



Semi-automated data collection for vehicle kilometers by public transport buses by territory

2024-NL-TRANSPORT— SMP-ESS-2024-TRANSPORT-STATS-IBA

Judith Nijenhuis

Math Hamers

Workpackage 3, deliverable D3.1

Co-funded by EU Grant related to SMP-ESS-2024-TRANSPORT-STATS-IBA



Co-funded by the
European Union

Disclaimer: Co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or Eurostat. Neither the European Union nor the granting authority can be held responsible for them.

CBS Heerlen

CBS-weg 11

6412 EX Heerlen

P.O. Box 4481

6401 CZ Heerlen

+31 45 570 60 00

www.cbs.nl

project number

PR004120

SVV

2025

Index

1.	Introduction	4
2.	Data	5
2.1	Data sources	5
2.2	Data quality and quantity	6
3.	Method	9
3.1	Population	9
3.2	Demarcation of cross-border bus lines	9
3.3	Old method: manual processing of the WIKI page	9
3.4	New method: semi-automated data collection from the WIKI page	10
4.	Results	12
4.1	Comparing the old and the new method	12
5.	Conclusions	15
5.1	Benefits of the new semi-automatic tool	15
5.2	Suggestions for further improvement of the tool	15
5.3	Continuation of the project	15

1. Introduction

Since 2011 Statistics Netherlands has estimated vehicle kilometres for buses using a comprehensive register of odometer readings. Our method first calculates kilometres per day between pairs of odometer readings for each individual vehicle. This data is then used to estimate kilometres per year for each vehicle. We weigh the results to correct for vehicles for which insufficient odometer readings are available. From next year onwards (when our statistical redesign is completed) we will replace the weighting process with imputation. The end result is an estimate for the total vehicle kilometres of all Dutch buses, both in the Netherlands and abroad. This estimate then has to be split by territory (kilometres on national territory and abroad).

The aim of this research undertaken as part of the grant is to improve our method for estimating kilometers of public transport buses in the Netherlands and for Dutch public transport buses abroad. Unfortunately, border crossings are not registered by odometers. This means we have to rely on other data sources to estimate vehicle kilometres per territory. We use different sources for coaches and public transport buses. Coaches are used for international travel more often, whilst public transport buses mostly drive within the national borders. There are however a select number of international bus lines operating in or from the Netherlands, operated by Dutch, German and Belgian companies. For these lines, data is collected from a WIKI page on cross-border public transport. The WIKI page contains a complete and up-to-date list of all international bus lines including relevant information on timetables, routes, type of bus etc. A major drawback however was our way of processing this data: it had to be entered into an excel sheet manually. This was a time-consuming and error-prone process.

Using this grant, we were able to replace the manual work by semi-automated data collection. We now have tooling which automatically copies all relevant information from the WIKI page into an excel sheet. The number of manual operations has been drastically reduced. This has been a time-saver, and it has improved the data quality by eliminating errors.

This report discusses the WIKI as a data source, goes into detail about both our old manual processing method and the new semi-automatic tool that was developed using the grant. We also present the results: public transport bus kilometres per territory for 2023 and 2024, calculated using the new tool. The report concludes with recommendations and plans for the future.

2. Data

This chapter discusses the data sources used to estimate vehicle kilometres by territory for public transport buses. The quality and quantity of the sources are also covered.

2.1 Data sources

Two sources are used to estimate bus kilometres:

- The license plate register of the Dutch Road Authority. This is a registry on Dutch vehicles and contains a wide range of vehicle characteristics. This is used to determine the Dutch bus population.
- The kilometre registration of the Dutch Road Authority. This register contains all available odometer readings for Dutch vehicles.

These two sources are combined to estimate the total traffic performance for Dutch buses. We then use other sources to break these kilometres down by territory:

- For public transport buses: WIKI page on cross-border public transport
- For coaches: holiday surveys – both national and foreign, lodging accommodation statistics, incoming tourism statistics, coach occupancy rates, cost development index for group transport.

This report focuses on the international kilometres for public transport buses, in particular improved data collection from the WIKI on cross-border buses to improve our estimates by territory.

2.1.1 The cross-border public transport WIKI

The [cross-border public transport WIKI](#) contains a list of all international bus lines to and from the Netherlands. Per bus line, it contains detailed information about for example the route, timetable, and type of bus. The information is complete and up to date. We can establish this by cross-checking the information from the WIKI page with a secondary source from CROW (*Knowledge centre for infrastructure, public space and traffic and transport*), see figure 2. The WIKI also contains an archive of terminated or modified routes. This archive is useful when, for example, a bus route is terminated sometime during the reporting year. We can then use the archive to process the information of the bus route until the date of termination.

One might wonder why the WIKI is used instead of the secondary data source from CROW. This is because CROW only collects data on bus routes operated by Dutch transport companies. But we also need data on bus lines operated by foreign companies, and the WIKI includes all.

Figure 1 below shows an example of a WIKI page of an international bus route. Every bus line has an individual page on the WIKI. The section in the top right corner states the name of the transport company, which we use to determine the nationality of the bus line. It also contains the type of bus line (regular, night bus, school bus, on-call bus etc.) and the type of bus (make and model).

Directly below this section is the timetable, split by peak- and off-peak, evening, week- and weekend days. The timetable is used to determine how often a bus covers the kilometres on its route.

On the left there is a section containing the bus route. All stops are included, as is the border-crossing. The route is used to calculate the kilometres travelled per trip, on each side of the border using Google Maps.

At the bottom of the page (not visible in figure 1) is a final paragraph. If applicable, information regarding changes to the bus route (e.g. route, stops, type of bus) is provided here. These changes have to be processed in our overview from their starting date.

Figure 1: An example of an international bus line on the WIKI page

Pagina
Overleg
Lezen
Brontekst bekijken
Geschiedenis weergeven
Doorzoek OV in Nederland Wiki

Lijn 62 Maastricht Station - Tongeren Station

Lijn 62 rijdt van Maastricht via Vroenhoven, Riemst, Herderen, Millen en Berg naar Tongeren.

Dienstregeling

- Op schooldagen rijdt in de ochtendspits één extra rit in beide richtingen.
- De halte *Statensingel* in Maastricht wordt door De Lijn nog aangeduid met de oude naam *Frontensingel/Maagdendries*.
- De halte *Dokter Bakstraat/Brusselse Poort* in Maastricht wordt door De Lijn nog aangeduid met de oude naam *Dokter Bakstraat*.

Geldigheid OV-chipkaart

- In de bussen van De Lijn kan NIET met de OV-chipkaart worden betaald. Er kan een kaartje worden gekocht bij de chauffeur.

Route

Halte
Maastricht, Station
Maastricht, Mosae Forum/Centrum
Maastricht, Boschstraat/Maagdendries
Maastricht, Statensingel (<i>Frontensingel/Maagdendries</i>)
Maastricht, Koningin Emmaplein
Maastricht, Sint Annalaan
Maastricht, Via Regia/Sint Annadal
Maastricht, Via Regia/Brusselse Poort
Maastricht, Dokter Bakstraat/Brusselse Poort
Maastricht, Keurmeestersdreef Oost
Maastricht, Gewantmakersdreef Zuid
Maastricht, Javastraat
Maastricht, Tongerseweg Kerkhof
Maastricht, Tongerseweg Wolder
Grens Nederland - België
Vroenhoven, Grens
Vroenhoven, Krijtstraat
Vroenhoven, Busstation
Riemst, Klein Lafeltstraat
Riemst, Busstation

Buslijn 62 in Nederland en België

Bekijk andere lijnen 62

Vervoerder(s)	De Lijn
Onderaannemers	De Morgenstond, Heidebloem
Vervoerregio	Limburg
Lijnsoort	Streekbus
Lijnkleur	62
Materieel	Jonckheere Transit 2000 MAN Lion's City Mercedes-Benz Citaro LE Iveco Crossway LE VDL Citea SLE-120

Frequenties	
Splite	1x/uur
Dal	1x/uur
Avond (werkdagen)	1x/uur (tot 20:50)
Zaterdag	1x/uur
Zaterdagavond	1x/uur (tot 21:15)
Zondag	1x/uur
Zondagavond	1x/uur (tot 20:05)

Vragen? Zie onze OV Vraagbaak

Lijn 62 stopt bij de halte Mosae Forum/Centrum in Maastricht.

2.2 Data quality and quantity

This paragraph discusses the quality and quantity of the three data sources that were used: the license plate register, the kilometre registration and the cross-border public transport WIKI.

2.2.1 License plate register

The license plate register of the Dutch Road Authority is used to determine the population of buses. Overall, the quality of this register is good, apart from relatively rare occasional registration errors. The register unfortunately does not distinguish between coaches and public transport buses. Therefore, we derive this distinction based on the owner's company name and the make and model of the bus.

In terms of quantity, it is a complete register of all Dutch buses (+/- 10,000), of which about 60 percent are used in public transport. The register contains data on vehicles currently in active use, but also of vehicles in company stock, and vehicles which were scrapped or exported during the previous year.

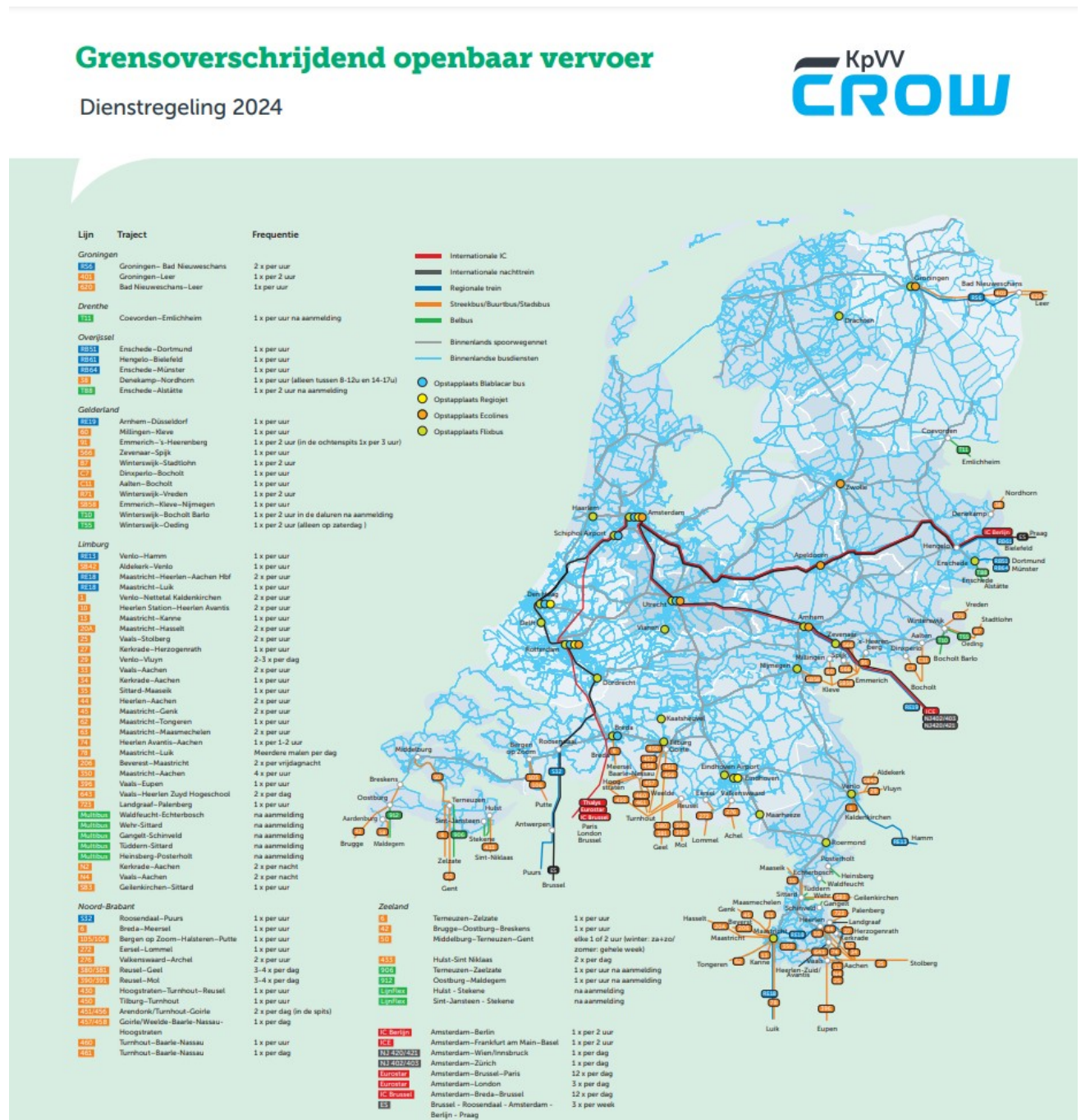
2.2.2 Odometer readings

Odometer readings are registered in the kilometre registration of the Dutch Road Authority. Overall, the quality of the register is good. It does however have certain data issues. Sometimes non-increasing odometer readings occur, and registration errors may lead to outliers. Also, for new vehicles there may be insufficient odometer readings available to estimate kilometres. But these types of issues are taken into account in our methodology. In terms of quantity: in 2024 this register contained sufficient odometer readings to estimate vehicle kilometres for 82,2 percent of all public transport buses.

2.2.3 The cross-border public transport WIKI

We use the cross-border public transport WIKI to estimate public transport bus kilometres by territory. The fact that we have to use this source instead of a register introduces a greater degree of uncertainty for our estimates. After all, the wiki is not a government administration, but a page maintained by volunteers. However, we do believe that the quality of the page is sufficient. We have cross-checked the bus lines on the WIKI with a secondary source from CROW. This cross-check showed that the overview of all cross-border lines on the WIKI is complete. The WIKI even contains some additional small on-call bus lines. One downside was that sometimes changes in routes and timetables were recorded on the WIKI page with a delay. In the event of changes to a bus route during the reporting year, for example a different route, different timetables, or if a line is discontinued, the wiki archive is consulted. The archive ensures that we can always find out what the old situation was like and when changes were made, so that we can properly incorporate all this information into our estimates regardless of when exactly we consult the WIKI.

Figure 2: Map of cross-border public transport by CROW



3. Method

3.1 Population

Vehicle kilometres are estimated for all Dutch buses that may have been on the road during the reporting year. This includes active vehicles, vehicles that have been discontinued (due to export or scrapping, among others) and vehicles that have been in the company stock.

Besides estimating how many kilometres Dutch buses drive on national territory and abroad, we also estimate kilometres for foreign buses on Dutch territory.

This report focuses on public transport buses. The distinction between coaches and public transport buses is derived using company name, type of owner, and vehicle characteristics like bus model.

We use the following definition for public transport buses: buses used for public passenger transport according to a timetable and for operating a scheduled service concession. This definition includes school buses that operate to a fixed time table and that are part of a scheduled service concession. It excludes a number of small on-call bus routes that are operated by taxi companies or volunteers. The definition also excludes international bus routes operated by e.g. Flixbus or Eurolines. Although these companies operate bus routes with a more or less fixed schedule, they are not part of a scheduled service concession.

3.2 Demarcation of cross-border bus lines

Which bus lines drive international routes? This is determined using information from the cross-border public transport WIKI, which contains detailed information on the bus routes. The route information states which border is crossed (either the German or the Belgian border). The WIKI also covers a few bus routes that were in fact no longer cross-border due to route alterations, these routes were of course excluded.

Cross-border bus lines which do not meet our definition of public transport (see the definition in paragraph 3.1) are excluded.

3.3 Old method: manual processing of the WIKI page

How did we collect and process data on international buses until now? The process consisted of 3 steps.

The first step was data collection. Once every year, we would check the WIKI on cross-border public transport bus lines. We would collect all relevant information: time tables, routes, the type of bus etc. and enter it into an excel sheet manually. The excel sheet was structured in such a way that it contained columns for all the necessary information. In practice, we would often take last year's excel sheet, and only update where necessary. This meant that most attention was focused on adding all the details of new bus lines and removing old lines. We also checked existing/continued bus lines to see if there were any changes. But when checking for updates on existing lines, a small change like an extra stop along the route or a change in frequency on Sunday afternoon was easily overlooked because we only did a visual inspection.

Besides directly copying data from the WIKI, some derivations had to be made. For example the nationality of the bus (derived from the name of the transport company). Or whether or not the type of bus used fit within our population of buses – so excluding on-call buses (this could be derived from the type of bus stated on the wiki). These derivations were done manually and everything was entered into the excel sheet.

Perhaps the most important derivation were the route kilometres per territory, that are not directly available on the WIKI. Although the WIKI lists every stop along the route, we still had to plot these stops in Google Maps

to calculate the length of the route on each side of the border. These kilometres were then entered into the excel sheet per bus line, to be used for further calculations.

Step two was processing all the collected data. This took place in the same excel sheet where the data from the WIKI had been entered. The excel sheet contained a lot of formulas to process the data. For example taking into account the nationality of the bus, the formulas in the excel sheet would automatically only fill kilometres for national or foreign buses. There was a formula that took into account the kilometres per territory and multiplied these with the time table and the number of weeks in which the bus was active, to calculate kilometres per territory. There were separate formulas for buses with a frequency per hour, and buses with a frequency per day.

Using these formulas could be a bit tricky: when adding new bus lines the right formulas had to be copied for your calculations to make sense – and mistakes were easily made.

Once the excel sheet was complete and kilometres had been calculated for every bus it was time for step 3: cross checking the results with last year's data. This helped to eliminate certain types of errors. For example, if an entire bus line was missing, it would immediately be noticed. But other types of mistakes would go unnoticed e.g. if you failed to update the time table for a specific bus.

The output of this process was an estimate of the total vehicle kilometres for public transport buses: of Dutch buses on foreign territory, and foreign buses on Dutch territory.

3.4 New method: semi-automated data collection from the WIKI page

Using this grant, we now have a tool which semi-automatically collects and processes data on cross-border bus lines from the WIKI. The tool was programmed in Visual basic for applications in Excel.

We start by providing the URLs of the cross-border bus lines to the tool, by placing them in a folder. The tool then loads the HTML code for all the pages (each page is an individual bus line). The tool then looks for specific elements in the pages, like the timetable and the type of bus. All collected data is automatically entered into an Excel spreadsheet for processing. This Excel sheet is similar to the one we already used in the old method, but now the tool automatically applies the correct formula's (for example the formula that calculates the kilometres based on the time table and route kilometres).

A limited number of tasks still has to be done by hand, for example placing the URLs of the cross-border bus routes in a folder. Unfortunately, an attempt to automatically plot the routes in Google Maps and read out the kilometres was unsuccessful. Too many errors were made when the locations of bus stops were automatically selected on the map. As a result, the automatically plotted routes were sometimes hundreds or thousands of kilometres off. Therefore, it was ultimately decided to continue plotting the routes manually – at least until these problems are fixed (see the recommendations in paragraph 5.1).

3.4.1 Improving the tooling

Of course, the first version of the data collection tool was not immediately perfect. When checking how the data was processed, and when analysing the results, we came across points of improvement. This paragraph addresses some of the problems we addressed whilst improving the tool.

Dealing with different types of buses: the first version of the tool could distinguish between regular line buses and on-call buses to determine whether a bus fell within our selected population or not. But other types of

lines could not be processed in the first version. The improved version could also process school bus lines (and immediately indicate that these were not running during weekends and holidays) and festival buses (only running for a week every year). The tool now also takes into account if a bus line changes from a regular bus line to an on-call bus line after a certain hour of the day.

Processing different notations in time tables: initially, the tool did not process every time table correctly. It was expected that every time table contained the number of departures per hour (eg. '2x/hour', or 1-2x/hour). The scraper had to be improved to correctly process notations like: 1x/2 hours or 2-3x/2,5 hours.

Calculating with start- and end dates: the WIKI contains information on the first and last day on which a bus line was active. The first version of the scraper had some trouble processing these dates, mainly because they are supplied in an irregular format. The improved version of the scraper could read different formats and use the dates to calculate the number of weeks in which the bus line was active during the reporting year.

Processing multi-nationality bus lines: most cross-border bus lines are operated by a single transportation company, either Dutch or foreign. But there are some exceptional cases where cross-border bus lines are alternately operated by a Dutch and a foreign public transport company. The tool can now identify these multi-nationality bus lines and automatically assign half the kilometres to Dutch buses and half to foreign buses.

An additional feature is that the tool now uses colour in the excel overview, which makes visual inspection a lot easier. For example, bus lines that are excluded from the population are given a separate colour. And cells that could not be filled because an error occurred are coloured red, indicating that action is required.

3.4.2 Validating the new tool

In order to validate the results of the new tool, the results for 2023 have been carefully compared with the results of the old manual processing of data from the wiki for reporting year 2023. For each bus line the estimates of the old and the new method were compared, and investigated if there was a difference. This led to improvement of the tool, discussed in the paragraph above. In graphs 1 and 2 below, both the results of the first version of the tool, and the final improved version are shown. But for most bus routes the tool offered improved data collection and processing, eliminating several small errors caused by manual processing in the past.

4. Results

Table 1 below shows vehicle kilometres of public transport buses by territory. The figures for 2018 – 2022 were estimated using our old manual processing method. The figures for 2023 and 2024 (in bold print) were estimated using the newly developed semi-automatic tool. The tool was used to improve the estimation of kilometres for Dutch buses abroad, and foreign buses in the Netherlands. The kilometres for Dutch buses in the Netherlands are just like before calculated by subtracting the kilometres of Dutch buses abroad from the total kilometres of Dutch buses. The total kilometres in the Netherlands are the sum of the kilometres of Dutch and foreign buses in the Netherlands.

Table 1. Vehicle kilometres of Dutch public transport buses by territory, 2018-2024

	Dutch buses			Foreign buses	Dutch and foreign buses
	In the Netherlands and abroad	In the Netherlands	Abroad	In the Netherlands	In the Netherlands
2018	470.778.285	469.607.455	1.170.830	1.710.468	471.317.923
2019	465.918.501	464.701.449	1.217.052	1.809.986	466.511.436
2020	435.239.862	434.418.350	821.513	1.607.719	436.026.069
2021	421.613.328	420.460.045	1.153.283	1.909.194	422.369.240
2022	396.981.098	395.889.392	1.091.705	1.977.538	397.866.931
2023	382.110.673	381.033.315	1.077.358	2.008.307	383.041.621
2024	405.100.648	404.191.221	909.428	1.989.126	406.180.346

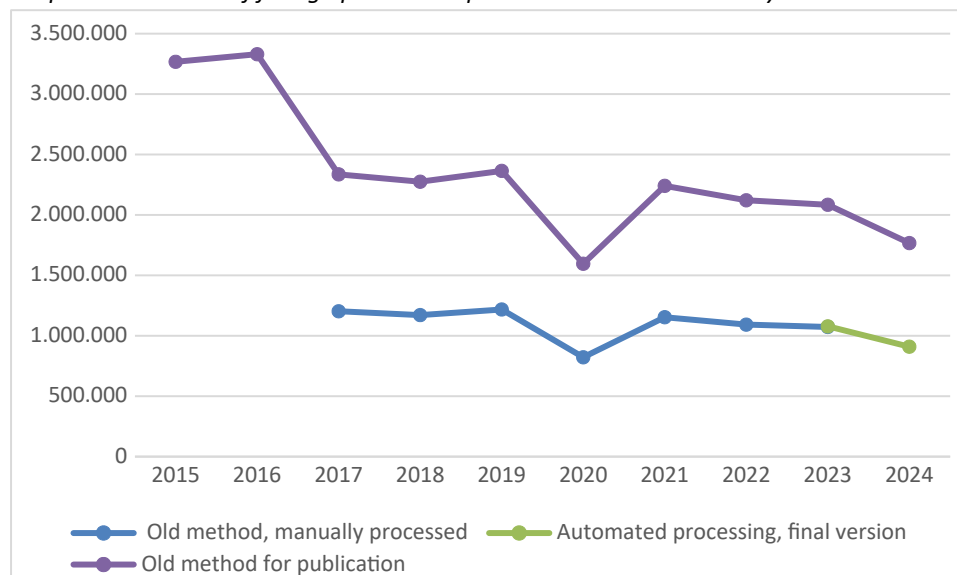
The figures show that the number of kilometres by foreign public transport buses in the Netherlands is approximately twice as high as the number of kilometres Dutch public transport buses drive abroad. The proportion of kilometres travelled by Dutch public transport buses abroad is very small (0.2% in 2024, 0.3% in 2023).

4.1 Comparing the old and the new method

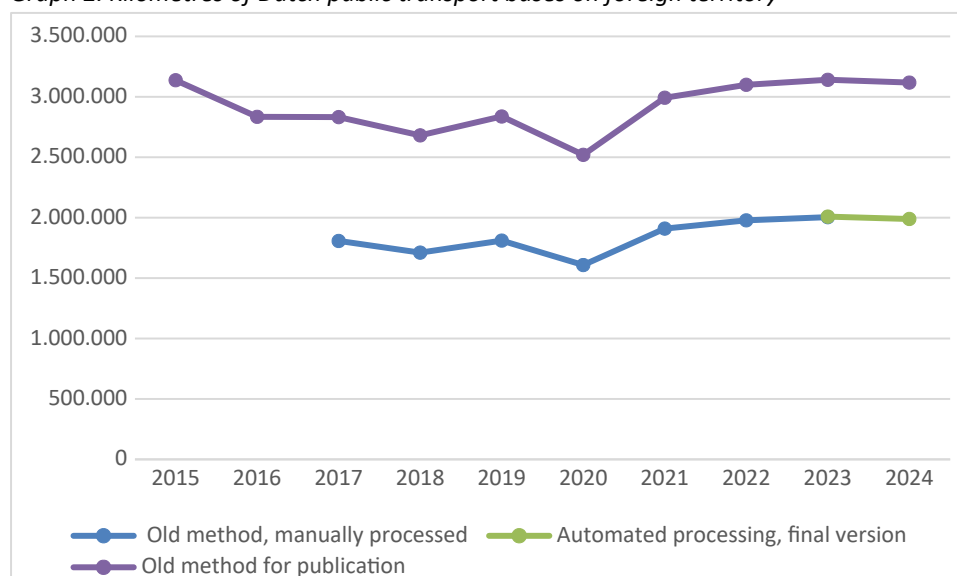
Graphs 1 and 2 below show kilometres of public transport buses by territory. Three versions have been included, we will go over the differences.

The blue line in the graphs below shows data based on the old method (manually collected and processed data from the WIKI page). The green line shows the data processed using the newly developed semi-automatic processing tool. As can be seen in the graphs, the level of the old manual estimate and the new level estimated using the new tool correspond well. There was a small difference (a few thousand kilometres), too small to be visible in the graphs.

Graph 1. Kilometres of foreign public transport buses on Dutch territory



Graph 2. Kilometres of Dutch public transport buses on foreign territory



Graphs 1 and 2 also show a purple line. Before 2017, Statistics Netherlands published public transport bus kilometres by territory, based on a previous version of the ‘old’ manual method. In 2017, we improved our method of data collection from the WIKI, and eliminated a number of errors. The result is presented in the graphs as the blue line. As can be seen in the graphs, the changes made a large difference. The level of the blue line is about half the level of the purple line. At the time, the level of the new method was expected to be

better, but we were not yet sure enough to switch to it. It would also have resulted in a major break in the series. Therefore, it was decided at the time to maintain the level of the old series (purple line) and update it with the year-on-year developments of the new series (blue line). The significant difference between the two versions led to considerable doubt on our part about the level of the old method (purple line) that we had always published. For this reason, no figures by territory have been published since 2017 on our online database StatLine. Except for an estimate of total kilometres on Dutch territory (by Dutch and foreign vehicles) that was absolutely necessary as input for the statistics on vehicle emissions.

The new tool that was developed using this grant has strengthened our confidence that the new, lower level of the estimates is indeed the better version (the blue line, and now the green). We intend to publish estimates at this level on StatLine, produced using the new semi-automatic tool in 2026. These will be published in the context of a new improved time series on vehicle kilometres that will be published in 2026 as part of our statistical redesign.

5. Conclusions

5.1 Benefits of the new semi-automatic tool

Using this grant, we have developed semi-automated tooling, in order to collect data on international buses from the cross-border bus line WIKI. This has resulted in three major gains:

First of all, improved data quality. Previously, the data was collected manually by copying relevant information from the WIKI into an excel spreadsheet. This was an error-prone method of data collection. Small mistakes like not registering an altered time table, or wrongly including a terminated bus line, impacted the kilometre estimates. The new tool simply copies all information, removing these types of errors.

The second gain: time saved. Previously, collecting and processing the data would collect approximately 3-4 days. Now, it can be done in half a day (running the script, checking the results and calculating kilometres).

And the final, and perhaps most important gain from this project: we can start publishing figures for public transport buses by territory again. Since 2017, no figures on public transport bus kilometres by territory have been published by Statistics Netherlands because we had serious doubts about the accuracy of our estimates.

We suspected that our previous estimate was far too high, but we were not sufficiently certain what the correct level was. Thanks to the new tool we have developed and all the validation research we have done during development, we are now convinced that the new level is correct. This gives us sufficient confidence in the quality to start publishing figures for public transport buses by territory again.

5.2 Suggestions for further improvement of the tool

Although we are content with the improvements this project has delivered, we also see opportunities for further improvements to the new tool. At the moment we are not sure when these will be addressed.

1. The new tool puts all the collected data in an excel sheet, which processes the data with formulas. We prefer not to use excel, and convert this step to processing in R. R could more easily be integrated into our new production system that is being built as part of our statistical redesign.
2. Although the new tool collects all data on international bus routes, the tooling is not yet successful in plotting all routes on a map to automatically calculate kilometres. For certain bus lines, the names of the bus stops yielded no results or wrong locations on the map, resulting in incorrect route kilometres. For now, this still has to be corrected by hand. We hope to perfect the plotting of routes at some point in the future. Automatically plotting routes is a technique we apply for other projects at Statistics Netherlands, so we can probably still make some progress by borrowing expertise within the organisation.
3. A minor improvement that could be made to the tool is the processing of bus routes that are discontinued during the year but are then restarted under the same route number. The tool can currently only process one bus with the same route number and name. In the rare event that this occurs, “discontinued” has to be added manually to the name of the old bus route.

5.3 Continuation of the project

Figures on public transport bus kilometers by territory, estimated using the new semi-automatic tool will be published on the Statistics Netherlands online database StatLine in 2026. The figures will be published as part of our new time series.

The semi-automatic processing tool has already become a part of our production process, that is currently part of a statistical redesign. The redesign includes improvements in several areas: from new tooling to improved methodology to new statistical output. One of the points addressed by the redesign is improving our sources and methods for the estimates of international vehicle kilometres. This grant has enabled us to make significant steps for the public transport buses. As a part of the redesign, we continue to look at improvements for other vehicle types as well. Work is currently underway to automate our method for calculating passenger car and coach kilometres by territory. These were also labour-intensive processes where a lot of work was done manually in Excel. We are also working to refine the method for goods vehicles and develop a more robust method for special vehicles. The experiences gained from this grant also contribute to those other improvement processes.