



Imputation Method For Passenger Cars

Author: Nardie Crijns

Reviewer: Hymke ten Have



**Co-funded by
the European Union**

Co-funded by EU Grant related to call SMP-ESS-20240TRANSPORT-STATS-IBA — Development of new transport statistics

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

SMP-ESS-2024-TRANSPORT-STATS-IBA — Development of new transport statistics

SVV

31 maart 2026

Index

1.	Summary	4
2.	Introduction	5
2.1	Goal	5
2.2	Missing Values	5
2.3	Challenge	6
3.	Data	7
4.	Method	10
4.1	Method Selection	10
4.2	Model Validation Strategy	10
4.3	Performance Metrics	11
4.4	Results	12
4.5	Results Missing Values Subgroups	16
5.	Future Steps	19

1. Summary

This document presents the challenge and solution of Statistics Netherlands on how to handle missing yearly vehicle kilometers of Dutch registered vehicles. The Traffic Performance statistic uses the insight of yearly vehicle kilometers, which is re-used by various other projects and statistics at Statistics Netherlands. The missing yearly vehicle kilometers are estimated using imputation. Eurostat and other NSO's face similar challenges when dealing with missing values. This document focuses on one approach to this problem by imputing missing values in accordance to the methods *mean imputation*, *median imputation*, *random hot deck* and *sequential hot deck*.

Two overarching methods are explored: imputing single values per strata and using a donor based imputation approach. The latter approach and specifically the *random hot deck* method displays the best results with regard to the group-level performance metrics. The missing yearly vehicle kilometers are simulated by splitting the response data into an 80% and 20% split for the train and test data respectively.

Next to the performance metrics, other visual validations are displayed to further inspire on the performance of the models. This should help the reader to explore other intuitive validation metrics when testing imputation methods in the statistical process.

The next steps revolve around bringing the *random hot deck* method into production. Not only for passenger cars, which this document focuses on, but for all vehicle types. Each vehicle type has its own selection of strata variables and categorical groups within this variable. In this way, the statistic Traffic Performance at Statistics Netherlands is improved by handling missing values in a scalable manner. This does not only provide a better result for the final statistic, but also for the underlying empirical data that is re-used by both internal and external researchers.

2. Introduction

2.1 Goal

One of the main statistics the Traffic and Transport (SVV) department of Statistics Netherlands publishes is the Traffic Performance statistic. This statistic is essential to publishing reliable statistics as it is used as input in many other statistics, like calculating vehicle emissions and additional statistical services to customers and other governmental agencies. SVV creates this statistic thrice for every reporting year: once with provisional data, once with further provisional data and finally with definitive data. The reasoning behind these recalculations are due the fact that SVV receives more odometer readings as time continues. The main insight of this statistic is to display the mean driven kilometers and the total driven kilometers of Dutch registered vehicles across various variables like *vehicle type*, *type of owner*, *fuel type*, *year of construction*, *regional division* or *business classification* (in case of a commercial vehicles subset). Additionally, the empirical data is used to examine trends and insights in various other projects and statistics. This document and supportive code repository focuses on Dutch registered passenger cars only for the sake of scope.

2.2 Missing Values

While many of the statistical process steps are essential to the publication of this statistic, the missing values of yearly vehicle kilometers and the imputation of these missing values are the focus of this document.

The odometer readings of passenger cars form the base of this statistic. Every passenger car that goes through inspection, (de-)registration, maintenance or any other form of checkup has its odometer read of which Statistics Netherlands receives data from the National Road Authority (RDW). In 2024, this process covers 91,3% of odometer readings for the passenger cars population. The other 8,7%, and a small additional group of passenger cars where no logical sequence is determined within the 91,3% response rate, are estimated by using a prediction.

Select subgroups can be recognized within Dutch passenger cars where often no response or illogical odometer sequences are observed. These are:

- Temporarily suspended passenger cars: In the Netherlands it is possible to temporarily suspend a vehicle for 1 to 3 years (although suspension periods can be consecutive and add up to longer periods of time). These vehicles are not allowed on the road and are often parked on private property for long periods of time. Although vehicles are allowed to drive to their periodic inspection whilst suspended, this rarely occurs. As a consequence, few odometer readings are available for temporarily suspended vehicles. Where odometer readings are available, the outcome is often that vehicles were stationary.
- Newly registered passenger cars, which are known as *young cars*, have significantly less odometer readings in the first 4 years. This is due to the fact that inspections and maintenance are not required. While some odometers are registered for these vehicles, unknown sequences in odometer readings appear more frequently in this subgroup.
- *Oldtimers* in the Netherlands abide to a special policy where vehicles between 30 and 50 years of age need only be checked once every two years. Vehicles older than 50 years are exempt from the mandatory inspection. As a result, SVV receives limited data on these vehicles meaning that non response is more common in this group.

2.3 Challenge

The old Traffic Performance methodology involves scaling the available vehicle kilometers to the entire vehicle population. This scaling is achieved using strata. Similar groups of vehicles are defined for each vehicle type, based on vehicle characteristics such as

- *Year of construction*
- *Weight class*
- *Fuel type*
- *Type of owner* (business or private)

Each individual vehicle with available yearly vehicle kilometers, based on the logical odometer readings, is assigned an increment factor which is based on a legacy method of the past. In the years between 2001 and 2011, Statistics Netherlands received only a sample of all registered odometer readings. Since 2012, all registered odometer readings are received, providing SVV with sufficient readings for over 85% of all vehicles to make an estimate, with up to 91,3% for passenger cars. Therefore, the conditions, assumptions on the data and the method are no longer appropriate to achieve the best result. As seen in section 2.2, some groups are poorly represented in the data. Especially for these groups, a new method for more reliable data is desired. Methodologists at Statistics Netherlands recommend imputation and Eurostat also indicates that imputation is the preferred method compared to using an increment factor.

In practice, SVV experiences the limitations of the old method as well. Regularly, requests are received to publish figures on various vehicle characteristics that are not included in our stratification. Other requests revolve around regional figures which are not included in the current method. As a result, the insights for these breakdowns are out of sync with the actual vehicle population and demand of customers. Finally, the databases are used by external researchers. It would be easier for them to work with a complete imputed dataset than with incremental factors.

The remainder of this document explains this challenge further starting with more details on the data at hand. Next, the current increment scaling method is explained. The approach and its fundamental shortcomings used in the statistical process for the Traffic Performance statistic are presented here. A solution to these shortcomings is proposed in chapter 4 where four alternative imputation methods are explored. The pros, cons and results are discussed per imputation method. Starting this year, future steps revolve around the deployment of the best performing model, the *random hot deck* method, into the statistical process. Other final remarks are presented at the end of this document.

3. Data

Two external data sources are used to estimate vehicle kilometers:

- The license plate register of RDW contains data on Dutch vehicles and a wide range of vehicle characteristics. This is used to determine the vehicle population. Vehicle characteristics like *vehicle type*, *fuel type* and *type of owner* that are used for the imputation method, originate from this dataset.
- The kilometers registration of RDW contains all available odometer readings for Dutch vehicles. Odometer readings are collected at a number of occasions, as seen in section 2.2.

For the analysis explained in this document, prediction is based on yearly vehicle kilometers of passenger cars for the year 2024. This year concludes 9.754.373 passenger cars of which 91,3% has valid yearly vehicle kilometers calculated in preceding steps by comparing odometer readings and status change of vehicles. The remainder of the vehicles are imputed by using the methods explained in chapter 4.

When estimating the value of yearly vehicle kilometers, only the most predictive explanatory variables are used. Various analyses have been done to support the choice between these variables, like the categories made within these variables and the sweet spot of the total explanatory variables per imputation method. During the analysis, different variables have been swapped in and out of the method to test not only individual prediction power, but also the correlation between variables. The amount of missing values is another point taken into account when choosing between explanatory variables for the method. Below in table 1, the explanatory variables are displayed.

<i>Explanatory variable</i>	<i>Number of groups</i>	<i>Included in the method as</i>
Fuel type	6	Yes, strata category
Year of construction	28	Yes, strata category *
Type of owner	3	Yes, strata category
Regional level	3	Yes, hierarchical category **
Manufacturing brand	3 – 6 clusters of categories	No
Number of cylinders	Continuous	No

Table 1: Explanatory Variables

* *Each individual year of construction is represented as its own group, except for old timers which have three separate groups clustered around a range of construction years*

** *Regional level: municipality, province (NUTS 2) and national (NUTS 0) levels are used as a hierarchical group to see on what level imputation may be done based on the imputation criteria*

While executing the analysis, separate functions and scripts are used to determine the most optimal split and number of categories for an explanatory variable. The predictive value of variables and the groups within these variables is determined. For example, 25 categories are used for the variable *Year of construction* per individual year between 2000 – 2024 while all years before 2000 are grouped into 3 separate categories concluding *old timers* with each having their own range of construction years. Additionally, the added value of including more variables to the imputation models are taken into account. *Manufacturing brand* and *number of cylinders* are examples that didn't make the final selection of variables. One of the main findings from the analysis is as more variables are added, the less influence they have on the independent variable.

Correlation of and shared predictive value of explanatory variables is another reason why the variable selection is capped at 4 variables. Finally, computation power of the application and comprehensibility of the model are the last arguments listed to limit the maximum amount of variables, and ultimately the total possible amount of strata within the method.

The imputation is executed based on the total of unique possible values of the strata used in the method. In practice a total of $6 * 28 * 3 = 504$ unique possible groups for imputation are used. 6 categories for *fuel type*, 28 categories for *year of construction* and 3 values for *type of owner*.

The fourth variable, *regional level*, is used to guide the level of the imputed independent variable and concludes the strata variables. This means that the possible values of *regional variables* (*municipality, province and national*) within the data are switched in and out of the imputation method based on the three criteria below. In other words, the *regional level* is used as a dynamic variable to the strata group and concludes the third source used in this project. In 2024, The Netherlands had 342 municipalities, 12 provinces and 1 national level. This means that when imputing on municipality level a total of $504 * 342 = 172.368$ unique strata groups are possible. Based on the size and response of these possible groups, the regional level for imputation is determined. In the most advantageous scenario the imputation is executed on municipality level, next on province and as a final resort the nationwide value is imputed for the 504 possible strata. Imputing on a lower regional level is beneficial as passenger cars in the same region drive similarly to one another. This is one of the findings of the explanatory variable analysis. Below are the rules used for the imputation criteria:

Municipality Level

Criteria 1: Group size (n) ≥ 30 and (n) < 50 and response rate $> 70\%$ - Imputation on municipality level

Criteria 2: Group size (n) ≥ 50 and response rate $> 50\%$ - Imputation on municipality level

Criteria 3: If none are applicable – Criteria 1 and 2 are checked again on province level of the strata

Province Level

Criteria 1: Group size (n) ≥ 30 and (n) < 50 and response rate $> 70\%$ - Imputation on province level

Criteria 2: Group size (n) ≥ 50 and response rate $> 50\%$ - Imputation on province level

Criteria 3: If none are applicable – Imputation on **National Level** for the strata

The value of *regional level* as the fourth strata variable is determined by the three criteria above. For example, all passenger cars within their strata groups are first matched to their corresponding municipality. *Criteria 1* and *2* are checked based on the size and response rate of the specific strata. If either *Criteria 1* or *2* is met, the imputation is based on the municipality level. If these are not met, the province value to which the municipality value belongs to is swapped with the municipality value. *Criteria 1* and *2* are checked again against the group size and response rate. Once more, if they are met, the imputation is based on the province level. If the criteria are not met, the national level is used for this specific strata. As an indication for passenger cars, roughly 95% of strata meet *criteria 1* or *2* on municipality level, meaning that imputation on municipality level is most often used.

The following example should make the response criteria and the relation to the strata imputation clear:

<i>Fuel type</i>	<i>Year of construction</i>	<i>Type of owner</i>	<i>Regional level</i>	<i>N</i>	<i>Response</i>
Diesel	2007	Company Stock	Municipality value	6	83,3%
Diesel	2007	Company Stock	Province value	41	90,2%

Table 2: Examples response criteria

The rows in table 2 depict the same strata with the combination *fuel type = Diesel, year of construction = 2007, type of owner = Company Stock*. The total amount of passenger cars within this group is equal to 6 with 5 of those having a response, indicating valid yearly vehicle kilometers (83,3%). The combination of this strata for this specific municipality does not satisfy *criteria 1* or *criteria 2*. Therefore, the municipality value is swapped for the provincial value to which this municipality belongs to and both *criteria 1* and *2* are checked again. In this instance, *criteria 1* for the province level is met since the group size (n) is greater than 30 but smaller than 50 as seen in row 2 of table 2. Additionally, the response rate is greater than 50% concluding a provincial wide imputation for this strata. This means that all other passenger cars with *fuel Type = Diesel, year of construction = 2007, type of owner = Company Stock* for the entire province are used to impute the missing values for yearly vehicle kilometers for this specific municipality strata. More information on how this outcome is implemented in the models is displayed in section 4.2.

4. Method

This chapter explains the four imputation models used in this project. The model validation and performance metrics are presented afterwards. Finally, the results are displayed per imputation model.

4.1 Method Selection

1. Mean Imputation

A straight forward but effective method is *mean imputation*. Here, the mean of the yearly vehicle kilometers is determined for each group and imputed for the missing vehicles of these groups. The imputation criteria are taken into account for this method, like all other methods, as seen in chapter 3.

2. Median Imputation

Similar to the *mean imputation*, the *median imputation* uses the same guidelines except that it imputes the median of the vehicles within the same strata group. The *median imputation* is added to the selection as it is less prone to outliers.

3. Random Hot Deck

The *random hot deck (RHD)* imputation method is a donor based imputation method. This means that missing values are replaced by values observed within the same strata group. Thus, *RHD* replaces a missing value for a similar observation within the group in a random fashion.

4. Sequential Hot Deck

The *sequential hot deck (SHD)* uses a step by step sequence of previously imputed values to determine the next imputation of missing values. Multivariate relationships are preserved in this method. This is especially useful when where multiple variables contain missing values, not only the dependent variable. Chapter 3 discussed using explanatory variables which consisted of none or few missing values. Thus, this method is less relevant, but a good benchmark for the *RHD* method and other use cases. Finally, the sequence is determined by the variable license plate. The logic to this sequence is flawed as the sequence is partly random and partly related to the date of issue. The sequence is viewed as a benchmark to the other methods.

4.2 Model Validation Strategy

For this project, the passenger cars with non-missing yearly vehicle kilometers are split into separate train and test sets with a 80 to 20 percent split. This results in 7.124.320 and 1.781.080 passenger cars in the train and test set respectively. The train set is used to train the actual models seen in the previous section which is then used to predict the values in the test set. The actual values of the test set are used as a reference on which the performance metrics are calculated. No vehicles are excluded from the train or test set except those with a higher daily driving distance of 500 kilometers per day. While these outliers are possible in theory, it is questionable if these distances are realistic in practice.

When imputing values for all four methods, the same strata variables are used across all possible combinations for *fuel type*, *year of construction*, *type of owner* and *Municipality*. The *regional level* and whether or not *criteria 1* or *2* are met determines what value is imputed (based on municipality, province or national level). Imputation is based on the daily vehicle kilometers of

passengers cars, which is the result of dividing the yearly vehicle kilometers by the amount of days the vehicle spent on the road. Initially, imputation is based on daily vehicle kilometers as it is possible for vehicles to be absent on parts of the year based on the vehicle status. For example, imported or exported vehicles, suspended vehicles or newly bought vehicles. For the final statistic, the yearly vehicle kilometers are reported by reversing the previous calculation by multiplying the daily vehicle kilometers by the amount of days spent on the road.

Mean or Median Imputation

- For *mean* and *median imputation*, if the addition of municipality value to the strata combination satisfies *criteria 1* or *2*, the mean or median of that specific group is imputed for the corresponding test set on municipality level.
- If the addition of municipality to the strata combination does not satisfy *criteria 1* or *2*, but does correspond to *criteria 1* or *2* when province is added instead of municipality, the mean or median of all passenger cars on province level for the strata is imputed for the strata combination with municipality level.
- If neither *criteria 1* or *2* is satisfied, the mean or median of all passenger cars that correspond to the values of the strata of the entire train subset is used to impute the test set.

Hot Deck Imputation

- For hot deck imputation, if the addition of the municipality value to the strata values satisfy *criteria 1* or *2*, *random* or *sequential hot deck* of that specific group is imputed for the test set on municipality level.
- If the addition of municipality to the strata combination does not satisfy *criteria 1* or *2*, but does correspond to *criteria 1* or *2* when province is added instead of municipality, *random* or *sequential hot deck* of all passenger cars on province level for the strata is imputed for the strata combination with municipality level.
- If neither *criteria 1* or *2* is satisfied, *random* or *sequential hot deck* of all passenger cars that correspond to the values of the strata of the entire train subset is used to impute the test set.

Concluding this section, it is good to note that one passenger car can be a donor for more than one passenger car in the test set. For example, it can be a donor for one passenger within its own strata and the same municipality and for another passenger car where it shares the same province or national level.

4.3 Performance Metrics

Four performance metrics are used to measure the results in order to validate the accuracy and correctness of the four imputation models:

1. Variance Ratio (VR): Measures how well the method retains the variance and distribution of the original data. A value closer to 1 indicates that the spread of the imputed values closely matches the observed data.
2. Relative Deviation (RD): Measures the relative difference between the imputed and observed values expressed as a proportion of the observed values in the test set. A value close to 0 indicates a smaller deviation and therefore higher accuracy.
3. R^2 (R-squared) measures the proportion of variance in the dependent variables, yearly vehicle kilometers, explained by the model. The benchmark is used to reflect the overall structure of the data not necessarily indicating small errors by the model.

4. RMSE (Root Mean Squared Error) measures the average prediction error of yearly vehicle kilometers while extra penalizing large errors due to the squaring of the result. This benchmark helps in reflecting large errors in the result.

The four performance metrics can be split into two groups:

- Group-level metrics: VR and RD calculate how well the aggregated variance and distributions are persevered on a group or strata level. These metrics focus on statistical consistency rather than individual row-level accuracy.
- Row-level metrics: R-squared and RMSE calculate the performance of the models in a row-level fashion. The metrics focus on the error or fit at an observational level.

For this analysis, group-level performance is considered superior to row-level performance as the final output and statistics are presented on an aggregated level. Therefore, even if models perform better overall on all metrics, the actual model that is preferred is the one that performance best on the group-level metrics. Micro data, row level data, is not the end goal of this document and ultimately the statistical or empirical output.

These benchmarks are used as not only prediction power is valuable, but also the variance of the yearly vehicle kilometers. The results on these benchmarks and why variance is valued next to accuracy is presented in the next section.

4.4 Results

Table 3 shows the results for each of the imputation methods across the four performance metrics. The results show a strong preference to the *RHD* imputation method compared to other models while focusing on the group-level performance metrics. While *mean imputation* performance best across all performance metrics, the *RHD* is the preferred method for the goals of this analysis.

Both the *mean* and *median imputation* present similar results in terms of the observational level performance metrics. The mean imputation further performs well on the group-level performance metric relative deviation. This is to be expected as imputing the mean will result in a small relative difference to the actual values on an aggregated level. The downside of imputing one value for all missing observations is seen in the variance ratio where the variance is underestimated and mitigates the underlying empirical distribution of the population. Because of this, the *mean* and *median imputation* are considered suboptimal. The downside of imputing one value per strata and the effect on variance is further emphasized by the similarity of the results of both the *mean* and *median imputation* models.

The *RHD* and *SHD* perform well on the group-level metrics, variance ratio and relative deviation, compared to *mean* and *median imputation* approaches. This indicates that hot deck methods better preserve the statistical distribution and group-level structure of the original data. With the sequence of the *SHD* being partly random, the performance of the *SHD* appears to be almost identical to *RHD*.

<i>Imputation Method</i>	<i>VR</i>	<i>RD</i>	<i>R²</i>	<i>RMSE</i>
Mean Imputation	0,25	0,05	0,240	25,820
Median Imputation	0,24	-11,46	0,238	26,206
Random Hot Deck	~1	~0	0,060	36,374
Sequential Hot Deck	~1	-0,06	0,067	36,078
<i>Interpretation</i>	<i>Closer to 1 is better</i>	<i>Closer to 0 is better</i>	<i>Higher is better</i>	<i>Lower is better</i>
<i>Calculation</i>	<i>Group-level</i>	<i>Group-level</i>	<i>Row-level</i>	<i>Row-level</i>

Table 3: Performance Metrics Results Per Imputation Method

To look at the results in another way, figure 1 displays the total yearly vehicle kilometers on the y-axis of the entire test set compared to the four models. The total of the test set is displayed in red, *RHD* in yellow, *SHD* in green, *mean imputation* in blue and *median imputation* in purple. Legend 1 on the following page can be used to interpret the results of figure 1, 2, 5, 6 and 7.



Legend 1: color coding test set and imputation models

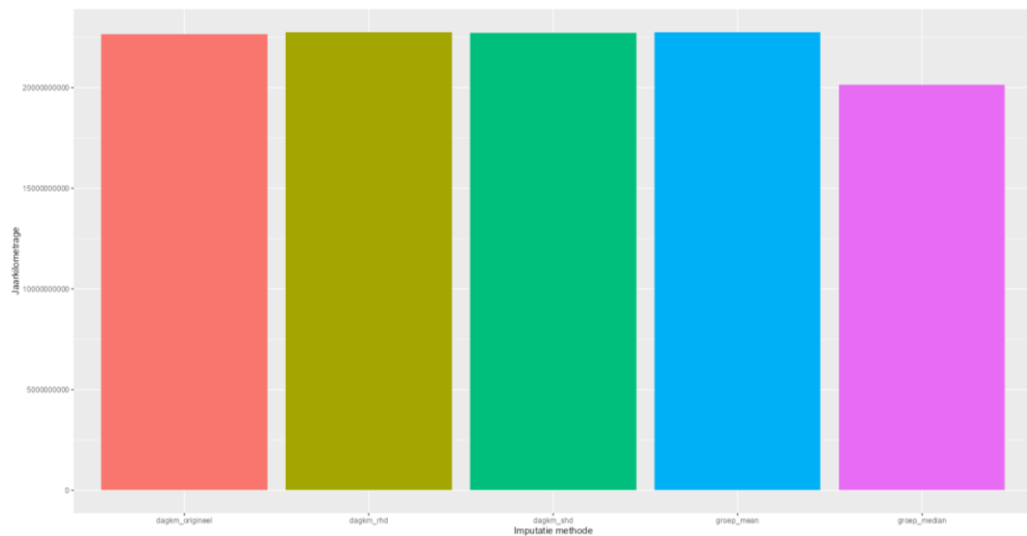


Figure 1: Total Yearly Vehicle Kilometers – In Support Of Relative Deviation

Compared to the total yearly vehicle kilometers (y-axis) of the test set (red), the most notable models are *RHD* (yellow) *SHD* (green) and *mean imputation* (blue). All three methods are closest compared to the actual yearly vehicle kilometers, see the RD results in table 3 for the exact differences. The *median imputation* is further off compared to the previous mentioned models. While this seems favorable for the *mean imputation* compared to the performance metrics, the problem with this estimate is noticeable in the next graph, figure 2. Here, the variance of the imputed values per model is compared to the variance of the test set. On the y-axis the daily vehicle kilometers is displayed. This image further support the results of the performance metrics indicating that the hot deck methods are better methods to achieve similar variance to the test set.

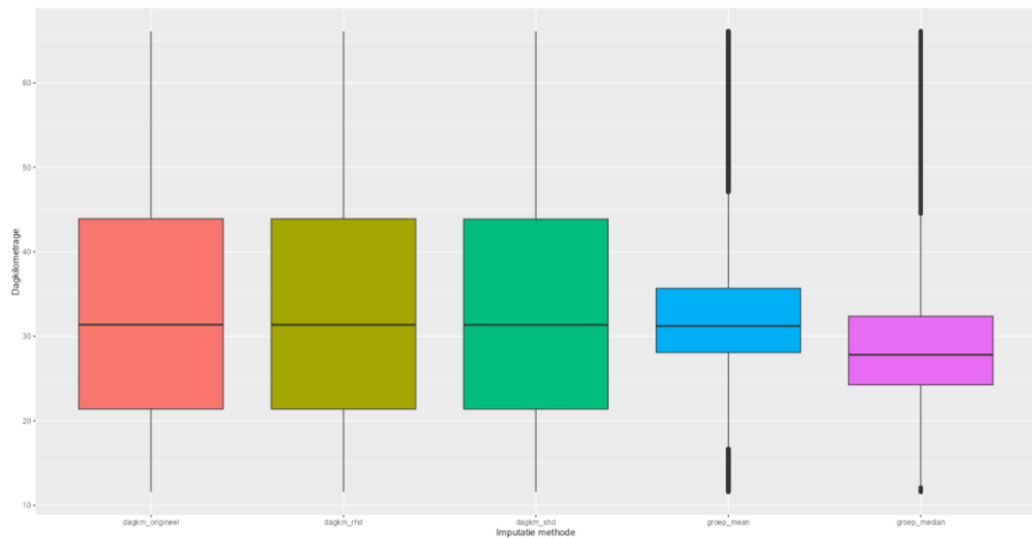


Figure 2: Boxplot Variance Daily Vehicle Kilometers – In Support Of Variance Ratio

The *mean imputation* (blue) is far off in terms of variance compared to the test set (red). While *RHD* (yellow), being random, displays a near identical distribution of values. While the total prediction of yearly vehicle kilometers for *mean imputation* might appear accurate on the surface in figure 1, SVV values the variance and confidence of the *RHD* method over the *mean imputation* for the following reasons:

- The empirical distribution of the population is maintained by distributing donor values randomly across the test set by using *RHD* instead of imputing one value per strata. This is relevant as other researchers and statistical analysts use this data across various other splits and groups to do their own analyses and research, allowing for better descriptive statistics.
- In practice, SVV notices a trend where clients and governmental agencies are increasingly more interested in regional data at the lowest level. While SVV cannot provide data on neighborhood and district data, due to no longer meeting the imputation criteria and other publication rules, the next lowest form of imputation is municipality level which is feasible in most cases. When choosing this method, it is key to note that strata groups become smaller and niche groups become more frequent. This means that the importance of individual vehicles increases for these smaller groups, indicating the importance of variance among these groups.

Figure 3 and 4 display the same overview as figure 1 and 2, but specified to only the variable *fuel type* and the methods *RHD* (yellow) and *mean imputation* (blue). Legend 2 is used to interpret the colors of figure 3 and 4. Here the downsides to imputing one value instead of a donor based method is further visualized. For figure 3, the groups with a high response, like *B* (gasoline) in the top-left and *D* (diesel) in the top-middle, display similar results. In contrast, imputing one value in a niche group like *X* (LNG and CNG) in the bottom-right performs worse than *RHD*. The same overview as figure 2 is again noticeable in figure 4 where the variance of *RHD* and *mean imputation* is compared to the test set. When diving deeper into actual strata level, this trend becomes even more apparent. Both images further support the arguments of the previous paragraph, especially the final point indicating that as groups become smaller the importance of individual vehicles increase.



Legend 2: color coding test set and imputation models

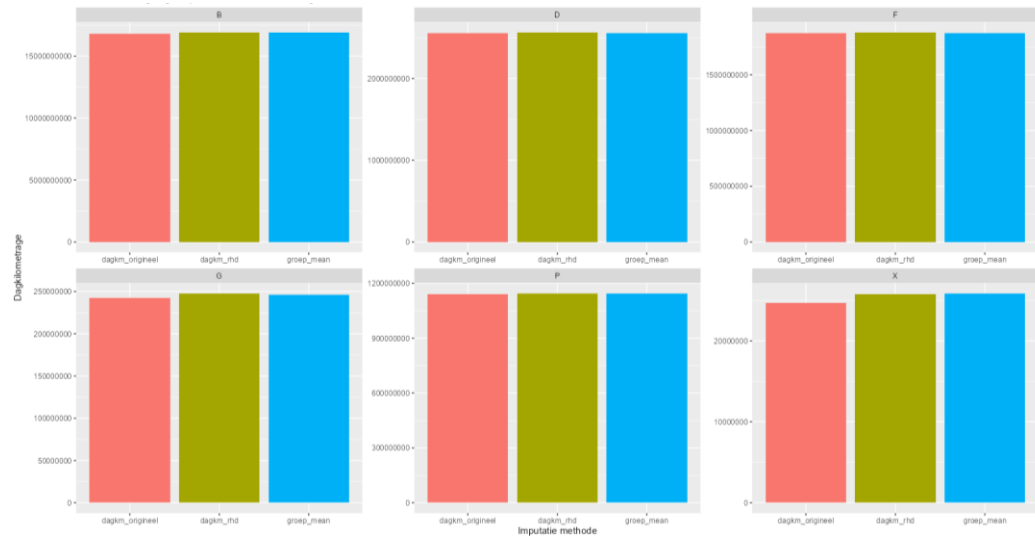


Figure 3: Total Yearly Vehicle Kilometers Per Fuel Type

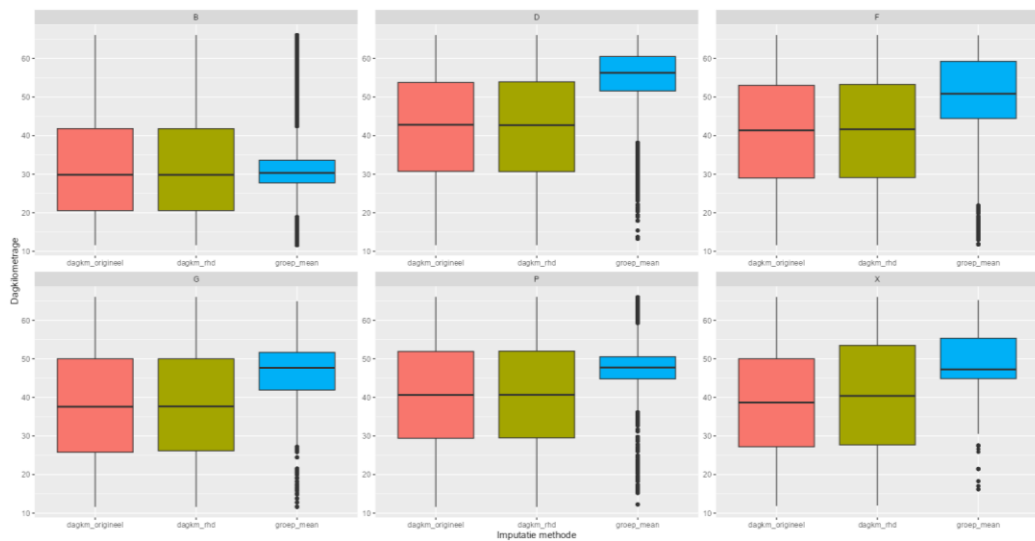


Figure 4: Boxplot Variance Daily Vehicle Kilometers Per Fuel Type

4.5 Results Missing Values Subgroups

In section 2.2, the recurring subgroups recognized within the missing values for yearly vehicle kilometers were discussed. In this final section of the results, *young cars* are displayed as a benchmark to the same visualizations and models presented before.

The conclusion for young cars is similar to those of other groups where missing yearly vehicle kilometers is more common. Again, the donor based imputation method RHD displays the best results. Even when the subgroups report significantly less odometer readings on which the yearly vehicle kilometers is based, this method continues to perform as seen previously. This conclusion is similar to the less common fuel types seen in figure 3 and 4. Thus, when groups become smaller or groups have less response, the individual importance of vehicles increases. This again increases the performance and variance of the imputed values by *RHD*. In the figures below, the test set (red), *RHD* (yellow), *SHD* (green), *mean imputation* (blue) and *median imputation* (pink) are displayed for vehicles younger than 1 year. In other words, passenger cars where year of construction is equal to 2024.

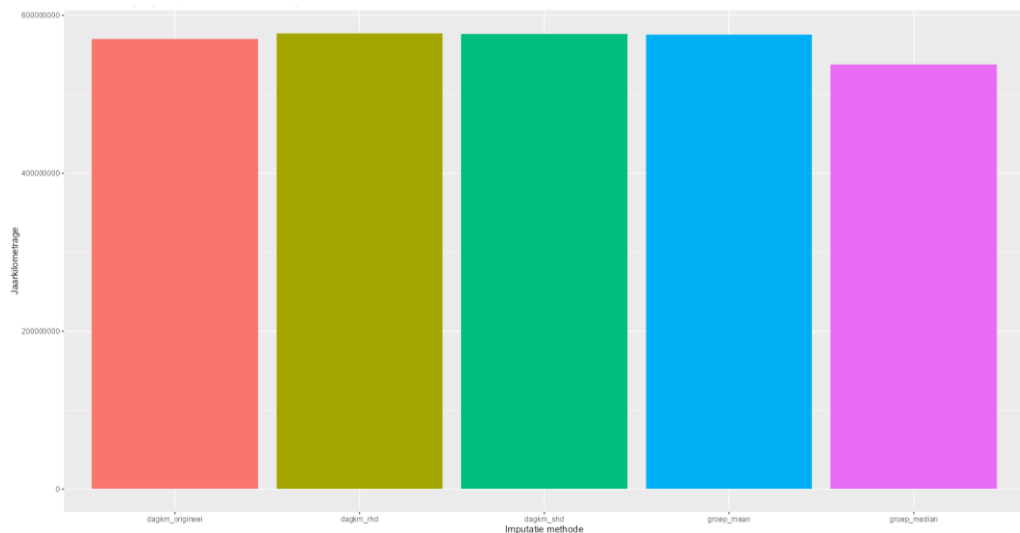


Figure 5: Total Yearly Vehicle Kilometers for passenger cars with a year of construction of 2024

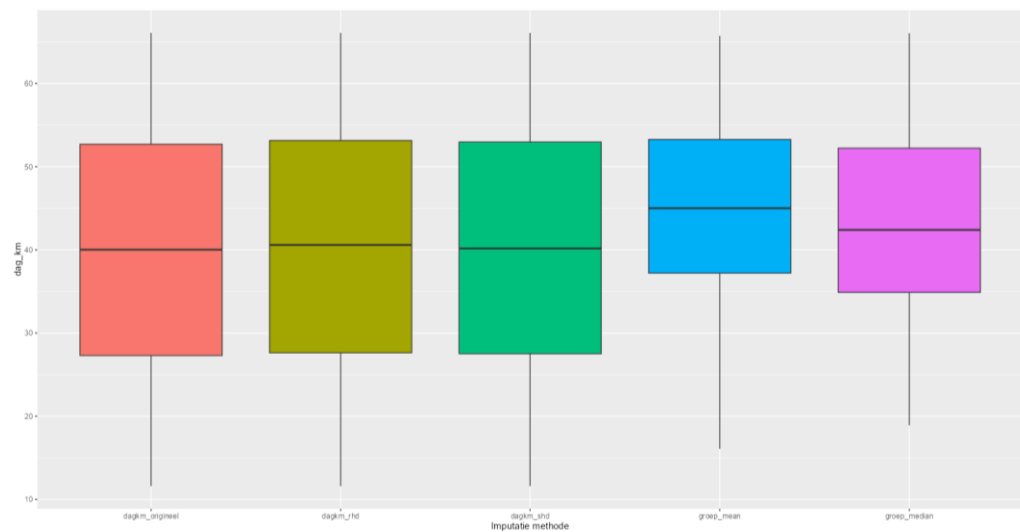


Figure 6: Boxplot Variance Daily Vehicle Kilometers for passenger with a year of construction of 2024

It should be noted that this figure paints a better picture than the actual reality. The train and test set are based on the response of vehicles where the Year of construction is equal to 2024. As seen in section 2.2, the group young cars contain more non response vehicles in proportion to other groups. The missing values in this group, which are not included in the train and test set, have their own characteristics. These characteristics are not represented in the imputation model e.g. young cars that do not have their odometer read. Thus, the response is a selective group within the entire group of young cars, which are divided into a train and test set . In other words, the non-response within 2024 is a subgroup of its own not registered by the RDW. This means that an accurate test set can only be created with more years of data, something which the current project lacks. Other years of data in accordance with the functional and technical design of the steps proceeding the imputation method still need to be created. This analysis has been executed in the past with other data that further support this conclusion. The results of this analysis shows that the boxplot of *mean* and *median imputation* appear similar to those displayed in figure 1 and 2, while *RHD* and *SHD* remain similar to the actual test set.

This trend is further displayed in the boxplots below in figure 7. Here the subsets of *Year of construction* between the years 2012 and 2023 are displayed. The trend mentioned in the previous paragraph is made clear here. As the *Year of construction increases* and the vehicles become younger:

- The total group size decreases which results in a better variance for the *mean* and *median imputation* compared to the test set
- The response rate decreases which results in a better variance for the *mean* and *median imputation* compared to the test set

The total origin of the trend seen in figure 7 is only partly due to these reasons. The total explanation is not yet completely investigated and would be part of future recommendations of this project.

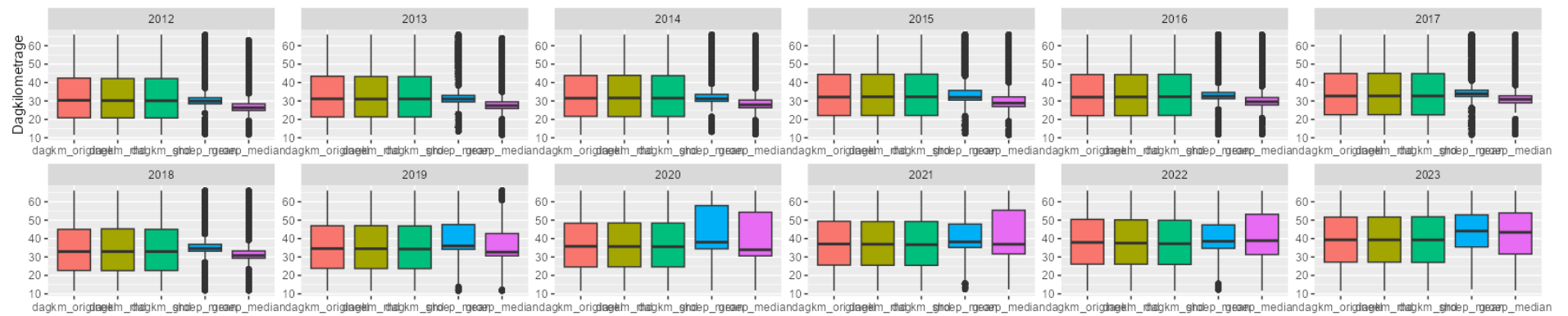


Figure 7: Boxplot Variance Daily Vehicle Kilometers for passenger cars where Year of construction is between 2012 and 2023

5. Future Steps

The future steps of this project are listed in this final chapter.

While the analysis of this document focuses on passenger cars, there are many more vehicle types distinguished at SVV. Each with its own variable selection for the strata and each with its own optimal categorical value split within the variables. For example, trucks use different explanatory variables compared to passenger cars and again different variables are used for buses. These need to be further explored within the project.

During the redesign of the Traffic Performance statistic, the *RHD* is included into the statistical process. For the first time, an imputation method is used in this statistic to predict actual missing values of the yearly vehicle kilometers. During the creation of the statistic of the following year, new odometer readings are incorporated to test the imputation of the previous years. Therefore, the next steps incorporated for this part revolve around using actual missing values instead of a predefined test set within the total response. The steps with regard to these developments are also next steps to implement in the statistical process.

Next, the process of imputation is packaged as a process step of the Traffic Transport statistic. This means the code is hosted on a server and receives input from the previous steps of the process. Finally, the output of the process provides input for the micro base where further statistics and analyses are executed. The hosting, technical implementation and positioning within the statistical process in production are still pending actions.

It should be noted that only four, of which two general approaches of imputation, methods are explored in this inquiry. Other models could be explored in a later stage, like the approaches of Random Forest or Generalized Linear Model. For this document, the scope allowed for a specific selection of models and the underlying analysis of variables selection, which is why no further methods are examined at this stage. Other technical implementations could have been chosen as well. For example, using k-fold cross-validation instead of a single split in train and test set or using more years of data to better estimate the yearly driving kilometers of *young cars*.