The main reason for the high percentage of points further than 20 meters from the road network are some new roads which are not yet included in the TOP10NL dataset.
The left part of figure 5.10 for example, is responsible for 230 points (0.3%) of the points further than 20 meters from the road network. These two examples therefore largely explain the number of off-route points.

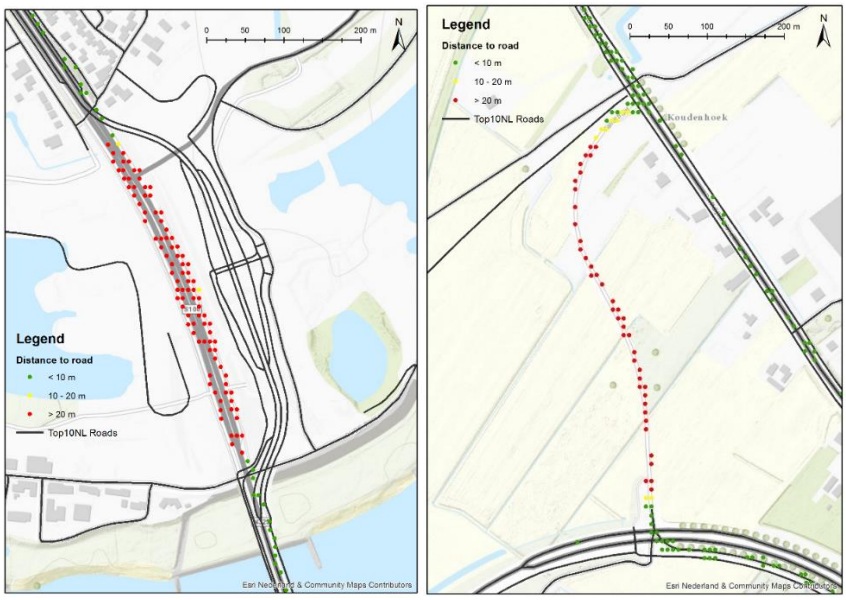
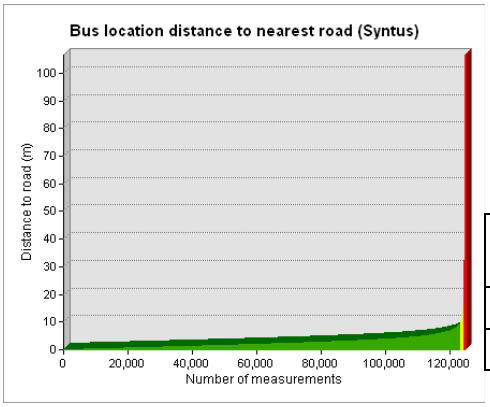


Figure 5.10 Missing roads in TOP10NL dataset near Nijmegen (examples Connexxion dataset)

Syntus

The spatial accuracy of the real-time bus points delivered by Syntus seems to be the best. Only 0.2% of the points is further than 20 meters from the road network. The distribution is presented in figure 5.11.



Vehicle distance to nearest road

< 10 m 10 - 20 m > 20 m

	< 10 m from road network	10 - 20 m from road network	> 20 m from road network
Points	123,415	918	273
% of total	99.0%	0.7%	0.2%

Figure 5.11 Bus locations distances to nearest road (Syntus)

The main reason for the points further than 20 meters from the road is again the lack of some roads in the dataset. Figure 5.12 consists of 106 points further than 20 meters from the road network. This is almost the half of all the points in this category.

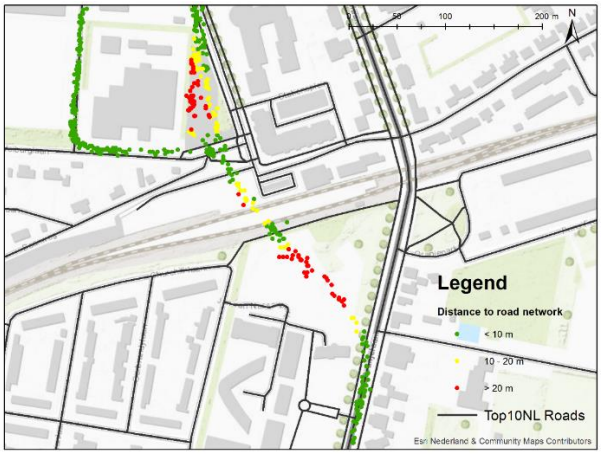
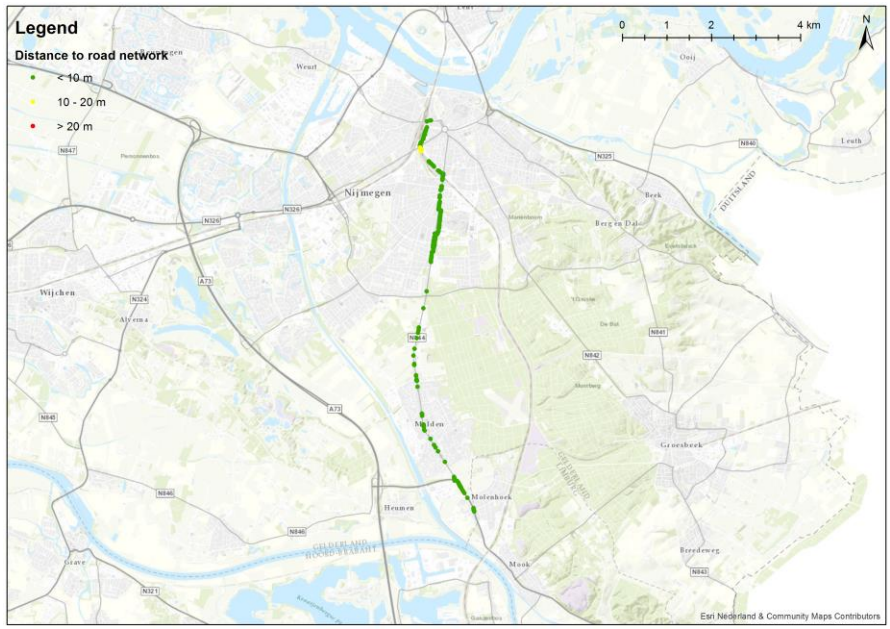
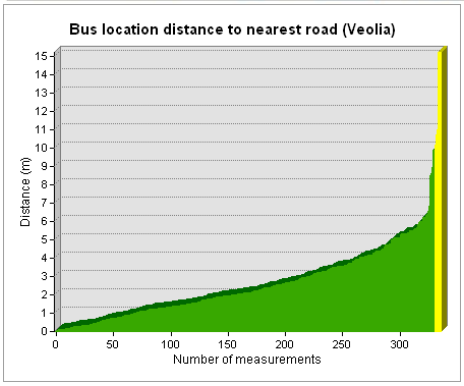


Figure 5.12 Missing roads in TOP10NL dataset in Harderwijk (examples Syntus dataset)

Veolia

The last transport operator is Veolia Transport. This operator has only one bus line within the province of Gelderland, but will still be reviewed. Only 4 points are more than 10 meters from the road network. This is apparently because of an inaccurate GPS. Figure 5.13 shows the points for Veolia.





Vehicle distance to nearest road

< 10 m 10 - 20 m > 20 m

	< 10 m from road network	10 - 20 m from road network	> 20 m from road network
Points	331	4	0
% of total	98.8%	1.2%	0.0%

Figure 5.13 Bus locations distances to nearest road (Veolia)

Overall it can be concluded that there are two main categories in the points further than 10 meters from the road network: points some meters too far from the road, probably due to an inaccurate GPS signal and points in a larger distance from the road network, due to missing roads in the reference dataset.

The datasets for the different transport operators are good enough to be used in the next part of the research. The percentage of points too far from the road is relatively low.

5.2 Analysis of the NDOV Interface 1 dataset

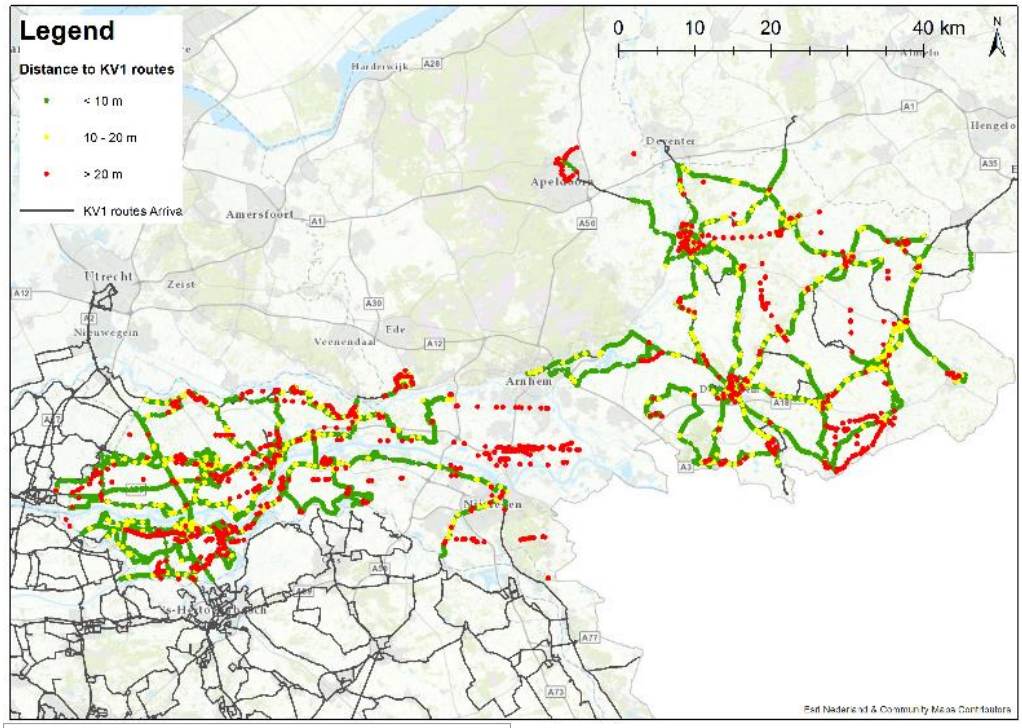
The next step is to compare the locations of the real-time bus locations to the planned bus routes. In the previous part was concluded that all the points were in acceptable distance from the road network. Therefore, all the points will be used, except for the deleted handheld points in the Arriva dataset.

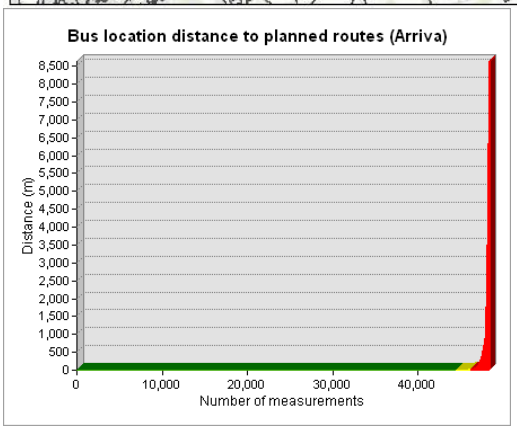
5.2.1 Distance to planned routes

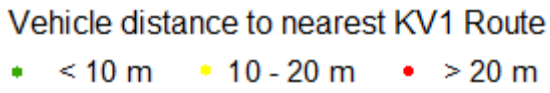
Because the points and routes are available per transport operator, the datasets will be reviewed per operator. For every transport operator, the distance from the real-time bus locations to the planned bus routes is calculated.

Arriva

Most of the real-time bus locations are within 10 meters from the planned bus routes. In some cases, buses seem to drive on the road, but do not follow a planned route.







	< 10 m from KV1 routes	10 - 20 m from KV1 routes	> 20 m from KV1 routes
Points	44,477	1,718	2,648
% of total	91.1%	3.5%	5.4%

Figure 5.14 Bus location distance to planned routes (Arriva)

Some of the unexplainable points created by the handheld devices still are included in the datasets. These are points within 10 meters from the road network, but not near planned routes. In some cases, the distances are several kilometers away from the planned routes.

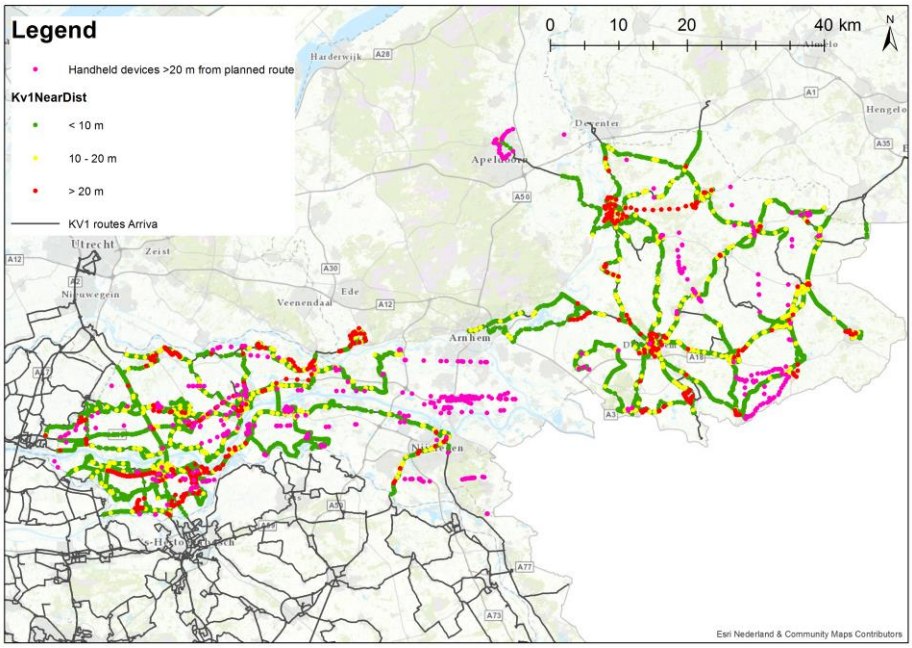


Figure 5.15 Arriva handheld devices too far from planned bus routes

Figure 5.15 shows the handheld devices further than 20 meters from the planned route in purple. Two patterns are reviewed below in figure 5.16. The figure on the left shows handheld devices clearly following a road, but not a planned route. By reviewing the schedule of this bus line it might become clear why these patterns are visible. One bus in the morning drives directly from Aalten to Dinxperloo (line on the right), other rides are only carried out after a passenger calls. Therefore, it might be possible for a bus to drive a shorter route than planned because no other passengers called at later stops.

The situation on the right shows some of the remaining points apparently not related to a route. These points are not deleted in the previous step because they actually are within 10 meters from the road network.

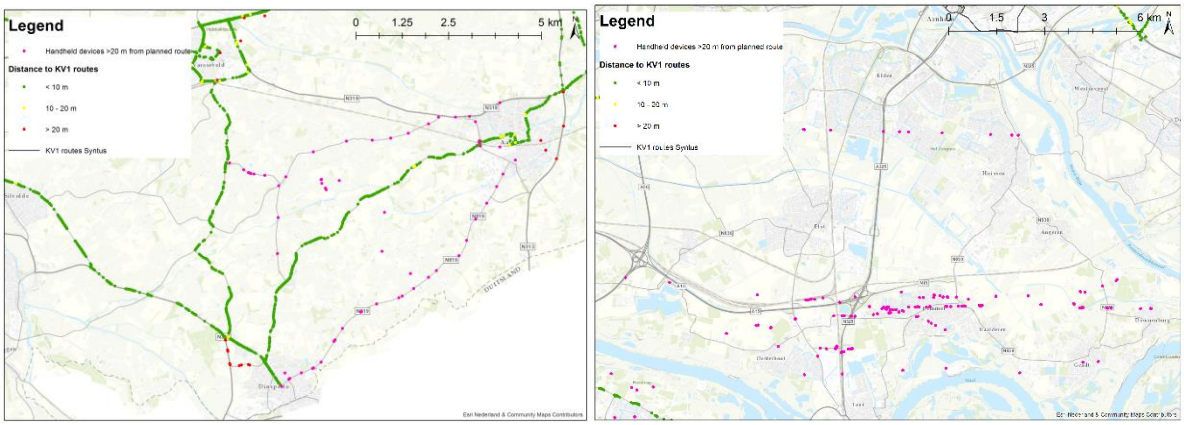
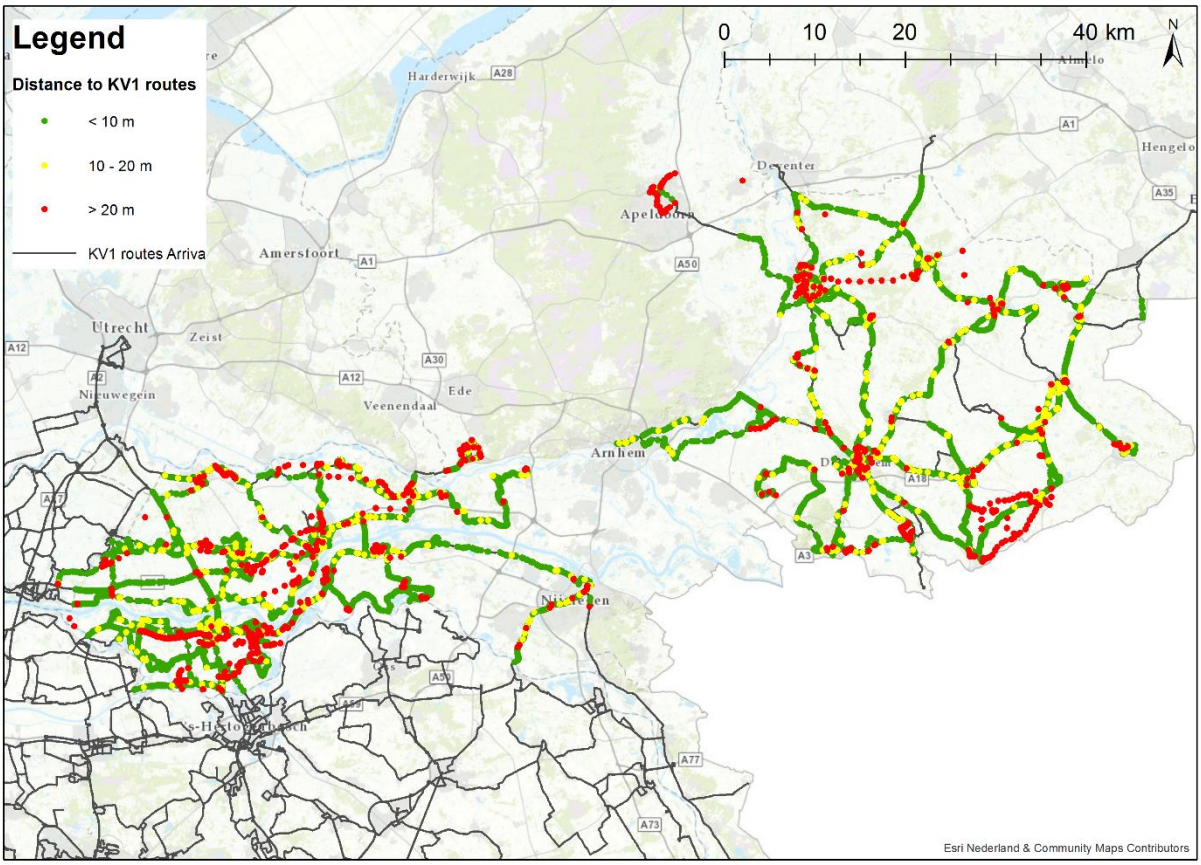
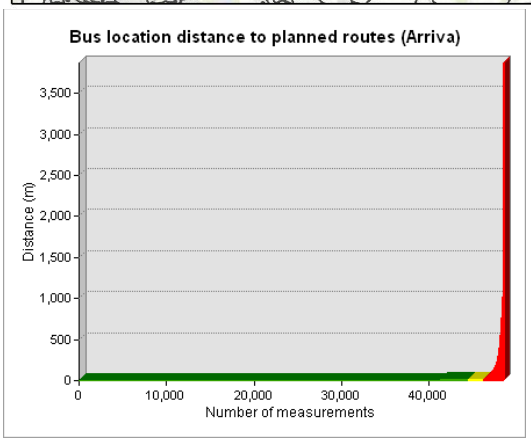
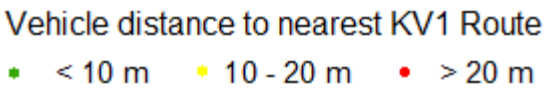


Figure 5.16 Arriva handheld devices too far from planned bus routes (details)

It was decided to manually delete the points which were clearly forming clusters or lines not related to a road network. Almost half of the handheld points was deleted: 413 out of 898 points were considered to be “ghost points”. Figure 5.17 shows the result after deleting these points.





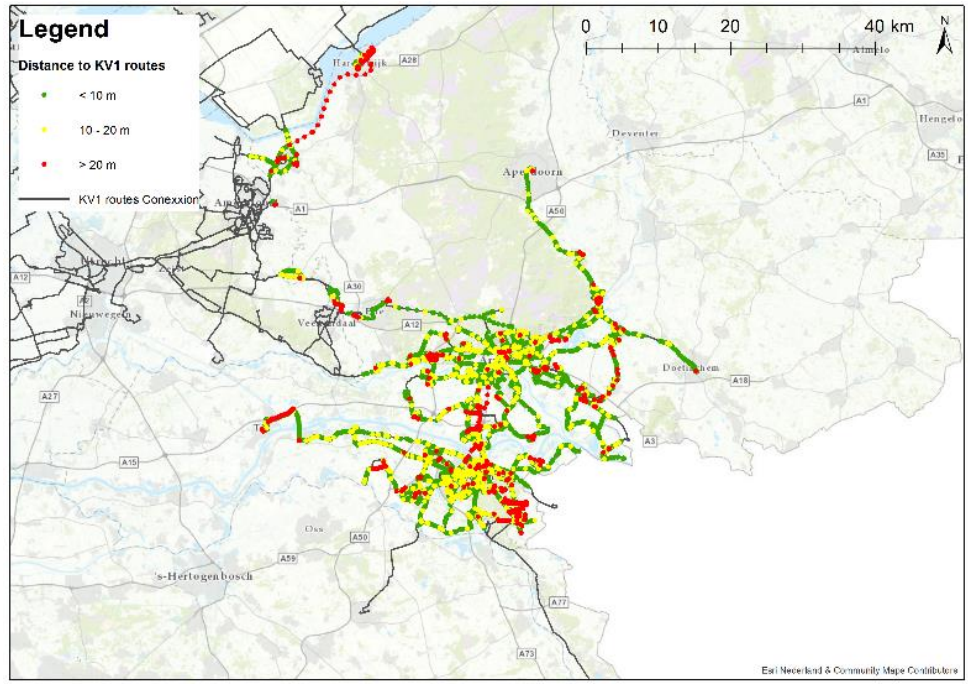


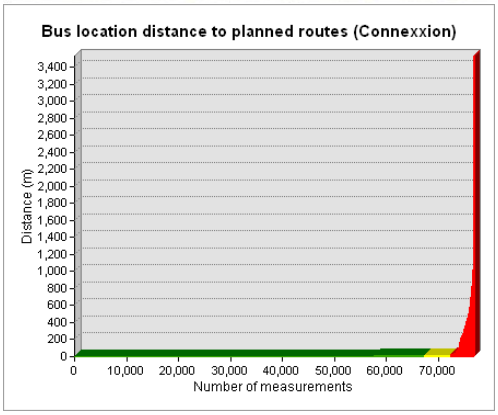
	< 10 m from KV1 routes	10 - 20 m from KV1 routes	> 20 m from KV1 routes
Points	44477	1718	2235
% of total	91.8%	3.5%	4.6%

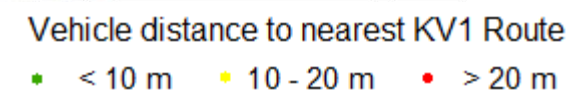
Figure 5.17 Arriva handheld devices too far from planned bus routes

Connexxion

The Connexxion dataset has some buses not driving on a planned route as well. Some of the buses are driving from or to a bus depot; other buses follow an alternative route. Figure 5.18 shows the point locations for Connexxion.







	< 10 m from KV1 routes	10 - 20 m from KV1 routes	> 20 m from KV1 routes
Points	67,529	5,082	4,926
% of total	87.1%	6.6%	6.4%

Figure 5.18 Bus location distance to planned routes (Connexxion)

The number of points further than 20 meters from the planned routes is relatively high (more than 6 percent). In some cases, it's clear why the bus doesn't follow the planned route, for example, the left part of figure 5.19. The planned route is via the highway, but at some moments the bus followed a parallel road, probably to avoid a traffic jam.

In the situation on the right, the situation is not that obvious. In this situation, a lot of the buses drove an alternative route as well, but the reason is not clear at all in this case.

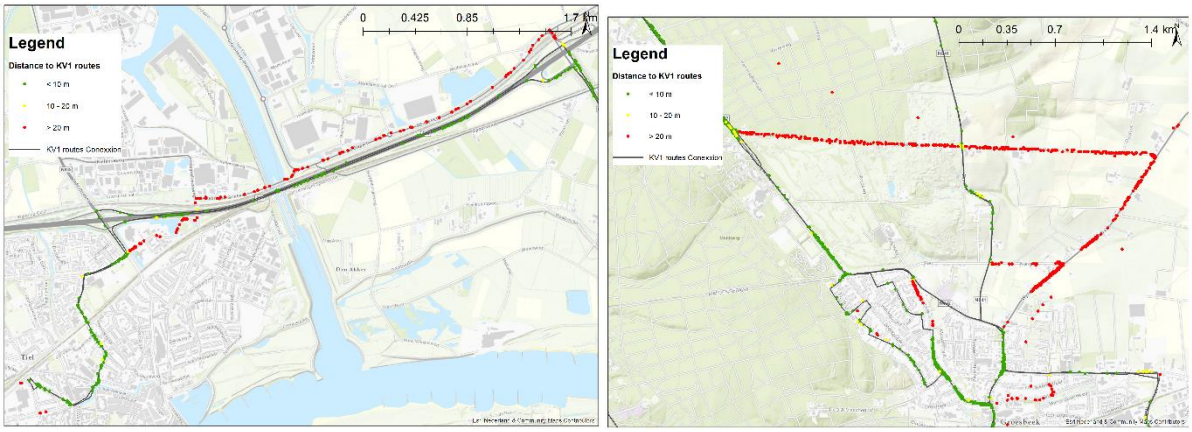
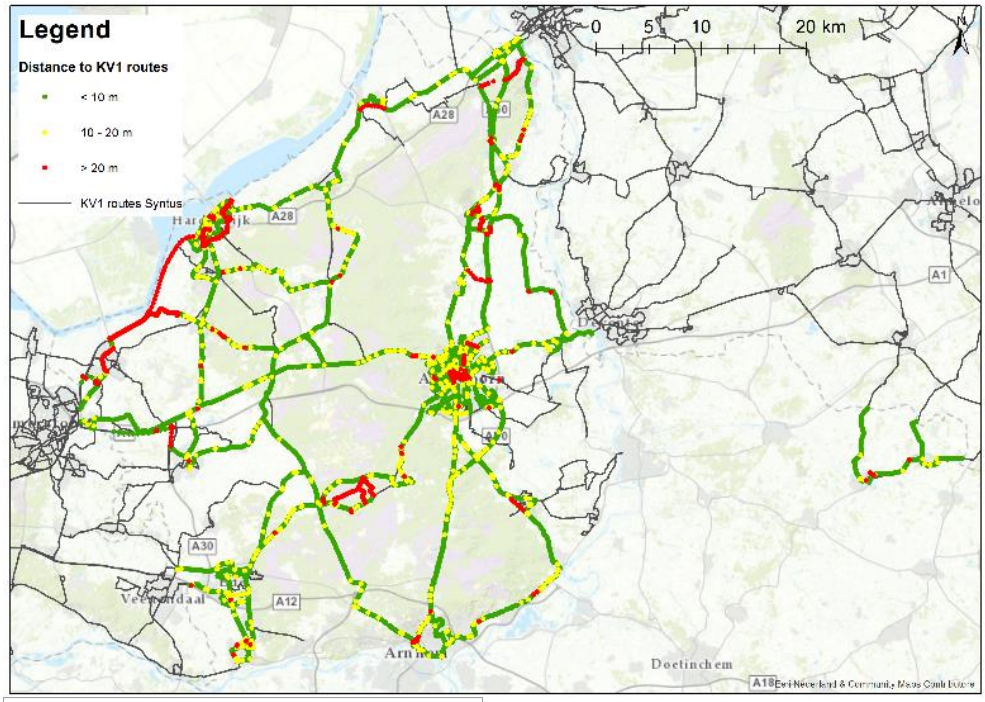
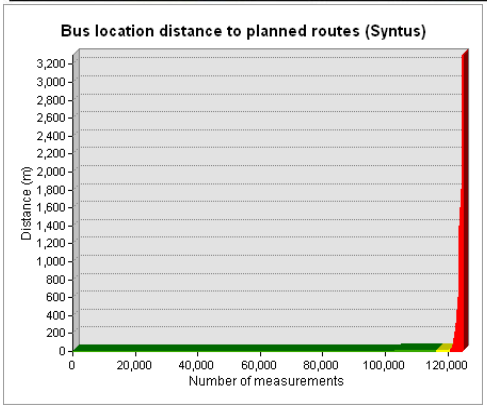


Figure 5.19 Bus location distance to planned routes (Connexxion, details)

Syntus

The Syntus dataset has the same issues as the Connexxion dataset. At several locations, the bus drives too far from the road for different reasons. Figure 5.20 shows several occasions where similar situations occur, such as the road between Amersfoort and Harderwijk left on the map.





Vehicle distance to nearest KV1 Route

< 10 m 10 - 20 m > 20 m

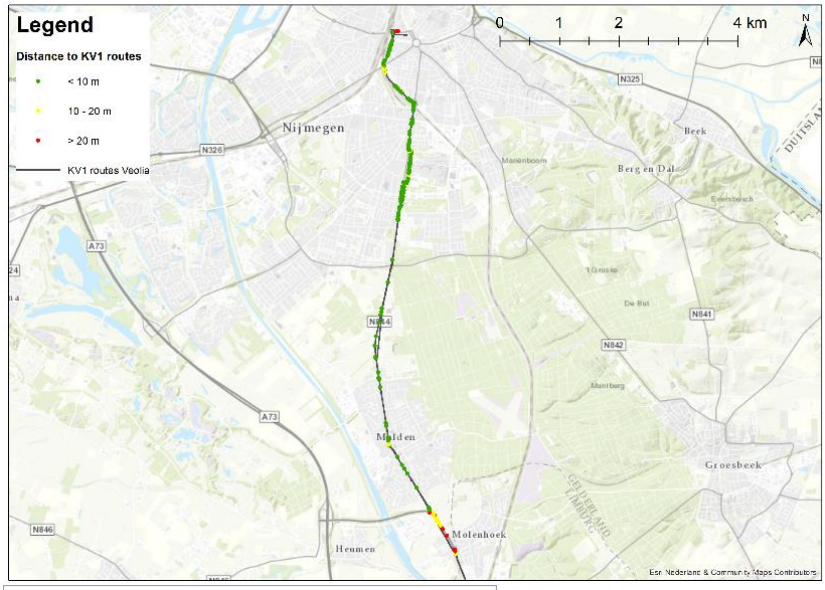
	< 10 m from KV1 routes	10 - 20 m from KV1 routes	> 20 m from KV1 routes
Points	116484	4410	4387
% of total	93.0%	3.5%	3.5%

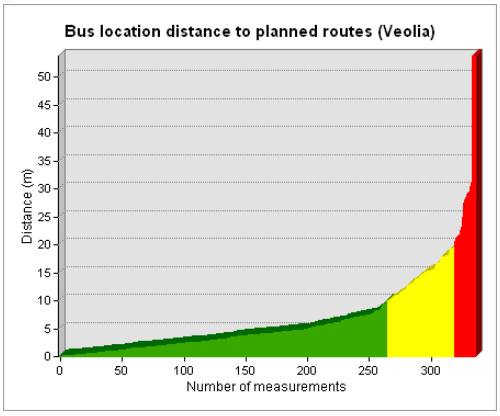
Figure 5.20 Bus location distance to planned routes (Syntus)

The number of points further than 10 meters from the road network is again relatively low, compared to the Arriva and Connexxion datasets.

Veolia

The Veolia dataset only has one bus line, but still, illustrates how the planned bus routes sometimes do not follow the roads very accurately. This causes a relatively large number of points further than 10 meters from the planned routes. Figure 5.21 shows the distribution of these points.





Vehicle distance to nearest KV1 Route

< 10 m 10 - 20 m > 20 m

	< 10 m from KV1 routes	10 - 20 m from KV1 routes	> 20 m from KV1 routes
Points	266	54	15
% of total	79.4%	16.1%	4.5%

Figure 5.21 Bus location distance to planned routes (Veolia)

5.2.2 OffRoute messages

Every real-time bus location point is either an OnRoute or OffRoute message. An OffRoute message is sent when, according to the transport operator, the bus is not following the planned route. Ideally, all the points further than 10 meters from the planned routes should be OffRoute points. Filtering out the OffRoute messages might explain cases where specific points are too far from the planned route.

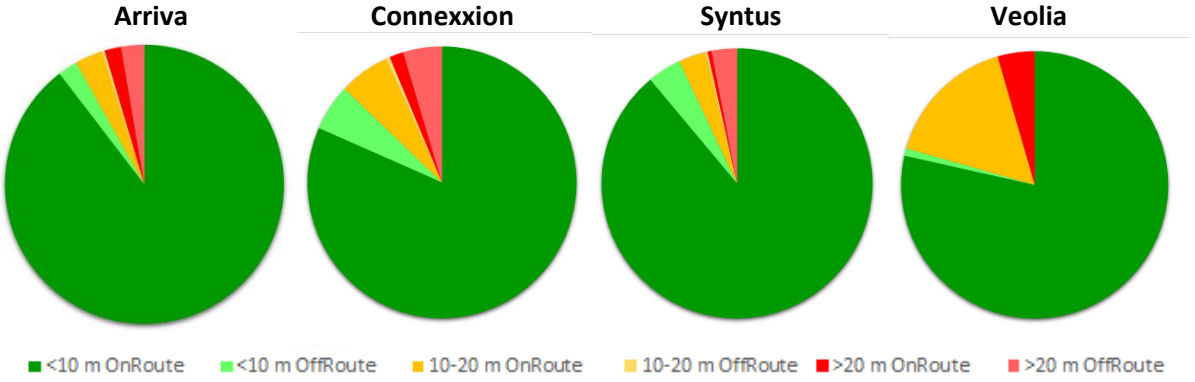


Figure 5.22 OffRoute and OnRoute messages per transport operator.

Table 5.2 shows the number of points within each category and the percentage in each category per transport operator.

	<10 m OnRoute		<10 m OffRoute		10-20 m OnRoute		10-20 m OffRoute		>20 m OnRoute		>20 m OffRoute	
Arriva	43384	90%	1093	2%	1534	3%	184	0%	930	2%	1305	3%
Connexxion	63230	82%	4299	6%	4697	6%	385	0%	1309	2%	3617	5%
Syntus	111424	89%	5060	4%	4053	3%	357	0%	607	0%	3780	3%
Veolia	263	79%	3	1%	54	16%	0	0%	15	4%	0	0%

Table 5.2 OnRoute and OffRoute points per transport operator

80-90 percent of all the points is within 10 meters from the road network and is an OnRoute message. 3-5 percent of the points is an OffRoute message and too far from the route network as well. There are however relatively large amounts points which are too far from a planned route, but which are still OnRoute message. Also, there are points which are within 10 meters from a planned bus route, but an OffRoute message. This case could, however, be explained by the fact that not the distance to a specific bus route is calculated, but the distance to the route network as a whole. For example, a bus driving a diversion on bus line 1, could still be within an acceptable distance from the route of bus line 2.

The maps below illustrate the locations where OnRoute and OffRoute points appeared, per transport operator. Arriva has clearly points categorized as OnRoute points at locations where no route is present in the planned bus route dataset, especially on the right side of the image (figure 5.23).

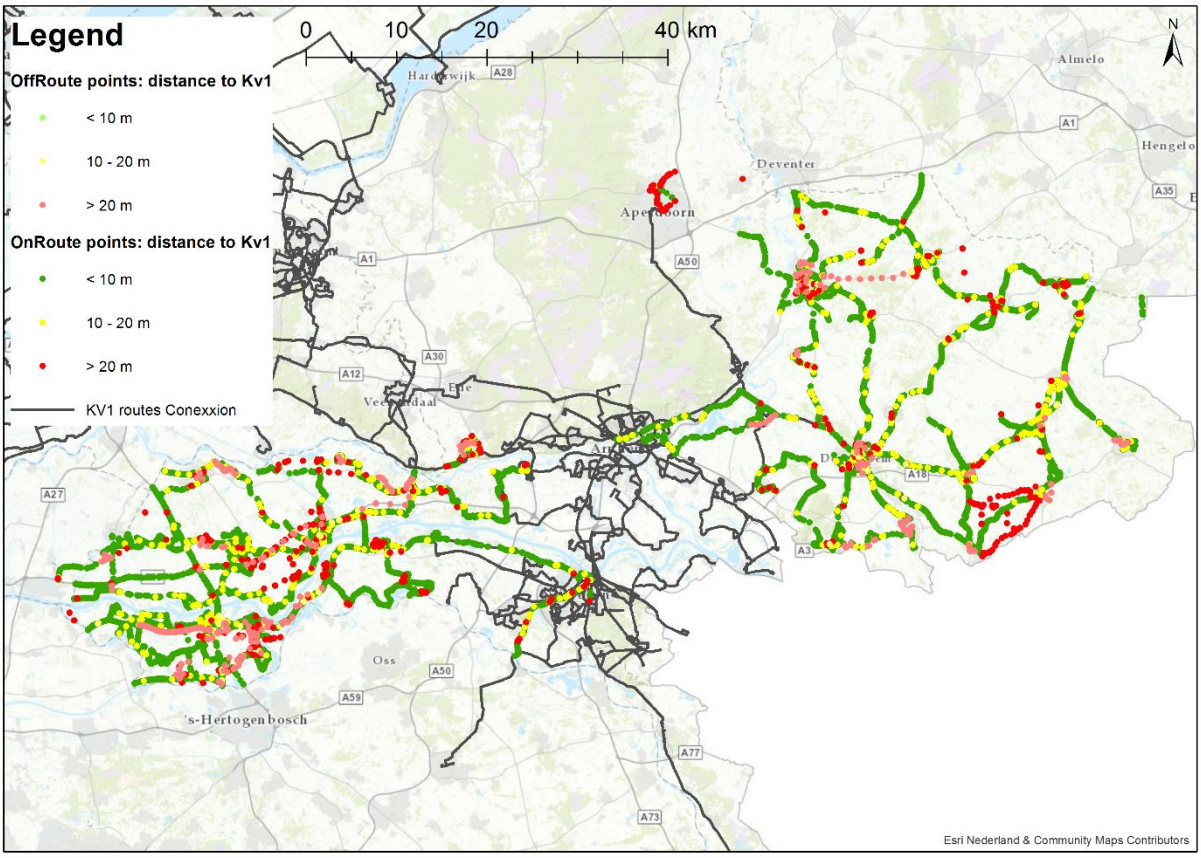


Figure 5.23: OffRoute and OnRoute point distance to KV1 (Arriva)

The next figures illustrate where points are too far from the bus route network. For example, in both the Connexxion (Figure 5.24) and Syntus (Figure 5.25) datasets buses drove alternative routes between Amersfoort and Harderwijk, not via a planned bus route.

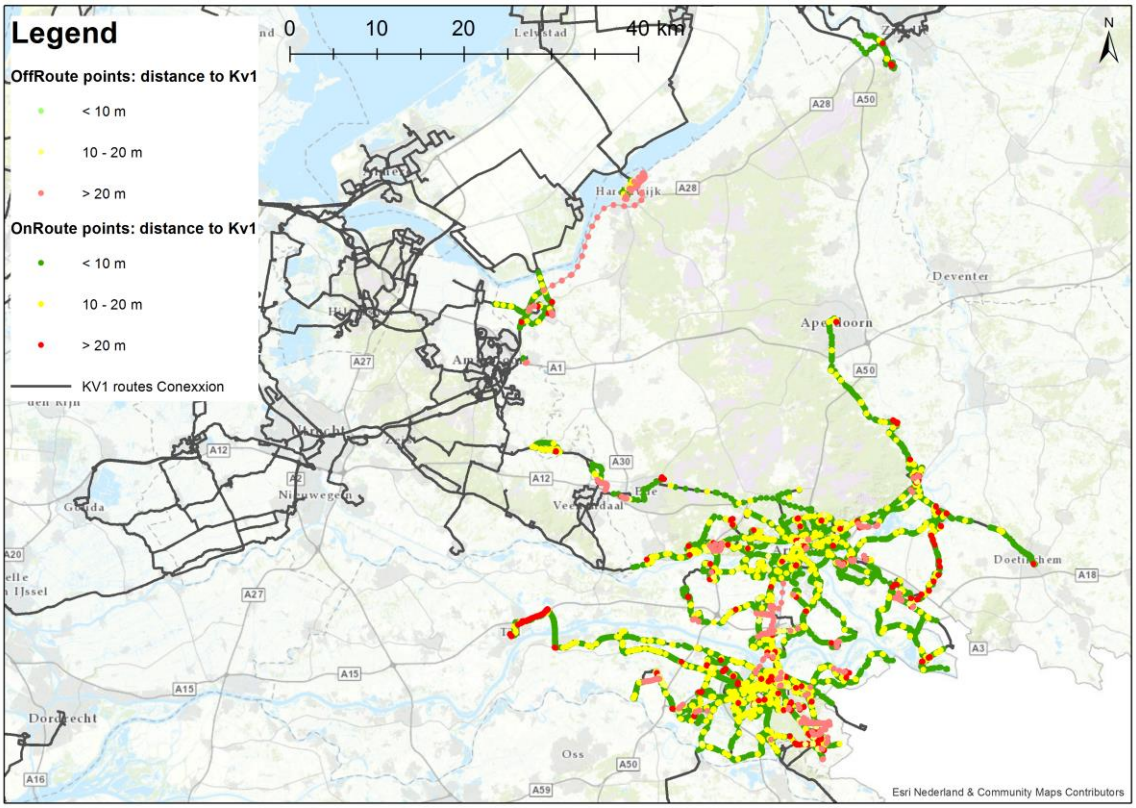


Figure 5.24: OffRoute and OnRoute point distance to KV1 (Connexxion)

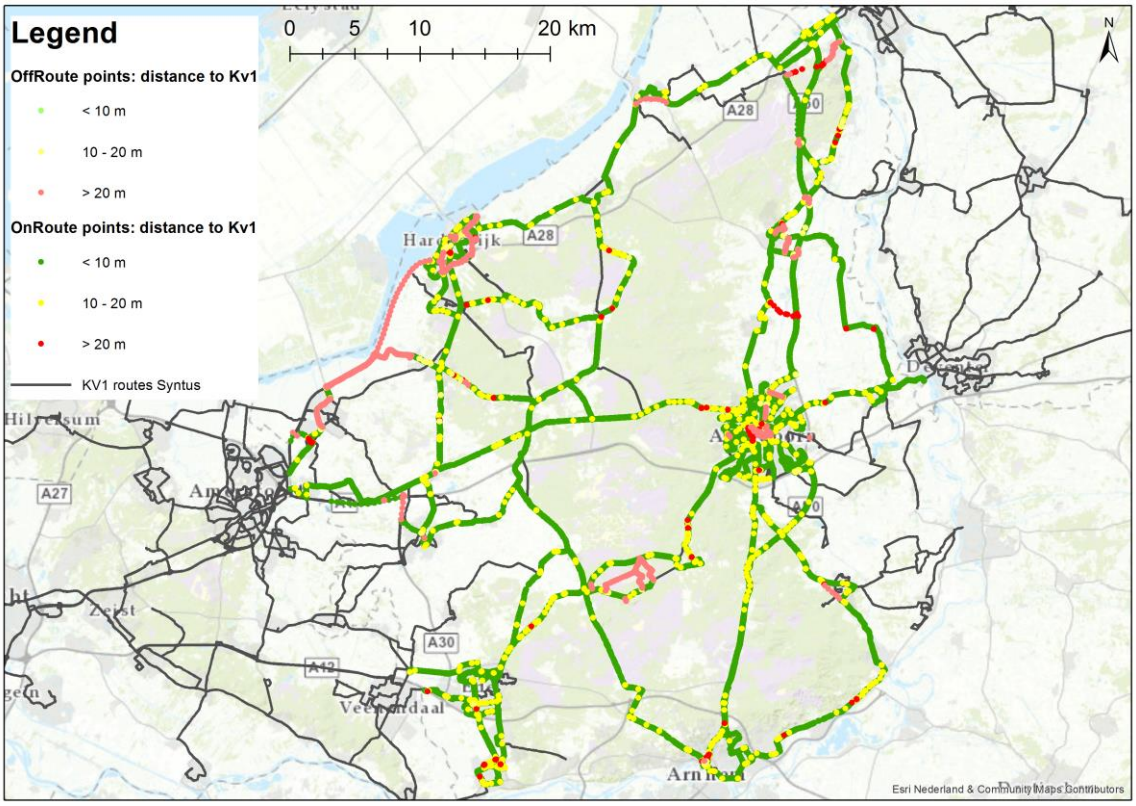


Figure 5.25: OffRoute and OnRoute point distance to KV1 (Syntus)

For the next step, it is decided to keep the OffRoute points in the dataset. One reason is the fact that the points do not seem to be very accurately classified as OnRoute or OffRoute all the time, another reason is to make a fair comparison between both the distance to both KV1 and OpenStreetMap dataset possible.

5.2.3 Differences between NDOV desks

Two parties function as NDOV portal: the REISinformatiegroep and OpenOV. It is concluded that both portals get exactly the same datasets delivered by the transport operators. Only the accessibility from both portals is slightly different. To access the OpenOV portal a form with a signature had to be signed, the REISinformatiegroep portal could be accessed without this. The structure of the portals is also different, but the accessibility is comparable. It would be a personal choice which of the portals works the best for every individual user.

5.3 Analysis of the OpenStreetMap dataset

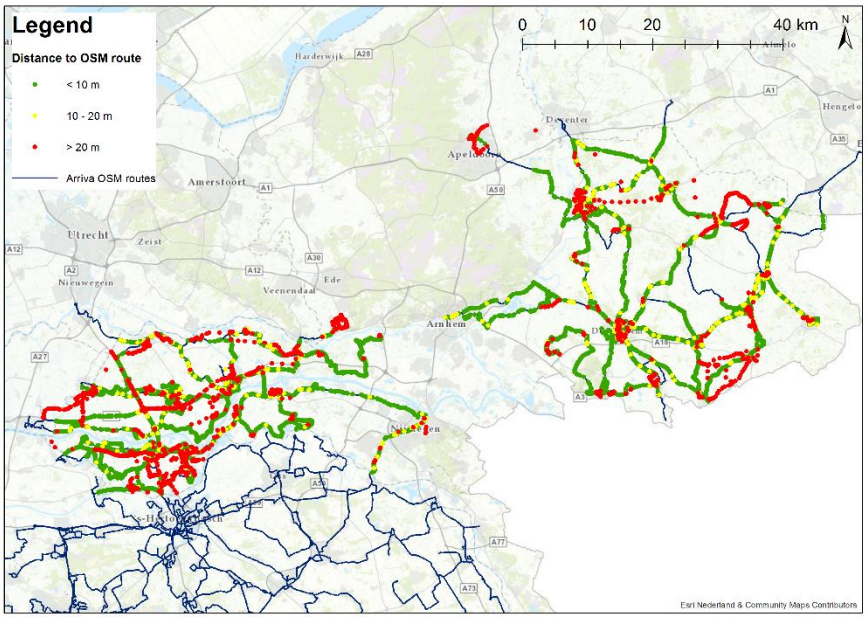
The last step is to assess the quality of the OpenStreetMap public transport routes. The distance from the real-time bus locations to the OpenStreetMap dataset will be calculated, similar to the process in the previous chapter. Also, the OpenStreetMap and Interface 1 datasets will be overlaid to find locations where they are different from each other.

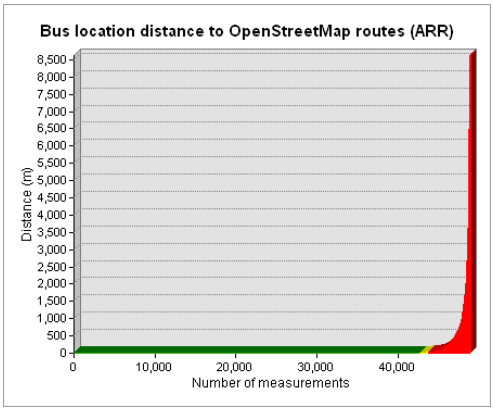
5.3.1 Distance to OpenStreetMap routes

The distance between the real-time bus locations and the OpenStreetMap routes in the same way as the distance to the planned bus routes in the previous step. The distance is calculated per transport operator, for both the OnRoute and OffRoute points.

Arriva

The Arriva dataset has at some points issues with the distance to the OpenStreetMap bus routes. The locations where the distance is too large are partially the same is the locations where this happened in the Interface 1 assessment. There are however some locations as well where the distance is too large but was not too large in the Interface 1 dataset. This will be studied in more detail at the end of this chapter.





Vehicle distance to nearest OSM route

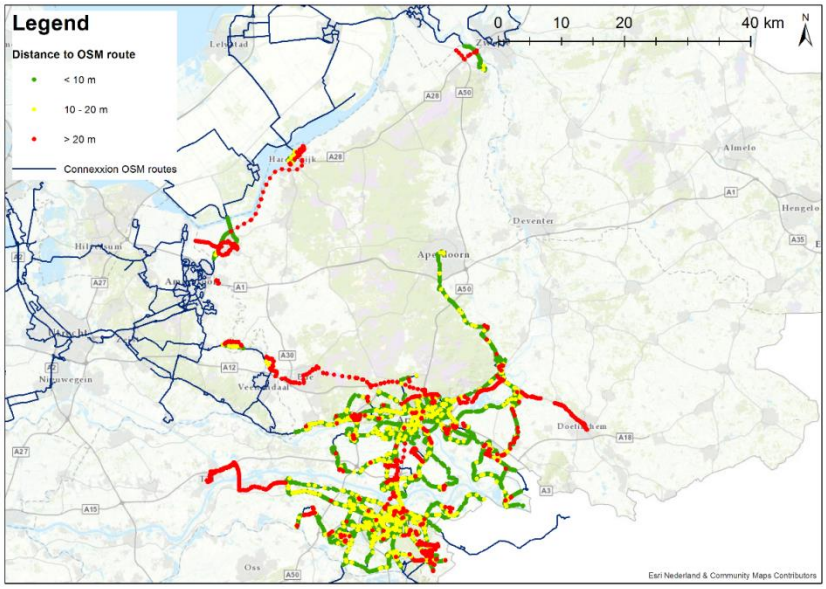
< 10 m 10 - 20 m > 20 m

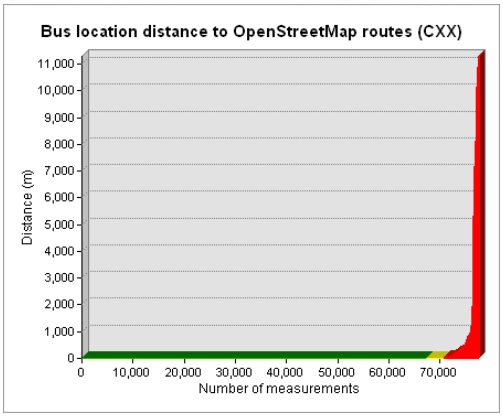
	< 10 m from OSM routes	10 - 20 m from OSM routes	> 20 m from OSM routes
Points	42736	1011	4683
% of total	88.2%	2.1%	9.7%

Figure 5.26 Bus location distance to OpenStreetMap routes (Arriva)

Connexxion

For the Connexxion dataset, the same patterns are visible compared to the Arriva dataset. At several locations, the distance to the OpenStreetMap dataset is larger than 20 meters. However, the number of points within 10 meters from the dataset is still 92 percent, which is relatively high.





Vehicle distance to nearest OSM route

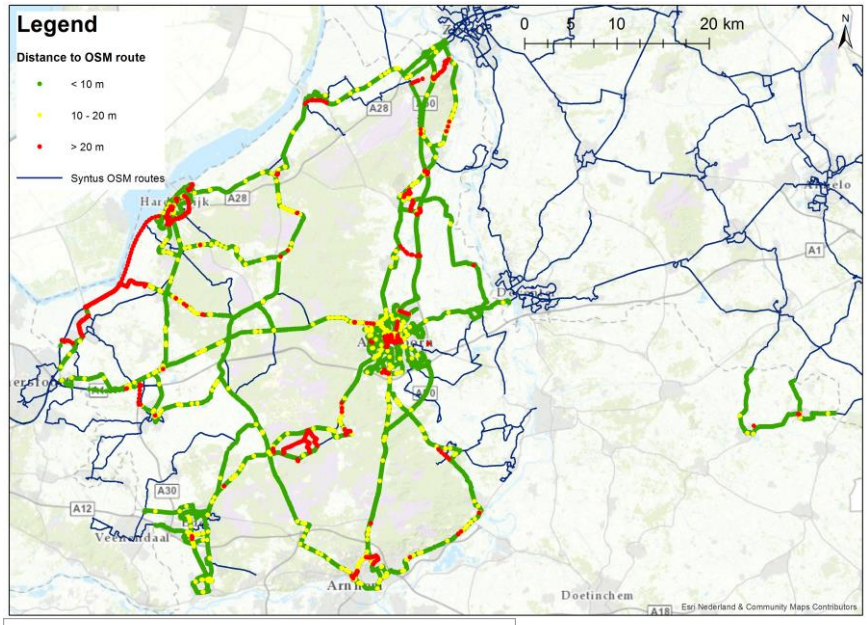
< 10 m 10 - 20 m > 20 m

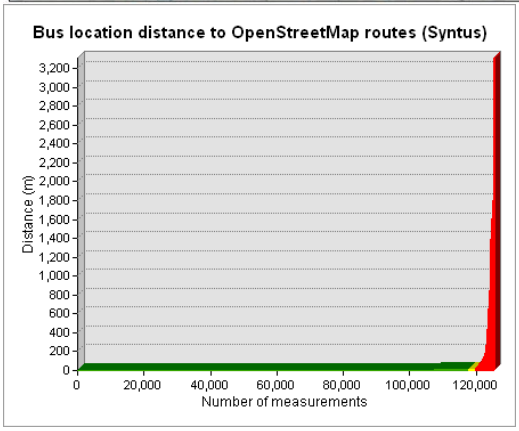
	< 10 m from OSM routes	10 - 20 m from OSM routes	> 20 m from OSM routes
Points	67282	3577	6678
% of total	91.9%	4.9%	9.1%

Figure 5.27 Bus location distance to OpenStreetMap routes (Connexxion)

Syntus

The Syntus dataset seems to be the best dataset compared to OpenStreetMap bus routes. 97 percent is within 10 meters from the OpenStreetMap dataset, even more than the number of points that was within 10 meters from the Interface 1 dataset. The locations where the distance is too large seem similar to the locations where the distance to the Interface 1 dataset was too large.





Vehicle distance to nearest OSM route

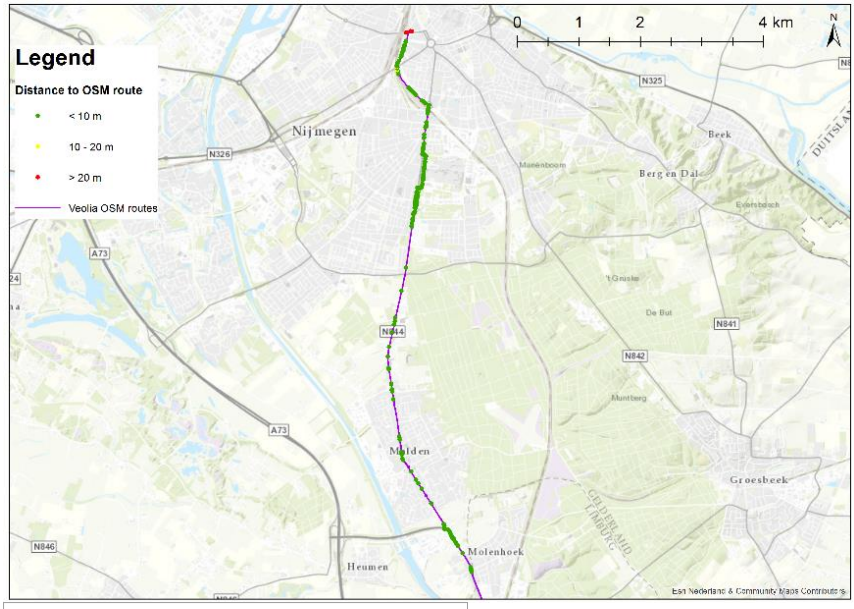
< 10 m 10 - 20 m > 20 m

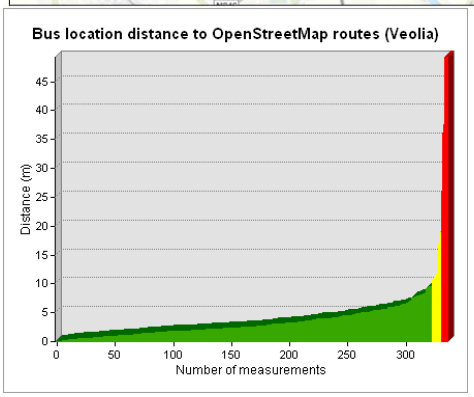
	< 10 m from OSM routes	10 - 20 m from OSM routes	> 20 m from OSM routes
Points	117651	2095	5535
% of total	97.1%	1.7%	4.6%

Figure 5.28 Bus location distance to OpenStreetMap routes (Syntus)

Veolia

The Veolia dataset only has a few points too far from the OpenStreetMap bus route near the bus station in Nijmegen.





Vehicle distance to nearest OSM route

< 10 m 10 - 20 m > 20 m

	< 10 m from OSM routes	10 - 20 m from OSM routes	> 20 m from OSM routes
Points	323	8	4
% of total	96.4%	2.4%	1.2%

Figure 5.29 Bus location distance to OpenStreetMap routes (Veolia)

5.3.2 Difference between distance to Interface 1 and OpenStreetMap datasets

Overall, 90.5 percent of the real-time bus locations are within 10 meters from the OpenStreetMap dataset. 90.8 percent was within 10 meters of the Interface 1 dataset, so the results are very comparable (Table 5.3). The number of points further than 20 meters from the network differs more. Almost 7 percent of the points is further than 20 meters from the OpenStreetMap network, compared to 5 percent of the points further than 20 meters from the Interface 1 network.

	< 10 m from KV1 routes		10 - 20 m from KV1 routes		> 20 m from KV1 routes	
	Points	Percentage	Points	Percentage	Points	Percentage
KV 1	228756	90.8%	11264	4.5%	11976	4.8%
OSM	227992	90.5%	6691	2.7%	17313	6.9%

Table 5.3: Point distance to KV1 and OpenStreetMap bus routes

The results of both the distance to Interface 1 and OpenStreetMap bus routes are displayed in figure 5.30. Per transport operator, the distribution is displayed.
The distance from the points to both the KV1 and OSM network is the best for Syntus. The percentage of points within 10 meters from the OpenStreetMap network is even slightly larger compared to the distance to the KV1 network.

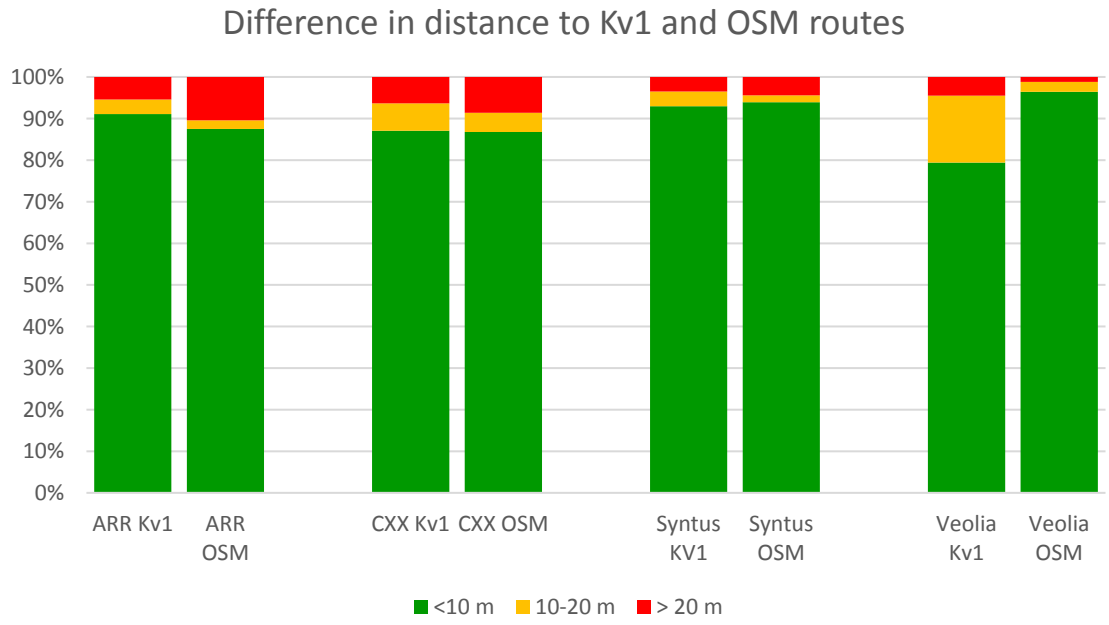


Figure 5.30: Difference in distance to Kv1 and OSM routes per transport operator

To find the locations where both datasets differ most, the bus routes will be overlaid in the next paragraph.

5.3.3 Overlay bus routes

To confirm and support the findings in the previous paragraph, another method is used to compare the bus routes. Buffers of 10 and 20 meters (acceptable distance to the road) are used on the Interface 1 dataset. For every OpenStreetMap bus route, it is calculated if the bus route is within these buffers (Table 5.4).

	Arriva	Connexxion	Syntus	Veolia
Within 10 meter buffer	95.5%	97.8%	97.8%	91.8%
Within 20 meter buffer	0.9%	1.1%	0.6%	3.0%
Not in buffer	3.6%	1.1%	1.6%	5.2%

Table 5.4: Overlap between KV1 and OpenStreetMap dataset per transport operator

There is a 98 percent overlap for the Connexxion and Syntus datasets. The Arriva dataset has a little less overlap. Figure 5.31 shows for every transport operator the OSM bus routes with the distance to the nearest KV1 bus route with the KV1 routes in blue displayed in the background.

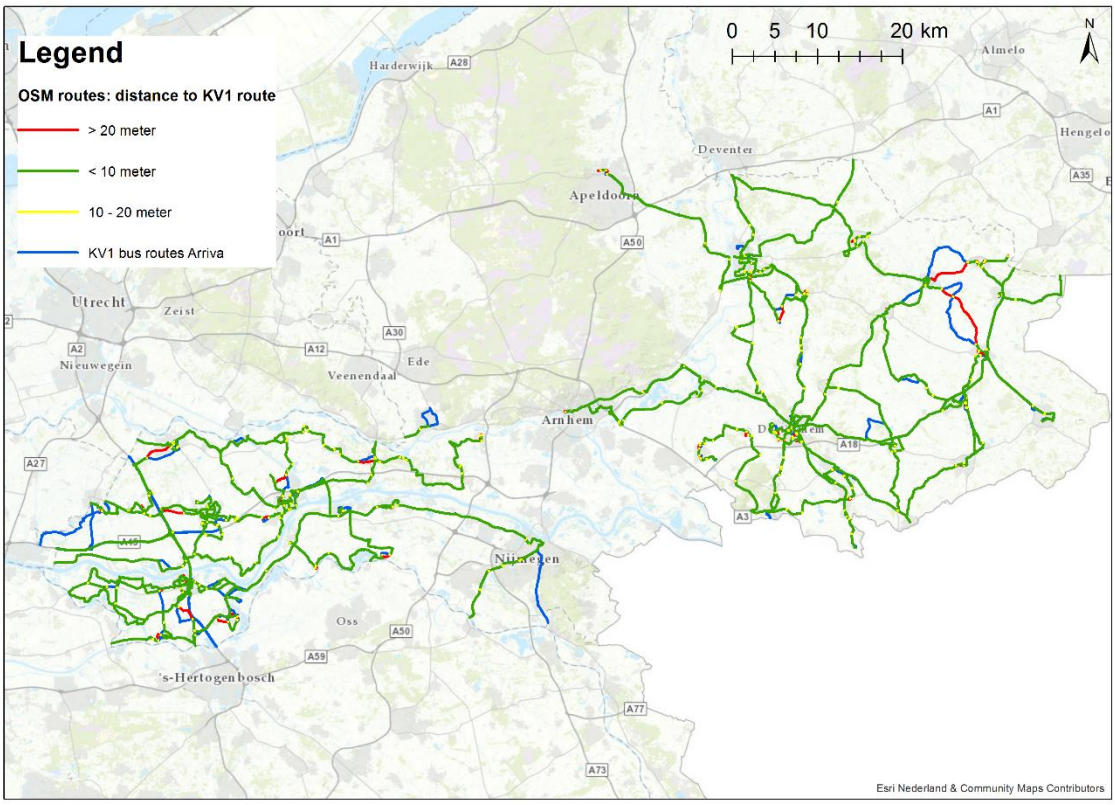


Figure 5.31 OpenStreetMap bus route distance to KV1 routes (Arriva)

The OpenStreetMap dataset for Arriva has some outdated bus lines, especially in the eastern part of the map. Apparently, a new bus route (in blue) replaced the old one (in red). The old routes are still included in OpenStreetMap and not updated yet.

For the Connexxion dataset some comparable results are visible, but overall the dataset is better up-to-date (Figure 5.32). Only at a few locations, the datasets shows differences.

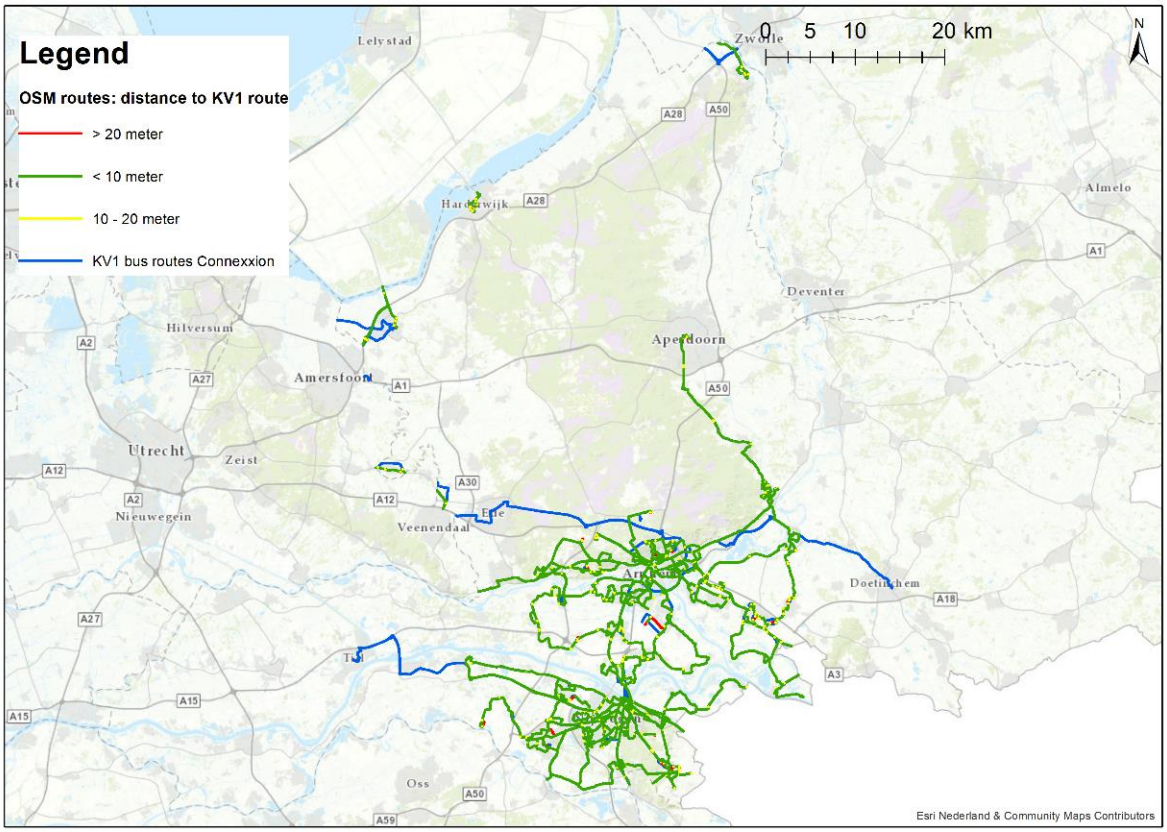


Figure 5.32 OpenStreetMap bus route distance to KV1 routes (Connexxion)

The Syntus datasets match well, with an exception for a few bus routes which are not in the KV1 dataset (figure 5.33).The most remarkable (longest) one appears to be a school bus which does not follow this route anymore since August 2016.

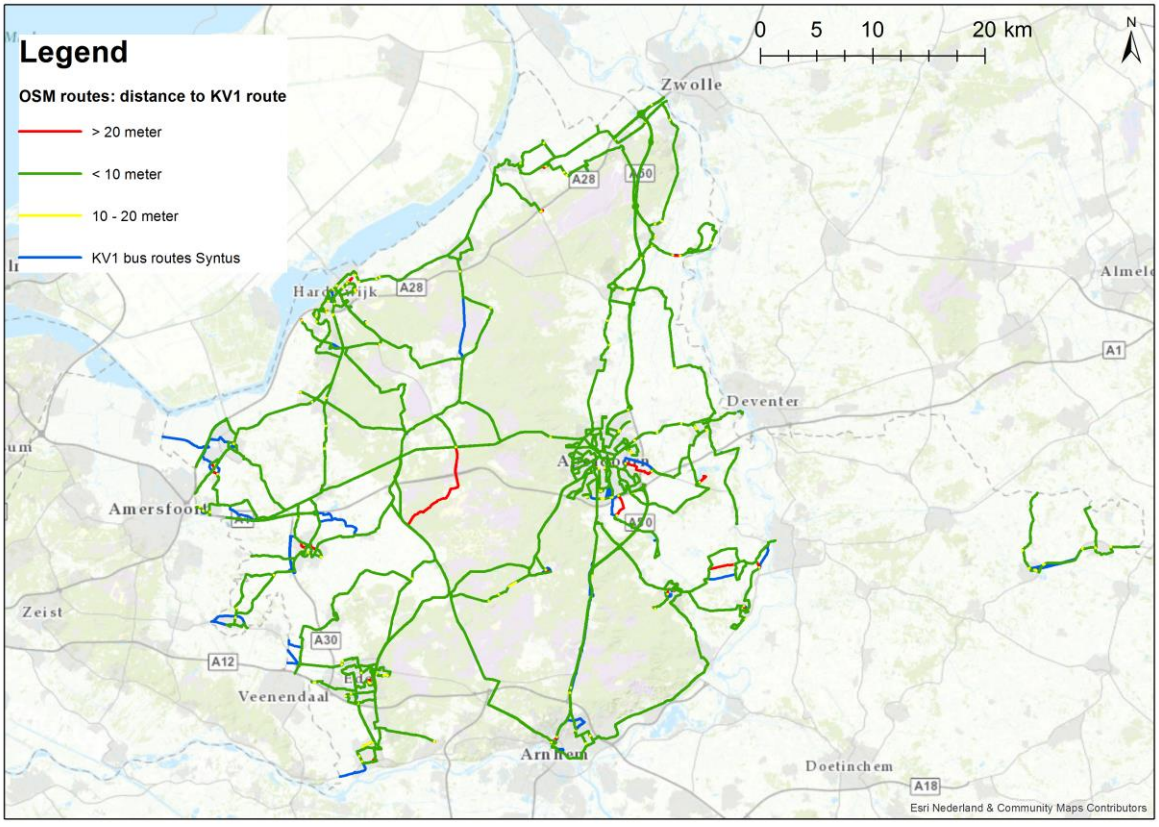


Figure 5.33 OpenStreetMap bus route distance to KV1 routes (Syntus)

Finally, figure 5.34 shows the overlaid bus routes for Veolia. This is only one bus route with only small differences between both datasets at few locations.

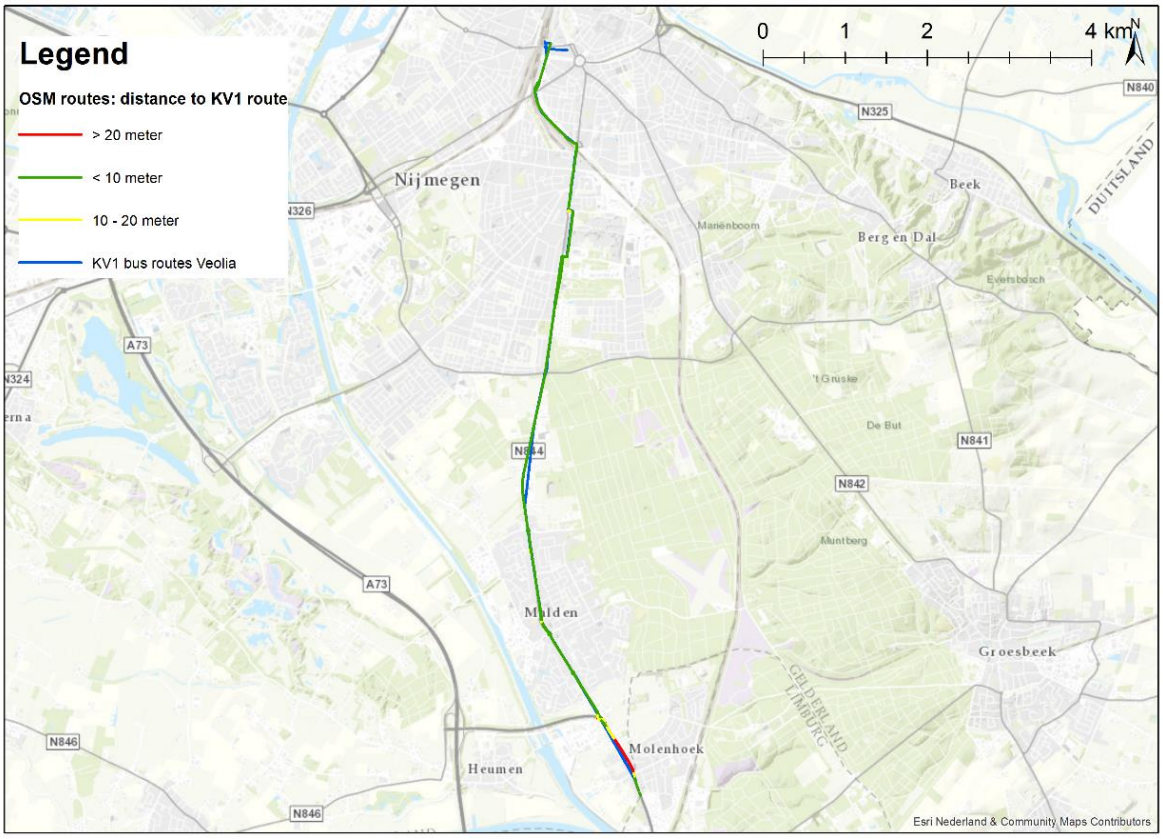


Figure 5.34 OpenStreetMap bus route distance to KV1 routes (Veolia)

6. Discussion

This chapter will review the methodology and results presented in the previous chapters. Some limitations will be addressed and the research will be placed in perspective to other studies on this topic.

6.1 Methodology

In this study, four objectives were formulated. The first objective was to understand the OpenStreetMap and NDOV data structures. After this, three objectives were formulated to assess the quality of the OpenStreetMap dataset. The approach was to first assess the quality of a reference dataset (real-time bus locations). Next, the quality of a dataset with the bus routes delivered by the transport operators was assessed. Finally, the quality of a dataset with the OpenStreetMap bus routes was assessed. This methodology seems to be the most accurate way to assess the spatial accuracy of the bus route data delivered by both the transport operators and OpenStreetMap.

Because of the limited time frame, the study focused only on the spatial accuracy of the dataset. Other elements of spatial data quality have not been tested. It would be interesting to see how quickly change over time is embraced, for example.

To test the temporal quality, the analysis can be repeated after a month with a new dataset, to see if the outcomes are comparable. Also, it would be interesting to see how long it takes before changes in the schedule are incorporated at OpenStreetMap.

The completeness and thematic accuracy of the dataset can be tested by performing statistical analysis on fields as bus line number, route etc. The used dataset had some issues in the comparable field for the same bus route. For example, for one bus route the OpenStreetMap dataset has elements name="Bus 45 Tiel Station - Wageningen Busstation" and ref="45". The same route in the KV1 dataset has elements route_short_name="45" and route_long_name="Tiel - Wageningen". However, the KV6 dataset with real-time bus locations only has a “LinePlanningNumber” element, which does not directly match the bus route number. This makes it difficult to compare these elements and the thematic accuracy as such.

The logical consistency of the dataset could be tested via network analysis. The dataset should not have gaps and should follow logical rules. This study was just a first step into the quality of routes in OpenStreetMap. However, when the dataset is going to be used in applications for bus route navigation, the logical consistency should be good as well.

6.2 Collection of the datasets

Three datasets were collected. The dataset with real-time bus locations and dataset with planned bus routes delivered by the transport operators were collected via the NDOV desks. To collect these two datasets, an agreement had to be signed with the OpenOV NDOV desk. The REISinformatiegroep desk can be accessed without sending a signed form. However, the contact with the OpenOV desks was more personal. Questions were answered quickly and a discussion group was available as well. The contact with the REISinformatiegroep was more formal.

The collection of the datasets was straightforward. With the right tools, it was easy to collect the real-time bus locations via Interface 6.
The creation of the Interface 1 dataset was the most difficult of the three datasets. The dataset was created using the GTFS data. This is a direct derivative of the original Interface 1 dataset, but it is not the original data. It is possible that using the original data results in a slightly different dataset.

The OpenStreetMap dataset was easy to access and collect as well. The followed method was the easiest and most convenient way to create the desired dataset.
Overall, the high accessibility of the datasets is a good sign. They are all open datasets and the high accessibility is one of the preconditions.

6.3 Real-time bus location quality

The quality assessment started with assessing the spatial accuracy of the real-time bus locations. Two problems occurred: the reference road dataset was not entirely up-to-date and there were some issues with handheld devices which formed strange routes.

6.3.1 Alternatives for TOP10NL dataset

Because of the problems with outdated roads in the TOP10NL dataset, a comparison was made with the distance to another open data road network. The National Road Dataset (Nationaal Wegenbestand, abbreviated as NWB) consists of all the roads in the Netherlands and can be downloaded as vector dataset.
The distance to this road network is calculated in the same way as done previously for the TOP10NL dataset (Table 6.1).

	Number of points	Average distance to TOP10NL roads	Average distance to NWB roads
Arriva	48269	2.88	3.68
Connexxion	76829	3.48	3.88
Syntus	124606	2.57	3.33
Veolia	335	2.60	3.08

Table 6.1: Distance to TOP10NL vs. NWB road network

The average distance from the points in the real-time bus location dataset to this NWB road network is for all the transport operators higher than the distance to TOP10NL roads.

Although the NWB dataset seems to be more up to date on certain occasions, there are also disadvantages. The NWB dataset does not have bus lanes included and is less detailed in other occasions. It can be concluded that the NWB dataset does not have enough advantages to be used as reference dataset instead of the TOP10NL dataset.

6.3.2 Handheld devices

In this study, handheld devices which were further than 10 meters from the road network were filtered out. It is not clear why these points existed in the dataset. Some of the followed vehicles followed a route very well, before starting to follow a “ghost route”, independent from the road network. It looked like the north or east coordinate stayed the same while the other coordinate continued according to the actually followed route. An explanation for this situation has not been found, but it’s clear that something is not going right.
Some of these handheld points were within 10 meters from the road network, but kilometers away from planned routes. They were clearly points where these buses had not been in reality and therefore they actually should be deleted from the dataset. The largest part of these points was deleted manually. A better way to do this was, of course, a rule to delete these points. However, it is difficult to formulate such a rule to delete these specific points. The chance that handheld devices which actually did form good routes are deleted, is too large.

6.4 Interface 1 bus route quality

The next step was to assess the quality of the Interface 1 data (bus routes delivered by transport operators). The number of points too far from the planned routes was relatively large compared to the number of points too far from the road network.
Some vehicles clearly followed routes which were not present in the dataset delivered by the transport operator. A possible reason is the presence of diversions. In most cases, a diversion is not included in the planned bus route dataset delivered via Interface 1. These buses driving another route than the planned route should, however, sent an OffRoute instead of an OnRoute message.
It was concluded that the real-time bus location dataset does not classify points as OnRoute or OffRoute points very accurately. It happens that points are included in the dataset as OnRoute points while they do not seem to follow a planned route at all. The classification of OffRoute and OnRoute messages is not set by the driver but happens automatically. If a bus does not follow the planned route it should automatically send an OffRoute message.

6.5 OpenStreetMap bus route quality

The quality of the OpenStreetMap bus routes appeared to be comparable to the quality of the dataset delivered by the transport operators. For transport operators Arriva and Connexxion, the OpenStreetMap route data is slightly less accurate compared to the KV1 data delivered by the transport operators. This was expected because OpenStreetMap data has to be kept up-to-date by a group of volunteers, while the NDOV data is published by the transport operator. Given this situation, the OpenStreetMap dataset is not bad at all: OpenStreetMap is by definition too late.
For Syntus, although the difference is small, more real-time bus locations are within 10 meters from the OpenStreetMap network than the Interface 1 network. A reason might be the fact that OpenStreetMap bus routes are built up using actual road segments already in OpenStreetMap, while Interface 1 data is delivered by transport operators and probably based on GPS tracks.

There were concerns about the quality of the OpenStreetMap bus routes because of the possibility for routes to break soon, but this does not seem to be the case. It is possible that on a micro level, the routes show small inconsistencies or gaps. In this study only spatial accuracy was taken into account. The dataset was used as is, test on topological correctness were not performed (connectivity issues for navigating e.g.). This would be interesting to investigate in further research.

6.6 Broader perspective

The quality of OpenStreetMap data was studied in previous research (Jackson et al. 2013; Mondzech & Sester 2011). It was concluded that the quality of OpenStreetMap was relatively high. Overall, the quality is better in urban areas compared to rural areas. This study tried to assess the quality of secondary data in OpenStreetMap, in this case, bus route data.
The results are in line with the results other research to the quality of OpenStreetMap. The quality of OpenStreetMap bus routes is comparable to the quality of the bus route data delivered by transport operators. However, there are some concerns about the up-to-date-ness of the bus route dataset. In a few cases, old bus routes which were already out of use for a couple of months were still in the OpenStreetMap dataset.
There wasn’t a clear relation with the level of urbanization. This is probably also because the Netherlands is a largely urbanized country.

7. Conclusions and recommendations

This chapter gives the main conclusions of the research per objective and some recommendations for future research.

7.1 Conclusions

Objective 1: To understand the OpenStreetMap and NDOV data structures.

The first objective was to study the data structures of OpenStreetMap and NDOV. The OpenStreetMap data model is structured in a straightforward way and relatively easy to understand. Three basic elements (nodes, ways, and relations) form the whole structure. This makes is much easier to understand than the NDOV data models. The main reason is that OpenStreetMap does not have data on timetables, operating hours and fares included.
The NDOV data model structures are much more complicated. A lot of situations can be handled. The OpenStreetMap dataset only has the option to display a route and optionally add some meta information.

Objective 2: To assess the quality of real-time bus locations delivered by transport operators.

The second objective was to assess the quality of the real-time bus locations by calculating the distance to a reference road network. The real-time bus location dataset does not only consist of buses but has trains included as well. Also, in some cases buses sent data while they are buffering near a bus station or in a bus depot.
Some vehicles appeared to be handheld devices and formed strange routes on the map, independently from the road network. A reason could not be found.
After removing the trains and handheld devices between 97-99 percent of the points was within 10 meters of the road network. In some cases, the reference road dataset (Top10NL) was not up-to-date enough and some roads seem to miss.

Objective 3: To assess the quality of public transport data delivered by transport operators via both NDOV desks (REISinformatiegroep and OpenOV).

The third objective was to assess the quality of the bus routes delivered by transport operators via NDOV Interface 1. On average, for all the operators, around 5 percent of the points is further than 20 meters from the planned bus routes. Compared to the number of points too far from the road network, this is a relatively large number.
In a few cases, buses followed routes which were not present in the dataset delivered by the transport operator. This can be because of diversions. This is only partially explained by reviewing differences between OnRoute and OffRoute points. This classification does not seem to be used very accurately by transport operators all the time.
The public transport data delivered by both NDOV desks of REISinformatiegroep and OpenOV were exactly the same, consequently, no differences in quality were found.

Objective 4: To assess the quality of public transport data in OpenStreetMap.

The last objective was to assess the quality of the OpenStreetMap bus routes. The quality of the OpenStreetMap dataset is comparable to the quality of the NDOV dataset. 90.5 percent of the points was within 10 meters from an OpenStreetMap bus route, compared to 90.8 percent within 10 meters from an Interface 1 bus route. The percentage of points further than 20 meters from the network was 6.9 percent from the OpenStreetMap bus routes versus 4.8 percent from the Interface 1 bus routes, making the Interface 1 dataset slightly better. For one of the transport operators however, the OpenStreetMap dataset was slightly better.

Overall the Interface 1 dataset was slightly better than the OpenStreetMap bus route dataset. This is not surprising because the Interface 1 data is delivered by the transport operators themselves. OpenStreetMap always lags behind, because it's collected and processed by volunteers. Keeping this in mind, the quality of the OpenStreetMap bus routes does not differ much from the Interface 1 bus routes.

7.2 Recommendations

This study focused on the spatial accuracy of the bus route data. Other elements of spatial data quality were beyond the scope of this study. However, it would be very interesting to study other elements, especially temporal quality. How fast does the OpenStreetMap data adapt to a new schedule with changed bus routes, for example. Other elements which could be tested are completeness, connectivity of the network and logical consistency of the different datasets.

Fitness for use is also an interesting element to keep in mind for further research. Formulating an use case and requirements for this specific use case could give new insights in the quality of the different datasets providing public transport route information.

Future research could also study the quality of other types of route data than public transport data in OpenStreetMap, for example walking and bicycle routes, which are included in OpenStreetMap in the same way as public transport data. It would also be interesting to compare the OpenStreetMap public transport routes for different regions, countries, or types of transport (bus, train, tram etc.).

References

Attard, J. et al., 2015. A Systematic Review of Open Government Data Initiatives. *Government Information Quarterly*. Available at: http://dx.doi.org/10.5281/zenodo.18592.

Barron, C., Neis, P. & Zipf, A., 2014. A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. , 18(6), pp.877–895.

Bennett, J., 2010. *OpenStreetMap*, Packt Publishing.

BISON, 2015a. Specificatie TMI8 Actuele ritpunctualiteit en voertuiginformatie. , (kv 6), pp.1–76.

BISON, 2015b. Specificatie TMI8 Dienstregeling Koppelvlak 1. , (november).

Girres, J.F. & Touya, G., 2010. Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS*, 14(4), pp.435–459.

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4), pp.211–221. Available at: http://www.springerlink.com/index/10.1007/s10708-007-9111-y.

Hadas, Y., 2013. Assessing public transport systems connectivity based on Google Transit data. *Journal of Transport Geography*, 33, pp.105–116. Available at: http://dx.doi.org/10.1016/j.jtrangeo.2013.09.015.

Haklay, M., 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. *Environment and Planning B: Planning and Design*, 37(4), pp.682–703.

Helbich, M., Amelunxen, C. & Neis, P., 2012. Comparative Spatial Analysis of Positional Accuracy of OpenStreetMap and Proprietary Geodata. *Proceedings of GI_Forum 2012: Geovisualization, Society and Learning*, pp.24–33. Available at: http://koenigstuhl.geog.uni-heidelberg.de/publications/2010/Helbich/Helbich_etal_AGILE2011.pdf.

Hochmair, H.H., Zielstra, D. & Neis, P., 2013. Assessing the Completeness of Bicycle Trail and Designated Lane Features in OpenStreetMap for the United States and Europe. *TRB 92nd Annual Meeting*, pp.1–21. Available at: http://trid.trb.org/view.aspx?id=1242969.

ISO, 2010. ISO 19157 Standard. Available at: https://www.iso.org/standard/32575.html.

Jackson, S. et al., 2013. Assessing Completeness and Spatial Error of Features in Volunteered Geographic Information. *ISPRS International Journal of Geo-Information*, 2(2), pp.507–530. Available at: http://www.mdpi.com/2220-9964/2/2/507/.

Keßler, C., Trame, J. & Kauppinen, T., 2011. Tracking editing processes in volunteered geographic information: The case of OpenStreetMap. *Conference on Spatial Information Theory: …*, pp.6–8. Available at: http://www.carsten.io/cosit11poster.pdf%5Cnhttp://www.carsten.io/iope2011.pdf.

Kounadi, O., 2009. *Assessing the quality of OpenStreetMap data*, Available at: ftp://ftp.cits.nrcan.gc.ca/pub/cartonat/Reference/VGI/Rania_OSM_dissertation.pdf.

Kulk, S. & van Loenen, B., 2012. Brave New Open Data World? *International Journal of Spatial Data Infrastructures Research*, 7, pp.196–206.

Longley, P.A. et al., 2010. *Geographic information systems & science*,

Ministerie van Infrastructuur en Milieu, 2013. *Stand van zaken project NDOV (Nationale Data Openbaar Vervoer)*,

Mondzech, J. & Sester, M., 2011. Quality analysis of OpenStreetMap data based on application needs. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 46(2), pp.115–125. Available at: http://dx.doi.org/10.3138/carto.46.2.115%5Cnhttp://utpjournals.metapress.com/content/d815571134382u31/?genre=article&id=doi%3A10.3138%2Fcarto.46.2.115.

Van Oort, P., 2006. *Spatial data quality: from description to application*, Available at: http://library.wur.nl/WebQuery/wdab/1788022.

OpenStreetmap Wiki, 2017. OpenStreetMap Wiki Relation:Route. Available at: http://wiki.openstreetmap.org/wiki/Relation:route [Accessed November 21, 2016].

Zandbergen, P.A., Ignizio, D.A. & Lenzer, K.E., 2011. Positional Accuracy of TIGER 2000 and 2009 Road Networks. *Transactions in GIS*, 15(4), pp.495–519.

Zielstra, D. & Zipf, A., 2010. A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. *13th AGILE International Conference on Geographic Information Science*, 1, pp.1–15. Available at: http://koenigstuhl.geog.uni-heidelberg.de/publications/2010/Zielstra/AGILE2010_Zielstra_Zipf_final5.pdf.

Appendix A: Interface 6 explanation

This Appendix describes how the Interface 6 format is structured. Every event comes with a message send by the vehicle. These messages are distributed in a XML format.

```
<?xml version="1.0" encoding="UTF-8"?>
<tmi8:VV_TM_PUSH
xmlns:tmi8c="http://bison.connekt.nl/tmi8/kv6/core"
xmlns:tmi8=”http://bison.connekt.nl/tmi8/kv6/msg”>
      <tmi8:SubsciberID> SUBSCRIBERID </tmi8:SubscriberID>
      <tmi8:Version> VERSION </tmi8:Version>
      <tmi8:DossierName> DOSSIERNAME </tmi8:DossierName>
      <tmi8:Timestamp> TIMESTAMP </tmi8:Timestamp>
      <tmi8:DOSSIER>[1]
            <tmi8:OBJECTNAME>[2]
                  RECORDDATA
                  RECORDEXTENSIE
            </tmi8:OBJECTNAME>
      </tmi8:DOSSIER>
</tmi8:VV_TM_PUSH>
```

[1] The XML can consist of as many dossiers as needed

[2] At this place in the XML, 1 or more messages are included. The possible messages are listed below, with some explanation.

Possible messages are Init, Arrival, Departure, Onstop, Offroute and Onroute.

Init

Vehicle gets a ride assigned.

```
<tmi8:INIT>
<tmi8:dataownercode>CXX</tmi8:dataownercode>
<tmi8:lineplanningnumber>L401</tmi8:lineplanningnumber>
<tmi8:operatingday>2016-03-10</tmi8:operatingday>
<tmi8:journeynumber>54</tmi8:journeynumber>
<tmi8:reinforcementnumber>0</tmi8:reinforcementnumber>
<tmi8:timestamp>2016-03-10T10:52:46+01:00</tmi8:timestamp>
<tmi8:source>VEHICLE</tmi8:source>
<tmi8:userstopcode>64308600</tmi8:userstopcode>
<tmi8:passagesequencenumber>0</tmi8:passagesequencenumber>
<tmi8:vehiclenumber>1201</tmi8:vehiclenumber>
<tmi8:blockcode>1100005</tmi8:blockcode>
<tmi8:wheelchairaccessible>UNKNOWN</tmi8:wheelchairaccessible>
<tmi8:numberofcoaches>1</tmi8:numberofcoaches>
</tmi8:INIT>
```

Arrival

A vehicle arrives at a bus stop.

```
<tmi8:ARRIVAL>
<tmi8:dataownercode>CXX</tmi8:dataownercode>
<tmi8:lineplanningnumber>M310</tmi8:lineplanningnumber>
<tmi8:operatingday>2016-03-10</tmi8:operatingday>
<tmi8:journeynumber>81</tmi8:journeynumber>
<tmi8:reinforcementnumber>0</tmi8:reinforcementnumber>
<tmi8:userstopcode>56430140</tmi8:userstopcode>
<tmi8:passagesequencenumber>0</tmi8:passagesequencenumber>
<tmi8:timestamp>2016-03-10T10:52:46+01:00</tmi8:timestamp>
<tmi8:source>VEHICLE</tmi8:source>
<tmi8:vehiclenumber>9193</tmi8:vehiclenumber>
<tmi8:punctuality>-73</tmi8:punctuality>
</tmi8:ARRIVAL>
```

Departure

A vehicle departs from a bus stop.

```
<tmi8:DEPARTURE>
<tmi8:dataownercode>CXX</tmi8:dataownercode>
<tmi8:lineplanningnumber>L014</tmi8:lineplanningnumber>
<tmi8:operatingday>2016-03-10</tmi8:operatingday>
<tmi8:journeynumber>25</tmi8:journeynumber>
<tmi8:reinforcementnumber>0</tmi8:reinforcementnumber>
<tmi8:userstopcode>64005160</tmi8:userstopcode>
<tmi8:passagesequencenumber>0</tmi8:passagesequencenumber>
<tmi8:timestamp>2016-03-10T10:52:46+01:00</tmi8:timestamp>
<tmi8:source>VEHICLE</tmi8:source>
<tmi8:vehiclenumber>3407</tmi8:vehiclenumber>
<tmi8:punctuality>47</tmi8:punctuality>
</tmi8:DEPARTURE>
```

Onstop

A vehicle is standing still at a bus stop.

```
<tmi8:ONSTOP>
<tmi8:dataownercode>CXX</tmi8:dataownercode>
<tmi8:lineplanningnumber>V102</tmi8:lineplanningnumber>
<tmi8:operatingday>2016-03-10</tmi8:operatingday>
<tmi8:journeynumber>11</tmi8:journeynumber>
<tmi8:reinforcementnumber>0</tmi8:reinforcementnumber>
<tmi8:userstopcode>72900013</tmi8:userstopcode>
<tmi8:passagesequencenumber>0</tmi8:passagesequencenumber>
<tmi8:timestamp>2016-03-10T10:52:46+01:00</tmi8:timestamp>
<tmi8:source>VEHICLE</tmi8:source>
<tmi8:vehiclenumber>5564</tmi8:vehiclenumber>
<tmi8:punctuality>0</tmi8:punctuality>
</tmi8:ONSTOP>
```

Onroute

A vehicle is currently driving on the planned route

```
<tmi8:ONROUTE>
<tmi8:dataownercode>CXX</tmi8:dataownercode>
<tmi8:lineplanningnumber>M340</tmi8:lineplanningnumber>
<tmi8:operatingday>2016-03-10</tmi8:operatingday>
<tmi8:journeynumber>66</tmi8:journeynumber>
<tmi8:reinforcementnumber>0</tmi8:reinforcementnumber>
<tmi8:userstopcode>55008510</tmi8:userstopcode>
<tmi8:passagesequencenumber>0</tmi8:passagesequencenumber>
<tmi8:timestamp>2016-03-10T10:52:46+01:00</tmi8:timestamp>
<tmi8:source>VEHICLE</tmi8:source>
<tmi8:vehiclenumber>3891</tmi8:vehiclenumber>
<tmi8:punctuality>-29</tmi8:punctuality>
<tmi8:distancesincelastuserstop>955</tmi8:distancesincelastuserstop>
<tmi8:rd-x>103342</tmi8:rd-x>
<tmi8:rd-y>487061</tmi8:rd-y>
</tmi8:ONROUTE>
```

Offroute

A vehicle is currently driving, but does not follow the planned route

```
<tmi8:OFFROUTE>
<tmi8:dataownercode>CXX</tmi8:dataownercode>
<tmi8:lineplanningnumber>B103</tmi8:lineplanningnumber>
<tmi8:operatingday>2016-03-10</tmi8:operatingday>
<tmi8:journeynumber>1036</tmi8:journeynumber>
<tmi8:reinforcementnumber>0</tmi8:reinforcementnumber>
<tmi8:timestamp>2016-03-10T10:52:54+01:00</tmi8:timestamp>
<tmi8:source>VEHICLE</tmi8:source>
<tmi8:userstopcode>51300100</tmi8:userstopcode>
<tmi8:passagesequencenumber>0</tmi8:passagesequencenumber>
<tmi8:vehiclenumber>4227</tmi8:vehiclenumber>
<tmi8:rd-x>125712</tmi8:rd-x>
<tmi8:rd-y>455201</tmi8:rd-y>
</tmi8:OFFROUTE>
```

Appendix B: real-time bus locations script

This is the full script used to collect the real-time bus locations.

```
#Inspiratie via https://github.com/StichtingOpenGeo/Koppelvlakken

# arend ligtenberg
# March 2016
# version 1.0
#!/usr/bin/env python2

#---------------------------------------------------------------------

from gzip import GzipFile
from cStringIO import StringIO
import zmq
import xml.etree.cElementTree as ET
from datetime import datetime
from datetime import date
from time import *


def get_elem_text(message, needle):
    ints = ['journeynumber', 'reinforcementnumber', 'passagesequencenumber',
'vehiclenumber', 'punctuality', 'blockcode', 'numberofcoaches',
'distancesincelastuserstop'] #, 'rd-x', 'rd-y'

    elem = message.find('{http://bison.connekt.nl/tmi8/kv6/msg}'+needle)
    if elem is not None:
        if needle in ints:
            if (needle == 'rd-x' or needle == 'rd-y') and elem.text == '-1':
                return None
            else:
                return int(elem.text)
        elif needle == 'wheelchairaccessible':
            return elem.text == 'ACCESSIBLE'
        else:
            return elem.text
    else:
        return elem

def parseKV6(message, message_type, needles=[]):
    result = {'messagetype': message_type}
    for needle in needles:
        result[needle.replace('-', '_')] = get_elem_text(message, needle)
    return result

def stripschema(tag):
    return tag.split('}')[-1]

def fetchfrommessage(message):
    global routetype
    message_type = stripschema(message.tag)
    required = ['dataownercode', 'lineplanningnumber', 'operatingday',
'journeynumber', 'reinforcementnumber', 'timestamp', 'source']
    #if message_type == 'DELAY':
    #    return parseKV6(message, message_type, required + ['punctuality'])
    #elif message_type == 'INIT':
    #    return parseKV6(message, message_type, required + ['userstopcode',
'passagesequencenumber', 'vehiclenumber', 'blockcode', 'wheelchairaccessible',
'numberofcoaches'])
    #elif message_type in ['ARRIVAL', 'ONSTOP', 'DEPARTURE']:
    #    return parseKV6(message, message_type, required + ['userstopcode',
'passagesequencenumber', 'vehiclenumber', 'punctuality'])
```

```
    if message_type == 'ONROUTE':
        routetype = 1
        return parseKV6(message, message_type, required + ['userstopcode',
'passagesequencenumber', 'vehiclenumber', 'punctuality',
'distancesincelastuserstop', 'rd-x', 'rd-y'])
    elif message_type == 'OFFROUTE':
        routetype = 2
        return parseKV6(message, message_type, required + ['userstopcode',
'passagesequencenumber', 'vehiclenumber', 'rd-x', 'rd-y'])
    #elif message_type == 'END':
    #    return parseKV6(message, message_type, required + ['userstopcode',
'passagesequencenumber', 'vehiclenumber'])

    return None


def isInt(s):
    try:
        int(s)
        return True
    except ValueError:
        return False

def loadInDatabase(result):
    #print result


    if 'dataownercode'  in result:
        vervoerder = result['dataownercode']
    else:
        vervoerder = ""

    if 'timestamp' in result:
        tijd = result['timestamp' ]
    else:
        timestamp = ""

    if 'rd_x' in result:
        x = result['rd_x']
        y = result['rd_y']
    else:
        x = -1
        y = -1

    if 'journeynumber' in result:
        lijnnummer = result['journeynumber']
    else:
        lijnnummer = -1

    if isInt(x) and isInt(y):
         xInt = int(x)
         yInt = int(y)
         if xInt > xMin:
             if xInt < xMax:
                 if yInt > yMin:
                     if yInt < yMax:
                         print 'vervoeder: %s, lijn: %s, tijd: %s, x: %s, y:
%s, type: %s' % (vervoerder, lijnnummer, tijd, x, y, routetype)
                         result_string = '%s,%s,%s,%s,%s,%s \n' % (vervoerder,
lijnnummer, tijd, x, y, routetype)
                         target.write(result_string)


def exit(e):
    print "\n So Long and Thanks for All the Fish..."
```

```
    target.close()
    subscriber.close()
    context.term()
    log_file = 'ndovlog'+datetime.now().strftime('%Y%m%d-%H%M%S')+'.txt'
    logfile = open(log_file, 'a')
    logfile.write (e)
    logfile.close()

def main():
    print 'start\n'
    output_file = 'result_'+datetime.now().strftime('%Y%m%d-%H%M%S')+'.csv'

    global subscriber, context, target, xMin, xMax,yMin,yMax
    xMin = 127711
    xMax = 256238
    yMin = 416761
    yMax = 504921
    target = open(output_file, 'a')
    context = zmq.Context()
    subscriber = context.socket(zmq.XSUB)
    subscriber.connect("tcp://pubsub.ndovloket.nl:7658")
    subscriber.send(chr(0x01) + "/") # 0x01 = subscribe, 0x00 = unsubscribe

    while True:
            multipart = subscriber.recv_multipart()
            address = multipart[0]
            contents = ''.join(multipart[1:])
            contents = GzipFile('','r',0,StringIO(contents)).read()
            #print("[%s] %s\n" % (address, contents))
            try:
                xml = ET.fromstring(contents)
                #print xml
            except Exception, e :
                exit(e)
            if xml.tag == '{http://bison.connekt.nl/tmi8/kv6/msg}VV_TM_PUSH':
                posinfo =
xml.findall('{http://bison.connekt.nl/tmi8/kv6/msg}KV6posinfo')
                if len(posinfo) == 0:
                    print "das niks gedaan...."
                else:
                    #results = []
                    for dossier in posinfo:
                        for child in dossier.getchildren():
                            if child.tag !=
'{http://bison.connekt.nl/tmi8/kv6/core}delimiter':
                                result = fetchfrommessage(child)
                                if result is not None:
                                    #print result
                                    #results.append(result)
                                    loadInDatabase(result)


if __name__ == '__main__':
    try:
        main()
    except:
        exit("handmatige stop")
```

Appendix C: Create Interface 1 dataset

The Interface 1 dataset consists of public transport routes collected by public transport operators themselves. Public transport operators are obligated to deliver this data to the NDOV loket, where it can be downloaded as open data. This includes real-time data on departure times, information about rates and routes of public transport connections.

Data delivered by NDOV is for example used by Google, using their General Transit Feed Specification (GTFS) standard. The easiest way to extract the NDOV data is via this Google standard. A GTFS Route shapes Toolbox for ArcGIS makes it possible to convert the GTFS data to shapefiles.

The method to follow is:

1. Download GTFS data from NDOV Loket
2. Open ArcGIS toolbox “Display GTFS Route Shapes” and select the download directory, which includes a file shapes.txt with the actual bus routes (a CSV file with TXT extension)
3. Run the script and save the result in a geodatabase.

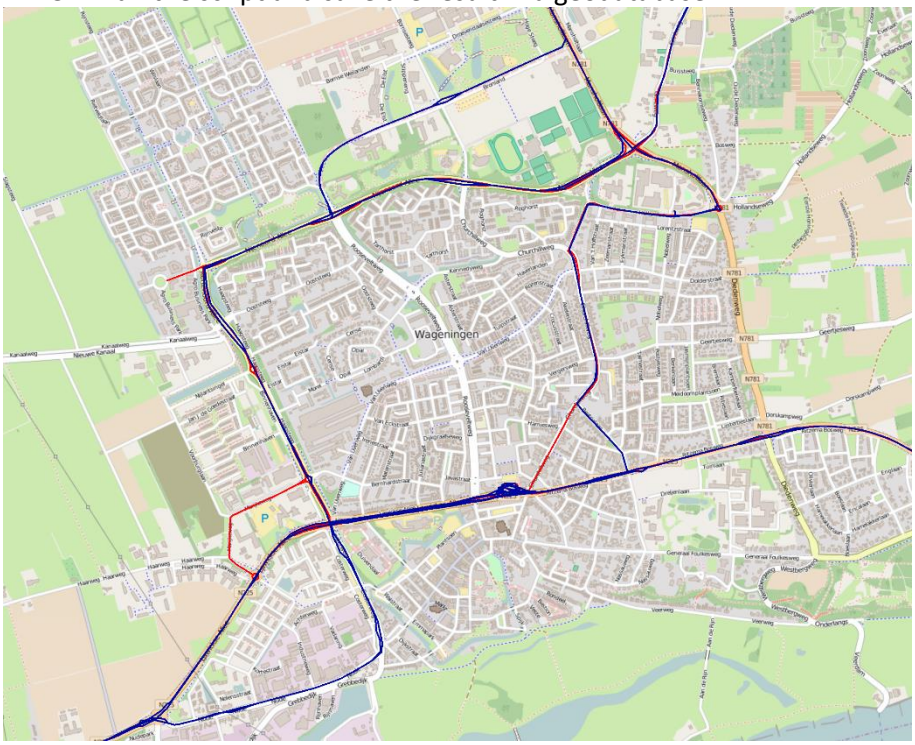
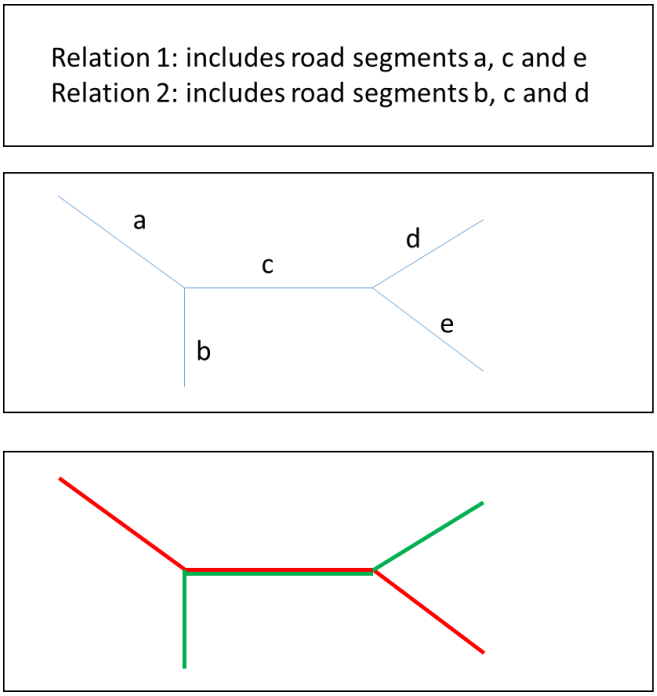


Figure C.1: Example of prepared dataset (OpenStreetMap data in red, NDOV data in blue)

Appendix D: Create OpenStreetMap dataset

To evaluate the spatial data quality of public transport routes in OpenStreetMap, a file has to be created consisting of all public transport routes in the study area. Via a widely used Java editor for OpenStreetMap, JOSM, it's possible to convert the items of a relation in OpenStreetMap to for example a GPX file. In this way, it's possible to manually create a dataset with all public transport routes in the area of interest. Attention has to be paid to the fact that relations of public transport routes in OpenStreetMap do not only consist of road elements but also include point data for bus stops and lines for platforms.

Every public transport route in OpenStreetMap is, in fact, a file describing all the roads involved in the route. The relation itself doesn't include the roads, only references to roads and is useless without information on the location of those roads (figure D.1a) However, the roads itself do not include information on bus routes involved. Directly saving a file from JOSM to a GPX, would remove all data on which bus line uses a specific road and will give only a dataset with roads where a bus runs over (figure D.1b). With the use of QGIS, it's possible to extract "multilinestrings", in fact, the OpenStreetMap relations with their associated road sections. The result is line data, very similar to the data available in the NDOV Loket (figure D.1c)



(a) OpenStreetMap relation information
Useless without road data

(b) OpenStreetMap road data
No information on bus routes included

(c) Creation of multilinestrings via QGIS

Figure D.1 multilinestrings in OpenStreetMap to represent bus routes

The method to follow is:

1. Download a bus route from the concession in JOSM (Java OpenStreeetMap Editor)
2. Download whole concession area (network relation) and check if all bus lines are included (some bus lines might be included in OpenStreetMap, but are not linked to the network relation)
3. Save as .osm file (formatted as XML document)
4. Open in QGIS. QGIS will recognize points (which are the bus stops), lines (which are the roads) and multilinestrings (which are the bus routes)
5. Export (1) points with tag “highway=bus_stop” and (2) multilinestrings to a geodatabase. This excludes all other point elements such as traffic lights and all other line elements such as platforms.